LEAD SCORING CASE STUDY

Prashant Singh C43/EPGDS/IIITB

Vaibhav Singh C43/EPGDS/IIITB

BUSINESS OBJECTIVE

- To help X Education select most promising leads ***(Hot Leads)***, i.e. the leads that are most likely to convert into paying customers.

    - *Selection of Hot Leads*
    - *Focused Marketing*
    - *Higher Lead Conversion Rate*

METHODOLOGY

- To build a Logistic Regression model that assigns lead scores to all leads such that the customers with higher lead score usually have a higher conversion chance and vice versa.
    - Target Lead Conversion Rate ≈ 80%

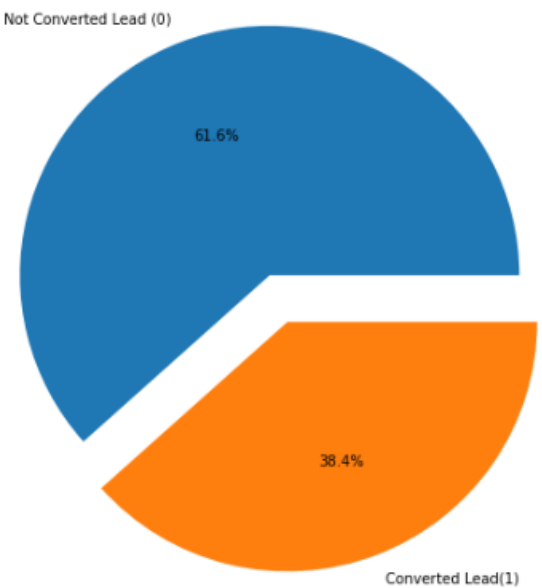
- Reading and Understanding the Data
  - Importing and Observing past data
- Data Cleaning
  - Missing value imputation
  - Removing duplicate data
  - and other redundancies
- Exploratory Data Analysis
  - Univariate and Bivariate analysis
- Data Preparation
  - Outlier treatment
  - Dropping unnecessary columns
  - Dummy variable creation
  - Feature standardization

- Model Building
  - Feature selection using RFE
  - Manual feature elimination based on p-values and VIFs
- Model Evaluation
  - Evaluating model based on various evaluation metrics
  - Finding the optimal probability threshold
  - Plotting ROC curve – Check AUC
  - Precision Recall Trade-Off
- Assigning Lead Scores
  - Finalizing the first model
  - Using predicted probabilities to
  - calculate Lead Scores:
  - Lead Score = Probability * 100

DATA VISUALIZATION

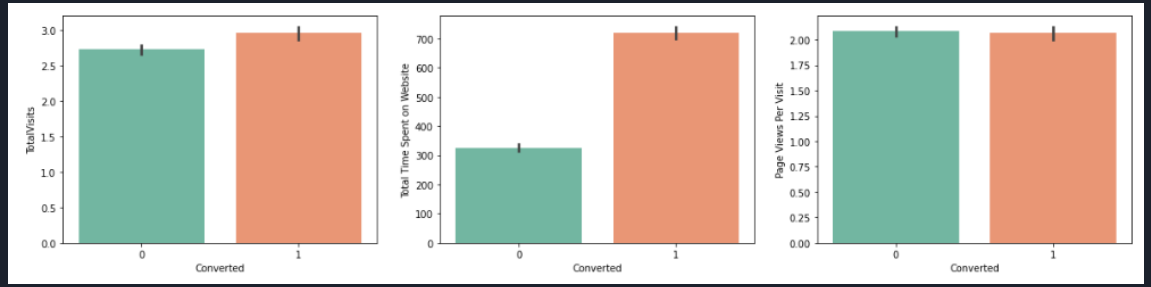
- To identify important features
- To get insights

