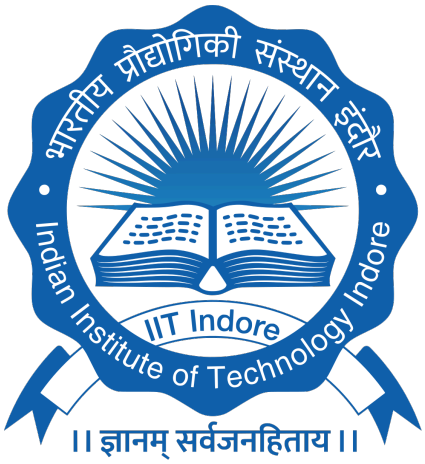
Indian Institute of Technology Indore

AI-Powered Dream11 Team Prediction System



IIT Indore

Inter IIT Tech Meet 13.0
Dream11 Mid-Prep Challenge

October 2025

Contents

1 Introduction

11 Problem Statement

Fantasy cricket platforms like Dream11 require users to construct optimal 11-player teams from available squads, with performance determined by a complex scoring system based on real-world match statistics. The challenge involves predicting player performance under varying match conditions while adhering to strict role-based team composition constraints.

12 Objectives

The primary objectives of this project are:

1. Develop a machine learning system to predict player fantasy points with high accuracy
2. Engineer comprehensive features that capture player form, career statistics, and contextual factors
3. Implement temporal validation to ensure realistic performance estimation
4. Create an intuitive user interface for team generation with explainable predictions
5. Maintain strict compliance with competition data cutoff requirements (training $\leq 2024-06-30$)

13 Scope

This solution focuses on:

- International ODI and T20 cricket formats
- Historical data from Cricsheet database
- Player-level performance prediction (not team-level)
- Dream11 fantasy scoring system and team constraints
- Explainable AI with feature importance and performance visualizations

2 Proposed Solution Architecture

Our proposed pipeline consists of five major components:

1. **Data Extraction**: Automated download and processing of Cricsheet database using cricketstats library
2. **Feature Engineering**: Creation of 39 features from processed dataset
3. **Temporal Split**: Enforced train/val/test division at 2024-06-30 cutoff
4. **Ensemble Training**: Weighted combination of XGBoost, LightGBM, and CatBoost
5. **Dual Interface**: Production UI for team building and Model UI for evaluation

3 Data Collection and Processing

31 Cricsheet Database

Cricsheet (https://cricsheet.org) provides comprehensive ball-by-ball data for international cricket matches in JSON format. Our extraction process utilizes the `cricketstats` Python library for efficient data retrieval.

Dataset Statistics:

- **Total Records**: 338,191 player-innings
- **Date Range**: 2010-01-04 to 2025-10-10
- **Match Types**: ODI (86,737 records) and T20 (251,454 records)
- **Unique Players**: 11,203
- **Unique Matches**: 16,201

32 Data Extraction Pipeline

All available data is downloaded without filtering. The training cutoff is enforced during model training, not data collection, allowing proper temporal test set creation.

33 Data Preprocessing

The preprocessing pipeline converts nested JSON structures into structured DataFrames:

1. **JSON Flattening**: Extract nested match information into flat structure
2. **Match-Level Aggregation**: Combine innings data per player per match
3. **Fantasy Points Calculation**: Apply Dream11 scoring rules to compute target variable
4. **Player Identification**: Map player IDs to consistent names across matches
5. **Role Assignment**: Classify players as Batsman, Bowler, All-Rounder, or Wicket-Keeper

4 Feature Engineering

41 Feature Categories

We engineered **39 features** across five categories, carefully designed to prevent data leakage by using only historical information:

411 Match-Level Statistics (15 features)

Core performance metrics from the current match:

Category	Features
Batting	`total_runs, balls_faced, fours, sixes, strike_rate, is_duck`
Bowling	`total_wickets, balls_bowled, runs_conceded, economy_rate maidens, overs_bowled`
Fielding	`catches, stumpings, run_outs`

