National Taipei University of Technology
Computer Science and Information Engineering

Principles and Applications of Data Science

Spring 2020

Semester Group Project Report

INDIAN PREMIER LEAGUE SCORE & WIN PREDICTOR

Name: **B MUKESH KUMAR, M VINEELASWATHI,**

SID: **108998405, 108998406**

Date:**07/01/2020**

ABSTRACT

Indian Premier League is a T20 League which was started in 2008 and now became the most awaited T20 cricket carnival. Since the IPL has large popularity, predicting the results of it is really important and to be more effective. The Solution of predicting the results can be done with the help of Time Series Analysis and the Machine Learning Algorithms and Techniques which reduce the Domain Knowledge. Data Analysis has to be done by taking the historical data and need to draw some conclusions by applying Machine Learning Techniques. The solution of predicting the match must be effective since, there is a lot enthusiasm for IPL seasons and winners of that Season. In this particular project the parameters like Venue of the match, Win or Loss of the Toss, ball to ball details, Batsman Strike Rate were taken in to consideration for which the machine learning techniques were applied and the results are predicted. The Data Sets of past 7 years are taken with the above parameters and preprocessing is done for the data. The Machine Learning Algorithms that we used in here are Random Forest and Logistic Regression for predicting the accurate results. Before predicting, we explored the data and analyzed it.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION TO IPL

IPL Stands for Indian Premier League, it belongs to the Cricket Community. The cricket game has different type of forms such as Test matches, Twenty20 Internationals, one day internationals etc. Among them IPL is also one of them and has major popularity. It is a Twenty-20 cricket competition league which is played in India for encouraging the young and dynamic players. The League was conducted every year of the month March, April or May and has a huge fan base among India. There are eight teams which represent eight cities which are chosen from an auction. These teams compete against each other for the trophy. The whole match depends on the luck for the team, players performance and lot more parameters that will be taken in to the consideration. The match that is played before the day is also will make a change in the prediction. The stakeholders are much more benefited due to the huge popularity and the huge presence of people at the venue.

1.2 DATASET

The matches about 636 records were taken into the consideration and applied to the models that are fit in it properly. The noisy data and null values is removed and the data preprocessing is done to the data and by training the models. The 80% of the data is taken for training set and models are trained with that dataset and the remaining data is used for validating of the models. The accuracy is one parameter where we can check whether the model has given proper prediction results or not. We are predicting the results (score and Win team) by testing the models with the 2019 dataset about 59 matches were taken.

1.3 REPORT DEALS WITH

In this project, we predict the teams' score and Match Winner in Indian Premiere League (IPL) matches by analyzing their characteristics and stats using supervised machine learning techniques. For this, we predict teams score based on innings performance separately as how many runs will a team score and how many runs the other team can take in a particular match.

CHAPTER 2

LITERATURE REVIEW AND REALTED WORKS

2.1 PLAYER PERFORMANCE PREDICTION

The Work that done in “Predicting Players' Performance in One Day International Cricket Matches Using Machine Learning” is predicting the winner using HBase which is later applied to the machine learning techniques. This tool is helpful for managing and selecting the players in next auction for a team.

https://airccj.org/CSCP/vol8/csit88310.pdf

2.2 STRIKE RATE PREDICTION

In the paper "Predicting the performance of batsmen in test cricket," the overall weight of a team is considered by taking each player performance. Six types of machine learning models were trained accordingly and used for predicting the outcome. Among them Random forest has given the highest accuracy.

https://www.jhse.ua.es/article/view/2014-v9-n4-predicting-the-performance-of-batsmen-in-test-cricket

2.3 NBA GAMES OUTCOME PREDICTION

Predicting the outcome of NBA games based on logistic regression – motivated by this work and got an idea of predicting the score and winner of IPL and the source code is given here

https://github.com/smidtfab/nba_score_analyzer

CHAPTER 3

PROBLEM STATEMENT

3.1 PROBLEM STATEMENT

Predicting the outcome of Indian Premier League (IPL) matches poses a challenging problem of interest to the research community as well as the general public. In this article, we formalize the problem of predicting IPL game results as a classification problem and apply the principle of Maximum Entropy to construct a model that fits to discrete statistics for IPL games, and then predict the outcomes of IPL playoffs using the model. Our results reveal that the model is able to predict the winning team with ___ accuracy, outperforming other classical machine learning algorithms that could only afford a maximum prediction accuracy of 70.6% in the experiments that we performed.

