Linköping University | Department of Computer and Information Science
Master's thesis, 30 ECTS | Statistics and Machine Learning
2023 | LIU-IDA/STAT-A--23/027--SE

Applications of Mixed Models to Predict Player Auction Prices in the Indian Premier League

- A study of Player Econometrics in Cricket

Jaskirat Singh Marar

Supervisor : Bayu Brahmantio
Examiner : Bertil Wegmann

LINKÖPING UNIVERSITY

Linköpings universitet
SE–581 83 Linköping
+46 13 28 10 00 , www.liu.se

Upphovsrätt

Copyright

Abstract

The Indian Premier League (IPL) is a professional T20 cricket league that attracts some of the best players from around the world. The auction of IPL players is a critical event that determines the composition of the teams for the upcoming season. In this thesis, I used mixed models to predict the auction prices of IPL players based on their performance data in the previous seasons of the IPL and their recent performance in other leagues. I collected data on the performance of players and their auction prices for the past seasons of the IPL and also other renowned T20 leagues and used this data to train and test the prediction models.

In this thesis, I use classical linear regression models as baseline models and later train linear mixed models (LMMs) and use Markov Chain Monte Carlo (MCMC) simulations of Generalized Linear Mixed Models (GLMM) to predict the auction prices of Indian Premier League (IPL) players. I modeled the players separately based on their roles, i.e., batsmen, bowlers, and all-rounders. I collected data on the performance of IPL players in previous seasons, including their batting and bowling statistics, and used this data as predictors to train and test the models.

I found that the LMMs and GLMM MCMC simulations generally outperformed the simple linear regression models in predicting the auction prices of IPL players. The mixed models also captured some relationships between the auction price and the predictors that could not be seen with the baseline models in each of the player roles. From this study shows that for batsmen, the player's ability to score quickly in recent competitions in addition to their overall batting experience in the IPL is considered important for a higher auction price. For bowlers, its the overall experience and consistency of play that is considered more important than the wickets they take. Additionally, the study shows that a player's *'Star'* status is a very important predictor for auction price regardless of the role that the player fulfills.

The findings have important implications for the IPL teams and players, as they can use this information to make informed decisions during the auction and the season. Overall, this study contributes to the growing body of research on sports analytics, cricket and prediction modeling and highlights the potential of *Mixed Models* for predicting auction prices in T20 cricket leagues like the IPL.

Keywords: Mixed Models, Longitudinal Data Modelling, Regression, MCMC, Sports, Cricket, IPL, Player Salary Prediction.

Acknowledgments

I would like to thank my supervisor, Bayu Brahmantio, for his frequent and knowledgeable support throughout the thesis study. He provided valuable insights and course corrections throughout the journey of writing this manuscript. He was very considerate of my questions and was a very delightful companion for discussing cricket. I would like to thank Bertil Wegmann, who was my examiner for this thesis. His key feedback during the review meetings and seminars was very critical for me to focus my study in the right direction. I'd like to thank Linus Kåge, who was the opponent in the defense of this thesis study. His insightful discussion and feedback was very critical in shaping this manuscript into its final form.

I would like to thank Oleg Sysoev, the course leader for the master thesis at LiU. His critical feedback during the reviews and his design of the master thesis program has been very helpful & appreciated as a student participant in the course. I would also like to thank Prof. Patrick Lambrix for his guidance and support in formulating the research strategy that was necessary to get this thesis started. His ideas and timely guidance also helped me find a lot of relevant background literature for this thesis and he was instrumental in generating ideas that shaped this thesis. I would like extend my gratitude to Isak Heitala who first introduced me to mixed models by accepting me as a student for a research project that utilized longitudinal models on game data from Football Manager 2022. The work I did while working on this project was instrumental in helping me design my thesis research questions and modelling proposals.

I would also like to thank my wife, Ruchita, for supporting me and encouraging me while I spent days and nights on this very long and extensive study that turned out to be one of the biggest commitments for me and my family.

Contents

List of Figures

List of Tables

Acronyms

ICC International Cricket Council

BCCI Board of Cricket Control in India

ODI One Day International

T20I Twenty20 International

AI Artificial Intelligence

ML Machine Learning

T20 Twenty20

NBA National Basketball Association

NFL National Football League

LM Linear Model

LMM Linear Mixed Model

GLMM Generalized Linear Mixed Model

MCMC Markov Chain Monte Carlo

MVN Multi Variate Normal

RMSE Root Mean Square Error

MAPE Mean Average Percentage Error

EDA Exploratory Data Analysis

SR Strike Rate

OLS Ordinary Least Squares

RBW Running Between Wickets

1 Introduction

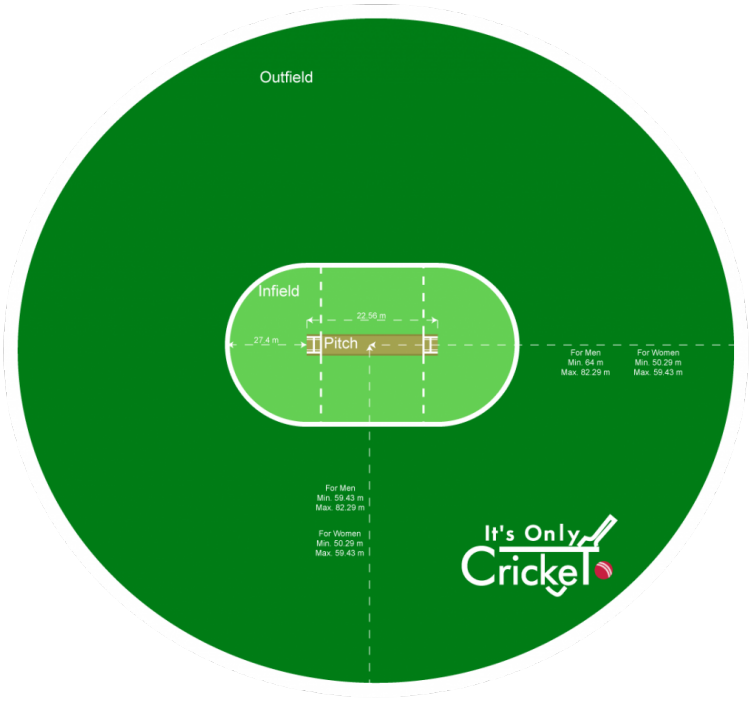
1.1 Background

Cricket is a professional sport that is played between two teams of 11 players each. The aim of the game is for one team to score more runs than the other team. The game is played on a large oval-shaped field with a rectangular strip of land in the middle called the *pitch* as depicted in figure 1.1. The pitch is 22 yards long and has two sets of wickets at either end. Each set of wickets consists of three wooden stumps and two bails. One team bats while the other team bowls and fields. The batsman's primary objective is to hit the ball with the bat and score runs by running back and forth between the wickets. The bowler's objective is to throw or *bowl* the ball towards the batsman in such a way that he cannot hit it, or to get him *out* or *dismissed* by hitting the wickets behind the batsman with the ball. There are different formats of cricket, such as Test cricket, One Day Internationals (ODIs), and Twenty20 (T20) cricket, each with their own set of rules and regulations. However, the basic principles of the game remain the same across all formats.

1.1.1 Evolution of Cricket

Cricket has existed as sport for centuries with its invention being traced back to Saxon or Norman times in the south-eastern part of England. First references to it being played by adults is made in 1611 with a dictionary from those times describing it as a boys' game [19]. The game started getting organized in England during the 17th century with the formation of county teams and generating professional employment for expert players. The spread of the game across the globe can be attributed to the British Empire. The first matches of significance were played between England and Australia which were highlighted by a tense match in 1882 at the The Oval in London where Australia came out the winners by beating England for the first time in England. What gave this match importance is that it sparked a satirical obituary in a British newspaper which stated that English cricket was dead and that its ashes were being taken to Australia. This gave rise to the first rivalry in Cricket and made way for a biennially contested series between England and Australia called the *'The Ashes'* that continues to be played to this day.

For the longest time, cricket was played in the *'test match'* format in which the match is played over 5 days until the introduction of limited over formats. In 1971, the first limited-



(a) Cricket Field and pitch dimensions [17]

(b) A Batsman playing the ball [16]

(c) A Bowler delivering a ball [16]

Figure 1.1: Basic Cricket rules & roles

overs international match or ODI was played in Australia at the Melbourne Cricket Ground. This was then followed up by the first ever World Cup being organized in 1975 that was played under the limited overs format rules. The event was highly successful and was made a regular part of the cricketing calendar and has since been repeated every 4 years. Over time gradual changes like coloured kits, white balls, evening games etc, have been introduced to modernize the game and make it more appealing for drawing larger crowds.

In the 21st century, this format of the game has since evolved into T20 cricket, which is now a very popular format of the game. The spread of this, the shortest format of the game, stems from the decline in interest in the longer duration format of test matches, which can last upto 5 days, and ODIs which typically last for 6-8 hours per game. A typical T20 match, however, gets over in about 3-4 hours. Crowds are more attracted to this format as it is more exciting and features a more attacking playstyle and more high points like fours and sixes. The rules of competitive cricket are very detailed and it is impractical to list them all here, but the following rules highlight some of the differences in the 3 most popular formats i.e. Tests, ODIs and T20s.

1. Test cricket:
    a) Each team has two innings to bat and bowl.
    b) There is no limit on the number of overs in an innings.
    c) The match can last up to five days.
2. ODI cricket:
    a) Each team has one innings to bat and bowl.
    b) Each innings is limited to 50 overs, an individual bowler can bowl upto 10 overs of 6 deliveries each.
    c) The match can last up to eight hours.
3. T20 cricket:
    a) Each team has one innings to bat and bowl.
    b) Each innings is limited to 20 overs, an individual bowler can bowl upto 4 overs of 6 deliveries each.
    c) The match can last up to four hours.

In addition to these basic comparative rules, each format has its own specific rules and regulations. For example, in T20 and ODI cricket, there are *powerplays* and fielding restrictions that are not present in Test cricket. Overall, each format of cricket offers a different style of play and different challenges for players. Test cricket is known for its strategic and patient approach, ODI cricket is known for its fast-paced action and high-scoring matches, and T20 cricket is known for its explosive hitting and innovative tactics.

1.1.2 Rise of Franchise Leagues

The T20 format has also made way for the introduction of franchise cricket leagues such as the Indian Premier League and the Big Bash League. These franchise leagues have changed the landscape of cricket with players from around the world now competing in multiple franchise league competitions in addition to their domestic competitions and playing for their country. Almost every major cricket playing nation now has its own T20 league competition. The largest and most valuable franchise league in cricket right now is the Indian Premier League (IPL) [27].

The IPL is a professional Twenty20 cricket league that is played in India. It was founded in 2008 and is organized by the Board of Control for Cricket in India (BCCI). It was built upon the model of successful franchise based leagues from other sports such as the English Premier League, Major League Baseball and the National Football League. The IPL is made up of ten teams that represent different cities or states in India. Each team is owned by a franchise, and players are bought and sold through an English auction system before the start of the season. The league is known for featuring some of the best players from around the world, as well as some of India's top domestic talent, made possible due to the fact that the salaries that the players receive are the highest as compared to any other cricket league in the world. The IPL season typically takes place in the months of April and May each year, and consists of a round-robin group stage followed by playoffs that feature knock-out matches. During the round-robin stage, each team plays the other competing teams, either twice - once at home and once away, or plays some opponents only once. The top four teams at the end of the round-robin stage qualify for the playoffs, which include 3 knock-out qualifier matches and a final.

The IPL auction precedes the playing season every year where franchise teams bid on players to build their team for the upcoming season. The auction is usually held a few months before the start of the tournament. The IPL auction is conducted in *English auction rules* [33] and works in the following manner:

1. Player registration: Before the auction, players from around the world register themselves for the IPL auction. The Board of Control for Cricket in India (BCCI) also sets a base price for each player, which serves as the minimum bid for that player.

2. Franchise team budgets: Each franchise team is given a budget, which they can use to bid on players. The budget varies from year to year, and depends on the total amount of money available for the auction.

3. Bidding process: The auction takes place in rounds, with each round featuring different players. Teams can bid on a player by placing a bid that is higher than the current highest bid. The bidding continues until no other team is willing to bid higher, at which point the player is sold to the highest bidder.

4. Unsold players: If a player does not receive any bids during the auction, he is considered unsold. Unsold players can still be signed by a team as a replacement player later in the tournament.

5. Player retention: Before the auction, each franchise team is allowed to retain a certain number of players from their previous season's squad. The number of players varies from year to year, and is decided by the BCCI.

1.2 Motivation

While the IPL is a lucrative league, there are significant disparities in revenue between teams. Some of the most popular teams, such as Mumbai Indians and Chennai Super Kings, earn significantly more revenue than less popular teams, such as Rajasthan Royals and Kings XI Punjab. This disparity in revenue can make it difficult for less popular teams to compete with the top teams in terms of player salaries and other expenses. While the IPL has a salary cap to prevent teams from overspending, this can also limit the ability of teams to attract top talent. Some teams with lower revenue streams may not be able to afford to pay top salaries, which can make it difficult for them to compete with the top teams. Hence, accurately predicting a player's auction value is a significant problem for a multitude of reasons, some of which are listed as follows:

1. Budget management: Each team in the IPL has a budget cap that they cannot exceed during the auction. Accurately predicting a player's auction value can help a team to manage their budget effectively and ensure that they do not overspend on any one player.
2. Team balance: A team's success in the IPL often depends on the balance of their squad. Accurately predicting a player's value can help a team to select players that complement each other's skills and strengths, and ensure that the team has a well-rounded squad.
3. Competitive advantage: The IPL is a highly competitive league, and every team wants to gain an edge over their rivals. Accurately predicting a player's value can help a team to identify undervalued players who may be able to provide a significant impact on the team's performance, while also avoiding overpaying for players who may not perform up to expectations.
4. Financial gain: Accurately predicting a player's value can also have financial implications, as teams can sell players at a profit if they perform well and increase in value over time. Conversely, if a team overpays for a player who does not perform up to expectations, they may lose money on the investment.

The objective of this thesis is to predict a player's auction value based on their current and past performances. Accurately predicting a player's auction value is crucial for teams in the IPL to ensure that they have a well-balanced squad, manage their budget effectively, gain a competitive advantage over their rivals, and potentially make financial gains in the long term.

1.3 Research questions

For this thesis two key questions to be answered are as follows:

1. *What are the best predictors for modelling a player's auction value?*

   The data available presents numerous performance statistics for each player but some of these statistics will be more important in terms of relevance to the format of the game, the balance of the team that is bidding for the player, the role that the player offers to a given team. These are some of the aspects that will determine which features are more important than others. Some metrics will also be derived and also prone to high correlation with other predictors.

2. *What is the best modelling choice for predicting the auction value against performance?*

   There are numerous predictive modelling techniques that can be used in the statistics and machine learning, but in this situation some will be more relevant than others. For instance, the player performance data is expected to have correlations since, the same players might be observed at different times or years. Hence a simple linear or non-linear regression may not be the most optimal model. We will also need to evaluate the performance of our models with baseline models that have been trained on similar data in other research studies with similar intended outcomes.

1.4 Delimitations

Cricket is a very complex team sport in which an individual player is just 1 out of a team of 11 players and as such quantifying their contribution to a team is not a trivial task. A player's valuation is sensitive to various factors and has no fixed template that can calculate a valuation. A player's form and record are just one side of their professional profile. There

are a multitude of other intrinsic and extrinsic factors that define the true value of a player like popularity, form, leadership etc.

In the context of this thesis, we are choosing to limit the scope of the research to including tangible performance indices that have been recorded with absolute certainty i.e. actual game data. In this study we will attempt to use these absolute measures as predictors and choose a prediction model which gives us the highest prediction accuracy using these tangible predictors. We will choose to not include intuitively relevant features such as player popularity, expert predictions, current team balance etc., which no doubt would be influencing a player's perceived value but at the same time, these features are highly subjective and would require extensive testing and tuning for them to be used to make reliable prediction models.

2 Theory

2.1 Literature Review

Compared to some popular sports like football or American sports like basketball, American football and baseball, cricket is an understudied sport from an analytics and machine learning perspective (ch 23. [2]). Most of this is a consequence of cricket being a sport that is primarily played in a handful of countries at a professionally competitive level. This leads to lower salaries and a lower interest level in observing the game from an analytical lens. The countries or region where the sport is most popular is the UK, Australia, New Zealand, Indian sub-continent countries, the Caribbean nations and South Africa. Every other country that plays cricket is considered a minor playing nation where cricket is not the top sport and also struggles to keep the sport economically viable for the local players. Most of the cricket playing nations other than India, Australia and England are not cash rich that is to say that their cricket governing bodies rely on these three countries and the ICC for funding and generating revenue by conducting bilateral matches. As a consequence, there is a relatively smaller interest in pursuing analytical studies on cricket. This has also resulted in a much smaller community that does active research on cricket statistics and analytics. There is however, comprehensive research on other American sports like the NBA, MLB and NFL.

2.1.1 Related Works

Stanek uses multiple statistical models, including ordinary least squares regression, fixed effects regression, random effects regression, and instrumental variables regression, to analyze the relationship between player performance, player salaries, and team revenues in the NBA [32]. The author discusses the statistical models used in detail, providing a thorough explanation of each model and its assumptions. The models used in the analysis include linear regression, fixed-effects regression, random-effects regression, and instrumental variable regression. The results of the thesis showed that player performance, as measured by various statistical metrics, had a significant positive impact on team revenues in the NBA. Specifically, the author found that player scoring, shooting efficiency, and minutes played had the strongest positive effects on team revenues. Additionally, the author found that player salaries had a significant positive effect on team revenues, but that this effect was partially mediated by player performance. The author also compared the performance of several dif-

ferent statistical models and found that a mixed-effects model provided the best fit for the data. Overall, the thesis provides evidence that investing in high-performing players can be a profitable strategy for NBA teams.

Most of the academic work that exists in cricket is focused on creating models that are able to predict match results or explain player performance metrics better. Irvine and Kennedy [20] aimed to identify the performance indicators that have the most significant impact on the outcome of International T20 cricket matches. The authors used principal component analysis (PCA) to identify the most important performance indicators. PCA is a multivariate statistical technique that reduces the dimensionality of a dataset by identifying the underlying patterns and correlations among the variables. The study found that a team's batting average, strike rate, and boundary percentage were the most important performance indicators for winning International Twenty20 cricket matches. Batting average refers to the average number of runs scored per batsman, while strike rate measures the number of runs scored per 100 balls faced. Boundary percentage refers to the proportion of runs scored through boundaries (four or six). In terms of bowling performance, the study found that a team's bowling average, economy rate, and the percentage of wickets taken by spin bowlers were the most important performance indicators for winning matches. Bowling average measures the average number of runs conceded per wicket taken, while economy rate measures the number of runs conceded per over bowled. Overall, the authors suggested that teams could use these performance indicators to identify their strengths and weaknesses, and adjust their strategies and tactics accordingly. For example, a team with a strong batting lineup may focus on setting a high target or chasing down a large total, while a team with strong bowling may focus on restricting the opposition's score and taking wickets at crucial moments. These inputs can also be used to target the right players in the IPL auctions that would fit with the team's balance for the upcoming season. The study also found that fielding statistics, such as the number of catches and run outs, had a lesser impact on the outcome of matches compared to batting and bowling performance.

Specific to the IPL and the auctions, a lot of research focuses on player salary and how identifying players that justify their salary or final auction bids. An early study on IPL player salaries was conducted where the author aimed to determine the value of cricketers in the IPL using a hedonic price model [23]. A hedonic pricing model is a statistical model used to estimate the economic value of a product or service based on its characteristics, such as its physical attributes, quality, and other relevant factors. In the context of real estate, for example, a hedonic pricing model might be used to estimate the value of a house based on its size, location, number of rooms, age, and other factors. In the context of cricket, a hedonic pricing model might be used to estimate the value of a player based on their performance statistics, such as batting average, strike rate, or bowling average, and other relevant factors, such as age and experience. The author found that the performance statistics of cricketers in the IPL had a significant impact on their auction prices, and therefore their value. The author also found that certain performance statistics, such as batting average and strike rate, had a greater impact on auction prices and value than others. This study also inspired a modified hedonic pricing model to predict the auction price of players in the IPL auction [6]. The authors proposed a modified hedonic pricing model that includes new variables, such as player nationality, player type (i.e., batsman or bowler), and team-specific variables, such as team performance and team owner's net worth. The authors argue that these variables could affect the auction price of players in the IPL and should therefore be included in the model. The authors found that the modified hedonic pricing model was able to more accurately predict the auction prices of players in the IPL than the traditional hedonic pricing model.

Sankaran aimed to compare the pay and performance of IPL bowlers using cluster analysis [30]. The author collected data on IPL bowlers from the 2011, 2012, and 2013 seasons and analyzed their performance using three performance metrics: economy rate, strike rate, and wickets taken per match. The author used cluster analysis, a statistical technique used to group similar data points together, to classify the bowlers into different clusters based on

their pay and performance. The analysis identified four clusters of bowlers based on their pay and performance: High Performers, Overpaid Performers, Underpaid Performers, and Low Performers. The results showed that the High Performers cluster had the highest average performance score and the second-highest average pay, while the Underpaid Performers cluster had the second-highest average performance score and the lowest average pay. In contrast, the Overpaid Performers cluster had the lowest average performance score and the highest average pay, while the Low Performers cluster had the lowest average performance score and the second-lowest average pay. The author concluded that there was a significant disparity between pay and performance among IPL bowlers, with some being overpaid and others being underpaid. The author suggested that teams could use the results of this analysis to optimize their team selection and spending, by investing more in high-performing, underpaid bowlers and reducing spending on low-performing, overpaid bowlers.

Another detailed study was conducted by Malhotra to develop a model to predict the auction prices of cricketers in the IPL and their economic value to their teams [26]. The author collected data on IPL auctions from 2008 to 2016 and analyzed the performance statistics of 420 cricketers who participated in at least one season of the IPL. The author used a combination of statistical techniques, including PCA and regression analysis, to identify the most important factors affecting auction prices and economic value creation. The author found that a cricketer's performance statistics, such as batting average, strike rate, bowling average, and economy rate, were the most important factors affecting their auction price and economic value. The author also found that a cricketer's nationality, age, and playing position also had a significant impact on their auction price and economic value. The author used Bayesian ridge regression as a type of multiple regression model to predict the auction prices. Bayesian ridge regression is a type of linear regression that is used to estimate the coefficients of the independent variables while reducing the risk of overfitting. The author found that Bayesian ridge regression was able to improve the accuracy of the model compared to traditional linear regression models, the model was able to accurately predict the auction prices of cricketers with an average error rate of less than 10%, and the economic value creation of cricketers with an average error rate of less than 5%.

A more recent study was carried out by Davis et al. who propose, expected run differential, as a metric for evaluating the impact of batting performance on winning the game [7]. They define expected run differential as the difference between the expected runs a team will score based on the average performance of the team's players, and the expected runs the team will concede based on the average performance of the opposition's players. The authors argue that expected run differential is a useful metric for evaluating batting performance in T20 cricket because it captures the impact of each player's performance on the team's overall performance, while taking into account the quality of the opposition. They use expected run differential to estimate the weights for each performance statistic in their overall performance score and find that batting performance, particularly strike rate, has the most significant impact on a player's overall performance. This proposed metric is a weighted sum of performance statistics to create an overall performance score for each player, where the weights are based on the impact of each statistic on winning the game. They use a data set of 1,161 T20 cricket matches played between 2007 and 2012 to estimate the weights and evaluate player performance.

In another paper the authors propose a new method for ranking cricketers in the IPL T20 format using machine learning techniques [8]. The authors argue that traditional methods of player evaluation, such as batting average and strike rate, do not capture the full complexity of player performance in T20 cricket, which includes multiple facets such as batting, bowling, and fielding. To address this, the authors propose a new Deep Performance Index (DPI) that incorporates multiple performance indicators for each player, including batting, bowling, fielding, and match-winning performances. The DPI is calculated using a combination of traditional statistical methods and machine learning techniques, including feature engineering, feature selection, and a deep neural network model. The authors evaluate the

effectiveness of the DPI using data from IPL seasons 2013-2018 and compare it with other commonly used player evaluation metrics, such as the Most Valuable Player (MVP) index and the player rating index (PRI). They find that the DPI outperforms these other metrics in predicting the actual rankings of IPL players and that it is more effective at capturing the complexity of player performance in T20 cricket. Of particular interest in this paper is the feature engineering which the authors utilize at deriving new metrics that help derive the final DPI. The authors derive 5 different metrics that define batting and bowling ability separately.

Saikia et al. developed an Artificial Neural Network (ANN) model to predict the performance of bowlers in the IPL based on their past performance data [29]. They included several performance indicators, such as economy rate, bowling average, number of wickets, and strike rate, in their analysis. They used this data to train and test an Artificial Neural Network (ANN) model. The ANN model used backpropagation as a learning algorithm, with a sigmoid activation function. The model was trained to predict the performance of bowlers in the IPL based on their past performance data. The authors used the root mean squared error (RMSE) and the mean absolute error (MAE) as performance metrics to evaluate the accuracy of the ANN model. The results showed that the ANN model was able to predict the performance of bowlers in the IPL with a high degree of accuracy. The RMSE and MAE values were found to be 0.64 and 0.44, respectively, indicating that the model was able to make accurate predictions. The authors also conducted a feature selection analysis to identify the most important performance indicators for predicting the performance of bowlers in the IPL. The results showed that the number of wickets and the economy rate were the two most important indicators for predicting the performance of bowlers.

2.2 Longitudinal Data Analysis

Data that is collected over time from the same individual or unit of analysis qualifies to be categorized as *longitudinal*. Common examples are usually related to medical data for instance, in a study of a group of individuals' weight loss over a year, longitudinal data would involve measuring the weight of each individual at multiple time points, such as every month or every quarter. Longitudinal data can be collected through a variety of methods, including surveys, interviews, observations, and medical tests. The data used for modelling in this thesis is categorized as longitudinal because it satisfies the following 2 conditions:

1. The data involves observations for multiple individuals (cricket players in this study)
2. Individual (player) has multiple observations over time (seasons and matches in this study)

The nature of longitudinal data implies that a set of observations for an individual in a longitudinal study are usually correlated. This is why longitudinal data requires a choice of model that allows inferences about regression parameters that also respect the correlation between the data. Even time series data has correlation between observations, but what makes longitudinal data different is that we can usually assume independence between individuals in the study. With longitudinal data, it is possible to make valid inferences using patterns across individuals that results in more robust conclusions versus time series.

The idea behind analyzing longitudinal data is to study trends and changes over time, as well as assess the effect of interest variables on the outcome of interest. Longitudinal data can also be used to make predictions about future outcomes based on past trends and patterns. In the context of this thesis, the aim is to predict a player's auction prices by taking into account their current and past performance and also to model the trends for each player separately. The trend for each player is relevant here because a player have played very well for a few years and then had a couple of years of bad form in between. This might lead to a player being completely absent from the top playing leagues as they rise back into

competitive consideration. Each player is unique and different in terms of how they play the game or even how they contribute. Statistically speaking, this refers to each player having their own random effects that define their playing ability.

While there are several methods in analyzing longitudinal data, two popular approaches are [10],

1. *Mixed effects models*: These models are used to analyze data with both fixed and random effects, where fixed effects are factors that are expected to influence the outcome variable for all individuals, and random effects account for individual-level variation within the population.

2. *Generalized estimating equations (GEE)*: These models are a type of statistical method used to analyze correlated data, where observations within a subject or cluster are expected to be correlated with each other. GEEs are a generalization of linear regression models that account for the correlation between observations. They estimate the regression coefficients and the correlation structure of the data simultaneously, and can be used to model various types of outcomes, including continuous, binary, and count data. The key advantage of GEEs is that they can accommodate missing data and can handle data that do not follow a normal distribution. GEEs are also robust to unspecified correlation structures, making them more flexible and applicable to a wide range of study designs.

2.3 Regressions Models for Longitudinal Data

2.3.1 General Linear Model

In this study we denote, Y_{ij} as the random variables of desired outcome i.e. Player Salary or Auction Price for the i_{th} player at the j_{th} time point. n_i are the total observations for the i_{th} player. m is the number of players in the data. For each y_{ij}, $\mathbf{x_{ijk}}$ is the vector of p explanatory variables for the i_{th} individual at j_{th} time point. It is assumed that y_{ij} are the realizations of Y_{ij}.

For a continuous outcome Y_{ij} a linear regression model with intercept can be written as (ch 4. [10]),

$$Y_{ij} = \beta_1 x_{ij1} + ... + \beta_p x_{ijp} + \varepsilon_{ij} \tag{2.1}$$

where, ε_{ij} is the error term of length n for each of the m players. For a classical linear model, the error terms would be assumed to be i.i.d. with a mean 0 but, for longitudinal data we would expect the error terms to be correlated within the same players.

The matrix formulation of (2.1) can be stated as,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.2}$$

where,

1. Y is the complete set of $N = nm$ observations from m players
2. $\mathbf{X}$ is the $N \times p$ matrix of explanatory variables
3. $\boldsymbol{\beta}$ is the $p \times 1$ regression coefficients estimated by the model

This model assumptions can be listed as follows,

1. Linearity: E[Y] is a linear function of the covariates.
2. Independence: The observations are assumed to be independent of each other. This means that the value of one observation does not influence the value of another observation.

3. Homoscedasticity: The variance of the response variable is assumed to be constant across all levels of the predictor variables. This assumption implies that the spread or dispersion of the response is consistent across the predictor values.

4. Normality: The error term (residuals) in the GLM is assumed to follow a normal distribution. This assumption implies that the distribution of the response variable is approximately normal for each combination of predictor values.

5. The same regressions coefficients are sufficient to explain the behavior of all the players in the data.

This would be evidently flawed for longitudinal data because it ignores the heterogeneity of behavior among different players. So we make the next adjustment of allowing each player to be modelled with their own slope and intercept.

2.3.2 Linear Mixed models

Fundamentally, what sets mixed models apart is the incorporation of parameters called *fixed* and *random effects*. *Fixed effects* as the name suggests do not vary, while, *random effects* are parameters that are random variables themselves. Linear mixed models are used when the outcome variable is continuous and the residuals follow a normal distribution. The linear mixed model is defined by a linear predictor function that includes fixed and random effects. We can extend the LM from (2.2) to construct a LMM as follows [5],

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{U} + \boldsymbol{\varepsilon} \tag{2.3}$$

where,

1. $\mathbf{Z}$ is the $N \times qm$ design matrix for q random effects and m groups or players

2. $\mathbf{U}$ is the $qm \times 1$ vector of q random effects for m groups or players

In a linear mixed model, we are actually modelling the conditional distribution of $\mathbf{Y}|\mathbf{U} = \mathbf{u}$ hence, the complete formulation of the model should be written as [4],

$$(\mathbf{Y}|\mathbf{U} = \mathbf{u}) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \sigma^2\mathbf{W}^{-1}) \tag{2.4}$$

where,

1. $\mathbf{W}$ is the diagonal matrix of known prior weights

2. $\mathbf{Z}$ is the $n \times q$ model matrix for the q-dimensional random effects variable, $\mathbf{U}$ whose value is being fixed at $\mathbf{u}$

3. The distribution of $\mathbf{U}$ is $\sim MVN(0, \Sigma)$, where Σ is a $q \times q$ variance-covariance matrix

4. Since Σ is a variance-covariance matrix, it should be positive semidefinite and is subsequently expressed as, $\Sigma_\theta = \sigma^2\Lambda_\theta\Lambda_\theta^T$. Here we introduce a $q \times q$ matrix form of *relative covariance factor* denoted by Λ_θ where θ is a variance-component parameter. σ is the same scale factor as in the conditional distribution.

2.3.2.1 Random Effects Matrix - Z

The random effects matrix, $\mathbf{Z}$ comprises of individual $\mathbf{Z}_i$ blocks where, $i = 1, 2, .., k$, $k \geqslant 1$. Each term of the random effects model matrix, $\mathbf{Z}_i$ is computed as,

$$\mathbf{Z}_i = (\mathbf{J}_i^T * \mathbf{X}_i^T)^T = \begin{bmatrix} \mathbf{J}_{i1}^T \otimes \mathbf{X}_{i1}^T \\ \mathbf{J}_{i2}^T \otimes \mathbf{X}_{i2}^T \\ \vdots \\ \mathbf{J}_{in}^T \otimes \mathbf{X}_{in}^T \end{bmatrix}$$

This $*$ product is called the Khatri-Rao [24] product and the $\otimes$ product is called the Kronecker product. $\mathbf{J}_{ij}^T \& \mathbf{X}_{ij}^T$ are the row vectors of the *Indicator matrix* of grouping factor indices, $\mathbf{J}_i$ and the *raw random effects model matrix*, $\mathbf{X}_i$. The subscript $i = 1, 2, ..., k$ denotes the specific random effect term.

2.3.2.2 Spherical Random Effects - U

For computational efficiency, the `lme4` package in R, which is used for training LMMs in this study, reformulates the model such that the *covariance parameter* θ appears only in the conditional distribution of the response vector given the chosen random effects, as opposed to the earlier formulation where θ appears only in the *marginal distribution* of the random effects. The reformulation is achieved by defining a *spherical random effects* variable, $\mathcal{U}$ which is distributed as, $\mathcal{U} \sim \mathcal{N}(0, \sigma^2 I_q)$. Consequently, $\mathbf{U}$ is set to, $\mathbf{U} = \Lambda_\theta \mathcal{U}$. This allows, $\mathbf{U}$ to take on the desirable distribution, $\mathbf{U} \sim \mathcal{N}(0, \Sigma_\theta)$. The motivation behind this transformation is that this reformulation allows us to work with singular covariance matrices, Λ_θ. The final reformulated LMM maybe now written as,

$$(\mathcal{Y}|\mathcal{U} = u) \sim \mathcal{N}(\mu_{\mathcal{Y}|\mathcal{U}=u}, \sigma^2 \mathbf{W}^{-1}) \tag{2.5}$$

where, $\mu_{\mathcal{Y}|\mathcal{U}=u} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\Lambda_\theta u + \boldsymbol{\varepsilon}$, is the conditional mean which is the vector of linear predictors.

