PREDICTING NATIONALS PARK ATTENDANCE

5/10/19

By:
Matt Dorris

I. INTRODUCTION

The Washington Nationals play in Southeast, D.C. along the Capitol Riverfront with fans mostly in the D.C. metropolitan area. When looking for entertainment, what causes fans to come to a Major League Baseball game, specifically one at Nationals Park? We were interested in analyzing trends with Nationals Park attendance to see if there was any predictive power on the attendance at games. Projecting stadium attendance can be useful for business advertisements, pricing tickets, ballpark promotions, and even projecting transportation in our nation's capital.

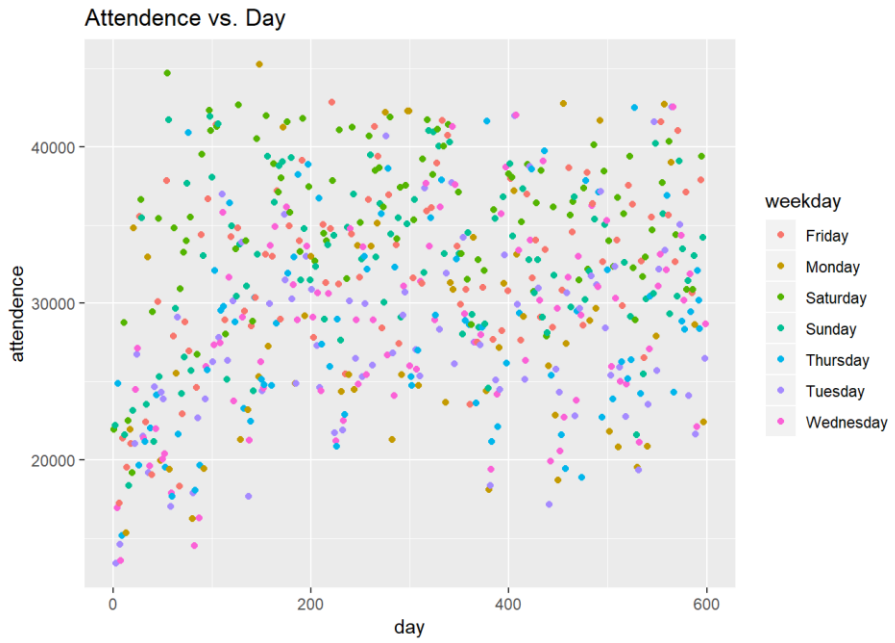
II. DATA

When searching for data, we were looking for certain characteristics what could be used to predict attendance. On any given day, we wanted to run a model to project attendance. There weren't many great data sets for the factors we wanted to look at so we decided to make a web scraper to pull data from www.baseball-reference.com for every Nationals home game for the last 8 years. Ultimately, our dataset consisted of the date, time of the game, day of the week, and temperature. We wanted to predict attendance using the other variables as predictor variables. The data set was broken up into a random training/testing 70/30 split to model the trees.

Variable	Description	Example
date	date that the game was played on	Saturday, April 2, 2011
attendance	recorded attendance at game	24875
day	the game in a sequence of all the games	1
weekday	day of the week the game was played on	Saturday
month	month the game was played on	June
time	if the game was in the day or night	day
temp	game time temperature of the game	85

Before we wanted to run any machine learning techniques, we wanted to do some exploratory data analysis to see if anything jumped out at us. After plotting the attendance vs.

time for the 8-year span of home games, we noticed something to look out for in our later studies. There appeared to be drastic changes in attendance for different days of the week. It appears people were more likely to show up on the weekends as opposed to weekdays. We will investigate this further later in the report, and back this claim up with data.

