National College of Ireland

Predictive Analysis in T-20 Cricket: Estimation and Prediction of fantasy points for IPL Players

MSc Research Project
MSc in Data Analytics

Pratheek Gogate
Student ID: X22159789

School of Computing
National College of Ireland

Supervisor: Mayank Jain

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Pratheek Gogate
Student ID:	X22159789
Programme:	MSc in Data Analytics **Year:** 2023
Module:	MSc Research Project
Supervisor:	Mayank Jain
Submission Due Date:	14/12/2023
Project Title:	Predictive Analysis in T-20 Cricket: Estimation and Prediction of fantasy points for IPL Players
Word Count:	7940 **Page Count:** 24

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: Pratheek Gogate

Date: 14/12/2023

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):	

Predictive Analysis in T-20 Cricket: Estimation and Prediction of fantasy points for IPL Players

Pratheek Gogate
X22159789

Abstract

This research investigates the utilisation of machine learning (ML) and deep learning (DL) models to estimate and forecast the performance of specific cricket players participating in the Indian Premier League (IPL). The study focuses on the estimation of their fantasy points from the batting perspective only. The study initially utilized ML methodologies, including Support Vector Machines (SVM), Linear Regression. Even though we got a good result with high R^2 value and low root mean square error (RMSE) values there were some drawbacks such as it used to compare with single predicted data point for evaluation and it used to consider current match data as well to estimate the fantasy points. So, the research is transitioned towards rolling window technique with integration of the DL techniques. By exclusively utilising historical data, this approach effectively eradicated both the drawbacks of previous approach. Although the outcomes of this methodological shift were not preferable to those of the initial ML models, it was still deemed more suitable for the predictive task. Potential factors that may have contributed to the subpar performance of DL models in this scenario include inadequate data volume or the requirement for a more features.

Keywords: Machine Learning, IPL Data, Fantasy Point Prediction, Linear Regression, SVM, Player Performance, LSTM, BiLSTM.

1 Introduction

1.1 Background

With presence of the large amount of sport related data and the powerful machine learning (ML) and deep learning (DL) techniques, it is an ever more intricate task to forecast the performance of players in different sports. Predicting sports performance has become an ever more intricate task in the age of data-driven decision making. The emergence of ML and DL models has provided novel opportunities for the examination of athlete performance metrics, specifically within the domain of fantasy sports. The present study investigates the predictive capability of diverse computational and statistical models in order to forecast fantasy cricket points, with a specific emphasis on notable cricketers (Balbudhe, et al., 2022). This study compares the performance of sophisticated DL techniques with that of conventional ML algorithms in the domain of time series forecasting.

1.2 Importance

It is imperative to comprehend the predictive capacities of various models within the domain of sports analytics. Precise forecasts not only augment the overall fantasy sports experience but also furnish valuable insights that can be applied to team administration, assessment of player performance, and fan involvement. As the intricacy of the data at hand escalates, it becomes evident that more advanced analytical methodologies are required. By comparing the performance of multiple predictive models in a sports context, this study makes a scholarly contribution by providing guidance for future methodological decisions in the field of cricket. Especially in a tournament like Indian Premier League (IPL) performance of each player has a huge impact where stakes will be high and predicting fantasy points can be beneficial both for the team and for the fantasy leagues. Fantasy points are basically something which tells how well the player has performed in that match by looking at his match statistics and additional points will be awarded when he hit sixes or boundaries or when he completes any milestones, and this is just from the batting perspective.

1.3 Research Question and Objective

The objective of the investigation is to provide responses to the subsequent inquiries:
RQ1: How can we estimate and predict the fantasy points for players in IPL?
RQ2: What are the merits and demerits of each technique with respect to its performance and approach?
Research Aim: The purpose of this study is to evaluate the effectiveness of various models in forecasting fantasy points. By conducting a comparative analysis, this research can contribute to the understanding of differences between both the approaches.

1.4 Outline of the Research

In the section 2 of the paper there is related work section, this section contains previous works which have been submitted in this area. Then the section 3,4 and 5 contain what are the framework used, how did this research was achieved and what are the techniques used during the research. then the section 6 contains result, which tells what are the results this project gave and how does they answer the research question. Then section 7 is conclusion section which summarises what was done during the research and what are the results achieved along with future works.

2 Related Work

In (V.S., et al., 2020) their study they had aimed to investigate the factual segmentation that which of the two teams would prevail in IPL for the ranking players. The analysis of the existing models and the ratings depending on the mathematical methods are assessed depending on the straightforward equation for studying the limitations of the team. These models formulated the standard that they have suffered from lack of effectiveness in the prediction and ranking. Scrutiny is the nature of the ball hit in the field game of IPL history is analysed for assessing the solution. The solution for the types of features and the comprehensive data analysis on IPL prediction provided up to 81% accurate data that denotes

that deep predictor reflected 12% toward chances of winning for Royal Challengers Bangalore (RCB) found in extensive data analysis.

2.1 Analysis of current trends

As per the views of (Kapadia, et al., 2019), analysis of the tree-based model in particular the Random Forest (RF) approach is more appropriate in predicting the results of IPL matches than other types of ML techniques. The distinctive perspective of the collected data in filter-based technique accumulation provided an overview that creating a precise prediction model is essential rather than tossing a featured subset for selecting the feet of the teams. Inside edition applies a machine for the prediction of cricket matches based on the toss choices.

Examination of the sports analytics and the proposition of the data visualisation enhances the process of player selection by having a hold on the decision-making of the managers in the game of IPL (Kanungo & Bomatpalli, 2019). This work has investigated the selection process of the best players for the team in RCB with the aid of statistical analysis of the coin toss. This provides a wider aspect towards decision makers by providing aesthetically happening education and data for choosing players that might contribute to the improvement of the overall team performance and their success.

The comparative analysis of the pole doubt of 840 fans during 18 months for two teams of Chennai Super Kings (CSK) and RCB monitoring their inclusive data while accessing the digital media content of the team (Sagar & Sharma, 2022). The study is very dense regarding the loyalty of the fans with the perception of content consumption, player participation, and community involvement. It is essential to clarify the kind of engagement that will be useful for enhancing fan interaction activity throughout the digital media platform engaging in targeted engagement and marketing initiatives for the sports team.

As per the views of (Prakash & Verma, 2022), the deep clear performance indicator investigates the data of the 2019 IPL and constructs an elevated approach with the formulation of K-means clustering and RF technique for providing accurate and comprehensive assessment of players' batting and bowling ability in the earlier innings. This software is essential for fantasy cricket players, fans, and coaches as well as managers as it enhances the inside of the player's performance and their team strength.

In the paper (Chittibabu & Sundararaman, 2023), proposed that a staged process of K means clustering and base book assignment logic advise the initial steps for determination of the starting page of players. The empirical demonstration of this methodology has decreased the number of Indian participants by 17.6% and foreign players by 31.1%. This outcome is appropriate in suggesting substantial improvement in the IPL auction process benefiting players as well as beating clubs for enabling a streamlined approach for auction players.

