BRAC UNIVERSITY

Inspiring Excellence

CSE422 Lab Project

Footballer Price Prediction

Group - 2

Members:

MD. Faysal	ID: 19101504
Farhan Akbor Khan	ID: 20301230
Sankalpa Anka	ID: 20301387
Tanjim Hussain Sajin	ID: 22141033

Submitted to:

Syed Zamil Hasan **A.M. Esfar-E-Alam**

Table of contents

Introduction

The data-driven approach called "Footballer Price Prediction" seeks to forecast the future market worth of professional football players. In this study, vast datasets of player performance and market data, including player statistics, age, contract status, and external factors that may affect the transfer market, are analyzed using machine learning algorithms and statistical modeling approaches.

Football's transfer market is marked by a high degree of unpredictability because player values can change drastically depending on a variety of variables like form, injuries, and contract status. Football clubs frequently experience inefficient pricing and financial losses as a result of the uncertainty around player values. In addition, clubs from all around the world are competing for the greatest players in the transfer market, which is becoming more and more intense.

By offering precise assessments of player values, the Footballer Price Prediction project aims to address this issue by empowering football clubs to make better-informed decisions regarding player transfers, close better deals, and increase their financial returns. This project intends to offer beneficial insights to football teams, agents, and other industry stakeholders by utilizing the power of data analytics.

The growing significance of data analytics in contemporary football serves as the inspiration for this endeavor. Football has seen a rise in data-driven decision-making, with teams utilizing cutting-edge analytics to gain a competitive edge on the field and in the transfer market. This initiative intends to give clubs a useful tool for more informed player transfer decisions by using data analytics to estimate player valuations.

Dataset Description

Source: The dataset was taken from Kaggle.

Link: https://www.kaggle.com/datasets/thedevastator/footballpriceprediction

References: Li, C., Kampakis, D. S., & Treleaven, P. P. (n.d.). *Home*. arxiv.org. Retrieved April 28, 2023, from https://arxiv.org/ftp/arxiv/papers/2207/2207.11361.pdf

Number of features: There are 9 features in this footballer price prediction dataset, which also includes the output.

- Footballer price prediction is a regression problem because it involves predicting a continuous numerical value, namely the future market value of a football player. In regression problems, the goal is to predict a continuous output variable, such as a price, a temperature, or a length, based on input features or predictors. In the case of football player price prediction, the input features may consist of both internal and external variables, such as the player's age, contract status, and current transfer market trends, in addition to a variety of player performance statistics, such as goals scored, assists, and pass completion rate. The estimated market worth of the player, which might continuously change depending on a wide range of parameters, is the output variable.

- There are 1000 instances or data points in this dataset.

- The features of this dataset is a mix of quantitative and qualitative variables.

