NBER WORKING PAPER SERIES

BOX OFFICE BUZZ: DOES SOCIAL MEDIA DATA STEAL THE SHOW FROM MODEL UNCERTAINTY WHEN FORECASTING FOR HOLLYWOOD?

Steven Lehrer
Tian Xie

Working Paper 22959
http://www.nber.org/papers/w22959

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2016

We wish to thank an anonymous referee, the editor (Bryan S. Graham), seminar participants at the Canadian Econometrics Study Group (CESG) 31st Annual Meeting, Hubei Province Quantitative Economics Society 2014 Annual Conference, NYU Shanghai 2014 Symposium on Data Science and Applications, Wuhan University, Xiamen University, and Chinese Academy of Sciences for helpful comments and suggestions. Lehrer wishes to thank SSHRC for research support. We are responsible for all errors. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Box Office Buzz: Does Social Media Data Steal the Show from Model Uncertainty When Forecasting for Hollywood?

Steven Lehrer and Tian Xie

NBER Working Paper No. 22959
December 2016
JEL No. C52,C53,M21

ABSTRACT

Substantial excitement currently exists in industry regarding the potential of using analytic tools to measure sentiment in social media messages to help predict individual reactions to a new product, including movies. However, the majority of models subsequently used for forecasting exercises do not allow for model uncertainty. Using data on the universe of Twitter messages, we use an algorithm that calculates the sentiment regarding each film prior to, and after its release date via emotional valence to understand whether these opinions affect box office opening and retail movie unit (DVD and Blu-Ray) sales. Our results contrasting eleven different empirical strategies from econometrics and penalization methods indicate that accounting for model uncertainty can lead to large gains in forecast accuracy. While penalization methods do not outperform model averaging on forecast accuracy, evidence indicates they perform just as well at the variable selection stage. Last, incorporating social media data is shown to greatly improve forecast accuracy for box-office opening and retail movie unit sales.

Steven Lehrer
School of Policy Studies
and Department of Economics
Queen's University
Kingston, ON K7L 3N6
CANADA
and NBER
lehrers@queensu.ca

Tian Xie
Wang Yanan Institute for Studies in Economics
Department of Finance
MOE Key Lab of Econometric
Xiamen, Fujian 361005, China
xietian001@hotmail.com

An online appendix is available at http://www.nber.org/data-appendix/w22959

1 Introduction

The power of social media was trumpeted when the citizens of Libya, Egypt and other Middle Eastern countries deposed their own government leaders and autocrats, in large part through social media utilization. While this may have heightened the awareness for those in governments across the world, businesses in every industry worldwide have been well aware of the potential of big data for conducting forecasts. Proponents in the burgeoning field of big data analytics often argue that to analyze data extracted from the social web,[1] software tools from fields such as predictive analytics and data mining, including penalization methods should be used. In this paper, we contrast the performance of penalization methods with computationally feasible econometric methods in demand forecasting using data from both the movie industry and social web.

One of the main challenges when comparing empirical approaches via demand forecasting exercises that exists even in the absence of social media data, is the absence of a unique well-accepted theory to guide model specification. We argue in these situations, it is better to simply accept that there is model uncertainty, rather than make difficult ad hoc decisions on the choice of parameters, which as a consequence leads to different potential models. Least squares model averaging provides a means to solve model uncertainty. Unlike the well-known AIC method which selects only one "winning" model from the set of approximation models, a model averaging estimator generates a weighted average model using all of the approximation models.

To select the weights for a model averaging estimator, we additionally extend the results of Hansen (2014) to develop an easy to implement econometric strategy that both minimizes asymptotic risk and is computationally efficient.[2] Further, motivated by the work of Bel-

[1] One of the main challenges researchers in this area face is determining the content from the approximately 350 million tweets and 6 billion Facebook messages per day. Appendix B discusses the tremendous growth in academic circles in using data extracted from social media to analyze the economy with these tools.

[2] This extension allows our empirical strategy to utilize the least squares model averaging estimator of

loni and Chernozhukov (2013) we introduce the idea of model averaging post LASSO.[3] By using the LASSO in a first step to reduce the dimensionality of the variables to include in subsequent models, computational savings are obtained, and this should reduce the bias of LASSO and thus possibly its risk.[4]

To compare the performance of alternative estimators, model selection criteria, and model averaging strategies, we use data from the film industry since it remains one of the most active industries in devoting resources to marketing departments to influence consumer sentiment towards their products on the social web.[5] Our empirical results suggest that there are substantial additional gains achieved by allowing for model uncertainty in the analysis. We find that there are gains from using model averaging after the LASSO relative to using OLS post LASSO or the LASSO in isolation. However, the LASSO only outperforms other variable selection and model selection strategies considered in this paper when forecasting retail movie unit sales. This indicates that variable selection mistakes do occur when certain potentially irrelevant regressors are eliminated from the potential set used to construct models and suggests that the LASSO penalty might be too strong. Thus, we contribute to the field of big data analytics by not just developing and applying new model averaging estimators to conduct forecasts, but also illustrate that supervised learning algorithms do not always outperform econometric approaches in variable and model selection.

