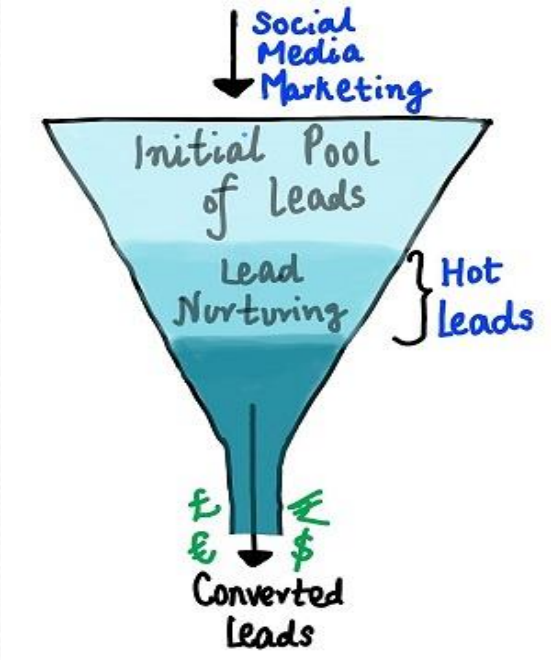
LEAD SCORE CASE STUDY

BY: CHANDANI, SAI SHARAN AND VISHNU KUMAR

PROBLEM STATEMENT

- An Education company (X) sells online courses to industry professionals.
- It needs help to select the most promising leads that will turn into profiting customers.
- The leads gotten by the company are many, but the conversion rate is poor. If we say they acquire 100 leads only 30 are getting converted.
- The company wants to know ‘Hot Leads’ that the most potential targets.
- A typical lead conversion funnel will look like this:



METHODOLOGY

- **In order to get the best solution of the problem, we need to first understand the problem and then understand the data provided to us.**
- **Then comes data cleaning and manipulation where we have to check the missing values and NaN values in the columns of the dataset, check duplicates, dropping and imputation of values is necessary.**
- **Check for outliers and handle them.**
- **Then we have to performed EDA and in that both Univariate analysis and Bivariate analysis.**
- **Feature scaling and dummy variable creation.**
- **Then we have to perform a classification technique called Logistic regression for model building and prediction.**
- **In the end, validation of the model, its presentation and conclusion.**

DATA MANIPULATION

- By using basic syntaxes like ‘.info’, ‘.shape’ and ‘.describe’ we understand the rows and columns present in the data and other statistical values.
- The total columns present in the dataset is 37 and rows are 9240.
- There are few columns that contains ‘select’ as a category where there was no information given, hence replaced it with ‘NaN’.
- There were columns that contained unique values, since they don’t provide any valuable information, hence dropped them. Eg. 'I agree to pay the amount through cheque', 'Get updates on DM Content', 'Update me on Supply Chain Content', 'Receive More Updates About Our Courses', 'Magazine‘.
- Few of the columns that didn’t provided much information were dropped such as ‘Lead Quality’, ‘Prospect ID’.

