Lead Score Case Study

Building a Logistic Regression Model to Identify High-Potential Leads and Assign Lead Scores for X Education to Maximize Conversion Rates.

Business Objective

The objective of this study is to help **X Education** optimize its **lead conversion strategy** by developing a **Logistic Regression model** that assigns a **Lead Score (0-100)** to each potential customer. This score will enable the company to:

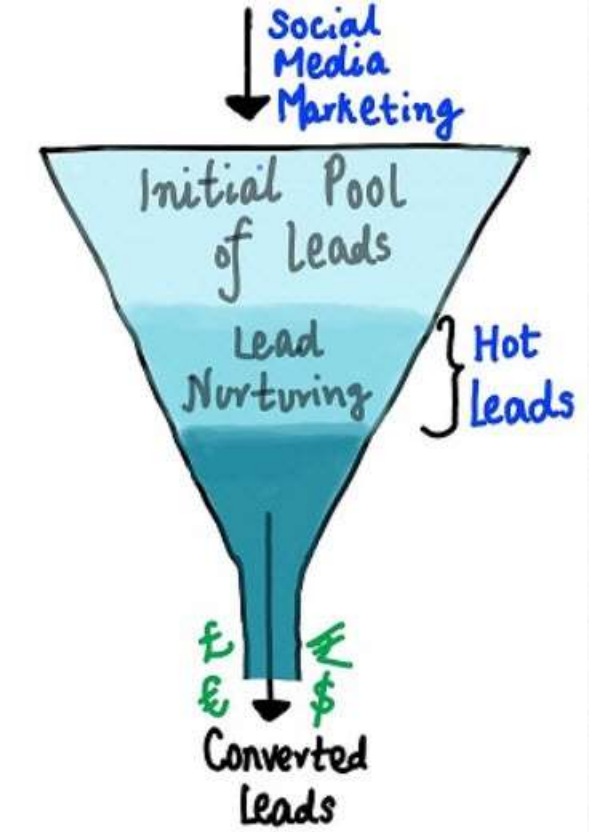
- Prioritize high-potential leads by identifying those most likely to convert.
- Optimize sales efforts by focusing on leads with higher conversion probabilities.
- Adapt to dynamic business requirements by adjusting the model based on different sales strategies, such as:
  - Aggressive lead conversion during peak hiring periods.
  - Minimizing unnecessary calls when sales targets are met.
- Provide data-driven recommendations to enhance decision-making and maximize revenue.

The model's insights will be integrated into the company's sales process, improving efficiency and ensuring targeted customer engagement.

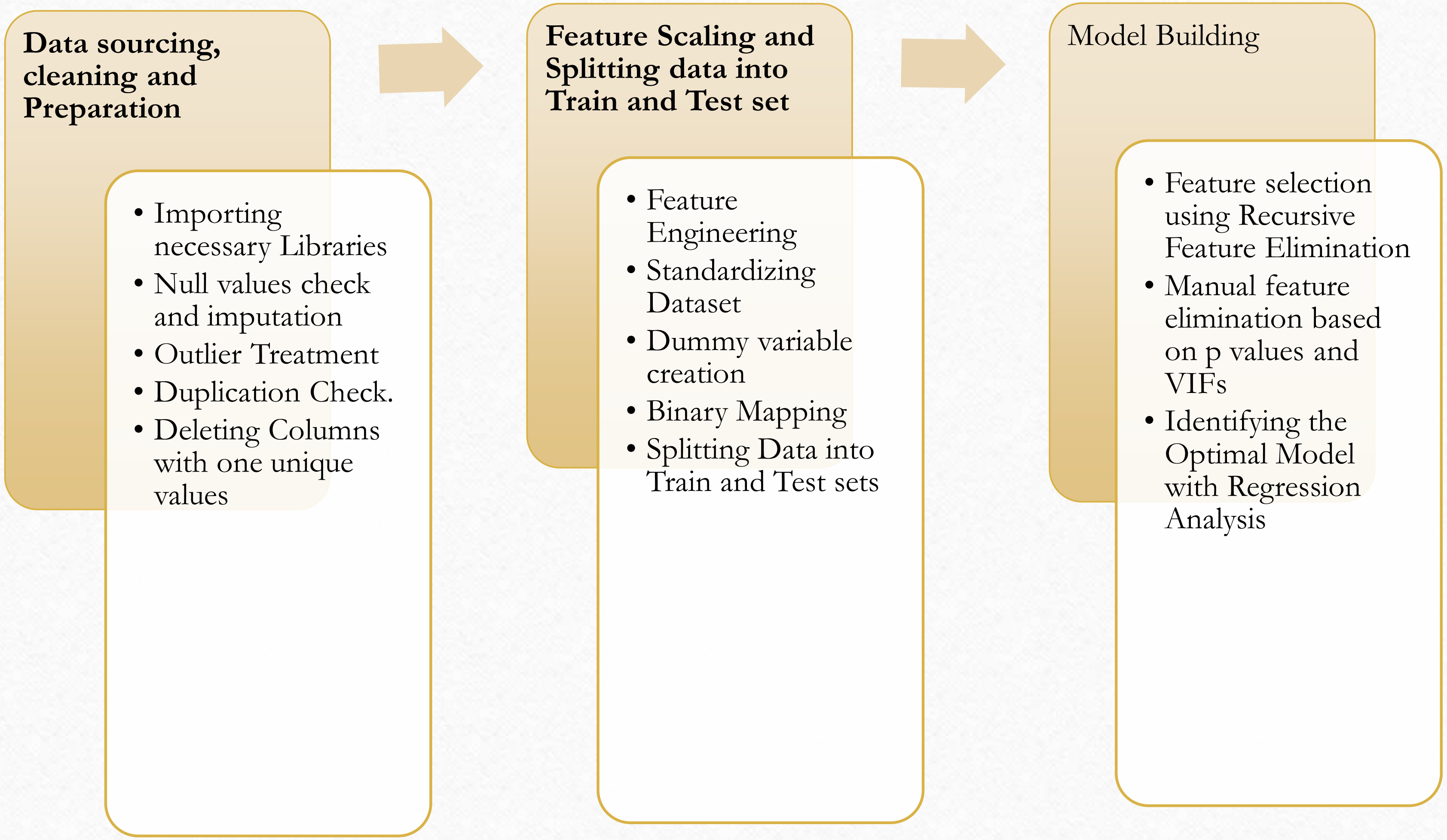
Problem Statement

X Education, an online course provider for industry professionals, receives a high volume of leads daily. However, its **lead conversion rate is significantly low**, with only about **30% of leads converting into customers**.

To improve efficiency, X Education aims to **identify high-potential leads, or ‘Hot Leads’**, who are most likely to enroll in a course. By accurately predicting these leads, the sales team can **prioritize outreach efforts**, focusing on the most promising prospects instead of engaging with all leads indiscriminately. Implementing a **data-driven lead scoring system** will help optimize the conversion process, ultimately **increasing enrollment rates and improving sales efficiency**.



Problem Solving Methodology



Data sourcing, cleaning and Preparation

- Importing necessary Libraries
- Null values check and imputation
- Outlier Treatment
- Duplication Check.
- Deleting Columns with one unique values

