LEAD SCORING CASE STUDY

PROBLEM STATEMENT

- X Education faces a low lead conversion rate of around 30% despite a high number of leads.

- Objective: Increase efficiency by identifying 'Hot Leads' for a higher conversion rate.

- Tasked with building a model to assign lead scores, prioritizing leads with a higher likelihood of conversion.

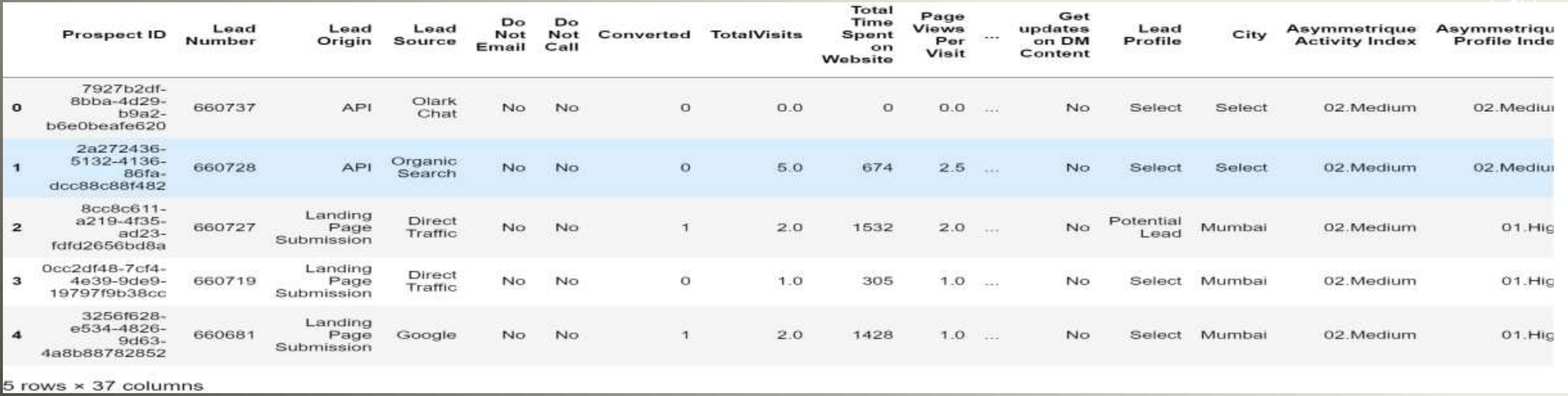
- CEO's target: Achieve an 80% lead conversion rate for improved sales efficiency.

BUSINESS OBJECTIVE

- Outline the approach briefly:
- Data Overview: 9000 data points, key attributes, 'Converted' as the target variable.
- Logistic Regression Model: Assign lead scores between 0 and 100.
- Results: Conversion predictions, evaluation metrics (accuracy, precision, recall, F1-score).
- Conclude with key recommendations for X Education based on the model's insights.
- Optionally, include a visual representation of the lead conversion process funnel.

DATA SET

- 9000 data points with various attributes: Lead Source, Total Time Spent, Total Visits, Last Activity, etc.
- Target variable: 'Converted' (1 for converted, 0 for not converted).
- Check categorical variables for levels, especially 'Select' (considered as null value).
- Refer to the data dictionary in the provided zip folder for detailed dataset insights.



	Prospect ID	Lead Number	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	...	Get updates on DM Content	Lead Profile	City	Asymmetrique Activity Index	Asymmetriqu Profile Inde
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	API	Olark Chat	No	No	0	0.0	0	0.0	...	No	Select	Select	02.Medium	02.Mediu
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	No	No	0	5.0	674	2.5	...	No	Select	Select	02.Medium	02.Mediu
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	...	No	Potential Lead	Mumbai	02.Medium	01.Hig
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	660719	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0	...	No	Select	Mumbai	02.Medium	01.Hig
4	3256f628-e534-4826-9d63-4a8b88782852	660681	Landing Page Submission	Google	No	No	1	2.0	1428	1.0	...	No	Select	Mumbai	02.Medium	01.Hig

5 rows × 37 columns

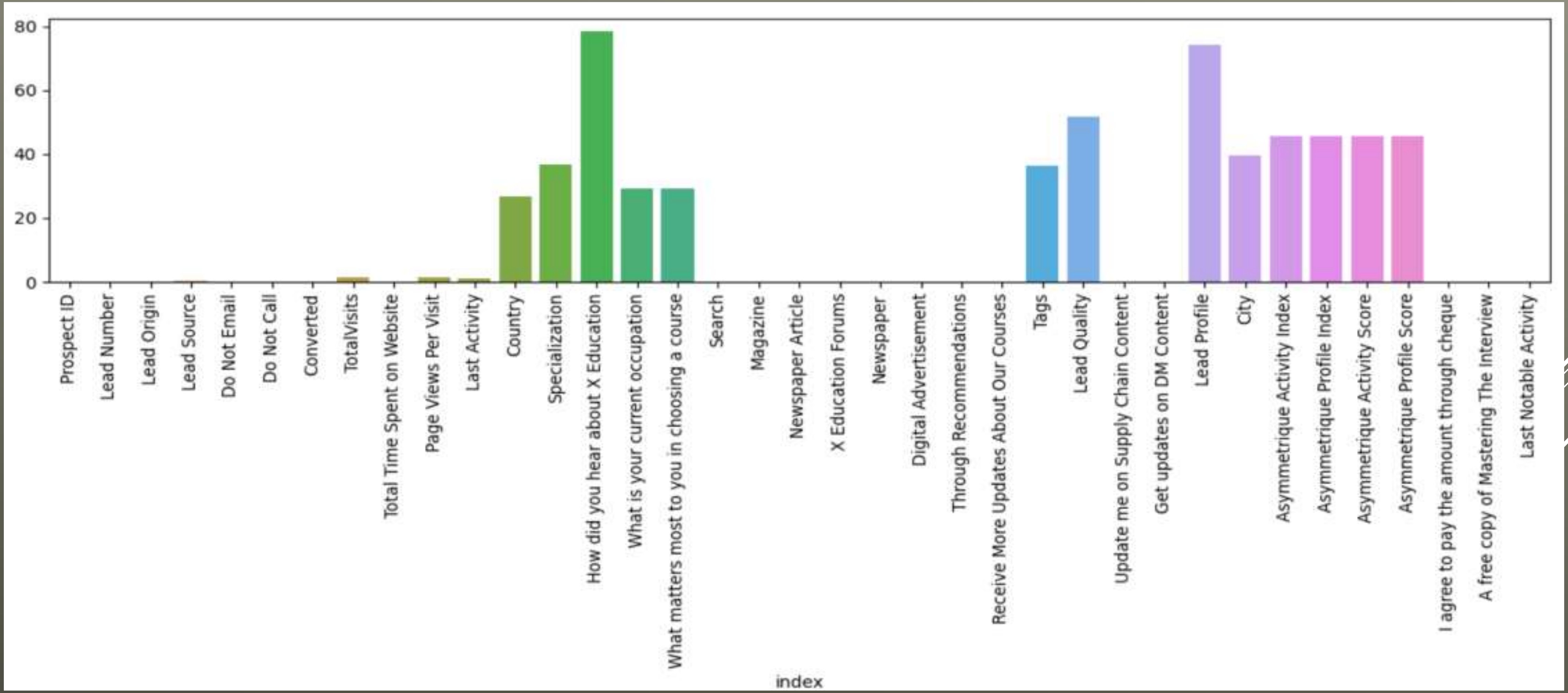
Approach & Methodology:

- Checking the missing values
- Handling outliers.
- Differentiates numerical columns and categorical columns.
- Univariate and Bivariate analysis.
- Correlations.
- Data Preparations
- Train Test Split
- Feature Scaling
- Model Building
- Checking Variance Inflation Factor (V.I.F)
- Confusion Matrix
- Plotting ROC Curve
- Finding optimal cut-off point
- Accuracy, Sensitivity, Specificity
- Precision And Recall

ASSUMPTIONS:

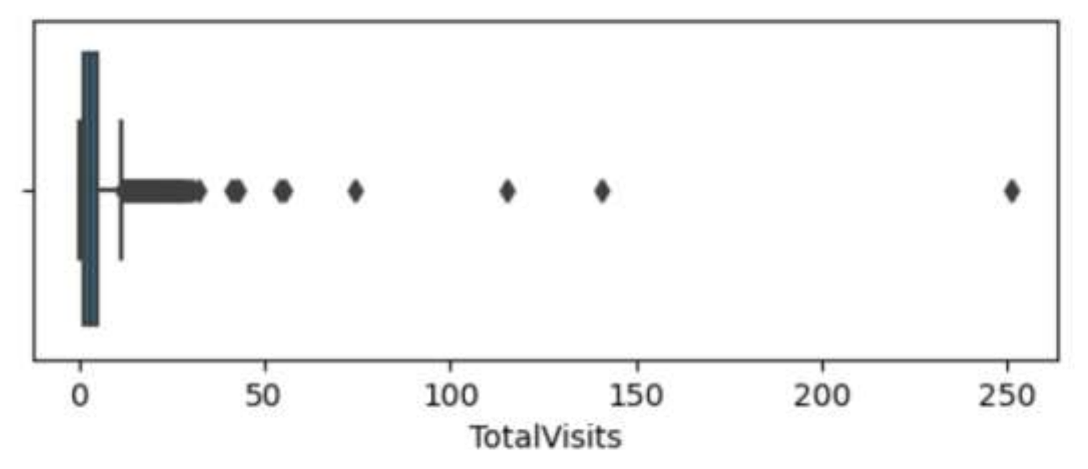
- In the data there are so many values as Select, that means visitor dosen't choose any thing so we replace Select as NAN.
- Dropping the columns 'How did you hear about X Education,Lead Profile', they have missing value more than 70%.
- There is variation in data in "Asymmetrique Activity Index", "Asymmetrique Profile Score", "Asymmetrique Activity Score", "Asymmetrique Profile Index". these four columns, and we were looking at the data in order to impute the NULL values (which are 45%). So we cant make a conclusive decision on this so we drop these columns.
- There are so many categorical columns having null values so we are replacing them with the mode.
- There are outliers in the numerical columns so we handle the outliers by capping them.

