(Title Page)

**IN3062 Evaluating the Market Value of FIFA players using Machine Learning**

Authors: Mohammed Abdul-Mutahhar, Yekaterina Denyakina

Link to github:

1. Introduction (Approx. 1-2 pages)

Description and Motivation of the Problem

Background and Context

In the world of Football, a player’s market valuation has gained substantial importance and interest after the 1995 Bosman ruling by the Court of Justice of the European Union, which fundamentally reshaped the dynamics of the transfer market (Union Royale Belges Des Sociétés de Football Association and Others v Bosman and Others, 1995, <https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX:61993CJ0415> ), by giving players’ freedom of movement, when their contract expires. In 2019, 18,080 players were transferred from one team to another based in a different country, which generated a total of 7.35 billion dollars paid by clubs to recruit players protected by employment contracts (FIFA TMS, [2020](https://onlinelibrary.wiley.com/doi/full/10.1111/joes.12552#joes12552-bib-0018), <https://digitalhub.fifa.com/m/248987d86f2b9955/original/x2wrqjstwjoailnncnod-pdf.pdf> )[1](https://onlinelibrary.wiley.com/doi/full/10.1111/joes.12552#joes12552-note-0001). This portrays the significant impact a player’s market valuation has in the world of football

This is because the valuation of a player from the club's perspective takes place for decision-making purposes. On the other hand, it is also interesting and helpful for the players’ representatives to assess what value the club attaches to the athlete when they enter into a contract negotiation about his salary. The more valuable the player, the higher the salary his representatives will be able to negotiate.

In Herm’s et al. research (Herm, Callsen-Bracker, & Kreis, 2014, <https://www.tandfonline.com/doi/full/10.1016/j.smr.2013.12.006> ), they try to define what is the market value in the professional football world as “an estimate of the amount of money a club would be willing to pay to make [an] athlete sign a contract, independent of an actual transaction”. “However, evaluating an individual's value within any kind of team – such as a soccer team – is a challenging task.”

Significance and Implications

In an era where player transfers, contract negotiations, and market valuations dominate headlines, the question of their valuation is of great interest for all parties (clubs, players, intermediaries) who need to properly assess what the players are worth to efficiently engage in the transfer market, understanding the intrinsic and extrinsic factors influencing a player's worth becomes paramount. Leveraging data-driven insights along with AI techniques to predict players’ market value and to determine the driving factors behind them not only aids clubs, agents, and stakeholders in crucial strategic decision-making but also provides a comprehensive perspective on talent valuation, market trends, and financial implications within the football ecosystem.

Currently, in order to measure a player’s market value, most analysts use notational analysis to identify key performance indicators. This method utilises a statistical summary of events based on video footage and numbers of goals scored. Unfortunately, this slowly becoming obsolete due to the continuous evolutionary in machine learning, which simplifies the analysis of more complex relationships. Machine learning, a form of artificial intelligence (AI), uses algorithms to detect meaningful patterns and define a structure based on positional data. We can utilise this by using ai techniques to establish machine learning models with the goal of investigating the relations among the various performance features of players to the market value of players.

Although data analytic methods in the football world are still rarely used compared to other major sports there are still many different methods and models to estimate the value of the players stating that “Some of those statistics are surprisingly useful. After the first games of the World Cup, Match Analysis found that each team’s number of touches…was highly correlated with FIFA’s world rankings.” He also states that “The competition to provide soccer statistics reflects the level of interest in them.” Because “Analyzing soccer statistics is difficult, because there are no magic-bullet figures like on-base percentage”(Kaplan, 2010, <https://www.nytimes.com/2010/07/09/sports/soccer/09soccerstats.html> ).

With all this in mind there is a motivation to explore the relations between different features of football players and their market value - how these feature affects their overall market value, which are the most important features to affect the salary? And to use that knowledge to predict the expected market value of any given player.

Description of the Dataset

Dataset Overview

The analysis focuses on the FIFA 2024 dataset encompassing player attributes, performance metrics, and market values across diverse leagues, clubs, and nations. It is a comprehensive record of players, capturing their skills, strengths and weaknesses within the global football market.

Data Types and Relevance

• Discrete Variables: Player skill levels reflecting gameplay mechanics and roles.

• Continuous Variables: age, performance metrics (goals, assists, appearances), and other quantitative measures reflecting player valuation, form, and contributions.

