Analysis and Prediction of Spectator Turnout at Baseball Games Based on Weather Effects

Sheel Patel
Rutgers University, New Brunswick
NetID: ssp189

Vidit Desai
Rutgers University, New Brunswick
NetID: vd235

1 INTRODUCTION

For a particular game-day in any professional event, the total turnout of spectators is important not only for the encouraging the home team but it has also some notable effects on the management. The inventories needs to be purchased in advance by the stadium, which should be in proportion to the total number of spectators attending. However, if there is incorrect data on the turnout there may be huge amount of either shortage or wastage. For example, lets say for a particular game the average attendance in that stadium is 35,000. So inventories like drinks or hot-dogs are purchased keeping that figure in mind. However, if the attendance turns out to be 25,000 there would be a huge amount of wastage. The opposite can also be true.

We are particularly focused on the effect of weather conditions on the overall turnout for a particular game. For that analysis, outdoor games are best for the consideration because if weather had any effect, it would be on these games. So, we have selected Major League Baseball(MLB) for the analysis.

To perform the analysis, We required all the data related to individual games in entire seasons such as the attendance in each match, the time, sky and weather conditions on that day. We used the dataset of retrosheet.org for our analysis. The details regarding the data processing and formatting will be discussed further in the report. We will also go over the linear regression model used to analyze the patterns and then make prediction of the future attendance given the conditions.

2 DATA COLLECTION AND PREPROCESSING

This section contains information regarding the raw data and preprocessing. We explain how and from where we obtained the all the data that was required for this project and what filtering and preprocessing was done to extract useful data.

2.1 Required Data

The project requires data for each MLB game such as location, teams playing, audience attendance and also the weather conditions on the match day.

2.2 Source

We collected the data from **Retrosheet** (retrosheet.org), which provides historical and current play-by-play game data for baseball games. We have collected data for each of the games played from 2000 to 2016. Theres one file for each home team and the matches they played in the entire season. The data contains each game match details including players and play-by-play records. More detailed description of the data can be obtained at http://www.retrosheet.org/eventfile.htm

2.3 Data Format

Files are ASCII text files consisting of a series of records. Each file has the home games for one team for one season. The file names are of the form YYYYTTT.EVL, where YYYY is the 4 digit year, TTT is the Retrosheet team abbreviation and L is the one letter league abbreviation. The record fields are comma separated and all players are identified by a unique 8 character code. Each game is associated with a twelve character ID record which identifies the date, home team, and

number of the game. For eg., ATL198304080.

There are up to 34 info records, each of which contains a single piece of information, such as the temperature, attendance, wind, etc. The record format is info,type,data. The records following the info records contain play by play instances. We arent using these data records and the info records are sufficient for our analysis. info records are of two general kinds, game-related and administrative.

Some examples of info records:
info,visteam,SDN
info,hometeam,ATL
info,date,1983/04/08
info,daynight,night
info,sky,cloudy
info,temp,54
info,windspeed,2

2.4 Preprocessing and Final Format

We filtered this available data in order to extract useful information and discard all the unnecessary data and write it in a clear format useful for our operations. We wrote a python script to filter the files in this way. The result was stored as a .txt file with comma separated values.

Example of these new records:

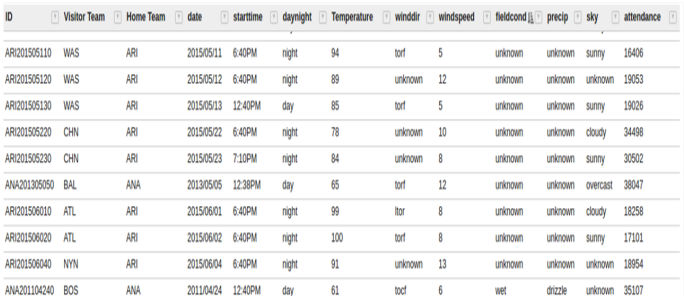
ANA201004220,DET,ANA,2010/04/22,7:08PM, night,56,tocf,4,wet,none,cloudy,37338

ANA201004230,NYA,ANA,2010/04/23,7:08PM, night,62,torf,10,unknown,drizzle,sunny,44002

These records have the following fields in this order:
id, visteam, hometeam, date, starttime, daynight,temp, winddir, windspeed, fieldcond, precip, sky, attendance

3 DATA ANALYSIS

The following figures are charts that represent the audience attendance plotted against various factors, for two teams: Anaheim Angels and Colorado Rockies, for the year 2015.
Spark and Visualizations using Apache Zeppelin.



