

Machine learning versus econometrics: prediction of box office

Yan Liu & Tian Xie

To cite this article: Yan Liu & Tian Xie (2019) Machine learning versus econometrics: prediction of box office, Applied Economics Letters, 26:2, 124-130, DOI: [10.1080/13504851.2018.1441499](https://doi.org/10.1080/13504851.2018.1441499)

To link to this article: <https://doi.org/10.1080/13504851.2018.1441499>



Published online: 21 Feb 2018.



Submit your article to this journal [↗](#)



Article views: 234



View Crossmark data [↗](#)

ARTICLE



Machine learning versus econometrics: prediction of box office

Yan Liu^a and Tian Xie^{b,c,d}

^aEconomic Development Research Center, Department of Finance, Economics and Management School, Wuhan University, Wuhan, China; ^bWang Yanan Institute for Studies in Economics (WISE), Xiamen University, Xiamen, China; ^cDepartment of Finance, School of Economics, MOE Key Lab of Econometrics, Ministry of Education, Xiamen University, Xiamen, China; ^dFujian Key Lab of Statistical Sciences, Xiamen University, Xiamen, China

ABSTRACT

In this note, we contrast prediction performance of nine econometric and machine learning methods, including a new hybrid method combining model averaging and machine learning, using data from the film industry and social media. The results suggest that machine learning methods have an advantage in addressing short-run noise, whereas traditional econometric methods are better at capturing long-run trend. In addition, once sample heterogeneity is controlled, the new hybrid method tends to strike a right balance in dealing with both noise and trend, leading to superior prediction efficiency.

KEYWORDS

Regression tree; bagging; model averaging; big data; social media sentiment

JEL CLASSIFICATION

C52; C53; M21

I. Introduction

The big data market continues to grow at a fast pace and the introduction of more sophisticated methods to conduct forecasts has been driven mostly by the machine learning literature (Varian 2014). In this article, we conduct a set of empirical experiments to contrast machine learning methods with standard econometric methods in a prediction context, by using data from the film industry in conjunction with social media data.¹ We extend the box office prediction exercises in Lehrer and Xie (2017) by considering eight widely used methods in both machine learning and econometrics literature. In addition, we propose a new hybrid method that combines features from both machine learning and econometric methods, based on the recent work of Hirano and Wright (2017) and Xie (2015).

The empirical exercises show two main results. First, the machine learning methods excel in short horizon prediction on average, whereas the standard econometric methods do better in long horizon. Second, the new hybrid method outperforms on subsamples where heterogeneity is limited, in both short and long horizon. The underlying intuition is the following. Machine learning methods

specialize in dealing with heterogeneity in data structures, and hence work well to capture short-run noise, i.e. nonlinearities or irregularities. In contrast, standard econometric methods are better with long-run trends, i.e. linear or regular patterns. As a result, when forecasting horizon lengthens or data heterogeneity decreases, standard econometric methods gain competitive edges over machine learning methods. By combining merits from two sides, our hybrid method strikes a good balance in both short and long horizon prediction.

In recent literature, a number of studies included social media data and machine learning in their analysis. Antenucci et al. (2014) and Toole et al. (2015) illustrated the potential of using data from the social network to measure economic indicators of labour market activity. Einav and Levin (2014) summarized the opportunities and challenges that confront economists wishing to use these large new data sets obtained from either the social web or administrative records. Mullainathan and Spiess (2017) provided an up-to-date overview on machine learning methods in economics, while Athey and Imbens (2017) demonstrated how machine learning methods can improve the performance of the standard econometric methods.

CONTACT Tian Xie  xietian001@hotmail.com  Rm.B114b, Economics Building, Xiamen University, Xiamen, China, 361005

¹Box office prediction is an important research topic and poses unique challenges for achieving high prediction accuracy. See Liu (2006), Chintagunta, Gopinath, and Venkataraman (2010) and Moretti (2011) for recent references.

This article is organized as follows. Section II describes the data set; Section III presents the prediction methods, including our new hybrid method, and the empirical design; Section IV discusses results; and Section V concludes.

