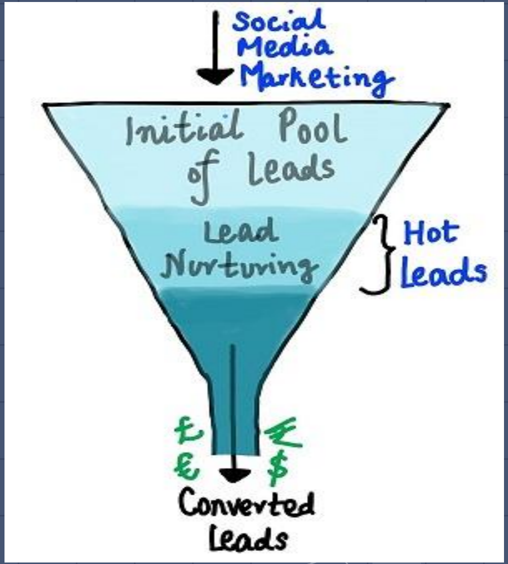
Lead Scoring – Case Study

Presented by:
Bharathy A
Shivansh Mishra

PROBLEM STATEMENT

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. 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. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

X Education has appointed us to help them 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 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.



There are a lot of leads generated in the initial stage, but only a few of them come out as paying customers. In the middle stage, we need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

Business Goals

The main goals for this case study are:

- Build 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. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- To adjust to if the company's requirement changes in the future so we will need to handle these as well.

Data Required for the Case Study

The leads.csv dataset from the past have been provided with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted. Another thing that we need to check out for are the levels present in the categorical variables. Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value.

Steps Involved in the Case Study

- Importing necessary Libraries
- Data Sourcing (Reading and understanding the data)
- Data Cleaning
- Visualization of the data
- Data Preparation for building the Model
- Building the Model
- Making Prediction on the Train dataset
- Evaluation of the Model
- Making Predictions on the Test Set
- Determining Feature Importance
- Final Observations
- Recommendations

Data Sourcing and Data Cleaning

- **Reading and understanding the data**
- Check for Duplicates
- Replace the 'Select' value in the categorical values to NaN.
- Check Percentage of Missing values for all columns
- Drop columns with a high percentage of missing values
- Drop categorical columns that are highly skewed
- Impute columns with less percentage of missing values

Visualisation of the data

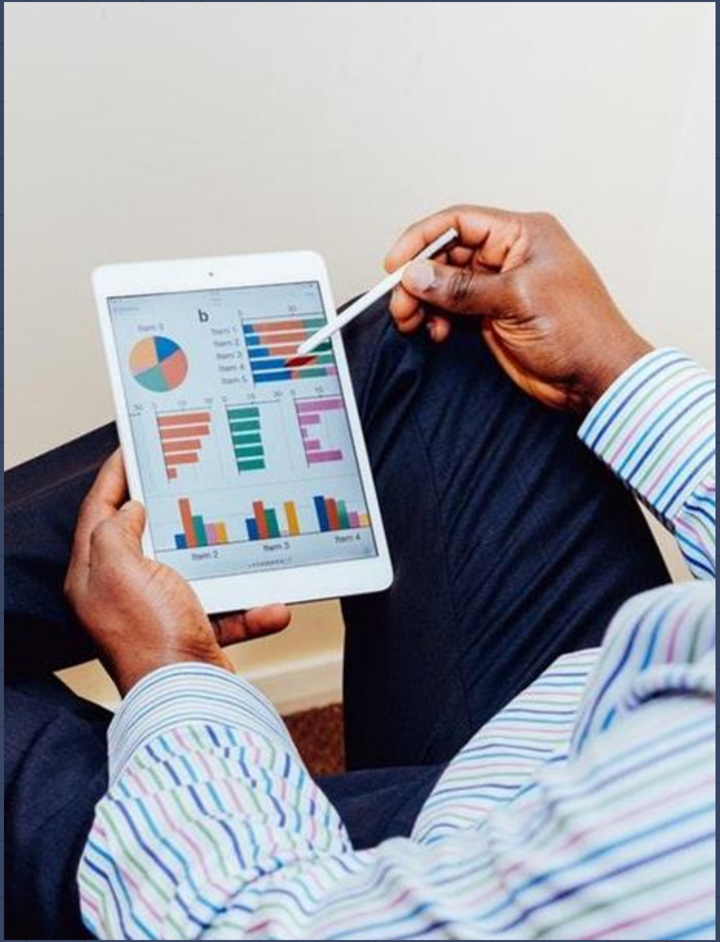
Numerical Variable Analysis and Outlier treatment

□ From the boxplots, we find that there are outliers present in the variables.

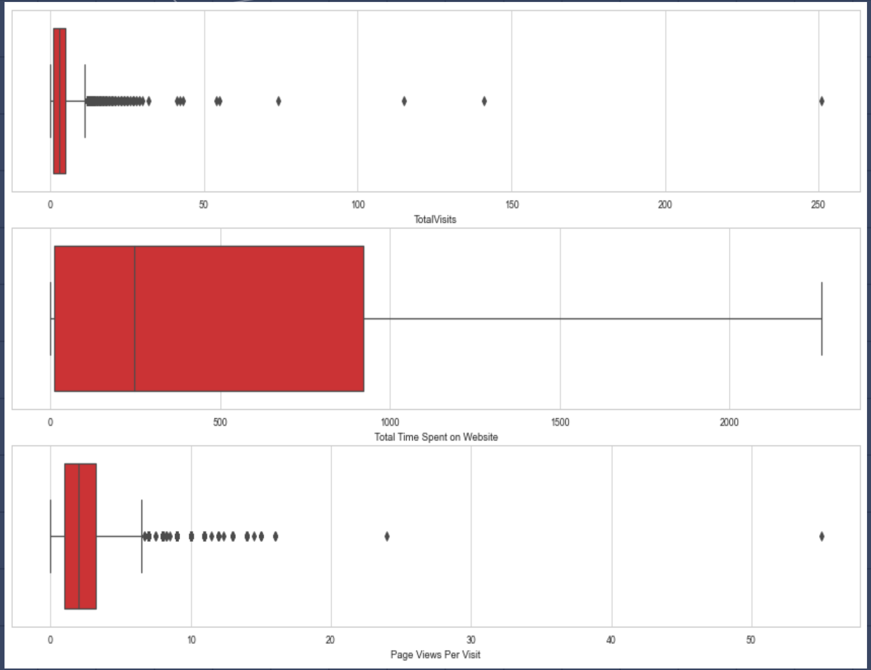
□ For 'TotalVisits', the 95% quantile is 10 but the maximum value is 251. Hence, we should cap these outliers at 95% value.

□ There are no outliers in 'Total Time Spent on Website'

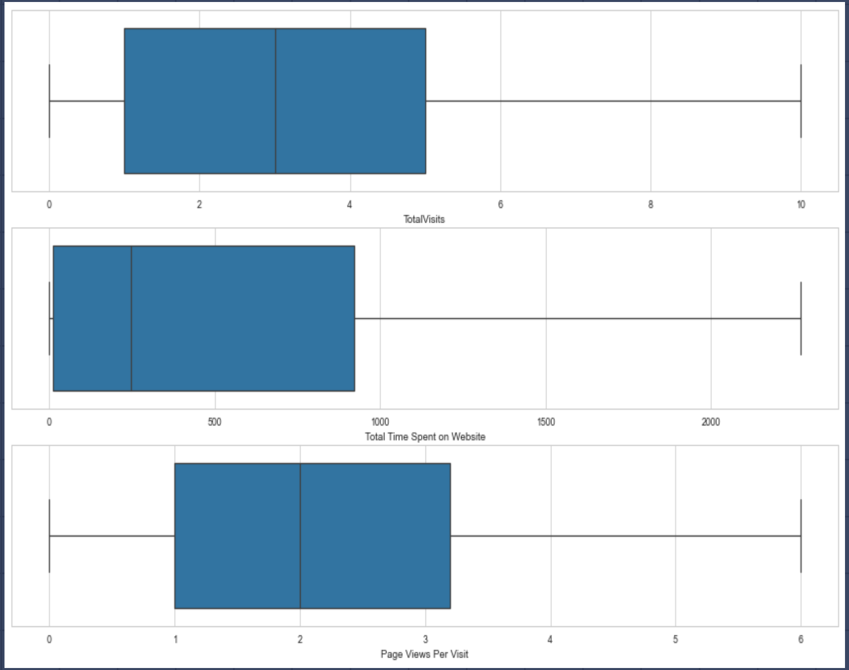
□ Again for 'Page Views Per Visit' we should cap outliers at 95% value.



