National College of Ireland

Prediction of Player Performance for IPL and analysing the attributes involved, using Explainable AI

MSc Research Project
Data Analytics

Ayushi Bajaj
Student ID: x20242638

School of Computing
National College of Ireland

Supervisor: Aaloka Anant

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Ayushi Bajaj
Student ID:	x20242638
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Aaloka Anant
Submission Due Date:	15/12/2022
Project Title:	Prediction of Player Performance for IPL and analysing the attributes involved, using Explainable AI
Word Count:	3493
Page Count:	15

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:	
Date:	28th January 2023

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):	

Prediction of Player Performance for IPL and analysing the attributes involved, using Explainable AI

Ayushi Bajaj
x20242638

Abstract

Cricket is the second most popular game across the globe. It is played in numerous formats but the most fast paced format is called T20 which has twenty overs. With the increase in its popularity the game has is been played in T20 format in various leagues. The most popular among all these leagues is Indian Premier League. A lot of money is involved in this league hence player selection becomes an important part of the game. As scores of cricket has high numerical influence computational analysis proves to very beneficial for prediction related to it. This paper provides the findings for optimal player selection and identifying the impact of features using Explainable Artificial Intelligence.The player has two primary roles in the game batting and bowling the secondary role involve fielding and wicket keeping. The paper analyses every player's performance in terms of total points and regression models are used to predict these performances in IPL. The models used are Random Forest, Decision tree, SVM and XGBoost. Then these models are compared based on RMSE score where XGBoost provides the most promising results. Later hyperparameter optimization is applied on XGBoost model using optuna which provides even more accurate results. Lastly, Explainable AI is performed using SHAP library to understand the role of factors affecting the predictions.

Keywords- Cricket, IPL, XGBoost, Decision Tree, SVR, Random Forest,Explainable AI

1 Introduction

Sports and statistics are very interconnected. Though every sports has a different set of rules and regulations, they all have a criteria of keeping the score to determine the winning side. With the increase in technology various methods of broadcasting methods have been introduced making sports more popular than ever. This has resulted in exponential growth in the commercialization of sports and various leagues are formed based on this. The sports teams have experts and analyst working in background to make out the most from the team and find factors to increase the efficiency of the players. Earlier this analysis was done manually but now with increase in machine learning tools and techniques data modeling and analysis has become a very important part of the game.As machine learning is helping in identifying foreseen factors affecting the game as well as players performance.

2002 saw the introduction of the statistical analysis to the sports world. This was known as "The Moneyball Theory". Traditionally players were manually analysed by coaches and scouts but the computational analysis introduced by William Lamar Beane

has become a preferred technique in every sports. It all began with baseball when it was discovered that different baseball players with similar performances had very high pay gap. The selection procedure for baseball and other team sports was revolutionized by computational analysis of those athletes.

After football, cricket is the second most popular sport. It is played in several formats. The game was originated in England and later gained its popularity during the industrial revolution. Several colonized nations played the English-originated game. This game used to take days to complete one match, but over the past 20 years, the dynamics have changed. The game was played in either one-day format or test cricket format where it lacked the attention of audience because of its dreadfully slow speed. With the introduction of Twenty20 cricket, the number of overs was narrowed down to twenty, making the game drastically shorter in time and making it more entertaining.

The Board of Control for Cricket in India (BCCI) organized the Indian Premier League in response to the popularity of T20 cricket (IPL). Since its founding in 2008, it has grown to be the world's largest cricket league. It has eight teams, and comprises of national as well as international players. It generates a significant sum of money from ticket sales, broadcasting, and branding because of its quick pace and popularity. For aspiring players, this has been an excellent venue to showcase their abilities.

The paper by Kadapa (2013) discusses the absurd sum of money invested in IPL. The 140 million TV viewers and revenue from sponsors and companies are both included. These academic works highlight the importance of choosing the ideal player for a club commercially and the increased importance of statics and analysis.

Analysis has been a great part of sports for the past 20 years but the predictions were done by black box testing and often the factors considered were never clearly indicated towards the factors affecting those predictions. Hence for better understanding of these features explainable AI can be used. Explainable AI are a set of frameworks that help in comprehending and understanding the predictions made by machine learning models. SHAP (SHapley Additive exPlanations) is a library used in python to implement explainable AI. SHAP is cooperative game theory approach. Here all the features are treated as players of a game and every players individual performance is analysed to determine team performance. Hence, contribution of every feature used in ML Model is projected to provide a better understanding of prediction.

1.1 Research Objective

What is the best model to determine the player performance in IPL, and can we identify the impact of features affecting the player performance using Explainable AI?

This research project will determine the regression model that works best for predicting cricketer's performance based on external factors. It will then discuss the factors affecting the models and the impact they create on result using explainable AI. Then the project will further discuss the prediction of player performance in upcoming season of IPL.

