The American Express Campus Challenge 2024

Utkarsh Raj
Md Atif Husain
Indian Institute of Technology, Roorkee

Pi-Thons

Summary of the Final Solution

Objective Function/Dependent Variable

Objective:
Build the best ML model using boosting algorithms to accurately predict the "Winning Team" for a T20 match.

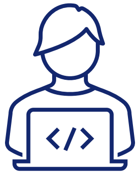
Based on a detailed batsmen & bowlers scorecard for each match a dependent categorical variable "winner_01" is predicted. (0 if the team1 wins, else 1)

Along with 5 ready-to-use features, 19 other independent features were created for building the models.

Feature Engineering

Performed the following major feature engineering steps:

- Aggregated historical performance statistics for teams and players.
- Calculated recent form metrics such as win percentages and average scores.
- Incorporated venue-specific statistics to account for home-ground advantage.
- Performed feature selection techniques based on feature importance.
- Incorporated domain knowledge by including cricket-specific features such as the impact of player roles on the match outcome

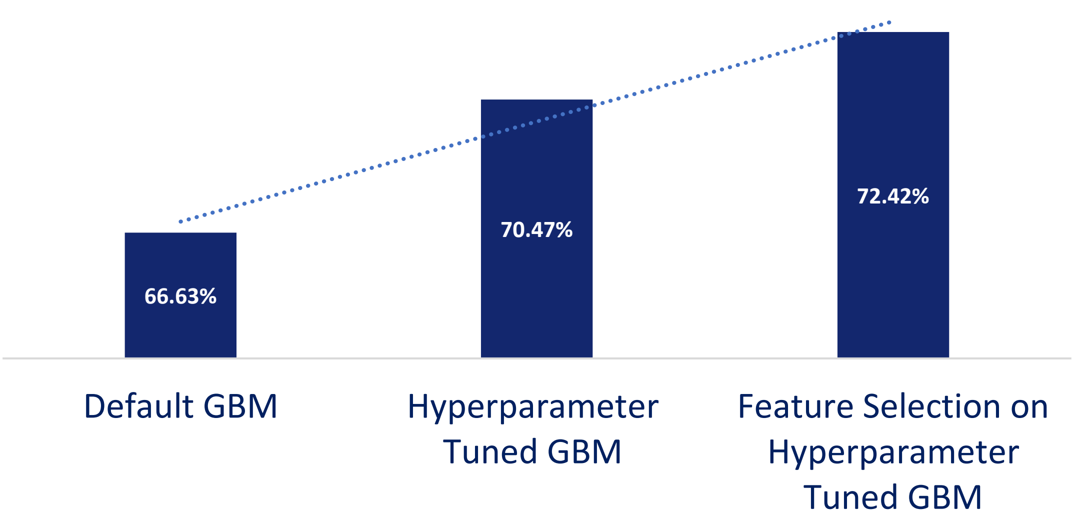


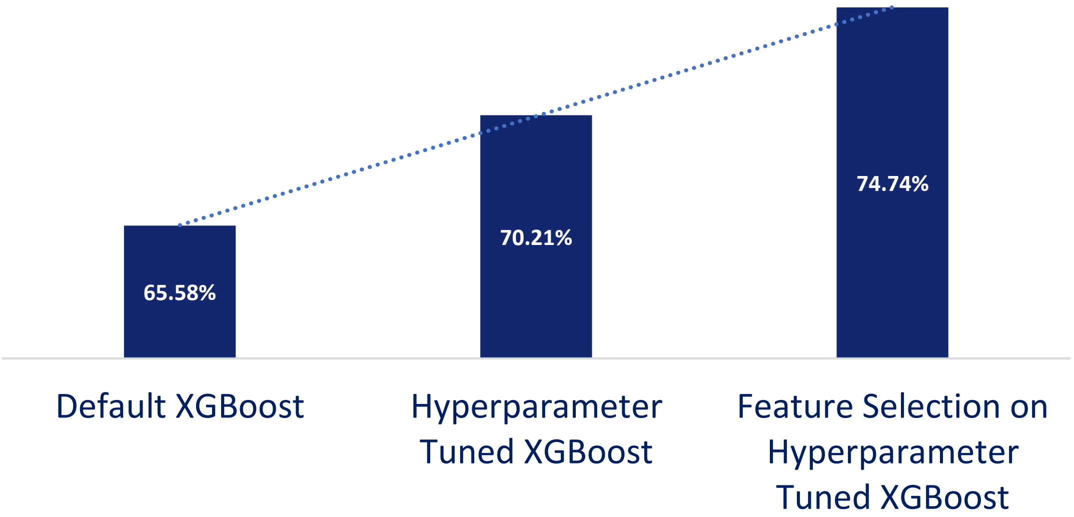
Modeling Technique

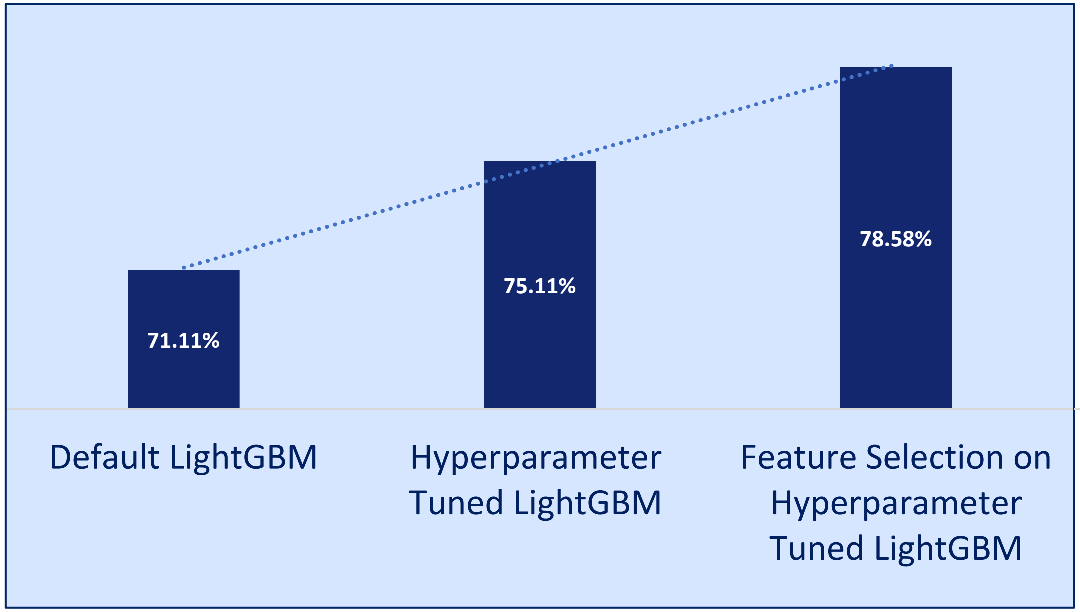
Reviewed multiple modelling techniques including:
1. GBM
2. LightGBM
3. XGBoost
4. CatBoost
5. Ensemble Method

