LEAD SCORE CASE STUDY

PRESENTED BY-
AVANNI GUDIMETLA
VAISHALI RAMACHANDRAN

PROBLEM STATEMENT

X Education sells online courses to industry professionals. They gets a lot of leads, however, their lead conversion rate is very poor.
For example, if they acquire 100 leads in a day, only about 30 of them are converted i.e, the lead conversion rate is only around 30%

To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.

If they successfully identify the ‘Hot Leads’, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than contacting everyone.

Business Objective:
X education wants to know the most promising leads and hence we need to build a Model which identifies the hot leads.

We 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.

PROBLEM SOLVING METHODOLOGY

We follow the below approach and steps to this problem :

- Data Cleaning
  - ✓ Handling duplicates in the dataset
  - ✓ Handing null values (NaN) in the dataset
  - ✓ Imputing the missing rows in the dataset
  - ✓ Dropping unwanted rows and columns.
  - ✓ Handling outliers

- Data Analysis using Exploratory Data Analysis (EDA)
  - ✓ Univariate data analysis by identifying the distribution of variables and checking value counts
  - ✓ Bivariate data analysis by identifying the correlation between the variables

- creating dummies for all categorical variables
- Feature scaling and data encoding
- Model Creating using the logistic regression classification technique
- Making Predictions
- Validation of the model
- Model Presentation
- Conclusions and Recommendations using the insights gathered.

DATA CLEANING

The main objective of this step is to clean and convert the dataset to a more presentable form by getting rid of null values and outliers in the data.

In our dataset, we observe that there are several columns which have Null values in them. We have dropped all those columns which have more than 30% of null values for easier analysis.

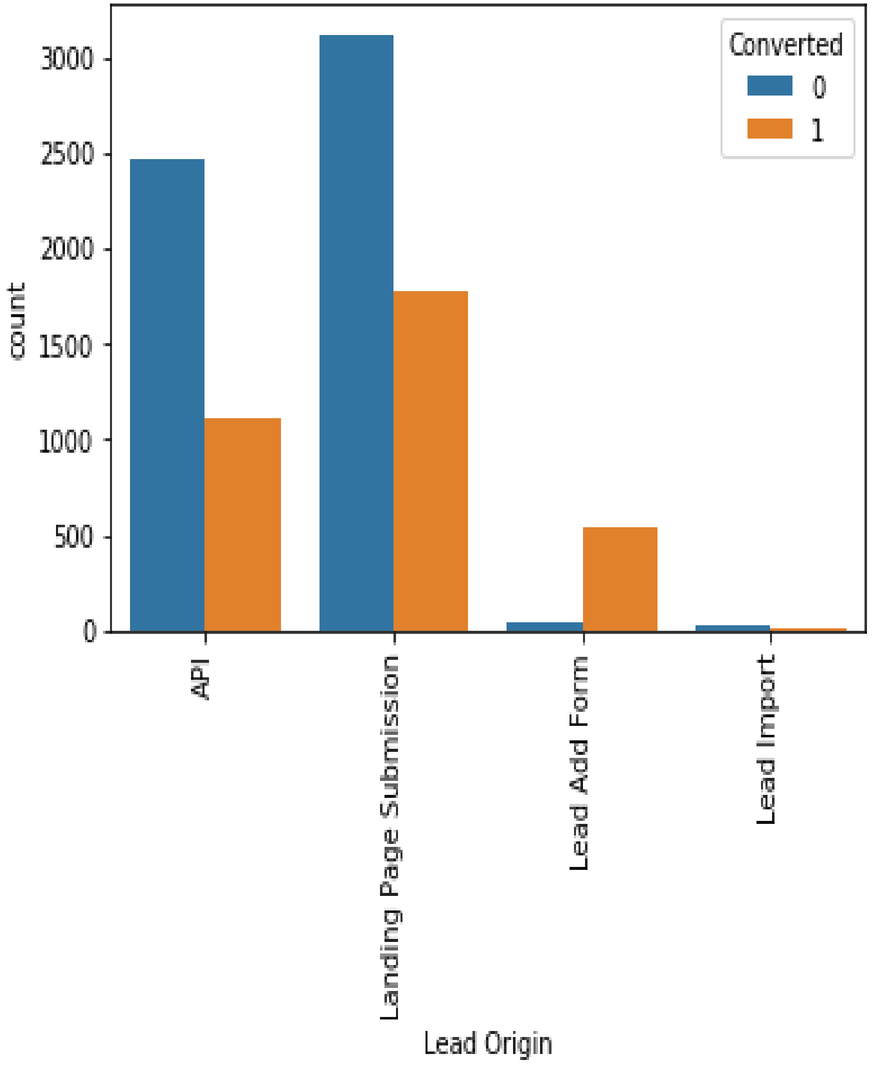
Also, we observe that some of the columns such as Specialization, Country, Current Occupation etc. have some values called ‘Select’ which indicate that the prospective customer has not made any selection and hence we treat these values also as Nulls.

