Mini-Project 2

MLB Playoff Qualification: A Data Science Perspective

By I. Font

Aim

Aim

Explore the factors influencing playoff qualification in Major League Baseball (MLB) by addressing the following questions:

1. How many games must a team win to qualify for the playoffs?

2. What run differential (runs scored minus runs allowed) is needed to achieve this win threshold?

3. What factors contribute to a team achieving the required run differential?

Aim

Explore the factors influencing playoff qualification in Major League Baseball (MLB) by addressing the following questions:

1. How many games must a team win to qualify for the playoffs?

2. What run differential (runs scored minus runs allowed) is needed to achieve this win threshold?

3. What factors contribute to a team achieving the required run differential?

Aim

Explore the factors influencing playoff qualification in Major League Baseball (MLB) by addressing the following questions:

1. How many games must a team win to qualify for the playoffs?

2. What run differential (runs scored minus runs allowed) is needed to achieve this win threshold?

3. What factors contribute to a team achieving the required run differential?

Aim

Explore the factors influencing playoff qualification in Major League Baseball (MLB) by addressing the following questions:

1. How many games must a team win to qualify for the playoffs?

2. What run differential (runs scored minus runs allowed) is needed to achieve this win threshold?

3. What factors contribute to a team achieving the required run differential?

The dataset

The dataset

CSV file from Kaggle with 1232 samples and 15 features

The dataset

CSV file from Kaggle with 1232 samples and 15 features

Two leagues: the National League and the American League

The dataset

CSV file from Kaggle with 1232 samples and 15 features

Two leagues: the National League and the American League

Data from 1962 to 2012. Missing entries for 1972, 1981, 1994 and 1995 (labor disputes)

The dataset

CSV file from Kaggle with 1232 samples and 15 features

Two leagues: the National League and the American League

Data from 1962 to 2012. Missing entries for 1972, 1981, 1994 and 1995 (labor disputes)

2 out of 10 teams went to playoffs in 1962 vs 10 out of 30 in 2012

The dataset

CSV file from Kaggle with 1232 samples and 15 features

Categorical variables: Team, League, Playoffs (binary)

Two leagues: the National League and the American League

Data from 1962 to 2012. Missing entries for 1972, 1981, 1994 and 1995 (labor disputes)

2 out of 10 teams went to playoffs in 1962 vs 10 out of 30 in 2012

The dataset

CSV file from Kaggle with 1232 samples and 15 features

Numerical variables: year, G, W, RS, RA, BA, OBP, SLG, OOBP, OSLG, RankSeason, RankPlayoffs

Categorical variables: Team, League, Playoffs (binary)

Two leagues: the National League and the American League

Data from 1962 to 2012. Missing entries for 1972, 1981, 1994 and 1995 (labor disputes)

2 out of 10 teams went to playoffs in 1962 vs 10 out of 30 in 2012

The dataset

CSV file from Kaggle with 1232 samples and 15 features

Two leagues: the National League and the American League

Data from 1962 to 2012. Missing entries for 1972, 1981, 1994 and 1995 (labor disputes)

2 out of 10 teams went to playoffs in 1962 vs 10 out of 30 in 2012

Categorical variables: Team, League, Playoffs (binary)

Numerical variables: year, G, W, RS, RA, BA, OBP, SLG, OOBP, OSLG, RankSeason, RankPlayoffs

OOBP and OSLG recorded from 1999. Missing values filled with team averages

The dataset

CSV file from Kaggle with 1232 samples and 15 features

Two leagues: the National League and the American League

Data from 1962 to 2012. Missing entries for 1972, 1981, 1994 and 1995 (labor disputes)

2 out of 10 teams went to playoffs in 1962 vs 10 out of 30 in 2012

Categorical variables: Team, League, Playoffs (binary)

Numerical variables: year, G, W, RS, RA, BA, OBP, SLG, OOBP, OSLG, RankSeason, RankPlayoffs

Data Leakage: RankSeason and RankPlayoffs only for teams that qualified.

OOBP and OSLG recorded from 1999. Missing values filled with team averages

The dataset

CSV file from Kaggle with 1232 samples and 15 features

Two leagues: the National League and the American League

Data from 1962 to 2012. Missing entries for 1972, 1981, 1994 and 1995 (labor disputes)

2 out of 10 teams went to playoffs in 1962 vs 10 out of 30 in 2012

Categorical variables: Team, League, Playoffs (binary)

Numerical variables: year, G, W, RS, RA, BA, OBP, SLG, OOBP, OSLG, RankSeason, RankPlayoffs

Data Leakage: RankSeason and RankPlayoffs only for teams that qualified.

Two features created: W_perc and R_diff

OOBP and OSLG recorded from 1999. Missing values filled with team averages

Method

How many games must a team win?

Method

How many games must a team win?

Logistic regression classification model

Method

How many games must a team win?

Logistic regression classification model

target variable: Playoffs

Method

How many games must a team win?

Logistic regression classification model

- target variable: Playoffs
- feature: W_perc (percentage of games won)

Method

How many games must a team win?

