Data Analysis and Winner Prediction of IPL (Indian Premier League)

Ashish Singh Chauhan
Mechanical Engineering
IIT Bombay
200020031
200020031@iitb.ac.in

***Abstract*— Cricket has been an integral part of the Indian culture for decades now, and the Indian Premier League (IPL) has been the torchbearer of this craze and culture for almost 14 years. The outcome of a particular match in the league depends on various variables like the playing 11 that day, previous performance, the venue, and other volatile variables like pitch and weather conditions. My work is based on models of linear regression and Exploratory Data Analysis from the data collected from trusted websites like Kaggle. I have also found out trends or relations among various variables in the data present to help analyze different matches. This analysis helps us to predict the outcome of any match provided that the variables mentioned are provided. This can help in further decisions like Fantasy Leagues and other spheres. The data collected is only for IPL, but the same process can be followed for similar data (if available) for different cricket tournaments across the globe.**

I. INTRODUCTION

Started in 2008, the Indian Premier League (IPL) is a professional Twenty20 cricket league, contested by ten teams based out of ten Indian cities. The Board of Control founded the league for Cricket in India (BCCI) in 2007. It is usually held between March and May of every year. The IPL is the most-attended cricket league globally and in 2014 was ranked 6th by average attendance among all sports leagues. The league began with eight teams and has had 14 seasons till now.

Data Science is extensively used in sports and cricket, for that matter, has a good amount of use of Data Science and Algorithms. Real-time data analytics can help gain insights even during the game for changing tactics by the team and by associated businesses for economic benefits and growth.

Out of the several projects that I saw related to data science, I chose this as mentioned I felt that IPL is an integral part of myself and the fact that it is not only limited to T-20 Cricket but also for other formats like ODIs and Test Matches. Moreover, a similar analysis can be performed for other sports like football, hockey and baseball, having somewhat similar parameters.

The data source that I have primarily used for drawing conclusions is from the online sources like Kaggle.

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 important 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 that which team wins the toss, which team bats first, which teams bowls 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 provide 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, but this is a very volatile parameter and can be ignored for now.

II. BACKGROUND AND PRIOR WORK

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. While basic work in this sector exists, as can be seen in the references, I tried to experiment with a diverse variety of models and a lot of variables in order to get the best possible prediction with the available data.

III. DATA AND METHODOLOGY

The dataset I have used is from the online repositories from a trusted data repository website named Kaggle.
I have primarily used four datasets i.e. two for each part of the code.

1. IPL Ball-by-Ball 2008-2020.csv
   The data shows all the ball-by-ball outcome of a particular match ranging from the striker, to the outcome of the ball like a single, a six or a wicket.
2. IPL Matches 2008-2020.csv
   The data shows all the details of a particular match like the date, the venue, and the winning team, the man of the match and umpires of the match.
3. matches.csv
   The data provides the details as match result, venues, umpires, win margin (in terms of wickets or in terms of runs).
4. deliveries.csv
   The data shows various variables like the players batting per bowl, the bowler, the outcome of each ball, reason of dismissal (if) and type of extra run conceived (if).

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 14 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.

Linear regression predicts the output of a continuous dependent variable. Therefore the outcome must be a continuous value. This model was used to prevent any chances of overfitting.

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)

As IPL is generally played between March and May of every year and this includes the disparity in 2020 (September- November) due to COVID-19 Pandemic. Many other tasks have been performed for Data Inspection and Cleaning like,

1. The name ‘Bangalore’ has been replaced with ‘Bengaluru’ everywhere in the dataset.
2. Missing values have been filled to prevent errors.

