1) **A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.**

Overall Summary:

Problem Statement:

An education company named X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.

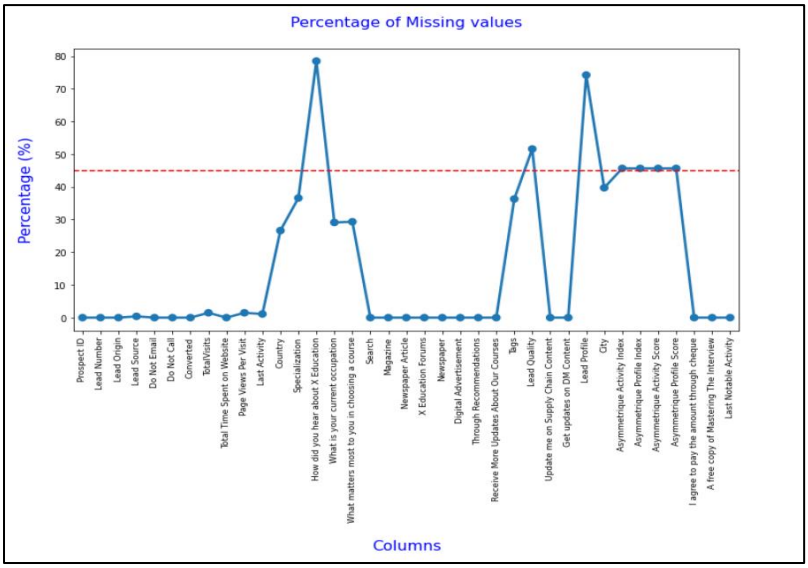
Objective:

- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The following are the steps used:

1. **Cleaning data:**

- Dropped columns that are having high percentage of missing values (cut off taken as 45%).



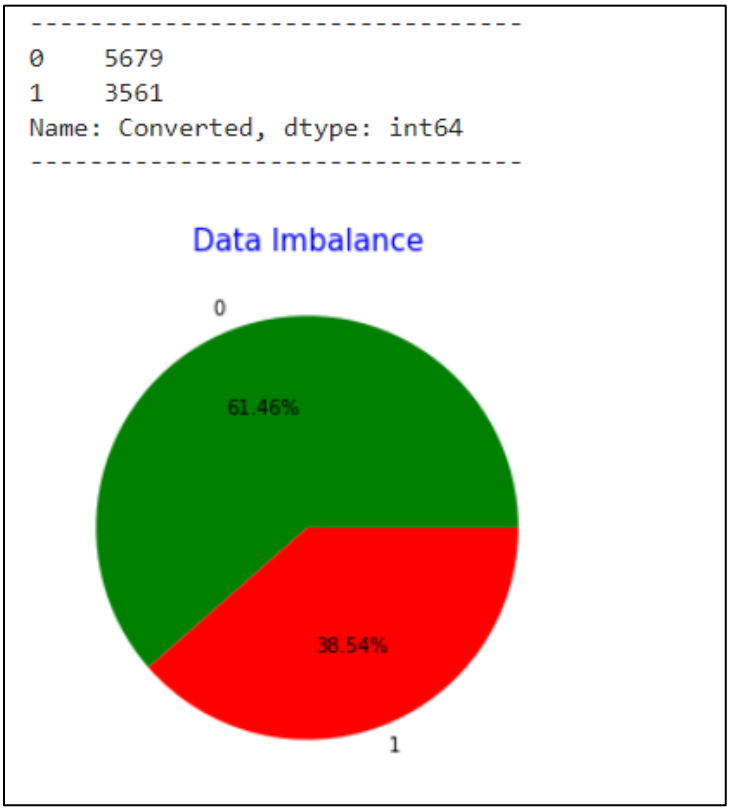
- We have checked data types, shape, info and summary of the loaded dataset.
- Check the number of unique categories in each categorical column. Here based on the skewness of the data we have dropped the columns.

Column Name	No. of Unique values	No. of Null values	Null Percentage
City	6	3669	39.71
Specialization	18	3380	36.58
What matters most to you in choosing a course	3	2709	29.32
What is your current occupation	6	2690	29.11
Country	38	2461	26.63
Lead Source	20	36	0.39
Lead Origin	5	0	0.00
Digital Advertisement	2	0	0.00
I agree to pay the amount through cheque	1	0	0.00
Get updates on DM Content	1	0	0.00
Update me on Supply Chain Content	1	0	0.00
Receive More Updates About Our Courses	1	0	0.00
Through Recommendations	2	0	0.00
Newspaper Article	2	0	0.00
Newspaper	2	0	0.00
X Education Forums	2	0	0.00
Magazine	1	0	0.00
Search	2	0	0.00
Do Not Call	2	0	0.00
Do Not Email	2	0	0.00

- For the columns with less percentage of missing in categorical columns, we have imputed the value with mode or with some other categorical variable like unknow, others value etc.
- For the columns with less percentage of missing in numerical columns, we have imputed the value with median (as those contains outliers, we have not imputed it with mean).
- Check the redundancy and quality of data.
- Check the null percentage row wise.
- Finally check the percentage of rows retained in data cleaning process.