2.3.2.3 Extensions and Limitations of LMMs

The LMM can be further extended to include additional random effects at different levels of nesting, such as multiple measurements within subjects, subjects nested within groups, and so on. The model can also be extended to include non-linear terms, interactions, and other more complex structures. The LMM is typically estimated using maximum likelihood estimation (MLE) or restricted maximum likelihood estimation (REML), which involves estimating the fixed and random effects coefficients as well as the residual variance from the observed data. The estimated coefficients can then be used to make inferences and predictions about the relationship between the response and the predictors.

But there are several limitations to choosing linear mixed models as a modelling choice. Firstly, they assume that the relationship between the outcome variable and the predictors is linear. This assumption may not hold in many cases, especially if the relationship is complex or nonlinear. Secondly, they assume that the residuals are normally distributed. If the residuals are not normally distributed, then the LMM may not provide an accurate description of the data. Also, LMMs assume that the variance of the residuals is constant across all levels of the predictors. This assumption may not hold in some cases, especially if the variance changes over time or varies by group. The authors, Diggle et al. suggest that parametric models can be used to address the limitations of LMMs by providing more flexibility in modeling the relationship between the outcome and predictors, handling non-normal distributions, and accommodating varying variances of the residuals.

2.3.3 Generalized Linear Mixed Models

Generalized Linear Mixed Models (GLMMs) as an extension of the LMMs, where the response variable may not be normally distributed, and the distribution belongs to the exponential family. GLMMs are useful when the variance of the response variable is not constant

or when there are non-linear relationships between the predictors and the response. Three different extensions of GLMs are captured and explained by Diggle et al. (ch. 7 [10]) for analyzing longitudinal data comprehensively:

1. *Marginal Models*: These models only consider the marginal distribution of the response variable, without considering the correlation structure of the data. Marginal models use the population-averaged or marginal mean response over time and treat the covariance structure between observations as a nuisance parameter. Marginal models are often used when the primary interest is in the population-average response and when the correlation structure of the data is not of primary interest.

2. *Random Effects Models*: Defined as a type of mixed model that assumes that the individual trajectories of a response variable within a population are related to one another through a set of unobserved or latent variables, referred to as random effects. These random effects represent the systematic variation between individuals in the population and are assumed to follow a specific distribution. There are 3 main assumptions in a random effects GLM:

   a) $Y_{ij}|\boldsymbol{U}_i$ follows a distribution from the exponential family with the pdf $f(y_{ij}|\boldsymbol{U}_i;\boldsymbol{\beta})$

   b) $Y_{i1}, ..., Y_{in}$ are independent observations given underlying random effects structure $\boldsymbol{U}_i$

   c) the underlying random effects, $\boldsymbol{U}_i$ are i.i.d. with pdf $f(\boldsymbol{U}_i; G)$

3. *Transition (Markov) Models*: When the choice is to model the outcome $Y_{ij}|Y_{ij-1}$ as a probability distribution. Transition models are useful when the variables of interest are categorical or ordinal, and the goal is to model the probability of moving from one state to another over time. Transition models can be thought of as a type of Markov model, where the transition probabilities depend on the current state but not on the past history.

2.3.3.1 MCMC simulations for GLMMs

While GLMMs offer a very flexible bouquet of modelling a wide variety of data, their limitations with non-Gaussian data leads to the issue of the likelihood not being available in the closed form. To solve this issue we can adopt Markov Chain Monte Carlo methods to sample from simpler distributions. A concise description of the application of MCMC routines to GLMMs in a Bayesian context is described as follows [22].

The model form has the following 3 components:

1. The probability of the data, y, given a latent variable, l, is given by the p.d.f. $f_i(y_i|l_i)$ for the i_{th} data point.

2. The latent variable is predicted by a linear mixed model, $l = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + e$, where $\mathbf{X}$ is the model matrix of the features, $\mathbf{Z}$ is the design matrix relating to the random predictors. The parameter vectors of the model are $\boldsymbol{\beta}$ & $\mathbf{u}$ also referred to as location effects, while e is the vector of residuals.

3. The parameters i.e. the location effects and residuals are assumed to belong to a multivariate normal distribution,

$$\begin{bmatrix}\boldsymbol{\beta}\\ \mathbf{u}\\ \mathbf{e}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\beta_0}\\ 0\\ 0\end{bmatrix}, \begin{bmatrix}\mathbf{B} & 0 & 0\\ 0 & \mathbf{G} & 0\\ 0 & 0 & \mathbf{R}\end{bmatrix}\right)$$

   where, fixed effects are given by a vector of prior means of $\boldsymbol{\beta_0}$ and covariance matrix $\mathbf{B}$. The default values are used in this study and they are specified as a zero mean

vector MVN prior distribution with a diagonal variance matrix with a large variance $(1e+10)$. **G** & **R** are the expected covariances of random effects and the residuals. The co-variance matrices by default are assumed to follow a conditional inverse-Wishart prior distribution (p.9 [15]) which takes 2 parameters `V` and `nu`.

Most models that are constructed with non-Gaussian data do not have a known distribution of the latent variable l. It is updated by using the Metropolis-Hastings algorithm, which generates a sequence of proposed parameter values based on a proposal distribution and accepts or rejects each proposal based on the ratio of the posterior distribution of the proposed parameter value and the current parameter value. The mixed model parameters, $\boldsymbol{\beta}$ & $\mathbf{u}$ are Gibbs sampled in a single block[12] as their covariance structures (**G** & **R**) both follow an inverse-Wishart distribution which is a conjugate prior.

3 Method

3.1 Data

The data used in this thesis study is taken from 2 sources. The data on auction prices and salaries of the various players have been taken from the official website of the Indian Premier League [25]. Along with the final auction prices, the website also gives the role and nationality of the player that was bought at the auction. The player performance data has been taken from a popular online database known as Cricsheet [28]. Cricsheet.com primarily focuses on providing ball-by-ball data for cricket matches, which includes details such as the bowler, batsman, runs scored, and the events that occurred on each ball (for eg. wicket taken or boundary scored). The website also provides data related to match information, such as team lineups, match location, and the match result. Due to the level of detail available for player performance data, the final data was constructed in multiple steps which are detailed now.

3.1.1 Data Compilation

The most relevant data for this modelling problem is the performance of the players in the past seasons of the IPL itself. While most of the related research has only considered the performance from the latest season of the IPL as the predictors of their modelling exercises, we are attempting to explore a new direction of analysis by considering the past IPL performance data for each player in two buckets:

1. *Previous Season performance*: If a player's auction price of season x is being considered, here we specify the performance data of the season $(x-1)$ as the relevant predictor for the auction price of season x.

2. *Cumulative Season performance*: Here we specify the aggregate performance data of the player upto and including season $(x-1)$ as the relevatn predictor for the auction price for season x.

It was mentioned in section 1.1.2 that the IPL is not a year round event and typically takes place in the months of April and May each year. This leaves players free to participate in other similar competitions throughout the year. Therefore, the player's performance in these other

competitions also becomes relevant to their perceived value for the upcoming IPL season during the auctions. The players' aggregate performance data from the match data of the last 12 months in leading upto the IPL auction date of year x of the following competitions were considered on the grounds of close resemblance to the IPL and/or recognition of significant competitive level of gameplay and talent involved in the competition.

1. Big Bash League: The top T20 league in Australia
2. Caribbean Premier League: The T20 league of the West Indies
3. Bangladesh Premier League: T20 league of Bangladesh
4. CSA T20 Challenge: T20 league organized in South Africa
5. Pakistan Super League: T20 league organized in Pakistan
6. Super Smash: T20 league organized in New Zealand
7. T20 blast: T20 league organized in England
8. Lanka Premier League: T20 league organized in Sri Lanka
9. International T20: T20 league organized in UAE
10. Syed Mushtaq Ali Trophy: Domestic T20 league organized in India that features only domestic players. This was done to include performance metrics of lesser known player with little to no international acclaim. Such players have no presence in cricket outside of India as the BCCI doesn't allow Indian players to play in foreign leagues in order to protect the IPL brand. Most of the domestic players as a result don't have any statistics outside of the IPL other than the domestic circuit in India.
11. The Hundred: This is a new format that is very similar to T20 which has been introduced in England. There are a few rule changes that make it different from the T20 format, but the gameplay and playstyle is very similar to the T20 format. As a result, player performance from this tournament have also been included.

One final category of match data considered relevant is the players' performances in T20I matches. These matches are the matches played at the highest level in which the countries select their best 11 players to compete with other countries. The skill and competition in these matches is the highest and features the most competitive results. A lot of teams scout international talent based on their performance in these type of matches. Although T20I are a very important part of the game, with respect to this thesis, these apply to a very a small subset of players. Mostly, overseas players in the auction will have game data for these type of matches.

3.1.2 Player Performance Data

For each of the competitions that are mentioned in the previous subsection, the following player performance indices have been chosen that completely capture the 2 main roles that a player can be required to fulfil in a given match i.e. Batting and Bowling. For each of these roles the following performance stats have been captured in table 3.1

Since, a lot of the players in the IPL are not present across all competitions, we aggregate each player's performance across all leagues in the previous year into one cumulative category which captures a player's performance across the metrics mentioned in the table for the entire year leading upto the auction. We do this by adding each metric of the competitions chosen with their corresponding metric from other competitions, for example, all if a player has played in leagues A, B and C and played 5 matches in each competition in the previous

Performance Measure	Description
bat_I	Total number of times the player batted. Note that this maybe different from total matches played as not all players playing in a match maybe required to bat
bat_R	Total Runs scored by the player
bat_B	Total Balls faced by the player
bat_Outs	Total number of times the player was dismissed by the opposition
bat_4s	Total number of 4s hit by the player
bat_6s	Total number of 6s hit by the player
bat_50plus	Total number of innings where the player was able to score more than 50 runs
bat_100plus	Total number of innings where the player was able to score more than 100 runs
bowl_I	Total number of times the player bowled. Note that this maybe different from total matches played as not all players playing in a match maybe required to bowl. Typically only 5-6 players bowl in a match
bowl_B	Total number of balls bowled by the player
bowl_R	Total number of runs scored off the bowler by the opposing team
bowl_W	Total number of wickets taken by the player
bowl_4W	Total number of times a bowler took >4 wickets in a match
bowl_4s	Total number of 4s conceded by the bowler
bowl_6s	Total number of 6s conceded by the bowler
bowl_Dots	Total number of balls bowled by the player where the opposition was not able to score any runs

Table 3.1: Raw Player Performance Metrics

year, then their aggregate performance data would show 15 matches played in the previous year across other competitions. We form a similar bucket for the combined performance in all seasons of IPL until the year of consideration.

The player performance data for modelling has been used in the long form, i.e. each player has a separate row for each year of participation in the IPL and the columns are a set of features that capture the player's batting and bowling performance metrics for the previous year across the IPL, all other competition, and the aggregate performance in the IPL till the auction date. A visualization of the performance data has been illustrated in table 3.2. The columns marked as *Last 12 Months* and *All IPL Seasons* serve as placeholders for each performance metric within this category of performance data aggregation. To understand the final features that are used in the modelling in this study we now go into detail about how the raw performance data from table 3.1 is transformed into the final feature set.

		Last 12 months		**All IPL Seasons**	
Player Name	**Season Num**	**Batting Ability features**	**Bowling Ability Features**	**Batting Ability features**	**Bowling Ability Features**
A	1				
A	2				
B	1				
B	2				
B	3				
C	1				
C	2				

Table 3.2: Player Performance Data Illustration

3.1.3 Feature Engineering - Derived Metrics

As explained, we are dealing with a varied list of player performance metrics that captures a player's contribution to their games. These input metrics in their own way define a player's ability which in turn allows teams and analysts to identify a player's worth. Some of these metrics are raw measurements which reflect an instance of play for example a ball bowled, a run scored of an individual delivery or a boundary scored by the batsmen. These metrics in aggregate give a cumulative performance perspective of a player, for example - total matches played, total runs scored or total wickets taken etc. These metrics lack the ability to capture the impact a player creates that is comparable with other players. For this study, we now define a set of metrics called - *Derived Metrics*.

The idea behind defining such metrics is to have a set of performance measures that allow us to compare players irrespective of the volume of matches these players play and allow us to evaluate the impact they produce when they play. Prakash et al. have proposed a list of metrics or indices for batting and bowling [9] that effectively identify top performers among the competition. These derived metrics have been defined as follows:

1. Batting Ability:

   a) A key ability in T20 cricket is to score runs quickly since each innings is very short. The fastest way to score runs is to score boundaries i.e. 4s or 6s. So we use the Hardhitter metric to capture how many boundaries a batsman is scoring in the time he spends in middle playing i.e., in the number of balls he faces.
   $HardHitter = \frac{(\text{Fours} + \text{Sixes})}{\text{Balls Faced by the player}}$

   b) For a batsman, one of the key challenges is to correctly judge the playing conditions and the match situation. In cricketing lingo, a batsman takes sometime to get his *eye in*. This is referred to the time a batsman spends before he feels comfortable in scoring at a faster rate. The time spent in gaining this comfort level can often lead to a lot of time lost for their team in accumulating more runs. What's more is that everytime a new batsman comes to middle, it means more time and momentum lost for their team. Hence, if a player is able to not lose his wicket and play for as long as possible, it often means less time lost where the team is scoring at a lesser rate. Thus, we define the metric of a *Finisher* as follows:
   $Finisher = \frac{\text{Count of times batsman was not-out}}{\text{Total Innings played}}$

   c) A popular index used to analyze batting performance is *strike rate*. It is used to measure how many runs would a batsman score per 100 balls faced. This is important in the context of T20 cricket, because the objective of the match is to score as many runs in the limited number of balls available. We use this index against the label *FastScorer*.
   $FastScorer = \frac{\text{Runs Scored}}{\text{Balls Faced}} \times 100$

   d) Another popular index to measure the consistency of a batsman's performance is called the batting average. It measures how many runs on average does a batsman score every time they bat. We use the label *Consistent* to define this derived metric as follows:
   $Consistent = \frac{\text{Total Rus Scored}}{\text{Count of outs}}$

   e) Sometimes matches are won due to significant individual performance that heavily influences the outcome of a match. We define a metric *BigScorer* which is defined as:
   $BigScorer = \frac{\text{Count of times>50 runs scored}}{\text{Count of times batted}}$

   f) Even though a batsman's objective is to score more runs at a faster pace, it is quite challenging to hit boundaries at will for a batsman. The bowler is bowling with

objective to stop the batsman from scoring and the fielders of the opposing team are also on the field trying to prevent boundaries. Therefore, another important contribution for a batsman is to accumulate runs even when they are unable to hit boundaries. We capture this by introducing the metric *Running Between Wickets or RBW* defined as follows:
$RBW = \frac{\text{Total Runs Scored} - (4 \times \text{Fours} + 6 \times \text{Sixes})}{\text{Total Balls Faced} - (\text{Fours} + \text{Sixes})}$

2. Bowling Ability:

   a) When evaluating bowlers, we need to look at their ability to stop the batsmen from the opposing team to score runs. Thus, a popular index called *Economy Rate* is used here as a derived metric. A bowler is allowed to bowl overs of 6 balls at a time so the economy rate is measured as runs given away per six balls or per over. So the metric is defined as:
   $Econ = \frac{\text{Runs given away}}{\text{Total balls bowled}} \times 6$

   b) A prized ability for bowlers is their ability to take wickets while giving away as few runs as possible. Thus we define *WicketTaker or Bowling Avg* as the bowlers average runs given before the bowler takes a wicket.
   $Avg = \frac{\text{Runs Given Away}}{\text{Total Wickets Taken}}$

   c) Another popular evaluation metric for bowlers is their ability to take as many wickets as possible whenever they bowl. So we define *Bowling Strike Rate* as the number balls bowled before the bowler takes a wicket.
   $BowlSR = \frac{\text{Total Balls Bowled}}{\text{Total Wickets Taken}}$

   d) Just like a batsman can influence the outcome of a match by scoring a big amount of runs, similarly a bowler can influence the match in favour of his team by taking a lot of wickets. We define *BigWickettaker* as the frequency with which a bowler takes 4 wickets or more everytime he bowls in a match.
   $BigWickettaker = \frac{\text{Count of times } >4 \text{ wickets taken}}{\text{Total times bowled}}$

   e) Just like we measure a batsman's ability to score runs in boundaries, we also measure a bowler's ability to not give away runs en masse. We define *46conceded* as the number of times the opposition is able to score boundaries off the bowler.
   $46conceded = \frac{\text{Total boundries given away}}{\text{Total balls bowled}}$

   f) A very crucial and strategic value that a bowler can create in a match, is by not letting the opposition score of his bowling. T20 is a fast paced game and as such balls off which no runs are scored are a rarity and a very precious occurrence for the bowling team. We define *Dots%* as a bowlers ability to ball dot-balls i.e. when no runs are scored.
   $Dots\% = \frac{\text{Total dot balls}}{\text{Total balls bowled}}$

An important advantage of using these derived metrics, is that even though we are creating these indices as a transformation of multiple raw indices, we are able to do so without introducing high correlations among the derived metrics for batsmen (3.1), bowlers (3.2) & all-rounders (3.3). These transformed *derived metrics* will now serve as the final features for the predictive modelling we will go into in the coming sections in this chapter.

3.1.4 Categorical Features

A final group of categorical features are included in the final feature list and they are described as follows:

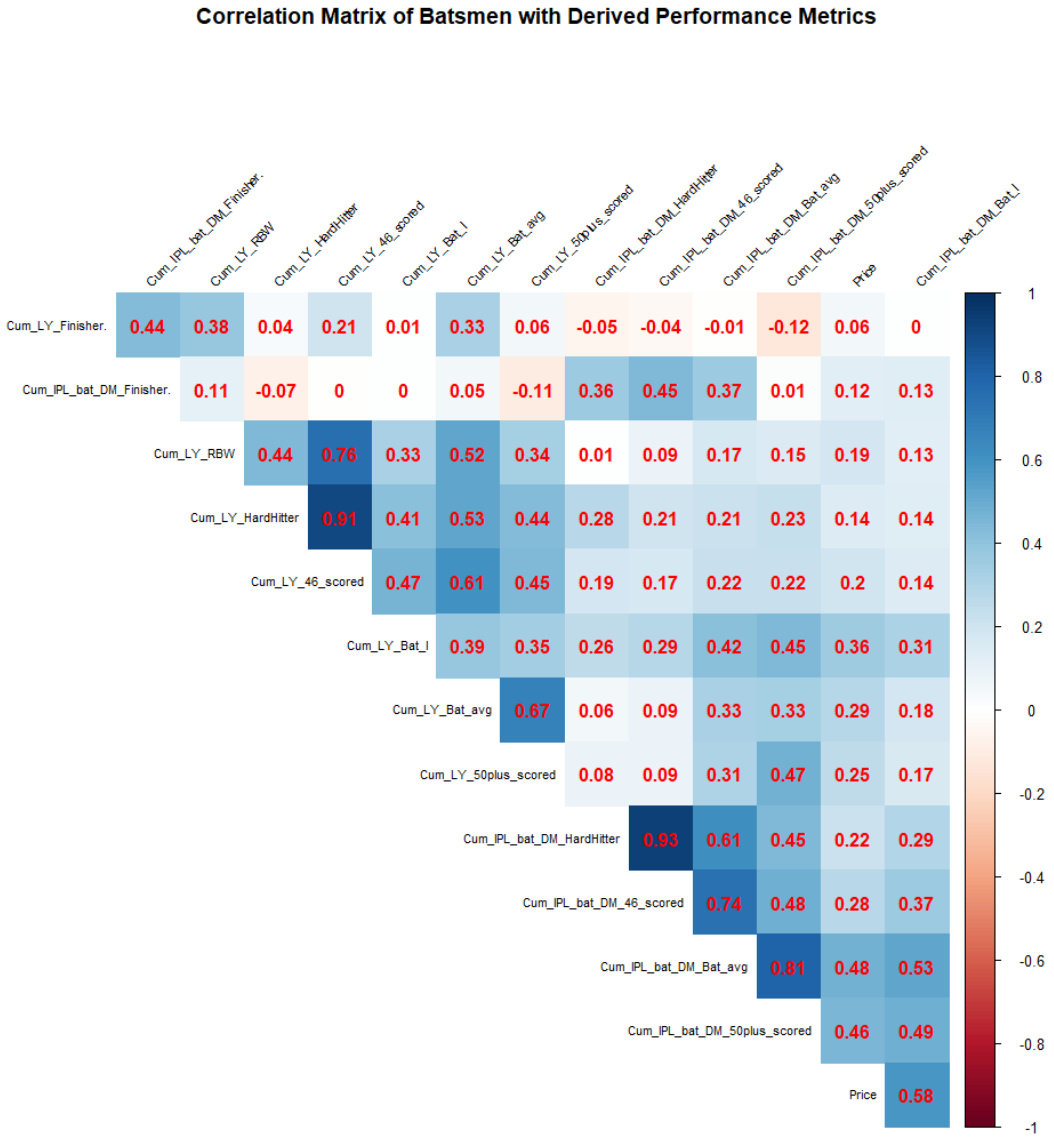


Figure 3.1: Correlation Heatmap for Batsmen

1. *Overseas Player*: This variable is used to indicate if the player is a player whose nationality is Indian or any other nationality. All international players i.e. non-Indians will have this variable value set to 1 and all domestic players i.e. Indians, will have this value set to 0. This variable is of interest because the IPL teams are only allowed to include a maximum of 4 international players in any match out of the total team size of 11. So, this becomes a strategic input for teams to consider when they bid for players in the auction and try to balance their team rosters.

2. *Wicket Keeper*: This variable is used to indicate if the player also has the additional skill of being able to take on the role of a *wicket keeper*. The wicket keeper plays a very important role when the team is fielding. The primary responsibility of a wicket keeper is to catch the ball when the batsman edges or misses it. They must have excellent reflexes, agility, and hand-eye coordination to take these catches. Additionally, a wicket keeper can attempt to dismiss a batsman by stumping them. This happens when the wicket keeper collects the ball and removes the bails while the batsman is out of their crease. There are many other layers to the contribution of a wicket keeper, but we

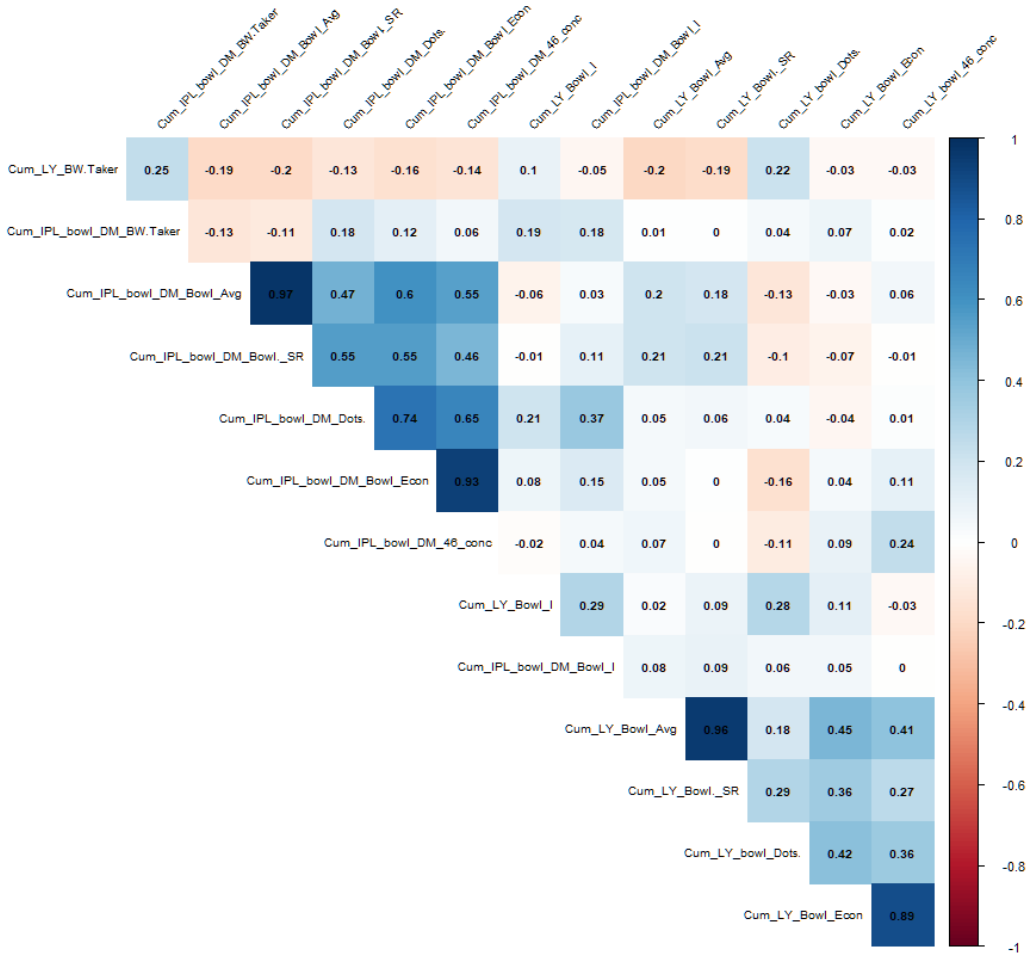


Figure 3.2: Correlation Heatmap for Bowlers

leave that for independent research on the part of the reader. In this study, we have been unable to include any fielding related performance data so a lot of wicket keeping skills will go unaccounted for. There is a possibility that many wicket keepers in the data might become outliers and affect a model's prediction accuracy. To address this somehwat, we will include this categorical feature which is set to 1, if the player is also a wicket keeper and is set to 0 otherwise.

3. *Active T20I Player*: This variable is included to see if the player has played in their national team in international matches against the national teams of other countries. This is considered relevant information as it is considered to be a significant achievement for a player to prove their skill levels enough to be picked to represent their country at the highest level of competition against the top players from other countries. This variable value is set to 1 if the player has played for their national team 1 or more times in the last 12 months and 0 otherwise.

4. *Star Player*: In a franchise league like the IPL, it is often the case that certain players are valued above others for their intrinsic value in addition to their play making skills. Examples of such players would be MS Dhoni, Virat Kohli, AB De Villiers, Keiron Pollard among others. What sets these players apart is that they may not be active players on the international circuit, for example MS Dhoni is a former captain of the Indian cricket team who retired from international cricket in 2019 but continues to play the IPL and

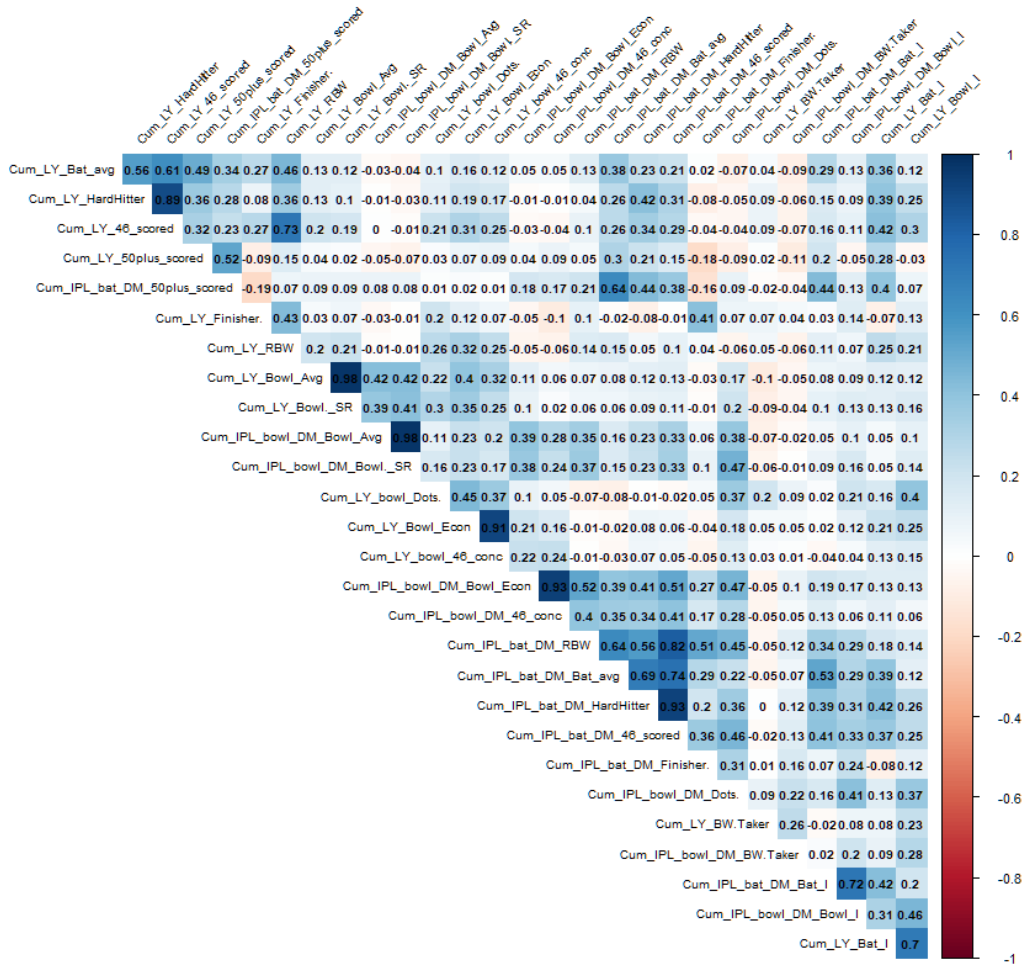


Figure 3.3: Correlation Heatmap for All-Rounders

draws a significant salary from his franchise even though he has not made a very significant contribution with the bat in the recent years. While there exists no well established algorithm to categorize players as such, for this study we adopt a simple methodology to capture this key factor. We consider players who have in the past played for their national teams (even if they havent played in the national team in the last 12 months), or if the player is an overseas player and has played for their national team in the last 5 years, then they are considered to be *star players* and have their value set to 1 for this categorical variable. All other players have their value set to 0 for this variable.

The illustration for visualizing the final transformed data with all the features can be seen in table 3.3. Note that this table is just an illustration and contains dummy data and doesn't demarcate each individual feature as described in the section explaining derived metrics. For illustration the table shows all the batting and bowling ability features as a single feature when in actuality each derived metric would be their own feature. It is emphasised again that the reader should note that this panel or longitudinal data will be used in its long form i.e. each observation of an individual player is a separate row as opposed to the wide-form of data where each separate observation of a player would be a separate column. This means that the data consists of multiple rows for the same player across seasons.

						Last 12 months		All IPL Seasons	
Player Name	**Season Num**	**Overseas Player**	**WK**	**Star Player**	**Active T20I Player**	**Batting Ability features**	**Bowling Ability Features**	**Batting Ability features**	**Bowling Ability Features**
A	1	1	0	0	0				
A	2	1	0	0	0				
B	1	1	1	1	1				
B	2	1	1	1	1				
B	3	1	1	1	1				
C	1	0	0	1	1				
C	2	0	0	1	1				

Table 3.3: Final Data Set Illustration

3.2 Exploratory Data Analysis

3.2.1 Data Preprocessing

The full data combined from all data sources of various competitions including past seasons of the IPL consists of 2185 combined observations of a total of 559 players observed over 15 playing seasons and 1 upcoming season of the IPL. For this study, we do not consider the data from the first 5 seasons of the data because the starting seasons of the IPL witnessed many rule changes and a lot of older players slowly phasing out of the league after they had helped setup the IPL brand. There were some rules in place that mandated higher salary for some star players irrespective of their performance levels. To avoid this outlier behavior, the first 5 seasons were dropped from the data used for modelling. After dropping observations from the first 5 seasons we have 1857 observations from 559 players. The next step was to identify those players for whom we don't have any performance data available. Inspection yielded that these are cases of players which might have played a different format of cricket like Tests or ODIs at the highest level based on which they received bids. For some domestic cases this might also be due to injuries that these players were not able to participate in any recent games. Some of the player might not have played in any of the competitions being included in the analysis. There were 181 such players with a total of 216 observations that had to be removed from the data.

Next we look at role split where the data consists of 576 bowlers, 530 all-rounders, 376 batsmen and 159 wicket-keepers as depicted in figure 3.4. Since, the data for wicket-keeping stats like stumpings affected and catches taken is not part of the data available for modelling, we will combine the wicket-keepers with batsmen since, wicket-keepers rarely ever bowl in a game and their secondary role in the team is that of a batsman. This will result in a total of 535 batsmen in the data. The player role categorization has been taken from the official website of the IPL [25] for most players. For the other players whose role info was not present from the official website, the role specification as mentioned on a top cricket database website, espncricinfo.com [3], has been used.

The players either belong to the Indian domestic circuit or are categorized as overseas players otherwise. In the compiled data we deal with 995 observations of Indian domestic players and 646 observations of overseas players. Another distinction that was created was that of *Star player*. The data includes 1005 observations of star players versus 636 observations of regular players. This skew can be reasoned with a higher churn that is expected from international stars. Their availability, form and injury status makes international players more prone to replacement and absenteeism from the IPL.

We look at some exploratory plots to understand the distribution of the salary data in figure 3.7. We can observe from the density plot in 3.5 that the batsmen and all-rounders tend to attract higher salaries as the density plot has a fatter tail at the higher price marks. The bowlers on average have much lower salaries and in very few cases does a bowler get a very high bid from the teams. This would lead to the initial belief that batting a more vital skill in T20 cricket than bowling. This can also be seen from the figure 3.6 which shows the salaries of all the players across the seasons with a trajectory for the top 10 earning players in the history of the league.

