IPL Match Winner Prediction Model and Statistical Analysis

Anand Kumar
Information Technology
NITK, Surathkal
anand.191it104@nitk.edu.in

Ashok Kumar
Information Technology
NITK, Surathkal
aashokkumar.191it210@nitk.edu.in

Vishal Kumar Verma
Information Technology
NITK, Surathkal
vishal.191it258@nitk.edu.in

***Abstract*—Now days, For every data analytic, Data Analysis is done with the help of ML models to examine the data set to extract the useful information from it and to find the conclusion according to the information. Data Analytic techniques and algorithms are mainly used by the commercial industries which provides them to take accurate business decisions. In the present date , the analytic is being used in the sports to predict various conclusion. Due to the involvement of money, team spirit, city loyalty and a massive fan following, the outcomes of matches are very important for all stake holders.**

In this project, We are taking the data from the 2008-2021 IPL matches which contains the match venue details, teams, players details, ball to ball details, and these data are analysed to find out the various conclusions which help in the improvement of a player's performance. There are many other features like how the venue or toss decision has influenced the winning of the match in previous years are also predicted. Different machine learning algorithms and data extraction model are used for the prediction like- XGBoost and Random Forest Classifier etc. The accuracy and the cross-validation score are also calculated using various machine learning algorithms. We must explore and visualise the data before prediction because data exploration and visualization are an important stage of a predictive model.

***Index Terms*—Cricket, Random Forest Classifier,Machine learning, Visualization,XGBoost,Ensemble learning.**

I. Introduction

We know that the Machine learning is a branch of Artificial Intelligence which aims solving real life problems and it provides the opportunity to learn without being explicitly programmed and it's based on the concept of learning from raw data. The advantages of ML methods are that it uses mathematical models, Heuristic learning, decision trees for decision making and knowledge acquisitions. Thus, it gives us a good idea how the data resembles and have different pattern in it.

ML in Sports and Cricket : ML is becoming quite a trend in sports analytic because of the ability of live as well as historical data. Decision making may be anything including which player to buy during an auction, which player will play for tomorrow's match, or something more strategic task like, building the tactics for upcoming matches based on players previous performance. ML can be used effectively in sports, both on-the-field and off-the-field. When it is on-the-field, ML applied to the analysis of a player's fitness level, design of offensive tactics, or decide short selection. It is also used in predicting the performance of a player, or a team, or the result of a match. On the other hand, the off-the-field scenario the business perspective of the sports, which include understanding sales pattern and assigning prices accordingly. The game of cricket is played in many formats One Day International, T20 and Test Matches.

ML in IPL: The Indian Premier League (IPL) is a T20 cricket tournament League introduced to promote cricket in India and thereby nurturing young and talented players. The League is an annual event which is like a festival for Indian people where teams name or on the name of Indian cities . The teams for IPL are selected by auction. Players auctions are not a new things in the sports world. However, in India, selection of a team of available players by the auction of players was done in Indian Premier League (IPL) for the first time. Due to the team spirit,involvement of money,a massive fan following,and city loyalty , the outcome of matches is very much important for all stakeholders. This, in turn, is dependent on the complex rule governing the game, the ability of a players and their performances on a given day and luck of the team (Toss). Many other factors, such as previous performance of the players play an important role in predicting the outcome of a cricket match. However, many factors are present that affect in predicting accurate results of a game. Moreover, the accuracy of prediction depends data size used . The model presented in the paper can find the performance of the different players. This model provides a overview of performance of the different players and a comprehensive details about almost all the important players and team's statistics which has been done using various libraries of python.

Various predictive models are also made for prediction of the result of a match, based on each player's past performance as well as some data related to match. The developed models can help people during IPL matches to find the strength of a team against another.

II. Literature Survey

There has been a lot of study related to Cricket matches in past years because of its so much popularity across the globe. And as we are dealing with IPL specific matches so related to that also several articles have been published by people some of which are mentioned here.

2004
GS Kantor and GDI Barr

They defined a criteria for selecting batsmen and comparing in limited overs cricket. They defined a new measure i.e. probability of getting out and used a 2-D graphical representation with batting average of the batsmen and Strike Rates . And, to the risk-return framework used in portfolio analysis, to obtain direct,useful, and comparative insights into batting performance, particularly in the context of the one-day game.

2016 ***Jhanwar and Paudi***

They are predicting the result of a cricket match by comparing the strength of the two teams. They analysed the performances of individual players of each team for this . They developed algorithms for the performances of bowlers and batsmen where they find the potential of player by analysing his past performances and then his current performances. Player independency factor have also been considered to predict the outcome of a match. They show that the k-Nearest Neighbour (KNN) algorithm gives better results compared to other classifiers.

