Lead Scoring Case Study

Submitted by,
Joel Jojo
Radhika Sharma
Vishal V Ravi

Problem Statement

- X Education is an education company that sells online courses to industry professionals.
- The company markets its courses through several websites, search engines like Google and also does advertisements through other mediums like newspapers, magazines etc.
- The company identifies leads when they fill up a form on their website, through their api etc.
- The sales team from the company contacts these identified leads and tries to convince them to enrol for the course.
- The current conversion rate is 30% i.e. if 100 leads are identified on a given day only 30 of them turn out to be paying customers in the end.
- The company wants to identify leads with higher potential of enrolling for the course so that sales team could concentrate on these 'hot leads' to increase the conversion rate.
- The CEO has set an ambitious target of increasing the conversion rate to 80%.

Goals of the Case Study

To build a logistic regression model to predict the chances of conversion of a certain identified lead.

Assign a Lead Score between 0 and 100 to each lead indicating its chances of conversion.

Setting a threshold value so that leads having a higher score than the threshold value can be identified as ‘Hot Leads’ and the sales team can concentrate more on these leads.

Solution Methodology

- Data Cleaning and Manipulation
  - Dropping columns containing only one unique value.
  - Checking the percentage of null values and dropping columns having more than 40% null values.
  - Dropping columns having unique values for each entry
  - Binning the features in a necessary manner.

- Outlier Analysis and Treatment
  - Check the distribution of the continuous variables visually and using percentiles.
  - Impute outliers

- EDA
  - Univariate Analysis:- Checking the value counts and the distribution of the variables.
  - Bivariate Analysis:- Checking the correlation and the pattern of relationships between different variables.

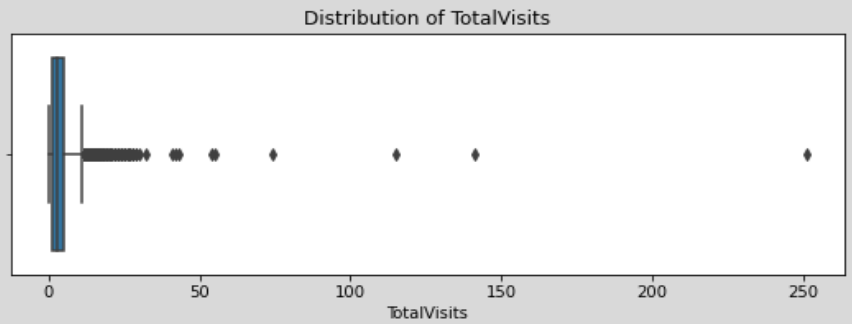
- Feature Scaling, Dummy Variable Creation and Train-Test Split
- Classification Technique: Logistic Regression for Model Making and Prediction
- Evaluation of Model and Presentation
- Conclusions and Recommendations

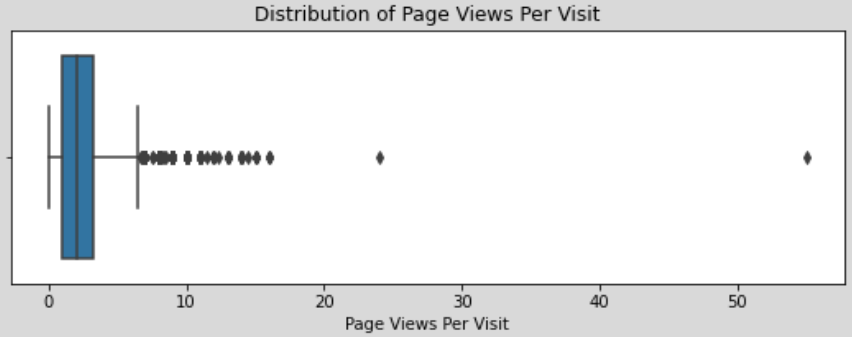
Data Manipulation

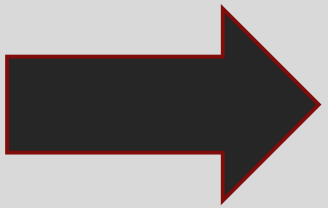
- Total No. of Rows in the Dataset: 9240
- Total No. of Columns in the Dataset: 37
- Columns with only one unique value like 'Magazine',‘ Receive More Updates About Our Courses', 'Update me on Supply Chain Content',‘ Get updates on DM Content', and ‘ I agree to pay the amount through cheque‘ were dropped.
- Columns having a percentage of null values greater than 40% were checked and dropped.
- Dropped columns ‘Prospect ID’ and ‘Lead Number’ as it has unique values for each entry.
- Dropped ‘Last Activity’ as it is highly redundant with ‘Last Notable Activity’.
- Replaced null values in columns 'Country', 'Specialization', 'What is your current occupation',‘ What matters most to you in choosing a course', 'Tags', and 'City‘ with ‘ not declared’.
- Dropped the rows with null values after all the previous steps as the percentage of data loss was only a mere 1.5% and then segmented the categorical and continuous variables.

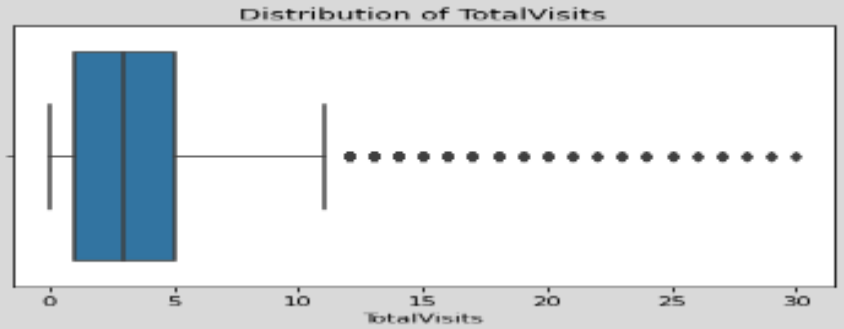
Outlier Treatment

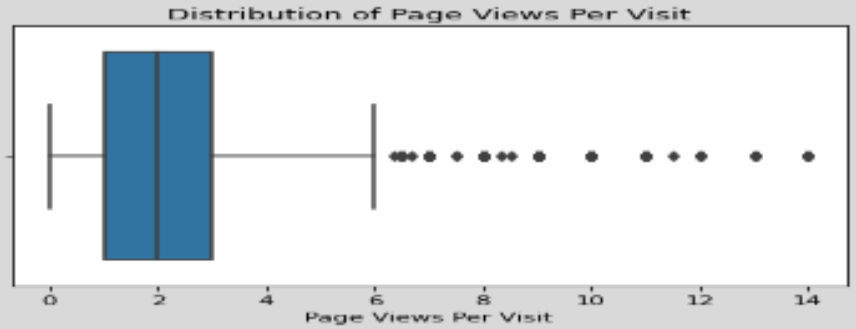
- The distribution of the continuous features was analysed visually by using the boxplot.
- The distribution was then analysed statistically using the describe() function at different percentiles.
- Outliers were detected in the 'TotalVisits' and "Page Views Per Visit‘ columns.
- Outliers were imputed at 99.9% using the quantile() function.





