Data Analysis of the Indian Premier League

By

Anik Chakraborty

22N0084

Mentors

Vivek Parmar, Ishaan Garg

Introduction

The field of sports analytics is becoming widely popular due to the competitive edge that it can give more opportunities both to sports teams as well as stakeholders involved in the sport and get analytics driven insights. Various data which are available such as the players and team statistics, environment conditions, etc are used to predict models which can help stakeholders make informed decisions on the game. The main objective is to improve the performance of the team and assist in creating strategies which would help the team perfectly counter its opponents. This can be done both prior to a game as well as dynamically as the game progresses. In recent times, it has been observed that the audience themselves are also interested in the data analysis that goes on in the game and hence, sports analysts try to present this data to the audience by making simplifications to it and making use of pictorial elements such as graphs and charts to capture their attention.

About Cricket

Cricket is a sport that is played by two teams, each having eleven members. A team consists of batsmen, bowlers, and all-rounders. The role of the batsmen is to score as many runs as possible in the limited time/overs available, while the bowlers try to restrict the score that the batsmen try to make. All-rounders are players that play both roles and have sufficient expertise in both batting and bowling. The performance of a team depends on various factors such as the constitution of the team in terms of types of players, the venue in which the match is being held, the environmental conditions, and the type of opponents that they're playing against. Data analytics can be made use of to help the teams management figure out which players to play in a specific match, the odds of them reaching a specific stage in a tournament, the environmental conditions that they're going to play in, etc. It can also be used during a match to help the team adjust their strategy according the state at which the match is in, to provide them a competitive edge against their opponent. These days, data science techniques are being made use of by every team that competes in the sport professionally. When used correctly, it can help teams bridge the gap in skill by formulating an effective strategy to counter their opponents.

About IPL

The Indian Premier League (IPL) is the biggest domestic cricket tournament all over the world. It is a 20-over format of the game that makes for short, fast-paced games which is one of the reasons for its massive fanbase. It is an annual tournament and has seen 13 such tournaments conducted so far. There are 8 teams involved in the tournament and the teams themselves consist of players from all around the world. The tournament generates a large revenue and has many stakeholders heavily invested in it. So, teams will do everything they can to get an edge over their opponents in a game. Data Analysis is now heavily used by all teams to try and gain this edge.

Scope and Overview

Section II talks about the Literature Review of the papers and resources referred. Further, in Section III, the datasets used for performing the analysis. Section IV talks about the Analysis Pipeline followed. This paper aims to create a forecasting model for teams to use during the match. Based on the scores data of a team and players at any stage, it tries to predict the final score of the team. The seasons under consideration are the 2008-2020 seasons. Due to the pandemic, the 2020 season was held in the UAE instead of India and provided a considerable challenge for the analysis as the data previously available was for Indian playing conditions which are considerably different from that of the UAE. In addition to this, the teams have changed considerably in their constitution as compared to the past seasons. Section V showcases the results obtained by the predictive models made. Section VI discusses the results obtained and Section VII concludes the paper along with further scope of research.

Datasets

The datasets used for analysis and prediction were collected from kaggle, where the data of all editions of the IPL so far was available. Two datasets have been used. One for overall matches data and one for ball-to-ball data for the full 2008-2020 period. Both the datasets are linked by the 'id' column which represents the matches uniquely. Some of the useful features present in the dataset are date of match, venue, run(s) and wicket (if any) on every ball, toss decision, batsman and bowler, result of match with margin etc. There are some minor discrepancies in data such as missing values in 'bowling team' column and duplicate team name but it doesn't hurt the predictions task as team data is also present in 'team1','team2' columns. The dataset consists of 2 lakh data points with 18 features in total. Data was also collected from www.cricsheet.org [8], which is a relatively small dataset with 76k data points but has useful features like current score, wickets, overs, striker and non-striker runs. These features are crucial to predict the final score of innings and hence is used for this particular prediction task in the work.

Analysis Pipeline

As observed from the literature survey conducted, a large majority of the predictive models that were made are used to predict the outcome of the match and this prediction is made before the start of the match. This prediction will be useful for the team to make long-term decisions for the team to perform better in the tournament as a whole but is not very useful during the match itself as no changes can be made to the team in the middle of a match. The work discussed in this paper seeks to fill in this gap by providing data to the team at various phases of the match so that the team can make informed decisions such as what batting order and bowling order to use for the rest of the game. Firstly, an exploratory analysis of the data is conducted to get a better understanding of what parameters affect the performance of the team as a whole as well as the individual contributions of the players.

Interesting Insights drawn from the datasets

Various manipulations are performed on the available datasets to extract some insightful information from them.

1. Maximum number of wins by any team in each season: In the period 2008-2019 different teams have performed differently. Some teams have performed well in almost every year where other teams lack consistency and performed well in some years but could not match expectations in other years. So here following table represents the top performing team of each season. This list is showing the dominance of Mumbai Indians and Chennai Super Kings in IPL, but if we talk about the dominance of a particular season then the performance of Rajasthan Royals in 2008 is still unmatched.

	year	team	wins
0	2008	Rajasthan Royals	13
0	2009	Delhi Capitals	10
0	2010	Mumbai Indians	11
0	2011	Chennai Super Kings	11
0	2012	Kolkata Knight Riders	12
0	2013	Mumbai Indians	13
0	2014	Kings XI Punjab	12
0	2015	Chennai Super Kings	10
0	2016	Sunrisers Hyderabad	11
0	2017	Mumbai Indians	12
0	2018	Chennai Super Kings	11
0	2019	Mumbai Indians	11

2. Most number of toss wins: We can see among all the teams Mumbai Indians has the highest number of toss wins followed by Kolkata Knight Riders.