Another very important observation can be made about the data from figure 3.6. We can see that not all players have equal number of observations. This is important to note and discuss. In a typical panel study, we expect all subjects to have the same number of observations for eg. patients in a weight loss study would be expected to have the same number of observations over the time period of the study. This case is slightly different because here not all players play in each season of the IPL. This means that some players will play in maybe a few seasons of the IPL, like SR Tendulkar, who played the initial few seasons and then retired. The most common example is of multiple players who fall out of form and favour and get overlooked. Only very rare cases of players who have played each season of the IPl exist like

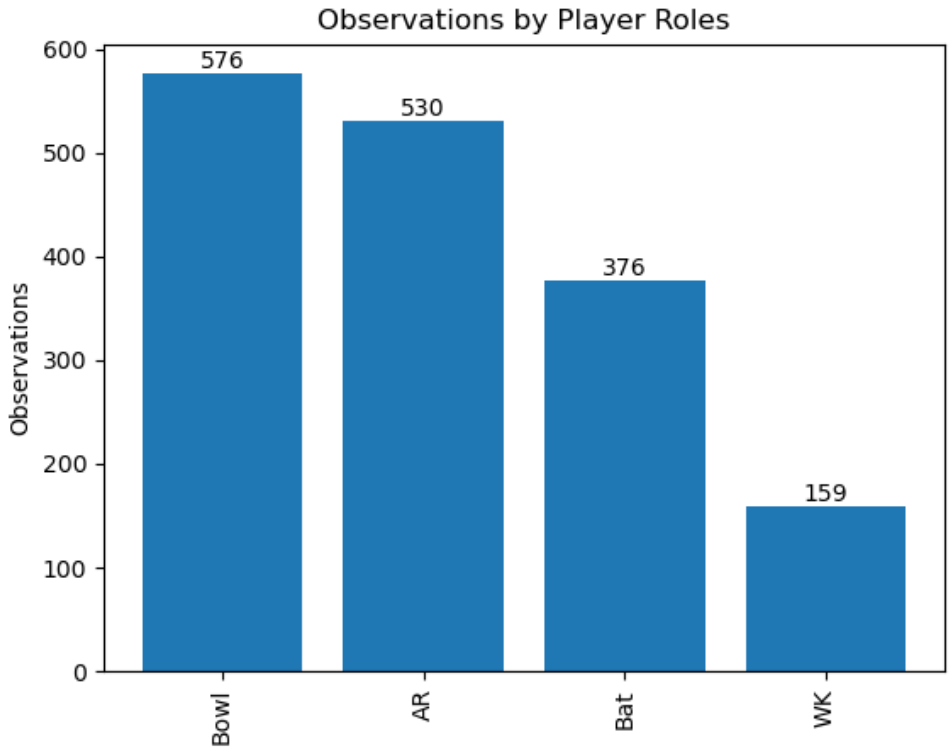


Figure 3.4: Player Role Distribution

MS Dhoni, who has played in every single season. Therefore, this is not a case of missing data where we can impute player salaries over time. It would actually be situationally incorrect to impute data as it would ignore the widely accepted principle of sports that as a player ages, their playing ability weakens over time. Thus, for this study, we will not employ any data imputation and train model with actual data.

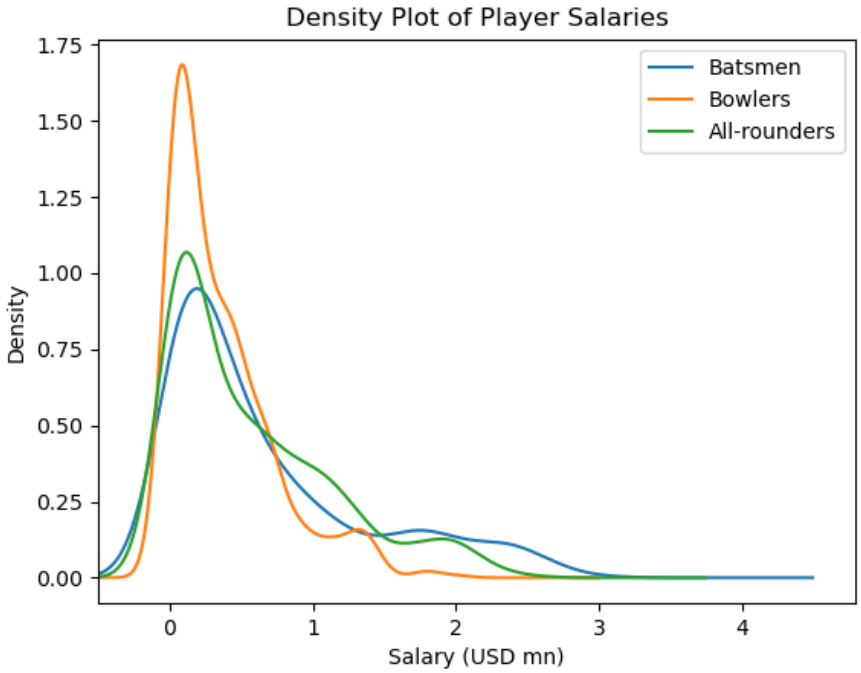


Figure 3.5: Salary Density by Player Role

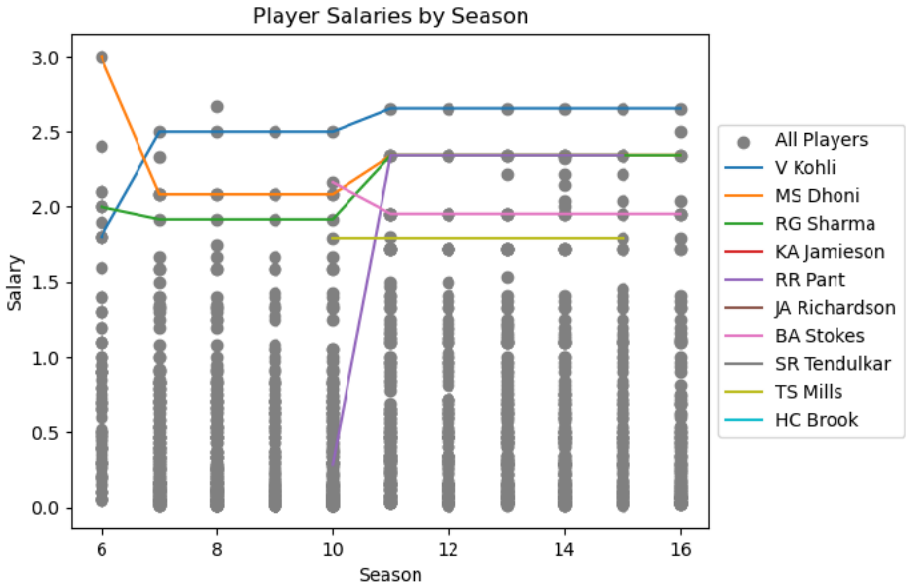


Figure 3.6: Top 10 Player Salaries vs Other Players

Figure 3.7: Player Salaries by seasons

The boxplots in figure 3.11 also show how for each role there are outliers of player that invite high bids and salaries each season. From the boxplots we also observe more clearly that the bowlers have a much smaller and lower range of salaries when compared to the role of batsmen. This also supports the previous reasoning that batsmen are more highly prized than bowlers in T20 cricket. All-rounder salaries follow a similar density curve as the batsmen since their batting ability is more highly prized than their bowling ability in most cases.

The other smaller and final preprocessing steps are explained below.

1. Player Salaries were transformed and rescaled to be expressed in million US dollar amounts.
2. Player performance measures were scaled with the exception of the categorical variables - wicketkeepers, overseas player, T20I player and star player.
3. The data was split by role separately i.e. batsmen, bowlers and all rounders, with only role relevant features being kept for each role. This was done to model each role separately according their role specific relevant features.
4. The data was finally split into *training and test sets*. Data from the seasons 6-12 was chosen to be the training dataset and data from the remaining seasons 13-16 was selected for the test dataset.

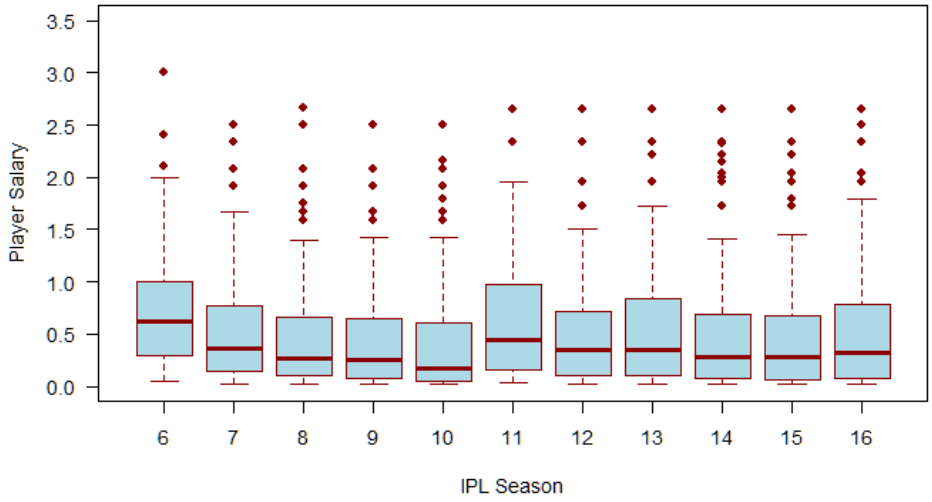


Figure 3.8: All Roles

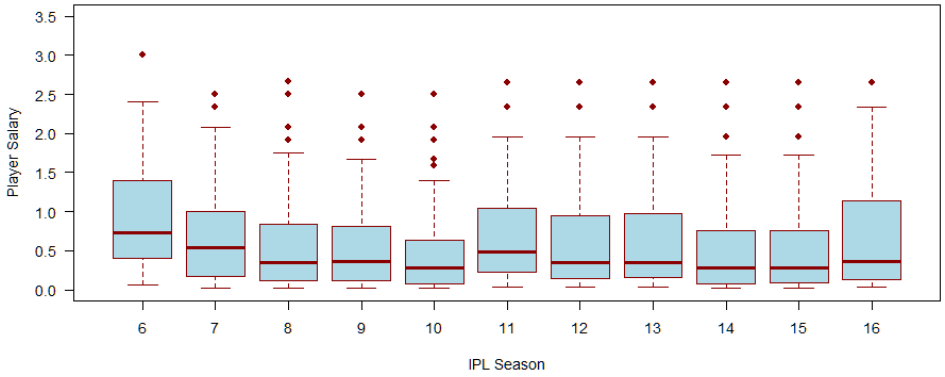


Figure 3.9: Batsmen only

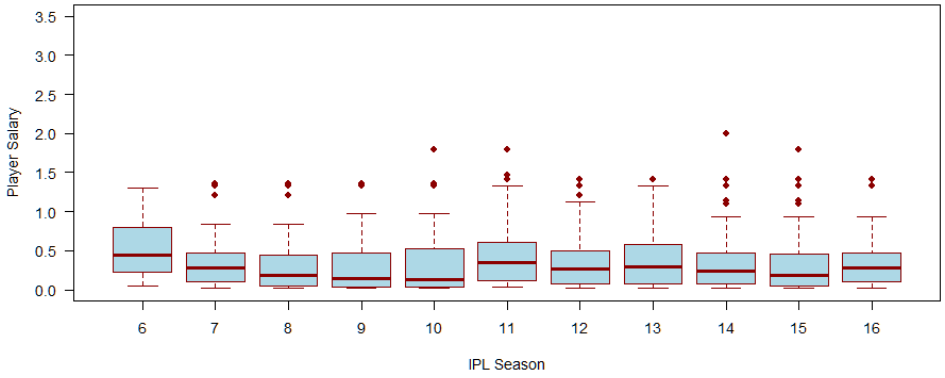


Figure 3.10: Bowlers only

Figure 3.11: Salary Boxplots by roles and seasons

3.2.2 Clustering

An observation from the data exploration was that we will be dealing with outliers for each player role in our analysis. A strategy that was explored to deal these outliers was to using clustering to identify similar players and model them together to handle outliers. By running a k-means algorithm and experimenting with different configurations, there were 4 sensible clusters found. The wordclouds for these clusters are shared in figure 3.12. A possible interpretation for these clusters can be:

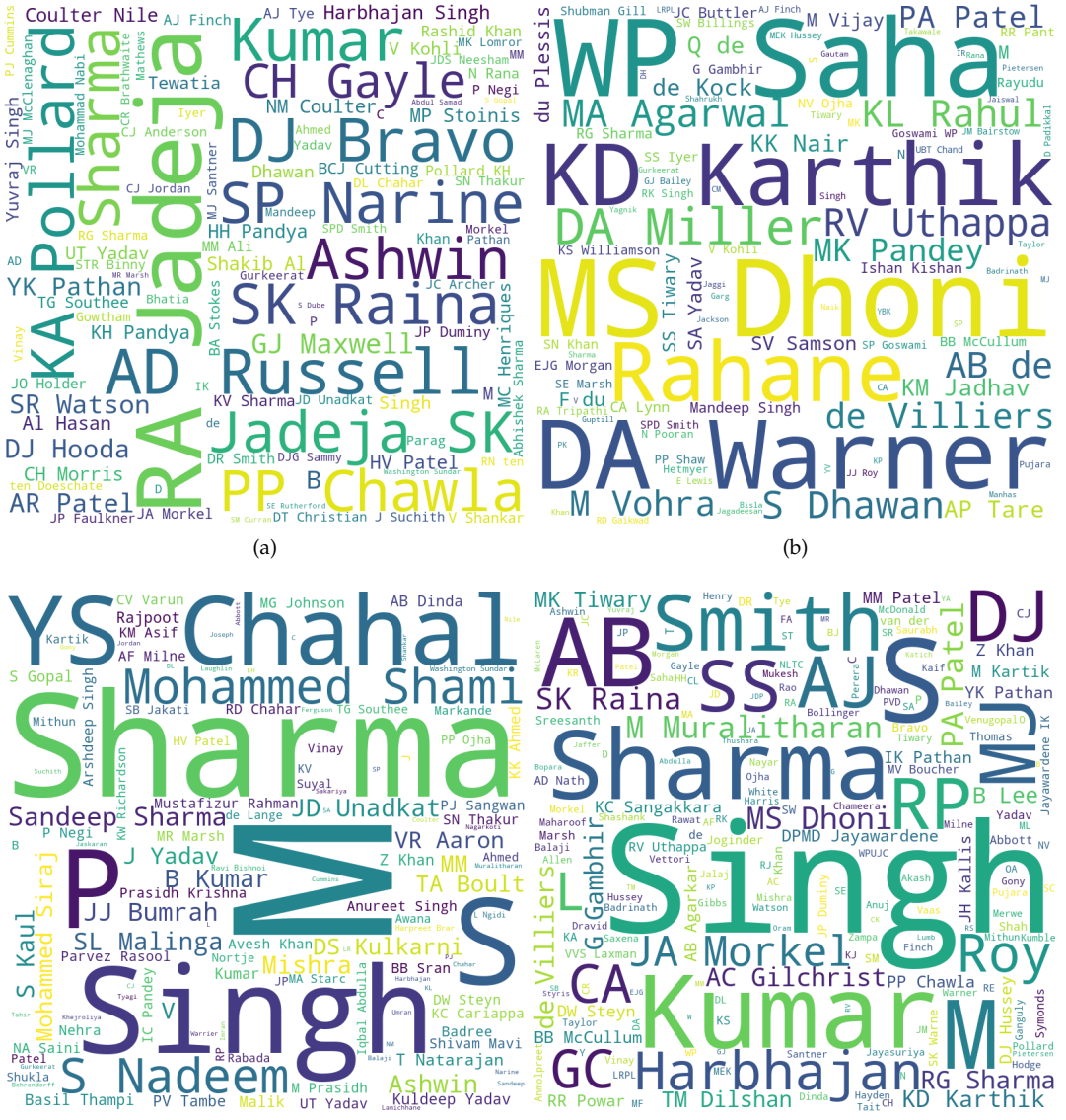
1. *Bowling All-rounders*: Highlighted by world renowned all-rounders such as KA Pollard, RA Jadeja and AD Russell among others.

2. *Pure Batsmen*: This cluster is highlighted by some well known specialist batsmen like DA Warner, DA Miller, KL Rahul. Interestingly, this cluster also features the wicketkeepers prominently. MS Dhoni, KD Karthik and WP Saha and well known wicketkeepers in the league and would have rarely, if ever, bowled in an IPL match. This clustering makes sense as most of the players would have non-existent bowling indices.

3. *Pure Bowlers*: This cluster is highlighted by players who rarely get to bat in an IPL match. Due to the nature of the game and rules, it is very likely that a lot of pure bowlers would have some batting data, but the data would be very infrequent and also very different from a pure batsmen. For eg. a pure bowler would very rarely have scored a significant amount of runs.

4. *All Rounders or Star Players*: This cluster is by far the most incoherent. We see a mix of batting all-rounders like JA Morkel, Yuvraj Singh with star wicketkeeper batsmen like AC Gilchrist and MS Dhoni. We also see M Muralitharan who is a legend of the game but a pure bowler. A standout feature of this cluster is that it has very high profile names, but no clear role boundaries.

The biggest issue arising out of using these clusters is that because batsmen and bowlers are mixed into different clusters, we can not avoid using less number of features which means we always ended up with higher error rates in prediction when using these clusters. Another very glaring shortcoming of using clusters was that since the data is in long form, the same players found themselves in multiple clusters. This is not ideal because then we are not really modelling the players on their cumulative performance rather we are modelling the same player multiple times. This is inadequate because in a real world scenario, a new prediction would be very hard to make since, the choice of which model to pick for predicting the latest price would be ambiguous and not clear. As a result, it was decided to not use the cluster outputs for the recommended modelling choices in this study.



(a) (b)

(c) (d)

Figure 3.12: k-means Clustering Word Clouds

3.3 Model Evaluation

This study requires extensive experimentation and as such will require a suitable choice of evaluation parameters. Since, this is a predictive modelling exercise, we will have a three step approach in evaluating and selecting the models that we train. Firstly, we will be looking at the error results to check the generalization ability of the trained model. Secondly, we will also be evaluating the models themselves by using information criterion for model selection. Finally, we will perform variable selection to identify the least complex model with significant predictors.

3.3.1 Prediction Accuracy Measures

1. *RMSE* or Root Mean Square Error is a scale dependent accuracy measure whose scale depends upon the scale of the data [18]. It is a suitable measure in this case as we are going to be comparing models that have been trained on the same training and test datasets. When the prediction error e_t, is given by the difference between actual values Y_t, and predictions F_t, i.e. $e_t = Y_t - F_t$, we can define,

$$RMSE = \sqrt{mean(e_t^2)}$$

2. *MAPE* or Mean Average Percentage Error is an accuracy measure based on percentage error, $p_t = e_t / Y_t \times 100$ [18]. While percentage based errors are mostly used to compare model performance across different datasets, it makes sense to use here because the prediction quantity is player salary which positive and non-zero. It also gives a more practical and intuitive interpretation of the accuracy of the predictions. Percentage error allows a quick understanding of how *off* the predictions are in terms of dollar salaries.

$$MAPE = mean(|p_t|)$$

3.3.2 Information Criteria

The most general summary of checking predictive fit is the log-likelihood or the log predictive density, $\log p(y|\theta)$. The log-likelihood is typically used to compare different models or to estimate the parameters of a model using maximum likelihood estimation. In general, the larger the log-likelihood value, the better the model fits the data. This is because the log-likelihood function takes into account both the accuracy of the model in predicting the observed data and the complexity of the model. The log-likelihood function penalizes complex models that overfit the data by assigning a lower probability to unlikely data points.

Information criteria are statistical measures used for model selection in the context of hypothesis testing and statistical inference. They provide a way to compare the performance of different models, taking into account both model complexity and goodness of fit. They are typically defined using *deviance*.The deviance D is defined by the expression, $D = -2log(Pr(y|\hat{\theta}))$ where, $\hat{\theta}$ is a point estimate (typically found by MLE) and posterior distribution $p_{post}(\theta)$ is fit to the data y [13].

Two of the most commonly used information criteria are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Both of these criteria are based on the likelihood function of the model, penalized by a term that increases with the number of parameters in the model. AIC and BIC can be used to select the *best* model among a set of candidate models. The model with the lowest AIC or BIC value is preferred, as it strikes a balance between model fit (captured by the likelihood function) and model complexity (captured by the penalty term). In Bayesian models, the predictive measure more commonly chosen is the DIC or Deviance Information Criteria [13].

1. **AIC** of a given model m is defined as $AIC(m) = -2\ logp(y|\hat{\theta}, m) + 2\ k$ where, k is the number of free parameters estimated during the MLE of the point estimate, $\hat{\theta}$ [1].

2. **BIC** of a given model m is defined as $BIC(m) = -2\ logp(y|\hat{\theta}, m) + k\ log\ n$ where n is the number of data points and k is the number of free parameters estimated during the MLE of the point estimate, $\hat{\theta}$ [31].

3. **DIC** is defined as $DIC = -2\ logp(y|\hat{\theta}_{bayes}) + 2\ p_{DIC}$ where, p_{DIC} is the effective number of parameters defined as, $p_{DIC} = 2(logp(y|\hat{\theta}_{bayes}) - E_{post}(logp(y|\theta)))$. The expectation term in this expression is the average of θ over its posterior distribution [13].

3.3.3 Variable Selection

Variable selection is carried out after the following considerations and inferences,

1. *p-values*: Observing p-values in the full model trained on all the features gives us an idea of the significant features in the model for predicting the auction price. In `R`, the default significance level for the `lm()` & `lmer()` class of models is set to 0.05 i.e. we consider coefficients with a p-value less than 0.05 to be statistically significant. Observing p-values is not enough on its own to be a variable selection method or various reasons. In their paper, Gelman and Loken address the issue of p-value misuse and its implications for variable selection [14]. They argue that traditional significance testing, based on p-values, often leads to inflated numbers of false positives and unreliable results. The paper emphasizes that the use of p-values for variable selection without appropriate correction for multiple comparisons can result in spurious findings and overfitting.

2. *Information Criterion*: To avoid the limitations of p-values, we also observe the information criterion for each trained model. We specifically look at the AIC and BIC values for the non-Bayesian models in the study. For the Bayesian models, we would turn to the DIC criterion. AIC and BIC can be used to select the *best* model among a set of candidate models. The model with the lowest AIC or BIC value is preferred, as it strikes a balance between model fit (captured by the likelihood function) and model complexity (captured by the penalty term).

3. *Stepwise Regression*: Stepwise selection involves performing a series of steps where at each step predictor variable is added or removed to the model according to their statistically significance level, typically using t-statistics for coefficients of these variables. There are forward as well as backward selection rules that decide whether the selection procedure starts with an empty model and adds variables, or starts with a full model and eliminates variables [21]. There is also a combined bi-directional approach where the procedure considers the statistical consequence of dropping a variable previously considered. We make use of the backwards selection procedure in our variable selection process with model significance determined by the AIC values.

3.4 Modelling

The objective of this study is to build a predictive model that can model auction prices against the player performance metrics. As a good practice, we gradually gravitate towards more complex models and start with training classical linear models first.

3.4.1 Linear Models

We start with fitting the classical linear regression models using OLS on the training data. We first train the full model using all the features and using variable selection we identify significant features that explain the auction price.

The next step will be to inspect the residuals and note the prediction errors. We inspect the residuals by plotting them against the fitted values and check for constant variance and

homoscedasticity assumption. We also check the normal Q-Q plot of the residuals to check for the normality assumption. Any significant deviation from the straight line will suggest a departure from normality. Finally, we post the model with the selected variables and also note the prediction accuracy over the test set and the AIC, BIC and log-likelihood values of the final model.

A linear model is not expected to be the best model choice for this study because of the nature of the data involved. Predominantly research on the topic of cricket and IPL related analytics and predictive modelling has stayed within the bounds of linear models. So this approach will give us a fair baseline to compare with the existing literature.

3.4.2 Longitudinal Models

After establishing a baseline prediction, we then move to the main focus of this thesis. We propose to train Mixed Models by specifying fixed and random effects. The first step will be to train a *Linear Mixed Model*. We use the `lme4` package [4] in R which provides the appropriate functions to fit and analyze linear mixed models, generalized linear mixed models and non-linear mixed models.

3.4.2.1 LMMs & GLMMs

The theory behind linear mixed models was already discussed earlier in section 2.3.2 of the thesis. The `lme4` package uses the `lmer` & `glmer` functions to train these models. These functions differ from the `lm` and `glm` in how the fixed and random effects are specified when training the models. The functions `lmer` & `glmer` take their first argument as a formula which for the most part is the same as the formula specification for a `lm` or `glm` function with the difference lying in how the random effects are specified. Each random effect term in the formula is of the form (`expression | group`) that is added in the formula after the fixed effects. While the package allows for construction of various nested or hierarchical models, for this thesis we stick to a random intercept model specified by (`1|g`) or a correlated random intercept and random slope model specified by (`x + (x|g)`)

We follow the same approach as we take while training the baseline models and first train a full model with all the features specified as the fixed effects. We specify a random intercept within the player names as the grouping variable. The resulting model would learn fixed effects coefficients for all the features and also learn a random slope for each player in the training set. It may happen some players that are present in the test set but not in the training set. For these cases, the predictions can be made by allowing new levels in the `predict()` in `R`. Performing variable selection in linear mixed-effects models is more challenging due to the complexity of the model and the presence of both fixed and random effects. We will employ stepwise regression to select variables and constantly monitor the AIC and BIC values of each model along with the generalization error from predicting on test data. The optimal model will be reported along with the fixed and random effects chosen in the process.

The final model will also be checked against the model assumptions by checking the diagnostic plots as in the baseline models. We will inspect the residuals by plotting them against fitted values and inspecting the variance as the fitted values increase. We will also inspect the normal Q-Q plot to check the normality assumption. We can also check the confidence intervals via Wald approximations for the fixed effects. Model comparison can be a little tricky since the model output doesn't include p-values for the fixed effects [4]. The alternative choice is to perform model comparison with the `anova` method [11].

3.4.2.2 MCMC simulations

The last category of modelling that we explore is MCMC simulations of GLMMs. In data exploration analysis in section 3.2.1, we observed that bowler salaries follow a different distribution than that of the batsmen and all-rounders. From the figure 3.5, we can see that both

batsman and all-rounder salaries don't follow a Gaussian distribution. This evidence is a motivation for evaluating the predictive performance of a GLMM which can bring a comparative model that deals with non-Gaussian data.

As we discussed earlier in section 2.3.3, the idea is to use MCMC simulations to overcome the limitations of computing the likelihood for non-Gaussian data for which there is no closed form. We use the default priors for the covariance structures of $\mathbf{B}, \mathbf{G}, \mathbf{R}$ and run experiments with different families to improve upon the results using `lmer()` procedures. Since these models are setup in Bayesian context, model selection criteria will shift to DIC and MAPE values on predictions. Also due to the Bayesian context, we now get a distribution of estimated parameters and as summary statistics we get the posterior mean, upper and lower 95% credible intervals for each effect. While checking the output module we will see an effective sample size and a *pMCMC* value for each fixed effect. *pMCMC* is not a p-value as such and is defined as two times the smaller of the two quantities, MCMC estimates of,

1. the probability that $a < 0$ or,
2. the probability that $a > 0$, where a is the parameter value.

One of the reliable ways to judge significance of each fixed effect is by observing the CIs and selecting the effects whose CIs don't span zero. But this can lead to scenarios where we end up discarding some features whose HPDinterval doesn't span zero. HPDinterval finds the closest points (x and y) for which $F(y) - F(x) = 0.95$ and $F()$ is the empirical cumulative distribution of a.

For random effects, the focus is on estimating variance, so the ideal choice in choosing a random effect is when the variance distribution is not peaking close to zero. To check convergence of the models, we can plot the `model$Sol` for fixed effects and `model$VCV` for random effects to inspect the MC for convergence. While it is possible to use parameterized priors, we stick to using flat priors in our experiments for simplicity.

Another important element of the MCMC convergence diagnostics is the effective sample size (ESS). ESS quantifies the number of independent and uncorrelated samples that carry the same amount of information as the original correlated samples. It provides an estimate of the effective amount of information obtained from the MCMC sampling process. A higher effective sample size means less correlated samples, which generally leads to more reliable estimation of model parameters, narrower credible intervals, and more accurate inference. Conversely, a lower effective sample size indicates that the MCMC chain is more correlated, leading to decreased precision and potentially misleading inference.

4 Results

4.1 Batsmen

4.1.1 Linear Model

We begin by training a classical linear model and train a full model using all the features. We then perform variable selection to identify significant predictors by,

1. observing the p-values of the hypothesis test that the coefficient is equal to 0.
2. performing bi-directional step regression and identify a model with significant features.

As a result we have 3 experimental models which we then compare using the `anova` method for model comparison. The results of model comparison are shown in listing 4.1. From the model comparison we can see that the most significant model is Model3 which was trained after choosing significant features by observing their p-values.

Listing 4.1: lm model comparison

```
> anova(lm1, lm1.step, lm3)
Analysis of Variance Table

Model 1: Price ~ Overseas_player + WK + Star_player + Active_T20I +
    Cum_LY_Bat_I + Cum_LY_HardHitter + Cum_LY_Finisher. +
    Cum_LY_46_scored + Cum_LY_Bat_avg + Cum_LY_50plus_scored +
    Cum_LY_RBW + Cum_IPL_bat_DM_Bat_I + Cum_IPL_bat_DM_HardHitter +
    Cum_IPL_bat_DM_Finisher. + Cum_IPL_bat_DM_46_scored +
    Cum_IPL_bat_DM_Bat_avg + Cum_IPL_bat_DM_50plus_scored +
    Cum_IPL_bat_DM_RBW
Model 2: Price ~ Overseas_player + Star_player + Active_T20I +
    Cum_LY_HardHitter + Cum_LY_46_scored + Cum_LY_Bat_avg +
    Cum_LY_RBW + Cum_IPL_bat_DM_Bat_I + Cum_IPL_bat_DM_Finisher. +
    Cum_IPL_bat_DM_50plus_scored
Model 3: Price ~ Overseas_player + Star_player + Active_T20I +
    Cum_IPL_bat_DM_Bat_I + Cum_IPL_bat_DM_Finisher.
  Res.Df     RSS Df Sum of Sq       F     Pr(>F)
1    324 74.559
2    332 75.292 -8    -0.7332 0.3983     0.9212
3    337 82.445 -5    -7.1531 6.2169 1.601e-05 ***
```

Therefore, the most significant model can be written in the equation form as follows:

$$\begin{aligned}\text{Price} = {} & 0.32 - 0.45(\text{Overseas\_player}_1) \\ & + 0.66(\text{Star\_player}_1) + 0.3(\text{Active\_T20I}_1) \\ & + 0.29(\text{Cum\_IPL\_bat\_DM\_Bat\_I}) + 0.07(\text{Cum\_IPL\_bat\_DM\_Finisher.})\end{aligned}$$

Predictor	**Description**
Overseas_player	If the player's nationality is Indian or other
Star_player	If the player has star status
Active_T20I	If the player has played any matches for his national team in the last year
Cum_IPL_bat_DM_Bat_I	Total number of innings played in player's IPL career
Cum_IPL_bat_DM_Finisher.	Number of times a player has remained not-out at the end of the batting innings in his IPL career

Table 4.1: Significant baseline model predictors for Batsmen

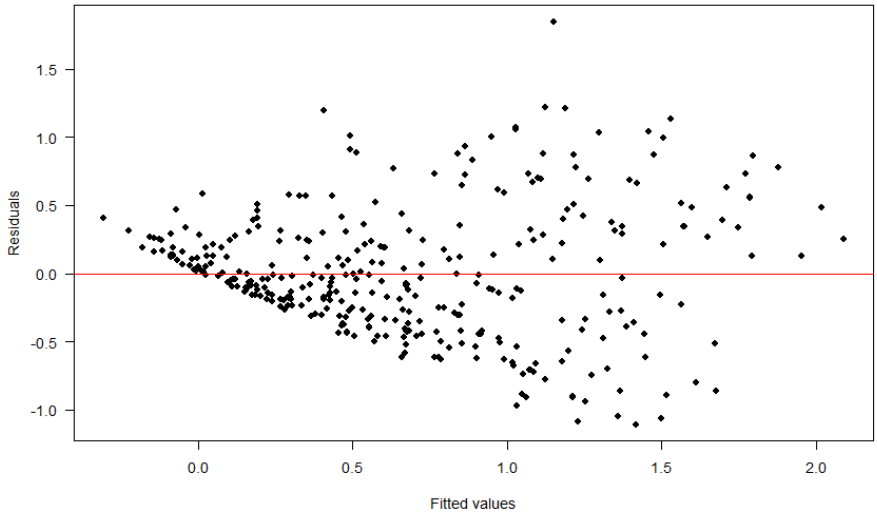
With this method, we identified a set of significant predictors that are detailed in table 4.1. Inspection tells us that a player's star status has the biggest impact to their predicted price. The other positive impact on auction price comes from the player being an active international team player. The cumulative experience of the batsman and a proven ability to finish games for their team also has a positive impact on the auction price. On the other hand, if a player does not belong to the domestic pool of Indian cricketers, then it has a negative effect on the auction price. This can be rationalized together with the other significant features that a batsman is valued on two major traits i.e., their proven consistency in the IPL and their ability to score reliably, not just scoring some runs and getting out.

The chosen baseline model was able to predict with a RMSE of 0.52 and a MAPE of 2.22 with the model AIC value of 498.41 as summarized in table 4.2. The other models scored lower IC values, but there is a very stark difference in the prediction accuracy metric MAPE for the test data. Hence, we choose to select this model as the baseline model.

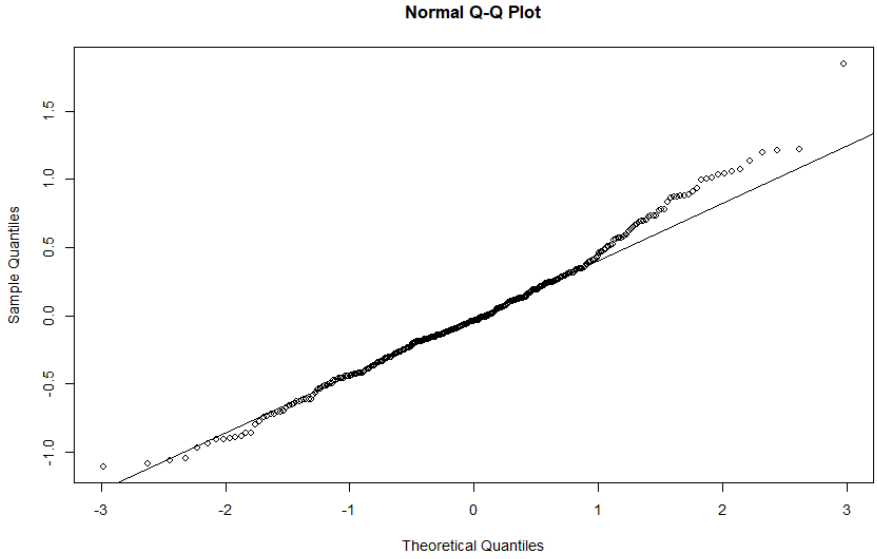
Model	**RMSE**	**MAPE**	**AIC**	**BIC**	**LL**
Full Model	0.5211702	2.881494	489.9252	566.6798	-224.9626
LM Var selection using step regression	0.5258855	2.901778	477.2817	523.3344	-226.6408
LM Var selection using p-values	0.5160106	2.224992	498.4121	525.2762	-242.2061

Table 4.2: Model Comparison: Linear Models - Batsmen

The residuals plot shows clear heteroscedasticity as we see the residuals diverge in a cone shape as the prediction keeps increasing as shown in plot 4.1(a). This is expected since we observed that some star players have attracted very high salaries each year while the median salary remains at a much lower level. The normal Q-Q plot on inspection also suggests that the normality condition is getting violated suggesting we might need to look at distributions other than Gaussian in our model choices.



