Predicting Ticket Price and Attendance of MLB Games Through Supervised and Unsupervised Learning Techniques

Hayrettin Uğur Çelebi*
Arjun Vivanti Jain†
Dan Jiang‡
Sai Prasad Shetty§

Fall 2018

Abstract

Professional sports teams in recent years have increasingly adopted dynamic pricing algorithms to price game tickets. In this paper we use techniques from supervised learning, unsupervised learning, and non-parametric estimation to build a predictive model using secondary market data from the 2007 MLB season to predict ticket prices and game attendance. We find that random forests exhibit optimal predictive performance with RMSE values of $18.14 for ticket price and 54.31 for attendance. This paper also proposes a seller strategy on StubHub based on opponent match-ups and game schedule.

Keywords: baseball, ticket price, attendance, secondary markets, Stubhub

Acknowledgements

We thank Professors Lorenzo Magnolfi and Christopher Sullivan for their help and feedback throughout the research process.

We also thank Mark Banghart for his invaluable expertise and time, both of which were instrumental in helping us write this paper.

*Hayrettin Uğur Çelebi: University of Wisconsin-Madison, Department of Economics, hcelebi@wisc.edu
†Arjun Vivanti Jain: University of Wisconsin-Madison, Department of Economics, avjain2@wisc.edu
‡Dan Jiang: University of Wisconsin-Madison, Department of Economics, dan.jiang@wisc.edu
§Sai Prasad Shetty: University of Wisconsin-Madison, Department of Economics, spshetty@wisc.edu

1 Introduction

The turn of the millennium has seen a massive uptick in the adoption of advanced analytics in professional sports. Major League Baseball pioneered this revolution when the Oakland Athletics used statistical analysis to revamp a dilapidated roster, laying the foundation for other teams across leagues to follow suit. In addition to putting together competitive rosters, teams in recent years have increasingly adopted complex algorithms to price game tickets taking into consideration team performance metrics, stadium section, game time, etc.

The data set used for this paper comes from Professor Andrew Sweeting's paper on dynamic pricing in secondary markets (Sweeting, 2012). The paper gives insight on how pricing of tickets mirrors theoretical models on dynamic pricing. Prof. Sweeting analyzes trends in ticket prices leading up to game days. Additionally, he looks at seller behavior in the days leading up to a baseball game and consumer search costs. Our work differs from Prof. Sweeting's paper in that we leverage machine learning techniques to build a model that predicts price and game attendance. Furthermore, while we provide a recommendation on seller strategy, we do not consider consumer costs.

We use techniques from supervised learning, unsupervised learning and non-parametric estimation to build a model with optimal predictive power. More specifically, we have incorporated cross-validation into ridge, lasso and principal component regression for our supervised and unsupervised methods, and use random forests as a non-parametric approach.

Our results show that random forests work best to predict baseball ticket prices and game attendance. While the performance of random forests is far superior to the parametric approaches, results from the latter are in line with each other and with OLS. The high performance of OLS is not unexpected given that the data set is near asymptotic in nature. In terms of seller strategy, we look at marginal profit to determine which Los Angeles Dodgers home games are best to flip tickets for on Stubhub.

This paper is organized as follows. Section 2 gives a brief background on Major League Baseball and ticket-selling on Stubhub. Section 3 describes the data while section 4 outlines our methodology. Section 5 presents our results with a comparison between root mean squared errors. Section 6 discusses seller strategy when listing tickets on Stubhub. Sections 7 and 8 conclude and credit references.

2 Background on Baseball

Major League Baseball (MLB) is one of America's four major professional sports leagues, and accounts for approximately $10 billion in yearly revenue. MLB divides its 30 teams into two leagues, American (AL) and National (NL). Each league contains three divisions: East, West, and Central; each division is comprised of five teams.

A single season of baseball is divided into the preseason, regular season, and postseason. In the preseason, which lasts from mid-February to late March, each team plays nearly 35 games. In the regular season, teams play 162 games from early April through September.

The postseason is played throughout October; AL and NL have separate brackets until the top AL team plays the top NL team in the World Series.

Each of the six divisions sends its highest ranked team based on regular season performance to the postseason. In addition, at the end of the regular season, each league has two wild card teams. Outside of the division leaders, these are the next two top teams from each league, and they play in a tiebreaker game for the final playoff spot.

A big focus of this paper is to try and predict ticket pricing based on historical ticket sales data. Official tickets to baseball games can be purchased in a variety of ways, with online marketplaces accounting for a steady increase in consumer traffic in recent years. Owned by EBay, StubHub is the MLB's official fan to fan ticket marketplace where fans can buy or resell tickets legally.

StubHub gives interested buyers the option to choose among a myriad of filters to purchase the ticket that best suits them. Consumers can choose based on price, premium suites, grouped tickets, parking lot passes, and can view the field virtually from the seat positions. Considering StubHub is a prime destination for resale tickets, transactions depend heavily on seller prices and buyer valuations, but unlike other marketplaces, has no auctions.

3 Data

The data set that has been used for this paper comes from Prof. Andrew Sweeting and his paper on dynamic pricing in secondary markets. For his paper, Prof. Sweeting obtained single game listing information for ticket sales from StubHub for the 2007 MLB season.