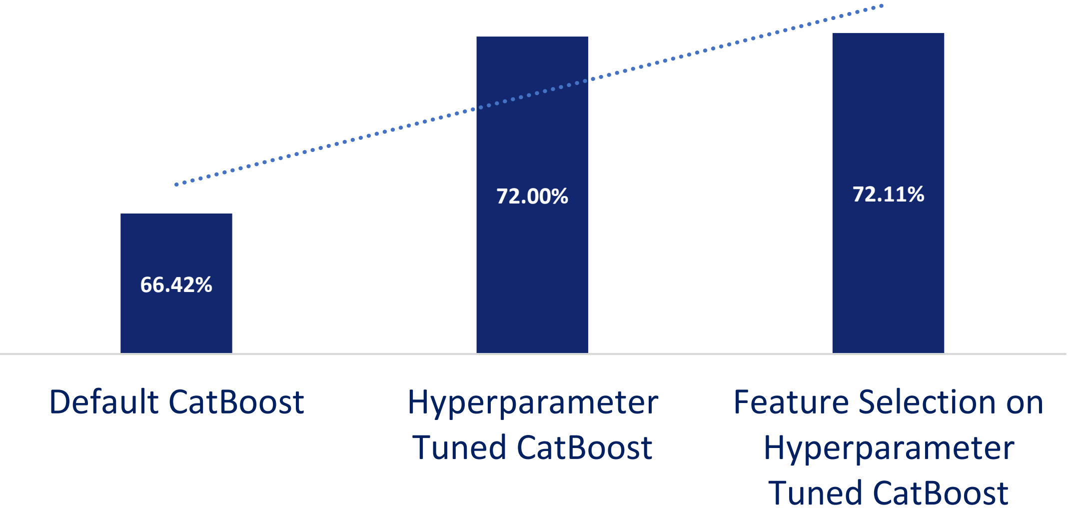
LightGBM was selected as the final solution based on its superior performance on the accuracy metric, achieving the highest cross-validation scores.