Not just team selectors like coaches, scouts, and club owners will benefit from this study, the players will also be able to analyze and improve their inferior talents with its assistance. Cricket has always used performance prediction, but relatively few experiments have been conducted to use Explainable AI to produce those predictions. Decision-making may also be aided by the rise in popularity of fantasy cricket and the websites that allow fans to build their own teams in accordance with matches.

2 Literature Review

There has been a lot of study done on sports and machine learning. Nowadays, statistical analysis is used to inform the majority of cricket judgments. The research article that follows explains the needs and efforts made in cricket and other team sports. In paper by Gerrard (2007) the moneyball hypothesis is covered. The benchmark analysis on player quantification is covered. Statistics were used to evaluate the players based on their prior results. It also sheds insight on how complex it is to judge the players in team sports. The theoretical foundation and examples of its usage revolutionized player selecting processes.

Professional football player performance is studied by in this paper by Barot et al. (2020). The player performances were evaluated both individually and in relation to the team. Players in the top 10 European leagues were the subjects of the analysis. Attackers and midfielders were used to categorize the performances. The study is a textbook illustration of player analysis using two independent variables that emphasize both total goals and goal assists. Individual performances were taken into account, including accurate passes, shots on target, and dual victories. Player age, height, and weight were also included in the philosophy of player selection. A brand value was created by the player selection, according to further study of the player's popularity on Twitter.The machine learning model displayed positive results, providing the strongest arguments for the goal-assist team, midfielder, and even opponent statistics. For this investigation, the Stochastic Gradient Descent Regressor model had an RMSE of 1.054 for goals and assists. This information could be provided to the teams and players to improve performance and raise understanding of the best fit for the team based on the opposition.

In a paper by Noh et al. (2021)Machine learning technique is focused on. MotionNode sensors are used to recognize human activity. To analyze statistical features, they suggest SMO-based random forests model. Results indicate that the suggested model is capable of identifying reliable human activities.

2.1 Review on Prediction in Cricket

In paper by Shobana & Suguna (2021) the role of team prediction is also discussed. It backs up the entire case for player analysis in team sports and is crucial in both player selection and player placement decisions for the best results. For the match analysis, data from prior games including player credit, runs scored, delivery specifics, and money gained are used to help the user comprehend the game.

A variation was seen in the paper by Passi & Pandey (2018), here, predictions were made on the number of runs that will be scored as well as the bowlers' performance in each over. both are categorization issues. The runs are taken while maintaining the record in a given range. For predicting those two questions, naive bayes, random forest, multiclass SVM, and decision tree classifiers are utilized. Among the other models, Random Forest provided the most promising outcomes. In paper by Rodrigues et al. (2019*a*), they talk about the need to shuffle in players dependent on the weather, the location of the game, and the stadium itself. They talk about how each of these things can affect a player's performance in an indirect way.Kapadia et al. (2020) talk about the project where the result of the match is anticipated. For making predictions, the article employs Naive Bayes, Random Forest, K-Nearest Neighbour (KNN), and Model Trees. In this assignment, Random Forest demonstrated the highest level of accuracy.

Player classification and performances were analysed by Ishi, Patil & Patil (2022) in their paper. They talked about how this strategy can make the jobs of coaches and

skippers much simpler. Similar to prior research, the players' individual and team performances were taken into account when classifying them. The players were primarily categorized as batters, bowlers, all-rounders (batting and bowling), and wicket keepers. The study was conducted using an international format over one day. The goal of the study was to develop a fair strategy for selecting an eleven-person team. Based on batting and bowling statistics, player performances were evaluated in relation to the strengths and weaknesses of the team. The accuracy for the principal distribution of players using CS-PSO (Cuckoo Search and Particle Swarm Optimization) was 97.14, 97.04, 97.28, 97.29, and 92.63 percent.

In an article Rodrigues et al. (2019*b*) to lessen the complexity of the scope, they talk about the cricket matches between India and Sri Lanka. They attempt to provide answers to questions regarding the team's chances of winning, the final score of the game, the next over partnership, the lineup of batters, and the bowler who will break a batting partnership. For coaches and other sports professionals in particular, the project was beneficial. The project had a score prediction MSE of 1.41 and a winning team classification accuracy of 84.4. In order to make decisions during any cricket match, they highlighted the value of these predictions in their conclusion.

Beal et al. (2019) describes cricket's popularity around the world. There is a race to develop better management, service, and evaluation with such influence. In every industry, prediction is essential to provide these attributes. Match predictions are based on a variety of factors, including player performances, team strengths, locations, and weather. The research evaluates team strength using a variety of machine learning techniques, including Random Forest, SVM, Naive Bayes, Logistic Regression, and Decision Tree. This forecast can help the team comprehend the situation better and can also be helpful for gambling software and sports media.

