CS419 - Introducing to Machine Learning
Project Report 2022

IPL Prediction using ML

GROUP 17

200020073	Prashul Vaishnav
200110042	Goransh Gattani
200110048	Ishita Tyagi
200110063	Laxita Karnawat
200020102	Pulkit Jindal

GitHub link : https://github.com/Astroid66/IPL-Machine-Learning-Project
Instructor : Prof. Abir De

1. INTRODUCTION

Sports analytics is a field that is becoming widely popular due to the competitive edge that it can give both to sports teams as well as stakeholders involved in the sport. Various data which is available such as the players and team statistics, environment conditions, etc is made use of to predictive models which can help stakeholders make informed decisions on the game. The main objective is to improve the performance of the team and assist in creating strategies which would help the team perfectly counter its opponents. This can be done both prior to a game as well as dynamically as the game progresses. In recent times, it has been observed that the audience themselves are also interested in the data analysis that goes on in the game and hence, sports analysts try to present this data to the audience by making simplifications to it and making use of pictorial elements such as graphs and charts to capture their attention.

About Cricket

Cricket is a sport that is played by two teams, each having eleven members. A team consists of batsmen, bowlers, and all rounders. The role of the batsmen is to score as many runs as possible in the limited time/overs available, while the bowlers try to restrict the score that the batsmen try to make. All rounders are players that play both roles and have sufficient expertise in both batting and bowling. The performance of a team depends on various factors such as the constitution of the team in terms of types of players, the venue in which the match is being held, the environmental conditions, and the type of opponents that they're playing against. Data analytics can be made use of to help the teams management 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 the state at which the match is in, to provide thema competitive edge against their opponent. These days, data science techniques are being made use of by every team that competes in the sport professionally. When used correctly, it can help teams bridge the gap in skill by formulating an effective strategy to counter their opponents.

About the Indian Premier League

The Indian Premier League (IPL) is the worlds biggest domestic cricket tournament. It is a 20-over format of the game that makes for short, fast-paced games which is one of the reasons for its massive fanbase. It is an annual tournament and has seen 13 such tournaments conducted so far. There are 8 teams involved in the tournament and the teams themselves consist of players from all around the world. The tournament generates a large revenue and has many stakeholders heavily invested in it. So teams will do everything they can to get an edge over their opponents in a game. Data Analysis is now heavily used by all teams to try and gain this edge.

Dataset

The datasets used for analysis and prediction were collected from www.kaggle.com , where the data of all editions of the IPL so far was available. Two datasets have been used. One for overall matches data and one for ball-to-ball data for the full 2008-2020 period. Both the datasets are linked by the 'id' column which represents the matches uniquely. Some of the useful features present in the dataset 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. There are some minor discrepancies in data such as missing values in 'bowling team' column and duplicate team name but it doesn't hurt the predictions task as team data is also present in 'team1','team2' columns. The dataset consists of 2 lakh data points with 21 features in total.

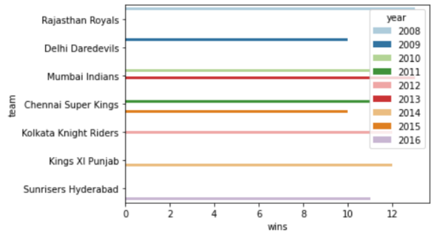
2. ANALYSIS PIPELINE

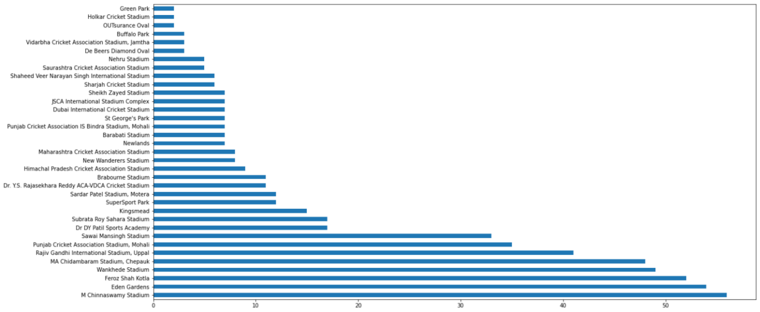
As observed from the literature survey conducted, a large majority of the predictive models that were made are used to predict the outcome of the match and this prediction is made before the start of the match. This prediction will be useful for the team to make long-term decisions for the team to perform better in the tournament as a whole but is not very useful during the match itself as no changes can be made to the team in the middle of a match. The work discussed in this paper seeks to fill in this gap by providing data to the team at various phases of the match so that the team can make informed decisions such as what batting order and bowling order to use for the rest of the game. Firstly, an exploratory analysis of the data is conducted to get a better understanding of what parameters affect the performance of the team as a whole as well as the individual contributions of the players.