Before treating the Outliers



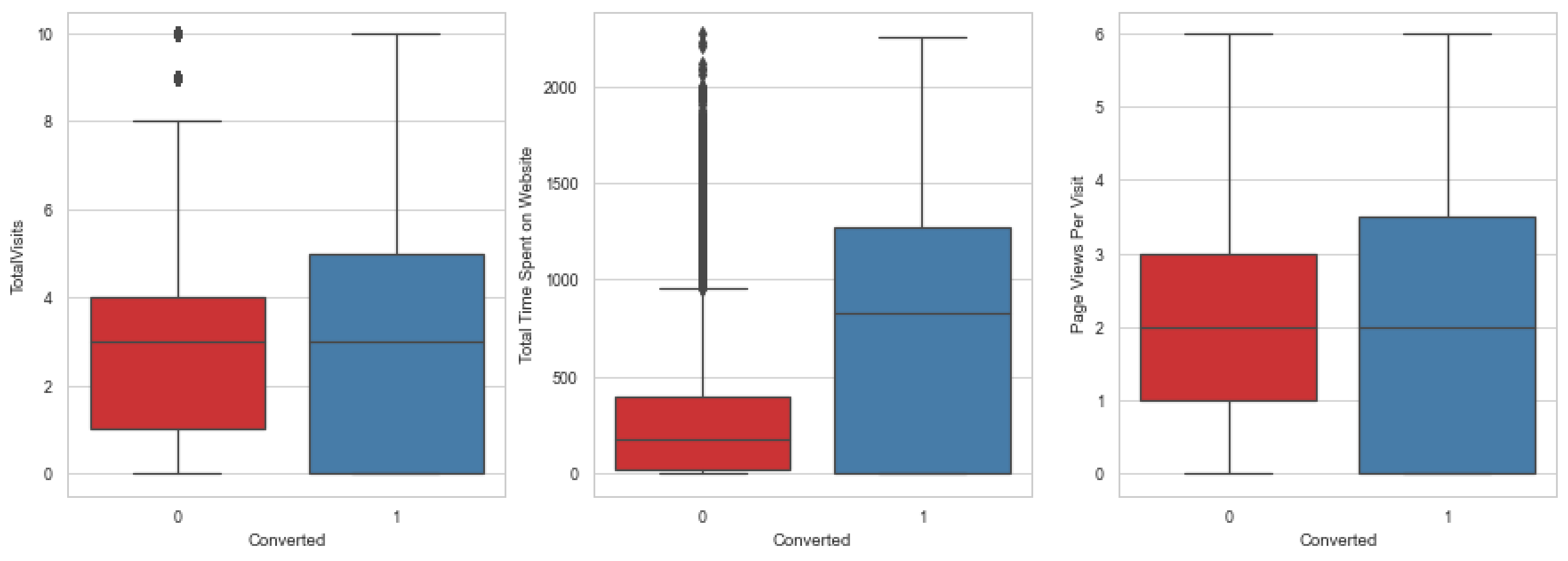
After treating the Outliers



Numerical variables against the target variable "Converted"

Insights:

- 'TotalVisits' has same median values for both converted and not converted leads. So no conclusion can be drawn from TotalVisits.
- People spending more time on the website are more likely to be converted.
- 'Page Views Per Visit' also has same median values for both outputs of leads. Hence no conclusion can be drawn from this.


TotalVisits
0
2
4
6
8
10
0
1
Converted
Total Time Spent on Website
0
500
1000
1500
2000
0
1
Converted
Page Views Per Visit
0
1
2
3
4
5
6
0
1
Converted

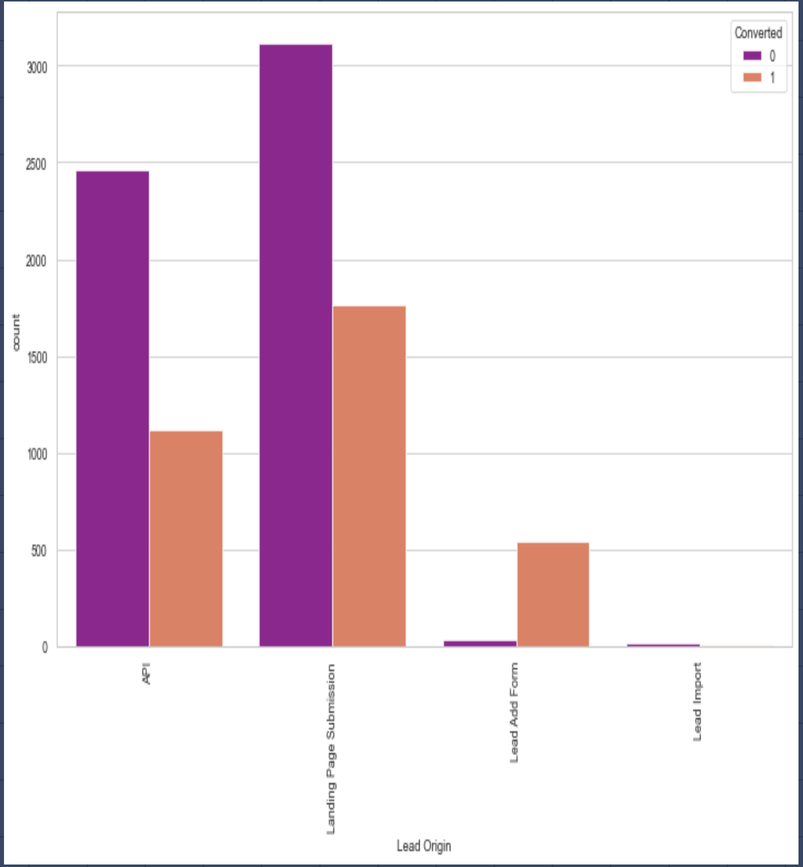
Visualization of Categorical Variables

Analysis of the categorical variables against the target variable "Converted“

Insights:

- □ 'API' and 'Landing Page Submission' generates the most leads but have less conversion rates.
- □ 'Lead Add Form' generates less leads but conversion rate is greater.
- □ 'Lead Import' d
- □ oes not seem to be very significant.
- □ To improve we should try to increase lead conversion rate for 'API' and 'Landing Page Submission', and generate more leads from Lead Add Form.

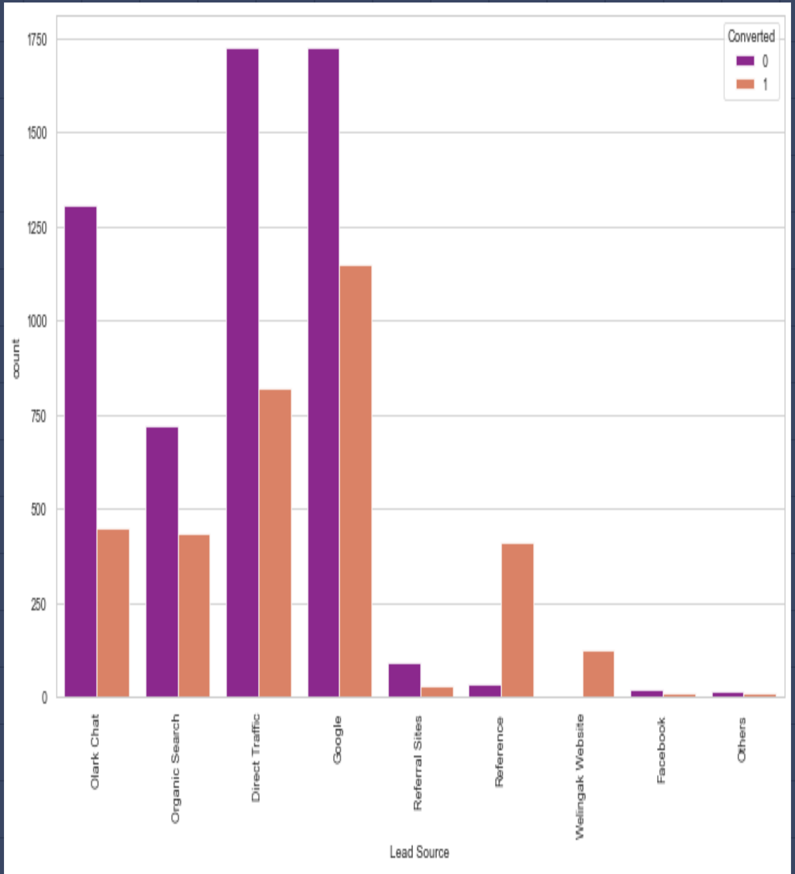
Lead Origin



Lead _Source

Insights:

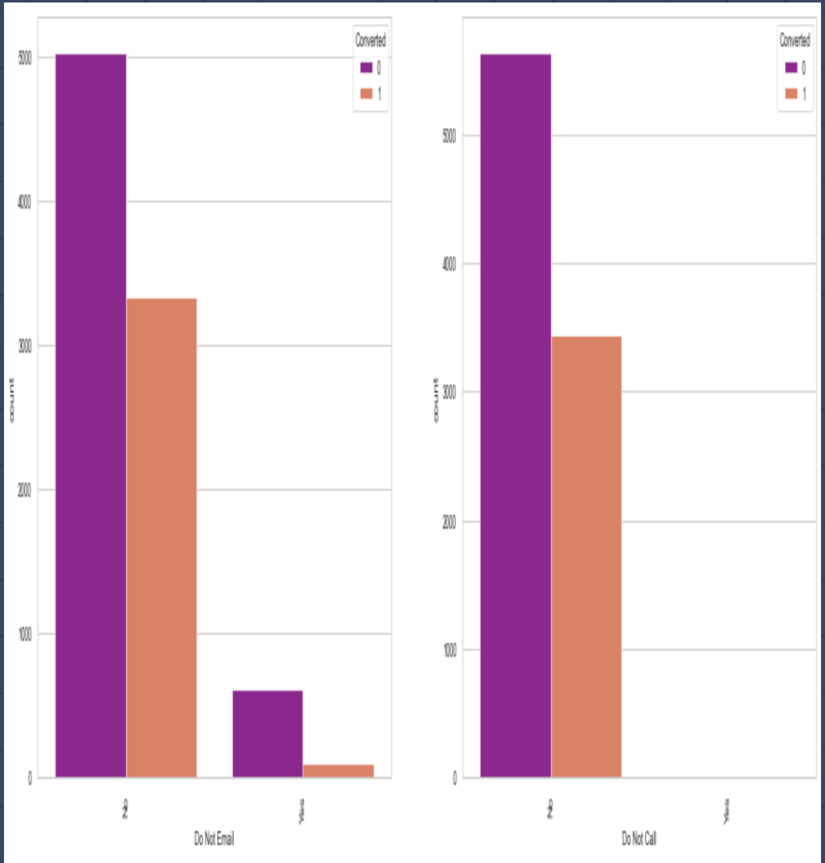
- □ There are both 'google' and 'Google'. Hence we have changed to 'Google'
- □ As it can be seen from the plot, number of leads generated by many of the sources are negligible. There are sufficient numbers till Facebook. So we have converted all these sources to one single category named 'Others'.
- □ 'Direct Traffic' and 'Google' generate maximum number of leads while maximum conversion rate is achieved through 'Reference' and 'Welingak Website'.
- □ To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organic search, direct traffic, and google leads.
- □ More leads has to be generated from reference and welingak website.



Do Not Email and Do Not Call

Insights:

- Most of the responses are 'No' for both the variables which generated most of the leads. Nothng significant can be inferred from these attributes.

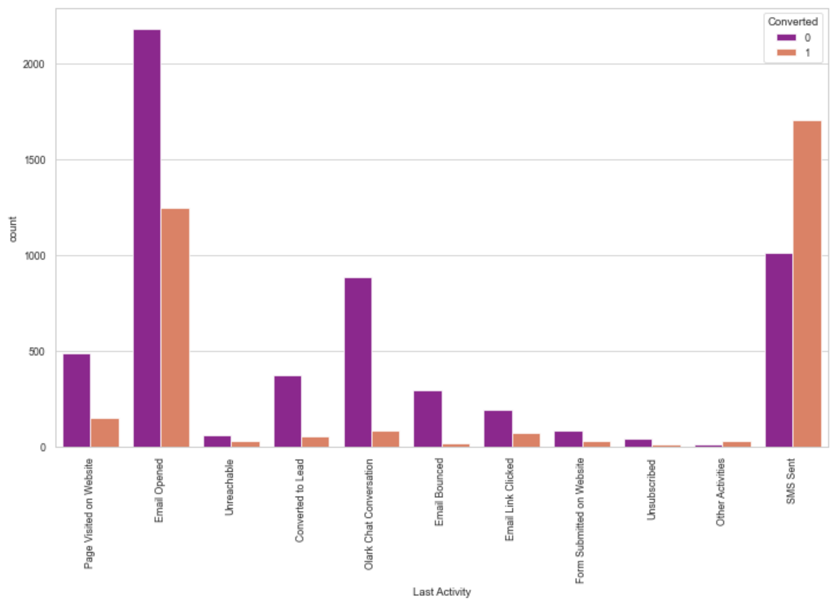


Last Activity

Insights:

□ Highest number of lead generated is 'Email Opened' while the maximum conversion rate is for the activity of 'SMS Sent'. Its conversion rate is significantly high.

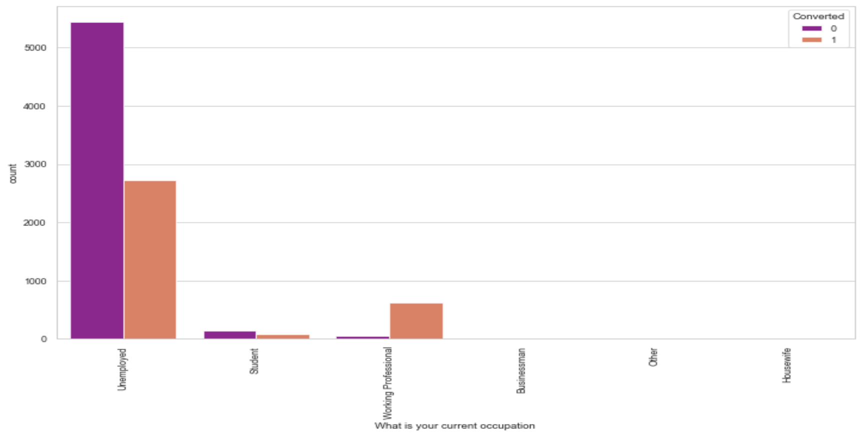
□ Categories after "SMS sent" are having very less data. So these have been be grouped into a new category named "Other Activities"



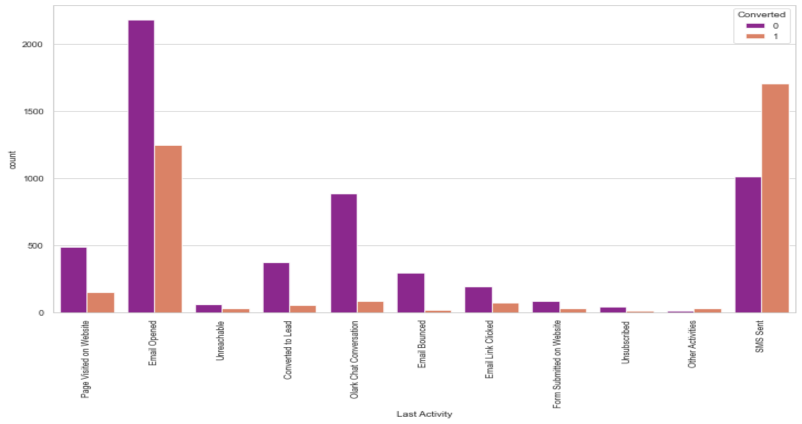
Insights:

- □ Most of the responses are for India. Remaining categories are not significant.
- □ Conversion rates are mostly same across different specializations.
- □ The highest conversion rate is for 'Working Professional'.
- □ High number of leads are generated for 'Unemployed' but conversion rate is low.

