Predicting Player Market Value in FIFA 20

Problem Statement:

Football is the most popular sport globally, with professional clubs and leagues heavily investing in players to build competitive teams. A player's market value serves as a critical indicator of their demand, skill level, and future potential, directly influencing transfer decisions, contract negotiations, and financial strategies of clubs. For this project, I aim to develop a predictive model capable of estimating the market value of professional football players using key performance metrics and attributes such as age, overall rating, potential, and skill-specific attributes.

Accurately predicting a player's market value can provide valuable insights for stakeholders within the football industry, including football clubs, scouts, sports analysts, and agents. Clubs can use this information to make data-driven decisions when scouting for new talent, negotiating transfers, or allocating budgets for player acquisitions. Agents and analysts can assess player performance to ensure fair valuations and identify undervalued players with high potential.

By analyzing features such as overall performance ratings, potential ratings, wages, and age, this project aims to uncover the factors that most influence player market value. Additionally, the model will highlight the relationships between player attributes and valuation trends, enabling clubs to strategically invest in players who offer high value for money.

The successful development of this model will improve the accuracy of player valuations for clubs and analysts, identify key drivers behind market value, such as performance metrics or player age, and provide actionable insights to optimize transfer decisions and investments in football talent.

Dataset Description:

Data for this project is sourced from Kaggle in .csv format for FIFA 20 player data, providing detailed information about football players, including their physical attributes, overall performance ratings, and specific skill metrics, which are essential for understanding their

market value. For this analysis, I have selected a subset of continuous variables that are most relevant for predicting a player's market value (value_eur). These variables include:

- Age: The player's age in years.
- Height (height_cm): Player's height in centimeters.
- Weight (weight_kg): Player's weight in kilograms.
- Overall: The overall rating of the player, representing their current skill level (0–100 scale).
- Potential: A rating that indicates the player's future potential in terms of performance (0–100 scale).
- Wage (wage_eur): The player's wage in Euros, reflecting their current market value to the club.
- Attacking Finishing: A skill metric that measures a player's ability to finish scoring opportunities.
- Skill Dribbling: Represents the player's dribbling ability.
- Movement Reactions: Measures a player's reaction time and situational awareness.
- Mentality Composure: Reflects a player's mental strength and calmness during key moments.

One challenge in constructing an accurate predictive model lies in the complexity and variability of factors that influence a player's market value. Market values are affected not only by quantifiable performance metrics but also by external factors such as club reputation, transfer trends, and player demand, which are not included in this dataset.

Nonetheless, by focusing on key features such as overall rating, potential, and wage, this project aims to identify the most impactful predictors of market value. Understanding how these features interact with each other and influence player valuations will be critical in building a robust regression model.

The dataset provides a strong foundation for analyzing the relationships between player performance metrics and their market value, offering valuable insights for clubs, scouts, and analysts looking to make data-driven decisions in the football transfer market.

Exploratory Data Analysis

Histograms of variables in filtered dataset

- age: The distribution is slightly right-skewed, with most players between 20 and 30 years old, which aligns with peak footballing years.
- height_cm: This is approximately normal, with most players having a height between 170 and 190 cm.
- weight_kg: The distribution is also normal, centered around 70–80 kg, reflecting typical player physiques.
- overall: A relatively normal distribution, with most players having an overall rating between 60 and 70, highlighting the dominance of average-rated players.
- potential: The distribution is slightly skewed toward higher values, indicating that many players have potential ratings between 70 and 80.
- **Value_eur (Target Variable)**: Highly right-skewed, with a majority of players having low market values and only a few with exceptionally high values. A log transformation can help spread out the data and make the distribution more interpretable for modeling.
- wage_eur: Similar to value_eur, this variable is highly right-skewed, with most players earning low wages and a few earning significantly higher amounts.
- attacking_finishing: The data is slightly skewed to the left, with a large concentration of players having lower finishing scores between 30 and 60.
- skill_dribbling: The distribution is bimodal, suggesting two clusters of players—those with lower dribbling skills and those with higher dribbling abilities around 60–70.
- movement_reactions: The distribution is approximately normal, peaking between 50 and 70, suggesting most players have average reaction times.
- mentality_composure: Similar to movement_reactions, this variable is normally distributed, centered around 50–70, indicating most players possess average composure.

