Dream11 Report - Team 84

Abstract

Fantasy sports have revolutionized fan engagement by blending passion with strategy, enabling users to create virtual teams and earn points based on players' real-world performances. This research focuses on enhancing user experience on Dream11, India's largest fantasy sports platform, through advanced predictive analytics and machine learning. A robust ensemble of algorithms, including XGBoost, CatBoost, Random Forest, and K-Nearest Neighbors (KNN), was employed to predict player fantasy points. Separate regression models tailored to T20, ODI, and Test cricket formats were developed, accounting for batting and bowling variations.To refine predictions, a probability-based model was introduced, integrating context-specific factors such as match conditions and player roles. Additionally, a knowledge graph captured head-to-head matchups, enriching insights with batter-bowler interaction metrics. By incorporating historical performance trends and contextual match variables, our model offers actionable recommendations, empowering users to make strategic, data-driven decisions.This research demonstrates the power of machine learning in addressing the challenges of creating optimal fantasy teams while ensuring feature explainability and user-friendly insights. The findings aim to bridge the gap between raw data and user understanding, enhancing the inclusivity and competitiveness of the fantasy sports ecosystem.

Keywords: XGBoost, CatBoost, Model Stacking, Polynomial Regressor, GenAI, Transformers

1 Introduction

Fantasy sports have changed the way fans connect with their favorite games, combining passion and strategy for a more interactive experience. Dream11, India's largest fantasy sports platform, leads this transformation. Launched in 2008 by Harsh Jain and Bhavit Sheth, Dream11 is a key brand of Dream Sports, with over 220 million users creating virtual teams for cricket, football, kabaddi, and more. Users earn points based on the real-life performances of players. Cricket, deeply cherished in India, stands out on Dream11, where fans showcase their knowledge by building teams that reflect their understanding of the game. Powered by advanced technologies like AI and machine learning, the platform ensures a smooth and engaging experience. Recognized by the Supreme Court of India as a "game of skill," Dream11 not only complies with legal standards but also encourages users to explore sports more strategically. By turning every match into an opportunity to play, predict, and win, Dream11 makes sports fandom more interactive and rewarding.

The core challenge for Dream11 lies in enhancing user engagement by improving the decision-making process for creating winning fantasy sports teams. Despite providing extensive player statistics, performance measures, and expert-curated recommendations, many users still find it challenging to make informed team-building choices, particularly those less familiar with advanced analytics. To address this, a robust machine learning (ML) model is needed to predict player performance based on historical data, contextual match factors, and game-relevant variables. This predictive model aims to empower users with actionable insights, enabling them to make strategic, data-driven team selections. Coupled with detailed feature explainability, such a solution would bridge the gap between raw data and user understanding, ensuring every player recommendation is transparent and meaningful.

Beyond predictive analytics, providing an intuitive and user-friendly interface is crucial to improving the team-building experience on Dream11. The solution should present the model's insights clearly, enabling users to understand player recommendations and make data-driven decisions with ease. By integrating tools for visualizing performance metrics and contextual factors such as match conditions, users can navigate the complexities of team creation more effectively. This streamlined approach ensures that even users with limited analytical expertise can participate confidently while catering to the needs of experienced users seeking advanced insights. Such enhancements align with Dream11's goal of creating an inclusive platform that offers strategic value, boosts user satisfaction, and reinforces its competitive edge in the dynamic fantasy sports market.

We employed advanced machine learning techniques, including XGBoost, CatBoost, Random Forest, and K-Nearest Neighbors (KNN), to predict Dream11 fantasy points using features like player performance metrics, venue data, and historical trends. Model stacking was implemented to combine predictions and enhance accuracy. Separate regression models for batting and bowling in T20, ODI, and Test formats addressed scoring differences, weighted by probabilities of a player batting or bowling. A knowledge graph captured head-to-head batter-bowler interactions, enriched with metrics like strike rate and dismissal rates. Integrating historical and recent performance metrics ensured robust, role-specific, and format-aware predictions.

2 Literature Survey

Sports analytics has transformed decision-making in many major sports, such as Baseball, Football, Basketball, Soccer, and Tennis, by leveraging data to optimize team strategies and player performance. However, Cricket, despite its rich historical data, has only recently begun to explore the power of analytics. This literature survey examines the evolution of sports analytics, its application in Cricket, and the key research developments that have shaped the current state of Cricket analytics. By drawing parallels with the advancements in other sports, particularly Baseball, this survey provides an overview of how data-driven approaches have influenced Cricket, with a focus on player performance prediction, team selection, and match outcome forecasting.

