Take Me Out to The Ballgame: Predicting Baseball Game Attendance

Li-Hao Liu
New York University
lhl302@nyu.edu

Zhen Liu
New York University
zl1471@nyu.edu

***Abstract*—Using the training data to create different models that predict attendance at Yankees Stadium with other attributes. In this case, we use the data from season 2011 to 2016. Attributes include baseball data such as standing, promotion and date. Also, we use weather data such as humidity, wind speed and the chance of precipitation into the model. By doing this, we can predict the attendance data for 2017 and the baseball club can use this prediction to increase their income.**

Keywords—analytics, baseball, weather, attendance

I. INTRODUCTION

There are a lot of factors that might affect attendance of a ballgame. From the view of the field, team record might have a big impact, if one team is winning, it is more likely that people might support that team. Another point is the player record. If one player has the chance to achieve a milestone in next game, people are most likely to go and witness that moment. From the view of the field, weather could be one of the factors since more people would like to go the ballgame with their family on a sunny day. On the other hand, if it is raining or snowing, people might want to stay at home just watch the game on television. There are still a lot of factors that may have the impact on attendance, we try to put as many attributes as possible and see which one would be the strongest element for high/low attendance.

II. MOTIVATION

Is there anything better than going to a baseball game? Provided it's during the day, sunny, above seventy degrees, on a Saturday, and there is a free hat giveaway for the first 30k fans. The point being that there are many factors that go into a fans decision to show up to the ballpark. Our goal for this project is to explore these variables with historical attendance data and make a prediction on the game-to-game attendance totals for the New York Yankees upcoming season.

For the baseball team, if they can predict the attendance, they can prepare the food or drinks more precisely. This helps them to save money. Moreover, they can adjust their ticket price or events to increase their income.

For the audience, they can predict the car or subway traffic status around the ballpark. Deciding whether to drive or take the subway or even go earlier.

III. RELATED WORK

"Moneyball," is a book and movie about how the Oakland Athletics, the United States major league baseball team won many more games than expected in 2002 given the team's weak revenue situation. The story is about how the team used data analytics to help them find the right players to help the team win, despite their inability to pay for top talent. The A's accomplished this feat by using data and statistics to help them make key personnel, resourcing and management decisions. This book provides us with another example of how data analytics is changing our world. General Manager Billy Beane hires a young data analytics expert named Peter Brand to help him extract insight from data to make baseball decisions. Instead of trusting his veteran assistant coaches who had logged over 80 years of collective experience to make key coaching decisions based on their subjective wisdom, Beane relied on 'hard,' objective, statistical data produced by his analyst to help him make important baseball decisions. Their groundbreaking approach ultimately changed how baseball and many other sports use data. [1]

Besides using baseball data, we can also use weather data to see the performance of a team. Based on the research, the results show that, overall, offensive production is higher in warm temperatures compared to cold temperatures. Across all populations, a lot of stats show significant increases while walks show significant decreases in warm temperatures compared to cold temperatures. Additionally, the American League shows a much stronger impact of temperature on the statistics than the National League.

Consistent with past findings, home runs were most affected by temperature, increasing from 1.79 per game in cold temperatures to 2.35 per game in warm temperatures for away and home batters combined. Runs scored, which is arguably the most important statistic of an MLB game, showed the second strongest response to temperature in this study, increasing from an average of 8.95 per game in cold temperatures to 10.08 per game in warm temperatures overall. Home batters playing for the Oakland Athletics benefit the most from warm weather, where runs scored, batting average, slugging percentage, on-base percentage, and home runs all show significant increases in warm temperatures compared to cold temperatures.

In conclusion, we can see that temperature does have an impact on baseball, this article mentioned its effect on on-field, therefore, we are interested in off-field, that is, the attendance of every ballgame. We assume the weather will be the main reason for attending a ballgame. [2]

Besides that, we selected different features such as promotion, team standing or months so we can know which is the most important factor when people are deciding to go to a ballgame.