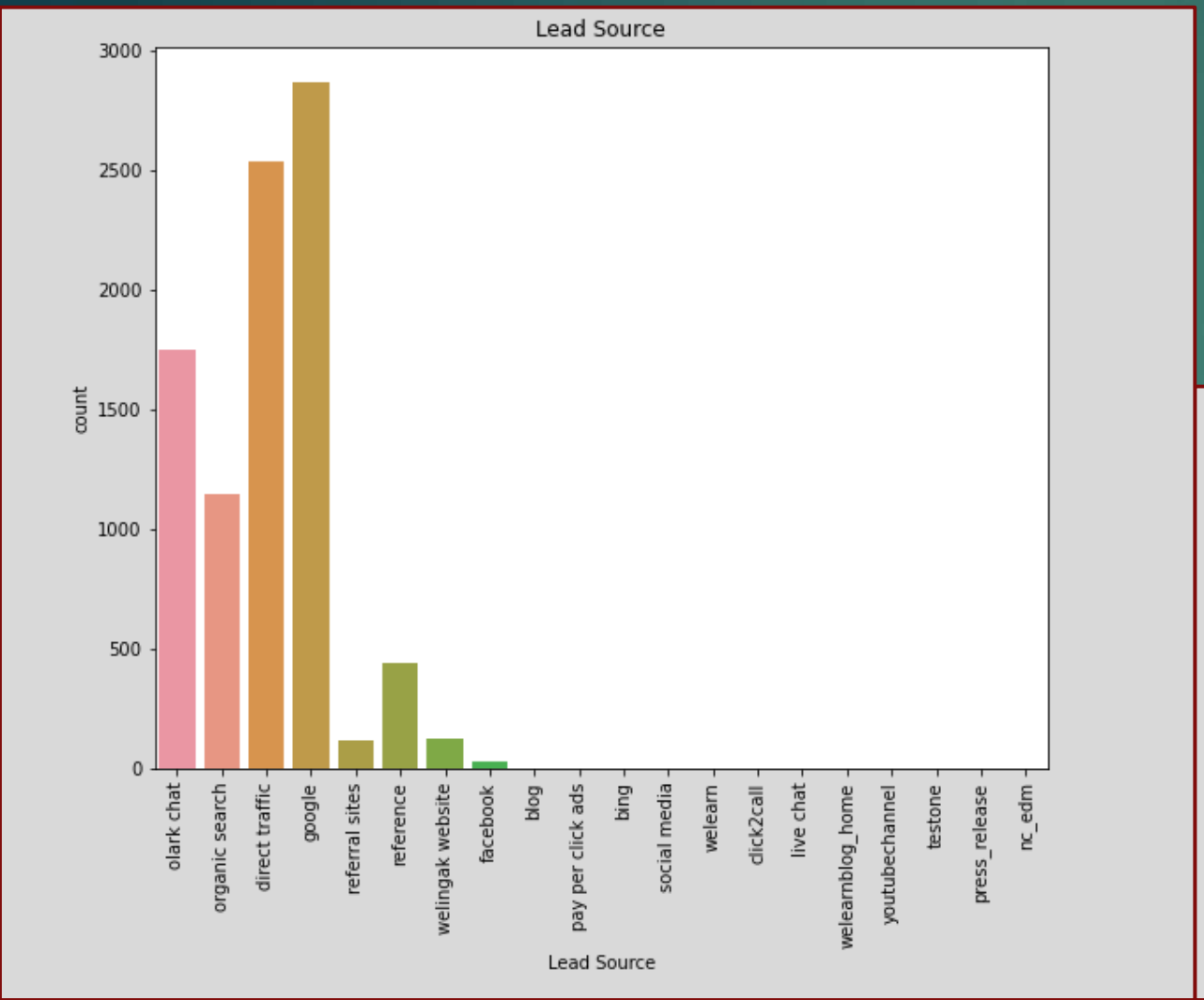




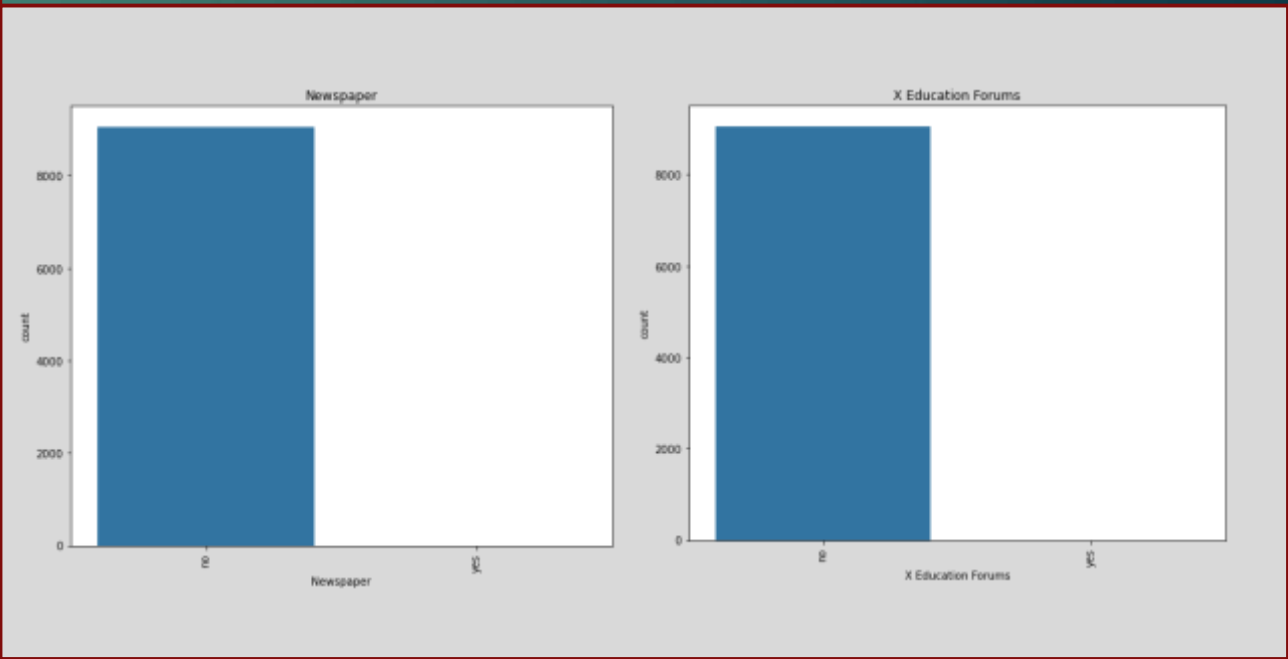


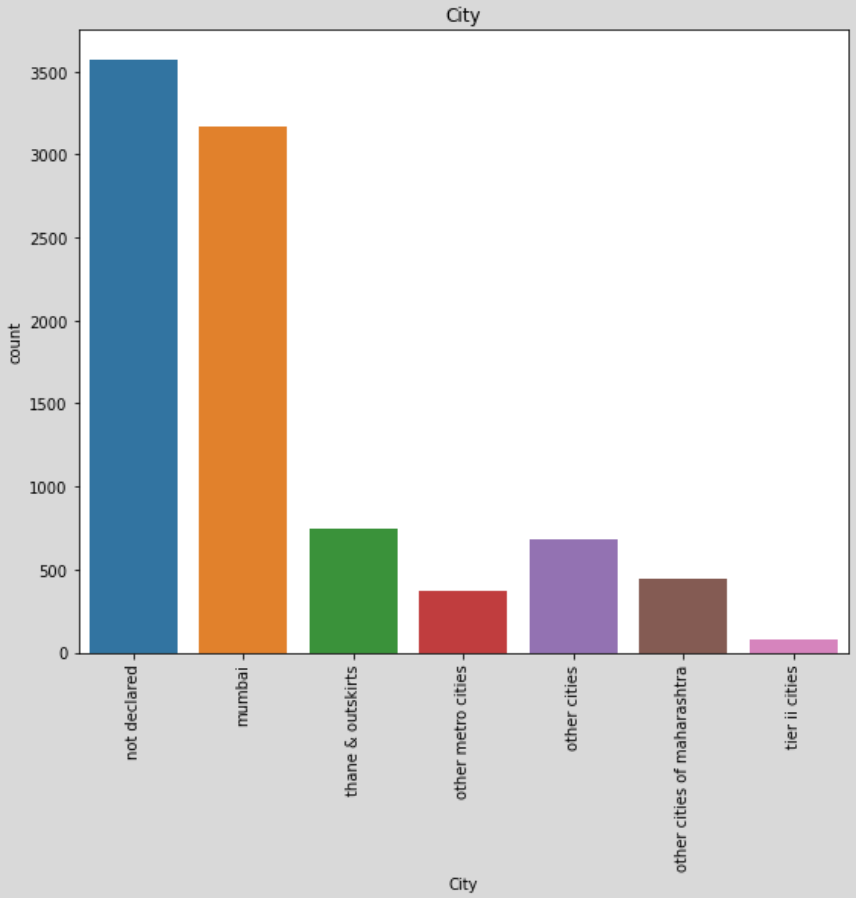
Exploratory Data Analysis

Univariate Analysis – Categorical Variables