(a) Restricted LM - Residuals vs Fitted plot showing heteroscedasticity



(b) Restricted LM - QQ plot violating normality

Figure 4.1: Restricted LM - Residuals

4.1.2 Linear Mixed Model

We first train a LMM by taking all features as fixed effects and take a random intercept with the grouping variable as the player name. The idea for choosing this design is that each player should be modelled with their own random intercept while keeping their performance measures as fixed. Like in the baseline LM case, we use stepwise regression on the full model to identify significant fixed effects. The resulting truncated model with the significant features is then experimented further with random effect combinations. The final model is chosen by model comparison using `anova` method. The model comparison is shown in listing 4.2.

Listing 4.2: LMM model comparison - Batsmen

```
> anova(lmm1.step, lmm1, lmm1.step.re, lmm1.step.re1, lmm1.step.re2, lmm1.step.re3)
Data: train
Models:
lmm1.step:
        Price ~ Season_num + Overseas_player + Star_player + Active_T20I +
                Cum_LY_HardHitter + Cum_LY_46_scored + Cum_LY_RBW +
                Cum_IPL_bat_DM_Bat_I + Cum_IPL_bat_DM_50plus_scored +
                (1 | Name2)
lmm1.step.re1:
        Price ~ Season_num + Overseas_player + Star_player + Active_T20I +
                Cum_LY_HardHitter + Cum_LY_46_scored + Cum_LY_RBW +
                Cum_IPL_bat_DM_Bat_I + Cum_IPL_bat_DM_50plus_scored +
                (Cum_IPL_bat_DM_Bat_I | Name2)
lmm1.step.re:
        Price ~ Season_num + Overseas_player + Star_player + Active_T20I +
                Cum_LY_HardHitter + Cum_LY_46_scored + Cum_LY_RBW +
                Cum_IPL_bat_DM_Bat_I + Cum_IPL_bat_DM_50plus_scored +
                (1 + Star_player + Cum_IPL_bat_DM_Bat_I | Name2)
lmm1.step.re3:
        Price ~ Season_num + Overseas_player + Star_player + Active_T20I +
                Cum_LY_HardHitter + Cum_LY_46_scored + Cum_LY_RBW +
                Cum_IPL_bat_DM_50plus_scored +
                (1 | Name2) + (Star_player + Cum_IPL_bat_DM_Bat_I | Name2)
lmm1.step.re2:
        Price ~ Season_num + Overseas_player + Star_player + Active_T20I +
                Cum_LY_HardHitter + Cum_LY_46_scored + Cum_LY_RBW +
                Cum_IPL_bat_DM_Bat_I + Cum_IPL_bat_DM_50plus_scored +
                (1 | Name2) + (Star_player + Cum_IPL_bat_DM_Bat_I | Name2)
lmm1:
        Price ~ Season_num + WK + Overseas_player + Star_player + Active_T20I +
                Cum_LY_Bat_I + Cum_LY_HardHitter + Cum_LY_Finisher. +
                Cum_LY_46_scored + Cum_LY_Bat_avg + Cum_LY_50plus_scored +
                Cum_LY_RBW + Cum_IPL_bat_DM_Bat_I + Cum_IPL_bat_DM_HardHitter +
                Cum_IPL_bat_DM_Finisher. + Cum_IPL_bat_DM_46_scored +
                Cum_IPL_bat_DM_Bat_avg + Cum_IPL_bat_DM_50plus_scored +
                Cum_IPL_bat_DM_RBW + (1 | Name2)

               npar    AIC    BIC  logLik deviance   Chisq Df Pr(>Chisq)
lmm1.step        12 420.44 466.49 -198.22   396.44
lmm1.step.re1    14 399.37 453.09 -185.68   371.37 25.0721  2  3.595e-06 ***
lmm1.step.re     17 401.44 466.68 -183.72   367.44  3.9266  3     0.26950
lmm1.step.re3    17 398.06 463.30 -182.03   364.06  3.3840  0
lmm1.step.re2    18 396.77 465.85 -180.39   360.77  3.2832  1     0.06999 .
lmm1             22 431.31 515.74 -193.65   387.31  0.0000  4     1.00000
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1        1
```

From the model comparison we can see that the most significant model seems to be model `lmm1.step.re1`. We next check the generalization accuracy and the IC values before choosing the best model. The accuracy and IC values are shown in table 4.3. From table 4.3 we can see that the model with the best accuracy and favourable scores is actually model `lmm1.step.re3`. We see that this model posts a significantly better accuracy and the sec-

ond lowest IC scores. Hence, we choose this model as the best LMM model and proceed to derive statistical inference.

Model: Random Effect	RMSE	MAPE	AIC	BIC	LL
lmm1: RI Name	0.4174570	2.133655	431.3061	515.7362	-193.6531
lmm1.step: RI Name	0.4203608	2.206477	420.4386	466.4914	-198.2193
lmm1.step.re: RI Name RS Star_player + Cum_IPL_bat_DM_Bat_I	0.3696429	2.029809	401.4400	466.6814	-183.7200
lmm.step.re1: RI Name RS Cum_IPL_bat_DM_Bat_I	0.3718562	2.089433	399.3666	453.0948	-185.6833
lmm1.step.re2: RI Name RS Star_player + Cum_IPL_bat_DM_Bat_I	0.3703225	1.976803	396.7727	465.8519	-180.3864
lmm1.step.re3: RI Name RS Star_player + Cum_IPL_bat_DM_Bat_I	0.3773605	1.910150	398.0559	463.2974	-182.0280

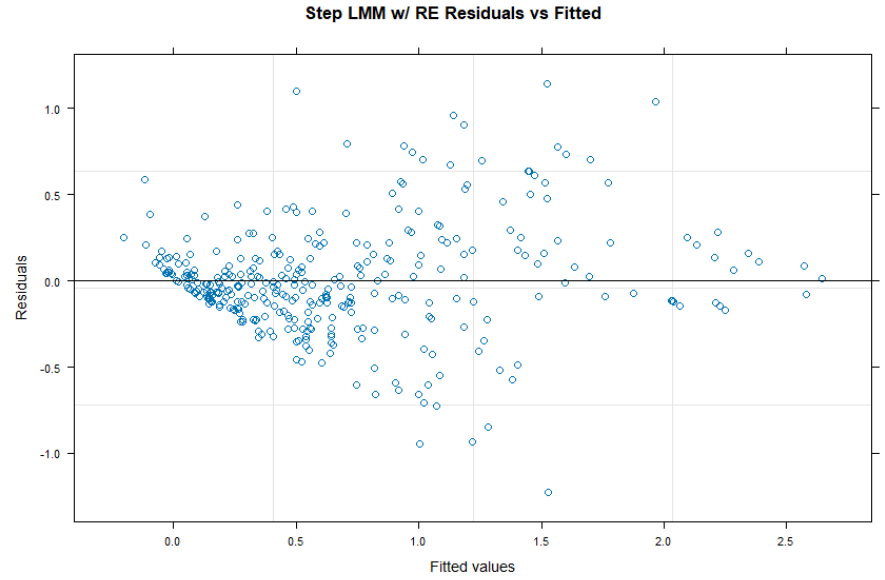
Table 4.3: Model Comparison: Linear Mixed Models - Batsmen

Figure 4.2 shows the residuals plotted for the chosen model. The residuals plot shows a very similar heteroscedasticity issue that we observed in the baseline LM model as well. The Q-Q normality plots also show deviation from normality for the outliers. For the chosen model, the summary is described in listing 4.3. From the *fixed effects* we can infer that again the most features are the star status and nationality of a player. As we saw in the baseline model, star status has a positive impact on player valuation while an overseas nationality has a negative effect. Within batting ability we see that the most significant feature is the ability of a batsman to score a large number of runs in a single match. The features relating to the frequency of scoring boundaries and accumulating runs in boundaries is also a favourable skill. We were able to find the best accuracy by introducing a correlated random intercept and slope for the attributes star player & cumulative innings batted in IPL, within the grouping variable player name. This implies that a player's total IPL experience along with their star status varies significantly between different players.

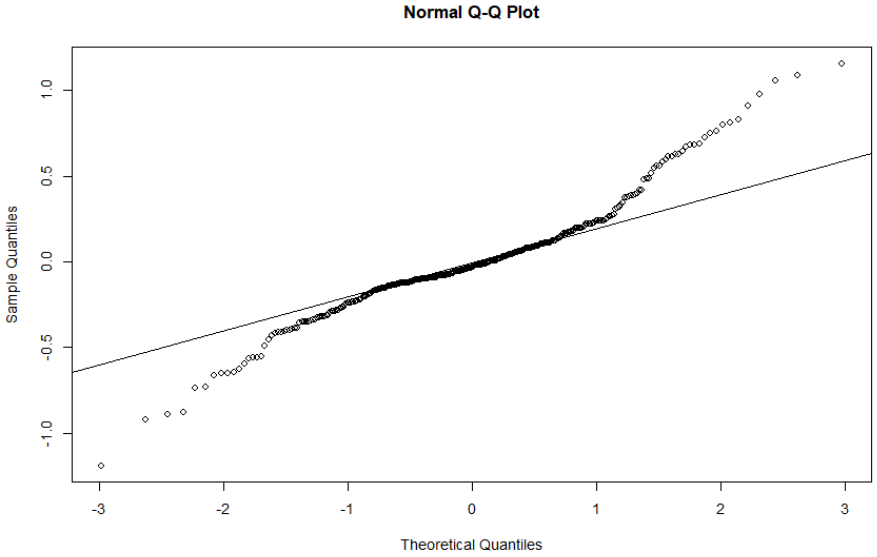
In the LMM, we see that the auction price is still significantly dependent on a player's star status and their nationality status as we saw in the baseline model. Additionally, we see that season is also a significant predictor. This makes sense, because some players that get retained by their respective franchise get pay raises over time. Another significant predictor identified by the LMM is the career aggregate of 50 plus scores for a batsman. This is another indicator that teams look for batsmen that have proven track record of making significant contributions to the team. Another point to note is that as a recency effect, the only significant factor seems to be the frequency of boundaries that the batsmen is hitting while batting. This model offers a significant improvement over the baseline model in terms of prediction accuracy.

Listing 4.3: LMM Final Model Summary - Batsmen

```
> summary(lmm1.step.re3)
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: Price ~ Season_num + Overseas_player + Star_player + Active_T20I +
    Cum_LY_HardHitter + Cum_LY_46_scored + Cum_LY_RBW + Cum_IPL_bat_DM_50plus_scored +
    (1 | Name2) + (Star_player + Cum_IPL_bat_DM_Bat_I | Name2)
   Data: train

     AIC      BIC   logLik deviance df.resid
   398.1    463.3   -182.0    364.1       326

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.5357 -0.4068 -0.0701  0.3795  3.2958

Random effects:
 Groups    Name                 Variance Std.Dev. Corr
 Name2     (Intercept)          0.00000  0.0000
 Name2.1   (Intercept)          0.02490  0.1578
           Star_player1         0.09984  0.3160   -0.05
           Cum_IPL_bat_DM_Bat_I 0.05618  0.2370    1.00 -0.05
 Residual                       0.12060  0.3473
Number of obs: 343, groups:  Name2, 106

Fixed effects:
                             Estimate Std. Error       df t value Pr(>|t|)
(Intercept)                   0.86530    0.12507 263.14399   6.918 3.47e-11 ***
Season_num                   -0.05060    0.01217 257.70625  -4.159 4.36e-05 ***
Overseas_player1             -0.60885    0.10924  54.26189  -5.573 8.06e-07 ***
Star_player1                  0.64553    0.10268  78.63338   6.287 1.69e-08 ***
Active_T20I1                  0.16771    0.06054 321.50651   2.770  0.00593 **
Cum_LY_HardHitter            -0.28666    0.12301 259.21427  -2.330  0.02056 *
Cum_LY_46_scored              0.43492    0.17365 214.14314   2.505  0.01301 *
Cum_LY_RBW                   -0.17361    0.08295 194.89974  -2.093  0.03766 *
Cum_IPL_bat_DM_50plus_scored  0.22501    0.03421 256.62549   6.577 2.68e-10 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1        1

Correlation of Fixed Effects:
            (Intr) Ssn_nm Ovrs_1 Str_p1 A_T20I C_LY_H C_LY_4 C_LY_R
Season_num  -0.919
Ovrss_plyr1  0.029 -0.017
Star_playr1 -0.346  0.180 -0.695
Activ_T20I1  0.120 -0.189 -0.137 -0.115
Cm_LY_HrdHt -0.097  0.033  0.111 -0.003  0.094
Cm_LY_46_sc  0.142 -0.074 -0.106 -0.014 -0.095 -0.978
Cum_LY_RBW  -0.115  0.062  0.095  0.005  0.072  0.905 -0.957
C_IPL__DM_5  0.198  0.010  0.086 -0.246 -0.164 -0.095  0.068 -0.053
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```



(a) LMM - Residuals vs Fitted plot



(b) LMM - QQ plot

Figure 4.2: LMM Residual Plots -Batsmen

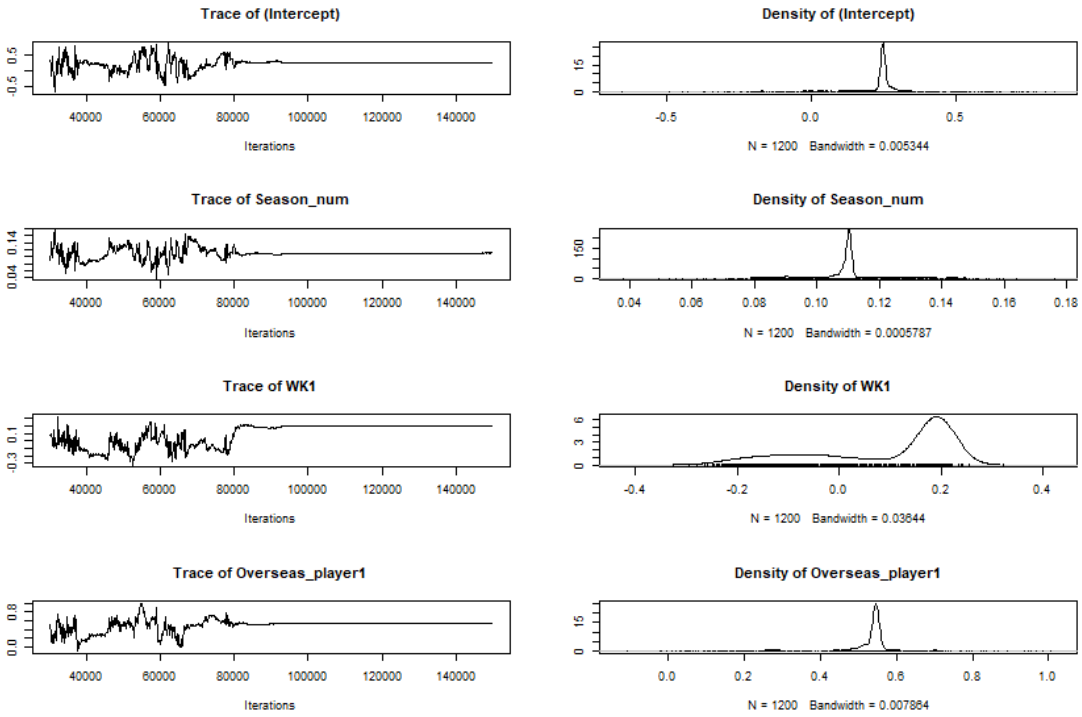
4.1.3 MCMC simulations of GLMM

Until now we have employed two linear classes of models under the Gaussian assumption. As the diagnostics for the models have shown, the normality assumption of the model gets violated repeatedly. So now we go with MCMC simulations of a Generalized Linear Mixed Model. We use the MCMC simulations to model a non-linear family of functions. After some experiments, it was observed that the best results were being delivered with the exponential family of distribution as the trait distribution setting. We choose 150000 iterations with a burn in of 30000, to generate markov chains and check for convergence. We inspect the density plots to check if the posterior means of the coefficients lies within 95% credible intervals and that the CIs do not overlap with zero. We also observe the HPD regions for each effect before deciding on variable selection.

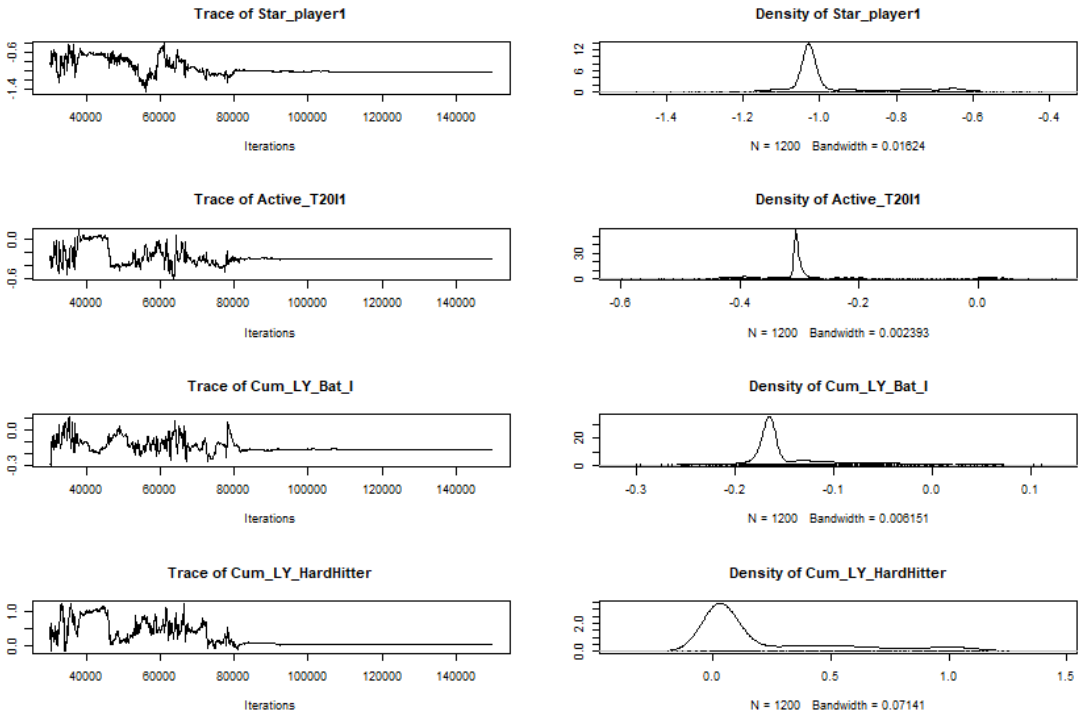
First we trained a full model using all the features for which the output is shown in listing 4.4 and observed the trace plots as shown in fig 4.3

Listing 4.4: MCMC Full Model Summary - Batsmen

```
> summary(MCMCglmm1)

 Iterations = 30001:149901
 Thinning interval  = 100
 Sample size  = 1200

 DIC: 208.6732

 G-structure:  ~Name2

      post.mean  l-95% CI u-95% CI eff.samp
Name2  0.005608 1.591e-16   0.0388     15.22

 R-structure:  ~units

      post.mean  l-95% CI u-95% CI eff.samp
units  0.001548 3.526e-07 0.009053     28.03

 Location effects: Price ~ Season_num + WK + Overseas_player + Star_player + Active_T20I +
Cum_LY_Bat_I + Cum_LY_HardHitter + Cum_LY_Finisher. + Cum_LY_46_scored +
Cum_LY_Bat_avg + Cum_LY_50plus_scored + Cum_LY_RBW + Cum_IPL_bat_DM_Bat_I +
Cum_IPL_bat_DM_HardHitter + Cum_IPL_bat_DM_Finisher. + Cum_IPL_bat_DM_46_scored +
Cum_IPL_bat_DM_Bat_avg + Cum_IPL_bat_DM_50plus_scored + Cum_IPL_bat_DM_RBW

                              post.mean   l-95% CI   u-95% CI eff.samp   pMCMC
(Intercept)                   0.2246390 -0.2077801  0.5537006   61.563 0.19000
Season_num                    0.1099426  0.0790511  0.1440200   37.095 < 8e-04 ***
WK1                           0.0872991 -0.1998745  0.2059781    4.236 0.60000
Overseas_player1              0.5023220  0.1352532  0.7192396   15.341 0.00667 **
Star_player1                 -0.9655935 -1.1563679 -0.5747453   11.728 < 8e-04 ***
Active_T20I1                 -0.2828862 -0.4327929  0.0224347   21.250 0.11333
Cum_LY_Bat_I                 -0.1438556 -0.2032750  0.0021546   23.070 0.05333 .
Cum_LY_HardHitter             0.2319472 -0.0008615  1.0407697    4.647 0.03500 *
Cum_LY_Finisher.              0.1524301  0.0397586  0.2013187   17.024 0.00167 **
Cum_LY_46_scored             -0.5283653 -1.5813872 -0.2123143    6.637 0.00667 **
Cum_LY_Bat_avg                0.0284888 -0.1466157  0.1063862    9.361 0.51000
Cum_LY_50plus_scored         -0.0362358 -0.1667514  0.0427144   22.540 0.20333
Cum_LY_RBW                    0.2819412  0.1092065  0.8316023   10.058 0.01333 *
Cum_IPL_bat_DM_Bat_I         -0.3205925 -0.5564451 -0.1731222   13.921 0.00167 **
Cum_IPL_bat_DM_HardHitter    -0.0494949 -0.6194691  0.6993807   40.360 0.45000
Cum_IPL_bat_DM_Finisher.     -0.0672014 -0.2520804  0.0033301   12.555 0.05833 .
Cum_IPL_bat_DM_46_scored      0.1986536 -1.0012664  1.0511950   37.391 0.41333
Cum_IPL_bat_DM_Bat_avg       -0.1665455 -0.3804852  0.3600055   23.925 0.37167
Cum_IPL_bat_DM_50plus_scored -0.1589020 -0.3975101  0.0608034   50.654 0.12667
Cum_IPL_bat_DM_RBW           -0.1883521 -0.5676879  0.3283273   41.400 0.34667
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1  1
```

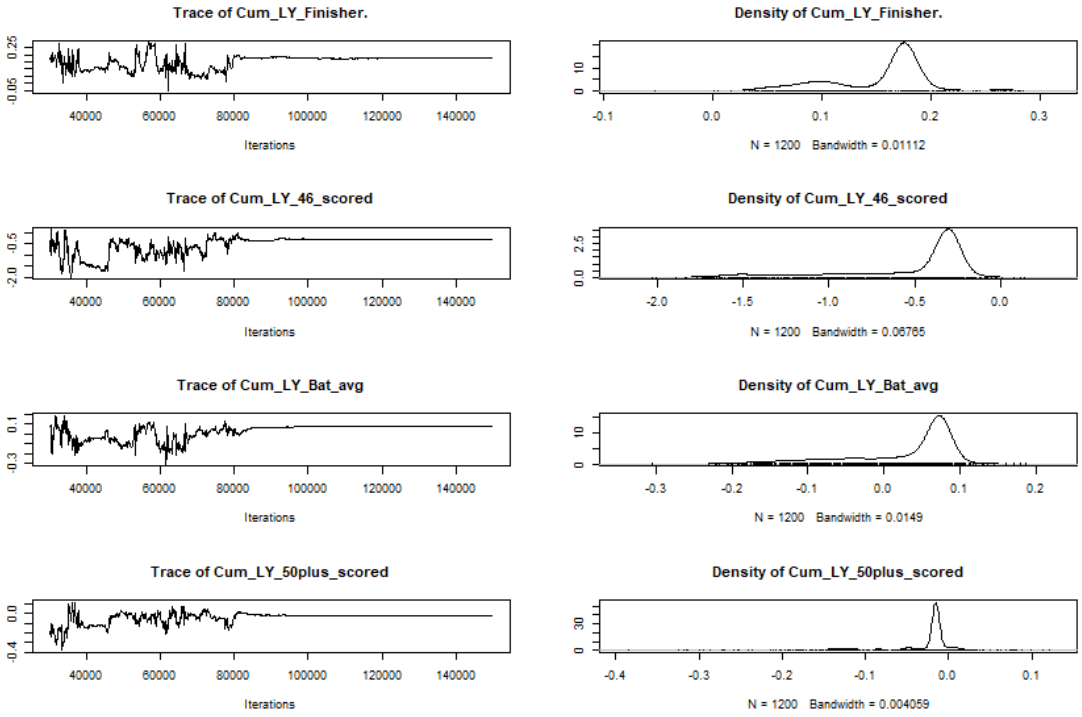


(a)

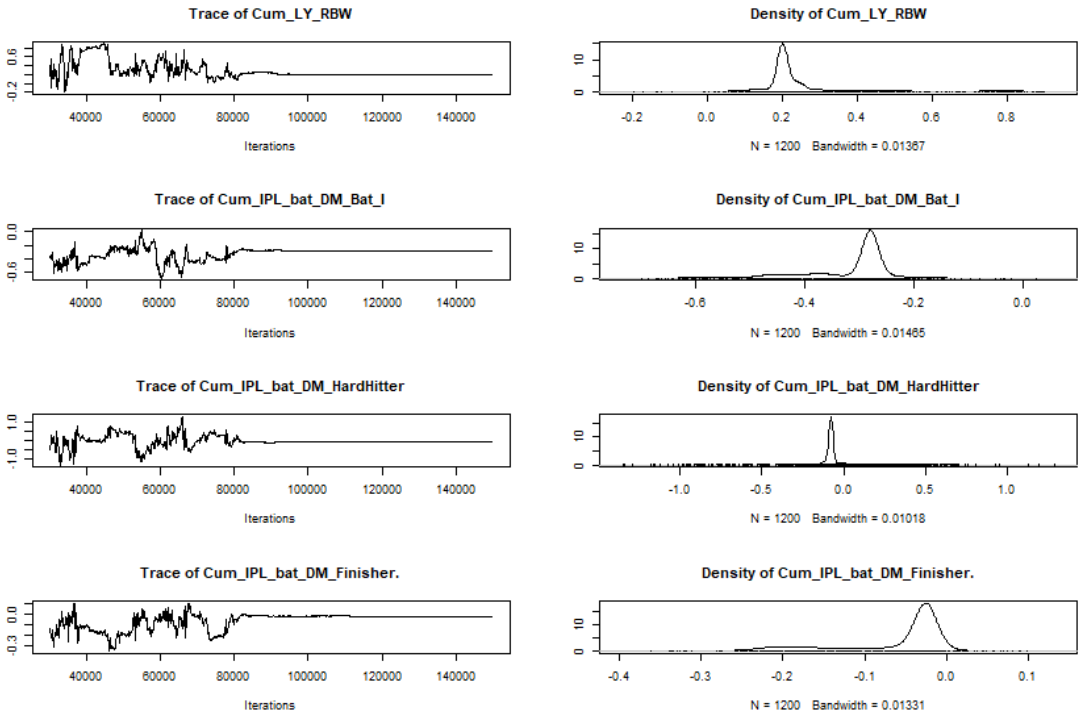


(b)

Figure 4.3: Trace plots and density estimates of posterior means - Batsmen 1/3

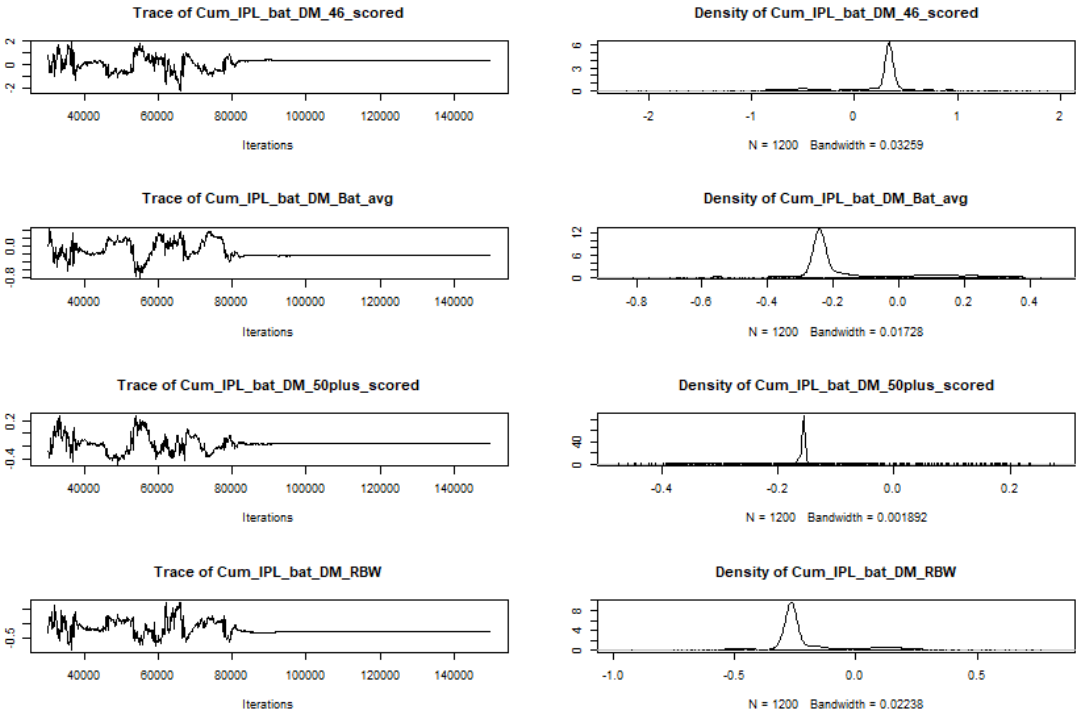


(c)

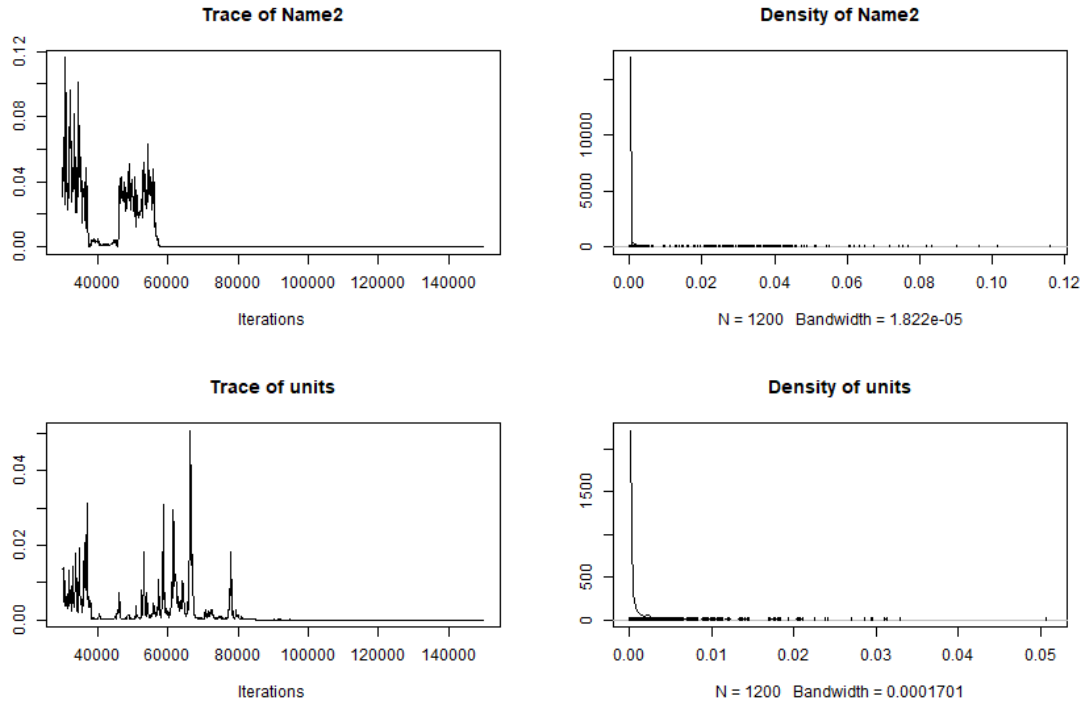


(d)

Figure 4.3: Trace plots and density estimates of posterior means - Batsmen 2/3



(e)



(f)

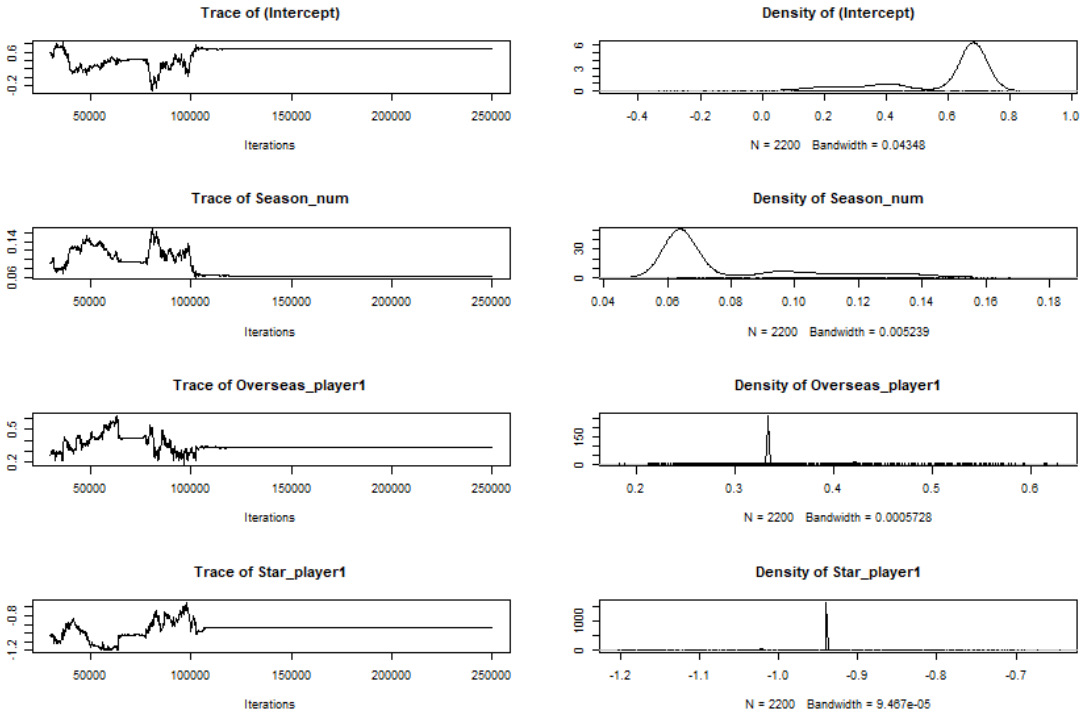
Figure 4.3: Trace plots and density estimates of posterior means - Batsmen 3/3

Listing 4.5: MCMC Final Model Summary - Batsmen

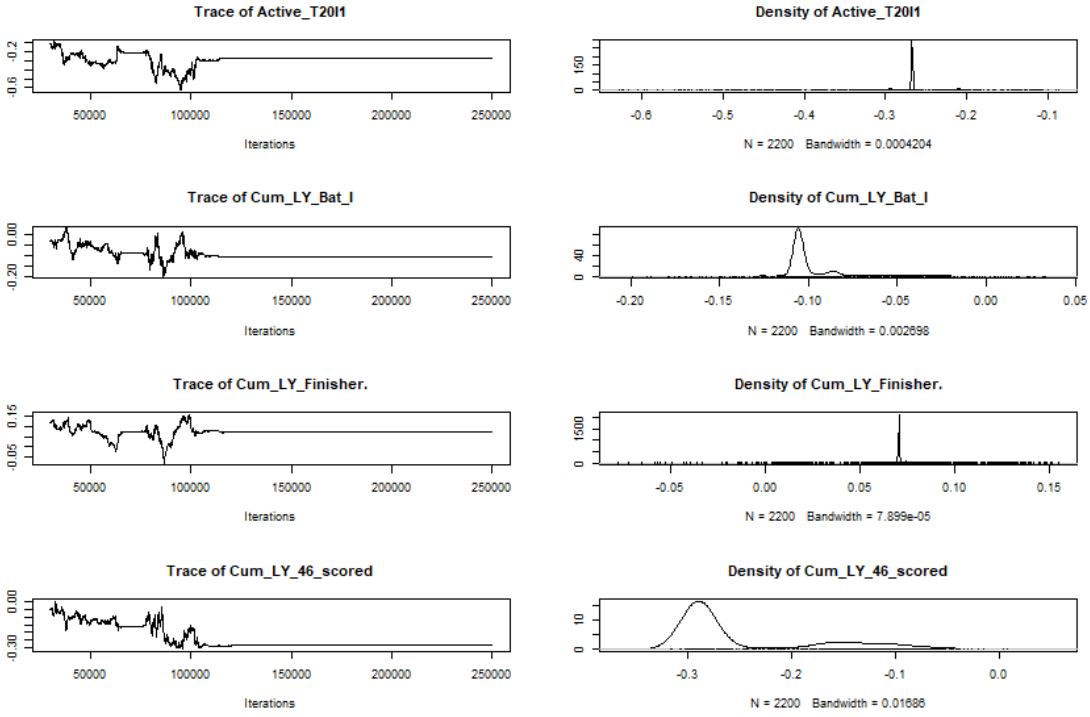
```
> summary(MCMCglmm2)