II. Data description

Following Lehrer and Xie (2017), we use data on all movies released in North America between 1 October 2010 and 30 June 2013 with budgets ranging from 20 to 100 million (U.S.) dollars. There are 94 movies in total, and their associated open box offices forms the dependent variable. With help from IHS film consulting unit, we obtain each movies characteristics including genre,² rating, budget (excluding advertising), and both the pre-determined number of weeks and number of screens the movie will be in theatres, based on forecasts by the film studios six weeks prior to opening. From the twitterverse, we compute the sentiment index specific to each film using an algorithm based on Hannak et al. (2012) that involves a textual analysis of movie titles and movie key words. We measure the volume of tweets over different time periods. The summary statistics are in Table 1.

Table 1. Summary statistics.*

Variable	Mean	Variable	Mean
<i>Dependent variable (in million dollars)</i>			
Box office	19.1098		
<i>Independent variables</i>			
<i>Genre</i>		<i>Core</i>	
Action	0.3723	Budget	49.9840
Adventure	0.1596	Weeks	13.7764
Animation	0.0745	Screens	2.9967
Comedy	0.4255	<i>Sentiment[†]</i>	
Crime	0.2660	T-27/-21	73.6871
Drama	0.3404	T-20/-14	74.0545
Family	0.0638	T-13/-7	74.3415
Fantasy	0.0745	T-6/-4	74.2604
Mystery	0.0851	T-3/-1	74.2972
Romance	0.1277	<i>Volume</i>	
Sci-Fi	0.0957	T-27/-21	0.1775
Thriller	0.2447	T-20/-14	0.1909
<i>Rating</i>		T-13/-7	0.2152
PG	0.1489	T-6/-4	0.2524
PG13	0.3723	T-3/-1	0.4130
R	0.4681		

*Median and SD are available upon request.

[†]Suppose the movie release date is T, then T-a/-b denotes average sentiment a days to b days before T.

Both *Genre* and *Rating* are dummy variables and the mean budget of the films is approximately \$50 million. The duration of a film in theatres varies from 4 to 30 weeks, and each film plays on average 3000 screens during the opening weekend. The volume of tweets increases sharply close to the release date, and the sentiment does not display a trend on average. As showed in Lehrer and Xie (2017), the sentiment index plays a major role in the open box prediction.

III. Methods and design

Regression tree and bagging

We first introduce two machine learning methods, both of which are widely used in the literature. The first one is the regression tree technique proposed by Breiman, Friedman, and Stone (1984). The trick in this method is to find the best split of a data sample into clusters recursively for best fitting subject to a pre-determined stopping rule.³ The resulting prediction function is then used for prediction on an evaluation sample set. Intuitively, the RT method can better address data heterogeneity by splitting the sample into clusters with distinct features, hence outperforming conventional regressions in prediction.

We also consider the bootstrap aggregation technique, known as bagging, developed in Breiman (1996). BAG involves a training process by creating new training sets through bootstrap. We draw B random samples with replacement from the original training set. For each bootstrap sample, we apply RT and obtain a forecast. Finally, we estimate the simple average of the B forecasts as the final forecast. BAG improves predictive accuracy by smoothing the prediction function from the original RT.

The hybrid of machine learning and econometrics

It is well known that machine learning methods have superior performance in short-run predictions, i.e. when the number of predicts is small relative to the size of the available data. Machine learning methods are very flexible in capturing structural heterogeneity

²There are 12 genres in total and one movie can have three genres at most.

³We follow the machine learning literature by applying the 10-fold cross validation criterion to determine the level of splitting in regression tree. The same criterion is applied in the bootstrap aggregation method later.

in the sample data to the maximum extent. By the same token, however, machine learning methods tend to perform badly in long-run prediction, when trends matter more than ‘noises’ due to short-run structural heterogeneity. This is in contrast to the traditional econometric methods, which, to some extent, specialize in identifying long-run trends. As a result, it is worthwhile to combine machine learning methods with traditional econometric methods, in an attempt to pick up both short- and long-run prediction power.

Hirano and Wright (2017) are one such attempt. They proposed a split-sample method to mitigate uncertainty about model selection. The core of SPLT is more in the econometric tradition, consisting of splitting the training set into two parts: one for model selection using AIC and the other for model estimation. The authors showed that adding a bagging step to the plain SPLT substantially improves its prediction performance. The bagging augmented SPLT method can be viewed as a hybrid of econometric and machine learning methods, and is implemented in our empirical exercises.

We modify SPLT to obtain a new hybrid method, denoted as SPLT_{PMA}, replacing the AIC method by the prediction model average method developed by Xie (2015), while keeping the bagging procedure. In SPLT_{PMA}, after an initial split, we apply PMA to the first subsample to obtain a weight structure over all candidate models, and then use the weights to calculate a weighted average model as the prediction model, where each candidate model is estimated on the second subsample. The detail of the algorithm is relegated to the Appendix.

