2018
WORLD CUP
Soccer Analysis

Structure

- Objective
- Data
- Visualization
- Data Proceeding And Features
- Model Building and Feature Importance
- Simulation

Objective

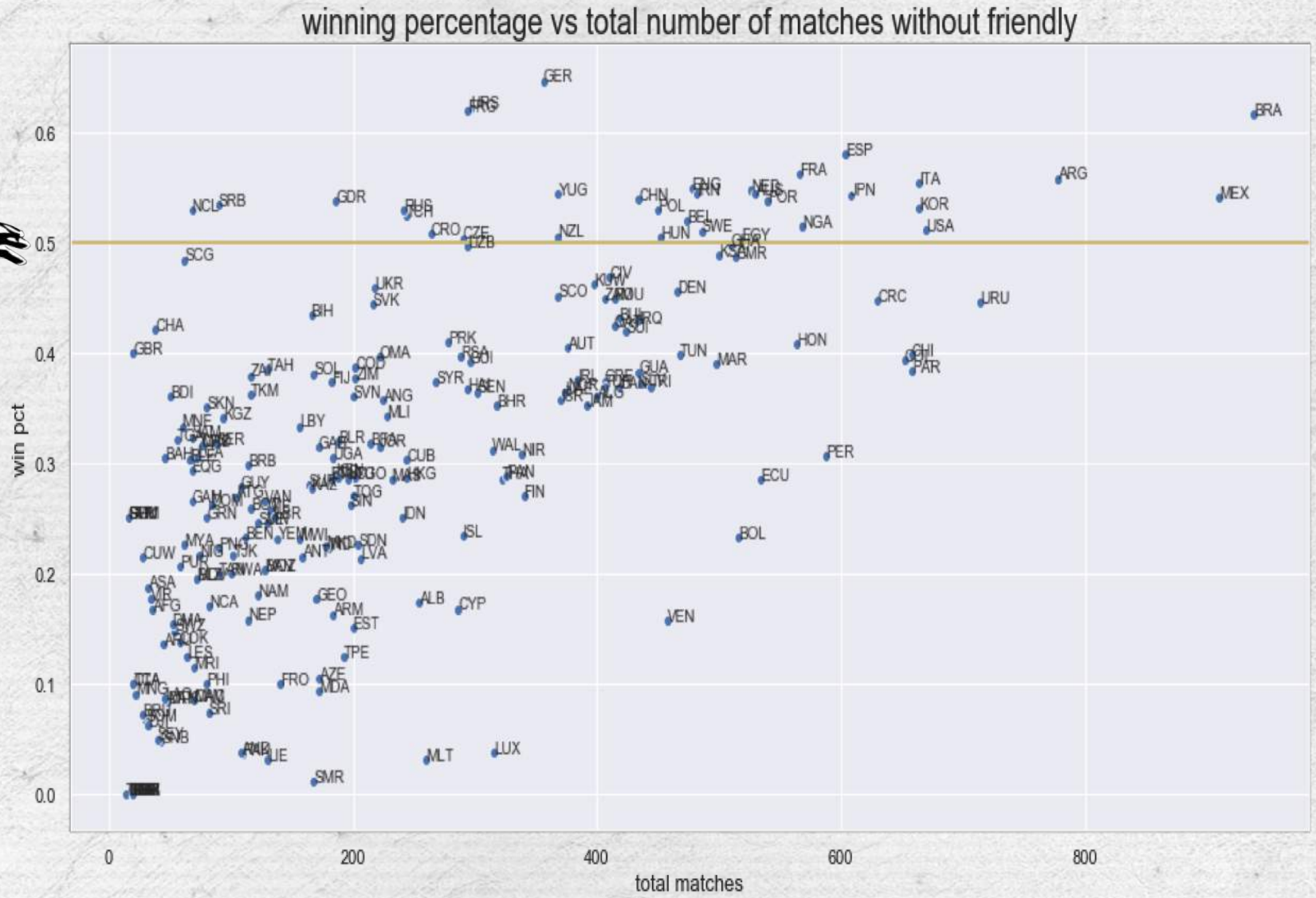
- Investigate factors affecting a soccer team's performance
- Build a Random Forest Regression Model to predict the goal difference of a match
- Simulate 2018 World Cup matches and generate a distribution of possible champions

