Playing Against the Elements: How Environmental Factors Influence Baseball Game Outcomes

Ryan Walter, Eliot Li, Yu Tao

Abstract

Baseball is a sport in which small changes in the game can result in great differences in the outcome and results of a game. Thus, the environment from weather, to the number of people in attendance, to how long the game is played affects the player and the game at a fundamental level. In this paper we examined how the environment and weather impacts the game statistics and observe trends in the data. Through the use of exploratory data analysis (EDA) and k-means clustering we found that there are underlying correlations between the environmental variables and resulting game statistics. Using this we created a neural network model that predicts the results of offensive, defensive, and important swing statistics with a high degree of accuracy. Using our results we demonstrate that weather plays a role in baseball outcomes and can affect the game in all aspects.

Keywords: baseball game performance, weather, environment, K-means clustering, neural network

Introduction

America's pastime has been an integral part of the culture and history of the nation. From the large stadiums under the lights to the back alleys in the crowded city, baseball enthralls people from all walks of life with its simple premise but high demand on skill. This world of baseball depends highly on numbers compared to other sports. Instead of fast paced consistent action that are found in sports such as basketball or soccer, baseball employs repeatable scenarios with slightly different conditions, which could be likened to an experiment in a lab. These small changes, the dependence upon balls and strikes, the fascination of how to improve ERA gave rise to one of the most statistically analyzed sports in history and brought in the "money ball" era where math plays as much of a role as gut feeling in decision making.

Statistics in baseball is not a novel concept, starting with Henry Chadwick with his initial statistics to convey to newspaper readers what happened in the game, statistics grew and evolved overtime (Schiff, 2008). Baseball statistics first came under public attention with the book and movie "Money Ball" in where the Oakland A's used statistics to select a new roster from undervalued players to create a winning team (Lewis, 2004). Since then, statistics in sports has increased tremendously as stakes become higher for both the participants and the viewers.

Although baseball has highly repeatable discrete events occurring in each game, not all variables can be controllable. Each game has its own unique circumstances from blistering heat, to blinding sunlight, to the heckling and boisterous crowds in a packed stadium. The environment for each game is unique and can affect the game in small but more pronounced ways. Long games in heat may lead to more balls instead of strikes, louder more engaged crowds may lead a player to become distracted causing an error, pouring rain may lead to lower hits and total runs in a game (Ashoff, 2018). Thus, understanding how the environment affects baseball and whether game statistics can be predicted could lead to a better understanding of the sport and a team's success in adversity.

Materials and Methods

Data Cleaning:

The data for baseball statistics were collected from an aggregated dataset from Retrosheets, a historical baseball statistic website (Retrosheets). The data was then filtered to select for all games from 2010-2015. Weather data was collected from NOAA for each city a baseball team plays in (National Centers for Environmental Information).

The data was then joined using the date and the city where the game took place. All missing data in the weather area related to precipitation was zero under the assumption that if that type of weather was not recorded it would be left NA by the measurement device and only recorded when it is present.

Exploratory Data Analysis:

The analysis focused on detecting the relationship between weather variables and game performances. A Single Variable Analysis is applied to weather data, from which the distribution of weather conditions, including precipitation, wind speed, maximum temperature, and minimum temperature was visualized for all matches. The heatmap was used to find the correlation between all variables. Besides the high correlation between game data and weather data themselves, the correlation between game data and weather data can be found as well. Box plot and violin plot were used to check the distribution of game performance and weather conditions between each other. One more day and night analysis was applied to take a brief look at if there exists a performance difference between day games and night games.

K-means Clustering:

The dataset was separated into two parts: the home team and visiting team. Each subset contains data on baseball performance measurements and weather records. K-means clustering serves as an exploratory stage where we explore the validity of our hypothesis and whether it is worth it for further investigation. We transformed the day/night variable into a binary indicator variable, and created a binary win variable indicating whether the visit team wins or not. K-means clustering algorithm is applied to both datasets with different numbers of clusters, ranging from 1 to 20. The total within-cluster sum of squares (WSS) is calculated for each number of clusters, and plotted against the number of clusters to determine the optimal number of clusters. Finally, the clustering results are visualized and a conclusion is drawn based on the optimal number of clusters.

