WiDS ‘22 - ‘23 Final Documentation

IPL Data Analysis and Winner Prediction(UID-19)

Mentors:- Vivek Parmar, Ishaan Garg

200040094 200100070

Team Member Name	Roll Number	Email-Id
Aditya Kumar	**210020005**	210020005@iitb.ac.in

Introduction to Problem Statement

• To perform **Exploratory Data Analysis** on IPL datasets and get some common insights like obtaining best performing teams, luckiest teams, most popular venues, best players, etc. through visualization using Seaborn, Matplotlib, and Plotly in Python.
• To implement **Predictive Data Analysis** using Linear, RF, and SVM Regressors to predict the final scores of an inning.
• Engineering features like the current score and batsman-bowler statistics from the raw data and curating Logistic Regression, SVM, and Decision Tree Classifiers for predicting match-winner with good accuracy

Existing Resources

Resources provided by our mentor:
WiDS IPL Data Analysis and Winner Prediction

Proposed Solution

Primarily in the project I have done the following things,
• I have studied the data and carried out exploratory data analysis to identify factors which might play an essential role in determining the winning team in a particular match.
• During EDA, I have also tried to identify trends in match results based on the fact which team wins the toss, which team bats first, and which teams bowl first
• I have further used a regression model to understand which factors play a role in what proportion in determining the match winner.

With this done, my primary aim is to use the Plotly library in Python to render interpretation efficiently using graphs. Performance data using visual analysis help select players for future matches and provides additional information about the player and team profiles. The aim is to provide detailed insights numerically and graphically to understand the tournament's history and make data-driven decisions like predicting the winning side of a particular match in the future with an acceptable accuracy solely based on the parameters mentioned above. I realize that the winners are decided by the squad playing at that time.

Methodology & Progress (Mention the work done week-wise)

Week	Start Date	End Date	Work
Week-1	15 December	21 December	Basics, Overview, Python Basics
Week-2	22 December	28 December	Data Cleaning, Preprocessing
Week-3	29 December	4 January	Exploratory Data Analysis(EDA)
Week-4	5 January	11 January	Learnt Feature Engineering, Classification Models
Week-5	12 January	15 January	Applied the previous week learnings
Week-6	16 January	22 January	Regression Models

Week-7	22 January	29 January	Final rechecking and report-making

Methodology

The data consists various variables whose name either has been used the same or has been changed according to the need. The number of teams has been a changing variable in these 15 years, with eight teams in the debut season. There has been a constant addition and deletion of various teams like

1. Deccan Chargers was active from 2008 to 2012.
2. Pune Warriors India was active from 2011 to 2013.
3. Kochi Tuskers Kerala only played one season in 2011.
4. Sunrisers Hyderabad was introduced in 2013 and is currently active.
5. Chennai Super Kings and Rajasthan Royals did not play in 2016 and 2017.
6. Gujarat Lions and Rising Pune Supergiant played only in 2016 and 2017.
7. Rising Pune Supergiant played as Rising Pune Supergiants in 2016 (i.e. deleted the last ‘s’ in 2017)

All this inspection and cleaning was done in the datasets to obtain a more usable form of data.

Methodology for Prediction

I experimented with various models and predictor variable combinations. I narrowed my choices down to 'team1', 'team2', 'total_runs', 'cum_wickets','field'(column which tells weather the winning team has elected to field or not), 'toss_winner', based on the correlations existing. In various models, I experimented with the various combinations of features used for predicting, the results of which are tabulated below.

First, I used a linear regression model to predict the final score at every instance of the match. Linear regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique.

Secondly, I experimented with the Random Forest Regressor. The Random forest or Random Decision Forest is a supervised Machine learning algorithm used for classification, regression, and other tasks using decision trees.

The Random forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a set of decision trees (DT) from a

randomly selected subset of the training set and then It collects the votes from different decision trees to decide the final prediction.

