NBER WORKING PAPER SERIES

THE BIGGER PICTURE: COMBINING ECONOMETRICS WITH ANALYTICS IMPROVE FORECASTS OF MOVIE SUCCESS

Steven F. Lehrer
Tian Xie

Working Paper 24755
http://www.nber.org/papers/w24755

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2018, Revised November 2020

We wish to thank Chris Hansen, seminar participants at the Young Econometricians around Pacific (YEAP) 2017 annual conference, the Canadian Econometrics Study Group (CESG) 2017 annual conference, Carleton University, Chinese Academy of Sciences, Northeastern University, Renmin University, Xiamen University, and Zhejiang University for helpful comments and suggestions. Xie's research is supported by the Natural Science Foundation of China (71701175), the Chinese Ministry of Education Project of Humanities and Social Sciences (17YJC790174), the Natural Science Foundation of Fujian Province of China (2018J01116), the Fundamental Research Funds for the Central Universities in China (20720171002, 20720171076, and 20720181050), and Educational and Scientific Research Program for Young and Middleaged Instructor of Fujian Province (JAS170018). Lehrer wishes to thank SSHRC for research support. The usual caveat applies. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

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

Steven F. Lehrer and Tian Xie

NBER Working Paper No. 24755
June 2018, Revised November 2020
JEL No. C52,C53,C55

ABSTRACT

There exists significant hype regarding how much machine learning and incorporating social media data can improve forecast accuracy in commercial applications. To assess if the hype is warranted, we use data from the film industry in simulation experiments that contrast econometric approaches with tools from the predictive analytics literature. Further, we propose new strategies that combine elements from each literature in a bid to capture richer patterns of heterogeneity in the underlying relationship governing revenue. Our results demonstrate the importance of social media data and value from hybrid strategies that combine econometrics and machine learning when conducting forecasts with new big data sources. Specifically, while both least squares support vector regression and recursive partitioning strategies greatly outperform dimension reduction strategies and traditional econometrics approaches in fore-cast accuracy, there are further significant gains from using hybrid approaches. Further, Monte Carlo experiments demonstrate that these benefits arise from the significant heterogeneity in how social media measures and other film characteristics influence box office outcomes.

Steven F. Lehrer
Dunning Hall 336, 94 University Avenue
Department of Economics
Queen's University
Kingston, Ontario
Canada
K7L 3N6
and NBER
lehrers@queensu.ca

Tian Xie
Rm.424
College of Business
Shanghai University of Finance and Economics
Shanghai, China, 200433
xietian001@hotmail.com

An online appendix is available at: http://www.nber.org/data-appendix/w24755

1 Introduction

Many speculate that in the near future, movie studios will find that predictive analytics may play just as large of a role as either the producer, director, and/or stars of the film when determining if it will be a success. Currently, predictive analytics that incorporate social media data are being predominately used for demand forecasting exercises in the film industry. Improved forecasts are valuable since they could increase capital investments by reducing investor uncertainty of the box office consequences and also help marketing teams tailor effective advertising campaigns. However, there remains skepticism as to whether social media data truly adds value to forecasting exercises.

While prior work by Bollen, Mao, and Zheng (2011), Goh, Heng, and Lin (2013) and Lehrer and Xie (2017), among others, present evidence of the value of social media in different contexts, the authors did not consider traditional off the shelf machine learning approaches such as regression trees, random forest, boosting, and support vector regression. These statistical learning algorithms do not specify a structure for the model to forecast the mean and often achieve predictive gains by allowing for nonlinear predictor interactions that are missed by conventional econometric approaches. Despite this benefit in modeling, the algorithms used to either construct hyperplanes or build tree based structures via recursive partitioning implicitly assumes homogeneous variance across the entire explanatory-variable space.[1]

Heteroskedasticity of data which may arise from neglected parameter heterogeneity can impact the predictive ability of many forecasting strategies. For example, the presence of heteroskedasticity can change the location of support vectors and how the data is partitioned, thereby influencing the structure of regression trees.[2] In this paper, we intro-

[1]More generally, each of OLS, regression trees, and Lasso methods rely on the unweighted sum of squares criterion (SSR), which implicitly assumes homoskedastic errors. It is well known that when this condition is violated and heteroskedasticity is present, the standard errors are biased influencing statistical inference procedures. Further, the objective function ensures that areas of high variability will contribute more to minimizing the unweighted SSR, and will therefore play a larger role when making predictions at the mean. As such, predictions for low-variance areas are expected to be less accurate relative to high variance areas. Therefore, heteroskedasticity might affect predictions at the mean, since the implicit weights to the data are determined by the local variance. Recent developments continue to use the SSR as a loss function but can generally accommodate richer forms of heterogeneity relative to parametric econometric models by accounting for limited forms of parameter heterogeneity.

[2]After all, the symmetrical loss function of support vector regression equally penalizes high and low misestimates and which observations constitute as being a support vector of the best fitting hyperplane are

duce new strategies for predictive analytics that are contrasted with existing tools from both the econometrics and machine learning literature to provide guidance on how to improve forecast accuracy in applications within the film industry. Thus, we contribute to a burgeoning literature in the emerging fields of data science and analytics that focuses on developing methods to improve empirical practice including forecast accuracy. For example, among other developments, Vasilios, Theophilos, and Periklis (2015) examine the accuracy of machine learning techniques when forecasting daily and monthly exchange rates, Wager and Athey (2018) propose variants of random forests to estimate causal effects, and Ban, Karoui, and Lim (2018) adopt machine learning methods for portfolio optimization.

Motivating our new hybrid strategies is that heteroskedasticity would be anticipated in many forecasting exercises that involve social media data for at least two reasons. First, the attributes of individuals attracted to different films will differ sharply, leading the data to appear as if coming from different distributions. Second, online respondents may have greater unobserved variability in their opinions of different films.[3]

Our proposed hybrid strategy considers heterogeneity that arises from heteroskedastic data with both least squares support vector regression and recursive partitioning methods. To illustrate, forecasts from regression trees traditionally use a local constant model that assumes homogeneity in outcomes within individual terminal leaves. Our hybrid approach allows for model uncertainty and undertakes model averaging within each terminal leaf subgroup. Thus, within each leaf subgroup the possibility of a heterogeneous relationship between the explanatory variables and the outcome being forecasted is considered. Recently, Pratola, Chipman, George, and McCulloch (2020) consider incorporating heteroskedasticity in the machine learning literature within a Bayesian framework. With support vector regression we also allow for model uncertainty and modify the criterion function to be based on a heteroskedastic error term. Using Monte Carlo exer-

influenced by heteroskedasticity since the data would indicate that the prediction errors differ for different ranges of the predicted value.

[3]In other words, if this unobserved variability in opinions is not modeled, heteroskedasticity may arise from neglected parameter heterogeneity; which is a form of an omitted variables problem. This link between neglected parameter heterogeneity and heteroskedasticity is not well known among practitioners but can be explained with the following example. If regression coefficients vary across films (perhaps the role of Twitter volume on box office revenue differs for a blockbuster action film relative to an art house drama), then the variance of the error term varies too for a fixed-coefficient model.

cises and an empirical application that focuses on measures of predictive accuracy, we provide researchers guidance on when to use this hybrid strategy with either recursive partitioning strategies or least squares support vector regression relative to the approach developed in Pratola, Chipman, George, and McCulloch (2020).

Our empirical examination of the predictive accuracy of alternative empirical strategies that forecast revenue for the film industry does not impose any sampling criteria and considers every movie released either in theatres or the retail environment over a three-year period. This data exhibits strong heteroskedasticity,[4] which likely arises since different films appeal to populations drawn from different distributions.

Our results first provide new insights on the trade-offs researchers face when choosing a forecasting method. With smaller sample sizes, we find improved performance benefits from using least squares support vector regression relative to other machine learning approaches. Recursive partitioning strategies including regression trees, bagging and random forests yield on average 30-40% gains in forecast accuracy relative to econometric approaches that either use a model selection criteria or model averaging approach. These large gains from statistical learning methods even relative to econometric estimators and penalization methods that implicitly account for heteroskedastic data, demonstrate the restrictiveness of linear parametric econometric models. These models remain popular in econometrics since as Manski (2004) writes "statisticians studying estimation have long made progress by restricting attention to tractable classes of estimators; for example, linear unbiased or asymptotic normal ones".

Second, our analysis uncovers additional gains of roughly 10% in forecast accuracy from our proposed strategy that allows for model uncertainty. These gains are exhibited across a variety of machine learning algorithms with i) alternative kernel functions for support vector regression and ii) both alternative hyperparameters and local objective functions to partition the data within a tree structure including random forest, bagging, M5', and least squares support vector regression. Monte Carlo experiments clarify why

[4]Results from Breusch-Pagan test are presented in appendix F.1 and sampling restrictions such as those in Lehrer and Xie (2017) may sidestep heteroskedasticity by reducing the heterogeneity in the data by only including films with similar budgets. Subsection F.13 in the appendix illustrates the improved forecasting accuracy of the new hybrid estimators proposed as well as random forest and bagging strategies relative to the estimators contrasted in Lehrer and Xie (2017).

these gains arise in our empirical application. We find that hybrid strategies are quite useful in settings where heteroskedasticity arises due to significant parameter heterogeneity, perhaps due to jumps or threshold effects, or simply neglected parameter heterogeneity in the underlying behavioral relationships. In this setting, hybrid strategies can explain a portion of the significant amount of heterogeneity in outcomes within each tree leaf.

Third, we find that there is tremendous value from incorporating social media data in forecasting exercises. Econometric tests find that including social media data leads to large gains in forecast accuracy. Variable importance calculations from machine learning methods show that measures of social media message volume account for up to 7 of the 10 most influential variables when forecasting either box office or retail movie unit sales revenue.

This paper is organized as follows. In the next section, we briefly review traditional econometric and machine learning strategies to conduct forecasting. We then introduce two computationally efficient strategies to aid managerial decision making by accommodating more general forms of heterogeneity than traditional methods. A discussion of Monte Carlo experiments in section 3 elucidates why an understanding of the source of heteroskedasticity is useful when selecting forecasting methods. The data used and design of the simulation experiments that compares forecasting methods is presented in section 4. Section 5 presents and discusses our findings that show the value of social media data and combining machine learning with econometrics when undertaking forecasts. We conclude in the final section.

2 Empirical Tools for Forecasting

Forecasting involves a choice of a method to identify the underlying factors that might influence the variable (y) being predicted. Econometric approaches begin by considering a linear parametric form for the data generating process (DGP) of this variable as

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

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.[5] The error term can be heteroskedastic, where $\sigma_i^2 = \mathbb{E}(e_i^2|x_i)$ denote the conditional variance that depends on x_i. Since the DGP in equation (1) is unknown, econometricians often approximate it with a set of M candidate models:

$$y_i = \sum_{j=1}^{k^{(m)}} \beta_j^{(m)} x_{ij}^{(m)} + e_i^{(m)}, \tag{2}$$

for $m = 1, ..., M$, where $x_{ij}^{(m)}$ for $j = 1, ..., k^{(m)}$ denotes the regressors, $\beta_j^{(m)}$ denotes the coefficients. The error $e_i^{(m)}$ now contains both the original error term (e_i) and a modeling bias term denoted as $b_i^{(m)} \equiv \mu_i - \sum_{j=1}^{k^{(m)}} \beta_j^{(m)} x_{ij}^{(m)}$.

