IITB DS203 Programming for Data Science
IPL Data analysis and Prediction

1. Prateek Jha (200040106)
2. Raghav Rander (200040113)
3. Sharad Vishwakarma(200040132)

I. Introduction

Machine learning is a subfield of artificial intelligence in which real-world issues may very well be solved. This approach does not necessitate programming and relies solely on data learning if the machine learns from past data and predicts the outcome accordingly. Decision trees, heuristic learning, knowledge acquisition, and mathematical models are all used in machine learning approaches. Today, there is a high demand for cricket, with many individuals concentrating on data analysis and prediction utilising machine learning technology. Machine learning is used to analyse and forecast IPL data, which is crucial in player selection, strategy making and can be used by people in fantasy sports leagues. The selection of players is influenced by a number of things. The national team is selected by the team board and coach, with the captain playing a larger part in the process. Examining the average scores of all the team's players vs the other team's players. As a result, the success of the winning team of individual players is mostly dependent on the average of the details of past encounters. The crew selects the greatest batting and bowling performances, as well as analyses all-around performances. The project begins by using Python to push IPL match data from games played between 2008 and 2020. Next, pre-processing, data analysis, and visualisation are performed. Finally, a model is developed that forecasts each team's final score and likelihood of winning. We employ machine learning techniques like Random Forest, Linear Regression, and Logistic Regression while creating models. Thus, these algorithms efficiently forecast the anticipated score of each batters, the projected overall score of the team, and the winner.

II. Objectives

The following are the objectives of the project:

- Exploratory data analysis to analyze visually the given data sets and make inferences through it
- To Determine important features, Drop Correlated features and Define important features needed to build the model
- Regression Models- Predicting the runs made by the team in 20 overs through different models
- Classification models- Predicting the team which will win the match
- A comparative study of the accuracy of the models such as Random Forest, Linear Regression, and Logistic Regression which have been implemented for the mentioned predictions.

III. Why it is an interesting project

In India, the majority of people think about cricket when they hear the word "sports." To make this sport more exciting, other formats like Test, ODI, and most recently Cricket T20 have been introduced. The Indian Premier League, or IPL, which is played amongst franchises in different Indian states, has recently gained popularity not just in India but also throughout the rest of the world. Due to the enormous amount of data produced by a single player to an entire line, data science and machine learning are playing an ever-growing role in cricket. To make predictions about things like the team's first inning score and the likelihood that the second team will win, we use the facts and statistics that are currently available. The teams themselves may be particularly interested in this study and the predictions produced in terms of performance analysis, strategy development, player selection, reverse strategizing opponents, etc. Fantasy leagues have expanded the usage of such information in the sporting world by allowing fans to create their own teams and compete for prizes.

IV. Data sources

The data-sets used for analysis and prediction were collected from www.kaggle.com.
Two data-sets have been used. One for overall matches data and one for ball-to-ball data for the full 2008-2019 period. Both the data-sets are linked by the 'id' column which represents the matches uniquely. Some of the useful features present in the data-set 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.

V. Data Pre-Processing

The data-sets used for analysis and prediction were collected from www.kaggle.com.Two data-sets have been used. One for overall matches data and one for ball-to-ball data for the full 2008-2019 period.Both the data-sets are linked by the 'id' column which represents the matches uniquely. Some of the useful features present in the data-set 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.

VI. EDA Inferences and figures

The following are the visual analysis results obtained by EDA:

The figure below shows which team has the maximum number of wins in a particular season. For instance Rajasthan Royals had 13 number of wins in 2008 which was the highest for that season.

