International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

Predictive Modelling Using Random Forest Classifier and Decision Tree Algorithm

Subhojit Paul[1], Ankan Paul[2], Sahil Sakar[3], Tuhin Sarkar[4], Debdoot Ganguly[5], Siddhartha Sen[6], Soumik Dutta[7]

[1, 2, 3, 4, 5, 6, 7]Department of Electrical Engineering, UEM-Kolkata, India

Abstract: ***In the 21st century with the development of programming languages new dimensions are opening in the line of work which are improving the quality of life and also making it simpler and better. Cricket has a great influence on the people of this country and is almost considered as a religion. With the introduction of Indian Premier League the excitement of people has reached new heights. People often debate which team will win the upcoming match and the championship. The main focus of our program is to predict the winner of the upcoming matches using Random Forest Regression method. The dataset is fed with the names of the teams and the team which won the toss and the winner. Based on this data we predict which team may win the next match. Random Forest Regression being a supervised learning algorithm helps in accurate and stable prediction and as such is very helpful in gaining the desired result.***

Keywords: Machine Learning, Random Forest Regression

1. Introduction

Cricket is a popular sport across the world and specially in India. It is played in various formats be it Test or One Day International or T20 International. People all over the world watch the sport and as such arguments and debates are always ongoing on the better team and which team has the highest possibility of winning. The winning chances of a team depend much on the team winning the toss. The toss decides whether the team will be batting first or bowling first. As there is a 50-50 probability of chances of winning of a particular team. The predictive learning algorithm analyzes this data and predicts the winner of the match.

2. Random Forest Regression

Random Forest is a supervised learning algorithm. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. One big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems. The random forest model is very good at handling tabular data with numerical features, or categorical features with fewer than hundreds of categories. One important note is that tree based models are not designed to work with very sparse features. Another great quality of the random forest algorithm is that it is very easy to measure the relative importance of each feature on the prediction.

3. Procedure

The main focus of our project is to predict the winner of the future matches based on their past encounters. The parameter in the input data set is the names of the two teams, the team which won the toss and whether it is a away match or a home match for the team. Based on this data we can find the likelihood of the team winning the match. For this purpose we follow this algorithm:

Step 1: Start the model.
Step 2: The data is fetched from datasets which is Pre-processed i.e. the data is encoded.
Step 3: The missing data are filled using imputing method.
Step 4: K-Fold cross validation is used to compare and select a model for the given predictive modeling because it is easy to understand, easy to implement and results in skill estimates that generally have a lower bias than other methods.
Step 5: Then Decision Tree Classifier repetitively divides the working area into sub part by identifying lines.
Step 6: Random Forest Classifier is used to predict the outcome of the IPL match.
Step 7: Output is provided along with pictorial representation of the process.
Step 8: Stop the model.

4. Output

We use NumPy and Pandas here for the purpose of using the given data-set. Firstly we fetch the 'number of toss winners and match winners by each team's from the following data-set. The analysis of the input data could be analyzed more efficiently through Bar-graphs. So we use matplotlib open source library as a better way of plotting the data in bar-graphs.

After fetching and analyzing the data from the data-set here we use the Scikit-learn tool to implement the machine learning in this data-set. We import the following tools- Logistic Regression, K-Fold,

Random forest classifier and Random forest regressor from Scikit-learn machine learning library to implement the prediction model. The prediction model explains the percentage of win and loses with the respect of winning the toss.

In this way after implementing the pie-chart prediction model, we use Seaborn library for implementing the statistical infographics based on the inputs of two teams, toss winner and the venue where the match has been played. Thus we can predict the outcome of any match.

