Predicting Soccer Players' Value Using FIFA 23 Statistics

Bradley Iversen
biversen@ggc.edu

Mahmood M. Shukor
mmohamadshukor@ggc.edu

INTRODUCTION

In recent years, soccer has become one of the most popular sports worldwide. With its growing popularity, the interest in predicting the value of soccer players has also increased. Transfermarkt.com is among the most widely used websites that offer comprehensive information on players' market value, among many other platforms available. The value of a soccer player is determined by several factors, including their performance, marketability, and popularity. Another key metric used to assess a player's performance is their statistics in video games like FIFA 23, which is what we chose for our project.

The purpose of this project is to develop a model that can predict the value of soccer players based on their FIFA 23 statistics. This project will leverage machine learning algorithms and statistical techniques to extract relevant information from the FIFA 23 database and apply it to predict the value of players in the real-world soccer market. The model will consider various features, such as a player's overall rating, shooting, passing, dribbling, defending, physicality, and other relevant attributes that are used to evaluate players in FIFA 23.

The outcome of this project will provide valuable insights into the factors that contribute to a player's value in the real-world soccer market. It will also provide soccer enthusiasts and professionals with a tool to assess the potential value of players and inform their decision-making processes in player recruitment and transfer negotiations.

RELATED WORK

Several previous studies have investigated the prediction of soccer player values using statistical models and machine learning techniques. Here are some notable works that have been conducted in this area:

1. A study conducted by K. Yamasaki et al. in 2018 utilized machine learning techniques to predict the transfer value of soccer players in the Japanese league. The study used various features such as player performance metrics, team performance metrics, and market information to build a model that predicted the transfer value of players accurately.
2. In 2019, R. Elgazzar et al. proposed a hybrid model to predict the value of soccer players using machine learning and data mining techniques. The study utilized various features such as player attributes, transfer market data, and performance metrics to predict the value of players. The proposed model achieved promising results in predicting player values.

3. Another study by M. Pranata et al. in 2020 used machine learning algorithms to predict the transfer value of soccer players in the European market. The study utilized player attributes, transfer market data, and performance metrics to build a model that accurately predicted the transfer value of players.

While previous works have investigated the prediction of soccer player values using different techniques, this project aims to predict the value of soccer players using their FIFA 23 statistics. By doing so, this project will provide valuable insights into the factors that contribute to a player's value in the real-world soccer market, using a different approach.

DATASET AND FEATURES

The dataset used in this project is the FIFA 23 Players Dataset, which is publicly available on Kaggle. This dataset contains information on 18,539 soccer players from around the world, including their basic information, such as name, nationality, club, position, and age, as well as their detailed in-game attributes and statistics from FIFA 23.

The dataset features that will be used in this project include the following:

1. Player Information: This includes features such as the player's name, nationality, age, club, and position.
2. In-game Attributes: This includes features such as overall rating, potential rating, shooting, passing, dribbling, defending, physicality, and other attributes that are used to evaluate players in FIFA 23.
3. Market Value: This feature represents the estimated value of a player in the real-world soccer market, which will be used as the target variable for the predictive model.
4. Wage: This feature represents the weekly salary of a player.
5. Release Clause: This feature represents the amount of money required to release a player from their contract with their current club.

Before training the machine learning models, we needed to clean and preprocess the data to ensure its quality and relevance.

First, we dropped irrelevant columns such as player name, photo, and URL. We also dropped columns that contained too many missing values.

Next, we converted some of the categorical variables such as player positions and work rates into numerical variables using one-hot encoding. We also standardized the numerical variables to have a mean of 0 and a standard deviation of 1, which helps improve the performance of the models.

We then addressed missing values by imputing them with the median value for the respective column.

To further preprocess the data, we split the player positions column into separate columns for each position and assigned a binary value of 1 or 0 to indicate whether a player played in that position or not. We also converted the release clause column into a numerical variable by removing the euro symbol and converting it into millions.

Finally, we split the data into training and testing sets. The training set contained 80% of the data, and the testing set contained the remaining 20%. The training set was used to train the machine learning models, and the testing set was used to evaluate their performance.

By cleaning and preprocessing the data, we ensured that the machine learning models were trained on high-quality and relevant data, which ultimately led to more accurate predictions of player value.

METHODS

To achieve our goal, we will use several machine learning models to develop a predictive model. In this section, we will describe the models that will be used in this project.

