Vishwakarma Government Engineering College, Chandkheda

Project Report

Project – IPL Score Prediction

Subject Name: Data Warehousing and Mining (Professional Elective Course – III)

Subject Code: 3161610

Prepared by:

Vishwa Chatarara (180170116005)

Yas Patel (180170116036)

Index

1. **Introduction**
2. **Literature Study**
3. **Problem Definition**
4. **Data Preprocessing**
5. **Modeling**
6. **Conclusion**
7. **References**

1) Introduction

Cricket is the second most popular sports in the world with billions of fans across India, UK, Pakistan, Africa, Australia, etc. It is an outdoor game played on a cricket field at 22-yard rectangular long pitch, between two teams consisting each of 11 players. It is played in three formats namely Test, One Day International (ODI) and Twenty Over International (T20). The **Indian Premier League** (**IPL**) is a professional Twenty20 cricket league, contested by eight teams based out of eight different Indian cities. The league was founded by the Board of Control for Cricket in India (BCCI) in 2007. It is usually held between March and May of every year and has an exclusive window in the ICC Future Tours Programme.

The IPL is the most-attended cricket league in the world and in 2014 was ranked sixth by average attendance among all sports leagues. In IPL each team takes its chance to bat, trying to score as many amounts of runs which can be scored in 20 overs while the other team fields for that much number of overs. Each chance is termed as an innings. The batsman looks for making runs by hitting the ball being bowled to him. The bowler on the other hand tries to get the batsman out. There are certain rules defined to get the batsman out by the bowlers or the fielders. Each batsman keeps on batting until he gets out. So, the innings of the batting team is over when either the 10 batsmen got out or the 20 overs have been bowled by the fielding team; in either of the situation the batting team now gets the chance of bowling and the bowling team gets the chance of batting. The team which scores more runs wins the match. Unlike other sports, cricket stadium's size and shape is not fixed except the dimensions of the pitch and inner circle which are 22 yards and 30 yards respectively. The cricket rules do not mention the size and the shape of the field of the stadium. Pitch and outfield variations can have a substantiate effect on batting and bowling. The bounce, seam movement and spin of the ball depends on the nature of the pitch.

The game is also affected by the atmospheric conditions such as altitude and weather. A unique set of playing conditions are created due to these physical differences at each

venue. Depending on these set of variations a particular venue may be a batsman friendly or a bowler friendly.

Currently, in an IPL match the projected scores can be seen displayed at the score card during the first innings, which is basically the final score of the batting team at the end of that innings if it scores according to the current run rate or a particular rate. Run rate is defined as the number of runs scored per the number of overs bowled. However, run rate is considered as the only criteria for calculating the final score. But there are other factors too which may affect the final score like number of wickets fallen, the venue and the batting team itself.

In this project, we are trying to predict powerplay score of each innings.

In the project XGboost, Random forest, Gaussian Naïve Bayes algorithms are implemented as they showed us more accurate results than common regression techniques. **XGboost** is an implementation of gradient boosted decision trees designed for speed and performance that is dominative competitive machine learning.

2) Literature Study

Very few have worked in statistically predicting the scores or the outcome of the ODI match. One such work called “Winning and Score Predicting (WASP)”, which has been done by Scott Brooker and Seamus Hogan at University of Canterbury as part of the PhD research project. It estimates about how well the average batting team will do against the average bowling team under given conditions and the current state of the game. In the first-innings it estimates the additional runs that can be scored with the given number of balls and wickets remaining. In the second innings it estimates the winning probability with the given number of balls and wickets remaining, runs scored at the given situation and the target given. The estimates have been made from a dynamic programming. Likewise, Raj and Padma analysed the Indian cricket team's ODI matches data and apply association rules on the attributes namely home or away game, toss, batting first or second and the match result. Swartz et al. use Markov Chain Monte Carlo methods to simulate ball by ball outcome of a match using a Bayesian Latent variable model. Depending on the ability of current batsman, bowler and game situation like number of balls delivered and number of wickets fallen, the outcome of the next ball had been predicted. But the model suffers from severe problems as noted by the authors themselves: the likelihood of a given batsman having previously faced a given bowler in previous games in the dataset is low.

