Khoury College of Computer Sciences
Northeastern University, Boston, MA
December 2022

Indian Premier League (IPL) Win Predictor

GROUP 12
Mahesh Babu Kommalapati
Sharanya Senthil
Sameer Prasad Koppolu

1 Summary:

1.1 Overview:

Twenty20 Cricket, abbreviated as T20, is a popular format of the sport of Cricket that originated in England in 2003. In recent years T20 appeals to millions of spectators across the globe as it is shorter than the classical cricket match (T20 matches feature just 20 overs per innings) but still manages to hold the thrill of the sport to a high degree. The Indian Premier League(IPL) is a T20 domestic cricket league held annually in India. In this context, we are focusing on predicting the winning team of a given match in the IPL, with respect to the players participating in the match.

1.2 Goals:

Our aim with this project is to predict the winning team for a given matchup in the 2022 IPL season. The winning team is chosen based on team impact scores that are calculated from the impact scores of the individual players that make up the team for a given matchup. In addition to the impact scores, the home team, the away team, the venue of the match, the winner of the toss, and their decision to bat or bowl first are also considered. The seasons from 2009 to 2021 form the training set while the 2022 season alone is the test set. Various machine learning models will be trained and test and their results will be reported. Additionally, a webpage is created that allows users to pick and choose their team and opposition from which the winning team is predicted.

1.3 Data Description:

The IPL T20 datasets were obtained from Kaggle.

The 'all_season_summary.csv' dataset holds the summary of every match in every season with 959 observations. It consists of:

1. **season**: The season in which the match was played represented by the year (example 2019, 2020, 2021, 2022). The values of this variable range from 2008 to 2022.
2. **id**: A unique match Identification number (ID) for every match that has occurred from 2008 to 2022. This variable is the Primary Key of the 'all_season_summary' dataset.
3. **name**: This variable holds the names of the two teams playing against one another in a given match in the format of $Team\ A\ v\ Team\ B$.
4. **short_name**: This variable holds the names of the two teams playing against one another in a given match. However, the name of each team is abbreviated using 2 to 4 characters in the following format
$< Team\ A\ Abbreviation > v < Team\ B\ Abbreviation >$.
5. **home_team**: The Abbreviated name of the Home Team of a given match.
6. **away_team**: The Abbreviated name of the Away Team of a given match.
7. **toss**_won: The Abbreviated name of the winner of the toss of a given match.
8. **decision**: The decision made by the toss winner to BOWL FIRST or BAT FIRST.

9. **winner**: Abbreviated name of the winning team of a given match.

10. **venue**: The stadium at which the match was played.

11. **home_playx1**: The variable containing the list of eleven players representing the home team in a given match.

12. **away_playx1**: The variable containing the list of eleven players representing the away team in a given match.

The ’all_season_batting _card.csv’ dataset has 14540 observations where every observation represents a specific batsman who played in a given match. From this dataset, the following variables are considered.

1. **season**: Same as that of the ’all_season_summary.csv’ dataset.

2. **match_id**: A unique number (ID) for every match that has occurred from 2008 to 2022 in which the given batsman has played.
   This variable is the Foreign Key of the ’all_season_batting _card.csv’ dataset.

3. **fullName**: This variable holds the full name of a batsman who batted in a given match.

4. **runs**: This variable holds the number of runs scored by a batsman in a given match.

5. **ballsFaced**: This variable holds the number of balls played by a batsman in a given match.

6. **isNotOut**: A flag variable that holds the boolean value True or False to indicate whether or not a batsman was dismissed.

7. **fours**: This variable holds the number of fours that a batsman has hit in a given match.

8. **sixes**: This variable holds the number of sixes that a batsman has hit in a given match.

The ’all_season_bowling _card.csv’ dataset has 11224 observations where every observation represents a specific bowler who bowled in a given match. From this dataset, the following variables are considered.

1. **season**: Same as that of the ’all_season_summary.csv’ dataset.

2. **match_id**: A unique match Identification number (ID) for every match that has occurred from 2008 to 2022 in which the given bowler has bowled in. This variable is the Foreign Key of the ’all_season_bowling _card.csv’ dataset.

3. **fullName**: This variable holds the full name of a batsman who bowled in a given match.

4. **overs**: This variable holds the number of overs bowled by a bowler in a given match.

5. **conceded**: This variable holds the number of runs conceded by a bowler in a given match.

6. **wickets**: This variable holds the number of wickets earned by a bowler in a given match.

7. **dots**: This variable holds the number of balls bowled by a bowler in a given match where no runs were conceded by the bowler.

8. **foursConceded**: This variable holds the number of fours conceded by a bowler in a given match.

9. **sixesConceded**: This variable holds the number of sixes conceded by a bowler in a given match.

