**CAPSTONE PROJECT FINAL REPORT**

**Gaurav Sood**

**PGP-DSBA Online Dec\_B’21**

**Date: 4th December, 2022**

**Table of Contents**

**1. Introduction 6**

**2. EDA and Business Implication 8**

**3. Data Cleaning and Pre-processing 21**

**4. Model Building 25**

**5. Model Validation 27**

**6. Final Interpretation/Recommendation 29**

**List of Tables**

Table no.1: Data Info ……………………………………………………………………………………………………….…………………. 08

Table no.2: Data Description ………………………………………………………………………………………………………….……. 09

Table no.3: Missing Value Check …………………………………………………………………………………………………………. 22

Table no.4: Continuous Variables ………………………………………………………………………………………………………… 23

Table no.5: Categorical Variables ………………………………………………………………………………………………………… 23

Table no.6: Dummy Variables ……………………………………………………………………………………………………………… 24

Table no.7: Combining Both Set of Variables ………………………………………………………………………………………. 24

Table no.8: Train vs Test Shape …………………………………………………………………………………………………………… 25

Table no.9: Logistic Regression Score ………………………………………………………………………………………………..… 25

Table no.10: Random Forest Score ……………………………………………………………………………………………………… 25

Table no. 11: Decision Tree Score ………………………………………………………………………………………………………… 25

Table no. 12: Model Evaluation …………………………………………………………………………………………………………… 27

Table no. 13: Actual Result vs Predicted Result …………………………………………………………………………………… 27

Table no. 14: Actual Result Count ……………………………………………………………………………………………………..… 28

Table no. 15: Predicted Result Count …………………………………………………………………………………………………… 28

**List of Figures**

Figure no. 1: Result Graph ………………………………………….……………………………………………………………………….. 10

Figure no. 2: Match Lighting Graph …..…………………………………………………………………………….…………………… 11

Figure no. 3: Match Format Graph ……………………………………………………………………………………..……………….. 12

Figure no. 4: Batting-Bowling Graph …..……………………………………………………………………………………………….. 13

Figure no. 5: Opponent Graph …………………………………………………………………………………………………………….. 14

Figure no. 6: Season Graph …………………………………………………………………………………………………………………. 15

Figure no. 7: Home-Away Graph ……………..………………………………………………………………………………………….. 16

Figure no. 8: Result vs Bowlers in Team Graph …………………………..……………………………………………………….. 17

Figure no. 9: Result vs All-rounders in Team Graph ……………………………………………………………………………… 17

Figure no. 10: Result vs Highest Wickets in Team Graph ….………………………..………………………………………… 18

Figure no. 11: Result vs Max Wickets in Over Graph ………………………….………………………………………………… 18

Figure no. 12: Correlation Heat-map ...........………………………………………………………………………………………… 19

Figure no. 13: Outlier Boxplot ……………………………………………………………………………………………………………… 21

**Problem Statement**

BCCI has hired an external analytics consulting firm for data analytics. The major objective of this tie up is to extract actionable insights from the historical match data and make strategic changes to make India win. Primary objective is to create Machine Learning models which correctly predicts a win for the Indian Cricket Team. Once a model is developed then you have to extract actionable insights and recommendation.

Also, below are the details of the next 5 matches, India is going to play. You have to predict the result of the matches and if you are getting prediction as a Loss then suggest some changes and re-run your model again until you are getting Win as a prediction. You cannot use the same strategy in the entire series, because opponent will get to know your strategy and they can come with counter strategy. Hence for all the below 5 matches you have to suggest unique strategies to make India win. The suggestions should be in-line with the variables that have been mentioned in the given data set. Do consider the feasibility of the suggestions very carefully as well.

1. 1 Test match with England in England. All the match are day matches. In England, it will be rainy season at the time to match.

2. 2 T20 match with Australia in India. All the match are Day and Night matches. In India, it will be winter season at the time to match.

3. 2 ODI match with Sri Lanka in India. All the match are Day and Night matches. In India, it will be winter season at the time to match.

**1. Introduction**

**a) Defining problem statement**

We are given a scenario wherein, the BCCI which is the highest authority for the sport of Cricket in India, has made use of an analytics consulting firm in order to help guide some of the decision-making that goes on within the team using the past set of historical data from matches played by the Indian Cricket team.