We have also dropped some of the columns which do not show enough variance as well as those columns which have only one unique value

The remaining missing column values have been imputed with the values that is most frequent in that column.

EXPLORATORY DATA ANALYSIS

```
(array([0, 1, 2, 3]), <a list of 4 Text xticklabel objects>)
```



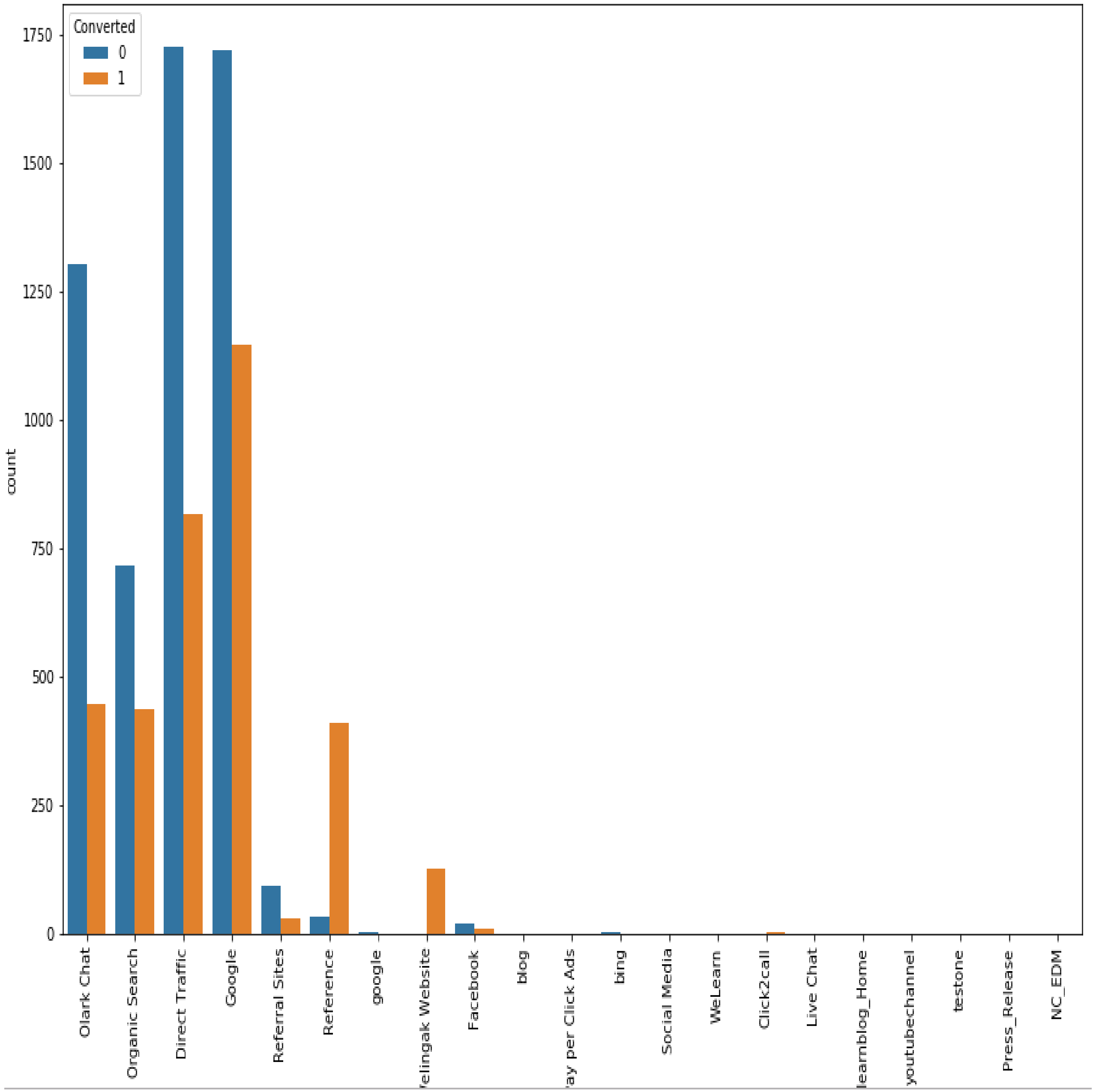
LEAD ORIGIN :

