Machine Learning in Baseball

Heathvonn Styles, Joon Jung, Karthikeya Manchala

1 Introduction

Data analytics in baseball, one of the most popular sports in the United States, was brought into mainstream conscience by Moneyball - a famous book written in 2003. However, there has been limited research on improving existing performance metrics in baseball, which largely rely on basic statistics and heuristics. This project aims to leverage machine learning to determine the most important factors that contribute to the three key aspects of the sport - pitching, batting and overall team performance - using Major League Baseball (MLB) data.

2 Datasets

Ball tracking data was derived from StatCast for the 2022 MLB, which contained over 700,000 pitches and 100 variables. A Batting Summary Statistics dataset for the 2022 MLB was downloaded from Kaggle and the team performance data was obtained from FanGraphs with MLB team-by-team data from 2001-2021.

3 Pitching Model

The objective of the pitching model is to predict the outcome of a pitch. This is a multiclass classification problem with the potential outcomes being strike, ball or hit. Pitchers with a high percentage of predicted strikes are likely to perform better.

3.1 Methodology

- Data was loaded using the pybaseball Python package. Variables relevant to pitching, before the impact point between bat and ball were selected, including initial velocity, acceleration, release spin, release coordinates and coordinates when ball reaches the batter. Missing data (4 percent) was dropped. Categorical variables were one-hot encoded.
- A train-test split of 70-30 was used. The minority class - hits - was upsampled to mitigate imbalance in the data. Basic linear regression, decision tree, random forest and XGBoost models were trained. Hyperparameters were tuned for the best performing model, judged on the accuracy and F-1 score metrics.

3.2 Findings

- Random forest was the best performing model with an accuracy of 0.73 and an F-1 score of 0.75. There was minimal improvement by tuning hyperparameters.
- After generating a confusion matrix, it was found that balls and strikes were predicted well, but hits were often misclassified as strikes - possibly due to the data imbalance.
- The most important features were coordinates when ball reaches the batter, velocity in the z-direction (dip), velocity in the x-direction (curve) and release spin rate.

4 Batting Model

The objective of the batting model was to predict batting average of batters using ball tracking data and more generic variables such as age, team and on base percent. Higher predicted batting average signfies better batting performance.

4.1 Methodology

Feature engineering was performed to find how batters perform against different types of pitches using tracking data. The engineered features were merged with the batting summary data. A train-test split of 70-30 was used. Linear regression and tree-based models were trained and evaluated on Mean Squared Error and Mean Absolute Error.

4.2 Findings

- Linear regression was the best performing model with a MSE of 0.002 and a MAE of 0.01.
- Age of 26-30 was found to be the peak age of batters.
- On base percentage had the highest feature importance. Batters with high number of hits and those versatile against high spin rates performed better.

5 Team Performance Model

The goal team performance model of is to predict the overall season runs based on other summary statistics of the team throughout the season.

5.1 Methodology

Exploratory data analysis was performed for all years (2000-2021) to determine which variables were most strongly correlated with season runs. A linear regression was then fitted on the data since there seemed to be plenty of variables that were strongly correlated and linearly related to season runs.

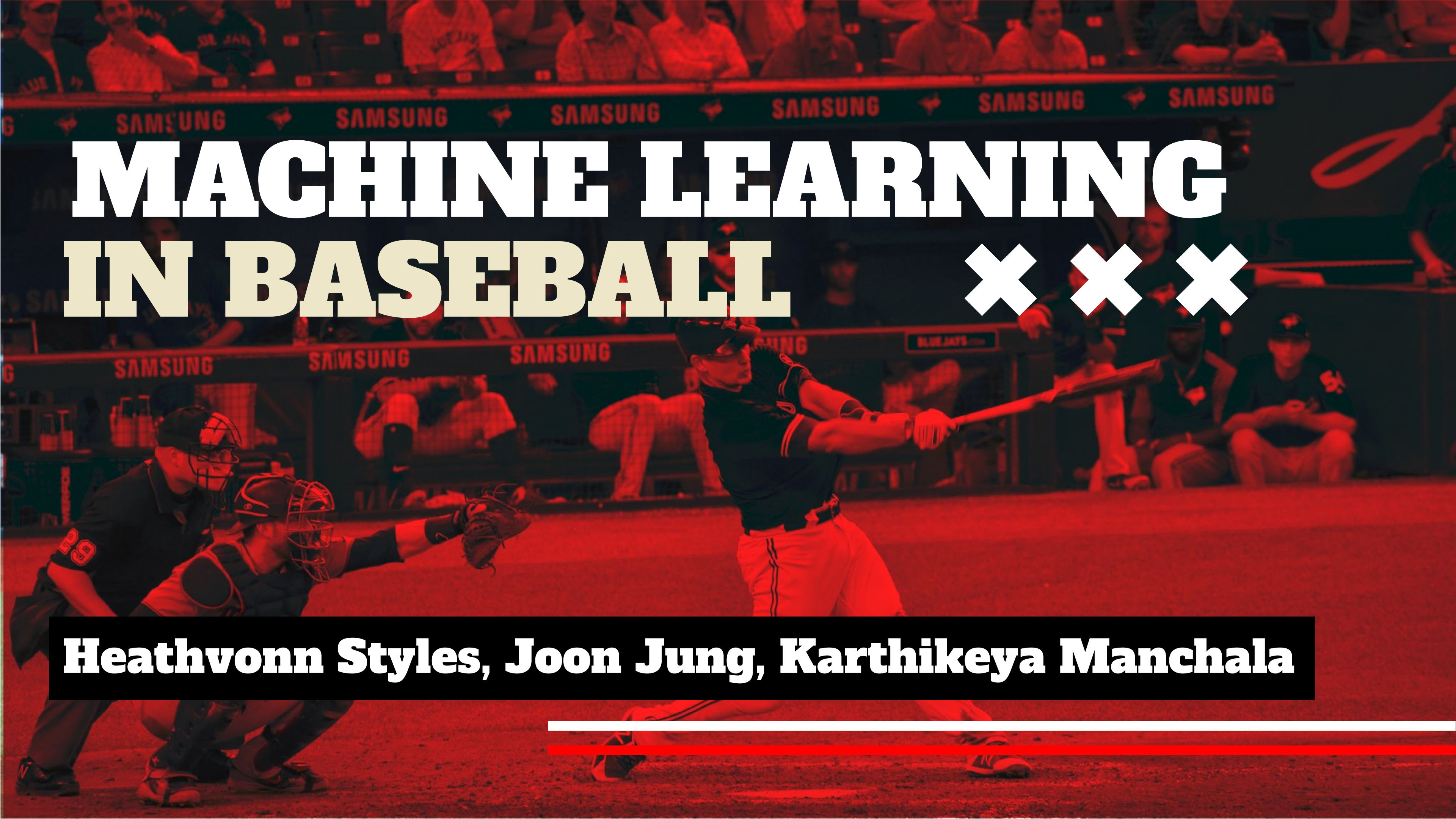
5.2 Findings

An R-squared of 0.91 was achieved using linear regression. Again, on base percent was very important in addition to isolated power. More common metrics like wRC+ and home runs, which are commonly used to judge baseball teams, proved to be less important.