During the research (Mohmmad, et al., 2020), included that ML methods, including Support Vector Machine (SVM) (Cortes, et al., 1995), Gaussian Fit-chime (GAU), and K-Nearest Neighbors (KNN) are applied in current training data sets for production of best classification results. Removal of unnecessary variables and increasing the performance and effectiveness of the data learning for grain prediction justify the prediction of game code techniques. The inclusion of linear regression (Box, 1976) and decision tree regression along with RF regression helps in adaptive posting algorithm formation for improving feature selection that in terms increases the accuracy of the player performance in real-time games.

This prediction will be reliable in addressing the monitoring features of RCB for improving their performance.

As per (Gokul & Malolan, 2023), creation of an optimisation model suggests a strategy for the likelihood of succeeding against the particular team. The methodology compares player's data across the white range of parameters for every opponent that determines the performance rating. The suggested method using historical data from 2008 to 2020 Discover model derived from this method and proposes that 70% equal to the real starting 11 across all the clouds in the league stage have a similarity. Despite that, only 7 3% of the team members differ in the team selection and advise a lining of the significant improvement that teams need to faster as per the analysis of the previous chord reflecting a value of 13.32%. This illustrates the effectiveness of the suggestive approach in developing a methodological selection for improving the performance of a team.

2.2 Evaluation of Selection process

In a paper (Pramoda, 2021), have presented that prediction of twenty-twenty (T20) cricket match results using ML techniques for monitoring the variations in their hybrid nature. The use of the RF method provided the most accurate result which is equivalent to 84.5% normal data and 76.61% for the normalized data. The hybrid third model outperforms the first two points 7 percent of the balanced data and unbalanced data for the creation of the decision tree approach. The inclusion of this hybrid model is successful in the prediction of the team winning Probability and it is beneficial for the team.

The paper by (Karkera, et al., 2020) incorporated that low-interest travel among the IPL team members influenced their performance. The interesting travel during the group rounds had a detrimental effect on the overall number of victories in the competition. This is established by looking at the code relation assessment between the team's cumulative travel distance and the quantity of home and away victories. The findings have highlighted the significance between the traveling factors in analyzing and success in IPL that offers a novel perspective towards the team members account in Planning and stargazing for future performance.

As per the views of (Bhatnagar & Batham, 2022), the prediction of the winning percentage and the inclusive standard of the Logistic regression algorithm rather than the other models provide comparatively low accuracy full stop the analysis of the recommended alert technique significantly outperforms the exhibiting algorithm and a provides outcome predictions with 80% accuracy in prediction of the winning percentage.

An analysis of the recent international and domestic performance of cricket players are examined for develop a clear idea formulation of the accurate model in a multicoated approach for monitoring the evaluated standards of their performance (Kumarapandiyan & Keerthiverman, 2020). The top players of the IPL T20 analysis are successful with this novel approach helping the team sports for detailed and reliable performance evaluation for future team development.

In the paper (Sarkar & Jana, 2020) they have included the relationship assessment between the captain's nationality and his team's success in the IPL in the previous 12 year span and has provided a clear understanding regarding the different table performance measured depending on the nationality of the team leader. The conclusion of this study has

guided the decision-making and strategy formulation that enhances the tournament's success by taking captain's nationality into account.

In (Gupta, 2022), included that the noble method for evaluating women's cricket batting performance in one-day internationals is only Guinea's best average scores and their batting strike monitoring. The inclusion of the principal component analysis has accounted for the consistency and strike race by monitoring the average score for assessment of the player's productivity. This method can be included in the IPL to enable fans, coaches, and management to make more informed decisions and judgments by providing a full picture of the statistical performance of players.

As per the views of (Sloane, 2020), the assessment of IPL T20 Matches, and the team batting first denotes the success rate of the team. The correlation between the total runs per total wickets lost and the total wickets taken during the match are related to the first player. All the attributes of the quoted number of wickets taken and number of wickets lost are positively correlated with the team's overall performance. Despite the higher success rate of teams batting first, the bowling statistics like the wicket taking and losing are the most crucial factors for winning the teams batting second.

As per the views of (Singh, et al., 2023), the determination of the effect of task and environmental constraints on cricket performance is assessed with the consideration of the studies related to pitch-type, pitch-length and the equipment on cricket performance. (Lutz, et al., 2019), strategically suggested that the inclusion criteria as per the K-Means clustering scores (Kmat) score ranging between 75% to 92% result in the demonstration of the environmental constraints such as pitch type constraint and equipment modification. The influence of the cricketer's batting session length and bowling practice has a high impact on the player's performance during the match.

In the analysis of the selective studies by monitoring the team performance from 1996 to 2019 with the inclusion of ML for predicting the match results has formulated adequate recommendations in terms of the benchmark data sales (Bunker & Susnjak, 2019). As per the views of (Skala & Zemková, 2022), features of the comparative discussion in terms of the accuracy performance and artificial neural networks for selecting the engineering perspective that derives the interdisciplinary collaborative measures are adequate in monetary approach of performance of certain teams.

As a part of research (Noorbhai, 2020), included that cricket coaching and batting when viewed through the lens of Fourth Industrial Revolution (4IR) with a reference to the performance of science and technology has provided an overview of the context of South Africa and recognition of the important technology in the end innovation in cricket has contributed to the industrial revolution. (Himagireesh, et al., 2023), strategically included that understanding the evolution of batting and factors that contributed to the successful batting approaches and leaving a gap of establishment between coaching manual coaching practices and skills for individual players are outlined for tangible example formulation in technology advancement in cricket coaching.

As per the views of (Naik, et al., 2022), the approach for the Graphics Processing Unit (GPU) best workstations and embedded platform in the sports vision analytics and the inclusive standards of artificial intelligence application has appropriate standards in recognition of the team's strategies and classifying their various events related to sports.

Application-specific tasks related to sports and the researcher's view regarding the team performance provide a check post approach in embedding their performance standard.

As a part of paper by (Noorbhai, 2020), incorporated that, the considered terms of the multi-criteria decision-making in this case for the tax's method have been utilized for weighing and performance factor analysis in terms of evaluating the performance measures for their selection of dream11 has improved the approach of team players.

In the paper proposed by (Balbudhe, et al., 2022) provided relevant information that performing a 13-minute color task is appropriate in assessing the effect of fatiguing exercise or mentally demanding tasks on human cognition and its impact on the decision-making and attention of the player selection has been provided. Hence, monitoring the physical attributes before the selection of the players should be adequate in reactive agility and sports-specific skill development.

