National College of Ireland

Concise Analysis of European Club Transfer Market Practices Using Seasonal Performance Analysis

MSc Research Project
Data Analytics

Saurav Jaglan
Student ID: 22105433

School of Computing
National College of Ireland

Supervisor: Dr.Athanasios Staikopoulos

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Saurav Jaglan
Student ID:	22105433
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Dr.Athanasios Staikopoulos
Submission Due Date:	14/12/2023
Project Title:	Concise Analysis of European Club Transfer Market Practices Using Seasonal Performance Analysis
Word Count:	6673
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Saurav Jaglan
Date:	29th January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Concise Analysis of European Club Transfer Market Practices Using Seasonal Performance Analysis

Saurav Jaglan
22105433

Abstract

Football stands as one of the most profitable sports in the world, and recently we have seen football clubs spending generous amounts of money to obtain best talents in the world. This big money spree does not always result in value for money, and clubs might even risk becoming bankrupt. This research involves the implementation of machine learning practices in modern football to strategize financial spendings on players. The research is divided into two parts, in the first, using Random Poisson and KNN for forcasting season end ranking of the leagues, based on the clubs ranking they might identify target player, the model was evaluated using real results by comparing predicted match outcomes vs real match outcomes,and the KNN model also achieved an average accuracy of 78%. Once the 1st stage was complete the 2nd stage aimed in predicting the value of players in different leagues using three different algorithms Gradient Boosting, Random Forest, and Decision Tree. It was found Gradient Boosting was the best suited algorithm which achieved an average accuracy of 91%. By assistance of these two models clubs can simulate their possible season end ranking, and estimate value of target players in different leagues during the transfer window and sign them without overspending.

Keywords: Football, Transfer market, Value, Player, Machine Learning.

1 Introduction

Football has grown into one of the world's most popular and highest-earning sports as said by "Finance Football"[1]. In recent years, the game has undergone significant changes: clubs have adopted new strategies, and new rules have been put in place to ensure the game is both safe and enjoyable. One major change, introduced by FIFA (the global football governing body), was the transfer market system. Since 2001, FIFA has allowed clubs to buy and sell players during two specific times of the year. Clubs have adjusted well to this system, using it not only to keep their teams stable but also to make a lot of money by selling players they do not need. Today's football scene shows more professionalism, a global presence, advanced playing tactics, better training, and the use of technology. However, the way players are transferred has not changed much over the years, many club owners, focused on winning championships, are ready to spend huge amounts of money to keep the best players. For them, spending money is not a big issue; their main goal is to win titles. But spending a lot of money does not always lead to

[1] `https://financefootball.com/2019/10/14/which-sport-makes-the-most-money/`

success, even if a player is very expensive, it does not mean he will always play at his best. A player's performance depends on many things, such as how well the team works together, how consistently he plays, and his history of injuries.

1.1 Motivation

The transfer market offers clubs a chance to not only bring in players but also to boost their finances. However, the success of this depends on the club's chosen strategy, football is evolving and becoming more modern, making it crucial to use technology to figure out the best approach during the transfer window. Over time, we have seen clear examples that spending a lot of money does not always lead to success. A prime example is Chelsea Football Club, as pointed out in an article[2] , "Money doesn't always buy success just ask Chelsea". Despite spending about £2.36 billion in their last two transfer windows, Chelsea did not get the results they hoped for. Clubs often take the risk of breaking financial fair play (FFP) rules to buy top players, and if caught, they face serious penalties. Not managing finances properly has also led to instances of bankruptcy. The rise of machine learning has opened new possibilities and ways of thinking in sports. It is time to focus on technical strategies rather than just spending big, and this can be achieved by using machine learning methods. This can be achieved by simulating league rankings, which will give football teams a possible season end league rank, knowing that clubs will be able to focus on their weak areas in case the simulated league rank is not what they expect to be. Once the club knows which field position, they want to improve they can predict value of players, in different leagues which could help them managing their budget.

1.2 Objective

This study aims to create a model that helps football clubs plan their transfer window strategies, focusing on buying players at a reasonable cost.

The model will first predict the seasonal rankings of the top three European leagues Bundesliga, Premier League, and Serie A. Using data from the past 6-7 seasons to forecast league standings, this will help identify the team position in their respective league. Now if the team is predicted to finish on the rank they expect, then they might look for players to increase the depth of their squad, but if this is not the case the clubs might need to identify players which will help them in improving the rank.

So the second model will predict the value of target player in the three leagues, this forecast will be based on player data, enabling clubs to decide if they should go ahead with the purchase, how much to invest, or whether they should consider a different player with similar skills but at a lower cost.

1.3 Research question

In what ways do league ranking simulations inform football clubs about their relative competitive standings, and in which ways the predictive analytics applied to estimate player values in various leagues can stop clubs overspending?

[2] `https://www.thisisanfield.com/2023/09/money-doesnt-always-buy-success-just-ask-chelsea/`

1.4 Structure of Paper

This segment of the introduction will provide brief layout of research paper.

- Section 2: This section outlines prior research and related works, focusing on their relevance to the present study while identifying both similarities and areas where knowledge is lacking.
- Section 3: The methodology section details the data mining techniques applied in this research and outlines the steps involved in the process.
- Section 4: This section emphasizes the various phases and techniques employed in conducting the research.
- Section 5: The implementation section covers the process from data acquisition to model creation, including a discussion on the tools and Python libraries utilized.
- Section 6: This section delves into the simulation of league rankings of the three leagues, along with discussing the varied estimated values of the players by using machine learning models.
- Section 7: This section addresses the conclusions drawn from this research and explores its potential future applications and scope.

2 Related Work

Numerous research efforts have historically focused on enhancing player performance. However, in team sports like football, which is one of the most popular and high-revenue-generating sports globally, the emphasis is not just on individual performance. Instead, as the term 'team sports' implies, it is about combining a group of players who each perform at their highest level, this section is focused into past studies that are closely related to this aspect of team sports, particularly football.

2.1 Predicting season outcome

This study AKTUĞ et al. (2022) aimed to forecast the final league rankings in the German Football League known as Bundesliga, Artificial Neural Network (ANN) algorithm was used to build the model. The research focused on analysing the times of goals scored and conceded during matches across three seasons by which it calculated the outcome of matches played by individual teams. The result indicated a high accuracy rate of almost 99% and mean squared error value of 0.00004, in predicting league standings, this shows the impact of scoring and conceding goals at specific times during the match.

The following research's Munđar and Šimić (2016) aim was to predict the season final ranking of the Croatian Football League using simulation modelling. It is focused on modelling the number of goals scored by each team as a Poisson random variable, using these predictions to simulate the rest of the season and estimate the final rankings. The final rankings were predicted based on around 1000 simulation runs for the season's second half. The methodology was validated using data from the 2014/15 season, which successfully predicting the rankings for the top five teams. This method offers a variable

way to predict final rankings during the season which also predict the probabilities of different ranking outcomes.