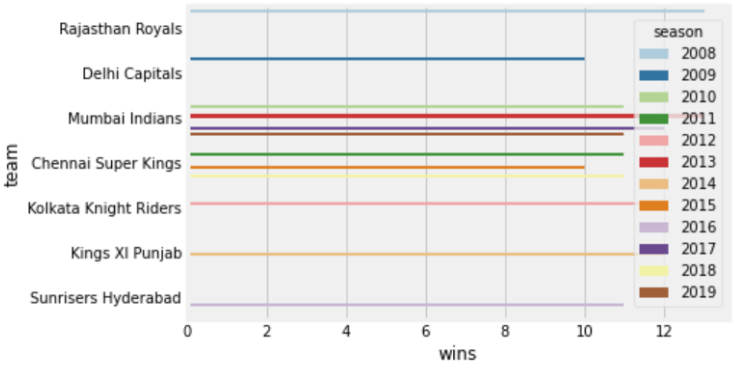

Fig. 1. Plot showing teams with maximum number of wins in a season

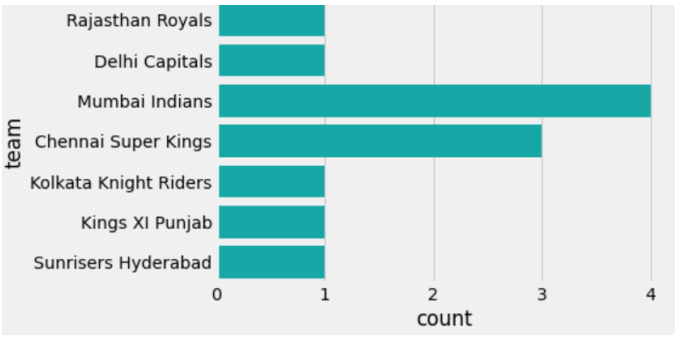

Fig. 2. Plot showing teams with number of titles won

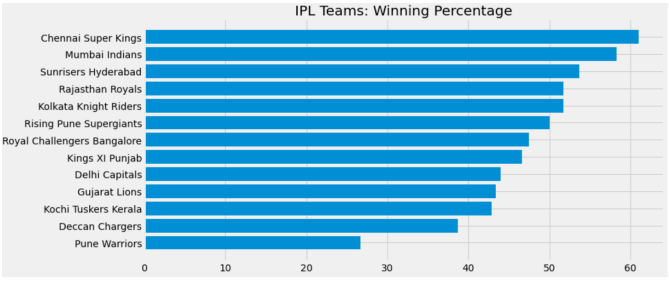

Fig. 3. Plot showing winning percentage of teams

Chennai Super Kings have won 60.97 percentage of their matches,which is the highest and Pune warriors india have the lowest with 26.67 percentage