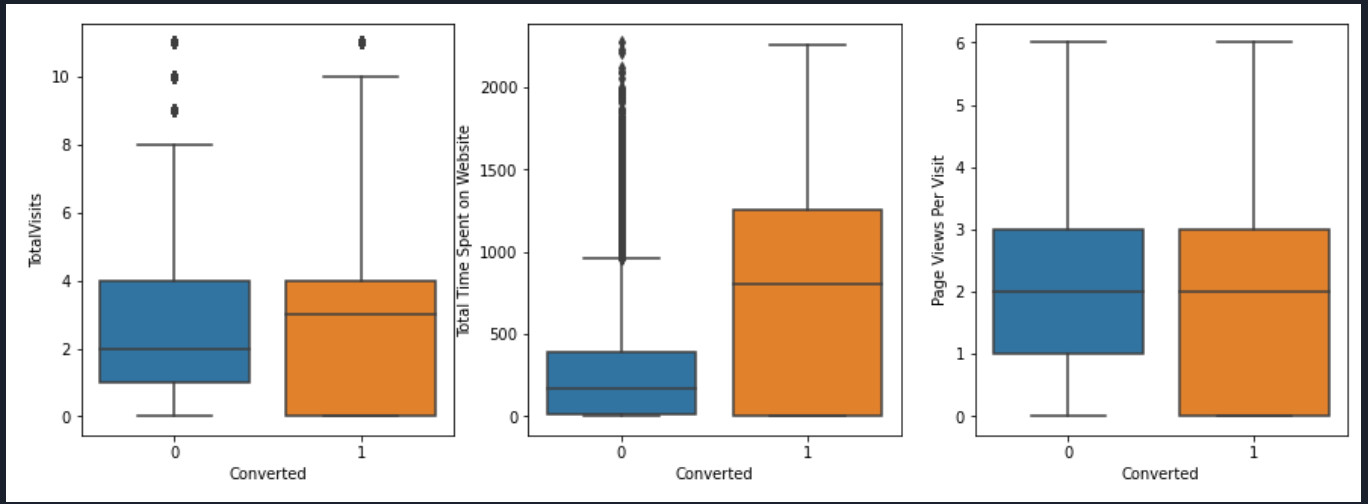
Data Imbalance analysis



Note:- As per above analysis Data is not much imbalanced 61.6% belong to Not Converted Leads and 38.4 % belong to Converted lead in Data.

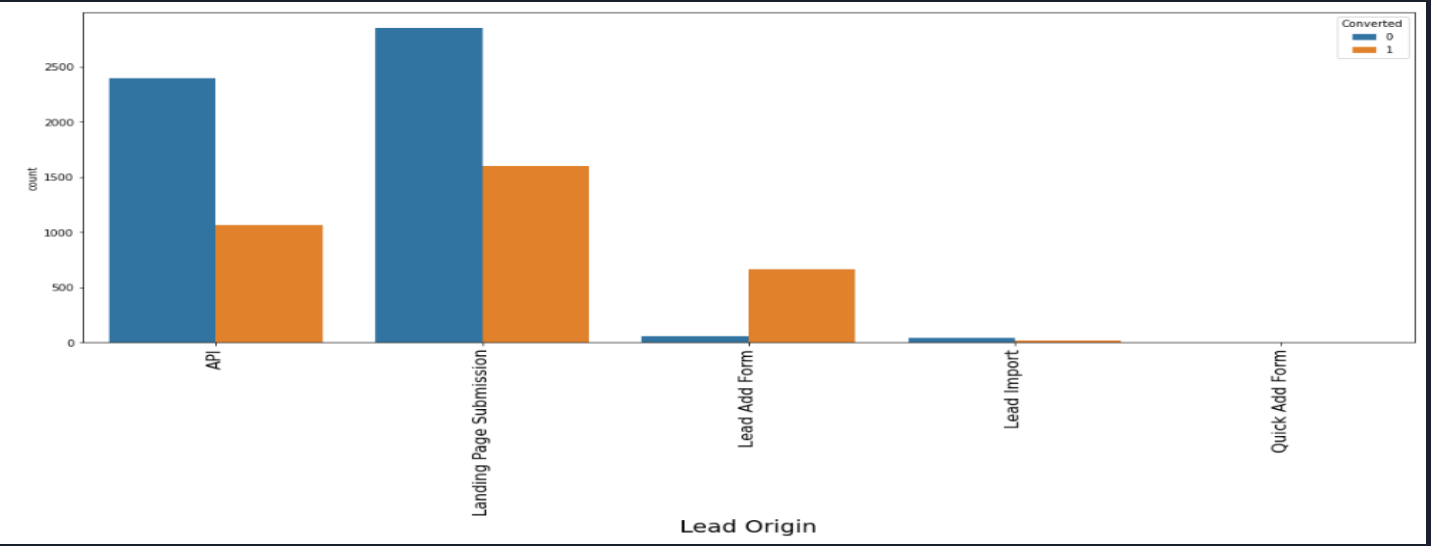
Numerical Variables





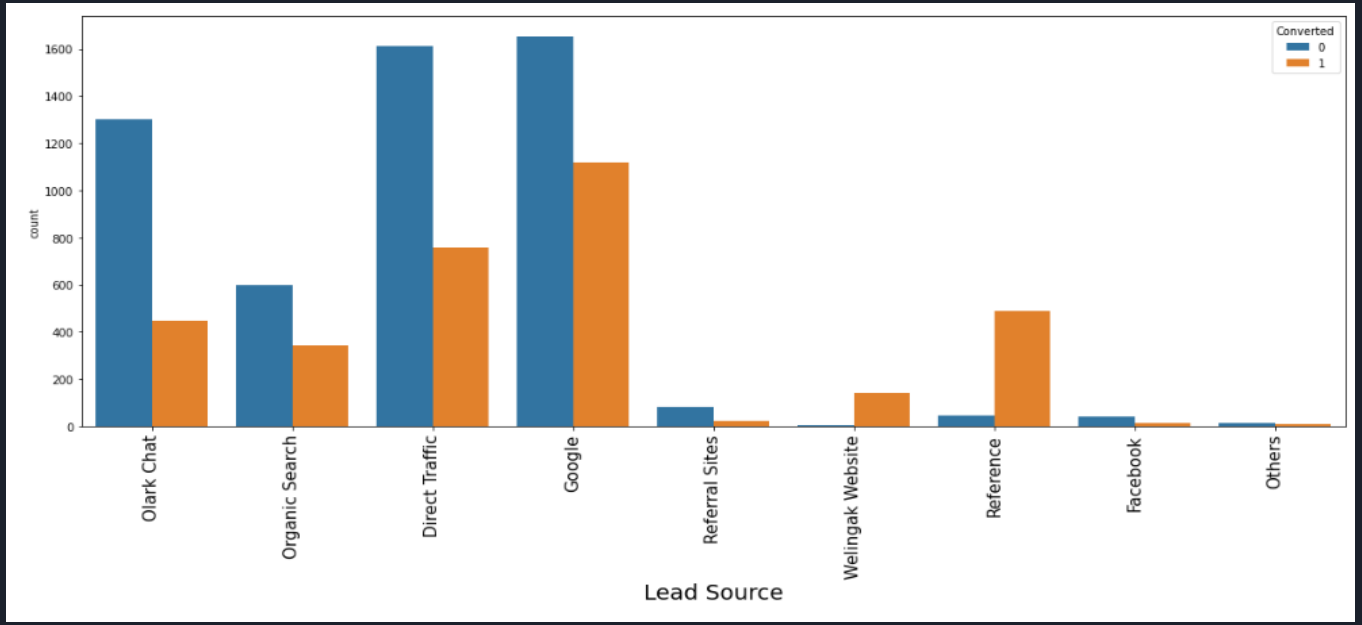
- People spending more time on websites are more likely to get converted.