ID	Visitor Team	Home Team	date	starttime	daynight	Temperature	winddir	windspeed	fieldcond	precip	sky	attendance
ARI201505110	WAS	ARI	2015/05/11	6:40PM	night	94	torf	5	unknown	unknown	sunny	16406
ARI201505120	WAS	ARI	2015/05/12	6:40PM	night	89	unknown	12	unknown	unknown	unknown	19053
ARI201505130	WAS	ARI	2015/05/13	12:40PM	day	85	torf	5	unknown	unknown	sunny	19026
ARI201505220	CHN	ARI	2015/05/22	6:40PM	night	78	unknown	10	unknown	unknown	cloudy	34498
ARI201505230	CHN	ARI	2015/05/23	7:10PM	night	84	unknown	8	unknown	unknown	sunny	30502
ANA201305050	BAL	ANA	2013/05/05	12:38PM	day	65	torf	12	unknown	unknown	overcast	38047
ARI201506010	ATL	ARI	2015/06/01	6:40PM	night	99	ltor	8	unknown	unknown	cloudy	18258
ARI201506020	ATL	ARI	2015/06/02	6:40PM	night	100	torf	8	unknown	unknown	sunny	17101
ARI201506040	NYN	ARI	2015/06/04	6:40PM	night	91	unknown	13	unknown	unknown	unknown	18954
ANA201104240	BOS	ANA	2011/04/24	12:40PM	day	61	tocf	6	wet	drizzle	unknown	35107