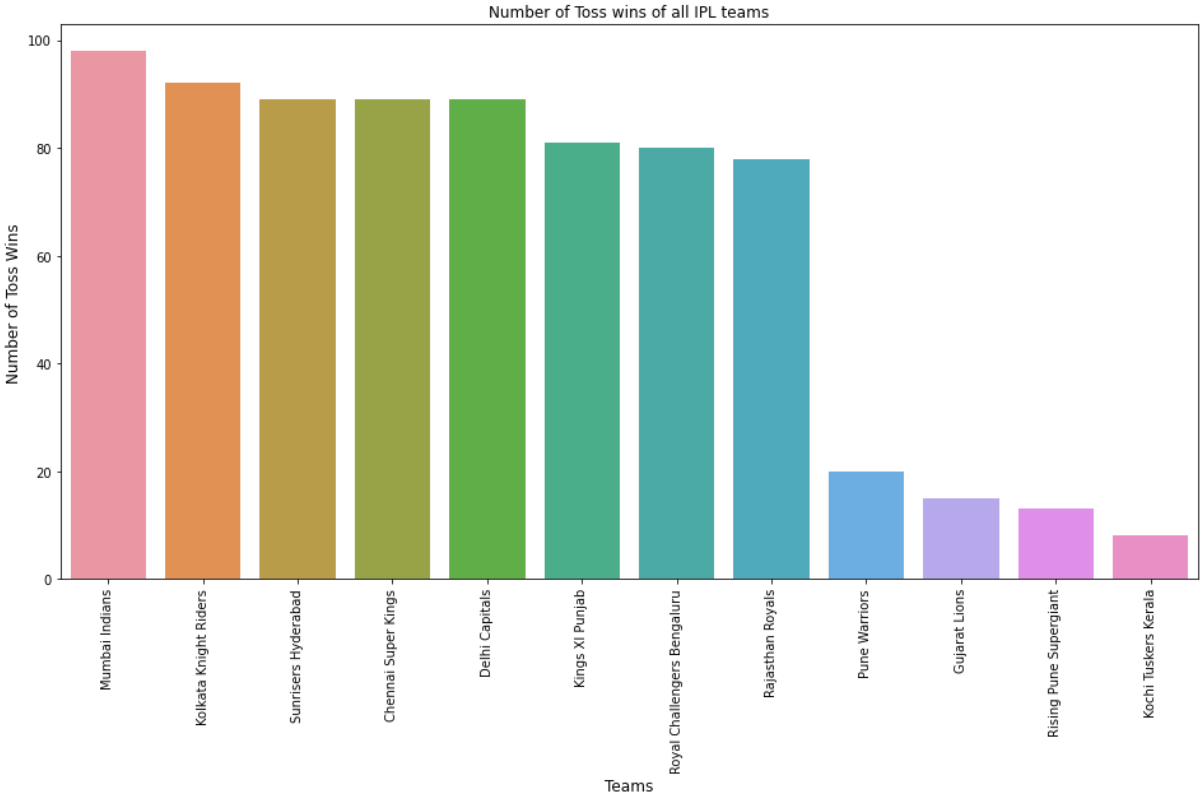

Fig 1. Total toss wins by teams

But since all teams have not played same number of matches, total number of toss wins is not a good measure of success. So, toss win percentage of each team has shown in another graph and it is showing that Chennai Super Kings have the highest toss win percentage followed by Rajasthan Royals. (Kochi Tuskers Kerala has the highest but they have played only one season)

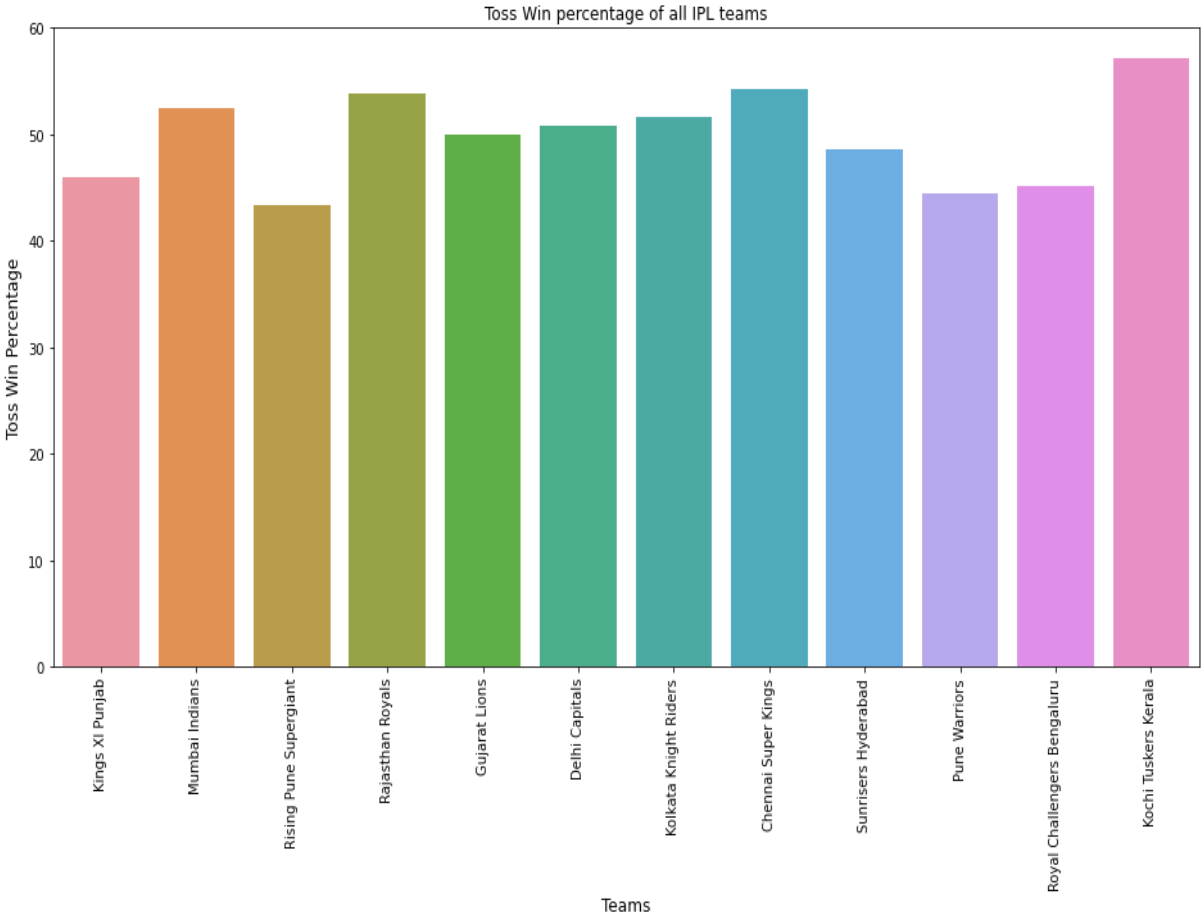

Fig 2. Toss win percentage of all IPL Teams

3. Top 10 greatest victories by runs and by wickets: In 2017 Mumbai Indians registered the biggest victory by 146 runs against Delhi Capitals in Delhi. All 10 biggest victories by runs are with almost more than 100 runs margin. All greatest victories by wickets are of 10 wickets margin.