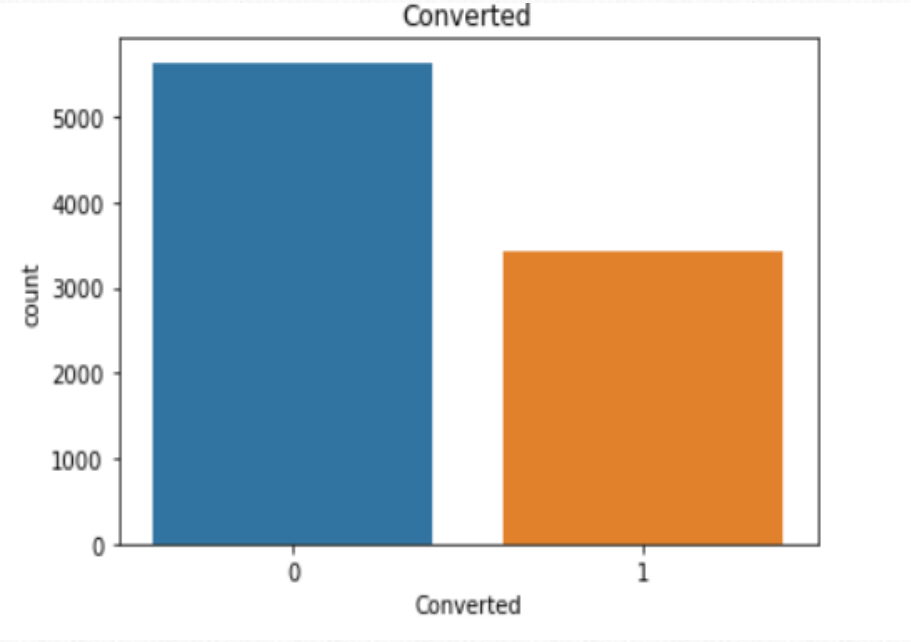
- We have chosen 35% and above missing value criteria, hence dropping the columns that have more than 35% of the missing values such as 'Tags', 'Asymmetrique Activity Index', 'Asymmetrique Profile Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score‘.
- There were few columns which contains NaN value but instead of dropping them we have grouped them together or filled with ‘Not gievn’ categories since these columns can provide valuable information.

EDA

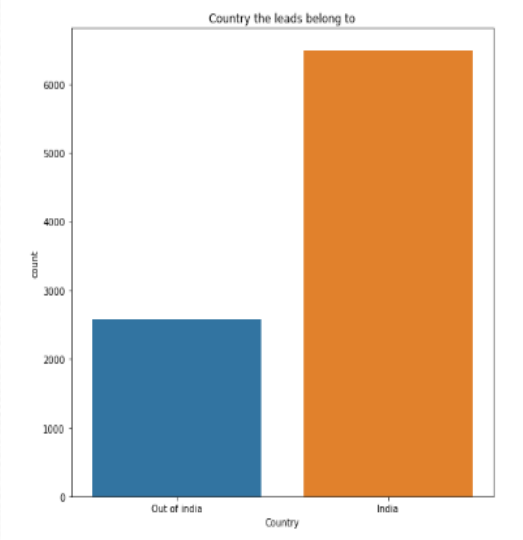
- After cleaning and manipulating the data, we have EDA in which we have done both univariate and bivariate analysis.
- Here 'Converted' is the target variable which seems balanced.

```
df.Converted.value_counts('Normalise')
```

```
0    0.61461
1    0.38539
Name: Converted, dtype: float64
```

