INDEPENDENT READINGS [2 CREDITS]

IPL Data analysis

Abstract

This project includes exploratory analysis and prediction of outcome of the matches based on the variables in IPL dataset.

Suyash Mhetre

Mhetresuyash7@gmail.com

Contents

Introduction:

IPL also known as Indian Premier league is a professional game where 8 teams come together to contest against each other to win the final cup. This league was founded in 2008 and reached the highest brand value in 2019 to 6.7 billion USD. Also, in 2015 the IPL contributed around 160 million USD to GDP of Indian economy. As we see that huge amount of money is involved in this league. It is important to figure out the factors responsible for performance of teams and wining probability of the teams.

Our main objective was predicting and visualize the impact of various factors on players and team's performance. Based on these visualizations it would help sponsors to make wise decision while make investments on team. We used sci-kit learn to import the prediction models. RandomForestClassifier was used to predict the outcome of the match. We will discuss factors used to classify the data going further in the paper in detail.

Literature Review and Exploratory analysis on IPL dataset:

According to the sports analytics column in "The Hindu" (A national newspaper in India), the central revenue generated by media rights and sponsorships reflects the huge growth to make IPL the world's best T20 league. Within the tenure of 2008 to 2019 the media rights have seen phenomenal growth from 31.57 million USD to 546.75 million USD. Therefore, it is very critical for title sponsor, official partners, umpire sponsor and strategic time out sponsor to make wise decisions before making investments on the teams. We tried to do exploratory analysis to get great insights about the data available in the open domain.

There is not a lot of research done on the IPL dataset. But, according to one of the researches done in 2018 by Laurentian University, Canada they are able to predict

the accuracy of performance of players using naïve bayes, decision trees, RandomForest, SVM. They have considered the factors like form of the player, consistency, batting average, bowling average and number of matches played. This research mainly focused on batting and bowling prediction of a player based on the idol situation. Factors like home ground advantage and weather were not considered due to unavailability of the data. Also, this research does not take in consideration the format of game. Because cricket is played into 3 formats One Day matches, Test matches and T20 format. We mainly focused our analysis on T20 format.

Data Sets and Data Engineering:

Data extraction is done by using Python libraries as shown in the picture. We have done data scraping from the websites that have data for IPL statistics from 2008-2019. BeautifulSoup is majorly used to clean and organize the data into csv format.

```
import requests
import urllib.request
import time
from bs4 import BeautifulSoup
import lxml.html as lh
import pandas as pd
```

Following are the links that we used to scrap the data:

1. https://www.iplt20.com
2. https://cricsheet.org/

These datasets have the IPL data for all the leagues played from 2008-2019. 2020 I PL was cancelled due to COVID19.We have matches dataset with 12,084 rows and deliveries dataset with 31,59,660 rows. Matches dataset includes the information a bout the matches of individual teams and their wins and losses.
While the deliveries have the dataset regarding each inning and performance of tea m and its players in each inning. Following is the table that explains few important variables in this dataset.

Variables	**Description**
Inning	It describes if the first set of batting was going or the second set.

Over	Describes the current ball number of th e over.
Non striker	Name of player playing on the other en d of the pitch for that current ball/deliv ery.
Is super over	Is a flag if the match was a tie and extr a super over was played or no
Toss winner	Who won the toss
Toss decision	Decision made by toss wining team to either field or bat first
Result	If the team won the match or was it a ti e
Duckworth lewis method	This is applied when it rains, and matc h gets cancelled in midway
Team	There are eight teams each representing a state in India. (Teams)
Player of the match	Player who has highest score in that match
Venue	Location where the match was played. (Home ground or not)
Umpire 1,2,3	Wicket decision given by any one of th ese umpires and their names
Win by wickets	If the team wins without loosing any wi cket
Win by runs	If the team batting first wins the match
seasons	Year of the IPL league
Extra runs	These are the runs given by wide balls, no balls, penalty, overthrows.