2.1 The Evolution of Sports Analytics

The roots of modern sports analytics can be traced to Michael Lewis's 2003 book Moneyball: The Art of Winning an Unfair Game, which chronicled the story of Billy Beane, the General Manager of the Oakland Athletics baseball team. Faced with limited resources, Beane turned to Sabermetrics, a statistical analysis method that focused on undervalued player metrics like on-base percentage and slugging percentage. This approach disrupted traditional scouting methods and led to a paradigm shift in player selection and team strategies across various sports. The success of Moneyball inspired other sports, such as Football and Basketball, to adopt similar analytical techniques.

Cricket, despite its deep historical roots with data dating back to 1697, was slower to embrace analytics. Unlike Baseball, with its discrete events and predictable structures, Cricket's complexity—stemming from diverse formats (ODI, T20, Test), varied playing conditions (pitch type, weather), and intricate player roles—presented challenges for applying data-driven approaches. However, the increasing popularity of Cricket, particularly in nations like India, Australia, and England, provided a fertile ground for the gradual introduction of analytics in the sport.

2.2 Early Adoption and Slow Rise of Cricket Analytics

Cricket's statistical data, though available for centuries, was initially underutilized for detailed performance analysis. The sport's complex dynamics, such as varying pitch conditions, diverse playing styles, and the impact of weather, made it difficult to implement straightforward predictive models. Early efforts in Cricket analytics focused on basic performance statistics, such as batting averages and bowling economy rates. However, these simplistic approaches did not match the granularity and predictive power seen in Baseball or other team sports.

In the 1990s and early 2000s, Cricket began to see more substantial efforts to apply data analysis to player and team performance. These early studies primarily aimed at understanding how various factors—such as batting position, bowling style, and playing conditions—affected overall performance. As Cricket's financial stakes increased, the potential for analytics to enhance match strategy and team selection became more apparent.

2.3 Key Contributions to Cricket Analytics

Significant studies have shaped the landscape of Cricket analytics. A notable early contribution was the use of machine learning techniques to predict player performance in One-Day Internationals (ODIs). For example, the paper Increased Prediction Accuracy in the Game of Cricket using Machine Learning demonstrated how machine learning algorithms, including Naïve Bayes, Random Forest, Support Vector Machines (SVM), and Decision Trees, could be employed to predict player performance. The Random Forest algorithm, in particular, showed the highest accuracy, becoming a key tool for player performance prediction.

Further advancements were made by researchers like Muthuswamy and Lam (2008), who employed neural networks to predict Indian bowlers' performance, accounting for match conditions, fatigue, and player form. This marked a move toward more sophisticated

models that incorporated the nuances of the game, such as player fatigue and environmental factors, which traditional statistical methods struggled to capture.

Other studies, such as A Machine Learning Approach to Analyze ODI Cricket Predictors, emphasized the importance of unpredictable factors like coin toss results, venue conditions, and player injuries. These studies underscored the need for more comprehensive models that could integrate these unpredictable elements for accurate match outcome predictions.

Additionally, the use of machine learning for team selection has also been explored. Bananki et al. introduced the use of SVM classifiers to predict match outcomes based on the categorization of players, allowing teams to optimize player composition and maximize their chances of winning. Such classification approaches helped balance team strength, which is crucial in Cricket, where team dynamics play a significant role in match outcomes.

2.4 Team Selection and Strategy in Cricket

Team selection, a critical aspect of Cricket, influences the outcome of matches. Research by Singla et al. and Agrawal et al. explored optimization algorithms for selecting the best team lineup, taking into account factors such as batting and bowling averages, fitness levels, and historical performance against specific teams. These studies showed how data-driven decision-making could enhance team management by incorporating a variety of performance metrics.

Regression models have also played a vital role in Cricket analytics. For example, Rodrigues et al. used regression analysis to predict continuous variables like runs scored or wickets taken, considering factors such as opposition strength and pitch conditions. This enabled the creation of models that could more accurately forecast player performance than traditional methods.

Saikia et al. expanded Cricket analytics by evaluating fielding performance—an often overlooked aspect of the game. Their work utilized composite indices based on match scorecards to assess fielding effectiveness, offering a more holistic approach to player evaluation.

