Predicting Playoff Teams for the MLB

Jose Fuentas, Loyola Marymount University, USA
Denali Tonn, Loyola Marymount University, USA

Abstract
The researchers utilized 47 years of team statistics to determine whether a team, amongst the 30 in the league, qualify for the playoffs. Using offensive and defensive features, the researchers attempted to analyze the Moneyball theory and determine which models were the most effective in categorizing the goal at hand. After applying and visualizing the data with 4 different models, they were able to predict with an accuracy rate of between 85%-90% and determine the impactfulness of each feature.

CCS Concepts: Machine Learning, Logistic Regression, Support Vector Machines, Random Forest

1 INTRODUCTION

Within the realm of predictive analytics, sports emerge as a captivating domain where uncovering intricate patterns and relationships within voluminous datasets defies traditional human-formed conclusions. Sports demand an array of decisions from participants, sparking a proliferation of applications ranging from predictive models to machine learning algorithms, databases, and nuanced scoring systems. Nowhere is this more apparent than in the realm of baseball, a sport steeped in strategy and precision. Baseball players showcase unique approaches to the game characterized by the various pitches, batting approaches, and strategic decisions that hold significant statistical weight.

2 THE MONEYBALL THEORY

The 2002 Oakland Athletics pioneered a transformative strategy to baseball after utilizing data analytics to construct a playoff-contending team. Faced with the loss of their three best players due to free agency, the organization created the Moneyball theory as a strategic framework in identifying and acquiring undervalued players. The Moneyball theory focused on two key metrics: slugging and on-base percentage, in appraising the value and performance of a player. The subsequent $41 million roster of questionably skilled players projected the Athletics to be one of the worst teams in baseball for years to come. However, the team achieved remarkable success from 2000 to 2006, averaging 95 wins, securing 4 American West titles, and making 5 playoff appearances. Following Oakland's success, a rapid utilization of data analytics pioneered how baseball organizations viewed strategy, team development, and player acquisition in correspondence to the Moneyball theory. Using insights from the results of the Moneyball theory, the researchers plan to determine what features are the most decisive factors in determining a playoff contending team.

3 Methodology

To show the statistical importance of offensive and defensive features, a series of machine learning models are devised to predict a team's ability to make the playoffs. Feature extraction is performed on historical team data to evaluate the impact of offensive and defensive metrics. A supervised learning algorithm able to predict more than eighty percent of playoff contending teams should indicate strong indications to what features are determining features for playoff contention. By applying a diverse machine learning model and algorithms to the extracted features, one can assess their accuracy and correlation and apply these principles to evaluating a team's projected performance.