Data

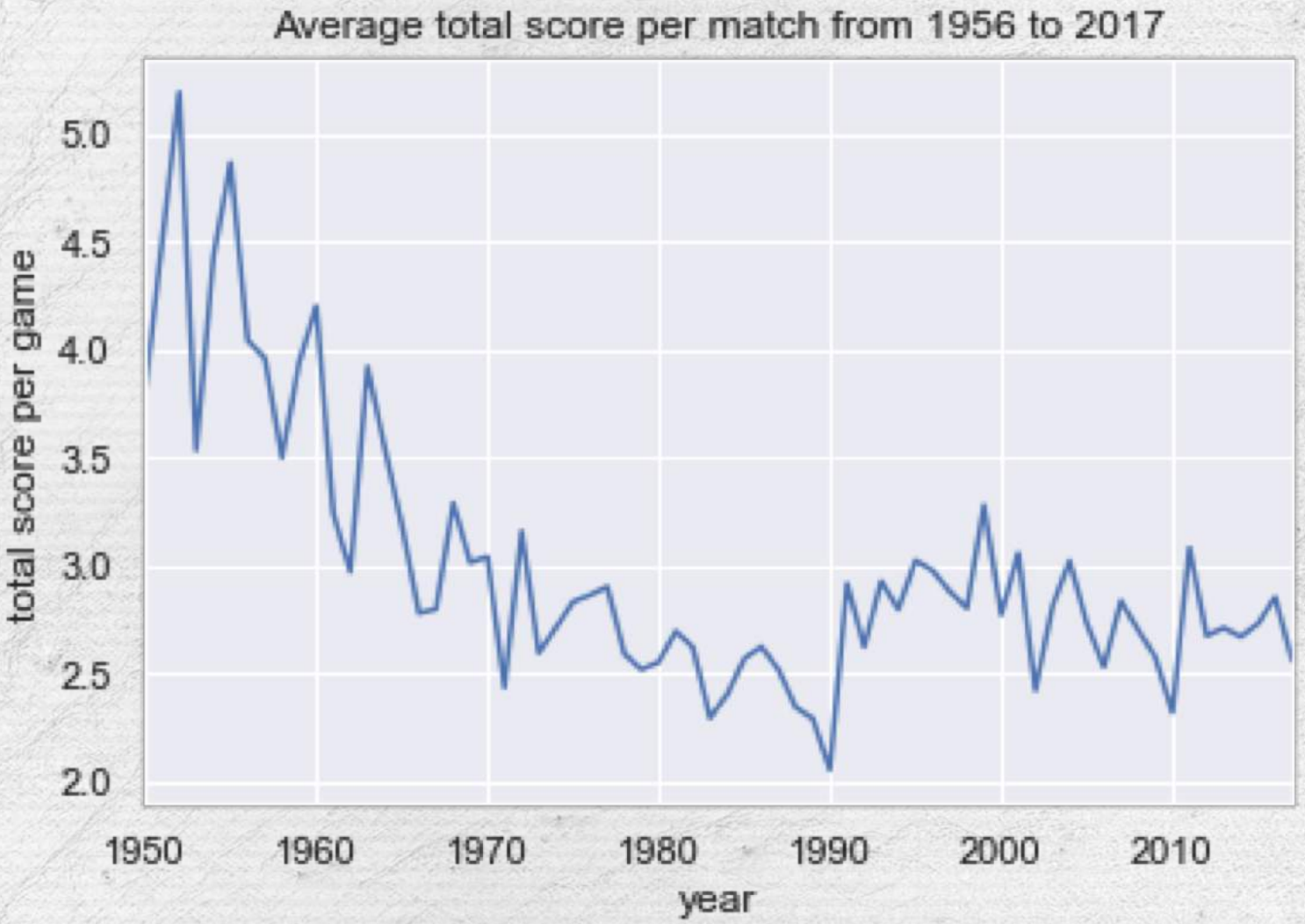
	date	team1	team1Text	team2	team2Text	venue	IdCupSeason	CupName	team1Score	team2Score	statText	resText
0	19500308	WAL	Wales	NIR	Northern Ireland	Cardiff, Wales	6	FIFA competition team qualification	0.0	0.0	NaN	0-0
1	19500402	ESP	Spain	POR	Portugal	Madrid, Spain	6	FIFA competition team qualification	5.0	1.0	NaN	5-1
2	19500409	POR	Portugal	ESP	Spain	Lisbon, Portugal	6	FIFA competition team qualification	2.0	2.0	NaN	2-2
3	19500415	SCO	Scotland	ENG	England	Glasgow, Scotland	6	FIFA competition team qualification	0.0	1.0	NaN	0-1
4	19500624	BRA	Brazil	MEX	Mexico	Rio De Janeiro, Brazil	7	FIFA competition team final	4.0	0.0	NaN	4-0

