PREDICTING FANTASY TEAM FOR IPL T20 CRICKET USING MACHINE LEARNING

Sivakalyan S
LJMU Student Id : 944621
MSc Data Science

Final Thesis Report

JUNE 2021

DEDICATION

This thesis is dedicated to both my parents. My father Mr. P Subbiah did not only raise and nurture me but also taxed himself dearly over the years for my education and intellectual development. My mother, Mrs. M Deivanayagi has been a source of motivation and strength during moments of despair and discouragement. Her motherly care and support have been shown in incredible ways recently.

I also dedicate this thesis to my sister Ms. S Isha Indhu and all my friends and relatives who constantly supported me in all means to complete this thesis.

ACKNOWLEDGMENT

Foremost, I would like to express my sincere gratitude to my project mentor for the continuous support during our study, for his patience, motivation, enthusiasm and immense knowledge. His guidance helped me through the time of research and writing of the thesis. I could not have imagined a better advisor and mentor for the project. I would also thank the review committee members for their constant inputs and support throughout the project. I am grateful to the college John Moore's University, Liverpool for providing the resources, giving me the wonderful opportunity to explore my framework. .

Lastly, I would also like to thank my parents and friends for their constant moral support and motivation throughout the project.

ABSTRACT

Cricket is one of the most admired games played all around the globe. A sport that is played between two opposing teams each of them having 11 players, a combination of batsmen, i.e., the players who bat, bowlers, i.e., the players who can bowl, and all-rounders, i.e., the players who can do both. The game is generally played in three formats such as Test, One Day Internationals, and T20. With the increase in the popularity of the T20 format, the gameplay has become complicated and therefore it becomes necessary to devise new batting and bowling techniques in a very short period due to the limited time available for the players to adapt to the changing match situations. The batsmen play a very crucial role by scoring as many runs as possible in an innings and the bowlers are expected to restrict the batsmen from scoring runs, either by dismissing the batsmen or containing the batsmen from scoring runs. The captain, coach, and team management find it difficult to identify the best playing 11 from a squad of 15 to 17 players. The best playing 11 is selected based on the player's performance against the opposition team in the venue of the match. The player's performance is measured using various metrics. This thesis addresses the above problem and helps the coaches in selecting the best playing 11 for an IPL match. The player performance metrics are calculated using the IPL dataset. Using the player performance metrics, regression models are used. These models include linear regression, ridge regression, lasso regression, decision tree, random forest, etc. These models are then used to predict the dream 11 score for each player. Based on these predicted score values, the best playing 11 which consists of 5 – 6 batsmen, 1-2 all-rounders, and 4-5 bowlers are identified. They are selected using the method of Linear programming which possesses various constraints such as the player cost and the no of players in each category. In batting and all-rounder categories the XGBoost model has performed well compared to all the other models such as linear, lasso, ridge, random forest, and Catboost. The R square values obtained are 98.02% and 98.2% respectively. The RMSE value for batting and all-rounder categories are 6.297 and 7.79 respectively. In the bowling category lasso, ridge, and linear regression has performed marginally better than the XGBoost model. The R square value is 99.9% but the lasso model has performed well in terms of RMSE score compared to the ridge and linear models. The value obtained from the lasso model is 0.008 while the values obtained from the ridge model and the linear model were 2.26 and 2.03 respectively. The random forest, Catboost, and XGBoost models are considered for the prediction of the fantasy team selection and it is observed that the XGBoost model has performed relatively better than random forest and Catboost. The team fantasy point predicted by XGBoost, Catboost and random forest are 405,400 and 400 respectively.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF EQUATION

LIST OF FIGURES

LIST OF ABBREVIATION

ABBREVIATION	FULL FORM
IPL	Indian Premier league
DEA	Data Development Analysis
RPO	Runs per over
T20	Twenty20 International
ODI	One Day International
CBR	Combined Bowling Ratio
HLM	Hierarchical Linear Modeling
ROC	Receiver Operating Curve
SVM	Support Vector Machine
DNN	Feed-Forward Deep Neural Network
XGBoost	Extreme Gradient Boosting
VIF	Variance Inflation Factor
PCA	Principal Component Analysis
PLS	Partial Least Squares
RSS	Residual Sum of Error
TSS	Total Sum of Error
RMSE	Root Mean Square Error
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion

CHAPTER 1

INTRODUCTION

1.1 Background

Cricket is an interesting and popular sport that is played between two opposing teams, each of which consists of 11 players. One team chooses to bat while the other is made to bowl (field) and one such session is called an innings. The game is played on an oval or round-shaped ground and in the middle of the ground, there lies a 22-yard-long pitch where the actual game is played. At both the extremes of the pitch, wickets along with three wooden stumps with two bails on top are kept. Each team consists of 11 players, all the 11 players from the bowling team and 2 players from the batting team will be actively involved at any particular time of the match. The batting team needs to score as many runs as possible before the end of their innings. The bowling team needs to restrict the batting team from scoring runs. The bowling team places their players in different fielding positions around the pitch to restrict the batsmen from scoring runs. After the end of the innings, the batting and bowling team will switch their roles. The team which has scored the maximum runs at the end of the match is considered the winner of the match. The end of the innings is reached when either 10 wickets from the batting team are dismissed or allocated overs before the innings gets completed. Six legal deliveries from the bowling team are together considered as one over.

Cricket is predominantly played in three formats: The first one being the Test Cricket, which generally spans for 5 days where each team is allowed to bat twice, and at the end of 4 innings, the team that has scored maximum runs will be considered the winner of the match. There is no limit on the number of overs for this format of the game, the team can bat until all the 10 wickets are dismissed by the bowling team. A maximum of 90 overs can be bowled in a day. The other two formats are limited over cricket ODI and T20. The ODI is played for 50 overs each, whereas the T20 is played only for 20 overs each. The most challenging format turns out to be T20. The team needs to adapt to the situation swiftly to get control of the match. The focus of this thesis is the Indian Premier League (IPL). It is the most attractive and attended cricket league in the world. IPL is played amongst 8 teams.

The Fantasy League is an online game; in which a virtual team with real players has to be selected by the user. The user will be awarded points based on the player's performance in the real match. The Fantasy League is almost played for all the sports such as cricket, football, baseball, hockey, auto racing, etc. Football and Baseball made significant progress in the

Fantasy League in the late 20th century and early 21st century. The Forbes report says that the market has expanded from 15 million people playing football fantasy with an average of $150 per person a year, which makes it a $ 1.5 billion market in 2003 to 32 million people playing with an average of $467 per person a year, making it a $ 1.5 billion market in 2013. The Fantasy League is played in two formats namely long and short. The long-format Fantasy League is played across a tournament where the user selects a team and is only allowed to make limited changes during the entire tournament. The shorter format is played before every match where the user is supposed to select 11 players from the 2 playing teams on a respective day. The advent of IPL in 2008 started attracting Fantasy Leagues to India. In the recent past, cricket fantasy has significantly grown into a huge market from 2 million users in 2016 to 90 million users in 2020. The major cricket Fantasy League platforms are Dream 11, Myteam11, My11circle, etc. where it is required to select a team of 11 consisting of 3-5 batsmen, 1-2 all-rounders, 3-5 bowlers, and 1-2 wicket-keeper.

In this thesis, the best dream 11 team has been predicted for the upcoming match by evaluating the performance of the player using the available IPL dataset (2008 – 2020). The performance of the player is measured based on their role in the team. It can be measured based on the contributions made by an individual player as a batsman, a bowler, and an all-rounder. The metrics for players vary depending on their roles. The metrics for batsmen are Batting Average, Strike Rate, Innings, Runs scored, no. of thirties, no. of fifties, etc. The corresponding metrics for bowlers are Innings, Overs, Wickets, Maiden, Bowling Average, 4 wicket haul, etc. Considering the above features, different models such as linear regression, ridge regression, lasso regression, decision tree regressor, random forest regressor, etc. are formulated to evaluate the dream 11 points for each batsman and bowler// score individually. Based on the points scored by the players and the cost incurred to select a particular player in the team, a Linear programming model is used to identify the best playing 11 such that the total cost is less than 100. The performance of each player varies based on the venue, opposition, pitch, form of the player, etc. The form of the player is a phrase used generally used to refer to a measure of the player's performance. In recent times, more weightage is given to the player who performs relatively better. The playing 11 for a team will vary for every match based on the player's performance.

During the literature review, various important player metrics were identified, and using these metrics a player's performance has been evaluated. Using the player's performance, the best playing 11 are identified with a given set of constraints. After identifying the team, the result

of the match is identified. The best players from both teams are selected to form the Fantasy League team.

1.2 Problem Statement

Selecting the right players for each match plays a significant role in a team's victory. In this research, identifying the best fantasy team with constraints on the cost of players and number of players to be selected in each category plays a vital role in winning the Fantasy League, the best fantasy 11 team for the match is predicted based on the batting, bowling, and fielding performances of players and other external factors such as venue, opposition, etc. The playing 11 consists of a combination of 5 – 6 batsmen including a wicketkeeper, 1 - 2 all-rounder, and 4 – 5 bowlers. The performance of the player is measured based on the role they play in the team such as batsmen, bowlers, and all-rounder.

The metrics for players vary depending on their roles. The metrics for batsmen are Batting Average, Strike Rate, Innings, Runs scored, no. of thirties, no. of fifties, etc. The metrics for bowlers are Innings, Overs, Wickets, Maiden, Bowling Average, 4 wicket haul, etc. the metrics for an all-rounder are considered from both the batting and bowling metrics. The ball-by-ball data is then aggregated based on the playing team, opposition team, venue, city, and player. The above metrics are calculated for each category respectively. The fantasy points have been calculated based on the points allocated for each of the metrics in the fantasy league portal.

Considering the above features and the total fantasy league points, different regression models such as linear regression, ridge regression, lasso regression, random forest regressor, Catboost, and XGBoost models have been trained to evaluate the relationship between the player's metrics and the fantasy points for each of the player role.

Based on the model trained three models, random forest regressor, Catboost regressor, and the XGBoost regressor were used to predict the player's performance for a particular match, 'Chennai Super Kings' vs 'Royal Challengers Bangalore' in 'Wankhede Stadium, Mumbai'. Based on the points scored by the players, a linear programming model is used to identify the best fantasy 11 consisting of 5 - 6 batsmen which include a wicketkeeper, 1 – 2 all-rounder, and 4 – 5 bowlers whose total cost is less than 100.

The performance of the players varies based on the venue, opposition, pitch, form of the player, etc. The form of the player is a phrase used generally used to refer to a measure of the player's performance. More weightage is given to the player who performs relatively better. The playing 11 for a team will vary for every match based on the player's performance.

1.3 Aim and Objectives

The main objective of this research is to identify the best model that can predict the maximum fantasy points scored by the player based on their performance against opposition in a particular venue, using the model and the player cost. The Linear programming model helps to identify the best fantasy 11, team.

Based on the aim of this research following research objectives are being proposed

- To identify the player's performance metrics for each category such as batsmen, bowler, and all-rounder individually.
- To train various models to evaluate the performance of the players.
- To analyze the relationship between the player's performance metrics with the fantasy points scored which helps in identifying the contribution of players to the fantasy team.
- To evaluate the performance of the model using R^2, RMSE, AIC, and BIC.
- To identify the best fantasy 11 team which consists of 5 - 6 batsmen 1 – 3 all-rounder, and 4 – 5 bowlers, and also include a wicketkeeper whose total cost is less than 100 using a linear programming model.

1.4 Research Questions

The following research questions are suggested for each of the research objectives as highlighted as follows.

- Prediction of the developed model
- Impact of the form of the players in the predicted model
- Assessing the scope of improvement to/of players in a match

1.5 Scope of the Study

The scope of the project is to identify the best fantasy team by maximizing the fantasy points with the constraint set on the total squad cost. The limitations of the study are that external factors like the toss result, climate, pitch are not considered while training the model. The physical health of the players and the mental state of the player is not part of the study. Other factors like the pressure and the non-striker influence on the performance of the striker are also not considered.

1.6 Significance of the Study

The result of the research will be of great benefit to the following people:

Fantasy players – The fantasy players will get a clear indication of the win percentage of the selected team. The players can use the win percentage to select different prize money competitions and accordingly take the risk.

Team Management – The Machine learning model's results helps to predict the players who will perform better against opposition in a specific venue. The team management will be benefitted by selecting the best 11 to play against the opposition.

Cricket Players – The players can improve their performance based on the performance analysis carried. This will help players identify their weaknesses and work on them to improve their game against a particular opposition.

1.7 Structure of the Study

In this study, the most popular regression machine learning algorithms, such as linear regression, ridge, lasso, random forest regression, Catboost, and XGBoost are applied over the IPL dataset and the results of the implemented models are compared using various evaluation metrics to find the best model that predicts the relationship between the player's performance metrics and the fantasy points. Using the three best models, the players for the match between 'Chennai super kings' vs 'Royal Challengers Bangalore' in Wankhede stadium, Mumbai are selected and the fantasy team with best 11 players are identified using the linear programming model.

The current study is organized as follows: Chapter 1 introduces cricket and fantasy league, discusses the problem statement which is going to be addressed, details upon the aim of the paper, and also addresses the research question to explain the main objective of this study followed by the significance and scope of this research. Chapter 2 presents a literature review with the recent explorative works involved in analyzing the player's performance, player's team selection, match results prediction, IPL, cricket fantasy, football fantasy. Chapter 3 presents the research methodology with a detailed explanation of the dataset, demonstrating the process of data aggregations, pre-processing, and also discusses the player's metrics. It describes the machine learning algorithms used in the study along with the evaluation criteria. It also describes the linear programming model. Chapter 4 presents the data aggregations, the preprocessing, and the model training and evaluation. Chapter 5 discusses the results of the +modeling and Chapter 6 presents the conclusion and recommendation of the study.

CHAPTER 2

RELATED WORK

2.1 Introduction

Cricket fantasy team selection has been the subject of the study in recent times where many researchers have not explored much, but the team selection using the ML technique has been the subject of the study for several experiments over the past few years. The learning from the player's performance will be extended to the fantasy team selection.

The Player's performance has been analyzed by identifying the relationship between the player's metrics and the player's performance. The above analysis is carried for each of the categories like batsmen, bowlers, fielding, and all-rounders individually. The batting performance studies analyze the maximum score that the player will be scoring in the next match using machine learning techniques. The bowler's performance study analyzes the no of wickets the players will take based on the condition and opposition. The combination of both batting and bowling is analyzed for the all-rounders. The players fielding performance studies the contribution in catching the ball in the caught wicket and part of the run-out.

Team selection from the player's performance has been studied based on the performance of the player and balance of the team with 5-6 batsmen with a minimum of 1 wicket-keeper, 1-3 all-rounders, and 3-5 bowlers. Selecting the right players for each match plays a significant role in a team's victory. Many researchers have predicted the match results even before the match using the team selection and the performance of the player against opposition in the given venue.

In this study, the fantasy team prediction is identified on the IPL dataset. Many researchers have analyzed various aspects of IPL such as team selection which is different from the normal team selection procedure where there is a constraint on the number of foreign players. Researchers have researched the player scout and the team dynamics to buy the players in auctions every three years.

Research has analyzed the various methods to predict fantasy team selection not only in cricket but also in football. The fantasy team has been selected based on the player cost and the constraint which insists that a team should consist of both the team players with a minimum of 4 players from both the team and constraints are set on each of the categories. At least 3 – 7 batsmen including 1 wicket-keeper, 1-3 all-rounders, and 3 – 5 bowlers must be picked. The football fantasy also predicts 11 players from the team with constraints on the player's position.

2.2 Player Performance

The Player's performance plays a very important role in the overall performance of the team. This section captures the different methods to evaluate the performance of the individual cricketers in different categories, i.e. batting, bowling, fielding, and overall performance of the team.

2.2.1 Batting Performance

(Wickramasinghe, 2014) predicts the performance of the batsmen on the test match using hierarchical linear modeling (HLM). Using this model, the author was able to identify how player level and team level characteristics influence the performance of the player. The data of the study includes test batsmen from 2006 – 2010 from nine test playing nations, namely India (Ind), Srilanka (SL), Pakistan (Pak), Australia (Aus), England (Eng), New Zealand (NZ), West Indies (WI), South Africa (SA) and Bangladesh (Ban). Any player who has played a minimum of 10 games and has an average of 25+ runs per over (RPO) is considered for this study.

The average runs of an individual player were included in the calculation only if that player had played more than 2 matches in a given series. The rank of the team is also calculated based on the average rank points for the 5 years, i.e. (2005 – 2010). The data has both longitudinal nature and a hierarchical structure of three levels.

After performing the HLM method, and after experimenting with different models the best-fitted model was chosen and the above result of 2 players from each team along with their actual, predicted score and the identified absolute error has been recorded in the above table. The absolute value error is high for 3 - 4 players because their current form was not taken into consideration.

Table 2.1 Predicted runs vs actual runs scored by a player (Wickramasinghe, I. P. (2014))

Player	Country	Predicted runs	Actual runs	Abs (error)
RT Ponting	Aus	47.33	40.09	7.24
MJ Clarke	Aus	42.61	42.80	0.19
Mushfiqur Rahim	Ban	37.14	36.50	0.79
Shakib Al Hasan	Ban	53.00	51.25	1.75
VVS Laxman	Ind	46.86	48.60	1.74
R Dravid	Ind	46.51	50.20	3.69
BB McCullum	NZ	38.43	39.25	0.82
LRPL Taylor	NZ	36.37	36.00	0.37
Taufeeq Umar	Pak	39.57	41.25	1.68
Younis Khan	Pak	43.36	59.00	15.64
AB de Villiers	SA	47.30	48.33	1.03
HM Amla	SA	48.27	59.75	11.48
DPMD Jayawardene	SL	50.53	54.00	3.47
KC Sangakkara	SL	45.51	30.66	14.85
S Chanderpaul	WI	44.55	48.20	3.65
MN Samuels	WI	35.95	36.00	0.05

(Shah, 2017) predicts a new measure to identify the player's performance by using a new measure called the performance index. The author has taken various factors into account such as the situations when the batsman scores run against a difficult bowling unit and when the bowlers pick up wickets when the team has a strong batting line-up. The aggregate of performance of batsmen against each bowler is taken as the total performance index for batsmen. The aggregate of performance of bowlers against each batsman is taken as the total performance index for bowlers.