Pramanik et al. (2022) in there paper particularly focuses on T20 cricket. They refer to this format as the most alluring type of cricket. In the past ten years, this format has grown in popularity and revenue. The prediction of the outcome of the game is another subject of this research. They used data from the Bangladesh Premier League (BPL) T20 competition for this investigation. This study supports the franchise while assisting investors in making a better investment. The technologists, sports analysts, and coaches stand to gain the most from such studies. They produced two data sets, one containing pre-match data and the other containing post-match data. K-Nearest Neighbor (KNN), one of five classifiers used in the investigation, fared better than the other four models.

Shakil et al. (2020) also discusses the prediction of match outcome and its importance in international cricket. From 2004 to 2018, cricket T20 and One Day International (ODI) matches were subjected to machine learning algorithms. ZeroR, Decision Tree, Random Forest, and XGBoost were the algorithms employed. It gave a good insight on how two different formats can have similar and contradictory outcomes during feature engineering.

In a paper by Ishi, Patil, Patil & Patil (2022) the outcome of ODI matches is forecast. For this approach, an ensemble algorithm utilizing voting and staking is developed. Additionally, logistic regression and SVM are carried out. The models were then contrasted based on recall value, F1-score, accuracy, and precision.

2.2 Review on Indian Premier League

In their paper Kansal et al. (2014) they describe how IPL's first team selection approach was entirely manual. The first teams were chosen at random, and the players' base salaries

were calculated. Based on players' performance in One-Day International (ODI) matches and T-20 data for both batting and bowling, the model developed here chooses players for those teams to be formed. Along with the player auction values, the model makes forecasts using a variety of variables. Even better, the model forecasts each player's value, which can aid in decision-making.

In a paper by Kalgotra et al. (2014) The prediction emphasizes on assembling a team that performs well while using the fewest resources possible. The players' prior performances were used as the source of the data for this study. The model produced a measure of each player's selection likelihood. The disadvantage of research based solely on IPL data is that there is little place for evaluation of new players, regardless of how effective they are in terms of performance.

In paper by Patel & Pandya (2019) in a fantasy cricket league, the key or forecast is addressed. Many people are taking part in it and profiting handsomely from it. By including people in selecting and bidding on their own squad, sports engage their fans more profoundly. The bowling statistics for the league from 2008 to 2018 are discussed in the publication. With the use of machine learning algorithms including Decision Tree, Random Forest, XGBoost, and Stacking, supervised learning is carried out in this project. The most effective technique for this method turned out to be stacking.

In their paper Das et al. (2021) discuss the modified hedonic model, which forecasts the cricket players' hedonic price equation. The value of the players and other aspects of design auctions are carefully assessed in this research. They have also used eXtreme Gradient Boosting-based models in addition to the Modified Hedonic model (XGBoost). The outcome of this study demonstrates that hedonic predicts with the fewest errors. In a paper by Pansare et al. (1900) They forecast their players' scores in a T20 game, which consists of two innings and 20 overs. Every inning consists of one team batting and one team bowling; the model developed for this project forecasts the first inning's score. For this first outcome, they applied the XGBoost model. The factors taken into account are run rate, lost wickets, match location, and batting team. The target runs for the batting team are then utilized, together with the same variables, to forecast the outcome of the second game. For the second model, they employed a logistic regression classifier. For the same data that was gathered from 2002 to 2014, a 5-over interval is produced. The research on cricket that has been done over many years. The majority of these studies used different machine learning models to predict match outcomes based on categorization. Several research are carried out for player selection. The research provides an approach to predict player's performance.

2.3 Explainable AI

There is very less related work found in explainable AI implemented for sports. In a paper by Wang et al. (2022), the match style and gameplay of the National Basketball Association(NBA) are analyzed using a clear artificial intelligence (AI) techniques. interpretability of the model is discussed in this paper. Results from experiments show how well the explainable AI algorithm analysis NBA results. Paper Lalwani et al. (2022) discusses the implementation of Explainable AI using SHAP libraries in volleyball. It explains how while predicting with DNN model there is explanation on how the selected feature affect the findings. It also focuses on how black box testing limits the finding of machine learning models as the in sports the features can play a better role than prediction.

The research on cricket that has been done over many years. The majority of these studies used different machine learning models to predict match outcomes based on categorization. Several research are carried out for player selection. The research provides an approach to predict player's performance. It also discusses an approach to find the key factors and their impact on those findings.

3 Methodology

The methodology for the experimentation part of the project is discussed in this section. The methodology selected is KDD. As shown in figure 1 all five steps of KDD are implemented. This methodology is suited for regression and prediction since it primarily concentrates on data mining rather than project management. In this section every aspect of KDD methodology is discussed.