All this inspection and cleaning was done in all the four datasets to obtain a more usable form of data. For example, the final output for IPL Matches 2008-2020 dataset has **816** rows and **17** columns.

	id	city	date	player_of_match	venue	neutral_venue	team1	team2	toss_winner
0	335982	Bengaluru	2008-04-18	BB McCullum	M. Chinnaswamy Stadium	0	Royal Challengers Bangalore	Kolkata Knight Riders	Royal Challengers Bangalore
1	335983	Chandigarh	2008-04-19	MEK Hussey	Punjab Cricket Association Stadium	0	Kings XI Punjab	Chennai Super Kings	Chennai Super Kings
2	335984	Delhi	2008-04-19	MF Maharoof	Feroz Shah Kotla	0	Delhi Daredevils	Rajasthan Royals	Rajasthan Royals
3	335985	Mumbai	2008-04-20	MV Boucher	Wankhede Stadium	0	Mumbai Indians	Royal Challengers Bangalore	Mumbai Indians
4	335986	Kolkata	2008-04-20	DJ Hussey	Eden Gardens	0	Kolkata Knight Riders	Deccan Chargers	Deccan Chargers

Fig. 1. Snapshot of the final table for IPL Matches 2008-2020 dataset

NOTE: The table above is just a snapshot, hence does not have all the columns of the actual dataset.

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. It was used by me in the winner prediction task.

Next I used a SVM Model. Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

IV. EXPERIMENTS AND RESULTS

Since we have two final objectives here, analysing the different variables present in a match and then using them in the results section to make a good model to predict outcome of any future match given that all the data regarding the variables mentioned are provided. The EDA gives us the idea about which variables are important in predictive tasks and which are not. The Elementary Data Analysis done can be seen as,

Meaningful Exploratory Data Analysis is what helps us in preventing problems like overfitting and underfitting.

1) Total Matches played in a season : As seen in Fig. 1, total matches increased in the years 2011, 2012 and 2013. The main reason behind this was addition of teams in IPL. Hence these three seasons shouldn't be used as a testing dataset. Thereon, the total games are almost constant with only slight differences mainly due to calling off the match due to weather conditions.

Fig. 1. Total Matches played in each season

2)Most Valuable Players: This gives us insights about the most impactful players over the years in each season depending upon the number of Man of the Match titles won.

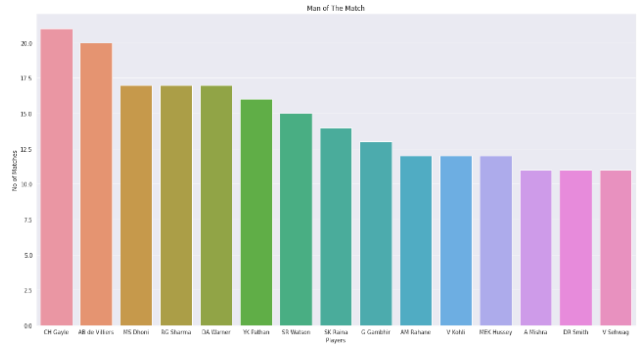


Fig2:Total Man Of Matches over the years

3)Max Wins:Current and Past strongest Teams.

Fig3: Maximum Number of maches won every season

4) Top Cities : The analysis of the data gives us that Mumbai is the favourite spot for the conduction of IPL matches probably because of its connectivity and good quality grounds.

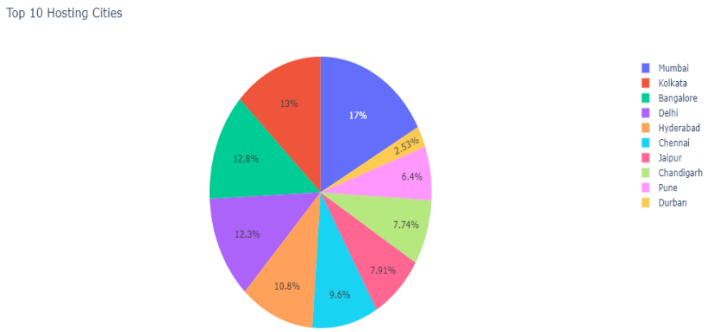


Fig4:Top Cities

5) Toss Analysis : In fig. 5, as we can see, the trend of toss decision has reversed over the years. Upto 2013, batting used to be considerate option but in the recent years fielding has completely dominated its counterpart.

