E0 259 : Data Analytics
ODI LIVE WIN PROBABILITY

Team Name : Fortune Tellers

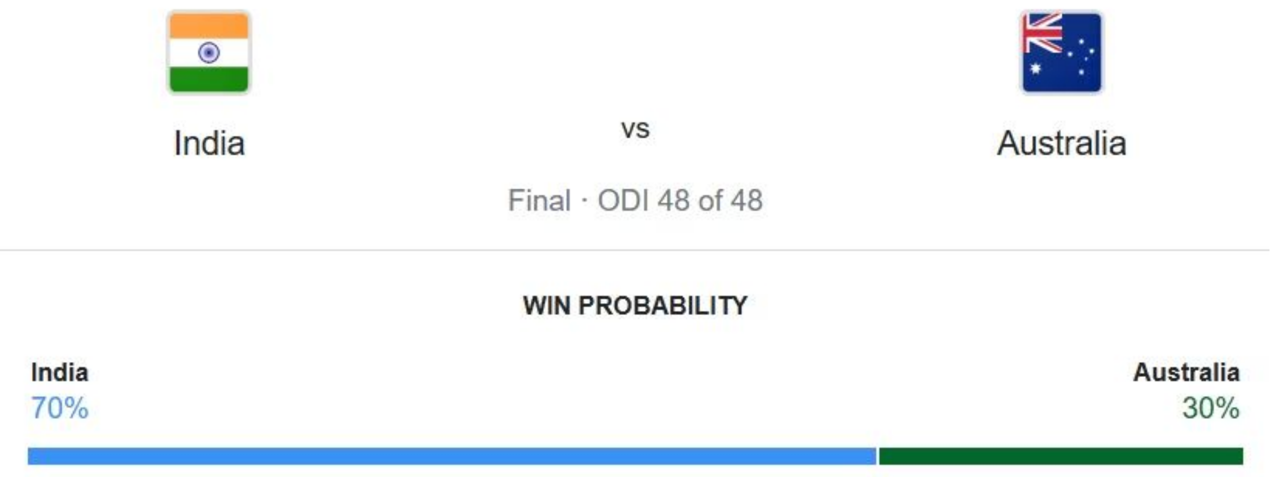
Team Number : 2

Team Members : Rahul Ranjan, M.V.S.AJAY, Smit Radadiya, Dhaval Chavda

Table of Contents

Problem Statement

To predict live win probability for ongoing match ball by ball.



Dataset Overview

Dataset is collected from official website (https://cricksheet.org)

Contains data of ODI men matches

Ball by Ball data for both innings

So many features like overs, run_scored, wickets_fallen etc.

Data Preprocessing

- Extracted Features
  - Head to head matches results
  - Last 5 matches result of batting team
  - Last 5 matches result of fielding team

- Removed extra features which are not relevant to model and not available in a live match like final score of innings.

Design Choices

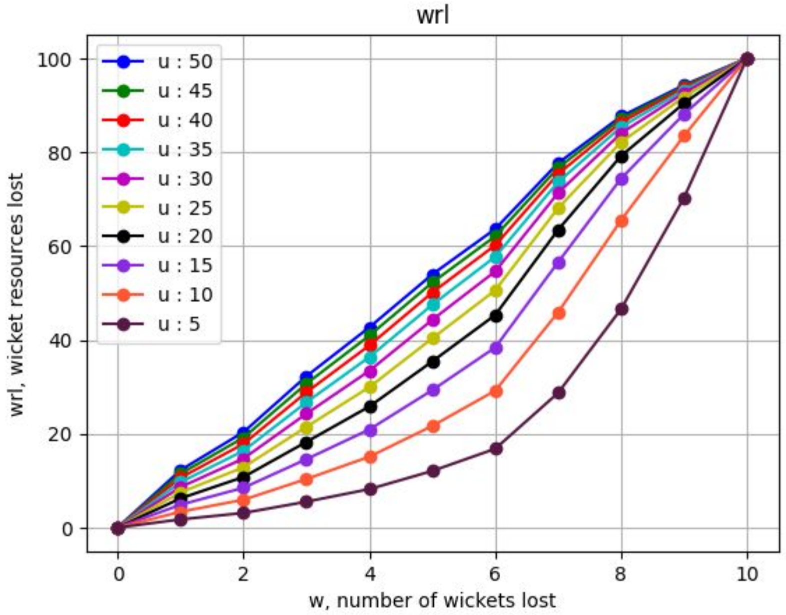
- Two distinct models for the two innings because
  - First innings objective is “to score as much as runs possible” whereas second innings objective is “to comfortably reach the target score”.
  - ‘Required_run_rate’ or ‘target_score’ is a relevant feature only for second innings.