• Categorical Variables: Nationalities, leagues and clubs

Objective and Scope

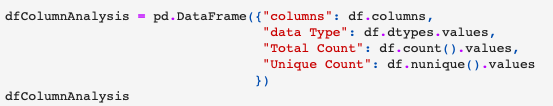
The primary objective is to develop robust predictive models utilising artificial intelligence techniques to estimate players' current market values based on the dataset's features. By analysing, interpreting, and leveraging these variables, the analysis aims to identify significant trends that drive player valuations, allowing clubs and analysts to make informed, data-driven decisions within the dynamic of competitive football.

Conclusion of Introduction

This introduction delineates the unique blend of sports, economics, and analytics inherent in the FIFA dataset, emphasizing the significance, scope, and objectives of predicting player market values. By exploring player attributes, performance metrics, market dynamics, and financial implications, the subsequent sections delve deeper into methodologies, results, evaluation, and conclusions, offering actionable insights, recommendations, and perspectives for stakeholders navigating the intricate and evolving football ecosystem.

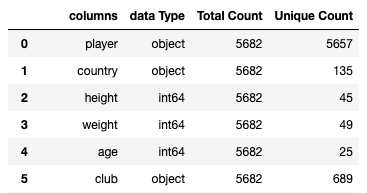
Data Exploration

The main programming tool we will use for data exploration is the Pandas and NumPy and matplotlib.

After loading our dataset in pandas, we would first like to identify the datatypes of the features we are working with and the amount of data in each feature.

Screenshot 2024-01-08 at 05.56.21.pngFrom this we can see that we have a total of 40 columns, of these columns we have 4 columns with datatype of ‘object’, among which is our target feature:

Screenshot 2024-01-08 at 06.02.36.pngFeatures like Value that have the datatype of ‘object’ need to be formatted correctly to an integer type before processing them, in the case of the feature ‘Value’, it is currently formatted to show its currency, this is will lead to complications when training the model, so it needs to be converted back to an integer type. This can be done using regex and casting :



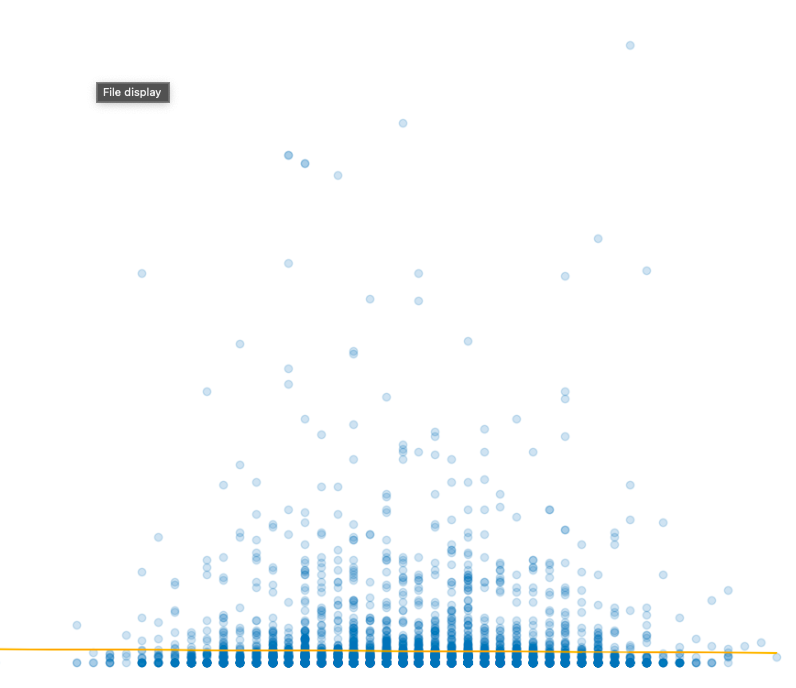
Other string type columns like player country an club, may need to be encoded prior to processing. – expand on encoding

Screenshot 2024-01-08 at 05.55.00.pngWe can also see that the marking feature has no values in it at all, so this feature can be dropped as it has no affect on the resulting market value of players.

