DECISION SCIENCES INSTITUTE

Predicting Attendance at Major League Soccer Matches: A Comparison of Two Techniques

ABSTRACT

Major League Baseball, the National Basketball Association, and the National Football League all have a large history of attendance, many potential predictor variables, and no shortage of analysts to analyze these data. Major League Soccer (MLS), on the other hand, is relatively new, has only a modest amount of data, and most teams and the league itself do not have analysts to build prediction models. We will use attendance data, box office data, venue city demographics, and weather data among other data to build two regression models to predict attendance at MLS matches. One model is a traditional ordinary least squares model and the other is a data mining model using random forests. The data mining model outperforms the least squares model on the main measure of mean squared error. It also outperforms the least squares model on the easier to understand mean absolute percent error. However, a significance test implies that the null hypothesis of equal forecast accuracy between the two techniques cannot be rejected.

KEYWORDS: Forecasting, R, Regression analysis, SPSS, Supervised learning

INTRODUCTION

Predicting attendance at sporting events has received much attention in the literature. For examples, see Brande and Tichen (1996), Goleman and Taylor (2012), and Hogan *et al*. (2012). It is important for teams to develop an accurate forecast of future attendance to plan staffing, inventory, and promotions. Most approaches use traditional statistical techniques such as ordinary least squares to produce the attendance predictions. We propose using a more contemporary approach, a data mining approach using random forests to produce the predictions. (See the Appendix for a simple explanation of random forest.) We then compare the performance of the two techniques with respect to error measures and conclude that overall the data mining approach outperforms ordinary least squares.

RESEARCH QUESTION

Given the myriad of data available to sports teams, in particular, MLS teams, is it possible to develop a useful prediction of attendance? Furthermore, when using a traditional ordinary least squares approach or a random forest data mining approach, will one emerge as clearly superior to the other?

DATA

We acquired 574 observations of MLS matches for the 2014 and 2015 seasons. The raw data consisted of 62 box office variables including the number of full tickets sold and the attendance at the match. The score of the match was not included in the data.

Initial Investigation of the Data

Most of the box office data were thought not to be useful for predicting attendance and were discarded. Additional data were acquired such as a surrogate for a team's popularity and weather data at the time of the match.

The SPSS package was used to develop the ordinary least squares model. The statistical language R was used in all other analyses. R does not have a feature to attach variable labels to variables. See Table 1 for labels for the variables used in this study.