Logistic regression classification model

- target variable: Playoffs
- feature: W_perc (percentage of games won)

Target variable is unbalanced 1:4

Method

How many games must a team win?

Logistic regression classification model

target variable: Playoffs

feature: W_perc (percentage of games won)

Target variable is unbalanced 1:4

Model was fitted with class_weight='balanced'

Method

How many games must a team win?

Logistic regression classification model

target variable: Playoffs

feature: W_perc (percentage of games won)

Target variable is unbalanced 1:4

Model was fitted with
class_weight='balanced'
and solver='saga'

Method

How many games must a team win?

Logistic regression classification model

target variable: Playoffs

feature: W_perc (percentage of games won)

Target variable is unbalanced 1:4

Model was fitted with
class_weight='balanced'
and solver='saga'

Optimal decision threshold was found
using precision_recall_curve( )

Method

How many games must a team win?

Logistic regression classification model

target variable: Playoffs

feature: W_perc (percentage of games won)

Target variable is unbalanced 1:4

Model was fitted with
class_weight='balanced'
and solver='saga'

Optimal decision threshold was found
using precision_recall_curve( )

Results

Method

How many games must a team win?

Logistic regression classification model

target variable: Playoffs

feature: W_perc (percentage of games won)

Target variable is unbalanced 1:4

Model was fitted with
class_weight='balanced'
and solver='saga'

Optimal decision threshold was found using precision_recall_curve( )

Results

	precision	recall	f1-score	support
0	0.99	0.87	0.93	290
1	0.68	0.96	0.79	80
accuracy			0.89	370
macro avg	0.83	0.92	0.86	370
weighted avg	0.92	0.89	0.90	370

Method

How many games must a team win?

Logistic regression classification model

target variable: Playoffs

feature: W_perc (percentage of games won)

Target variable is unbalanced 1:4

Model was fitted with
class_weight='balanced'
and solver='saga'

Optimal decision threshold was found using precision_recall_curve( )

Results

	precision	recall	f1-score	support
0	0.99	0.87	0.93	290
1	0.68	0.96	0.79	80
accuracy			0.89	370
macro avg	0.83	0.92	0.86	370
weighted avg	0.92	0.89	0.90	370

Predicted well teams that did not qualify for playoffs.

Method

How many games must a team win?

Logistic regression classification model

target variable: Playoffs

feature: W_perc (percentage of games won)

Target variable is unbalanced 1:4

Model was fitted with
class_weight='balanced'
and solver='saga'

Optimal decision threshold was found using precision_recall_curve( )

Results

	precision	recall	f1-score	support
0	0.99	0.87	0.93	290
1	0.68	0.96	0.79	80
accuracy			0.89	370
macro avg	0.83	0.92	0.86	370
weighted avg	0.92	0.89	0.90	370

Predicted well teams that did not qualify for playoffs.

Captured 96% of the teams that went to the playoffs

Method

How many games must a team win?

Logistic regression classification model

target variable: Playoffs

feature: W_perc (percentage of games won)

Target variable is unbalanced 1:4

Model was fitted with
class_weight='balanced'
and solver='saga'

Optimal decision threshold was found using precision_recall_curve( )

Results

	precision	recall	f1-score	support
0	0.99	0.87	0.93	290
1	0.68	0.96	0.79	80
accuracy			0.89	370
macro avg	0.83	0.92	0.86	370
weighted avg	0.92	0.89	0.90	370

Predicted well teams that did not qualify for playoffs.

Captured 96% of the teams that went to the playoffs

32% of the teams predicted as qualifying actually didn't.

Method

How many games must a team win?

