

2018 INFORMS International Conference
Taipei, Taiwan, June 17–20, 2018
© 2018 INFORMS| ISBN 978-0-9906153-1-6
https://doi.org/10.1287/infp.2018.XXXX
pp. 000–000

Discovering the Factors Influencing Professional Baseball Box Office using Machine Learning

Hwai-Jung Hsu
Department of Information Engineering and Computer Science, Feng Chia University, Taichung 40724, Taiwan hjhsu@mail.fcu.edu.tw

Abstract In the past, the data analyses in professional baseball, i.e *sabermetrics*, focus on evaluation of players and organizing a team for winning more games because good players bring wins and winning brings business. However, winning might not be sufficient. The previous study about annual box office in Chinese Professional Baseball League shown that the franchise reputation and playing styles matter. In this paper, the factors influencing per game attendance in CPBL are further studied using machine learning techniques. As a result, we found that the game statistics have no influences to attendance. As long as the weather is during weekend, fans would attend the ball park to support the franchise highly address the local business.

Keywords Sabermetrics, Data-driven Analysis of Causes, Professional Baseball Box Office

1. Introduction

In the past, the data analyses in professional baseball, i.e *sabermetrics*, focus on evaluation of players and organizing a team for winning more games. It is because most franchises believe that good players bring wins and winning brings fans [3]. However, not only winning influences the professional baseball game attendance but also outcome uncertainty, size, and quality of the stadium, and playing styles do [1]. The stories of Oakland Athletics told us that winning does not always bring business. During 1998 to 2008, the Athletics won 89 games in average with a tiny payroll of 53 million US dollars (which is merely one-third of the payroll of New York Yankees). Billy Beane, the GM (General Manager) of Oakland Athletics that time, beats 99.4% of MLB (Major League Baseball) managers in efficiency [7]. Nevertheless, Beane's strategies in winning the most games with a relative low payroll also generated 1,600 attendance loss per game, i.e. Athletics lost 2.5 million US dollars in box-office per game in exchange of such winning-payroll efficiency [7].

Except for MLB, Hsu et. al found that in CPBL, the domestic professional baseball league in Taiwan, the reputation of the franchise and the playing styles with excitement like home runs, base stealing, and excellent fielding, etc. play the most important roles in annual number of attendance [4]. However, annual number of attendance is rough in granularity. In this paper, we further analyze the factors influencing *per game* attendance within a year in CPBL. In [3], Davis shown that weather can be essential to the number of per game attendance in MLB. Therefore, besides the statistics collected and calculated from games, time, place, and weather are also considered for establishing models predicting the per game attendance of CPBL in 2016 and 2017. Less important features are then removed from the models using backward feature elimination, and finally the features influencing per game attendance in CPBL are revealed.

FIGURE 1. The progress of feature selection for 2016.

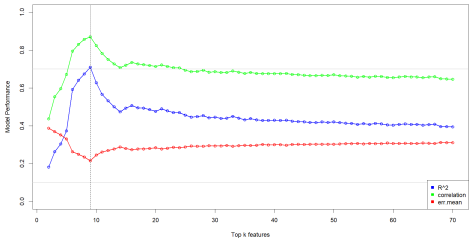
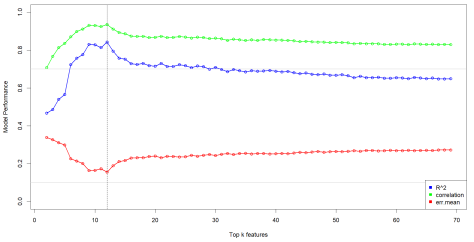


FIGURE 2. The progress of feature selection for 2017.



This paper is organized as following: An overview of the data used for analysis is made in section 2. In section 3, we described how backward feature selection is applied for obtaining a prediction model with essential features. Finally, the results is shown and discussed in section 4, and a conclusion is made in section 5.

2. Data Overview

The raw data for game information and game statistics are are crawled from the official site of CPBL [2]. For the total 480 games in 2016 and 2017, totally 11 game information, including the *dates (year, month, and day), weekdays, parks, game durations, number of attendance, whether the game is delayed, and the names for home and guest teams* are gathered. As for the game statistics, 29 game statistics including *number of runs, hits, errors, at bats, runs batted in (RBI), 2Bs, 3Bs, home runs, grounded into double play (GIDP), hitters being walked, hit by pitch, stroke-outs, sacrifice bunt (SAC), sacrifice flies (SF), stolen bases (SB), counter base stealing, innings of pitches, batters faced, pitches, strikes, opponent hits, opponent home runs, walks allowed, hits by pitch made, strike-outs get, wild pitches, balks, runs allowed, and earned runs* are collected for both the home and the guest teams.

The raw-data are then further calculated into 34 cumulative seasonal and recent (the last 10 games) statistics for both teams competing to each other in a game, including *winning percentage, batting average (BA), on base percentage (OBP), slugging rate (SLG), home runs per at bat, walks to stroke-out ratio, earned runs average (ERA), strike-outs per 9 innings, walks allowed per 9 inning, opponents' OBP, SLG, home runs per 9 innings, and average number of SB, GIDP, SAC and average durations per game.*

Besides the game information and statistics, the weather data, including *temperature, humidity, air pressure, wind speed, and rainfall* recorded by the stations of the Central Weather Bureau are obtained from data bank for atmospheric and hydrologic research [6]. We used the weather data gathered by the stations located around the game park (within 5 km radius) during the game day afternoon for each CPBL game in 2016 and 2017. All the 5 weather data used are averaged as the indicators of the weather of the game.

