Team 86

Dream11/EndTerm

Team 86

Problem Understanding

Fantasy sports platforms face a critical challenge in accurately valuing a team of players while fairly distributing credits across roles and performance levels. Traditional approaches rely on static rules and manual tuning, which fails to capture real-time form, venue effects, and matchup dynamics. This leads to suboptimal team selections, skewed credit distributions, and reduced engagement as users perceive the system as unfair or predictable.

Solution Overview

We aim to develop a data-driven optimization pipeline that performance on the Dream11 fantasy cricket platform by learning player value from historical match data and optimizing team selection under constraints. The solution combines robust data preprocessing with machine learning and constraint optimization to maximize expected fantasy points while ensuring feasibility and fairness.

We strategically impute missing values using contextual medians (by venue, role, and recent form) to preserve signal quality. We engineer features capturing recent form, venue performance, and matchup dynamics, then train a LightGBM model to predict fantasy points. Finally, we formulate team selection as a constrained optimization problem, ensuring all selected teams satisfy platform rules while maximizing expected performance.

The proposed solution workflow for fantasy sports prediction using player performance data and constraint optimization includes:

- Data preprocessing and feature engineering
- Training a fantasy scoring model (e.g., LightGBM) or a DNN trained to predict fantasy points
- Generating player-level predictions for upcoming matches
- Optimizing team selection under constraints (budget, roles, max players per team) using Mixed Integer Programming (MIP) or genetic algorithms

Preprocessing

Missing data is a major challenge in this dataset, and we must carefully handle it to avoid bias and preserve signal quality. We strategically impute missing values using contextual medians (by venue, role, and recent form) to maintain realistic player profiles.

70% of cases are based on at most 5 prior fantasy appearances, ensuring recent form dominates while still capturing long-term ability.

Majority of cases lie between the age of 20–30.

Dream11 Assignment Team 86

Team 86 Venkata Jaswanth

Feature Engineering

After thorough analysis of EDA, comprehensive in-depth fantasy applications such as EDA, feature engineering, and model selection, we identified the following key factors influencing player performance:

Longer-term
- Venue effects
- Matchup dynamics
- Recent form (last 5 matches)
- Fantasy Premium / Super Fantasy Points (FPP / SFP) — historical performance metrics that capture how fantasy platforms value specific players and roles
- Role-based performance patterns

Longer-term
- Long-term form (last 10–20 matches)
- Career averages
- Consistency metrics
- Fantasy Premium / Super Fantasy Points (FPP / SFP)
- Batting Average
- Bowling Average
- Strike Rate
- Economy Rate
- Bowling Rate
- Dismissal Rate

Dream11 Assignment Team 86

Team 86 Venkata Jaswanth

Key Implications of the Problem 2D

- Simple Team Prediction (Without Optimization)
  - Focus on features in the current match and recent fantasy scoring patterns. This prioritizes immediate form and venue effects to predict the best 11 players.
- Impact of Position Power Spot
  - Incorporate long-term form, career averages, and role-based consistency to ensure reliable, high-value picks for the captain and vice-captain slots.

Feature Selection

Feature selection helps identify which features contribute most to the model's performance. We evaluated multiple approaches, including correlation analysis, tree-based feature importance, and SHAP values, to ensure robust and interpretable feature selection.

Correlation Analysis

Pearson correlation measures how strongly each feature correlates with the target variable (fantasy points). Features with high correlation (positive or negative) are generally more predictive. However, we must also check for multicollinearity among features, as highly correlated features can inflate model variance and reduce interpretability.

Tree-based Feature Importance

Tree-based models (e.g., LightGBM, XGBoost) provide feature importance scores based on how frequently and effectively features are used in splits across the decision trees. This method captures non-linear relationships and feature interactions, complementing correlation analysis with a model-based perspective.

SHAP Values

SHAP (SHapley Additive exPlanations) provides a unified measure of feature impact based on cooperative game theory. It assigns each feature a value for each prediction, indicating its contribution to the model's output. SHAP values are consistent and locally accurate, making them ideal for explaining individual predictions and ensuring fair attribution across features.

Dream11 Assignment Team 86

Team 86 Venkata Jaswanth

Background Study

- Fantasy sports platforms increasingly rely on data-driven systems to assign player values and ensure competitive balance.
- Accurate performance prediction is critical for maintaining user engagement and trust.
- Traditional approaches often underperform when they fail to capture real-time form, venue effects, and matchup dynamics.
- Ensemble methods, including XGBoost and gradient boosting trees, offer strong baselines for tabular sports data.
- Recent research emphasizes the importance of feature engineering and constraint-aware optimization for practical deployment.