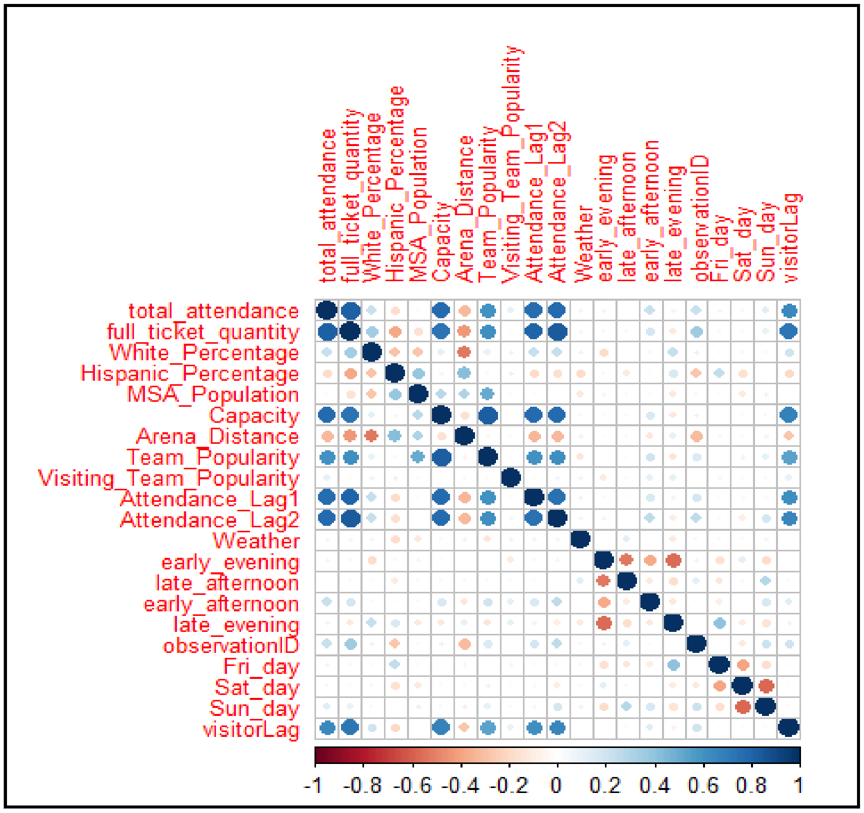
Table 1: Labels for R Variables

R Variable	Label	Comment
Arena_Distance	Arena distance from downtown	
Attendance_Lag1	Total attendance lagged one home match	
Attendance_Lag2	Total attendance lagged two home matches	
Capacity	Venue seating capacity	
early_afternoon	Early afternoon match	11:00 am through 2:30 pm start time
early_evening	Early evening match	5:00 pm through 7:30 pm start time
Fri_day	Friday match	Binary indicator variable
full_ticket_quantity_log	Log of full season ticket sales	
Hispanic_Percentage	Hispanic proportion of MSA population	Home venue MSA Hispanic proportion
late_afternoon	Late afternoon match	3:00 pm through 4:30 pm start time
late_evening	Late evening match	8:00 pm through 11:00 pm start time
MSA_Population	Metropolitan Statistical Area population	MSA population of home venue
Other_day	Weekday match	Baseline indicator variable
Sat_day	Saturday match	Binary indicator variable
Sun_day	Sunday match	Binary indicator variable
Team_Popularity	Home team popularity	Relative value based on number of Google searches
total_attendance	Attendance	Target variable
Visiting_Team_Popularity	Visiting team popularity	Relative value based on number of Google searches
visitorLag	Visitor lagged attendance	Attendance at last home match with this visitor
Weather	Weather	Binary indicator variable; 0 okay, 1

		inclement
White_Percentage	White proportion of MSA population	Home venue MSA white proportion

An initial look at correlations can be seen in Figure 1.

Figure 1: Correlation Plot of the Variables Used in the Study.
Larger circles indicate greater correlation. Blue is positive correlation and red is negative correlation. The legend appears at the bottom of the figure.

