Andrea Fernandez

Predicting UEFA Champions League

MSDS 696 Data Science Practicum II

Regis University

Introduction

The UEFA Champions League consists of 32 teams that are divided up into 8 groups, A-H, and is held every year. There is an initial stage of qualifying rounds which results in the 32 total teams in the competition. The qualifying stages begin in July and the final is taken place in May, which makes this a year-round tournament. Teams that can participate in this league are from various countries in Europe and are considered the top club teams in the world. They take the best teams from each country's specific soccer league and have them compete against one another to produce the best club team in Europe. Due to the nature of this league, it is considered extremely competitive and is one of, if not, the biggest accomplishment for a player or club's career. The Champions League promotes so much excitement towards the end of the tournament, but with machine learning we can extend the anticipation.

Research Question

The purpose of this project is to use machine learning to predict the outcome for competitive leagues like The Champions League. Every year new teams qualify as well as new teams being created in general. This tournament first began in 1955 and has grown so much since then, and the biggest teams in Europe have taken it over. There are teams that competed in the 1950's that I have never heard of as well as countries that no longer participate. There are so many great teams involved now and player transfers happen every season which changes the dynamic of the competition. With all these changes and inconsistencies with teams, my main goal is to produce a model that can accurately predict the outcome of this tournament with machine learning.

Data and Cleaning

I began this project by compiling enough data to make various models and ended up with 2 main datasets. The first dataset has of 7 columns with 127 rows and consisted of teams that have been in the final of every Champions League. It specified if the team was the "winner" or the "runner-up" and provided information like formation, the coach that season, the mvp, and the country the team is from. The second dataset I used was much bigger with dimensions of 23 columns by 6553 rows. This dataset had every game that has ever been played in the history of The Champions League up until the 2015/2016 season. There were a lot of variables included like the score at half-time and full-time, the home and visiting team, the total goals for both teams, as well as aggregate scores. Aggregate scores are important because they aide the situation of teams ending in a tie. Scoring at the other teams home gains the advantage and it acts as a combined score. The dataset also includes the country that each team comes from as well as basic information like date, the season, and the round that was being played.

The first dataset that has only winner or runner-up data had zero missing values or NA values. There were only a few "unknown" data points, but since the dataset consisted of only 127 rows and the unknowns weren't in the most important columns, I did not exclude them. The second dataset of every game had a lot of missing values and NA values that needed to be fixed. I decided to switch the NA values for character values to keep the data from the earliest days of the tournament. The Champions League has changed over 60 years with the addition of Group stages due to the tournament growing. I replaced some NA values to "No info", "No extra time", and "No penalties" and was able to keep the dataset at its original length. There were also some foreign characters that were deleted from team names due to the different languages, but the names were still recognizable without them.

Exploratory Data Analysis

The process of data exploration began with understanding the shape of the datasets as well as the type of columns in each one. Once I understood what the dataset looked like, I started to dig deeper so find information that would be useful to the project. I produced various results that listed the top teams for various topics. My EDA communicated that Real Madrid has won the most titles as well as been in the final the most out of all teams. Spain has been represented the most in this competition as the winner or runner-up, which is where Real Madrid is located. This league has had 53 countries represented all over Europe and some parts of Africa and Asia. Although, most of those countries do not compete in this competition anymore, it is interesting to see how diverse this league has been. The main countries that have proven to be the most successful as the winner or runner-up have only been from 13 countries and are all in Europe. Some other interesting visualizations showed how the most common scores from every game has been 1-0 and second most being 1-1. This shows how competitive these games are and how it is not easy to score on the opposing team.