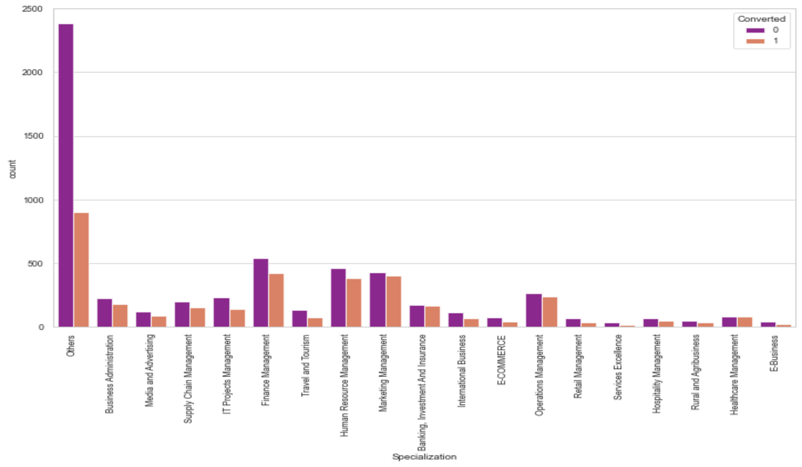
What is your current occupation



Country



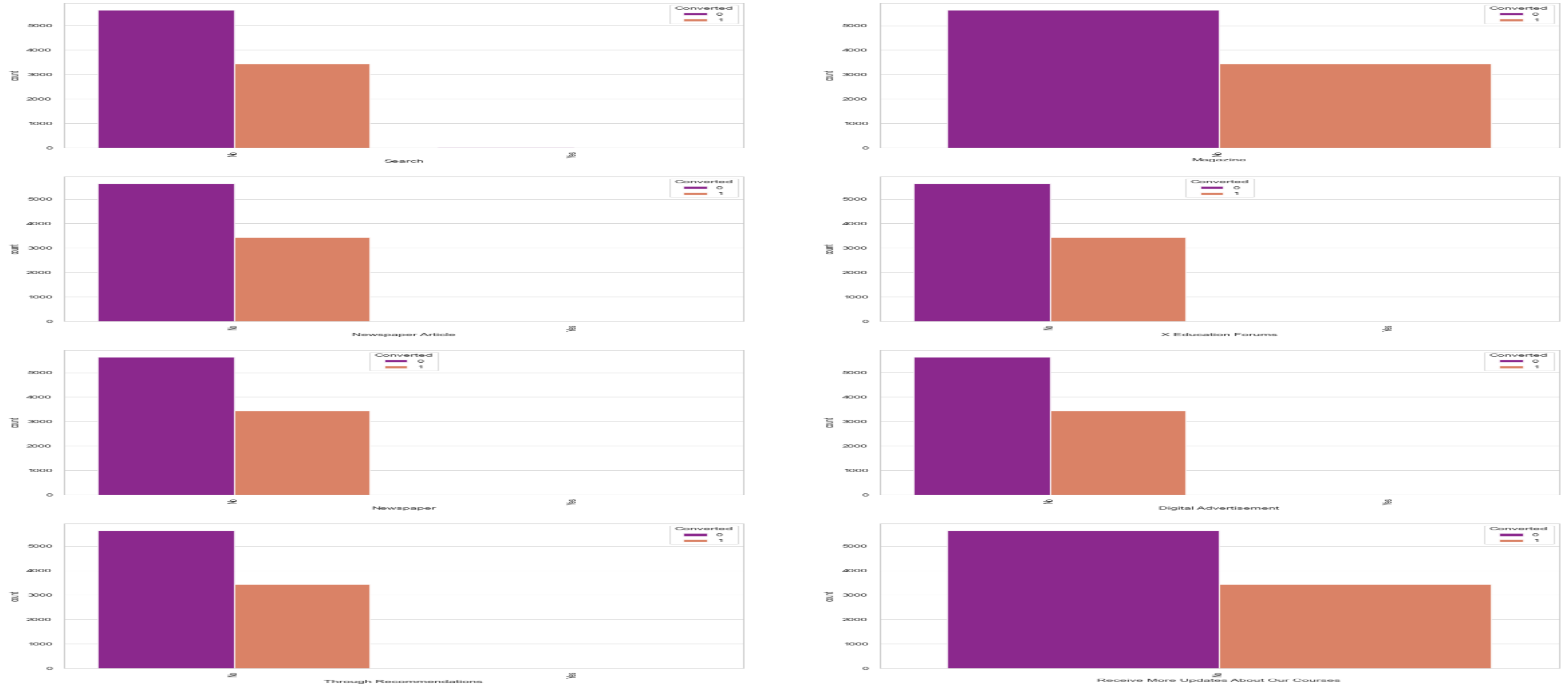
Specialization



'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', 'Receive More Updates About Our Courses'

Insight:

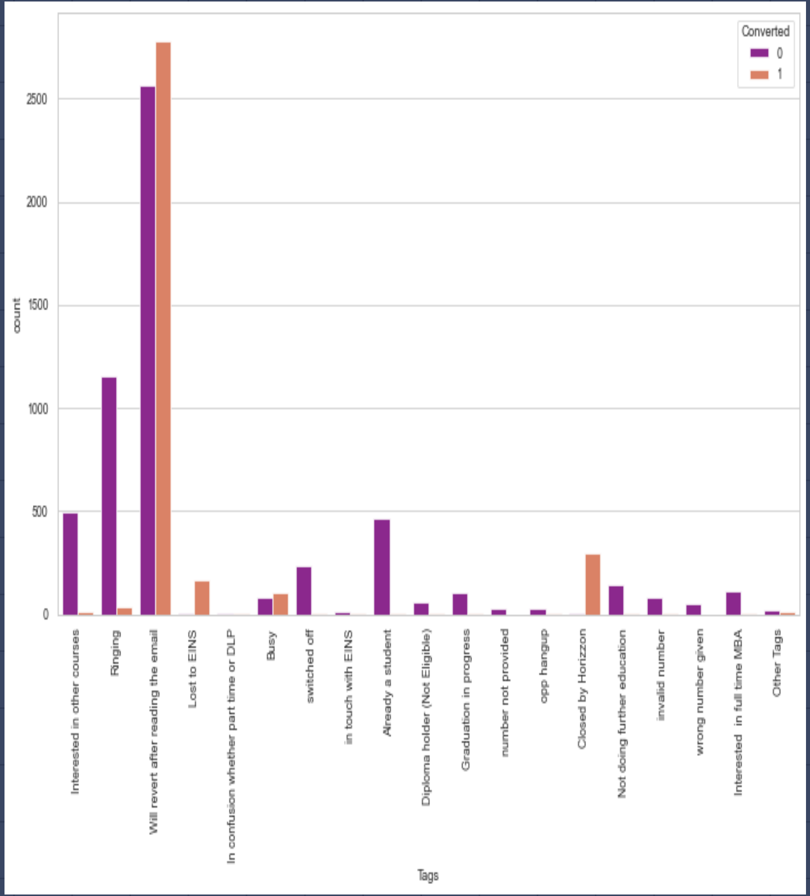
Most of the responses are 'No' for both the variables which generated most of the leads. Nothng significant can be inferred from these attributes.



Insights:

- In Tags, categories after 'Interested in full time MBA' have very few leads generated. Hence we have combined them into one single category as "Other Tags".
- Most leads generated and the highest conversion rate are both from the attribute 'Will revert after reading the email'.
- In Lead quality, 'Might be' has the highest conversion rate while 'Worst' has the less conversion rate.