(Bhattacharjee et al., 2018) predicts the batting partnership of the batsmen using the Pressure Index (PI). The pressure index, for the second innings, is calculated based on the current required run rate (CRRR) whereas the initial required run rate (IRRR), uses a different approach based on the fall of wickets for the first innings. The pressure Index calculation is performed on the World Cup 2016 data and it identifies the top 10 partnerships in the first and second

innings. This prediction can help us in understanding the batting order of the team which is expected to increase the overall performance of the team.

(Deep et al., 2016) predicts the ranking of the IPL players using a deep performance index that uses Machine Learning. The Index calculation is different and it aims to find the most valuable batsmen and most valuable bowlers. With the help of the obtained index and other different metrics, one can rank the players within multiple categories. For instance, within the batsmen category, there are other subcategories such as openers, middle-order batsmen, finisher inexperienced, etc. Similarly, within the bowling category, there are subcategories like fast, medium pace, spin bowlers, etc.

(Barr and Kantor, 2004) predicts batsmen performance by the method of examining a batsman's performance in the one-day cricket game two-dimensionally as an alternative to the largely used one-dimensional model concerning runs per innings adopted conventionally.

2.2.2 Bowling Performance

(Lemmer, 2012) predicts the performance of the bowlers as it varies based on the condition of the pitch. The pitch plays a pivotal role in measuring the performance of the bowler wherein a batting-friendly pitch gives a good economy of about 7-8 runs per over (RPO) whereas in a bowling-friendly pitch it can be around 4-5 runs per over (RPO). To encounter this problem, the author has come up with the combined bowling rate (CBR) equation which takes into consideration, three important bowling metrics which are economy rate (E), bowling average (A), and bowling strike rate (S).

$$CBR = 3/(1/A + 1/E + 1/S) \quad [2.1]$$

The CBR value weights are modified based on the format of the game. In limited overs, the economy is given higher weightage than the other. The significance of the bowler who takes the wicket of the 4^{th} or 5^{th} batsman in the batting order is more than a bowler who takes the wicket of the $10^{th} – 11^{th}$ batsman in the batting order. Therefore, the weightage that is given to the corresponding wickets also varies accordingly. The CBR method is applied to the T20 World Cup 2010. This CBR helps in identifying which player has performed best after the tournament but it will not be able to identify the current best player and also can't predict the best-suited player for the next match.

Table 2.2 Ranking of bowlers according to CBR* score (Lemmer, H. H. (2012))

Rank	Name	O	R	$R^{\#}$	W	W*	CBR*	$CBR^{\#}$	A	E	S
1	C Langeveldt	16.0	104	104.36	11	12.87	7.29	7.30	9.45	6.50	8.73
2	DJG Sammy	13.4	72	64.36	6	5.91	8.72	7.97	12.00	5.27	13.67
3	DP Nannes	26.0	183	176.46	14	17.94	8.45	8.24	13.07	7.04	11.14
4	GP Swann	22.0	144	136.66	10	13.33	8.66	8.34	14.40	6.55	13.20
5	SW Tait	23.4	131	113.27	9	8.7	9.73	8.64	14.56	5.53	15.78
6	MG Johnson	22.2	145	133.45	10	11.57	9.37	8.81	14.50	6.49	13.40
7	KAJ Roach	12.0	77	75.18	5	6.24	9.27	9.11	15.40	6.42	14.40
8	A Nehra	20.0	156	157.23	10	13.29	9.26	9.30	15.60	7.80	12.00
9	M Morkel	15.0	119	123.55	8	10.02	9.33	9.56	14.88	7.93	11.25
10	SPD Smith	23.0	163	164.77	11	12.73	9.63	9.71	14.82	7.09	12.55
11	Saeed Ajmal	22.2	169	168.84	11	13.06	9.78	9.77	15.36	7.57	12.18
12	N McCullum	19.0	124	124.17	7	8.63	10.05	10.06	17.71	6.53	16.29
13	SCJ Broad	20.5	140	133.31	8	9.07	10.48	10.11	17.50	6.72	15.62
14	R Sidebottom	21.3	160	163.55	10	11.69	10.07	10.22	16.00	7.44	12.90
15	NO Miller	12.0	63	54.79	2	2.14	11.80	10.42	31.50	5.25	36.00
16	AD Mathews	12.0	83	82.46	4	5.35	10.59	10.54	20.75	6.92	18.00
17	M Aamer	23.0	152	142.74	8	8.54	11.14	10.61	19.00	6.61	17.25

O = No. of overs R = No. of runs $R^{\#}$ = Adjusted no. of runs W = No. of wickets W* = Sum of weights of wickets
CBR* = Combined bowling rate $CBR^{\#}$ = Combined bowling rate adjusted A = Bowling average E = Economy rate S = Strike rate

(Muthuswamy and Lam, 2008) predicts the performance of Indian bowlers against seven international teams against which the Indian cricket team plays most frequently. They have used backpropagation network (BPN) and radial basis network function (RBNF) techniques to do the task. Using this neural network, the author is trying to predict the number of runs that the bowler is going to concede and then classify the number of wickets the bowler would take in 2 categories which are 0-2 wickets and 3+ wickets.

2.2.3 Fielding performance

(Saikia et al., 2012) predicts the Fielding performance of the player by using a double edge method. The various parameters that are considered for the fielding performance of the player are done by analyzing the video of the game after evaluating every ball that is played in each match. The subjective weights are assigned to the fielder based on the level of difficulty of the fielding performance and also the based on the batsmen. The two different fielding measures are calculated on the world T20 2007 tournament and the players are ranked based on both measures. The measures are preparatory fielding performance measures and fairer fielding performance measures.

Table 2.3 Fielding performance of seven Indian players on T20 world cup final 2007 (Saikia, H. et al. (2012))

SI. no.	Players name	No. of balls fielded	Scores of the based cricketers on		Ranks of the based cricketers on	
			Preparatory Measure (FP_j^1))	Fairer Measure (FP_j^2))	Preparatory Measure (FP_j^1))	Fairer Measure (FP_j^2))
1	MS Dhoni	42	0.396	0.894	7	3
2	S Sreesanth	6	0.516	0.951	3	1
3	R Uthappa	11	0.563	0.891	2	4
4	RP Singh	7	0.466	0.69	4	6
5	R Sharma	7	0.445	0.796	5	5
6	Y Singh	13	0.398	0.685	6	7
7	I Pathan	18	0.601	0.936	1	2

The preparatory measure identifies the overall impact of the game based on the player's fielding performance and the fairer measure identifies the performance of the fielder in the match irrespective of the position of the field.

2.2.4 Team Performance

(Passi et al., 2018) developed a model for increasing prediction accuracy in cricket using machine learning. Player selection plays a vital role in any sport. It is generally accepted that the performance of the players differs upon the location, his current form, physical health, and even depends on the opposition team's strength and performance. A committee including the coach and the captain decides, from a squad of 15-20 members, the 11 players who will be playing the forthcoming matches. This is based on analyzing each player's statistics and their characteristics to choose the best 11 players for every single match. The batsman is expected to score as many runs as possible, while a bowler is expected to prevent the batsman from scoring runs and thereby he also needs to take more wickets. This paper aims to estimate the player's performance by trying to predict the number of runs a batsman can take and the number of wickets a bowler can take on that particular day given the situation. The experiment was carried out using the following classifiers such as multiclass SVM, decision tree, random forest, naïve

Bayes. Out of which random forest was found to be most accurate after comparison. The accuracy obtained was 90.74% for the batsman and 92.25% for the bowler.

Table 2.4 Predicting Runs (Passi, K. and Pandey, N. (2018))

Classifier	**Accuracy (%)**			
	60% train 40% test	**70% train 30% test**	**80% train 20% test**	**90% train 10% test**
Naïve Bayes	43.08	42.95	42.47	42.50
Decision Trees	77.93	79.02	79.38	80.46
Random Forest	89.92	90.27	90.67	90.74
SVM	50.54	50.85	50.88	51.45

Table 2.5 Predicting Wickets (Passi, K. and Pandey, N. (2018))

Classifier	**Accuracy (%)**			
	60% train 40% test	**70% train 30% test**	**80% train 20% test**	**90% train 10% test**
Naïve Bayes	57.05	57.18	57.48	58.12
Decision Trees	84.40	85.12	85.99	86.50
Random Forest	90.68	91.26	91.80	92.25
SVM	67.45	67.53	68.35	68.78

(Managea et al., 2020) developed an approach to classify all-rounders by machine learning techniques. This was done by dividing them into batting all-rounders and bowling all-rounders by their performance statistics. Performance was evaluated on the metrics of accuracy and area under the receiver operating characteristic curve (ROC). It was shown that various machine learning techniques can be used by administrators and team managers to select the players. (Tyagi et al., 2020) developed a predictive model for a cricket game using various machine learning algorithms. Companies spend a huge amount of money to book a slot during the match for advertising their products. Due to a short game, the interest of the people is lost, therefore it becomes a risk for advertising companies. Therefore, a model is developed to analyze the number of balls to be delivered and with the help of it, it also determines the duration of the match. Player performance plays a vital role in deciding this factor.

2.3 Team Selection

Now that the player's performance has been analysed, all the ways in which these metrics will cumulatively affect the team selection has been discussed in the following section. The 11 players to be selected will have a lot of constraints as it is mandatory for the team to have a minimum of 5 bowlers and 1 wicket-keeper. (Bhattacharjee and Saikia, 2013) predicts the best dream 11 teams, after the end of the tournament using batting, bowling, and the wicket keeper's performance as measures. A balanced optimal team is formed based on the performance of a player throughout the tournament. Based on the individual group's performance measure an excel optimization model is used to predict the best 11 played in IPL 5.

Table 2.6 Best playing 11 in IPL 2012 (Bhattacharjee, D. and Saikia, H. (2013))

Player	Team	Type
L Balaji	KKR	Bowler
P Negi	DD	Bowler
AB Dinda	PWI	Bowler
A Mishra	DC	Bowler
CH Gayle	RCB	Batsman
G Gambhir	KKR	Batsman
KP Pietersen	DD	Batsman
RG Sharma	MI	Batsman
MS Dhoni	CSK	Wicket Keeper
A Mahmood	KXIP	All-rounder
SR Watson	RR	All-rounder

(Faez et al., 2011) predicts a model which selects a high-performance team with a restricted budget. Batting and bowling strength play a vital role in the overall performance of the team and an optimum trade-off needs to be reached for the formation of a good team. The author proposes a multi-objective approach using NSGA-II to optimize the batting and bowling performance of the team, based on the optimization approach the final team is also formed. (Amin and Sharma, 2014) predicts the team by using a linear programming technique called the data envelopment analysis (DEA). The evaluation determines both efficient and inefficient players based on the DEA scores. DEA rank prediction is applied to the IPL 2011 and the best team is selected based on the ranking.

Table 2.7 Cricket team chosen for IPL IV using DEA

Capabilities	Players in the team
Batsmen	Virender Sehwag
	Chris Gayle
	Paul Valthathy
	Micheal Hussey
	Tirumalsetti Suman
	Subraminian Badrinath
Bowlers	Lasith Malinga
	Rahul Sharma
	Mitchell Marsh
	Daniel Vettori
	Ali Murtaza
All-rounders	Yuvraj Singh
	Yusuf Pathan
Wicketkeepers	MS Dhoni
	AB de Villiers

(Iyer and Sharda, 2009) predicts the team selection for the world cup using the non - linear modeling such as neural networks. These different models forecast the performance of the player's batting and bowling individually. They use two different ways of experimenting to evaluate the overall and the current form of the players by considering the statistics of the last few years.

(Omkar and Verma, 2003) predicts the team using a genetic algorithm. It calculates the fitness of the player. Based on the fitness of the player they can identify the best team for the tournament from a pool of players. The fitness parameter varies for each category of the players. In the case of a batsman, his fitness is measured upon the number of runs he has scored, while the bowler's fitness is identified as the number of wickets he has taken and so on.

(Ahmeda et al., 2013) proposed a scheme to select a cricket team using a decision-making approach. Selection is based on the efficiency of a player to fit all roles. The main factor that is considered while selecting a good team will especially depend on the batting and bowling strength of the team.

2.4 Result Prediction

Based on the individual player's performance and the team selection, the following section discusses different techniques that try to predict the results of the match or tournament.

(Akarshe et al., 2019) predicts the cricket score using machine learning techniques. Records of data are taken and the model predicts the final score or result of a match. With the help of machine learning algorithms, it analyses the pre-stored data and enhances the prediction system with accurate results. (Basit et al., 2020) proposed to predict the winner of ICC T20 Cricket world cup using machine learning algorithms. Random forest achieved an accuracy of 80.86% and 64.68 of R-mean score which was better than other methods.

(Jayalath, 2018) suggested a method for ODI cricket predictors using machine learning approaches. It is done using the classification and regression tree (CART) algorithm and logistic regression. Factors to be considered are toss result, day-night game, and home-field advantage. (Thenmozhi et al., 2019) implemented a model to predict the winner for an ongoing match using machine learning algorithms like the random forest, K-Nearest Neighbor, support vector machine. Cricket is a sport with plenty of data and these data are processed further to achieve the desired prediction. The results obtained are binary. It gives information about whether or not the match was won by the home team. The accuracies obtained are separately provided for each team. Overall, the accuracy was found to be 75.5%. (Somaskandhan et al., 2017) propounded to identify the optimal set of attributes with the help of machine learning techniques to increase the chance of winning. The team owners will decide the players based on these attributes. Data about every ball was collected from the past IPL tournaments. Support Vector Machine (SVM) obtained the best accuracy of 81%. (Aburas et al., 2018) developed a predictive model for the Cricket World Cup with the help of the KNN Intelligent Big Data Approach. The KNN and data reduction algorithm are employed for this model. MySQL performs the cleaning and cleansing of the data from the datasets.

(Mustafa et al., 2017) presented a model based upon opinions on social networks for predicting the cricket match outcome. Freedom of expression can be done through social media and one such event is a cricket match. Predicting the winner of a match is done through collecting knowledge from micro-posts posted on social media. Features like fan's sentiment, total no of tweets, fans score prediction are taken into account. Effectiveness of Support Vector Machine (SVM) has an edge over other classifiers. (Kanhaiya et al., 2019) proposed a cracked cricket pitch analysis with the help of image processing and machine learning. The pitch analysis is indirectly used to predict winners or losers of the match. The study is done on the location of the crack, its shape, and further analyzes the depth of the crack and its effects on the batsman or bowler.

(Vidisha and Bhatia, 2020) suggested machine learning recommendations for cricket. It was done by analyzing the pitch and predicting the player's performance and predicting the results

of the match. Because of its ability to produce accurate results with small samples, Naïve Bayes produced the best results. (Lemmer et al., 2014) predicts the outcome of the matches based on two models in the T20 cricket series, the author identified that model did not predict the winner of the match properly when both the teams win a respective match. A new adjusted measure was identified to compensate for the inconsistent result. The adjusted measure improved the results from 56.8% to 74.6% for the first model and from 52.7% to 70.8% for the second model.

2.5 IPL

This section researches the Indian Premier League (IPL). IPL player selection is slightly different from the normal team selection where a maximum of 4 foreign players and the correct combination of Indian and foreign players is necessary in the IPL team selection. The auction happens every 3 years where the entire base of the team is changed so it is very difficult to form a consistent team in IPL.

(Vistro et al., 2019) proposed a model to predict IPL match winner even before the match had started, using data from the previous seasons of IPL. It was done with the help of machine learning algorithms and data analytics. Every team should work on their strengths and their performances to make sure that they are equipped enough to win. Apart from the role the venue plays, the player's performance and weather are various other factors that are to be considered for predicting the winner. Various machine learning techniques were applied on both test and training datasets. Out of which, the decision tree classifier achieved an accuracy of 94.87%. (Kansal *et al.*, 2014) predicts the base price of the player in the IPL auction. The players are divided into batsmen, bowlers, and all-rounders respectively and the performance is evaluated on both ODI and T20 Metrics using the Multi-Layer Perceptron model, Naïve Bayes, etc. (Saikia and Bhattacharjee, 2011) predicts the performance of the all-rounders using their strike rate and economy rate. These metrics are modeled using various models such as Naïve Bayes, Multinomial Regression, etc. to classify the all-rounder's into two category performers, batting all-rounder or bowling all-rounder.

(Kapadia et al., 2019) presented a model for sports analytics of a cricket match using machine learning. Based on the historical match data of the IPL, machine learning algorithms are applied to predict cricket match results. Features from the dataset have been identified and the various machine learning techniques are applied on two sets, one being the team that got the advantage of playing at their home ground while the other is based on the toss decision. For the home team featured set, the Random Forest model was found to be more accurate and precise when

compared to other models, whereas for the toss featured set none of the machine algorithms produced accurate results.

2.6 Cricket Fantasy

This section is similar to that of the team selection that has already been discussed above. Cricket fantasy is different from IPL team selection where a maximum of 4 foreign players in the squad can be selected, whereas fantasy team selection does not have any constraint on the number of foreign players. Also in the cricket fantasy, you select 11 players based on the assumption that they will score high and pick more wickets whereas, the Fantasy team can be selected with the allocated money, and the other constraints are the requirement of a minimum of 3 batsmen, 1 wicket-keeper, 1 all-rounder, and 3 bowlers to form the team.

(Karthik et al., 2020) proposed to predict winners of fantasy cricket contests by machine learning techniques. The growing popularity of cricket in recent years has resulted in the emergence of formats such as T20 and T10, which vary from the test and one-day formats. The craze for both of these cricket formats has now spread to online fantasy cricket league games. Dream11, as well as many other similar apps, are some of the most common ones discussed in this context. Putting together a dream team of 11 players requires a mix of skills, ideas, and luck. It's difficult to predict a winner among all the contestants based on previous records. A feed-forward deep neural network (DNN) classifier was used to predict the winner of a fantasy league cricket contest for the top three slots. The data obtained was from the dream11 app in which a private contest was played during the 12th IPL season in 2019. Later this data was processed further. In terms of prediction accuracy, the proposed DNN method outperformed all other classifiers by at least 13%, 8%, and 9% for first, second, and third winning positions, respectively.

