Cricket Match Prediction and Analysis (T20,IPL)

Maliha Tasnim Aurini(14101051): Computer Science and Engineering Department
BRAC University

Shitab Mushfiq-ul Islam(14101088) :Computer Science and Engineering Department
BRAC University

Fahima Akter(14301015): Computer Science and Engineering Department
BRAC University

***Abstract*—IPL matches dataset is used for analysis and generating a predictive model. The data analytics is performed to compare features that influences the outcomes of matches(win/lose). The features that are considered are city, toss decision ,toss winner and venue. A predictive model is developed using Machine Learning algorithm to get highest possible accuracy. Then feature importance retrieved from predictive model applied reveals the importance of features in determining match outcomes(win/lose).These feature importance is then compared against actual data to check for validity. Model used is Random Forest Classifier which gave 88% accuracy.**

Keywords—IPL India Premier League

I. INTRODUCTION

1.1 Project Goal: The goal of the project is to predict the percentage of winning a match by a specific team in comparison with its opponent by applying Logistic Regression algorithm. The primary aim is to show the predictions and comparison between two teams based on selective features. The reason behind using the algorithm is that it is one of most suitable ones to show the relation between variables which are dependent on each-other such as the features of the match.

1.2 Literature Review: Many researches have been going in this regard to get predictions more accurately with different approaches which are significant enough to play a role in prediction. In [1], the authors applied several algorithms which are Bayesian, Native Bayesian, SVM, Classification and Regression Tree to find out the rising stars of a cricket team meaning who will be doing better for test matches individually
as well as a whole with the team based on few classifiers. They also discussed the problems they faced with each algorithm for different cases for calculation purpose of the performance of the players.

Another prediction analysis has been done by Oliver G. Stevenson and Brendon J. Brewer in [2]. Here, they emphasized on Bayesian algorithm to predict best opening batsman pair in test matches with higher potential who will be able to score high amount of run. For this purpose, they used the dataset of new eland players to predict the batsman order using hierarchical structure.

In[3], The authors worked to enhance the efficiency of the umpiring system in a cricket match. The region of the authors interest is using a Haar-cascade-classifier and later the particular gesture is identified using Logistic Regression algorithm. The process would remove the manual updating of scorecards and reduce the game duration noticeably. In addition, it excludes the pre-requisite of wearing special gloves that include sensors. The efficiency of the algorithm is then cross-checked with the training and test data. Their method proved to be a very simple but efficient algorithm for umpires gesture detection.

II. METHODS

The dataset that has been used to prove the method is the Indian Premier League T20 cricket match data. The main goal is to predict the outcome of a match. The data that has been used is from 2008 to 2016. The variable that are included in our dataset are year, team 1, team2, winner ,toss win, toss loss, venue, umpire, toss decision, man of the match, win by wicket, win by run, result. Among all this variables, the focus we are giving on year team1, team2, toss win toss loss, venue and result for the outcome prediction of the IPL T20 match. The dataset is acquired from Kraggle an open source dataset resource site.[3]

Pandas data frame, one dimensional labeled array capable of holding any data type with axis labels or index, was used here. The Pandas data frame was used to read the data into or algorithm. At first, we read the data through pandas data frame. After that we used the encode technique to update our

dataset in such way all the NaN values declared in the winner column turns into draw instead of Nan. The null values in the city or venue column is replaced with Dubai through encoding technique. The process has been done for null value so that we can get the result more accurately. The other parameters included with the null value are updates randomly. After that, randomly generated unique id was given to each row so that we can easily identify the rows.

After the preprocessing of the data, the algorithm can now identify the teams that has been won the toss more and the teams won have own the match most. Based on it divided the dataset into two parts. Through mathplot function a graph was generated [figure 1].

From the dataset, a predictive model was made. Each model is made up of a number of predictors, which are variables that are likely to influence future results. Once data has been collected for relevant predictors, a statistical model is formulated. The variables influencing the predictive model is the city, venue and toss decision. Generic function is used for making a classification model and accessing performance. K-fold is for cross validation.

Firstly, Linear regression was used based on toss win and match winning. From this we get the result which is toss winner has least chances of winning matches[figure 2].Random forest classifier is used to classify the data depending on team1, team2, venue, toss winner, city, toss decision. If we ignore teams, Venue seems to be one of important factors in determining winners followed by toss winning, city. From the above prediction on features, we notice toss winner has least chances of winning matches but does the current stats shows the same result. Toss winning does not guarantee a match win from analysis of current stats and thus prediction feature gives less weightage to that.

III. Results

Testing the algorithms with our testing dataset. top 2 team analysis based on number of matches won against each other and how venue affects them. Previously we noticed that CSK won 79, RCB won 70 matches. Now let us compare venue against a match between CSK and RCB we find that CSK has won most matches against RCB in MA Chidambaram Stadium, Chepauk, Chennai. RCB has not won any match with CSK in stadiums St George's Park and Wankhede Stadium, but won matches with CSK in Kingsmead, New Wanderers Stadium. It does prove that chances of CSK winning is more in Chepauk stadium when played against RCB. Proves venue is important feature in predictability.