Figure 1: Screenshot of record table

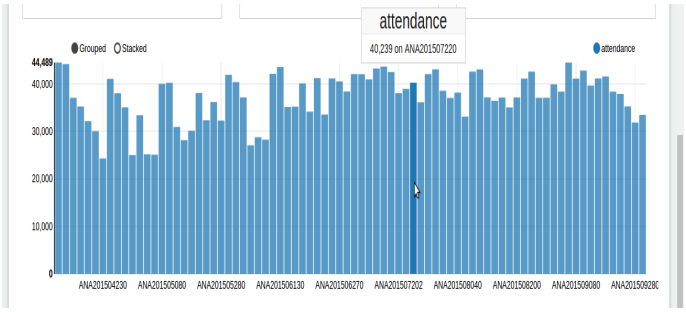


Figure 2: Anaheim Angels, 2015, Per-game Attendance

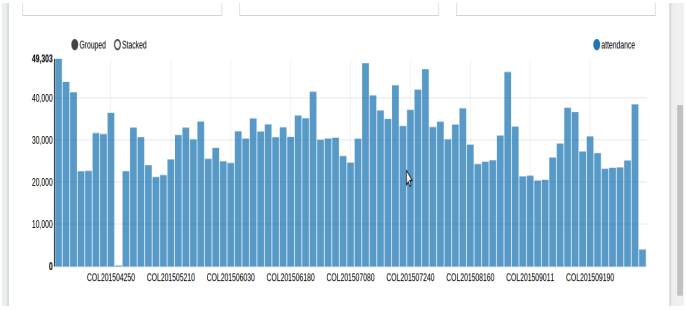


Figure 3: Colorado Rockies, 2015, Per-game Attendance

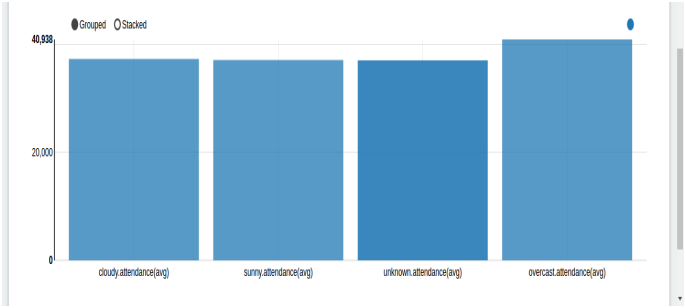


Figure 4: Anaheim Angels, 2015, Avg Attendance vs Sky Conditions

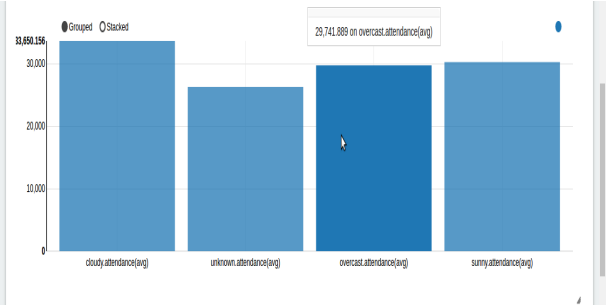


Figure 5: Colorado Rockies, 2015, Avg Attendance vs Sky Conditions

Figure 6: Anaheim Angels, 2015, Day/Night Avg Attendance

Figure 7: Colorado Rockies, 2015, Day/Night Avg Attendance

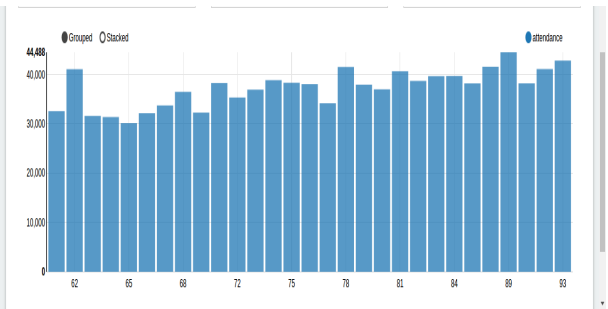


Figure 8: Anaheim Angels, 2015, Avg Attendance vs Temperature

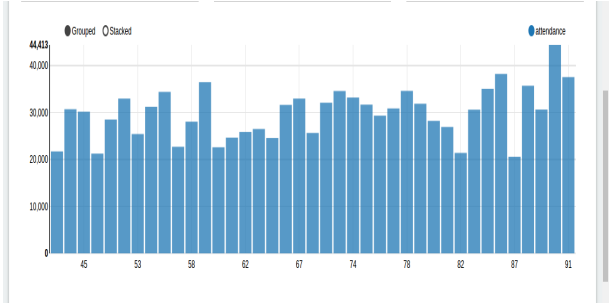


Figure 9: Colorado Rockies, 2015, Avg Attendance vs Temperature

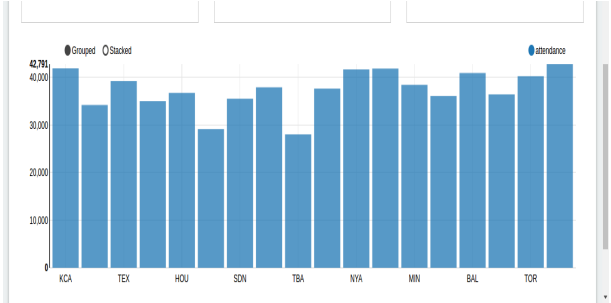


Figure 10: Anaheim Angels, 2015, Avg Attendance based on Visiting Team

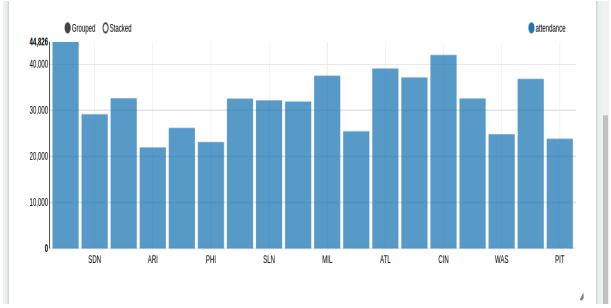


Figure 11: Colorado Rockies, 2015, Avg Attendance based on Visiting Team

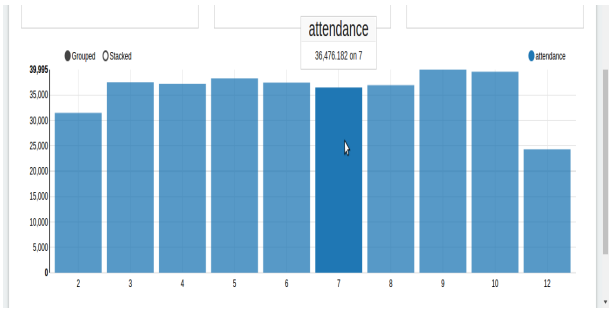


Figure 12: Anaheim Angels, 2015, Avg Attendance vs Wind Speed

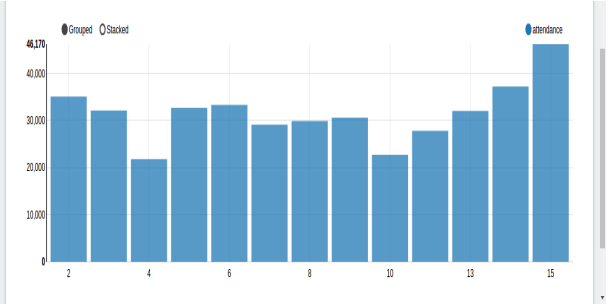


Figure 13: Colorado Rockies, 2015, Avg Attendance vs Wind Speed

