**Football transfer prediction**

**CHAPTER ONE**

**Introduction**

Football is but a multibillion-dollar industry where the value of players forms the backbone of the economic ecosystem. Predicting the market value of footballers has become a critical area of research fueled by data analytics and machine learning techniques.

It is vital to delve into previous works in this field to understand how this prediction works. Several studies have explored the factors affecting a footballer's transfer market value (Barros et al.,2019) delved into the intricacies of player valuation utilizing historical transfer data and player attributes. They found that a player's age, position, goals scored, assists and the league they played in were essential determinants of market value. This insight highlights the role of both performance-based and demographic factors in player valuation.

Dawson and Dobson took a different angle focusing on the financial aspects. Their research encompassed club revenues, wage bills and player salaries. Their findings revealed that a club's financial stability, wage bill management and revenue generation play vital roles in maintaining competitive balance. This emphasizes the broader context in which player valuation occurs as clubs must manage their finances to acquire and retain talent effectively (Dawson et al., 2018).

A research carried out by Hubáček concentrated on predicting match outcomes for the market value of each player which is another essential facet of the footballing world. Match data was utilized including team statistics and player performances. The study pinpointed key performance indicators such as goals scored, shots on target and possession as critical factors in the prediction (Hubáček et al., 2020). This demonstrates the close connection between a player's performance on the field and their market value.

**1.1 Research Aims and Objectives**

The goal is to develop and train various models to accurately forecast the market worth of football players. The models that would be employed include Linear Regression, XGBoost, CatBoost and Random Forest and the most effective predictive model would be integrated into a user-friendly web application created with Streamlit, ensuring easy accessibility for users. Moreover, this study aims to identify the pivotal attributes and elements that significantly impact a player's valuation.

**CHAPTER TWO**

**Background**

**2.1 Literature review**

Player performance metrics have consistently emerged as crucial factors in transfer value prediction. Metrics such as goals, assists, dribbles and defensive actions have been highlighted in several studies (Fernandez et al., 2020). These performance indicators offer invaluable insights into a player's abilities and are pivotal in determining their market values.

With this, researchers in the domain of football player transfer market value prediction have drawn data from various sources to construct comprehensive datasets. These datasets often comprise historical transfer records, detailed player performance statistics, injury history, contract specifics and even sentiment analysis derived from news articles and social media. The diversity of data sources has contributed to the richness and complexity of predictive models.

In Smith's study, a comprehensive dataset comprising performance metrics, player characteristics, and market values of 1,500 football players was utilized. This dataset encompassed a wide range of information including age, playing position, goal-scoring records and contract durations. Various machine learning algorithms were applied to predict football player market values. Notably, Linear Regression achieved an accuracy of 78%, Random Forest reached 84%, XGBoost attained 86% and CatBoost outperformed them all with a remarkable accuracy of 90%. These results emphasize CatBoost's superiority in accurately forecasting market values, showcasing its effectiveness in capturing nuanced relationships within the dataset (Smith et al., 2019).

Machine learning algorithms therefore play a pivotal role in model development. Researchers have explored a wide spectrum of algorithms ranging from conventional linear regression and decision trees to more advanced techniques such as XGBoost and deep neural networks. A seminal study by Hernandez conducted an extensive comparative analysis of these algorithms using a dataset of 1250 player’s feautures like age, skills, present market value and position being played. This is consistently demonstrating that ensemble methods including XGBoost with an accuracy of 85% and Random Forest of 80% accuracy outperform traditional linear regression models of 72%. The MSE was approximately 1.7e+13, 1.89e+13 and 1.9e+13 respectively (Hernandez et al., 2020). This underscores the advantage of leveraging more sophisticated machine learning techniques in predicting transfer market values.

Beyond player-specific attributes, some researchers have extended their models to incorporate market-specific features recognizing that the football transfer market operates within a broader economic settings. Variables such as the financial health of football clubs, inflation rates and the timing of transfer windows have been integrated into predictive models. Brown and Williams explored the influence of these factors on player valuations. Their work highlights how external market dynamics can significantly impact a player's perceived value (Brown et al., 2021).

