Lead Scoring Case Study

Presented By –
Kaivalya Degaonkar
Khushi Somaiya
Kritik Sahu

Problem Statement

X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Approach

We have build this model using Logistic regression along with RFE, to get top features and based on that we have provided recommendations to the company.

Steps Followed

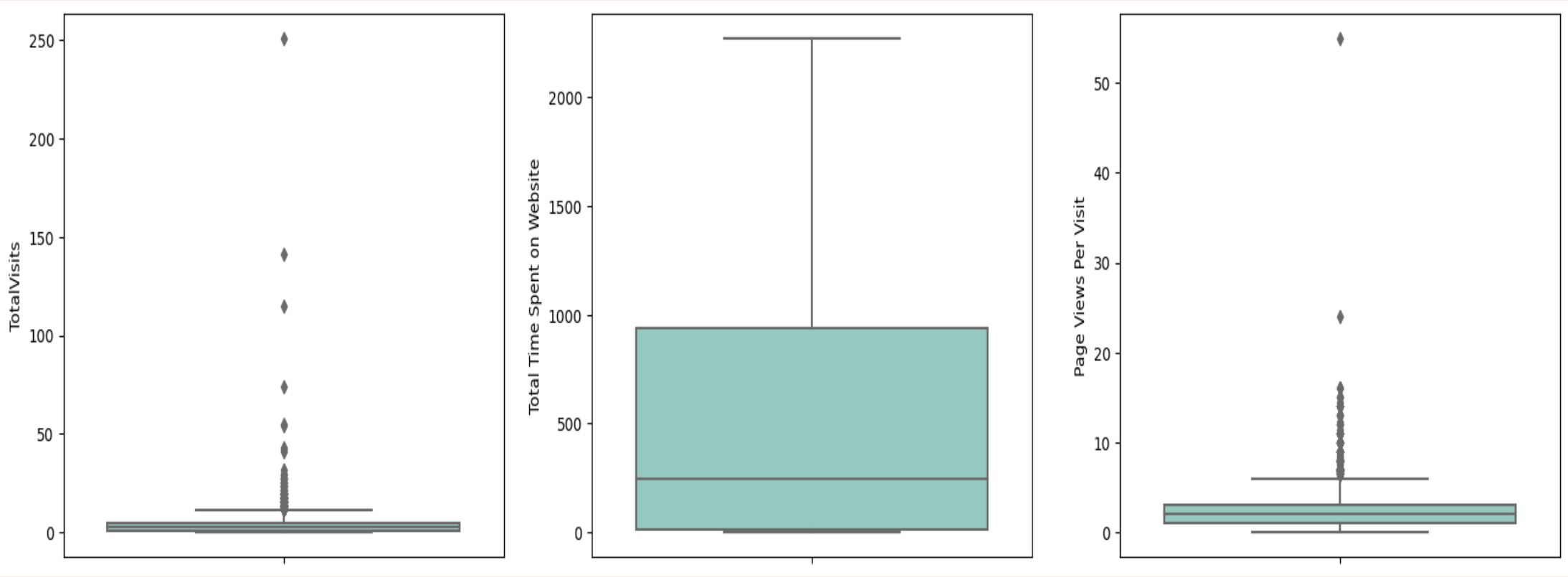
1. Data Understanding and EDA
2. Data Preparation and Test-Train Split
3. Model Building
4. Model Evaluation

Data Understanding and EDA

- In the given dataset there were total of **9240 records** with **37 attributes.**
- Data contains high number of missing values which we have handled by dropping columns with more then 30% missing data.
- Imputed data with mode or median according to the need.
- Other activities like outliers' treatment, fixing invalid data, handling class imbalance were carried out.
- About **98%** of data was retained.
- Finally we have cut down to **9029 records** and **13 attributes.**

