Data Analysis of the India Premier League

Vedanth S Saoor
IIT Bombay

Introduction

Analytics is a way of systematically analysing data. It is used for discovery, interpretation of patterns that exist in the data. It has been applied to variety of fields such as marketing, management, finance and sports as well. Sports analytics has become widely popular in the following days. Using relevant information like historical performances and individual player statistics can facilitate decision making process during or prior to the matches that can ultimately provide a competitive advantage to a team or an individual. As over the years has the technology has advanced and collection of in-depth data has become relative easier, apart from assisting in the decision making and efforts towards improving the performance of team or a player. The off-field analytics helps the sports organizations or teams increase the ticket and merchandise sales and also improve fan engagement.

About Cricket

The game of Cricket was first played in the 16th century in England. It is a sport that is played by two teams with each teams consisting of 11 players. In a match there are two innings with one team batting in one inning and the other team batting in the next inning. The decision of whether a team will bat first or field first is done by a coin toss. There are different of formats of this game like ODI, T20 and Test. In ODI's and T20's there are limited over's in which the bowling team tries to restrict the batting team to reasonable score and the team batting second has to score a run more than opposition team total. There are several factors that affect team's performance like the individual team-mates form, the pitch conditions, the environmental conditions like the dew factor, venue where the match is being played like the team playing in its home ground tend to have some advantage and finally the team against they are playing.

Data analytics can assist the decision making of the teams management by helping them 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, for example the modern technologies allow analyst to draw up the wagon wheel of the batsman's knock and help the bowlers change their lengths and also field.

About Indian Premier League

The Indian Premier League (IPL) was founded in the year 2007 by the BCCI. It is biggest and the most popular domestic cricket league in the world with the brand value reaching to 6.3 Billion dollars in 2019. IPL is 20-over format with 8 teams involved in the tournament with each team consists of both Indian players as well as foreign players.

Overview

The report is structure in the following format. Firstly, the datasets that have been used for the analysis are stated. Secondly, Exploratory Data analysis (EDA) that is conducted is stated. The major focus of the project was to create a forecasting model in order to predict the final score. In order to obtain a good result data pre-processing is carried out along with feature engineering that would assist in predicting the final score more accurately. For the predictive model we have used various machine learning and deep learning models. Third, the results that are obtained from the predictive model are given. Finally, Conclusion and future scope of this project are stated.

Datasets

The dataset that was use for analysis was taken from www.kaggle.com, where the data is given for 12 seasons starting from the year 2008 to 2019. There are two datasets that are named as deliveries and matches all in csv format with approximately 17900 and 756 data points respectively. The matches file consists of 18 features in total, some of the features are as follows: season, city, team1, team2, player_of_match, toss_winner, etc. The other file consists of 21 features in total; some of the features are as follows: team1, team2, total_runs, dismissed_kind, batsman, non-stirker, etc. There were some missing values in the data which was removed or replaced by an appropriate value in the data pre-processing stage.

Analysis

We conduct an Exploratory data analysis (EDA) on the data as it allows us to discover any patterns in the data, to spot anomalies in the data, find out if any data is missing and also provide better understanding of the features that are important for the predictive models. The insights that are drawn from the data after careful and calculated manipulation of the data are stated in the following.

1) Maximum Number of wins by any team in particular season
   It can be observed from Fig 1 the most number of wins by a team in a season. It can be observed that the most number of wins by any team in any given season is 13 wins. It can be observed that Mumbai Indians (MI) tend to appear to have most number of wins for 5 seasons followed by Chennai Super Kings (CSK) with most number of wins in only 2 seasons.

2) Stadium that has hosted most number of IPL matches
   From Fig 2 it can be seen that the most of the IPL matches have been played at Mumbai with 101 wins followed by Kolkata and Delhi with 77 and 73 number of matches that are played at that venue. The least number of matches that is 2 matches are played at the venue Bloemfontein which is South Africa.