In practice, researchers have a set of plausible models, and do not know with certainty which model is correct, or the best approximation for the task at hand. The traditional solution is empirical model selection, which provides an evidence-based rule (e.g. Akaike information criterion) for selecting one model from a set of feasible models. Rather than selecting one model among a set of M linear candidate models, empirical model averaging approaches allow the researcher to remain uncertain about the appropriate model specification and take a weighted average of results across the set of plausible models to approximate the DGP in equation (1).[6]

In the context of model averaging, the critical question is how to select the weights for each candidate model. Formally, assume that the M candidate models that approximate the DGP are given by $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{e}$, where $\boldsymbol{y} = [y_1, ..., y_n]^\top$, $\boldsymbol{\mu} = [\mu_1, ..., \mu_n]^\top$ and $\boldsymbol{e} = [e_1, ..., e_n]^\top$. We define the variable $\boldsymbol{w} = [w_1, w_2, ..., w_M]^\top$ as a weight vector in the unit simplex in $\mathbb{R}^M$,

$$\mathcal{H} \equiv \left\{ w_m \in [0,1]^M : \sum_{m=1}^{M} w_m = 1 \right\}. \tag{3}$$

Numerous optimization routines have been developed by econometricians to estimate these weights and each routine aims to strike a balance between model performance and

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

[6] That is, define the estimator of the m^{th} candidate model as $\hat{\boldsymbol{\mu}}^{(m)} = \boldsymbol{X}^{(m)}(\boldsymbol{X}^{(m)^\top}\boldsymbol{X}^{(m)})^{-1}\boldsymbol{X}^{(m)^\top}\boldsymbol{y} = \boldsymbol{P}^{(m)}\boldsymbol{y}$, where $\boldsymbol{X}^{(m)}$ is a full rank $n \times k^{(m)}$ matrix of independent variables with $(i,j)^{th}$ element being $x_{ij}^{(m)}$ and $\boldsymbol{P}^{(m)} = \boldsymbol{X}^{(m)}(\boldsymbol{X}^{(m)^\top}\boldsymbol{X}^{(m)})^{-1}\boldsymbol{X}^{(m)^\top}$. 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. See Steel (2019) for a recent survey of the model averaging literature.

complexity of the individual models. Once the optimal weights (w_m) are obtained, the forecast from the model averaging estimator of $\boldsymbol{\mu}$ is

$$\hat{\boldsymbol{\mu}}(\boldsymbol{w}) = \sum_{m=1}^{M} w_m \hat{\boldsymbol{\mu}}^{(m)} = \sum_{m=1}^{M} w_m \boldsymbol{P}^{(m)} \boldsymbol{y} = \boldsymbol{P}(\boldsymbol{w}) \boldsymbol{y}. \tag{4}$$

This forecast is a weighted average of the forecasts of the individual candidate models, which is why model averaging can equivalently be described as forecast combination.

Data mining techniques developed within the machine learning literature can also be used for forecasting. Unlike many econometric approaches that begin by assuming a linear parametric form to explain the DGP, supervised learning algorithms do not ex-ante specify a structure for the model to forecast the mean and build a statistical model to make forecasts by selecting which explanatory variables to include. For example, decision trees create a form of a top-down, flowchart-like model that recursively partitions a heterogeneous data set into relatively homogeneous subgroups in order to make more accurate predictions on future observations. Each partition of the data is called a "node", with the top node called the "root" and the terminal nodes called "leaves".

There are many algorithms to build decision trees but one of the oldest for continuous outcome variables is known as the regression tree (RT) approach developed by Breiman, Friedman, and Stone (1984). RT uses a fast divide and conquer greedy algorithm to recursively partition the data. Formally, at node τ containing n_τ observations with mean outcome $\bar{y}(\tau)$ of the tree can only be split by one selected explanatory variable into two leaves, denoted as τ_L and τ_R. The split is made at the variable where $\Delta \equiv \text{SSR}(\tau) - \text{SSR}(\tau_L) - \text{SSR}(\tau_R)$, reaches its global maximum;[7] where the within-node sum of squares is $\text{SSR}(\tau) = \sum_i^{n_\tau} (y_i - \bar{y}_\tau)^2$. This splitting process continues at each new node until the $\bar{y}(\tau)$ at nodes can no longer be split since it will not add any additional value to the prediction. Forecasts at each final leaf l are the fitted value from a local con-

[7] Intuitively, this procedure may appear to operate like forward stepwise regression where at each step, the procedure adds an independent variable based on the reduction in the sum of squares error caused by the action in the full sample until a stopping criterion is met. However, variables are added in a more flexible manner with regression trees. For continuous covariates, an equally-spaced grid covering the range of possible values is usually considered, thereby allowing for highly nonlinear models with potentially complex interactions within the subsamples by node following each split. Implicitly it is assumed that there are no unobservables relevant to the estimation.

stant regression model

$$y_i = a + e_i^*, \quad i \in l, \tag{5}$$

where e_i^* is the error term and a stands for a constant term. The least square estimate of $\hat{a} = \bar{y}_{i \in l}$. In other words, after partitioning the dataset into numerous final leaf nodes, this approach approximates the DGP with a series of discontinuous flat surfaces forming an overall rough shape. Further, the forecast assumes any heterogeneity in outcomes within each subgroup is random, which can appear unsatisfying from the perspective of the econometrician.

The statistical learning literature has noted both this drawback in how forecasts are made,[8] along with drawbacks in how splits within the tree are made, leading to further refinements. Hastie, Tibshirani, and Friedman (2009) discuss enhancements including ensemble methods that combine estimates from multiple models or trees to reduce the variance of predictions from individual regression trees. For example, bootstrap aggregating decision trees (a.k.a. bagging) proposed in Breiman (1996) and random forest developed in Breiman (2001) are randomization-based ensemble methods that draw a parallel to model averaging.[9] In bagging, trees are built on random bootstrap copies of the original data, producing multiple different trees. Bagging differs from random forest only in the set of explanatory factors being considered in each tree. To determine the best split at each node of the tree, random forests only consider a random subset of the predictor variables rather than the full set used with bagging. With both strategies, the final forecast is obtained as an equal weight average of the individual tree forecasts.

Studies within the statistical learning literature (see e.g. Loh and Shih, 1997; Kim and Loh, 2003; Hothorn, Hornik, and Zeileis, 2006) have concluded that the split selection process is biased towards selecting variables with many split points. This critique appears imprecise and we argue that any split to minimize Δ with heteroskedastic data will be

[8] For example, algorithms e.g. Chaudhuri, Huang, Loh, and Yao (1994) use weighted polynomial smoothing techniques to smooth forecasts between leaf nodes.

[9] Since individual trees are constructed sequentially, very small perturbations in the sample can lead to a different tree structure used for forecasting. The main idea of ensemble methods is to introduce random perturbations into the learning procedure by growing multiple different decision trees from a single learning set and then an aggregation technique is used to combine the predictions from all these trees. These perturbations help remedy the fact that a single tree may suffer from high variance and display poor forecast accuracy. See appendix A for more details.

biased to regions of variables with high heteroskedasticity, since they will contain more split points relative to regions of low heteroskedasticity. Thus, heteroskedastic data may lead to not choosing the "correct" first split of the root node and could subsequently lead the tree to follow a suboptimal path.[10]

To summarize, forecasts from recursive partitioning and model averaging methods are computationally expensive but differ in three important way. First, how the DGP in equation (1) is approximated differs and both bagging and random forest do not make any assumptions about the probabilistic structure of the data. Second. optimal weights across models are calculated using equation (3) from predictions using the full sample in model averaging strategies. The weight of each leaf in the tree forecast is simply determined by the sample proportion in each leaf. Third, final predictions from a regression tree rule out heterogeneity and any model uncertainty in each final leaf $\bar{y}(\tau)$ of the tree.

This lack of heterogeneity and computational considerations motivate our two proposed extensions for forecasting with social media data. The next subsection proposes an improved method to select candidate models for model averaging estimators. The subsection that follows proposes a hybrid strategy that combines model averaging with both a recursive partitioning algorithm and least squares support vector regression. With the former hybrid approach, heterogeneity is considered when making predictions in each tree leaf.

The presence of heteroskedasticity cannot be combated by taking a log-transformation on the outcome variable. Silva and Tenreyro (2006) point out that such a nonlinear transformation of the dependent variable will generate biased and inconsistent OLS estimates since the transformation changes the properties of the heteroskedastic error term creating correlation with the covariates. Similarly, this transformation will also influence where splits occur with recursive partitioning algorithms, thereby generating different subgroups. Initial splits would continue to be biased in regions of high heteroskedasticity,

[10] In the statistical learning literature, the critique that minimizing Δ to determine splits by the greedy approach of Breiman, Friedman, and Stone (1984) leads to choosing locations of local, rather than global optimality with each split is disscussed in Murthy, Kasif, and Salzberg (1994); Brodley and Utgoff (1995); Fan and Gray (2005); and Gray and Fan (2008). Subsequent work to build trees involve new algorithms that search for the best combination of splits one to two more levels deeper before selecting a split rule. These more global algorithms involve larger computational costs since they need to look several steps ahead in the tree.

which is likely regions containing more low revenue films due to the log transformation.

Last, not all algorithms developed in the statistical learning literature that approximate the DGP involve the construction of tree structures. As discussed in appendix section B.5, support vector regression (SVR) solves a convex quadratic programming problem to obtain a best fitting hyperplane that minimizes the distance between the actual and predicted outcome variable within a predefined or threshold error value. The vector points closest to the hyperplane are known as the support vector points and are the only observations that contribute to the forecast from the algorithm. Since SVR is computationally challenging, Suykens and Vandewalle (1999) proposed least squares support vector regression (SVR_{LS}) that modifies the optimization problem to find the hyperplace within the threshold error values by solving a set of linear equations under a squared loss function. A portion of the objective function of SVR_{LS} contains the SSR where homoskedasticity is assumed and as such is both subject to the critique motivating the study.

Applications of SVR and SVR_{LS} are common in the engineering and computer science communities since they show high degrees of forecast accuracy, particularly in settings with a low ratio of sample size to covariates. Despite their strong performance in settings common to many business applications, these algorithms are intermittently used in practice since the output from these algorithms is difficult to interpret, thereby presenting challenges when communicating results to a layperson audience. This challenge arises in part since the SVR and SVR_{LS} algorithms involve converting the original mapping of the data into a higher dimensional Hilbert space.

2.1 A New Strategy for Model Screening

The empirical performance of any model averaging estimator crucially depends on the candidate model set. Yet, a potential drawback of constructing a candidate model set by considering the full permutation of all regressors is that the total number of candidate models increases exponentially with the number of regressors. To narrow down the list of candidate models, a screening step can be undertaken. As shown in Wan, Zhang, and Zou (2010), Xie (2015), Zhang, Zou, and Carroll (2015), among others, by either keeping the total number of candidate models small or letting the total number of candidate

models converge to infinity slow enough, provides a necessary condition to maintain the asymptotic optimality of model averaging estimators.[11] We follow the insights in Zhang, Yu, Zou, and Liang (2016) who established the asymptotic optimality of Kullback-Leibler (KL) loss based model averaging estimators with screened candidate models. We define $\mathcal{M}$ and $\tilde{\mathcal{M}}$ to respectively be the candidate model set prior to, and following model screening; in which $\tilde{\mathcal{M}} \subseteq \mathcal{M}$. The weight vector space solved via an optimization routine under $\tilde{\mathcal{M}}$ can be written as

$$\tilde{\mathcal{H}} = \left\{ \boldsymbol{w} \in [0,1]^M : \sum_{m \in \tilde{\mathcal{M}}} w_m = 1 \text{ and } \sum_{m \notin \tilde{\mathcal{M}}} w_m = 0 \right\}. \tag{6}$$

Note that the resultant weight vector, denoted as $\tilde{\boldsymbol{w}}$, under $\tilde{\mathcal{M}}$ is still $M \times 1$, however, models that do not belong in $\tilde{\mathcal{M}}$ are assigned zero weight.

We define the average squared loss as $L(\boldsymbol{w}) = (\hat{\boldsymbol{\mu}}(\boldsymbol{w}) - \boldsymbol{\mu})^\top (\hat{\boldsymbol{\mu}}(\boldsymbol{w}) - \boldsymbol{\mu})$ where $\hat{\boldsymbol{\mu}}(\boldsymbol{w})$ is defined in equation (A10). We present the following set of assumptions

Assumption 1 *We assume that there exists a non-negative series of v_n and a weight series of $\boldsymbol{w}_n \in \mathcal{H}$ such that*

(i) $v_n \equiv L(\boldsymbol{w}_n) - \inf_{\boldsymbol{w} \in \mathcal{H}} L(\boldsymbol{w})$,

(ii) $\xi_n^{-1} v_n \to 0$,

(iii) $Pr(\boldsymbol{w}_n \in \tilde{\mathcal{H}}) \to 1$ *as* $n \to \infty$,

where $\tilde{\mathcal{H}}$ is defined in (6) and ξ_n is the (lowest) modified model risk defined in equation (A28).

