Appendix for "The Bigger Picture: Combining Econometrics with Analytics Improve Forecasts of Movie Success"

October 2020

Abstract

This is the online appendix for Lehrer and Xie (2020). Six sections are included, which provide further details on the data collection as well all of the econometric estimators and machine learning algorithms listed in table 2 of the main text including the newly proposed model averaging least squares support vector regresssion strategy. The appendix also contains all formal proofs of the econometric theory, review of related empirical literature focused on how online reviews influence revenue outcomes, and additional intuition explaining why the hybrid strategies yield improvements in forecast accuracy. Most importantly, all of the robustness exercises and additional results that are referenced in Lehrer and Xie (2020) are provided with a brief discussion. Last, a detailed study illustrates how the hybrid strategy can be used with a different algorithm to make splits in the tree structure to generate new empirical findings is provided.

JEL classification: C52, C53, D03, M21
Keywords: Machine Learning, Model Specification, Heteroskedasticity, Heterogeneity, Social Media, Big Data

A Review of Popular Machine Learning Tools for Forecasting

Algorithms in machine learning often build forecasting models by a series of data-driven decisions that optimize what can be learnt from the data to subsequently make predictions. Proponents of machine learning algorithms point to their improved performance in out of sample forecast exercises and stress the intuition on why they perform well, but do not consider their small sample or asymptotic properties.

The majority of machine learning tools used for forecasting implicitly assume homoskedastic data and ex ante we would expect their performance to deteriorate with heteroskedastic data. In this section, we summarize why we make this conjecture with six alternative strategies. First, estimates from the least absolute selection and shrinkage operator (Lasso) of Tibshirani (1996) is obtained by minimizing the l_1-penalized least squares criterion. Much research has investigated the model selection performance of the Lasso and found that it performs well under sparse and homoskedastic regression models. This result is unsurprising since the criterion involves the unweighted sum of squares and a penalty to make the model sparse. Thus, with heteroskedastic data and that objective function it may place more weight on high variance regions at the expense of low variance areas. Further, as some parameter estimates are shrunk relative to traditional OLS estimates, some omitted variable bias may arise.

Breiman, Friedman, and Stone (1984) introduced the classification and regression decision trees (CART). A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch (or tree leaf) represents the outcome of the test and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represents classification rules. In the application in the paper, we concentrate on the regression tree case, since our predicted outcomes are real number.

A regression tree (RT) recursively partition data into groups that are as different as possible and fit the mean response for each group as its prediction. The variable and splitting point are chosen to reduce the residual sum of squares (SSR) as much as possible after the split as compared to before the split.[1] That is, similar to stepwise regression the first split is akin to choosing which variable should be first included in the model. With regression trees, splits can continue within each subgroup until some stopping rule

[1] As mentioned in the main text, in RT, a node τ contains n_τ of observations. Each node can only by split into two leaves, denoted as τ_L and τ_R, each contains subsets of n_L and n_R observations with $n_\tau = n_L + n_R$. Define the within-node sum of squares as $\text{SSR}(\tau) = \sum_i^{n_\tau} (y_i - (\bar{y}_\tau)\bar{)}^2$, where $\bar{y}(\tau)$ is the mean of those cases. We split the n_τ observations of node τ into τ_L and τ_R if the following value reach its global maximum: $\Delta = \text{SSR}(\tau) - \text{SSR}(\tau_L) - \text{SSR}(\tau_R)$. Each tree leaf τ_L or τ_R can be treated as a new node and continue with the splitting process. We start from the top of the tree (full sample) and apply the same approach to all subsequent nodes. Once a tree is constructed, the full sample is split into a number of leaves. Each leaf contains a subset of the full sample and the accumulation of all leaves is the full sample.

is reached. This could lead to overfitting and as such, in practice the full trees are pruned using a cost-complexity criterion. This criterion takes into account the amount of squared error explained by each sub-tree plus a penalty chosen by cross-validation for the number of terminal nodes in the sub-tree in an attempt to trade-off tree size and over-fitting.

Forecasts from RT involve calculating the average of the associated observations of the dependent variable in each leaf calculated and treated as the fitted value of the regression tree. Hastie, Tibshirani, and Friedman (2009) provide evidence that in practice, predictions from RT have low bias but large variance. This variance arises due to the instability of RT as very small changes in the observed data can lead to a dramatically different sequence of splits, and hence a different prediction. This instability is due to the hierarchical nature; once a split is made, it is permanent and can never be "unmade" further down in the tree. Variations of RT have been shown to have better predictive abilities and we now briefly outline the procedures of two popular approaches known as bagging and random forest.

Bootstrap aggregating decision trees, or bagging, was proposed by Breiman (1996) to improve the classification by combining classifications of randomly generated training sets. Given a standard data set $\{y_i, \boldsymbol{X}_i\}$ with $i = 1, ..., n$, bagging generates B new training sets $\{y_i, \boldsymbol{X}_i\}^b$ for $b = 1, ..., B$, in which each set is a random sample of size n replacement from the original training set $\{y_i, \boldsymbol{X}_i\}$. By sampling with replacement, some observations may be repeated and for large n the set $\{y_i, \boldsymbol{X}_i\}^b$ is expected to have the fraction $(1 - 1/e) \approx 63.2\%$ of the unique examples of $\{y_i, \boldsymbol{X}_i\}$. Each data set will construct one regression tree that is grown deep and not pruned. In a forecasting exercise, we first obtain forecasts from each tree that similar to RT has a high variance with low bias. The final forecast takes the equal weight averages of these tree forecasts and by averaging across trees, the variability of the prediction declines. Much research has found that bagging, which combines hundreds or thousands of trees, leads to sharp improvements by over a single RT.

A challenge that bagging faces is that each tree is identically distributed and in the presence of a single strong predictor in the data set, all bagged trees will select the strong predictor at the first node of the tree. Thus, all trees will look similar and be correlated. The bias of bagged trees is identical to the bias of the individual trees but the variance declines even when trees are correlated as B increases.

To reduce the chance of getting correlated trees, Breiman (2001) developed the random forest method. Random forest is similar to bagging, as both involve constructing B new trees with bootstrap samples from the original data set. But for random forest, as each tree is constructed, we take a random sample (without replacement) of q predictors out of the total K^{total} ($q < K^{total}$) predictors before each node is split. This process is repeated for each node and the default value for q is $\lfloor 1/3K^{total} \rfloor$. In our application, we fixed q at specific numbers of explanatory variables to consider. Note that if $q = K^{total}$, random forest is equivalent to bagging. Eventually, we end up with B trees and the final random forest estimate is calculated as the simple average of forecasts from each tree.

Research has found that random forests do a good job at forecasting when the number of relevant variables in the set K is large. After all, if there are many irrelevant variables the chance of a split on something relevant becomes low. Yet, by randomly selecting predictors they produce trees with much lower degrees of correlation than bagging.

In the next section of the appendix, we consider four other methods including boosted regression trees that use a sequential process of fitting regression trees (without bootstrap sampling) to determine the weights of each tree in the forest. This relaxes the equal weight assumption implicit in the final forecast of random forest and bagging, but the method still relies on homoskedasticity in determining the initial splits at each node. We also describe in greater detail strategies based on Bayesian adaptive regression tree require researchers to assign priors including a functional form of the residual. A new strategy in that branch of the literature does account for heteroskedasticity.

We also describe in subsequent sections of the Appendix, the use (and discuss the empirical performance) of algorithms that use linear regression in the terminal leaves instead of estimating a local constant model. These methods include strategies that build tress using greedy algorithms as well as alternatives such as the M5' model tree method of Quinlan (1992) with an extension that allows for linear regression functions at the nodes and the scalable linear regression tree (SECRET) algorithm of Dobra and Gehrke (2002). Last, for space considerations, we did not consider two other methods developed in the machine learning literature. Artificial neural networks and multivariate adaptive regression splines also have algorithms that make decisions assuming homoskedasticity.[2] In summary, heteroskedastic data is not considered with many popular tools in the machine learning literature and its presence may bias the algorithms to operate in regions with higher variance at the expense of regions of low variance.

B Review of Alternative Machine Learning Strategies

In the previous section, we discussed the most commonly employed tree structured modeling techniques that recursively partition a data set into relatively homogeneous subgroups in order to make more accurate predictions on future observations. Yet, as described in the main text the CART induction algorithm of Breiman, Friedman, and Stone (1984) has received critiques in the machine learning literature related to greediness, split selection bias and the simplistic formation of prediction rules in the terminal leaf nodes. Our main text adds to this literature by first pointing out how heteroskedastic data results in split selection bias and by proposing an improvement that generalizes how prediction rules are made in the terminal leaves. In this section of the Appendix, we discuss prior

[2]Briefly, with artificial neural networks the weights for each node that correspond to different explanatory variables are estimated by minimizing the residual sum of squares; this approach is called back-propagation. With multivariate adaptive regression splines, terms are added to the regression model if they give the largest reduction in the residual sum of squares and to prevent over-fitting a backward deletion process is used to make the model sparse.

proposed algorithms developed in the machine learning literature to address the above three critiques. That said, we would like to reinforce that the addition of model averaging to the terminal leaves is new to this literature and in our application we witnessed large gains in forecast accuracy relative to BART and boosting and similar gains but with low additional computational costs relative to Bayesian approaches that explicitly consider heteroskedasticity. Further, in the next section of the appendix, we briefly show that our hybrid strategy also yields gains with these more computationally intensive algorithms.

B.1 Boosting Tree

Boosting provides a popular alternative ensemble method to random forests. Boosting is also based on multiple regression trees and no linearity assumptions are made about the DGP. To motivate boosting strategies, recall that random forests use a series of discontinuous flat surfaces forming an overall rough shape to approximate complex DGPs. The effectiveness of this strategy relies on interpolation allowing the rough shape to capture highly local features of the outcome data in a robust manner. Random forest predictions can substantially reduce the bias in fitted values relative to traditional econometric approaches due to their flexible modeling. This flexibility does come with a risk of overfitting the data since none of the trees in the forest are pruned.

Boosting is an alternative machine learning strategy developed to accommodate highly local features of the data and does not engage in overfitting. Boosting works in a sequential tree building manner that re-weights the residuals from the prediction of the first $K-1$ trees to create the Kth tree. Intuitively, this method produces a very accurate prediction rule by combining rough and moderately inaccurate rules of thumb. The reason why overfitting is unlikely is that boosted trees are typically grown small. The maximum depth of variable interactions is often set to be less than 4 or 5. With each individual tree growing to a low depth ensures that it cannot explain all of the variation in the data, and thereby allows the new trees in the sequence to "catch" the patterns that the previous ones missed. Put simply, observations with large residuals in the $K-1$th tree receive more weight in the construction of the Kth tree. The idea of boosting dates back to Valiant (1984) and the first algorithm was developed in Schapire (1990). In our application, we employ the gradient boost algorithm of Izenman (2013) to build boosted trees.

More formally, boosting is a way of fitting an additive expansion in a set of elementary basis functions. A basis function expansion takes the form

$$f(\boldsymbol{X}) = \sum_{k=1}^{K} \delta_k b(\boldsymbol{X}; \gamma_k),$$

where $\delta_k, k = 1, 2, ..., K$ are the expansion coefficients, and $b(\boldsymbol{X}; \boldsymbol{\gamma}) \in \mathbb{R}$ are usually simple functions of the multivariate argument $\boldsymbol{X}$, characterized by a set of parameters γ. Typically these models are fit by minimizing a loss function averaged over the training data,

such as the squared-error,

$$\min_{\{\delta_k,\gamma_k\}_1^K} \sum_{i=1}^{n} L\left(y_i, \sum_{k=1}^{K} \delta_k b(\boldsymbol{x}_i;\gamma_k)\right). \tag{A1}$$

Forward stagewise modeling approximates the solution to (A1) by sequentially adding new basis functions to the expansion without adjusting the parameters and coefficients of those that have already been added. This is outlined in the following algorithm:

Algorithm A.1.1: Forward Stagewise Boosting

1. Initialize $f_0(\boldsymbol{X}) = 0$.
2. For $k = 1$ to K:
   (a) Compute
   $$(\beta_k, \gamma_k) = \arg\min_{\beta,\gamma} \sum_{i=1}^{n} L\big(y_i, f_{k-1}(\boldsymbol{x}_i) + \delta_k b(\boldsymbol{x}_i;\gamma)\big).$$
   (b) Set $f_k(\boldsymbol{X}) = f_{k-1}(\boldsymbol{X}) + \delta_k b(\boldsymbol{X};\gamma_k)$.

At each new tree k, one solves for the optimal basis function $b(\boldsymbol{X};\gamma_k)$ and corresponding coefficient δ_k to add to the current expansion $f_{k-1}(\boldsymbol{X})$. This produces $f_k(\boldsymbol{X})$, and the process is repeated. Previously added terms are not modified.

For squared-error loss

$$L\big(\boldsymbol{y}, f(\boldsymbol{X})\big) = \big(\boldsymbol{y} - f(\boldsymbol{X})\big)^2, \tag{A2}$$

one has

$$\begin{aligned} L\big(y_i, f_{k-1}(\boldsymbol{x}_i) + \delta b(\boldsymbol{x}_i,\boldsymbol{\gamma})\big) &= \big(\boldsymbol{y}_i - f_{k-1}(\boldsymbol{x}_i) - \delta b(\boldsymbol{x}_i;\boldsymbol{\gamma})\big)^2 \\ &= \big(r_{ik} - \delta b(\boldsymbol{x}_i;\boldsymbol{\gamma})\big)^2, \end{aligned}$$

where $r_{ik} = \boldsymbol{y}_i - f_{k-1}(\boldsymbol{x}_i)$ is simply the residual of the current model on the i^{th} observation. Thus, for squared-error loss, the term $\delta_k b(\boldsymbol{x}_i;\gamma_k)$ that best fits the current residuals is added to the expansion at each step.

B.1.1 Gradient Tree Boosting

Regression trees partition the space of all joint predictor variable values into disjoint regions μ_j, $j = 1, 2, ..., J$, as represented by the terminal nodes of the tree. A local constant

model is then estimated to generate a prediction in each terminal node. A tree can be formally expressed as

$$T(\boldsymbol{X}, \Theta) = \sum_{j=1}^{J} \gamma_j \mathbb{I}(\boldsymbol{X} \in \mu_j),$$

with parameters $\Theta = \{\mu_j, \gamma_j\}_{j=1}^{J}$, where γ_j is the forecast in terminal node j. As with regression trees, γ_j is simply the mean outcome of all observations in j. Formally, the parameters are found by minimizing the risk

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{j=1}^{J} \sum_{\boldsymbol{x}_i \in \mu_j} L(y_i, \gamma_j).$$

The boosted tree model is a sum of all trees,

$$f_k(\boldsymbol{X}) = \sum_{k=1}^{K} T(\boldsymbol{X}; \Theta_k)$$

induced in the forward stagewise manner (Algorithm A.1.1) described in the previous subsection. At each step in the forward stagewise procedure one must solve

$$\hat{\Theta}_m = \arg\min_{\Theta_m} \sum_{i=1}^{n} L\big(y_i, f_{k-1}(\boldsymbol{x}_i) + T(\boldsymbol{x}_i; \Theta_k)\big). \tag{A3}$$

for the region set and constants $\Theta_k = \{\mu_{jk}, \gamma_{jk}\}_1^{J_k}$ of the next tree, given the current model $f_{k-1}(\boldsymbol{X})$.

Given the regions μ_{jk}, finding the optimal constants γ_{jk} in each region (i.e. leaf of the boosted tree) is typically straightforward:

$$\hat{\gamma}_{jk} = \arg\min_{\Theta_k} \sum_{i=1}^{n} L\big(y_i, f_{k-1}(\boldsymbol{x}_i) + \gamma_{jk}\big).$$

Finding the regions is difficult, particularly for a single tree. However, with squared-error loss, the solution is straightforward. It is simply the regression tree that best predicts the current residuals $y_i - f_{k-1}(\boldsymbol{x}_i)$, and $\hat{\gamma}_{jk}$ is the mean of these residuals in each corresponding region.

Fast approximate algorithms for solving (A3) with any differentiable loss criterion can be derived by analogy to numerical optimization.

$$\hat{\boldsymbol{f}} = \arg\min_{\boldsymbol{f}} L(\boldsymbol{f}),$$

where the "parameters" $\boldsymbol{f} \in \mathbb{R}^n$ are the values of the approximating function $f(\boldsymbol{x}_i)$ at

each of the n data points $\boldsymbol{x}_i$. The solution is a sum of component vectors

$$\boldsymbol{f}_K = \sum_{k=0}^{K} \boldsymbol{h}_k, \qquad \boldsymbol{h}_k \in \mathbb{R}^n,$$

where $\boldsymbol{f}_0 = \boldsymbol{h}_0$ is the initial guess, and each successive $\boldsymbol{f}_k$ is induced based on the current parameter vector $\boldsymbol{f}_{k-1}$. Each increment vector $\boldsymbol{h}_k = -\rho_k \boldsymbol{g}_k$, where ρ_k is a scalar and $\boldsymbol{g}_k \in \mathbb{R}^n$ is the gradient of $L(\boldsymbol{f})$ evaluated at $\boldsymbol{f} = \boldsymbol{f}_{k-1}$. The components of the gradient $\boldsymbol{g}_k$ are

$$g_{ik} = -\left[\frac{\partial L(y_i, f(\boldsymbol{x}_i))}{\partial f(\boldsymbol{x}_i)}\right]_{f=f_{k-1}}.$$

The step length ρ_k is the solution to

$$\rho_k = \arg\min_{\rho} \; L(\boldsymbol{f}_{k-1} - \rho \boldsymbol{g}_k).$$

The current solution is then updated

$$\boldsymbol{f}_k = \boldsymbol{f}_{k-1} - \rho_m \boldsymbol{g}_k$$

and the process repeated at the next iteration.

The following Algorithm A.1.2 summarizes the steps involved in the generic gradient tree-boosting algorithm for continuous outcomes. Specific algorithms can also be obtained by inserting different loss criteria $L(\boldsymbol{y}, f(\boldsymbol{X}))$. Notice that the first line of algorithm A.1.2 initializes the tree to be a global optimal constant model, which is just a single terminal node tree. The components of the negative gradient that is next computed as described in line 2(a) of algorithm A.1.2 are referred to as generalized or pseudo residuals, r. The procedure is more computationally expensive than bagging and generates shorter trees. The rationale is that allows new trees in the sequence can correct patterns that the previous trees missed, and this would not be possible if the trees were grown to be deep.

Algorithm A.1.2: Gradient Tree Boosting Algorithm

1. Initialize $f_0(\boldsymbol{X}) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$.
2. For $m = 1$ to K:
   (a) For $i = 1, 2, ..., n$ compute
   $$r_{ik} = -\left[\frac{\partial L(y_i, f(\boldsymbol{x}_i))}{\partial f(\boldsymbol{x}_i)}\right]_{f=f_{k-1}}.$$
   (b) Fit a regression tree to the targets r_{ik} giving terminal regions μ_{jk}, $j = 1, 2, ..., J_k$.
   (c) For $j = 1, 2, ..., J_k$ compute
   $$\gamma_{jk} = \arg\min_{\gamma} \sum_{\boldsymbol{x}_i \in \mu_{jk}} L(y_i, f_{k-1}(\boldsymbol{x}_i) + \gamma).$$
   (d) Update $f_k(\boldsymbol{X}) = f_{k-1}(\boldsymbol{X}) + \sum_{j=1}^{J_k} \gamma_{jk} \mathbb{I}(\boldsymbol{X} \in \mu_{jk})$.
3. Output $\hat{f}(\boldsymbol{X}) = f_K(\boldsymbol{X})$.

The number of trees grown is chosen via cross-validation and one can also use bagging samples or the full sample to conduct boosting. As noted in the prior section, each tree is built by minimizing the residual sum of squares and will face the same negative impact as heteroskedasticity. Each new regression tree is fit to the residuals of the predictions from the weighted sum of the previous trees and may correct for some heteroskedasticity. Yet, the short depth of these trees suggest that the consequences of heteroskedastic data can be severe since all the initial splits in boosted regression trees will be in high variance areas. Thus, the consequences of heteroskedastic data for boosting is that it may not consider potentially beneficial splits in low variance areas due to the objective function in the algorithm.

B.2 Bayesian Approaches: BART, BART$_{BMA}$ and H-BART

A second and increasingly popular alternative to random forest is Bayesian additive regression trees (BART) developed by Chipman, George, and McCulloch (2010). The popularity of BART arises from the possibility of undertaking statistical inference. This Bayesian approach has the flavor of a semiparametric model in econometrics and models the relationship between the outcome variable and explanatory variables as nonparametric, however a parametric residual enters the equation in an additively separable manner.

In this subsection, we review the conventional BART, BART under Bayesian model averaging (BMA) by Hernández, Raftery, Pennington, and Parnell (2018), and the heteroskedastic BART by Pratola, Chipman, George, and McCulloch (2019), in details.

B.2.1 Likelihood function for BART

Let's first introduce some notations, $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T$, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T$ and $\mathbf{x}_i = [\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \ldots, \mathbf{x}_{i,p}]^T$ for $i = 1, 2, \ldots, n$. Then the sum-of-trees model can be more explicitly expressed as

$$y_i = \sum_{j=1}^{m} g\left(x_i; T_j, M_j\right) + \epsilon_i, \quad \epsilon_i \sim N\left(0, \sigma^2\right) \tag{A4}$$

where m is the number of trees, for each binary regression tree T_j and its associated terminal node parameters $M_j = [\mu_{1j}, \mu_{2j}, \ldots, \mu_{n_j^g j}]^T$ with n_j^g as the number of leaf nodes in T_j, $g\left(x_i; T_j, M_j\right)$ is the function which assigns $\mu_{ij} \in M_j$ to x_i. From (A4), the BART model can be rewritten as

$$\mathbf{y}|\mathbf{X}, \mathcal{T}, \mathcal{M}, \sigma^2 \sim N(\sum_{j=1}^{m} J_j \mathbf{M}_j, \sigma^2 \mathbf{I}) \tag{A5}$$

where $\mathcal{T} = [T_1, T_2, \ldots, T_m]$, $\mathcal{M} = [M_1^T, M_2^T, \ldots, M_m^T]$, J_j is a $n \times n_j^g$ binary matrix whose (i, k) element denotes the inclusion of observation $i = 1, 2, \ldots, n$ in terminal node $i = 1, \ldots, n_j^g$ of tree T_j.

Then from (A5), the likelihood function of BART model can be specified as

$$p\left(\mathbf{y}|\mathbf{X}, \mathcal{T}, \mathcal{M}, \sigma^2\right) = \frac{1}{(2\pi)^{n/2}} \exp\left\{ \frac{\left(\mathbf{y} - \sum_{j=1}^{m} J_j \mathbf{M}_j\right)^T \left(\mathbf{y} - \sum_{j=1}^{m} J_j \mathbf{M}_j\right)}{2\sigma^2} \right\} \tag{A6}$$

where the parameters in BART model (A5) is defined as $(\mathcal{T}, \mathcal{M}, \sigma^2)$. We will specify the prior distribution for the parameters.

B.2.2 Priors for BART

The assumption of the prior distribution for $(\mathcal{T}, \mathcal{M}, \sigma^2)$ is that $\{(T_1, M_1), \ldots, (T_m, M_m)\}$ and σ are independent and that $(T_1, M_1), \ldots, (T_m, M_m)$ are independent of each other.

Then, the prior distribution can be written as

$$\begin{aligned} p\left[\mathcal{T},\mathcal{M},\sigma\right] &= p\left[\left(T_1,M_1\right),\ldots,\left(T_m,M_m\right),\sigma\right] = p\left[\left(T_1,M_1\right),\ldots,\left(T_m,M_m\right)\right]p(\sigma) \\ &= \left[\prod_{j=1}^{m} p\left(T_j,M_j\right)\right]p(\sigma) = \left[\prod_{j=1}^{m} p\left(M_j|T_j\right)p\left(T_j\right)\right]p(\sigma) \\ &= \left[\prod_{j=1}^{m}\left\{\prod_{i=1}^{n_j^g} p\left(\mu_{ij}|T_j\right)\right\}p\left(T_j\right)\right]p(\sigma). \end{aligned} \tag{A7}$$

For the entire tree T_j, $j = 1,2,\ldots,m$, the prior is specified as

$$\mathbb{P}(T_j) = \prod_{\eta^j \in H^j_{\text{terminals}}}\left(1-\mathbb{P}_{\text{SPLIT}}\left(\eta^j\right)\right)\prod_{\eta^j \in H^j_{\text{internals}}}\mathbb{P}_{\text{SPLIT}}\left(\eta^j\right)\prod_{\eta^j \in H^j_{\text{internals}}}\mathbb{P}_{\text{RULE}}(\eta^j) \tag{A8}$$

where η^j denotes the node (internal node and terminal node), $H^j_{\text{terminals}}$ denotes the set of terminal nodes and $H^j_{\text{internals}}$ denotes the internal nodes of T_j.

$\mathbb{P}_{\text{SPLIT}}\left(\eta^j\right)$ is the probability of splitting on a given node

$$\mathbb{P}_{\text{SPLIT}}\left(\eta^j\right) = \frac{\alpha}{\left(1+d_{\eta^j}\right)^{-\beta}} \tag{A9}$$

where d_{η^j} is the depth (number of parent generations) of node η^j, $\alpha \in (0,1)$ and $\beta \in [0,\infty)$. Following Chipman, George, and McCulloch (2010), the default values are $\alpha = .95$ and $\beta = 2$ in (A8). With this kind of prior, trees with $1,2,3,4$, and ≥ 5 terminal nodes receive prior probability of $0.05, 0.55, 0.28, 0.09$, and 0.03, respectively which puts most probability on tree sizes of 2 or 3, trees. $\mathbb{P}_{\text{RULE}}(\eta^j)$ is the probability of splitting rule at internal node η^j which consists of two parts

$$\mathbb{P}_{\text{RULE}}(\eta^j) = \mathbb{P}_{sx}(\eta^j)\mathbb{P}_{sxp}(\eta^j) \tag{A10}$$

where $\mathbb{P}_{sx}(\eta^j)$ is the distribution of split variable at η^j and $\mathbb{P}_{sxp}(\eta^j)$ is the distribution of splitting points conditional on selected split variable at η^j. Chipman, George, and McCulloch (2010) proposed uniform distributions for $\mathbb{P}_{sx}(\eta^j)$ and $\mathbb{P}_{sxp}(\eta^j)$, then

$$\mathbb{P}_{\text{RULE}}(\eta^j) = \frac{1}{p_{sx}^{\eta^j}} \times \frac{1}{p_{sxp}^{\eta^j}} \tag{A11}$$

where $p_{sx}^{\eta^j}$ and the number of available splitting variables at η^j and $p_{sxp}^{\eta^j}$ is the number of unique splitting points at η^j.

In our exercises, we set the tree prior $\mathbb{P}(T_j)$ in our paper as in (A8) with $\alpha = 0.95$, $\beta = 2$ and $\mathbb{P}_{\text{RULE}}(\eta^j)$ as in (A11). We tried other values and find that the results are more sensitive to α than β.

Chipman, George, and McCulloch (2010) proposed to use conjugate normal prior for $\mu_{ij}|T_j$

$$\mu_{ij}|T_j \sim N(\mu_\mu, \sigma_\mu^2), \tag{A12}$$

and then the induced prior of the condition expectation of y_i, $E[y_i|\mathbf{x}_i, \mathcal{T}, \mathcal{M}]$ can be written as

$$E[y_i|\mathbf{x}_i, \mathcal{T}, \mathcal{M}] \sim N(m\mu_\mu, m\sigma_\mu^2). \tag{A13}$$

It is highly probable that $E[y_i|\mathbf{x}_i, \mathcal{T}, \mathcal{M}]$ is between $y_{\min}$ and $y_{\max}$, the observed minimum and maximum of y_i in the data. Chipman, George, and McCulloch (2010) proposed to choose μ_μ and σ_μ so that $N\left(m\mu_\mu, m\sigma_\mu^2\right)$ assigns substantial probability to the interval $(y_{\min}, y_{\max})$. This can be conveniently done by choosing

$$\begin{aligned} m\mu_\mu - k\sqrt{m}\sigma_\mu &= y_{\min} \\ m\mu_\mu + k\sqrt{m}\sigma_\mu &= y_{\max} \end{aligned}$$

for some pre-selected value of k. The solution is as follows

$$\mu_\mu = \frac{y_{\max} + y_{min}}{2m}, \quad \sigma_\mu = \frac{y_{\max} - y_{\min}}{2k\sqrt{m}},$$

then (A12) can be rewritten as

$$\mu_{ij}|T_j \sim N\left(\frac{y_{\max} + y_{min}}{2m}, \left(\frac{y_{\max} - y_{\min}}{2k\sqrt{m}}\right)^2\right). \tag{A14}$$

It can be shown from (A14) that

$$\mathbb{P}\left[y_{\min} < E[y_i|\mathbf{x}_i, \mathcal{T}, \mathcal{M}] < y_{\max}\right] = \mathbb{P}\left[-k < \frac{E[y_i|\mathbf{x}_i, \mathcal{T}, \mathcal{M}] - m\mu_\mu}{\sqrt{m}\sigma_\mu} < k\right] = 2\Phi(k) - 1 \tag{A15}$$

where $\Phi(k)$ is the CDF of standard normal distribution. For example, $k = 2$, $k = 5$ and $k = 10$ would yield 95%, 99% and 100% prior probability that $E[y_i|\mathbf{x}_i, \mathcal{T}, \mathcal{M}]$ is in the interval $(y_{\min}, y_{\max})$ respectively. Following Chipman, George, and McCulloch (2010), we choose $k = 2$ in our paper. We carried out exercises with both $k = 5$ and $k = 10$, the results are qualitatively unchanged.

Chipman, George, and McCulloch (2010) also proposed to use a conjugate prior, the

inverse chi-square distribution for σ

$$\sigma^2 \sim \frac{\nu\lambda}{\chi^2_\nu}, \tag{A16}$$

where a data-informed prior approach is used to guide the specification of ν and λ. In this case, it aims to assign substantial probability to the entire region of plausible values of σ while avoiding over-concentration and over-dispersion. Given the value of $\nu \geq 3$, the value of λ is determined by

$$\mathbb{P}(\sigma < \hat{\sigma}) = q \tag{A17}$$

where q is some pre-specified probability such as 75%, 90% or 99%, $\hat{\sigma}$ is taken as the sample standard deviation of $\mathbf{y}$ or the residual standard deviation from a least squares linear regression of $\mathbf{y}$ on the original $\mathbf{X}$. Then the prior of σ is determined by ν and q. We follow Chipman, George, and McCulloch (2010) and set $\nu = 2$ and $q = 0.9$ in our paper.

B.2.3 Posterior Distribution for BART

The prior distribution would induce the posterior distribution

$$\begin{aligned} &\mathbb{P}\left[\left(T_1, M_1\right), \ldots, \left(T_m, M_m\right), \sigma | \mathbf{X}, \mathbf{y}\right] \\ &\propto p\left[Y | \left(T_1, M_1\right), \ldots, \left(T_m, M_m\right), \sigma\right] \times p\left[\left(T_1, M_1\right), \ldots, \left(T_m, M_m\right), \sigma\right] \end{aligned} \tag{A18}$$

which can be simplified into two major posterior draws using Gibbs sampling. First, draw m successive

$$\mathbb{P}\left[\left(T_j, M_j\right) | T_{-j}, M_{-j}, y, \sigma\right] \tag{A19}$$

for $j = 1, \ldots, m$, where T_{-j} and M_{-j} consist of all the tree structures and terminal nodes except for the jth tree structure and terminal node; then, draw

$$p\left[\sigma | \left(T_1, M_1\right), \ldots, \left(T_m, M_m\right), y\right] \tag{A20}$$

from $IG\left(\frac{v+n}{2}, \left\{v\lambda + \sum[y - f(x)]^2\right\}/2\right)$. This is the tree generating process by MCMC.

In this paper, we ran 100 burn-in draws and kept 1000 subsequent draws to represent the posterior.

B.2.4 Bayesian Additive Regression Trees using Bayesian Model Averaging (BART-BMA)

The likelihood function of BART-BMA is the same as BART in (A6). Following Hernández, Raftery, Pennington, and Parnell (2018), the tree prior and the σ prior are given by (A8)

and (A16) with $\nu = 3$, $q = 0.9$. But the prior for μ_{ij} is now specified as

$$\mu_{ij}|T_j \sim N\left(\frac{y_{\max} + y_{min}}{2m}, \frac{(y_{\max} - y_{\min})^2 \sigma^2}{a}\right). \tag{A21}$$

where $a = 3$ in our paper, which is the same as Hernández, Raftery, Pennington, and Parnell (2018).

The posterior likelihood of BART-BMA can be written as

$$\begin{aligned} &\mathbb{P}\left[\mathcal{T}, \mathcal{M}, \sigma|\mathbf{X}, \mathbf{y}\right] \\ &\propto \mathbb{P}\left[\mathbf{y}|\mathbf{X}, \mathcal{T}, \mathcal{M}, \sigma\right] \mathbb{P}\left[\mathcal{T}, \mathcal{M}, \sigma\right] \\ &= \mathbb{P}\left[\mathbf{y}|\mathbf{X}, \mathcal{T}, \mathcal{M}, \sigma\right] \mathbb{P}\left[\mathcal{M}|\mathcal{T}, \sigma\right] \mathbb{P}\left[\sigma|\mathcal{T}\right] \mathbb{P}\left[\mathcal{T}\right] \end{aligned} \tag{A22}$$

where $\mathbb{P}\left[\mathcal{M}|\mathcal{T}, \sigma\right]$ is specified by using (A21). By changing the prior of μ_{ij} from (A14) to (A21), we can integrate out the parameters $\mathcal{M}$ and σ from (A22) to get the marginal likelihood of $\mathbf{y}$ conditional on $\mathbf{X}$ and a set of trees $\mathcal{T}$

$$\begin{aligned} &\mathbb{P}\left[\mathbf{y}|\mathbf{X}, \mathcal{T}\right] \\ &= \int \mathbb{P}\left[\mathbf{y}, \mathcal{M}, \sigma|\mathbf{X}, \mathcal{T}\right] d\mathcal{M} d\sigma \\ &= \mathbb{P}\left[\mathcal{T}\right] \int \mathbb{P}\left[\mathbf{y}|\mathbf{X}, \mathcal{T}, \mathcal{M}, \sigma\right] \mathbb{P}\left[\mathcal{M}|\mathcal{T}, \sigma\right] \mathbb{P}\left[\sigma|\mathcal{T}\right] d\mathcal{M} d\sigma. \end{aligned} \tag{A23}$$

From (A23), the BIC for a set of trees $\mathcal{T}$ is

$$\text{BIC}_{\mathcal{T}} = -2 \log \mathbb{P}\left[\mathbf{y}|\mathbf{X}, \mathcal{T}\right] + B \log n \tag{A24}$$

where B is the number of parameters in $\mathcal{T}$. Instead of MCMC, BART-BMA uses a greedy tree growing algorithm (similar to classical CART) to generate $\mathcal{T}$ with the value of BIC into Occam's Window.

For grid search method, each variable $\mathbf{x}_p$ in data set $\mathbf{X}$ is split into grid gridsize $+ 1$ equally spaced partitions within the range of $\mathbf{x}_p$ and each partition value is then used as a potential split point. Increasing **gridsize** finds better solutions but makes the algorithm slower. Hernández, Raftery, Pennington, and Parnell (2018) proposed to select **gridsize** $= 15$ since it struck a good balance and gave good performance in most cases. Trees are greedily grown using the best percentage (denoted as "numcp") of the total splitting rules based on their residual squared error. Following Hernández, Raftery, Pennington, and Parnell (2018), numcp $= 20\%$.

As it is not possible to perform an exhaustive search of the model space especially when p is large, Hernández, Raftery, Pennington, and Parnell (2018) proposed to use a greedy and efficient version of BMA called Occam's window (Madigan and Raftery, 1994). Here only the models which fall within Occam's Window are selected using

$$\text{BIC}_k - \arg\min_{l \in \ell} \left(\text{BIC}_l\right) \leq \log(o) \tag{A25}$$

where ℓ indexes the sets of trees accepted into Occam's Window (the set of sum of trees models with the highest posterior probabilities to date).

In our paper, we set $o = 1000$ following Hernández, Raftery, Pennington, and Parnell (2018). Regarding the number of trees, Hernández, Raftery, Pennington, and Parnell (2018) claimed that $m = 5$ generally works well. However, empirical results under $m = 5$ are much worse than $m = 20$ in our exercises. In this paper, we set $m = 20$.

B.2.5 Heteroskedastic BART Via Multiplicative Regression Trees (HBART)

The HBART model can be specified as follows

$$y_i = \sum_{j=1}^{m} g\left(\mathbf{x}_i; T_j, M_j\right) + s\left(\mathbf{x}_i\right) z_i, \quad z_i \sim N\left(0,1\right) \tag{A26}$$

with $s\left(\mathbf{x}_i\right)$ specified as

$$s^2(\mathbf{x}_i) = \prod_{l=1}^{m'} h\left(\mathbf{x}_i | T_l', M_l'\right) \tag{A27}$$

where T_j, M_j are the same as in (A4), T_l' encodes the structure of the l-th tree for the variance and $M_l' = [s_{1l}^2, s_{2l}^2, \ldots, s_{n_l^h l}^2]^T$ with n_l^h as the number of leaf nodes in T_l'.

The prior for HBART is specified as below

$$\begin{aligned}
\mathbb{P}\left[\mathcal{T}, \mathcal{M}, \mathcal{T}', \mathcal{M}', \sigma\right] &= \mathbb{P}\left[\left(T_1, M_1\right), \ldots, \left(T_m, M_m\right), \left(T_1', M_1'\right), \ldots, \left(T_{m'}', M_{m'}'\right), \sigma\right] \\
&= \mathbb{P}\left[\left(T_1, M_1\right), \ldots, \left(T_m, M_m\right)\right] \mathbb{P}\left[\left(T_1', M_1'\right), \ldots, \left(T_{m'}', M_{m'}'\right)\right] \mathbb{P}(\sigma) \\
&= \left[\prod_{j=1}^{m} \mathbb{P}\left(T_j, M_j\right)\right] \left[\prod_{l=1}^{m'} \mathbb{P}\left(T_l', M_l'\right)\right] \mathbb{P}(\sigma) \\
&= \left[\prod_{j=1}^{m} \mathbb{P}\left(M_j | T_j\right) \mathbb{P}\left(T_j\right)\right] \left[\prod_{l=1}^{m'} \mathbb{P}\left(M_l' | T_l'\right) \mathbb{P}\left(T_l'\right)\right] \mathbb{P}(\sigma) \\
&= \left[\prod_{j=1}^{m} \left\{\prod_{i=1}^{n_j^g} \mathbb{P}\left(\mu_{ij} | T_j\right)\right\} \mathbb{P}\left(T_j\right)\right] \left[\prod_{l=1}^{m'} \left\{\prod_{k=1}^{n_l^h} \mathbb{P}\left(s_{kl}^2 | T_l'\right)\right\} \mathbb{P}\left(T_l'\right)\right] \mathbb{P}(\sigma).
\end{aligned} \tag{A28}$$

Following Pratola, Chipman, George, and McCulloch (2019), the priors $\mathbb{P}[T_j]$ and $\mathbb{P}[\mu_{ij}|T_j]$ are the same as in BART model. The prior of T_l', $\mathbb{P}[T_l']$ is also specified the same as in (A8). For the prior of s_{kl}^2, Pratola, Chipman, George, and McCulloch (2019) specified it as

$$s_{kl}^2 | T_l' \sim \frac{\nu' \lambda'}{\chi_{\nu'}^2}, \tag{A29}$$

then from (A27) and (A29), the prior of $s^2(\mathbf{x}_i)$ is

$$s^2(\mathbf{x}_i) \sim \prod_{l=1}^{m'} s_l^2 \tag{A30}$$

with $s_l^2 \sim \frac{\nu'\lambda'}{\chi^2_{\nu'}}$, i.i.d. Pratola, Chipman, George, and McCulloch (2019) proposed to choose a prior in the heteroskedastic model to match the prior in the homoskedastic case by matching the prior means. It can be shown that $E\left[\sigma^2\right] = \frac{\nu\lambda}{\nu-2}$, and $E\left[s(\mathbf{x}_i)^2\right] = \prod_{l=1}^{m'} E\left[s_l^2\right] = \lambda^{m'}\left(\frac{\nu'}{\nu'-2}\right)^{m'}$. Then ν' and λ' can be determined by separately matching the " λ piece" and the "ν piece" such as $\lambda' = \lambda^{\frac{1}{m'}}, \nu' = 2/\left(1-\left(1-\frac{2}{\nu}\right)^{1/m'}\right)$.

Following Pratola, Chipman, George, and McCulloch (2019), in our paper, we set $\nu = 10$ and λ to be the sample variance of $\mathbf{y}$ to specify the ν' and λ'. And we also set $m = 200$ and $m' = 40$. We ran 100 burn-in draws and kept 1000 subsequent draws to represent the posterior.

