Introduction

Baseball is one of America's most popular sports. Major League Baseball is the league with the best players from around the world. MLB is watched by millions all over the world. There are 30 MLB teams that play 162 games in a season. Popular teams can attract millions of fans that attend their home games each season. It can be helpful to understand what makes fans attend games in person. Is it a team's geographic location, a winning team, or perhaps people just like to see a lot of home runs. I will explore a team's statistics to see what effects if any it will have on attendance. Having a successful winning MLB team can be very expensive, but it might not lead to an increase in attendance. From a business perspective, it is good to know how much a team's performance can affect attendance.

Data

Data is from <http://www.seanlahman.com/baseball-archive/statistics/>

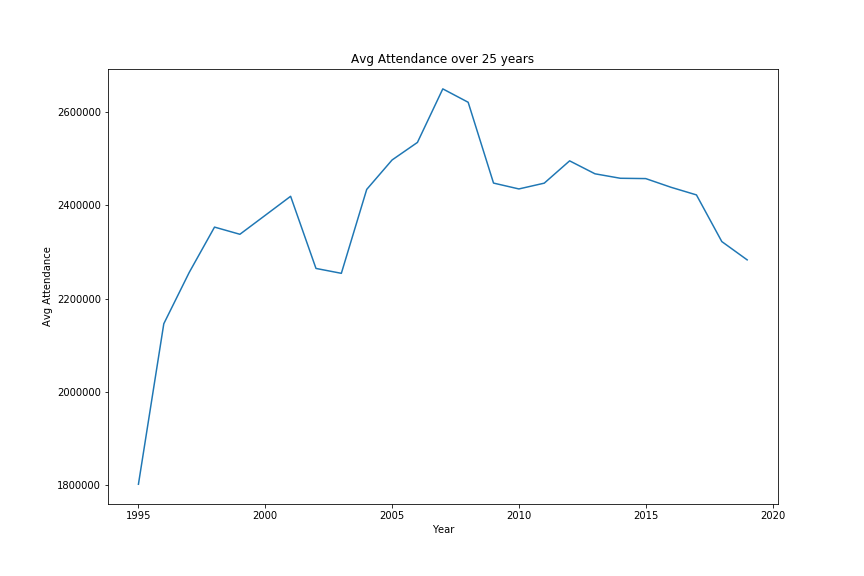
It contains statistics from 1871-2019. The teams.csv contains the teams statistic for every season. I will use the seasons from 1995-2019 when the wild card was introduced to the league. This would give us 25 years of stats on 30 teams (28 teams for 3 years). The column names are in table. Explanation of each stat is in the readme file that comes with the data.

