Soccer Predictive Analytics

By Chizoba Obasi

Motivation

The person that said winning isn't everything, never won anything.
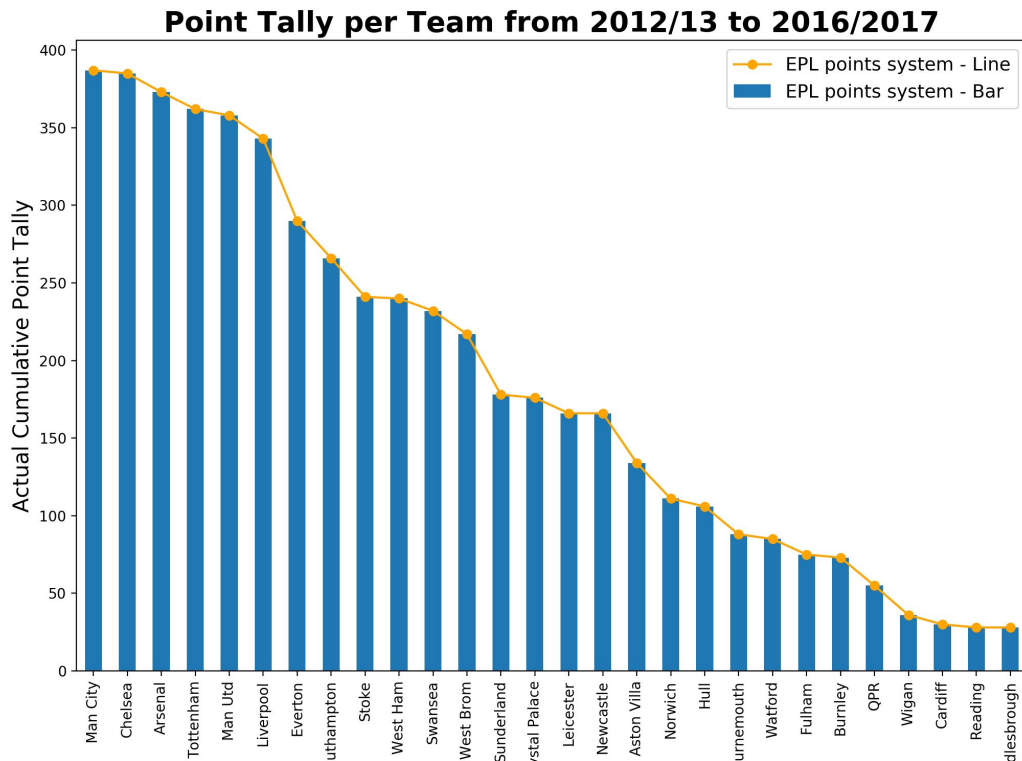
Mia Hamm
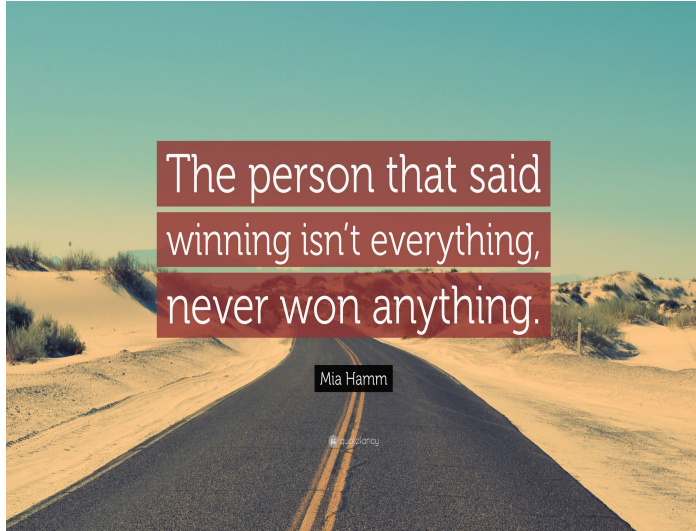
- **English Premier League (EPL)**
  - **Most watched soccer league**
  - **Most profitable soccer league**
  - **212 territories**
  - **643 million homes**
  - **Potential TV audience of 4.7 billion people**
  - **€ 2.2 billion/yr in domestic and int'l rights**
- **High level of uncertainty, unpredictability and variability in soccer**
- **Sports betting industry - a major one**