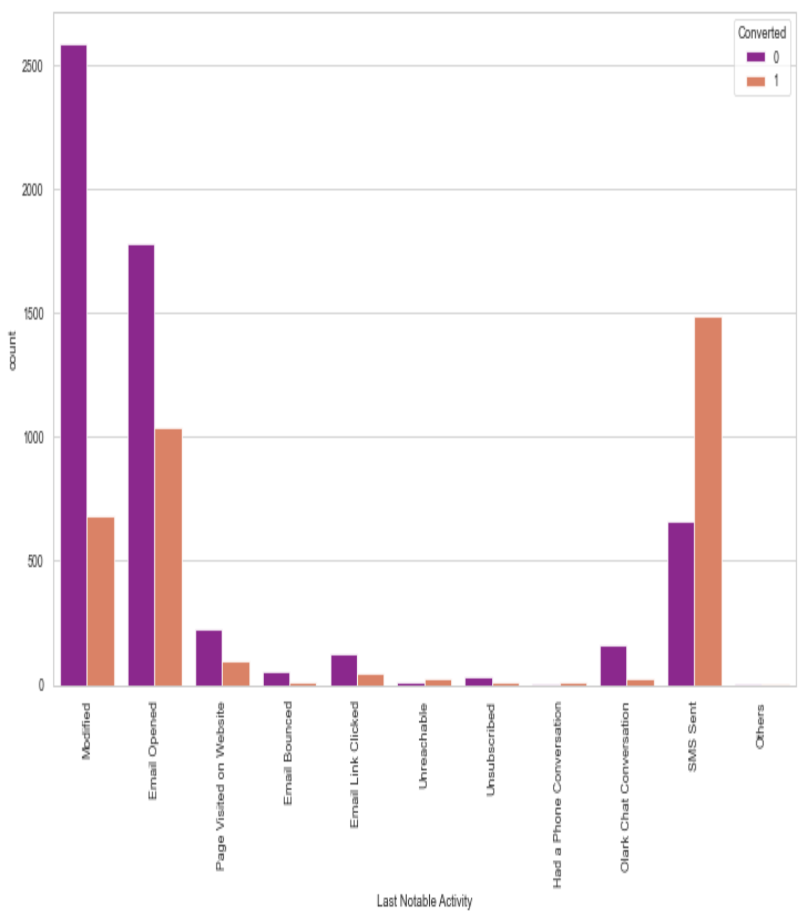
Tags

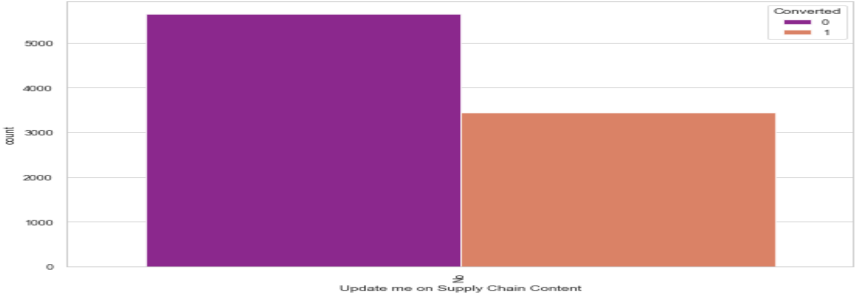


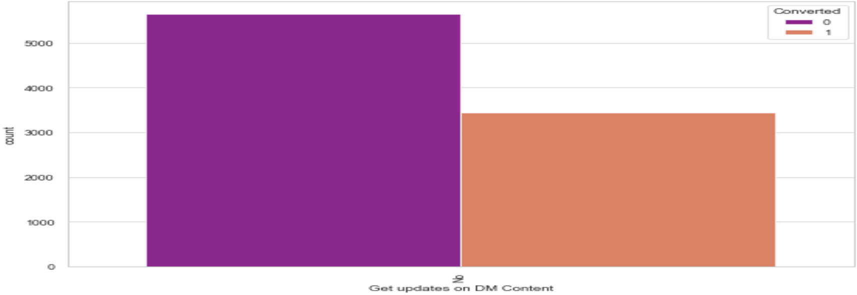
'Update me on Supply Chain Content', 'Get updates on DM Content', 'City', 'I agree to pay the amount through cheque', 'A free copy of Mastering The Interview', 'Last Notable Activity'

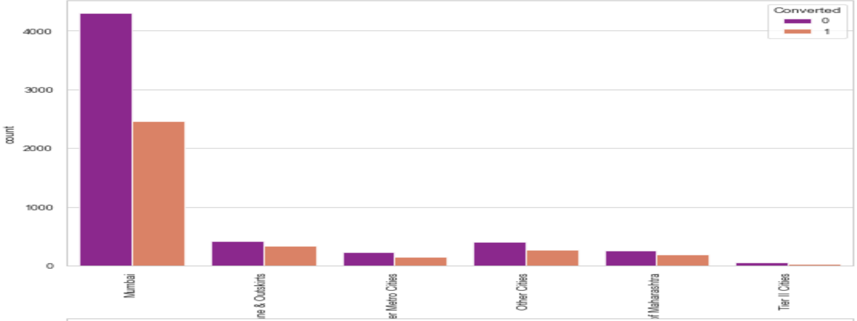
Insights:

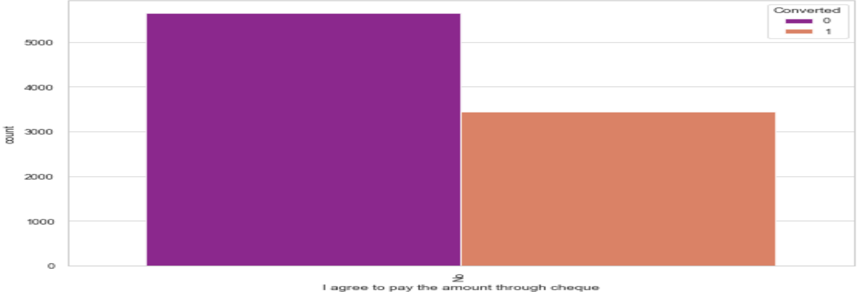
- Most of these variables have only one significant category 'NO' and hence they are insignificant in analysis.
- In City, most of the leads are generated for 'Mumbai'.
- In 'Last Notable Activity', we have combined categories after 'SMS Sent' to a new category 'Others' as they have very less data. The most generated leads are for the category 'Modified' while most conversion rate are for 'SMS Sent' activity.

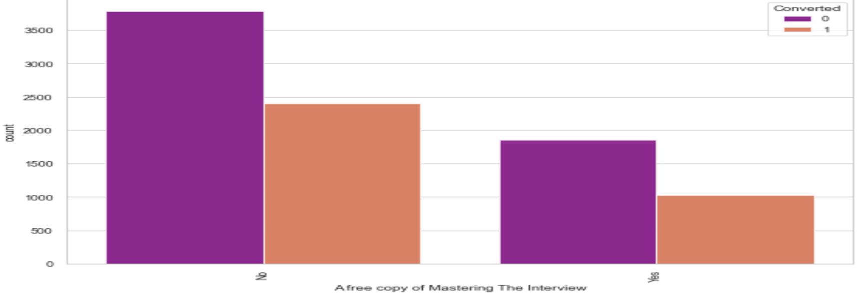


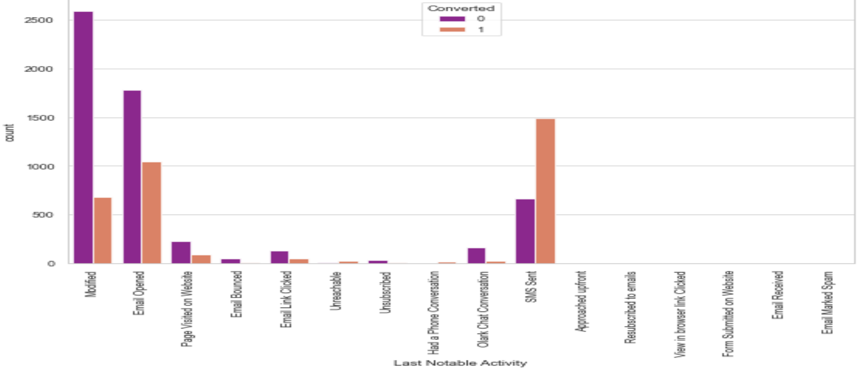

Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Update me on Supply Chain Content


Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Get updates on DM Content


Converted
0
1
count
4000
3000
2000
1000
0
Mumbai
ne & Outskirts
er Metro Cities
Other Cities
f Maharashtra
Tier II Cities


Converted
0
1
count
5000
4000
3000
2000
1000
0
No
I agree to pay the amount through cheque


Converted
0
1
count
3500
3000
2500
2000
1500
1000
500
0
No
Yes
A free copy of Mastering The Interview


Converted
0
1
count
2500
2000
1500
1000
500
0
Modified
Email Opened
Page Visited on Website
Email Bounced
Email Link Clicked
Unreachable
Unsubscribed
Had a Phone Conversation
Olark Chat Conversation
SMS Sent
Approached upfront
Resubscribed to emails
View in browser link Clicked
Form Submitted on Website
Email Received
Email Marked Spam
Last Notable Activity

Data Preparation for building the Model

- Creating Dummy variables for Categorical Data:

  There are few columns which are having more than 2 categorical values. In order to perform logistic regression dummy variables has to be created for these categorical columns.
  - Create Dummy variable
  - Drop original variable for which the dummy was created
  - Drop first dummy variable for each set of dummies created.

