Predicting Major League Baseball Games using Machine Learning

Daniel W. Sager

University at Buffalo

School of Engineering and Applied Sciences

CSE474/574 Introduction to Machine Learning: Final Report

Source Code: https://github.com/DanSager/PredictMLB

1. Abstract

Among popular sports in America, baseball tends to have one of the closest relations to the world of statistics. As a result, determining the victor of a baseball game before the game is even started is not only possible but can be accurate to a certain extent. One such way to analyze this is by constructing a machine-learning model that can more accurately predict the outcome of a Major League Baseball game than the average baseball fan. As a control sample, three average baseball fans have made their predictions for the entire 2019 Arizona Diamondbacks season. Supervised learning across five different machine-learning classifiers has also been used in an attempt to achieve the highest accuracy possible. Based on the results, it is clear that even with this simple model, machines possess a higher level of accuracy than the control sample and have the potential to reach even higher degrees of accuracy.

2. Introduction

According to an article published in 2020 by SportsShow, Baseball is the second most popular sport in America based on TV Ratings, falling short only to Football [1]. Baseball, also known as America's pastime, has been greatly influenced in recent years by the increased use of statistics to increase the chance of winning. This idea can be best shown by nothing other than the Oakland Athletics 2002 season where general manager Billy Beane used statistics to a much greater degree than ever before to assemble a competitive team, a story that was popularized by the novel and movie titled Moneyball [2].

In contrast to America's other most popular sports leagues, such as the National Basketball Association or the National Hockey League that each has 82 games per team per season, each team in the Major League Baseball organization competes in 162 games before playoffs. This not only provides a great amount of data, but it allows for a large number of future testing opportunities as the dataset has many opportunities to update during the season, allowing for predictions to constantly be influenced.

Therefore, through the construction of a machine-learning model that can more accurately predict the outcome of a Major League Baseball game than the average baseball fan, further analysis can be conducted on the predictability of baseball game outcomes. Training data will be

greatly limited to pure statistics on previous games because of the very large amount of potential influences. As a result, a problem such as this would be impossible to achieve perfect accuracy; however, being able to slowly increase accuracy as the sample size increases would prove to be a successful experiment.

3. Methodology and Data

The single most important prerequisite to this experiment is reliable and accurate data on past MLB games.

3.1 Dataset Acquisition

To begin, securing datasets that not only provided enough data but also provided desired stats while omitting unrelated information was achieved by creating a custom web scraper for baseball-reference.com [3-4]. This requires using the urllib and Beautiful Soup libraries for python to obtain data on every game played in the MLB each season. Data was then stored using sqlite3 into individual databases for each season, ranging from 2012 to 2019. Before the 2012 season, the Miami Marlins were known as the Florida Marlins and wanting to avoid any further issues, data has been limited to seasons since the name change.

```
def extract_data():
    """
    Extract data from the baseball-reference.com reguarding team stats in addition to their pitchers.
    """
    teams = ['ARI', 'ATL', 'BAL', 'BOS', 'CHC', 'CHW', 'CIN', 'CLE', 'COL', 'DET', 'HOU', 'KCR', 'LAA', 'LAD', 'MIA',
             'MIL', 'MIN', 'NYM', 'NYY', 'OAK', 'PHI', 'PIT', 'SDP', 'SEA', 'SFG', 'STL', 'TBR', 'TEX', 'TOR', 'WSN']
    years = ['2012', '2013', '2014', '2015', '2016', '2017', '2018', '2019']
```

Fig 1. Teams and years in dataset.

Table: ARISchedule

	num	date	team	opponent	home	runs	runsallowed	innings	day	pitcher	opp_pitcher	win
	Fi...	Filter	...	Filter	Fi...	F...	Filter	Fil...	...	Filter	Filter	Fil...
1	1	2019-03-28	ARI	LAD	0	5	12	9	1	Zack Greinke	Hyun-Jin Ryu	0
2	2	2019-03-29	ARI	LAD	0	5	4	13	0	Matt Andriese	Yimi Garcia	1
3	3	2019-03-30	ARI	LAD	0	5	18	9	0	Zack Godley	Kenta Maeda	0
4	4	2019-03-31	ARI	LAD	0	7	8	9	1	Yoshihisa Hira...	Dylan Floro	0
5	5	2019-04-01	ARI	SDP	0	10	3	9	0	Merrill Kelly	Matt Strahm	1

Fig 2. First five games in Arizona 2019 schedule.