As shown later, our hybrid method SPLT_{PMA} performs quite well in general compared with eight widely used methods in both econometrics and machine learning. Moreover, it outperforms SPLT in all cases, and delivers the best prediction accuracy when sample heterogeneity is limited. Our hybrid method SPLT_{PMA} thus offers a potential way to combine advantages of both machine learning and econometrics.

Assessing prediction efficiency

Following Lehrer and Xie (2017), the efficiencies of different methods are assessed by an exercise that shuffles the experiment data sample of size n , into an evaluation set of size n_E and a training set of size $n_T = n - n_E$. Effectively, n_E/n_T represents the (relative) forecasting horizon, where a longer horizon means less information for prediction. For a training set, we apply nine prediction methods:

- (1) a general unrestricted model using all variables available,
- (2) a GUM that ignores social media data,
- (3) a model selected by the general to specific method,
- (4) a model selected by AIC,
- (5) a model selected by the PMA technique proposed by Xie (2015),
- (6) the RT method from “Regression Tree and Bagging”,
- (7) the BAG method from “Regression Tree and Bagging”,

Table 2. Relative efficiency of unconditional prediction.

n_E	n_E/n_T (%)	GUM	MTV	GETS	AIC	PMA	RT	BAG	SPLT	SPLT _{PMA}
<i>Panel A: MSFE</i>										
1	1.08	1.1154	2.7680	1.9078	1.3104	1.1190	0.2456	0.5237	1.2237	1.0000
2	2.17	1.3729	2.9176	1.8195	1.2781	1.2048	0.4808	0.7410	1.2184	1.0000
5	5.62	1.1765	2.6507	1.6568	1.1793	1.0882	0.7601	0.8307	1.1817	1.0000
10	11.90	1.3500	2.6403	1.6581	1.1309	1.0314	0.9337	0.9064	1.1489	1.0000
20	27.03	1.5205	2.6377	1.7708	1.0726	1.0017	1.1680	1.0749	1.1900	1.0000
30	46.88	1.8085	2.5700	1.8062	1.0384	0.9966	1.1609	1.0821	1.2010	1.0000
40	74.07	2.1973	2.7174	1.8582	0.9548	0.9079	1.1810	1.0771	1.2130	1.0000
<i>Panel B: MAFE</i>										
1	1.08	1.1448	1.8034	1.4972	1.2409	1.1466	0.5372	0.7845	1.1991	1.0000
2	2.17	1.2516	1.8148	1.4197	1.2264	1.1645	0.7184	0.8891	1.1881	1.0000
5	5.62	1.1668	1.7651	1.3706	1.1784	1.1122	0.8406	0.9359	1.1706	1.0000
10	11.90	1.2280	1.7913	1.4256	1.1693	1.1131	0.8977	0.9757	1.1651	1.0000
20	27.03	1.2757	1.7916	1.4407	1.1562	1.1109	1.0076	1.0452	1.1808	1.0000
30	46.88	1.3778	1.7840	1.4858	1.1584	1.1119	1.0467	1.0681	1.2012	1.0000
40	74.07	1.5067	1.7503	1.5303	1.1320	1.0711	1.0892	1.1043	1.2000	1.0000

Notes: Bold numbers indicate the best prediction performance across nine methods for a given horizon (row).

- (8) the SPLT method from “The Hybrid of Machine Learning and Econometrics”, and
 (9) the SPLT_{PMA} method from “The Hybrid of Machine Learning and Econometrics”.

Additional details on methods (1)–(5) are available in Lehrer and Xie (2017). The performance of these methods are evaluated by calculating the mean squared forecast errors and mean absolute forecast errors on the evaluation set, respectively.

The exercise is carried out 10,001 times for different n_E . For methods (7)–(9), we set $B = 200$ for bootstrap. We split the sample in half for methods (8) and (9) following Hirano and Wright (2017).

IV. Results and discussion

Unconditional prediction

We first present results on the unconditional prediction exercise. In this case, the training sets are random subsets of the full sample and the evaluation sets include all genres. Table 2 reports the median MSFE and MAFE from the 10,001 duplications, for $n_E = 1, 2, 5, 10, 20, 30$, and 40. To ease interpretation, we normalize the median MSFE and MAFE of each method by those of SPLT_{PMA}. Values larger than 1 thus indicate inferior performance relative to SPLT_{PMA}.