2.5 Current Trends in Cricket Analytics

The use of analytics tools like Yorkpy [27] has significantly contributed to the ease of performing Cricket analytics. Yorkpy assists in transforming raw match data into usable features, facilitating tasks like data preprocessing, feature engineering, and model training. This tool, alongside other software platforms, has enabled researchers to streamline the process of building predictive models, making it easier to explore vast datasets and extract meaningful insights.

Despite these advancements, Cricket analytics faces several challenges. The complex nature of the sport, which involves numerous external variables such as weather, pitch conditions, player fatigue, and home-ground advantage, makes accurate prediction models difficult to build. These variables are often unpredictable, adding layers of uncertainty to match outcomes. Furthermore, the integration of different data sources, such as historical player performance, real-time match data, and environmental factors, presents significant challenges in terms of data preprocessing and model training.

The application of sports analytics in Cricket has grown significantly over the past few decades, with increasing use of machine learning, regression analysis, and optimization algorithms. These data-driven approaches have led to more informed decisions regarding player performance, team selection, and match predictions. However, the unpredictable nature of Cricket, with variables like weather and pitch conditions, continues to pose challenges in building accurate models. As technology advances and more data becomes available, Cricket analytics is poised to further refine its predictive capabilities, enhancing the sport's strategic decision-making and performance optimization.

3 Data

The dataset provided comprised ball-by-ball cricket match data, encompassing detailed event-level information for numerous matches. To efficiently process and analyze this extensive dataset, the data was structured and ingested into an SQLite3 database. Match details and player-specific statistics for all matches were systematically stored within this database to facilitate streamlined querying and data management.

SQL queries were employed to extract insights, execute complex queries, and generate refined datasets tailored for various analyses. Additionally, CSV files were generated from these queries, enabling their use in machine learning model training and subsequent statistical evaluations. This systematic preprocessing ensured optimal data organization and accessibility for further exploration and analysis.

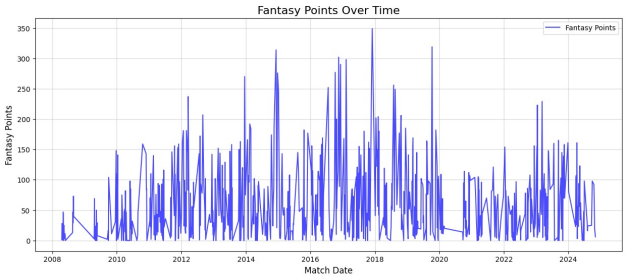

Figure 1. Fantasy Points of V Kohli over the Years

During the data analysis phase, we focused on mapping the fantasy points of players for each match. The resulting graph exhibited rapid fluctuations in fantasy points across matches. However, these variations remained within consistent ranges over specific segments of time, resembling patterns observed in stock market analysis. To model these trends effectively, we experimented with various regression techniques. Our aim was to capture the temporal consistency and abrupt changes accurately, enabling better predictions and insights into player performance trends over time. This approach laid the foundation for selecting and fine-tuning the appropriate regression models.

4 Methodology

We started by employing various machine learning models to predict fantasy points for players based on their recent and historical performance metrics. These models were chosen for their ability to handle complex data and deliver reliable predictions. Below, we detail the models used in this study:

A) **XGBoost:** Known for its speed and reliability, XGBoost minimizes overfitting through regularization by optimizing a loss function that includes a complexity term. This feature enables it to effectively capture non-linear player performance patterns, making it well-suited for this prediction task.

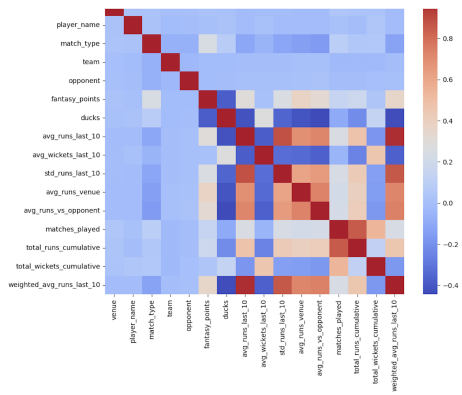

Figure 2. Correlation among Features

B) **CatBoost:** CatBoost employs gradient-boosted decision trees to reduce prediction bias, especially for categorical data such as player-specific and venue-specific statistics. It utilizes a greedy strategy and target-based encoding to improve accuracy and mitigate overfitting, ensuring better performance on diverse data.

