First Innings IPL Score Prediction

**Abstract**

Cricket is one of the famous outdoor sports that contain a large set of statistical data in real world. As IPL games rise in popularity, it is necessary to examine the possible predictors that affect the outcome of the matches. In this paper, the several years’ data of IPL containing the players details, match venue details, teams, ball to ball details, is taken and analyzed to draw various conclusions which help in the improvement of a player’s performance. The major aim of this project is to predict the First Inning score of an Indian Premier League (IPL) Match using machine learning algorithms. The algorithms used in this project include- Decision Tree Regressor, Linear Regression, Random Forest Regression, Lasso Regression, Support Vector Machine Regression, Neural Network Regression.

**Introduction**

Data mining tools predict the future trends and behaviors, which gives an opportunity to predict the outcome of an IPL (Indian Premier League) match using data mining algorithms. Data mining algorithms have been applied to the IPL dataset and the knowledge from each algorithm has been obtained and analyzed thoroughly as the results are obtained with good accuracy performance. Cricket is one of the most popular sports. The International Cricket Council (ICC) out listed 106 cricket playing nations representing 10 belongs to the full members, 37 of them are

associates, and the remaining 59 are considered to be affiliate members. The game of cricket is played in various formats, i.e., One Day International, T20 and Test Matches. The Indian Premier League (IPL) is a Twenty-20 cricket tournament league established with the objective of promoting cricket in India and thereby nurturing young and talented players. The teams for IPL are selected by means of an auction. Players’ auctions are not a new phenomenon in the sports world. However, in India, selection of a team from a pool of available players by means of auctioning of players was done in Indian Premier League (IPL) for the first time. This in turn, is dependent on the complex rules governing the game, luck of the team (Toss), the ability of players and their performances on a given day. A way of predicting the outcome of the matches between various teams can aid in the team selection process.

In this project, a Machine Learning model is presented that adapts a Regression approach to predict the score of the First Inning of an IPL Match using Decision Tree Regressor, Linear Regression, Random Forest Regression, Lasso Regression, Support Vector Machine Regression, Neural Network Regression and finally predict the scores of IPL using the best model possible.

**Problem Survey**

Since the dawn of the IPL in 2008, it has attracted viewers all around the globe. High level of uncertainty and last moment nail biters has urged fans to watch the matches. Within a short period, IPL has become the highest revenue generating league of cricket. Data Analytics has been a part of sports entertainment for a long time. In a cricket match, we might have seen the score line showing the probability of the team winning based on the current match situation.

In Machine Learning, the problems are categorized into 2 groups mainly: Regression Problem and Classification problem. The Regression problem deals with the kind of problems having continuous values as output while in the Classification problem the outputs are categorical values. Since the output of winner prediction is a categorical value, the problem which we are trying to solve is a Classification problem. We present a model to-

* Understand the dataset.
* Clean the data.
* Analyze the candidate columns to be Features.
* Process the features as required by the model/algorithm.
* Train the model/algorithm on training data.
* Test the model/algorithm on testing data.
* Tune the model/algorithm for higher accuracy.

The basic idea in developing this project is to successfully predict the first innings scores of IPL using different machine learning models.

**Dataset Description and Pre-processing**

The Dataset contains ball by ball information of the matches played between IPL Teams of Season 1 to 10, i.e. from 2008 to 2017. This section of the project includes the dataset chosen to build the model and the pre-processing steps used to clean and analyse the data. The dataset contains 76014

Entries with 15 columns. This Machine Learning model adapts a Regression approach to predict the score of the First Inning of an IPL Match.

**Description of dataset-**

* mid: Unique match id.
* date: Date on which the match was played.
* venue: Stadium where match was played.
* Bat-team: Batting team name.
* Boll-team: Bowling team name.
* batsman: Batsman who faced that particular ball.
* bowler: Bowler who bowled that particular ball.
* runs: Runs scored by team till that point of instance.
* wickets: Number of Wickets fallen of the team till that point of instance.
* overs: Number of Overs bowled till that point of instance.
* runslast5: Runs scored in previous 5 overs.
* wicketslast5: Number of Wickets that fell in previous 5 overs.
* striker: max(runs scored by striker, runs scored by non-striker).
* non-striker: min(runs scored by striker, runs scored by non-striker).
* total: Total runs scored by batting team at the end of first innings.

