UEFA Champions League Prediction Model

Abdul Alhusaini, Leonardo Astorga, Mehlam Saifudeen

Abstract

In this project, we created a predictive model to compare different metrics and their importance in predicting the winner of the UEFA Champions League. The data we used to predict future winners of this league was called from fbref.com, a database of soccer statistics. The data collected to build the model comes from the previous four seasons. The specific types of data collected include shooting percentages, goalkeeping percentages, passing percentages, as well as other relevant data. The aim of our project was to predict and explain the features that were most directly correlated with a winner in the champions league. Prior to our construction of the model, we hypothesized that the variables that were most important for predicting winners were wins and losses. However, after actually running the model, we discovered that the most important figures were league rankings, wins, and points scored.

Github repository link to the code:
https://github.com/mehlu22/UEFA-Champions-League-Prediction-Model

Introduction

The outcome of any form of competition that results in one victor has been of historical context and falls under the umbrella of game theory. More specifically to our final project, we researched which factors have historically contributed more in deciding the winner of a UEFA Championship League. This league holds an annual tournament where the best soccer clubs (32) in Europe meet in a grueling round robin group stage qualifier; the winners qualify for a double-legged knockout format and a single-leg final. With the rapid increase in viewership of football (soccer) games around the world, professional soccer players now have more potential financial gain to relocate teams. In the past, a team's roster would remain relatively consistent through the years, making it significantly easier to predict the winner of the championship match. Additionally, players in lower markets now have access to the same world class physical fitness programs as higher market teams, making it possible for any player to compete at an all-star level, and for any team to become a championship contender.

In addition to the players, there are also pressures that those off the field must face due to the changing conditions of the sport. For the members of team management, they must make the informed decision of investing more capital into their current roster, or keeping them in reserves for a future season. For fans financially invested into sport, they must face risk with each position they take. For everyday fans, they must decide whether watching a match in-person outweighs the opportunity cost of not watching the match, depending on the potential outcome of the match.

With the predictive model we have constructed, we aimed to solve all of these problems, by finding which factors lead to the greatest success among football clubs in the UEFA championship league. Management is able to analyze whether their current team is aligned with the metrics required to be successful in the championship league. Sporting speculators will have information needed to predict the outcome of a matchup. Regular fans will know if their team is the likely winner of a matchup, making it more enjoyable for in-person viewership.

1.1 Data

The data collected for our analysis was collected from fbref.com, a database containing features such as shooting, passing, goalkeeping, as well as many other relevant metrics. This data allows us the opportunity of seeing how different metrics perform as a classifier for success against different football team matchups. There is a data set available for each season, our model takes into account the previous four seasons to predict the importance of each metric. Using web scraping methodologies, we are able to collect the data from fbref and import them into Pandas dataframes and NumPy arrays which provide ease for data analysis.

1.2 Approach

Fundamentally, it is known that the team that scores the most points will win, which implies that a team that is better at scoring is more likely to win a matchup. However, the outcome of the match is much more complex and dependent on other factors. To create our model we utilized several Sci-kit learn functions. The first function we used was the train_test_split() which allows us to split our arrays of data into randomly sized train and test subsets.

Using these train and test subsets we were able to run multiple machine learning algorithms on them, also provided by Sci-kit. The first algorithm we used to fit our data was linear regression. The root-mean-squared-error (RMSE) produced for the train and

test dataset were 3.03 x 10^{-15} and 0.0668, respectively. The next algorithm we fit to our data was random forest. The RMSE produced for the train and test dataset were 0.4435 and 1.1141, respectively. The final algorithm we used to fit our data was gradient boost. The RMSE produced for the train and test dataset were 0.000149 and 0.8613, respectively.

Using the gradient boost's fit on the data we were able to sort through the features with respect to their importance in predicting the winner of a champions league matchup.

1.3 Summary and Insights

Our findings indicate that the metric with the highest correlation to victory was ranking among the entire league. Following this metric was average points per game, then win/loss percentage.
31.5% of the importance can be attributed to ranking among the league. 27.48% of the importance can be attributed to average points per game. 22.84% of the importance can be attributed to win/loss percentage. It should be taken that these figures have an inherent margin of uncertainty since no predictive model will always be correct. All of these figures appear to align with our beliefs since historically, ranking, ppg, and W-L have played a big part in the success of a football club in the UEFA Championship League.

Methods

The raw data is obtained from https://fbref.com/en/comps/8/Champions-League-Stats, which is a website devoted to tracking statistics for football teams and players from around the world. FBref was created by Sports Reference, the team behind popular stats. We scraped the data from the web and using the various functions and libraries of python we were able to store and clean up the relevant data into the code.