CHAPTER 4

DATA PROCESSING

.

4.1 DATA PRE-PROCESSING

4.1.1 Data Collection

The Data from the past ten years (2008-2017) was taken for the Analysis and the Variables are selected from the Data. Pandas Library is used for the transformation of the data in to the required form. It is an open source library that is used for analyzing and manipulating the data. About 636 matches were taken in to the consideration for analyzing and predicting the outcome.

4.1.2 Calculating Traditional Attributes

As mentioned in the previous section, the stats of the players such as average, strike rate etc. are not available directly for each game, we calculated these attributes from the innings by innings list using aggregate functions and mathematical formulae. These attributes are generally used to measure a player's performance. These attributes are as follows:

4.1.3 Batting Attributes

No. of Innings: The number of innings in which the batsman has batted till the day of the match. This attribute signifies the experience of the batsman. The more innings the batsman has played the more experienced the player is.

Batting Average: Batting average commonly referred to as average is the average number of runs scored per innings. This attribute indicates the run scoring capability of the player.

Average = Runs Scored / Number of times dismissed

Strike Rate (SR): Strike rate is the average number of runs scored per 100 balls faced. This attribute indicates how quickly the batsman can score runs.

*Strike Rate: (Runs Scored / Balls Faced) * 100*

4.1.4 Opposition

Opposition describes a player's performance against a particular team. All the traditional attributes used in this formula are calculated over all the matches played by the player against the opposition team in his entire career till the day of the match.

Formula for batting:

$$Opposition = 0.4262*average + 0.2566*no.\ of\ innings + 0.1510*SR + 0.0787*Centuries + 0.0556*Fifties - 0.0328*Zeros$$

Formula for bowling:

$$Opposition = 0.3177*no.\ of\ overs + 0.3177*no.\ of\ innings + 0.1933*SR + 0.1465*average + 0.0943*FF$$

4.1.5 Venue

Venue describes a player's performance at a particular venue. All the traditional attributes used in this formula are calculated over all the matches played by the player at the venue in his entire career till the day of the match.

Formula for batting:

$$Venue = 0.4262*average + 0.2566*no.\ of\ innings + 0.1510*SR + 0.0787*Centuries + 0.0556*Fifties + 0.0328*HS$$

Formula for bowling:

$$Venue = 0.3018*no.\ of\ overs + 0.2783*no.\ of\ innings + 0.1836*SR + 0.1391*average + 0.0972*FF$$

4.2 DATA CLEANING

A large number of values of Opposition and Venue were zero. This is because a player has not played any match against a particular team or at a venue before the day of

play. We treated such values as missing values and replaced them with the class average of corresponding attributes.

4.3 DATA VISUALIZATION

4.3.1 Venue Analysis

The data which has been collected is used for visualizing for the better understanding of the information. Matplotlib Library is used here for visualizing the graphs for team wins in different cities according to their venues and the player strike rate from the seasons 2008 to 2017. The graph for the team wins according to their respective cities in 2016 and the strike rates were shown in the fig 1