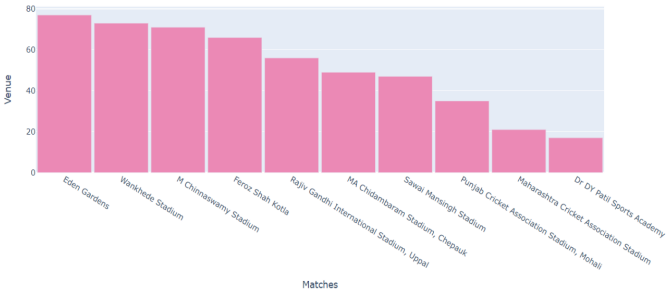

Fig. 4. Plot showing number of matches played at different venues

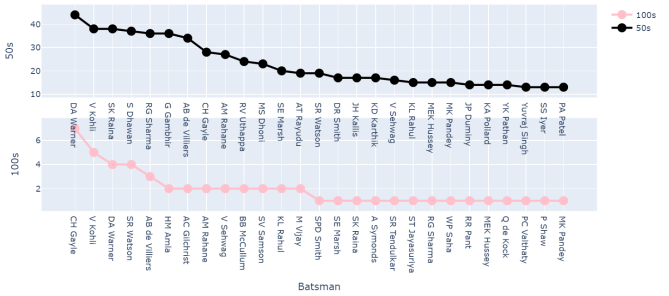

Fig. 5. Plot showing centuries and half centuries by top batsmen

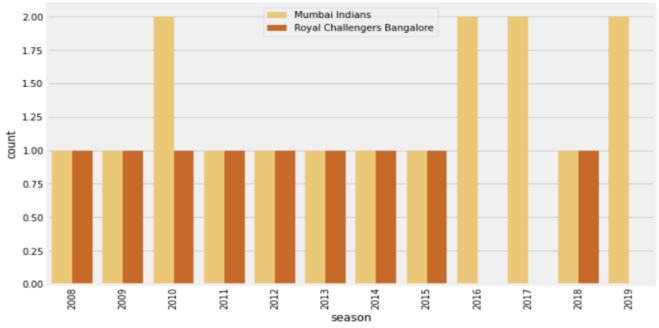

Fig. 6. Bi-histogram of encounter between Mumbai Indians and RCB

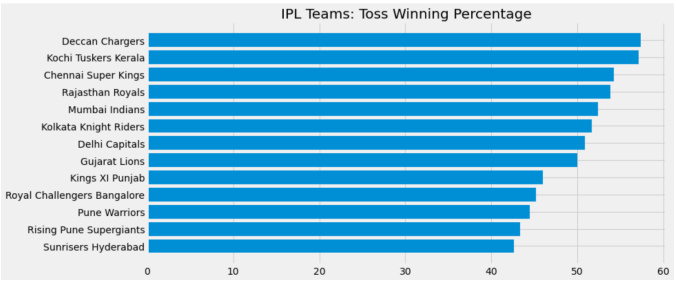

Fig. 7. Toss winning percentage of each team

As seen in the above graph,the luckiest team in IPL was Deccan Chargers with 57.33 toss winning percentage.The top 2 teams in this list are in bottom 3 in the match winning percentage list. So the data also suggest that winning toss is not a match defining factor.

VII. Feature Engineering

Used label encoder to convert the data into numerical form. The exact features which are used for prediction are listed as follows:
a)Venue:
b)batsman:
c)Cumulative runs
d)Cumulative wickets
e)bowler
f)over
g)ball
h)batsman runs
Engineered the features Cumulative runs and Cumulative wickets which gives us the data of current score and current wickets.This is of no use if we did not had the data of the current over and ball.This immensely helps in making a better model.Venue is extremely important to predict the final score as in cricket scores vary extremely between venues.Batsman batting(on strike) and bowler bowling bowling also makes an impact.Batsman runs gives the data of the runs scored by the batsman who is in strike Also created Custom Accuracy for score predicting models which is engineered in such a way that it shows accuracy to be 1 when predicted score is within 10 runs

VIII. Results

A. Models to Predict Scores

We created few regression models to predict the runs scored by team in first innings based on features like Team batting, Team bowling, Venue etc Result and accuracy of those models are stated in table below:

ML Models	Accuracy of Predicted Scores
Gradient Boosting Regresor	64
Support Vector Regressor	53
Random Forest Regressor	65
Extra Tree Regressor	67
Linear Regression	50

B. Models to Predict Winners in Percentage

Engineered some classification models to predict the winner of the match, since IPL match rarely ends with draw or No results, We classify our prediction as Binary classification and results were :

ML Models	Accuracy of Predicted Winner
Gaussian NB	42
Random Forest Classifier	60
Logistic Regression	35
Decision Tree	51

IX. Conclusion and Future work

This paper provides useful insights from IPL data-set about what are the best performing teams and players.The Best performing players of IPL can be listed with the most MoM awards analysis and best performing batsman can be seen by the number of 50s and 100s they made in this time. Sponsors can focus on which stadiums host the IPL matches most to analyse the audience in those areas specifically and make their plans accordingly The prediction of final score at any given moment of match is currently done with the help of Current Run Rate (CRR), while it is one of the useful features, it doesn't take into account what are the remaining overs and scores of the batsmen at crease. The models proposed in the work take these features into account to predict the final score given these features at any point in the game.Future Work can be pre-training the neural network models on an ODI or T20 international data-sets and then fine tuning them for IPL predictions as direct training with data-sets is not possible due to different formats and playing conditions

Acknowledgment

We would like to express my profound gratitude to Prof. Amit Sethi, Prof. Manjesh K Hanawal and Prof. Sunita Sarawagi for their contributions to the completion of my project.
I would like to express my special thanks to all the Teaching Assistants of the course for their time and efforts they provided throughout the semester. Their useful advice and suggestions were really helpful to us during the project's completion. We are eternally grateful to you.
We would like to acknowledge that this project was completed entirely by us and not by someone else

References

Git-hub repositories
www.youtube.com/codebasics
www.kaggle.com
https://scikit-learn.org/stable
www.geeksforgeks.org