ECE 208 Final Project Proposal

Team Members: Haotian, Guangzhou, Yicheng

1 TOPIC

Our project proposal is to analyze the data from the soccer statistics websites and transfer market including players' personal information and data like goals and assists. The dataset's targets are players' transfer market value.

2 MOTIVATION

Based on several reasons, we select this topic and soccer data as the main analysis source: Firstly, soccer is a highly popular sport all over the world. Our project result could be tested and utilized in various areas and soccer leagues; Secondly, with modern technologies and professional data collecting, soccer data especially in the main soccer leagues are highly usable and standardized. Thirdly, a good prediction model of players' market value can improve the efficiency of the transfer market among clubs because it saves time for negotiation and worries about fair prices. Fourthly, it is a useful tool for finding young players with great growth potential who probably hold high market value in the next several years.

3 METHOD

1) Visualization: The first step will be to comprehensively understand the distribution of features and underlying data patterns of the dataset through basic statistical plots like scatter plots and histograms. The second step will be to find some strong-correlated features and find a suitable way to visualize them.
2) Data Processing: Based on an initial review of the dataset, we anticipated that issues such as missing values, coding inconsistencies, etc. would need to be addressed. We need to find a way to ensure the quality of the data.
3) Model Selection: To predict the market value of soccer players, a sizable dataset is required. Additionally, it's essential to validate and compare the performance across these models to determine which one offers the best results for our dataset. Our initial approach will be to deploy Linear Regression, which is particularly suitable for linear relationships, such as the correlation of a player's age and performance statistics. Next, the Neural Network is advantageous for large datasets, which is able to model more intricate and non-linear relationships, as well as interactions among various features. Lastly, we suggest incorporating the Random Forest Regressor. This model, which constructs numerous decision trees during its training phase, is adept at handling complex datasets.

4 TIME TABLE

Date of start	Content and Assignments	Who	Due
Mar 25, 2024	Collect raw data	David	Apr 1, 2024
Apr 1, 2024	Process data (cleaning, feature engineering)	Haotian	Apr 8, 2024
Apr 8, 2024	Test and compare different training methods	Yicheng	Apr 15, 2024
Apr 15, 2024	Train with the best performance models and tune	David	Apr 22, 2024
Apr 22, 2024	Discover more (model ensembling, dataset pipelining, and etc.)	Haotian	Apr 29, 2024
Apr 29, 2024	Get a final model and wrap up	Yicheng	May 6, 2024
May 6, 2024	Prepare for presentation	Together	May 7, 2024