We have been provided the schedule of the Indian Cricket team for the next 5 matches, which consist of:

1. 1 morning-to-evening Test match with team England in England during monsoon season.
2. 2 day-night T20 matches with Australia in India during winter season; and
3. 2 day-night ODIs with Sri Lanka in India, also during winter season.

With the past dataset that we have been handed over, we are supposed to create Machine Learning models that would predict the result of the above 5 matches, individually. In case of a loss, we must make suitable adjustments to the team so that the model gives a winning result for team India. Furthermore, we cannot use the same strategy in more than one match as the opponent team will be prepared with counter-plays to our gameplan. Hence, we are also required to have a substantial knowledge of the sport of cricket as well to go along with our technical and statistical expertise.

**b) Need of the study/project**

Our mission here is to create finely tuned models using our knowledge of Machine Learning that would give the result as a win for team India. Then, with that information and with the conditions that we set in our model, we must present our valuable inferences and offer data driven solutions to the relevant cricketing authorities.

What we are doing here is taking part in Predictive Sports Analytics, i.e. making predictions using sports data. One such use case in cricket is to **predict the number of runs a batsman scores**against an opponent in a particular match. This in turn, helps the team management along with the captain to select the best team for every match and catered to every opponent.

Sports Analytics has exploded onto the scene in the past decade. It has absolutely become a game-changer. Using analytics in sports impacts the decision making of a team and can even alter the future of the club (or country in this instance). It has become a must-have aspect to explore for every team in every major sport in order to steer the odds in their favour.

The impact of such an undertaking could be truly wide-ranging.

From an on-the-field perspective, it could help solidify the team to prepare them for future international tournaments (World Cup, Champions Trophy, etc.) It could also help each and every player to know more about themselves and their teammates, whether it be their strengths and weaknesses, how they react in certain situations and so on.

India is a cricket-mad nation. We are a country of over 1.4 billion people and it is by far our most popular sport. When we play our rivals Pakistan, it is said to be the most watched event in all of sports, even surpassing the American Superbowl. So, the business and social implications of team India winning are quite clear.

1. Wins on the cricket field result in more interest in the cricket team. This means more jersey and merchandise sales for team India from fans willing to spend.
2. Wins from team India could also result in re-negotiation of TV and broadcasting deals from the BCCI with the likes of Star Sports.
3. More eye balls on the product mean more revenue to be generated from advertising.
4. Increase in sponsorships (or better sponsorships) not only for the Indian Cricket team, but for players as well individually in the form of advertisements.
5. More revenue to be generated from tickets sold to people attending future matches, if sufficient public interest in the team has been generated from good performances.
6. Feel-good factor for all Indians.
7. Bragging rights over competitive cricketing sides.
8. Lastly, as the BCCI is a huge body, they may recommend our consulting firm for future business opportunities with their partners if they are satisfied with our work.

**2. EDA and Business Implication**

First, we loaded all the necessary libraries for model building.

**Data Info**

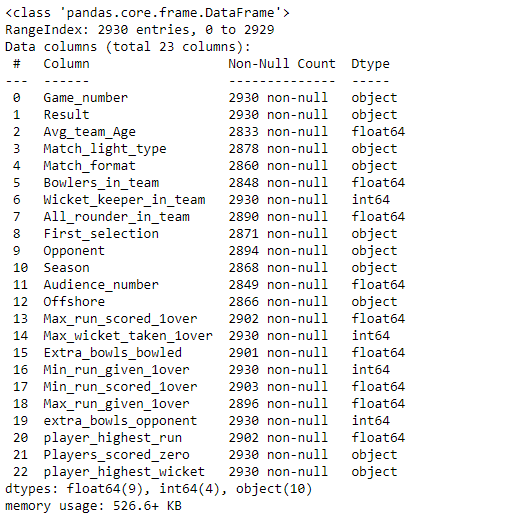


Table no.1: Data Info

The given dataset is found to have 23 variables over 2930 observations, with 9 of those variables being of float type, 4 of those being integer type and 10 being object type.