This information is available on the "Buy" page on StubHub. Each listing on the website displays a fixed price decided by the seller (new or resale) with additional information on seat characteristics like seat location (stadium section, row, etc.), number of seats available, and indicators for possibility of purchasing multiple seats. Apart from the posted fixed price, no other information about the seller is available on the website. StubHub retains a 25% commission, which makes up their margin on each ticket sale, excluding shipping costs.

For the data set we have used, of the 161 variables that were avaiable in Prof. Sweeting's data, we dropped 21 variables that were either identical to other variables in the data or repeated, and narrowed down our covariate set to 78 variables that were most relevant to our research question. Given that Prof. Sweeting's paper dealt with EBay data as well, it is not unreasonable to drop variables related to EBay postings.

Of the 78 variables, we dropped observations that had missing values in any column (2.5% of data sample). The missing observations were not considerably different in ticket price or attendance (our dependent variables). Therefore, dropping these listings should not have a major effect on our findings.

In econometric applications, it is common to drop independent variables that have a high correlation with the dependent variables as they may bias the estimates for coefficients of interest. In machine learning, we're interested in predictive power and therefore, don't concern ourselves with inference but we do want to address instances of high serial correlation because the correlation may not be purely coincidental - some of these variables may be

proxies for the dependent variable. This could pose a problem for prediction because a few variables out of the entire covariate set could potentially explain most of the variation in the response variable, thereby overstating the predictive power of the model. Following this reasoning, we have dropped variables that may be considered proxies for ticket price or attendance, or those that were determined ex-post, i.e., after the game had taken place.

Table 1 below provides the descriptive statistics for crucial variables in the final data set used for this paper, including both of the dependent variables (transaction price and game attendance). Table 2 provides the descriptive statistics for each team.

Table 1: Descriptive Statistics

Variable	N	Mean	St. Dev.	Min.	Max.
Number of Seats	2,147,156	2.92	1.20	1	6
Ticket Face Value ($)	2,147,156	39.98	28.52	5	312
Seller Posted Price ($)	2,147,156	73.25	68.45	0.21	767.55
Transaction Price ($)	2,147,156	86.18	80.52	0	903
Game Attendance (1000s)	2,147,156	40.94	9.42	8.20	56.44

At first glance, it may seem unusual that the average transaction price is higher than the average posted price, but it must be remembered that transaction prices include handling fees and shipping costs, both of which StubHub is responsible for. Additionally, the closeness of estimates for the mean prices and standard deviations indicate that, even though the average ticket is priced reasonably, prices vary over a wide range. This is also evident from the figures for minimum and maximum price.

There are several things to note in Table 2 below. First, the Colorado Rockies are not listed. This is because they were the only team in 2007 to practice a form of dynamic pricing (Sweeting, 2012).

There are an extremely high number of observations for the big market teams, like the New York Yankees, Chicago Cubs, and San Francisco Giants. This makes sense because the most popular teams generate highest revenue in ticket sales, which translates to the highest number of tickets changing hands.

The highest average attendance is for the New York Yankees and Los Angeles Dodgers. This is understandable because these teams are not only two of the most popular in Major League Baseball, they have the two of the largest stadiums.

A striking detail in average price is that it is far higher for the Boston Red Sox than for any other team. A reasonable explanation behind this is that Fenway Park, the stadium of the Boston Red Sox, is much smaller than stadiums of the other popular teams. Because the Red Sox are one of MLB's most storied franchises, there is high demand for a much smaller number of available tickets, driving prices far higher than for other big name teams.

Table 2: Descriptive Statistics by Team

Teams	N	Average Attendance (in thousands)	Average Price
Arizona Diamondbacks	36,201	31.274	75.79
Atlanta Braves	32,594	37.037	72.20
Baltimore Orioles	44,813	32.918	93.30
Boston Red Sox	101,723	36.684	165.36
Chicago White Sox	155,904	33.522	78.31
Chicago Cubs	221,470	40.400	93.65
Cincinnati Reds	16,785	28.494	63.07
Cleveland Indians	24,871	32.172	68.55
Detroit Tigers	110,013	38.832	70.55
Miami Marlins	5,484	19.687	71.43
Houston Astros	40,228	38.082	91.24
Kansas City Royals	7,599	22.447	49.95
Los Angeles Angels	107,574	41.827	73.67
Los Angeles Dodgers	102,996	47.875	74.25
Milwaukee Brewers	18,771	37.789	59.58
Minnesota Twins	8,514	28.928	64.05
New York Mets	145,294	48.187	77.42
New York Yankees	380,215	53.144	88.78
Oakland Athletics	34,790	26.100	74.60
Philadelphia Phillies	65,074	39.428	80.65
Pittsburgh Pirates	4,869	24.384	58.98
San Diego Padres	30,574	36.120	86.34
San Francisco Giants	169,908	39.737	85.35
Seattle Mariners	28,414	36.368	104.9
St. Louis Cardinals	116,093	43.840	65.12
Tampa Bay Rays	16,910	21.766	93.71
Texas Rangers	41,785	31.511	98.41
Toronto Blue Jays	9,504	32.354	100.89
Washington Nationals	68,186	24.553	92.59

Figure 1 below shows that ticket price and attendance vary similarly with the day of the week the game was played. Games that occur on weekends (Friday, Saturday, Sunday) have the highest ticket prices and attendance; games that occur on weekdays have lower ticket prices and attendance. This is to be expected because the ratio of leisure to labor is generally much higher over the weekend when compared with weekdays.

