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

June 2019

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 recursive partitioning strategies greatly outperform dimension reduction strategies and traditional econometrics approaches in forecast 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.

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

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 and random forests.[1] These statistical learning algorithms do not specify a structure for the model to forecast the mean and often achieve predictive gains relative to conventional econometric approaches.[2] Despite this benefit in modeling, the algorithm used to build tree based structures via recursive partitioning implicitly assumes homogeneous variance across the entire explanatory-variable space.[3]

[1] More recent work by Cui, Gallino, Moreno, and Zhang (2018) and Lehrer, Xie, and Zhang (2018) consider these off the shelf methods and each present evidence that the value of social media information increases with the machine learning technique's level of sophistication.

[2] Gains from statistical learning arise by allowing for nonlinear predictor interactions that are missed by common econometric estimators. Subsection F.13 in the Appendix provides an illustration of the improved forecasting accuracy of random forest and bagging strategies relative to the estimators contrasted in Lehrer and Xie (2017). Further, the authors restricted the films considered in their exercise on the basis of a budget criteria, which reduces the amount of heterogeneity in the data. We extend their empirical exercise in this paper to address both issues and demonstrate the practical benefits of the new hybrid estimators proposed.

[3] More generally, both 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

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 how the data is partitioned thereby influencing the structure of regression trees. In this paper, we introduce new strategies for predictive analytics that are contrasted with existing tools from both the econometrics and machine learning literature to first give guidance on how to improve forecast accuracy in applications within the film industry.[4] Motivating our 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.

We propose hybrid strategies that first use recursive partitioning methods to develop subgroups and then undertake model averaging within these terminal groups to generate forecasts. Traditionally, forecasts from regression trees use a local constant model that assumes homogeneity in outcomes within individual terminal leaves. By allowing for model uncertainty in the leaves, richer forms of heterogeneity in the relationships between independent variables and outcomes within each leaf subgroup is allowed. To the best of our knowledge, only Pratola, Chipman, George, and McCulloch (2018) consider incorporating heteroskedasticity in the machine learning literature within a Bayesian framework. In our empirical application, we find significant computational advantages from using our hybrid strategy relative to the approach developed in Pratola, Chipman, George, and McCulloch (2018), while achieving nearly identical predictive accuracy as measured by mean square forecast error.

Our empirical examination of the predictive accuracy of alternative empirical strate-

by accounting for limited forms of parameter heterogeneity.

[4]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 forecasting accuracy of machine learning techniques on forecasting daily and monthly exchange rates, Wager and Athey (2017) propose variants of random forests to estimate causal effects, and Ban, Karoui, and Lim (2018) adopted machine learning methods for portfolio optimization.

gies 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,[5] which likely arises since different films appeal to populations drawn from different distributions. The results provide new insights on the trade-offs researchers face when choosing a forecasting method. Recursive partitioning strategies including regression trees, bagging and random forests yield on average a 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, we find additional gains of roughly 10% in forecast accuracy from our proposed strategy that allows for model uncertainty in each subgroup leaf relative to algorithms that estimate a single leaf-specific model. These gains are exhibited across a variety of algorithms including random forest, bagging and M5', that each create leaves by maximizing a local objective function that partitions the data within the tree structure. Monte Carlo experiments clarify why these gains arise in our empirical application. We find hybrid strategies are quite useful in settings where heteroskedasticity arises due to significant parameter heterogeneity, perhaps due to jumps or threshold effects, or simply

[5]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. We should also stress that another reason one needs to account for heteroskedasticity is parameter heterogeneity, which is a form of an omitted variables problem. However, the link between neglected parameter heterogeneity and heteroskedasticity are not well known among practitioners, but can be easily 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 science fiction film relative to an art house drama), then the variance of the error term varies too for a fixed-coefficient model. This link between neglected heterogeneity and heteroskedasticity has implications for specification tests and Chesher (1984) demonstrates that the well-known information matrix (IM) test due to White (1982) can be interpreted as a test against random parameter variation.

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 leaf of a bagging tree.[6]

Third, our analysis finds tremendous value from incorporating social media data in forecasting exercises. Econometric tests find that the inclusion of social media data leads to large gains in forecast accuracy. Calculations of variable importance from recursive partitioning methods show that measures of social media message volume account for 6 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 first briefly review traditional econometric and machine learning strategies to conduct forecasting. We then propose two new 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

[6]We also find larger gains relative to trees built using a boosting algorithm. This may arise since boosting builds trees that are quite short and thus have more observations (and heterogeneity) in the leaves. Further, our analysis finds that adding model averaging post variable selection by penalization methods or using a model screening approach leads to small gains relative to traditional econometric approaches.