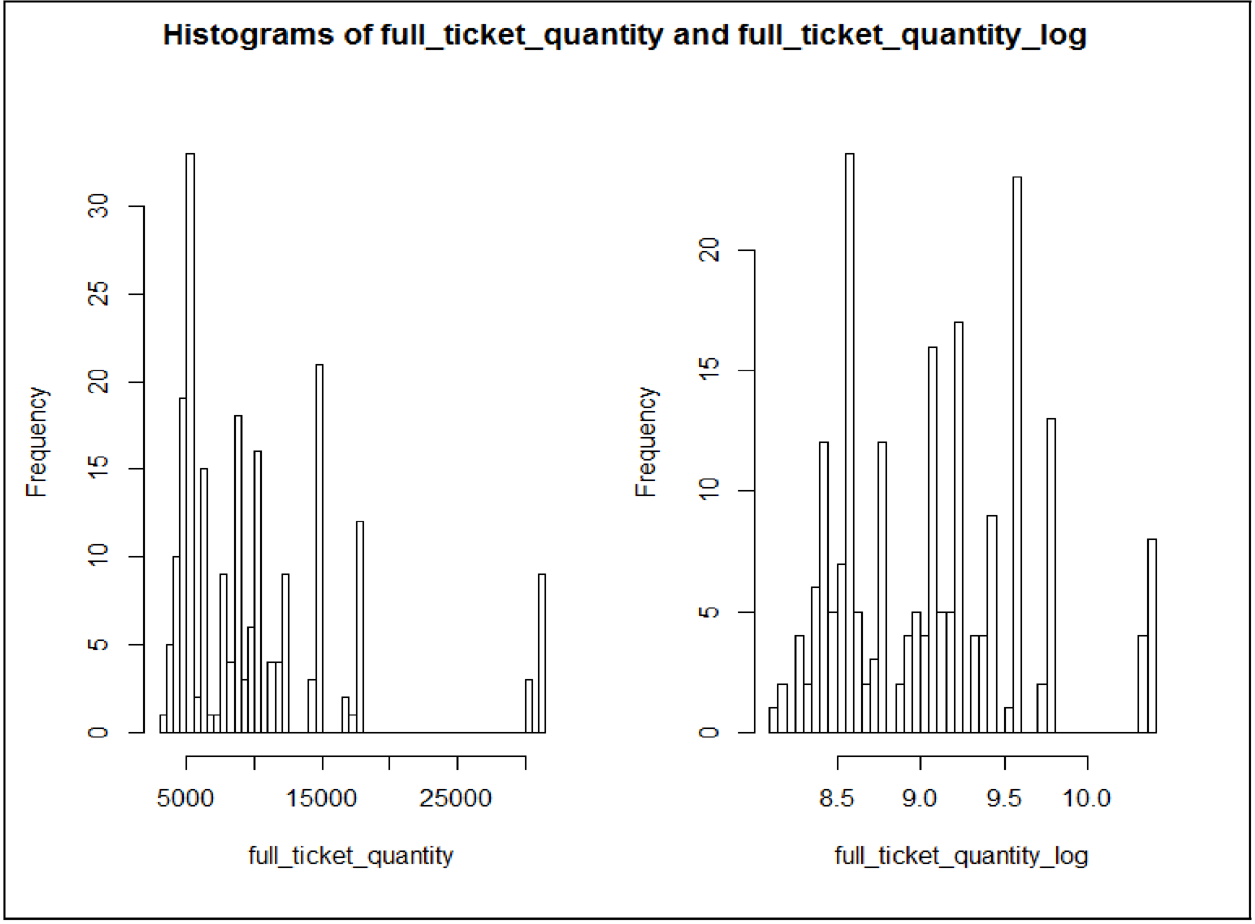


An examination of the first row of Figure 1 reveals those independent variables that are highly correlated with the target variable, total_attendance.

Transform the Full Ticket Quantity Variable

The distribution of the variable full_ticket_quantity skews to the right as can be seen in the left panel of Figure 2. A natural log transform of that variable can be seen in the right panel of Figure 2. The right panel is not as wide and not as skewed as the left panel indicating that the transformed variable should be used instead of the untransformed one.

Figure 2: Distribution of full_ticket_quantity and distribution of ln(full_ticket_quantity)



ORDINARY LEAST SQUARES MODEL

Ordinary least squares can suffer from multicollinearity. See Table 2 for the variance inflation factors (VIF) resulting when total_attendance is modelled as a function of the independent variables in Table 2. The VIF quantifies