ARTICLE IN PRESS
International Journal of Forecasting ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Predicting the failures of prediction markets: A procedure of decision making using classification models

Chung-Ching Tai [a], Hung-Wen Lin [b,*], Bin-Tzong Chie [c], Chen-Yuan Tung [d]

[a] Department of Economics, Tunghai University, Taichung, Taiwan
[b] Department of Finance, Nanfang College of Sun Yat-Sen University, Guangzhou, China
[c] Department of Industrial Economics, Tamkang University, New Taipei City, Taiwan
[d] Graduate Institute of Development Studies, National Chengchi University, Taipei, Taiwan

ARTICLE INFO

Keywords:
Combining forecasts
Support vector machine
Decision trees
Principal component analysis
Discriminant analysis
Imbalanced data
Oversampling
SMOTE

ABSTRACT

Prediction markets have been an important source of information for decision makers due to their high ex post accuracies. Nevertheless, recent failures of prediction markets remind us of the importance of ex ante assessments of their prediction accuracy. This paper proposes a systematic procedure for decision makers to acquire prediction models which may be used to predict the correctness of winner-take-all markets. We commence with a set of classification models and generate combined models following various rules. We also create artificial records in the training datasets to overcome the imbalanced data issue in classification problems. These models are then empirically trained and tested with a large dataset to see which may best be used to predict the failures of prediction markets. We find that no model can universally outperform others in terms of different performance measures. Despite this, we clearly demonstrate a result of capable models for decision makers based on different decision goals.

1. Introduction

Prediction markets (PMs), defined as speculative markets that provide forecasts for future events or variables, have been broadly used to predict outcomes in elections, sporting events, economic performance, and other fields in recent years. Ever since the establishment of *Iowa Electronic Markets* (IEM) in 1988, prediction markets have been praised for their high accuracy in predicting future events. The success of IEM was documented in Berg, Nelson, and Rietz (2008), and other studies also demonstrated that prediction markets have higher accuracy over other forecasting methods (Gürkaynak & Wolfers, 2005; Leigh & Wolfers, 2006; Ortner, 1998; Pennock, Lawrence, Giles, & Nielsen, 2001; Servan-Schreiber, Wolfers, Pennock, & Galebach, 2004; Tung, Chou, Lin, & Lin, 2011; Vaughan Williams & Reade, 2016a; Wolfers & Leigh, 2002; Wolfers & Zitzewitz, 2004).

Prior assessment of PMs' prediction accuracy of any particular future event is important for individuals, organizations, and even societies to manage risks and seize opportunities. Given the overall high accuracy of PMs observed in past studies, it appears that we should be confident of the prediction accuracy of any particular future event currently traded in a PM. However, the failures of PMs in events such as *Brexit* and the 2016 U.S. Presidential Election, remind us of the risk of basing our decision on a PM's prediction simply because PMs have shown high accuracy in the past.[1] The reason is that the current prediction market for a particular event might differ considerably

[1] Since prediction markets gave non-zero probabilities to "Leave" in Brexit and "Trump" in the 2016 U.S. presidential election, one can defend prediction markets by contending that they did not really fail from the probability point of view. Because what PMs collected were people's beliefs, the discrepancy between PMs' predictions and the actual results

* Corresponding author.
E-mail address: d97724008@ntu.edu.tw (H.-W. Lin).

https://doi.org/10.1016/j.ijforecast.2018.04.003
0169-2070/

Please cite this article in press as: Tai, C.-C., et al., Predicting the failures of prediction markets: A procedure of decision making using classification models. International Journal of Forecasting (2018), https://doi.org/10.1016/j.ijforecast.2018.04.003.

from past ones in terms of their market properties. To decision makers, what is needed is not the *ex post* overall accuracy for past PMs, but an *ex ante* assessment of the prediction accuracy of the current PM. Can we provide such assessment for prediction markets as we do for other prediction methods, such as surveys and opinion polls?

Berg, Nelson, and Rietz (2003) proposed three distinct methods to measure the forecast standard errors for prediction markets: the *structural model* developed by Berg, Forsythe, and Rietz (1997), the *time-series model*, and the *implied volatility model*. The structural model tries to predict a vote-share (VS) PM's absolute average prediction error with linear regression. This regression model is intuitive because it uses variables related to election properties and market properties to explain the prediction errors. Berg et al. (2003) tested the structural model with two out-of-sample predictions from IEM's 1996 and 2000 US presidential election markets but found it performed poorly in the tests. The time-series model assumes that the evolution of the spread forecasts (differences in two candidate contracts' market prices) follows an AR(1) process if VS PMs are efficient. In this case, one can estimate the mean and standard deviation of the error term using market data and then compute the standard error of the forecast accordingly. Even if there is empirical evidence of underreaction or overreaction in financial markets, Berg et al. (2003) still showed that the time-series model can be used to compute the confidence intervals for VS PM predictions over time. The third model—the implied volatility model—uses both the VS and winner-take-all (WTA) PMs to compute the implied volatility of the vote-share forecast. In a two-candidate election event, the VS market price can be regarded as the forecast of the mean, and the WTA market price measures the probability that the vote share exceed 50%. Under the assumption that the vote share is normally distributed, one can use an inverse normal function to compute the implied variance of the vote share. Nonetheless, there are two practical limitations in applying this model. First, one needs to have both a VS and a WTA market for the target event. Secondly, this method applies only when there are two candidates in an election.

We notice three issues regarding Berg et al. (2003)'s methods. First, their methods were designed only for vote-share PMs. Second, to provide assessments of PMs' predictions is one thing, whether those assessments are helpful in decision is another. Last but most importantly, they only considered information contained within market prices. In theory, PMs' predictions will be accurate if they aggregate dispersed information efficiently. Although, in general, prediction markets can be regarded as efficient, there is also evidence of prediction biases (Sauer, 1998; Vaughan Williams, 1999).[2] Prior literature has identified many factors influencing market efficiency, such as market types (Ali, 1979; Asch & Quandt, 1987; Bird & McCrae, 2008; Blackburn & Peirson, 1995; Cain, Law, & Peel, 1997; Franck, Verbeek, & Nesch, 2010; Gabriel & Marsden, 1990, 1991; Lo & Busche, 2008; Smith, Paton, & Vaughan Williams, 2009; Vaughan Williams & Paton, 1997), trading volume (Busche & Walls, 2000; Walls & Busche, 2003), sources of information or how information is disseminated through social media (Brown, Rambaccussing, Reade, & Rossi, 2016; Croxson & Reade, 2014; Vaughan Williams & Reade, 2016b), etc. These findings suggest a new route to the assessment of PMs' prediction accuracy: we should consider additional information which is not contained within market prices when evaluating PM's predictions.

In response to these issues, Lin, Tung, and Yeh (2013) proposed a pioneering method which, drawing on a large set of non-price factors, combines *principal component analysis* (PCA) and *discriminant analysis* (DA) to judge winner-take-all markets' predictions ex ante. For any winner-take-all contract, the PM predicted that a candidate would be elected if its contract price is the highest among all candidate contracts. Then, the PCA-DA model predicts whether PMs' binary predictions are correct or not based on a set of market variables. In the out-of-sample test, their PCA-DA model exhibited high accuracy in identifying correct PM predictions—for those being predicted as correct PM predictions, 97.72% of them were indeed correct. However, their PCA-DA model performed poorly in identifying incorrect PM predictions—only 19.58% of the PM predictions being categorized as incorrect predictions by PCA-DA were indeed incorrect.

To improve the ability of predicting the correctness of PM's predictions, Tai, Chih, Lin, and Tung (2016) extended Lin et al. (2013)'s tool box and included three additional models—the *Logit model*, *decision tree* (DT), and *support vector machine* (SVM)—to predict whether a WTA PM's prediction would be correct in advance. The idea proposes that different models exploit information in different ways, thereby enabling us to make the best use of information if we combine these models' predictions in a proper way. Tai et al. (2016) compared several combination methods and found that in general, combined models performed better than single models. Furthermore, their study suggests that the best combined models would change under different decision goals; a decision maker can determine a utility function first and then select the combined model which meets their need.

Although Tai et al. (2016)'s method can be readily used to predict any winner-take-all PM, they ignored an important issue concerning the proper identification of incorrect market predictions—the *imbalanced dataset*. An imbalanced dataset refers to a dataset where the classification categories are not approximately equally represented (Chawla, 2005). There are two main reasons why we should pay attention to an imbalanced dataset. First, the cost of misprediction of the minority class is usually higher than that of the majority class, e.g. in medical domains (Rahman & Davis, 2013) or credit scoring problems (Crone & Finlay, 2012). Secondly, an imbalanced dataset will lead classification algorithms to focus on the majority class, because for those algorithms it would be easier to reduce the overall loss or errors in identifying the majority class. In general,

merely reflected the surprise nature of these two events *ex ante*. We thank an anonymous referee for indicating this point of view.

[2] For example, the favourite-longshot bias (Berkowitz, Depken, & Gandar, 2017; Busche, 2008; Busche & Hall, 1988; Gandar, Zuber, & Dare, 2000; Gandar, Zuber, & Lamb, 2001; Ottaviani & Sørensen, 2008, 2010; Schnytzer & Weinberg, 2008; Snowberg & Wolfers, 2010; Snyder, 1978; Vaughan Williams & Paton, 1998; Woodland & Woodland, 1994).

you need to have enough samples for both categories (correct and incorrect) so that the models can learn how to identify them well.

For the purpose of identifying PM failures in advance, it is clear that we do suffer a lot from these two problems. For example, although they only constitute two cases of failure since the birth of IEM and would not drastically lower the overall accuracy of prediction markets, the failures of PMs to predict Brexit or the 2016 US Presidential Election were very costly to decision makers in many societies and greatly harmed PMs' credibility. Also, prediction markets usually have high overall accuracy rates. For WTA markets, this means that for the market samples we have, almost all of them made correct predictions. As a result, a classification model can still have a high accuracy rate even if it erroneously predicts that all PMs' predictions are correct. Requesting an algorithm to control Type II errors will not be very helpful because there are not enough data points for it to learn the true structure. Consequently, many of Tai et al. (2016)'s models tended to predict that all PM predictions were correct in the out-of-sample tests. This observation suggests that the imbalanced dataset did hinder their models from learning to identify incorrect PM predictions.

