Summary Report

We have to build the model which can assign the lead score to individual from 0 to 100 and higher the score better is the conversion rate also we have to increase the accuracy of conversion rate of 30% to 80%.

We begin the process by cleaning the data by removing the columns which are having the same type of value and which are of no use for predicting the model, also we removed the null values by filling with zeros because considering such huge datasets we would be fine if we treat this null values as zero's.

Then we perform the univariate and bivariate analysis on the remaining data sets and found some interesting patterns and found that these following variables have some impact on tha converted : Lead Origin','Lead Source','Last Notable Activity','Lead Number','Lead Profile','What matters most to you in choosing a course','What is your current occupation','How did you hear about X Education','Specialization','Last Activity','Tags','Page Views Per Visit','Total Time Spent on Website','TotalVisits','Converted'

By using this we created the dummy varibles and we try to fit the logistic regression model .

Then by using the RFE we select features by recusively considering the smaller and smaller sets of feature and we found out following variables:

'Lead Number', 'Lead Origin', 'Lead Source', 'TotalVisits',

'Total Time Spent on Website', 'Page Views Per Visit', 'Last Activity',

'Specialization', 'How did you hear about X Education',

'What is your current occupation',

'What matters most to you in choosing a course', 'Tags', 'Lead Profile',

'Last Notable Activity'

We again run the logistic model on these variable and get minimum p value that is less than 0.05.

Then we started making prediction in probabilities and we are making prediction when the probability is greater than 50% we convert it to 1 that is hot lead.

After testing the data on model we evaluate the model by building the confusion matrix on y_pred_final and x_pred_final variable and we got the overall accuracy rate of 72.11% .

```
             precision  recall   f1-score   support

 0               0.72       0.87      0.79        1677

 1                0.71      0.49      0.58         1095

avg / total    0.72       0.72     0.71          2772
```

At last we run our model using the training dataset we got nearly same result with overall accuracy of **72.11%**

```
                precision    recall   f1-score   support

 0                 0.72      0.87       0.79      1677

 1                 0.71      0.49      0.58      1095

avg / total     0.72      0.72      0.71      2772.
```