Predicting Transfer Market Value of Football Players in European Top Leagues

Background

In the dynamic and competitive landscape of football, a highly dynamic and unpredictable sport, the ability to accurately predict market values hold immense significance for clubs, agents, and enthusiasts.

As in many other industries, football clubs also look for profiting, and it will depend on the specific club structure to really differentiate to what extent you focus on business over sporting objectives. Clubs such as Ajax, tend to put extra focus on the development of players in their academy levels to later develop high-profile players that can be sold for high margins. Other clubs such as Borussia Dortmund or RB Leipzig, do great in scouting to attract young low-profile players to develop them and sell them for a higher margin. Other clubs such as Real Madrid are known for rather buying the “end-product”.

Regardless of the overall business and sporting strategy of the club, it is clear that player market values play a crucial role in the strategic plans of many professional football institutions, with clubs, players, agents, brands and even fans posing as main stakeholders.

Player market value serves as a crucial metric guiding such institutions decisions. Since teams seek to maintain a competitive edge and maximize their resources, the need for advanced predictive models that are able to account for player performance metrics and its relationship with player’s market valuation becomes crucial.

In this project, the objective is to leverage machine learning models to forecast player market values.

Data Collection

Data Sources

The selection of reliable and consistent data sources is fundamental for the purposes of this project, as the quality of the data will have a direct implication on the final performance of our models. Therefore, two main data sources were identified:

1. Transfermarkt.com

2. FBREF.com

Transfermarkt.com serves as a repository of player information from various regions worldwide. It encompasses a great variety of player details, including player valuations, personal profiles, general performance statistics, clubs and competitions, agency profiles, length of contract, etc.

FBREF.com, on the other hand, focuses in a wide array of performance statistics quantifying aspects of the game, for both teams and players. They sourced their data from Opta, a very well-known and reputable football data provider.

Initially, we intended to develop web scrapping pipelines for the automation of the extraction of the data. However, for time purposes, we were forced to seek datasets containing the necessary information. We found Transfermarkt data in a Kaggle repository and extracted the FBREF data from a football analytics guru, Ben Griffis, from his GitHub account. You can find the links in the project's repository.

This data was later consolidated into one single dataset ready for subsequent analyses.

Feature	**Type**	**Description**	**Potential Relevance**
Player Name	Categorical	Metadata	Metadata
Age	Numerical (Discrete)	Age of the player in years (int)	Younger players might have higher market potential
Position	Categorical	Playing position of the player	Different positions might have varying market values
Goals per 90s	Numerical (Continuous)	Average Goal per game of the player	Ratio of goals might have strong impact on market value
Market Value	Numerical (Continuous)	Market value of the player in million pounds.	Target variable for prediction

Table 1: Feature Definition & Potential Relevance

Special Considerations

Data Selection Criteria

The dataset utilized for this project is limited to players with presence in attacking areas of the game, including midfielders and attackers. This decision was taken due to the higher availability of performance metrics for these categories of players. Other player positions, such as defenders and goalkeeper have been ruled out from the analysis. However, this is merely based on the data available, since the author of this project (me haha) played for several years as a left-back in his academy years. There is no bad sentiment towards defenders! They are as important as other players.

External Factors Affecting Market Value

It is also essential to denote that a player's market value is influenced by far various factors beyond performance metrics alone. While player performance does play a significant role, other aspects such as contract duration, marketing and publicity appeal, and overall physical appearance can also have an effect in a player's market value. However, these aspects-beyond-the game are not quantified in our dataset, as we solely focused on the player performance.

Challenges in Data Integration

Some challenges raised when consolidating the data from our data sources. One these challenges was regarding to the mapping and merging of such datasets. Inconsistencies in player names across the sources posed a significant obstacle, especially with players with non-English characters in their names, such as in the case of Martin Odegaard or Mezut Ozil (Ødegaard & Özil, respectively). Despite the availability of resources for more efficient mapping and merging process, we proceed to rely only on those players with consistent names that made it through the mapping & merging on a first instance. This decision unfortunately reduced the sample size considerably, but still to a decent number of records for the analysis.

