International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 05 | May -2020 ISSN: 2582-3930

Cricket Matches Prediction using Machine Learning

Ishank Singh[1], Akash Mishra[2], Faraz Naqvi[3], Mohd. Afzal[4],

[1,2,3,4]B.Tech. 4[rth Year] Student, RKGIT, Ghaziabad, U.P.

[5]Fourth Year, Assistant Professor, Department of Computer Science and Engineering RKGIT, Ghaziabad, U.P.

ABSTRACT-

The objective of this paper is to emphasize on how cricket prediction works. Cricket is a growing sport in terms of money and value. It, therefore, needs a very good technical infrastructure and models to predict the outcomes of things. It can be used by the teams, players, fans, and brokers who want to invest in sports. The data and the predictions would focus mainly on the matches in the Indian Premier League (IPL). When we examine these data, it is very necessary to evaluate what is the best cricket team prediction and worst-case team prediction. It makes it easier to compare both situations by using Machine learning algorithms. An optimal machine learning model is necessary to match the prediction based on certain factors and conditions like past performances, home crowd advantage, climate conditions, performance at the specific venue, teams, and current forms of the players and teams.

1. INTRODUCTION

In this paper, we discuss machine learning models to predict the outcomes of the IPL matches. Cricket is unpredictable because at each stage of the game the momentum shifts to one of the two teams. Considering all these unpredictable scenarios there is a huge interest among the fans to do some prediction either at the start of the game or during the game. Many fans also play betting games to win money. So, with these possibilities, this reports the problem of predicting the match results before the game has started based on the statistics and data available from the data set. The dataset features were used as inputs to four different classification algorithms: naive Bayes, logistic regression, random forests, and gradient boosted decision trees. The datasets, data dependencies, and machine learning models have performed well in these scenarios. Characteristics and results of each model and direction for future research are discussed.

2 DATASET AND MODEL

The first step was to select the Indian Premier League (IPL) T20 cricket competition as the case study for this work. T20 cricket is a major growth-area within cricket currently and attracts a huge amount of media interest globally. It is relatively new as a format compared to one day and test cricket and so the coaches and players alike should be more engaged to learn in this format. League Stage games sometimes do not end with one side winning. We will either have 'no outcome if the weather stops the game from being finished or they can be tied' if both teams end up with the same score. In the knockout rounds in such cases, matches are decided by super over ensuring that there is always a winner.

2.1 Factor 1: Only Team Data

Statistical features are formed that represent the 'form' or skill level of a team. These datasets and charts are based on how teams as a unit perform overall in a match or set of matches.
Such cases include how teams have performed at a venue, how teams have performed at a phase of the season, how teams have performed against an opposition etc. The past success of the team can be considered to predict the outcome of the

match. History of games at a venue, how did the teams perform, performance at that specific venue, Performance against the specific opposition, and experience at the specific venue.

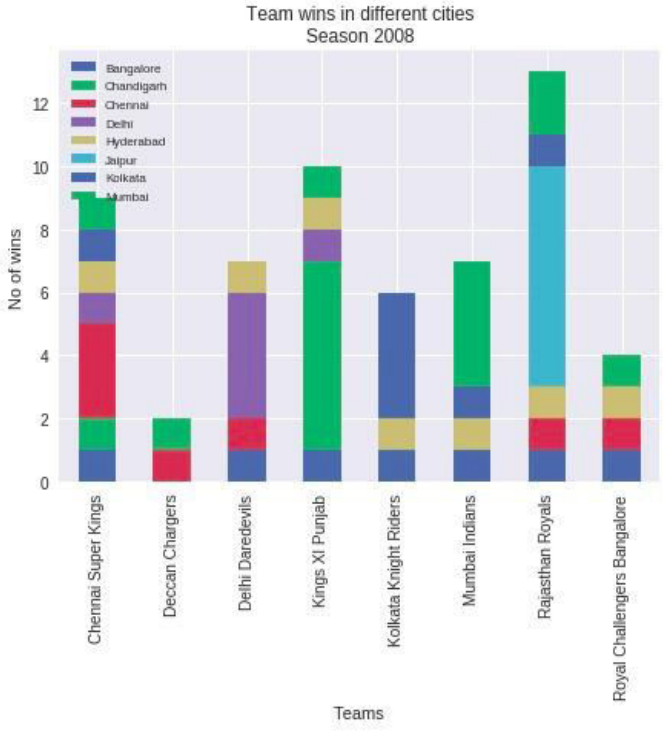

Chart -1: Team wins based on cities

The performance of individual teams in a season can also add up to odds of performance in the next upcoming season.

