IPL Predictor

by Aniket Anand

IPL Predictor

ANAND SINGHANIA
Department of Computer Science and Engineering
PES University
Bengaluru-85, India
asinghania71@gmail.com

ANIKET ANAND
Department of Computer Science and Engineering
PES University
Bengaluru-85, India
aniket.anand1304@gmail.com

HARSH CHOUDHARY
Department of Computer Science and Engineering
PES University
Bengaluru-85, India
choudhary.harsh6@gmail.com

***Abstract*— The Paper presents a Random Forest Model to predict the outcome of the Indian Premier League, 2019. The model uses data scrapped from various sources. The data is cleaned and processed. The model is trained on the last ten seasons of the Indian Premier League data and is tested on the eleventh season. The Model is based on the methods from Machine Learning. The prediction is as good as forty two matches out of fifty nine matches.**

Keywords—Cricket, Prediction Model, Machine Learning (Random Forest), Data Analytics.

I. INTRODUCTION

India is the second largest country in terms of population with more than 1.33 billion people. Cricket is considered more than a sport in this country of a billion people. India is one of the most religiously and ethically diverse nations in the world. In a country with such diversities, Cricket is that one thing which unites people of different religion, caste and culture. People in India are captured by the fever of cricket making it the most loved sport of the country.

There are various forms of cricket played across the world. ODI (One-Day International), Test Matches, and the latest Twenty-20. The IPL (Indian Premier League) is a Tewnty-20 cricket tournament league. The main aim of this league was to promote cricket in India and to discover young talents from across all the districts of India. This league was started in 2008 and has been taking place every year since then.

The Indian Premier league is governed by the Board of Control for Cricket in India (B.C.C.I). When the IPL was launched, the companies started pouring in millions because they were aware of the instant returns. Players were bought for huge amount of money and lot of film stars and Corporate showed their interest in owning a team of their own and rest is History.

Indian Premier League (IPL) is a very famous league across the world. Last season had record breaking views. More than 130 streamed the league through their digital devices and 410 million people watched it directly on television. So it was very interesting for us to work on the data of a sport with such a huge fan base. It was very difficult for us to predict the outcome of such an unpredictable sport. Statistics and Machine learning are the two most important building blocks of approach.

Indian Premier League has mostly eight teams competing with each other. The first stage is league based and the teams with highest score qualify for the playoffs. The top four teams make their way to the playoffs. The top two teams compete with each other and the winning team is directly selected for the finals. The next two teams compete with each other and the winner competes with the team in the second position and thus the second finalist is selected. The two finalists compete with each other and the winning team creates history and is declared as the Champion of that particular season.

Every season around sixty matches take place and our aim was to predict at least forty matches correctly. Various Machine Learning algorithms were tested and one with the best results was used to deal with the problems. The accuracy percentage varied vastly for different algorithms and Random Forest Classifier was proved to be the best suited algorithm for our dataset. Decision Tree Classifier, Support Vector Machine (Support Vector Classifier) and Naive Bayes were the few rejected algorithms.

The data of the last ten seasons were collected. The collection of the data wasn't a big deal but cleaning of the data set was one of the important challenges that came on our way. Data was scrapped from various sources and thus cleaning and matching of data was an important task.

Data processing was the next task on our way. Ball by ball data of the last ten seasons was used to process the data figure out number of matches played by a player till now. Total Number of runs scored by each player. Based on the statistics, we grouped the players into bowlers and batsman. Player vs. Player comparisons were made so figure out the performance of a player against every other player present. This was an important step because with this we were able to calculate the relative strength of the teams.

The next task was to predict playing eleven of each team. Player performance was used as a major factor for this prediction. Player performance was calculated using ball by ball data. Team consisted of at least five top bowlers according to the number of matches played in the last two seasons.

Training of the dataset was the next step. Random Forest Classifier was used for data training. Data of last 10 seasons was trained and finally predictions were made. Prediction of various matches were recorded and analyzed to predict the accuracy of our system.

A simple UI was designed to display the results and to make the system user friendly.

II. DESIGNING DATASET

A. Feature Selection

Data was scrapped from various sources and the important features were noted down. The main features selected were Venue, the two playing teams, toss winner, toss decision. Playing eleven of the two playing teams were predicted and based on that the relative Eigen values were calculated which was used as an important feature in prediction. Number of important features was not large and thus features reduction using Principal Component Analysis was avoided. Feature selection was an important step as it affected the prediction largely. We are considering every ball ever bowled in the history of Indian Premier League till 2018. This increases the size of the dataset largely. We had to deal with more than 15000 balls. We considered where each ball was bowled (Venue), how much runs were given on that particular ball, wicket taken on that particular ball, who was on strike on that particular ball and who was on the other end on that particular ball.