B.3 Linear Regression Tree Algorithms

All decision tree algorithms discussed above, base their forecasts on a set of piecewise local constant model. In this subsection, we first describe our implementation of model trees that estimate linear models in the leaf nodes. Numerous researchers in machine learning have developed algorithms[3] that estimate regression models in the leaf nodes to not just aid in prediction, but also simplify the tree model structure. That is, these researchers often suggest that the gains in prediction from using a piecewise linear model could allow one to grow shorter trees that are more parsimonious. Not surprisingly, ex ante from an econometrics perspective the success of these linear tree algorithms clearly depend on both the source and amount of heterogeneity in the underlying data.

Perhaps the best known of the linear regression tree algorithms is the M5 algorithm of Quinlan (1992) that was further clarified in the M5' algorithm of Wang and Witten (1997). The M5 algorithm builds subgroups using the same algorithm as RT (Breiman, Friedman, and Stone, 1984), but a multiple regression models is estimated in the terminal node. The model in each leaf only contains the independent variables encountered in split rules in the leaf node's sub-tree and are simplified to reduce a multiplicative factor to inflate estimated error.[4] In our application, we do not directly consider M5 but consider

[3] See Quinlan (1992), Chaudhuri, Lo, Loh, and Yang (1995), Kim and Loh (2003), Vens and Blockeel (2006), among others

[4] Note, we did not consider the extension by Torgo (1997) that undertakes non-parametric kernel regression in the terminal nodes since there are large computational costs. Our model averaging approach can be viewed as an approximation to the kernel regression and is easier to carry out since it involves simply estimating a suite of linear regression models.

two strategies that use RT to create subgroups subject to the restriction that each terminal node contains at least as many observations as the total number of explanatory variables included in a naive model in each leaf. This naive model is then estimated either by OLS or the LASSO and allows all the coefficients to vary across the terminal leaves.

We additionally examined the performance of the M5′ model tree that uses a different criteria to construct splits in the tree. Splits are based on minimizing the intra-subset variation in the output values down each branch. In each node, the standard deviation of the output values for the examples reaching a node is taken as a measure of the error of this node and calculating the expected reduction in error as a result of testing each attribute and all possible split values. The attribute that maximizes the expected error reduction is chosen. The standard deviation reduction (SDR) is calculated by

$$SDR = sd(S) - \sum_i sd(S_i) \times |S_i|/|S|,$$

where S is the set of examples that reach the node and S_is are the sets that result from splitting the node according to the chosen attribute (in case of multiple split). As usual, the splitting process will terminate if the output values of all the instances that reach the node vary only slightly or only a few instances remain.

Similar to M5 once the tree has been grown, M5′ estimates a multivariate linear model in each tree leaf that only includes variables that were used in the subtree of this node. Thus, the M5′ model tree is also analogous to using piecewise linear functions in each leaf.

A different algorithm that applies linear regression in the terminal nodes developed by Dobra and Gehrke (2002) has been named SECRET for Scalable EM and Classification based Regression Tree. SECRET differs from M5 and M5′ in how terminal leaves are constructed. Tree nodes are split in a two-stage process that first uses the EM algorithm to cluster observations and quadratic discriminant analysis is then used to identify split points within the clusters.

Unlike linear discriminant analysis, quadratic discriminant analysis does not assume homoskedastic data. However, quadratic discriminant analysis assumes that outcomes for each class identified by the split point is normally distributed. Prior work (e.g. Clarke, Lachenbruch, and Broffitt (1979)) has found that quadratic discriminant analysis performs poorly at determining split points when the distributions are highly skewed. By relaxing the box office budget condition it would not be surprising that the raw social media volume data is highly skewed (most films attract little attention) and as such in our application we would anticipate that SECRET would split in the wrong locations.

For completeness, the estimation algorithm for SECRET is presented below:

Algorithm: Scalable Linear Regression Tree (SECRET)

1. normalize datapoints to unitary sphere
2. find two Gaussian clusters in regressor–output space (EM)
3. label datapoints based on closeness to these clusters
4. for each split attribute, find best split point and determine its gini gain
5. let X be the attribute with the greatest gini gain and Q the corresponding best split predicate set
6. For one split S, partition data D into D_1 and D_2 based on Q and label node S with split attribute X
7. create children nodes S_1, S_2 of S and build leaves (S_1, D_1) (S_2, D_2)
8. Repeat steps 6-7 until stopping rule is satisfied

Once the tree has been constructed, we find the best linear regressor that fits the training data for each leaf. Similar to M5', SECRET searches through the original dataset and identify the subset that falls into each leaf. A simple regression model is formed with these datapoints and solved.

Returning to heteroskedastic data, the main differences in the splitting criteria between RT and M5 is the use of variance as the splitting criteria versus standard deviation reduction (SDR). Yet, the consequence of heteroskedastic data being biased to regions of the data are similar. RT and M5 and M5' also differ in the mechanism to estimate the leaf value, but as we discuss in the next section of the Appendix, this mechanism is nested within our hybrid approach. We will demonstrate hybrid versions of M5' outperform M5' in forecast accuracy with heteroskedastic data below.

Table A1 provides a list of the statistical software packages we use and their online source for many of the algorithms described above. Each of the methods described above that are not listed in the table are available in many platforms and in general we used either Matlab or R implementation of the respective algorithm.

B.3.1 Additional Comments on the Relative Performance of Linear Regression Tree Algorithms

In Table A2, we compare the proposed MAB and MARF (with 15 variables only) with the four linear regression tree algorithms described in the preceding section. Note that for each of the above algorithms, we conduct bagging and random forest (with 15 variables)

Table A1: Description of Machine Learning Packages

Method	Package Description
RT	Standard `fitrtree` package in MATLAB.
BAG,RF	Standard `TreeBagger` package in MATLAB.
BOOST	Standard `fitrensemble` package in MATLAB.
BART	`BART` package by Robert McCulloch, Rodney Sparapani, Robert Gramacy, Charles Spanbauer, Matthew Pratola, Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. Package is available in CRAN.
BART_{BMA}	`bartBMAnew` package by Eoghan O'Neill https://github.com/EoghanONeill/bartBMAnew
HBART	`rbart` package by Robert McCulloch, Matthew Pratola, and Hugh Chipman Package is available in CRAN.
M5′	`M5PrimeLab` package by Gints Jekabsons. http://www.cs.rtu.lv/jekabsons/
SECRET	`SECRET` package by Alin Dobra and Johannes Gehrke. http://himalaya-tools.sourceforge.net/Secret/
SVR	Standard `fitrsvm` package in MATLAB.
SVR_{LS}	`LS-SVMlab` package by K. De Brabanter, P. Karsmakers, F. Ojeda, and C. Alzate, J. De Brabanter, K. Pelckmans, B. De Moor,J. Vandewalle, J.A.K. Suykens, http://www.esat.kuleuven.be/sista/lssvmlab/

in our estimation. Therefore, methods (i) and (ii) can be regarded as a type of special bagging and random forest with pruned trees. Recall that using either OLS or Lasso to estimate the model in the terminal node requires that the sample size must be greater than the number of explanatory variables in the model. In contrast, MAB and MARF are more flexible since we can adjust the total number of variables within each model based on the number of observations that fall within each terminal node.

Estimation results that correspond to our main exercise are presented in Table A2. In all cases, our results indicate that MAB has the best performance. In general, the more naive linear regression tree algorithms display quite poor performance. However, we find that Lasso based methods outperform the OLS based methods, since the latter ignores subsample model uncertainty. The performance of M5′ method is similar to the Lasso based naive bagging.

Last, SECRET is designed for large datasets and as discussed in the preceding section the transformation of regression to classification problem crucially depends on the use of two Gaussian clusters. Given that our data contains numerous variables that are highly skewed, it is not a surprise that SECRET performs quite poorly.

B.4 Comparing Conventional M5′ Estimators with Model Averaging M5′ Hybrids

In this subsection, we demonstrate the wider applicability of our proposed hybrid strategy and examine if combing model averaging with M5′ methods can lead to improve-

Table A2: Comparing Our Hybrid Tree Approaches with Linear Regression Tree Algorithms Described in Section B.2

n_E	NB_{OLS}	NRF_{OLS}	NB_{LA}	NRF_{LA}	$BAG_{M5'}$	$RF_{M5'}$	SECRET	MAB	MARF	Benchmark
Panel A: Open Box Office										
Mean Squared Forecast Error (MSFE)										
10	1.6609	1.7734	0.8750	0.8274	0.7099	0.6902	0.8147	**0.5066**	0.5356	1.0000
20	2.2955	2.3302	0.8248	0.7881	0.8705	0.8352	0.9058	**0.7315**	0.7787	1.0000
30	2.7054	2.7654	0.8040	0.7675	0.9789	0.9292	1.0270	**0.7531**	0.8694	1.0000
40	2.9108	3.0032	0.7997	0.7572	1.0425	1.0171	1.0134	**0.9145**	1.0348	1.0000
Mean Absolute Forecast Error (MAFE)										
10	0.9289	0.9478	0.9508	0.9302	0.7460	0.7543	0.8324	**0.6232**	0.6742	1.0000
20	1.0748	1.0820	0.9460	0.9197	0.7922	0.7933	0.8975	**0.6955**	0.7495	1.0000
30	1.1631	1.1735	0.9497	0.9235	0.8177	0.8223	0.9785	**0.7042**	0.7733	1.0000
40	1.2215	1.2430	0.9573	0.9318	0.8399	0.8466	10469	**0.7625**	0.8157	1.0000
Panel B: Movie Unit Sales										
Mean Squared Forecast Error (MSFE)										
10	2.8963	2.8966	1.9840	1.9815	0.8805	0.9579	0.9575	**0.7307**	0.9168	1.0000
20	3.6710	3.6283	3.2619	3.2311	0.9598	1.0610	0.9649	**0.7009**	1.0564	1.0000
30	5.0171	5.1724	2.8294	2.8132	1.0378	1.1442	0.9576	**0.7494**	1.1702	1.0000
40	7.6604	7.9057	2.8138	2.7831	1.0467	1.1406	1.0706	**0.8626**	1.1832	1.0000
Mean Absolute Forecast Error (MAFE)										
10	1.2420	1.2274	1.5719	1.5678	0.8536	0.9064	0.9572	**0.7461**	0.9098	1.0000
20	1.3818	1.3737	1.6626	1.6456	0.8827	0.9384	1.0854	**0.7564**	0.9313	1.0000
30	1.6283	1.6482	1.6359	1.6219	0.9103	0.9620	0.9003	**0.7954**	0.9722	1.0000
40	2.0066	2.0142	1.6187	1.6116	0.9307	0.9754	1.0419	**0.8211**	0.9805	1.0000

Note: Bold numbers denote the strategy with the best performance in that row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the HRC^p method presented in the last column. The benchmark is identical to the main paper to facilitate further comparisons. The subscript in $MARF_q$ stands for the number of covariates randomly chosen at each node to consider as the potential split variable. All bagging and random forest estimates involve 100 trees.

ments in forecast accuracy. The no free lunch theorem of Wolpert and Macready (1997) states that since the relative performance of the RT and M5′ optimization algorithm varies across forecasting exercises, with neither dominating in all scenarios. However, we aim to show the value of undertaking the hybrid strategy relative to M5′ in this section. Specifically, we propose and consider the following six additional hybrid strategies

(i) $BAG^0_{M5'}$: M5′ bagging with each tree leaf estimated by simple average;

(ii) $RF^0_{M5'}$: M5′ random forest with each tree leaf estimated by simple average;

(iii) $BAG_{M5'}$: M5′ bagging with each tree leaf estimated by linear regression;

(iv) $RF_{M5'}$: M5′ random forest with each tree leaf estimated by linear regression;

(v) $MAB_{M5'}$: M5′ bagging with each tree leaf estimated by model averaging;

(vi) $MARF_{M5'}$: M5′ random forest with each tree leaf estimated by model averaging.

Strategies (i) and (ii) are similar to conventional bagging and random forest with the exception that the splitting rule follows the standard deviation reduction in M5′ instead of the SSR reduction in CART. Strategies (iii) and (iv) are the standard M5′ methods we discussed in Table A2. Strategies (v) and (vi) follows the idea of MAB and MARF that we proposed in the main text, in which we apply model averaging technique to the M5′ bagging and random forest. All tuning parameters are set to the default settings and we only consider random forest with a set of 15 explanatory variables chosen randomly to determine the split at each node.

Note that for methods (iii) and (iv), not all variables are needed in the linear model and only variables that are reference by the splitting process are included. Therefore, for methods (v) and (vi), we apply model averaging to each leaf using only the variables included in that leaf. We construct the candidate model set using the HRMS method introduced in Xie (2017).

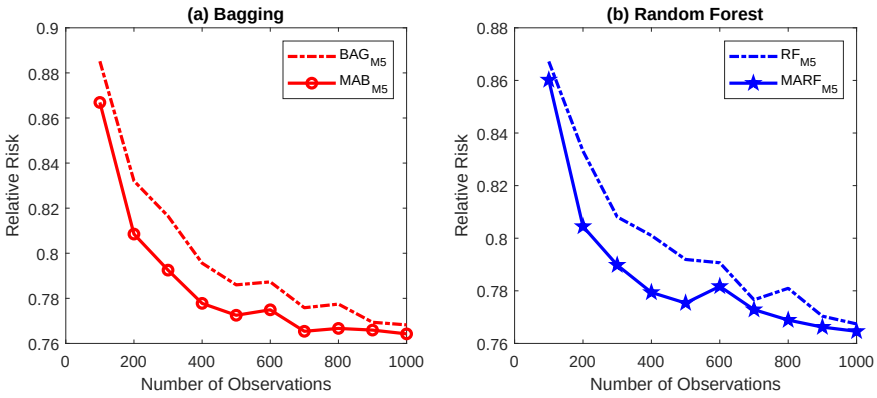
We repeat the empirical exercises undertaken in the main text using the six strategies discussed above and compare their relative performance to the benchmark HRC^p method. The results are presented in Table A3. Notice that the M5′ with simple average methods does not have good performance relative to other M5′ methods. Improved performance relative to the benchmark is not achieved in any case with the exception of retail movie unit sales with MSFE being the risk function. In general, model averaging M5′ hybrid methods has good performance relative to all of the other strategies. On several occasions, they perform similarly to M5′ with linear regression counterparts.

Using Monte Carlo simulation, we also compare conventional M5′ estimators to model averaging M5′ hybrids, following a identical set-up to that described in Section 3 of the main text. We reproduce Figures 2 and 3 on BAG$_{M5'}$, RF$_{M5'}$, MAB$_{M5'}$, and MARF$_{M5'}$. The results are plotted in figure A1. We see that with random heteroskedasticity, the gains from adding model averaging to M5′ appear quite small, particularly in large samples. However, when heteroskedasticity arises due to neglected parameter heterogeneity, we find larger gains from incorporating model averaging in place of linear regression within the terminal nodes. This likely arises since with neglected parameter heterogeneity, we are able to use multiple models, each allowing the same covariate to have different effects to explain more of the heterogeneity in the outcomes within the leaves than under the random heteroskedasticity scenario. Notice all that since relative to random forest, M5′ grows shorter trees are grown, we actually exhibit larger gains with model averaging as the sample size increases, since there are more observations and hence more candidate models available in each terminal node.

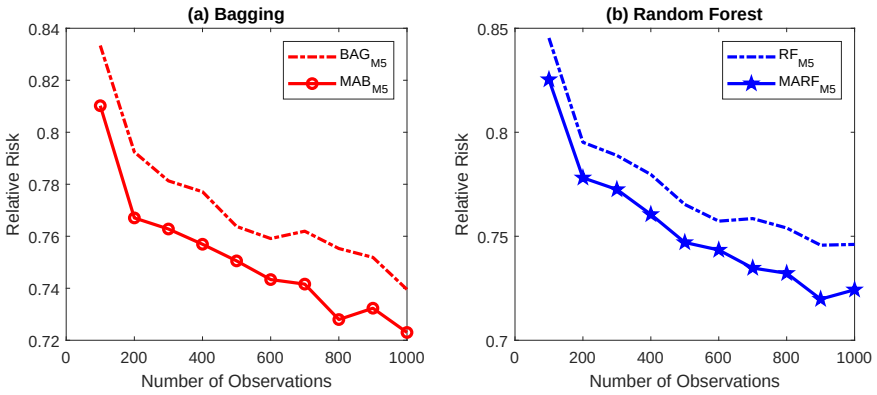
Last, the relative variable importance by M5′ hybrids can also be calculated in the same fashion as MAB and MARF. The role of social media variables in the top 10 are similar to the random forest results presented in the main text. However, the genres of the films now play a larger role in explaining retail movie unit sales, which may arise since the subgroups are formed by a different objective function to determine splits in the tree.

Figure A1: Relative Performance of M5′ and Model Averaging M5′ Learning

A. Random Heteroskedasticity



B. Parameter Heterogeneity



C. Risk Comparison under Different Scenarios

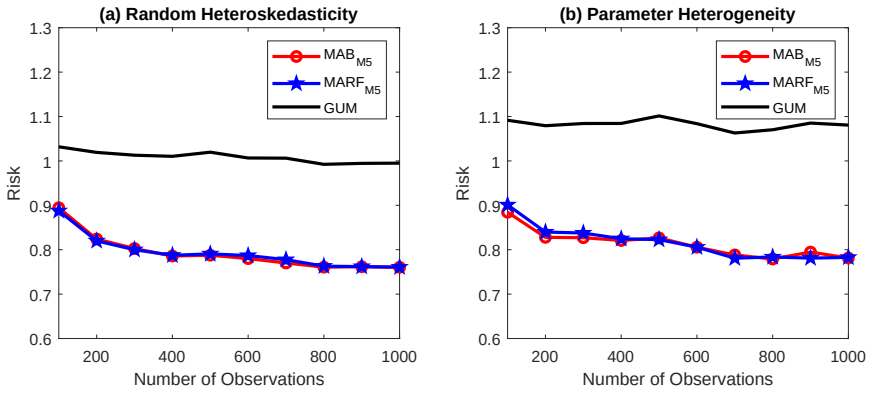


Table A3: Comparing Conventional M5′ Estimators with Model Averaging M5′ Hybrids

n_E	$BAG^0_{M5'}$	$RF^0_{M5'}$	$BAG_{M5'}$	$RF_{M5'}$	$MAB_{M5'}$	$MARF_{M5'}$	Benchmark
Panel A: Open Box Office							
Mean Squared Forecast Error (MSFE)							
10	0.7421	0.7974	0.7150	**0.7035**	0.7043	0.7040	1.0000
20	0.9761	0.8778	0.8605	0.8317	0.8579	**0.8249**	1.0000
30	0.9038	0.9674	0.8790	0.9230	**0.8762**	0.9180	1.0000
40	1.0767	1.2133	1.1980	1.0487	1.1004	1.0138	**1.0000**
Mean Absolute Forecast Error (MAFE)							
10	0.8024	0.7793	0.7462	0.7508	**0.7203**	0.7485	1.0000
20	0.8463	0.8167	0.8028	**0.7962**	0.7968	0.7870	1.0000
30	0.8183	0.8386	0.8182	0.8341	**0.8098**	0.8313	1.0000
40	0.8620	0.9022	0.8327	0.8488	**0.8032**	0.9048	1.0000
Panel B: Movie Unit Sales							
Mean Squared Forecast Error (MSFE)							
10	1.0750	1.0748	0.8874	0.9206	0.8536	**0.8443**	1.0000
20	1.1462	1.2109	0.9845	1.0963	**0.9782**	1.0728	1.0000
30	1.0464	1.0311	**0.9724**	1.0379	1.0790	1.0135	1.0000
40	1.0091	1.0606	**0.9816**	1.0730	0.9997	1.0272	1.0000
Mean Absolute Forecast Error (MAFE)							
10	0.9743	0.9023	0.8626	0.9049	**0.8589**	0.8942	1.0000
20	0.8597	0.9766	0.8841	0.9397	0.8766	**0.8744**	1.0000
30	0.8480	0.9309	0.9577	0.9218	**0.8405**	0.9124	1.0000
40	0.9049	0.9237	0.9532	0.9463	**0.8944**	0.9358	1.0000

Note: Bold numbers denote the strategy with the best performance in that row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the HRC^p method presented in the last column.

The difference in how splits are conducted and the shorter tree may explain why the results in the lower panel of Table A3 appear less promising. Similar to how applied econometrics studies often carry out robustness of logit regression results with linear probability models, social scientists and practionners should likely investigate how results differ based on different split rules used to build tree structures. Yet, our results continue to find that irrespective of the split rules, there are gains from using model averaging in the terminal nodes.

B.5 Support Vector Regression

Support vector regression (SVR) is an extension of the support vector machine (SVM) classification method of Vapnik (1996) to consider a real-valued outcome variable as in the classical regression problem. Prior to describing SVR, we will provide some intuition for how SVM works in a binary classification setting. SVM was introduced in Boser, Guyon, and Vapnik (1992) and provides a learning algorithm that infers functional relationships in the underlying dataset by following the structural risk minimization induction principle (formally defined below) of Vapnik (1996). The major difference with classical regression or tree based methods is that SVM allows for complex nonlinear relationships

Table A4: Relative Importance of the Predictors by M5′ Hybrids

Ranking	With Twitter Variables		Without Twitter Variables	
	$MAB_{M5'}$	$MARF_{M5'}$	$MAB_{M5'}$	$MARF_{M5'}$
Panel A: Open Box Office				
1	Screens	Screens	Screens	Screens
2	Volume: T-21/-27	Volume: T-1/-3	Rating: R	Budget
3	Volume: T-7/-13	Budget	Rating: PG	Rating: R
4	Budget	Volume: T-4/-6	Genre: Adventure	Genre: Horror
5	Volume: T-1/-3	Volume: T-14/-20	Budget	Rating: PG13
6	Volume: T-14/-20	Volume: T-7/-13	Genre: Horror	Rating: PG
7	Volume: T-4/-6	Volume: T-21/-27	Genre: Comedy	Genre: Comedy
8	Weeks	Genre: Horror	Genre: Fantasy	Genre: Adventure
9	Sentiment: T-1/-3	Genre: Adventure	Genre: Action	Weeks
10	Sentiment: T-14/-20	Sentiment: T-1/-3	Weeks	Genre: Animation
Panel B: Movie Unit Sales				
1	Genre: Biography	Screens	Screens	Screens
2	Genre: Mystery	Budget	Weeks	Weeks
3	Screens	Weeks	Budget	Budget
4	Weeks	Volume: T+8/+14	Rating: R	Genre: Adventure
5	Budget	Volume: T-21/-27	Genre: Comedy	Genre: Fantasy
6	Volume: T+8/+14	Volume: T+1/+7	Genre: Horror	Genre: Drama
7	Volume: T+0	Volume: T+15/+21	Genre: Thriller	Genre: Comedy
8	Volume: T+1/+7	Volume: T+0	Rating: PG	Rating: R
9	Volume: T-4/-6	Volume: T+22/+28	Genre: Animation	Rating: PG13
10	Genre: Animation	Volume: T-1/-3	Rating: PG13	Genre: Family

Note: This table presents the rank order of the importance of the predictors for film revenue by the respective machine learning.

by transforming the original data into a higher dimensional space via an a priori chosen mapping.

Intuitively, SVM finds a hyperplane in this higher dimensional space that perfectly separates the two classes (i.e. two values of the outcome). That is, all data points in one class lie above the hyperplane and all points below rely in the other class. However, there might be more than one separating hyperplane and the preferred one is that which maximizes the distance to the closest point. The distance from the hyperplane to the nearest point is called the margin and all data points on the boundary of the margin are called support vectors. These data points are the sole ones that contribute to the forecast.

Drucker, Burges, Kaufman, Smola, and Vapnik (1996) introduced SVR and following Hastie, Tibshirani, and Friedman (2009, Chapter 12), we first assume that our DGP can be expressed by a linear equation as

$$y_i = f(\boldsymbol{X}_i) + e_i = \boldsymbol{X}_i\boldsymbol{\beta} + e_i = \beta_0 + \tilde{\boldsymbol{X}}_i\boldsymbol{\beta}_1 + e_i \tag{A31}$$

for $i = 1, ..., n$, where $\boldsymbol{X}_i = [1, \tilde{\boldsymbol{X}}_i]$ is a $1 \times (k+1)$ input vector and $\boldsymbol{\beta} = [\beta_0, \boldsymbol{\beta}_1^\top]^\top$ are the coefficients. By assuming a linear DGP we do not have to transform the data into a higher dimensional space and the goal is to find coefficients that support a function $f(\boldsymbol{X}_i)$ of the input variables $\boldsymbol{X}_i$ (this is often referred to as feature vector in the SVR literature) that deviates from the target variable by a value no greater than a predetermined ϵ for each

observation; that is as flat as possible. In other words, ϵ can be viewed as the maximum margin for the hyperplane.

To solve the maximization problem described above we transform it into a minimization problem of a quadratic cost function. Support vector regression, like OLS, seeks to minimize a function of residuals, but as we discuss it penalizes residuals in a different way. Instead of maximizing the margin, SVR minimizes the Euclidean norm of the coefficient vector. The reformulation into a quadratic cost function does not change the optimization problem but assures that all training data only occur in form of a dot product between vectors. Formally, SVR solves the following problem to estimate $\boldsymbol{\beta}$ through the minimization of

$$H(\boldsymbol{\beta}) = \sum_{i=1}^{n} V_\epsilon\big(y_i - f(\boldsymbol{X}_i)\big) + \frac{\lambda}{2}\|\boldsymbol{\beta}_1\|^2, \tag{A32}$$

where

$$V_\epsilon(e_i) = \begin{cases} 0 & \text{if } |e_i| < \epsilon \\ |e_i| - \epsilon & \text{otherwise} \end{cases} \tag{A33}$$

is an ϵ-insensitive error measure. The ϵ-insensitive error is a loss function for residuals that imposes no penalty on residuals smaller than $|\epsilon|$ and penalizes residuals linearly in the degree to which they exceed $|\epsilon|$. Examining function (A32) we observe that the size of coefficient is penalized using the squared l_2 norm, which reduces the variance of the model and hence overfitting. While the parameter ϵ is (usually) predetermined, λ is a more traditional regularization parameter that can be estimated by cross-validation. Therefore, SVR can be regarded as another form of penalized regression whose performance can be influenced by the values of its tuning parameters, ϵ and λ.

Let $\hat{\boldsymbol{\beta}} = \big[\hat{\beta}_0, \hat{\boldsymbol{\beta}}_1^\top\big]^\top$ be the minimizers of function (A32), the solution function can be shown to have the form

$$\begin{aligned} \hat{\boldsymbol{\beta}}_1 &= \sum_{i=1}^{n} (\hat{\alpha}_i^* - \hat{\alpha}_i)\tilde{\boldsymbol{X}}_i^\top, \\ \hat{f}(\boldsymbol{X}) &= \sum_{i=1}^{n} (\hat{\alpha}_i^* - \hat{\alpha}_i)\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}_i^\top + \hat{\beta}_0 \boldsymbol{\iota}_n, \end{aligned}$$

where $\boldsymbol{\iota}_n$ is an $n \times 1$ vector of ones, $\tilde{\boldsymbol{X}}$ is the $n \times k$ matrix with each row being $\tilde{\boldsymbol{X}}_i$ for $i = 1, ..., n$, and the parameters $\hat{\alpha}_i$ and $\hat{\alpha}_i^*$ are the nonnegative multiplier of the following Lagrangian equation

$$\min_{\hat{\alpha}_i, \hat{\alpha}_i^*} \; \epsilon \sum_{i=1}^{n} (\hat{\alpha}_i^* + \hat{\alpha}_i) - \sum_{i=1}^{n} y_i(\hat{\alpha}_i^* - \hat{\alpha}_i) + \frac{1}{2} \sum_{i=1}^{n} \sum_{i'=1}^{n} (\hat{\alpha}_i^* - \hat{\alpha}_i)(\hat{\alpha}_{i'}^* - \hat{\alpha}_{i'}) \tilde{\boldsymbol{X}}_i \tilde{\boldsymbol{X}}_{i'}^\top$$

subject to the constraints

$$0 \leq \hat{\alpha}_i^*, \hat{\alpha}_i \leq 1/\lambda, \quad \sum_{t=1}^{n} (\hat{\alpha}_i^* - \hat{\alpha}_i) = 0, \quad \hat{\alpha}_i \hat{\alpha}_i^* = 0$$

for all $i = 1, ..., n$. The non-zero values of $\hat{\alpha}_i^* - \hat{\alpha}_i$ for $i = 1, ..., n$ are referred to as the support vector. Only these observations that are at least ϵ away from the predicted hyperplane support it; and contribute to the forecast. Thus, the solution is sparse and the prediction variance is reduced. In other words, outliers are removed since the linear penalization of residuals beyond ϵ ensures these observations do not contribute to the prediction.[5] Since our application is focused strictly on forecasting, we are not concerned that a limitation of SVR is that procedures to undertake statistical inference methods require distributional assumptions that are rarely satisfied in practice.

B.5.1 SVR with a Nonlinear DGP

We next extend the above linear SVR framework to a more general case where the DGP can take a more general form such as

$$y_i = f(\boldsymbol{X}_i) + e_i, \tag{A34}$$

where f is unknown to the researcher. As with SVM, to allow for complex nonlinear relationships and approximate f , the original data is transformed from k-dimensional feature space into a new higher dimensional space feature space whose dimensions depends on an a priori chosen mapping scheme. That is, suppose we reexpress

$$y_i = f(\boldsymbol{X}_i) + e_i = \beta_0 + h(\tilde{\boldsymbol{X}}_i)\boldsymbol{\beta}_2 + e_i, \tag{A35}$$

where $h(\cdot) : \mathbb{R}^k \to \mathbb{R}^q$ is a set of basis functions which can be infinite dimensional (there is a relation with the existence of a Hilbert space (Courant and Hilbert, 1953) and $\boldsymbol{\beta}_2$ is the coefficient of the nonlinear SVR that is identical to $\boldsymbol{\beta}_1$ if the DGP is linear. In this new $\mathbb{R}^q$ space, the relationship between the outcome and the new feature vector $h(\tilde{\boldsymbol{X}}_i)$ is believed to be in linear form. Intuitively, we can then use the same support vector regression algorithm to find the separating hyperplane for a linear relationship (i.e. equation (A31)) on this transformed version of the data allowing us to get a non-linear algorithm.

To estimate the non-linear algorithm requires a kernel based procedure that can be interpreted as mapping the data from the original input space into a potentially higher dimensional "feature space", where linear methods may then be used for estimation.

[5] In other words the support vector has a somewhat local interpretation whereas the least absolute deviation estimator (i.e. conditional median regression) has a strictly local interpretation since it employs an absolute value loss function. That said, SVR does not have the same robustness to outliers as a conditional quantile estimator, but it is much less susceptible to outliers than linear regression or LASSO estimators.

The kernel function allow for non-linear relationships and this data transformation is achieved if the kernel satisfies conditions given by Mercer's theorem (Mercer, 1909).[6] The use of kernels enables us to avoid paying the computational penalty implicit in the number of dimensions, since it is possible to evaluate the training data in the feature space through indirect evaluation of the inner products. This is often referred to as the kernel trick.

As such, the kernel function is essential to the performance of SVR since it contains all the information available in the model and training data to perform supervised learning; with the sole exception of having measures of the outcome variable. By letting q grow large ($q >> k$), and through appropriate choices of $h(\cdot)$, SVR gains many of the advantages of series estimators that are frequently used in nonparametric econometrics. Formally, we define the kernel function $K(\cdot)$ as the linear dot product of the nonlinear mapping,

$$K(\tilde{\boldsymbol{X}}_i, \tilde{\boldsymbol{X}}_{i'}) = h(\tilde{\boldsymbol{X}}_i)h(\tilde{\boldsymbol{X}}_{i'})^\top.$$

In our analysis, we contrast various kernel choices, including the linear kernel, the Gaussian kernel (sometimes referred to as "radial basis function" and "Gaussian radial basis function" in the support vector literature), and polynomial kernels with different orders:

$$\begin{aligned}
\text{Linear} \quad &: \quad K(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{X}}_{i'}) = \tilde{\boldsymbol{X}}_i \tilde{\boldsymbol{X}}_{i'}^\top, \\
\text{Gaussian} \quad &: \quad K(\tilde{\boldsymbol{X}}_i, \tilde{\boldsymbol{X}}_{i'}) = \exp\left(-\frac{\|\tilde{\boldsymbol{X}} - \tilde{\boldsymbol{X}}_{i'}\|^2}{2\sigma_x^2}\right), \\
\text{Polynomial} \quad &: \quad K(\tilde{\boldsymbol{X}}_i, \tilde{\boldsymbol{X}}_{i'}) = (\gamma + \tilde{\boldsymbol{X}}_i \tilde{\boldsymbol{X}}_{i'}^\top)^d,
\end{aligned}$$

where the hyperparameters σ_x^2, γ, and d can be tuned through cross-validation.

As before, to solve the maximization problem, we transform it into a minimization problem of the a quadratic cost function. In this setting, we estimate the coefficients $\boldsymbol{\beta} = \left[\beta_0, \boldsymbol{\beta}_2^\top\right]^\top$ through the minimization of

$$H(\boldsymbol{\beta}) = \sum_{i=1}^{n} V_\epsilon\big(y_i - f(\boldsymbol{X}_i)\big) + \frac{\lambda}{2}\|\boldsymbol{\beta}_2\|^2. \tag{A36}$$

Note that $\boldsymbol{\beta}_2$ is implicit and can be infinite dimensional. The solution of equation (A36) now has the form

$$\hat{f}(\boldsymbol{X}) = \sum_{t=1}^{n} (\hat{\alpha}_i^* - \hat{\alpha}_i) K(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{X}}_i) + \hat{\beta}_0 \boldsymbol{\iota}_n$$

[6] Mercer's theorem states that for $K(\cdot)$ to be a valid (Mercer) kernel, it is necessary and sufficient that for any finite input data, the corresponding kernel matrix is both symmetric and positive semi-definite. The elements of this kernel matrix are given by the dot-product in the transformed feature space.

with $\hat{\alpha}_t^*$ and $\hat{\alpha}_t$ being the nonnegative multiplier of the following Lagrangian equation

$$\min_{\hat{\alpha}_i, \hat{\alpha}_i^*} \epsilon \sum_{t=1}^{n} (\hat{\alpha}_i^* + \hat{\alpha}_i) - \sum_{i=1}^{n} y_i (\hat{\alpha}_i^* - \hat{\alpha}_i) + \frac{1}{2} \sum_{i=1}^{n} \sum_{i'=1}^{n} (\hat{\alpha}_i^* - \hat{\alpha}_t)(\hat{\alpha}_{i'}^* - \hat{\alpha}_{i'}) K(\tilde{\boldsymbol{X}}_i, \tilde{\boldsymbol{X}}_{i'}).$$

This appears similar to the solution of the SVR case in the linear DGP setting and the main addition being the kernel function. In the nonlinear setting, the optimization problem corresponds to finding the flattest function in feature space, rather than input space.

B.6 Least Squares Support Vector Regression

SVR requires solving the regression problem by means of convex quadratic programing in addition to the researcher having to specify i) two hyperparameters ϵ and λ, and (ii) a kernel function together with its respective parameters. Suykens and Vandewalle (1999) proposed a modification to the classic SVM that eliminate the hyperparameter ϵ and replaces the original ϵ-insensitive loss-function with a least-square loss function. This is known as least squares SVM, which solves a set of linear equations to find the minimum of the cost function without requiring quadratic programming. The least squares SVM is more computationally efficient than classic SVM and is computationally capable to deal with large datasets with high dimensionality. Subsequently, for continuous outcome variables Suykens, Gestel, Brabanter, Moor, and Vandewalle (2002) extended this idea to develop least squares SVR, henceforth SVR_{LS}.

Similar to the minimization problem in equation (A32), the SVR_{LS} considers minimizing

$$H(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2 + \frac{\lambda}{2} \|\boldsymbol{\beta}_2\|^2. \tag{A37}$$

Thus, the classic SVR formulation is modified at two points. The error term e_i in the constraint (A33) is now denoted by an equality to emphasize that it represents the true deviation between actual values and forecasts in the SVR_{LS} formulation, rather than the inequality constraints that define the maximum margin variable needed to ensure feasibility of SVR. Second, in equation (A37) a squared loss function replace the ϵ-insensitive loss function in SVR.

Notice, that the cost function in equation (A37) consists of a residual sum of squares (SSR) fitting error as well as a regularization term. This is also a standard procedure for the training of multi-layer perceptrons and the formulation of equation (A37) can also be regarded as a nonparametric ridge regression function formulated in the feature space. SVR_{LS} involves choosing a kernel function and kernel parameters, and determining the regularization parameter usually via cross validation. These settings of the hyperparameters can be viewed as model selection with SVR_{LS}.

SVR_{LS} has a global and unique solution that can be found with computationally efficient numerical optimization methods. An important feature of the solution is that every data point now becomes a support vector, whereas SVR has sparse solution. This can be clearly illustrated by considering the conditions for optimality. We define $\boldsymbol{H}$ to be the $n \times q$ basis matrix where $q > n$. For ease of exposition, we let the intercept term β_0 be absorbed in $h(\cdot)$. The coefficient $\boldsymbol{\beta}$ can be estimated by minimizing the following penalized least squares criterion

$$H(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{H\beta})^\top (\boldsymbol{y} - \boldsymbol{H\beta}) + \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2.$$

The solution $\hat{\boldsymbol{\beta}}$ should satisfy $-2\boldsymbol{H}^\top(\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{\beta}}) + \lambda\hat{\boldsymbol{\beta}} = 0$ and the prediction

$$\hat{f}(\boldsymbol{X}) = \boldsymbol{H}\hat{\boldsymbol{\beta}} = \left(\boldsymbol{HH}^\top + \frac{\lambda}{2}\boldsymbol{I}_n\right)^{-1} \boldsymbol{HH}^\top \boldsymbol{y} \equiv \boldsymbol{Py}. \tag{A38}$$

where $\boldsymbol{I}_n$ is a $n \times n$ identity matrix and

$$\boldsymbol{P} \equiv \left(\boldsymbol{HH}^\top + \frac{\lambda}{2}\boldsymbol{I}_n\right)^{-1} \boldsymbol{HH}^\top \tag{A39}$$

is a $n \times n$ matrix. Note that the $n \times n$ matrix $\boldsymbol{HH}^\top$ is the kernel matrix that is positive definite and each element of the matrix is a symmetric continuous function. Equation (A38) implies that although the matrix $\boldsymbol{H}$ is implicit, we can make predictions as long as the kernel we predetermined is explicit. We consider linear, Gaussian, and polynomial kernels for SVR_{LS} in this paper. It is also worth noting that Foxall, Cawley, Talbot, Dorling, and Mandic (2002) and Cawley, Talbot, Foxall, Dorling, and Mandic (2004) each considered extending SVR_{LS} to allow for potential heteroskedasticity and introduced a regularized kernel regression model for such a setting. In the next subsection, we discuss how to extend SVR_{LS} to allow for model uncertainty and in one step additionally estimate model averaging weights in both homoskedastic and heteroskedastic settings.

B.7 Model Averaging SVR_{LS}

Since SVR_{LS} considers minimizing a least squares loss function, there is possibility to incorporate least squares model averaging. In contrast, the classic SVR deals with the ϵ-insensitive loss-function. The conventional least squares model averaging framework is incompatible with the classic SVR. Moreover, the sparseness of SVR ensures that different support vectors are found for different models and hyperparameter combinations, presenting additional challenges for interpreting a model averaging SVR hybrid. Combining the classic SVR with model averaging may require results from an emerging area of research that extends model averaging to functional data (see e.g. Zhang, Chiou, and

Ma 2018), which we leave for further research. As such, we propose a hybrid model averaging estimation based on SVR_{LS} only in this section.

In practice, the DGP presented in equation (A35) is unknown and we approximate it with a set of M candidate models. In our application in the main text, the candidate model set is constructed by the model screening strategy discussed in Section 2.1 of the text. Alternatively, the full combination of all potential regressors could be used. We denote the input variables by model m as $\boldsymbol{X}_i^{(m)}$, in which the $1 \times k^{(m)}$ vector $\boldsymbol{X}_i^{(m)}$ is a subset of $\boldsymbol{X}_t$ that includes all potential explanatory variables. The m^{th} candidate model can be written as

$$y_i = f(\boldsymbol{X}_i^{(m)}) + e_i^{(m)},$$

where the superscript m indicates variables associate with model m. Let $\hat{f}(\boldsymbol{X}^{(m)}) = \hat{f}^{(m)}$ for $m = 1, ..., M$ be the set of predictions corresponding to different candidate models. Define the weight vector $\boldsymbol{w} = [w^{(1)}, ..., w^{(M)}]^\top$ which follows the weight set definition in $\mathcal{H}$:

$$\mathcal{H} \equiv \left\{ w_m \in [0,1]^M : \sum_{m=1}^{M} w_m = 1 \right\}.$$

Then, the weighted average prediction is $\hat{f}(\boldsymbol{w}) = \sum_{m=1}^{M} w^{(m)} \hat{f}^{(m)} = \boldsymbol{P}(\boldsymbol{w})\boldsymbol{y}$, where $\hat{f}^{(m)} = \boldsymbol{P}^{(m)}\boldsymbol{y}$ following equation (A38), $\boldsymbol{P}(\boldsymbol{w}) = \sum_{m=1}^{M} w^{(m)} \boldsymbol{P}^{(m)}$, and $\boldsymbol{P}^{(m)}$ is the $\boldsymbol{P}$ matrix defined in equation (A39) for model m.