**Data Description**

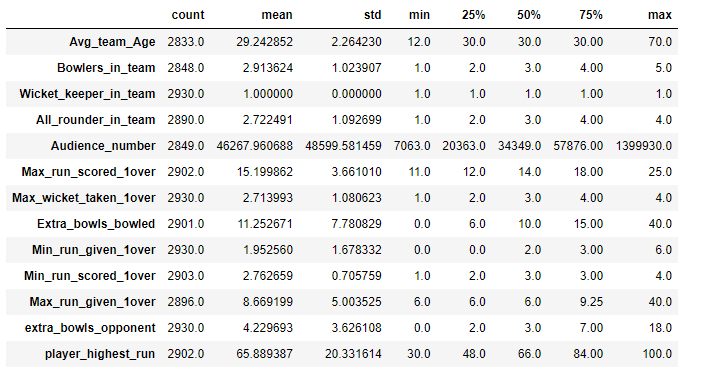


Table no.2: Data Description

From the above table, we can make some basic inferences:

* The average team age is about 29 years.
* The team usually plays with 3 bowlers and 3 all-rounders in the team, and with 1 wicket-keeper.
* Average attendance for team India matches is about 46,000.
* The team usually bowls around 11-12 extras in an innings, however has once conceded a whopping 40 extras.
* The highest scoring batsman usually scores about 66 runs in an innings. The maximum scored by a batsman is 100, completing his century in the process.

**Result Plot**

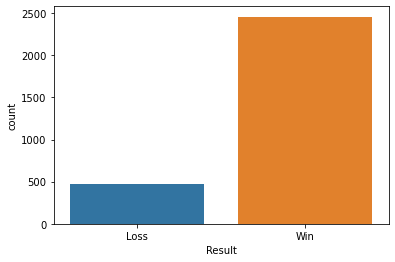


Figure no.1: Result Graph

From the above plot, it is clear that team India wins a vast majority of their matches, with roughly 2,500 wins as compared to just 500 losses.

**Match Lighting Plot**

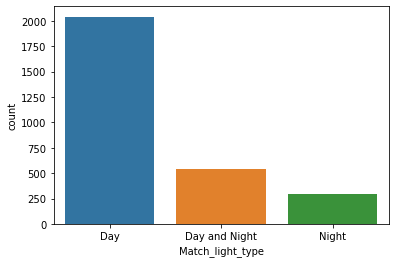


Figure no.2: Match Lighting Graph

The above plot shows us that most matches are played during the day, about 2,000 of them in the given dataset. There are about 500 matches that are day-night and about 300 that are purely night matches

**Match Format Plot**

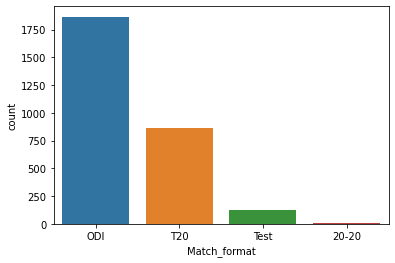


Figure no.3: Match Format Graph

One-Day Internationals are by far the most popular form of cricket, with roughly 1,900 matches in total. The Twenty Over format comes in second with about 800 matches, with Test cricket being the least played with just about 200 matches in total. There’s also been a data compiling error here, as the 20-20 field should be lumped into the T20 field.

**Batting-Bowling Plot**

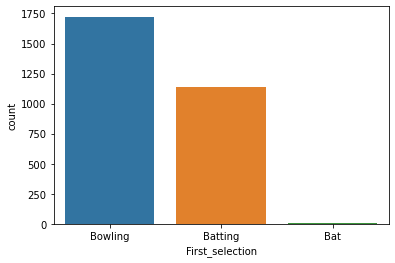


Figure no. 4: Batting-Bowling Graph

Team India is found to be bowling first in most of their matches, about 1,700 times as compared to batting first in roughly 1,100 matches. There is data compilation error here as well, with the Bat field needed to be merged with the Batting field. Keep in mind, this does not tell us about who won the toss and who elected to bat or bowl first, simply the outcome is given.