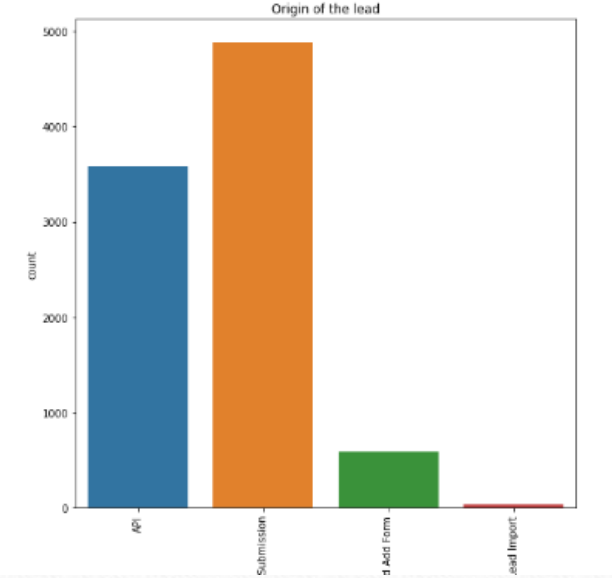


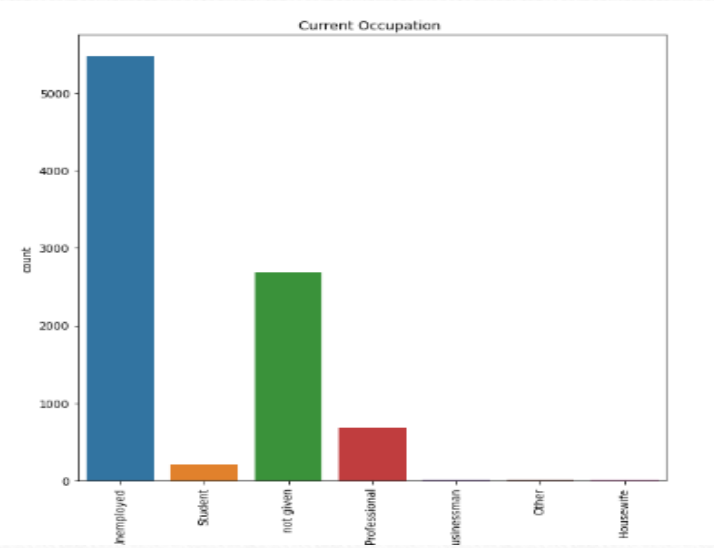
CATEGORICAL DATA VARIENCE UNIVARIATE ANALYSIS



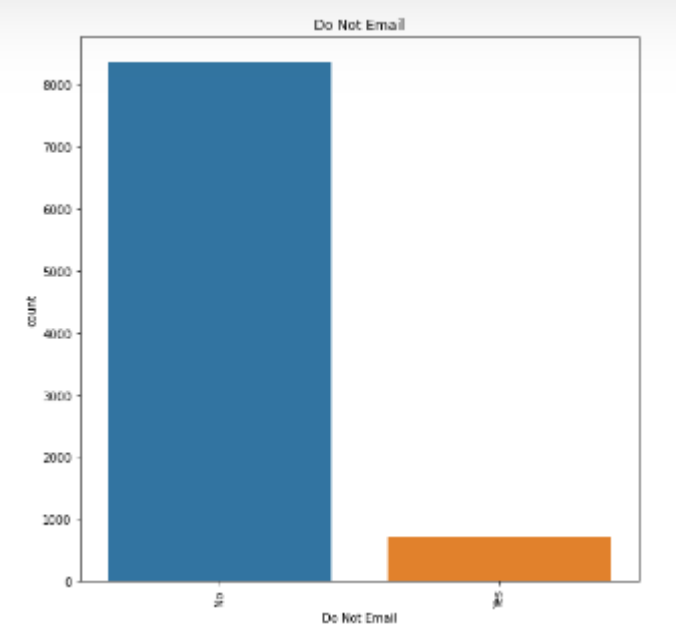
From this we can say that there are many leads from India.

The customers from landing page submission has the highest lead.



