Big Data Analytics Final Report
Application of Machine Learning in Baseball

Leo Lam
Statistics Department
Columbia University
lkl2129@columbia.edu

Wanting Cui
Statistics Department
Columbia University
wc2619@columbia.edu

Yuhan Zha
Statistics Department
Columbia University
yz3284@columbia.edu

Abstract - **Machine learning is widely used in baseball prediction. This study aims to construct a classification model for the prediction of award-winning players in order to reveal some potential hidden future baseball stars from a large pool of players. In addition, this study creates a career peak prediction model for the team managers to apply during player selection process in order to predict whether the players have passed their career peak. Furthermore, the study proposes a salary prediction model for the players to evaluate their current contracts on whether they are being underpaid. Lastly, the study performs unsupervised machine learning techniques in categorizing different pitchers. All models result in promising and accurate performances. Detail results can be referred to Experiment Results Section.**

I. INTRODUCTION

Baseball is the third most popular sport in the United State with more than 500 million fans globally [1,2]. The first official baseball game in the United State can be traced back to 1845, and since then, the popularity of baseball has been increasing and remained as one of the most likeable sports in the US [1]. The high popularity of baseball allows the Major League Baseball (MLB) to generate more than $10 billion in the US alone last year with an average MLB player salary at $4.47 million [3,7]. According to a statistical report, there are more than 15 million professional and amateur baseball players in American in 2017 [4]. Besides America, baseball is also extremely fashionable in Japan, Taiwan, U.K., Canada, Cuba, and many other countries [5]. The high global attention of baseball allows MLB to be the highest season attendance of any sports league in the world reaching over 73 million online and television audiences in 2015 [6]. As a result, considering the high renown, fever, and income of becoming a professional baseball player, there are many passionate talents trying to join the MLB teams every year, which results in competitive player selection procedures, especially in the well-known professional teams.

Award Winner Prediction

Many talented baseball players with strong potentials are not recruited into the professional teams due to the flaws in the current player selection system. A typical path to become one of the very limited players in the teams of Minor League Baseball and Major League Baseball is to join the baseball camps during high school or college, and during the times in the camps, potential candidates often have to expose themselves in social media in order to spread their names out into the baseball field to catch attentions of professional team recruiters [8]. An immediately disadvantage of this system is as some passionate players may not have the privilege and opportunity to join baseball camps at such a young age, their chances of being a professional player can be significantly reduced. Furthermore, even if they get into the typical routine of being a professional player, they may not be a socially active person to catch enough attentions from the team managers. This study provides a prediction model to determine whether a player will become an award-winning candidate. The motivation behinds this goal is that as the current player selection procedures consider so many factors that are not baseball-skill-related, this study aims to construct an evaluation system for the professional teams such that they can recruit potential players purely based on their baseball techniques and talents using the model proposed in this study. To achieve this goal, a machine learning classification model is constructed for the prediction on whether a player will be an award winning player.

Salary and Career Peak Prediction

Many excellent and potential players may be being underpaid by their current contract. While the famous baseball stars may have their own agent in monitoring their salary and contract, the ordinary professional players may

not have this luxury. As a result, these less fortunate players may not know how much do their skills worth, and they can be signing their next contract at a much lower price as compared to other players at the same skill level. In the opposite, some players may have already passed their peak of their baseball career but still having an unreasonably expensive contract such that the teams are over paying the players who can no longer continue to highly contribute the team [9]. This may prevent a leading team from continuing its glory as well as may prevent other good players to receive the payment that they deserve. Therefore, this paper also concentrates in the construction of an evaluation method in the determination of the salary of the players purely based on their performance as well as an evaluation method in the judgement of whether the player has already passed his career peak. These methods can allow a team manager in allocating the right amount of budget in hiring potential players that are currently underpaid and before their career peak while reducing the the waste of resources in overpaying the players who have already past their career peak.

Pitcher Category Clustering Analysis

Excessive baseball data in the modern system can cause misleading and subjective interpretation, which reduces the analytic clarity and and accuracy of the analysis. There are a increasing number of baseball datasets becoming available online allowing baseball participants to analyze. However, different analysts may hold different subjective perspectives in the use of the data, which results in numerous different views on the evaluations of players [10]. As for pitcher types alone, there exist many ways in categorizing different pitchers. Some would suggest that there are three different pitcher types, and some may argue that there are only eight while some may propose that there exist 13 [11, 12, 13]. Different analysts may use different categories to evaluate pitchers causing different evaluation results. The massed analytics methods can easily cause confusions in selecting the optimal type of pitchers that matches the interest of the team. Therefore, this study aims to apply unsupervised clustering method in the categorization of the various types of pitchers in order to provide an objective and statistically based pitcher category, which can generalize the evaluation of the pitchers. K-means clustering is performed in order to reveal the similarity among different pitchers so that those with high similarity can be grouped into a single pitcher category.