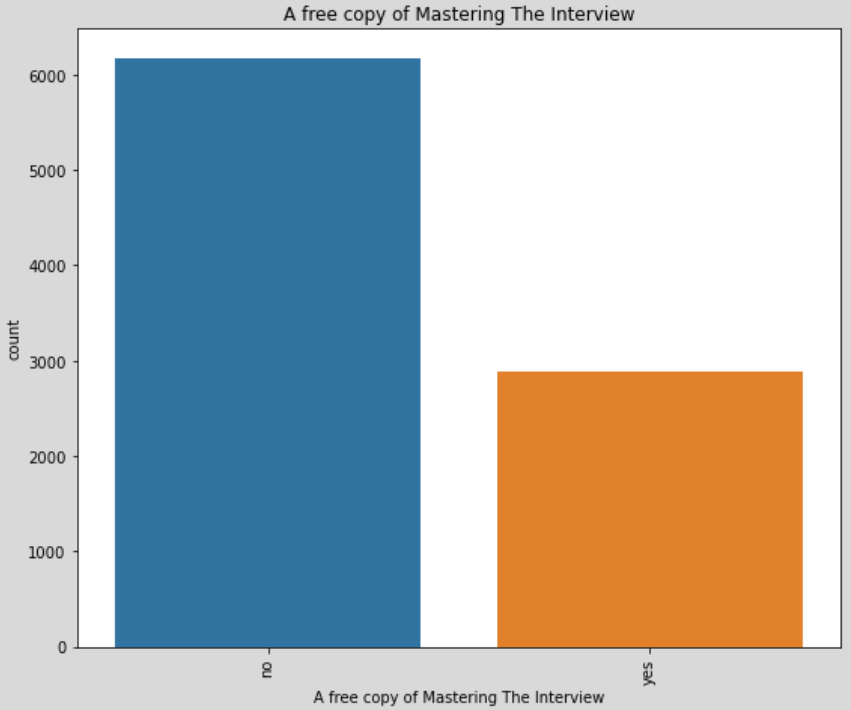
Most of the leads have been sourced from Google and direct traffic and very few have come through newspapers, X Education Forums, and Youtube channels which means the company can cut its budget here.



