Predicting Sports Outcomes with a Rank-And-Choose Variable Selection Process

Tao Lin, Ricardo Trujillo, Haibo Wang
R. Sanchez, Jr. School of Business, Texas A&M International University, Laredo, Texas, USA

Abstract - *Predicting outcomes of games or championship is of great interest to fans, sponsors and media. There are many factors in determining the outcomes of each game or championship. We develop a logistics regression model with a rank-and-choose variable selection process in this study. We implemented this model in open source R language and computed the predictive variable efficiencies from historical data, then test the predictive model for 2015 NBA championship series data with the comparison of Microsoft Bing search engine prediction. The approach we were using for solving this problems includes: 1. Gathering the historical performance data of all teams from the official stats website; 2. Using LASSO method to select independent variables that are strongly correlative to the dependent variables; 3. Building a regression model with R with the historical data, then plug in the data of this season to accomplish the prediction.*

Keywords: Prediction; Sports games; Variable selection; Liner regression; LASSO; R

1 Research Design And Methodology

1.1 Data collection:

We collected the historical team performance data from the official statistics website: NBA.com/stats. We mainly focus on the historical performance data of the champions of each season.

1.2 Research tools:

We use a popular statistics analysis tool—R to do the analysis work. With R, we use the data we collected to select variables, build regression model, test accuracy then plug the data to predict.

1.3 Model Evaluation:

We used LASSO method for selecting variables, by adjusting the model, overall error rate is about 4.89%, and got the same result with Microsoft Bing search engine.

2 System Modeling And Results

2.1 Data collection:

Our data is collected from the official website of NBA (http://stats.nba.com/). As mentioned above, we collect the history performance data including the number of winning games and some remarkable technique indicators. As the independent variables for prediction, we also collected the performance data of the playoff teams in season 2014-2015.

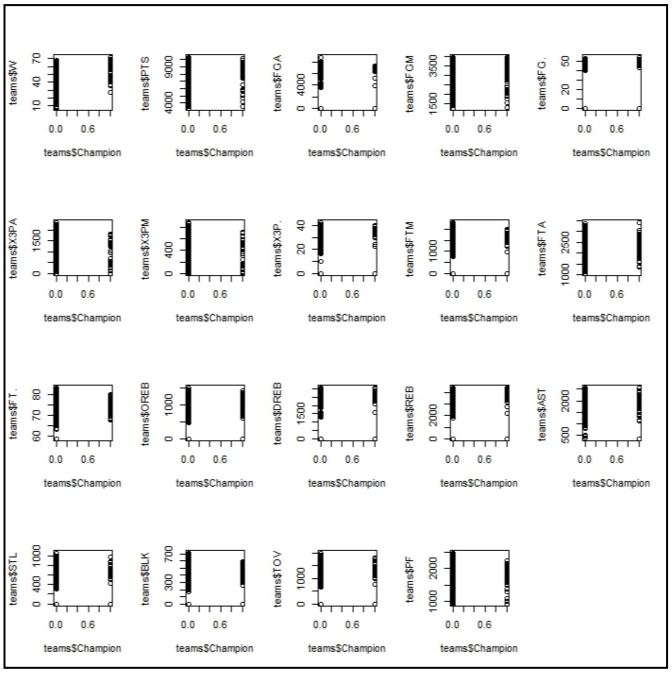
2.2 Data understanding:

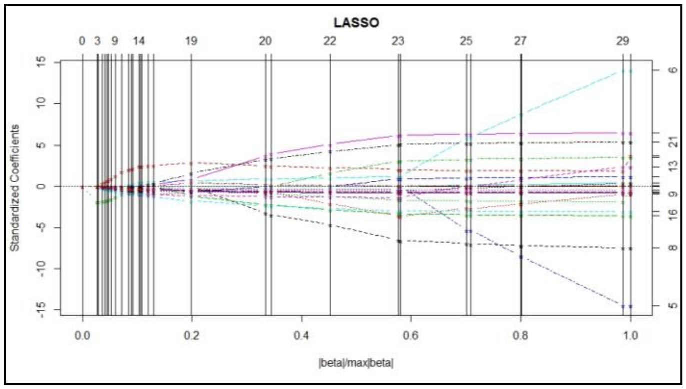
For understanding the relationship of the independent variables, we used multiple plot to find out the correlations between each technique indicators and champion winner.

2.3 Model analyzing:

LASSO is a good tool to find out the relationship among variables and a good combination which can provide good results. In this case, the graph shows there are several key variables and made significant contribution to the model.

To optimize our model, we made 3 cases with different number of variables. At the same time, we made a function to





research factors effects of the sample ratio, adjustment and sampling randomness.