B. Organizing the Dataset

Prediction for the IPL requires a good organized dataset. In this tournament, firstly each team plays with all the other teams twice, once in each team's home stadium. For an instance, if Royal Challengers Bangalore played against Chennai Super Kings in Chinnaswamy Stadium, then next game between Royal Challengers Bangalore and Chennai Super Kings must take place in M.A. Chidambaram Stadium, Chennai. Data was studied and organized accordingly. Organization of data was important to make sure that the feature 'Venue' is used impeccably.

C. Problem with the Dataset

Dataset is the most important element of this project. This project is totally based on the quality of data. The stronger the dataset, stronger is the level of prediction. We are mostly focussing on ball to ball data. This leads to a lot of problems. We are considering every ball ever bowled in the history of Indian Premier League till 2018. This increases the size of the dataset largely. We had to deal with more than 15000 balls. We considered where each ball was bowled (Venue), how much runs were given on that particular ball, wicket taken on that particular ball, who was on strike on that particular ball and who was on the other end on that particular ball. Dealing with such an amount of data was a difficult task and it had to be dealt very carefully in order to achieve maximum accuracy of prediction. In such a huge amount of data, it's impossible to have complete values for every field. Hence missing data values had to be taken care of. Missing values were replaced with appropriate values referred from the web or NULL values. Inconsistent naming was another problem in the dataset. Players and teams had different spellings in different fields of years. This problem was rectified manually with utmost attention. Data had abbreviations which made the study difficult and complicated. The abbreviations were replaced by full values in order to avoid complexity in future use.

Serial Number	Machine Learning Algorithm used for Training	Accuracy (%)
1.	Support Vector Machine (SVC)	41.791045
2.	Decision Tree Classifier	41.791045
3.	Naive Bayes	20.895522
4.	Support Vector Machine (Linear)	25.373134
5.	Random Forest	52.238806

III. Machine Learning Models

Several Machine Learning algorithms were tested and accuracy of the respective models was studied. The best suited algorithm was chosen to deal with the future problems.

Algorithms such as Decision Tree Classifier, Support Vector Machine (Support Vector Classifier) and Naive Bayes were tested. The accuracy percentage varied vastly for different algorithms. Table showing accuracy for various algorithms is mentioned in the table given

.

A. Player Points

Player points are calculated according to the runs scored against the bowlers of the opposite team. This means that player points are calculated for every game with respect to the runs scored by each player against each bowler of the different team. This favours the accuracy as we consider a players performance against the respective opponent team. A player can play with different forms against different teams depending on the type of bowlers present in that team. This is one important feature to increase the accuracy of prediction. This player points is further used to calculate the Eigen values which is an important feature for prediction.

B. Team Weightage

Player points of all the players in the current teams are used to calculate the team weightage. Since player points are calculated for every match with respect to the bowlers of the opponent team, team weightage is also calculated for every match separately depending upon the team they are playing with. This helps us to get a greater accuracy. Team weightage is calculated my summing up player points of every individual player in a team by total number of players playing. Team weightage is an important factor as it corresponds to the Eigen values.

C. Random Forest Classifier

Random Forest Algorithm is unexcelled in accuracy among any other algorithms being used. Different algorithm is suited for different data problems. In this case (and in most) Random Forest Algorithm gives the best result. Random Forest Classifier runs best for such large dataset. It can handle multiple variables without variable deletion. It gives you an estimate of what variables are important for classification. It generates an internal unbiased estimate of the generalization error as the forest building progresses. It has an effective method for estimating missing data and

maintains accuracy when a large proportion of the data are missing. It has method for balancing error in class population unbalanced datasets. In this generated forest can be saved for future use. Prototypes are computed that gives information about the relation between the variables and the classification. It computes proximities between pairs of cases that can be used in clustering, locating outliers or (by scaling) gives interesting views of the data. The capabilities of the above can be extended to unlabelled data, leading to unsupervised clustering, data views and outlier detection. It offers an experimental method for detecting variable interactions. Random Forest Classifier gave an accuracy of 52.24% initially. The accuracy was however improved and extended to 71.1%. The prediction with this percentage gave satisfactory result.

IV. Features Weightage

Different features have different weightage in this model. The weightage is calculated for various features using Random Forest Classifier.

We can see that the Eigen Values of the two playing teams have the most weightage in our model. Venue is the second most important feature according to our model. Toss decision and toss winner are the two least important feature according to our model.

Serial Number	Feature	Weights
1.	City	0.202
2.	Toss Winner	0.055
3.	Toss Decision	0.059
4.	Team1 Eigen Value	0.337
5.	Team2 Eigen Value	0.345

V. Result

The final model was made and it was tested on Indian Premier League 2019. The result was satisfactory. We achieved our goal of predicting at least forty matches correctly. However, the greater the better, we are still trying to make small amendments in order to get better results. The tabular representation of the result is given below for reference. The table shows match number as column one, the next column is for Home team, next column is for the Away team, next column stands for the actual winner of the game as recorded from the official webpage of the Indian Premier League and last column shows the predicted result of the match according to the developed model.