Evidently, RT has superior performance to all others when n_E is small, i.e. short horizon, particularly for $n_E = 1$. Further, BAG performs similarly to RT. As the forecast horizon increases, the performance of econometric methods improves, especially for PMA. Exploring the two hybrid methods, we first notice that SPLT is always worse than PMA, RT and BAG. In contrast, our newly proposed SPLT_{PMA} does not only outperform SPLT, but actually dominates pure machine learning and econometric methods when $n_E = 20$ under MSFE and $n_E = 20, 30$, and 40 under MAFE.

Underlying these results is the following simple intuition. Data heterogeneity implies many nonlinearities in the data, which matters most in short horizon prediction. Because machine learning methods perform very well in detecting these noises, they deliver more accurate forecast in short run. In contrast, for long horizon, what matters is mainly the trends, and traditional econometric methods are better at identifying these patterns. Moreover,

although linearity plays a larger role in long run, the relevant trends can still quite vary across samples. This renders a single regression specification unable to cope well with such a model uncertainty problem, which in turn is better handled by the model averaging procedure. As a result, PMA and SPLT_{PMA} outperform in long-run prediction.

Conditional prediction

Although the average forecasting efficiency across all genres in the previous subsection is illustrative on the merits of machine learning and econometric

Table 3. Relative efficiency of conditional prediction.

n_E	PMA	RT	BAG	SPLT	SPLT _{PMA}
Panel A: genre-comedy					
<i>Panel A.1: MSFE</i>					
1	2.0790	2.1953	2.3613	1.1340	1.0000
2	2.7655	2.5110	2.8558	1.1501	1.0000
5	1.4104	1.5053	1.4737	1.1244	1.0000
10	1.5193	1.5528	1.6109	1.1850	1.0000
20	1.2443	1.3669	1.3897	1.2153	1.0000
30	1.1123	1.1729	1.1766	1.2147	1.0000
<i>Panel A.2: MAFE</i>					
1	1.4419	1.4816	1.5367	1.1340	1.0000
2	1.5943	1.5921	1.6243	1.1501	1.0000
5	1.2458	1.3325	1.3460	1.1244	1.0000
10	1.2433	1.2986	1.2871	1.1850	1.0000
20	1.1866	1.2515	1.2377	1.2153	1.0000
30	1.0990	1.1627	1.1727	1.2147	1.0000
Panel B: genre-drama					
<i>Panel B.1: MSFE</i>					
1	1.7932	1.8875	1.8561	1.1020	1.0000
2	2.5155	2.6516	2.4397	1.1451	1.0000
5	1.6946	1.7557	1.6958	1.2013	1.0000
10	1.5306	1.6005	1.5657	1.2546	1.0000
20	1.4143	1.4722	1.4405	1.2632	1.0000
30	1.4196	1.4848	1.4605	1.2764	1.0000
<i>Panel B.2: MAFE</i>					
1	1.7932	1.8875	1.8561	1.1020	1.0000
2	2.5155	2.6516	2.4397	1.1451	1.0000
5	1.6946	1.7557	1.6958	1.2013	1.0000
10	1.5306	1.6005	1.5657	1.2546	1.0000
20	1.4143	1.4722	1.4405	1.2632	1.0000
30	1.4196	1.4848	1.4605	1.2764	1.0000
Panel C: genre-action					
<i>Panel C.1: MSFE</i>					
1	2.7464	2.3105	2.6074	1.1708	1.0000
2	1.0115	0.9670	0.9254	1.1420	1.0000
5	0.9144	0.8669	0.8561	1.2200	1.0000
10	0.8192	0.7697	0.7550	1.2356	1.0000
20	0.7355	0.6871	0.7255	1.2636	1.0000
30	0.7557	0.7073	0.7026	1.2864	1.0000
<i>Panel C.2: MAFE</i>					
1	1.6572	1.5200	1.6147	1.1708	1.0000
2	1.1049	1.0674	1.0686	1.1420	1.0000
5	1.0196	0.9658	0.9634	1.2200	1.0000
10	0.9757	0.9294	0.9309	1.2356	1.0000
20	0.9374	0.8962	0.9181	1.2636	1.0000
30	0.9555	0.9192	0.9131	1.2864	1.0000

Note: Bold numbers indicate the best prediction performance across nine methods for a given horizon (row). Results on relative efficiencies of GUM, MTV, GETS and AIC are available upon request.