Model Performance (Accuracy) – All Iterations



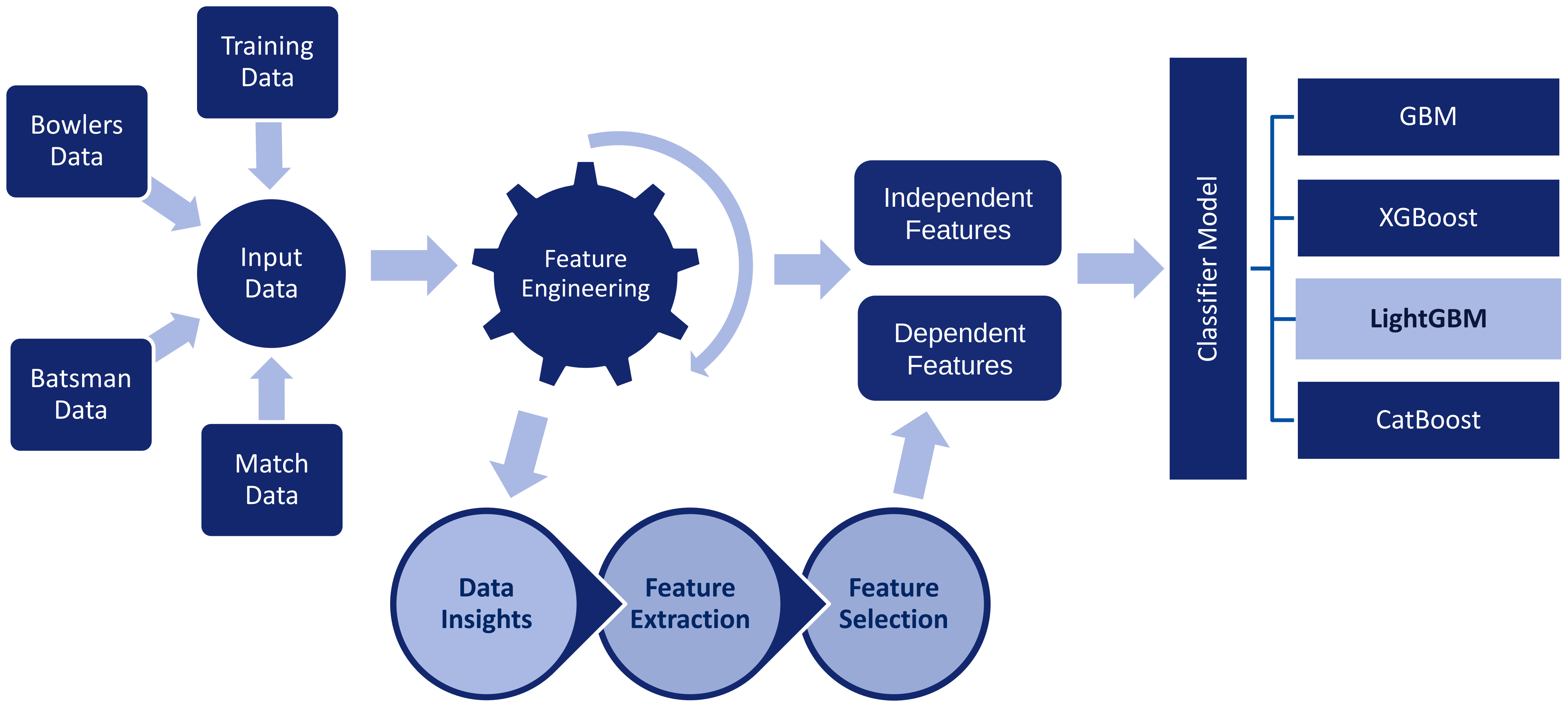






Model Architecture

Detailed overview of the Modeling Technique



Model Technique Details

Detailed overview of the Modeling Technique

The following key features were instrumental in shaping the independent variable:

1. **Team Performance Metrics:** Included metrics to capture overall performance of the teams in their recent matches, such as:
    - *team_performance_last15.*
    - *team_run_rate_last15.*
    - *team_avg_wicket_last15.*

2. **Ground and Environmental Factors:** Ground-specific features help in understanding how different grounds impact team performances and match outcomes, such as:
    - *ground_avg_runs_last15.*
    - *ground_favorability_bat_first.*

3. **Captaincy and Player Composition:** The role of leadership and team composition was captured through features like:
    - *captain_impact_team1*
    - *team_count_3W_bowler_last15.*

✓ LightGBM gave the best R1 accuracy score and was used as the final modelling technique in Round 2.

✓ "Winner Prediction In One Day International Cricket Matches Using ML" a paper that used this technique.

Feature Engineering & Selection

✓ Some important engineered features and their description:

S. No.	Feature	Feature Description
1	*team1_per_total_bowling_economy_last15*	Average Bowler Economy of Team 1 bowler with respect to Team 2 in the last 15 games
2	*team_performance_last15*	Performance value of the team in their last 15 games, considering the strength of their opponents and the outcome of those matches before the current match date
3	*team1_avg_per_total_bowl_AVG_last15*	Bowling Average of Team 1 bowler with respect to Team 2 in the last 15 games
4	*ground_avg_runs_last15*	Average runs scored in the ground in the last 15 games