Logistic regression classification model

target variable: Playoffs

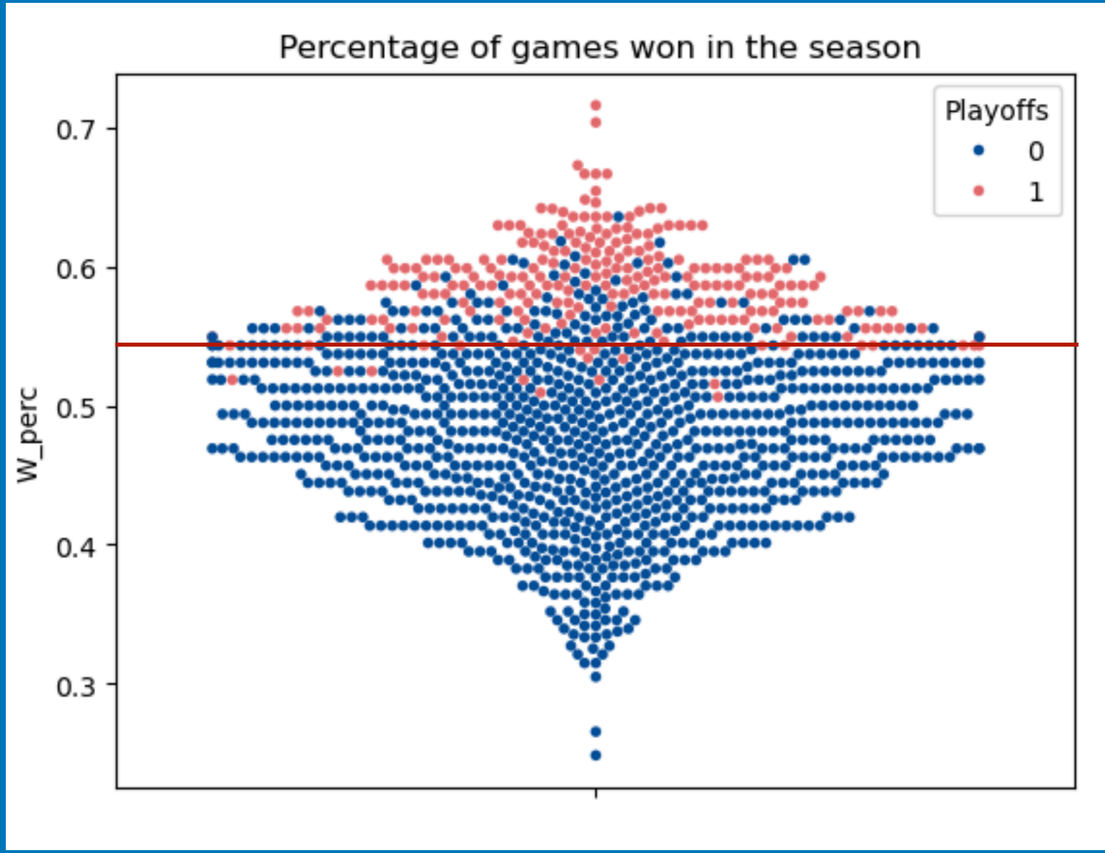
feature: W_perc (percentage of games won)

Target variable is unbalanced 1:4

Model was fitted with
class_weight='balanced'
and solver='saga'

Optimal decision threshold was found using precision_recall_curve( )

Results



Method

How many games must a team win?

Logistic regression classification model

target variable: Playoffs

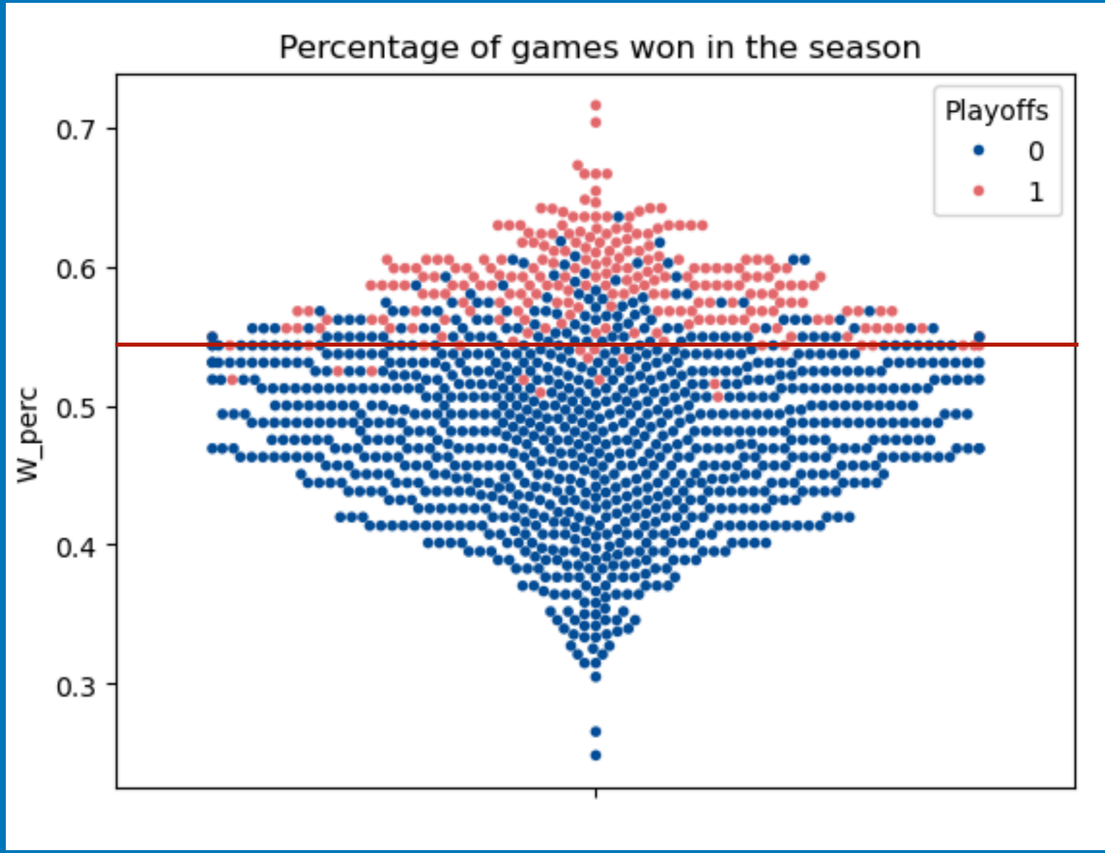
feature: W_perc (percentage of games won)

Target variable is unbalanced 1:4

Model was fitted with
class_weight='balanced'
and solver='saga'

Optimal decision threshold was found using precision_recall_curve( )

Results



At decision point: W_perc = 0.54, therefore must win 88 out of the 162 games

Method

What run differential is needed to win 88 games?