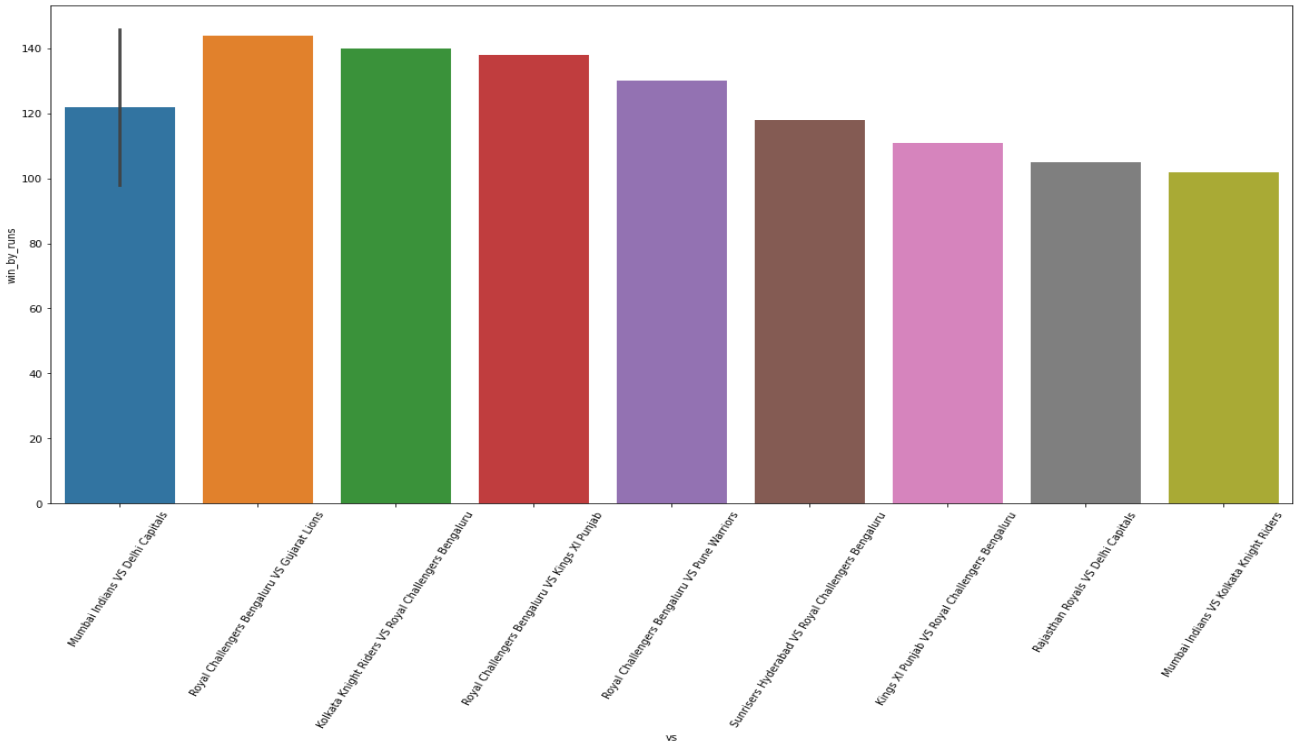

Fig 3. Top 10 wins by runs

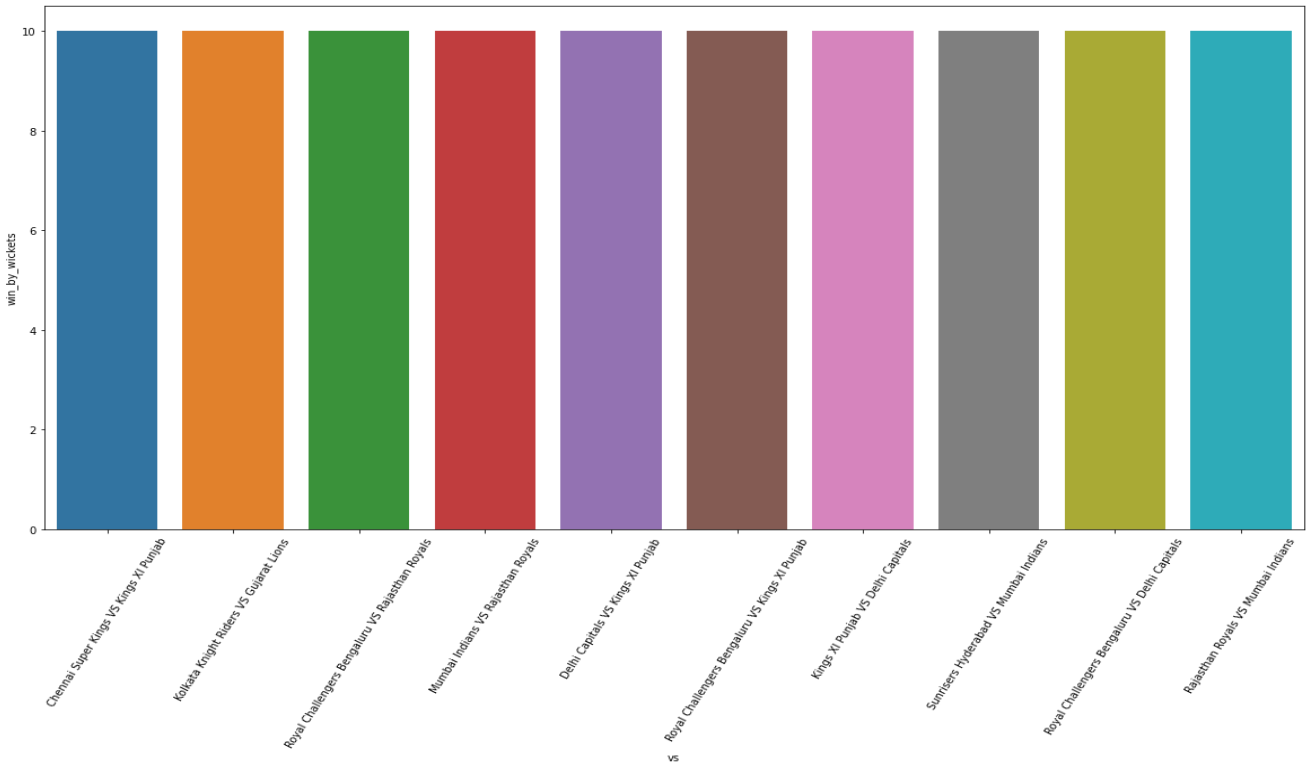

Fig 4. Top 10 wins by wickets

4. Strike Rates of batsman in different overs/phases of the game: Following graph shows how top performing batsman usually structure their inning, how they play in in different phases of game and their mean strike rate. This is influenced by the batting strength of his team and his batting position.