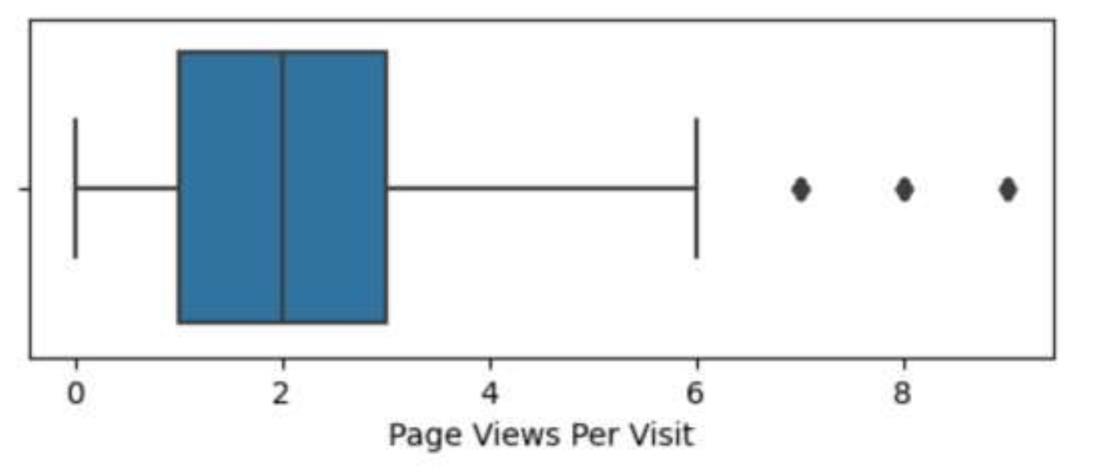
MISSING VALUES:

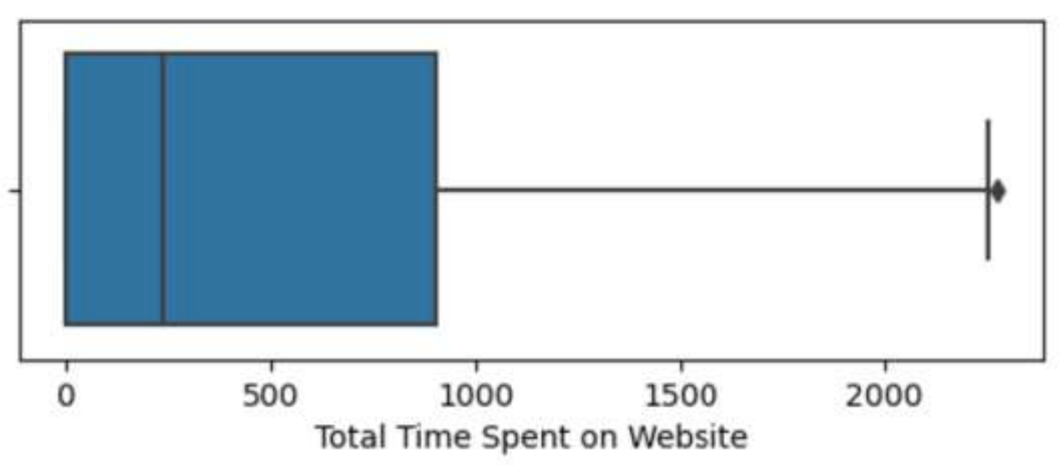


Outliers In Data Set:

There are some insights of outliers in the numerical columns.



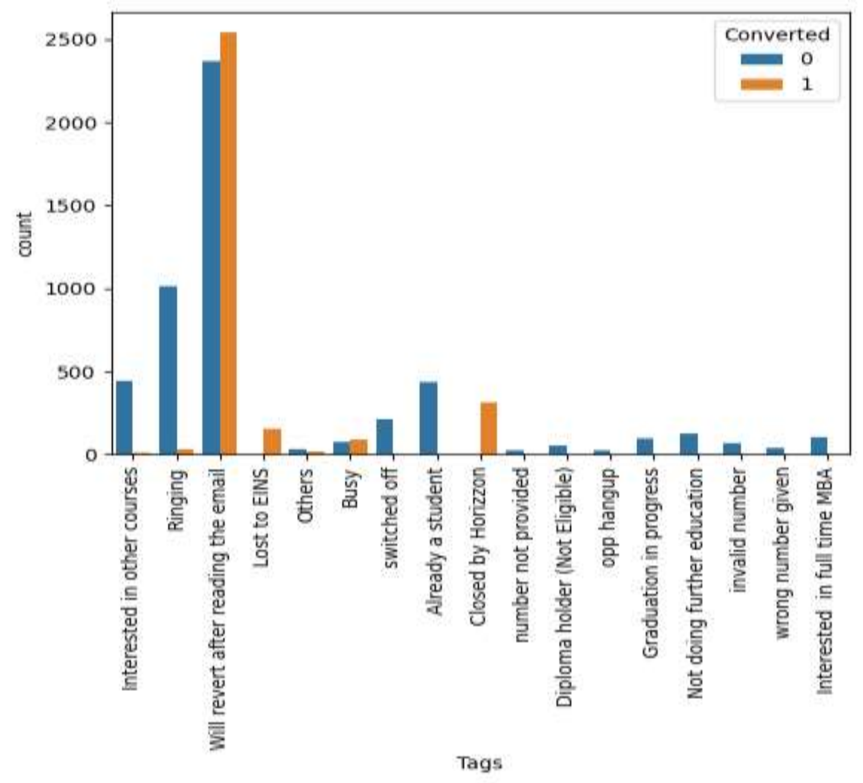


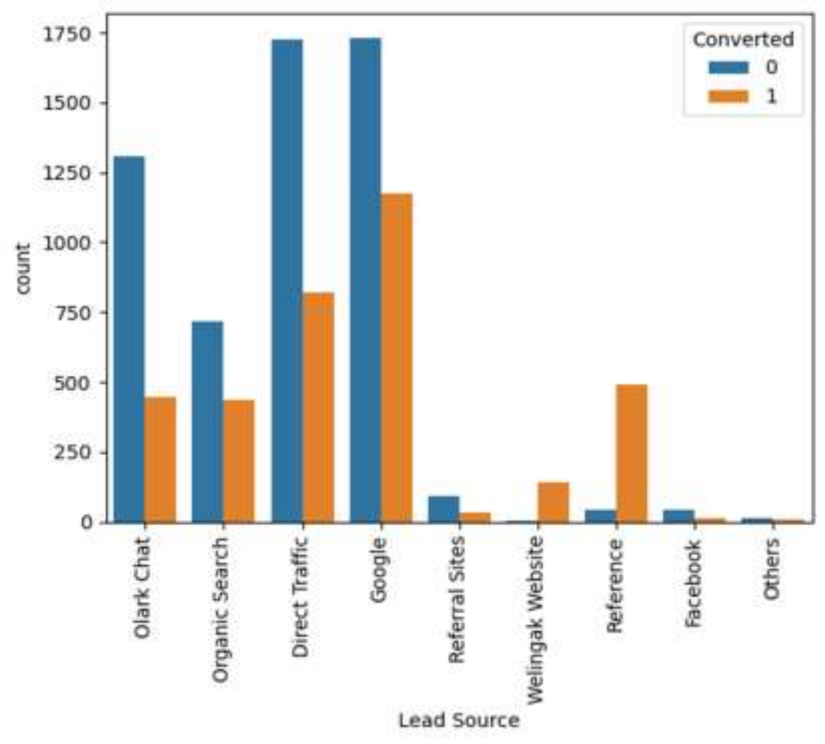


Univariate and Bivariate analysis

Some Univariate and Bivariate analysis was performed on the columns here some insights:

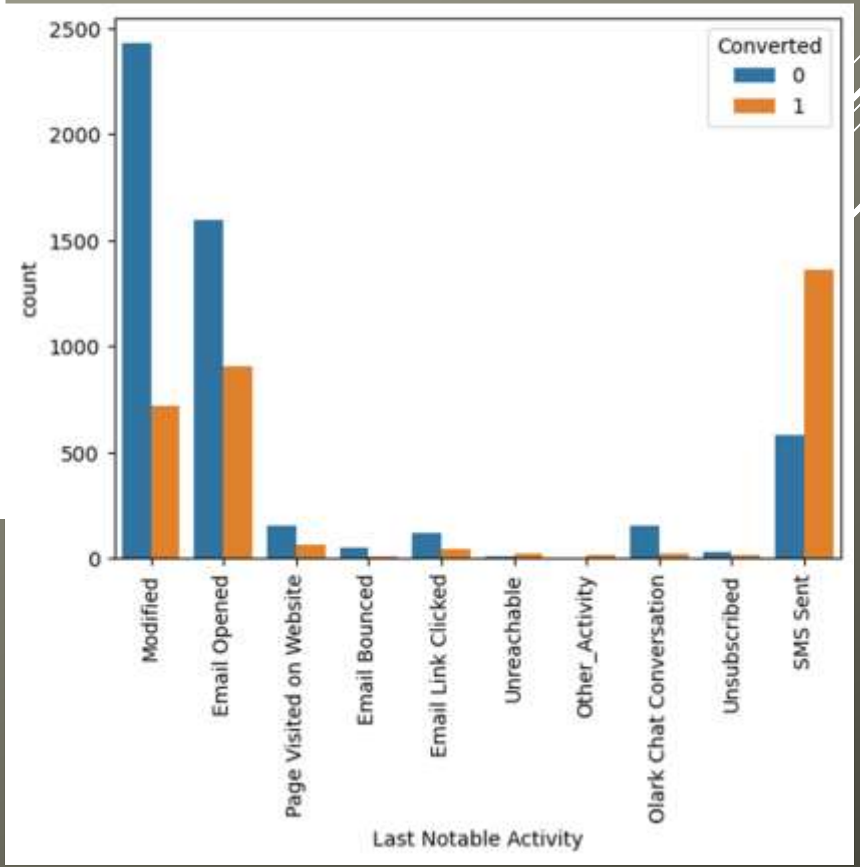
Tags

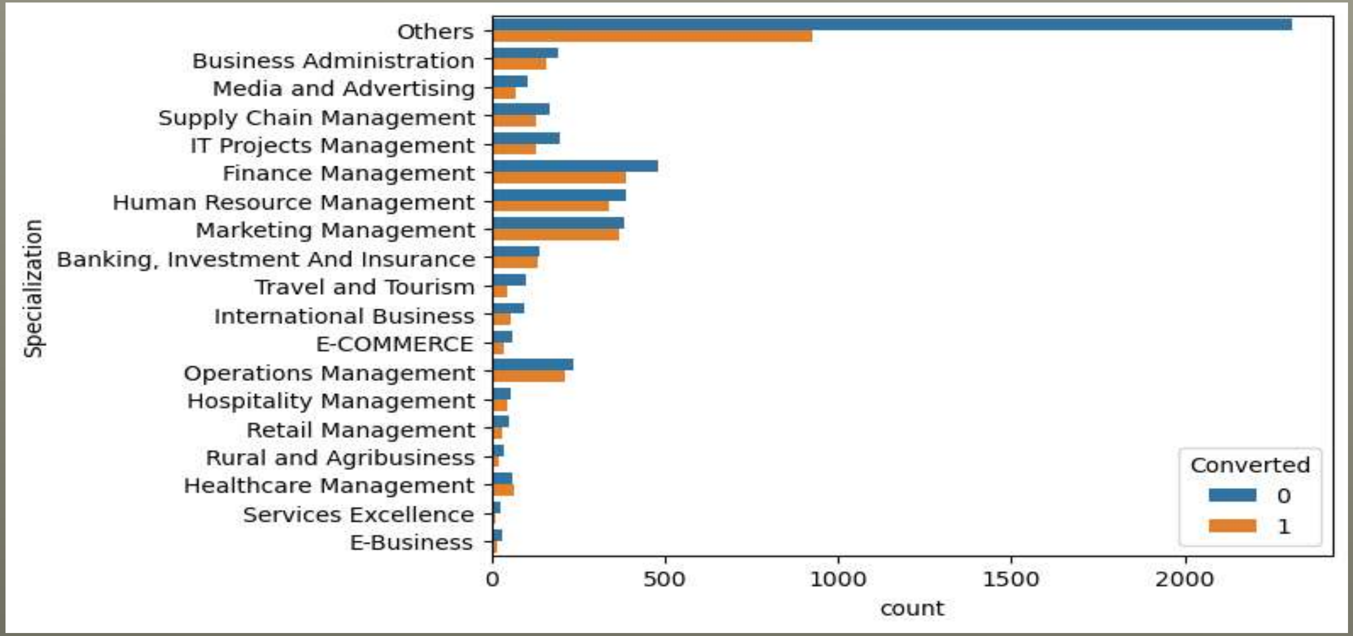




Lead Source

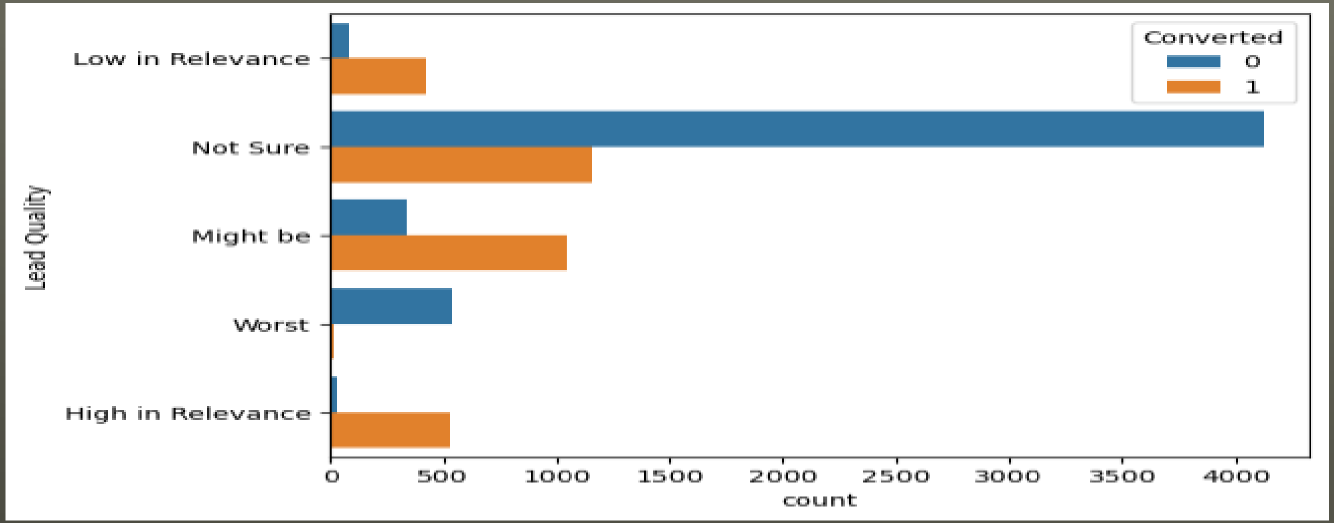
Lead Notable Activity

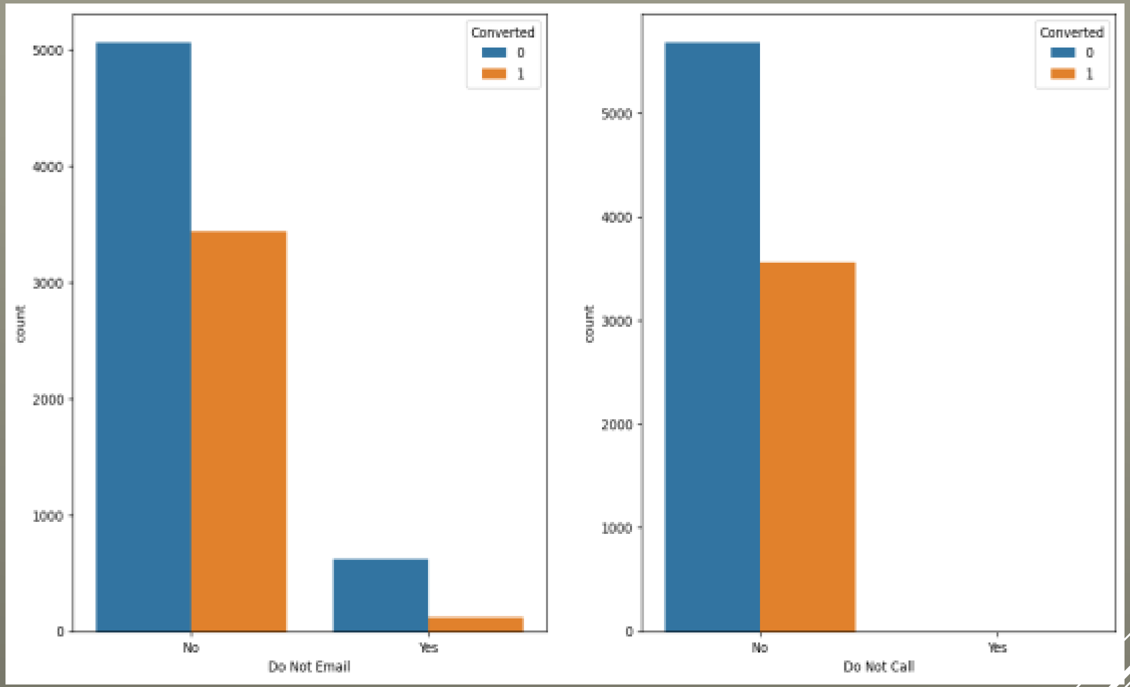




Specialization

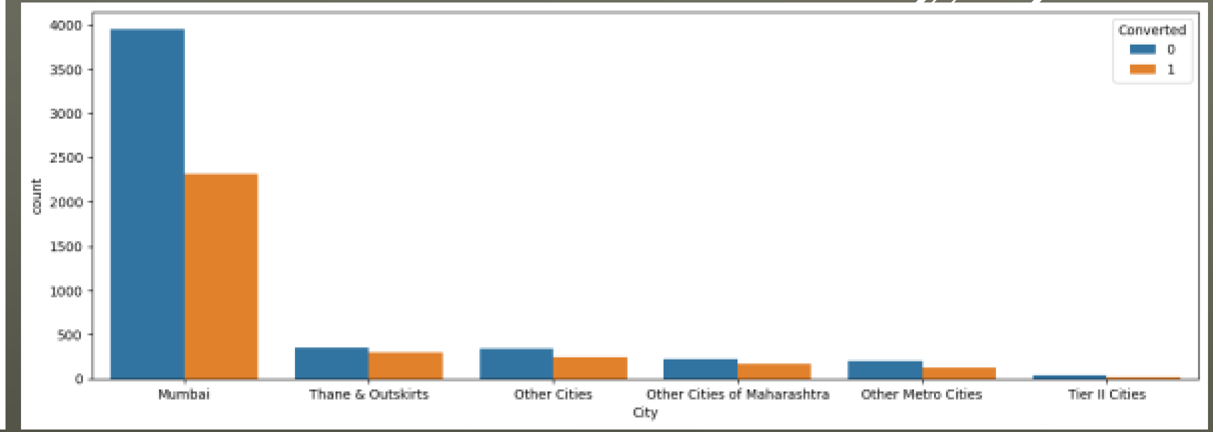
Lead Quality

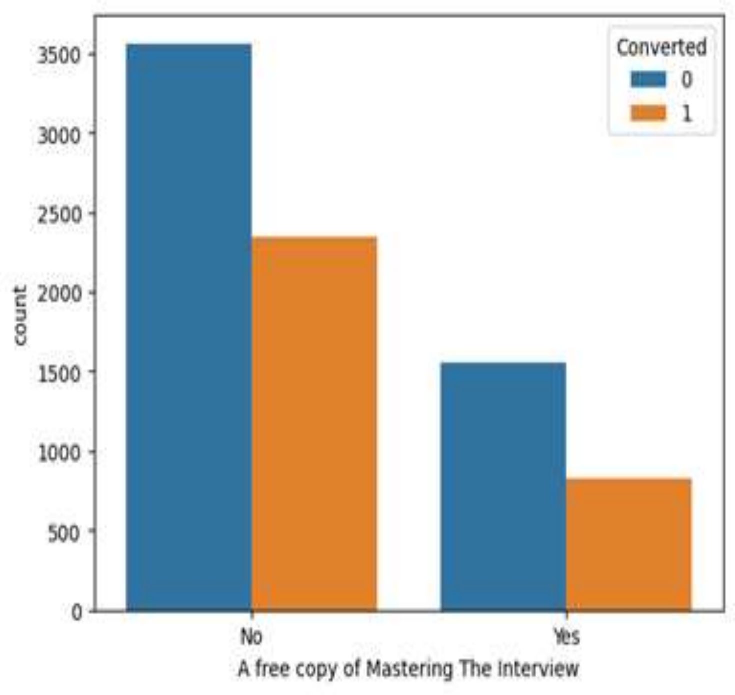




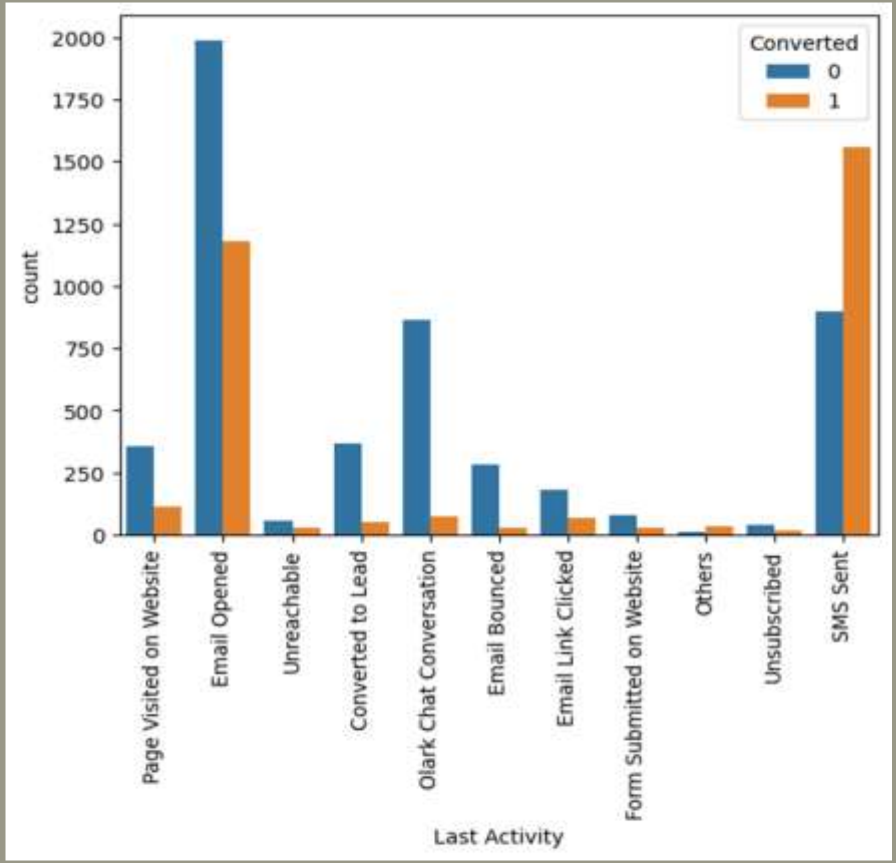
Do Not Email And Do Not Call

City



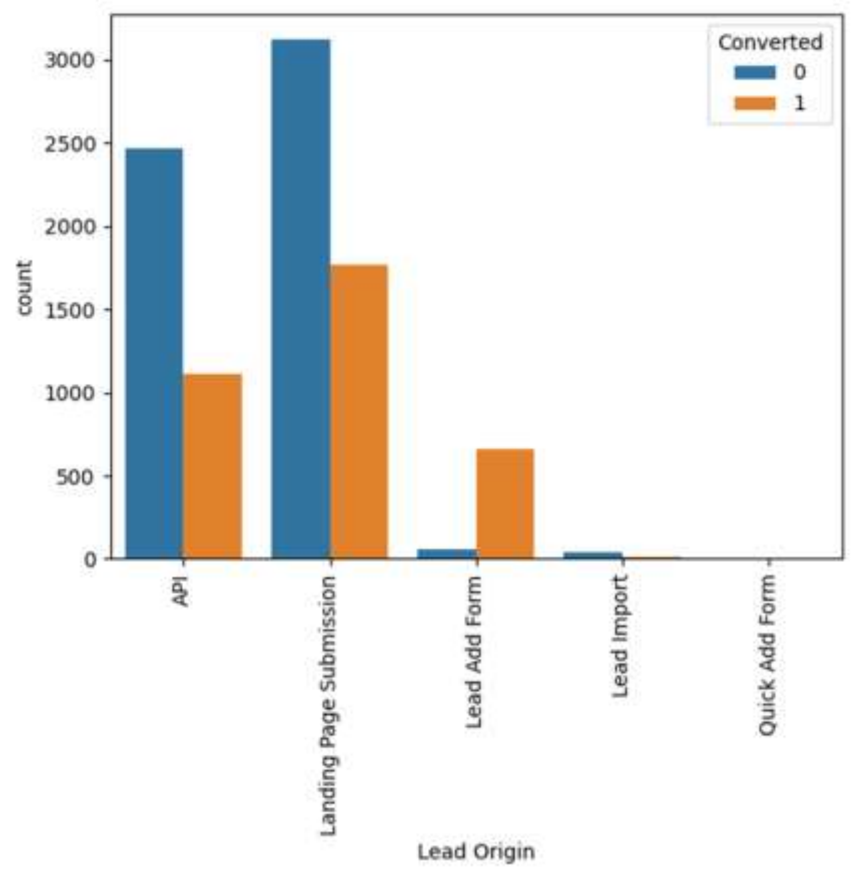


A Free Copy Of Mastering The Interview



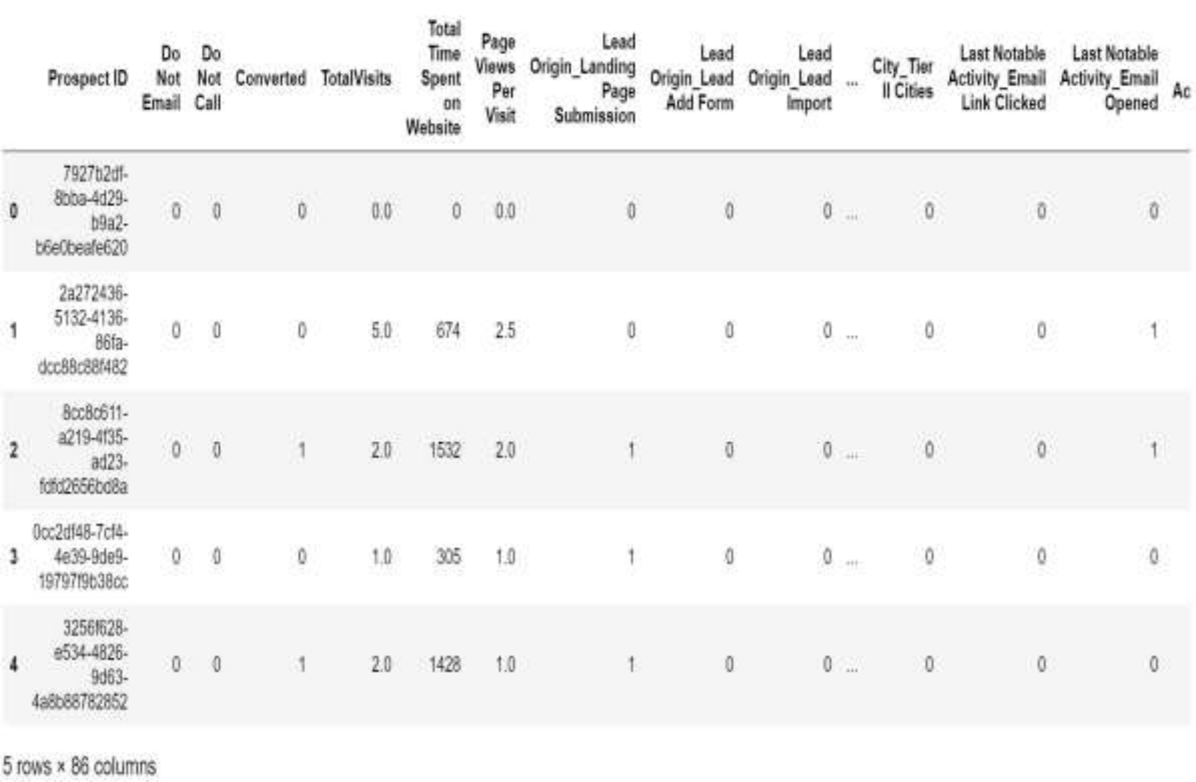
Last Activity

Lead Origin



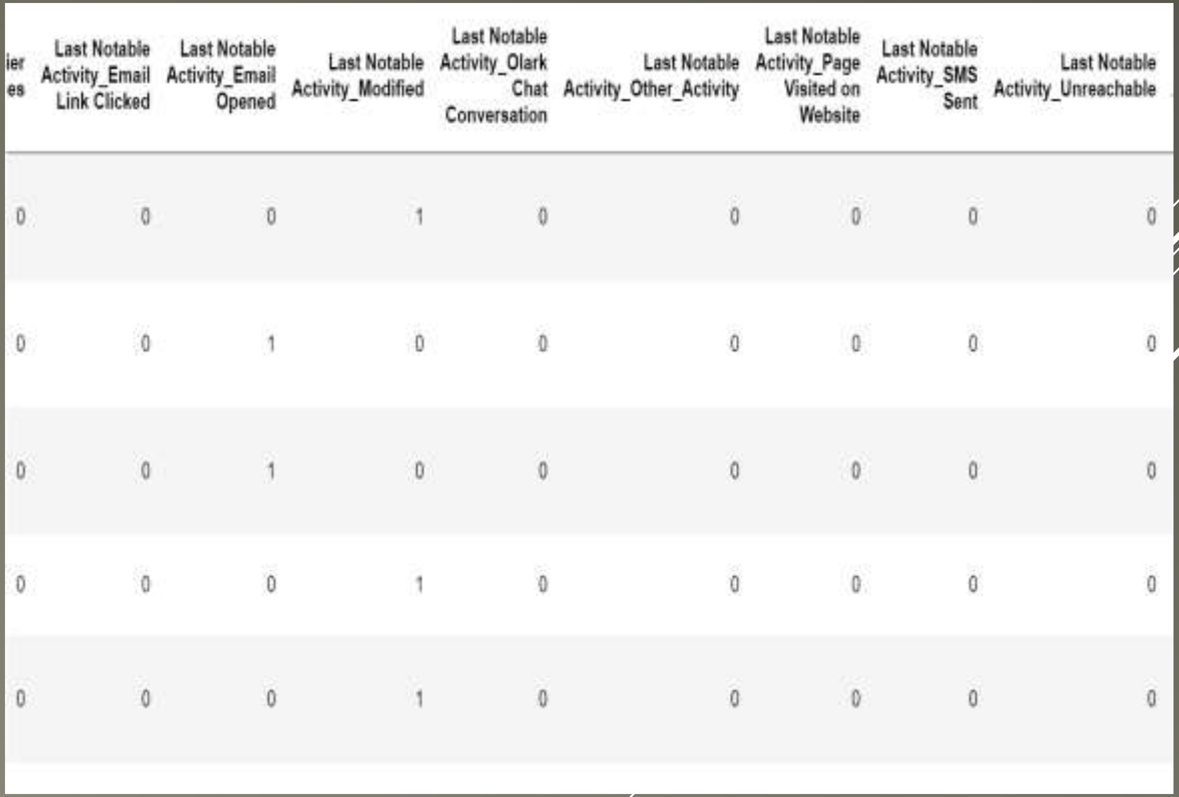
