Prediction of Soccer Games

FootWizard – Data Sorcerers

Abhi Savaliya
asavaliy@sfu.ca

Chirayu Shah
cshah@sfu.ca

Sagar Parikh
parikh@sfu.ca

1. Introduction

The fanbase of Soccer games is humongous. There are different leagues followed by Football fans. There are more than 20 leagues in the European Football World alone! However, we aim to focus on EPL (English Premier League) - the most followed league. It has 20 teams including Chelsea, Liverpool, Manchester City etc. We chose this league because it's one of the most unpredictable league. Regular player transfers, injuries, training and various other features affect the statistics of players are quite. Even the team considered as underdogs can win the league (Leicester City - 2016) whereas the best team may lose the most. Team squads change every year which means squad & form of 2008 team might be completely different as compared to 2019 team which makes it nearly impossible to predict the future matches. This led us to build FootWizard - an interactive & accurate Football Match Outcome Prediction tool.

2. Motivation and Background

- There is a huge war going on among the football fans. “Who will Win Today’s Match?” - Asking this question can lead to a war among the football fans. One may say we got the best defence while other with best attack. Even a single comment can lead to a war. Hence, it shows that the football is one of the most followed sport. It has huge fan base which includes all riches as well as poors including celebrities.
- We, being huge fans of football, were quite interested in this project. We have checked many websites where we can predict the outcomes of the matches. We also aimed to recommend the best betting platform to maximize the betting profits on the matches. However, predicting matches is a complex task. We aimed at predicting the betting outcomes, however, we aim it as a future scope as it requires real-time predictions.
- We have chosen to work with English Premier League as it’s the toughest league with no fixed favorites. Like in 2015-16 season, Leicester City who were considered underdogs of the league won it!
- With recent advancements in the Machine Learning world, we now have the chance to try and predict the outcome of a match with innovative methods.

3. Problem Statement

As discussed earlier, Football is one of the most followed as well as unpredictable sport around the world. Being huge fans of the game ourselves we are really enthusiastic to showcase our product. With recent advancement in the Machine Learning domain is it possible to make accurate match predictions. As this sport has a huge fan-base, more questions, more data to work with. As a result, we aim to answer few of the following questions:

- Is it possible to predict which team is going to win the match?
- Which team has the most possible chance to win?
- Is it possible to use machine learning models to predict the outcomes?
- Can we achieve a higher accuracy through the machine learning models?
- Do we really need a huge dataset to predict the outcomes?
- What are the key features in predicting the matches?
- Does more features help in getting higher accuracy?

We had the following challenges:

- There is an abundance of features and feature selection is a very challenging task. For example, Chelsea FC's 2010 team is totally different from 2019's squad. So to predict this, the feature engineering plays a crucial role.
- Algorithm selection - A variety of classification algorithms are available. Yet to select the best, we need to try and test everything.
- Requires to update the statistics of each player as their performance is increased. Getting the most recent data is quite challenging.
- Total number of features available are 114. This would basically create a complex model which will surely overfit. Hence, feature selection is quite important task.
- The size of the dataset is quite small since each season, each team would play against a particular opponent only twice (one home and one away match). Thus, sample size is quite small and there are high chances of overfitting. We have to avoid that by carefully engineering features and eliminating overfitting characteristics.
- Scraping the Latest Data, as the dataset provided ranges from 2008-2016. Scraping is easy but making it consistent with the historical data is tough. For instance, the new match data for 2017 and 2018 season did not include the team squads. Thus, we had to separately fetch those from another data source and then merge both of them together to create a consistent data source for the model.

