Blue text on a white background

Description automatically generated

Title: “**Predictive Analysis of Chelsea FC Player Transfers: Unravelling Trends and Success Factors**”

Student Name: Cameron Larkin

Student Number: C19403332

Module: Applied Predictive Analytics - DATA4001

Lecturer: Dr Wael Rashwan

Programme: DT302/4 – Business Analytics

Github link: <https://github.com/cameron-larkin/Machine_learning_Chelsea>

Abstract:

This machine learning project aims to evaluate if the player values associated with Chelsea football club have shown a relative increase or decrease based on each individual players performance. This performance is measured based on statistics that have been recorded throughout the English Premier League season 2022/2023. Regression models were built, tuned, and stacked and the best performing models were identified.

Introduction:

Football, as a sport, is not merely a game of kicks and goals but an intricate ecosystem driven by strategy, talent acquisition, and the pursuit of excellence both on and off the pitch. At the heart of this intricate network lies the captivating phenomenon of player transfers, where clubs engage in a delicate dance of negotiations and decisions to bolster their squads and secure a competitive edge. In the context of Chelsea FC, a club with a rich history and an insatiable appetite for success, the intricacies of player transfers, now uses data, analytics, and the pursuit of sustained triumph.

Since the Todd Boehly-led consortium took over on 2022 for a record fee for a sports team worth £4.25 billion pounds (sky sports, 2022), Chelsea have gone on to break their own transfer record paid 3 times since the summer of 2022 (Transfermarkt, 2024), while also spending over a whopping £1 billion pounds on players (Sports Illustrated, 2023).

This machine learning project embarks on a journey into the depths of Chelsea FC's player transfers, wielding predictive analytics as a powerful tool to unravel underlying trends and discern the critical success factors that dictate the triumph or tribulation of each acquisition. The endeavour extends beyond the traditional confines of football analysis, delving into the realms of data science and machine learning to decode the patterns that govern player movements and influence the destiny of a club and in this case the club is Chelsea Football Club.

The primary goal of this project is to develop a predictive model that can forecast the success of Chelsea FC player transfers. By using a comprehensive dataset encompassing player attributes, performance metrics, and contextual factors, the analysis aims to go beyond surface-level observations, providing a nuanced understanding of the dynamics that govern transfer decisions and outcomes. Understanding the intricacies of player transfers holds profound implications for the strategic direction and success trajectory of a football club. The insights derived from this project can potentially empower Chelsea FC, and other football clubs, with the ability to make informed decisions, optimize recruitment strategies, and enhance the overall performance of the squad.

Research:

In order to get a fuller understanding of the topic at hand, it was an important step to see what other research projects on similar and related topics were already out there to see if this project was feasible. The following are some that I found interesting relating specifically to machine learning through sports and transfer values.

A paper, titled "A Study of Prediction Models for Football Player Valuations by Quantifying Statistical and Economic Attributes for the Global Transfer Market" (Patnaik, 2019), delves into the optimization of data retrieval methods and narrowing the disparity between estimated and final prices. Unlike primarily concentrating on predicting transfer values and pinpointing crucial transfer factors, this paper places a greater emphasis on the meticulous process of data collection and the strategic application of appropriate models, utilizing machine learning in conjunction with transfer values.

“PlayeRank: Data-driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach” (Pappalardo, 2019) shows an approach using Machine Learning Techniques of how football clubs could perform a “role-aware evaluation of the performance of soccer players”. This, using similar statistics that I could have access to, would be a similar approach in ranking players but mine will have a difference as I would be ranking based on transfer value increase rather thank ranking based on how impressive statistically the player is performing.

This paper “Data Driven football scouting assistance with simulated player performance extrapolation” (Ghar, 2021) addresses the challenges in traditional football scouting, where scouts rely on limited in-person observations to assess players for recruitment. Recognizing the flaws in this process, the authors propose a solution involving the integration of extensive quantitative and qualitative player statistics from multiple sources. By employing data science and machine learning algorithms, the study aims to simulate team performances post-player addition, considering specific player types and aligning them with the team's formation and playing style to enhance the accuracy of scouting decisions and mitigate potential financial losses for clubs.