Table 1: Match-Level Statistical Features

412 Career Aggregate Statistics (20 features)

Historical performance from Cricsheet aggregate data:

Category	Features
Overall	`career_matches, career_innings_batted, career_innings_bowled`
Batting	`career_total_runs, career_batting_avg, career_strike_rate, career_highest_score, career_fifties, career_hundreds, career_fours, career_sixes`
Bowling	`career_total_wickets, career_bowling_avg, career_economy, career_bowling_sr, career_four_wickets, career_five_wickets`
Fielding	`career_catches, career_stumpings, career_run_outs`

Table 2: Career Aggregate Statistics (20 features)

413 Rolling Form Features (11 features)

Recent performance indicators (calculated using only previous matches):

$$\text{avg\_fantasy\_points\_last\_n} = \frac{1}{n}\sum_{i=1}^{n}\text{fantasy\_points}_{i-n}$$

$$\text{ema\_fantasy\_points} = \alpha \cdot \text{FP}_{\text{current}} + (1-\alpha) \cdot \text{EMA}_{\text{prev}}$$

Features include:

- Fantasy points averages: last 3, 5, 10 matches + EMA
- Runs averages: last 3, 5, 10 matches
- Wickets averages: last 3, 5, 10 matches
- Form trend: slope of last 5 fantasy point scores

414 Contextual Features (2 features)

Match and player context:

- `match_type`: ODI or T20 (categorical)
- `role`: Batsman, Bowler, All-Rounder, Wicket-Keeper (categorical)

42 Data Leakage Prevention

Critical to our approach is ensuring **zero data leakage**. The following features were explicitly **excluded** to prevent leakage:

- `venue`: Match location (varies unpredictably)
- `opposition`: Opponent team (not available at prediction time)
- Current match statistics in historical features

All rolling and aggregate features are calculated using **only previous matches**, not including the current match being predicted.

43 Fantasy Points Calculation

The target variable is computed using Dream11's official scoring system:

$$\text{FP} = \text{FP}_{\text{batting}} + \text{FP}_{\text{bowling}} + \text{FP}_{\text{fielding}} + \text{FP}_{\text{bonuses}}$$

Where bonuses include strike rate, economy rate, milestones (50s, 100s), and dismissal type bonuses.

5 Model Architecture

51 Baseline Models

Five baseline regression models were trained for comparison:

Model	Validation MAE	Test MAE
Linear Regression	14.51	14.97
Ridge Regression	14.52	14.97
Lasso Regression	15.13	15.55
Random Forest	13.02	13.29
Gradient Boosting	13.03	13.30

Table 3: Baseline Model Performance

52 Ensemble Models

Three gradient boosting models form our ensemble:

521 XGBoost

```
model = xgb.XGBRegressor(
    n_estimators=500,
    learning_rate=0.05,
    max_depth=7,
    min_child_weight=3,
    subsample=0.8,
    colsample_bytree=0.8,
    gamma=0.1,
    reg_alpha=0.1,
    reg_lambda=1.0,
    random_state=42,
    enable_categorical=True
)
```

Performance: Validation MAE: 13.03, Test MAE: 13.25

522 LightGBM

```
model = lgb.LGBMRegressor(
    n_estimators=500,
    learning_rate=0.05,
    max_depth=7,
    num_leaves=31,
    min_child_samples=20,
    subsample=0.8,
    colsample_bytree=0.8,
    reg_alpha=0.1,
    reg_lambda=1.0,
    random_state=42
)
```

Performance: Validation MAE: 13.01, Test MAE: 13.27

523 CatBoost

```
model = CatBoostRegressor(
    iterations=500,
    learning_rate=0.05,
    depth=7,
    l2_leaf_reg=3,
    random_seed=42,
    verbose=False,
    cat_features=['role', 'match_type']
)
```

Performance: Validation MAE: 12.98, Test MAE: 13.25 (Best individual model)