Assumption 1(i) defines v_n to be the distance between a model risk given by $\boldsymbol{w}_n$ and the lowest possible model risk. Assumption 1(ii) is a convergence condition. It requires that ξ_n goes to infinity faster than v_n. The final item of assumption 1 implies the validity of our selected model screening techniques. When the sample size goes to infinity, the chance that the model screening technique accidentally omits at least one useful model

[11] Moreover, Hansen (2014) and Zhang, Ullah, and Zhao (2016) point out that to satisfy the conditions on the global dominance of averaging estimators over the unrestricted least-squares estimator, the number of candidate models should be limited by screening. Screening ensures that not every possible model is estimated.

goes to 0. This condition is easily satisfied by imposing mild screening conditions, while keeping the candidate models in $\tilde{\mathcal{M}}$ to be as many as allowed.

The following theorem establishes the asymptotic optimality of Mallows-type model averaging estimators under a screened model set.

Theorem 1 *Let Assumption 1 be satisfied, then under the conditions that sustain the asymptotic optimality of Mallows-type model averaging estimators under given (unscreened) candidate model set, as $n \to \infty$, we have*

$$\frac{L(\tilde{\boldsymbol{w}})}{\inf_{\boldsymbol{w}\in\mathcal{H}} L(\boldsymbol{w})} \xrightarrow{p} 1, \tag{7}$$

The proof appears in appendix D.7. Theorem 1 states that using screened model set $\tilde{\mathcal{M}}$, the model averaging estimator $\tilde{\boldsymbol{w}}$ is asymptotically optimal in the sense of achieving the lowest possible mean squared error (model risk); even compared to a model averaging estimator that used all potential candidate models in its set.

2.2 New Hybrid Approaches: Model Averaging Learning Methods

In an influential paper, Belloni and Chernozhukov (2013) suggest applying the OLS estimator after variable selection by the Lasso, thereby introducing a two-step hybrid machine learning and econometrics estimator.[12] In this paper, we propose using recursive partitioning algorithms in the first step to build RT structures and then apply econometric estimators that allow for model uncertainty in place of equation (5) when forecasting. We denote this procedure as model averaging regression tree (MART), which is the building block of many of the proposed hybrid approaches. MART generalizes linear regression trees that have been shown to yield improvements over the local constant model in equation (5), by allowing multiple regression models to explain outcomes within each leaf.

Formally, following the recursive partitioning procedure, at each tree leaf there may be a sequence of $m = 1, ..., M$ linear candidate models, in which regressors of each model

[12]Penalization methods such as the Lasso have objective functions designed to reduce the dimensionality of explanatory variables. The post Lasso strategy can be viewed as a model screening method since it limits the number of explanatory variables and hence dimensionality of the candidate models. Lehrer and Xie (2017) extend this idea and proposed using model averaging in place of the OLS estimator in the second step. The set of candidate models considered in that step are restricted to those constructed with variables selected by the first step Lasso. See appendix D.6 for further details on these strategies and all Lasso estimators considered.

m is a subset of the regressors belonging to that tree leaf. The regressors $\boldsymbol{X}_{i\in l}^{(m)}$ for each candidate model within each tree leaf is constructed such that the number of regressors $k_l^{(m)} \ll n_l$ for all m. Using these candidate models, model averaging obtains

$$\underset{(K\times 1)}{\hat{\boldsymbol{\beta}}_l(\boldsymbol{w})} = \sum_{m=1}^{M} w^{(m)} \underset{(K\times 1)}{\tilde{\boldsymbol{\beta}}_l^{(m)}}, \tag{8}$$

which is a weighted averaged of the "stretched" estimated coefficient $\tilde{\boldsymbol{\beta}}_l^{(m)}$ for each candidate model m. Note that the $K \times 1$ sparse coefficient $\tilde{\boldsymbol{\beta}}_l^{(m)}$ is constructed from the $k_l^{(m)} \times 1$ least squares coefficient $\hat{\boldsymbol{\beta}}_l^{(m)}$ by filling the extra $K - k_l^{(m)}$ elements with 0s.

To implement this strategy, the predicting observations $\boldsymbol{X}_t^p$ with $t = 1, 2, ..., T$ are dropped down the regression tree. For each $\boldsymbol{X}_t^p$, after several steps of recursive partitioning, we end up with one particular tree leaf l. We denote the predicting observations in tree leaf l as $\boldsymbol{X}_{t\in l}^p$. The forecast for all observations can then be obtained as

$$\hat{\boldsymbol{y}}_{t\in l} = \boldsymbol{X}_{t\in l}^p \hat{\boldsymbol{\beta}}_l(\boldsymbol{w}). \tag{9}$$

This strategy preserves the original recursive partitioning process and within each leaf allows observations that differ in characteristics to generate different forecasts $\hat{\boldsymbol{y}}_{t\in l}$.

Model averaging bagging (MAB) applies this process to each of the B samples used to construct a bagging tree. The final MAB forecast remains the equal weight average of the B model averaged tree forecasts. Model averaging random forest (MARF) operates similarly with the exception that only k predictors out of the total K predictors are considered for the split at each node. The candidate model set for each leaf is constructed with the k regressors used to split the nodes that generated this leaf l, whereas each of the K regressors are considered with MAB.[13] This restriction on the number of predictors also affects how $\hat{\boldsymbol{\beta}}_l(\boldsymbol{w})$ is calculated since it is averaged only over those leafs in the forest where it was randomly selected. The intuition of this hybrid strategy can be applied to almost any machine learning algorithm including ones with a different objective function

[13] If the full sample contains n observations, the tree leaf l contains a subset $n_l < n$ of the full sample of y, denoted as y_i with $i \in l$. Also, the sum of all n_l for each tree leaf equals n. The mean of $y_{i\in l}$ is calculated, denoted as $\bar{y}_{i\in l}$. The value $\bar{y}_{i\in l}$ is the forecast of $\boldsymbol{X}_{t\in l}^p$. It is possible that different predicting observations $\boldsymbol{X}_t^p$ and $\boldsymbol{X}_s^p$ with $t \neq s$ will end up with the same tree leaf, therefore, generating identical forecasts.

to determine splits within a tree such as M5.

A hybrid strategy that mimics the model averaging estimator described earlier is also possible with statistical learning strategies that generate hyperplanes such as SVR_{LS}. This hybrid approach estimates each candidate model by SVR_{LS}. Next, inspired by Ullah and Wang (2013), we define a criteria function for a model averaging SVR_{LS} (MASVR_{LS}) strategy that estimates the model averaging weights

$$C_n(\boldsymbol{w}) = \sum_{i=1}^{n} \hat{e}_i^2(\boldsymbol{w}) + 2\sum_{i=1}^{n} (\hat{e}_i^2(\boldsymbol{w}))^2 p_{ii}(\boldsymbol{w}). \tag{10}$$

An important feature of this criteria function is that it directly considers heteroskedasticity since $\hat{e}_i(\boldsymbol{w})$ is the averaged SVR_{LS} residual and $p_{ii}(\boldsymbol{w})$ is the i^{th} diagonal element of the averaged projection matrix that is similar to how $\boldsymbol{P}(\boldsymbol{w})$ was defined in equation (4). Further details of this approach are provided in appendix B.7, which also introduces a criteria function for MASVR_{LS} with a homoskedastic error term that would offer computational benefits relative to equation (10). That said, both criteria functions for MASVR_{LS} face the same limitations as SVR_{LS} in regards to both interpretation of the output and performance in terms of computational speed in a setting with many observations.

2.3 A Simple Illustration

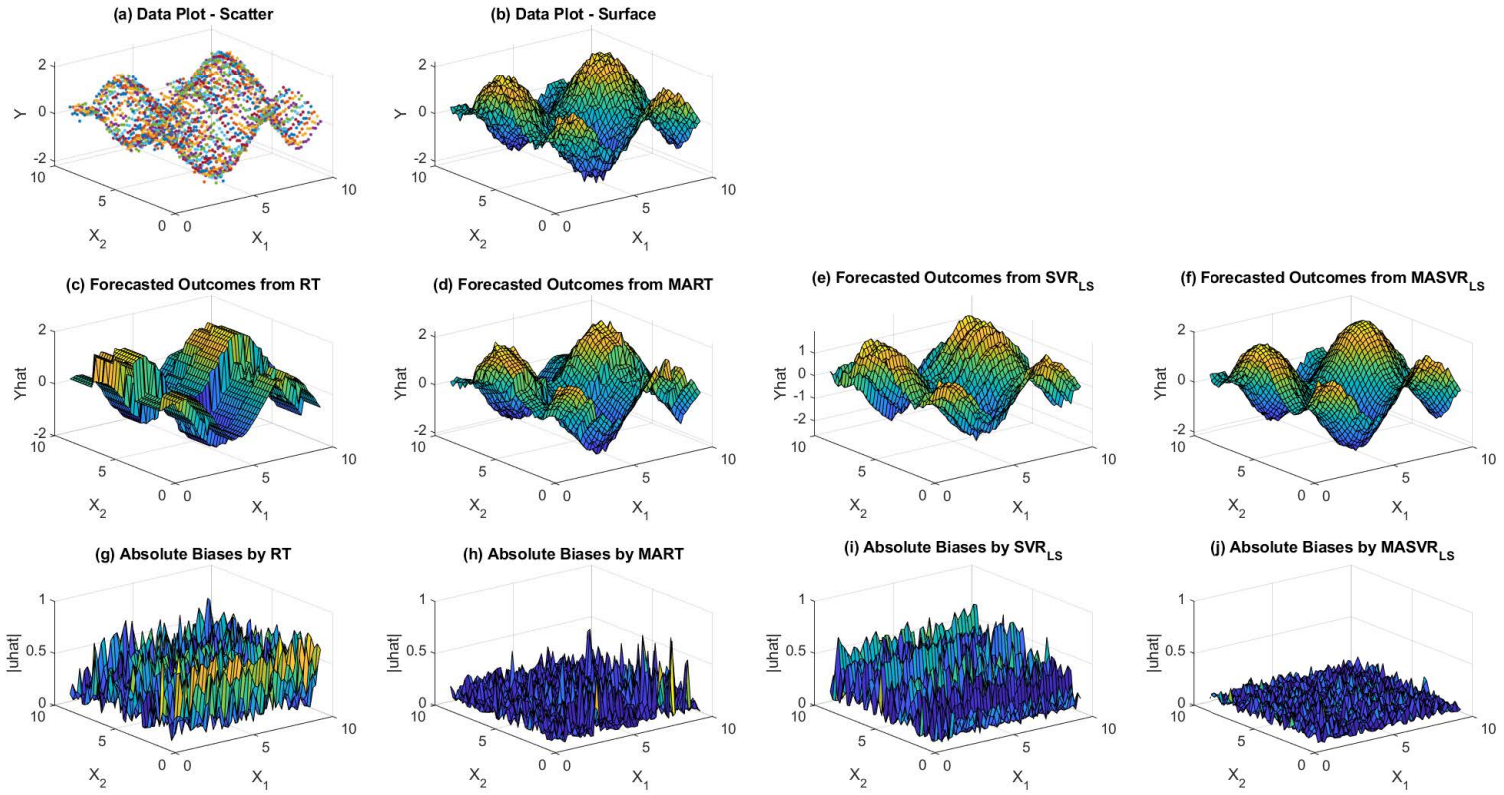
To illustrate the benefits of allowing for heterogeneity due to model uncertainty via the proposed MART and MASVR_{LS} hybrid procedures, we simulate data drawn from a nonlinear process. Panels (a) and (b) of figure 1 respectively present the scatter plot and surface plot of training data generated by

$$y_i = \sin(X_{1i}) + \cos(X_{2i}) + e_i,$$

where $X_{1i} \in [1, 10]$, $X_{2i} \in [1, 10]$, and e_i is a Gaussian noise with mean 0 and variance 0.01.