6 Limitations and Future Work

The pitching model could not differentiate between Strikes and Hits as well as we would have liked, perhaps because we did not have data of the full ball trajectory. In addition, our pitching and batting models were trained on a single season's data and trends in baseball may change over time. Perhaps, using video data, we could use Computer Vision to predict Strikes and Hits, to improve our pitching model. For this project, we lacked the skills and computational resources, but this could be a future endeavor. We could test the predictiveness of our models on different years and compare them with existing models.

Overall, the project experience was very fruitful - it confirmed some theories we had and surprised us occasionally which made us look more closely at our methods. We gained valuable exposure to working with the full ML pipeline to answer some complex and open-ended questions.



MACHINE LEARNING IN BASEBALL

Heathvonn Styles, Joon Jung, Karthikeya Manchala

Background Information

High Popularity

The Major League Baseball is one of the most popular leagues in the world

High Stakes

The livelihoods of players, coaches, TV companies and sponsors depend on the sport

Demand for Analytics

Data analytics is used for player recruitment, in-game strategy and match evaluation

Lack of Context

Currently, only basic statistics and heuristics are used for performance analysis

Lack of ML research

There is a shortage of research on machine learning models that leverage the wealth of open source data

Project Overview

01 Pitcher Performance

We developed an xStrike model to rank pitchers using advanced ball tracking data.

02 Batter Performance

We used findings from our xStrike model to develop an xBattingAvg model - an advanced statistic that contextualizes batting average.

03 Team Performance

We developed an xRuns model to determine the key factors that drive team success.

Datasets

Batting Stats Dataset

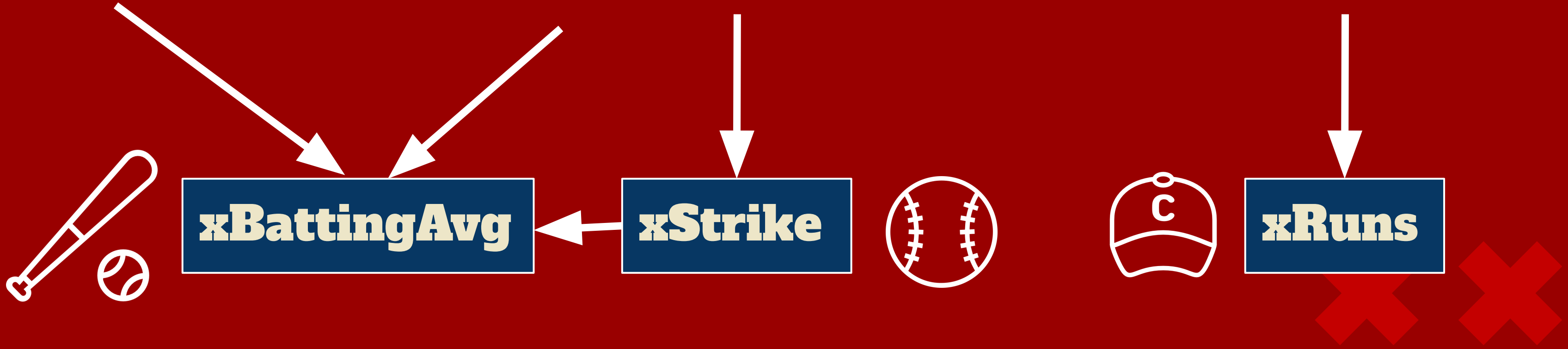
Source: Kaggle
Rows: 989
Columns: 22
Summary stats of all batters in the 2022 MLB season.

Pitching Ball Tracking

Source: StatCast
Rows: 765,445
Columns: 100
Tracking data from every pitch recorded in the 2022 MLB season.

Team Summary Dataset

Source: FanGraphs
Rows: 660
Columns: 23
Team aggregate stats for the 30 MLB teams from 2000-2021