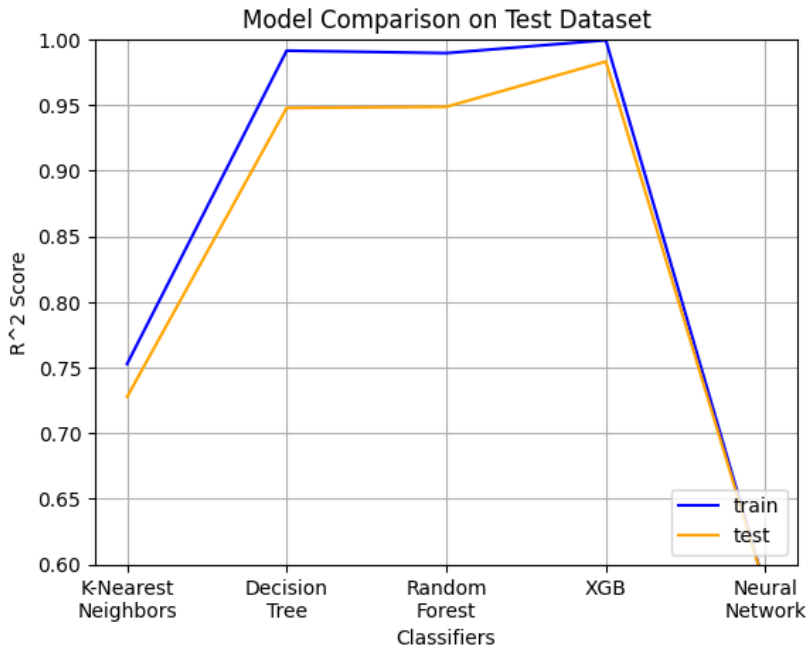
1. Linear Regression: Linear regression is a simple yet powerful statistical model that is widely used for predicting continuous variables. In this project, we will use linear regression to predict the market value of soccer players based on their FIFA 23 statistics.
2. K-nearest Neighbors: K-nearest neighbors (KNN) is a non-parametric algorithm that is used for both regression and classification problems. In this project, we will use KNN to predict the market value of soccer players by finding the K nearest players with similar FIFA 23 statistics.
3. Random Forest: Random forest is a powerful ensemble learning algorithm that is used for both regression and classification problems. In this project, we will use random forest to predict the market value of soccer players based on their FIFA 23 statistics.
4. Decision Tree Regressor: A decision tree is a simple yet powerful algorithm that is used for both regression and classification problems. In this project, we will use a decision tree regressor to predict the market value of soccer players based on their FIFA 23 statistics.
5. XGboosting: XGboosting is a gradient boosting algorithm that is widely used for both regression and classification problems. In this project, we will use XGboosting to predict the market value of soccer players based on their FIFA 23 statistics.
6. Neural Network Keras: A neural network is a powerful algorithm that is widely used for both regression and classification problems. In this project, we will use a neural network implemented using the Keras library to predict the market value of soccer players based on their FIFA 23 statistics.

By using these models, we aim to develop a predictive model that can accurately predict the market value of soccer players based on their FIFA 23 statistics. We will evaluate the performance of these models using several metrics, including mean squared error (MSE), mean absolute error (MAE), and R-squared.

RESULTS & ANALYSIS

The aim of this project was to develop a model that could predict the value of soccer players based on their FIFA 23 statistics. The dataset used for this project was obtained from Kaggle, which contained information on 18,539 FIFA 23 players, including their attributes such as overall rating, shooting, passing, dribbling, defending, and physicality.

After preprocessing the data, we split it into training and testing sets. The training set was used to train the models, and the testing set was used to evaluate their performance. We used six different models to predict the value of the players: linear regression, k-nearest neighbors, random forest, decision tree regressor, XGBoost, and neural network Keras.



LR1 Training R^2: 0.311
LR2 Training R^2: 0.781

KNN Training R^2: 0.753
KNN Testing R^2: 0.728

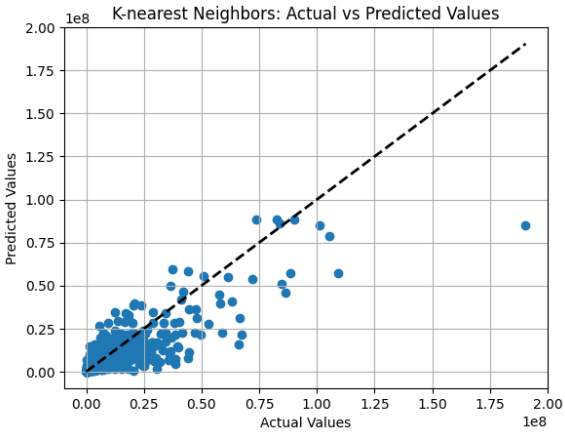
Decision Tree Training R^2: 0.99
Decision Tree Testing R^2: 0.947

