Lead Scoring Case Study

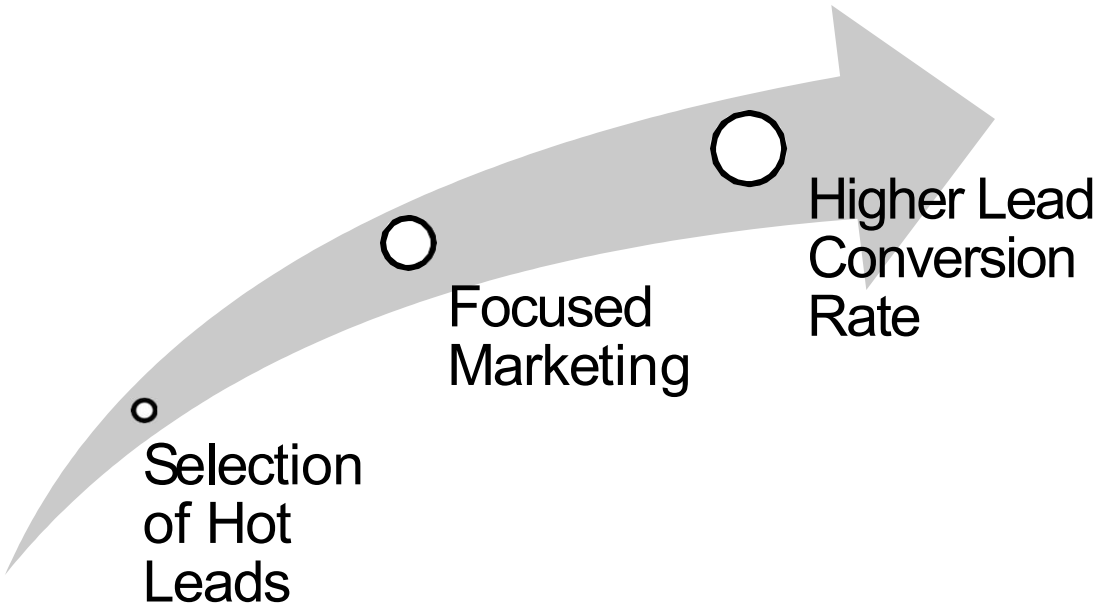
Presentation Made By-
Miss Ruchi Gupta
Mr. Rudresh Sishodia
Mr. Sri Nath S

Problem Statement:

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor i.e., 30%.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objective:

- X Education has appointed us to help them select the most promising leads(**HOT Leads**), i.e. the leads that are most likely to convert into paying customers.
- The Conversion rate should rise from 30% to 80% and above.

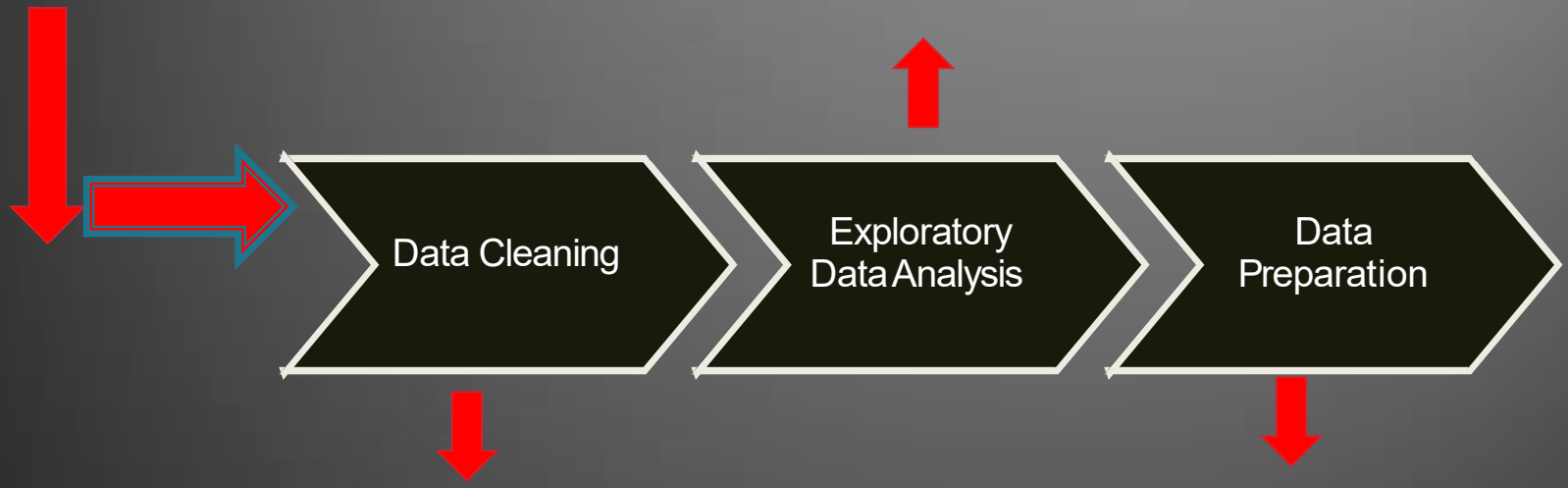


Methodology:

- To build a logistic regression model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. Target Lead Conversion Rate $\approx$ 80%