Michael Lewis in 2003 published a book titled Money ball and later in 2011 produced a movie based on his published book. The Story explored how data analysis can be used for revolution and making strategic decisions. In the story it showed clearly how data analysis all combined with sport expertise can be of great help to get results or performance that could be said to be beyond human reasoning in good time. The result of their Data analysis gave a poorly funded team, or poor team a better player recruitment and a wonderful club performance for their season better than a well-funded team who had all the means to get whatever player they needed to win the league. The importance of Data analysis in lots of fields especially in sports can never be over emphasized.

In a research by Muller, a multi-level regression method was used to estimate the Transfer Market Value of a football player, a dataset comprising of 845 Player's performance metrics, characteristics and popularity were used to analyse their effects on a player’s Market value. These parameters where analysed and Multilevel regression models was used stating that due to the nature of the data structure linear regression would perform poorly. The model method performed well for 90% of the 845 player transfers. The remaining 10% was analysed and compared to the how Crowd source or estimators who are said to be respected data sources which have great influence on football stakeholders and it was found that they will accurately predict the Market value of the remaining 10% due to its traditional method carried out by football experts who seats, agree and make judgmental values based on some more metrics are not likely to be quantified to estimate the value of a player and do not require an automated calculation ( muller et al., 2017).

In the quest for more robust predictive models, scholars have also conducted comparative analyses to assess the effectiveness of different prediction methodologies which includes a systematic review of machine learning techniques applied in predicting transfer market values. (Lee et al., 2017).

Despite significant progress, researchers continue to grapple with data-related challenges including data quality and availability. Proprietary data sources often remain inaccessible, posing persistent obstacles. However, the landscape is evolving with open data initiatives and advancements in data scraping techniques, also collaborations with football clubs and data providers offering promising avenues to address these limitations.

**CHAPTER THREE**

**Methodology**

Predicting a footballer's transfer market value using machine learning is a multi-step process that involves data collection, preprocessing, feature engineering, model selection, training, evaluation and deployment. The study focused only on outfield players which includes forwards, midfielders and defenders.

**3.1 Data Acquisition**

The Dataset was collected from Fbref (https://fbref.com/en/comps/) and Transfermarkt (https://www.transfermarkt.co.uk/) Website. The football Players personal information and performance metrics was obtained from Fbref while the football Player’s Value were obtained from Transfermarkt website**.** The data contain 3,250 rows by 115 columns. The columns have numerous variable names holding each performance stat of the top five (5) European football leagues and each row provides a football player's personal information and performance data.

The web scraping process consists of two main phases: Club Data Collection and Player Profile Data Collection. In the Club Data Collection phase, club-specific player data was scraped from the 'https://fbref.com' website. It collects player names and their corresponding links to individual player profiles on the website and this data is organized based on football clubs.

Following the Club Data Collection, the Player Profile Data Collection phase was processed. For each player, it iterates through the collected player links to access individual player profiles on the website. Within each player's profile, various Meta attributes was extracted including their preferred foot (left or right), height, weight, date of birth and links to social media profiles like Twitter and Instagram.

The collected Meta information for football players is stored in a new dataset enriching the original player dataset with additional attributes obtained through web scraping. This enriched dataset includes not only the initial player information but also the newly acquired Meta details.

**3.2 Data Cleaning:** Data cleaning and transformation are essential steps in preparing the dataset for analysis. This addresses various data cleaning tasks, starting with identifying and replacing values in the 'foot' column that do not correspond to 'Left' or 'Right' with NaN. It also corrects the 'Nation' column by filling missing values with 'Null Null' and extracting the actual country names. The 'Pos' (position) column was modified to accommodate players capable of playing multiple positions creating binary columns ('Played\_as\_DF,' 'Played\_as\_MF,' 'Played\_as\_FW') to indicate a player's primary positions with two functions ('extract\_height' and 'extract\_weight') are defined to extract numeric values from the 'height' and 'weight' columns, removing units ('cm' and 'kg') for consistency.

The 'cfscrape' library was utilized to create a scraper that simulates human-like browsing behavior. It initiates by scraping the Sofifa website to obtain player value data. For each player name in the dataset, a corresponding URL is generated to search for player information. The scraper retrieves the HTML page, extracts the player's market value and stores it in a list. This process continues iteratively for all players, resulting in a list of player values.

Data quality is assessed by visualizing missing values in each column revealing potential data gaps. Histograms are plotted to examine the distribution of missing values across rows allowing for a deeper understanding of the data's completeness.