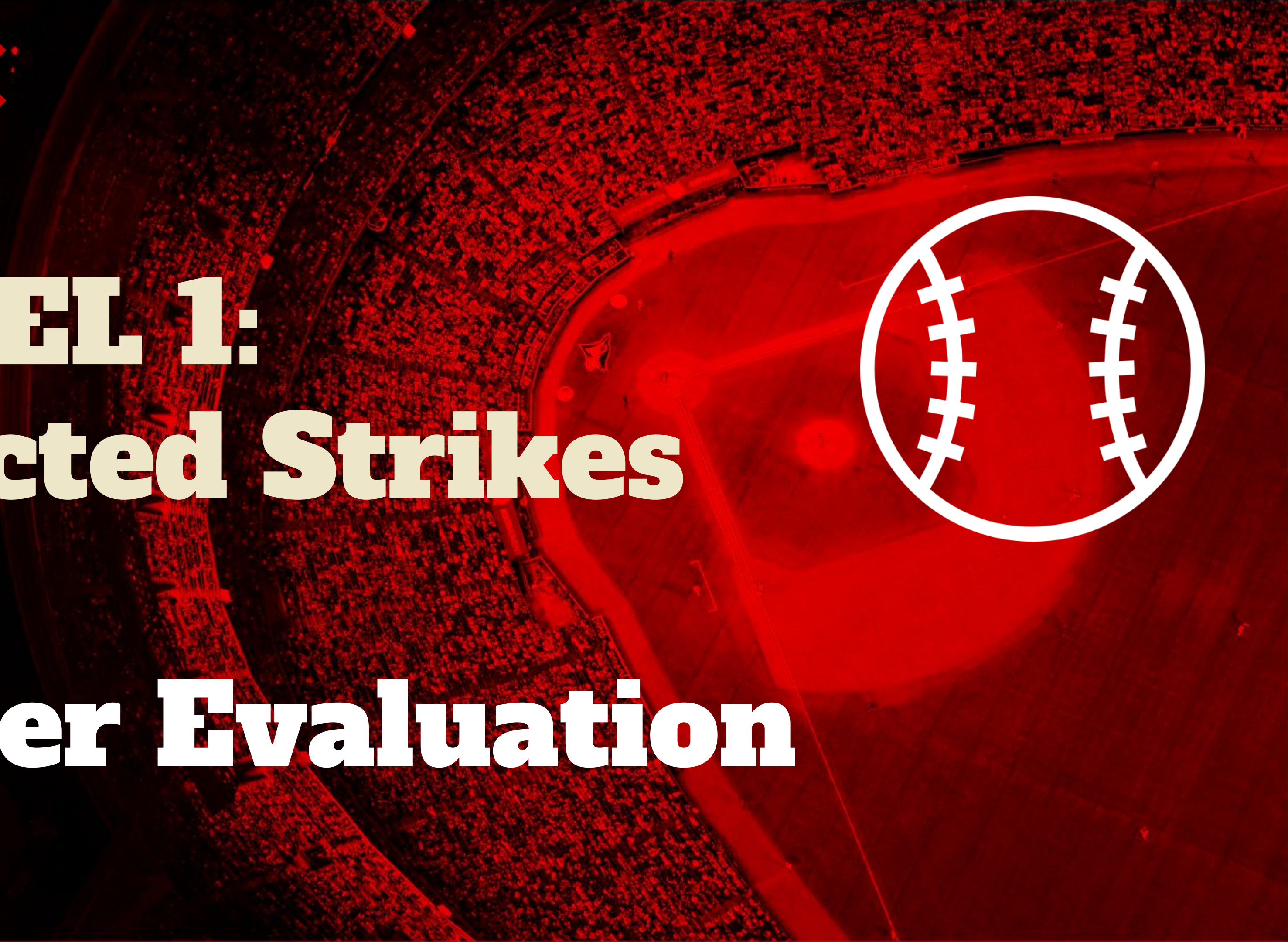
xBattingAvg

xStrike

xRuns

MODEL 1: Expected Strikes

Pitcher Evaluation



Data Curation & Processing

01 Load Data from API

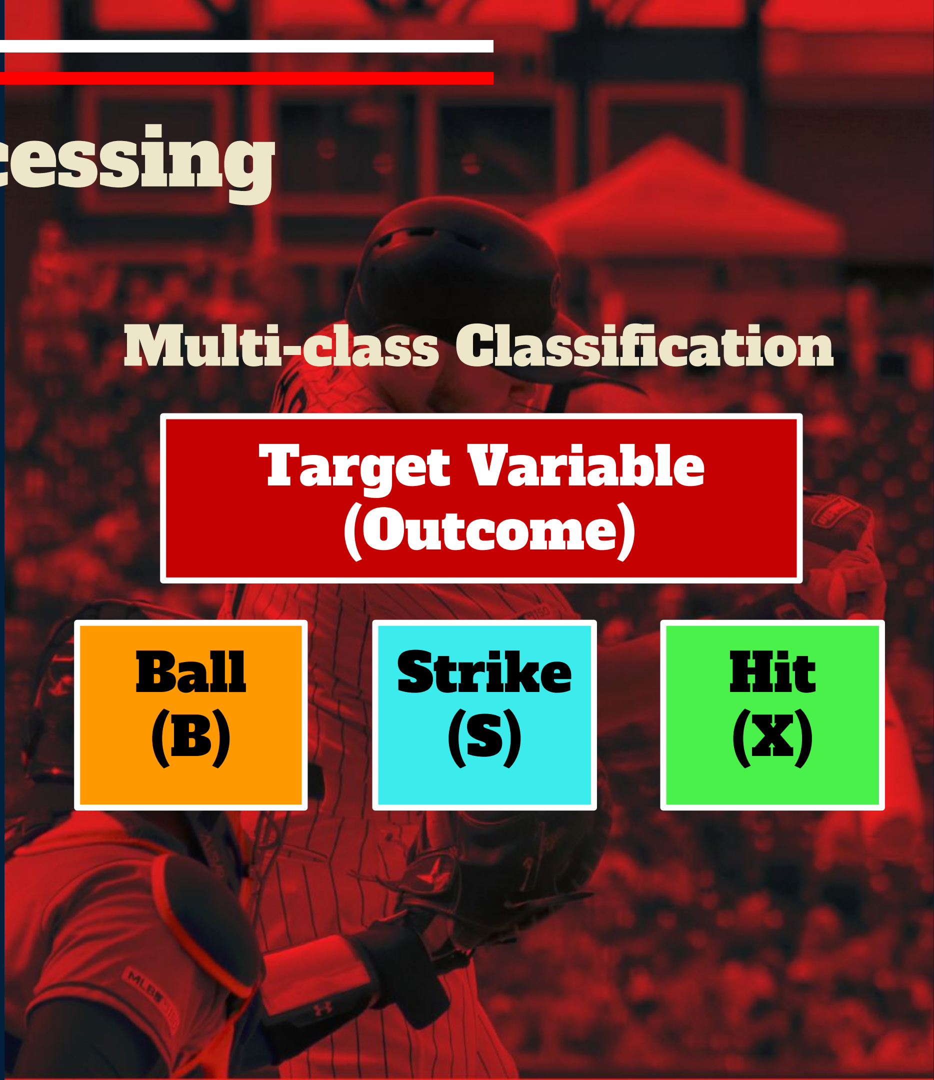
The `pybaseball` Python package was used to load data on every pitch in the 2022 MLB season.

02 Variable Selection

All variables relevant to the pitch before the ball is struck were kept, such as release speed, acceleration, spin as well as pitcher hand, pitch type etc. Missing data was dropped (4%).

03 One-Hot Encoding

Categorical variables were one-hot encoded.



Multi-class Classification

Target Variable (Outcome)

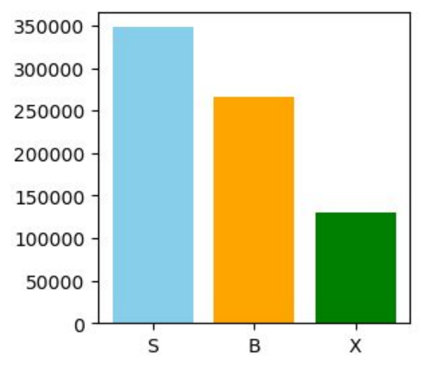
Ball (B)

Strike (S)

Hit (X)

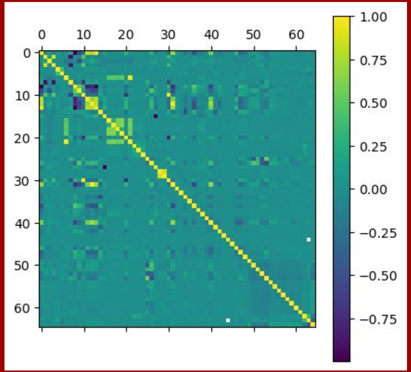
Exploratory Data Analysis

Data Imbalance



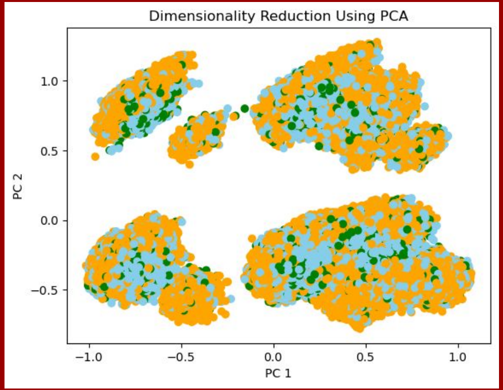
Outcomes are not evenly distributed - strikes are very common and hits occur rarely

No Strong Correlations



Very few variables very highly correlated with each other and in particular, with the target variable. Non-linear model?