This paper is organized as follows. Section 2 describes the data set we are using and how

Xie (2015) that computes empirical weights through numerical minimization of a prediction model averaging (PMA) criterion. The PMA method has been shown in Xie (2015) to be (i) asymptotically optimal in the sense of achieving the lowest possible mean squared error, which applies both to nested and to non-nested approximation models; and (ii) to exhibit very good finite sample performance, particularly when the sample size is small.

[3]The LASSO and related methods have received enormous attention in big data analytics since the seminal work of Tibshirani (1996) since this estimator can be applied when the number of regressors exceeds the number of observations.

[4]There are many different risk functions (e.g. absolute error loss, Kullback-Leiber loss, L_p loss, among others) used to evaluate how good a prediction obtained from a given estimator is. In this paper, we follow the general practice of using mean squared error (MSE) loss and refer to it as commonly in the text.

[5]Moretti (2011) provides convincing evidence that not only is social learning an important determinant of sales in the movie industry, but that the effects of positive buzz on revenue are more pronounced for consumers with larger social networks.

sentiment regarding films is measured. Section 3 details the least squares model averaging estimator and our extensions. To measure the relative prediction efficiency of the model averaging method relative to other model specification methods commonly employed to make demand forecasts, we use a simulation based exercise to compare the accuracy of out of sample forecasts for each strategy. Our empirical results are presented and discussed in Section 4. Our main findings are that while social media data can improve box office predictions for Hollywood studios irrespective of the method employed, model uncertainty appears important for this industry. Section 5 presents our conclusions. This paper provides a clear illustration of the potential benefits for those in data science of collaborating with applied econometricians,[6] who have a long history of developing estimable models of human behavior.

2 Data Description

We collected data on all movies released in North America between October 1, 2010 and June 30, 2013 with budgets ranging from 20 to 100 million dollars. The IHS film consulting unit provided information on the characteristics of each film including the genre, the rating,[7] budget excluding advertising and both the pre-determined number of weeks and screens the film studio forecasted six weeks prior to opening that the specific film will be in theaters. In our analysis, we consider two measures of retail sales of films that differ on the timing of consumer purchases. We examine initial demand using opening weekend box office and total sales of both DVD and Blu-Rays upon initial release.

Purchasing intentions are measured from the universe of Twitter messages by calculating

[6] We concur with Einav and Levin (2014) that interpreting social media data is quite challenging and in the absence of collaborating with researchers experienced in analyzing source of plausible identifying variation, this limitation will remain an important feature of incorporating this data.

[7] Film ratings are assigned by the Motion Picture Association of America and there are very few G rated movies in our sample. See Table 1 for the list of film genres utilized in our analysis. Note our sample contains few sequels and Appendix E.2 demonstrates seasonality and sequels do not play a significant role.

Table 1: Summary Statistics

Variable	Open Box		Movie Unit	
	Mean	Std.Dev	Mean	Std.Dev
Genre				
Action	0.3723	0.4860	0.3671	0.4851
Adventure	0.1596	0.3682	0.1646	0.3731
Animation	0.0745	0.2639	0.0759	0.2666
Comedy	0.4255	0.4971	0.4304	0.4983
Crime	0.2660	0.4442	0.2532	0.4376
Drama	0.3404	0.4764	0.3671	0.4851
Family	0.0638	0.2458	0.0759	0.2666
Fantasy	0.0745	0.2639	0.0633	0.2450
Mystery	0.0851	0.2805	0.0886	0.2860
Romance	0.1277	0.3355	0.1013	0.3036
Sci-Fi	0.0957	0.2958	0.1013	0.3036
Thriller	0.2447	0.4322	0.2405	0.4301
Rating				
PG	0.1489	0.3579	0.1646	0.3731
PG13	0.3723	0.4860	0.3671	0.4851
R	0.4681	0.5017	0.4557	0.5012
Core Parameters				
Budget (in million)	49.9840	20.3961	51.1076	20.7681
Weeks	13.7826	5.4631	13.9747	5.7042
Screens (in thousand)	2.9967	0.5200	2.9751	0.5473
Sentiment				
T-21/-27	73.6871	3.0737	73.4635	3.3572
T-14/-20	74.0545	2.4099	73.9789	2.5458
T-7/-13	74.3415	1.7985	74.2909	1.8175
T-4/-6	74.2604	2.0787	74.1940	2.1580
T-1/-3	74.2972	2.0516	74.2246	2.1297
T+0			74.3067	2.1654
T+1/+7			74.4563	1.8822
T+8/+14			73.8944	2.9500
T+15/+21			74.1226	2.5739
T+22/+28			74.3700	1.9751
Volume				
T-21/-27	0.1775	0.9293	0.2011	1.0128
T-14/-20	0.1909	0.9055	0.2149	0.9867
T-7/-13	0.2152	0.8965	0.2385	0.9764
T-4/-6	0.2524	1.1280	0.2830	1.2289
T-1/-3	0.4130	1.1025	0.4528	1.1980
T+0			1.2025	3.1132
T+1/+7			0.6248	1.3385
T+8/+14			0.3328	1.0752
T+15/+21			0.2547	0.9562
T+22/+28			0.2116	0.9596