Proposed Methodology

Exploratory Data Analysis

General Exploration

The final dataset comprises 1911 instances and 312 features, with no missing values. The data is related to player performance for the last 5 seasons (2017-2022). Among the features, there are 8 categorical (binary) features, 1 metadata (Name of Player), and 302 numerical features related to player performance statistics.

As seen in Table 1, the target feature 'market_value_in_eur' is represented as a numeric variable, indicating the player's market value in millions of British pounds. The descriptive statistics for this feature are as follows:

- Mean: 6.27 million GBP
- Mode: 1.0 million GBP
- Median: 2.0 million GBP
- Min Value: 0.050 million GBP
- Max Value: 180.000 GBP

Given the overall descriptive statistics, there are some signs of clear skewness on the data within the distribution, which will be further addressed later in the report.

Data Analysis

After implementing a general data exploration, in which our main intention is to familiarize ourselves with the surface statistics of the data, it is time to start getting some insights and interpreting some of the trends we can derive from it.

Since we are trying to predict market value of a player, it is relevant to start with some of the features that may have some importance to the target feature. A good starting point would be to see if there is any difference in market values by league. Given my domain knowledge, even though La Liga had arguably the strongest two teams for a long time (FC Barcelona & Real Madrid), the English Premier League (EPL) had elite players spread across different teams, given its bigger audience, reach and the money being invested by Arabic enterprises and millionaires that allow many clubs to invest heavily in their squad. Other factors such as the arrival of elite coaches such as Pep Guardiola or Jurgen Klopp, and their strong and offensive sporting projects also contributed to better performance, spectacle, and therefore an increase in revenue.

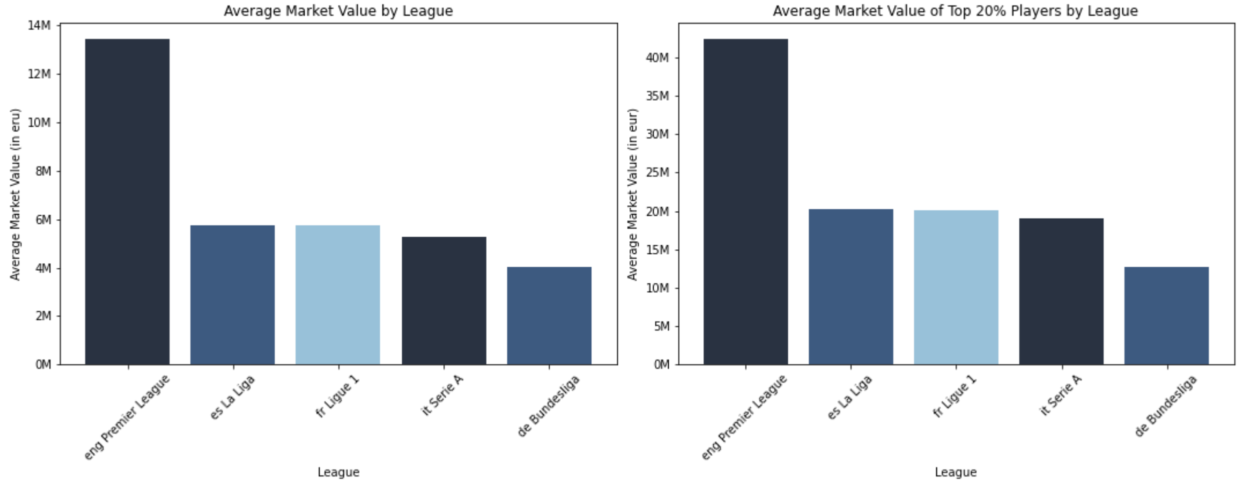


Figure 1: Average Market Value per League

As depicted in Figure 1 and correctly inferred, the average market value of EPL players is considerably higher than those of their counterparts. Additionally, the more we aggregate our players, for instance by aggregating based on the TOP 20% for each league, the smaller the gap appears to be, but not considerably, which aligns with my initial empirical hypothesis.