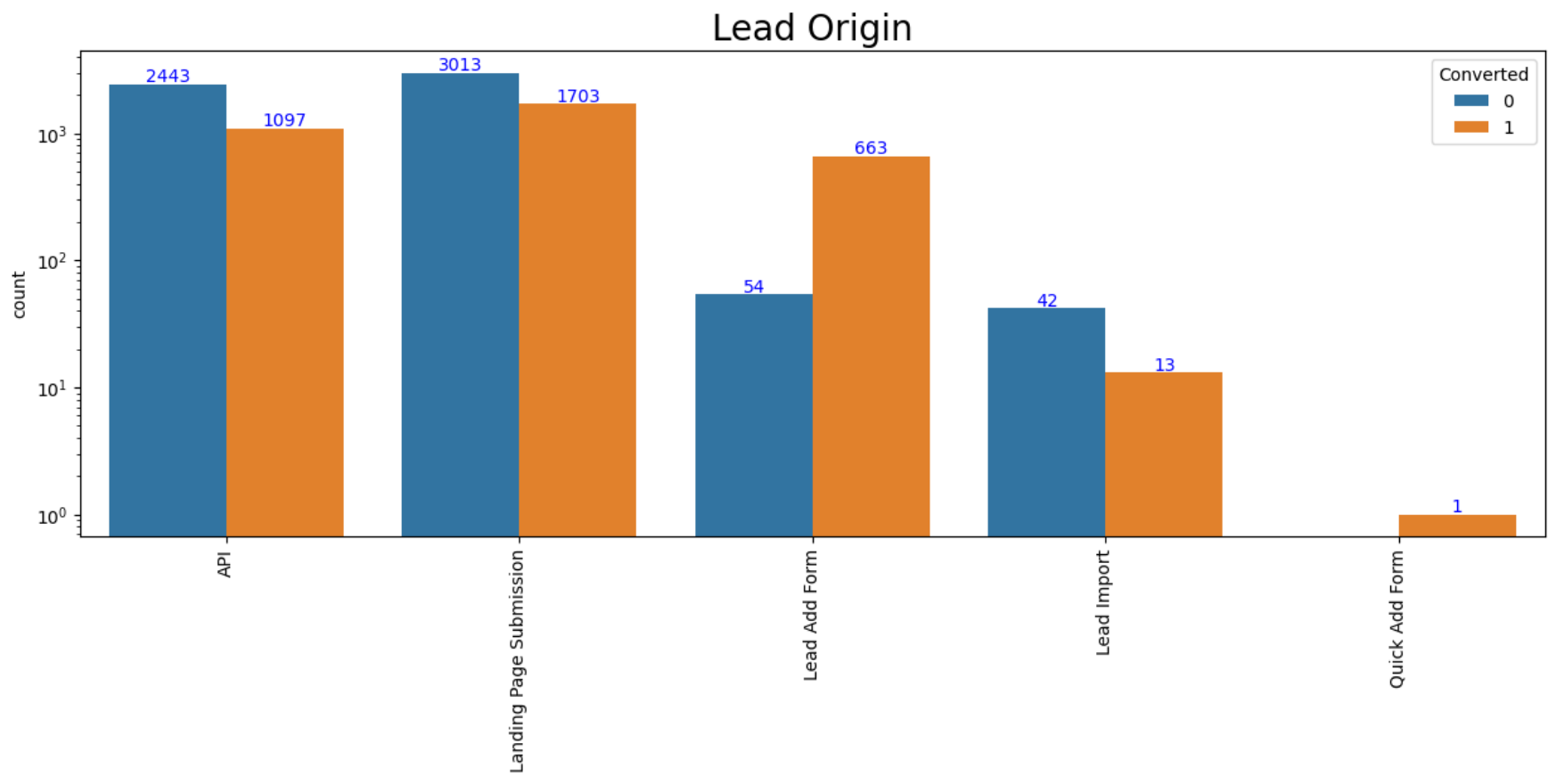
Outlier Analysis

- We did some univariate analysis and then outlier treatment these were some potential outliers we did capping of 99%