2 Methods:

2.1 Data Preprocessing:

Three datasets were used for this project, (all_season_summary.csv, all_season_batting _card.csv, and all_season_bowling _card.csv). As part of the data preprocessing, additional parameters were calculated for batsmen and bowlers in order to arrive at their respective impact scores. Once, the impact scores for both bowlers and batsmen are calculated, their respective relative impact scores are calculated. Thereafter, the relative impact scores of each player for a given season are aggregated with respect to the team that they play for in a given matchup in a given season. This results in the Team Impact Score. Following this, the data is partitioned into a training and test set. The training set ranges from 2009 to 2021 seasons, while the test set is only the 2022 season as the aim is to predict the winner of every matchup in the 2022 season. Finally, different Machine Learning models are used and the accuracy of each model on the test set is reported.

2.1.1 Evaluating Batting Impact

For a given player, the Batting Impact Score is calculated for every season. It is calculated using the following parameters:

- **Batting Strike Rate**: This parameter depicts how fast a batsman scores runs. For a given season, the runs scored and balls faced by a batsman are aggregated and then used to calculate the Batting Strike Rate for that batsman for the season overall.
  Batting Strike Rate = 100 * (Runs Scored)/(Balls Faced)

- **Batting Average**: This parameter depicts how much a batsman scores on average before getting out. For a given season, the runs scored and the number of times a batsman has gotten out are aggregated and then used to calculate the Batting Average for that batsman for the season overall.
  Batting Average = (Runs Scored)/(Number of times Dismissed)

- **Quality**: This parameter depicts the performance of a batsman in a given season relative to the average batsman who has a Batting Strike Rate of 130 and a Batting Average of 30 per season.
  Quality = (Batting Strike Rate + Batting Average) / (Average Batsman Strike Rate + Average Batsman Batting Average)

- **Frequency**: This parameter represents the average number of balls faced by a batsman in a given season. It is calculated as follows.
  Frequency = (Balls Faced in a Season) / (Matches Played in a Season)

- **Batting Player Impact**: This is a preliminary impact score that is calculated using Quality and Frequency. It represents the impact a batsman has in a season. It is given by the following formula.
  Batting Player Impact = Quality * Frequency

- **100s Factor**: The Hundreds Factor represents the average number of Hundreds a batsman has scored in a given season.

- **50s Factor**: The Fifties Factor represents the average number of Fifties a batsman has scored in a given season.

- **Batting Impact Score**: This is the final Batting Impact score of a batsman in a given season. It is calculated as follows.
  Batting Impact Score: Batting Player Impact + (Batting Player Impact * 100s Factor) + (Batting Player Impact * 50s Factor/2)

2.1.2 Evaluating Bowling Impact

For a given player, the Bowling Impact Score is calculated for every season using the following parameters.

- **Bowling Economy**: This parameter depicts the rate at which a bowler concedes runs. For a given season, Bowling Economy Rate is given as follows.
  Bowling Economy Rate = (Runs Conceded by Bowler in a Season)/(Overs Bowled by Bowler in a Season)

- **Bowling Strike Rate**: This parameter depicts the rate at which a bowler earns wickets. For a given season, Bowling Strike Rate for a bowler is as follows.
  Bowling Strike Rate = (Balls Bowled by Bowler in a Season)/(Wickets Earned by Bowler in a Season)

- **Quality**: This parameter depicts the performance of a bowler in a given season with respect to the average bowler. It is given as follows.
  Quality = 100/(Bowling Economy * Bowling Strike Rate)
  Here, 100 is the Quality of the Average Bowler. The Average Bowler has a Bowling Economy of 8 and a Bowling Strike Rate of 12. The product of these two parameters is then rounded off to 100.

- **Frequency**: Frequency for a bowler depicts the average number of balls that a bowler has bowled in a given season. It is given as follows.
  Frequency = (Balls Bowled in a Season)/(Matches Played in a Season)

- **Bowling Player Impact**: This is a preliminary impact score that is calculated using Quality and Frequency. It represents the impact a bowler has in a season. It is given by the following formula.
  Bowling Player Impact = Quality * Frequency

- **4 Wicket Factor**: This parameter depicts the average number of 4 wicket hauls that a bowler has earned in a given season. A four-wicket haul happens when a bowler earns 4 wickets in a single match.

- **5 Wicket Factor**: This parameter depicts the average number of 5 wicket hauls that a bowler has earned in a given season. A five-wicket haul happens when a bowler earns 5 wickets in a single match.
- **Bowling Impact Score**: This is the final Bowling Impact score of a bowler in a given season. It is calculated as follows.
  Bowling Impact Score = Bowling Player Impact + (Bowling Player Impact * 5WF) + (Bowling Player Impact * 4WF/2)
  Here 5WF and 4WF are the 5 Wicket and 4 Wicket Factors respectively.