The aim of this paper is to provide a procedure which helps the decision maker acquire a useful model to foresee the correctness of PMs' predictions. We tackle the imbalanced dataset problem by generating artificial samples where the number of incorrect records were largely increased. We propose a new procedure in which models were trained with artificial samples before the out-of-sample test. Our goal is to discover models which have high-accuracy in predicting the correctness of PMs' predictions, especially those PMs which failed to predict the actual outcomes.

The rest of this paper is organized as follows. Section 2 presents our methodology. Section 3 reports the training results and outcome of the out-of-sample test. Section 4 discusses the choice of the best prediction model. Section 5 concludes.

2. Methodology

In this paper, we propose a procedure consisting of the following stages:

(1) **Sample adjustment**
(2) **Training single models**
(3) **Model combinations**
(4) **Out-of-sample tests**
(5) **Selecting the best model**

In this section, we introduce the elements of our methodology, including:

- The data and variables
- The strategies to tackle the imbalanced dataset problem
- The four single models we used to predict the correctness of our WTA markets
- The methods of model combination
- Methods by which we conducted the out-of-sample test.

How to select the best model will be discussed in Section 4.

2.1. Data

This paper uses data from *xFuture* (http://xfuture.org), which is an online prediction market system established in July 2006, and operated by a Taiwanese company. XFuture attracts Chinese-speaking participants worldwide and is the largest Chinese prediction market system. As of September 2015, xFuture had issued 115,564 contracts of 6,111 contract sets, with more than 190 thousand registered members, and 6 million transactions.

The data used in this paper consists of 650 futures contracts of Taiwan's elections from the xFuture database. These elections in Taiwan include the 2006 mayoral elections, the 2008 legislators' elections, the 2008 presidential election, the 2009 county magistrate and mayoral elections, the 2009–2011 legislator by-elections, and the 2010 five-metropolis mayoral elections. Among these 650 winner-take-all markets, only 35 of them lead to incorrect predictions, which means that the overall accuracy is 94.6%.

2.2. Variable description

The accuracy of PMs' predictions depends on whether markets work efficiently to aggregate disperse information possessed by market participants. There are a lot of potential factors which could influence market efficiency. In this paper, we consider six types of factors which were mentioned in the PM literature. The first type of factor is *marginal traders*. Forsythe, Nelson, Neumann, and Wright (1992), Forsythe, Rietz, and Ross (1999), and Oliven and Rietz (2004) emphasize the importance of marginal traders, which significantly affect PMs' prediction accuracy. There is, however, still no consensus on the operational definition of marginal traders, and obtaining the corresponding data is problematic. On the basis of Luckner, Weinhardt, and Studer (2006)'s methodology, we define two proxy variables for marginal traders: the traders with previously better transaction results, *vis-à-vis* all traders, and the ratio of the number of transactions with limit orders to the total number of transactions.

From previous studies, there are an additional five types of factors affecting the prediction accuracy of PMs: *market consensus* (Berg et al., 1997; Gruca, Berg, & Cipriano, 2005), *market properties* (Berg et al., 1997; Gruca et al., 2005; Ho & Chen, 2007; Kambil & Heck, 2002; Ledyard, 2006; Luckner et al., 2006; Snowberg, Wolfers, & Zitzewitz, 2005), *the difficulty of predicting election outcomes* (Forsythe et al., 1999; Rhode & Strumpf, 2004; Wolfers & Zitzewitz, 2004), *manipulation* (Deck, Lin, & Porter, 2013), and *market price* (Wolfers & Zitzewitz, 2006). As a result, we have 40 variables belonging to six categories, as summarized in Table 1. Among them, the Avatar variables, defined as the ratios of players with multiple accounts to manipulate market prices (Lin, Tung, Lin, & Cho, 2015), are proxy variables of the degree of price manipulation. The basic descriptive statistics of all variables are listed in Table A.10 in Appendix.

Table 1
Definitions of input variables.

Variable type	Variable name	Variable description
Marginal traders	*GP_share_lyc_R*	For traders involved in this contract, the ratio of the top R performing traders of the last calendar year to the total number of traders till the election eve, $R = 100, 200, 300$.
	GP_share_lyc_S%	For traders involved in this contract, the ratio of the top $S\%$ performing traders of the last calendar year to the total number of traders till the election eve, $S = 1, 5, 10$.
	GP_share_365d_T	For traders involved in this contract, the ratio of the top T performing traders of the last 365 days to the total number of traders till the election eve, $T = 100, 200, 300$.
	GP_share_365d_U%	For traders involved in this contract, the ratio of the top $U\%$ performing traders of the last 365 days to the total number of traders till the election eve, $U = 1, 5, 10$.
	GP_share_30d_V	For traders involved in this contract, the ratio of the top V performing traders in the last 30 days to the total number of traders till the election eve, $V = 100, 200, 300$.
	GP_share_30d_W%	For traders involved in this contract, the ratio of the top $W\%$ performing traders of the last 30 days to the total number of traders till the election eve, $W = 1, 5, 10$.
	Limit_ratio_volume	Ratio of the number of transactions with limit orders to the total number of transactions.
Degree of market consensus	*WBAS*	$\frac{\sum \text{offer price} \times \text{selling volume} - \sum \text{bid price} \times \text{buying volume}}{\sum \text{selling volume} + \sum \text{buying volume}}$
	Buy_sell	Ratio of shares to buy to shares to sell.
Market properties	*Trades*	Number of transactions of this contract.
	Traders	Number of traders of this contract.
	Days	Number of days between the day that the contract is firstly traded and the election eve.
	Volume	Traded volume of this contract.
	Two_way	Ratio of two-way traders to the total number of traders.
	IP_share	Ratio of traders in Taiwan to the total number of traders. We define traders in Taiwan as those who used IP addresses registered in Taiwan to trade.
	Traded_order_ratio	Ratio of the total number of traded orders to the total number of orders.
Difficulty of predicting a topic	*Highest-price*	The contract with the highest weighted average price in a contract set.
	NC	Number of contracts for the linked contract set.
	Price_gap	The difference of the highest price and the second highest price of contracts in a contract set.
Avatar	*Avatar_ratio_3*	Assume an avatar account exists if at least three traders share the same password. This variable is the ratio of the number of avatar accounts involved in this contract to the total number of traders of this contract.
	Avatar_Xd_ratio_3	Assume an avatar account exists if at least three traders share the same password. This variable is the ratio of the number of avatar accounts involved in this contract to the total number of traders of this contract in the last X days before liquidation, $X = 15, 30, 365$.
	Avatar_volume_ratio_3	Assume an avatar account exists if at least three traders share the same password. This variable is the ratio of the number of transactions made by avatar traders to the total number of transactions of this contract.
	Avatar_volume_Yd_ratio_3	Assume an avatar account exists if at least three traders share the same password. This variable is the ratio of the number of transactions made by avatar traders to the total number of transactions of this contract in the last Y days before liquidation, $Y = 15, 30, 365$.
Properties of election contracts	P^w	The weighted average price of each election contract on the last trading day.

2.3. Sample adjustment methods

Among the 650 PM contracts in our dataset, only 35 of them lead to incorrect predictions. Consequently, our data consists of two imbalanced classes: correct (615 contracts, 94.6%) and incorrect (35 contracts, 5.4%). There are two general ways in the literature to overcome the problems which result from an imbalanced dataset: *undersampling* and *oversampling* (Chawla, 2005). Undersampling means downsizing the majority class by dropping some data points from the majority class so as to match the number of data points in the minority class. Oversampling means increasing the data points of the minority class so as to match the number of data points in the majority class. Although previous studies show mixing results regarding which method is better (Batista, Prati, & Monard, 2004; Chawla, 2003; Drummond & Holte, 2003; Maloof, 2003), undersampling is not suitable for our study because we would have too few (only 35) cases for both classes.

For the method of oversampling, there are two techniques commonly used in the literature: *random oversampling* and the *synthetic minority over-sampling technique* (SMOTE). Random oversampling means that cases in the minority class are randomly chosen and replicated a number of times, then these new cases are added to the training dataset. While there is a concern that random oversampling will lead to overfitting of the selected cases (Crone & Finlay, 2012), we adopt a modified version of

this technique: we replicate the whole set of the minority class so as to mitigate the overfitting problem as much as possible. SMOTE was proposed by Chawla, Bowyer, Hall, and Kegelmeyer (2002), and it deals with the overfitting problem in a different way. SMOTE artificially creates synthetic examples rather than oversampling the same set of examples with replacement—it generates a new data point by taking the weighted average in the feature space of two selected data points in the k nearest neighborhood. By doing so, SMOTE effectively generalizes the minority class to mitigate the overfitting problem.

In this paper, we adopt both oversampling and SMOTE to expand our minority class—the PM contracts which lead to incorrect prediction—in the training dataset. However, our goal is not to compare the pros and cons of these two techniques. We planned to utilize simple methods to help uncover the best model in predicting the failures of PMs, and these two techniques are simpler than other models in creating balanced datasets.[3] Using simple techniques to achieve our goal is important as it may allow decision makers to dynamically generate correctness predictions for our prediction markets in the future.