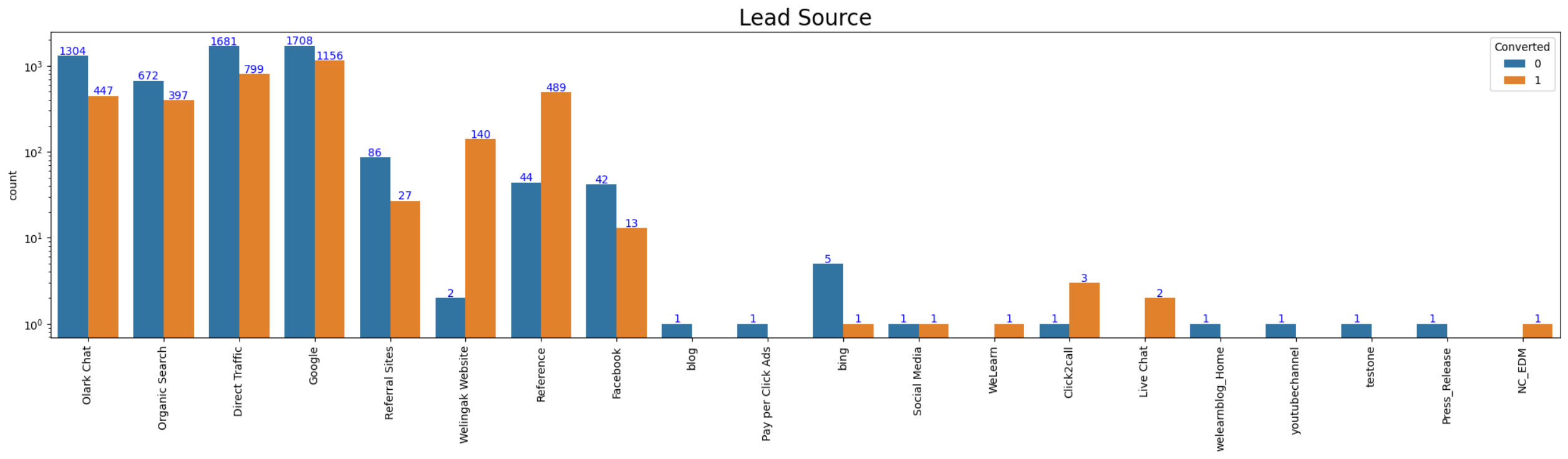
Lead Origin vs. Converted

- API and Landing Page Submission bring higher number of leads as well as conversion.
- Lead Add Form has a very high conversion rate but count of leads are not very high.



