Predicting the winner of an IPL Match

Sai Pranav M
Computer Science and Engineering
PES University
Bangalore, India
saipranavmadineni@gmail.com

Rohan Bennur
Computer Science and Engineering
PES University
Bangalore, India
rohan.bennur@gmail.com

***Abstract*—All real-life sports contain a vast amount of statis-tical information regarding individual players, team, games and seasons. Cricket is one of the most popular sports played in the world. Each team has a desire for winning and choosing the best playing eleven has been the most challenging task. Indian Premier League (IPL) is being played for the past 10 years. Abundant data is being gathered and mining over this data can lead to some hidden patterns in understanding the likelihood of a team winning and in also selecting the best playing eleven of a team. In addition, years of experience, current form, the impact of the players, batsmen available and the number of runs to be chased can have an influence on the outcome; hence these factors are likewise expected and compared. Hence, a model can be developed which can predict the outcome of the match at various stages of the game State-of-the-art 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 an outdoor sports game played between two teams and the team that scores maximum runs wins the game. This game is played at domestic and international levels and is played across three different formats: one day international(50-over match), Test match (5-day match) and T20(20-20) format respectively. Cricket is followed and loved by more than a billion people all around the world. Hence there is a need fora data mining model to make prediction to predict the winner of the fixture.

The recent T20 format has gained huge recognition through-out the globe. India winning the inaugural T20 World Cup in 2007, lead to a massive foundation for the introductory edition of Indian Premier League in 2008. This league was followed by the immense population due to the fast-paced fixtures. Out of all domestic leagues around the world, IPL is leading in terms of money, entertainment, footfall, popularity, number of views, etc. After the enormous success of IPL, various other countries also started a similar type of franchise league. In IPL, players from various nationalities feature in a different franchise. Selection of the players depends upon the auction which is carried prior to each season. The franchise with the highest bidder for a particular player gets the ownership to feature that player. There is no upper limit for bid price but the overall budget is limited.

In this project, prediction of a match is considered before the start of the game depending upon the impact-factor of each player, playing in that respective team. The winning prediction would be done utilising various algorithms of Machine Learning.

Use of Artificial Intelligence, machine learning, deep learning and data science makes life easier in every aspect. Use of machine learning and predicting the outcomes before the match actually played will allow the players as well as the coaches to analyse the improvement areas. Machine learning is booming and firmly identified with (and frequently covers with) computational insights, which also focuses on prediction-making through the use of technology. It has solid connections to numerical improvement; which hypothesis conveys strategies and application areas to the field. Machine learning is some of the time conflated with data mining where the latter subfield concentrates more on exploratory information analysis and is known as supervised learning.

II. CRICKET STATISTICS

Statistics mainly fall in two categories: Batting and Bowling. These two statistics have most influence on final outcome of match.

A. *Batting Statistics*

All Statistics related to batting are:

1. Innings: The number of innings player played.

2. Not Outs: The number of times the batsman remains not out.

3. Runs: The number of runs scored by batsman.

4. Highest Score: The highest score ever made by the batsman in the past.

5. Batting Average: The batting average of batsman.

6. Centuries: The number of times batsman score 100 runs.

7. Half-Centuries: The number of times batsman score 100 runs.

8. Balls Faced: The total number of balls faced by batsman.

9. Strike Rate: Strike Rate = (100 * Runs) / Ball Faced

10. Run Rate: The average number of runs scores by team per 6 balls.

B. *Bowling Statistics*

Statistics related to bowlers are:

1. Over: Six balls bowled by bowler is an over.

2. Maiden Overs: The number of maiden overs (In which the bowler conceded zero runs in an over) bowled.

3. Wickets: The number of wickets taken by bowler.

4. No-Balls: The number of no-balls bowled by bowler.

5. Wide: The number of wide balls bowled by bowler.

6. Bowling Average: The average number of runs conceded per wicket.

7. Strike Rate: The average number of balls bowled per wicket taken.

8. Economy Rate: The average number of runs conceded per over.

9. Five Wickets in an Innings (5w): The number of innings in which the bowler took five wickets or more.

10. Ten Wickets in a Match (10w): The number of matches in which the bowler took ten wickets.

Apart from batting and bowling there are other statistics as well which affect the final outcome of cricket match. Other important statistics are: Toss Win, Weather, Partnership, Format of the match, Venue, Pitch condition, Age of the player, right-handed or left-handed, Bowling Speed, Play in Pressure, Opponent team, Importance of match (e.g. Final Match is important than other match), One day or Day night match, Injury.

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 final weights of the models are chosen such that top 6 performing batsmen and bowlers match with predicted top 6 players.

Here they've distributed the training dataset to be all the matches played between 2010 and testing dataset to be all the matches played after 2014.

SVM, Random Forests, Logistic Regression, Decision Trees and kNN have been used for binary classification.

They claim that the kNN algorithm yields better results as compared to other classifiers.

Takeaway-This paper conveys relying completely on the historical data is not only insufficient, but also fallacious since it does not portray the current competence of a team.

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

In this paper a model has been proposed that has two methods, first predicts the score of first innings not only on the basis of current run rate but also considers number of wickets fallen ,venue of the match and batting team. The second method predicts the outcome of the matching the second innings considering the same attributes as of the former method along with the target given to the batting team. These two methods have been implemented using Linear Regression Classifier or Q-Learning base decision tree approach and Naïve Bayes Classifier for first innings and second innings respectively. In both methods, 5 over intervals have been made and at each interval above mentioned attributes have been updated. The dataset consists of all the matches played between 2002 and 2014 of every team independently.