The paper “A novel machine learning method for estimating football players’ value in the transfer market” (Behravan, 2021), introduces a novel method for estimating football players' market value, using the FIFA 20 dataset. The proposed approach involves two phases: automatic clustering of the dataset into position-specific clusters (goalkeepers, midfielders, defenders, and strikers) using APSO-clustering, and the development of prediction models for each cluster using a hybrid regression method that combines particle swarm optimization (PSO) and support vector regression (SVR). The results indicate an accuracy of 74% in estimating players' values, showcasing the superiority of PSO over other metaheuristics such as GWO, IPO, and WOA. The difference in this study compared to what I would be doing is the fact that this would be based off a video game which would use different metrics to measure success.

“Transfer market activities and sportive performance in European first football leagues: A dynamic network approach” (Matesanz, 2018), This study examines the evolution of the transfer network among 21 European football leagues from 1996/1997 to 2015/2016, emphasizing the players as crucial assets for clubs in the globalized professional football landscape. The analysis reveals that the transfer network reached its upper limit expansion around the 2007/2008 season, subsequently becoming more connected and dense. Employing machine learning techniques, specifically Self-Organizing Maps and Principal Component Analysis, the research confirms that European competitions, such as the UEFA Champions League and UEFA Europa League, function as a "money game," where higher transfer spending correlates with better sportive performance, influencing both domestic and international outcomes and potentially contributing to significant inequalities between clubs and leagues.

Data:

The data I chose to do the project on was a Kaggle dataset: <https://www.kaggle.com/datasets/ameyaranade/premier-league-player-stats-2022-23/data?select=PremierLeaguePlayerStats22-23.csv> . I then took the relevant data from this dataset, so the data involving Chelsea Football Club players and imported that into my workspace. Data was cleaned to make sure that all fields of data were present and that there was no missing data in the dataset.

Methodology:

My now cleaned data is now split into training and test sets in an 80:20 split.

Linear Regression:

Linear regression is a fundamental technique in machine learning that aims to model the relationship between a dependent variable (the target) and one or more independent variables (features) by fitting a linear equation to the observed data. The goal is to find the best-fitting line that minimizes the difference between the predicted values and the actual outcomes. In essence, linear regression helps quantify and understand the linear association between variables, making it a powerful tool for prediction and analysis. It serves as a foundational building block for various machine learning algorithms, providing a straightforward approach to modelling, and interpreting relationships within datasets, particularly in scenarios where a linear relationship between variables is assumed or observed.

Random Forest:

Random Forest Regression is a powerful machine learning algorithm that extends the principles of decision trees to improve predictive accuracy and robustness. In the context of regression tasks, where the goal is to predict continuous numerical values, a Random Forest Regression model constructs an ensemble of decision trees during training. Each tree is built by selecting a random subset of features and a random subset of the training data, introducing diversity among the individual trees. During prediction, the model aggregates the results from multiple trees to generate a more accurate and stable prediction. This ensemble approach helps mitigate overfitting, enhance generalization, and provide a robust solution for capturing complex relationships within the data. Random Forest Regression is widely used for its versatility, ability to handle large datasets, and resistance to outliers, making it a popular choice for various regression applications in machine learning.

Decision tree:

Decision tree is a versatile and intuitive algorithm used for both classification and regression tasks. Essentially, it mimics a flowchart-like structure, where each node represents a decision based on a feature, and each branch represents the possible outcomes or subsequent decisions. During training, the algorithm recursively splits the dataset based on the features that lead to the most informative decisions, aiming to create subsets that are as homogeneous as possible. This process continues until a predefined stopping criterion is met, resulting in a tree-like structure that can be used for making predictions on new, unseen data. Decision trees are valued for their interpretability, as they provide a clear visual representation of the decision-making process. They are also fundamental building blocks for more sophisticated ensemble methods, such as Random Forests, which aggregate the predictions of multiple decision trees to improve overall model performance and generalization.