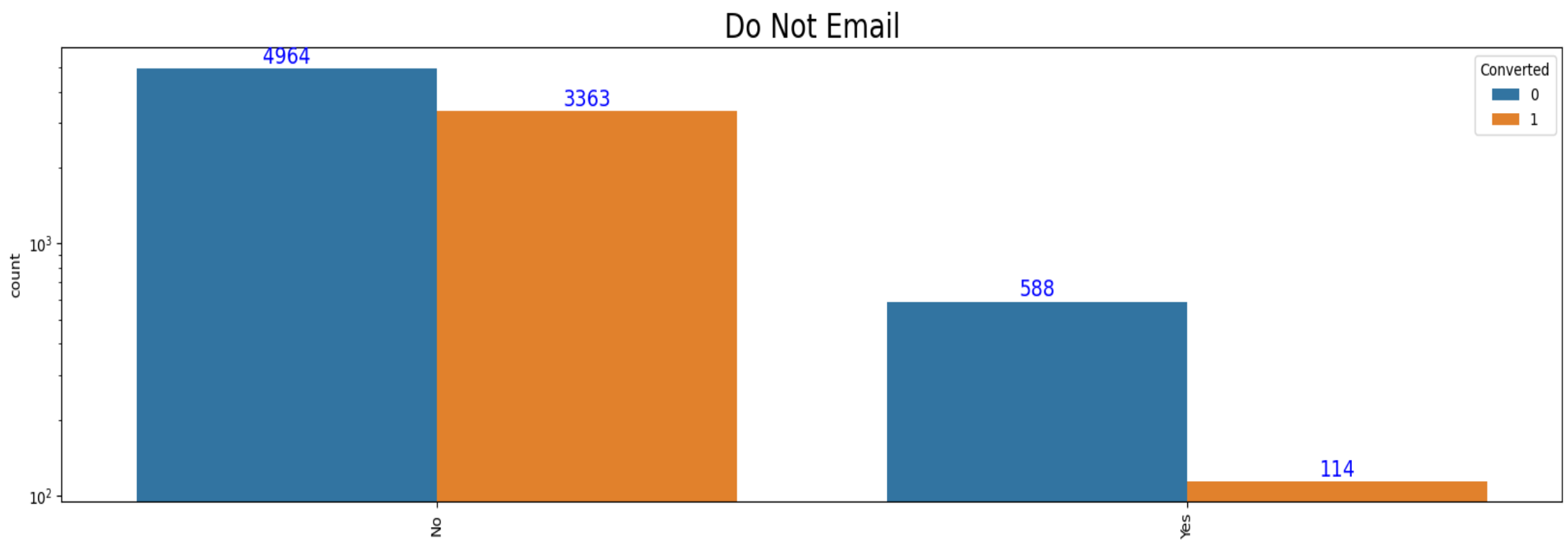
Lead Source vs. Converted

- Google and Direct traffic generates maximum number of leads.
- Conversion rate of 'Reference' and 'Welingak Website' leads is high.