→ Landing Page Submission and API have very high number of leads but in comparison, the count of converted leads is quite less

→ Lead Add Form has very less count of leads, but among that, the number of converted leads is very significant

→ Number of converted leads from Lead Import is pretty negligible

→ To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission and generate more leads from Lead Add Form.



LEAD SOURCE:

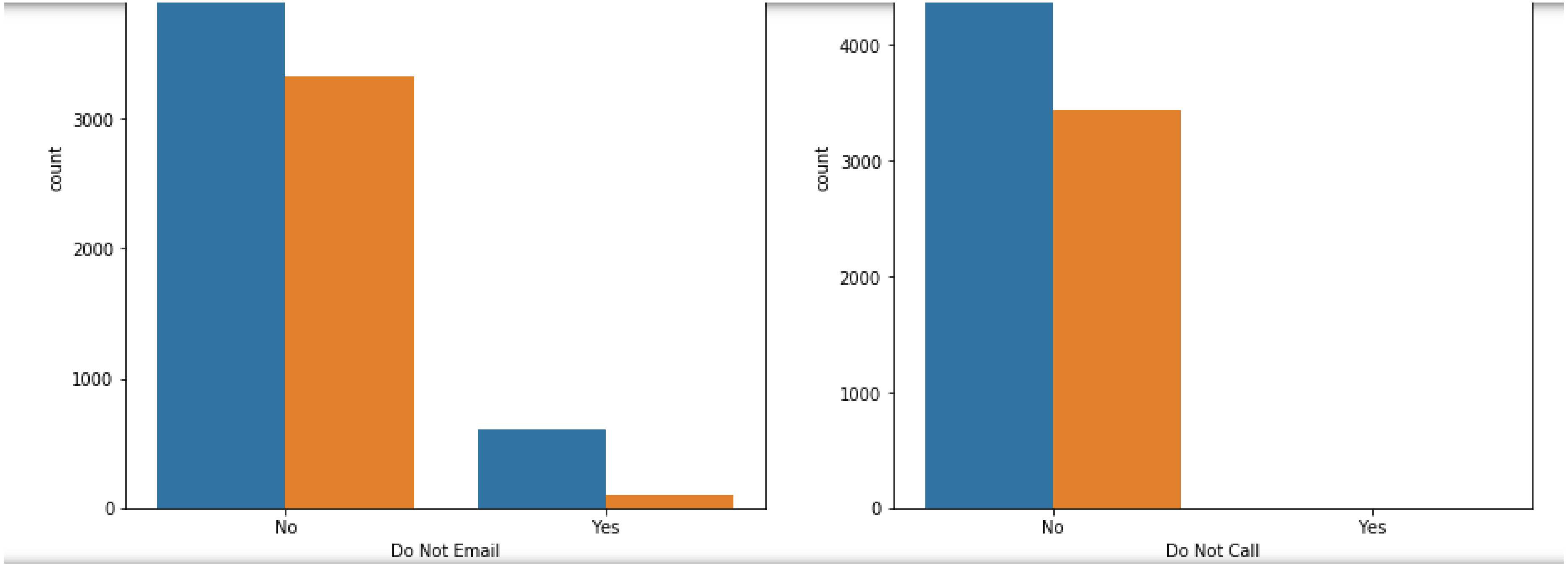
→ Google and Direct traffic generates maximum number of leads.