IV. DESIGN

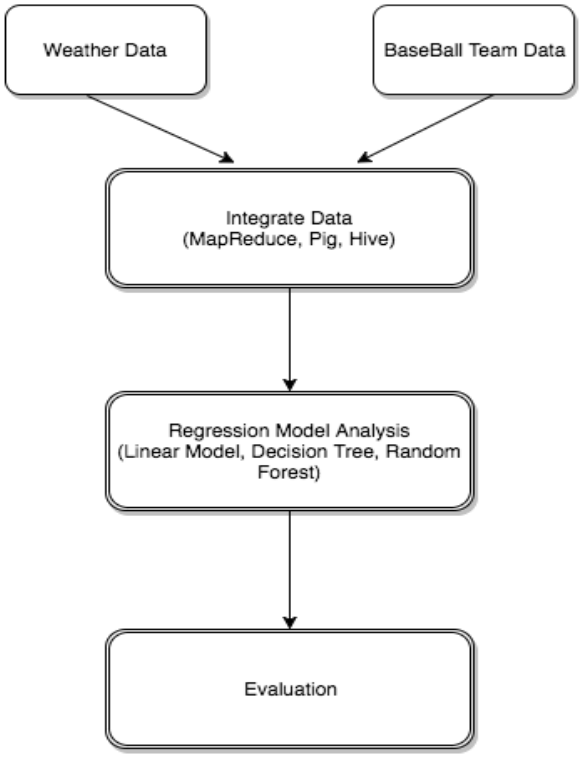
In this section, we will present our approach to analysis baseball game attendance based on past years team statistics and weather record.

a. Overview

The overall approach is depicted in Figure 1. We first ingested the baseball team statistics and weather record into HDSF.

Secondly, we cleaned and formatted the data before further processing. We replaced the missing values in weather record with the average value in certain months and deleted the unnecessary heading information in source data. We also designed the database schemata this step.

Thirdly, we ran Map reduce to join the weather record table and baseball record table together. Since HDFS support pig, hive, and to run MapReduce. We finished the joining step by both java code and hive script. For this step, we generated our final CSV file which is our reference to do the model analysis. Last, we use Spark to build some models.



b. Experiment Description

In this part, we will discuss the experiment to implement the design at part a.

Data Source Description and Exploration

Since we decided to do a 5-year time window analysis between baseball team record and weather attributes.
A comprehensive and precise weather data source is necessary. We decided to use "Weather Underground" as our weather data provider. To get the raw data we need, we first selected the location nearest the ballpark. For example, the station in Upper Manhattan is the one we want, since we chose Yankees Stadium for our example. Then we select each calendar year as the range to get the weather record from 2011 to 2016. The weather source data provides information about temperature, humidity, wind speed, precipitation and daily events such as rain, snow, and fog. One problem with our weather raw data is that the events and precipitation have missing values for some daily record. We are going to see these events as none.

Another part of our source data is the baseball team record. We choose baseballreference.com, the most authoritative data for our reference. To get the raw data we need, we first select the team, such as New York Yankees, and then get each year's schedule and results from the team data page, (http://www.baseballreference.com/teams/NYY/2016.shtml).
The file type of team record is CSV file. In next section, we will discuss our approach to clean and format the source data.

Data Cleaning and Preprocessing

In this section, we will explain our approach doing data cleaning and preprocessing before data integration. We've been noted that there are missing values in daily weather record that exist. We used Pig Latin to replace missing value with monthly. [4]

ReplacingWithZero.pig

```
A = LOAD 'weather.csv' USING PigStorage(',');
B = FOREACH A GENERATE $0, (($2 IS NULL) ? 0 : $2), (($8 IS NULL) ? 0 : $8) , (($14 IS NULL) ? 0 : $14) , (($17 IS NULL) ? 0 : $17), $21;
STORE B INTO 'output_weather_replace_zero';
```

As we discussed earlier, we also noted that there are duplicate headings in baseball team's raw data. Below is our approach to delete duplicate headings using pig.

DeleteDuplicateHeading.pig

```
A = LOAD 'game.csv' USING PigStorage(',');
B = DISTINCT A;
STORE B INTO 'output_remove_duplicate';
```

Since both of our raw data from weather and game record are quite comprehensive and our goal is to do big data analytics

between major weather attributes such as temperature and baseball game attendance, we removed the unrelated attributes that we don't need by using MapReduce. [3]

First, we copied the data into Hadoop HDFS, then executed MapReduce code.