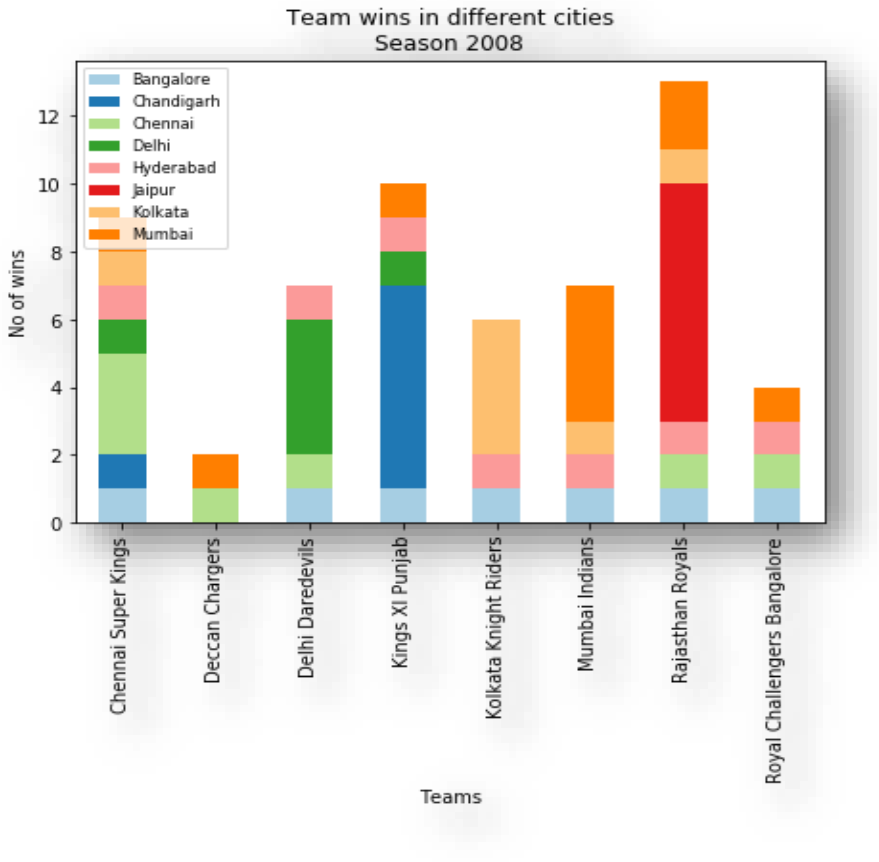


Figure 1: Team wins in different Cities

4.3.2 Strike Rate Analysis

The visualization for the players strikes rate from the past seasons were shown below in the graph. The strike rate of a player also plays a major role in the prediction

so, that in the next season the prediction becomes obvious if the player with high strike rate will been in any team and that team has a huge winning chances. The strike rate was shown in the below fig 2

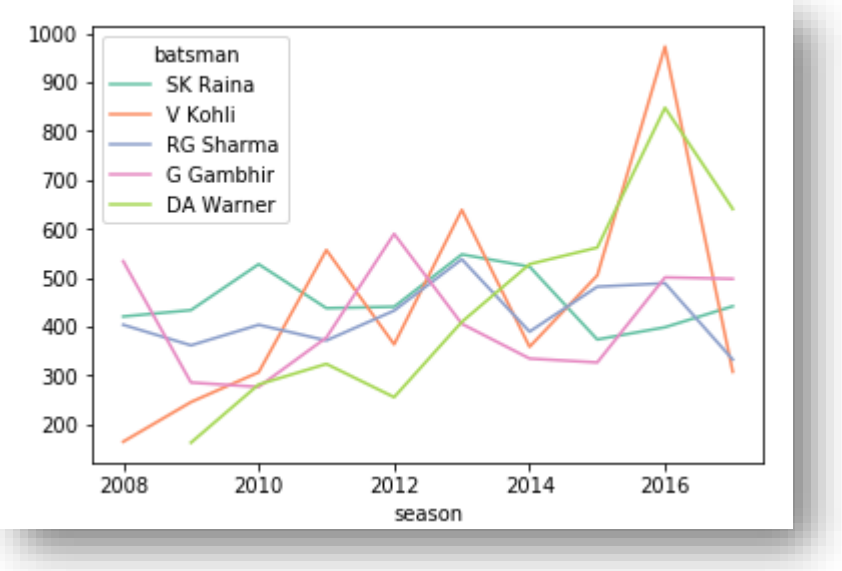


Figure 2: Strike Rate of Batsmen

4.3.3 Toss Analysis

Winning toss plays a key role in the match, in order to draw a clear picture on toss analysis has been made on individual years, fig 3 below shows the win and loss by a team with respect to toss.