Method

What run differential is needed to win 88 games?

Univariate linear regression model

Method

What run differential is needed to win 88 games?

Univariate linear regression model

target variable: W (games won)

Method

What run differential is needed to win 88 games?

Univariate linear regression model

target variable: W (games won)

feature: R_diff (runs scored - runs allowed)

Method

What run differential is needed to win 88 games?

Univariate linear regression model

target variable: W (games won)

feature: R_diff (runs scored - runs allowed)

Target and feature variables have strong positive correlation (0.94)

Method

What run differential is needed to win 88 games?

Univariate linear regression model

target variable: W (games won)

feature: R_diff (runs scored - runs allowed)

Target and feature variables have strong positive correlation (0.94)

Both target and feature variable are normally distributed

Method

What run differential is needed to win 88 games?

Univariate linear regression model

target variable: W (games won)

feature: R_diff (runs scored - runs allowed)

Target and feature variables have strong positive correlation (0.94)

Both target and feature variable are normally distributed

Results

Method

What run differential is needed to win 88 games?

Univariate linear regression model

target variable: W (games won)

feature: R_diff (runs scored - runs allowed)

Target and feature variables have strong positive correlation (0.94)

Both target and feature variable are normally distributed

Results

R^2 training data = 0.87, R^2 test data = 0.89

Method

What run differential is needed to win 88 games?

Univariate linear regression model

target variable: W (games won)

feature: R_diff (runs scored - runs allowed)

Target and feature variables have strong positive correlation (0.94)

Both target and feature variable are normally distributed

Results

R^2 training data = 0.87, R^2 test data = 0.89

MSE = 18.1 or average prediction error of 4.25 games

Method

What run differential is needed to win 88 games?

Univariate linear regression model

target variable: W (games won)

feature: R_diff (runs scored - runs allowed)

Target and feature variables have strong positive correlation (0.94)

Both target and feature variable are normally distributed

Results

R^2 training data = 0.87, R^2 test data = 0.89

MSE = 18.1 or average prediction error of 4.25 games

Predicted a run differential of 69 (rounded up) for winning 88 games

Method

What run differential is needed to win 88 games?

Univariate linear regression model

target variable: W (games won)

feature: R_diff (runs scored - runs allowed)

Target and feature variables have strong positive correlation (0.94)

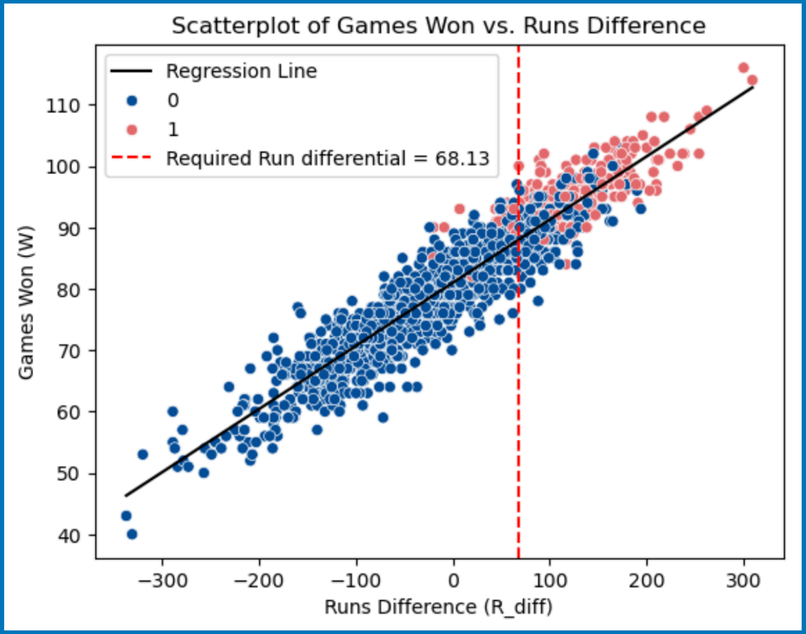
Both target and feature variable are normally distributed

Results

R^2 training data = 0.87, R^2 test data = 0.89

MSE = 18.1 or average prediction error of 4.25 games

Predicted a run differential of 69 (rounded up) for winning 88 games



Method

What run differential is needed to win 88 games?