Importing and Observing the past data providedby the Company

Univariate and Bivariate analysis

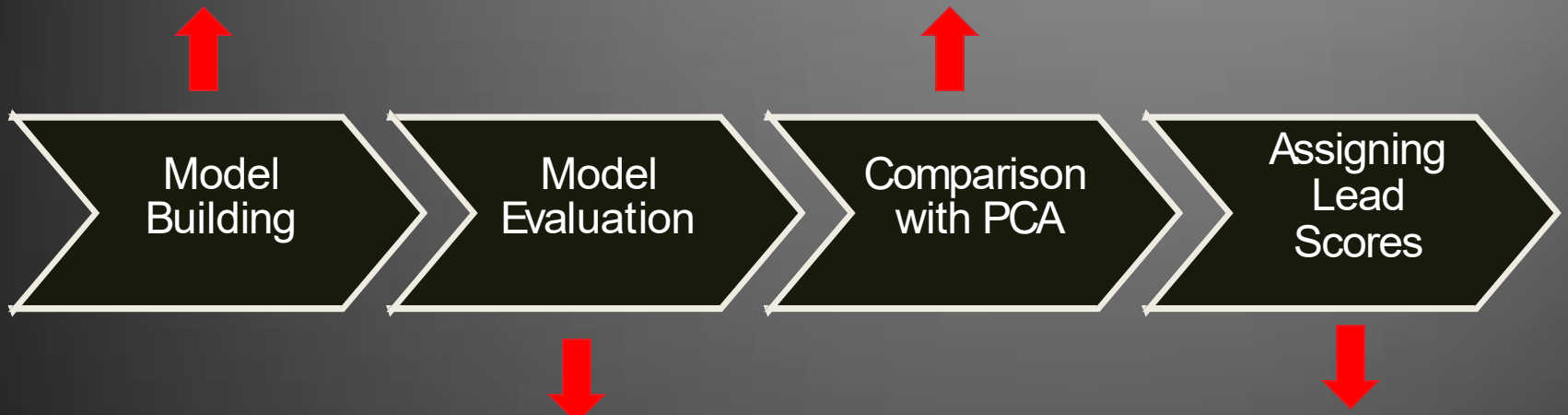


- Missing value imputation
- Removing duplicate data and other redundancies

- Outlier treatment
- Dropping unnecessary columns
- Dummy variable creation
- Feature standardization

- Feature selection using RFE
- Manual feature elimination based on p-values and VIFs

- Building another model using PCA
- Comparing the two models



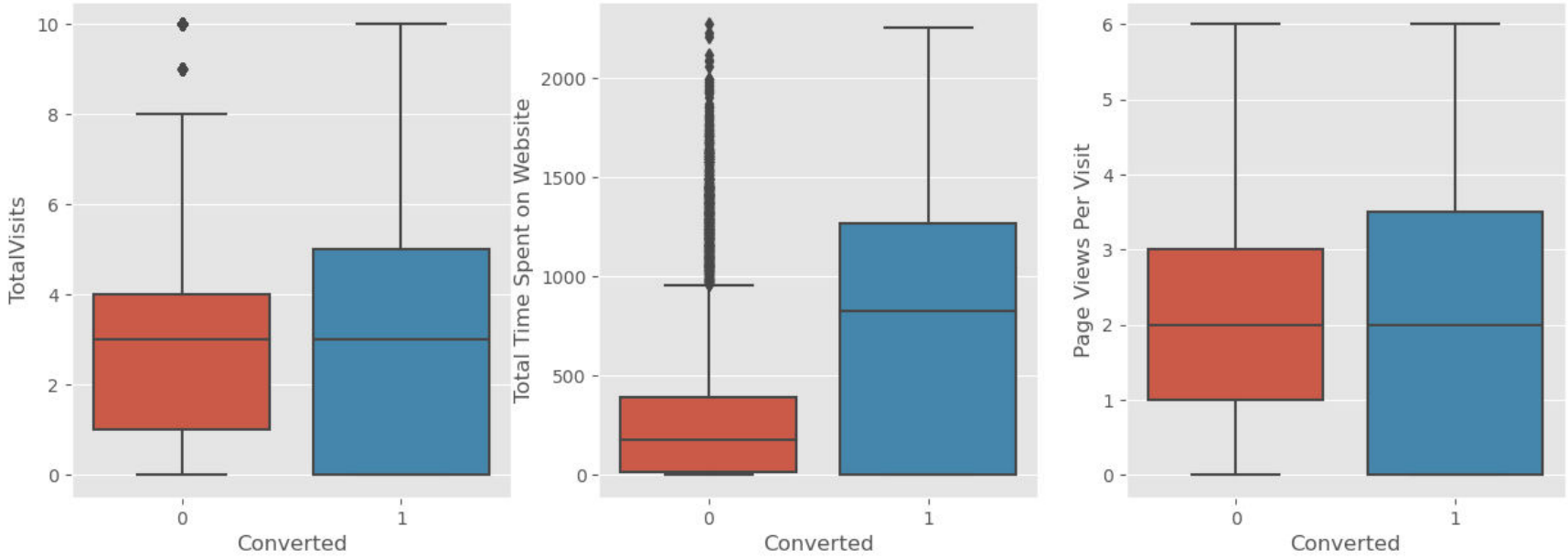
- Evaluating model based on various evaluation metrics
- Finding the optimal probability threshold

- Finalizing the first model
- Using predicted probabilities to calculate Lead Scores:
  Lead Score = Probability * 100

Data Visualization:

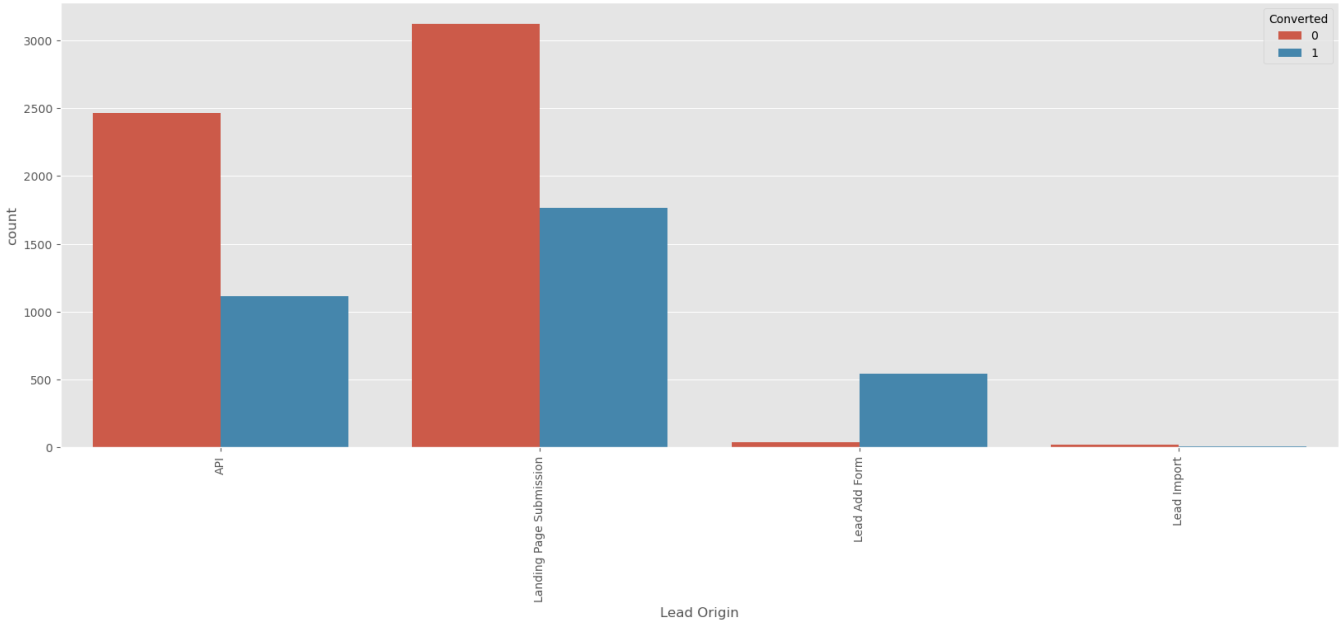
- To identify important features.
- To get insights.
- To analyze the dataset and draw meaningful conclusions.

Numerical Variables:



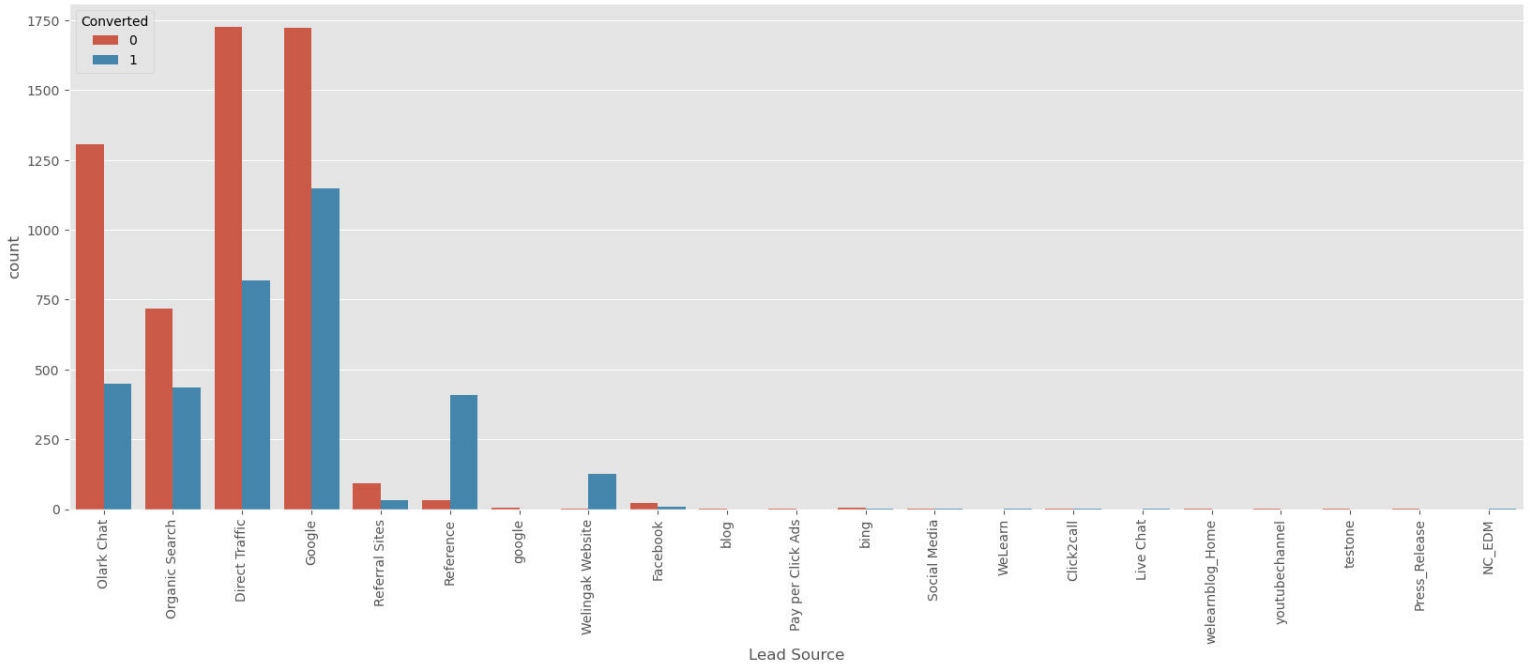
- People spending more time on website are more likely to get converted.