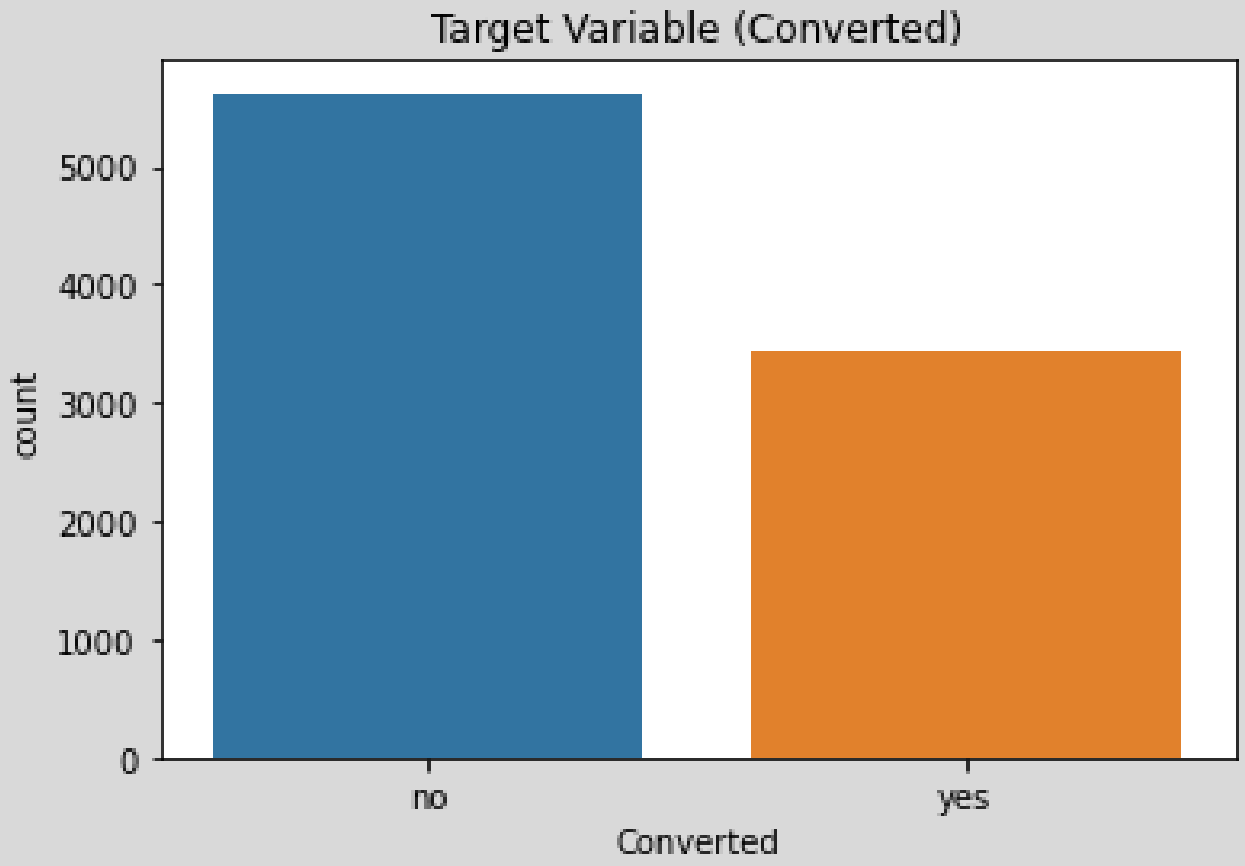


The most number of leads are from the city of Mumbai which indicates marketing campaigns should be concentrated there more.

Many people have subscribed to the free copy of '**Mastering the Interview**' which makes it a good medium, for advertisement.



Data Imbalance

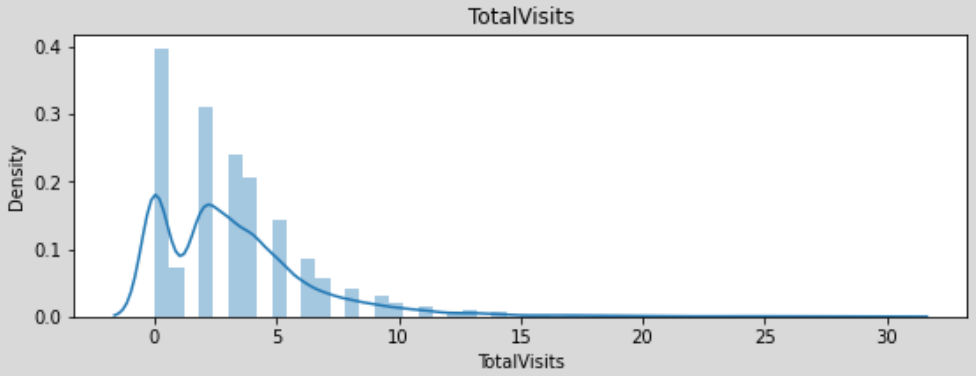


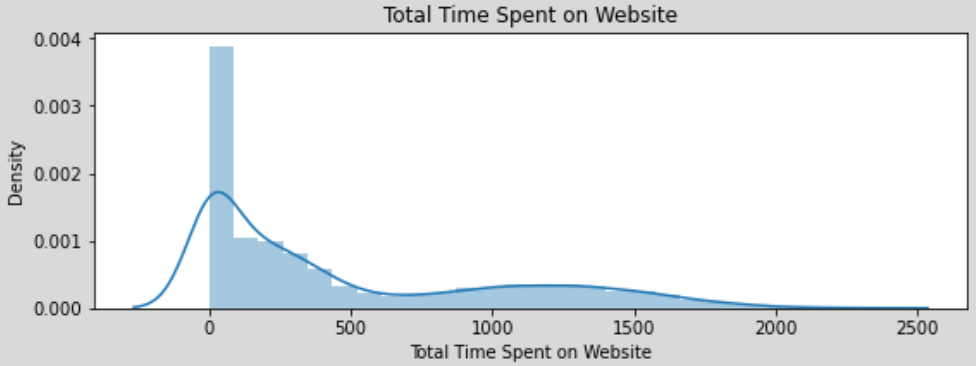
Data Imbalance has been detected in

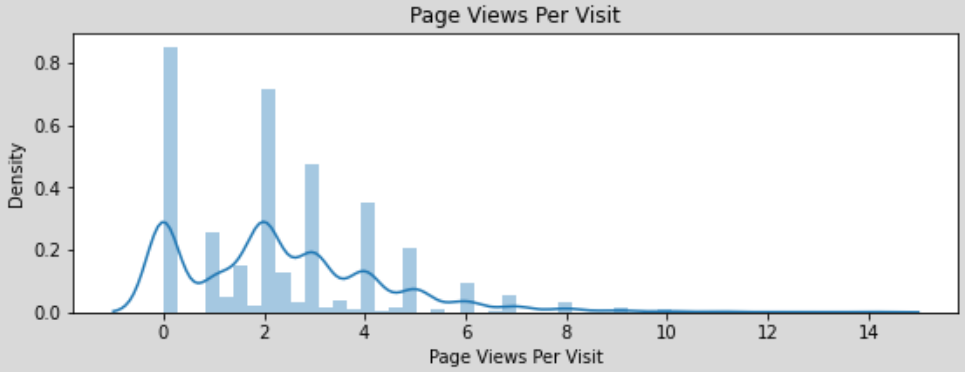
the Ratio:

'Converted' : "Not Converted = 61:100

Univariate Analysis – Continuous Variables

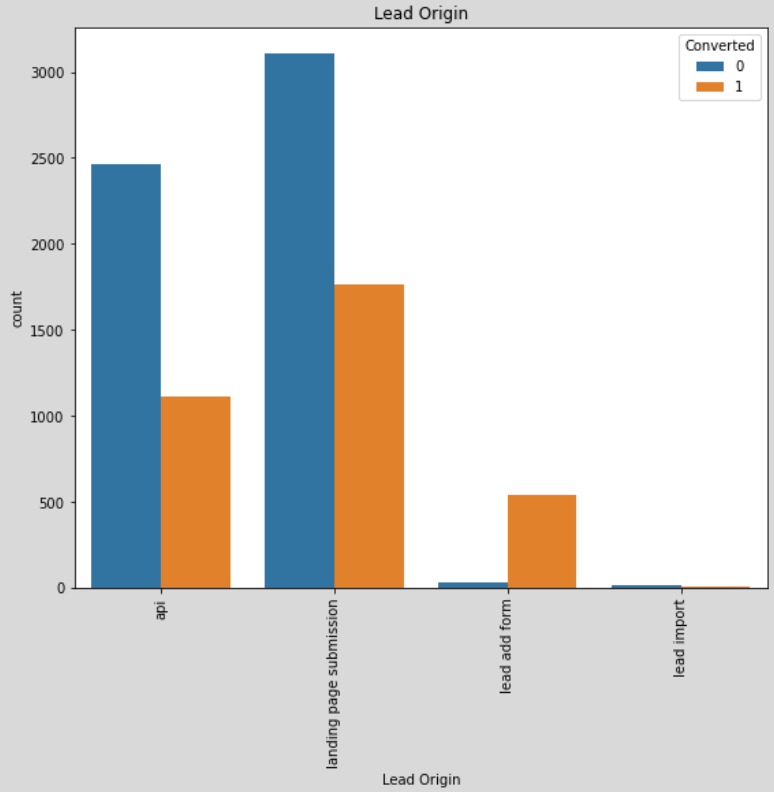




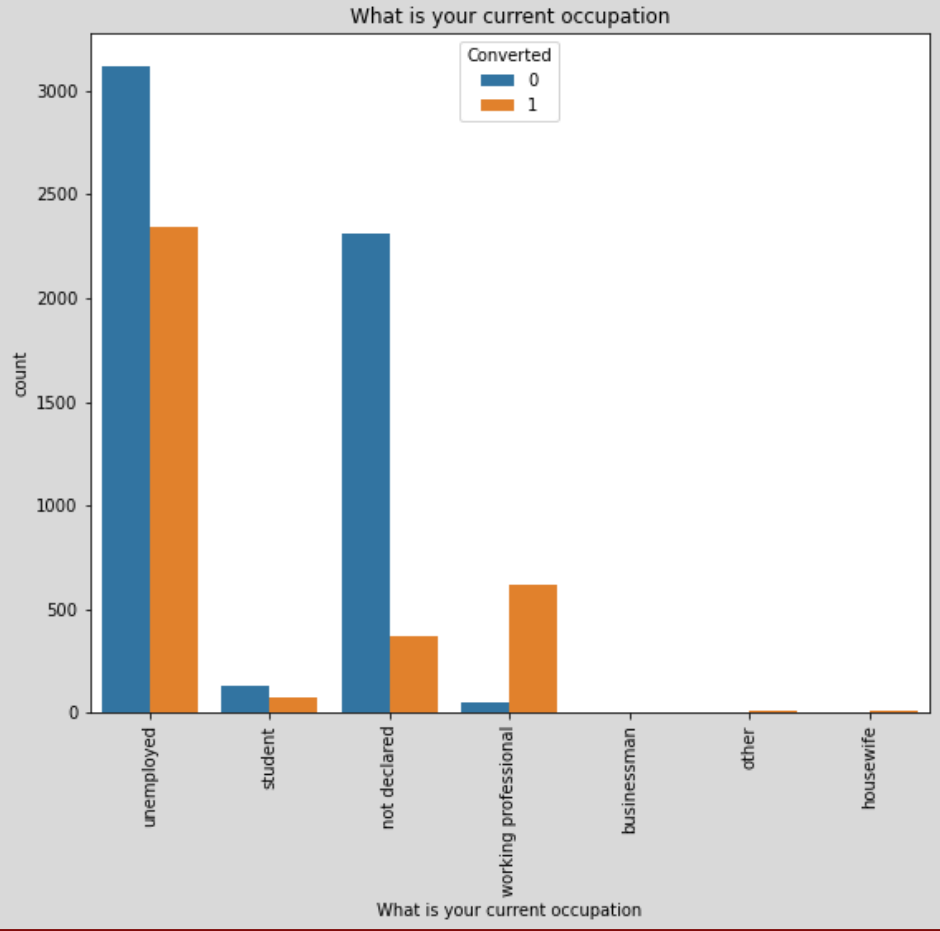


1. Many customers made a maximum of approx. 10 visits to the website.
2. A large group of customers have spent less than 8 hrs (500 mins) at maximum on the website.
3. Most customers have less than 5 page views per visit to the website.

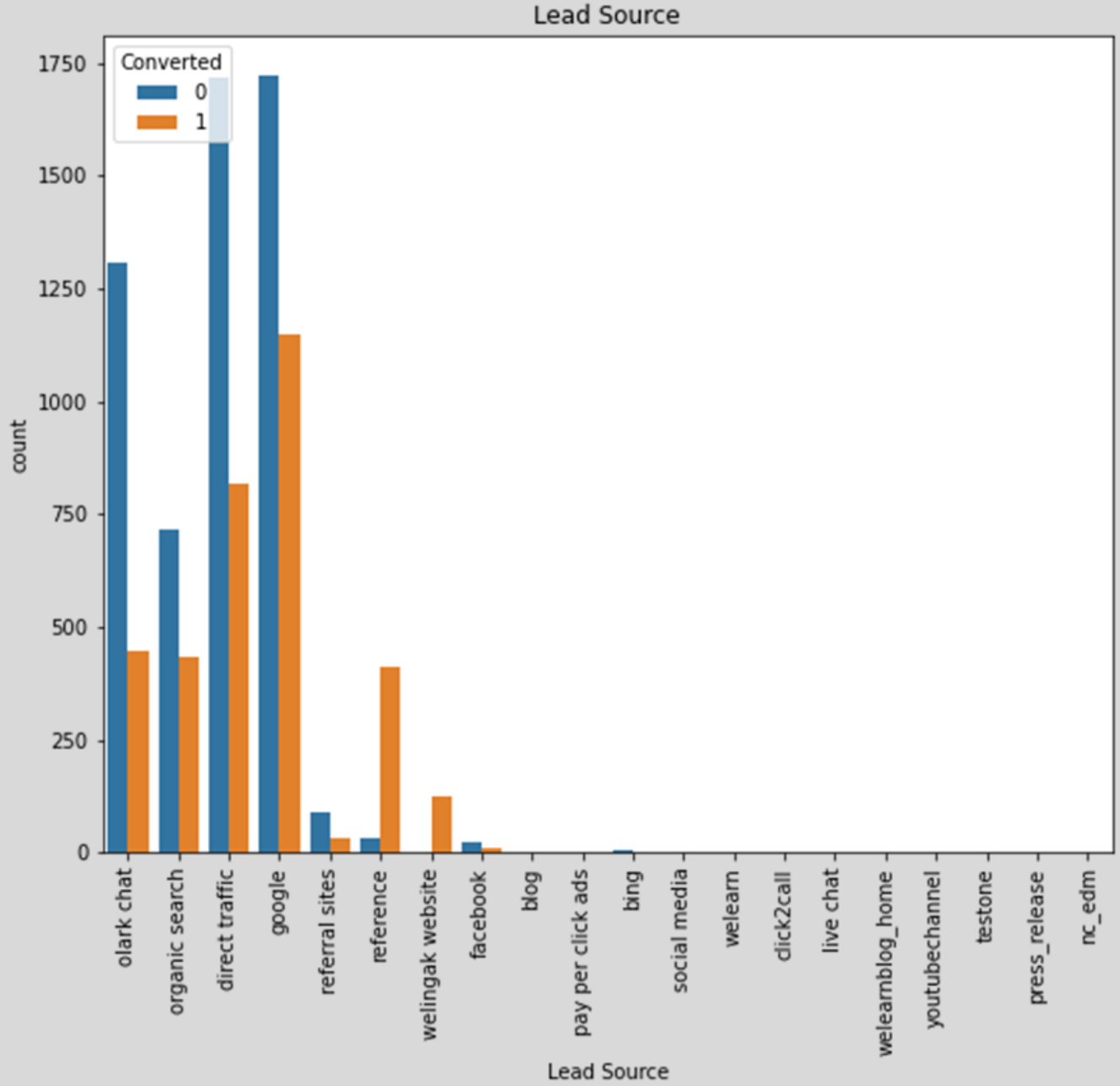
Bivariate Analysis – Categorical Variables to target variable