3 Research Methodology

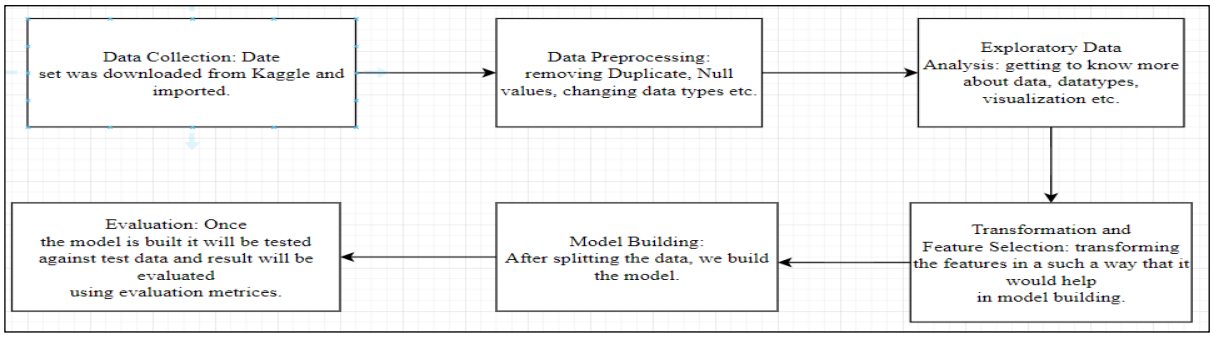

Figure 1: Flow diagram for estimation and prediction of fantasy points for players

3.1 Data Collection: In this research ball by ball dataset was used which was taken from Kaggle[1] and this data contains all the information on that particular ball during the match like who was on a strike, who bowled, what was the result of that ball, where was the match played etc and using this, data was filtered for particular players whom we want and then analyse this data**.**

3.2 Preprocessing and Exploratory Data Analysis of the Dataset: Once the data is fetched, data needs to be initially cleaned which includes eliminating duplicate values and take necessary actions when there are null values. In the figure 3 we can see that there were no Null values present in the data. Then some columns like boundary count, sixer count, half century count, century counts, total runs in that match were calculated and added to the dataset and thus making a player specific dataset and then Fantasy points are calculated for each match and that column will be added to the data for further analysis. Then data was visualised with the help of graphs to understand more about the data and get more information on the patterns of the data.

[1] https://www.kaggle.com/datasets/shrutisaxena/ipl-dataset-ball-by-ball-dataset/

Table 1:Number of Null Values for each feature

start date	0
Match id	0
venue	0
innings	0
bowling team	0
ball	0
Runs off bat	0
Boundary count	0
Sixer count	0
Half century count	0
Century count	0
Strike rate	0
Fantasy points	0

Visualisation: Visualisation helps a data analyst in many ways, it actually helps in finding more about the data and also helps to understand the patterns in the data and best thing about visualisations is that with the help of graphs and charts we can explain the stats to a person who does not even know much about that data set. Here are few graphs and charts from the dataset which would help to know more about the dataset.

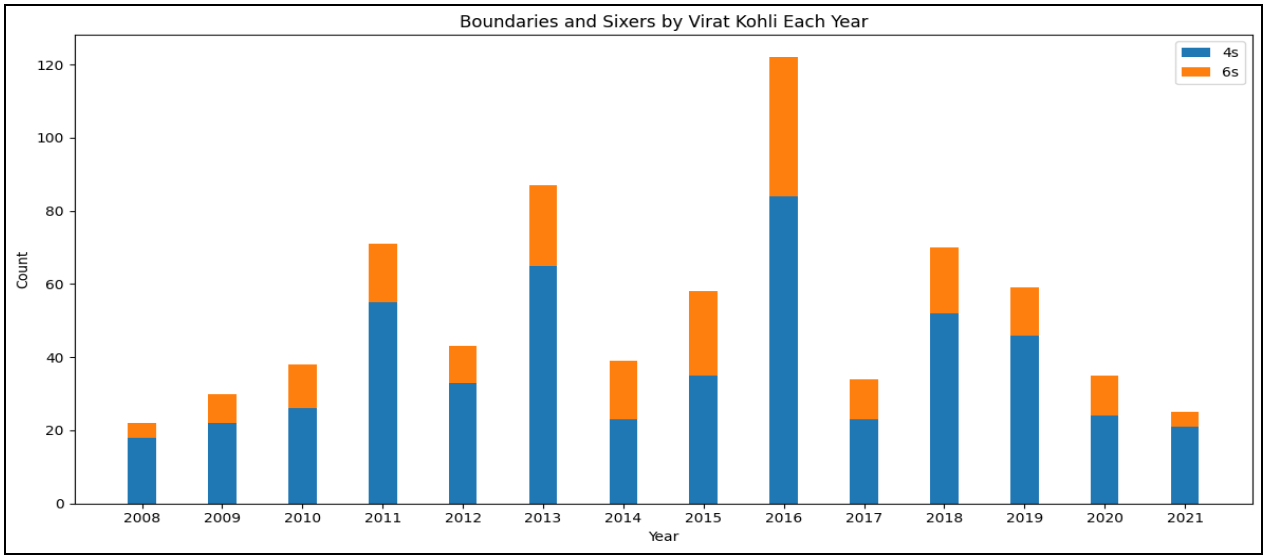

Figure 2: Number of Boundaries scored by Virat Kohli every year

Looking at figure 2 we can definitely say that 2016 is the year where Kohli was in his prime form as he hit almost 80 boundaries and 40 sixes. But in 2021 that number has been reduced by so much that he just scored around 20 boundaries and handful of sixes, even in 2008 the case was similar but since it was his careers starting, it can be ignore because he later just kept on improving.

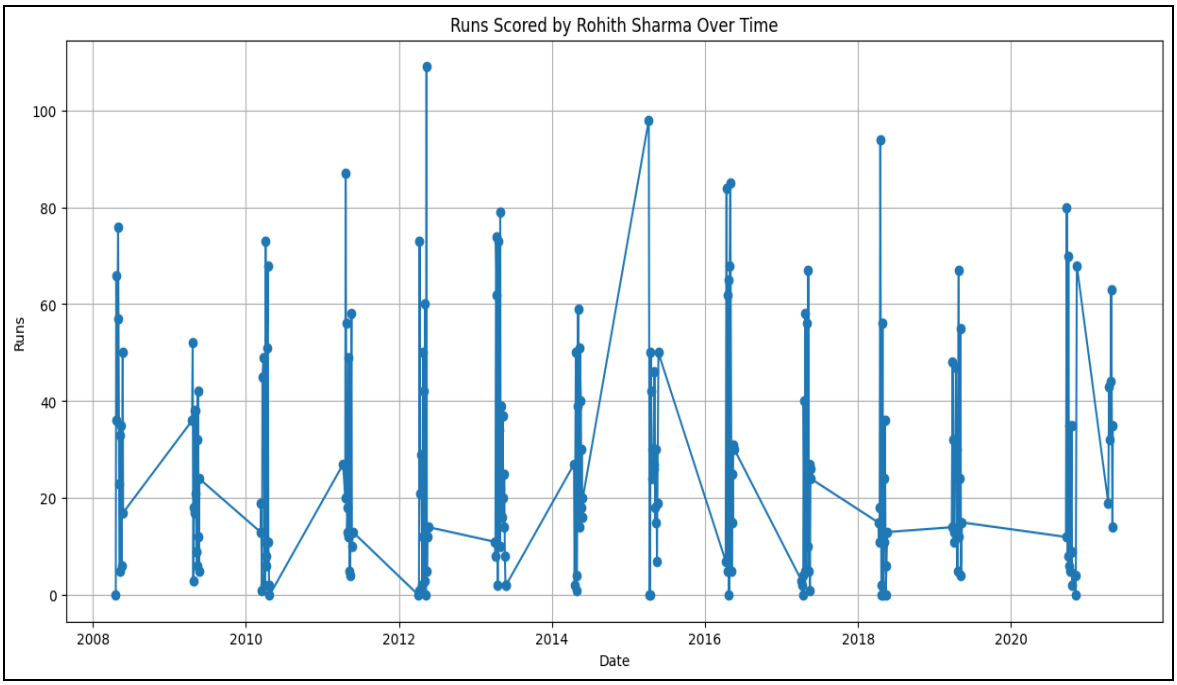

Figure 3: Runs scored by Rohith in each match

So while looking at figure 3, overall Rohith Sharma mostly scores less than 40 runs in a match and he usually scores 2-3 times more than 60 in a year and also he scored a century in 2012, so this are the inferences which can be drawn from this graph.