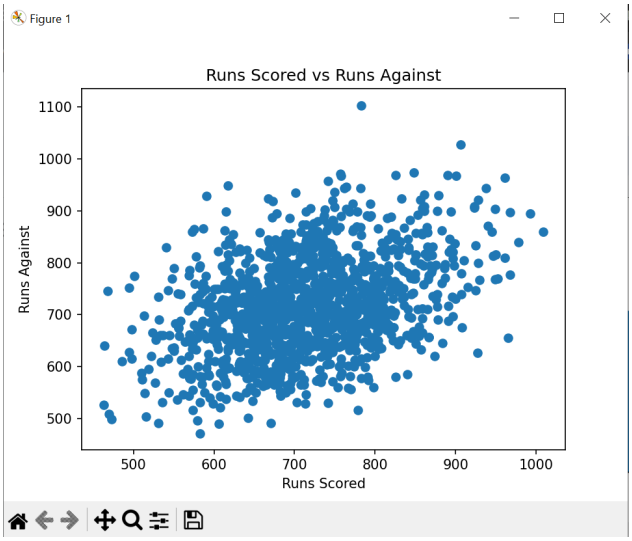
3.1 Data

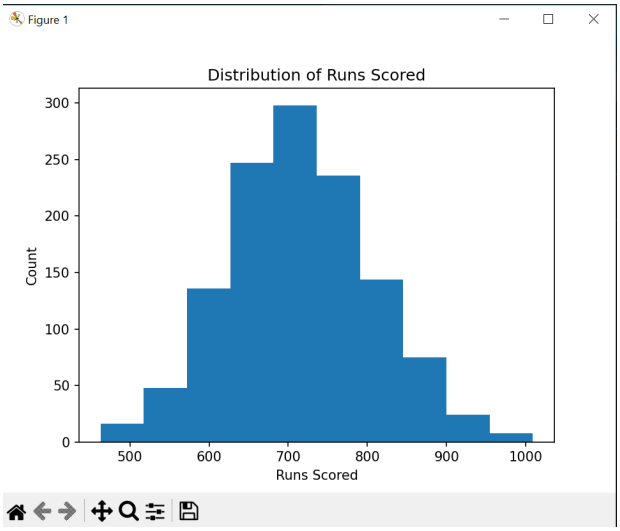
This project utilized a labeled dataset sourced by MIT using information provided by the widely referenced baseball-reference.com. This site is often referenced by major media organizations and contains sabermetric data for every player in the MLB. The reliability of the platform can be traced back to its original development by Sean Forman as part of his Ph.D. dissertation in applied math and computational science at the University of Iowa. Originally built as a web interface for the widely used Lahman Baseball database, the site has become a prominent resource in the baseball community[1]. The CSV file used in this project was curated by MIT professor Dr. Dimitris Bertsimas in 2017 as a component of his lecture on how data analytics can be used to predict a baseball world series champion[2].

3.2 Visualization

Prior to the implementation of hyperparameter tuning, the researchers visualized different components of the dataset's features. Runs scored and runs against were hypothesized to be the most impactful features due to their inherent correlation to winning games. The scatter plot labeled 'Runs scored vs Runs against' displays a tightly packed variation of points and the lack of outliers suggest a relatively balanced playing field. The distribution of points indicate that higher scoring teams tend to have lower runs against, suggesting that factors such as fielding and pitcher are important attributes in the outcome of winning games.

The graph labeled 'Distribution of runs scored' displays a normal distribution. The majority of the data is centralized around the mean indicating that teams often score runs in close proximity to the mean. In the projection of these features, it is not common to have outliers and displays a balance between the offensive and defensive tendencies of baseball.





3.3 Modeling

Using the following feature set, the following models will be tried:

- Logistic Regression
- Random Forest
- Support Vector Machine (SVM)
- K-nearest Neighbors

With these different models, if there is an inherent pattern that can be observed, these models should be able to find it. In addition to just using the default values given to the researchers by the Sklearn library, the researchers also hypertuned each of the models above using GridSearchCV, which helped in trying out different hyperparameters and also being able to grab the best result out of the different hypertuning. The specifics of the hyperparameters for each of the models will be described in more detail in the Parameters Section 4.1 below.

4. OUTCOME

From the modeling behaviors observed after training, the researchers were able to determine some valuable information from feature analysis. Before even trying to hyper-tune the models, the researchers found that the accuracy score for each of the models was already pretty high, with the lowest model, Random Forest receiving an 85% accuracy while the highest model, Logistic Regression, got a 90% accuracy. With this in mind, the researchers decided to find the weights or coefficients of each of the features for each of the models.

Analyzing the coefficients of the Logistic Regression Model, it was observed that the only impactful feature was ‘Runs Scored’ and indicated the obvious inclination that winning more games is associated with scoring more runs. The coefficients of the SVM model provide more insight, ranking the features in order of importance: slugging, on-base percentage, and run scored. A large emphasis on SLG and OBP heavily supported the Moneyball theory and showed that these attributes were the main factors in scoring runs.

The usage of Ablative analysis, the concept of removing one feature at a time to see impactfulness, revealed that the two most impactful metrics for the random forest model were ‘Runs Against’ and ‘Runs Scored.’ Using this information, Run’s observed had the highest importance while the least important metrics were Opponents' Slugging and Opponents' On-Base Percentage. This issued the importance of having more defensive features and statistics since baseball is both an offense and defense game. Such features would be important in determining the best way to limit the offensive capabilities of opposing teams.

4.1 Parameters

4.1.1 *Logistic Regression Classifier*

In logistic regression, the hyperparameter C plays a pivotal role in striking a balance between the smoothness of the decision boundary and the accurate classification of training points. A smaller C encourages a smoother, more generalizable boundary, prioritizing overall pattern capture and preventing overfitting by tolerating some misclassifications. Conversely, larger C values allow for a more intricate decision boundary that precisely fits the training data, risking

overfitting by placing higher emphasis on classifying individual points correctly. Our C values ranged from as small as 0.01 all the way up to 10000.