Neural Network:

The data was then filtered to only include the game statistics chosen for prediction, the weather and environment statistics. Three different predictor variables were chosen to be used in the neural network. The first is penalized strikes, the total number of strikes in a game minus the total number of walks in game, was chosen to see how well a pitcher was performing while also taking into account the number of poor at bats there were for the pitcher. Next we examined the total number of hits in a game to see how well the offense in a game was. The final statistic we examined was the total number of errors which looked at whether the weather could predict how many mess ups occur in a game.

The neural network was created with 5 fold cross validation and the hyper parameters were chosen using a grid search (James, Witten, Hastie, & Tibshirani, 2021). The results were examined using root mean square error and MAE. The best model was then visualized.

Results

Exploratory Data Analysis:

Table 1 (left). **Weather data summary table**. The weather statistics for numerical variables, include AWND(wind speed), PRCP(precipitation), SNOW(snowfall), TMAX(maximum temperature), TMIN(minimum temperature), WDF2(direction of fastest 2-minute wind), WDF5(direction of fastest 5-second wind), WESD(water equivalent of snow), WSF2(fastest 2-minute wind speed), and WSF5(fastest 5-second wind speed)

Table2 (right). **Baseball game performance summary table**. The variables include bats, hits, homeruns, walks, RBI, walks, strikeouts, stolen bases, left on base, pitchers used, wild pitches, balks, and errors. Calculated the mean value and standard deviation for each variable.

	Mean	Standard Deviation
AWND	7.89	3.34
PRCP	0.10	0.32
SNOW	0.00	0.03
TMAX	78.37	13.64
TMIN	60.37	12.33
WDF2	204.16	97.80
WDF5	201.21	101.01
WESD	0.00	0.00
WSF2	18.64	6.20
WSF5	24.71	12.31

	Mean	Standard Deviation
bats	68.11	7.77
hits	17.30	5.07
homeruns	1.88	1.49
rbi	8.02	4.18
walks	6.06	2.89
strikeouts	14.93	4.27
stolen_bases	1.19	1.27
left_on_base	13.81	3.97
pitchers_used	7.93	2.32
wild_pitches	0.68	0.86
balks	0.06	0.26
errors	1.19	1.14

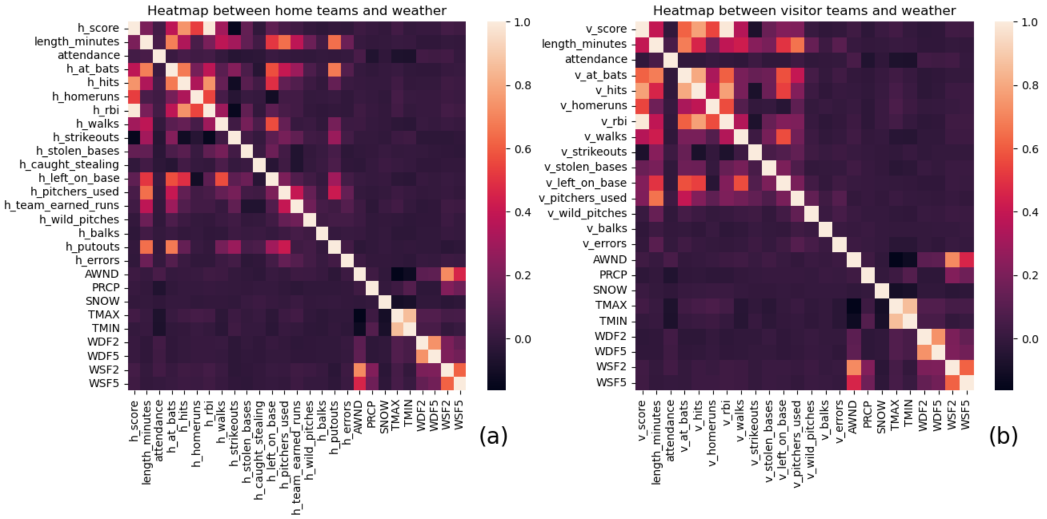

Figure 1. **Heatmap for all variables in the dataset.** A light color represents a high correlation between variables, and a dark color indicates a low correlation. The heatmap (a) on the left shows the correlations between home team statistics and weather variables and the heatmap (b) on the right shows the correlation between visitor team statistics and weather variables.