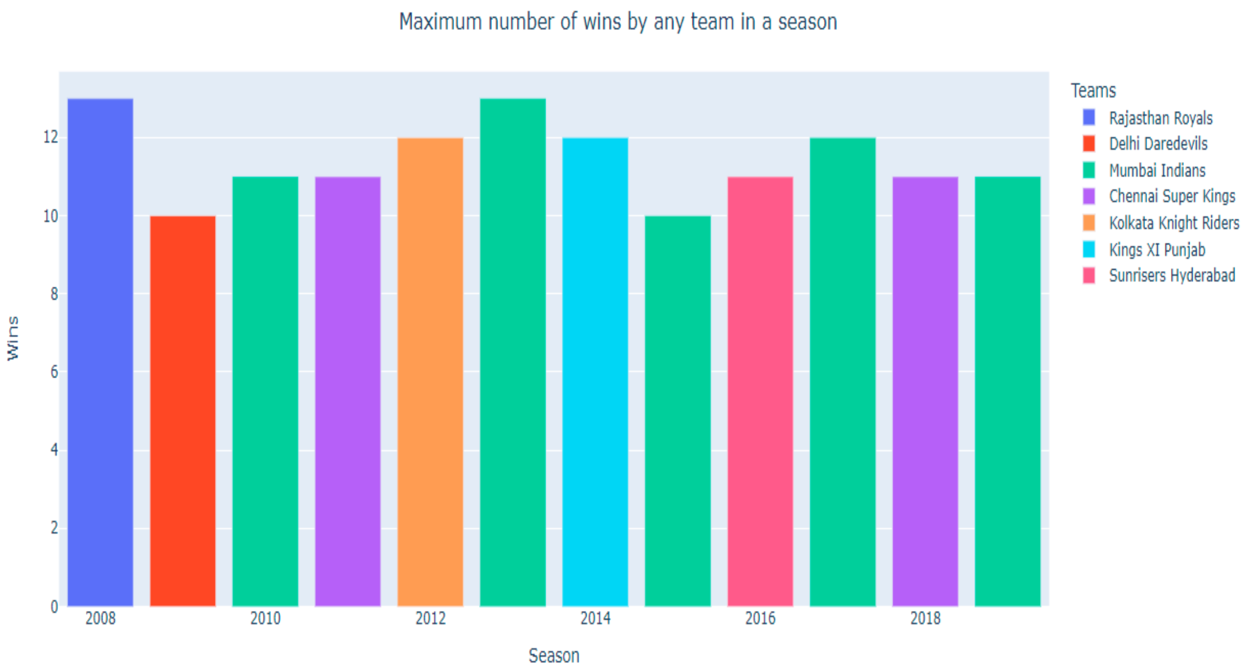


Fig 1: Maximum number of wins by any team in season

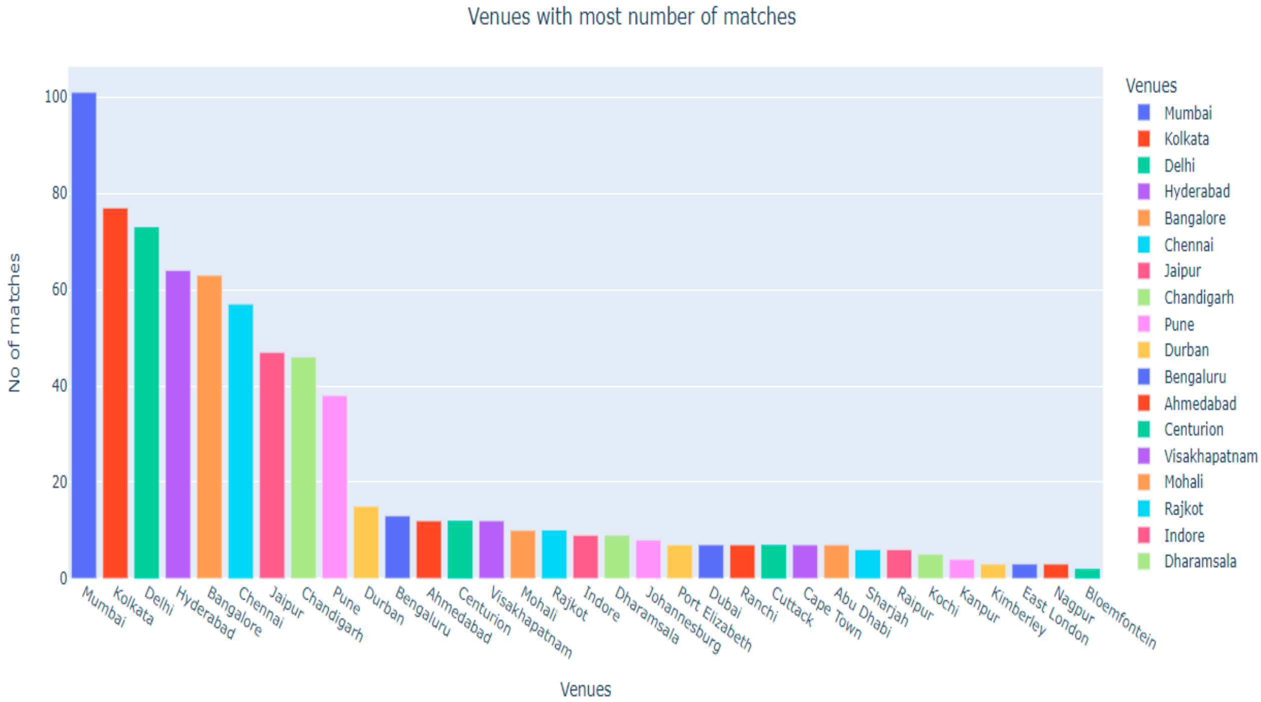


Fig 2: Venues with most number of Matches

3) Which team has won the most number of matches won

In Fig 3 we can see the most number of wins by all the teams in the IPL. The number matches won is MI with 109 followed by CSK with 100 wins. When we look at the win percentage the standings are as shown in Fig 4 with the no.1 position for the highest win% is taken by CSK followed by Delhi capitals (DC) with 60% and then MI with 58.59% until the year 2019.

