LEAD SCORING CASE STUDY REPORT

HOANG NGOC TIEN – MASTER OF DATA IN DATA SCIENCE PROGRAME

PROBLEM & BUSINESS GOALS

X Education is and education company who sells online courses to industry professionals. They acquires leads through marketing campaign on several websites and past referrals, then sale team will contact leads by making calls, writing email, etc. Problem is that the conversion rate of this process is very poor (around 30%). The company wants to make the process more efficient by identifying the most potential leads (hot leads) to focus more on communicating.

BUSINESS GOALS

Using lead dataset provided by X Education Company, build a logistic regression model to predict hot leads by assigning a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

The CEO wants the lead conversion rate to be around 80%

THE APPROACH

1. Data Understanding and Preparation

- Get familiar with the dataset's structure, variables and their meaning, target variable
- Handle missing values: Impute missing values using appropriate methods.

2. Exploratory Data Analysis (EDA)

- Visualize the distribution of the target variable: Understand the proportion of converted vs. non-converted leads
- Examine distribution of features: Analyze the distributions of numerical and categorical features with respect to converting status.
- Identify outliers and anomalies.
- Explore correlations: Check correlations between variables in each target variable’s segment.
- Identify trends: Observe any patterns or trends that might be associated with lead conversion.

THE APPROACH

3. Data Preparation for Model Building

- Treating categorical variables
- Splitting the Data into Train and Test sets.
- Scaling the Data

4. Building Model and Making Predictions

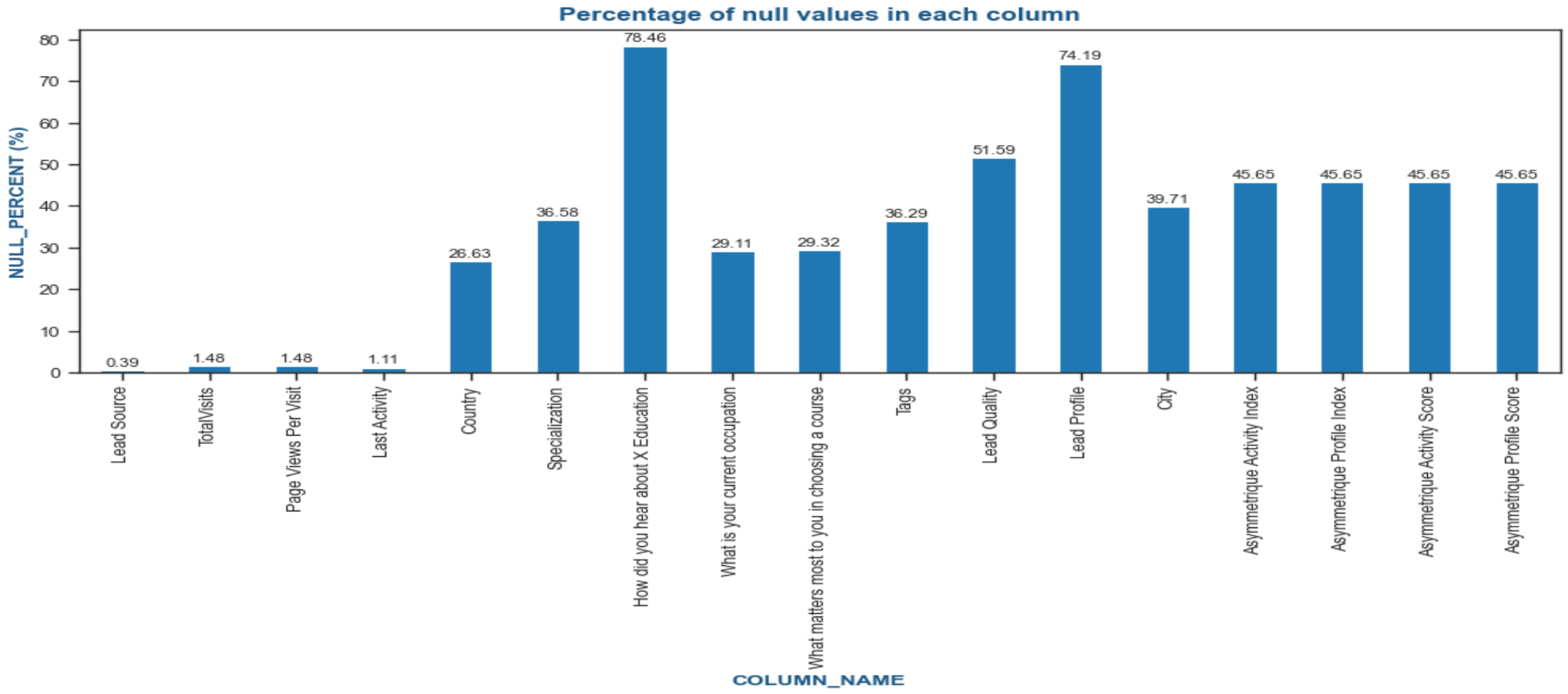
- Building logistic regression model
- Making predictions.
- Lead Scoring