Data Preparation:

After binary mapping and dummification we got the data set ready for model building, here is some insight of data set :



	Prospect ID	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	...	City_Tier II Cities	Last Notable Activity_Email Link Clicked	Last Notable Activity_Email Opened	Ac
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	0	0	0	0.0	0	0.0	0	0	0	...	0	0	0	
1	2a272436-5132-4136-86fa-dcc88c88f482	0	0	0	5.0	674	2.5	0	0	0	...	0	0	1	
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	0	0	1	2.0	1532	2.0	1	0	0	...	0	0	1	
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	0	0	0	1.0	305	1.0	1	0	0	...	0	0	0	
4	3256f628-e534-4826-9d63-4a8b88782852	0	0	1	2.0	1428	1.0	1	0	0	...	0	0	0	

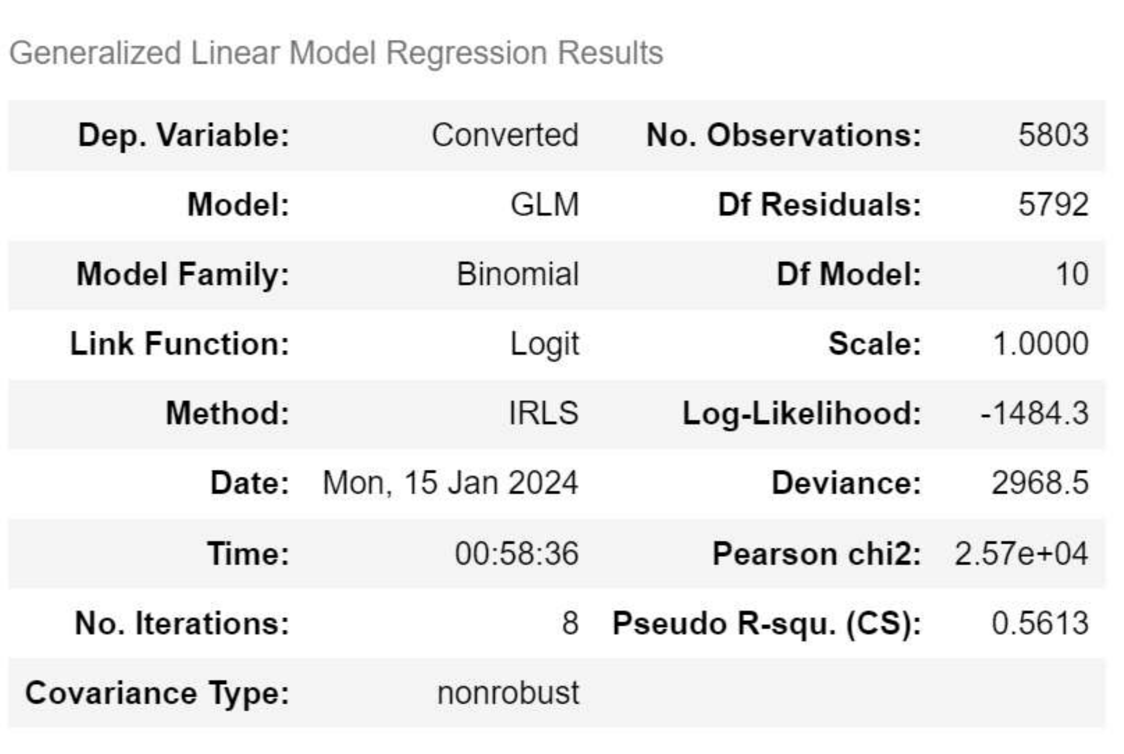
5 rows × 86 columns



ier es	Last Notable Activity_Email Link Clicked	Last Notable Activity_Email Opened	Last Notable Activity_Modified	Last Notable Activity_Olark Chat Conversation	Last Notable Activity_Other_Activity	Last Notable Activity_Page Visited on Website	Last Notable Activity_SMS Sent	Last Notable Activity_Unreachable
0	0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0
0	0	0	1	0	0	0	0	0