Lead Origin



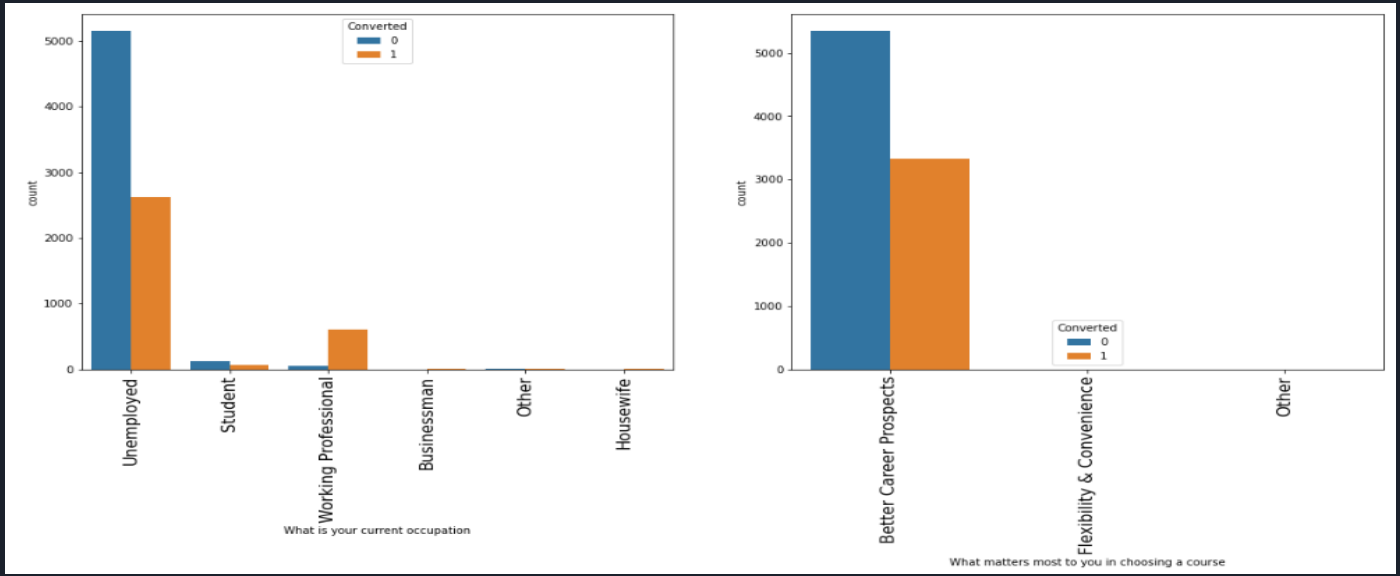
- ‘API’ and ‘Landing Page Submission’ generate the most leads but have less conversion rates.
  - Focus on the increasing conversion rate for ‘API’ and ‘Landing Page Submission.
- ‘Lead Add Form’ generates fewer leads but the conversion rate is great.
  - Focus on increasing leads generation using the ‘Lead Add Form’.