(Singla et al., 2020) suggested integer optimization for selecting the dream 11 cricket team. Due to cricket's popularity in India, Fantasy Cricket has gained the most users when compared to other sports. As it provides a huge amount of data it is fairly easy for analyzing. A Dream 11 player must choose the right combination of players to maximize his/her points and, as a result, win cash prizes. The paper proposes a Dream 11 Fantasy team for the upcoming match using a retrospective approach to team selection based on real-world data obtained from player’s records from the previous 10 matches. Integer Programming is used, and the Gurobi library in Python was used for implementing it. The model is also recalculated to account for a Dream 11 User's risk aversion, and it appears to outperform Machine Learning and Integer programming models which concentrate only on ranking metrics and/or overall scores.

(Patel et al., 2019) developed a model to predict IPL player's performance using machine learning approaches. In a Fantasy League, the most critical role is the player's selection. A player's success is influenced by a variety of factors, including the opposing team, the location, his current form, and so on. Analyzing previous records ball by ball (2008-2018) data, the performance of players is predicted in IPL matches using supervised machine learning techniques. The batsman's runs and the bowler's wickets are divided into various ranges. Decision Tree, Random Forest, XGBoost, Stacking are used for the prediction of the players. The most reliable classifier for the problem was discovered as the stacking technique.

(Perera et al., 2015) suggests how to find the best team lineup in a cricket match of Twenty20 format. The lineup is made up of three parts which include the selection of a team, the order of batting, and the order of bowling. The difference between the expected runs scored and the expected runs allowed for a given lineup using match simulation is estimated. Simulated annealing is then used to optimize the lineup over a large combinatorial space. The composition of a best Twenty20 lineup generally ends up providing non-traditional roles for players. An 'all-star' lineup of international Twenty20 cricketers is obtained as a byproduct of this technique. (Naha, 2019) discusses fantasy cricket. As smartphone technology progressed, the digital craze for Fantasy leagues also reached its peak and its influence was felt all over the world. Concerns about fantasy cricket being a gambling cover have surfaced as the number of Indians playing the game has increased in recent years. This paper concludes whether or not the fantasy cricket is skill-based game. And it also discusses about the extent knowledge and luck requirements for assured success in a fantasy gaming. Methods of autoethnography and phenomenology have been used. (Das,2014) suggested for Fantasy Cricket League Selection and Substitution uses an Integer Optimization Framework. Moneyball 1 is a Fantasy league game that uses binary integer programming to generate optimal team sequences in Fantasy leagues, especially cricket. Some of the characteristics of team selection in sport are very similar to those in finance. It also includes an automated back testing system for cricket performance measure testing, as well as a tool for automatically retrieving rich cricket statistics.

Figure 2.1 Dream 11 team selection (www.cricketadictor.com)

2.7 Football Fantasy

The football fantasy is also similar to cricket fantasy where it is required to select a team of 11 based on the formation and strategy that is going to use against the opposition. In this section, a brief introduction about football fantasy is discussed

(Nicholas Bonello et al., 2019) proposed an idea to predict the performances in fantasy football by analyzing the previously recorded data. This is done by including feedback from the humans into the model. Data considered are the past performance records of the players, fixture difficulty ratings, market analysis for betting, experts, and public views through social networking sites and blogs. This helps to provide a better understanding of the situation to select players. Over 6.5 million players every season compete against themselves to gain the most number of points. As the popularity of the fantasy premier league keeps increasing, obtaining a decent rank is equally difficult. Gradient boosting machines were found to be most effective and optimal for the prediction of Fantasy leagues. The main reason for choosing this approach is because it is more effective in dealing with unbalanced data. The data set is available publicly

and is collected from various sources such as blogs, tweets, betting odds and further worked upon removing players who are injured and those whose appearance was not recorded in recent times. According to their positions, the data set is split. Based upon their position and various factors it helps us to predict the player's form. Experiments carried out in the English premier league 2018/2019 season, showed that it outflanked by more than 300 points on average of 11 points each week.

(Arseniy Stolyarov et al., 2017) developed a model for the formation of a fantasy football squad using machine learning tools. Over 4.5 million participants were present in the 2016/2017 fantasy premier league. This fact denotes it is the most competitive fantasy league. To succeed in the FPL, a high degree of competitive pressure necessitates making the right decisions over time. This role is especially challenging since there is an infinite number of potential actions, and the actions must be taken in the face of uncertainty. As a result, a model that is powerful enough to perform well in these conditions is required. The method used is Integer Programming. The model is required to solve the problem in which an objective function sums the number of points expected by a machine learning algorithm and the constraints define the corresponding game rules to maximize the number of points scored in a given round. The dataset used is from the official fantasy premier league website which includes data for combining players for each game week. The XGBoost model is used, which is one of the most successful modern data science tools, to predict expected points.

(Brendan Dwyer et al., 2010) studied the Fantasy league's participation. Modern sports consumption has expanded to include a variety of activities such as event participation, television viewing, and subscriptions. Fantasy league participation is one of these forms of sports fan consumption. Although the sport has grown tremendously in popularity over the last few decades, little is understood about who participates and what effect the participation has on sports product and service use. As a result, this study aimed to look into the various modes of sports consumption shown by Fantasy league participants.

2.8 Discussion

The batting performance is predicted by using various methods like hierarchical linear modeling. The HLM identifies the relationship between the team level and the performance of the player. The performance index evaluates the performance of the player against a strong bowling unit. The pressure index is calculated based on the required run rate and current rate. ML models are used to predict the performance of the player. The combined bowling rate takes the economy rate, bowling average, and bowling strike rate into account to evaluate the

performance of the bowler. Using backpropagation and radial basis network function to predict the number of runs the bowler is going to concede and classify based on the number of wickets the bowler picks. The fielding performance is predicted by using the double edge method.

The best 11 players are selected based on the number of runs the batsmen score and the number of wickets the bowler pick. The model is trained using SVM, decision tree, random forest, and Navies Bayes. The random forest model is more accurate with 90.74% for batsmen and 92.25% for the bowlers. The machine learning model will classify the all-rounders into batting and bowling all-rounders based on a dominant performance in the respective field.

Various optimization techniques like excel optimization, NGSA-II, data envelopment analysis (DEA) are used to predict the best team based on the performance measure. Neural networks are used to estimate the performance of batting and bowling individually. Machine learning algorithms are used to predict the winner of the match.

The random forest model has achieved an accuracy of 80.86% to predict the winner of the T20 world cup. Other classification algorithms such as logistic, SVM, KNN, etc. are used to predict the match winners using the past data.

IPL team selection varies from the normal team selection where the constraints on the number of foreign players are four. Models are predicted to identify the price of the player in the auction. The IPL match results are predicted even before the toss of the match using a machine learning algorithm. The decision tree algorithm performed relatively better than other algorithms with an accuracy of 94.87%. The IPL all-rounder's performance is predicted using strike rate and economy.

The fantasy team selection contains a mix of both playing teams with a minimum of 4 players from each of the teams. The team should consist of 3-5 batsmen, 1-3 wicket-keeper, 1-3 all-rounder, and 3-5 bowlers with a maximum of 11 players. Each player selected will be associated with a player cost and the total team cost shouldn't exceed 100. A feed-forward deep neural network classifier predicts the winner of the dream 11 contest winner.

Similar to Cricket fantasy the football fantasy also expects the player to select 11 players with constraints on the position of players. XGBoost algorithm is used to predict the fantasy point scored by the player at the end of the season of the premier league.

2.9 Summary

The end–to–end method from raw data to identifying the best team in a fantasy league for IPL is not available, only the individual components in the entire process are available such as analyzing the performance of the player, identifying the best playing team with a maximum of

4 foreign players for both the playing teams, forming the fantasy team from the 22 players pool. So, in this thesis, the end-to-end from identifying the player performance metrics to identifying the best fantasy team will be identified.

Based on all the research, our work in this thesis is to use an ML algorithm to predict the best dream 11 consisting of 5-6 batsmen,1-2 all-rounders, and 4-5 bowlers. In this study, the best dream 11 teams against a given opposition can be identified. This player performance prediction has been extended to select the team with a given point in the Fantasy league. The detailed study in this work has captured various aspects that can impact the performance of the player.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Fantasy team selection is becoming popular in recent years, the fantasy team selection is played across various platforms and applications, where identifying the best set of players from both teams will yield the maximum fantasy points. The cricket team fantasy is becoming very popular in India with the advent of the Indian premier league.

The dataset for this study is the IPL (2008 – 2021 mid-season). A brief introduction about the dataset is discussed in this chapter and the aggregation and the preprocessing techniques required to be validated in the dataset before the performance metrics calculation are also explained. The performance of the player is measured based on their role in the team such as batsmen, bowlers, and all-rounders. These player's performance metrics calculation logics are explained in detail.

Machine learning is a branch of Artificial Intelligence that instills the ability to learn into a system based on a data set used for training, as opposed to the traditional method of coding deriving the results. Establishing a system that can understand learning requires a range of methods and techniques. In machine learning methods, there are three types of learning: supervised, unsupervised, and reinforcement learning. A collection of data instances is used to train the system and are named to give the right outcome in supervised learning. Unsupervised learning, on the other hand, lacks predetermined data sets and no idea of the desired result, making the target more difficult to achieve.

One of the most popular supervised machine learning techniques is regression. It identifies the relationship between the independent and dependent variables. The regression models help and identify the performance of the players like the no. of runs the batsmen will score, the no. of wickets the bowler will take, etc. As a result of the related work, linear, lasso, ridge, random forest, Catboost, XGBoost regression algorithms are used to train the data to predict maximum fantasy points. A detailed summary of these regression algorithms is explained in this chapter.

Basis the relationship between the dependent and independent variable the output from the regression model is fed into the linear programming to identify the best fantasy 11. Linear programming is also explained in this chapter.

3.2 Dataset

The data contains around the IPL dataset from 2008-2021 (match 30) as the 2021 season is not yet complete. The dataset contains ball to ball information about all 845 matches in different cities across three different countries. The dataset contains all the columns as stated below:

Table 3.1 Data Features

match_id	runs_off_bat
Season	extras
start_date	wides
Venue	no balls
Innings	byes
Ball	leg byes
batting_team	penalty
bowling_team	wicket_type
Striker	player_dismissed
non_striker	other_wicket_type
Bowler	other_player_dismissed

Most of the fields above should, hopefully, be self-explanatory, but some clarification is also given:

- "innings" contains the number of innings within the match. A match would normally have 2 innings, such as a T20 or ODI, then any innings of more than 2 can be regarded as a super over.
- "ball" is a combination of the over and delivery. For example, "0.3" represents the 3rd ball of the 1st over.
- "wides", "no balls", "byes", "leg byes", and "penalty" contain the total of each type of extras or are left blank if not relevant to the delivery.
- If a wicket occurred on delivery then "wicket type" will contain the method of dismissal, while "player_dismissed" will indicate who was dismissed. There is also the, admittedly remote, possibility that a second dismissal can be recorded on the delivery (such as when a player retires on the same delivery as another dismissal occurs). In this case, "other_wicket_type" will record the reason, while "other_player_dismissed" will show who was dismissed.

3.3 Player Statistics

The performance of the player is measured based upon their role in the team such as batsmen, bowlers, and all-rounders. The measures for players vary depending on their roles. Using the data obtained these measures are calculated. In this study, batting and bowling performance are only considered.

3.3.1 Batting Metrics

Innings: Innings refer to the period in which an individual player has played in his entire career. Higher the innings the better the experience of the player.

Runs Scored: Total Number of runs the batsmen has scored in his career. The more the batsmen score better the player.

Batting Average: The mean score batsmen scored in an innings. It can be mathematically expressed as follows:

$$Average = \frac{Runs\ scored}{Numbers\ of\ Innings\ Played} \qquad [3.1]$$

Batting Strike Rate (SR): The mean runs scored by a batsman for every 100 balls played. The higher the SR quicker the runs get scored. In the limited-overs, especially in T20, this becomes an important metric for the selection of a player.

$$Strike\ Rate = \frac{Runs\ scored}{Number\ of\ balls\ faced} \times 100 \qquad [3.2]$$

Thirties Score: Sum of occasions in which the batsman has a total score of "30+" in an innings. The consistency of the player can be evaluated using this metric. Higher the thirties score more consistent the player is in the T20 format.

Fifties Score: Sum of occasions in which the batsman has a total score of "50+" in an innings. The consistency of the player can be evaluated using this metric. Higher the fifties score more consistent the player in the T20 format. This metric has more precedence than the thirties metric.

Zeros: Sum of occasions in which the player was dismissed even before securing a run.

Highest: Maximum runs that a batsman has scored in an innings in his entire career. The highest score in each venue and opposition can be used as an additional metric while considering the

opposition and venue. The probability of a player scoring more in his home ground is likely high compared to other venues or the player will score maximum runs in specific pitch conditions as well.

Average of Dots: The average number of balls where a player has not scored in a run. The lower the average the better the player. In T20 format, this cannot be high as the number of overs is minimum.

3.3.2 Bowling Metrics

Innings: The sum of occasions where the player has bowled a minimum of 1 ball in their entire career. Higher the innings the better the experience of the player.

Overs: The sum of total overs the bowler has bowled in his entire career. Higher the number of overs the better the experience of the player.

Wickets: Total number of occasions the bowler has dismissed the batsmen in his entire career. Higher the wickets better the player.

Maiden: The number of overs where the bowler has not even conceded a single run. The higher the metric the better the player.

Bowling Average: Mean runs conceded by the bowler to pick a wicket. The Lower the Bowling average the better the bowler who will concede less for picking up a wicket. It can be mathematically expressed as below:

$$Bowling\ Average = \frac{Number\ of\ Runs\ conceded}{Number\ of\ wickets\ taken} \qquad [3.3]$$

Bowling Strike Rate: It is defined as the mean number of balls that a bowler has taken to pick a wicket. This plays a pivotal role in analyzing the number of wickets the bowler can pick on an average in a match. Lower the metric attributes higher the bowler will wicket frequently. It can be mathematically expressed as below:

$$Strike\ Rate = \frac{Number\ of\ balls\ bowled}{Number\ of\ wickets\ taken} \qquad [3.4]$$

4 Wicket Haul: Total sum of the occasion where the sum of wickets taken by the bowler in an innings is equal to or greater than 4. Higher the 4 wicket haul more consistent the player in the T20 format

Economy Rate: Average number of runs the bowler concedes in a single over, lower the economy rate better the bowler performance.

3.4 Data Pre-Processing

These metrics are calculated from the IPL ball-by-ball data which has been discussed above. These batting, bowling, and fielding measures are then calculated. In addition to these, other player parameters like current form, consistency, partnership, opposition, venue, non - striker will be taken into account for selecting the best team. These metrics are calculated using the Analytic Hierarchy process where a pair-wise comparison is done with all of the pre-calculated metrics and the basis that the additional measures are also introduced. All the attributes are normalized using a standard normalizer where all the metrics are normally distributed data with zero mean and unit variance.

3.5 Machine Learning Algorithm

3.5.1 Supervised learning

Supervised learning method where the past data is used to train the model:

Regression - The output variable used to predict is a continuous variable e.g. runs scored by the batsmen.

Classification - The output variable to predict is a categorical variable e.g. whether or not a player is selected in a team.

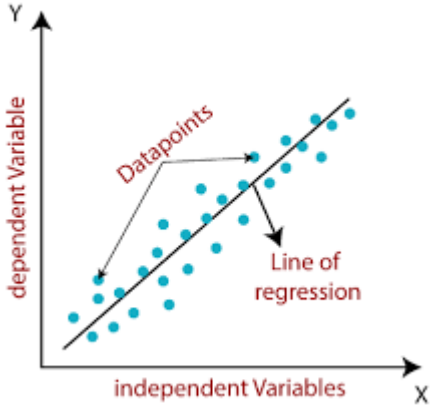
For our study, models such as linear regression, lasso regression, ridge regression, decision tree, and Random decision forest have been used.

3.5.2 Linear Regression

The most elementary regression model is the simple linear regression which explains the linear relationship between the dependent variable (Target variable) and one independent variable using a straight line

$$Y = \beta_0 + \beta_1 X \qquad [3.5]$$

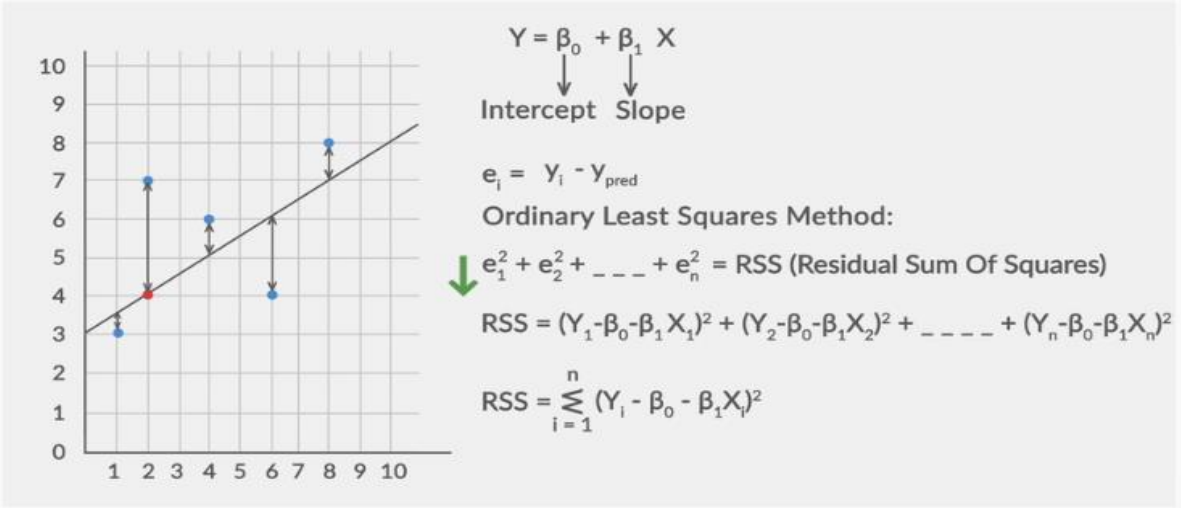
Figure 3.1 Simple linear regression



Best fit Line

The best fit line is identified by minimizing the error term RSS (Residual sum of error) which is the sum of squares of residuals for each point. The residual for a point is identified by subtracting the actual value from the predicted value.