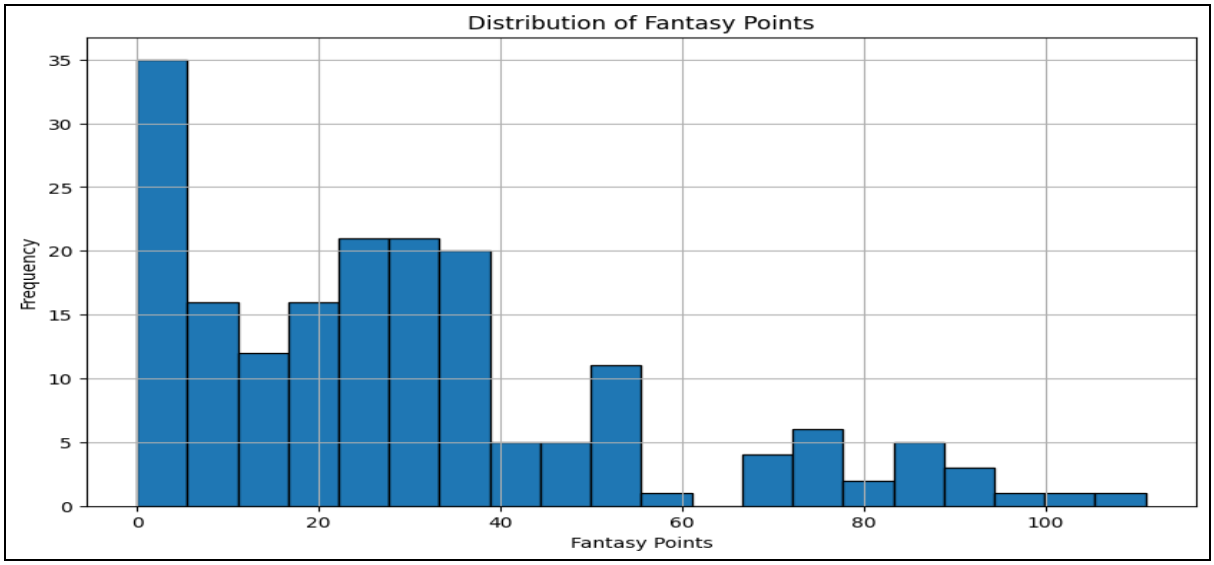

Figure 4: Fantasy points distribution of Dhoni in different frequencies

Looking at figure 4 we can tell is most often than not when he scores he will give you 20-40 fantasy points so in a way that he is a safe player and you can keep him in fantasy teams and anyways his wicket keeping will be a added bonus.

3.3 Transformation: There were some non-numerical features which had to be converted to numeric type so that it would be easy during the model building step and hence those were transformed into numeric values and for the prediction purpose time data had to be converted to input and for that purpose rolling window technique was used with taking 10 as the window size which means 1- 10 matches data will be

considered as input and prediction will be done for the 11th match and next 2-11 will be input and prediction is done for match 12 and so on.

3.4 Model Building: After doing preprocessing and all the necessary transformations we initially built model with Linear regression and SVM and then considering its drawbacks we also built deep learning models with the data we which we got using rolling window technique.

3.5 Evaluation: evaluation of the model was performed using R^2 value, Mean Squared Error (MSE) and Mean Absolute Error (MAE) where R^2 value should be in between 0 and 1 and value near to 1 and low MSE and MAE indicates it as a good fit and in other cases model is not considered as good fit.

4 Design Specification

The code for this project has been customized to utilize a Kaggle-sourced dataset in order to analyze Virat Kohli's, Rohith Sharma's, Mahendra Singh Dhoni's (MS Dhoni) cricket performances and generate fantasy points. The data preprocessing in the python-based implementation is carried out using pandas, with a specific emphasis on the performance data of Kohli, Rohith and Dhoni. The process of key feature engineering involves augmenting the dataset with additional metrics such as boundary and sixer counts, centuries, and half-centuries. To visualize performance trends for the purpose of insightful exploratory data analysis, Matplotlib is utilized. The predictive modelling component encompasses SVM and Linear Regression. There are also the DL models have been created to predict the fantasy points of the players. Scikit-Learn is utilized to facilitate the training and evaluation of models on divided data sets. Efficacy of a model is evaluated utilizing metrics such as R^2, Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE) in (Bhatnagar & Batham, 2022). In addition, Shapley Additive Explanations (SHAP) values are calculated to provide a more nuanced comprehension of the influence that features have on predictions. Developed with jupyter notebook compatibility in mind, this code provides an interactive framework that facilitates in-depth examination, rendering it an all-encompassing instrument for analyzing cricket performance and forecasting fantasy points.

4.1 Models

4.1.1 Linear Regression: It is a statistical method which is used to model a relationship between a target variable with one or more independent variables by fitting a linear equation to observed data. It assumes that there is a linear relationship between dependent and independent variable so that it can find best fitting line which has least difference in predicted and actual value. The goal here is to predict continuous numerical results based on input features on the basis of coefficient which basically tells the impact of each input variable on the target variable.

4.1.2 SVM: it is a supervised machine learning algorithm which is used both in classification and regression tasks. It works in a such a way that it finds the optimal

hyperplane or decision boundary which best separates the different classes or also predicts the continuous outcomes by maximizing the margin between the classes. SVM identifies support vectors, which are basically the data points which are located near to the decision boundary, to effectively categorize or predict new data based on their position relative to this boundary.

4.1.3 **LSTM (Long Short-Term Memory):** LSTM (Sanger, 1989**)** is a type of recurrent neural network (RNN) architecture which is particularly designed to handle the vanishing gradient problem in traditional RNNs. It is capable of learning long-term dependencies in sequential data one by one utilizing memory cells and different kind of gates to regulate the flow of information. here basically based on the output of one step the output is stored in the memory cells and then based on that resulting gates will allow the flow and it continues. LSTMs are very good in capturing and remembering patterns in sequences, making them popular for tasks such as speech recognition, language modeling, and time series prediction.

4.1.4 Bidirectional LSTM: A Bidirectional LSTM is basically an extension of LSTM which processes the input sequences both in forward and backward directions, allowing the model to capture information from past and future contexts at a same time. This bidirectional flow increases the understanding of sequential data, enabling the model to consider context from both directions and improve its performance in tasks that gets benefit from broader context of understanding, such as machine translation and sentiment analysis.

4.1.5 GRU: It is another type of recurrent neural network which is very similar to LSTM but with a more streamlined architecture. It also addresses the problem of vanishing gradient and accommodates learning long-range dependencies in sequential data. GRUs have less parameters compared to LSTMs, they use reset and update gates to control the flow of information. They are computationally less expensive and are used in tasks similar to LSTMs, such as natural language processing and time series analysis.