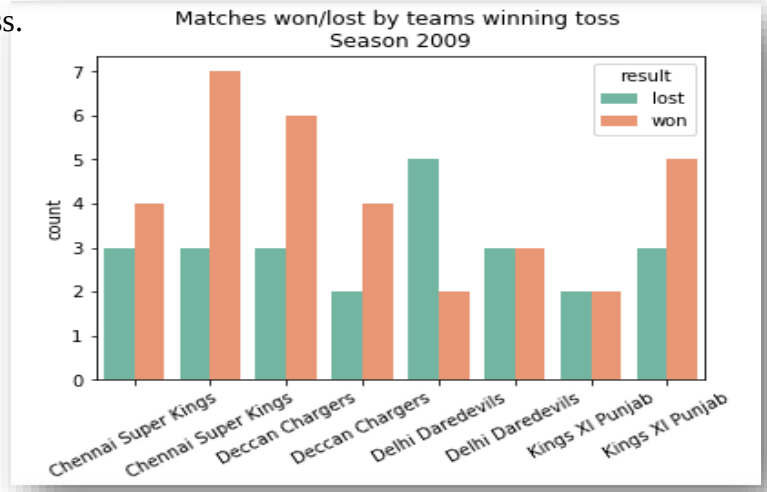


Figure 3: Match Result with respect to Toss

CHAPTER 5

PROPOSED MODEL

For generating the prediction models, we used supervised machine learning algorithms. In supervised learning algorithms, each training tuple is labeled with the class to which it belongs. We used decision trees, random forest and multiclass support vector machines for our experiments. These algorithms are explained in brief.

5.1 DECISION TREE

Decision tree induction is the process of creating decision trees for class-labeled training tuples. A decision tree is basically a tree structure like a flowchart. Each internal node of the tree represents a test on an attribute and each branch is the outcome of the test. Each leaf node is a class label. The first node at the top of the tree is the root node. To classify a given tuple X, the attributes of the tuple are tested against the decision tree starting from the root node to the leaf node which holds the class prediction of the tuple. Ross Quinlan introduced a decision tree algorithm called ID3 in his paper. Later he introduced a successor of ID3 called C4.5 in to overcome some shortcomings such as over-fitting. Unlike ID3, C4.5 can handle both continuous and discrete attributes, training data with missing values and attributes with differing costs. In a basic decision tree induction algorithm, all the training tuples are at the root node at start. The tuples are then partitioned recursively based on selected attributes. The attributes are selected based on an attribute selection method which specifies a heuristic procedure to determine the splitting criterion. The algorithm terminates if all the training tuples belong to the same class or there are no remaining attributes for further partitioning or all training tuples are used. ID3 uses the attribute selection measure called information gain, which is simply the difference of the information needed to classify a tuple and the information needed after the split. These two can be formularized as follows:

Expected information needed to classify a tuple in the training set D

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Where; pi is the nonzero probability that a tuple in D belongs to class Ci. Information needed after the splitting (to arrive at the exact classification)

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

where A is the attribute on which the tuples are to be partitioned. Then, information gain

$$Gain(A) = Info(D) - Info_A(D)$$

The attribute with highest information gain is selected as the splitting attribute. C4.5 uses gain ratio as the attribute selection measure. Gain ratio is an extension to information gain in a sense because it normalizes information gain by using a split information value;

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)}$$

The attribute with the highest gain ratio is selected as the splitting attribute.

5.2 RANDOM FOREST CLASSIFIER

Random Forests is an ensemble method for classification and regression. Random forests are a set of decision trees where each tree is dependent on a random vector sampled independently and with the same distribution of all the trees in the forest. The algorithm generates a number of decision trees creating a forest. Each decision tree is generated by selecting random attributes at each node to determine the split. Tim Kam Ho introduced the first method for random forests using random subspace method in his paper. Later,