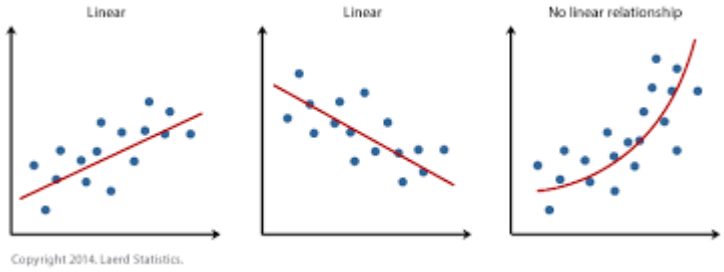
Figure 3.2 Best fit line



Assumption of linear regression

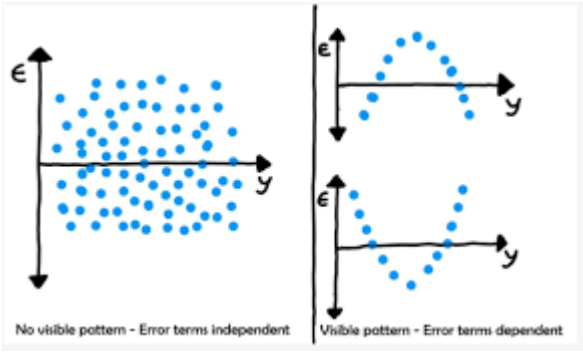
Linearity of residual there should be a linear and additive relationship between the dependent and independent variable. If a linear model is fit to a non-linear, non-additive set, then the regression model will fail to capture the trend, thus resulting in an inefficient model. This will lead to erroneous predictions on unseen data.

Figure 3.3 Linearity of Residual



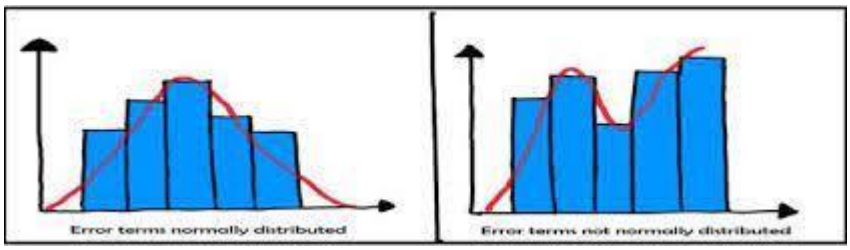
The independence of residual error terms should not be dependent on one another. There should be no correlation between the residual (error) terms.

Figure 3.4 The independence of residual



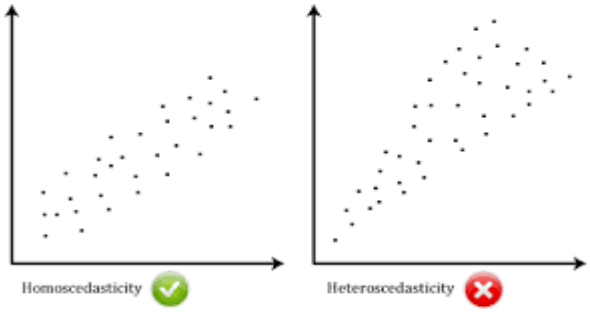
Normal distribution of residual – the mean of residual should follow a normal distribution with a mean equal to zero or close to zero. This is done to check whether the selected line is the line of best fit.

Figure 3.5 The normal distribution of residual



The equal variance of residual – the error term must have a constant variance. This phenomenon is known as homoscedasticity. The presence of non-constant variance is referred to as heteroscedasticity.

Figure 3.6 The equal variance of residual



3.5.3 Multi-linear regression

Multilinear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of the multilinear regression is to determine the best equation which can predict the value of Y using multiple dependent variables

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i \qquad [3.6]$$

Assumptions

All the assumptions made in the linear regression such as Linearity of residual, Independence of residual, Normal distribution of residual, Equal variance of residual holds for multiple linear regression along with few additional assumptions.

Overfitting

When more and more variables are added to a model, the model may become far too complex and usually ends up memorizing all the data points in the training dataset. This phenomenon is known as the overfitting of a model. Overfitting causes the model to become specific rather than generic. This is usually high training accuracy and low test accuracy.

Multicollinearity

It is the phenomenon where a model with several independent variables may have some variables interrelated. When the variables are correlated to each other, they can easily explain each other, and thus, their presence becomes redundant. Multicollinearity can be detected by

1. Pairwise Correlation – Checking.
2. Variance Inflation Factor – Pairwise correlations may not always be useful as one variable cannot completely explain some other variable, but some of the variables combined might be able to do that. Thus, to check these sorts of relations between

variables, one can use VIF. VIF explains the relationship of one independent variable with all the other independent variables.

$$VIF_i = \frac{1}{1 - R_i^2} \qquad [3.7]$$

i refers to the ith variable which is being represented as a linear combination of the rest of the independent variables.

VIF > 10 indicates definite removal of the variable.
VIF > 5 indicates variable needs inspection.
VIF < 5 indicates the variable is good to go.

Once multicollinearity is deducted in the dataset then the following methods can be used to deal with the same:

- A high correlated variable can be dropped,
- The business interpretable variable can be picked,
- New interpretable features can be derived using correlation variables.
- Variable transformation can be transformed by Principal component analysis (PCA) or partial least squares (PLS).

The linear regression models are complex and overfits. To overcome, the phenomenon regualrisation is used to identify the relationship between the complexity of the model with the usefulness in a learning context.

3.5.4 Regularized Regression

A predictive model must be as simple as possible. There is an important relationship between the complexity of a model and its usefulness in a learning context because of the following reasons:

• Simpler models are usually more generic and are more widely applicable (are generalizable)
• Simpler models require fewer training samples for effective training than the more complex ones

Regularization

It is the process used to penalize the complexity of the model i.e., encourage the model to be as simple as possible and perform well on the training data. Usually, an additional regularization term is added to the cost function with the error term, while minimizing the cost function. The regularization term can be as follows.

1. Ridge regression (sum of the square of coefficients)
2. Lasso regression (sum of absolute values of coefficient)

Ridge Regression

In the ridge regression, an additional term of the sum of squares of the coefficients is added to the cost function along with the error term.

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2 + \lambda \sum_{j=0}^{p} w_j^2 \qquad [3.8]$$

Lasso Regression

In the ridge regression, an additional term of the sum of absolute values of the coefficients is added to the cost function along with the error term.

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2 + \lambda \sum_{j=0}^{p} |w_j| \qquad [3.9]$$

When there are large number of features with less data-sets (with low noise), **linear** regressions may outperform Decision **trees**/random forests. In general cases, Decision **trees** will be having **better** average accuracy. For categorical independent variables, decision **trees** are **better than linear regression**.

Decision Tree

A decision tree creates regression or classification models using a tree data structure. It incrementally splits the dataset into smaller groups while simultaneously developing a decision tree. A tree containing decision nodes and leaf nodes is the result. Each branch of a decision node represents a value for the characteristic under consideration. A numerical target choice is represented by a leaf node. The best predictor is represented by the root node, which is the

topmost decision node in a tree. Both category and numerical data can be handled by decision trees.

Algorithm

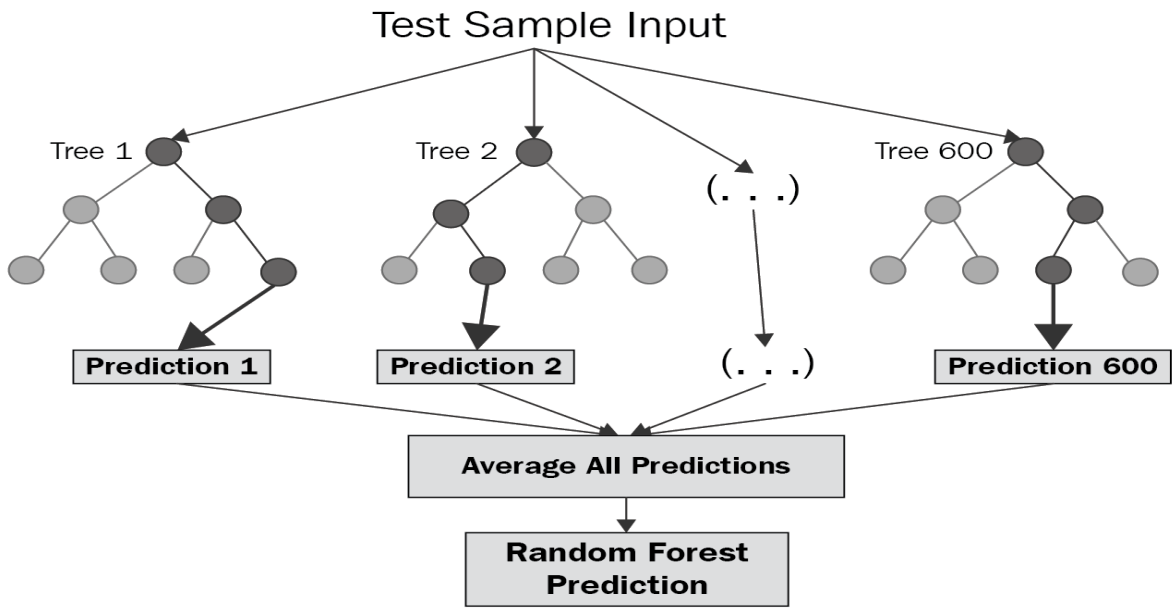
J. R. Quinlan's ID3 technique for generating decision trees uses a top-down, greedy search through the space of feasible branches with no backtracking. By replacing Information Gain with Standard Deviation Reduction, the ID3 method can be used to build a decision tree for regression.

Decision trees are very dependent on the data they are trained on. If the training data is modified, the decision tree that results can be considerably different, as may the predictions. Furthermore, they are computationally expensive to train, have a high risk of overfitting, and tend to discover local optima because they can't go back after splitting. The ability to overcome these flaws is due to the effectiveness of integrating many decision trees into a single model known as random forest.

3.5.5 Random Forest

It is a Supervised Learning technique that employs the ensemble learning method for classification and regression models. It is a bagging method rather than a boosting method. These trees run parallel in the random forest algorithm. During the creation of the trees, there is no interaction between them. It is a meta-estimator (i.e., it integrates the results of numerous forecasts) that mixes numerous decision trees with certain useful alterations. At each node, the number of characteristics that can be divided is limited to a certain fraction of the total known as a hyperparameter. When generating splits, each tree takes a random sample from the original data set, adding a layer of unpredictability that inhibits overfitting. These changes assist to keep the trees from becoming overly linked.

Figure 3.7 Structure for random forest



Random Forest has the following features and benefits:

- It generates a very accurate classifier for various data sets.
- Works well with huge databases.
- Can handle numerous input variables without deleting any of them.
- It calculates the importance of several variables in the classification.
- It offers a method for guessing the missing data that works well and retains accuracy even when a large part of the data is missing.

Random Forest Drawbacks:

- For some datasets with noisy classification/regression tasks, random forests have been found to overfit.
- Random forests are biased when data includes categorical variables with differing numbers of levels. As a result, random forest variable significance scores are unreliable for this type of data.

3.5.6 Bagging

Bagging or bootstrapped sampling is an ensemble method. It is used when the goal is to reduce the variance of the algorithm. Bootstrapping means creating several random subjects of the data (about 30% -70%) from training samples chosen randomly from replacement. Each collection subset data is then used to build their tree using a random sample of features while splitting a

node. Aggregation implies combining the result of different models, resulting in an ensemble of different models. Average of all the predictions from different models are then used which is robust than a single model. Bagging is just a sampling technique not specific to the random forest.

3.5.7 Boosting

Boosting is an ensemble learning technique to build strong regression from several weak learners in series. Boosting also plays a crucial role in the bias-variance trade-off, unlike bagging, the boosting takes care of both variance and bias in a more effective way.

These are a few boosting algorithms:

- Gradient Boost
- XGBoost
- Catboost

Gradient Boost

Gradient boosting is one of the powerful techniques for building predictive models. The main objective of gradient boosting is to minimize the loss function by using a gradient descent optimization algorithm. Gradient boosting is the greedy algorithm and can overfit a training set very quickly.

Gradient boosting works on three components:

Loss function - the role of the loss function is to identify how well the data is predicting the output with the test data.

Weak learner - the weak learners are the models which regress the data poorly compared to its random guess.

Additive model – the process of adding the weaker model in each iteration to reduce the loss function until the model no longer improves on the external validation.

Gradient boosting can be improved by subsampling, shrinkage, and early stop.

XGBoost

The XGBoost is an extension to the gradient boosting decision tree and is specially designed to improve speed and performance.

XGBoost features:

Regularized learning – the regularization allows the model to smooth the final learned weights to avoid overfitting. The regularized will tend to select the model with simple and predictive functions.

Gradient Tree Boost – The tree model cannot be optimized using the traditional optimization technique, instead the model is trained in an additive model.

Shrinkage – after each iteration the model is shrunk by the factor 'n', the shrinkage reduces the influence of each tree and leaves space for further trees to improve the model.

XGBoost is a faster algorithm compared to other algorithms because of its parallel and distributed computing. XGBoost dominates the structural and tabular dataset on regression and classification models. It repetitively leverages the patterns in residuals, strengthens the model with weak predictions, and make it **better**. By combining the advantages from both **random forest** and gradient boosting, **XGBoost** gave a prediction error ten times lower **than** boosting or **random forest** in my case.

Catboost

Catboost is a relatively new open-source machine learning algorithm developed by Yandex in 2017. One of Catboost's core edge is its ability to integrate a variety of different data types such as image, audio, and text features into one data frame. The idiosyncratic way of handling the categorical variables will require a minimum transformation of categorical variables.

Catboost builds upon the theory of decision and gradient boosting. The main idea of boosting is to sequentially train many weak models and through greedy search creates a strong competitive model. The gradient boosting fits the tree sequentially, the fitted trees will learn from the mistakes of the former trees and reduce the error, this process of adding new functions is continued until no further minimization is possible. Catboost grows on oblivious trees, which means that the trees are imposing the rules that all nodes at the same level test the same predictor with the same condition. Catboost offers immense flexibility with its approach of handling sparse and categorical variables with fast training time.

3.6 Evaluation Metrics

The strength of the model can be assessed by

1. R^2 or coefficient of determination

2. Mean Squared Error

R^2 or coefficient of determination

The accuracy of the model is predicted using the R^2 statistics. R^2 explains the variation of the actual data with predicted data. The value R^2 always lies between 0 – 1. It explains how well the actual output is close/similar to the model. The R2 is represented mathematically by

$$R^2 = 1 - \frac{RSS}{TSS} \qquad [3.10]$$

RSS = Residual sum of squares

TSS = Total sum of squares

RSS (residual sum of squares) – It is defined as the sum of squares of actual value with predicted value, lower the RSS value better the model fit.

$$RSS = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \qquad [3.11]$$

TSS (total sum of squares) – It is defined as the sum of squares of actual value with the mean value.

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad [3.12]$$

Root Mean Square Error

Root Mean square error measures the absolute goodness of the fit. MSE is calculated by the sum of squares of prediction error which is the real output subtracted from the predicted output and divided by the number of points.

$$RMSE = \sqrt{\frac{1}{N}\left(\sum_{i=1}^{n}\left(y_i - y_i\right)^2\right)} \qquad [3.13]$$

Information Criterion

The IC or information criterion is derived from the field of frequentist and Bayesian probability. Any selection method scoring lowest means less information is lost and hence a best model.

It is advised to maximize the likelihood by adding more parameters which may end up in making model more complex and over-fitted. Thus, AIC/BIC add a penalty for additional parameters. That's how it's maintaining balance.

Akaike Information Criterion (AIC)

The relationship between the maximum likelihood and measure of information loss

$$AIC = 2K - \log(L) \qquad [3.14]$$

k= number of independent variables to build model

L= maximum likelihood estimate of model

Bayesian Information Criterion

Bayesian information criterion (BIC) is a criterion for model selection among a finite set of models. It is based, in part, on the likelihood function, and it is closely related to AIC.

When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. The BIC resolves this problem by introducing a penalty term for the number of parameters in the model. The penalty term is larger in BIC than in AIC.

BIC has been widely used for model identification in time series and linear regression. It can, however, be applied quite widely to any set of maximum likelihood-based models

$$BIC = \log(n)K - 2\log(L) \qquad [3.15]$$

k= number of independent variables to build model

L= maximum likelihood estimate of model

n = sample size (#observations)

log-base = e(natural log)

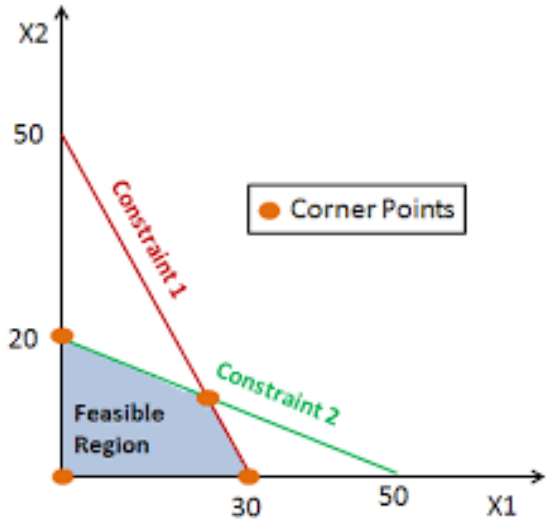
3.7 Linear Programming

Linear programming or linear optimization is the best outcome in a mathematical model where the objective function and the constraints have a linear relationship.

To formulate the linear programming

- Decision variable – the unknown variables
- Objective function – the linear function represents the quantities to be maximized or minimized.
- Constraints – the system of equalities or inequalities representing the restrictions on decision variables.
- Non-negativity restrictions – the value of the decision variable should be greater than or equal to zero.