Univariate linear regression model

target variable: W (games won)

feature: R_diff (runs scored - runs allowed)

Target and feature variables have strong positive correlation (0.94)

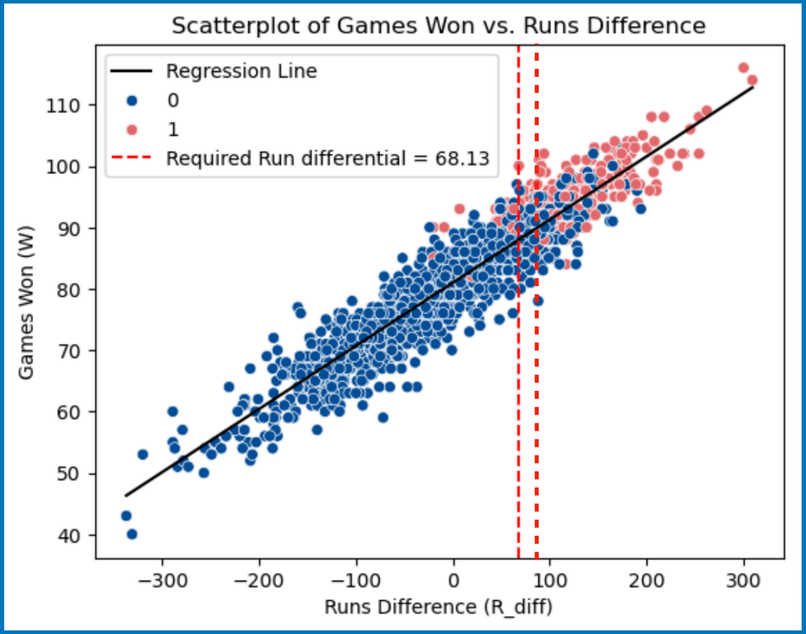
Both target and feature variable are normally distributed

Results

R^2 training data = 0.87, R^2 test data = 0.89

MSE = 18.1 or average prediction error of 4.25 games

Predicted a run differential of 69 (rounded up) for winning 88 games



Best closer to 100 !

Method

What factors contribute to a team achieving a high run differential?

Method

What factors contribute to a team achieving a high run differential?

Multivariate linear regression model

Method

What factors contribute to a team achieving a high run differential?

Multivariate linear regression model

target variable: R_diff

Method

What factors contribute to a team achieving a high run differential?

Multivariate linear regression model

target variable: R_diff

features: BA, OBP, SLG, OOBP, OSLG

Method

What factors contribute to a team achieving a high run differential?

Multivariate linear regression model

target variable: R_diff

features: BA, OBP, SLG, OOBP, OSLG

Team statistics showed positive correlation with target

Method

What factors contribute to a team achieving a high run differential?

Multivariate linear regression model

target variable: R_diff

features: BA, OBP, SLG, OOBP, OSLG

Team statistics showed positive correlation with target

Opposite team stats showed negative correlation with target

Method

What factors contribute to a team achieving a high run differential?

Multivariate linear regression model

target variable: R_diff

features: BA, OBP, SLG, OOBP, OSLG

Team statistics showed positive correlation with target

Opposite team stats showed negative correlation with target

Features were highly correlated, hence feature selection algorithm implemented

Method

What factors contribute to a team achieving a high run differential?

Multivariate linear regression model

target variable: R_diff

features: BA, OBP, SLG, OOBP, OSLG

Team statistics showed positive correlation with target

Opposite team stats showed negative correlation with target

Features were highly correlated, hence feature selection algorithm implemented

Results

Method

What factors contribute to a team achieving a high run differential?

Multivariate linear regression model

target variable: R_diff

features: BA, OBP, SLG, OOBP, OSLG

Team statistics showed positive correlation with target

Opposite team stats showed negative correlation with target

Features were highly correlated, hence feature selection algorithm implemented

Results

With all five features:
R^2 training data = R^2 test data = 0.41

Method

What factors contribute to a team achieving a high run differential?

Multivariate linear regression model

target variable: R_diff

features: BA, OBP, SLG, OOBP, OSLG

Team statistics showed positive correlation with target

Opposite team stats showed negative correlation with target

Features were highly correlated, hence feature selection algorithm implemented

Results

With all five features:
R^2 training data = R^2 test data = 0.41

MSE = 7371 or average prediction error of 86 runs

Method

What factors contribute to a team achieving a high run differential?

Multivariate linear regression model

target variable: R_diff

features: BA, OBP, SLG, OOBP, OSLG

Team statistics showed positive correlation with target

Opposite team stats showed negative correlation with target

Features were highly correlated, hence feature selection algorithm implemented

Results

With all five features:
R^2 training data = R^2 test data = 0.41

MSE = 7371 or average prediction error of 86 runs

```
Model's coefficients:
BA: -33.21272940796096
OBP: 3058.404797159456
SLG: 370.4550729331356
OOBP: -2294.337810737113
OSLG: -786.0395836364761
intercept:  [-45.91184706]
```

Method

What factors contribute to a team achieving a high run differential?