Results

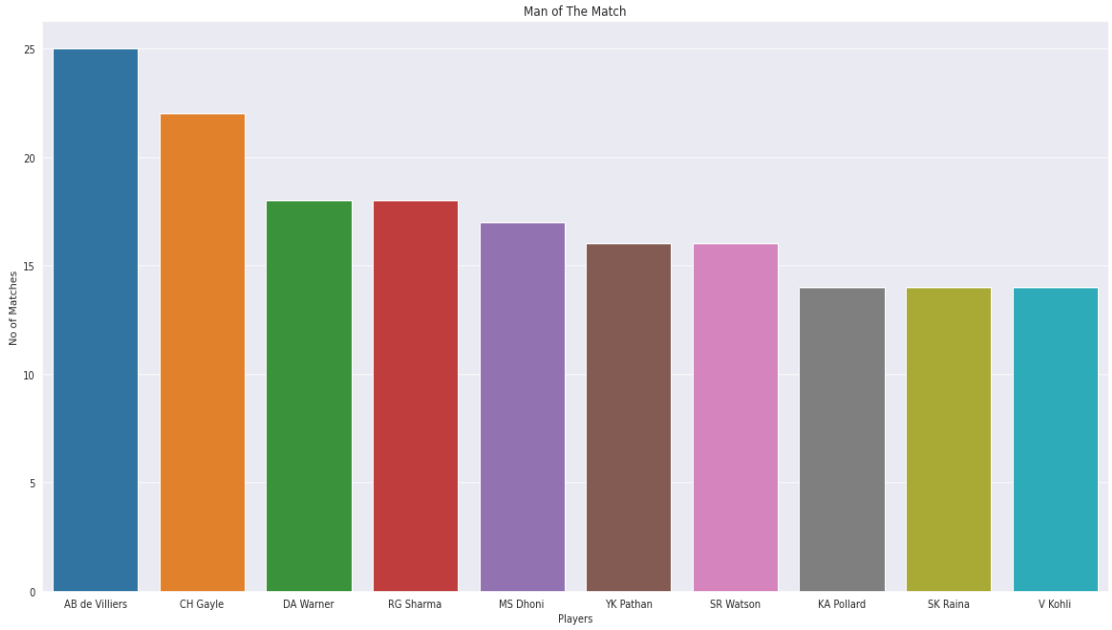
Drive link:
https://drive.google.com/drive/folders/1TcVesYdqMqkg1oXaEnQeVQxsAng1p6Cg?usp=sharing

The Elementary Data Analysis done can be seen as,

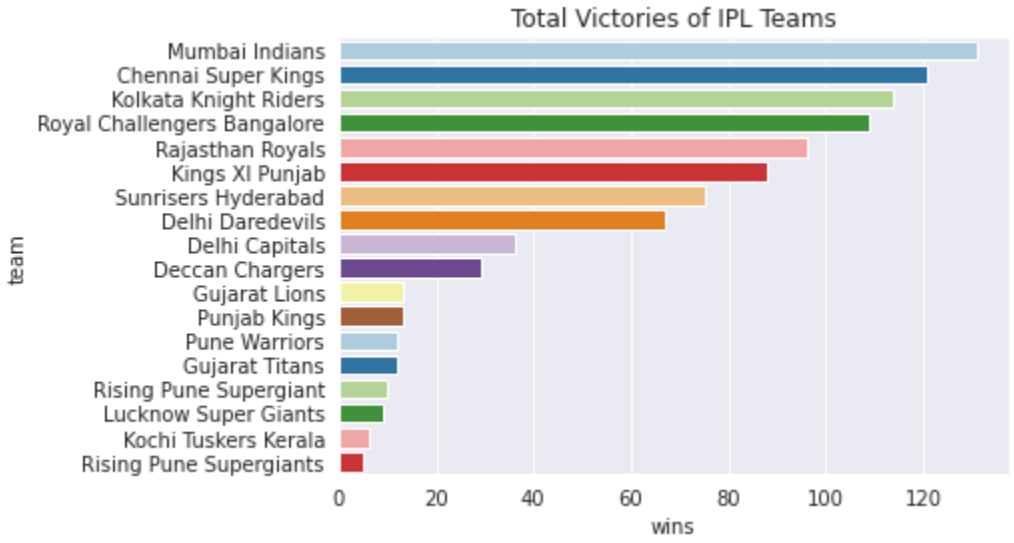
- Maximum Number of wins by any teams in any particular season:

```
Season   WinningTeam
2007/08  Rajasthan Royals          13
         Kings XI Punjab           10
         Chennai Super Kings        9
         Delhi Daredevils           7
         Mumbai Indians             7
                                   ..
2022     Punjab Kings               7
         Kolkata Knight Riders      6
         Sunrisers Hyderabad        6
         Chennai Super Kings        4
         Mumbai Indians             4
Name: WinningTeam, Length: 126, dtype: int64
```