3) Data Preprocessing

Here, is the data taken from Kaggle :

Dataset link: https://www.kaggle.com/patrickb1912/ipl-complete-dataset-20082020

The CSV File provided follows the "NEW" format.
The "new" format consists of a single row for each delivery in a match (or group of matches).
The first row of each CSV file contains the headers for the file, with each subsequent row providing details on a single delivery. The headers in the file are:

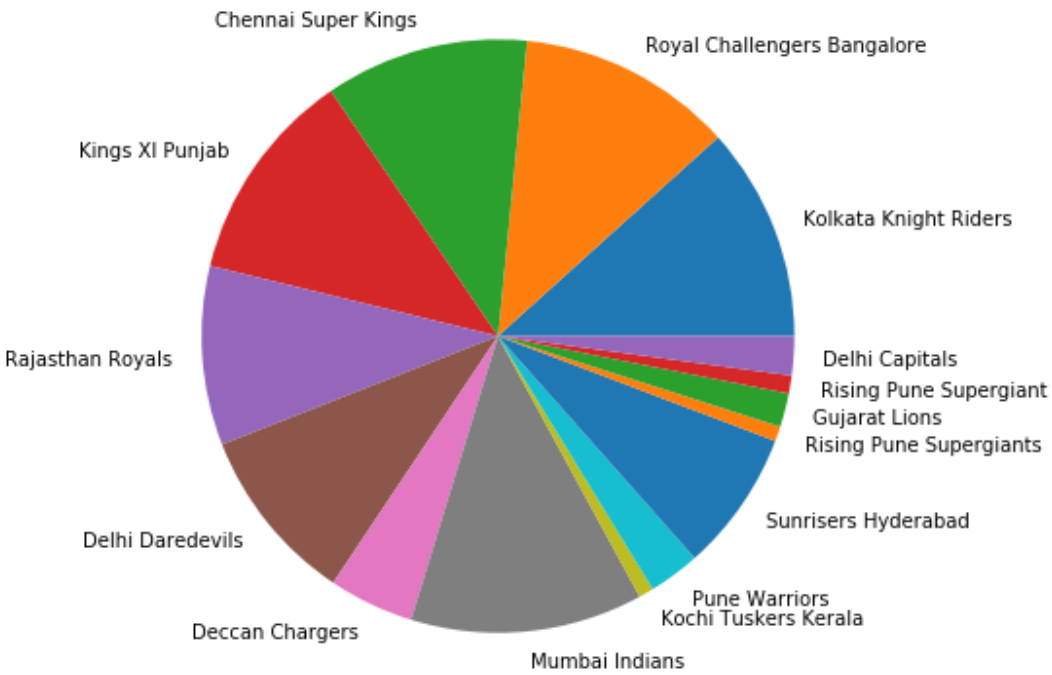
- match_id
- season
- start_date
- venue
- innings
- ball
- batting_team
- bowling_team
- striker
- non_striker
- bowler
- runs_off_bat
- extras
- wides
- noballs
- byes
- legbyes
- penalty
- wicket_type
- player_dismissed
- other_wicket_type
- Other_player_dismissed

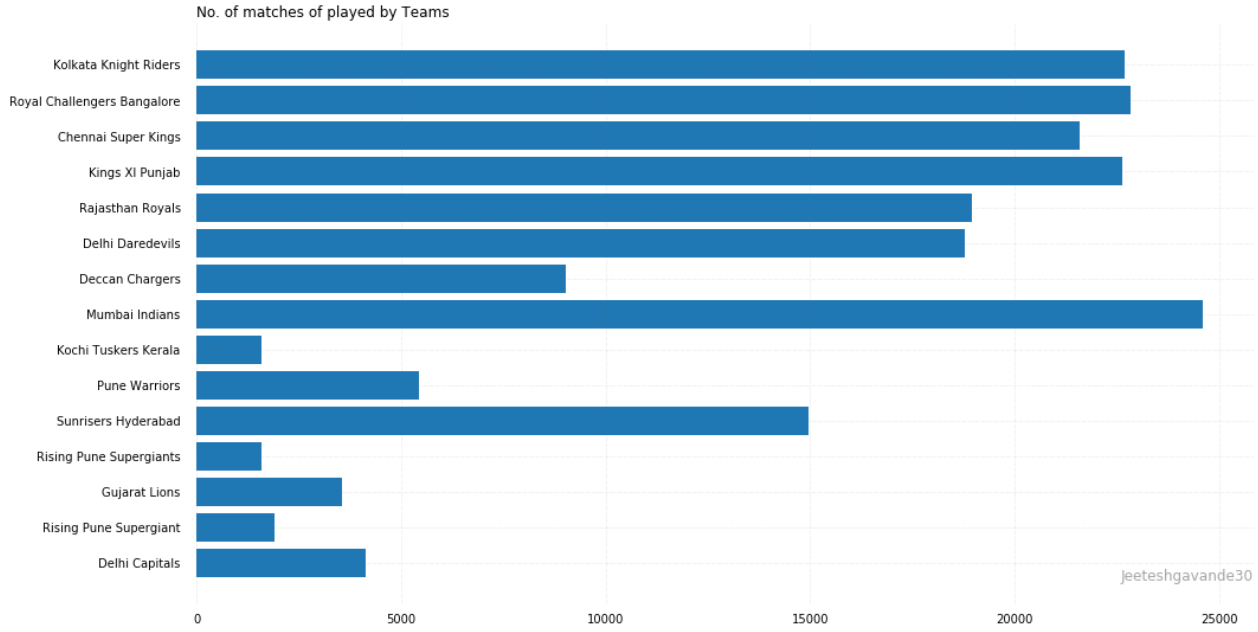
Most of the fields above should, hopefully, be self-explanatory, but some may benefit from clarification