4) Which player has won the most number of Man of the matches awards

It can be seen in Fig 8 that the most number of Man of match award is gone to Chris Gayle followed by AB devillers who have both played for RCB for a very long time. Despite the splendid performance, RCB has failed to win the IPL trophy. This tends to show that RCB's bowling performance is the reason why the choke in the final stages of the tournament.

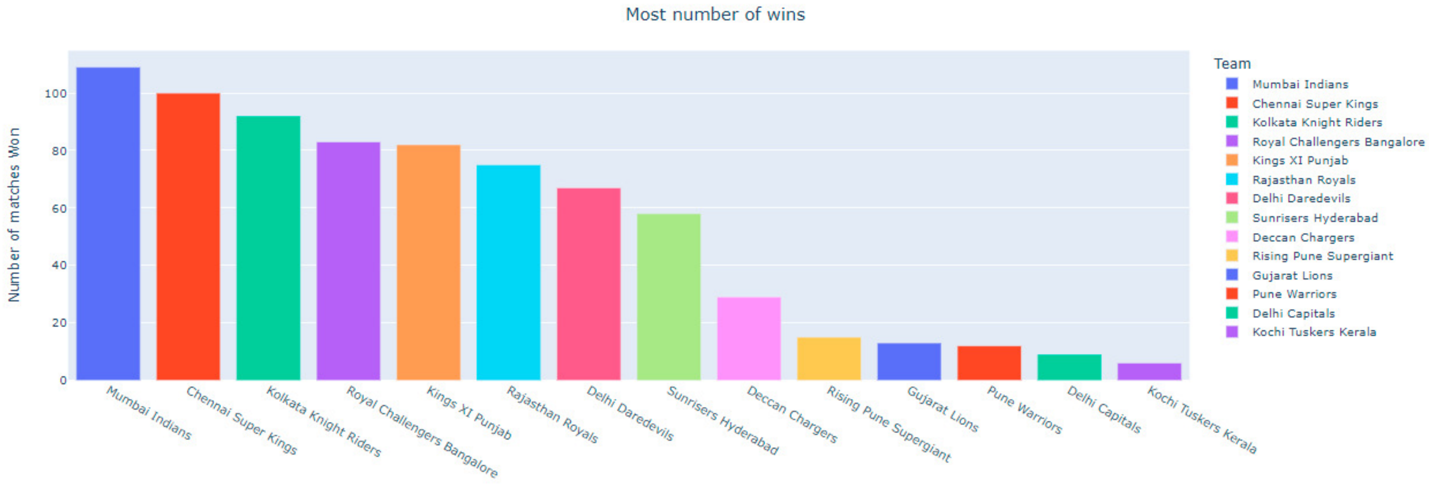


Fig 3: Most Number of wins

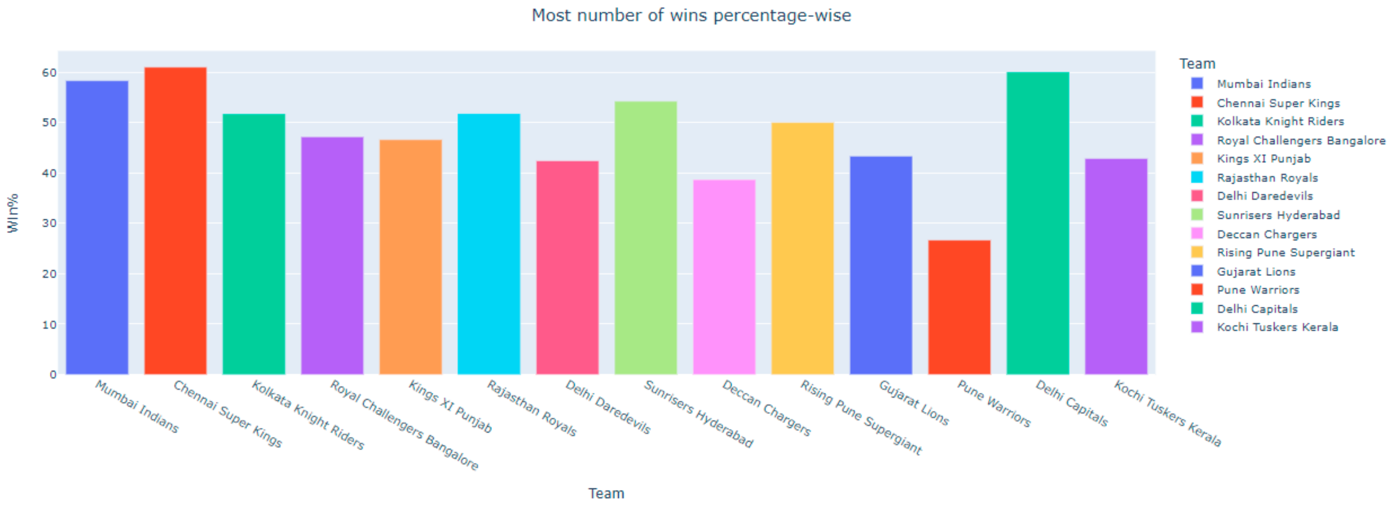


Fig 4: Most Number of Wins percentage-wise

5) What are the 10 greatest victories

In Fig 5 we can see that the greatest win margin in terms of runs was between MI vs DD, with run margin of 146 in which MI has been victorious. Also from the Fig it can be observed that MI has appeared in the plot. Hence MI has one of the strongest bowling attack

6) Which team has won the most tosses

In Fig 6 we can observed that the most number of toss won is by MI with 98 wins followed by KKR with 92 wins. As early when we see the %win the picture is different. We observed in Fig 7 that Delhi capitals have highest %wins followed by Deccan chargers with 57.33%.