The weight vector is essential in model averaging estimation. Ullah and Wang (2013) presents recent developments in model selection and model averaging for parametric and nonparametric models. Ullah and Wang (2013) argues that for nonparametric models with prediction function following the formulation of Equation (A38), one can estimate the model averaging weights by applying Mallows criterion when the error term exhibits homoskedasticity.

$$C_n(\boldsymbol{w}) = \sum_{i=1}^{n} \hat{e}_i^2(\boldsymbol{w}) + 2\sigma^2 \sum_{i=1}^{n} p_{ii}(\boldsymbol{w}) \tag{A40}$$

We extend the above criterion and propose estimating $\boldsymbol{w}$ by minimizing the following criterion with $\boldsymbol{w} \in \mathcal{H}$ based on a heteroskedastic error term:

$$C_n(\boldsymbol{w}) = \sum_{i=1}^{n} \hat{e}_i^2(\boldsymbol{w}) + 2 \sum_{i=1}^{n} \left(\hat{e}_i(\boldsymbol{w})\right)^2 p_{ii}(\boldsymbol{w}), \tag{A41}$$

where $p_{ii}(\boldsymbol{w})$ is the i^{th} diagonal term in $\boldsymbol{P}(\boldsymbol{w})$ and $\hat{e}_i(\boldsymbol{w})$ is the i^{th} element in

$$\hat{\boldsymbol{e}}(\boldsymbol{w}) = \sum_{m=1}^{M} w^{(m)} \hat{\boldsymbol{e}}^{(m)} = (\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{w}))\boldsymbol{y}. \tag{A42}$$

Equation A42 represents the averaged SVR_{LS} residuals. Estimating $\boldsymbol{w}$ by Criterion (A41)

is a convex optimization process. Criteria (A41), which also appears in an identical fashion as equation (10) in the main text, can be regarded as the HPMA criteria for nonparametric models.

We discuss the HPMA criteria in further detail in Section D.5 and briefly this extends the PMA used in the second step of our hybrid procedure with regression trees, to allow for heteroskedastic error terms. The Mallows criterion is often used as a stopping rule for various forms of stepwise regression and it is asymptotically equivalent to the squared error. Hansen (2007) proves that the model average estimator that minimizes the Mallows criterion also minimizes the squared error in large samples. In the next section of the Appendix, we provide intuition on why incorporating model averaging yields gains and remind the reader of figure 1 in the main text for simulation evidence.

C Further Intuition on the Hybrid Method for Tree Based Algorithms

Hansen (2020) stresses that the default in empirical work should be to assume that the errors in regression models are heteroskedastic, not the converse. He further claims the definition of heteroskedasticity in most econometrics textbooks adds confusion. He argues that rather than framing heteroskedasticity as the case where the variance of the regression error varies across observations, researchers should view it simply as the conditional variance depends on observables.

There was very little work in the field of econometrics on the consequences of heteroskedastic data when Breiman, Friedman, and Stone (1984) was first published. Thus, the choice of a homogeneous variance across the entire explanatory-variable space in Breiman, Friedman, and Stone (1984) was innocuous, even though it corresponds to imposing a homoskedasticity assumption. This algorithm would split in the correct places with homoskedastic data in a computationally efficient manner.

The focus of Breiman, Friedman, and Stone (1984) was on computational efficiency and assuming homoskedasticity is common in many econometric innovations. Hansen (2020) writes "Homoskedasticity greatly simplifies the theoretical calculations, and it is therefore quite advantageous for teaching and learning. It should always be remembered, however, that homoskedasticity is never imposed because it is believed to be a correct feature of an empirical model, but rather because of its simplicity." This quote motivates our consideration of the challenges posed by heteroskedastic data for forecasting with statistical learning algorithms.

The idea of using a local constant model within terminal nodes in Breiman, Friedman, and Stone (1984) is also well justified by econometrics, since it is well known that the best predictor for an outcome y (in the class of constants) is the unconditional mean as it minimizes the mean squared prediction error. Yet, with the availability of covariates a

researcher can gain improvements in mean squared prediction error. In the econometrics literature, it has been proven under assumptions of data having finite variances and independent variation, that given a realized value of x, the conditional mean $\mathbb{E}(y|x)$ is the best predictor of y. The linear conditional expectation function ($\mathbb{E}(y|x)$) is quite flexible since it can include interaction terms and non-linear terms. The challenge that arises in practice is there is uncertainty about which covariates to include when we calculate the conditional mean. Put differently, the functional form is typically unknown. Many of the linear regression tree estimator discussed in the preceding section require a low dimensional set of covariates for their implementation and assume they have the correct functional form.

The reason we advocate model uncertainty is that the conditional expectation function for different types of films may not only contain different sets of covariates but may also have the parameters vary across the models. This parameter heterogeneity is what causes heteroskedasticity as is well known from a random coefficient model allows the returns to each explanatory variable to vary in the population. Further, the linear random coefficient model implies a linear conditional expectations function with a heteroskedastic error. The idea of the hybrid strategy can basically be viewed as allowing there to be different weights to models with different parameters in subgroups of films that are ex-post to tree splitting thought to be as similar as possible. By averaging across multiple candidate models, our strategy provides a better approximation since it nests all the simpler conditional expectation function (given by a single candidate model) and the unconditional mean (which itself is another candidate model.

C.1 Additional Monte Carlo Evidence on the Effect of Parameter Heterogeneity

In this subsection, we demonstrate how heterogeneity can affect the effectiveness of recursive partitioning strategies that engage in forecasting outcomes based on the input variables. We consider a simplified version of the DGP (10) from the main text

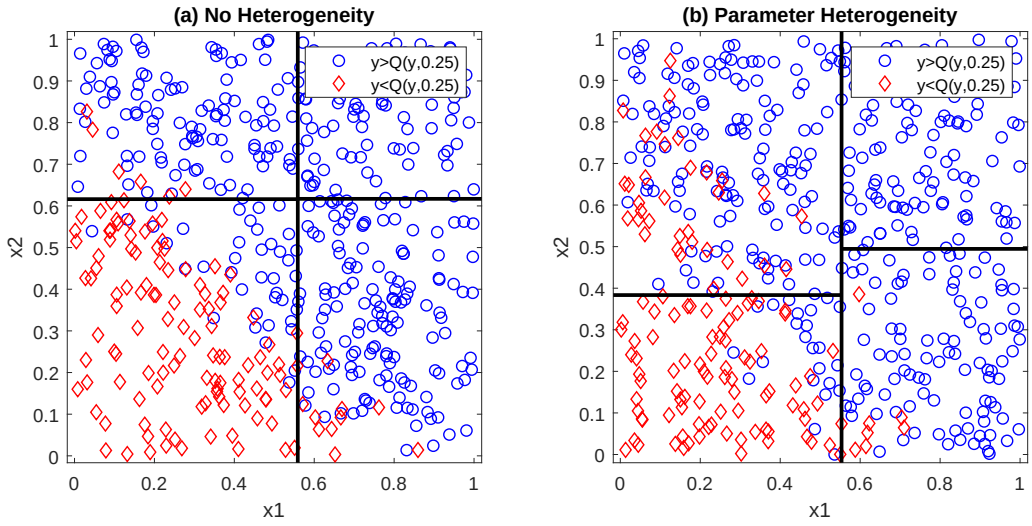
$$y_i = \beta_1 x_{1i} + (\beta_2 + r \cdot \sigma_i) x_{2i} + e_i$$

for $i = 1, ..., 500$. We set both x_i's follow $U(0,1)$, $\beta_1 = \beta_2 = 1$ and $\sigma_i \sim \mathrm{N}(0,1)$. We let the error term $e_i \sim \mathrm{N}(0, 0.01)$ which has a smaller impact on the DGP. We consider the following two scenarios:

1. **No Heterogeneity**: we set the parameter $r = 0$ and eliminate heterogeneity.
2. **Parameter Heterogeneity**: heterogeneity in the coefficient on x_{2i} for each observation is created by setting $r = 1/5$. Note that the expected value of the coefficients on x_{2i} is identical to the one used in Scenario 1 of the main text.

For simplicity, we restrict the number of splits to be three which generates four partitions for each scenario. Since both x_i's are generated from the same distribution, the splits should ideally divide the plots into four rectangles of equal size. The results presented in Figure A2 consider subplot (a) corresponding to no heterogeneity and subplot (b) which considers parameter heterogeneity. For each subplot, the horizontal and vertical axes represent x_1 and x_2, respectively. Scatter plots with red diamonds correspond to the pairs of x_i's associated with y_i's on the lower 25% quantile; and blue circles represent the remaining pairs of x_i's.

Figure A2: Demonstrate the Effect of Heteroskedasticity on Initial Split Location



As Figure A2(a) shows, when there is no heterogeneity, the regression tree divides the plots into four similar-sized rectangles with most of the diamonds located in the lower left quadrant. However, once we allow parameter heterogeneity, as demonstrated in Figure A2(b), the splits are uneven and the diamonds are mixed with circles, which hints at the ineffectiveness of conventional recursive partitioning under heterogeneity. When we remove the restriction of the number of splits to allow more than four groups, we continue to find more heterogeneity in outcomes in the leaves of trees that use data generated in the parameter heterogeneity scenario relative to the no heterogeneity scenario.

D Econometric Methods

In this section, we first provide details on how we implement each econometric strategy considered in the simulation experiment considered in Section 4.1 of the data. This section also provides a review of several existing heteroskedasticity-robust model averaging methods and several of the Lasso based methods. We summarize the theoretical conclusions and provide details on the computational algorithm used for each method.

D.1 Traditional econometric approaches to model building

The first estimator considered in the simulation experiment is a general unrestricted model (GUM) that includes every independent variable in our data set linearly. OLS estimation is used and our data set contains fewer covariates (when we exclude the possibility of both higher order and interaction terms). A special case of the GUM is MTV, which is simply the subset of the GUM that excludes all of the Twitter generated sentiment and volume explanatory variables. The MTV model is used as a first step to gain an idea of the importance of the inclusion of social media data in explaining variation in film revenues. It is well-established in the econometrics literature that a kitchen sink model such as GUM (and MTV) will lose efficiency if some of the explanatory variables are irrelevant in the sense of its exclusion would not affect the unbiasedness of the OLS estimator.

To determine which regressors should remain in the estimating equation of GUM, model selection methods have been developed in the econometrics literature. Among the best known of the methods is Akaike's Information Criterion (AIC) that summarizes the quality of a model by trading fit, measured by the maximized log likelihood, against complexity, measured by the number of estimated parameter. This correction introduced in Akaike (1973) is available in every statistics package and uses asymptotic theory to construct an analytical bias correction that arises since using the maximized log-likelihood, based on the estimated parameters, is an upward biased estimator of the Kullback-Leibler Divergence, the quantity that measures how the probability distribution of the approximation model is different from the true DGP.

An alternative approach to selecting a model using a subset of regressors from GUM is the general to specific method (GUM) of Hendry and Nielsen (2007). This approach begins with a GUM that nests restricted models and, thus, allows any restrictions to be tested with t-tests, and F-tests. Using the results of these specification tests, one will move from the GUM to a smaller, more parsimonious, specific model. This general to specific (GETS) strategy does not suffer from path dependence that plagues a backwards stepwise approach since it only removes explanatory variables if the new model is a valid reduction of the GUM based on a researcher pre-specified p-value. If competing models are selected, encompassing tests or information criteria such as the AIC can be used to select a final model.

Last, the prediction model averaging proposed by Xie (2015) is described in the main text. The weights that are applied to all potential candidate models are determined as the analog of the prediction criterion of Amemiya (1980). In the remaining subsections of this section, we provide extended discussions of the remaining dimension reduction, econometric and model screening approaches whose performance is evaluated in our study.

D.2 Jackknife Model Averaging

Hansen and Racine (2012) proposed a jackknife model averaging (JMA) estimator for the linear regression model. The model set-up is identical to that provided in section 2. Hansen and Racine (2012) demonstrate the asymptotic optimality of the JMA estimator in the presence of heteroskedasticity and suggest selecting the weights by minimizing a leave-one-out cross-validation criterion

$$\text{JMA}(\boldsymbol{w}) = \frac{1}{n}\boldsymbol{w}^\top \tilde{\boldsymbol{E}}^\top \tilde{\boldsymbol{E}}\boldsymbol{w} \quad \text{with} \quad \hat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in \boldsymbol{H}^*}{\arg\min}\ \text{JMA}(\boldsymbol{w}), \tag{A43}$$

where $\tilde{\boldsymbol{E}} = [\tilde{\boldsymbol{e}}^1, ..., \tilde{\boldsymbol{e}}^M]^\top$ is an $n \times M$ matrix of jackknife residuals and $\tilde{\boldsymbol{e}}^{(m)}$ stands for the jackknife residuals of model m.

The jackknife residual vector $\tilde{\boldsymbol{e}}^{(m)} = \boldsymbol{y} - \tilde{\boldsymbol{\mu}}^{(m)}$ for model m requires the estimate of $\tilde{\boldsymbol{\mu}}^{(m)}$, where its i^{th} element, $\tilde{\mu}_i^{(m)}$, is the least squares estimator $\hat{\mu}_{\backslash i}^{(m)}$ computed with the i^{th} observation deleted. In practice, $\tilde{\boldsymbol{e}}^{(m)}$ can be conveniently written as $\tilde{\boldsymbol{e}}^{(m)} = \boldsymbol{D}^{(m)}\hat{\boldsymbol{e}}^{(m)}$, where $\hat{\boldsymbol{e}}^{(m)}$ is the least squares residual vector and $\boldsymbol{D}^{(m)}$ is the $n \times n$ diagonal matrix with the i^{th} diagonal element equal to $(1 - h_i^{(m)})^{-1}$. The term $h_i^{(m)}$ is the i^{th} diagonal element of the projection matrix $\boldsymbol{P}^{(m)}$.

Hansen and Racine (2012) assume $\boldsymbol{H}^*$ to be a discrete set of $\left\{0, \frac{1}{N}, \frac{2}{N}, ..., 1\right\}$ for some positive integer N. Obtaining $\boldsymbol{w}$ following equation (A43) with condition $\boldsymbol{w} \in \boldsymbol{H}^*$, is a quadratic optimization process. Note that while there is a difference between our continuous $\mathcal{H}$ set defined in equation (A48) and $\mathcal{H}^*$, this should be negligible in practice since N can take any value.

D.3 Heteroskedasticity-Robust C^p Model Averaging

Liu and Okui (2013) also use the same model set-up to propose the heteroskedasticity-robust C^p (HRC^p) model averaging estimator for linear regression models with heteroskedastic errors. They demonstrate the asymptotic optimality of the HRC^p estimator when the error term exhibits heteroskedasticity. Liu and Okui (2013) propose computing the weights by the following feasible HRC^p criterion

$$\text{HR}C^p(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{P}(\boldsymbol{w})\boldsymbol{y})^\top(\boldsymbol{y} - \boldsymbol{P}(\boldsymbol{w})\boldsymbol{y}) + 2\sum_{i=1}^{n}\hat{\boldsymbol{e}}_i^2 p_{ii}(\boldsymbol{w}) \tag{A44}$$

with $\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in\mathcal{H}}{\arg\min}\ \text{HR}C^p(\boldsymbol{w})$. Obtaining $\boldsymbol{w}$ following (A44) with condition $\boldsymbol{w} \in \mathcal{H}$ is a quadratic optimization process.

Equation (A44) includes a preliminary estimate $\hat{e}_i$ that must be obtained prior to estimation. Liu and Okui (2013) discuss several ways to obtain $\hat{e}_i$ in practice. When the models are nested, Liu and Okui (2013) suggest using the residuals from the largest model. When the models are non-nested, they recommended constructing a model that contains all the regressors in the potential models and use the predicted residuals from the estimated model. In addition, a degree-of-freedom correction on $\hat{e}_i$ is recommended to improve finite-sample properties. For example, when the m^{th} model is used to obtain $\hat{e}_i$, we can use

$$\hat{\boldsymbol{e}} = \sqrt{n/(n-k^{(m)})}(\boldsymbol{I} - \boldsymbol{P}^{(m)})\boldsymbol{y}$$

instead of $(\boldsymbol{I} - \boldsymbol{P}^{(m)})\boldsymbol{y}$ to generate the preliminary estimate $\hat{e}_i$.

D.4 Iterative HRC^p Model Averaging

Liu and Okui (2013) also consider an iterative procedure in the presence of too many regressors, a common feature of big data sources. The procedure takes the following steps

1. Begin with an initial estimate $\hat{\sigma}_i$ using one selected model (Liu and Okui (2013) recommended using the largest model). This initial estimate can always be written as $\hat{\sigma}_i(\hat{\boldsymbol{w}}^0)$, with $\boldsymbol{w}^0$ being a special weight vector such that the selected model is assigned weight 1 and 0s for all other models.

2. Plug $\hat{\sigma}_i(\hat{\boldsymbol{w}}^0)$ in the HRC^p criterion function defined in equation (A44) and obtain the next round $\hat{\boldsymbol{w}}^1$.

3. Using $\hat{\boldsymbol{w}}^1$, we obtain the average residual $\hat{e}_i(\hat{\boldsymbol{w}}^1)$ and hence $\hat{\sigma}_i(\boldsymbol{w}^1)$. We then use $\hat{\sigma}_i(\boldsymbol{w}^1)$ to generate the next round weight vector.

4. Repeat steps (2) and (3) until weight vector $\hat{w}^j$ is obtained that satisfies $\left|\widehat{\text{HRC}^p}(\hat{\boldsymbol{w}}^j) - \widehat{\text{HRC}^p}(\hat{\boldsymbol{w}}^{j-1})\right| \leq \varphi$, where φ is a predetermined tolerance level (usually a small number).

A problem with this iterative process is that it can be computationally demanding, since multiple steps of quadratic optimization are required. To overcome this problem, we can either choose a relatively large φ or fix the total number of iterations.

D.5 Hetero-robust Prediction Model Averaging (HPMA) Method

Our setup is similar to both Wan, Zhang, and Zou (2010) and Liu and Okui (2013) by allowing the candidate models to be non-nested. We observe a random sample (y_i, x_i) for

$i = 1, ..., n$, in which y_i is a scalar and $x_i = (x_{i1}, x_{i2}, ...)$ is countably infinite. We consider the following data generating process (DGP)

$$y_i = \mu_i + e_i, \quad \mu_i = \sum_{j=1}^{\infty} \beta_j x_{ij}, \quad \mathbb{E}(e_i|x_i) = 0 \tag{A45}$$

for $i = 1, ..., n$ and μ_i can be considered as the conditional mean $\mu_i = \mu(x_i) = \mathbb{E}(y_i|x_i)$ that is converging in mean square.[7] We assume the error term to be heteroskedastic by letting $\sigma_i^2 = \mathbb{E}(e_i^2|x_i)$ denotes the conditional variance which is allowed to depend on x_i.

Now we consider a set of M candidate models. We allow the M models to be non-nested. The m^{th} candidate model that approximates the DGP in equation (A45) is

$$y_i = \sum_{j=1}^{k^{(m)}} \beta_j^{(m)} x_{ij}^{(m)} + b_i^{(m)} + e_i, \tag{A46}$$

for $m = 1, ..., M$, where $x_{ij}^{(m)}$ for $j = 1, ..., k^{(m)}$ denotes the regressors, $\beta_j^{(m)}$ denotes the coefficients, and $b_i^{(m)} \equiv \mu_i - \sum_{j=1}^{k^{(m)}} \beta_j^{(m)} x_{ij}^{(m)}$ is the modeling bias.

Define $\boldsymbol{y} = [y_1, ..., y_n]^\top$, $\boldsymbol{\mu} = [\mu_1, ..., \mu_n]^\top$, and $\boldsymbol{e} = [e_1, ..., e_n]^\top$. The DGP in equation (A45) can be presented by $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{e}$. Let $\boldsymbol{X}^{(m)}$ be a full rank $n \times k^{(m)}$ matrix of independent variables with $(i, j)^{th}$ element being $x_{ij}^{(m)}$. The estimator of $\boldsymbol{\mu}$ from the m^{th} model is

$$\hat{\boldsymbol{\mu}}^{(m)} = \boldsymbol{X}^{(m)} \left(\boldsymbol{X}^{(m)^\top} \boldsymbol{X}^{(m)} \right)^{-1} \boldsymbol{X}^{(m)^\top} \boldsymbol{y} = \boldsymbol{P}^{(m)} \boldsymbol{y},$$

where $\boldsymbol{P}^{(m)} = \boldsymbol{X}^{(m)} (\boldsymbol{X}^{(m)^\top} \boldsymbol{X}^{(m)})^{-1} \boldsymbol{X}^{(m)^\top}$ for all M. Similarly, the residual is $\hat{\boldsymbol{e}}^{(m)} = \boldsymbol{y} - \hat{\boldsymbol{\mu}}^{(m)} = (\boldsymbol{I}_n - \boldsymbol{P}^{(m)}) \boldsymbol{y}$ for all m. Since $\boldsymbol{P}^{(m)}$ is $n \times n$ for each m, we follow standard model averaging procedure and construct an averaged projection matrix $\boldsymbol{P}(\boldsymbol{w})$:

$$\boldsymbol{P}(\boldsymbol{w}) = \sum_{m=1}^{M} w^{(m)} \boldsymbol{P}^{(m)}, \tag{A47}$$

where $\boldsymbol{P}(\boldsymbol{w})$ is a weighted average of all potential $\boldsymbol{P}^{(m)}$. Due to its structure, $\boldsymbol{P}(\boldsymbol{w})$ is symmetric but not idempotent. The variable $\boldsymbol{w} = [w^1, w^2, ..., w^M]^\top$ is a weight vector we defined in the unit simplex in $\mathbb{R}^M$,

$$\mathcal{H} \equiv \left\{ \boldsymbol{w} \in [0, 1]^M : \sum_{m=1}^{M} w^{(m)} = 1 \right\}. \tag{A48}$$

[7] Convergence in mean square implies that $\mathbb{E}(\mu_i - \sum_{j=1}^{k} \beta_j x_{ij})^2 \to 0$ as $k \to \infty$.

Then, the model averaging estimator of μ is

$$\hat{\boldsymbol{\mu}}(\boldsymbol{w}) = \sum_{m=1}^{M} w^{(m)} \hat{\boldsymbol{\mu}}^{(m)} = \sum_{m=1}^{M} w^{(m)} \boldsymbol{P}^{(m)} \boldsymbol{y} = \boldsymbol{P}(\boldsymbol{w}) \boldsymbol{y}. \tag{A49}$$

Similarly, we define the averaged residual as

$$\hat{\boldsymbol{e}}(\boldsymbol{w}) = \sum_{m=1}^{M} w^{(m)} \hat{\boldsymbol{e}}^{(m)} = (\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{w})) \boldsymbol{y}. \tag{A50}$$

The performance of a model averaging estimator crucially depends on its choice of the weight vector w. Xie (2015) proposed a predictive model averaging (PMA) method that selects w through a convex optimization of a PMA criterion function of Amemiya (1980). One merit of the PMA method is that no preliminary estimates are required. The limitation of the PMA method is that the error term is required to be homoskedastic.

In the spirit of Liu and Okui (2013), we extend the PMA method to a heteroskedastic-robust predictive model averaging (HPMA) method with the following criterion function

$$\text{HPMA}(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{P}(\boldsymbol{w})\boldsymbol{y})^{\top} (\boldsymbol{y} - \boldsymbol{P}(\boldsymbol{w})\boldsymbol{y}) + 2 \sum_{i=1}^{n} \left(\hat{\boldsymbol{e}}_i(\boldsymbol{w})\right)^2 p_{ii}(\boldsymbol{w}), \tag{A51}$$

where $\boldsymbol{P}(\boldsymbol{w})$ is defined in (A47), $\hat{\boldsymbol{e}}_i(\boldsymbol{w})$ is the i^{th} element in $\hat{\boldsymbol{e}}(\boldsymbol{w})$ defined in equation (A50), $p_{ii}(\boldsymbol{w})$ is the i^{th} diagonal term in $\boldsymbol{P}(\boldsymbol{w})$. We estimate the weighting vector following

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w} \in \mathcal{H}}{\arg\min}\, \text{HPMA}(\boldsymbol{w}).$$

Similar to PMA, obtaining $\hat{w}$ from HPMA with restrictions $\boldsymbol{w} \in \mathcal{H}$ is a convex optimization process.

D.5.1 Asymptotic Optimality

In this subsection, we investigate the asymptotic properties of the HPMA estimator of w. We demonstrate that the proposed HPMA estimator is asymptotically optimal, in the sense of achieving the lowest possible mean squared error.

Let the average squared error loss and the corresponding l_2 type risk be

$$L(\boldsymbol{w}) = (\hat{\boldsymbol{\mu}}(\boldsymbol{w}) - \boldsymbol{\mu})^{\top} (\hat{\boldsymbol{\mu}}(\boldsymbol{w}) - \boldsymbol{\mu}), \tag{A52}$$

$$R(\boldsymbol{w}) = \mathbb{E} L(\boldsymbol{w}), \tag{A53}$$

where $\hat{\boldsymbol{\mu}}(\boldsymbol{w})$ is defined in equation (A49). To prove the optimality of HPMA, we assume the following regularity conditions similar to those demonstrated in Liu and Okui (2013),

Assumption AS 1 *There exists* $\epsilon > 0$ *such that* $\min_{1\leq i\leq n} \sigma_i^2 > \epsilon$.

Assumption AS 2 $\mathbb{E}(e_i^{4G}|\boldsymbol{x}_i) \leq \kappa < \infty$ *for some integer* $1 \leq G < \infty$ *and for some* κ.

Assumption AS 3 $M\xi^{-2G}\sum_{m=1}^{M}\left(R(\boldsymbol{w}_m^0)\right)^G \to 0$ *as* $n \to \infty$, *where* $\xi \equiv \inf_{w\in\mathcal{H}} R(\boldsymbol{w})$ *and* $\boldsymbol{w}_m^0$ *is a vector whose* m^{th} *element is 1 and all other elements are 0s.*

Assumption AS 4 $\max_{1\leq m\leq M}\max_{1\leq i\leq n} p_{ii}^{(m)} = O(n^{-1/2})$, $p_{ii}^{(m)}$ *is the* i^{th} *diagonal element of* $P^{(m)}$.

Assumptions AS 1-AS 4 correspond to Assumptions 2.1-2.4 in Liu and Okui (2013). Assumptions AS 1 and AS 2 establish bounds on the error terms and conditional moments. Assumptions AS 3 is a convergence condition that requires ξ goes to infinity faster than M and $\max_m R(\boldsymbol{w}_m^0)$. Assumptions AS 4 is a standard convergence condition on projection matrices.

Assumption AS 5 $\max_{1\leq m\leq M}\xi^{-1}\tilde{p}\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}^{(m)})\boldsymbol{\mu} \xrightarrow{p} 0$ *and* $\max_{1\leq m\leq M} M^2\xi^{-2G}\tilde{p}^{2G}\big(\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}^{(m)})\boldsymbol{\mu}\big)^G \xrightarrow{p} 0$, *where* $\tilde{p} \equiv \sup_{w\in\mathcal{H}}\max_{1\leq i\leq n}(p_{ii}(\boldsymbol{w}))$.

Assumption AS 6 $\max_{1\leq m\leq M}\xi^{-1}\tilde{p}\boldsymbol{e}^\top\boldsymbol{P}^{(m)}\boldsymbol{e} \xrightarrow{p} 0$, $\max_{1\leq m\leq M}\xi^{-1}\tilde{p}tr(\boldsymbol{P}^{(m)}\boldsymbol{\Omega}) \xrightarrow{p} 0$, *and* $\max_{1\leq m\leq M} M^2\xi^{-2G}\tilde{p}^{2G}\big(tr(\boldsymbol{P}^{(m)})\big)^G \xrightarrow{p} 0$, *where* $\tilde{p}$ *is defined in Assumption AS 5 and* $\boldsymbol{\Omega}$ *is an* $n\times n$ *diagonal matrix with* σ_i^2 *being its* i^{th} *diagonal element.* .

Assumption AS 5 requires that the bias from the worst potential model is small and Assumption AS 6 states that the associated variance be small. Similar requirements can be found in Wan, Zhang, and Zou (2010), which implies that some pre-selection procedures are always needed not just for the sake of computational efficiency, but also to maintain asymptotic optimality.[8] Finally, we demonstrate the optimality of HPMA estimator in the following Theorem.

[8]Frequentist model averaging usually involves a constraint optimization (quadratic, convex, etc.) process that can be quite computationally demanding when the set of approximation models is large. A pre-selection procedure can reduce the total number of models by removing some poorly constructed models following certain criteria, therefore, improves computation efficiency. On the other hand, conditions like Assumptions AS 5 and AS 6 are frequently used (Wan, Zhang, and Zou (2010), Liu and Okui (2013), Xie (2015), etc) in demonstrating asymptotic optimality. As argued in Wan, Zhang, and Zou (2010), a necessary condition for Assumptions AS 5 and AS 6 type conditions to hold is removing some poorly constructed models (by a pre-selection procedure) before commencing the model averaging process. See Xie (2017) for a detailed discussion of various pre-selection methods for frequentist model averaging.

Theorem 1 *Let Assumptions AS 1-AS 6 hold, as $n \to \infty$, we have*

$$\frac{L(\hat{\boldsymbol{w}})}{\inf_{w\in\mathcal{H}} L(\boldsymbol{w})} \xrightarrow{p} 1, \tag{A54}$$

where $L(\boldsymbol{w})$ is defined in equation (A52) and $\hat{w}$ is the HPMA estimator.

Proof of Theorem 1 Our proof follows Liu and Okui (2013) and Xie (2015). Let $\bar{\boldsymbol{P}}(\boldsymbol{w})$ be a diagonal matrix whose i^{th} diagonal element is $p_{ii}(\boldsymbol{w})$. Let $\hat{e}_i(\boldsymbol{w})$ as the i^{th} element of $\hat{\boldsymbol{e}}(\boldsymbol{w})$. Because:

$$\begin{aligned}
\widehat{\mathrm{HPMA}}(\boldsymbol{w}) &= (\boldsymbol{y}-\hat{\boldsymbol{\mu}}(\boldsymbol{w}))^\top(\boldsymbol{y}-\hat{\boldsymbol{\mu}}(\boldsymbol{w})) + 2\sum_{i=1}^{n}\hat{\boldsymbol{e}}_i^2(\boldsymbol{w})p_{ii}(\boldsymbol{w}) \\
&= \mathrm{HRC}^p(\boldsymbol{w}) + 2\left(\sum_{i=1}^{n}\hat{e}_i^2(\boldsymbol{w})p_{ii}(\boldsymbol{w}) - \mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{P}(\boldsymbol{w}))\right).
\end{aligned}$$

where $\mathrm{HPC}^p(\boldsymbol{w})$ takes another form of the heteroskedasticity-robust model averaging method Liu and Okui (2013) proposed in (A44) such that

$$\mathrm{HPC}^p(\boldsymbol{w}) = (\boldsymbol{y}-\boldsymbol{P}(\boldsymbol{w})\boldsymbol{y})^\top(\boldsymbol{y}-\boldsymbol{P}(\boldsymbol{w})\boldsymbol{y}) + 2\mathrm{tr}\big(\boldsymbol{\Omega}\boldsymbol{P}(\boldsymbol{w})\big), \tag{A55}$$

where $\boldsymbol{\Omega}$ is an $n\times n$ diagonal matrix with σ_i^2 being its i^{th} diagonal element.

Theorem 1 of Liu and Okui (2013) showed that under Assumptions AS 1 to AS 3

$$\sup_{\boldsymbol{w}\in\mathcal{H}}\left\{\mathrm{HRC}^p(\boldsymbol{w})\Big/R(\boldsymbol{w})\right\} \xrightarrow{p} 0.$$

Therefore, we just need to prove that

$$\sup_{w\in\mathcal{H}}\left\{\left|\sum_{i=1}^{n}\hat{e}_i^2(\boldsymbol{w})p_{ii}(\boldsymbol{w}) - \mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{P}(\boldsymbol{w}))\right|\Big/R(\boldsymbol{w})\right\} \xrightarrow{p} 0 \tag{A56}$$

LHS of equation (A56) can be rewritten as

$$\begin{aligned}
&\sup_{w\in\mathcal{H}}\left\{\left|\sum_{i=1}^{n}\hat{\boldsymbol{e}}_i^2(\boldsymbol{w})p_{ii}(\boldsymbol{w}) - \mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{P}(\boldsymbol{w}))\right|\Big/R(\boldsymbol{w})\right\} \\
\leq\ &\sup_{w\in\mathcal{H}}|\hat{\boldsymbol{e}}(\boldsymbol{w})^\top\bar{\boldsymbol{P}}(\boldsymbol{w})\hat{\boldsymbol{e}}(\boldsymbol{w}) - \mathbb{E}(\boldsymbol{e}^\top\bar{\boldsymbol{P}}(\boldsymbol{w})\boldsymbol{e})|/\xi \\
\leq\ &\sup_{w\in\mathcal{H}}\{|\hat{\boldsymbol{e}}(\boldsymbol{w})^\top\bar{\boldsymbol{P}}(\boldsymbol{w})\hat{\boldsymbol{e}} - \boldsymbol{e}^\top\bar{\boldsymbol{P}}(\boldsymbol{w})\boldsymbol{e}| + |\boldsymbol{e}^\top\bar{\boldsymbol{P}}(\boldsymbol{w})\boldsymbol{e} - \mathbb{E}(\boldsymbol{e}^\top\bar{\boldsymbol{P}}(\boldsymbol{w})\boldsymbol{e})|\}/\xi.
\end{aligned} \tag{A57}$$

where $\boldsymbol{e}(\boldsymbol{w})$ is defined in (A50), $\bar{\boldsymbol{P}}(\boldsymbol{w})$ is an $n\times n$ diagonal matrix with $p_{ii}(\boldsymbol{w})$ being its i^{th}

diagonal element, and ξ is defined in Assumption AS 3. The first term in (A57) is

$$\begin{aligned} & \hat{\boldsymbol{e}}(w)^\top \bar{\boldsymbol{P}}(w)\hat{\boldsymbol{e}}(w) - \boldsymbol{e}^\top \bar{\boldsymbol{P}}(w)\boldsymbol{e} \\ = \; & \boldsymbol{\mu}^\top (\boldsymbol{I} - \boldsymbol{P}(w))\bar{\boldsymbol{P}}(w)(\boldsymbol{I} - \boldsymbol{P}(w))\boldsymbol{\mu} + 2\boldsymbol{\mu}^\top (\boldsymbol{I} - \boldsymbol{P}(w))\bar{\boldsymbol{P}}(w)(\boldsymbol{I} - \boldsymbol{P}(w))\boldsymbol{e} \\ & + \boldsymbol{e}^\top (\boldsymbol{I} - \boldsymbol{P}(w))\bar{\boldsymbol{P}}(w)(\boldsymbol{I} - \boldsymbol{P}(w))\boldsymbol{e} - \boldsymbol{e}^\top \bar{\boldsymbol{P}}(w)\boldsymbol{e}. \end{aligned}$$

We have

$$\sup_{w\in\mathcal{H}} \boldsymbol{\mu}^\top (\boldsymbol{I} - \boldsymbol{P}(w))\bar{\boldsymbol{P}}(w)(\boldsymbol{I} - \boldsymbol{P}(w))\boldsymbol{\mu}/\xi \leq \tilde{p} \max_{1\leq m\leq M} \boldsymbol{\mu}^\top (\boldsymbol{I} - \boldsymbol{P}^{(m)})\boldsymbol{\mu}/\xi \xrightarrow{p} 0 \tag{A58}$$

by Assumption AS 6. Next, we consider the term

$$\boldsymbol{e}^\top (\boldsymbol{I} - \boldsymbol{P}(w))\bar{\boldsymbol{P}}(w)(\boldsymbol{I} - \boldsymbol{P}(w))\boldsymbol{e} - \boldsymbol{e}^\top \bar{\boldsymbol{P}}(w)\boldsymbol{e} = -2\boldsymbol{e}^\top \boldsymbol{P}(w)\bar{\boldsymbol{P}}(w)\boldsymbol{e} + \boldsymbol{e}^\top \boldsymbol{P}(w)\bar{\boldsymbol{P}}(w)\boldsymbol{P}(w)\boldsymbol{e},$$

where

$$\sup_{w\in\mathcal{H}} \boldsymbol{e}^\top \boldsymbol{P}(w)\bar{\boldsymbol{P}}(w)\boldsymbol{P}(w)\boldsymbol{e}/\xi \leq \tilde{p} \max_{1\leq n\leq M} \boldsymbol{e}^\top \boldsymbol{P}^{(m)}\boldsymbol{e}/\xi \xrightarrow{p} 0. \tag{A59}$$

by Assumption AS 6. For the term $\boldsymbol{e}^\top \boldsymbol{P}(w)\bar{\boldsymbol{P}}(w)\boldsymbol{e}$, we note that

$$\begin{aligned} \mathbb{E}(\boldsymbol{e}^\top \boldsymbol{P}(w)\bar{\boldsymbol{P}}(w)\boldsymbol{e}) &= \mathbb{E}\left(\boldsymbol{e}^\top \sum_{m=1}^{M} w^{(m)}\boldsymbol{P}^{(m)}\bar{\boldsymbol{P}}(w)\boldsymbol{e}\right) = \sum_{m=1}^{M} \mathbb{E}(\boldsymbol{e}^\top w^{(m)}\boldsymbol{P}^{(m)}\bar{\boldsymbol{P}}(w)\boldsymbol{e}) \\ &= \sum_{m=1}^{M} \mathbb{E}(w^{(m)}\boldsymbol{e}^\top P^{(m)}\bar{\boldsymbol{P}}(w)\boldsymbol{e}) = \sum_{m=1}^{M} \mathbb{E}(w^{(m)} tr(\boldsymbol{P}^{(m)}\bar{\boldsymbol{P}}(w)\boldsymbol{e}\boldsymbol{e}^\top) \\ &= \sum_{m=1}^{M} w^{(m)} tr(\bar{\boldsymbol{P}}(w)\boldsymbol{P}^{(m)}\boldsymbol{\Omega}). \end{aligned}$$

Therefore,

$$\sup_{w\in\mathcal{H}} \mathbb{E}(\boldsymbol{e}^\top \boldsymbol{P}(w)\bar{\boldsymbol{P}}(w)\boldsymbol{e})/\xi \leq \max_{1\leq m\leq M} \xi^{-1}\tilde{p} tr(\boldsymbol{P}^{(m)}\boldsymbol{\Omega}) \xrightarrow{p} 0$$

by Assumption AS 6. Moreover, using Chebyshev's inequality and Theorem 2 of Whittle (1960), for any $\delta > 0$, we have

$$\begin{aligned} & \Pr\left\{\sup_{w\in\mathcal{H}} \left|(\boldsymbol{e}^\top \boldsymbol{P}(w)\bar{\boldsymbol{P}}(w)\boldsymbol{e}) - \mathbb{E}(\boldsymbol{e}^\top \boldsymbol{P}(w)\bar{\boldsymbol{P}}(w)\boldsymbol{e})\right|/\xi > \delta\right\} \\ \leq \; & \sum_{l=1}^{M}\sum_{m=1}^{M} \mathbb{E}\left\{\frac{[(\boldsymbol{e}^\top \boldsymbol{P}^{(l)}\bar{\boldsymbol{P}}(w_m^0)\boldsymbol{e}) - \mathbb{E}(\boldsymbol{e}^\top \boldsymbol{P}^{(l)}\bar{\boldsymbol{P}}(w_m^0)\boldsymbol{e})]^{2G}}{\delta^{2G}\xi^{2G}}\right\} \\ \leq \; & \delta^{-2G}\xi^{-2G} \sum_{l=1}^{M}\sum_{m=1}^{M} C_1 \left\{\sum_{i=1}^{n}\sum_{j=1}^{n} (p_{ij}^{(l)})^2 p_{ii}^2(w_m^0)[\mathbb{E}(e_i^{4G})]^{1/2G}[\mathbb{E}(e_i^{4G})]^{1/2G}\right\}^G \end{aligned}$$

$$
\begin{aligned}
&\leq C_1 \max_{1\leq j\leq n} \mathbb{E}(e_i^{4G})\delta^{-2G}\xi^{-2G}\tilde{p}^{2G}\sum_{l=1}^{M}\sum_{m=1}^{M}\left\{\sum_{i=1}^{n}\sum_{i=1}^{n}(p_{ij}^{(l)})^2\right\}^G \\
&= C_2 \max_{1\leq l\leq M} \delta^{-2G}\xi^{-2G}M^2\tilde{p}^{2G}[\mathrm{tr}(\boldsymbol{P}^{(l)})]^G \to 0
\end{aligned}
$$

by Assumption AS 6, where C_1 is a constant and $C_2 \equiv C_1 \max_{1\leq i\leq n}\mathbb{E}(e_i^{4G})$ is a bounded constant according to Assumption AS 2. It follows that

$$
\sup_{\boldsymbol{w}\in\mathcal{H}} (\boldsymbol{e}^\top \boldsymbol{P}(\boldsymbol{w})\bar{\boldsymbol{P}}(\boldsymbol{w})\boldsymbol{e})/\xi = o_p(1). \tag{A60}
$$