Web Application

Data visualization is efficient in delivering complex statistical analytics and machine learning algorithms to audiences who do not have a strong background in statistics or data science. Despite the high accuracy and comprehensive inference of many machine learning models, they are often very complicated ideas and would require some knowledge of statistics and data science to fully understand the logic behind them. As the target audiences of this study are baseball participants who tend to have limited knowledge in statistics and data science, such as team managers and baseball players, it is essential to ensure the delivery method of this study is accessible and understandable for them. Therefore, a web application with concise and modern design is built. For the predictive model, users only need to input the player statistics of a player onto the web app, and the app will return the prediction result. The web app also consists some interactive tables on baseball player and team statistics, which allow users to filter and sort the information. Besides building a web app, this study also applies data visualization in exploratory data analytics on other baseball related information, which is also displayed on the web application.

The machine learning models proposed in this study would only require player statistics to predict players' salary, career peak, and whether the player is a potential award winner. As the current baseball player evaluation methods consider many factors that are not skill-related, this study aims to provide an alternative approach that concentrates solely on performances and skillsets of the players in a hope to provide a more effective and transparent methods in player selection and salary determination process.

II. RELATED WORKS

Award Winner Prediction

There are very limited resources on the prediction of award winning baseball players online. Instead, there exist researches in the prediction of potential Hall of Fame winners in baseball. However, as Hall of Fame is extremely selective and cannot be a good reflection of the general population of baseball players. Some of these studies also use the number of award that players receive to predict whether he will win Hall of Fame. However, in this case, the award related variables are highly correlated with Hall of Fame, the response variable. In order to propose an approach that is purely concentrated in the baseball

performance of the players, this study proposes an alternative model that uses only player statistics and background information for predictors and the number of awards that a player wins as the response variable. Furthermore, feature engineering is performed to eliminate meaningless variables so that the model can concentrate in the important variables. The related works can be found in Reference [16, 17].

Salary and Career Peak Prediction

Theoretically, teams prefer to give large contracts to players before their career peak to gain more wins and save budget. However, in the real baseball field, it is common that team managers sign a player after the player achieves his career peak, which implies the player is in a declining stage. Therefore, predicting whether the player is still improving is important for the team manager. Inspired by a research to predict player decline [18], this paper achieved a high accuracy using machine learning models to predict the career peak of baseball hitters. On the other hand, in order to save budget, team managers need to know whether a player is underpaid or overpaid when they draw up a contract. Even though there exist multiple researches to predict salaries of major league baseball players, it is uncommon to evaluate whether players are undervalued or overvalued in the incoming contract. To solve this problem, this paper combined the salary prediction and value evaluation.

III. SYSTEM OVERVIEW

The findings of this study are presented in a web application method. The web app consists five major parts: interactive tables of the data, EDA findings from k-means and PCA, predictive model in baseball player salary, predictive model in baseball player career peak, and predictive model in award baseball player. To achieve the goals of this study, various datasets are used.

Datasets

Lahmen's Baseball Database is the primary data source of this study. This database contains pitching, hitting, and fielding statistics for Major League Baseball from 1871 through 2017 with 28 datasets. It includes data from the two current leagues (American and National), the four other "major" leagues (American Association, Union Association, Players League, and Federal League), and the National Association of 1871-1875. The major tables that are used in this project are MASTER, Batting, Pitching, Teams, Appearances, AwardsPlayers and Salaries. Master table contains player names, date of birth and biographical information. Batting is the batting statistics with 22 variables and pitching is pitching statistics with 30 variables. Teams contains each team's yearly statistics and its standing. It also indicates which division and league the team belongs to. Appearances shows details on the positions a player appeared at based on year and team. AwardsPlayers records awards won by players and the corresponding year. And salaries contains each player salary data of each year. All the tables can be joined together through playerID or teamID.

In addition to Lahmen's Baseball Database, Statcast is also used in this study. Statcast is a state-of-the-art tracking technology that allows for the collection and analysis of a massive amount of baseball data [15]. The data are created through radar and camera systems create Statcast. It provides detail information on each pitch, bat, and defense such as detail statistics on every pitch by pitchers with their opposing pitchers and batters, pitch type, release and perceived velocity, launch angles, spin rate, pitch, and received coordinates. This system was installed in 2015 and it recorded over 2.1 million pitches within 3 years. This information is used in pitcher type clustering and pca analysis.