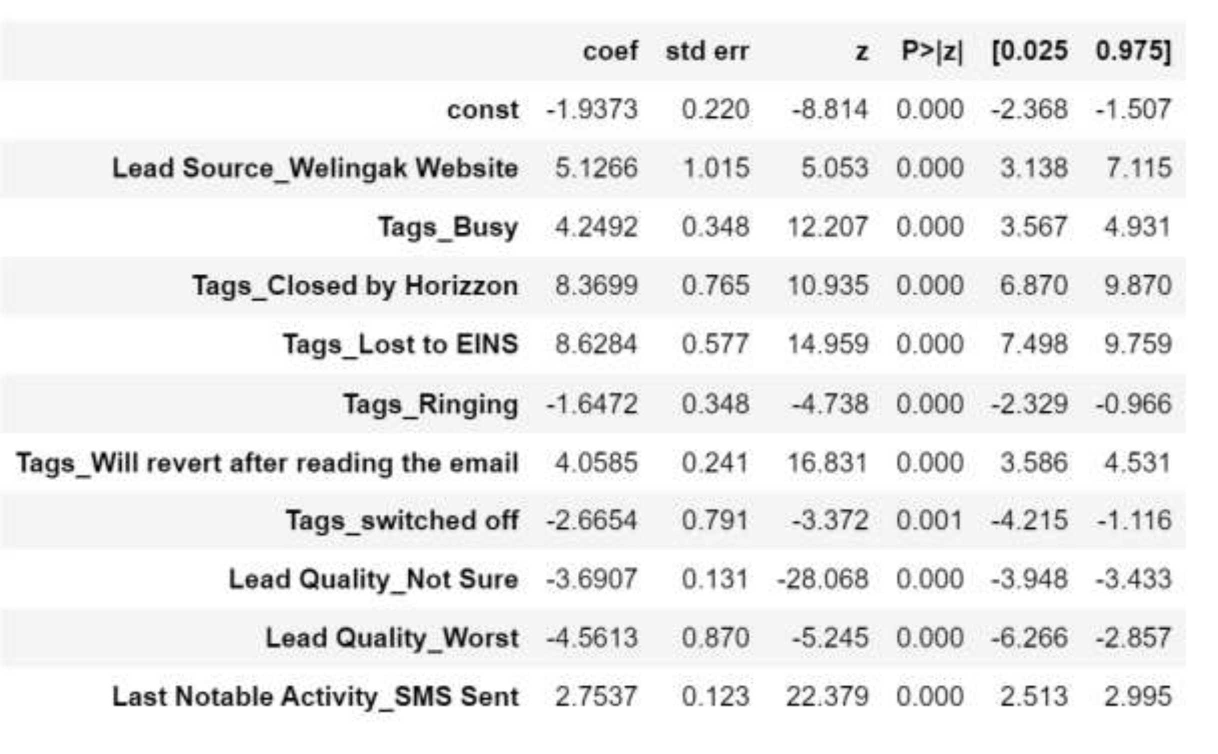
Model Building

After creating a RFE we got are model as shown below



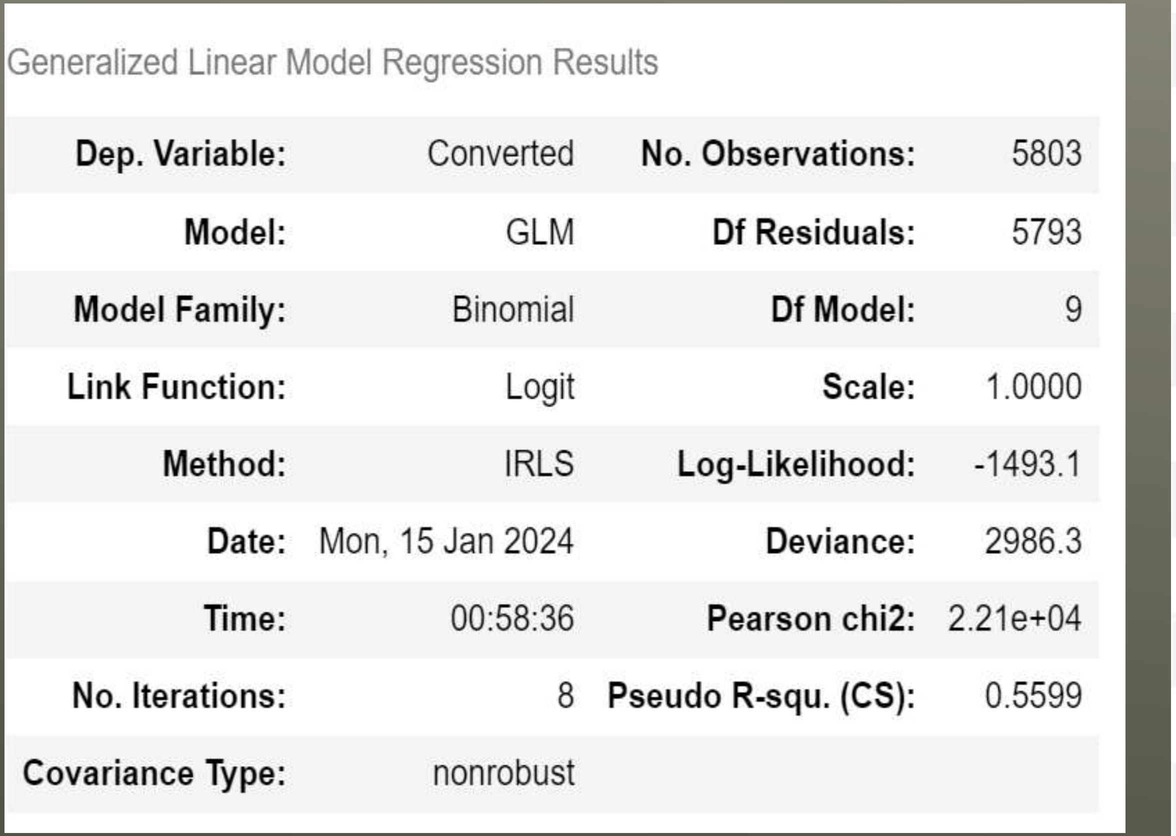
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	5803
Model:	GLM	**Df Residuals:**	5792
Model Family:	Binomial	**Df Model:**	10
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1484.3
Date:	Mon, 15 Jan 2024	**Deviance:**	2968.5
Time:	00:58:36	**Pearson chi2:**	2.57e+04
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.5613
Covariance Type:	nonrobust		



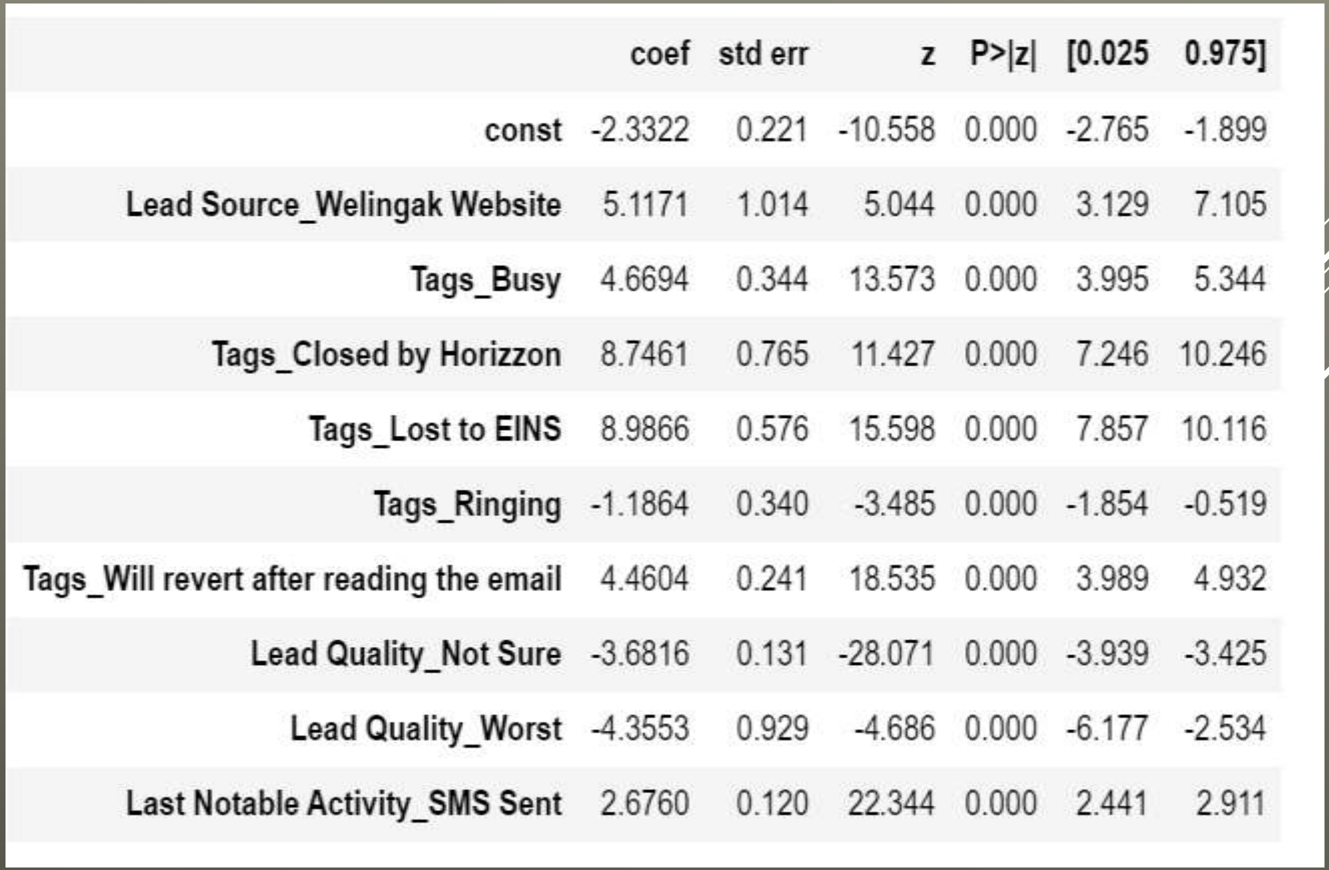
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.9373	0.220	-8.814	0.000	-2.368	-1.507
Lead Source_Welingak Website	5.1266	1.015	5.053	0.000	3.138	7.115
Tags_Busy	4.2492	0.348	12.207	0.000	3.567	4.931
Tags_Closed by Horizzon	8.3699	0.765	10.935	0.000	6.870	9.870
Tags_Lost to EINS	8.6284	0.577	14.959	0.000	7.498	9.759
Tags_Ringing	-1.6472	0.348	-4.738	0.000	-2.329	-0.966
Tags_Will revert after reading the email	4.0585	0.241	16.831	0.000	3.586	4.531
Tags_switched off	-2.6654	0.791	-3.372	0.001	-4.215	-1.116
Lead Quality_Not Sure	-3.6907	0.131	-28.068	0.000	-3.948	-3.433
Lead Quality_Worst	-4.5613	0.870	-5.245	0.000	-6.266	-2.857
Last Notable Activity_SMS Sent	2.7537	0.123	22.379	0.000	2.513	2.995