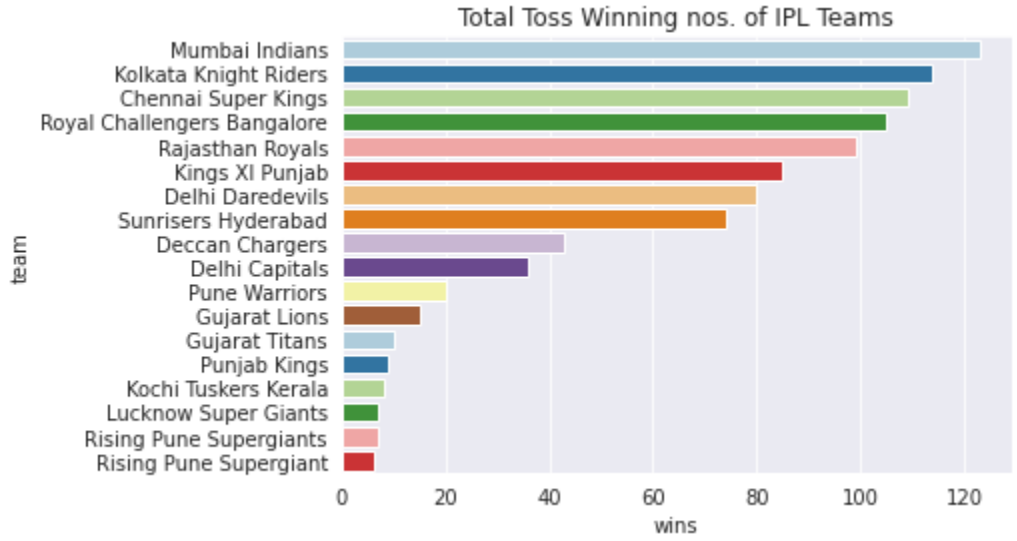
- Players to get most number of ‘Man of the Match’ award:



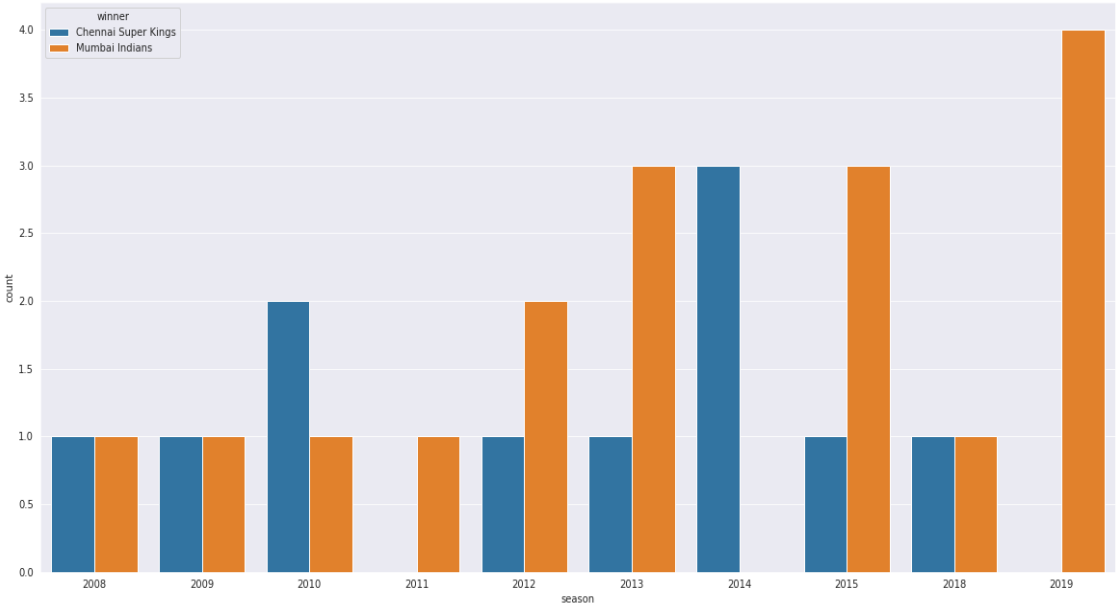
- Most successful Teams:



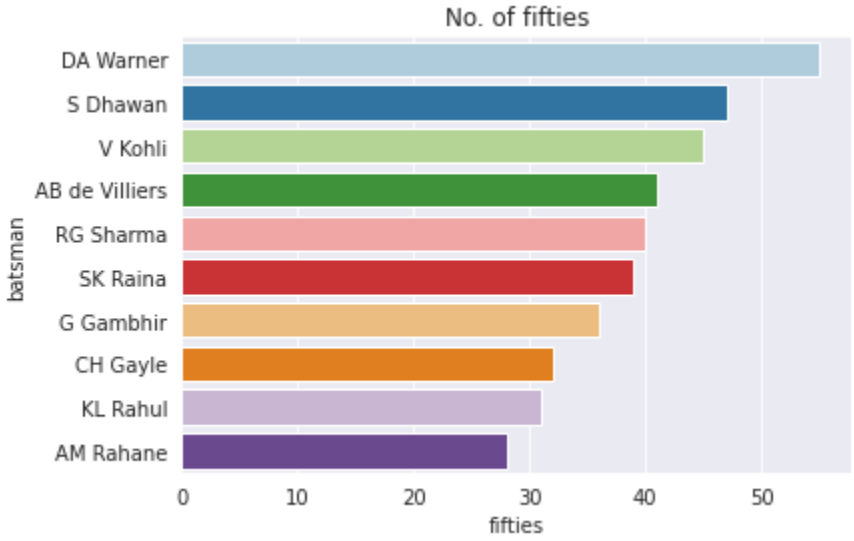
- Teams winning most number of tosses:



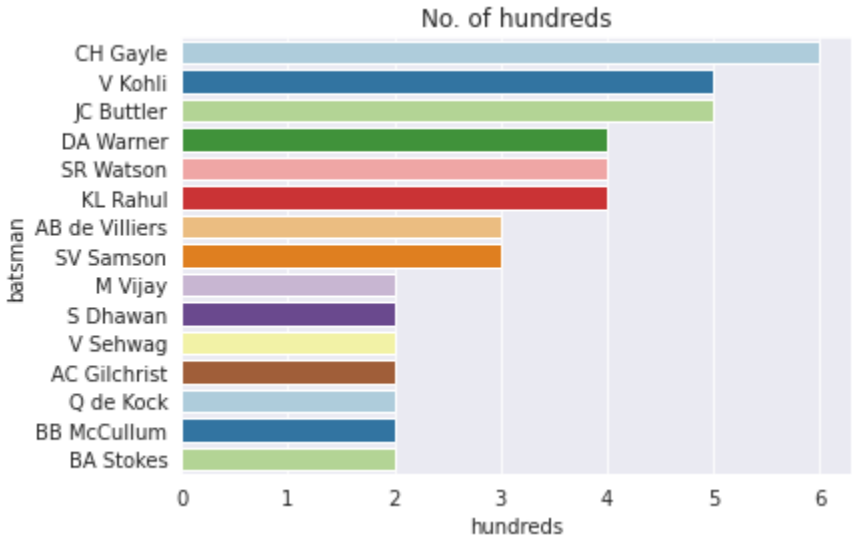
- Team1(CSK) V/S Team2(MI):