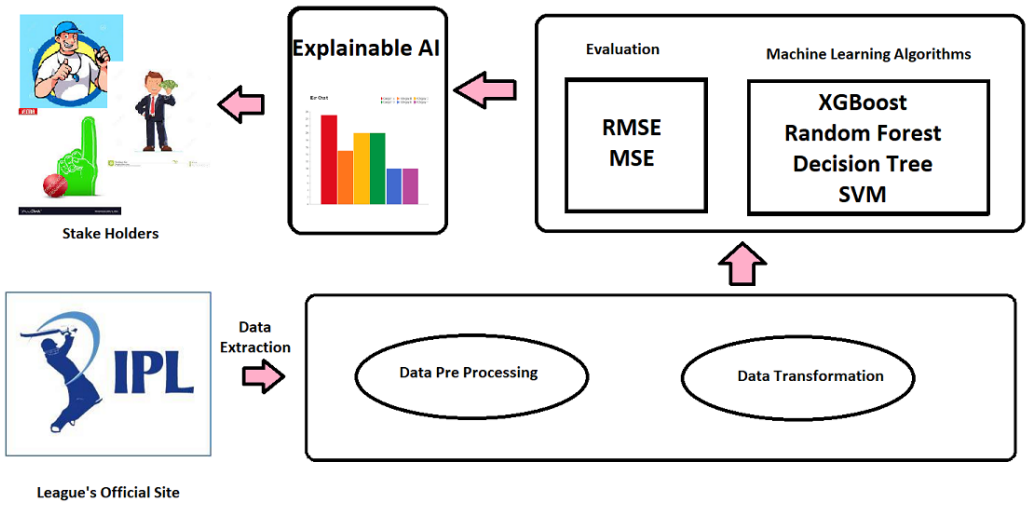

Figure 1: KDD Methodology

3.1 Data Selection

In this step the data relevant to the research work is extracted. The data selected in this process was found in a open source and provided a good insight related to player performances. Multiple data sets are selected for the research step from Kaggle [1]. These data sets include the players past performance, their team, details regarding matches, And cost and role of players.

3.2 Pre-Processing

After selection and extraction of data. The data is processed, and preliminary information is derived from it. The data from players past performance is collaborates into single scoring for a uniform understanding of player's performance in batting and bowling.

[1]https://www.kaggle.com/competitions/ipl-2020-player-performance

3.3 Data Transformation

As multiple data sets are used in this research these data sets are incorporated into one data frame. This task is done in this step. Multiple features of the data are in string and can be categorised for the model to run more properly. Later correlation is found between total points and other features to find most relevant features for modelling the data.

3.4 Data Mining

Once the data is processed and transformed it is used for data mining. The data is modeled to perform predictive analysis. Regression models are used to train 80 percent of data and is tested on 20 percent of data. The models used are:

- **Decision Tree Regression Model:** The graphical representation of a result and the decision that led to it is called a decision tree. It is presented in a sequential fashion, which facilitates visualization. It reveals the underlying connections between the features, which increases its predictive value and makes it more appropriate for usage with Explainable AI.
- **Random Forest Regression Model:** Large input parameters are no problem for random forest, and it does a good job of estimating the key variables. As a result, this model is chosen to serve as a comparative strategy for the above mentioned tree-based algorithm.
- **Support vector machine (SVM) Regression Model:** SVM is supervised modelling method that transforms the data using the kernel trick and determines the best class boundary on the basis of this transformation. SVM has the advantage of being able to capture intricate correlations between dataset variables.
- **XGBoost Regression Model:** Extreme Gradient Boosting is a considerably quicker and more sophisticated gradient boosting method that enhances model performance by using various regularization to lessen overfitting. It can be useful for increase the accuracy.

3.5 Evaluation

After the model has been trained, it is assessed using metrics like Mean Square Error (MSE) and Root Mean Square Error (RMSE). Which model is best is determined by the values of these criteria. For a better understanding of the forecast, explainable AI is used to compare and analyze the components having the greatest impact after the model is complete.

4 Design Specification

The project is divided into a three-tier system. Figure 1 Shows how the KDD is performed in three parts in this research. The three tiers are created based on the techniques being performed on the data. Following are the descriptions of those three tiers.

4.1 Data Tier

With the data layer being where the data is processed and transformed after being retrieved from Kaggle. After data extractions the data is being analysed and transformed to connect multiple datasets. This includes removing outliers, interlinking datasets, and transforming the data features into numeric values hence preparing them for modelling.

4.2 Application Tier

The application tier is where the data is modeled and compared. Machine learning algorithms are being performed and compared in this layer. The models are then compared based on their RMSE and MSE scores.

4.3 Presentation Tier

Finally, the business layer where Explainable AI is performed, and the performance of players is predicted in the form of total score for upcoming matches that can help coach, sponsors, and other stakeholders to utilize in making decisions is created using the elements affecting the players' performance.