After Removing the variables with high p-value finally we got are final model as shown below:



Generalized Linear Model Regression Results

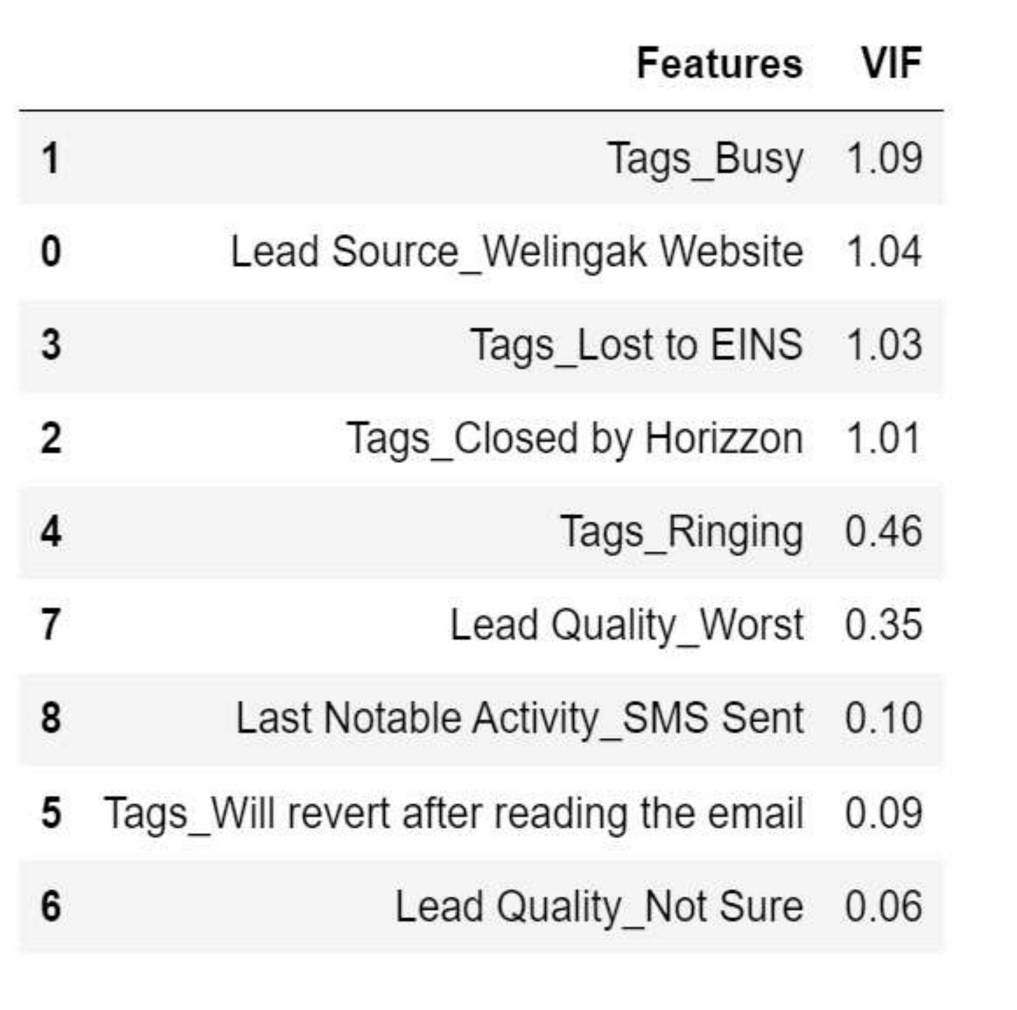
Dep. Variable:	Converted	**No. Observations:**	5803
Model:	GLM	**Df Residuals:**	5793
Model Family:	Binomial	**Df Model:**	9
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1493.1
Date:	Mon, 15 Jan 2024	**Deviance:**	2986.3
Time:	00:58:36	**Pearson chi2:**	2.21e+04
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.5599
Covariance Type:	nonrobust		



	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.3322	0.221	-10.558	0.000	-2.765	-1.899
Lead Source_Welingak Website	5.1171	1.014	5.044	0.000	3.129	7.105
Tags_Busy	4.6694	0.344	13.573	0.000	3.995	5.344
Tags_Closed by Horizzon	8.7461	0.765	11.427	0.000	7.246	10.246
Tags_Lost to EINS	8.9866	0.576	15.598	0.000	7.857	10.116
Tags_Ringing	-1.1864	0.340	-3.485	0.000	-1.854	-0.519
Tags_Will revert after reading the email	4.4604	0.241	18.535	0.000	3.989	4.932
Lead Quality_Not Sure	-3.6816	0.131	-28.071	0.000	-3.939	-3.425
Lead Quality_Worst	-4.3553	0.929	-4.686	0.000	-6.177	-2.534
Last Notable Activity_SMS Sent	2.6760	0.120	22.344	0.000	2.441	2.911

Variance Influence Factor (V.I.F)

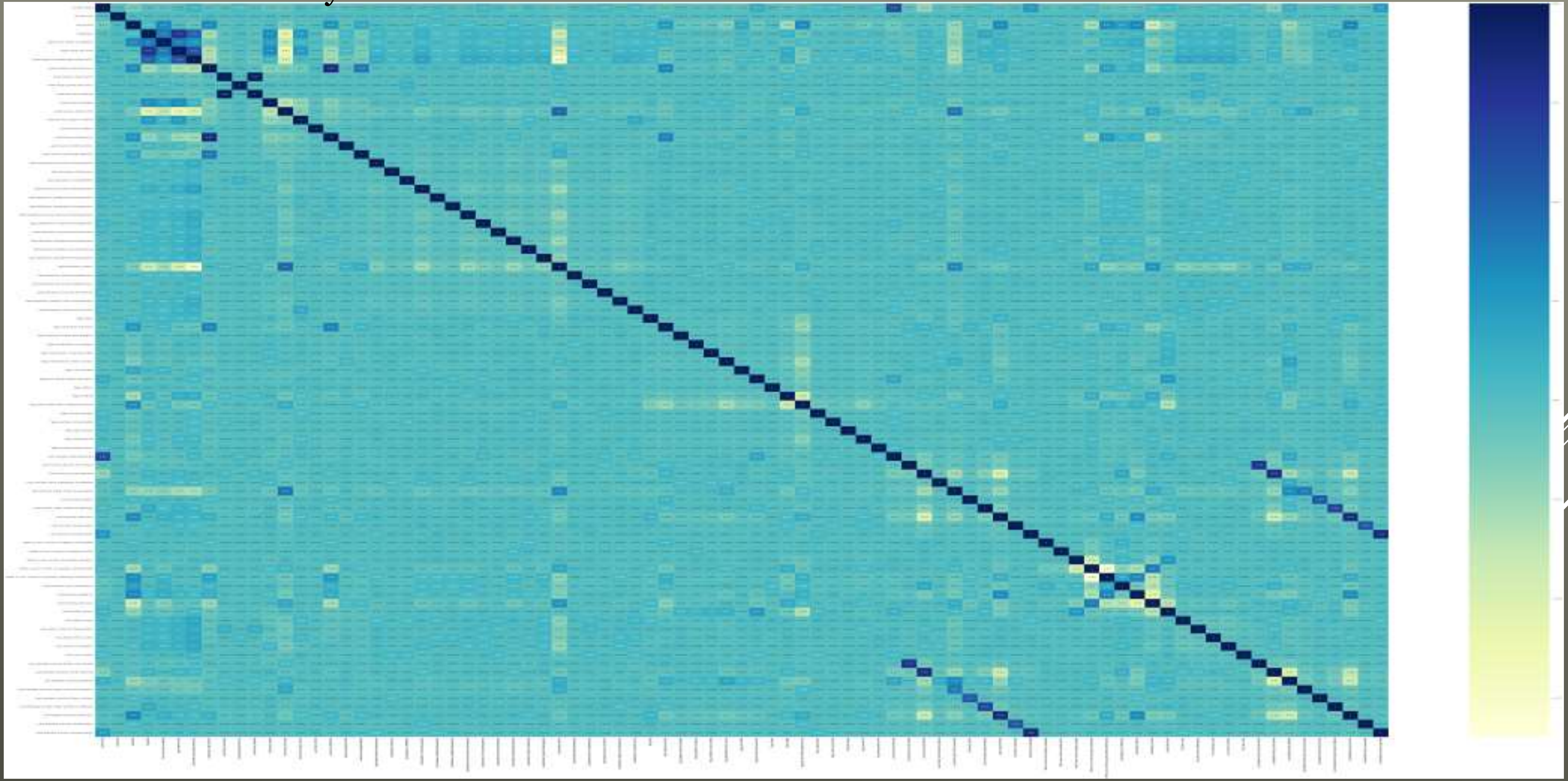
V.I.F Values for the final model.