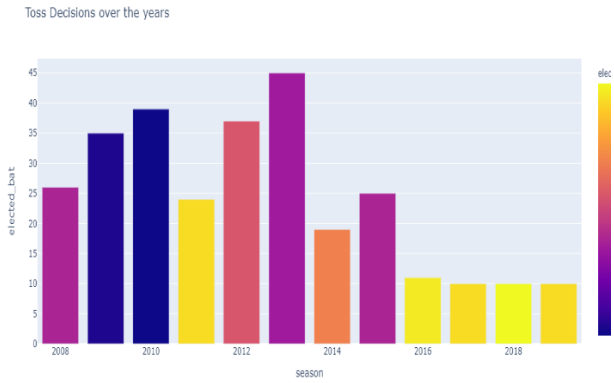


Fig5: Toss Trends

6)Team rivalry : Year wise statistics of teams rivalry are also plotted. It can help a team identify against whom are they strongest and weakest in the recent years providing them a win/loss ratio. Fig. 6 includes CSK vs MI analysis over different seasons.

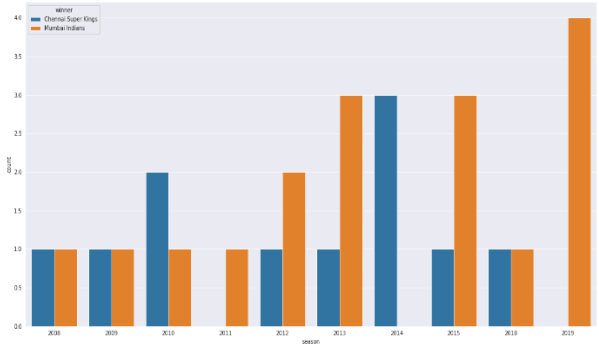


Fig 6:CSK v/s MI over the years

7)Most 50s and 100s: Figure 7 shows the top 25 players over the history of the IPL on the basis of number of 100s scored whereas Figure 8 shows the top 25 players over the history of the IPL on the basis of number of 50s scored.

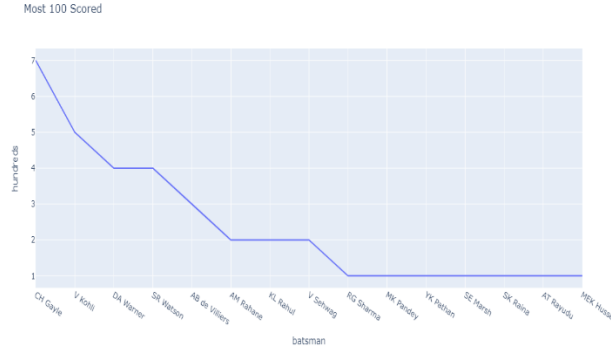


Fig 7: Most Number Of 100s

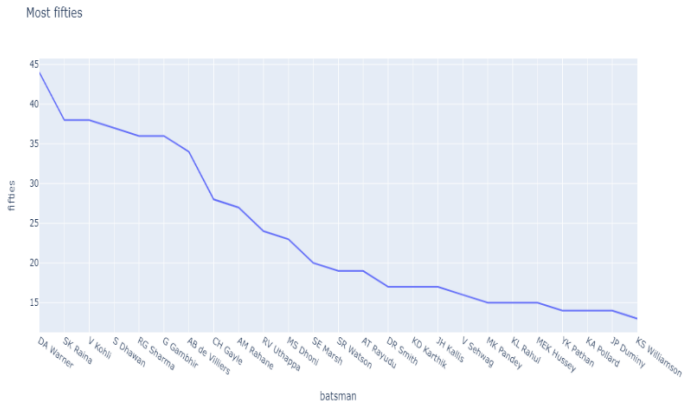


Fig 8:Most Number of 50s

7)Greatest victory on the basis of margin of runs over the years:Fig 9 shows the 5 greatest victories of all time.

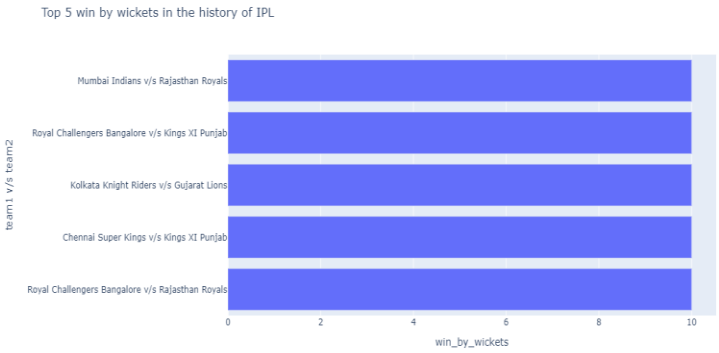


Fig 9: Greatest victory of all time in IPL on the basis of run margin

8)Greatest victory on the basis of number of wickets over the years: Fig 10 shows the 5 greatest victories of all time.

