IPL Data Analysis

by Trupti Jayanna

Submission date: 30-Nov-2020 03:51PM (UTC+0530)
Submission ID: 1460119381
File name: TEN_IPL_DATA_ANALYSIS-FINAL_REPORT.pdf (201.47K)
Word count: 3266
Character count: 16976

IPL Data Analysis

Trupti J
CSE Department
PES University
Bangalore, India
trupti7014@gmail.com

Neha Ganesh Shastri
CSE Department
PES University
Bangalore, India
neha.ganesh.shastri@gmail.com

Esha Arun
CSE Department
PES University
Bangalore, India
eshaarun2310@yahoo.com

ABSTRACT

Data Analytics is a fast-moving, up and coming branch of computer science that helps in solving various issues and problem statements. It helps with the decision of choosing a particular strategy and modus operandi. Data Analysis can make all the difference in the world when it comes to strategic thinking and informed problem-solving. In this paper, we use analytics to solve cricket-based problems.

I. INTRODUCTION

The Indian Premier League is a major white-collar cricket league that is conducted during the summer every year in the months of April and May. It consists of eight teams, each one representing either a major city or state in the Indian province. The Indian Premier League was founded and initiated by the Board of Control Cricket in India (BCCI) in 2008. Presently, the Indian Premier League consists of eight teams where each team plays against an opposing team in the league two times in a Round-Robin format in the initial phase of the Indian Premier League. At the conclusion of the primary league stage, the top four teams on the scoreboard, qualify for the playoffs. The top 2 teams (on the scoreboard) from the general phase then play against one another in the as first qualifying match of the Indian Premier League, with the team winner going right away to the premier final and the team loser, fortunately, getting another chance to redeem themselves by virtue of their position on the scoreboard, another chance to qualify for the Indian Premier League Finals, by playing the second qualifying match with the winner of the eliminator match that was played previously to determine the team that will play against the qualifying loser. Contemporarily, the teams ranked 3rd and 4th in the Indian Premier League phase, play against each other in an eliminator match, and the emerging winner from this particular match will go on to play the loser from the initially played qualifying match. The team that emerges the victor from this second qualifying match will move onto the final to play the winner of the first qualifying match in the finals of the Indian Premier League, where the victor is elegantly crowned as the champion of the Indian Premier League Season. So far, we have had 12 seasons of the IPL Tournament leaving plenty of room for analyses and predictions. Thus, the problems we try to solve and the questions we try to answer should be familiar to anyone who knows and understands Cricket. We try to predict various outcomes based on the statistics given to us to optimize our results and increase the rate of accuracy in prediction.

II. BACKGROUND

Match Statistics have always played a significant role in various sporting events. In fact, we can go so far to say literally 'All Sports'. Games and sports analytics have, over the years, continuously and consistently been predominant and a mega resounding hit in moulding success for many teams and players in various events. Various data analysts and strategists have already done a

plethora of analyses based on this particular field and we have done an extensive literature survey, studying as many models and visualizations as possible to gain an insight that would help us develop our own perspective to carry out our own unique analysis to help us with all the predictions and calculations that we have planned. The Indian Premier League's cricket teams depend strongly on qualified and competent data analysts to decide their strategy for an upcoming match. As it has been mentioned and understood with increased confidence previously, sports analytics is being relied upon with increasing importance and will continue to play an important role in how teams decide and work on their strategy and modus operandi, pick their teams, how they conduct the game, etc. Gaming and sports analytics and premier quality data visualization can often play a crafty role in choosing the best lay-up of a team. With the unsurprising and unprecedented increase in data creation and computing and collection – sports teams, coaches, sponsors, owners, and even gamblers, along with franchise owners all around us are tapping into the seemingly endless treasure of data and knowledge waiting to be analysed at our fingertips. Cricket, as one can understandably imagine, is rich and heavy with data points. The battle played between bat and ball, is played across different levels, different formats, etc... Players are told where the opponent typically pitches during the normal and death overs, players are shown various analyses and visualizations to help them understand the important and weak spots of opposition batsmen, and so on. The plethora of ball-by-ball analyses of matches can produce unprecedented and surprising hidden unforeseen, game-changing insights and thoughts, such as batting partnerships, batting order, bowler to be decided for that particular match, etc.