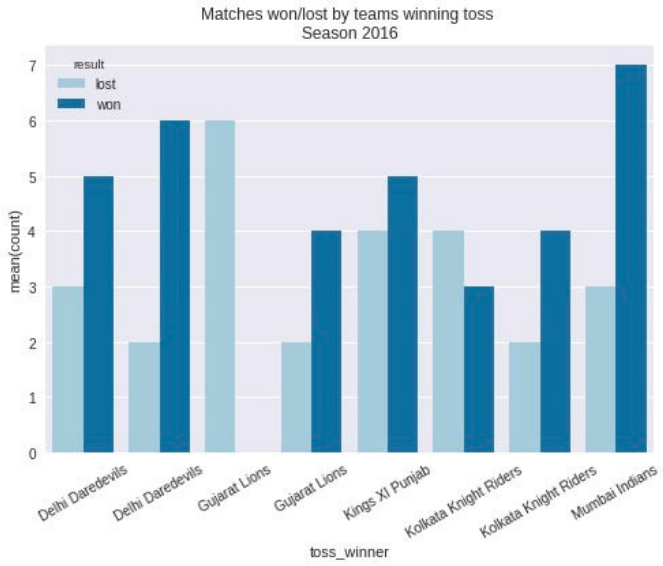

Chart -2: Teams performance in a season.

2.2 Factor 2: Only Player Data

These factor features were formed by calculating individual performances for each member of a team for batting, fielding, and bowling. This feature space was narrowed using methods that were shown to function well during the first modelling process. The optimal algorithms were retrained on this new data set to see if further improvements were possible.

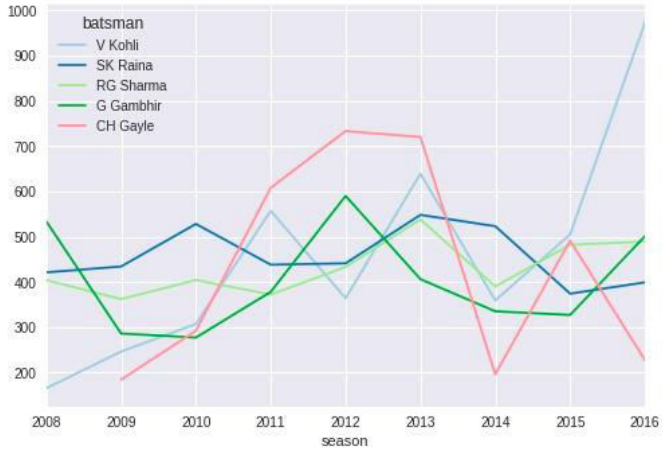

Chart 3: Individual batting performances in IPL.

.

3. PROBLEM STATEMENT

In this problem, we must design a machine learning model that could be used to predict a cricket match. The model could analyse multiple factors like the toss, batting side player performance, current team form, home advantage, etc.

4. ALGORITHMS

The Machine Learning algorithms which can be used for the necessity of the completion of this project are-:

4.1 K Mean Clustering algorithm

To predict score made by a batsman we make use of a clustering algorithm called k-means. It is a simple but effective algorithm. K-means clustering aims to partition n observations into k clusters in which each observation is located in a cluster with the nearest

mean acting as a cluster prototype. The algorithm is as follows:

Step 1: Begin with a decision on the value of k being the number of clusters.

Step 2: Position any initial partition to classify the data into clusters k. You can randomly or systematically assign the training samples as follows:

• Take the first k training sample as a single-element cluster.