Results and Discussion

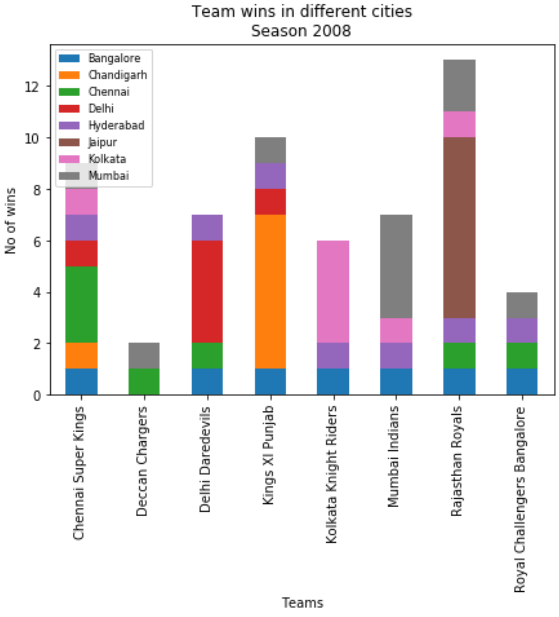
Analysis

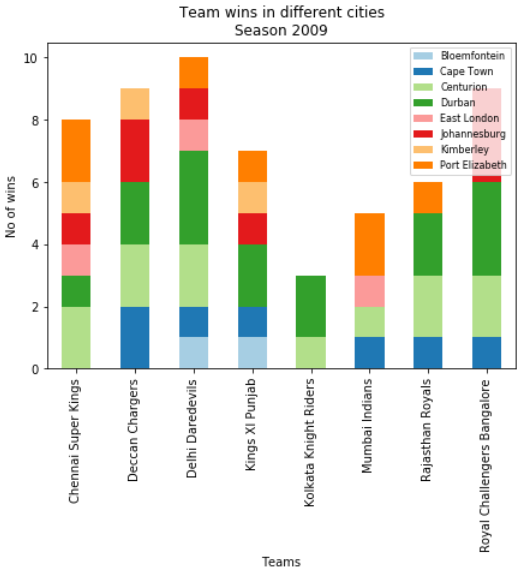
In this section we visualize the data to get an insight whether the factors like home ground matches, wining the tosses really reflect on the winning of the teams. We have analyzed the data for every season and found out that these factors do have effect on the winning team.

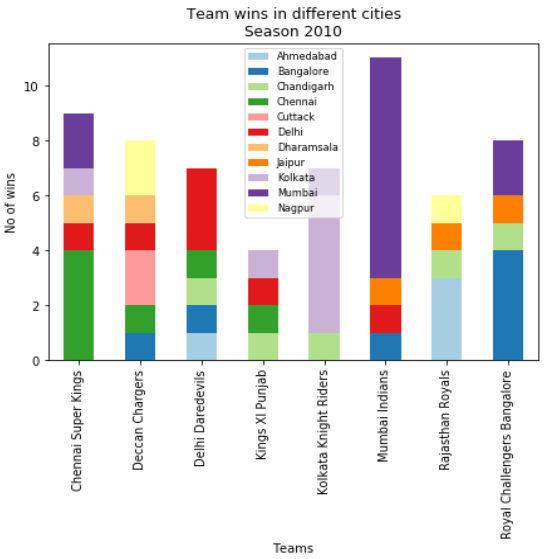
a. **Home ground advantages:**

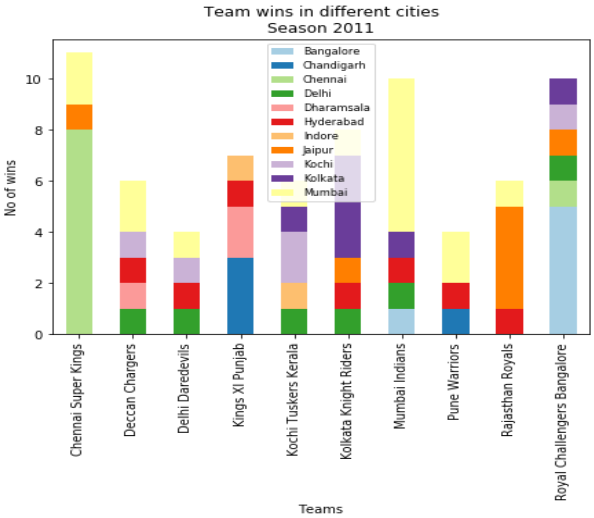
If we check for each team, we can clearly find out that winning teams had great advantage of home ground. For example, If we consider 2008 3 teams have won the highest number of matches when they were played at their home ground, viz: Chennai super kings in Chennai, Kings XI Punjab in Chandigarh(Capital of Punjab), Rajasthan