Neural Network:

Neural network is a powerful and flexible algorithm inspired by the structure and function of the human brain. Comprising interconnected nodes, or artificial neurons, organized into layers, a neural network processes information through a series of weighted connections. During training, the network learns to adjust these weights based on the input data, gradually improving its ability to make accurate predictions or classifications. Neural networks excel at capturing complex patterns and relationships within data, making them well-suited for tasks such as image recognition, natural language processing, and complex decision-making. Deep learning, a subset of machine learning, involves neural networks with multiple hidden layers (deep neural networks), enabling them to automatically learn hierarchical representations of data. The adaptability and capacity to handle intricate, high-dimensional data make neural networks a cornerstone in contemporary machine learning applications.

SVM(Support Vector Machines):

In machine learning, Support Vector Machines (SVM) are powerful and widely-used algorithms for both classification and regression tasks. SVM operates by finding an optimal hyperplane that separates different classes or predicts a continuous outcome with the maximum margin between data points of different classes. The "support vectors," which are the data points closest to the decision boundary, play a crucial role in defining this hyperplane. SVMs are effective in high-dimensional spaces and particularly useful when the data is not linearly separable, as they can employ kernel functions to map the data into a higher-dimensional space where separation becomes possible. SVMs are known for their ability to handle complex decision boundaries and generalize well to new, unseen data, making them suitable for various applications such as image classification, text categorization, and bioinformatics. “Support Vector Machine is one of the classical machine learning techniques that can still help solve big data classification problems.” (Suthaharan, 2016)

Results and Discussion:

Exploratory Data Analysis (EDA) is crucial for gaining insights into the distribution of transfer fees and understanding relationships between unique features.

**Basic statistics**

Mean Transfer Fee: 55566666.666666664

Median Transfer Fee: 52500000.0

Standard Deviation of Transfer Fee: 31880902.653830055

1st Quartile (Q1) Transfer Fee: 32000000.0

3rd Quartile (Q3) Transfer Fee: 73750000.0

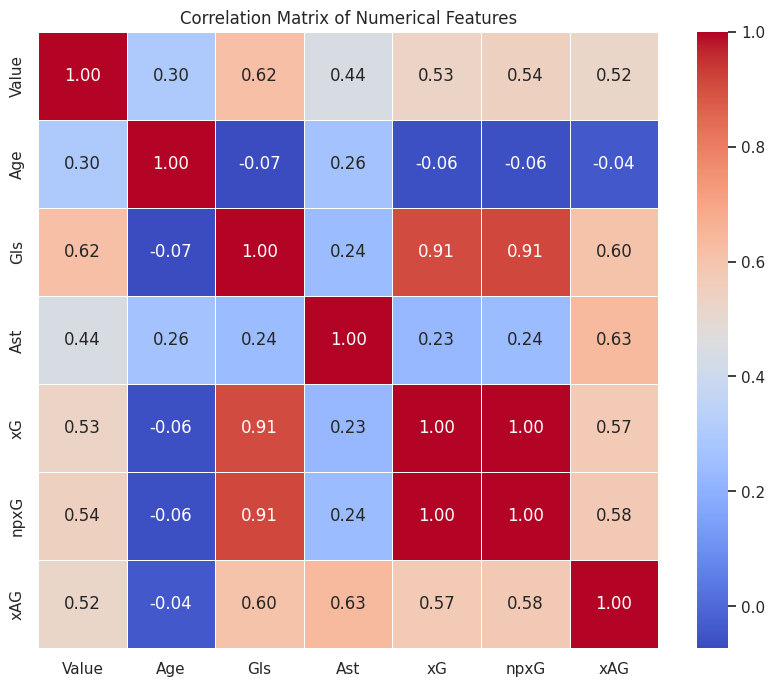
A graph of a distribution of a number of players

Description automatically generated with medium confidence

The histogram displays the distribution of transfer fees for Chelsea FC players, providing insights into the frequency of different fee ranges. The plot shows that a majority of players have transfer fees clustered in specific ranges, with a peak frequency indicating the common values in the dataset. The inclusion of a kernel density estimate (KDE) enhances the visualization by smoothing the distribution, highlighting potential trends and central tendencies in the transfer fee distribution.