- **Over 30,000 international soccer matches**
- **From 1950 to 2017**
- **Game results and type of the game (CupName)**

Visualization



Visualization



Data Preprocessing and Features

Features (52 in total):

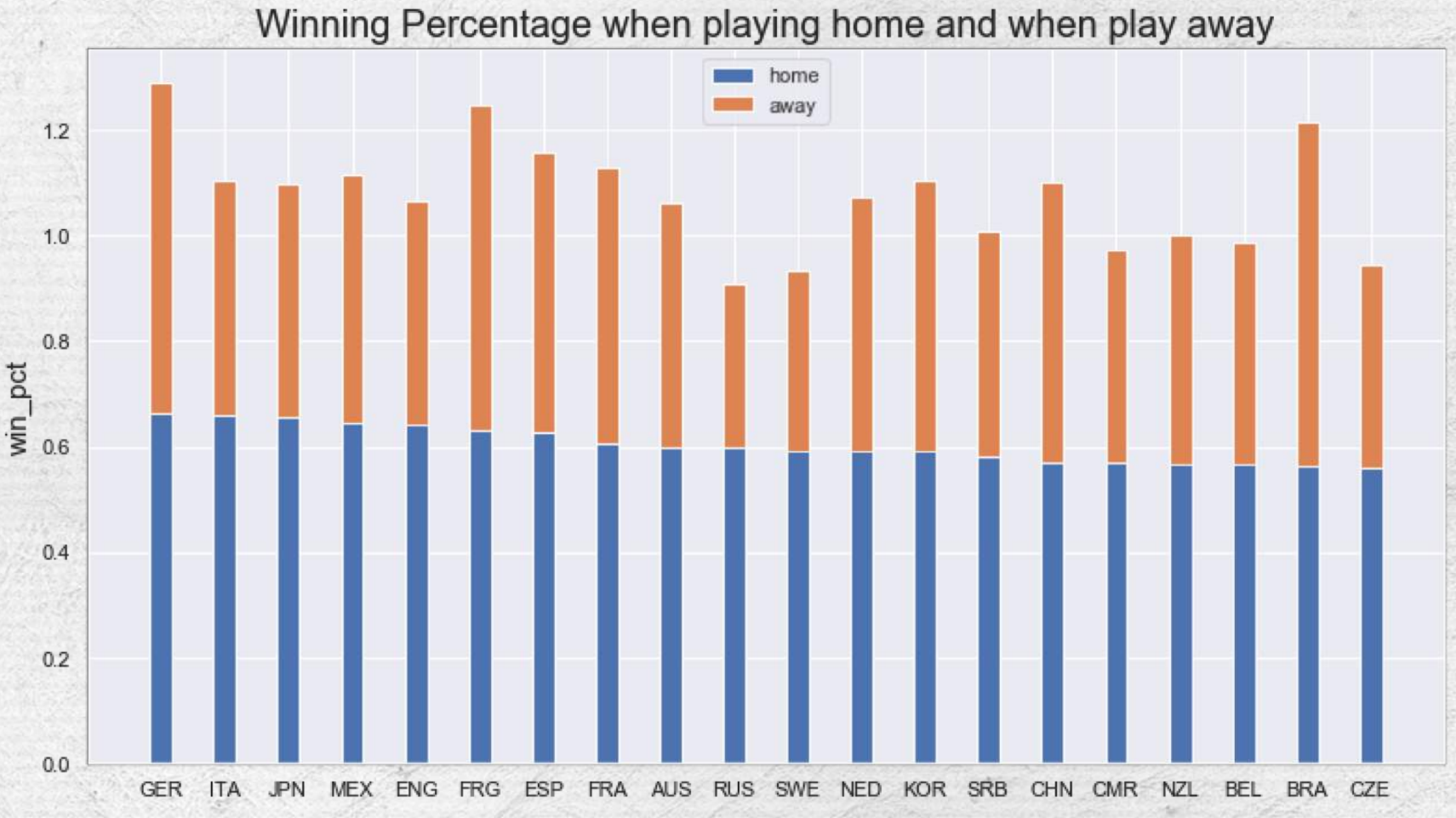
1. Recent Performance (wins, loses, draws etc.)
2. Home Advantages
3. League Strength (regional strength):