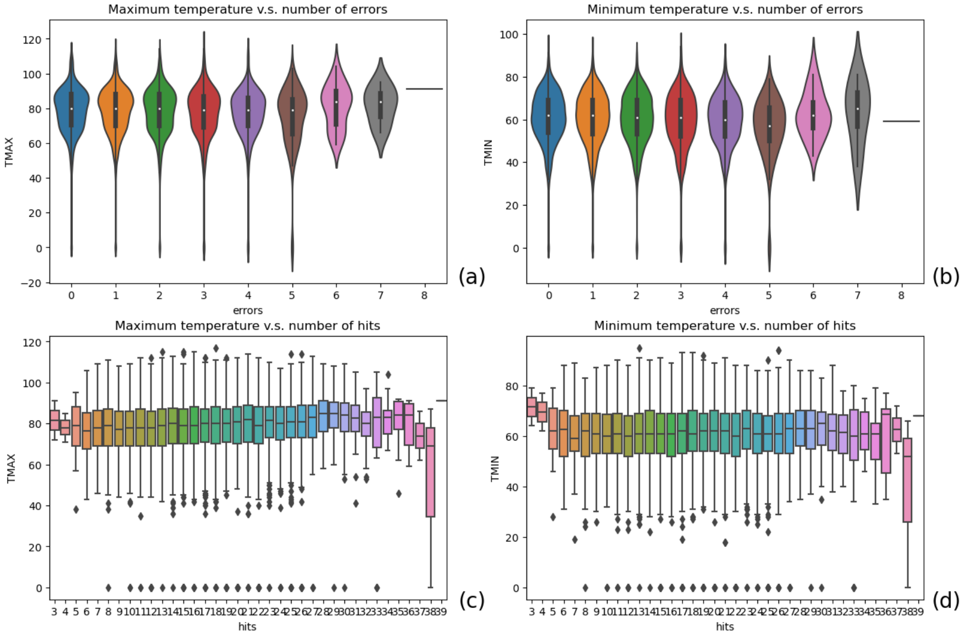

Figure 2. **Box plot and violin plot for visualizing the distribution between weather and game statistics.** Detailed distribution for the number of errors and the number of hits versus temperature are presented here. The violin plots (a) and (b) display the density, first quantile, median, and second quantile of temperature for a specific number of errors. The boxplots (c) and (d) show the outliers, first quantile, median, and second quantile for a given number of hits.

K-means Clustering:

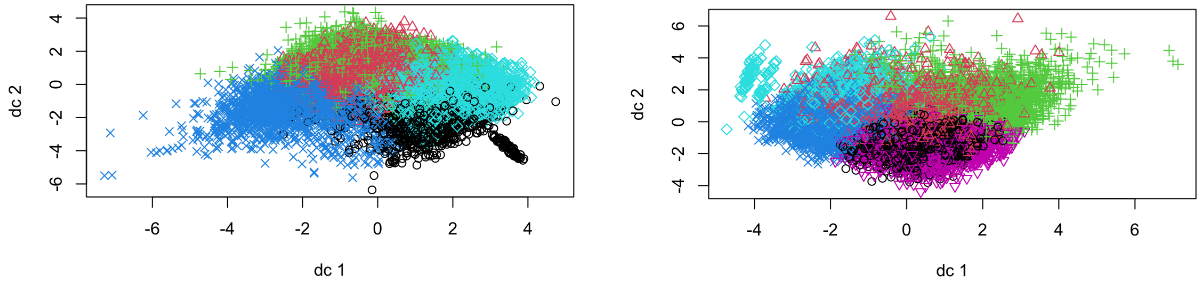

Figure 3. **K-means clustering**. Based on the clustering analysis conducted in this analysis, it was found that the optimal number of clusters for the home and visit team analysis was 6. This analysis could provide valuable insights into how weather conditions affect the performance of the home and visit teams in baseball games. Further investigation could be conducted to understand the characteristics of each cluster and how different weather conditions may affect the team's performance.