2016
Lakshmi, Prakash, and Patvardhan

They present a Deep Mayo Predictor model for predicting the outcomes of the matches in IPL. Defined Bowling index and Batting index to rank performance of player for their models to predict IPL matches outcomes.

2018
Kalpdrum Passi and Nirav Kumar Pandey

They found various factors that affects the outcome of IPL. Prediction accuracy in terms of the number of wickets taken by the bowler in each team and runs scored by batsman . For predictive analytics, they used Dataiku and Weka . Both these tools are a collection of ML algorithms for data mining and provide some pre-processing functionalities.

"A Criterion for comparing in selecting Batsmen in limited overs cricket"

2019
Ayesha Choudhary and Rabindra Lamsal

Gives, a method to calculate the weightage of team on the basis of past performance of player in IPL using linear regression.They also calculate performance of player in upcoming matches by using linear regression .For calculating the points earned by each player based on their past performances a multivariate regression-based model was formulated. "Predicting Players' Performance in One Day International Cricket Matches Using Machine Learning"

III. Methodology

In this project we basically have taken the data up to 2021 matches which has the information of IPL match details and another one is about ball by ball information of the matches. Basically we've analysed the data from 2008-2020 dataset that was available and analysed it thoroughly to draw some inferences about the statistics of players, teams, and venues etc. And when coming to predicting the outcomes of the a match we have predicted the outcomes for 2021 mathces and compared the our results with the actual results of the mathces, this same thing can be further enhanced for the predicting the outcomes of upcoming matches with slight changes in the implementation of our model.

The methodology of our project mainly consists following stages: -

- Data reading and it's cleaning pre-processing
- Pre-processing the data.
- Doing statistical analysis of various factors mentioned above such as team's performance, player's performance etc.
- Drawing various inferences from it.
- Working for the prediction model
- Final Prediction of winners.

A pictorial representation of methodology can be shown as below :

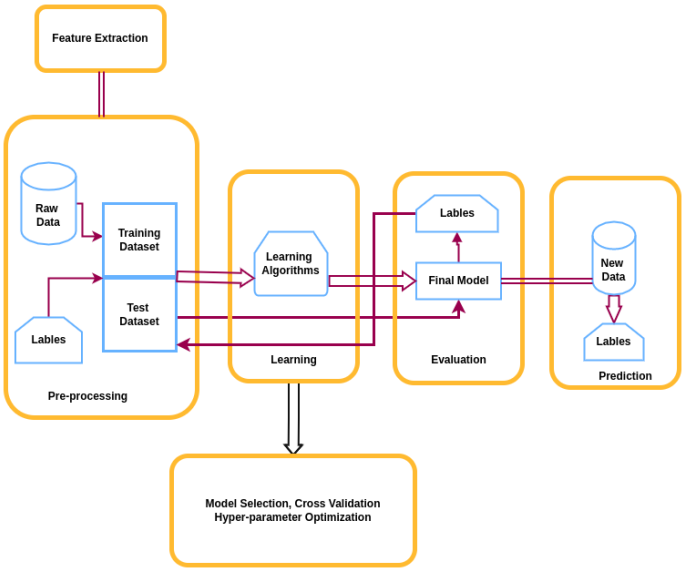

Fig. 1. Work flow

Some major steps are described below :

Data Pre-Processing

A real-world data generally contains missing values,noises, and maybe in an unusable format which cannot be directly used for ML models. Data Pre-Processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

Our Data-processing involves the following steps:

- Getting the data set till 2020.
- Manually noting down the match details of 2021 matches as it was not available.
- Checking for the missing values and filling them with appropriate values if possible.
- Encoding Categorical Data to make it suitable for the input of ML models

In Data Pre-processing phase, the data is inconsistent, noisy and incomplete. Checking for NULL values, we found:

city	13
player_of_match	4
winner	4
result	4
result_margin	17
eliminator	4
method	797

Fig. 2. Null Values

After inspecting the dataset we found that the columns winner, result, player_of_match has 4 NaNs.

We found that the matches that were tied due to rains have nans in these columns. We droped the NaNs rows as it will not affect our analysis.

To correct the inconsistencies data is to be filled with missing values. In Data Cleansing phase, by maintaining consistency across the data set ,validation of data is done and by adding related information to the dataset, enhancement of data is done. To achieve optimal results the data preparation is significant. This involves choosing an outcome measure to evaluate different predictor variables. In data Encoding phase, label each term with short names and encode them as numerical values for predictive modeling as implemented below.

Performing the statistical analysis

Data Visualization is used to perform Exploratory Data Analysis (EDA). Here we are dealing with large numbers of data, building graphs is the best way to explore and communicate the findings. Visualization is very much helpful for us to identify the patterns and trends in our data, which leads to clearer understanding regarding the model and it also reveals important insights.