Figure 2 below illustrates that game attendance increases throughout the summer months, but decreases in September (the final month of the MLB regular season, but the first month after the end of summer). It seems there is no discernible relationship between the month of the game and ticket prices.

Figure 1: Average ticket price (left) and attendance (right) by day of week game was played

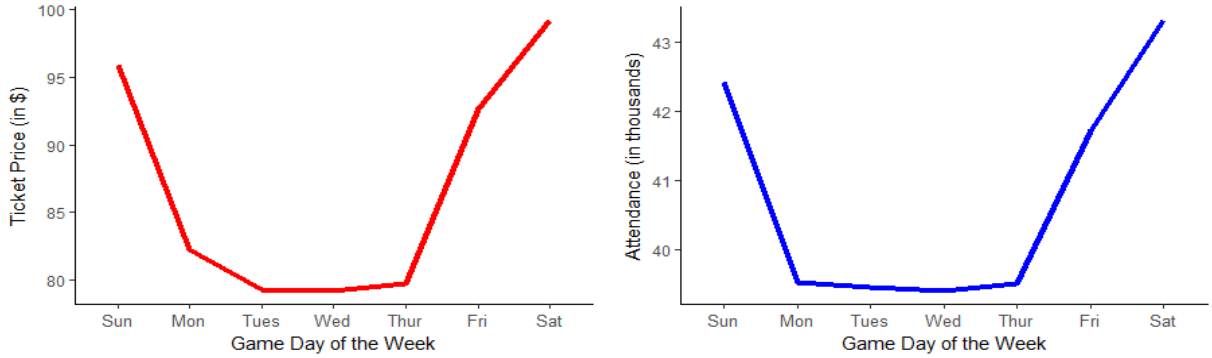


Figure 2: Average ticket price (left) and attendance (right) by month game was played

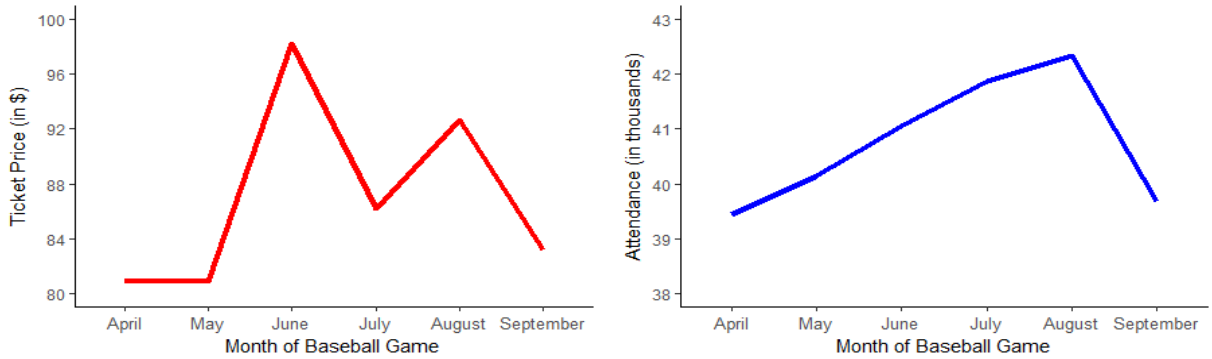


Figure 3: Average ticket price (left) and attendance (right) by win percentage of home team

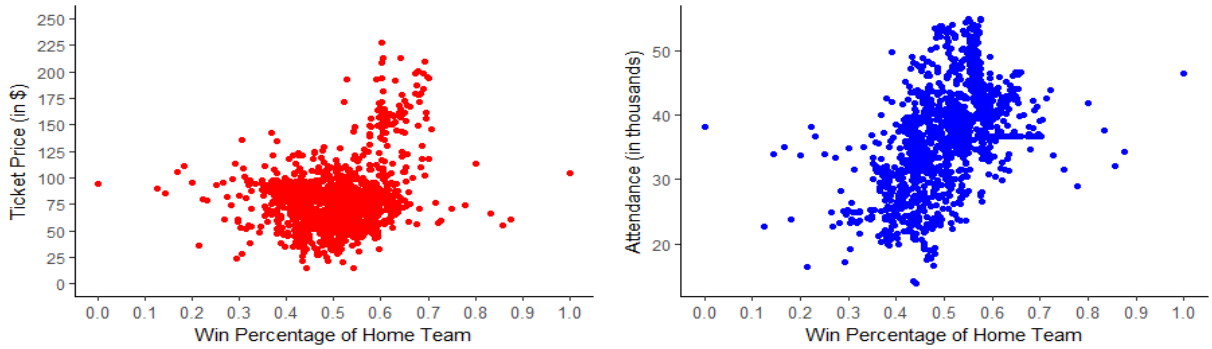


Figure 4: Average ticket price by seat row order

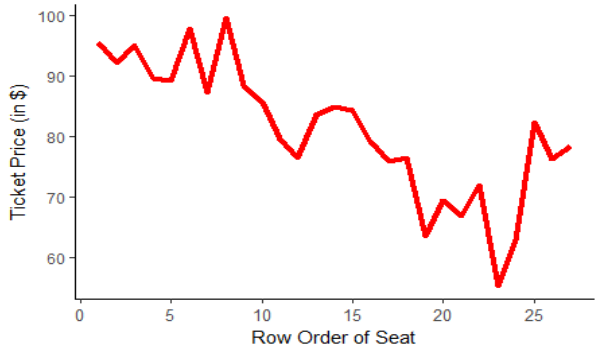


Figure 3 above depicts the relationship between ticket price, attendance, and the win percentage of the home team. One important point to note is that the data is highly concentrated around win percentages of 0.5. This is because by the end of the season, the best teams usually have a 0.66 win percentage and the worst teams usually have a 0.33 win percentage. Very high and very low win percentages only occur at the beginning of the regular season. There is a general positive relationship between home team win percentage and ticket price, which makes sense because there is higher demand for better teams. There is similarly a positive relationship between home team win percentage and attendance for the same reason.