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.[7] 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:

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

for $m = 1, ..., M$, where $x_{ij}^{(m)}$ for $j = 1, ..., k^{(m)}$ denotes the regressors, $\beta_j^{(m)}$ denotes the coefficients. The residual now contains both the original error term 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).[8]

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

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

[8] 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.

the DGP are given by $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{e}$, where $\boldsymbol{y} = [y_1, ..., y_M]^\top$, $\boldsymbol{\mu} = [\mu_1, ..., \mu_M]^\top$ and $\boldsymbol{e} = [e_1, ..., e_M]^\top$. 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 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".

One of the more popular algorithms is classification and regression decision trees (CART) introduced by Breiman, Friedman, and Stone (1984) that 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 $\Delta = \text{SSR}(\tau) - \text{SSR}(\tau_L) - \text{SSR}(\tau_R)$, reaches its global maximum;[9] 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 constant regression model

$$y_i = a + u_i, \quad i \in l, \tag{5}$$

where u_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, the forecast assumes any heterogeneity in outcomes within each subgroup is random. From the perspective of the econometrician, this can appear unsatisfying.

The statistical learning literature has noted both this drawback in how forecasts are made,[10] along with drawbacks in how splits within the tree are made, leading to further refinements. First, Hastie, Tibshirani, and Friedman (2009) discuss that individual regression trees are not powerful predictors relative to ensemble methods since they exhibit large variance.[11] Ensemble methods that combine estimates from multiple models or trees exist in both the machine learning and econometrics literature. Bootstrap aggregating decision trees (aka bagging) proposed in Breiman (1996) and random forest developed in Breiman (2001) are randomization-based ensemble methods that draw a parallel to model averaging.[12] In bagging, trees are built on random bootstrap copies of the orig-

[9] 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, with regression trees variables are added in a more flexible manner since every cut-point in each independent variable is considered 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.

[10] This approach approximates the DGP with a series of discontinuous flat surfaces forming an overall rough shape. Our hybrid strategy smooths the shape and is more general than algorithms that use weighted polynomial smoothing techniques to smooth forecasts between leaf nodes, see e.g. Chaudhuri, Huang, Loh, and Yao (1994).

[11] Put differently, since trees are constructed sequentially very small small perturbations in the sample used to construct a tree can lead to a very different tree model used for forecasts.

[12] The main idea is to introduce random perturbations into the learning procedure by growing multiple

inal data, producing multiple different trees. Bagging differs from random forest only in the set of explanatory factors being considered in each tree. That is, rather than consider which among the full set of explanatory variables leads to the best split at a node of the tree, random forests only consider a random subset of the predictor variables for the best split. With both strategies, the final forecast is obtained as an equal weight average of the individual tree forecasts.

Second, within the statistical learning literature studies have concluded that the split selection process is biased towards selecting variables with many split points due to the greater possibility for significantly different partitions to be found (see e.g. Loh and Shih, 1997; Kim and Loh, 2003; Hothorn, Hornik, and Zeileis, 2006). This critique appears imprecise and we argue that any split to minimize Δ with heteroskedastic data will be biased to regions of variables with high heteroskedasticity at the expense of regions of low heteroskedasticity. Thus, heteroskedastic data can lead to perhaps not choosing the "correct" first split of the root node can lead the rest of the tree down a sub-optimal path.[13]

To summarize, forecasts from recursive partitioning and model averaging methods are computationally expensive but differ in three important ways. 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. The remaining two differences relate to how predictions are weighted across the different models/trees. 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. Second, final predictions from regression trees rule out any model uncertainty in each final leaf $\bar{y}(\tau)$ of the tree.

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.

[13] 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 (Murthy, Kasif, and Salzberg, 1994; Brodley and Utgoff, 1995; Fan and Gray, 2005; 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.

This lack of heterogeneity and computational considerations motivate our two proposed extensions for forecasting with social media data. The first extension considers an improved method to select candidate models for model averaging estimators. The second extension proposes a hybrid strategy that combines recursive partitioning with model averaging to allow for heterogeneity in forecasts when the final leaf subgroup consists of observations that differ in some observed covariates.

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

2.1 A New Strategy for Model Screening

The empirical performance of any model averaging estimator crucially depends on the candidate model set. Let $\mathcal{M}$ denote the candidate model set before screening. In practice, one possible approach to construct the candidate model set is to consider a full permutation of all regressors. One obvious drawback of this approach is that the total number of candidate models increases exponentially with the number of regressors. 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 to be 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.[14] While most ex-

[14]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 and every possible model should not be estimated.

isting research assumes a pre-determined candidate model set, a recent paper by Zhang, Yu, Zou, and Liang (2016) established the asymptotic optimality of Kullback-Leibler (KL) loss based model averaging estimators with screened candidate models. Following this insight, we define $\tilde{\mathcal{M}}$ to be the candidate model set 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 (A10). We present the following set of assumptions

Assumption 1 *We assume that there exist 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) is the definition of v_n, which is the distance between a model risk 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 techniques accidentally omit at least one useful model 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 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, we have*

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

as $n \to \infty$.