Forecasts of y calculated from RT, MART, SVR_{LS} and MASVR_{LS} with the training data are presented in panels (c) to (f) of figure 1, respectively. Since RT forecasts assume homogeneity within leaves, the surface plot in panel (c) appears similar to a step-function.

Figure 1: Simulation Evidence Illustrating the Gains of the Hybrid Approach That Combines Model Averaging with Regression Trees



Note: Plot (a) presents a scatter plot of the simulated data, plot (b) is the corresponding surface plot, plots (c)-(f) display the forecasted shape by RT, MART, SVR_{LS}, and $MASVR_{LS}$, and plots (g) and (j) present the absolute value of forecast errors against the two explanatory variables for each forecasting strategy, respectively.

In contrast, by allowing for heterogeneity in the forecasts within each leaf, the surface plot from MART in panel (d) more closely mimics the variation in the joint distribution in the underlying data. The SVR_{LS} forecast shape in panel (e) looks similar to the MART forecast with what appears to be sharper folds, whereas the MASVR_{LS} forecast in panel (f) appears to have the smoothest surface plot of the forecasted outcome.

Panels (g) through (j) of figure 1 respectively plot the forecast errors from RT to MASVR_{LS} against both X_1 and X_2. Comparing the height of these figures shows that the absolute biases from MART and MASVR_{LS} are respectively less than half of the biases obtained from RT and SVR_{LS}. The reduced height occurs throughout the space spanned by X_1 and X_2 demonstrating that gains are achieved by allowing for richer relationships to capture parameter heterogeneity either in each tree leaf or support vector. In the next section, a Monte Carlo study provides further insights on when the hybrid procedures that allow for model uncertainty improve forecasts relative to traditional strategies developed in the statistical learning literature.

3 Monte Carlo Study

Similar to Liu and Okui (2013), we consider the following DGP

$$y_i = \mu_i + e_i = \sum_{j=1}^{\infty} (\beta_j + r \cdot \sigma_i) x_{ji} + e_i \tag{11}$$

for $t = 1, ..., n$. The coefficients are generated by $\beta_j = cj^{-1}$, where c is a parameter that we control, such that $R^2 = c^2/(1 + c^2)$ that varies in $\{0.1, ..., 0.9\}$. The parameter σ_t is drawn from a $N(0, 1)$ and the scale variable r introduces potential heterogeneity to the model. We set $x_{1i} = 1$ and the other x_{ji}s follow $\text{N}(0, 1)$. Since the infinite series of x_{ji} is infeasible in practice, we truncate the process at $j_{\max} = 10,000$ without violating our assumption on the model set-up.[14] We assume that the full set of 10,000 x_{ji}s is not entirely feasible. Two scenarios that represent random heteroskedasticity and heteroskedasticity that arises due

[14]That is, variables with close-to-0 coefficients (i.e. x_{ji} with $j > j_{\max}$) can be ignored since they barely influence the dependent variable. This simulation design aims to mimic a big data environment, where the number of covariates is large. Last, all results are robust to alternative values of the scale variable r.

to neglected parameter heterogeneity are considered. Formally,

A. Random Heteroskedasticity: we set the parameter $r = 0$, eliminating heterogeneity and pure random heteroskedasticity is created by drawing $e_i \sim N(0, x_{2i}^2)$.

B. Parameter Heterogeneity: heterogeneity in β for each observation is created by setting $r = 1/5$ and drawing $e_i \sim N(0, 1)$.

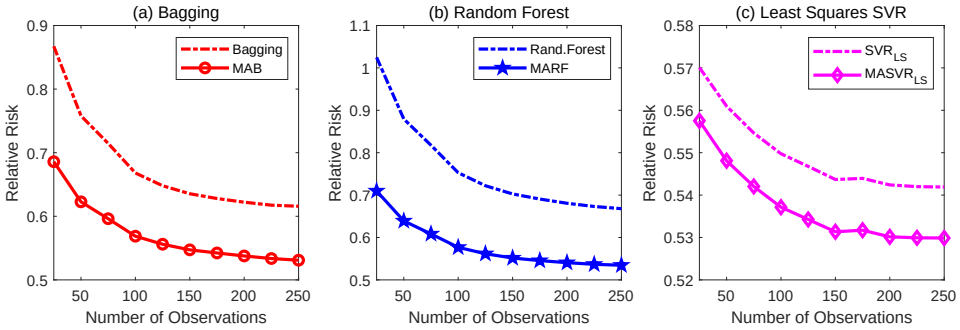
With this DGP, we compare the performance of conventional learning methods and model averaging learning methods using their risks.[15] Panels A and B of figure 2 respectively present results for the random heteroskedasticity and the parameter heterogeneity scenario. In each figure, the number of observations is presented on the horizontal axis, the relative risk is displayed on the vertical axis and dash-dotted (solid) lines respectively represent the machine learning strategy and (the hybrid extension). The results indicate that: i) the model averaging learning method performs better than their respective conventional learning method in all values of n; ii) as sample sizes increase, all methods tend to yield smaller risks; and iii) MASVR_{LS} has the best relative performance in all cases, particularly when sample sizes are small. Overall, we observe smaller relative risks in the parameter heterogeneity scenario.

Since the results in figure 2 panel A are relative to OLS estimates of a generalized unrestricted model (henceforth GUM) that utilizes all the independent variables, the panels of figure 3 present the absolute risks for each model averaging learning methods along with the risks of the GUM under random heteroskedasticity and parameter heterogeneity. In each figure, MAB, MARF, MASVR_{LS} and GUM are presented by circle-, star-, diamond-, and solid lines, respectively. The ranking of the methods is identical and GUM yields significantly higher risks in the parameter heterogeneity scenario. This suggests that conventional regressions suffer from efficiency loss in the presence of effect heterogeneity. Yet

[15] Specifically, $\text{Risk}_i \equiv \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_i^L - \mu_i\right)^2$, where μ_i is the true fitted value (feasible in simulation) and $\hat{\mu}_i^L$ is the fitted value obtained by a specific learning method for for $L =$ Regression Tree, Bagging, MAB, Random Forest, MARF, SVR_{LS} and MASVR_{LS}. For each sample size, we compute the risk for all methods and average across 100,000 simulation draws. For bagging and random forest, we set the total number of bootstraps as $B = 20$. For random forest, we randomly draw 2 regressors out of 5 to split each node. The same settings apply to the model averaging learning methods. For all model averaging learning methods, the candidate model set for each leaf contains all feasible combinations of the regressors. To ease interpretation, we normalize all risks by the risk from OLS estimates of the generalized unrestricted model.

Figure 2: Relative Performance of Conventional and Model Averaging Learning

A. Random Heteroskedasticity



B. Parameter Heterogeneity

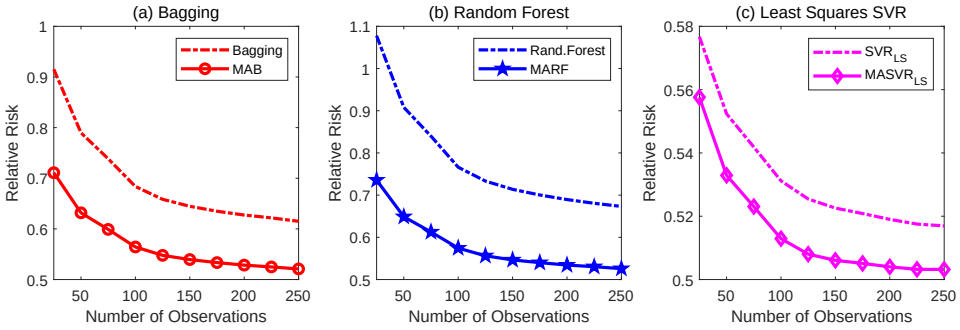
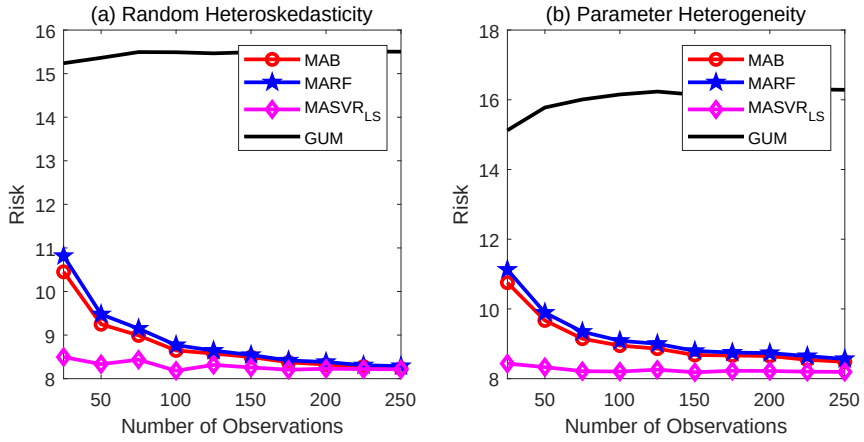


Figure 3: Risk Comparison under Different Scenarios



the statistical learning methods are more resistant to this form of neglected heterogeneity, since model uncertainty was acknowledged and treated by the hybrid algorithm.

In summary, the results from the Monte Carlo experiments suggest that there are benefits from the hybrid strategies when there exists significant parameter heterogeneity, perhaps due to jumps or threshold effects. Econometric strategies that use the mean or average marginal effects simply do not allow for good forecasts when there is large heterogeneity in effects both within and across subgroups. Intuitively, this additional heterogeneity shifts to the residual, creating new outliers that change the effective weighting on different observations.[16] In contrast, recursive partitioning methods rule out heterogeneity by assigning an equal weight to each observation within a subgroup.

4 Empirical Exercise

4.1 Data

We collected data on the universe of movies released in North America between October 1, 2010 and June 30, 2013. As detailed in appendix E, with the assistance of the IHS film consulting unit, the characteristics of each film were characterized by a series of indicator variables describing the film's genre,[17] the rating of a film's content provided by the Motion Picture Association of America's system (G, PG, PG13 and R), film budget excluding advertising and both the pre-determined number of weeks and screens the film studio forecasted the specific film will be in theatres measured approximately six weeks prior to opening. In our analysis, we focus on initial demand with both opening weekend box office ($n = 178$) and total sales of both DVD and Blu-Rays ($n = 143$) upon initial release.

[16] Appendix C.2 presents Monte Carlo evidence which shows that splits in trees occur at different locations and that there is more variation in outcomes in the final leaves with heteroskedastic data relative to homoskedastic data. Related, appendix F.4 presents evidence that the performance of model screening approaches and model averaging or Lasso methods that directly consider heteroskedasticity is invariant to the source of heteroskedasticity. In practice, we find minimal gains from modifying model screening, model averaging and Lasso approaches to allow for heteroskedasticity. This finding may appear surprising at first, but recall that the theoretical benefits of most model screening methods relate to efficiency.

[17] In total, we have 14 genres: Action, Adventure, Animation, Biography, Comedy, Crime, Drama, Family, Fantasy, Horror, Mystery, Romance, Sci-Fi, and Thriller.

