IPL DATASET ANALYSIS AND SCORE-ACCURACY PREDICTION

ABSTRACT

In this study machine learning algorithms were applied on real-world IPL datasets and to find out their useability on certain kind of patterns in the data. On applying certain kind of visualization techniques and analysis modules, the data presented a much larger dimensional view and scope of information on the basis of which many inferences can be inferred.

INTRODUCTION

Cricket is played and watched by millions of people around the world. It has fascinated sports' lovers and quite unique in character and ever growing in popularity. The **Indian Premier League (IPL)** is a professional Twenty20 Cricket League in India contested annually by franchise teams representing various Indian cities. The **IPL** is the most attended cricket league in the world and rank sixth among all sports leagues. Every year people are very curious to find out which cricket team is going to win the game and experts make predictions about various aspects of the Twenty20 Matches on the bases of data analytics using machine learning technology.

Machine learning is a sub- domain of Computer Science field which tries to modulate and predict the solutions of real-life problems in the hope of a better version than the previous state. the abundance of data aggregated over the past few years has generated the need for data analytics in this field. Large datasets of IPL match and its intricacies makes the task of comprehending accumulated large datasets quite complex hence, machine learning has come as a saviour for many scholars and professionals.

Machine learning problems may be categorized in three distinct categories:

- Regression Problem Statement
- Classification Problem Statement
- Clustering Data Analysis

SCOPE OF THE STUDY AND SELECTION OF DATASETS

The dataset chosen is a sport-driven dataset where it comprises the information of the matches played in the Indian Premier League (IPL) from 2008-2016.The dataset has been taken from Kaggle, which is the world's largest data science community. The description of each of the following attributes is as follows:

ID -Match ID

Season – season/year of the IPL being played

City- Name of the city

Date- Date of the game played

Team1- Team 1 of the IPL for that particular match

Team2- Team 2 of the IPL for that particular match

Toss Winner - Winner of the toss (Either Team1 or Team 2)

Toss Decision – The decision taken by the winner of the toss

Result- Result whether it was Normal, Draw or Abandoned

Dl applied – Duck-Worth Lewis Applied or not

Winner – Winner of the match

Win by Runs – The margin of runs by which a team won batting first

Win by Wickets – The margin of wickets by which a team won batting second

Player of Match – The man of the match

Venue- Stadium Name

Umpire1 – Name of Umpire1

Umpire2 – Name of Umpire2

Umpire3 – Name of Umpire3

PRE-REQUISITES OF THE STUDY

1. **Python -** Python is a popular object-oriented programming language having high-level capabilities. Python has few keywords, simple structure, and a clearly defined syntax. This allows the user to pick up the language quickly. Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh and one can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient. Also, it supports GUI

applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.

2. **Jupyter Notebook** - The Jupyter Notebook App is a server-client application that allows editing and running notebook documents via a web browser. The Jupyter Notebook App can be executed on a local desktop requiring no internet access or can be installed on a remote server and accessed through the internet.

   Also, the other libraries used for this particular dataset for analysing, visualizing and pre-processing the data are supported by the Jupyter Notebook and are as follows:

   - Numpy
   - Pandas
   - Scikit-Learn
   - Seaborn
   - Matplotlib

READING AND VIEWING DATA

After reading the data, the different attributes are checked to find out what they have to offer so as to get a subtle look before the start of the analysis.

```
In [5]: ipl_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 636 entries, 0 to 635
Data columns (total 18 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   id               636 non-null    int64
 1   season           636 non-null    int64
 2   city             629 non-null    object
 3   date             636 non-null    object
 4   team1            636 non-null    object
 5   team2            636 non-null    object
 6   toss_winner      636 non-null    object
 7   toss_decision    636 non-null    object
 8   result           636 non-null    object
 9   dl_applied       636 non-null    int64
 10  winner           633 non-null    object
 11  win_by_runs      636 non-null    int64
 12  win_by_wickets   636 non-null    int64
 13  player_of_match  633 non-null    object
 14  venue            636 non-null    object
 15  umpire1          635 non-null    object
 16  umpire2          635 non-null    object
 17  umpire3          0 non-null      float64
dtypes: float64(1), int64(5), object(12)
memory usage: 89.6+ KB
```

Fig.1 – Information of the different attributes of the dataset

METHODOLOGY AND DATA ANALYSIS

For analysis of data exploratory approach has been used. After looking into the spread of the data, some key aspects of significance and some unwarranted assumptions were explored during the analysis.