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 creating the first two-step hybrid machine learning and econometrics estimator.[15] In this paper, we suggest using classification algorithms in the first step to build tree 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 the proposed hybrid approaches.

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

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

[15] Penalization methods such as the Lasso have objective functions designed to reduce the dimensionality of explanatory variables. 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.

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. This approach generalizes linear regression trees that were found to yield improvements over the local constant model in equation (5), by allowing there to be more than a single model to explain outcomes in each leaf.

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 classification, we end up with one particular tree leaf l. We denote the predicting observations that are classified in tree leaf l as $\boldsymbol{X}_{t\in l}^p$. The forecast for all observations can then be obtained as

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

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

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. With fewer predictors, the candidate model set for each leaf does not potentially consider each of the K regressors as in MAB, but rather is constructed with the k regressors used to split the nodes that generated this leaf l.[16] This restriction affects how $\hat{\boldsymbol{\beta}}_l(\boldsymbol{w})$ is calculated as it is averaged only over those leafs where it was randomly selected. The intuition of this hybrid strategy can be applied to any machine learning algorithm including ones with a different objective function to determine splits within a tree such as M5.

[16] 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, generates identical forecasts.

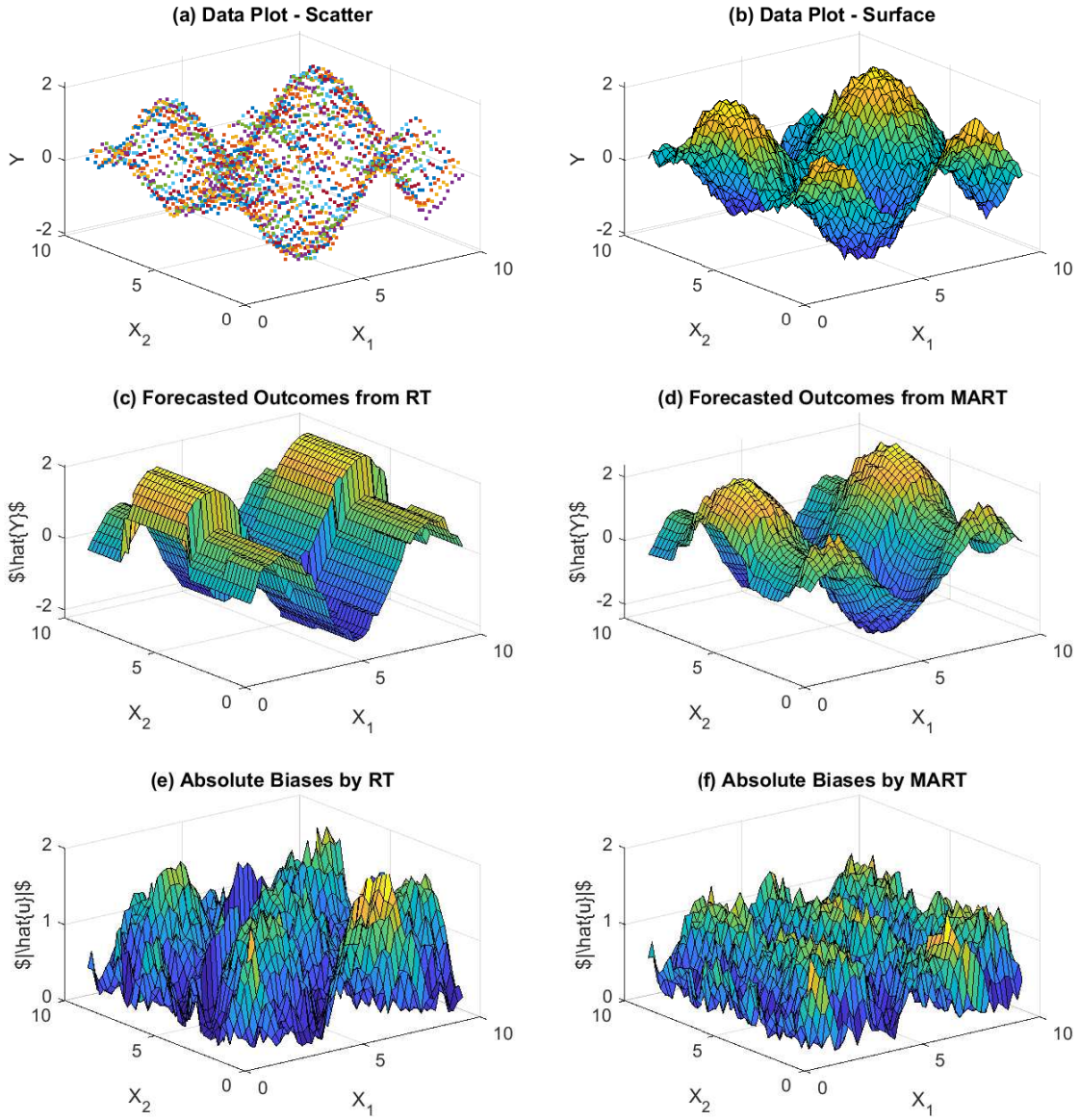
To illustrate the benefits of allowing for heterogeneity due to model uncertainty in each tree leaf in the forest via this two-step hybrid procedure, we simulate data drawn from a non-linear process. Panels (a) and (b) of Figure 1 respectively present the scatter plot and surface plot of training data generated by

