Lead Scoring Case Study

X Education

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.
- Although X Education gets a lot of leads, its lead conversion rate is very poor. The typical lead conversion rate at X education is around 30%.
- An X Education need 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%.

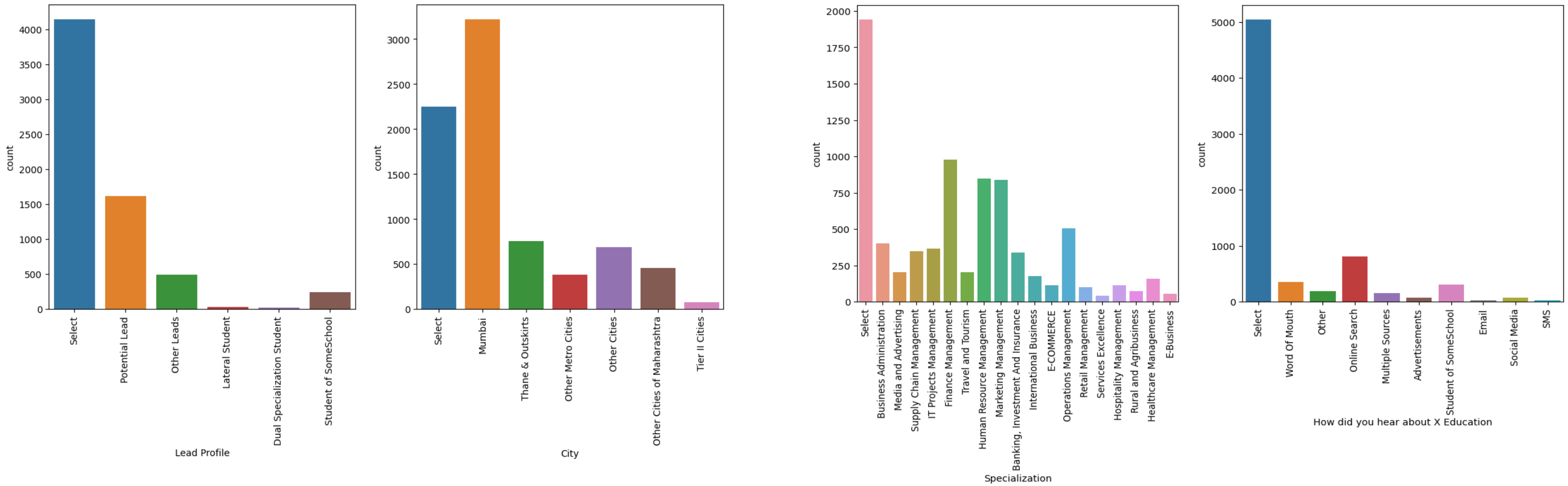
Approach taken for Solution

- Importing libraries and dataset
- Data cleaning and data manipulation
  - Check for imbalanced data
  - Check for columns with Select value and Impute the data
  - Check for duplicate data
  - Check for missing values and impute/drop the columns
  - Check for outliers and impute data
- Exploratory data analysis
  - Univariate analysis
  - Bivariate analysis
- Data Preparation
  - Dataset split into Train and Test sets
  - Feature scaling and Dummy variables and encoding the data
- Model Building using RFE and make use of StatsModel
  - Build logistic regression model and delete the variables which are not useful using the p-value and VIF values
- Model Evaluation
  - Confusion Matrix and ROC curve
  - Optimal cutoff point for Accuracy Sensitivity and Specificity
- Model Prediction on test dataset
- Precision and Recall View
  - Precision and Recall tradeoff
- Model Prediction on test dataset