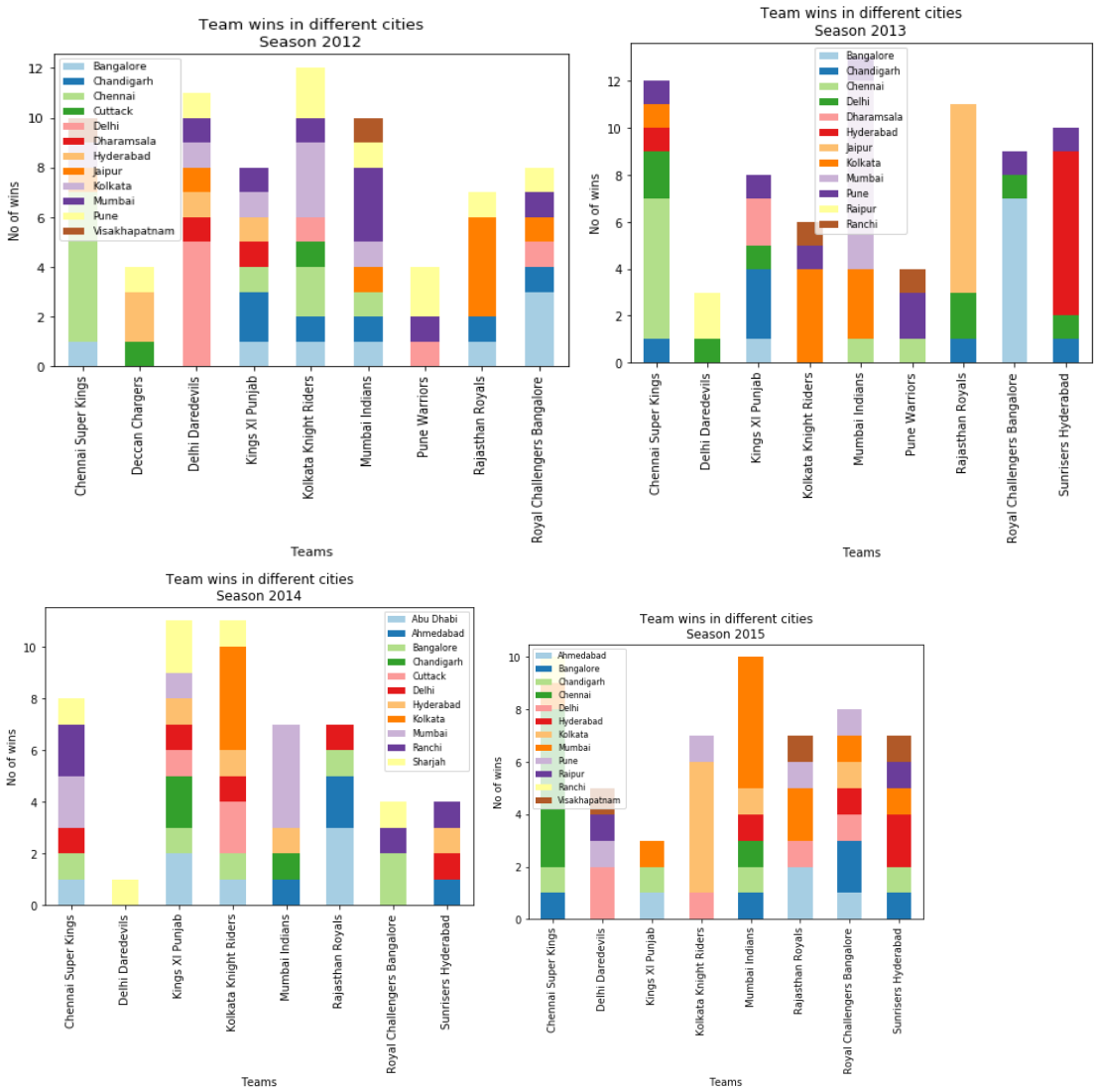
Royals in Jaipur(City in Rajasthan). Also, Rajasthan Royals shows the highest number of matches won in home-ground and result was that they had won the IPL season 2008. Similarly goes for Mumbai Indians won the IPL season 2013 with the greatest number of matches won in home-ground. But this does not work for 2010 as the chart shows that Mumbai Indians won the greatest number of matches in Mumbai but performed really bad in other places. Hence, Chennai Super kings were the winners of IPL 2010. To see this further we checked the performance of top players of each team. This will help tackle the above issue where we can find that players of winning team also play important factors along with the home ground factor to decide whether the team will win or no.