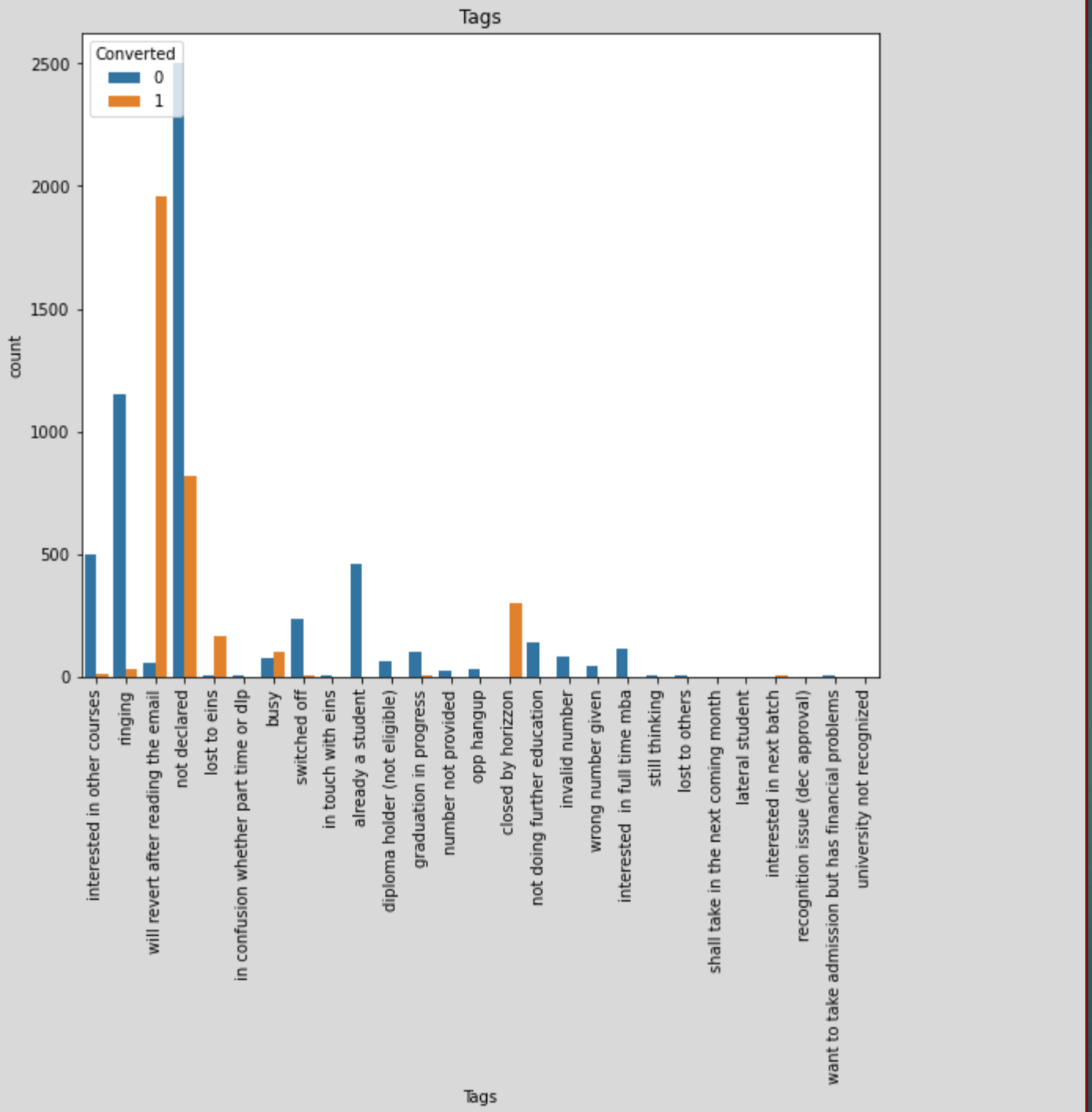
The rate of conversion for leads that have been ascertained through the lead add form are high.



Leads who are Working professionals have a high chance of getting converted.