- **"innings"** contains the number of the innings within the match. If a match is one that would normally have 2 innings, such as a T20 or ODI, then any innings of more than 2 can be regarded as a super over.
- **"ball"** is a combination of the over and delivery. For example, "0.3" represents the 3rd ball of the 1st over.
- If a wicket occurred on a delivery then **"wicket_type"** will contain the method of dismissal, while "player_dismissed" will indicate who was dismissed.
- There is also the, admittedly remote, possibility that a second dismissal can be recorded on the delivery (such as when a player retires on the same delivery as another dismissal occurs). In this case **"other_wicket_type"** will record the reason, while **"other_player_dismissed"** will show who was dismissed.

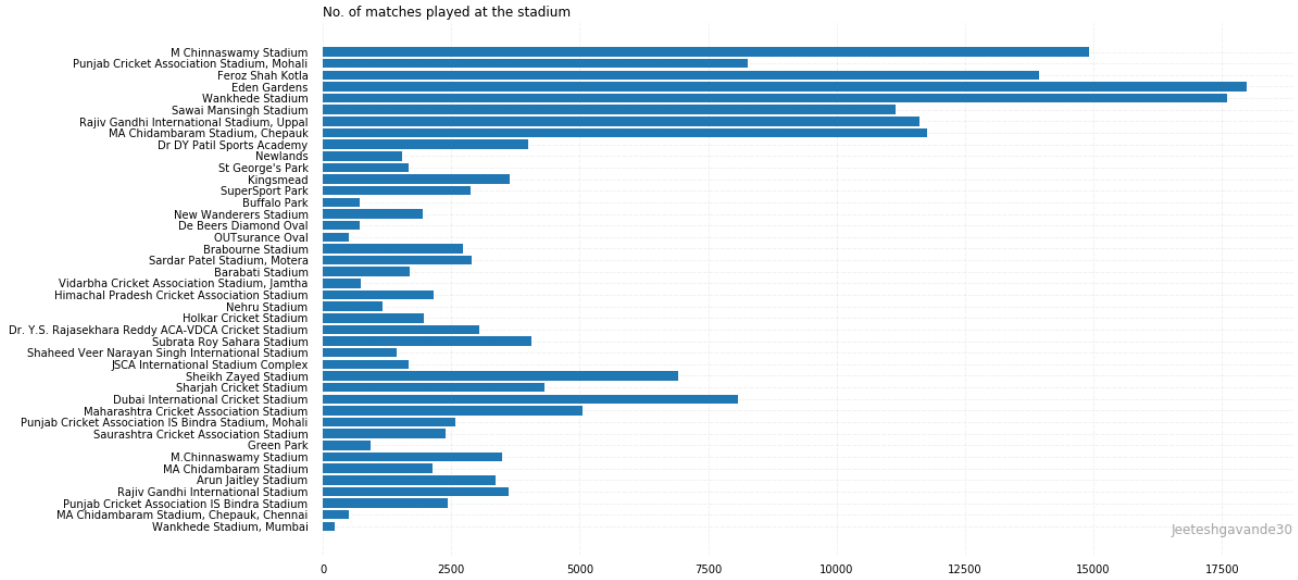
Data Visualization:

Teams vs No. of matches played.