It is not unreasonable to think that away team record also determines ticket price and attendance. However, a majority of fans attending a baseball game will be supporting the home team. Although not presented, the data show little to no relationship between away team win percentage, ticket prices, and attendance.

Figure 4 exhibits a downward trend in how average prices change with row number. A point to note here is that in this data set, seats in rows 27 and above have been clubbed together under one group. The downward trend in prices is not unexpected because one would expect seats closer to the field to be priced more competitively. The spike we see in rows above 25 could be attributed to the location of the more premium box locations in the upper sections of the stadium.

4 Methodology

The goal of this paper is to come up with a predictive model that dictates ticket pricing and game attendance based on ticket listing characteristics obtained from StubHub. Given that the literature in this field has numerous predictive techniques, we have attempted to include techniques from supervised learning, unsupervised learning and non-parametric methods to come up with a model that gives us optimal predictive power.

The methods that have been used in this paper are ridge, lasso, regularized ridge and lasso, principal component analysis (PCA), and random forests. These techniques are in no way exhaustive but allow us to test different machine learning techniques without placing restrictions on functional form assumptions.

4.1 Shrinkage Methods

Subset selection techniques help choose the set of variables that best describe the dependent variable. However, this process can get cumbersome when working with a large set of covariates. Since our goal is optimizing predictive power and not the interpretation of coefficients, an alternative to subset selection is a technique that utilizes all covariates available, but *regularizes* or shrinks coefficient estimates towards zero. The mmost well known techniques to achieve shrinkage are ridge and lasso regressions.

We know from least squares estimation (OLS) that the residual sum of squared errors (RSS) is defined as follows:

$$RSS = \sum_{n=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_i x_{ij} \Big)^2 \tag{1}$$

In equation (1) y_i is the dependent variable and x_i the set of covariates. The ridge and lasso regressions work differently in that they add a *tuning parameter* that determines the degree to which coefficients are reduced.

$$RSS = \sum_{n=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \Big)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2}$$

$$RSS = \sum_{n=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \Big)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3}$$

(2) represents the regularization equation for ridge regression, where the square of coefficient estimates are penalized and therefore, is also called L2 regularization because it penalizes the L2 norm of coefficients. (3) gives the equation for lasso that penalizes the L1 norm of coefficient estimates. Using a value of zero for lambda results in regular OLS.

Considering the StubHub data set has observations for the dependent variables in our analysis (ticket price and attendance), ridge and lasso regressions help us deal with this supervised learning problem. Additionally, as mentioned above, these methods let us work with the entire set of covariates rather than a subset of variables.

Another technique we have used to complement the shrinkage methods described above is *k-fold cross validation*. Through cross validation, the training sample we work with is split up into groups or "k-folds" (in our analysis we use 10 folds). The method is then run with all but one fold and an MSE computed using the fold left out. This procedure is repeated k times and the cross validation estimate is calculate as the mean of the MSE obtained from each fold:

$$CV_k = \frac{1}{k} \sum_{i=1}^{k} MSE_i \tag{4}$$

An important element of ridge and lasso regressions is the selection of the tuning parameter. One can run the regressions to see how MSE estimates change with the different values of λ. However, since our focus is to come up with a model that most closely predicts ticket prices and attendance, we use cross validation to obtain the value of λ that gives the least MSE for ridge and lasso.

4.2 Principle Component Analysis

In this section we use principal component analysis to reduce the dimension of our model. For this purpose, instead of regressing the dependent variable (ticket price or game attendance) on the explanatory variables directly, we use the principal components of the explanatory variables as regressors. This method is widely known as principal component regression

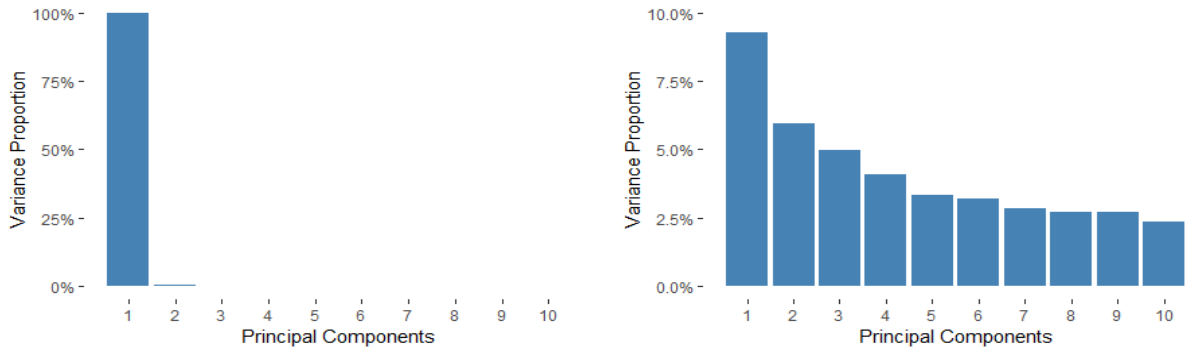
(PCR). We utilize PCR methodology to exclude some of the low-variance principal components in the regression step. Moreover, regressing on only a subset of all the principal components, we try to reduce the dimension of our model through lowering the effective number of parameters. This is particularly useful for our high-dimensional data set, such as Professor Andrew Sweeting's Stubhub data.