A study Corona et al. (2019) utilised a Bayesian approach to probabilistic sports forecasting, particularly for the UEFA Champions which is different from a league competition. The study utilized probabilistic forecasting models for both individual matches and the overall tournament, this approach was based on prior research that employed match-level forecasting models followed by Monte Carlo simulations of the tournament to estimate outcome probabilities. A key innovation of this study was the incorporation of a Bayesian approach to address the uncertainty surrounding parameter estimates in the underlying match-level forecasting model, this also enhanced the model's accuracy and reliability. To complete the Bayesian formulation, the study defined a prior distribution for the regression coefficients and employed MCMC (Markov Chain Monte Carlo) methods to generate an approximate sample of values from the posterior distribution.

The paper Joseph et al. (2006) compares multiple machine learning models, including a decision tree (MC4), Naive Bayesian learner, Data Driven Bayesian learner, K-nearest neighbour, and an expert-constructed Bayesian Network (BN). It is focused on Tottenham Hotspur Football Club, providing detailed insights into the performance of these models in a real-world, specific context. The K-Nearest Neighbour (KNN) model exhibited high classification errors, ranging from 61.11% to 68.52% across different seasons and models, similarly the Naive Bayesian Learner also showed significant classification errors, with reductions around 26.31% and 15.78% for the general and expert models in the 1995/1996 season, respectively but when using disjoint training and test data sets, the Naive Bayesian Learner's error rate was 61.19% for the general model and 64.26% for the expert model. These findings highlight the challenges in accurately predicting football match outcomes using these machine learning models.

These studies marked the initial integration of machine learning in modern football, recognizing that every club seeks insights into their potential success. The introduction of new technology in sports is essential. While the research was intriguing, it fell short in comparing different leagues. Predicting and comparing a single league's outcomes with actual rankings is a solid approach to assess the model. However, the analysis should extend beyond just one league for a more comprehensive evaluation.

2.2 Predicting match outcome

In the study Baboota and Kaur (2019) on predictive analysis of professional soccer matches, various machine learning models were assessed for their ability to forecast match outcomes. The Radial Basis Function Support Vector Machine (RBF SVM) excelled in predicting home wins with a high recall of 0.80, though it was less effective for draws. The Random Forest model achieved a moderate accuracy of 0.57 and displayed balanced performance across all match outcomes. Gaussian Naive Bayes, with a mean accuracy of 0.519, performed surprisingly well in predicting draws. However, it was the Gradient Boosting model that stood out as the top performer, surpassing others in overall effectiveness.

The study Rodrigues and Pinto (2022) applies a variety of machine learning algorithms to a comprehensive dataset of 1900 matches from the English Premier League, spanning five seasons, out of those key algorithms include Naive Bayes, K-nearest neighbors, Random Forest, Support Vector Machines, and others, with the Boruta algorithm aiding in critical variable selection. The methodology involves a two-phase prediction process,

initially using 18 variables, then exploring combinations to optimize hit rates. Notably, the research found Support Vector Machine as the most effective algorithm, achieving 61.32% accuracy and significant betting profits.

The research Ren and Susnjak (2022) offers an in-depth exploration of using machine learning algorithms to forecast football match results, focusing on the English Premier League from 2019-2021. It employs a diverse array of techniques, including CatBoost, Decision Trees, Gradient Boost, and others, enhanced by feature selection guided by Shapley Additive explanations (SHAP). The study categorizes matches using the Kelly Index to gauge uncertainty levels, highlighting the dynamic nature of football. Notably, it assesses the models' accuracy, precision, recall, and F1 Score, with ensemble-based algorithms generally showing superior performance.

In a research Yang (2019) the author evaluated four machine learning algorithms: Random Forest, Support Vector Machine, Naïve Bayes, and K-Nearest Neighbour, to predict football match outcomes. The study found Random Forest to be the most effective, utilizing player-specific features like influence, creativity, threat, and BPS index, achieving an impressive 80.8% accuracy, the model demonstrated a strong capability in match prediction. The research provides valuable insights for practical applications, assisting coaches, fans, and gamers in formulating strategies based on player attributes and their impact on match outcomes.

The studies indicated that relying on simple algorithms does not yield satisfactory results. In some cases, there was a strong recommendation for using ensemble methods. Many of these research efforts wisely chose to gather data across multiple seasons rather than limiting it to a single season. However, these models primarily focused on predicting single match outcomes rather than forecasting the entire season. Additionally, there was a noticeable absence of player value evaluation. Assessing player values could significantly assist clubs in strategizing for the transfer window, enabling them to reinforce their squads more effectively.

2.3 Estimating players value

In their innovative study Al-Asadi and Tasdemır (2022) the authors explored the use of machine learning to estimate football players' market values based on FIFA 20 video game data, employing different regression models like Linear Regression, Multiple Linear Regression, Decision Trees, and Random Forests, the research focused on analyzing player data to predict market values. The model's performance was evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2) across a dataset of 17,980 players. While the study demonstrated a reasonable accuracy in its predictions, it notably lacked specific values for its evaluation metrics.

In 2015 a study Nazari and Azari (2021) in which, the authors employ an innovative network analysis to dissect the European football transfer market of summer 2014. Analysing 436 transfers across three major leagues, the study quantifies transfer values using a novel Google-based standard and validates these against actual fees, achieving a notable Pearson correlation of 0.68. The findings challenge traditional assumptions, revealing a random transfer pattern and highlighting the capitalistic nature of the English Premier League. While the study is based on Google search data and its focus on a single transfer window present limitation.

The research Majewski (2021) presents a pioneering approach in sports economics by using econometric models to identify football players as brands. Data from the 150 most

valuable forwards was analysed, and utilizing feasible generalized least square (FGLS) and ordinary least square (OLS) methods for robust statistical verification. It successfully pinpointed notable player-brands like Lionel Messi and Cristiano Ronaldo, demonstrating how brand variables significantly increase a player's theoretical market value, while insightful, the research's focus on forwards and reliance on a single data source limits its broader application.

This research Hoey et al. (2021) critically evaluate the effectiveness of the European football transfer system in redistributing revenues. Their groundbreaking empirical investigation, a first in this field, reveals that the system offers only a marginal reduction in revenue inequality, primarily benefiting middle-market clubs over smaller ones. Analyzing over €10 billion in transfer fees, the study challenges the conventional belief that the system substantially aids financial parity among clubs. While proposing hypotheses on the limited revenue redistribution, the research underscores the need for alternative models to more equitably distribute resources in European football, thus providing a substantial contribution to the discourse on sports economics and club management.

The Table 1 collectively delve into the complex dynamics of the football transfer market, examining various aspects such as player valuations, transfer patterns, and market strategies. A common theme across these studies is the application of advanced statistical models like regression analysis and clustering techniques to analyze data cross several years. These papers highlight the significant influence of player performance, geographical and financial factors, and strategic market engagements on transfer activities. They reveal a trend where clubs prefer international purchases and national sales for profit maximization. Additionally, a recurring finding is the substantial impact of transfer strategies on a team's performance in leagues.

The transfer market plays a crucial role, particularly highlighting the financial disparity between smaller and wealthier clubs. Smaller clubs often struggle to generate significant revenue, while wealthier clubs face fewer financial constraints. Implementing machine learning approaches could offer smaller clubs a strategic advantage in the transfer market. By understanding which clubs are willing to invest substantially in players, smaller clubs could strategically navigate the market. This approach would enable them to retain promising young talent and maximize revenue, making the most of their resources and opportunities.