2.4. Single models

As mentioned earlier, we have 40 independent variables to describe each of the 650 PM contracts in our dataset. However, there could be problems such as overspecification and multicollinearity if we put all of these variables into a linear regression model. One solution to this problem is to check these variables one by one, as did by Berg et al. (1997). Alternatively, we can use multivariate analysis or machine learning algorithms to discover potential information contained within a large set of variables. On the basis of Chen, Tung, Tai, Chie, Chou, and Wang (2011), Lin et al. (2013), and Tai et al. (2016)'s experimentations, this paper adopts four models to predict the correctness of each future contract traded in winner-take-all PMs: *Logit*, *principal component analysis plus discriminant analysis* (PCA-DA), *support vector machine* (SVM), and *decision tree* (DT). We choose these four models for their broad applications in classification problems. As to our knowledge and literature survey, Logistic regression and PCA-DA are two standard linear classification procedures, while SVM and Decision tree are two widely adopted non-linear methods. Also, these four models had good performance in predicting PM accuracy in the aforementioned studies.[4]

2.4.1. Logit

Our goal is to judge whether or not WTA markets' predictions will be correct, therefore Logit is a reasonable choice because it represents a corresponding model to Berg et al. (1997)'s linear regression model. In the Logit model, we estimate the coefficients of the following function:

$$\ln\left(\frac{\hat{y}}{1-\hat{y}}\right) = \hat{a} + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \cdots \hat{b}_n x_n, \tag{1}$$

where $\hat{y}$ is the estimated probability of being a correct PM, and x_n are the n explanatory variables. After that, if a WTA contract has an estimated odds larger than the sample mean, Logit model will classifies it as correct, otherwise incorrect.[5]

2.4.2. PCA-DA

The PCA-DA model first uses PCA to reduce the dimensionality of the original data and determine a smaller set of principal components. These components are then fed to linear discriminant functions thereby generating the final classification judgments. The PCA-DA model has been applied to forecast many different problems (Dillon, Mulani, & Frederick, 1989; Jombart, Devillard, & Balloux, 2010; Newman & Sheth, 1985; Rosen & Granbois, 1983; Wilton & Pessemier, 1981; Zhao, Chellappa, & Krishnaswamy, 1998).

The PCA-DA model will first map a vector of n explanatory variables into a vector of m principal components:

$$y_k = w_{k1}x_1 + w_{k2}x_2 + \cdots + w_{kn}x_n, \tag{2}$$

where y_k is the kth principal component, w_{ki} is the ith loading of the explanatory variable for the kth component. After that, the model uses these principal components to construct discriminant functions as follows:

$$\begin{aligned} f^1 &= \alpha_0 + \alpha_1 y_1 + \alpha_2 y_2 + \cdots + \alpha_m y_m, \\ f^2 &= \beta_0 + \beta_1 y_1 + \beta_2 y_2 + \cdots + \beta_m y_m, \end{aligned} \tag{3}$$

where f^1 is the discriminant variable for being correct, and f^2 is the discriminant variable for being incorrect. PCA-DA estimates f^1 and f^2 for each PM contract, and classifies that contract as correct or incorrect depending on which variable has a larger value.

2.4.3. Support vector machine

Support vector machine (SVM), proposed by Cortes and Vapnik (1995), is a supervised learning model which captures multivariate and nonlinear relationships hidden in the data. Support vector machines (SVMs) are widely used for classification problems (Delen, Cogdell, & Kasap, 2012; State, Cocianu, & Fusaru, 2010; Weron, 2014). In this paper, we use radial basis function (RBF) as the kernel function

[3] There are more complex hybrid models than the two techniques we used in this paper. For example, some studies combined oversampling and undersampling, while the others integrated SMOTE with other approaches. See Chawla (2005) for an early summarization. On the contrary, oversampling can be easily done with any statistical software, and there is a ready-made R package for SMOTE so that one can easily use it with a single command.

[4] Despite this it doesn't mean that these four methods are the only reasonable choices. Each forecasting model has its own advantages and weakness. For example, with recent development in deep learning technique, one may expect that artificial neural networks (ANN) could be considered as well.

[5] Regarding the issue of multicollinearity, we refer to Greene (2017) and use statistical software STATA to select the independent variables so as to reduce its undesirable effect. To be precise, once STATA detects a high collinearity between two independent variables, it drops one of them. Furthermore, instead of identifying key explanatory variables, our goal is simply to classify PM's predictions as correct or incorrect. The main consequences of multicollinearity, namely, the direction and significance of coefficients, will not cause a huge problem in our study because, according to statistical theory, accurate forecasts may still be possible if the nature of the collinear relationship remains the same within new sample observations (Griffiths, Hill, & Judge, 1993).

Table 2
The confusion matrix of classification models' predictions.

		Predicted outcome	
		PM's prediction is correct (O)	PM's prediction is incorrect (X)
Actual outcome	PM's prediction is correct (T)	TP	FN
	PM's prediction is incorrect (F)	FP	TN

Note: "TP" denotes true positive cases, "FP" denotes false positive cases, "FN" denotes false negative cases, and "TN" denotes true negative cases.

and use LIBSVM proposed by Chang and Lin (2011). 10-fold cross-validation was used to determine the hyperparameters of SVM.

2.4.4. Decision trees

The decision tree method originated from Morgan and Sonquist (1963)'s *Automatic Interaction Detector* (AID) program and became popular after the introduction of Breiman, Friedman, Olshen, and Stone (1984)'s *Classification and Regression Trees* (CART) program. The tree models have evolved into a family of concurrent techniques (Loh, 2014), and we adopted CART as our fourth model because it remains the current standard of tree modeling (Su, Azuero, Cho, Kvale, Meneses, & McNees, 2011).[6] 10-fold cross-validation was used to determine the optimal tree size.

2.5. Evaluating PM's predictions

Following the highest-price principle for our winner-take-all markets, we predict that a candidate would be elected if their corresponding contract has the highest price in the election contract set.[7] After the election, we acknowledge that our PM is correct if that candidate were indeed elected. Similarly, the PM is confirmed correct if a candidate's corresponding contract price was not the highest one in the contract set and that candidate was not elected in the end, either. We use "T" to denote the group of PM contracts whose predictions were proven to be correct, and "F" to denote the contracts which lead to incorrect predictions.

Since our goal is to determine the correctness of PMs' predictions in advance, we use the four models mentioned earlier to classify PMs' predictions into two groups: Group "O"—PMs being determined to have correct predictions, and Group "X"—PMs being determined to have incorrect predictions. Based on the aforementioned definitions, Table 2 summarizes the classification results of our models.

Table 2 is the confusion matrix of our models' predictions. We use three indexes to evaluate our classification models:

(1) The Accuracy Rate of Correct Identification (ARCI) $= \frac{TP}{TP+FP}$,
(2) Accuracy Rate of Incorrect Identification (ARII) $= \frac{TN}{FN+TN}$,
(3) Specificity $= \frac{TN}{FP+TN}$.

As the names suggest, by using ARCI and ARII we want our models to be as accurate as possible when they classify a PM's prediction to be correct or incorrect. Specificity, on the other hand, measures the ratios of incorrect PMs being successfully identified.

2.6. Model combination

In order to improve the capability of predicting the correctness of WTA markets, this paper introduces a method of combined forecasts. Clemen (1989) collected more than 200 papers to support the advantage of combined forecasts and demonstrated that a simple mean of all forecasting tools could be a superior forecasting solution. On the basis of their survey on election predictions, Armstrong (2001) and Graefe, Armstrong, Jones, and Cuzán (2014) reached the same conclusion that combined forecasts could be better than any individual forecasting method.

Combining single models' predictions is straightforward. Nevertheless, integrating binary predictions into a single conclusion when models have different opinions is not a trivial task. We propose two distinct rules to integrate predictions from our single models:

- The Majority Rule: If the majority of the models predict that a PM would be correct, the combined prediction is "correct", otherwise the combined prediction is "incorrect".
- The Consensus Rule: If all models predict that a PM would be correct, the combined prediction is "correct"; if all models predict that a PM would be incorrect, the combined prediction is "incorrect".

To be precise, let f be the weighted average of a single model's prediction, and F be the combined prediction:

$$f = \sum_{i=1}^{4} w_i f_i \tag{4}$$

where w_i represents the weightings for the four single models. The majority rule says

$$F^M = \begin{cases} 1 & \text{if } f \geq 0.5, \\ 0 & \text{if } f < 0.5. \end{cases} \tag{5}$$

The consensus rule says

$$F^C = \begin{cases} 1 & \text{if } f = 1, \\ 0 & \text{if } f = 0, \\ \text{no conclusion} & \text{if } 0 < f < 1, \end{cases} \tag{6}$$

[6] We also tried another popular tree-based algorithm—random forest, but it demonstrated no superiority over standard decision tree model in both the training and the out-of-sample test.

[7] An election contract set is a set of contracts which represents the candidates in the same election event. For example, if there are three candidates in a mayor's election, we will set up an election contract set which consists of three contracts representing the three candidates.

Table 3
Original training sample and testing sample.

Set	Election events	Number of contracts	Sample division
1	2006 mayoral elections	45 = 43 + 2	Original training sample
2	2008 legislators elections	288 = 266 + 22	(333 = 309 + 24)
3	2008 presidential election	78 = 74 + 4	Testing sample (317 = 306 + 11)
4	Other elections in between 2008 and 2009	58 = 55 + 3	
5	2009 county magistrate and mayoral elections	48	
6	2009-2011 legislator by-elections	25 = 21 + 4	
7	2010 five-metropolis mayoral elections	86	
8	Other elections in 2010	22	
	Total	650 = 615 + 35	

Note: The number of contracts for Set 1, 2, 3, 4, and 6 are expressed as "total number of contracts = number of correct contracts + number of incorrect contracts." Contracts belonging to Set 5, 7, and 8 are all correct.

where F^C is the prediction of the combined model following the consensus rule. $F = 1$ means that the target PM contract is classified as "O"; $F = 0$ means that the target PM contract is classified as "X".

There are two further issues concerning combined forecasts. First, which and how many models should we combine? Secondly, what kind of weighting rule produces the most accurate results? For the first question, we tried out all combination possibilities. As for the second question, Armstrong (2001) suggests equal weightings unless there is sufficient evidence that particular forecasting tools should be given higher weights based on their previous performance. Consequently, we implemented four different weighting rules to discover the best combined model to predict our WTA markets:

- Equal weightings
- ARCI-weighting: weightings determined based on the models' ARCI during training.
- ARII-weighting: weightings determined based on the models' ARII during training.
- Specificity-weighting: weightings determined based on the models' specificity during training.

To determine the weightings, we first trained our single models using a subset of our 650 PM contracts (in-sample test). After training, the weightings were calculated based on different performance indexes; then we combined models according to different weighting rules.

2.7. Out-of-sample test

The data used in this paper consists of 650 prediction markets held between 2006 and 2010. Table 3 presents the detailed composition of our dataset. Among the eight subsets of elections, only five of them have incorrect PMs—a total of 35 PM contracts which lead to incorrect predictions. Consequently, we divide our data into training and testing datasets, as demonstrated in Table 3: subset 1 and 2 (which contains 24 incorrect PM contracts) constitute the training dataset and others (which contains 11 incorrect PM contracts) are left for the out-of-sample test.