There are three main steps in PCR. In the first step, we separate our dependent variables (ticket price and game attendance) and performed PCA on the training set of our covariates. After determining the principle components, we use ten-fold cross validation to determine the optimal number of components to include in our final model. In the final step, we use the resulting model coefficients in the test sample to predict ticket price and game attendance, and calculate a mean squared error for the test set. The 'pls' R package was very helpful in performing these three steps properly.

Because principal component analysis can be considered an exercise in maximizing variance, it is crucial to scale our variables. By performing PCA, we are projecting the original data onto directions that maximize the variance. When we have large differences between the scales of variables, the PCA procedure will put more weight on variables with high variance. This problem can be seen in the plot on the left in Figure 5 below, which shows the proportion of total variance explained by different principal components when the data is not normalized. As we can see from the graph only the first component seems like to explain all the variance in our data set.

However, when the data is normalized (subtract from each variable its mean and divide by its standard deviation), it is clear from the graph on the right below that other components make bigger contributions. In essence, scaling the data allows for inclusion of more variation in other principal components.

Figure 5: Proportion of variance explained by raw PCs (left) and scaled PCs (right)



To determine the optimal number of principal components, we perform a ten-fold cross validation process. The 'pls' package in R contains two primary strategies for choosing optimal number of components. In the first strategy we choose the model with fewest components that are less than one standard deviation away from the overall best model. The second strategy employs a permutation approach, and essentially tests whether adding a new component is beneficial or not. For the sake of shorter computation time we apply only the first strategy.

4.3 Random Forests

The non-parametric method of random forests is a popular ensembling technique used to improve the predictive performance of decision trees. This is accomplished by averaging over trees to reduce variance.

Decision trees are considered a very simple modelling technique with easy interpretation. However, a major drawback is that they have poor predictive performance and inadequate generalization to the test set. For this reason, we use random forests instead, which involves producing multiple trees that yield a single consensus prediction.

The main idea of random forests is to improve variance by reducing correlation between decision trees. This is achieved in the tree-growing process by randomly selecting the input variables and averaging over multiple bootstrapped training samples.

The following give the basic steps involved in performing the random forest algorithm:

1. Pick N random observations from the training set to form a bootstrap sample.
2. Build a decision tree based on the bootstrap sample by randomly selecting a subset of variables and then picking the best variable and split-point among them.
3. Choose a number of trees to use and repeat steps 1 and 2.

In the case of a regression problem, for a new observation, each tree in the forest predicts a value for the outcome. The final value can be calculated by taking the average of the outcomes predicted by all the trees in the forest.

If there is a highly non-linear and complex relationship between the features and the response, then random forests may outperform standard linear regression approaches. If, however, the relationship between the features and the response is quite linear, then a linear regression will likely work well and may outperform a method such as random forests that does not exploit the linear structure of the data. The relative performance of random forests and classical linear regression can be assessed by estimating the mean squared error of the test set. This will be discussed further in section 5.

Random forests have some considerable advantages over other models we consider. For example, random forests using trees closely mirror human decision-making with regard to choosing between different options. In addition, trees can easily handle qualitative predictors without the need to create dummy variables. Despite these advantages, some of the more noticeable downsides of random forests include more complicated implementation and higher computational cost relative to linear regression.

5 Results

As mentioned in the previous section, we have used models from supervised learning, unsupervised learning and non-parametric techniques to predict ticket prices and game attendance based on data from the 2007 MLB season. In this section we report the *root mean squared errors* (RMSE) instead of the mean squared errors (MSE) because the RMSE gives us a measure that is on the same scale as the dependent variables while the MSE is essentially a variance measure. Table 3 contains the results of the root mean squared values obtained

from models we have discussed in the previous section. All models have been trained using 80% of the data set, with the remaining 20% serving as the test set.

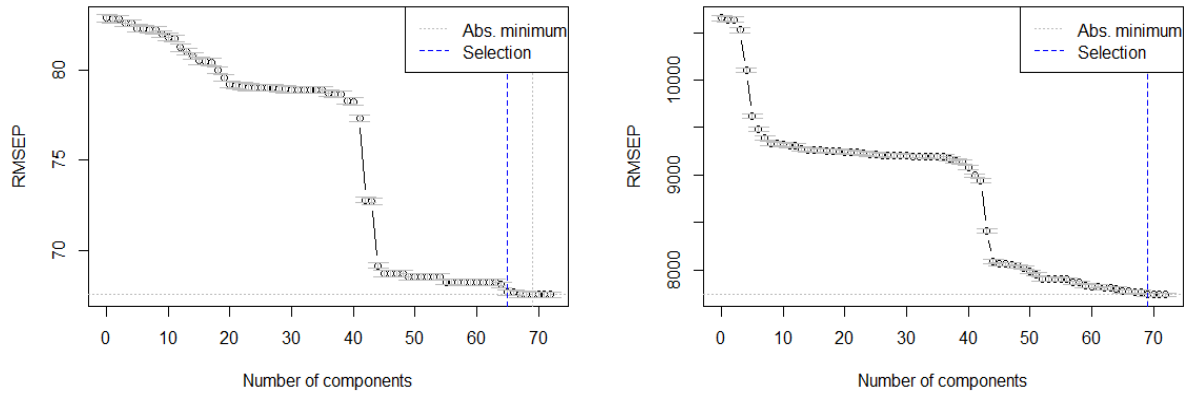
Table 3: Root Mean Squared Values for Predictive Models