the severity of multicollinearity in an ordinary least squares regression analysis. It provides an index that measures how much the variance of an estimated regression coefficient is increased because of collinearity.

Table 2: Variance Inflation Factor

Variable	VIF
full_ticket_quantity	8.727
Capacity	7.561
Team_Popularity	6.146
Attendance_Lag2	4.675
Attendance_Lag1	3.726
MSA_Population	2.785
early_evening	2.762
Sun_day	2.592
Sat_day	2.571

visitorLag	2.409
late_afternoon	2.352
Fri_day	1.988
Arena_Distance	1.971
early_afternoon	1.956
White_Percentage	1.830
Hispanic_Percentage	1.768
Weather	1.224
Visiting_Team_Popularity	1.120

The Linear Model

Based on the correlations shown in Figure 1 and adjusting for the multicollinearity indicated in Table 2, the ordinary least squares prediction model for total_attendance is shown in Table 3.

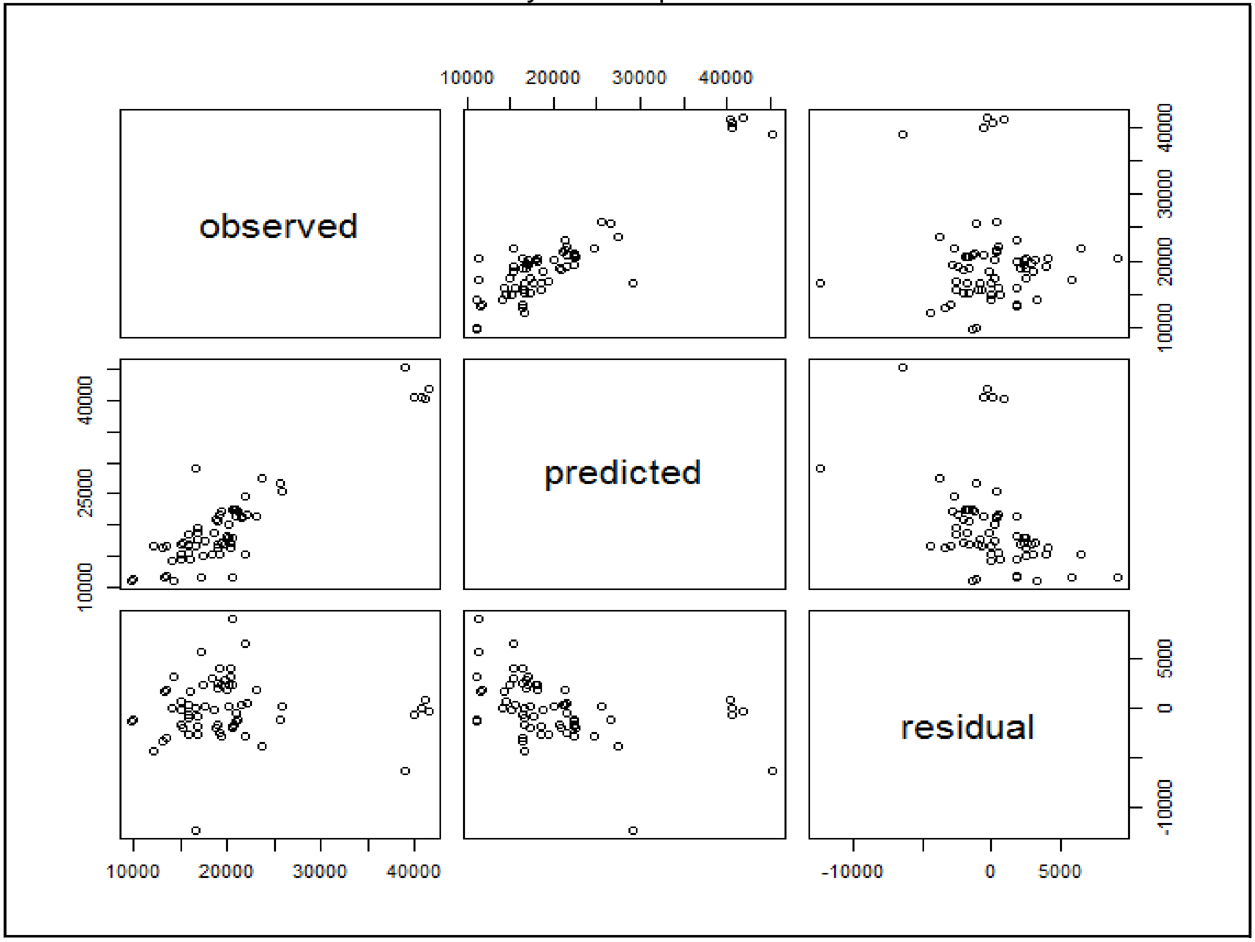
Table 3: Ordinary Least Squares Model for Total Attendance