We adopt two methods to tackle the imbalanced data problem: oversampling and SMOTE. Our goal is to have roughly the same amount of correct and incorrect PM contracts in our training dataset. By oversampling, we replicate the 24 incorrect contracts in the original training dataset twelve times so that we have an additional 288 incorrect contracts. As a result, the new training dataset is composed of 309 correct contracts and 312 incorrect contracts, a total of 621 PM contracts.

By SMOTE, we create synthetic samples from the 24 incorrect contracts by twelve times so that we have the same number (312) of incorrect contracts as the oversampling method. Meanwhile, we also slightly increase the correct contracts from 309 to 311 with random oversampling.[8] As a result, we have a total of 623 contracts in the new training dataset, in which we have roughly the same number of correct and incorrect contracts (311:312).

Now we have two different training datasets from the oversampling method and SMOTE. The whole experimental process is as follows:

(1) We use these two training datasets to train our single models separately and record their performances in terms of ARCI, ARII, and specificity.
(2) We calculate the weightings according to different weighting rules based on single models' performances during training, and then generate a set of combined models on the basis of different combination rules as described in Section 2.6.
(3) We test all the single and combined models using the testing dataset, which contains 306 correct contracts and 11 incorrect contracts.
(4) We compare all models' performances and identify the best one.

[8] Chawla et al. (2002) proposed that researchers can use SMOTE to expand the minority class and undersample the majority class at the same time to reverse the learning bias. However, our goal is to have a balanced learning between the "correct" and "incorrect" class. Therefore, we oversample the "correct" class a little bit to match the expanded "incorrect" class.

Table 4
Training results for single models.

	Training dataset	TP	FP	FN	TN	ARCI	ARII	Specificity
SVM	Original	307	20	2	4	0.939	0.667	0.167
	Oversampling	307	0	2	312	1.000	0.994	1.000
	SMOTE	311	3	0	309	0.990	1.000	0.990
PCA-DA	Original	287	18	22	6	0.941	0.214	0.250
	Oversampling	283	208	26	104	0.576	0.800	0.333
	SMOTE	293	196	18	116	0.599	0.866	0.372
Logit	Original	303	7	6	17	0.977	0.739	0.708
	Oversampling	257	26	52	286	0.908	0.846	0.917
	SMOTE	287	20	24	292	0.935	0.924	0.936
DT	Original	307	20	2	4	0.939	0.667	0.167
	Oversampling	282	0	27	312	1.000	0.920	1.000
	SMOTE	290	17	21	295	0.945	0.934	0.946

3. Empirical results

In this section, we overlook all the models used and evaluate their capability in predicting the failures of PMs. We evaluate them in terms of the three indexes introduced in Section 2.5; an additional index of identified PM failures is also used to judge the quality of our classification models.

The first step of our analysis is to train our single models with three training datasets: the original training dataset, the oversampling training dataset, and the SMOTE training dataset. Table 4 reports the training results of each single model using the three different training datasets. Note how oversampling and SMOTE improve the models' ability to identify incorrect contracts. The specificities of SVM, DT, and Logit surpass 90% with these two sampling techniques. Nevertheless, we should focus on models' performances in the out-of-sample predictions to effectively gauge these models' capabilities.

Before the out-of-sample test, we build combined models following two different rules (the majority or consensus rule) with four different weighting rules (equal weighted, ARCI-weighted, ARII-weighted, and specificity-weighted), using training performances from three different training datasets. We also consider different combinations of single models. Tables 5–7 presents the results of out-of-sample predictions from single models and various combined models, which were trained with the original dataset, oversampling dataset, and SMOTE dataset, respectively.

Table 5 reports the out-of-sample prediction results for models trained with the original training dataset. There are several points we may ascertain from this table. First, almost all models, whether they are single or combined models, have very high ARCI, meaning that they are very accurate in predicting that a PM would lead to correct predictions. This may result from the fact that our PM has a high overall accuracy; even a classification model which mistakenly classifies all PM contracts to be correct would have a ARCI as high as 96.5% (306/317 = 0.965).

Secondly, the models have quite different ARII—they can be as high as 1 or as low as 0.088. There are two models with ARII of 1: M12 and C134 under equal weightings. M12 is a model which combines SVM and PCA-DA's predictions with equal weightings under the majority rule, and C134 is a model which combines SVM, Logit, and DT with equal weightings under the consensus rule. It seems that they are accurate when classifying a PM contract as incorrect. However, their specificity is low (0.091 and 0.2, respectively). Furthermore, out of the 11 incorrect PM contracts in the testing sample, they only identify 1 of them (TN = 1). Consequently, we may conclude that M12 and C134 cannot effectively predict when a PM will fail. In fact, this is not just a feature of M12 and C134. When trained with the original dataset, no matter whether they are combined models or not or how they are combined, all models in Table 5 have either low ARII or low specificity.

Thirdly, except for the specificity, using different weighting rules does not foster much improvement. The average ARCIs of equal weighted combined models and ARCI-weighted combined models are 0.976 and 0.978; the average ARIIs for equal weighted combined models and ARII-weighted combined models are 0.572 and 0.428; the average specificity for equal weighted combined models and specificity-weighted combined models are 0.184 and 0.286, but they are only marginally different (p-value = 0.08909, Wilcoxon rank-sum test).

Table 6 reports the out-of-sample prediction results for models trained with the oversampling training dataset. We make the following observations. First, we find that using different weighting rules does not bring much improvement, except for the specificity. The average ARCIs of equal weighted combined models and ARCI-weighted combined models are 0.980 and 0.983; the average ARIIs for equal weighted combined models and ARII-weighted combined models are 0.606 and 0.481; the average specificity for equal weighted combined models and specificity-weighted combined models are 0.240 and 0.351, but they are not significantly different.

Secondly, all models have very high ARCI, but what interests us is whether we can find better models to classify our WTA markets relative to models in 5. In general, we do have higher ARII and specificity by using the oversampling training dataset. The average ARII increases from 0.371 to 0.403, and the specificity increases from 0.306 to 0.326 (but not significantly different). However, this is only a general comparison. In practice, a decision maker will need to choose a specific model to classify WTA markets. Do we have a better model after training our models with the oversampling dataset?

Table 5
Out-of-sample test results—models trained with the original dataset.

Model	Equal weighted								ARCI weighted							
	TP	FP	FN	TN	ARCI	ARII	Specificity	NC	TP	FP	FN	TN	ARCI	ARII	Specificity	NC
(1) SVM	305	10	1	1	0.968	0.500	0.091	0								
(2) PCA-DA	275	8	31	3	0.972	0.088	0.273	0								
(3) Logit	288	6	18	5	0.980	0.217	0.455	0								
(4) DT	278	5	28	6	0.982	0.176	0.545	0								
M12	306	11	0	0	0.965	NA	0.000	0	275	8	31	3	0.972	0.088	0.273	0
M13	306	10	0	1	0.968	1.000	0.091	0	288	6	18	5	0.980	0.217	0.455	0
M14	305	10	1	1	0.968	0.500	0.091	0	305	10	1	1	0.968	0.500	0.091	0
M23	305	10	1	1	0.968	0.500	0.091	0	288	6	18	5	0.980	0.217	0.455	0
M24	302	8	4	3	0.974	0.429	0.273	0	275	8	31	3	0.972	0.088	0.273	0
M34	300	7	6	4	0.977	0.400	0.364	0	288	6	18	5	0.980	0.217	0.455	0
M123	305	9	1	2	0.971	0.667	0.182	0	305	9	1	2	0.971	0.667	0.182	0
M124	301	7	5	4	0.977	0.444	0.364	0	301	7	5	4	0.977	0.444	0.364	0
M134	299	7	7	4	0.977	0.364	0.364	0	299	7	7	4	0.977	0.364	0.364	0
M234	297	5	9	6	0.983	0.400	0.545	0	297	5	9	6	0.983	0.400	0.545	0
M1234	305	9	1	2	0.971	0.667	0.182	0	300	7	6	4	0.977	0.400	0.364	0
C123	257	4	0	0	0.985	NA	0.000	56								
C124	251	5	0	0	0.980	NA	0.000	61								
C134	266	4	0	1	0.985	1.000	0.200	46								
C234	239	4	1	1	0.984	0.500	0.200	72								
C1234	239	4	0	0	0.984	NA	0.000	74								
Model	ARII weighted								Specificity weighted							
	TP	FP	FN	TN	ARCI	ARII	Specificity	NC	TP	FP	FN	TN	ARCI	ARII	Specificity	NC
M12	305	10	1	1	0.968	0.500	0.091	0	275	8	31	3	0.972	0.088	0.273	0
M13	288	6	18	5	0.980	0.217	0.455	0	288	6	18	5	0.980	0.217	0.455	0
M14	305	10	1	1	0.968	0.500	0.091	0	305	10	1	1	0.968	0.500	0.091	0
M23	288	6	18	5	0.980	0.217	0.455	0	288	6	18	5	0.980	0.217	0.455	0
M24	278	5	28	6	0.982	0.176	0.545	0	275	8	31	3	0.972	0.088	0.273	0
M34	288	6	18	5	0.980	0.217	0.455	0	288	6	18	5	0.980	0.217	0.455	0
M123	305	9	1	2	0.971	0.667	0.182	0	288	6	18	5	0.980	0.217	0.455	0
M124	301	7	5	4	0.977	0.444	0.364	0	301	7	5	4	0.977	0.444	0.364	0
M134	299	7	7	4	0.977	0.364	0.364	0	288	6	18	5	0.980	0.217	0.455	0
M234	297	5	9	6	0.983	0.400	0.545	0	288	6	18	5	0.980	0.217	0.455	0
M1234	299	7	7	4	0.977	0.364	0.364	0	288	6	18	5	0.980	0.217	0.455	0

Note: (1) The top-left panel presents the results from models based on equal-weighted methods; the top-right panel is for the ARCI-weighted models; the bottom-left panel is for the ARII-weighted models; the bottom-right panel is for the specificity-weighted models. (2) For combined models, the prefix "M" denotes models following the majority rule, prefix "C" denotes models following the consensus rule; the numbering denotes the constituent models—"1" denotes the SVM model, "2" denotes the PCA-DA model, "3" denotes the Logit model, and "4" denotes the DT model. (3) The results do not vary under different weighting rules for models using consensus rule, so we only report them once in the Equal weighted panel. (4) NC means "no conclusion", see Eqs. (5) and (6) for definitions.