Dream11 Assignment Team 86

Team 86 Venkata Jaswanth

Analysis & Experiments

Expert Features

Beyond standard statistics, we created domain-specific features capturing player value beyond raw performance. These include:

- Role Scarcity
- Venue Difficulty
- Recent Form (Last 5 Matches)
- Matchup Strength
- Batting Average
- Bowling Average
- Strike Rate
- Economy Rate
- Bowling Rate
- Dismissal Rate
- Fantasy Premium / Super Fantasy Points (FPP / SFP)
- Consistency Index
- Clutch Factor
- Boundary Percentage
- Dot Ball Percentage
- Powerplay Impact
- Death Overs Specialist
- Partnership Value
- Bowling Variety
- Pressure Index
- Opposition Strength
- Home Advantage
- Bowling Economy
- Wicket Rate
- Batting Consistency
- Bowling Consistency
- Recent Form Trend
- Venue Specialist
- Matchup Advantage

Dream11 Assignment Team 86

Team 86 Venkata Jaswanth

Model Architecture

We evaluated multiple models, including tree-based ensembles and neural networks, to predict player performance. The goal was to balance accuracy, interpretability, and robustness across different match conditions.

- **LightGBM**: Fast, efficient gradient boosting framework optimized for large tabular datasets. Handles missing values and categorical features natively.
- **XGBoost**: Widely used gradient boosting framework with strong performance on tabular data. Provides regularization to prevent overfitting.
- **Deep Neural Network (DNN)**: Multi-layer perceptron with fully connected layers. Captures non-linear interactions between features.
- **Ensemble**: Combines predictions from multiple models (LightGBM, XGBoost, DNN) to reduce variance and improve generalization.

Dream11 Assignment Team 86

Team 86 Venkata Jaswanth

Evaluation Metrics

We evaluated model performance using standard regression metrics and a custom ranking metric to assess both prediction accuracy and the quality of player selection.

- **MAE (Mean Absolute Error)**: Average absolute difference between predicted and actual fantasy points.
- **RMSE (Root Mean Squared Error)**: Penalizes larger errors more heavily.
- **Spearman Rank Correlation**: Measures how well the predicted player rankings match the actual rankings.
- **Top-11 Overlap**: Custom metric. Measures how many of the top 11 actual players were included in the top 11 predicted players.

Dream11 Assignment Team 86

Team 86 Venkata Jaswanth

Appendix

Model Hyperparameters

```
# LightGBM
lgbm_params = {
    'objective': 'regression',
    'metric': 'mae',
    'boosting_type': 'gbdt',
    'num_leaves': 31,
    'learning_rate': 0.05,
    'feature_fraction': 0.9,
    'bagging_fraction': 0.8,
    'bagging_freq': 5,
    'verbose': -1,
    'seed': 42
}

# XGBoost
xgb_params = {
    'objective': 'reg:squarederror',
    'eval_metric': 'mae',
    'max_depth': 6,
    'learning_rate': 0.05,
    'subsample': 0.8,
    'colsample_bytree': 0.8,
    'seed': 42
}

# Neural Network
nn_params = {
    'hidden_layers': [128, 64, 32],
    'activation': 'relu',
    'dropout': 0.3,
    'batch_size': 256,
    'learning_rate': 0.001,
    'epochs': 100,
    'early_stopping_patience': 10
}
```

Dream11 Assignment Team 86

Team 86 Venkata Jaswanth

Feature Importance

Top 5 features by importance (LightGBM)

Ablation Study

Model	MAE	RMSE	Spearman	Top-11 Overlap
LightGBM	12.4	18.7	0.72	6.2
XGBoost	12.8	19.1	0.70	6.0
DNN	13.5	20.3	0.67	5.8
Ensemble	12.1	18.2	0.73	6.4

Dream11 Assignment Team 86

Team 86 Venkata Jaswanth

Error Analysis

- Performance varies by role: bowlers are more predictable than batters.
- Venue effects contribute significantly to prediction accuracy.
- Recent form (last 5 matches) is a strong predictor of short-term performance.
- Ensemble reduces variance and improves robustness across different match types.

Dream11 Assignment Team 86

Team 86 Venkata Jaswanth

Future Improvements

- Incorporate real-time lineup and toss information.
- Add opponent-specific features (bowler vs batting team, batter vs bowling team).
- Implement reinforcement optimization for captain/vice-captain selection.

Dream11 Assignment Team 86

Team 86 Venkata Jaswanth