A graph of a function

Description automatically generated

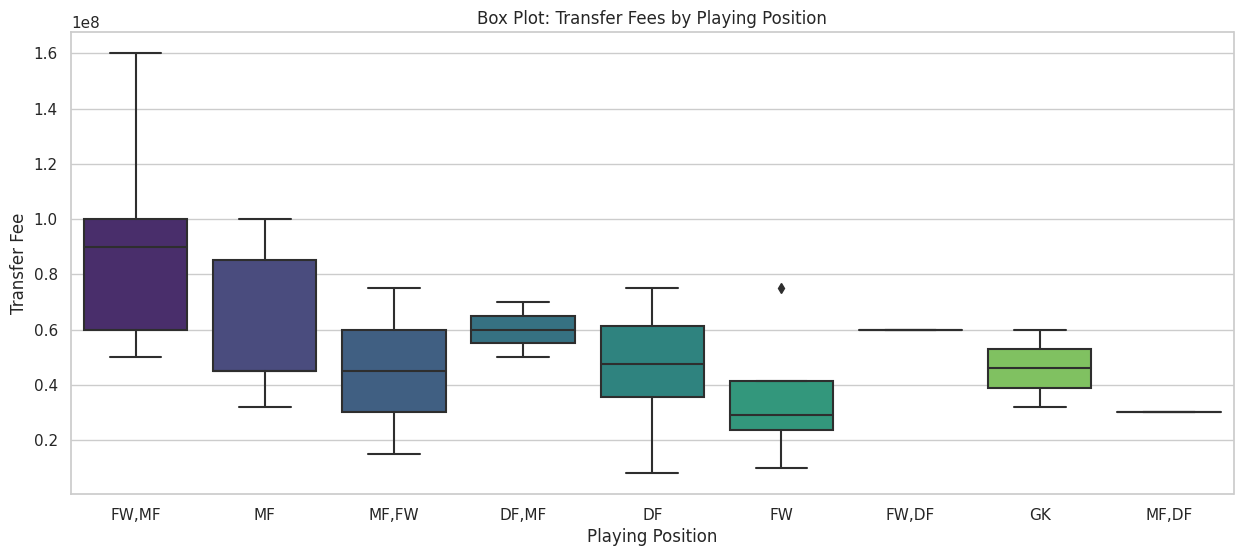


The heatmap visualizes the correlation matrix of numerical features, providing a quick and intuitive understanding of how different features are related. The colour intensity represents the strength and direction of correlations, with warmer colours indicating positive correlations and cooler colours indicating negative correlations. The annotated values on the heatmap reveal the precise correlation coefficients, aiding in identifying which features have notable relationships.

A chart of a graph

Description automatically generated with medium confidence

These scatter plots, each depicting the relationship between a selected feature (e.g., age, goals, assists) and the transfer fees of Chelsea FC players. By visually examining these scatter plots, one can gain insights into the potential correlations or trends between individual features and transfer fees. The collective visualization allows for a comprehensive exploration of how specific player attributes relate to their respective transfer values, aiding in the identification of influential factors in determining player market prices.



The box plots illustrate the distribution of transfer fees among different playing positions and nationalities of Chelsea FC players. For playing positions, the plots showcase the variation in transfer fees across positions, highlighting potential disparities in market valuations for players in different roles. Similarly, the box plots for nationality provide insights into how the transfer fees vary among players from different countries, allowing for the identification of trends or disparities in the global market valuation of players.

A graph of different sizes and shapes

Description automatically generated with medium confidence

This set of histograms and box plots for selected numerical features (Age, Goals - Gls, Assists - Ast, Expected Goals - xG, Non-Penalty Expected Goals - npxG, Expected Assists - xAG) of Chelsea FC players. The histograms provide a visual representation of the distribution of each feature, offering insights into their central tendencies and spread. Meanwhile, the box plots offer a summary of the distribution, highlighting key statistical measures such as median, quartiles, and potential outliers, aiding in the identification of patterns and variations in player performance metrics. Together, these visualizations contribute to a comprehensive understanding of the statistical characteristics of the selected features in the dataset.