2.1.3 Evaluating Relative Player Impact

- We use Relative Player Impact for Batsmen and Bowlers to calculate a Team's Impact Score based on the playing eleven that is present in a given match in a given season.
- To calculate the Relative Batting/Bowling Impact of a player in a given season, we first calculate the Modified Player Impact of that player, where the Modified Batting/Bowling Impact of a player is obtained by simply considering the Batting/Bowling Player Impact Score of a player in the i^{th} season to be that of the $(i-1)^{th}$ season.
- Additionally, the 2008 season of the IPL is dropped as it was the very first season of the league. All missing values for the Modified Batting/Bowling Impact of a player are imputed using the mean of the already calculated values of the Batting/Bowling Impact of that player.
- Relative Batting/Bowling Impact score for each player in each season is calculated with its respective modified impact scores in order to consider the change in it across the seasons that the player has participated in.

$$Let\ M_i\ =\ Modified\ Batting/Bowling\ Player\ Impact\ Score\ of\ i^{th}\ Season$$

$$Let\ R_i\ =\ Relative\ Batting/Bowling\ Player\ Impact\ Score\ of\ i^{th}\ Season$$

$$Therefore,$$

$$R_i\ =\ \begin{cases} R_{i-1} + M_i & ,\ if\ M_i > M_{i-1} \\ R_{i-1} - \left|M_i - M_{i-1}\right| & ,\ if\ M_i < M_{i-1} \end{cases}$$

- Upon obtaining the Modified Batting/Bowling Impact score for each player in each season, we then calculate the Relative Batting/Bowling Impact score for each player in each season. This is done because we need to take into account any increase or decrease in the Modified Batting/Bowling Impact score for each player across the seasons of the IPL that the player has participated in.
- For the first season that the player has participated in (excluding the 2008 season), the Relative Batting/Bowling Impact Score is calculated as follows. The Top 10 Players with the highest Relative Batting Impact Scores are given in Figure 1 in the Appendix. The Top 10 Players with the highest Relative Bowling Impact Scores are given in Figure 2 in the Appendix.

2.1.4 Predictor and Response Variables

- As mentioned earlier, the Team Impact Score for Batting and for Bowling for a given season is calculated as an aggregation of the Relative Batting/Bowling Impact scores for each player on that team in that season. However Relative Batting and Bowling Impact scores have a right-skewed distribution as shown in Figures 3 and 4 in the Appendix.
- Log Transformations on the Relative Batting and Bowling Impact scores decrease the skewness as shown in Figures 5 and 6 in the Appendix. Thereafter, the transformed Relative Batting and Bowling Impact scores are summed up for each player in each team for a given season to give the Team Impact Scores.
- The next set of predictors are the home and away teams, toss winner, toss decision, and venue. These are categorical variables and are One Hot Encoded. The dependent variable is 'winner' and this variable is Label Encoded.
- Observations containing the teams Royal Challengers Bangalore (RCB)^{1}, Chennai Super Kings (CSK)^{2}, Rajasthan Royals (RR)^{3}, Delhi Capitals (DC)^{4}, Sunrisers Hyderabad (SRH)^{5}, Punjab Kings (PBKS)^{6}, Kolkata Knight Riders (KKR)^{7}, Mumbai Indians (MI)^{8}, Gujarat Titans (GT)^{9}, and Lucknow Super Giants (LSG)^{10} are kept because these teams are the most consistently participating teams in the IPL since 2008.
- GT^{9} and LSG^{10} are considered because these teams, although new, take part in the 2022 IPL season on which we predict the winner of each matchup. The remaining teams from the dataset are not considered as they have participated irregularly across the IPL seasons. The number of matches played by the consistent teams from 2008 to 2022 is shown in Figure 7 in the Appendix. When LSG^{10} or GT^{9} participates, their impact score is considered along with the difference with the corresponding opposition. If this difference is greater than 3.5, (value chosen through trial and error) then the team with a higher score is expected to win.
- The Top 10 venues in terms of the number of matches held in which the consistent teams have participated are present in Figure 8 in the Appendix.

2.1.5 Data Partitioning

The 'all_season_summary.csv' dataset is partitioned into a training and test set after the impact scores for the home teams and away teams are calculated. Thereafter, the dataset is partitioned. The training set contains observations from the 2009 season till the 2021 season while the test set contains only the observations of the 2022 season.

2.2 Data Modelling:

2.2.1 Logistic Regression Model:

- Logistic regression models the probability that the target belongs to a particular category. The logistic function is p(X) = $e^{\beta_0+\beta_1 X}/(1+e^{\beta_0+\beta_1 X})$
- It can be written as P(Y=1|X) or P(Y=0|X) where P(Y|X) is approximated as the sigmoid function applied to a linear combination of the input features.

2.2.2 Random Forest Classifier:

- Random forest is an ensemble method that uses the divide and conquer approach on decision trees generated on a randomly split dataset.
- Each tree depends upon the independent random sample. Each tree votes, and the most popular class is chosen as the final result.
- The prediction result with the most votes is chosen as the final prediction.

2.2.3 AdaBoost Classifier:

- AdaBoost is a boosting algorithm that improves the prediction power by training a set of weak models(stump/slow learner - just one node with two leaves), by compensating the weakness of its predecessors.
- One stump can have more priority than the other in the final classification. There might also be chances that independent variables may predict the classification at a higher rate when compared to the other variables.

2.2.4 Extreme Gradient Boost Classifier:

- XGBoost concentrates on the distribution of features across all data points in a leaf thereby reducing the search space of possible feature splits.
- It boosts the model performance and the execution speed.

