Football Top League Placement Prediction

Waris Barakzai, Gary Wu

April 2020

1 Project Description

The prediction of outcomes in professional football (soccer) leagues is something of great importance to fans of the sport and betting agencies alike. By using advanced statistics, teams can look to target specific characteristics in players that can affect overall team quality without putting a strain on financial resources. For example, teams such as Getafe in La Liga target tall, athletic players that may not reach the technical standards of much more lavish teams like Barcelona, but execute the game plan with the intensity and style desired by their manager and therefore obtain their desired wins. In this project, we seek to classify teams that will be promoted to the Champions League and the Europa League, as well as teams that will be relegated from the top five European Leagues. In addition, we seek to predict the positioning of the league tables of several leagues. By using advanced statistics, we can prevent from using financial backing as an overestimating predictor variable and instead focusing on the characteristics and play style of a team that are most likely to generate wins.

Questions we seek to answer include: Which teams are promoted to the Champions and Europa League and which teams are relegated to the lower division? Can we accurately predict the positioning of each team in each of the top five European leagues (Premier League, La Liga, Bundesliga, Serie A and Ligue 1)?

2 Data

The dataset used was pulled from Football Reference. Football Reference shares many advanced statistics from each of Europe's top 5 leagues as well as others starting from the 2017-18 season:

- `https://fbref.com/`

3 Models Used

3.1 Multiple Linear Regression

We used different models to predict positioning and answer the question laid ahead of us. We first used multiple linear regression to predict the exact ranking of the teams in the top five leagues. Our predictor variables are everything in our data except squad, wins, losses and draws(208 features in total for each league) and our target variable is the positioning of each team. We split the data using a 80/20 ratio and trained the data using scikit-learn's built in multiple linear regression model and successfully obtained a predictor array, y_{pred}. We calculated the accuracy, and also computed the RMSE value(Root Mean Square Error) for our model. It turned out that the accuracy for each of the top five leagues was relatively low since we were predicting the exact positioning of the teams(0.09 accuracy for Bundesliga, 0.17 accuracy for Laliga, 0.08 accuracy for Ligue1, 0.08 accuracy for Premier League and 0.08 accuracy for Serie A). However, the RMSE values we got indicated that our test data fit the model we trained, which mean that multiple linear regression was a good model for our data set.(RMSE for Bundesliga 3.74, Laliga: 6.22, Ligue1:7.27, Premier League:5.96, Serie A:5.36). Since the accuracy was low, we decided to turn this into a classification problem.

We then tried to predict whether a team will qualify to European competition (the Champions League/Europa League) or relegated to the lower division based off their final rankings. We used the same split ratio and the predictor variables are the same. Our target variable becomes a categorical value from 1-4 which represents the cup qualification of a team by the end of the season. In this case, 1 indicates that a team will stay in the league and be qualified to play in the Champions League next season, 2 indicates that a team will stay in the league and play in the Europa League next season, 3 means the team will stay in the league and 4 means that the mean will be relegated to the lower division. We calculated the accuracy, and also computed the RMSE value(Root Mean Square Error) for our model. In this case, the accuracy and RMSE values for each league was improved.(Accuracy: 0.125 for Bundesliga, 0.125 for Laliga, 0.125 for Ligue1, 0.25 for Premier League and 0.375 for SerieA. RMSE: 0.98 for Bundesliga, 1.71 for Laliga, 0.84 for Ligue1, 2.19 for Premier League and 0.91 for Serie A) Based on the above stats, we conclude that Multiple Linear Regression is good as a baseline, but other models may be able to obtain better results.