- Splitting the data:

  The dataset is split into
  - Train data (70%)
  - Test data (30%)

- Scaling the features:
  - The features have been scaled using StandardScaler

Building the Model

The following steps are followed in building a model:

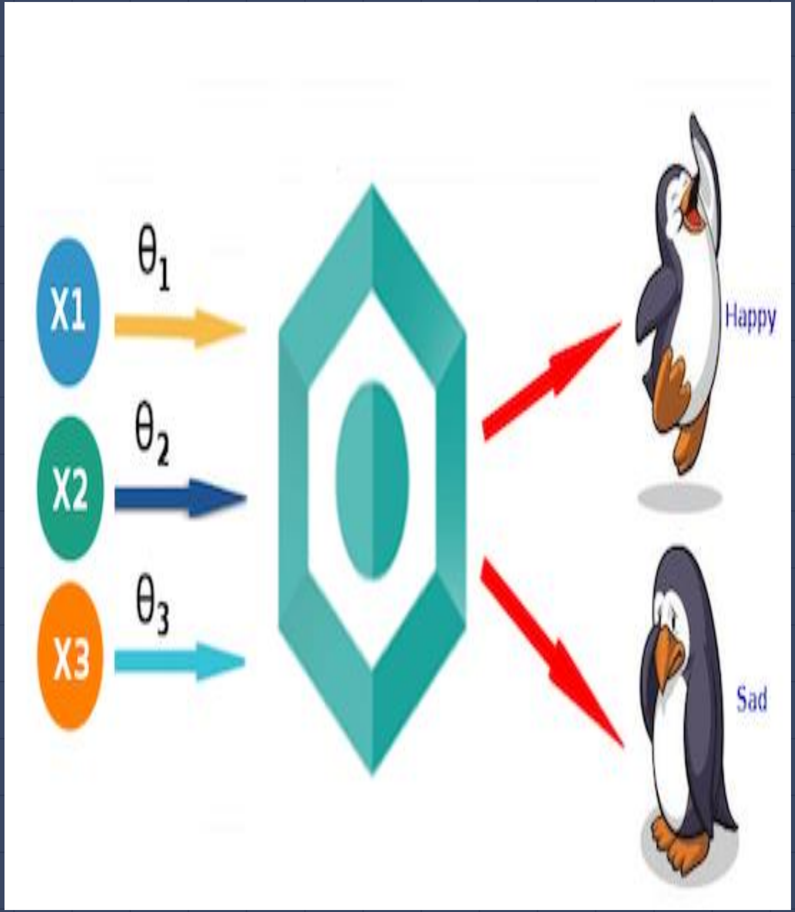
□ Import the necessary packages for model preprocessing and model building

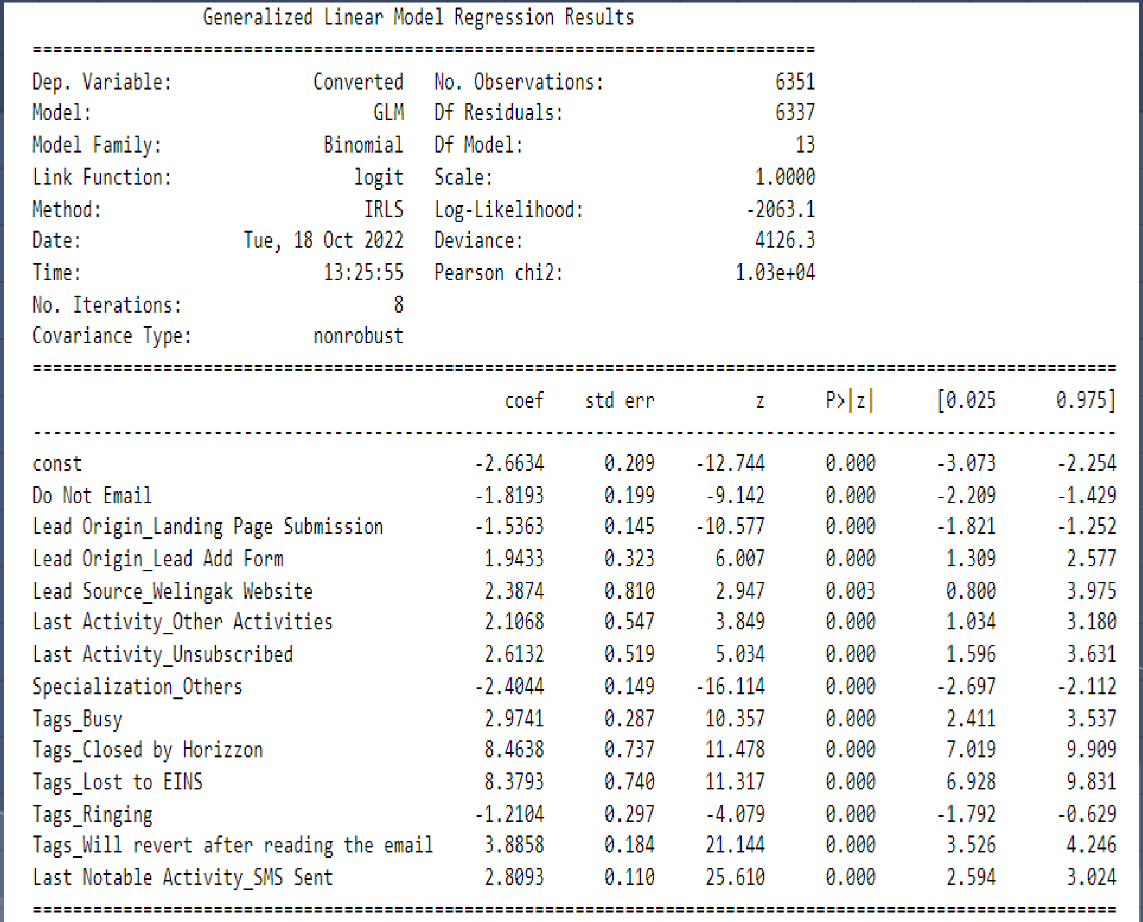
□ Build the model using a combination of automatic and manual processing

□ Start the model with RFE features (automatic) and drop the features,

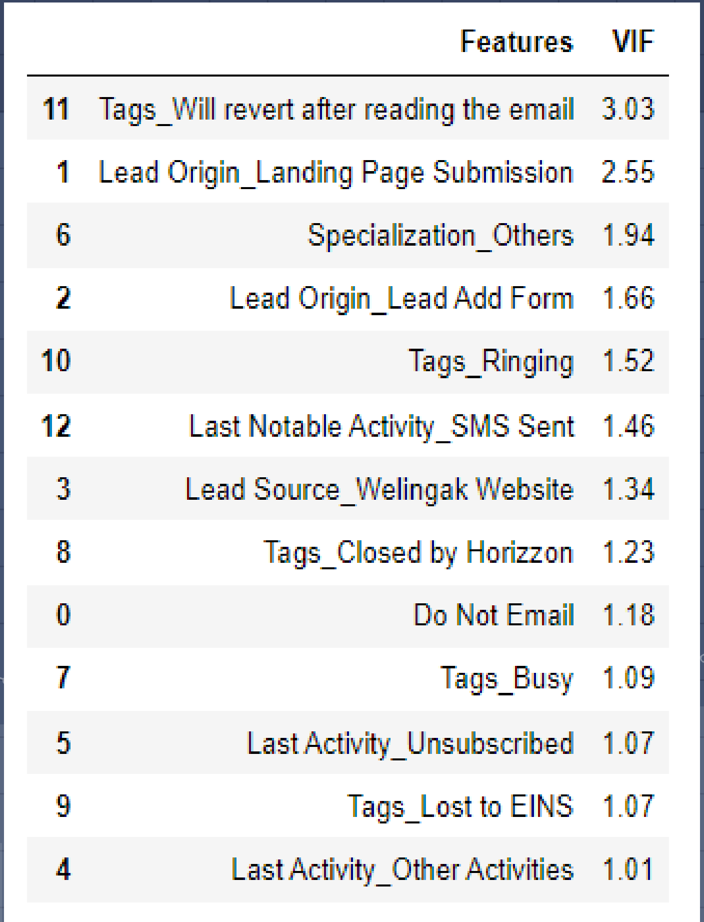
□ Use manual processing for feature reduction by dropping one feature at a time using the p value and VIF.