→ Conversion of leads through welingak website is high. X Education can focus on more advertising and marketing in this websites as it could lead to a lot of conversions

→To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

→ Reference does best in conversion of leads in comparison with its count of leads

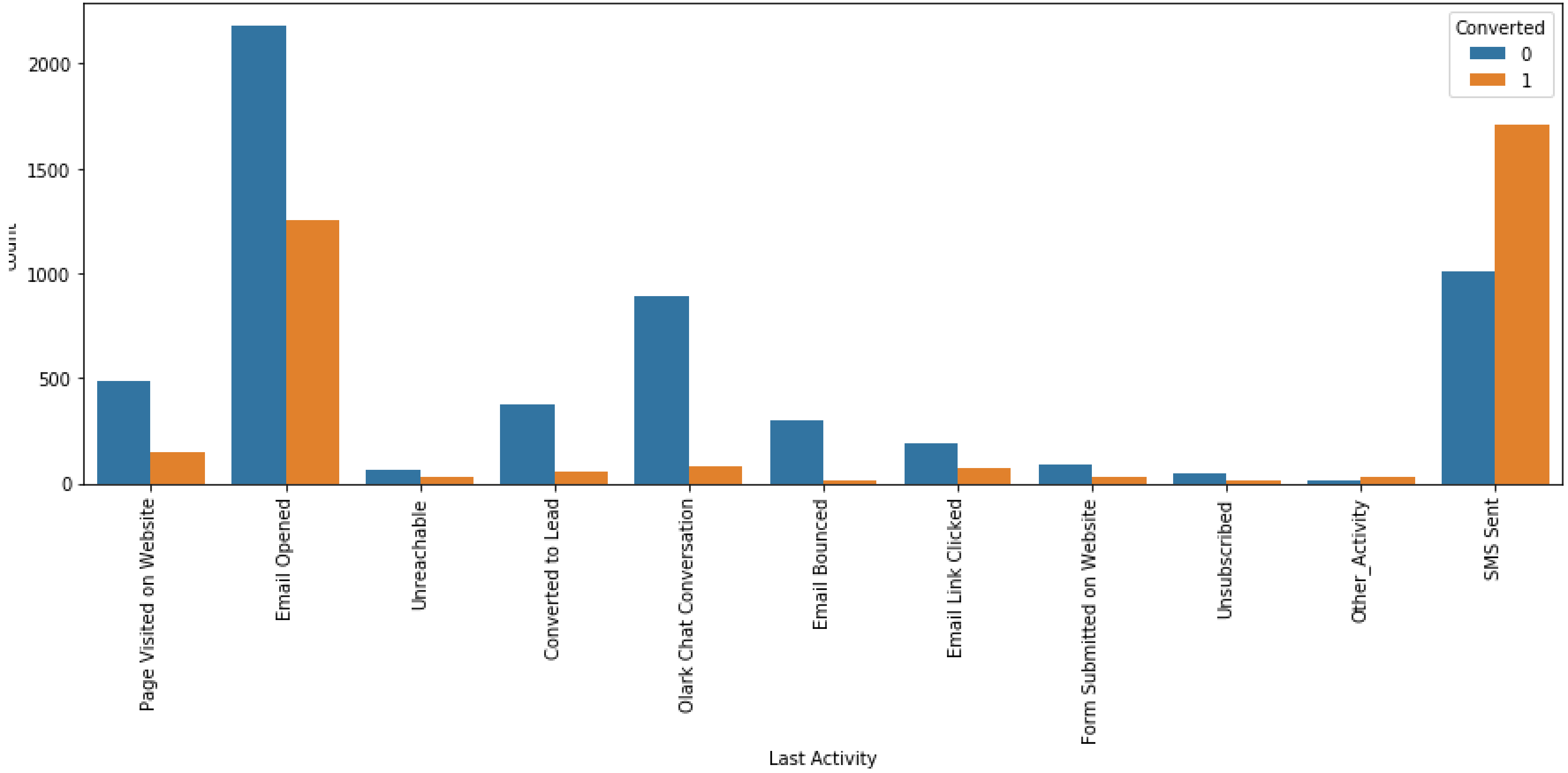
→ More focus can be given to references as it could lead to potential ‘Hot Leads’.



Do not Email - An indicator variable selected by the customer wherein they select whether of not they want to be emailed about the course or not.
Do not Call - An indicator variable selected by the customer wherein they select whether of not they want to be called about the course or not.

We see that majority of them have opted for No for both calling and emailing them about courses and we see that inspite of not opting for calls and emails, they have chosen the course.
However, those who have opted for calls and emails have rarely been converted into leads.