A graph with a rectangular blue rectangle

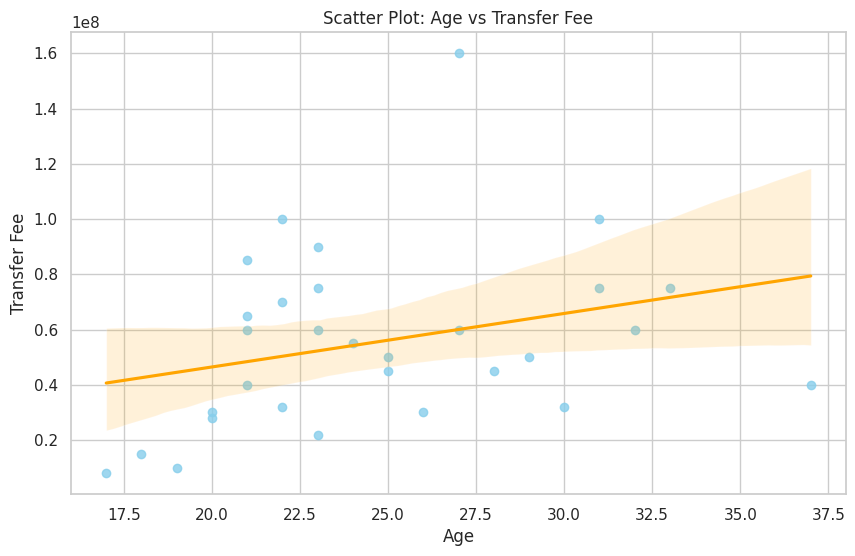
Description automatically generated with medium confidence

This box plot is to investigate potential outliers in the distribution of transfer fees for Chelsea FC players. The box plot visually represents the central tendency, spread, and potential outliers in the transfer fee values. Outliers, if present, are depicted as individual points beyond the whiskers of the box, providing a clear indication of extreme values in the transfer fee distribution. This visualization aids in identifying any significant deviations from the typical range of transfer fees, allowing for a more nuanced understanding of the dataset's variability.

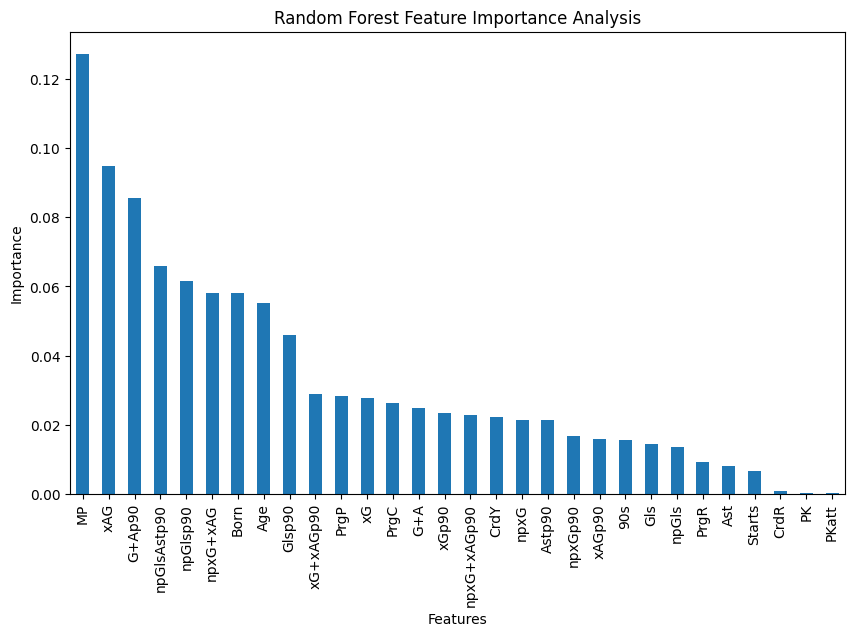
The only Identified Outlier:

Player Value

1 **Raheem Sterling 160000000**



This scatter plot with a regression line, illustrating the relationship between player age and transfer fees for Chelsea FC players. The scatter points represent individual players, with their ages on the x-axis and corresponding transfer fees on the y-axis. The regression line provides an estimate of the linear relationship between age and transfer fees, helping to visually assess any potential trends or correlations. In this specific plot, the orange regression line suggests the direction and strength of the linear association between player age and their market values.