Feature Engineering

1. Wicket Resource Left (wrl)

- Unequal weight of wickets lost at the start and end of the innings.

$$wrl = \frac{Z(u,0) - Z(u,w)}{Z(u,0)}$$



Feature Engineering

2. Form difference

- weighted mean of match outcomes over their last five games.

$$form = \frac{\sum_{t=1}^{5} w(t,\theta)y_t}{\sum_{t=1}^{5} w(t,\theta)}, \text{ where } w(t,\theta) = (1-\theta)^{t-1} \text{ and } 0 < \theta < 1.$$

3. Predicted Remaining score

- Using knn neighbours, we are calculating this feature.
- Adding this feature to input features to predict the win probability.

Methods

- Duckworth Lewis Par score
- Basic Logistic Regression
- Dynamic Logistic Regression
- XGBoost

DLS Par score

- Calculate DLS par score
    - 1st inning
        - Par score = $Z(u, w) = Z0(w)[1 - \exp\{-Lu/Z0(w)\}]$
    - 2nd inning
        - Par score = Resources used * target
        - Resources used = $Z(u, w)/Z(N, 10)$

          Where; u = overused, w = wickets fallen, N = 50

DLS Par score

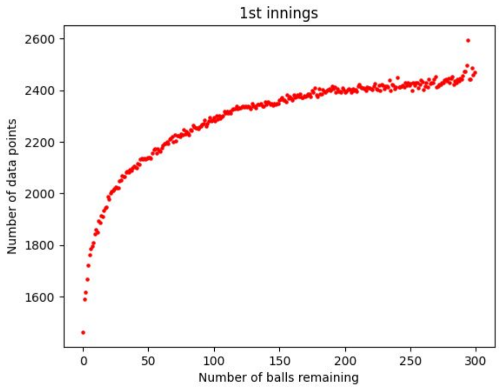
- Using Par score and run scored finding probability of winning of batting team
  - If both same then 50%-50% for both teams
  - If Run scored > par scored
    - Probability = (Run scored - Par score)/Run scored
  - Else
    - Probability = 1 - (Par scored - Run score)/Par scored

Basic Logistic Regression

- Trained different Logistic Regression model for 1st and 2nd innings
- Used features like Overs used, Run scored, wicket falen, Current run rate etc. to train model
- The final output is the predicted probability for the batting team
- $f(x) = \frac{1}{1+e^{-x}}$ where $x = w^T x_d$ where x_d is the data point and w is the vector of coefficients.
- f(x) gives us a value between 0 and 1, which can be interpreted as probability.

Dynamic Logistic Regression

- In one inning total 300 balls to be bowled and 10 wickets is available
- Now considering each scenario separately 0-300 = 301 and 0-10= 11
  - Total 301 * 11 cases
  - Training 3311 different models(Basic Logistic Regression) with case specific data
- To predict the probability
  - we use **u** and **w** to get corresponding model
  - Using that model we predict the probability



XGBoost

- XGBoost (eXtreme Gradient Boosting) is an ensemble learning method that combines the predictions of multiple weak models (decision trees) to create a robust and accurate model.

- XGBoost with Binary Logistic Regression objective.

- Loss function : Cross entropy loss
- Trained two separate models for both innings.

- XGBoost provides insights into feature importance, helping identify key factors influencing win probability.

Performance of individual methods

Accuracy

Model	First Innings	Second Innings
DLS	38.34%	43.23%
Basic LR	74.57%	92.07%
Dynamic LR	77.78%	85.60%
XGBoost	77.78%	92.43%
XGBoost (with engineered features)	**94.45%**	**97.7%**