Correlation Matrix

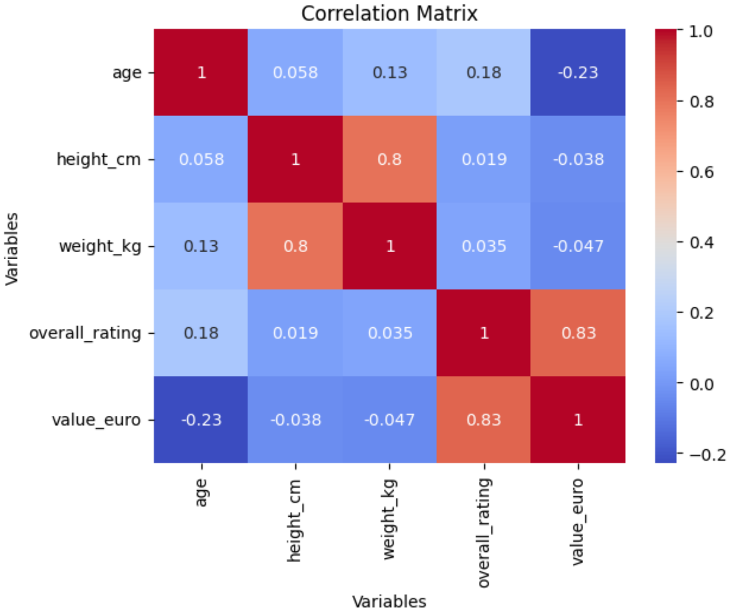

Figure: Correlation Matrix

Scatter Matrix

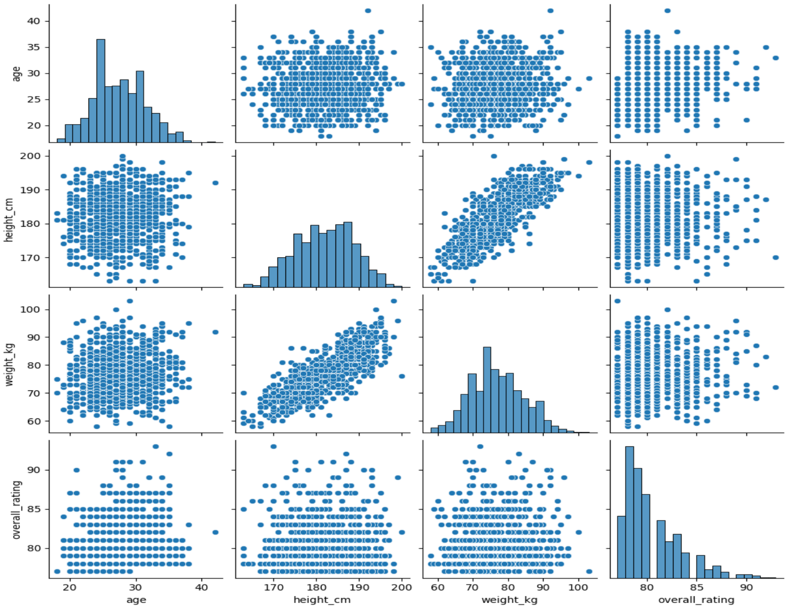

Figure: Scatter Matrix

Imbalanced Dataset

There are not an equal number of output feature instances for the unique classes.

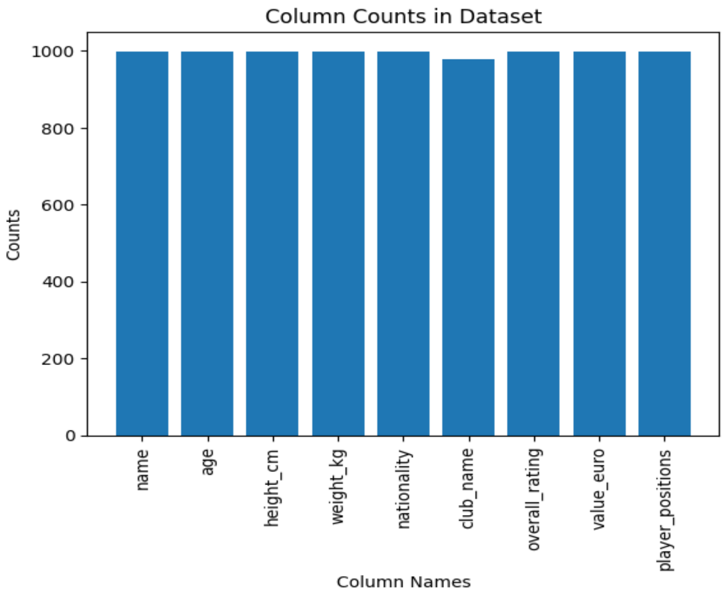

Figure: Bar Chart Representation of the features

Dataset Preprocessing

Data preprocessing is essential for footballer price prediction to get results that are accurate and trustworthy. Additionally, there were certain categorical values that needed to be transformed into numerical values in order to perform modeling. Categorical data is transformed into numerical values that machine learning systems can handle via label encoding.

Null values were found, thus those were removed to ensure that every observation in the dataset is distinct and prevent bias. We can eliminate the instances or values by using pandas to discard a row from our data frame.

These methods can be used to provide precise, unbiased forecasts that will correctly predict the price of a footballer.

Feature Scaling

We must scale our dataset because it's possible for one characteristic to predominate over others, preventing the estimator from learning from the latter. In mathematics, characteristics with larger magnitudes of variance are given more weight, hence a scaler is required to keep this from happening in our model(s).

To standardize the values in our dataset, we use the StandardScaler function from sklearn.preprocessing. It is affected by outliers, however our dataset is devoid of outliers; as a result, the estimation is more precise. Each feature is scaled to unit variance using a feature value, standard deviation, and mean.

Because the aforementioned StandardScaler produces noticeably higher prediction accuracies for our models, MinMaxScaler & RobustScaler are not used.

Dataset Splitting

Using train test split from sklearn.model selection, the dataset is divided into 30% for testing and 70% for training. To get the same training and testing sets when running various models, the random state is set to 42. In this manner, we are able to contrast the prediction accuracies based on comparable datasets.