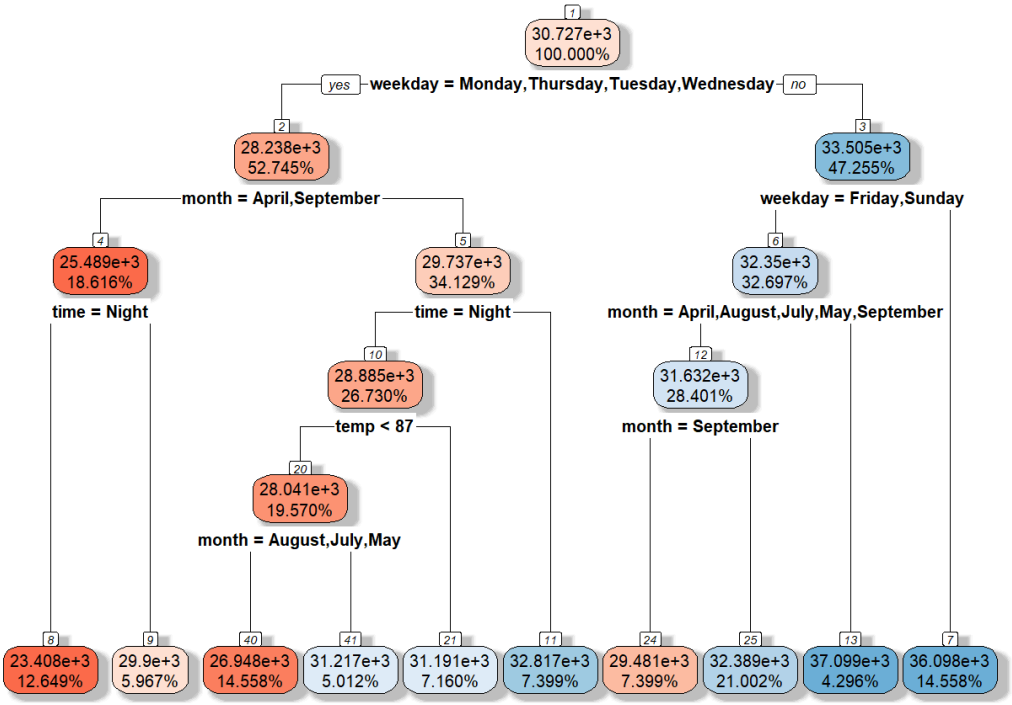


III. REGRESSION TREES

1. Full Tree

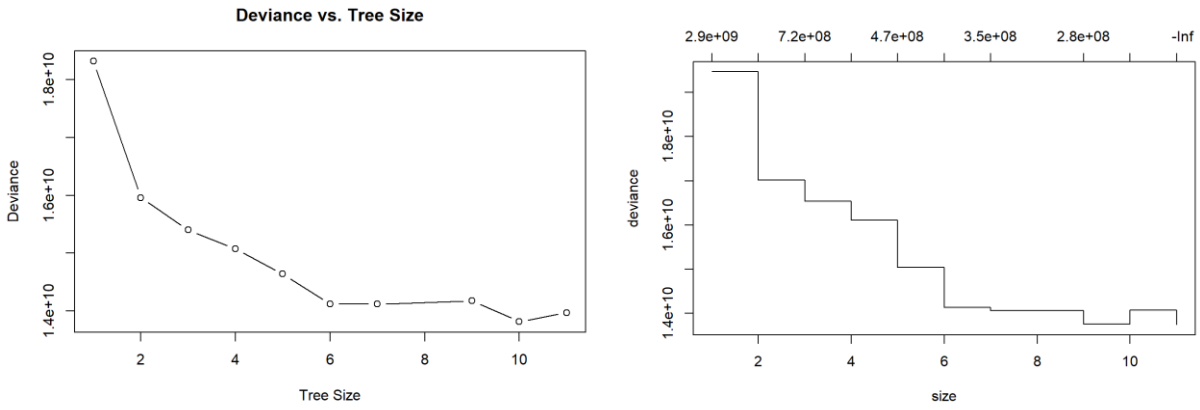
Our response variable, attendance, is continuous, and we wanted to characterize the average behavior of attendance as a function of both the categorical and numerical predictors. Regression trees split the predictor space into regions based on the other variables, so the response variable is well represented in each region. These regions are used to make decisions on

the data and can be very useful for predictions. For our dataset, first we implemented a regression tree using all the predictors for attendance. The root mean square error value was 5678. We can clearly see how day of the week and month have the strongest impact.