Variable	B
(Constant)	-42230.0000
full_ticket_quantity_log	4163.0000
Capacity	0.7981
Team_Popularity	-35.7700
Attendance_Lag1	0.1854
Attendance_Lag2	0.1894
visitorLag	0.0234

Residual Analysis for the Ordinary Least Squares Model

Figure 3 shows correlation plots for observed, predicted, and residual values.

Figure 3: Correlation Plots of Observed, Predicted, and Residual for Ordinary Least Squares Model



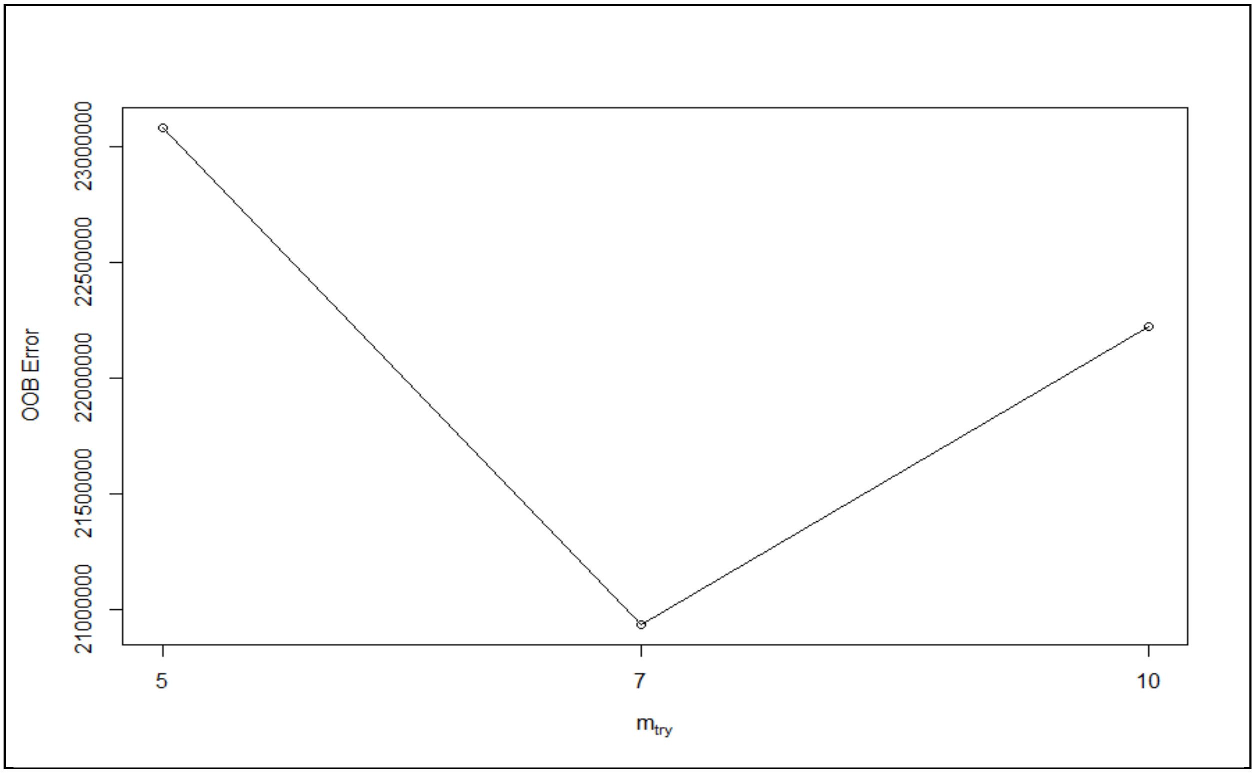
DATA MINING WITH RANDOM FOREST MODEL

Random forest is a bagging technique that reduces the variance of a statistical learning method. The test error is estimated with the out-of-bag (OOB) observations eliminating the need to perform cross validation (James *et al.* (2013)). With random forest, it is not necessary to pre-select the variables that should be in the model (JEquiha (n.d.)) and multicollinearity does not present a problem with this technique (Welling (n.d.)). See Breiman and Cutler (n.d.) for a thorough discussion of random forests.

Tuning the Random Forest Model

Not all the predictor variables are used when producing a tree in the forest. The model can be tuned to minimize the OOB error by selecting an appropriate number of predictor variables to use, mtry. Figure 4 shows that the best number of predictors to use is seven.

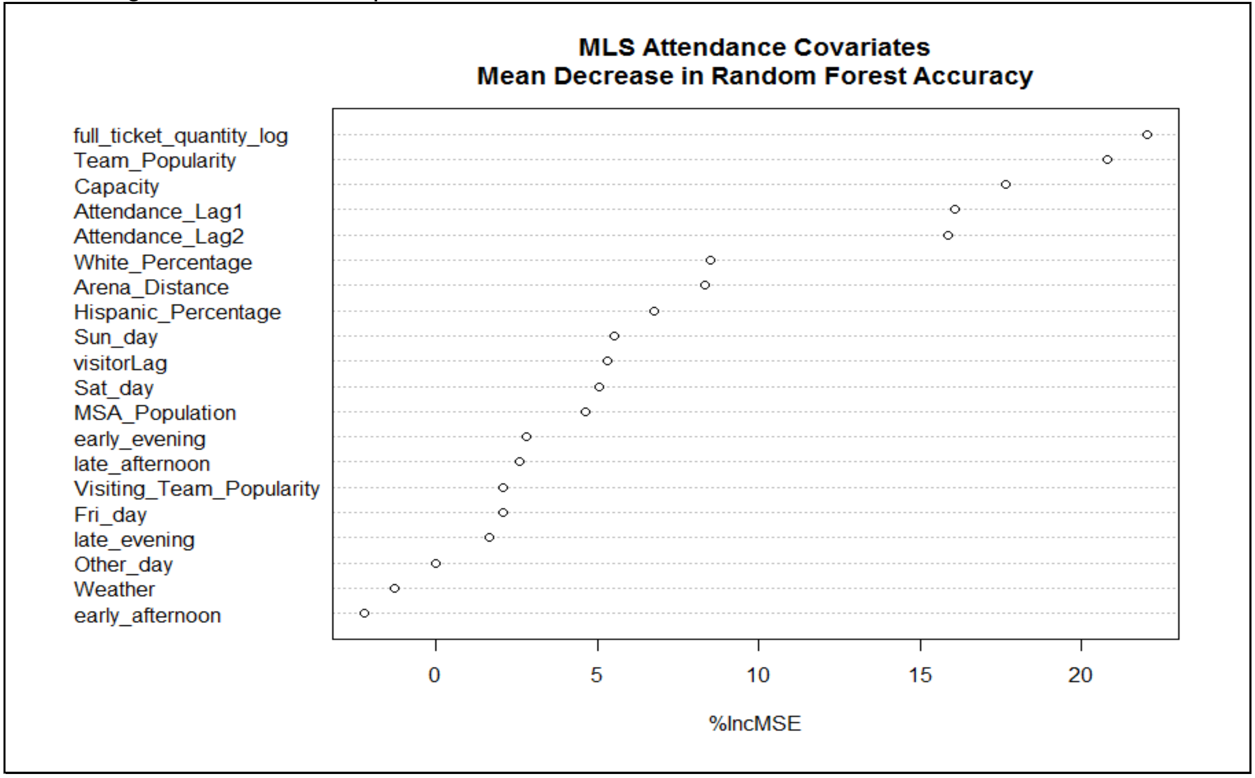
Figure 4: Lowest Out-of-Bag Error Occurs When Using Seven Predictor Variables in the Random Forest Model



Variable Importance

Figure 5 presents the importance of the variables in affecting mean squared error (MSE). Variables with high %IncMSE are more important than others.