Following is part of our Mapper code

```
if (filePath.contains("weather.csv")) {
  try {
     context.write(new Text(values[0]), new Text("w#" + values[2] + DELIMITER + values[8] + DELIMITER + values[14] + DELIMITER + values[17] + DELIMITER + values[21]));
   } catch (IndexOutOfBoundsException e) {
       System.out.println(values[0]);
   }
}
else if (filePath.contains("game.csv")) {
  try {
     context.write(new Text(values[0]), new Text("g#" + values[1] + DELIMITER + values[2] + DELIMITER + values[3] + DELIMITER + values[4] + DELIMITER + values[5] + DELIMITER + values[6] + DELIMITER + values[7] + DELIMITER + values[8] + DELIMITER + values[9] + DELIMITER + values[10] + DELIMITER + values[11]));
  } catch (IndexOutOfBoundsException e) {
       System.out.println(values[0]);
      }
```

Following is part of our Reducer code:

```
List<String> linkU = new LinkedList<String>();
    List<String> linkL = new LinkedList<String>();

    for (Text tval : values) {
      String val = tval.toString();
      if (val.startsWith("w#")) {
        linkU.add(val.substring(2));
      } else if (val.startsWith("g#")) {
        linkL.add(val.substring(2));
      }
    }

    for (String u : linkU) {
      for (String l : linkL) {
        context.write(key, new Text(u + DELIMITER + l));
      }
    }
```

In the Mapper phase, we select the columns we need from weather and game data. In the Reducer phase, we join the two results by using the date as the key.

Although Hadoop supports various kind of data techniques, we also use Hive to parallel processing the input big data and integrate the useful weather attribute and baseball game record together same as what we've been done using MapReduce Java API mentioned earlier in this part.

Below are our major steps using Hive.

First, merged the baseball game record for certain team and create an external table named as bosLog. The figure about is the schema for bosLog.

```
+-------------+-----------+---------+--+
|  col_name   | data_type | comment |
+-------------+-----------+---------+--+
| date        | date      |         |
| weekend     | int       |         |
| promotion   | int       |         |
| tm          | string    |         |
| opp         | string    |         |
| wl          | string    |         |
| rank        | int       |         |
| gb          | string    |         |
| dorn        | string    |         |
| attendance  | int       |         |
| streak      | string    |         |
+-------------+-----------+---------+--+
```

Since the input data files include duplicate headings, we deleted these heading using the query below:

```
INSERT OVERWRITE TABLE bosLog
SELECT * FROM bosLog
WHERE tm not in ('Tm');
```

Then we create an external table to store weather record for the 5-year range in HDFS near the ballpark. Next, we use right outer join query to join the attributes we select from the weather record table with bosLog table. The result returns all the information for attributes mentioned above after running MapReduce to join two tables.

The table below shows the schema for the final CSV file which will be used to do further analysis.

Attribute Name	Description	Type	Range
Date	Date of the game	Date	
Mean Temperature	Day mean temperature in Fahrenheit	Integer	35-89
Mean Humidity	Day mean humidity	Integer	23-90
Mean Visibility	Day mean visibility in mile	Integer	4-10
Mean Wind Speed	Day mean wind speed in MPH	Integer	1-12
Events	Rain, Snow or Fog	Text	Max 10 chars
Home/Road	Blank: home game, @: road game	Text	Max 1 char
Opp	Opponent team	Text	Max 3 chars
W/L	Win or Loss	Text	Max 1 char
R	Run	Integer	0-16
RA	Run allowed	Integer	0-13
W-L	Team record	Text	Max 6 chars
Rank	Rank in division	Integer	1-5
GB	Games behind the leader in division	Float	1-12.5
D/N	Day or Night game	Text	Max 1 char
Streak	Recently win/loss record	Text	Max 10 chars
Promotion	Has a special event or not	Text	Max 1 char
Attendance	Attendance of that game	Integer	12192-48339