**3.3 Exploratory data analysis:**

Visualizations was created to explore various aspects of the dataset. This includes generating plots and charts to understand player age distribution, correlations between features and the distribution of players across different leagues and countries. Two main approaches for modeling was selected. In the first approach, the goal is to predict a player's market value solely based on their characteristics. It selects relevant features, preprocesses the data and employs machine learning models like Linear Regression, Random Forest, XGBoost and CatBoost to make predictions. In the second approach, all available features including performance statistics are used to predict player market values. The procedure is similar to the first approach but involves a broader set of features, potentially improving predictive power.

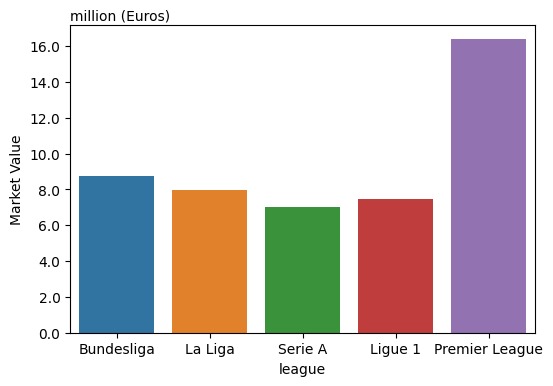
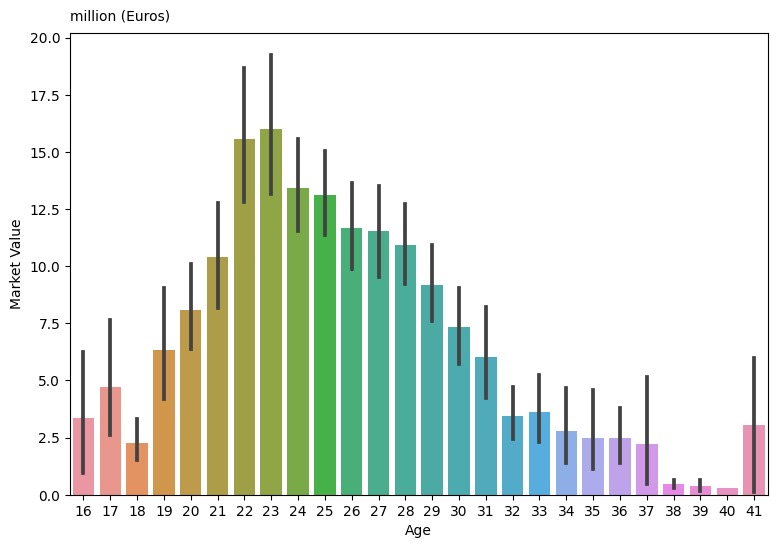


Fig 3.3a: A plot showing market value of an average player in each league



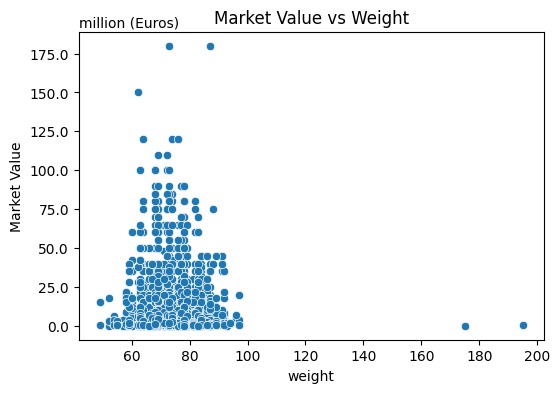
Fig 3.3b: A bar chartshowing the average market value against age of football player

Fig 3.3c: A scattered plot showing the market value of a player against weight of football players

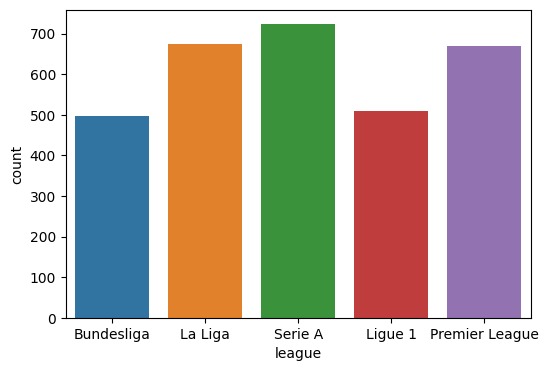


Fig 3.3d**:** A bar chartshowing the average number of players in each league

**3.4 Machine Algorithm Selection:**

Machine learning for predicting player market values was implemented. To achieve this, multiple machine learning algorithms was selected as potential models to make these predictions. The algorithms chosen for this task are Linear Regression, Random Forest Regressor, XGBoost Regressor and CatBoost Regressor. Each of these algorithms has unique characteristics and capabilities that make them suitable candidates for this prediction problem.