Neural Network:

The neural network model resulted in a low RMSE and MAE for all three target variables: penalized strikes, total number of hits, and total number of errors. Penalized strikeouts had the best performance with the lowest RMSE and MAE followed by Total Hits with the second best. Total Number of Errors had the worst model performance shown by the higher RMSE and MAE compared to the other variables (Table 3).

Table 3. **Evaluation of Model Performance**. Evaluation statistics of root mean square error and mean absolute error were applied to Penalized Strikeouts, Total number of hits, Total number of errors. Penalized strikeouts had the best results followed by hits then errors.

	Penalized Strikeouts	Total Number of Hits	Total Number of Errors
RMSE	0.1063396	0.1181028	0.1403144
MAE	0.08397531	0.093066	0.1098034

The results of the neural network were plotted and showed a positive correlation for the target variables as seen in figure 4 (Figure 4).

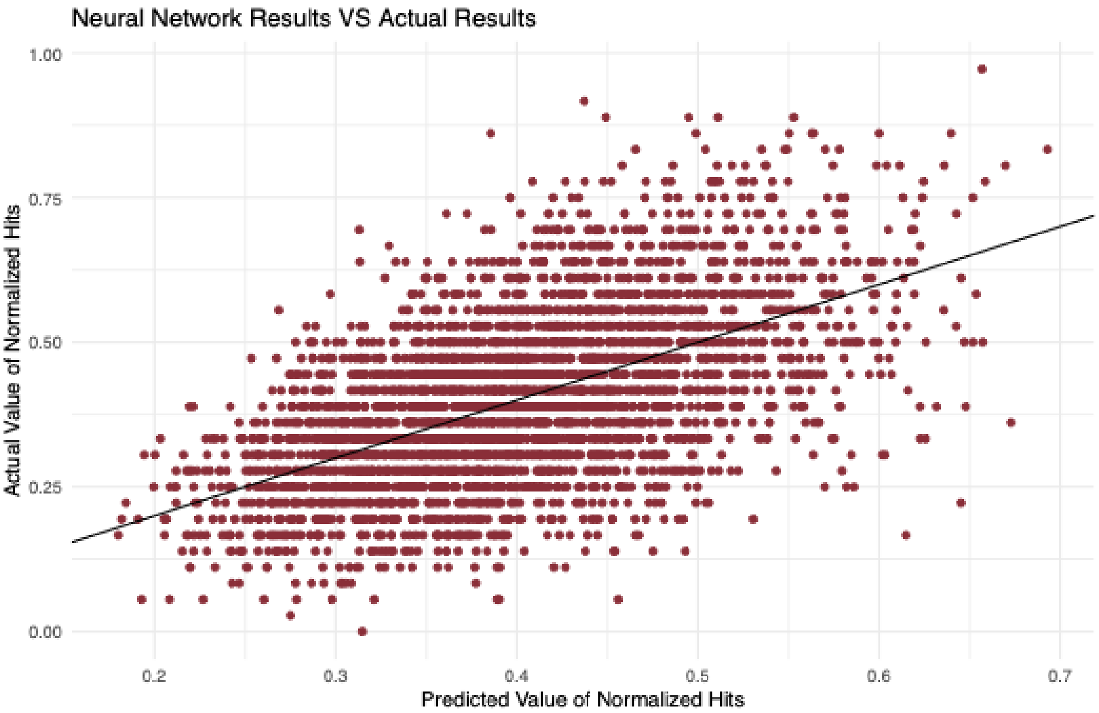

Figure 4. **Predicted Number of Hits vs. Actual Hits**. Total number of normalized hits were predicted using the neural network then plotted. The trend line represents perfect prediction. The points form a positive correlation and follow a trend similar to the line.