**Opponent Plot**

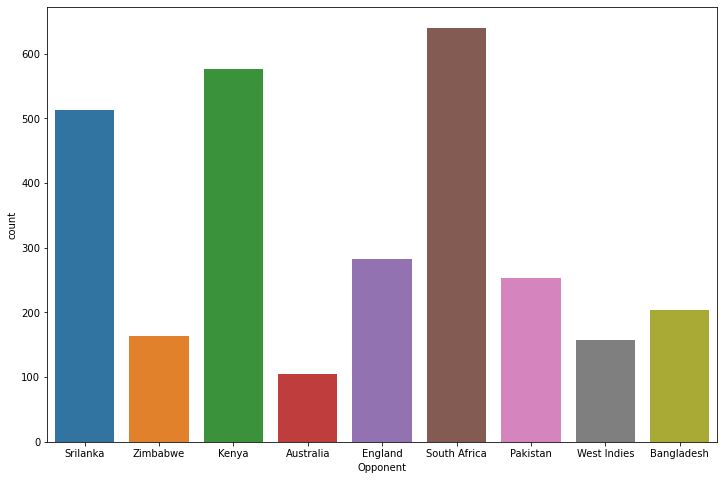
****

Figure no.5: Opponent Graph

The colourful bar graph above shows us the number of times team India has faced each opponent.

* South Africa is found to be the number 1 opponent, with over 600 matches between the two cricketing nations.
* Kenya and Sri Lanka are next in line, with 500 matches each.
* Australia is the rarest opponent, with just about 100 matches in total.

**Season Plot**

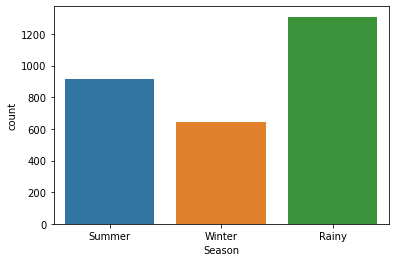
****

Figure no.6: Season Graph

* Most matches played by team India occur during the rainy season, over 1,200 matches in total.
* Summer season sees the next highest number of matches, with about 900 matches.
* Winter season is dead last with just above 600 matches.

**Home-Away Plot**

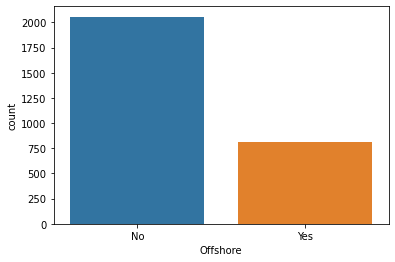


Figure no.7: Home-Away Graph

* India plays the vast majority of its matches at home, with approximately 2,000 matches in total.
* About 800 of its matches are played abroad.
* However, not all these matches could be classified as ‘Away’ matches. Some of them could be on neutral ground as well.

**Relationship between Result and no. of Bowlers in Team**

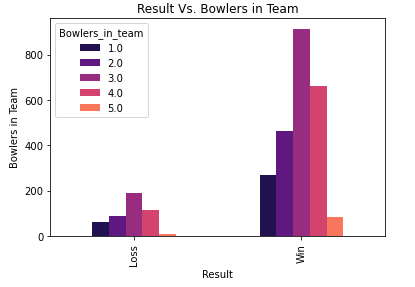


Figure no.8: Result vs Bowlers in Team Graph

If we extrapolate the losses to match the number of wins proportionally, there is not much of an advantage to be found by having more number of bowlers in the team that would result in more wins.

**Relationship between Result and no. of All-rounders in Team**

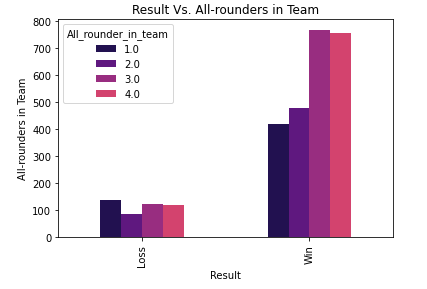


Figure no.9: Result vs All-rounders in Team Graph

There is found to be quite a significant link between having more all-rounders in the team resulting in wins than having fewer all-rounders. This points to the versatility that all-rounders possess, having the ability to be plugged into any role necessary in a variety of situations and being able to perform well leading to more wins.

**Relationship between Result and Wickets in Team**

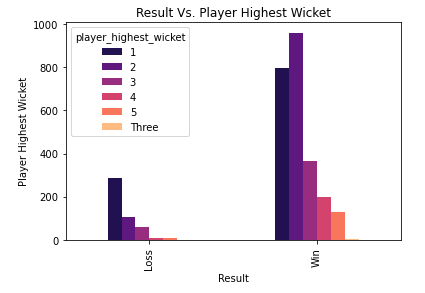


Figure no.10: Result vs Highest Wickets in Team Graph

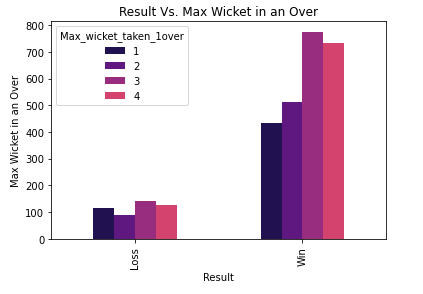


Figure no.11: Result vs Max Wicket in Over Graph

Naturally, more wickets attained leads to increase in amount of wins. There is clear evidence of this in the above plots.