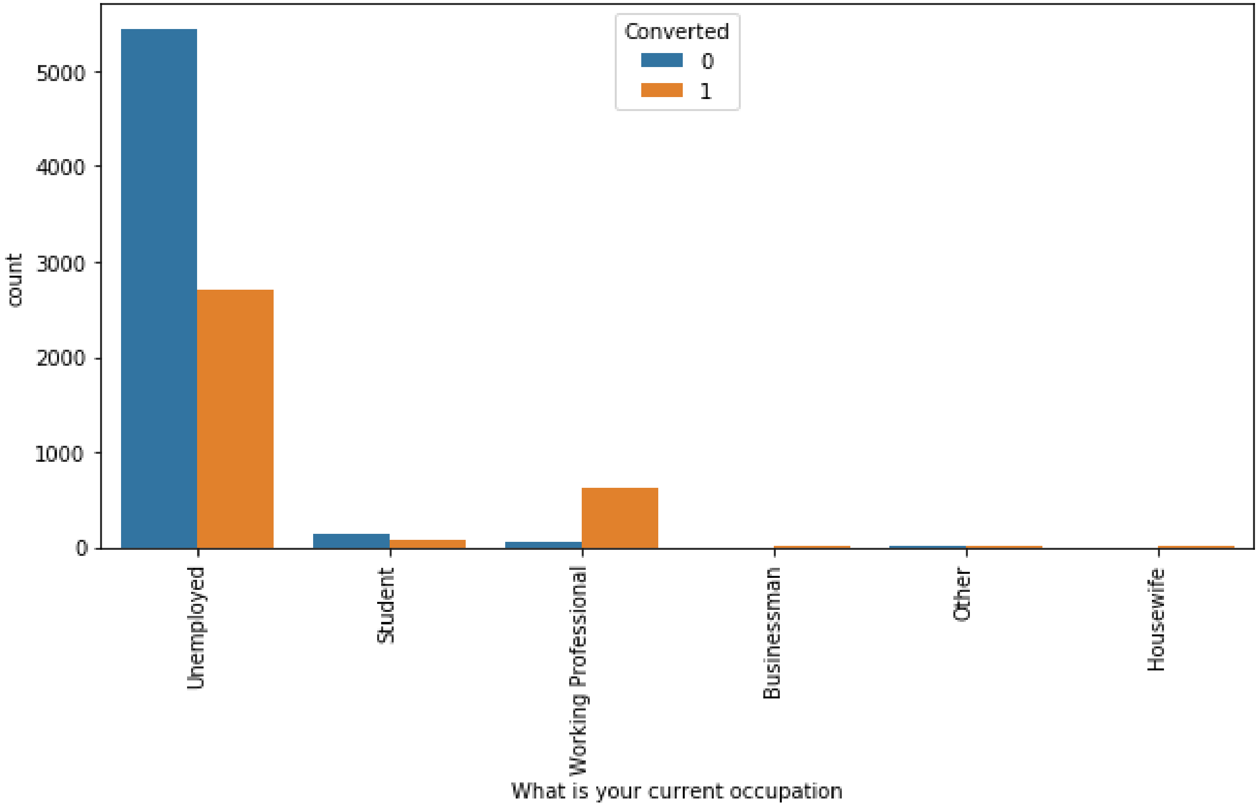
It would hence be better to use other modes of promoting the course and avoid calls and emails.



LAST ACTIVITY:

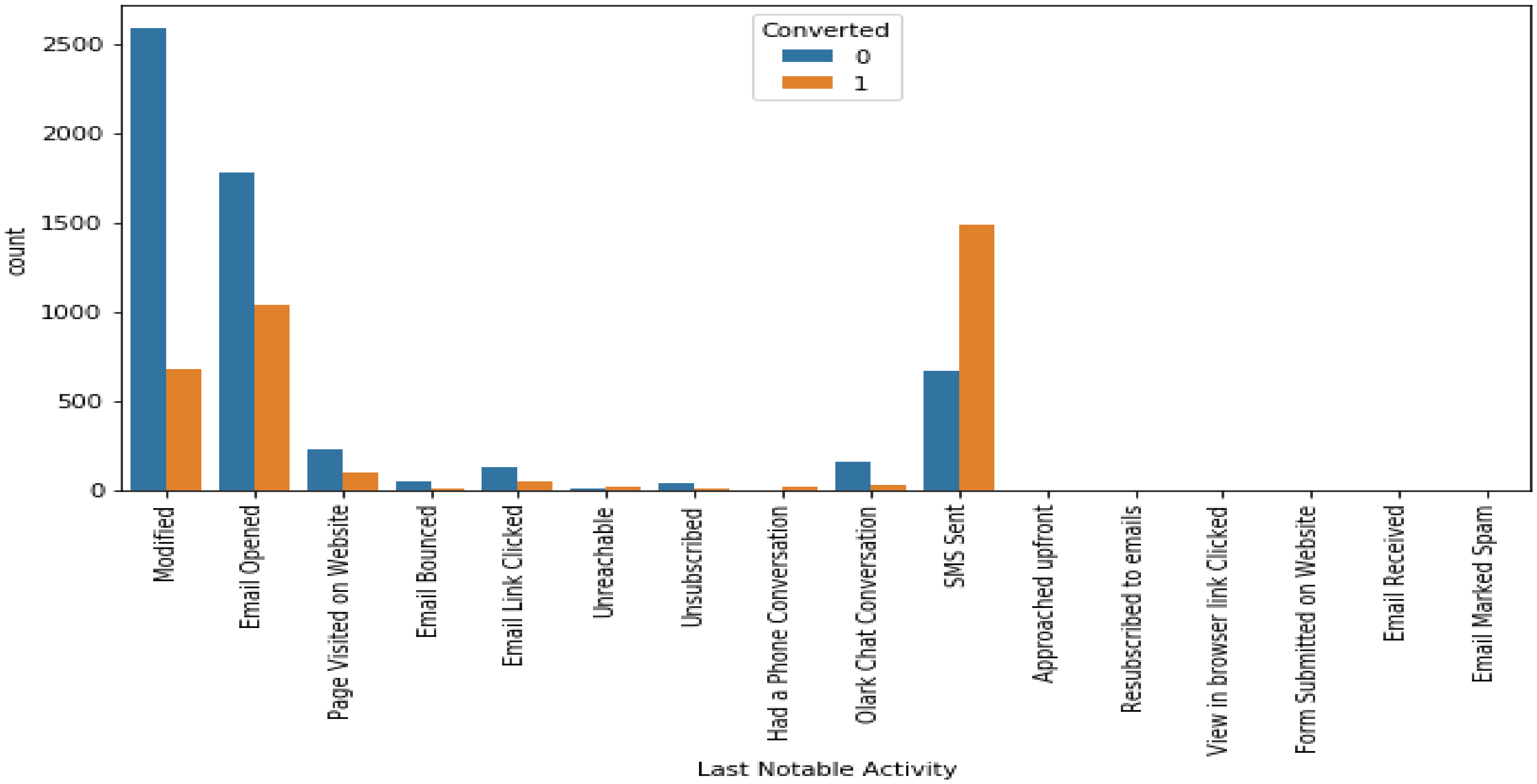
The Last activity performed by most users is opening their emails and we see that the conversion rate here could definitely be improved by working towards making the emails look more visually appealing to the customers and including relevant data which catches the eye. This would result in more conversions.

On the other hand, we see that a number of customers for whom ‘SMS Sent’ is the last activity , the conversion rates are pretty good and for such customers we can keep pushing promotional SMSes



CURRENT OCCUPATION :

We see that most of the people who are leads are mostly ‘Unemployed’ .However, the number of converted leads is very less compared to the count and this can be improved by offering tailor-made courses to suit their needs and conducting market surveys .
We also see that working professionals have a good chance of joining this course.



The Last notable activity performed by any student is Modified but the maximum conversions happen to those leads whose last notable activity is

MODEL BUILDING

DATA PREPARATION FOR MODEL BUILDING :

This step is a prerequisite to model building and once we have the clean dataset we start with binary encoding for categorical variables which have only one of the two values ‘Yes ‘ and ‘No’ and we convert those to 1 and 0.

After this, we create dummy variables for all the other categorical variables and after creating dummies, we then drop it from the main dataframe and merge the dummies with the main dataframe.