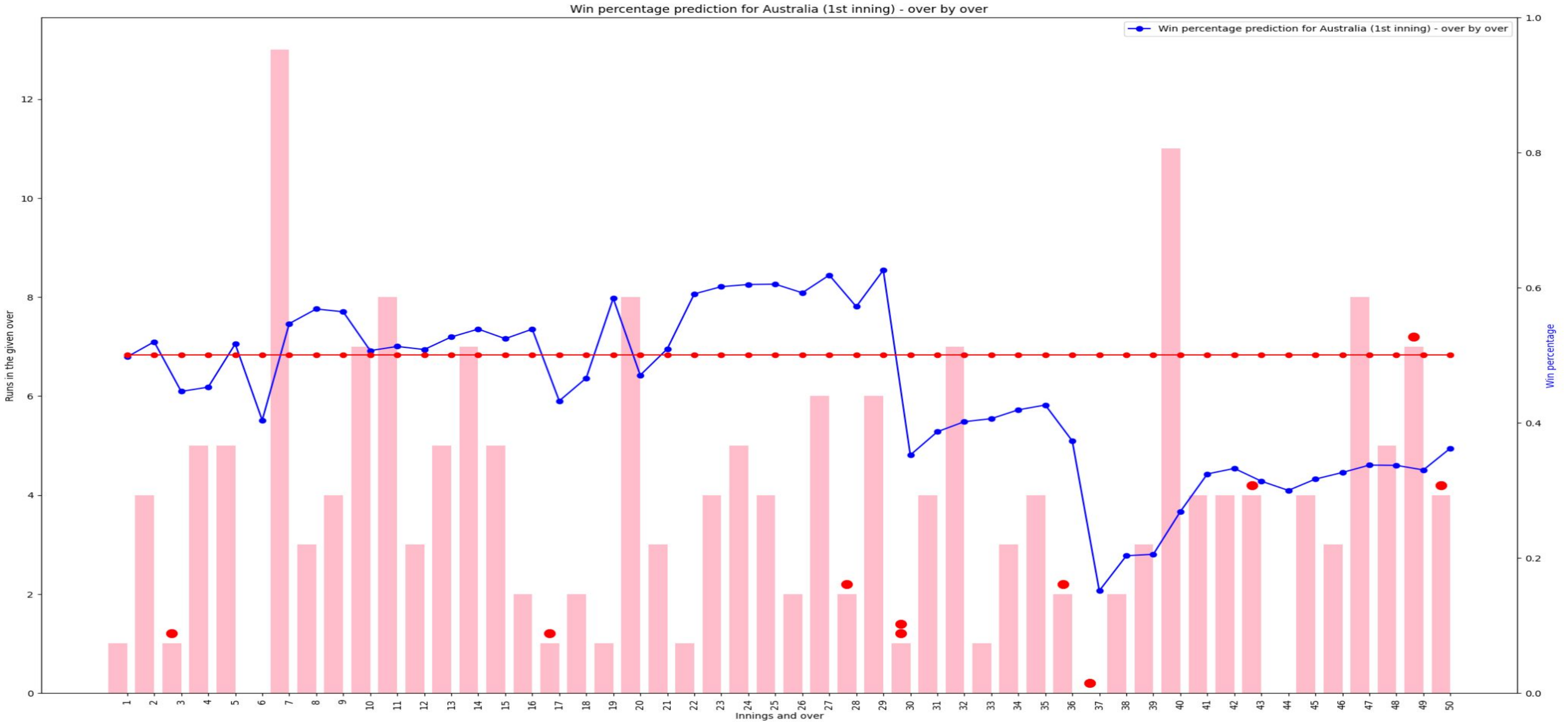
Results

Event : ICC ODI World cup 2023 league matches

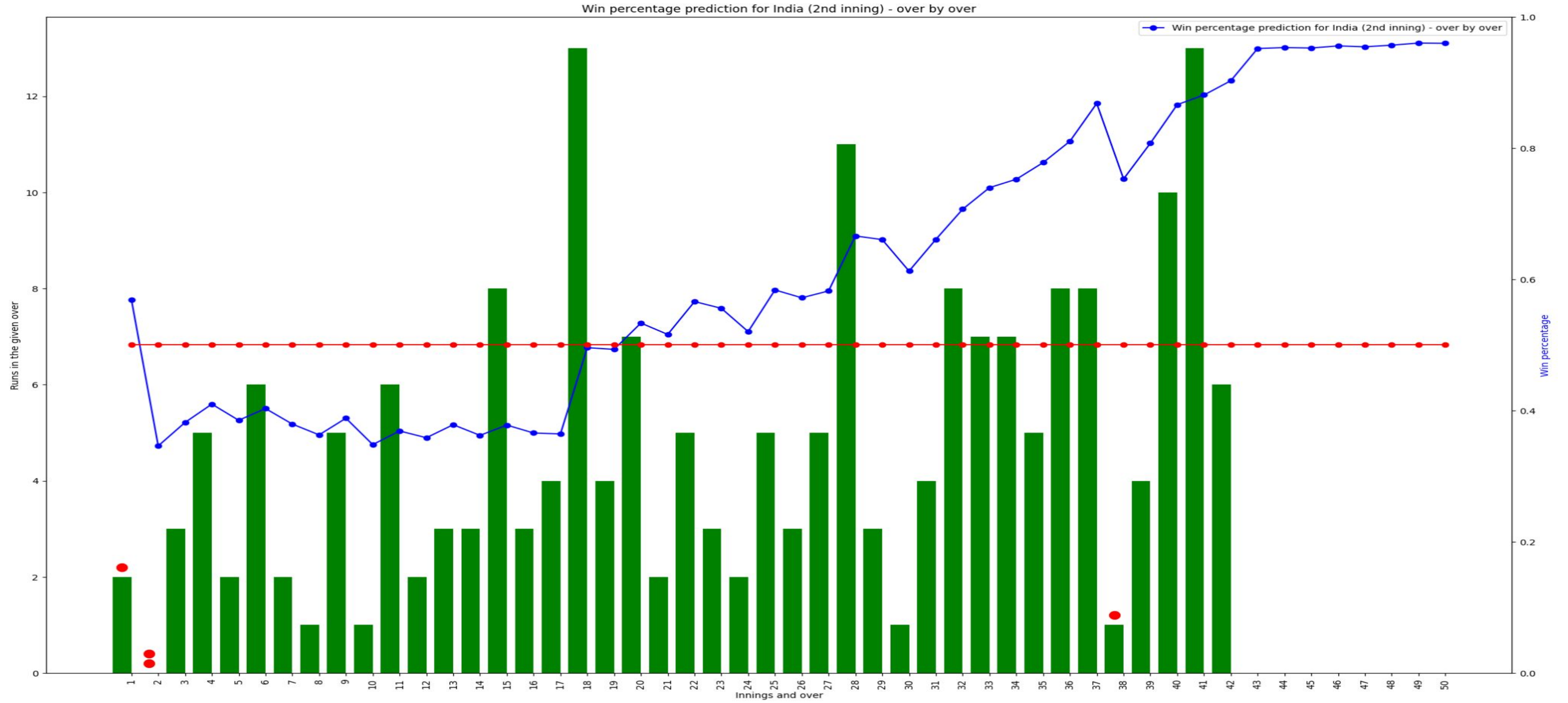
