Random Forest Classification Model

Premier League Win-Loss
Prediction
Seasons 2016 - 2018

Project Team

- Dragan Bogatic

- Howard Edwards

- Jordan Gross

- Ryan Reisner

Project goal

- Make accurate predictions about Premier League game Win-Loss based on historical game statistics using Machine Learning Model

Objectives

- Research available sources of historical soccer game data

- Download data in standardized format

- Clean the data

- Select model features

- Select prediction variable

- Transform data to format to fit machine learning model

Objectives cont'd

- Select prediction model

- Train and test the model

- Evaluate prediction performance of the model

- Fine-tune model hyper-parameters to improve model prediction

- Evaluate results

- Conclusions

Resources

- Data source: RapidAPI

- Data: Premier League Seasons 2016 - 2018

- Data format: CSV

- Dataset type: detailed fixture game stats by season game

- Dataset size: approx. 1.4 mil. rows by 33 columns

Machine Learning Model

- Model selected: Balanced Random Forest Classifier (BRF)

- Reduces overfitting

- Proven performer in classification prediction problems

- Works with categorical and continuous values

- Uses Ensemble Learning technique (many weak learners strong together)

Base BRF Model Features and Prediction

- Total features in dataset: 32

- 100 Trees Random Forest

- Train data: seasons 2016-2017

- Test data: season 2018

- Prediction: h_result (Win-Loss)
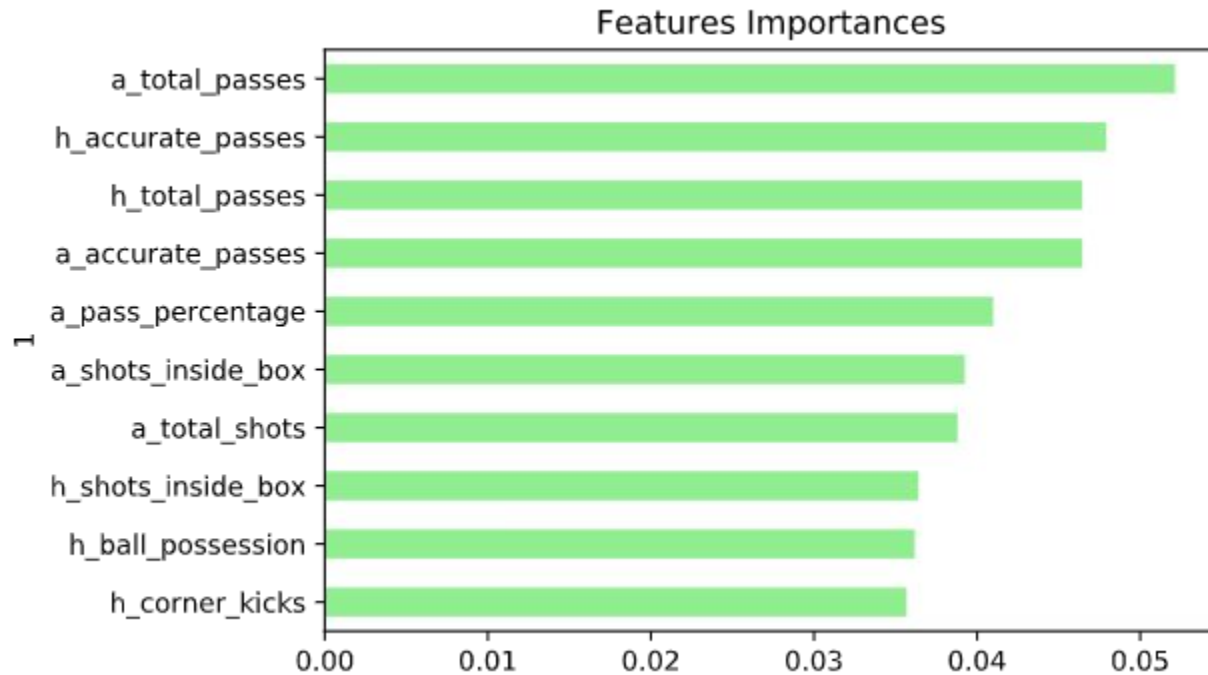
Initial BRF Model Confusion Matrix

- 100 Trees
- 32 Features

	Predicted 0	Predicted 1
Actual 0	129	70
Actual 1	66	115

Base BRF Model Accuracy

`Balanced Accuracy Score : 0.4285714285714286`

Most Important Features

Features Importances

Fine-Tuned BRF Model Features and Prediction

- Total features in dataset: 6 most important from base BRF

- 128 Trees Random Forest

- Features selected for the base model:
  - `h_accurate_passes`
  - `h_total_passes`
  - `h_pass_percentage`
  - `h_ball_possession`
  - `h_total_shots`
  - `h_fouls`