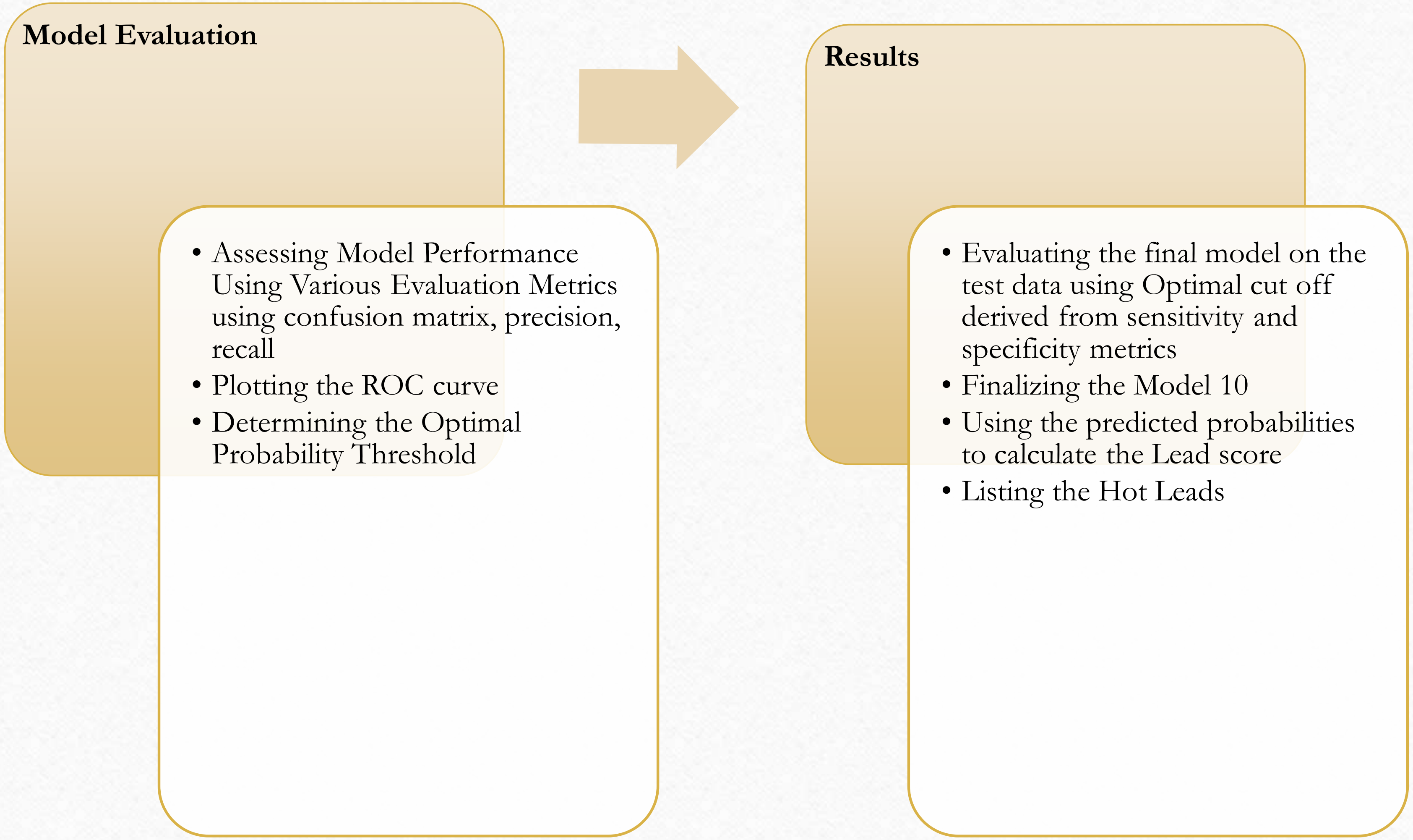
Feature Scaling and Splitting data into Train and Test set

- Feature Engineering
- Standardizing Dataset
- Dummy variable creation
- Binary Mapping
- Splitting Data into Train and Test sets

Model Building

- Feature selection using Recursive Feature Elimination
- Manual feature elimination based on p values and VIFs
- Identifying the Optimal Model with Regression Analysis

Problem Solving Methodology



Model Evaluation

- Assessing Model Performance Using Various Evaluation Metrics using confusion matrix, precision, recall
- Plotting the ROC curve
- Determining the Optimal Probability Threshold

Results

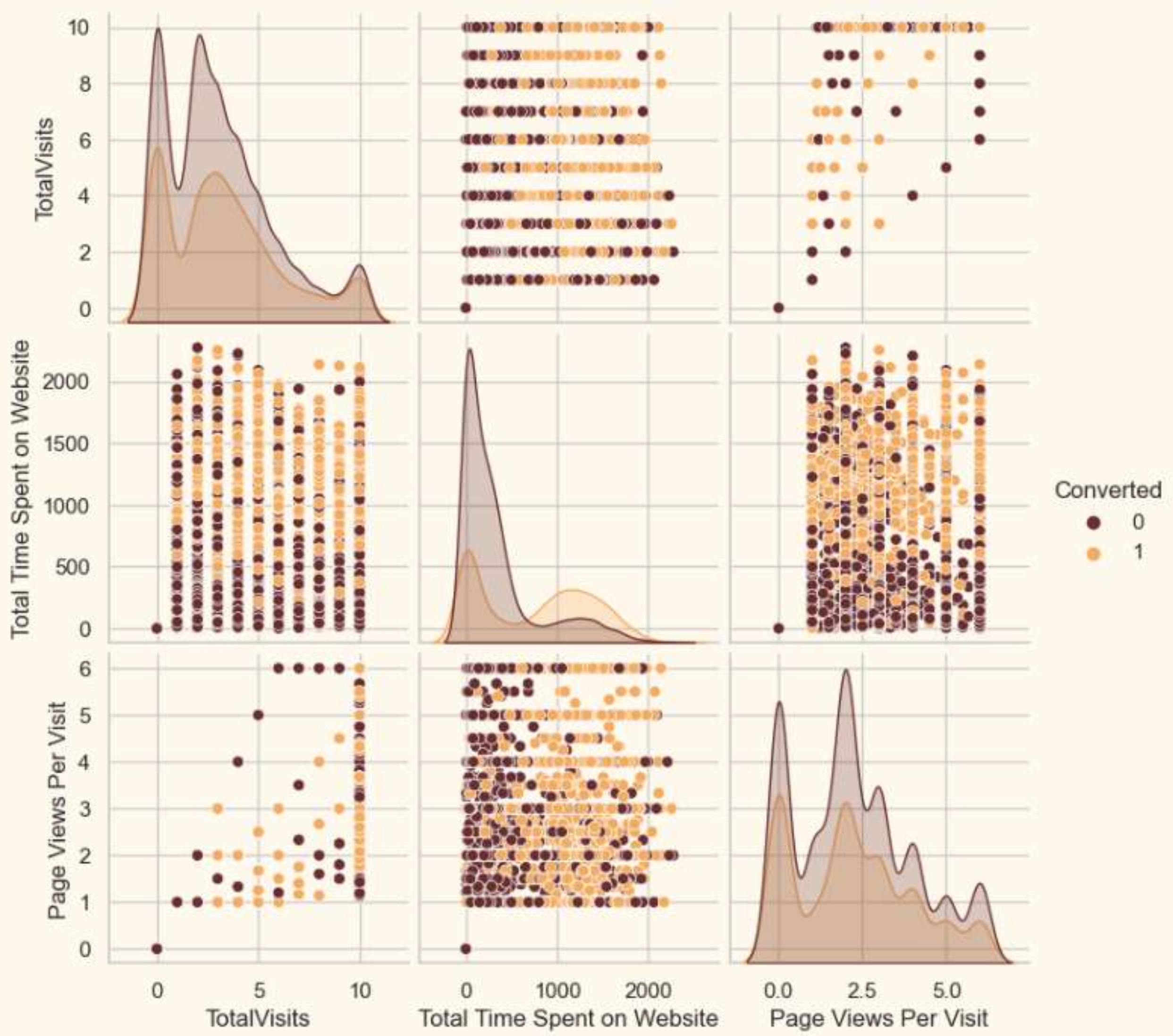
- Evaluating the final model on the test data using Optimal cut off derived from sensitivity and specificity metrics
- Finalizing the Model 10
- Using the predicted probabilities to calculate the Lead score
- Listing the Hot Leads

Data Visualization

Relative plot of Numerical variables

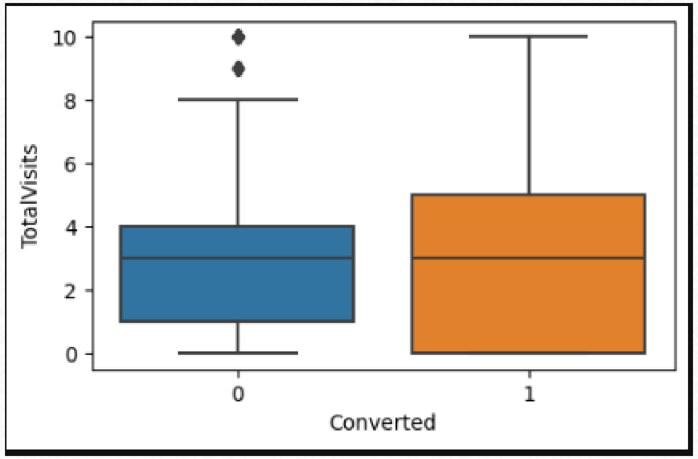
The following inferences can be drawn from the plot:

- Time spent on the website seems to be a stronger indicator of conversion compared to total visits or page views per visit.
- Many users who convert have relatively low visits but spend a good amount of time on the website.
- High visits with low time spent indicate non-engaged users who do not convert.