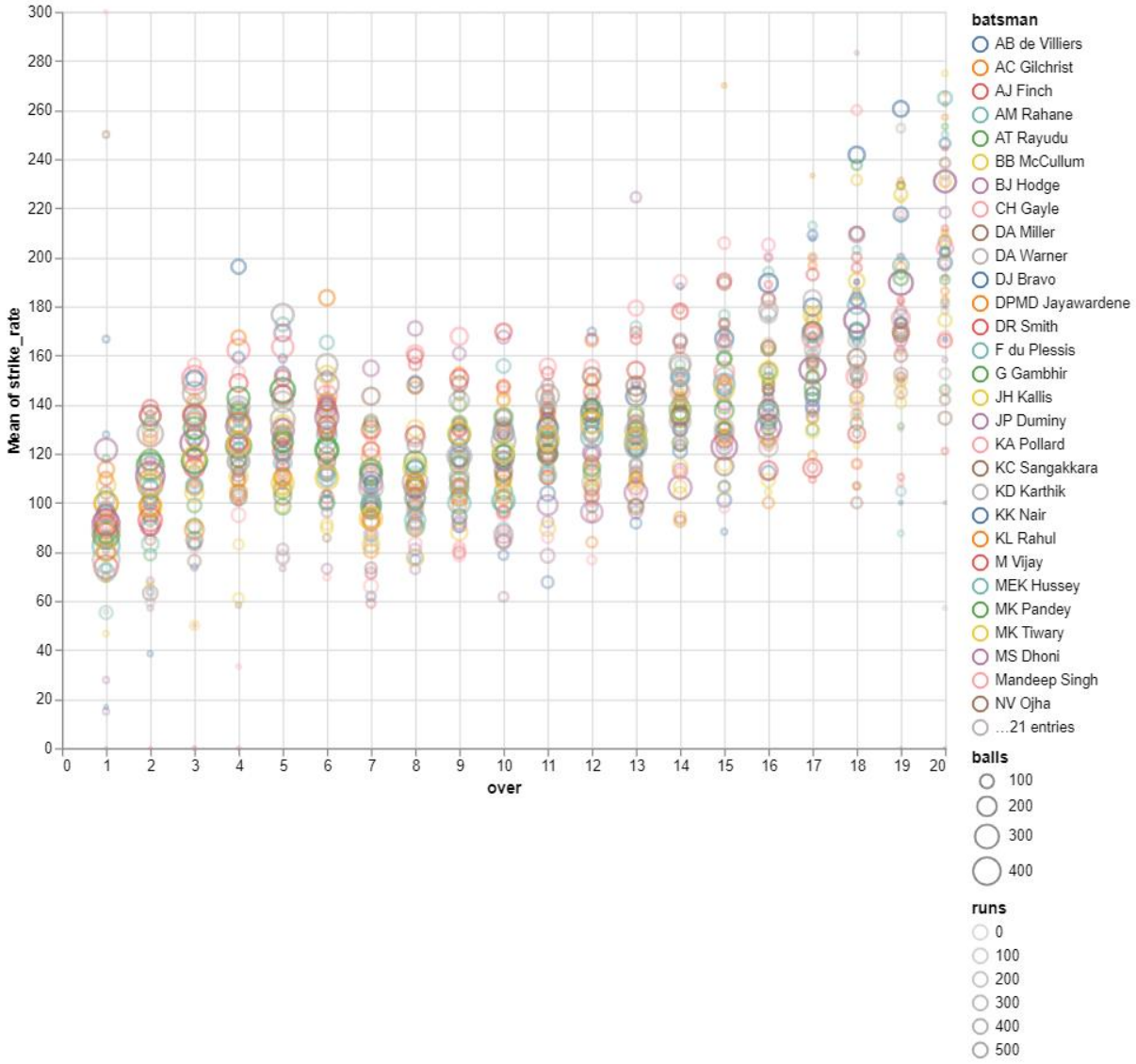

Fig 5. Strike Rates of batsman in different overs/phases of the game

5. Most 50's and 100's scored: In the period 2008-2019 Chris Gayle has scored most number (7) of centuries followed by Virat Kohli and David Werner has scored most number (44) of half-centuries followed by Virat Kohli. So, usually batsman who bats in higher order have dominated the list.