- Data analysis revealed that the most key players who turned out be the game changers for their team franchise were certainly the man of the match for the matches played during the period 2008-2016. Analysis was performed on the data to identify the top 5, top 10, top 20 players with highest number of men of the match awards during the season played from 2008 to 2016. It was noted that **C H Gayle, Y K Pathan, DA Warner, AB De Villers, Rohit Sharma** were amongst the main key players contributing to the target variable **'winner'**.

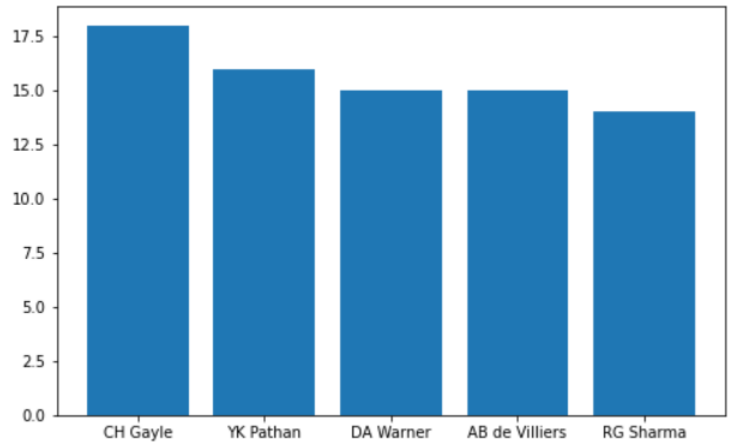
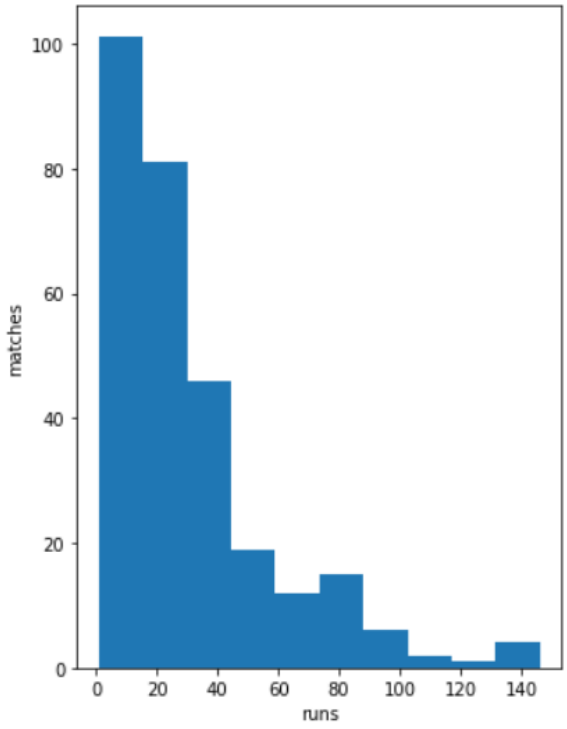


Fig. 2 – Plotting of a bar graph with most no. of MOM

- Another key thing to take into account was that the 3 out of top 5 teams with most wins throughout the seasons were the teams who have won the greatest number of tosses won. It indicates that winning the toss had a slight positive magnitude upon the result i.e. which team was going to won the match.

- Also, it was noted that only less than 12 matches the team batting first won by a margin of >=120 runs. Making it seem that most of the time, it was a close call between the two teams competing.



up

Fig.3- Histogram depicting margin of runs win

- The top 5 teams who were better at defending their total were **Mumbai Indians, Chennai Super Kings, Kings XI Punjab , Kolkata Knight Riders, Royal Challengers Bangalore** indicating that they had a strong bowlers contingent to defend their total in the second innings.