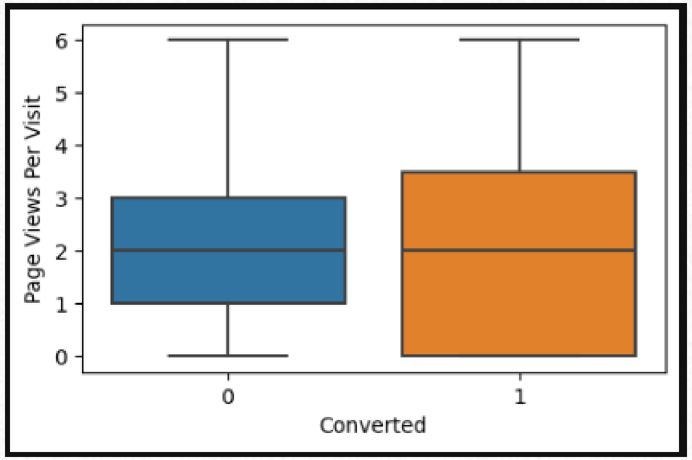


Outlier treatment of Numerical columns

Total Visits vs Converted

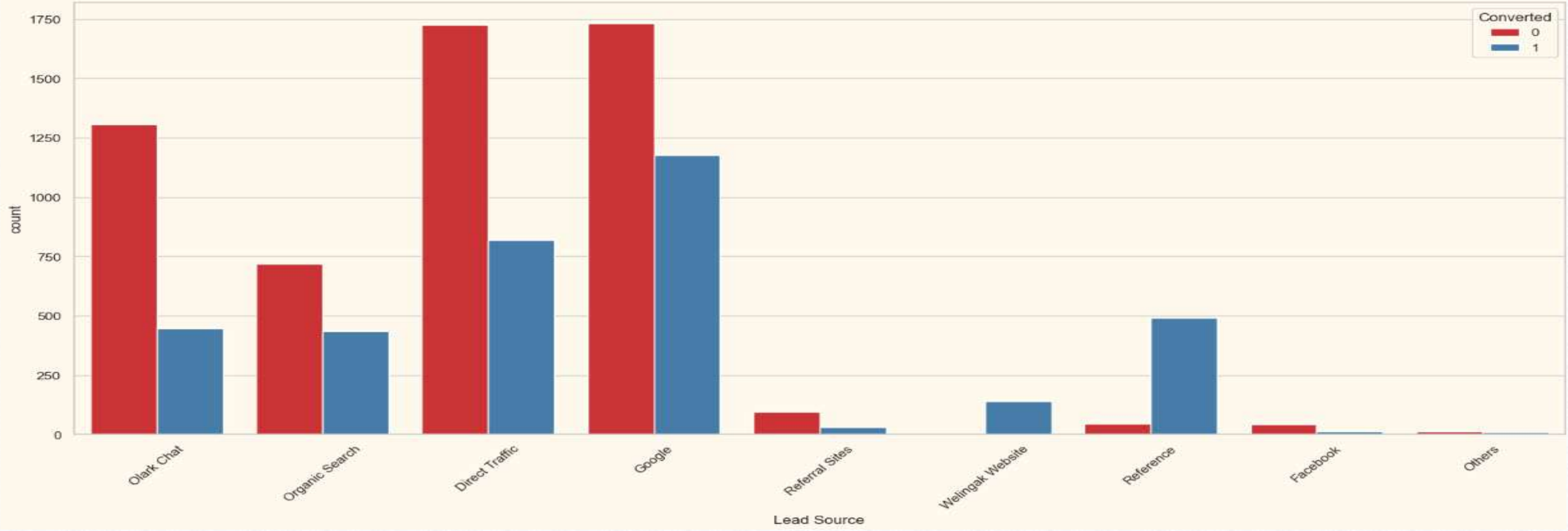


Pages views per visit vs Converted



Since the data for both the above features were highly skewed, we will cap the outliers to 95%. The missing percentage is very low and the data is skewed, using the median imputation is the best approach. It will prevent distortion caused by extreme values while preserving the original distribution.

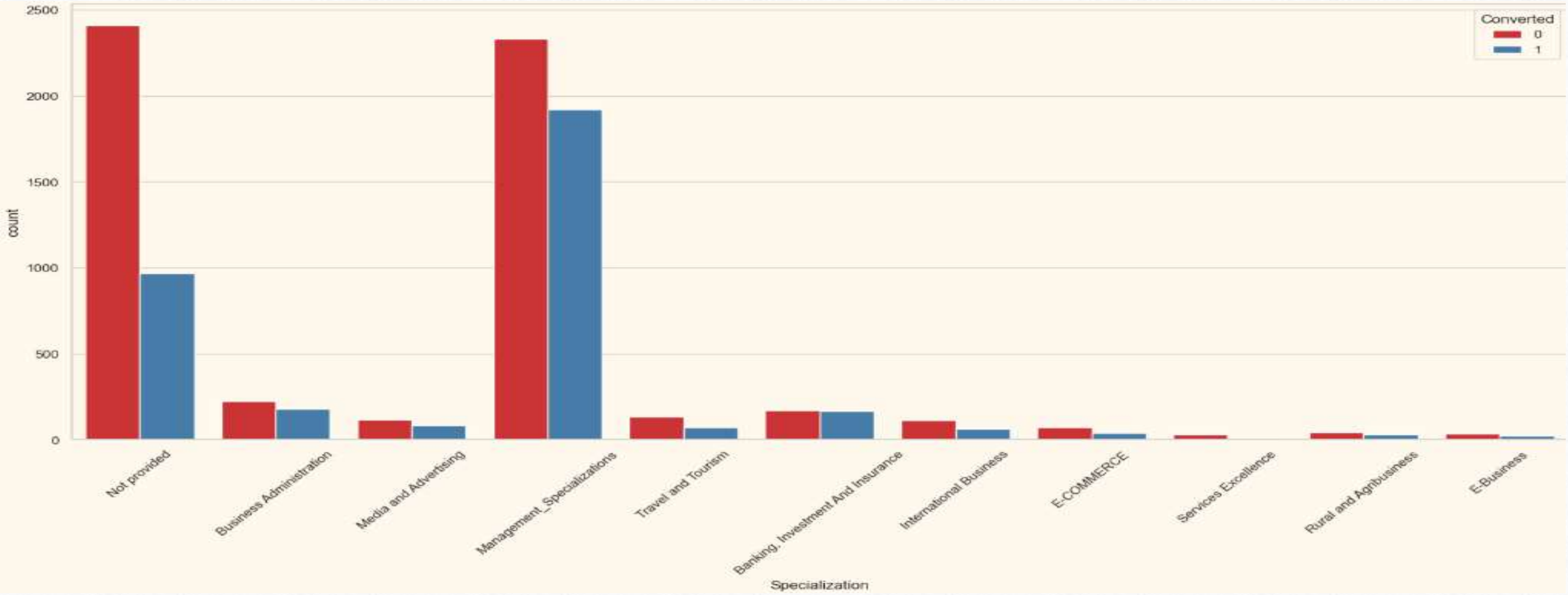
Lead Source vs Converted



Inferences

- Feature Engineering was done for this column, the ones with the very low value counts were grouped together and a new category "Others" were created to reduce dimensionality
- It can be deduced from the plot that Welingak Website and References have very high conversion rate

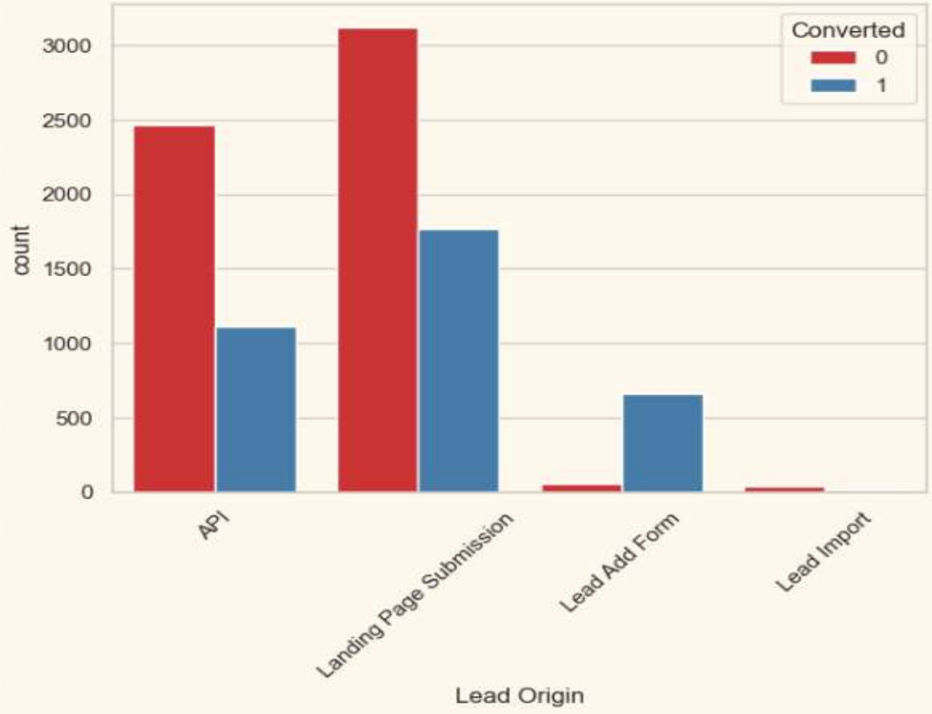
Specialization vs Converted



Inferences:

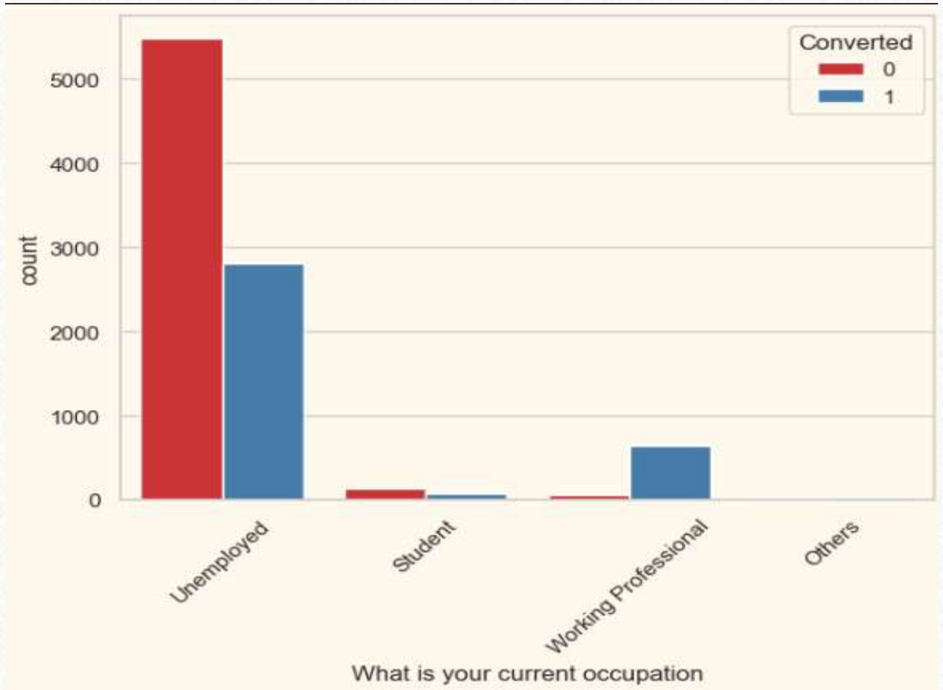
- Feature Engineering was done on this column and all the managements specializations were grouped together
- Banking, investment and Insurance have high conversion rate

Lead Origin Vs Converted



Inferences:
To improve overall lead conversion rate, focus should be on improving lead conversion rate of API and Landing Page Submission.
Also, generate more leads from Lead Add form since they have a very good conversion rate

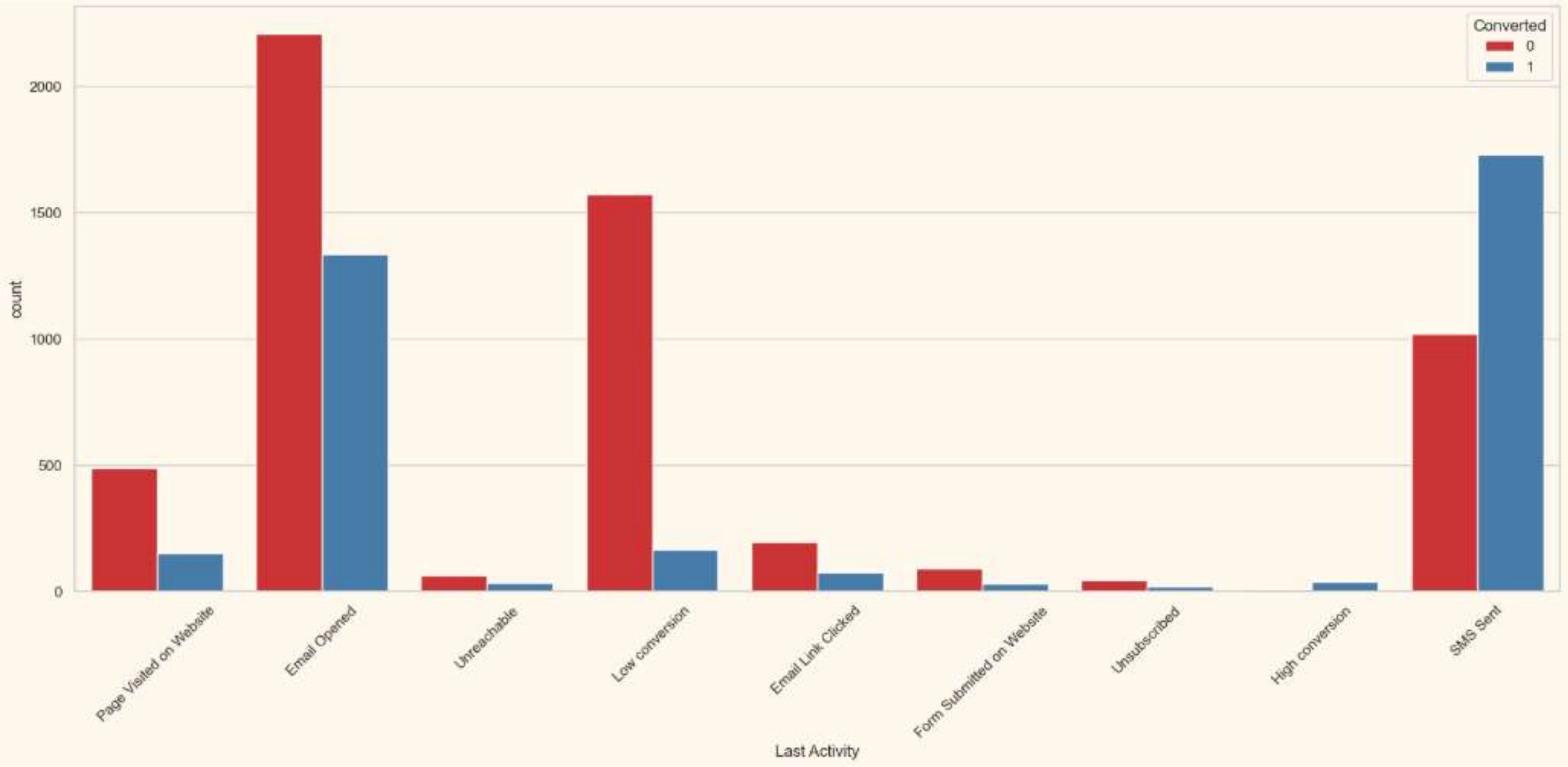
Occupation Vs Converted



Inferences

- Conversion rate for Working professionals is highest 92%.
- To improve overall lead conversion rate, focus should be on improving lead conversion of unemployed .
- Also , generate more leads from Working Professionals.
- "Businessman" ,"Housewives and "Other" were merged to "Others" to reduce dimensionality