Lead Origin:



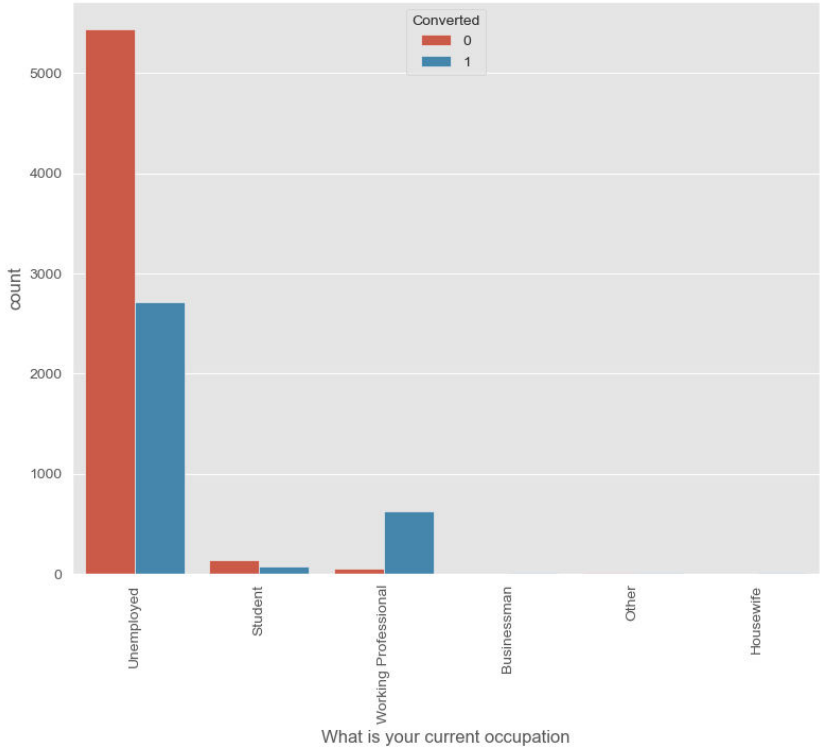
- ‘API’ and ‘Landing Page Submission’ generate the most leads but have less conversion rates, whereas ‘Lead Add Form’ generates less leads but conversion rate is great.
- Try to increase conversion rate for ‘API’ and ‘Landing Page Submission’, and increase leads generation using ‘Lead Add Form’.

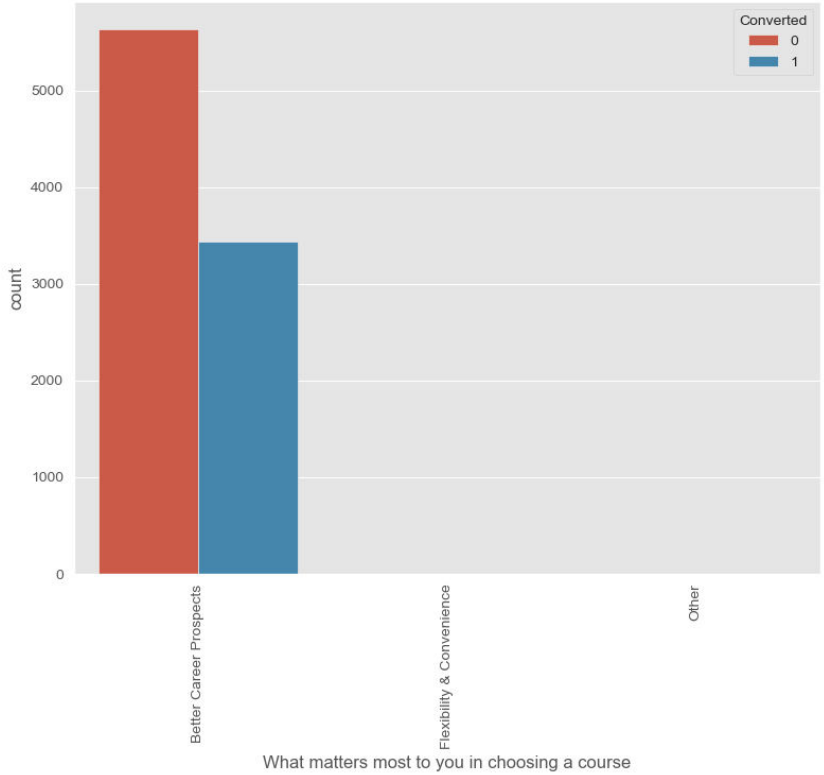
Lead Source:



- Most leads are generated through 'Direct Traffic' and 'Google'.
- Very high conversion rates for lead sources 'Reference' and 'Welingak Website'.

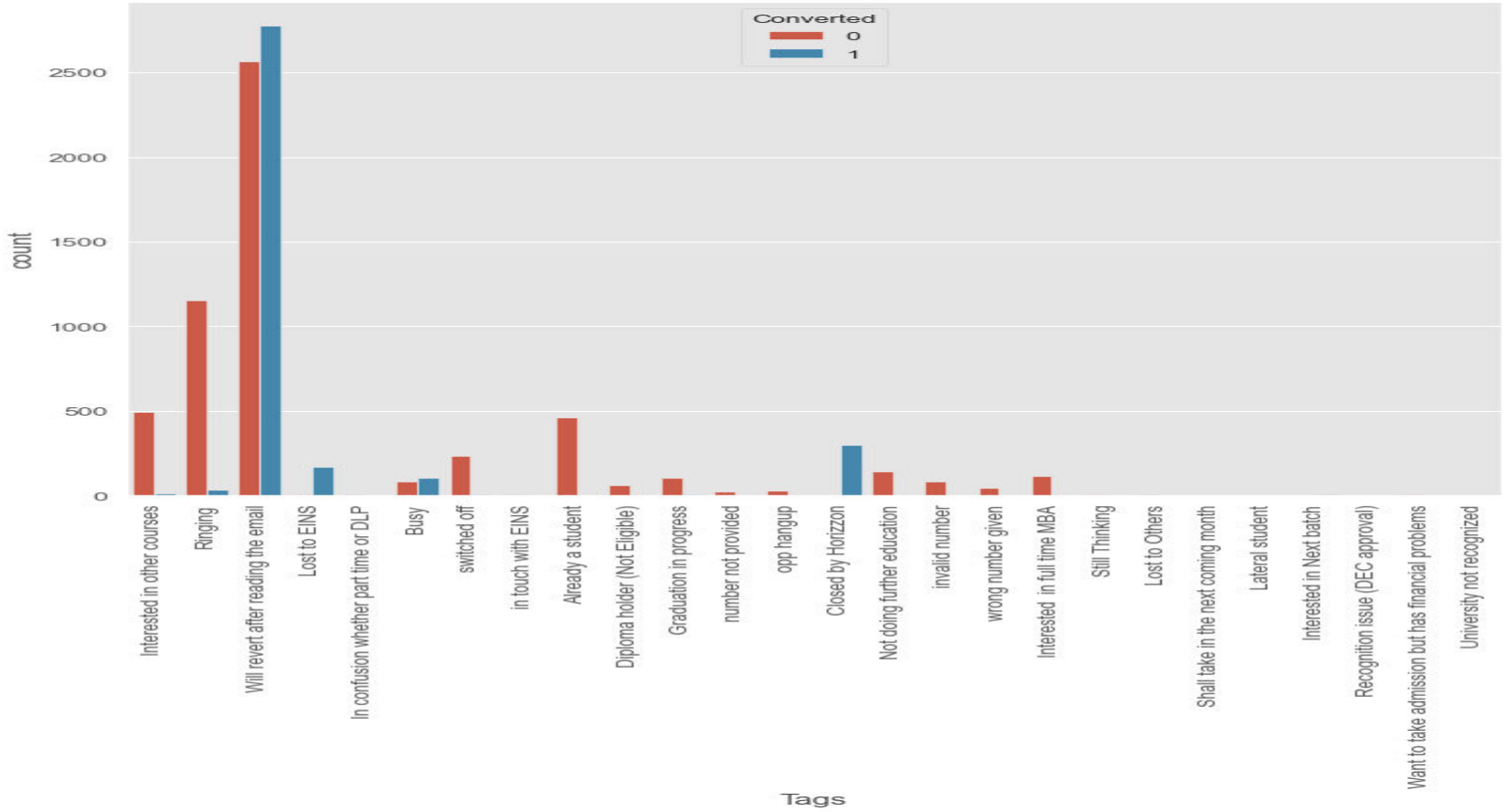
Current Occupation:





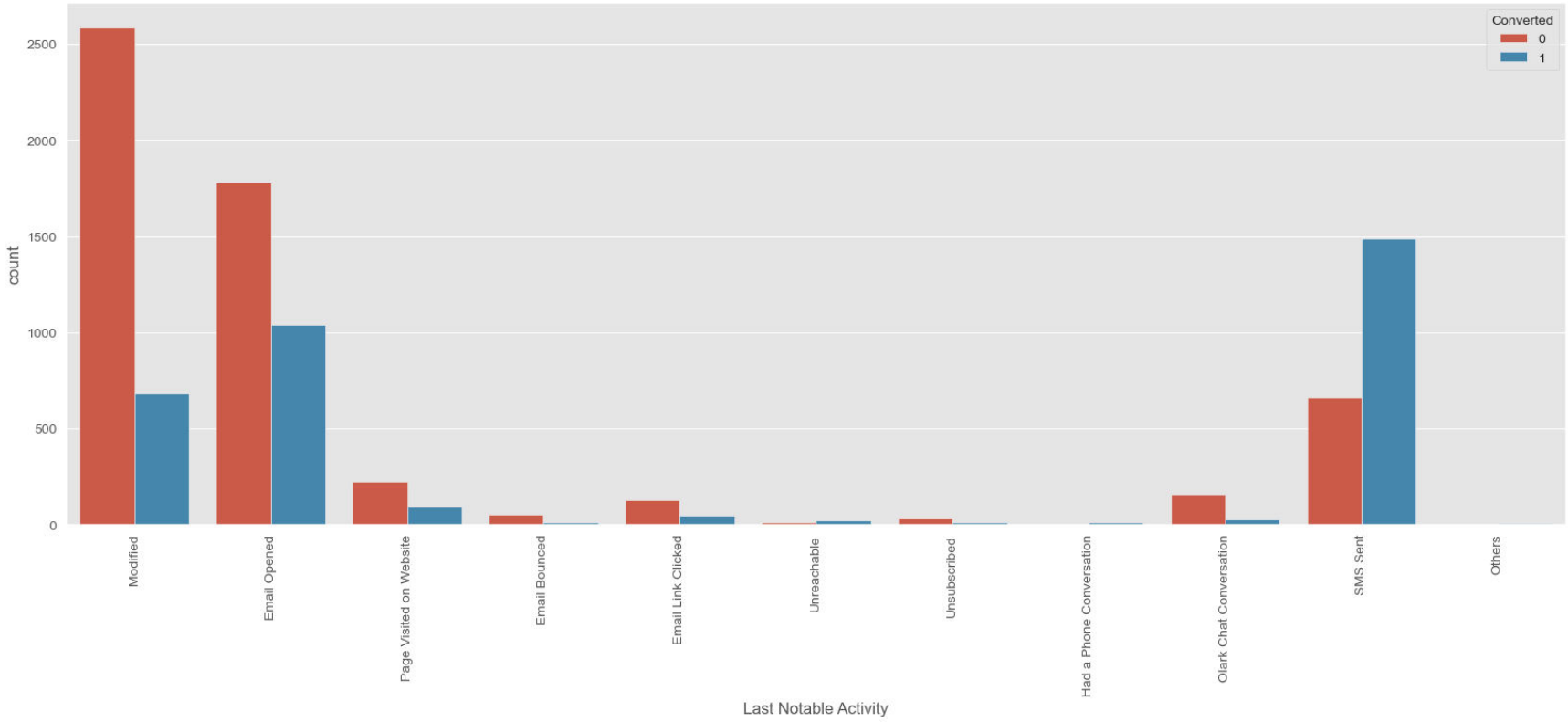
- Working Professionals have higher chance of conversion.

Tags:



- Tags with **'Will revert after reading the email', 'Closed by Horizon', 'Lost to EINS', and 'Busy'** have higher chance of conversion.