Regardless of the reasons why this happens, I am interested in knowing what are the performance characteristics of those players with higher market values within each league.

As stated initially, the data was gathered to contain players in midfield and attacking positions. Therefore, I'm interested in analyzing what areas of the pitch are those high value players occupying.

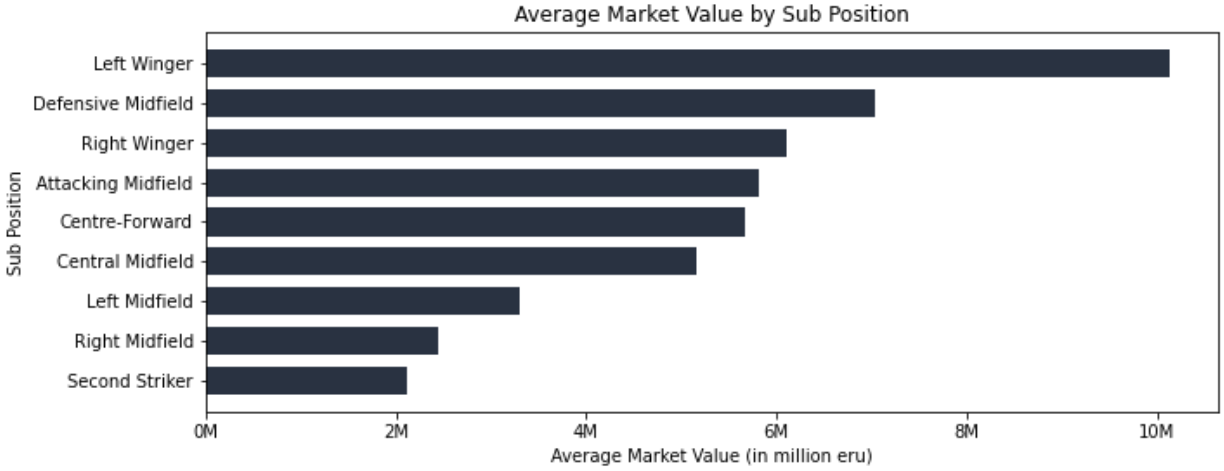


Figure 2: Average Market Value by Position

By looking into Figure 2, we clearly visualize that Left-wingers have the highest average market value, by a considerable margin of about $4M with respect to the other positions. Many great players and known 'ballers' have played in this position in the latest years. Some of the most prominent ones from the 2000s decade include Ronaldinho, Robinho, Ribery, Henry, etc. These players tend to be a total spectacle on the pitch given its production ability, but most importantly their progression skills such as dribbling and pace.

Surprisingly, defensive midfielders also rank high. Defensive midfielders are probably the most misunderstood players out there, but they are super important for the overall compactness of a team. They don't only have to offer the linkage between defense and attack in possession and progression, but also have to be the sweepers and the stoppers defensively speaking, in a position in which stakes are high and there is usually little time and space to maneuver.

Right wingers, attacking midfielders and center forwards also follow along very closely. Since we are already diving deep into the exploration, let's visualize the players with the highest value for a random year in our dataset.

