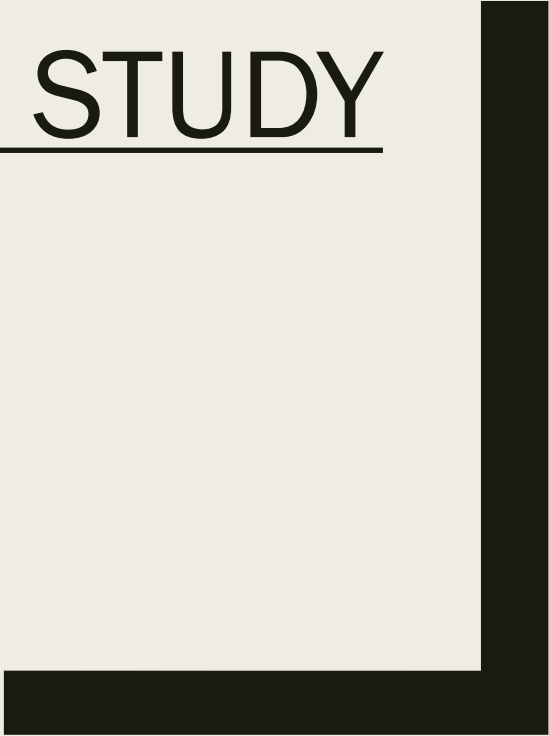# LEAD SCORING CASE STUDY

By : – Priyanka Gangwar

Pravin Kumar

Priya Dubey

(DS – C52)

Problem Statement

An EdTech company named as X Education needs help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us 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 CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Business Goals:

- Building a Logistic Regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

- CEO of company has given an approximate target lead conversion rate to be around 80%.

Approach

➤ Data Understanding

➤ Data Cleaning

  - *Identifying and treating missing values*
  - *Identifying non-relevant columns*
  - *Outlier analysis on numerical variables*