the sentiment specific to a particular film using an algorithm developed by Hannak et al. (2012). This algorithm involves textual analysis of movie titles and movie key words. In a Twitter message mentioning a specific film title or key word, sentiment is calculated by examining the strength of the emotion words and icons contained within.[8] The overall sentiment score for each film is a weighted average of the sentiment of the scored words

[8]Full details on the external validity and how the sentiment index for each film is calculated are provided in Appendix A.1. This appendix summarizes evidence from evaluations of the sentiment inference algorithm, demonstrating a high degree of accuracy in sentiment prediction. For open box office, the volume of Twitter message is 1,100,439; for DVD, this number is 3,433,413 messages.

in all the messages associated, and indicates the propensity for which there is a positive emotion tweet related to that movie. Since opinions regarding a film likely vary over time, we measured volume of Twitter messages and calculated sentiment over different time periods.[9]

Summary statistics for our sample are presented in Table 1. Since certain movies were not released in either DVD or Blu-Ray format, the total number of observations for the DVD and Blu-Ray sales is slightly smaller than that for open box office. Notice that the mean budget of films analyzed for each outcome is approximately 50 million. On average, these films were in the theater for 14 weeks and played on roughly 3000 screens during the opening weekend. Not surprisingly, given trends in advertising, the volume of Tweets increases sharply close to the release date, peaking that day and decreasing steadily afterwards. We consider volume separate from sentiment in our analyses since the latter may capture perceptions of quality, whereas volume likely just proxies for popularity.[10]

3 Model Averaging

Researchers who ignore model uncertainty implicitly assume their selected model is the "true" one that generated the data $(y_i, \boldsymbol{x_i}) : i = 1, ..., n$, where y_i and $\boldsymbol{x_i} = [x_{i1}, x_{i2}, ...]$ are real-valued.[11] We assume the data generating process for an outcome y_i is given as

$$y_i = \mu_i + u_i, \tag{1}$$

where $\mu_i = \sum_{j=1}^{\infty} \beta_j x_{ij}$, $\mathbb{E}(u_i|\boldsymbol{x_i}) = 0$ and $\mathbb{E}(u_i^2|\boldsymbol{x_i}) = \sigma^2$. Since researchers and analysts in the film industry often have little knowledge of the true data generating process, they

[9] Suppose the movie release date is T, we calculate sentiment in ranges suggested by the IHS film consulting unit. For example, for a typical range, T–a/–b denotes a days to b days before date T. Similarly, T+c/+d means c days to d days after date T, which are additionally used for forecasting the retail unit sales.

[10] Intuitively, if herd behavior is important, volume drives box office revenue, whereas Chintagunta et al. (2010) and Liu (2006) suggest sentiment may affect revenue for those who make decisions based on quality.

[11] While there is a burgeoning theoretical literature, Breiman and Spector (1992) describes the certitude that many researchers have with respect to model uncertainty as the "quiet scandal" in statistical research.

generally select one model from a sequence of linear approximation models $m = 1, 2, ..., M$. An approximation model m using $k^{(m)}$ regressors belonging to $\boldsymbol{x_i}$ such that

$$y_i = \sum_{j=1}^{k^{(m)}} \beta_j^{(m)} x_{ij}^{(m)} + u_i^{(m)} \quad \text{for } i = 1, ..., n, \tag{2}$$

where $\beta_j^{(m)}$ is a coefficient in model m and $x_{ij}^{(m)}$ is a regressor in model m. Approximation models can be either nested or non-nested and model averaging approaches first involve solving for the smoothing weights across the set of approximation models based on a specific criterion.

Formally, the DGP (1) and approximation model (2) can be represented in matrix forms: $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{u}$ and $\boldsymbol{y} = \boldsymbol{X}^{(m)}\boldsymbol{\beta}^{(m)} + \boldsymbol{u}^{(m)}$, where $\boldsymbol{y}$ is $n \times 1$, $\boldsymbol{\mu}$ is $n \times 1$, $\boldsymbol{X}^{(m)}$ is $n \times k^{(m)}$ with the ij^{th} element being $x_{ij}^{(m)}$, $\boldsymbol{\beta}^{(m)}$ is $k^{(m)} \times 1$ and $\boldsymbol{u}^{(m)}$ is the error term for model m. For an approximation model m, the least squares estimate of $\boldsymbol{\mu}$ from model m can be written as $\hat{\boldsymbol{\mu}}^{(m)} = \boldsymbol{P}^{(m)}\boldsymbol{y}$, where $\boldsymbol{P}^{(m)}$ is a projection matrix. Let $\boldsymbol{w} = \left[w^{(1)}, ..., w^{(M)}\right]^\top$ be a weight vector in the unit simplex in $\mathbb{R}^M$, $\boldsymbol{H_M} \equiv \left\{\boldsymbol{w} \in [0,1]^M : \sum_{m=1}^M w^{(m)} = 1\right\}$, which is a continuous set. We define the model average estimator of $\boldsymbol{\mu}$ as