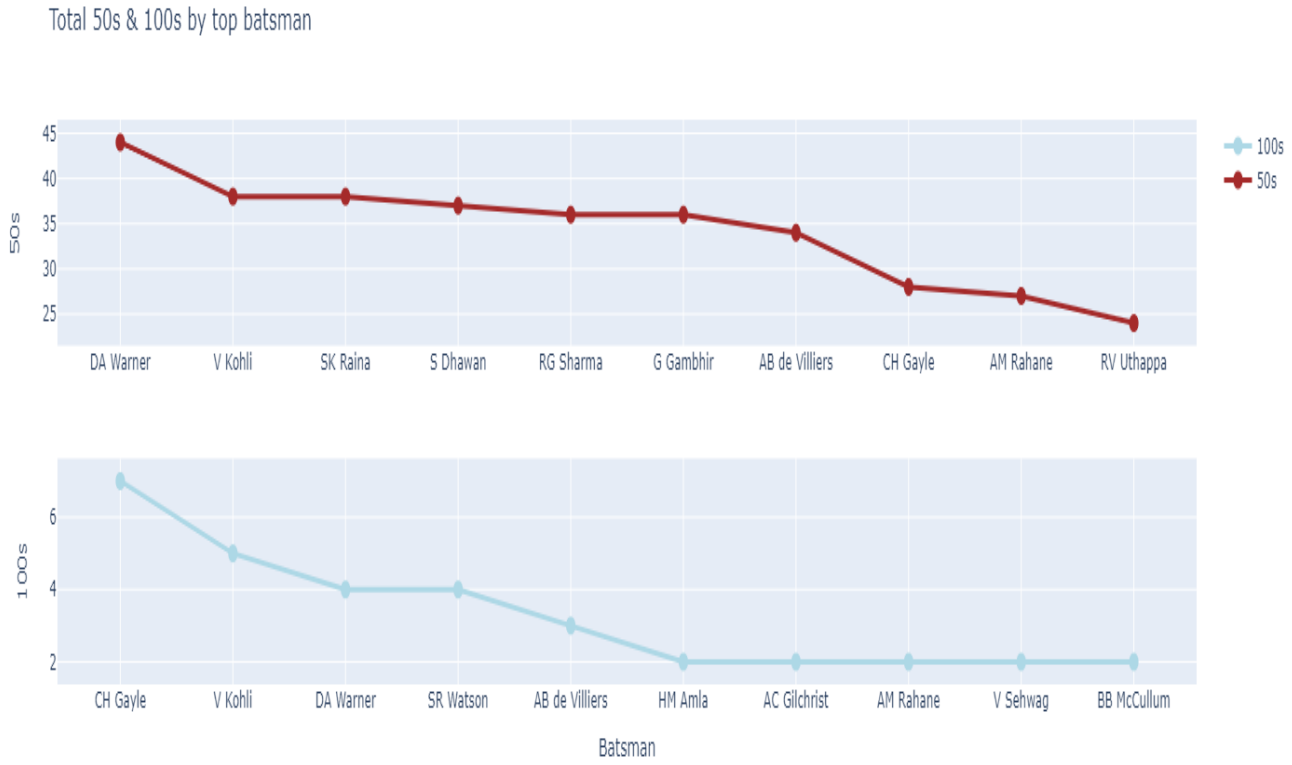

Fig 6. Total 50s and 100s scored by Top Batsmen

6. Most successful teams: We can see most wins by all teams ever competed in IPL. As being the most crowned team of IPL, Mumbai Indians has the highest number of wins as expected. What is interesting to note is that despite not competing in 2 full tournaments (due to ban), CSK are just 14 wins away from MI in terms of total wins, a proof why MI-CSK rivalry is most popular in IPL.

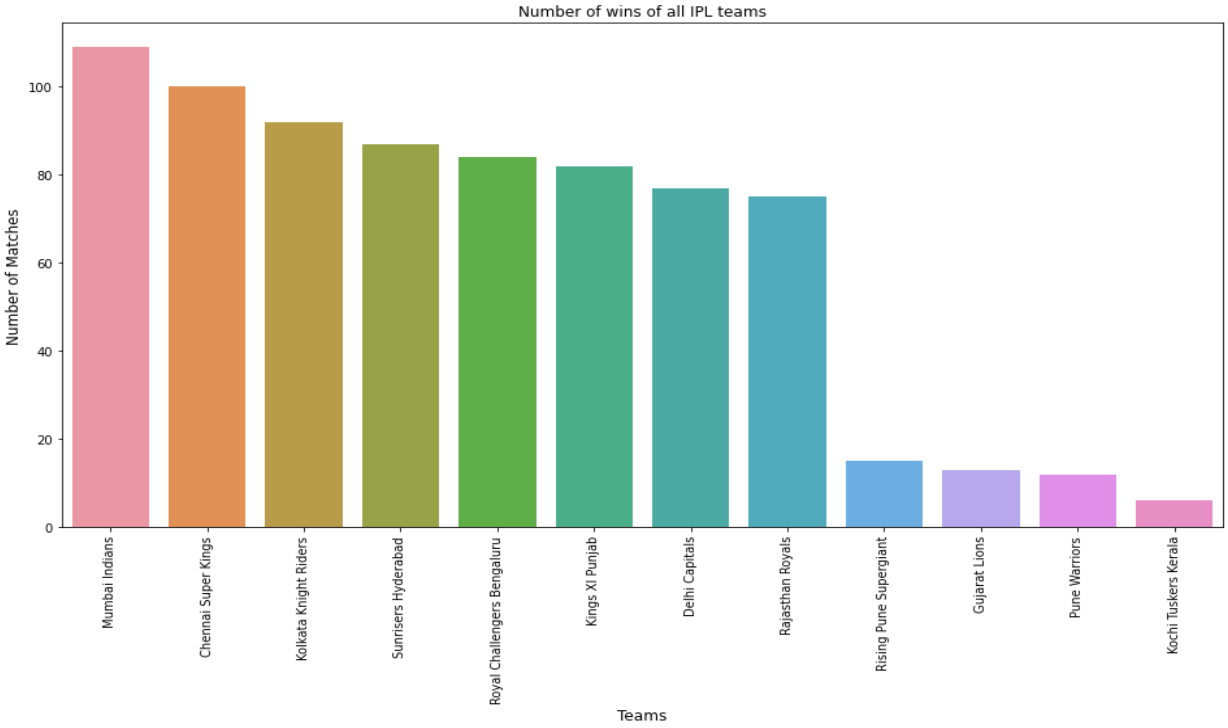

Fig 7. Total no. of wins of all IPL Teams

But since all teams have not played same number of matches, total number of wins is not a good measure of success. So win percentage of each team has shown in another graph and it is showing that Chennai Super Kings have the highest win percentage followed by Mumbai Indians.