Last Notable Activity:



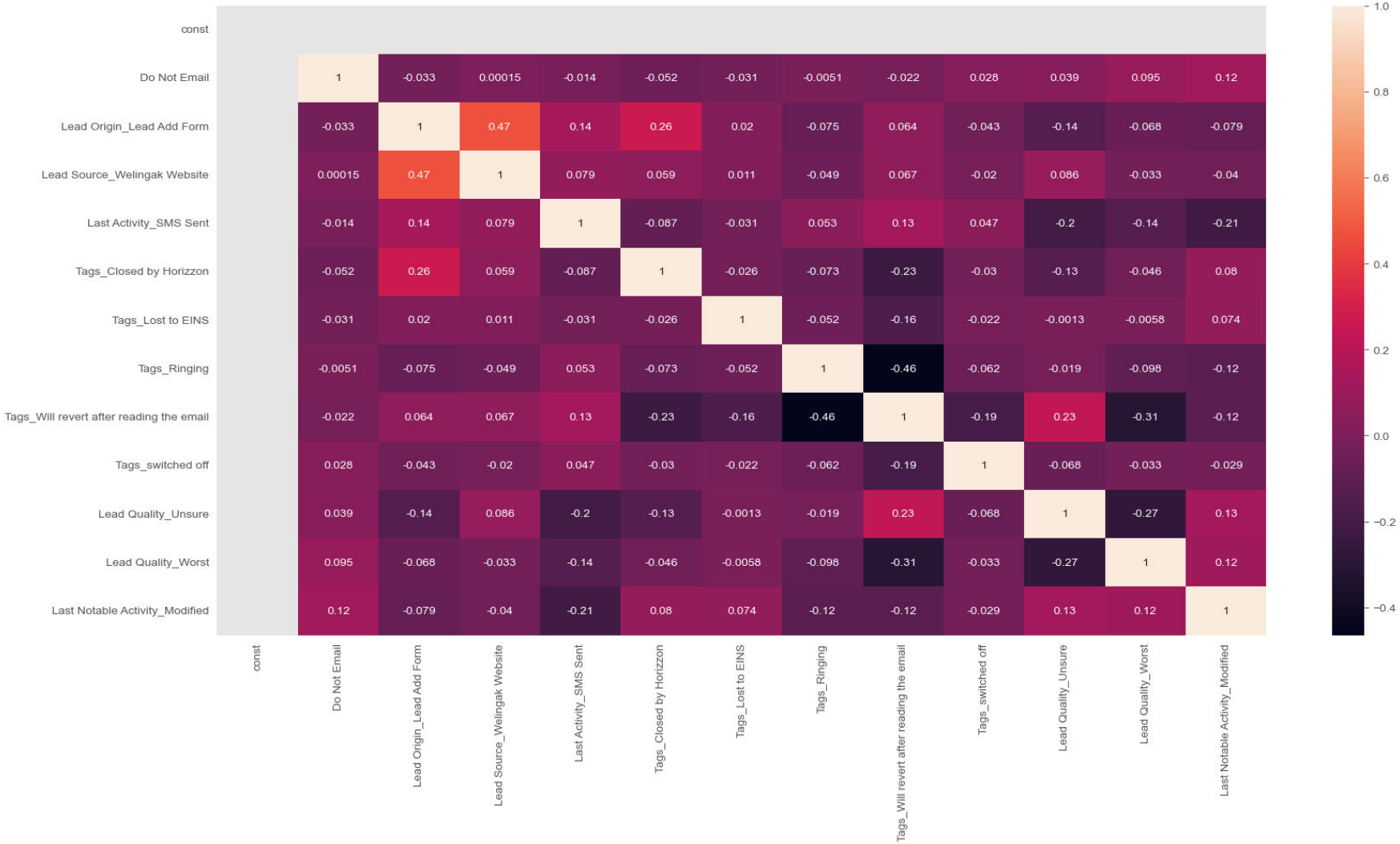
- The last notable activity ‘SMS Sent’ has the highest conversion rate.

MODEL EVALUATION

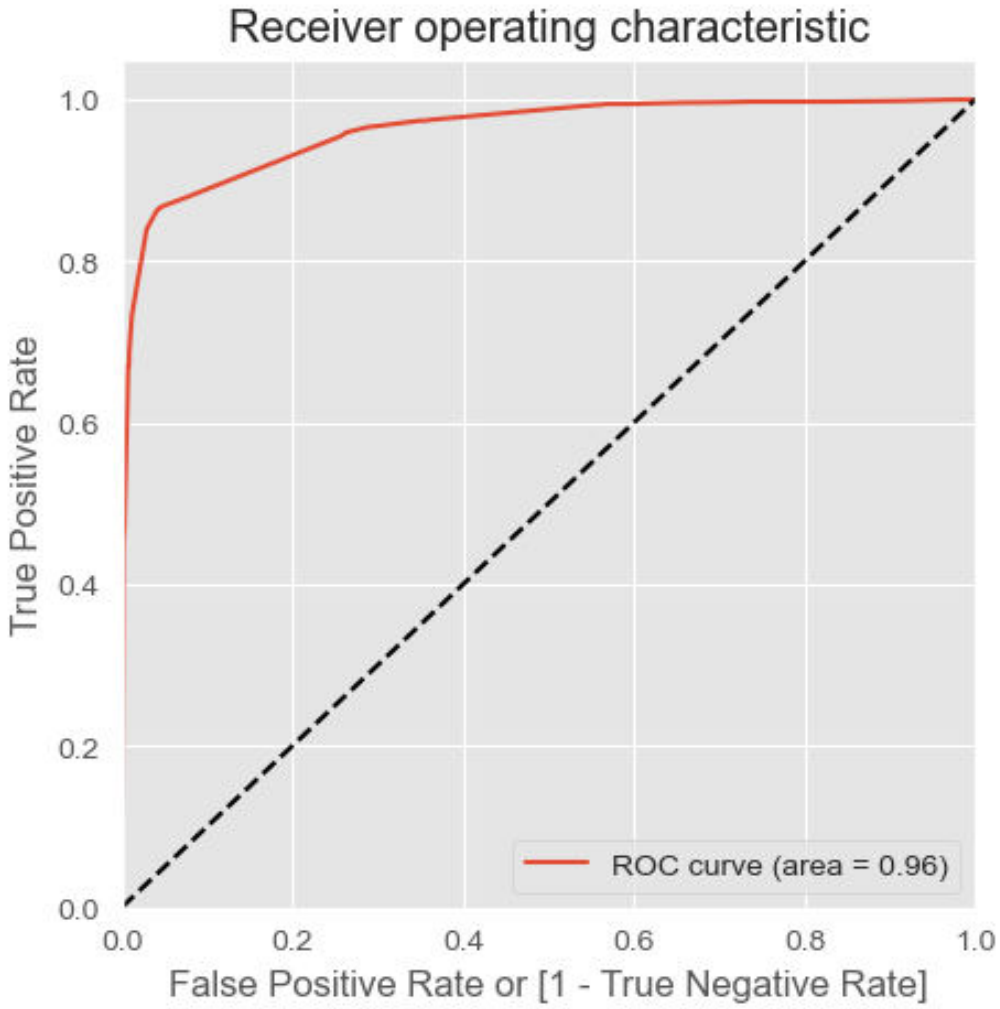
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6338
Model Family:	Binomial	Df Model:	12
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1455.2
Date:	Thu, 21 Sep 2023	Deviance:	2910.3
Time:	13:07:26	Pearson chi2:	1.81e+05
No. Iterations:	9	Pseudo R-squ. (CS):	0.5831
Covariance Type:	nonrobust		