$$\boldsymbol{\mu}(\boldsymbol{w}) \equiv \sum_{m=1}^{M} w^{(m)} \hat{\boldsymbol{\mu}}^{(m)} = \sum_{m=1}^{M} w^{(m)} \boldsymbol{P}^{(m)} \boldsymbol{y}. \tag{3}$$

By defining the weighted average projection matrix $\boldsymbol{P}(\boldsymbol{w})$ as $\boldsymbol{P}(\boldsymbol{w}) \equiv \sum_{m=1}^M w^{(m)} \boldsymbol{P}^{(m)}$, equation (3) can be simplified to $\boldsymbol{\mu}(\boldsymbol{w}) = \boldsymbol{P}(\boldsymbol{w})\boldsymbol{y}$. Thus, the effective number of parameters to be solved is defined as $k(\boldsymbol{w}) \equiv \sum_{m=1}^M w^{(m)} k^{(m)}$.[12]

The prediction model averaging (PMA) estimator of Xie (2015) can be understood as the model averaging analog of the prediction criterion of Amemiya (1980). Following Xie

[12]Note that $k(\boldsymbol{w})$ is not necessarily an integer and is a weighted sum of the $k^{(m)}$.

(2015), the vector of empirical weight $\hat{\boldsymbol{w}}$ is the solution to

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in\boldsymbol{H}_M}{\arg\min}\, \mathrm{PMA}_n(\boldsymbol{w}) = \underset{\boldsymbol{w}\in\boldsymbol{H}_M}{\arg\min}\, \big(\boldsymbol{y}-\boldsymbol{\mu}(\boldsymbol{w})\big)^\top \big(\boldsymbol{y}-\boldsymbol{\mu}(\boldsymbol{w})\big)\left(\frac{n+k(\boldsymbol{w})}{n-k(\boldsymbol{w})}\right), \tag{4}$$

where $\boldsymbol{\mu}(\boldsymbol{w})$ and $k(\boldsymbol{w})$ are defined above. The PMA estimator is asymptotically optimal in the sense of achieving the lowest possible mean square error.[13]

3.1 Strategies that Reduce Asymptotic Risk

Recently, Hansen (2014) derived conditions under which the asymptotic risk of an averaging estimator is globally smaller than the unrestricted least-squares estimator. In Appendix C.2 we supplement Theorem 3 in Hansen (2014), allowing this finding to be applied to a broader set of least squares model averaging estimators including the PMA estimator. Imposing Assumptions 1 to 6 and Lemmas 1 and 2 defined in Appendix C.1 permits us to state the following theorem:

Theorem 1 *Let Assumptions 1 – 6 hold. We have*

$$R(\hat{\boldsymbol{\beta}}_A, \boldsymbol{\beta}) < R(\hat{\boldsymbol{\beta}}_{LS}, \boldsymbol{\beta}), \tag{5}$$

where $\hat{\boldsymbol{\beta}}_{LS}$ and $\hat{\boldsymbol{\beta}}_A$ are defined in equations (A1) and (A3) respectively.

Theorem 1 indicates that, under relatively mild restrictions, if we group regressors into sets of four or larger (Assumption 6), the averaging estimator $\hat{\boldsymbol{\beta}}_A$ always yields smaller asymptotic risk than the unrestricted least-squares estimator $\hat{\boldsymbol{\beta}}_{LS}$. By grouping regressors, the total number of potential models is reduced, leading to gains in computational efficiency.

[13] See Xie (2015) for a detailed proof. For computational convenience, we can re-express the PMA in (4) as $\mathrm{PMA}_n(\boldsymbol{w}) = \boldsymbol{w}^\top \hat{\boldsymbol{U}}^\top \hat{\boldsymbol{U}}\boldsymbol{w}\left(\frac{n+\boldsymbol{k}^\top\boldsymbol{w}}{n-\boldsymbol{k}^\top\boldsymbol{w}}\right)$. where $\hat{\boldsymbol{U}}$ is an $n\times M$ matrix consisting of $n\times 1$ vectors of residuals for each of the m models (i.e. $\hat{\boldsymbol{U}} \equiv [\hat{\boldsymbol{u}}^{(1)}, \hat{\boldsymbol{u}}^{(2)}, ..., \hat{\boldsymbol{u}}^{(M)}]$) and $\boldsymbol{k}$ is a $M\times 1$ vector of the number of parameters from each model, such that $\boldsymbol{k} \equiv [k^{(1)}, k^{(2)}, ..., k^{(M)}]^\top$.