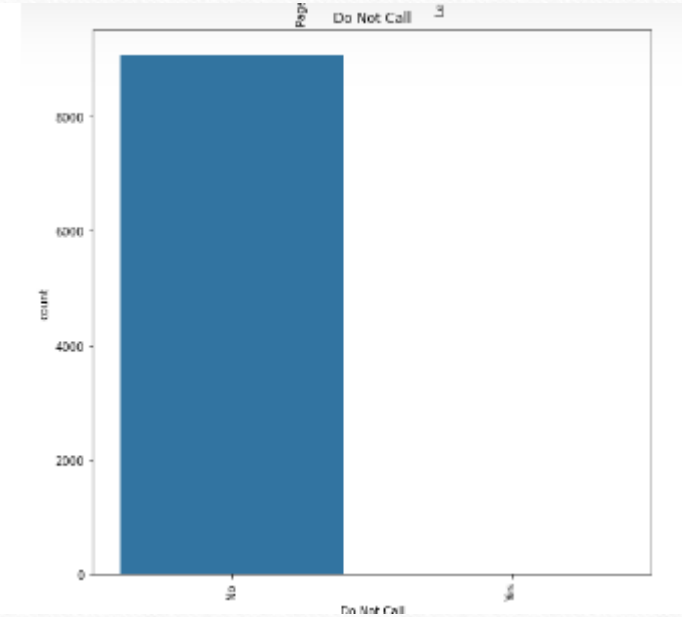


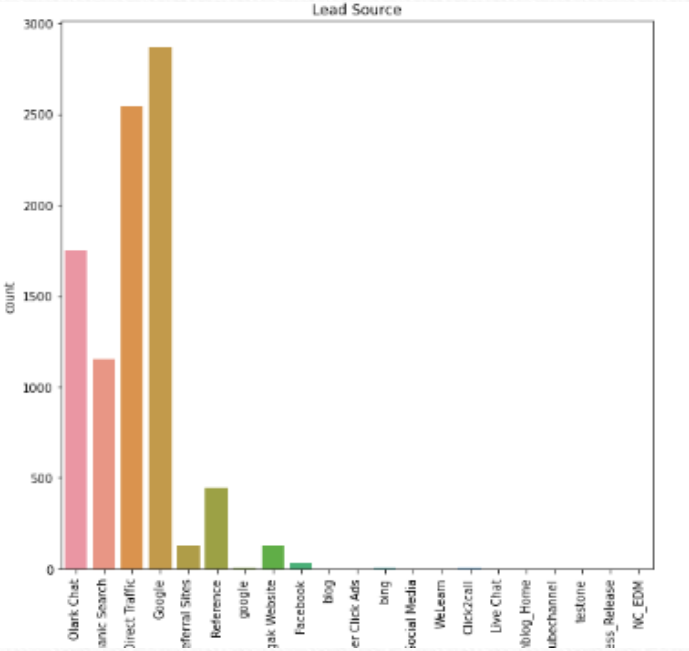
The customers that have highest lead capacity are unemployed.



There are many leads who do not want to get email.

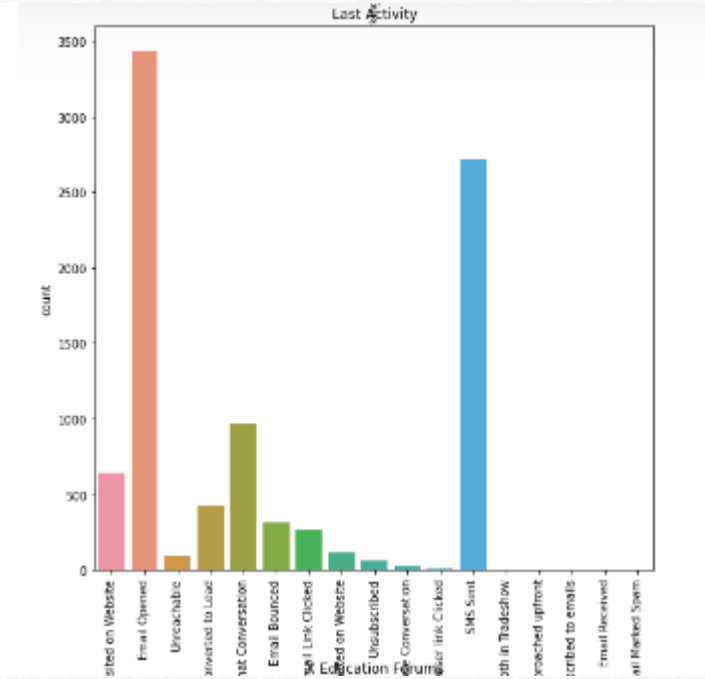
There are many leads who do not want to get a call.