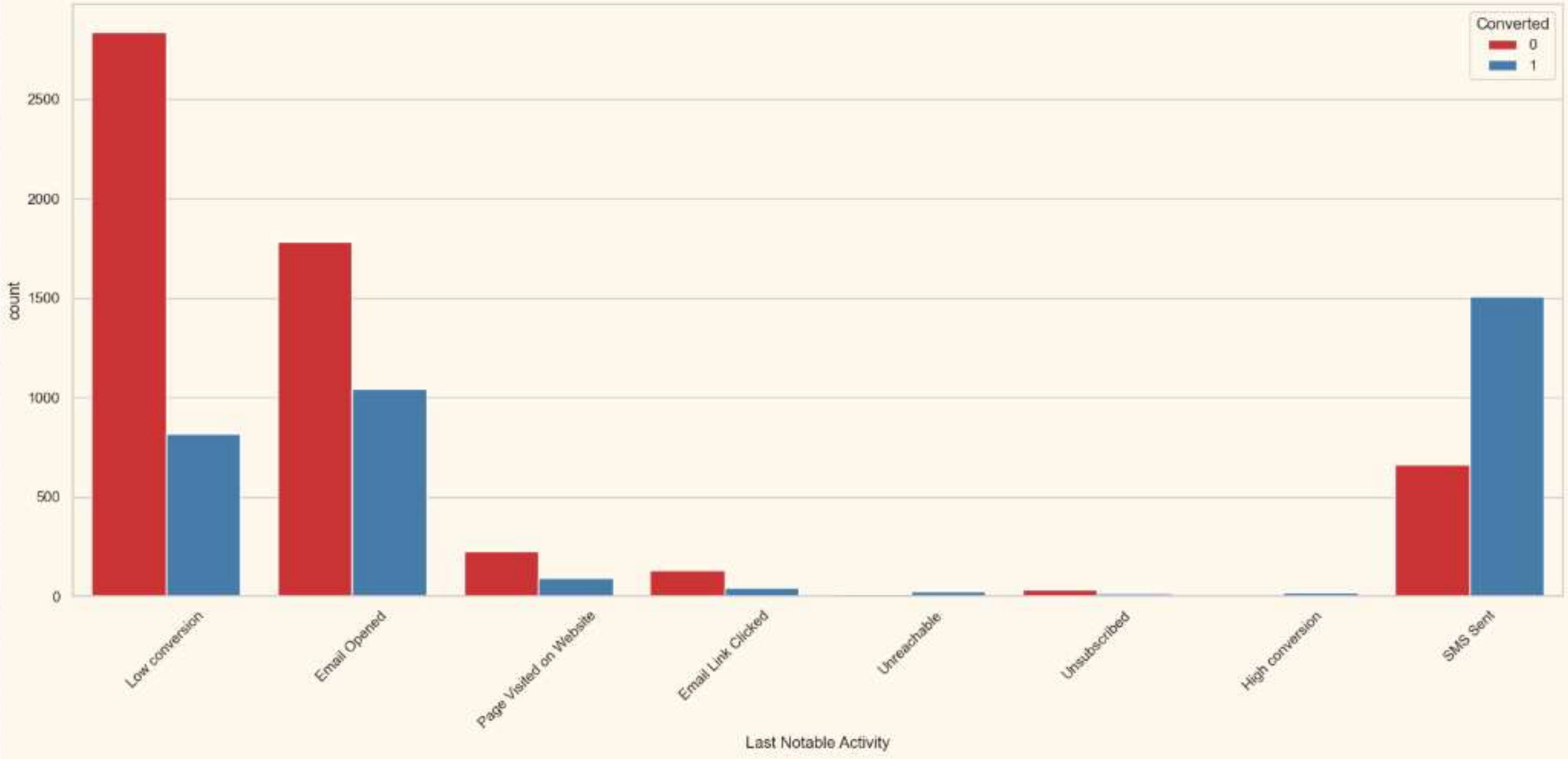
Last Activity Vs Converted



Inferences

- Maximum leads are generated from Email opened and SMS sent, though the conversion rate is not that good as compared to the number of leads generated.
- To improve overall lead conversion rate, focus should be on improving lead conversion of people with low conversion (-olark chat conversation)as the leads are high but conversion are very less.

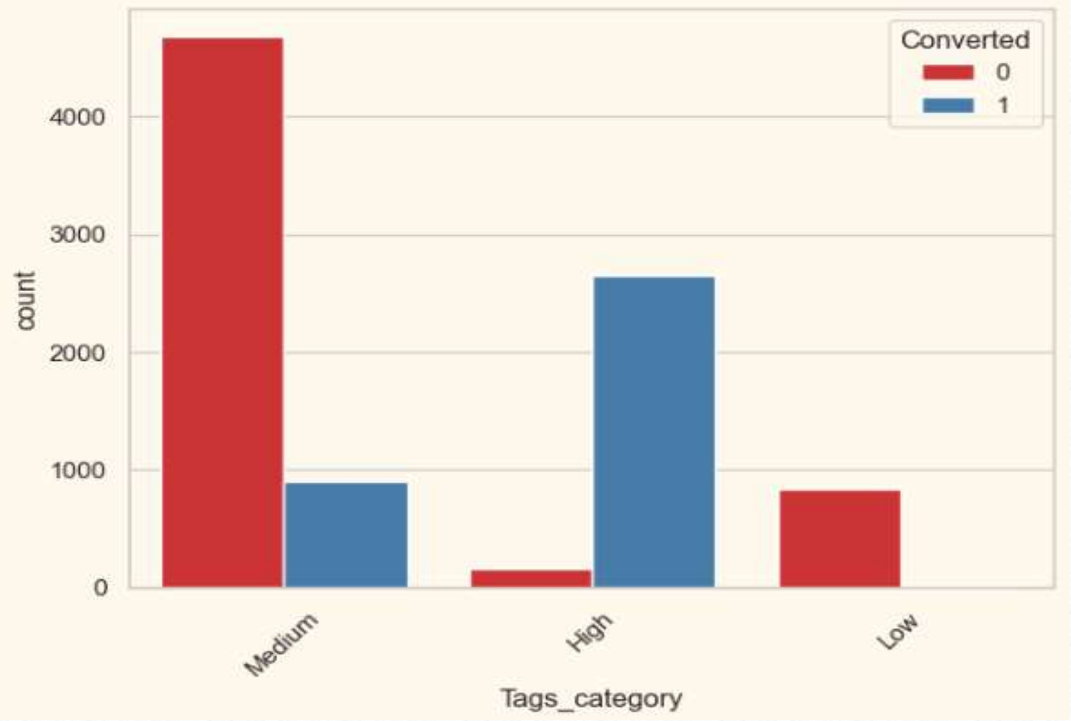
Last Notable Activity Vs Converted



Inferences

- Last Notable Activity is specific to students and good conversion rate can be seen for SMS sent, telephonic conversation
- To improve overall lead conversion rate, focus should be on improving lead conversion of people with low conversion (Olark chat, Modified and email bounced)as the leads are high but conversion are very less.

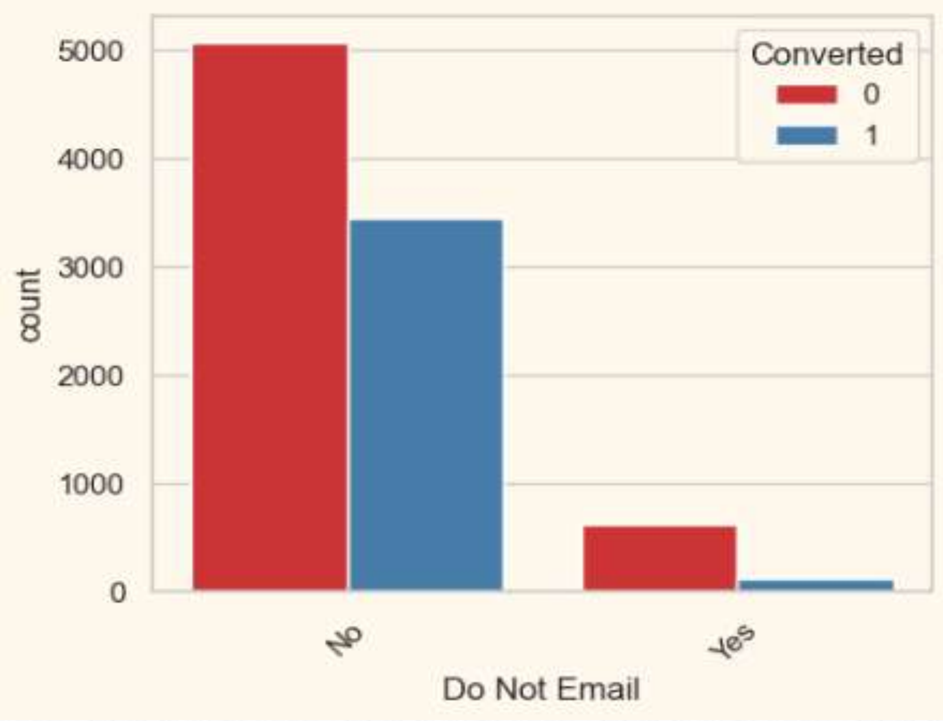
Tag_Category vs Conversion



Inferences:

- Feature Engineering was done on column Tags and a new column Tags_category was created which has only 3 categories based on the conversion rates.
- Business focus should be on High as is has very high conversion rate on the contrary other categories have poor conversion rate

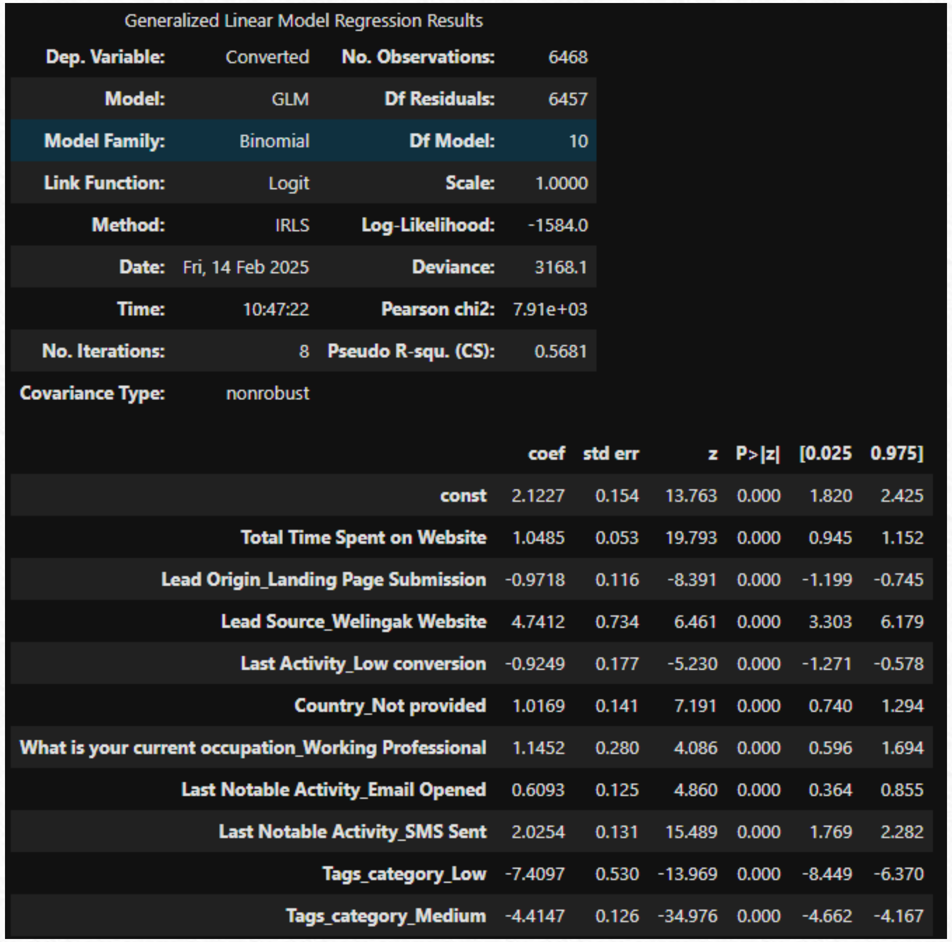
Do Not Email vs Conversion



Inferences:

- It can be inferred that leads don't prefer being contacted via email
- Moreover leads not opting for this service have high conversion rate

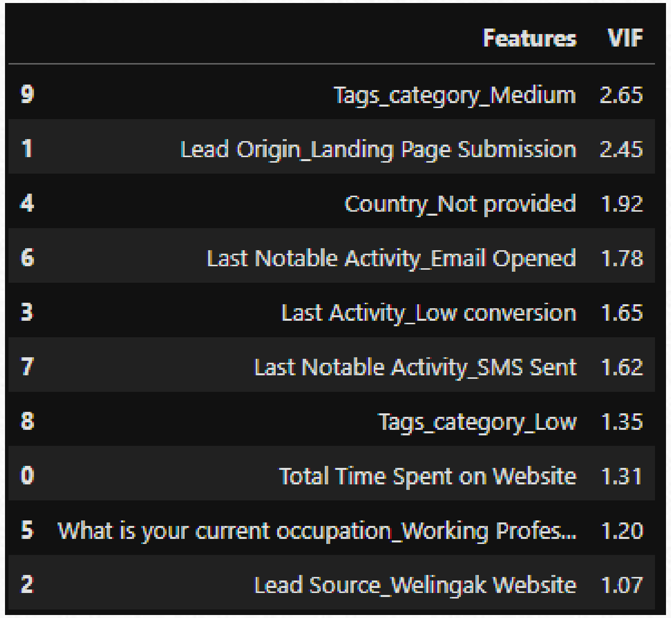
Final Model GLM Regression Results and VIF values



Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6457
Model Family:	Binomial	**Df Model:**	10
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1584.0
Date:	Fri, 14 Feb 2025	**Deviance:**	3168.1
Time:	10:47:22	**Pearson chi2:**	7.91e+03
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.5681
Covariance Type:	nonrobust		

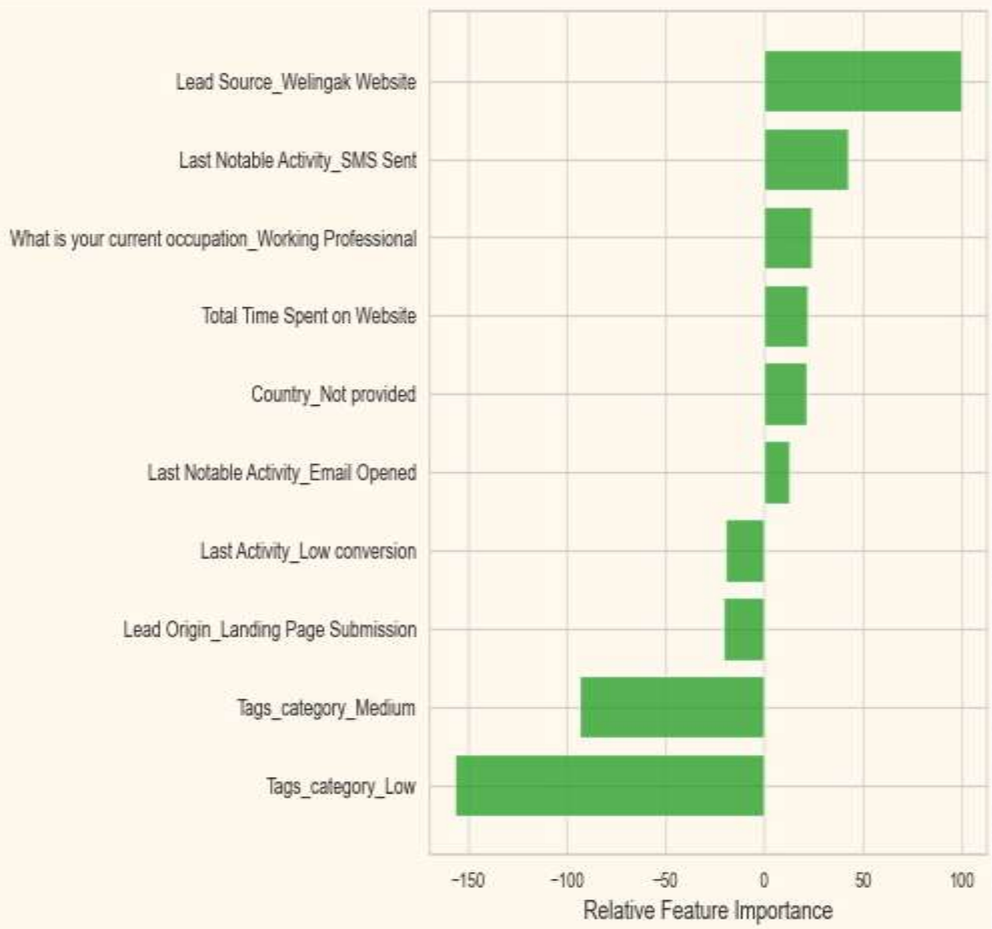
	coef	std err	z	P>|z|	[0.025	0.975]
const	2.1227	0.154	13.763	0.000	1.820	2.425
Total Time Spent on Website	1.0485	0.053	19.793	0.000	0.945	1.152
Lead Origin_Landing Page Submission	-0.9718	0.116	-8.391	0.000	-1.199	-0.745
Lead Source_Welingak Website	4.7412	0.734	6.461	0.000	3.303	6.179
Last Activity_Low conversion	-0.9249	0.177	-5.230	0.000	-1.271	-0.578
Country_Not provided	1.0169	0.141	7.191	0.000	0.740	1.294
What is your current occupation_Working Professional	1.1452	0.280	4.086	0.000	0.596	1.694
Last Notable Activity_Email Opened	0.6093	0.125	4.860	0.000	0.364	0.855
Last Notable Activity_SMS Sent	2.0254	0.131	15.489	0.000	1.769	2.282
Tags_category_Low	-7.4097	0.530	-13.969	0.000	-8.449	-6.370
Tags_category_Medium	-4.4147	0.126	-34.976	0.000	-4.662	-4.167



	Features	VIF
9	Tags_category_Medium	2.65
1	Lead Origin_Landing Page Submission	2.45
4	Country_Not provided	1.92
6	Last Notable Activity_Email Opened	1.78
3	Last Activity_Low conversion	1.65
7	Last Notable Activity_SMS Sent	1.62
8	Tags_category_Low	1.35
0	Total Time Spent on Website	1.31
5	What is your current occupation_Working Profes...	1.20
2	Lead Source_Welingak Website	1.07

Inferences:
All the features listed in the GLM regression model are significant as there p value is very small. VIF values are also below threshold value 5.