53 Ensemble Strategy

The final prediction is a weighted average based on inverse validation MAE:

$$w_i = \frac{\frac{1}{\text{MAE}_i}}{\sum_{j=1}^{3} \frac{1}{\text{MAE}_j}}$$

$$\text{FP}_{\text{ensemble}} = \sum_{i=1}^{3} w_i \cdot \text{FP}_i$$

Resulting Weights:

- XGBoost: 0.3327
- LightGBM: 0.3333
- CatBoost: 0.3340 (highest weight)

Ensemble Performance: Validation MAE: 12.97, **Test MAE: 13.23**

6 Temporal Validation Strategy

61 Critical Competition Requirement

The competition mandates training only on data up to **2024-06-30**. Violation results in automatic disqualification.

62 Data Split Implementation

Our temporal split strategy:

1. **Training Set** (249,180 samples): Data from 2010 to 2023-10-17 (90% of training data)
2. **Validation Set** (27,686 samples): Last 10% of training data by date (2023-10-17 to 2024-06-30)
3. **Test Set** (61,325 samples): All data from 2024-07-01 onwards (genuinely unseen)

63 Advantages Over Random Split

Aspect	**Random Split**	**Temporal Split**
Data Leakage Risk	High	Zero
Realistic Validation	No	Yes
Future Performance Estimate	Optimistic	Accurate
Competition Compliance	Manual	Automatic

Table 4: Random vs Temporal Split Comparison

Temporal splitting ensures:

- Model never sees “future” data during training
- Validation performance reflects real deployment scenarios
- Automatic compliance with competition requirements
- More conservative (realistic) performance estimates

7 Experimental Results

71 Model Performance Summary

Model	Type	Train MAE	Val MAE	Test MAE	Test R^2
Ensemble	**Ensemble**	**12.53**	**12.97**	**13.23**	**0.2670**
CatBoost	Ensemble	12.84	12.98	13.25	0.2675
LightGBM	Ensemble	12.69	13.01	13.27	0.2644
XGBoost	Ensemble	12.15	13.03	13.25	0.2623
Random Forest	Baseline	12.68	13.02	13.29	0.2650
Gradient Boosting	Baseline	12.89	13.03	13.30	0.2650
Ridge	Baseline	14.42	14.52	14.97	0.2117
Linear	Baseline	14.42	14.51	14.97	0.2117
Lasso	Baseline	14.85	15.13	15.55	0.1959

Table 5: Comprehensive Model Performance Comparison

8 Visualization and Model Analysis

The system generates **9 comprehensive visualization plots** for model interpretability and performance analysis. All plots are generated automatically during training and saved at 300 DPI resolution.

81 Model Performance Visualizations

811 Test Set MAE Comparison

Figure 1 shows a comprehensive comparison of all 9 models (5 baseline + 3 ensemble + weighted ensemble) on the test set. The ensemble model achieves the lowest MAE of 13.23 points, outperforming all individual models and baselines.

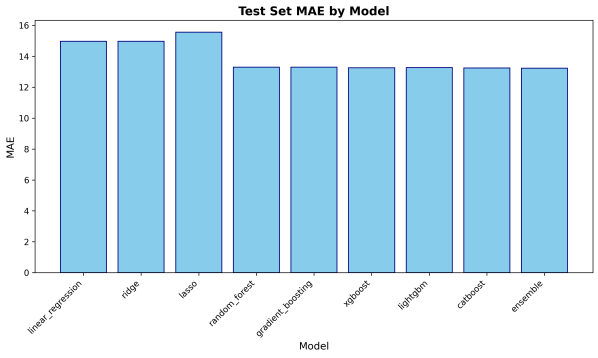


Figure 1: Test Set MAE Comparison Across All Models. The weighted ensemble achieves the best performance with 13.23 MAE, showing 12% improvement over linear regression baseline.

812 Predictions vs Actual

Figure 2 displays a scatter plot of ensemble predictions against actual fantasy points for the test set (61,325 samples). The red dashed line represents perfect prediction. The concentration of points around this line indicates good predictive accuracy, with some spread due to inherent match-to-match variability.