➤ Exploratory Data analysis : Univariate, Bivariate and multivariate

➤ Data Preparation

  - *Dummy Value creation*
  - *Feature scaling of numerical variables*

➤ Splitting of Train and Test data set

➤ Logistic regression Model Building

➤ Model Evaluation

➤ Conclusions

Missing values in Leads data set
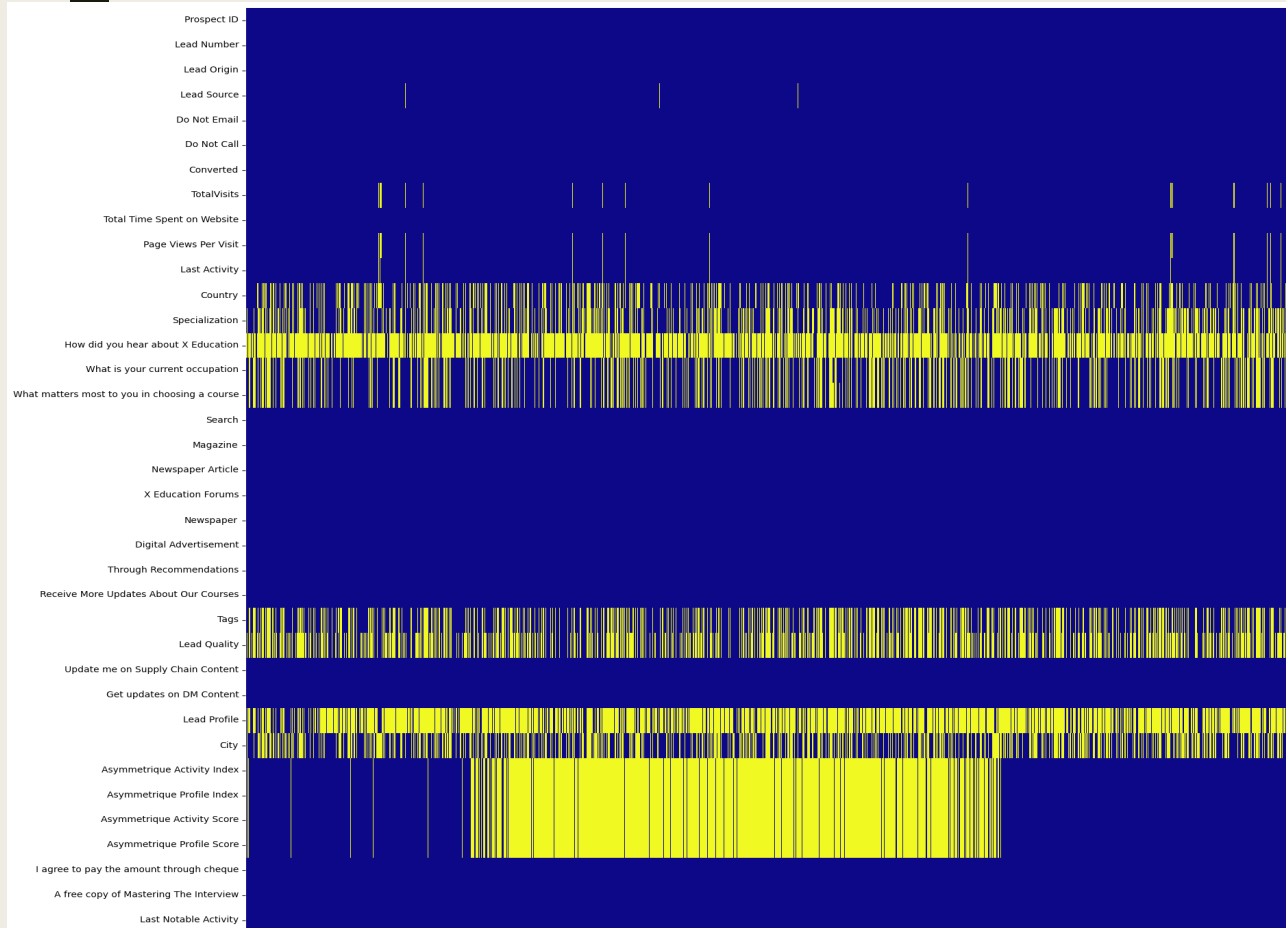

Top variables with missing values

```
In [12]: (leads_df.isnull().sum()/leads_df.shape[0]*100).sort_values(ascending=False)

Out[12]: How did you hear about X Education          78.463203
         Lead Profile                                74.188312
         Lead Quality                                51.590909
         Asymmetrique Profile Score                  45.649351
         Asymmetrique Activity Score                 45.649351
         Asymmetrique Activity Index                 45.649351
         Asymmetrique Profile Index                  45.649351
         City                                        39.707792
         Specialization                              36.580087
         Tags                                        36.287879
         What matters most to you in choosing a course  29.318182
         What is your current occupation             29.112554
         Country                                     26.634199
         Page Views Per Visit                         1.482684
         TotalVisits                                  1.482684
         Last Activity                                1.114719
         Lead Source                                  0.389610
         Receive More Updates About Our Courses       0.000000
         I agree to pay the amount through cheque     0.000000
```

Numerical Attribute Analysis

```
In [174]: leads_df["Converted"].value_counts()

Out[174]: 0    5581
          1    3423
          Name: Converted, dtype: int64
```
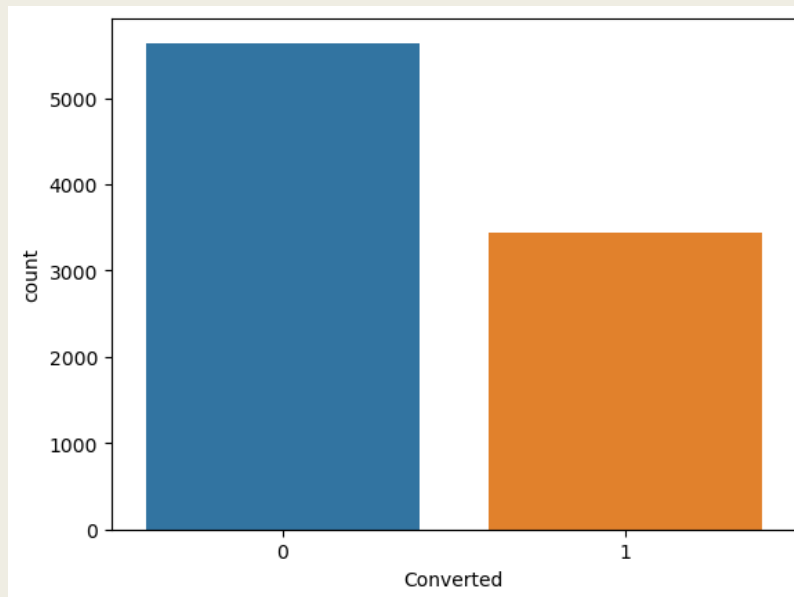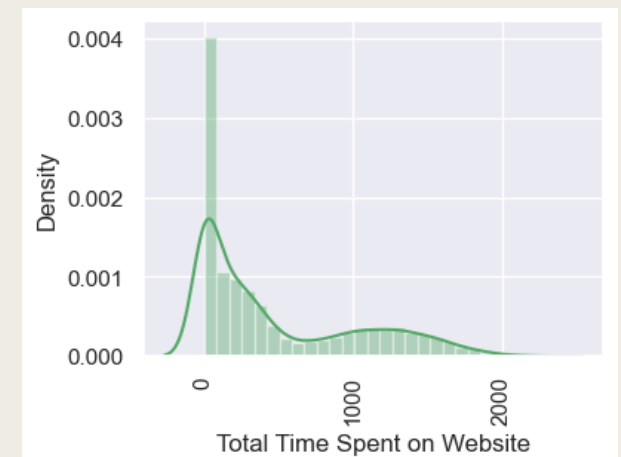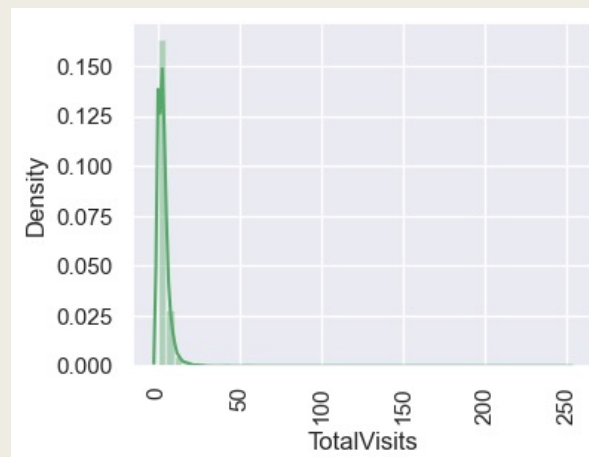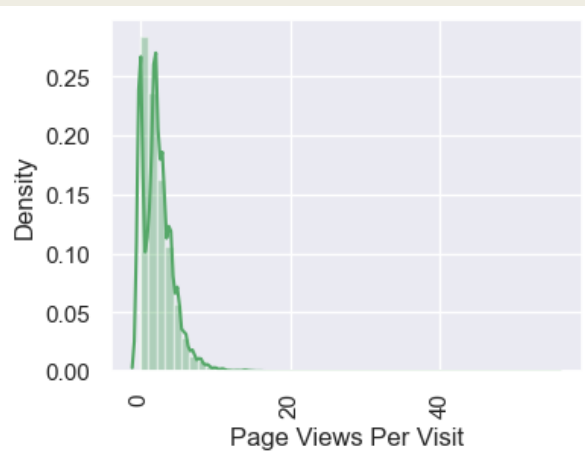