Multivariate linear regression model

target variable: R_diff

features: BA, OBP, SLG, OOBP, OSLG

Team statistics showed positive correlation with target

Opposite team stats showed negative correlation with target

Features were highly correlated, hence feature selection algorithm implemented

Results

With all five features:
R^2 training data = R^2 test data = 0.41

MSE = 7371 or average prediction error of 86 runs

```
Model's coefficients:
BA: -33.21272940796096
OBP: 3058.404797159456
SLG: 370.4550729331356
OOBP: -2294.337810737113
OSLG: -786.0395836364761
intercept:  [-45.91184706]
```

Method

What factors contribute to a team achieving a high run differential?

Multivariate linear regression model

target variable: R_diff

features: BA, OBP, SLG, OOBP, OSLG

Team statistics showed positive correlation with target

Opposite team stats showed negative correlation with target

Features were highly correlated, hence feature selection algorithm implemented

Results

With all five features:
R^2 training data = R^2 test data = 0.41

MSE = 7371 or average prediction error of 86 runs

```
Model's coefficients:
BA: -33.21272940796096
OBP: 3058.404797159456
SLG: 370.4550729331356
OOBP: -2294.337810737113
OSLG: -786.0395836364761
intercept:  [-45.91184706]
```

Feature selection algorithm only dropped BA:
R^2 =0.41, adjusted R^2 =0.40

Method

What factors contribute to a team achieving a high run differential?

Multivariate linear regression model

target variable: R_diff

features: BA, OBP, SLG, OOBP, OSLG

Team statistics showed positive correlation with target

Opposite team stats showed negative correlation with target

Features were highly correlated, hence feature selection algorithm implemented

Results

With all five features:
R^2 training data = R^2 test data = 0.41

MSE = 7371 or average prediction error of 86 runs

```
Model's coefficients:
BA: -33.21272940796096
OBP: 3058.404797159456
SLG: 370.4550729331356
OOBP: -2294.337810737113
OSLG: -786.0395836364761
intercept:  [-45.91184706]
```

Feature selection algorithm only dropped BA:
R^2 =0.41, adjusted R^2 =0.40

Three of the four best features had outliers

Future improvements

Future improvements

- Investigate another classification (NB, SVM) models and consider adding inter-league and intra-division data

Future improvements

- Investigate another classification (NB, SVM) models and consider adding inter-league and intra-division data
- Address multicollinearity with Lasso or Ridge regularisation

Future improvements

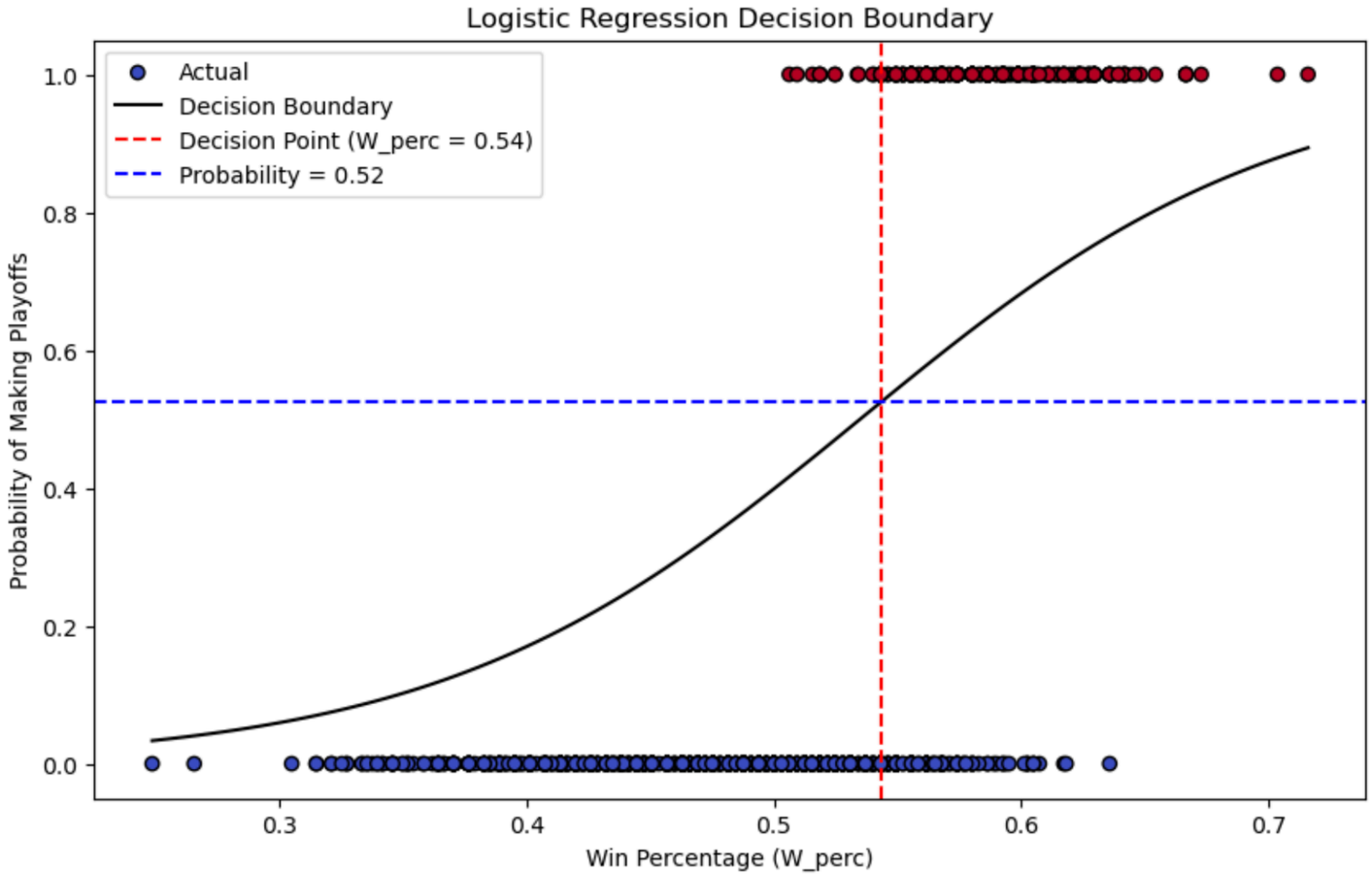
- Investigate another classification (NB, SVM) models and consider adding inter-league and intra-division data
- Address multicollinearity with Lasso or Ridge regularisation
- Reduce the effect of outliers with robust scaling or transformations