The overall size of the datasets that have been used in this project, is around 950MB. The combination of these datasets consists comprehensive and detail information of the majority of the professional baseball players.

This study consists three parts: award winner prediction, salary and career peak prediction, and EDA along with clustering analysis. The major presentation method of the findings is by web application.

Award Winner Prediction

A three-class classification machine learning model is constructed using multinomial logistic regression and random forest. There exist 23 predictors on player statistics and background information. The response variable is total award, which consists class 0 - players without award, class 1 - players with some awards but not Hall of Fame, and class 2 - players with many awards or Hall of Fame winner.

Salary and Career Peak Prediction

A two-class classification machine learning model is constructed for the career peak prediction to determine whether the player is before (class 0) or after (class 1) his career peak using logistic regression and XGBoost. This model mainly based on the OPS of the player, which is widely considered one of the more important evaluation for hitters. The basic information such as age, weight, height, and games statistics are also included in the model to increase accuracy.

A regression machine learning model is constructed for the average annual salary prediction to determine whether an MLB player is underpaid or overpaid for their incoming MLB contract using random forest. This model correlates a player's game information and WAR statistics before signing their active contract and the annual salary of the contract they signed. The prediction of average annual salary is less than the true value indicates this player is undervalued as he supposed to gain more money. Otherwise, the prediction of annual salary is more than what the player received, the player is defined as overpaid, as he got more than what he supposed to be.

Pitcher Category Clustering Analysis

K-means clustering and PCA are used to analyze different pitchers and generalize pitchers into different pitching categories. PCA is used to further analyze numerous pitcher related variables in order to reveal the characteristics of different pitchers in different clusters from K-mean clustering. Additional EDA graphs are plotted for a better understanding of the teams and players in both American League and National League.

Web Application

Web application is constructed for interactive data visualization and accessible information delivery method purposes. Google cloud is used to publish the web application online. However, a limitation of such publishing method is google cloud instance must be remaining connected for the online linkage. An alternative method to run the application is to run it on a local machine. Detail instruction can be referred to the GitHub page of this study.

IV. ALGORITHMS

Award Winner Prediction

The baseball database consists massive detail information on the players. Therefore, only datasets with information on award, appearance, and people are used. The original data are present in an annual format such that it consists the performance and award information of the player by years. To match the interest of this study, the sum of the player statistics (number of pitches, hits, runs, etc.) variables over years are calculated. After computation of the sum of the player appearance information (how often do they serve in different position on the field), the relative frequencies of these variables are as well constructed. The merged and processed dataset contains 94 variables before feature selection and each row consists information of a unique player. Within these 94 variables, many of them are meaningless and redundant. Hence, some of these variables are removed. Then, random forest is applied for the selection of important variables. The final model consists 23 predictors. The model uses total award that each player receives as a response variable. Out of the various awards, Hall of Fame is often considered to be the most influential award and the highest honor that a baseball player can achieve. Therefore, this study aims to classify three types of players: normal players who do not receive any award, excellent players who receive some award, and outstanding players who receive many awards and are potential candidates of Hall of Fame. The model uses players who debut from 1920 to 2000 as there are significantly fewer awards before 1920 while some players who debut after 2000 still have not received nomination of many award, especially Hall of Fame. This study uses two different test datasets. The first test dataset is the prediction of award winner player that are in the same period as the training dataset (1920 - 2000). The second test dataset is the prediction of award winner player that debut in between 2000 and 2017. A major challenge of this part of the study is as there are only about 14% of the players receiving award with only 7% of them receiving more than one awards, the dataset is imbalanced. Therefore, the training data is oversampled such that each of the three classes can contain the same number of data. The final dataset consists 16500 rows of training data.

Multinomial logistic regression and random forest are used to perform three-class classification on players who do not receive any award (total award = 0), players who receive some awards (0 < total award < 21), and players who receive many awards and are Hall of Fame candidate (total award > 20). To give a heavier weight of the players who receive Hall of Fame, an addition of 20 awards will be rewarded to Hall of Fame winners. Compare to random

forest, multinomial logistic regression can provide statistical inference of the variables as well as the relation between each predictor to the response with its relative significance. However, as a tradeoff, the predictive power of multinomial logistic regression tends to be lower than that of random forest. In addition to the high predictive power, random forest can also provide a list of important variables and is shown to be more robust toward imbalanced data, which is the case of this problem. While multinomial logistic regression does not have any tuning parameter, random forest consists three major tuning parameters: number of trees, maximum depth of each tree, and number of variables used in each tree. Cross validation is performed in order to reveal the optimal parameters.

