European Soccer Predictions

Final Report

By Israel Booth and Tuyen Huynh

With the FIFA World Cup occurring this year, the purpose of this report is to investigate the probability of winning a soccer match. There are two parts to this research. The first part consists of predicting the players' rating based on their performances, and the second part deals with predicting the game winner. The work was ultimately split with Tuyen doing the work regarding players and predicting their stats and Israel working with predicting soccer game outcomes.

First, we investigated the player ratings. The data utilized was taken from Kaggle, titled as "European Soccer Database", consisting of the years 2008 to 2016. The specific dataset utilized was "Player Attributes," and it was in SQLite format, then converted to CSV format for convenience purposes. We imported the data into a Jupyter Notebook and began cleaning it by dropping columns that were not seemingly correlated with any trend, getting rid of empty observations, and removing variables that were not finite. The dependent variable is set as the overall rating of a player. There were 33 independent variables, listed here: Crossing, Finishing, Heading Accuracy, Short passing, Volleys, Dribbling, Curve, Free kick accuracy, Long passing, Ball control, Acceleration, Sprint speed, Agility, Reactions, Balance, Shot power, Jumping, Stamina, Strength, Long shots, Aggression, Interceptions, Positioning, Vision, Penalties, Marking, Standing tackles, Sliding tackles, Diving (GK), Kicking (GK), Positioning (GK), and Reflexes (GK). Each variable was a float value that could have been as low as 0 and as high as 100, but the stats truly ranged from 50 to 99.

Afterwards, the data was split into 70-30 partitions, with 70% of the data in the training data and the rest as testing data. Two methodologies were used, the multiple linear regression model and the decision trees. The linear model was then fit using the training data. A scatterplot was created to visualize the correlation between predicted overall and actual overall rating; the R^2 is approximately 0.784, showing that 78% of the actual overall rating can be explained by the model. The model then used the testing data, and the R^2 was extremely similar, at .7796. Research was then done with the decision trees. The decision tree model was fitted with the training data, and the resulting R^2 score was 0.96. The decision tree model was then tested using the test data, and the score was 0.999, almost a perfect model. We then graphed the decision tree with Reactions being at the top, Ball control next, then Standing tackle, and then Marking. It shows that a player's reaction is the determining factor of the overall rating with reactions, ball control, standing tackle, and marking coming in afterwards. Overall, the decision tree was far more accurate compared to the multiple linear regression model, and we were happy with the results of our experiment so far.

We then began with trying to predict actual soccer games. We started with taking all the game data from a dataset created and maintained by the site with the acronym FBREF,

assumingely standing for Football Reference. This included data from the late 1990's to 2023. We also took data from EA sports FIFA game form the gears 2017-2022, so the years actually used from the game data set were from the same time frame. Initial experimentation was done only using the football game data, with no data taken from the FIFA dataset. This resulted in the prediction odds being almost at random, though the data seemed to find that 50% of the time the home team would win, with a draw occurring 25% of the time, and an away win happening 25% of the time. While this was interesting to find, it did not have the results we had hoped for. We continued research with the dataset using the FIFA player data. We ran through the dataset, taking all the players who were on a team during a game and averaging their overall scores, then inserted that into the dataset. This was not implemented in an efficient manner, leading to the time it took to do all of these calculations took too much time. We opted instead to only use he data from 2017, to get some form of result in a timely manner.

The results were lackluster. The lowest overall team score (64.24%) had about a 2.5% difference from the highest overall team score (66.86%). This made the use of these results negligible, and were not going to have a major impact on the resulting model. The model was trained nonetheless, and the results were as predicted. By this time, the timeline to try moving in a different direction for further research was not enough to get results properly, and we opted to keep what we had so far as the resulting research from our project. Further investigation using only the top players of the teams for each game may yield better results, as it can be seen that teams with stat players may have a better chance of winning games on the merit of their best on the field.