$$Y = \sin(X_1) + \cos(X_2) + u,$$

where $X_1 \in [1, 10]$, $X_2 \in [1, 10]$, and u is a Gaussian noise with mean 0 and variance 0.01. Forecasts of Y calculated from RT and MART with the training data are presented in Panels (c) and (d) of Figure 1. Since RT forecasts assume homogeneity within leaves, the surface plot in Panel (c) appears similar to a step-function. 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.

Panels (e) and (f) of figure 1 respectively plot the forecast errors from RT and MART against both X_1 and X_2. Comparing the height of these figures shows that the absolute biases from MART are less than half of the biases obtained from RT. The reduced height occurs throughout the space spanned by X_1 and X_2 demonstrating that the gains are achieved by allowing for richer relationships in each tree leaf. In the next section, we conduct a formal Monte Carlo study to provide further insights on when allowing for model uncertainty may improve forecasts from recursive partitioning strategies.

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) and (d) display the forecasted shape by RT and MART, and plots (e) and (f) present the absolute value of forecast errors against the two explanatory variables for each forecasting strategy, respectively.

3 Monte Carlo Study

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

$$y_t = \mu_t + e_t = \sum_{j=1}^{\infty} (\beta_j + r \cdot \sigma_t) x_{jt} + e_t \tag{10}$$

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 introduces potential heterogeneity (depends on values of the scale variable r) to the model. We set $x_{1t} = 1$ and other x_{jt}s follow $\mathrm{N}(0, 1)$. Since the infinite series of x_{jt} is infeasible in practice, we truncate the process at $j_{\max} = 10,000$ without violating our assumption on the model set-up.[17] We assume that the whole 10,000 x_{jt}s set is not entirely feasible and we can only observe the first 20 regressors. Two scenarios designed to represent pure random heteroskedasticity and heteroskedasticity that arises due to neglected parameter heterogeneity are considered. Formally,

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

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

With this DGP, we compare the performance of conventional learning methods and model averaging learning methods using their risks.[18] We assume that the first $K = 5$

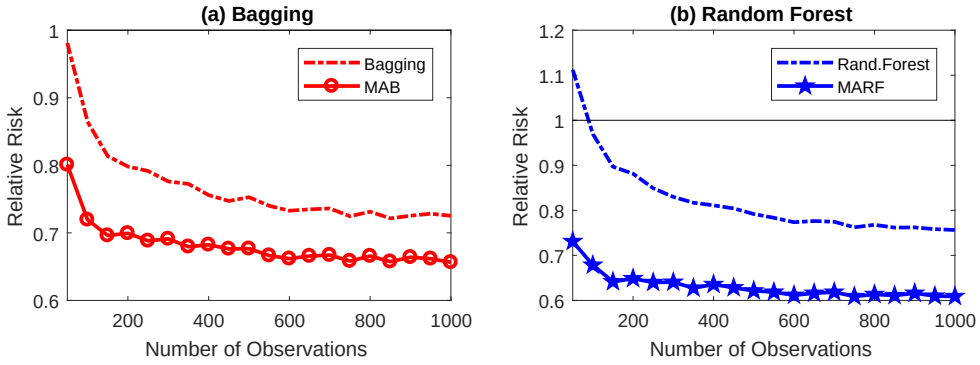
[17] The simulation design aims to mimic a big data environment, where the number of covariates is large. Variables with close-to-0 coefficients can be ignored since they barely influence the dependent variable. Such is the case for x_{jt} with $j > j_{\max}$. All results are robust to alternative values of the scale variable r.

[18] 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, and MARF. For each sample size, we compute the risk for all methods and average across 1,000 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 of the generalized unrestricted model.