5 Implementation

The specifications for hardware and software are covered in this section in brief. In this part, the dataset is described in great depth. It addresses the analysis and processing of the data. Additionally, it explains how points are awarded to players and examines the criteria used to evaluate their performance. Additionally, other models that were considered for the prediction were evaluated. To evaluate the results and determine the most accurate model, various models are utilized. The implementation procedure is depicted in Figure 2.

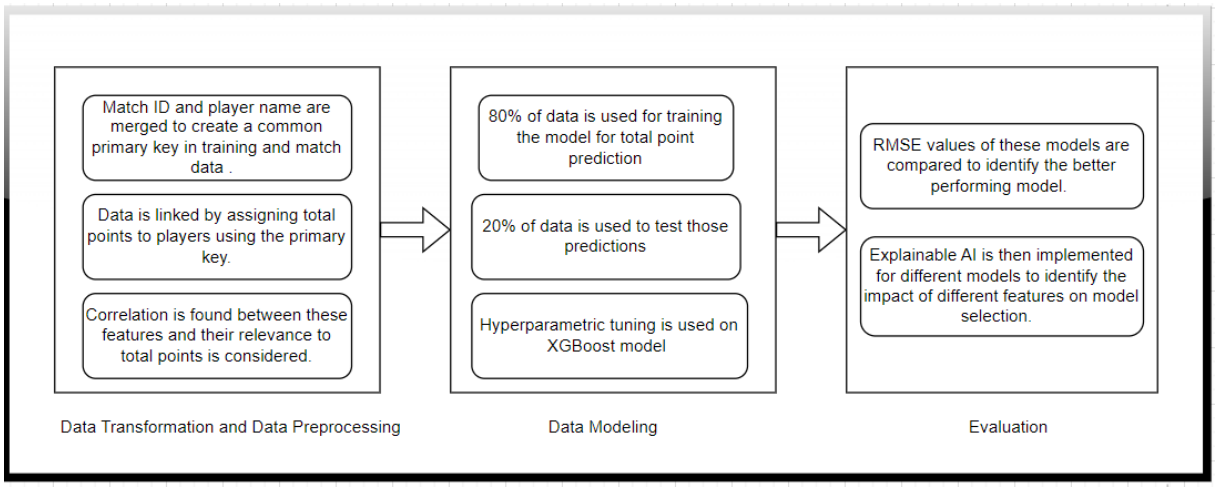

Figure 2: Implementation process

5.1 Hardware and Software Specifications

- **Hardware Specification:** Following hardware specifications are used in this research.
  Processor: Core i5 11th Gen
  Operating System: Windows 11
  RAM : 8.00 GB

- **Software Specification:** The research is implemented using Jupyter Notebook. Data cleaning, processing, analysing and modelling is done using python libraries. Explainable AI is also performed using python libraries.

5.2 Data Understanding

For the prediction of player's performance 5 data sets are used. These files contain the past performance of the player, match record from past five years, role and price of players from the past season, upcoming match schedule and team details. The game is divided in two parts batting and bowling which is reflected in the point system. Later both these points are added. Following formula was used by Barot et al. (2020) to analyse and score players and was mentioned in the data set provided by Kaggle.

- Batting Points = Runs scored by batsman in the match + Number of boundaries + 2 * Number of sixes + 8 {if the player scores a fifty} + 16 {if player scores a hundred} - 2 {if player gets dismissed without scoring a run}
- Bowling Points = 25 * Number of wickets taken by the bowler + 8 {if player did a 4 wicket haul} + 16 {if player did 5 wicket haul} + 8 * Number of maiden overs bowled
- Total Points = Bowling Points + Batting Points

5.3 Data Processing and Transformation

After extracting and loading the data various data components are checked as this process includes cleaning the raw data into making something more suitable for modelling. As machine learning applied in this research project the data in csv file has to be processed for better accuracy. Data consistency is checked and outliers are removed from then a preliminary analysis is performed on it. This preliminary analysis is mostly done on the past performance of players which is stored in data file named training.csv as it contains total points

- **Missing Values:** Since missing values in datasets are a key barrier to machine learning modeling, we have checked for them at this stage. There is a considerable likelihood of receiving incorrect results if missing values are kept in the data without being handled. Therefore, it is crucial to correctly handle missing values. All attributes had false returns after checking for them. Therefore, there are no null data in our dataset. Consequently, we shall continue with the next phase.

- **Data Collaboration:** A primary key is created using the match number and player ID to interlink all individual datasets.Then a dataframe is created merging all the data using the primary key. This data frame is then used for feature engineering as all the datasets are unified to include maximum features relevant in them.

- **Feature Engineering:** Some properties in the dataset, including city, date, umpire names, etc., are undesired. Retaining these extra properties could cause the runtime to slow down. Additionally, it will affect how well machine learning models work. Therefore, the final dataframe does not use these properties. The characteristics were then narrowed to six key features.