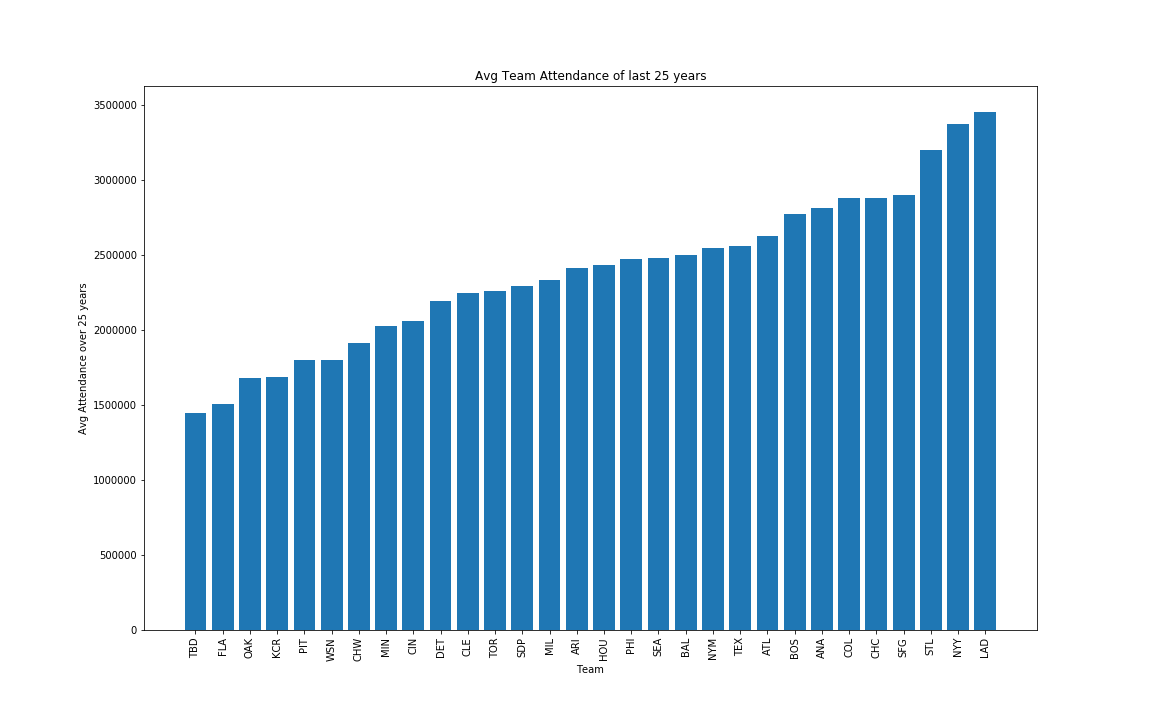
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| yearID | lgID | teamID | franchID | divID | Rank | G |
| Ghome | W | L | DivWin | WCWin | LgWin | WSWin |
| R | AB | H | 2B | 3B | HR | BB |
| SO | SB | CS | HBP | SF | RA | ER |
| ERA | CG | SHO | SV | IPouts | HA | HRA |
| BBA | SOA | E | DP | FP | name | park |
| attendance | BPF | PPF | teamIDBR | teamIDlahman45 | teamIDretro |  |