Log-transformed market value histogram

The log transformation of value_eur has successfully addressed the extreme right-skewness of the original market value distribution, resulting in a more symmetrical and bell-shaped curve. The majority of player values now fall within a compressed range, with fewer extreme outliers

influencing the data. This transformation improves the interpretability of the data and ensures better performance for regression models by stabilizing variance and reducing the dominance of large values.

Heatmap and Scatter Plots

A heatmap was generated to see how the variables correlated with each other, particularly for value_eur. Variables with a high correlation coefficient with value_eur were then studied in more depth using scatterplots with respect to value_eur. These variables were wage, overall rating, potential rating, mentality composure, and movement reactions.

Market Value vs Overall Rating: There is a clear upward trend, where higher overall ratings are associated with higher market values. The relationship becomes stronger for ratings above 80, showing exponential growth.

Market Value vs Wage: A positive linear relationship is evident, with higher wages correlating strongly with higher market values. However, there is more spread among the data, particularly at higher wage levels.

Market Value vs Potential Rating: Similar to the overall rating, higher potential ratings are strongly associated with higher market values, particularly for values above 80, where market value increases steeply.

Market Value vs Mentality Composure: A positive trend exists, where players with higher composure scores tend to have higher market values. However, the spread is more pronounced for lower composure values.

Market Value vs Movement Reactions: A strong upward trend is visible, where higher reaction scores correlate with higher market values. Players with reactions above 70 see a sharp increase in market value.

Predictive Modeling

To predict player market values, I decided to use multiple different regression models and see which one performs the best in predicting these and accounting for the variation in my data and

the fluctuations in value. For each of these models, I decided to utilize an 80-20 train-test split, training my model on 80% of the data and then testing it on the remaining 20%.

Baseline Model

I evaluated the success of each of my regression models by comparing it to the Baseline Model's Root Mean Squared Error. My Baseline Model has a MSE of 31195891849332.703. Since this number is very large, I will be using the Root MSE. This came out to be 5585328.266926905.

Multiple Regression Model:

I chose to run a Multivariate Regression to predict the dependent variable value_eur (player market value) using multiple independent variables such as wage_eur, overall, and potential. This approach helps leverage the collective influence of the predictor variables to better estimate market values compared to the simplistic baseline model.

The Root Mean Squared Error (RMSE) for the training data is 2,582,326.74, while the testing data achieved a slightly better RMSE of 2,418,188.95. Both values are significantly lower than the RMSE of the Baseline Model, demonstrating that the Multivariate Regression outperformed the baseline by effectively capturing the relationships between features and the target variable.

Feature importance:

- wage_eur has the largest positive influence on market value, making it the most important predictor in this model.
- overall also has a strong positive effect, showing that higher player ratings contribute significantly to market value.
- Other variables like movement_reactions and attacking_finishing have moderate positive coefficients, indicating smaller but meaningful contributions.
- Features such as potential, skill_dribbling, and age exhibit negative coefficients, suggesting that they have an inverse relationship with market value under this model.

The y-intercept is -8,156,375.37, representing the baseline prediction when all predictor variables are zero, although this has limited real-world applicability.

Overall, the Multivariate Regression Model has performed better than the Baseline Model by incorporating the relationships between multiple predictors and the target variable, resulting in improved predictive accuracy. This demonstrates the model's ability to leverage key performance metrics and player attributes to predict market value effectively.