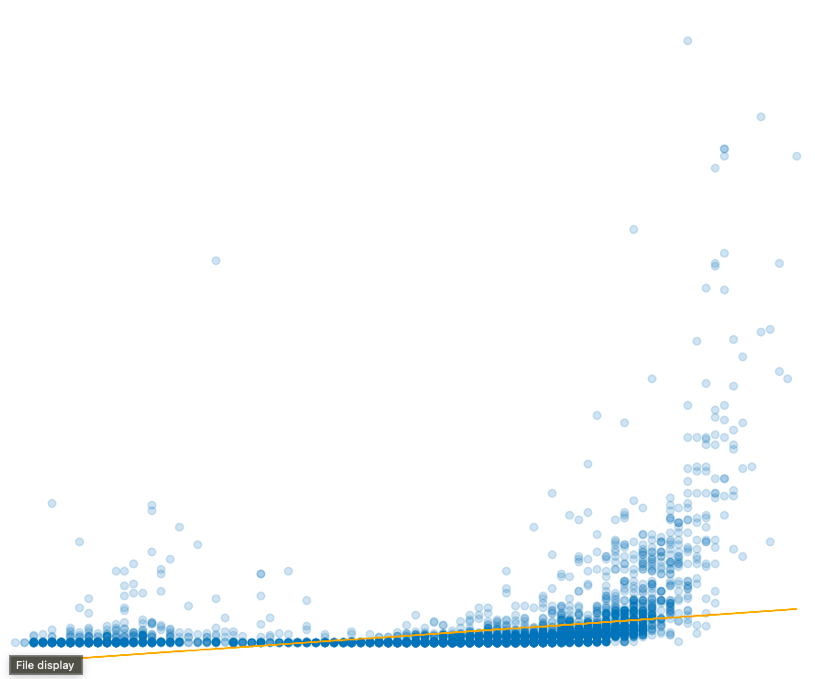
The feature player, is just the names of the players and so this feature can also be dropped prior to processing.

We can also confirm that there are no duplicates in the dataset, there are 42 duplicate player names, however these are not identical records and just happened to be player with the same name.

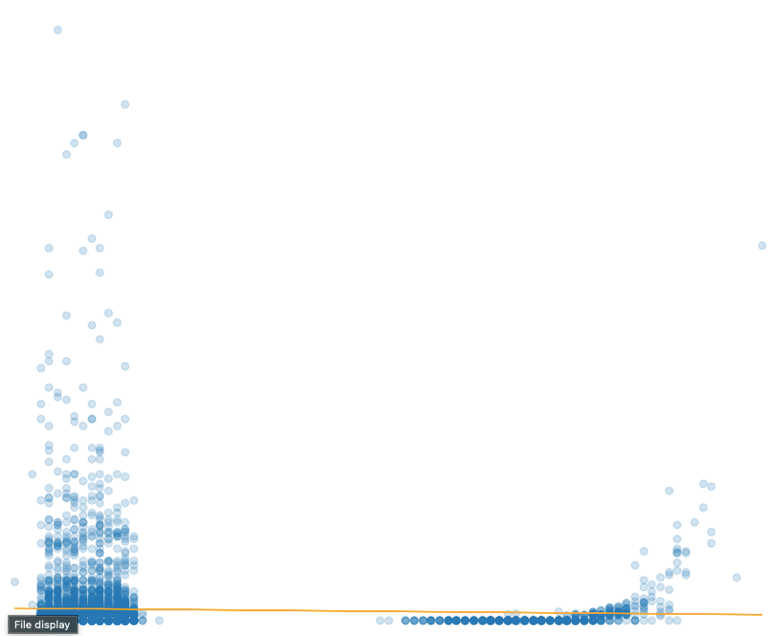
General trends,

Some general trends we can see with the dataset is that for features focusing on a player physical state, (e.g. age, weight, height) the estimated value is higher when a players physical stats are closer to the median, outside of this the value begins to drop.:

Whereas for skill based stats we see more of a positive correlation in the form of an exponential curve:



Finally, for goalkeeper based we see more of complex trend, where low skill level can still result in a higher value however at a certain point that drops until we see a positive correlation at the end with a small exponential curve