Observations - "BIG" team consistency

Point Tally per Team from 2012/13 to 2016/2017

Legend:
- EPL points system - Line
- EPL points system - Bar

Y-axis: Actual Cumulative Point Tally

X-axis teams: Man City, Chelsea, Arsenal, Tottenham, Man Utd, Liverpool, Everton, uthampton, Stoke, West Ham, Swansea, West Brom, Sunderland, rstal Palace, Leicester, Newcastle, Aston Villa, Norwich, Hull, urnemouth, Watford, Fulham, Burnley, QPR, Wigan, Cardiff, Reading, dlesbrough

- Manchester City
- Chelsea
- Arsenal
- Tottenham
- Manchester United
- Liverpool

❖ EPL point system
  ➢ Win - 3 points
  ➢ Draw - 1 point
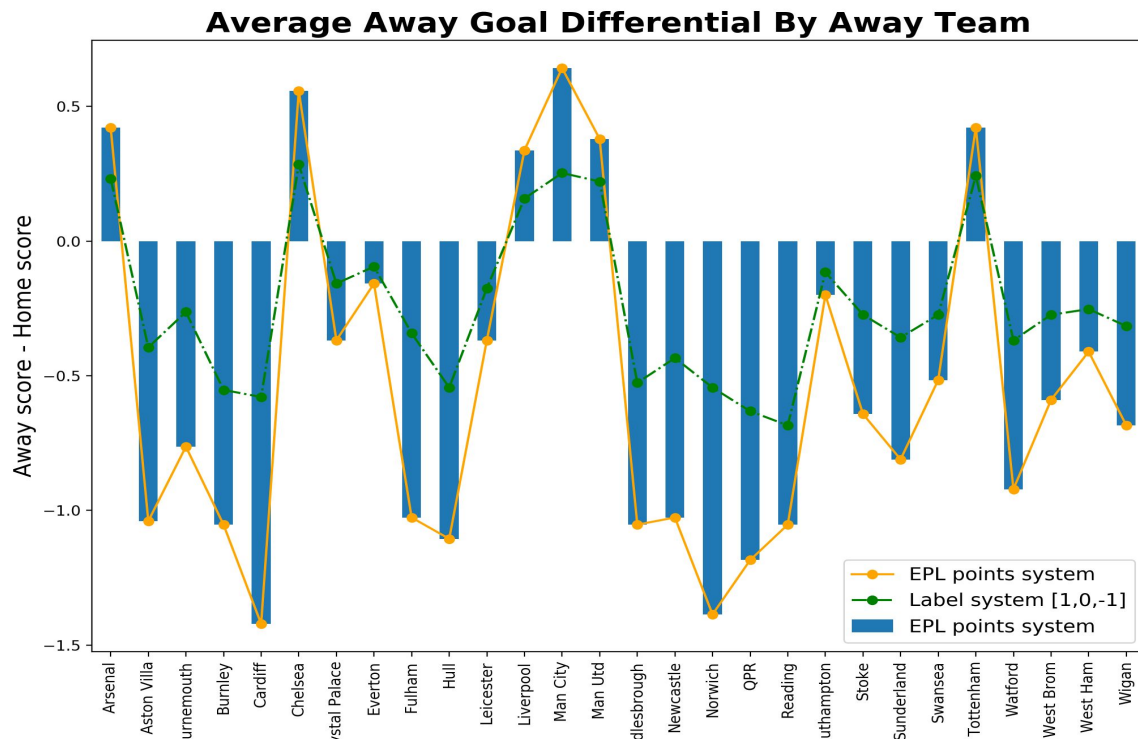 ➢ Loss - 0 points