No. of matches played at Particular venue.

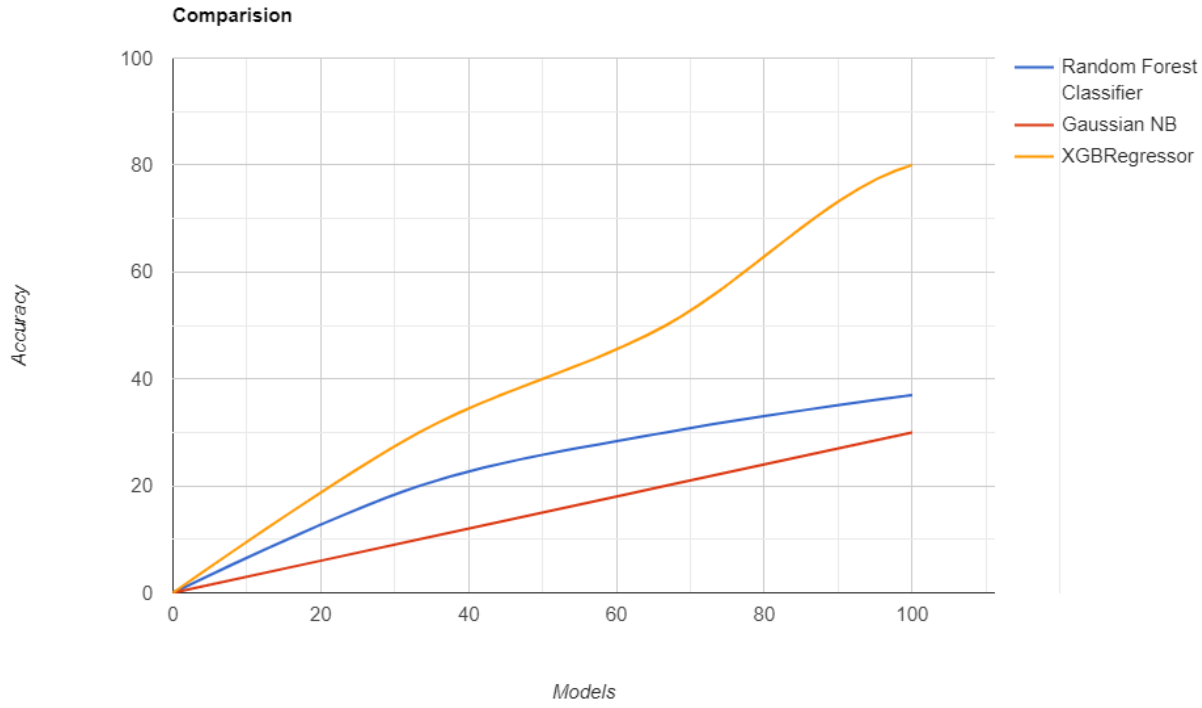


Firstly, Data consist of 1 to 20 overs, but we are only interested in powerplay so we have discarded the other overs. Then we have removed some unnecessary columns like ‘season','start_date','wides','noballs','byes','legbyes','penalty','wicket_type','player_dismissed','other_wicket_type','other_player_dismissed' for the better prediction. We have considered that total no. of batsmen that have batted whole powerplay may impact much than individual, same for the bowlers. and then also we have calculated the total runs and total wickets from the data and saved the new refine data into new csv file for the modeling.

4) Modeling

Since we are having much data, we have divided our data into 70% for the training and 30% for the testing purpose.

For the modeling purpose, we have used three different model Gaussian Naïve Bayes, Random Forest Classifier, XGBRegressor, From the three models, XGBResgressor has shown us the highest accuracy which is 92%. And other showed us not good results.



5) Conclusion

If we use the use the features like 'match_id', 'venue', 'innings', 'ball', 'batting_team','bowling_team','no_of_wickets' to predict power play score, XGBRegressor can show us good results, but it can be still improved by using last 2 overs runs or considering other parameters.

6) References :

Seamus Hogan (2012) Cricket and the Wasp: Shameless self promotion (Wonkish). http://offsettingbehaviour.blogspot.co.nz/2012/11/cricket-andwasp-shameless-self.htmlAccessed 2 March 2015.

K. Raj and P. Padma. Application of association rule mining: A case study on team India. In International Conference on Computer Communication and Informatics (ICCCI), pages 1{6, 2013.