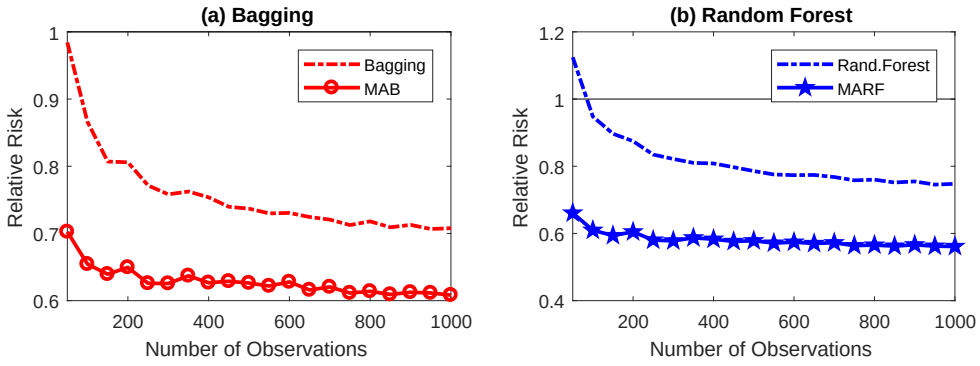
regressors are observed in both scenarios and fix the control parameter $c = 2$ when generating the true coefficients. Figure 2 panels A and B present results respectively for the random heteroskedasticity and parameter heterogeneity scenario. In each figure, the number of observations is presented on the horizontal axis, the relative risk is displayed on the vertical axis and dash-dotted (solid) lines respectively represent bagging and random forest (the model averaging counterpart). The results indicate that: i) the model averaging learning method performs much 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) MARF has the best relative performance in all cases. Overall, we observe smaller relative risks in the parameter heterogeneity scenario.

Figure 2: Relative Performance of Conventional and Model Averaging Learning

A. Random Heteroskedasticity

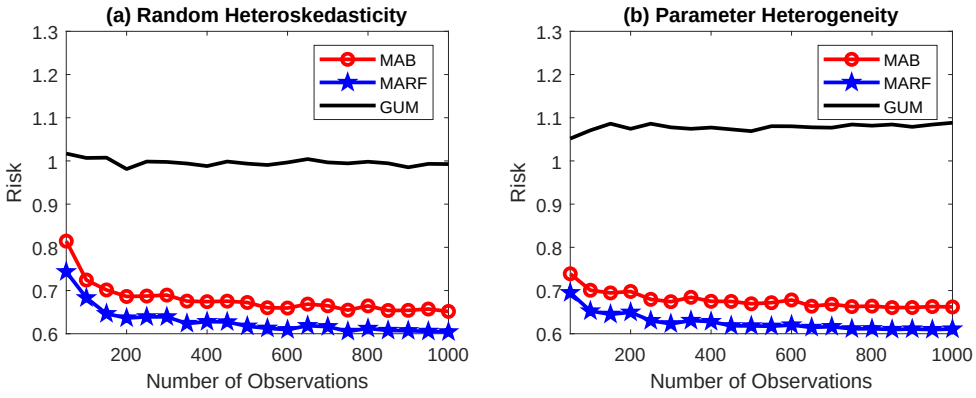


B. Parameter Heterogeneity



Since the results in figure 2 panel A are relative to a generalized unrestricted model (henceforth GUM) that utilizes all the independent variables, we next present absolute risks for all model averaging learning methods along with the risks of the GUM in figure 3. Figure 3(a) and (b) presents results for the absolute risks under random heteroskedasticity and parameter heterogeneity, respectively. In each figure, MAB, MARF, and GUM are presented by circle-, and star-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 heterogeneity. Yet the statistical learning methods are immune to heterogeneity, since it has been acknowledged and treated during the classification process.

Figure 3: Risk Comparison under Different Scenarios



In summary, the results from the Monte Carlo experiments suggest that hybrid strategies may be beneficial when there is 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.[19] In contrast, recursive partitioning methods provide equal weights across

[19] Appendix C.2 presents Monte Carlo evidence that splits in trees occur at different locations and there is more variation in outcomes in the final leaves with heteroskedastic data relative to homoskedastic data.

observations ruling out heterogeneity within groups.[20]

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. We extend the analysis in Lehrer and Xie (2017) that concentrated solely on movies with budgets ranging from 20 to 100 million dollars and consider the full suite of films released during this period.[21] With the assistance of the IHS film consulting unit the characteristics of each film were characterized by a series of indicator variables to describe the film's genre,[22] the rating of a film's content provided by the Motion Picture Association of America's system,[23] 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 examine the initial demand by using the actual opening weekend box office ($n = 178$) and total sales of both DVD and Blu-Rays ($n = 143$) upon initial release.