4. Data Science Pipeline

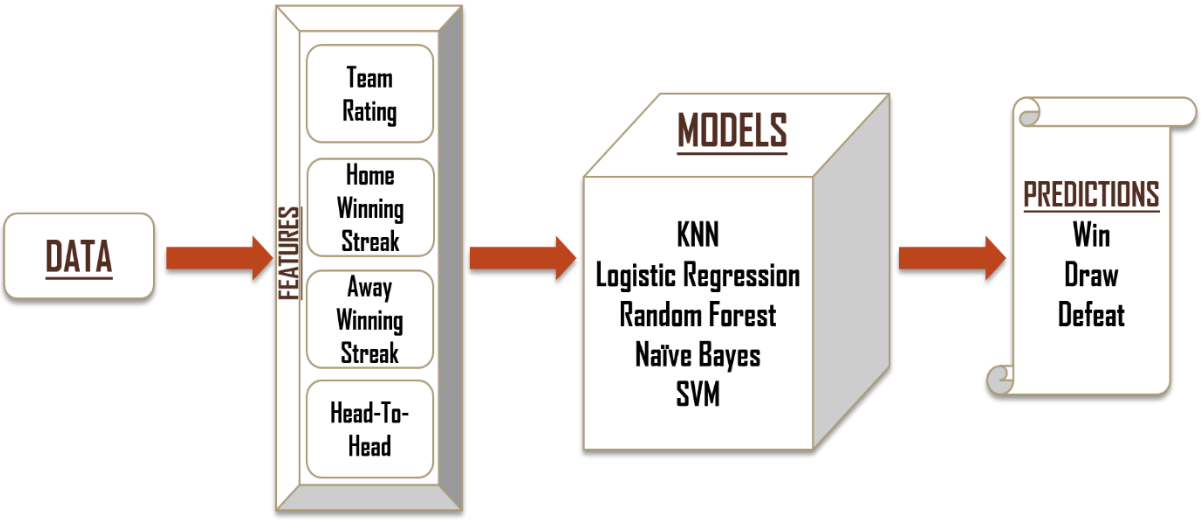
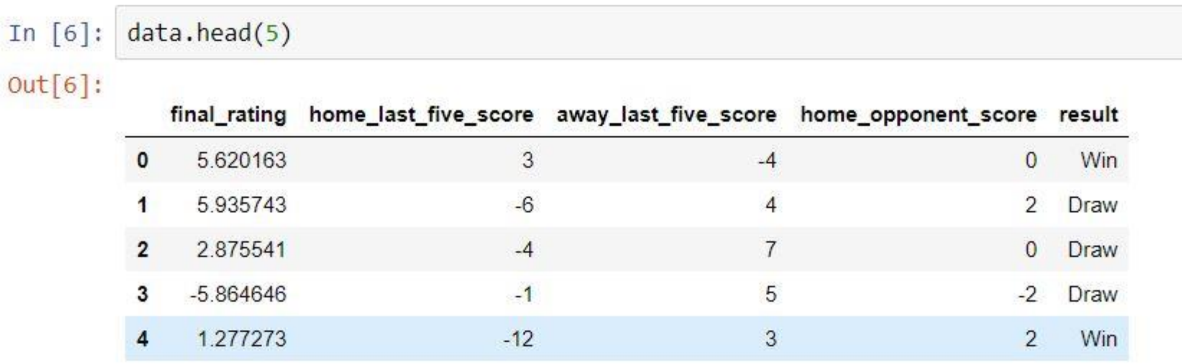


Figure 1: Data Science Pipeline

4.1 Data integration: The first step in the pipeline is to collect the required data from multiple sources. However, collection can be easy task but making it consistent to a single data is quite challenging. Some may have different IDs while others may be missing values. Hence, consistency is challenged. We handled the missing values for squads of home and away teams. We also take the latest available player rating to take the average team rating. The sources: FIFA19, Kaggle and few small repositories.

4.2 Feature Scaling : Team ratings, Home and away streaks and Head-to-head performances all are in different ranges. This does not play well with the model since the model would give more importance to the heavier weighted features. To combat this, we perform feature scaling on all the features before the training and prediction process.

Here are the final features and the output label(result) that are used for modelling. We can see that all of them are in different ranges and it is important to bring them to the same scale.