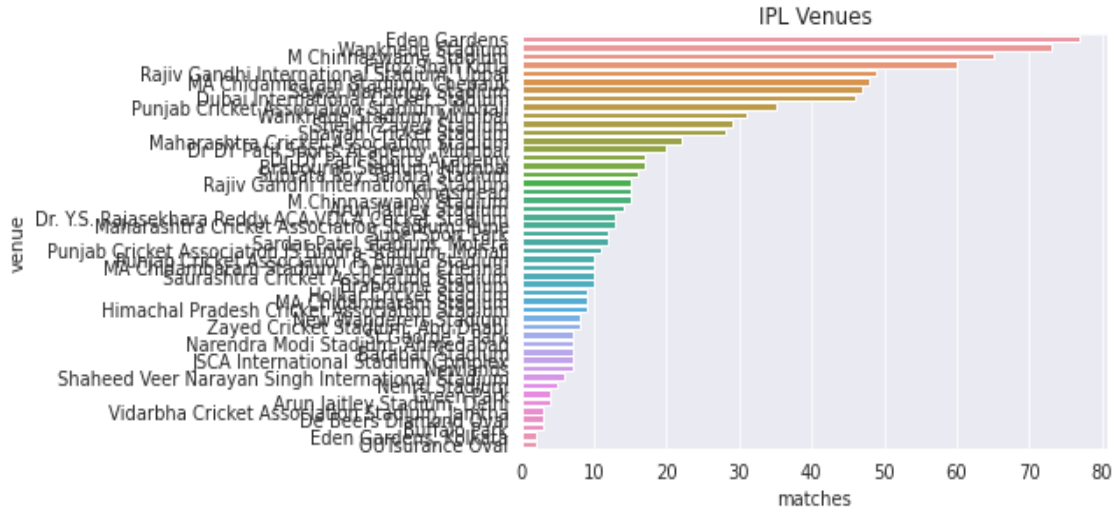
- Most number of 50s:



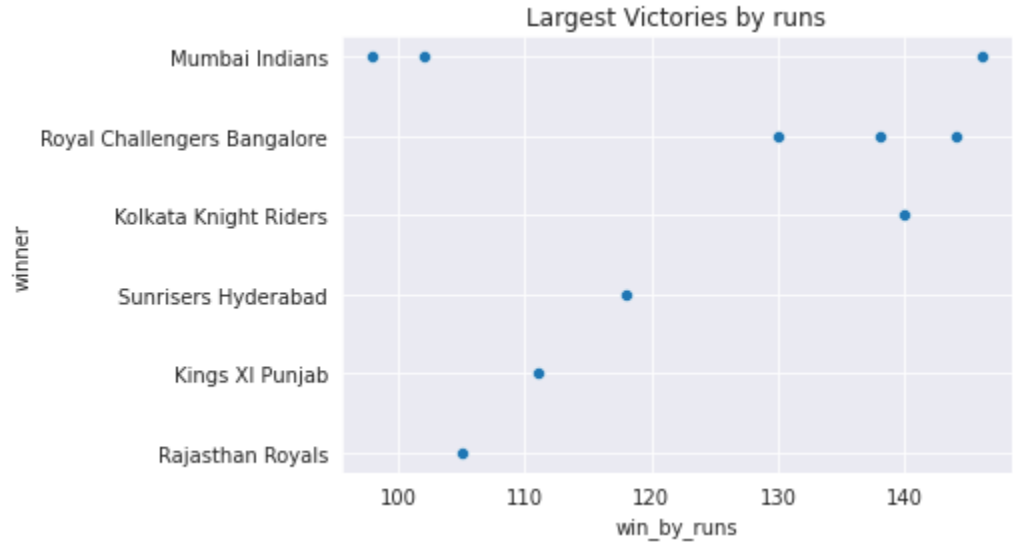
- Most number of 100s:



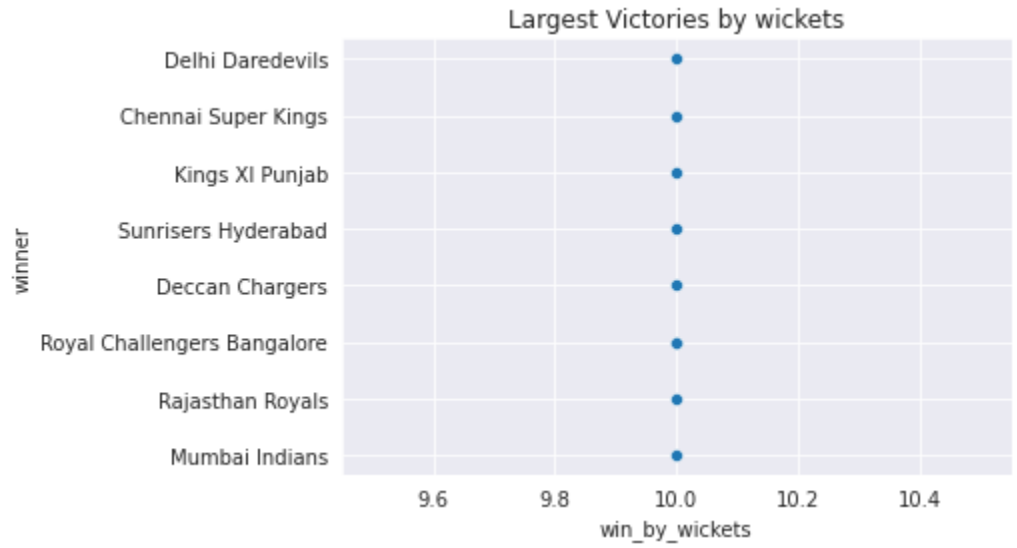
- Number of matches hosted by a particular venue:



- Largest victories by runs:



- Largest victories by wickets:



Learning Value

This project provides useful insights from the IPL dataset about what are the best performing teams and players along with the toss analysis and its importance in winning games for the strongest and weakest teams. Best performing players of IPL can be listed with the most MoM awards analysis. Toss decisions have more or less no influence on winning.
Random Forest Regressor had the best accuracy among the three models used to predict the final score at a given instant of the match.
Even though the accuracy is not high enough to be extremely useful, owing to the limited domain of data available and a variety of factors IPL matches depend on, it gives a basic idea about the strategies and methodologies used in designing a solution to this Machine Learning problem.

Tech-stack Used

- Statistics and Python Basics
- Google Colab
- Required Python libraries (Numpy, Pandas, Sklearn, Matplotlib, Seaborn)
- Plotly (Another highly interactive plotting library)
- ML Models
- Loss function
- Regression analysis
- Classification analysis

Mentor's resources(Most useful🙂) :

WiDS IPL Data Analysis and Winner Prediction

Suggestions for others

A basic understanding of Machine Learning using Python is sufficient to follow my solution report. One can familiarise oneself with the various libraries used in the code easily. The report also explains the various prediction models used, in the methodology section.

References and Citations

1. IPL Score Prediction using Deep Learning - GeeksforGeeks
2. https://www.youtube.com/playlist?list=PLLssT5z_DsK-h9vYZkQkYNWcItqhlRJLN
3. https://www.youtube.com/playlist?list=PLu0W_9lII9ai6fAMHp-acBmJONT7Y4BSG

Thank you