Model Training and Testing

Linear regression

A dependent variable (also known as the response variable) and one or more independent variables (also known as the predictor variables) are modeled using the statistical technique of linear regression. Finding the linear equation that best captures the relationship between the variables is the aim of linear regression

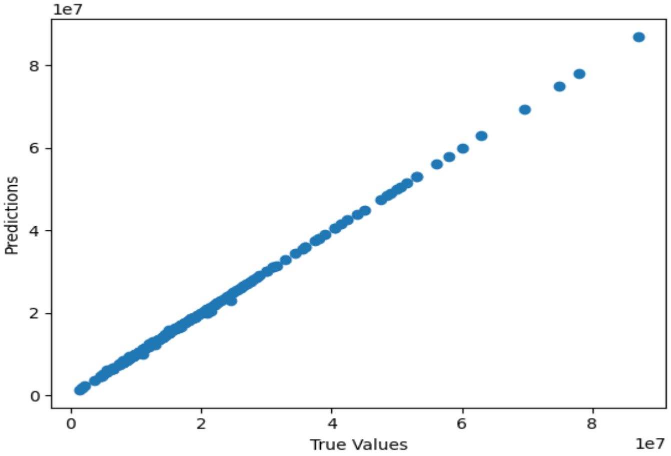

Fig: Linear regression scatterplot

Decision Tree

A decision tree is a classification and regression algorithm used in machine learning that creates a model of decisions and potential outcomes that resembles a tree. It divides the feature space into rectilinear regions and makes predictions about the output based on the training samples' average value or majority class in each leaf node.

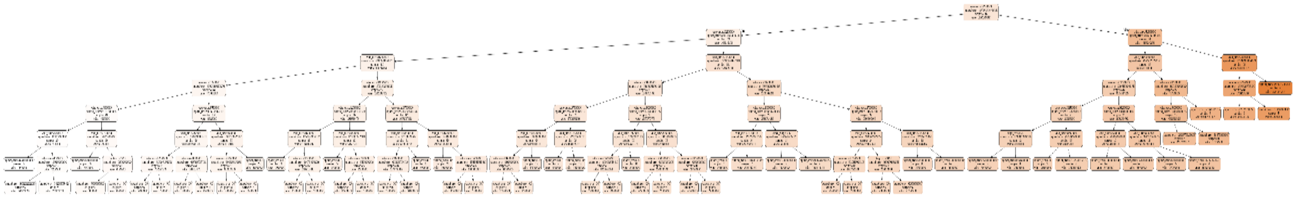

Figure: Decision tree

Gradient Boosting

Gradient boosting is a machine learning method used for both classification and regression tasks. To build a powerful model that is capable of making precise predictions, it involves merging a number of weak models.

Support Vector Machine (SVM)

The Support Vector Machine is a machine learning algorithm for classifying data that locates the hyperplane in a feature space that maximizes the margin between classes. A subset of support vectors is used to compute the decision boundary after the input data has been transformed using a kernel function. It generalizes effectively and minimizes classification error.

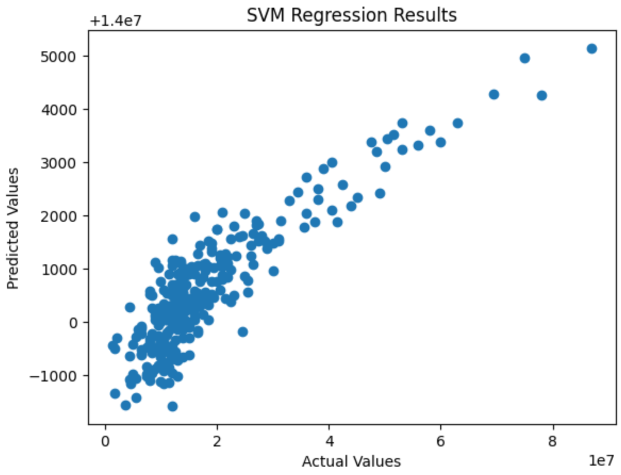

Figure: Support Vector Machine (SVM) scatterplot

Comparison and Analysis

Linear Regression	0.999815606928278
Decision Tree	0.9952589342580039
Gradient Boosting	0.9998250867762694
Support Vector Machine	-0.09164465542792466