Match 1: India vs Australia (League match)

Match 2: Australia vs Afghanistan

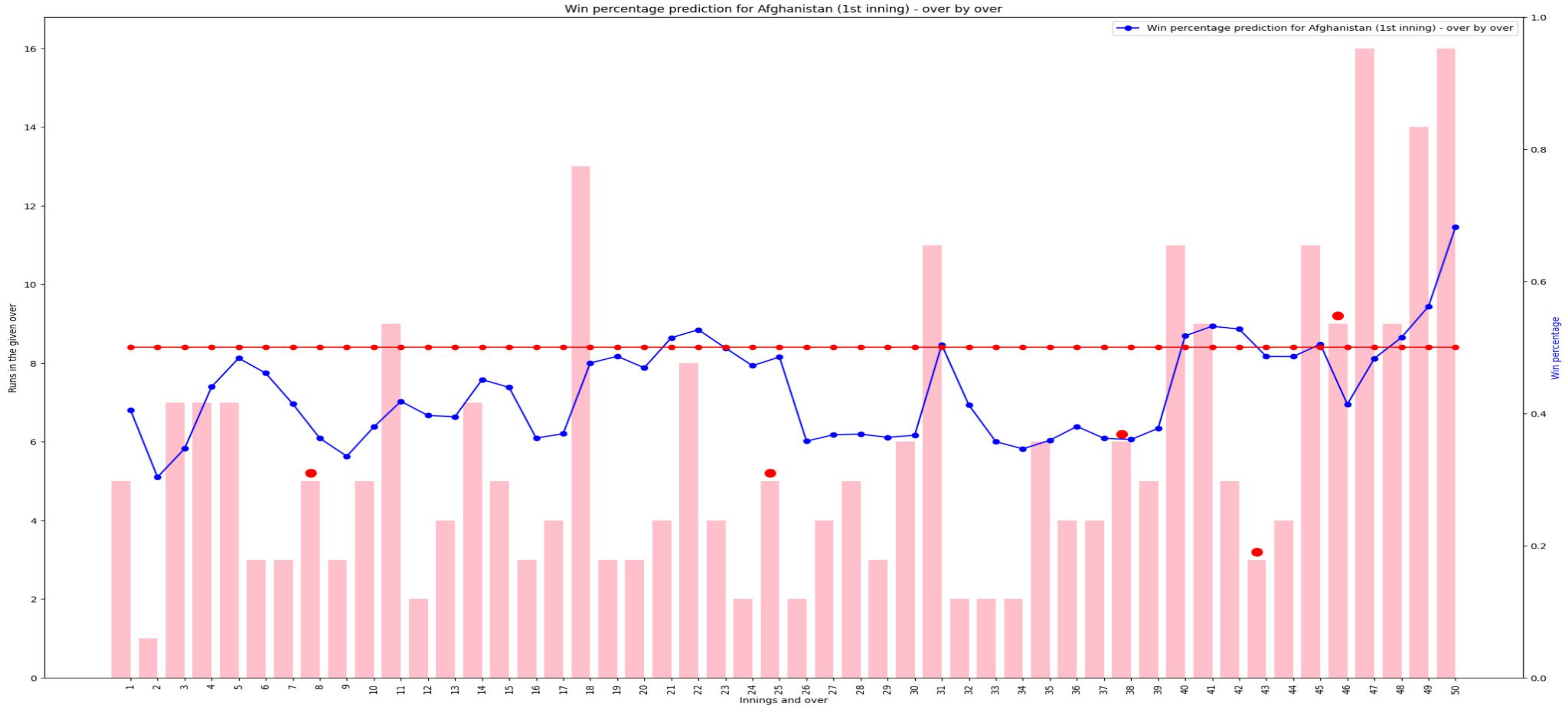
XGBoost : Match 1, 1st Inning (AUS Batting)