 Iterations = 30001:249901
 Thinning interval  = 100
 Sample size  = 2200

 DIC: 198.9353

 G-structure:  ~Name2

       post.mean  l-95% CI  u-95% CI eff.samp
Name2  5.89e-06 1.344e-16 1.799e-06     32.2

 R-structure:  ~units

       post.mean  l-95% CI u-95% CI eff.samp
units  0.000187 1.256e-09 0.001128    30.46

 Location effects: Price ~ Season_num + Overseas_player + Star_player + Active_T20I +
Cum_LY_Bat_I + Cum_LY_Finisher. + Cum_LY_46_scored +
Cum_LY_RBW + Cum_IPL_bat_DM_Bat_I + Cum_IPL_bat_DM_46_scored +
Cum_IPL_bat_DM_Bat_avg + Cum_IPL_bat_DM_50plus_scored + Cum_IPL_bat_DM_RBW

                           post.mean l-95% CI u-95% CI eff.samp   pMCMC
(Intercept)                  0.56962  0.14119  0.72934    6.284 0.03909 *
Season_num                   0.07943  0.06362  0.13505    4.525 < 5e-04 ***
Overseas_player1             0.34917  0.21544  0.46069   13.603 < 5e-04 ***
Star_player1                -0.95410 -1.19799 -0.82470   10.800 < 5e-04 ***
Active_T20I1                -0.28106 -0.50141 -0.18252   14.089 < 5e-04 ***
Cum_LY_Bat_I                -0.09428 -0.11988 -0.01742   30.634 0.02455 *
Cum_LY_Finisher.             0.07110  0.01791  0.13193   32.365 0.03273 *
Cum_LY_46_scored            -0.24394 -0.29749 -0.08813    4.337 0.00182 **
Cum_LY_RBW                   0.14670 -0.11677  0.20723    2.675 0.29273
Cum_IPL_bat_DM_Bat_I        -0.40230 -0.46776 -0.37264   17.360 < 5e-04 ***
Cum_IPL_bat_DM_46_scored     0.18157 -0.18227  0.29833    1.847 0.37273
Cum_IPL_bat_DM_Bat_avg      -0.06563 -0.27724  0.02455   10.510 0.19091
Cum_IPL_bat_DM_50plus_scored -0.06339 -0.12500  0.06325   18.733 0.13636
Cum_IPL_bat_DM_RBW           -0.25839 -0.38226  0.11748    1.289 0.38818
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1  1
```

The summary of the final model after variable selection is given in listing 4.5 along with the trace plots and density charts in figure 4.4. The significant predictors that stand out are the season number, Star status, nationality and T20I playing status. From a batting ability perspective, aggregate batting innings in the player's career i.e. their total experience in batting in the league is significant. This is followed by boundaries scored in the last year & their ability to remain not out. The effect of proportion of runs scored in boundaries is also deemed significant by the model which doesn't change the overall profile of batsmen that are being valued by the teams i.e. Consistent boundary scorers with a proven track record in the IPL & a premium on the star status.

The prediction performance on the test data of the MCMC GLMM was better with a RMSE of 0.50 and a MAPE of 1.38 which is a significant improvement. The random effect that gives the best generalization is the grouping variable Player Name. The MCMC model performance is shown in the table 4.4.

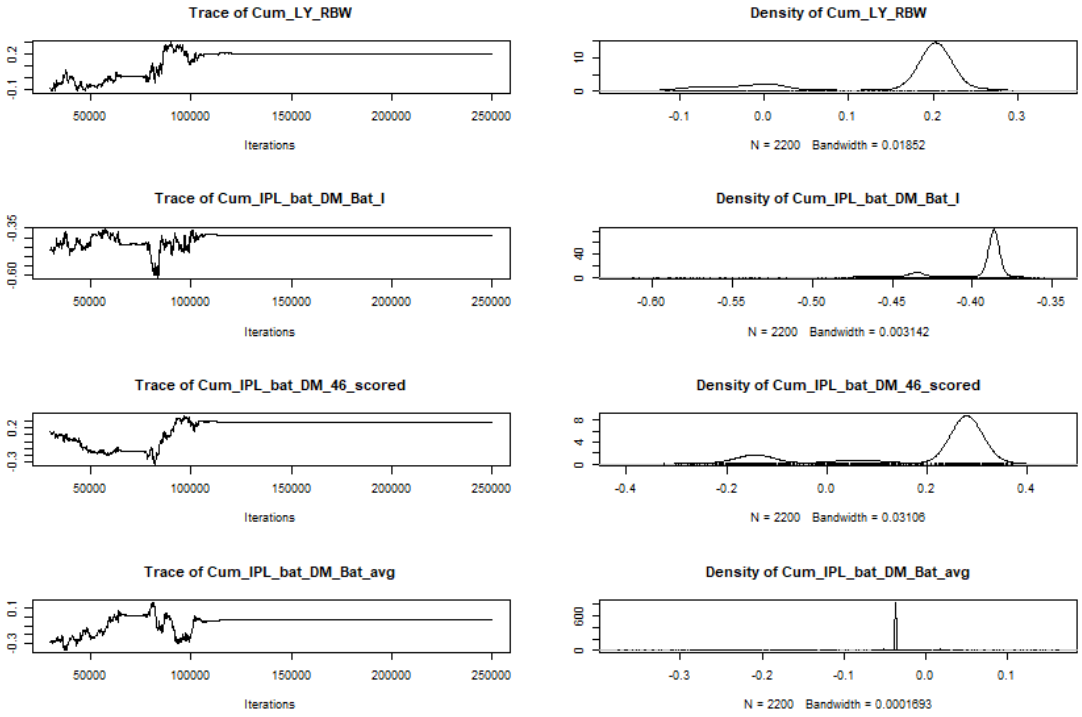


(a)

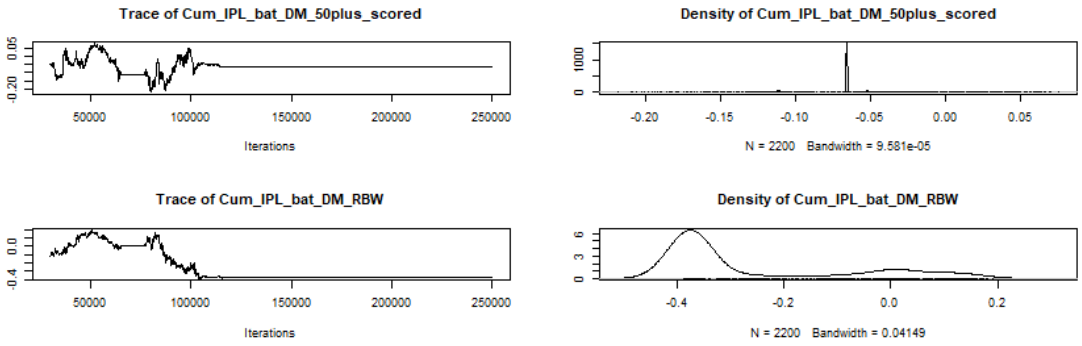


(b)

Figure 4.4: Trace plots and density estimates of posterior means - Batsmen 1/3

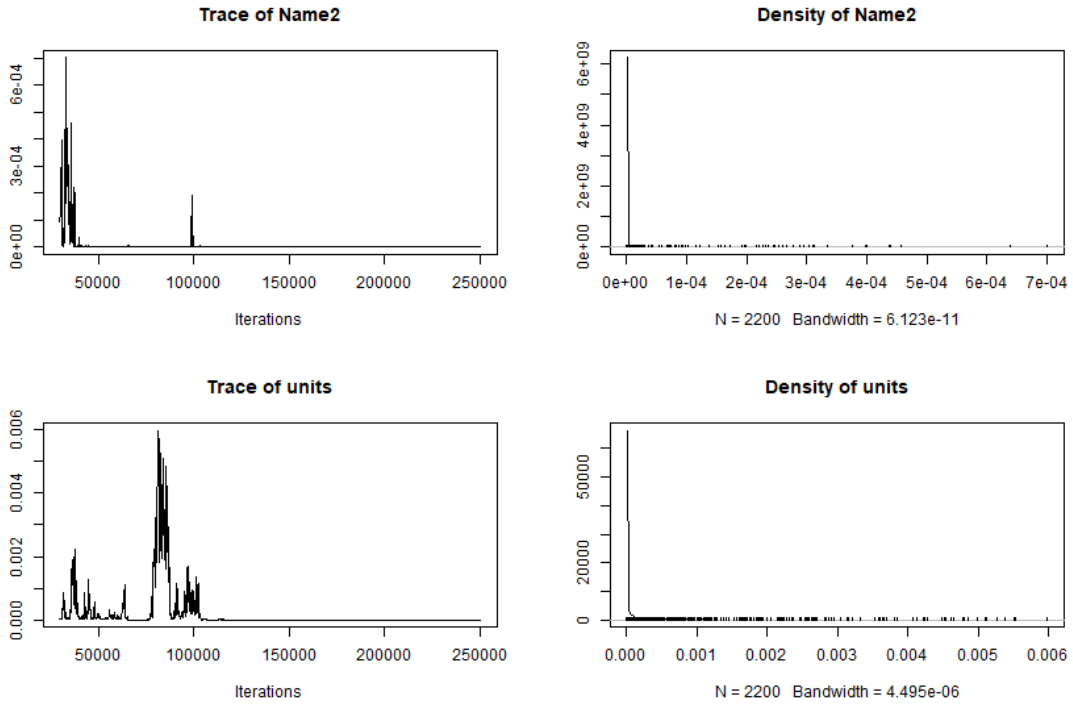


(c)



(d)

Figure 4.4: Trace plots and density estimates of posterior means - Batsmen 2/3



(e)

Figure 4.4: Trace plots and density estimates of posterior means - Batsmen 3/3

Model	**RMSE**	**MAPE**	**DIC**
MCMCglmm1: Full Model	0.5076035	1.278538	208.6732
MCMCglmm1: Variable Selection Model	0.5040844	1.375288	198.9353

Table 4.4: Model Comparison: MCMC GLMM Model Comparison - Batsmen

Model	RMSE	MAPE
Full Model	0.5211702	2.881494
LM Var selection using step regression	0.5258855	2.901778
LM Var selection using p-values	0.5160106	2.224992
Full LMM-Bat: RI Name	0.4174570	2.133655
Step LMM-Bat: RI Name	0.4203608	2.206477
Step LMM-Bat: RI Name RS Star_player + Cum_IPL_bat_DM_Bat_I	0.3696429	2.029809
Step LMM-Bat: RI Name RS Cum_IPL_bat_DM_Bat_I	0.3718562	2.089433
Step LMM-Bat: True RI Name RS Star_player + Cum_IPL_bat_DM_Bat_I	0.3703225	1.976803
Step2 LMM-Bat: True RI Name RS Star_player + Cum_IPL_bat_DM_Bat_I	0.3773605	1.910150
MCMCglmm1: Full Model	0.5076035	1.278538
MCMCglmm1: w/ variable selection	0.5040844	1.375288

Table 4.5: Model Comparison: LM vs LMM vs MCMC GLMM - Batsmen

4.1.4 Summary of results - Batsmen

To summarize we look at the generalization errors from the three model classes that we have used i.e. baseline classical linear model, linear mixed model and the MCMC simulations of a GLMM. For each model class we have experimented with different configurations and choices of fixed and random effects. We now look at all the models together summarized in table 4.5. We see that the best predictive model is the MCMC GLMM which gives a significantly better generalization on the test data compared to the other model classes.

4.2 Bowlers

4.2.1 Linear Model

We now share the results for bowlers by following the same protocol as we did in the case of batsmen in section 4.1. We begin by training a linear model using all the predictors and identify significant features using the same method as earlier. The model comparison of the linear models is shown in ANOVA Table listing 4.6. Of the three experiments that we ran for the baseline model, none of the models were significant so we choose the baseline model by observing on the generalization error, which is model3, which was the model trained using stepwise regression. The resulting model is as follows:

$$\begin{aligned}\text{Price} = {} & 0.29 - 0.11(\text{Overseas\_player}_1) \\ & + 0.17(\text{Star\_player}_1) + 0.12(\text{Active\_T20I}_1) \\ & + 0.08(\text{Cum\_LY\_Bowl\_I}) + 0.14(\text{Cum\_IPL\_bowl\_DM\_Bowl\_I}) \\ & - 0.04(\text{Cum\_IPL\_bowl\_DM\_Bowl\_Econ}) - 0.03(\text{Cum\_IPL\_bowl\_DM\_Bowl.\_SR}) \\ & + 0.05(\text{Cum\_IPL\_bowl\_DM\_Dots.})\end{aligned}$$

Listing 4.6: LM Model Comparison (bowlers)

```
Analysis of Variance Table

Model 1: Price ~ Overseas_player + Star_player + Active_T20I + Cum_LY_Bowl_I +
    Cum_LY_Bowl_Econ + Cum_LY_Bowl_Avg + Cum_LY_Bowl._SR + Cum_LY_BW.Taker +
    Cum_LY_bowl_46_conc + Cum_LY_bowl_Dots. + Cum_IPL_bowl_DM_Bowl_I +
    Cum_IPL_bowl_DM_Bowl_Econ + Cum_IPL_bowl_DM_Bowl_Avg + Cum_IPL_bowl_DM_Bowl._SR +
    Cum_IPL_bowl_DM_BW.Taker + Cum_IPL_bowl_DM_46_conc + Cum_IPL_bowl_DM_Dots.
Model 2: Price ~ Overseas_player + Star_player + Active_T20I + Cum_LY_Bowl_I +
    Cum_IPL_bowl_DM_Bowl_I
Model 3: Price ~ Overseas_player + Star_player + Active_T20I + Cum_LY_Bowl_I +
    Cum_IPL_bowl_DM_Bowl_I + Cum_IPL_bowl_DM_Bowl_Econ + Cum_IPL_bowl_DM_Bowl._SR +
    Cum_IPL_bowl_DM_Dots.
  Res.Df     RSS  Df Sum of Sq       F Pr(>F)
1    350 30.855
2    362 31.540 -12  -0.68478  0.6473 0.8012
3    359 31.020   3   0.52056  1.9683 0.1184
```

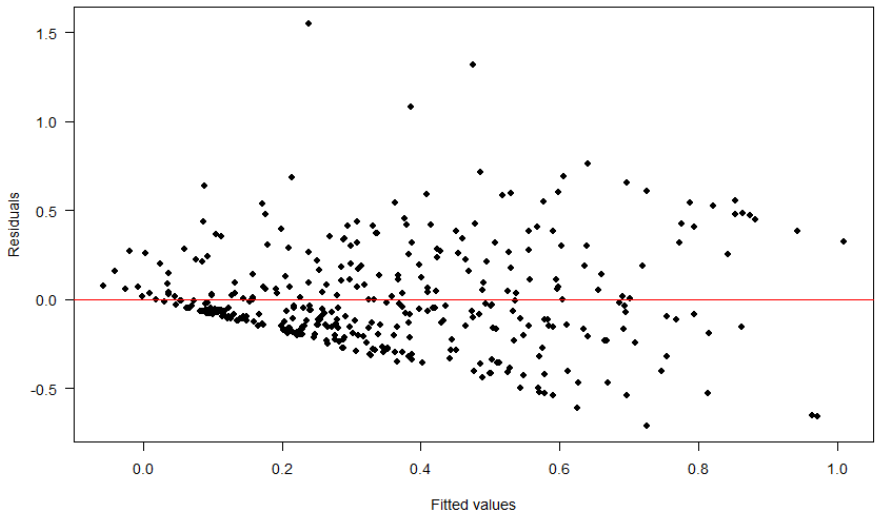
The baseline model suggests that the player's current status as a star, their nationality and their recognition as a national representative is significant with a negative effect due to overseas status, as was the case for batsmen. The significant playing metrics are the player's frequency of playing in the last year as well as in their IPL career. This points to bowlers being prized for their overall experience and their game readiness. This also means that the teams look at bowlers as work horses. They want players who don't break down due to injuries and can play reliably over a longer stretch of games. This is a rational strategy as bowlers who are playing more T20s can be expected to be earning their place in the team by performing consistently and with in the expectations of the team. T20s are a batsman's format and bowlers are often an afterthought. The other important bowling abilities deemed significant are their ability to stop the opposition from scoring too much and taking wicket regularly. The prediction baseline comparison is shown in table 4.6 where we can see that the chosen baseline model has lower IC score than than the full model which supports the choice of baseline model.

We also look at residuals plot for the checking the homoscedasticity and normality condition in figure 4.5. we see that the residuals plot shows the same heteroscedasticity issue that we saw in the case of batsmen as we see the residuals diverge in a cone shape as the prediction keeps increasing as shown in plot 4.5(a). The reasoning is expected to stay the

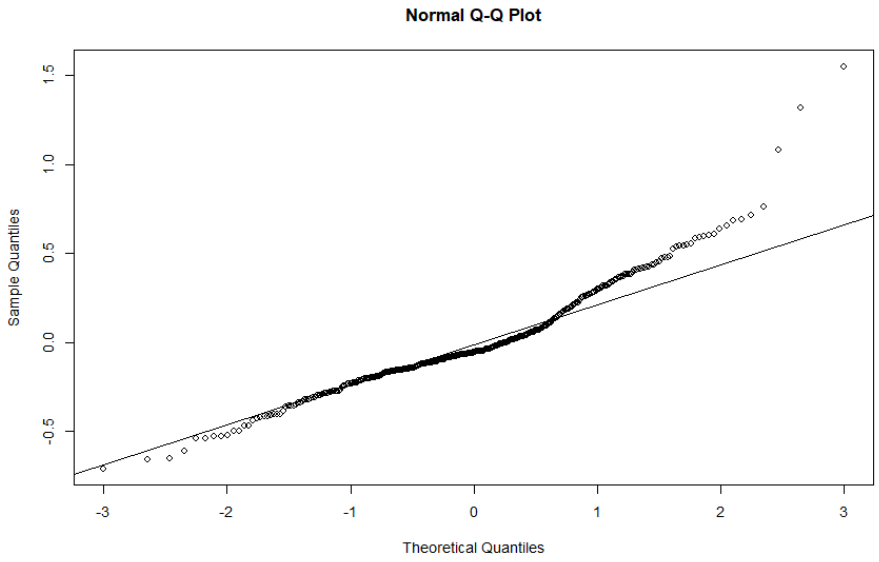
Model	RMSE	MAPE	AIC	BIC	LL
Full LM-Bowlers	0.3498010	2.689711	170.1501	244.4037	-66.07505
VarSel - pval LM-Bowlers	0.3548858	2.730100	154.2280	181.5845	-70.11398
VarSel - Step LM-Bowlers	0.3470013	2.664598	154.1035	193.1844	-67.05177

Table 4.6: Model Accuracy & IC Comparison: LM Bowlers

same since we observed that some star players have attracted very high salaries each year while the median salary remains a much lower level. The normal Q-Q plot on inspection also suggests that the normality condition is getting violated suggesting we might need to look at distributions other than Gaussian in our model choices later on.



(a) Baseline LM - Residuals vs Fitted plot showing heteroscedasticity



(b) Baseline LM - QQ plot violating normality

Figure 4.5: Baseline LM - Residuals

4.2.2 Linear Mixed Model

We first train a LMM with all features as fixed effects with a random intercept for the grouping variable Player Name. Then we employ stepwise regression on the full LMM to find the significant fixed effects and experiment with different random effects. We will choose the best model using the `anova` method. We run multiple experiments with different random effects but in the listing 4.7 we report the results of 3 experiments - full model, variable selected model with random intercept and variable selected model with random slope for a chosen random effect.

Listing 4.7: LM Model Comparison (bowlers)

```
> anova(lmm7, lmm7.step, lmm8)

Data: train
Models:
lmm7.step: Price ~ Season_num + Star_player + Active_T20I + Cum_LY_Bowl_I +
                Cum_IPL_bowl_DM_Bowl_I + (1 | Name2)
lmm8: Price ~ Season_num + Star_player + Active_T20I + Cum_LY_Bowl_I +
                Cum_IPL_bowl_DM_Bowl_I + (1 + Cum_LY_Bowl_I | Name2)
lmm7: Price ~ Season_num + Overseas_player + Star_player + Active_T20I +
                Cum_LY_Bowl_I + Cum_LY_Bowl_Econ + Cum_LY_Bowl_Avg + Cum_LY_Bowl._SR +
                Cum_LY_BW.Taker + Cum_LY_bowl_46_conc + Cum_LY_bowl_Dots. +
                Cum_IPL_bowl_DM_Bowl_I + Cum_IPL_bowl_DM_Bowl_Econ +
                Cum_IPL_bowl_DM_Bowl_Avg + Cum_IPL_bowl_DM_Bowl._SR +
                Cum_IPL_bowl_DM_BW.Taker + Cum_IPL_bowl_DM_46_conc +
                Cum_IPL_bowl_DM_Dots. + (1 | Name2)

           npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
lmm7.step     8 111.65 142.92 -47.827   95.653
lmm8         10  97.01 136.09 -38.505   77.010 18.643  2  8.948e-05 ***
lmm7         21 127.29 209.36 -42.646   85.293  0.000 11          1
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1
```

From the model comparison we can see that the most significant model is `lmm8`. This model was trained by identifying significant fixed effects using step regression on the full model and then experimenting with multiple random effects before choosing a combination of a model with random intercept with a random slope for `Cum_LY_Bowl_I`. The ANOVA test shows that this model is significant so, we next check the generalization accuracy and the IC values before choosing the best model. The accuracy and IC values are shown in table 4.7. Here, we can see that the model with the best accuracy and favourable scores is still `lmm8`. We see that this model posts a significantly better prediction accuracy and the lowest IC scores. Hence, we choose this model as the best LMM model and proceed to derive statistical inference.

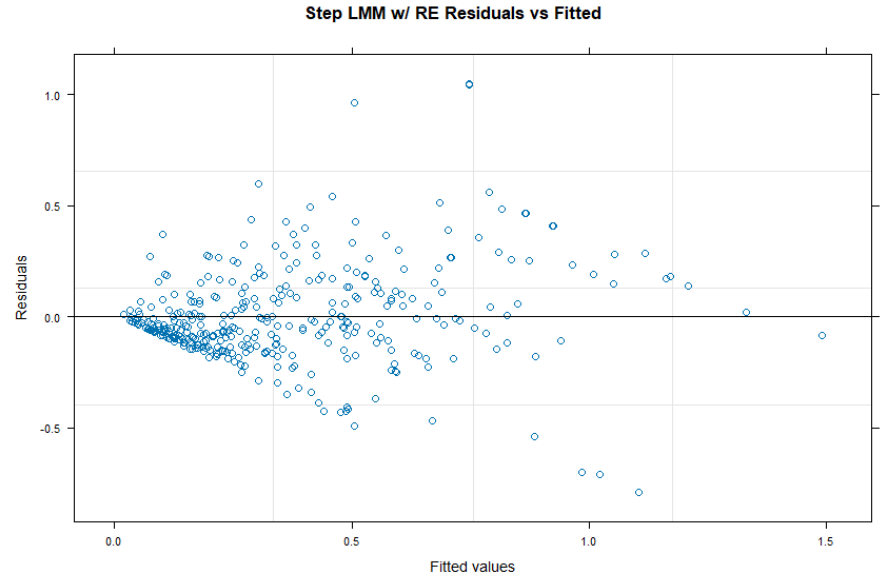
Model	RMSE	MAPE	AIC	BIC	LL
Full LMM-Bowl: RI Name	0.3116282	1.831372	127.29279	209.3625	-42.64640
Step LMM-Bowl: RI Name	0.3103053	1.896722	111.65306	142.9177	-47.82653
Step LMM-Bowl: RI RS Cum_LY_Bowl_I Name	0.3058359	1.712630	97.01015	136.0910	-38.50508

Table 4.7: Model Accuracy & IC Comparison: LMM Bowlers

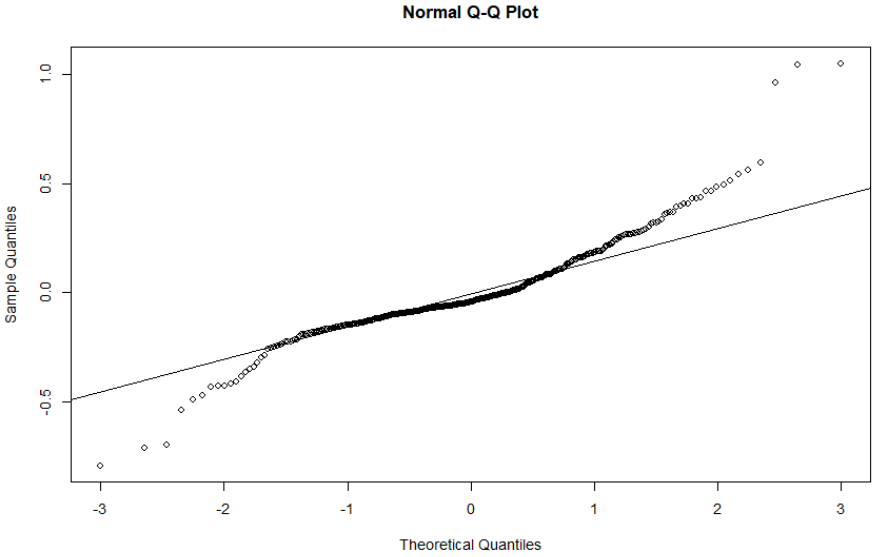
The final LMM output is shown in listing 4.9 for inspection of fixed and random effects. The LMM results mirror those of the baseline model quite closely with the significant predictors being a player's total experience in bowling in the IPL as well as in other competitions in the last year. Thus, the identified profile of a bowler being a reliable workhorse in the IPL is reinforced. Figure 4.6 shows the residuals plot, shows the recurring heteroscedasticity issue that we saw in the previous section. The Q-Q plot for the bowlers shows even more deviations that what we saw in the case of batsmen.

Listing 4.8: LMM Final Model Output (bowlers)

```
> summary(lmm8)
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: Price ~ Season_num + Star_player + Active_T20I + Cum_LY_Bowl_I +
                    Cum_IPL_bowl_DM_Bowl_I + (1 + Cum_LY_Bowl_I | Name2)
   Data: train

     AIC      BIC   logLik deviance df.resid
    97.0     136.1    -38.5     77.0       358

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.3758 -0.4588 -0.1821  0.4016  4.4783

Random effects:
 Groups   Name          Variance Std.Dev. Corr
 Name2    (Intercept)   0.029393 0.1714
          Cum_LY_Bowl_I 0.006007 0.0775   0.97
 Residual               0.055033 0.2346
Number of obs: 368, groups:  Name2, 128

Fixed effects:
                         Estimate Std. Error         df t value Pr(>|t|)
(Intercept)              0.421747   0.077964 288.048479   5.410 1.33e-07 ***
Season_num              -0.015757   0.007588 348.222834  -2.077  0.03858 *
Star_player1             0.108712   0.039164 116.892955   2.776  0.00641 **
Active_T20I1             0.093113   0.035538 349.444357   2.620  0.00917 **
Cum_LY_Bowl_I            0.069418   0.022071  84.439003   3.145  0.00229 **
Cum_IPL_bowl_DM_Bowl_I   0.131715   0.021758 165.479200   6.054 9.19e-09 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1        1

Correlation of Fixed Effects:
            (Intr) Ssn_nm Str_p1 A_T20I C_LY_B
Season_num  -0.923
Star_playr1 -0.251  0.042
Activ_T20I1 -0.004 -0.074 -0.270
Cm_LY_Bwl_I  0.216 -0.029 -0.089 -0.228
C_IPL__DM_B  0.232 -0.219  0.094 -0.009 -0.227
```



(a) LMM - Residuals vs Fitted plot



(b) LMM - QQ plot

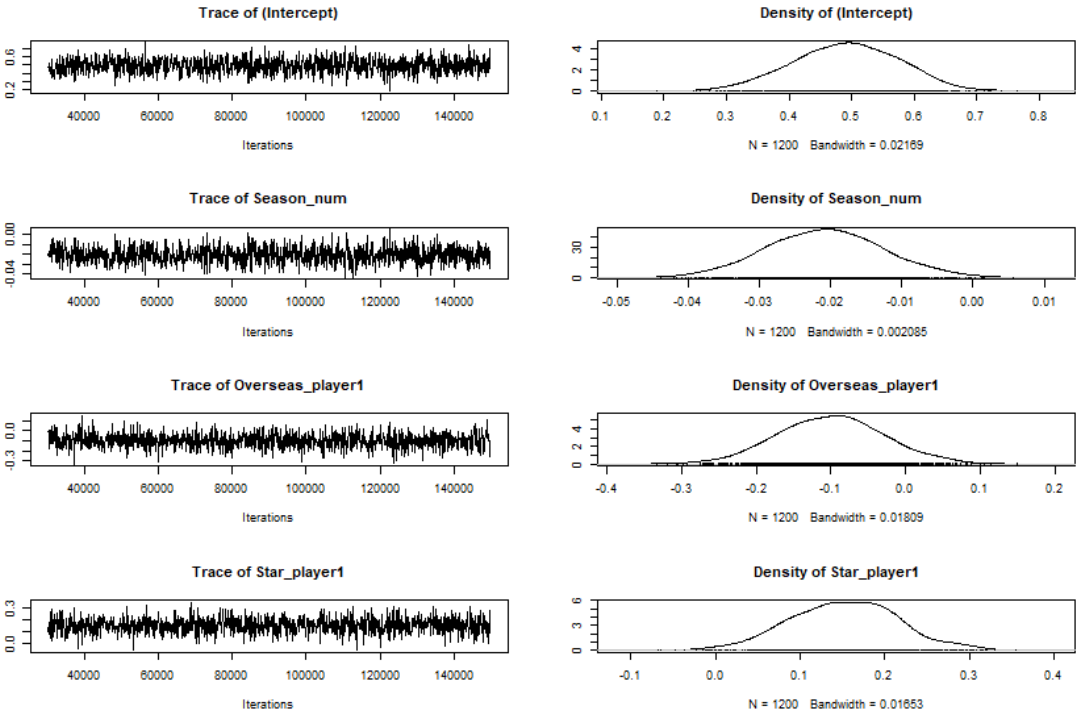
Figure 4.6: LMM Residual Plots (bowlers)

4.2.3 MCMC simulations of GLMM

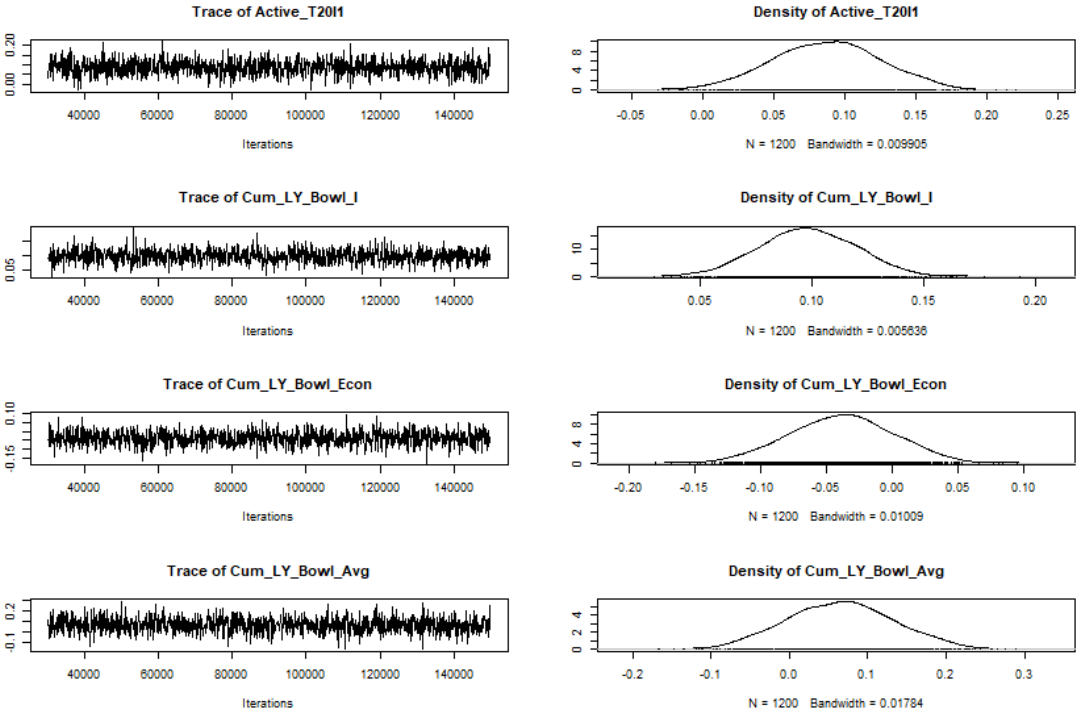
In the next step we run MCMC simulations of GLMMs for the case of bowlers. While experimenting with different family of distributions we find that the best results are achieved with the *Gaussian* family as the trait distribution setting. As in the previous section, we find that 150000 iterations with a burn in of 30000, to generate markov chains and leads to convergence. We inspect the density plots to check if the posterior means of the coefficients lies within 95% credible intervals and that the CIs do not overlap with zero. We also observe the HPD regions for each effect before finalizing variable selection for the fixed effects. First we trained a full model using all the features for which the output is shown in listing 4.10 and observed the trace plots as shown in fig 4.8

Listing 4.9: MCMC Full Model Summary - Bowlers

```
> summary(MCMCglmm1)

 Iterations = 30001:149901
 Thinning interval  = 100
 Sample size  = 1200

 DIC: 80.46458

 G-structure:  ~Name2

      post.mean l-95% CI u-95% CI eff.samp
Name2   0.03197  0.01711  0.04762     1200

 R-structure:  ~units

      post.mean l-95% CI u-95% CI eff.samp
units   0.05778  0.04795  0.06842     1686

 Location effects: Price ~ Season_num + Overseas_player + Star_player + Active_T20I +
                     Cum_LY_Bowl_I + Cum_LY_Bowl_Econ + Cum_LY_Bowl_Avg + Cum_LY_Bowl._SR +
                     Cum_LY_BW.Taker + Cum_LY_bowl_46_conc + Cum_LY_bowl_Dots. +
                     Cum_IPL_bowl_DM_Bowl_I + Cum_IPL_bowl_DM_Bowl_Econ +
                     Cum_IPL_bowl_DM_Bowl_Avg + Cum_IPL_bowl_DM_Bowl._SR +
                     Cum_IPL_bowl_DM_BW.Taker + Cum_IPL_bowl_DM_46_conc + Cum_IPL_bowl_DM_Dots.