A second strategy to potentially reduce asymptotic risk is using model averaging post LASSO, which is in the spirit of Belloni and Chernozhukov (2013).[14] In the first step, a LASSO estimator selects the explanatory variables via a shrinkage procedure that predominantly eliminates potentially irrelevant variables. In the second step, LASSO coefficient estimates are dropped and the variables selected by LASSO are used with any model averaging estimator. This procedure allows for both model uncertainty and is more data dependent in conducting variable selection relative to common econometric strategy that require a complete model specification. Further, this proposed estimation strategy yields improvement in computational efficiency relative to traditional model averaging approaches since the LASSO zeros out coefficient values in the first step, thereby reducing the numbers of potential regressors and the total number of potential models used in the second step.[15]

3.2 Assessing Prediction Efficiency

Following Hansen and Racine (2012), the relative prediction efficiency of different estimators with different sets of covariates is assessed via an experiment that shuffles the original data with sample n, into a training set of size n_T and an evaluation set of size $n_E = n - n_T$. Using the training set, 11 different estimation strategies are used to make forecasts. For each strategy, we next use the estimates to forecast box office and retail unit sales for the evaluation set. The forecasts from these strategies are then evaluated by calculating mean

[14]The least absolute shrinkage selection operator (LASSO) of Tibshirani (1996) not only estimates regression coefficients but also acts as a variable selection device. LASSO coefficients are the solutions to an l_1-optimization problem (see Appendix D.2 for details) that minimizes the sum of the OLS objective function with a penalty for model size, which is the sum of the absolute values of the estimated regression coefficients. Intuitively, the estimator shrinks several of the non-zero coefficient estimates towards zero to satisfy a sparsity condition; at the cost of potentially introducing shrinkage bias. Belloni and Chernozhukov (2013) suggest discarding the LASSO estimates and using OLS post LASSO to estimate the coefficients on the remaining variables.

[15]Future theoretical and Monte Carlo research is needed to understand the optimal penalty when using the LASSO for variable selection prior to model averaging. It is well-established that by zeroing out potentially relevant regressors, the LASSO may generate bias in the LASSO coefficients. Therefore, with model averaging post LASSO this would also result in having fewer potential approximation models.

squared forecast error (MSFE) and mean absolute forecast error (MAFE):

$$\begin{aligned} \text{MSFE} &= \frac{1}{n_E}(\boldsymbol{y}_E - \boldsymbol{x}_E\hat{\boldsymbol{\beta}}_T)^\top(\boldsymbol{y}_E - \boldsymbol{x}_E\hat{\boldsymbol{\beta}}_T), \\ \text{MAFE} &= \frac{1}{n_E}\left|\boldsymbol{y}_E - \boldsymbol{x}_E\hat{\boldsymbol{\beta}}_T\right|^\top \boldsymbol{\iota}_E, \end{aligned}$$

where $(\boldsymbol{y}_E, \boldsymbol{x}_E)$ is the evaluation set, n_E is the number of observations of the evaluation set, $\hat{\boldsymbol{\beta}}_T$ is the estimated coefficients by a particular model based on the training set, and $\boldsymbol{\iota}_E$ is a $n_E \times 1$ vector of ones. The 11 estimation strategies include

(i) a general unrestricted model (GUM) using all the independent variables available,

(ii) a general unrestricted model that ignores social media data (MTV),

(iii) a model selected by Hendry and Nielsen (2007) general to specific method (GETS),

(iv) a model selected using the Akaike Information Criterion Method (AIC),

(v) the model selected using Mallows model averaging (MMA) proposed by Hansen (2007),

(vi) the model selected by Jackknife model averaging (JMA) (Hansen and Racine, 2012),[16]

(vii) the model selected by group MMA method developed in Hansen (2014) ($\text{MMA}_{g1,g2}$),

(viii) the model selected by group PMA estimator, developed in Section 3.1 ($\text{PMA}_{g1,g2}$),

(ix) the model selected using the (PMA) estimator developed in equation (4),

(x) the OLS post LASSO estimator of Belloni and Chernozhukov (2013) with 10, 12, and 15 explanatory variables selected by the LASSO ($\text{OLS}_{10,12,15}$),

(xi) the PMA model averaging post LASSO estimation strategy proposed in Section 3.1 with 10, 12, and 15 explanatory variables selected byc the LASSO ($\text{PMA}_{10,12,15}$).

The above exercise is carried out 10,001 times for different n_E. For strategies (v) to (ix), a simplified version of the automatic general-to-specific approach of Campos, et al.