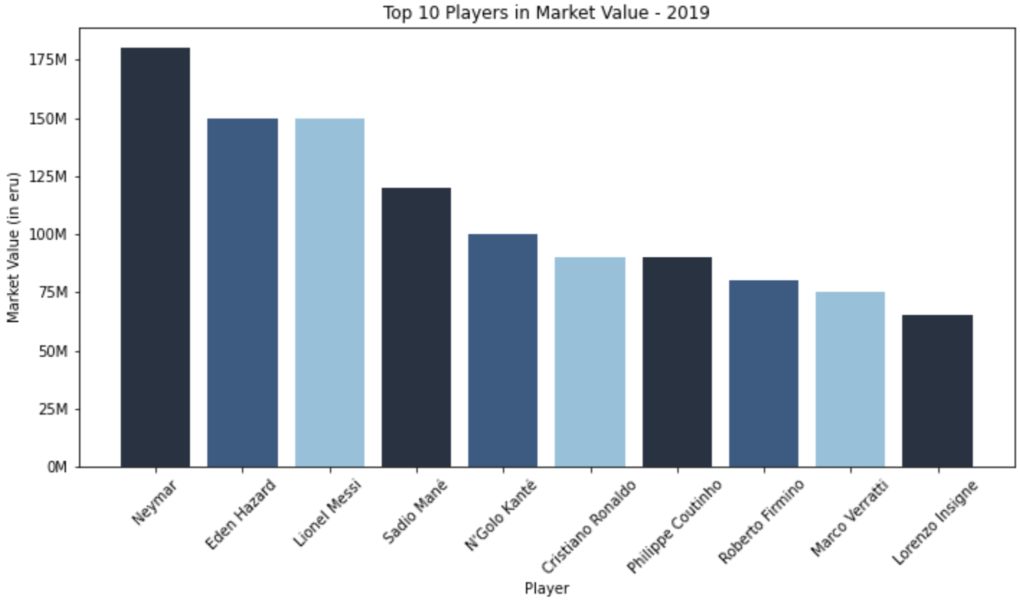


Figure 3: Top 10 Players in Market Value

In the year 2019, a 27-year-old Neymar Jr was the player with highest valuation in the market with a worth of $175M, followed by Eden Hazard and Lionel Messi. Two years earlier (2017), PSG had broken the market with the acquisition of then FC Barcelona’s Neymar, for a total price of $263M, and he became the face of French football.

Eden Hazard, a year before (2018), had just performed extremely well in the World Cup by leading his home country Belgium into semifinals by defeating precisely Neymar’s Brazil in the quarterfinals.

Regarding Messi: he is the greatest of all time. Even at his advanced age he was amongst the top given his contribution to his team, which in that period was FC Barcelona, where he won the league but failed to go through the semifinals against Liverpool in Anfield.

Additionally, notice how 5 out of the 10 players are left-wingers.

But what are these characteristics that contribute the most to a player’s market value? Let’s look at overall correlation metrics.

Feature	**Correlation**
ThruBalls	0.572945
ProgCarries	0.567267
GCA	0.567214
GCAPassLive	0.555344

PenAreaCmp	0.547783
CarriesToFinal3rd	0.535508
Att3rdTouch	0.535358
npxG+xA	0.526606
SCAPassLive	0.523227
PrgCarryDist	0.518252

Table 2: Feature Correlation Scores

Given the information seen in Table 2, we can reflect on some of the TOP predictors. The TOP predictor according to the correlation coefficient is *ThruBalls*, representing the number of times a player performs a through ball, which is a pass into open space between two defenders for an attacker to receive the ball, thus a line-breaking pass.

This aspect is particularly crucial as line-breaking passes are very risky but are the most efficient way to break defensive lines (and consequently opposition players). Since the ball always moves faster than a player, it is ideal. However, breaking defensive lines or just advancing in position by carrying the ball, thus Progressive Carries (as our TOP 2 feature), is also very important, and it makes total sense for both to be up there. Goal Creating Actions also rank high, and it is interesting how all of these variables are related to the final progression and closeness to the objective of any team: scoring goals.

Another relevant feature is npxG+xA, which stands for non-penalty expected goals + expected assists. This metric has been recently developed by the application of machine learning algorithms that focus on the quality of the opportunities created rather than the quantity of them.

Expected Goals (xG) differs from Goals Scored or Shots performed because of interpretability. While Goals and Shots score rather reflect the actual number of actions performed, xG values reflect the probability of a shot resulting in a goal based on historical data and the characteristics of the shot itself (such as positioning, distance to the goal, etc.). This metrics are important since they offer football spectators and analyses new ways to watch and assess performance in football.