Europe: 1.0 // South America: 1.0 // North America: 0.88 // Asia: 0.86 // Africa: 0.86 // Oceania: 0.85

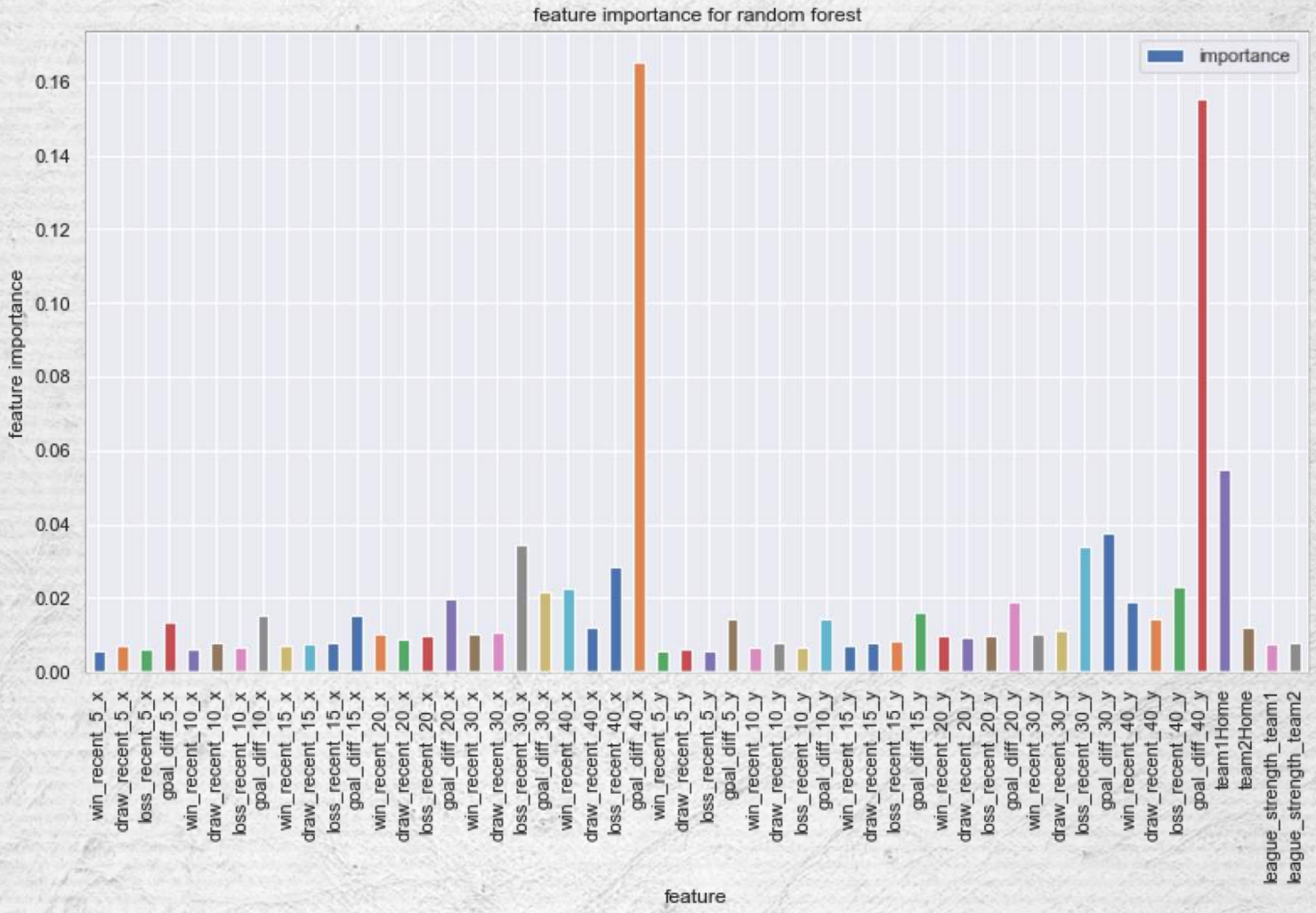
	win_recent_5_x	draw_recent_5_x	loss_recent_5_x	goal_diff_5_x	win_recent_10_x	draw_recent_10_x	loss_recent_10_x	goal_diff_10_x	win_recent_15_x
0	0.6	0.2	0.2	0.6	0.4	0.5	0.1	0.7	0.466667
1	1.0	0.0	0.0	3.2	0.9	0.1	0.0	3.3	0.800000
2	1.0	0.0	0.0	2.0	0.8	0.2	0.0	1.8	0.800000
3	0.6	0.4	0.0	1.2	0.5	0.3	0.2	0.5	0.533333
4	0.6	0.4	0.0	1.6	0.8	0.2	0.0	2.3	0.800000