[16] We also tried the more generalized JMA by Zhang, Wan, and Zou (2013) and the results are similar.

(2003) was used for model screening.[17] Last, as detailed in Appendix D.1, when examining the empirical performance of strategies (vii) and (viii), we group regressors based on either economic intuition (g_1) or statistical logic (g_2).

4 Results and Discussion

Table 2 reports the median MSFE and MAFE from the relative out-of-sample prediction efficiency experiment for evaluation sets $n_E = 10, 20, 30, 40$ for open box office and $n_E = 10, 15, 20, 25$ for movie unit sales.[18] To ease interpretation, we normalize the MSFEs and MAFEs, respectively, by the MSFE and MAFE of the PMA. For open box office, all entries are larger than one indicating inferior performance of the respective estimator relative to PMA.

There are several findings worth stressing. First, when comparing the results of PMA with results of MTV (models without Twitter data), we see that the prediction efficiency increases by more than 147% using MSFE as criterion when $n_E = 10$. As n_E increases, the prediction efficiency of using PMA improves even more (204% when $n_E = 40$). This result provides the first piece of evidence demonstrating the importance of using social media data in this forecasting exercise.[19]

Second, the results in Table 2 demonstrate the importance of considering model uncertainty as seen when comparing GUM (no model uncertainty) to PMA. The prediction

[17] This approach examines each of the 16,777,216 potential models for open box office and 4,294,967,296 potential models for retail movie sales by estimating the p-values for tests of statistical significance. If the maximum of these p-values exceeds our benchmark values (0.1 for open box office and 0.65 for retail movie sales), we exclude the corresponding model. After pre-selection, we respectively obtain 95 and 56 models for open box office and movie unit sales. Note Wan, Zhang, and Zou (2010) showed that screening methods are necessary in practice to remove some poor models prior to model averaging. Appendix E.5 presents a comparison with the backward elimination procedure of Claeskens et al. (2006).

[18] The size of evaluation set for retail movie unit sales is smaller because we have fewer observations since some films were not released on DVD/Blu-Rays.

[19] Additional results demonstrating the importance of social media data including the need to have two distinct measures, are presented in Appendix E.1.

Table 2: Results of Relative Prediction Efficiency

n_E	GUM	MTV	GETS	AIC	MMA	JMA	PMA_{g1}	PMA_{g2}	MMA_{g1}	MMA_{g2}	OLS_{10}	OLS_{12}	OLS_{15}	PMA_{10}	PMA_{12}	PMA_{15}	PMA
Panel A: Open Box Office																	
								Mean Squared Forecast Error (MSFE)									
10	1.3444	2.4790	1.6738	1.0678	1.0532	1.0701	1.1400	1.1107	1.1466	1.1232	1.1101	1.0476	1.0250	1.0272	1.0229	1.0232	**1.0000**
20	1.5154	2.5949	1.8056	1.0727	1.0681	1.0765	1.2567	1.1859	1.2547	1.1741	1.1313	1.0685	1.0790	1.0598	1.0335	1.0343	**1.0000**
30	1.7602	2.7021	1.8946	1.0714	1.0632	1.0706	1.3863	1.2409	1.3537	1.2107	1.1230	1.0469	1.0491	1.0319	1.0082	1.0304	**1.0000**
40	2.3096	3.0373	2.1219	1.0689	1.0492	1.0514	1.6908	1.3381	1.5025	1.2209	1.1644	1.0859	1.0513	1.0300	1.0269	1.0344	**1.0000**
								Mean Absolute Forecast Error (MAFE)									
10	1.1068	1.5421	1.2757	1.0335	1.0406	1.0425	1.0403	1.0503	1.0450	1.0534	1.0606	1.0587	1.0565	1.0613	1.0484	1.0410	**1.0000**
20	1.1593	1.5924	1.3104	1.0385	1.0353	1.0424	1.0721	1.0653	1.0747	1.0656	1.0732	1.0681	1.0555	1.0729	1.0617	1.0592	**1.0000**
30	1.2274	1.5862	1.3378	1.0352	1.0353	1.0485	1.1048	1.0759	1.0990	1.0698	1.0755	1.0526	1.0375	1.0442	1.0384	1.0427	**1.0000**
40	1.3739	1.5863	1.3978	1.0367	1.0225	1.0296	1.1947	1.1031	1.1460	1.0741	1.0846	1.0573	1.0353	1.0331	1.0331	1.0316	**1.0000**
Panel B: Movie Unit Sales																	
								Mean Squared Forecast Error (MSFE)									
10	3.1730	2.3046	2.3535	1.1175	1.1044	1.1948	1.0565	1.0813	0.9945	**0.9917**	1.0860	1.0383	1.0240	1.0653	1.0102	1.0164	1.0000
15	4.4860	2.1385	2.5779	1.1082	1.0868	1.1045	1.0472	1.0141	0.9904	1.0457	0.9500	0.8439	0.8354	0.8409	0.7788	**0.7642**	1.0000
20	6.8644	2.0439	3.3449	1.1135	1.0761	1.1224	1.0192	1.0728	1.0393	1.0598	0.9655	0.8282	0.8025	0.7321	0.6654	**0.6571**	1.0000
25	12.8320	1.9039	4.6958	1.1058	1.0790	1.0941	1.0099	1.0298	1.0401	1.0924	0.8868	0.8435	0.8078	0.6681	**0.6110**	0.6198	1.0000
								Mean Absolute Forecast Error (MAFE)									
10	1.5002	1.5920	1.5348	1.0656	1.0270	1.0275	1.2717	**0.9750**	1.2110	1.0857	1.1464	1.1749	1.1772	1.1393	1.1280	1.1462	1.0000
15	1.7520	1.5478	1.5887	1.0611	1.0192	1.0215	1.2367	**0.9918**	1.2411	1.0751	1.0937	1.1089	1.1001	1.0415	1.0245	1.0314	1.0000
20	2.1635	1.5193	1.7490	1.0642	1.0124	1.0194	1.1959	1.0518	1.2528	1.0596	1.0702	1.0766	1.0741	1.0171	1.0104	1.0131	**1.0000**
25	2.8765	1.4875	2.0091	1.0632	1.0046	1.0091	1.1824	1.0404	1.2437	1.0428	1.0824	1.0756	1.0587	1.0136	1.0003	**0.9970**	1.0000