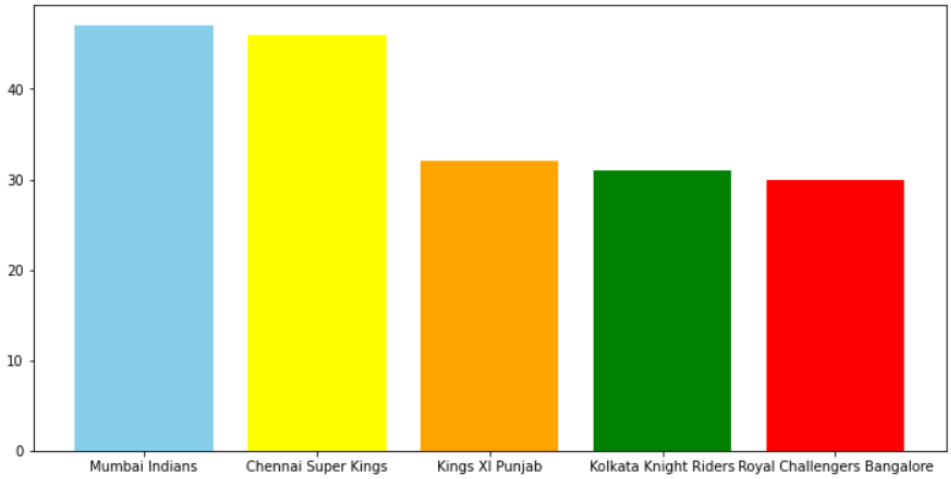


Fig.4- Bar graph for most wins batting first

On the other hand, teams batting second with the most wins comprised **Kolkata Knight Riders, Mumbai Indians, Royal Challengers Bangalore, Delhi Dare devils, Rajasthan Royals**. And by the no. of wickets, they won from were most ly by **6-7 wickets** in about **135 matches** making them clearly dominant in the ga me for the most part.

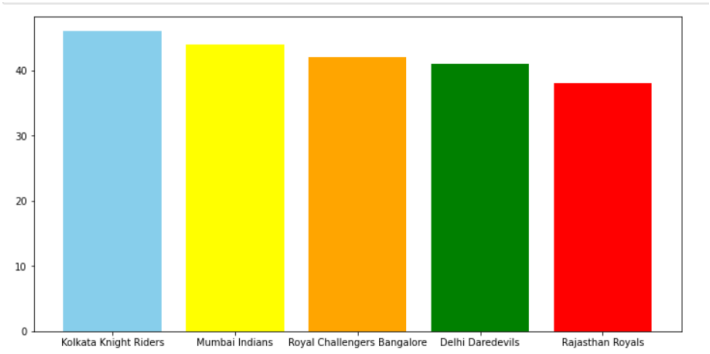


Fig.5 – Most wins batting second

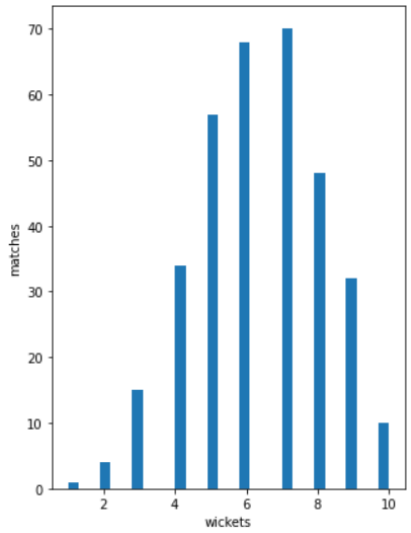


Fig.6-No of wickets by which the team won

- The most played venues across all seasons have been shown in the graph below, and to explore the entity venue Feroz Shah Kotla was chosen as a n example to show for what purpose after winning the toss which stadium was most suitable for.

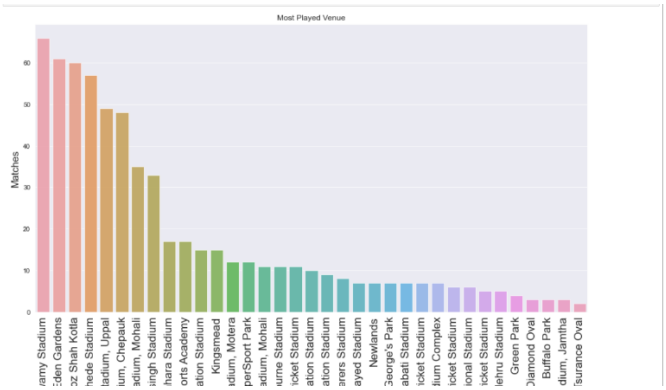


Fig-6 venue with most matches

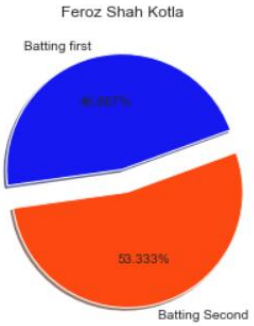


Fig-7 Feroz Shah Kotla

As seen the stadium Feroz Shah Kotla was much suitable for batting second afte r winning the toss as compared batting first.