Midfielders		**Attackers**	
Feature	**Correlation**	**Feature**	**Correlation**
ShortPassCmp	0.547498	**GCA**	0.654209
ReceivedPass	0.546694	**ThruBalls**	0.640853
Ground	0.541801	**PenAreaCmp**	0.629516
PassTarget	0.539434	**GCAPassLive**	0.622725
ShortPassAtt	0.538295	**ProgCarries**	0.618764
Carries	0.537001	**npxG+xA**	0.608576

PassesCompleted	0.527678	**CarriesToFinal3rd**	0.600482
LivePass	0.518891	**SCA**	0.594239
MedPassCmp	0.514711	**Att3rdTouch**	0.590201
LiveTouch	0.510093	**xA**	0.586586

Table 3: Feature Correlation Scores by Position

If you recall from previous paragraphs, the data contain players for both midfield and attacking positions. Hence, some bias in the top predictors may exist if they aggregate all of them together.

Since, they generally perform different roles within the pitch, for instance midfielders are more interested in moving the ball into the final 1/3, while attackers are more interested in performing actions in that final 1/3, it does make sense that the TOP predictors vary depending on the category. Table 3 is a representation of that, and from a feature selection perspective, we need to make sure to account the position as a feature for the models to be able to catch it.

Additionally, in a same fashion, game dynamics are different depending on the league. For example, La Liga (Spain) is known for it is wide technical, slower-paced and positional football, whereas the Bundesliga is known for its very direct approach and counter-attacking football. These aspects may be of importance for determining market value overall.

Pre-processing

With that being said, the first thing needed to do would be to precisely encode the aforementioned categorical variables. For these purposes, we will utilize dummy encoding. After this has been applied, we proceed to check the normality of our data.

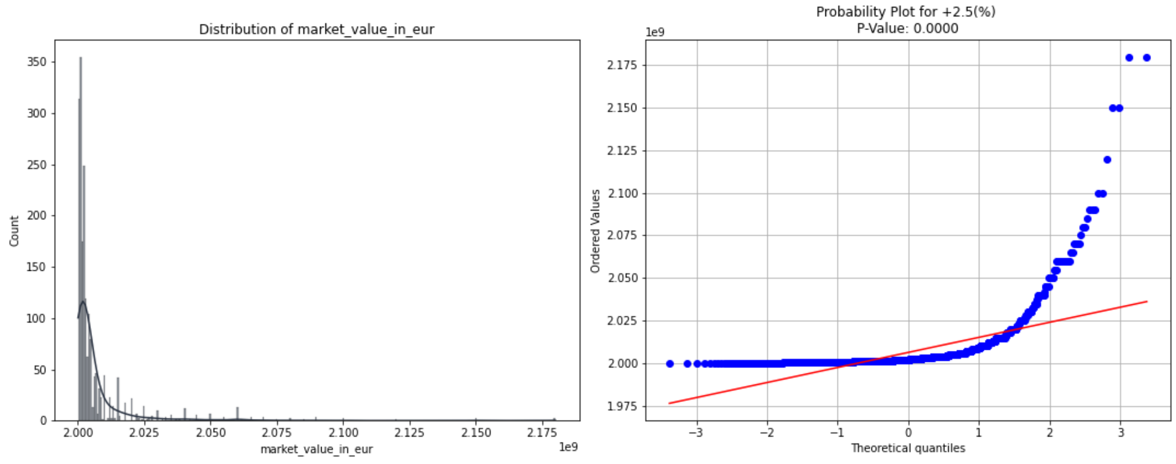


Figure 4: Normality Assessment for Target Feature

Unlike non-parametric models that do not require normality assumptions (Sahngun, 2016), linear models tend to assume that the dependent feature and the linear model's residuals follow a normal distribution (Schmidt & Finan, 2018). Figure 4 plots the target feature's distribution, and we can clearly see that it is highly right skew, with a p-value of less than 0.001, thus it appears not to follow the normal distribution.