Observations - Home Advantage

Average Home Goal Differential By Home Team
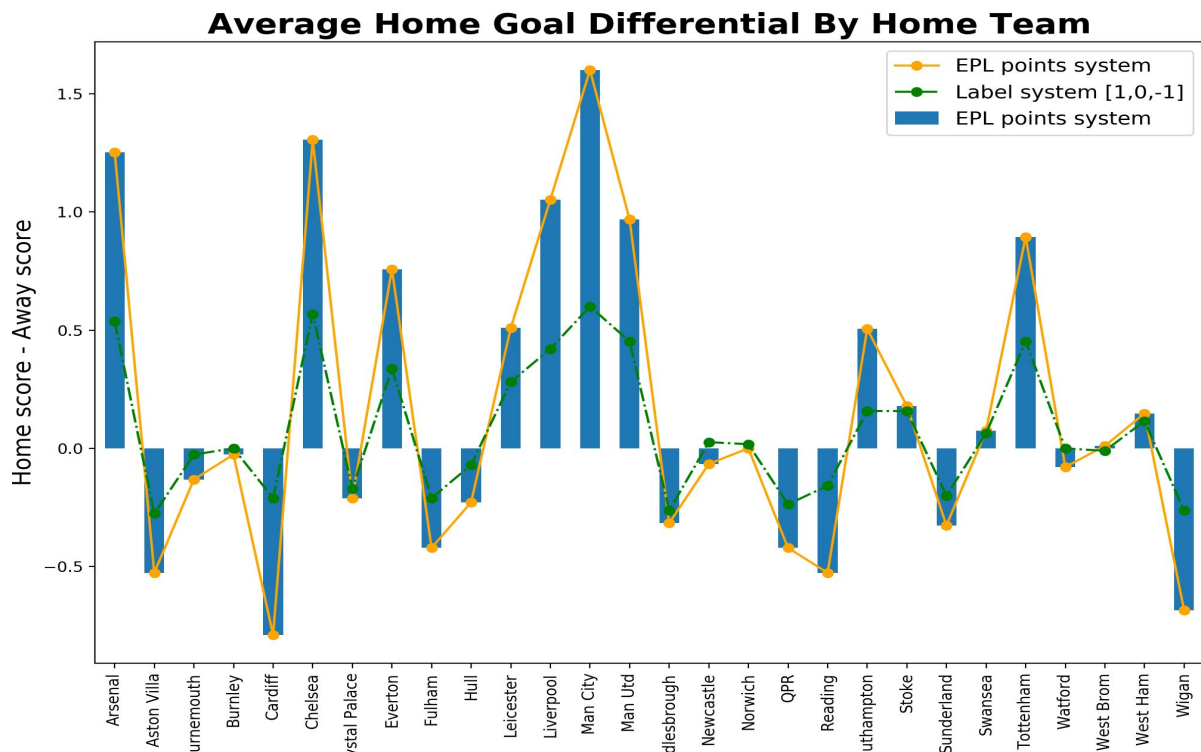
❖ EPL point system
  ➢ Win: 3 points
  ➢ Draw: 1 point
  ➢ Loss: 0 points