b. Player performance advantage:

Following are the graphs that represent the statistics of batsmen and bowlers. If we want to use the players individual performance as a factor to predict the result of a match, then following charts prove to us that it must be the performance of both the batsmen and bowlers and no individual player. A very good example is, Figure 1 where we can see the Virat Kohli as has the highest performance in 2016 but his team did not win the IPL league because we can see bowlers of his team are not in the top 10 according to Fig.3

Figure 1

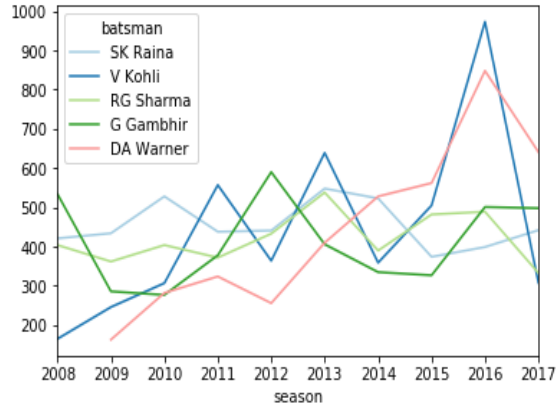


Figure 2

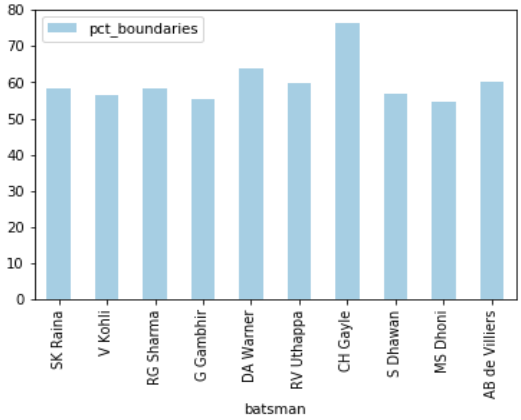


Figure 3

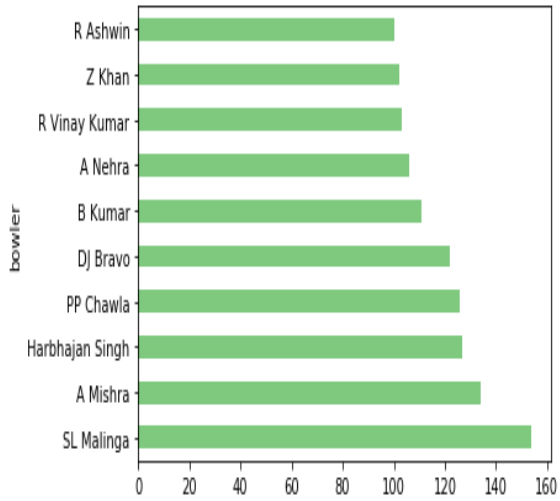
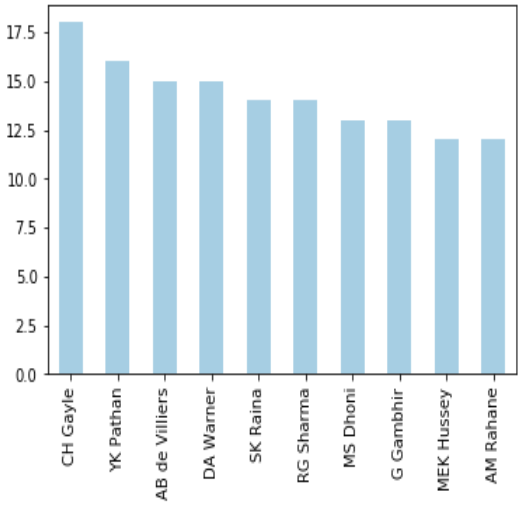
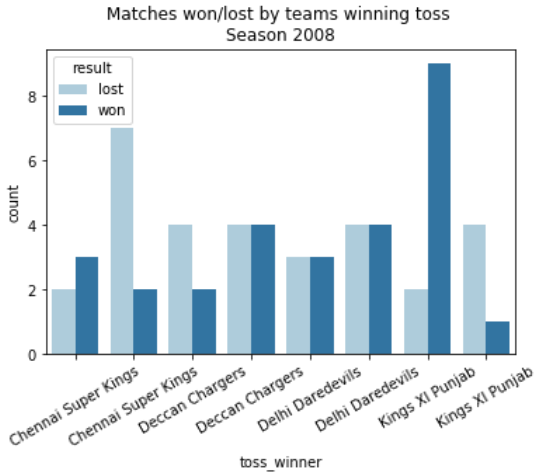