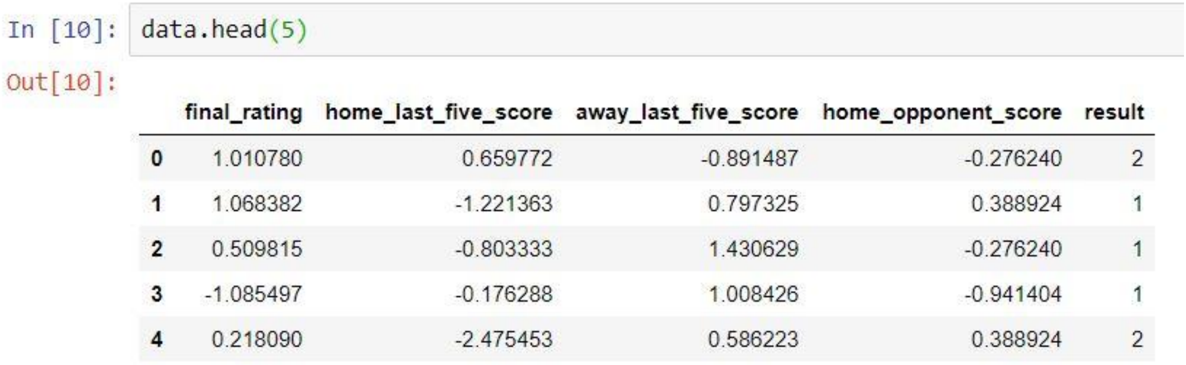


```
In [6]: data.head(5)
Out[6]:
```

	final_rating	home_last_five_score	away_last_five_score	home_opponent_score	result
0	5.620163	3	-4	0	Win
1	5.935743	-6	4	2	Draw
2	2.875541	-4	7	0	Draw
3	-5.864646	-1	5	-2	Draw
4	1.277273	-12	3	2	Win

Figure 2: Final Features included in the model

Here are the features after scaling them to a range of -1 to 1. Also, the result column has been label encoded so that {Win,Draw,Defeat} is subsequently converted to {2,1,0} values so that the model can map the values.



```
In [10]: data.head(5)
Out[10]:
```

	final_rating	home_last_five_score	away_last_five_score	home_opponent_score	result
0	1.010780	0.659772	-0.891487	-0.276240	2
1	1.068382	-1.221363	0.797325	0.388924	1
2	0.509815	-0.803333	1.430629	-0.276240	1
3	-1.085497	-0.176288	1.008426	-0.941404	1
4	0.218090	-2.475453	0.586223	0.388924	2

Figure 3: After feature scaling and Label Encoding

4.3 Feature engineering: The core of any machine learning model is smart feature engineering. This plays an immense role in the performance of the model. Developing and taking in smart features as the input data is of the utmost importance, specially in this case where we cannot take the post-match data of each match such as features because our goal is to predict the game before it's completed. For feature engineering, we used statistical methods to select the right choice which the highest impact over the predictor. We made use of 95% confidence level (p-value 0.05) to get the right features from the 114 attributes.

a. Team Ratings -

- Team strength is a huge deciding factor in the outcome of the match for any team. To quantify the strength of the team for our model, we take the playing 11 for both home and away teams and take an average of their latest overall ratings obtained from the FIFA dataset. The overall rating is the combination of every aspect of the player such as attack, defense, and so on for each season.
- The FIFA dataset contained multiple ratings values for each player. So we decided to take the latest rating of each player available before the match in the dataset. This way, we input the most updated team strength feature for each match row.
- Also, this eliminates a strong influence of any one individual in favor of the average of the whole team. That is, the model is only concerned with the ratings of the players and not the players themselves. This means, technically that if a player is replaced by someone with the same rating, the team strength remains the same as far as model is concerned.
- At last, we take the difference of home team rating and away team rating and input this as a feature. This means, if this difference is positive, the home team is stronger than the away team player strength wise.