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

[20]The theoretical benefits of most model screening methods relate to efficiency. Appendix F.4 presents evidence that model screening approaches and model averaging or Lasso methods that additionally consider heteroskedasticity do not seem to perform differently whatever the source of heteroskedasticity, and in practice yield minimal gains to approaches that treat the data as homoskedastic.

[21]Movies with budgets above 100 million dollars are usually regarded as "Blockbusters" and many ''Arthouse" movies usually have budgets below 20 million dollars. Further details on the data collection are provided in subsection E of the Appendix.

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

[23]Specifically, films in our sample were assigned ratings of PG, PG13, and R. There are very few movies in our data set that were given a G rating.

al. (2012) that involves textual analysis of movie titles and movie key words.[24] 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.[25] 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 just proxy for popularity.[26]

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.[27]

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 in the theatre for 14 weeks and played on roughly 3000

[24]This algorithm 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

[25]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.

[26]We consider both measures since 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 and Gopinath, Chintagunta, and Venkataraman (2013) study the effects of blogs and advertising on local-market movie box office performance.

[27]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.

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

screens. Not surprisingly, given trends in advertising, the volume of Tweets increases 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 methods that combine econometrics with machine learning algorithms to capture richer patterns of heterogeneity. Table 2 lists each estimator analyzed in the exercise and the 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 the estimates from each strategy and then forecast the outcomes for the evaluation set. With these forecasts, we evaluate each of the forecasting strategies by calculating mean squared forecast error (MSFE) and 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

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 (2018) 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
Panel E: Newly Proposed Hybrid Methods[a]		
(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

[a] The Appendix also contains a section examining a hybrid M5′ algorithm.

following model screening procedure based on the GETS method to reduce the set of potential candidate models for model selection and model averaging methods. First, based on the OLS results presented in table A4, we restrict that each potential model contains a constant term and 7 (11) relatively significant parameters for open box office (movie unit sales). Second, 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.[28] While this restriction that rules out many potential candidate model may appear severe, it has been found in numerous applications including Lehrer and Xie (2017), that only a handful of models account for more than 95% of the total weight of the model averaging estimate.[29] Last, the tuning parameter for Lasso strategies was chosen to fix the number of explanatory variables selected (i.e. OLS_{10} indicates OLS with 10 variables selected by the Lasso).

5 Empirical Results

The two panels of table 3 report the median MSFE and MAFE from the prediction error exercise outlined in the preceding section for the 10 different econometric strategies listed in panel A of table 2. Each row of the table considers a different size for the evaluation set and to ease interpretation all MSFEs and MAFEs are normalized by the MSFE and MAFE of the HRC^p. Panel A of table 3 presents results for forecasting open box office and panel B demonstrates results corresponding to forecasting retail movie unit sales. Notice that for open box office, all remaining entries for MSFE are larger than one, indicating

[28] This approach explores through the whole set of potential models and examine each model using the 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 pre-selection 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.

[29] See appendix F.5 for a detailed discussion of the model averaging weights and top 5 models for both open box office and movie unit sales in our experiment.

inferior performance of the respective estimator relative to HRC^p. In general, the three model averaging approaches and the model selected by AIC perform nearly as well as HRC^p. For movie unit sales, HPMA yields the best results in the majority of experiments. However, the gains from using HPMA in place of PMA appear quite small.

The results in table 3 also stress the importance of social media data for forecast accuracy. Models that ignore social media data (MTV) perform poorly relative to all other strategies. Additional experiments makes clear that both social media measures are needed.[30] In contrast to Lehrer and Xie (2017) we find that the post-Lasso methods listed in table 2,[31] including the double-Lasso method, OLS post Lasso and model averaging post Lasso perform poorly relative to HRC^p in this application.

Table 4 considers the performance of alternative model screening strategies listed in panel B of table 2 relative to HRC^p. We observe small gains in forecast accuracy from model screening relative to the benchmark HRC^p. The hetero-robust methods yields slightly better results than homo-efficient methods for forecasts of box office opening. In contrast, when forecasting retail movie unit sales, the homo-efficient ARMS demonstrates better results than the other screening methods.[32] Taking these findings together with the results contrasting PMA to HPMA table 3 illustrate that there are small gains in practice from using econometric approaches that accommodate heteroskedasticity.[33]

Table 5 demonstrates that are very large gains in prediction efficiency of the recursive partitioning algorithms relative to the benchmark HRC^p. The subscript below RF and

[30] In appendices F.3, F4.1, and F.6, we carried out additional experiments to evaluate the forecast accuracy of alternative strategies with only a single social media measure. In each case, the evidence demonstrates markedly lower degrees of forecast accuracy relative to the corresponding exercise with two measures, thereby providing robust evidence of the need to account for both sentiment and volume. Last, Appendix F.14 presents tables of absolute bias of each strategy that correspond to tables 3-5.