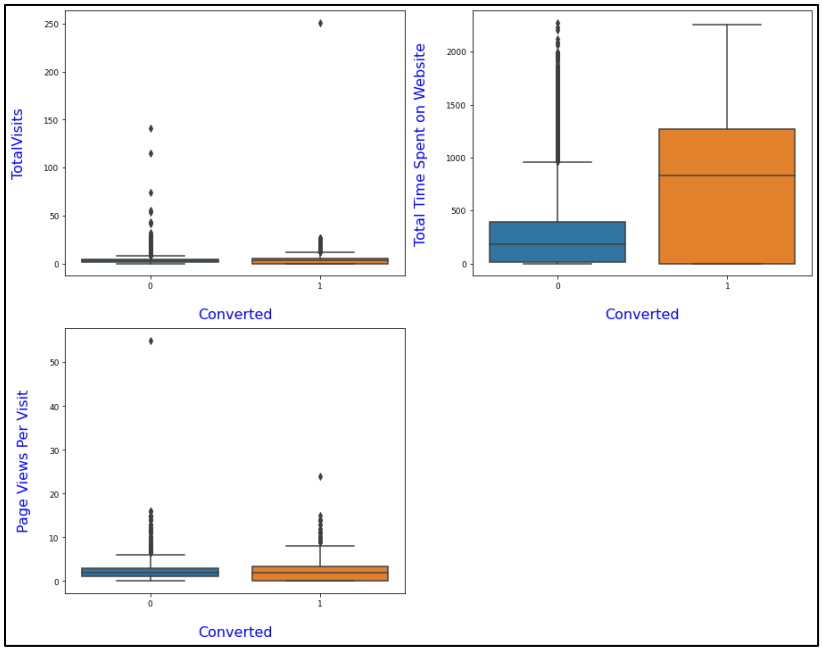
2. Data Imbalance Check:

- We have performed a data imbalance check on Target column. Then we found that the is almost balance based on Target: Converted column. Almost 62% is marked as ‘Not Converted’ and 39% is marked as ‘Converted’.



3. **EDA (Exploratory Data Analysis):**

- We have done outlier analysis on continuous columns and will be treating it later.



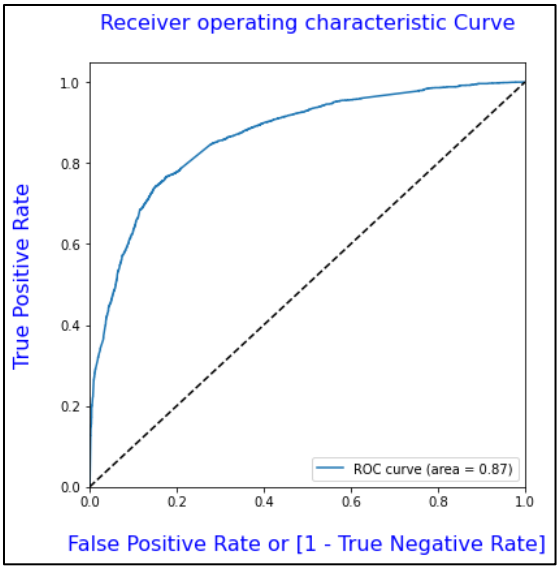
- We have done some Univariate analysis (Based on Target: Converted column)
- We have done some Bivariate Analysis (Based on Target: Converted column)

4. **Data Preparation:**

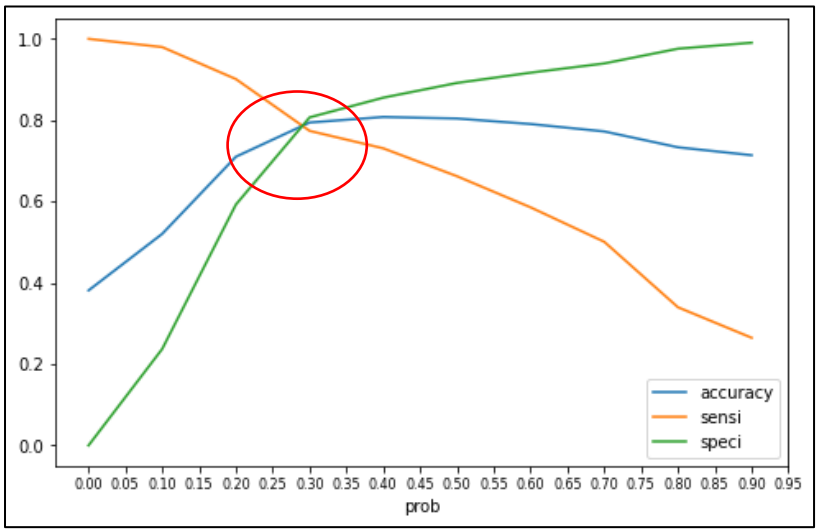
- We have treated Outliers that we found in earlier steps.
- Create dummies for all categorical columns.
- Perform train-test split (70% train set and 30% test set).
- Perform scaling using standard scaler.