The figure below shows that Project workflow which has been thoroughly explained below if anyone would like to create a similar or a even a modified version of the prediction model with the same data using the same or different factors:

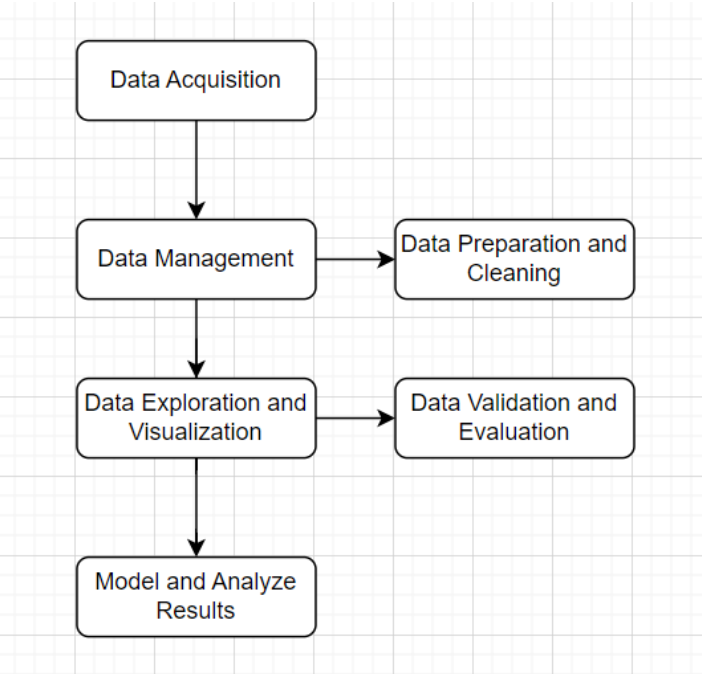

Figure 1: Project Workflow

The workflow that we used to store the data and analyze it in the following section is:

1. **Data Acquisition:**
   The data was scraped from the link provided above which contains all the relevant statistics for each of the soccer teams and players in every European team and league. We scrapped the data from the years 2018-2019 to 2021-2022 for each of the 32 teams participating in the UEFA Champions League tournament.

2. **Data Management:**
   The relevant data which would help create a prediction model from the website was scrapped such as number of goals scored, trophies won, head to head records, recent form, number of points scored, goals for and against a team etc. was stored in numpy arrays and pandas dataframes due to the flexibility in performing various CRUD operations and at the same time allowing us to create test and training data sets to analyze using machine learning algorithms.

3. **Data Preparation:**
   In order to prepare the data to create figures and models the following steps were taken:
   a. Inspecting the data: Pandas functions such as head(), info(), and describe() to get a sense of the structure and content of the data.

b. Cleaning the data: The data was searched for and handled missing values, duplicates, and other issues in the data. In order to do this we used pandas functions such as isnull(), dropna(), and drop_duplicates() to identify and remove invalid or unnecessary data.
c. Formatting the data: Changing the data types of columns, renaming columns, and reshaping the data using functions such as astype(), rename(), and pivot_table().
d. Filtering the data: Use of pandas functions such as where(), query(), and isin() to select only the rows and columns that are relevant to the analysis.
e. Aggregating the data: Use of pandas functions such as groupby(), pivot_table(), and crosstab() to summarize and reshape the data in a way that is more useful for analysis.
f. Normalizing or standardizing the data: Use of NumPy functions such as mean(), std(), and subtract() to scale the data so that it is on the same scale and comparable across features. This was one of the most important steps in data preparation as it helps prepare the data for any form of analysis to be conducted on it later.

4. Data Exploration:

In order to explore the data, the following steps were taken:

a. Visualizing the data: Used libraries such as Matplotlib and Seaborn to create visualizations of the data. This helped us get a sense of the overall structure and relationships in the data that was scrapped. We visualized some graphs that look at the number of points scored by each team for each year
b. Calculating summary statistics: Used pandas functions such as mean(), median(), min(), and max() to calculate basic statistics about the data. Using other pandas functions such as histo() to visualize the distribution of the data helped us understand the impact of each of the statistics for each team and which one would have more of an impact on the overall prediction model.
c. Splitting the data into training and test sets: Used pandas functions such as train_test_split() to divide the data into a training set, which will be used to build the model, and a test set, which will be used to evaluate the model's performance using the machine learning algorithms as stated earlier. This was again done in the machine learning algorithms in order to obtain the RMSE values.
d. Exploring and comparing different model architectures and hyperparameter settings: Used libraries such as scikit-learn to build and test various prediction models, and use techniques such as

cross-validation and grid search to find the best model architecture and hyperparameter settings.