- ✓ From the data set out of approximate 9000 leads there are 3423 leads which are Converted into clients while 5581 are not converted.
- ✓ The actual conversion ratio is approximate **38%**.

Contd….

Numerical Attributes Analysis

- <u>TotalVisits</u> - This variable indicates number of visits by a Lead, most of its value are ranging between 0 and 10 with few outlier values.
- <u>Total Time Spent on Website</u> - It indicates total time spent by a Lead on X education website. The dist plot for this variable is spread 0 till 2000 with peak being at 0.
- <u>Page Views Per Visit</u> - This gives an average page views by a Lead during the visits. This value is mostly ranging from 0 to 8 with two peaks at 0 and 2.

Outlier Analysis

Distribution of TotalVisits

Distribution of Total Time Spent on Website

Distribution of Page Views Per Visit
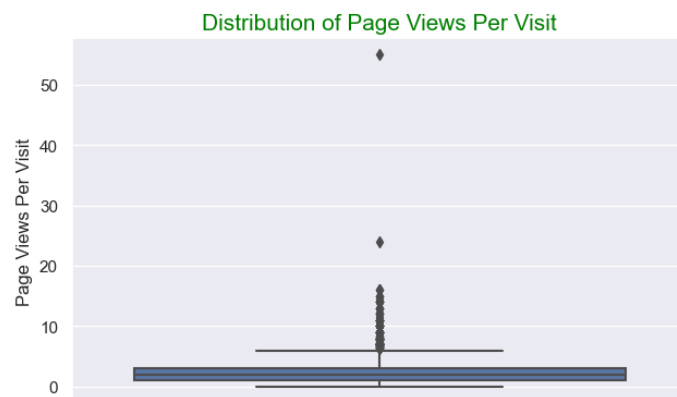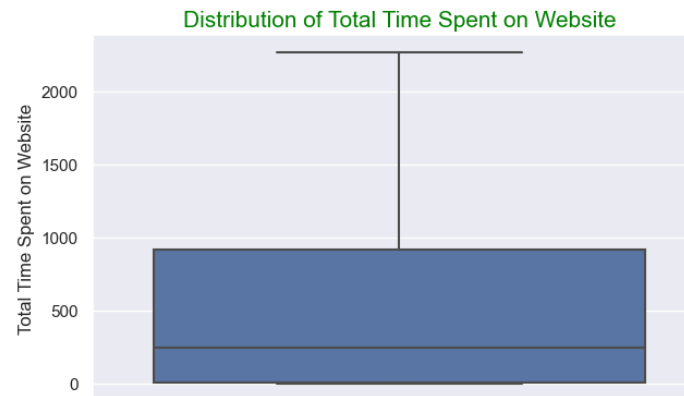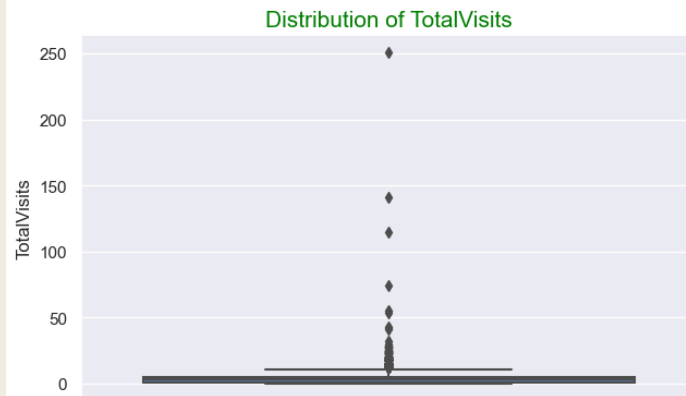
- **Total Visits** : As we can see the 75% quantile of this feature is up to 5 whereas outlier value ranges from 50 and above. So we will treat the outlier by capping the value to 99%.