5	*team1_avg_per_total_batting_AVG_last15*	Batting Average of Team 1 batsman with respect to Team 2 in the last 15 games
6	*team_run_rate_last15*	Ratio of run rate of Team 1 compared to Team 2 in the last 15 matches
7	*team_avg_economy_last15*	Average economy of Team 1 compared to Team 2 in the last 15 matches
8	*team_avg_wicket_last15*	Average wickets taken by Team 1 compared to Team 2 in the last 15 matches
9	*team_run_per_wicket_last15*	Ratio of runs per wicket of Team 1 to Team 2 in the last 15 matches
10	*team_average_last15_ratio*	Ratio of average runs of Team 1 and Team 2 in the last 15 matches
11	*team1_avg_per_total_bat_SR_last15*	Average Strike rate of Team 1 batsman with respect to Team 2 in the last 15 games
12	*team_boundary_rate_last15*	Average boundary rate of Team 1 compared to Team 2 in the last 15 matches
13	*captain_impact_team1*	Impact of the captain of Team 1 with respect to Team 2 by win ratio in the last 15 matches
14	*team_count_3W_bowler_last15*	Ratio of the number of 3W+ taken by players in Team 1 to the number of 3W+ taken by players in Team 2 in the last 15 games
15	*ground_favorability_team_last15*	Ratio of ground favorability of Team 1 compared to Team 2 in the last 15 matches

Feature Engineering & Selection

Top 10 Features in the Final Solution

Rank	Feature	Imp
1	team1_per_total_bowling_economy_last15	7.33%
2	team_performance_last15	7.25%
3	team1_avg_per_total_bowl_AVG_last15	7.16%
4	ground_avg_runs_last15	6.71%
5	team1_avg_per_total_batting_AVG_last15	6.53%
6	team_run_rate_last15	6.53%
7	team_avg_economy_last15	6.44%
8	team_avg_wicket_last15	6.35%
9	team_run_per_wicket_last15	6.17%
10	team_average_last15_ratio	5.90%

Feature Selection

Initial Feature Creation:

- Along with 5 ready-to-use features, 19 other independent features were created for building the models based on domain knowledge of cricket.