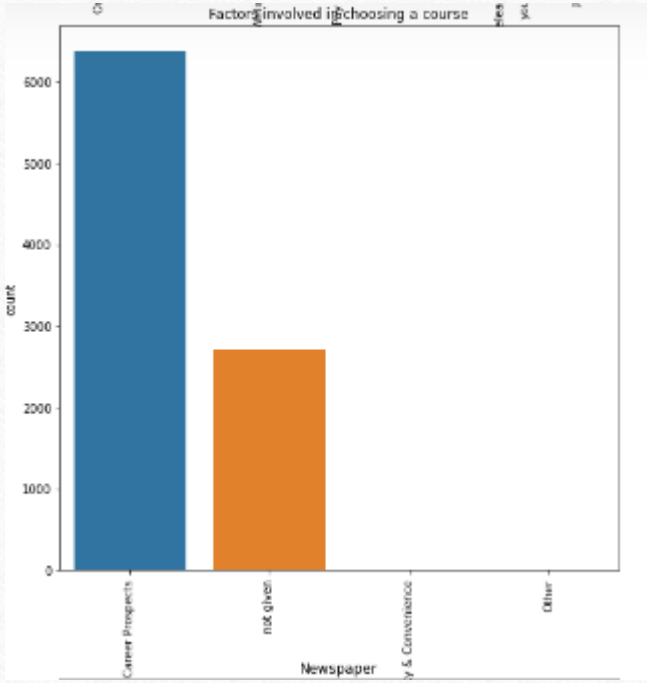


Majority source of the lead is google and direct traffic.

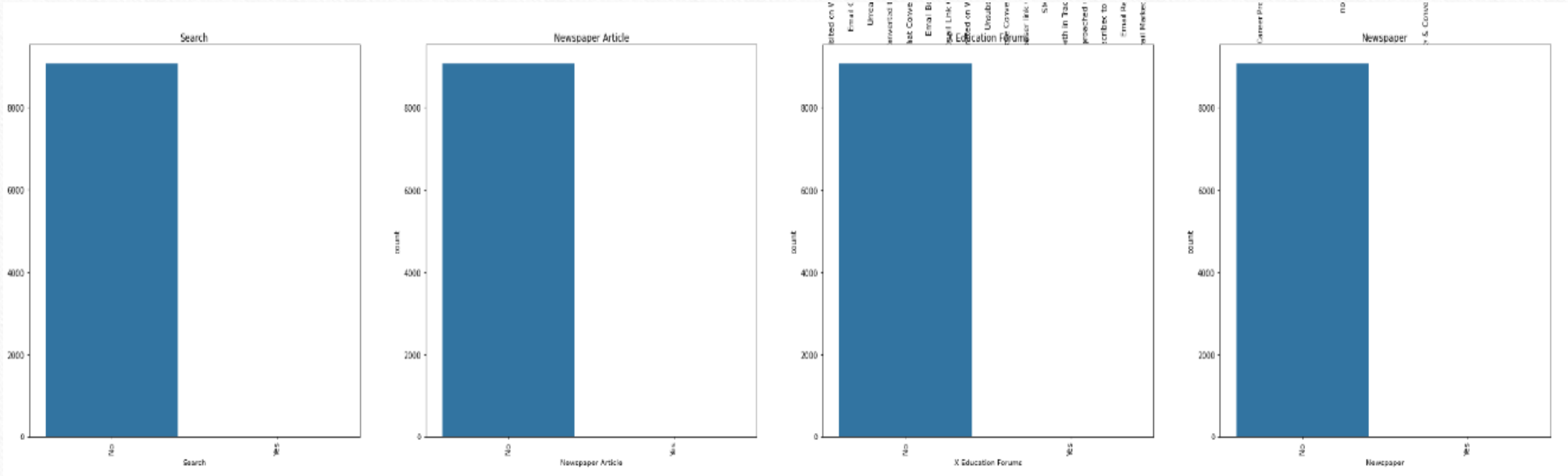
The conversion rate of people with SMS sent is more and customers whose last activity was email opened are in majority.