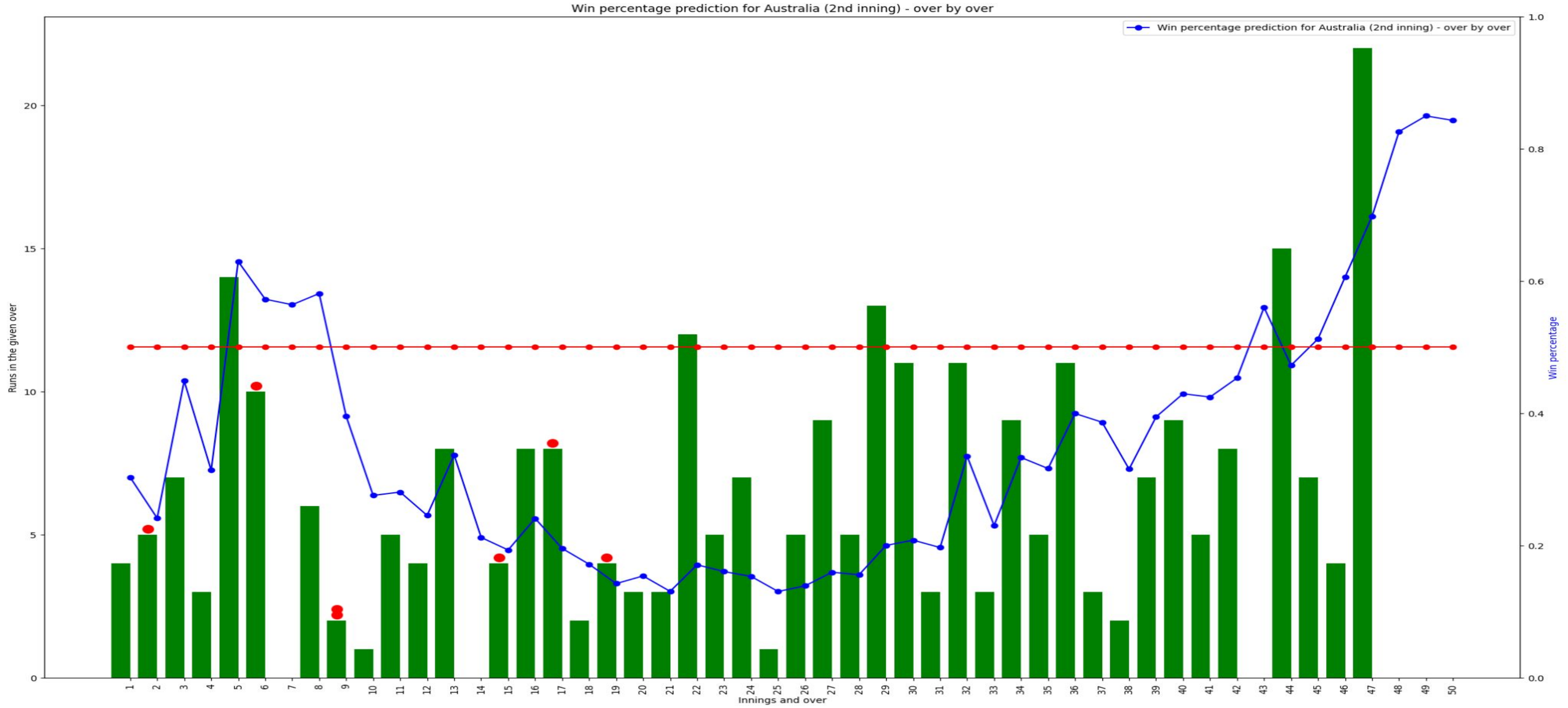
XGBoost : Match 1, 2nd Inning (IND Batting)