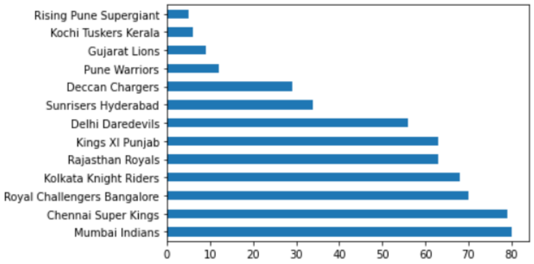
Interesting Insights drawn from the datasets

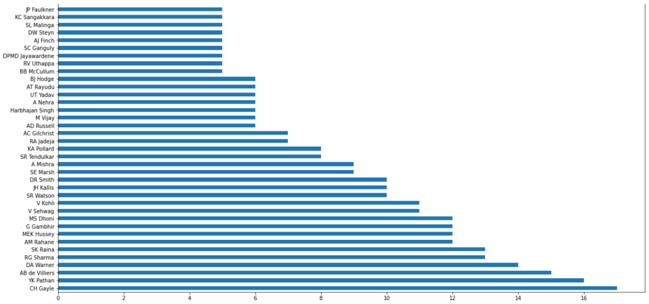
Various manipulations are performed on the available datasets to extract some insightful information from them.

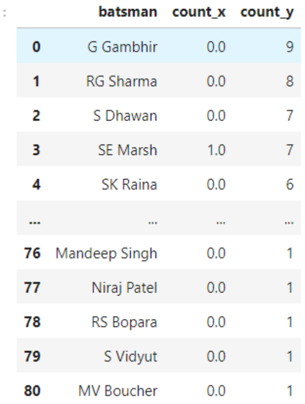
1) Maximum number of wins by any team in particular seasons: Fig. 1,shows the team with maximum number of wins in each season.
2) Stadium hosted the most number of IPL matches: In Fig.2 we can clearly see that M Chinnswamy Stadium has hosted the most no of IPL matches.
3) Team winning the most number of matches: In Fig.3 we can clearly see that Mumbai Indians has won most number of matches followed by Chennai Super Kings.
4) Player winning the most number of Man of the Match awards: In Fig.4 we can clearly see that CH Gayle is the most influential player followed by Y K Pathan.
5) Team wining the most number of tosses: In Fig.5 we can see that Mumbai Indians has won most number of tosses followed by Kolkata Knight Riders.
6) Most 50s and 100s scored by batsmen: In Fig.6 we can see that most no of 100s is scored by G. Gambhir while most no of 50s is scored by SE Marsh.



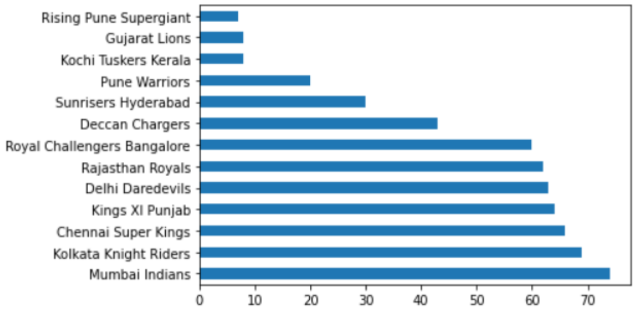








	batsman	count_x	count_y
0	G Gambhir	0.0	9
1	RG Sharma	0.0	8
2	S Dhawan	0.0	7
3	SE Marsh	1.0	7
4	SK Raina	0.0	6
...	...	...	...
76	Mandeep Singh	0.0	1
77	Niraj Patel	0.0	1
78	RS Bopara	0.0	1
79	S Vidyut	0.0	1
80	MV Boucher	0.0	1