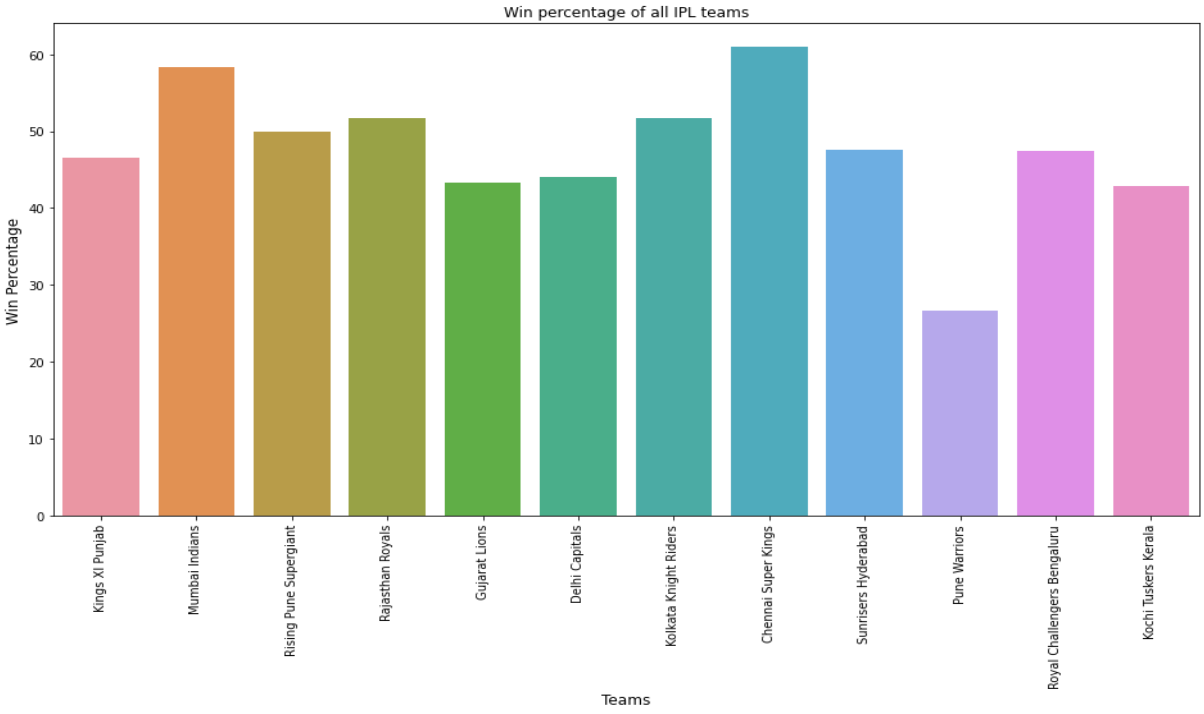

Fig 8. Win Percentage of all IPL Teams

7. Most Influential Players: Using Man of the Match (MoM) data available in the dataset, we can extract players with most MoM awards in IPL. This analysis shows us that

despite RCB not winning no IPL yet, three of RCB players feature in top 20 list with AB de Villiers and CH Gayle at number 1 and 2 respectively. This highlights the incompetence of RCB's bowling which hampers their title-winning chances. (only top 20 players are shown in fig)

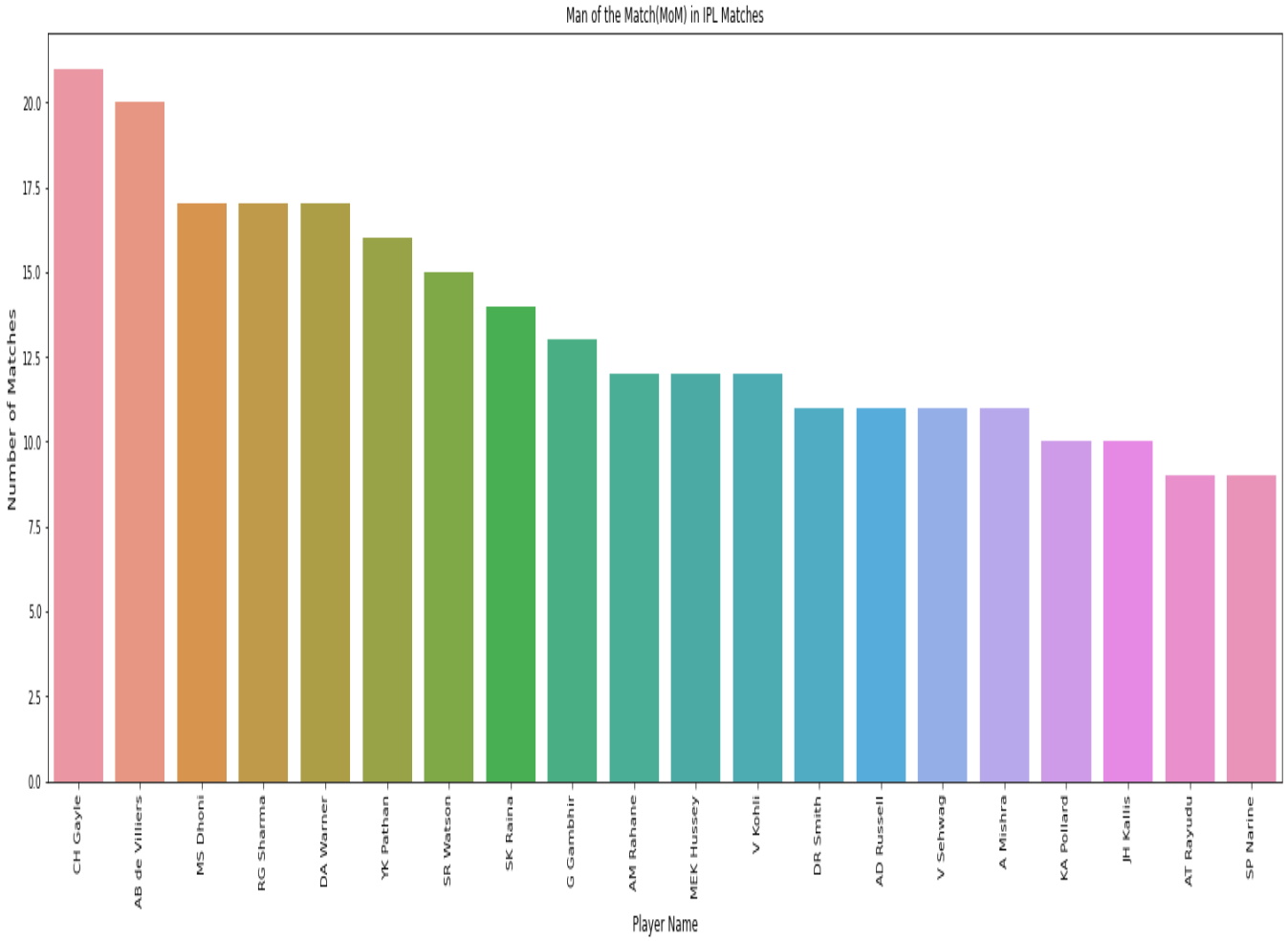

Fig 9. Man of the Match in IPL matches

8. Top hosting venues or stadiums: We can see Eden Gardens (Kolkata) and M Chinnaswamy Stadium (Bengaluru) as the joint top venues to host IPL matches, followed by Wankhede Stadium (Mumbai). Despite only having one tournament there, two cities of South Africa also show up in the top 20 list due to limited number of grounds available in the country