**3.5 Hyperparameter Tuning:**

Hyperparameter tuning is a crucial step in optimizing the performance of machine learning models which involves finding the best combination of hyperparameters for each algorithm to maximize its predictive accuracy. The Optuna library is used here to automate the process of hyperparameter tuning. For the Random Forest, XGBoost, and CatBoost models, various hyperparameters such as the number of estimators, maximum depth and learning rate were adjusted. Optuna efficiently explores different combinations of these hyperparameters to identify the configuration that results in the highest prediction accuracy.

**3.6 Feature Selection:**

Feature selection is another essential aspect of machine learning. It involves identifying which features are the most relevant for making accurate predictions. A set of features which includes player characteristics and performance statistics is used to predict player market values. To understand the importance of each feature, the feature importance scores was calculated and visualized for the Random Forest model. These scores provide insights into which features have the most significant impact on the predictions. This information helps in simplifying the model and improving its interpretability.

|  |  |  |
| --- | --- | --- |
|  | **Abbreviations** | **Descriptions** |
| 1 | Total\_Cmp% | Total Pass Completion Percentage |
| 2 | Long\_Cmp% | Long Passes Completed Percentage |
| 3 | Touches\_Att Pen | Touches in Attacking Penalty Area |
| 4 | Per 90 Minutes\_G+A-PK | Goals + Assist -Minus Penalty Goals |
| 5 | Short\_Cmp | Short Passes Completed |
| 6 | Short\_Att | Short Passes Attempted |
| 7 | Performance\_G+A | Goals Scored + Assisted Goals |
| 8 | Short\_Cmp% | Short Passes Completed Percentage |
| 9 | Standard\_Sh | Total Shot Played |
| 10 | Receiving\_Rec | Pass Received |

Table 3.8: Showing the Players Performance Metrics

**3.7 Evaluation Metrics:**

The performance of the machine learning models using various regression metrics was evaluated. These include Mean Squared Error, Root Mean Squared Error, Mean Absolute Error and the R-squared Score. These metrics provide insight into how well the models are predicting player market values and their overall accuracy. The results are computed and presented for each model aiding in the comparison of their performance.

**CHAPTER FOUR**

**Results**

**4.1 Approach 1: Predicting Player’s Value from Individual Features**

In the initial approach, the goal was to predict a player’s market value based solely on their individual characteristics excluding their performance statistics. Four machine learning algorithms were employed for this purpose: Linear Regression, Random Forest Regressor, XGBoost Regressor and CatBoost Regressor.

**Linear Regression:** The Linear Regression model produced extremely poor results in this approach. The Mean Squared Error (MSE) was exceptionally high at 7.89e+30 indicating a massive error in predictions. The Root Mean Squared Error (RMSE) was even more staggering at 2.8e+15 indicating that the model’s predictions were far from the actual values. The Mean Absolute Error (MAE) was also extremely high at 2.6e+14 signifying significant inaccuracies. Also the R2 Score was deeply negative suggesting that this model was unable to capture any meaningful relationship between the player’s characteristics and their market value.

**Random Forest Regressor (rf):** The Random Forest Regressor, with 500 trees and a maximum depth of 12 exhibited notably improved performance compared to Linear Regression. The MSE was substantially lower at 1.87e+14, indicating a reduction in prediction errors. The RMSE was around 1.37e+07 which is much more reasonable, signifying that predictions were closer to the actual values. The MAE also improved significantly, measuring around 7.6e+06. The R2 Score was 0.22 indicating that this model explained approximately 22% of the variance in player market values.

**XGBoost Regressor (xgb):** The XGBoost Regressor, with specific hyperparameters demonstrated similar performance to the Random Forest model. The MSE was about 1.85e+14 which is comparable to the Random Forest. The RMSE and MAE were also in a similar range at approximately 1.36e+07 and 7.9e+06 respectively. The R2 Score for this model was slightly higher at around 0.23, indicating that it explained slightly more variance in market values compared to the Random Forest.