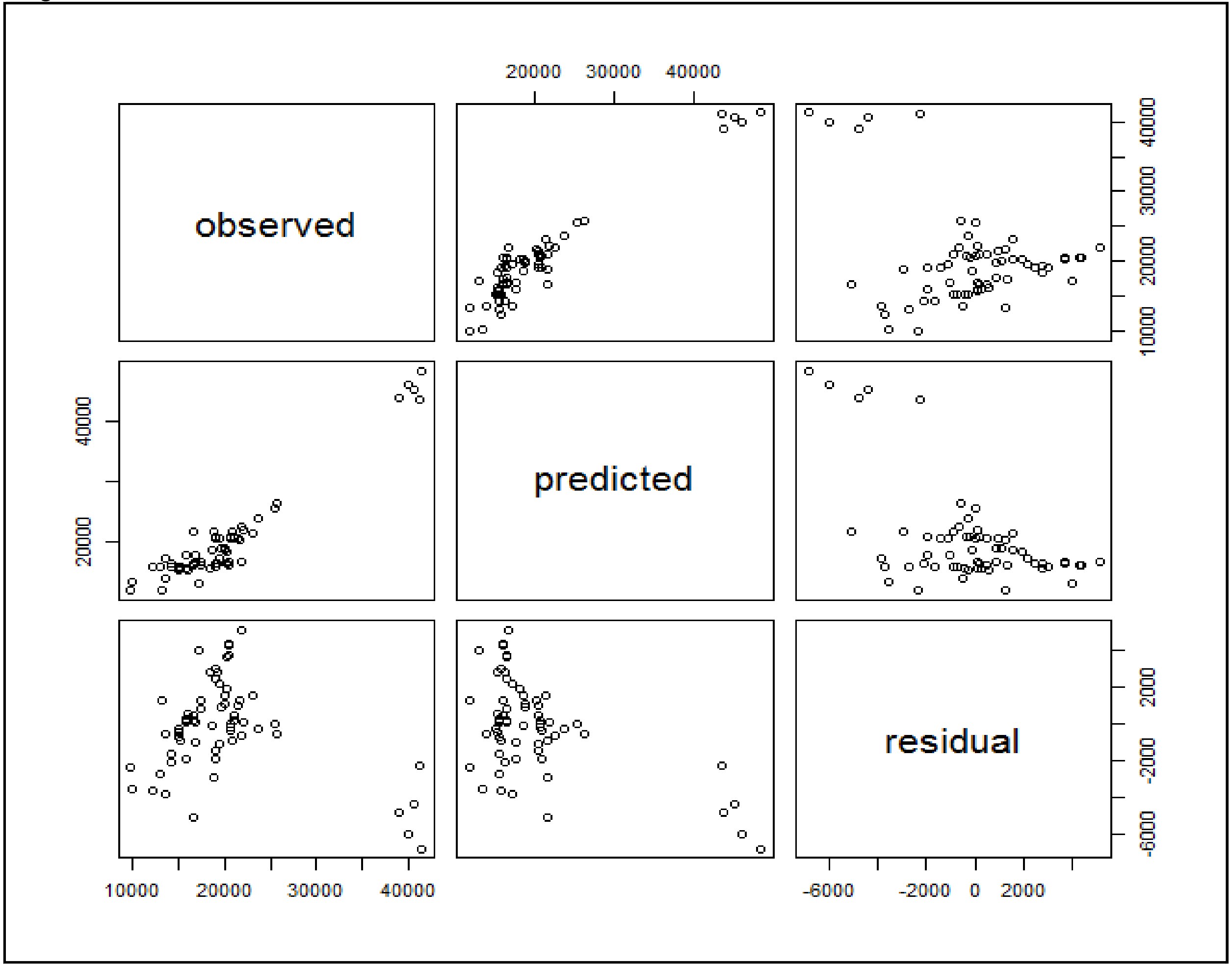
Figure 5: Relative Importance of Predictor Variables in the Random Forest Model



Residual Analysis for the Random Forest Model

Figure 6 shows correlation plots for observed, predicted, and residual values.

Figure 6: Correlation Plots of Observed, Predicted, and Residual for the Random Forest Model



COMPARISON OF FORECASTS AND ERRORS FOR THE TWO TECHNIQUES

Table 4 shows the actual attendance, the ordinary least squares predicted attendance, and the random forest predicted attendance for five home matches of the Chicago Fire in 2015.

Table 4: Comparison of Forecasts to Actual Attendance

Home Team	Visiting Team	Match Date	Actual Attendance	Ordinary Least Squares	Random Forest
Chicago Fire	Dallas	8/2/2015	14,209	14,835	15,823
Chicago Fire	New York	8/26/2015	11,196	13,942	12,066
Chicago Fire	Orlando	9/19/2015	20,280	15,127	16,459
Chicago Fire	New England	10/3/2015	16,694	14,519	15,731
Chicago Fire	New York	10/25/2015	19,850	15,931	15,997

Error Analysis

Three errors are reported in Table 5 -- mean error, mean squared error (MSE), and mean absolute percent error (MAPE).