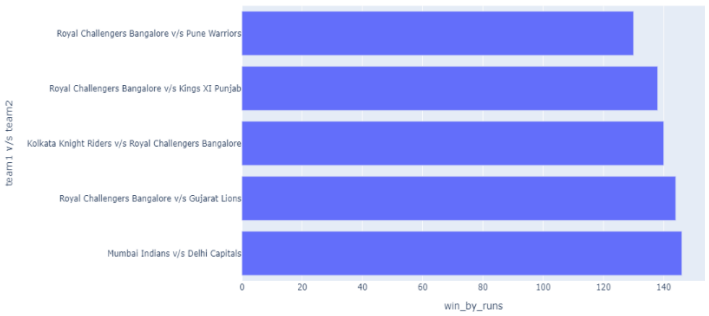
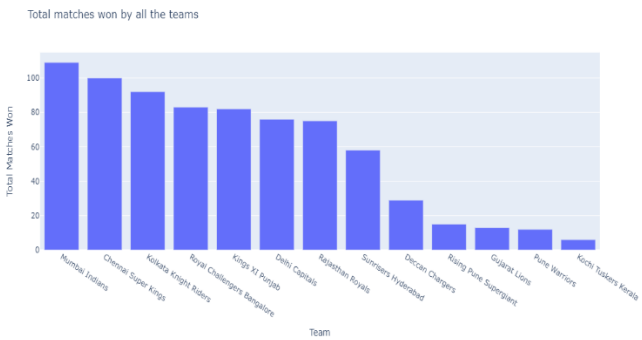


Fig 10 :Greatest Victories of all time in IPL.

9)Total Matches won by all the teams over the years:



10)Toss Win percentage: In this analysis I found that Deccan Chargers is the most lucky team in the history of IPL with the toss win percentage of 57.33. Further also found that Sunrisers Hyderabad is the least lucky team with the toss win percentage of 42.99%.

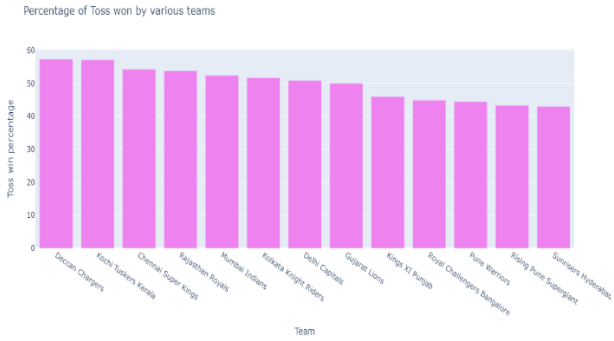


Fig 12: Toss win percentages of various teams

IV EVALUATION

The EDA helped us to reduce the number of features from 34 to 16 of which 12 will be used for modelling.

A. Metrics Used

1) Machine Learning: For prediction of total runs, R^2 score and MSE (Mean Square Error) were used. A new accuracy metric was defined which gave accuracy if the score is within a limit of +-10. This accuracy function was used throughout the total runs prediction model.
   For predicting the winner, classifiers were trained and the default accuracy model of scikit learn was used. Teams were used as classes which were label encoded on the basis of their strength concluded from EDA.
2) Deep Learning: For predicting total runs, the accuracy function was used same as in the Machine Learning. In addition, validation and training loss were calculated at each epoch.
   For predicting the winner, NN classification was used and its default accuracy model. Similarly, validation and training loss was calculated at each epoch.

B. Models Used - Score Prediction

1) Linear Regression : Linear regression is the simplest of model. This model was chosen to avoid any chances of overfitting.
   Results: As expected, results weren't much fascinating with 48.86% accuracy on the splitted data obtained through train split function.
2) Random Forest Regressor: The tree growing in Random Forest happen in parallel which saves a lot of time

Hyperparameter Tuning: As even with the tuning the model was highly overfitting. Hyperparameter tuning on the split data set gave the best parameter as n_estimators=500.