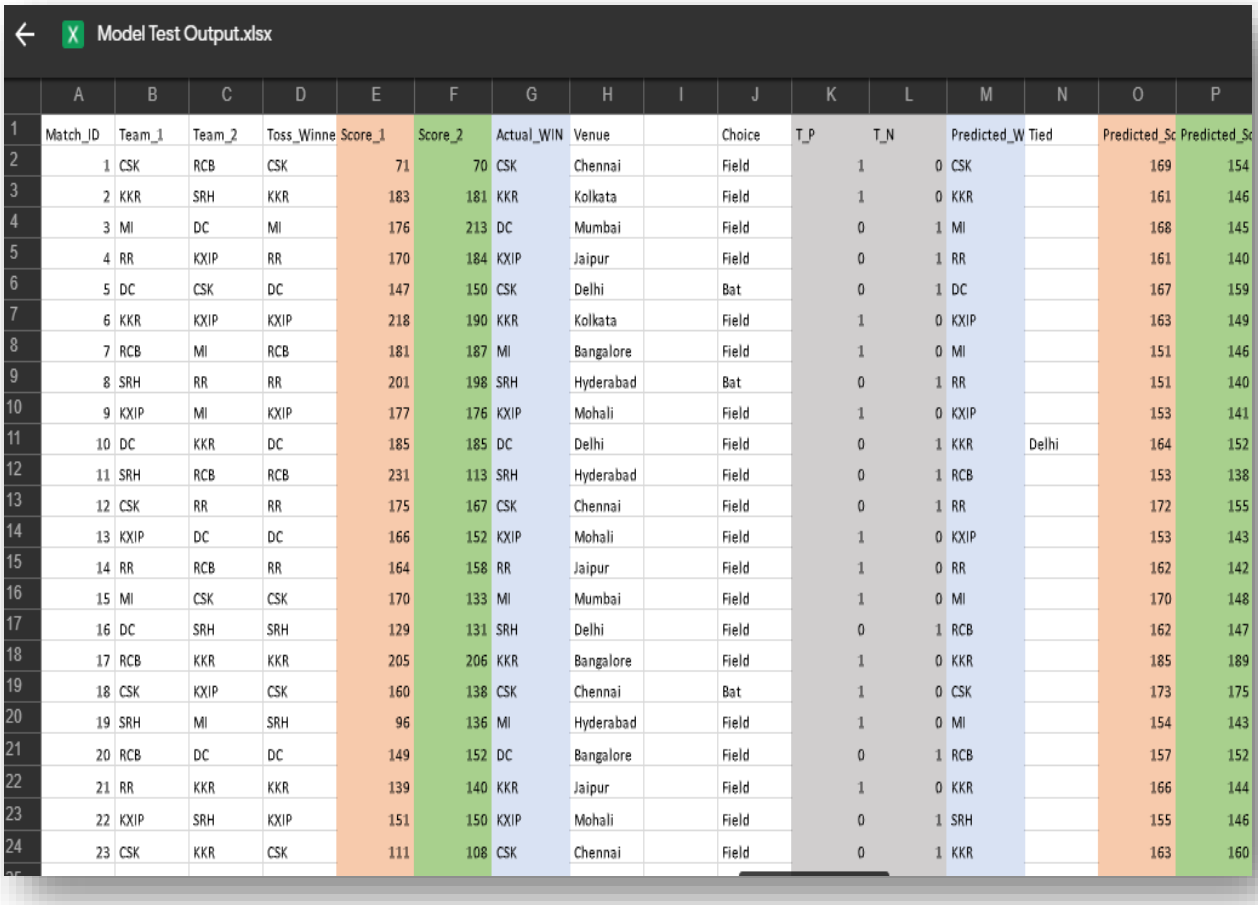
Breiman Leo extended the algorithm in his paper and this method was official known as Random Forests. The general procedure to generate decision trees for random forests starts with a dataset D of d tuples. To generate k decision trees from the dataset, for each iteration k, a training set Di of d tuples is sampled with replacement from the dataset D. To construct a decision tree classifier, at each node, a small number of attributes from the available attributes are selected randomly as candidates for the split at the node. Then Classification and Regression Trees (CART) method is used to grow the trees. The trees are then grown to maximum size and are not pruned. CART is a non-parametric decision tree induction technique that can generate classification and regression trees. CART recursively selects rules based on variables' values to get the best split. It stops splitting when it detects that no further gain can be made or some pre-determined stopping conditions are met.

CHAPTER 6

RESULTS & ANALYSIS

6.1 TEST DATA

For testing the model, we have used completely new data set i.e., 2019_Data_Set. Our model is trained with 2008_2017 data set, for testing purpose we used this data_set and the results of the model are as follows.