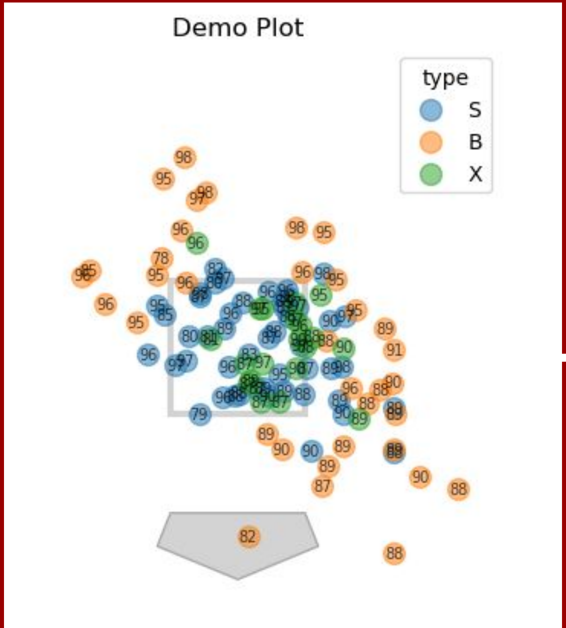
Inconclusive PCA



37 of the 62 predictors were required to explain 99% of the variance. The outcome variable is scatter across clusters

Hypothesis

The coordinates when the ball reaches the plate are most important, but other factors such as spin and dip will be necessary for a better performing model.



Model Building

01 Upsampling minority class

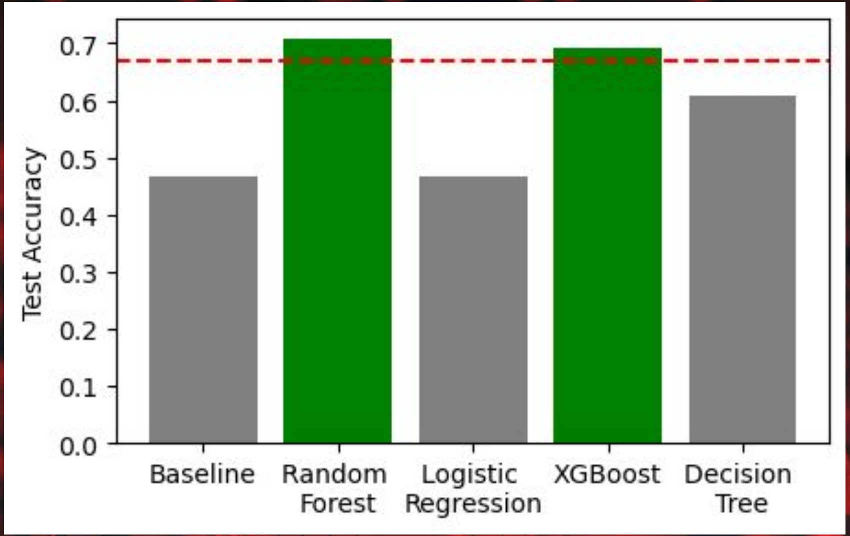
We used a train-test split with test size of 30%. We upsampled Hits (X) in the training data to mitigate the effect of imbalanced data.

02 Model Selection

We trained a basic Logistic Regression, Decision Tree, Random Forest and XGBoost. The baseline model - predicting majority class always resulted in 46% accuracy.

03 Hyperparameter Tuning

We selected the best baseline model and used trial & error to optimize performance.



Random Forest was the best model

Test Accuracy: 0.73
Test F-1 score: 0.75

Evaluation

Confusion Matrix

		Actual		
		B	S	X
Predicted	B	70478	16366	3354
	S	9148	81282	29497
	X	186	6730	6185

We did well at predicting Balls (B) and Strikes (S) but were poor at predicting Hits (X). Since we are focusing on Strikes, we can let this pass, but perhaps we should have dealt with data imbalance better.

Accuracy and F-1 Score

Achieving an accuracy of >80% is extremely hard despite having a huge sample size. Past research papers achieved accuracies of ~67%, so we did well.

Key Findings

Feature Importances

The Top 5 most important features in the Random Forest were:

- plate_x: The x-coordinate when the ball crosses the plate
- plate_z: The height of the ball when it crosses the plate.
- vz_0: The initial velocity in the downward direction - i.e., how much the ball is dipping.
- vx_0: The initial velocity in the sideways direction - i.e., how much the ball is curving.
- release_spin_rate: The revolutions imparted on the ball.

When recruiting pitchers, teams should look for **ACCURATE** pitchers who can impart both **DIP** and **CURVE** on the ball. Variety of pitches is important!

Top Pitchers in 2022 (min. 1000 pitches)

Rank	Pitcher	xStrikes %
1	Edwin Diaz	0.59
2	Emmanuel Clase	0.58
3	Chris Martin	0.57
4	Joe Jimenez	0.57
5	Jhoan Duran	0.56
6	Kenley Jansen	0.56
7	Jordan Romano	0.56
8	Caleb Thielbar	0.55

One of the most important players for the New York Mets!

Our Expected Strikes model suggests that these pitchers had the best underlying numbers in the 2022 season.

This information could be useful for future recruitment, deciding player salaries and potential trades!

Players with higher xStrikes than actual Strikes were perhaps unlucky and this could be taken into account.

MODEL 2: Expected Batting Avg

Batter Evaluation

Data Curation & Processing

01 Merge Datasets

The StatCast and Kaggle datasets were merged on Player Name. Some data wrangling was required for merging without data loss.

02 Feature Engineering

New features were created. These were the differentials in the average of velocities, spin rates, launch angle, etc of balls faced by batters for successful & unsuccessful plays.

03 Binning Data

Some numerical variables such as Age were binned and one-hot encoded.

Regression Problem

Target Variable (Batting Average)

Basic Predictors (Age, Team, Innings, Balls Faced, etc.)