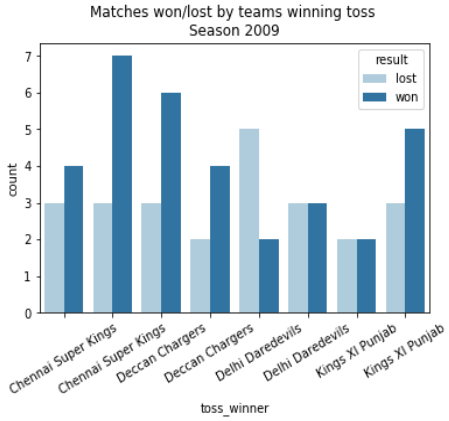
Figure 4

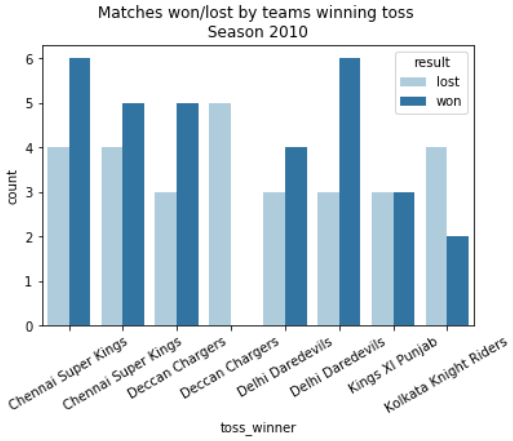


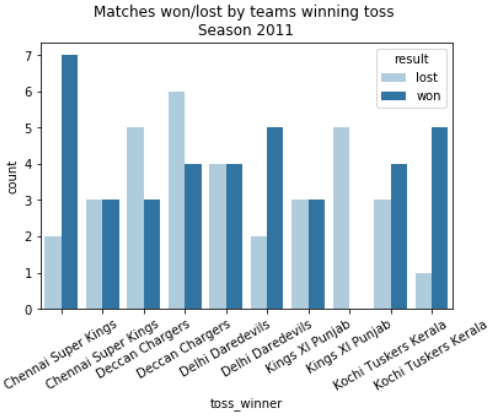
c. Wining Toss advantage:

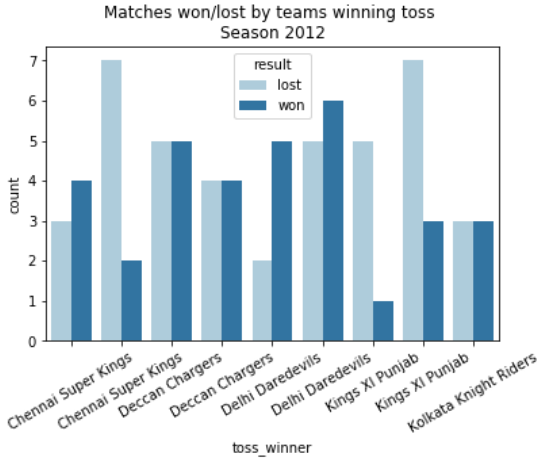
When we look at the individual charts for every IPL season we can visually see that there is some advantage to teams who won the toss. But, it is not a significant advantage. To see that clearly the pie chart shows that there is 56% of chance that the team would win if they won the toss. And this is not a good factor we can use to predict the wining of the team