Model Test Output.xlsx

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Match_ID	Team_1	Team_2	Toss_Winne	Score_1	Score_2	Actual_WIN	Venue		Choice	T_P	T_N	Predicted_W	Tied	Predicted_Sc	Predicted_Sc
2	1	CSK	RCB	CSK	71	70	CSK	Chennai		Field	1	0	CSK		169	154
3	2	KKR	SRH	KKR	183	181	KKR	Kolkata		Field	1	0	KKR		161	146
4	3	MI	DC	MI	176	213	DC	Mumbai		Field	0	1	MI		168	145
5	4	RR	KXIP	RR	170	184	KXIP	Jaipur		Field	0	1	RR		161	140
6	5	DC	CSK	DC	147	150	CSK	Delhi		Bat	0	1	DC		167	159
7	6	KKR	KXIP	KXIP	218	190	KKR	Kolkata		Field	1	0	KXIP		163	149
8	7	RCB	MI	RCB	181	187	MI	Bangalore		Field	1	0	MI		151	146
9	8	SRH	RR	RR	201	198	SRH	Hyderabad		Bat	0	1	RR		151	140
10	9	KXIP	MI	KXIP	177	176	KXIP	Mohali		Field	1	0	KXIP		153	141
11	10	DC	KKR	DC	185	185	DC	Delhi		Field	0	1	KKR	Delhi	164	152
12	11	SRH	RCB	RCB	231	113	SRH	Hyderabad		Field	0	1	RCB		153	138
13	12	CSK	RR	RR	175	167	CSK	Chennai		Field	0	1	RR		172	155
14	13	KXIP	DC	DC	166	152	KXIP	Mohali		Field	1	0	KXIP		153	143
15	14	RR	RCB	RR	164	158	RR	Jaipur		Field	1	0	RR		162	142
16	15	MI	CSK	CSK	170	133	MI	Mumbai		Field	1	0	MI		170	148
17	16	DC	SRH	SRH	129	131	SRH	Delhi		Field	0	1	RCB		162	147
18	17	RCB	KKR	KKR	205	206	KKR	Bangalore		Field	1	0	KKR		185	189
19	18	CSK	KXIP	CSK	160	138	CSK	Chennai		Bat	1	0	CSK		173	175
20	19	SRH	MI	SRH	96	136	MI	Hyderabad		Field	1	0	MI		154	143
21	20	RCB	DC	DC	149	152	DC	Bangalore		Field	0	1	RCB		157	152
22	21	RR	KKR	KKR	139	140	KKR	Jaipur		Field	1	0	KKR		166	144
23	22	KXIP	SRH	KXIP	151	150	KXIP	Mohali		Field	0	1	SRH		155	146
24	23	CSK	KKR	CSK	111	108	CSK	Chennai		Field	0	1	KKR		163	160

Figure 4: 2019_Data_Set

6.2 SCORE PREDICTOR

In this we will be analyzing the predicted score of team_1 and team_2 with respect to their innings and it will compared with the actual score data. In the figure below