Model	Ticket Price ($)	Game Attendance
Ordinary Least Squares	65.04	7,382.94
Ridge & 10 Fold Cross Validation	65.52	7,497.97
	(4.61)	(527.54)
Lasso & 10 Fold Cross Validation	65.05	7,385.32
	(0.004)	(1.22)
Principal Component Regression	69.08	7,739.09
Random Forest	18.14	54.31

Note: the values in parentheses are the tuning parameters that give minimum RMSE.

We have used OLS as a baseline model here to comare results from the other methods. It may be interesting to note that OLS and ridge/lasso perform comparably on the test data set, with an RMSE value of approximately $65 for ticket prices and ∼7,400 for game attendance. As a data set becomes asymptotic, OLS does a good job with prediction. Even though we have used a 80%-20% split for training (∼1.6 million observations) and test (∼500k observations), OLS produces comparable RMSE estimates for both. With cross-validation it is important to note that the tuning parameter for ridge and lasso is selected on the basis of the performance of the regression with the *training sample*. Unlike OLS, when dealing with such large sample sizes, there is no saying if the RMSE estimates for cross-validation will be comparable for the training and test sets.

Figure 6: PCR - RMSE Values for Ticket Price (left) and Attendance (right)



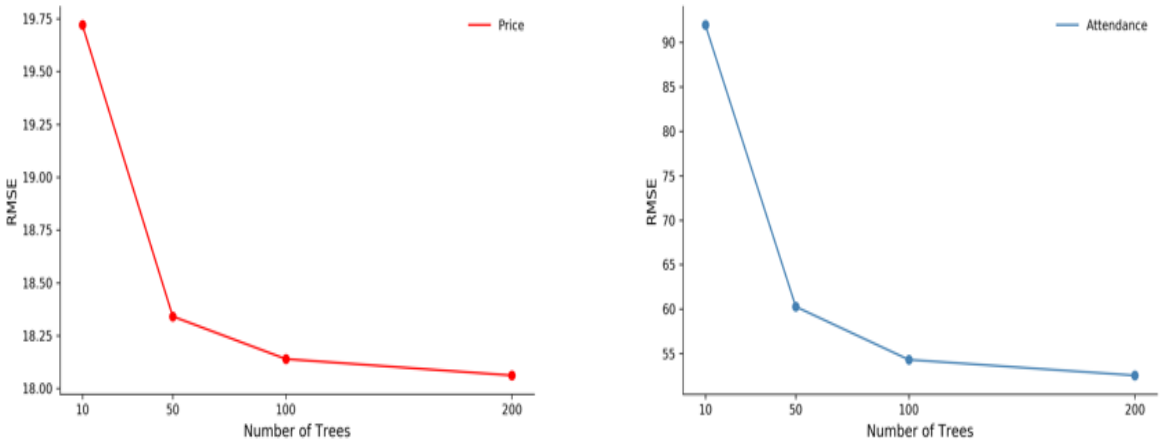
Principal Component Regression represents the unsupervised learning technique used in this paper as an alternative to supervised learning procedures. Figure 6 above shows that

root mean square values corresponding to different principal components that we include in our model. The grey dotted line in the graph gives the number of principal components that correspond to the global minimum of RMSE for the respective dependent variables. The blue line points to the selection of the one-sigma method that we mentioned in our methodology section. According to these results we selected 65 principal components to predict game ticket price and 69 principal components to predict attendance. For both ticket price and attendance, PCR results in the highest RMSE. However, considering PCR is an unsupervised learning technique that is used for dimension reduction, the selection of a smaller number of principal components instead of the entire set of covariates does not compromise the test sample errors in a significant way. From the table 3, it can be seen that the PCR results are mostly in line with the results from parametric methods.

To apply random forest, we simply construct B regression trees using B bootstrapped training sets and average the resulting predictions. A random sample of m predictors is chosen as split candidates from the full set of p predictors. These trees are grown deep, and are not pruned. Hence each individual tree has high variance but low bias. Averaging these B trees reduces the variance. Typically we choose $m = \frac{1}{3}p$; in essence the number of predictors considered at each split is approximately equal to a third of the total predictors.

In Figure 7 below the root mean squared error is shown as a function of number of trees, the number of bootstrapped training sets used. Random forests were applied with 25 features, which is approximately one-third of all 77 features. We can see that the error values decrease with the increase in number of trees. Beyond 100, as the number of trees grows, the rate of decrease in error diminishes, so we use 100 trees for our model.