4.1.2 *Support Vector Machine Classifier*

In SVM, similar to Logistic regression it also uses the hyperparameter C which again, serves as the regularization parameter, akin to its role in logistic regression. Additionally, SVM introduces the choice of kernel type, with "Linear" seeking linear patterns and "Rbf" accommodating non-linear patterns, providing greater flexibility in capturing intricate relationships within the data and further trying to find that balance.

4.1.3 *Random Forest*

Random Forest encompasses two crucial hyperparameters, namely N-estimators and maximum depth. The former denotes the number of trees in the forest, with higher values generally improving performance at the cost of efficiency and which the chosen values were between 50-200. The latter, max depth, determines the maximum depth of each tree, influencing the complexity of relationships captured. For the dataset in use, where excessive depth may not be necessary, controlling this hyperparameter becomes crucial to balance computational efficiency with model performance and the chosen range for this was between 0-20.

4.1.4 *K-nearest Neighbors*

K Nearest Neighbors involves the hyperparameter N-neighbors, dictating the number of neighbors considered when making predictions. This parameter influences the granularity of the model's decision boundaries, with smaller values capturing local variations and larger values resulting in smoother, more generalized predictions. Another hyperparameter was the choice of the weight function where the researcher decided on two options. “Uniform" which assigned equal weight to all neighbors regardless of how far they were and "Distance" giving more weight to closer neighbors, emphasizing the significance of nearby instances in prediction.

4.2 Metrics

In pursuit of forecasting playoff outcomes for baseball teams, the researchers employed a comprehensive multi-metric evaluation approach. This approach considered key metrics such as accuracy, precision, recall, and F1 score. Accuracy, as a fundamental metric, provides an overall assessment of the model's performance in successfully labeling playoff and non-playoff teams. Precision was utilized to assess the reliability of the model is playoff classification to ensure a low rate of false positives. Concurrently, recall is of paramount importance, evaluating the model's ability to capture all teams that actually made the playoffs while minimizing false negatives. The F1 score, serving as a harmonized blend of precision and recall, strikes a balance between these metrics. It offers a comprehensive performance evaluation that accounts for both false positives and false negatives. This diverse set of metrics allowed a nuanced understanding of the model's strengths and weaknesses in various aspects, facilitating a robust assessment of its efficiency in predicting playoff outcomes for baseball teams.

4.3 Evaluation

In summary, the model exhibits notable strength in accurately identifying teams destined for the playoffs. It showcases a high precision that attests to its reliability when making positive predictions. This is particularly advantageous in situations where minimizing false positives is paramount, as the consequences of misclassifying a non-playoff team as playoff-bound can be significant. However, the model's performance reveals a potential limitation in terms of recall, suggesting that it may overlook a notable proportion of actual playoff-bound teams. This trade-off between precision and recall highlights the inherent challenge of achieving a balance between minimizing false positives and false negatives. The F1 score, serving as a comprehensive metric, encapsulates this balance, providing a nuanced evaluation of the model's overall effectiveness.

The model was very efficient in accurately predicting non-contending teams and was slightly less efficient for contending teams. Considering the objective at hand, this yielded better results than utilizing random chance which would have yielded a 50% accuracy score in predicting play-off contending teams. The researchers believe the models are a step in the right direction in attempting to predict which teams will or won't make the playoffs.

4.3.1 *Data Drawback.* In addition to possible problems with the models, the data could also have an issue. The labeled dataset contains 47 years worth of team statistics that may not account for statistical outliers as rules and regulations changed the game. Baseball analytics was not recorded as thoroughly until the institutionalization of sabermetrics in the 1980s and the robust Statcast system in 2015. One inherent flaw that was observed within the dataset was that before the year of 1999, OOBP and OSLG, were not recorded. As a result, the researchers used a SimpleImputer from scikit-learn to fill in the missing values in the dataset to give the models more complete information.