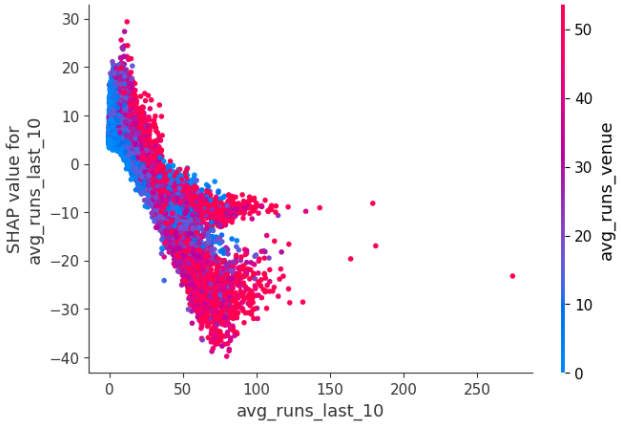

Figure 3. SHAP Dependence Plot

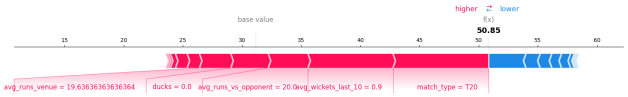

Figure 4. SHAP Force Plot

C) **Random Forest:** This robust ensemble method uses randomly selected features for splitting nodes at each decision point. This randomness reduces overfitting and variance, allowing the model to capture diverse performance metrics, including recent form and venue-specific conditions, effectively.

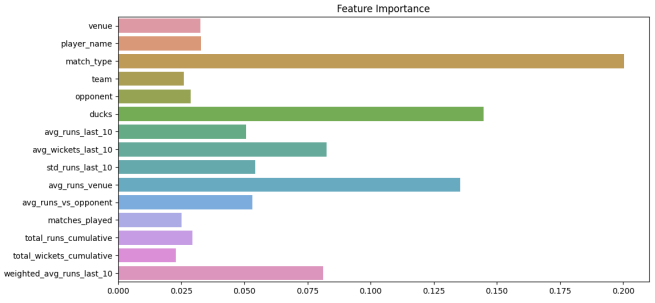

Figure 5. Feature Importance Map

D) **K-Nearest Neighbors (KNN):** KNN is a non-parametric, distance-based algorithm that predicts outcomes by identifying the closest data points (neighbors) in the feature space. For predicting fantasy points, it leverages recent performance metrics, venue-specific data, and

overall statistics of players to calculate similarity. The algorithm assigns predictions based on the average fantasy points of the nearest neighbors, providing a simple yet interpretable baseline for comparison with more complex models.

4.1 Model Stacking

To enhance the prediction accuracy and leverage the strengths of multiple machine learning algorithms, we implemented a model stacking approach. Model stacking is an ensemble learning technique that combines the predictions of multiple base models to produce a final prediction. The key idea behind stacking is to use a meta-model to learn from the predictions of base models and select the best-performing model or combination of models.

In our study, the base models included XGBoost, CatBoost, Random Forest, and K-Nearest Neighbors (KNN). Each of these models was trained independently on the same dataset, utilizing features such as recent player performance metrics, venue-specific statistics, and overall historical data. Their predictions were then fed into a meta-model, which was tasked with identifying the most reliable predictions from the base models.

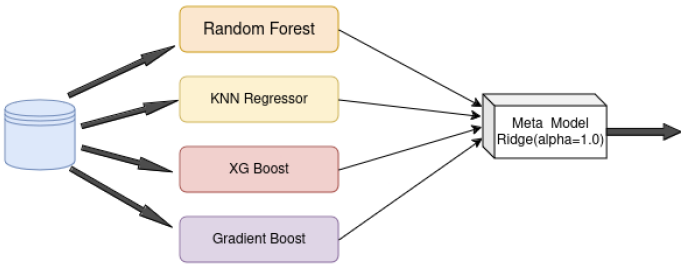

Figure 6. Stacking Model Architecture

The stacking approach allowed us to address individual model limitations, such as overfitting or bias towards certain feature distributions, by combining their outputs. This ensemble strategy proved effective in capturing complex relationships in the data and provided a robust framework for predicting fantasy points. The meta-model, typically a simple algorithm like logistic regression or a weighted averaging scheme, was optimized to ensure minimal error and maximum predictive performance.