4.2 Evaluation Metrices:

4.2.1 R^2 (Coefficient of Determination): It is a statistical measure which represents the proportion of variance in the dependent variable (target) that is explained by the independent variables (features) in a regression model. It ranges from 0 to 1, where 1 indicates that the model will perfectly predict the target variable based on the features, while lower values indicate less accuracy in prediction. R^2 value helps assess how well the model fits the data. So practically if the value is near to 1 then it is said to be a good fit model and if the value is near to zero then the model is said to be not good.

4.2.2 MSE: It is a most commonly used metric measure used in calculating the average squared difference between the actual and predicted values in a regression problem. It basically calculates the average of the squared differences between predicted and

actual values. Larger errors are magnified more due to squaring, making MSE sensitive to outliers. Lower MSE values indicate better model performance.

4.2.3 MAE: It is almost same as MSE only difference is that here direction of the error is not considered and also since it does not square the error value it is less sensitive to outliers. The lower value of MAE indicates model performance is good.

4.3 Tools and Libraries

4.3.1 Programming Language:

During the research python was used as a programming language in this project as this has huge number of libraries which is a going to help especially in projects related to data. This research used pandas, numpy for data exploration purpose. Then Mat plot library was used for visualisation purpose and then Scikit-learn was used for model building and evaluation purpose.

4.3.2 Libraries

Scikit-learn: In this research for implementing the Linear regression, SVM model, LSTM, Bidirectional LSTM, GRU and in preprocessing steps, this library was used. And this also been used while evaluation of models.
Pandas: This was used for efficient data handling.
Matplotlib: This was used for creating visualizations to explore more about the data and to help in result interpretation.
NumPy: This was mainly used for numerical and array operations.

5 Implementation

This section of the report explains more about how the result was found starting from first step.

5.1 Importing Libraries and Dataset

Libraries

- pandas for data manipulation and analysis.
- numpy for numerical operations.

Data Import

The dataset was imported using pandas library, which is a versatile function that can handle different separators, quoting rules, etc.

5.2 Initial Data Exploration

Data Inspection

Tail function of pandas is used to look at the last five entries of the dataset to understand its structure and contents.

The columns attribute is used to examine the DataFrame's column headers, which include match details, ball-by-ball data, and various cricket match attributes.

5.3 Data Cleaning and Preprocessing

Data Cleaning

Initially checked for duplicate values and null values using “isnull” function and since the dataset did not have any such missing values it helped in the research to not to spend much time on it and rather address on the other parts of the project.

Filtering Data for a Specific Player

The dataset is filtered to include only the rows where 'V Kohli' is the 'striker'. This step is crucial for focusing the analysis on Kohli's batting performance. Similarly, it was done for the other 2 players. A group by operation followed by agg (aggregate) function is used to calculate the number of balls faced ('ball': 'count') and runs scored ('runs_off_bat': 'sum').The results are reset to flatten the indexing and make the grouped data accessible for further analysis.

5.4 Feature Engineering

- **Calculating Boundary Counts**

New columns boundary_count and sixer_count are created by applying a lambda function that checks the runs_off_bat for each ball and counts fours and sixes, respectively.
These columns are then aggregated to get the total boundaries and sixes for each match.

- **Merging Aggregated Data**

The original batting stats and the aggregated boundary and sixer counts are merged based on several common columns to create a comprehensive data frame.

- **Calculating Half-Centuries and Centuries**

New binary columns half_century_count and century_count are added to flag whether Kohli scored a half-century or a century in each match.

- **Calculating Strike Rate**

The strike rate is computed by dividing the total runs scored by the number of balls faced and multiplying by 100 to convert it into a percentage.

5.5 Calculation of Fantasy Points

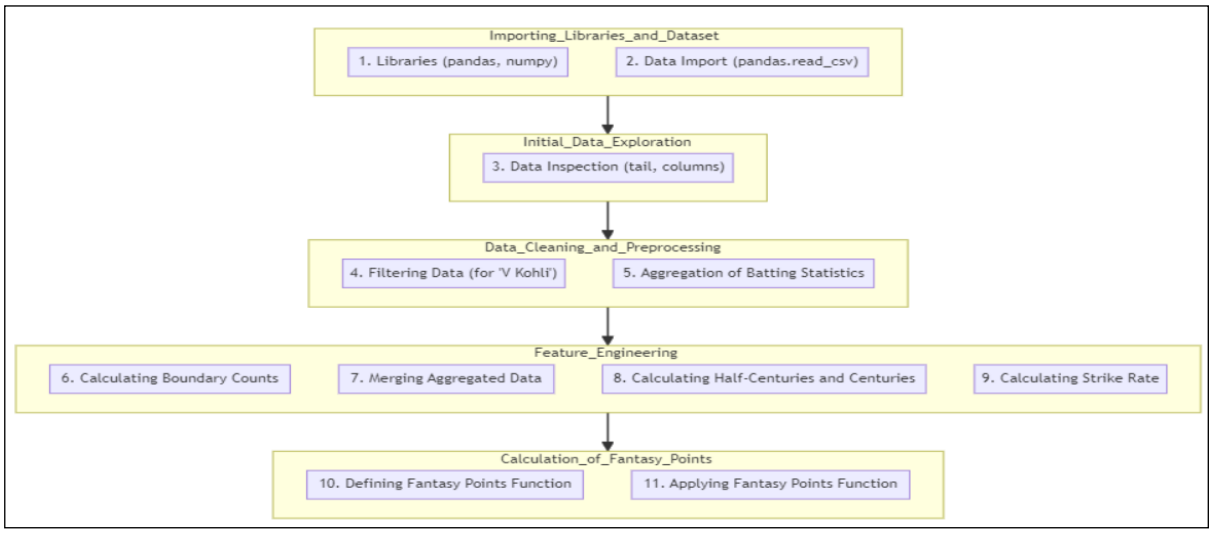

Figure 5: Flow diagram explaining how fantasy points are calculated

Defining Fantasy Points Function:

A function calculate_fantasy_points is defined to compute fantasy points based on runs, boundaries, sixes, half-centuries, centuries, and strike rates. Bonus points are added for milestones. For each runs scored player will be awarded with 1 point and 1 point bonus for every boundary the player scores and 2 points for sixer and 8 points for half century and 16 points for the century, in figure 5 we can see that flow how the fantasy points are calculated.

Applying Fantasy Points Function

The apply method is used to calculate fantasy points for each row in the data frame.

5.6 Implementation of the Rolling Window technique

Initialization

Lists are initialized to store the rolling predictions (rolling predictions) and the corresponding actual values (actual values) from the training set.

Rolling Window Loop

A for-loop iterates over the training dataset with a rolling window of size 2 (can be adjusted as needed). For each iteration, a subset of the training data (X_train_subset, y_train_subset) is used to train a temporary Linear Regression model (lr_temp).

- **Prediction:** The trained model predicts the next value (X_next_match) in the series. The prediction and actual value are appended to their respective lists and for prediction we used deep learning models.
- **Metrics Calculation:** After predictions are made for the rolling windows, performance metrics are calculated including RMSE, MAE.