Data collection, cleaning

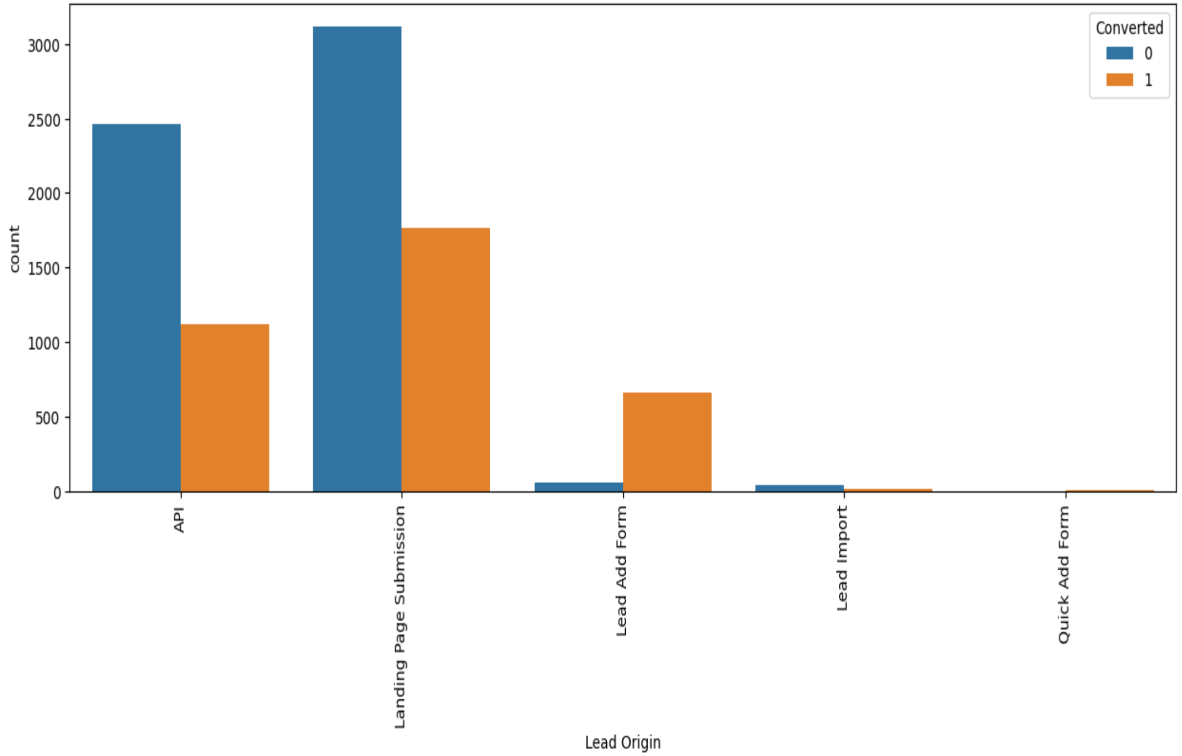
- The dataset used for the problem is “leads.csv” which has 37 columns(features)
- Initially 5 columns which has more than 40% missing data has been dropped
  - Asymmetrique Activity Index
  - Asymmetrique Profile Index
  - Asymmetrique Activity Score
  - Asymmetrique Profile Score
  - Lead Quality
- Further 14 columns which are having unique values and highly imbalanced columns has been removed
  - Prospect ID, Lead Number, Magazine, Receive More Updates About Our Courses, Update me on Supply Chain Content, Get updates on DM Content, I agree to pay the amount through cheque, Do Not Call, Search, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations
- On few columns the missing values are filled with median or mode accordingly based on datatype of the columns.
  - Page Views Per Visit
  - TotalVisits

EDA: Handling “Select” value in the columns



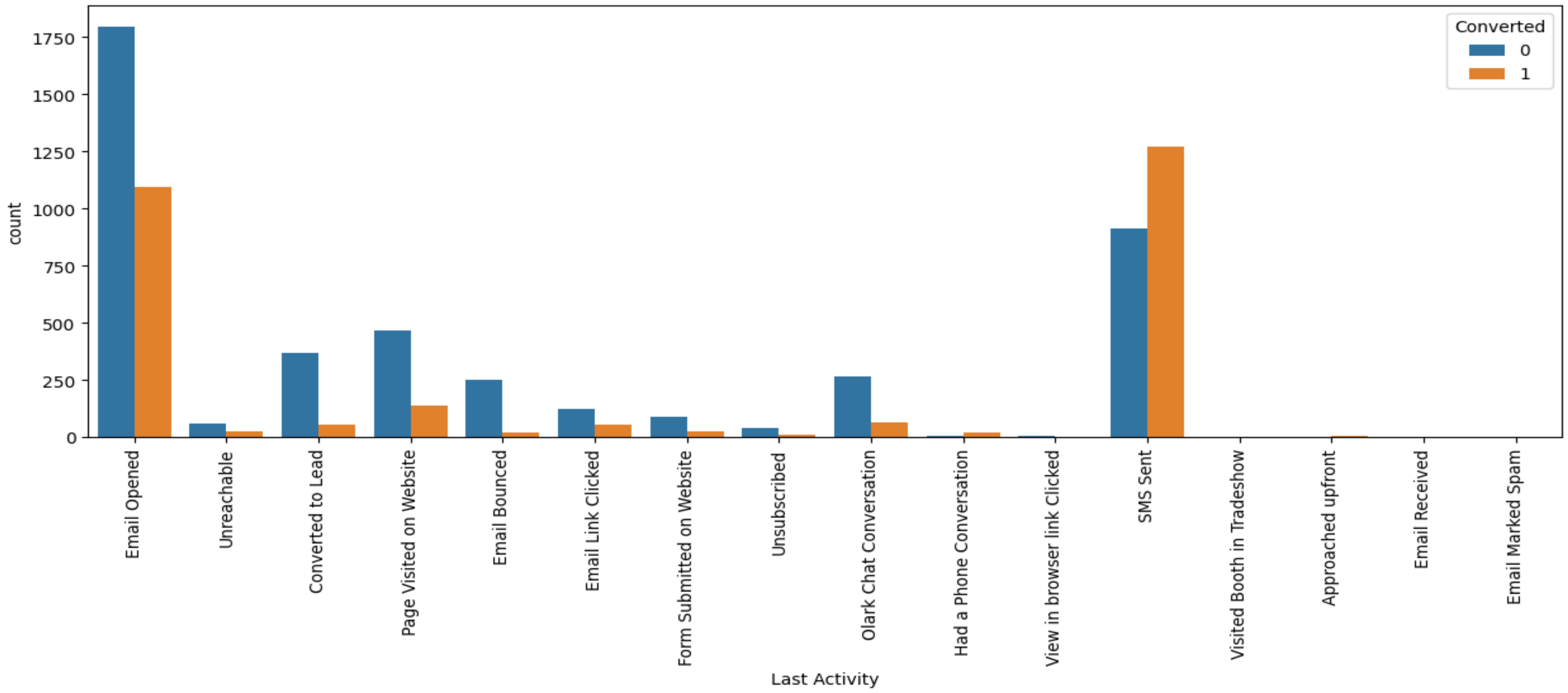
- The “Select” value has been treated as same as NaN (As it was not been selected by user(s)). The columns which are having these values are
  - Specialization
  - How did you hear about X Education,
  - Lead Profile,
  - City
- Once the columns are updated, we further performed data missing/imputation treatment to the data
- 2 columns - 'How did you hear about X Education', 'Lead Profile’, are having more than 40% missing data has been dropped