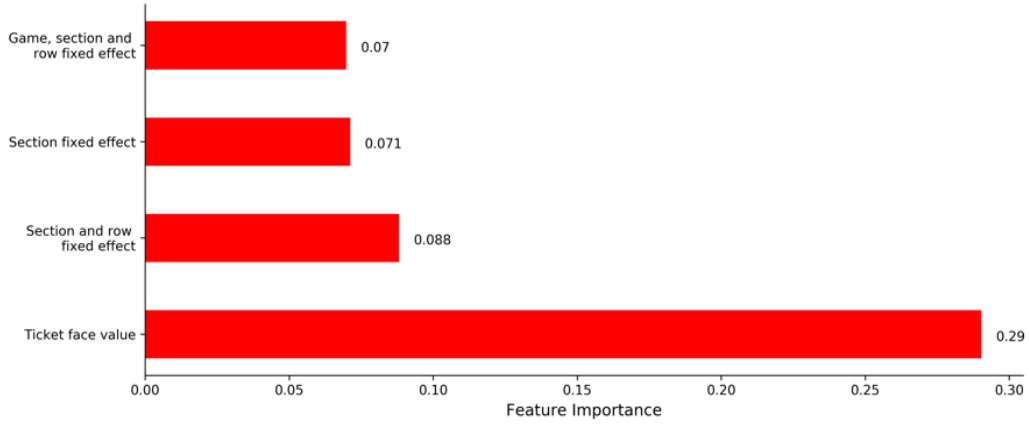
Figure 7: RMSE by Number of Trees for Ticket Price (left) and Attendance (right)



We can also obtain an overall summary of the importance of each feature with respect to the predictability of the response variable. The feature importance reported here is a measure of the mean decrease impurity, which results from splits over that feature averaged over all trees. A large value indicates an important feature. Take price prediction as an example, from figure 8 we can see the relative importance of different covariates in price

prediction. The top 4 important features are 'Ticket face value', 'Section and row fixed effect', 'Section fixed effect', 'Game, section, and row fixed effect.'

Figure 8: Covariate Importance in Price Prediction



It's quite intuitive that the face value of the ticket will play an important part in determining final transaction price, which contributes to 29% mean decrease impurity among all the variables. The other features are identifiers for games and seat characteristics. Team match ups are crucial for price, and understandably so. For example, top teams or big market teams generate a good amount of audience interest, driving ticket prices up. Also, seating sections and rows can dictate the quality of game watching experience and will inevitably be a major predictor for price.

6 Seller Strategy for Los Angeles Dodgers

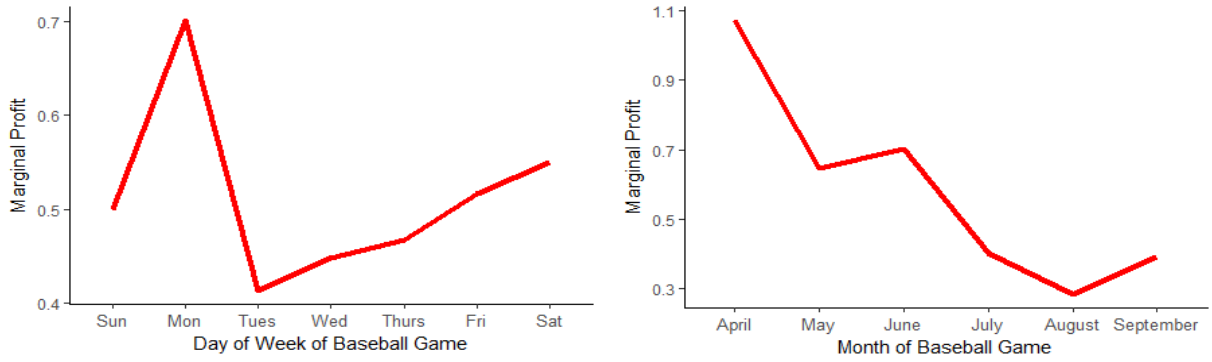
In this section, we provide a recommendation on seller strategy for Los Angeles Dodgers tickets. This recommendation includes the type of games a seller on Stubhub should resell tickets for, including the day of the week and month of the year the game is played, how well the Dodgers are playing, and who the away team is.

Figure 1 from Section 3 shows the average ticket prices for games occurring on each day of the week; Figure 2 shows the average ticket prices for games occurring in each month of the regular season. While these graphs show that on average, the highest selling tickets are for games on weekends and during the summer months of June and August, these graphs are misleading for the purposes of seller strategy.

This is because sellers are less concerned with how much a ticket sells for, they are primarily focused on profit. For example, compare ticket A with face value \$80 that sells for \$85 and ticket B with face value \$40 that sells for \$50. Ticket A clearly sells for more than ticket B, but a seller prefers to flip ticket B for a \$10 profit, compared to a \$5 profit from ticket A. Furthermore, consider ticket C with face value \$100 that sells for \$110. While the profit is the same for tickets A and C, a much larger fraction of ticket face value is made

back in ticket A than in ticket C. For these reasons, the graphs below consider marginal profit instead of transaction price. The marginal profit is defined as the difference between the transaction price and the ticket face value, divided by the ticket face value.