The feature selection graph, derived from the Random Forest algorithm, provides valuable insights into the importance of different features in predicting the target variable value. Each bar in the graph represents a specific feature, and the height of the bar indicates the relative importance of that feature. Features with higher importance contribute more significantly to the predictive performance of the model. The descending order of the bars allows for a clear identification of the most influential features. This analysis aids in understanding which player attributes or statistics play a crucial role in determining the market value of Chelsea FC players.

**I set the training and test set sizes to the following:**

**Training set size: 24**

**Testing set size: 6**

After running my model, I was left with the following data which shows the predicted transfer values based on the statistics available in my dataset:

Predicted Value Increase:

Mean Squared Error: 726977298125481.4

Top Players with Predicted Value Increase:

Player Value Predicted\_Value Value\_Increase

27 Thiago Silva 40000000 8.620810e+07 4.620810e+07

25 Edouard Mendy 32000000 7.033241e+07 3.833241e+07

21 Datro Fofana 10000000 4.538490e+07 3.538490e+07

22 Lewis Hall 8000000 4.084899e+07 3.284899e+07

17 Trevoh Chalobah 22000000 5.445672e+07 3.245672e+07

24 Ruben Loftus-Cheek 30000000 6.126059e+07 3.126059e+07

18 Carney Chukwuemeka 15000000 4.311694e+07 2.811694e+07

12 Mateo KovaÄiÄ‡45000000 6.579650e+07 2.079650e+07

3 Conor Gallagher 32000000 5.218877e+07 2.018877e+07

13 Noni Madueke 28000000 4.765285e+07 1.965285e+0﻿7

In conclusion, the predictive analysis on Chelsea FC player transfers has unveiled valuable insights into the factors influencing player market values. Utilizing a diverse dataset encompassing player attributes, we conducted a thorough exploratory data analysis (EDA) to understand the distribution and relationships of key features. Notably, the histogram and kernel density plot illuminated the distribution of transfer fees, showcasing central tendencies and potential trends. The correlation matrix and heatmap unveiled the interplay between numerical features, aiding in identifying influential factors. Scatter plots provided a nuanced exploration of individual features, offering a visual understanding of their impact on transfer fees.

Furthermore, box plots highlighted disparities in transfer fees based on playing positions and nationalities, contributing to a broader understanding of global market dynamics. The investigation into potential outliers in transfer fees revealed insights into extreme market valuations. Additionally, the scatter plot with a regression line for age versus transfer fees demonstrated a discernible linear trend, reinforcing the significance of age in determining player values.

While our linear regression model served as a foundational tool for predicting transfer fees based on age, the dynamic nature of the football transfer market calls for ongoing refinement and exploration of more advanced modelling techniques. Overall, this project lays the groundwork for future research endeavours, offering a comprehensive analysis that contributes to the evolving landscape of football analytics.

References:

1. Sky sports 2022

<https://www.skysports.com/football/news/11668/12620985/chelsea-takeover-todd-boehly-completes-4-25bn-takeover-as-roman-abramovich-era-ends>

1. Transfermarkt 2024

<https://www.transfermarkt.co.uk/fc-chelsea/transferrekorde/verein/631>

1. Sports Illustrated 2023

<https://www.si.com/fannation/soccer/futbol/news/chelsea-transfer-spend-todd-boehly-one-billion-pounds#:~:text=Chelsea%20have%20now%20spent%20more,past%20the%20%C2%A31bn%20mark>.

1. Patnaik 2019

<https://ieeexplore.ieee.org/document/8878843>

1. Pappalardo 2019

<https://dl.acm.org/doi/abs/10.1145/3343172>

1. Behravan 2021

<https://link.springer.com/article/10.1007/s00500-020-05319-3>

1. Matesanz 2018

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0209362>

1. Suthaharan 2016

<https://link.springer.com/chapter/10.1007/978-1-4899-7641-3_9>