Salary and Career Peak Prediction

The Batting data is used to predict the career peak of players. This prediction is mainly based on On-base Plus Slugging (OPS), which represents two important skills for the player to get on base and to hit [19]. BA, OBP, and SLG are calculated by using hits (H), base on ball (BB), times hit by pitch (HBP), at bats (AB), sacrifice flies (SF), and total bases (TB). And OPS is calculated by the sum of On-base Percentage (OBP) and Slugging Average (SLG). The dataset is grouped by player and team, a cumulative game played is calculated and the OPS is calculated each year among their career life. If the latest season OPS is less than the highest OPS before, the player is defined as after his peak. Otherwise, if the latest season OPS is greater than highest OPS before, the player is defined as before his peak as he may achieve a higher OPS in the feature. The Batting dataset is merged with People dataset to take a consideration on player's salaries. Since salaries could increase each year due to inflation, this paper adjusted inflation rate to keep all salaries in 2018 value.

Logistic regression and XGBoost models are applied to predict whether the player is after (1) or before (0) their career peak. The dataset is imbalance due to the number of after and before are unequal, therefore a oversampling techniques is used to change the training data into balance. Both of logistic regression model and XGBoost are able to dealing with binary classification task.

A new salary dataset from USA TODAY includes the salary and years of an active contract for MLB players. Combining with players WAR and game statistics from BASEBALL REFERENCE, the real average annual salary of the contract is calculated, and the active year is extracted to make prediction on the average annual salary. If the prediction is less than the true average annual salary on the contract, the player is overvalued. Otherwise, if the prediction is greater than the true average annual salary, the player is undervalued.

Regression models are applied to make prediction on average annual salary. The pearson correlation and feature importance are checked to select features. Random forest regression model provides feature importance to do feature selection. All features less than 0.05 importance are removed from the model. In the end, 'WAR' ,'WAR_rep', 'IPouts', 'IPouts_start', 'waa_win_perc_rep', 'BIP', 'pyth_exponent_rep', and 'xRA_def_pitcher' are selected from 49 variables. In addition to random forest, a XGBoost model also used to make prediction.

Pitcher Category Clustering Analysis

The statcast database contains all the consists all the pitches thrown from 2015 to 2017 seasons, including games from both American League and National League. Grouping by year and name, there are around 750 around unique pithers each year. Players with the same name are renamed into name+sequential numbers for differential purposes. For example, there are 2 pitchers named 'Chris Smith', and these pitchers are renamed into 'Chris Smith1' and 'Chris Smith2' correspondingly. After some preliminary EDA, pitchers with less than 10 inning pitched per season are deleted to eliminate noises, since sometimes fielders may fill in pitching position for strategic purposes. Based on variable 'pitch_type', A new dataset 'Pitched Type' is created based on the proportions of different types of balls each thrown each year. The datatset contains 13 pitch type: four-seam fastball (FF), two-seam fastball (FT), cutter (FC), sinker (FS), slider (SL), changeup (CH), curveball (CU), knuckle-curve (KC), knuckleball (KN), eephus (EP) and pitch out (PO).

Due to high correlations of variables in the original statcast dataset, PCA is performed before clustering. Later k-mean clustering is performed in an attempt to cluster pitchers based on the their performances. Elbow method is used to determine to the optimal amount of clusters. Similarly, k-mean clustering is also performed on the 'Pitched Type' dataset, aiming to differentiate players based on their pitching styles.

V. SOFTWARE PACKAGE DESCRIPTION

Python

Python, a programming language, is used for data preprocessing, reformatting, model training, and data visualization. Python consists various type of modules and packages including sklearn for machine learning related applications, numpy for multi-dimensional arrays and matrices applications, pandas for data manipulation and analysis applications, matplotlib for data visualization and plotting applications, pybaseball for baseball related data extraction, flask for web application construction and linkage between python models and html interface, and dask for building analytical web application with interactive data table.

HTML & CSS

HTML and CSS are used to create web application template layout and optimize application design.

Google Cloud Platform

Google Cloud Platform, a cloud computing service, is used to run the model code from python in cloud linking to the external terminal, which allows publishing of the web application.