Previous studies in this domain typically relied on data from either very recent seasons (spanning the last 2-3 years) or significantly older datasets. This approach raised concerns about the accuracy of the models used in simulating football league ranks and match outcomes. Furthermore, many of these studies were confined to analysing a single league, lacking a comparave analysis across different leagues. Additionally, earlier research primarily concentrated on forecasting the value of players, rather than examining how a player's value might vary across different leagues, recognizing that a player's market value is not uniform across all leagues.

Paper by	**Aim**	**Models Used**	**Evaluation Matrices**	**Result**
Van den Berg (2011)	To investigate the determinants of transfer fees in the English Premier League, focusing on the valuation of players as a form of human capital.	Regression analysis approach	Sample Size=42, Adjusted R Squared=0.76 Sample Size=279, Adjusted R Squared=0.43 Sample Size=149,Adjusted R Squared=0.51	The study found that individual player performance and ability are significant determinants of transfer fees.
Wand (2022)	To construct and analyze a trade network between football clubs between 1992 and 2020.	Linear Regression Analysis	R^2 above 0.88	The transfer activity patterns of football clubs are influenced by geographical and financial similarities.
Rossetti and Caproni (2016)	To analyze club market strategies over 25 years of football-related data within the UEFA confederation.	K-means Clustering K-mode Clustering	No specific evaluation metric, focused on pattern identification	Clubs prefer buying players internationally and selling them nationally for major profits.
Dieles (n.d.)	To explore if and to what extent a relation exists between a team's role in the transfer market and their performance within a football league.	Multiple Linear Regression	Best model Adjusted R Squared value 0.77	Relationship between a team's role and engagement in the transfer market and their performance in football leagues, suggesting that a strategic approach to transfers could benefit a team's league performance

Table 1: Summary of Research Papers on Football Transfer Fees

3 Methodology

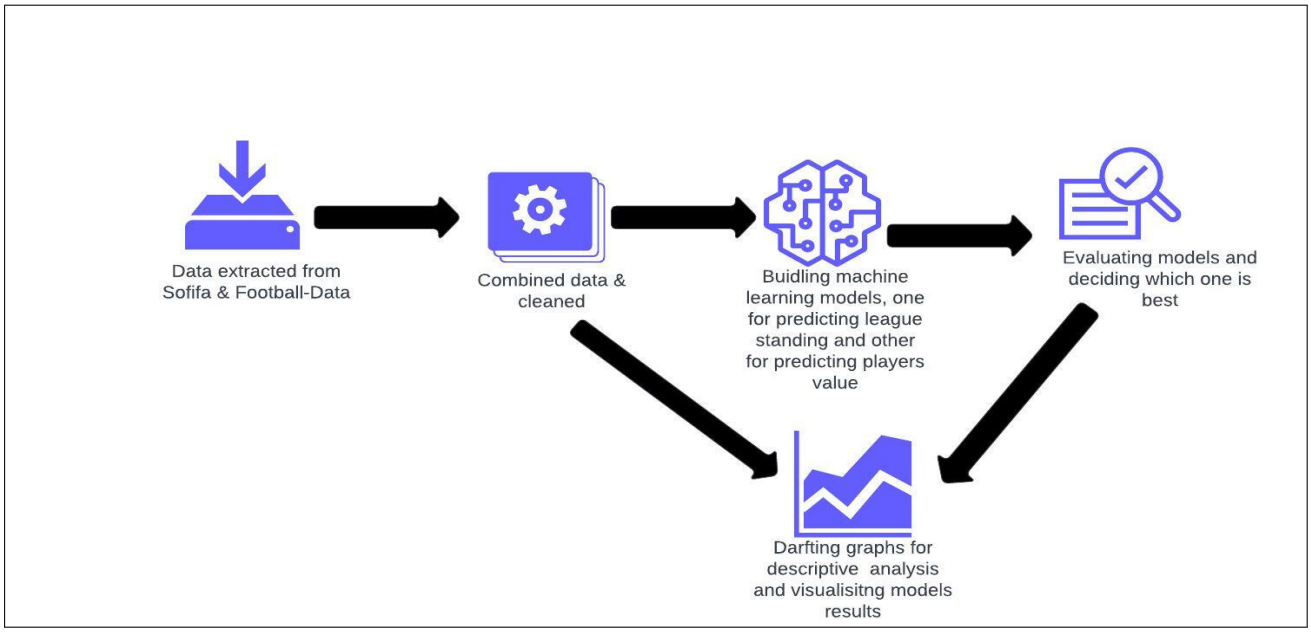

Figure 1: KDD Methodology

The objective of the research question suggested the adoption of the Knowledge Discovery in Databases (KDD) methodology for the entirety of this research project as shown in Fig.1. The KDD method, a comprehensive data mining approach, comprises multiple steps that, when implemented, assist in unveiling significant information from the data set. The primary stages incorporated in this research included:

3.1 Data Selection

The initial step, following the understanding of the research aim, include identifying data sets that will allow answering the research question. In this study, two distinct data sets were selected, the first dataset contained historical match data from 2016 to 2022 taken from "Football-Data"[3], covering three different football leagues: the Bundesliga, Premier League, and Serie A. The second dataset comprised data on players participating in these leagues, detailing their attributes and current market valuations taken from website"Sofifa"[4].

3.2 Data Preprocessing

The subsequent phase involved data cleansing and the potential creation of additional columns derived from existing ones as necessary. In the case of the matches data, there was no requirement for cleaning, as this dataset did not have missing values or outliers, encoding of this data was done to convert columns contiaing strings into numeric. However, the players' data required thorough cleaning, new columns were generated in this dataset, and those unnecessary were eliminated.

[3] `https://www.football-data.co.uk/`
[4] `sofifa.com`

3.3 Data Transformation

This stage include converting data into a structured format, preparing it for initial analysis and subsequent use in creating a machine learning model. In this research, the data, originally in CSV (comma-separated values) format, was transformed into an XLSX format.

3.4 Descriptive Analysis

This phase focused on conducting an initial analysis of the data, including its graphical representation, prior to the modelling process. Descriptive analysis plays a crucial role in extracting important information in this case the average home goals and away goals scored by a team was calculated. The purpose was to gain insights into which features might be pivotal and should therefore be incorporated into the construction of the machine learning model.

3.5 Data Mining

The emphasis of this stage is on selecting and developing the appropriate model by employing machine learning algorithms. For this study, a simulation model was constructed using the first dataset, this model aimed to simulate the season-end standings of the leagues based on historical data.

Meanwhile, the second dataset was utilized to create a model designed to estimate players' values based on their attributes. A critical component of this phase also involved evaluating and fine-tuning the models to increase their accuracy, for KNN model it can be done by increasing or decreasing number of Nearest Naighbours and to set "weights" to either "Uniform" or "Distance", for Gradient Boosting, the learning rate can be modified and the number of estimator can be increased or decreased. This step is crucial for ensuring that the models are robust and reliable in their predictive capabilities.