The points on which we are visualising the data:

- How many matches are played each year in IPL ?
- How many matches did the teams played throughout the IPL(2008-2020)?
- Does winning the toss affects the outcome of a match for a team?
- Which team has the highest win percentage? in general and based on winning the toss batting first, winning toss and fielding first etc.
- Which batsman hit the most number of 6s,4s throughout the IPL till now.
- Which bowler gave the most number of 6s,4s and also who is the highest wickets taker till now?
- Who bowled highest number of maiden overs who has the highest strike rate etc.
- Who has won the most number of Man Of the Matches.
- Which player scored the most number of centuries?
- And so many others...

Following are the figures depicting the statistical analysis

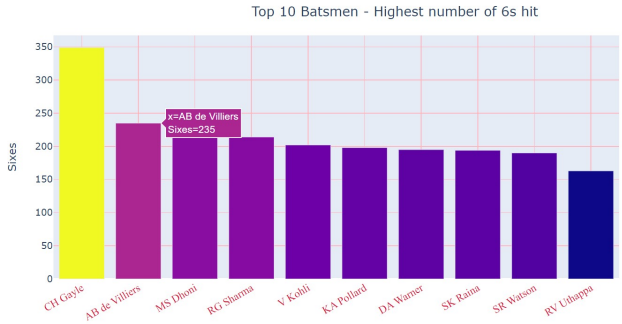

Fig. 3. Highest Sixes

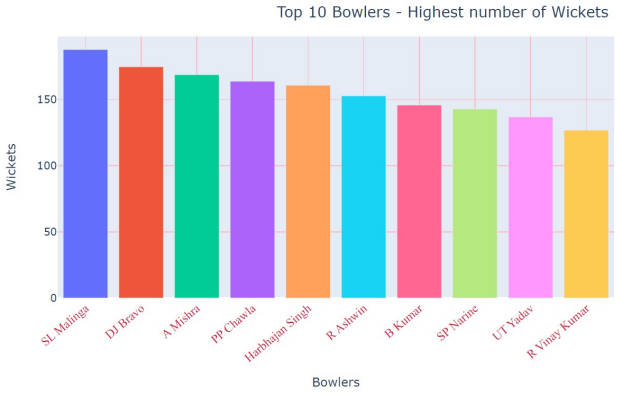

Fig. 4. Highest Wickets

Fig. 5. Player of the Match

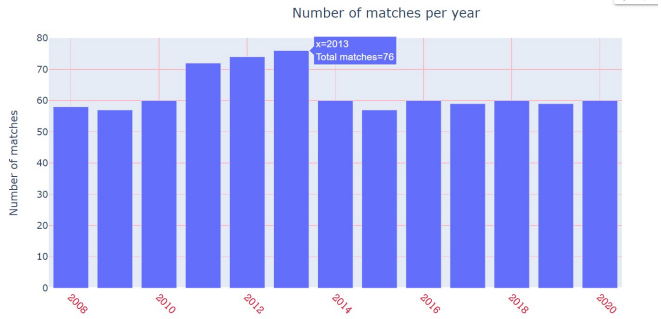

Fig. 6. Mathces per year

For making the visualisation much more attractive we've used plotly library of python which provides easier ways to make interactive bar plots, count plots, etc.

Inferences from the Statistical analysis and making predictions

After doing the complete statistical analysis of the data that we have used it becomes very much easy for us to draw inferences from it.

Like people have myth in mind that winning the toss have a huge impact on the winning probability of the match etc, but in our analysis we found that there are teams which have won more number of matches when they lost the toss as compared to when they won the toss.

Then coming to team's analysis we have shown for any particular team winning the toss and choosing to field first affects its winning probability and vice versa, also which team has played most number of matches till now. How many times they have won the IPL title etc.

Then coming the stats of the player's like who has taken the highest number of wickets, who has hit the most number of 4s and 6s till now, who have the record of highest number of centuries in the IPL, which bowler conceded highest 6s,4s who bowled most number of maiden overs etc.

Then the last part is working with various classifiers to predict the winner of the matches of 2021 IPL matches based on the model that we trained using the data up to 2020.

IV. Results

In the result part we have analysed various aspects of the dat set as mentioned above and drawn various inferences from it.

And then coming to predicting the winners for the IPL matches of 2021 matches we have used the data of 2008-2020 to train our classifier model and used them to make predictions on the data set.