VI. EXPERIMENT RESULTS

EDA & Pitcher Category Clustering Analysis

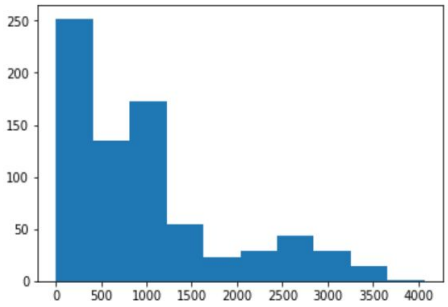

Figure 1 - Histogram of Pitch Throw in 2017:
As shown in the histogram, the bins between 0 and 1000 are significantly higher than the bins beyond 1000. This implies pitches at 1000 may be a cutoff number of experienced and non-experienced pitchers. To investigate deeper into the hidden information from the histogram, Figure 2 is plotted.

Justin Verlander	4065
Chris Sale	3605
Gio Gonzalez	3514
Rick Porcello	3465
Jose Quintana	3423
Chris Archer	3406
Jon Lester	3375
Luis Severino	3365
Kevin Gausman	3357
Zack Greinke	3326

Figure 2 - Table of Top 10 Pitch Throw Pitchers in 2017:
The table consists the pitchers with the highest pitch throw in 2017. As hypothesized in Figure 1, the pitchers with higher number of pitches are the famous and experienced professional players. For instance, Justin Verlander, pitcher with the highest pitch throw, is the fourth highest paid MLB player in 2017 with $28,000,000 salary[14].

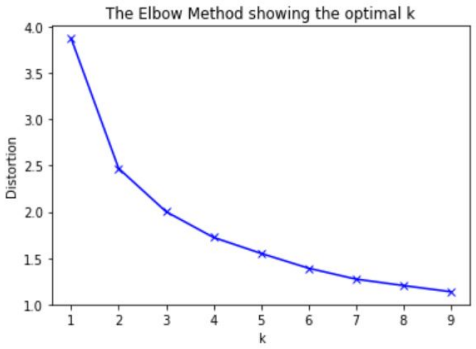

Figure 3 - Plot of Elbow Method on K-Means Clustering on Pitcher Statistics:
K-means clustering is performed to reveal the hidden correlation among various of pitchers using pitcher statistics. A parameter that is required for k-means clustering is the number of clusters, which elbow method can often provide a clear answer. According to Figure 3, two clusters should be used.

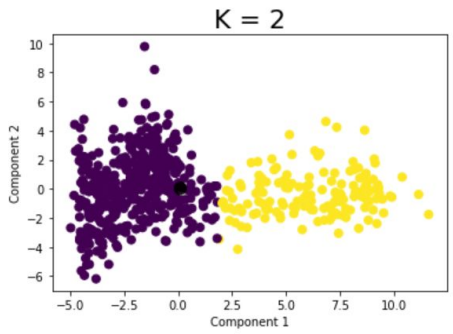

Figure 4 - Plot of K-Means Clustering Plot on Pitcher Statistics:
This clustering plot is constructed based on the result of PCA such that x-axis represents first principal component while y-axis represents second principal component. As shown in the plot, the two clusters are nicely separated into two clusters. The yellow cluster does not show much variation in y-axis. This suggests much of the variation of the yellow cluster can be explained by first component. In contrast, the purple cluster shows noticeable variation in both x-axis and y-axis, which indicates this cluster depends on both first and second components. To further investigate the result of the yellow cluster, the major variables that are represented by the first principal component are analyzed in Figure 5.

Label	BF	AB	Pit	IP	H	R	W	L	ER	GS
0	190.294248	169.026549	749.462389	44.149779	42.267699	22.816372	2.199115	2.336283	21.183628	2.221239
1	636.128378	572.378378	2473.033784	149.133784	146.979730	77.020270	9.439189	8.932432	71.378378	25.628378

Figure 5 - Mean Values of the Important Variables in the First Principal Component:
Based on the mean differences of the variables between the two clusters, GS has the most significant difference. GS represents “Game Started”, which is an important variable in distinguishing between starting pitcher and relief pitcher. Starting pitcher is often considered to be the most important pitchers among the team who would start the game and participate in at least four innings. Hence, cluster 0 would represent the category of relief pitchers while cluster 1 would represent the category of starting pitchers. To further maximize the value of the cluster analysis, the composition of the two clusters in both National and American League are investigated in Figure 6 and 7.