- **Data Encoding:** Data must be presented in a numerical format since machine learning models include mathematical computation. There is a concern that the machine learning model will produce inaccurate results if we feed it raw data. Thus, before feeding any machine learning model, input must be transformed into a numerical representation. Venue, season, player name, player's team, and rival team name are transformed into a numerical representation for this study.

In the next step this dataframe is used to test and train the model.

5.4 Data Mining

After data preparation the data is divided into two parts testing and training. The training part contains part 80 percent of the data and then the prediction has been tested on the test data which contains 20 percent of the data.

A wide range of literature reviews have been conducted in section 2 in order to construct machine learning models for forecasting a cricket player's performance. The most reliable machine learning algorithms, according to the researchers, are SVM, Random Forest, Decision Tree, and XGBoost. To find the most efficient model, the model RMSE value is checked.

SVM outperformed all other models. The results were improved by Shakil et al. (2020) in their paper by applying hyperparameter tuning on XGBoost and taking 30 percent as test and 70 percent as training data. Hence, hyperparameter tuning is applied on XBoost which then provided the best results.

6 Evaluation

This study offers two distinct types of evaluations of the model results. One assessment compares the models based on their RMSE scores. This is traditional approach in the field of regression models, it offers useful insight into which models make more accurate predictions. The second evaluation method will be explainable AI, in which the model's variables is analysed and their effects compared and analyzed.

6.1 Evaluation of Machine learning models

Performance evaluation in data mining is dependent on the kind of issue answered and the outcomes. The validation method used in this study was therefore selected to support both the research objective and the related work that is reviewed. The literature review, as well as the widely used performance measures for regression model in this project, served as the foundation for the results and evaluations made in this study. When the RMSE number is zero, there was no mistake in the prediction, hence the model with least RMSE score is considered to be a better model. This section offers a thorough examination of the findings from various stages of the experiment that led to this study.

- **Decision Tree Regression Model:** Decision tree showed the best RMSE value for the training data, but the value was too high for test data. Hence there is a high gap between the test and training results.
- **Random Forest Regression Model:** Similar to Decision Tree there is huge gap in RMSE value for training and testing. Hence the predictions are not very reliable for random forest regression model.
- **Support vector machine (SVM) Regression Model:** SVM had RMSE score very similar for its training and testing data making the prediction model more reliable the ones prior to them.
- **XGBoost Regression Model:** XGBoost gave decent results later hyper parametric tuning was applied to it using optuna. The results and RMSE score was better than all other models.

6.2 Explainable AI

The second evaluation technique used is explainable AI. As the research is done on players using explainer() function from SHAP library. Machine learning algorithms' output and outcomes can now be understood and trusted by human users thanks to a set of procedures and techniques known as explainable artificial intelligence (XAI). An AI model, its anticipated effects, and potential biases are all described in terms of explainable AI. The entire calculating procedure is transformed into what is known as a "black box," which is difficult to understand. The data are used to generate these black box models. Furthermore, nobody, not even the engineers or data scientists who developed the algorithm, is able to comprehend how the AI algorithm came to a particular conclusion.

SHAP stands for SHapley Additive exPlanations. SHAP instances are the mathematical values representing the impact of a feature in prediction for one dataset. The average of these instances can then be used to calculate the SHAP score that provides the overall influence of a feature on the overall prediction.

Figure 3 compares the outcomes of the decision tree and XGBoost using bar graph. It was foreseen that the player's prior performance would be the greatest deciding point in performance prediction. The impact of other elements is quite different for both graphs, as external factors play a larger part in decision tree than in XGBoost.

Figure 3: Bar graphs representing factors affecting the model

In figure 4 the positive and negative aspects of impact is presented using bee swarm graph. A single dot is used to indicate each instance of the explanation in each aspect of the figure. Venue has a greater negative impact on a player's performance.

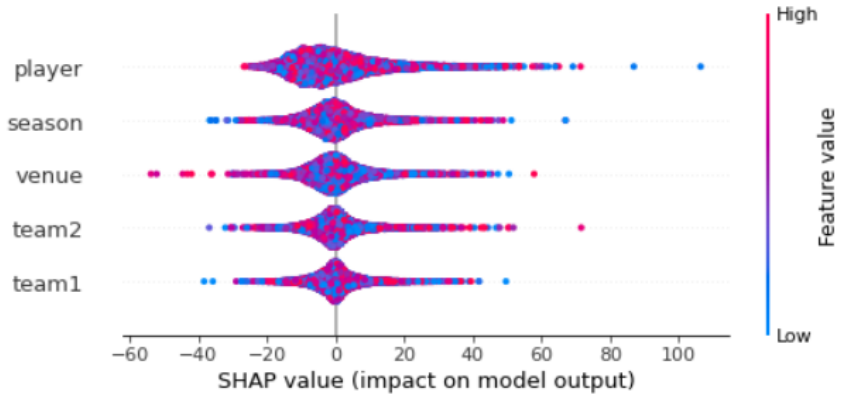

Figure 4: Beeswarm graph presenting the positive and negative impact for Decision Tree