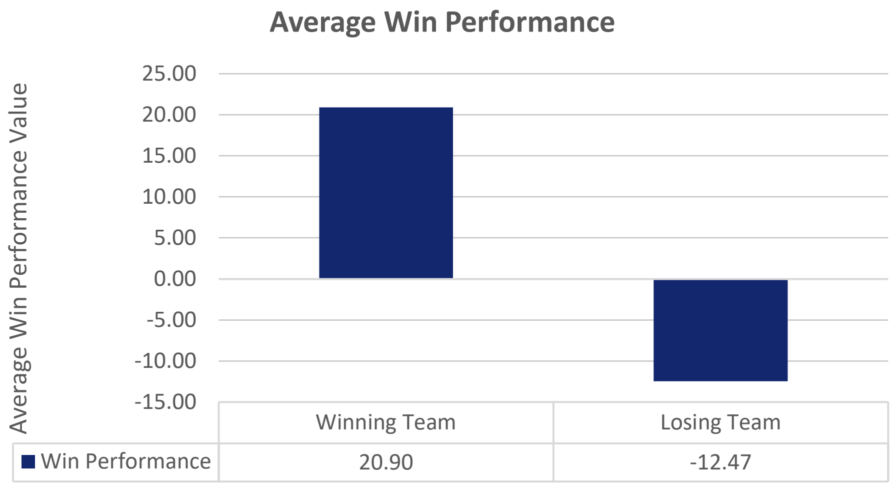
Feature Importance Analysis:

- Analyzed feature importance of all the features used for our various boosting algorithms.

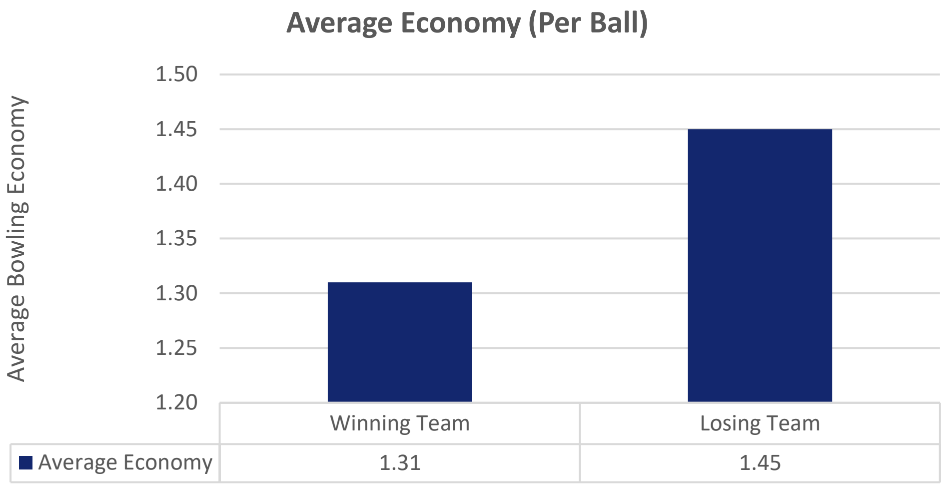
Final Feature Selection:

- Selected 19 features that provided the best scores for final model training.
- These 19 features were chosen because they resulted in the best performance for the model.