Advanced Features (Differentials of Launch Angle, Velocities, Spin, Acceleration, etc.)

Feature Engineering

events
field_out
single
force_out
home_run
strikeout
double
walk
NaN

STEP 1:
GROUP BY
Event

events	release_speed	launch_speed	spin_rate	....	release_pos
field_out	91.23	52.11	123.42	....	0.39
single	90.34	103.11	120.03	....	1.12
force_out	89.33	48.89	129.99	....	-0.43
home_run	92.34	123.09	100.45	....	-0.89
strikeout	90.07	91.23	107.63	....	0.78
double	87.21	100.06	111.72	....	-1.12
walk	90.07	23.93	118.43	....	-1.02
NaN	NaN	NaN	NaN	....	NaN

STEP 2:
Drop missing data and identify successful plays

success	release_speed	launch_speed	spin_rate	....	release_pos
unsuccessful	91.23	52.11	123.42	....	0.39
successful	90.34	103.11	120.03	....	1.12
unsuccessful	89.33	48.89	129.99	....	-0.43
successful	92.34	123.09	100.45	....	-0.89
unsuccessful	90.07	91.23	107.63	....	0.78
successful	87.21	100.06	111.72	....	-1.12
successful	90.07	23.93	118.43	....	-1.02

Feature Engineering

success	release_speed	launch_speed	spin_rate	....	release_pos
unsuccessful	91.23	52.11	123.42	....	0.39
successful	90.34	103.11	120.03	....	1.12

STEP 3: Compute Differences

Engineered Features

Batter Name	release_speed	launch_speed	spin_rate	....	release_pos
Judge, Aaron	1.11	-51.00	23.39	....	-0.73

STEP 4: Repeat for all players

Batter Name	release_speed	launch_speed	spin_rate	....	release_pos
Judge, Aaron	1.11	-51.00	23.39	....	-0.73
Ohtani, Shohei	1.02	-23.45	-0.90	....	0.00
....	....	....	....	....	
Alvarez, Yordan	-0.89	72.38	20.04	....	0.45

Model Building

01 Choosing a Metric

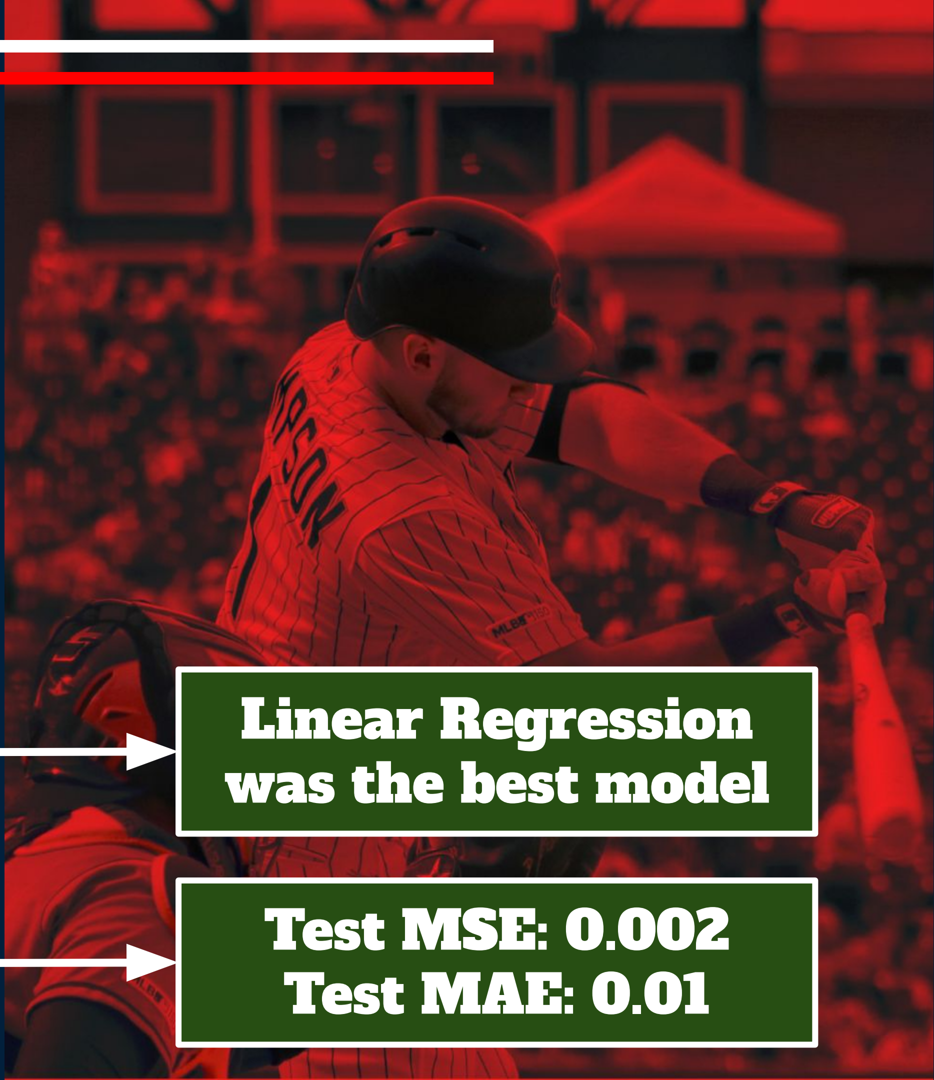
We decided to use Mean Squared Error as the metric to optimize, but we also trialled Mean Absolute Error

02 Model Selection

A basic Linear Regression model was trained. We also tried tree-based models, but Linear Regression still produced the best results.

03 Trialing different feature sets

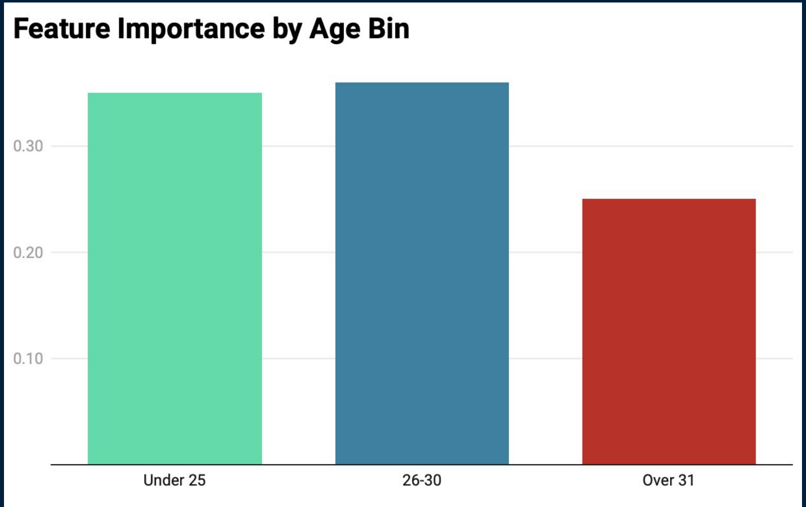
We used different combinations of feature sets, for example with and without binning / one-hot encoding to optimize performance.