5. **Data Validation and Evaluation:**
    a. Training the model on the training data: Using the training data we tried to fit the model to the data. This involved adjusting the model's hyperparameters to optimize its performance.
    b. Making predictions on the test data: Using the trained model we tried to make predictions on the test data.
    c. Analyzing the results: Looking at the results of the evaluation we tried to understand how well the model is performing and where it may be making errors. This can help you identify any areas where the model needs improvement, especially for the machine learning algorithms.
    d. Fine-tuning the model: Making adjustments to the model, such as changing the model architecture or adjusting the hyperparameters, in order to improve its performance.
    e. Validating the model: Use cross-validation and other techniques to ensure that the model is robust and not overfitting the data.
    f. Evaluate the final model: Use the test data to evaluate the final machine learning model to be used based on comparisons of its performance to the machine learning models available.

6. **Data Modelling and Visualization**:
    In order to model the data provided, for it to make sense we had to find a specific type of correlation between the information given which could be later used in the Machine Learning Algorithms. To address this, we found the correlation between the different inputs in the data sets. As for the three different data sets provided using the library Seaborn, a python data visualization library based on matplotlib.It provides a high level interface for drawing attractive and informative statistical references.
    Using the correlation to apply it to Machine Learning algorithms was the most difficult part of the project.we had to learn through various examples how to train the data sets and apply it to various machine learning models. The Machine Learning models we decided to choose to train the data sets were:

    a. Linear Regression: is the supervised machine learning model which finds the best fit linear line between the independent and dependent variable, that means it finds the linear relationship between the dependent and independent variable.

b. Random Forests: Random Decision Forests is an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.
c. Gradient Boost: Gradient boosting is a machine learning technique used in regression and classification tasks, amongst others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.

By applying the 3 Machine Learning Models to my 3 different data sets we got 9 different RMSE Values(Root Mean Square Error is a method of measuring the difference between values predicted by a model and their actual values) which were as follows:

Linear Regression:

```
RMSE for the train dataset: 3.0306096457764417e-15
RMSE for the test dataset: 0.06684421219526265
```

Random Forests:

```
RMSE for the train dataset: 0.4756299081468801
RMSE for the test dataset: 1.0340575419192106
```

Gradient Boost:

```
RMSE for the train dataset: 0.0001493318669904948 3
RMSE for the test dataset: 0.9724537302789279
```

How we used these RMSE values obtained above in order to form our prediction model, as well as visualizing the graphs and data is further explained in the Results and Discussion section.