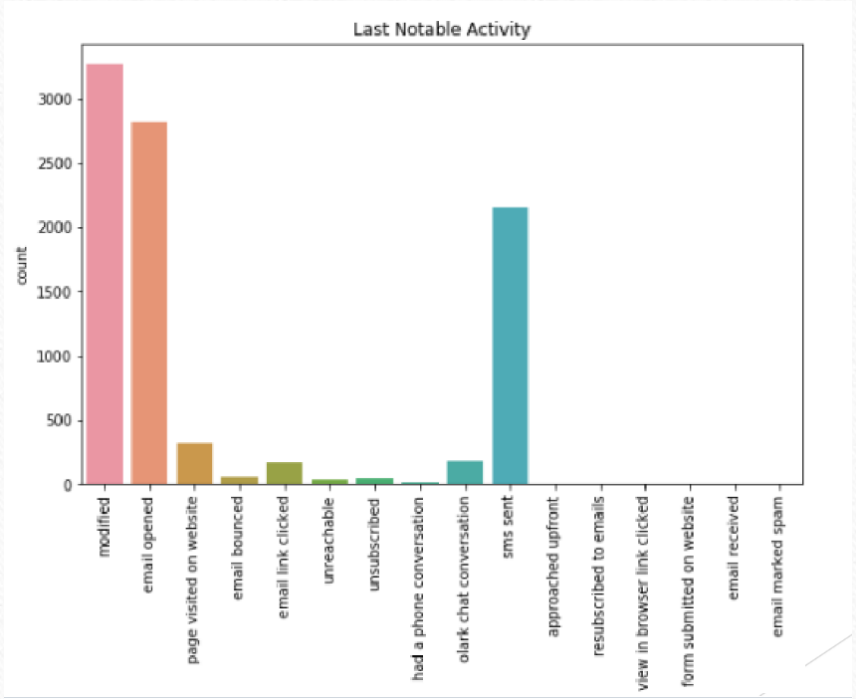
Better career prospects is the factor involved with choosing a course.



Search, Newspaper article, X education forums, Newspaper does not add any value to the data.

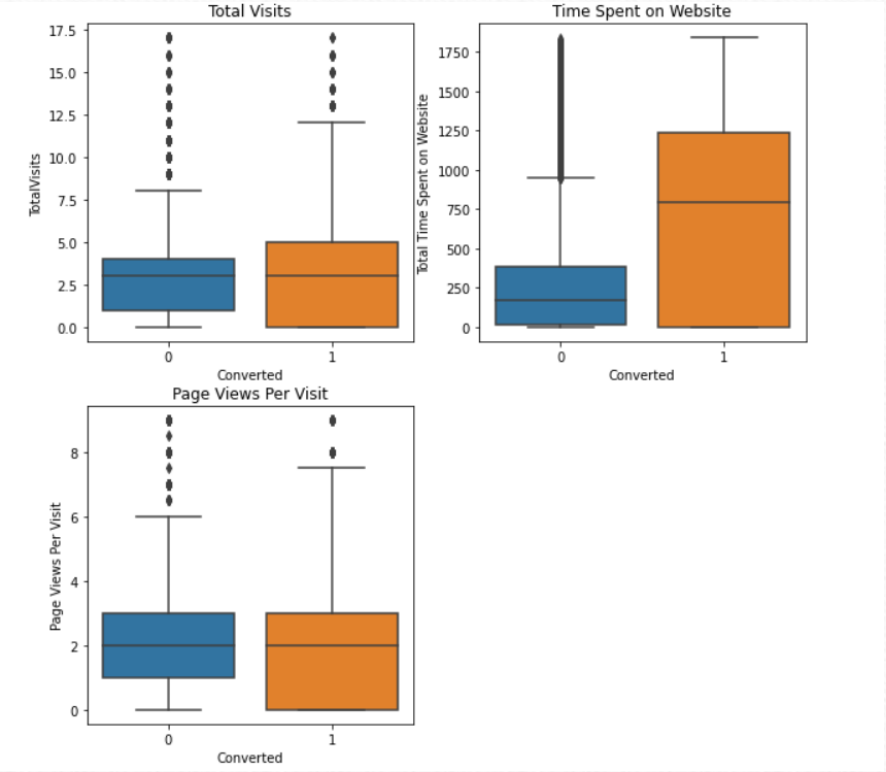


The last notable activity is email opened , SMS sent that has high conversion rate.



BIVARIATE ANALYSIS

- From this we can say that the conversion rate is more the category where people spend more time on website.
- The probability of total visits is between 15-20. but the average total visit between converted and non converted is found to be same.
- The max probability of page views per visit is 3-5 but average for converted and non converted is found to be same.