- **Total Time Spent on Website** : No outlier treatment is required.

- **Page Views Per Visit** : This feature have quite few outliers, so we will capping the outliers to 99% value for analysis.

Categorical Attributes Analysis
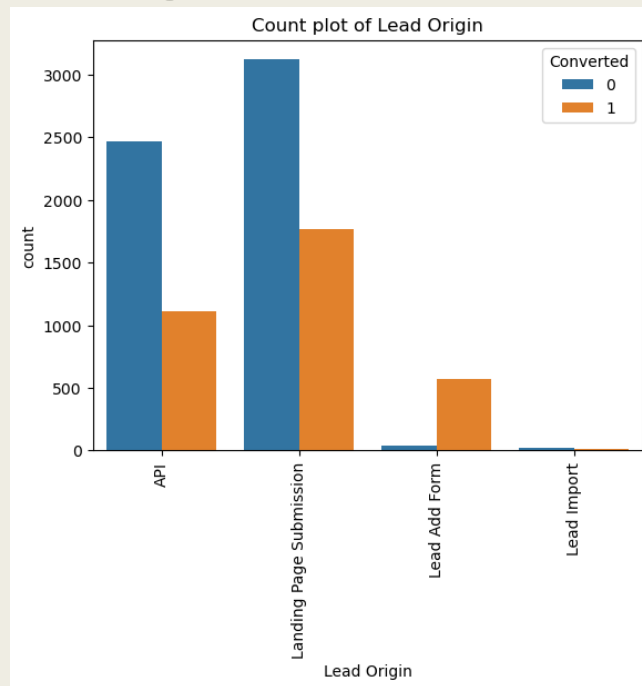

Lead Origin vs Converted
Maximum incoming and converted leads are from Landing Page submission. To improve the lead conversion rate, we have to focus more on leads coming from 'API' and 'Landing Page Submission'.
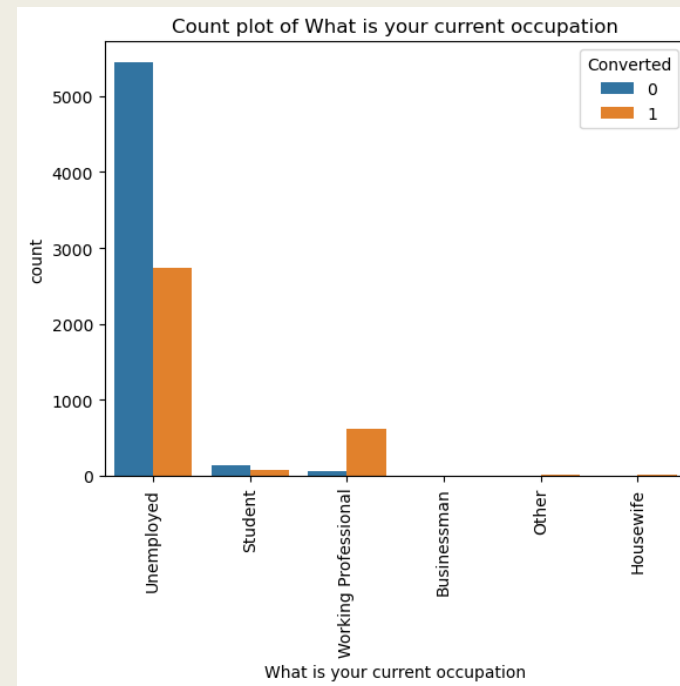
Current Occupation vs Converted
Most of the leads are obtained from "Unemployed" section,
however conversion rate for "Working Professional" is quite higher.