Figure 3.8 Objective function for linear programming (www.datacamp.com)



Optimization is crucial for investigating complex issues, identifying and solving problems, and data-driven decision-making. Many practical problems can be modeled and solved with mathematical optimization techniques, where inputs and constraints are obtained from the machine learning model's predictions.

3.8 Summary

The IPL (2008 – 2021) data has information about the 845+ matches played so far. The data does not consider the super over played by the players in this study. The IPL data is aggregated and preprocessed. Individual datasets are created for the batting, bowling, and all-rounder.

The player's performance measures vary depending on their roles. Using the data, the batting measures such as innings, runs scored, balls faced, no. of 4s, no. of 6s, no. of not out, strike rate, average, bowling measures like no of wickets taken, no of instances where the bowler has taken 3, 4 and 5+ wickets has been taken, bowling average, economy, strike rate, etc. The combination of batting and bowling is combined in the all-rounder dataset. A brief introduction summary about the machine learning algorithm gives an understanding of the methods that are going to be used in the study and the evaluation metrics to evaluate the performance of these models. The accuracy defines how well the test data performs based on the trained model, apart from the accuracy the evaluation metrics such as RMSE, AIC and BIC are taken into consideration. The Model with high R^2 value, low RMSE, AIC and BIC. The lower these numbers better the model.

The random forest, Catboost, and XGBoost algorithm are used to predict the fantasy points scored by the players in the match where ‘Chennai Super Kings’ is playing against ‘royal challengers Bangalore’ in ‘Wankhede stadium, Mumbai’ and the output is then fed into the linear programming model to identify the best fantasy team with the given constraints.

CHAPTER 4

ANALYSIS

4.1 Introduction

Based on the methodology, the data is preprocessed by checking for the duplicate values in each of the columns. The first and second innings of the match are taken into consideration for this study. The super over statistics is not taken into consideration for this study.

The data aggregations from the IPL data are aggregated into three different categories batting, bowling, and all-rounders. Multi-collinearity is a case where a model with several independent variables, may have some variables highly correlated. When the variables are correlated to each other, they can easily explain each other. The presence in the model becomes redundant. Pairwise correlation and variance inflation factor VIF can be used to identify multicollinearity.

Machine learning models usually get more complex and usually end up with memorizing all the data points in the training dataset. This phenomenon is known as overfitting of a model. Overfitting causes the model to become specific rather than generic. This is can be observed from the metrics as models exhibit high training accuracy and low test accuracy. To overcome the overfitting, the data is split into three sets: train, validation and test. The train data is to train the model and the validation data is used to validate the model for overfitting if the training and the validation accuracy is close which means the model is trained properly.

The processed data are trained using various regression models like linear regression, ridge, lasso, random forest, Catboost and XGBoost. The results are validated by accuracy, root mean square error, Akaike information criterion, Bayesian information criteria. The relation between the player's metrics and the fantasy points is identified.

The fantasy team is selected for the match between 'Chennai Super Kings' and 'Royal Challengers Bangalore' in 'Wankhede Stadium, Mumbai'. The ball-by-ball data is used to predict the fantasy points of every player of the squad and using the linear programming the best fantasy team is chosen. The constraints on the team selection are minimum 4 players from both the teams playing, 3-7 batsmen including maximum of 3 wicket keeper, 1-3 all-rounders and 3-5 bowlers with the total cost of the team not exceeding 100.

4.2 Data Preprocessing

The duplicate values in the ‘venue’, ‘batting team’ and ‘bowling team’ columns have been harmonized to their respective values e.g., ‘Rising Pune Supergiant’ and ‘Rising Pune Supergiants’ are harmonized to a common value. The ‘innings’ column value 1 and 2 is taken into consideration for this project, the other values that correspond to the super over has not been taken into consideration. Columns like date of the match ‘start date’, ‘other wicket type’ and ‘other player dismissed’ are dropped from the data.

4.3 Data Aggregation

The ball-to-ball data is used to prepare three categories: batting, bowling and all-rounder. These three categories are prepared as follows:

4.3.1 Batting Dataset

The ball-to-ball data should be converted into an aggregated format so that the player’s performance against an opponent in a given venue can be identified. The data is aggregated on 'venue’, ‘City’, ‘batting team’, ‘bowling team’ and ‘striker’ by calculating these metrics such as ‘innings’, ’runs scored’, ‘balls faced’, ‘4s’, ‘6s’, ‘dots’, ‘30+’, ‘50+’, ‘100+’, ‘outs’, ‘not out’ and ‘maximum score’.

Value Computation

Three columns have been calculated from the aggregated data:

Average = Runs Scored / Not out [3.1]

Strike rate = Runs Scored / Balls faced [3.2]

Table 4.1 Dream 11 points for batting

Features	Points
Runs	1
4s	1
6s	2
30+	4
50+	8
100+	16
Ducks	-2

Points = (runs *1) + (4s *1) + (6s *2) + (30+ * 4) + (50+ * 8) + (100+ *16) – (ducks *2) [4.1]

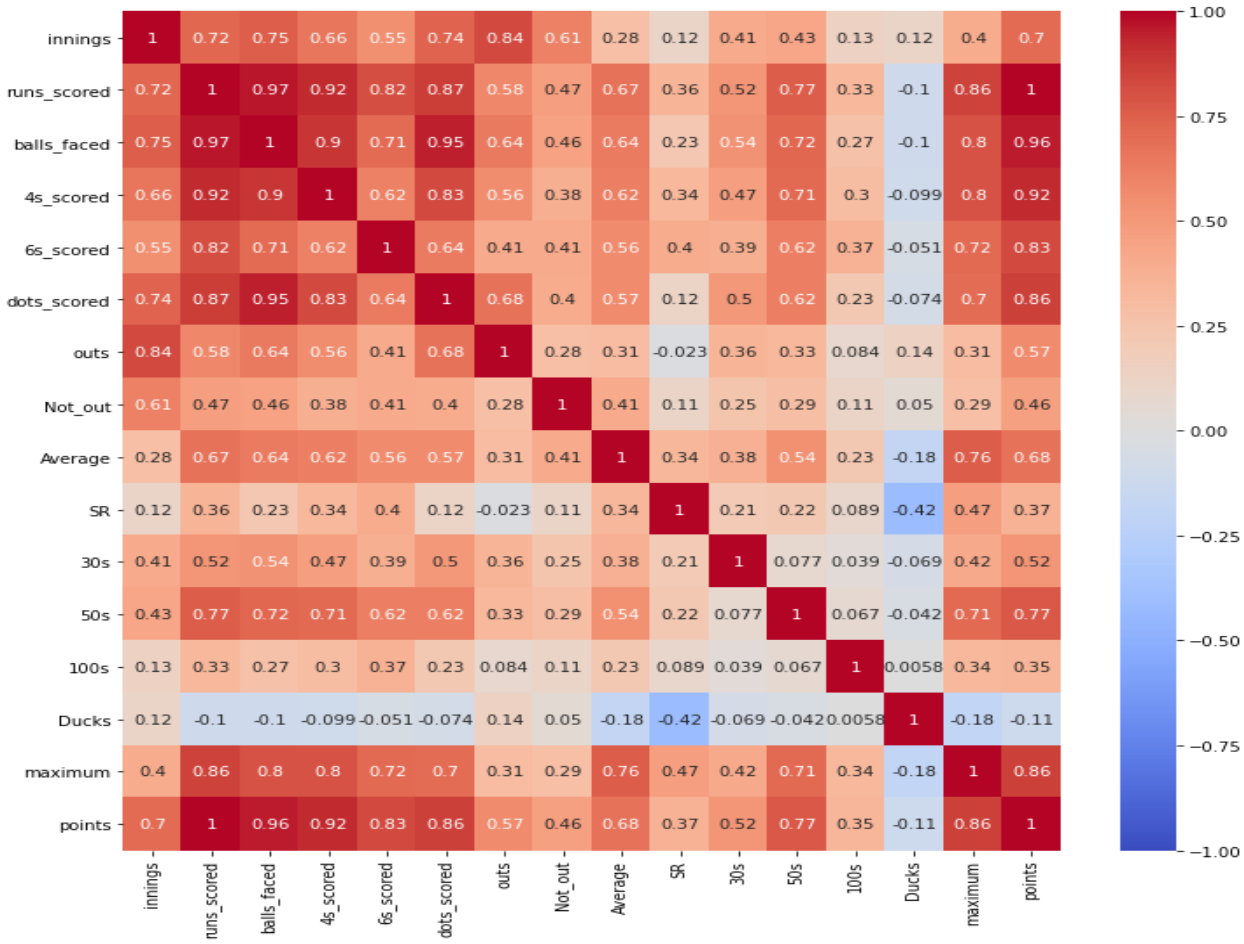
Missing Value

The data had some missing values in ‘4s’, ‘6s’, ‘dots’, ‘30+’, ‘50+’, ‘100+’, ‘Strike Rate’, ‘Average’, ‘outs’, ‘not out’ and the missing data have been replaced by zero. The average had an ‘infinite’ value which has been replaced by 0.

Pairwise Correlation

The pairwise correlation between different pairs of independent variables can throw useful insights into multicollinearity. The pairwise correlation alone will not be able to reduce multicollinearity.

Figure 4.1 Pairwise correlation for batsmen



Variance Inflation Factor

The variance inflation factor (VIF) is calculated to identify all the features which will be enough to encounter multicollinearity. The value of VIF for all features needs to be less than 5 so, to obtain the features such as ‘runs scored’, ‘balls faced’, ‘innings’, ‘maximum’, ‘4s’, and ‘dots’ are removed.

Table 4.2 VIF features for batting

Features	VIF
6s	3.59
Average	3.14
outs	2.56
50+	2.49
Strike Rate	2.34
30+	1.78
Not Out	1.4
100+	1.32
Ducks	1.19

The ‘6s’, ‘Average’, ‘Outs’, ‘50+’, ‘Strike Rate’, ‘30+’, ‘Not Out’, ‘100+’ and ‘Ducks’ features have been considered for the modeling.

Data Split

The batting data is split into three sets which are train, validation, and test in 80%, 10%, and 10% respectively. Train data is used to train the model and validation data is used to validate the model to reduce the overfitting. The Target variable is the points.

4.3.2 Bowling Dataset

The ball-to-ball data should be converted into an aggregated format so that it can be used to identify the bowler’s performance against an opponent in a given venue. The data is aggregated on 'venue’, ‘City’, ‘batting team’, ‘bowling team’ and ‘striker’ by calculating these metrics such as ‘runs_conceeded’, ‘runs_conceeded_balls’, ‘extras’, ‘extras_balls’, ‘byes’, ‘byes_balls’, ‘legbyes’, ‘legbyes_balls’, ‘total_runs’, ‘total_balls’, ‘wickets’, ‘maiden’.

Value Computation

Four columns have been calculated from the aggregated data:

Average = Total balls bowled/ Wickets [3.3]

Strike rate = Total Runs conceded / Wickets [3.4]

Economy = Total Runs conceded * 6 / Total balls bowled [4.2]

Table 4.3 Dream 11 points for bowling

Features	Points
wickets	25
wickets (lbw or bowled)	8
3 wickets	4
4 wickets	8
5+ wickets	16
maiden	12

Points = (wickets *****25) **+** (wickets_lbw_bowled ***** 8) **+** (‘3+’ ***** 4) **+ (‘**4+’ ***** 8) **+** (‘5+’ ***** 16) **+** (maiden ***** 12) [4.3]

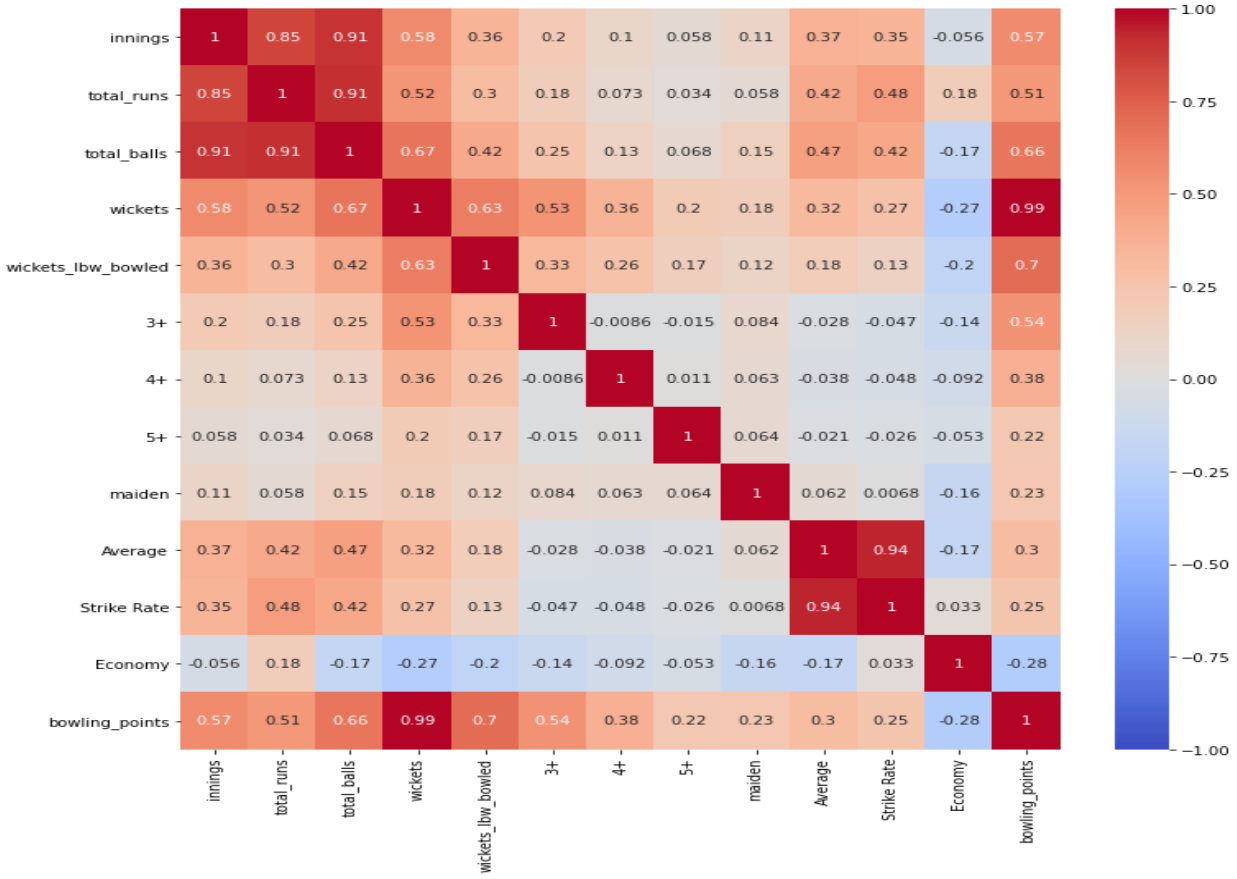
Missing Value

The data had some missing values in ‘wickets’, ‘wickets_lbw_bowled’, ‘dots’, ‘3+’, ‘4+’, ‘5+’, ‘maiden’, ‘Bowling Average’, ‘Strike Rate’, ‘Economy’, the missing data have been replaced by zero. The Bowling Average’ had ‘infinite’ values which have been replaced by 0.

Pairwise Correlation

The pairwise correlation between different pairs of independent variables can throw useful insights into multicollinearity. The pairwise correlation alone will not be able to reduce multicollinearity.

Figure 4.2 Pair wise correlation for bowlers



Variance Inflation Factor

The variance inflation factor (VIF) is calculated to identify all the features which will be sufficient to encounter multicollinearity. The value of VIF for all features needs to be less than 5 so, to obtain the features such as ‘total balls bowled’, ‘bowling average’, ‘innings’, ‘total runs conceded’ are removed.

The ‘wicket’, ‘bowling strike rate’, ‘wickets (lbw or bowled)’, ‘3 wickets’, ‘Economy’, ‘5+ wicket’, ‘maiden’ features have been considered while modeling.

Table 4.4 VIF features for bowling

Features	VIF
Wickets	4.61
bowling strike rate	2.19
wickets (lbw or bowled)	2.04
3 wickets	1.81
Economy	1.78
4 wickets	1.35
5+ wicket	1.11
maiden	1.07

Data Split

The batting data is split into three sets which are train, validation, and test in 80%, 10%, and 10% respectively. Train data will train the model and validation data will validate the model this will reduce the overfitting. The Target variable is the points.

4.3.3 All-rounder Dataset

The ball-to-ball data should be converted into an aggregated format so that it can be used to identify the all-rounder performance against an opponent in a given venue. The data is aggregated on 'venue', 'City', 'batting team', 'bowling team' and 'player' by calculating these metrics such as 'innings', 'runs scored', 'balls faced', '4s', '6s', 'dots', '30+', '50+', '100+', 'outs', 'not out' and 'maximum score', 'runs_conceeded', 'runs_conceeded_balls', 'extras', 'extras_balls', 'byes', 'byes_balls', 'legbyes', 'legbyes_balls', 'total_runs', 'total_balls', 'wickets', 'maiden'.