Noting that $\mathbb{E}[\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}(\boldsymbol{w}))\bar{\boldsymbol{P}}(\boldsymbol{w})(\boldsymbol{I}-\boldsymbol{P}(\boldsymbol{w}))\boldsymbol{e}] = 0$, we again use Chebyshev's inequality and Theorem 2 of Whittle (1960) to show that

$$
\begin{aligned}
&\Pr\left\{\sup_{\boldsymbol{w}\in\mathcal{H}}\left|\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}(\boldsymbol{w}))\bar{\boldsymbol{P}}(\boldsymbol{w})(\boldsymbol{I}-\boldsymbol{P}(\boldsymbol{w}))\boldsymbol{e}\right|/\xi > \delta\right\} \\
\leq\ & \sum_{l=1}^{M}\sum_{m=1}^{M}\mathbb{E}\left\{\frac{[\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}^{(l)})\bar{\boldsymbol{P}}(\boldsymbol{w}_m^0)(\boldsymbol{I}-\boldsymbol{P}^{(l)})\boldsymbol{e}]^{2G}}{\delta^{2G}\xi^{2G}}\right\} \\
\leq\ & \delta^{-2G}\xi^{-2G}M\sum_{m=1}^{M}C_3\left\{\sum_{i=1}^{n}\gamma_{im}^2[\mathbb{E}(e_i^{2G})]^{1/G}\right\}^G,
\end{aligned}
$$

where γ_{im} is the i^{th} element of $\max_{1\leq l\leq M}\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}^{(l)})\bar{\boldsymbol{P}}(\boldsymbol{w}_m^0)(\boldsymbol{I}-\boldsymbol{P}^{(l)})$, and C_3 is a constant. We now have that

$$
\delta^{-2G}\xi^{-2G}M\sum_{m=1}^{M}C_3\left\{\sum_{i=1}^{n}\gamma_{im}^2[\mathbb{E}(e_i^{2G})]^{1/G}\right\}^G \leq \delta^{-2G}\xi^{-2G}M\sum_{m=1}^{M}C_4\left\{\sum_{i=1}^{n}\gamma_{im}^2\right\}^G,
$$

where $C_4 \equiv C_3\max_{1\leq i\leq n}\mathbb{E}(e_i^{2G})$ is a bounded constant according to Assumption AS 2 and

$$
\begin{aligned}
\sum_{i=1}^{n}\gamma_{im}^2 &= \max_{1\leq l\leq M}\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}^{(l)})\bar{\boldsymbol{P}}(\boldsymbol{w}_m^0)(\boldsymbol{I}-\boldsymbol{P}^{(l)})(\boldsymbol{I}-\boldsymbol{P}^{(l)})\bar{\boldsymbol{P}}(\boldsymbol{w}_m^0)(\boldsymbol{I}-\boldsymbol{P}^{(l)})\boldsymbol{\mu} \\
&\leq \max_{1\leq l\leq M}(\tilde{p})^2\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}^{(l)})\boldsymbol{\mu}.
\end{aligned}
$$

Therefore, it holds that

$$
\delta^{-2G}\xi^{-2G}\sum_{m=1}^{M}C_4\left\{\sum_{i=1}^{n}\gamma_{jm}^2\right\}^G \leq \max_{1\leq l\leq M}\delta^{-2G}\xi^{-2G}C_4\tilde{p}^{2G}M^2\left\{\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}^{(m)})\boldsymbol{\mu}\right\}^G \to 0
$$

by Assumption AS 5. Therefore, we have

$$\sup_{w\in\mathcal{H}}\left|\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{P}(\boldsymbol{w}))\bar{\boldsymbol{P}}(\boldsymbol{w})(\boldsymbol{I}-\boldsymbol{P}(\boldsymbol{w}))\boldsymbol{e}\right|/\xi \xrightarrow{p} 0. \tag{A61}$$

By (A58), (A59), (A60), and (A61), we have that the first term in (A57)

$$\sup_{w\in\mathcal{H}}\left|\hat{\boldsymbol{e}}(\boldsymbol{w})^\top\bar{\boldsymbol{P}}(\boldsymbol{w})\hat{\boldsymbol{e}}(\boldsymbol{w})-\boldsymbol{e}^\top\bar{\boldsymbol{P}}(\boldsymbol{w})\boldsymbol{e}\right|/\xi \xrightarrow{p} 0. \tag{A62}$$

Similarly, for the second term in (A57), using Chebyshev's inequality and Theorem 2 of Whittle (1960), for any $\delta>0$, we have

$$\begin{aligned}
&\Pr\left\{\sup_{w\in\mathcal{H}}\left|\boldsymbol{e}^\top\bar{\boldsymbol{P}}(\boldsymbol{w})\boldsymbol{e}-\mathbb{E}\left(\boldsymbol{e}^\top\bar{\boldsymbol{P}}(\boldsymbol{w})\boldsymbol{e}\right)\right|/\xi>\delta\right\}\\
\leq\ & \sum_{m=1}^{M}\mathbb{E}\left\{\frac{[\boldsymbol{e}^\top\bar{\boldsymbol{P}}(w_m^0)\boldsymbol{e}-\mathbb{E}\left(\boldsymbol{e}^\top\bar{\boldsymbol{P}}(\boldsymbol{w}_m^0)\boldsymbol{e}\right)]^{2G}}{\delta^{2G}\xi^{2G}}\right\}\\
\leq\ & \delta^{-2G}\xi^{-2G}\sum_{m=1}^{M}C_5\left\{\sum_{i=1}^{n}p_{ii}^2(\boldsymbol{w}_m^0)[\mathbb{E}(e_i^{4G})]^{1/G}\right\}^G\\
\leq\ & C_6\max_{1\leq j\leq n}\mathbb{E}(e_i^{4G})\delta^{-2G}\xi^{-2G}\sum_{m=1}^{M}\left\{\sum_{i=1}^{n}p_{ii}^2(\boldsymbol{w}_m^0)\right\}^G\\
\leq\ & C_6\delta^{-2G}\xi^{-2G}\sum_{m=1}^{M}[\mathrm{tr}[\left(\boldsymbol{P}(\boldsymbol{w}_m^0)\right)^2]]^G\\
=\ & C_6\delta^{-2G}\xi^{-2G}\left(\inf_{1\leq i\leq M}\sigma_i^2\right)^{-G}\sum_{m=1}^{M}[\mathrm{tr}[\inf_{1\leq i\leq M}\sigma_i^2\left(\boldsymbol{P}(\boldsymbol{w}_m^0)\right)^2]]^G\\
\leq\ & C_6\delta^{-2G}\xi^{-2G}\left(\inf_{1\leq i\leq M}\sigma_i^2\right)^{-G}\sum_{m=1}^{M}[\mathrm{tr}[\boldsymbol{\Omega}\left(\boldsymbol{P}(\boldsymbol{w}_m^0)\right)^2]]^G\\
=\ & C_6\delta^{-2G}\xi^{-2G}\left(\inf_{1\leq i\leq M}\sigma_i^2\right)^{-G}\sum_{m=1}^{M}[\mathrm{tr}[\boldsymbol{\Omega}\boldsymbol{P}(\boldsymbol{w}_m^0)]]^G\\
\leq\ & C_6\delta^{-2G}\xi^{-2G}\sum_{m=1}^{M}\left[R(\boldsymbol{w}_m^0)\right]^G\to 0,
\end{aligned}$$

where C_5 is a constant and $C_6\equiv C_5\max_{1\leq i\leq n}\mathbb{E}(e_i^{4G})$ is a bounded constant according to Assumption AS 2. The last inequality is due to

$$R(\boldsymbol{w}_m^0)\ =\ \mathbb{E}(L\left(\boldsymbol{w}_m^0\right))=\mathbb{E}\left[\left(\boldsymbol{P}^{(m)}\boldsymbol{y}-\boldsymbol{\mu}\right)^\top\left(\boldsymbol{P}^{(m)}\boldsymbol{y}-\boldsymbol{\mu}\right)\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left(\boldsymbol{P}^{(m)}(\boldsymbol{\mu}+\boldsymbol{e})-\boldsymbol{\mu}\right)^{\top}\left(\boldsymbol{P}^{(m)}(\boldsymbol{\mu}+\boldsymbol{e})-\boldsymbol{\mu}\right)\right] \\
&= \mathbb{E}\left[\left(\left(\boldsymbol{P}^{(m)}-\boldsymbol{I}\right)\boldsymbol{\mu}-\boldsymbol{P}^{(m)}\boldsymbol{e}\right)^{\top}\left(\left(\boldsymbol{P}^{(m)}-\boldsymbol{I}\right)\boldsymbol{\mu}-\boldsymbol{P}^{(m)}\boldsymbol{e}\right)\right] \\
&= \boldsymbol{\mu}^{\top}\left(\boldsymbol{P}^{(m)}-\boldsymbol{I}\right)^{\top}\left(\boldsymbol{P}^{(m)}-\boldsymbol{I}\right)\boldsymbol{\mu}-2\mathbb{E}\left[\boldsymbol{\mu}^{\top}\left(\boldsymbol{P}^{(m)}-\boldsymbol{I}\right)^{\top}\boldsymbol{P}^{(m)}\boldsymbol{e}\right]+\mathbb{E}\left[\boldsymbol{e}^{\top}\boldsymbol{P}^{(m)}\boldsymbol{e}\right] \\
&= \boldsymbol{\mu}^{\top}\left(\boldsymbol{P}^{(m)}-\boldsymbol{I}\right)^{\top}\left(\boldsymbol{P}^{(m)}-\boldsymbol{I}\right)\boldsymbol{\mu}+\operatorname{tr}[\boldsymbol{\Omega}\boldsymbol{P}^{(m)}] \\
&= \boldsymbol{\mu}^{\top}\left(\boldsymbol{P}^{(m)}-\boldsymbol{I}\right)^{\top}\left(\boldsymbol{P}^{(m)}-\boldsymbol{I}\right)\boldsymbol{\mu}+\operatorname{tr}[\boldsymbol{\Omega}\boldsymbol{P}(\boldsymbol{w}_m^0)] \\
&\geq \operatorname{tr}[\boldsymbol{\Omega}\boldsymbol{P}(\boldsymbol{w}_m^0)],
\end{aligned}
$$

where $\boldsymbol{P}(\boldsymbol{w}_m^0)=\boldsymbol{P}^{(m)}$ and the expectation is conditional on X. Therefore, we have

$$
\sup_{w\in\mathcal{H}}\left|\boldsymbol{e}^{\top}\bar{\boldsymbol{P}}(\boldsymbol{w})\boldsymbol{e}-\mathbb{E}\left(\boldsymbol{e}^{\top}\bar{\boldsymbol{P}}(\boldsymbol{w})\boldsymbol{e}\right)\right|/\xi \xrightarrow{p} 0. \tag{A63}
$$

Results of (A62) and (A63) imply that condition (A56) hold. This completes the proof.

D.6 Lasso, Post Model Selection by Lasso, and Double Lasso

Consider the linear regression model:

$$
y_i=\boldsymbol{x}_{0i}^{\top}\boldsymbol{\beta}_0+\sum_{j=1}^{p}x_{ji}\beta_j+u_i
$$

for $i=1,...,n$, where $\boldsymbol{x}_{0i}$ is $k_0\times 1$ and x_{ji} is scalar for $j\geq 1$. Let

$$
\begin{aligned}
\boldsymbol{\beta} &= \left[\boldsymbol{\beta}_0^{\top},\beta_1,...,\beta_p\right]^{\top} \\
\boldsymbol{x}_i &= \left[\boldsymbol{x}_0^{\top},x_{1i},...,x_{pi}\right]^{\top}
\end{aligned}
$$

and define the matrices $\boldsymbol{X}$ and $\boldsymbol{y}$ by stacking observations. The OLS estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{y}$. Consider a constrained least-squares estimate $\tilde{\boldsymbol{\beta}}$ subject to the constraint $\beta_1=\beta_2=...=0$. The Lasso estimator shrinks $\hat{\boldsymbol{\beta}}$ towards $\tilde{\boldsymbol{\beta}}$ by solving

$$
\hat{\boldsymbol{\beta}}^{L}=\arg\min_{\boldsymbol{\beta}}\frac{1}{2n}\sum_{i=1}^{n}(y_i-\boldsymbol{x}_i^{\top}\boldsymbol{\beta})^2+\lambda\sum_{j=1}^{p}|\beta_j|, \tag{A64}
$$

where λ is the tuning parameter that controls the penalty term. In practice, researchers either assign λ to take on a specific value or use k-fold cross-validation to determine the

optimal λ. A common choice is to pick λ to minimize 5-fold cross-validation. In general, the benefits from applying the Lasso in place of OLS exist in settings where either the number of regressors exceeds the number of observations since it involves shrinkage, or in settings where the number of parameters is not small relative to the sample size and some form of regularization is necessary.

The drawback of k-fold cross-validation is its lack of computational efficiency. For example, using five-fold cross-validation, the Lasso computation procedure will need to be carried out over 200 times. This computational inefficiency becomes especially significant when either the sample size is large or the number of variables is large. Thus, we follow Belloni and Chernozhukov (2013) and ex ante pick the number of explanatory variables that will not have their coefficient shrunk to zero, a form of post model selection by Lasso.

The double-lasso regression is similar to the post model selection by Lasso. The goal is to identify covariates for inclusion in two steps, finding those that predict the dependent variable and those that predict the independent variable of interest. Without loss of generality, we focus on the case with a single focal independent variable of interest, x_{0i}, and we want to know how it relates to dependent variable y_i. The double-Lasso variable selection procedure can be carried out as follows:

Step 1. Fit a lasso regression predicting the dependent variable, and keeping track of the variables with non-zero estimated coefficients:

$$y_i = c_1 + \sum_{j=1}^{p} x_{ji}\beta_j + u_i,$$

where c_1 is a constant.

Step 2. Fit a lasso regression predicting the focal independent variable, keeping track of the variables with non-zero estimated coefficients:

$$x_{0i} = c_2 + \sum_{j=1}^{p} x_{ji}\beta_j + u_i,$$

where c_2 is a constant. If x_{0i} is an effectively randomized treatment, no covariates should be selected in this step.

Step 3. Fit a linear regression of the dependent variable on the focal independent variable, including the covariates selected in either of the first two steps:

$$y_i = c_3 + x_{0i}\beta_0 + \sum_{k \in A} x_{ki}\beta_k + u_i,$$

where c_3 is a constant, A is the union of the variables estimated to have non-zero coefficients in Steps 1 and 2.

D.7 More Details on the Econometric Theory

In this section, we prove the asymptotic optimality of Mallows-type model averaging estimator under the constraint of screened model set. Our proof is inspired by the work of Zhang, Yu, Zou, and Liang (2016) who demonstrated the asymptotic optimality of Kullback-Leibler (KL) type model averaging estimators under screened model set. We extend their results, allowing their findings to be applied to a broader set of model averaging estimators.

We impose the following condition on the total number of candidate models.

Condition 0 *The total number of candidate models M is finite.*

We require the total number of candidate models to be finite such that they do not increase with the sample size. Note that (A48) does not hold with an infinite M.

We then lay out the following conditions that have been verified in the existing literature such as White (1982).

Condition 1 *We have $\|\boldsymbol{X}^\top \boldsymbol{\mu}_0\| = O(n)$ and $\|\boldsymbol{X}^\top \boldsymbol{\epsilon}\| = O_p(n^{1/2})$.*

Condition 2 *Conditions that sustain the asymptotic optimality of Mallows-type model averaging estimators (homoskedasticity or heteroskedasticity-robust) under given unscreened candidate model set in the original paper.*

Note that our proof is built upon the conditions that sustain the asymptotic optimality of Mallows-type model averaging estimators under given unscreened candidate model set. For example, see either equations (7) and (8) in Wan, Zhang, and Zou (2010), or assumptions 1 to 3 in Xie (2015), or assumptions 2.1 to 2.7 in Liu and Okui (2013). Condition 2 corresponds to these suppositions and would change slightly as we adopt different model averaging estimators.

For each approximation model m, we can define its mean squared error as

$$L(\boldsymbol{\beta}_m) \equiv (\boldsymbol{\mu}(\boldsymbol{\beta}_m) - \boldsymbol{\mu}_0)^\top (\boldsymbol{\mu}(\boldsymbol{\beta}_m) - \boldsymbol{\mu}_0), \tag{A65}$$

where $\boldsymbol{\mu}_0$ is the true value and $\boldsymbol{\mu}(\boldsymbol{\beta}_m) = \boldsymbol{X}\boldsymbol{\beta}_m$. Note that in our definition, all $\boldsymbol{\beta}_m$ for $m = 1, ..., M$ are $k \times 1$ vector, in which certain coefficients are set to 0 if the associated independent variables are not included in model m. Let $\boldsymbol{\beta}_m^*$ be the coefficient that minimizes equation (A65) such that $\boldsymbol{\beta}_m^* = \arg\min L(\boldsymbol{\beta}_m)$. The coefficient vector $\boldsymbol{\beta}_m^*$ minimizes the mean squared error of model m with respect to the true prediction value $\boldsymbol{\mu}_0$, which is different from $\hat{\boldsymbol{\beta}}_m$ that minimizes the sum squared residual (SSR) of model m.

We define the following averaged coefficients

$$\hat{\boldsymbol{\beta}}(\boldsymbol{w}) \equiv \sum_{m=1}^{M} w_m \hat{\boldsymbol{\beta}}_m \quad \text{and} \quad \boldsymbol{\beta}^*(\boldsymbol{w}) \equiv \sum_{m=1}^{M} w_m \boldsymbol{\beta}_m^*$$

Since $\hat{\boldsymbol{\mu}}(\boldsymbol{w}) = \sum_{m=1}^{M} w_m \boldsymbol{X}\hat{\boldsymbol{\beta}}_m = \boldsymbol{X}\sum_{m=1}^{M} w_m \hat{\boldsymbol{\beta}}_m = \boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{w})$, we define $\boldsymbol{\mu}^*(\boldsymbol{w}) \equiv \boldsymbol{X}\boldsymbol{\beta}^*(\boldsymbol{w})$ and the associated mean squared error can be written as

$$L^*(\boldsymbol{w}) = \left(\boldsymbol{\mu}^*(\boldsymbol{w}) - \boldsymbol{\mu}_0\right)^\top \left(\boldsymbol{\mu}^*(\boldsymbol{w}) - \boldsymbol{\mu}_0\right). \tag{A66}$$

We then define the ξ_n as

$$\xi_n = \inf_{\boldsymbol{w}\in\mathcal{H}} L^*(\boldsymbol{w}), \tag{A67}$$

which is the lowest possible value of $L^*(\boldsymbol{w})$ under set $\mathcal{H}$. Although the mean squared error $L^*(\boldsymbol{w})$ is based on a different averaged coefficients $\boldsymbol{\beta}^*(\boldsymbol{w})$, it is closely related to the $L(\boldsymbol{w})$ defined in (A52).

We impose the following condition on ξ_n

Condition 3 $n\xi_n^{-2} = o(1)$.

Condition 3 requires that ξ_n grows at a rate no slower than $n^{1/2}$. This condition is identical to the Condition (C.3) of Zhang, Yu, Zou, and Liang (2016) and is also implied by Conditions (7) and (8) of Ando and Li (2014).

Lemma 1 *Given Conditions 1-3, we have*

$$\sup_{\boldsymbol{w}\in\mathcal{H}} \frac{|L(\boldsymbol{w}) - L^*(\boldsymbol{w})|}{L^*(\boldsymbol{w})} = o_p(1), \tag{A68}$$

$$\sup_{\boldsymbol{w}\in\mathcal{H}} \frac{|C(\boldsymbol{w}) - \sum_{i=1}^{n}\sigma_i^2 - L^*(\boldsymbol{w})|}{L^*(\boldsymbol{w})} = o_p(1). \tag{A69}$$

Proof of Lemma 1 In line with the Theorem 3.2 of White (1982), under regularity conditions such that A1-A6 of White (1982) hold, it is straightforward to show that $\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m^* = O_p(n^{-1/2})$. Therefore,

$$\hat{\boldsymbol{\beta}}(\boldsymbol{w}) - \boldsymbol{\beta}^*(\boldsymbol{w}) = \sum_{m=1}^{M} w_m\left(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m^*\right) = O_p(n^{-1/2}) \tag{A70}$$

holds uniformly for $\boldsymbol{w} \in \mathcal{H}$.

By Taylor expansion and Condition 1,

$$L^*(\boldsymbol{w}) = L(\boldsymbol{w}) + 2\boldsymbol{X}^\top\left(\boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{w}) - \boldsymbol{\mu}_0\right)\left(\boldsymbol{\beta}^*(\boldsymbol{w}) - \hat{\boldsymbol{\beta}}(\boldsymbol{w})\right) + o_p(1)$$

$$= \quad L(\boldsymbol{w}) + O_p(n^{1/2}) + o_p(1),$$

which implies the order of $\sup_{\boldsymbol{w}\in\mathcal{H}} |L(\boldsymbol{w}) - L^*(\boldsymbol{w})|$ must be smaller or equal to $O_p(n^{1/2})$. Given Condition 3, we can obtain (A68).

Moreover, for Mallows-type criterion, we have

$$\begin{aligned} C(\boldsymbol{w}) &= \left(\boldsymbol{y} - \hat{\boldsymbol{\mu}}(\boldsymbol{w})\right)^\top \left(\boldsymbol{y} - \hat{\boldsymbol{\mu}}(\boldsymbol{w})\right) + 2\sigma^2 k \\ &= L(\boldsymbol{w}) + \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon} + 2\boldsymbol{\epsilon}^\top \left(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}}(\boldsymbol{w})\right) + 2\sigma^2 k \\ &= L^*(\boldsymbol{w}) + \left(L(\boldsymbol{w}) - L^*(\boldsymbol{w})\right) + \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon} + 2\boldsymbol{\epsilon}^\top \boldsymbol{X}\left(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}(\boldsymbol{w})\right) + 2\sigma^2 k. \end{aligned}$$

Therefore, by Condition 2,

$$\sup_{\boldsymbol{w}\in\mathcal{H}} |C(\boldsymbol{w}) - L^*(\boldsymbol{w})| \leq \sup_{\boldsymbol{w}\in\mathcal{H}} |L(\boldsymbol{w}) - L^*(\boldsymbol{w})| + 2\sup_{\boldsymbol{w}\in\mathcal{H}} |\boldsymbol{\epsilon}^\top \boldsymbol{X}\left(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}(\boldsymbol{w})\right)| + \sum_{i=1}^n \sigma_i^2 + o_p(1).$$

Note that the term $\sum_{i=1}^n \sigma_i^2$ can be simplified as $n\sigma^2$ if we assume homoskedasticity. Following Condition 1 and results in (A68), we have the order of $\sup_{\boldsymbol{w}\in\mathcal{H}} |C(\boldsymbol{w}) - \sum_{i=1}^n \sigma_i^2 - L^*(\boldsymbol{w})|$ being smaller than $O_p(n^{1/2})$. Hence, we obtain (A69) and complete the proof. ∎

Once Lemma 1 is established, we can prove Theorem 1 with the following steps.

Proof of Theorem 1 Our proof follows Zhang, Yu, Zou, and Liang (2016). Define $a(\boldsymbol{w}) = C(\boldsymbol{w}) - \sum_{i=1}^n \sigma_i^2 - L(\boldsymbol{w})$. As demonstrated in Lemma 1, Assumption 1, and Conditions 1 to 3, it is straightforward to show that, as $n \to \infty$

$$\sup_{\boldsymbol{w}\in\mathcal{H}} \left|\frac{a(\boldsymbol{w})}{L^*(\boldsymbol{w})}\right| \xrightarrow{p} 0, \tag{A71}$$

$$\sup_{\boldsymbol{w}\in\mathcal{H}} \left|\frac{v_n}{L^*(\boldsymbol{w})}\right| \xrightarrow{p} 0, \tag{A72}$$

$$\sup_{\boldsymbol{w}\in\mathcal{H}} \left|\frac{L^*(\boldsymbol{w})}{L(\boldsymbol{w})}\right| \xrightarrow{p} 1. \tag{A73}$$

Therefore,

$$\sup_{\boldsymbol{w}\in\mathcal{H}} \left|\frac{L^*(\boldsymbol{w})}{L(\boldsymbol{w}) - v_n}\right| \leq \left\{1 - \sup_{\boldsymbol{w}\in\mathcal{H}} \left|\frac{L(\boldsymbol{w}) - L^*(\boldsymbol{w})|}{L^*(\boldsymbol{w})}\right| - \sup_{\boldsymbol{w}\in\mathcal{H}} \left|\frac{v_n}{L^*(\boldsymbol{w})}\right|\right\}^{-1} \xrightarrow{p} 0, \tag{A74}$$

as $n \to \infty$. Then, we expand equation (7) of Theorem 1 as

$$\Pr\left\{\left|\frac{\inf_{\boldsymbol{w}\in\mathcal{H}} L(\boldsymbol{w})}{L(\tilde{\boldsymbol{w}})} - 1\right| > \delta\right\}$$

$$
\begin{aligned}
&= \Pr\left\{\left|\frac{\inf_{\boldsymbol{w}\in\tilde{\mathcal{H}}}\left(L(\boldsymbol{w})+a(\boldsymbol{w})\right)-a(\tilde{\boldsymbol{w}})-\inf_{\boldsymbol{w}\in\mathcal{H}}L(\boldsymbol{w})}{L(\tilde{\boldsymbol{w}})}\right|>\delta\right\}\\
&= \Pr\left\{\left|\frac{\inf_{\boldsymbol{w}\in\tilde{\mathcal{H}}}\left(L(\boldsymbol{w})+a(\boldsymbol{w})\right)-a(\tilde{\boldsymbol{w}})-\inf_{\boldsymbol{w}\in\mathcal{H}}L(\boldsymbol{w})}{L(\tilde{\boldsymbol{w}})}\right|>\delta, \boldsymbol{w}_n\in\tilde{\mathcal{H}}\right\}\\
&\qquad + \Pr\left\{\left|\frac{\inf_{\boldsymbol{w}\in\tilde{\mathcal{H}}}\left(L(\boldsymbol{w})+a(\boldsymbol{w})\right)-a(\tilde{\boldsymbol{w}})-\inf_{\boldsymbol{w}\in\mathcal{H}}L(\boldsymbol{w})}{L(\tilde{\boldsymbol{w}})}\right|>\delta, \boldsymbol{w}_n\notin\tilde{\mathcal{H}}\right\} \qquad \text{(A75)}
\end{aligned}
$$

By definitions of conditional and joint probabilities, we have

$$
\begin{aligned}
&\text{Right-hand-side of equation (A75)}\\
\leq\ & \Pr\left\{\left|\frac{\inf_{\boldsymbol{w}\in\tilde{\mathcal{H}}}\left(L(\boldsymbol{w})+a(\boldsymbol{w})\right)-a(\tilde{\boldsymbol{w}})-\inf_{\boldsymbol{w}\in\mathcal{H}}L(\boldsymbol{w})}{L(\tilde{\boldsymbol{w}})}\right|>\delta\middle|\boldsymbol{w}_n\in\tilde{\mathcal{H}}\right\}\Pr(\boldsymbol{w}_n\in\tilde{\mathcal{H}})\\
&\qquad + \Pr(\boldsymbol{w}_n\notin\tilde{\mathcal{H}})\\
\leq\ & \Pr\left\{\left|\frac{L(\boldsymbol{w}_n)+a(\boldsymbol{w}_n)-a(\tilde{\boldsymbol{w}})-\inf_{\boldsymbol{w}\in\mathcal{H}}L(\boldsymbol{w})}{L(\tilde{\boldsymbol{w}})}\right|>\delta\middle|\boldsymbol{w}_n\in\tilde{\mathcal{H}}\right\}\Pr(\boldsymbol{w}_n\in\tilde{\mathcal{H}})+\Pr(\boldsymbol{w}_n\notin\tilde{\mathcal{H}})\\
\leq\ & \Pr\left\{\left|\frac{L(\boldsymbol{w}_n)+a(\boldsymbol{w}_n)-a(\tilde{\boldsymbol{w}})-\inf_{\boldsymbol{w}\in\mathcal{H}}L(\boldsymbol{w})}{L(\tilde{\boldsymbol{w}})}\right|>\delta\right\}+\Pr(\boldsymbol{w}_n\notin\tilde{\mathcal{H}}). \qquad \text{(A76)}
\end{aligned}
$$

Following the definition of v_n defined in Assumption 1(i), we have

$$
\begin{aligned}
&\text{Right-hand-side of (A76)}\\
=\ & \Pr\left\{\left|\frac{v_n+a(\boldsymbol{w}_n)-a(\tilde{\boldsymbol{w}})}{L(\tilde{\boldsymbol{w}})}\right|>\delta\right\}+\Pr(\boldsymbol{w}_n\notin\tilde{\mathcal{H}})\\
\leq\ & \Pr\left\{\left|\frac{v_n}{L(\tilde{\boldsymbol{w}})}\right|+\left|\frac{a(\boldsymbol{w}_n)}{L(\tilde{\boldsymbol{w}})}\right|+\left|\frac{a(\tilde{\boldsymbol{w}})}{L(\tilde{\boldsymbol{w}})}\right|>\delta\right\}+\Pr(\boldsymbol{w}_n\notin\tilde{\mathcal{H}})\\
\leq\ & \Pr\left\{\sup_{\boldsymbol{w}\in\mathcal{H}}\left|\frac{v_n}{L(\boldsymbol{w})}\right|+\left|\frac{a(\boldsymbol{w}_n)}{\inf_{\boldsymbol{w}\in\mathcal{H}}L(\boldsymbol{w})}\right|+\sup_{\boldsymbol{w}\in\mathcal{H}}\left|\frac{a(\boldsymbol{w})}{L(\boldsymbol{w})}\right|>\delta\right\}+\Pr(\boldsymbol{w}_n\notin\tilde{\mathcal{H}})\\
\leq\ & \Pr\left\{\sup_{\boldsymbol{w}\in\mathcal{H}}\left|\frac{v_n}{L^*(\boldsymbol{w})}\right|\sup_{\boldsymbol{w}\in\mathcal{H}}\left|\frac{L^*(\boldsymbol{w})}{L(\boldsymbol{w})}\right|+\sup_{\boldsymbol{w}\in\mathcal{H}}\left|\frac{a(\boldsymbol{w})}{L^*(\boldsymbol{w})}\right|\sup_{\boldsymbol{w}\in\mathcal{H}}\left|\frac{L^*(\boldsymbol{w})}{L(\boldsymbol{w})-v_n}\right|\right.\\
&\qquad \left.+\sup_{\boldsymbol{w}\in\mathcal{H}}\left|\frac{a(\boldsymbol{w})}{L^*(\boldsymbol{w})}\right|\sup_{\boldsymbol{w}\in\mathcal{H}}\left|\frac{L^*(\boldsymbol{w})}{L(\boldsymbol{w})}\right|>\delta\right\}+\Pr(\boldsymbol{w}_n\notin\tilde{\mathcal{H}}). \qquad \text{(A77)}
\end{aligned}
$$

According to Conditions (A71), (A72), (A73), (A74), and Assumption 1(iii), we obtain that the right-hand-side of (A77) converge to 0 as $n\to\infty$. This completes the proof. ∎

E Additional Details on Data Collection and Related Literature

The data for this project was initially assembled in conjunction with the IHS Film Consulting unit for a industry driven exercise. Using the IHS annual screen digest, characteristics of all 178 films that were released in movie theaters and 143 films that were released for sale on DVD/Blu Ray were first collated. These characteristics were used to determine which words could be used in hashtags associated with each specific film. A member of IHS with expertise in the media industry examined this hashtag list and on occasion supplemented the terms.

Since Janys Analytics was simultaneously measuring the sentiment in all Twitter messages with current and historical Twitter data at the hourly level for separate projects with IHS, we used queries with the historical data to extract all messages that involved the specific terms presented on the hashtag list. The size of the historical data is large and all Twitter messages at the hourly level are stored in separate datasets on cloud computers. The queries led us to create daily datasets for all Twitter messages that were extracted based on terms from the hashtag list. These datasets were subsequently analyzed to provide our volume and sentiment measures as described below.

At this time, Janys Analytics adapted the Hannak, Anderson, Barrett, Lehmann, Mislove, and Riedewald (2012) algorithm to provide IHS with a hourly measure of purchasing intentions on Twitter. We used this algorithm on our extracted subset of Twitter messages to calculate the sentiment specific to each film. The algorithm involves textual analysis of movie titles and movie key words that were placed on the hashtag list. In each Twitter message that mentions a word from the hashtag list, sentiment is calculated by examining the emotion words and icons that are captured in the same Twitter message.

In total, each of 75,065 unique emotion words and icons that appeared in at least 20 of the universe of tweets between January 1st, 2009 to September 1st, 2009 are given a specific value that is determined using emotional valence. Emotional valence is a term frequently used in psychology that refers to the intrinsic attractiveness (positive valence) or aversiveness (negative valence) of an event, object, or situation. This algorithm calculates the sentiment index for the film as a weighted average of the sentiment of the scored words in all of the messages associated with a specific film during a time period is then calculated. This overall sentiment score can be interpreted as a measure of the propensity for which there is a positive emotion tweet related to that movie. Last and as discussed in the main text, since opinions regarding a film likely vary over time with the release of different marketing devices to both build awareness and increase anticipation, IHS film consulting unit suggested to calculate sentiment over different time periods.

We use the sentiment data before the release date in equations that forecast the opening weekend box office. After all, reverse causality issues would exist if we include sentiment data after the release date. Post box opening tweets are used to measure sentiment and are only considered in equations that forecast DVD and Blu-Ray sales.

We should note that the algorithm does not consider the network structure of Twitter. It is likely that there is what psychologists term as belief polarization and research in computer science has shown that individuals give more weight to messages from those that are considered strong ties relative to weak ties. That said, there has been substantial evaluation of the sentiment inference algorithm developed by Janys Analytics for IHS. Hannak, Anderson, Barrett, Lehmann, Mislove, and Riedewald (2012) compare this sentiment inference methodology score to one calculated by users of Amazon Mechanical Turk and that they are strongly positively correlated with $\rho = 0.6525$. An additional advantage is that the sentiment inference algorithm is easy to regularly update to readjust the frequency at which a specific word is associated with a positive emotion in calculating the initial values that enter the sentiment calculator to adjust to potential changes in the Twitter user population.

Briefly, we note the demographics of Twitter users differs markedly from the national population. Mislove, Jorgensen, Ahn, Onnela, and Rosenquist (2011) document that these users are predominately male and located in urban areas, but point out these calculations are based on self-reported profiles. Further, these authors note that the male bias is declining rapidly. Despite the self-selection of these users, we next discuss why we believe that this sample of users is likely highly correlated with the characteristics of moviegoers and DVD purchasers so is relevant to study. After all, research in marketing indicates that everyday consumers often seek like-minded amateurs' opinions (for example, Hannak, Anderson, Barrett, Lehmann, Mislove, and Riedewald (2012) and Holbrook (1999).

We examined monthly reports from iHS Markit, Stastica and the Mintel group that were initially independently prepared for industry sources during this sampling period. These reports provide information concerning the characteristics of moviegoers and DVD purchasers. In October 2013, Mintel group reported survey results suggestion that while there was not a gender difference in the probability of visiting a cinema in the last six months, men were roughly 35% more likely (23% vs. 15%) to be categorized as frequent moviegoers. To be categorized as a frequent moviegoer suggested having paid to watch at least four movies in the past month. Interestingly, the consulting companies conclude that higher attendance is not just due to the release of films that target a younger male demographic, but a larger share is due to theaters' technology enhancements such as 3D or IMAX.

Within these reports, a subset of survey respondents claim to be regular Twitter users. While Twitter does not provide information at a geographic level below self-reported state, this survey of consulting reports is suggestive that the male bias in Twitter population is consistent with heavy moviegoers who were slightly less likely to be deterred from theaters based on parking, location (or attracted by technology offerings). Across social media sites, a movie theater's proximity to home or work only influenced 45% of Twitter users attendance and the sole outlet with a statistically significant difference was Flickr.[9]

[9]Among all survey repondents nearly 57% of suggested that the proximity to a moviegoers' home or work is an important consideration; but young men (18-34) were the least likely to consider this a factor.

The reports indicate that the characteristics of DVD purchasers were similar to moviegoers with the sole exception of presence of children in the household,[10] which increased the likelihood of DVD purchases. Unfortunately, profiles on Twitter data do not allow us to determine if the user has a child. In 2012, box office sales in North America totaled $10.8 billion (Germain 2013) whereas retail home movie sales were $18 billion that included movie purchases and rentals (Orden 2013). Given the size of this market, film studios need to jointly consider when to release the DVD and the film in theatres. This topic has attracted significant attention and is reviewed by Ahmed and Sinha (2016) who additionally present a model of the optimal time to release a DVD following the box office release.

The pattern of sales for DVDs is different than theatre tickets presenting a second hurdle for film studios who effectively need to develop a pricing strategy to segment the market. Mortimer (2007) presents a discussion of the optimal pricing of home movies retail movie using a theoretical model. There is less empirical evidence on who buys a DVD and when, ranging from pre-orders to initial release to sales later on when the DVD is repackaged and offered at discounted prices in big box retailers to also target an audience seeking a second viewing.[11] The majority of empirical research in this market focuses on renting and not purchases but DVD sales at Amazon.com have been shown to increase immediately following a public TV broadcast of a film (Smith and Telang, 2009) or considered how piracy (i.e. via peer-to-peer file sharing networks) may influence retail movie unit sales (see e.g. Rob and Waldfogel (2007) or Smith and Telang (2010)).

E.1 Related Literature

Our study provides empirical evidence that relates to an interdisciplinary literature that explores how external information affects purchasing decisions in the movie industry. We briefly summarize several branches and explain how our findings contribute to this literature in this section as well as mention directions for further research. Prior to summarizing the extant literature, we wish to stress that few studies (discussed in the previous section) explore retail movie unit sales in isolation.[12] Among studies that focus on box-office sales, roughly half of them consider gross revenue as the outcome variable; whereas the remainder including our study aim to forecast opening weekend box office.[13]

[10] Industry analysts claim that this is not a surprise since these activities should be considered as complements to each other. Individuals who see movies in the theatre gain knowledge through trailers about movies that will be later available in the other medium and vice versa.

[11] Ingnoring this second viewing audience, Hui et al. (2008) suggest that sales for many DVD titles would follow a exponential-decay pattern. Our data does not consider the later sales if a DVD is rereleased at low prices in big box retailers.

[12] As such, we do not discuss this research also in part due to recent trends that have seen the global Blu-Ray and DVD market decline sharply in size since consumers are either buying digital or on demand copies of films or have a subscription to an online streaming service such as Netflix.

[13] We do not consider studies that consider the length of a movieâ€™s run in theatres as the outcome of interest, beyond noting that Moul (2007) considered the consequences of hetereskesdasticity with this

We chose this outcome since opening weekend box office is well-known to be crucial to the industry and the results are frequently reported in the popular press. Further, Einav (2007) reports that first week revenues make up to 40% of a film's total box office sales.

E.1.1 Initial Econometric Models and Investigation of Expert Reviews

The majority of published research focuses heavily on understanding what characteristics of films are associated with their earnings, with less attention paid to developing models for forecasting. Empirical research that investigate the determinants of box office revenues within the economics literature date back to Prag and Casavant (1994), who used a large sample of 652 films released in the US market. Their main findings were that advertising expenditures were positively related to total box office earnings, whereas films that were classified as dramas had a negative association. Numerous subsequent studies suggest a role played by the explanatory variables that were suggested by the IHS-Markit Film unit to include as potential predictors and controls in the MTV models. The importance of these control variables is grounded by evidence in the literature as Basuroy, Chatterjee, and Ravid (2003) finds that motion picture association rating has a strong relationship with revenue, an effect Terry, Terry, and De'Armond (2011) postulate is due to the potential audience size. Related to audience size is screens, that is unsurprisingly positively correlated with earnings (see e.g. Neelamegham and Chintagunta 1999; Basuroy, Chatterjee, and Ravid 2003; Moon, Bergey, and Iacobucci 2010). Film genre has been shown to also correlate with revenue as Dahl and DellaVigna (2008) and Prieto-Rodriguez, Gutierrez-Navratil, and Ateca-Amestoy (2014) report the action and violence genres yield better performance.