Discussion

Exploratory Data Analysis:

Based on preliminary statistical analysis of the weather data (Table 1) and game data (Table 2), it was found that the average wind speed was 7.89 meters per second, the average maximum temperature was 78.37 Fahrenheit, and the average minimum temperature was at 60.37 Fahrenheit. Among all the game statistics, the highest standard deviation was observed in the total number of bats, which reached 7.77. While, the lowest standard deviation was observed in the total number of balks, which was 0.26.

It is clear to see the high correlations between game statistics and game statistics, and weather variables and weather variables from Figure 1. As for the relationship between gaming performances and weathers, most of them are at a lower level. But it is still worth noting that hits, home runs, RBI, strikeouts, and pitchers have some correlation with the maximum temperature and the wind speed.

Figure 2 (a) and (b) indicates that the number of errors approaches 7 when the median maximum temperature is over 80, and the median minimum temperature is over 60. When the total number of hits is ranged from 6 to 30, it has a slightly positive correlation with the median of temperature, as we can see from Figure 2 (c). The relationship between minimum temperature and number of hits is not detected.

K-means Clustering:

The groupings of the data into clusters in this analysis suggest that there may be certain weather combinations that lead to certain results in the game statistics (Figure 3). This observation could potentially indicate a cause-effect relationship between weather and game outcomes. For example, a cluster of games may have higher wind speeds and lower temperatures which lead to lower offensive outputs from both the home and visiting teams.

However, it's important to note that this is merely speculation, and a more in-depth analysis would be required to validate these assumptions and to establish a causal relationship between weather and game outcomes. Furthermore, other factors such as team strategies, player performance, and injuries could also impact game outcomes.

Neural Network:

The RMSE and the MAE were low for all three target variables (Table 3). This suggests that weather is able to predict game stats to a high degree of accuracy. This high accuracy of the model may be due to the nature of the environment and weather data. Many of the weather variables are indicators of certain weather types and thus some are more infrequent then others. Thus, when a certain weather event occurs it highly correlates with certain results in pitching, hitting, or fielding. Additionally, the data between games is very similar thus small differences are magnified leading to it being more accurately predicted by the model.

Additionally, the evaluation of the model shows that errors were the least accurate. This was contrary to our initial hypothesis that errors were highly affected by weather and the data. This may be due to errors being a limited event in baseball that occurs infrequently thus there is not enough data relating to errors for it to be predicted.

Conclusions

Although errors were not as highly correlated to weather and the environment as compared to hits and pitching, we were able to show that there is a relationship between all areas of baseball and the surrounding game environment. Our results show that there are underlying relationships that cause groupings in the data which can be used to predict the results of offensive and defensive statistics for the game. Future research needs to be conducted to examine the effects the environment has on the game of baseball through the use of more accurate environmental data collection specific for the game itself. Overall baseball is not only affected by the players participating in the game but also the wider world the game takes place in.

References

Ashoff, T. (2018). *Understanding the Relationship Between Weather Conditions and Home Run Rates in the MLB.* [PDF]. Retrieved from https://dspace.mit.edu/bitstream/handle/1721.1/120271/1083220855-MIT.pdf?sequence=1&isAllowed=y

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning an introduction to statistical learning: With applications in R* (2nd ed.). New York, NY: Springer.

Lewis, M. (2004). *Moneyball: The art of winning an unfair game*. New York, NY: WW Norton.

National Centers for Environmental Information. (n.d.). *Access.* https://www.ncei.noaa.gov/access.

Retrosheets. (1996). *Data Downloads.* https://www.retrosheet.org/#.

Schiff, A. J. (2008). *"The father of baseball": A biography of Henry Chadwick*. Jefferson, NC: McFarland & Co..