The third observation is that, C234, which is a model combining PCA-DA, Logit, and DT's predictions with equal weightings under the consensus rule, can reach the highest possible ARCI, ARII, and specificity. The fact that C234 has ARCI, ARII, and specificity of 1 means that it may provide extremely accurate predictions about the correctness of PM's predictions. However, it also has 116 NC cases, which means that it can only provide definite predictions for 201 PM contracts out of 317 contracts in the testing sample. It turns out that C234 is a very conservative model, which only classifies a PM contract when all three single models it considers have the same opinions. As a result, it only identified 1 out of 11 PM failures in the testing sample (TN = 1).

Fourthly, by comparing Tables 5 and 6, we find that after training with the oversampling dataset, the single models become more assertive in predicting PM's failures, while combined models become more conservative. This can be seen in terms of ARII and the number of incorrect PM contracts being identified. For the four single models, the average number of incorrect PM contracts being identified increases from 3.75 to 4.75, but their average ARII decreases from 0.246 to 0.131. By contrast, for combined models, the average number of incorrect PM contracts being identified decreases from 3.29 to 3.14, but their average ARII increases from 0.382 to 0.427.

Lastly, acknowledging this fact leads us to another model which attempts to predict the failures of PMs in a different way. The DT model trained with the oversampling dataset has the second highest specificity (0.909) in Table 6, but it also succeeds in identifying almost all PM failures in the testing dataset (10 out of 11). It achieves this by aggressively classifying suspicious contracts as incorrect and therefore has a very low ARII (only 0.167).

Table 7 reports the out-of-sample prediction results for models trained with the SMOTE training dataset. Still, all models have a very high ARCI. Additionally, we make the following observations. First, using different weighting rules than usual does not render much improvement, except for the specificity. The average ARCIs of equal weighted combined models and ARCI-weighted combined models are 0.978 and 0.982; the average ARIIs for equal weighted combined

Table 6
Out-of-sample test results—models trained with the oversampling dataset.

Model	Equal weighted								ARCI weighted							
	TP	FP	FN	TN	ARCI	ARII	Specificity	NC	TP	FP	FN	TN	ARCI	ARII	Specificity	NC
(1) SVM	306	11	0	0	0.965	NA	0.000	0								
(2) PCA-DA	255	6	51	5	0.977	0.089	0.455	0								
(3) Logit	281	7	25	4	0.976	0.138	0.364	0								
(4) DT	256	1	50	10	0.996	0.167	0.909	0								
M12	306	11	0	0	0.965	NA	0.000	0	306	11	0	0	0.965	NA	0.000	0
M13	306	11	0	0	0.965	NA	0.000	0	306	11	0	0	0.965	NA	0.000	0
M14	306	11	0	0	0.965	NA	0.000	0	306	11	0	0	0.965	NA	0.000	0
M23	306	10	0	1	0.968	1.000	0.091	0	281	7	25	4	0.976	0.138	0.364	0
M24	292	6	14	5	0.980	0.263	0.455	0	256	1	50	10	0.996	0.167	0.909	0
M34	300	8	6	3	0.974	0.333	0.273	0	256	1	50	10	0.996	0.167	0.909	0
M123	306	10	0	1	0.968	1.000	0.091	0	306	10	0	1	0.968	1.000	0.091	0
M124	292	6	14	5	0.980	0.263	0.455	0	292	6	14	5	0.980	0.263	0.455	0
M134	300	8	6	3	0.974	0.333	0.273	0	300	8	6	3	0.974	0.333	0.273	0
M234	286	4	20	7	0.986	0.259	0.636	0	286	4	20	7	0.986	0.259	0.636	0
M1234	306	10	0	1	0.968	1.000	0.091	0	300	8	6	3	0.974	0.333	0.273	0
C123	230	3	0	0	0.987	NA	0.000	84								
C124	219	1	0	0	0.995	NA	0.000	97								
C134	237	0	0	0	1.000	NA	NA	80								
C234	200	0	0	1	1.000	1.000	1.000	116								
C1234	200	0	0	0	1.000	NA	NA	117								
Model	ARII weighted								Specificity weighted							
	TP	FP	FN	TN	ARCI	ARII	Specificity	NC	TP	FP	FN	TN	ARCI	ARII	Specificity	NC
M12	306	11	0	0	0.965	NA	0.000	0	306	11	0	0	0.965	NA	0.000	0
M13	306	11	0	0	0.965	NA	0.000	0	306	11	0	0	0.965	NA	0.000	0
M14	306	11	0	0	0.965	NA	0.000	0	306	11	0	0	0.965	NA	0.000	0
M23	281	7	25	4	0.976	0.138	0.364	0	281	7	25	4	0.976	0.138	0.364	0
M24	256	1	50	10	0.996	0.167	0.909	0	256	1	50	10	0.996	0.167	0.909	0
M34	256	1	50	10	0.996	0.167	0.909	0	256	1	50	10	0.996	0.167	0.909	0
M123	306	10	0	1	0.968	1.000	0.091	0	306	10	0	1	0.968	1.000	0.091	0
M124	292	6	14	5	0.980	0.263	0.455	0	292	6	14	5	0.980	0.263	0.455	0
M134	300	8	6	3	0.974	0.333	0.273	0	300	8	6	3	0.974	0.333	0.273	0
M234	286	4	20	7	0.986	0.259	0.636	0	286	4	20	7	0.986	0.259	0.636	0
M1234	306	10	0	1	0.968	1.000	0.091	0	300	8	6	3	0.974	0.333	0.273	0

Note: (1) The top-left panel presents the results from models based on equal-weighted methods; the top-right panel is for the ARCI-weighted models; the bottom-left panel is for the ARII-weighted models; the bottom-right panel is for the specificity-weighted models. (2) For combined models, the prefix "M" denotes models following the majority rule, prefix "C" denotes models following the consensus rule; the numbering denotes the constituent models—"1" denotes the SVM model, "2" denotes the PCA-DA model, "3" denotes the Logit model, and "4" denotes the DT model. (3) The results do not vary under different weighting rules for models using consensus rule, so we only report them once in the Equal weighted panel. (4) NC means "no conclusion", see Eqs. (5) and (6) for definitions.

models and ARII-weighted combined models are 0.308 and 0.253; the average specificity for equal weighted combined models and specificity-weighted combined models are 0.235 and 0.320, but are not significantly different.

Secondly, using the SMOTE training sample improves the specificity, but not the ARII. The average ARIIs using the original dataset, the oversampling dataset, and the SMOTE dataset are 0.371, 0.403, and 0.244, respectively, while the average specificities using the three datasets are 0.306, 0.326, and 0.352, respectively.

Thirdly, training with the SMOTE dataset does bring some general benefits in terms of the number of detected PM failures. The average numbers of PM failures being identified by single model with the original, oversampling, and SMOTE dataset are 3.75, 4.75, and 4.75; and the average numbers by combined models are 1.88, 1.69, and 2.25, respectively. This observation signifies that using the SMOTE dataset makes both single and combined models more aggressive in predicting that a PM will fail.

Lastly, is there any model which outperforms the ones we found in Tables 5 and 6? It turns out that we cannot find any model with ARCI, ARII, or specificity of 1. Individually, the models using the SMOTE dataset do not have better performances than their counterparts in Table 6. However, models such as DT, C134, M23, and M24 do have higher ARCI or specificity than models using the original dataset.

In summation, the following general remarks clarify the application of our techniques in building classification models:

- Combined models are more conservative than single models in predicting PM failures.
- Using equal weightings when combining models is good enough to generate reliable combined models.
- Using the oversampling dataset to train our models enhances their ARII and specificity, suggesting that models are better at identifying PM failures in a more balanced way.
- Using the SMOTE dataset to train our models improves specificity but lowers the ARII, suggesting that models are more aggressive in predicting the PM failures.

Table 7
Out-of-sample test results—models trained with the SMOTE dataset.

Model	Equal weighted								ARCI weighted							
	TP	FP	FN	TN	ARCI	ARII	Specificity	NC	TP	FP	FN	TN	ARCI	ARII	Specificity	NC
(1) SVM	303	10	3	1	0.968	0.250	0.091	0								
(2) PCA-DA	249	6	57	5	0.976	0.081	0.455	0								
(3) Logit	284	7	22	4	0.976	0.154	0.364	0								
(4) DT	253	2	53	9	0.992	0.145	0.818	0								
M12	306	11	0	0	0.965	NA	0.000	0	303	10	3	1	0.968	0.250	0.091	0
M13	305	10	1	1	0.968	0.500	0.091	0	303	10	3	1	0.968	0.250	0.091	0
M14	303	10	3	1	0.968	0.250	0.091	0	303	10	3	1	0.968	0.250	0.091	0
M23	302	10	4	1	0.968	0.200	0.091	0	284	7	22	4	0.976	0.154	0.364	0
M24	292	6	14	5	0.980	0.263	0.455	0	253	2	53	9	0.992	0.145	0.818	0
M34	297	7	9	4	0.977	0.308	0.364	0	253	2	53	9	0.992	0.145	0.818	0
M123	301	9	5	2	0.971	0.286	0.182	0	301	9	5	2	0.971	0.286	0.182	0
M124	289	5	17	6	0.983	0.261	0.545	0	289	5	17	6	0.983	0.261	0.545	0
M134	295	7	11	4	0.977	0.267	0.364	0	295	7	11	4	0.977	0.267	0.364	0
M234	285	3	21	8	0.990	0.276	0.727	0	285	3	21	8	0.990	0.276	0.727	0
M1234	302	9	4	2	0.971	0.333	0.182	0	295	7	11	4	0.977	0.267	0.364	0
C123	229	3	0	0	0.987	NA	0.000	85								
C124	210	2	0	0	0.991	NA	0.000	105								
C134	240	2	1	1	0.992	0.500	0.333	73								
C234	198	2	3	1	0.990	0.250	0.333	113								
C1234	198	2	0	0	0.990	NA	0.000	117								
Model	ARII weighted								Specificity weighted							
	TP	FP	FN	TN	ARCI	ARII	Specificity	NC	TP	FP	FN	TN	ARCI	ARII	Specificity	NC
M12	303	10	3	1	0.968	0.250	0.091	0	303	10	3	1	0.968	0.250	0.091	0
M13	303	10	3	1	0.968	0.250	0.091	0	303	10	3	1	0.968	0.250	0.091	0
M14	303	10	3	1	0.968	0.250	0.091	0	303	10	3	1	0.968	0.250	0.091	0
M23	284	7	22	4	0.976	0.154	0.364	0	284	7	22	4	0.976	0.154	0.364	0
M24	253	2	53	9	0.992	0.145	0.818	0	253	2	53	9	0.992	0.145	0.818	0
M34	253	2	53	9	0.992	0.145	0.818	0	253	2	53	9	0.992	0.145	0.818	0
M123	301	9	5	2	0.971	0.286	0.182	0	301	9	5	2	0.971	0.286	0.182	0
M124	289	5	17	6	0.983	0.261	0.545	0	289	5	17	6	0.983	0.261	0.545	0
M134	295	7	11	4	0.977	0.267	0.364	0	295	7	11	4	0.977	0.267	0.364	0
M234	285	3	21	8	0.990	0.276	0.727	0	285	3	21	8	0.990	0.276	0.727	0
M1234	300	9	6	2	0.971	0.250	0.182	0	295	7	11	4	0.977	0.267	0.364	0