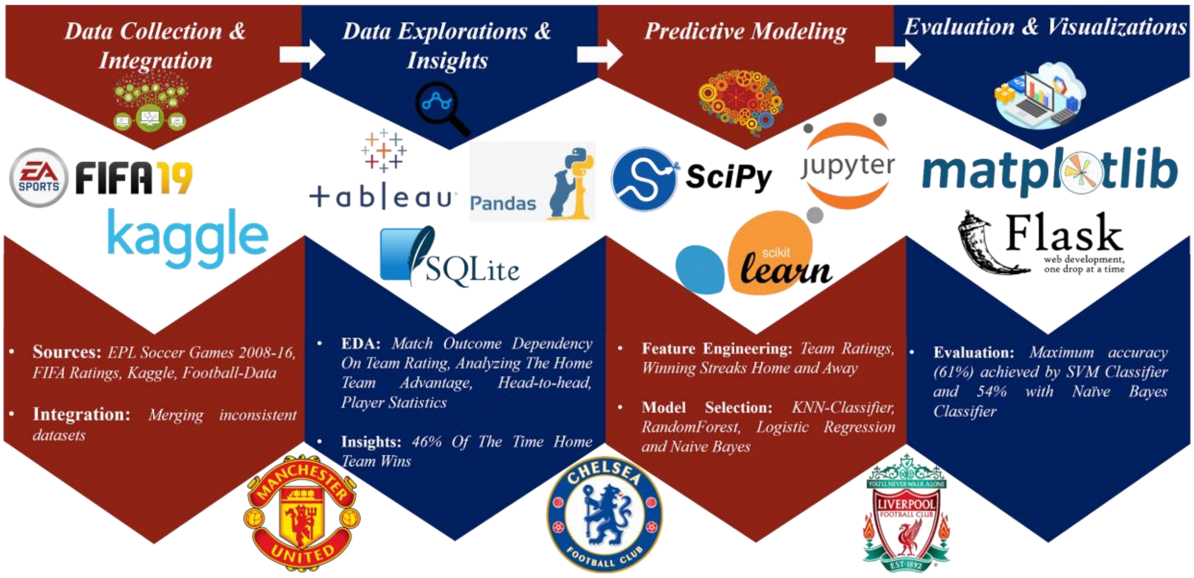
b. Winning Streak -

- Winning momentum plays an important factor in deciding the outcome of a match. If a team has won last five matches, then there are high chances that it would win the next one too.
- For this feature, we consider the last five matches played by the team and we take the sum of goal difference (to factor the impact of win/defeat) as a feature.
- This is done for both home and away teams. For eg. if the score-line of the last five matches of the home team was 5-4, 4-2, 2-1, 1-4, 3-5 then the value of the feature would be 1+2+1+(-3)+(-2)=-1.

c. **Head-to-Head -**

- Head-to-Head performance sometimes outplays the winning streak or player strength of team. If a team has better performance against a team, then again it plays a crucial role in predicting the outcome as it shows the best form of a team against a specific.
- This feature is created by fetching the last five matches played between the same opponents at that venue.
- The sum of difference of goals (home goals - away goals) for these last five matches is taken as a feature in the model.

5. Methodology - Tools & Technologies



Data Collection & Integration

- **Sources:** *EPL Soccer Games 2008-16, FIFA Ratings, Kaggle, Football-Data*
- **Integration:** *Merging inconsistent datasets*

Data Explorations & Insights

- **EDA:** *Match Outcome Dependency On Team Rating, Analyzing The Home Team Advantage, Head-to-head, Player Statistics*
- **Insights:** *46% Of The Time Home Team Wins*

Predictive Modeling

- **Feature Engineering:** *Team Ratings, Winning Streaks Home and Away*
- **Model Selection:** *KNN-Classifier, RandomForest, Logistic Regression and Naive Bayes*

Evaluation & Visualizations

- **Evaluation:** *Maximum accuracy (61%) achieved by SVM Classifier and 54% with Naïve Bayes Classifier*

1. Data Collection and Integration

- **Kaggle - We utilized Kaggle as our data source**

European soccer database which comprised of match data and player dataset from 2008-2016. This was our main source of data which consisted of almost 25,000 matches with 10,000 players. The Players and Teams' attributes were sourced from EA Sports' FIFA video game series, including the weekly updates. Betting Odds were also a part of the data from 10 providers with other detailed match events such as goal, possession, corner, cross, fouls, cards.

- **Kaggle FIFA Dataset 17 & 18 – Latest Players' Data**

The the European soccer database was huge. Soccer is a sport with continuously changing team dynamics. E.g. If a player was playing at Chelsea in 2014 it might not be necessary that he still plays there. Team ratings are also one of the most important factor in prediction of the winners and getting the latest Player Attributes was of utmost importance. The latest player dataset was available from Kaggle for 2017 and 2018.