Takeaways: The dataset used is dynamic , 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 which can be used on online platforms like fantasy leagues for getting maximum points.

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

This research paper 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 and spread.

Spread represents and advantage in the number of points given to one of the teams in order to equalize their chances for victory.

For each match two groups of attributes have been used . The first group consists of standard basketball statistics which are 141 in number and the second group of attributes consists of information about league standings. Fresh information about the games that were played previous day, as well as new statistical data for all teams were downloaded and using this model calculates the prediction for future games.

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

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

Takeaway-This model used 141 attributes in the first group of attributes which led to increased epochs for the model in reaching 67% accuracy.

IV. PROBLEM STATEMENT

Cricket is one of the famous outdoor sports that contain a large set of statistical data in real world. With millions of fans following the IPL matches , the prediction of outcome of a match and the best player of a match before it takes place is a real word problem. The predicted data can then be used to create teams on platforms like Dream11 or MPL.

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 w 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 1 (deliveries.csv):

- Match_id: Assigns a unique id for every match played in the IPL
- Inning: Tells if the fist set of batting was going on or second. 1: First innings 2: Second Innings
- Batting_team: The team name which is currently batting.
- Bowling_team: The team name which is currently bowling.
- Over: Describe the current over number
- Ball: describe the current bowl no of the current over.
- Batsman: Name of the batsman on striking end.
- Non_striker: Name of the batsman on non-striking end
- Bowler: bowler name:
- Is_super_over: To indicate if it is a super over or a normal over.

The variables in file2 (matches.csv):

- Id: Unique id for every match
- Season: The year of the IPL
- City: the city in which the match is played
- Date: The date on which the match is played
- Team1: Team name of team 1
- Team 2: Team name of team 2
- Toss_winner: The team name that won the toss in that match
- Toss decision: The decision taken by the team winning the toss (bat or field)
- Result: normal or a tie
- dl_applied: Whether DL was applied or not
- Winner: The winner of the match
- Win_by_runs: The 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 outcome 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 to ratio 8:2 and the models were trained.

The variables '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 'city' and 'winner' of matches.csv
- The missing values in city were replaced with a neutral city "Dubai" which isn't the home team for both the 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 a Python-based Deep Learning framework which is the high-level API of TensorFlow. It can be configured to execute over Theano, TensorFlow or CNTK. Being an open-source Deep Learning framework, Keras models are developed by stacking layers and connecting graphs. It can be classified as semi-supervised learning which is an amalgam of both supervised and unsupervised learning. Supervised machine learning is used for la-belled data whereas unsupervised machine learning is used on nonlabelled data, therefore, semi-supervised learning algorithms are trained on a combination of both labelled and unlabeled data. Keras can be used for building neural networks. Multi-layer Perceptron will be built for multiclass classification

.

2) ***Decision Trees:*** Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute. This process is then repeated for the subtree rooted at the new node.

3)Naïve Bayes: This technique is based on the Bayes Theorem. Naive Bayes model is easy to implement and is used for larger datasets. It is used to predict the probability of the dependent variable based upon different attributes and mainly used in the problems which are having multiple classes. Naive Bayes is formulated as:

P(A/B) = [P(B/A) X P(A)]/P(B)

Where P(A/B) = Posterior probability of class given predictor; P(A) = Probability of prior class; P(B/A) = Likelihood of probability of predictor given class; P(B) = Prior probability of predictor

4) ***Support* Vector Machine:** A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two-dimensional space this hyperplane is a line dividing a plane in two parts where each class lay in either side.

5) ***K Nearest Neighbors:*** A k-nearest-neighbor algorithm, often abbreviated k-nn, is an approach to data classification that estimates how likely a data point is to be a member of one group or the other depending on what group the data points nearest to it are in.

6) ***AdaBoost Classifier:*** An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

7)***Gradient Boosting Classifier:*** Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting. Gradient boosting models are becoming popular because of their effectiveness at classifying complex datasets.

8) ***Random Forest Classifier:*** Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

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 '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%.

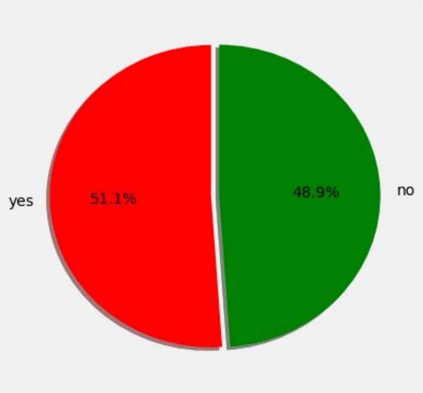
But when the model was tested on the testing data, Gradient Boosting worked the best with a testing accuracy of 57.812% followed by Random Forest Classifier with 55.469% and then Decision Trees with a testing accuracy of 44.531%. So, the Gradient Boosting Classifier is the best model for predicting the 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 Winner of an IPL 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 decision taken after winning the toss plays a good role in predicting the winner but it was later found that it hardly contributes 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 WORK

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.

Evaluating the performances of players is not a straight-forward task. Manually analysis of all the past record of each player is practically impossible. So intelligent system to predict the performance of the players based on their past record can be helpful for team management and team selectors.

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