Count plot of Lead Source

Count plot of Last Activity
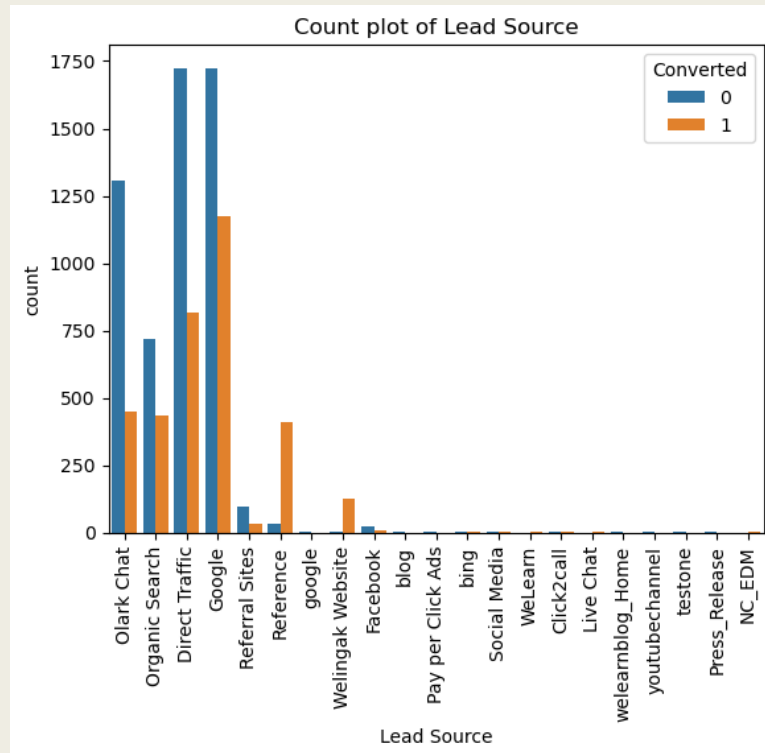
Lead Source vs Converted
Maximum number of leads are obtained by 'Google' and 'Direct Traffic' with more conversion rate for 'Google'. Conversion rate Reference and Welingak website is quite high.
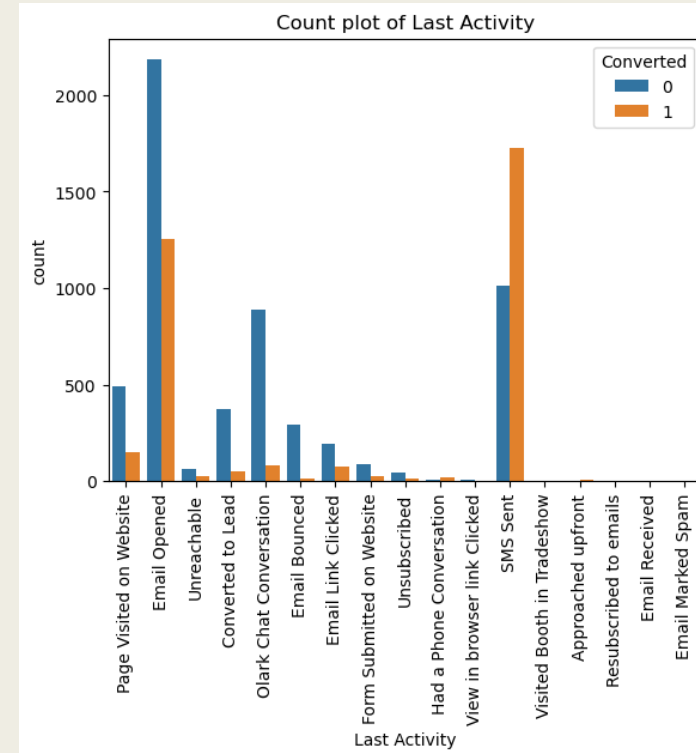
Last Activity vs Converted
Most number of leads obtained in from "EMail Opened" and "SMS Sent" categories. Conversion rate for "SMS Sent" is higher than any of the categories.