- Prediction: h_result (Win-Loss)

Modified BRF Model Accuracy

Balanced Accuracy Score : 0.6418001610261251

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	129	70
Actual 1	66	115

Classification Report

	f1-score	precision	recall	support
0	0.654822	0.661538	0.648241	199.000000
1	0.628415	0.621622	0.635359	181.000000
accuracy	0.642105	0.642105	0.642105	0.642105
macro avg	0.641619	0.641580	0.641800	380.000000
weighted avg	0.642244	0.642525	0.642105	380.000000

Most Important Features

Features Importances
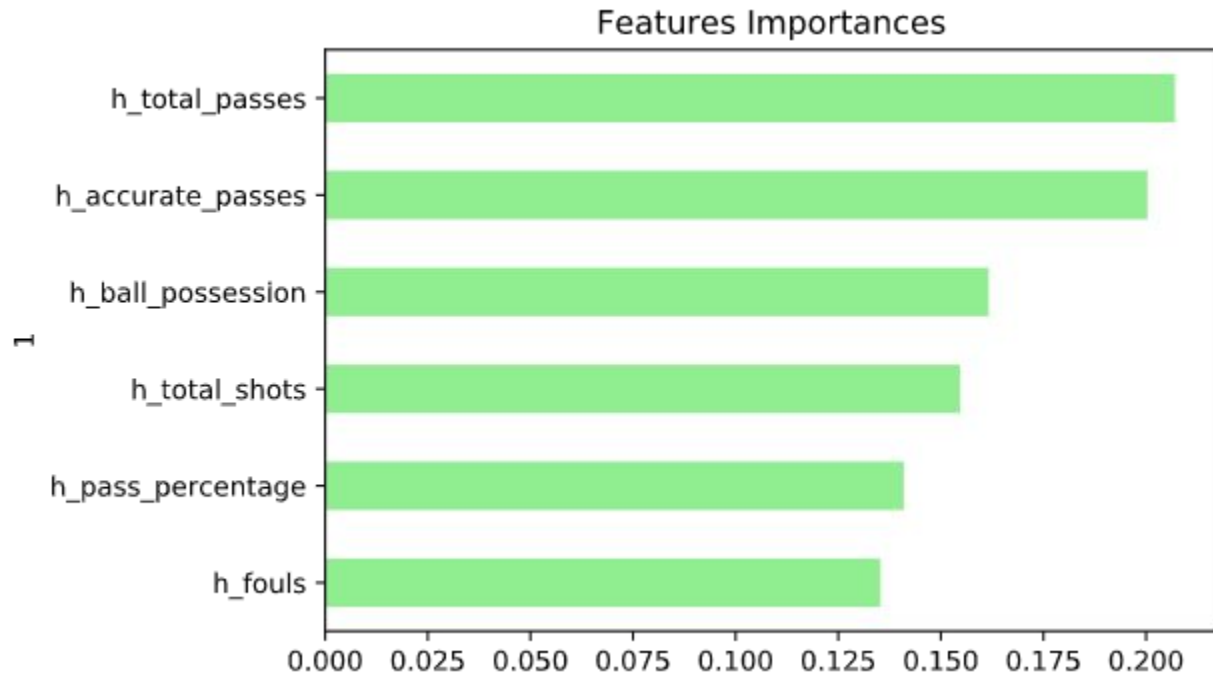
Alternative Classification Models

- Easy Ensemble Classifier

- SMOTEENN Model

Easy Ensemble Classifier Accuracy Score

`Balanced Accuracy Score : 0.608900857880563`

Easy Ensemble Classification Report

```
Classification Report
                 pre       rec       spe        f1       geo       iba       sup

           0    0.62      0.64      0.57      0.63      0.61      0.37       199
           1    0.59      0.57      0.64      0.58      0.61      0.37       181

avg / total    0.61      0.61      0.61      0.61      0.61      0.37       380
```

SMOTEENN Model Accuracy Score

SMOTEEN Balanced Accuracy Score = 0.4908381687442739

SMOTEENN Classification Report

	pre	rec	spe	f1	geo	iba	sup
0	0.52	0.52	0.46	0.52	0.49	0.24	199
1	0.47	0.46	0.52	0.47	0.49	0.24	181
avg / total	0.49	0.49	0.49	0.49	0.49	0.24	380

Conclusions

- Random Forest best performing classification model

- Fine-tuning hyper-parameter (n estimators) improved significantly prediction ability of the Base BRF model

- Longer data history would potentially improve model score

- Alternative classification models underperformed BRF model

Conclusions cont'd

- Overall accuracy of Modified BRF model (0.61) puts prediction

  odds in our favor

- BRF model could benefit from more data history to potentially

  further improve accuracy