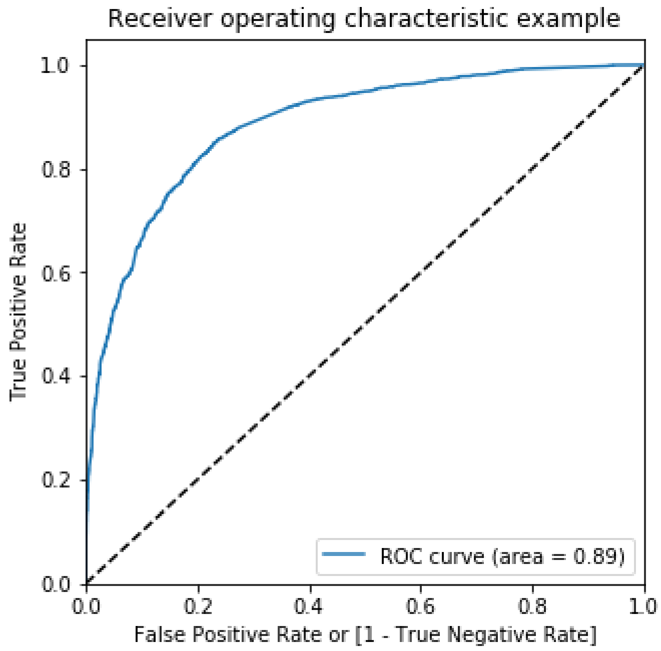
FEATURE SELECTION :

Stats Model and RFE used to identify Model features.
VIF was used to validate features as shown.
We created ROC curve to find the strength of the model

STEPS FOLLOWED IN MODEL BUILDING

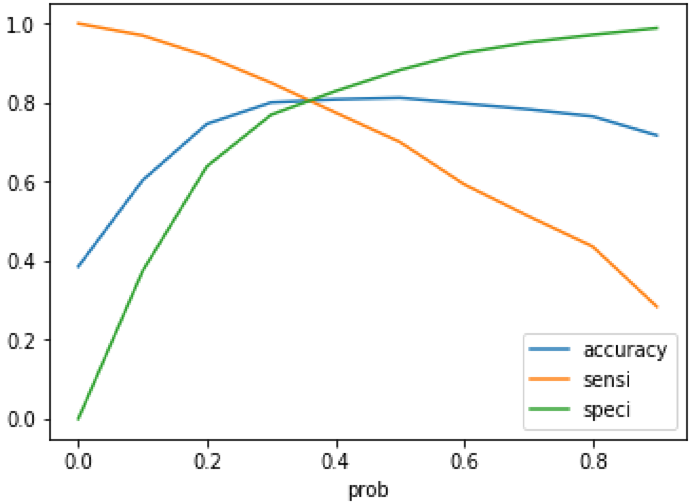
1. Convert the data set into train and test sets.
2. Check the correlation among the variables by plotting a heatmap.
3. Perform scaling to bring all the variables onto a common scale.
4. Build the logistic regression model
5. Select the features using Recursive Feature Elimination (RFE) method.
6. observe the Generalized Linear Regression Model (GLM) results.
7. Drop the unwanted columns and repeat the above step.
8. Check the Confusion Matrix and the overall model accuracy.
9. Check VIF and drop any unwanted columns. Repeat above steps as required.
10. Compute the metrics for Sensitivity, Specificity, False Positive Rate, Positive Predicted value and Negative Predicted Value.
11. Plot the ROC curve
12. Find the optimal cut-off point and calculate accuracy, sensitivity and specificity for the various probability cut-offs.
13. Assign lead score to each customer
14. Compute the Precision and Recall and Precision and Tradeoff Recall
15. Make predictions on the test set

PLOTTING THE ROC CURVE



This curve shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

Figure shows the area under the curve is around 0.9



Optimal cut-off probability is that probability where we get balanced sensitivity and specificity.

Accuracy, Sensitivity and Specificity are here plotted for various probabilities.

In our graph, the value is 0.4

SUMMARY

We found out that 0.4 is the minimal cut offs.

Lead Score between 0-100 is assigned to each of the customer to check the hot leads.

- ➢ The overall accuracy of the model is 80 % as was the objective of X Education’s CEO
- ➢ The sensitivity of the logistic regression model is 74%
- ➢The specificity of the model was 83%
- ➢ The precision metric of the model was 78 %.