Count plot of Tags

Count plot of Lead Quality
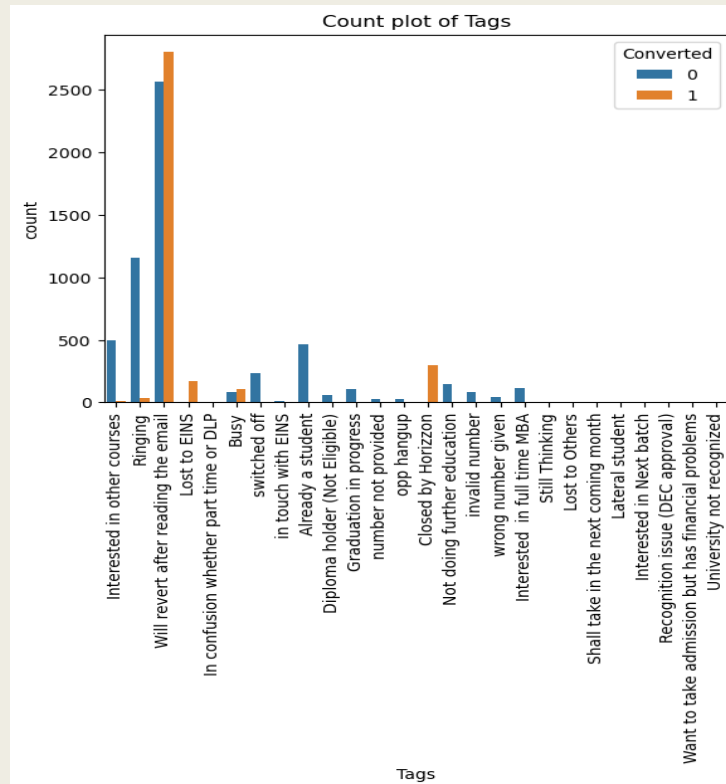
Tags vs Converted
This column signify the current status of leads and most of the leads have status as 'Will revert after reading the email', also conversion rate of this status is highest than other categories.
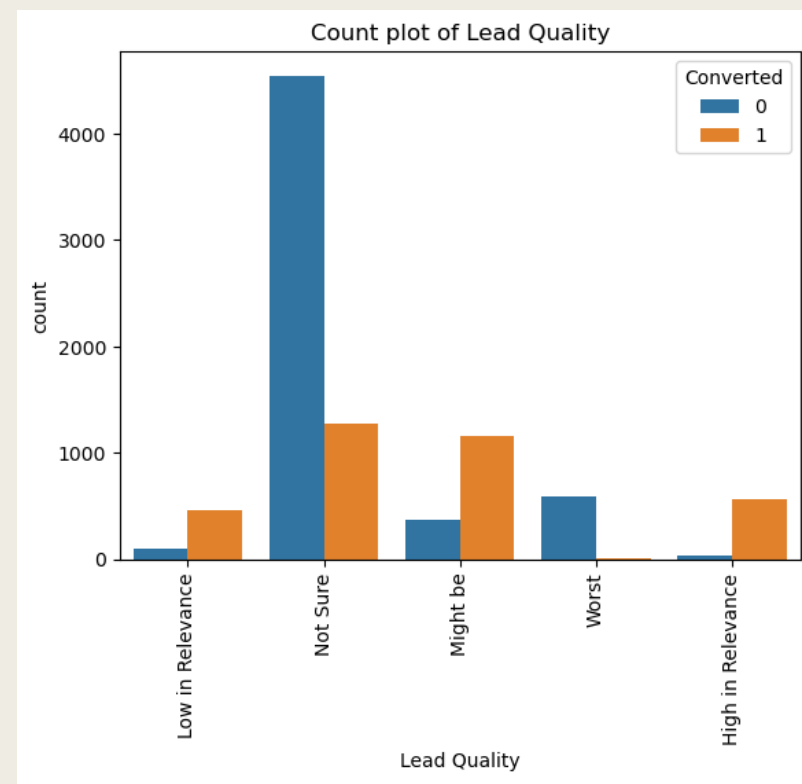
Lead Quality vs Converted
"Might Be" category has quite good conversion rate with respect to other categories.