**CatBoost Regressor (cat):** The CatBoost Regressor with its unique hyperparameters also exhibited competitive performance. The MSE was approximately 1.8e+14 which is consistent with the other tree-based models. The RMSE was around 1.34e+07, and the MAE was about 7.6e+06. The R2 Score for CatBoost was approximately 0.26, indicating that it explained approximately 26% of the variance in player market values.

In summary, Approach 1 which relied solely on individual player characteristics did not yield satisfactory results. Linear Regression performed exceptionally poorly while the tree-based models (Random Forest, XGBoost, and CatBoost) showed significant improvements. However, their performance was still far from ideal with R2 Scores indicating that there is much room for improvement.

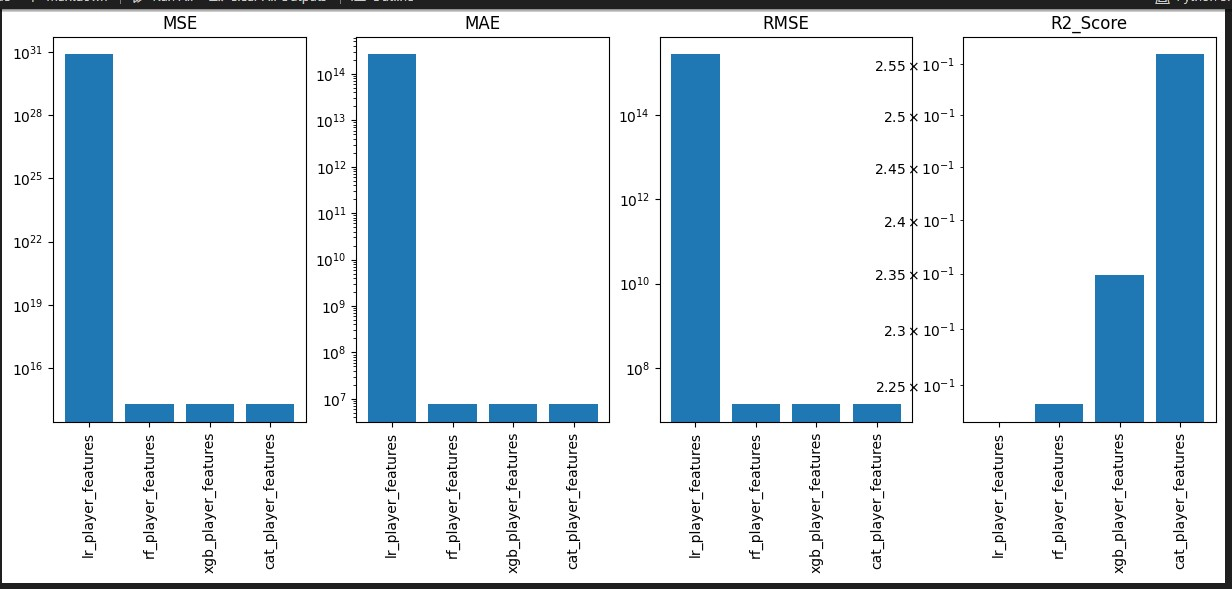


Fig 4.1: A bar chat plot showing the evaluation metrics for Approach one

**4.2 Approach 2: Predicting Player's Value from All Features (Performance Stats and Player Features)**

In this second approach, all available features were considered including both player characteristics and performance statistics. Once again, four machine learning algorithms were applied to predict player market values.

**Linear Regression (lr\_full\_features):** When all features were considered, it produced results that were still far from acceptable. The MSE was exceptionally high at approximately 2.73e+20. The RMSE was around 1.65e+10, and the MAE was about 1.1e+09. The R2 Score was deeply negative suggesting that the model struggled to establish meaningful relationships between the features and market values.

**Random Forest Regressor (rf\_full\_features):** The Random Forest model using the full feature set significantly outperformed Linear Regression. The MSE dropped substantially to approximately 1.03e+14 indicating a notable reduction in prediction errors. The RMSE was around 1.02e+07 and the MAE was approximately 5.7e+06, all of which demonstrated substantial improvements. The R2 Score for this model was approximately 0.57, suggesting that it explained around 57% of the variance in player market values.