                           post.mean  l-95% CI  u-95% CI eff.samp  pMCMC
(Intercept)                0.491536  0.331553  0.653784     1200 <8e-04 ***
Season_num                -0.020796 -0.037449 -0.005457     1200 0.0117 *
Overseas_player1          -0.099980 -0.233246  0.046705     1200 0.1583
Star_player1               0.151414  0.037614  0.293067     1200 0.0233 *
Active_T20I1               0.087172  0.016090  0.162993     1180 0.0217 *
Cum_LY_Bowl_I              0.098242  0.059297  0.142700     1200 <8e-04 ***
Cum_LY_Bowl_Econ          -0.038472 -0.109490  0.041266     1200 0.3383
Cum_LY_Bowl_Avg            0.065173 -0.067795  0.197663     1200 0.3400
Cum_LY_Bowl._SR           -0.065946 -0.206667  0.067476     1200 0.3400
Cum_LY_BW.Taker            0.011333 -0.020118  0.043341     1200 0.5000
Cum_LY_bowl_46_conc        0.024969 -0.063094  0.101677     1200 0.5550
Cum_LY_bowl_Dots.          0.004630 -0.037491  0.047064     1200 0.8500
Cum_IPL_bowl_DM_Bowl_I     0.121952  0.071923  0.176406     1200 <8e-04 ***
Cum_IPL_bowl_DM_Bowl_Econ -0.024063 -0.148834  0.105024     1200 0.7117
Cum_IPL_bowl_DM_Bowl_Avg   0.007354 -0.282537  0.301881     1200 0.9633
Cum_IPL_bowl_DM_Bowl._SR  -0.028101 -0.322434  0.250782     1441 0.8567
Cum_IPL_bowl_DM_BW.Taker   0.018948 -0.021963  0.061507     1200 0.3617
Cum_IPL_bowl_DM_46_conc   -0.008565 -0.137436  0.117373     1755 0.9083
Cum_IPL_bowl_DM_Dots.      0.030708 -0.048851  0.126620     1314 0.5250
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1  1
```

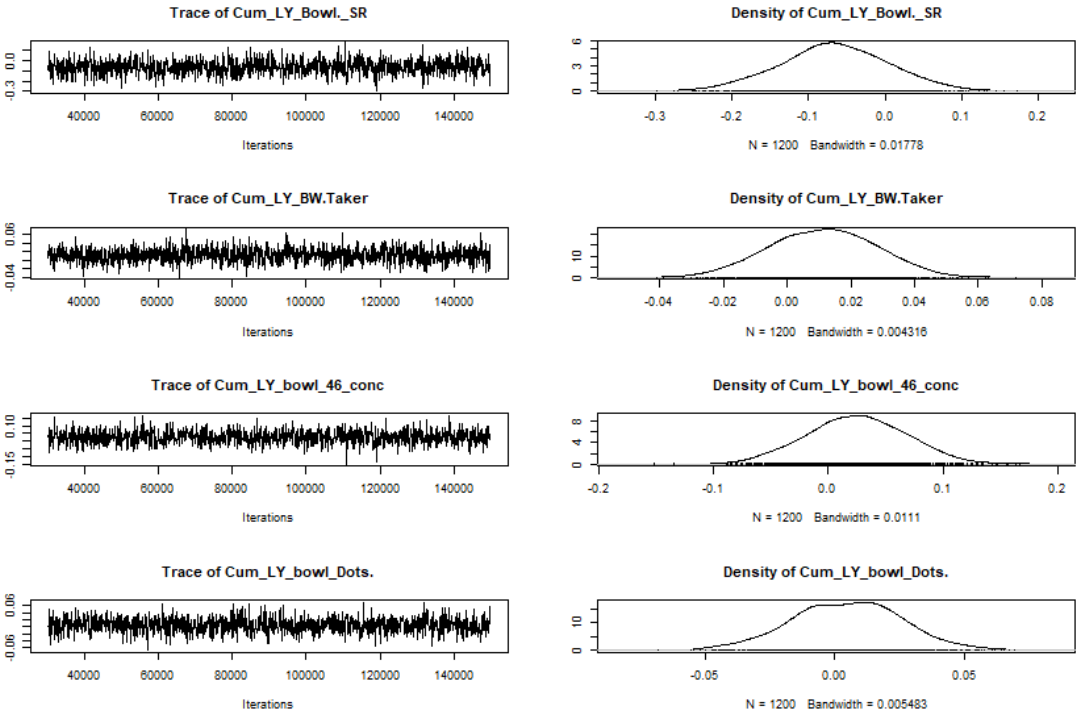


(a)

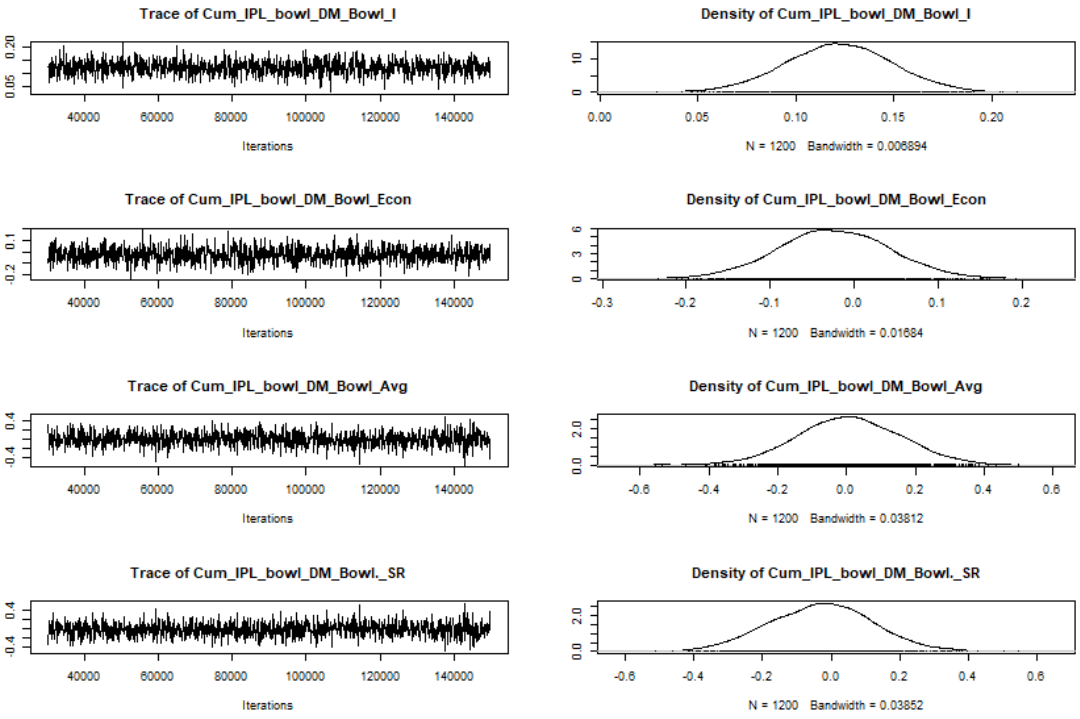


(b)

Figure 4.7: Trace plots and density estimates of posterior means - Bowlers 1/3

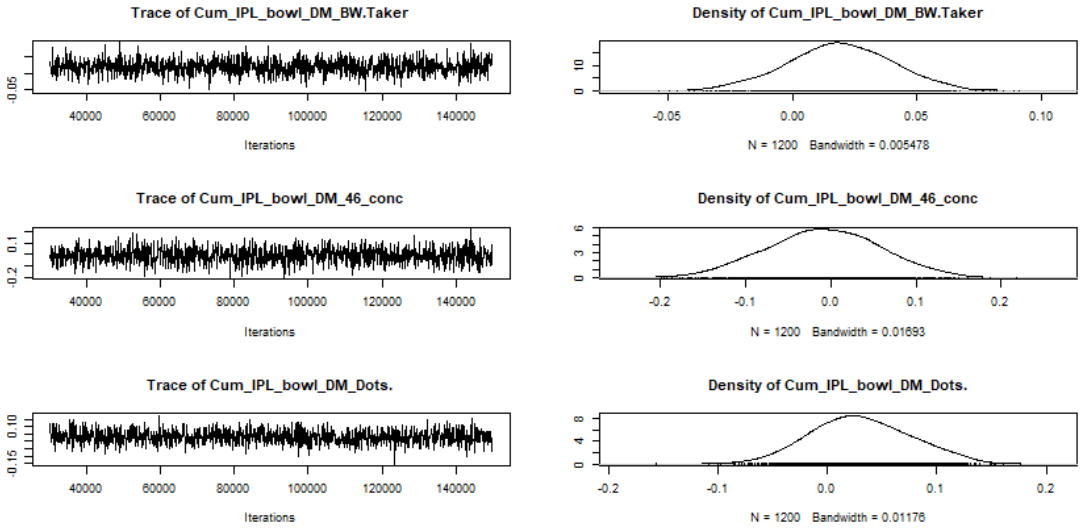


(c)

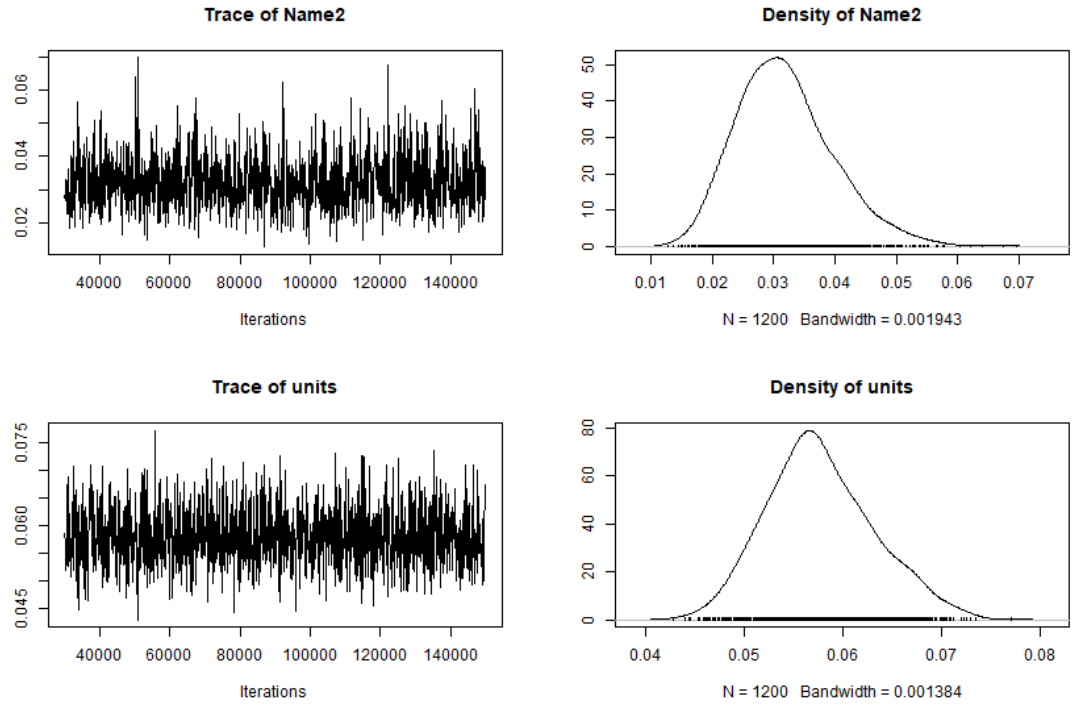


(d)

Figure 4.7: Trace plots and density estimates of posterior means - Bowlers 2/3



(e)



(f)

Figure 4.7: Trace plots and density estimates of posterior means - Bowlers 3/3

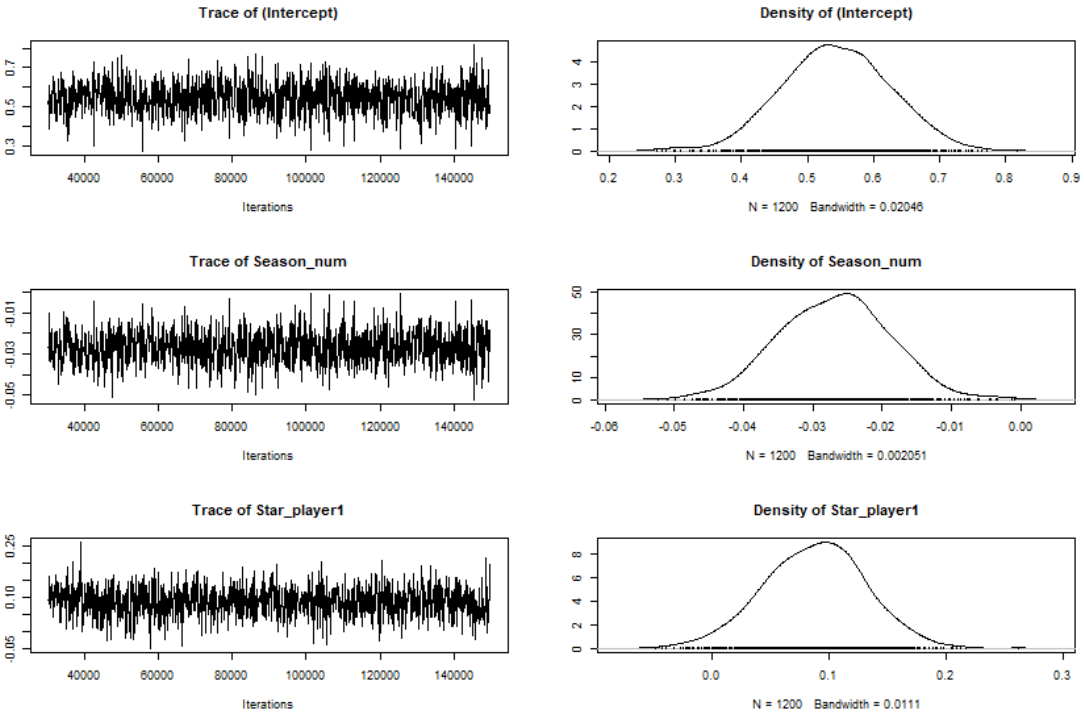
From observing the CI intervals and the HPD on the density plots, we were able to identify significant fixed effects and proceed with experimenting to find a significant random effect. The random effect which yielded the lowest generalization error is a random slope for Active T20I status within the grouping variable Player Name. The final model is summarized in listing 4.10 along with the trace plots in figure 4.8. The model selection can be motivated from the summary of results of training MCMC GLMM class of models in table 4.8 where we see a comparison of 4 model experiments. We started with a full model trained on the Gaussian Family trait which was then found to yield better results than using a model with the Exponential family trait. We then have to last two models where variable selection was applied for smaller list of fixed effects and a different combination of random effect chosen in both cases.

Listing 4.10: MCMC Final Model Summary - Bowlers

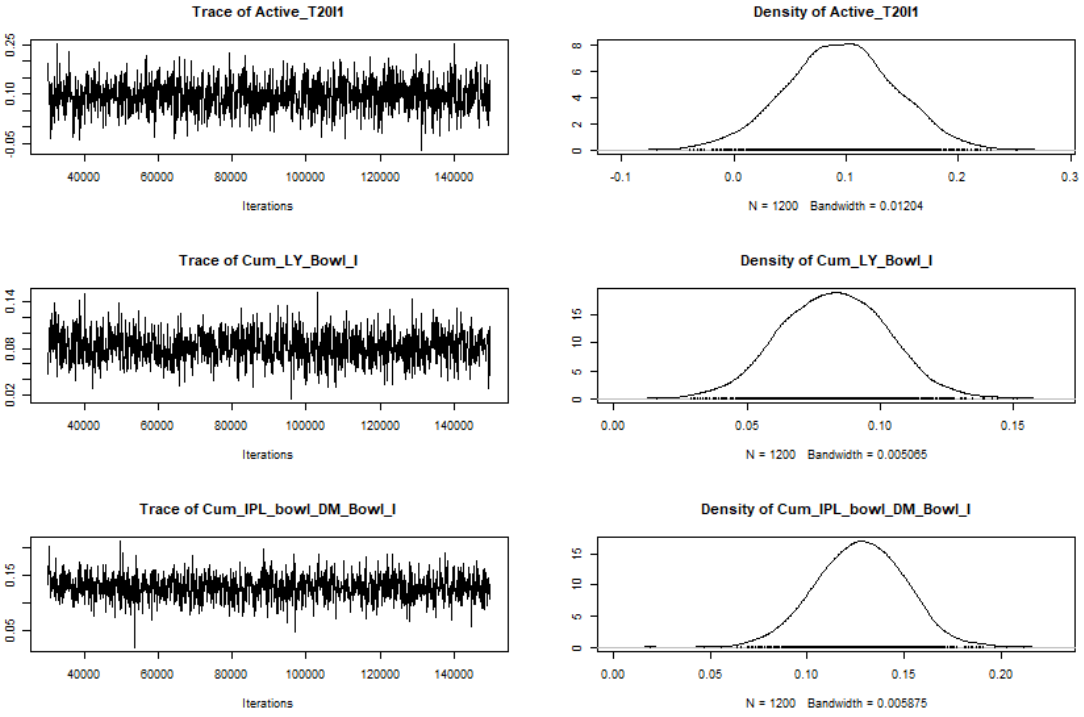
```
> summary(MCMCglmm1.VS1)

 Iterations = 30001:149901
 Thinning interval  = 100
 Sample size  = 1200

 DIC: 40.88814

 G-structure:  ~Active_T20I:Name2

                  post.mean l-95% CI u-95% CI eff.samp
Active_T20I:Name2   0.04312  0.02596  0.06023     1479

 R-structure:  ~units

      post.mean l-95% CI u-95% CI eff.samp
units   0.04955  0.03932  0.05977     1194

 Location effects: Price ~ Season_num + Star_player + Active_T20I + Cum_LY_Bowl_I + Cum_IPL_bowl_DM_Bowl_

                         post.mean  l-95% CI  u-95% CI eff.samp  pMCMC
(Intercept)              0.543044  0.383960  0.695316     1200 <8e-04 ***
Season_num              -0.026842 -0.041851 -0.011459     1200 <8e-04 ***
Star_player1             0.087968  0.001960  0.170492     1200 0.0433 *
Active_T20I1             0.094936 -0.006361  0.182759     1200 0.0533 .
Cum_LY_Bowl_I            0.082583  0.048097  0.124661     1200 <8e-04 ***
Cum_IPL_bowl_DM_Bowl_I   0.127561  0.084457  0.172582     1200 <8e-04 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1  1
```

Model	RMSE	MAPE	DIC
MCMCglmm1 Gaussian Family: RE Name2	0.3398747	1.996473	80.46458
MCMCglmm2 Exponential Family: RE Name2	0.4841747	2.219615	-198.41731
MCMCglmm1.VS Gaussian Family: RE Name2	0.3362299	1.941232	68.63488
MCMC glmm1 Gaussian Family: RS Active_T20I:Name2	0.3404402	1.778658	40.88814

Table 4.8: Model Accuracy & IC Comparison: MCMC GLMM (Bowlers)

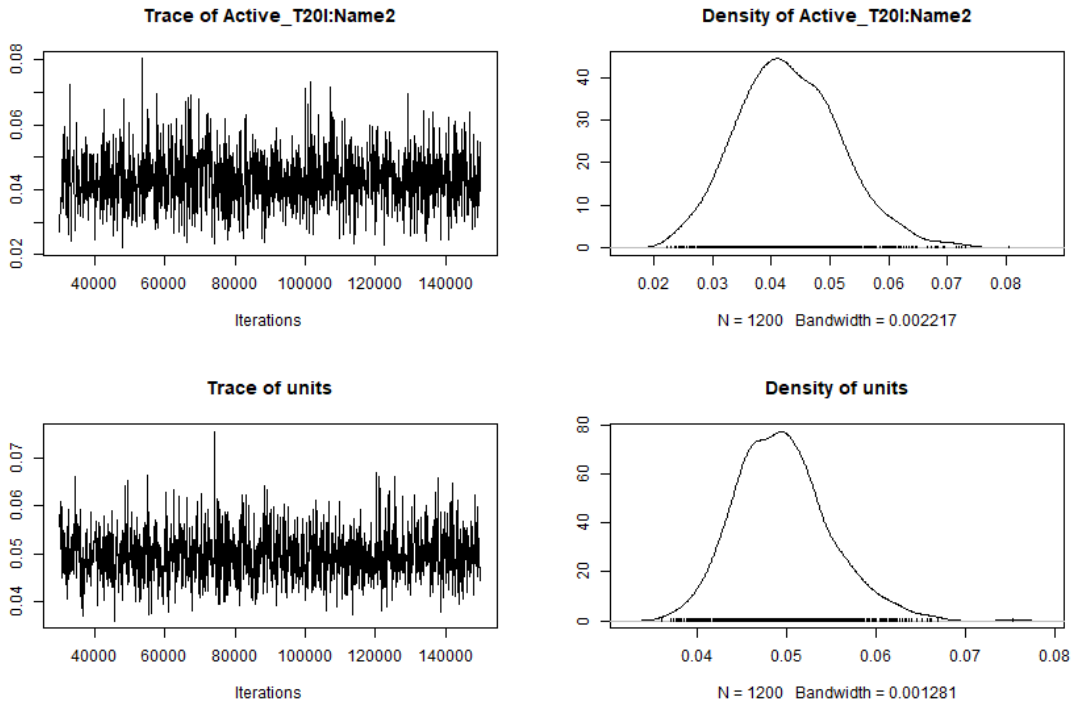


(a)



(b)

Figure 4.8: Trace plots and density estimates of posterior means - Bowlers 1/2



(c)

Figure 4.8: Trace plots and density estimates of posterior means - Bowlers 2/2

The MCMC glmm model also identifies the exact same significant predictors as we saw in the previous two model classes. There is also the identification of season number as a significant predictor which lends further evidence to a player's demonstrated ability to perform consistently over time that is prized by the teams. Thus, the overall profile for bowlers then is identified to be that of consistent and reliable work horses who are rewarded for their endurance over the long competition.

4.2.4 Summary of Results - Bowlers

The overall results from the all prediction models of bowlers are compared in table 4.9. We see that in contrast to what we found in the case of batsmen, its actually the LMM that gives us a better predictive model with lower error result on the test data. But even the MCMC simulations result in a very significant model with accuracy that is near to the best Linear Mixed Model. What is interesting to note is that the all the models identify the same performance metrics as significant thus, making the choice of predictors very clear.

Model	**RMSE**	**MAPE**
Full LM-Bowlers	0.3498010	2.689711
VarSel - pval LM-Bowlers	0.3548858	2.730100
VarSel - Step LM-Bowlers	0.3470013	2.664598
Full LMM-Bowl: RI Name	0.3116282	1.831372
Step LMM-Bowl: RI Name	0.3103053	1.896722
Step LMM-Bowl: RI RS Cum_LY_Bowl_I Name	0.3058359	1.712630
MCMCglmm1 - Gaussian Family Full Model	0.3398747	1.996473
MCMCglmm2 Exponential Family: RE Name2	0.4841747	2.219615
MCMCglmm1 Gaussian Family with Variable Selection: RE Name2	0.3362299	1.941232
MCMCglmm1 Gaussian Family with Variable Selection: RS Active_T20I RE Name2	0.3404402	1.778658

Table 4.9: Model Accuracy Comparison: LM vs LMM vs MCMC GLMM (Bowlers)

4.3 All-Rounders

4.3.1 Simple Linear Model

The linear baseline model for all-rounders is a more complex problem because we are now interested in both batting and bowling performance measures for these players. The full linear model trained on all the features was compared to the models trained after variable selection using p-values and by performing stepwise regression. The ANOVA comparison for the 3 models is shown in listing 4.11 which shows that the most significant model is the stepwise regression model.

Listing 4.11: LM Model Comparison (All-Rounders)

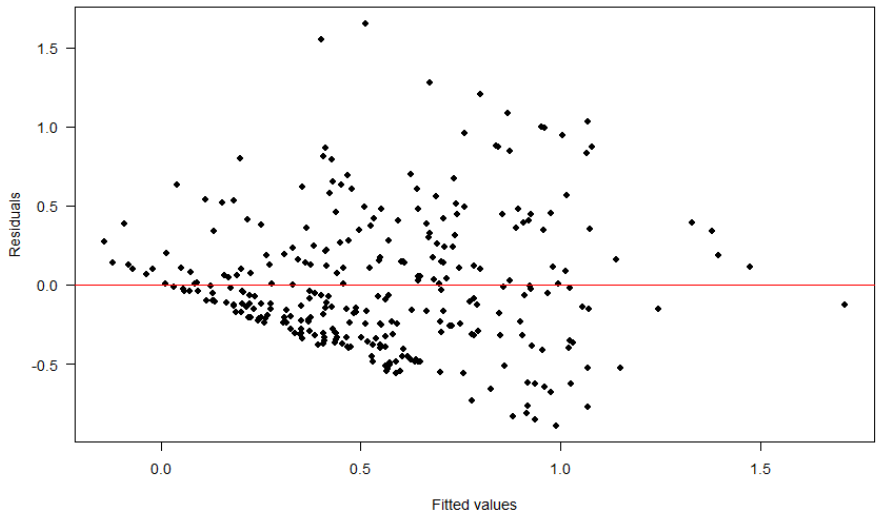
```
> anova(lm6, lm6.pval, lm6.step)
Analysis of Variance Table

Model 1: Price ~ Overseas_player + Star_player + Active_T20I + Cum_LY_Bat_I +
    Cum_LY_HardHitter + Cum_LY_Finisher. + Cum_LY_46_scored +
    Cum_LY_Bat_avg + Cum_LY_50plus_scored + Cum_LY_RBW + Cum_LY_Bowl_I +
    Cum_LY_Bowl_Econ + Cum_LY_Bowl_Avg + Cum_LY_Bowl._SR + Cum_LY_BW.Taker +
    Cum_LY_bowl_46_conc + Cum_LY_bowl_Dots. + Cum_IPL_bat_DM_Bat_I +
    Cum_IPL_bat_DM_HardHitter + Cum_IPL_bat_DM_Finisher. + Cum_IPL_bat_DM_46_scored +
    Cum_IPL_bat_DM_Bat_avg + Cum_IPL_bat_DM_50plus_scored + Cum_IPL_bat_DM_RBW +
    Cum_IPL_bowl_DM_Bowl_I + Cum_IPL_bowl_DM_Bowl_Econ + Cum_IPL_bowl_DM_Bowl_Avg +
    Cum_IPL_bowl_DM_Bowl._SR + Cum_IPL_bowl_DM_BW.Taker + Cum_IPL_bowl_DM_46_conc +
    Cum_IPL_bowl_DM_Dots.
Model 2: Price ~ Star_player + Cum_LY_50plus_scored + Cum_LY_bowl_46_conc
Model 3: Price ~ Overseas_player + Star_player + Cum_LY_HardHitter + Cum_LY_46_scored +
    Cum_LY_50plus_scored + Cum_LY_RBW + Cum_LY_Bowl_I + Cum_LY_Bowl_Econ +
    Cum_LY_Bowl_Avg + Cum_LY_Bowl._SR + Cum_LY_bowl_46_conc +
    Cum_IPL_bat_DM_Bat_I + Cum_IPL_bat_DM_Bat_avg + Cum_IPL_bat_DM_50plus_scored +
    Cum_IPL_bowl_DM_Bowl_Avg + Cum_IPL_bowl_DM_Bowl._SR + Cum_IPL_bowl_DM_BW.Taker +
    Cum_IPL_bowl_DM_46_conc + Cum_IPL_bowl_DM_Dots.
  Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
1    274 52.640
2    302 72.730 -28   -20.090 3.7347 7.654e-09 ***
3    286 53.728  16    19.002 6.1819 1.080e-11 ***
```

Not unexpectedly, the baseline linear model identifies many performance metrics across both roles as significant. The model equation is specified below.

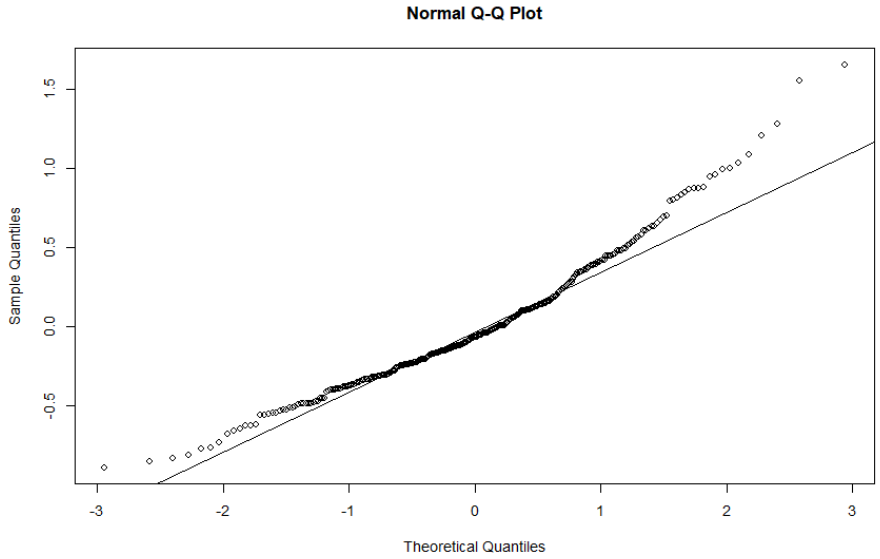
$$
\begin{aligned}
\text{Price} = {} & 0.37 - 0.15(\text{Overseas\_player}_1) \\
& + 0.36(\text{Star\_player}_1) + 0.26(\text{Cum\_LY\_HardHitter}) \\
& - 0.33(\text{Cum\_LY\_46\_scored}) - 0.08(\text{Cum\_LY\_50plus\_scored}) \\
& + 0.17(\text{Cum\_LY\_RBW}) + 0.11(\text{Cum\_LY\_Bowl\_I}) \\
& - 0.16(\text{Cum\_LY\_Bowl\_Econ}) - 0.25(\text{Cum\_LY\_Bowl\_Avg}) \\
& + 0.29(\text{Cum\_LY\_Bowl.\_SR}) + 0.18(\text{Cum\_LY\_bowl\_46\_conc}) \\
& + 0.1(\text{Cum\_IPL\_bat\_DM\_Bat\_I}) + 0.06(\text{Cum\_IPL\_bat\_DM\_Bat\_avg}) \\
& + 0.08(\text{Cum\_IPL\_bat\_DM\_50plus\_scored}) - 0.37(\text{Cum\_IPL\_bowl\_DM\_Bowl\_Avg}) \\
& + 0.35(\text{Cum\_IPL\_bowl\_DM\_Bowl.\_SR}) + 0.04(\text{Cum\_IPL\_bowl\_DM\_BW.Taker}) \\
& + 0.05(\text{Cum\_IPL\_bowl\_DM\_46\_conc}) - 0.1(\text{Cum\_IPL\_bowl\_DM\_Dots.})
\end{aligned}
$$

The wide spread of significant predictors in this model makes it hard to draw a coherent profile that would explain the valuation for the all-rounders. Upon inspecting the residuals in figure 4.9, we can see that the problem of heteroscedasticity exists for this set of players as well. We have already established that we are dealing with outliers in this data which is a

probable cause for this behavior. The Q-Q plots for all-rounders show even more deviations and susceptibility to outliers.



(a) Baseline LM - Residuals vs Fitted plot showing heteroscedasticity



(b) Baseline LM - QQ plot

Figure 4.9: Baseline LM Residual plots (All rounders)

The model comparison with prediction accuracy and IC values can be seen in table 4.10. We can see that the stepwise regression model for variable selection yields the best accuracy and lowest IC values and is a good choice for a baseline model.

4.3.2 Linear Mixed Model

We start with a full linear mixed model for all-rounders, and we use stepwise regression to identify significant fixed effect while keeping the random effect fixed as random intercept for each player. We then experiment with multiple random effects but report the model

Model	RMSE	MAPE	AIC	BIC	LL
Full LM	0.5552673	2.723436	395.7967	518.6750	-164.8983
LM Variable Selection using p-values	0.5798854	3.386221	438.7202	457.3381	-214.3601
LM Variable Selection using step regression	0.5541425	2.870196	378.0550	456.2503	-168.0275

Table 4.10: Model Comparison: LM (All-Rounders)

with the best prediction accuracy. We also train another model for comparison with variable selection for fixed effects done using p-values from the full model using random intercept within player name as the random effect. The models are then compared using the ANOVA procedure and the results are shared in the listing 4.12. We can see that the most significant model is `lmm9.step.re` which was trained by choosing fixed effects through stepwise regression and the chosen random effect is a random intercept and a random slope for `Cum_LY_50plus_scored` within the player name group. The models are also compared on their prediction accuracy and IC values in table 4.11. From this table we can see that `lmm9.step.re` has very nearly the lowest accuracy and IC values among all the models. Hence, this model is recommended as the best LMM model choice.

Listing 4.12: LMM Model Comparison (All-Rounders)