**Data Pre-processing**

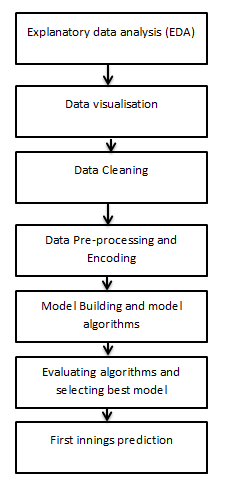
Data pre-processing is a data mining technique which is used to transform the raw data in a useful and efficient format. From the dataset, we have decided to process the data by-

* Handling categorical features
* Performing Label Encoding
* Performing One Hot Encoding and Column Transformation

Data cleaning- The data can have many irrelevant and missing parts. To handle this part, data cleaning is done

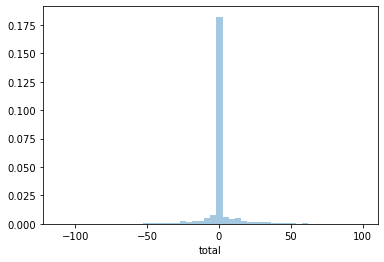
* Removing unwanted columns
* Keeping only consistent teams
* Removing the first 5 overs data set in every match
* Removing outliers

**Implementation/Flow of the project**

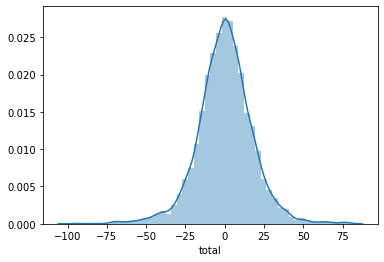


**Results and Graphs**

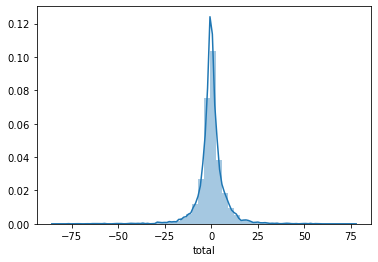
The results presented in this section include the accuracy graphs of all the machine learning models used in this project.

****

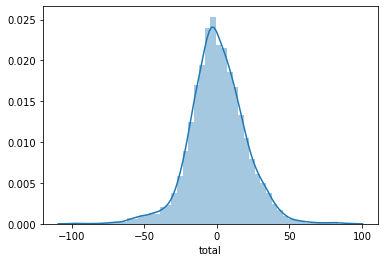
**Fig 1: Decision tree output**

****

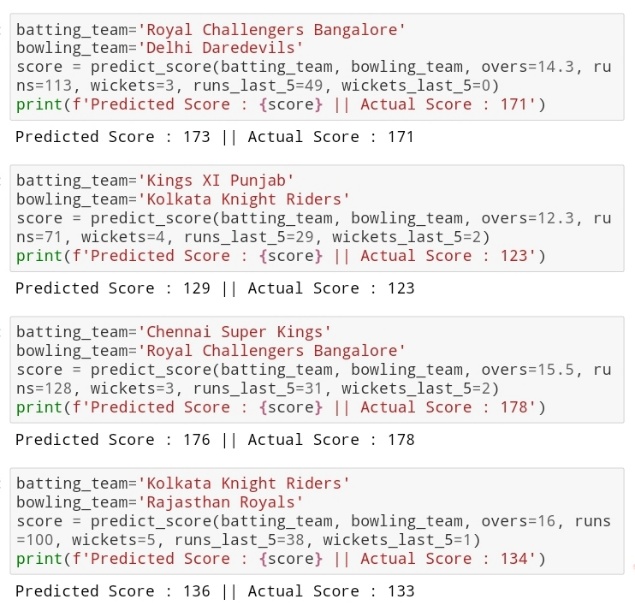
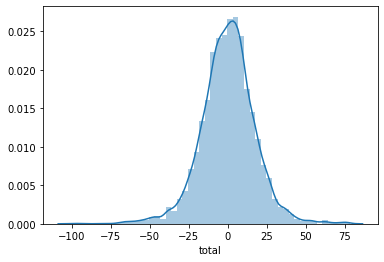
**Fig 2: Losso regressor output**

****

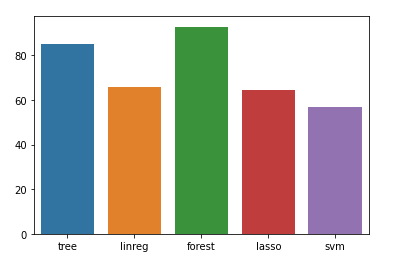
**Fig 3: Random Forest output**

****

**Fig 4: SVM output**

****

**Fig 5: Linear Regression output**



**Fig 6: Comparison of models**

|  |  |
| --- | --- |
| Algorithm | Accuracy |
| Linear Regression | 65.65% |
| Lasso Regressor | 64.53% |
| SVM | 56.86% |
| Decision Trees | 85.09% |
| Random Forest | 93.43% |

**Table 1: Accuracies of different ML models**

Random Forest performed the best, closely followed by Decision Tree. So we will be choosing Random Forest for the final model prediction.

First innings IPL-2021 final score prediction-

**Fig 7: Final predicted score**

**Conclusions**

From the results presented in the above section, the final machine learning model that we chose to predict the IPL first innings score is Random Forest algorithm because, its accuracy has as been high as 93.43%. The predicted IPL first innings of 2021 score in the above section prove that the model built is successful in meeting the project requirements.

**Group 10**

**AP18110010142 P Retheshna - CSE C**

**AP18110010144 R Nikitha - CSE C**

**AP18110010456 M VijayaDurga-CSE G**

**AP18110010220 T Saahas Reddy-CSE D**

**AP18110010660 K Sailesh - CSE J**