Results and Discussion

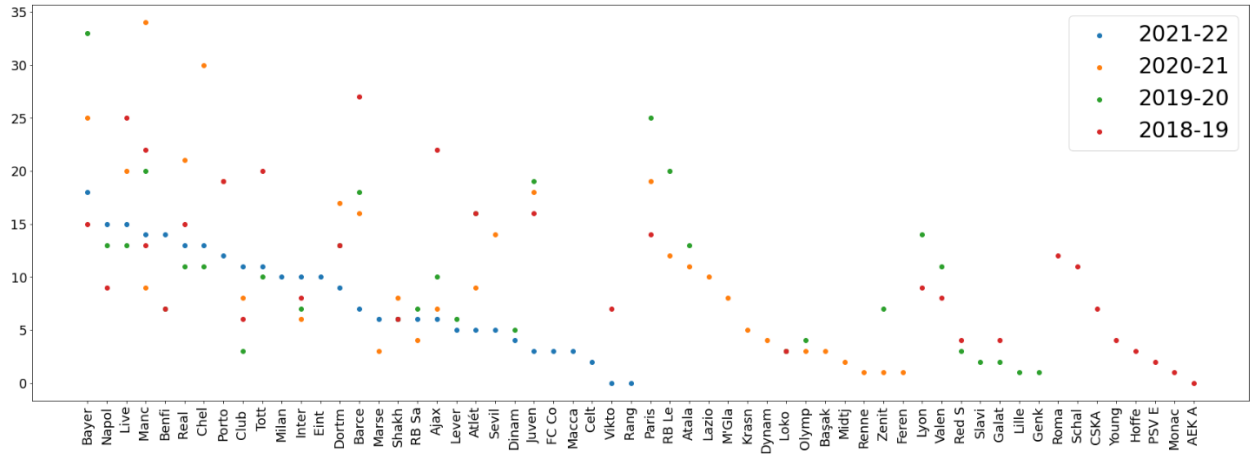

Figure 2: Initial Visualization of Data-set

In Figure 1, we can see the numbers of points that were scored by each of the 32 teams that participated in the UEFA Champions League from the years 2018 - 2021. This box plot generated from the web scraping done on the data found in the website we used to collect our data. From this box plot alone, we can see the number of points scored by each team in addition to the goal difference and number of wins throughout the season can be an important factor in determining the winner of the UEFA champions league. This provides us with a specific correlation between the information provided (the data from the website) which we used later on to optimize our Machine Learning algorithms. From this box plot alone we can see that Manchester has extra dots, however, this is done by design; we wanted to bunch together Manchester United and Manchester City.

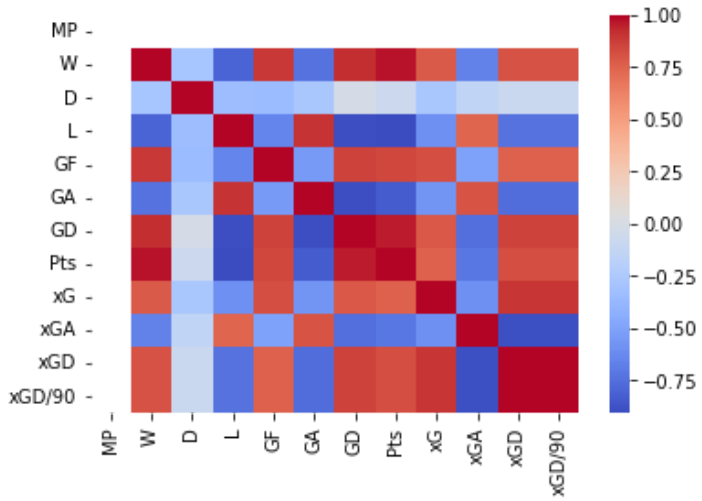

Figure 3: ***Heat map***

In Figure 3, the heat map displays the number of wins, number of goals and points are strongly positively correlated but goals against a team (GA) and goal difference (GD) is not strongly correlated because the heat map shows the importance of each factor in determining the winner of the championship league. Some of these include number of wins, losses, draws, goals scored (GF), goal difference (GD), expected values denoted by (x..), number of points etc. This heat

map helps us better understand the importance of each factor in terms of how red each square is, that is the redder a square is, the more important a factor it is in determining the winner. An example of this can be the number of points, and the number of wins. In addition to this the heat map helps us find whether there is a positive or a negative correlation between 2 factors. Besides the diagonal line we can compare the 2 factors to see how they are positively or negatively correlated to each other depending on if the square is more red or more blue. To re-emphasize the results shown by the heat map, the more red an intersection is, the more potential that intersection has in predicting a team's overall success. In essence, the intersections of higher value correspond to the factors that relate the most. The intersection between the number of wins and the number of points has a high predictive value, allowing us to anticipate that this intersection has the highest probability of determining a team's chances of winning. Therefore, based on these results, the number of points has the strongest correlation.

	Significance of Each Factor		
Rk_GR	0.377046	Squad_Milan	0
W	0.224426	Squad_Porto	0
Pts	0.178052	Squad_RB Sa	0
Rk_R16	0.100746	Squad_Real	0
GD	0.051087	Squad_Sevil	0
L	0.025166	Squad_FC Co	0
GA	0.021806	Squad_Inter	0
GF	0.007631	MP	0
xGD	0.004032	Squad_Eint	0
Squad_Bayer	0.003474	Squad_Dortm	0
xGA	0.003342	Squad_Dinam	0
xG	0.00241	Squad_Club	0
Squad_Napol	0.000353	Squad_Benfi	0
Squad_Barce	0.000296	Squad_Ajax	0
Squad_Vikto	0.000069	Squad_ Tott	0
D	0.000061	Squad_ Rang	0
xGD/90	0.000001	Squad_ Manc	0
Squad_Juven	0	Squad_ Live	0
Squad_Lever	0	Squad_ Chel	0
Squad_Macca	0	Squad_ Celt	0
Squad_Marse	0	Squad_Atlét	0
Squad_Shakh	0		

Tabie 1: ***Significance of each factor for a prediction***

In Table 1, the importance of each factor in creating a prediction model for the winner of the UEFA Champions League. The table shows us that the number of wins, number of points and goal ratio are the most significant than others. With the different values of RMSE, we chose the smallest value out of them which belonged to the Machine

Learning model of Gradient Boost. We went ahead with using gradient boost given the fact that linear regression had a lower RMSE value as the data given out by the linear regression model was minute and insignificant to the point that it could not have been plotted on a graph to understand and when analyzing for Random Forest and Gradient Boost we chose Gradient Boost. The table shows the importance of each factor in creating a prediction model for the winner of the UEFA Champions League. We proceeded with plotting out the most important value depicted through our Machine Learning model.