III. PROBLEM STATEMENT

The calculation of the best move by the team owner/franchise against a certain opponent, the choice of the perfect bowler/ wicket-keeper/ batsman who would be the ideal person to face a certain opponent, and the player who would work best under a stressful or pressurising situation. Our main focus vis-a-vis this dataset is that given a situation, in a T-20 cricket match, we would like to predict the expected performance of any given match in the situation against a certain situation/condition/batsman/wicket-keeper or a venue/bowler/team or other previously unforeseen conditions. We would also mainly like to predict the winning team in a given match and situation. Prediction involves persistent and meticulous analysis and research of a myriad of datasets as every calculated prediction is made after weeks of dedicated research. It is imperative for every IPL team to have a qualified analyst to study every single match and advise the coach on the best strategy to be adopted for each opposing team that they would be facing which would also have to be updated regularly with every single match. This is a very fascinating and truly intriguing process.

IV. LITERATURE SURVEY

[1] Predicting the Outcome of Indian Premier League (IPL) Matches Using Machine Learning by Rabindra Lamsal and Ayesha Choudhary Link-https://www.researchgate.net/publication/327904009_Predicting_Outcome_of_Indian_Premier_League_IPL_Matches_Using_Machine_Learning.

Assumptions made: The various factors that were considered for this analysis include - away(guest) team, host(home) team, toss winner, toss decision, the weight of the home team, the weight of the opposing team. These were considered to potentially influence the win probability of a team by a considerable amount and hence used for the prediction. The official website of the Indian Premier League has a Player Points section where every player is awarded points based on these 6 salient features:

(i) number of fours scored in each match
(ii) number of catches in each match
(iii)number of wickets taken in each match
(iv) number of stumpings in the match
(v) number of dot balls bowled in each match
(vi) number of sixes scored in each match

The prediction model used in this paper makes use of multivariate regression to calculate the attack points of each player in the league and comprehensively, using computational mathematics, decide the overall strength of each player and team. Every player of the team is sorted in descending order according to their number of appearances in previous matches of the same season. This is done to find the average strength of each battling team. To carefully deduce how the management assigned points to each player based on these 6 features, MVR was the crafty model that was used on the players' points data to find the various different weights for each of these salient features. Results Observed: Out of all the different ML models, the MLP classifier performed the best,

with a classification accuracy of 71.7% and an F measure of 0.73. Based on this classification accuracy that was observed and reported, the MLP classifier was followed by Random Forests models, Logistic Regression models, and SVC SVM classifiers. However, it was also observed that Extreme Gradient Boosting and Naive Bayes classifiers performed poorly in predicting the results of the 2018 premier league matches.

Lacuna: In this work, it has been carefully assumed that every player of the team is sorted in descending order according to their number of appearances; while using this assumption, it can be inferred that the playing 11 of the team may be inconsistent in a number of players (bowlers and batsmen). Our models could not rely entirely on the approach and insight gained from the above paper. We have made use of 2 datasets--deliveries.csv and matches.csv to carry out our analyses and make the concerned predictions. We have faced numerous challenges in terms of accessibility and usability of our chosen datasets as they needed an extensive amount of cleaning before being available for model testing and analysis. We have used and tested several models so that we could compare and later choose the best model. We have also not tried out the RFE (Recursive Feature Elimination) model as we were confident about other models giving us stronger and credible results. We had also tried out 'Random Forest Regressor' to experiment and see whether it is in fact a stronger estimator than the 'Random Forest Classifier' in our analysis as instead of a specified kind of classification, we have principally just tried to extrapolate the values to give us the desired result. This is in addition to certain other models that have also been used to test the validity and credibility of our foreseen prediction.

V. PROPOSED SOLUTION

[i] PRE-PROCESSING AND DATA VISUALIZATION

Data prepping and filtering can take a considerably significant amount of wasteful processing time. The presence of too much noisy or unnecessary and unwarranted extra information present or unreliable data, then knowledge discovery during the training phase becomes cumbersome. Data Pre-Processing includes meticulous cleaning, data transformation, and boundary selection, etc... The final cleaned product of data preprocessing is the final training set which can then confidently be used further to analytically design and train models to predict certain warranted values.

