A
Project
Report on

Data
Analytics

Estimation of the first innings score in an IPL match based on the venue and team.

Submitted by-

Prabhath Akula
(190C2020029)

Pavan Kumar Bellamkonda
(190C2020028)

Under the guidance of

Dr. Yogesh Gupta
(Associate Professor)

BML MUNJAL
UNIVERSITY™
FROM HERE TO THE WORLD

Department of Computer Science and Engineering
SCHOOL OF ENGINEERING AND TECHNOLOGY
BML MUNJAL UNIVERSITYGURGAON-122413, INDIA

Acknowledgement

This project was a very good opportunity for some experiential learning in the field of data analytics, apart from the project we implemented we were able to know how real-world projects actually takes place and how the data is collected, how the work is divided etc.

We are happy that we did a project which we can actually present in our resumes for future use.

Index

Abstract

The project is analyzing a dataset of ball-to-ball data of each match taken from the sources Cricsheet, Kaggle and some web scraping.

The project analyzes the data and calculates the overall score of the match of a particular team in a particular ground from the data available and stores the data calculated above.

Then we are applying some regression techniques on the stored data and predicts the score of first innings of a particular team in that particular ground.

Motivation

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. Being a cricket fan, visualizing the statistics of cricket is mesmerizing.

Hence, we decided to get my first hands-on experience by building a classifier to predict the winning team.

‘

1. INTRODUCTION

A typical Twenty-20 game is completed in about three hours, with each innings lasting around 75–90 minutes and a 10–20-minute interval. Each innings is played over 20 overs and each team has 11 players. This is much shorter than previously existing forms of the game, and is closer to the timespan of other popular team sports. It was introduced to create a fast-paced form of the game, which would be attractive to spectators at the ground and viewers on television. Since its inception the game has been very successful resulting in its spread around the cricket world and spawned many premier cricket league competitions such as the Indian Premier League. On most international tours there is at least one Twenty20 match and all Test-playing nations have a domestic cup competition. One of the International Cricket Council's (ICC) main objectives in recent times is to deliver real-time, interesting, storytelling stats to fans through the Cricket World Cup app or website. Players are what fans obsess most about so churning out information on each player's performance is a big priority for ICC and also for the channels broadcasting the matches.

Hence to solving an exciting problem such as determining the first innings score of T20 match based on the venue and the team playing would have considerable impact in the way cricket analytics is done today

2. PROBLEM STATEMENT

Build a model to predict the first innings score of any IPL match with respect to the venue of the match and the team batting first.

3. METHODOLOGY

Coming to the methodology of the project, we had the data set of ball-by-ball data of every ipl match from 2008. so firstly, we have to convert the ball-by-ball data to the summarized match by match data, to do this our model iterated from top to bottom of each cell of the match data's first 120 legal deliveries hence calculating the score of the first innings of that match taking the venue and the team playing into the consideration. The process is repeated for each and every match thus obtaining the match summarized data of team playing, venue and the first innings score.

With this the whole bundle of a data is converted in to simple to use and effective data. from there we took the average scores with respect to venue and the team and making it a dictionary in python so that we can easily access the variable instead of repeatedly taking from the data frame. Now we designed the linear regression model and the random forest regression model with respective attributes and we fit the data into the model and compiled it. we tested the accuracy of the model with the training set with is divided at the beginning of the model creation and accuracies of 83% and 87% occurred for linear regression and random forest models respectively.

Now we compiled the program making ready for the prediction.

4. RESULTS AND DISCUSSION

As discussed in the methodology section the processes of linear regression and the random forest regression have been used to create the relation between the two parameters and later predict the output. Both the process did the prediction reasonably well.

Linear regression model had the accuracy of around 83.7 percent with the training set whereas the random forest regression model had the accuracy of 87% .

We personally think that the reason for the random forest model having more accuracy than that of the linear model is because the random forest model takes the prediction from the average of different trees predicting the output and hence will consider more possibilities and where as the linear regression model is \just a best fit line for the venue average scores and the team average scores so the coefficient depends on not only the avg value of 1 venue but also using the data of other venues when predicting the score so the regression model is slightly off the courser when compared to the random forest model .

5. FUTURE SCOPE

- Add columns in dataset of top batsmen and bowlers of all the teams.
- Add columns that consists of striker and non-striker's strike rates.
- Implement this problem statement using Artificial Neural Network (ANN).

……………………………….END OF REPORT………………………………