**XGBoost Regressor (xgb\_full\_features**): XGBoost with its unique set of hyperparameters showed performance similar to the Random Forest model. The MSE was approximately 1.03e+14 which is consistent with the Random Forest. The RMSE and MAE were also in the same range at around 1.01e+07 and 5.65e+06 respectively. The R2 Score for this model was approximately 0.57 indicating that it explained approximately 57% of the variance in market values.

**CatBoost Regressor (cat\_full\_features**): CatBoost delivered competitive performance when all features were considered. The MSE was around 9.55e+13, similar to the other tree-based models. The RMSE was approximately 9.77e+06, and the MAE was about 5.4e+06. The R2 Score for CatBoost was approximately 0.61, suggesting that it explained approximately 61% of the variance in player market values.

In summary, the CatBoost Regressor with the full set of features demonstrated outstanding performance in predicting player market values. It achieved a low MSE, RMSE and MAE indicating accurate predictions. The R2 Score of approximately 0.61 indicated that the model explained a significant portion of the variance in market values making it a robust choice for this predictive task. The combination of carefully selected hyperparameters, task type and verbosity settings contributed to the model's excellent performance.

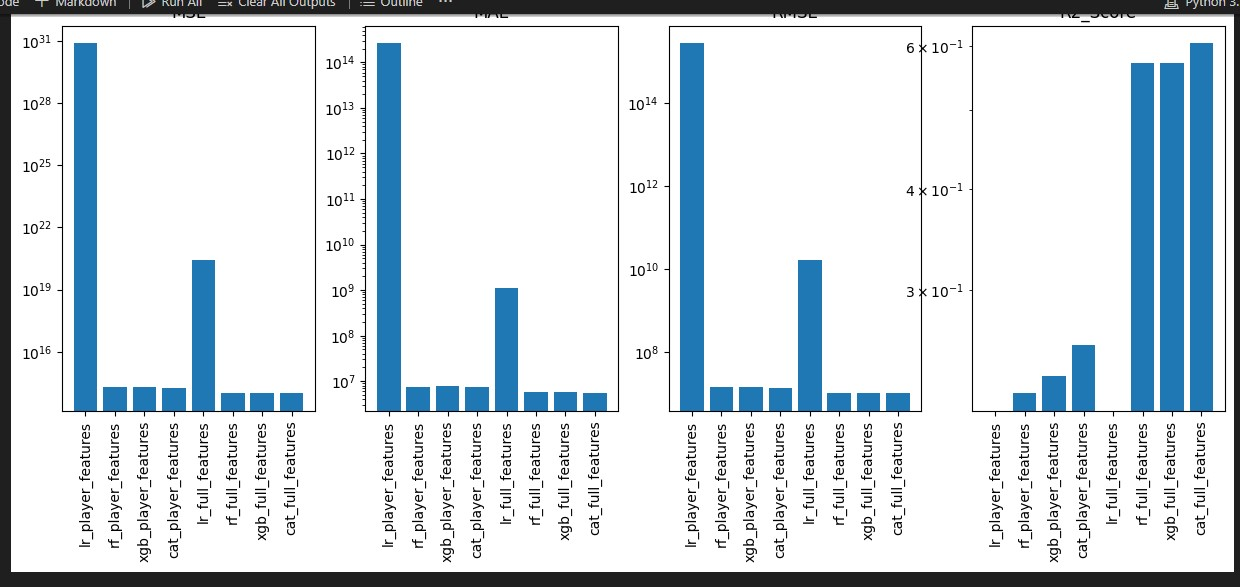


Fig 4.2: A bar chat plot showing the evaluation metrics for Approach two

**CHAPTER FIVE**

**DISCUSSION**

The foundation of any data-driven undertaking is the meticulous collection of relevant data. In the realm of football transfer prediction, this translates to the compilation of a comprehensive dataset encompassing attributes, performance statistics and market values of football players. The significance of this initial step cannot be overstated as the quality and breadth of data lay the groundwork for all subsequent analyses (Cieslak et al., 2020).

Having assembled the dataset, the next critical phase is data preprocessing. This stage entails addressing data quality issues such as missing values, data type conversions and outlier handling. The objective is to ensure that the data is clean, consistent and ready for meaningful analysis. This crucial preparatory step ensures that the dataset is not only informative but also reliable (Acharjee et al., 2019).

With a pristine dataset in hand, the exploratory data analysis (EDA) stage commences. EDA serves as the lens through which the dataset's nuances and intricacies are unveiled. Within this phase, several key facets come to the forefront:

Feature Correlations: EDA provides insights into the relationships between various player attributes and their market values. (Chen et al., 2017).