**Correlation Heat-map**



Figure no.12: Correlation Heat-map

There is not much of a correlation to be found amongst the variables in the given dataset.

* However, one interesting observation is the relationship between extra bowls bowled to each of Audience number as well as to Max runs given in 1 over.
* This could indicate the pressure factor felt by bowlers with a larger audience in the stadium, or simply not being able to focus enough (being distracted).
* The pressure aspect of the game could be applied to the max runs given in 1 over as well. If bowlers are being hammered in the over they might feel less confident in subsequent deliveries or subsequent overs.
* Of course, the extras themselves contribute to runs in the over as well. Every wide ball or no ball adds an extra run to the opposition, without even counting the delivery.
* The opponents are also discovered to be suffering from the same issues to some extent.

**3. Data Cleaning & Pre-processing**

First, we check for outliers in the data through the use of a boxplot.

**Outlier Boxplot**

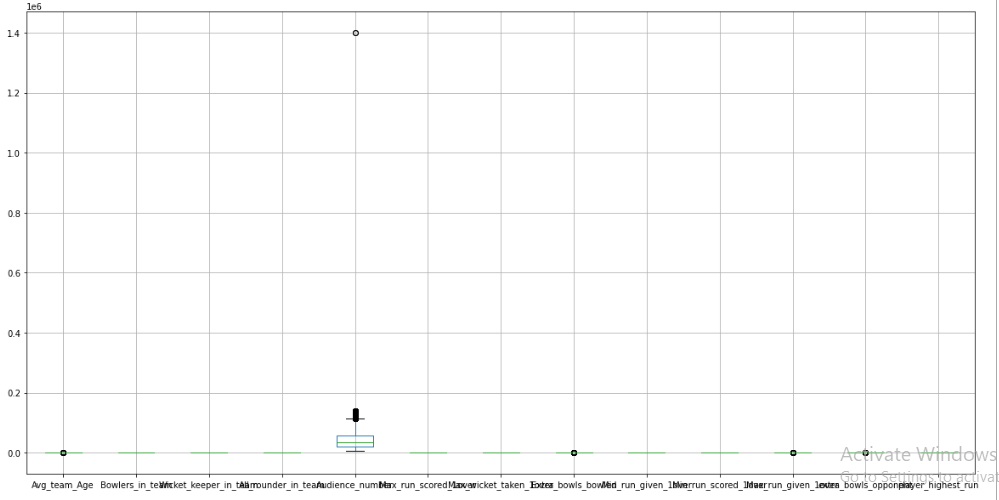


Figure no.13: Outlier Boxplot

* The only variable found to be containing outliers is Audience\_number.
* As this is an external factor/outcome and not one related to team composition or team performance, we need not treat these outliers.
* All the other variables are relatively grouped together.

**Checking for Missing Values**

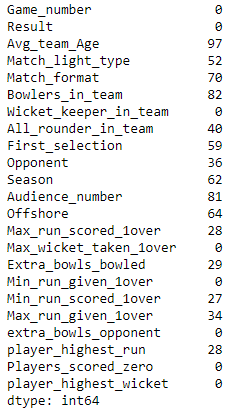


Table no.3: Missing Value Check

There are a number of missing values in the dataset, a total of 789 missing values across 15 fields.

We shall treat these missing values after the variable transformation process and right before our model is built. Missing values can bias the results of the machine learning models and/or reduce the accuracy of the model.