Accuracy: 22.184%
Cross-Validation Score: 21.666%
Accuracy: 89.601%

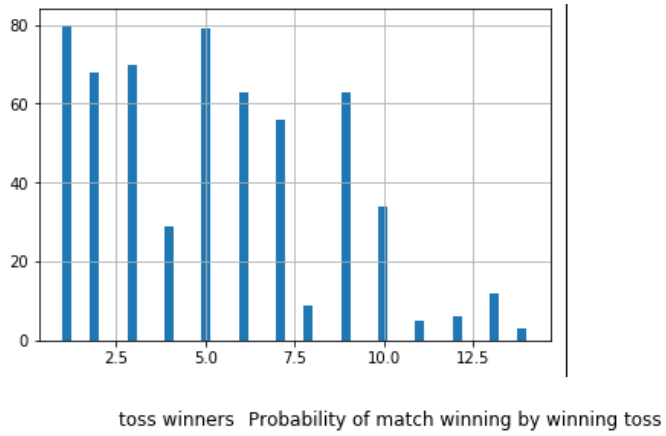

Figure 1

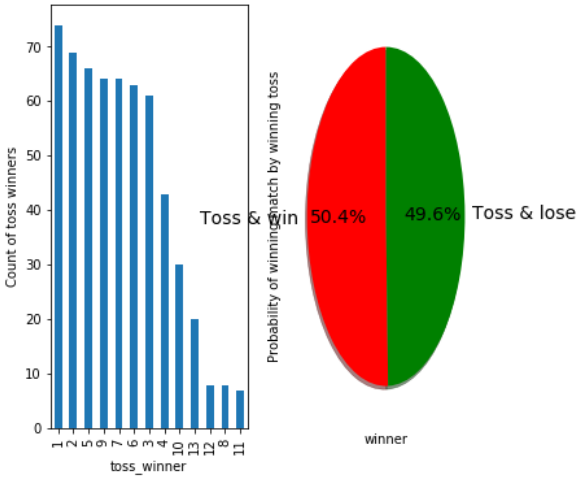

Figure 2

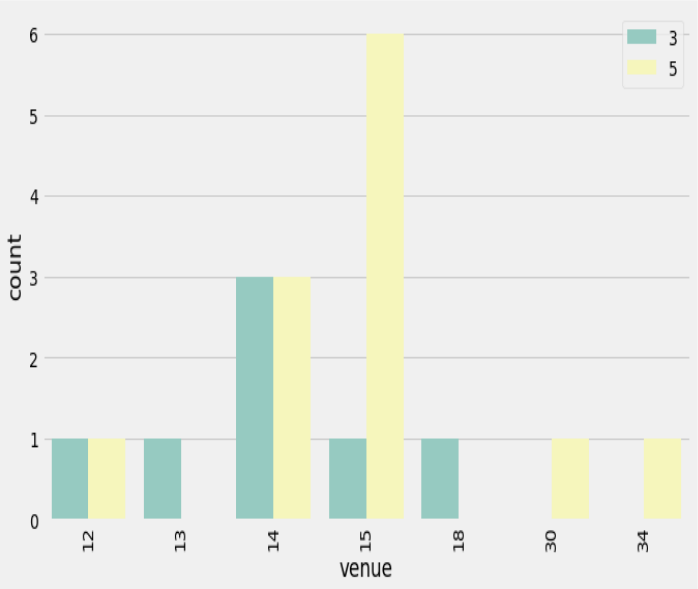

Figure 3

IV. Discussion

Major observation in the pattern is that winning the toss cannot be the main point for predicting the match future. Predicting the decision of winning the match mostly depends on the venue and city and stadium. From this outcome in general we can say whether a team is able to win the game based on the ground it is playing. Supported evidence is given in the figure 3. We have tested the algorithm for 2 new teams and the analysis shows that city and venue is very crucial for predicting the outcome.

V. Conclusion

The strongest and most important observation of the project prediction of a cricket match does not depend on only winning the toss and the decision of the toss. It also depends on the city and the venue on which the team is playing. The reader should remember that venue plays an significant result on the winner team.

Csv is used for analysis and generating a predictive model.

The data analytics is performed to compare features that influe nces theoutcomes of matches(win/lose).The features that are c onsidered are 'city','toss decision','toss winner' and'venue'.A pr edictive model is developed using Machine Learning algorith m to gethighest possible accuracy.Then feature importance ret rieved from predictive model applied reveals theimportance of features in determining match outcomes(win/lose).These feat ure importance is then compared against actual data to check f or validity. Model use this Random Forest Classifier Which gave 88% accuracy.

References

[1] [1] Ahmad, H., Daud, A., Wang, L., Hong, H., & Dawood, H. (2017). Prediction of Rising Stars in the Game of Cricket, 7. doi:10.1109/ACCESS.2017.2682162

[2] [2] Pradeep, I., & Singhe, W. (2017). Predicting the Performance of Batsmen in Test Cricket, 9(4), 4104-4124. doi:10.1109/ACCESS.2017.2682162

[3] [3] Shahjalal, M. A., Ahmad, Z., Rayan, R., & Alam, L. (2017). An Approach to Automate the Scorecard in Cricket with Computer Vision and Machine Learning, 17557499 . doi:10.1109/EICT.2017.8275204

[4] "CricAI: A Classification Based Tool to Predict the Outcome in ODI Cricket",Amal Kaluarachchi, Aparna S. Varde,Department of Computer Science,Montclair State University, Montclair, NJ, USA

[5] A. Joseph, N. E. Fenton, and M. Neil, "Predicting football resultsusing bayesian nets and other machine learning techniques," Know.-Based Syst., vol. 19, pp. 544–553, November 2006. [Online]. Available:http://dl.acm.org/citation.cfm?id=1222216.1222263

[6] "A Neural Network Method for Prediction of 2006 World Cup,Football Game",Kou-Yuan Huang, Senior Member, IEEE and Wen-Lung Chang

[7] J. Hucaljuk and A. Rakipovic, "Predicting football scores using machine learning techniques," in MIPRO, 2011 Proceedings of the 34th International Convention,May 2011, pp. 1623–1627.

[8] A. Joseph, N. E. Fenton, and M. Neil, "Predicting football results using bayesian nets and other machine learning techniques," 2005.[Online].Available: http://www.dcs.qmw.ac.uk/norman/papers/Spurs-2.pdf