Co-relation Matrix

Before processing and building model, the impact each significant attribute has on the other attributes was evaluated for choosing the most distinct feature for building the model.

On applying the co-relation matrix, it was found that toss winner teams and the venue seemed to have some impact on deciding whether a team is going to

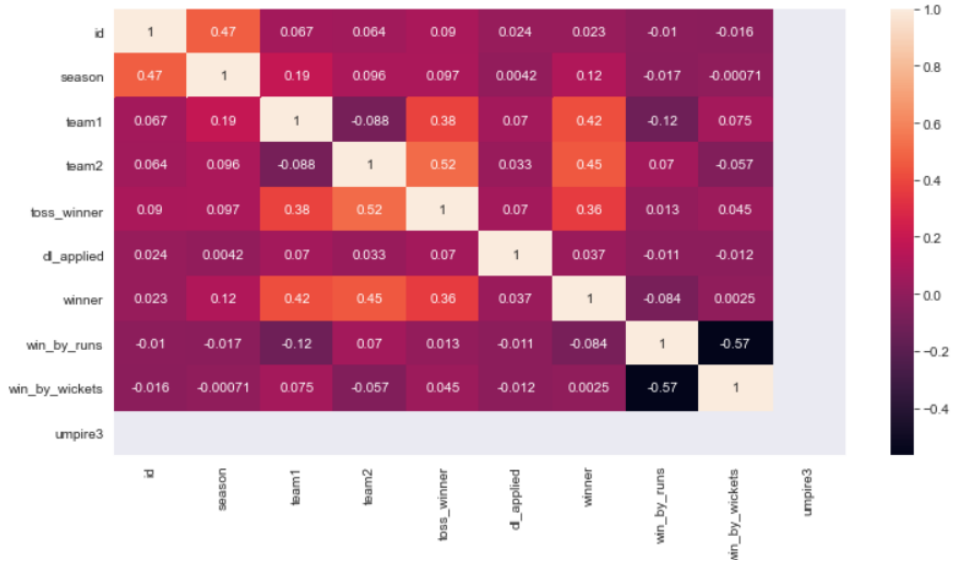


Fig-8 heat map for co-relation matrix

Tackling Null Values

The dataset does present some null values but these are minute and do not seem to have a significant impact on the data so they can either be dropped or be fille d by desired choice or be imputed into the transform data.

Encoding for categorical attributes

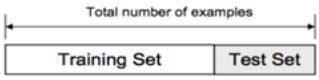
The data presents a challenge as it was inferred after the analysis that many attri butes were not comparable as they were categorically different with respect to many. To tackle this problem, I applied the method of encoding where I have ta ken ‘team1’ , ‘team2’ , ‘toss_winner’ , ‘winner’ and encoded with desired varia bles . Also for the rest of the taken attributes such as ‘city' , ‘venue’ , ‘ toss_deci

sion' and applied label encoding to make the task easy. All the attributes had a n umerical value and were managed easily.

Building Model

Train-Test Split: The data was split into training and test sets. The model with 80% of the samples was trained and with the remaining 20% was tested. It was done to assess the model's performance on unseen data. To split the data, train_test_split function was used which was by scikit-learn library. We can check this by printing the size of our training and test set to verify if the splitting has occurred properly.

One key thing that we need to maintain is that after making the split in the data to train and test the data remains stratified for accurate results.



Choosing the models

Logistic Regression Model: - Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

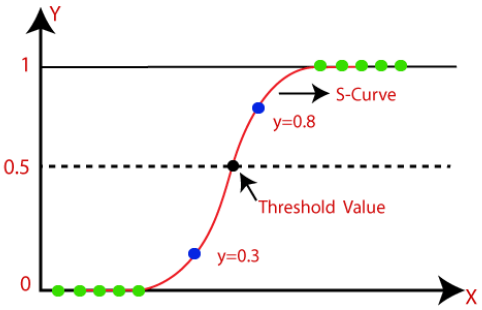


Fig-10 Logistic Regression Graph

The model gives us an accuracy of **0.31** with cv score values as **[2.81452814 3 .49088449 3.34018903 3.22275327 3.15757772]** followed by the RMSE value of **3.2298529375681624**, which is not good enough so we choose a rather stron gest model to evaluate our accuracy and RMSE score.