3.6 Testing & Visualisation

The final step, as implied by its name, involves testing the developed models and visualizing the results, this stage is crucial for assessing the performance and effectiveness of the models. For the first model comparison of actual match outcomes vs predicted match outcomes was done, and models accuracy was measured using evaluation matrices.

For the second model the price of players was compared to the current original price and there was a comparison between predicted value of a single player in the three different leagues which was represented graphically using Matplotlib library.

The visualization of the results plays a significant role in interpreting the outcomes, making it easier to communicate findings and insights derived from the models.

4 Design Specification

The research is segmented into two distinct components as shown in Fig.2, the first segment focuses on simulating the end-of-season league standings, offering clubs valuable insights into their potential finishing positions within the league, derived from historical

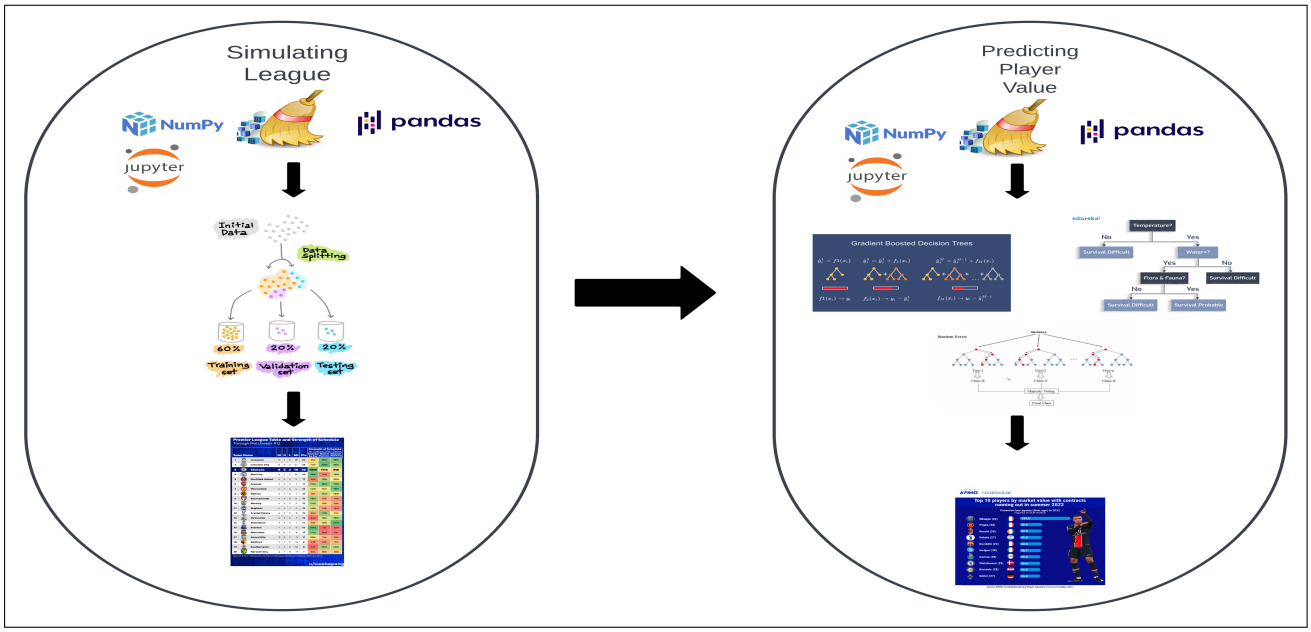

Figure 2: Desing Specification

data analysis. The second component is dedicated to estimating the values of players across various leagues, considering their individual attributes.

4.1 Simulating league standings at the end of season

To simulate the seasonal league standings, the initial method employed involved calculating the outcome of each match played by a team and determining the number of wins and losses using Random Poisson , followed by drafting a league standing. In this system, a team earns three points for every win, while a draw results in both teams receiving a single point, and no points are awarded for a loss. Recognizing that relying solely on one approach might not be sufficient, a KNN (K-Nearest Neighbours) model was also constructed using the same dataset to verify the results of the initial method. The effectiveness of this model was evaluated using metrics such as Precision, Accuracy, and F1 score. Apart from evaluating the result by the use of evaluation matrices it was also compared with real data.

4.2 Predicting Players Value

The second phase of the research involved estimating the value of players in each league based on their attributes. To achieve this, a series of models were developed using advanced algorithms Random Forest, Gradient Boosting, and Decision Tree. These algorithms were then subjected to evaluation using regression evaluation metrics, including Mean Average Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Accuracy. This approach allowed for a comprehensive analysis of player valuations, leveraging the strengths of various predictive modelling techniques.

5 Implementation

In this section in depth discussion of how the KDD steps were implemented and followed are discussed, by following these steps, it was possible to answer the research question.

	Date	HomeTeam	AwayTeam	FullTime	Halftime	HomeGoals	HomeGoalsHalftime	HomeShots	HomeShotsOnTarget	HomeCorners	...	HomeYellowCard
0	05/08/2022	Ein Frankfurt	Bayern Munich	A	A	1	0	8	2	5	...	
1	06/08/2022	Augsburg	Freiburg	A	D	0	0	10	2	5	...	
2	06/08/2022	Bochum	Mainz	A	D	1	1	16	3	3	...	
3	06/08/2022	M'gladbach	Hoffenheim	H	D	3	1	18	8	4	...	
4	06/08/2022	Union Berlin	Hertha	H	H	3	1	18	8	5	...	

Figure 3: Data-sets 1

	Name	Age	Overall rating	Potential	Value	Wage	Total attacking	Finishing	Total skill	Dribbling	...	Total goalkeeping	GK Diving	GK Handling	GK Kicking	GK Positioning	GK Reflexes
0	V. Osimhen ST	24	88	91	€126.5M	€120K	394	91	338	83	...	53	14	14	10	9	6
1	L. Martínez ST	25	87	89	€101.5M	€150K	400	90	375	84	...	48	11	8	8	8	13
2	M. Maignan GK	27	87	90	€78M	€90K	131	15	186	31	...	428	85	82	87	85	89
3	N. Barella CM	26	86	88	€83.5M	€135K	376	78	397	81	...	54	11	5	13	14	11
4	W. Szczęsny GK	33	86	86	€22M	€125K	86	12	99	11	...	415	86	83	73	86	87

Figure 4: Data-sets 2

5.1 Data Acquisition

Two different data sets were required for this research, the first data was taken from Football-Data[5] Fig.3 , it comprised of past matches of football teams in different leagues. Three leagues were selected for this study Serie A, Premier League, and Bundesliga. So last 6 seasons (2016-2022) data was taken from this open source. The data contained:

- Date of Match
- Home Team
- Away Team
- Full Time
- Halftime
- Home Goals
- Home Goals Halftime
- Home Shots
- Home Shots On Target
- Home Corners
- Home Fouls
- Home Yellow Cards
- Home Red Cards
- Away Goals
- Away Goals
- Halftime
- Away Shots

[5] `https://www.football-data.co.uk/englandm.php`

- Away Shots On Target
- Away Corners
- Away Fouls
- Away Yellow Cards
- Away Red Cards
- Value
- Wage

The second dataset was taken from the online open source Sofifa [6], data of players playing in the chosen leagues was taken. The data contained players attribute and value as shown in Fig.3.