As we can see in the chart, the attendance per game for Anaheim Angles is evenly distributed compared to the the attendance per game graph for Colorado Rockies. So analysis on Anahein Angels may not give interesting insights on the effect of various factors on it. However teams like Colorado Rockies might be affected by the conditions and might give some relationships or pattern. So we plot the average attendance to various entities and check the difference and the same effect is noticed in these observations i.e. even distribution for Anahein Angels and somewhat variations in Colorado Rockies. Detailed search[3] showed that the Anahein Angels have greater fanbase compared to Colorado Rockies and therefore their games are almost full. So we will continues our analysis on team with less fanbase and more variation in attendances.

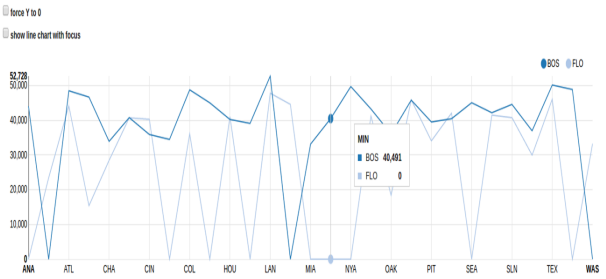


Figure 14: Variability in Attendance BOS vs FLO

The graph in Figure 14 compares the variability in attendance between two teams: BOS- Boston Red Sox, a large market team and FLO- Miami Marlins, a small market team. The Red Sox have a much higher and consistent attendance whereas the Marlins see a comparatively low and much fluctuating.