Exploratory Data Analysis: Lead Origin



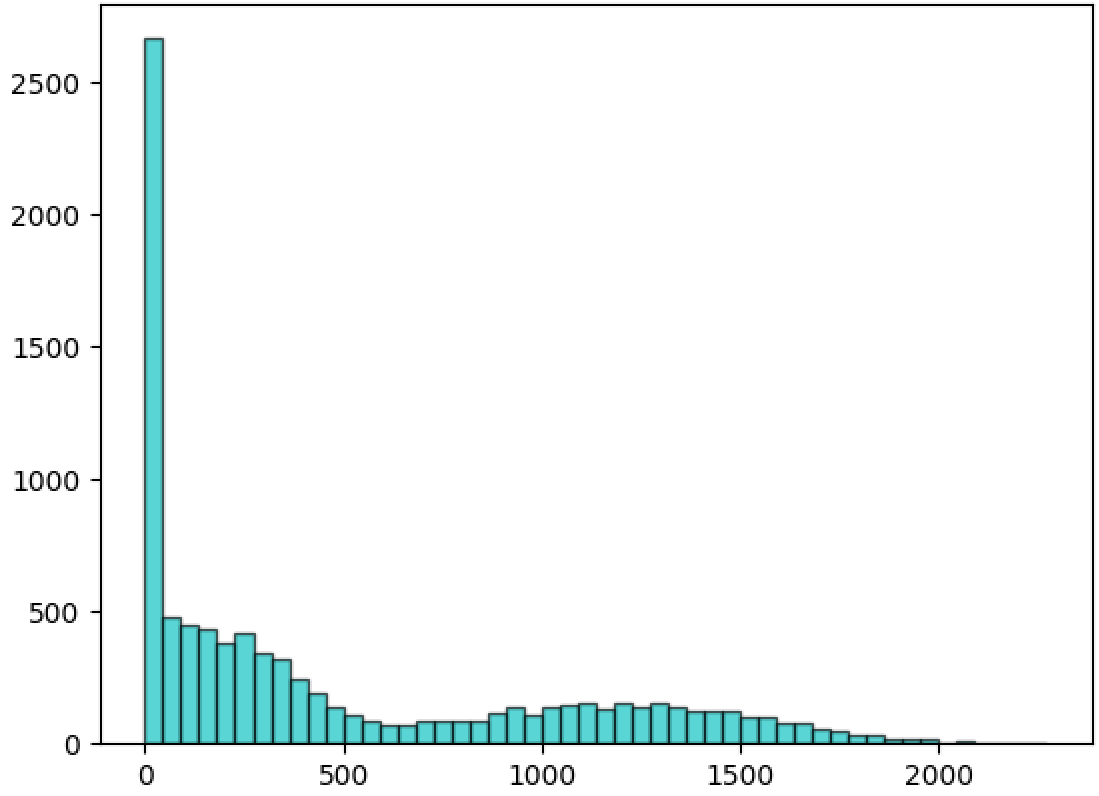
- Most leads are from API & Landing Page Submission
- Most converted leads are from "Lead Add Form".
- There are negligible leads from “Lead Import” & “Quick Add Form”