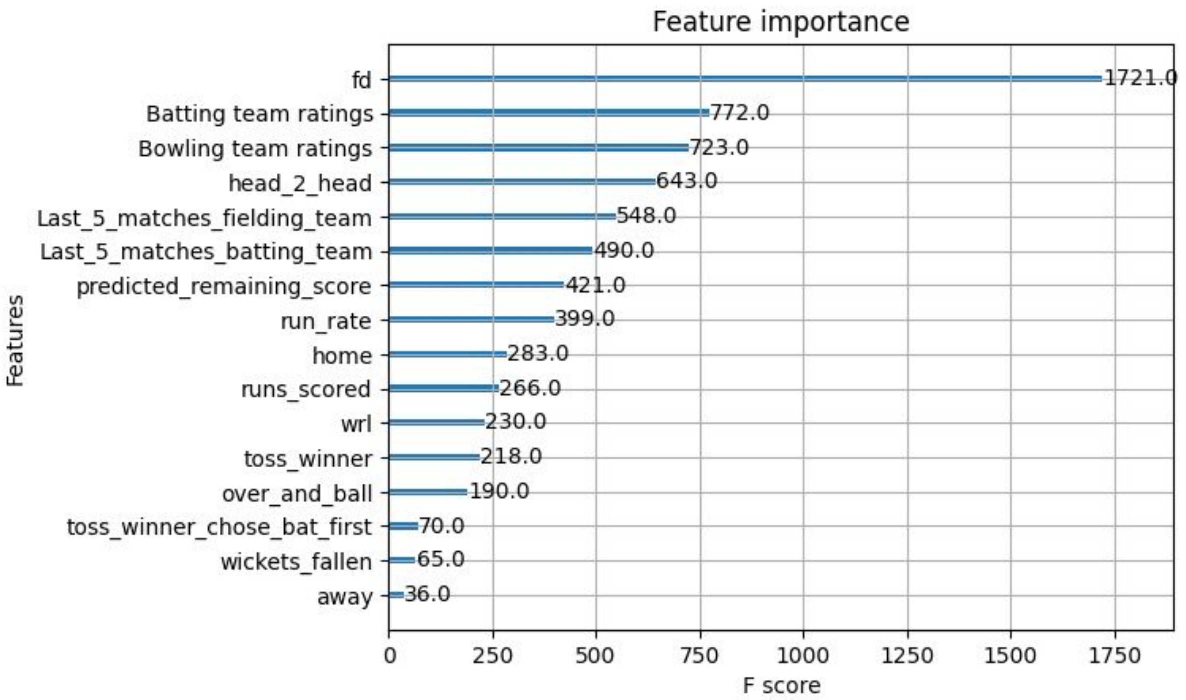
XGBoost : Match 1, 1st Inning (AFG Batting)