Note: (1) The top-left panel presents the results from models based on equal-weighted methods; the top-right panel is for the ARCI-weighted models; the bottom-left panel is for the ARII-weighted models; the bottom-right panel is for the specificity-weighted models. (2) For combined models, the prefix "M" denotes models following the majority rule, prefix "C" denotes models following the consensus rule; the numbering denotes the constituent models—"1" denotes the SVM model, "2" denotes the PCA-DA model, "3" denotes the Logit model, and "4" denotes the DT model. (3) The results do not vary under different weighting rules for models using consensus rule, so we only report them once in the Equal weighted panel. (4) NC means "no conclusion", see Eqs. (5) and (6) for definitions.

In addition to these general remarks, what really interests decision makers is not the techniques themselves, but the model discovered in our procedure. The final stage of our procedure is to select the most useful model to predict PM's correctness. We discuss this issue in the next section.

4. The best prediction models (under different criteria)

We have four measures to evaluate a model's ability to predict the correctness of PMs' predictions: ARCI, ARII, specificity, and NC (number of inconclusive cases). First of all, if a model can attain high ARCI, ARII, and specificity at the same time, its classifications will be very accurate no matter the type of prediction it is going to make about a PM. Table 8 lists the models that attain the highest values for these three target measures, with the highest values for these indexes all being 1.000. Among these models, **C234-O** is the only one whose ARCI, ARII, and specificity reach 1.000 at the same time.

C234-O combines the PCA-DA, Logit, and DT model under the consensus rule and is trained with the oversampling dataset (see Table 6 for detailed numbers). C234-O's high accuracy results from the rule it uses to combine models. It is a very conservative classification model because it only gives a prediction when all the constituent models have the same opinion. Its high accuracy is achieved at the expense of its applicability. Note that the number of NC (no conclusion) cases of C234-O is large—it cannot give definite predictions to 116 out of the 317 PM contracts. Despite this, C234-O may be employed by decision makers who want to minimize the risk of making erroneous judgments.

Although C234-O is very accurate, decision makers may suffer from one obvious problem in practice: it cannot provide definite predictions in every case (NC = 116). As a result, it only identifies 1 out of the 11 PM failures in the testing dataset (Table 3). For decision makers, it could be more essential to utilize a model that can always give predictions and has a high accuracy as well. For this reason, we discard models with positive NC cases, and select another set of candidate models in Table 9.

Table 8
Selection of the best model: no constraints.

ARCI-targeted				ARII-targeted				Specificity-targeted			
Model	ARCI	NC	TN	Model	ARII	NC	TN	Model	Specificity	NC	TN
C134-O	1.000	116	1	M13	1.000	0	1	C234-O	1.000	116	1
C234-O	1.000	117	1	C134	1.000	46	1				
C1234-O	1.000	117	0	M23-O	1.000	0	1				
				M123-O	1.000	0	1				
				M1234-O	1.000	0	1				
				C234-O	1.000	116	1				

Note: (1) We attach a letter "O" to a model to denote the model trained with the oversampling dataset; we attach a letter "S" to a model to denote the model trained with the SMOTE dataset; models without any letter attached are the ones trained with the original dataset. (2) NC stands for "no conclusion", and TN stands for "true negative".

Table 9
Selection of the best model: NC = 0.

ARCI-targeted				ARII-targeted				Specificity-targeted			
Model	ARCI	NC	TN	Model	ARII	NC	TN	Model	Specificity	NC	TN
DT-O	0.996	0	10	M13	1.000	0	1	DT-O	0.909	0	10
M24-O	0.996	0	10	M23-O	1.000	0	1	M24-O	0.909	0	10
M34-O	0.996	0	10	M123-O	1.000	0	1	M34-O	0.909	0	10
				M1234-O	1.000	0	1				

Note: (1) We attach a letter "O" to a model to denote the model trained with the oversampling dataset; we attach a letter "S" to a model to denote the model trained with the SMOTE dataset; models without any letter attached are the ones trained with the original dataset. (2) NC stands for "no conclusion", and TN stands for "true negative".

From Table 9, it is clear that there is no model which has high ARCI, ARII, and specificity at the same time. DT-O, M24-O, and M34-O have almost perfect ARCI and very high specificity. In fact, M24-O and M34-O can be entirely represented by **DT-O**, because these two combined models simply reflect DT-O's predictions under the majority rule with performance-based weightings (Table 6). DT-O is a single model (decision tree) trained with the oversampling dataset (see Table 6 for detailed numbers). If decision makers want to identify PM failures ex ante as much as possible, DT-O would be the first choice because it can identify 10 out of the 11 PM failures in the testing sample (specificity = 0.909).

In spite of DT-O's excellent ability in identifying PM failures, its success actually comes with a cost of prediction accuracy: Its ARII is only 0.167, which means that it reacts disproportionately to dubious PM contracts and therefore classifies too many contracts as incorrect (Table 6). The four models in the middle panel of Table 9 provide an alternative choice. These four models have exactly the same performance in every index (Tables 5 and 6), so we take M13 as the representative. **M13** is a combined model that integrates SVM and Logit's predictions following the majority rule and trained with the original dataset (see Table 5 for detailed numbers). If a decision maker wants a classification model with a high accuracy in predicting PM failures, and which always gives unequivocal predictions, M13 would be the ideal model because it has ARII of 1.000, but no NC cases.

To summarize, we find that there doesn't exist an optimal model which outperform others in every aspect within our study. We select **C234-O**, **DT-O**, and **M13** due to their superiority in distinct performance measures. If a decision maker is very cautious and doesn't want to make any erroneous conclusion when interpreting PM predictions, they may find C234-O the most applicable. If a decision maker wants to predict PM failures as accurately as possible for every case, M13 is an effective choice. If the decision makers want to identify PM failures as much as possible, even if they may sometimes have miscalculations, DT-O is the most suitable model. Of course, decision makers can always propose new criteria to meet their diverse needs and goals. For example, one may look for a model which has a medium ARII and specificity, or even design a composite objective function as the model selection criterion. No matter the nature of their purpose, we argue that decision makers will find a favorable prediction model of PM failures ex ante, by adhering to our methods.

5. Conclusion

Accurate forecasts are an essential part of superior decision making and risk management for individuals, organizations, societies and states. PMs have been praised for their ex post prediction accuracies of future events, and it appears that PMs could greatly assist critical decision making and risk management. Although historically, PM predictions are quite accurate on average, sometimes markets could be inefficient. Our knowledge about the patterns of market inefficiency is scarce. For example, in betting markets, there is no universal consistency in prediction bias from all published studies (Vaughan Williams, 1999). There is a need of prediction models which can identify market failures in advance, and our procedures can be used to find out such models.

For decision makers, actions must be taken to better manage the risk ex ante when facing uncertain future events. As a result, it is risky if real-world investors or decision makers take actions based on PMs' predictions

when we have almost no clues about when PMs may fail. Therefore, we need analytical tools to assess the accuracy or correctness of PM predictions prior to the events themselves. Such assessments could then serve as a basis for advantageous decision making and risk management.

This paper examined a procedure determining a reliable model that can predict a winner-take-all market's prediction correctness ex ante. Following this procedure, our aim was to discover a model that classified a WTA PM as correct or incorrect in advance with high accuracy. We also proposed three performance measures—ARCI, ARII, and specificity—to assess a model's classification accuracy.

This procedure consisted of five stages:

(1) sample adjustment
(2) model training
(3) model combinations
(4) out-of-sample tests
(5) selection of the best model.

In the sample adjustment stage, we deal with the imbalanced data problem commonly known in classification problems with two techniques: oversampling and SMOTE. By oversampling, we duplicated the records of PM failures in the training dataset several times, so that we have roughly the same sizes for the correct and incorrect categories. By SMOTE, we increased the records of PM failures in the training dataset by artificially creating synthesized data. In the model training stage, four single classification models—SVM, PCA-DA, Logit, and DT—were constructed and trained with our datasets to assess the prediction correctness of our WTA PMs. We used data from xFuture, which is the biggest Chinese prediction market platform based in Taiwan, to train and test our models. Our dataset has 650 election contracts from elections held between 2006 and 2010, and half of them were used as the source of the training dataset. In the model combination stage, we used two different rules—the majority rule and the consensus rule—to combine the predictions from the four trained models using different weighting rules. In the out-of-sample test, we examined the prediction performances of our single and combined models in terms of the three performance indexes. In the model selection stage, we evaluated and compared the classification models using different criteria based on different decision goals.

We do acknowledge that within the scope of our research there may be limitations in our methods. First, our models use a large set of explanatory variables, to which an ordinary decision maker may not have access. Secondly, it is possible that different models would be selected using different training and testing datasets, even under the same decision objective. As a result, there may not be a "best" model that can classify all prediction markets universally well. To understand how these two issues impact the generalizability of our findings, with more experiments using additional datasets with a cautious selection of variables, is a task for the future.