DATA PROCESSING FOR MODELLING

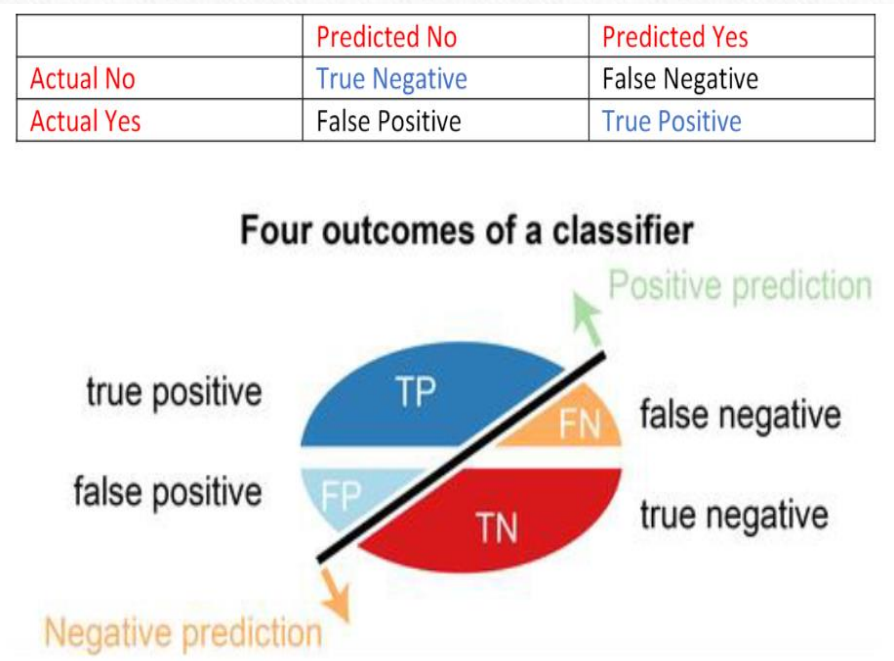
- **Numerical variables are normalised.**
- **Dummy variables were created for object type variable.**
- **Total rows for analysis: 9074**
- **Total columns for analysis: 78**

TRAIN/TEST SPLIT

- **After splitting the data into train and test by 70:30 ratio, use RFE as feature selection with MinMaxScaler with 15 variables as output.**
- **Model building by removing the variable whose p value > 0.05 and VIF > 5.**
- **Then we have performed predictions on test data.**
- **The overall accuracy attained is 81.6%.**

CONFUSION METRICS

	Predicted No	Predicted Yes
Actual No	True Negative	False Negative
Actual Yes	False Positive	True Positive




```
TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives
```

```
#sensitivity

TP/(TP+FN)
```

```
0.6901052631578948
```

```
# Calculating the specificity
TN/(TN+FP)
```

```
0.8913480885311871
```

Recall = True Positives/(True Positives +False Negatives)

Precision = True Positive/ (True Positives +False Positives)

Recall

TP / TP + FN

```
TP/(TP+FN)
```

```
0.6901052631578948
```

Precision

TP / TP + FP

```
##### Precision
TP / (TP + FP)
```

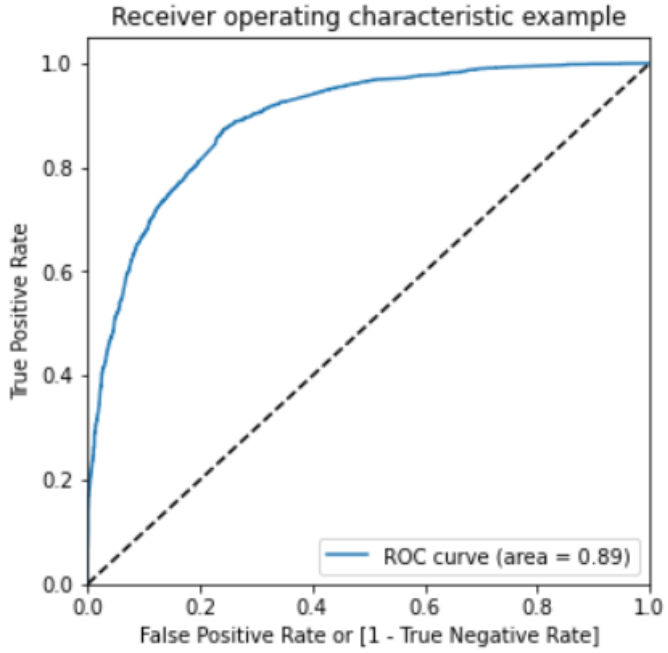
```
0.791405118300338
```

FINDING THE OPTIMAL CUT OFF

ROC curve represents how much the model is able to distinguish between the classes.