• Allocate each of the remaining (N-k) training samples to the nearest centroid cluster. Recomputed the centroid of the increasing cluster after every assignment.

Step 3: Take every sample in the sequence, compute its distance from the centroid of each of the clusters. If the sample is not currently in the cluster with the closest centroid, move this sample to the cluster and update the cluster centroid accepting the new sample.

Step 4: Repeat Step 3 until the convergence is reached, that is until the pass through the training sample triggers no new assignments.

Working - The list of players gets populated automatically based on team selection. There is also an option to select a venue for the match. By pressing the fetch button, pre-stored database data is retrieved and divided into two clusters based on random cluster center assumptions by the program. Estimated data collected from multiple users is thus stored in the database. On clicking the Load button this data is retrieved and clustered along with past data.

After the final iteration, the cluster which contains a maximum number of entries, both past as well as expected data, is selected. The cluster centre of that particular cluster is the expected score and also the result of the system.

4.2 Prediction Modelling using Multiple Linear Regression

Regression is an inherently mathematical method commonly used in data mining. Linear regression is typically among the first few subjects that people select when studying predictive modeling. In this technique, the dependent variable is constant, the independent variable(s) may be constant or discreet and the structure of the regression line is linear. Multiple linear regression tries to model the relationship between two or more explanatory variables and the response variable by adapting the linear equation to the data being examined.

Multiple linear regression equations are the following:

$$Y = b0 + b1X1 + b2X2 + ... + bpXp$$

If Y is the projected or expected value of the dependent variable, X1 through Xp are separate independent or predictor variables, b0 is the value of Y when all independent variables (X1 through XP) are equal to zero, and b1 through bp is the approximate regression coefficient. The increasing regression coefficient reflects a change in Y relative to a one-unit change in the corresponding independent variable. For example, in multiple regression cases, b1 is the change in Y relative to a single unit change in X1, keeping all other independent variables constant (i.e. when the remaining independent variables are held at or set at the same value). Again, statistical tests can be conducted to determine whether each regression coefficient is significantly different from zero. In our model, we have taken into account batsmen in-growth, bowler, and wickets and run-rate. The equation that we used is V(b, w)=r(b, w)+p(b, w)V(b+1,w+1)+(1-p(b, w)))V(b+1,w) Since V(b*,w)=0 where v is match variable and b would be the bowler and w his wicket. By measuring v(b, w) we would have an estimated forecast of the result.

4.3 Logistic Regression

This is commonly used for classification issues. Logistic regression does not involve a linear relationship between dependent and independent variables. This can accommodate various types of relationships as it applies a non-linear log transformation to the expected odds ratio. To order to prevent over-fitting and under-fitting, all relevant variables should be included. A good approach to achieving this procedure is to use a step-by-step method to estimate logistic regression. It needs large sample sizes because the maximum likelihood estimates are less efficient at small sample sizes than the ordinary least square. Independent variables should not be associated with each other, i.e. no multi-collinearity. Nevertheless, there are possibilities for incorporating the interaction effects of categorical variables in the study and the model. If the value of the dependent variable is ordinal, it is called the Ordinal Logistic Regression. If the dependent variable is multiclass, it is known as Multinomial Logistic Regression.

5. CONCLUSIONS

The paper discusses the problem of predicting the outcome of an IPL match using machine learning models. The problem of match predictions which are faced a lot by the fans in such fantasy cricket is addressed in this paper with various machine learning models.

This paper aims to create an interest among the fans because of the rise of a new version of cricket which is the IPL and also to give correct predictions to them in order to play the fantasy cricket leagues. This could also help the teams in order to distinguish the opposition team forms and also to build the required strategy to win a game.

REFERENCES

1. Kumar, S. (2014). Rankings: A batting index for limited-overs cricket.
2. D. Roy Choudhury, Preeti Bhargava, Reena and SamtaKain, "Use of Artificial Neural Networks for Predicting the Outcome of Cricket Tournaments ".
3. Pardee, M. (1999). An artificial neural network approach to college football prediction and ranking.