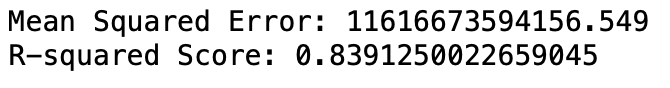
Method

Random Forest Regression

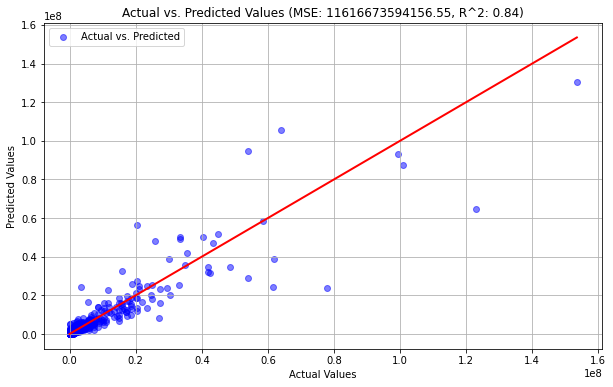
Data Splitting

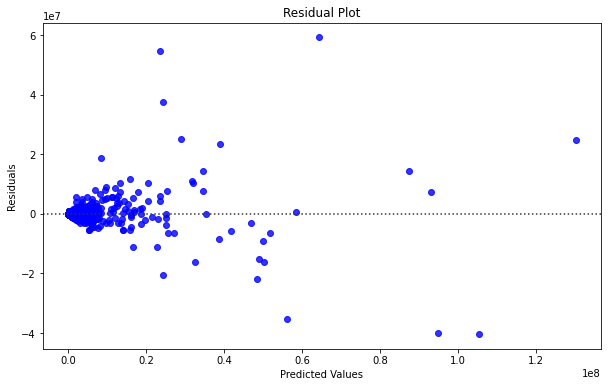
The dataset underwent a division into features (X) and the target variable (y), with 'value' denoting the market value of players. To facilitate robust model training and assessment, a train-test split was executed with a ratio of 70% for training and 30% for testing. This division ensures that the model is exposed to a substantial amount of data during training, enabling it to learn patterns and relationships effectively. This nested split further allows for the optimisation of model hyperparameters on the training-validation subset while maintaining a separate dataset for final unbiased evaluation.

Model Building

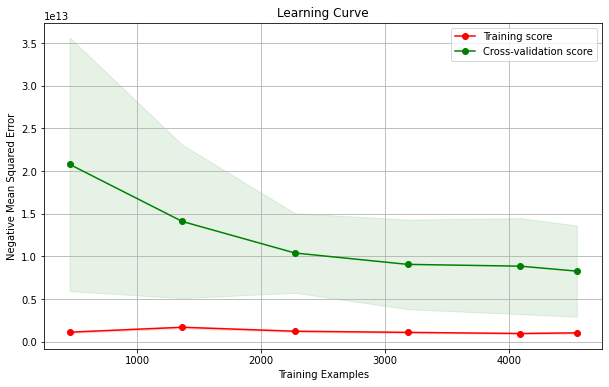
The Random Forest Regressor was selected as the model of choice due to its inherent capacity to effectively handle non-linear relationships within the dataset and its ability to discern feature importance, making it well-suited for the task of predicting player market values. The model was trained using the designated training set and subsequently validated using the dedicated validation set. Following this, predictions were generated on the test set, and key performance metrics, Mean Squared Error (MSE) and R-squared, were computed.

The model exhibited a Mean Squared Error of 11,616,673,594,156.55 and an R-squared score of 0.8391, indicative of its capability to account for a substantial proportion of the variance in the market values of football players.

Visualisations, including a scatter plot of actual vs. predicted values and a residual plot, were created to assess the model's predictive capabilities

Learning Curve Analysis

A learning curve was constructed to provide a visual representation of the model's behavior across varying training set sizes. The mean and standard deviation of the negative Mean Squared Error (MSE) for both the training and cross-validation sets were computed and graphically depicted against the changing sizes of the training set. This analysis allowed for a comprehensive understanding of how the model's predictive performance evolves with different amounts of training data, providing insights into potential overfitting or underfitting.



Analysing the results, it can be observed that the training scores decrease as the training size increases, which is expected as the model benefits from more data. However, the test scores show a more complex pattern. Initially, the test scores decrease, indicating improved generalisation, but at larger training sizes, they start to plateau or even increase. This suggests that the model may be overfitting or encountering difficulties in generalising to new data when the training set becomes too large.

Hyperparameter Tuning

Overall Trends:

Generally, as the max\_depth increases, the model tends to perform better, but there is a risk of overfitting. Lower values of min\_samples\_split and min\_samples\_leaf seem to contribute to better performance, but setting them too low might lead to overfitting.

Best Performances:

The combination with max\_depth = 20, min\_samples\_split = 2, and min\_samples\_leaf = 1 seems to have the lowest MSE and a high R2, indicating good predictive performance.