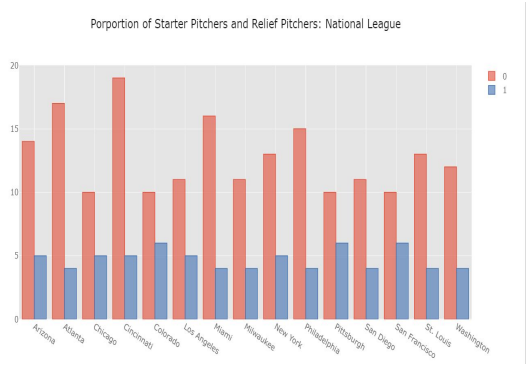

Figure 6 - Bar Chart of Proportion of Starter and Relief Pitchers in National League

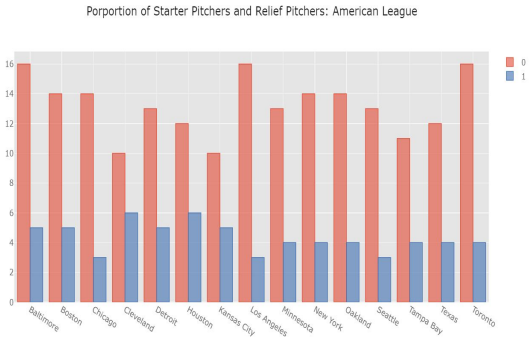

Figure 7 - Bar Chart of Proportion of Starter and Relief Pitchers in American League

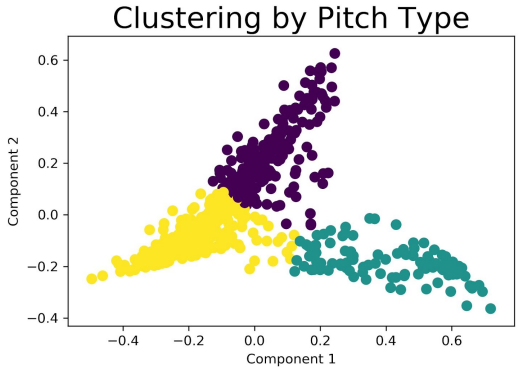

Figure 8 - Cluster Plot of Pitcher Category by Using Pitch Type Data:
Cluster analysis is performed to generalize different types of pitchers using the frequencies of each pitching type that the players use. As shown from the plot, the model categorizes the pitchers into three clusters. The composition of different types of pitches that the pitchers in each cluster use is analyzed in Figure 9.

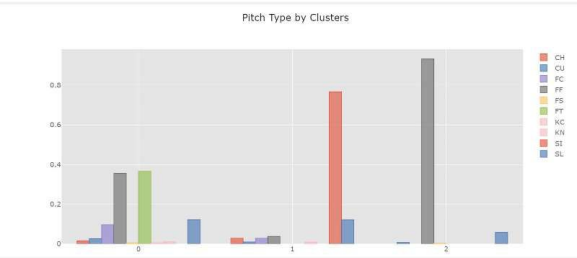

Figure 9 - Bar Chart of Pitcher Type by Cluster Analysis: Cluster 0 consists the most diverse pitchers such that they concentrate in various pitching type including FF – four seam fastball, FT – two seam fastball, and FC – cutter fastball. The pitchers in cluster 1 are less diversify as most of them concentrate in CH – changeup along with some concentrations of SL – slider. Pitchers in cluster 2 is the least diverse with a high concentration in FF – fastball.

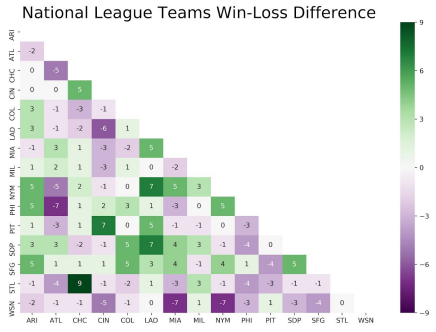

Figure 10 - Heatmap of National League Teams Win-Loss Difference:
The heatmap consists of the win-lose difference of the baseball teams in the National League. A higher number (green) represents a team wins a higher number of games while a smaller numbers (purple) represents a team loses a higher number of games.

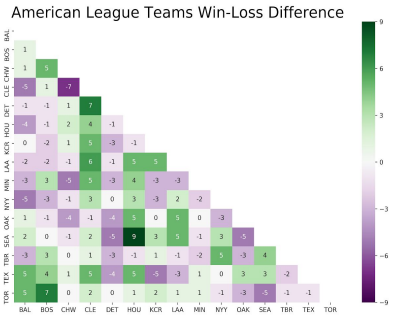

Figure 11 - Heatmap of American League Teams Win-Loss Difference:
The heatmap consists of the win-lose difference of the baseball teams in the American League. A higher number (green) represents a team wins a higher number of games while a smaller numbers (purple) represents a team loses a higher number of games.