3 Results:

- The following tables represent the integrated results of all the models :

Models along with accuracy	
Model	Accuracy
AdaBoost Classifier	0.62
Logistic Regression	0.59
Random Forest Classifier	0.56
XGBoost Classifier	0.55

- The AdaBoost Classifier has the highest accuracy with the following Classification Report and an F1 score of 0.59.

Classification Report of AdaBoost Classifier			
Team	Precision	Recall	F1-Score
Chennai Super Kings (CSK)	0.50	0.75	0.60
Delhi Capitals (DC)	0.67	0.57	0.62
Gujarat Titans (GT)	0.73	0.92	0.81
Kolkata Knight Rider (KKR)	0.40	0.33	0.36
Lucknow Super Giants (LSG)	0.73	0.89	0.80
Mumbai Indians (MI)	0.50	0.75	0.60
Punjab Kings (PBKS)	0.67	0.57	0.62
Royal Challengers Bangalore (RCB)	0.55	0.67	0.60
Rajasthan Royals (RR)	0.67	0.20	0.31
Sunrisers Hyderabad (SRH)	0.60	0.50	0.55

- As shown in Figures 9 and 10, the webpage is created on Heroku using Flask and allows the user to pick the home and away teams along with the toss winner, toss decision, and venue from which the winning team is predicted.

4 Discussions:

- The data preprocessing and model building results in predicting the winners of every match in the 2022 IPL season. It also explains the impact of each player by calculating the player impact score.
- After evaluating all the models, the AdaBoost Classifier is the best-fit model for the test data with an accuracy of 62% and an F1-Score of 0.59.
- The best model accuracy obtained is 62% is reasonably good though not great because we do not have historical data to evaluate the performance of the debutant players of the 2022 IPL Season as well the absence of historical data for GT[9] and LSG[10] before 2022.
- In the future, the number of matches played will increase, and thereby retraining the model can give better results.
- Franchise owners, Team management, and bettors of cricket can benefit from this project as they can quantify a player's impact. Results from this project can be utilized by them which can lead to better roster building.

5 Statement of Contribution

- Mahesh Babu Komalapatti: Feature Engineering, Model Building, and Webpage development.
- Sameer Prasad Koppolu: Data Cleaning and Exploratory Data Analysis (EDA).
- Sharanya Senthil: Model Selection and Metrics Evaluation.