Note: Bold numbers indicate the strategies with the best performance in that simulation experiment denoted by the row of the table. The remaining entries provide the ratio of the degree of the respective forecast error metric of the estimator using specific estimation approach denoted in the column relative to results using the PMA method presented in the last column. A complete description of the implementation of different strategies appear in Appendix D.3.

efficiency is improved by 34% when $n_E = 10$ and by 131% when $n_E = 40$. All model selection and model averaging methods in our exercise yield better forecasts than GUM, which indicates the lack of prediction efficiency when ignoring model uncertainty.

Third, all of the different grouping estimators presented in columns PMA_{g1} to MMA_{g2} perform poorly relative to PMA and other model selection and model averaging methods that considers a weighted combination of all potential models. In general, estimators perform better when the grouping is done based on statistical logic (g_2) rather than economic intuition (g_1). Consistent with both Theorem 1 and Theorem 3 of Hansen (2014), the different grouping estimators yield smaller MSFE than GUM, which is the unconstrained least squares estimator.

Fourth, exploring the two empirical strategies that use the LASSO for the variable selection, we first find that irrespective of the evaluation set or number of regressors selected by the first step LASSO, model averaging post LASSO outperforms OLS post LASSO. However, comparing model averaging post LASSO to PMA is quite striking in that PMA outperforms the post LASSO strategy with box office openings but does not do so with retail unit sales. In fact, the gains from model averaging post LASSO in forecasting retail movie unit sales are incredibly large when $n_E \geq 15$. Yet, the worse performance with box office openings suggests that if the true coefficients do not satisfy a strong sparsity condition and are forced to equal zero, this estimator will not have lower MSFE.[20] Since ex-ante, a researcher will neither know the true data generating process nor whether the strong sparsity condition holds, conducting model averaging with different variable selection algorithms appears worthwhile as a minimum to serve as a robustness exercise.

Panel B of Table 2 presents results for retail movie unit sales, where we find strong

[20] In Appendix E.1.2 we repeat this exercise with additional explanatory variables selected by the LASSO and continue to find PMA outperforming model averaging post LASSO. By using the LASSO and forcing one to drop certain coefficients and resulting approximation models, these results suggest that several important models are lost. We should add that ex-ante, we anticipated model averaging post LASSO to perform worse with retail movie unit sales since there are more potential variables and we restricted the LASSO to select a small subset; thereby reducing the variables that satisfy the strong sparsity condition.

evidence (i) supporting using model averaging to deal with model uncertainty when comparing the forecasting results of GUM to PMA; (ii) of gains from including social media data, since MTV yields worse performance than any model averaging and model selection method considered; and (iii) of improved performance from model averaging estimators that use grouping. This contrasts with box office openings, in which PMA was the preferred estimator in all cases. The improvement from PMA to grouping estimators is often quite marginal in forecast accuracy topping out at 3%, but the computational gains are substantial. Last, we observed mixed results on how to optimally group regressors into subsets,[21] suggesting a remaining issue that researchers might confront in practice.[22]

By exploring the variables selected by the LASSO with either outcome, additional evidence of the relative importance of social media measures for forecasting is found. When the LASSO respectively selects 10, 12 and 15 variables for open box office, 4, 4, and 5 of which are social media measures; whereas 5, 7, and 9 are social media measures for retail movie unit sales. This indicates that among the 10 variables with the strongest links to the industry outcomes considered, 40 or 50% of them are obtained from social media, rather than traditional data sources that describe the characteristics of the film itself.[23]

While these results show the practical advantages of using model averaging for forecasts within this industry, there are clear computational costs relative to conventional approaches. Put simply, implementing the model averaging method can be time consuming when the total number of potential models is very large. This is mainly due to the optimization routine irrespective of the software employed. To illustrate, consider the box office opening

[21] For example, we observe that MMA_{g2} offers the best performance when $n_E = 10$ with both MSFE and MAFE criteria, whereas MMA_{g1} has best performance when $n_E = 15$ for MSFE and MAFE cases, and PMA_{g2} has best performance when $n_E = 10, 15$ when using MAFE as the criteria.