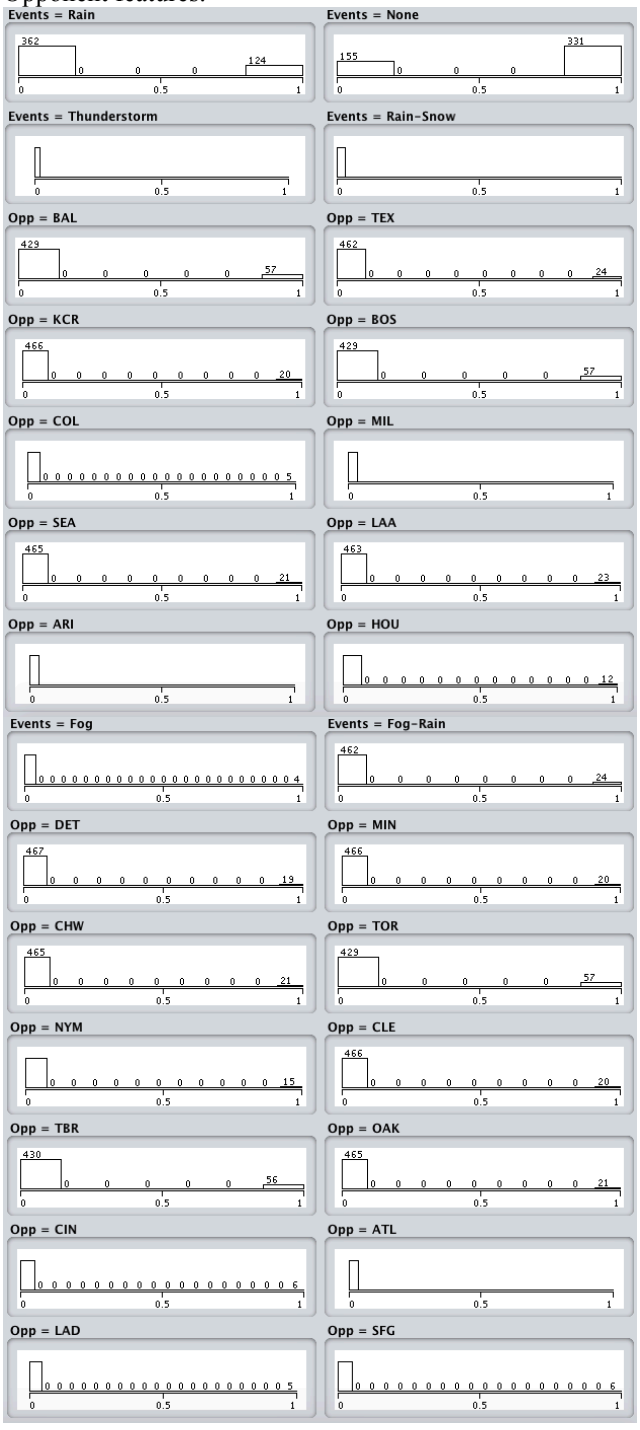
V. RESULT

In this section, we will present the experiment result of our own approach.

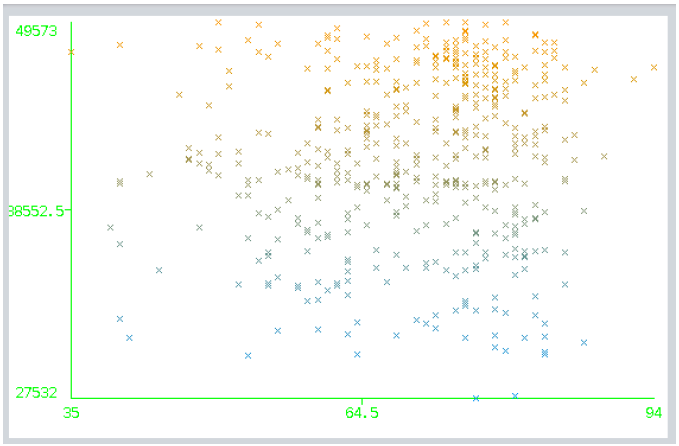
a. Data Visualization

We first visualize the data and try to find some insights into factors and attendance.

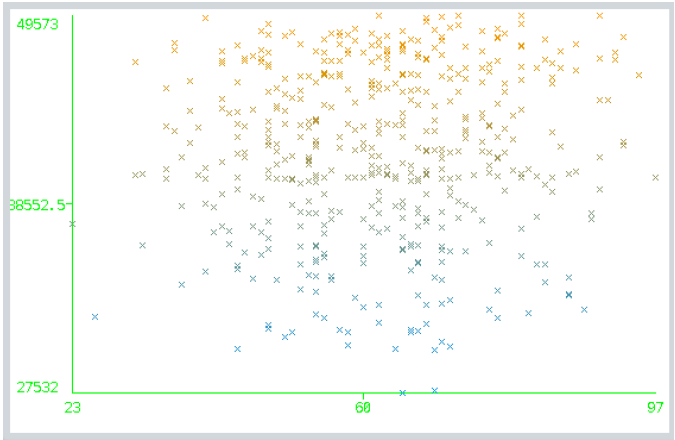
The below figure shows the number of data for Event and Opponent features.