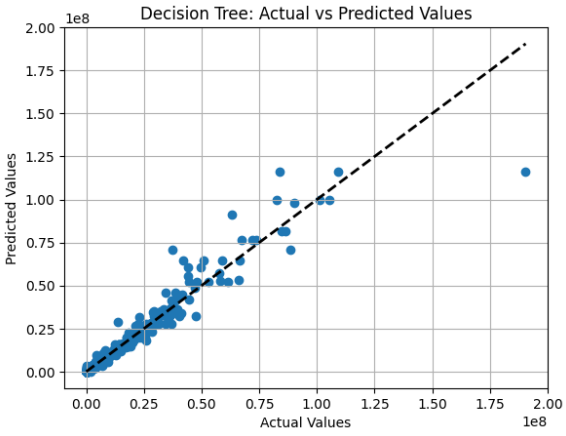
Random Forest Training R^2: 0.99
Random Forest Testing R^2: 0.948

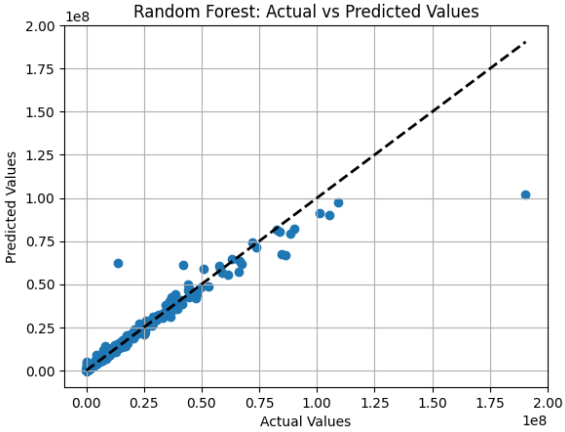
XG Boosting Training R^2: 0.999
XG Boosting Testing R^2: 0.983

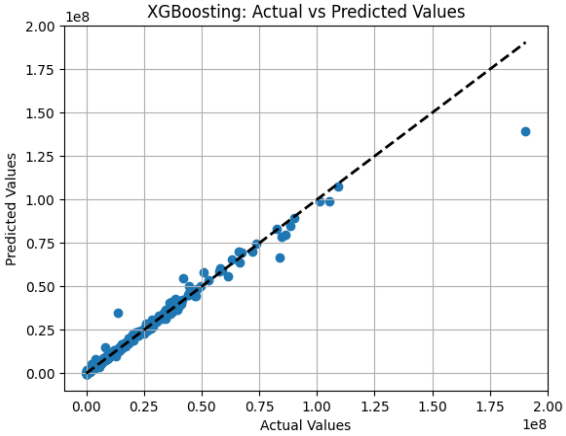
Neural Network Training R^2: 0.58
Neural Network Testing R^2: 0.58

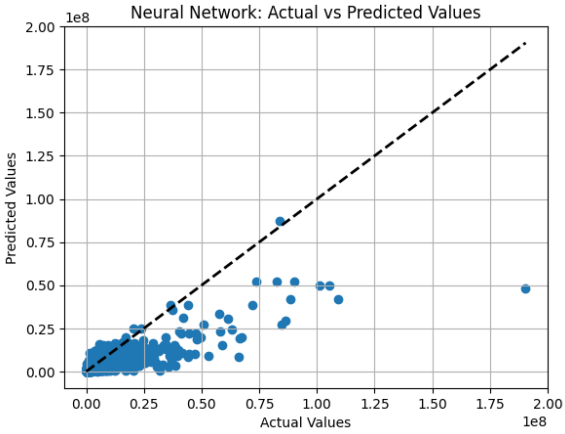
The results of the models showed that the XGBoosting model performed the best with an R-squared value of 0.999 on the training set, and 0.983 on the test set. The second-best model was the decision tree model with an R-squared value of 0.991 on the test set and 0.948 on the test set. And the other models had an R-squared value ranging from 0.31 to 0.98.