Last Activity



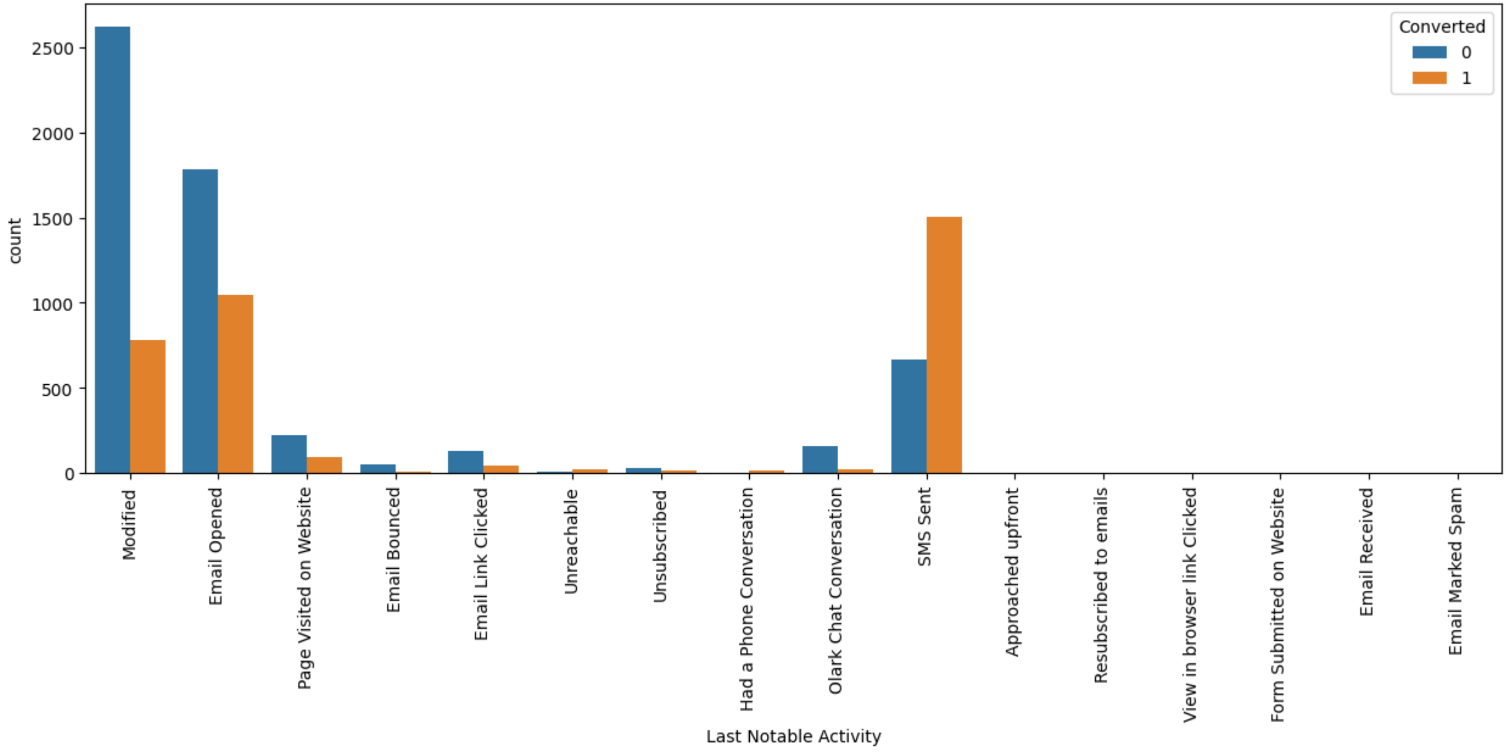
- The major converted leads are from Email & SMS Channel(s) as we can see the rate of "Email Opened" & "SMS sent“
- These options can be used efficiently for better converted ratio