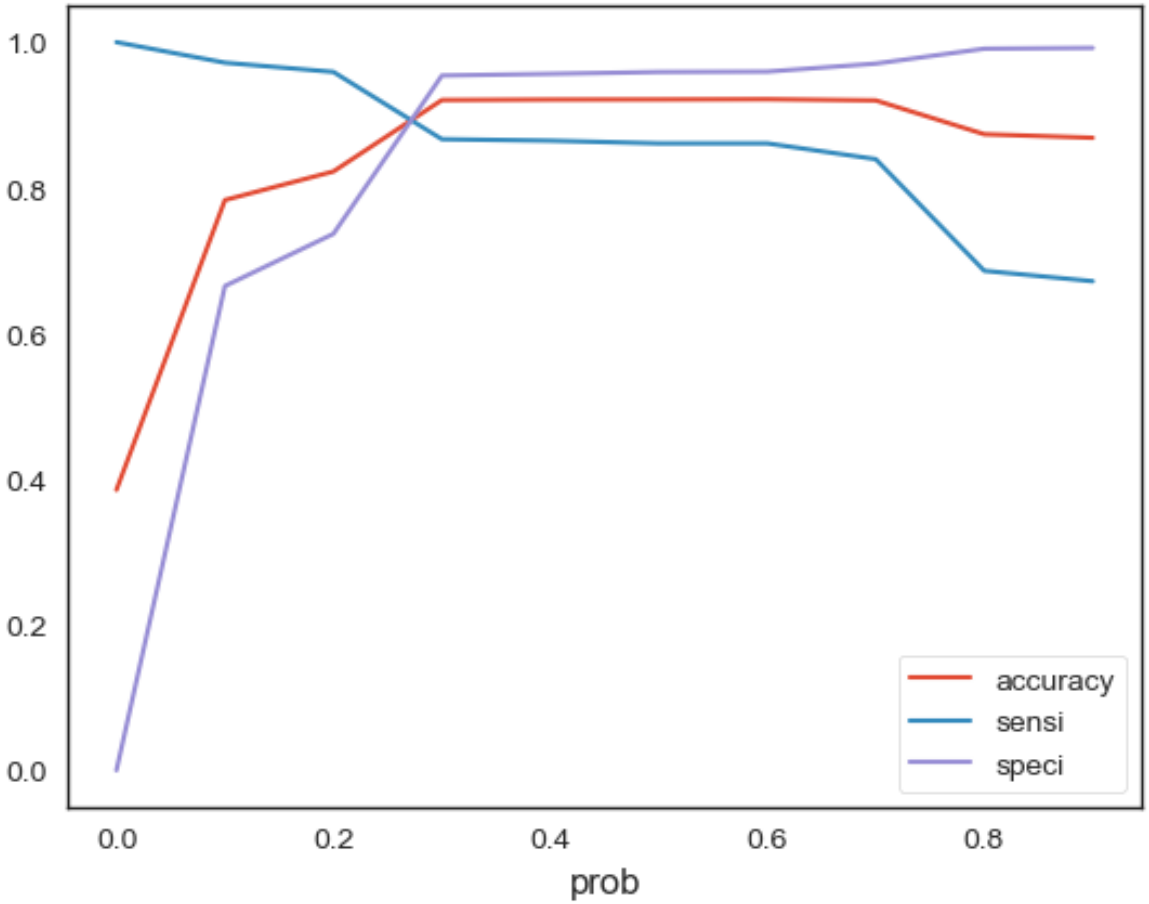
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.5423	0.146	-10.558	0.000	-1.829	-1.256
Do Not Email	-1.2511	0.230	-5.439	0.000	-1.702	-0.800
Lead Origin_Lead Add Form	0.8170	0.422	1.936	0.053	-0.010	1.644
Lead Source_Welingak Website	3.5621	0.847	4.204	0.000	1.901	5.223
Last Activity_SMS Sent	2.0339	0.105	19.344	0.000	1.828	2.240
Tags_Closed by Horizzon	9.5337	0.780	12.220	0.000	8.005	11.063
Tags_Lost to EINS	10.3529	0.769	13.471	0.000	8.847	11.859
Tags_Ringing	-2.6778	0.273	-9.824	0.000	-3.212	-2.144
Tags_Will revert after reading the email	4.6184	0.199	23.171	0.000	4.228	5.009
Tags_switched off	-2.9532	0.537	-5.504	0.000	-4.005	-1.902
Lead Quality_Unsure	-4.2330	0.157	-27.031	0.000	-4.540	-3.926
Lead Quality_Worst	-3.7989	0.887	-4.283	0.000	-5.537	-2.060
Last Notable Activity_Modified	-1.7342	0.114	-15.152	0.000	-1.959	-1.510



- In the final Model, **correlation** between features are **negligible.**



- Area under curve (auc) is approximately 0.96 which is very close to ideal auc of 1.