draw_recent_15_x	...	team1Home	team2Home	league_strength_team1	league_strength_team2	goal_diff	team1	team2	team1Text	team2Text	result
0.333333	...	True	False	1.0	1.00	0.0	WAL	NIR	Wales	Northern Ireland	0
0.066667	...	True	False	1.0	1.00	4.0	ESP	POR	Spain	Portugal	1
0.133333	...	True	False	1.0	1.00	0.0	POR	ESP	Portugal	Spain	0
0.200000	...	True	False	1.0	1.00	-1.0	SCO	ENG	Scotland	England	2
0.200000	...	True	False	1.0	0.88	4.0	BRA	MEX	Brazil	Mexico	1

Home Advantage



Model Building and Feature Importance



- The best number of trees is: 100
- The square root of mean squared error on test set: 1.982
- The mean absolute error on test set: 1.539

- Goal Difference Accuracy: 0.26020607995929756
- Win-Draw-Lose Accuracy: **0.5389**

Simulate 2018 Russia World Cup

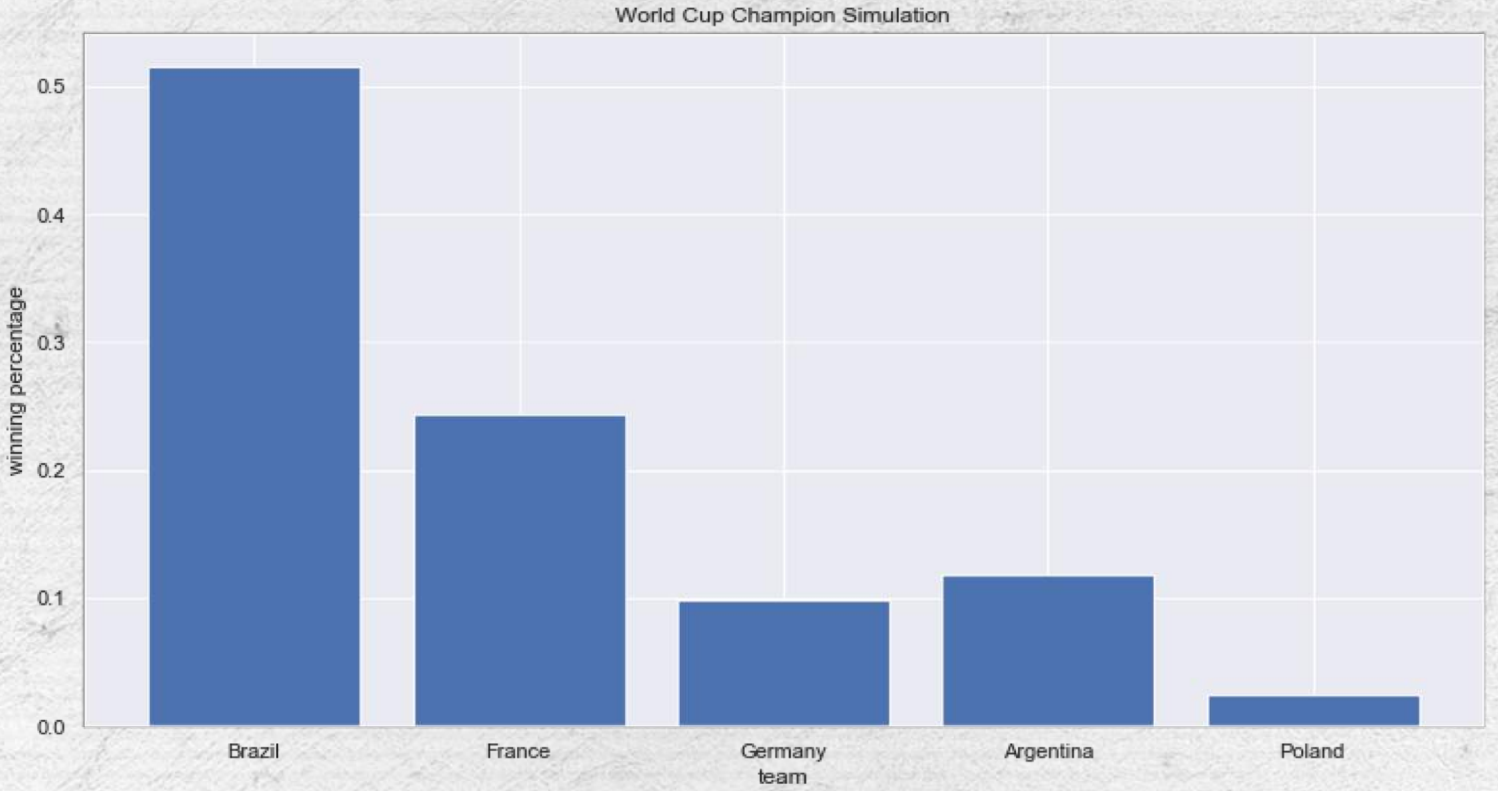
1. Group Stage (32 → 16)
2. Knockout Stages
   1. Round of 16
   2. Quarter finals