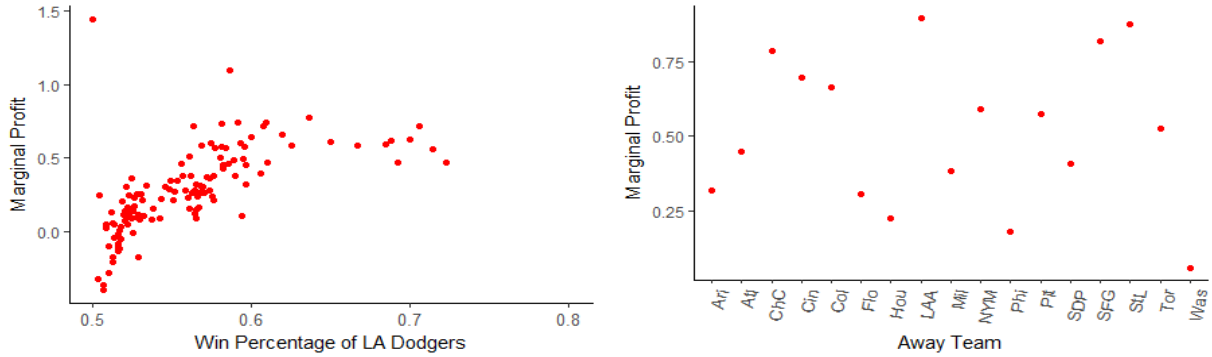
Figure 9: Marginal Profit by Day of Week (left) and Month (right)



The graph to the left above indicates that marginal profit is much higher on Mondays than any other day (note that average ticket price is highest in Fridays, Saturdays, and Sundays). This is in part due to much lower ticket face value for games occurring on Mondays, so even though tickets don't sell for very much, the marginal profit is much larger than that of tickets for weekend games.

The graph to the right above indicates that marginal profit is much higher in April than any other month of the regular season. Figure 2 shows that average ticket price is lowest in April, which means marginal profit seems to be highest in April for the same reason it is highest on Mondays. This also makes sense, because April marks the very beginning of the regular season; teams will not sell tickets for high prices until it is known there will be substantial demand for them.

Figure 10: Marginal Profit by LAD Win Percentage (left) and Away Team (right)



The graph to the left above shows that marginal profit increases as the win percentage of the Los Angeles Dodgers increases. From Figure 3 in Section 3, average ticket price also seems to generally increase as win percentage increases. As the Dodgers' win percentage increases, we expect the face value of tickets to increase by much less than the seller price

on Stubhub. This points to the trend in the graph: an increase in marginal profit despite an increase in average ticket price.

The graph to the right above shows that marginal profit is highest when the Los Angeles Angels are the away team. While this is very interesting, there is much to speculate as to why this is the case. Perhaps because the Angels play in the same city as the Dodgers, more fans are able to support the away team at a Dodgers home game. Their valuation for a ticket is higher than the average home team fan because they are devoted enough to support their team even in an away game (and heckle the home team). This may contribute to a higher marginal price when selling to fans of the Los Angeles Angels, and give sellers incentives to sell to fans of the away team.

Thus we conclude from this analysis that sellers of Los Angeles Dodgers tickets should flip tickets on Stubhub for games in which the marginal profit will be the highest. Sellers might expect the optimal games for which to flip tickets to be on Mondays and in April (the beginning of the regular season). For games after April, sellers would fair best flipping tickets for games in which the win percentage of the Dodgers is high. Lastly, sellers should target games in which the Los Angeles Angels are the away team.

7 Conclusion

In this paper we have formulated a model that predicts ticket prices and game attendance for baseball games based on StubHub listing data from the 2007 Major League Baseball season. We incorporate supervised and unsupervised learning techniques along with non-parametric analysis. More specifically, we use cross-validation in ridge, lasso, and principle component regression, and look at random forests independently.

We find that Random Forests provide optimal predictive power as evidenced by an RMSE of just $18.13 for ticket price and 48.14 for game attendance. These are exceptional results considering the average ticket price for the 2007 season was $86.18 and average attendance at 40,940. We can also conclude that Principal Component Analysis serves as a reliable alternative to supervised learning techniques as it's dimension reduction capabilities does not drastically affect the ability to predict the relevant response variable.

In terms of seller strategy for Los Angeles Dodgers tickets, we examine marginal profit, because sellers are more concerned with profit than actual transaction price. This is found by looking at the difference between transaction price and ticket face value, and then dividing by the ticket face value. We find that sellers might expect the optimal games for which to flip tickets to be on Mondays and in April. For games after April, sellers would fair best flipping tickets for games in which the win percentage of the Dodgers is high, as seen in the positive relationship between win percentage and marginal profit. Lastly, sellers should target games in which the Los Angeles Angels are the away team, as the marginal profit is highest in these games.

There is much potential for future research in this area. The models we have described can be extended to include covariate interactions. Creating interactions for such a large covariate set could be computationally intensive, but could also help explain more of the

variation in the response variables. Additionally, these models can also be extended to other professional sports or even to fantasy sports where prediction of game performance metrics could prove to be particularly profitable. Aside from extending our methodology to other sports, it would be interesting to compare the different models on a basis other than the root mean squared error.

For seller strategy, we have concentrated on sales decisions based on opponent matchups and game schedule. The StubHub data set has identifiers for tickets sold by the number of days before a game. However, it would be interesting to see how prices change based on when a seller posts tickets on StubHub as opposed to when the sale is made. Availability of such data could help determine optimal time frame before games to post tickets on StubHub.

8 References

Hastie, T., Tibshirani, R., & Frieman, J. (2001). The Elements of Statistical Learning. *Springer Series in Statistics* New York, NY, USA.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. *Springer Texts in Statistics* New York, NY, USA.

Sweeting, A. (2012). Dynamic Pricing Behavior in Perishable Goods Markets: Evidence from Secondary Markets for Major League Baseball Tickets. *Journal of Political Economy*, 120(6), 1133-1172.