Total Time Spent on Website



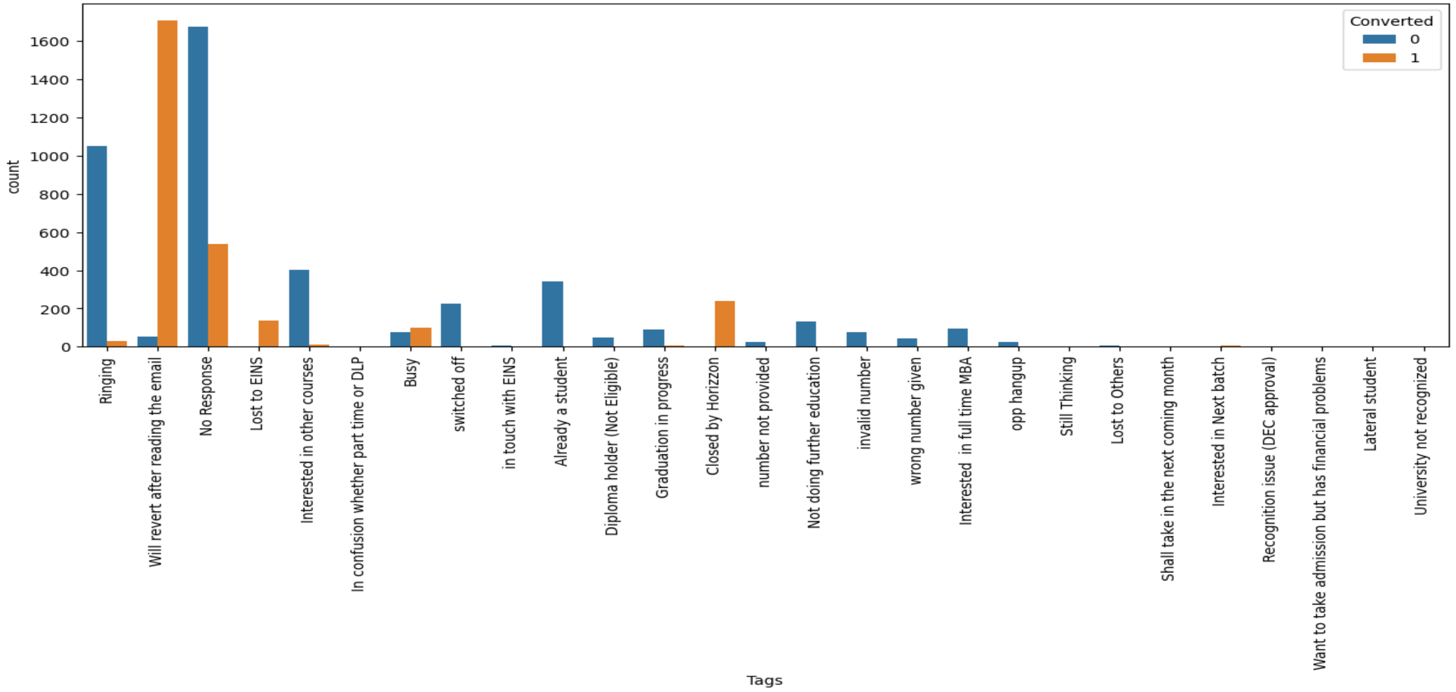
- Maximum number of the leads are spent less than 45 mins i.e. ~2650 customers
- More than 2000 customers spent between 45 mins & 272 mins

Last Notable Activity



- Most of the leads and converted leads are from Email & SMS Channel(s)
- High rate of converted leads are via SMS sent
- There are no/negligible leads from Approached upfront, Resubscribed to emails, View in browser link, Form submitted on website, Email received and Email marked as spam