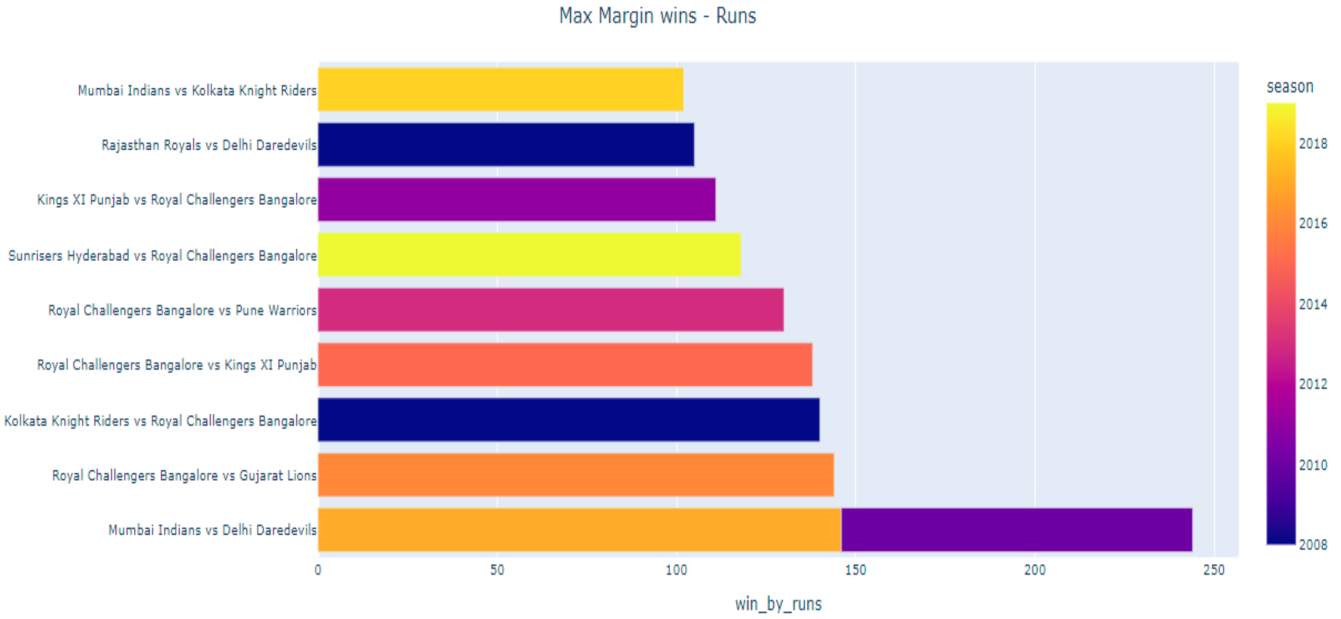


Fig 5: Max Margin wins -Runs

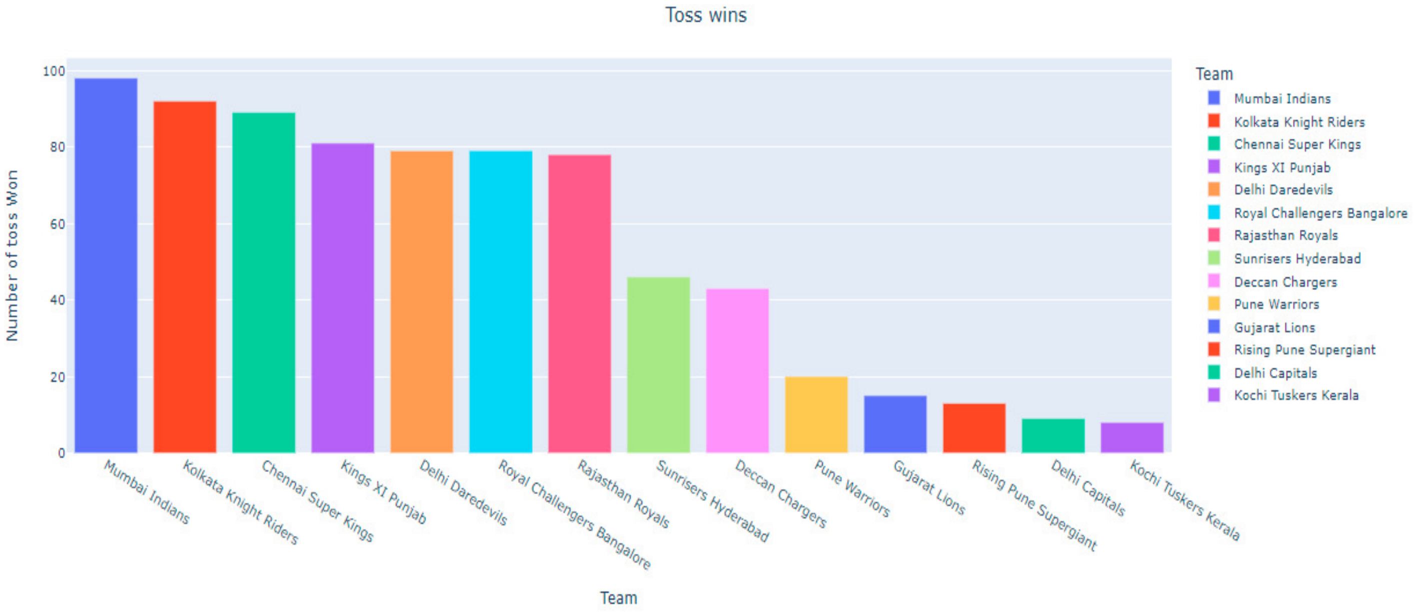


Fig 6: Toss wins

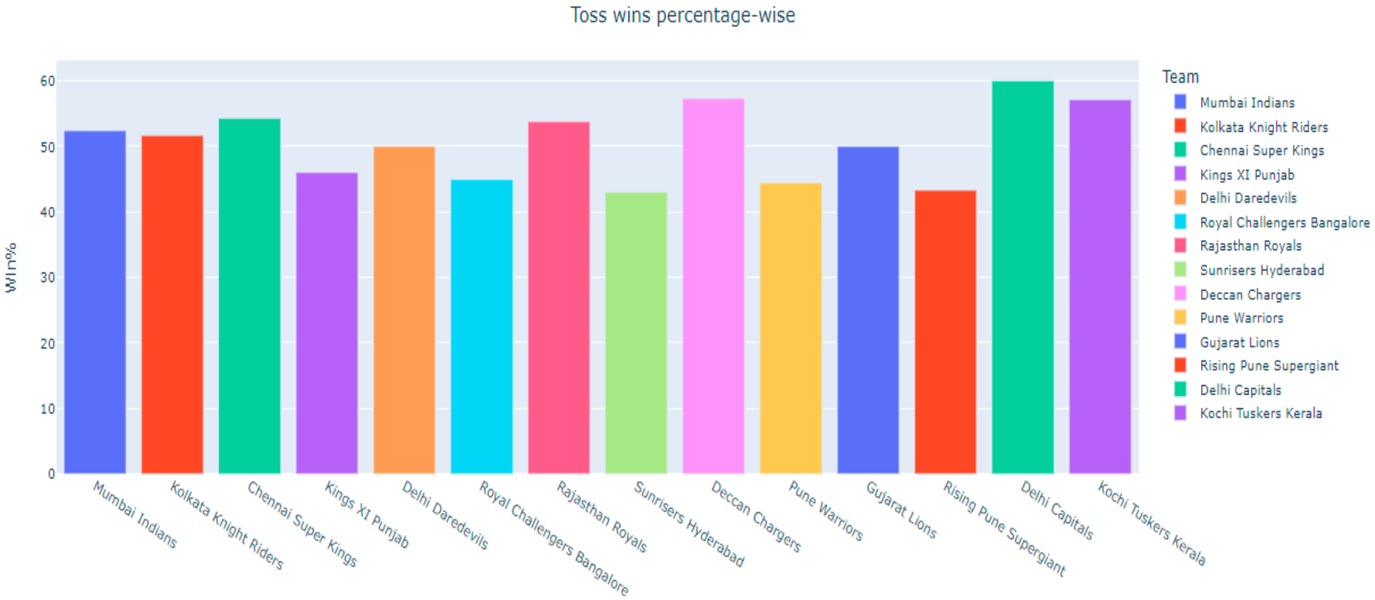


Fig 7: Toss wins percentage-wise

7) Most 50’s and 100’s scored
From Fig 10 and 11 It can be seen that most number of 100’s is scored by CH Gayle with 7 centuries followed by V Kohli with 5 centuries. From Fig it can be seen that most number of 50’s is scored by DA Warner with 44 half centuries followed by V Kohli with 38 half centuries. Again be observed that despite a good batting order RCB tend to fail to win the IPL.