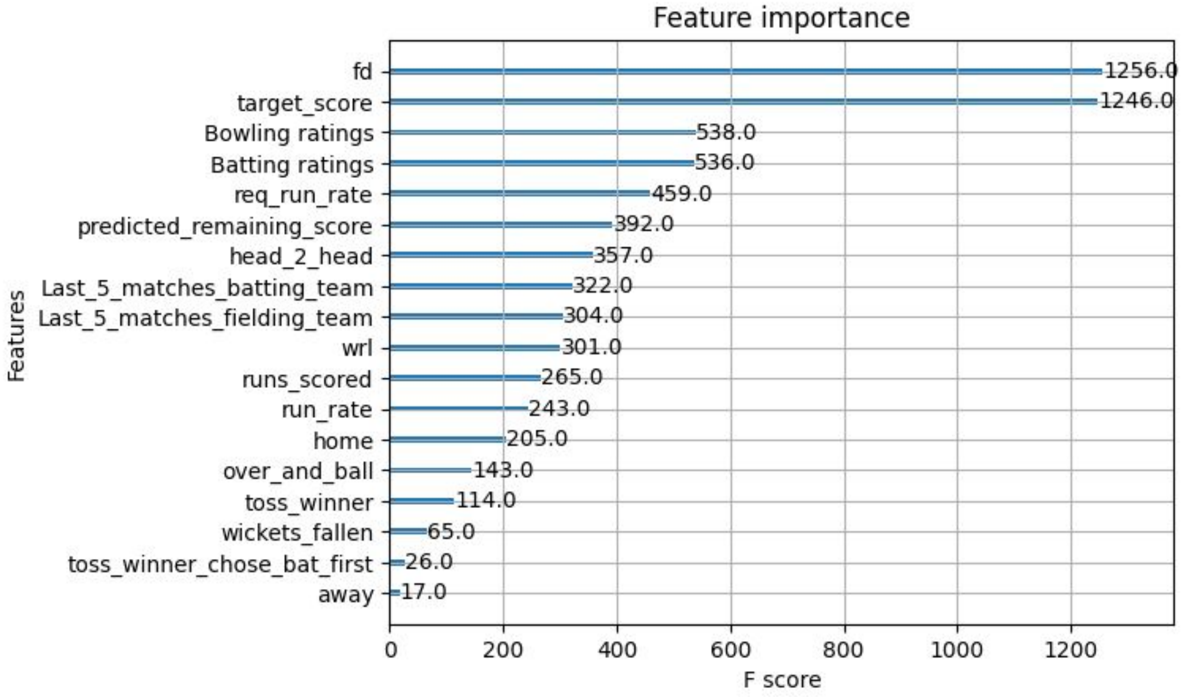
XGBoost : Match 2, 2nd Inning (AUS Batting)



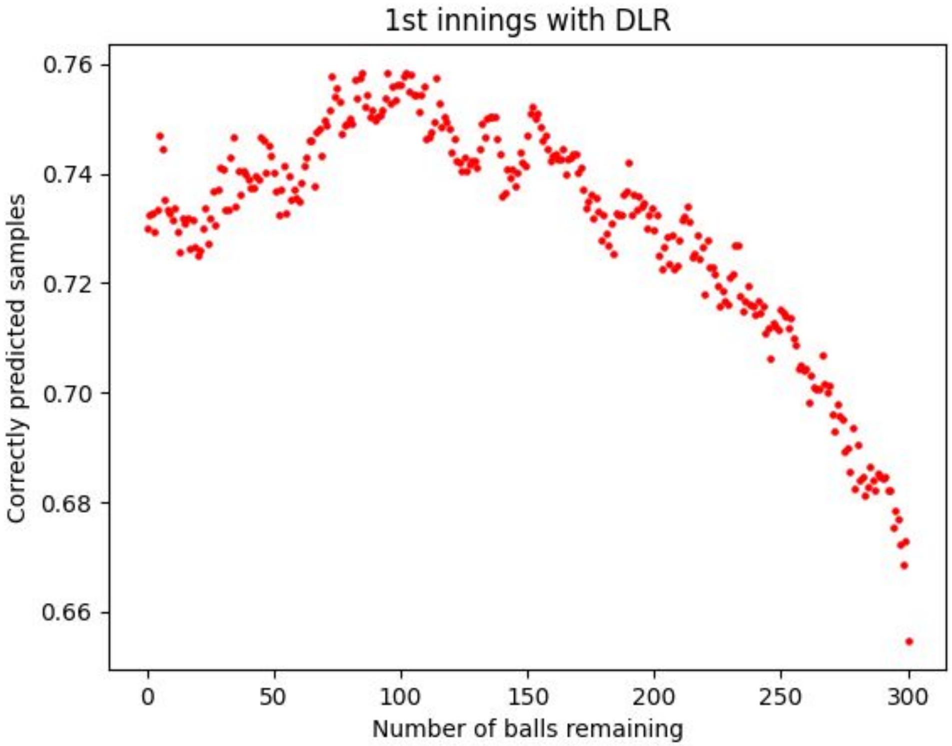
Feature Importance for 1st inning of XGBoost