Determining Feature Importance



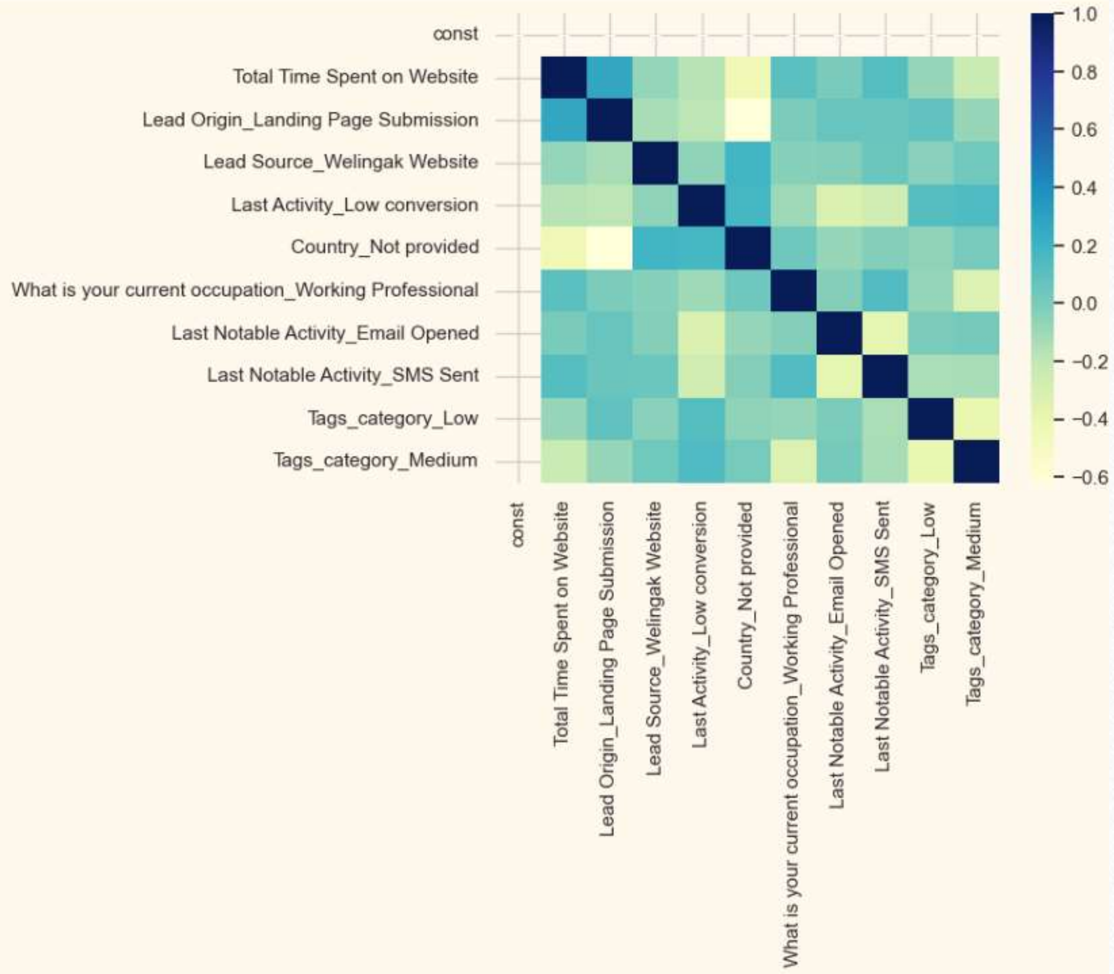
Positive Indicators:

- Lead_Source Welingak Website
- Last Notable Activity_SMS sent
- What is your current occupation_working professional
- Total time spent on website

Negative Indicators:

- Last Activity_Low conversion
- Lead Origin Landing page submission
- Tag category Medium and Low

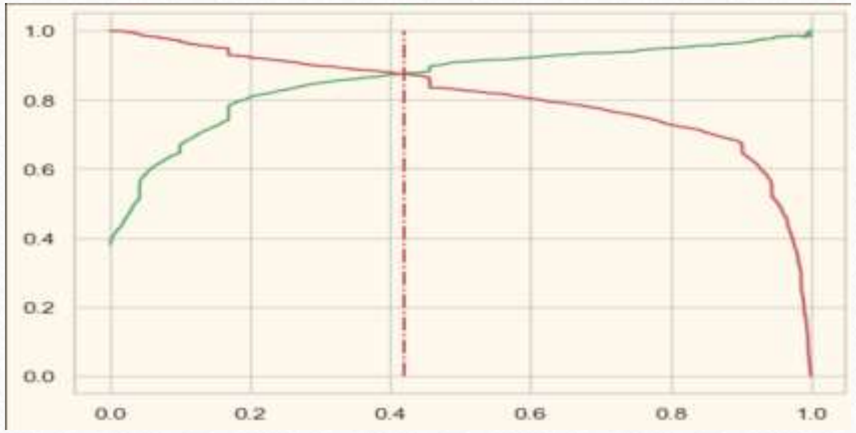
Correlation between independent variables of the Final Model

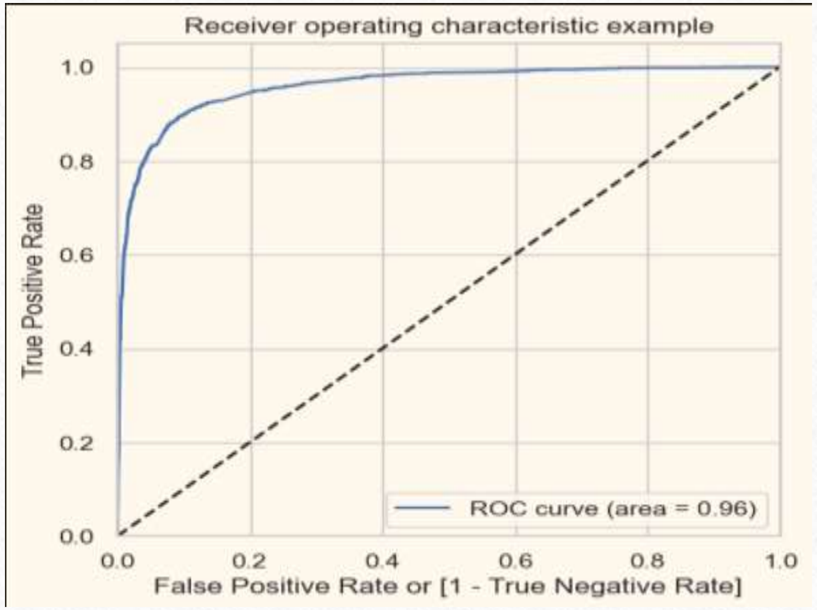


Inferences:

- From the heatmap, there does not appear to be any strong correlation (close to 1 or -1) between the features.
- Most of the correlation values seem to be moderate to weak, indicated by lighter shades of blue and green