In figure 5 the heatmap plot function generates a graph with SHAP instances on the x-axis, the model attributes on the y-axis, and the SHAP values are plotted on a color scale. By default, the samples are arranged in a hierarchical clustering according to the similarity of their explanations. As a result, samples that were grouped together because they had the same model attribute

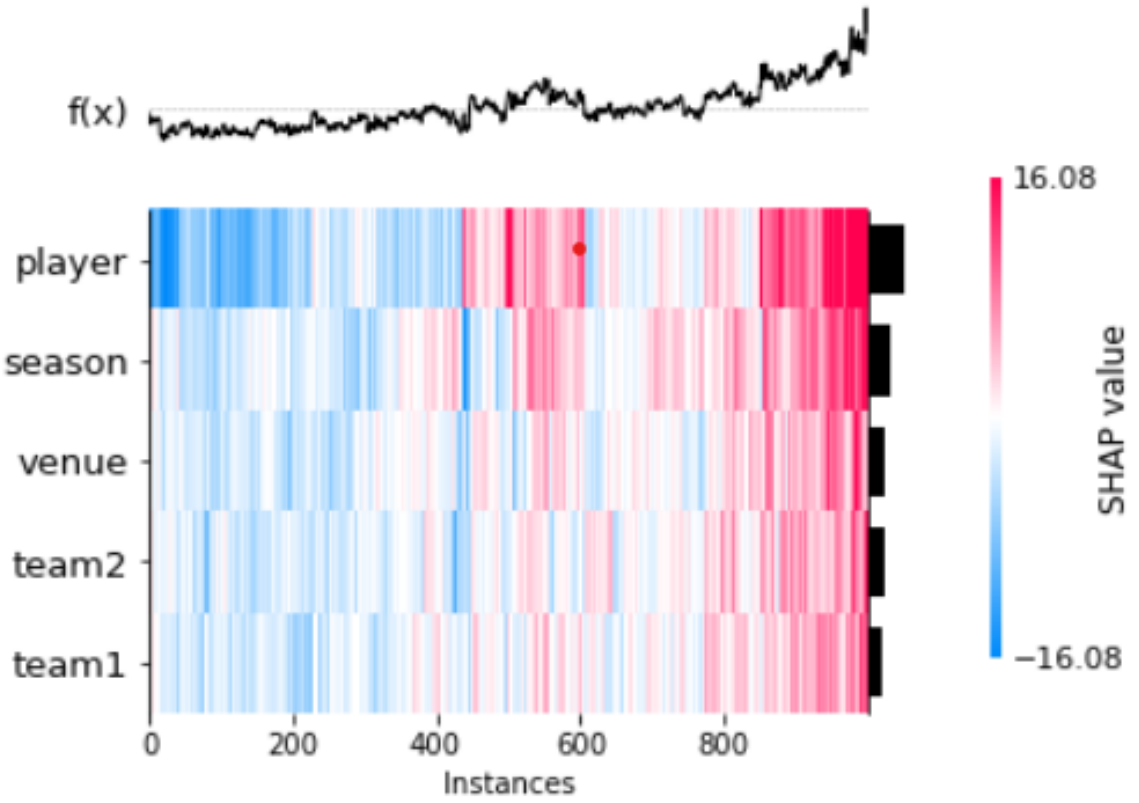

Figure 5: Heat Map

7 Discussion

The computational approach of scoring a sportsperson aid in assessing and comparing their performances. Leagues like these have a lot of commercial value and sports have a monetary worth. The selection of players becomes one of the most important factors as a result. Performance of a player is influenced by a number of variables. These variables can be evaluated and provide the player a clear sense of what to do with these external factors.

Since these players were evaluated after every match, the scores could be used to forecast how well they would do overall over the course of an entire season. Therefore, 2020 match information is used to anticipate the player's performance in those matches. Later average of those performances is calculated to compare and rank those players. Additionally, a data set on the player's price and role is included. The stakeholders may benefit from this execution, by picking players with similar roles and points in a more economical way. This provides a financial feasibility to the selection process as was done by Paul DePodesta and William James in the" Moneyball Theory".

player	role	Price in Dollar	
A Mishra	Bowler	873480.0	31.000000
A Zampa	Bowler	543453.0	39.666668
AB de Villiers	Wicketkeeper batsman	1600000.0	41.571430
AD Russell	Allrounder	908768.0	51.500000
AJ Finch	Top-order batsman	575323.0	32.769230

Name: Total Points, dtype: float32

Figure 6: Performance prediction for next season

Figure 6 shows the results after scoring the players for upcoming season. As we can the price difference between two players of similar role and points. Hence the decision making becomes easy for the selectors.