4.2 Probability Model

To further refine our prediction of fantasy points, we incorporated a probability-based approach. This method recognizes that players may not always bat or bowl in a match, and hence assigning probabilities to these actions provides a more realistic measure of their contribution. Additionally, we trained separate models for each format—T20, ODI, and Test—as the fantasy point scoring systems vary significantly across these formats. A single model for all formats might fail to distinguish these differences effectively.

4.2.1 Model Training and Selection

Based on the features identified earlier, we trained six separate polynomial regression models: three for batting (one for each format) and three for bowling. This separation allows the models to learn format-specific performance trends effectively. During prediction, the appropriate batting and bowling models are selected based on the match format.

4.2.2 Incorporating Probabilities

We calculated the probabilities of a player batting (P_{batting}) and bowling (P_{bowling}) in each format. These probabilities were then used to weight the predicted fantasy points from the respective models. The final predicted fantasy points (FP) for a player in a match are calculated as:

$$FP = P_{\text{batting}} \cdot \text{Predicted Points from Batting Model} + P_{\text{bowling}} \cdot \text{Predicted Points from Bowling Model.}$$

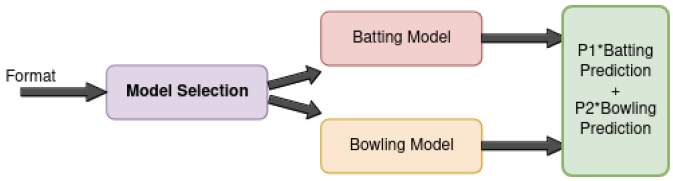

Figure 7. Model M1

4.2.3 Illustrative Examples

This approach dynamically adjusts to player roles:

- **Virat Kohli:** As a specialist batsman, his P_{batting} is close to 1, while P_{bowling} is near 0. This results in his fantasy points being almost entirely derived from the batting model.

- **Jasprit Bumrah:** For a specialist bowler, P_{bowling} will be close to 1, and P_{batting} near 0, making his bowling performance the primary contributor to his fantasy points.

- **All-Rounders:** For players like Ben Stokes or Ravindra Jadeja, both probabilities will have significant values,

resulting in a combination of batting and bowling points contributing to their total fantasy points.

4.2.4 Analysis of the Approach

This probability-based approach adds flexibility and realism to the prediction system by tailoring predictions to a player's role and format. By training separate models for batting and bowling in each format, we address the differences in fantasy point scoring mechanisms across formats. Incorporating probabilities ensures that predictions remain realistic, aligning with players' actual contributions in a match. This method is particularly effective in handling the varied roles and contributions of players, making the prediction system robust and adaptable.

4.3 Head-to-Head Matchups

4.3.1 Understanding the importance of H2H Statistics

While the previous models provided robust predictions by analyzing historical performances of players, they inherently carry a limitation: they do not account for matchups between players in a live match scenario. Cricket is a dynamic sport where player performance is significantly influenced by the strength and skill of the opposition.

For instance, consider **Virat Kohli**, whose overall batting statistics are exemplary. If in a match, he faces a bowling attack comprising players like **Mitchell Starc**, **Jasprit Bumrah**, and other equally formidable bowlers, the probability of him scoring high fantasy points decreases. This is because his potential performance is directly constrained by the skill level of these bowlers.

Conversely, if another batter faces a team with weaker bowlers, their likelihood of scoring more runs and fantasy points increases, even if their overall historical stats are not as impressive as Kohli's.

This observation underscores the importance of incorporating **head-to-head matchups** into the prediction system. By analyzing player-specific interactions—such as Kohli's record against specific bowlers or the bowler's performance against top-order batsmen—we can refine predictions. A batter with favorable matchups against a weaker bowling attack is likely to contribute more significantly to a fantasy team than one with strong historical stats but a challenging opposition.

Integrating head-to-head matchup analysis allows us to make predictions that are closer to real-world outcomes, providing a strategic edge in selecting the optimal fantasy team.

4.3.2 Knowledge Graph

To incorporate head-to-head matchup analysis into our prediction system, we constructed a knowledge graph, where the **main node** represents the player for whom the prediction is being made, and the **neighboring nodes** represent the opponents in the match.