K-Nearest Neighbors (KNN) Regression

I chose to use K-Nearest Neighbors (KNN) Regression because, unlike linear regression, KNN does not assume a linear relationship between the predictor variables and the target variable. KNN makes predictions based on the similarity of instances in the feature space, which allows it to capture non-linear relationships and localized patterns in the data. For example, if player market values are influenced differently within clusters of players with similar overall, wage_eur, or potential scores, KNN can effectively identify and utilize these local variations to make predictions. This flexibility makes KNN a strong candidate when linear assumptions may not hold true.

The RMSE for my training data is 1342925.14 and my RMSE for my testing data is 1472397.31. My training data performed better than my testing data. For both the training data and testing data, however, the RMSE is much lower than the baseline model, identifying better performance. Additionally, for both test and train, the KNN performs better than the multiple variable regression model. I speculate this was the case because KNN can capture non-linear relationships between the features and the target variable, value_eur. Unlike linear regression, which assumes a straight-line relationship, KNN makes predictions based on the similarity of data points, leveraging local patterns and clusters in the feature space. This flexibility allows it to adapt better to the underlying structure of the data, especially when features like wage_eur, overall, and potential influence market value in a non-linear manner.

The feature importance table shows that wage_eur is the most critical predictor of player market value, followed by potential and overall, which aligns with their intuitive importance. These top predictors indicate that higher wages, skill ratings, and potential strongly determine a player's market value. In contrast, features like height_cm and weight_kg have much lower importance, suggesting that physical attributes have a minimal role. This ranking highlights that performance-based metrics are the primary drivers of market value in the KNN model.

Decision Tree Regression Model

I used a Decision Tree Regression model because it can efficiently handle non-linear relationships between the predictors and the target variable value_eur. Decision Trees are also capable of automatic feature selection, prioritizing the most relevant features and ignoring less important ones during the splitting process. The tree visualization provides an intuitive interpretation of the decision-making process by showing splits at each node based on feature thresholds, making the model highly interpretable.

The optimal depth for the tree was selected to avoid overfitting while maintaining generalizability. With a max depth of 3 (or lower complexity), the Decision Tree balances bias and variance effectively. The RMSE for the training set was 1,694.99, and for the test set, it was 870,940.28. The significant improvement in RMSE over the previous models (baseline, multivariable regression, and KNN) demonstrates that the Decision Tree model can better

capture the underlying relationships in the data. The tree's ability to segment the data into regions with locally accurate predictions explains this performance advantage.

The Decision Tree's feature importance table highlights overall (0.7798) as the most critical predictor, followed by potential (0.1472) and age (0.0233). This suggests that player ratings (both current and future potential) and age play a dominant role in determining market value, aligning with expectations for player valuation. Lower-ranked features such as weight_kg, height_cm, and mentality_composure contribute minimally, indicating their limited predictive power. Unlike KNN, which is sensitive to all features (including irrelevant ones), the Decision Tree naturally reduces noise by focusing only on the most impactful predictors during splits.

The model has outperformed previous models because it effectively identifies and prioritizes the strongest predictors (overall and potential) while ignoring irrelevant features. Unlike KNN, which struggles when noisy or irrelevant features influence distance calculations, the Decision Tree simplifies the decision process, reducing overfitting and improving generalization. Additionally, the ability to handle interactions between features, such as overall and potential, provides an added advantage. This highlights why the Decision Tree is currently the best-performing model for predicting player market value.

Random Forest Regression Model

To improve upon the Decision Tree model, I implemented a Random Forest Regression model, which aggregates predictions from multiple decision trees to reduce overfitting and improve robustness. Using GridSearchCV, the best hyperparameters were determined as max_depth = 20 and n_estimators = 100. These values strike a balance between model complexity and generalization by controlling the depth of each tree and the number of trees in the forest.

The Root Mean Squared Error (RMSE) for the Random Forest model on the training set is 273,770.22, while on the test set it is 538,946.89. While this model has strong performance, the notable gap between training and test RMSE suggests some degree of overfitting, likely due to the depth of the trees and the model’s ability to capture noise in the training data. Nonetheless, the Random Forest outperformed the baseline, multivariable regression, KNN, and Decision Tree models in terms of test RMSE.