In team sports, home-field advantage refers to the perceived advantage a team has when playing at their home field [5]. While this phenomenon is not always true, there is often a noticeable statistical difference in the team's win percentage based on where they are playing. Therefore, recognizing this information while making predictions is worthwhile.

Table: WinLossSplit

	team	overall	home	away
	Filter	Filter	Filter	Filter
1	ARI	0.525	0.543	0.506
2	ATL	0.599	0.617	0.58
3	BAL	0.333	0.309	0.358
4	BOS	0.519	0.469	0.568
5	CHC	0.519	0.63	0.407

Fig 3. The first five teams 2019 Win-Loss split for overall, home, and away.

3.2 Classifier Training and Feature Selection

To secure the highest prediction accuracy possible, it is vital to choose the classifier that will consistently perform the greatest. To carry this out, a total of five different machine-learning classifiers have been chosen from the scikit-learn and XGBoost Python libraries. These include Logistic Regression (LR), Support Vector Classifier (SVC), KNeighbors Classifier (KNC), Random Forest Classifier (RFC), and XGBoost (XGB). Because each of these classifier's time to train and time to predict tends to become rather large as the dataset increases, eventually

exceeding the acceptable execution time, each model is only trained before the season is simulated and never updated as the season progresses. In a measure to ensure that higher accuracy could be achieved by updating after every game, a second Logistic Regression classifier, labeled 'Logistic Regression Previous', is trained using a smaller training dataset that is then updated and retrained after every game is played. Lastly, a final overall census is made by taking the average prediction of the first five classifiers.

Although originally, providing as much data to the classifiers seemed like a good idea, it became obvious as time went on that they were beginning to over train the classifiers and did not provide adequate results. To avoid this, included features were instead hand-selected from the available data to achieve the highest accuracy. Of the originally desired features, the number of innings, runs, runs allowed, pitcher WLP (win/loss percentage), pitcher ERA (earned run average), opponent pitcher WLP, and opponent pitcher ERA have been removed. The features that remain include season, team, opponent, home or away, day or night, pitcher WHIP (walks plus hits per inning pitched), pitcher FIP (fielding independent pitching), opponent pitcher WHIP, opponent pitcher FIP, team's win-loss percentage based on location and opponent's win-loss percentage based on location. When testing yearly pitching stats have been replaced with the previous year's stats, because each respective season had not been theoretically played yet.

3.3 Control Group

The control group includes data from three average baseball fans attempting to predict the 2019 Arizona Diamondback's season. To obtain this data, a python script iterates through every game in the season, presenting the user with data that is known about the game before it is played. This includes team, opponent, home or away, pitchers and their stats, and win lost split based on the location of the game being played.

4. Results Analysis

To begin, let us have a look at how the classifiers stacked up against each other. To do so, they have each been used to stimulate 2017, 2018 and 2019 MLB seasons for every team in the league. At the end of each season, the predictions were compared against the actual results of the games predicted, their totals have been added up and their results are shown below.