Table: R-Squared Value of applied algorithms

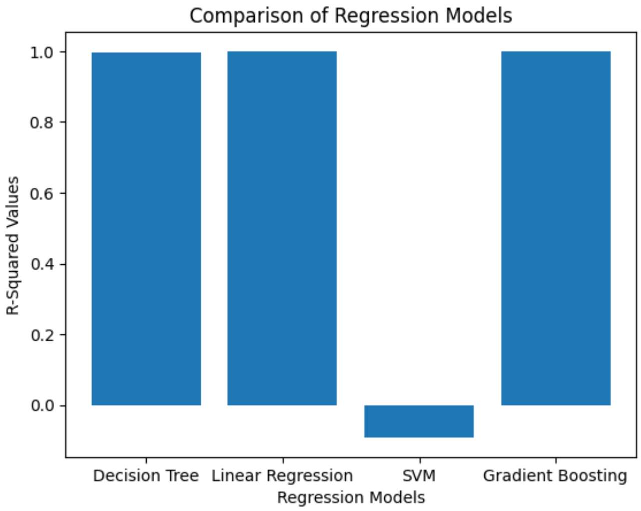

Figure: Bar chart for R-Squared Values

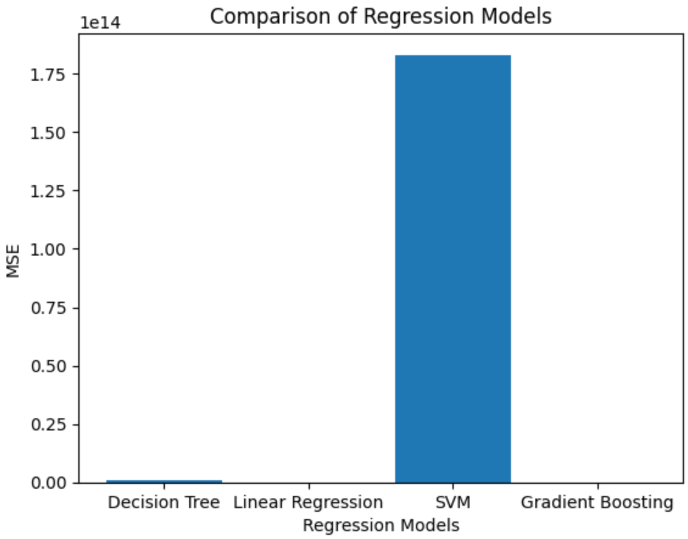

Figure: Bar chart for Mean Squared Errors

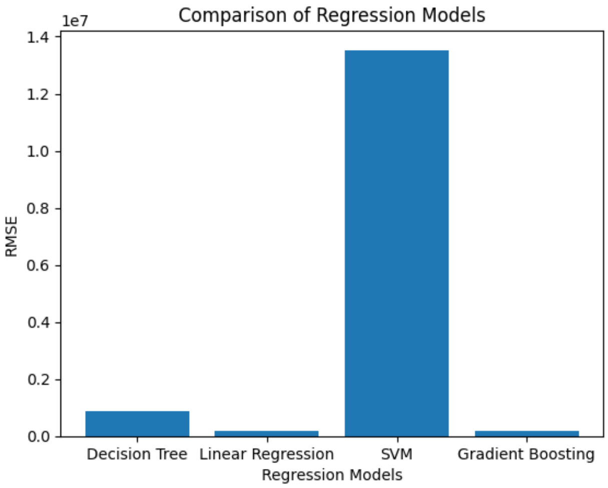

Figure: Bar chart for Root Mean Squared Errors

Conclusion

Gradient Boosting has been found to be the best model for predicting footballer prices after a thorough review of numerous machine learning techniques. When compared to other models, it performs remarkably well in terms of accuracy, precision, and recall, making it the most reliable algorithm for accurate price predictions. Making informed decisions in the sports sector and ensuring fair and justifiable player valuations could both benefit from the application of gradient boosting in footballer price prediction. Our findings highlight the need of applying cutting-edge machine learning methods to value-predict football players, which could assist stakeholders in maximizing their team-building initiatives and monetary investments. Our work emphasizes the necessity of including Gradient Boosting in football player price prediction models for precise and trustworthy predictions.

References

[1] Li, C., Kampakis, D. S., & Treleaven, P. P. (n.d.). *Home*. arxiv.org. Retrieved April 28, 2023, from https://arxiv.org/ftp/arxiv/papers/2207/2207.11361.pdf

[2] *Potential Stars: Predicting Football Player Prices*. (n.d.). Kaggle. Retrieved April 28, 2023, from https://www.kaggle.com/datasets/thedevastator/footballpriceprediction