The ROC curve is 0.89 which is a good value.

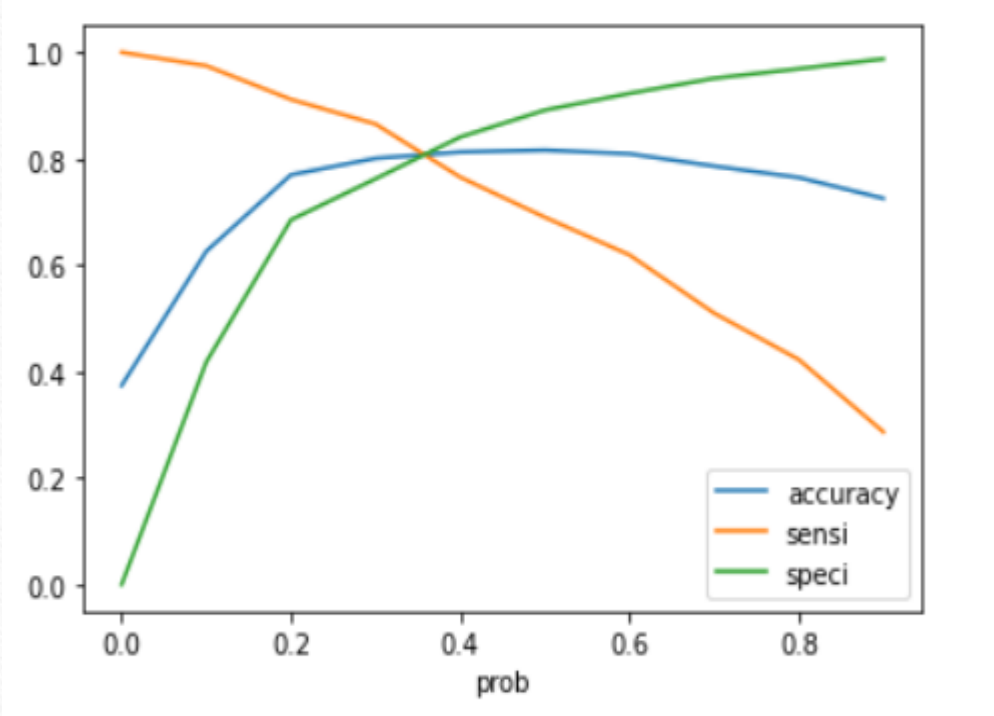
AUC is found to be 0.89 which states that the model is quite stable.



The are under the RoC curve is 0.89 which is a good value.

The optimal cut off probability is that probability where we get balanced sensitivity and specificity.

The optimal probability is 0.35.



- We can conclude that the model 5 with 0.35 as a cut-off is delivering the Recall value of 78.7% on training dataset and 80% on the test dataset. This can be considered as a reasonable performance.
- With the obtained cut off we got an accuracy of 80%, specificity of 81% and sensitivity of 79% for train data.
- For the test data we got the results of Accuracy 80%, specificity 82% and sensitivity 80%.

GENERATION OF SCORE COLUMN

- **In order to help the company we have used their past data and implemented a machine learning model to calculate the scores of the leads. The scores are in the range of 0-100, suppose the lead has a higher score, that means that they are more likely to purchase a course from them.**
- **So, whenever lead data comes in, they can find out the score using the model and understand the potential of the lead.**

CONCLUSION

- **Company should focus on the following features to increase the leads:**

  (i) what is the lead source : google, direct traffic, organic search , welingak website.

  (ii) the total number of visits and total time spent on website.

  (iii) when was the last activity either on SMS or chat conversation

  (iv) when the lead origin is lead add format.

  (v) when they have working profession.

by keeping these things in mind the company can convert more customers into leads and hence gain profit from it.

…The end…