Predicting Football Matches' Results in the English Premier League with the Aid of Different Machine Learning Models

Zeyad Manaa
Aerospace Engineering
s-zeyadmanaa@zewailcity.edu.eg

Rana Elsemary
Aerospace Engineering
s-rana.elsemary@zewailcity.edu.eg

Abdelrahman Kotb
Aerospace Engineering

Abstract

The English Premier League (EPL), which is televised in 212 countries to millions of people, is the most popular club-based soccer league in the world. In this report, an attempt to predict match results in the EPL is made using two machine learning algorithms; Linear Regression (LR), Support Vector Machines (SVM), Nearest Neighbour Classifier (KNN), Decision Tree, and Random Forest based on the data provided by [1]. Data Scraping and Cleaning throughout systematic procedure are made in order to merge EPL data from 2000 to 2021 with time span of 21 seasons. Furthermore, an exploratory data analysis is operated on the data set to get some insights about the data as well as some features correlations. The results showed that the best model is SVM with test accuracy of 68 %, whereby the lowest model's score goes to KNN with test accuracy of 52 %. Models can be manipulated and get better results by integrating sentiments (e.g. witter sentiments) and enlarging the data, as well as involving totally predictable data.

1 Introduction

The English Premier League (EPL), based in England, is the most popular soccer league in the world. 4.7 billion people watched it in the 2010-11 season, according to estimates. As a result, it's no wonder that the TV rights are worth £1 billion per season. There are 20 teams competing for first place in the EPL. The last three teams are excluded and replaced with better-performing teams from lower EPL divisions. Every team play two rounds with the same team, one at home and the other away. Throughout total, there are $2 \times {}^{20}C_2 = 380$ games in a season. A season is defined as the time period between August and May of the following year.

1.1 Challenges

The considerable occurrence of draws; (*neither team wins*) makes projecting outcomes somehow difficult as compared to other sports. As shown in figure 1, the aggregate draw percentage is 25.21% which is quite large.

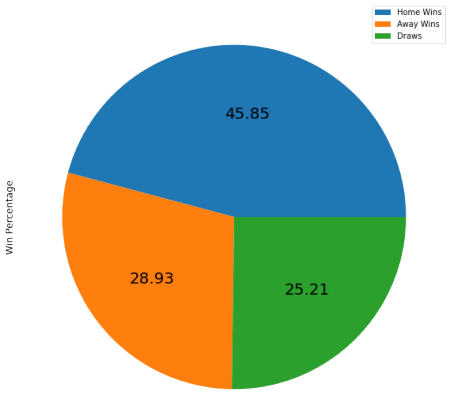

Figure 1: Aggregate Win Percentage

Some data collection process difficulties are, as well, presented as a challenging part. Data files provided in [1] does not have the same structure for data.

2 Approach and Methodology

2.1 Data

The data utilized in the models are retrieved from [1], which is a free repository of football outcomes that provides historical data. Data are provided through the portal starting from season 1993-1994 till season 2020-2021 in separate files for each season. Relevant data starts from season 2000-2001. So, the data we worked on is ranging from season 2000-2001 till 2020-2021.

The same data are modified by Kaggle [2] with excellent way to get useful features from it. It will be used for building the model, the original data set is used for data exploration and visualization. Some of features included in this data set are, Full Time Result (FTR), Home Team Point (HTP), Away Team Points (ATP), Home Team Match no. Result (HMn), Away Team Match no. Result (AMn), Difference between scored and received goals for home team (HTGD), difference between scored and received goals for away team (ATGD), and Difference between HTP and ATP (DiffFromPts). This can be shown in figure 2.

	FTR	HTP	ATP	HM1	HM2	HM3	AM1	AM2	AM3	HTGD	ATGD	DiffFormPts
6835	H	1.078947	1.842105	L	L	L	D	W	W	-0.289474	0.710526	-0.263158
6836	NH	0.947368	2.552632	W	D	W	W	D	W	-0.473684	2.052632	-0.131579
6837	NH	0.868421	0.789474	L	L	L	L	D	D	-0.710526	-0.894737	-0.052632
6838	H	1.947368	1.236842	W	L	W	W	L	L	0.973684	-0.078947	0.078947
6839	H	1.026316	1.289474	D	W	L	D	W	W	-0.578947	-0.315789	-0.105263

Figure 2: Updated Data by kaggle.

Furthermore, the data set reduced the problem to two class classification problem instead of three class by making the FTR have two attributes which are H, and NH.