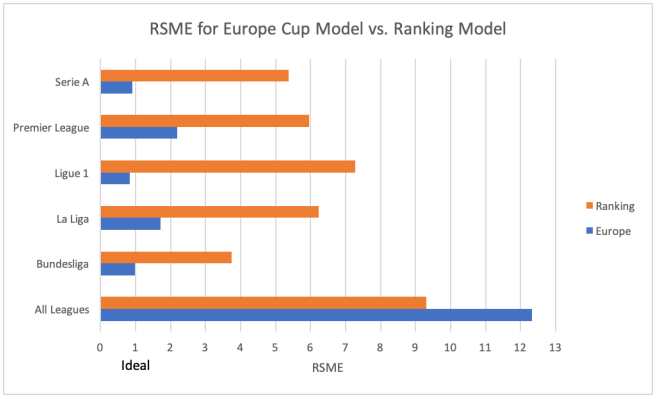

Figure 1: Multiple Linear Regression Model RSMEs

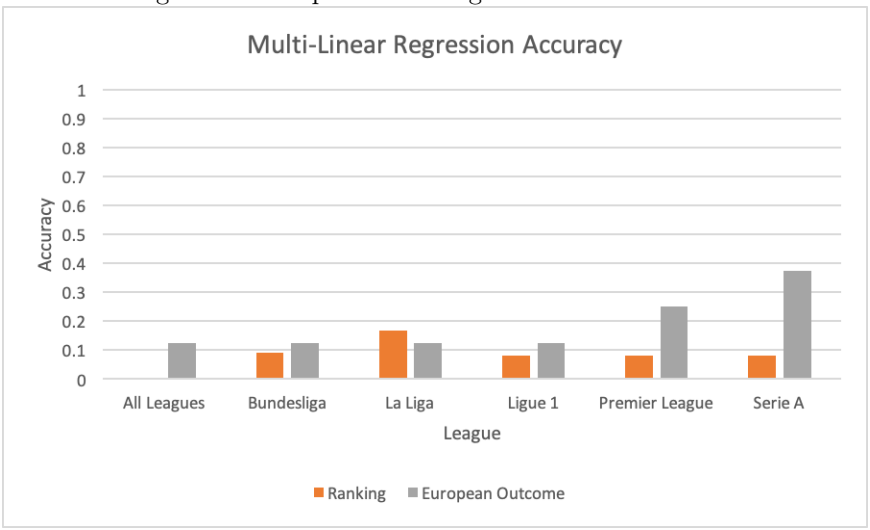

Figure 2: Multiple Linear Regression Model Tabulated Accuracy

3.2 Support Vector Machines and Kernel Methods

After testing Multiple Linear Regression, the next model we tried was Support Vector Machines. In order to train the model using support vector machines, three things were necessary to take into consideration: The columns with NaN data, the kernel methods being used and performance of the model in a unified, trans-league classifier vs. having separate models for each league to better account for the difference in play style between leagues. In order to account for this, several Support Vector Machine (SVM) models were trained. The primary difference between each model included: Whether data that was not tracked prior to the 2018-2019 season would be imputed (NaNs assigned by inference, in our case by the column mean), whether the model was trained on a multi-league dataset or trained per each league and whether a linear or gaussian kernel was used. Finally, each model was trained twice for a different target

variable. The first target variable modeled for was to predict the exact ranking of each team while the other was to predict the outcome of the team at the end of the season (whether the team would qualify European cup tournaments such as the champions league or Europa league or whether they'd stay in the same league or get relegated). After training for these different models, it was determined that the Gaussian model dropped accuracy to 0, so linear kernels produced the best outcomes. Also, composite league models were surprisingly successful in accurately in predicting European cup outcomes, although not as successful in predicting exact league rankings. The exact measured accuracy obtained for each model are as follows:

	Ranking				European Outcome	
	Linear-Imputed	Linear-Not Imputed	Gaussian-Imputed	Gaussian-Not Imputed	Linear-Not Imputed	Linear-Imputed
All Leagues	0.15	0.15	0	0	0.519	0.444
Bundesliga	0.2	0.2	0	0	0.5	0.417
La Liga	0.167	0.167	0	0	0.375	0.5
Ligue 1	0.167	0.333	0	0	0.333	0.417
Premier League	0.167	0.167	0	0	0.5	0.5
Serie A	0.167	0.167	0	0	0.417	0.542