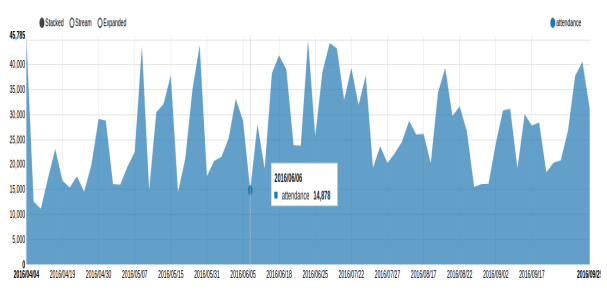


Figure 15: Attendance when Home Team: Baltimore Orioles

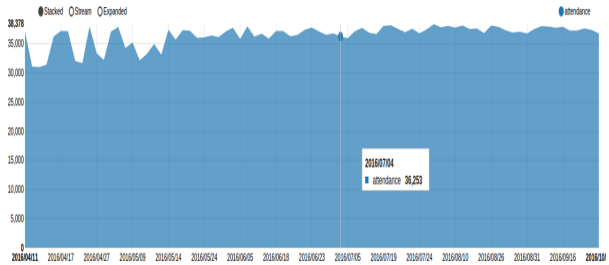


Figure 16: Attendance when Home Team: Boston Red Sox

The graphs in Figure 15 and 16 show the attendance on particular dates when the home team were Baltimore Orioles(small market team) and Boston Red Sox(large market team) respectively. As we can see, there are a lot of spikes in attendance rates for the Orioles. The spikes for Orioles are on the weekends and the attendance is pretty low on weekdays. In comparison, the attendance rate is much linear for Red Sox, with very low difference between days. This shows that fans come to almost all games for Red Sox and it doesn't matter if it's a weekday or weekend. This is because of the team market size and popularity.

The following are the standard deviations in some of the teams in order from less to more marketable teams:
Blue Jays 11785.46
Orioles 10442.70
Tigers 6219.05
Marlins 5928.10
Giants 2958.11
Red Sox 1780.72

This variability in data for different teams because of popularity and other factors lead us to considering the home and visiting teams as

predictors for regression. Including them as predictors improved our predictions a lot compared to the results we were getting before utilizing them.

	Start-time	Day-night	Temperature	Wind-speed	attendance
mean	5.6552029259	0.3268640868	73.4358777725	7.6091316659	30437.5182869
std	2.87773858	0.4690809478	10.8961944917	5.1549874611	10021.7370765
min	1	0	23	0	8701
max	12	1	107	58	56000

Figure 17: Statistics for predictors

Figure 17 shows the Mean, Standard Deviation, Minimum, Maximum of various predictors. Temperature varies from 23F to 107F with constant fluctuations and a mean of about 73F. We represented day as 1 and night as 0 and see that most of the games were played during the day. The attendance ranges from as low as 8701 and goes as high as 56000.

4 MACHINE LEARNING

Our aim was to predict the attendance for a particular game based on given weather conditions. So based on our dataset and requirements we considered Regression algorithm for the purpose of predicting because it suited the best for our dataset. The following sections includes the details of how we fit and predict data, the results obtained and the insights gained.

4.1 Prediction

We used Linear Regression for prediction using temperature, windspeed, condition , Visiting team , precipitation as predictors for fitting it in our model . We used the existing sklearn librairies [4] for carrying out prediction.

We were first considering all the games played since 2000 to 2016 for generating our model. But adding these many data points skewed our results. So we just considered the data for 2010 - 2016 seasons for all the involved teams. We have a total of 17006 data points in our regression model.

The first step in this was to fit our data into the the model. For training the model we used the predefined fit(X,Y) function in scikit-learn library. We have many categorical values compared to the numerical values. We normalized tempearture, start time and windspeed to get better result. We created 1-0 vectors for all categorical values depending on the sky being sunny, cloudy, overcast, the month the game was played and the for each team visiting as well as home. So we had total of 72 predictors for our model.

For the purpose fo testing, we split the data 70-30 for training and testing. Attendance was included as the 2nd parameter in the function which is the value to be predicted.

4.2 Results

We obtained the following results from the regression model generated from our data.

Actual	Predicted
23524	29935
30784	33081
36244	34504
25748	25152
25236	21237

Table 1: Prediction Results

The Table 1 shows a sample of our predictions. We were able to make near accurate predictions for most cases with some off predicted values. The RMSE(Root Mean Squared error) was about 6823, depending on sampling of training data.

The Table 2 shows the coefficients for the weather predictors that we obtained from regression. These coefficients offer us significant insights about dependency of attendance on the predictors.