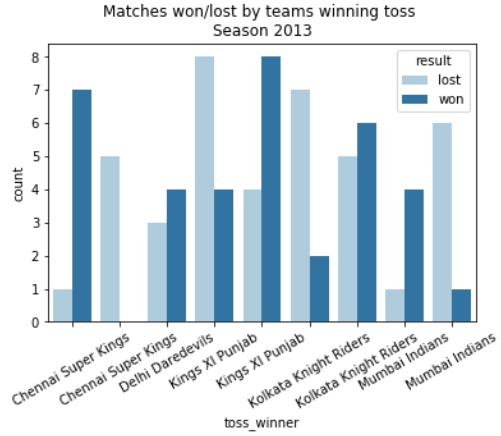

Matches won/lost by teams winning toss
Season 2008
result
lost
won
count
Chennai Super Kings
Chennai Super Kings
Deccan Chargers
Deccan Chargers
Delhi Daredevils
Delhi Daredevils
Kings XI Punjab
Kings XI Punjab
toss_winner

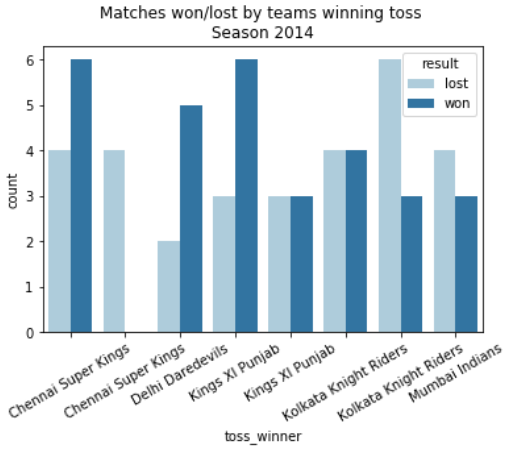

Matches won/lost by teams winning toss
Season 2009
result
lost
won
count
Chennai Super Kings
Chennai Super Kings
Deccan Chargers
Deccan Chargers
Delhi Daredevils
Delhi Daredevils
Kings XI Punjab
Kings XI Punjab
toss_winner

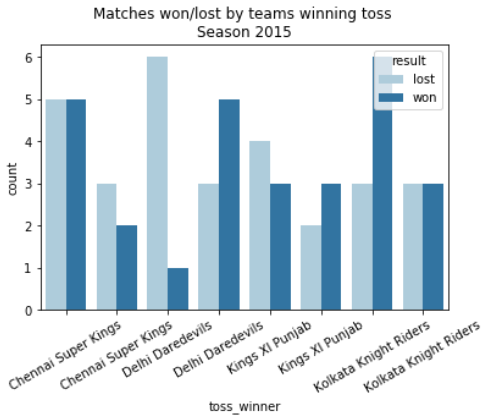

Matches won/lost by teams winning toss
Season 2010
result
lost
won
count
Chennai Super Kings
Chennai Super Kings
Deccan Chargers
Deccan Chargers
Delhi Daredevils
Delhi Daredevils
Kings XI Punjab
Kolkata Knight Riders
toss_winner

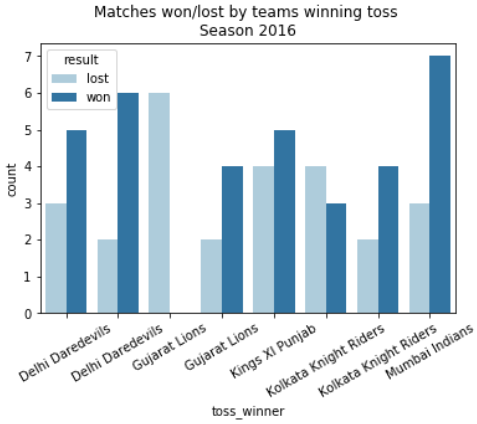

Matches won/lost by teams winning toss
Season 2011
result
lost
won
count
Chennai Super Kings
Chennai Super Kings
Deccan Chargers
Deccan Chargers
Delhi Daredevils
Delhi Daredevils
Kings XI Punjab
Kings XI Punjab
Kochi Tuskers Kerala
Kochi Tuskers Kerala
toss_winner