Thus, we postulate that forecasting models of film genre may differ in their specification by genre. Specification uncertainty provides additional motivation for allowing model uncertainty with either econometric or machine learning strategies. In other words, if the models that are frequently developed in the literature were used to forecast revenue for the film industry, they may neglect important parameter heterogeneity by film genre. Using more flexible machine learning algorithms than Lehrer and Xie (2017) we would anticipate significant gains in forecast accuracy since these algorithms capture many dimensions of this heterogeneity that is due to neglected non-linearities.

In marketing, the concept of word of mouth (WOM) is used to define the act of consumers providing information about goods, services, brands, or companies to other consumers. WOM is often referred to as a form of social learning or as an endogenous peer

outcome. Specifically, Moul (2007) investigates the word-of-mouth effect on individual film demand using a nested logit model, where word-of-mouth presents through the heteroskedasticity and serial correlation in the error term of the model. The results suggest that word of-mouth can explain roughly i) 38% of the variance in the unobservables, and ii) 10% of variation in consumer expectations. This study provides an additional model-based rationale for why heteroskedasticity needs to be considered when conducting forecasts for the film industry. In our study and evidence surveyed below, word of mouth is directly measured by the respective authors and is often the focus of interest.

effect in the economics literature. The potential role of WOM when forecasting for the film industry was initially considered in Hirschman and Pieros (1985) who presented a counter-intuitive result that positive criticism is negatively correlated with gross box office success. This finding did not replicate with data drawn in later periods and subsequent research repeatedly finds that expert reviews have a positive and significant effect on consumption (see e.g. Eliashberg and Shugan 1997; Chintagunta, Gopinath, and Venkataraman 2010; Moon, Bergey, and Iacobucci 2010, among others).

Research on whether WOM influences film revenue also highlight the potential of specification uncertainty from potentially neglecting important parameter heterogeneity. For example, Koschat (2012) finds evidence of heterogeneous effects of the valence of critics' reviewer ratings on opening weekend box office by film genre. Similarly, Gemser, Oostrum, and Leenders (2007) compares the effect of the valence and size of critical reviews in Dutch newspapers on art-house films versus mainstream films. They find that art house films benefit from having more reviews and suggest that coverage of any sort is better than no coverage at all. This finding mirrors some of our own presented in table 7 of the main text that the volume of WOM matters and in a differential manner by film budget.

The timing of when expert reviews appear does vary, in part since not every film critic is able to see a specific movie at the same point in time. This variation in when reviews are reported was considered in studies that forecast gross revenue by examining if there were large gains in including covariates that captured the valence in critics' reviews that occurred after a film's release. Specifically, Neelamegham and Chintagunta (1999) use Bayesian methods and demonstrated that their model which included post-release WOM surpassed the forecast accuracy of the model of Sawhney and Eliashberg (1996) that only used pre-release WOM information among the predictors, by 45 - 71%.

Despite this gain in forecast accuracy, in our paper we use pre-release WOM measures only. Post-theatre release WOM metrics can be used to forecast retail movie unit sales. We seek to avoid endogeneity concerns that arise from reverse causality between box-office (and retail movie unit sales) revenue and reviews. Each variable that we consider as a potential predictor including the measures of WOM, budget and screens are predetermined. To the best of our knowledge, Holbrook and Addis (2008) were the first to both note this endogeneity concern in the literature on WOM from expert reviewers on movie revenue and propose estimating a triangular system of equations.[14] More recently, Lee, Hosanagar, and Tan (2015) provide additional evidence that users' ratings of the quality of a movie are influenced by prior ratings of online user groups in a heterogeneous manner that depends on the movie's popularity. To overcome challenges posed by the endogeneity of social influence on user's reviews that may also explain user's purchasing decision, we follow the standard in the econometrics literature of using predetermined variables when estimating a peer effect given its link to WOM discussed above.

[14]Few studies deal with the endogeneity of WOM and later in this section we discuss Chintagunta, Gopinath, and Venkataraman (2010) which is an exception.

E.1.2 Incorporating Social Media in Forecasting Models

Over the last 15 years, research on the effects of WOM on movie-going has changed its focus from expert reviews to consider how qualitative aspect of online buzz (i.e., how customers perceive or feel about a product) influences the consumer decision-making process. This qualitative aspect is frequently referred to as electronic word of mouth (eWOM) and include Internet communications ranging from user reviews, tweets, blog posts, "likes", "pins" images and video testimonials. In practice, researchers operationalized eWOM in extant academic literature in multiple ways that we argue capture different aspects of their decision making process. In our study, we distinguish among the following eWOM metrics: volume and sentiment (also commonly referred to as valence). Recently, Houston, Kupfer, Hennig-Thurau, and Spann (2018) argue that valence measures prior to film release capture pre-release consumer buzz that proxies the extent to which consumers' are interested in a new product. This is argued to be a different phenomena than a valence measure post film release reinforcing our need to use a predetermined measure of eWOM as a predictor.

The majority of research studies that measure eWOM do so from a single source, such as a movie review site such as Yahoo!Movie (see e.g. Chintagunta, Gopinath, and Venkataraman 2010; Liu 2006), an online forum of blog posts (e.g. Gopinath, Chintagunta, and Venkataraman 2013), Twitter (see e.g. Rui, Liu, and Whinston 2011; Kaplan 2012)[15] or a single firm (Sonnier, McAlister, and Rutz, 2011). Results show that various social media signals ranging from online review and rating systems to sentiment analysis have a significant predictive value to predict performance in the box office. For example, Vujič and Zhang (2018) find that negative sentiment in eWOM is damaging to box office revenues. Chakravarty, Liu, and Mazumdar (2010) find professional movie reviews and popular word-of-mouth are related to consumers' frequency of movie attendance. Similarly, Gopinath, Chintagunta, and Venkataraman (2013) report that release day performance of a movie is impacted most by prerelease blog volume and advertising, whereas post-release performance is influenced by post-release blog valence and advertising. eWOM on social media sites ranging from Twitter and Weibo predicts more ticket sales (see e.g. Rui, Liu, and Whinston 2011; Ding, Cheng, Duan, and Jin 2016; Vujič and Zhang 2018; among others). Last, as an alternative metric of eWOM, studies have found that the amount of specific discussion pre-release on Twitter (Chintagunta, Gopinath, and Venkataraman, 2010) and Lehrer and Xie (2017) as well as the number of followers of an actor on Twitter (Vosoughi, Mohsenvand, and Roy, 2017) has predictive power for the movie success on box office.

In the majority of studies summarized above measures of WOM and eWOM are treated as exogenous. As noted above, to sidestep endogeneity concerns we use a strategy of using a predetermined measure that is frequently used in the economics literature focused

[15]With Twitter data as a covariate, Kaplan (2012) forecasts gross revenue, whereas our interest is to predict opening week earnings.

on the estimation of peer effects. An alternative empirical strategy that is also when estimating peer effects is to use instrumental variables, where the instruments are measures of peer characteristics and not oneself. Chintagunta, Gopinath, and Venkataraman (2010) follow such an approach and use measures such as average critic score for competing movies in the prior week, the average star power of competing movies in the previous week and the average proportion of movies of the same genre as the focal movie playing in the prior week. Each of these instruments is predetermined and in the empirical microeconomics literature the plausibility of the exclusion restriction assumption of instrumental variables is frequently debated. Since our focus is not on estimating parameters but on conducting forecasts,[16] we treat our eWOM measures as exogenous.

A branch of the marketing literature explores how to better incorporate pre-release buzz data in forecasting models and both Bandari, Asur, and Huberman (2012) and Xiong and Bharadwaj (2014) draw a distinction between cumulative measures versus including measures to capture buzz dynamics. While our analysis due to endogeneity concerns discussed above only considers pre-release measures defined over ad-hoc intervals in the pre-release period, *F*-tests of the coefficients on these variables in the GUM model presented in appendix table A5 would reject any restriction that these variables had an identical effect. In addition, the results in table 7 suggest there is further heterogeneity by movie type across the budget distribution. This is not a surprise as one of the motivations for our study is that different types of people will likely create pre-release buzz for different types of films.[17] Taken together, our results suggest that researchers need to be quite flexible in how they include pre-release buzz data both as there are many dimensions of heterogeneity and many machine learning algorithms can help identify these non-linear interactions between observed variables.

Many of the findings of the effects of eWOM on box office outcomes summarized above appear to mirror the results discussed in the previous subsection on the positive effects of WOM in expert reviews on these outcomes. Similarly, research suggests the possibility of neglected parameter heterogeneity in econometric models since studies the explored whether there are heterogeneous effects of eWOM by either film genre or movie market segmented into mainstream versus non-mainstream films (Yang, Hu, Winer, Assael, and Chen, 2012) have found evidence of significant differences. For example, Yang, Hu, Winer, Assael, and Chen (2012) provide evidence that the effect of eWOM volume on box office revenue is greater for mainstream movies and suggest this arises since films are

[16]To the best of our knowledge, Dellarocas, Zhang, and Awad (2007) were the first to show using a modified Bass diffusion model with data from the film industry that adding online movie ratings to their revenue-forecasting model significantly improves the model's predictive power.

[17]This finding is not unique to our valence and volume measures. Ding, Cheng, Duan, and Jin (2016) find with another valence measure, the pre-release â€œlikeâ€ on Facebook has a significantly positive impact on box office performance, whose effect increases closer to the release date. Similarly, Mestyán, Yasseri, and Kertész (2013) finds that the impact of online movie review sentiment on gross revenue becomes statistically significant three weeks after release. Specifically, 1-star review have a larger significant decline in revenue relative to the significant gains from 5-star reviews; whereas 2â€“4 star reviews have no significant impact.

experience goods with uncertain quality, eWOM volume provides a proxy for the credibility of the quality of the product, thereby increasing their confidence in attendance. Our empirical results appear consistent with this interpretation and as discussed in the introduction of the main text, there are several plausible rationales for why one should anticipate specification uncertainty with eWOM data.

The majority of studies outlined above measure eWOM from a single social network platform or in case of multiple platforms,[18] treat each platform as a distinct silo (Shruti, Roy, and Zeng, 2014). Both Tsao (2014) and Basuroy and Ravid (2014) contrast the role played by movie reviews by professional critics and ordinary consumers. They reach conflicting results where Tsao (2014) finds that without considering an interaction effect, potential moviegoers attach greater importance to consumer reviews than they do critical reviews. Consistent with Vujič and Zhang (2018) the influence of negative consumer reviews on movie selection is stronger than that of positive consumer reviews. In contrast, Basuroy and Ravid (2014) find that internet reviews matter, but expert opinions carry more weight. The difference in these findings likely arises since Basuroy and Ravid (2014) follow Chintagunta, Gopinath, and Venkataraman (2010) and treat eWOM as endogenous and address endogeneity in their work which may explain the different results. Among studies that contrasted eWOM from Twitter, YouTube, Yahoo! Movies and blogs on box office revenue, Baek, Oh, Yang, and Ahn (2017) showed that Twitter is the most influential platform in the initial stages of release, while Yahoo! Movies and blogs were more influential in the later stages.[19] Thus, for the outcome under consideration, we are using the most relevant source for eWOM .

In summary, while there has been some attempt to examine the role of eWOM on a movie's box office performance, there is a lack of consensus on factors affecting the movie box office performance. There is clear need for future work to relax the assumption that eWOM measures from alternative social media platforms should be treated independent of each other. However, findings often differ across studies for other reasons on how eWOM is measured.

E.1.3 Differences Across Studies Due to Machine Learning Algorithms

Machine learning algorithms are used in our study to first code sentiment in social media messages and subsequently to undertake forecasts. How messages/ reviews are coded and aggregated varies sharply across papers. For example, Liu (2006) simply manually coded each post as positive, negative, or neutral; whereas Lehrer and Xie (2017) used an algorithm due to Hannak, Anderson, Barrett, Lehmann, Mislove, and Riedewald (2012) that created a continuous measure of the emotion content of each message. Ravi (2015)

[18] Schweidel and Moe (2014) do not explore the film industry but provide evidence that consumers' sentiments towards a particular brand does differ across different social media platforms.

[19] Evidence from Shruti, Roy, and Zeng (2014) cast doubt on the influence of Facebook "likes" on movie revenue.

points out that an important limitation of the algorithms frequently used for sentiment analysis is they struggle with irony and sarcasm. Recent advances in the natural processing literature use deep learning to reduce this concern. Rui, Liu, and Whinston (2011) differ from earlier studies by weighting each eWOM in a tweet by the number of followers the author of each WOM message.[20]

In either case, the aggregated valence measure should be viewed as being measured with error. We note that our interest is strictly in conducting forecasts and this with regression models any bias in measuring sentiment would be classical in nature.[21]

Based on guidance from the IHS-Markit film consulting unit we were careful in using data in our sample to avoid periods of time where bots or fake reviews were prevalent on Twitter. This provides a small window to explore the effect of eWOM on Twitter as earlier periods were associated with very few users and later periods require one to develop algorithms to identify fake reviews. In the data science literature, there is evidence from Mayzlin, Dover, and Chevalier (2014) who contrast hotel reviews on Expedia.com and TripAdvisor.com, that biased user reviews impede review usefulness.[22] We note this as a direction for further research to also consider possibly disentangling the effect of eWOM from actual users versus so-called bots who are hypothesized to spread biased information online.

Turning to forecasting strategies, Kim, Hong, and Kang (2015) as well as Lehrer and Xie (2017) each find that the utilization of the combination of data from social media and machine learning-based algorithms made noticeable improvements to forecasting accuracy. Similarly Hur, Kang, and Cho (2016) illustrate the benefits of using both support vector regression, neural networks and a regression tree algorithm relative to multiple regression with respect to forecast accuracy in the Korean film industry.[23] In general, the sample sizes in these forecasting exercises as well as the studies reviewed in the prior subsection are small and generally consist of 150-200 films.[24]

Our study considers time-varying measures of two eWOM measures calculated from

[20]Recent work by Lehrer, Xie, and Zeng (2020) propose a new method on how to weight messages when aggregating social media data based on the time messages were posted. This strategy is shown to offer an advantage of handling parameter instability that may reflect jumps, which can introduce asymptotic bias to the averaging estimate.

[21]Thus, in a regression coefficient on any sentiment variables would be biased towards zero in absolute value and statistical insignificance. Thus, all evidence on the importance of sentiment likely reflects a lower bound but as shown in Tables A9 and A24, this covariate still improves forecast accuracy and is often chosen by the Lasso as one of the most important variables to forecast revenue.

[22]The idea is that there are no restrictions on who can post a review on TripAdvisor whereas Expedia makes it more challenging for fake reviewers to be uploaded.

[23]Many other papers explore a single machine learning algorithm, For example, when developing a predictive algorithm for motion picture revenues, Antipov and Pokryshevskaya (2017) and Zhou, Zhang, and Yi (2017) respectively only use a random forest-based model or deep neural networks.

[24]Briefly, Liu (2006), Chintagunta, Gopinath, and Venkataraman (2010), Elberse and Eliashberg (2003), Vujič and Zhang (2018) uses samples of 40, 148, 164, and 158 movies respectively; each collected over different time periods.

social media data as opposed to a single sentiment measure obtained from reviews and introduces hybrid strategies that combine econometrics with machine learning to conduct forecasts. Most importantly, we also explore a fuller suite of machine learning algorithms and provide insights on the trade-offs of using algorithms with the small samples used in these forecasting exercises. In the next subsection, we elaborate on this point and restate our main contribution to the empirical literature that investigates the association of characteristics of movies with film revenue.

E.1.4 Clarifying our Contributions to the Literature on the Role of eWOM on Movie Revenue and Forecasting Outcomes for Hollywood

In summary, research on forecasting film revenue tends to find that different eWOM measures are important predictors, irrespective of how they are measured. There is increased value to allowing for time-varying eWOM measures and substantial evidence of heterogeneity in the effects of eWOM predictors on the outcomes being considered. This heterogeneity was also exhibited in the earlier literature that considered WOM measures. The idea that WOM is reflected by heteroskedasticity underlies the empirical strategy used in Moul (2007) but is not considered in prior work that undertook forecasts. This prior work often found gains from using machine learning strategies to undertake these forecasts relative to econometric approaches. However, a full suite of machine learning algorithms was not considered and the majority of work did not illustrate the robustness of their findings to the choice of hyperparameters. In addition, the effect on retail movie unit sales upon initial release was rarely considered.

Our study considers heteroskedatic data and introduces new hybrid strategies for this setting with both revenue outcomes. Our empirical results confirm earlier work on initial box office earnings by illustrating heterogeneous effects of WOM/eWOM and can provide an explanation for why there are large gains with machine learning approaches. This will further highlight the value of SVR_{LS} and MASVR_{LS} since there will be even more covariates (or features) to consider as potential predictors and this estimator can accommodate such settings and allow for a very rich set of potential non-linear interactions. In section F.21 of the appendix, we discuss and illustrate how machine learning strategies can suggest the type of nonlinearities to include in linear econometric models to generate new insights on which characteristics explain film revenue. Last, as discussed in the concluding section of the main text, future work can also consider richer measures of eWOM from multiple social media platforms as well as richer data from a single social media platform data such as Twitter.

F More Empirical Results

This appendix consists of numerous subsections that provide further analyses and robustness checks of our main findings. OLS estimates of the GUM model are provided

in subsection F.1. Breusch-Pagan tests are provided in Table A5 show strong evidence of heteroskedasticity for both open box office and movie unit sales.

The first piece of evidence pertaining to using two social media measures versus one is obtained by comparing estimates across tables in subsection F.3 (see tables A7 and A8). Further evidence is shown in subsection F.17 and tables A9 and A23 for the proposed MAB and MASVRLS strategies. In subsection F.2 and F.6 we provide evidence of the relative prediction efficiency for double Lasso and Lasso based Strategies respectively. As observed in tables A6 and A12 – A14, the benchmark HRC^p outperforms all Lasso based methods considered. Finally, the evidence contrasting tables A12 – A14 present further evidence for why two social media measures are preferred to either one.

In subsection F.4, a Monte Carlo study is used to shed further light on the relative performance of ARMS and ARMSH under different scenarios related to what is the source of heteroskedasticity. Related, in subsection F.7 we present additional analyses that contrasts which models (and their contents) are selected by ARMS to ARMSH. These sections explain when differences between these methods could occur and why in our application, there were many similarities. Related to F.7, in subsection F.5 we present weights of, and contents of the top 5 models selected by the HRC^p estimator. These results continue to show that in practice, the model averaging estimator gives lots of weight to very few of all the potential models and is consistent with other applications of these methods including in policy oriented applications such as crime deterrence (Durlauf, Navarro, and Rivers, 2016).

In subsection F.8, we provide evidence that even when we restrict machine learning strategies to use the identical set of predictors as model screening choices made for model averaging that recursive partitioning methods yield more accurate forecasts. This shows that much of the gains we observed in our application come from the restrictiveness of the linear model and that additional gains can still be obtained by allowing for model uncertainty and considering that the data is heteroskedastic. Subsection F.9 provides formal evidence that the proposed MASVR$_{LS}$ method significantly outperforms other forecasting strategies considered in the main text (tables 3 – 5).

Subsection F.10 considers adding a model averaging flavor to a single regression tree (MART). For space considerations, we did not include this in the main text since as seen in the single figure A4 presented in subsection F.10, the MART method is outperformed by both MAB and MARF by a large margin in both heteroskedasticity scenarios. Thus, similar to the discussion in the statistical learning literature that forecasts from RT are unreliable and both bagging and random forest present improvement, we advocate only adding model averaging to strategies that used bagging or random forest to create subgroups.

As is well known, a model that fits well in sample may not be good for forecasting—a model may fit well in-sample, only to turn out to be useless in prediction. Consequently, it is common practice to select the model based on pseudo-out-of-sample fit from a sequence of recursive or rolling predictions. In subsection F.11, we compute the centered

R^2 of the main exercise, when applying the CART algorithm to the 10,001 training sets on the GUM (with twitter variable) and MTV (without twitter variable). We report the mean, median, 2.5% quantile, and 97.5% quantile of the R^2s under different values of n_E for both open box office and movie unit sales scenarios. We note that GUM yields higher R^2 than MTV in all cases. Note the p-values of F tests of their inclusion of Twitter variable are always below 0.01 even at the 2.5 percentile. These results continue to illustrate the importance of social media data in our forecasting exercises.

Subsection F.12 describes the selected parameters by the OLS-post-Lasso method. These results as described in a footnote in the main text reinforce the value of social media data in our application.

The table in subsection F.13 revisits the subset of data initially analyzed in Lehrer and Xie (2017). The authors placed a strong restriction on film budget, thereby limiting the amount of heterogeneity in their data. Only dimension reduction and econometric approaches were considered in that paper. We revisit the data using the off the shelf statistical learning methods of bagging and random forest with different numbers of explanatory variables considered. These methods do not assume a functional form and achieve large gains in accuracy relative to the benchmark and best performing estimator reported in Lehrer and Xie (2017). This analysis led us to consider a larger set of machine learning strategies in the main text, where we additionally relax the sampling restriction to include all films released over that sample period.

Subsection F.14 presents the results of the forecasting experiment in terms of absolute units of the loss function, as opposed of degree of loss to relative to the chosen baseline estimator. These results are measured in millions of dollars and the general finding is that the statistical learning methods generally reduce the variation relative to econometric approaches. The addition of model averaging in the leaves does not lead to further gains in efficiency but achieves higher MSFE and MAFE by reducing the bias. These results show that statistical learning methods are less variable and the incorporation of model averaging in place of the local constant model achieves gains by increasing the accuracy of forecasts. These results appear consistent with the patterns illustrated in the main text when discussing Figure 1.

In subsection F.15 we demonstrate that forecasts with HBART become more accurate as the number of bootstrapped samples increase, whereas the gains from MARF are smaller. This section provides a sense of the computational power needed to benefit from HBART. While the two algorithms rely on different splitting rules and tuning parameters, this section shows that the performance of HBART in our application also relies on setting the number of bootstrapped samples to be larger.

In Appendix F.16, we use MASVR_{LS} to better understand how social media should be accounted for.

To motivate the exercise in appendix F.17, theorems 4.3 and 4.4 of Scornet (2017) demonstrate that the consistency of random forests with binary outcome is valid for any

value of the number of covariates to split. This may seem surprising since with few covariates, the computational cost of the procedure is small compared to RT but the splitting direction may deviate sharply from the best splitting direction. Probst, Boulesteix, and Bischl (2019) provide evidence among the hyperparameters to tune in an application with 38 datasets that this tuning this parameter provides the biggest average improvement in forecasting among all possible RF hyperparameters. Our findings are that a moderate number of covariates is needed for forecasting open box office but a larger number is needed for retail movie unit sales. This suggests the default may be small with RF and MARF for retail movie unit sales, but the performance gain is minimal relative to using MASVR$_{LS}$ with defaults in our application. Most importantly, we do not find large differences in the results as the number of variables changes sharply.

In appendix F.18 attention is paid to different kernels to be used with SVR and SVR$_{LS}$. Probst and Boulesteix (2018) suggest that there is more benefits from changing hyperparameters of SVM algorithms than random forest. This tuning exercise allows us to examine model averaging weights that treat the error as homoskedastic using the Mallow's criterion (equation A40) and using the HPMA with the heteroskedastic error (equation A41). The findings are presented in Figure A9 and show that the loss function of SVR$_{LS}$ is preferred over SVR and that kernels that allow for more nonlinearities provide larger gains. As with the model selection results, we find very small gains when using HPMA instead of the Mallow's criterion reinforcing the largest gains come from neglected nonlinearities that lead to a heteroskedastic error with linear econometric models. Last, appendix F.19 presents the most comprehensive examination of different hyperparameter choices for alternative machine learning algorithms. We do not find large differences to the main results as hyperparameters change suggesting that the default values of the hyperparameters specified in software packages work reasonably well. We view these exercises as demonstrating the robustness of the main findings and stress a caveat that our investigation was not exhaustive so there remains a possibility that there are particular specific combinations of hyperparameters with each algorithm that may lead to changes in the ordering of forecast accuracy in the empirical horserace presented in the text.

Appendix F.20 presents estimates of a RT for box office revenue and illustrates how to incorporate the suggested nonlinearities into an estimable model. This allows a researcher to conduct statistical inference and examine marginal effects to supplement the analysis of variable importance in Tables 6 and 7 of the main text. In this section, we also discuss alternative machine learning strategies that empirical researchers could undertake to understand how covariates explain film revenue outcomes and contribute to the literature on how eWOM affects film outcomes surveyed in Appendix E.1. The results suggest that there are statistically significant threshold effects in the social media measures that are measured in the week prior to box office opening on film revenue. While the results find evidence for numerous significant nonlinear relationships between specific explanatory variables and film revenue, they also cast doubt on the importance of complex interactions between non-linear terms of the explanatory variables in our application.

Appendix F.21 presents the main results illustrated in Figures 4 and 5 of the main text

in tabular form. A brief discussion that mirrors the main text is provided so the subsection is self-contained.

F.1 OLS Estimates of the GUM Model

Table A5: OLS Estimates of the Generalized Unrestricted Model

Variable	Open Box Office		Movie Unit Sales	
	Coefficient	Robust S.E.	Coefficient	Robust S.E.
Genre				
Action	-1.6895	3.0838	-0.0622	0.1194
Adventure	4.6542	3.7732	-0.0967	0.1588
Animation	-1.9354	5.6046	0.8167*	0.3609
Biography	0.1229	4.2324	-0.0109	0.2015
Comedy	-0.9595	3.7382	-0.1287	0.1859
Crime	2.6461	2.7335	-0.0931	0.1052
Drama	-1.7884	3.6083	0.0139	0.1092
Family	2.6236	6.7679	-0.4118	0.3503
Fantasy	12.8881*	4.9159	0.5634	0.3937
Horror	3.0486	2.4376	-0.3655*	0.1441
Mystery	3.3377	2.4852	0.1414	0.1243
Romance	-2.5919	3.3696	-0.0986	0.0921
Sci-Fi	-0.3705	2.6569	0.0336	0.1391
Thriller	0.8643	2.9379	0.0306	0.1301
Rating				
PG	2.8901	5.4757	-0.6290	0.4196
PG13	1.8691	6.8517	-0.8369	0.5112
R	2.6378	6.6841	-0.7490	0.4964
Core Parameters				
Budget	0.1182*	0.0399	0.0035*	0.0016
Weeks	0.3738	0.2768	0.0447*	0.0109
Screens	6.1694*	1.3899	0.3215*	0.0526
Sentiment				
T-21/-27	-0.1570	0.6610	-0.0148	0.0241
T-14/-20	-0.9835	0.9393	-0.0040	0.0304
T-7/-13	-1.2435	1.0695	0.1802	0.1104
T-4/-6	0.2277	1.1775	-0.1708*	0.0842
T-1/-3	2.5070*	0.7509	-0.0422	0.0839
T+0			0.2172*	0.0864
T+1/+7			-0.0927*	0.0399
T+8/+14			0.0212	0.0234
T+15/+21			0.0085	0.0291
T+22/+28			-0.0808	0.1072
Volume				
T-21/-27	-97.5186*	31.6624	-1.6863	0.9608
T-14/-20	19.4109	38.6929	0.0724	1.1598
T-7/-13	-45.2885	30.9011	-1.8770	1.1417
T-4/-6	86.2881*	27.2008	2.5302*	0.7184
T-1/-3	18.9664*	5.1687	-1.2437*	0.4167
T+0			0.4423*	0.1064
T+1/+7			-0.2006	0.2404
T+8/+14			1.1195	0.9779
T+15/+21			0.4945	0.6281
T+22/+28			-0.3414	0.3104
Breusch-Pagan Statistic	**249.9485**		**207.3698**	
Breusch-Pagan p-value	**$<$0.0001**		**$<$0.0001**	
R-square	**0.7973**		**0.8016**	

Note: * indicates the associated variable is significant at 5% level.

F.2 Performance of Double-Lasso Strategy in Simulation Experiment

Table A6: Comparing Hetero-robust and Homo-efficient Model Screening Methods

n_E	OLS_{10}	OLS_{12}	OLS_{15}	HRC^p_{10}	HRC^p_{12}	HRC^p_{15}	Benchmark
Panel A: Open Box Office							
Mean Squared Forecast Error (MSFE)							
10	1.4388	1.5229	1.1787	1.4181	1.5075	1.1564	**1.0000**
20	1.6213	1.6090	1.2135	1.5898	1.5814	1.1854	**1.0000**
30	1.7625	1.6869	1.2597	1.7322	1.6714	1.2344	**1.0000**
40	1.8172	1.7028	1.2622	1.7745	1.6768	1.2548	**1.0000**
Mean Absolute Forecast Error (MAFE)							
10	1.2064	1.2131	1.0778	1.1962	1.2054	1.0680	**1.0000**
20	1.2356	1.2208	1.0880	1.2262	1.2173	1.0841	**1.0000**
30	1.2420	1.2273	1.0882	1.2331	1.2192	1.0833	**1.0000**
40	1.2475	1.2330	1.0845	1.2360	1.2187	1.0766	**1.0000**
Panel B: Movie Unit Sales							
Mean Squared Forecast Error (MSFE)							
10	1.3855	1.4254	1.4699	1.3645	1.3892	1.4364	**1.0000**
20	1.3562	1.3960	1.4022	1.3321	1.3651	1.3730	**1.0000**
30	1.2831	1.3096	1.3088	1.2733	1.2909	1.2821	**1.0000**
40	1.1793	1.2094	1.2499	1.1573	1.1807	1.2210	**1.0000**
Mean Absolute Forecast Error (MAFE)							
10	1.2604	1.2731	1.2840	1.2514	1.2616	1.2683	**1.0000**
20	1.2345	1.2541	1.2626	1.2273	1.2365	1.2472	**1.0000**
30	1.2014	1.2190	1.2314	1.1920	1.2053	1.2169	**1.0000**
40	1.1682	1.1878	1.2051	1.1565	1.1706	1.1880	**1.0000**

Note: Bold numbers denote the strategy with the best performance in that row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the HRC^p method presented in the last column. OLS_q and HRC^p_q stand for OLS and HRC^p with q number of covariates selected by the Lasso.

F.3 Additional Evidence on the Importance of Social Media Data

Table A7: OLS Estimates of Models with Sentiment Only

Variable	Open Box		Movie Unit	
	Coefficient	Robust S.E.	Coefficient	Robust S.E.
Genre				
Action	-11.8297	5.1756	-0.5991	0.2118
Adventure	1.8903	9.0801	-0.2221	0.3721
Animation	-8.6157	7.2188	0.3618	0.3987
Biography	-10.6777	7.3202	-0.3815	0.3079
Comedy	-6.1906	4.4094	-0.3875	0.2136
Crime	5.6338	3.7323	0.1658	0.1751
Drama	-4.3020	4.9879	-0.2661	0.1924
Family	-1.3797	8.0709	-0.3123	0.3741
Fantasy	19.2129	10.5968	0.8570	0.4906
Horror	-0.8574	4.6042	-0.6504	0.2190
Mystery	-4.1597	3.1965	-0.1284	0.1412
Romance	-1.3851	4.4953	0.1232	0.1784
Sci-Fi	0.6611	6.1694	0.1187	0.2989
Thriller	1.4062	5.2588	0.0971	0.2140
Rating				
PG	7.6872	7.0093	-1.0293	0.4639
PG13	21.6049	10.7996	-0.5286	0.5447
R	19.5326	10.5227	-0.5796	0.5433
Core Parameters				
Budget	0.1525	0.0827	0.0064	0.0033
Weeks	1.3267	0.5057	0.0943	0.0204
Screens	13.8708	2.9586	0.5949	0.1233
Sentiment				
T-21/-27	0.9289	0.7021	-0.0195	0.0292
T-14/-20	-0.7583	0.7503	0.0373	0.0366
T-7/-13	-1.1656	1.6137	0.3103	0.1303
T-4/-6	0.9664	2.1090	-0.0694	0.1113
T-1/-3	-0.1460	1.1729	-0.0401	0.1418
T+0			0.1238	0.1668
T+1/+7			-0.1016	0.0603
T+8/+14			0.0649	0.0372
T+15/+21			-0.0992	0.0411
T+22/+28			-0.1859	0.1286
R-square	**0.5322**		**0.6488**	

Note: * indicates the associated variable is significant at 5% level.

Table A8: OLS Estimates of Models with Volume Only

Variable	Open Box		Movie Unit	
	Coefficient	Robust S.E.	Coefficient	Robust S.E.
Genre				
Action	-1.7845	3.0495	-0.1049	0.1163
Adventure	4.8425	3.7630	0.0347	0.1508
Animation	-3.8178	5.2420	0.6189	0.3508
Biography	0.5099	4.4590	-0.1050	0.2038
Comedy	-0.5934	3.8404	-0.1896	0.1556
Crime	3.1958	2.6371	0.0043	0.0961
Drama	-1.9479	3.5767	-0.0280	0.1078
Family	3.6903	6.5546	-0.3090	0.3424
Fantasy	13.3327	4.9812	0.5544	0.3864
Horror	3.6698	2.5120	-0.2299	0.1305
Mystery	2.6945	2.5712	-0.0145	0.1100
Romance	-2.5929	3.4036	-0.0859	0.0909
Sci-Fi	-0.5145	2.7094	0.0015	0.1279
Thriller	0.6968	3.0682	-0.0407	0.1181
Rating				
PG	1.8990	5.2023	-0.3739	0.3662
PG13	1.6943	6.7034	-0.5650	0.4418
R	2.3396	6.4815	-0.5206	0.4475
Core Parameters				
Budget	0.1142	0.0396	0.0029	0.0016
Weeks	0.4335	0.2705	0.0424	0.0114
Screens	6.9067	1.4856	0.3422	0.0557
Volume				
T-21/-27	-97.6733	30.6043	-1.5188	0.9072
T-14/-20	21.1375	36.7023	-0.0649	1.1053
T-7/-13	-39.7233	31.2763	-1.6555	1.1440
T-4/-6	81.3088	27.3566	2.1988	0.6776
T-1/-3	18.1939	4.9561	-1.4011	0.3762
T+0			0.4675	0.1007
T+1/+7			-0.2659	0.2455
T+8/+14			1.6392	0.8910
T+15/+21			0.2306	0.5984
T+22/+28			-0.2764	0.3631
R-square	**0.8224**		**0.8445**	

Note: * indicates the associated variable is significant at 5% level.

F.4 Using Monte Carlo Study to Understand How Different Sources for Heteroskedasticity Affect Strategies

We found in forecasts of retail movie unit sales that the difference in the performance between PMA and HRC^p in table 3 in conjunction with the relative improved performance of ARMS presented in table 4 to be surprising. A potential explanation for these findings is the source of heteroskedasticity in the data. We examine the performance of five different model screening methods that are implied in the subscripts of the following model sets: $\mathcal{M}^K_{\text{GETS}}$, $\mathcal{M}^K_{\text{Lasso}}$, $\mathcal{M}^K_{\text{ARMS}}$, $\mathcal{M}^K_{\text{ARMSH}}$, and $\mathcal{M}^K_{\text{HRMS}}$.[25] Using data generated by the Monte Carlo design described in section 3, we compare the risks of each method:

$$\text{Risk}_i \equiv \frac{1}{n}\sum_{t=1}^{n}\left(\hat{\mu}_t(\mathcal{M}_i^K) - \mu_t\right)^2 \qquad \text{for } i = \text{GETS, Lasso, ARMS, ARMSH, and HRMS,}$$

where μ_t is the true fitted value (feasible in simulation) and $\hat{\mu}_t(\mathcal{M}_i^K)$ is the average fitted value obtained by HRC^p using specific candidate model set. Four different sample sizes ($n = 100, 200, 300$, and 400) are considered and the risk for each method - sample size pair is averaged across 10,000 simulation draws.

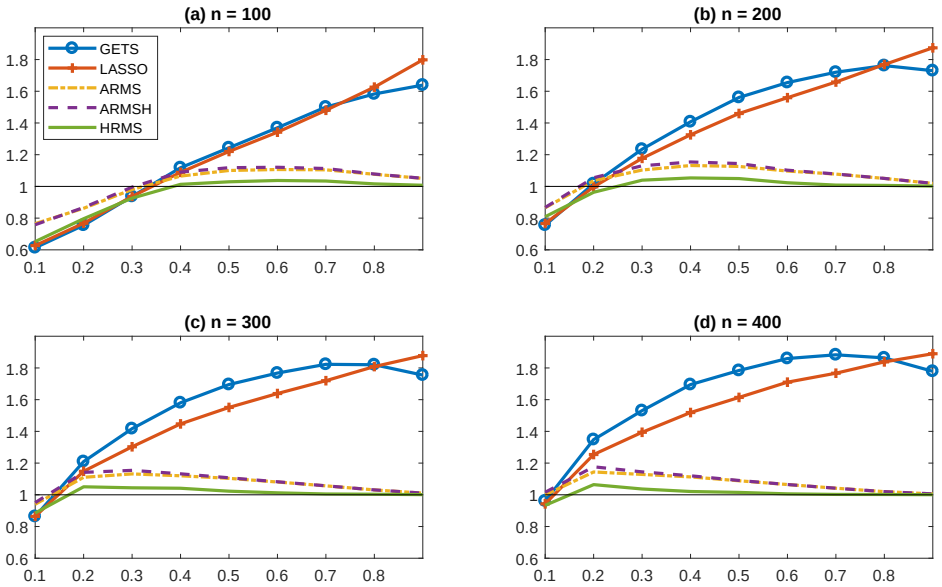
Figures A3 presents the results from this exercise where we normalize the risks by the risk of the infeasible optimal model. Each line presents the relative risks of each model screening method associated with R^2 from 0.1 to 0.9, respectively. Each sub-panel (a) to (d) presents the results for different sample sizes.

In virtually every panel of figures A3, HRMS has the best performance. In the random heteroskedasticity scenario, GETS and Lasso perform well only when R^2 is low. As R^2 increases, the relative improved performance of ARMS, ARMSH, and HRMS emerges. The performance of both ARMS and ARMSH more closely mimics HRMS at larger sample sizes. However, in simulations where heteroskedasticity arises due to neglected parameter heterogeneity both GETS and Lasso perform poorly, particularly when there is strong correlation among the regressors. The performance of both screening methods is relatively poorer when either the sample size or R^2 increases. In contrast, ARMS and ARMSH yield consistently better results that are similar with increasing n and R^2. Note that for both cases, ARMS and ARMSH yield quite similar results. The results in figures A3 point out that the performance of both GETS and Lasso rely heavily on homoskedasticity.

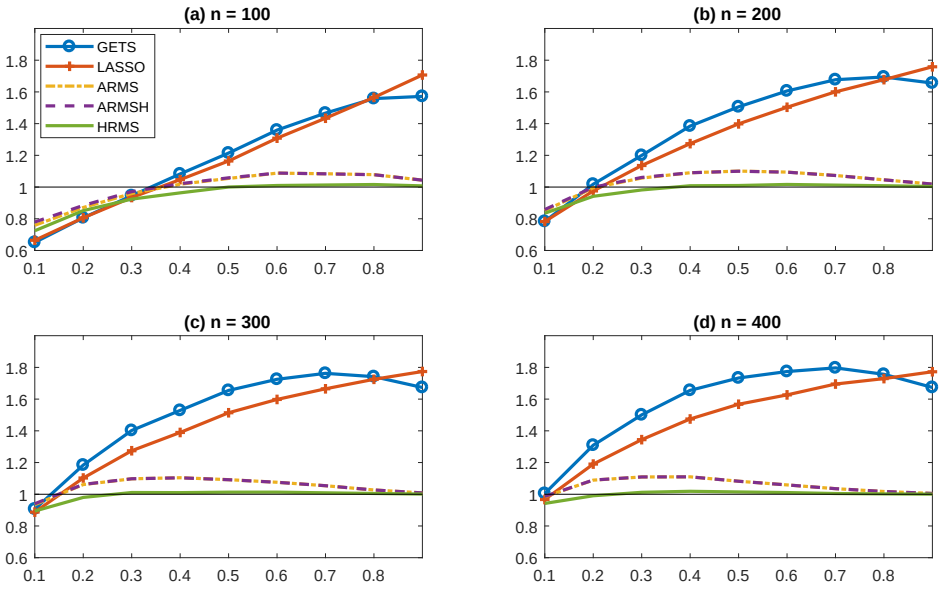
[25] A full permutation of the $K = 20$ regressors leads to a total of 1,048,575 candidate models (the null model is ignored). In our experiments, the pre-determined parameters for GETS and ARMS(H) are $p = 0.1$ and $M' = 20$ respectively, whereas we manipulate the tuning parameter for Lasso and select 5 predictors. We construct $2^5 - 1 = 31$ models based on permutation of the selected parameters.