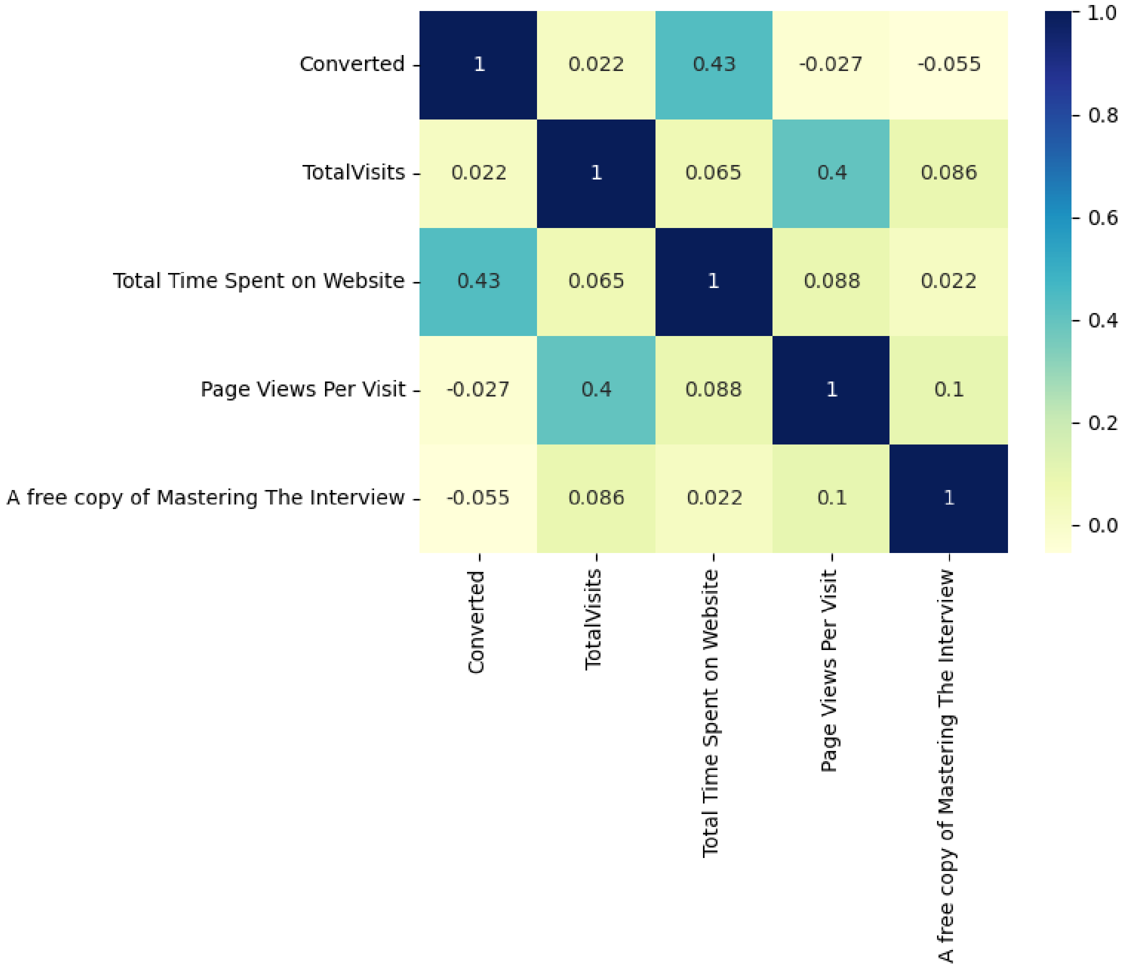
Tags



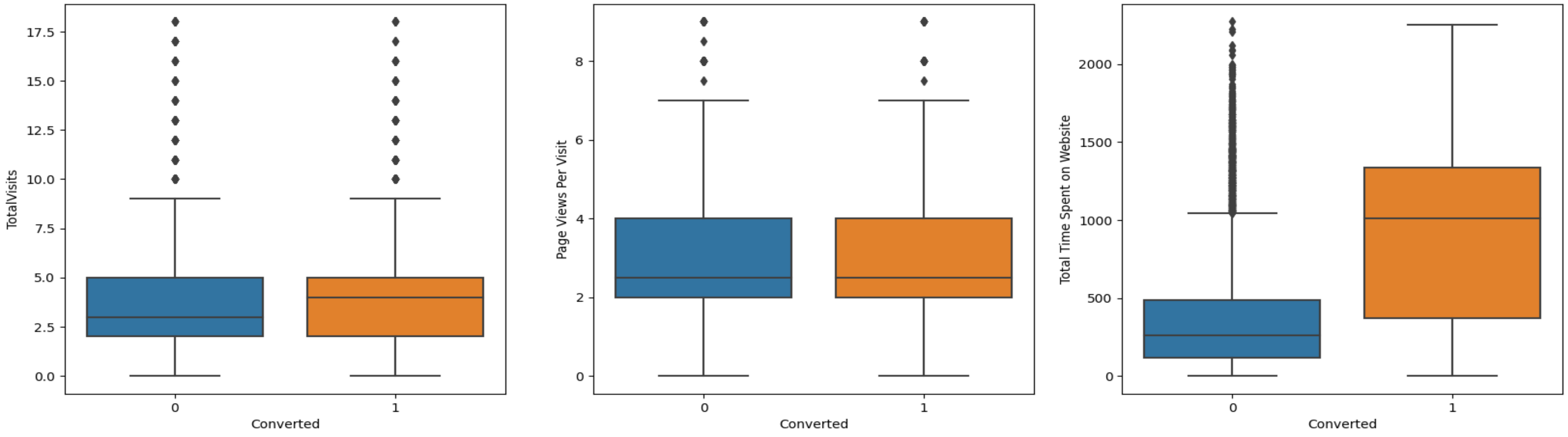
- Most of the converted leads ratio are from the customers who "will revert after reading email", "Closed by Horizzon" & "No Response".
- Almost 1000 leads which are received from "Ringing“, the converted ratio is very less.
- Almost 200 leads are converted which are from “Lost to EINS”

Correlation

- There are no strong correlations between the numerical columns.
- There is good correlation between "Total time spent on website" and "Converted“
- There is good correlation between "Page Views Per Visit" and "TotalVisits"
- Least correlation is between Converted and TotalVisits.



Outliers: Numerical columns analysis



The outliers in the “TotalVisits”, “Page views per visit” & “Total Time spent on website” has been removed. The datapoints which are above 99% & less than 1% has been considered as an outliers in these columns.

Model building

- Splitting the dataset into train and test dataset
- Scaling numerical columns using StandardScaler
- Model building using RFE to find most significant features and use Stats model for analysis
- Check VIF for the existing features
- Eliminate the variable with high p-value and high VIF value
- Rebuild the model after removing the variables which are not useful
- Predict the train dataset
- Calculate Accuracy, Specificity, Sensitivity using Confusion matrix
- Plot ROC curve
- Find the optimal cutoff point and adjust the probabilities if required
- Predict using test dataset and evaluate metrics - Accuracy, Specificity, Sensitivity
- Perform precision and recall tradeoff
- Calculate Accuracy, Specificity, Sensitivity using Confusion matrix
- Precision and Recall analysis on test dataset predictions.