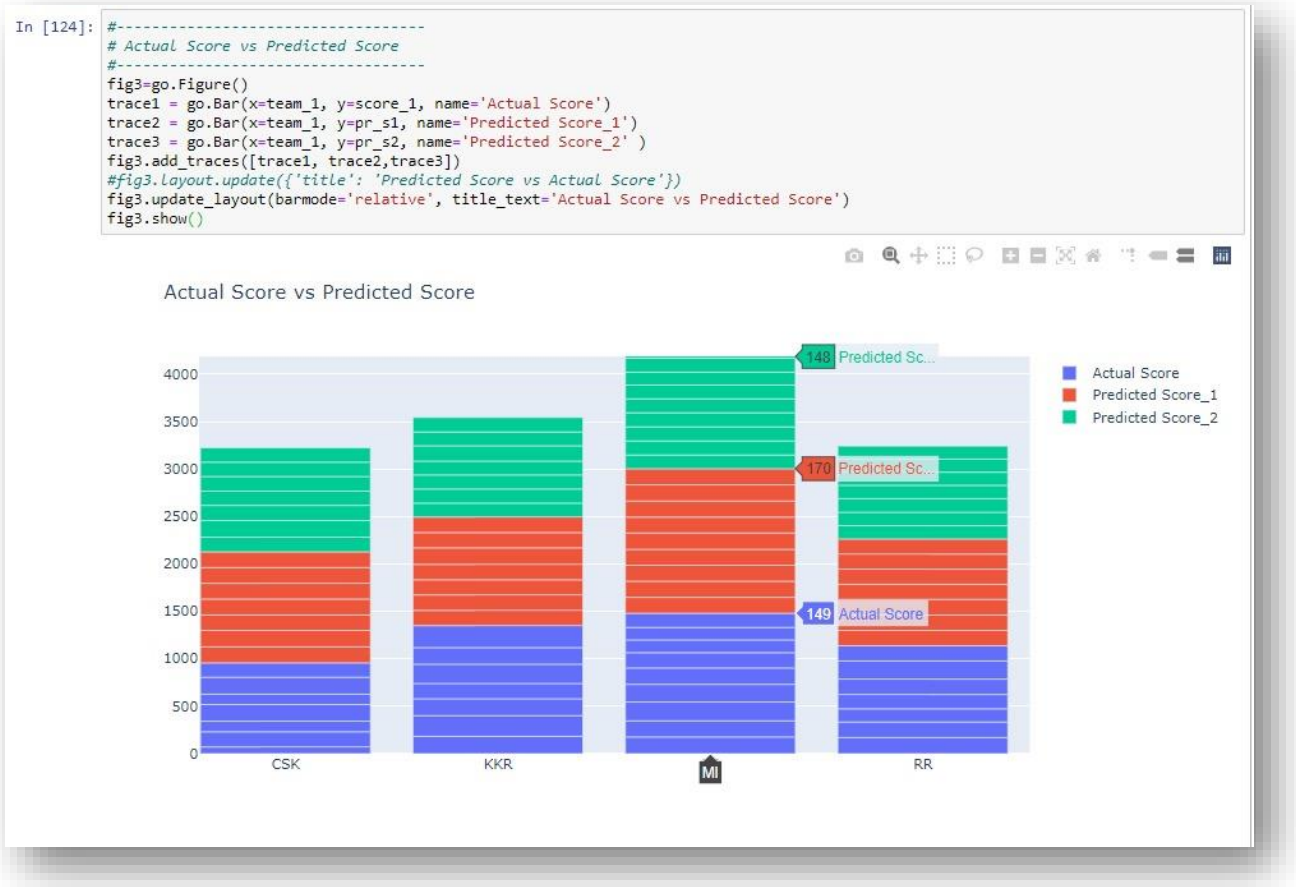


Figure 5: Actual vs Predicted Score of Team

As we see in the above figure, blue stack indicates the actual score by the 2019_data_set where as the red stack indicates the predicted score for team_1 and green stack indicates the predicted score of opposite teams. In most of the cases the predicted score is near to the actual score, but in cases like the venue or innings change may varies the score value.

Figure 6: Actual vs Predicted Score of 2019 Series

6.3 WIN PREDICTOR

Win Predictor is predicted based on the cross validation of individual teams with each other, based up on the trained data with previous ten years data, the match winner be predicted.