The correct procedure to follow would be to first build a model without transformation, analyze the residuals of the model, plot them, and then decide whether a transformation is needed. However, this may come along with some other complications. Perhaps the bigger complication would be that, if transformation is needed then our target feature would become less explainable, since it would change the target estimate and hence bias point estimates (Schmidt & Finan, 2018). Additionally, it is well-known that even in scenarios in which the residuals are deviating from normality, the model still produces valid results (Schmidt & Finan, 2018). Therefore, it was decided not to apply any kind of transformation.

For high dimensional data, and given the different magnitudes of other different features, standardization by variable ranges is needed to control the dependency of such magnitudes (Tanioka & Yadohisa, 2012). This statement makes total sense, as the magnitudes of some variables will be naturally higher than other's as important and this difference could lead to several bias.

For example, we will take two of the top predictors for Attackers: GCA (Goal Creating Actions) & xA (Expected Assists). While GCA quantify the number of actions preceded to a shot or approximation, xA assess the quality of a shot by quantifying the probability of a shot resulting in a goal after such pass was made. Therefore, xA values tend to be lower in comparison to GCA, and this magnitude could lead to aforementioned bias.

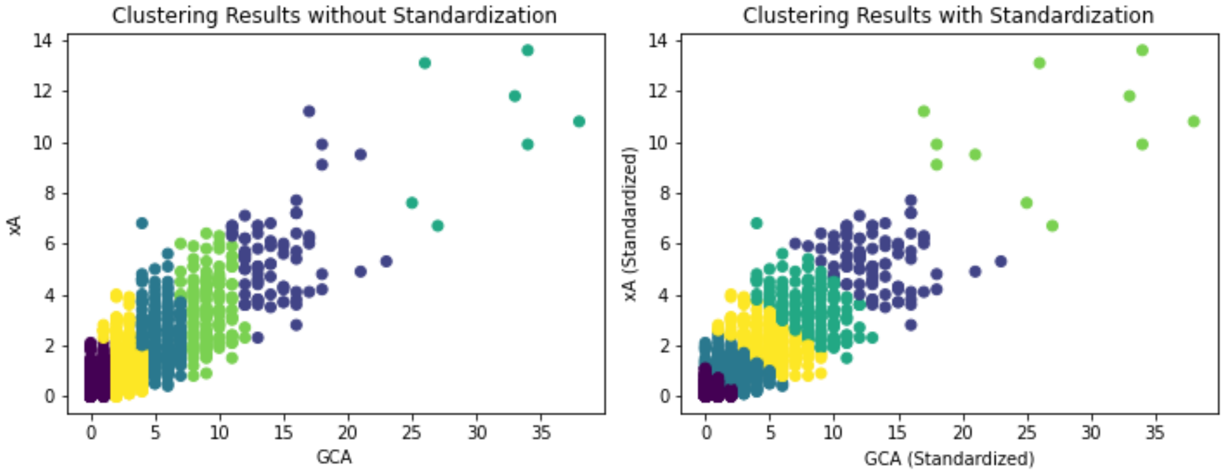
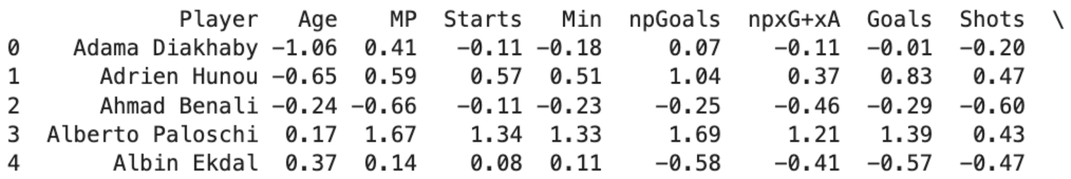


Figure 5: Clustering based on GCA vs xA (Non-Standardization vs Standardization)

Figure 5 is a clear representation of how magnitudes affect the model's performance, in this case k-means clustering. Even though k-means is an unsupervised model and doesn't necessarily work directly on classification, it is a great visual representation on how models can be biased because of difference in magnitudes. Hence, feature standardization was performed to ensure no magnitude bias is present.