We also divide the data set into training group and test group by the sampling ratio, then use the training data to build a Generalized Linear Model.

2.4 Model testing:

By adjusting the model and trying the factors, the most optimized model will includes all key variables with the specific set of model factors.

2.5 Building Model and Prediction

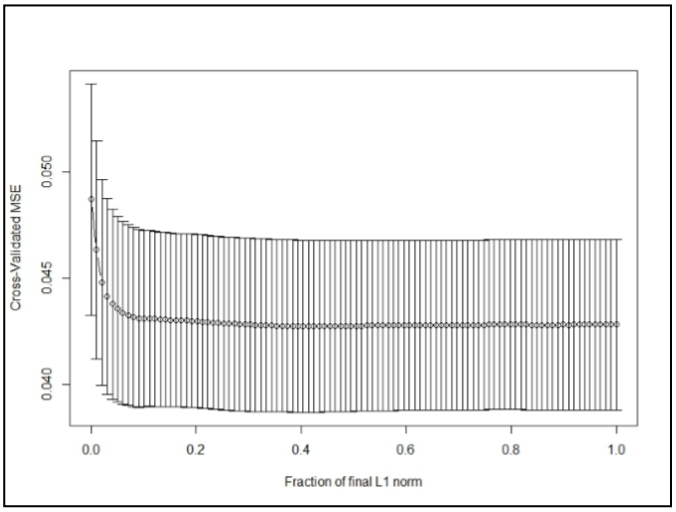
We used the 14-15 seasons play-off teams technique indicators to plug in the model, the result shows the Golden State Worriers is most possibly to be the champion of season 2014-2015.

3 Conclusions

As the conclusion, we used the all the key variables such as number of winning, total points, field shoot attempt, 3 points shoot accuracy, offence and defense rebound, steal and block.

We put the sampling rate and adjustment to 80%, also we use seed 4 as the sampling randomness. In this condition, the error rate will be 4.89% which is the optimized condition.

We also gathered the performance data of the playoff teams and plug into the model, and we have the prediction that the Golden state worriers is most likely to be the champion of season 2014-2015.



4 Future Work

We will focus on improve our model's accuracy by reevaluating other technique indicators, try to find more efficient indicators.

We will also try to use this method to do more researches on other popular sports event.

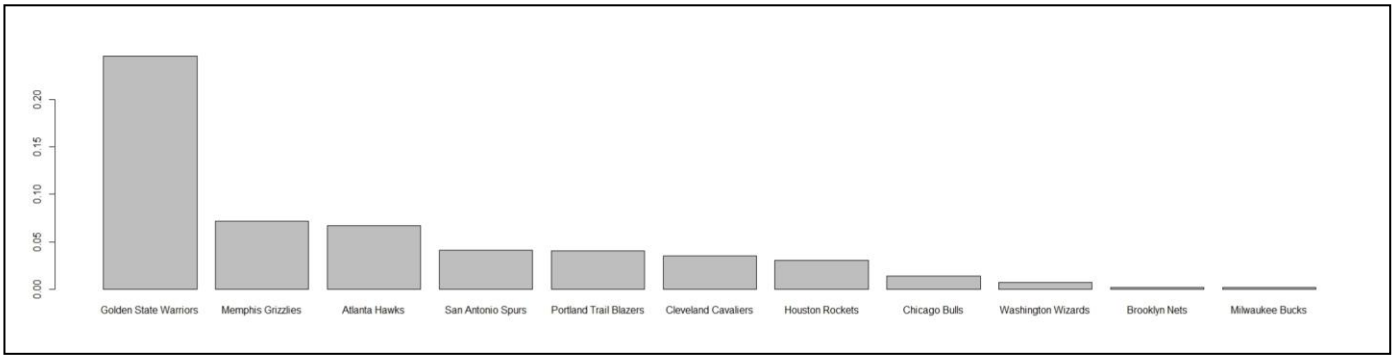
5 References

[1] Kabacoff, R. , (2011). R in action: Data analysis and graphics with R. [Online]. [May. 10, 2015].

[2] Ledolter, Johannes. , Data Mining and Business Analytics with R..

[3] Wikimedia Foundation,Linear Regression [Online]. Available: https://en.wikipedia.org/wiki/Linear_regression

[4] NBA.com,NBA.com/Stats [Online]. Available: NBA.com/Stats.


Golden State Warriors
Memphis Grizzlies
Atlanta Hawks
San Antonio Spurs
Portland Trail Blazers
Cleveland Cavaliers
Houston Rockets
Chicago Bulls
Washington Wizards
Brooklyn Nets
Milwaukee Bucks