Figure 3: Model Accuracy Comparison Table

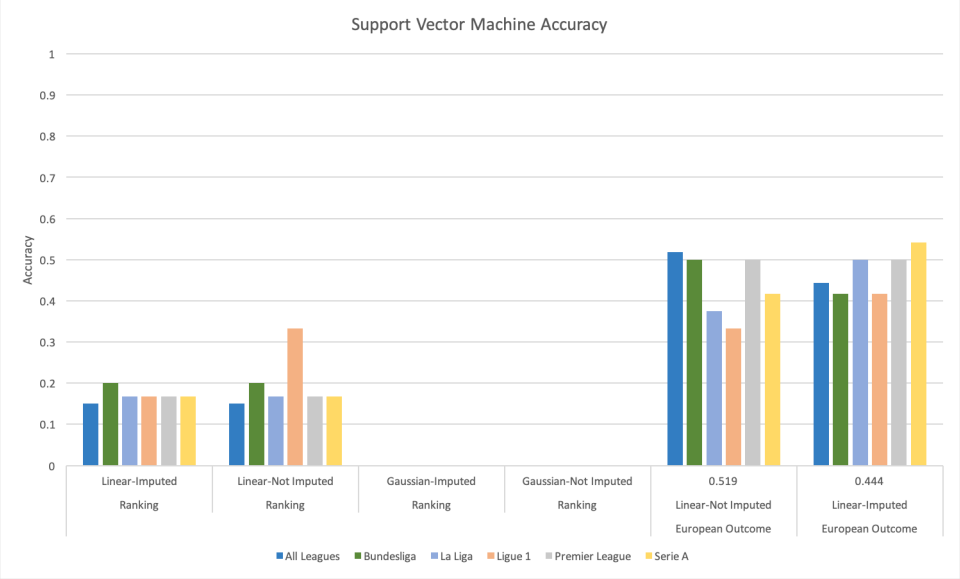

Figure 4: Support Vector Machine Model Tabulated Accuracy

3.3 Neural Networks

We then tried to apply neural networks to our data set. We used keras to build our neural network model and we used a 90/10 split ratio for our data. We added 8 hidden layers and one output layer, used binary crossentropy for the loss function and sigmoid activation function. We fit the model with epochs = 8, batch size = 10, the input size = (208,), and we obtained the following accuracy

for each of the top five leagues. (Accuracy: Bundesliga 0.25, SerieA accuracy 0.22, Premier League accuracy 0.22 ,Ligue 1 accuracy 0.17, Laliga 0.22) The accuracy obtained was similar to the multiple linear regression model, but our neural network model could be improved by using one hot encoding to label the output classes. Due to time constraint, we are currently using integers to represent different classes and this representation could not accurately predict the results through neural network.

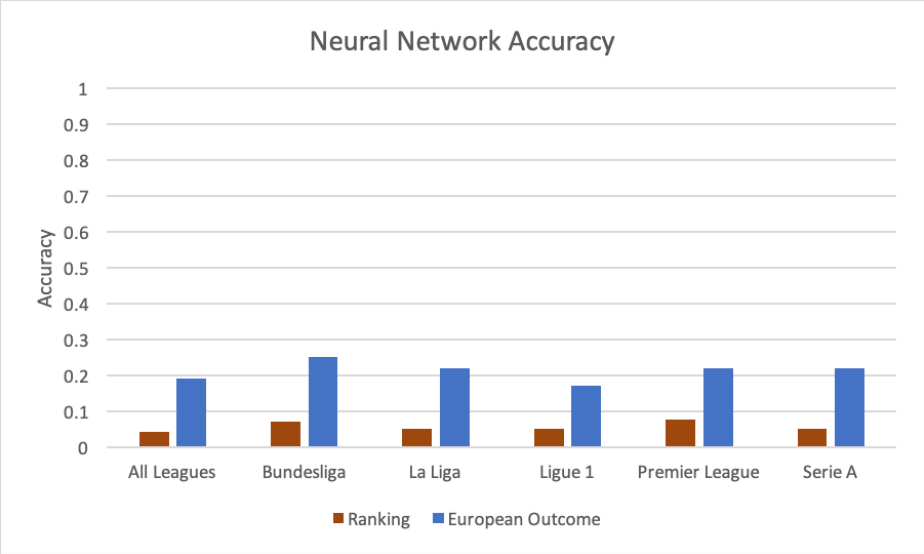

Figure 5: Neural Network Model Tabulated Accuracy

4 Possible Improvements

- Incorporate wages into the model. Include a regularization penalty to prevent overfitting to financial data
- In depth Feature Selection (Tried Variance Threshold with different logarithmic sequence of thresholds with no improvement)
- Try more Kernel Methods
- Test different ratios for the Train:Test Split
- Use one hot encoding for class labels
- Try different neural network loss functions

5 Conclusion

In conclusion, the SVM model that we used give relatively high accuracy(around 0.5) in predicting whether a team will be qualified to participate in European competition or relegated to lower divisions, while other models generate acceptable accuracy(from 0.15-0.3) for the amount of data available. Since there is only limited amount of advanced statistics in soccer (3 seasons in total), the overall accuracy of our models was not that high. We predict that as time goes on and advanced statistics is embraced more in professional soccer model accuracy will continue to improve in the future. Due to time constraint, there are some questions that we would have liked to include but could not such as more refined feature selection to determine which values are most impactful as well as deeper insights into the data such as whether referee officiating plays a large role in match outcomes. The results could be extremely useful for teams that are participating in European competition as they can use the results to analyze the play style of teams from different countries.

6 References

1. https://repository.usfca.edu/cgi/viewcontent.cgi?article=1016context=at
2. https://scrapy.org/