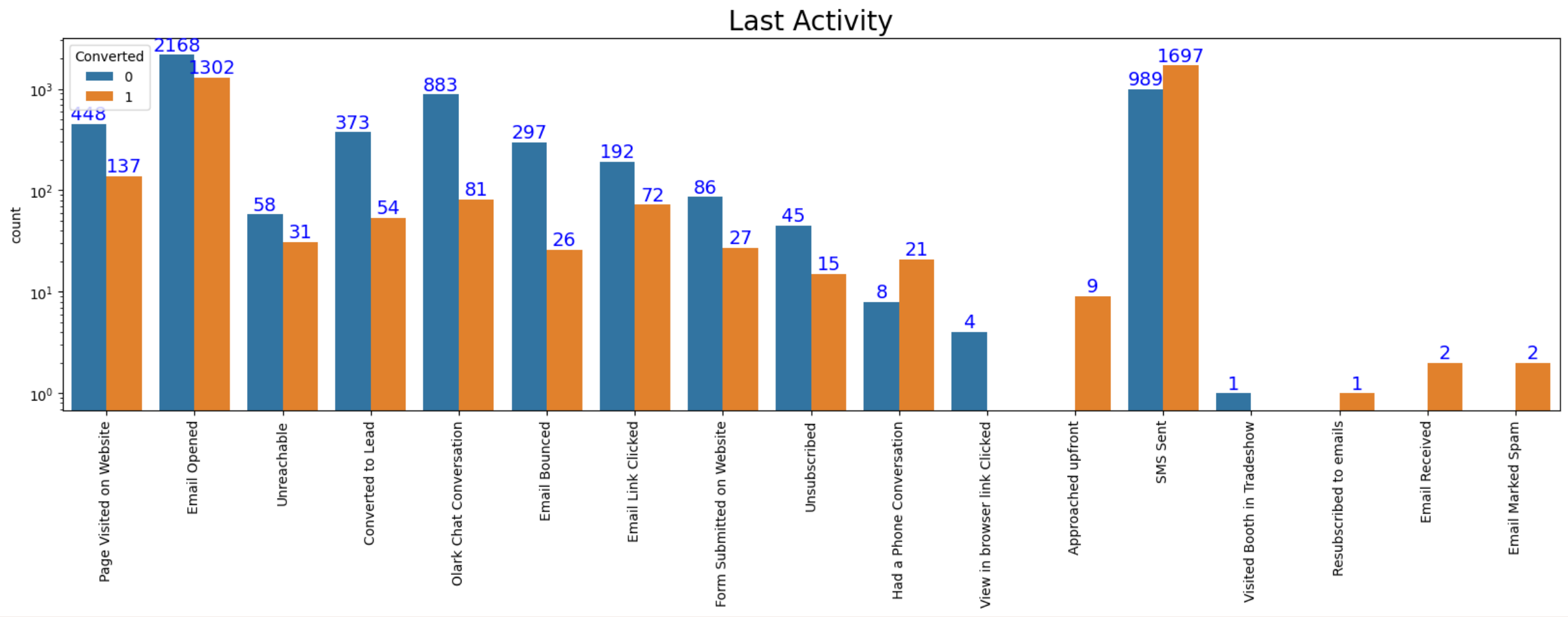
Do Not Email vs. Converted

- People who opted for mail option are becoming more leads.



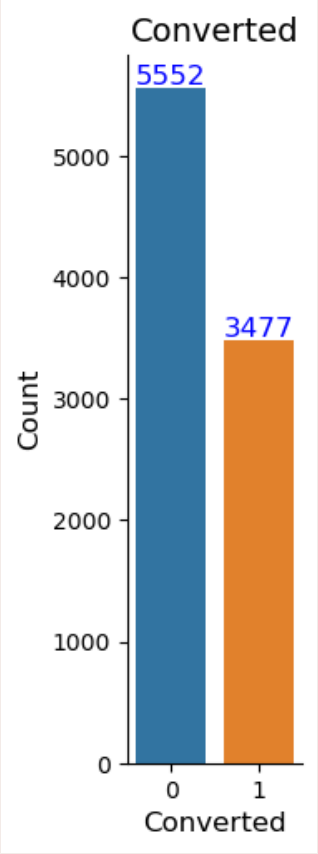
Lead Activity vs. Converted

- Conversion rate for last activity of 'SMS Sent' is ~63%.
- Highest last activity of leads is 'Email Opened' .



Target Variable Class Imbalance Check

- There doesn't seem to be data imbalance in the 'Target Variable'.