To measure purchasing intentions from the universe of Twitter messages (on average, approximately 350 million tweets per day) we consider two measures. First, the sentiment specific to a particular film is calculated using an algorithm based on Hannak et al. (2012) that involves textual analysis of movie titles and movie key words. In each Twitter message that mentions a specific film title or key word, sentiment is calculated by examining the emotion words and icons that are captured within.[18] The sentiment index for a film is the average of the sentiment of the scored words in all of the messages associated with a specific film. Second, we calculate the total unweighted volume of Twitter messages for each specific film. We consider volume separate from sentiment in our analyses since the latter may capture perceptions of quality, whereas volume may proxy for interest.[19]

Across all the films in our sample, there is a total of 4,155,688 messages to be assessed. There is a large amount of time-varying fluctuations in both the number of, and sentiment within the Twitter messages regarding each film. Some of this variation reflects responses to the release of different marketing campaigns designed to both build awareness and increase anticipation of each film. Thus, in our application we define measures from social media data over different time periods. That is, suppose the movie release date is T, we separately calculate sentiment in ranges of days within the window corresponding to 4 weeks prior to and subsequent the release date.[20]

Summary statistics are presented in table 1. The mean budget of films is respectively approximately 61 and 63 million for the open box office and retail unit sales outcome. On average, these films were released in theatres for 14 weeks and played on roughly 3000 screens. Not surprisingly, given trends in advertising, the volume of Tweets increases

[18] In total, each of 75,065 unique emotion words and icons that appeared in at least 20 tweets between January 1st, 2009 to September 1st, 2009 is given a specific value that is determined using emotional valence. Note that Twitter messages were capped at 140 characters throughout this period. These messages often contain acronyms and Twitter specific syntax such as hashtags that may present challenges to traditional sentiment inference algorithms. The algorithm we use was developed by Janys Analytics for IHS-Markit was also used for the initial reported measures of the Wall Street Journal-IHS U.S. Sentiment Index

[19] Prior work by Liu (2006) and Chintagunta, Gopinath, and Venkataraman (2010) suggest that sentiment in reviews affect subsequent box office revenue. Similarly, Xiong and Bharadwaj (2014) finds that pre-launch blog volume reflects the enthusiasts' interest, excitement and expectations about the new product.

[20] For a typical range, T–a/–b, it stands for a days before date T (release date) to b days before date T. 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. Similarly, T+c/+d means c days to d days after date T, which are additionally used for forecasting the retail unit sales. Similarly, to reduce concerns related to reverse causality, we ensure that we do not include any Twitter data post release of the Blu-Ray.

Table 1: Summary Statistics

Variable	Open Box Office ($n = 178$)		Retail Unit Sales ($n = 143$)	
	Mean	**Std. Dev.**	**Mean**	**Std. Dev.**
Genre				
Action	0.3202	0.4679	0.3357	0.4739
Adventure	0.2416	0.4292	0.2378	0.4272
Animation	0.0843	0.2786	0.0909	0.2885
Biography	0.0393	0.1949	0.0420	0.2012
Comedy	0.3652	0.4828	0.3776	0.4865
Crime	0.1966	0.3986	0.1818	0.3871
Drama	0.3483	0.4778	0.3706	0.4847
Family	0.0562	0.2309	0.0629	0.2437
Fantasy	0.1011	0.3023	0.0909	0.2885
Horror	0.1180	0.3235	0.1049	0.3075
Mystery	0.0899	0.2868	0.0909	0.2885
Romance	0.1124	0.3167	0.0979	0.2982
Sci-Fi	0.1124	0.3167	0.1119	0.3163
Thriller	0.2416	0.4292	0.2517	0.4355
Rating				
PG	0.1461	0.3542	0.1608	0.3687
PG13	0.4213	0.4952	0.4126	0.4940
R	0.4270	0.4960	0.4196	0.4952
Core Parameters				
Budget (in million)	60.9152	56.9417	63.1287	56.5959
Weeks	13.9446	5.4486	14.4056	5.7522
Screens (in thousand)	2.9143	0.8344	2.9124	0.8498
Sentiment				
T-21/-27	73.5896	3.2758	73.4497	3.5597
T-14/-20	73.6999	3.0847	73.7530	3.0907
T-7/-13	73.8865	2.6937	73.9411	2.6163
T-4/-6	73.9027	2.7239	73.8931	2.8637
T-1/-3	73.8678	2.8676	73.7937	3.0508
T+0			73.8662	3.0887
T+1/+7			73.8241	3.1037
T+8/+14			73.4367	3.8272
T+15/+21			73.7001	3.3454
T+22/+28			74.0090	2.7392
Volume				
T-21/-27	0.1336	0.6790	0.1499	0.7564
T-14/-20	0.1599	0.6649	0.1781	0.7404
T-7/-13	0.1918	0.6647	0.2071	0.7377
T-4/-6	0.2324	0.8400	0.2494	0.9304
T-1/-3	0.4553	0.9592	0.4952	1.0538
T+0			1.5233	3.2849
T+1/+7			0.6586	1.1838
T+8/+14			0.3059	0.8290
T+15/+21			0.2180	0.7314
T+22/+28			0.1660	0.7204

sharply close to the release date and peaks that day. Following a film's release we find a steady decline in the amount of social web activity corresponding to a film.

4.2 Simulation Experiment Design

To examine the importance of incorporating data from the social web either using traditional estimators or an approach from the machine learning literature, we follow Hansen and Racine (2012) and conduct the following experiment to assess the relative prediction

efficiency of different estimators with different sets of covariates. The estimation strategies that we contrast can be grouped into the following categories i) traditional econometric approaches, ii) model screening approaches, iii) and iv) machine learning approaches, and v) newly proposed hybrid methods that combine econometrics with machine learning algorithms to capture richer patterns of heterogeneity. Table 2 lists each estimator analyzed in the exercise. Online Appendices A, B, and D provide further details on each econometric estimator and machine learning strategy considered.

The experiment shuffles the original data with sample n, into a training set of n_T and an evaluation set of size $n_E = n - n_T$. Using the training set, we obtain parameter estimates from each strategy that are then used to forecast outcomes for the evaluation set. With these forecasts, we evaluate each of the forecasting strategies by calculating the mean squared forecast error (MSFE) and the mean absolute forecast error (MAFE):

$$\begin{aligned} \text{MSFE} &= \frac{1}{n_E}(y_E - x_E\hat{\beta}_T)^\top(y_E - x_E\hat{\beta}_T), \\ \text{MAFE} &= \frac{1}{n_E}\left|y_E - x_E\hat{\beta}_T\right|^\top \iota_E, \end{aligned}$$

where (y_E, x_E) is the evaluation set, n_E is the number of observations of the evaluation set, $\hat{\beta}_T$ is the estimated coefficients by a particular model based on the training set, and ι_E is a $n_E \times 1$ vector of ones. In total, this exercise is carried out 10,001 times for different sizes of the evaluation set, $n_E = 10, 20, 30, 40$.

In total, there are $2^{23} = 8,388,608$ and $2^{29} = 536,870,912$ potential candidate models for open box office and movie unit sales respectively. This presents computational challenges for the HRC$_p$ and other model averaging estimators. Thus, we conducted the following model screening procedure based on the GETS method to reduce the set of potential candidate models for model selection and model averaging methods. Based on the OLS results presented in table A4, we restrict each potential model to contain a constant term and 7 (11) relatively significant parameters for open box office (movie unit sales). Next, to control the total number of potential models, a simplified version of the automatic general-to-specific approach of Campos, Hendry, and Krolzig (2003) is used for model screening.[21] While this restriction may appear severe by ruling out many poten-

[21]This approach explores through the whole set of potential models and examine each model using the

Table 2: List of Estimators Evaluated in the Prediction Error Experiments

Panel A: Econometric Methods		
(1)	GUM	A general unrestricted model that utilize all the independent variables described above
(2)	MTV	A general unrestricted model that does not incorporate the Twitter based sentiment and volume variables
(3)	GETS	A model developed using the general to specific method of Hendry and Nielsen (2007)
(4)	AIC	A model selected using the Akaike Information Criterion method
(5)	PMA	The model selected using the prediction model averaging proposed by Xie (2015)
(6)	HPMA	The model selected using a heteroskedasticity-robust version of the PMA method discussed in appendix D.5
(7)	JMA	The model selected by the jackknife model averaging (Hansen and Racine, 2012)
(8)	HRC_p	The model selected by hetero-robust C_p (Liu and Okui, 2013)
(9)	$\text{OLS}_{10,12,15}$	The OLS post Lasso estimator of Belloni and Chernozhukov (2013) with 10, 12, and 15 explanatory variables selected by the Lasso
(10)	$\text{HRC}^p_{10,12,15}$	The HRC_p model averaging post Lasso estimation strategy with 10, 12, and 15 explanatory variables selected by the Lasso
Panel B: Model Screening		
(1)	GETS_s	Three threshold p-values are selected, as $p = 0.24, 0.28$, and 0.32 for open box office, and $p = 0.30, 0.34$, and 0.38 for movie unit sales
(2)	ARMSH	The modified hetero-robust adaptive regression by mixing with model screening method of Yuan and Yang (2005)
(3)	HRMS	The hetero-robust model screening of Xie (2017)
(4)	Double-Lasso	We set all tuning parameters in the two steps as equal, and we control the tuning parameter so as to select a total of 10, 12, and 15 parameters
(5)	Benchmark	The GETS method we used in previous experiments, that is, $p = 0.3$ and 0.35 for open box office and movie unit sales, respectively
Panel C: Popular Machine Learning strategies		
(1)	RT	Regression tree of Breiman, Friedman, and Stone (1984)
(2)	BAG	Bootstrap aggregation of Breiman (1996) with $B = 100$ bootstrap samples and all of the K^{total} covariates
(3)	RF	Random forest of Breiman (2001) with $B = 100$ bootstrap samples and $q = \lfloor 1/3K^{total} \rfloor$ covariates
Panel D: Advanced Machine Learning Methods		
(1)	Gradient Boosting	quadratic loss function with $B = 100$ learning cycles
(2)	BART	Bayesian additive regression trees by Chipman, George, and McCulloch (2010) with default setting and $B = 100$
(3)	HBART	heteroskedasticity-robust BART by Pratola, Chipman, George, and McCulloch (2020) with default setting and $B = 100$
(4)	BART-BMA	Bayesian model averaging BARTby Hernández, Raftery, Pennington, and Parnell (2018) with default setting with $B = 100$
(5)	Linear regression tree	we apply OLS to each leaf created by conventional CART
(6)	M5′	proposed by Quinlan (1992), combines regression tree with linear regression at the nodes
(7)	SECRET	scalable linear regression tree algorithm by Dobra and Gehrke (2002), similar to M5′ but solve the problem from the perspective of classification
(8)	SVR	Support vector machine for regression by Drucker, Burges, Kaufman, Smola, and Vapnik (1996)
(9)	SVR_{LS}	Least squares support vector regression by Suykens and Vandewalle (1999)
Panel E: Newly Proposed Hybrid Methods		
(1)	MAB	Hybrid applying the PMA method on subgroups created by BAG, $B = 100$ bootstrap samples and all of the K^{total} covariates
(2)	MARF	Hybrid applying the PMA method on subgroups created by RF, $B = 100$ bootstrap samples and $q = \lfloor 1/3K^{total} \rfloor$ covariates
(3)	MASVR_{LS}	Hybrid applying the PMA and HPMA method to SVR_{LS}, where heteroskedasticity is considered in criteria function as in equation (10)

tial candidate model, numerous applications including Lehrer and Xie (2017), find that only a handful of models account for more than 95% of the total weight of the model averaging estimate.[22] Last, by altering the covariate set, one could additionally use this experiment to examine the importance of incorporating data from the social web on any of the econometric or machine learning strategies.

To facilitate replication we use default values of hyperparameters from standard and well-established software packages listed in appendix table A1 for each machine learning method. The tuning parameter for Lasso strategies is chosen to fix the number of explanatory variables selected (i.e. OLS_{10} indicates OLS with 10 variables selected by the Lasso). In appendices F6, F18, F19 and F20, we demonstrate the robustness of our findings presented in section 5 to using alternative hyperparameter values that deviate from the defaults.

5 Empirical Results

The results of the prediction errors exercise outlined in the preceding section are illustrated in figures 4 and 5 for open box office and movie unit sales respectively. The top panel of each figure presents the median MSFE and the bottom panel displays results for the median MAFE. In each panel, there are four lines that correspond to different sizes of the evaluation set and each point on the line presents the result for the listed estimator along the x-axis for that evaluation set size. The estimators are generally listed in order based on improvements in forecast accuracy, with the sole exception of the newly proposed hybrid methods being placed adjacent to their conventional machine learning approach. This reordering facilitates an examination of the marginal benefits of allowing for model uncertainty via each hybrid approach. Note that the values after RF and MARF refer to the number of randomly chosen explanatory variables used to determine a split

following rule: we first estimate the p-values for testing each parameter in the model to 0. If the maximum of these p-values exceeds our benchmark value, we exclude the corresponding model. In this way, we are deleting models with weak parameters from our model set. We set the benchmark value to equal to 0.3 and 0.35 for open box office and movie unit sales respectively, which is a very mild restriction. These preselection restrictions lead us to retain 105 and 115 potential models for open box office and retail movie unit sales respectively. Note, we did investigate the robustness of our results to alternative benchmark values and in each case the results presented in the next section are quite similar.