- **Football.co.uk – Latest Matches data 2017 – 2018**

We are working with predicting Premier League matches for our project. The European match dataset we did not consist of the latest match and player data. We got the latest premier league match data from Football.co.uk.

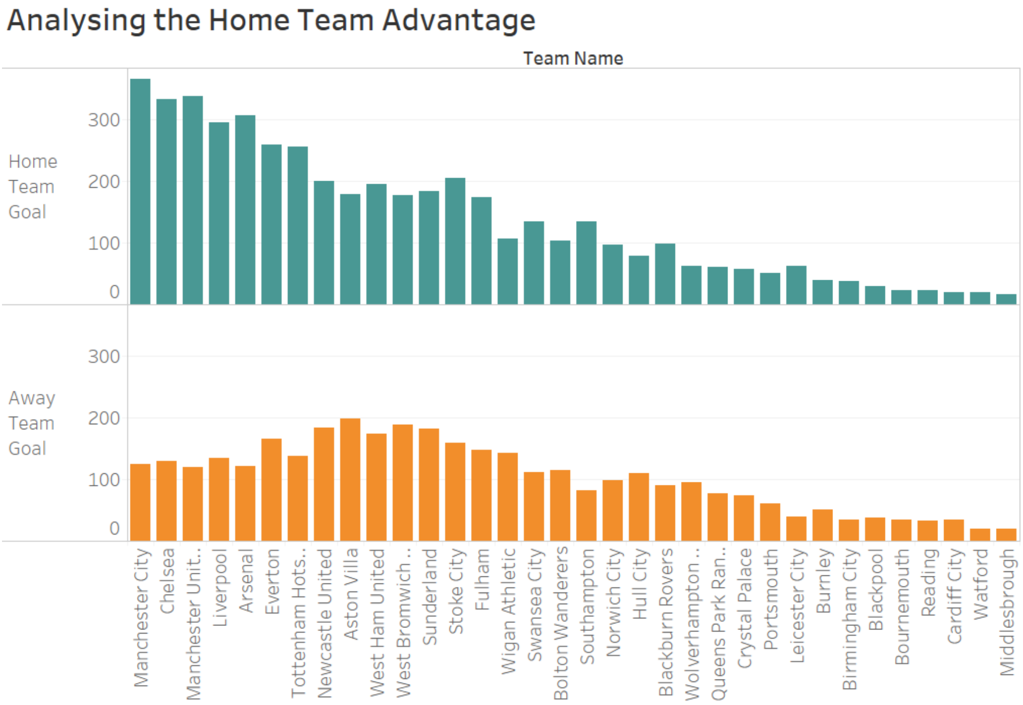
We integrated all these datasets to make the dataset complete and consistent.

2. Data Exploration and Insights

The data was initially in SQLite format which had to be converted to .csv format. All other datasets we acquired were in .csv format. We were working on the English Premier League, so the first plan of attack was to integrate all the data and filter out all the matches for Premier League.

2.1 Analysing the Home Team Advantage

The first step of any Machine Learning model preparation is to understand and gain insights from the data we're working with. Detecting patterns and correlations between data help us in modeling. In soccer, there is huge impact and advantage for the home team. Even a not so good performing club defeats the better opponent because of the home support and advantage. We start with exploring the Home team advantage using Tableau.



This is a real finding. Just look at the home team advantage it seems that every time a club is playing at home it wins! Further analysing the statistics we found out that 46% of the time the Home Team wins. Without considering anything if we just go with this insight we would be predicting a match right 46% of the time. It does not include even the draws.

2.2 Team Ratings

On paper if a team squad is stronger than its opponent there's a high chance that the stronger team wins. The overall rating is gained by calculating the average of all the players playing 15 including substitutes. We try to explore the number Win, Draw and Defeats vs. Team rating to find out the advantage of a stronger squad.

