Predicting the winner of an IPL Match

Sai Pranav M
CSE
PES University
saipranavmadineni@gmail.com

Rohan Bennur
CSE
PES University
rohan.bennur@gmail.com

***Abstract*—All sports contain a vast amount of useful statistical information with respect to teams, players, games and editions. Cricket being one of the most popular sports in the world, every team has a desire to win a match. Indian Premier League (IPL) is the most watched cricket competitions in the world. A huge amount of data is collected and mining over this data can lead to some hidden patterns in understanding the likelihood of a team winning. So, a model can be developed which predicts the outcome of the match. The best data mining and Machine Learning techniques will be utilized in the making of the model. The performances of data mining techniques such as Deep Learning using Keras, Decision Trees, Naïve Bayes, Support Vector Machines, K Nearest Neighbors, AdaBoost, Gradient Boosting and Random Forest will be implemented using different metrics at different stages and accuracy of the models can be studied. These can be extensively useful in predicting the winner of an IPL Match**

Keywords: Cricket, IPL, Machine Learning, , Keras, Deep Learning, Decision Trees, KNN, SVM, Naïve Bayes, AdaBoost, Gradient Boosting, Random Forest

I. INTRODUCTION

Cricket is one of the most loved sports in the world. It is played with a bat and a ball that is played between two teams with 11 players in a team. The batsmen scores runs by hitting the ball that is bowled at the wicket from a 22-yard distance. The bowlers and the fielders try to prevent these runs by dismissing the batsmen by either getting him bowled (where the ball hits the wickets behind the batsmen), getting caught(where the ball hit by the batsmen is caught by the fielders before it pitches on the ground, or by getting run-out where a fielder takes the bails off the wickets before the batsmen completes his run.

Cricket is played in a variety of forms ranging from T20, with every team batting for twenty overs to ODI, where each team bats for fifty overs, to a Test Match, where two teams play for five days with two inning each until 10 of the players get bowled out. T20 is becoming the most popular format among the audience and IPL is at the pinnacle of entertainment. IPL features players from across the world with up to four foreigners in each team.

In this project, a match is predicted before the start of the game. The winning prediction would be done utilising various algorithms of Machine Learning.

. Use of machine learning algorithms to predict the result before the match has been played will allow the management team to analyse the improvement areas. Machine learning is trending and is full of interesting computational insights, which uses technology to make predictions.

II. CRICKET STATISTICS

Cricket Statistics can be divided in two categories: Batting and Bowling.

A. Batting Statistics

1. **Innings:** part of a game during which each team has an opportunity to score runs
2. **Runs:** total number of runs scored by a player in his career
3. **Average:** average of the runs scored by the batsman
4. **Strike Rate:** number of runs scored per ball * 100
5. **Highest Score:** maximum number of runs scored by a batsman in a single innings
6. **Centuries:** number of times a batsman has scored more than hundred runs in an innings
7. **Balls Faced:** number of balls faced by a batsman.
8. **Run Rate:** Total number of runs scored by a team/ number of balls played

B. Bowling Statistics

1. **Maiden:** the total number of overs in which a bowler hasn't conceded a single run
2. **Overs:** A single over is completed when a bowler delivers six balls.
3. **No-Balls:** total number of times a bowler oversteps the crease or the he bowls a ball which goes above the waist line of the batsman without pitching on the ground.
4. **Average:** average number of balls bowled before picking a wicket
5. **Economy:** number of runs given by a bowler in his bowling career/ total number of overs bowled
6. **Five Wicket hauls:** number of times a bowler has taken 5 or more wickets in an innings
7. **Wickets:** number of times a bowler has dismissed a batsman
8. **Wide:** number of times a bowler has bowled a wide

Apart from the above-mentioned statistics, there are few statistics which influence the winner of a match. They are- Winner of the toss, Format of the match, Weather forecast, Location, Player's age, Condition of the pitch, Wind Speed and Direction, right-handed or left-handed, Speed of Bowling, Crowd Support, team facing, match importance (e.g. The final of a tournament), Time of playing of the match also play a major role in determining the winner of a match.

III. PREVIOUS WORK REVIEW AND ASSUMPTION

A. *Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach [1]*

In this paper, the researchers have used three features i.e. Toss, Venue, Strength A/B (relative strength of team A wrt to team B) to detect a match winner.

They use 4 models here for prediction of a match winner. The models are trained to predict the best 6 performing batsman and bowlers in a match.

The dataset is distributed in such a way that the training dataset is a set of all the matches played between the year 2010 and 2014 and the testing dataset is a set of all the matches played after the year 2014.

Support Vector Machines, Logistic Regression, Random Forests Classifier, Decision Trees with stumps and k nearest neighbours have been considered for predicting the best 6 players in batting and bowling.

They claim that the k nearest neighbours algorithm gives best results among the classifiers mentioned.

Takeaway-The testing results show that we cannot completely depend on the historical data as it doesn't show the current capability of a team.