Lead Source



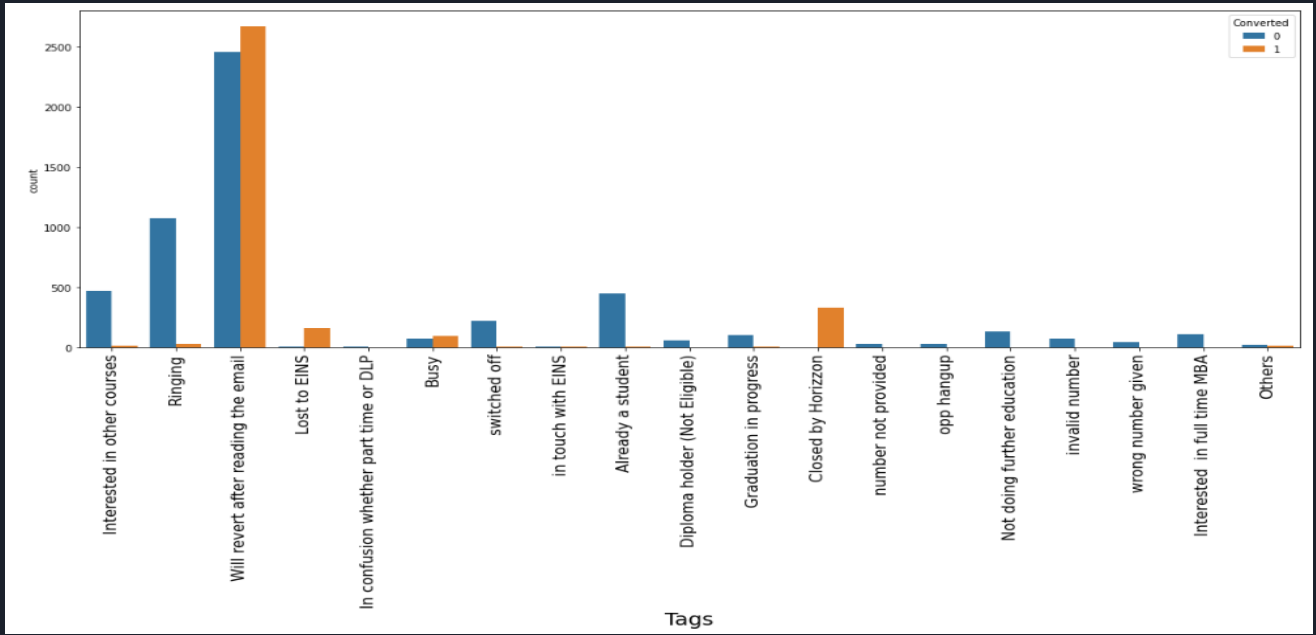
- Very high conversion rates for lead sources ‘Reference’ and ‘Welingak Website’.
- Most leads are generated through ‘Direct Traffic’ and ‘Google’.

Current Occupation



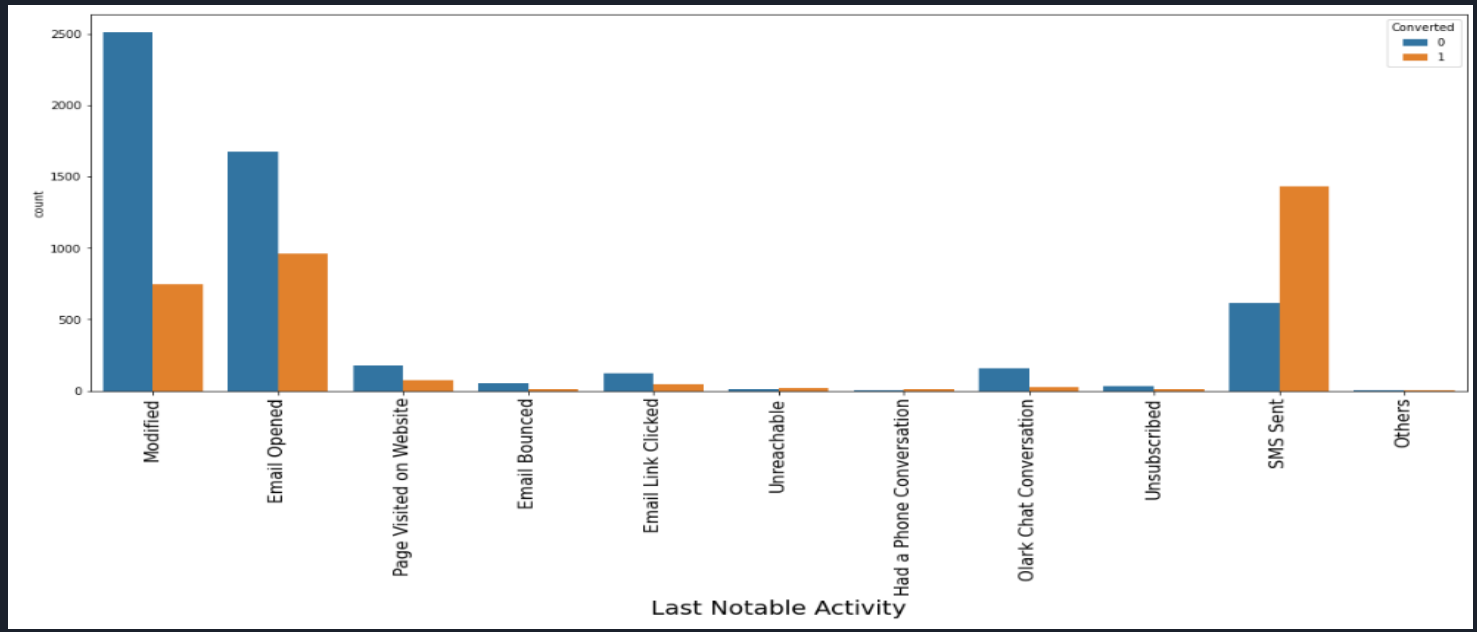
- Working Professionals are most likely to get converted.
- People choosing these coursed due to Better Career Prospects most leads belong to the same.

Tags



- *High conversion rates for tags 'Will revert after reading the email', 'Closed by Horizon', 'Lost to EINS', and 'Busy'.*

Last Notable Activity



- Highest conversion rate is for the last notable activity ‘SMS Sent’.
- We can also be focused on the “Email Opened” Conversion rate can be increased by well-drafted emails and relevant information-only emails to leads.