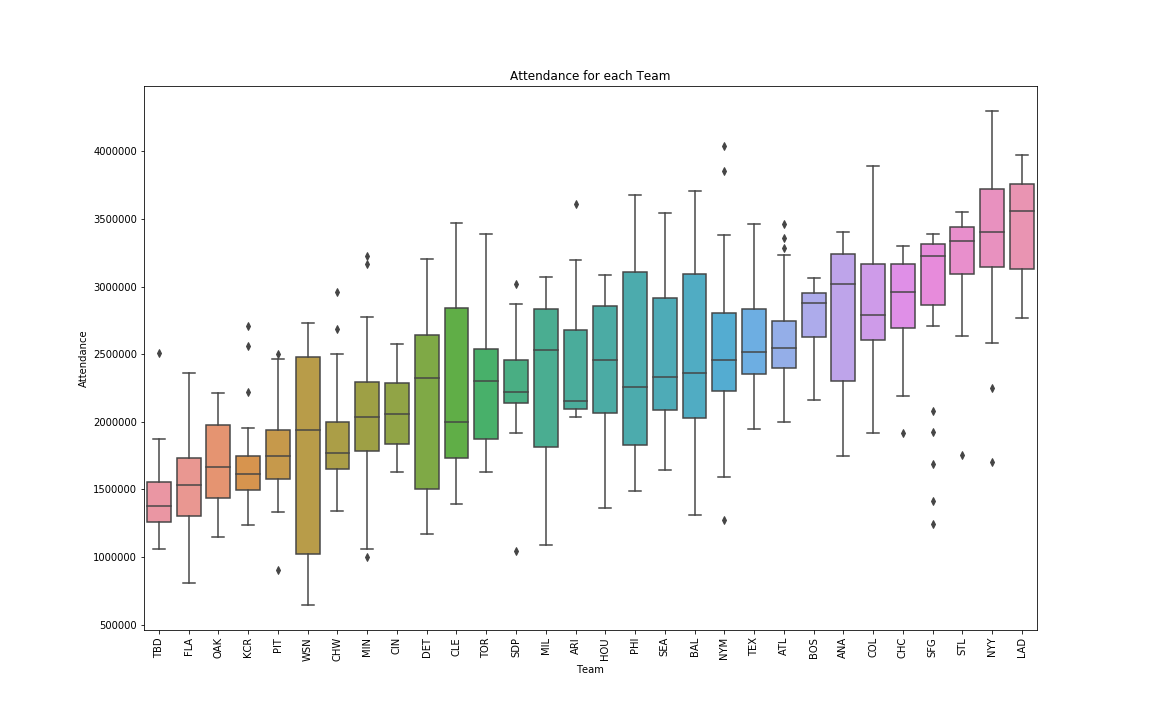
Methodology

I removed some of the columns that are just IDs that link to other databases. I kept ‘franchID’ to identify each franchise/team instead of using ‘teamID’, ‘name’, or ‘park’. Reason is some teams have name changes or different IDs or they have new parks/stadiums.. The franchise ID was consistent in identifying the teams. Although changing park/stadiums may have an affect on attendance, I don’t think there are enough stadium changes over the 25 years. Most new stadiums were built near the old stadiums anyways.

Let’s look at the changes in attendance over 25 years.

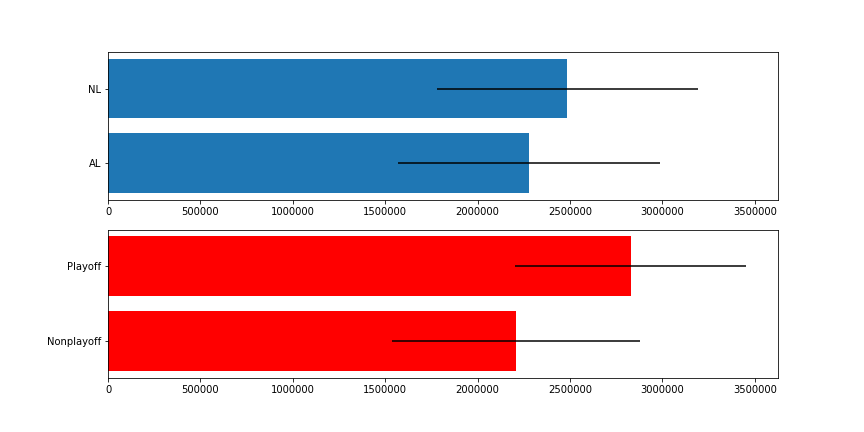


Some of the data are categorical like team, league and division winners, wild card winners, league champions and world champions. For teams, there are 30 unique teams, 2 of them were expansion teams that started playing in 1998. Below is the bar graph that shows the mean attendance for each team over 25 years (22 for the two expansion team)



A boxplot of the team attendance with the same ordering as the bar graph.

MLB is split into two leagues. The American league and National league. The difference between each league is the DH position that would bat for pitchers in the American league. Some fans enjoy some of the strategies of the National league that involve substitution of the pitcher during key batting opportunities. Let’s group the attendance data by their league and see what’s the difference.