5.7 Model Building

Initially models were built with the conventional approach and then DL algorithms were used to deal with the time series data. We used balls, runs scored, boundary count, sixer count, strike rate, half century count, century count, venue, innings, bowling team as the features.

Linear Regression (LR): A simple linear model that attempts to predict the dependent variable as a linear combination of the independent variables.

SVM with Linear Kernel: A more complex model that can capture linear relationships in higher-dimensional space.

SVM with Polynomial Kernel: An SVM model that can capture non-linear relationships.

DL Model: The deep learning models such as LSTM, BiLSTM and GRU models are also implemented for predicting the fantasy point of players.

5.8 Model Evaluation

The performance of the models is evaluated using metrics such as RMSE, MAE, MSE, and R2 Score. These metrics provide a quantitative measure of how well the models are able to predict the test data. MAE measures the average magnitude of errors between predicted and actual values without considering their direction. MSE is like MAE but squares the differences, penalizing larger errors. R^2 score indicates the proportion of the variance in the dependent variable that is predictable from the independent variables.

Final Output

The performance metrics for each model are printed, providing insights into which model performs best based on the given data and chosen features. The lower the MAE and MSE, and the higher the R^2 value, the better the model performs.

6 Evaluation

6.1 Case study 1: Fantasy Point estimation using conventional methods:

6.1.1 Player 1: Virat Kohli

Linear Regression, SVM with Linear Kernel, and SVM with Polynomial Kernel:

Table 2: Performance Metrics Values with conventional Methods for Virat Kohli

	Results		
Model/ Results	MAE	MSE	R Square Value
Linear Regression	0.45	0.85	0.99
SVM with Linear	0.35	0.96	0.99
SVM with Polynomial	7.68	88.50	0.85

MSE and MAE: These values were notably low for these models, indicating that the predictions were, on average, very close to the actual fantasy points scored. A low RMSE and MAE suggest high accuracy in the predictions.

R^2 value: The R^2 values were high (for linear Regression 0.99, SVM Linear Kerneyl 0.99 and for SVM polynomial Kernel 0.85), implying that a significant portion of the variance in the fantasy points was successfully captured by the models. A high R^2 value indicates that the model explains a large proportion of the variance in the response variable.

6.1.2. Player 2: MS Dhoni

Linear Regression and SVM Models

Good Performance: The initial approach using linear regression and SVM (both linear and polynomial kernels) yielded very good results. This suggests that the features selected for these models (such as runs off bat, balls faced, boundary count, sixer count, strike rate, etc.) have a strong linear relationship with the target variable (fantasy points).

High R^2 Scores: The high R^2 scores from these models indicate a strong correlation between predicted and actual fantasy points. The obtained R^2 values of linear regression, SVM Linear Kernel and SVM Polynomial Kernel are 0.99, 0.99 and 0.87 respectively which are a quite good value. This suggests that these models are capturing the majority of variance in the data.

Low Error Metrics: The low MAE and MSE values for these models further reinforce their accuracy and reliability in predicting fantasy points. As per the results it can be interpreted that MAE value of linear regression and SVM linear kernel model are 0.43 and 0.22 respectively.

Table 3: Performance Metrics Values with conventional Methods for MS Dhoni

	Results		
Model/ Results	MAE	MSE	R Square Value
Linear Regression	0.43	0.44	0.99
SVM with Linear	0.22	0.48	0.99
SVM with Polynomial	6.76	72.14	0.87

6.1.3 Player 3: Rohith Sharma

Linear Regression, SVM Linear, and SVM Polynomial

The initial models included Linear Regression and two variants of SVM - one with a linear kernel and another with a polynomial kernel.

These models showed excellent performance with high R2 scores, suggesting they were able to capture the relationship between the features (like runs scored, boundaries, sixer counts, etc.) and the target variable (fantasy points) effectively.

Table 4: Performance Metrics Values with conventional Methods for Rohith Sharma

	Results		
Model/ Results	MAE	MSE	R Square Value
Linear Regression	0.71	5.67	0.99
SVM with Linear	0.53	6.23	0.99
SVM with Polynomial	5.38	45.18	0.95

The choice of features and the data preprocessing steps played a crucial role in achieving these results. The high accuracy could also indicate that the models were successful in capturing linear and non-linear relationships in the data.

Key Insights:

- The features such as runs scored, number of balls faced, boundary and sixer counts, and strike rates are strong predictors for fantasy points.

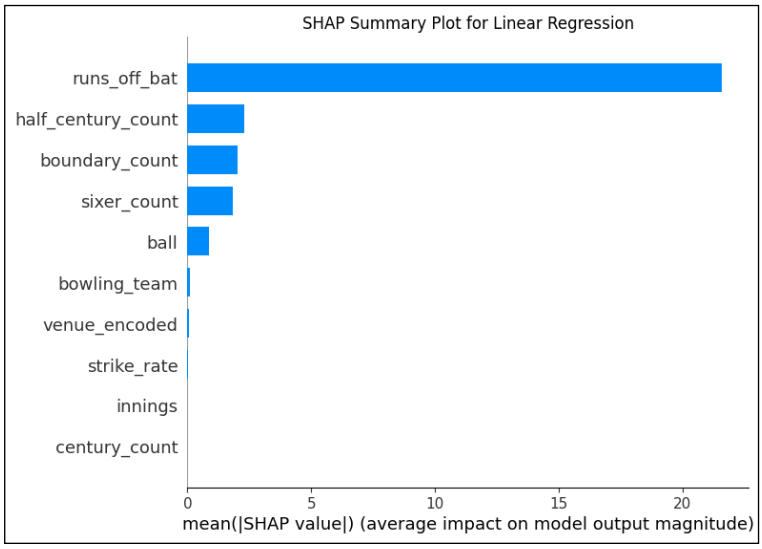

Figure 6: SHAP summary plot for Linear Regression

- The models were robust enough to handle variations in performance and could predict fantasy points with reasonable accuracy.
- The results indicate that simple models can sometimes be very effective, especially when the relationships in the data are not overly complex.

6.2 Case study 2: Prediction using Rolling window technique with the help of DL.

6.2.1 DL Models for Time Series Forecasting for Kohli:

LSTM, Bidirectional LSTM and GRU:

Table 5: Results of Virat Kohli for time series data

	Training Data		Testing Data	
Model/ Results	MAE	RMSE	MAE	RMSE
LSTM	26.05	32.52	26.24	31.29
Bidirectional LSTM	26.50	32.39	25.43	30.66
GRU	25.62	32.82	25.33	32.11

RMSE and MAE: For these models, the RMSE and MAE values were higher compared to the traditional models (more than 20 and 30). The RMSE values of LSTM, BiLSTM and GRU models are 31.29, 30.66 and 32.11 respectively which are quite high. Similarly, the MAE values of LSTM, BiLSTM and GRU models are 26.24, 25.43 and 25.35 respectively on test data. This suggests that the predictions were less accurate and had greater deviations from the actual values.

6.2.2 DL Models for Time Series Forecasting for Dhoni

Table 6: Results of Dhoni for time series data

	Training Data		Testing Data	
Model/ Results	MAE	RMSE	MAE	RMSE
LSTM	11.37	14.87	24.75	30.27
Bidirectional LSTM	13.39	17.59	19.58	27.28
GRU	17.08	22.15	19.30	27.02

Windowed Approach: For time series forecasting, a windowed approach was used where data from 10 consecutive matches was used to predict the fantasy points for the 11th match. In this approach we get so many data points to compare so that it can be effectively evaluated.