□ Build the model and fit the training data.





```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6351
Model:                            GLM   Df Residuals:                     6337
Model Family:                Binomial   Df Model:                           13
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -2063.1
Date:                Tue, 18 Oct 2022   Deviance:                       4126.3
Time:                        13:25:55   Pearson chi2:                 1.03e+04
No. Iterations:                     8
Covariance Type:            nonrobust
============================================================================================================
                                               coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------
const                                       -2.6634      0.209    -12.744      0.000      -3.073      -2.254
Do Not Email                                -1.8193      0.199     -9.142      0.000      -2.209      -1.429
Lead Origin_Landing Page Submission         -1.5363      0.145    -10.577      0.000      -1.821      -1.252
Lead Origin_Lead Add Form                    1.9433      0.323      6.007      0.000       1.309       2.577
Lead Source_Welingak Website                 2.3874      0.810      2.947      0.003       0.800       3.975
Last Activity_Other Activities               2.1068      0.547      3.849      0.000       1.034       3.180
Last Activity_Unsubscribed                   2.6132      0.519      5.034      0.000       1.596       3.631
Specialization_Others                       -2.4044      0.149    -16.114      0.000      -2.697      -2.112
Tags_Busy                                    2.9741      0.287     10.357      0.000       2.411       3.537
Tags_Closed by Horizzon                      8.4638      0.737     11.478      0.000       7.019       9.909
Tags_Lost to EINS                            8.3793      0.740     11.317      0.000       6.928       9.831
Tags_Ringing                                -1.2104      0.297     -4.079      0.000      -1.792      -0.629
Tags_Will revert after reading the email     3.8858      0.184     21.144      0.000       3.526       4.246
Last Notable Activity_SMS Sent               2.8093      0.110     25.610      0.000       2.594       3.024
============================================================================================================
```

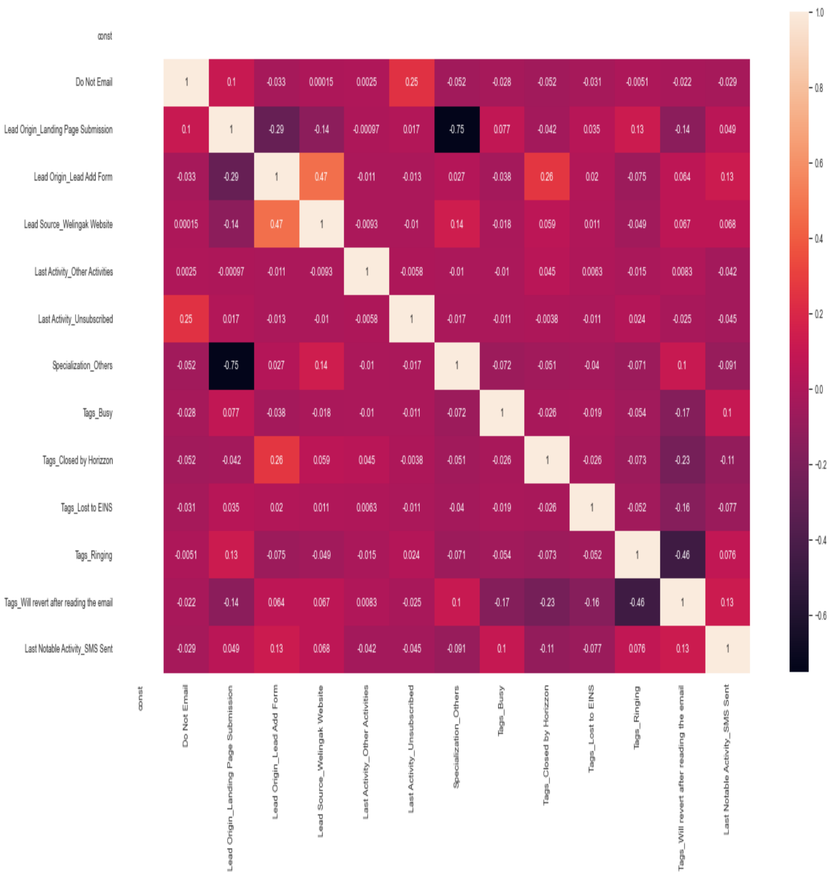


	Features	VIF
11	Tags_Will revert after reading the email	3.03
1	Lead Origin_Landing Page Submission	2.55
6	Specialization_Others	1.94
2	Lead Origin_Lead Add Form	1.66
10	Tags_Ringing	1.52
12	Last Notable Activity_SMS Sent	1.46
3	Lead Source_Welingak Website	1.34
8	Tags_Closed by Horizzon	1.23
0	Do Not Email	1.18
7	Tags_Busy	1.09
5	Last Activity_Unsubscribed	1.07
9	Tags_Lost to EINS	1.07
4	Last Activity_Other Activities	1.01

Correlation:

Insights:

- From VIF values and heat maps, we can see that there is not much multicollinearity present. All variables have a good value of VIF. These features seem important from the business aspect as well. So we need not drop any more variables and we can proceed with making predictions using this model as final model.

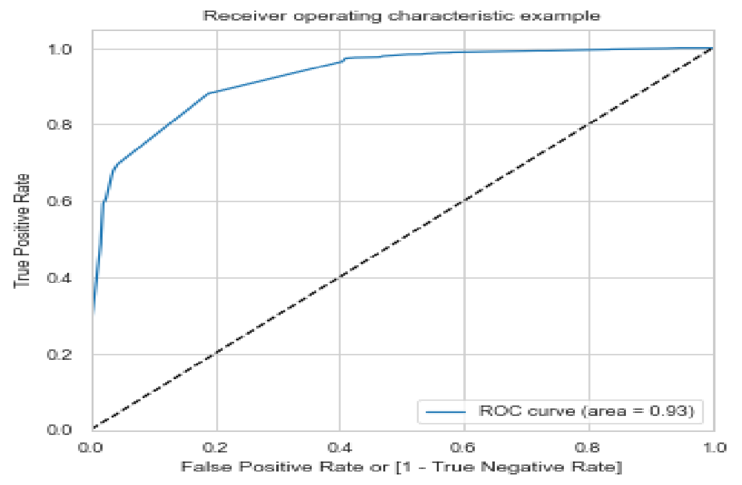


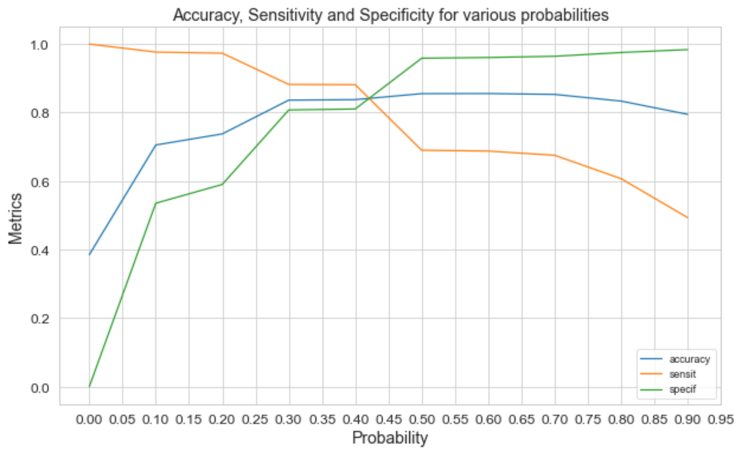
	const	Do Not Email	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Source_Welingak Website	Last Activity_Other Activities	Last Activity_Unsubscribed	Specialization_Others	Tags_Busy	Tags_Closed by Horizzon	Tags_Lost to EINS	Tags_Ringing	Tags_Will revert after reading the email	Last Notable Activity_SMS Sent
const														
Do Not Email		1	0.1	-0.033	0.00015	0.0025	0.25	-0.052	-0.028	-0.052	-0.031	-0.0051	-0.022	-0.029
Lead Origin_Landing Page Submission		0.1	1	-0.29	-0.14	-0.00097	0.017	-0.75	0.077	-0.042	0.035	0.13	-0.14	0.049
Lead Origin_Lead Add Form		-0.033	-0.29	1	0.47	-0.011	-0.013	0.027	-0.038	0.26	0.02	-0.075	0.064	0.13
Lead Source_Welingak Website		0.00015	-0.14	0.47	1	-0.0093	-0.01	0.14	-0.018	0.059	0.011	-0.049	0.067	0.068
Last Activity_Other Activities		0.0025	-0.00097	-0.011	-0.0093	1	-0.0058	-0.01	-0.01	0.045	0.0063	-0.015	0.0083	-0.042
Last Activity_Unsubscribed		0.25	0.017	-0.013	-0.01	-0.0058	1	-0.017	-0.011	-0.0038	-0.011	0.024	-0.025	-0.045
Specialization_Others		-0.052	-0.75	0.027	0.14	-0.01	-0.017	1	-0.072	-0.051	-0.04	-0.071	0.1	-0.091
Tags_Busy		-0.028	0.077	-0.038	-0.018	-0.01	-0.011	-0.072	1	-0.026	-0.019	-0.054	-0.17	0.1
Tags_Closed by Horizzon		-0.052	-0.042	0.26	0.059	0.045	-0.0038	-0.051	-0.026	1	-0.026	-0.073	-0.23	-0.11
Tags_Lost to EINS		-0.031	0.035	0.02	0.011	0.0063	-0.011	-0.04	-0.019	-0.026	1	-0.052	-0.16	-0.077
Tags_Ringing		-0.0051	0.13	-0.075	-0.049	-0.015	0.024	-0.071	-0.054	-0.073	-0.052	1	-0.46	0.076
Tags_Will revert after reading the email		-0.022	-0.14	0.064	0.067	0.0083	-0.025	0.1	-0.17	-0.23	-0.16	-0.46	1	0.13
Last Notable Activity_SMS Sent		-0.029	0.049	0.13	0.068	-0.042	-0.045	-0.091	0.1	-0.11	-0.077	0.076	0.13	1