Linear Regression was the best model

Test MSE: 0.002
Test MAE: 0.01

Effect of Age



We examined in detail how Age affects xBattingAvg.

Younger players performed better while older players appear to be liabilities. It seems like 26-30 is the peak age for batters.

Key Findings

Feature Importances

The Top 5 most important features were:

- **OBP**: The percentage of plate appearances that result in the batter reaching base safely
- **H**: Total Hits.
- **effective_speed**: TA metric that combines pitch velocity and spin rate to provide a measure of how challenging a pitch is for the batter to track
- **outs_when_up**: refers to the number of outs that have occurred in the inning when a particular event or statistic is being recorded for a player
- **hit_distance_sc**: used to measure the power or distance of a hit

When recruiting batters, teams should look for players in the age range **26 - 30**, with a **high OBP (On-Base Percentage),** who can get plenty **hits** and handle **high effective speeds** while imparting the greatest **distance** on the ball.

Top Batters in 2022 (min. 500 pitches)

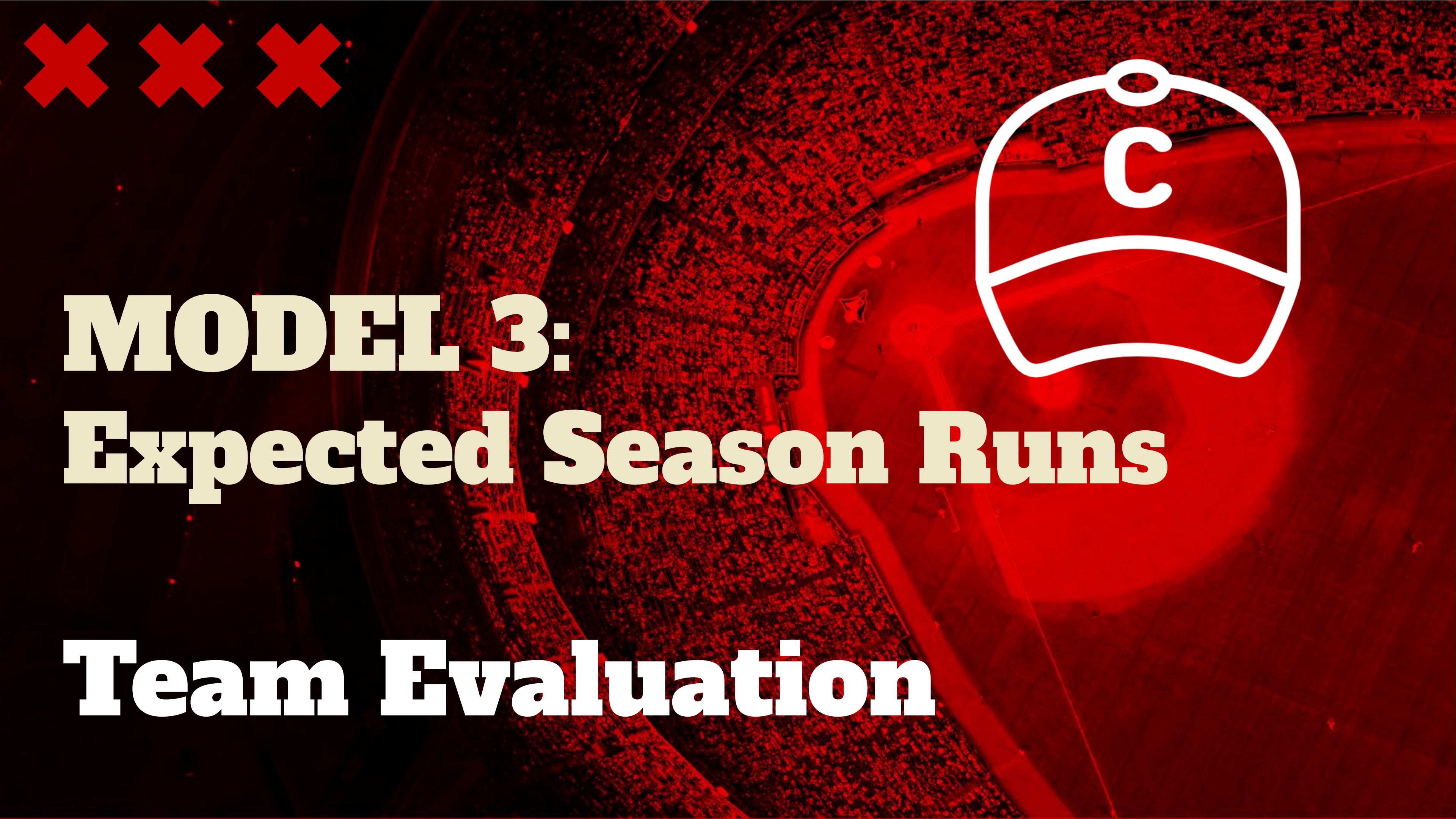
Rank	Pitcher	xBattingAvg
1	Jeff McNeil	0.325
2	Freddie Freeman	0.324
3	Joey Meneses	0.323
4	Paul Goldschmidt	0.317
5	Luis Arraez	0.315
6	Aaron Judge	0.310
7	Xander Bogaerts	0.307
8	Yordan Alvarez	0.306

Won 2 All Star Nods
and the Batting Title!

Our Expected Batting Average model suggests that these batters had the best underlying numbers in the 2022 season.

This information could be useful for future recruitment, deciding player salaries and potential trades!

Players with higher xBattingAvg than actual Batting Average were perhaps unlucky and this could be taken into account.



MODEL 3: Expected Season Runs

Team Evaluation

Data Curation & Processing

01 Combine Year-Wise Data

Since this model requires data from multiple seasons, CSV files had to be merged from each of the seasons..

02 Feature Selection

Variables that summarize a team's performance over the course of the season were used as predictors for the model.