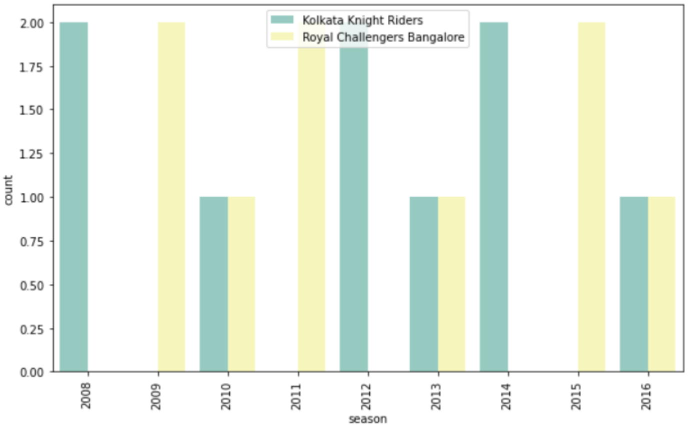


Classification Task

In this section we will see the prediction of winner of the match. We are using matches data set for predicting the winner of match.

The features which are used for predicting are:

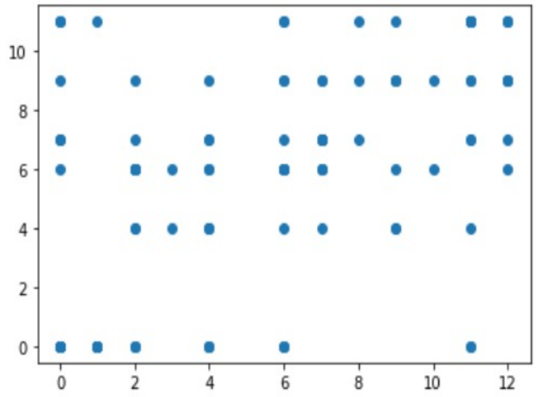
- team1
- team2
- team1 toss win
- team1 win
- venue



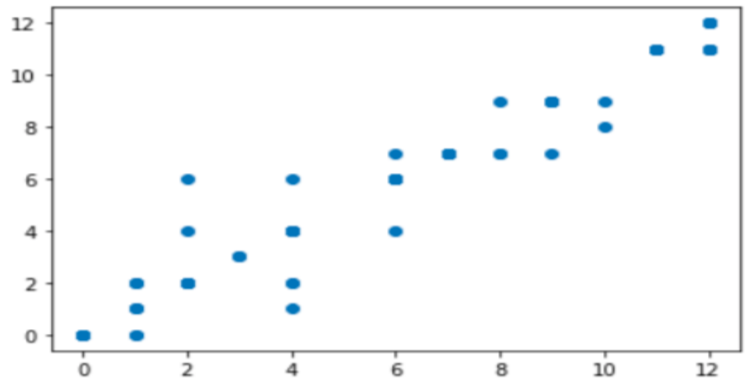
The columns team1 toss win and team1 win are introduced in the data set because they are very useful in predicting the winner of the match.

1) Prediction:

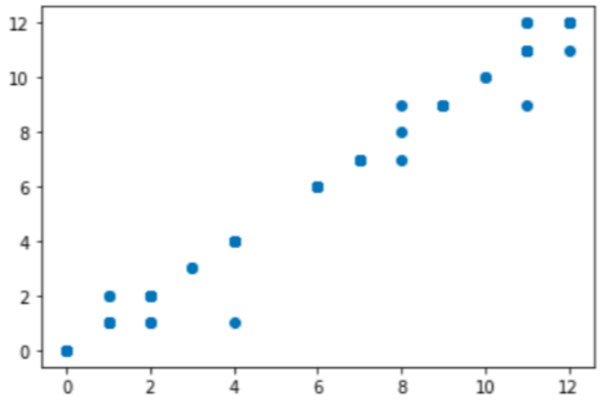
- **SVM:** A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text. Result: We get accuracy of 31.3043% on test set when we use Support Vector Classifier(SVC).



