

---

# 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
- Reduces overfitting
- Proven performer in classification prediction problems
- Works with categorical and continuous values
- Uses Ensemble Learning technique (many weak learners strong together)

# Model Features and Prediction

- Total features in raw dataset: 32
- 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)



# Model Accuracy

**Balanced Accuracy Score : 0.6418001610261251**

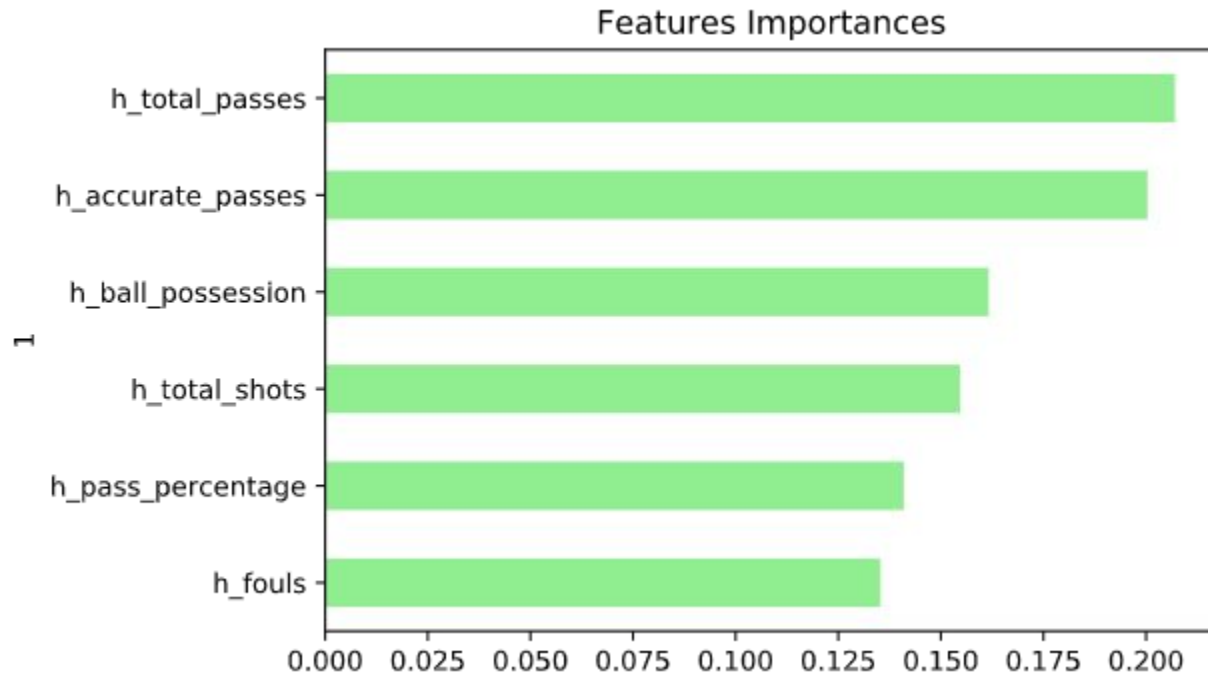
# Confusion Matrix

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

# Classification Report

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

# Most Important Features



# 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