Value Computation

Total Points = (runs *1) + (4s *1) + (6s *2) + (30+ * 4) + (50+ * 8) + (100+ *16) – (ducks *2) + (wickets *25) + (wickets_lbw_bowled * 8) + ('3+' * 4) + ('4+' * 8) + ('5+' * 16) + (maiden * 12) [4.4]

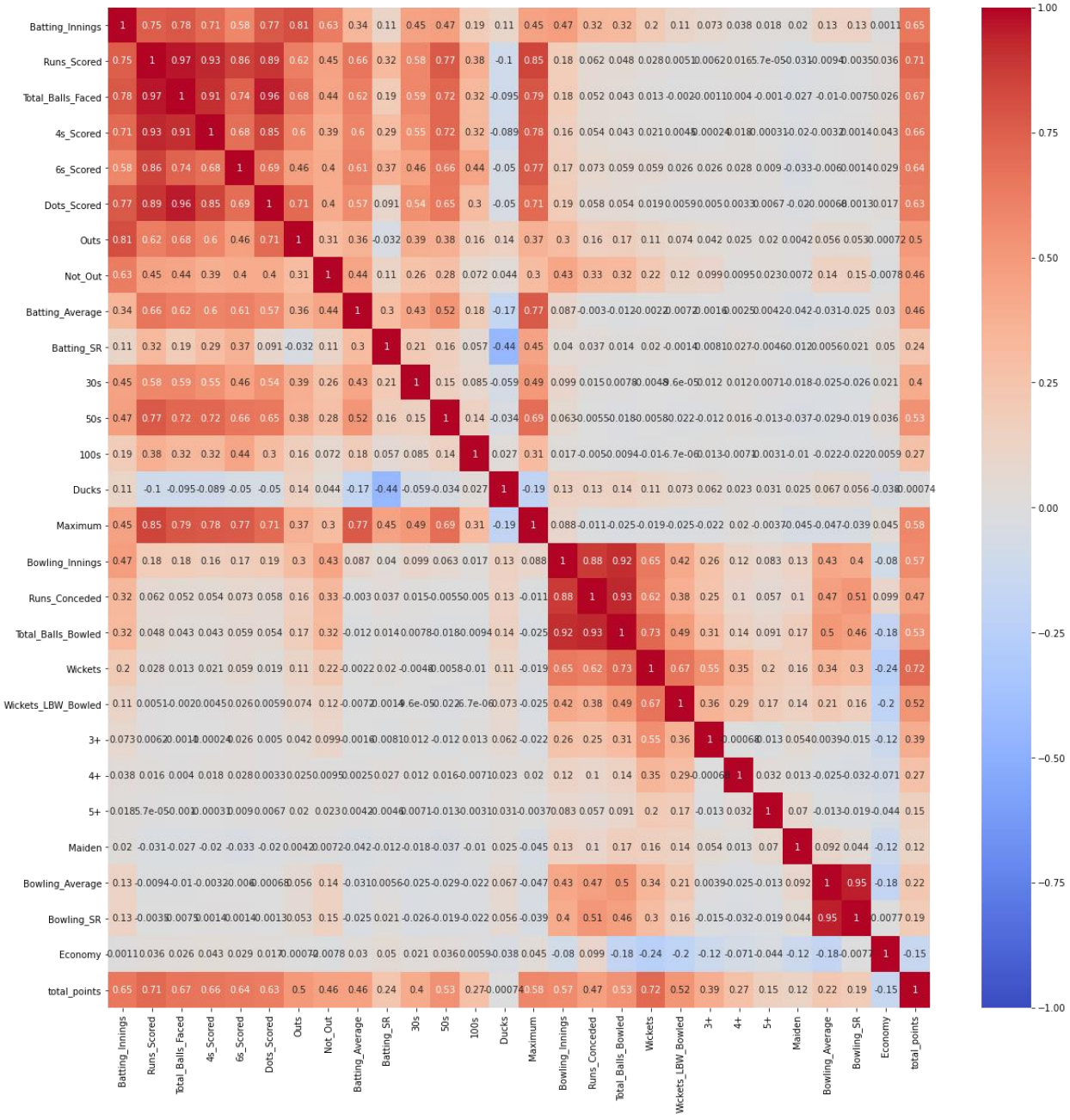
Missing Value

The data had some missing values in ‘4s’, ‘6s’, ‘dots’, ‘30+’, ‘50+’, ‘100+’, ‘Strike Rate’, ‘Average’, ‘outs’, ‘not out’ and the missing data have been replaced by zero. The average had an ‘infinite’ value which has been replaced by 0.

Pairwise Correlation

The pairwise correlation between different pairs of independent variables can throw useful insights into multicollinearity. The pairwise correlation alone will not be able to reduce multicollinearity.

Figure 4.3 Pairwise correlation for all-rounders



Variance Inflation Factor

The variance inflation factor (VIF) is calculated to identify all the features which will be enough to encounter the multicollinearity. The value of VIF for all features needs to be less than 5 so, to obtain the features such as ‘total balls bowled’, ‘bowling average’, ‘innings’, ‘total runs conceded’ are removed.

Table 4.5 VIF features for all-rounders

Features	VIF
Wickets	4.95
6s_Scored	4.56
Economy	4.09
Batting_SR	4.07
Outs	2.82
Batting_Average	2.79
50s	2.38
Wickets_LBW_Bowled	2.22
Bowling_SR	2.07
3+	1.84
30s	1.76
Not_Out	1.52
100s	1.43
Ducks	1.42
4+	1.32
5+	1.1
Maiden	1.07

These ‘wicket, ‘6_scored’, ‘economy’, ‘batting strike rate’, ‘outs’, ‘50+ score’, ‘30+ score’, ‘100+ score’, ‘ducks’, ‘batting average’, ‘not out’, ‘bowling strike rate, ‘wickets (LBW or bowled), ‘3 wickets’, ‘4 wickets’ ‘5+ wicket’, ‘maiden’ features have been considered for the modeling.

Data Split

The batting data is split into three sets which are train, validation, and test in 80%, 10%, and 10% respectively. Train data will train the model and validation data will validate the model this will reduce the overfitting. The Target variable is the points.

4.4 Linear Programming

Once the modeling is performed to predict the maximum points. The trained model is fed into the ball-to-ball data and predicts the dream 11 scores based on the playing team and opposition. In this study, the dream 11 team for 'Royal Challengers Bangalore' vs 'Chennai Super Kings' in 'Wankhede Stadium, Mumbai' has been predicted. 34 players from both the teams are fed into 'Catboost', 'XGBoost', and Random Forest

Linear Programming Algorithm

Objective: To select a team of 11 players out of a squad of 22 players such that it maximizes the Dream 11 points scored in a game

Maximize

$$\text{Total Points} = \left(\sum_{i=1}^{22} \text{Player}_i * \text{RunShare}_i * \text{TotalRuns} * \text{RunPoints}_i + \text{Player}_i * \text{WicketsShare}_i * \text{TotalWickets} * \text{WicketsPoints}_i\right)(N_i)$$

Such That –

Total cost of purchasing the players -

$$\sum_{i=1}^{22} Player_i * \text{Cost}_i < \text{MAXCOSTLIMIT}\ (100) \quad [4.5]$$

Constraint on the type of players who can be selected: Batsmen, All Rounders and Bowlers

$$\text{MINBATSMEN} \leq \sum_{i=1}^{22} Player_i * Batsmen_i \leq \text{MAXBATSMEN} \quad [4.6]$$

$$\text{MINALLROUDNERS} \leq \sum_{i=1}^{22} Player_i * Allrounder_i < \text{MAXALLROUNDERS} \quad [4.7]$$

$$\text{MINBOWLERS} \leq \sum_{i=1}^{22} Player_i * Bowler_i < \text{MAXBOWLERS} \quad [4.8]$$

N_i = 2 if player$_i$ has max points,
N_i = 1.5 if player$_i$ has max points, otherwise [4.9]
N_i = 1

$$RunPoints_i = Blended\ points\ per\ run\ scored\ based\ on\ historical\ strike\ rate \quad [4.10]$$

$$WicketsPoints_i = Blended\ points\ per\ wickets\ taken\ based\ on\ historical\ economy\ rate \quad [4.11]$$

One way to do this is solving it using a mix integer linear programming method. Slightly complicated and take time to code, however, a brute force algorithm is used and an approximation of the results has also been tried. In that case, a maximum of C (22,11), 705,4312 iterations are required. To estimate the objective function, the following equations are used:

$$RunShare_i = \frac{e^{RunPotential_i}}{\sum_{i=1}^{22} e^{RunPotential_i}} \quad [4.12]$$

$$WicketsShare_i = \frac{e^{WicketPotential_i}}{\sum_{i=1}^{22} e^{WicketPotential_i}} \quad [4.13]$$

$$RunPotential_i = \sum_j \gamma_j * PlayerFeature_j \quad [4.14]$$

$$WicketPotential_i = \sum_j \mu_j * PlayerFeature_j \quad [4.15]$$

To estimate the run potential and wicket potential of a player, the historical performance of a player, batting position and impact of other players in the team are considered:

$$TotalRuns = \sum_j \alpha_j * RunFeature_j \quad [4.16]$$

$$TotalWickets = \sum_j \beta_j * WicketFeature_j \quad [4.17]$$

4.5 Summary

The duplicate values such as 'rising pune supergiant' and 'rising pune supergiants' have been harmonized to a common name. The data is harmonized on the venue, batting team, and bowling team. The missing value such as 'NA' is replaced with 0 in the numerical column and replaced with 'all others' in the categorical column. The data aggregation for batting, bowling, and all-rounders was created using the suitable player metrics respectively.

The pairwise correlation and the variance inflation factor were calculated for each of the datasets and only the metrics with VIF less than 5 were considered. The batting metrics which are considered for the modeling are '6s', 'Average', 'Outs', '50+', 'Strike Rate', '30+', 'Not Out', '100+' and 'Ducks'. The bowling metrics which are considered for modeling are wicket, 'bowling strike rate', 'wickets (lbw or bowled), '3 wickets', 'Economy', '5+ wicket', 'maiden'. The all-rounder metrics which are considered for modeling are 'wicket, '6_scored', 'economy', 'batting strike rate', 'outs', '50+ score', '30+ score', '100+ score', 'ducks', 'batting average', 'not out', 'bowling strike rate, 'wickets (LBW or bowled), '3 wickets', '4 wickets' '5+ wicket',

‘maiden’. The model is trained using these features with the fantasy point for each of the categories.

Random forest, Catboost, and the XGBoost model are used to predict the fantasy team for the match against ‘Chennai Super Kings’ and ‘Royal Challengers Bangalore’ in ‘Wankhede stadium, Mumbai’. The model predicts the fantasy point that a player will score in the match and the output of the model for 34 players was fed into the linear programming model to predict the 11 fantasy team players for each of the models individually.

CHAPTER 5

RESULTS AND DISCUSSION

5.1 Introduction

The linear regression, lasso, ridge, random forest, Catboost, and XGBoost models have been trained individually on each dataset. The results of these models are evaluated based on the R^2, RMSE, AIC, and BIC scores.

Random forest, Catboost, and the XGBoost model are used to predict the fantasy team for the match against 'Chennai Super Kings' and 'Royal Challengers Bangalore' in 'Wankhede Stadium, Mumbai'. The model predicts the fantasy point that a player will score on the match and the output of the model for 34 players was fed into the linear programming model to predict the 11 fantasy team players for each of the models individually.

5.2 Modelling

5.2.1 Batsmen's Modelling

The batting pre-processed data is modeled using different models like linear regression, lasso regression, ridge regression, random forest, Catboost regression, and XGBoost regression. The evaluation metrics are R^2, RMSE, AIC, and BIC scores.

Table 5.1 Evaluation metrics analyzes for batsmen

Batsmen Model	R^2	RMSE	AIC	BIC
Linear Regression	0.9488	10.13	4370.44	4423.72
Lasso Regression	0.9487	10.13	4371.02	4424.31
Ridge Regression	0.9485	10.15	4374.56	4427.35
Random Forest Regression	0.9379	11.49	4550.60	4603.90
Catboost Regression	0.7893	20.54	5698.77	5752.06
XGBoost Regression	0.9802	6.297	3477.87	3531.17

The linear regression, lasso regression, and ridge regression have similar evaluation metrics like the R^2 value which is around 94.8% and the root means square error which is 10.13. Random forest performs relatively poorly when compared to linear regression. Models which

use gradient boosting like Catboost exhibit deteriorated performance, whereas XGBoost performs better than all the models with an R^2 value of 98.02 and the root mean square error is 6.297. The AIC and BIC values for XGboost are 3477.87 and 3531.17 respectively.

5.2.2 Bowler's Modelling

The bowling pre-processed data is modeled using different models like linear regression, lasso regression, ridge regression, random forest, Catboost regression, and XGBoost regression. The evaluation metrics are R^2, RMSE, AIC, and BIC scores.

Table 5.2 Evaluation metrics analyzes for bowlers

Bowler's Model	R^2	RMSE	AIC	BIC
Linear Regression	0.999	2.03	-44808.95	-44757.88
Lasso Regression	0.999	0.008	-10805.19	-10754.07
Ridge Regression	0.999	2.26	-16381.48	-16330.41
Random Forest Regression	0.995	2.65	1518.54	1569.60
Catboost Regression	0.889	12.07	4433.80	5673.34
XGBoost Regression	0.995	2.65	1520.01	1575.08

The linear regression, ridge regression, and lasso regression have similar evaluation metrics like the R^2 value which is around 99.9% and the root mean square error for lasso regression is 0.008 and the corresponding value for the other models is 2.03 and 2.26 respectively. Random forest performs relatively poorly when compared to the linear regression, with R^2 = 99.5% and RMSE = 2.65 while the gradient boosting algorithms like Catboost exhibit deteriorated performance with R^2 = 88.9% and RMSE = 12.07 while XGBoost performs similar to random forest with R^2 = 99.5% and RMSE = 2.65. The AIC and BIC values for linear regression are -44808.95 and -44757.88 respectively.

5.2.3 All-Rounder's Modelling

The all-rounder pre-processed data is modeled using different models like linear regression, lasso regression, ridge regression, random forest, Catboost regression, and XGBoost regression. The evaluation metrics such as R^2, RMSE, AIC, and BIC scores.

Table 5.3 Evaluation metrics analyzes for all-rounders

All-rounder's Model	R^2	RMSE	AIC	BIC
Linear Regression	0.975	9.26	1838.99	1883.12
Lasso Regression	0.975	9.26	1838.70	1882.82
Ridge Regression	0.976	9.25	1837.31	1881.44
Random Forest Regression	0.918	17.05	2359.68	2403.81
Catboost Regression	0.657	34.98	2922.87	2966.99
XGBoost Regression	0.982	7.79	1697.91	1742.03

The linear regression, ridge regression, and lasso regression have similar evaluation metrics like the R^2 value which is around 97.5%, and the RMSE which is equal to 9.25. The Random forest is relatively performing poorly when compared to the linear regression with the R^2 = 91.8% and the RMSE = 17.05 adding the gradient boosting like Catboost has deteriorated the performance of the model to R^2 = 65.7% and RMSE = 34.98, whereas, the XGBoost has performed better than all the other models with R^2 = 98.2% and RMSE = 7.79. The AIC and BIC values for XGBoost are 1697.91and 1742.03 respectively.

5.3 Linear Programming

Based on the predictions made on the aggregated data, random forest, Catboost and XGBoost models run on the ball-to-ball data to identify the dream 11 points the player will get and identifies whether or not the player should be considered. The study has considered ‘Chennai Super Kings’ vs ‘Royal Challengers Bangalore’ in ‘Wankhede Stadium, Mumbai’ for the player prediction. The 34 players from the teams were used for identifying the best 11 players with a minimum of 3 and a maximum of 7 batsmen including a wicket-keeper, minimum of 3 and maximum of 5 bowlers, minimum of 1, and a maximum of 3 all-rounders was identified for each of the models with the predicted dream 11 points.

5.3.1 Random Forest

The match data is trained using the random forest model based on all the previous meeting between the team and will evaluate the predicted points scored by the player and the linear programming model identifies the best 11 with all the constraint.

Table 5.4 Fantasy team using random forest

Player	Role	Team	Player Cost (Crores)	Dream 11 pts RF
AB de Villiers	Batsmen	Royal Challengers Bangalore	10	41
F du Plessis	Batsmen	Chennai Super Kings	9	41
Devdutt Padikkal	Batsmen	Royal Challengers Bangalore	9	36
AT Rayudu	Batsmen	Chennai Super Kings	9	35
V Kohli	Batsmen	Royal Challengers Bangalore	10.5	34
SM Curran	All-Rounder	Chennai Super Kings	8.5	41
RA Jadeja	All-Rounder	Chennai Super Kings	8.5	41
MM Ali	All-Rounder	Chennai Super Kings	8	40
L Ngidi	Bowler	Chennai Super Kings	8.5	36
Imran Tahir	Bowler	Chennai Super Kings	9.5	29
YS Chahal	Bowler	Royal Challengers Bangalore	9	28
		Total	99.5	400

The random forest model predicts the dream 11 players. 7 players from ‘Chennai Super Kings’ and 4 players from ‘Royal Challengers Bangalore’. The team contains 3 all-rounders, 5 batsmen including a wicket-keeper and 3 bowlers. The total cost of player is 99.5. The total dream 11 points scored by the team is 400.

5.3.2 XGBoost

The match data is trained using the XGBoost model based on all the previous meeting between the team and will evaluate the predicted points scored by the player and the linear programming model identifies the best 11 with all the constraint.

Table 5.5 Fantasy team using XGBoost

Player	Role	Team	Player Cost in Crores	Dream 11 pts
AB de Villiers	Batsmen	Royal Challengers Bangalore	10	42
F du Plessis	Batsmen	Chennai Super Kings	9	40
Devdutt Padikkal	Batsmen	Royal Challengers Bangalore	9	35
AT Rayudu	Batsmen	Chennai Super Kings	9	34
V Kohli	Batsmen	Royal Challengers Bangalore	10.5	34
SM Curran	All-Rounder	Chennai Super Kings	8.5	42
MM Ali	All-Rounder	Chennai Super Kings	8	41
RA Jadeja	All-Rounder	Chennai Super Kings	8.5	41
L Ngidi	Bowler	Chennai Super Kings	8.5	36
Imran Tahir	Bowler	Chennai Super Kings	9.5	30
M Siraj	Bowler	Royal Challengers Bangalore	8	28
		Total	98.5	405

The XGBoost model predicts the dream 11 players. 7 players from ‘Chennai Super Kings’ and 4 players from ‘Royal Challengers Bangalore’. The team contains 3 all-rounders, 5 batsmen including a wicket-keeper and 3 bowlers. The total cost of player is 98.5. The total dream 11 points scored by the team is 405.

5.3.3 Catboost

The match data is trained using the catboost model based on all the previous meeting between the team and will evaluate the predicted points scored by the player and the linear programming model identifies the best 11 with all the constraint.