We divided the variables based on their typing:

1. Continuous variables, which are of float and integer type; and
2. Categorical variables, which are of object type.

**Continuous Variables**

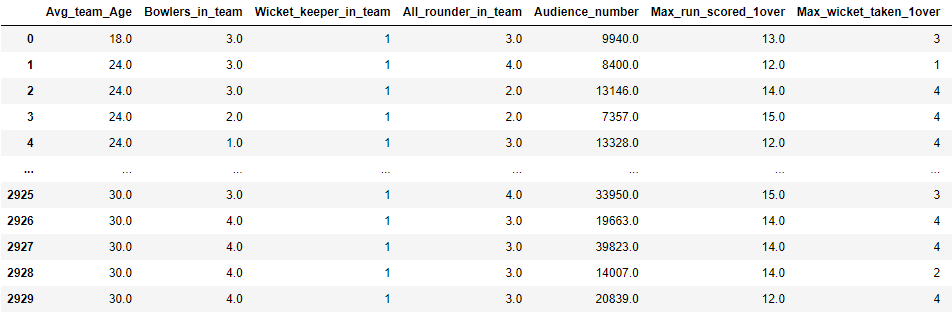


Table no.4: Continuous Variables

**Categorical Variables**

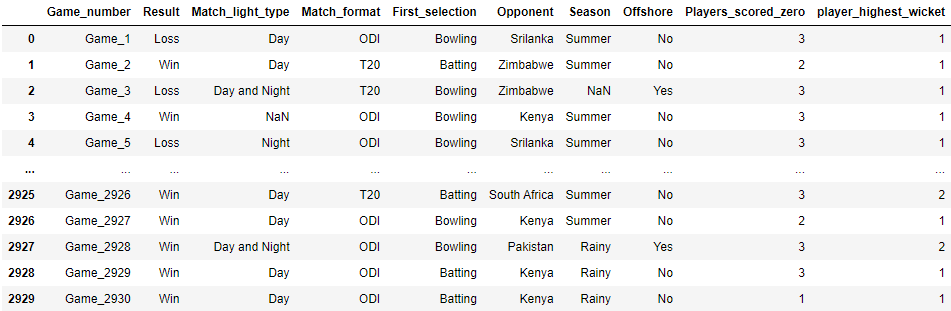


Table no.5: Categorical Variables

Here, we will drop the variable ‘Game\_number’ as it is irrelevant.

**Creating Dummy Variables**

After checking the number of categorical values in each column, we move on to create dummy variables in binary or in multi classification variables for each categorical variable to help us to predict in a better way.

This step is crucial, as categorical variable transformation is important for any models and its selection matters a lot for model performance. It is a way to make the data work better in our model.

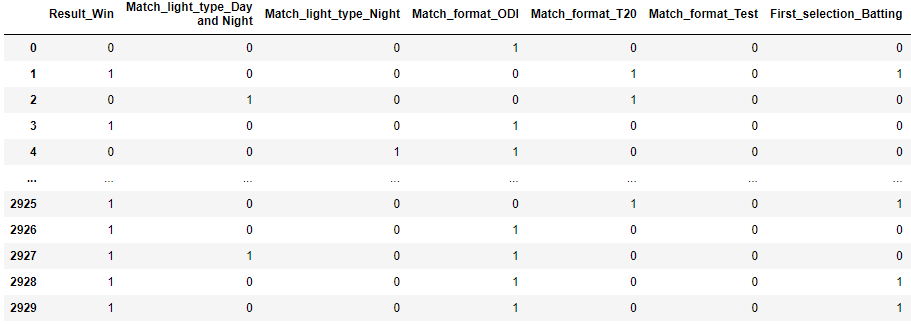


Table no.6: Dummy Variables

Here, we have our categorical variables presented in a binary format.

Next, we combine the continuous and categorical variables together to create a new dataframe.

Then, we created 'x' variable for training and test purpose, where we removed ‘Result’ from the dataframe.

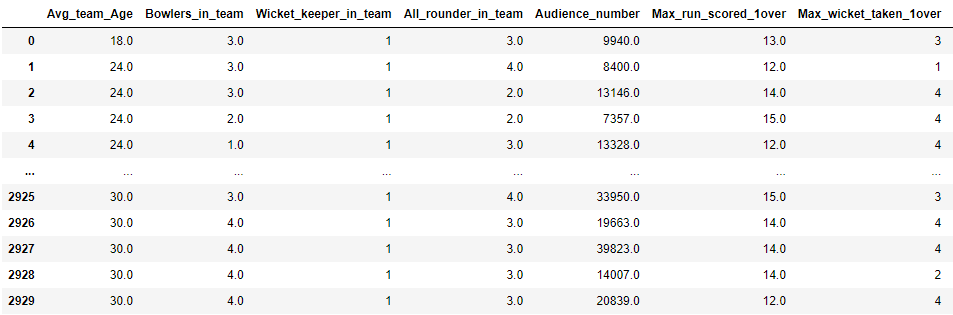


Table no.7: Combining both set of Variables

Next, we convert all the float values into integer values and then proceed to the Model building part of our project.