Award Winner Prediction

Recall, precision, and F1-score are used to evaluate the award winner prediction model. As the dataset used in this part of the study is imbalanced, accuracy is not a good evaluation method since the model can simply classify the entire test data into the majority class and can already achieve a high accuracy. Therefore, recall and F1-score are used instead. Recall is the correctly predicted positive observation to the all observation in the actual class. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. As recall and precision is in a tradeoff situation such that a high recall often links with a low precision, F1-score is introduced as a balance in between recall and precision. The model validation is performed in two ways. The first validation method is by splitting data in between 1920 and 2000 into training data and test data and evaluating the model using the test data, which is from the same time period. The second validation method is by using data from 2001 to 2017 as the test data and evaluate the result. The recall, precision, and F1-score of the test dataset of the same time period (1920 to 2000) with multinomial logistic regression are 0.84, 0.89, 0.86 (weight average), respectively, and 0.9, 0.91, and 0.9 (weight average), respectively, for random forest. The tuning parameters used here are maximum depth = 19, maximum features = 5, and number of trees = 3000. The result of random forest forest is noticeably better than that of multinomial logistic regression. Therefore, random forest is further validated using test dataset from 2001 to 2017. The recall, precision, and F1-score of the test dataset of 2001 to 2017 with random forest are 0.9, 0.89, and 0.88, respectively, for random forest (weight average). The top seven important variables determined by random forest are number of years that the player plays, number of games that the player play, number of assists, number of runs, number of hits, number of strikeouts, and number of at bats. As these variables are are cumulative, this suggests the more experiences the player is would imply a higher likelihood of winning awards.

Salary and Career Peak Prediction

The training and test dataset are randomly separate by 70% and 30%. For career path prediction, the recall, precision, and F1-score of the test dataset with XGBoost are 0.90, 0.90, 0.90, respectively. The result of XGBoost is better than logistic regression and random forest with a high accuracy and recall. The tuning parameters here are learning_rate=0.1, n_estimators=10, max_depth=5, min_child_weight=3, gamma=0.2, subsample=0.6, colsample_bytree=1.0, objective='binary:logistic', nthread=4, scale_pos_weight=1, seed=27. For salary prediction, the average annual salary is applied log transformation as it significantly right skewed. As a result, the RMSE for XGBoost is within 0.74. The best parameters of are learning_rate =0.05, n_estimators=1000, max_depth=3, min_child_weight=0 ,gamma=0, subsample=0.7, colsample_bytree=0.7,objective= 'reg:linear', nthread=4, scale_pos_weight=1, and reg_alpha=0.00006.

VII. CONCLUSION

The study successfully constructs a comprehensive analysis of baseball players including accurate prediction models of the players' salary, career peak, and award. This study further provides a statistics based categorization of the different types of pitchers. In addition, to increase the accessibility of the massive information provided by this study, a web application is built for the baseball participants so that they can easily use the functions of this study without requiring any coding, data science, or statistics knowledge.

Future Study

As time is limited, this study has met certain limitations. Although there are many roles in baseball, this study only concentrates in the analysis of pitcher. In the future, this study can extend it's concentrations in analysis of other baseball roles, such as team managers, batters, and fielders.

Contribution

Each of the authors contributed equally in this study:

Leo Lam	Award Prediction and HTML Layout
Yuhan Zha	Career Peak Prediction, Salary Prediction, and HTML Layout
Wanting Cui	EDA, PCA, and Clustering Analysis

ACKNOWLEDGEMENT

The authors would like to thank Professor Ching-Yung Lin

APPENDIX

Variables Names of the Dataset [20]

Name	Explanation	Name	Explanation
W / L	Wins / Losses	C	Catcher
ATS	Record Against The Spread	1B	First Base
Slug	Slugging Percentage	2B	Second Base
Ho	Home record	3B	Third Base
Aw	Away Record	SS	Short Stop
O/U	Over/Under Record	LF	Left Field
AF	Average Runs For	CF	Centre Field
AA	Average Runs Against	RF	Right Field
BA	Batting Average	DH	Designated Hitter
SLG	Slugging Percentage	SP	Starting Pitcher
HR	Home Runs For	RP	Relief Pitcher
ERA	Earned Run Average		
OBP	On Base Percentage		
Home-Away	Home Score - Away Score	SB%	Stolen Base Percentage
H Starter	Home Starter in that particular game	QS%	Quality Start Percentage
A Starter	Away Starter in that particular game	TWL	Team Win - Team Loss
LOB:R	Left On Base to Runs ratio	W/L%	Winning Percentage
OPS	Slugging Percentage + On Base Percentage	vs. R	vs. Right-handed Pitchers
AVG	Batting Average for that game	vs. L	vs. Left-handed Pitchers
Starter	Team's Starter for that game	Start	Starters
IP	Innings the starter pitched	Rel	Relievers
Opp Starter	Innings the starter pitched	R/9	Runs per nine innings
H	Hits Allowed by the starter	K	Strikeouts
R	Runs Allowed by the starter	Doub	Doubles