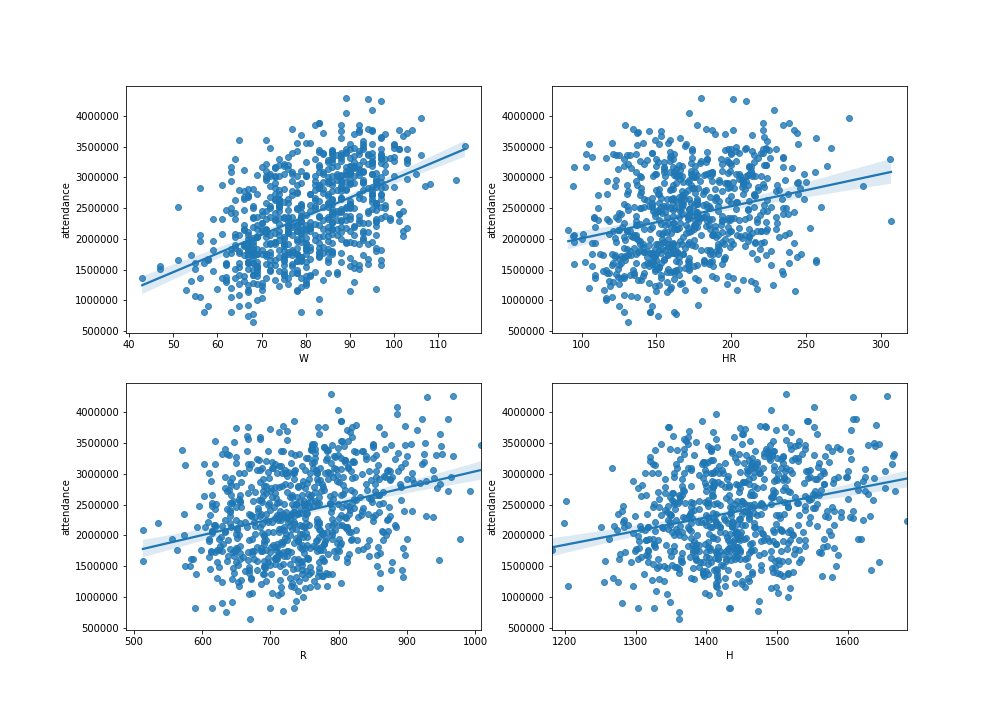


There doesn’t seem to be much difference in attendance between the two leagues. Playoff teams on the other hand shows significantly higher attendance than non playoff teams. As far as I know, the data for attendance is only for regular season games. So the attendance stat doesn't have the added playoff game attendance.

For numerical stats I used pandas corr function to measure the pearson’s coefficient between some of the stats and attendance. I sorted the order from most positively correlated to negatively correlated in the table below

|  |  |  |  |
| --- | --- | --- | --- |
| W | 0.500461 | SB | -0.029760 |
| R | 0.312266 | 3B | -0.061749 |
| H | 0.274978 | SO | -0.074900 |
| HR | 0.269356 | CS | -0.139414 |
| BB | 0.237644 | E | -0.232152 |
| 2B | 0.112769 | ER | -0.263962 |
| SF | 0.110917 | RA | -0.286550 |
| HBP | 0.041127 | ERA | -0.299198 |
|  |  | L | -0.451873 |

Here are 4 scatter plots of the 4 highest correlated stat with their linear regression lines.



Data would be used as features in regression models to predict the attendance. Teams would be one hot encoded, which gives 30 features. Binary categories such as lgID and DivWin, WcWin, LgWin, WsWin would be converted to (‘N’, ‘Y’) = (0, 1) and (‘NL’, ‘AL’) = (0, 1). Shape of the input dataset is (744, 64). 744 samples with 64 features. Input data set is randomly split into 80% training set and 20% test set. For the neural network inputs, training and test set is scaled with the mean and standard deviation from the training set. The regression models used are linear regression, ridge regression, support vector machines, decision tree, random forest, gradient boost, and fully connected neural networks with different architectures. The metrics used are root mean squared error and r2 score. The results are in the table in the results section.

Results

|  |  |  |
| --- | --- | --- |
| Model | RMSE | R^2 |
| Linear Regression | 423805.97 | 0.619 |
| Ridge Regression | 427167.02 | 0.613 |
| SVM | 788233.89 | -0.317 |
| Decision Tree | 712579.58 | -0.076 |
| Random Forest | 549887.11 | 0.359 |
| Gradient Boost | 461068.33 | 0.549 |
| Dense Neural Network  2 hidden layers  1st 256 nodes with sigmoid activation  2nd 128 nodes with relu activation | 388487.77 | 0.680 |

I am surprised by how well both linear regression (linear and ridge) did compared to everything besides the neural network. Maybe SVM and decision tree overfit the training data. The neural network took about 300 epochs with batch size of 1 to converge. Various architecture hyperparameters were tried. From 1 to 3 hidden layers, node size from 16 to 256, activation functions tried were relu, sigmoid and tanh.

Discussion

Even though the number of wins has a pretty good correlation with attendance, using it as the sole input for linear regression gave a rmse of 615467. Comparing that with the 64 features input which score an rmse of 423805 for linear regression. From looking at the bar graph of team and their attendance, we can see the big market teams like NY, LA, SF are at the top of the attendance. These are the big market teams that are in populous cities. Looking at the boxplot they have consistently high attendance. The stats used were team stats, which ignores the star players for each team. Adding something like the number of all-star players from each team might improve the models.

Conclusion

I created a neural network that would predict the season attendance of a baseball team in the MLB. A winning team that can make it to the postseason will help increase attendance. The team hitting and pitching stats probably don’t make as much of a difference as long as they lead to wins. A baseball friendly market is more important. So don’t start a team in Florida.