TEAM 1	TEAM 2	TOSS WINNER	CODE PREDICTION	ACTUAL WINNER
SRH	RCB	RCB	SRH	SRH
GL	KKR	KKR	KKR	KKR
RPS	KXIP	KXIP	KXIP	KXIP
GL	SRH	SRH	SRH	SRH
KKR	MI	MI	MI	MI
KKR	SRH	SRH	SRH	KKR
SRH	KXIP	KXIP	SRH	SRH
RCB	GL	GL	GL	RCB
KXIP	MI	MI	MI	MI
KXIP	KKR	KXIP	KKR	KKR

Figure 1: Comparitive study of toss winners and match winners

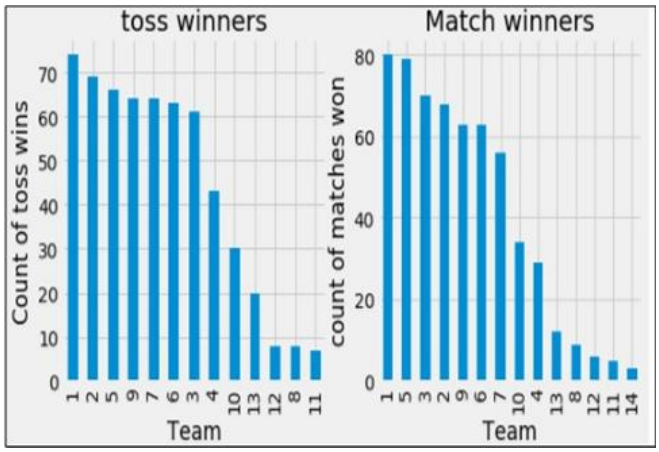

Figure 2: Comparison of real vs obtained result

5. Conclusion

To begin with machine learning comes into picture it is the general term for when computers learn from data - there are lots of different ways that machines can learn - the algorithms can be grouped into supervised, unsupervised, and reinforcement algorithms. Thus using such innovative and efficient algorithm which in turn has improved the accuracy level to a desired extent. Using regression technique in this project especially Random Forest Regression has really helped us to achieve all the favourable outcomes necessary. The above technique being a supervised learning algorithm uses the concept of decision tree to build and merge in order to get an accurate and stable prediction. In this work, we have used one of the machine learning algorithms which is Random Forests Regression, in order to predict the outcome of the Indian Premier League match. We used the dataset containing data of matches from beginning of the IPL season 1 to season 10. These number of matches details were obtained after putting the dataset through cleaning and pre-processing. The above technique being efficient in handling tabular data with numerical features with skylearn providing a great tool that measures a features importance by looking at a tree nodes has made things much simpler while predicting the IPL match winners and other events likely using its predictive algorithm. Thus displaying the nest course of action or outcomes as desired. Such algorithms finds its place in collaborative learning deeper Personalization, cognitive services in future.

References

[1] Aha,D. Kibler,D. (1991): Instance-Based Learning Algorithms. In Machine Learning, vol. 6 - 1. Kluwer Academic Publishers.
[2] Breiman, L., Friedman,J.H., Olshen,R.A. & Stone,C.J. (1984): Classification and Regression Trees, Wadsworth Int. Group, Belmont, California, USA.
[3] Clark, P., Niblett, T. : Induction in noisy domains, in Proc. of the 2th European Working Session on Learning ; Author: Bratko,I. and Lavrac,N. (eds.); Sigma Press, Wilmslow, 1987.
[4] Dillon,W., Goldstein,M. (1984) : Multivariate Analysis methods and applications; Author: John Wiley & Sons.
[5] A new heuristic of the decision tree induction , Authors:Ning Li ,Li Zhao, Ai-Xia Chen,Qing-Wu Meng,Guo-Fang Zhang ,International Conference on Machine Learning and Cybernetics 12-15 July 2009 ,DOI: 10.1109/ICMLC.2009.5212227
[6] Using Support Vector Machines to Classify Student Attentiveness for the Development of Personalized Learning Systems , Authors: Manus Ross,Corey A. Graves,John W. Campbell, Jung H. Ki, 2013 12th International Conference on Machine Learning and Applications, DOI: 10.1109/ICMLA.2013.66
[7] The application of machine learning algorithm in underwriting process, Authors: Yi Tan, Guo-Ji Zhang,2005 International Conference on Machine Learning and Cybernetics, Date of Conference: 18-21 Aug. 2005,Date Added to IEEE Xplore; 07 November2005,DOI: 10.1109/ICMLC.2005.1527552