Nevertheless, despite the constraints of our methods, this study clarifies significant benefits to the utilization of prediction markets. Through this procedure, we found that combined models are more conservative than single models in predicting the failures of PM, and equal weighting is good enough to generate reliable combined models. Furthermore, oversampling and SMOTE techniques both improved models' ability to identify PM failures in advance, though by different means. Lastly but most importantly, we demonstrated that there are models which predict PM failures with high accuracy. C234-O is a very accurate model in the sense that it never makes erroneous classifications, although it may not provide definite predictions for every PM contract. M13 can provide predictions for every WTA contact and has the same accuracy as C234-O with a lower specificity. DT-O, on the other hand, succeeds in predicting almost all PM failures in advance. We therefore conclude that our procedure can effectively discover reliable models whose predictions can successfully serve as a basis for decision making and risk management.

Our procedure is flexible in the following sense. First, which model is the "best" classification model actually depends on decision makers' intentions, and decision makers can always propose new criteria to meet their distinct needs and goals. Secondly, excepting the Logit model, our procedure does not make assumptions about the data. As a result, one may improve the accuracy of combined models by introducing more classification models. Thirdly, this procedure can be used during the course of an actual election, wherein the decision makers can evaluate a prediction market in real time.

Unlike vote-share markets of which statistical models can be built to evaluate the prediction accuracy, how to assess the prediction accuracy of winner-take-all markets is a challenging task for both researchers and practitioners of prediction markets. With a large set of explanatory variables which extract information not contained within market prices, our first contribution is to show that reliable models can be discovered to provide assessments of PM predictions in advance. Secondly, while market price itself contains information about people's beliefs, our results suggest that non-price factors contains valuable information about market efficiency. A future step would be an anatomical examination of these models. With hindsights learned from these models, we may contribute to the understandings of why PMs fail from the theoretical perspective.

Acknowledgments

This research was supported by Ministry of Science and Technology, Taiwan, R.O.C., grant number MOST 104-2410-H-004-093-MY2. The first author also acknowledges financial support from Ministry of Science and Technology, Taiwan, R.O.C., grant MOST 104-2410-H-029-002.

Appendix. Descriptive statistics of the explanatory variables

See Table A.10.

Table A.10
Descriptive statistics of the explanatory variables.

Variable	Mean	Maximum	Minimum	Standard deviation	Kurtosis	Skewness
GP_share_lyc_100	0.002	0.083	0	0.007	70.51	7.69
GP_share_lyc_200	0.003	0.083	0	0.008	39.35	5.59
GP_share_lyc_300	0.003	0.083	0	0.009	30.36	4.89
GP_share_lyc_1	0.003	0.083	0	0.009	29.73	4.82
GP_share_lyc_5	0.02	0.2	0	0.032	5.3	2.03
GP_share_lyc_10	0.078	0.5	0	0.081	3.23	1.46
GP_share_365d_100	0.001	0.04	0	0.003	68.42	7.21
GP_share_365d_200	0.001	0.04	0	0.004	34.27	5.16
GP_share_365d_300	0.004	0.5	0	0.023	338.44	16.39
GP_share_365d_1	0.004	0.5	0	0.023	339.23	16.41
GP_share_365d_5	0.019	0.5	0	0.037	54.62	5.42
GP_share_365d_10	0.051	0.5	0	0.065	9.61	2.4
GP_share_30d_100	0.031	0.5	0	0.055	13.14	3.06
GP_share_30d_200	0.04	0.5	0	0.06	8.85	2.49
GP_share_30d_300	0.1	1	0	0.127	5.56	2
GP_share_30d_1	0.001	0.083	0	0.007	83.58	8.47
GP_share_30d_5	0.008	0.143	0	0.02	16.72	3.8
GP_share_30d_10	0.036	0.5	0	0.058	10.49	2.73
Limit_ratio_volume	0.841	1	0.246	0.139	−0.33	−0.62
WBAS2_all	39.275	99.879	−98.766	33.498	−0.02	−0.14
Buy_sell	247.439	94315.283	0	4132.237	431.28	20.01
Trades	559.549	36057	1	2385.413	96.73	8.57
Traders	128.432	11074	2	564.435	230.12	13.26
Days	108.755	675	15	131.771	6.43	2.47
Volume	1.67E+05	2.40E+07	1	1.16E+06	299.01	15.83
Two_way	0.212	1	0	0.166	0.48	0.49
IP_share	0.259	1	0	0.194	1.72	1.18
Traded_order_ratio	0.506	0.818	0.029	0.163	−0.48	−0.4
Highest-price	85.106	100	20.32	17.684	1.27	−1.32
NC	6.017	19	2	4.332	2.51	1.82
Price_gap	67.737	99.98	0.379	32.581	−0.99	−0.68
Avatar_ratio_3	0.447	1	0	0.326	−0.92	0.57
Avatar_15d_ratio_3	0.324	1	0	0.224	−0.55	0.1
Avatar_30d_ratio_3	0.334	1	0	0.216	−0.36	0.09
Avatar_365d_ratio_3	0.373	1	0	0.185	0.09	0.08
Avatar_volume_ratio_3	0.655	1	0	0.332	−0.86	−0.66
Avatar_volume_15d_ratio_3	0.566	1	0	0.4	−1.52	−0.34
Avatar_volume_30d_ratio_3	0.589	1	0	0.385	−1.37	−0.43
Avatar_volume_365d_ratio_3	0.654	1	0	0.333	−0.87	−0.65
P^w	25.254	100	0	37.307	−0.46	1.12

References

Ali, M. M. (1979). Some evidence on the efficiency of a speculative market. *Econometrica*, *47*(2), 387–392.

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 417–439). Kluwer Academic Publishers.

Asch, P., & Quandt, R. E. (1987). Efficiency and profitability in exotic bets. *Economica*, *54*(215), 278–298.

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations*, *6*(1), 20–29.

Berg, J., Forsythe, R., & Rietz, T. (1997). What makes markets predict well? Evidence from the Iowa electronic markets. In W. Albers, W. Güth, P. Hammerstein, B. Moldovanu, & E. van Damme (Eds.), *Understanding strategic interaction: Essays in honor of Reinhard Selten* (pp. 444–463). New York: Springer.

Berg, J. Nelson, F., Rietz, T. (2003). Accuracy and forecast standard error of prediction markets. Working paper, Henry B. Tippie College of Business Administration, University of Iowa. Retrived 2017, from https://www.biz.uiowa.edu/faculty/trietz/papers/forecasting.pdf.

Berg, J., Nelson, F., & Rietz, T. (2008). Prediction market accuracy in the long run. *International Journal of Forecasting*, *24*(2), 285–300.

Berkowitz, J. P., Depken, C. A., II, & Gandar, J. M. (2017). A favorite-longshot bias in fixed-odds betting markets: Evidence from college basketball and college football. *Quarterly Review of Economics and Finance*, *63*, 233–239.

Bird, R., & McCrae, M. (2008). Efficiency of racetrack betting markets: Australian evidence. In D. B. Hausch, V. S. Lo, & W. T. Ziemba (Eds.), *Efficiency of racetrack betting markets* (pp. 575–582). Singapore: World Scientific.

Blackburn, P., Peirson, J. (1995). Betting at British racecourses: An analysis of semi-strong efficiency between bookmaker and tote odds. Working paper 95/4. Department of Economics, University of Kent at Canterbury.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.

Brown, A., Rambaccussing, D., Reade, J. J., Rossi, G. (2016). Using social media to identify market inefficiencies: Evidence from Twitter and Betfair. Working papers 2016-002. The George Washington University, Department of Economics, Research Program on Forecasting. URL: https://ideas.repec.org/p/gwc/wpaper/2016-002.html.

Busche, K. (2008). Efficient market results in an Asian setting. In D. B. Hausch, V. S. Lo, & W. T. Ziemba (Eds.), *Efficiency of racetrack betting markets* (pp. 615–616). Singapore: World Scientific.

Busche, K., & Hall, C. D. (1988). An exception to the risk preference anomaly. *Journal of Business*, *61*(3), 337–346.

Busche, K., & Walls, W. D. (2000). Decision cost and betting market efficiency. *Rationality and Society*, *12*, 477–492.

Cain, M., Law, D., & Peel, D. A. (1997). Insider trading and market efficiency in British racetrack betting. *Salford Papers in Gambling Studies*, *1*, 97–101.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*,

27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chawla, N. V. (2003). C4.5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure. In *ICML Workshop on Learning from Imbalanced Data Sets II*.

Chawla, N. V. (2005). Data mining for imbalanced datasets: an overview. In O. Maimon, & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (pp. 853–867). Boston, MA: Springer.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, S.-H., Tung, C.-Y., Tai, C.-C., Chie, B.-T., Chou, T.-C., & Wang, S. G. (2011). Prediction markets: a study on the Taiwan experience. In L. V. Williams (Ed.), *Prediction markets: Theory and applications* (pp. 137–156). Routledge.

Clemen, R. T. (1989). Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Crone, S. F., & Finlay, S. (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting*, *28*(1), 224–238.

Croxson, K., & Reade, J. J. (2014). Information and efficiency: Goal arrival in soccer betting. *Economic Journal*, *124*(575), 62–91.

Deck, C., Lin, S., & Porter, D. (2013). Affecting policy by manipulating prediction markets: experimental evidence. *Journal of Economic Behavior & Organization*, *85*, 48–62.

Delen, D., Cogdell, D., & Kasap, N. (2012). A comparative analysis of data mining methods in predicting NCAA bowl outcomes. *International Journal of Forecasting*, *28*(2), 543–552.

Dillon, W. R., Mulani, N., & Frederick, D. G. (1989). On the use of component scores in the presence of group structure. *Journal Consumer Research*, *16*(1), 106–112.

Drummond, C., Holte, R. (2003) C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *ICML Workshop on Learning from Imbalanced Data Sets II*.

Forsythe, R., Nelson, F., Neumann, G. R., & Wright, J. (1992). Anatomy of an experimental political stock market. *American Economic Review*, *82*(5), 1142–1161.