ER	**Earned Runs Allowed by the starter**	**Trip**	**Triples**
SO	**Strikeouts by the starter**	**$**	**Units Won or Lost**
BB	**Base on Balls allowed by the starter**	**Line**	**Line for the game**
PIT	**Total Pitches by the starter**	**$ Won**	**Units Won**
P/IP	**Pitches divided by the number of Innings Pitched**	**$ Loss**	**Units Lost**
G/F	**Number of Ground Ball outs divided by the Fly Ball outs**	**SB**	**Stolen Bases**
OBA	**Opposition Batting Average**	**CS**	**Caught Stealing**
WHIP	**Walks and Hits per Inning Pitched**		
GB:FB	**Ground Ball to Fly Ball Ratio**		

REFERENCES

[1] Wikipedia: The Free Encyclopedia. *Sports in the United States*. Retrieve from https://en.wikipedia.org/wiki/Sports_in_the_United_States

[2] worldatlas. *The Most Popular Sports In World*. Retrieve from https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html

[3] statista. *Major League Baseball (MLB) - Statistics & Facts*. Retrieve from https://www.statista.com/topics/968/major-league-baseball/

[4] statista. *Number of participants in baseball in the United States from 2006 to 2017 (in millions)**. Retrieve from https://www.statista.com/statistics/191626/participants-in-baseball-in-the-us-since-2006/

[5] Wikipedia: The Free Encyclopedia. *History of baseball outside the United States*. Retrieve from https://en.wikipedia.org/wiki/History_of_baseball_outside_the_United_States#Baseball_World_Cup

[6] Wikipedia: The Free Encyclopedia. *Major League Baseball*. Retrieve from https://en.wikipedia.org/wiki/Major_League_Baseball

[7] statista. *Average player salary in Major League Baseball from 2003 to 2018 (in million U.S. dollars)*. Retrieve from https://www.statista.com/statistics/236213/mean-salaray-of-players-in-majpr-league-baseball/

[8] work.chron.com. *How to Become a Minor League Baseball Player.* Retrieve from https://work.chron.com/become-minor-league-baseball-player-26651.html

[9] The Atlantic. *Why Baseball Players Are Actually Underpaid.* Retrieve from https://www.theatlantic.com/business/archive/2012/04/why-baseball-players-are-actually-underpaid/255512/

[10] K. Isley (*2006, 28 February*), The Hardball Times. *On Defense: Subjective Data, Objectively Considered.* Retrieve from https://www.fangraphs.com/tht/on-defense-subjective-data-objectively-considered/

[11] E. Fischer (2014, 24 October), DearSportsFan. *What different kinds of pitchers are there in baseball?*. Retrieve from http://dearsportsfan.com/2014/10/24/different-kinds-pitchers-baseball/

[12] BASEBALL REFERENCE. *Pitcher.* Retrieve from https://www.baseball-reference.com/bullpen/Pitcher

[13] Stackexchange.com. *Why are there so many pitcher types in baseball?*. Retrieve from https://sports.stackexchange.com/questions/5765/why-are-there-so-many-pitcher-types-in-baseball

[14] Wikipedia: The Free Encyclopedia. *List of highest paid Major League Baseball players*. Retrieve from https://en.wikipedia.org/wiki/List_of_highest_paid_Major_League_Baseball_players#Highest_annual_salaries_in_2017

[15] m.mlb.com. *Statcast.* Retrieve from http://m.mlb.com/glossary/statcast

[16] D. Poston (2017, 20 June), DataCamp. *Scikit-Learn Tutorial: Baseball Analytics Pt2*. Retrieve from https://www.datacamp.com/community/tutorials/scikit-learn-tutorial-baseball-2

[17] A. Rubino (2017, 15 May), NYC DATA SCIENCE ACADEMY. *Predicting the Baseball Hall of Fame*. Retrieve from https://nycdatascience.com/blog/student-works/predicting-baseball-hall-fame/

[18] Baseball Data Science. *Machine Learning to Predict Player Decline.* Retrieve from http://www.baseballdatascience.com/machine-learning-to-predict-player-decline/

[19] Wikipedia: The Free Encyclopedia. *On-base plus slugging*. Retrieve from
https://en.wikipedia.org/wiki/On-base_plus_slugging
[20] Baseball Almanac. *Baseball Abbreviations.* Retrieve from
http://www.baseball-almanac.com/stats4.shtml