Data Preparation and Test-Train Split

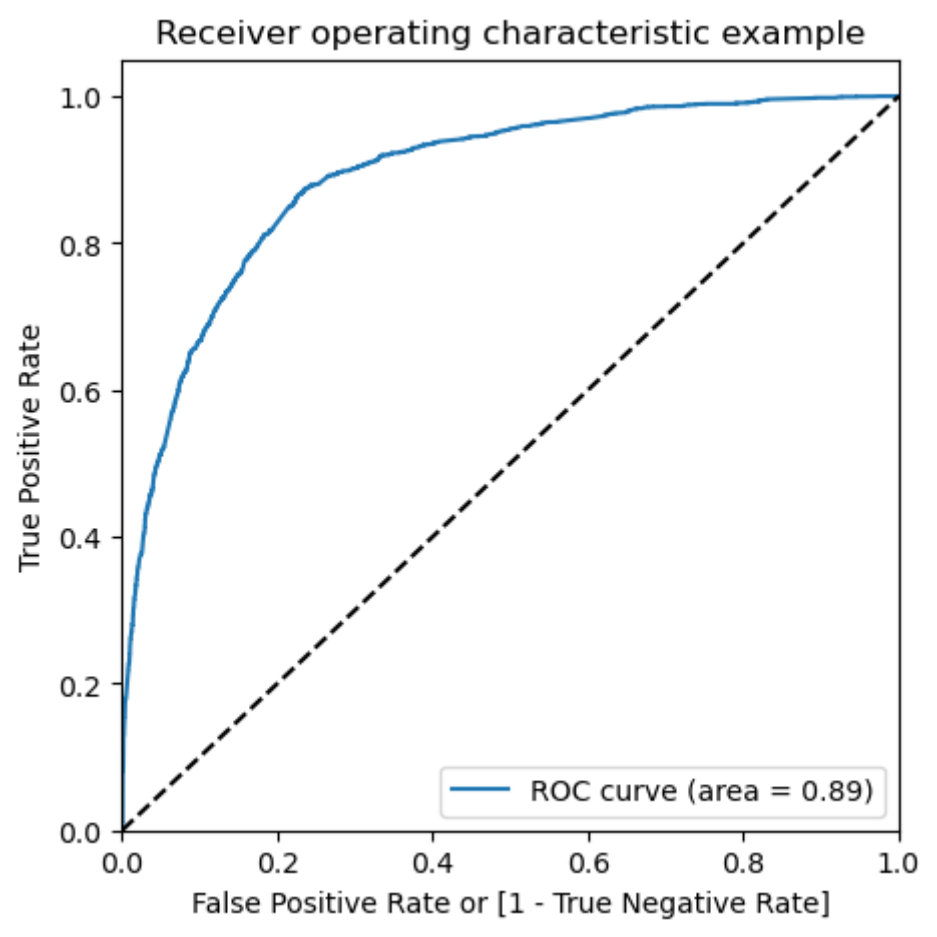
- We proceeded with encoding the categorical data with multiple levels into dummy variables.
- We also did binary mapping for Yes/No columns.
- Splitting Train & Test Sets: 70:30 ratio.
- Feature Scaling using Standardization.

Model Building

- The training data is fed into a Generalized Linear Model (GLM) to build a logistic regression model.
- The ineffective variables are eliminated using RFE and VIF.
- Total 4 models were built before reaching final Model 5 which was stable with (p-values < 0.05). No sign of multicollinearity with VIF < 5.
- lrm5 was selected as final model which was used for making prediction on train and test set.