[22] See appendix F.5 for further discussion including the top 5 models in our experiment.

at each node. The raw data that generated each figure is presented in online appendix tables F.27 to F.29, which provides direct comparison of each forecasting strategy relative to the benchmark HRC_p estimator. As such, the benchmark HRC_p estimator is presented at the far left of each figure.

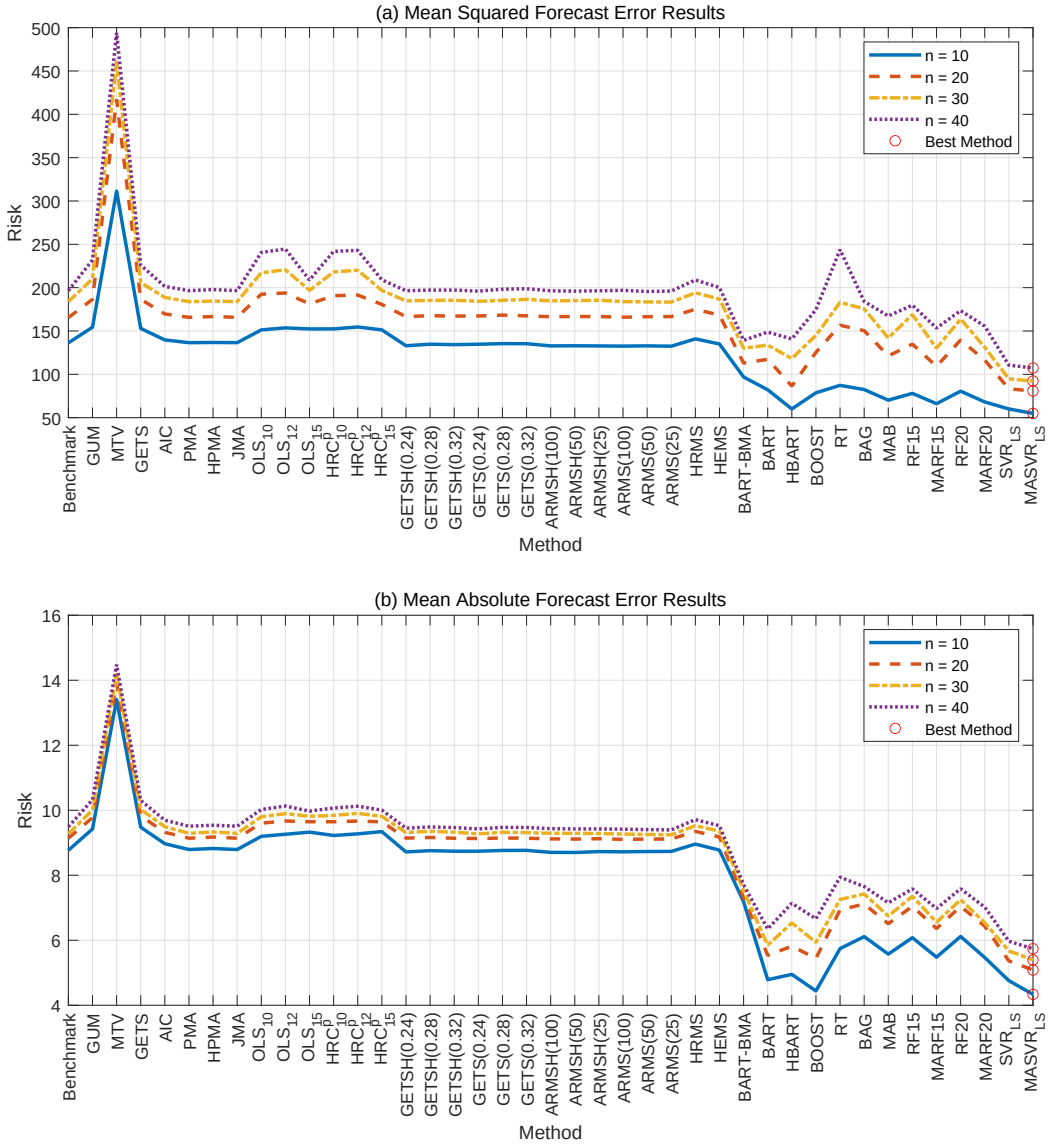
Our proposed MASVR_{LS} is presented at the far right of each figure since it demonstrates the best performance when evaluated by either MSFE or MAFE for both outcomes. Immediate to the left is the traditional SVR_{LS} approach that offers the second-best performance. Adding model averaging tends to lead to gains of 10% between SVR_{LS} and MASVR_{LS}. Results from the SPA test of Hansen (2005) in appendix F.9, present significant evidence of the superior predictive ability of the MASVR_{LS} method over each of the other ML tree based algorithms considered.

That said, for both outcomes when n_E is small, any of the machine learning methods considered in the exercise have dominating performance over the HRC_p as well as econometric estimators and penalization methods. Popular approaches from the statistical learning literature such as bagging and random forest greatly outperform the benchmark. In addition, we find gains of approximately 10% by adding model averaging to bagging that are of a similar order to incorporating model uncertainty with SVR_{LS}.

Comparing the results between figures 4 and 5, 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. 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.

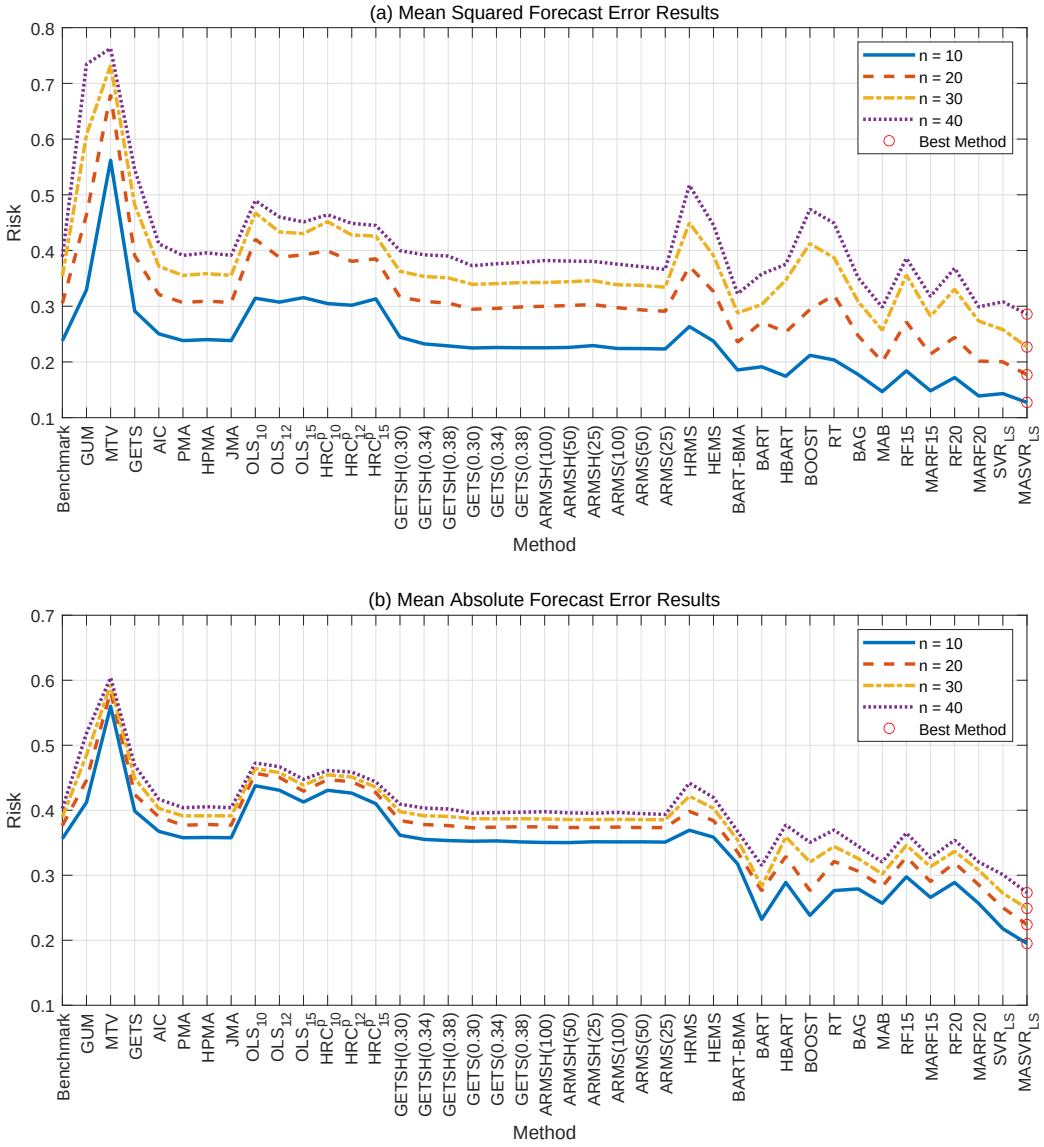
The far-left hand side of the x-axis in each figure is populated by the traditional econometric estimators listed in table 2. These estimators perform poorly relative to the other forecasting strategies. We observe that the three model averaging approaches and the model selected by AIC perform nearly as well as the benchmark HRC_p. Similarly, we observe small gains in forecast accuracy from the suite of model screening approaches

Figure 4: Results of Prediction Efficiency on Open Box Office



Note: Descriptions of each estimator presented in the horizontal axis of each figure are provided in table 2. The risks on the y-axises in the top and bottom panels represent the median of 10,001 MSFEs and MAFEs, respectively.

Figure 5: Results of Prediction Efficiency on DVD and Blu-Rays Sales



Note: Descriptions of each estimator presented in the horizontal axis of each figure are provided in table 2. The risks on the y-axises in the top and bottom panels represent the median of 10,001 MSFEs and MAFEs, respectively.

undertaken relative to the benchmark HRC_p. We find that there are very small gains from using HPMA in place of PMA. We also observe slightly better results from using a hetero-robust model screening method relative to the 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.[23] Taking these findings together, we conclude that there are small gains in our exercise from using econometric approaches that accommodate heteroskedasticity.[24] This finding differs with machine learning methods as we consistently find improved performance by allowing for heteroskedasticity with HBART relative to BART.

In each figure, we find that forecasts from a linear model that excludes social media data (MTV) exhibit the worst performance. This result is the first evidence that we present, which stresses the importance of social media data for forecast accuracy in the film industry. Additional experiments discussed in appendices F.3, F4.1, and F.6 make clear that to bolster forecast accuracy, both social media measures are needed. However, in contrast to Lehrer and Xie (2017) we find that the post-Lasso methods listed in table 2, 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.

Taking the evidence in figures 4 and 5 together with the Monte Carlo results presented in figure 3(a) and figure 3(b), leads us to conclude that the improved performance of SVR_{LS} in our empirical exercises arises due to the small sample size. Intuitively, tree-based strategies that perform a computationally greedy search over the covariates are restricted to making fewer splits in the tree structure given the sample size. In contrast, the optimization algorithm of SVR_{LS} is better able to learn the nonlinear decision surface. Despite the above advantage, there remain other trade-offs across machine learning algorithms that may include computational considerations.[25] With small samples, we il-

[23] Interestingly as presented in appendix F.7, the ARMS and ARMSH approaches select nearly identical weights and models.

[24] In appendix F.4, we use the Monte Carlo design introduced in section 3 to additionally evaluate whether the source of heteroskedasticity can explain some of these surprising results.