Figure A3: Comparing Model Screening Methods with Simulated Data

Scenario A. Random Heteroskedasticity



Scenario B. Parameter Heterogeneity



F.4.1 Prediction Comparison Using One Set of Measures

Table A9: Evaluating the Importance of Twitter Variable using MAB

n_E	Include Both	Sentiment Only	Volume Only	Include None	Benchmark
Panel A: Open Box Office					
		Mean Squared Forecast Error (MSFE)			
10	**0.5066**	0.8659	0.6009	1.5271	1.0000
20	**0.7315**	0.9111	0.8242	1.6091	1.0000
30	**0.7531**	0.9463	0.9654	1.8287	1.0000
40	**0.9145**	0.9934	1.0810	2.1822	1.0000
		Mean Absolute Forecast Error (MAFE)			
10	**0.6232**	0.7505	0.6635	0.9881	1.0000
20	**0.6955**	0.8531	0.7428	1.0911	1.0000
30	**0.7042**	0.9057	0.7940	1.2939	1.0000
40	**0.7625**	0.9653	0.8151	1.3988	1.0000
Panel B: Movie Unit Sales					
		Mean Squared Forecast Error (MSFE)			
10	**0.7307**	0.9235	0.8683	1.4882	1.0000
20	**0.7009**	0.9621	0.9038	1.6761	1.0000
30	**0.7494**	0.9796	0.9325	1.7988	1.0000
40	**0.8626**	0.9744	0.9757	1.9982	1.0000
		Mean Absolute Forecast Error (MAFE)			
10	**0.7461**	0.8397	0.7970	1.0981	1.0000
20	**0.7564**	0.8861	0.8287	1.1532	1.0000
30	**0.7954**	0.9052	0.8525	1.2887	1.0000
40	**0.8211**	0.9311	0.8617	1.4109	1.0000

Note: Bold numbers denote the strategy with the best performance in that row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the HRC^p method presented in the last column.

F.5 Weights of, and Contents of the Top 5 Models Selected by the HRC^p Estimator

Table A10: Describing the 5 Highest Weight Models: Open Box Office

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**HRC^p**
Weight in HRC^p	0.3862	0.2159	0.1755	0.0945	0.0816	
Genre						
Action				x		x
Adventure	x		x		x	x
Animation						x
Biography						x
Comedy				x		x
Crime	x					x
Drama				x		x
Family						x
Fantasy	x	x	x	x	x	x
Horror	x	x			x	x
Mystery		x	x		x	x
Romance		x	x			x
Sci-Fi						x
Thriller						x
Rating						
PG						x
PG13						x
R						x
Core						
Budget	x	x	x	x	x	x
Weeks	x	x	x	x	x	x
Screens	x	x	x	x	x	x
Sentiment						
T-21/-27						x
T-14/-20	x			x	x	x
T-7/-13		x	x			x
T-4/-6						x
T-1/-3	x	x	x	x	x	x
Volume						
T-21/-27	x	x	x	x	x	x
T-14/-20						x
T-7/-13	x	x	x	x	x	x
T-4/-6	x	x	x	x	x	x
T-1/-3	x	x	x	x	x	x
R^2 w/ SV.	0.8265	0.8249	0.8258	0.8248	0.8259	0.8230
R^2 w/o SV.	0.4836	0.4796	0.4789	0.4911	0.4795	0.7383

Note: x denotes that explanatory variable is included in the particular model, SV denotes social media data and HRC^p refers to a specific model averaging method.

Table A11: Describing the 5 Highest Weight Models: Retail Movie Unit Sales

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**HRC^p**
Weight in HRC^p	0.2977	0.1645	0.1558	0.1447	0.0989	
Genre						
Action						x
Adventure						x
Animation	x	x	x	x	x	x
Biography						x
Comedy			x			x
Crime						x
Drama						x
Family	x	x	x	x	x	x
Fantasy	x	x	x	x	x	x
Horror	x	x	x	x	x	x
Mystery	x	x			x	x
Romance						x
Sci-Fi						x
Thriller		x				x
Rating						
PG	x	x	x	x		x
PG13	x	x	x	x	x	x
R	x	x	x	x		x
Core						
Budget	x	x	x	x	x	x
Weeks	x	x	x	x	x	x
Screens	x	x	x	x	x	x
Sentiment						
T-21/-27						x
T-14/-20			x			x
T-7/-13	x	x	x	x	x	x
T-4/-6	x	x	x	x	x	x
T-1/-3					x	x
T+0	x	x	x	x	x	x
T+1/+7	x	x	x	x		x
T+8/+14		x				x
T+15/+21						x
T+22/+28					x	x
Volume						
T-21/-27	x	x	x	x	x	x
T-14/-20						x
T-7/-13	x	x	x	x	x	x
T-4/-6	x	x	x	x	x	x
T-1/-3	x	x	x	x	x	x
T+0	x	x	x	x	x	x
T+1/+7						x
T+8/+14		x				x
T+15/+21	x		x	x		x
T+22/+28						x
R^2 w/ SV.	0.8512	0.8517	0.8530	0.8503	0.8362	0.8450
R^2 w/o SV.	0.5976	0.6024	0.6027	0.5976	0.5918	0.7002

Note: x denotes that explanatory variable is included in the particular model, SV denotes social media data and HRC^p refers to a specific model averaging method.

F.6 Further Comparison of the Relative Prediction Efficiency for Lasso-based Strategies

Table A12: Further Comparison of the Relative Prediction Efficiency (with Both Sentiment and Volume)

n_E	OLS_{10}	OLS_{11}	OLS_{12}	OLS_{13}	OLS_{14}	OLS_{15}	HRC^p_{10}	HRC^p_{11}	HRC^p_{12}	HRC^p_{13}	HRC^p_{14}	HRC^p_{15}	HRC^p
Panel A: Open Box Office													
							Mean Squared Forecast Error (MSFE)						
10	1.1464	1.1704	1.1671	1.1778	1.1132	1.1221	1.1462	1.1642	1.1647	1.1717	1.1094	1.1203	**1.0000**
20	1.1620	1.1809	1.1803	1.1830	1.0943	1.0992	1.1606	1.1771	1.1797	1.1755	1.0815	1.0826	**1.0000**
30	1.1922	1.2092	1.2067	1.2113	1.0731	1.0696	1.1899	1.2068	1.2037	1.2092	1.0636	1.0624	**1.0000**
40	1.2076	1.2295	1.2174	1.2233	1.0608	1.0633	1.2027	1.2197	1.2141	1.2199	1.0573	1.0537	**1.0000**
							Mean Absolute Forecast Error (MAFE)						
10	1.0529	1.0591	1.0669	1.0689	1.0623	1.0632	1.0430	1.0595	1.0576	1.0687	1.0593	1.0594	**1.0000**
20	1.0603	1.0657	1.0692	1.0767	1.0556	1.0549	1.0506	1.0631	1.0689	1.0750	1.0551	1.0546	**1.0000**
30	1.0568	1.0619	1.0669	1.0722	1.0560	1.0558	1.0473	1.0528	1.0576	1.0719	1.0542	1.0538	**1.0000**
40	1.0591	1.0663	1.0673	1.0734	1.0549	1.0537	1.0578	1.0654	1.0641	1.0720	1.0536	1.0530	**1.0000**
Panel B: Movie Unit Sales													
							Mean Squared Forecast Error (MSFE)						
10	1.3737	1.2921	1.3495	1.3456	1.3621	1.3757	1.3558	1.2784	1.3354	1.3434	1.3512	1.3704	**1.0000**
20	1.3756	1.2772	1.2811	1.2457	1.2578	1.2768	1.3448	1.2459	1.2697	1.2432	1.2568	1.2651	**1.0000**
30	1.3001	1.2388	1.2086	1.1616	1.1666	1.1814	1.2728	1.2282	1.2012	1.1530	1.1644	1.1822	**1.0000**
40	1.2306	1.1718	1.1609	1.1135	1.1364	1.1454	1.2069	1.1565	1.1486	1.1093	1.1281	1.1398	**1.0000**
							Mean Absolute Forecast Error (MAFE)						
10	1.2303	1.2058	1.2161	1.1581	1.1534	1.1600	1.2229	1.1974	1.2155	1.1575	1.1523	1.1564	**1.0000**
20	1.2096	1.1844	1.1958	1.1386	1.1427	1.1436	1.2036	1.1760	1.1890	1.1369	1.1411	1.1398	**1.0000**
30	1.1887	1.1656	1.1735	1.1182	1.1204	1.1195	1.1794	1.1569	1.1675	1.1161	1.1180	1.1149	**1.0000**
40	1.1704	1.1469	1.1557	1.0989	1.1064	1.1086	1.1600	1.1364	1.1459	1.0959	1.1005	1.1027	**1.0000**

Note: Bold numbers denote the strategy with the best performance in that row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the HRC^p method presented in the last column. OLS_q and HRC^p_q stand for OLS and HRC^p with q number of covariates selected by the Lasso.

Table A13: Further Comparison of the Relative Prediction Efficiency (with Sentiment Only)

n_E	OLS_{10}	OLS_{11}	OLS_{12}	OLS_{13}	OLS_{14}	OLS_{15}	HRC^p_{10}	HRC^p_{11}	HRC^p_{12}	HRC^p_{13}	HRC^p_{14}	HRC^p_{15}	HRC^p
Panel A: Open Box Office													
							Mean Squared Forecast Error (MSFE)						
10	1.1111	1.1240	1.1428	1.1403	1.1389	1.1528	1.0865	1.0922	1.1077	1.1022	1.1068	1.1084	**1.0000**
20	1.0836	1.0940	1.1102	1.1121	1.0887	1.0896	1.0802	1.0766	1.0912	1.1010	1.0795	1.0842	**1.0000**
30	1.0648	1.0700	1.0888	1.0871	1.0799	1.0840	1.0641	1.0643	1.0787	1.0809	1.0702	1.0772	**1.0000**
40	1.0732	1.0779	1.1027	1.1099	1.0902	1.0909	1.0727	1.0768	1.0939	1.0916	1.0777	1.0795	**1.0000**
							Mean Absolute Forecast Error (MAFE)						
10	1.0305	1.0422	1.0485	1.0528	1.0552	1.0652	1.0302	1.0325	1.0368	1.0281	1.0318	1.0457	**1.0000**
20	1.0314	1.0399	1.0467	1.0535	1.0556	1.0647	1.0276	1.0311	1.0323	1.0369	1.0413	1.0456	**1.0000**
30	1.0303	1.0378	1.0474	1.0522	1.0542	1.0669	1.0256	1.0298	1.0318	1.0364	1.0382	1.0421	**1.0000**
40	1.0355	1.0468	1.0542	1.0615	1.0592	1.0719	1.0281	1.0343	1.0398	1.0402	1.0411	1.0475	**1.0000**
Panel B: Movie Unit Sales													
							Mean Squared Forecast Error (MSFE)						
10	1.0179	1.0323	1.0391	1.0299	1.0494	1.0257	1.0152	1.0037	1.0030	1.0192	1.0151	1.0068	**1.0000**
20	1.0462	1.0589	1.0635	1.0528	1.0639	1.0362	1.0388	1.0515	1.0557	1.0429	1.0509	1.0303	**1.0000**
30	1.0308	1.0406	1.0501	1.0376	1.0445	1.0199	1.0273	1.0296	1.0342	1.0338	1.0328	1.0168	**1.0000**
40	1.0111	1.0214	1.0307	1.0291	1.0309	1.0094	1.0019	1.0204	1.0263	1.0227	1.0233	1.0033	**1.0000**
							Mean Absolute Forecast Error (MAFE)						
10	1.0180	1.0219	1.0216	1.0212	1.0394	1.0330	1.0063	1.0101	1.0073	1.0088	1.0192	1.0132	**1.0000**
20	1.0044	1.0132	1.0162	1.0194	1.0366	1.0242	1.0072	1.0049	1.0062	1.0056	1.0166	1.0115	**1.0000**
30	1.0013	1.0100	1.0145	1.0195	1.0327	1.0253	1.0010	1.0072	1.0014	1.0019	1.0148	1.0122	**1.0000**
40	1.0081	1.0042	1.0089	1.0149	1.0300	1.0214	1.0028	1.0032	1.0013	1.0099	1.0052	1.0023	**1.0000**

Note: Bold numbers denote the strategy with the best performance in that row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the HRC^p method presented in the last column. OLS_q and HRC^p_q stand for OLS and HRC^p with q number of covariates selected by the Lasso.

Table A14: Further Comparison of the Relative Prediction Efficiency (with Volume Only)

n_E	OLS_{10}	OLS_{11}	OLS_{12}	OLS_{13}	OLS_{14}	OLS_{15}	HRC^p_{10}	HRC^p_{11}	HRC^p_{12}	HRC^p_{13}	HRC^p_{14}	HRC^p_{15}	HRC^p
Panel A: Open Box Office													
							Mean Squared Forecast Error (MSFE)						
10	1.0614	1.0391	1.0312	1.0315	1.0309	1.0380	1.0551	1.0351	1.0297	1.0224	1.0255	1.0362	**1.0000**
20	1.0817	1.0181	1.0074	1.0122	1.0069	1.0137	1.0791	1.0102	0.9984	1.0108	1.0041	1.0121	**1.0000**
30	1.1556	1.0217	1.0176	1.0200	1.0207	1.0263	1.1517	1.0159	1.0107	1.0131	1.0117	1.0205	**1.0000**
40	1.1705	1.0267	1.0179	1.0198	1.0170	1.0271	1.1689	1.0227	1.0104	1.0164	1.0172	1.0199	**1.0000**
							Mean Absolute Forecast Error (MAFE)						
10	1.0058	1.0113	1.0109	1.0113	1.0116	1.0115	1.0012	1.0037	1.0067	1.0119	1.0036	1.0018	**1.0000**
20	1.0228	1.0150	1.0160	1.0131	1.0117	1.0163	1.0148	1.0120	1.0069	1.0078	1.0045	1.0137	**1.0000**
30	1.0343	1.0122	1.0147	1.0149	1.0172	1.0212	1.0249	1.0075	1.0091	1.0125	1.0159	1.0158	**1.0000**
40	1.0280	1.0169	1.0194	1.0203	1.0213	1.0247	1.0264	1.0084	1.0104	1.0186	1.0166	1.0196	**1.0000**
Panel B: Movie Unit Sales													
							Mean Squared Forecast Error (MSFE)						
10	1.2868	1.2680	1.2518	1.1204	1.0814	1.0996	1.2614	1.2570	1.2493	1.1113	1.0772	1.0969	**1.0000**
20	1.2641	1.2501	1.2383	1.1429	1.0971	1.0951	1.2537	1.2472	1.2332	1.1340	1.0883	1.0879	**1.0000**
30	1.1739	1.1650	1.1541	1.0774	1.0604	1.0389	1.1700	1.1549	1.1439	1.0704	1.0522	1.0304	**1.0000**
40	1.1208	1.1178	1.1126	1.0543	1.0504	1.0125	1.1103	1.1082	1.1092	1.0474	1.0408	1.0093	**1.0000**
							Mean Absolute Forecast Error (MAFE)						
10	1.1268	1.1229	1.1274	1.0750	1.0752	1.0715	1.1236	1.1128	1.1177	1.0668	1.0728	1.0724	**1.0000**
20	1.1125	1.1080	1.1138	1.0688	1.0631	1.0547	1.1096	1.0970	1.1043	1.0610	1.0632	1.0461	**1.0000**
30	1.0874	1.0886	1.0918	1.0492	1.0479	1.0439	1.0803	1.0828	1.0820	1.0490	1.0455	1.0364	**1.0000**
40	1.0784	1.0833	1.0877	1.0487	1.0474	1.0425	1.0768	1.0803	1.0786	1.0463	1.0434	1.0367	**1.0000**

Note: Bold numbers denote the strategy with the best performance in that row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the HRC^p method presented in the last column. OLS_q and HRC^p_q stand for OLS and HRC^p with q number of covariates selected by the Lasso.

F.7 Comparing ARMS and ARMSH

From the exercises in the main text, we notice that ARMS and ARMSH provide similar results in many cases. Although ARMSH is hetero-robust, ARMS and ARMSH end up with similar candidate model sets. In the following table A15, we show the 5 highest weight models estimated by HRC^p using candidate model sets screened by ARMS and ARMSH respectively. For each model screening method, an "x" denotes the associated explanatory variable is included in the particular model. Each model screening method contains a candidate model set of 100 selected models. Estimated model weights are presented in the last row for each method.

Table A15: Describing the 5 Highest Weight Models Using Model Sets Screened by ARMS and ARMSH

	ARMS					ARMSH				
	M1	**M2**	**M3**	**M4**	**M5′**	**M1**	**M2**	**M3**	**M4**	**M5′**
Genre										
Action										
Adventure	x			x	x	x			x	x
Animation										
Biography										
Comedy										
Crime	x					x				
Drama										
Family										
Fantasy	x	x	x	x	x	x	x	x	x	x
Horror	x			x	x	x			x	x
Mystery		x	x	x			x	x		x
Romance		x					x			
Sci-Fi										
Thriller										
Rating										
PG										
PG13										
R										
Core										
Budget	x	x	x	x	x	x	x	x	x	x
Weeks	x	x	x	x	x	x	x	x	x	x
Screens	x	x	x	x	x	x	x	x	x	x
Sentiment										
T-21/-27										
T-14/-20	x		x	x	x	x		x	x	x
T-7/-13		x					x			
T-4/-6										
T-1/-3	x	x	x	x	x	x	x	x	x	x
Volume										
T-21/-27	x	x	x	x	x	x	x	x	x	x
T-14/-20										
T-7/-13	x	x	x	x	x	x	x	x	x	x
T-4/-6	x	x	x	x	x	x	x	x	x	x
T-1/-3	x	x	x	x	x	x	x	x	x	x
Weights	0.4278	0.3914	0.1296	0.0332	0.0155	0.4283	0.4220	0.1038	0.0291	0.0168

Note: x denotes that explanatory variable is included in the particular model. The above exercise is carried out by using the top 100 models screened by ARMS and ARMSH respectively for open box office.

The top 5 models for each method accumulates more than 95% of the total weights. Moreover, we notice that the top 5 models for each method are identical with the same

ranking. This explains why in our prediction experiment, ARMS and ARMSH yield quite similar results in terms of forecast accuracy. In Subsection F.4, we conduct a Monte Carlo study to shed further light on the relative performance of ARMS and ARMSH under different scenarios related to what is the source of heteroskedasticity.

F.8 Performance of Recursive Partitioning Methods Using Identical Variables to Model Screening/Averaging Strategies

In the empirical exercises, we restrict that each potential model contains a constant term and 7 (11) relatively significant parameters for open box office (movie unit sales) based on the OLS results presented in table A5. To examine if our findings are driven by pre-selection, we compare the performance of recursive partitioning methods to econometric strategies using identical set of selected 7 (11) parameters. Results are presented in table A16.

As usual, we report the median MSFE and MAE of different strategies listed in panel A of table A16 for each evaluation set of different sizes $n_E = 10, 20, 30, 40$. Panel A presents results for forecasting open box office and panel B demonstrates results for forecasting movie unit sales. To ease interpretation, in each row of table A16 we normalize the MSFEs and MAFEs, respectively, by the MSFE and MAFE of the HRC^p.

For both panels, table A16 demonstrates that there are very large gains in prediction efficiency of the recursive partitioning algorithms relative to the benchmark HRC^p, although such gains are not as large as those demonstrated in table 5, in which the recursive partitioning methods use all the potential variables available. Take the MSFE results under $n_E = 10$ in panel A for example, Reg.Tree shows approximately 37% increase in prediction efficiency in table 5 and 20% increase in table A16. The results indicate that the pre-selected 7 (11) variables play crucial roles in predicting the open box office (movie unit sales). On the other hand, the other potential variables also jointly provide significant predicting power. In summary, the gains from machine learning strategies that use recursive partitioning over econometric methods is not due to differences in the set of predictors.

F.9 Test for Superior Predictive Ability (SPA) of the MASVR$_{\text{LS}}$ Method

In this subsection, we perform the SPA test of Hansen (2005) to examine if the MASVR$_{\text{LS}}$ method we proposed demonstrates superior predictive ability over all the other methods listed in this paper. We consider both the squared forecast error (SFE) and the absolute forecast error (AFE) as the quantities for comparing predictive ability. We set the results of MASVR$_{\text{LS}}$ as the benchmark.

The null hypothesis of the SPA test states that the average performance of the benchmark is as good as the best average performance across the other competing methods. The

Table A16: Results of Relative Prediction Efficiency between Recursive Partitioning Methods Using Selective Variables and the Benchmark Method

n_E	Reg. Tree	Bagging	Random Forest			Benchmark
			RF_{10}	RF_{15}	RF_{20}	
Panel A: Open Box Office						
Mean Squared Forecast Error (MSFE)						
10	**0.8020**	0.9501	0.8155	0.8542	0.9559	1.0000
20	1.0149	0.9287	0.8560	**0.8540**	0.8940	1.0000
30	1.1125	0.8611	0.8679	**0.8525**	0.9940	1.0000
40	1.3306	1.1571	1.2549	1.1343	1.2340	**1.0000**
Mean Absolute Forecast Error (MAFE)						
10	**0.7794**	0.8487	0.7865	0.7973	0.8641	1.0000
20	0.8079	0.7635	0.7571	**0.7359**	0.7507	1.0000
30	0.8780	**0.8487**	0.8536	0.8670	0.8909	1.0000
40	**0.8501**	0.8539	0.8649	0.8837	0.8914	1.0000
Panel B: Movie Unit Sales						
Mean Squared Forecast Error (MSFE)						
10	0.9236	0.9580	**0.9009**	0.9151	0.9571	1.0000
20	1.0261	0.9600	0.9439	**0.9053**	0.9557	1.0000
30	1.2982	**0.9810**	1.0447	1.0652	1.1236	1.0000
40	1.1213	1.0037	0.9886	**0.9761**	0.9834	1.0000
Mean Absolute Forecast Error (MAFE)						
10	**0.8390**	0.9794	0.9201	0.9525	0.9443	1.0000
20	0.8409	**0.8303**	0.8448	0.8388	0.8563	1.0000
30	0.9485	**0.9103**	0.9431	0.9220	0.9250	1.0000
40	0.8905	0.8367	**0.8332**	0.8456	0.8398	1.0000

Note: Bold numbers denote the strategy with the best performance in that row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the HRC^p method presented in the last column. The subscript in RF_q stands for the number of covariates randomly chosen at each node to consider as the potential split variable. All bagging and random forest estimates involve 100 trees.

alternative is that there is at least one competing method has better average performance than the benchmark. We estimate the p-values under the two forecast error quantities for open box office and movie unit sales. Large p-values signify the superior predictive ability of the $MASVR_{LS}$ method over others.

Results for different n_E values are presented in table A17 and all the p-values are larger than 5% implying the superior predictive ability of the $MASVR_{LS}$ method over others. This is particularly true for the AFE case in which the p-values are as high as 1 in all cases. The p-values for open box office under SFE are relatively smaller than other cases which coincides with the MSFE results demonstrated in table 5.

F.10 Model Averaging Regression Tree

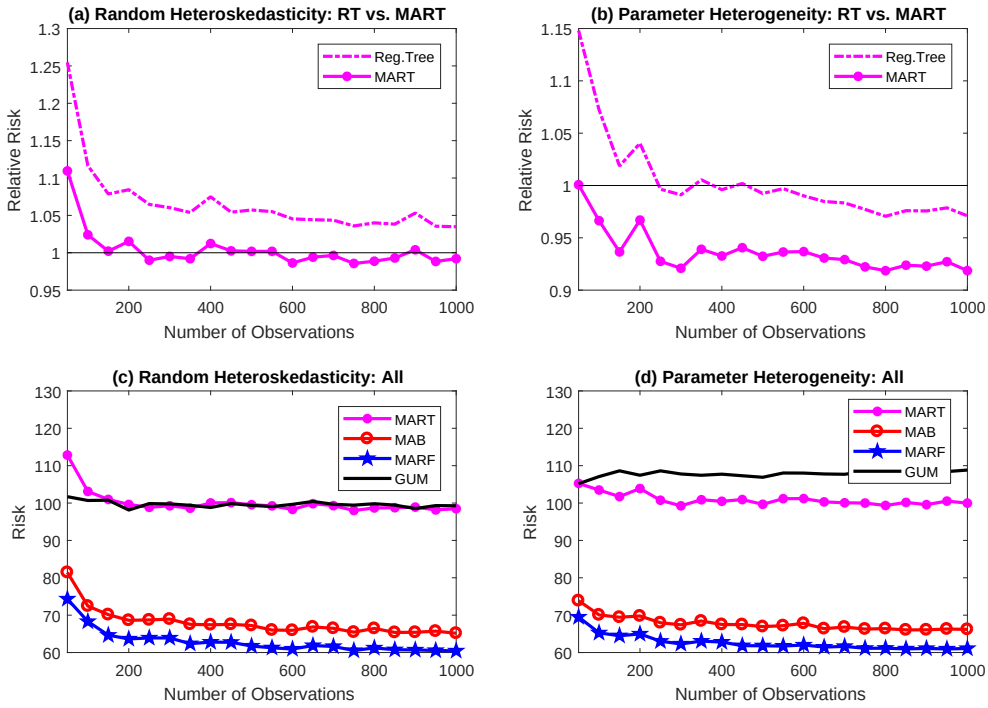
This subsection considers adding a model averaging flavor to a single regression tree (MART). We duplicate the Monte Carlo simulations in section 3 and the MART method is

Table A17: SPA Test Results of the MASVR_{LS} Method

n_E	Open Box Office		Movie Unit Sales	
	SFE	AFE	SFE	AFE
10	0.4424	1.0000	1.0000	1.0000
20	0.1772	1.0000	0.2975	1.0000
30	0.2145	1.0000	1.0000	1.0000
40	0.0987	1.0000	0.1266	1.0000

represented by the lines with dots in figure A4. Although MART dominates RT for both heteroskedasticity scenarios in figures A4(a) and A4(b), it is clear in figures A4(c) and A4(d) that the MART method is outperformed by both MAB and MARF by a large margin in both scenarios. In fact under random heteroskedasticity MART performs similarly to OLS estimation of GUM. This reinforces our claim that gains to adding model averaging to recursively partitioned subgroups occurs when there is systemic heterogeneity perhaps due to parameter heterogeneity. The MART method only outperforms GUM under parameter heterogeneity.

Figure A4: Risk Comparison under Different Scenarios



F.11 Centered R^2 on the Training Set

Table A18: Centered R^2 Description on 10,001 Training Sets

n_E	Method	Mean	Median	2.5% Q	97.5% Q
Panel A: Open Box Office					
10	GUM	0.8666	0.8718	0.7721	0.9329
	MTV	0.5862	0.5852	0.4931	0.6871
20	GUM	0.8685	0.8743	0.7704	0.9355
	MTV	0.5891	0.5869	0.4946	0.6964
30	GUM	0.8715	0.8773	0.7740	0.9390
	MTV	0.5936	0.5918	0.4951	0.7011
40	GUM	0.8745	0.8811	0.7719	0.9416
	MTV	0.5976	0.5958	0.4948	0.7100
Panel B: Movie Unit Sales					
10	GUM	0.9037	0.9071	0.8258	0.9596
	MTV	0.7251	0.7248	0.6309	0.8188
20	GUM	0.9087	0.9127	0.8305	0.9628
	MTV	0.7338	0.7335	0.6370	0.8305
30	GUM	0.9137	0.9184	0.8322	0.9679
	MTV	0.7423	0.7416	0.6447	0.8407
40	GUM	0.9205	0.9255	0.8396	0.9729
	MTV	0.7523	0.7516	0.6497	0.8547

F.12 Selected Parameters by OLS-post-Lasso

Table A19: Describing the Selected Parameters by OLS-post-Lasso

Method	1^{st} Quartile		2^{nd} Quartile		3^{rd} Quartile		4^{th} Quartile		Full Sample	
	Sentiment	Volume	Sentiment	Volume	Sentiment	Volume	Sentiment	Volume	Sentiment	Volume
Panel A: Open Box Office										
OLS_{10}	5	1	3	2	4	1	5	2	6	2
OLS_{12}	6	1	5	2	4	1	6	2	7	2
OLS_{15}	7	2	6	2	5	1	6	2	8	3
Panel B: Movie Unit Sales										
OLS_{10}	7	1	8	1	8	2	6	2	7	2
OLS_{12}	9	1	9	1	9	2	8	2	8	2
OLS_{15}	11	1	10	1	10	2	11	2	10	2

Note: Each entry in the table lists the number of respective social media variables chosen as one of the first 10 predictors among all variables in different budget subsamples 10, 12, or 15.

F.13 Revisiting the Results of Lehrer and Xie (2017) with Popular Machine Learning Estimators

This section illustrates that random forest and bagging yield more accurate forecasts on the Lehrer and Xie (2017) subsample of films with budgets between 20 to 100 million.

Table A20: Revisit Lehrer and Xie (2017) with Machine Learning Estimators

n_E	GUM	MTV	GETS	AIC	MMA	PMA_{g1}	PMA_{g2}	MMA_{g1}	MMA_{g2}	OLS_{12}	PMA_{12}	BAG	RF	PMA
							Mean Squared Forecast Error (MSFE)							
10	1.5215	2.8983	1.3623	1.2265	1.1503	1.3006	1.3306	1.2753	1.2904	1.2506	1.3366	1.1450	**0.9706**	1.0000
20	1.5180	2.7006	1.3593	1.0712	1.0226	1.2835	1.1846	1.2457	1.1368	1.1111	1.2186	1.0108	**0.9947**	1.0000
30	1.5958	2.4315	1.4160	1.0322	1.0024	1.2454	1.1127	1.1764	1.0341	0.9963	1.0367	**0.8725**	0.9908	1.0000
40	2.1040	2.6592	1.7339	1.0389	1.0005	1.6176	1.2492	1.4283	1.0930	0.9852	0.9889	**0.8879**	0.9953	1.0000
							Mean Absolute Forecast Error (MAFE)							
10	1.2354	1.7888	1.2387	1.1731	1.1269	1.1666	1.2227	1.1681	1.2198	1.2114	1.2284	1.1225	**0.9734**	1.0000
20	1.1892	1.7204	1.1766	1.0909	1.0490	1.1109	1.1317	1.0995	1.1103	1.1143	1.1623	1.0428	**0.9775**	1.0000
30	1.2104	1.5971	1.1643	1.0113	1.0038	1.0873	1.0701	1.0652	1.0508	1.0210	1.0585	0.9728	**0.9667**	1.0000
40	1.3514	1.5825	1.2692	1.0027	1.0072	1.1945	1.0991	1.1281	1.0408	1.0221	1.0095	0.9736	**0.9695**	1.0000

Note: Bold numbers indicate the strategies with the best performance in that simulation experiment denoted by the row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the PMA method presented in the last column. OLS_q and PMA_q stand for OLS and PMA with q number of covariates selected by the Lasso.

F.14 Expressing Results on Forecast Accuracy in Absolute Values

This subsection presents the results of the forecasting experiment in terms of absolute units of the loss function, as opposed of degree of loss to relative to the chosen baseline estimator. These results are measured in millions of dollars and the general finding is that the statistical learning methods generally reduce the variation relative to econometric approaches. The addition of model averaging in the leaves does not lead to further gains in efficiency but achieves higher MSFE and MAFE by reducing the bias. These results show that statistical learning methods are less variable and the incorporation of model averaging in place of the local constant model achieves gains by increasing the accuracy of forecasts. These results appear consistent with the patterns illustrated in the main text when discussing Figure 1.

F.15 Comparing Computational Efficiency between HBART and MARF

In this section, we compare the performance of HBART and MARF. Note that the two algorithms rely on different splitting rules and tuning parameters. They construct each regression tree in different fashion. The code we use for each strategy is written with different software (see Appendix table A1) so we do not wish to compare their computational efficiency by CPU time. Rather, in this section we consider the following sensitivity test.

Both methods draw random trees at a given number. Denote such number as B. The number of trees in a forest is a parameter that is not tunable in the classical sense. Theoretically, the prediction accuracy of a method increases with B. In practice, however, one may want to set B as small as possible for the sake of computational efficiency. Probst and Boulesteix (2018) theoretically prove with RF that more trees are always better since they reduce the mean squared error. The convergence rate, and thus the number of trees needed to obtain optimal performance, depends on the dataset's properties. Using a large number of datasets, Probst and Boulesteix (2018) show empirically that with RF the biggest performance gain can often be achieved when growing the first 100 trees. Similarly, Hastie, Tibshirani, and Friedman (2009) demonstrate that the benefit from increasing B is minimal, when B is large enough.

In this exercise, we compare the improvement in results when $B = 20$ versus $B = 100$ for HBART and MARF. The computation time increases linearly with the number of trees so we would expect that to build trees with $B = 20$ would be five times faster than $B = 100$. We also anticipate that the prediction accuracy increases for both methods as B increases. Table A22 present the percentage gain in prediction accuracy in terms of MSFE (top panel) and MAFE (bottom panel) as the size of the evaluation set increases. Columns 2 and 3 shows the result for open box office prediction for HBART and MARF and columns 4 and 5 show the corresponding results for predicting retail movie unit sales with HBART and MARF, respectively.

Table A21: Results of Table 5 in Absolute Values

n_E	BART	BART_{BMA}	HBART	BOOST	Reg.Tree	Bagging	Random Forest		MAB	Random Forest		SVR_{LS}	MASVR_{LS}	Benchmark
							RF_{15}	RF_{20}		MARF_{15}	MARF_{20}			
Panel A: Open Box Office														
10	5.5425	7.8076	5.5574	5.2273	7.0491	7.0978	7.1192	7.0508	6.4905	6.3401	6.3265	5.2596	5.0270	9.7105
	(3.6050)	(3.3786)	(3.1614)	(3.6817)	(4.6419)	(3.8684)	(3.8236)	(3.7440)	(3.6941)	(3.4896)	(3.4971)	(2.9922)	(3.2709)	(5.4446)
20	5.8970	7.6891	6.1402	5.7685	7.6393	7.5170	7.4441	7.4303	7.0023	6.7856	6.8190	5.7361	5.4863	9.8558
	(2.6682)	(2.3187)	(2.4083)	(2.6648)	(3.5742)	(2.7851)	(2.7448)	(2.7128)	(2.7797)	(2.5874)	(2.5970)	(2.6052)	(3.1294)	(4.6146)
30	6.0433	7.7769	6.8616	6.2848	7.8910	7.7925	7.7119	7.6620	7.1281	6.9514	6.9596	6.0493	5.7780	10.2581
	(2.0817)	(2.0108)	(2.1307)	(2.4912)	(3.1535)	(2.5738)	(2.5302)	(2.5040)	(2.4546)	(2.3285)	(2.3571)	(2.5469)	(2.6176)	(4.8701)
40	6.5397	8.0275	7.4240	6.7887	8.3260	7.8980	7.8616	7.8095	7.3811	7.2072	7.2334	6.2665	6.1214	10.1545
	(2.0622)	(1.9269)	(1.9735)	(2.3033)	(2.7467)	(2.1213)	(2.1240)	(2.0859)	(2.0733)	(2.0121)	(1.9969)	(2.0981)	(3.0438)	(3.6305)
Panel B: Movie Unit Sales														
10	0.2650	0.3421	0.3184	0.2713	0.3254	0.3182	0.3375	0.3288	0.2900	0.2990	0.2934	0.2500	0.2299	0.3980
	(0.1734)	(0.1393)	(0.1613)	(0.1860)	(0.1929)	(0.1599)	(0.1686)	(0.1647)	(0.1531)	(0.1545)	(0.1529)	(0.1648)	(0.1607)	(0.2330)
20	0.2907	0.3527	0.3428	0.2948	0.3467	0.3255	0.3454	0.3363	0.2983	0.3068	0.3014	0.2663	0.2409	0.4084
	(0.1254)	(0.1051)	(0.1143)	(0.1355)	(0.1443)	(0.1151)	(0.1202)	(0.1182)	(0.1092)	(0.1112)	(0.1075)	(0.1224)	(0.1214)	(0.1868)
30	0.2991	0.3640	0.3711	0.3314	0.3627	0.3406	0.3597	0.3510	0.3157	0.3254	0.3192	0.2918	0.2643	0.4225
	(0.1079)	(0.0920)	(0.1065)	(0.1212)	(0.1238)	(0.1006)	(0.1041)	(0.1015)	(0.0969)	(0.0968)	(0.0960)	(0.1239)	(0.1075)	(0.2002)
40	0.3231	0.3795	0.3911	0.3616	0.3844	0.3570	0.3753	0.3666	0.3305	0.3395	0.3343	0.3188	0.2946	0.4243
	(0.0987)	(0.0902)	(0.0941)	(0.1069)	(0.1164)	(0.0900)	(0.0923)	(0.0918)	(0.0870)	(0.0865)	(0.0857)	(0.1293)	(0.1600)	(0.1561)

Note: These results are also used in the construction of Table 5, that reported the ratio of the entry relative to our benchmark HRC^p. Each entry in this table reflects the absolute bias of the estimator with the standard deviation in parentheses.

Table A22: Prediction Accuracy Improvement by Percentage

	Open Box Office		*Movie Unit Sales*	
n_E	HBART	MARF	HBART	MARF
	MSFE		MSFE	
10	18.2506	3.7511	38.2600	3.9410
20	12.3448	3.6438	37.7286	7.1008
30	10.7863	6.2474	19.9060	5.7431
40	10.2459	7.5833	29.6064	2.7005
	MAFE		MAFE	
10	18.8853	1.3436	15.6816	5.5966
20	10.7998	3.7752	8.4059	2.3489
30	3.6224	2.4628	6.3800	2.6138
40	2.0861	2.8604	5.8877	2.5064

Each entry presents the ratio of forecast accuracy with B=100 relative to B=20 where forecast accuracy is measured by MSFE and MAFE.

The results document that when the size of the evaluation set is small (corresponding to a large training set) that there are large improvements in forecast accuracy with HBART when increasing B from 20 to 100. This implies the necessity of setting B to a relatively large number for HBART in our exercises. On the other hand, we note that improvement for MARF by setting B to 100 is relatively small comparing to $B = 20$.

There is a trade-off between accuracy and CPU time for both methods. An interpretation of the results is that MARF (or RF) do not require as large a value of B to obtain accurate forecasts and is therefore more computationally efficient. This finding may be important for practitioners if computational speed is an issue. Last, the results indicate when the training set is small, there are smaller gains from increasing B with HBART when forecasting open box office, which is not surprising since there is both substantial variation in this outcome measure and a small sample size for the training set.

F.16 Using MASVR$_{\text{LS}}$ to Understand How Social Media Should be Accounted For

In this section, we apply MASVR$_{\text{LS}}$ to forecast film outcomes with data sets the contain different combinations of the Twitter data. First, analogous to the MTV estimates in the main text we consider conducting forecasts without sentiment and volume data. We next repeat exercise presented in subsection F.3 by conducting forecasts (ii) without Twitter sentiment data; (iii) without Twitter volume data; and (iv) using both Twitter measures. Results relative to the benchmark model are presented in Table A23.

The results show that if one incorporates volume alone they gain more accurate forecasts that using sentiment measures with MASVR$_{\text{LS}}$. In other words, forecasts calculated without sentiment yield higher prediction accuracy than MASVR$_{\text{LS}}$ without volume. MASVR$_{\text{LS}}$ without Twitter data has the worst overall performance and it performs only marginally worse than MASVR$_{\text{LS}}$ without volume data. What is striking, is how

Table A23: Results of Relative Prediction Efficiency for MASVR_{LS} by Different Sets of Social Media Explanatory Variables

n_E	Include None	Volume Only	Sentiment Only	Include Both	Benchmark
Panel A: Open Box Office					
Mean Squared Forecast Error (MSFE)					
10	0.6173	0.4930	0.5181	0.4128	1.0000
20	0.7083	0.5346	0.7058	0.4798	1.0000
30	0.9753	0.5897	0.9654	0.4954	1.0000
40	1.2911	0.7098	1.0355	0.5538	1.0000
Mean Absolute Forecast Error (MAFE)					
10	0.6319	0.5348	0.6000	0.5002	1.0000
20	0.7000	0.6379	0.7061	0.5536	1.0000
30	0.9506	0.6454	0.9349	0.5727	1.0000
40	1.0142	0.6776	1.0031	0.6057	1.0000
Panel B: Movie Unit Sales					
Mean Squared Forecast Error (MSFE)					
10	0.7038	0.5883	0.6757	0.5820	1.0000
20	0.7481	0.6818	0.7320	0.5964	1.0000
30	0.9962	0.7536	0.9858	0.6974	1.0000
40	1.2675	0.8259	1.0554	0.7931	1.0000
Mean Absolute Forecast Error (MAFE)					
10	0.6590	0.5795	0.6559	0.5494	1.0000
20	0.6856	0.6201	0.6644	0.6015	1.0000
30	0.9541	0.6743	0.9387	0.6487	1.0000
40	1.1921	0.7458	1.0859	0.6879	1.0000

well MASVR_{LS} without Twitter data compared to OLS estimation of the MTV model presented in the main text. When the evaluation set is small, MASVR_{LS} without Twitter data outperforms all of the econometric estimators considered in the main text. This stresses the strength of MASVR_{LS} to capture nonlinearities even with fairly small datasets. Last, and consistent with the findings in section F.3, MASVR_{LS} that incorporates both Twitter measures has the best performance, further reinforcing the need to use both measures that likely capture different dimensions of consumer demand.