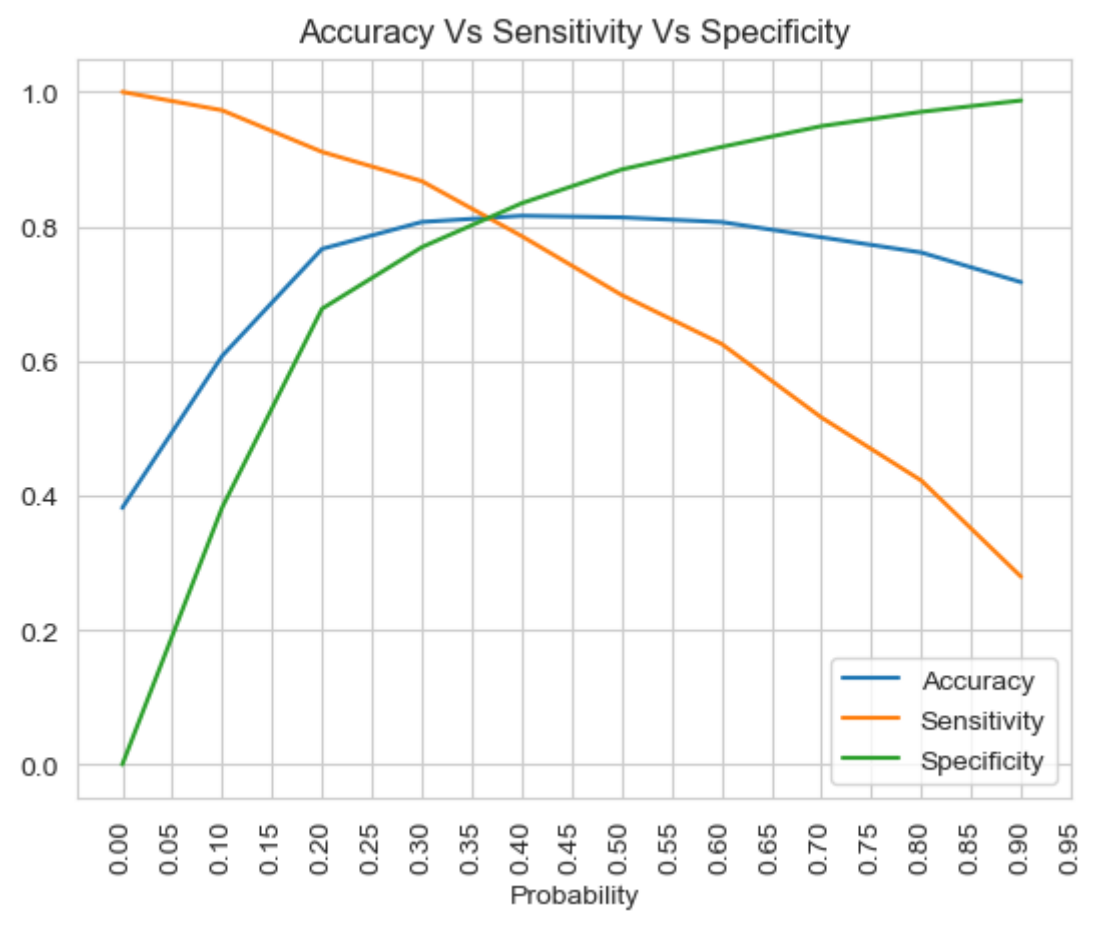
Model Evaluation

- ROC Curve: The ROC Curve area value should be close to 1. We have 0.89 which is a good value.