Table 5: Comparison of Errors for Least Squares and Random Forest

Error	Least Squares	Random Forest	Least Squares/ Random Forest
mean error	-46.70159	-180.2708	0.259
mean squared error (MSE)	9410736	6247593	1.506
mean absolute percent error (MAPE)	12.03868	9.68865	1.243

For both techniques, the mean error shows that the predicted attendance over-predicts the observed. An adjustment can be made for the mean error when producing forecasts for future matches. MSE is the usual measure used to compare two or more techniques. Here the MSE for least squares is 50 percent higher than that for random forest. Lastly, MAPE, a measure more easily understood by non-statisticians, is 24 percent higher for least squares than for random forest.

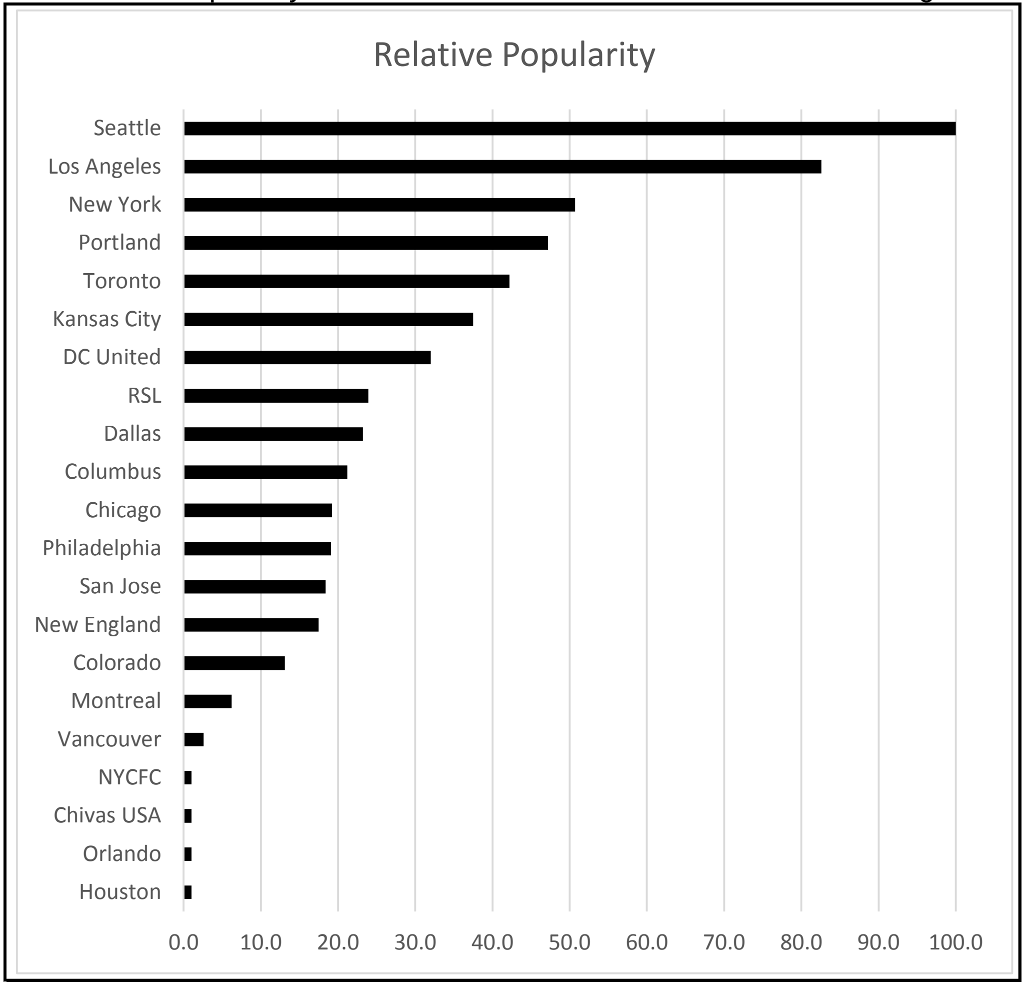
Interestingly, although the MSEs are considerably different, when applying the Diebold-Mariano test to the two-sided null hypothesis, the p-value is 0.3905 implying that the null hypothesis of equal forecast accuracy cannot be rejected.

SUMMARY AND SUGGESTION FOR FUTURE RESEARCH

We have shown the amount of effort to produce attendance forecasts using two techniques. The traditional ordinary least squares technique requires more effort than does the random forest technique and overall the random forest technique produces better results. Additionally, random forest is freely available as a package to the open-sourced R statistical language but is only available to SPSS as an extra module.