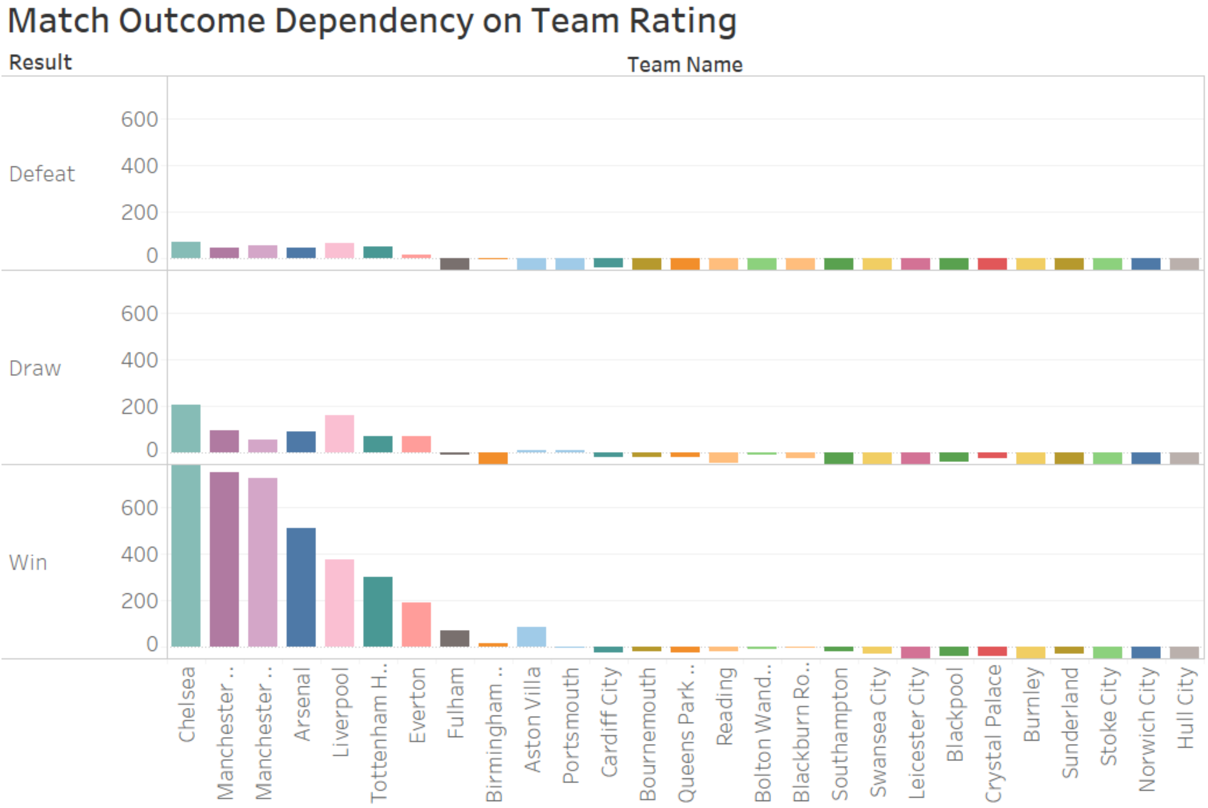


Figure:

The results were quite evident and expected. Stronger the team squad, more the number of wins. The team ratings plotted above show us that stronger teams win and draw more games than the weaker teams on paper.

2.3 Player strengths exploration

There are always top strong individuals which affect and enhance the performance of the whole team. We explored the top players of premier league and their attributes using Bokeh. Some of the visualizations are shown below.

3. Predictive Modelling

After exploring data and finding patterns comes the Machine Learning modelling. This is always the tricky and challenging part of developing a data product. Along-with smart feature engineering as we explored above selecting the appropriate classification algorithm is of the utmost importance. To predict the Win, Draw and Lose we use Support Vector Machines, Logistic Regression, Naive Bayes, KNeighborsClassifier and Random Forest.

4. Machine Learning Models

4.1 KNeighborsClassifier

Despite its simplicity, KNN can outperform more powerful classifiers and is used in a variety of applications such as economic forecasting, data compression and genetics. KNN falls in the supervised learning family of algorithms. Informally, this means that we are given a labelled dataset consisting of training observations (x,y) and would like to capture the relationship between x and y. In our case KNN gave an accuracy of 45% which was one of the lowest and not suitable for our goal.