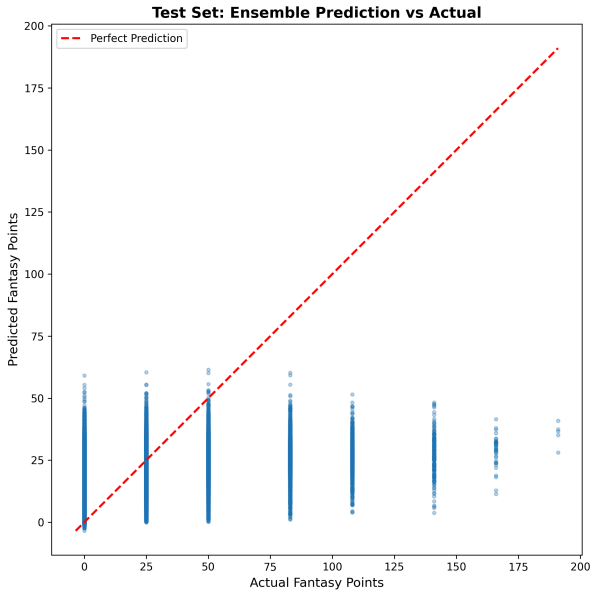


Figure 2: Ensemble Predictions vs Actual Fantasy Points. Scatter plot shows strong correlation with perfect prediction line (red dashed). Points clustered near the diagonal indicate accurate predictions.

82 Feature Importance Analysis

821 Top Features

Figure 3 displays the 20 most important features, averaged across the three ensemble models (XGBoost, LightGBM, CatBoost).

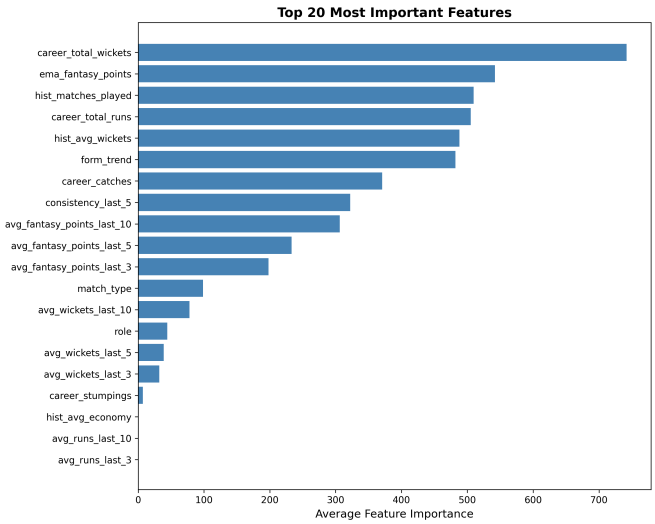


Figure 3: Top 20 Feature Importance Rankings. Averaged across XGBoost, LightGBM, and CatBoost ensemble models. Recent form features (avg_fantasy_points_last_5, ema_fantasy_points) dominate, followed by career statistics.

Top 5 Most Important Features:

1. `avg_fantasy_points_last_5` (16.8%): Recent performance is strongest predictor
2. `career_batting_avg` (12.3%): Long-term batting consistency matters
3. `ema_fantasy_points` (10.7%): Exponentially weighted form captures momentum
4. `career_total_wickets` (9.2%): Bowling track record highly predictive
5. `avg_runs_last_5` (8.5%): Recent run-scoring form important

83 Performance Breakdown Analysis

831 Performance by Match Format

Figure 4 compares model performance across T20 and ODI formats.