5.2 Data Preprocessing

Cleaning data and formatting it was essential before moving forward, both the dataset was first formatted into table using Excel Power query which is a feature of Microsoft Excel which automate the process of data transformation. The first dataset did not required cleaning as it accurately contained matches data but to be safe cleaning steps were implemented:

- Checked null values, which in this case they were not present.
- Checked if any value was missing, but the data had no missing values.
- Checked duplicate values, which were not present in the dataset.

The second dataset required cleaning and feature engineering, the dataset was unstructured and to begin with it was formatted into a table and following steps were performed to make it ready for further analysis:

- Checked null values, missing, values, duplicate values, which were not found in this dataset.
- The column Team Contract was removed.
- The column height, weight, value, and wage was converted from general to numerical format.
- The column foot was converted into binary categorical from, Right foot was represented by 0 and Left foot by 1.

5.3 Descriptive Analysis

Descriptive analysis provides initial insight of the data, what it can be expected from it, and it also aid in selecting important features in the dataset for model building. The descriptive analysis of the first dataset revealed that in last 6 years Bayern Munich has highest winning rate and scored the most goals in Bundesliga among all the teams playing the competition which mean they might not require a striker as their goal scoring record is very good, they can look to improve the midfield or the defence. In Serie A Juventus has the highest winning rate, but it is Napoli leading the goal scoring in the competition, which shows a concerning point for Juventus, they might look for a good