[25] In appendix F.16, we compare the computational efficiency of HBART and MARF in a sensitivity test that varies the relative gains as the number of bootstrap samples increase.

lustrate in appendix F.17 there are large gains from using either SVR_{LS} and $MASVR_{LS}$ that arise from their ability to capture nonlinearities relative to other approaches even in the absence of Twitter data. Further, the optimization algorithm of SVR ensures that it makes a local forecast that is specific to the statistic under investigation (i.e. conditional mean of Y). In contrast, both the econometric estimators considered including OLS and tree based algorithms are able to make a prediction across all covariate values in sample.

While the small sample size may explain why SVR_{LS} approaches perform well, we next evaluate a potential explanation for the improved performance of statistical learning tree based approaches relative to all of the econometric strategies. That is, the full suite of predictors is considered when building each tree, whereas model screening reduced the number of predictors to offer a computational advantage by limiting the number of candidate models for model averaging estimators. In appendix F.8, we repeat the prediction errors exercise above, where we additionally restrict the set of predictors to be identical for both the support vector and recursive partitioning strategies as the model screening and model averaging approaches. We continue to find large gains in forecast accuracy from random forest and bagging relative to the econometric approaches as well as dominant performance from SVR_{LS}.

Briefly, among the alternative machine learning strategies, we believe the improved performance of the hybrid tree-based strategies relative to HBART and boosting in each figure arises since the latter strategies build short trees and substantial heterogeneity remains in the terminal nodes. The hybrid approach nests the conventional local constant model and allows for more candidate models (and thereby) heterogeneity in terminal leaves with more observations. Similarly, the regression function used in each terminal leaf of popular linear regression tree algorithms is nested and contained among the multiple multivariate functions used to conduct forecasts in each terminal leaf in the hybrid approach. Further, with some linear regression tree algorithms, the fixed multivariate function in the terminal leaf may involve more covariates than observations available in the terminal leaf. Model averaging allows the researcher to consider all possible candidate models that involve at least as many covariates as one plus the number of observations in the respective terminal leaf.

As noted in section 4, we demonstrate the robustness of our findings to the choice of alternative hyperparameters. Specifically, appendix F.6 and F.18 respectively consider different parameters for Lasso and MARF. In general, we find no major differences in performance with either strategy, with the exception of few covariates are selected either because q is small or a large penalty is imposed with Lasso. This result and the small difference between Lasso and econometric model selection stresses that gains from machine learning in this application are not primarily due to regularization. Further, the results complement those presented in appendices B.2 and F.20 that explore changes in hyperparameters for numerous machine learning algorithms and illustrate small gains if hyperparameters are selected by cross-validation methods versus slight changes in the default values. Allowing for heteroskedasticity always leads to improved performance between BART and HBART and with the criteria function used for MASVR_{LS}. In addition, the small differences between SVR and SVR_{LS} suggest that the change in loss function also explain a small amount of gains relative to allowing for nonlinearities with the machine learning strategies. Last, in appendix F.19, we find small differences in forecast accuracy with using different kernel functions with SVR methods, although there are gains when we allow for nonlinearities by using a polynomial kernel in place of a linear kernel.

5.1 Relative Importance of the Explanatory Variables

Recursive partitioning and SVR algorithms were developed to make predictions and not understand the underlying process of how predictors correlate with outcomes. Empirical strategies have since been developed to identify which predictor variables are the most important in making forecasts.[26] The most important variables are the ones leading to the greatest losses in accuracy. For example, with bagging and random forests, each tree is grown with its respective randomly drawn bootstrap sample and the excluded data from the Out-Of-Bag sample (OOB) for that tree. The OOB sample is used to evaluate the tree or support vectors without the risk of overfitting since the observations did not build the tree. To determine importance, a given predictor is randomly permuted in the OOB

[26] Variable importance is often computed by applied researchers but the theoretical properties and statistical mechanisms of these algorithms are not well studied. To the best of our knowledge, Ishwaran (2007) presents the sole theoretical study of tree-based variable importance measures.

sample and the prediction error of the tree on the modified OOB sample is compared with the prediction error of the tree in the untouched OOB sample. This process is repeated for both each tree and each predictor variable. The average of this gap in prediction errors across all OOB samples provides an estimate of the overall decrease in accuracy that the permutation of removing a specific predictor induced.

We calculate variable importance scores using the MAB, MARF and MASVR_{LS} strategies.[27] The first three columns of table 3 include the social media variables as predictors for each hybrid approach. We find that these predictors account for between 3 to 7 of the top 10 most important predictors for open box office and movie unit sales in panels A and B, respectively. These results complement the comparison of forecast accuracy between the GUM and MTV models and reinforce the importance of including social media data to improve forecast accuracy irrespective of the estimation strategy. With MASVR_{LS} we find that social media measures are particularly important for forecasting retail movie unit sales, where four different volume measures are considered among the six most important predictors. With MAB and MARF, volume related variables are found to have a greater association with revenue outcomes than sentiment measures. These results suggest that the amount of social media buzz is more important than the emotional content when forecasting revenue outcomes.[28] Last, we observe that different forecasting strategies yield different rankings of the importance of each predictor both when social media measures are included as well as excluded from the specification as shown in the last 3 columns of table 3.

We next examine if there is heterogeneity in the variable importance measures across the film budget distribution. Motivating this exercise is the conjecture that sentiment may play a larger role for small budget films since they may benefit more from word of

[27] We consider both MAB and MARF since Strobl et al. (2008) showed that using mean decreased accuracy in variable importance with random forests is biased and could overestimate the importance of correlated variables. This bias exists if random forest did not select the correct covariate, but rather chose a highly correlated counterpart in a bootstrapped sample. This bias should not exist with bagging strategies that use all available predictors. Since Genuer, Poggi, and Tuleau-Malot (2010) could not replicate Strobl, Boulesteix, Kneib, Augustin, and Zeileis (2008)'s finding, we report both MAB and MARF.

[28] While the Lasso can be used to select variables to include in a regression model it does not rank them. In table A18, we report the numbers of Twitter sentiment and volume variables selected by Lasso in various samples. The results show that the Lasso also favors the inclusion of sentiment variables in almost all subsamples. This difference in the importance of social media variables selected may explain the uneven prediction performance of Lasso-related estimators in the appendix table F.27.

mouth or critical reviews. Table 4 presents estimates of the variable importance scores for films located in different budget quartile. Notice that constructed buzz measures are highly important for large budget films, but the volume of messages is key for many films in lower budget quartiles. The evidence in this study suggests that each social media measure captures a different dimension of purchasing intentions. Social media measures account for a smaller fraction of the most important predictors of box office opening for films in the second quartile of budget.

The striking difference in the ranking of the importance of social media variables across the budget distribution suggests model uncertainty arises due to parameter heterogeneity. This finding extends prior work that contrasts forecasting strategies with data from the film industry that is summarized in appendix E1.4 by considering a wider variety of algorithms and illustrating the robustness to choice of hyperparameters. The improved forecast accuracy of tree based and SVR methods show that the nonlinearities these methods generate are responsible for the significant improvements relative to econometric approaches. Further, the hybrid procedures yield further gains since this parameter heterogeneity is neglected with traditional strategies.

Although the variable importance measure differs from an estimate of a marginal effects of each predictor on revenue outcomes, our findings contribute to a large interdisciplinary literature surveyed in appendix E.1 that provides an understanding of whether online word-of-mouth explains box office openings and retail movie unit sales. The evidence in this study is consistent with i) Gopinath, Chintagunta, and Venkataraman (2013) who find considerable heterogeneity in the effects of online content, and ii) both Bandari, Asur, and Huberman (2012) and Xiong and Bharadwaj (2014) who stress the importance of measuring the dynamics in online buzz for forecasting film revenue.

Machine learning strategies can also inform researchers on which nonlinearities to include in the specification of an empirical model. Appendix F.21 provides an illustration of this idea by using a RT structure to suggest which interactions and nonlinear terms should be included in the specification of a linear model to explain film revenue. Estimates of the more flexible specification yield additional new findings that show there are threshold effects of social media measures on box office opening revenue. This result adds further

Table 3: Relative Importance of the Predictors

Ranking	With Twitter Variables			Without Twitter Variables		
	MAB	MARF	$MASVR_{LS}$	MAB	MARF	$MASVR_{LS}$
Panel A: Open Box Office						
1	Screens	Volume: T-1/-3	Volume: T-1/-3	Screens	Screens	Screens
2	Volume: T-1/-3	Budget	Screens	Budget	Rating: PG	Weeks
3	Volume: T-21/-27	Volume: T-4/-6	Budget	Genre: Comedy	Genre: Comedy	Genre: Adventure
4	Volume: T-4/-6	Screens	Genre: Drama	Genre: Animation	Genre: Adventure	Rating: R
5	Budget	Volume: T-14/-20	Genre: Fantasy	Genre: Adventure	Weeks	Genre: Animation
6	Volume: T-7/-13	Weeks	Sentiment: T-14/-20	Rating: R	Rating: R	Budget
7	Volume: T-14/-20	Sentiment: T-14/-20	Weeks	Genre: Fantasy	Genre: Fantasy	Genre: Drama
8	Weeks	Rating: R	Genre: Adventure	Rating: PG	Rating: PG13	Genre: Fantasy
9	Genre: Adventure	Genre: Fantasy	Sentiment: T-21/-27	Genre: Horror	Genre: Horror	Rating: PG13
10	Rating: PG13	Sentiment: T-1/-3	Rating: R	Rating: PG13	Genre: Animation	Genre: Horror
Panel B: Movie Unit Sales						
1	Screens	Volume: T-4/-6	Volume: T+0	Screens	Screens	Screens
2	Volume: T+0	Genre: Adventure	Screens	Weeks	Weeks	Weeks
3	Budget	Volume: T+0	Volume: T-1/-3	Budget	Budget	Budget
4	Weeks	Volume: T+15/+21	Weeks	Genre: Comedy	Genre: Comedy	Genre: Adventure
5	Volume: T+8/+14	Genre: Fantasy	Volume: T-4/-6	Genre: Fantasy	Genre: Fantasy	Genre: Action
6	Volume: T-21/-27	Screens	Volume: T+8/+14	Genre: Adventure	Genre: Adventure	Genre: Fantasy
7	Volume: T+15/+21	Genre: Family	Budget	Rating: R	Rating: R	Genre: Comedy
8	Genre: Comedy	Volume: T+22/+28	Sentiment: T+22/+28	Genre: Horror	Genre: Thriller	Genre: Drama
9	Genre: Fantasy	Budget	Sentiment: T+0	Genre: Thriller	Genre: Horror	Rating: R
10	Genre: Animation	Genre: Animation	Sentiment: T-7/-13	Genre: Mystery	Genre: Family	Rating: PG13