Model building – Final Model

Model7

In [527]:
```
# Creating model with resultant col values
X_train_sm = sm.add_constant(X_train[col])
logm7 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
res = logm7.fit()
res.summary()
```

Out[527]:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	4935
Model:	GLM	**Df Residuals:**	4925
Model Family:	Binomial	**Df Model:**	9
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-955.87
Date:	Tue, 22 Oct 2024	**Deviance:**	1911.7
Time:	09:15:46	**Pearson chi2:**	7.32e+03
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.6139
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.5652	0.184	-24.873	0.000	-4.925	-4.205
Total Time Spent on Website	1.1445	0.063	18.121	0.000	1.021	1.268
Lead Origin_Lead Add Form	2.3928	0.556	4.306	0.000	1.304	3.482
Last Activity_Converted to Lead	-1.1530	0.336	-3.434	0.001	-1.811	-0.495
Tags_Busy	3.5070	0.276	12.686	0.000	2.965	4.049
Tags_Closed by Horizzon	9.2238	1.027	8.978	0.000	7.210	11.237
Tags_Lost to EINS	7.9949	0.635	12.584	0.000	6.750	9.240
Tags_No Response	2.8151	0.182	15.467	0.000	2.458	3.172
Tags_Will revert after reading the email	7.2901	0.249	29.269	0.000	6.802	7.778
Last Notable Activity_SMS Sent	2.0766	0.138	15.006	0.000	1.805	2.348

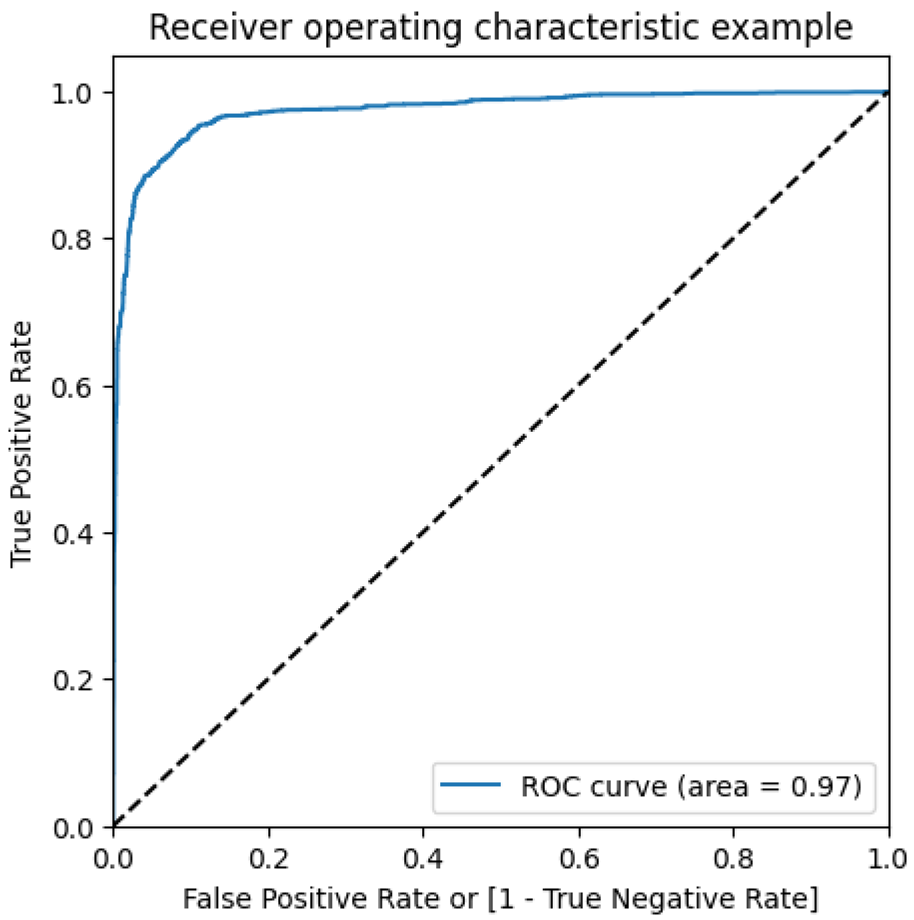
Checking VIFs

In [528]:
```
# Create a dataframe that will contain the names of all the feature variables and their respective VIFs
vif = pd.DataFrame()
vif['Features'] = X_train[col].columns
vif['VIF'] = [variance_inflation_factor(X_train[col].values, i) for i in range(X_train[col].shape[1])]
vif['VIF'] = round(vif['VIF'], 2)
vif = vif.sort_values(by = "VIF", ascending = False)
vif
```