6 References:

1. Data Source:
   https://www.kaggle.com/datasets/rajsengo/indian-premier-league-ipl-all-seasons/code

2. Performance metrics were taken from:
   https://www.krick3r.com/2016/02/t20-performance-impact-method.html

3. Research paper:
   https://arxiv.org/pdf/2209.06346.pdf

7 Appendix:

7.1 Code Repository Link:

https://github.com/sameerprasadkoppolu/Group-12-IDMP-Project

7.2 Data Visualizations:

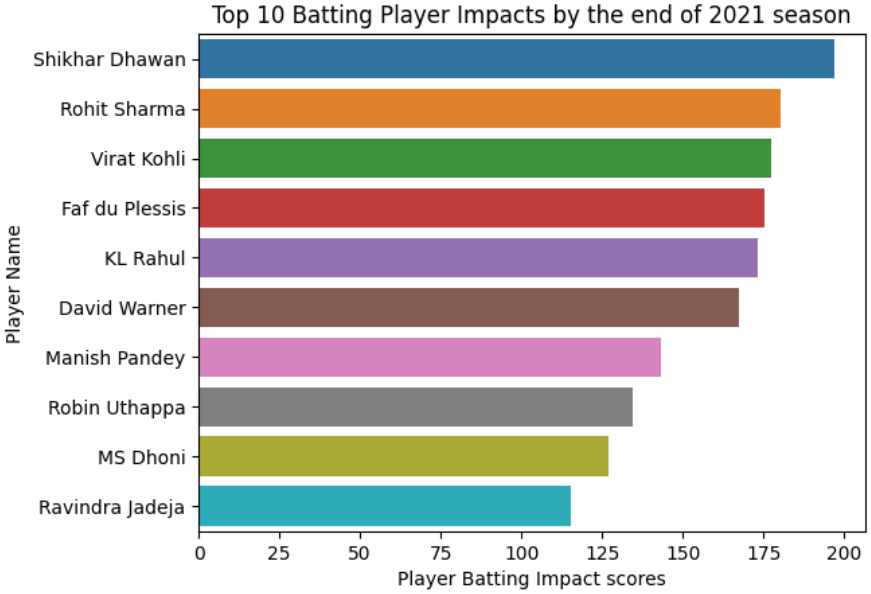


Figure 1: Top 10 Players with the highest Relative Batting Impact Scores

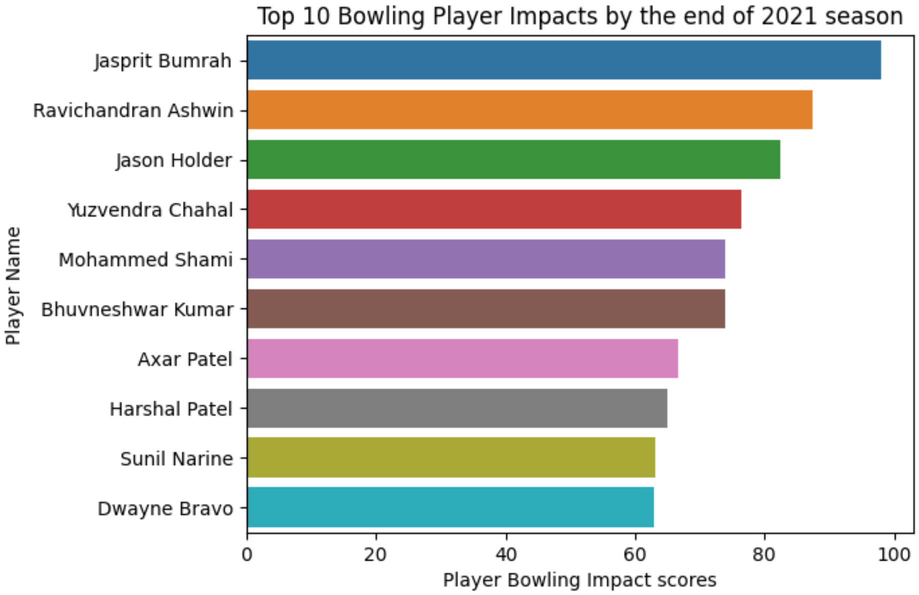


Figure 2: Top 10 Players with the highest Relative Bowling Impact Scores

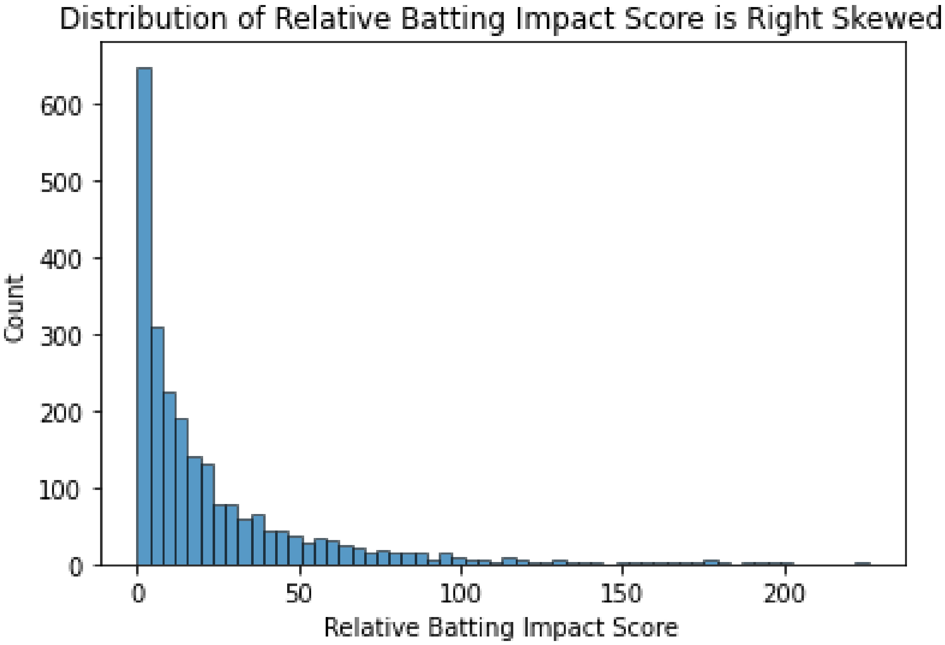


Figure 3: Relative Batting Impact Score Distribution

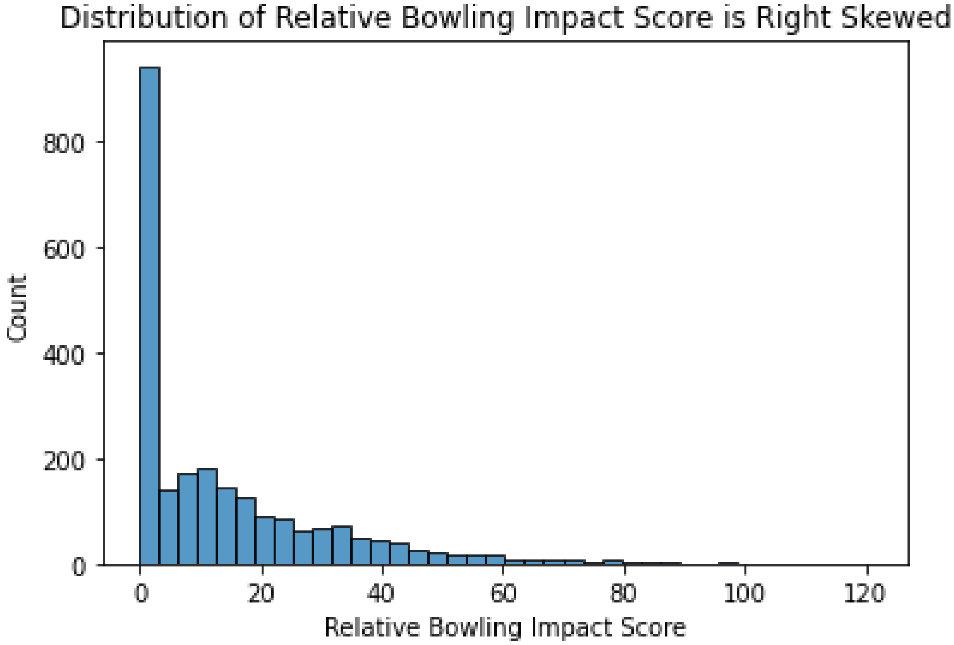