[6] `https://sofifa.com/`

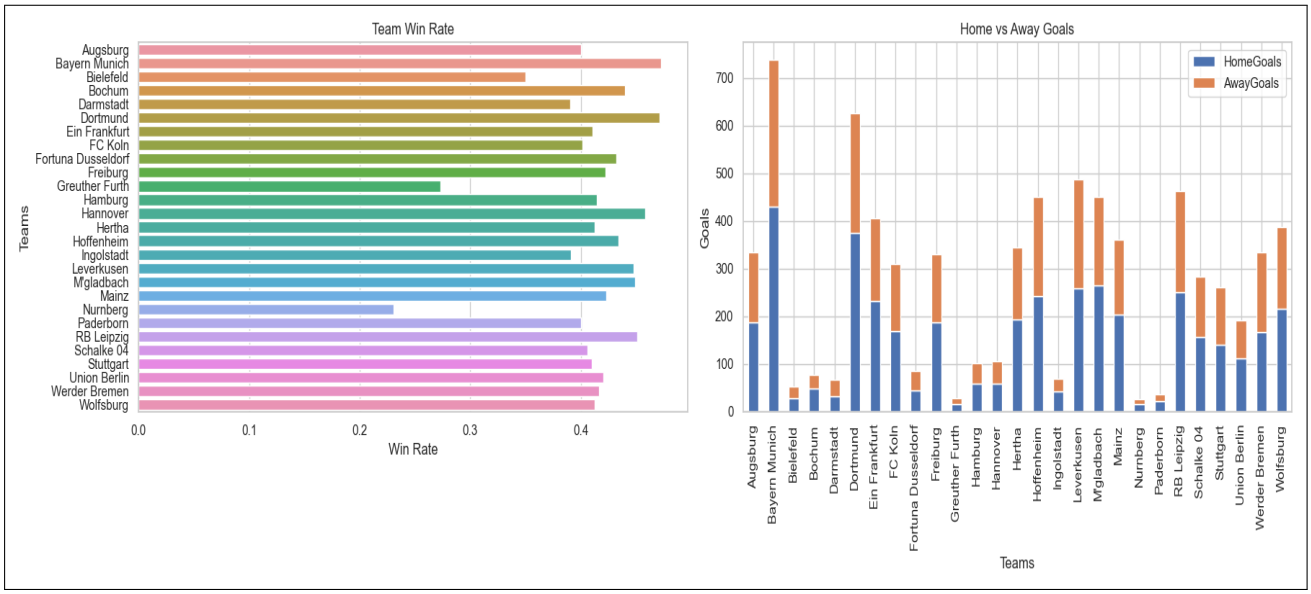

Figure 5: Descriptive Analysis Bundesliga

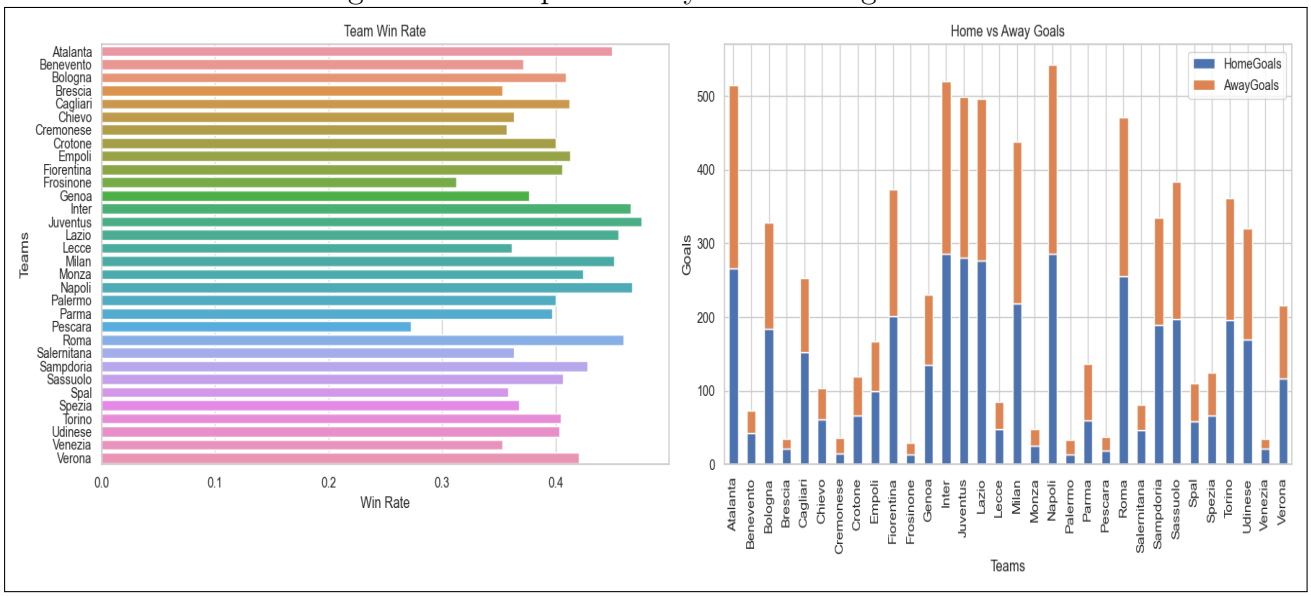

Figure 6: Descriptive Analysis Serie A

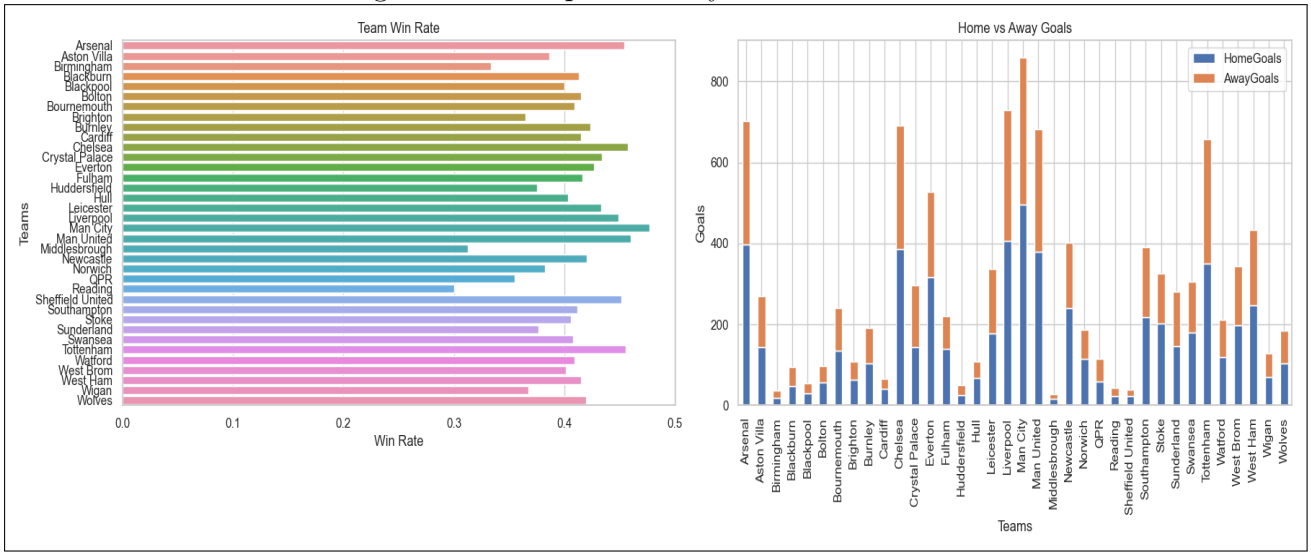

Figure 7: Descriptive Analysis Premier League

striker so they can improve their goal scoring record. In Premier League Manchester City is dominating having highest number of goals in the competition and winning rate as well which is similar case to the Bundesliga, and same for Manchester city they might look for improving defence or midfield. This information can be verified in Fig.7.

5.4 Model Building

	Age	Overall rating	Potential	Value	Wage	Total attacking	Finishing	Total skill	Dribbling	Ball control	...	Total goalkeeping	GK Diving	GK Handling	GK Kicking	(Positioni
0	0.523810	1.00	0.869565	0.888476	1.000000	1.000000	1.000000	0.948718	0.857143	0.961538	...	0.098086	0.071429	0.096386	0.091954	0.1395
1	0.476190	0.90	0.826087	0.654274	0.757576	0.868354	0.700000	0.991453	0.857143	0.935897	...	0.112440	0.071429	0.156627	0.045977	0.1511
2	0.904762	0.85	0.739130	0.066913	0.303030	0.200000	0.088889	0.279202	0.238095	0.435897	...	1.000000	0.976190	1.000000	1.000000	1.0000
3	0.333333	0.85	0.869565	0.605947	0.357576	0.101266	0.055556	0.128205	0.166667	0.282051	...	0.966507	1.000000	0.987952	0.804598	0.9651
4	0.238095	0.80	0.826087	0.617100	0.478788	0.673418	0.500000	0.706553	0.654762	0.782051	...	0.102871	0.119048	0.108434	0.091954	0.1162

Figure 8: Standarized Data

Two different models were built, the first model was responsible to simulate the seasonal league standings, and the second estimated the value of the players based on their attributes, but before building the model the data was standardized the result can be seen in the Fig.8.

Model Simulating League Rank

The simulation model was firstly built using Random Poisson, a function which simulated a match outcome between 2 teams was created and by the help of Random Poisson, away goals and home goals were generated, based on average goals scored by the team at home matches and away matches respectively, then it was decide which was the winning team(team scoring more goals). There were some advantage factors, for example a team playing at home ground had more advantage over a team playing away, also if a team scored a good number of goals in past, it was expected considering having the same team it will keep scoring almost at the same level in future as well. The use of Random Poisson was made because of the statistical nature of the data and because of its nature of forecasting probabilities which in this scenario is the scoring of goals.

To verify the results from the Random Poisson algorithm a KNN algorithm was implemented, and evaluated using evaluation matrices such as Precision, Recall, F1-Score and also the results of forecasted matches was compared to past real matches outcome. The KNN pipe line was created using preprocessor and 5 neighbours, which provided the best result compared to when number of neighbours was 3, and 4. The reason for using KNN was because it can make predictions without the use of an testing data and can simulate predictions. It works on the principle of similarity so if a team is performing good and scoring good amount of goals KNN will keep the performance of the team constant for future simulation. Considering that all team players performs at the same level.

Model Predicting Players Price

The second model aim was to estimate players value, and for these 3 major algorithms were implemented to identify which among these was the best suitable for these specific predictions Decision Tree, Random Forest, and Gradient Boosting. The reason for choosing these 3 algorithms was that they can handle the complex, non-linear relationships often found in sports data, and provide insights into which features are most predictive of a player's market value. The choice of multiple models also allows for a comparison of different approaches, ensuring a more comprehensive analysis.

6 Evaluation

The evaluation step is important to make sure the model is good. It's not just about using evaluation tools, but also about doing tests and making predictions and discussing the result.