MODEL EVALUATION

Final Model Summary

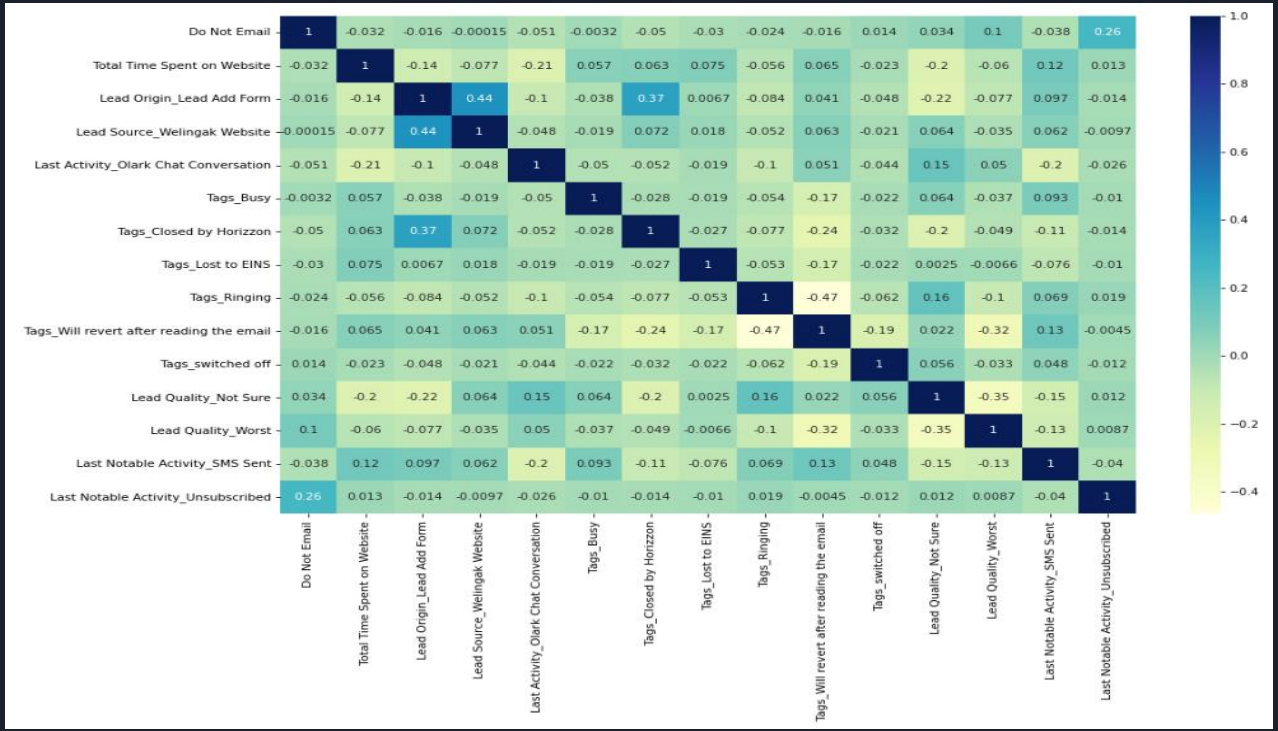
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6075
Model:	GLM	Df Residuals:	6059
Model Family:	Binomial	Df Model:	15
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1332.4
Date:	Tue, 18 Oct 2022	Deviance:	2664.7
Time:	17:45:24	Pearson chi2:	2.02e+04
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.8791	0.257	-11.194	0.000	-3.383	-2.375
Do Not Email	-1.4278	0.241	-5.914	0.000	-1.901	-0.955
Total Time Spent on Website	3.2143	0.211	15.201	0.000	2.800	3.629
Lead Origin_Lead Add Form	1.6422	0.310	5.293	0.000	1.034	2.250
Lead Source_Welingak Website	3.9826	1.060	3.757	0.000	1.905	6.060
Last Activity_Olark Chat Conversation	-1.3718	0.223	-6.147	0.000	-1.809	-0.934
Tags_Busy	3.8096	0.350	10.884	0.000	3.124	4.496
Tags_Closed by Horizzon	7.7226	0.786	9.824	0.000	6.182	9.263
Tags_Lost to EINS	8.7797	0.660	13.302	0.000	7.486	10.073
Tags_Ringing	-1.1891	0.350	-3.397	0.001	-1.875	-0.503
Tags_Will revert after reading the email	4.0649	0.262	15.518	0.000	3.552	4.578
Tags_switched off	-2.0816	0.681	-3.056	0.002	-3.417	-0.746
Lead Quality_Not Sure	-3.2070	0.135	-23.803	0.000	-3.471	-2.943
Lead Quality_Worst	-3.4657	0.851	-4.072	0.000	-5.134	-1.798
Last Notable Activity_SMS Sent	2.5165	0.129	19.487	0.000	2.263	2.770
Last Notable Activity_Unsubscribed	1.9940	0.660	3.021	0.003	0.700	3.288