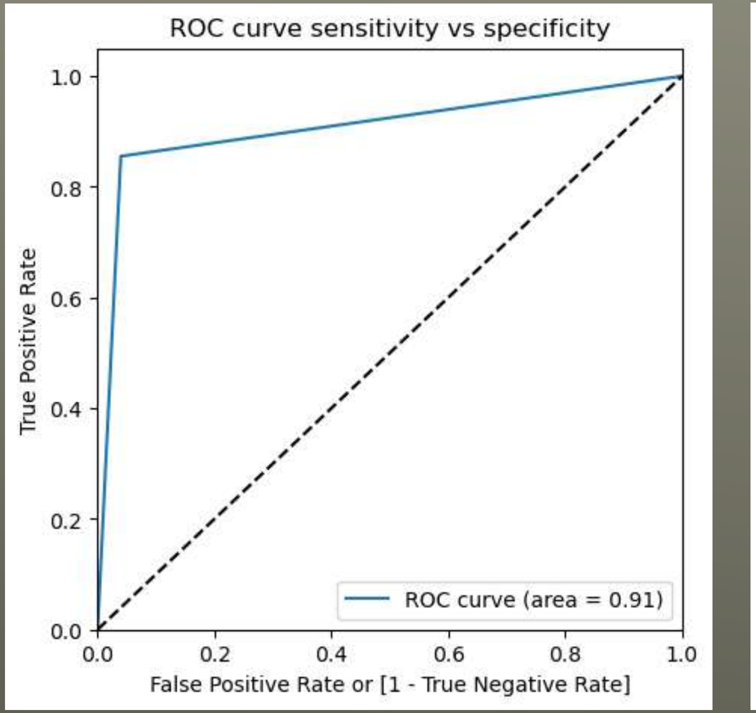
	Features	VIF
1	Tags_Busy	1.09
0	Lead Source_Welingak Website	1.04
3	Tags_Lost to EINS	1.03
2	Tags_Closed by Horizzon	1.01
4	Tags_Ringing	0.46
7	Lead Quality_Worst	0.35
8	Last Notable Activity_SMS Sent	0.10
5	Tags_Will revert after reading the email	0.09
6	Lead Quality_Not Sure	0.06

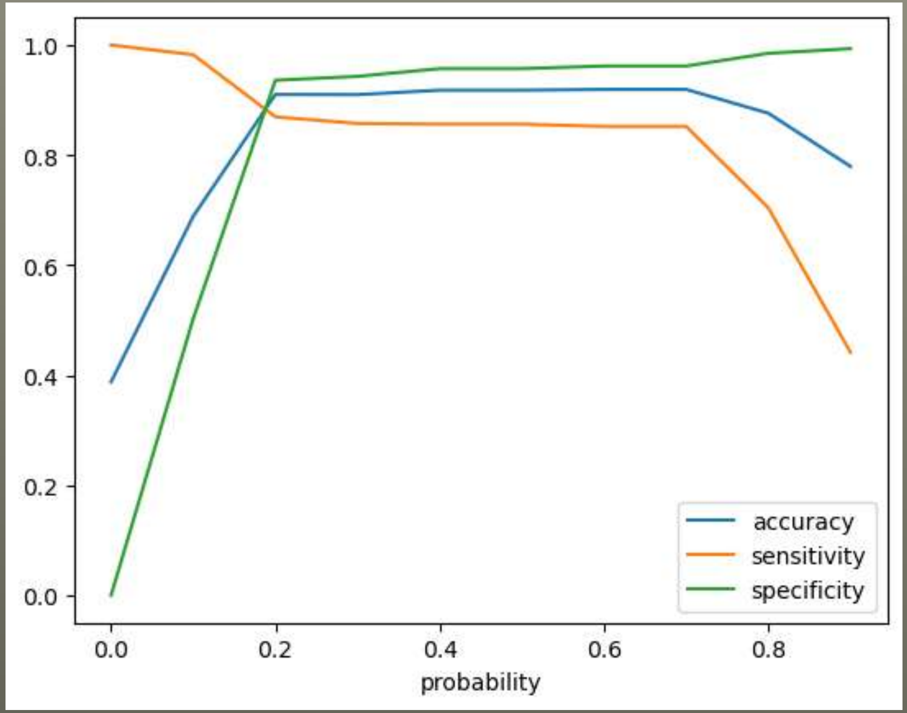
Correlations

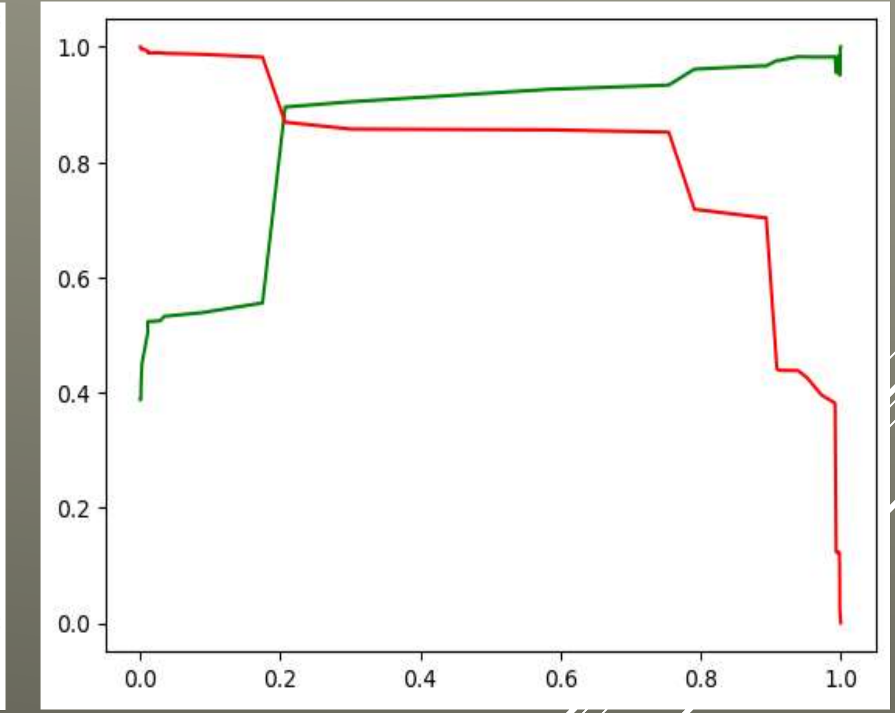
There are not many high corelations except the ones like Last Activity_Unsubscribed and Last Notable Activity_Unsubscribed. These are the type of corelations that dont make sense as they are actually same variables told differently.



ROC Curve   Optimal Cut-off point   Probability







ROC Curve 0.91

From the above curve 0.2 is the optimum probability as that's where the accuracy, sensitivity and specificity coincide.

Probability

Train Data

Accuracy	91.78%
Sensitivity	0.8693
Specificity	0.9358
False Positive Rate	0.0641
Positive Predictive value	0.8956
Negative Predictive value	0.9187
True Positive rate	0.8693
False Positive rate	0.0641
Precision	0.8956
Recall	0.8693

Test Data

Accuracy	91.31%
Sensitivity	0.8705
Specificity	0.9381
False Positive Rate	0.0618
Positive Predictive value	0.8918
Negative Predictive value	0.9251
True Positive rate	0.8705
False Positive rate	0.0618
Precision	0.8918
Recall	0.8705

Conclusions:

- ✓ This logistic regression model in our analysis primraily focuses on estimating the probability of a particular value for the target variable instead of directly forecasting the target column value for every lead. Also, a threshold is used to derive the predicted value for the target variable.
- ✓ In our model, the logistic regression is applied to predict the likelihood of a lead's conevrsion into customer.
- ✓ 0.233 is the optimum probability as thats where the accuracy, sensitivity and specificity coincide, so any lead with probability greater than 0.233 will be classified as a "Hot Lead" whereas any lead below this values will be a cold lead.
- ✓ Our finalized logistic model comprises 12 features with coefficients:

1. Lead Source_Welingak Website	5.1171
2. Tags_Busy	4.6694
3. Tags_Closed by Horizzon	8.7461
4. Tags_Lost to EINS	8.9866
5. Tags_Ringing	-1.1864
6. Tags_Will revert after reading the email	4.4604
7. Lead Quality_Not Sure	-3.5353
8. Lead Quality_Worst	-4.3553
9. Last Notable Activity_SMS Sent	2.6760

Recommendation:

1. **Targeted Email Campaigns:**
   - Avoid 'Do Not Email' leads (-1.4006) to refine targeting.
   - Tailor communication strategies to prevent opt-outs.

2. **Leverage Welingak Website:**
   - Allocate more resources to leads from Welingak Website (coef: 3.9789).
   - Maximize marketing efforts on this high-converting source.

3. **Prioritize Specific Tags:**
   - Focus on 'Busy' (coef: 2.4518), 'Closed by Horizzon' (coef: 8.0902), and 'Lost to EINS' (coef: 7.2135).
   - Develop targeted content and engagement strategies for these categories.

4. **Mitigate Negative Tags:**
   - Address 'Ringing' (coef: -1.7804) and 'switched off' (coef: -2.3905) impact.
   - Tailor interventions to re-engage leads with these tags.

5. **Enhance Olark Chat:**
   - Improve effectiveness of Olark Chat (coef: -2.0735) in lead conversion.
   - Provide additional support or incentives during chat interactions.

6. **Refine Lead Quality Assessment:**
   - Reevaluate criteria for 'Lead Quality Worst' (coef: -2.4141) to minimize false negatives.
   - Enhance lead scoring system for better reflection of potential conversions.

7. **Maximize SMS Sent Activities:**
   - Capitalize on positive impact (coef: 2.5791) of 'Last Notable Activity_SMS Sent'.
   - Increase frequency of SMS-based interactions for improved conversion rates.

8. **Overall Strategy:**
   - Continuously monitor and adjust strategies based on model performance.
   - Train sales team to effectively use lead scoring system.

9. **Cross-Functional Collaboration:**
   - Foster collaboration between marketing, sales, and customer support teams.
   - Share insights and align efforts to maximize lead conversion opportunities.

10. **Customer Feedback Integration:**
    - Incorporate customer feedback into the model for real-time adjustments.
    - Enhance predictive power by integrating external feedback.

Thank you