03 Outlier Detection

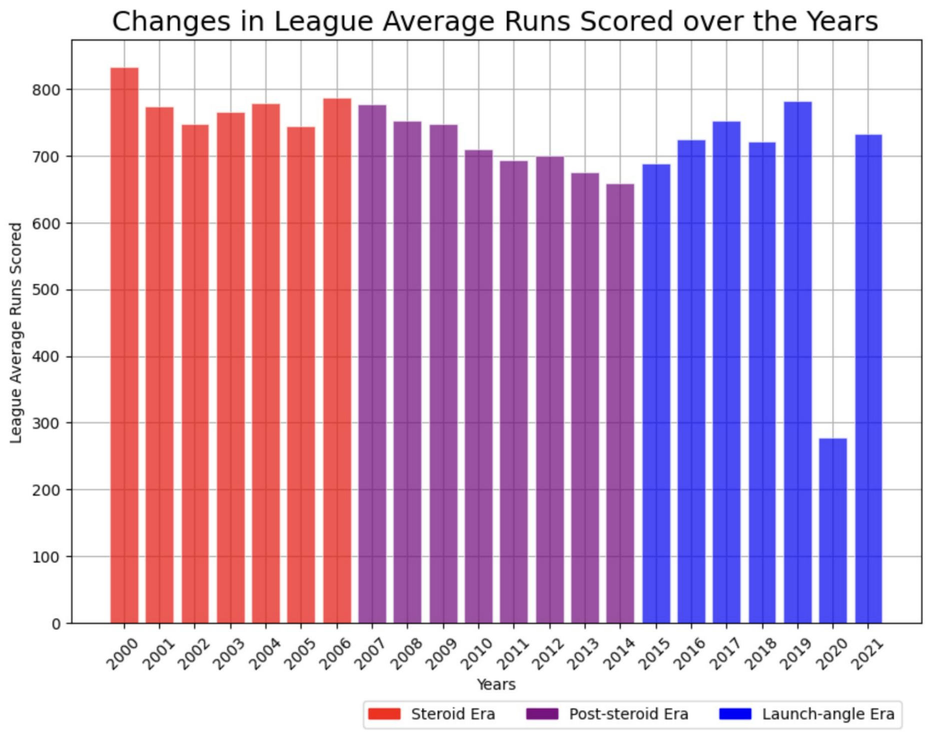
Since the 2020 season was affected by COVID, there were some outliers which were dropped.

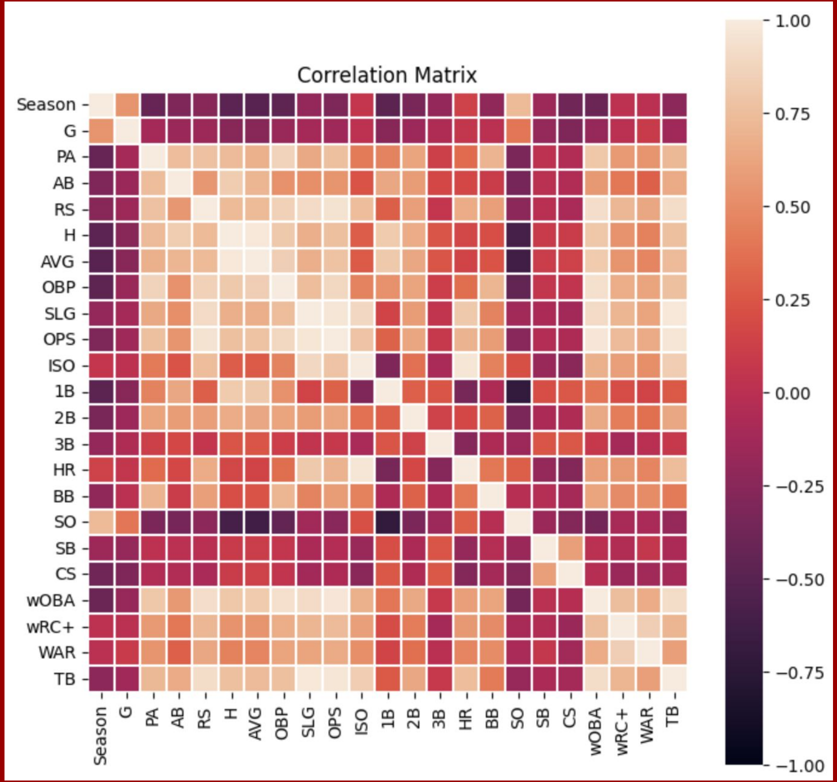
Regression Problem

Target Variable (Season Runs)

Predictors (Matches, Singles, Doubles, Home Runs, Batting Avg, wRC+)

Exploratory Data Analysis





Model Building

01 Choosing a Metric

We decided to use R-squared as the metric we want to optimize.

02 Model Selection

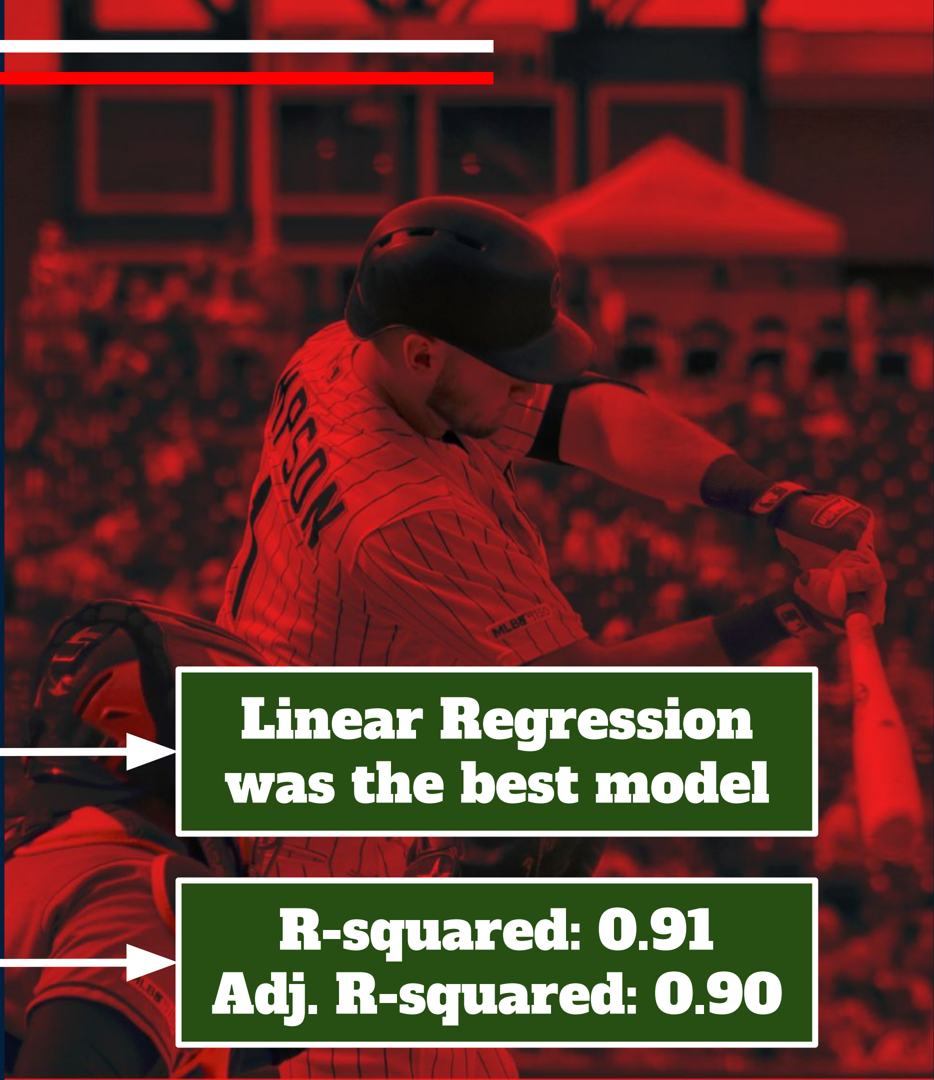
A basic Linear Regression model was trained given this is a fairly straightforward regression task with variables linearly related with the target