4.2 RandomForest

Being a classification problem into 3 categories: win, defeat and draw, we aimed at using decision trees. However, to get an accurate model, we made use of RandomForest. Tweaks like using large number of trees (500-1000) showed better results as it gives accurate accuracies. With the help of this model, the overall performance achieved is 48.52%

4.3 Logistic Regression

One of the most basic yet best model to predict the classes using multinomial regression. The accuracy achieved by this model is 51.56%.

4.4 Naive Bayes

Since Naive Bayes makes use of probability and considers independency of the attributes, we aim to choose the features independently. Hence, Naive Bayes seems to be a better solution as it considers independent probability. The final accuracy achieved is 53.74%.

4.5 SVM

SVM gave the highest accuracy with 61%. However, this model also keeps in mind the shots-on-target which can only be calculated after the match outcome or on real-time basis. However, for real-time prediction, SVM proved to be the best model

From the above results, it can be seen that the overall accuracy is not yet achieved over 70%. Our models scored maximum of 61%. It do predicts well, but also shows that predicting the football matches is really tough and may be unpredictable. The reasons can be injuries incurred, transfers or even coaches planning. However, a deep dive into machine learning models can be used but through the research we can easily say that predicting is quite tough.

6. Evaluation

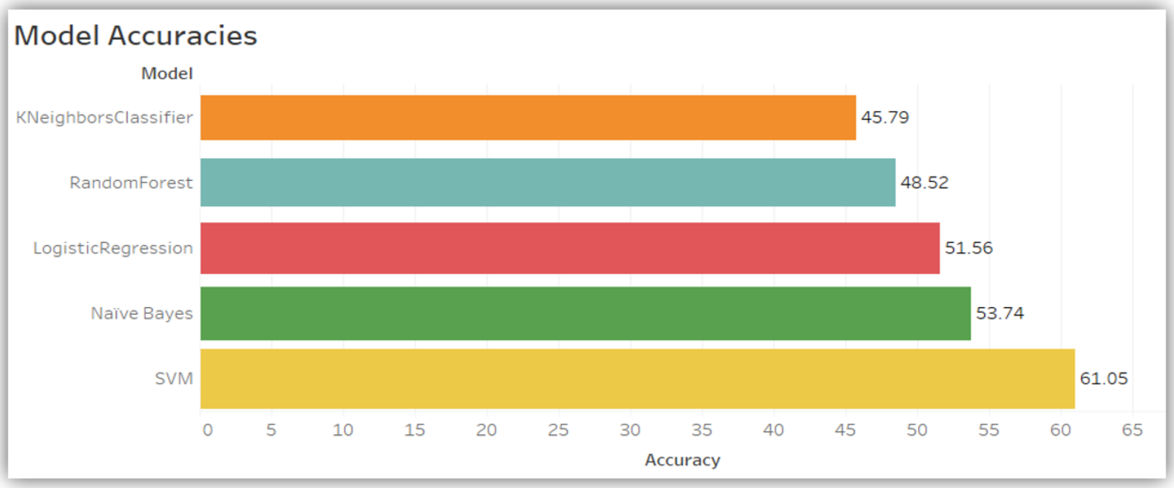
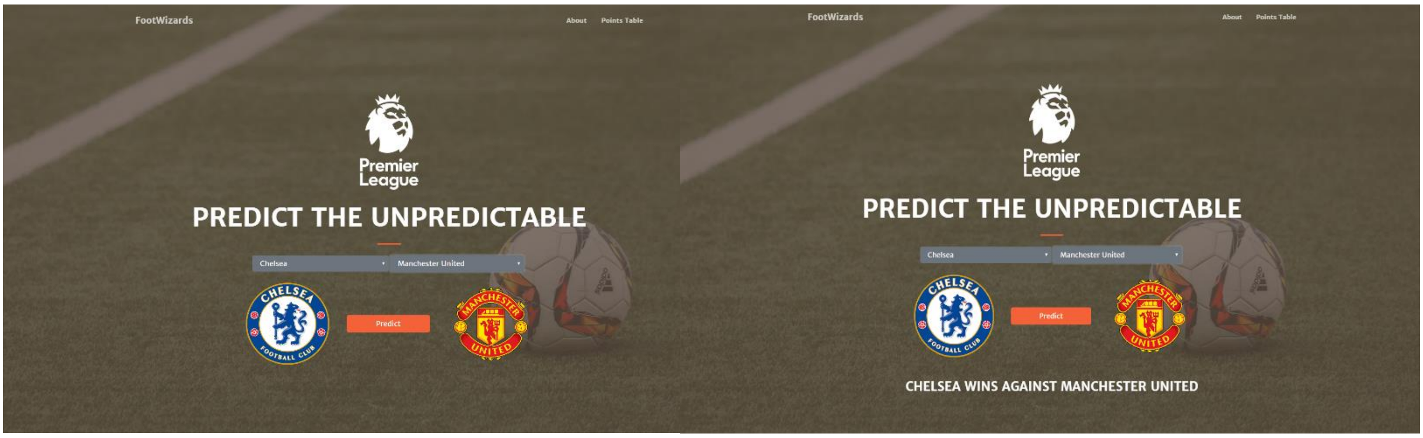