Table 5.6 Fantasy team using Catboost

Player	Role	Team	Player Cost in Crores	Dream 11 pts
SM Curran	All-Rounder	Chennai Super Kings	8.5	40
MM Ali	All-Rounder	Chennai Super Kings	8	40
GJ Maxwell	All-Rounder	Royal Challengers Bangalore	9	37
AB de Villiers	Batsmen	Royal Challengers Bangalore	10	42
F du Plessis	Batsmen	Chennai Super Kings	9	42
Devdutt Padikkal	Batsmen	Royal Challengers Bangalore	9	36
V Kohli	Batsmen	Royal Challengers Bangalore	10.5	35
SK Raina	Batsmen	Chennai Super Kings	10	35
L Ngidi	Bowler	Chennai Super Kings	8.5	36
Imran Tahir	Bowler	Chennai Super Kings	9.5	29
M Siraj	Bowler	Royal Challengers Bangalore	8	28
		Total	100	400

The Catboost model predicts the dream 11 players. 6 players from 'Chennai Super Kings' and 5 players from 'Royal Challengers Bangalore'. The team contains 3 all-rounders, 5 batsmen including a wicket-keeper and 3 bowlers. The total cost of player is 100. The total dream 11 points scored by the team is 400.

5.4 Summary

In batting and all-rounder categories the XGBoost model has performed well compared to all the other models such as linear, lasso, ridge, random forest, and Catboost models. The R^2 values are 98.02% and 98.2% respectively. The root mean square error identifies the deviation between the actual with the predicted value, the lower the value the better the model. The root mean square error values for batting and all-rounder categories are 6.297 and 7.79 respectively. In the bowling category lasso, ridge, and linear regression has performed marginally better than

the XGBoost model. The R^2 values are 99.9% but the lasso has performed well based on the RMSE score compared to the ridge and linear regressions. The value for lasso is 0.008 whereas the values for ridge and linear are 2.26 and 2.03 respectively.

The AIC and BIC values for XGboost algorithm for batsmen data are 3477.87 and 3531.17, linear regression for bowler's data are -44808.95 and -44757.88 and BIC values for XGBoost all-rounder's data are 1697.91and 1742.03 respectively.

The random forest, Catboost, and XGBoost model are considered for the prediction of the fantasy team selection and XGBoost has performed relatively better than random forest and Catboost. The team fantasy point predicted by XGBoost is 405 whereas the points scored by random forest and Catboost are 400. The XGBoost predicted the team would yield the maximum points and opportunity to win the match.

XGBoost is performing better than all the models because it executes the sequential tree building using the parallelized implementation. This algorithm has been designed to make efficient use of hardware resources. This is accomplished by cache awareness by allocating internal buffers in each thread to store gradient statistics. Further enhancements such as 'out-of-core' computing optimize available disk space while handling big data-frames that do not fit into memory. The model is also enhanced based on

- Regularization
- Sparsity
- Weighted Quantile Sketch
- Cross validation

CHAPTER 6

CONCLUSION

6.1 Introduction

Selecting the best players plays an important role in a team's victory. The best 11 is selected based on the player's performance and the team balance. The dream 11 team is selected by analyzing the combination of 22 players from either of the playing teams to yield the maximum points.

The dream 11 teams should comprise at least a minimum of 1 and a maximum of 3 wicket-keepers, a minimum of 3 and a maximum of 5 batsmen, a minimum of 1 and a maximum of 3 all-rounders, and a minimum of 3 and a maximum of 5 bowlers.

6.2 Discussion and Conclusion

In this study, the data was divided into three categories: batsmen, bowlers, and all-rounder against opposition and venue. The player's performance was analyzed in each of the categories. Six regression algorithms were used and compared. XGBoost regression turned out to be the most accurate model for all three datasets with an accuracy of 98.0%, 99.9%, and 98.3% for predicting the fantasy points for batsmen, bowlers, and all-rounders respectively.

Results from the Catboost regression were surprising because it achieved an accuracy of just 78.9%, 88.9%, and 65.7% for predicting the fantasy points for batsmen, bowlers, and all-rounders respectively. Basis the modeling on the aggregated data, the models were used to identify the dream 11 team for the match against 'Chennai Super Kings' and 'Royal Challengers Bangalore' in 'Wankhede Stadium, Mumbai'.

11 players were identified based on the fantasy points that they will score on the match and player cost using linear programming with the constraints on the number of batsmen, bowlers, and all-rounders. Three models were used to predict the fantasy score for 34 players from which the best 11 players were identified with the total points the team will get. XGBoost regression model has predicted that the team will score 405 fantasy points, whereas the random forest and Catboost has predicted that the team will score 400 fantasy points.

6.3 Future Scope

The discussed approach can be extended further by including fielding performances like the catches taken, instrumental in run-out wickets, etc. Similar studies can be carried out for other formats of the game i.e. test and one day international. The inclusion of external factors such

as pitch condition, humidity, dew, etc., and player's partnership metrics, the influence of the non-striker, and pressure on the player might be included. The fantasy point can be extended to various other sports like football, baseball, hockey, etc.

Fantasy sport has two types of tournament firstly, the tournament is played daily where the user selects his/her team on a day-to-day basis and have a winner for each of the tournament. The other one is where the user selects a team and gets only a hundred changes that are allowed for the entire tournament. The user with the maximum fantasy points at end of the tournament is the winner.

This study takes into consideration, the first kind of tournament. The team selection can be extended to creating a dashboard where the player can track the win percentage and analyze the pattern of the team selection. The dashboard contains all the information of the matches played by the user. The various information can be used to understand the pattern to play a specific match.

The player recommendation module will help us to gain more profit based on the information available in the dashboard. E.g. if the user has a 90% winning margin when the user plays matches with 'Mumbai Indians' as a team. The dashboard will provide the suggestion for the user based on the available historical data.

The fantasy team selection can be used based on the fantasy app where after completion of each match, the application provides a dream team of the match. Models can be trained based on the dream team information available for every match and predict the fantasy team. The team selected will be purely based on the maximum points scored by the player in the application.

The current study can be extended to the second type of fantasy tournament also.

REFERENCE

1. Wickramasinghe, I. P. (2014) 'Predicting the performance of batsmen in test cricket', *Journal of Human Sport and Exercise*, 9(4), pp. 744–751. doi: 10.14198/jhse.2014.94.01.
2. Shah, D. P. (2017) 'New performance measure in Cricket', *IOSR Journal of Sports and Physical Education*, 04(03), pp. 28–30. doi: 10.9790/6737-04032830.
3. Bhattacharjee, D., Lemmer, H. H, Saikia, H and Mukherjee, Diganta. (2018). 'Measuring performance of batting partners in limited overs cricket', Journal for Research in Sport, Physical Education and Recreation. 40. 1-12.
4. Deep, C., Patvardhan, C. and Singh, S. (2016) 'A new Machine Learning based Deep Performance Index for Ranking IPL T20 Cricketers', *International Journal of Computer Applications*, 137(10), pp. 42–49. doi: 10.5120/ijca2016908903.
5. Barr, G. D. I. and Kantor, B. S. (2004) 'A Criterion for Comparing and Selecting Batsmen in Limited Overs Cricket', Operational Research Society, vol. 55, no. 12, pp. 1266-1274.
6. Lemmer, H. H. (2012) 'Individual Match Approach to Bowling Performance Measures in Cricket', 34(2), pp. 95–103.
7. Muthuswamy S. and Lam, S. S. (2008) 'Bowler Performance Prediction for One-day International Cricket Using Neural Networks', Industrial Engineering Research Conference
8. Saikia, H. et al. (2012) 'A double weighted tool to measure the fielding performance in cricket', *International Journal of Sports Science and Coaching*, 7(4), pp. 699–713. doi: 10.1260/1747-9541.7.4.699.
9. Passi, K. and Pandey, N. (2018) 'Predicting Players' Performance in One Day International Cricket Matches Using Machine Learning', (December), pp. 111–126. doi: 10.5121/csit.2018.80310.
10. Manage, A. B. W., Kafle, R. C. and Wijekularathna, D. K. (2020) 'Classification of all-rounders in limited over cricket - a machine learning approach', *Journal of Sports Analytics*, 6(4), pp. 295–306. doi: 10.3233/jsa-200467.
11. Tyagi, S. *et al.* (2020) 'Enhanced Predictive Modeling of Cricket Game Duration Using Multiple Machine Learning Algorithms', *2020 International Conference on Data Science and Engineering, ICDSE 2020*. doi: 10.1109/ICDSE50459.2020.930081

12. Bhattacharjee, D. and Saikia, H. (2013) 'Selecting the Optimum Cricket Team after a Tournament.', *Asian Journal of Exercise & Sports Science*, 10(2), pp. 77–91.
13. Faez, A. Jindal and K. Deb, "Cricket team selection using evolutionary multi-objective optimization," in International Conference on Swarm, Evolutionary, and Memetic Computing, Berlin, 2011.
14. Amin, G. R. and Sharma, S. K. (2014) 'Cricket team selection using data envelopment analysis', *European Journal of Sport Science*, 14(SUPPL.1). doi: 10.1080/17461391.2012.705333.
15. Iyer, S.R. and Sharda, R., 2009. Prediction of athletes performance using neural networks: An application in cricket team selection. Expert Systems with Applications, 36(3), pp.5510-5522.
16. Omkar, S. N. and Verma, R. (2003) 'Cricket team selection using genetic algorithm', International Congress on Sports Dynamics (ICSD2003), pp. 1–3.
17. Ahmed, F., Deb, K. and Jindal, A. (2013) 'Multi-objective optimization and decision making approaches to cricket team selection', *Applied Soft Computing Journal*, 13(1), pp. 402–414. doi: 10.1016/j.asoc.2012.07.031.
18. Vistro, D. M., Rasheed, F. and David, L. G. (2019) 'The cricket winner prediction with application of machine learning and data analytics', *International Journal of Scientific and Technology Research*, 8(9), pp. 985–990.
19. Kansal, P. et al. (2014) 'Player valuation in Indian premier league auction using data mining technique', Proceedings of 2014 International Conference on Contemporary Computing and Informatics, IC3I 2014, pp. 197–203. doi: 10.1109/IC3I.2014.7019707.
20. Saikia, H. and Bhattacharjee, D. (2011) 'On classification of all-rounders of the Indian premier league (IPL): A Bayesian approach', Vikalpa, 36(4), pp. 51–66. doi: 10.1177/0256090920110404.
21. Kapadia, K. *et al.* (2019) 'Sport analytics for cricket game results using machine learning: An experimental study', *Applied Computing and Informatics*. doi: 10.1016/j.aci.2019.11.006.
22. Akarshe, S., Khade Nikhil Bankar, R. and Khedkar Prashant Ahire Student Professor, P. (2019) 'Cricket Score Prediction using Machine Learning Algorithms', *GRD Journals-Global Research and Development Journal for Engineering*, 5(1), pp. 1–4. Available at: www.grdjournals.com.
23. Basit, A. *et al.* (2020) 'ICC T20 Cricket World Cup 2020 Winner'.

24. Jayalath, K. P. (2017) ‘A machine learning approach to analyze ODI cricket predictors’, *Journal of Sports Analytics*, 4(1), pp. 73–84. doi: 10.3233/jsa-17175.
25. Thenmozhi, D. *et al.* (2019) ‘MoneyBall - Data mining on cricket dataset’, *ICCIDS 2019 - 2nd International Conference on Computational Intelligence in Data Science, Proceedings*, pp. 6–10. doi: 10.1109/ICCIDS.2019.8862065.
26. Somaskandhan, P. *et al.* (2018) ‘Identifying the optimal set of attributes that impose high impact on the end results of a cricket match using machine learning’, *2017 IEEE International Conference on Industrial and Information Systems, ICIIS 2017 - Proceedings*, 2018-Janua, pp. 1–6. doi: 10.1109/ICIINFS.2017.8300399.
27. Aburas, A. A., Mehtab, M. and Mehtab, Y. (2018) ‘Cricket world cup predictions using KNN intelligent bigdata approach’, *ACM International Conference Proceeding Series*, pp. 18–22. doi: 10.1145/3277104.3277117.
28. Ul Mustafa, R. *et al.* (2017) ‘Predicting the Cricket match outcome using crowd opinions on social networks: A comparative study of machine learning methods’, *Malaysian Journal of Computer Science*, 30(1), pp. 63–76. doi: 10.22452/mjcs.vol30no1.5.
29. Kanhaiya, K., Gupta, R. and Sharma, A. K. (2019) ‘Cracked Cricket Pitch Analysis (Ccpa) Using Image Processing and Machine Learning’, 3(1).
30. Vidisha and Bhatia, V. (2020) ‘A review of Machine Learning based Recommendation approaches for cricket’, *PDGC 2020 - 2020 6th International Conference on Parallel, Distributed and Grid Computing*, pp. 421–427. doi: 10.1109/PDGC50313.2020.9315320.
31. Lemmer, H. H., Bhattacharjee, D. and Saikia, H. (2014) ‘A consistency adjusted measure for the success of prediction methods in cricket’, *International Journal of Sports Science and Coaching*, 9(3), pp. 497–512. doi: 10.1260/1747-9541.9.3.497.
32. B. E. Boser, I. M. Guyon and V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers," in Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, 1992.
33. J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques, 3rd Edition ed., Waltham: Elsevier, 2012.
34. Karthik, K. *et al.* (2021) ‘Analysis and Prediction of Fantasy Cricket Contest Winners Using Machine Learning Techniques’, *Advances in Intelligent Systems and Computing*, 1176(September), pp. 443–453. doi: 10.1007/978-981-15-5788-0_43.

35. Das, D. (2014) ‘Moneyballer: An Integer Optimization Framework for Fantasy Cricket League Selection and Substitution’. Available at: http://en.wikipedia.org/wiki/.
36. Singla, S. and Shukla, S. S. (2020) ‘Integer Optimisation for Dream 11 Cricket Team Selection International Journal of Computer Sciences and Engineering Open Access Integer Optimisation for Dream 11 Cricket Team Selection’, (November), pp. 0–6. doi: 10.26438/ijcse/v8i11.16.
37. Patel, N. and Pandya, M. (2019) ‘IPL Players Performance Prediction’, *International Journal of Computer Sciences and Engineering*, 7(5), pp. 478–481. doi: 10.26438/ijcse/v7i5.478481.
38. Perera, H., Davis, J. and Swartz, T. B. (2016) ‘Optimal lineups in Twenty20 cricket’, *Journal of Statistical Computation and Simulation*, 86(14), pp. 2888–2900. doi: 10.1080/00949655.2015.1136629.
39. Bonello, N. *et al.* (2019) ‘Multi-stream data analytics for enhanced performance prediction in fantasy football’, *arXiv*.
40. Stolyarov, A. and Vasiliev, G. (2017) ‘Predict to Succeed: Optimal Sequential Fantasy Football Squad Formation Using Machine Learning Tools’, pp. 1–3.
41. Dwyer, B. and Drayer, J. (2010) ‘Fantasy Sport Consumer Segmentation: An Investigation into the Differing Consumption Modes of Fantasy Football Participants’, *Sport Marketing Quarterly*, 19(April), pp. 207–216.
42. Naha, S. (2019) ‘Flight of fantasy or reflections of passion? Knowledge, skill and fantasy cricket’, *Sport in Society*, 0(0), pp. 1–13. doi: 10.1080/17430437.2019.1607012

APPENDIX – B

Predicting Best Playing 11's for IPL T20 Cricket using Machine Learning

SIVAKALYAN S

LJMU STUDENT ID : 944621

MSC DATA SCIENCE

Research Proposal

SUPERVISOR : AAKARSH MALHOTRA

Table of Contents

List of Tables

List of Equation

Abstract

Cricket is one of the most admired games played all around the globe. A sport that is played between two opponent teams each of them having 11 players, a combination of batsmen, the players who bat, bowlers, the players who can bowl and all-rounders, the players who can do both. The game is generally played in three formats such as Test, One Day International and T20. With increase in the popularity of the T20 format, the game play has become complicated and therefore it becomes necessary to devise new batting and bowling techniques in a very short span of time due to the limited time available for the players to adapt to the changing match situations. The batsmen play a very crucial role by scoring the as many runs as in an innings and the bowlers are expected to restrict the batsmen from scoring runs, either by dismissing batsmen or containing the batsmen from scoring runs. The captain, coach and the team management find it difficult to identify the best playing 11 from the squad of 15 to 17 players. The best playing 11 is selected based on the player's performance against the opposition team in the venue of the match. The player's performance is measured using various metrics. This thesis predicts the best playing 11 for an IPL team. The player performance metrics are calculated using the IPL dataset. Using the player performance metrics, we use Logistic regression, Naive Bayes, Random Forest, SVM etc. model to predict a value between 0 – 1. Higher the value better the player. Basis these values the best playing 11 which consists of 5 – 6 batsmen, 1-2 all-rounder and 4-5 bowler is selected with maximum of 4 foreign players.

1. Introduction

Cricket is an interesting and popular sport played between two opponent teams each consisting of 11 players in which one team chooses to bat and the other to field (Bowl) and one such session is called an innings. The game is played on an oval or round shaped ground and in the middle of the ground there lies a 22-yard-long pitch where the actual game is played. At both the extremes of the pitch, wickets along with three wooden stumps with two bails on top are kept.

Each team consists of 11 players, all the 11 players from bowling and 2 players from the batting team will be actively involved in any particular time of the match. The batting team needs to score maximum runs possible in end of their innings. The bowling team needs to restrict the batting team from scoring runs. The bowling team places themselves in different position around the pitch to restrict the batsmen from scoring runs. After the end of innings, the batting and bowling team will swap their roles. The team who scored the maximum runs in the end of the innings is considered as the winner of the match. The end of innings is considered when either of 10 wickets from the batting team is dismissed or allocated overs for the innings gets completed. An over consist of 6 legal balls.

The batsmen can score runs by hitting a boundary, by hitting the ball over the boundary line in air or ground if the balls cross the boundary line without touching any of the playing area its considered as a 6 or its considered as 4. The other way of scoring runs is to place the between the fielders and reach the opposite end of the pitch. If the batsman able to cross the pitch once its considered as a solitary run, he can run as much as possible before the fielding team fields the ball. An over consists of 6 legal balls, the ball is considered as legal when he bowls with a stipulated area of the pitch. The bowler can dismiss the batsmen in different ways such bowled where the bowls hits the wicket when the batsmen tries to score runs, caught when the ball is caught by the fielder directly from the bat, run-out is when the fielder hits the wickets before the batsmen reaching the crease while scoring the runs.