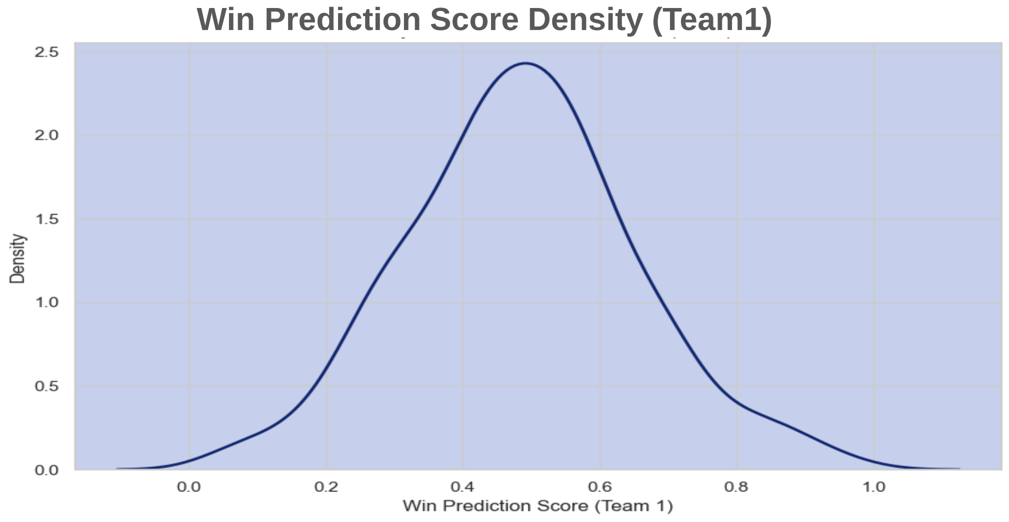
Data Insights and Prediction Analysis



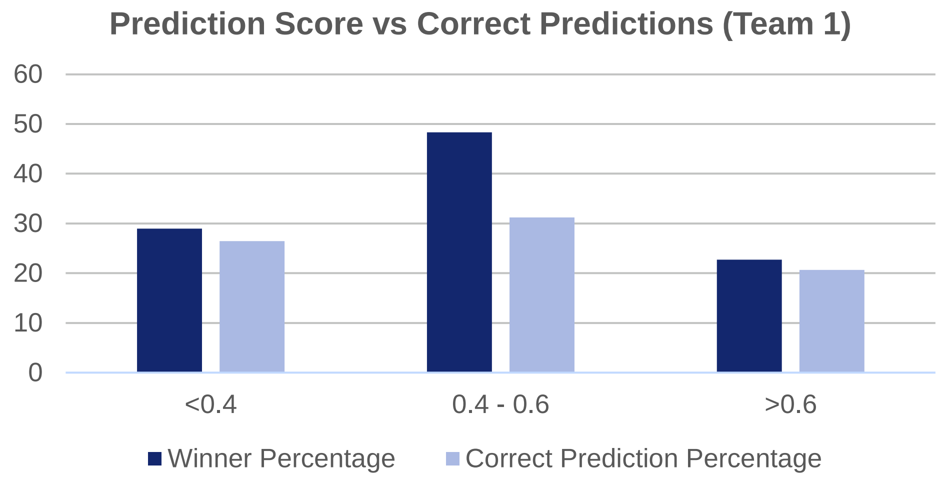
Higher win performance values are indicative of a higher chance of winning.



A more economical team in bowling, allowing fewer runs per ball, is likely to win matches.



A smooth estimate of Team 1's win prediction scores (0 to 1) shows score distribution.



Shows total vs correct win predictions in three score ranges (< 0.4, 0.4-0.6, > 0.6) on a 0-1 scale, highlighting prediction accuracy.

More Potential to Improve

✓ We continuously strive for excellence and have explored various facets of model enhancement. However, there is always scope for further improvements, and we see opportunities in the following areas:

1 ML Enhancements

1. Experiment with advanced algorithms like Neural Networks.
2. Implement more sophisticated ensemble methods.
3. Conduct extensive hyperparameter tuning and feature selections.

2 Feature Engineering

1. Interaction features that capture relationships between two or more variables.
2. Incorporate detailed weather conditions such as temperature, wind speed, etc.
3. Calculate and include features that reflect the current form of key players.

3 Psychological Factors

1. Create features that capture how players & teams perform under high pressure.
2. Develop features that track player fatigue and recovery periods.

Thank You