2.2 EDA

From figure 1, home team has a wining probability of 45.85%. This, without any prediction, gives a prediction accuracy of 45.85% if you predicted the home team as a winner. Figure 3, verify this findings discussed by figure 1 as the percentage of home team wining alternate between nearly $50\% - 45\%$.

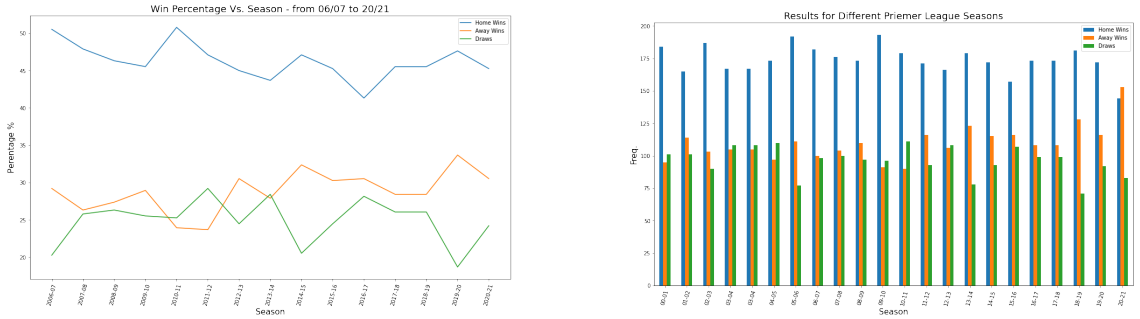

Figure 3: Different seasons result statistics

Provided in figure 4 the correlation between the selected features. It is noted how some features are linearly dependant on others.

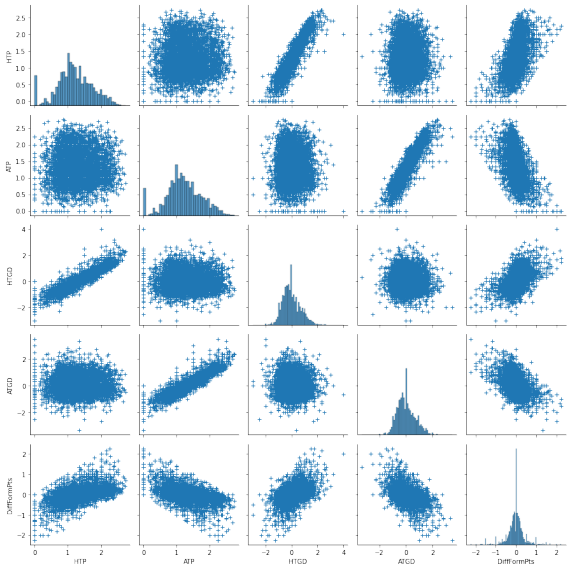

Figure 4: Pair Plot of the selected features

To further get more sense of the correlation between the selected features a heat map of the features correlations can be found in figure 5.

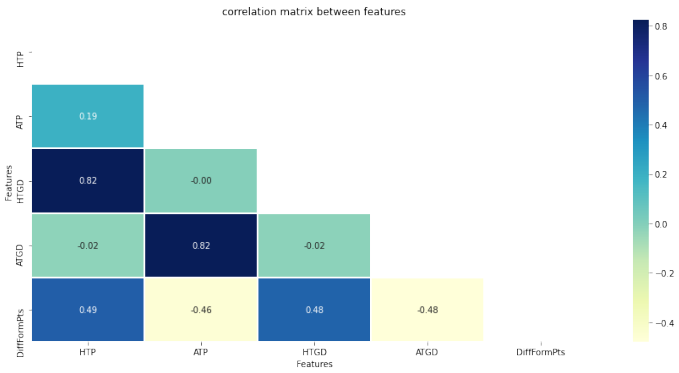

Figure 5: Heat map for correlation between the features

2.3 Models

Five models will be implemented; Linear Regression, Support Vector Machines, KNN and Decision Tree and Random Forest. Each model will be evaluated using two metrics F1 score, and accuracy. We will compare between them to get out with the suitable model for our project.

3 Tools and Libraries

Libraries

- Pandas.
- NumPy.
- Matplotlib.
- Seaborn.
- Sklearn.

4 Model evaluation

4.1 Linear Regression

LR Model Output

```
Training a LogisticRegression using a training set size of 6790...
F1 score and accuracy score for training set:  0.5939 , 0.6470.
F1 score and accuracy score for test set:  0.5532 , 0.5800
```

4.2 Support Vector Machines

Different SVM models are tested with different kernels with the same C = 1000

SVM Models Output with Different Kernels

```
Linear kernel ⟶ accuracy = 62%
Poly kernel ⟶ accuracy = 54%
rbf kernel ⟶ accuracy = 66%
```

A comparison between the three kernels' SVM model is shown in figure 6.