In our first dataset (deliveries.csv), We begin the process of data cleaning by removing the outliers from the 'innings' column. In our second dataset (matches.csv), the Pandas DataFrame's dropna() function has been used to remove rows and columns with NaN values from the 'umpire1', 'umpire2', and 'umpire3' columns. We also drop the column containing the names of umpires who officiated the matches as this information is observed to be redundant. Following this, we proceed forward and remove records having null values in the 'winner' column. Another very important part of data pre-processing involves the removal of redundant team names. Over the years, the names of the IPL teams have changed due to various reasons in regard to sponsors and ownership. This particular issue is something that should not be overlooked and needs to be dealt with utmost care and precision. We have replaced 'Sahara Pune Warriors' with 'Rising Pune Supergiants', 'Deccan Chargers' with 'Sunrisers Hyderabad', and 'Delhi Daredevils' with 'Delhi Capitals'. In this dataset, we have also observed the column containing the names of the cities have a considerable amount of NaN and Null values. But interestingly, the 'venue' column was observed to have no null values, and hence, we filled the null values in the column containing city names from the information given in the column containing the venue of the cricket match. Data Visualization provides a graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. For our visualization, we have made use of various visualization tools such as heatmaps, grids, line graphs, bar graphs, scatter plots, matrix charts, etc… to represent our data for the purpose of understanding and appreciating its significance in order to further comprehend the elegance of our representation and find relative comfort while designing and building models.

[ii] MODELS TESTED

We have tested various models to work and understand the best estimate for the final calculated prediction. We have begun the entire process by encoding data in order to make it easier for the ML models to run efficiently without any confusion. We

have coded a generic function, based on whose design, we have run all our classification models and calculated our performance assessment scores in terms of percentage of accuracy. We have also calculated the Cross-Validation, which basically primarily resamples evaluated complex ML models on a limited data sample. This cross-validation method involves a single parameter k representing the number of groups that a given data sample can be chosen to split into.

In all our models, we have used 'winner' as our dependent variable[Y] and 'team1', 'team2', 'venue', 'toss_winner', 'city' and, 'toss decision' as our independent variables, based on which we have predicted the winner of the match. We will elaborate briefly on each ML model that we have used.

[a] **Logistic Regression**- Logistic Regression is a machine learning model that is a sigmoid, statistical ML model that in its primitive form, uses a sigmoid function to model a dependent(Y) variable, along with the presence of various other complex mathematical extensions.

In our LR model, we observed the accuracy to be 32.713%, and the Cross-Validation Score to be 30.319%.

[b] **Naive Bayes Classifier**- The Naive Bayes model is a primitive technique for building classification models. These models cleverly assign class labels to problem instances which are later craftily represented as vectors of values that determine features, where the class labels are drawn from a finite set.

In our NBC (Naive Bayes Classifier) model, we observed the accuracy to be 19.548%, and the Cross-Validation Score to be 17.420%.

[c] **K-Nearest Neighbour Classifier**- K-Nearest Neighbour is a type of instance-based machine learning model, where the function is approximated locally and all calculations and computations are deferred until the evaluation of the function. Since this algorithm relies on boundaries for classification, the accuracy of the training data can be improved significantly and dramatically after normalizing the dataset.

In our KNN model, we observed the accuracy to be 60.771%, and the Cross-Validation Score to be 41.622%.

[d] **Support Vector Machine**- A support vector machine (SVM) is an ML MODEL that computes data for intelligent regression analysis. SVM sorts data into one of two categories- called Binary Classification, which was what we required for our analysis.

In our SVC SVM model, we observed the accuracy to be 43.085%, and the Cross-Validation Score to be 40.293%.

[e] **Random Forest Classifier**- Random Forest Classifier can be used for various different tasks including binary classification and complex regression. It is an extensive ensemble method, which means that a random forest model is made up of a large number of smaller decision trees, which each produce their own predictions. These are called Estimators.

In our Random Forest Classifier model, we observed the accuracy to be 86.170%, and the Cross-Validation Score to be 49.601%.

[f] **Decision Tree Classifier-** The decision tree ML model makes use of multiple alggorithms with the abstract of node based data splitting in terms of making a decision. We have made use of various inbuilt greedy decision tree algorithms to understand and get the most optimal solution.

In our Decision Tree Classifier model, we observed the accuracy to be 86.835%, and the Cross-Validation Score to be 51.064%.