Random Forest

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called **Bootstrap Aggregation**, commonly known as **bagging**. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

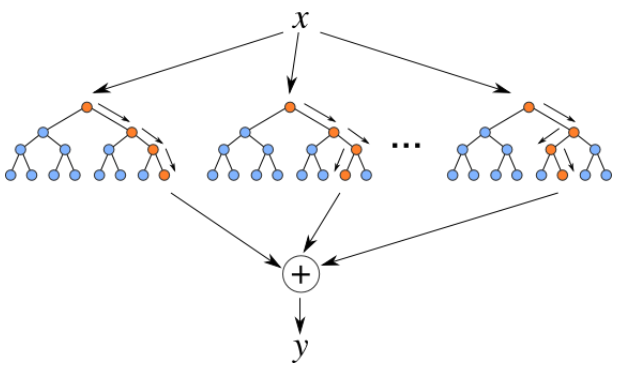


Figure-11 Random Forest

The model has been trained with **100 esembles(trees)** giving us the mean accur acy of **0.91(91%)** with cv scores as **[2.81452814 3.49088449 3.34018903 3.222 75327 3.15757772]** with a RMSE score of **3.205186526793562**.

Result : Testing the Model on Test Data

When we apply the Random Forest model on the test data set we get an accurac y of **0.95(95%)** with an RMSE score of **3.305838855157762.** This means the ra ndom forest model predicts 95% percent of the time correctle on which team is going to win a given match.

References

[1] "Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition. Datasets: Coronary Heart Disease Dataset." Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition.

[2] Hastie, Trevor, Robert Tibshirani, and J. H. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction: With 200 Full-color Illustrations. New York: Springer, 2001."

[3] Norving, Peter, and Stuart Russel, "Artificial Intelligence: A Modern Approach. S.l.: Pearson Education Limited, 2013."

[4] Witten, I. H., and Eibe Frank," Data Mining: Practical Machine Learning Tools and Techniques. Amsterdam: Morgan Kaufman, 2005."

[5] Schölkopf, Bernhard, Christopher J. C. Burges, and Alexander J. Smola," Advances in Kernel Methods: Support Vector Learning. Cambridge, MA: MIT Press, 1999."

[6] MICHAEL AARON WHITLEY, "Using statistical learning to predict survival of passengers on the RMS Titanic by Michael Aaron Whitley, 2015."

[7] Kunal Vyas, Zeshi Zheng, Lin Li, "Titanic- Machine Learning From Disaster- 2015."

[8] Andy Liaw and Metthew Wiener, "Classification and Regression by Random Forest", vol. 2/3, December 2002.

[9] Dr. Neeraj Bhargava, Girja Sharma," Decision Tree Analysis on J48 Algorithm for Data Mining", Volume 3, Issue 6, June 2013.

[10] Stuart J. Russell, Peter Norvig,” Artificial Intelligence: A Modern Approach, Pearson Education”, 2003, p.p. 697- 702.

[11] Machine Learning Benchmarks and Random Forest Regression, Segal, Mark R, 2004.

[12] A. Ng. CS229 Notes. Stanford University, 2012

[13] Kaggle. (2012). Housing: Machine Learning from Disaster.

[14] Kuhn, J., and Johnson, K. (2013),” Applied Predictive Modeling”, 1st edition. New York: Springer

[15] john D. Kelleher, Brain Mac Namee, Aoife D’Arcy,” Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms”

[16] Lonnie Stevans, David L. Gleicher, “Who Survived the Titanic? A logistic regression analysis”-Article in International Journal of Maritime History, December 2004

[17] S. Cicoria, J. Sherlock, M. Muniswamaiah, L. Clarke, "Classification of Titanic Passenger Data and Chances of Surviving the Disaster", Proceedings of Student-Faculty Research Day CSIS, pp. 1-6, May 2014.

[18] Corinna Cortes, Vlasdimir Vapnik, “Support-vector networks”, Machine Learning, Volume 20, Issue 3, pp 273-297.

[19] Prediction of Survivors in Titanic Dataset: A Comparitive Study using Machine Learning Algorithms, Tryambak Chatterlee, IJERMT-2017.

[20] Zhenyan Liu, Yifei Zeng, Yida Yan, Pengfei Zhang and Yong Wang, Machine Learning for Analyzing Malware, Journal of Cyber Security and Mobility, Vol: 6 Issue: 3, July 2017.