This analysis did not consider the win/loss record of the home team although team_popularity, shown in Figure 7, acted as a surrogate. Future research should explicitly include a measure of winning.

Figure 7: Relative Popularity of a Team as a Function of the Number of Google Searches



Although the differences in accuracy are not statistically significant, they are managerially significant in that random forest produces errors that are 50 percent lower than are least squared errors.

APPENDIX

The following is an analogy-based explanation of random forest by Edwin Chen (n.d.):

> Suppose you’re very indecisive, so whenever you want to watch a movie, you ask your friend Willow if she thinks you’ll like it. To answer, Willow first needs to figure out what movies you like, so you give her a bunch of movies and tell her whether you liked each one or not (i.e., you give her a labeled training set). Then, when you ask her if she thinks you’ll like movie X or not, she plays a 20 questions-like game with IMDB, asking questions like “Is X a romantic movie?”, “Does Johnny Depp star in X?”, and so on. She asks more informative questions first (i.e., she maximizes the information gain of each question), and gives you a yes/no answer at the end.
>
> Thus, **Willow is a decision tree for your movie preferences**.

But Willow is only human, so she doesn't always generalize your preferences very well (i.e., she overfits). To get more accurate recommendations, you'd like to ask a bunch of your friends, and watch movie X if most of them say they think you'll like it. That is, instead of asking only Willow, you want to ask Woody, Apple, and Cartman as well, and they vote on whether you'll like a movie (i.e., **you build an ensemble classifier**, aka a forest in this case).

Now you don't want each of your friends to do the same thing and give you the same answer, so you first give each of them slightly different data. After all, you're not absolutely sure of your preferences yourself – you told Willow you loved Titanic, but maybe you were just happy that day because it was your birthday, so maybe some of your friends shouldn't use the fact that you liked Titanic in making their recommendations. Or maybe you told her you loved Cinderella, but actually you really really loved it, so some of your friends should give Cinderella more weight. So instead of giving your friends the same data you gave Willow, you give them slightly perturbed versions. You don't change your love/hate decisions, you just say you love/hate some movies a little more or less (formally, **you give each of your friends a bootstrapped version of your original training data**). For example, whereas you told Willow that you liked Black Swan and Harry Potter and disliked Avatar, you tell Woody that you liked Black Swan so much you watched it twice, you disliked Avatar, and don't mention Harry Potter at all.

By using this ensemble, you hope that while each of your friends gives somewhat idiosyncratic recommendations (Willow thinks you like vampire movies more than you do, Woody thinks you like Pixar movies, and Cartman thinks you just hate everything), the errors get canceled out in the majority. Thus, **your friends now form a bagged (bootstrap aggregated) forest of your movie preferences**.

There's still one problem with your data, however. While you loved both Titanic and Inception, it wasn't because you like movies that star Leonardo DiCaprio. Maybe you liked both movies for other reasons. Thus, you don't want your friends to all base their recommendations on whether Leo is in a movie or not. So, when each friend asks IMDB a question, only a random subset of the possible questions is allowed (i.e., **when you're building a decision tree, at each node you use some randomness in selecting the attribute to split on**, say by randomly selecting an attribute or by selecting an attribute from a random subset). This means your friends aren't allowed to ask whether Leonardo DiCaprio is in the movie whenever they want. So, whereas previously you injected randomness at the data level, by perturbing your movie preferences slightly, now you're injecting randomness at the model level, by making your friends ask different questions at different times.

And so, **your friends now form a random forest.**

REFERENCES

Baade, R. A., & Tiehen, L. J. (1990). An analysis of major league baseball attendance, 1969-1987. Journal of Sport & Social Issues, 14(1), 14-32.

Breiman, L, & Cutler, A., (n.d.). retrieved from http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.

Chen, E., (n.d.). retrieved from blog.echen.me/2011/03/14/laymans-introduction-to-random-forests/

Goleman, A., & Taylor, Z. M. (2012). The Effects and Consequences of Advancing Technology in the Sports Industry: The Declining Attendance Rates at NFL Games.

Hogan, V., Massey, P, & Massey, S. (2012). Analysing match attendance in the European rugby cup. University College Dublin Centre for Economic Research Working Paper Series.

James, G., Witten, D., Hastie, T., & R. Tibsshirani. (2013). An introduction to statistical learning with applications in R. New York, NY: Springer. pp. 319-321.

JEquiha, (n.d.). retrieved from http://stats.stackexchange.com/questions/99527/random-forest-for-large-number-of-variables-and-predictions.

Welling, S., (n.d.). retrieved from http://stats.stackexchange.com/questions/141619/wont-highly-correlated-variables-in-random-forest-distort-accuracy-and-feature.