[31] 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. Full details on how these estimators are implemented is available in appendix D.6.

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

[33] 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. This includes (i) the difference in the performance between PMA and HRC^p in table 3 when forecasting retail movie unit sales, and (ii) the relative improved performance of ARMS presented in table 4.

Table 3: Results of Relative Prediction Efficiency by MSFE and MAFE

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

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

Table 4: Comparing Hetero-robust and Homo-efficient Model Screening Methods

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

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

MARF refer to the number of randomly chosen explanatory variables used to determine a split at each node. For both outcomes when n_E is small, machine learning methods have dominating performance over the HRC_p. Popular approaches such as bagging and random forest greatly outperform the benchmark. However, our proposed MAB has the best performance when evaluating by MSFEs and adding model averaging tends to lead to gains of 10% between bagging and MAB.[34] While regression tree yields the lowest relative MAFE, random forest methods, both conventional and model averaging, have moderate performance in all cases. Note that as n_E increases, all learning methods observe decreases in performance. Last, note that the large gains in performance of all strategies in table 5 relative to the results presented in tables 3 and 4.

A potential explanation for the improved performance of statistical learning approaches relative to all of the econometric strategies is that the full suite of predictors is considered. Recall, that due to computational challenges we undertook model screening to reduce the number of candidate models for model averaging estimators and by so doing reduced the number of predictors. In appendix F.8, we reconsider table 5 where we restrict the set of predictors to be identical for the 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. This suggests that the gains in forecast accuracy are not from allowing a larger dimension of predictor variables, but rather likely are obtained by relaxing the linearity assumption imposed by the econometric estimator considered when constructing candidate models.

Table 5 compares the performance of our hybrid strategies to the suite of advanced machine learning strategies listed in panel D of Table 2. We continue to find improved performance of our hybrid strategy relative to these alternative algorithms with the potential sole exception of HBART in the box office opening MSFE scenario, which exhibit marginally smaller MSFE and MAFE for retail movie unit sales. However, HBART is computationally expensive and takes over a week to yield results, roughly three to four

[34] In appendix F.9, we present results from the SPA test of Hansen (2005) that provide significant evidence of the superior predictive ability of the MAB method over the other ML algorithms considered.

Table 5: Results of Relative Prediction Efficiency Between Machine Learning and Model Averaging Learning

n_E	BART	BART_{BMA}	HBART	BOOST	Reg.Tree	Bagging	Random Forest			MAB	Random Forest			Benchmark
							RF_{10}	RF_{15}	RF_{20}		MARF_{10}	MARF_{15}	MARF_{20}	
Panel A: Open Box Office														
Mean Squared Forecast Error (MSFE)														
10	0.5853	0.5561	**0.4575**	0.5606	0.6142	0.5883	0.5504	0.5755	0.5313	0.5066	0.5628	0.5356	0.5519	1.0000
20	0.7333	0.7211	**0.5903**	0.6869	0.8834	0.8622	0.7967	0.7901	0.8157	0.7315	0.7898	0.7787	0.8272	1.0000
30	0.7616	0.7407	**0.6562**	0.7685	1.0214	0.8589	0.8654	0.8353	0.8476	0.7531	0.8467	0.8694	0.8802	1.0000
40	0.8251	0.7652	**0.7302**	0.8899	1.3120	0.9833	0.9870	0.9395	0.9994	0.9145	0.9531	1.0348	0.9915	1.0000
Mean Absolute Forecast Error (MAFE)														
10	0.7395	0.7152	0.6416	0.6278	0.6792	0.7036	0.7000	0.7018	0.6652	**0.6232**	0.6940	0.6742	0.6703	1.0000
20	0.8182	0.7866	0.7120	0.6983	0.7443	0.7690	0.7665	0.7513	0.7474	**0.6955**	0.7607	0.7495	0.7444	1.0000
30	0.8144	0.8001	0.7452	0.7116	0.7897	0.7789	0.7899	0.7655	0.7625	**0.7042**	0.7734	0.7733	0.7654	1.0000
40	0.8588	0.8432	0.7809	0.7855	0.8390	0.8080	0.8245	0.8084	0.8072	**0.7625**	0.8052	0.8157	0.8113	1.0000
Panel B: Movie Unit Sales														
Mean Squared Forecast Error (MSFE)														
10	0.9813	0.9251	0.7363	1.0063	1.1288	0.8299	0.9163	0.8678	0.8780	**0.7307**	0.8849	0.9168	0.8674	1.0000
20	1.0080	0.9461	0.7755	1.1014	1.0547	0.8563	1.0581	0.8900	0.9551	**0.7009**	0.9559	1.0564	0.8884	1.0000
30	1.0098	0.9548	0.8079	1.1476	1.1610	0.9610	1.1687	0.9820	1.0682	**0.7494**	1.0677	1.1702	0.9849	1.0000
40	1.0951	0.9547	0.8687	1.3274	1.2739	1.0085	1.1828	1.0456	1.1065	**0.8626**	1.1061	1.1832	1.0391	1.0000
Mean Absolute Forecast Error (MAFE)														
10	0.8984	0.8661	0.7874	0.7764	0.8407	0.8294	0.9109	0.8479	0.8625	**0.7461**	0.8727	0.9098	0.8493	1.0000
20	0.9463	0.8976	0.8355	0.8316	0.8453	0.8302	0.9334	0.8622	0.8791	**0.7564**	0.8775	0.9313	0.8611	1.0000
30	0.9779	0.9444	0.8590	0.9055	0.8924	0.8668	0.9690	0.8974	0.9225	**0.7954**	0.9205	0.9722	0.8968	1.0000
40	1.0029	0.9991	0.8996	0.9910	0.9361	0.8914	0.9817	0.9191	0.9455	**0.8211**	0.9456	0.9805	0.9181	1.0000