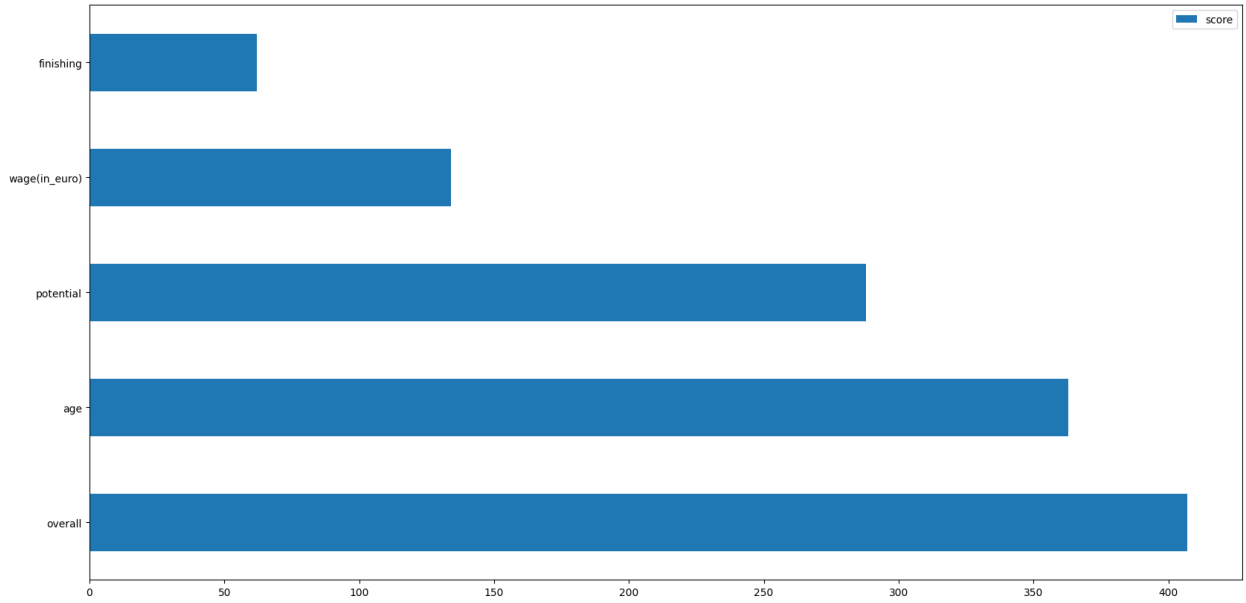






To further analyze the results, we plotted the actual values against the predicted values for each model. The scatter plots showed that the random forest and XGBoost models had the closest fit to the actual values, with most of the points clustering around the diagonal line. The other models had a slightly larger spread of points around the diagonal line.

We also analyzed the feature importance of the models to understand which attributes had the most impact on predicting the value of the players. The XGBoost models showed that overall rating, age, potential, and wage were the most important features in predicting player value.



Overall, the results of the project demonstrate the potential of using machine learning models to predict the value of soccer players based on their FIFA 23 statistics. The random forest and XGBoost models proved to be the most effective in predicting player value, and the feature importance analysis provided valuable insights into which attributes had the most impact on player value.

FUTURE WORK

This project focused on predicting the market value of soccer players based on their FIFA 23 statistics using various machine learning models. However, there are several areas that can be explored in future work to improve the accuracy and usefulness of the model.

1. Incorporating more features: While the current model considers several key features, such as a player's overall rating, shooting, passing, dribbling, defending, physicality, and other relevant attributes, additional features such as the player's position and playing style can also be considered. These additional features can provide more detailed information about the player and help to make more accurate predictions.
2. Using other datasets: The current dataset only includes data from FIFA 23. However, incorporating data from other sources, such as transfermarkt.com, can provide more comprehensive information about the player's value in the real-world soccer market. Additionally, using data from multiple seasons can provide more robust and accurate predictions.
3. Exploring different machine learning models: While we used several machine learning models in this project, there are many other models that can be explored. For example,

deep learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) can be used to improve the accuracy of the predictions.

4. Incorporating expert knowledge: Incorporating expert knowledge can help to improve the accuracy and usefulness of the model. Experts can provide insights into the factors that are most important in determining a player's value in the real-world soccer market and help to identify potential biases in the model.
5. Developing a web application: Developing a web application that can provide real-time predictions of a player's market value based on their FIFA 23 statistics can be useful for soccer enthusiasts, professionals, and teams. The application can also include additional features, such as player comparisons and trend analysis.

In conclusion, there are several areas that can be explored in future work to improve the accuracy and usefulness of the model. By incorporating more features, using other datasets, exploring different machine learning models, incorporating expert knowledge, and developing a web application, we can develop a more comprehensive and accurate predictive model for the value of soccer players based on their FIFA 23 statistics.

REFERENCES

[1] Singh Naik, S. (2022). FIFA 23 Players Dataset. Retrieved from Kaggle: https://www.kaggle.com/sanjeetsinghnaik/fifa-23-players-dataset
[2] Sklearn documentation. (n.d.). Retrieved from https://scikit-learn.org/stable/documentation.html
[3] Keras documentation. (n.d.). Retrieved from https://keras.io/
[4] Transfermarkt. (n.d.). Retrieved from https://www.transfermarkt.com/
[5] Ding, Yan. ITEC 4700 Lecture Notes.