8) Comparison between batsmen
Fig 9 is a comparison of two batsmen in both batting and bowling aspects. We can see a comparison of Gayle and Warner. We observe that Warner has more number of matches player, runs scored , number of boundaries and number of 50’s scored. On the other hand Gayle has more number of 100’s and 6’s against his name. Gayle also has taken 12 wickets in the IPL has Gayle is an all rounder as he is a part time Right off-break bowler.

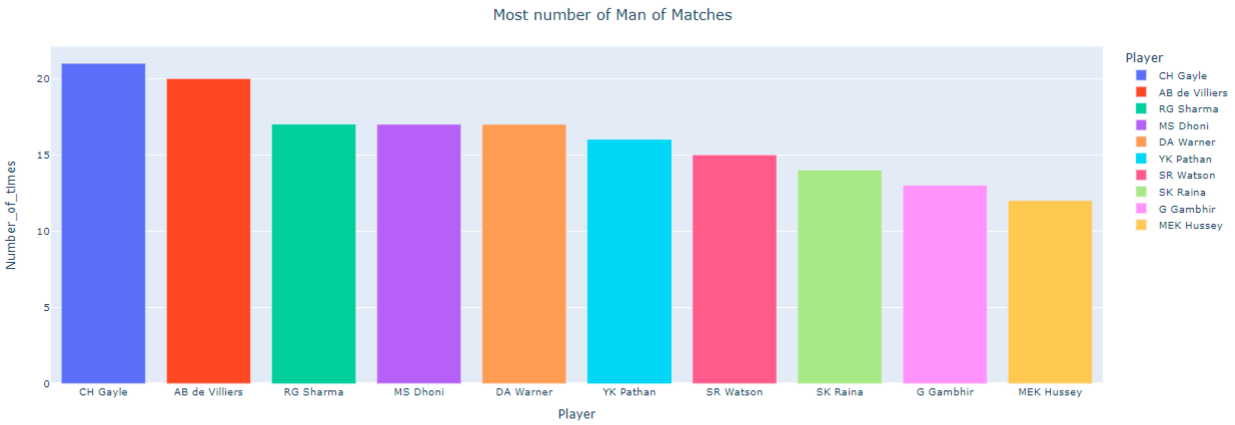
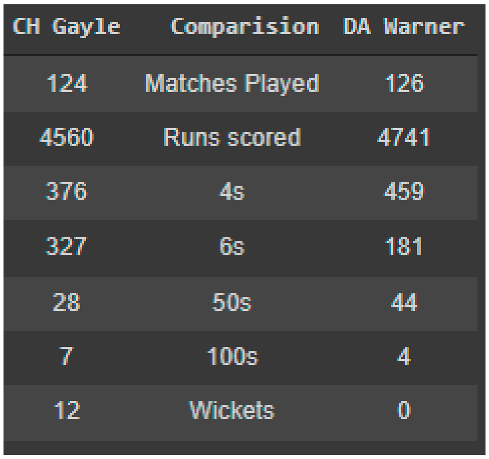


Fig 8: Most Number of MoM



CH Gayle	Comparision	DA Warner
124	Matches Played	126
4560	Runs scored	4741
376	4s	459
327	6s	181
28	50s	44
7	100s	4
12	Wickets	0