Figure 4: Relative Bowling Impact Score Distribution

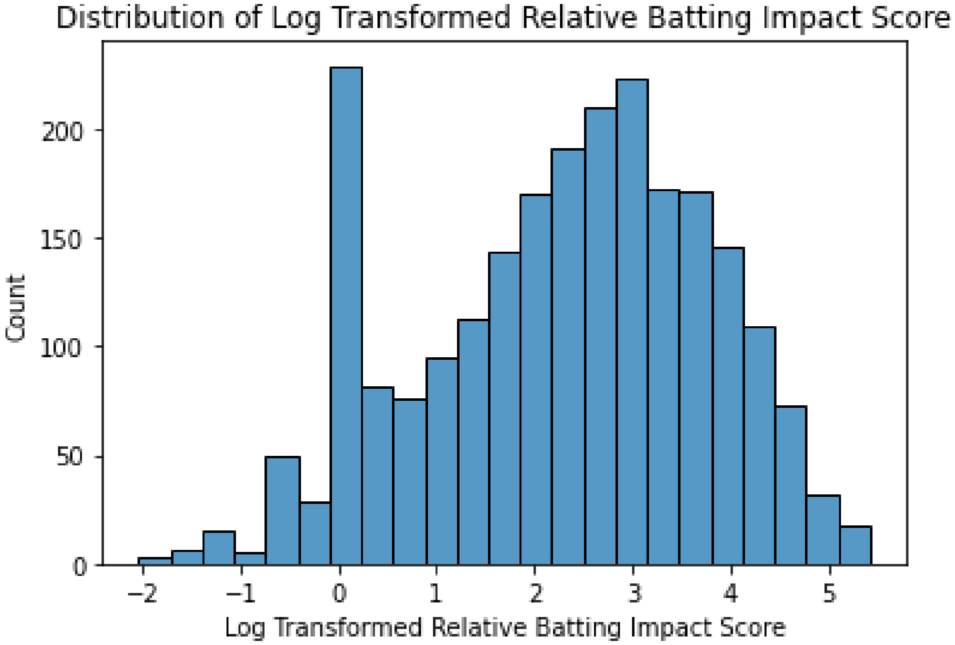


Figure 5: Log Transformed Relative Batting Impact Score Distribution

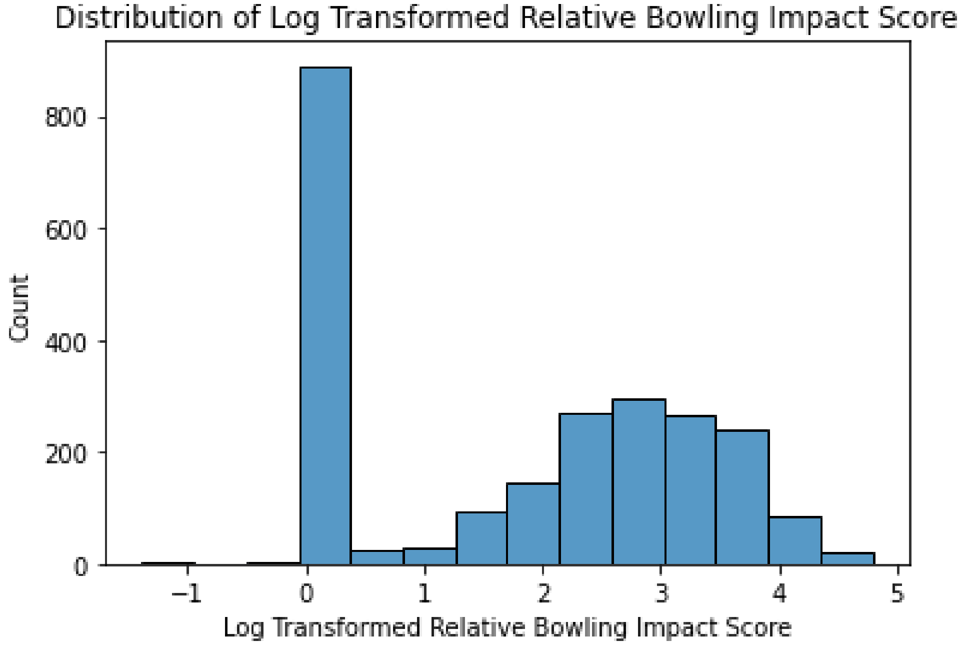


Figure 6: Log Transformed Relative Bowling Impact Score Distribution

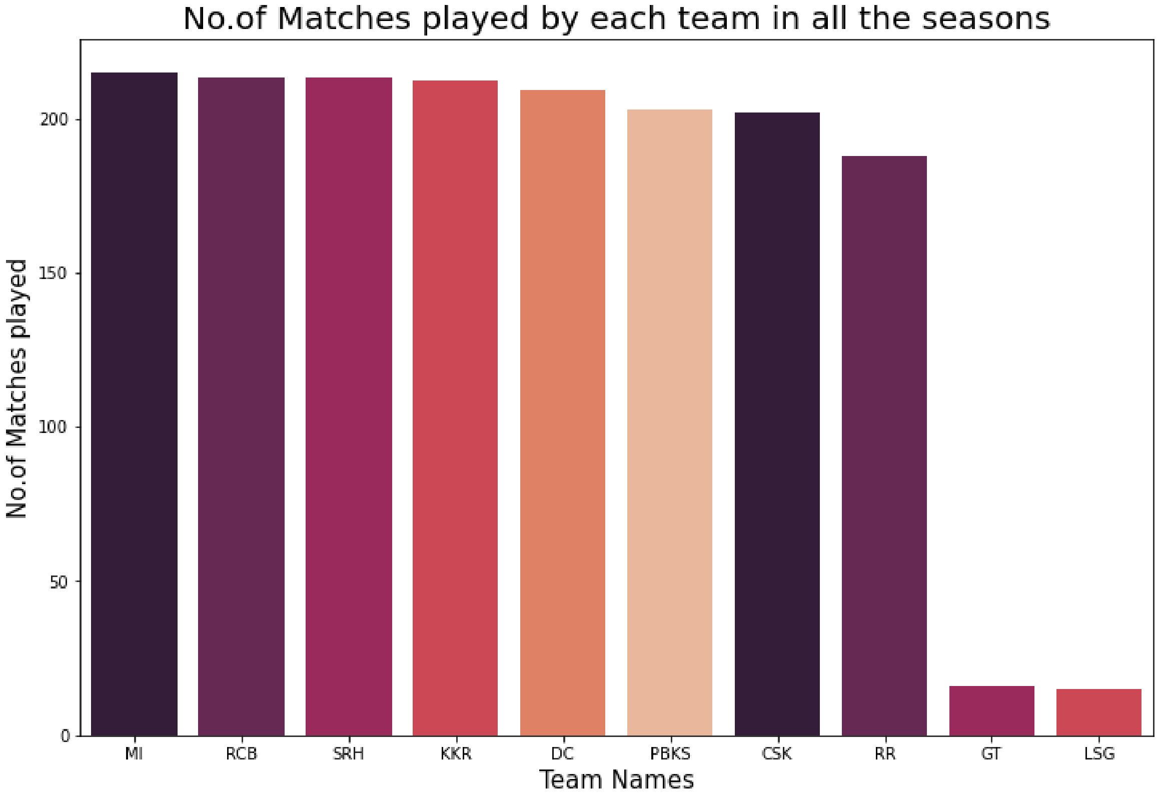


Figure 7: Number of Matches Played by Each Team Across All IPL Seasons (2008-2022)

