PROJECT II PROPOSAL

UNIVERSITY OF MIAMI
FINTECH BOOTCAMP
2023

by Daniel Molnar

Summary of the Project

(Coach's Corner)

This project aims to develop a machine learning model that predicts the final standings of the Premier League football season. The prediction will be based on several factors including the value of each player on each team, the total sum of all players, the sum of the players' value by their position type (eg.: goalkeeper, defender, midfielder, attacker). We will also take the age of each player, and the remaining contract length of each player in to consideration and see how it changes the outcome of the predictive models.

Scope and Purpose

(The Warm Up)

The purpose of this project is to explore how various player-level factors can predict a team's success in the Premier League. This information can be useful to stakeholders in the football industry, such as clubs, players, fans, analysts, and even the bookies. It could also potentially be used to inform decision-making about player transfers, contract negotiations, and team strategies.

Data Collection Methods and Sources

(Kick Off!)

The data for this project will be sourced from various sports analytics and statistics websites and databases, such as Transfermarkt, Whoscored, and the official Premier League website. Data to be collected will include the current and past Premier League standings, players' market values, player positions, player ages, and contract lengths.

Possible Limitations of the Data

(First Half)

While the data sources are reputable and widely used, there are potential limitations to be aware of. For instance, players' market values are estimates and can be subject to significant variability and uncertainty. Furthermore, football performance is influenced by a multitude of factors, some of which may not be captured in the data, such as players' health and fitness, coaching quality, team chemistry, and even luck.

The Type of Data to be Used and Data Preparation Methods

(Second Half)

The data used in this project will be primarily numerical (e.g., market values, ages, contract lengths) and categorical (e.g., player positions, team names). To prepare the data for analysis, I will perform cleaning to handle missing values and outliers, and I might encode categorical variables into a suitable format for machine learning algorithms. We will also carry out exploratory data analysis to understand the distribution and relationships of our variables.

Predictions and Machine Learning Model Comparisons

(Interviews)

I plan to experiment with different supervised machine learning models to predict the final Premier League standings, such as linear regression, decision trees, random forest, and gradient boosting. Then I will evaluate each model based on appropriate metrics, such as Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE), and compare their performance. I will discover using ensemble methods to combine the predictions of multiple models as they may provide better results. Finally, I will interpret the results to identify the most important features driving team success in the Premier League.

This proposal provides a roadmap for developing a machine learning model to predict Premier League standings. It is my belief that by harnessing data and machine learning, we can gain valuable insights into the factors influencing team performance in football.