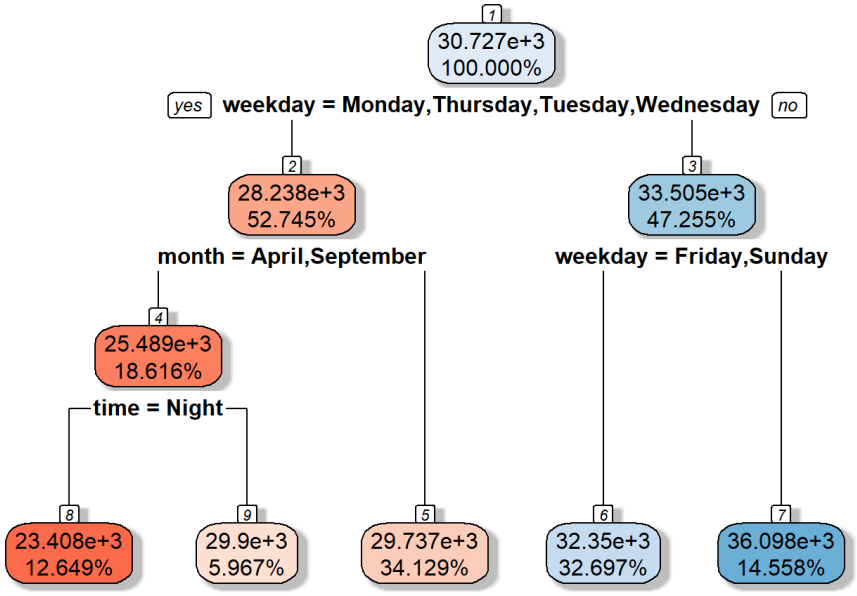
2. Cross Validation/ Regularization

Because trees often over-fit the data, it is important to apply regularization techniques. We want to ensure that the model is a good fit, without being too complex. To find the optimal number of nodes in a tree, cross validation is essential. Pruning is a method that aims to find the optimal number of nodes, while comparing the deviance of each additional node. Ideally, a model with the first lowest deviance is the best model. Below are plots of deviance vs. size of the tree. As we can see, a model with 5 nodes is optimal when regarding the fit/complexity tradeoff.



3. Reduced Tree

A tree with 5 nodes was optimal for predicting attendance in terms of being a good fit and not being too complex. The root mean square error value was 5732, which was only slightly worse than the full tree. Again, we can see the significance of the day of the week on the attendance at the game. The weekends tend to have much higher importance, followed by month, time and temperature, respectively. The model is now a simpler representation of the data.

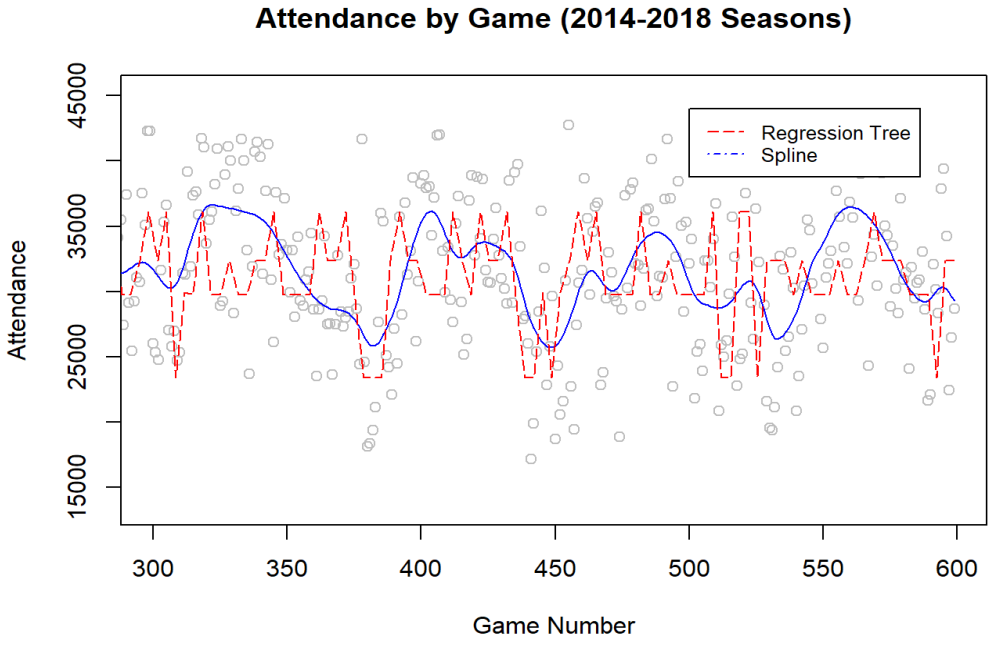


IV. DATA ANALYSIS

A few takeaways from the regression tree: when predicting attendance, the day of the week is the most important variable, followed by month, time, and temperature, respectively. This supports our earlier exploratory scatterplot which appeared to have high attendance on Friday, Saturday, and Sunday. The root mean square error value was 5732 which is relatively low and close to prediction values from the splining model.

```
Variable importance
weekday   month    time    temp
     45      26      23       7
```

Modeling attendance with regression turned out to be well suited for the data. Below is a plot of both the regression tree and a spline over the span of the last 4 seasons. We can see that the models produce somewhat similar results.



V. CONCLUSION

Predicting attendance at an MLB game can be rewarding and can lead to many benefits. If you are on the Nationals ticket sales side, you have the extra analytics to price tickets. If you are on the city of DC transportation board, you can project when to have more METRO stations open and when it might be more crowded. From the marketing side, you have insights for when the best time to have promotions or free giveaways. The next question is, how can we use predictive analytics to help bring the Washington Nationals their first ever World Series.