[22] Determining the optimal way to group regressors is beyond the scope of this paper. We leave this for future research.

[23] Further analyses in Appendix E.1 demonstrate that while the sentiment variables play a larger role in increasing forecast accuracy, the inclusion of the volume variables does explain substantially more variation in both outcome variables. This difference is not surprising since an individual themselves is not exposed to the full volume of messages on Twitter, just the sentiment within a subset. Thus, sentiment is more likely to influence individual decisions, whereas volume can better predict aggregate outcomes.

weekend example. With our data, there is a total number of 29 potential parameters in the general unrestricted model. Even if we were to fix 5 parameters in every model, it still implies a total of $2^{24} = 16,777,216$ potential models, since each model utilizes different combinations of explanatory variables and estimates the corresponding parameters. Our analysis suggests that researchers should use algorithms from both the machine learning and econometrics literature to determine which of the potential variables and models are reasonable to include.[24]

5 Conclusion

In summary, our evidence suggests that social media data and model uncertainty should equally share the billing on the top of the marquee for Hollywood forecasts. Our empirical exercise provides mixed evidence on the use of the LASSO relative to a more traditional econometric strategy when conducting variable selection. While there are tremendous gains from model averaging post LASSO in forecasting retail movie unit sales, our analysis is suggestive that this strategy critically relies on the strength of the sparsity condition. Thus, we suggest that future researchers verify the robustness of their findings using model averaging approaches that select variables using at least one algorithm from each of the econometrics and data science literature.[25] Since there is a need for forecasts to help inform planning for management and administrators in many industries beyond film, we believe the tools

[24]Further, additional analyses in Appendix E.3 uncovers that only 5 of the thousands of models estimated accounted for over 90% of the resulting PMA estimator. Thus, concerns regarding model selection in empirical practice in this setting may appear small. But in Appendix E.3, the gains in forecast accuracy from PMA to any of these 5 models are shown to be non-trivial; reinforcing the importance of model uncertainty. However, the issue of variable selection appears important since each of the top 5 models presented in Appendix E.3 contain more than 15 variables. Yet, in Table 2, model averaging post LASSO (PMA_{12} and PMA_{15}) outperform PMA with retail movie sales. Since variable selection influences the set of potential approximation models, more guidance is needed to determine whether the traditional LASSO penalty is either too strict or too lenient. We leave this to future research.

[25]In a follow-up paper, Lehrer and Xie (2016) develop and consider additional model screening strategies with both homoskedastic and heteroscedastic data and suggest that heteroscedasticity is likely an important feature of social media data.

developed and illustrated in this paper can help managerial decision making.

References

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

Belloni, A., and V. Chernozhukov (2013): "Least Squares After Model Selection in High-dimensional Sparse Models," *Bernoulli*, 19(2), 521–547.

Breiman, L., and P. Spector (1992): "Submodel Selection and Evaluation in Regression. The X-random Case," *International Statistical Review*, 60(3), 291–319.

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.

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.

Claeskens, G., C. Croux, and J. Venkerckhoven (2006): "Variable Selection for Logit Regression Using a Prediction-Focused Information Criterion," *Biometrics*, 62, 972–979.

Einav, L., and J. Levin (2014): "Economics in the Age of Big Data," *Science*, 346(6210).

Hannak, A., E. Anderson, L. F. Barrett, S. Lehmann, A. Mislove, and M. Riedewald (2012): "Tweetin in the Rain: Exploring Societal-scale Effects of Weather on Mood," *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pp. 479–482.

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

——— (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.

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

Lehrer, S. F., and T. Xie (2016): "Forecasting with Social Media Data in the Presence of Heteroscedasticity," *Working Paper*.

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

Moretti, E. (2011): "Social Learning and Peer Effects in Consumption: Evidence from Movie Sales," *Review of Economic Studies*, 78(1), 356–393.

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

Wan, A. T., X. Zhang, and G. Zou (2010): "Least Squares Model Averaging by Mallows Criterion," *Journal of Econometrics*, 156(2), 277 – 283.

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

Zhang, X., A. T. Wan, and G. Zou (2013): "Model Averaging by Jackknife Criterion in Models with Dependent Data," *Journal of Econometrics*, 174(2), 82–94.