Models Used: Various deep learning models like LSTM, GRU, and Bidirectional LSTM were used. These models are well-suited for capturing complex patterns and dependencies in sequential data like time series.

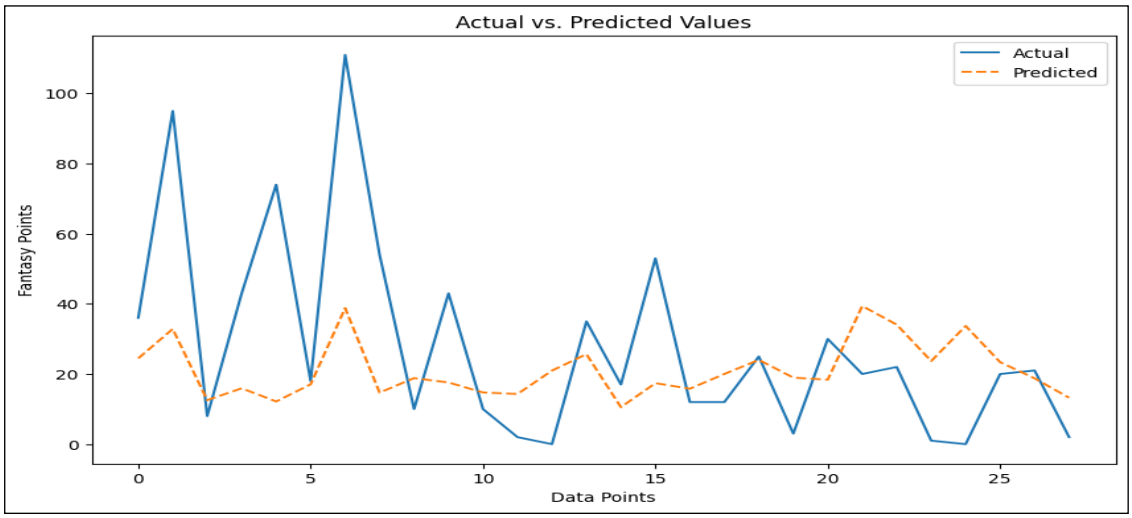

Figure 7: Actual vs predicted values using BiLSTM

As per the obtained performance metrics shown above, it can be said that the LSTM, BiLSTM and GRU models have the MAE and RMSE values are quite high and looking at the figure 7 we can see that there is so much variation in the expected vs actual values.

Poor Performance: Despite the sophistication of these models, the results were not as good as expected. This could be due to several reasons:

- Overfitting: DL models, due to their complexity and large number of parameters, might overfit the training data, especially if the dataset is not large enough and, in this case, it was true because even though IPL has started in 2008, each player has played around only 200 matches till now and hence the dataset is not large enough.
- Hyperparameter Tuning: In this project it was tried to improve the results by hyperparameter tuning but even then could not improve the results by much, may be more hyperparameter tuning is required.

- Data Preprocessing: The way the data is pre-processed and fed into the models can significantly impact the performance. For time series data, aspects like normalization, handling missing values, and choosing the right window size are crucial.
- Model Complexity: The complexity of deep learning models, while beneficial for capturing non-linear patterns, might be unnecessary for this specific dataset. Sometimes, simpler models perform better due to their generalization capabilities.

6.2.3 DL Models for Time Series Forecasting for Rohith Sharma:

Table 7:Results of Rohith Sharma for time series data

	Training Data		Testing Data	
Model/ Results	MAE	RMSE	MAE	RMSE
LSTM	25.79	32.22	26.76	31.28
Bidirectional LSTM	25.75	31.85	25.54	31.41
GRU	25.94	33.43	25.61	32.66

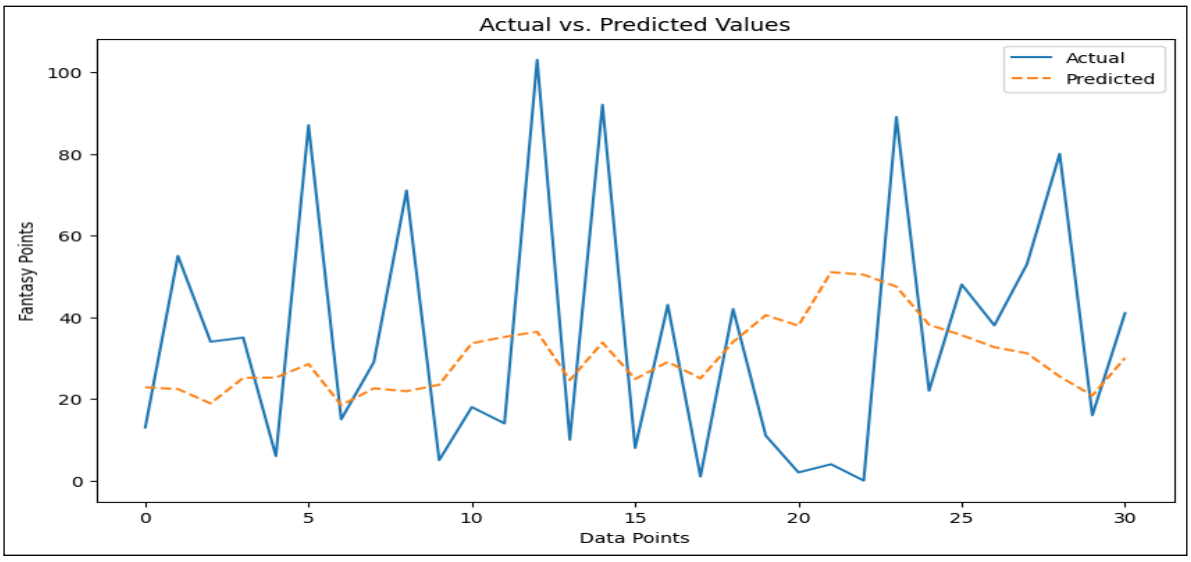

Figure 8: Actual vs predicted values using BiLSTM

DL Models:

For time series forecasting, a window of data from 10 matches was used to predict the fantasy points for the 11th match. Several deep learning models, including LSTM (Long Short-Term Memory), Bidirectional LSTM, and GRU (Gated Recurrent Unit) were employed. Surprisingly, the performance of these models was significantly lower than the initial machine learning models.

Interpretation

Conventional Model's Strength: The conventional models' low RMSE and MAE values and high R^2 values indicate strong estimating performance. These models were effective in making accurate estimation and explaining the variance in fantasy points.

DL Model's Challenges: The higher RMSE and MAE values for the DL models indicate that these models struggled with the prediction task. This could be due to the complex nature of time-series data in sports or possible overfitting issues.

6.3 Discussion

Critique and Contextualization of Experiments

In contrast to the prevailing expectation that deep learning models would thrive in intricate prediction tasks, we did not get the expected result. Conventional models such as SVM and Linear Regression gave good results in estimation, and this is consistent with the findings reported by (Kapadia, et al., 2019) but there are some disadvantages like considering current match results to estimate the fantasy points and having only one data point to compare.