```
> anova(lmm9, lmm9.step, lmm9.pval, lmm9.step.re)

Data: train

Models:
lmm9.step: Price ~ Star_player + Cum_IPL_bat_DM_Bat_I + Cum_IPL_bat_DM_Bat_avg +
                       Cum_LY_Bowl_I + (1 | Name2)
lmm9.pval: Price ~ Star_player + Cum_LY_50plus_scored + Cum_LY_bowl_46_conc +
                       Cum_IPL_bowl_DM_Bowl_Avg + Cum_IPL_bowl_DM_Bowl._SR + (1 | Name2)
lmm9.step.re: Price ~ Star_player + Cum_IPL_bat_DM_Bat_I + Cum_IPL_bat_DM_Bat_avg +
                       Cum_LY_Bowl_I + (1 + Cum_LY_50plus_scored | Name2)
lmm9: Price ~ Season_num + Overseas_player + Star_player + Active_T20I +
                  Cum_LY_Bat_I + Cum_LY_HardHitter + Cum_LY_Finisher. + Cum_LY_46_scored +
                  Cum_LY_Bat_avg + Cum_LY_50plus_scored + Cum_LY_RBW + Cum_IPL_bat_DM_Bat_I +
                  Cum_IPL_bat_DM_HardHitter + Cum_IPL_bat_DM_Finisher. +
                  Cum_IPL_bat_DM_46_scored + Cum_IPL_bat_DM_Bat_avg + Cum_IPL_bat_DM_50plus_scored +
                  Cum_IPL_bat_DM_RBW + Cum_LY_Bowl_I + Cum_LY_Bowl_Econ + Cum_LY_Bowl_Avg +
                  Cum_LY_Bowl._SR + Cum_LY_BW.Taker + Cum_LY_bowl_46_conc + Cum_LY_bowl_Dots. +
                  Cum_IPL_bowl_DM_Bowl_I + Cum_IPL_bowl_DM_Bowl_Econ + Cum_IPL_bowl_DM_Bowl_Avg +
                  Cum_IPL_bowl_DM_Bowl._SR + Cum_IPL_bowl_DM_BW.Taker + Cum_IPL_bowl_DM_46_conc +
                  Cum_IPL_bowl_DM_Dots. + (1 | Name2)

             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
lmm9.step       7 335.85 361.92 -160.93   321.85
lmm9.pval       8 360.87 390.66 -172.44   344.87  0.000  1     1.0000
lmm9.step.re    9 339.72 373.23 -160.86   321.72 23.156  1  1.494e-06 ***
lmm9           35 358.02 488.34 -144.01   288.02 33.700 26     0.1428
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1        1
```

Model	RMSE	MAPE	AIC	BIC	LL
Full LMM: RE Name	0.4630323	2.447581	358.0190	488.3444	-144.0095
Step LMM: RE Name	0.4381216	2.119528	335.8529	361.9180	-160.9264
LMM Variable selection using p-values: RI Name	0.4541876	2.194546	360.8746	390.6633	-172.4373
Step LMM: RI RS Cum_LY_50plus_scored | Name	0.4388995	2.130147	339.7187	373.2310	-160.8594

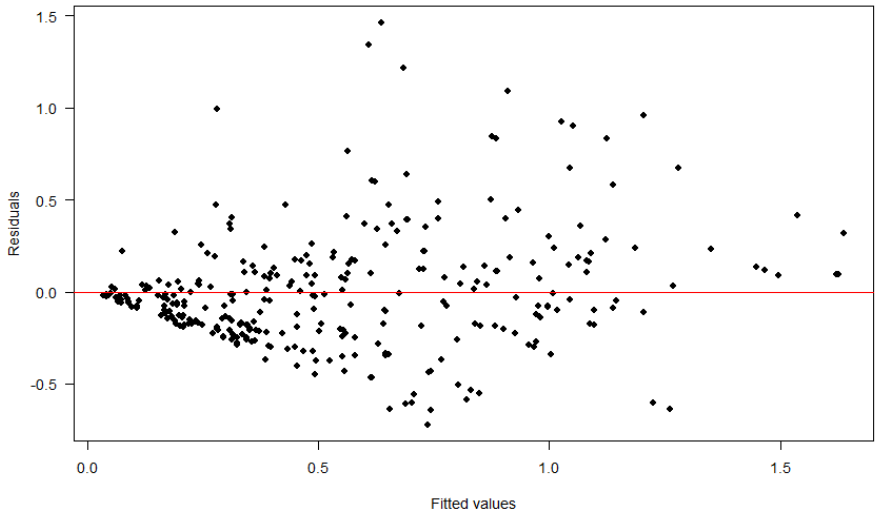
Table 4.11: Model Accuracy & IC Comparison: LMM (All-Rounders)

The final model output is shown in the listing 4.13, which shows that the significant predictors for the fixed effects are their cumulative batting and bowling experience along with their batting average and star status. This gives a sharper picture of the profile of a highly valued all-rounder in the league. The residual plots show significant deviation from normality in the Q-Q plot for outliers. The heteroscedasticity is present in chosen model as well suggesting a possible need to improvement to the model by transforming a few of the predictors.

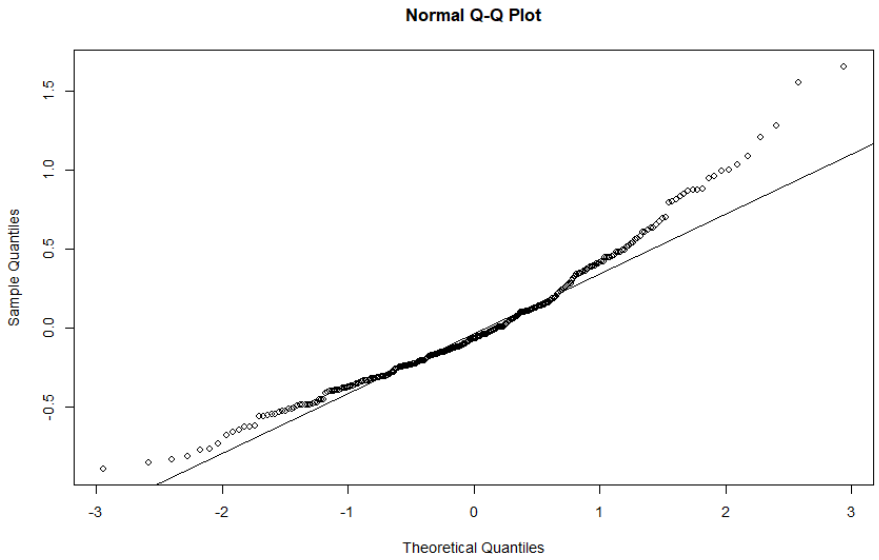
Listing 4.13: Final LMM (All-Rounders)

```
> summary(lmm9.step.re)
Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: Price ~ Star_player + Cum_IPL_bat_DM_Bat_I + Cum_IPL_bat_DM_Bat_avg +
                    Cum_LY_Bowl_I + (1 + Cum_LY_50plus_scored | Name2)
   Data: train

     AIC      BIC   logLik deviance df.resid
   339.7    373.2   -160.9    321.7      297

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.0308 -0.5173 -0.1309  0.3589  4.1247

Random effects:
 Groups   Name                 Variance  Std.Dev. Corr
 Name2    (Intercept)          0.0677966 0.26038
          Cum_LY_50plus_scored 0.0001536 0.01239  1.00
 Residual                      0.1258369 0.35473
Number of obs: 306, groups:  Name2, 95

Fixed effects:
                        Estimate Std. Error        df t value Pr(>|t|)
(Intercept)              0.35304    0.06586 105.23654   5.361 4.93e-07 ***
Star_player1             0.25320    0.07934 111.86541   3.191  0.00184 **
Cum_IPL_bat_DM_Bat_I     0.11568    0.03998 188.82960   2.894  0.00426 **
Cum_IPL_bat_DM_Bat_avg   0.06790    0.03036 261.15443   2.236  0.02617 *
Cum_LY_Bowl_I            0.08335    0.02964 281.81484   2.813  0.00526 **
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Correlation of Fixed Effects:
            (Intr) Str_p1 C_IPL__DM_B_I Cm_IPL__DM_B_
Star_playr1 -0.838
C_IPL__DM_B_I  0.113 -0.015
Cm_IPL__DM_B_  0.044 -0.073 -0.395
Cm_LY_Bwl_I    0.268 -0.289 -0.176        -0.035
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```

It can be summarized from this model that an all-rounder is valued for their contribution with bat and bowl in each match along with an expectation to score runs regularly throughout the competition. The random effect that yields the best model is the player's ability to score big runs every few innings i.e. teams expect all-rounders to make a significant contribution with the bat every few matches.



(a) LMM - Residuals vs Fitted plot showing heteroscedasticity



(b) LMM - QQ plot

Figure 4.10: LMM Residual Plots (All-rounders)

4.3.3 MCMC simulations of GLMM

We start with training the full model for all rounders as shown in listing 4.14. The Exponential family trait was chosen because the distribution of all-rounders closely resembled that of the batsmen as shown in figure 3.5 in section 3.2. For comparison we also modelled the full model with the Gaussian family trait but the results were not favorable. The trace plots in figure 4.11 for the full model help us identify the significant fixed effects by observing the CIs and HPD regions on the density plots. We eliminate the predictors that whose CIs cross over 0 and the HPD region does not reside over 0.

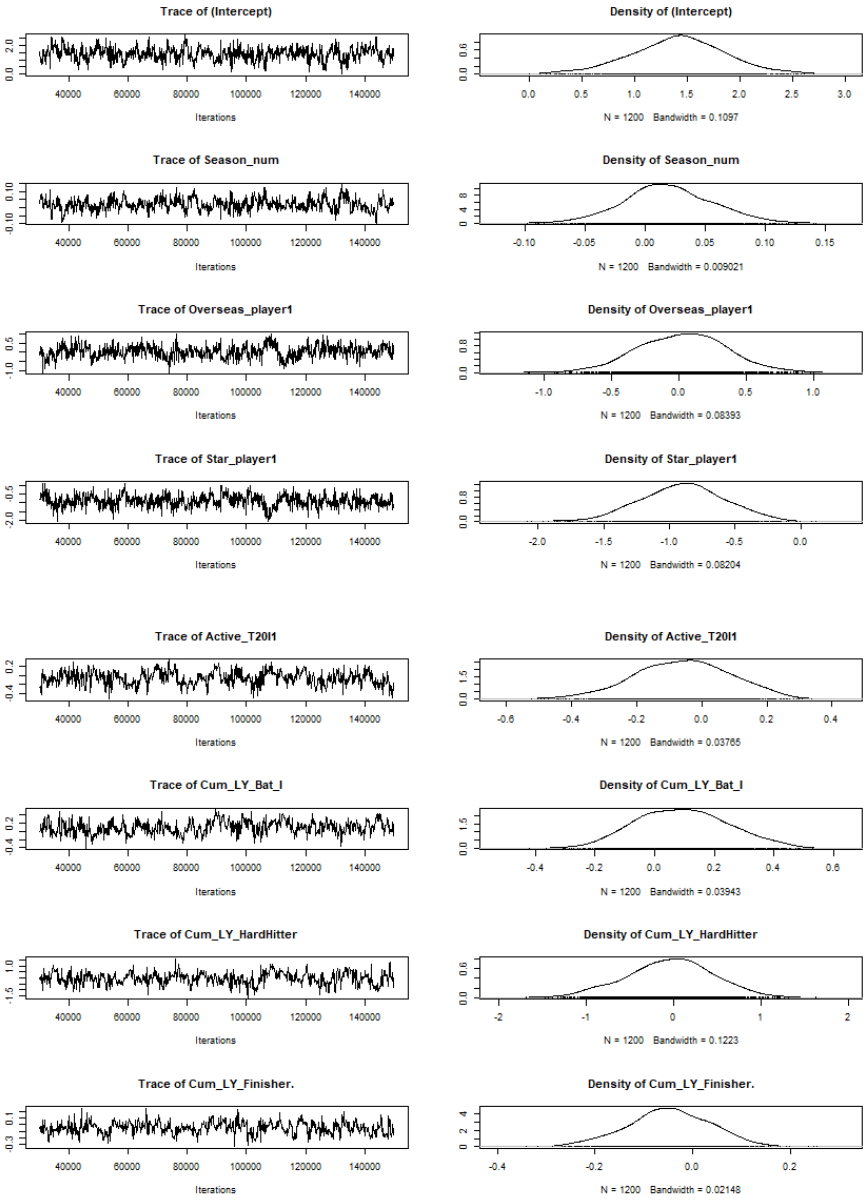


Figure 4.11: Trace plots and density estimates of posterior means - All-rounders 1/5

After variable selection the three models are compared in table 4.12 with their prediction accuracy and the DIC values. We can see that based on the lowest prediction accuracy and DIC value, the best model is `MCMCglmm6.vs1` which was trained after variable selection with

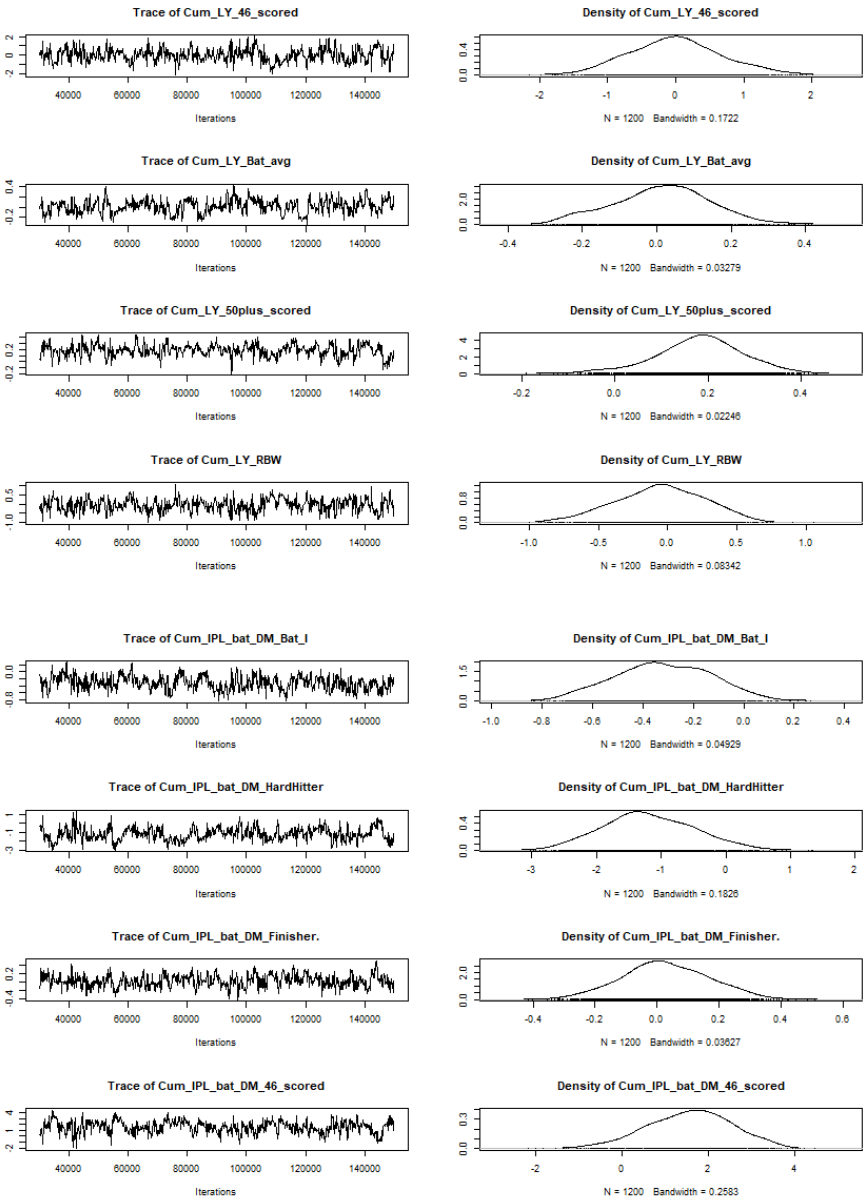


Figure 4.11: Trace plots and density estimates of posterior means - All-rounders 2/5

the exponential family and a random intercept for each player. We describe this final model in the listing 4.15 along with the trace plots for the final model in figure 4.12.

Model	**RMSE**	**MAPE**	**DIC**
MCMCglmm6 Full Model Exponential Family: RE Name2	0.7396640	2.522691	153.2253
MCMCglmm6 Exponential Family: Variable Selection using CI with RE Name2	0.5210748	2.367108	127.8410
MCMC GLMM-AR Full Model Gaussian Family: RE Name2	0.5622114	2.736657	309.2352

Table 4.12: Model Comparison: MCMC GLMM (All-Rounders)

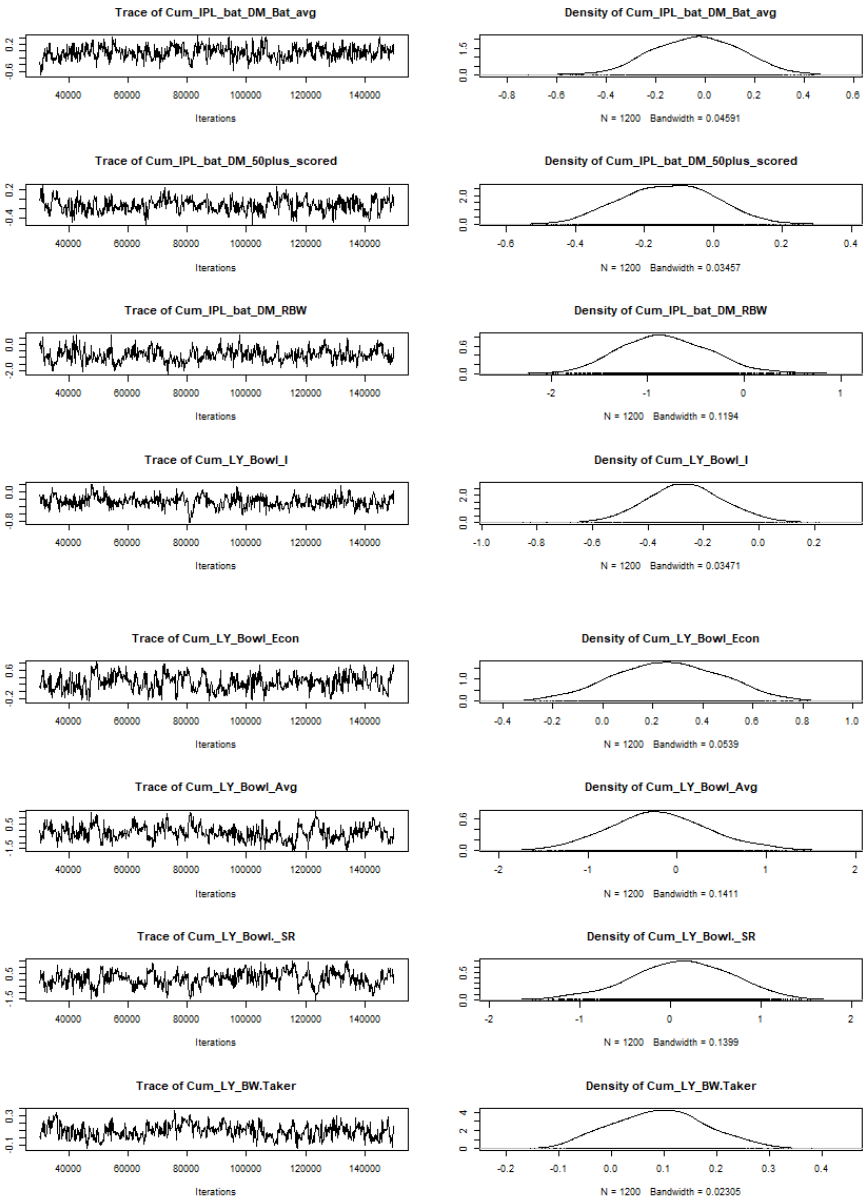


Figure 4.11: Trace plots and density estimates of posterior means - All-rounders 3/5

From the final model we can infer that the MCMC GLMM suggests that the ideal profile for an all-rounder takes into account specific aspects of both batting and bowling. In terms of batting ability they are valued for their overall batting experience in the league along with their ability to score runs quickly by hitting boundaries. Their bowling ability is valued for their recent form and the ability to not give away too many boundaries.

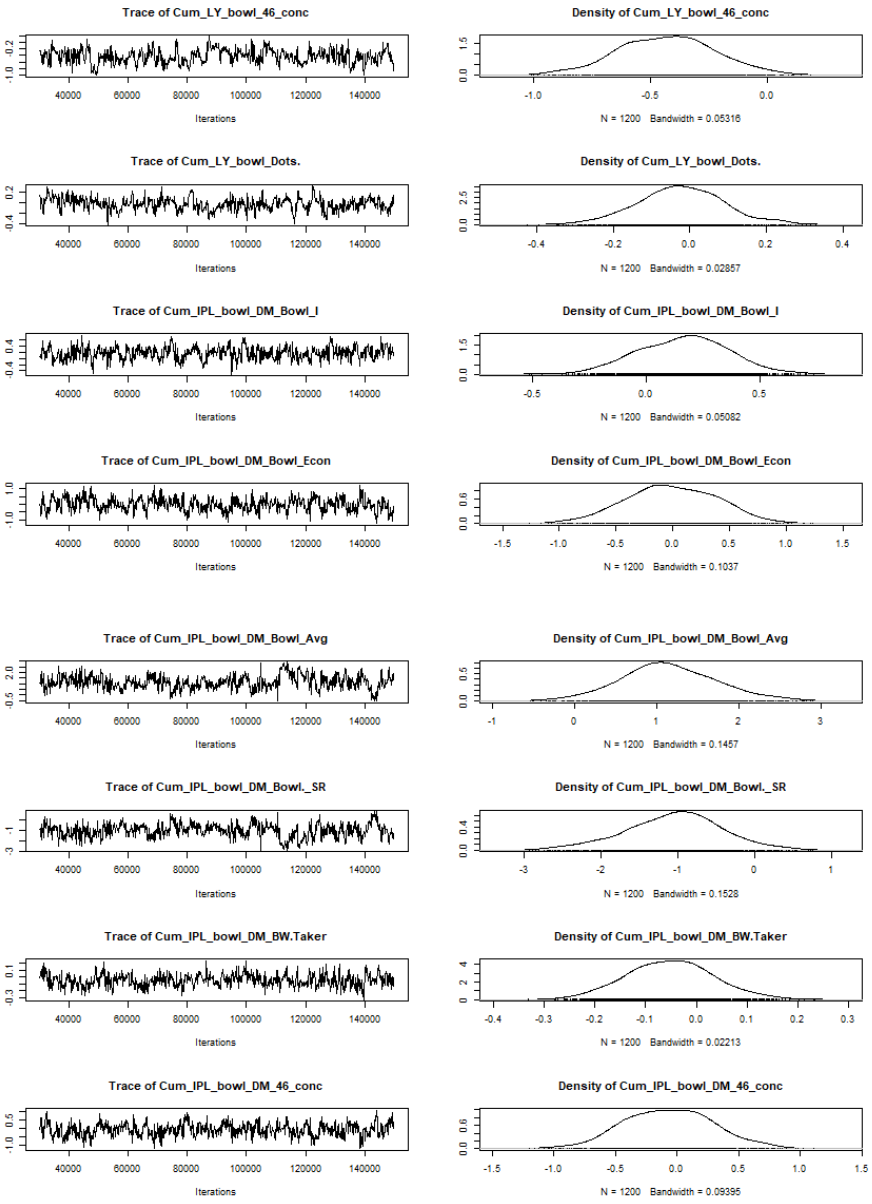


Figure 4.11: Trace plots and density estimates of posterior means - All-rounders 4/5

Listing 4.14: Full MCMC GLMM (All-Rounders)

```
> summary(MCMCglmm6)

 Iterations = 30001:149901
 Thinning interval  = 100
 Sample size  = 1200

 DIC: 153.2253

 G-structure:  ~Name2

       post.mean l-95% CI u-95% CI eff.samp
Name2     0.3302    0.1268    0.5887     73.42

 R-structure:  ~units

       post.mean  l-95% CI u-95% CI eff.samp
units  0.009296 0.0002623  0.03407      128.2

 Location effects: Price ~ Season_num + Overseas_player + Star_player + Active_T20I + Cum_LY_Bat_I + Cum
    Cum_IPL_bowl_DM_Bowl_I + Cum_IPL_bowl_DM_Bowl_Econ + Cum_IPL_bowl_DM_Bowl_Avg + Cum_IPL_bowl_DM_Bowl.

                              post.mean   l-95% CI   u-95% CI eff.samp   pMCMC
(Intercept)                   1.4097800  0.6125252  2.4463137   166.18 0.00167 **
Season_num                    0.0193638 -0.0551864  0.0922908   142.56 0.59333
Overseas_player1              0.0248513 -0.5925835  0.6960682    98.77 0.91333
Star_player1                 -0.8993209 -1.5073089 -0.2274831   114.30 0.00500 **
Active_T20I1                 -0.0586547 -0.3347683  0.2255977    76.81 0.70667
Cum_LY_Bat_I                  0.0931001 -0.1765045  0.4139170    82.39 0.58500
Cum_LY_HardHitter            -0.0428393 -1.0592851  0.8392314   121.90 0.96333
Cum_LY_Finisher.             -0.0477376 -0.2231411  0.1086936   156.07 0.58000
Cum_LY_46_scored             -0.0007902 -1.3048100  1.3329254   148.80 0.97833
Cum_LY_Bat_avg                0.0138460 -0.2427122  0.2390640   106.05 0.88667
Cum_LY_50plus_scored          0.1808187 -0.0012796  0.3822835   127.92 0.08333 .
Cum_LY_RBW                   -0.0533364 -0.6752704  0.5637790   166.24 0.87333
Cum_IPL_bat_DM_Bat_I         -0.3188564 -0.6733797  0.0486547   121.47 0.08667 .
Cum_IPL_bat_DM_HardHitter    -1.1864319 -2.4749753  0.2305226    95.26 0.10500
Cum_IPL_bat_DM_Finisher.      0.0285767 -0.2408110  0.3090933   140.24 0.85000
Cum_IPL_bat_DM_46_scored      1.5580951 -0.2584318  3.5960301    85.34 0.13167
Cum_IPL_bat_DM_Bat_avg       -0.0342637 -0.3729368  0.3072869   147.56 0.84000
Cum_IPL_bat_DM_50plus_scored -0.1296949 -0.3777290  0.1311047   124.55 0.34000
Cum_IPL_bat_DM_RBW           -0.8157939 -1.7083551  0.0603710   129.38 0.08500 .
Cum_LY_Bowl_I                -0.2643450 -0.5257820  0.0371893   120.88 0.06833 .
Cum_LY_Bowl_Econ              0.2598284 -0.1480117  0.6593228   133.73 0.21500
Cum_LY_Bowl_Avg              -0.1990496 -1.3229523  0.8481995    71.58 0.68667
Cum_LY_Bowl._SR               0.1246247 -0.9204004  1.2201969    74.69 0.79833
Cum_LY_BW.Taker               0.0921987 -0.0834452  0.2543014   109.83 0.33833
Cum_LY_bowl_46_conc          -0.4301925 -0.8558220 -0.0256279   125.63 0.04167 *
Cum_LY_bowl_Dots.            -0.0239279 -0.2406051  0.2247473   123.98 0.82500
Cum_IPL_bowl_DM_Bowl_I        0.1637501 -0.2428216  0.5160849   183.97 0.43000
Cum_IPL_bowl_DM_Bowl_Econ    -0.0082600 -0.7805283  0.7610781   131.88 0.96167
Cum_IPL_bowl_DM_Bowl_Avg      1.1498772  0.0410832  2.4278259    87.26 0.05167 .
Cum_IPL_bowl_DM_Bowl._SR     -1.0557895 -2.4583044  0.1206940    82.14 0.08667 .
Cum_IPL_bowl_DM_BW.Taker     -0.0509652 -0.2199976  0.1183289   155.67 0.54333
Cum_IPL_bowl_DM_46_conc      -0.0660969 -0.6978677  0.7108245   141.21 0.87167
Cum_IPL_bowl_DM_Dots.         0.1822360 -0.1087481  0.4845616   110.10 0.23500
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
```

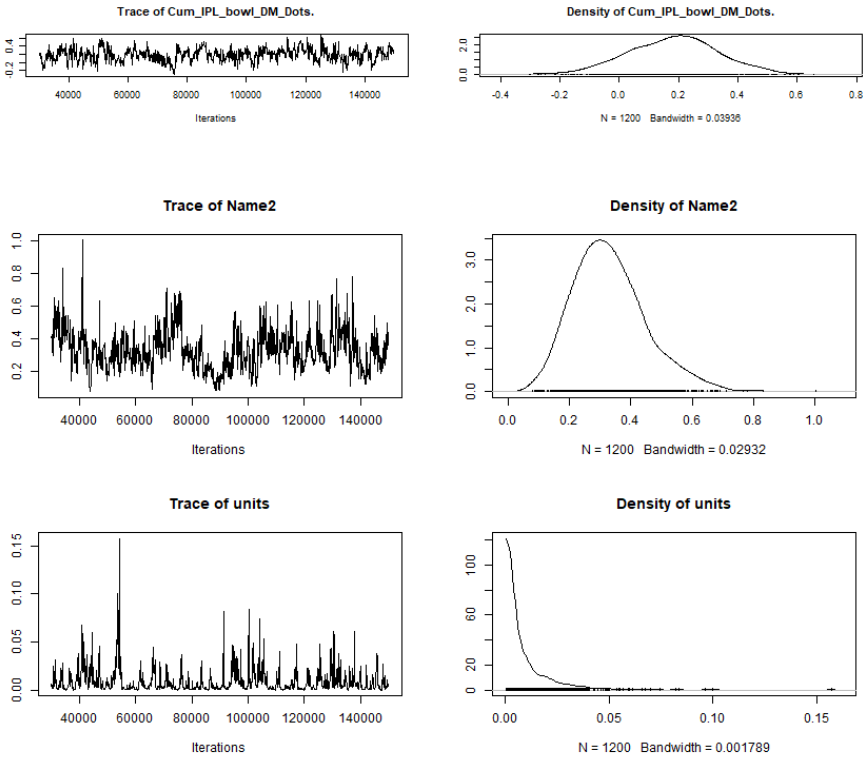


Figure 4.11: Trace plots and density estimates of posterior means - All-rounders 5/5

Listing 4.15: Final MCMC GLMM (All-Rounders)

```
> summary(MCMCglmm6.vs1)

 Iterations = 30001:149901
 Thinning interval  = 100
 Sample size  = 1200

 DIC: 127.841

 G-structure:  ~Name2

      post.mean l-95% CI u-95% CI eff.samp
Name2    0.3579    0.159   0.6067    77.85

 R-structure:  ~units

      post.mean l-95% CI u-95% CI eff.samp
units  0.008626 0.000253  0.02981      157

 Location effects: Price ~ Star_player + Cum_LY_50plus_scored + Cum_IPL_bat_DM_Bat_I +
                                         Cum_IPL_bat_DM_HardHitter + Cum_IPL_bat_DM_46_scored +
                                         Cum_IPL_bat_DM_RBW + Cum_LY_Bowl_I + Cum_LY_bowl_46_conc +
                                         Cum_IPL_bowl_DM_Bowl_Avg + Cum_IPL_bowl_DM_Bowl._SR

                           post.mean   l-95% CI   u-95% CI eff.samp  pMCMC
(Intercept)                1.604e+00  1.250e+00  1.978e+00   158.27 <8e-04 ***
Star_player1              -9.264e-01 -1.362e+00 -4.910e-01   144.27 <8e-04 ***
Cum_LY_50plus_scored       1.019e-01 -5.315e-02  2.464e-01   116.52 0.2017
Cum_IPL_bat_DM_Bat_I      -2.307e-01 -4.563e-01  3.425e-05    93.16 0.0417 *
Cum_IPL_bat_DM_HardHitter -1.187e+00 -2.365e+00  2.209e-02   107.48 0.0517 .
Cum_IPL_bat_DM_46_scored   1.533e+00 -1.577e-01  3.239e+00   112.90 0.0783 .
Cum_IPL_bat_DM_RBW        -8.171e-01 -1.615e+00 -7.218e-02   121.20 0.0467 *
Cum_LY_Bowl_I             -1.196e-01 -2.902e-01  3.705e-02   110.87 0.1567
Cum_LY_bowl_46_conc       -1.906e-01 -3.404e-01 -4.197e-02   109.90 0.0217 *
Cum_IPL_bowl_DM_Bowl_Avg   5.072e-01 -2.831e-01  1.308e+00    89.93 0.2033
Cum_IPL_bowl_DM_Bowl._SR  -4.087e-01 -1.260e+00  3.911e-01    95.36 0.3167
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1  1
```