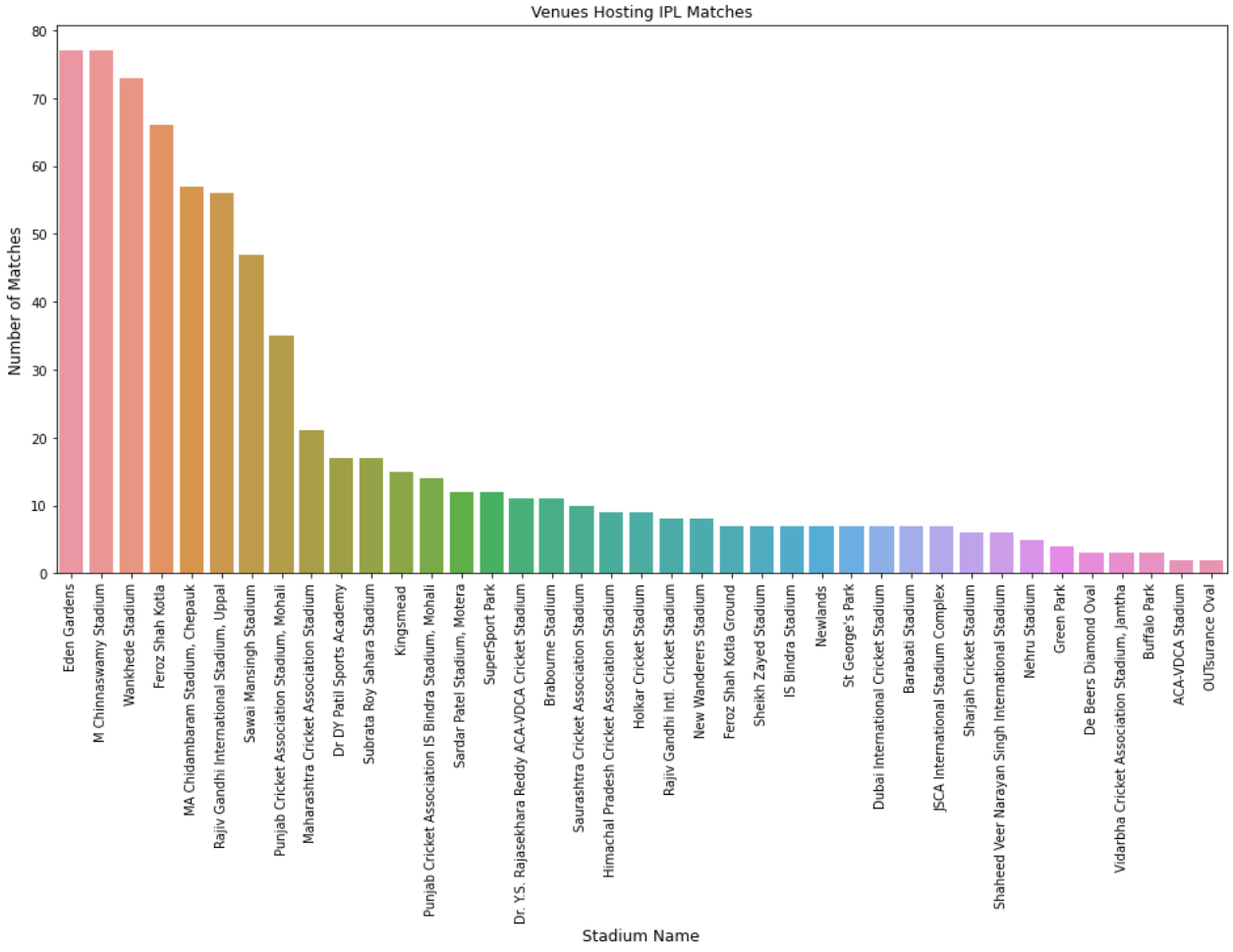

Fig 10. Total matches hosted by venues

9. <u>How two teams performed against each other:</u> A function is defined that helps to get how two teams have performed against each other in all seasons.

   For example, here is a graph of how Chennai Super Kings and Mumbai Indians have played against each other.