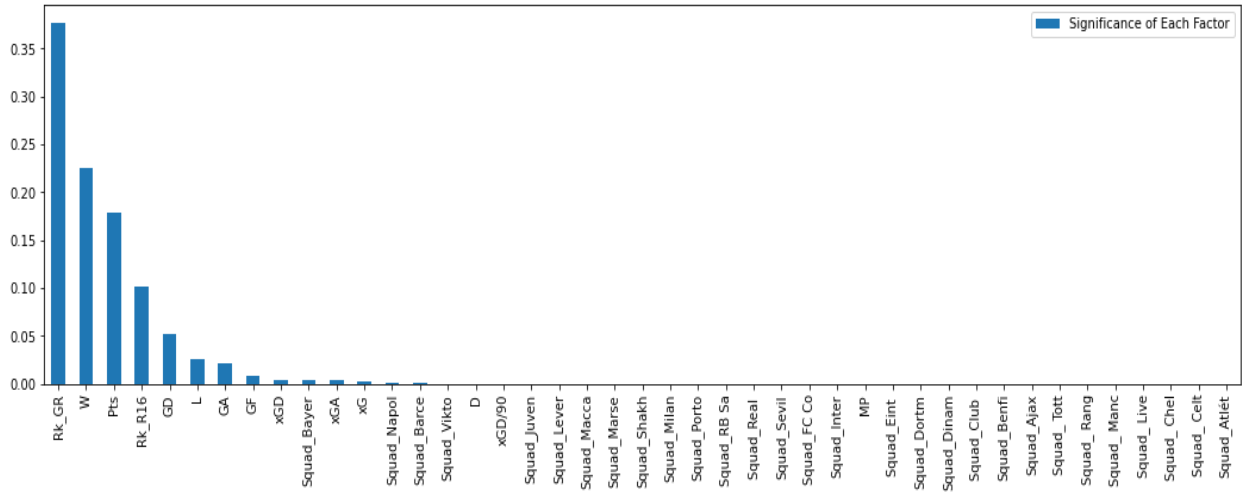

Figure 4: ***Visualization of Significant Factors***

It can be clearly interpreted from the graph that, out of all the parameters that have been applied, the goal ratio and number of wins that a team possess holds the most importance in decoding a team's success in the UEFA Champions League. It can also be observed that the number of points that a team has is also significant - which indicates how closely related the team's wins and points are since every win results in a team getting 3 points. Goals scored by a team (GF) and the team's goal difference also hold tiny contributions because a higher scoring team is usually expected to collect more points. We have also added a few parameters that are completely redundant in order to test and prove that our machine learning algorithms indeed mine out the components that are most significant. Hence these are some factors that can determine a team's success in UEFA champions league based on current trends and historical events.

Conclusion

Utilizing models such as the heatmap, as well as machine learning modules such as gradient boost, random forest, and logistic regression we have determined that the most

likely factors in predicting the winner of the UEFA Champions League are league rankings, points scored, and number of wins. Although this is a different conclusion than what he had hypothesized, the results appear very probable. If we look at the previous winners of the UEFA league, they have always been in the top ranks of the EU soccer teams. Logically, a team being in the top ranks also implies that they have increasing scoring and winning capabilities in individual matches. While our results are sufficient, all models can be improved. A few ways our model can be enhanced is by increasing the size of our train and test dataset, as well as including data from an extended number of years. In conclusion, building this model has given us a better understanding of what is required to win the UEFA championship, in addition to a more holistic view of the types of research and algorithms required to make predictions on a number of varying datasets. Through our model we predicted that Real Madrid would be the winner of the UEFA Champions League in the year 2022 and this was the result that was seen as well.

Roles

Mehlam - Role was to use the machine learning algorithms in order to analyze the data and create RMSE values which helped us predict the winner of the champions league tournament.

Abdul - Collected data necessary for data analyzation through web scraping. Conducted data analysis using machine learning algorithms.

Leonardo - Data Representation and Visualisation as graphs, scatter plots, and heatmaps.

References

1. https://fbref.com/en/comps/8/Champions-League-Stats