KNN Classification Report:

	precision	recall	f1-score	support
0	0.89	0.95	0.92	242
1	0.76	0.59	0.67	66
accuracy			0.87	308
macro avg	0.83	0.77	0.79	308
weighted avg	0.87	0.87	0.87	308

Classification Report for K-nearest neighbors shows our precision pretty high while our recall and f-1 score drops, reflecting a full picture

5. ETHICAL CONCLUSIONS

An ethical consideration in predicting playoff teams is the potential impact on sports betting and gambling. Accurate predictions may attract individuals engaging in betting activities, raising concerns about the ethical implications of influencing gambling decisions. The use of predictive models in the context of sports betting could inadvertently contribute to problematic gambling behaviors, posing risks of addiction and financial harm to individuals. Ethical practices necessitate responsible communication about the limitations and uncertainties of predictions, discouraging the undue reliance on model outputs for gambling purposes. Ensuring transparency about the predictive model's purpose and promoting awareness of the potential risks associated with sports betting are essential components of ethical decision-making in the realm of sports analytics.

6. CONCLUSIONS

With the culmination of work done on the importance of offensive and defensive features in considering the play-off potential of a team, the researchers were able to determine that the Moneyball theory corroborated with the conclusions drawn from our dataset. The results indicated that the offense of a team in respect to the features of slugging and on-base-percentage were the most valuable features in evaluating the play-off contention of a team. Furthermore, visualizing the data from each model revealed limitations of the models and potential improvements that can be made through the implementation of in-depth defensive attributes and statistics.

6.1 Future Work

The researchers believe with additional information of individual players statistics and the incorporation of a larger amount of defensive statistics such as pitching and fielding would greatly improve accuracy and valuable insight. The results derived from the dataset coincided with the revelations the Oakland Athletics experienced through the Moneyball theory and correlated with the style of baseball of that era. The current era of baseball centralizes around the offensive principle of power hitters evaluated by the slugging percentile of its hitters. The researchers believe that the Moneyball theory profoundly impacted the style of baseball from there on and believe diving deeper in the history of baseball may reveal profound results. Although data is limited from the pre-steroid era (i.e. before 1993), the researchers want to observe the variation in feature importance when small-ball style of play was more prevalent and determine its impact on play-off contention. Despite these concerns, the most important implementation is defensive attributes due to its direct impact on the offensive capabilities of an opposing team. The researchers believe the usage of pitcher and fielding statistics are very influential in determining the significance of defense in baseball and its potential to be a determining factor for a play-off contending team.

6.2 Applications

The applicability of the researcher’s project can be utilized in any sport with clean and reliable data. Current MLB teams have their own databases and utilize data analytics in strategy and the development of players within their organizations. Further work in implementing larger amounts of data and features can reveal important information that can give an edge to a team.

Developing an in-depth understanding of the importance of features and data gathering can allow teams to trade or sign players in their journey to create a play-off contending team.

6.3 Lessons Learned

The researchers originally desired to pick the most successful model in predicting the play-off potential of a team. The following tips were surmised from development.

- Data
    - The researchers' data revealed a need for more defensive attributes and the dataset was not satisfactory in considering the importance of defensive features.
        - The lack of data before 1999 for OOBP and OSLG was an indicator that the project was lacking defensive information to make a fair evaluation of the importance of defense in play-off contending teams
- Implementation of in-depth defensive features
    - The researchers recognize that the defensive features of their dataset was the opponent's offensive statistics. This does the not provide the insight to how MLB teams can limit the offensive capabilities of opponents.
    - Implementation of various pitching and fielding statistics to the model will allow researchers to determine the impact of the defense in baseball and its impact in predicting playoff contending teams.

7. ACKNOWLEDGEMENTS

Thanks Professor Korpusik for helping us and guiding us in our direction for this final project. The researchers would also like to thank sklearn, seaborn, and the pandas library for helping us visualize and capture the data.

8. REFERENCES

[1] Baseball-Reference.com. https://en.wikipedia.org/wiki/Baseball-Reference.com

[2] MIT OpenCourseWare. 2017. Logistic Regression. https://ocw.mit.edu/courses/15-071-the-analytics-edge-spring-2017/pages/logistic-regression/assignment-3/predicting-the-baseball-world-series-champion/

[3] University of Wisconsin-Extension. Moneyball Proves the Importance of Big Data and Big Ideas. https://uwex.wisconsin.edu/stories-news/moneyball-proves-importance-big-data-big-ideas/