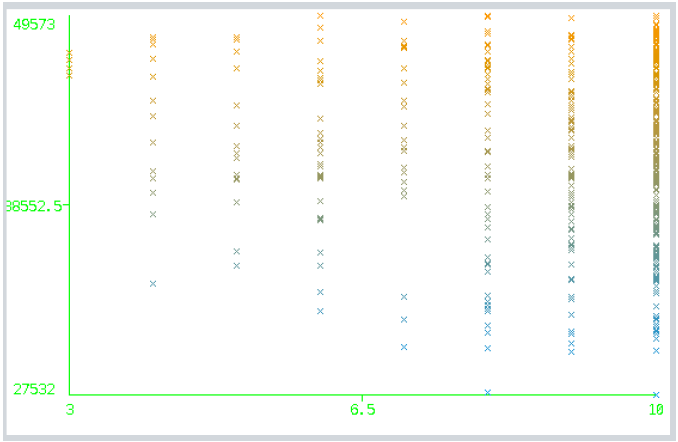
We try to find insights from this graph and since we want to see if weather factors play an important role in baseball game attendance. We visualize the relationship between baseball game attendance and weather indicator using the graphs below where we used important weather attributes, mean temperature, mean humidity, mean visibility, mean wind speed separately as X coordinate means the value of that attribute and Y coordinate means the number of attendance.