   3. Semi finals
   4. Final

	team1	team2	result
0	Russia	Saudi Arabia	win
1	Egypt	Uruguay	lose
2	Morocco	Iran	draw
3	Portugal	Spain	lose
4	France	Australia	win
5	Argentina	Iceland	win

```
{'A': {'Russia': 7, 'Saudi Arabia': 3, 'Egypt': 0, 'Uruguay': 7},
 'B': {'Portugal': 6, 'Spain': 9, 'Morocco': 1, 'Iran': 1},
 'C': {'France': 7, 'Australia': 4, 'Peru': 2, 'Denmark': 2},
 'D': {'Argentina': 9, 'Iceland': 0, 'Croatia': 6, 'Nigeria': 3},
 'E': {'Brazil': 9, 'Switzerland': 4, 'Costa Rica': 1, 'Serbia': 2},
 'F': {'Germany': 9, 'Mexico': 1, 'Sweden': 4, 'Korea Republic': 2},
 'G': {'Belgium': 6, 'Panama': 0, 'Tunisia': 3, 'England': 9},
 'H': {'Poland': 7, 'Senegal': 0, 'Colombia': 6, 'Japan': 4}}
```

Simulate 2018 Russia World Cup

```
{'Brazil': 5155,
 'France': 2439,
 'Germany': 987,
 'Argentina': 1178,
 'Poland': 241}
```



Model Building and Feature Importance

Thanks!