In order to compare the the influences among different features, all the numeric features are normalized using standard score, and all the categorical features are transformed into dummy variables.

3. Model Establishment and Feature Selection

The models predicting per game attendance of CPBL in 2016 and 2017 are trained separately. First, the recursive feature elimination (RFE) method in R package caret [5] is adopted for feature ranking. The features which are eliminated earlier in the process of RFE are considered less important and vice versa. After all the features are ranked, we train models to predict the per game attendance with the top k features (k is ranging from 2 to 69 which is the number of all features). 10-fold cross validation is used for evaluating the model

FIGURE 3. The predictive results for 2016.

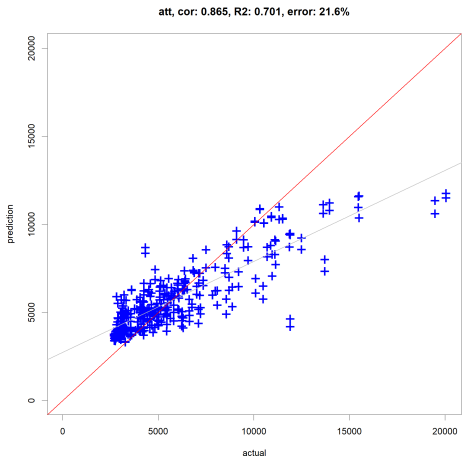
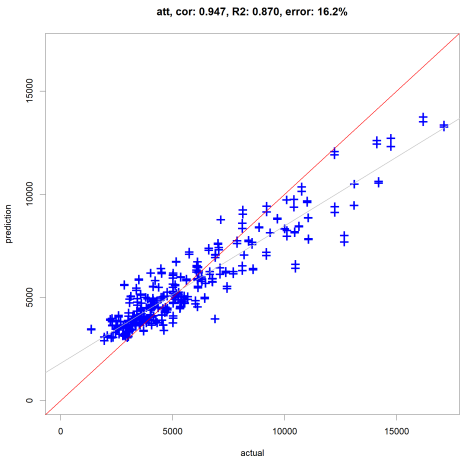


FIGURE 4. The predictive results for 2017.



performance, and the model with the highest coefficient of determination (R Squared) is selected as the best model for the year.

Figure 1 and Figure 2 shows the results of feature selection. As a result, the best predictive models for 2016 uses top 9 features and the 2017 model uses the top 12 features for prediction. The details of the results are discussed in the next section.

4. Results and Discussion

The best predictive model for 2016 uses 9 features including whether the game was played at Taoyuan or Tainan park, whether the opponent is Brother Elephants, whether the game was played on Saturday, and the 5 weather data, temperature, wind speed, air pressure, humidity, and amount of rainfalls. The predictive results for 2016 model is shown in Figure 3, and the R^2 of 2016 model is 0.701.

On the other hand, The best predictive model for 2017 uses 12 features including whether the game was delayed, whether the game was played at Taoyuan, Taichung, or Douliu park, whether the game was played on Saturday, Sunday, Thursday, or Friday, and the 4 of the 5 weather data, temperature, wind speed, air pressure, humidity. The predictive results for 2017 model is shown in Figure 4, and the R^2 of 2017 model is amazingly 0.842.

For both of the years, no game statistics (no matter it is seasonal or recent) are selected into the best predictive model. The opponent factor only shows in the 2016 model, but time, place and weather are essential to both of the years. The attendance increases on weekends and decreases during weekdays. Taoyuan park, the home park of Lamigo Monkeys which is the franchise being famous in running home game marketing well, has positive influences to the box-office. Weather is the most important factors affecting whether fans would go watching a game at the ball park. Without the weather data the best predictive model for both the years has only R^2 less than 0.6.

5. Conclusion

In this paper, we give a brief analysis of the factors influencing the per game attendance in CPBL. Different from the similar analysis made for annual box office, no game features are included in the models with the best predictive performance to the per game attendance. Besides, since the Taoyuan park is run by the franchise address the local business the most. Our results shows that as long as the weather on weekends is good, fans are willing to attend the ball park running by the franchise dedicating to local business the most. In the future,

we will further collect data about franchise business to understand how business activities like promotion, advertisement, and commodities may influence the attendance.

References

[1] Ahn SC, Lee YH (2014) Major league baseball attendance: Long-term analysis using factor models. *Journal of sports economics* 15(5):451–477.
[2] CPBL (2018) URL http://www.cpbl.com.tw/.
[3] Davis MC (2008) The interaction between baseball attendance and winning percentage: A var analysis. *International Journal of Sport Finance* 3(1):58.
[4] Hsu HJ, Huang CH (2017) Using big data to analyze the correlation between baseball players' performance, the matching teams, and fans attendance of chinese professional baseball league. *Physical Education Journal* 50(S):79–90, ISSN 1024-7297.
[5] Kuhn M (2018) caret: Classification and regression training. URL https://CRAN.R-project.org/package=caret.
[6] Laboratories NAR (2014) Data bank for atmospheric and hydrologic research.
[7] Regan CS (2012) The price of efficiency: Examining the effects of payroll efficiency on major league baseball attendance. *Applied Economics Letters* 19(11):1007–1015.