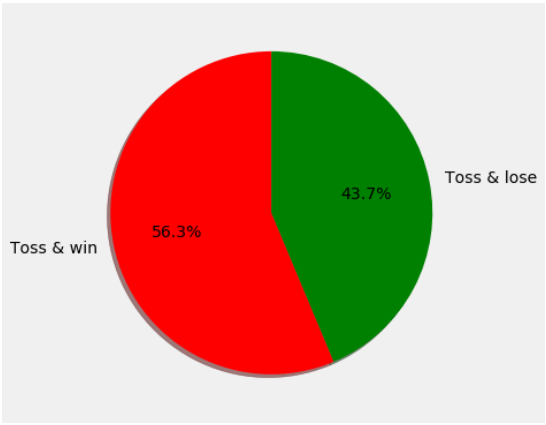

Matches won/lost by teams winning toss
Season 2012
result
lost
won
count
Chennai Super Kings
Chennai Super Kings
Deccan Chargers
Deccan Chargers
Delhi Daredevils
Delhi Daredevils
Kings XI Punjab
Kings XI Punjab
Kolkata Knight Riders
toss_winner


Matches won/lost by teams winning toss
Season 2013
result
lost
won
count
Chennai Super Kings
Chennai Super Kings
Delhi Daredevils
Kings XI Punjab
Kings XI Punjab
Kolkata Knight Riders
Kolkata Knight Riders
Mumbai Indians
Mumbai Indians
toss_winner









d. Prediction Model

To this end we have seen few important factors like venue, players statistics, toss winners. Now, we will check how these factors effect the prediction model to predict the wining team. For creating the prediction model, we have used ***[venue, toss winners, city, players stats].*** Later we tested this model on test data using above factors to predict the accuracy of model to predict the wining team. The results were as follows:

1. The individual accuracy to predict against toss winners was 23% which is very poor. Further testing the model on every individual factor gave prediction accuracy below 35%.
2. Secondly, I tried using all the factors to predict the winners among two teams and the accuracy increased to 88%. This concludes that to predict the winners we need to consider huge number of factors. As the game is very much dynamic and to get a good amount of accuracy data needs to be more mature. Currently, we have data for only 10 seasons and each table has less than 1 lakh rows.
3. Algorithms used was Random Forest Classifier to create the classification model since we have a binary output win/loss. Python sklearn library has a inbuilt method RandomForestClassifier() which takes in the inputs as n_estimators which creates the number of decision trees, in this case the decision tress take the feature samples

and record samples randomly. We need to specify the target variable. In this case **[winner]** is the target variable.

Limitations and Future scope

Cricket is a huge game. And after doing this overall exploratory analysis and finding patterns regarding the teams and players performance. We can say that to predict the outcome of game. Instead of only focusing on Algorithms it is more important to have data related to weather conditions, ground conditions, form of the team, format of the game, venue. Because cricket is a very unpredictable sport. Also, not a lot of research is done in this sport so before deeper analytics researchers will have to work on data creating and collection. A huge dataset must be created for better outputs. In future scope weather conditions and individual player performance in all formats of matches need to be studied. Because a player can preform good in test format matches but might do good in T20 format. This kind of study needs to be done in future analysis.

References

https://www.news18.com/cricketnext/news/analyzing-the-economics-of-the-indian-premier-league-1944069.html

https://www.statista.com/topics/4543/indian-premier-league-ipl/

https://sportstar.thehindu.com/cricket/ipl/ipl-indian-premier-league-revenue-growth-media-rights-sponsorship-deals/article30360212.ece

https://www.academia.edu/36381573/INCREASED_PREDICTION_ACCURACY_IN_THE_GAME_OF_CRICKETUSING_MACHINE_LEARNING