Forsythe, R., Rietz, T. A., & Ross, T. W. (1999). Wishes, expectations and actions: a survey on price formation in election stock markets. *Journal of Economic Behavior & Organization*, *39*(1), 83–110.

Franck, E., Verbeek, E., & Nesch, S. (2010). Prediction accuracy of different market structures — bookmakers versus a betting exchange. *International Journal of Forecasting*, *26*(3), 448–459.

Gabriel, P. E., & Marsden, J. R. (1990). An examination of market efficiency in British racetrack betting. *Journal of Political Economy*, *98*(4), 874–885.

Gabriel, P. E., & Marsden, J. R. (1991). An examination of market efficiency in British racetrack betting: errata and corrections. *Journal of Political Economy*, *99*(3), 657–659.

Gandar, J. M., Zuber, R. A., & Dare, W. H. (2000). The search for informed traders in the totals betting market for National Basketball Association games. *Journal of Sports Economics*, *1*(2), 177–186.

Gandar, J. M., Zuber, R. A., & Lamb, R. P. (2001). The home field advantage revisited: A search for the bias in other sports betting markets. *Journal of Economics and Business*, *53*(4), 439–453.

Graefe, A., Armstrong, J. S., Jones, R. J., Jr., & Cuzán, A. G. (2014). Combining forecasts: an application to elections. *International Journal of Forecasting*, *30*(1), 43–54.

Greene, W. H. (2017). *Econometric analysis* (8th ed.). Pearson.

Griffiths, W. E., Hill, R. C., & Judge, G. G. (1993). *Learning and practicing econometrics* (1st ed.). Wiley.

Gruca, T. S., Berg, J. E., & Cipriano, M. (2005). Consensus and differences of opinion in electronic prediction markets. *Electronic Markets*, *15*(1), 13–22.

Gürkaynak, R., Wolfers, J. (2005). Macroeconomic derivatives: an initial analysis of market-based macro forecasts, uncertainty, and risk. Working paper 2005-26, Federal Reserve Bank of San Francisco. Retrieved 2017, from http://www.nber.org/chapters/c0355.pdf.

Ho, T.-H., & Chen, K.-Y. (2007). New product blockbusters: the magic and science of prediction markets. *California Management Review*, *50*(1), 144–158.

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, *11*(1), 94.

Kambil, A., & Heck, E. V. (2002). *Making markets: How firms can design and profit from online auctions and exchanges*. Cambridge, MA: Harvard Business School Press.

Ledyard, J. O. (2006). Designing information markets for policy analysis. In R. Hahn, & P. Tetlock (Eds.), *Information markets: A new way of making decisions* (pp. 37–66). AEI-Brookings Joint Center.

Leigh, A., & Wolfers, J. (2006). Competing approaches to forecasting elections: economic models, opinion polling and prediction markets. *Economic Record*, *82*, 325–340.

Lin, H.-W., Tung, C.-Y., Lin, J.-W., & Cho, T.-C. (2015). Which factors best predict election results? A case study of 2008-2010 Taiwan election prediction markets. *Taiwan Journal of Democracy*, *12*(2), 87–122 (in Chinese).

Lin, H.-W., Tung, C.-Y., & Yeh, J. (2013). Multivariate methods in assessing the accuracy of prediction markets ex ante based on the highest-price criterion. *Journal of Prediction Markets*, *7*(3), 29–44.

Lo, V. S., & Busche, K. (2008). How accurately do bettors bet in doubles? In D. B. Hausch, V. S. Lo, & W. T. Ziemba (Eds.), *Efficiency of racetrack betting markets* (pp. 465–468). Singapore: World Scientific.

Loh, W.-Y. (2014). Fifty years of classification and regression trees. *International Statistical Review*, *82*(3), 329–348.

Luckner, S., Weinhardt, C., & Studer, R. (2006). Predictive power of markets: a comparsion of two sports forecasting exchanges. In T. Dreier, R. Studer, & C. Weinhardt (Eds.), *Information Management and market engineering* (pp. 187–195). Karlsruhe: Karlsruhe University Press.

Maloof, M. A. (2003). Learning when data sets are imbalanced and when costs are unequal and unknown. In *ICML workshop on learning from imbalanced data sets II*.

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data and a proposal. *Journal of the American Statistical Association*, *58*(302), 415–434.

Newman, B. I., & Sheth, J. N. (1985). A model of primary voter behavior. *Journal Consumer Research*, *12*(2), 178–187.

Oliven, K., & Rietz, T. A. (2004). Suckers are born but markets are made: individual rationality, arbitrage, and market efficiency on an electronic futures market. *Management Science*, *50*(3), 336–351.

Ortner, G. (1998). Forecasting markets: an industrial application. Working paper, Technical University of Vienna.

Ottaviani, M., & Sørensen, P. N. (2008). The favorite-longshot bias: An overview of the main explanations. In D. B. Hausch, & W. T. Ziemba (Eds.), *Handbook of sports and lottery markets* (pp. 83–101). Amsterdam: North-Holland.

Ottaviani, M., & Sørensen, P. N. (2010). Noise, information, and the favorite-longshot bias in parimutuel predictions. *American Economic Journal: Microeconomics*, *2*(1), 58–85.

Pennock, D. M., Lawrence, S., Giles, C. L., & Nielsen, F. A. (2001). The real power of artificial markets. *Science*, *291*(5506), 987–988.

Rahman, M. M., & Davis, D. N. (2013). Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing*, *3*(2), 224–228.

Rhode, P. W., & Strumpf, K. S. (2004). Historical presidential betting markets. *Journal of Economic Perspectives*, *18*(2), 127–141.

Rosen, D. L., & Granbois, D. H. (1983). Determinants of role structure in family financial management. *Journal Consumer Research*, *10*(2), 253–258.

Sauer, R. D. (1998). The economics of wagering markets. *Journal of Economic Literature*, *36*(4), 2021–2064.

Schnytzer, A., & Weinberg, G. (2008). Testing for home team and favorite biases in the Australian rules football fixed-odds and point spread betting markets. *Journal of Sports Economics*, *9*(2), 173–190.

Servan-Schreiber, E., Wolfers, J., Pennock, D. M., & Galebach, B. (2004). Prediction markets: does money matter? *Electronic Markets*, *14*(3), 243–251.

Smith, M. A., Paton, D., & Vaughan Williams, L. (2009). Do bookmakers possess superior skills to bettors in predicting outcomes? *Journal of Economic Behavior & Organization*, *71*(2), 539–549.

Snowberg, E., & Wolfers, J. (2010). Explaining the favorite-longshot bias: Is it risk-love or misperceptions? *Journal of Political Economy*, *118*(4), 723–746.

Snowberg, E., Wolfers, J., & Zitzewitz, E. (2005). Information (in)efficiency in prediction markets. In L. V. Williams (Ed.), *Information efficiency in financial and betting markets* (pp. 366–386). Cambridge, UK: Cambridge University Press.

Snyder, W. W. (1978). Horse racing: Testing the efficient markets model. *Journal of Finance*, *33*(4), 1109–1118.

State, L., Cocianu, C., & Fusaru, D. (2010). A survey on potential of the support vector machines in solving classification and regression problems. *Informatica Economica*, *14*(3), 128–139.

Su, X., Azuero, A., Cho, J., Kvale, E., Meneses, K. M., & McNees, M. P. (2011). An introduction to tree-structured modeling with application to quality of life (QOL) data. *Nursing Research*, *60*(4), 247–255.

Tai, C.-C., Chih, P.-T., Lin, H.-W., & Tung, C.-Y. (2016). Assessing the accuracy of prediction markets: single versus combined identification models. *Taiwan Economic Review*, *44*(3), 413–474 (in Chinese).

Tung, C.-Y., Chou, T.-C., Lin, J.-W., & Lin, H.-Y. (2011). Comparing the forecasting accuracy of prediction markets and polls for Taiwans presidential and mayoral elections. *Journal of Prediction Markets*, *5*(3), 1–26.

Vaughan Williams, L. (1999). Information efficiency in betting markets: A survey. *Bulletin of Economic Research*, *51*(1), 1–39.

Vaughan Williams, L., & Paton, D. (1997). Does information efficiency require a perception of information inefficiency? *Applied Economics Letters*, *4*(10), 615–617.

Vaughan Williams, L., & Paton, D. (1998). Why are some favourite-longshot biases positive and others negative? *Applied Economics*, *30*(11), 1505–1510.

Vaughan Williams, L., & Reade, J. J. (2016a). Forecasting elections. *Journal of Forecasting*, *35*(4), 308–328.

Vaughan Williams, L., & Reade, J. J. (2016b). Prediction markets, social media and information efficiency. *Kyklos*, *69*(3), 518–556.

Walls, W. D., & Busche, K. (2003). Breakage, turnover, and betting market efficiency: New evidence from Japanese horse tracks. In L. Vaughan Williams (Ed.), *The economics of gambling* (pp. 43–66). New York: Routledge.

Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, *30*(4), 1030–1081.

Wilton, P. C., & Pessemier, E. A. (1981). Forecasting the ultimate acceptance of an innovation: the effects of information. *Journal Consumer Research*, *8*(2), 162–171.

Wolfers, J., & Leigh, A. (2002). Three tools for forecasting federal elections lessons from 2001. *Australian Journal of Political Science*, *37*(2), 223–240.

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives*, *18*(2), 107–126.

Wolfers, J., Zitzewitz, E. (2006). Prediction markets in theory and practice. NBER Working papers 12083. National Bureau of Economic Research, Inc. URL: https://ideas.repec.org/p/nbr/nberwo/12083.html.

Woodland, L. M., & Woodland, B. M. (1994). Market efficiency and the favorite-longshot bias: The baseball betting market. *Journal of Finance*, *49*(1), 269–279.

Zhao, W., Chellappa, R., Krishnaswamy, A. (1998). Discriminant analysis of principal components for face recognition. In *Proceedings of international conference on automatic face and gesture recognition* (pp. 336–341).

Please cite this article in press as: Tai, C.-C., et al., Predicting the failures of prediction markets: A procedure of decision making using classification models. International Journal of Forecasting (2018), https://doi.org/10.1016/j.ijforecast.2018.04.003.