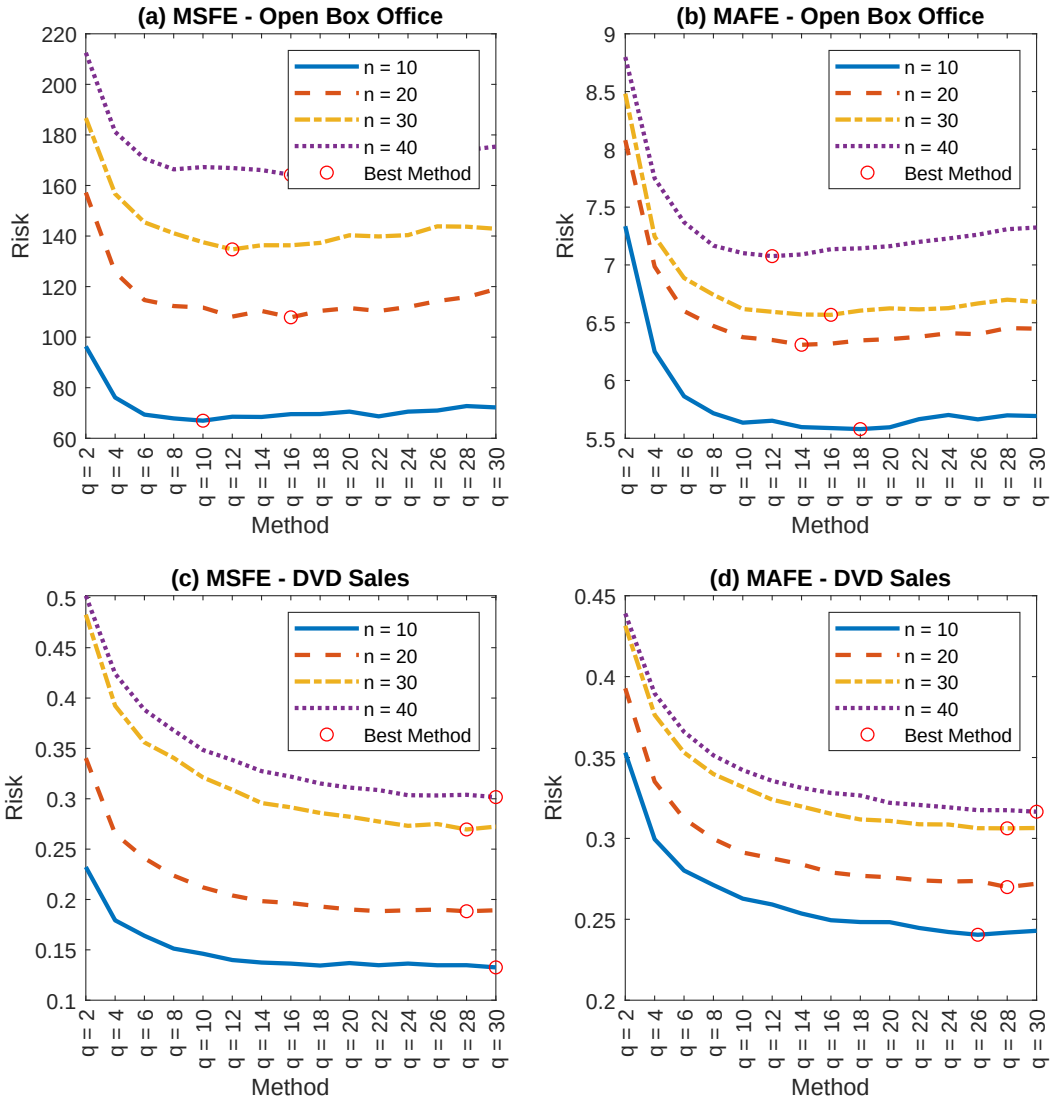
F.17 The Robustness of MARF Results to Different Values of q

In this section, we show the results by MARF with different values of q presented in the x-axis of each subplot in Figure A5. Probst, Boulesteix, and Bischl (2019) provide evidence in an application with 38 datasets that among all the potential hyperparameters to tune with RF, tuning q yields the biggest average improvement in forecasting. From a computational perspective, having a small q, the speed should be fast but there is a chance the splitting direction is far from optimal.

We consider a list of q varies from 2 to 30. The results show that the best performing MARF for open box office are the ones with moderate values of q. For DVD sales, the

best performing MARFS are those with large q. Yet, , there is not large differences in the results with the higher q than the default used for either outcome.

Figure A5: Risk with MARF with Different Values of q



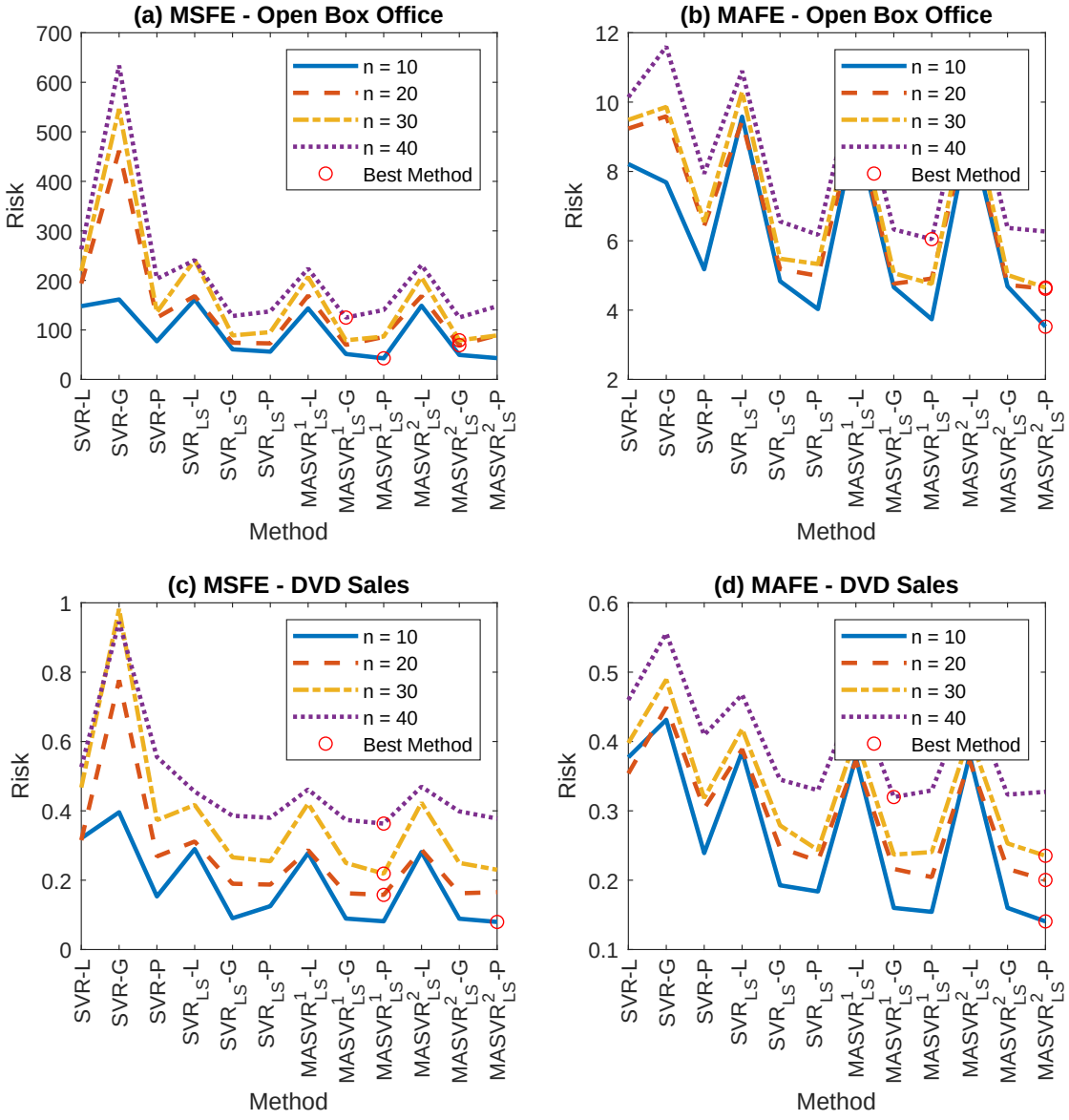
F.18 Examining SVR Methods with Different Kernels

Probst et al. (2018) suggest that there is more benefits from changing hyperparameters of SVM algorithms than random forest. In this section, we compare various SVR methods with different kernels using default hyperparameters. We also compare results by differences in the loss function of SVR versus SVR_{LS} and whether we under MASVR_{LS} with a homoskedastic error using the Mallow's criterion (equation A40) or heteroskedastic error using the HPMA criteria (equation A41).

1. SVR-L: conventional SVR with linear kernel;
2. SVR-G: conventional SVR with Gaussian kernel;
3. SVR-P: conventional SVR with polynomial kernel;
4. SVR_{LS}-L: least squares SVR with linear kernel;
5. SVR_{LS}-G: least squares SVR with Gaussian kernel;
6. SVR_{LS}-P: least squares SVR with polynomial kernel;
7. $\text{MASVR}^1_{\text{LS}}$-L: model averaging SVR_{LS} with linear kernel by PMA;
8. $\text{MASVR}^1_{\text{LS}}$-G: model averaging SVR_{LS} with Gaussian kernel by PMA;
9. $\text{MASVR}^1_{\text{LS}}$-P: model averaging SVR_{LS} with polynomial kernel by PMA;
10. $\text{MASVR}^2_{\text{LS}}$-L: model averaging SVR_{LS} with linear kernel by HPMA;
11. $\text{MASVR}^2_{\text{LS}}$-G: model averaging SVR_{LS} with Gaussian kernel by HPMA;
12. $\text{MASVR}^2_{\text{LS}}$-P: model averaging SVR_{LS} with polynomial kernel by HPMA;

We present the results in figure format in Figure A6. The results show that MASVR_{LS} always yield the best performance irrespective of which kernel was employed. The differences between $\text{MASVR}^1_{\text{LS}}$ and $\text{MASVR}^2_{\text{LS}}$ appear to be marginal. In general, SVR_{LS} has better performance than conventional SVR. Finally, the linear kernel has the worst overall performance among SVR_{LS} and MASVR_{LS} methods.

Figure A6: Comparing Various SVR Methods with Different Kernels



F.19 Further Results Exploring Robustness to Alternative Hyperparameters

Within the social sciences, there is increased attention paid towards developing a pre-analysis plans to deal with concerns ranging from specification search and failure to replicate. The idea of a pre-analysis plan may also appear relevant with machine learning algorithms as it pertains to the selection of both hyperparameters and methods to choose specific tuning parameters. In this paper, we began with the default settings with well-established software. We have conducted a detailed investigation of the sensitivity of our conclusions to alternative hyperparameters for every machine learning algorithm This investigation focused heavily on small changes from the default values to guide subsequent larger changes. Many of these additional exercises are included in the online appendix and due to the length of the current appendix output from the remaining checks can be made available from the authors upon request.

In this section, we consider alternative hyperparameters rather than the default conventional parameter settings for a large number of machine learning methods examined in the main text. Since our main findings stress the performance of MARF, SVR-type methods, and HBART, robustness to different values of the hyperparameters is explored in greater detail in Section F.17, F.18, and F.15, respectively.

We replicate the empirical exercises in the main text and consider the following hyperparameter setting.

1. For BOOST, we grow deeper and shallower trees by setting the number of minimum leaf size at 10 and 20 instead of the default setting 15. We also choose the optimized leaf size between 10 and 20 by 5-fold cross-validation. We denote the methods as BOOST′, BOOST″, and BOOST‴, respectively.

2. For RT, BAG, and RF, we grow shallower trees by setting the number of minimum leaf size at 5 instead of the deep tree default setting 1. We denote these methods as RT′, BAG′, RF′, respectively. .

3. For SVR-P, we set the polynomial order to 2 instead of the default value 3.

4. For SVR_{LS}-G, we fixed $\sigma_x^2 = 10$ instead of the default estimated value by 5-fold cross-validation. We consider two different penalty coefficients $\lambda = 1$ and $\lambda = 10$. The latter is more sensitive to the magnitude of the coefficients. We denote the methods under the two settings as SVR_{LS}-G′ and SVR_{LS}-G″, respectively.

5. Similarly, for SVR_{LS}-P, we also fixed the hyperparameters at $\gamma = 1$ and $d = 2$ instead of the optimized value estimated by 5-fold CV. We consider two different penalty coefficients $\lambda = 1$ and $\lambda = 10$. We denote the methods under the two settings as SVR_{LS}-P′ and SVR_{LS}-P″, respectively.

Results are presented in Table A24. As we can see, the results are similar to those in the main text and the appendix. Comparing to the results in Section F.18, the new SVR_{LS}-P' and SVR_{LS}-P'' are very similar.

It is important to reiterate that the conclusions of any machine learning algorithm horse race could be based on selection of hyperparameters. Our main findings in the text are drawn from pre-determined parameters and generally using cross validation techniques. The robustness exercise has increased our confidence in the main findings as the general ranking across forecasting strategies in terms of accuracy as measured by either MAFE or MSFE remains stable provided the tuning parameters are selected from an appropriate range that surrounds the defaults.

It remains possible that either haphazardly or using a grid search to choose tuning parameters, that we may find certain ranges of hyperparameters where the ordering in terms of forecast accuracy changes. Despite this caeat we did a comprehensive investigation of the sensitivity of our findings to hyperparameter choice with the simulated data used in the Monte Carlo exercises presented in section 3 of the main text. The results continuously found large gains from incorporating model averaging with either a tree based or least squares SVR strategy and that the gains from the hybrid strategy with least squares SVR greatly exceed the hybrid strategy with regression trees as well as HBART, particularly when the sample sizes are less than or equal to 400. Future work is needed to investigate if there are further gains from alternative methods to make SVR_{LS} sparse such as pruning after training and then retraining (i.e. Suykens 2000) or following Hong, Zhang, Ye, Cai, He, and Wang (2019) who suggest using a simplex basis function as the kernel function to obtain sparse SVM_{LS} models.

Table A24: Robustness Check to Alternative Hyperparameters

n_E	BOOST′	BOOST″	BOOST‴	RT′	BAG′	RF′	SVR-P′	SVR_{LS}-G′	SVR_{LS}-G″	SVR_{LS}-P′	SVR_{LS}-P″	Benchmark
Panel A: Open Box Office												
						Mean Squared Forecast Error (MSFE)						
10	0.6323	0.9525	0.5823	0.7690	0.7104	0.6698	0.7253	0.7817	1.4808	0.6652	0.5241	1.0000
20	0.6579	1.0374	0.7012	1.0231	0.8802	0.8595	0.8433	0.8600	1.9102	0.7863	0.6144	1.0000
30	0.7308	1.1681	0.7798	1.1456	0.9542	0.9894	0.9044	0.9702	3.0475	0.7993	0.6278	1.0000
40	0.7623	1.2667	0.8232	1.2719	1.0224	1.0238	0.9357	0.9563	2.6548	0.8176	0.6498	1.0000
						Mean Absolute Forecast Error (MAFE)						
10	0.6476	0.9237	0.6211	0.7958	0.7572	0.7464	0.6832	0.8943	1.3518	0.5731	0.5657	1.0000
20	0.6409	0.9468	0.6882	0.8331	0.7688	0.7809	0.7057	0.8939	1.4069	0.6038	0.5928	1.0000
30	0.6901	1.0146	0.7110	0.8739	0.8117	0.8319	0.7501	0.9578	1.5059	0.6365	0.6224	1.0000
40	0.7311	1.0731	0.8054	0.9206	0.8403	0.8472	0.7951	0.9616	1.4610	0.6800	0.6542	1.0000
Panel B: Movie Unit Sales												
						Mean Squared Forecast Error (MSFE)						
10	0.8721	1.3526	0.9871	1.2687	0.8803	0.9407	0.9087	1.0843	2.4703	0.6899	0.6159	1.0000
20	0.9419	1.4259	1.0767	1.1750	0.9358	1.0260	1.0250	1.1877	3.3292	0.7968	0.7073	1.0000
30	0.9938	1.4447	1.1032	1.2758	0.9869	1.1177	1.1069	1.2481	2.9861	0.8947	0.8208	1.0000
40	1.1597	1.6046	1.5071	1.3760	1.0771	1.1749	1.2823	1.4025	3.1122	1.0110	0.9384	1.0000
						Mean Absolute Forecast Error (MAFE)						
10	0.7259	1.0980	0.8071	0.9525	0.8686	0.9013	0.7658	1.0451	1.7345	0.5670	0.5640	1.0000
20	0.8012	1.1619	0.8932	0.9544	0.8629	0.9210	0.8104	1.0927	1.7978	0.6441	0.6386	1.0000
30	0.8413	1.1961	0.9261	1.0044	0.9046	0.9483	0.8740	1.1083	1.8046	0.6986	0.6945	1.0000
40	0.9199	1.2812	1.0012	1.0488	0.9338	0.9888	0.9517	1.1468	1.7940	0.7636	0.7520	1.0000

F.20 Hybrid Approaches to Explain Instead of to Predict

In Section E.1 , we surveyed the literature that explores relationships which either explain the role of eWOM on film revenue or predict film revenue where eWOM is an explanatory variable. This distinction is important since it is well-known that machine learning algorithms are designed to optimize predictive performance. In this section, we consider whether and how they can also inform studies that seek to understand a marginal effect of eWOM on film revenue. A marginal effect differs in their interpretation from the variable importance metrics reported in Tables 6 and 7 of the main text.[26] In addition, many researchers wish to undertake statistical inference on this estimated marginal effect.

As stressed in section 5.1 of the main text, the relationship between explanatory variables and either movie studio revenue outcome variable that we considered is complex and includes interactions. The improved performance of the algorithms evaluated in table 5 of the main text arises since the influence of the explanatory variables on the prediction surface is not additive but more complex since the features in the prediction model interact with each other. The presence of these interactions explains why more complex algorithms (including tree-based algorithms) tend to perform very well relative to the strategies considered in Tables 3 and 4 of the main text.

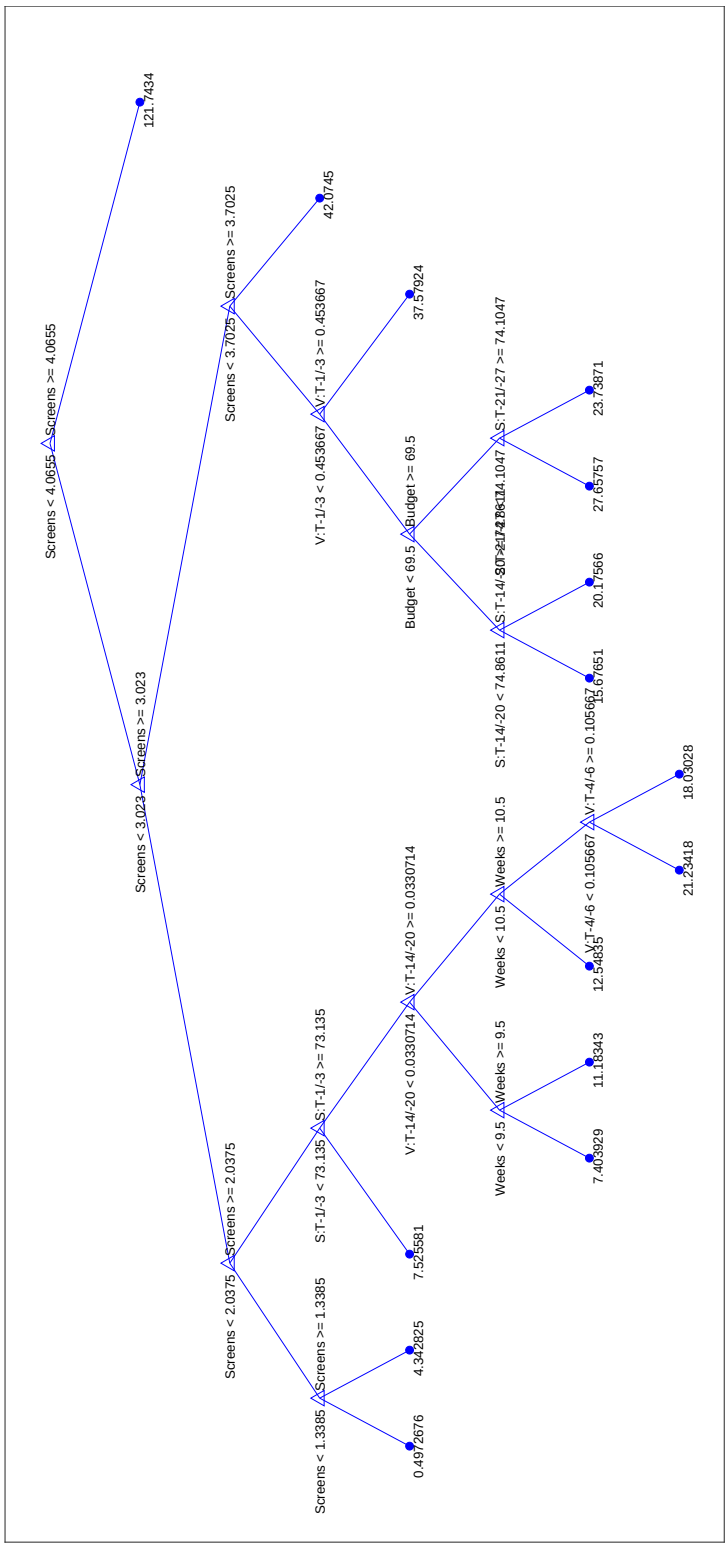
To demonstrate, we present an illustration of the estimated regression tree in Figure A7 for film revenue. To conserve space, the regression tree in Figure A7 is pruned. We can collect information on all the splitting nodes and construct dummy variables accordingly. For example, at the root node, the variable Screen is used to split the data and the cut point is 4.0655. The root node can be represented by the indicator $\mathbb{I}_{\text{Screen}\in(-\infty,4.0655)}$ which is a $n \times 1$ dummy variable such that elements satisfying Screen $<$ 4.0655 equal 1 and 0 otherwise. In other words, the first few splits of the tree show that the relationship between screens and box office opening weekend revenue is explained by a step function, where the step size was chosen by the RT algorithm.

If we examine the tree in its entirety, we observe that there are 14 splitting nodes in Figure A7. Each splitting node contains two directions allowing us to create indicators, for example, Screen $<$ 4.0655 and Screen $\geq$ 4.0655 at the root node, this can be easily achieved by combining the indicator $\mathbb{I}_{\text{Screen}\in(-\infty,4.0655)}$ with the constant term. Therefore, we need one and only one indicator for each splitting node. We can also include interactions between these indicator variables to separate out the common effect of the early split from the differential effect between the two branches at the later split. In other words, the estimates of a RT allow one to consider both indicator variables as well as potentially their interactions to naturally capture complex interactions that can be considered to be incorporated in linear econometric models.

With knowledge of where these nonlinearities are exhibited, we consider two types

[26] We do not consider partial dependence plots (Friedman, 2001) since they obfuscate heterogeneous relationships that result from strong interaction effects.

Figure A7: An Illustration of Pruned Regression Tree



of analyses.[27] First, we revisit the GUM model and rather than enter the screen variable as a continuous linear regressor, we replace it with indicators that capture the estimated nonlinear step function suggested by the earlier nodes of Figure A7. Further, we also estimate a combined GUM model that adds each of the 14 indicators for splits at a node with the original explanatory variables. Note that we replace the indicators based on variable Screen by a step function, of which the components are represented by $\mathbb{I}_{\text{Screen}\in\ (-\infty, 1.3385)}$ to $\mathbb{I}_{\text{Screen}\in[3.7025,4.0655)}$, respectively. The last component $\mathbb{I}_{\text{Screen}\in\ [4.0655,\ +\infty)}$ is left out as the reference term. OLS estimates of models that contain these sets of regressors are presented with the original GUM model in Table A25.

Second, we can also use regularization methods including OLS-post Lasso to determine which interactions to include. Since the algorithm for RT finds the optimal split at each node, the solution depends crucially on each previous step. As such, rather than consider all of the interactions suggested in Figure A7, we consider the potential set to not only solely consider all the indicators denoting a split at each node but all of the potential two by two interactions of these node indicators. With 14 indicators and the constant term, this yields a total of 16,384 different potential combinations. The corresponding $n \times 16384$ raw indicator matrix has rank 96 and after we remove columns that induce multicollinearity, we obtain a final $n \times 96$ matrix of indicators and their products. This matrix is combined with the original explanatory variables used in the GUM model to form the input variable set. We then apply OLS-post-LASSO to reduce the dimensionality of the explanatory variables. This strategy has the advantage of considering several of the two-way interactions proposed in Figure A7. We consider different choices for the penalty term and Table A26 reports estimates from models where the penalties selected 13 and 9 explanatory variables in addition to the constant.

Table A26 contrasts the original OLS estimates (and robust standard errors) of the GUM model in column 3 with a model that replaces screens as a linear regressor with the step function suggested in the initial nodes of Figure A7 in column 2 and a combined GUM model that adds all the single node split indicators in column 1. Since the models are nested, we can conduct specification tests between the restricted GUM model and the models in columns 1 and 2 of Table A25. In each case, the specifications with the nonlinearities suggested by RT is preferred. Accounting for nonlinearities leads each of the film rating indicators as well as the family genre movies to become statistically significant in which the estimated magnitude of is more than triple the size of the GUM model estimate. The results show that the positive effect of screens is driven by films that are slated to open at over 4065.5 locations.

Turning to the effects of social media variables, we find that the effects of volume in both T-1/-3 and T-4/-6 is highly nonlinear as and the effect of T-1/-3 fall by over 50% in size once this nonlinearity is accounted for. Interestingly, none of the linear sentiment variables remain statistically significant once we account for nonlinearities in column 1. Taken together, these results point to the importance of there being thresholds in both

[27] The approach we illustrate can also be used with other tree based algorithms including M5′.

social media measures beyond which there are significant gains in box office opening revenue. This suggests the effect of eWOM in the week prior to opening on immediate box office revenue is highly non-linear. This is a finding that prior research did not report since measures obtained from social media are often restricted to have a linear effect and this could explain the large differences in the variable importance findings we observed with MAB and MASVRLS in table 6 relative to Lasso results presented in Appendix Table F12.

The columns of Table A26 present OLS post Lasso estimates and standard errors from models that can choose only 13 and 9 explanatory variables from an expanded set including the original GUM model as well as set of indicators and their two by two interactions. This table reinforces the importance of nonlinearities since 8 of the 13 variables selected of model 1 and 5 of the 9 variables selected by model 2, are from the splits suggested in Figure A7. Further, similar to the results in table A25, we see the importance of including social media measures collected in the week prior to opening on immediate earnings and that nonlinearities in these measures are present. The results also suggest that family films have a large statistically significant effect on opening weekend box office, an effect that was not apparent with the linear GUM model. In addition, the role of screens and ratings are not statistically significant. Without theory, the aid of RT in Figure A7 is needed to identify how to model these nonlinearities. While only one 2×2 interaction is kept by the Lasso in model 1, its effect is not statistically significant. This result suggests that the nonlinearities exhibited in the underlying data may not be complex in this application.

The above strategies consider the use of machine learning to suggest appropriate nonlinearities for when a researcher wishes to explain the effect of multiple explanatory variables on the outcome of interest. Yet, in many empirical papers the aim of the researcher is to provide an understanding of how changes in a specific explanatory variable of interest influence outcomes, where we condition on all potential observed confounding variables. If a researcher is interested in the effect of a single explanatory variable on the outcome, the framework proposed in Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) for estimating treatment effects using a machine learning algorithm is appealing. This approach involves three estimation steps to obtain a $\sqrt{n}$ consistent estimate of this parameter of interest and most importantly valid confidence intervals can be constructed. Two of the three estimation steps involve machine learning methods so that the researcher does not need to make strong assumptions about the functional form of the model. Naturally, exogeneity still must hold to obtain unbiased and consistent estimates, but unlike with linear regression methods one does not need to worry that they did not correctly specify the functional form. The three step estimation procedure does remove any regularization bias from using a machine learning algorithm with Neyman orthogonality and also deals with any possible bias from overfitting using sample-splitting.

In summary, this section outlines two strategies of how a researcher whose primary interest is to explain how a covariate influences outcomes can use machine learning to understand the complex functional form of the model that underlies an accurate forecast

Table A25: OLS Estimates of Models of the GUM model and Expanded GUM Models with Nonlinearities in Regressors Suggested by Figure A7

Variable	Model 1		Model 2		GUM	
	Coefficient	Robust S.E	Coefficient	Robust S.E	Coefficient	Robust S.E
Indicator						
$\mathbb{I}_{\text{Screen}\in(-\infty,1.3385)}$	-56.6300	14.1695	-65.0731	11.7315		
$\mathbb{I}_{\text{Screen}\in[1.3385,2.0375)}$	-57.8002	13.2526	-62.2067	11.8837		
$\mathbb{I}_{\text{Screen}\in[2.0375,3.0230)}$	-53.0591	11.8713	-55.4585	11.5559		
$\mathbb{I}_{\text{Screen}\in[3.0230,3.7025)}$	-50.8725	11.3095	-49.6291	11.0167		
$\mathbb{I}_{\text{Screen}\in[3.7025,4.0655)}$	-39.0310	10.3892	-37.8028	10.7892		
$\mathbb{I}_{\text{S:T-1/-3}\in(-\infty,73.1350)}$	-4.2889	3.8930				
$\mathbb{I}_{\text{V:T-1/-3}\in(-\infty,0.4537)}$	-10.8812	4.5893				
$\mathbb{I}_{\text{V:T-14/-20}\in(-\infty,74.8611)}$	-3.6523	2.1559				
$\mathbb{I}_{\text{Budget}\in(-\infty,69.5000)}$	-9.3064	3.1220				
$\mathbb{I}_{\text{Weeks}\in(-\infty,9.5000)}$	0.5816	2.6241				
$\mathbb{I}_{\text{Weeks}\in[9.5000,10.5000)}$	-4.3726	2.6547				
$\mathbb{I}_{\text{S:T-14/-20}\in(-\infty,74.8611)}$	-1.7690	2.5688				
$\mathbb{I}_{\text{S:T-21/-27}\in(-\infty,74.1047)}$	-0.2664	2.3623				
$\mathbb{I}_{\text{V:T-4/-6}\in(-\infty,0.1057)}$	6.9371	2.2008				
Original Variable						
Action	-3.7873	2.5497	-2.6888	2.4691	-1.6895	3.0838
Adventure	4.7358	3.5770	4.7643	3.4842	4.6542	3.7732
Animation	-8.5375	4.1268	-6.8518	4.4028	-1.9354	5.6046
Biography	4.4317	3.6187	0.1811	3.5270	0.1229	4.2324
Comedy	-1.4082	3.6390	-1.6254	3.3573	-0.9595	3.7382
Crime	1.2918	2.2978	1.9982	2.1794	2.6461	2.7335
Drama	-3.6228	3.0274	-2.9136	2.9511	-1.7884	3.6083
Family	11.1923	4.8659	9.1639	5.2832	2.6236	6.7679
Fantasy	9.6846	3.3023	9.4915	3.7590	12.8881	4.9159
Horror	1.9164	2.2681	1.7449	2.3979	3.0486	2.4376
Mystery	2.1528	2.4246	2.4099	2.4099	3.3377	2.4852
Romance	0.9066	2.5781	-1.2987	2.9160	-2.5919	3.3696
Sci-Fi	0.5080	3.2638	-1.0704	2.8807	-0.3705	2.6569
Thriller	1.9914	3.2745	0.4561	2.8743	0.8643	2.9379
PG	13.3045	7.3352	10.8373	6.6338	2.8901	5.4757
PG13	14.0101	7.2553	12.1703	7.0504	1.8691	6.8517
R	16.2428	7.2569	13.7877	7.0876	2.6378	6.6841
Budget	-0.0114	0.0499	0.0535	0.0397	0.1182	0.0399
Weeks	0.0013	0.2890	0.3335	0.2221	0.3738	0.2768
Screens	2.9629	2.4792			6.1694	1.3899
S:T-21/-27	-0.1745	0.5421	-0.2272	0.6159	-0.1570	0.6610
S:T-14/-20	-0.4624	0.7546	-0.0958	0.8689	-0.9835	0.9393
S:T-7/-13	-0.2220	0.8379	-0.9921	0.9688	-1.2435	1.0695
S:T-4/-6	-0.2745	1.0264	0.3448	1.0422	0.2277	1.1775
S:T-1/-3	0.7616	0.7575	1.4817	0.7185	2.5070	0.7509
V:T-21/-27	-80.6518	30.3456	-90.4421	31.6774	-97.5186	31.6624
V:T-14/-20	18.8944	38.5757	33.1787	41.0211	19.4109	38.6929
V:T-7/-13	-34.8571	24.5486	-33.4308	25.6964	-45.2885	30.9011
V:T-4/-6	70.8133	25.6501	65.0400	26.0461	86.2881	27.2008
V:T-1/-3	8.9706	4.2187	11.8043	4.3072	18.9664	5.1687
R-square		0.8532		0.8455		0.7973

Table A26: OLS-post-LASSO Estimates of Models whose Input Set Includes GUM Regressors and Indicators From Figure A7

Variable	Model 1		Model 2	
	Coefficient	Robust S.E	Coefficient	Robust S.E
Indicator				
$\mathbb{I}_{\text{Screen}\in(-\infty,1.3385)}$	-42.2872	13.2098	-56.5420	15.1019
$\mathbb{I}_{\text{Screen}\in(-\infty,2.0375)}$	-0.0514	1.8105	-3.1432	1.9225
$\mathbb{I}_{\text{Screen}\in(-\infty,3.7025)}$	-10.3197	4.5716	-12.1738	4.0516
$\mathbb{I}_{\text{S:T-1/-3}\in(-\infty,73.1350)}$	-4.6171	1.7198		
$\mathbb{I}_{\text{V:T-1/-3}\in(-\infty,0.4537)}$	-9.9209	4.7654	-15.6941	5.5244
$\mathbb{I}_{\text{Budget}\in(-\infty,69.5000)}$	-6.5649	3.6927	-7.0845	4.3126
$\mathbb{I}_{\text{Weeks}\in(-\infty,10.5000)}$	-0.0838	4.3860		
$\mathbb{I}_{\text{Screen}\in(-\infty,2.0375)} * \mathbb{I}_{\text{Weeks}\in(-\infty,10.5000)}$	-3.7377	4.4305		
Original Variable				
Family	13.0173	5.1538	13.3995	5.8497
R	0.0231	0.0529	-0.0016	0.0605
Weeks	4.4881	0.8043	4.6789	0.8435
S:T-1/-3	-14.5734	3.1721		
V:T-4/-6	14.4344	3.8020	4.1407	4.0937
R-square		0.8418		0.8006

with the data. For researchers seeking to understand the effects of multiple explanatory variables we are proposing a hybrid strategy where the machine learning algorithm is first used to identify the complex interactions that need to be accounted for. We illustrate this approach using RT and future work can show how to adopt it to work with SVR with linear kernels. Statistical inference can be undertaken, although procedures such as Cattaneo, Jansson, and Ma (2019) can be considered to deal with the inclusion of many covariates in a first-step estimate entering a two-step estimation procedure. We believe these estimation strategies could be valuable in settings where theory does not provide guidance to an appropriate model to explain how the data was generated.

F.21 Results of Relative Prediction Efficiency by MSFE and MAFE

The two panels of table A27 report the median MSFE and MAFE from the prediction error exercise outlined in the preceding section for the 10 different econometric strategies listed in panel A of table 2 in the main text. Each row of the table considers a different size for the evaluation set and to ease interpretation all MSFEs and MAFEs are normalized by the MSFE and MAFE of the HRC^p. Panel A of table A27 presents results for forecasting open box office and panel B demonstrates results corresponding to forecasting retail movie unit sales. Notice that for open box office, all remaining entries for MSFE are larger than one, indicating inferior performance of the respective estimator relative to HRC^p. In general, the three model averaging approaches and the model selected by AIC perform nearly as well as HRC^p. For movie unit sales, HPMA yields the best results in the majority of experiments. However, the gains from using HPMA in place of PMA appear quite small.

The results in table A27 also stress the importance of social media data for forecast accuracy. Models that ignore social media data (MTV) perform poorly relative to all other strategies. In contrast to Lehrer and Xie (2017) we find that the post-Lasso methods, including the double-Lasso method, OLS post Lasso and model averaging post Lasso perform poorly relative to HRC^p in this application. This likely arise since all movies released are considered rather than only those with budgets ranging from 20 to 100 million dollars, thereby increasing the presence of heteroskedasticity in the data.

Table A28 examines the performance of alternative model screening strategies listed in panel B of table 2 in the main text relative to HRC^p. We observe small gains in forecast accuracy from model screening relative to the benchmark HRC^p. The hetero-robust methods yields slightly better results than homo-efficient methods for forecasts of box office opening. In contrast, when forecasting retail movie unit sales, the homo-efficient ARMS demonstrates better results than the other screening methods. Taking these findings together with the results contrasting PMA to HPMA table A27 illustrate that there are small gains in practice from using econometric approaches that accommodate heteroskedasticity.

Table A29 demonstrates that there are very large gains in prediction efficiency of either the recursive partitioning algorithms or the suite of advanced machine learning strategies listed in panel D of table 2 of main text relative to the benchmark HRC^p. The subscript below RF and MARF refer to the number of randomly chosen explanatory variables used to determine a split at each node. For both outcomes when n_E is small, machine learning methods have dominating performance over the HRC$_p$. Popular approaches such as bagging and random forest greatly outperform the benchmark. However, our proposed MASVR$_{\text{LS}}$ has the best performance when evaluated by either MSFE or MAFE. We find larger gains from the hybrid strategy involving support vector regression instead of tree based strategies with open box revenue relative to retail movie unit sales. However, the percentage gain in forecast accuracy is higher for retail movie unit sales due to the smaller sample size. We find the relative performance of HBART to the tree based procedures improves with the larger sample used to predict DVD and Blu-Ray sales. Adding model averaging tends to lead to gains of 10% between either SVR$_{\text{LS}}$ and MASVR$_{\text{LS}}$ or bagging and MAB. Random forest methods, both conventional and model averaging, have moderate performance in all cases. Note that as n_E increases, all statistical learning methods observe decreases in performance. We also stress that there are large gains in performance of all strategies in table A29 relative to the results presented in tables A27 and A28.

Figures A8 to A10 correspond to the results of tables A27 to A29, respectively. In each figure, subplots (a) to (d) correspond to MSFE-Open Box, MAFE-Open Box, MSFE-DVD, and MAFE-DVD, where the solid, dashed, dash-solid, and dots line represent the results of $n = 10, 20, 30$, and 40, respectively. We list the methods in the x-axis and the y-axis reports the risks. We highlight the best method for each n with a circle.

Table A27: Results of Relative Prediction Efficiency by MSFE and MAFE

n_E	GUM	MTV	GETS	AIC	PMA	HPMA	JMA	OLS_{10}	OLS_{12}	OLS_{15}	HRC^p_{10}	HRC^p_{12}	HRC^p_{15}	HRC^p
Panel A: Open Box Office														
Mean Squared Forecast Error (MSFE)														
10	1.1035	2.3032	1.2357	1.0274	1.0022	1.0018	1.0274	1.1223	1.1390	1.1208	1.1205	1.1335	1.1068	**1.0000**
20	1.1328	2.5704	1.2208	1.0246	1.0030	1.0028	1.0221	1.1634	1.1757	1.0833	1.1638	1.1717	1.0863	**1.0000**
30	1.1561	2.5402	1.2305	1.0253	1.0022	1.0012	1.0153	1.2067	1.2284	1.0769	1.2021	1.2209	1.0807	**1.0000**
40	1.1892	2.4835	1.2198	1.0215	1.0018	1.0016	1.0054	1.2160	1.2338	1.0580	1.2161	1.2314	1.0556	**1.0000**
Mean Absolute Forecast Error (MAFE)														
10	1.0597	1.5235	1.1300	1.0194	1.0012	**0.9999**	1.0060	1.0515	1.0589	1.0669	1.0543	1.0591	1.0691	1.0000
20	1.0751	1.5317	1.1258	1.0174	1.0013	**0.9998**	1.0084	1.0543	1.0604	1.0602	1.0562	1.0611	1.0615	1.0000
30	1.0814	1.5251	1.1373	1.0168	1.0026	1.0003	1.0137	1.0571	1.0681	1.0548	1.0588	1.0658	1.0564	**1.0000**
40	1.0929	1.5275	1.1376	1.0207	1.0002	1.0013	1.0038	1.0551	1.0665	1.0564	1.0560	1.0666	1.0555	**1.0000**
Panel B: Movie Unit Sales														
Mean Squared Forecast Error (MSFE)														
10	1.4183	2.4468	1.5231	1.0499	1.0013	1.0019	1.0183	1.3730	1.3481	1.3531	1.3524	1.3302	1.3449	**1.0000**
20	1.5010	2.2299	1.5895	1.0514	**0.9979**	0.9998	1.0263	1.3951	1.2665	1.2617	1.3695	1.2546	1.2498	1.0000
30	1.6988	2.1005	1.5836	1.0455	**0.9943**	0.9981	1.0218	1.3341	1.2393	1.2071	1.3104	1.2348	1.2047	1.0000
40	1.8518	1.9312	1.5235	1.0444	**0.9964**	1.0013	1.0227	1.2205	1.1579	1.1364	1.1947	1.1420	1.1252	1.0000
Mean Absolute Forecast Error (MAFE)														
10	1.1507	1.5950	1.2693	1.0296	1.0015	1.0016	1.0149	1.2354	1.2284	1.1634	1.2297	1.2211	1.1599	**1.0000**
20	1.1863	1.5342	1.2792	1.0266	1.0007	1.0009	1.0146	1.2047	1.1852	1.1365	1.1980	1.1772	1.1310	**1.0000**
30	1.2333	1.5388	1.2886	1.0312	1.0024	1.0013	1.0144	1.1904	1.1735	1.1165	1.1791	1.1642	1.1137	**1.0000**
40	1.2828	1.4793	1.2861	1.0244	**0.9983**	1.0009	1.0157	1.1551	1.1435	1.0952	1.1458	1.1365	1.0900	1.0000

Note: Bold numbers denote the strategy with the best performance in that row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the HRC^p method presented in the last column. OLS_q and HRC^p_q stand for OLS and HRC^p with q number of covariates selected by the Lasso.