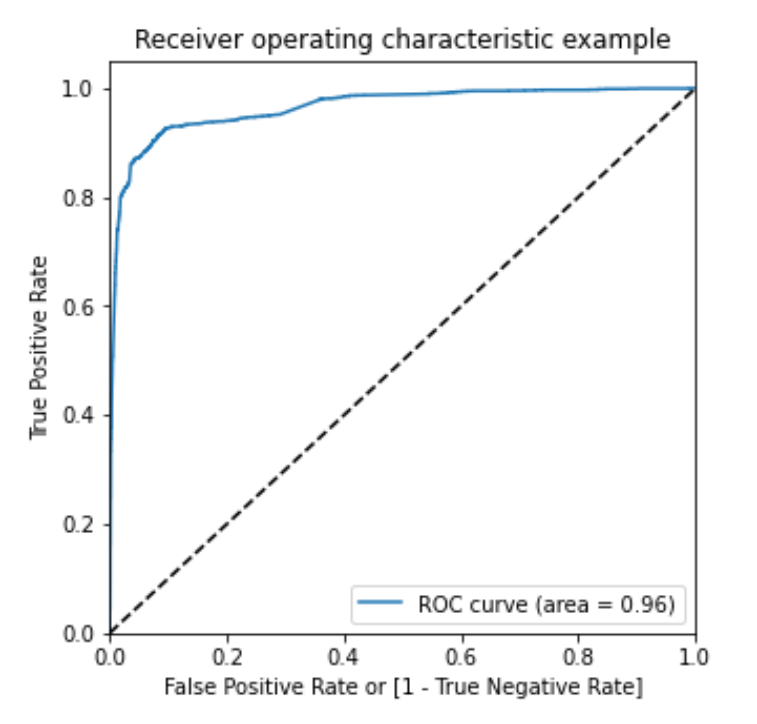
- All P- values are zero.

Heatmap



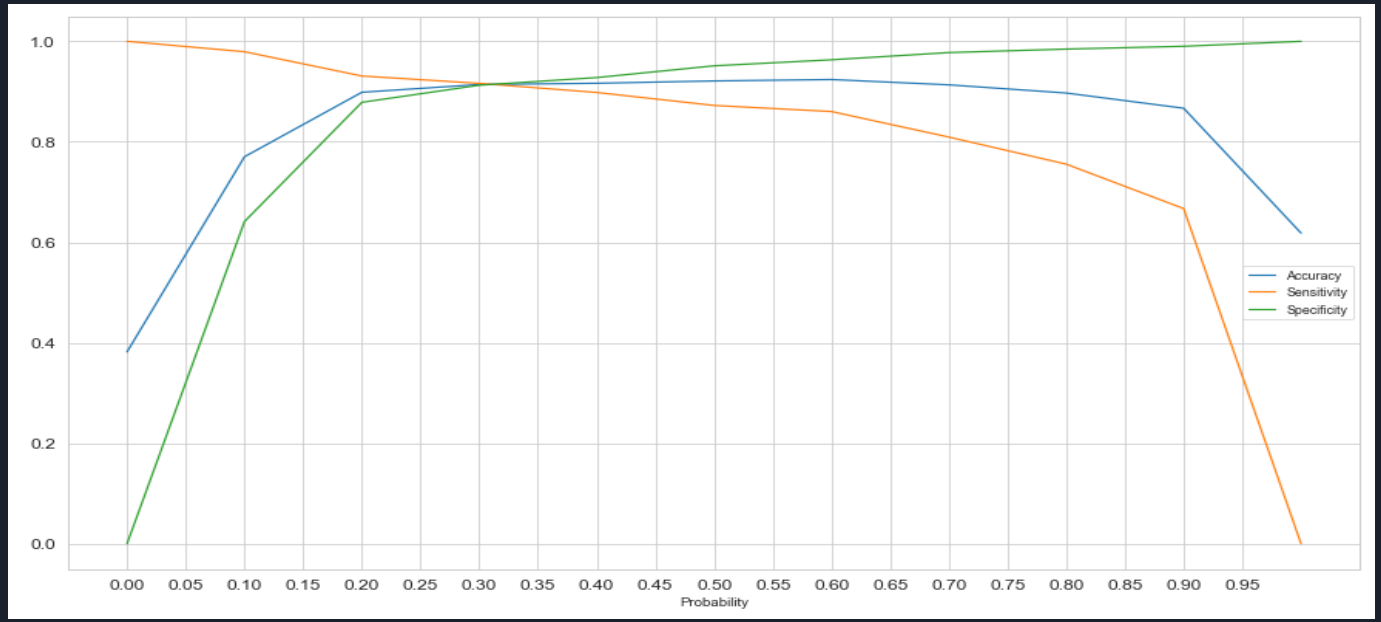
- Correlations between features in the final model are negligible.