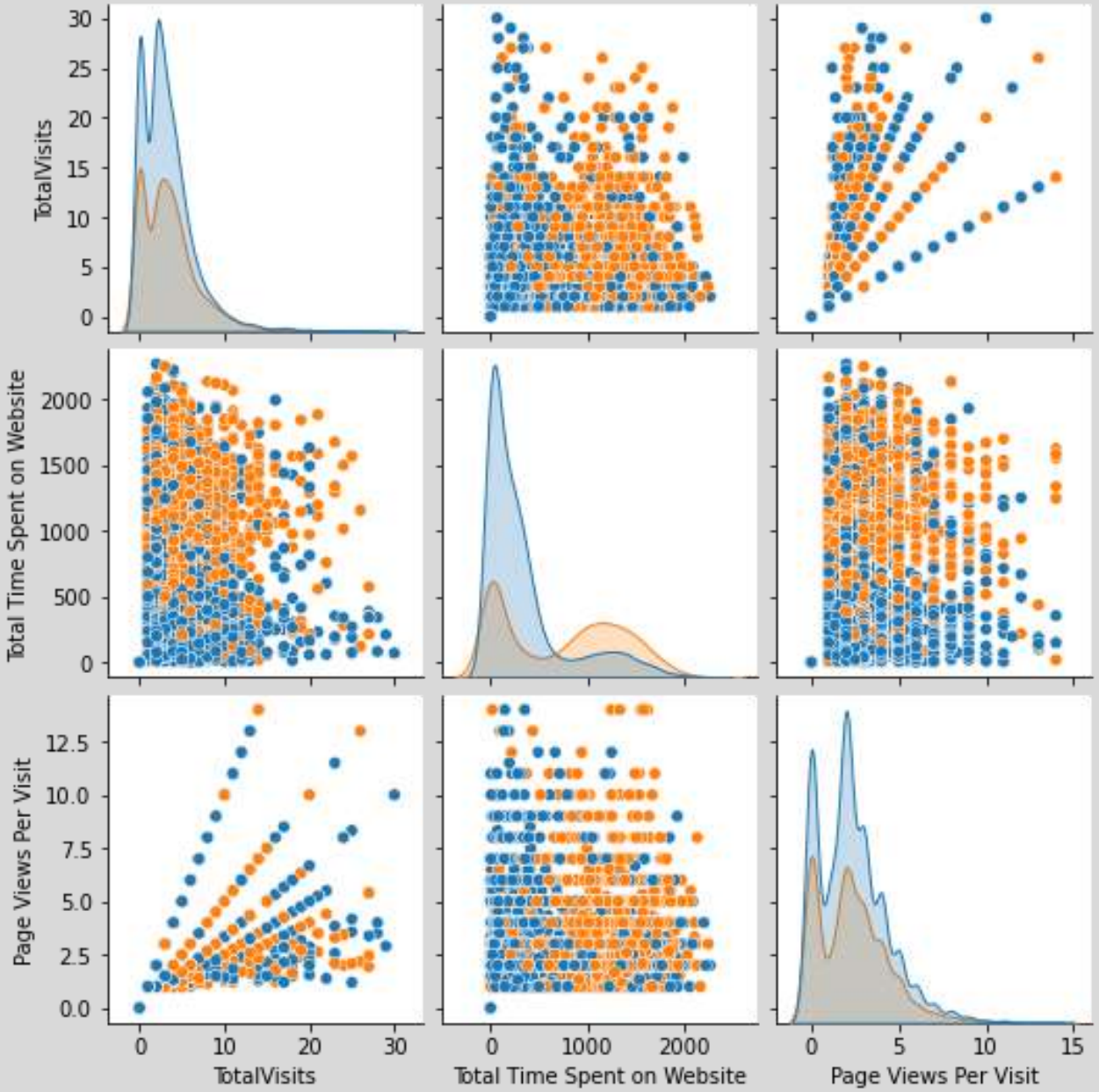


Students coming through references and through the welingak website are mostly getting converted.



Leads with the tag 'will revert after reading the email' have a very high conversion rate.

Bivariate Analysis – Continuous Variables



There seems to be a correlation between 'TotalVisits' and 'Page Views Per Visit‘ in the pair plot which was ascertained by the heatmap.

Data Pre-processing

- Dummy Variables were created for the categorical variable
- Number of Rows after Data Cleaning and Creation of dummies: 9055
- Number of Columns after Data Cleaning and Creation of dummies: 97 (dropping the original variables)
- The Training and Testing Dataset was split in 70:30 Ratio.
- The dataset was scaled using MinMaxScaler.

Model Building

- Used Recursive Feature Elimination for feature selection.
- Selected 15 features using RFE
- Built model using the Generalized Linear Models Method from the statsmodels library by eliminating features having p-values greater than 0.05 and VIF values greater than 5.

Finalized Model – Meeting the Required Parameters

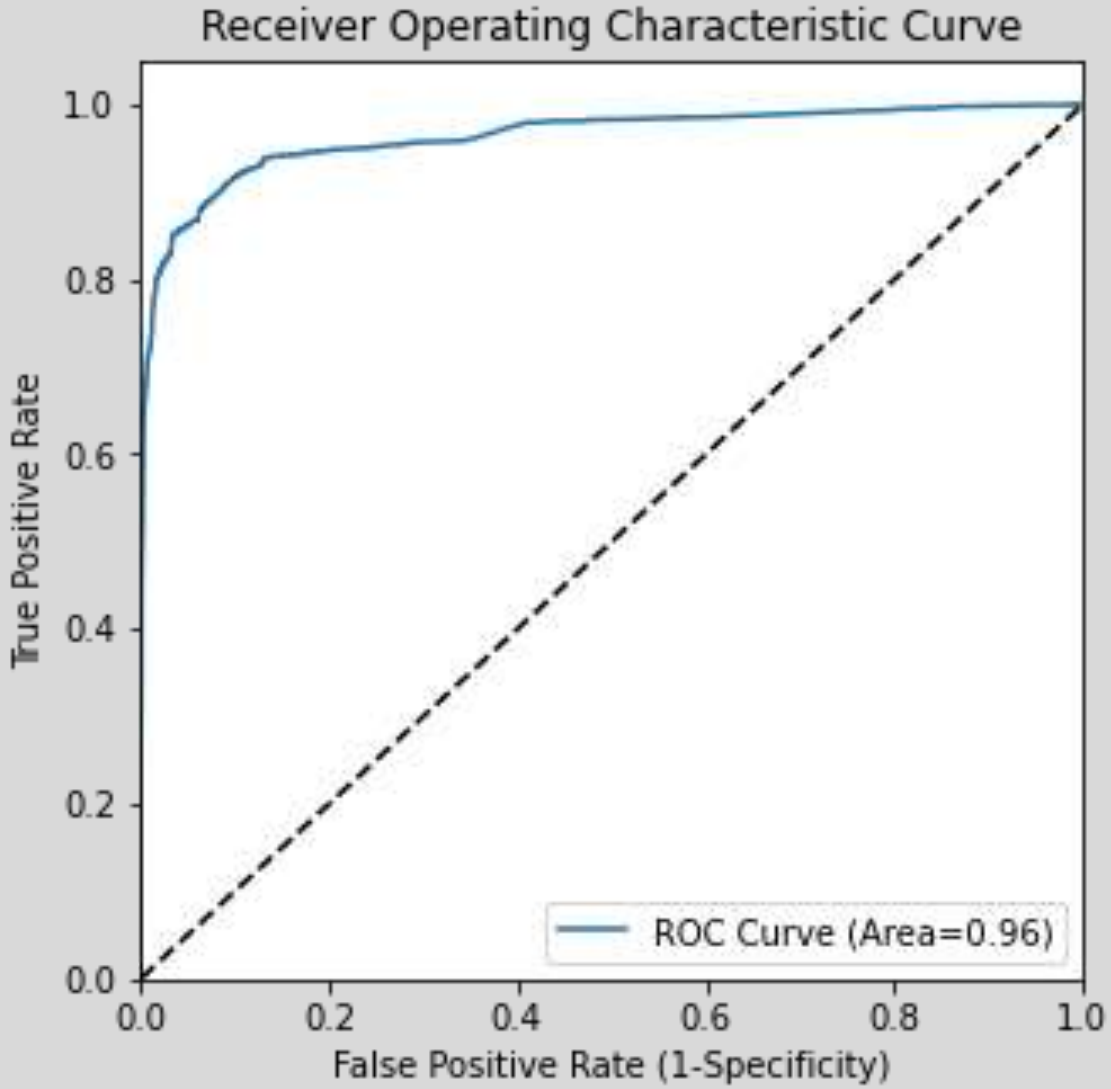
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6338
Model:	GLM	Df Residuals:	6326
Model Family:	Binomial	Df Model:	11
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1372.3
Date:	Sun, 16 Apr 2023	Deviance:	2744.6
Time:	12:09:27	Pearson chi2:	8.14e+03
No. Iterations:	8	Pseudo R-squ. (CS):	0.5905
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.5771	0.097	-26.626	0.000	-2.767	-2.387
Total Time Spent on Website	3.4522	0.205	16.812	0.000	3.050	3.855
Lead Origin_lead add form	1.6819	0.392	4.289	0.000	0.913	2.450
Lead Source_welingak website	4.3478	1.094	3.974	0.000	2.203	6.492
Do Not Email_yes	-1.2829	0.224	-5.738	0.000	-1.721	-0.845
What matters most to you in choosing a course_not declared	-0.7321	0.113	-6.463	0.000	-0.954	-0.510
Tags_closed by horizzon	6.6048	1.012	6.528	0.000	4.622	8.588
Tags_lost to eins	5.9444	0.727	8.175	0.000	4.519	7.370
Tags_ringing	-3.6886	0.253	-14.566	0.000	-4.185	-3.192
Tags_switched off	-4.0762	0.607	-6.716	0.000	-5.266	-2.887
Tags_will revert after reading the email	4.3696	0.178	24.520	0.000	4.020	4.719
Last Notable Activity_sms sent	2.8408	0.122	23.261	0.000	2.601	3.080