Table A28: Comparing Hetero-robust and Homo-efficient Model Screening Methods

n_E	GETS						ARMS						HRMS	HEMS	Benchmark
	Hetero-robust			Homo-efficient			Hetero-robust			Homo-efficient					
	$(p = 0.24$	0.28	0.32)	$(p = 0.24$	0.28	$0.32)^{\dagger}$	$(L = 100$	50	25)	$(L = 100$	50	25)			
	$(p = 0.30$	0.34	0.38)	$(p = 0.30$	0.34	$0.38)^{\ddagger}$									
Panel A: Open Box Office															
							Mean Squared Forecast Error (MSFE)								
10	0.9992	1.0040	0.9999	0.9954	0.9989	1.0021	0.9825	0.9813	**0.9751**	0.9820	0.9834	0.9926	1.0121	1.0172	1.0000
20	**0.9878**	1.0005	0.9996	0.9809	0.9971	1.0190	0.9994	0.9971	0.9908	1.0005	1.0000	0.9951	1.0143	1.0136	1.0000
30	**0.9927**	0.9991	1.0007	0.9939	1.0019	0.9997	0.9947	0.9929	1.0006	0.9987	1.0015	0.9998	1.0466	1.0283	1.0000
40	0.9921	0.9983	1.0025	0.9671	0.9990	1.0075	1.0045	0.9874	**0.9842**	1.0010	1.0094	1.0066	1.0449	1.0296	1.0000
							Mean Absolute Forecast Error (MAFE)								
10	1.0019	1.0034	1.0025	**0.9809**	1.0114	1.0037	0.9890	0.9930	1.0002	0.9904	0.9875	1.0008	1.0135	1.0143	1.0000
20	0.9955	0.9994	0.9986	0.9932	0.9978	1.0118	0.9944	0.9968	0.9956	0.9898	0.9894	**0.9863**	1.0042	1.0000	1.0000
30	0.9992	1.0015	1.0011	**0.9814**	1.0124	0.9881	0.9990	0.9976	1.0022	0.9988	0.9966	0.9972	1.0098	1.0059	1.0000
40	0.9974	1.0031	1.0020	0.9912	1.0113	0.9930	0.9954	**0.9886**	0.9930	0.9950	0.9938	0.9914	1.0172	1.0072	1.0000
Panel B: Movie Unit Sales															
							Mean Squared Forecast Error (MSFE)								
10	1.0370	1.0008	0.9940	1.0338	0.9799	0.9880	0.9620	0.9577	0.9598	0.9613	0.9504	**0.9328**	1.0481	1.0380	1.0000
20	1.0388	1.0002	0.9912	1.0374	1.0033	1.0097	0.9675	0.9713	0.9682	0.9482	0.9318	**0.9271**	1.1770	1.1245	1.0000
30	1.0309	1.0003	0.9913	1.0290	1.0010	1.0019	0.9765	0.9811	0.9843	0.9471	0.9394	**0.9344**	1.1491	1.1072	1.0000
40	1.0113	0.9977	0.9985	1.0063	1.0023	1.0004	0.9600	0.9519	0.9615	0.9316	0.9370	**0.9202**	1.2418	1.1842	1.0000
							Mean Absolute Forecast Error (MAFE)								
10	1.0122	0.9988	0.9923	1.0036	0.9881	0.9926	0.9778	0.9728	**0.9681**	0.9819	0.9828	0.9773	1.0242	1.0067	1.0000
20	1.0215	1.0001	0.9953	1.0059	1.0025	0.9859	0.9818	0.9818	0.9808	0.9814	0.9809	**0.9766**	1.0544	1.0340	1.0000
30	1.0203	1.0000	0.9966	1.0038	1.0000	1.0026	1.0014	0.9952	0.9919	0.9915	0.9920	**0.9866**	1.0637	1.0477	1.0000
40	1.0134	0.9997	0.9956	1.0213	1.0079	1.0011	0.9809	0.9742	0.9722	0.9725	0.9714	**0.9689**	1.0787	1.0664	1.0000

Note: Bold numbers denote the strategy with the best performance in that row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the HRC^{p} method presented in the last column.

Table A29: Results of Relative Prediction Efficiency Between Machine Learning and Model Averaging Learning

n_E	BART	BART_{BMA}	HBART	BOOST	Reg.Tree	Bagging	Random Forest		MAB	MA Random Forest		SVR_{LS}	MASVR_{LS}	Benchmark
							RF_{15}	RF_{20}		MARF_{15}	MARF_{20}			
Panel A: Open Box Office														
Mean Squared Forecast Error (MSFE)														
10	0.5853	0.5561	0.4575	0.5606	0.6142	0.5883	0.5755	0.5313	0.5066	0.5356	0.5519	0.4494	**0.4111**	1.0000
20	0.7333	0.7211	0.5903	0.6869	0.8834	0.8622	0.7901	0.8157	0.7315	0.7787	0.8272	0.4930	**0.4783**	1.0000
30	0.7616	0.7407	0.6562	0.7685	1.0214	0.8589	0.8353	0.8476	0.7531	0.8694	0.8802	0.5132	**0.4996**	1.0000
40	0.8251	0.7652	0.7302	0.8899	1.3120	0.9833	0.9395	0.9994	0.9145	1.0348	0.9915	0.5695	**0.5524**	1.0000
Mean Absolute Forecast Error (MAFE)														
10	0.7395	0.7152	0.6416	0.6278	0.6792	0.7036	0.7018	0.6652	0.6232	0.6742	0.6703	0.5464	**0.4980**	1.0000
20	0.8182	0.7866	0.7120	0.6983	0.7443	0.7690	0.7513	0.7474	0.6955	0.7495	0.7444	0.5833	**0.5529**	1.0000
30	0.8144	0.8001	0.7452	0.7116	0.7897	0.7789	0.7655	0.7625	0.7042	0.7733	0.7654	0.6032	**0.5737**	1.0000
40	0.8588	0.8432	0.7809	0.7855	0.8390	0.8080	0.8084	0.8072	0.7625	0.8157	0.8113	0.6294	**0.6052**	1.0000
Panel B: Movie Unit Sales														
Mean Squared Forecast Error (MSFE)														
10	0.9813	0.9251	0.7363	1.0063	1.1288	0.8299	0.8678	0.8780	0.7307	0.9168	0.8674	0.6394	**0.5685**	1.0000
20	1.0080	0.9461	0.7755	1.1014	1.0547	0.8563	0.8900	0.9551	0.7009	1.0564	0.8884	0.6762	**0.5979**	1.0000
30	1.0098	0.9548	0.8079	1.1476	1.1610	0.9610	0.9820	1.0682	0.7494	1.1702	0.9849	0.7648	**0.6717**	1.0000
40	1.0951	0.9547	0.8687	1.3274	1.2739	1.0085	1.0456	1.1065	0.8626	1.1832	1.0391	0.8518	**0.7905**	1.0000
Mean Absolute Forecast Error (MAFE)														
10	0.8984	0.8661	0.7874	0.7764	0.8407	0.8294	0.8479	0.8625	0.7461	0.9098	0.8493	0.6139	**0.5498**	1.0000
20	0.9463	0.8976	0.8355	0.8316	0.8453	0.8302	0.8622	0.8791	0.7564	0.9313	0.8611	0.6770	**0.6067**	1.0000
30	0.9779	0.9444	0.8590	0.9055	0.8924	0.8668	0.8974	0.9225	0.7954	0.9722	0.8968	0.7097	**0.6488**	1.0000
40	1.0029	0.9991	0.8996	0.9910	0.9361	0.8914	0.9191	0.9455	0.8211	0.9805	0.9181	0.7644	**0.6940**	1.0000

Note: Bold numbers denote the strategy with the best performance in that row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the HRC^p method presented in the last column. The subscript in RF_q and MARF_q respectively stand for the number of covariates randomly chosen at each node to consider as the potential split variable. All bagging and random forest estimates involve 100 trees.

Figure A8: Demonstrate Results of Table 3 in Figures

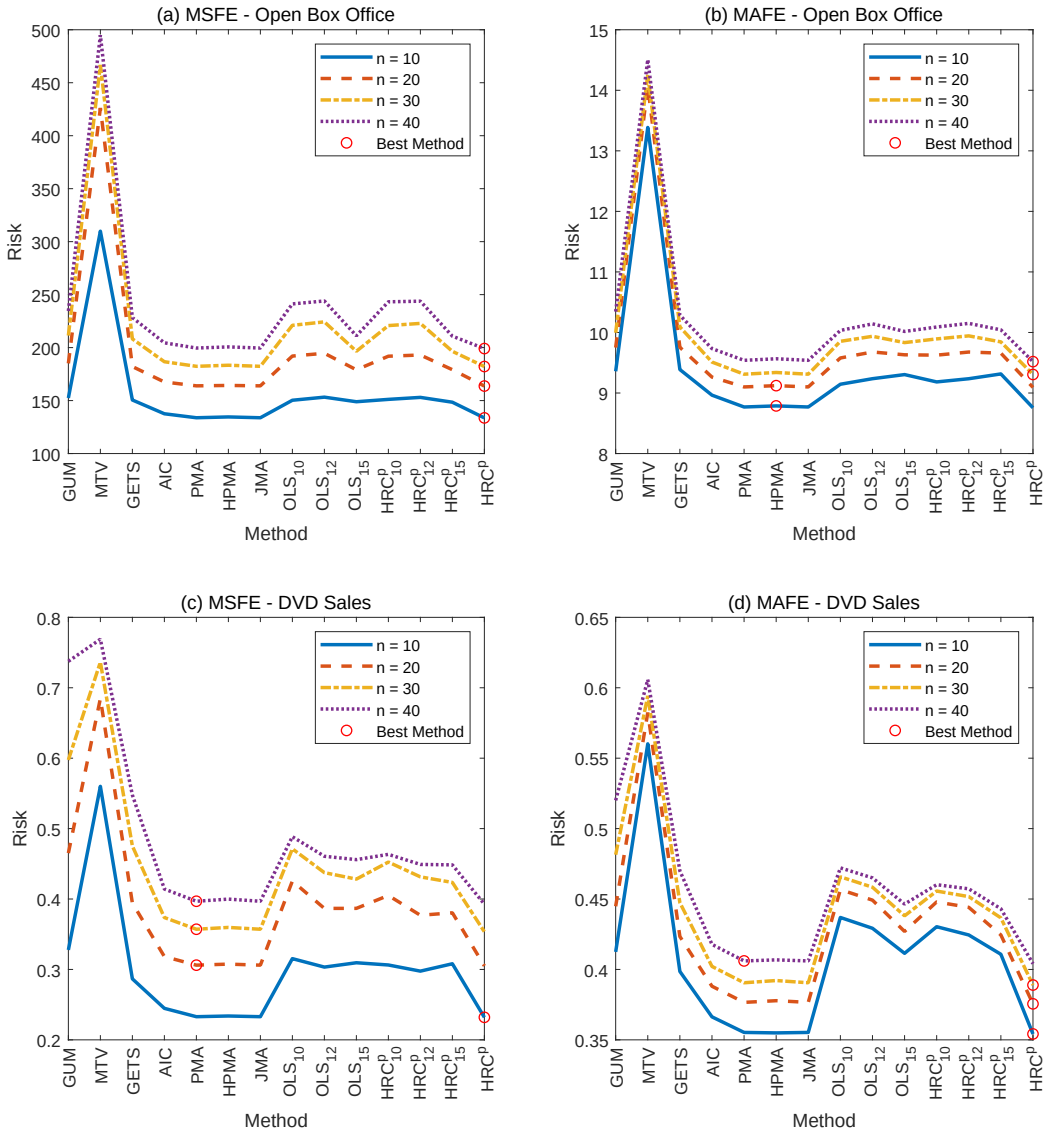


Figure A9: Demonstrate Results of Table 4 in Figures

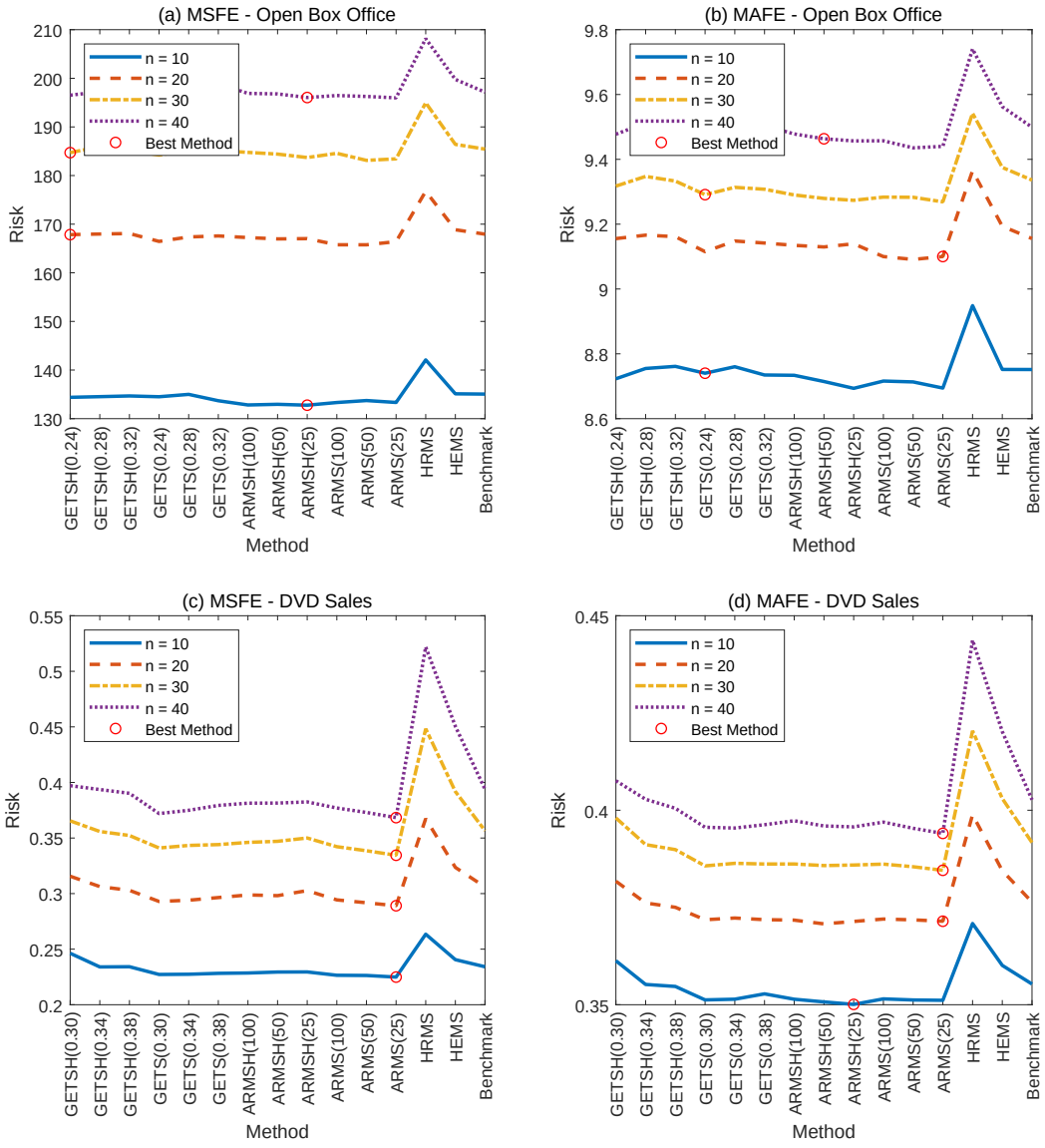
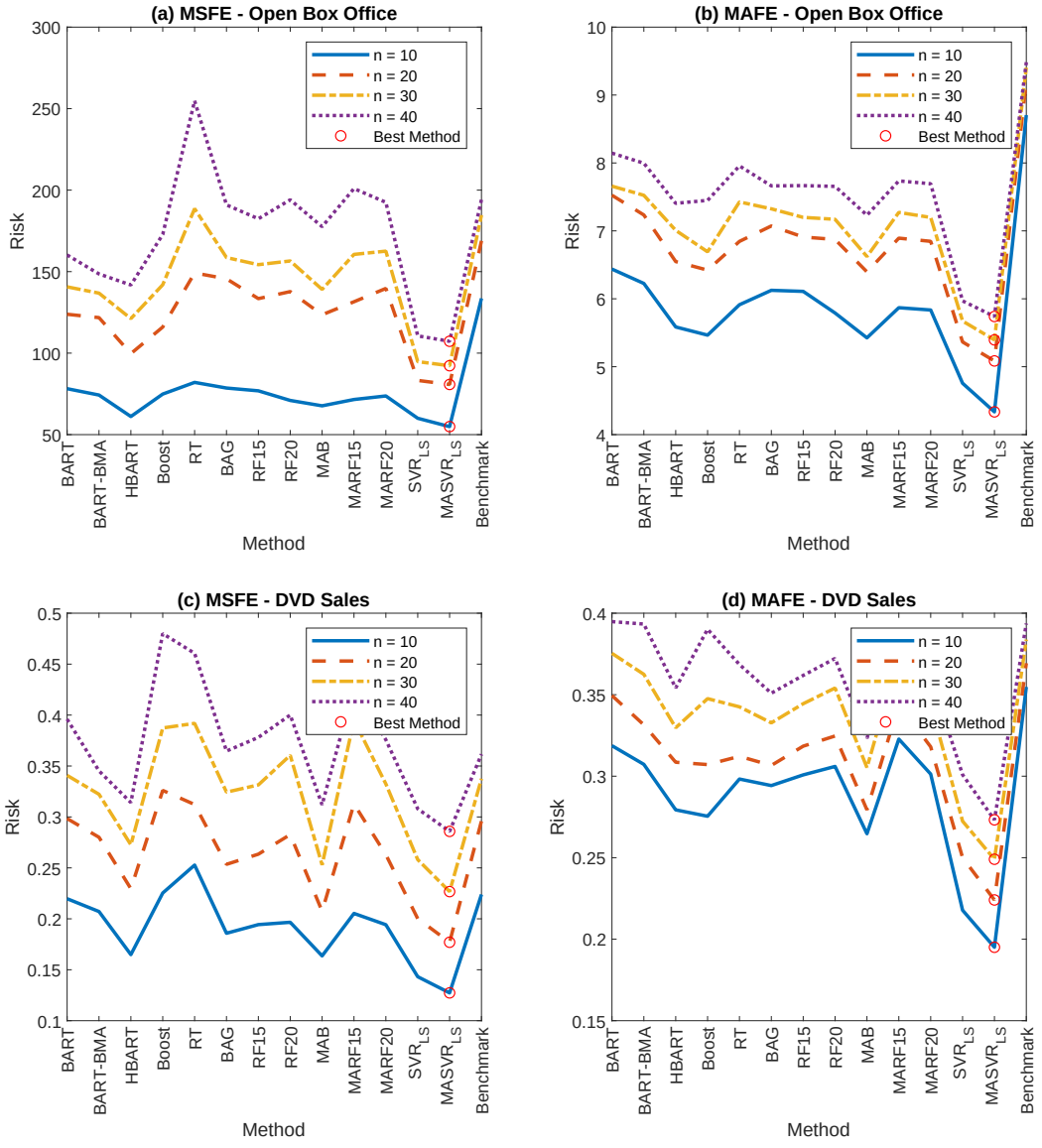


Figure A10: Demonstrate Results of Table 5 in Figures



References

AHMED, S., AND A. SINHA (2016): "When It Pays to Wait: Optimizing Release Timing Decisions for Secondary Channels in the Film Industry," *Journal of Marketing*, 80(4), 20–38.

AKAIKE, H. (1973): "Information theory and an extension of the maximum likelihood principle," *Second International Symposium on Information Theory*, pp. 267–281.

AMEMIYA, T. (1980): "Selection of Regressors," *International Economic Review*, 21(2), 331–354.

ANDO, T., AND K.-C. LI (2014): "A Model-Averaging Approach for High-Dimensional Regression," *Journal of the American Statistical Association*, 109(505), 254–265.

ANTIPOV, E. A., AND E. B. POKRYSHEVSKAYA (2017): "Are box office revenues equally unpredictable for all movies? Evidence from a Random forest-based model," *Journal of Revenue and Pricing Management*, 16(3), 295–307.

BAEK, H., S. OH, H.-D. YANG, AND J. AHN (2017): "Electronic word-of-mouth, box office revenue and social media," *Electronic Commerce Research and Applications*, 22, 13–23.

BANDARI, R., S. ASUR, AND B. HUBERMAN (2012): "The Pulse of News in Social Media: Forecasting Popularity," *ICWSM 2012 - Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*.

BASUROY, S., S. CHATTERJEE, AND S. A. RAVID (2003): "How Critical Are Critical Reviews? The Box Office Effects of Film Critics, Star Power, and Budgets," *Journal of Marketing*, 67, 103–117.

BASUROY, S., AND S. A. RAVID (2014): "How relevant are experts in the internet age? Evidence from the motion pictures industry," *Working Paper*.

BELLONI, A., AND V. CHERNOZHUKOV (2013): "Least Squares after Model Selection in High-Dimensional Sparse Models," *Bernoulli*, 19(2), 521–547.

BOSER, B. E., I. M. GUYON, AND V. N. VAPNIK (1992): "A Training Algorithm for Optimal Margin Classifiers," in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pp. 144–152. ACM Press.

BREIMAN, L. (1996): "Bagging Predictors," *Machine Learning*, 26, 123–140.

——— (2001): "Random Forests," *Machine Learning*, 45, 5–32.

BREIMAN, L., J. FRIEDMAN, AND C. J. STONE (1984): *Classification and Regression Trees.* Chapman and Hall/CRC.

CATTANEO, M. D., M. JANSSON, AND X. MA (2019): "Two-step Estimation and Inference with Possibly Many Included Covariates," *Review of Economic Studies*, 86(3), 1095–1122.

CAWLEY, G., N. TALBOT, R. FOXALL, S. DORLING, AND D. MANDIC (2004): "Heteroscedastic kernel ridge regression," *Neurocomputing*, 57, 105–124.

CHAKRAVARTY, A., Y. LIU, AND T. MAZUMDAR (2010): "The Differential Effects of Online Word-of-Mouth and Critics Reviews on Pre-Release Movie Evaluation," *Journal of Interactive Marketing*, 24(3), 185–197.

CHAUDHURI, P., W.-D. LO, W.-Y. LOH, AND C.-C. YANG (1995): "Bagging Predictors," *Generalized Regression Trees*, 5, 641–666.

CHEN, X., Y. CHEN, AND C. WEINBERG (2012): "Learning about movies: The impact of movie release types on the nationwide box office," *Journal of Cultural Economics*, 37(9), 359–386.

CHERNOZHUKOV, V., D. CHETVERIKOV, M. DEMIRER, E. DUFLO, C. HANSEN, W. NEWEY, AND J. ROBINS (2018): "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21(1), C1–C68.

CHEVALIER, J., Y. DOVER, AND D. MAYZLIN (2018): "Channels of Impact: User Reviews When Quality Is Dynamic and Managers Respond," *Marketing Science*, 37.

CHINTAGUNTA, P., S. GOPINATH, AND S. VENKATARAMAN (2010): "The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation Across Local Markets," *Marketing Science*, 29(5), 944–957.

CHIPMAN, H. A., E. I. GEORGE, AND R. E. MCCULLOCH (2010): "BART: Bayesian Additive Regression Trees," *The Annals of Applied Statistics*, 4.

CLARKE, W. R., P. A. LACHENBRUCH, AND B. BROFFITT (1979): "How non-normality affects the quadratic discriminant function," *Communications in Statistics - Theory and Methods*, 8(13), 1285–1301.

COURANT, R., AND D. HILBERT (1953): *Methods of Mathematical Physics (Vol.1)*. Interscience Publishers, New York, NY, USA.

DAHL, G., AND S. DELLAVIGNA (2008): "Does Movie Violence Increase Violent Crime?," *Quarterly Journal of Economics*, 124(2), 677–734.

DELLAROCAS, C., X. M. ZHANG, AND N. AWAD (2007): "Exploring the Value of Online Product Reviews in Forecasting Sales: The Case of Motion Pictures," *Journal of Interactive Marketing*, 21(4), 23 – 45.

DING, C., H. CHENG, Y. DUAN, AND Y. JIN (2016): "The Power of the "Like" Button: The Impact of Social Media on Box Office," *Decision Support Systems*, 94, 77–84.

DOBRA, A., AND J. GEHRKE (2002): "SECRET: A scalable linear regression tree algorithm," in *In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 481–487. ACM Press.

DRUCKER, H., C. J. C. BURGES, L. KAUFMAN, A. J. SMOLA, AND V. VAPNIK (1996): "Support Vector Regression Machines," in *Advances in Neural Information Processing Systems 9*, ed. by M. C. Mozer, M. I. Jordan, and T. Petsche, pp. 155–161. MIT Press.

DURLAUF, S. N., S. NAVARRO, AND D. A. RIVERS (2016): "Model uncertainty and the effect of shall-issue right-to-carry laws on crime," *European Economic Review*, 81, 32 – 67.

EINAV, L. (2007): "Seasonality in the U.S. Motion Picture Industry," *The RAND Journal of Economics*, 38, 127 – 145.

ELBERSE, A., AND J. ELIASHBERG (2003): "Demand and Supply Dynamics for Sequentially Released Products in International Markets: The Case of Motion Pictures," *Marketing Science*, 22(3), 329–354.

ELIASHBERG, J., AND S. SHUGAN (1997): "Film Critics: Influencers or Predictors?," *Journal of Marketing*, 61(2), 68–78.

FOXALL, R., G. CAWLEY, N. TALBOT, S. DORLING, AND D. MANDIC (2002): "Heteroscedastic regularised kernel regression for prediction of episodes of poor air quality," in *ESANN 2002, 10th Eurorean Symposium on Artificial Neural Networks, Bruges, Belgium*, pp. 19–24.

FRIEDMAN, J. H. (2001): "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, 29(5), 1189–1232.

GEMSER, G., M. OOSTRUM, AND M. LEENDERS (2007): "The impact of film reviews on the box office performance of art house versus mainstream motion pictures," *Journal of Cultural Economics*, 31, 43–63.

GERMAIN, D. (2013): "2012 Box Office Hits Record $10.8 Billion; Ticket Sales Increase For First Time In 3 Years," *Huff Post Entertainment*.

GOPINATH, S., P. CHINTAGUNTA, AND S. VENKATARAMAN (2013): "Blogs, Advertising and Local-Market Movie Box-Office Performance," *Management Science*, 59, 1–20.

HANNAK, A., E. ANDERSON, L. F. BARRETT, S. LEHMANN, A. MISLOVE, AND M. RIEDEWALD (2012): "Tweetin ' in the Rain: Exploring Societal-scale Effects of Weather on Mood," *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pp. 479–482.

HANSEN, B. E. (2007): "Least Squares Model Averaging," *Econometrica*, 75(4), 1175–1189.

——— (2020): *Econometrics,* pp. 286–301. University of Wisconsin. Department of Economics.

HANSEN, B. E., AND J. S. RACINE (2012): "Jackknife Model Averaging," *Journal of Econometrics*, 167(1), 38–46.

HANSEN, P. R. (2005): "A Test for Superior Predictive Ability," *Journal of Business & Economic Statistics*, 23(4), 365–380.

HASTIE, T., R. TIBSHIRANI, AND J. FRIEDMAN (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

HENDRY, D. F., AND B. NIELSEN (2007): *Econometric Modeling: A Likelihood Approach,* chap. 19, pp. 286–301. Princeton University Press.

HERNÁNDEZ, B., A. E. RAFTERY, S. R. PENNINGTON, AND A. C. PARNELL (2018): "Bayesian Additive Regression Trees using Bayesian model averaging," *Statistics and Computing*, 28(4), 869–890.

HIRSCHMAN, E. C., AND A. PIEROS (1985): "RELATIONSHIPS AMONG INDICATORS OF SUCCESS IN BROADWAY PLAYS AND MOTION PICTURES," *Journal of Cultural Economics*, 9(1), 35–63.

HOLBROOK, M., AND M. ADDIS (2008): "Art versus commerce in the movie industry: a Two-Path Model of Motion-Picture Success," *Journal of Cultural Economics*, 32, 87–107.

HOLBROOK, M. B. (1999): "Popular Appeal Versus Expert Judgments of Motion Picture," *Journal of Consumer Research*, 26(2), 144–155.

HONG, B., W. ZHANG, J. YE, D. CAI, X. HE, AND J. WANG (2019): "Scaling Up Sparse Support Vector Machine by Simultaneous Feature and Sample Reduction," *Journal of Machine Learning Research*, 20, 1–39.

HOUSTON, M., A.-K. KUPFER, T. HENNIG-THURAU, AND M. SPANN (2018): "Pre-release consumer buzz," *Journal of the Academy of Marketing Science*, 46, 338–360.

HUI, S. K., J. ELIASHBERG, AND E. I. GEORGE (2008): "Modeling DVD Preorder and Sales: An Optimal Stopping Approach," *Marketing Science*, 27(6), 1097–1110.

HUR, M., P. KANG, AND S. CHO (2016): "Box-office Forecasting based on Sentiments of Movie Reviews and Independent Subspace Method," *Information Sciences*, 372, 608–624.

IZENMAN, A. J. (2013): "Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning," New York: Springer.

KAPLAN, J. J. (2012): "Turning followers into dollars: The impact of social media on a movie's financial performance," *Undergraduate Economic Review*, 9(1), 1–12.

KIM, H., AND W.-Y. LOH (2003): "Classification Trees With Bivariate Linear Discriminant Node Models," *Journal of Computational and Graphical Statistics*, 12(3), 512–530.

KIM, K., S. YOON, AND Y. K. CHOI (2018): "The effects of eWOM volume and valence on product sales – an empirical examination of the movie industry," *International Journal of Advertising*, 38(3), 471–488.

KIM, T., J. HONG, AND P. KANG (2015): "Box office forecasting using machine learning algorithms based on SNS data," *International Journal of Forecasting*, 31(2), 364–390.

KOSCHAT, M. (2012): "The Impact of Movie Reviews on Box Office: Media Portfolios and the Intermediation of Genre," *Journal of Media Economics*, 25(1), 35–53.

LEE, Y.-J., K. HOSANAGAR, AND Y. TAN (2015): "Do I Follow My Friends or the Crowd? Information Cascades in Online Movie Ratings," *Management Science*, 61(9), 2241–2258.

LEHRER, S., T. XIE, AND T. ZENG (2020): "Does High-Frequency Social Media Data Improve Forecasts of Low-Frequency Consumer Confidence Measures?," *Journal of Financial Econometrics*, p. in press.

LEHRER, S. F., AND T. XIE (2017): "Box Office Buzz: Does Social Media Data Steal the Show from Model Uncertainty When Forecasting for Hollywood?," *The Review of Economics and Statistics*, 99(5), 749–755.

LIU, Q., AND R. OKUI (2013): "Heteroskedasticity-robust C_p Model Averaging," *The Econometrics Journal*, 16, 463–472.

LIU, Y. (2006): "Word-of-Mouth for Movies: Its Dynamics and Impact on Box Office Revenue," *Journal of Marketing*, 70(3), 74–89.

MAYZLIN, D., Y. DOVER, AND J. CHEVALIER (2014): "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," *American Economic Review*, 104(8), 2421–2455.

MERCER, J. (1909): "Functions of Positive and Negative Type, And Their Connection with the Theory of Integral Equations," *Philosophical Transactions of the Royal Society A*, 209, 415–446.

MESTYÁN, M., T. YASSERI, AND J. KERTÉSZ (2013): "Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data," *PloS one*, 8(8), 1–8.

MISLOVE, A., S. L. JORGENSEN, Y.-Y. AHN, J.-P. ONNELA, AND J. N. ROSENQUIST (2011): "Understanding the Demographics of Twitter Users," *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pp. 554–557.

MOON, S., P. BERGEY, AND D. IACOBUCCI (2010): "Dynamic Effects Among Movie Ratings, Movie Revenues, and Viewer Satisfaction," *Journal of Marketing*, 74, 108–121.

MORTIMER, J. H. (2007): "Price Discrimination, Copyright Law, and Technological Innovation: Evidence from the Introduction of DVDs*," *The Quarterly Journal of Economics*, 122(3), 1307–1350.

MOUL, C. (2007): "Measuring Word of Mouth's Impact on Theatrical Movie Admissions," *Journal of Economics & Management Strategy*, 16(4), 859–892.

NEELAMEGHAM, R., AND P. CHINTAGUNTA (1999): "A Bayesian Model to Forecast New Product Performance in Domestic and International Markets," *Marketing Science*, 18(2), 115–136.

ORDEN, E. (2013): "Home Movie Sales Log Rare Increase," *Wall Stree Journal*, [Online; posted Jan. 8, 2013].

PRAG, J., AND J. CASAVANT (1994): "An Empirical Study of Determinants of Revenues and Marketing Expenditures in the Motion Picture Industry," *Journal of Cultural Economics*, 18(3), 217–235.

PRATOLA, M. T., H. A. CHIPMAN, E. I. GEORGE, AND R. E. MCCULLOCH (2019): "Heteroscedastic BART via Multiplicative Regression Trees," *Journal of Computational and Graphical Statistics*, Forthcoming.

PRIETO-RODRIGUEZ, J., F. GUTIERREZ-NAVRATIL, AND V. ATECA-AMESTOY (2014): "Theatre allocation as a distributor's strategic variable over movie runs," *Journal of Cultural Economics*, 39(1), 65–83.

PROBST, P., AND A.-L. BOULESTEIX (2018): "To Tune or Not to Tune the Number of Trees in Random Forest," *Journal of Machine Learning Research*, 18, 1–18.

PROBST, P., A.-L. BOULESTEIX, AND B. BISCHL (2019): "Tunability: Importance of Hyperparameters of Machine Learning Algorithms," *Journal of Machine Learning Research*, 20, 1–32.

QUINLAN, J. R. (1992): "Learning With Continuous Classes," pp. 343–348. World Scientific.

RAVI, K. (2015): "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," *Knowledge-Based Systems*, 89(1), 14–46.

ROB, R., AND J. WALDFOGEL (2007): "Piracy on the Silver Screen," *The Journal of Industrial Economics*, 55(3), 379–395.

RUI, H., Y. LIU, AND A. WHINSTON (2011): "Whose and What Chatter Matters? The Impact of Tweets on Movie Sales," *Decision Support Systems*, 55(4), 863–870.

SAWHNEY, M., AND J. ELIASHBERG (1996): "A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures," *Marketing Science*, 15(2), 113–131.

SCHAPIRE, R. E. (1990): "The strength of weak learnability," *Machine Learning*, 5(2), 197–227.

SCHWEIDEL, D., AND W. MOE (2014): "Listening In on Social Media: A Joint Model of Sentiment and Venue Format Choice," *Journal of Marketing Research*, 51(4), 387–402.

SCORNET, E. (2017): "Tuning parameters in random forests," *ESAIM: Procs*, 60, 144–162.

SCORNET, E., G. BIAU, AND J.-P. VERT (2015): "Consistency of Random Forests," *The Annals of Statistics*, 43(4), 1716–1741.

SHRUTI, S. ROY, AND W. ZENG (2014): "Influence of social media on performance of movies," pp. 1–6.

SMITH, M., AND R. TELANG (2009): "Competing With Free: The Impact of Movie Broadcasts on DVD Sales and Internet Piracy," *MIS Quarterly*, 33(2), 321–338.

SMITH, M. D., AND R. TELANG (2010): "Piracy or promotion? The impact of broadband Internet penetration on DVD sales," *Information Economics and Policy*, 22(4), 289–298.

SONNIER, G., L. MCALISTER, AND O. RUTZ (2011): "A Dynamic Model of the Effect of Online Communications on Firm Sales," *Marketing Science*, 30(4), 702–716.

SUYKENS, J. (2000): "Least squares support vector machines for classification and nonlinear modelling," *Neural Network World*, 10, 29–47.

SUYKENS, J., AND J. VANDEWALLE (1999): "Least Squares Support Vector Machine Classifiers," *Neural Processing Letters*, 9.

SUYKENS, J. A. K., T. V. GESTEL, J. D. BRABANTER, B. D. MOOR, AND J. VANDEWALLE (2002): *Least Squares Support Vector Machines*. World Scientific Publishing Company, Singapore.

TERRY, N., N. TERRY, AND D. DE'ARMOND (2011): "The determinants of domestic box office performance in the motion picture industry," *Southwestern Economic Review*, 32, 137–148.

TIBSHIRANI, R. (1996): "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

TORGO, L. (1997): "Functional Models for Regression Tree Leaves," in *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pp. 385–393, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

TSAO, W.-C. (2014): "Which type of online review is more persuasive? The influence of consumer reviews and critic ratings on moviegoers," *Electronic Commerce Research*, 14, 559–583.

ULLAH, A., AND H. WANG (2013): "Parametric and Nonparametric Frequentist Model Selection and Model Averaging," *Econometrics*, 1, 157–179.

VALIANT, L. G. (1984): "A Theory of the Learnable," *Commun. ACM*, 27(11), 1134–1142.

VAPNIK, V. N. (1996): *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA.

VENS, C., AND H. BLOCKEEL (2006): "A Simple Regression Based Heuristic for Learning Model Trees," *Intell. Data Anal.*, 10(3), 215–236.

VOSOUGHI, S., M. MOHSENVAND, AND D. ROY (2017): "Rumor Gauge: Predicting the Veracity of Rumors on Twitter," *ACM Transactions on Knowledge Discovery from Data*, 11(4), 50:1–36.

VUJIČ, S., AND X. ZHANG (2018): "Does Twitter chatter matter? Online reviews and box office revenues," *Applied Economics*, 50(34), 3702–3717.

WAN, A. T., X. ZHANG, AND G. ZOU (2010): "Least squares model averaging by Mallows criterion," *Journal of Econometrics*, 156(2), 277–283.

WANG, Y., AND I. H. WITTEN (1997): "Inducing Model Trees for Continuous Classes," in *In Proc. of the 9th European Conf. on Machine Learning Poster Papers*, pp. 128–137.

WHITE, H. (1982): "Maximum Likelihood estimation of Misspecified Models," *Econometrica*, 50(1), 817–838.

WHITTLE, P. (1960): "Bounds for the Moments of Linear and Quadratic Forms in Independent Variables," *Theory of Probability and Its Applications*, pp. 302–305.

WOLPERT, D. H., AND W. G. MACREADY (1997): "No free lunch theorems for optimization," *IEEE Transactions on Evolutionary Computation*, 1(1), 67–82.

XIE, T. (2015): "Prediction Model Averaging Estimator," *Economics Letters*, 131, 5–8.

——— (2017): "Heteroscedasticity-robust Model Screening: A Useful Toolkit for Model Averaging in Big Data Analytics," *Economics Letter*, 151, 119–122.

XIONG, G., AND S. BHARADWAJ (2014): "Prerelease Buzz Evolution Patterns and New Product Performance," *Marketing Science*, 33(3), 401–421.

YANG, S., M. M. HU, R. S. WINER, H. ASSAEL, AND X. CHEN (2012): "An Empirical Study of Word-of-Mouth Generation and Consumption," *Marketing Science*, 31(6), 952–963.

ZHANG, X., J.-M. CHIOU, AND Y. MA (2018): "Functional prediction through averaging estimated functional linear regression models," *Biometrika*, 105(4), 945–962.

ZHANG, X., D. YU, G. ZOU, AND H. LIANG (2016): "Optimal Model Averaging Estimation for Generalized Linear Models and Generalized Linear Mixed-Effects Models," *Journal of the American Statistical Association*, 111(516), 1775–1790.

ZHAO, S., X. ZHANG, AND Y. GAO (2017): "Model averaging with averaging covariance matrix," *Economics Letter*, 145, 214–217.

ZHOU, Y., L. ZHANG, AND Z. YI (2017): "Predicting movie box-office revenues using deep neural networks," *Neural Computing and Applications*, 31, 1855–1865.