**4. Model Building**

Here we are splitting the data into training and testing model, where training data is 80% and testing data is 20%.

**Train and Test Dimensions**



Table no.8: Train and Test Shape

**Logistic Regression Model**



Table no.9: Logistic Regression Score

After using Logistic Regression, we are able to find out that we are getting an accuracy score of 83.78% (approximately) which is considered as quite good.

**Random Forest Model**



Table no.10: Random Forest Score

After using Random Forest, we are able to find out that we are getting an accuracy score of 100% which is better than the Logistic Regression model we built earlier.

**Decision Tree Model**



Table no.11: Decision Tree Score

After using Decision Tree, we are able to find out that we are getting an accuracy score of 100% which is better than Logistic Regression and equal to Random Forest.

**Choosing our Model**

Ultimately, we decided to go ahead with the Random Forest model.

* We chose this over the Logistic Regression model due to having a better score.
* Although we had a 100% score on both our Random Forest and Decision Tree models, we chose Random Forest over Decision Tree due to the fact that a decision tree combines some decisions, whereas a random forest combines several decision trees.
* Decision Trees are graphs that illustrate all possible outcomes of a decision using a branching approach. In contrast, the Random Forest algorithm output are a set of Decision Trees that work according to the output.
* Machine learning engineers and data scientists often use the Random Forest algorithm because they are so accurate and because modern computers and systems can usually handle large, previously unmanageable datasets.

**5. Model Validation**

We can evaluate the model individually on their score and we can see that Random Forest and Decision Tree are better options to choose.

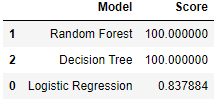


Table no.12: Model Evaluation

**Predicting on Random Forest.**

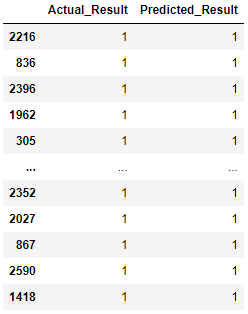


Table no.13: Actual Result vs Predicted Result

**Actual Result vs Predicted Result**

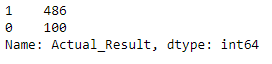


Table no.14: Actual Result Count



Table no.15: Predicted Result Count

Actual Results ended up with 486 wins to 100 losses as compared to the Predicted Results of 508 wins to 78 losses.

**6. Final Interpretation/Recommendation**

A number of Business insights and subsequent recommendations can be made from exploring and analysing the given data.

* The management with the coaches and the captain must focus on team performances rather than individual performances in order to ensure wins. These wins consequently result in higher revenue generated across the board, whether it be for the BCCI or the team itself through sponsorships, ad revenue, player advertisements, merchandise sales, greater attendance numbers and much more.
* An increase in number of bowlers played was not found to be leading to more wins, they more or less stayed the same in proportion.
* However, an increase in all-rounders played was found to have a far greater impact in guaranteeing wins. This is due to their ability to play multiple roles in the team, in batting and bowling. If one bowler is having a bad day, he could be rotated out and given less overs in favour of an all-rounder. Conversely, if the top order collapses, the all-rounders could help stabilise the situation and give more of a fighting chance to the team than if it were a bowler in their place.
* Bowlers should be bowling more of aggressive, wicket-taking deliveries; as increase in wickets was evident in leading the team to more number of wins.
* On the other hand, the management needs to make use of sports psychologists for the team. In bowlers’ case, they are found to be giving away silly runs by bowling more extras than usual if the attendance is high and also if they are already being scored on more. Mitigating/dealing with pressure is absolutely crucial in this era of sports, even more so with the advent of social media.
* The BCCI could also schedule more matches (with the help of the ICC and other cricketing boards) against strong teams with historical rivalries to India, such as Pakistan and Australia instead of playing matches again and again with the likes of South Africa, Sri Lanka and Kenya which could lead to boredom for the average fan . This could drum up interest in the sport at home.

**THE END**