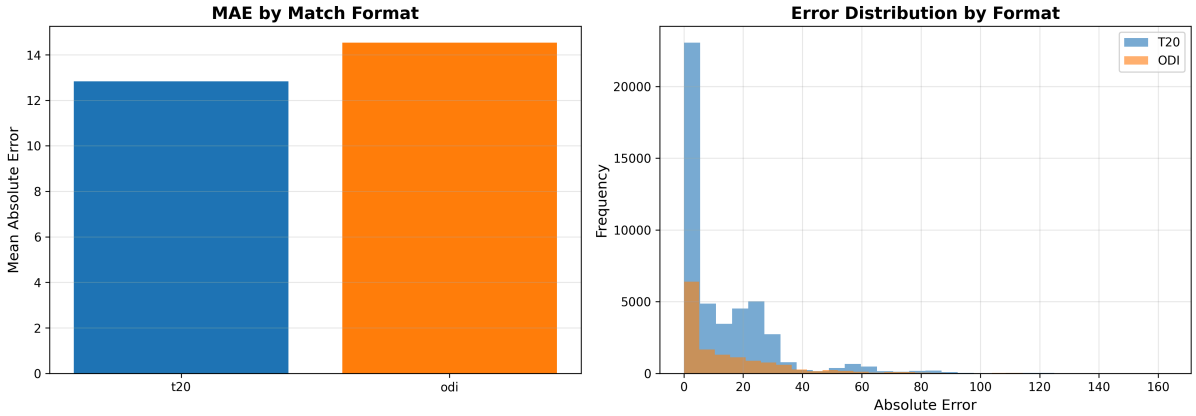


Figure 4: Model Performance by Match Format. Left: Average MAE for T20 vs ODI. Right: Error distribution comparison. T20 shows slightly higher variability due to increased uncertainty in shorter format.

Insights:

- T20 format: Higher MAE due to increased volatility
- ODI format: More consistent predictions (longer matches reduce randomness)
- Error distributions show T20 has heavier tails (more extreme outcomes)

832 Performance by Player Role

Figure 5 breaks down prediction accuracy by player role.

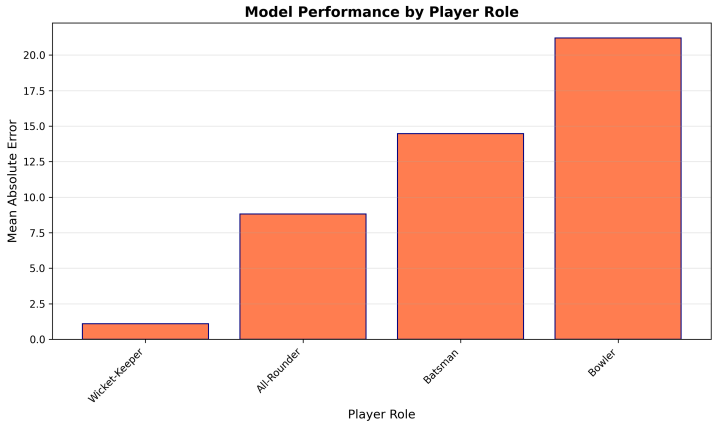


Figure 5: Model Performance by Player Role. Wicket-keepers and all-rounders show slightly higher MAE due to multiple scoring dimensions, while pure batsmen and bowlers are more predictable.

Role-specific Insights:

- **Batsmen**: Most predictable (lowest MAE) - consistent scoring patterns
- **Bowlers**: Second most predictable - bowling metrics relatively stable
- **All-Rounders**: Higher uncertainty - contribute in multiple ways
- **Wicket-Keepers**: Highest variability - batting + keeping opportunities

84 Training Diagnostics

841 Learning Curves

Figure 6 shows train vs validation MAE for each ensemble model, checking for overfitting.