ROC curve



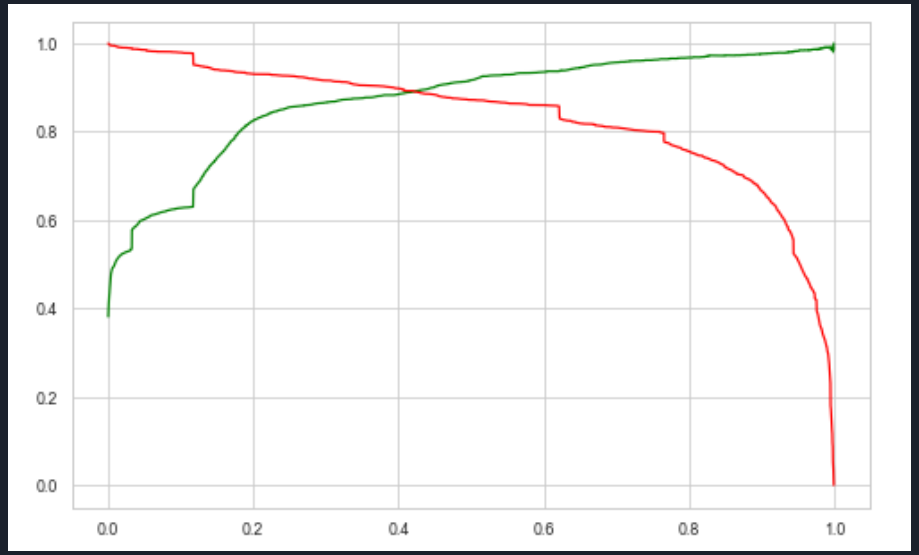
- Area under curve = 0.96

Finding Optimal Threshold



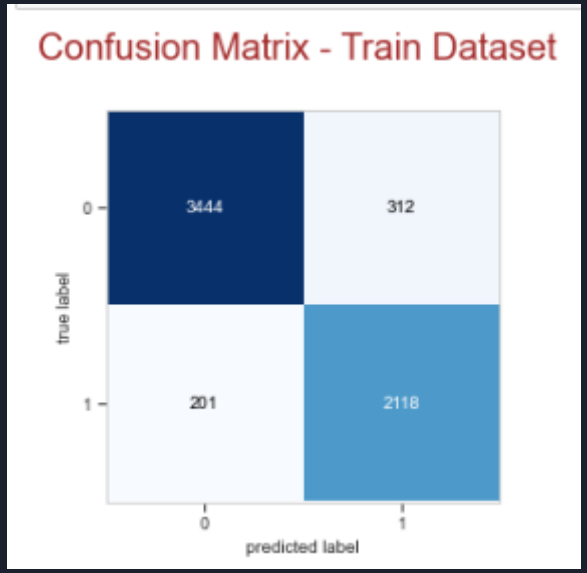
- Graph showing changes in Sensitivity, Specificity and Accuracy with changes in the probability threshold values.
- optimal Probability cutoff = 0.32

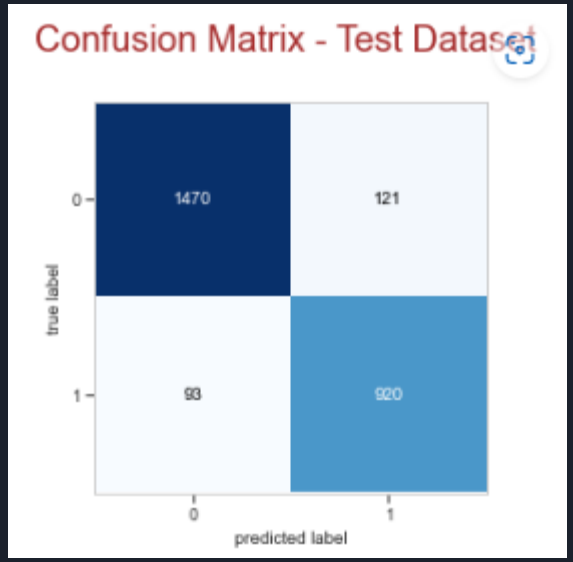
Precision and Recall trade-off



- From the precision-recall graph above, we get the optical threshold value as close to 0.41

Confusion Matrix





- For train set
- For test set

Final Results

Data	Train Set	Test Set
Accuracy	0.9155	0.9178
Sensitivity	0.9133	0.9081
Specificity	0.9169	0.9239
False Positive Rate	0.083	0.076
Positive Predictive Value	0.871	0.8837
Negative Predictive Value	0.944	0.9404
AUC	0.96	0.96