Out[528]:

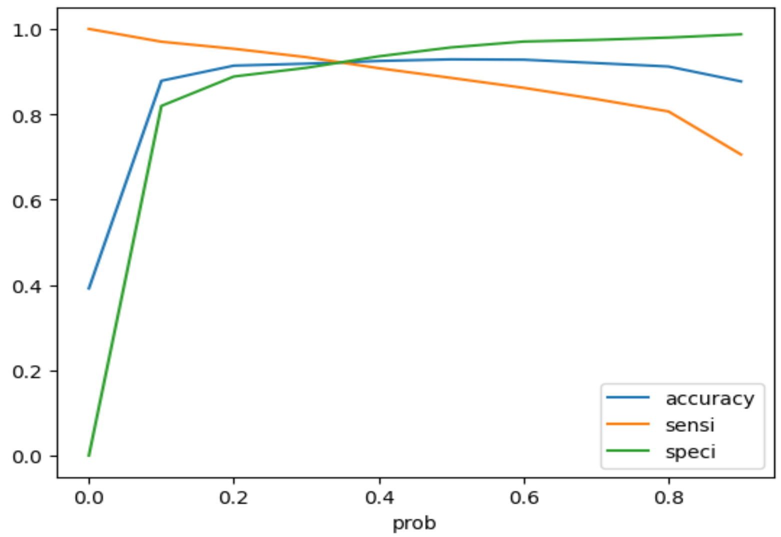
	Features	VIF
7	Tags_Will revert after reading the email	1.45
8	Last Notable Activity_SMS Sent	1.38
1	Lead Origin_Lead Add Form	1.26
4	Tags_Closed by Horizzon	1.16
0	Total Time Spent on Website	1.15
6	Tags_No Response	1.08
2	Last Activity_Converted to Lead	1.04
3	Tags_Busy	1.04
5	Tags_Lost to EINS	1.01

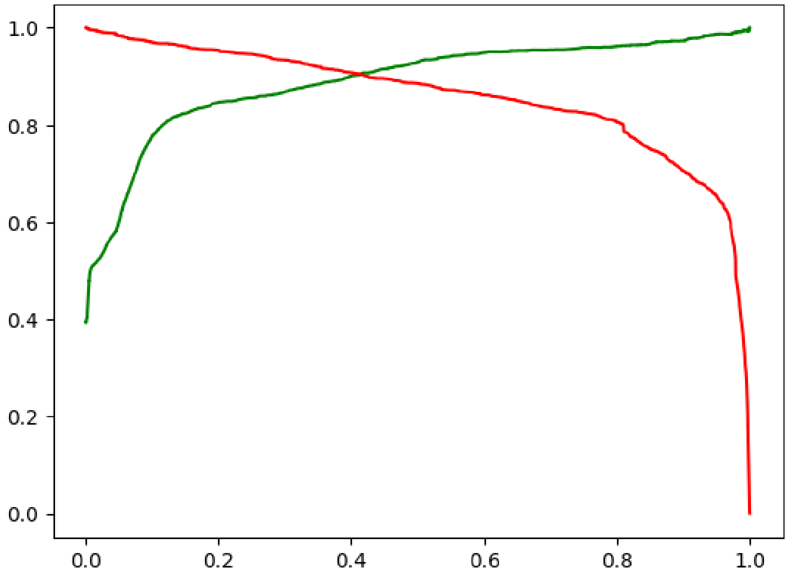
Model building – ROC curve



- The area under ROC curve is 0.97

Model Evaluation(Train Dataset)





Accuracy Sensitivity and Specificity

Accuracy: 92.41%
Sensitivity: 91.61%
Specificity: 92.53%

Confusion matrix

[[2778, 224],
[162 , 1771]]

Precision and Recall(After tradeoff)

Precision: 90.72%
Recall: 90.11%

Model Evaluation (Test Dataset)

Accuracy Sensitivity and Specificity

Accuracy: 91.11%
Sensitivity: 88.55%
Specificity: 92.67%

Confusion matrix

[[1215, 96],

[92 , 712]]

Precision and Recall(After tradeoff)

Precision: 90.21%
Recall: 87.18%

Conclusions and Recommendations

- To improve the potential lead conversion rate X-Education will have to mainly focus important features responsible for good conversion rate are :-
    - **Total Time Spent on Website**: The customers spending more time on website can turn to be potential leads.
    - **Lead Origin_Lead Add Form**: Leads who have engaged through 'Lead Add Form' having higher conversion rate so company can focus on it.
    - **Last Activity_Converted to Lead**: The last activity by the customer who successfully converted to lead.
    - **Tags_Closed by Horizzon**: The converted leads that are closed by Horizzon as they play major role in conversion.

Thanks

From Team

- Giridhar Challa
- Ganesh Behera
- Vinay Kumar Sharma