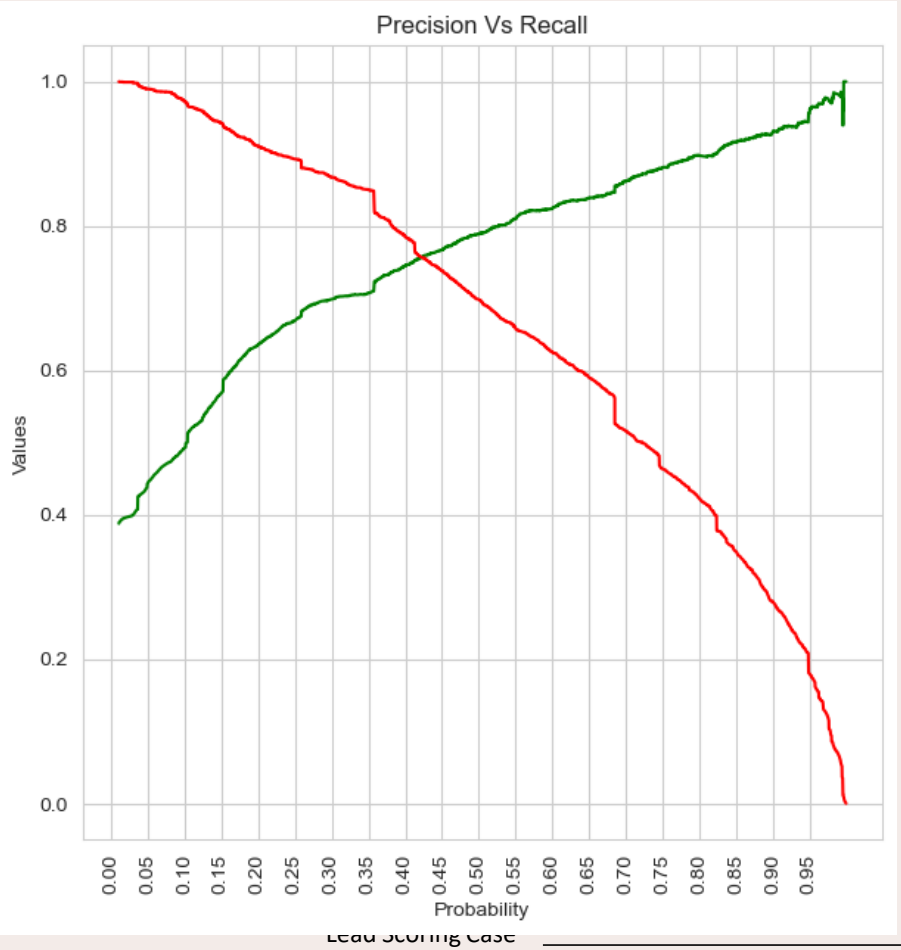
Model Evaluation

- Accuracy vs. Sensitivity vs. Specificity trade off: Prediction was done on the test data frame and with an optimum cut off of 0.37 and accuracy, specificity and sensitivity, all came around 81%.



Model Evaluation

- Precision vs. Recall trade off: Precision – recall method was also used to recheck and a cut off of 0.42 was found with Precision around 75% and recall around 75% on the test data frame.
- At the 0.42 cut off, accuracy came as 81.37.



Model Evaluation

- After Observing both 0.37 and 0.42 Cutoffs: 0.37 Gives a bit higher accuracy score.
- All the metrics seems good.
- Also, false positive rate is lower for 0.37 cutoff which will help in reduction of falsely predictions.

Training Data Set		Vs	Test Data Set	
Sensitivity is	: 81.17	|	Sensitivity is	: 81.14
Specificity is	: 81.53	|	Specificity is	: 80.95
True Positive Rate is	: 81.17	|	True Positive Rate is	: 81.14
False Positive Rate is	: 18.47	|	False Positive Rate is	: 19.05
Precision is	: 73.05	|	Precision is	: 73.43
Recall is	: 81.17	|	Recall is	: 81.14
Accuracy score is	: 81.39	|	Accuracy score is	: 81.03

- The difference b/w train and test data's performance metrics is under 2%. This means that the final model did not overfit training data and is performing well.
- High Sensitivity will make sure that all possible leads who are likely to convert are correctly predicted, whereas high Specificity will ensure that the leads that are on the brink of the probability of getting converted or not are not selected.
- Based on the business requirement, we can increase or decrease the probability threshold value which in turn will decrease or increase the Sensitivity and increase or decrease the Specificity of the model as required.

Top 3 Factors which can help in generating more successful leads:

- Lead Origin_Lead Add Form
- Lead Source_Welingak Website
- What is your current occupation_Working Professional

Recommendations

- Lead Add Format should be focused on more to generate more leads.
- More budget/spend can be done on Welingak Website in terms of advertising, etc.
- Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay fees.
- If the X Education has more man power for 2 months so, they can get in contact with more customers than before to try increasing the conversion rate.
  To increase the conversion rate which in turn means to increase sensitivity, we can lower the threshold value in our model. The table in next slide can be referred to know the values of sensitivity based on different threshold(probability) values for our model.
- If the company has already achieved its target for a quarter before the deadline and wants to avoid unnecessary calls, we can try increasing the value of specificity.
  Increasing specificity may increase the chances of misinterpretation of possible conversions as non-conversions but this way the sales team can focus on only extremely necessary phone calls. The sales team can use the saved time to work for their new tasks.
  To increase the value of specificity, we can try increasing the threshold value. The table in next slide can be referred to check for values of different specificity based on different threshold(probability) values for our model.

'Accuracy', 'Sensitivity' and 'Specificity' for various probability cutoffs

Probability	Accuracy	Sensitivity	Specificity
0.0	0.381487	1.000000	0.000000
0.1	0.607437	0.972625	0.382195
0.2	0.766614	0.910825	0.677667
0.3	0.806646	0.867275	0.769250
0.4	0.815823	0.785566	0.834485
0.5	0.813449	0.698051	0.884625
0.6	0.806487	0.625467	0.918138
0.7	0.783861	0.515968	0.949092
0.8	0.761234	0.422231	0.970325
0.9	0.717089	0.279137	0.987209

Thank you