Building Models

The model building process began with deciding to use 3 classification models, which were a logistic regression, naïve bayes classification, and a decision tree. I chose classification models because I not only wanted the model to predict outcomes, but this competition also has qualifiers, as well as groups in which teams are sorted into. The model would need to have the ability to categorize some data as well as predict where scores and teams are placed in the bracket. Although the data was clean, it still needed to be prepped before splitting the data. The first step was making a feature set and a target variable for the full-time score to produce score outcomes. All the numeric columns in the dataset needed to be standardized which included 6 in

total. The next step was to convert the remaining object columns with characters into dummy variables that way the model would run and not stop due to errors. I was able to display the new dataset and see that all the input was numeric and scaled appropriately. I used sklearn to split the data into training and testing sets as well as shuffle it as an extra randomization. I made the test size 100 due to how large the dataset was and made the random state equal to 2. Now that the data was split, the process of training it began with training it for a certain amount of time and then stopping and moving on to the next set. The training portion also included code to produce the F1 score as well as the accuracy values for each model. The final step was to set up each model with their parameters needed, and then run it with the training and testing data.

Results

Once all 3 models were run on all the training and testing data, the results were printed as F1 score and accuracy score. From the results, the training set size was 6454 and ran for different times depending on the model. The logistic regression model ran the longest for training the data at 20.18 seconds, while the prediction only took 0.003 seconds. The F1 and accuracy was 94% in total and proved to be a solid model for predicting this dataset. The naïve bayes trained the model much faster at 0.0199 seconds and predicted in 0.0193 seconds. The scores for the F1 and accuracy were higher at 100% for both, which is another great model to predict. The last model was the decision tree using the XBG classifier that trained for 6.05 seconds and predicted in 0.05 seconds. This model had the highest rate of accuracy and F1 score of 99.8%, which meant it predicted the outcome almost exactly as the actual values.

Although these models were sufficient as is, I went ahead and changed the parameters of the decision tree to assess the differences. The first parameter adjustment was slightly lower than the initial model at 99% and 98%. The second adjustment proved to be less successful as the first

2 with an F1 score and accuracy score of 96%. Once I determined the best model was the initial decision tree XGBoost classifier, I saved the model for future use.

Conclusion

Overall, this project was successful in the ability to predict the outcome of The Champions League regardless of the inconsistencies and changes overtime. The models that I was able to create had very high accuracy rates at predicting the outcomes of games. This information can be very useful for sports betting, while being specific enough to predict the end score. The models could also be utilized by the organizations and teams that participate in this tournament to strategize their formations for each game. Companies like Nike and other sponsorships would benefit from the prediction of the winner by making merchandise to sell before the final has even taken place. These models can be used for various purposes for predicting the outcomes of The Champions League.

Sources

Brownlee, Jason. “LOOCV for Evaluating Machine Learning Algorithms.” *Machine Learning Mastery*, 26 Aug. 2020, machinelearningmastery.com/loocv-for-evaluating-machine-learning-algorithms/.

Brownlee, Jason. “Save and Load Machine Learning Models in Python with Scikit-Learn.” *Machine Learning Mastery*, 27 Aug. 2020, machinelearningmastery.com/save-load-machine-learning-models-python-scikit-learn/.

gwats10. “gwats10/Superbowl_2020_Predictions: Rank Every NFL Team by Likelihood That They Win the 2020 Superbowl Based on Regular Season Statistics and Regular Season Statistics from Past Winners.” *GitHub*, github.com/gwats10/Superbowl_2020_Predictions.

Ramachander, Rajesh. “Data Science: Data Cleansing and Visualization for Beginners Using Python.” *Medium*, Medium, 11 Sept. 2020, rajesh-r.medium.com/data-science-data-cleansing-and-visualization-for-beginners-using-python-3f55323768f1.

Sharma, Sagar. “Visualizing Cross-Validation Code.” *KDnuggets*, www.kdnuggets.com/2017/09/visualizing-cross-validation-code.html.

Tuwani, Rudraksh. “RudrakshTuwani/Football-Data-Analysis-and-Prediction.” *GitHub*, github.com/RudrakshTuwani/Football-Data-Analysis-and-Prediction.