Figure 4: Accuracies of various classifiers

7. Data Product



CHELSEA WINS AGAINST MANCHESTER UNITED

Figure 5: Web UI

Here is a snapshot of our final WebApp/Data Product - FootWizard. We made use of Flask, HTML, CSS and JavaScript to built the Web UI. We made use of HCI principles to keep the web ui as simple as possible to get prediction in extremely less number of clicks. Within 3 clicks, user can get the results. User need to select the home team and the away team from the dropdowns. Once selected, clicking on "Predict" gives the result. Currently FootWizard can predict matches between teams in English Premier League but it can easily be extended to other European Football Leagues such as Bundesliga, La Liga and so on. We predicted 3 upcoming matches and each time model predicted the right team.

Home	Away	Date	Actual Decision	Our Prediction	Result
Burnley	Cardiff	13th Apr 19	Burnley	Burnley	True
Man United	West Ham	13th Apr 19	Man United	Man United	True
Fulham	Everton	13th Apr 19	Everton	Everton	True
Wolves	Man United	13th Apr 19	Wolves	Man United	False

8. Lessons Learnt

- **Preprocessing:** Many people underestimate this step in the data science pipeline. Making the datasets consistent as well handling missing values was an important step which took a lot of time and effort.
- **Choosing the right features:** We learnt that whichever model we chose, if we don't have the right features for the model, the accuracies would not go up. Thus we carefully crafted the four features that we used for our model.
- **Importance of feature scaling:** We first inserted the features in the model as it is, but the outcome wasn't as expected. After some research, we found out the need for feature scaling to nullify dominance of any features in the model.
- **Front end back end integration using Flask:** This was something new for all of us. Integrating Machine Learning model in an end-to-end Web Application was challenging and fun at the same time.

9. Summary

We aimed at predicting the outcomes of the EPL matches on the basis of their previous records based on winning streaks, head-to-head and overall rating. We implemented these models using Machine Learning techniques and found that SVM provides the best accuracy among the the other 4 techniques which was 61%.

Currently, we have predicted the Football Matches outcomes. However, betting is next challenge as it involves predicting the matches with higher accuracy and predict the dynamic odds in real-time. We plan to recommend best platform to maximize the betting profits.

10. Future Work

- Currently all the predictions are made taking into account the match statistics available before the game.
- We found that if we consider shots on target and corners also with the existing features, an accuracy of almost 66% is achievable.
- Getting the live data in an interval of every 15 minutes, which is then continuously fed to the model a great deal of improvement is possible.
- Correct match predictions and betting odds analysis can be utilized for developing betting platforms which are getting very popular nowadays.
- Also, the present scope of FootWizard is limited to EPL matches which can be easily extended to any European Football Leagues.

11. References/Related Work

1. https://www.kaggle.com/hugomathien/soccer
2. http://www.football-data.co.uk/
3. http://sofifa.com
4. https://www.kaggle.com/airback/match-outcome-prediction-in-football
5. https://towardsdatascience.com/predicting-premier-league-odds-from-ea-player-bfdb52597392