5. **Perform Modelling:**

- Use techniques like RFE to perform variable selection (Total 20 features selected using RFE). Then perform manual elimination based on p-value (<0.05) and VIF (<5). Finally selected 12 features.
- Build a Logistic Regression model based on 12 features.
- Build the AUC ROC curve and find the area under the curve which is 0.87 in our case on train dataset.



- Find the specificity, sensitivity and Accuracy curve based on confusion matrix and then select the optimal cutoff threshold from the curve intersection point of the above different metrics. Which is 0.288 in our case based on train dataset.



- Find the different train model evaluation metrics like sensitivity/Recall, precession, Accuracy etc. The recall score in our case is (0.78061) 78%.

```
------------------------------------------------------------------
The Recall score of the model is 0.7806163828061639
The Specificity score of the model is  0.796351824087956
The Precision score of the model is  0.7025547445255474
The False positive rate of the model is  0.20364817591204398
The Negative predicted value of the model is  0.8548819742489271
------------------------------------------------------------------
```

- Now predict the probability based on test data.

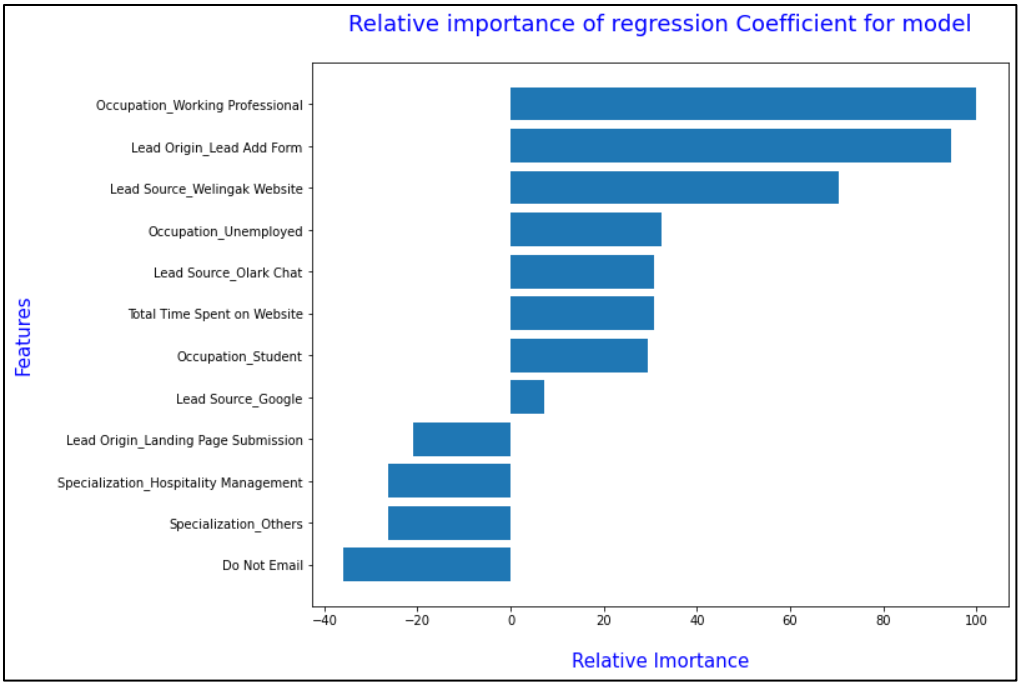
- Check the model performance metrics like sensitivity/Recall, precession, Accuracy etc. over the test dataset. The recall score in our case for test data set is (0.781735) 78% test data.

```
------------------------------------------------------------
The Recall score of the model is 0.7817351598173516
The Specificity score of the model is  0.7853309481216458
The Precision score of the model is  0.7039473684210527
The False positive rate of the model is  0.2146690518783542
The Negative predicted value of the model is  0.846401028277635
------------------------------------------------------------
```

- Finally Generate the lead score variable on test model.

6. It was found that the variables that mattered the most based on which the leads are most likely to convert into paying customers (In descending order):

- Occupation_Working Professional
- Lead Origin_Lead Add Form
- Lead Source_Welingak Website
- Occupation_Unemployed
- Lead Source_Olark Chat
- Total Time Spent on Website
- Occupation_Student
- Lead Source_Google
- Lead Origin_Landing Page Submission
- Specialization_Others
- Specialization_Hospitality Management
- Do Not Email



Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.

END