For our model we got this regression variance score:
R^2 (coefficient of determination): 0.53

From Figure 18, we can see that our prediction model closely follows the actual data. As we can see from the crests and troughs, although there some shift in some of the values, our prediction

Predictors	Coefficients
starttime	-295.702
daynight	1335.260
Temperature	82.07
windspeed	-47.79
rain	1166.79
sunny	85.50
cloudy	-149.13
dome	-1140.62
overcast	205.65

Table 2: Coefficients of Predictors

model is accurately able to follow the trends in the actual attendance.

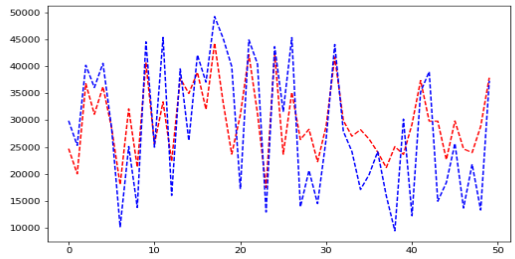


Figure 18: Plot of Prediction() vs Actual() Attendance

Some of our prediction results are varying from the actual attendance to such extent as we have very few numerical data and most of the predictors such as sky, sunny, cloudy, visiting teams, home teams are all categorical data and hence not giving much accuracy to our results. Also the fact that there are many other significant factors that we are not considering for the analysis including the city population, the rivalry, city income, location, etc., which do have a major affect on the game and the overall turnout.
We also found that some teams have a pretty loyal and large fanbase who attend the games in almost any condition. We saw this trend mostly in popular large market teams. Weather, time and weekend factors affected attendance for medium to small market teams to a much more extent.

4.3 Insigts

We gained some interesting insights from our regression model. We saw that with every 10F increase in temperature, there was an increase of about 820 attendees. On average there are about 1335 more people attending in day matches than night matches. There is a low increase of 85 to 205 in attendance when the sky conditions are sunny or overcast. Surprisingly, we saw that on average there are 1166 more spectators on rainy days.
We saw that windspeed didn't affect the attendance to a great extent. A lot of the games were played in domes but the attendance in such games would be on average 1140 less that of open stadiums, suggesting that some fraction of people prefer open stadiums.

5 FUTURE WORK

Till now we have the filtered data for all games played till now.
the next step is analyzing which teams does not have varying attendances and filter those team . We will only take this team under consideration for further analysis as the other teams will not have significant effect of the weather or other entities.After that, next is applying different formulas and plotting the results to find the pattern or relations.

We are planning to take 3-game average as one of the parameters for prediction. We will also explore other parameters for the same.

We will be looking into different Machine Learning algorithm and techniques which fits best to our system ad develop a tool which predicts the attendance.

6 CONCLUSION

We started this project with goal of how weather affects the overall turnout for an openground sports game. By analyzing the data for MLB games, our results do reflect some of the affects of the weather conditions including temperature and rain on the overall attendance. However, There were some variation in the predicted values compared to the actual values due to the fact that even though weather has its effect on the attendance there are other factors that have more influence over it which was not included into consideration in our work.
The factors such as Local Rivalries, City Popu-

lation, Average City income, Demographic location, etc does have major impact for a particular game.So if we would have more time, for the analysis and pivoted from our original idea to just include weather and considrerd more predictors as above for calculating the overall result.

7 RELATED WORK

1. J. James Reade,"Modeling and Forcasting Football Attendance",Oxonomics 2 (2007) 27-32. @ 2007 The Author. Journal compilation

2. E. Webb,"Predicting Day-to-Day Variability in Baseball Attendance to Support Staffing",BLOG: http://cantstoperic.com/

3. Robert J. Lemke, Matthew Leonard, Kelebogile Tlhokwane , "Estimating Attendance at Major League Baseball Games for the 2007 Season", Journal of Sports Economics 2010 11: 316 originally published online 6 August 2009

8 REFERENCES

1. $http://www.retrosheet.org/game.htm$

2. $https://zeppelin.apache.org/docs/0.7.0/interpreter/spark.html$

3. $http://baseballot.blogspot.com/2016/07/how-many-fans-does-each-mlb-team-have.html$

4. $http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html$