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

Table 4: Heterogeneity in the Relative Importance of Predictors by Film Budget

Ranking	MAB	MARF	$MASVR_{LS}$	MAB	MARF	$MASVR_{LS}$
Panel A: Open Box Office						
	1^{st} Quartile			2^{nd} Quartile		
1	Screens	Genre: Drama	Volume: T-7/-13	Screens	Sentiment: T-1/-3	Screens
2	Weeks	Weeks	Volume: T-1/-3	Sentiment: T-1/-3	Volume: T-14/-20	Genre: Thriller
3	Genre: Drama	Rating: PG13	Genre: Drama	Budget	Weeks	Weeks
4	Genre: Comedy	Genre: Comedy	Volume: T-4/-6	Volume: T-21/-27	Screens	Volume: T-7/-13
5	Genre: Horror	Rating: R	Screens	Weeks	Genre: Romance	Rating: PG
6	Rating: R	Volume: T-4/-6	Volume: T-21/-27	Genre: Sci-Fi	Rating: PG	Rating: PG13
7	Rating: PG	Screens	Weeks	Genre: Romance	Genre: Sci-Fi	Genre: Sci-Fi
8	Genre: Adventure	Genre: Crime	Volume: T-14/-20	Rating: PG	Genre: Biography	Budget
9	Genre: Animation	Volume: T-1/-3	Sentiment: T-21/-27	Genre: Crime	Genre: Fantasy	Genre: Adventure
10	Genre: Family	Genre: Romance	Genre: Horror	Genre: Biography	Genre: Mystery	Genre: Romance
	3^{rd} Quartile			4^{th} Quartile		
1	Budget	Budget	Budget	Volume: T-4/-6	Volume: T-4/-6	Volume: T-4/-6
2	Volume: T-21/-27	Sentiment: T-1/-3	Volume: T-14/-20	Screens	Volume: T-1/-3	Screens
3	Sentiment: T-1/-3	Genre: Comedy	Genre: Sci-Fi	Budget	Budget	Volume: T-7/-13
4	Screens	Volume: T-14/-20	Sentiment: T-1/-3	Volume: T-1/-3	Genre: Fantasy	Volume: T-1/-3
5	Genre: Comedy	Genre: Action	Volume: T-21/-27	Volume: T-7/-13	Sentiment: T-14/-20	Budget
6	Volume: T-14/-20	Rating: R	Sentiment: T-4/-6	Volume: T-21/-27	Volume: T-14/-20	Sentiment: T-21/-27
7	Genre: Action	Rating: PG13	Genre: Thriller	Genre: Fantasy	Weeks	Genre: Fantasy
8	Rating: PG13	Sentiment: T-4/-6	Sentiment: T-7/-13	Genre: Drama	Genre: Family	Volume: T-14/-20
9	Sentiment: T-7/-13	Volume: T-21/-27	Screens	Genre: Family	Genre: Drama	Sentiment: T-4/-6
10	Genre: Animation	Genre: Drama	Genre: Family	Genre: Action	Screens	Sentiment: T-14/-20
Panel B: Movie Unit Sales						
	1^{st} Quartile			2^{nd} Quartile		
1	Screens	Volume: T-4/-6	Volume: T-21/-27	Screens	Genre: Horror	Screens
2	Weeks	Sentiment: T+15/+21	Volume: T+8/+14	Weeks	Sentiment: T-7/-13	Weeks
3	Genre: Romance	Sentiment: T+22/+28	Screens	Genre: Horror	Genre: Drama	Genre: Horror
4	Sentiment: T+22/+28	Sentiment: T+0	Volume: T-4/-6	Rating: PG	Genre: Adventure	Volume: T-21/-27
5	Sentiment: T+15/+21	Volume: T+0	Volume: T+15/+21	Genre: Sci-Fi	Rating: R	Volume: T+0
6	Genre: Animation	Genre: Drama	Volume: T-14/-20	Genre: Adventure	Rating: PG	Volume: T+8/+14
7	Genre: Family	Genre: Thriller	Volume: T-1/-3	Genre: Crime	Volume: T-7/-13	Sentiment: T-14/-20
8	Genre: Fantasy	Genre: Romance	Volume: T-7/-13	Genre: Biography	Genre: Romance	Genre: Comedy
9	Rating: PG	Rating: PG	Sentiment: T-1/-3	Genre: Fantasy	Genre: Biography	Volume: T-4/-6
10	Rating: PG13	Genre: Animation	Weeks	Sentiment: T-1/-3	Genre: Comedy	Volume: T-14/-20
	3^{rd} Quartile			4^{th} Quartile		
1	Budget	Sentiment: T+15/+21	Budget	Screens	Volume: T-4/-6	Screens
2	Screens	Genre: Fantasy	Screens	Volume: T+0	Volume: T-14/-20	Volume: T-21/-27
3	Weeks	Genre: Adventure	Weeks	Volume: T+8/+14	Genre: Adventure	Volume: T+0
4	Genre: Fantasy	Sentiment: T-7/-13	Genre: Horror	Genre: Animation	Genre: Fantasy	Volume: T+8/+14
5	Rating: R	Volume: T-7/-13	Rating: R	Volume: T-21/-27	Volume: T+0	Sentiment: T+15/+21
6	Genre: Horror	Rating: R	Sentiment: T-7/-13	Genre: Comedy	Genre: Animation	Sentiment: T-7/-13
7	Genre: Sci-Fi	Genre: Comedy	Genre: Drama	Volume: T-14/-20	Volume: T+15/+21	Genre: Fantasy
8	Rating: PG13	Genre: Romance	Genre: Sci-Fi	Volume: T-4/-6	Sentiment: T+22/+28	Volume: T-4/-6
9	Genre: Mystery	Budget	Volume: T-1/-3	Weeks	Sentiment: T-7/-13	Genre: Drama
10	Genre: Biography	Weeks	Sentiment: T+8/+14	Genre: Action	Sentiment: T+0	Volume: T-14/-20

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

emphasis of the need for researchers to flexibly consider multiple metrics collected from social media data, since they may proxy for alternative dimensions of consumer demand.

6 Conclusion

The film industry is characterized by substantial uncertainty and De Vany and Walls (2004) report that just 22% of films among 2,000 movies exhibited between 1984 and 1996, either made a profit or broke-even. Since social media can be used to gauge interest in movies before they are released as well as provide measures of potential audience response to marketing campaigns, there is excitement in this industry about using this new data source in forecasting exercises. Not only can a new data source potentially improve forecasts, so too can adopting either recursive partitioning or SVR algorithms developed for data mining applications. Using data from the film industry we find significant gains in forecast accuracy from using these algorithms in place of either dimension reduction or traditional econometrics approaches.

Despite the clear practical benefits from using machine learning, we suggest that heteroskedastic data may hinder the performance of many algorithms. We propose a new hybrid strategy that applies model averaging to observations in either each support vector or within each leaf subgroup created by a statistical learning algorithm. Our empirical investigation demonstrates that irrespective of the machine learning algorithm, there are significant gains in forecast accuracy from the proposed hybrid strategy. We find larger gains from the hybrid strategy involving least squares 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. Further, our analysis casts doubt that there are gains from modifying traditional econometric approaches, penalization methods or model screening methods to account for heteroskedasticity.

Monte Carlo experiments shed further light on why these additional gains are achieved. Evidence from these simulations show that gains from combining model averaging with either recursive partitioning or support vectors are obtained when heteroskedasticity

arises due to neglected parameter heterogeneity. Last, we find benefits from incorporating social media in forecasting exercises for the film industry, in part since up to 7 of the 10 most influential variables when using statistical learning algorithms originate from this new data source.

A challenge facing researchers in machine learning is known as the no free lunch theorem of optimization due to Wolpert and Macready (1997). This is an impossibility theorem that rules out the possibility that a general-purpose universal optimization strategy exists. The optimal strategy depends not just on the sample size and what is being forecasted, but also the structure of the specific problem under consideration that is generally unknown ex-ante to the analyst. Yet, we argue that since heteroskedastic data is the norm in the real world, our proposed hybrid strategy with either tree based structures or least squares support vector regression may both add significant value and can complement the HBART strategy developed in Pratola, Chipman, George, and McCulloch (2020).

To subsequently advance the literature on how social media influences film industry revenue, a potential direction would consider less aggregated Twitter volume and sentiment score measures as explanatory variables. For example, one could measure mood from subset(s) of tweets based on subgroups characterized by either number of followers or demographic characteristics or even whether the Twitter message has a positive or negative orientation. By unpacking the social media sentiment into its components, one could understand what type of emotions conveyed in individual tweets is associated with purchasing decisions. Future work is also needed to understand the statistical properties of hybrid strategies as well as developing formal tests that can detect the source of heteroskedasticity in settings with many covariates, to help guide practitioners choice of strategy. In addition, developing diagnostics that can evaluate forecasting strategies on the basis of not just the bias and efficiency of the estimator, but also the forecasting strategy's computational complexity should prove fruitful to aid in business decision making.

References

BAN, G.-Y., N. E. KAROUI, AND A. E. B. LIM (2018): "Machine Learning and Portfolio Optimization," *Management Science*, 64(3), 1136–1154.

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*.

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

BOLLEN, J., H. MAO, AND X. ZHENG (2011): "Twitter Mood Predicts the Stock Market," *Journal of Computational Science*, 2(1), 1–8.

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.

BRODLEY, C. E., AND P. E. UTGOFF (1995): "Multivariate decision trees," *Machine Learning*, 19(1), 45–77.

CAMPOS, J., D. F. HENDRY, AND H.-M. KROLZIG (2003): "Consistent Model Selection by an Automatic Gets Approach," *Oxford Bulletin of Economics and Statistics*, 65(s1), 803–819.

CHAUDHURI, P., M.-C. HUANG, W.-Y. LOH, AND R. YAO (1994): "Piecewise-polynomial regression trees," *Statistica Sinica*, 4, 143–167.

CHINTAGUNTA, P. K., 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.

DE VANY, A. S., AND W. WALLS (2004): "Motion picture profit, the stable Paretian hypothesis, and the curse of the superstar," *Journal of Economic Dynamics and Control*, 28(6), 1035–1057.

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.

FAN, G., AND J. B. GRAY (2005): "Regression Tree Analysis Using TARGET," *Journal of Computational and Graphical Statistics*, 14(1), 206–218.

GENUER, R., J.-M. POGGI, AND C. TULEAU-MALOT (2010): "Variable Selection Using Random Forests," *Pattern Recognition Letters*, 31(14), 2225 – 2236.

GOH, K.-Y., C.-S. HENG, AND Z. LIN (2013): "Social Media Brand Community and Consumer Behavior: Quantifying the Relative Impact of User- and Marketer-Generated Content," *Information Systems Research*, 24(1), 88–107.

GOPINATH, S., P. K. CHINTAGUNTA, AND S. VENKATARAMAN (2013): "Blogs, Advertising, and Local-Market Movie Box Office Performance," *Management Science*, 59(12), 2635–2654.

GRAY, J. B., AND G. FAN (2008): "Classification tree analysis using TARGET," *Computational Statistics & Data Analysis*, 52(3), 1362 – 1372.

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. (2014): "Model Averaging, Asymptotic Risk, and Regressor Groups," *Quantitative Economics*, 5, 495–530.

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.

HOTHORN, T., K. HORNIK, AND A. ZEILEIS (2006): "Unbiased Recursive Partitioning: A Conditional Inference Framework," *Journal of Computational and Graphical Statistics*, 15(3), 651–674.

ISHWARAN, H. (2007): "Variable Importance in Binary Regression Trees and Forests," *Electronic Journal of Statistics*, 1, 519–537.

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.

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.

LOH, W.-Y., AND Y.-S. SHIH (1997): "Split Selection Methods for Classification Trees," *Statistica Sinica*, 7, 815.

MANSKI, C. F. (2004): "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72(4), 1221–1246.

MURTHY, S. K., S. KASIF, AND S. SALZBERG (1994): "A System for Induction of Oblique Decision Trees," *Journal of Artificial Intelligence Research*, 2.

PRATOLA, M. T., H. A. CHIPMAN, E. I. GEORGE, AND R. E. MCCULLOCH (2020): "Heteroscedastic BART via Multiplicative Regression Trees," *Journal of Computational and Graphical Statistics*, 29(2), 405–417.

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

SILVA, J. M. C. S., AND S. TENREYRO (2006): "The Log of Gravity," *The Review of Economics and Statistics*, 88(4), 641–658.

STEEL, M. F. (2019): "Model Averaging and its Use in Economics," *Journal of Economic Literature*, p. forthcoming.

STROBL, C., A.-L. BOULESTEIX, T. KNEIB, T. AUGUSTIN, AND A. ZEILEIS (2008): "Conditional Variable Importance for Random Forests," *BMC Bioinformatics*, 9(1), 307.

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

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

VASILIOS, P., P. THEOPHILOS, AND G. PERIKLIS (2015): "Forecasting Daily and Monthly Exchange Rates with Machine Learning Techniques," *Journal of Forecasting*, 34(7), 560–573.

WAGER, S., AND S. ATHEY (2018): "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 113(523), 1228–1242.

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

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.

YUAN, Z., AND Y. YANG (2005): "Combining Linear Regression Models: When and How?," *Journal of the American Statistical Association*, 100(472), 1202–1214.

ZHANG, X., A. ULLAH, AND S. ZHAO (2016): "On the dominance of Mallows model averaging estimator over ordinary least squares estimator," *Economics Letters*, 142, 69–73.

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.

ZHANG, X., G. ZOU, AND R. J. CARROLL (2015): "Model Averaging Based on Kullback-Leibler Distance," *Statistica Sinica*, 25, 1583.