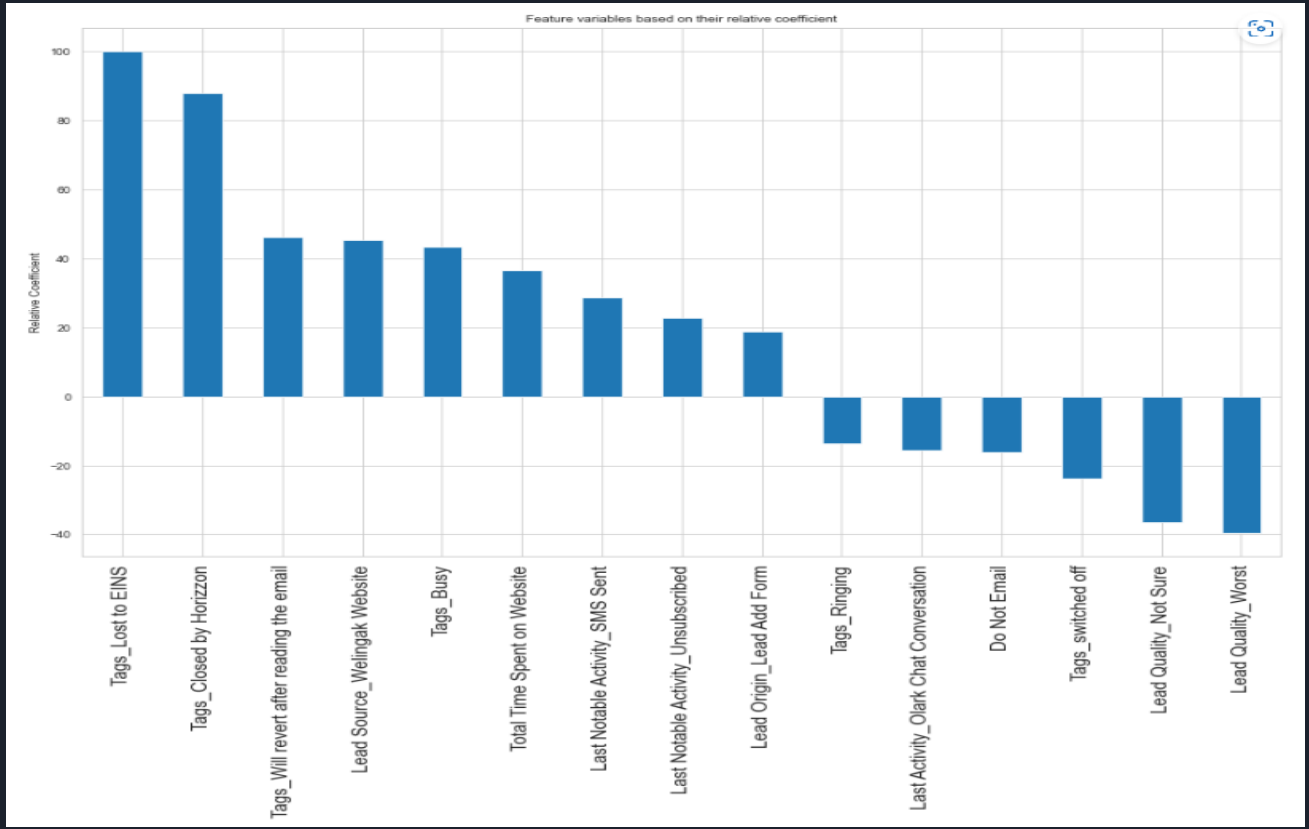
Train Set Classification Report

	precision	recall	f1-score	support
0	0.94	0.92	0.93	3756
1	0.87	0.91	0.89	2319
accuracy			0.92	6075
macro avg	0.91	0.92	0.91	6075
weighted avg	0.92	0.92	0.92	6075

Test Set Classification Report

	precision	recall	f1-score	support
0	0.94	0.92	0.93	1591
1	0.88	0.91	0.90	1013
accuracy			0.92	2604
macro avg	0.91	0.92	0.91	2604
weighted avg	0.92	0.92	0.92	2604

Relative Importance of Features



INFERENCES

Feature Importance

- Three variables which contribute most towards the probability of lead conversion in decreasing order of impact are:
  - Tags_Lost to EINS
  - Tags_Closed by Horizon
  - Tags_Will revert after reading the email

- These are dummy features created from the categorical variable Tags.
- All three contribute positively towards the probability of lead conversion.
- These results indicate that the company should focus more on the leads with these three tags.

Case 1:
The company has interns for 2 months. They wish to make the lead conversion more aggressive. They want almost all of the potential leads to be converted and hence, want to make phone calls to as many of such people as possible.

Solution:

- $Sensitivity = True\ Positives / (True\ Positives + False\ Negatives)$
- *Sensitivity can be defined as the number of actual conversions predicted correctly out of the total number of actual conversions. As we saw earlier, sensitivity decreases as the threshold increases.*
- *High sensitivity implies that our model will correctly predict almost all leads who are likely to convert. At the same time, it may overestimate and misclassify some of the non-conversions as conversions.*
- *As the company has extra manpower for two months and wants to make the lead conversion more aggressive, it is a good strategy to go for high sensitivity. To achieve high sensitivity, we need to choose a low threshold value.*

Case 2:

At times, the company reaches its target for a quarter before the deadline. It wants the sales team to focus on some new work. So during this time, the company's aim is to not make phone calls unless it's extremely necessary.

Solution:

- $Specificity = True\ Negatives / (True\ Negatives + False\ Positives)$
- *Specificity can be defined as the number of actual non-conversions predicted correctly out of a total number of actual non-conversions. It increases as the threshold increases.*
- *High specificity implies that our model will correctly predict almost all leads who are not likely to convert. At the same time, it may misclassify some of the conversions as non-conversions.*
- *As the company has already reached its target for a quarter and doesn't want to*
- *make unnecessary phone calls, it is a good strategy to go for high specificity.*
- *It will ensure that the phone calls are only made to customers who have a very high probability of conversion. To achieve high specificity, we need to choose a high threshold value.*

Recommendations

- *By referring to the data visualizations, focus on*
  - *Increasing the conversion rates for the categories generating more leads and*
  - *Generating more leads for categories having high conversion rates.*
- *Pay attention to the relative importance of the features in the model and their positive or negative impact on the probability of conversion.*
- *Based on varying business needs, modify the probability threshold value for identifying potential leads.*

THANK YOU