methods, in practice, the film industry is more concerned with the prediction accuracy for films with specific attributes. Here, we present results on the conditional prediction exercise, and we focus on a particular attribute – genre. The design is similar to the previous one, except that the evaluation sets are randomly drawn from movies of identical genre, with $n_E = 1, 2, 5, 10, 20$ and 30. We consider three subsamples: comedy, drama and action, and Table 3 reports the results.⁴

On the one hand, for predictions conditional on comedy and drama, our new hybrid method $SPLT_{PMA}$ dominates all other methods, both in short run and long run. On the other hand, for action movies, pure machine learning methods command higher accuracy in most cases. Underlying the apparent disparity between these two cases is another simple intuition. For a brief explanation, we first observe from Figure 1 that relative to the

subsamples of comedy and drama, the subsample of action is clearly more heterogeneous, as once a movie belongs to action, it is also very likely to be either a thriller or crime, or both. Greater heterogeneity in action thus leads to better performance of machine learning methods. In contrast, if sample heterogeneity is controlled, as in comedy and drama, the new hybrid method $SPLT_{PMA}$ turns out to provide the correct balance between capturing short-run noise and identifying long-run trend.

V. Conclusion

Our empirical results illustrate the superiority of machine learning methods in detecting irregular patterns or ‘noises’ due to data heterogeneity for short-run prediction, and demonstrate the ability of more standard econometric methods in identifying regular trends which matter more in long-run

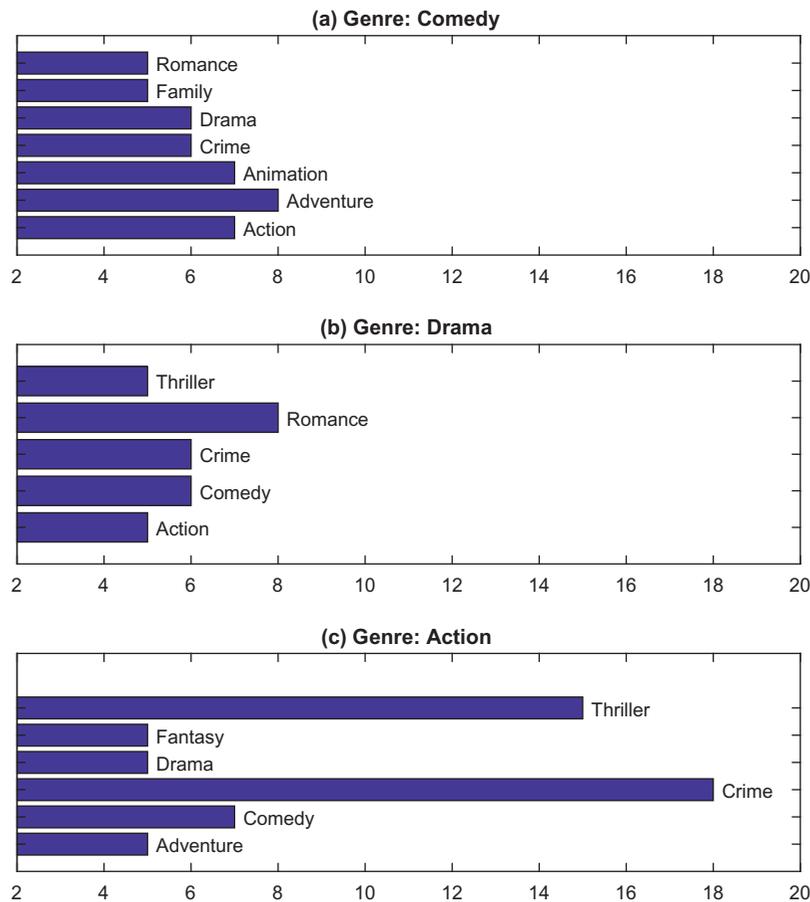


Figure 1. Heterogeneity on selected genres. (a) Genre: comedy. (b) Genre: drama. (c) Genre: action.

⁴The sample sizes are 40, 32 and 35, respectively, and they are the three genres with sample size larger than 30. To save space, we omit results on GUM, MTV, GETS and AIC. None of the methods gives the highest prediction accuracy.

prediction. Moreover, our new hybrid prediction method $SPLT_{PMA}$ performs well on average, and reveals potential to outperform both pure machine learning and standard econometric methods, simultaneously in short- and long-run prediction, once data heterogeneity is controlled.