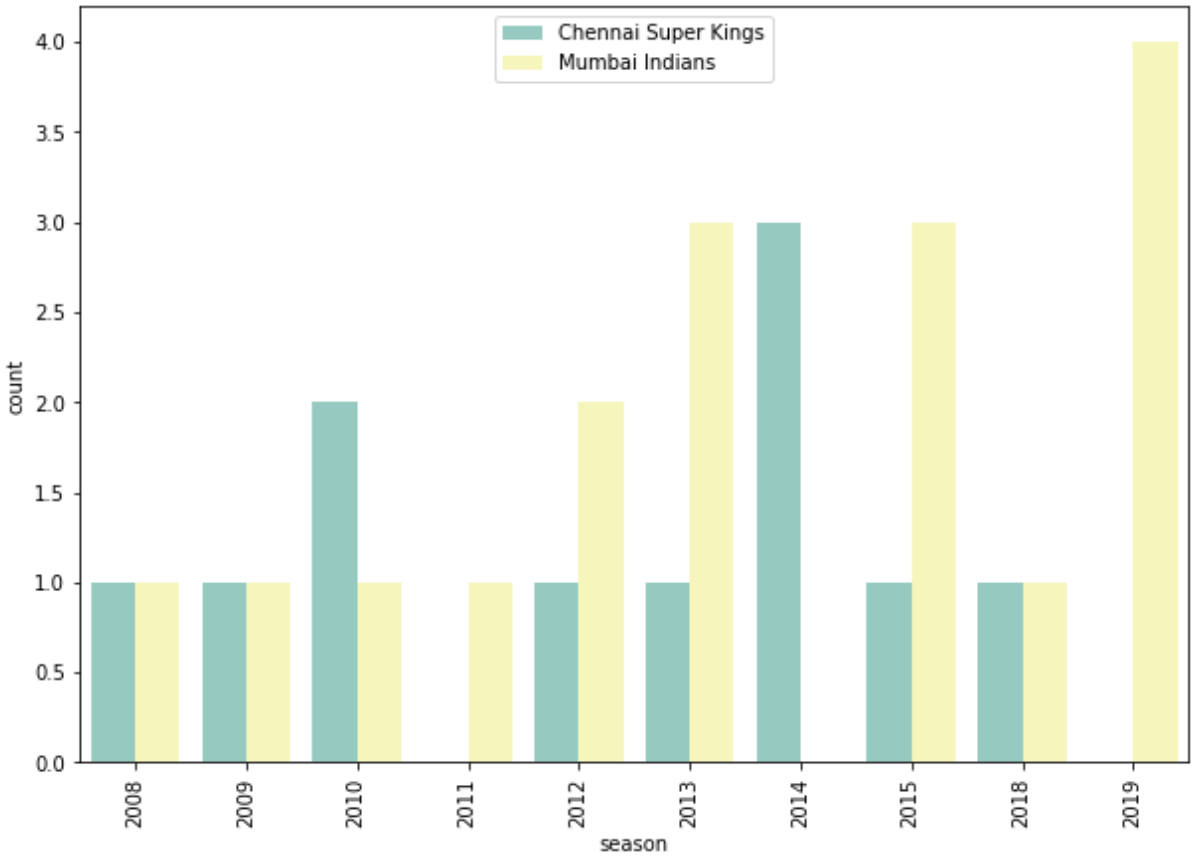

Fig 11. Mumbai Indians vs Chennai Super Kings all time record

Predictive Analysis

Final Score Prediction

In this section, this paper presents a method to model the final score of the team which is presently batting, based on the current score of the team. The exact features which are used for prediction are listed as follows:

- Current runs
- Current wickets
- Overs completed
- Striker's runs
- Non-striker's runs
- Bowler
- Runs in last 5 overs
- Wickets in last 5 overs

Current runs and wickets are the most influential in determining what will be the final score of the inning as more wickets fallen would mean the team won't be able to muster up more runs and more runs at any position would automatically mean higher score possibility due to runs being added cumulatively to the final score. Current score alone won't be of much help if we don't know current overs as combining these two the network can know current net run rate. Striker and non-striker's current score would also be helpful as set batsmen would be crucial to elevate the teams' score. Runs in last 5 overs and wickets in last 5 overs are also important features.

Prediction

- **Neural Networks (NNs)**: Neural networks are the modern times way-to-go prediction models. Neural networks are based on how the human nervous system works; neural units are basic processing junctions of a neural network. Each unit receives an input, applies an activation function to it and sends processed output to next layers' units. Several such layers of units are stacked together and the neural network learns intricate details of data itself such as to achieve satisfactory results on target labels. Fig shows the architecture of the NN used for prediction. Batchnorm is applied after each linear operation before applying leaky relu as activation function. Batchnorm is used to prevent exploding of units' activations and consequently the gradients. Leaky relu prevents 'dead neurons' as it always has a slope for gradient computation. Adam optimizer is used for backpropagation and regularization purposes. Mini-batches of 128 size are used for mini-batch gradient descent. The training is performed on Google Colab with GPU runtime and the model is built in Pytorch, a deep learning framework. The total number of learnable parameters in the model are 100k and it is run for 200 epochs with 0.001 as the learning rate. These hyperparameters for learning are chosen after running many experiments. Metrics used: Mean squared error is used as a loss

function for backpropagation. R2 score and custom accuracy (predicted score being in margin of 10 of actual final score) is used for evaluating the model.

Results: After 200 epochs, the results obtained are as follows: Train Loss: 0.0027 — Val Loss: 0.0032 — Train R2: 0.5974 — Val R2 : 0.5211 — Val Acc: **62.3512%.** So, the accuracy is around 62% which is not usable for practical purposes. Limited amount of data is the primary reason why neural networks are not giving decent results even after training for more than 1 hour.

- **Random Forest:** Random Forests achieve a reduction in overfitting by combining many weak learners (Decision Tree) that underfit because they only utilize a subset of all training samples. Best number of estimators and maximum depth of trees are found to be 300 and 50 respectively.

  Results: Choosing the number of estimators to be 5000 and maximum depth as 14, we get the accuracy as **66.84%** and R2 score as 0.69 on testing data.
- **Linear Regression:** Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

  In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

  Results: The accuracy score is **45.06%.** So it is not a good fit.
- **Support Vector Machine**: The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points.

  To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

  Results: The accuracy score is 44.18%

	Model	R-Sq	Accuracy (%)
0	Linear Regression	0.50	45.06
1	Random Forest	0.69	66.84
2	SVM	0.47	44.18

Fig 12. Summary of results obtained from different models

As seen in above table, the Random Trees Regressor offers the best performance and is the best ML frameworks to use for predicting the final score of the batting team.

Winner Prediction

In this section, this paper presents a method to predict the winner before any match.Here, exact features which are used for prediction are listed as follows:

- Team 1
- Team 2
- Toss winning team
- Toss decision
- DL-applied
- Venue

Prediction

- **Logistic Regression**: Logistic Regression is a Machine Learning method that is used to solve classification issues. It is a predictive analytic technique that is based on the probability idea. The classification algorithm Logistic Regression is used to predict the likelihood of a categorical dependent variable. The dependant variable in logistic regression is a binary variable with data coded as 1 (yes, True, normal, success, etc.) or 0 (no, False, abnormal, failure, etc.).
  The goal of Logistic Regression is to discover a link between characteristics and the likelihood of a specific outcome. For example, when predicting whether a student passes or fails an exam based on the number of hours spent studying, the response variable has two values: pass and fail
  A Logistic Regression model is similar to a Linear Regression model, except that the Logistic Regression utilizes a more sophisticated cost function, which is known as the "Sigmoid function" or "logistic function" instead of a linear function.

  Results: The accuracy is **0.6358**. So, it is not a good fit.

- **Support Vector Machine:** It is also used for classification problem. Accuracy score **0.6159** and it is also not a good fit.

- **Random Forest:** It is also used for classification problem. Accuracy score **0.5497** and it is also not a good fit.

Though no model gives a very good fit, still logistic regression gives the best fit.

	Model	Accuracy (%)
0	Logistic Regression	63.58
1	SVM	61.59
2	Random Forest	54.97
3	Decision Trees	58.28

Fig 13. Summary of results of different models

ACKNOWLEDGMENT

I would like to express our very great appreciation to my mentors Vivek Parmar and Ishaan Garg for this opportunity to explore the application of data science to cricket. I learned a lot of beneficial things during the working and completion of this project. My programming and data science skills have surely become better thanks to my mentors.