VI. RESULTS

Observed low accuracy on either classification or ML model, would mean that our classes are not very well separable with the current features that we have used. This can be remedied by finding better features. We would need to intelligently move forward and experiment with other models with more complex decision boundaries. A recurring issue in estimating LR models is a failure of the likelihood maximization algorithm to converge. In the majority of cases, this lack of accuracy that is observed is a direct consequence of data patterns, ie. quasi-complete separation. Also, Logistic Regression only roughly estimates a linear boundary. Therefore, when there is a non-linear separation of labels to be considered, LR models fail badly. This is why we observe a low accuracy percentage.

Though SVMs are comparatively better and more versatile models. Still, we observe the accuracy score to be comparatively lesser. This could be because of 2 main reason:

[i] SVMs work better and faster when there are plenty of samples for each class, if there are only a few samples for some classes then a KNN Classifier might work better than an SVC SVM. KNN models can learn very quickly while SVMs take a little longer, but learn better and more vigorous frontiers because of its inherent ability to maximize the boundary margin.

[ii] When data is not linearly separable, we would have to use SVM with a kernel. But unfortunately, with the presence of large amounts of data, with the need for a kernel might lead to some important performance issues. There are approximations that could be used to overcome this, the most famous one being the Nystrom approximation but one has to be careful because a kernel SVM with the Nystrom approximation may not perform as well as some other algorithms. The option to pre-compute a kernel matrix here is arguably infeasible and the consistent computation of the kernel every time it is needed might be tiring and expensive.

We also observe a low accuracy when we try the Naive-Bayes model. The main issue that is observed with the Naive-Bayes is, if there are no observed obvious occurrences of a class label and the value of an attribute together, then the frequency-based probability estimate becomes 0. This problem is more clearly observed during Conditional Independence.

KNN models tend to perform well, even overperform, when there are many instances and very few dimensions, but the model designer has to be very careful about the performance issue as a brute-force version of KNN can be very slow when there is a mountain of data being reported. Seeing how SVM is generally observed to be a better model than KNN, this result is not very surprising. So KNN is good when there are many points, but at the same time, it becomes too slow.

Random Forest Classifier deals with the process of splitting that is applied on a random subset of primary features. This articulately means that at each split of the tree, the model intelligently takes into consideration, only a small subset of features rather than all of the extensive, innumerable features of the model. Nevertheless, it fell short of our decision tree classifier.

The Decision Tree Classifier is our most optimal model and gave us the best results for our prediction. A whopping 86.835% accuracy was observed, making it an unbeatable and most reliable model with respect to our analysis.

VII. CONCLUSION

[i] REFERENCES

[i] D. Jyothsna, K. Srikant, *Analyzing and Predicting the outcome of IPL Cricket Data,* April 2019

[ii] L. Rabindra, A. Choudhary, *Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning,* 2018

TEAM MEMBERS' CONTRIBUTION

1) Trupti J - Exploratory Data Analysis, Model Testing, and Video Presentation

2) Neha Shastri - Literature Survey, Model Testing, Intermediate, and Final Report Design Structure

3) Esha Arun - Literature Survey, Intermediate, and Final Report Design Structure and Analytical Framework

IPL Data Analysis

ORIGINALITY REPORT

13%	8%	3%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	en.wikipedia.org Internet Source	3%
2	deepai.org Internet Source	3%
3	Submitted to University of Hertfordshire Student Paper	1%
4	www.youtube.com Internet Source	1%
5	Vidit Kanungo, Tulasi B. "Data visualization and toss related analysis of IPL teams and batsmen performances", International Journal of Electrical and Computer Engineering (IJECE), 2019 Publication	1%
6	Submitted to Victorian Institute of Technology Student Paper	1%
7	Alberto Cosimato, Roberto De Prisco, Alfonso Guarino, Delfina Malandrino et al. "The Conundrum of Success in Music: Playing it or	1%

	Talking About it?", IEEE Access, 2019 Publication	
8	Submitted to North Lake College Student Paper	1%
9	Submitted to The Robert Gordon University Student Paper	<1%
10	Submitted to University College London Student Paper	<1%
11	Submitted to University of Strathclyde Student Paper	<1%
12	export.arxiv.org Internet Source	<1%
13	ascelibrary.org Internet Source	<1%

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		