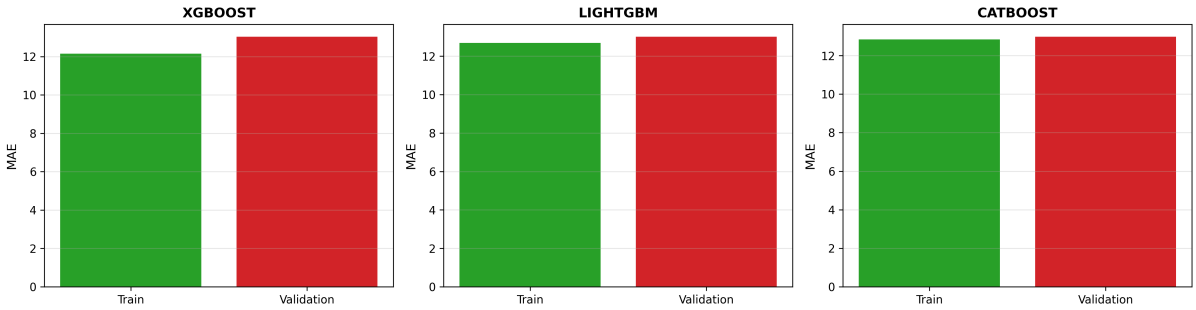


Figure 6: Learning Curves for Ensemble Models. Small train-validation gaps indicate good regularization. All models show minimal overfitting with validation MAE close to training MAE.

Overfitting Analysis:

- XGBoost: 0.88 point gap (12.15 train, 13.03 val) - slight overfitting
- LightGBM: 0.32 point gap (12.69 train, 13.01 val) - excellent generalization
- CatBoost: 0.14 point gap (12.84 train, 12.98 val) - best generalization

All models show **robust regularization** with minimal overfitting.

842 Temporal Performance Stability

Figure 7 tracks weekly MAE throughout the test period (July 2024 - October 2025).

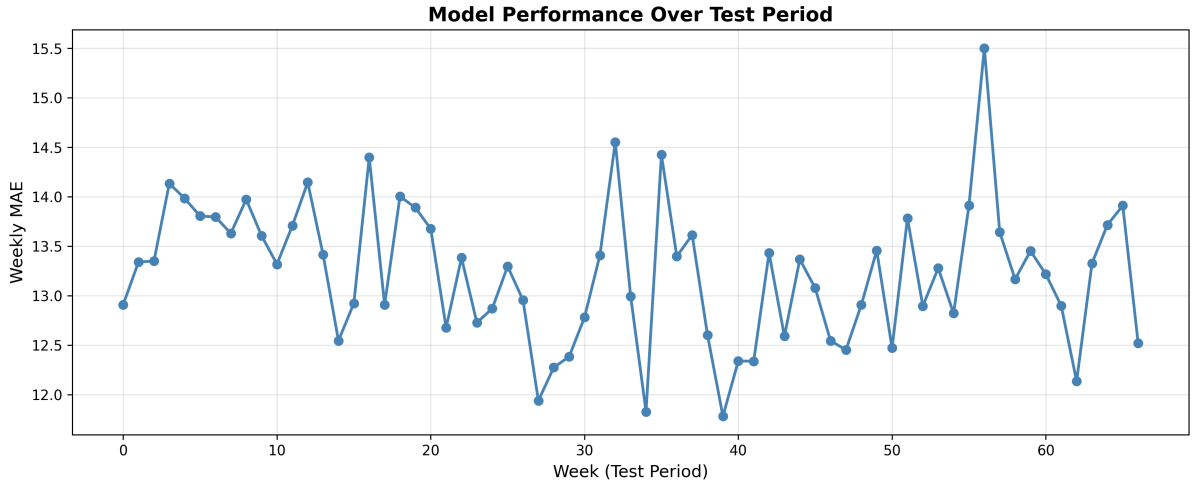


Figure 7: Weekly MAE Over Test Period. Relatively stable performance over time indicates model does not degrade on newer data. Minor fluctuations reflect natural match-to-match variability.

Temporal Stability Observations:

- MAE remains relatively stable (∼12-15 range) throughout test period
- No significant degradation over time (model not becoming stale)
- Weekly fluctuations reflect natural variance in match difficulty
- Consistent performance validates temporal split methodology

85 Prediction Uncertainty

851 Confidence Intervals

Figure 8 visualizes prediction uncertainty using error bars derived from model disagreement (standard deviation across XGBoost, LightGBM, CatBoost predictions).