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

times as long as the hybrid strategy.

Briefly, we believe the improved performance of the hybrid strategies relative to BART and boosting 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 involves 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 1 plus the number of observations in the respective terminal leaf. Last, the hybrid approach is quite flexible and can be applied with other algorithms including M5' (see Appendix C for additional evidence and intuition) that partition the data into subgroups.

5.1 Relative Importance of the Factors

While recursive partitioning algorithms were developed to make predictions and not understand the underlying process of how predictors correlate with outcomes, strategies have since been developed to identify which predictor variables are the most important in making forecasts.[35] The importance of each predictor variable is first computed at the tree level, and the scores are averaged across all trees to obtain the final, global importance score for the variable.[36] The most important variables are the ones leading to the

[35] 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.

[36] 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 can be used to evaluate the tree 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 sample and the prediction error

greatest losses in accuracy.

We calculate variable importance scores using the MAB and MARF strategies where we include and exclude the social media variables as predictors.[37] Table 6 reports the top 10 most important predictors for open box office and movie unit sales in panels A and B, respectively. The results with both strategies reinforce the importance of social media data and volume related variables are found to have a greater association with revenue outcomes than sentiment measures. Further, the predetermined budget and screens as well as weeks in theatre are important predictors. Taken together, these results suggest that the amount of social media buzz is more important than the emotional content when forecasting revenue outcomes.[38]

To examine whether sentiment plays a larger role for small budget films that may benefit more from word of mouth or critical reviews, we calculated variable importance scores for films located in different budget quartile. The results are presented in table 7. 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. In summary, the evidence in this study continues to point to the inclusion of both social media measures and different forecasting strategies yield different rankings of the importance of each measure.

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.

[37] 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. However, it should also be noted that the finding in Strobl et al. (2008) were not replicated in Genuer, Poggi, and Tuleau-Malot (2010).

[38] 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 tables 3 and 4

Table 6: Relative Importance of the Predictors

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

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

6 Conclusion

The film industry is characterized by substantial uncertainty and De Vany and Walls (2004) report that only 22% of films 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 algorithms developed in the machine learning literature for data mining applications. Using data from the film industry we find significant gains in forecast accuracy from using recursive partitioning strategies instead 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 each leaf subgroup cre-

Table 7: Heterogeneity in the Relative Importance of Predictors by Film Budget

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

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.

ated by a statistical learning algorithm. Our empirical investigation first demonstrates significant gains in forecast accuracy from the proposed hybrid strategy. Second, 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 recursive partitioning 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 6 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 on the structure of the specific problem under consideration and is generally unknown ex-ante to the analyst. Yet, we argue that our proposed hybrid strategy may add significant value since heteroskedastic data is the norm in the real world. After all, our findings that irrespective of the optimization strategy used to build trees, gains in forecast accuracy are achieved from using model averaging in place of either a local constant or linear regression model, reinforcing the potential practical value of using a hybrid strategy.

Future work is 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.

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.

CHESHER, A. (1984): "Testing for Neglected Heterogeneity," *Econometrica*, 52(4), 865–872.

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.

CUI, R., S. GALLINO, A. MORENO, AND D. J. ZHANG (2018): "The Operational Value of Social Media Information," *Production and Operations Management*, 27(10), 1749–1769.

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.

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.

LEHRER, S. F., T. XIE, AND X. ZHANG (2018): "Twits versus Tweets: Does Adding Social Media Wisdom Trump Admitting Ignorance when Forecasting the CBOE VIX?," *Working Paper*.

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 (2018): "Heteroscedastic BART Via Multiplicative Regression Trees," *Working Paper*.

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.

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 (2017): "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, forthcoming.

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

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

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.