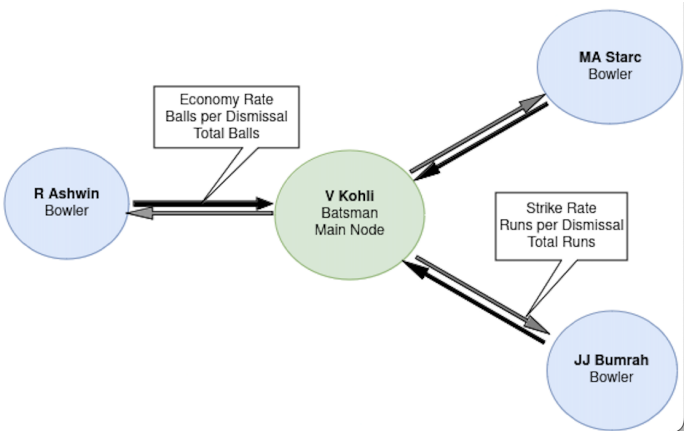

Figure 8. Knowlege Graph

- If the main node is a **Batsman**, the neighboring nodes are the **Bowlers** in the opposition team. - If the main node is a **Bowler**, the neighboring nodes are the **Batsmen** in the opposition team.

Edges in the Knowledge Graph The knowledge graph contains two types of edges, each carrying meaningful statistics:

1. **Batter to Bowler Edge**: This edge captures the interaction between a batter and a bowler. The characteristics of this edge are:

- **Runs per Dismissal:** Average runs scored by the batter before getting dismissed by the bowler.
- **Strike Rate:** Runs scored per 100 balls against the bowler.
- **Total Runs:** Total runs scored by the batter against the bowler.
- **Total Balls:** Total balls faced by the batter from the bowler.

2. **Bowler to Batter Edge**: This edge captures the interaction from the bowler's perspective and includes:

- **Dismissal Rate:** Frequency of dismissals by the bowler against the batter.
- **Economy Rate:** Average runs conceded by the bowler per over against the batter.

- **Balls per Dismissal:** Average number of balls bowled to dismiss the batter.
- **Total Dismissals:** Total number of times the bowler has dismissed the batter.

Graph Utilization By modeling the interactions between batsmen and bowlers as a knowledge graph, we gain a structured representation of the dynamic matchups that directly influence player performance. This allows for a more nuanced prediction model, where the strength of the edges provides critical insights into the expected outcomes of specific player interactions during the match.

4.3.3 Balancing Head-to-Head Stats with Recent Performance

While head-to-head matchups provide valuable insights into how players are likely to perform against specific opposition, it is essential to recognize that the predicted fantasy points based on these stats may not fully reflect a player's current form. For instance, **Virat Kohli**, despite being an exceptional batter with a strong overall record, may not always perform at his best against the most formidable bowlers. The head-to-head stats alone might suggest a lower performance in certain matchups, but this does not account for Kohli's ability to perform well even in challenging situations and his recent performances.

Therefore, it is crucial to combine the head-to-head analysis with the player's most recent performances. To achieve this, we included **actual fantasy points** from **Virat Kohli**'s last 10 matches as additional features. Specifically, we calculated the **mean** and **standard deviation** of his fantasy points over these recent matches to provide a more dynamic representation of his current form.

By incorporating both the historical head-to-head stats and the recent performance metrics, we ensure that the predictions are grounded in the player's current abilities while also accounting for the context of the upcoming match. This combination helps strike a balance between past performance and current form, improving the accuracy of the fantasy point predictions.

5 Performance and Knowledge Fusion Pipeline

The complete pipeline for predicting fantasy points involves multiple stages, each designed to capture a different aspect of player performance, both historical and real-time. This pipeline integrates predictions from the probability model, knowledge graph embeddings, and recent performance metrics to provide an accurate and dynamic prediction of fantasy points.

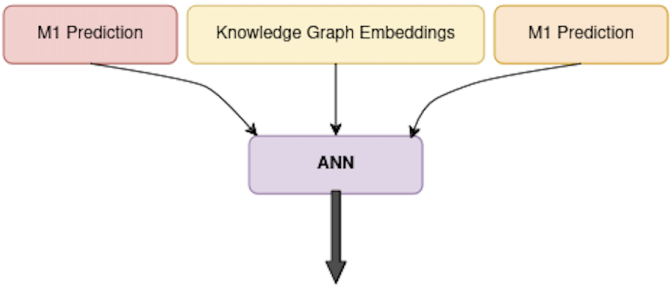

Figure 9. Performance and Knowledge Fusion Pipeline

5.1 Input Features

To generate the most reliable fantasy point predictions, we consider a variety of features that provide a comprehensive understanding of a player's potential in an upcoming match. The key input features include:

- **Probability Model Predictions:** We first use the predictions from our probability model, which incorporates player performance data from previous matches. The model accounts for the batting and bowling probabilities of a player in each format (T20, ODI, Test), and combines the predicted fantasy points based on these probabilities. This provides a base prediction that reflects the player's historical performance and format-specific tendencies.
- **Embeddings from the Knowledge Graph:** To capture the head-to-head matchups and player interactions, we utilize embeddings from the knowledge graph. The graph highlights the interactions between the batsman and the bowler (or vice versa), incorporating key statistics such as batting vs. bowling performance, dismissal rates, strike rates, and economy rates. The graph helps us model the effect of specific matchups, where, for example, a strong bowler can significantly impact a batter's predicted fantasy points.
- **Mean and Standard Deviation of Recent Fantasy Points:** To account for a player's current form and ability to perform under different match situations, we include the mean and standard deviation of their fantasy points from the last 10 matches. These features are particularly important for adjusting predictions, as they reflect the player's recent performance and adaptability in various conditions. A player like Virat Kohli,

who has demonstrated consistency but may also face tough opponents, will have his performance adjusted accordingly based on this dynamic feature.

5.2 Artificial Neural Network (ANN)

Once the input features have been compiled, they are fed into an Artificial Neural Network (ANN) for final prediction. The ANN is a powerful model capable of learning complex non-linear relationships between the features and the target variable (fantasy points). The network is trained on historical data to understand how the various input features interact and affect the overall fantasy points.

The output of the ANN is the predicted fantasy points for the player, which is used to inform the final selection for the Dream11 team. By learning from past performance data, head-to-head matchups, and recent form, the ANN generates a robust prediction that accurately reflects the player's potential performance in the upcoming match.

This final step in the pipeline ensures that the predictions are dynamic and account for the complex relationships between player attributes and match outcomes.

6 Results

Model	MAE	MPAE (%)
Complete Pipeline	120.48	20.13
Model Stacking	143.50	11.47
XGBoost	164.44	8.21
CatBoost	175.98	6.89
Linear Regression	192.91	3.34
Random Forest	178.78	6.72

Table 1. Comparison of Models Based on MAE and MPAE

6.1 Analysis of Results

The **Complete Pipeline** demonstrated the best performance, with the lowest MAE of 120.48 and an MPAE of 20.13%. This indicates the effectiveness of combining multiple methodologies—probability-based regression, knowledge graph embeddings, and recent performance statistics—integrated into an ANN. This holistic approach ensures the model captures both historical trends and contextual factors such as matchups and recent form, making it highly robust and versatile.

Model Stacking: Model stacking, which combines predictions from multiple base models (XGBoost, CatBoost, Random Forest, etc.), emerged as the second-best approach. It leverages the strengths of individual models while mitigating their weaknesses, resulting in an improved MAE of 143.50 compared to standalone methods. However, it falls short of the Complete Pipeline due to its inability to incorporate advanced features like knowledge graph embeddings and probability-based adjustments.

Standalone Models: Among the standalone models: - **XGBoost** and **CatBoost** performed well, with XGBoost achieving lower MAE and MPAE. These gradient-boosted tree models effectively handle structured data but lack the contextual insights provided by the Complete Pipeline. - **Random Forest** exhibited slightly higher MAE and MPAE, reflecting its limitations in capturing non-linear relationships as efficiently as gradient-boosted approaches. - **Linear Regression**, with the highest MAE (192.91) and the lowest MPAE (3.34%), highlighted the inability of simple linear models to handle complex feature interactions and non-linear dependencies inherent in player performance data.

6.2 Discussion

The **Complete Pipeline**'s superior performance underscores the importance of integrating diverse data sources and methodologies to capture the multifaceted nature of player performance. By combining historical stats, contextual match data, and recent performance metrics, it ensures a comprehensive understanding of the factors influencing fantasy points.

7 Future Enhancements

One promising direction for future work is the application of transformers in analyzing and predicting fantasy points. Initially, we experimented with a transformer model using basic numerical encoding for inputs. However, this approach yielded poor results, as it lacked the capacity to capture complex patterns in the data. By incorporating embeddings—a standard technique for transformers—we achieved a significant reduction in error, showcasing the model's improved ability to learn intricate relationships. However, the increased model size resulted in extremely long training times, with each epoch taking over three hours. This limited us to only a few epochs. Future efforts could focus on optimizing the transformer architecture or utilizing more advanced hardware to better exploit the model's

potential.