Fig 9: Comparison of Two players

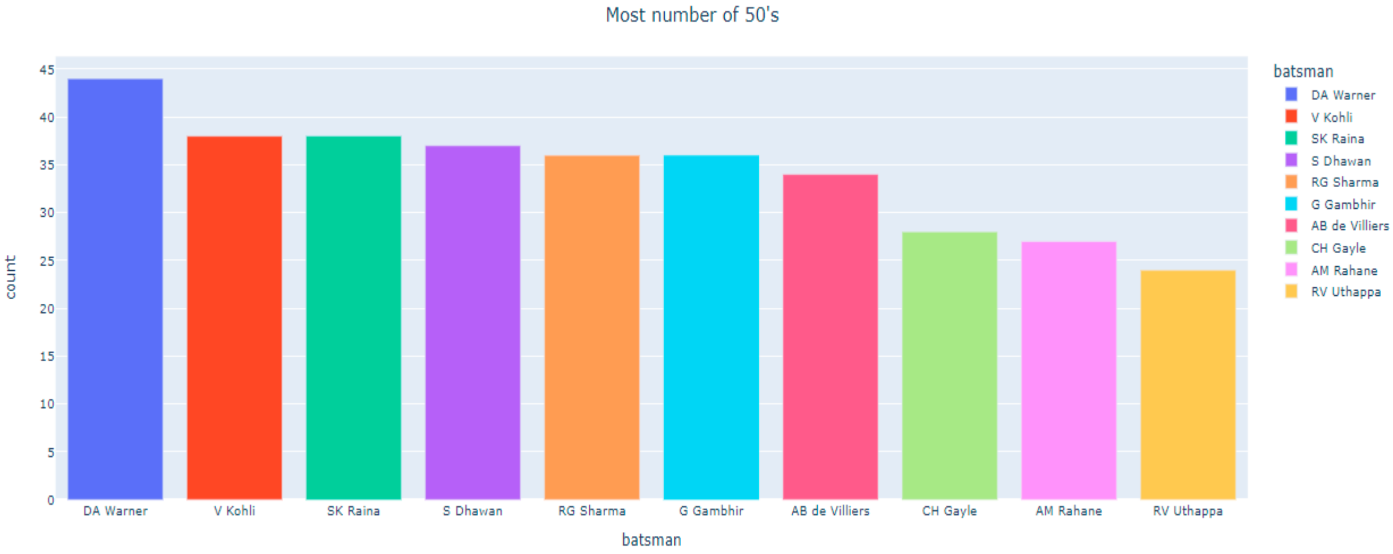


Fig 10: Most number of 50’s

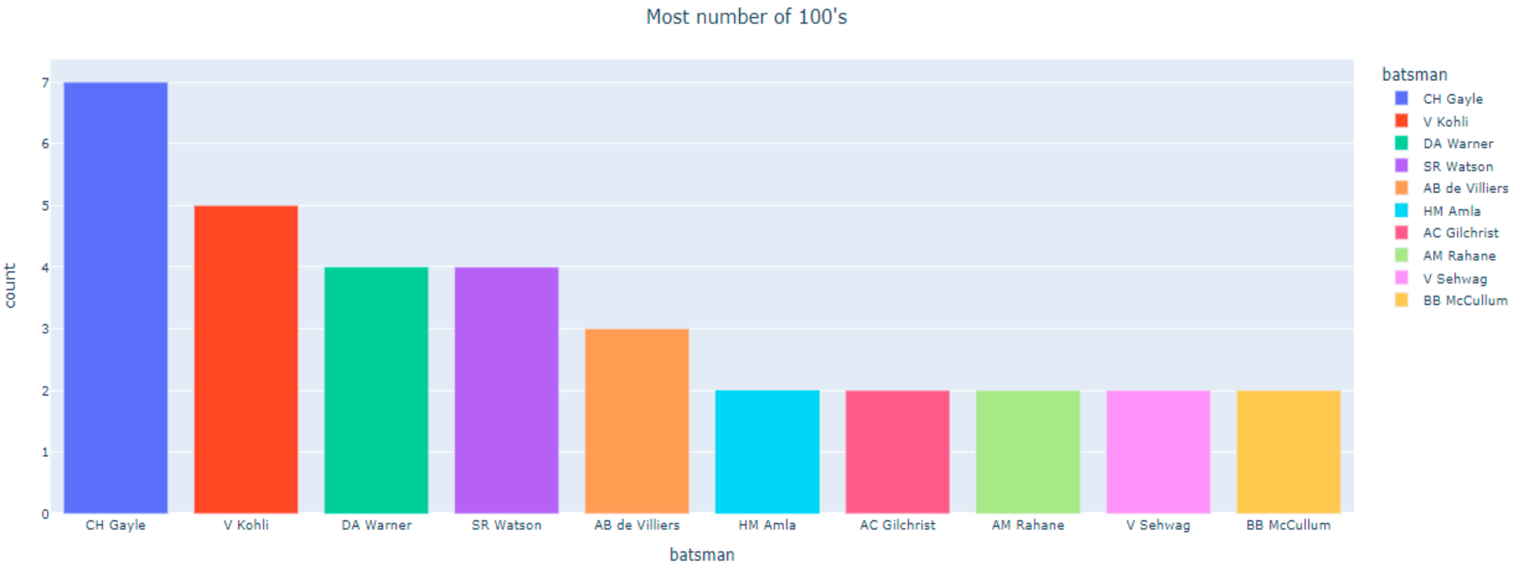
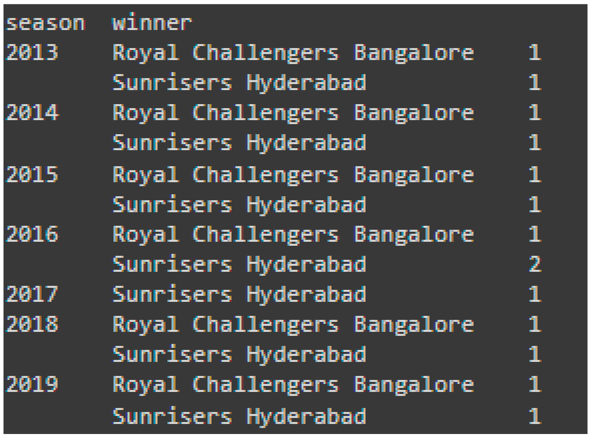


Fig 11: Most number of 100’s

9) Comparison of two teams:

The following Fig shows the comparison of two teams on how they have performed against each over all the seasons so far. It can be seen that during the stage group both teams tend have same number of wins. But SRH tend to get better off RCB during the playoffs.



```
season  winner
2013    Royal Challengers Bangalore    1
        Sunrisers Hyderabad            1
2014    Royal Challengers Bangalore    1
        Sunrisers Hyderabad            1
2015    Royal Challengers Bangalore    1
        Sunrisers Hyderabad            1
2016    Royal Challengers Bangalore    1
        Sunrisers Hyderabad            2
2017    Sunrisers Hyderabad            1
2018    Royal Challengers Bangalore    1
        Sunrisers Hyderabad            1
2019    Royal Challengers Bangalore    1
        Sunrisers Hyderabad            1
```

Fig 12: Comparison of two teams

Results

In this report we tend to use machine learning algorithms to predict the final score. The dataset used for training the machine learning model is stated in the earlier section. The data contains number of features; we use the following features for the prediction:

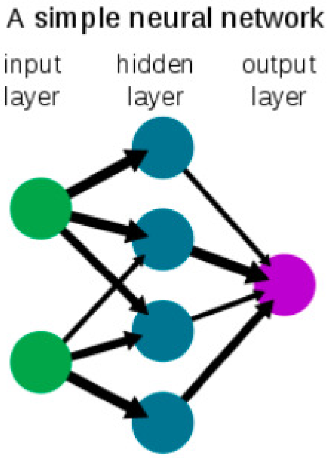
- Current runs
- Current Wickets
- Current Over's
- Final Score

Current runs and Current Wickets are one of the most important features for predicting the final score. These features were not provided in the dataset. Hence these features were created. Just alone the current score is not enough to predict the final score.

Neural Networks:

A Neural Network is a network of neurons. They tend to mimic the human brain. The neural network consists of input layer, hidden layer and finally output layer. Each neuron has an activation function. There are different types of activation functions like linear, ReLu and Leaky ReLu .

For this project we have used three hidden layers with decreasing number of neurons with each neuron having a **ReLu** activation function. We have split the data to training and testing and used the training data set to train the neural networks followed by testing it using the testing dataset.



The loss on the training and validation/testing set is shown below and its can be seen that the training set loss decrease as the number of Epochs increase. For training the neural network we have completed **200 epochs**. We have used **RMSE**- Root Mean Squared Error, which turns out to be **0.137**. Further the hyperparameters such as the activation function, number of hidden layer and number of units can be altered to meet the required accuracy and to avoid overfitting and underfitting.

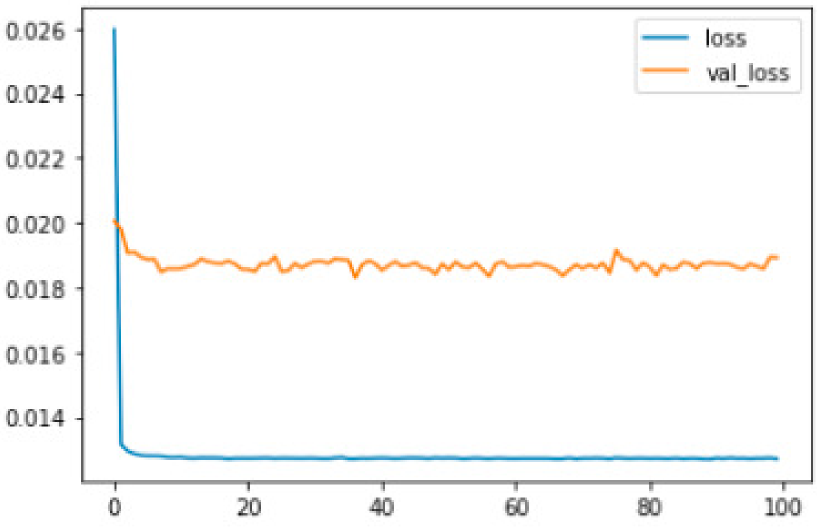


Fig 13: Training and Validation error

Linear Regression and Random Forest Regression:

We use the Linear Regression and Random Forest Regression algorithms/models to predict further score. Similar to above us split the data into training and test set. Train the model on the training set and then testing the model on the training set with custom loss function.

The custom loss function counts the number of times the model has given a correct prediction within a given threshold on the testing dataset. Using the custom made loss function we have the following accuracy as follows **73.53%** using the Linear Regression and **65.21%** using the Random Forest Regression

Conclusion and Future Scope:

During the project we get valuable insight about whom and what the best performing teams and players are. We can also get insight on the most influential and the entertaining player in the IPL, the trends of rivalry existing in the league. The data we infer can assist the team management to make some changes to the team. For example we can observe that RCB has best batsmen yet they fail to lift the IPL trophy there by raising concerns for its bowling attack.

In future scope one can do the hyperparameter tuning and try obtaining a model that best fits the dataset and yet is generalized such that there is no underfitting and overfitting. Also try other Machine learning algorithms.