Model Evaluation Train dataset





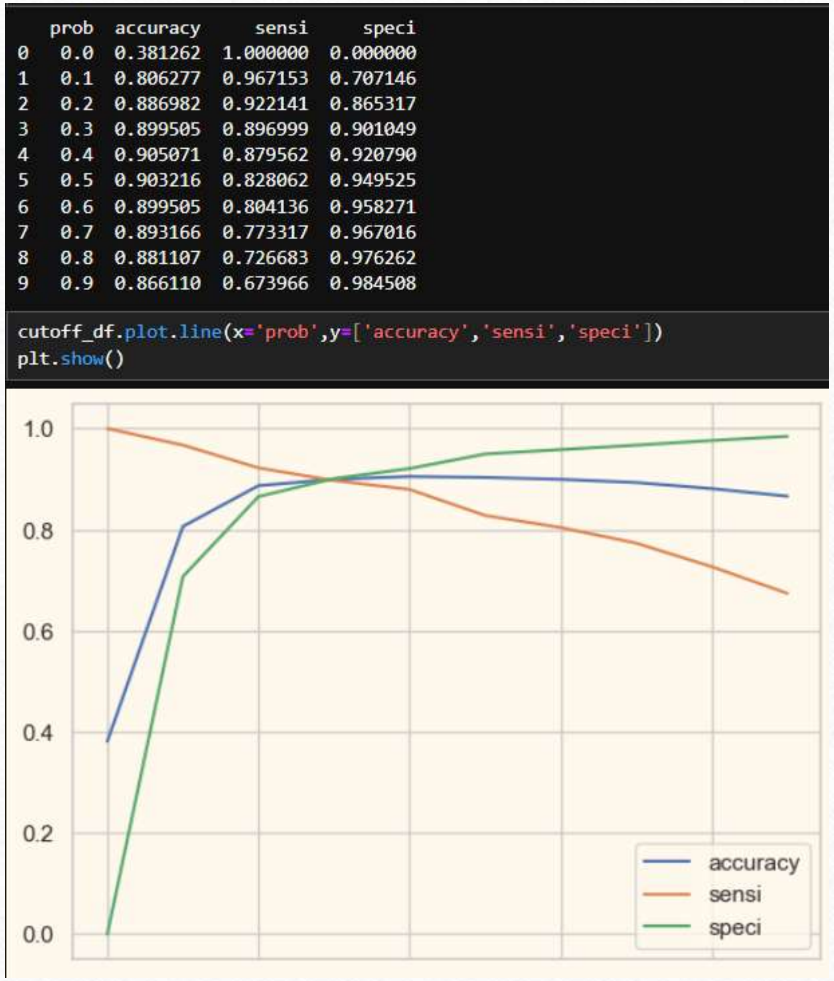
Precision Recall curve

- This plot appears to show Sensitivity (True Positive Rate) and Specificity (True Negative Rate) as a function of the probability threshold.

ROC Curve: Its interpretation evaluates the performance of Classification model

- High AUC (0.96) indicates excellent model in distinguishing between classes
- Curve Shape (Far from the Diagonal Line) suggest suggests the model correctly identifies **most of the positive cases** with minimal false positives

Optimal Threshold



```
   prob  accuracy     sensi     speci
0   0.0  0.381262  1.000000  0.000000
1   0.1  0.806277  0.967153  0.707146
2   0.2  0.886982  0.922141  0.865317
3   0.3  0.899505  0.896999  0.901049
4   0.4  0.905071  0.879562  0.920790
5   0.5  0.903216  0.828062  0.949525
6   0.6  0.899505  0.804136  0.958271
7   0.7  0.893166  0.773317  0.967016
8   0.8  0.881107  0.726683  0.976262
9   0.9  0.866110  0.673966  0.984508
```

```
cutoff_df.plot.line(x='prob',y=['accuracy','sensi','speci'])
plt.show()
```

Inferences:

- The graph depicts optimal cut off of 0.3 based on accuracy, sensitivity and specificity. They all intersect at this point and take a value of around 90%.

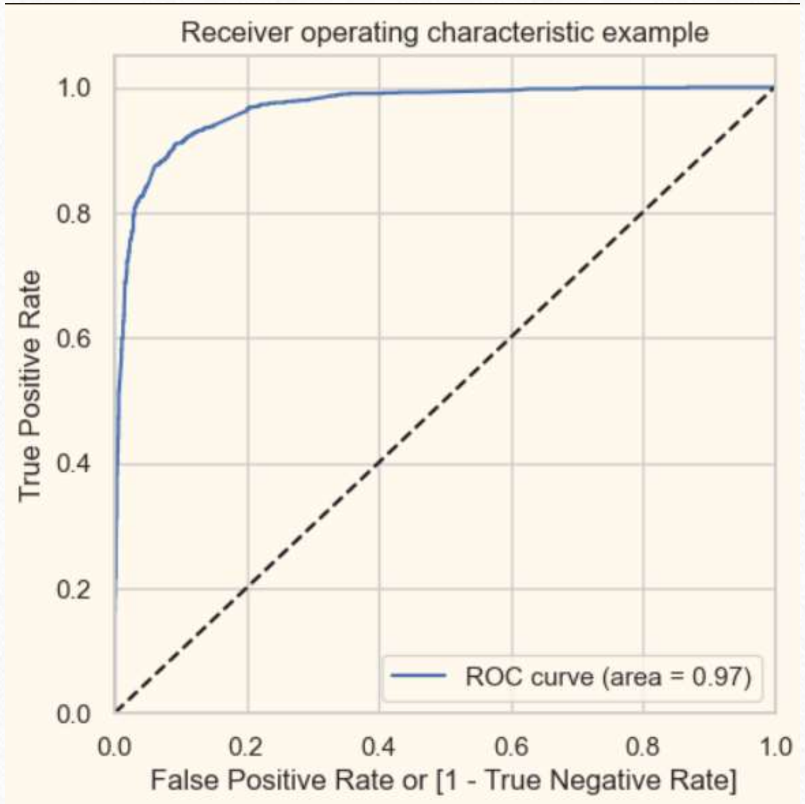
Confusion Matrix

3800	202
424	2042

Train Data:

Accuracy : 90.32%
Sensitivity : 82.80%
Specificity : 94.45 %

Model Evaluation Test dataset



High AUC (0.97) indicates excellent model in distinguishing between classes

Confusion Matrix

1525	152
102	993

Test Data:

Accuracy : 90.83%
Sensitivity : 90.68%
Specificity : 90.36 %

The above value shows that the model is performing well on unseen data .Good Sensitivity indicates that Model is also doing well in identifying more converted leads.

Recommendations

- Education Company "X Education" should focus on Lead Source_Welingak Website as the customers visiting this website are more likely to convert into paying customers.
- Leads whose Last Activity was "SMS Sent" are hot leads and should be focused
- Working Professional are more likely to be potent customers as there conversion rates are also high and they want to upgrade themselves for better job prospects
- The more time the customer spent on the website the better are there conversion prospects
- Customers who have not provided there countries have high conversion rate and should be focused
- Leads whose "Last Notable Activity_Email Opened" should be focused as they can prove to be potent conversion
- Tag_Category Medium and Low have very poor conversion rates as compared to Tags_Category High
- Last Activity_Low conversion mainly contains ('Converted to Lead','Olark Chat Conversation','Email Bounced') they have very less conversion rates
- Lead Origin_Landing Page Submission has very poor conversion rate, so better to focus on other Lead origin ways.

Apart this, referring to the Data Visualization, it is imperative that company should focus on:

- Increasing the conversion rate of the Features generating more leads and
- Generating more leads for the features displaying high conversion rates.

Other business objectives:

Since X Education wants to maximize lead conversion and make phone calls to as many potential leads as possible they should take the following steps:

- Prioritize recall (sensitivity) over precision. This ensures they capture most of the potential leads (customers predicted as 1) and avoid missing out on possible conversions. By revising the threshold to 0.3 our recall Improved from 82.83% to 90.78% (More potential conversions are identified)
- Specificity slightly decreased from 94.95% to 90.94% (More false positives). But still it's high so model is still rejecting irrelevant customers well.
- Instead of calling all predicted leads at once, sort them by conversion probability and call high-confidence leads first. This will optimize intern's time
- Implement multi touch follow up strategy: Use a CRM system to track interactions and automate follow-ups.
- Optimize call strategy by dividing the interns into shifts to maximize the coverage over different time slots(morning, afternoon, evening)

Strategy post meeting targets:

Since the company has already met its targets for the quarter and now wants to **avoid unnecessary phone calls,**

- The focus should be on **precision (positive predictive value, PPV)** rather than recall.
- This could be achieved by raising the threshold to (0.7 or 0.8) will classify fewer customers as leads, but those who remain will have a **higher probability of conversion**.