Additionally, standardization still preserves feature interpretability, a very important connotation for our purposes.



```
             Player   Age    MP  Starts   Min  npGoals  npxG+xA  Goals  Shots  \
0    Adama Diakhaby -1.06  0.41   -0.11 -0.18     0.07    -0.11  -0.01  -0.20
1      Adrien Hunou -0.65  0.59    0.57  0.51     1.04     0.37   0.83   0.47
2      Ahmad Benali -0.24 -0.66   -0.11 -0.23    -0.25    -0.46  -0.29  -0.60
3  Alberto Paloschi  0.17  1.67    1.34  1.33     1.69     1.21   1.39   0.43
4       Albin Ekdal  0.37  0.14    0.08  0.11    -0.58    -0.41  -0.57  -0.47
```

Figure 6: Standardization Subset Output

Machine Learning Algorithms

The data was later split into training and test sets. The split was performed by random assignment using scikit-learn library. However, in practice it would have been nice to predict last season's (2021-2022) player market values, since objectively speaking the whole point is to predict the 'next' season player value. However, since we are trying to analyze results first, the randomly split data align with our purposes.

To find the best regression model, different algorithms were tested in order to verify which one provides the better error performance. The algorithms used along with the software of its development are listed as follows:

- Simple Linear Regression (Python, Scikit-learn)
- Ridge Regression (Python, Scikit-learn)
- Lasso Regression (Python, Scikit-learn)
- Random Forest (Python, Scikit-learn)
- K Nearest Neighbors (Python, Scikit-learn)
- Neural Networks (Orange Software)

Prediction Results – Phase 1

When separating the data into two different sets for training and testing, a major issue occurred in which the model was outputting error messages related to the values of some of the features.

More specifically, some values were being interpreted as either NaN or infinite. After conducting investigations, we realized that the problem relied on the binary variables only. Some troubleshooting was performed; however, given time constraints we decided to just remove the binary features and continue with the modeling on Phase 1, and later build models in a different software with the entire dataset on Phase 2.

The Mean Absolute Deviation (MAE) was chosen for the assessment given its interpretability, as in the context of football analytics, stakeholders tend to be interested in seeing the actual number of errors in the known units, rather than having to interpret other less explainable ones such as the MSE (Mean Squared Error).

We first started by building a Simple Linear Regression model to assess the accuracy of the predictions, in which we got a high number on the MAE, indicating a poor performance.

Some regularization were applied as it can be seen in the Ridge Regression and Lasso Regression models, with penalizations of $\alpha = 200$ and $\alpha = 1$ respectively, and some improvements were seen by introducing an amount of bias to reduce the variance.

Algorithm	**MAE**
Linear Regression	7.38
Ridge Regression	5.93
Lasso Regression	5.49
Random Forest	**4.19**
K Nearest Neighbors	4.46

Table 4: Phase 1 Test Scores

K Nearest Neighbors model was the next on the list, in which different values for hyperparameter k were tested to select the one that was capturing the underlying relationships of the data the best in relation to their error performance. A k value of 20 was chosen for development, providing a better performance than the previous Regression models.

Random Forest Regressor was later developed and by using cross-validation the optimal number of estimators was selected to be 50, which produced the lowest-value MAE and consequently the best model among the 5 chosen.

Additionally, it is important to highlight that all models developed were built by carefully considering measurements for avoiding under or overfitting to the data. More specifically, we were very interested in knowing the difference in error metrics between training and testing, and also by incorporating cross-validation to assess the difference in error performance.

Prediction Results – Phase 2

As mentioned before, the same models with the addition of Artificial Neural Networks were built using Orange Software, but this time by incorporating the binary variables for both Position and League of each player.