- **Random Forest Classifier:** A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree. Result: We get accuracy of 80.00% on test set when we use Random Forest Classifier.



- **Decision Tree Classifier:** Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. Result: We get accuracy of 88.6957% on test set when we use Random Forest Classifier.



B. Regression Task

In this section we will see the prediction of final score of the match.Here we have used both matches and ball by ball dataset for predicting the final score of match.
The features which are used for predicting are:

- over cur
- batsman runs
- total runs y
- CUMSUM runs
- CUMSUM wickets
- batting team
- bowling team
- venue

- is wicket

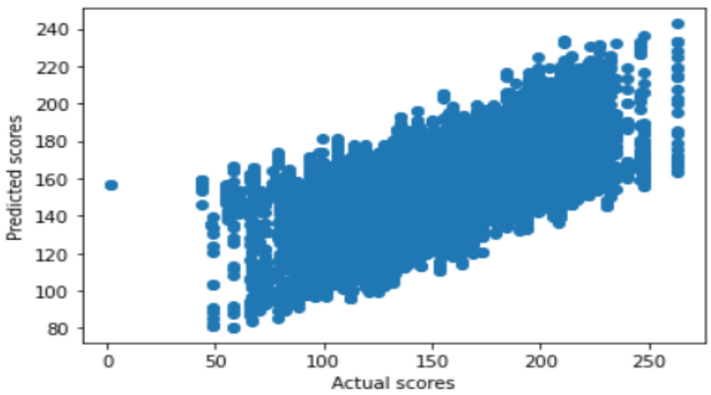
over cur has been defined by merging over and current ball for one useful feature. CUMSUM runs, CUMSUM wickets, is wicket are the new columns introduced in the dataset useful for prediction

.

1) Prediction:

- **Gradient Boosting Regressor:** 'Boosting' refers to one by one adding weak learner sub-models or more specifically decision trees (weak learner denoting a learner performing slightly better than chance). Gradient descent is applied after calculating loss and the next tree is added such as to take maximum descent towards optimum value (the gradient direction)

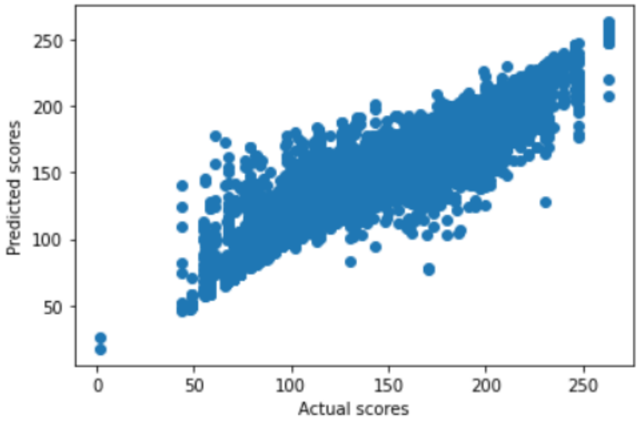
Results:
Accuracy on testing data: 41.94 %
Mean Absolute Error: 16.61
Root Mean Squared Error: 22.40
Custom Accuracy: 41.84%



Random Forest Regressor:

The tree growing in Random Forests happens in parallel which is a key difference between AdaBoost and Random Forests. Random Forests achieve a reduction in overfitting by combining many weak learners that underfit because they only utilize a subset of all training samples.

Results:
Accuracy on testing data: 85.60%
Mean Absolute Error: 7.05
Root Mean Squared Error: 11.15
Custom Accuracy: 77.68%



VI. DISCUSSION

For the Classification task we get maximum accuracy using Decision Tree Classifier of 88.69% followed by Random Forest Classifier of 80.00% for predicting the winner of match. For the Regression task we get maximum accuracy using Random Forest Regressor of 85.6% for predicting the final score of the match.