Linear Regression was the best model

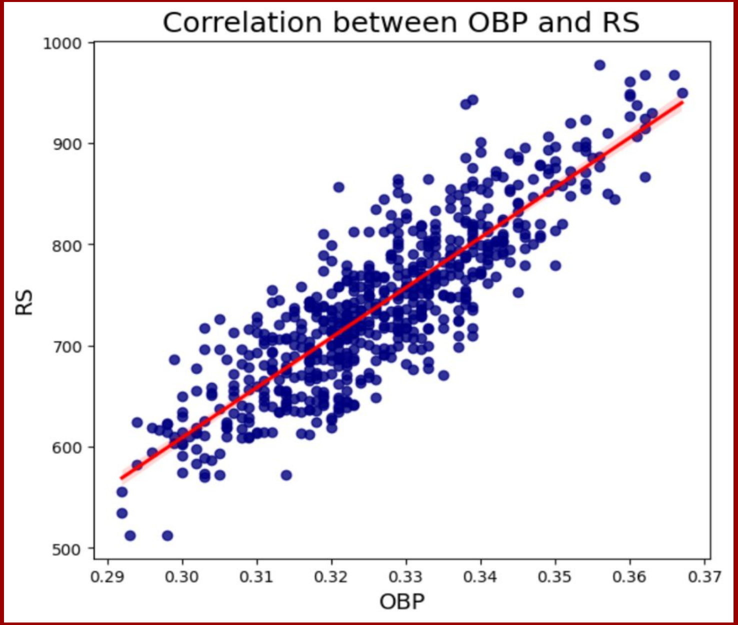
03 Train-Test Split

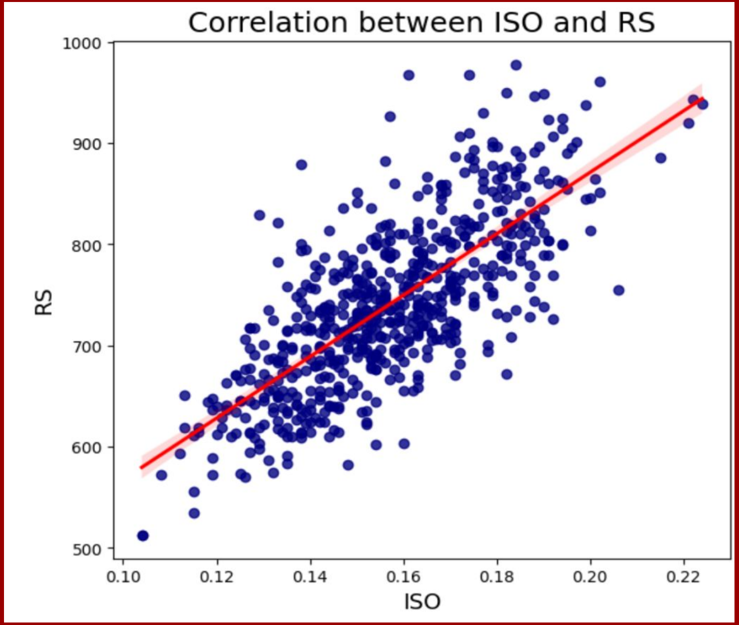
We chose earlier years to train the model and later years to predict on to simulate a real-world use of this model.

R-squared: 0.91
Adj. R-squared: 0.90



Feature Selection





On Base Percentage and Isolated Power are top two highly positively correlated with Runs Scored

Evaluation

Features	R-squared	Ad-R^2
[H, ISO, OBP]	0.913	0.912
[H, ISO, OBP, wRC+]	0.913	0.912

Adding wRC+ as an additional feature did not increase the R-squared, and adjusted R-squared value

"Your goal shouldn't be to buy players, it should be to buy wins. And in order to buy wins, you need to buy runs."

- Quote from Moneyball

Key Findings

Feature Importances

The most important in order were:

- **OBP**: On Base Percentage
- **ISO**: Isolated Power metric.
- **H**: Number of hits.
- **wRC+**: Weighted Runs Created Plus. (No impact)

These findings were surprising since wRC+ is typically the most important statistic used to judge teams.

Although Home Runs look glorious, they contribute less to overall season performance compared to smaller hits!

Conclusions

Ball Tracking

Ball tracking data like curve, dip and spin is important for recruiting pitchers.

Versatility is Key

Pitchers who can deliver a variety of pitches and batters who can handle high effective speeds do better.

Home Runs Don't Win Titles

While there may be a temptation to go for Home Runs, they don't contribute much to season performance.

Batters in Peak Age

Recruit batters who are in their peak age, i.e., 26-30.

High OBP

A high On Base Percentage is key for both batter performance and overall team performance.

Limitations

- Our pitching model could not differentiate between Strikes and Hits as well as we would have liked, perhaps because we did not have data of the full ball trajectory.
- Our pitching and batting models were trained on a single season's data. Trends in baseball may change over time.
- When predicting season performance, we always have to deal with a small sample size, which could affect our analysis.

Future Work

- Perhaps, using video data, we could use Computer Vision to predict Strikes and Hits, to improve our xStrike model. For this project, we lacked the skills and computational resources, but this could be a future endeavor.
- We could test the predictiveness of our models on different years and compare them with existing models.
- There is scope for more research on what contributes to higher win percentage, apart from Runs Scored.

THANK YOU

A special thank you to Professor Isayev and Kamal for giving us the opportunity to work on this project and also for helping us develop the skills required to accomplish the objectives we set out to achieve!