Difficulties Presented by DL Models: Several factors may contribute to the suboptimal performance of DL models (LSTM, BiLSTM, GRU) when it comes to predicting fantasy points. These include the possibility of overfitting, suboptimal hyperparameter tuning.

Contextual Relevance: When considering these results in the context of prior investigations, it becomes clear that although advanced machine learning methods provide sophisticated instruments for analyzing data, their efficacy is dependent on the characteristics of the data and the particular circumstances surrounding the issue.

Suggesting Future Directions: Subsequent investigations ought to concentrate on enhancing the process of feature selection, testing different window sizes in time series models, and investigating hybrid models that amalgamate the merits of conventional and DL approaches. The dynamic nature of the incorporation of technology and analytics in sports is consistently evolving. To stay abreast of this development, it is imperative to conduct ongoing research.

7 Conclusion and Future Work

The primary aim of this study was to explore and analyse the performance of selected cricket players in the IPL, focusing on the estimation and prediction of their fantasy points. The research initially employed ML techniques such as SVM, linear regression, and DL methods such as LSTM, BiLSTM and GRU. These models provided insightful results by analysing various features like strike rate, runs, balls faced, boundary and sixer counts, which are crucial in calculating fantasy points. However, this approach included current match data and also a single data point for evaluation and hence there was need for a more robust predictive model.

To address these issues, the study transitioned to a rolling window technique, employing deep learning techniques that relied exclusively on historical performance data, excluding current match variables. This shift was aimed at creating a predictive model based solely on past performance, thereby reducing the risk of overfitting. Although the DL models did not outperform the initial ML techniques in terms of accuracy, this methodological change was considered more suitable for the nature of sports performance prediction.

The study underscores the challenges in sports analytics, particularly in contexts with limited data or when additional features may be required for enhanced accuracy. It also highlights the importance of selecting the right approach based on the specific objectives and constraints of the analysis. While the DL models did not yield superior results, the approach was deemed more appropriate for long-term predictive accuracy, laying a foundation for future research in this area. The findings of this research have significant implications for fantasy sports and betting sectors, where understanding and predicting player performance is of paramount importance.

Future Work

Additional research should investigate the integration of more intricate data points, including situational variables such as weather and pitch conditions, player psychological factors, and team dynamics, in order to augment the accuracy of predictions. It may also be beneficial to conduct research on how these models can potentially adjust to swift transformations in team compositions and player configurations. Commercialization of these predictive models is a possibility in the fantasy sports and wagering industries, where precise forecasts can have a substantial influence on both user satisfaction and business performance. An additional research endeavour might involve the construction of more sophisticated, yet comprehensible models that integrate state-of-the-art machine learning methods with conventional statistical methods in order to provide a more comprehensive perspective on cricket analytics.

References

Balbudhe, P., Khandelwal, B. & Solanki, S., 2022. Automated Training Techniques and Electronics Sensors Role in Cricket: A Review. *IOP Science,* p. 1310–1333.

Bhatnagar, V. & Batham, D., 2022. Estimating the Chances of Winning IPL Match using Machine Learning. *JOURNAL OF SOFTWARE ENGINEERING TOOLS & TECHNOLOGY TRENDS,* Volume 9.

Box, G. E., 1976. Science and Statistics. *Journal of the American Statistical Association,* pp. 791-799.

Bunker, R. & Susnjak, T., 2019. The Application of Machine Learning Techniques for Predicting Results in Team Sport: A Review. *Journal of Artificial Intelligence Research.*

Chittibabu, V. & Sundararaman, M., 2023. Base price determination for IPL mega auctions: A player performance-based approach. *Journal of Sports Analytics.*

Cortes, C. & Vapnik, 1995. Support-vector networks. p. 273–297.

Gokul, G. & Malolan, S., 2023. Determining the playing 11 based on opposition squad: An IPL illustration. *Journal of Sports Analytics,* Volume 9, pp. 191-203.

Gupta, K., 2022. An integrated batting performance analytics model for women's cricket using Principal Component Analysis and Gini scores. *Decision Analytics Journal,* Volume 4.

Himagireesh, C., P.V., P., B., B. & Taj, T., 2023. Selection of dream-11 players in T20 cricket by using TOPSIS method. *Management Science Letters,* 13(4), pp. 257-264.

Kanungo, V. & Bomatpalli, T., 2019. Data visualization and toss related analysis of IPL teams and batsmen performances. *Researchgate,* Volume 9, pp. 4423-4432.

Kapadia, K., Jaber, H. A., Thabtah, F. & Hadi, W., 2019. Sport analytics for cricket game. *Emerald,* pp. 1-11.

Karkera, P., Bagchi, A. & Bhattacharya, D., 2020. Indian Premier League – Effect of Distance Travelled by Teams during Group Stages on their Home and Away Wins. 23(17).

Kumarapandiyan, G. & Keerthiverman, S., 2020. A STATISTICAL ANALYSIS FOR PREDICTING THE TOP. *MUK Publications,* Volume 24.

Lutz, J. et al., 2019. Wearables for Integrative Performance and Tactic Analyses: Opportunities, Challenges, and Future Directions. *Int J Environ Res Public Health.*

Mohmmad, S. et al., 2020. Survey on machine leaning based game predictions. *IOP Publishing.*

Naik, M. B. T., Hashmi, D. M. F. & Bokde, D. N. D., 2022. A comprehensive review of computer vision in sports: Open issues, future trends and research directions. *Applied Sciences,* 12(9).

Noorbhai, D. H., 2020. A systematic review of the batting backlift technique in cricket. *Journal of Human Kinetics,* pp. 207-223.

Prakash, C. D. & Verma, S., 2022. A new in-form and role-based Deep Player Performance Index for player evaluation in T20 Cricket. *Decision Analytics Journal,* Volume 2.

Pramoda, K., 2021. *T20 cricket match score and winning team prediction using machine learning techniques,* Colombo: s.n.

Sagar, M. & Sharma, J., 2022. Community Engagement with Social and Digital Media Content: A Study on Online fan. *Journal of Content, Community & Communication,* Volume 16.

Sanger, T. D., 1989. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Science Direct,* pp. 459-473.

Sarkar, D. A. & Jana, S., 2020. Role of Captain's Nationality in Team's Success: A Case of Indian Premier League. *Shodhsamhita,* pp. 1-13.

Singh, U., Ramachandran, A. K., Doma, K. & Connor, J. D., 2023. Exploring the influence of task and environmental constraints on batting and bowling performance in cricket: A systematic review. *Sage Journals Home,* 18(6).

Skala, F. & Zemková, E., 2022. Effects of Acute Fatigue on Cognitive Performance in Team Sport Players: Does It Change the Way They Perform? A Scoping Review. *Appl. Sci. 2022, 12(3), 1736;*.

Sloane, S., 2020. *Analysis of Performance Indicators in IPL Twenty20 Cricket from 2015 to 2017,* BLOEMFONTEIN: s.n.

V.S., A. K., Mishra, A. S. & B, V., 2020. Comprehensive Data Analysis and Prediction on IPL using Machine Learning. *International Journal on Emerging Technologies,* pp. 218-228.