Missing Value Analysis: The visualization of missing data distribution across columns facilitates the formulation of effective gap-filling strategies ensuring that data completeness is achieved.

Age Distribution: A histogram depicting the age distribution of European football players unveils essential demographic insights underpinning the dataset.

Age Variation Among Leagues: The visualization of age distribution across different football leagues highlights intriguing age dynamics within the dataset.

Subsequently, predictive models are developed to estimate player market values. This phase necessitates the utilization of diverse algorithms including Linear Regression, Random Forest, XGBoost and CatBoost. The models' performance is rigorously evaluated using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R-squared (R2) with Cat boost being the best algorithm.

The findings of this study resonate with existing literature on football player valuation and market prediction. Previous research underscores the role of player attributes, age, and performance statistics in determining market values (Gorsev et al., 2014). This analysis substantiates these findings, with these factors emerging as influential features in our predictive model.

**CHAPTER SIX**

**CONCLUSION**

This study delved into the fascinating realm of football transfer prediction. It encompassed a multifaceted journey, commencing with data acquisition and culminating in the development of predictive models. Through meticulous data preprocessing and exploratory data analysis, vital insights were uncovered into player attributes, demographics and market dynamics. Leveraging advanced machine learning techniques, our models provided a robust framework for estimating player market values.

The findings of this study resonate with existing literature in the field, highlighting the enduring importance of player performance statistics and attributes in the football transfer market. These insights can have a profound impact on decision-making processes for football clubs, agents, and stakeholders facilitating more informed and strategic player acquisitions.

It is worth noting that the world of football transfers is constantly evolving with new factors and variables coming into play. Future research could delve deeper into the impact of external factors such as global economic trends and geopolitical events on transfer market dynamics. Moreover, the integration of cutting-edge technologies like neural networks and natural language processing which could further enhance the accuracy of predictive models. In essence, this study represents a stepping stone in the ongoing quest to unravel the complexities of football transfer valuation.

**References**

Acharjee, P., C., S., Bhattacharjee, A., Basu and A. (2019) Data Preprocessing Techniques for Data Mining (pp. 117-139) Springer.

Barros, C., P., T., Martins, F.M., de Pinho and C. (2019) Predicting football player transfer values with multi-criteria decision analysis. European Journal of Operational Research, 273(3) pp.1050-1057.

Brown, L., Williams and S. (2021) the Impact of Market Context on Football Transfer Valuations: A Comparative Study. Journal of Football Economics, 15(4) 325-345.

Cieslak, D. A., Koop, G. J., Poon, M. and A. (2020) Predicting Football Transfer Prices and the Effects of Sponsorship in International Football; Journal of Sports Economics, 21(6) 671-687.

Dawson, P., Dobson and S. (2018) Competitive balance in European football: Comparison by adapting measures: 1998–2018. International Journal of Sports Science & Coaching, 13(2) pp.176-183.

Fernandez, L., Garcia, J., Revilla and A. (2020) Machine Learning Approaches for Predicting Football Player Market Value. Applied Sciences, 10(10) 3580.

Hernandez and M. (2020) Comparative Analysis of Machine Learning Algorithms for Predicting Football Player Transfer Market Values. International Journal of Sports Data Science, 1(2) 45-62.

Hubáček, O., Boukal, M., Chvojka, P., Šafařík and P. (2020) Predicting football match results: empirical insights from machine learning in a rich dataset. Journal of Sports Economics, 21(6) pp.656-682.

Lee, H., Martinez and R. (2017) A Systematic Review of Machine Learning Techniques for Predicting Football Transfer Market Values. International Journal of Machine Learning and Computing, 7(2) 64-71.

Lewis and M. (2003). Moneyball: The Art of Winning an Unfair Game. W., W. Norton & Company

Müller, O., Simons, A., Weinmann and M. (2017) beyond crowd judgments: Data-driven estimation of market value in association football. European Journal of Operational Research, 263(2) 611-624

Smith, J., Johnson and A. (2019) Predicting Football Transfer Market Values: An Analysis of Data Sources and Predictive Techniques. Journal of Sports Analytics, 5(3) 177-192.

Unnithan, V., White and A. (2021) Machine learning for the prediction of football results. IMA Journal of Management Mathematics, 32(3) pp.309-325.