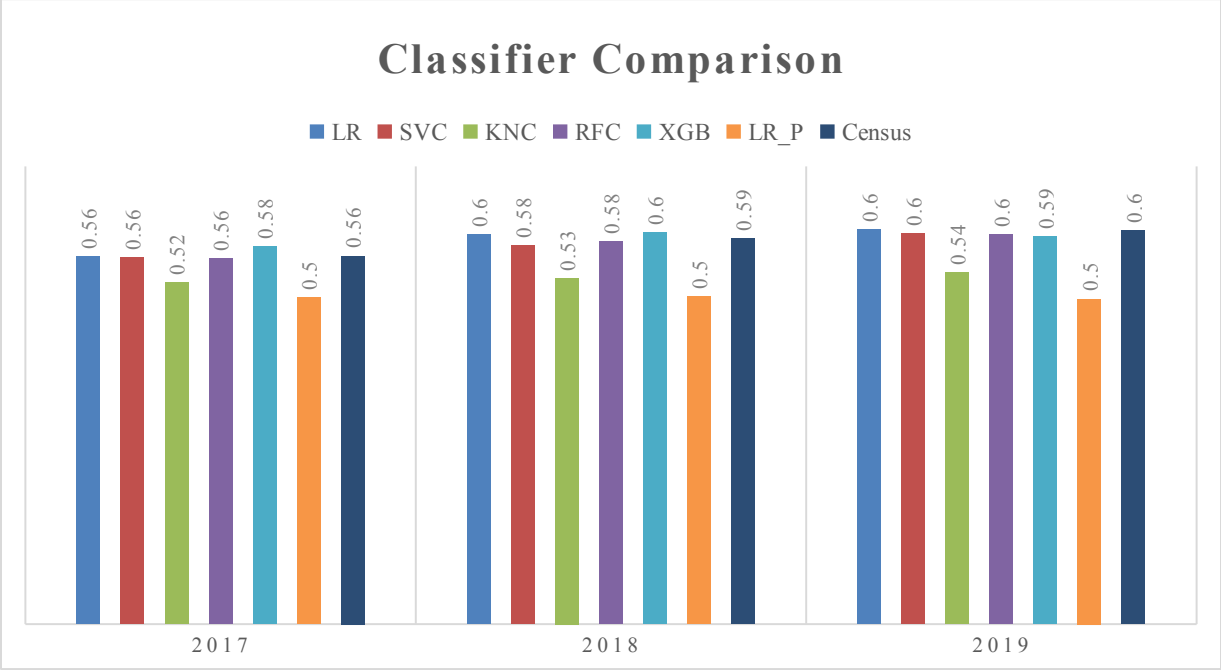

Fig 4. Comparing each classifier used against one another.

Consequently, Logistic Regression is consistently outperforming every other classifier. That is, of course, except for the overall census, who is either leading or ever so slightly trailing behind Logistic Regression. It is clear that more testing is required to make a more conclusive decision on which is superior, however, based on the data we do have; the census appears to be the safer bet as it relies on multiple classifiers. Therefore, it is less susceptible to an outlier throwing a curveball. The accuracy of the two classifiers over the 2019 season can be seen in the following graph.

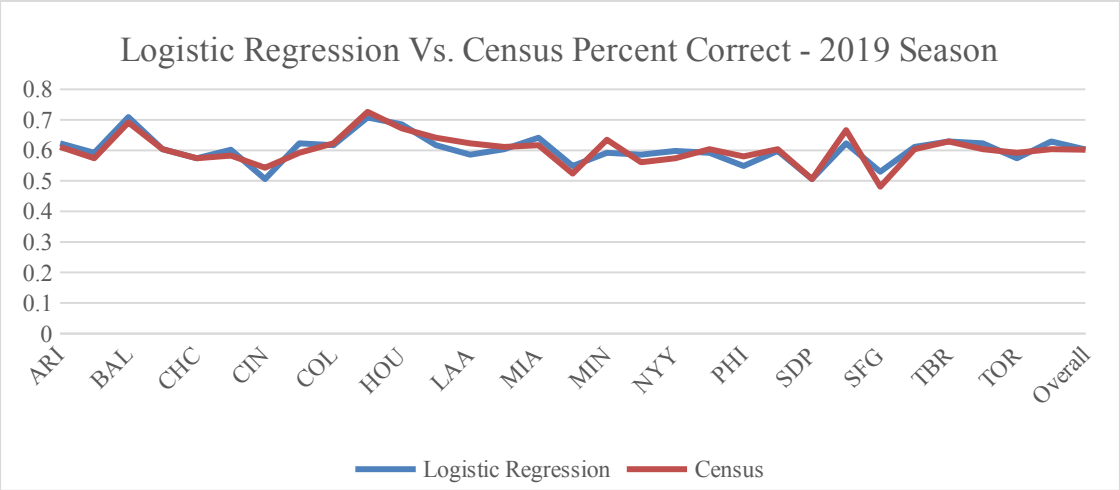

Fig 5. Logistic Regression versus the overall Census across every teams 2019 season.