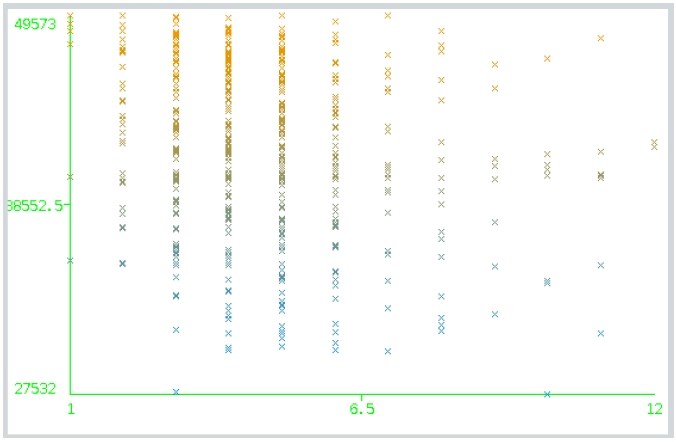
X: Mean Temperature, Y: Attendance



X: Mean Humidity, Y: Attendance



X: Mean Visibility in Miles, Y: Attendance

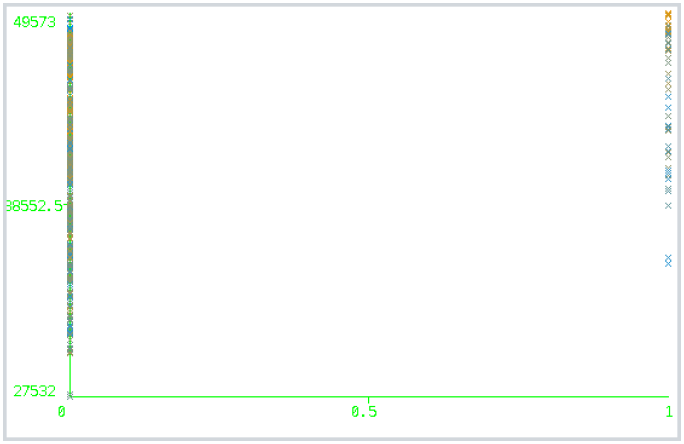


X: Mean Wind Speed in MPH, Y: Attendance

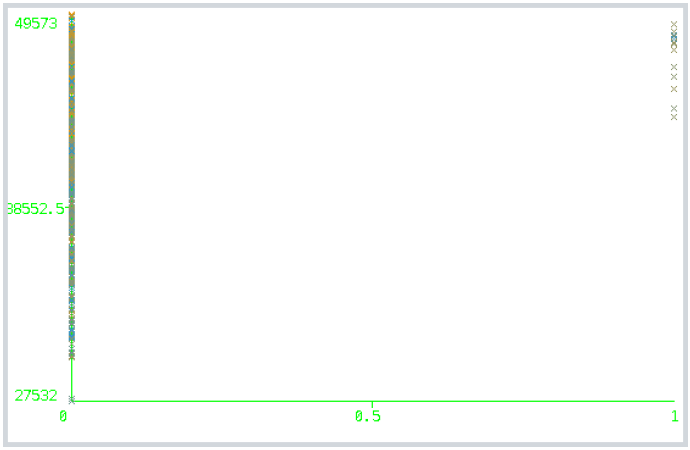
As we can see in the graph, the temperature has a positive correlation with attendance. One possible reason is that baseball season is from May to October. The season is either spring or summer. Besides, during the summer, many parents may bring their children to baseball games.

However, we cannot conclude the relationship between baseball games and other weather indicators such as humidity, visibility and wind speed from other graphs above.

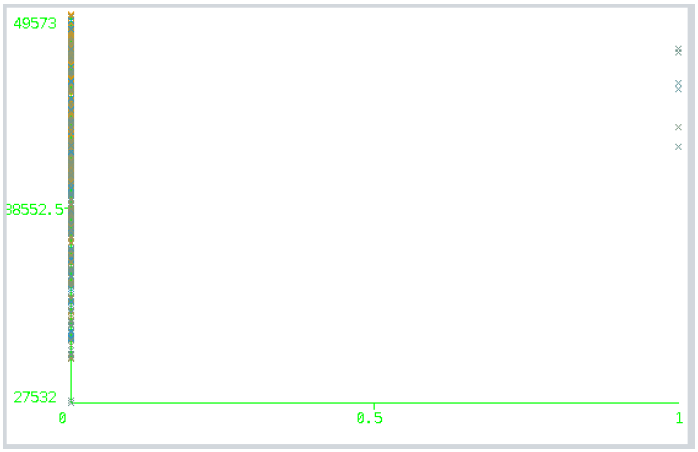
Another assumption we have is that certain opponents are related to attendance. As we can see in the statistic graphs below, opponents such as New York Mets, Boston Red Sox, and Cincinnati Reds all have more attendance than other teams.



NYY vs BOS (Y: Attendance)



NYY vs NYM (Y: Attendance)



NYY vs CIN (Y: Attendance)

For Boston Red Sox, we believe it is because the Yankees and Red Sox rivalry is so popular that it brings excitement for people.

For New York Mets, we think it is because they are from New York as well and have a lot of fans.

For Cincinnati Reds, we can take a close look at the data.

Date	Mean TemperatureF	Mean Humidity	Mean VisibilityMiles	Mean Wind SpeedMPH	Events	Weekend	Promotion	Gm#	Tm	Opp	W-L	Rank	GB	D/N	Attendance	Streak
2012-5-18	63	49	10	4		0	1	39	NYY	CIN	21-18	4	4.5	N	42015	+
2012-5-19	67	49	10	3		1	1	40	NYY	CIN	21-19	4	5.5	D	45302	-
2012-5-20	68	44	10	6		1	1	41	NYY	CIN	21-20	4	5.5	D	45622	--
2014-7-18	73	55	10	4		0	1	95	NYY	CIN	48-47	2	4.0	N	47372	+
2014-7-19	72	63	10	5		1	1	96	NYY	CIN	49-47	2	4.0	D	47606	++
2014-7-20	72	65	10	4		1	1	97	NYY	CIN	50-47	2	3.0	D	43115	+++

First, there were only six games during six years, this could have a bias. Second, all six games were hosted on Friday, Saturday and Sunday so it is reasonable that these games had more attendance than others.

b. Model Analysis

In this part, we will discuss our analytics model and the results we get using Spark and MLlib. [5]

First, we use ChiSqSelector for feature selection.

```
df_columns=["Gm#","MeanTemperatureF","MeanHumidity","MeanVisibilityMiles","MeanWindSpeedMPH","Weekend","Promotion","W","L","Rank","GB","EventsVec","OppVec","DayVec","StreakVec"]

assembler = VectorAssembler(inputCols = df_columns, outputCol = "features")

output = assembler.transform(final_df)

output = output.select("features", "Attendance")

selector = ChiSqSelector(numTopFeatures = 10, featuresCol = "features", outputCol = "selectedFeatures", labelCol = "Attendance")

result = selector.fit(output).transform(output)

model = selector.fit(output)
```

As the result, we know top 10 features are Gm#, MeanTemperatureF, MeanHumidity, MeanVisibilityMiles, MeanWindSpeedMPH, W, L, Rank, GB and Opp=LAD.