Cricket is predominately played in three formats: Test Cricket which span for 5 days where each team is allowed to bat twice and in the end of 4 innings, the team scored maximum runs will be the winner of the match. There is no over limit for this format of the game, the team can bat until all the 10 wickets are dismissed by the bowling team. Maximum of 90 overs to be bowled in a day. The other two formats are limited over cricket ODI and T20. The ODI is played for 50 overs each whereas the T20 is played only for 20 overs each. The most

challenging format turns out to be T20. The team needs to adapt to the situation swiftly to get control of the match. This focus of this thesis is Indian Premier League (IPL). It is most attracted and attended cricket league in the world. IPL is played amongst 8 teams. In this thesis we are going predict the best playing 11 before the toss of the match by evaluating the performance of the player using the available IPL dataset (2008 – 2020). The performance of the player is measured based upon their role in the team such as batsmen, bowler and All – Rounder. The metrics for players vary depending on their roles. The metrics for batsmen are Batting Average, Strike Rate, Innings, Runs scored, no of Thirties, No of fifties etc. The metrics for bowlers are Innings, Overs, Wickets, Maiden, Bowling Average, 4 wicket haul etc. Considering the above features we formulate different models for batsmen and bowler individually. The model output will vary from 0 – 1. The higher the number better the player performs in his respective roles. Basis the model output we will identify the best combination 11 consists of 5 - 6 batsmen which includes a wicket keeper, 1 – 2 all-rounder and 4 – 5 bowlers with maximum of 4 foreign players considering the venue, opposition and balance of the team, form of the player etc. The form of the player is considered as how well the player performs in the recent times we give more weightage to the player who performs relatively better in recent times.

2. Problem Statement

Selecting the right players for each match plays a significant role in a team's victory. In the research, we are predicting the best playing 11 for the match considering the batting, bowling, and fielding performances of players and other external factors such as venue, opposition, etc. The playing 11 consists of a combination of 5 – 6 batsmen includes a wicket - keeper, 1 - 2 all-rounder, and 4 – 5 bowlers with maximum of 4 foreign players.

3. Related Work

A wide online search has produced only very few articles or research papers related to predicting a player’s performance. A very few researchers have studied the game of cricket players. (Passi and Pandey, 2018) predict the best combination of playing 11 suited for that match basis the player’s performance against the opposition, venues etc. They have achieved high accuracy using Random Forest. (Wickramasinghe, 2014) predicts the performance of the batsmen on the test match using the HLM Model. This model takes into consideration factors such as height, series number, current team rank etc. (Bhattacharjee, Lemmer, Saika, and

Mukherjee, 2018) predicts the batting partnership of the batsmen using pressure index. The study identifies the pressure generated by the partners who are playing considering the number of runs required, number of overs left and number of wickets left etc. (Bhattacharjee and Saikia, 2013) predicts the best dream 11 team after the end of the tournament using batting , bowling and wicket keeper performance measures. (Faez, Jindal and Deb, 2011) predicts the best playing 11 using optimized decision making based of various parameters for batsmen, bowler and fielder. (Lemmer, Bhattacharjee and Saikia, 2014) predicts method that identifies to adjust the success prediction methods in cricket. (Amin and Sharma, 2014) predicts the team by using a linear programming called envelopment data analysis. (Saikia, Bhattacharjee and Lemmer, 2012) predicts the Fielding performance of the player by using a double edge method. (Lemmer, 2012) predicts how the performance of the bowlers vary based on the condition of the pitch. (Muthuswamy and Lam, 2008) predicts the performance of Indian bowlers against seven international teams against which the Indian cricket team plays most frequently. They have used back propagation network and radial basis network function to predict how many runs a bowler is likely to concede and how many wickets a bowler is likely to take in a given ODI match (Shah, 2017) predicts a new measure to identify the player performance by using a new measure called performance index (Barr and Kantor, 2004) predicts batsmen performance by method of examining a batsman's performance in the one-day cricket game two-dimensionally as an alternative to the largely one dimensional concern with runs per innings adopted conventionally. (Iyer and Sharda, 2009) predicts the team selection for the world cup using the non - linear modelling such as neural networks. These different models forecast the performance of the payer's batting and bowling individually. They used two different ways experiments to evaluate the overall and the current form of the players by considering last few years of statistics. (Kansal *et al.*, 2014) predicts the base price of the player in IPL auction. The players are divided into batsmen, bowlers and all-rounders respectively and the performance is evaluated on both ODI and T20 Metrics using the Multi-Layer Perceptron model, Naïve Bayes etc. (Saikia and Bhattacharjee, 2011) predicts the performance of the all-rounders using their strike rate and economy rate. These Metrics are modeled using various models such as Naïve Bayes, Multinomial Regression, etc. to classify the all-rounder's into four category performer, batting all-rounder, bowling all-rounder and under performer. (Deep, Patvardhan and Singh, 2016) predicts the ranking of the IPL players using a deep performance Index using Machine Learning. The Index calculation is different for the Most valuable batsmen and most Valuable bowler. Basis the index and other different metrics they are able to rank the players within multiple categories in batsmen such openers, middle order batsmen, finisher inexperienced etc.

In bowling the categories are fast, medium pace, spin etc. (Omkar and Verma, 2003) predicts the team using genetic algorithm. It calculates the fitness of player. Basis the fitness of the player they are able to identify the best team for the tournament from a pool of players. The fitness parameter varies for each category of the players such batsmen's fitness is considered as the number of runs he has scored, the bowlers fitness is identified as the number of wickets he has taken and so on.

4. Research Questions

The following research questions are suggested for each of the research objective as highlighted as follows.

- Prediction of the developed model
- Impact of Form for the players in the predicted model
- Assessing the scope of improvement to players in a match

5. Aim and Objectives

The main objective of this research is to Identify the best playing 11 for a given match before the toss of the match.

The research objectives of this study which are as follows:

- Data Preprocessing
- Obtaining the player metrics for batsmen, bowlers and all-rounders respectively
- Formulate the ML models for each categories of players respectively
- Basis the ML output predict the best playing 11

6. Expected output

The output is expected in different levels of this thesis are:

- Calculated player performance metrics from raw data.
- Calculated metrics are feed into the ML model and values between 0 – 1 is obtained.
- The 0-1 value is used to predict the best playing 11 team consists of 5-6 batsmen,1-2 all-rounders and 4-5 bowlers with maximum of 4 foreign players.
- Comparison between the model predicted team vs team played in the match.

7. Significance of the project

The significant contributions of this thesis are:

- Present a model that can evaluate the performance of batsmen.
- Present a model that can evaluate the performance of bowlers.
- Formulated new metrics based that will improve the model.
- Compare the accuracies of different models on calculated features from dataset. These comparisons can be used to identify the best playing 11 without any bias

8. Scope of the project

The Scope of the project is to preprocess dataset. calculate the derived performance metrics from the ball – by – ball data and consider other features like venue, opposition, partnership, form etc. These features are used in ML algorithms to calculate the value between 0 – 1. Using the ML output we predict the best playing 11 consists of 5-6 batsmen,1-2 all-rounders and 4-5 bowlers with maximum of 4 foreign players before the toss. We evaluate the performance of the player basis the predicted 11 vs the current match. When the player gets injured or player does not play the match. These scenarios are not considered in the scope of the project.

The extended scope of this thesis is to feed in ball - by - ball data to identify the identify the best playing 11 using multiple parameters into consideration like form, Non – striker etc.

9. Research Methodology

Data of all IPL matches played from 2008 - 2020 is considered in the research. All the player names are synced in all the 'Deliveries' datasets and 'Summary' dataset. Detailed information of 'Deliveries' dataset is available in Table -1 and 'Summary' dataset in Table 2.

Table 9.1 – Ball by Ball events of the match information for IPL (2008 – 2020)

Column Name	Data Description
match_id	Unique Match id
inning	Indicates whether its 1st or 2nd innings
batting_team	Batting team in that innings
bowling_team	Bowling team in that innings
over	Which over of the innings
ball	Which ball of the innings
batsman	Batsmen is playing that ball
non_striker	Batsmen is in the Non- strilker end playing that ball
bowler	Bowler bowling the over
is_super_over	Is that Super over ?
wide_runs	The number of runs scored from wide
bye_runs	The number of runs scored from bye
legbye_runs	The number of runs scored from legbye
noball_runs	The number of runs scored from noball
penalty_runs	The number of runs scored from penalty
batsman_runs	The number of runs scored from batsman
extra_runs	The number of runs scored from extra
total_runs	The number of runs scored from total
player_dismissed	Player Dismissed
dismissal_kind	Dismissal kind
fielder	Fielder involved in the wicket

Table 9.2 Overall Summary of the Match

Summary Cols	Data Description
id	Unique Match id
season	The season of the IPL
city	Venue of the game
date	Matchday
team1	Home Team of the match
team2	Opposition team of the match
toss_winner	Team which won the toss
toss_decision	Decision taken by the Team which won the toss
result	Match Result
dl_applied	Whether or not the DL method is applied
winner	Team who won the match
win_by_runs	The total runs by which the team won
win_by_wickets	The total wickets by which the team won
player_of_match	Best player of the game
venue	Venue where we have played the match
umpire1	1st umpire
umpire2	2nd Umpire
umpire3	3rd Umpire

Table – 2 contains details of 816 IPL matches and Table 1 contains 1.9L ball by ball information for the 816 IPL matches Both tables can be joined by using the 'match_id' from table – 1 and 'id' from table-2. Data of Table 1 will used to obtain all the batting, bowling metrics.

Player Statistics

The performance of the player is measured based upon their role in the team such as batsmen, bowlers and All – Rounders. The measures for players vary depending on their roles. Using the Data obtained we will calculate these measures.

Batting Metrics

Innings: Innings refers to the period in which an individual player has played in his entire career. Higher the innings the better the experience of the player.

Runs Scored – Total Number of runs the batsmen has scored in his career. The more the batsmen score better the player

Batting Average: The mean score the batsmen scored in an innings. It can be mathematically expressed as follows

$$Average = \frac{Runs\ scored}{Numbers\ of\ Innings\ Played} \qquad [9.1]$$

Strike Rate (SR): The mean runs scored by a batsman for every 100 balls played. Higher the SR quicker the runs get scored. In the limited overs, especially in T20 this becomes an important metric fir selection of a player.

$$Strike\ Rate = \frac{Runs\ scored}{Number\ of\ balls\ faced} \times 100 \qquad [9.2]$$

Thirties Score: Sum of occasions in where the batsman has a total score of "30+" in an innings. Consistency of the player can be evaluated using this metric. Higher the thirties score more consistent the player in the T20 format

Fifties Score: Sum of occasions in where the batsman has a total score of "50+" in an innings. Consistency of the player can be evaluated using this metric. Higher the fifties score more consistent the player in the T20 format. This metric has more precedence than the thirties metric

Zeros Sum of occasions in where the player was dismissed even before securing a run.

Highest: Maximum runs that a batsman has scored in an innings in his entire career. The highest score in each venue and opposition can be used as an additional metric while considering the opposition and venue. The probability of a player scoring more in his home ground is likely high compared to other venues or the player will score maximum runs in a specific pitch conditions as well.

Additional Metrics

Average of Dots – The average number of balls where a player has not scored in run. The lower the average better the player. In T20 format, this cannot be high as the number of overs is minimum.

Bowling Metrics

Innings: The sum of occasions where the player has bowled a minimum of 1 ball in entire career. Higher the innings the better the experience of the player.

Overs: The sum of total overs the bowler has bowled in his entire career. Higher the overs the better the experience of the player.

Wickets – Total number of occasions the bowler has dismissed the batsmen in his entire career. Higher the wickets better the player.

Maiden – The number of over where the bowler has not even conceded a single run. The higher the metric the better the player.

Bowling Average: Mean runs conceded by the bowler to pick a wicket. The Lower the Bowling average better the bower who will concede less for picking up a wicket. It can be mathematically expressed as below

$$Bowling\ Average = \frac{Number\ of\ Runs\ conceded}{Number\ of\ wickets\ taken} \quad [9.3]$$

Bowling Strike Rate: It defined as the mean number of balls that a bowler has taken to pick a wicket. This plays a pivotal role in analyzing number of wickets the bowler can pick on an average in a match. Lower the metric attributes higher the bowler will wickets frequently. It can be mathematically expressed as below

$$Strike\ Rate = \frac{Number\ of\ balls\ bowled}{Number\ of\ wickets\ taken} \quad [9.4]$$

4 Wicket Haul: Total sum of occasion where the sum of wickets taken by the bowler in an innings is equal to or greater than 4. Higher the 4 wicket haul more consistent the player in the T20 format

Economy Rate – Average number of runs the bowler concedes in a single over, lower the economy rate better the bowler performance.

Fielding

No of catches: The total number of catches that the player has caught. (The total number of balls caught by the player that dismisses the batsman.) Batting and bowling capabilities will not be solely important, fielding plays a vital role in a team's victory as well.

In addition to these other player parameters like Current form, Consistency, Partnership, opposition, Venue, Non - striker will be taken into account for selecting the best team. These calculated measures will be input features for the ML Algorithm Logarithmic Regression, Linear Regression, Lasso and Ridge etc. Basis the output from the ML we identify the best 11 consists of 5-6 batsmen,1-2 all-rounders and 4-5 bowlers with maximum of 4 foreign players.

10. Resource Required

Hardware Resources

- GPU for Deep Learning
- Ram – 16 GB
- Processor – i7-9750H CPU

Software Resources

- Python
- Keras
- Tensorflow
- PyTorch
- Pandas
- Numpy
- SciPy

11. Resource Plan

Table 3 – Gantt Chart for Report

Tasks/Period	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Data Preprocessing																	
Calculation of Batting Metrics																	
Predicting Models for Batting performance																	
Identify 6-7 best batsmen																	
Interim Report																	
Calculation of bowling Metrics																	
Predicting Models for Bowling performance																	
Identify 3 - 4 best bowlers																	
All - Rounder Prediction																	
Output Evaluation																	
Final Report																	

Period - Week Ending dates, Period starts at 1 and ends 17 the 1 and 17 week are represented below:

1 – Week Ending 25th Jan 2021

17 – Week Ending 17th May 2021

Reference

[1] Passi, K. and Pandey, N. (2018) 'Predicting Players' Performance in One Day International Cricket Matches Using Machine Learning', (December), pp. 111–126. doi: 10.5121/csit.2018.80310.

[2] Wickramasinghe, I. P. (2014) 'Predicting the performance of batsmen in test cricket', *Journal of Human Sport and Exercise*, 9(4), pp. 744–751. doi: 10.14198/jhse.2014.94.01.

[3] Bhattacharjee, D., Lemmer, H. H, Saikia, H and Mukherjee, Diganta. (2018). 'Measuring performance of batting partners in limited overs cricket', Journal for Research in Sport, Physical Education and Recreation. 40. 1-12.

[4] Bhattacharjee, D. and Saikia, H. (2013) 'Selecting the Optimum Cricket Team after a Tournament.', *Asian Journal of Exercise & Sports Science*, 10(2), pp. 77–91.

[5] A. Faez, A. Jindal and K. Deb, "Cricket team selection using evolutionary multi-objective optimization," in International Conference on Swarm, Evolutionary, and Memetic Computing, Berlin, 2011.

[6] Lemmer, H. H., Bhattacharjee, D. and Saikia, H. (2014) 'A consistency adjusted measure for the success of prediction methods in cricket', *International Journal of Sports Science and Coaching*, 9(3), pp. 497–512. doi: 10.1260/1747-9541.9.3.497.

[7] Amin, G. R. and Sharma, S. K. (2014) 'Cricket team selection using data envelopment analysis', *European Journal of Sport Science*, 14(SUPPL.1). doi: 10.1080/17461391.2012.705333.

[8] Saikia, H., Bhattacharjee, D. and Lemmer, H. (2012) 'A double weighted tool to measure the fielding performance in cricket', *International Journal of Sports Science and Coaching*, 7(4), pp. 699–713. doi: 10.1260/1747-9541.7.4.699.

[9] Lemmer, H. H. (2012) 'Individual Match Approach To Bowling Performance Measures in Cricket', 34(2), pp. 95–103.

[10] Muthuswamy S. and Lam, S. S. (2008) 'Bowler Performance Prediction for One-day International Cricket Using Neural Networks', Industrial Engineering Research Conference.

[11] Shah, D. P. (2017) ‘New performance measure in Cricket’, *IOSR Journal of Sports and Physical Education*, 04(03), pp. 28–30. doi: 10.9790/6737-04032830.

[12] Barr, G. D. I. and Kantor, B. S. (2004) ‘A Criterion for Comparing and Selecting Batsmen in Limited Overs Cricket’, Operational Research Society, vol. 55, no. 12, pp. 1266-1274.

[13] Iyer, S.R. and Sharda, R., 2009. Prediction of athletes performance using neural networks: An application in cricket team selection. Expert Systems with Applications, 36(3), pp.5510-5522.

[14] Kansal, P. et al. (2014) ‘Player valuation in Indian premier league auction using data mining technique’, Proceedings of 2014 International Conference on Contemporary Computing and Informatics, IC3I 2014, pp. 197–203. doi: 10.1109/IC3I.2014.7019707.

[15] Saikia, H. and Bhattacharjee, D. (2011) ‘On classification of all-rounders of the Indian premier league (IPL): A Bayesian approach’, Vikalpa, 36(4), pp. 51–66. doi: 10.1177/0256090920110404.

[16] Deep, C., Patvardhan, C. and Singh, S. (2016) ‘A new Machine Learning based Deep Performance Index for Ranking IPL T20 Cricketers’, International Journal of Computer Applications, 137(10), pp. 42–49. doi: 10.5120/ijca2016908903.

[17] Omkar, S. N. and Verma, R. (2003) ‘Cricket team selection using genetic algorithm’, International Congress on Sports Dynamics (ICSD2003), pp. 1–3.