Making Prediction on the Train dataset

The following steps are done after building the model:

- □ Get the predictions on the training dataset with the final model
- □ Use the cut-off with 0.5 for the initial predictions
- □ Derive the Classification report and Classification metrics with the initial cutoff and predictions
- □ Derive the Area under the ROC curve for the initial cut-off and predictions





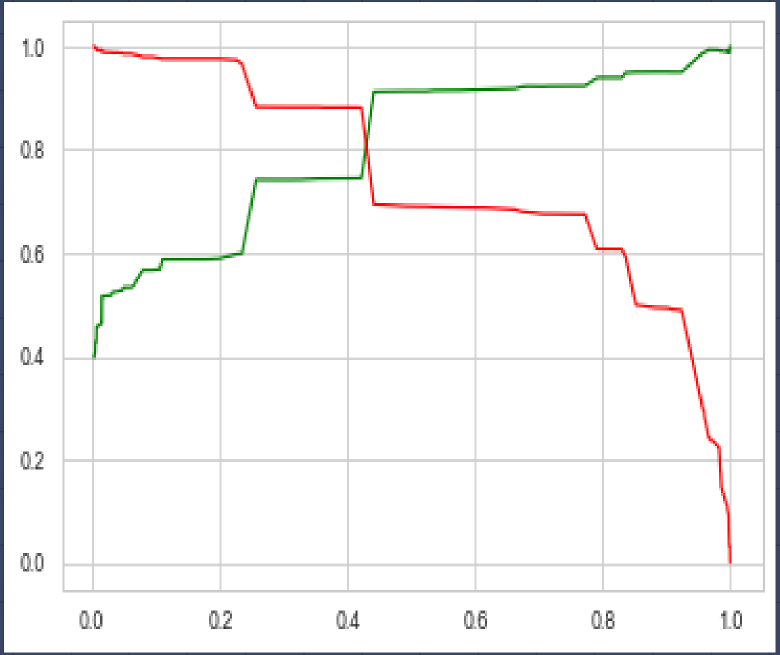
- □ Calculate the predicted values for the different cut-offs to arrive at the optimal cutoff.
- □ Plot the Sensitivity / Specificity curve for the different cut-offs and identify the optimal cut-off
- □ Get the final_Predictions and the metrics for the Predictions with the optimal cut-off
- □ Assign a Lead Score to the Training dataset based on the Conversion probability of the final_Predictions
- □ Measuring the Precision Recall Trade-off

Training Data set:

- □ Accuracy: 0.84%
- □ Sensitivity: 0.88%
- □ Specificity: 0.81%

Overall the Accuracy and other metrics yield similar values for both the cutoffs. We'll use the cutoff of 0.42 as our threshold.

precision vs recall



Making Predictions on the Test Set

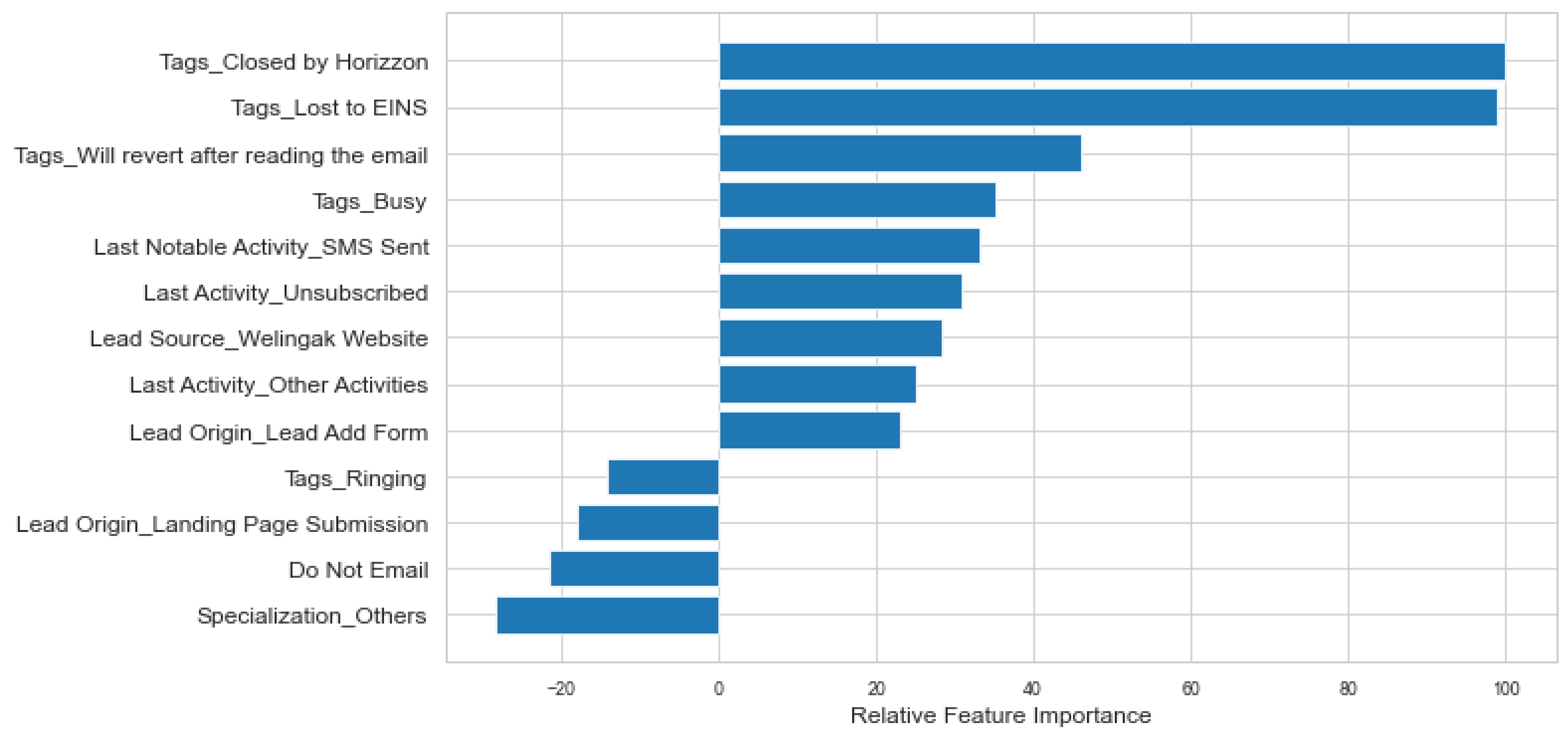
The next step is to validate the model with the test dataset. The following are the steps involved:

- Fit the Numeric features of the Test dataset with the Scaler method
- Making Predictions on the X_test dataset
- Create a Dataset with the Prospect ID and the conversion probability for the test dataset
- Generate the Lead Score for the test dataset based on the predicted probability from the model
- Get the final Predicted values using the optimal threshold value
- Get the Final evaluation Metrics for the test dataset with the actual converted values and final predicted values
- **Finally assign the Lead Score to the Testing data**

The Final Evaluation Metrics for the test Dataset:

- Accuracy: 0.82%
- Sensitivity: 0.86%
- Specificity: 0.80%

Determining Feature Importance



Final Observations

The Final Evaluation Metrics for the train Dataset:

- Accuracy: 0.84%
- Sensitivity: 0.88%
- Specificity: 0.81%

The Final Evaluation Metrics for the test Dataset:

- Accuracy: 0.82%
- Sensitivity: 0.86%
- Specificity: 0.80%

The Model seems to predict the Conversion Rate very well and we should be able to make good calls based on this model

Recommendations

Following are the selected features which are more significant in predicting the conversion.

- Features having positive impact on conversion probability:
- Features with Positive Coefficient Values:
- Tags_Closed by Horizzon
- Tags_Lost to EINS
- Tags_Will revert after reading the email
- Tags_Busy
- Last Activity_Unsubscribed
- Last Notable Activity_SMS Sent
- Lead Source_Welingak Website
- Last Activity_Other Activities
- Lead Origin_Lead Add Form

Features having negative impact on conversion probability:

- Features with Negative Coefficient Values:
- Specialization_Others
- Do Not Email
- Lead Origin_Landing Page Submission
- Tags_Ringing

THANK
YOU !