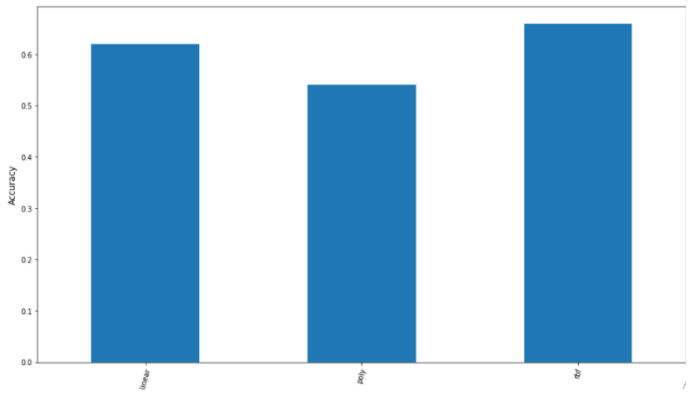

Figure 6: SVM Different Kernels

The hyper-parameter C tuning resulted in 100 as the best value for C to be used.

SVM Models Output with Kernel

```
Training a SVM using a training set size of 6790...
F1 score and accuracy score for training set:  0.6161 , 0.6804.
F1 score and accuracy score for test set:  0.5714 , 0.6400.
```

4.3 KNN

The hyper-parameter tunning for the N parameter resulted in N = 9 gives the best accuracy for the KNN Model.

KNN Model Output

```
Training a KNeighborsClassifier using a training set size of 6790...
F1 score and accuracy score for training set:  0.7529 , 0.7761.
F1 score and accuracy score for test set:  0.4000 , 0.5200.
```

4.4 Decision Tree

Decision Tree Model Output

```
Training a DecisionTreeClassifier using a training set size of
6790...
F1 score and accuracy score for training set:  0.9774 , 0.9792.
F1 score and accuracy score for test set:  0.5778 , 0.6200.
```

4.5 Random Forest

Random Forest Model Output

```
Training a DecisionTreeClassifier using a training set size of
6790...
F1 score and accuracy score for training set:  0.9774 , 0.9792.
F1 score and accuracy score for test set:  0.5778 , 0.6200.
```

5 Discussion

5.1 Discussion

Figure 7 shows a comparison between the models.

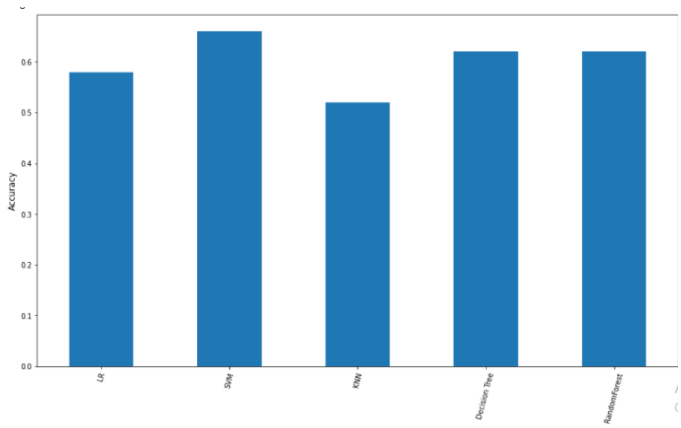

Figure 7: Final Models Comparison

From this graph we conclude that SVM model is the best model in accuracy and the fitting to use for our project to bet the best accuracy.

5.2 Scores on Kaggle

LR on Kaggle Model

```
F1 score and accuracy score for test set:  0.59 , 0.65.
```

SVM on Kaggle Model

```
F1 score and accuracy score for test set:  0.70 , 0.54
```

Random Forest on Kaggle Model

```
F1 score and accuracy score for test set:  0.58 , 0.64.
```

6 Conclusion

To sum up, this project helped us to learn more about data collection and how to find the suitable features and neglect the irrelevant ones. also we learnt that not all models can work better for certain models. Model can be better if data are larger, Twitter sentiment analysis are integrated into the models. Moreover, the used features should be completely predictable.

References

[1] "Historical football results and betting odds data." `http://football-data.co.uk/data.php`, 2021. [Accessed: 2020-12-19].

[2] "English premier league." `https://www.kaggle.com/saife245/english-premier-league`, 2019. [Accessed: 2020-12-19].