The model's parameters and hyperparameters remain the same as those in Phase 1, not necessarily because they might give the best results, but to serve as a benchmark to compare the impact of the binary variables on the overall performance of the improved model in relation to the previous ones.

The Neural Networks were left with 2 hidden layers of 30 neurons each. The performance of these models are shown in Table 5.

Algorithm	**MAE**	**MSE**	**RMSE**
Linear Regression	14.13	26,114.38	161.60
Ridge Regression	5.63	101.73	10.09
Lasso Regression	5.25	96.50	9.82
Random Forest	**3.97**	**77.51**	**8.80**
K Nearest Neighbors	4.15	103.134	10.16
Neural Networks	4.15	145.801	12.08

Table 5: Phase 2 - Test Scores

As suspected, the improvement in error performance of most of the models suggest that they are performing better when being exposed to the entirety of the data, hence highlighting the importance of accounting the necessary features that may have an impact on the target variable as correctly inferred in the Data Analysis section.

Discussion

The Random Forest model continued to be the best model in Phase 2, yielding an overall MAE value of 3.97 million GBP. While this performance demonstrates the model's relative strength within the context of our analysis, it may not be ideal for practical expectations.

If we recall from the Data Analysis section, the mean of the target variable was 6.27 million GBP, and the median was 2.0 million GBP. With a MAE of 3.97 million GBP, this implies that the model is unable to correctly predict market value by a margin higher than the median of the data, and we can make some analysis on why this is happening:

Presence of Outliers

Given that the dataset exhibits a wide range of values in the target feature (minimum value of 0.050 million GBP and maximum of 180 million GBP), the data is prone to the influence of outliers, potentially distorting model predictions.

Non-Normality of Target Feature

As previously noted, the non-normal distribution of the target feature definitely complicates prediction accuracy. Transformation methods such as Box-Cox were considered to target this issue, however its impact on model interpretability led to our decision to avoid it for explainability purposes.

Selective Filtering

Limiting the analysis on players that meet certain requirements, such as having a minimum market value of 10 million GBP could enhance model performance by focusing on a more representative subset of high-performing players.

Feature Selection

Technique such as Boruta Algorithm or Stepwise Feature Selection were considered but discarded due to time constraints and computational efficiency. They could be very useful in reducing dimensionality of the data and keeping only relevant variables, thus enhancing the model's performance.

Influence of Non-Performance Factors

As stated before, the low performance could be related to non-performance values such as marketing, publicity, contract length, and social media presence of the players. Features that are not quantifiable in our dataset.

Conclusion and future work

Football performance on the pitch, although it is very important, is not enough for determining the player's market value. There is a higher indication that other factors such as marketing and contract length of the player's contract have a high relevance on the target feature. By focusing only on performance metrics, the model does not produce high levels of prediction accuracy. However, by separating players by league and positions, we have slightly better results, which are still not good enough given my domain knowledge.

This project, though, serves as a basis for future work, in which other factors such as the ones mentioned should be incorporated and re-evaluated alongside other measures such as feature selection or data normalization for better prediction accuracy.

References

Schmidt, A. F., & Finan, C. (2018). Linear regression and the normality assumption. Journal of Clinical Epidemiology, 98, 146–151. https://doi.org/10.1016/j.jclinepi.2017.12.006

Nahm FS. Nonparametric statistical tests for the continuous data: the basic concept and the practical use. Korean J Anesthesiol. 2016 Feb;69(1):8-14. doi: 10.4097/kjae.2016.69.1.8. Epub 2016 Jan 28. PMID: 26885295; PMCID: PMC4754273.

Tanioka, K., Yadohisa, H. (2012). Effect of Data Standardization on the Result of k-Means Clustering. In: Gaul, W., Geyer-Schulz, A., Schmidt-Thieme, L., Kunze, J. (eds) Challenges at the Interface of Data Analysis, Computer Science, and Optimization. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-24466-7_7