B. *Prediction of Live Cricket Score and Winning [2]*

In this paper the model is trained to predict the number of runs that will be scored by a team in the first innings based on current run-rate and number of wickets lost by the batting team. This methods have been implemented using Q-Learning base decision tree or Linear Regression Classifier approach, Linear Regression and Naïve Bayes Classifier. For the above approach, a five over interval has been created and at each interval the attributes have been updated. The dataset consists of all the matches played in the years ranging between 2002 and 2014 of every team.

Takeaways: The dataset used is dynamic as it is updated after every 5 over interval.

Using prediction like Naive Bayesian algorithm, they've also predicted the best 11 players from each team.

C. *The Use of Data Mining for Basketball Matches Outcomes Prediction [3]*

This research has been chosen since it uses binary classification algorithms that is required for sport result prediction.

The problem intended to solve is the outcome of the match..

For each match, two groups of attributes have been used . One group consists of a few statistics for basketball which are 141 in number and the second group of attributes consists of information about league standings. New information on the matches that were conducted on the day before, along with a new set of data for all teams were obtained and using this model the prediction for future games was calculated.

For the development of regression and classification, RapidMiner is selected as a tool.

The Naïve Bayes algorithm and multivariate linear regression is used in predicting the outcome and spread for a team respectively.

Takeaway-In the above paper 141 attributes in the first group of attributes which led to overfitting of data.

IV. PROBLEM STATEMENT

Cricket is one of the famous outdoor sports that contain a large set of statistical information in real world. With millions of fans following the IPL matches , the prediction of outcome of a match is gaining huge importance. With the emerging popularity of IPL, numerous online platforms have come up where the user is made to pick the winning team. The predicted data can then be used in these online platforms to gain points.

V. PROPOSED SOLUTION AND METHODOLOGY INVOLVED

All the variables of the dataset were not considered for building the models because they tend to result in a misleading outcome because they either tend to underfit or overfit the data.

The variables in the dataset are defined as follows:

The variables in file (matches.csv):

- Id: Unique id for every match
- Season: The year of the IPL
- City: the city where the match takes place
- Date: date on which the match takes place
- Team1: Team name of team 1
- Team 2: Team name of team 2
- Toss winner: Name of the team that won the toss in that match
- Toss decision: The decision taken by the team captain either to bat or field
- Result: normal or a tie
- dl_applied: Whether DL was applied or not
- Winner: The match winner
- Win_by_runs: number of runs the winning team has won by if it batted first
- Win_by_wickets: The number of runs the winning team has won by if it bowled first.

The matches.csv file was used to build the model to predict the result of a match as is it consisted of all the necessary variables for predicting the winner of a match.

For training the model, the data has been split into train and test data in the ratio 80:20 and the models were trained.

The variables 'season', 'team1', 'team2', 'city', 'toss_decision' and 'toss_winner' were chosen for building the models.

A. Data Pre-Processing

Data cleaning and Pre-processing embark significant importance in data mining. Data cleaning is defined as the process of identifying inaccurate data and then replacing it with appropriate values or by zeros, by using various data wrangling tools.

The dataset [4] consists of two files, matches.csv and deliveries.csv.

- There were few missing values in the 'city' and 'winner' columns of the dataset
- The missing values in 'city' were replaced with a neutral city "Dubai" which isn't the home team for either of the participating teams.
- The missing values in 'winner' were replaced with a 'draw'. This was probably due to the match not taking place because of rain or other obstructions.
- There were few outliers, which were ignored

B. Building the Models

The models built were:

1) ***Deep Learning Model with keras*:** Keras is one of the most powerful API of TensorFlow. The models are developed by connecting graphs stacking layers. It is a form of semi-supervised learning that is a combination of both unsupervised and supervised learning. Unsupervised learning is used to draw results from datasets that are not labelled. Supervised learning a type of machine learning with a task of learning a function that maps input data to target labelled data. Keras is extensively used in the construction of neural networks. For multiclass classification, Multi-Layer perceptron can be constructed.
.

2) ***Decision Trees:*** Decision tree is highly useful for prediction and classification. It is a flowchart like structure in the form of a tree, where each node indicates a computation on an attribute. The branch represents a result of the test, and each terminal node contains a label for the class.
They classify instances by arranging the attributes down a tree starting at the top of the tree, that is root to a terminal node, which establishes a classification for an instance.

3) ***Naïve Bayes:*** This technique is based on the Bayes Theorem with an assumption that there is independence among variables. It assumes the presence of a target variable which is unrelated to any other variable. Naive Bayes is formulated as:

P(C) = Probability of the first class;
P(C/D) = [P(D/C) X P(C)]/P(D)
Here, P(D) = The probability of predictor
P(D/C) = The probability of a predictor variable when the class has been given;
P(C/D) = The probability of class given predictor

4) ***Support* Vector Machine:** An SVM is a classifier which consists of a hyperplane for separation. When the training data that has been labeled has been given, this SVM gives a hyperplane as output which divides the new data into categories.

5) ***K Nearest Neighbours:*** A k nearest neighbour is an algorithm which classifies the data points by finding the distance between the data point and a centroid of a group.