The feature importance results emphasize overall (0.822) as the most significant predictor of market value, followed by potential (0.1157) and age (0.0225). This aligns with earlier observations that player ratings and potential heavily influence market value. Meanwhile, features like height_cm and weight_kg contribute minimally, which confirms their limited predictive power. Random Forest effectively identifies and prioritizes important features while ignoring irrelevant ones during splits, which reduces noise and improves predictive accuracy.

Despite the strong test performance, the Decision Tree model remains our best model due to its balance between bias and variance (training model only). Random Forest, while powerful, slightly overfits the data, as evidenced by the significant training-test RMSE gap. Nevertheless, its ability to model complex, non-linear relationships and interactions among features highlights its value as a robust modeling approach.

Summary of Findings and Final Ranking

In my analysis of player market values, all the models demonstrated improved performance over the baseline predictor, underscoring their effectiveness in capturing the relationships between features and the target variable. The models, ranked in terms of performance based on test RMSE, are as follows: Random Forest Regression, Decision Tree Regression, K-Nearest Neighbors (KNN) Regression, and Multiple Linear Regression.

Performance of the Random Forest Model:

The Random Forest Regression model emerged as the most effective, delivering the lowest test RMSE of 538,946.89. By aggregating predictions from multiple decision trees, the model reduced variance and improved accuracy. The ensemble nature of Random Forest allowed it to better capture complex, non-linear relationships in the data, outperforming all other models.

Strength of the Decision Tree Model:

The Decision Tree Regression model provided a strong balance between bias and variance, achieving a test RMSE of 870,940.28. While it did not outperform the Random Forest model, its interpretability and focus on key predictors made it a valuable approach. The Decision Tree was able to identify the hierarchical importance of features, which was consistent with findings across all models.

Feature Importance:

Across all models, overall (player rating) emerged as the most influential feature, with a notably high importance score of 0.822 in the Random Forest model. Other significant contributors include potential and age, reinforcing the idea that player ability and growth potential are key determinants of market value. Features like height_cm and weight_kg consistently exhibited minimal influence, indicating their limited predictive power.

Model Comparisons:

KNN Regression achieved moderate performance but struggled with feature scaling and sensitivity to irrelevant features, resulting in slightly higher RMSE compared to Decision Tree and Random Forest. Multiple Linear Regression was the weakest performer, primarily due to its inability to capture nonlinear relationships between the features and target variable.

The Random Forest model's superior performance highlights the advantage of ensemble methods in capturing intricate patterns within the data, reducing variance, and enhancing predictive accuracy. The Decision Tree model also stood out for its simplicity and interpretability, making it a close competitor. Key features like overall, potential, and age consistently proved to be the most significant predictors of player market value, aligning with expectations regarding player performance and valuation. These findings provide a robust understanding of feature influence and model performance, offering valuable insights for future predictive modeling and player valuation analyses.

Next Steps and Improvements

To further enhance the predictive capabilities of the models and gain deeper insights into player market values, I would explore additional features and modeling approaches. Incorporating data on player performance metrics such as recent match statistics, injury history, and consistency over time could significantly improve predictions. These dynamic performance indicators would provide real-time context to complement the static attributes already used in the current analysis.

Another potential improvement involves integrating team and league-level data. Features such as the club's overall reputation, league competitiveness, and recent team performance could influence player valuations, as players from prestigious clubs or competitive leagues often have higher market values. Additionally, incorporating contract length and transfer history could shed light on how contractual obligations and transfer market activity impact player worth.

Lastly, to ensure robustness and generalizability, I would implement cross-validation strategies on larger, more diverse datasets that include players across multiple seasons or leagues. This would mitigate any overfitting issues observed in the current models and offer a broader perspective on player valuations over time. By integrating these improvements, the models could achieve even more accurate predictions while offering deeper insights into the factors driving player market values.

https://github.com/SameerKamran1/Data-Bootcamp-Final-Project