Multivariate Analysis

- Top 20 highly correlated variables

fetching top 20 correlated variables

```python
In [67]: correlation_0 = leads_df.corr().abs().unstack().sort_values(kind='quicksort')
         correlation_0 = correlation_0.dropna()
         correlation_0 = correlation_0[correlation_0 != 1.0]

         round(correlation_0.sort_values(ascending=False).head(20),2)
```

```
Out[67]: Last Notable Activity_Unsubscribed              Last Activity_Unsubscribed                       0.88
         Last Activity_Unsubscribed                     Last Notable Activity_Unsubscribed               0.88
         Last Notable Activity_Email Opened             Last Activity_Email Opened                       0.86
         Last Activity_Email Opened                     Last Notable Activity_Email Opened               0.86
         Last Notable Activity_SMS Sent                 Last Activity_SMS Sent                           0.85
         Last Activity_SMS Sent                         Last Notable Activity_SMS Sent                   0.85
         What is your current occupation_Working Professional  What is your current occupation_Unemployed  0.85
         What is your current occupation_Unemployed     What is your current occupation_Working Professional  0.85
         Lead Origin_Lead Add Form                      Lead Source_Reference                            0.85
         Lead Source_Reference                          Lead Origin_Lead Add Form                        0.85
         Last Notable Activity_Email Link Clicked       Last Activity_Email Link Clicked                 0.80
         Last Activity_Email Link Clicked               Last Notable Activity_Email Link Clicked         0.80
         Specialization_Not Provided                    Lead Origin_Landing Page Submission              0.76
         Lead Origin_Landing Page Submission            Specialization_Not Provided                      0.76
         Page Views Per Visit                           TotalVisits                                      0.75
         TotalVisits                                    Page Views Per Visit                             0.75
         Last Activity_Page Visited on Website          Last Notable Activity_Page Visited on Website    0.68
         Last Notable Activity_Page Visited on Website  Last Activity_Page Visited on Website            0.68
         Lead Quality_Might be                          Lead Quality_Not Sure                            0.60
         Lead Quality_Not Sure                          Lead Quality_Might be                            0.60
         dtype: float64
```
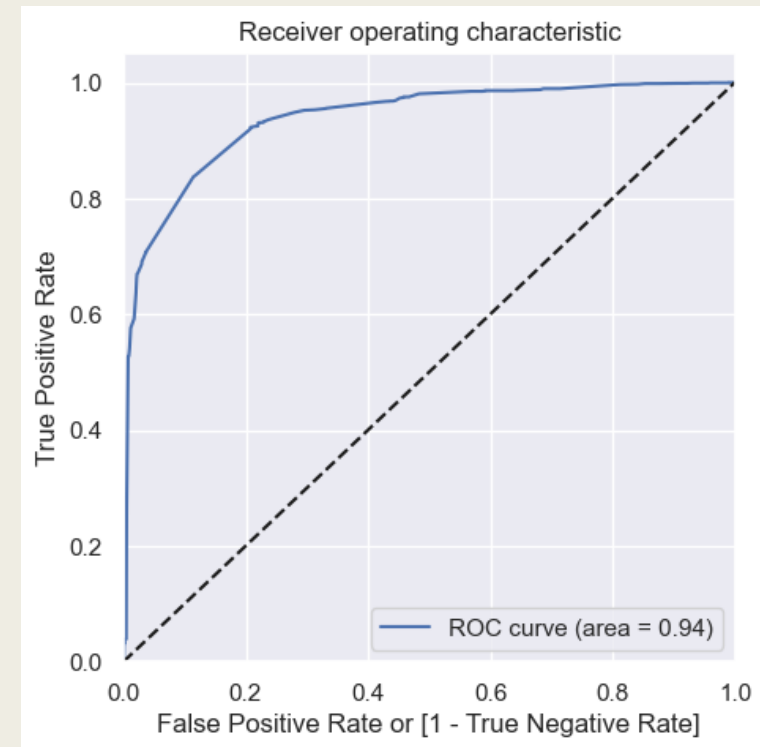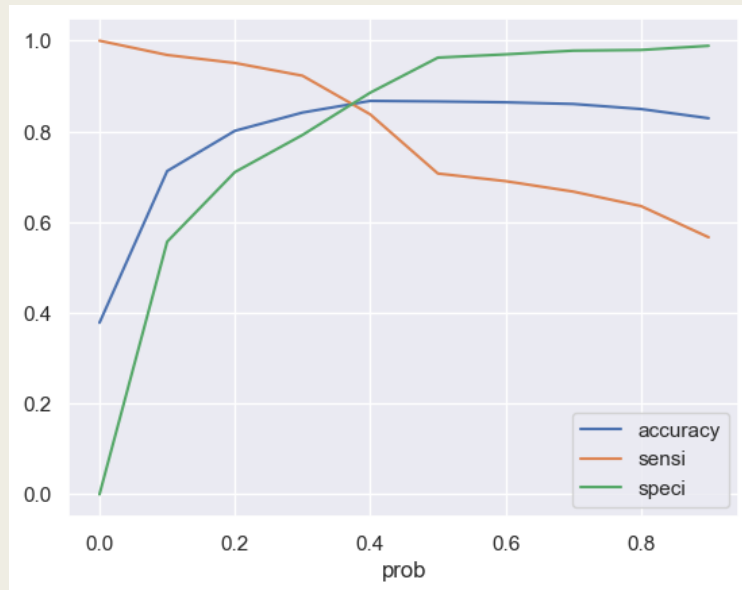
ROC Curve

An ROC curve demonstrates several things:

- It shows the trade-off 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.

Receiver operating characteristic

ROC curve (area = 0.94)

❖ The ROC Curve should be a value close to 1. We are getting a good value of 0.94 indicating a good predictive model.

Finding Optimal Cut-off point

Sensitivity-Specificity-Accuracy Curve
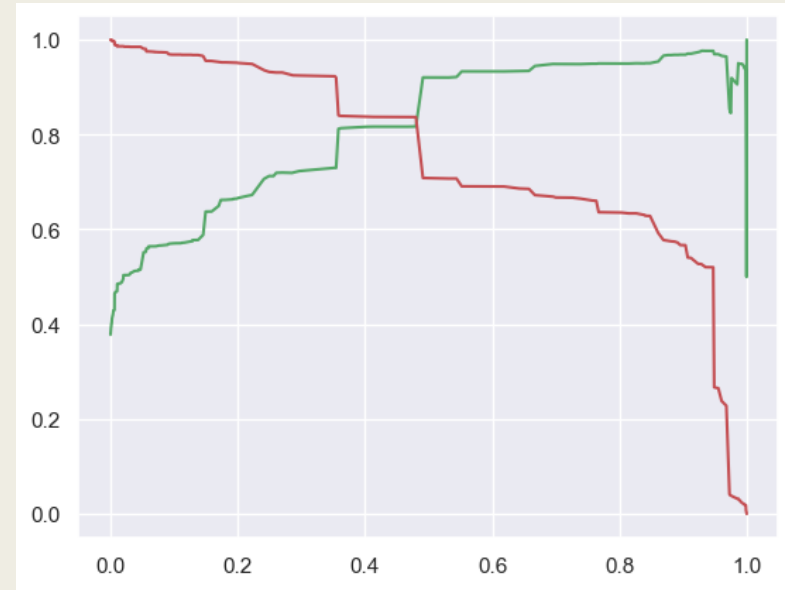
Precision-Recall Curve