In addition to the methodological contributions, our results also have managerial and practical implications. For the film industry, box office prediction is a major task in management. In order to improve prediction accuracy, recent research and industry practice have focused on utilizing social media data (Liu 2006; Chintagunta, Gopinath, and Venkataraman 2010; Moretti 2011). However, given pronounced structural heterogeneity in the social media data, traditional econometric methods perform poorly typically. Against this background, our results demonstrate the usefulness of machine learning methods in general, and the potentially significant gains of the hybrid method, such as our $SPLT_{PMA}$, in particular, for utilizing the social media data efficiently and improving the box office prediction accuracy.

Acknowledgments

We wish to thank seminar participants at the 2017 Young Econometricians around the Pacific (YEAP) conference, Chinese Academy of Sciences, Renmin University, and Xiamen University for comments and suggestions. Any errors are our own.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This paper is partially supported by the National Natural Science Foundation of China under Grant Nos. 71661137003, 71503191, 71701175 and 91646206, the Chinese Ministry of Education Project of Humanities and Social Sciences Grant No. 17YJC790174, the Fundamental Research Funds for the Central Universities under Grant Nos. 20720171002 and 20720171076, and Educational and scientific research program for young and middle-aged instructor of Fujian province, No. JAS170018.

References

Antenucci, D., M. Cafarella, M. Levenstein, C. Ré, and M. D. Shapiro. 2014. "Using Social Media to Measure Labor

Market Flows." Working Paper 20010. National Bureau of Economic Research.

- Athey, S., and G. W. Imbens. 2017. "The State of Applied Econometrics: Causality and Policy Evaluation." *Journal of Economic Perspectives* 31 (2): 3–32. doi:10.1257/jep.31.2.3.
- Breiman, L. 1996. "Bagging Predictors." *Machine Learning* 26: 123–140. doi:10.1007/BF00058655.
- Breiman, L., J. Friedman, and C. J. Stone. 1984. *Classification and Regression Trees*. Chapman and Hall/CRC.
- 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. doi:10.1287/mksc.1100.0572.
- Einav, L., and J. Levin. 2014. "Economics in the Age of Big Data." *Science* 346 (6210): 1243089–1243090. doi:10.1126/science.1243089.
- 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." In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, 479–482.
- Hirano, K., and J. H. Wright. 2017. "Forecasting with Model Uncertainty: Representations and Risk Reduction." *Econometrica* 85 (2): 617–643. doi:10.3982/ECTA13372.
- Lehrer, S. F., and T. Xie. 2017. "Box Office Buzz: Does Social Media Data Steal the Show from Model Uncertainty When Forecasting for Hollywood?" *Review of Economics and Statistics* 99 (5): 749–755. doi:10.1162/REST_a_00671.
- Liu, Y. 2006. "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue." *Journal of Marketing* 70 (3): 74–89. doi:10.1509/jmkg.70.3.74.
- Moretti, E. 2011. "Social Learning and Peer Effects in Consumption: Evidence from Movie Sales." *Review of Economic Studies* 78 (1): 356–393. doi:10.1093/restud/rdq014.
- Mullainathan, S., and J. Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106. doi:10.1257/jep.31.2.87.
- Toole, J. L., Y.-R. Lin, E. Muehlegger, D. Shoag, M. C. González, and D. Lazer. 2015. "Tracking Employment Shocks Using Mobile Phone Data." *Journal of the Royal Society Interface* 12 (107): 20150185. doi:10.1098/rsif.2015.0185.
- Varian, H. R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28 (2): 3–28. doi:10.1257/jep.28.2.3.
- Xie, T. 2015. "Prediction Model Averaging Estimator." *Economics Letters* 131: 5–8. doi:10.1016/j.econlet.2015.03.027.

Appendix. Details of the $SPLT_{PMA}$ method

The algorithm consists of the following steps:

- Step 1. Draw a random sample with replacement from the original training set.
- Step 2. Split the sample into two parts.

Step 3. Apply the PMA method to the first subsample and obtain a weight structure on all candidate models.

Step 4. Estimate each candidate model on the second subsample, and obtain the candidate forecast on the evaluation set.

Step 5. Use the weights in Step 3 to calculate the model average forecast using candidate forecasts in Step 4.

Step 6. Repeat Steps 1–5 by B times.

Step 7. The final forecast is the simple average of B model average forecasts.