	Features	VIF
0	Total Time Spent on Website	1.84
9	Tags_will revert after reading the email	1.83
1	Lead Origin_lead add form	1.68
10	Last Notable Activity_sms sent	1.51
2	Lead Source_welingak website	1.31
5	Tags_closed by horizzon	1.22
4	What matters most to you in choosing a course_...	1.16
7	Tags_ringing	1.13
3	Do Not Email_yes	1.07
6	Tags_lost to eins	1.05
8	Tags_switched off	1.04

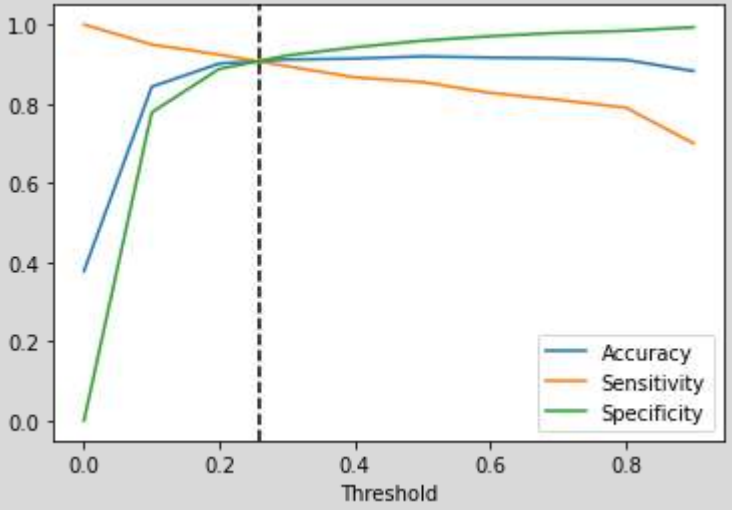
ROC Curve



The ROC Curve tended to the upper-left corner and had an AUC Score of 0.96 which indicated that the model was very good.

Optimal Threshold and Evaluation

Accuracy-Sensitivity-Specificity

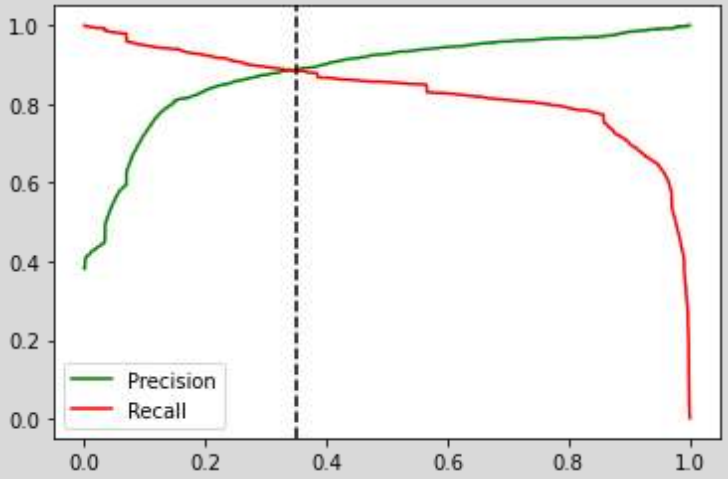


Optimum Threshold Value = **0.26** (Balanced Accuracy, Sensitivity and Specificity indicated by the intersecting point)

Performance on Test Set

- Accuracy - 91.2%
- Sensitivity - 90.5%
- Specificity - 91.6%

Precision-Recall



Optimum Threshold Value = **0.35** (Balanced Precision and Recall indicated by the intersecting point)

Performance on Test Set

- Accuracy - 91.9%
- Precision - 89.7%
- Recall - 89.0%

Conclusion

- The Lead Score was computed for each of the entries and a threshold value of 35 was fixed to segregate the hot leads from the cold ones based on the results from the Accuracy-Sensitivity-Specificity Trade-off and the Precision-Recall Trade-off.
- It was found that the characteristics that mattered the most in the potential buyers are (In descending order) :
  1. Leads having the tag 'closed by horizzon‘
  2. Leads having the tag 'lost to eins‘
  3. Leads having the tag 'will revert after reading the email‘
  4. Leads that have been sourced from the welingak website.
  5. More Total Time Spent on Website
  6. Leads having the last notable activity as 'sms_sent‘
  7. Leads that have been ascertained through the lead add form.
- The model produced predictions with very good accuracy and also performed well with the other evaluation metrics such as sensitivity, specificity, precision, recall, etc..
- Leads ascertained through the lead add form, coming through references and through the welingak website, who are working professionals or whose tag is ‘will revert after the reading the email have high chances of getting converted.