We have considered various classifiers for predicting the outcomes and based on their performaces we moved to some good modern classifiers, classifiers used are listed below :

- SVM
- KNN
- Decision Tree
- Random Forest
- Ensemble classifier XGBOOST

Above classifiers are listed in increasing order of the accuracy scored obtained which like using svm we obtained we received 30% , with KNN it increased to 35% with Decision tree 49% with Random classifier we obtained 56% then comes the ensemble classifier XGboost which gave accuracy of 64% then we performed hyper parameter tuning of it which result in accuracy of around 87-90% with best parameters being considered.

And then we saved our prediction in a csv file which can be referred for later uses.

Various figures shows our results.

Information related to prediction model

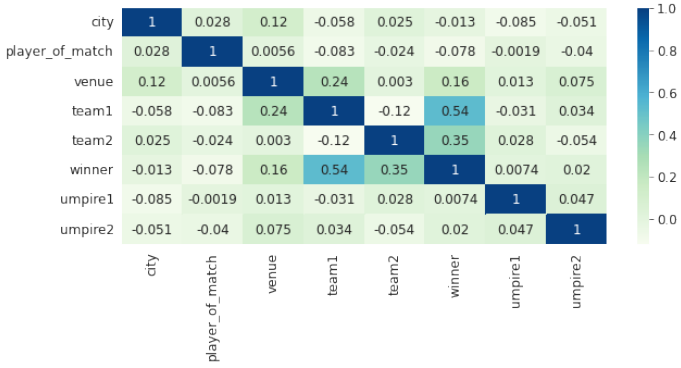

Fig. 7. Correlation between different features

```
from sklearn.tree import DecisionTreeClassifier

clf = DecisionTreeClassifier()
clf = clf.fit(X_train,y_train)

y_pred = clf.predict(X_test)


# Model Accuracy, how often is the classifier correct?
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.49056603773584906
```

```
rfc2 = RandomForestClassifier(max_depth=4)
rfc.fit(X_train,y_train)
rfc.score(X_test,y_test)
```

```
0.5518867924528302
```

Fig. 8. Decision tree and random forest

```
model =XGBClassifier(n_estimators = 100,colsample_bytree=0.3,
                     learning_rate=0.3,gamma=0.2,max_depth=7)
kfold = StratifiedKFold(n_splits=10, random_state=7,shuffle=True)
results = cross_val_score(model, X_train, y_train, cv=kfold)


print("Average score: %.2f%% (%.2f%%)" % (results.mean()*100, results.std()*100))
```

```
Average score: 77.77% (3.69%)
```

Fig. 9. XGB with Kfold

```
xbg = XGBClassifier(n_estimators=100,colsample_bytree=0.3,
                    learning_rate=0.3,gamma=0.2,max_depth=3)
xbg.fit(train_x,train_y)
```

```
XGBClassifier(colsample_bytree=0.3, gamma=0.2, learning_rate=0.3,
              objective='multi:softprob')
```

```
xbg.score(X_test,y_test)
```

```
0.9056603773584906
```

Fig. 10. XGB with hyper paramter tuning

V. Conclusion

In this project we did a comprehensive study about the IPL data right from the beginning till 2021 and saw various stats related to teams, players, various factors which may contribute to the winning probability of a match and also we explored various ML techniques for classification purpose, and learnt about some modern ensemble classifiers also.

So mainly this project is all about doing the statistical analysis of IPL matches and seeing the various information related to teams, players, venues, toss decisions etc and then based on all such features working with various classifiers for predicting the winners of IPL matches.

References

[1] Rameshwari Lokhande and P.M. Chawan, "Live Cricket Score and Winning Prediction", International Journal of Trend in Research and Development, Volume 5(1), ISSN: 2394-9333

[2] Abhishek Naiket. Al, "Winning Prediction Analysis in One-Day-International (ODI) Cricket Using Machine Learning Techniques", IJETCS, vol. 3, issue 2, ISSN:2455-9954, April 2018

[3] Jhanwar and Paudi, "Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach" 2016

[4] Rabindra Lamsal and AyeshaChoudhary, "Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning".

[5] Lakshmi, Prakash, and Patvardhan, "Batting index and Bowling index to rank player's performance for their models to predict outcomes of IPL matches".

[6] Passi, Kalpdrum Pandey, Niravkumar. (2018) "Predicting Players' Performance in One Day International Cricket Matches Using Machine Learning" 111-126. 10.5121/csit.2018.80310.

[7] Abhishek Naiket. Al, "Winning Prediction Analysis in One-Day-International (ODI) Cricket Using Machine Learning Techniques", IJETCS, vol. 3, issue 2, ISSN:2455-9954, April 2018.

[8] Ujwal U J et. At, "Predictive Analysis of Sports Data using Google Prediction API" International Journal of Applied Engineering Research", ISSN 0973-4562 Volume 13, Number 5 (2018) pp. 2814-2816.