Project Executive Summary

Cody Tipton and Rafael Magaldi

The soccer transfer market is highly competitive and volatile, with player valuations frequently fluctuating. Clubs have access to two transfer windows per season, when thousands of transfers take place worldwide, moving billions of dollars. Clubs often compete to sign the best talent within their budget, so having accurate player valuations is vital when determining how much to spend on a new player. These valuations allow clubs to make informed financial decisions and ensure sustainable investments that pay off in the field. Our objective with this project is to **use in-game statistics to predict the market value of soccer players**.

Our data is sourced from three websites: **Transfermarkt** (which has the transfer fees, market values, and general stats about players worldwide); **Sofascore** (which contains all the player and team stats in every professional game, including a sofascore rating measuring how well a player did in a match); and **Understat** (which contains some general stats about each player and game in some of the top European leagues). For the Transfermarkt data, we were able to use a dataset available on Kaggle that is kept up-to-date, so we had current information on the market values of all players. As for the player stats, we performed web scraping to gather them from Understat (using Beautiful Soup) and Sofascore (using Selenium and Sofascore's API). We chose to gather data from the top 10 strongest soccer leagues in Europe, across as many seasons as they had stats for (ranging from 5 to 11 seasons depending on the league). Finally, we combined these data sources, leaving us with a **dataset of over 10k players and all their in-game stats, as well as their personal information** like name, height and date of birth.

Our stakeholders are European soccer clubs across various leagues, including both large, established clubs and smaller, emerging teams. They have a vested interest in optimizing the buying and selling of players to enhance both sporting performance and financial sustainability. These clubs seek data-driven insights to avoid overpaying for talent, identify undervalued players, and make informed transfer decisions. Ultimately, their goal is to maximize returns on player investments, maintain competitive squads, and achieve long-term financial health within the volatile football transfer market. Another potential niche for the models developed in this project are fantasy soccer players, who seek accurate player valuations to optimize their fantasy teams.

As we wanted to create a predictive model, the relevant KPIs we measured for all models were: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and R^2. We also needed to make sure our model has generalization power when presented to new data, so we measure all of these on both the training and testing sets, and compare them.

Given the player stats on each match, the approach we decided to follow was to aggregate career-wide statistics for each player (using sums or averages, depending on the feature). We also chose to use the last known market value for each player as the target, and adjusted it for inflation when it was not current. We noted the target had high skewness and kurtosis, which was causing some negative market value predictions with linear regression, so to make the distribution closer to a normal one, we decided to apply ln(1+x) to the market values. This also has the benefit of helping reduce outliers, as now the range of the target is between the values of 10 and 20. We then split the dataset into training (80%) and testing (20%) sets, ensuring diversity for player profiles and making sure there are proportional amounts of players for each position in each set.

We tested a wide range of models, including: Linear Regression with either L1 (Lasso) or L2 (Ridge) regularization, or both; K-Nearest Neighbors; Decision Trees Regression; Random Forest Regression; and Gradient Boosting Regression. We performed hyperparameter tuning on these models, and also varying approaches to feature engineering like position-specific features and choosing the minimum amount of minutes played for players to be included in the model. The main conclusion was that the best model we created was a gradient boosting regression, which performed the best out of all the models tested, **improving upon the baseline by about 24%**, while only slightly overfitting.

The main difficulties we faced were related to dealing with outliers (either players with very low minutes played, or players with very high market values). They can reduce the generalization power of our models, and cause significant overfitting when using models other than linear regression. Given these considerations, in the future we aim to expand this project by exploring the following avenues: gathering data from more competitions and/or creating synthetic data for training; including features related to “decisiveness” and “star power” of players; more hyperparameter tuning; creating different models for players in each country; giving more weight to most recent statistics; and testing different modeling approaches, like time series.