8 Conclusion and Future Work

At last it can be concluded that among all four regression models Random Forest regression worked the best after hyperparameter tuning.On applying explainable AI Random Forest regression gave more comprehensible model. As expected, the player's past performance has the greatest role in predicting their future performances but the external factors affecting these performance are the season, the team they are playing with and against as well as venue also play a vital role. Lastly, a comprehensible comparison can be drawn for player performance and their price.

For future work more factors like pitch type, popularity of the player etc. could be added to the model. Team selection can also be conducted using this research. This can be very beneficial for scouts and team owners as they can use the features as a blueprint along with their expertise to select a player.

Acknowledgement

It was an enlightening experience to research about the sport I've been following since

childhood. I am most grateful to my mentor Mr Aaloka Anant. He introduced me to concepts like explainable AI and helped in every step of the project.

I am also very grateful for help and support I've received from my peers throughout the semester. Lastly, I would like to thank my father for introducing me to the world of cricket from a very early age.

References

Barot, H., Kothari, A., Bide, P., Ahir, B. & Kankaria, R. (2020), Analysis and prediction for the indian premier league.

Beal, R., Norman, T. J. & Ramchurn, S. D. (2019), 'Artificial intelligence for team sports: a survey', *The Knowledge Engineering Review* **34**, e28.

Das, N. R., Priya, R., Mukherjee, I. & Paul, G. (2021), Modified hedonic based price prediction model for players in ipl auction.

Gerrard, B. (2007), 'Is the moneyball approach transferable to complex invasion team sports?'.

Ishi, M., Patil, D. J., Patil, D. N. & Patil, D. V. (2022), 'Winner prediction in one day international cricket matches using machine learning framework: An ensemble approach', *Indian Journal of Computer Science and Engineering* **13**, 628–641.

Ishi, M., Patil, J. & Patil, V. (2022), 'An efficient team prediction for one day international matches using a hybrid approach of cs-pso and machine learning algorithms', *Array* **14**.

Kadapa, S. (2013), 'How sustainable is the strategy of the indian premier league-ipl? a critical review of 10 key issues that impact the ipl strategy', *International Journal of Scientific and Research Publications* **3**.

Kalgotra, P., Sharda, R. & Chakraborty, G. (2014), 'Predictive modeling in sports leagues: An application in indian premier league', *SSRN Electronic Journal* .

Kansal, P., Kumar, P., Arya, H. & Methaila, A. (2014), Player valuation in indian premier league auction using data mining technique.

Kapadia, K., Abdel-Jaber, H., Thabtah, F. & Hadi, W. (2020), 'Sport analytics for cricket game results using machine learning: An experimental study', *Applied Computing and Informatics* (ahead-of-print).

Lalwani, A., Saraiya, A., Singh, A., Jain, A. & Dash, T. (2022), 'Machine learning in sports: A case study on using explainable models for predicting outcomes of volleyball matches'.

Noh, B., Youm, C., Goh, E., Lee, M., Park, H., Jeon, H. & Kim, O. Y. (2021), 'Xgboost based machine learning approach to predict the risk of fall in older adults using gait outcomes', *Scientific Reports* **11**.

Pansare, J., Khande, S., Oswal, A., Munsiff, Z., Choudhary, S. & Kumbhar, V. (1900), 'Cricket score prediction using xgboost regression'.
URL: *www.irjmets.com*

Passi, K. & Pandey, N. (2018), 'Increased prediction accuracy in the game of cricket using machine learning', *arXiv preprint arXiv:1804.04226* .

Patel, N. & Pandya, M. (2019), 'Ipl players performance prediction', *International Journal of Computer Sciences and Engineering* **7**, 478–481.

Pramanik, M. A., Suzan, M. M. H., Biswas, A. A., Rahman, M. Z. & Kalaiarasi, A. (2022), Performance analysis of classification algorithms for outcome prediction of t20 cricket tournament matches, pp. 1–7.

Rodrigues, N., Sequeira, N., Rodrigues, S. & Shrivastava, V. (2019*a*), Cricket squad analysis using multiple random forest regression, *in* '2019 1st International Conference on Advances in Information Technology (ICAIT)', pp. 104–108.

Rodrigues, N., Sequeira, N., Rodrigues, S. & Shrivastava, V. (2019*b*), Cricket squad analysis using multiple random forest regression, *in* '2019 1st International Conference on Advances in Information Technology (ICAIT)', IEEE, pp. 104–108.

Shakil, F. A., Abdullah, A. H., Momen, S. & Mohammed, N. (2020), Predicting the result of a cricket match by applying data mining techniques, Vol. 1295.

Shobana, G. & Suguna, M. (2021), 'Sports prediction based on random forest algorithm'.

Wang, Y., Liu, W. & Liu, X. (2022), 'Explainable ai techniques with application to nba gameplay prediction', *Neurocomputing* **483**, 59–71.