From the data we can see that we are able to predict forty two matches correctly out of fifty nine matches played. Thus we can see that this model is predicting 71.2% matches correctly. The previous referred models were able to predict at max 39 matches correctly. So the results prove that this model is better than a lot of available models. This is good achievement but getting higher accuracy is even better.

Match No		IPL 2019		
	Team1	*Team2*	*Winner*	*Prediction*
0	**RCB**	**CSK**	**CSK**	**CSK**
1	**SRH**	**KKR**	**KKR**	**KKR**
2	**DC**	**MI**	**DC**	**DC**
3	**KXIP**	**RR**	**KXIP**	**KXIP**
4	**DC**	**CSK**	**CSK**	**CSK**
5	**KKR**	**KXIP**	**KKR**	**KKR**
6	**MI**	**RCB**	**MI**	**RCB**
7	**RR**	**SRH**	**SRH**	**RR**
8	**MI**	**KXIP**	**KXIP**	**KXIP**
9	**KKR**	**DC**	**DC**	**DC**
10	**SRH**	**RCB**	**SRH**	**SRH**
11	**CSK**	**RR**	**CSK**	**CSK**
12	**KXIP**	**DC**	**KXIP**	**KXIP**
13	**RCB**	**RR**	**RR**	**RCB**
14	**MI**	**CSK**	**MI**	**CSK**
15	**DC**	**SRH**	**SRH**	**SRH**
16	**RCB**	**KKR**	**KKR**	**KKR**
17	**CSK**	**KXIP**	**CSK**	**CSK**
18	**MI**	**SRH**	**MI**	**MI**
19	**RCB**	**DC**	**DC**	**DC**
20	**RR**	**KKR**	**KKR**	**KKR**
21	**SRH**	**KXIP**	**KXIP**	**SRH**
22	**KKR**	**CSK**	**CSK**	**CSK**
23	**KXIP**	**MI**	**MI**	**MI**
24	**RR**	**CSK**	**CSK**	**CSK**
25	**KKR**	**DC**	**DC**	**DC**
26	**MI**	**RR**	**RR**	**RR**
27	**KXIP**	**RCB**	**RCB**	**RCB**
28	**KKR**	**CSK**	**CSK**	**CSK**
29	**DC**	**SRH**	**DC**	**SRH**

Match No		IPL 2019		
	Team1	Team2	Winner	Prediction
30	RCB	MI	MI	MI
31	KXIP	RR	KXIP	KXIP
32	CSK	SRH	SRH	SRH
33	MI	DC	MI	DC
34	RCB	KKR	RCB	KKR
35	MI	RR	RR	MI
36	KXIP	DC	DC	DC
37	KKR	SRH	SRH	KKR
38	RCB	CSK	RCB	CSK
39	RR	DC	DC	RR
40	SRH	CSK	CSK	CSK
41	RCB	KXIP	RCB	RCB
42	KKR	RR	RR	RR
43	MI	CSK	MI	CSK
44	SRH	RR	RR	SRH
45	DC	RCB	DC	DC
46	KKR	MI	KKR	KKR
47	SRH	KXIP	SRH	SRH
48	CSK	DC	CSK	CSK
49	MI	SRH	MI	MI
50	KXIP	KKR	KKR	KKR
51	RR	DC	DC	RR
52	SRH	RCB	RCB	RCB
53	CSK	KXIP	KXIP	CSK
54	KKR	MI	MI	KKR
55	CSK	MI	MI	MI
56	SRH	DC	DC	DC
57	DC	CSK	CSK	CSK
58	MI	CSK	MI	MI

REFERENCES

Data Analytics based Deep Mayo Predictor for IPL

(Source:-International Journal of Computer Applications (0975 – 8887) Volume 152 – No.6, October 2016)

- C. Deep Prakash Dayalbagh Educational Institute Agra
- C. Patvardhan Dayalbagh Educational Institute Agra
- C. Vasantha Lakshmi Dayalbagh Educational Institute Agra

Predicting Outcome of Indian Premier League (IPL) Matches Using Classification Based Machine Learning Algorithm

- Rabindra Lamsal, Ayesha Choudhary
- School of Computer & Systems Sciences, Jawaharlal Nehru University

IPL Predictor

ORIGINALITY REPORT

9%	7%	6%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	surfertas.github.io Internet Source	5%
2	ijarcsse.com Internet Source	1%
3	Submitted to Oklahoma State University Student Paper	1%
4	www.iig.asia Internet Source	<1%
5	icds19.jinnah.edu Internet Source	<1%
6	Submitted to Symbiosis International University Student Paper	<1%
7	K.S Daya, N.D Kataria, V.G Das, G.S Tyagi, G.P Srivastava. "Phenomenological model for the anomalous peaks in the microwave surface resistance of high temperature superconductors", Physica C: Superconductivity, 2002 Publication	<1%

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		