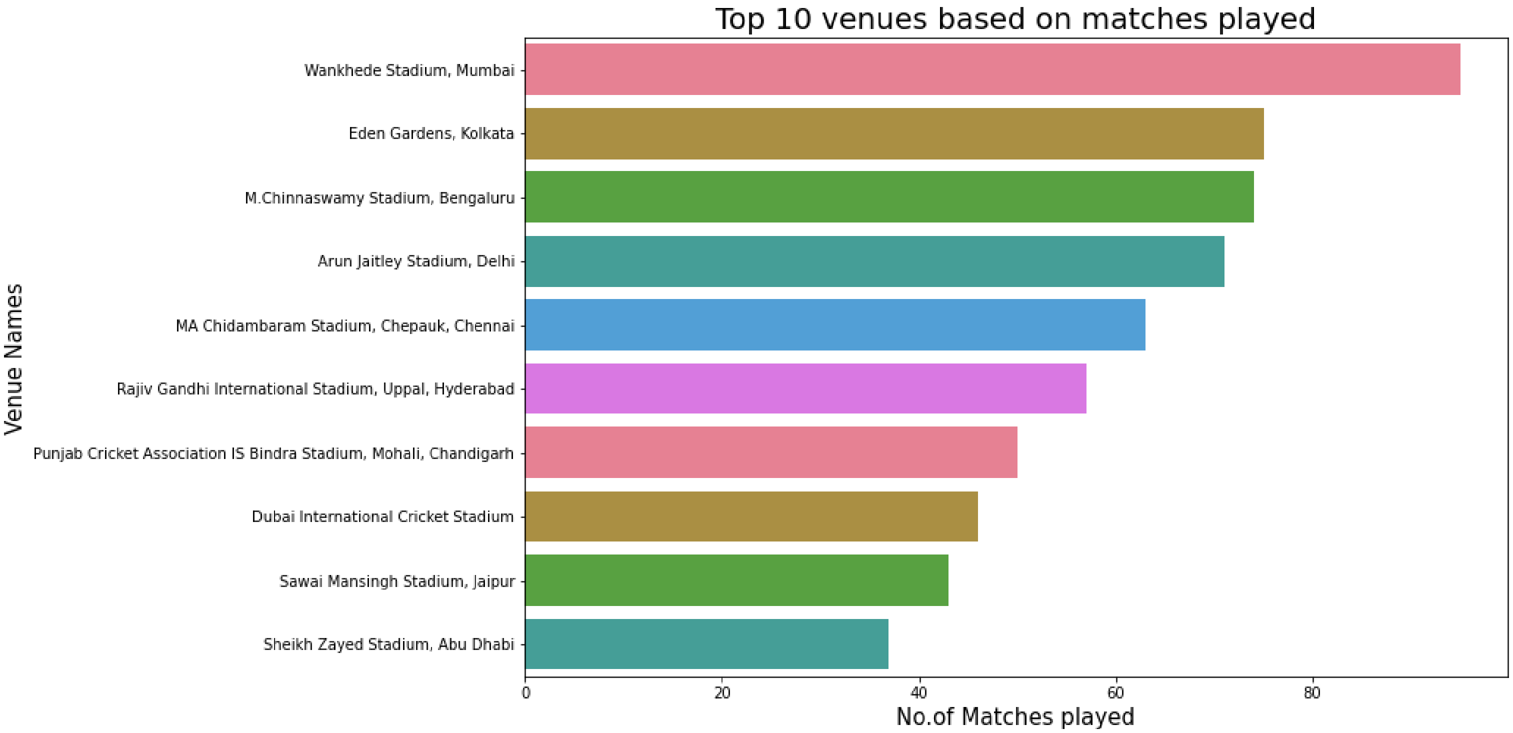
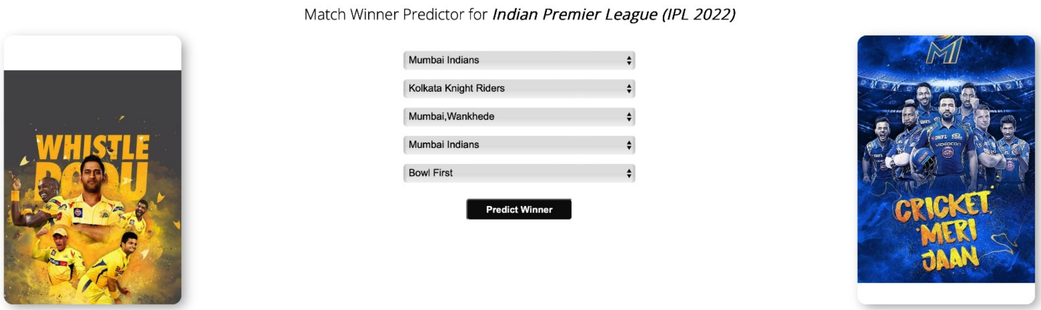


Figure 8: Top 10 Venues with the Highest Number of Matches Played Across All IPL Seasons (2008-2022)



Match Winner Predictor for *Indian Premier League (IPL 2022)*

Mumbai Indians

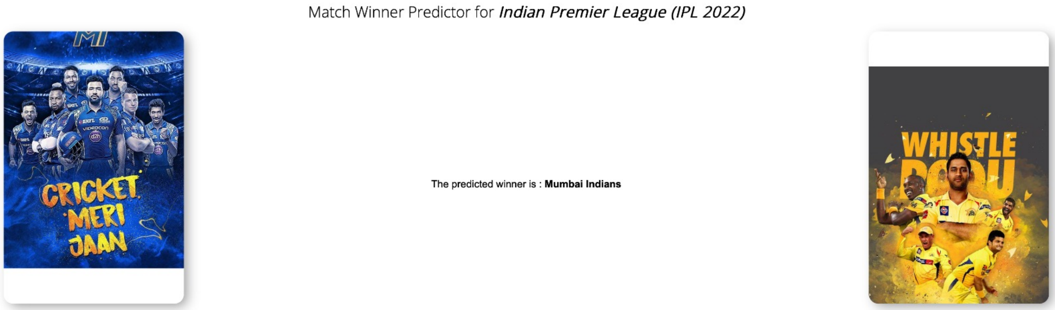
Kolkata Knight Riders

Mumbai,Wankhede

Mumbai Indians

Bowl First

Figure 9: Webpage to Allow User to choose Home Team, Away Team, Toss Winner, Toss Decision, and Venue



Match Winner Predictor for *Indian Premier League (IPL 2022)*

The predicted winner is : **Mumbai Indians**

Figure 10: Webpage that displays the Predicted Winner to the User