Conclusions:
- With Sensitivity-Specificity-Accuracy plot, we are getting 0.38 optimum value for cutoff probability while In Precision-Recall Curve 0.48 looks optimal.
- We are taking 0.38 as the optimum point as a cutoff probability and assigning Lead Score in training and test data.

Variables from final model impacting conversion rate

Top 3 variables which are positively correlated from this model are :
- Lost to EINS from Tags
- Lead Source
- Busy from Tags

Top 3 variables with negative correlation are:
- Lead_Quality_Worst
- Invalid Number from Tags
- Switched Off from Tags

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.3426	0.151	15.477	0.000	2.046	2.639
Lead Origin_Landing Page Submission	-0.5190	0.085	-6.109	0.000	-0.686	-0.352
Lead Origin_Lead Import	-3.6312	0.856	-4.242	0.000	-5.309	-1.954
Lead Source_Other_source	3.1436	0.423	7.440	0.000	2.315	3.972
Lead Source_Referral Sites	-0.8321	0.367	-2.265	0.024	-1.552	-0.112
Last Activity_Email Bounced	-2.2161	0.418	-5.307	0.000	-3.034	-1.398
Last Activity_Olark Chat Conversation	-1.7551	0.184	-9.541	0.000	-2.116	-1.395
Tags_Busy	1.8412	0.238	7.726	0.000	1.374	2.308
Tags_Interested in full time MBA	-3.2584	0.793	-4.109	0.000	-4.813	-1.704
Tags_Interested in other courses	-3.7352	0.452	-8.267	0.000	-4.621	-2.850
Tags_Lost to EINS	6.0285	0.574	10.503	0.000	4.904	7.153
Tags_Ringing	-3.4079	0.252	-13.536	0.000	-3.901	-2.914
Tags_Will revert after reading the email	1.0663	0.141	7.558	0.000	0.790	1.343
Tags_invalid number	-3.8946	1.056	-3.688	0.000	-5.964	-1.825
Tags_switched off	-3.8501	0.550	-6.997	0.000	-4.928	-2.772
Lead Quality_Not Sure	-3.4888	0.112	-31.106	0.000	-3.709	-3.269
Lead Quality_Worst	-5.7068	0.500	-11.402	0.000	-6.688	-4.726
Last Notable Activity_Modified	-1.1704	0.097	-12.096	0.000	-1.360	-0.981
Last Notable Activity_Page Visited on Website	-0.7826	0.217	-3.612	0.000	-1.207	-0.358

Evaluation Metrics on Train and Test Data sets

- Accuracy, Sensitivity and Specificity of test data set are almost similar as train data set indicating no overfitting/underfitting.
- The model has fulfilled all the assumption of Logistic regression model.

- Also the target conversion ration by CEO is achieved by this model i.e. 83% which indicates that this is indeed a good model.

Metrics for Train Data

- *Accuracy* : 86.68%
- *Sensitivity* : 83.85%
- *Specificity* : 88.41%
- *Precision* : 81.50%
- *Recall* : 83.92%

Metrics for Test Data

- *Accuracy* : 86.23%
- *Sensitivity* : 83.25%
- *Specificity* : 88.09%
- *Precision* : 81.43%
- *Recall* : 83.31%

- **Actual Conversion Ratio = 38.02%**
- **Predicted Conversion Ratio = 83.25%**

Conclusions:

- We have build this Logistic regression model to predict the convert probability of incoming leads based on a cutoff value.

- The final model has **18 features** predicting the target variable.

- Optimum cutoff was obtained from **Sensitivity-Specificity-Accuracy** plot and value was chosen as **0.38.**

- So, any lead with greater than 0.38 of convert probability will be treated as Hot Lead (customer will convert) and lead with convert prob less than 0.38 will be taken as Cold lead.

- The final **Conversion Ratio** calculated on test data set is approx. **83%** which is quite good as X education wanted an expected value around 80%.

- **Accuracy, Sensitivity and Specificity** values of test data set are around **87%, 83% and 88%** respectively.

- The final model has Sensitivity of around 83%, which means the model is able to predict 83% customers out of all the converted customers.

- As per final model (Logm4) the top predictor variables are:

✓   Tags_Lost to EINS

✓   Lead Source_Other_source

✓   Tags_Busy