Another avenue for enhancement involves incorporating pitch type as a key feature to improve prediction accuracy. Current results indicate that accounting for pitch characteristics could significantly enhance the model's ability to capture match conditions. Additionally, clustering matches using techniques like K-means clustering based on pitch types—such as spin-friendly, pace-friendly, or balanced pitches—could help the model learn performance variations more effectively. This segmentation would allow for more tailored predictions, accounting for how different pitch types influence player performance.

Semantic analysis also offers a valuable opportunity to refine predictions further. By combining insights from web scraping and natural language processing, the model can analyze real-time data, such as news about player form, injuries, or team strategies. Additionally, weather conditions and the impact of powerplay overs can be integrated into the analysis. Weather factors like humidity, temperature, or rain probability can influence player performance, while powerplay overs often see higher scoring rates and strategic decisions that affect match outcomes. By incorporating these dimensions, the model would gain a comprehensive understanding of player dynamics, creating predictions that reflect real-world scenarios with greater accuracy.

8 Conclusion

This study highlights the complexities and opportunities in predicting fantasy points for cricket players, combining innovative approaches with robust analytical models. Beginning with a well-structured dataset of player performance, match conditions, and contextual factors, we evaluated multiple models to identify the most effective prediction pipeline. While initial attempts with transformers showed potential by reducing errors with embeddings, the approach proved computationally infeasible due to its high complexity, requiring over three hours per epoch and limiting further experimentation.

A comparative analysis of different models provided valuable insights into their performance. Among these, a complete pipeline integrating various techniques emerged as the most effective, achieving the lowest Mean Absolute Error (MAE) of 120.48 and a Mean Percentage Absolute Error (MPAE) of 20.13%. Other models, including stacking, XGBoost, and CatBoost, demonstrated competitive performance but were ultimately outperformed by the complete pipeline. Simpler models like linear regression, while computationally efficient, lacked the predictive accuracy necessary for this task.

This work not only showcases the potential of advanced machine learning in fantasy sports but also identifies avenues for improvement. Future efforts could refine predictions by incorporating additional contextual factors like pitch types, weather, and game phases (e.g., powerplay), and leveraging semantic analysis of current player news through web scraping. By addressing computational challenges and exploring these enhancements, this framework can evolve into a powerful tool, enabling fantasy cricket enthusiasts to make more informed decisions and enhancing their engagement with the game.

References

[1] S. Choudhari, N. Wagholikar, A. Swaminathan and S. Kurhade, "Dream11 IPL Team Recommendation using Machine Learning and Skill-Based Ranking of Players," 2022 International Conference for Advancement in Technology (ICONAT), Goa, India, 2022, pp. 1-6, doi: 10.1109/ICONAT53423.2022.9725819. keywords: Error analysis;Predictive models;Forecasting;Optimization;Random forests;Sports;Meteorology;Dream 11;catboost;fielding;skill-based ranking;linear optimization.

[2] S, Sachin HV, Prithvi Nandini, C. (2022). Data Science Approach to predict the winning Fantasy Cricket Team Dream 11 Fantasy Sports. 10.48550/arXiv.2209.06999.

[3] Dhanday, Sukhdayal Ranjan, Sandeep. (2024). Fantasy Sports Meets Reality: An Analytical Approach to Craft Probable Winning Dream11 Team Using IPL Data Insights.

[4] M. Balpande, K. Mahajan, J. Bhandarkar, B. Kapadne and G. Borse, "Machine Learning Based IPL Fantasy Cricket Dream11 Best Team Prediction," 2024 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2024, pp. 1-6, doi: 10.1109/ESCI59607.2024.10497335. keywords: Machine learning algorithms;Virtual groups;Forestry;Predictive models;Prediction algorithms;Boosting;Real-time systems;Dream 11;IPL;Machine Learning;XGBoost;CatBoost;Random Forest;ESPN CricInfo;team statistics;historical data;sports analytics.

[5] Gupta, Akhil. (2019). Time Series Modeling for Dream Team in Fantasy Premier League. 10.48550/arXiv.1909.12938.

[6] Singla, Saurav Shukla, Swapna. (2020). Integer

Optimisation for Dream 11 Cricket Team Selection. INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING. 8. 1-6. 10.26438/ijcse/v8i11.16.