Lastly, to see how a control group of human participants would compare against various machine-learning classifiers, a final test was established. Each would attempt to predict the 2019 Arizona Diamondbacks season provided only data that would be available before the game has begun. Based on the findings, Logistic Regression, and the overall Census, the two samples we are taking into serious consideration, both outperformed all three of the human control group members by predicting the correct victor more often.

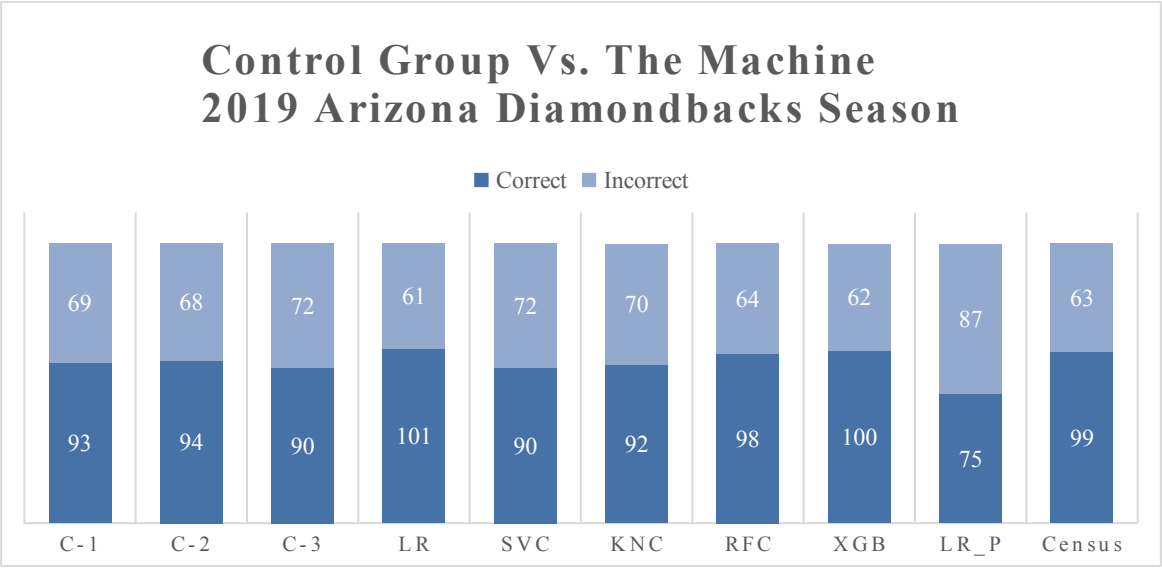

Fig 6. Three control-group (C) participates versus various machine-learning classifiers.

5. Conclusion

One of the most impressive things about modern technology is the scope of its abilities. Through this experiment, we have seen that even a moderately trained machine-learning algorithm can more accurately predict the outcomes of baseball games than life long fans. To the machine, our dataset is just a set of random numbers but can still make very impression assumptions about them. Without a doubt, a more fine-tuned classifier is capable of producing results of even higher accuracy. However, what has been achieved here is nothing short of the original goal, which is outperforming average baseball fans in predicting the outcome of Major League Baseball games.

6. References

[1] Das, Sourav. "Top 10 Most Popular Sports in America 2020 (TV Ratings)." *Sportsshow.net*, 21 Jan. 2020, sportsshow.net/most-popular-sports-in-america/.

[2] Grier, Kevin, and Tyler Cowen. "The Economics of Moneyball." *Grantland*, 9 Dec. 2011, grantland.com/features/the-economics-moneyball/.

[3] "2018 New York Mets Schedule." *Baseball*, www.baseball-reference.com/teams/NYM/2018-schedule-scores.shtml.

[4] "Noah Syndergaard Stats." *Baseball*, www.baseball-reference.com/players/s/syndeno01.shtml.

[5] "Home Field Advantage Definition - Sporting Charts." *SportingCharts.com*, www.sportingcharts.com/dictionary/mlb/home-field-advantage.aspx.

[6] llSourcell. "LlSourcell/Predicting_Winning_Teams." *GitHub*, 23 Aug. 2017, github.com/llSourcell/Predicting_Winning_Teams.