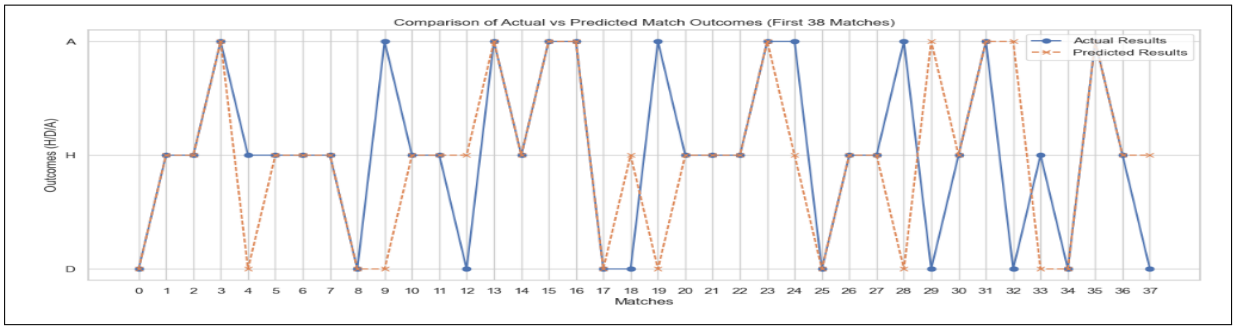

Figure 9: Actual vs predicted match Bundesliga

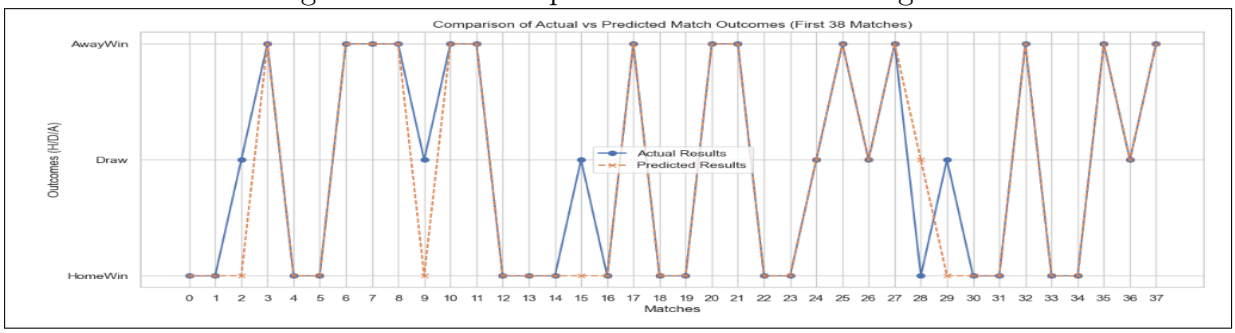

Figure 10: Actual vs predicted match Premier League

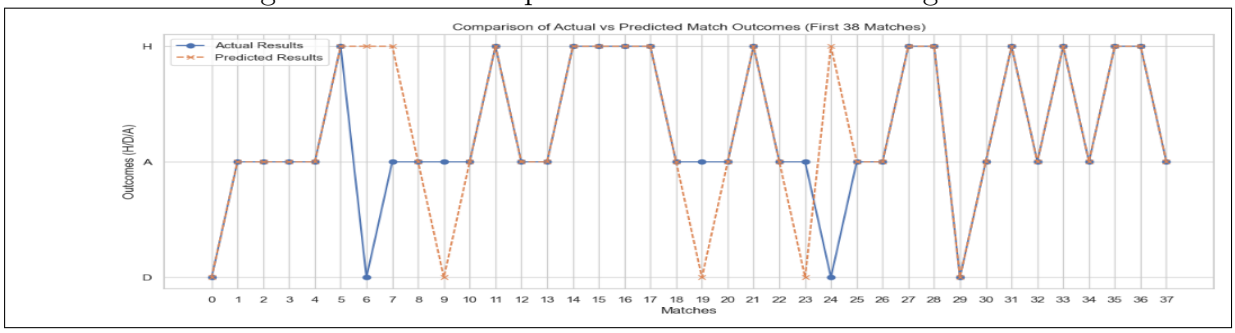

Figure 11: Actual vs predicted match Serie A

6.1 League Ranking Simulation Model

To evaluate the outcome achieved by Random Poisson, a KNN model was built to bring comparison between the two models, and the result obtained were very similar with some exceptions, KNN was evaluated based on Precision, F1-Score, Accuracy, and Recall. Apart from relying on evaluation matrices, a comparison of actual outcome of matches vs original match result was done, it was found that out of 38 matches played by a team during a complete league season, the model predicted 30 matches accurately, and the results can be seen in Fig.9 for Bundesliga,Fig.10 for Premier League, and Fig.11 for Serie A Based on three different data sets of average accuracy of 79%, 76%, and 79% respectively which can be considered a good result, the F1 score were 73%, 75%, and 74% in the same order. The Recall value averaged around 73% for every model. Based on the results of the model and comparison with the actual matches vs predicted matches it was identified the model can predict wins and losses more accurately than the draws.

	Model	Dataset	MAE	MSE	RMSE	R2
0	Random Forest	Bundesliga	0.023717	0.004349	0.065949	0.821077
1	Gradient Boosting	Bundesliga	0.015574	0.002755	0.052485	0.886676
2	Decision Tree	Bundesliga	0.036059	0.009073	0.095251	0.626765
3	Random Forest	Premier League	0.013836	0.000815	0.028547	0.929654
4	Gradient Boosting	Premier League	0.010825	0.000396	0.019912	0.965777
5	Decision Tree	Premier League	0.022936	0.001824	0.042704	0.842584
6	Random Forest	Serie A	0.014082	0.000818	0.028604	0.960273
7	Gradient Boosting	Serie A	0.011484	0.000518	0.022762	0.974842
8	Decision Tree	Serie A	0.030314	0.004039	0.063552	0.803891

Figure 12: Evaluation of models to forecast players value

6.2 Players Value Predictive Model

In the second model, 3 different algorithms Decision Tree, Gradient Boosting, and Random Forest were implemented to estimate the value of players in the three different European Leagues. To evaluate these regression-based algorithms, evaluation matrices such as MAE, MSE, RMSE and Accuracy were calculated as shown in Fig.12. The Decision Tree was the least performing algorithm among the three it achieved only 62% accuracy for the players playing in Bundesliga and 84% and 80% for premier league and Serie A respectively. The Random Forest algorithm which is an enhanced version of Decision Tree was successful in securing better results than the previous algorithm, it achieved 82%, 92%, and 96% accuracy respectively and compared to the previous algorithm the MAE was also lower 23%, 13% and 4% only. Now the best performing algorithm was the Gradient Boosting algorithm which achieved and excellent accuracy values of 88%, 96% and 97% and MAE values of 1.5%, 1% and 1% respectively which make this the best among the 3. Comparing these results with the past studies such as Nazari and Azari (2021),Majewski (2021), and Van den Berg (2011) the model achieved a satisfactory accuracy, which is not limited, with more accurate data and improved modelling it can be increased further.

	Team	Points	Goals For	Goals Against	Goal Difference
0	Bayern Munich	115	153	64	89
1	Dortmund	106	120	80	40
2	Leverkusen	90	86	58	28
3	Hoffenheim	87	90	74	16
4	Wolfsburg	84	78	60	18

	Matches	Wins	Draws	Losses	Points
Bayern Munich	38	27	6	5	87
Dortmund	38	23	6	9	75
Freiburg	38	22	8	8	74
RB Leipzig	38	22	6	10	72
Leverkusen	38	18	9	11	63

	Team	Points	Goals For	Goals Against	Goal Difference
0	Juventus	125	124	59	65
1	Inter	125	133	78	55
2	Napoli	120	124	68	56
3	Atalanta	112	119	85	34
4	Lazio	111	113	85	28

	Matches	Wins	Draws	Losses	Points
Napoli	38	27	9	2	90
Inter	38	24	6	8	78
Milan	38	23	6	9	75
Juventus	38	21	10	7	73
Lazio	38	22	6	10	72

	Team	Points	Goals For	Goals Against	Goal Difference
0	Man City	144	168	79	89
1	Liverpool	133	136	89	47
2	Man United	128	116	75	41
3	Bolton	119	86	63	23
4	Southampton	117	98	75	23

	Matches	Wins	Draws	Losses	Points
Man United	38	26	7	5	85
Chelsea	38	26	4	8	82
Southampton	38	24	7	7	79
Man City	38	20	13	5	73
Arsenal	38	22	6	10	72

Figure 13: Simulated League Rankings

6.3 Discussion

The initial model, utilizing Random Poisson and validated by KNN, projected the seasonal rankings of football clubs based on historical data. This approach helps clubs anticipate their potential league standings, if the predicted league rank is not satisfactory the clubs will have to investigate what did they lack. In the Bundesliga, Bayern Munich led with the highest goal count, securing the top position in the standings. In contrast, for Serie A, despite Napoli scoring the most goals, the Random Poisson model projected Juventus as the likely league winner, but with a close point gap of only five. However, the KNN model contradicted this by favoring Napoli as the champions. The Premier League witnessed a similar pattern. Manchester City, with the highest goal tally, was predicted by the Random Poisson model to dominate the league. Contrarily, the KNN model suggested a different outcome, placing Manchester City in fourth and anticipating Manchester United to lead the league as shown in Fig.13.