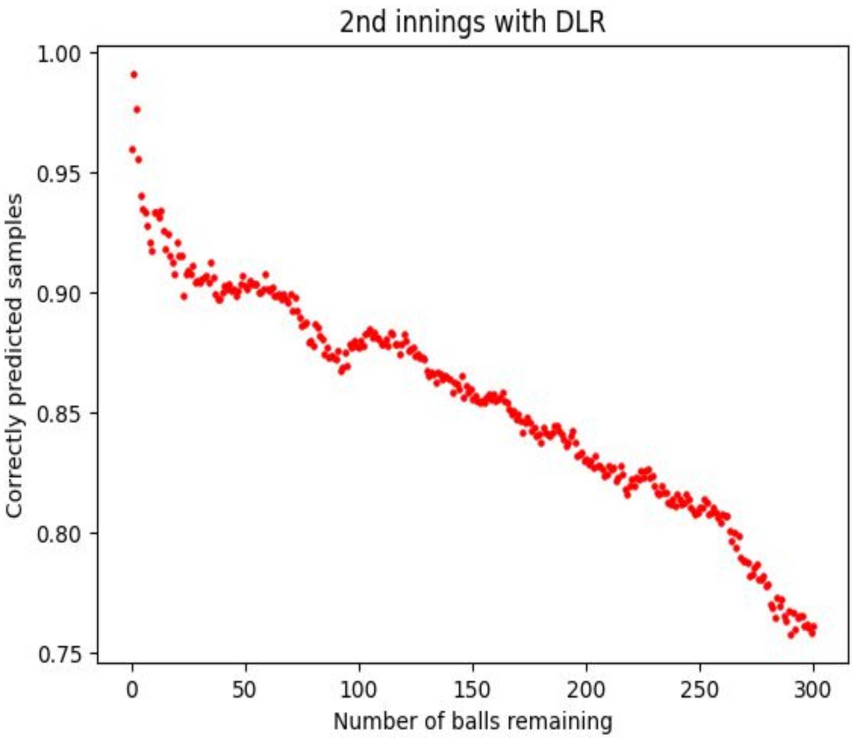


Feature Importance for 2nd inning of XGBoost



Interesting Observations with DLR method





Conclusion

→ The best model came out to be XGboost and the worst was the DLS par score method.

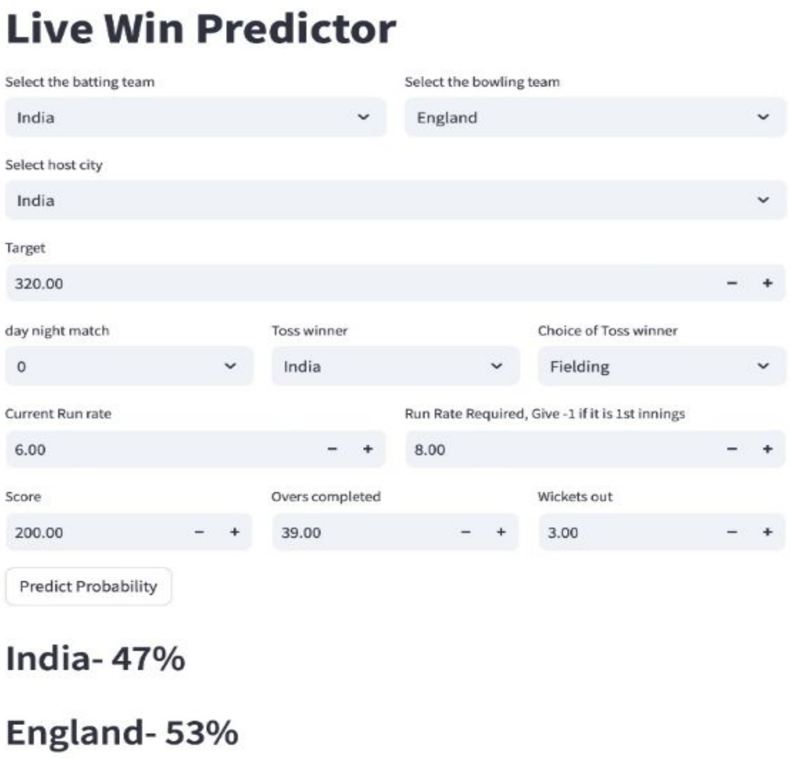
→ The three models namely basic logistic regression, dynamic logistic regression and XG boost depend on quality of the features, hence, depend upon good **feature engineering**.

→ From the importance score, we can observe that features wicket resource lost(wrl), form difference(fd) and predicted score significantly improved the results of XGboost model.

Future work

- Additional features, such as **partnership dynamics**, **pitch conditions, current ICC rankings of players, and historical data on interactions between bowlers and batsmen**, were not incorporated in this iteration but are acknowledged as potential contributors to enhanced predictive accuracy. Future iterations of the model could explore the inclusion of these features for more comprehensive and precise win probability predictions.

Front end tool



Thankyou

–Team Fortune Tellers