Results: Selecting the parameters obtained from GridSearchCV, we get accuracy as 75.58% on the split data set obtained using train_test_split function on the IPL 2008-2020 Complete ball by ball dataset.

3)Neural Networks(NNs): Neural Networks are the modern prediction models. Analogies similar to human nervous systems are drawn. Each unit is considered as a neuron which fires an output through an activation function upon receiving an input. Several layers of units are stacked together to predict the results.
Input layer has number of units equivalent to the number of features which are passed to the second layer through activation functions. Functions like sigmoid, RelU and Leaky RelU are used. Adam Optimizer and SGD (Stochastic Gradient Descent) are used for back-propagation. SGD is usually not chosen since it suffers vanishing gradients problem. Dropout regularization is used to prevent overfitting.

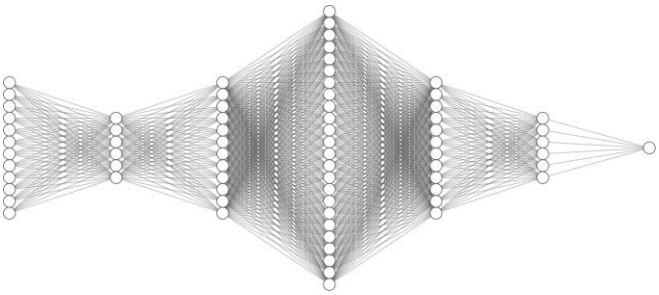


Fig: Neural Networks

7 layers were build with 5 of them being hidden layers. A total of 85k parameters were trained. Training the model on the same input as used in previous models didn't come out to be of much use as an accuracy of around only 49% came. Limited amount of data is the primary reason why neural network isn't performing well over the dataset.

C Models Used- Winner Prediction

1) Random Forest Classifier: All parameters same as mentioned above.
   Results: Accuracy Score:93.83%

2) Support Vector Classifier(SVC):Since it is a classification problem, SVC is used.
   Results: Accuracy Score: 24.83%
3) Neural Networks: While neural networks weren't fruitful in predicting final score, quite the opposite happened in the classification problem. A 6 layer model was trained with 4 hidden units. Stochastic gradient descent was used as the optimizer.
   Results: Accuracy score: 90.47%

V. Learning, Conclusions and Future Work

This paper provides useful insights from IPL dataset about various teams & players along with the toss analysis and its importance in winning games for the strongest and weakest teams. Sponsors can do marketing in only certain cities citing highest spectators involved. The prediction of final score is done on the basis of teams involved in the match, the current over/ball, current score as well as the total wickets which have fallen, toss winner and its decision.

Random Forest Regressor had the best accuracy among the three models used to predict the final score at a given instant of the match. Further Random Forest Classifier had the best accuracy among all the three models used to predict Winner.

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.

The future scope includes having a more extensive dataset containing more information about the IPL than the one which we have right now. The dataset does not consist of various volatile variables like pitch condition, weather condition, playing eleven, recent performance of each player, etc.

More collection of data and analysis will help in getting more concrete answers in future and if combined with the above mentioned variables in suitable proportion, i.e. some kind of algorithm if can be developed forming some relation of all the variables, this model can be used extensively in future IPL matches, even by the teams themselves to have a better insight about the matches and changes they should make. This model, can further be applied to various other formats of cricket like ODI(s) and Test Matches. Also, to other sports like football, hockey and other sports having such similar variables.
Moreover, this model if completed with the said variables can also be used by general audience in many other spheres like Fantasy Leagues, an industry worth more than $ 1 billion just for IPL, and much more billions for other cricket formats and sports mentioned above.

Acknowledgment

I would like to thank Analytics Club, IIT Bombay for hosting WIDS(Winter in Data Science) Bootcamp and providing a bunch of students with a mentor and a project along with great appreciation to my mentor, Prayas Jain to provide with the best resources enriching this experience.

References

[1] Aur élien Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition
[2] IPL Cricket Match prediction,Praveen Sridhar
[3] IPL Score prediction using Deep Learning,
[4] url: https://www.geeksforgeeks.org/ipl-score-prediction-using-deep-learning/