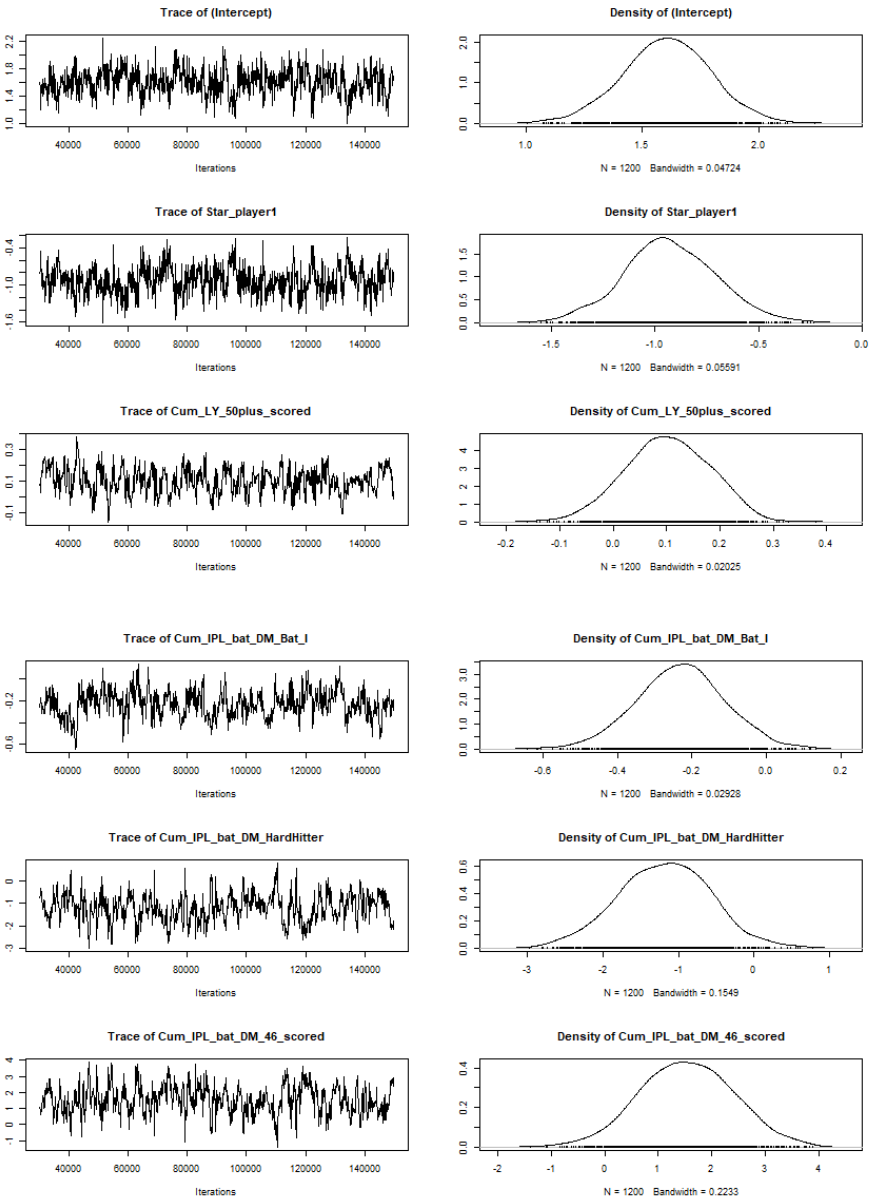


Figure 4.12: Trace plots and density estimates of posterior means - All-rounders 1/3

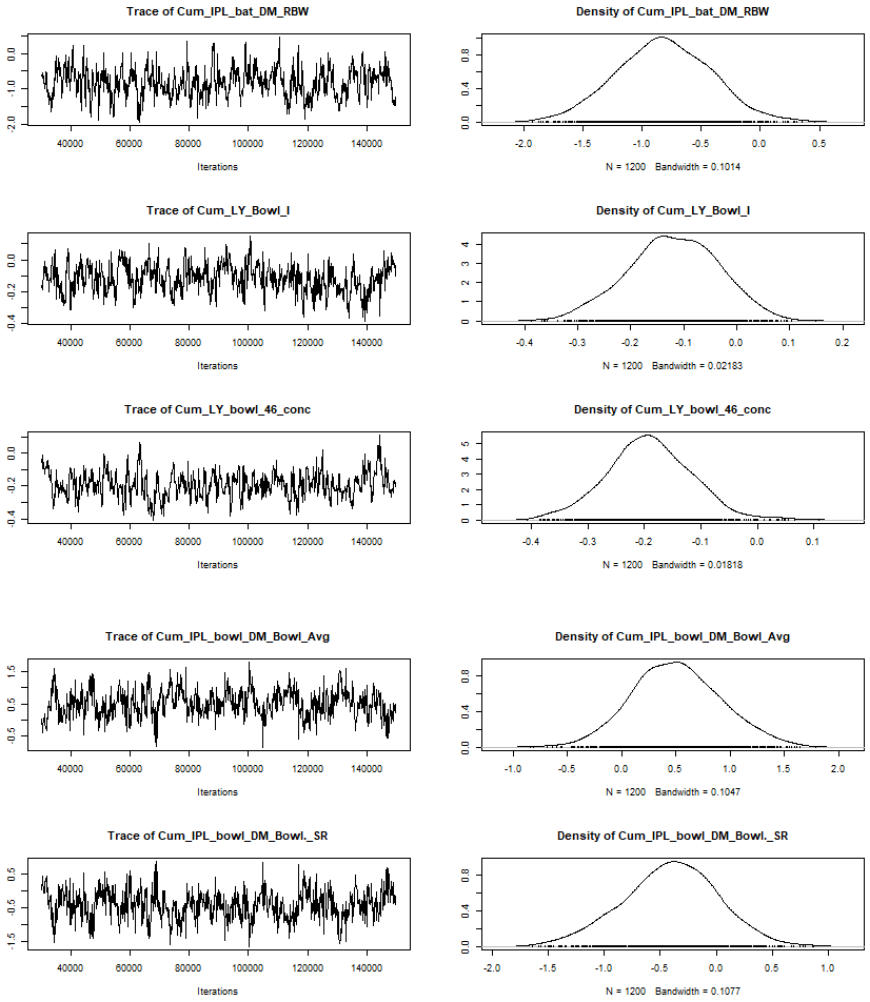


Figure 4.12: Trace plots and density estimates of posterior means - All-rounders 2/3

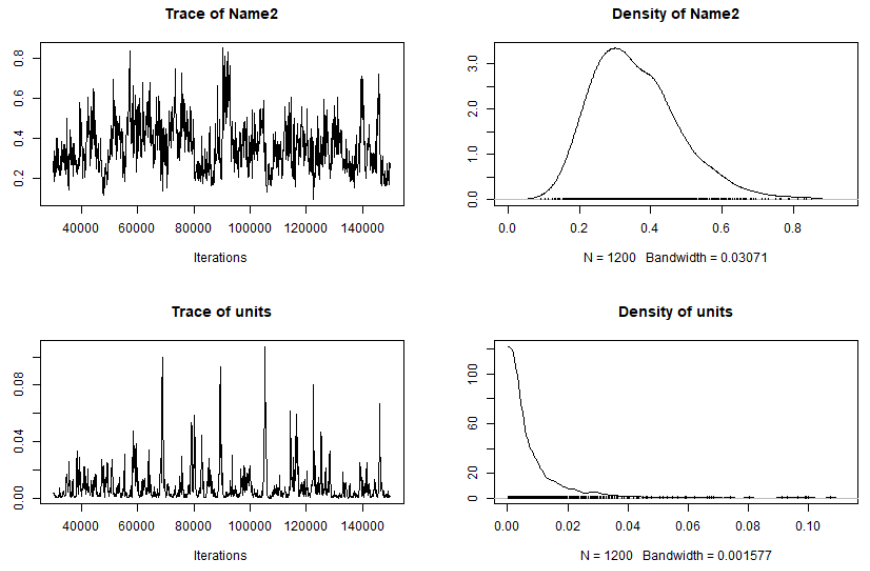


Figure 4.12: Trace plots and density estimates of posterior means - All-rounders 3/3

4.3.4 Summary of Results - All Rounders

The models from all three model classes - baseline, LMM and MCMC GLMM are compared in the table 4.13. We see that like in the case of bowlers it is not the MCMC simulations that lead to the best predictive model, for all-rounders it is the LMM that gives us the lowest generalization error result on the test data. The resulting profile from the LMM and MCMC simulated models was also slightly different, with the MCMC models giving a more specific insight into both the batting and bowling ability metrics that affect the valuation of an all rounder.

Model	**RMSE**	**MAPE**
Full LM	0.5552673	2.723436
LM Variable Selection using p-values	0.5798854	3.386221
LM Variable Selection using step regression	0.5541425	2.870196
Full LMM: RE Name	0.4630323	2.447581
Step LMM: RE Name	0.4381216	2.119528
LMM Variable selection using p-values: RI Name	0.4541876	2.194546
Step LMM: RI RS Cum_LY_50plus_scored | Name	0.4388995	2.130147
MCMCglmm6 Full Model exponential Family: RE Name2	0.7396640	2.522691
MCMCglmm6 exponential Family: Variable Selection using CI with RE Name2	0.5210748	2.367108
MCMC GLMM-AR Full Model Gaussian Family: RE Name2	0.5622114	2.736657

Table 4.13: Model Comparison: LM vs LMM vs MCMC GLMM (All-Rounders)

5 Discussion

This study has been compiled by running a series of experiments for each playing role and in this section I will share my interpretation of the models used and their caveats in the context of the game and its rules and nuances. In accordance with the structure of analysis followed in this study, I will discuss each of the 3 roles and their results separately. But before I go into role specific insights an important observation from the results in general should be discussed separately. In a franchise league based sporting competition like the IPL, the team owners and managers will always try to manage their spending budget by balancing their team composition. They do so by having a mix of high profile superstars that typically cost a lot of money to get, and a group of domestic (or sometimes international) players who don't cost a lot of money but can be crucial to a team's balance. This is the most important reason why the data that we deal with has a very wide range of player salaries. It is also the reason why outliers are such an integral part of this study. This is also why in almost every model that has been reported in the results section, star player has been identified as a significant predictor. Across model classes, this predictor has remained very significant for all the roles. The other 2 categorical variables in the data, Overseas player & Active T20I player, have also been observed as important features across different models. This makes sense because an overseas player is critical to the team's balance because the IPL has a rule that a team is not allowed to field more than 4 overseas players in any match. So teams have this extra constraint to process when considering spending money on overseas players and how they would fit into the team. A player's active T20I status can mean that the player is currently in good enough form to be picked to represent their national team at the highest level of play. This makes them equivalent to a star player and garners extra attention from team managements and sets up a potential bidding war for their acquisition.

5.1 Results

5.1.1 Batsmen

Looking at the results for the batsmen, in the baseline model we saw that all 3 categorical variable except WK are significant. We see that a batsman's cumulative experience of batting in the IPL is what is valued most.

Looking at the LMM, we find that in addition to the categorical predictors - nationality, Star and T20I status, a significant predictor is found to be scores of 50 plus i.e. the number of times a player has made a very significant contribution to the team with the bat throughout their entire career in the IPL. The other significant features are the ability to score quickly by hitting boundaries. This is very intuitive as T20 is known to be a very fast paced format.

Moving to MCMC simulations, we observe that along with nationality, Star and T20I status, the other important predictor is a player's cumulative experience of batting in the league. This shows that teams value a player for their longevity in the IPL. We also observe the significance of scoring boundaries for a player's value. A different predictor to stand out was that of Finisher i.e. in recent history how frequently is the player closing out the match for his team. In the context of the game this is also a very prized role as the player that remains not-out at the end of the match has successfully blocked the opposition from gaining his wicket and thus taken away their ability to swing the game in the favor of the opposition.

5.1.2 Bowlers

In the classical linear models for bowlers we didn't find a significant model but in the chosen baseline model we saw that the categorical predictors again proved to be significant, again supporting the argument that teams put a premium on players with a higher star value & international experience. In the baseline linear model we see that the most important predictors are the number of times the bowler bowled in the last whole year across competitions and his cumulative bowling experience in all the IPL seasons preceding the current one. This leads us to think that rather than putting emphasis on bowling performance, the teams are prioritizing players with overall playing experience and fitness. How many wickets the bowler is taking or runs he is giving away is not as important in the pricing decision.

When we move to the LMMs, we see the same predictors come out as the significant fixed effects. The best model utilizes the cumulative bowling experience in the preceding year as the random effect within player names. So, the LMM backs up the baseline model inference that the teams value bowling experience of the players more than their in-game contribution.

Unsurprisingly, the MCMC simulation estimations also support these predictors as the most important. An important observation about the MCMC simulations for the bowlers is that models trained with family parameter as Gaussian, proved to give more accurate predictions. In the EDA we had observed a slightly less skewed distribution for the bowler saalries as compared to the batsmen and the all-rounders, both of which had a thicker right hand tail. This could be the reason why the Gaussian trait specification performs much better here.

A contextual explanation about these results can be that the teams are willing to invest on bowlers that can reliably bowl in the match that they play in. To elaborate, the more number of matches the bowler is bowling is reflective of the team captain and management's confidence in the said bowler to deliver his overs with satisfactory returns. If a player is bowling too expensively and/or not taking wickets, then it stands to reason that the team will sideline the bowler and not let him participate in the team's bowling effort. Also, it is often said that T20 is a batsman's game, thus implying that a bowler's perceived importance to winning the game is considerably lesser than that of a batsman.

5.1.3 All Rounders

In the case of All-rounders we deal with a higher complexity since all-rounders are evaluated on both their batting and bowling performance. So the number of predictors nearly doubles leading to a higher challenge in explaining the model. We see this upfront in the baseline linear model, where we find many significant features for the baseline model. This suggests that while all-rounders are expected to perform both roles, but what drives their valuation is unclear.

The LMM however, reveals a much sharper mix of both batting and bowling significant predictors. The teams value the all-rounders ability to make regular contributions with the bat but more importantly, their ability to bowl regularly and reliably in matches. The random effect leading to the best model was the number of 50plus scores in the last year. Thus, for all-rounders there is definitely a mix of performance measures from both roles, but with a skew towards batting. This suggests that batting ability is more valued than the bowling function by a team.

The MCMC simulations shift the focus towards a player's ability to score fast with bat and not let the opposition score off them when they bowl. From the fixed effects estimations we see that the most significant predictors are regular batting, fast scoring & bowling participation along with wicket taking capability of the bowler as the profiling criteria for all-rounders.

5.2 Method

The methodology adopted in this thesis study was to apply Mixed Models and show how this branch of statistics can be beneficial for analyzing longitudinal data in the context of sports, specifically cricket. The general approach designed at the start was to start by establishing a baseline by using the widely accepted method in analyzing cricket data i.e. classical linear models. As outlined in the theory section 2.1.1, multiple papers have used linear models as their prediction choice. My approach here was to use this as a baseline and then gravitate towards the more complex Mixed Models.

Since, we start with a variety of predictors for both batting and bowling roles, to get better results, it was decided to simplify the models by modelling each role separately. This makes sense intuitively since a batsman is only evaluated on his batting performance and not on his bowling even if he may have bowled in some matches and even taken a few wickets in between. The important point here is that a player's evaluation will always be based on his relevant skills. As such, it was decided to tackle each role separately. The batsmen would be modelled only based on batting indices and the bowlers only on the bowling indices. For all-rounders we would need both sets of indices as they are evaluated on both metrics. It was considered to forgo this step and model all roles together but then we identified the first problem which is that pure batsmen and bowlers will not have any data of their secondary role i.e. batsmen who have never bowled and vice versa. This led to the model having to deal with a lot of missing values that cannot be imputed. Thus, it made sense to work with all the roles separately.

The linear baseline models give us a start to identify significant predictors. We discussed variable selection at length in section 3.3.3 and saw the benefits in identifying significant predictors across model classes in the results chapter. This strategy of dropping insignificant features works well enough for linear models, but when we move to mixed modelling, more caution needs to be exercised. This is because in a mixed model, the predictors' combined significance changes due to our choice of fixed and random effects. A LMM with a certain random effect structure can report a different set of significant fixed effects than another random effect choice. Hence in the experiments, it was important to keep rotating between fixed effects and random effects and not discarding the fixed effects arbitrarily based on the output of a different LMM. The final models that were reported in the results section are a result of multiple experiments with different structures of fixed and random effects that resulted in the lowest prediction errors and also reported appropriate IC values.

5.3 The work in a wider context

When this study was undertaken, it was highlighted that cricket remains an understudied sport as compared to some mainstream sports like football, NBA, NFL etc. To repeat, most

of the existing literature and study revolves around trivial problem statements that don't employ very scientifically challenging methodologies. This study has sought to add to the studies surrounding cricket by tackling the prediction of IPL player auction prices in the following two ways:

1. *Mixed Models* are very good at handling panel data. By bringing this class of models to cricket data, it is a deliberate effort to update the way cricket data is analyzed. This is significant because it is not only relevant to this particular problem statement. Other academic studies such as identifying key performance measures for a player, or identifying which players justify their salary, which are common research questions in sports analytics will benefit from this approach as game data for players and teams will mostly be longitudinal.

2. *Data depth* is a term I would coin here to talk about the choice we make in using relevant data for a prediction exercise such as this. Most studies we came across only deal with current season data and/or data from a particular competition or season. We haven't come across examples of researchers choosing to combine data from multiple competitions to their model input. This was strange because it seems intuitive that a player's worth and performance are a sum of their current and past form. It would be incomplete to say that a player is only valued for how they have performed in terms of numbers alone. Take for example the case of an Australian player named Cameron Green. He is a young player who has recently arrived onto the international scene and is lauded by experts to be a bright prospect for the future. Statistically, this player has very little game data in the T20 cricket format. And yet, he was one of the highest bids in the IPL auction for 2023 season. In the scope of this study, it is not feasible to include subjective factors such as popularity and expert opinions or internet chatter for that matter. But these are still relevant inputs to how players are valued.

   The effort made in this thesis to address this issue somewhat is that we have taken game data from other competitions across the year where these player participate. This is done in the anticipation that with richer inputs, we are identifying more meaningful relationships between the prediction variable and the predictors. We have achieved this by combining the game statistics from various competitions into an aggregate annual performance index which we use as input alongside the historical performance indices of the player in the previous seasons of the IPL. But incorporating qualitative input such as expert opinions can be a very good starting point for an independent study into player valuations.

5.4 Limitations

One of the most important limitation of this approach has already been talked about, which is the incapability of this study to include subjective inputs from experts and internet buzz. Another key limitation to think about is that we have combined the player performance data from various competitions into an aggregate metric. This aggregation may not be ideal for the following reasons:

1. The quality of competition varies drastically. Some T20 leagues like The Big Bash League in Australia are fiercely competitive while others like the Lanka Premier League or the Bangladesh Premier League might not be so. An illustration of this came around the year 20109-2015 when cricket leagues emulated the Champions League format of football by getting top teams from various T20 leagues together for a competition. The results can be argued upon, but 4 out of 6 times the eventual winners were an IPL team. So, a player's performance in a B-tier competition may not be a fair indication of their capability in a A-tier competition like the IPL.

2. The playing conditions in cricket matter and vary a lot. One of the most interesting things about cricket is that each country has its own unique set of challenges to the players in terms of the playing conditions that it offers. Traditionally, South Africa, England, New Zealand and Australia (collectively called SENA countries in cricket) offer playing surfaces that assist pace bowlers. In contrast, playing conditions in India are much more suitable for spinners and batsmen. These conditions are another variable that is very difficult to account for. Its another reason why combining performance stats from different playing conditions can offer misleading conclusions about a player's ability.

At various points during the study, we pointed out the presence of outliers in the data and heteroscedasticity in the analysis. While attempts were made by avoiding correlations and dropping insignificant features, we haven't been able to properly address outliers in this study. Part of the reason lies in the insufficiency of data. Because of the decided approach to split the data by player roles, we have very limited training data for modelling. This can lead to problems like overfitting and some model assumptions being violated like we witnessed. A known way of addressing heteroscedasticity is by transformation of predictor variables. This is a time consuming exercise that was consciously ignored to favour the completion of the study in the restricted time frame. But it is also presents an opportunity to formulate and pursue an independent study in this direction with the aim of improving the prediction accuracy of the models.

The current model flexibility only extends to basic transformations of the raw performance data and then using this transformed data into mixed models. An unexplored direction for this task is the transformation of player performance data to explore higher order polynomial relationships to account of non-linear relationships of the predictors with the auction price. Another class of modelling that promises good results is General Additive Models and Splines.

Lastly, we would like to point out the inherent challenges of analyzing MCMC GLMMs. The combination of fixed and random effect interactions increases the model complexity manifold and it is very hard to deconstruct these models and derive statistical inference from them. Analyzing MCMC convergence is a tough area as it is and thus it is impressed upon the user to have a very strong understanding of longitudinal data and its modelling choices before being dissect the statistical inferences from the sparsity involved in MCMC simulations of GLMMs.

5.5 Ethical Considerations

The ethical aspects of this study are not straightforward but still relevant. As an academic study, the proposed method and results can have some unforeseen consequences such as:

1. Fairness and discrimination: The prediction model should not be based on discriminatory factors such as race, gender, or nationality. The model should be designed in such a way that it treats all players equally and does not discriminate against any particular group. While it is not clear, but if there are any existing biases that influence how players from a certain background maybe treated currently in their pricing evaluations by teams, then these biases would be a part of the model. Say for example, if players from an associate nation i.e. a nation that does not have the status of playing test cricket like Ireland, then even a highly talented player from these countries might be getting a lower salary.

2. Impact on player salaries and career: The study's results may have significant implications for player salaries and career prospects. If the study's predictions are inaccurate, they could harm players' future earnings and career opportunities. The results we have posted in this study have also shown that some predictions are much less accurate than

others. Without proper due diligence of the results, a player's auction value can be grossly misinterpreted which can result in a team not managing to obtain the services of the player through the auction because of an incorrect bid, or even that the player gets ignored in the auction totally by all the teams.

3. Misuse of the study findings: The study's findings may be misused by individuals or organizations to manipulate player salaries or gain an unfair advantage in the IPL auction. It is essential to ensure that the study's findings are not used to harm players or violate the principles of fair play in the sport. IPL is also no stranger to controversy and match fixing. The results can also be used to target players by bookies to create unlawful and unethical events that would lead the game into disrepute.

6 Conclusion

In this thesis we set out to explore a previously unexplored modelling technique in the context of predictive modelling in cricket. It was discussed and established by quoting several studies that there is a vacuum in the field of cricket analytics of applying appropriate methods in dealing with longitudinal player performance data. We proposed to use the widely used classical linear models as baseline methods and then challenge the outputs using a more robust alternative - Mixed Models.

At the start we had outlined the following key questions to be answered in the study:

1. *What are the best predictors for modelling a player's auction value?*
2. *What is the best modelling choice for predicting the auction value against performance?*

To answer the first question, we have for each player role identified the most important predictors and also identified the best random effect structures. In each case, we have provided an insightful view on which predictors from the category of batting and bowling ability are considered most important in the valuation of each player role. For all the roles, a player's status as a star player, international level player and an overseas player proved to be important features. For Batsmen specifically, we showed that their overall batting experience and their ability to score fast along with being reliable is what the teams look for most in batsman. For bowlers, we found that, they are expected to be workhorses who can reliably bowl without too many injuries that might hamper their ability to play the whole season. We saw evidence that batting ability is considered to be the more favoured skill for IPL teams. This was backed up when we saw the results for all-rounders. While they are valued for both their batting and bowling, their batting skills are more in focus.

To answer the second and arguably the more important research question, We have explored mixed models in depth in this study. We made use of the uncorrelated intercept and random slope structures and avoided the nested random effect structures for keeping the models simple and to avoid dealing with sparse covariance structures. Variable selection was done using popular scientific methods and dodel evaluation and selection was done by observing both prediction error estimates and the information criteria. For the frequentist models, we have made use of the AIC & BIC scores while for the Bayesian framework MCMC GLMMs we have used the DIC. For prediction errors we chose to employ MAPE alongside RMSE to have an intuitive understanding of the errors and the prediction quality.

In this study, we have for each role, managed to improve our prediction accuracy over the baseline models with LMMs and then attempted to improve them further by deploying MCMC GLMMs. Barring the batsmen role, for both bowlers and all-rounders the LMMs reported lower error results than MCMC simulations. But, the performance of the best performing MCMC models was quite similar to the best performing LMMs. This can be possibly due to the stochastic nature of the MCMC methods themselves as the results different each time we train the models.

Overall, we have shown that mixed models are a favourable alternative to the common approach of using simple linear regression methods for cricket data. We have tested numerous configurations of fixed and random effects to present very intuitive results which make it easy to decipher the prediction criteria for each player role.

6.1 Future Work

At various points in this manuscript, the limitations of the study design and data insufficiency issues have been discussed and pointed out in depth. The IPL is a relatively young and small league compared to leagues from other sports such as EPL, NFL, NBA. The combination of volatility during the initial years of the IPL being launched and the limited player pool in cricket and the volume of data available for this study makes it very difficult to take full advantage of a robust mechanism such as mixed models which have a well documented history of being very powerful tools in modelling panel data. Typically, mixed models work very well with very large volumes of data and have been widely employed in biomedical research. With the right settings, they can very easily outperform modern neural networks as well. As the field of sports analytics keeps growing and the volume of data keeps improving, we expect more studies to fine tune this approach and improve the prediction results even further.

A totally fresh and novel approach in tackling this problem would be to approach it from an Auction theory perspective. A very important perspective that we have not talked about at all in this study is that the team managements themselves have their own prior knowledge about a player that is unique and different from other teams. Each team decides their auction strategy keeping in mind their current team composition and their budget constraints. Each team has their own army of scouts that come with their own personalized view on each player. This explanation is far from sufficient in describing the true complexity behind auction theory that labels auctions as Bayesian games of incomplete information. A study from this perspective would a completely novel approach in the sports analytics field as a whole not just for cricket.

Bibliography

[1] H. Akaike. "A new look at the statistical model identification". In: *IEEE Transactions on Automatic Control* 19.6 (1974), pp. 716–723. DOI: 10.1109/TAC.1974.1100705.

[2] J. Albert, M.E. Glickman, T.B. Swartz, and R.H. Koning. *Handbook of Statistical Methods and Analyses in Sports*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, 2017. ISBN: 9781351678964. URL: https://books.google.se/books?id=%5C_ZcnDwAAQBAJ.

[3] Sambit Bal, ed. *Cricketers - Player Lists, Stats, Videos, Photos*. URL: https://www.espncricinfo.com/cricketers.

[4] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1 (2015), pp. 1–48. DOI: 10.18637/jss.v067.i01. URL: https://www.jstatsoft.org/index.php/jss/article/view/v067i01.

[5] J. Bruin. *newtest: command to compute new test @ONLINE*. Feb. 2011. URL: https://stats.oarc.ucla.edu/other/mult-pkg/introduction-to-linear-mixed-models/#:~:text=Linear%5C%20mixed%5C%20models%5C%20are%5C%20an,or%5C%20patients%5C%20from%5C%20within%5C%20doctors..

[6] Nayan Ranjan Das, Ratna Priya, Imon Mukherjee, and Goutam Paul. "Modified Hedonic Based Price Prediction Model for Players in IPL Auction". In: *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. 2021, pp. 1–7. DOI: 10.1109/ICCCNT51525.2021.9580108.

[7] Jack Davis, Harsha Perera, and Tim B. Swartz. "Player evaluation in Twenty20 cricket". In: *Journal of Systems Architecture* 1 (2015), pp. 19–31.

[8] Chaitanya Deep, Chellapilla Patvardhan, and Sushobhit Singh. "A new Machine Learning based Deep Performance Index for Ranking IPL T20 Cricketers". In: *International Journal of Computer Applications* 137 (2016), pp. 42–49.

[9] Chellapilla Deep Prakash, C. Patvardhan, and Sushobhit Singh. "A new Machine Learning based Deep Performance Index for Ranking IPL T20 Cricketers". In: *International Journal of Computer Applications* 137 (Mar. 2016), pp. 42–49. DOI: 10.5120/ijca2016908903.

[10] P. Diggle, P. Heagerty, K.Y. Liang, and S. Zeger. *Analysis of Longitudinal Data*. Oxford Statistical Science Series. OUP Oxford, 2013. ISBN: 9780199676750. URL: https://books.google.se/books?id=ur0BlXPuOukC.

[11] Nicholas Galwey. "Introduction to Mixed Modelling: Beyond Regression and Analysis of Variance". In: *Introduction to Mixed Modelling: Beyond Regression and Analysis of Variance: 2nd Edition* (Aug. 2014), pp. 213–214. DOI: 10.1002/9780470035986.

[12] L. A. Garcia-Cortes and D Sorensen. "Alternative implementations of Monte Carlo EM algorithms for likelihood inferences". In: *Genetics, Selection, Evolution : GSE* 33 (2001), pp. 443–452.

[13] Andrew Gelman, Jessica Hwang, and Aki Vehtari. *Understanding predictive information criteria for Bayesian models*. 2013. arXiv: 1307.5928 [stat.ME].

[14] Andrew Gelman and Eric Loken. "The Statistical Crisis in Science". In: *American Scientist* 102 (Nov. 2014), p. 460. DOI: 10.1511/2014.111.460.

[15] Jarrod D. Hadfield. "MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package". In: *Journal of Statistical Software* 33.2 (2010), pp. 1–22. DOI: 10.18637/jss.v033.i02. URL: https://www.jstatsoft.org/index.php/jss/article/view/v033i02.

[16] Matt Harris. *Rules of Cricket*. URL: https://www.itsonlycricket.com/rules-of-cricket.

[17] Matt Harris. *The Cricket Field*. URL: https://www.itsonlycricket.com/cricket-pitch-or-field.

[18] Rob J. Hyndman and Anne B. Koehler. "Another look at measures of forecast accuracy". In: *International Journal of Forecasting* 22.4 (2006), pp. 679–688. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2006.03.001. URL: https://www.sciencedirect.com/science/article/pii/S0169207006000239.

[19] International Cricket Council ICC. *History of cricket*. URL: https://www.icc-cricket.com/about/cricket/history-of-cricket/early-cricket.

[20] Scott Irvine and Rodney Kennedy. "Analysis of performance indicators that most significantly affect International Twenty20 cricket". In: *International Journal of Performance Analysis in Sport* 17.3 (2017), pp. 350–359. DOI: 10.1080/24748668.2017.1343989. eprint: https://doi.org/10.1080/24748668.2017.1343989. URL: https://doi.org/10.1080/24748668.2017.1343989.

[21] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. "Linear Model Selection and Regularization". In: *An Introduction to Statistical Learning: with Applications in R*. New York, NY: Springer US, 2021, pp. 225–288. ISBN: 978-1-0716-1418-1. DOI: 10.1007/978-1-0716-1418-1_6. URL: https://doi.org/10.1007/978-1-0716-1418-1_6.

[22] Had Jarrod. "MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package". In: *Journal of Statistical Software* 33 (Feb. 2010).

[23] Ajit Karnik. "Valuing Cricketers Using Hedonic Price Models". In: *Journal of Sports Economics* 11.4 (2010), pp. 456–469. DOI: 10.1177/1527002509350442. eprint: https://doi.org/10.1177/1527002509350442. URL: https://doi.org/10.1177/1527002509350442.

[24] C. G. Khatri and C. Radhakrishna Rao. "SOLUTIONS TO SOME FUNCTIONAL EQUATIONS AND THEIR APPLICATIONS TO CHARACTERIZATION OF PROBABILITY DISTRIBUTIONS". In: 2016.

[25] Indian Premier League. *IPL T20 Player Auctions*. URL: https://www.iplt20.com/auction.

[26] Gaurav Malhotra. "A comprehensive approach to predict auction prices and economic value creation of cricketers in the Indian Premier League (IPL)". In: *Journal of Intelligent Fuzzy Systems* 8 (July 2022), pp. 1–22. DOI: 10.3233/JSA-200580.

[27] Akshay Ramesh. *IPL world's second-most valued sporting league in terms of per match value, digital rights costiler than TV*. June 2022. URL: https://www.indiatoday.in/sports/cricket/story/ipl-media-rights-digital-value-more-than-tv-2nd-most-valuable-jay-shah-1962410-2022-06-14.

[28] Stephen Rushe. URL: https://cricsheet.org/matches/.

[29] Hemanta Saikia, Dibyojyoti Bhattacharjee, and H. Hermanus Lemmer. "Predicting the Performance of Bowlers in IPL: An Application of Artificial Neural Network". In: *International Journal of Performance Analysis in Sport* 12.1 (2012), pp. 75–89. DOI: 10.1080/24748668.2012.11868584. eprint: https://doi.org/10.1080/24748668.2012.11868584. URL: https://doi.org/10.1080/24748668.2012.11868584.

[30] Srikanth Sankaran. "Comparing Pay versus Performance of IPL Bowlers: An application of Cluster Analysis". In: *International Journal of Performance Analysis in Sport* 14.1 (2014), pp. 174–187. DOI: 10.1080/24748668.2014.11868713. eprint: https://doi.org/10.1080/24748668.2014.11868713. URL: https://doi.org/10.1080/24748668.2014.11868713.

[31] Gideon Schwarz. "Estimating the Dimension of a Model". In: *The Annals of Statistics* 6.2 (1978), pp. 461–464. DOI: 10.1214/aos/1176344136. URL: https://doi.org/10.1214/aos/1176344136.

[32] Tyler Stanek. "Player Performance and Team Revenues: NBA Player Salary Analysis". In: 2016.

[33] Wikipedia contributors. *English auction — Wikipedia, The Free Encyclopedia*. [Online; accessed 4-May-2023]. 2022. URL: https://en.wikipedia.org/w/index.php?title=English_auction&oldid=1105726478.