❖ Label Class system
  ➢ Win: 1
  ➢ Draw: 0
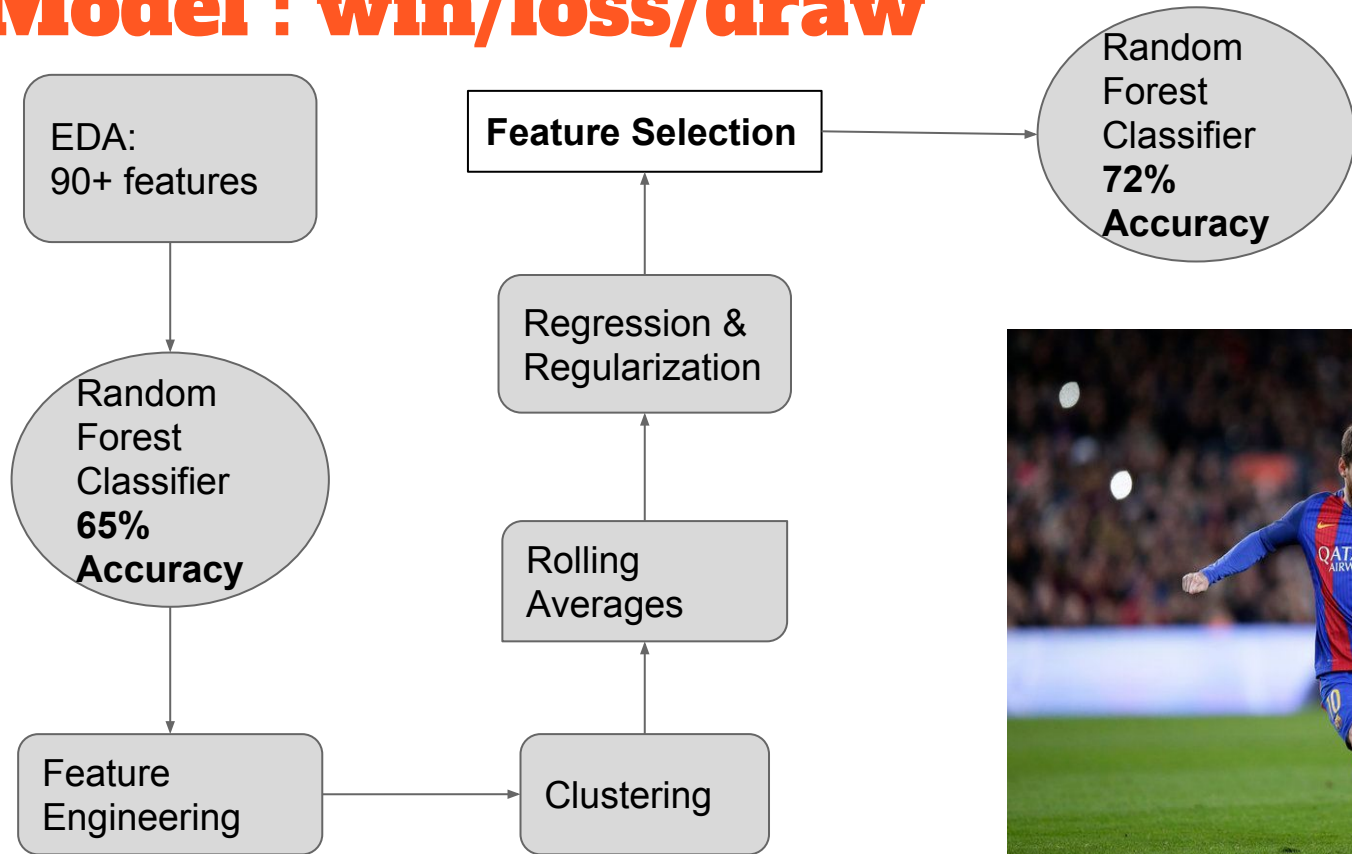 ➢ Loss: -1

Observations - "BIG" team effect

Average Away Goal Differential By Away Team

- ❖ EPL point system
  - ➢ Win: 3 points
  - ➢ Draw: 1 point
  - ➢ Loss: 0 points

- ❖ Label Class system
  - ➢ Win: 1
 - ➢ Draw: 0
  - ➢ Loss: -1

Model : win/loss/draw

EDA:
90+ features

Random
Forest
Classifier
65%
Accuracy

Feature
Engineering

Clustering

Rolling
Averages

Regression &
Regularization

Feature Selection

Random
Forest
Classifier
72%
Accuracy

Model Results

Random Forest Classification Prediction Accuracy Per Class

%	Predicted		
TRUE	**Loss**	**Draw**	**Win**
Loss	**80.70**	14.04	5.26
Draw	21.28	**36.17**	42.55
Win	5.23	8.14	**86.63**

Technologies
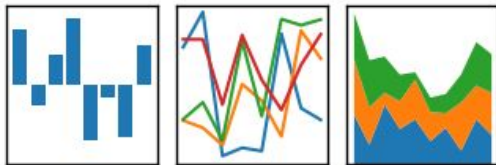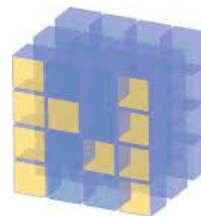
Conclusion

Recommendations

★ **Feature selection** is very important for modeling improvement
★ More **informed feature engineering**
★ **Larger dataset** - more concrete and justifiable model selection
★ **Individual player statistics** and impact on team tactics, and results
  ○ "Star" player effect
  ○ Player injuries, tiredness/rest
★ Accuracy evaluation based on **high win/loss chance** (>= +/- 2 goal differentials)

© CanStockPhoto.com - csp47415763

Questions ?

linkedin.com/in/chizoba-obasi

co14@utexas.edu