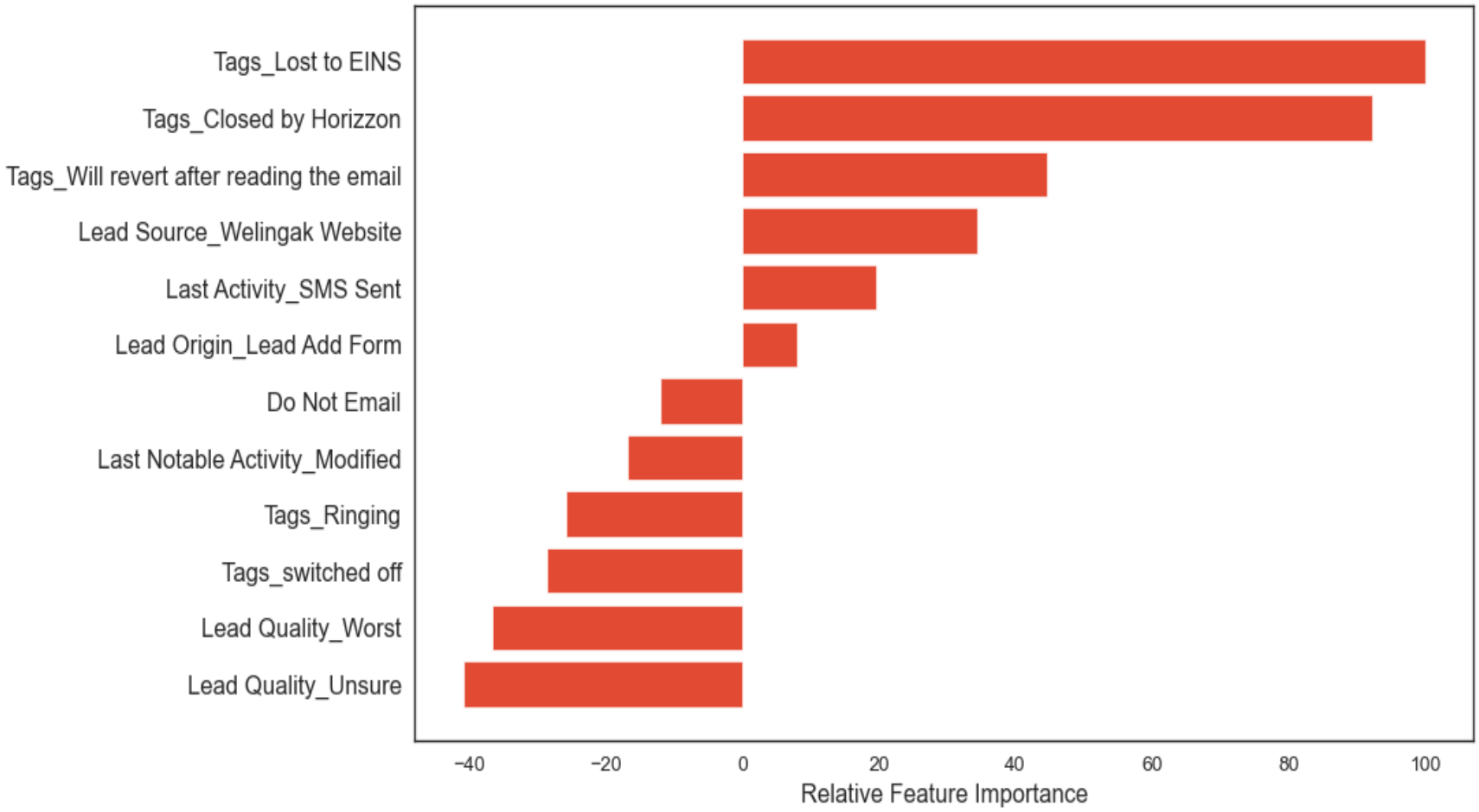


- Optimal cutoff = 0.25(appx.)
- Graph showing changes in Sensitivity, Specificity and Accuracy with changes in the probability threshold values

Final Results:

Data	Train set	Test set
Accuracy	0.8223	0.8968
Sensitivity	0.9591	0.8574
Specificity	0.7367	0.9192
False Positive Rate	0.0775	0.0807
Positive Predictive Value	0.8765	0.8582
Negative Predictive Value	0.9243	0.9187
AUC	0.9614	0.9506

Relative Importance of Features:



CONCLUSION

Feature Importance:

- Three variables which contribute most towards the probability of a lead conversion in decreasing order of impact are:
  - ✓ **Tags_Lost to EINS**
  - ✓ **Tags_Closed by Horizzon**
  - ✓ **Tags_Wil revert after reading the email**
- These are dummy features created from the categorical variable Tags.
- All the above mentioned variables contribute positively towards the probability of a lead conversion.
- These results indicate that the company should **focus more on the leads with these three tags** to increase its conversion rate.

Recommendation:

- Pay attention to the relative importance of the features in the model and their positive or negative impact on the probability of conversion of a lead.
- Based on the different business needs, modify the probability threshold value accordingly for identifying potential leads.
- By referring to the data visualizations, X Education should focus on:
  - Generate more leads for the categories having higher conversion rate
  - Increase conversion rate for the categories for which they have more leads.

THANKK YOU