Then we use Gm#, MeanTemperatureF, MeanHumdity, MeanVisibilityMiles, MeanWindSpeedMPH, Weekend, D/N, Promotion, W, L, Rank and GB as variables to do regression analysis below.

We first build an Linear Regression to model the realtionship between attedendance and the variables we mentioned earlier.

```
lr = LinearRegression(labelCol = "Attendance")

model = lr.fit(trainingData)

predictions = model.transform(testData)
```

As the below results shows the R-squared value is 0.174886.

```
Linear Regression result:
Root Mean Squared Error (RMSE): 4719.71
Mean Squared Error (MSE): 2.22757e+07
Mean Absolute Error (MAE): 3833.15
R-squared: 0.174886
```

Then, we build a decision tree regression model to evaluate the relationshi between attendance and attributes mentioned above.

```
dt=DecisionTreeRegressor(featuresCol="indexedFeatures", labelCol="Attendance")

pipeline = Pipeline(stages=[featureIndexer, dt])

model = pipeline.fit(trainingData)
```

The R-squared value is 0.19305.

```
Decision Tree Regression result:
Root Mean Squared Error (RMSE): 4457.18
Mean Squared Error (MSE): 1.98664e+07
Mean Absolute Error (MAE): 3596.12
R-squared: 0.19305
```

Last, we create a random forest regression model to evaluate the relation.

```
rf=RandomForestRegressor(featuresCol="indexedFeatures", labelCol="Attendance")

pipeline = Pipeline(stages=[featureIndexer, rf])

model = pipeline.fit(trainingData)
```

The R-squared value is higher than the results returned by linear regression and decision tree regression.

```
Random Forest Regression result:
Root Mean Squared Error (RMSE): 4143.32
Mean Squared Error (MSE): 1.71671e+07
Mean Absolute Error (MAE): 3478.12
R-squared: 0.298574
```

Combing the results together, we can not conclude that weather condition, weekend or weekdays, and promotion will highly affect the baseball ticket sales. Further detailed data resources might help us to get more results and a more accurate result and find the true factors that affect attendance. More comprehensive records and more useful attributes will help us to achieve our goal.

VI. Future Work

The biggest problem of this project is that there are only 81 home games for every team, which means that the data is rare. To make our models more accurate, we need to get all the data for the rest of the MLB teams so that we can include every team and not just the Yankees. We also want to put more attributes in our training data which can help us to build a more accurate and complex model to predict the attendance numbers for the 2017 season. Although multiple linear regression is easy to understand and interpret, we believe a more sophisticated supervised learning technique will yield better results so we want to build more different models.

VII. Conclusion

We have developed an effective approach to identify the relationship between baseball game team attendance and other attributes based on most updated big data methods. We evaluated our assumption that weather condition will highly affect the baseball game attendance based on data sources from MLB and weather domain. The results of our experiments showed that there is no significant correlation between baseball game attendance and weather condition or other factors like promotion or weekend/weekdays. In the future，we plan to use our approach to verifying our assumption for all team in MLB with more attributes in our training data if given access

to a more comprehensive data source. We believe our approach will provide a good prediction for the 2017 baseball game attendance.

REFERENCES

[1] B. Evan, P. Turgoose, A. Redshaw, M. LaScola, and D. Howlett. "Big Data Is “Moneyball” on Steroids." February 2016.

[2] B. Lee, D. Koch and A.K. Panorska. "The Impact of Temperature on Major League Baseball". October 2013.

[3] T. White. Hadoop: The Definitive Guide. O’Reilly Media Inc., Sebastopol, CA, May 2012.

[4] A. Gates. Programming Pig. O’Reilly Media Inc., Sebastopol, CA, October 2011.

[5] H.Karau, A.Konwinski, P.Wendell and M.Zaharia. Learning Spark: Lightning-Fast Big Data Analysis. O’Reilly Media Inc., January 2015.