Future improvements

- Investigate another classification (NB, SVM) models and consider adding inter-league and intra-division data
- Address multicollinearity with Lasso or Ridge regularisation
- Reduce the effect of outliers with robust scaling or transformations
- Enhance dataset quality by collecting more granular data like advanced team metrics and player-level statistics

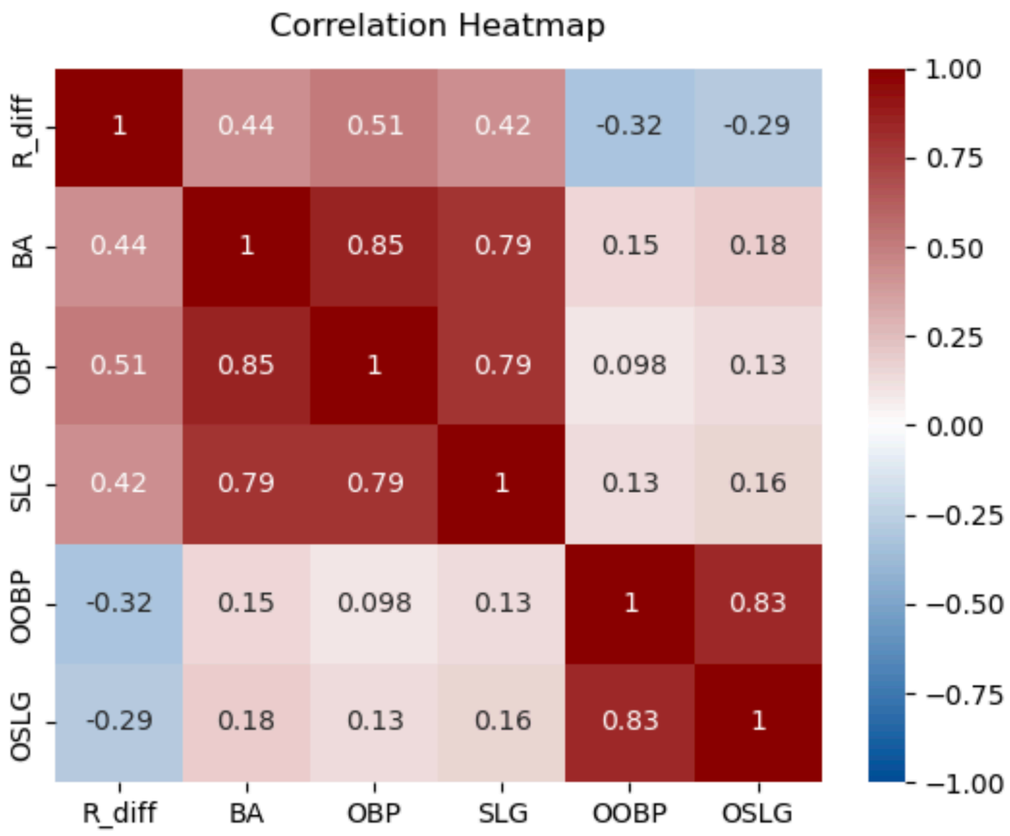
Appendix

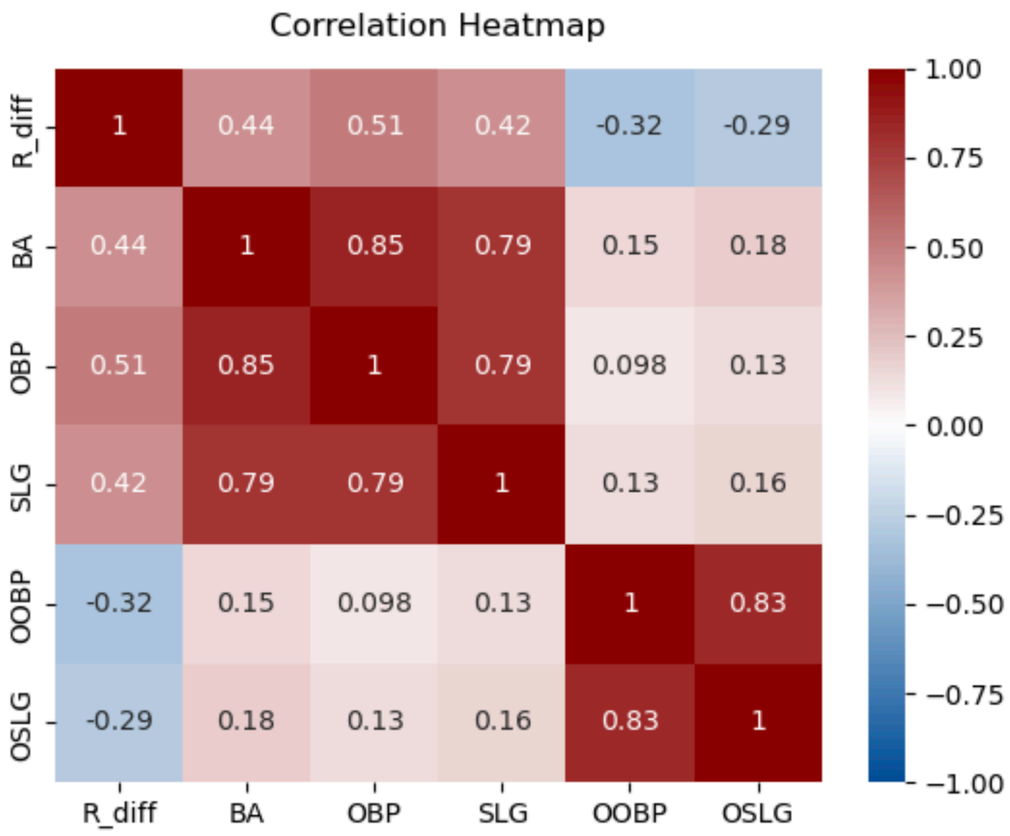
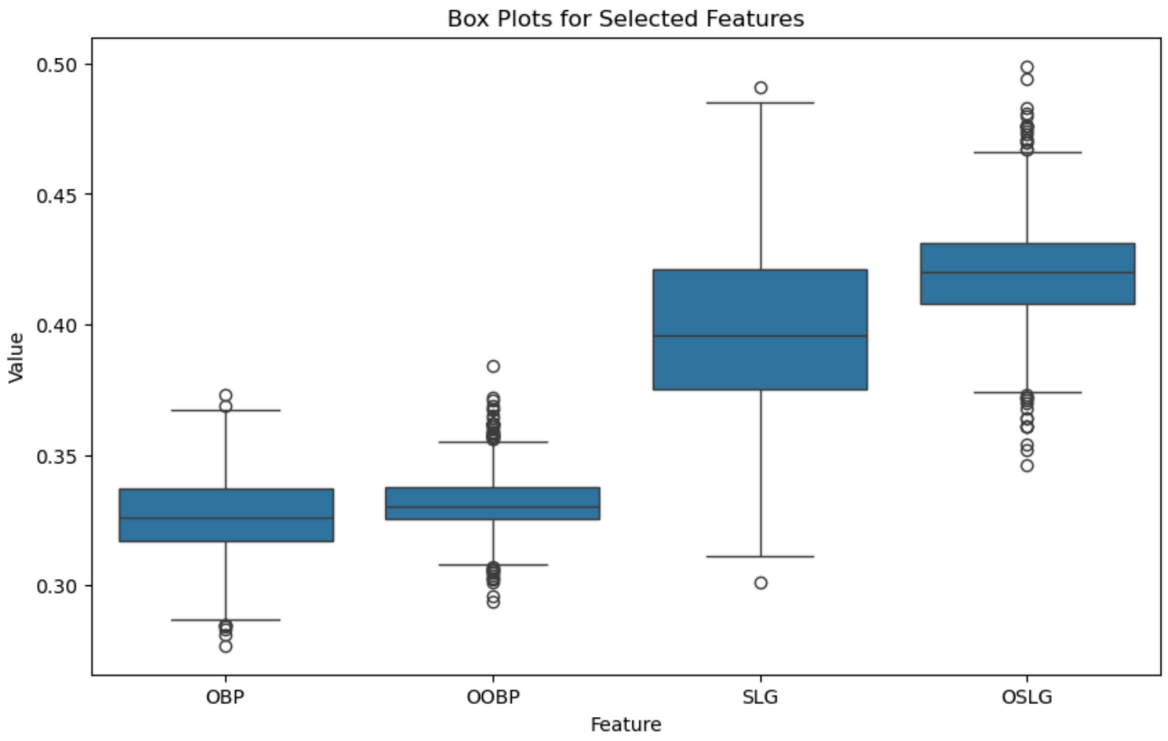
Classification model



Appendix

Multivariate regression model



Appendix

Multivariate regression model