Once identified the target player, the second model employing three algorithms with Gradient boosting as the most effective, achieving an average accuracy of 90%, was designed to estimate player values across these leagues. For example, striker Harry Kane's value was estimated at 88M in the Bundesliga, 75M in the Premier League, and 81M in Serie A. The second example was Victor Osimhen, a striker currently for Napoli football club, the values forcasted for him were 78.5M IN Bundesliga, 80.93M in Premier League and almost 100M in Serie A as shown in Fig.14.

The modelling is limited using historical data which can vary from real time data generated in matches and players attributes changes in every match depending if they perform good or not.

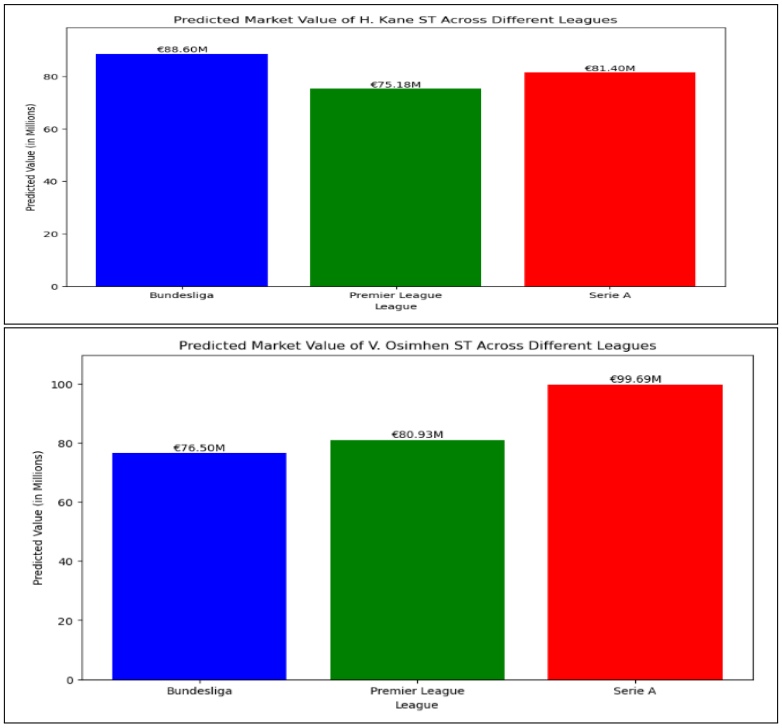

Figure 14: Predicted Players Value

7 Conclusion and Future Work

The study focused on developing strategies for the transfer window, enabling the acquisition of players at reasonable prices and preventing overspending on individual. This model can be a tool of reference it will never replace human knowledge, and can be used whether they aim to sign a single high-value player or multiple players by considering alternative options.

The first model, combining Random Poisson and KNN algorithms, was effectively used to forecast seasonal rankings, on comparing the forecasted match data with the actual match result, the model predicted correctly 30 matches out of 38. The second model employed Gradient Boosting to predict the value of players across different leagues. The player taken for example here was Harry Kane, previously, who was estimated to be valued 88.6m in bundesliga, 75.18m in Premier League, and 81.40M in Serie A. The actual price paid by Bayern Munich(Club playing in Bundesliga) to sign him was 86M as stated by Skysport[7]. These models highlight the significant role that machine learning can play in enhancing decision-making processes within the football industry. The scope of this study can be expanded beyond merely predicting player values. Expanding the model and adding a feature to recommend players of comparable skill and potential in scenarios where a club is unable to sign a high-valued, talented player. This would offer clubs alternative options, ensuring they can still acquire proficient players who fit their budget and strategic needs.

[7] `https://shorturl.at/dlsG5`

8 Acknowledgement

I extend my intenseappreciation to my supervisor, Dr. Athanasios Staikopoulos, for his invaluable contribution to this work. His vast knowledge, consistent motivation, and enduring patience have been key throughout my research and the composition of this thesis. I am also deeply grateful to my family, friends, and colleagues for their unwavering support and encouragement during my academic journey.

References

AKTUĞ, Z. B., Serkan, İ., Hasan, A. and KILIÇ, F. (2022). The estimation of german football league (bundesliga) team ranking via artificial neural network model, *Turkish Journal of Sport and Exercise* **24**(1): 22–29.

Al-Asadi, M. A. and Tasdemır, S. (2022). Predict the value of football players using fifa video game data and machine learning techniques, *IEEE Access* **10**: 22631–22645.

Baboota, R. and Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for english premier league, *International Journal of Forecasting* **35**(2): 741–755.

Corona, F., Forrest, D., Tena, J. d. D. and Wiper, M. (2019). Bayesian forecasting of uefa champions league under alternative seeding regimes, *International Journal of Forecasting* **35**(2): 722–732.

Dieles, T. (n.d.). Identifying successful football teams in the european transfer market: a network science approach.

Hoey, S., Peeters, T. and Principe, F. (2021). The transfer system in european football: A pro-competitive no-poaching agreement?, *International journal of industrial organization* **75**: 102695.

Joseph, A., Fenton, N. E. and Neil, M. (2006). Predicting football results using bayesian nets and other machine learning techniques, *Knowledge-Based Systems* **19**(7): 544–553.

Majewski, S. (2021). Football players' brand as a factor in performance rights valuation, *Journal of Physical Education and Sport* **21**(4): 1751–1760.

Munđar, D. and Šimić, D. (2016). Croatian first football league: teams' performance in the championship, *Croatian Review of Economic, Business and Social Statistics* **2**(1): 15–23.

Nazari, R. and Azari, S. (2021). Predicting market value of iranian football players using linear modeling techniques, *Research in Sport Management and Marketing* **2**(1): 41–53.

Ren, Y. and Susnjak, T. (2022). Predicting football match outcomes with explainable machine learning and the kelly index, *arXiv preprint arXiv:2211.15734* .

Rodrigues, F. and Pinto, Â. (2022). Prediction of football match results with machine learning, *Procedia Computer Science* **204**: 463–470.

Rossetti, G. and Caproni, V. (2016). Football market strategies: Think locally, trade globally, *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, IEEE, pp. 152–159.

Van den Berg, E. (2011). The valuation of human capital in the football player transfer market, *Rotterdam: ErasmusUniversity* .

Wand, T. (2022). Analysis of the football transfer market network, *Journal of Statistical Physics* **187**(3): 27.

Yang, R. (2019). Using supervised learning to predict english premier league match results from starting line-up player data.