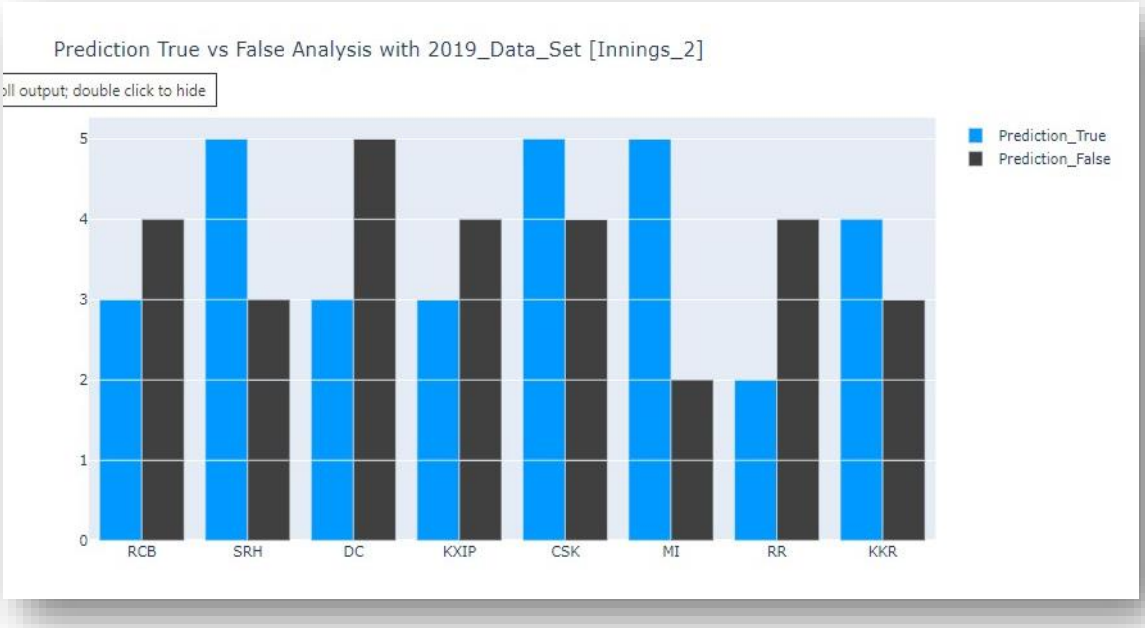


Figure 7: True Positive vs True Negative of Innings 1

True positives and true negatives of innings 2 is as follows, there is a issue with particular case where matches will be suspended because of rain, these matches score is calculated ‘0’ for both true positives and true negatives.

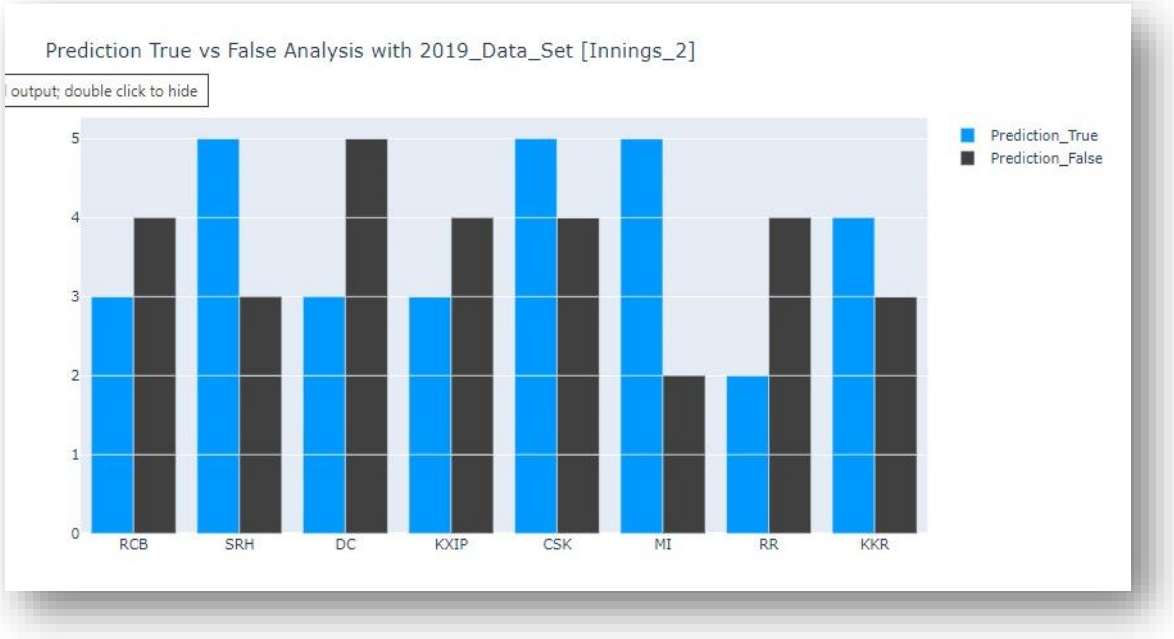


Figure 8: True Positive vs True Negative of Innings 2

CHAPTER 7

FUTURE WORK & SCOPE

The problem with IPL is every year a month before the series start, players will go for auction so every year there is a chance for varying 3 ~ 5 players from the team among the 11 players. All this work has been carried out with an assumption of at least 70% of the team is retained. A solution should be formulated in order to overcome this issue, this results in low prediction accuracy of the model.

Resolving this issue with a better model will form as scope for the future work. This troubleshooting can boost up the model with better accuracy.

REFERENCES:

[1] https://airccj.org/CSCP/vol8/csit88310.pdf

[2] https://www.jhse.ua.es/article/view/2014-v9-n4-predicting-the-performance-of-batsmen-in-test-cricket

[3] https://github.com/smidtfab/nba_score_analyzer

[4] https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7489605

[5] https://datascience.stackexchange.com/questions/12101/machine-learning-technique-to-calculate-weighted-average-weights

[6] http://localhost:8888/notebooks/Documents/Data%20Science%20Project/IPL-RESULT-PREDICTOR.ipynb