6) ***AdaBoost Classifier:*** This is a classifier which starts its process by trying to fit a classifier on the initial data and then tries to fit the other copies of the classifier on the same data. Although, here the weights of the instances which were classified incorrectly have been adjusted so that the next classifiers focus more on much tougher cases.

7) ***Gradient Boosting Classifier:*** They are a set of ML algorithms that merge a lot of learning models which are very weak, together, to construct a model for prediction that is very strong. Gradient boosting classification is usually done using decision trees. They classify highly complex datasets considerably better than other models because this combines the strength of a lot of weak learners.

8) ***Random Forest Classifier:*** Random forests is a model based on ensemble learning that is used for regression, classification and wide variety of jobs that function by building a large number of of decision trees at the time of training the data. They output the class that appears the greatest number of times in the classification method and the average of the classes in regression method. They solve the problem of overfitting in decision tress.

These Models were chosen to test which best predicts the outcome of an IPL match by computing their accuracy on the training and testing dataset.

VI. EXPERIMENTAL RESULTS AND OTHER OBSERVATIONS

A. Model Accuracy

All the above models were trained using winner as the target variable and 'season', 'team1', 'team2', 'city', 'toss_decision' and 'toss_winner' were chosen as predictor variables. The accuracy on the testing data and training data was computed for all the models to choose the best model.

1. Deep Learning Model with Keras

Training Accuracy: 72.24%

Testing Accuracy: 45.31%

2. ***Decision Trees***

   Training Accuracy: 99.606%

   Testing Accuracy: 44.531%

3. ***Naïve Bayes***

   Training Accuracy: 36.220%

   Testing Accuracy: 32.812%

4. ***Support Vector Machine***

   Training Accuracy: 15.157%

   Testing Accuracy: 11.719%

5. ***K Nearest Neighbors***

   Training Accuracy: 49.803%

   Testing Accuracy: 35.938%

6. ***AdaBoost Classifier:***

   Training Accuracy: 21.063%

   Testing Accuracy: 22.656%

7. ***Gradient Boosting Classifier:***

   Training Accuracy: 99.606%

   Testing Accuracy: 57.812%

8. ***Random Forest Classifier:***

   Training Accuracy: 99.606%

   Testing Accuracy: 55.469%

It was observed that Decision Trees, Gradient Boosting Classifier and Random Forest Classifiers worked really well with the training data with a training accuracy of 99.6065%.

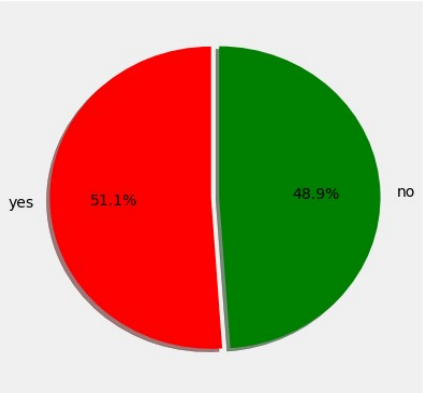
But when the model was tested on the testing data, Gradient Boosting worked the best with a testing accuracy of 57.812% followed by Random Forest Classifier with 55.469% and then Decision Trees with a testing accuracy of 44.531%. So, the Gradient Boosting Classifier is the best model for predicting the winner of an IPL Match.

B. Other Observations:

The importance of the predictor variables used for predicting the winner was computed and the results are as follows:

```
team2           0.360726
team1           0.248415
city            0.122728
venue           0.104026
toss_winner     0.089894
season          0.068179
toss_decision   0.006032
```

It can be seen that the toss winner is hardly correlated with the match winner. This means that the toss winner is not necessarily the match winner.



VII. CONCLUSION

The goal was to predict the team which would win a given Indian Premier League match.

After considering various models, and computing their accuracies, Gradient Boosting Classifier proved to be the best model in predicting the winner. It was initially assumed that the winner of the toss and decision taken after winning the toss plays a major role in predicting the winner but after building the model it was found that they hardly contribute to the winner of a match. Since a cricket match, especially an IPL match of the T20 format is highly unpredictable, an accuracy of 57.812% from the Gradient Boosting is more than a decent value.

VIII. FUTURE WORKP

With increaing popularity of IPL, predicting the best players for intelligent team selection which will not only benefit the IPL team but also the audience who can benefit from using this analysis to build teams on online fantasy premier league platforms like Dream 11 and MPL.

IX. REFERENCES

[1] Madan Gopal Jhanwar, and Vikram Pudi, "Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach" *MLSA@PKDD/ECML,* 2016

[2] Rameshwari A. Lokhande and Pramila M. Chawan, "Prediction of Live Cricket Score and Winning", International Journal of Trend in Research and Development, Volume 5(4), July 2018

[3] Dragan Miljković, Ljubiša Gajić, Aleksandar Kovačević, Zora Konjović, "The Use of Data Mining for Basketball Matches Outcomes Prediction", Intelligent Systems and Informatics (SISY), 2010 8th International Symposium, September 2010

[4] Dataset: kaggle.com/manasarg/ipl