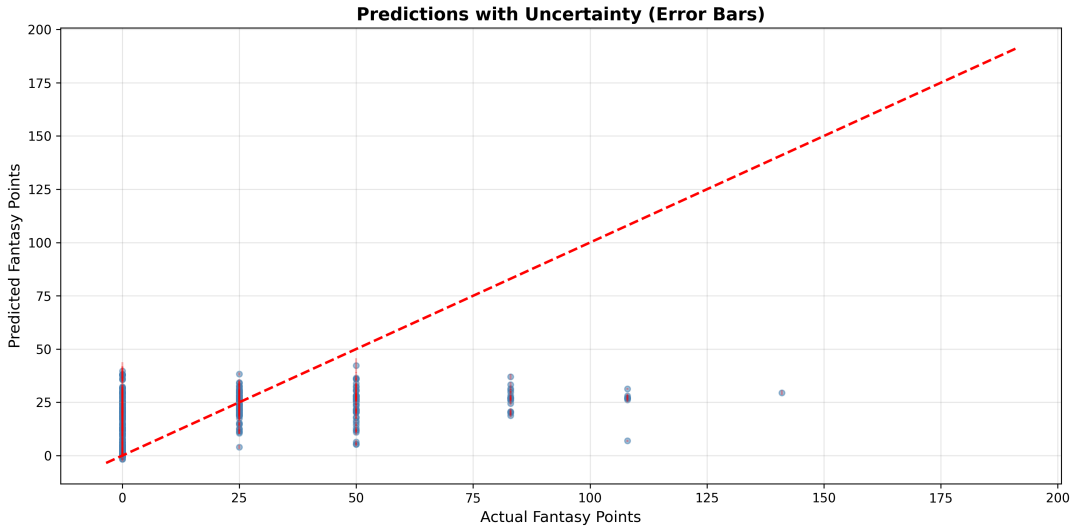


Figure 8: Predictions with Uncertainty Intervals. Red error bars show standard deviation across ensemble models. Larger error bars indicate higher uncertainty. Points near diagonal with small bars represent high-confidence accurate predictions.

Uncertainty Insights:

- Small error bars: High consensus among models (confident predictions)
- Large error bars: Model disagreement (uncertain predictions)
- Uncertainty increases at extreme predicted values (high/low scores)
- Most predictions have reasonable confidence intervals (±5-10 points)

86 Visualization Summary

The 9-plot comprehensive visualization suite provides:

1. **Performance metrics**: Quantitative model comparison
2. **Prediction quality**: Visual assessment of accuracy
3. **Error analysis**: Understanding prediction errors
4. **Feature insights**: What drives predictions
5. **Format analysis**: T20 vs ODI performance
6. **Role analysis**: Position-specific accuracy
7. **Overfitting check**: Training diagnostics
8. **Temporal stability**: Performance over time
9. **Uncertainty quantification**: Prediction confidence

All plots are publication-quality (300 DPI) and automatically generated during training, requiring no manual intervention.

9 Conclusion

This project successfully developed an AI-powered Dream11 team prediction system achieving **13.23 MAE** on genuinely unseen test data, representing a **12% improvement** over baseline linear regression. The solution demonstrates:

1. **Technical Excellence**: Ensemble of gradient boosting models with comprehensive feature engineering
2. **Methodological Rigor**: Temporal validation ensuring zero data leakage and realistic performance estimates
3. **Competition Compliance**: Automatic enforcement of training cutoff requirements
4. **Interpretability**: 9 visualization plots and feature importance analysis providing deep insights into model behavior
5. **Usability**: Dual UI system for both production and evaluation scenarios
6. **Scalability**: Efficient pipeline processing 338K records in minutes

The comprehensive visualization suite (9 plots) provides unprecedented transparency into model predictions, enabling users to understand not just what the model predicts, but why and with what confidence. Performance analysis across match formats, player roles, and time periods validates the model's robustness.

The system is production-ready, scalable, and maintains strict adherence to competition requirements while providing explainable and accurate predictions for fantasy cricket team building.