5. Model Evaluation & Interpretation

- Evaluating model.
- Final model Interpretation

1. DATA UNDERSTANDING AND PREPARATION

Handling missing values



1. DATA UNDERSTANDING AND PREPARATION

Handling missing values:

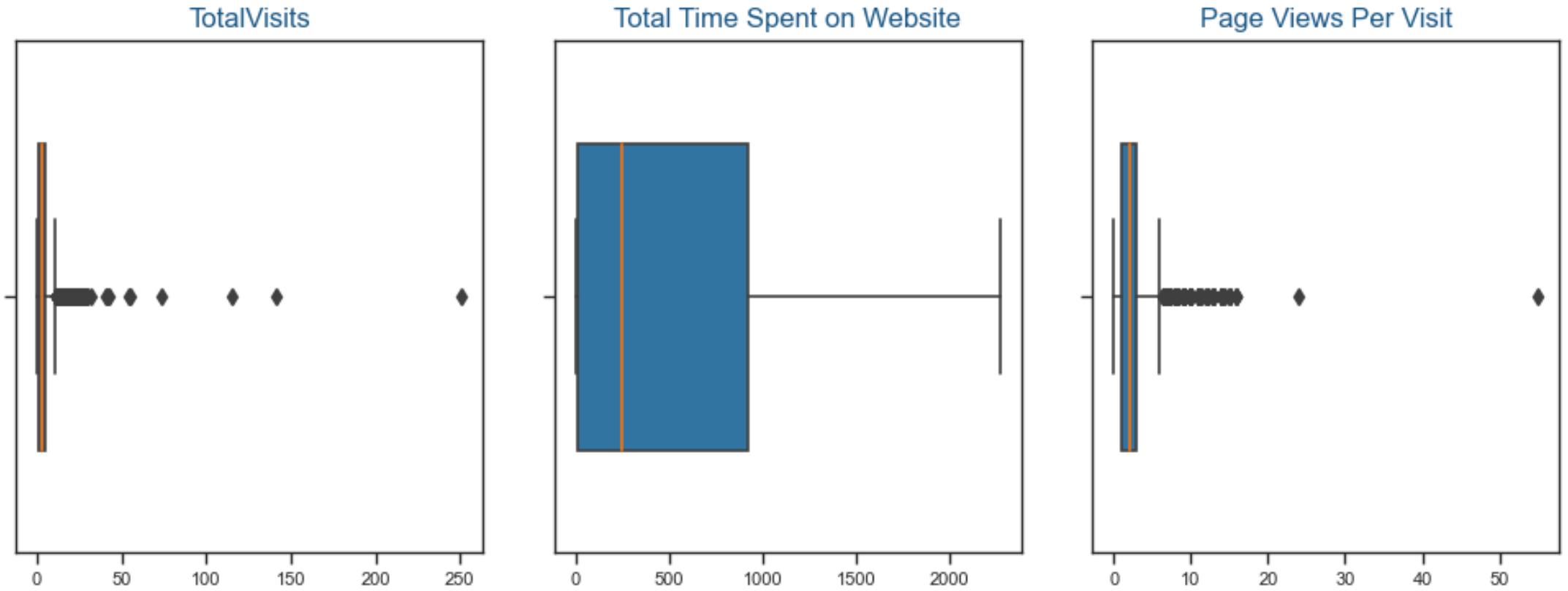
- Drop columns which have >40% null values
- Handling Methods with other columns as below:

COLUMN_NAME	NULL_PERCENT	VARIABLE_TYPE	HANDLING_METHOD
Lead Source	0.39	categorical	remove data
TotalVisits	1.48	numerical	replace with median value
Page Views Per Visit	1.48	numerical	replace with median value
Last Activity	1.11	categorical	remove data
Country	26.63	categorical	replace with mode value
Specialization	36.58	categorical	replace with 'Others'
What is your current occupation	29.11	categorical	replace with mode value
What matters most to you in choosing a course	29.32	categorical	replace with mode value
Tags	36.29	categorical	replace with 'Others'
City	39.71	categorical	replace with mode value

2. EXPLORATORY DATA ANALYSIS

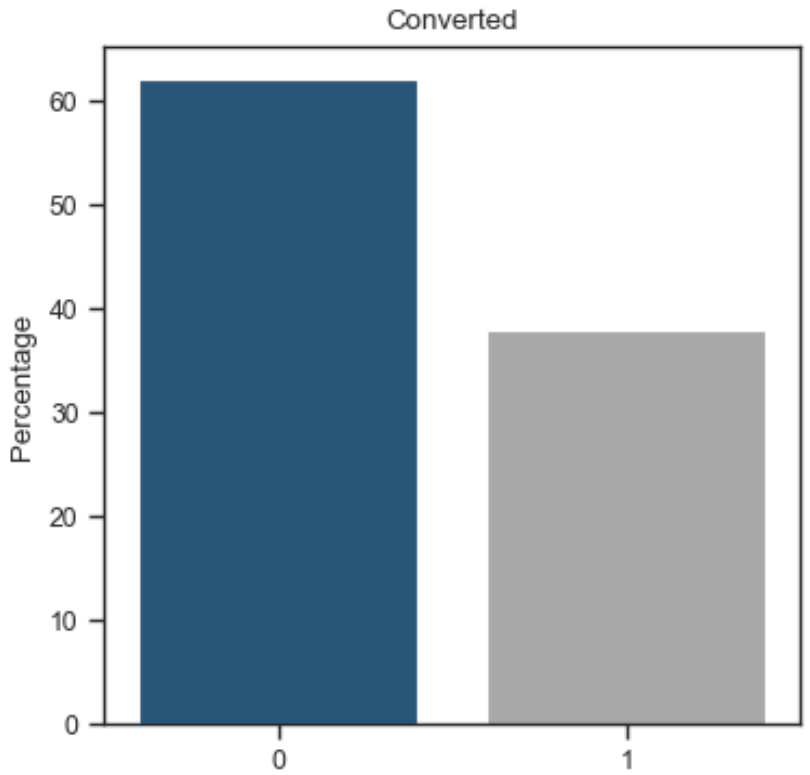
- Outliers
- Data imbalance
- Examine distribution of features: Analyze the distributions of numerical and categorical features with respect to default status.
- Explore correlations: Check correlations between features and the target variable as well as among features themselves.
- Identify trends: Observe any patterns or trends that might be associated with default behavior.

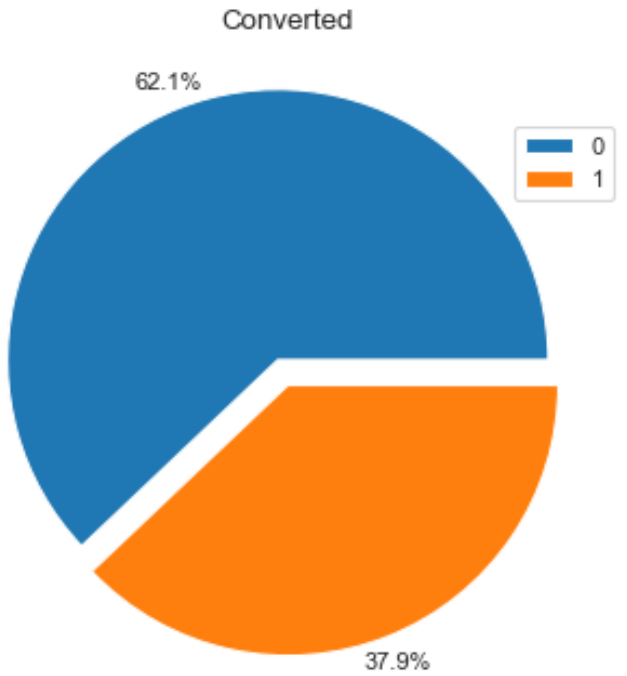
EDA - OUTLIER ANALYSIS



- Some columns contain values much bigger than 95th percentile values => outliers
- We can replace outliers with median value

EDA - DATA IMBALANCE



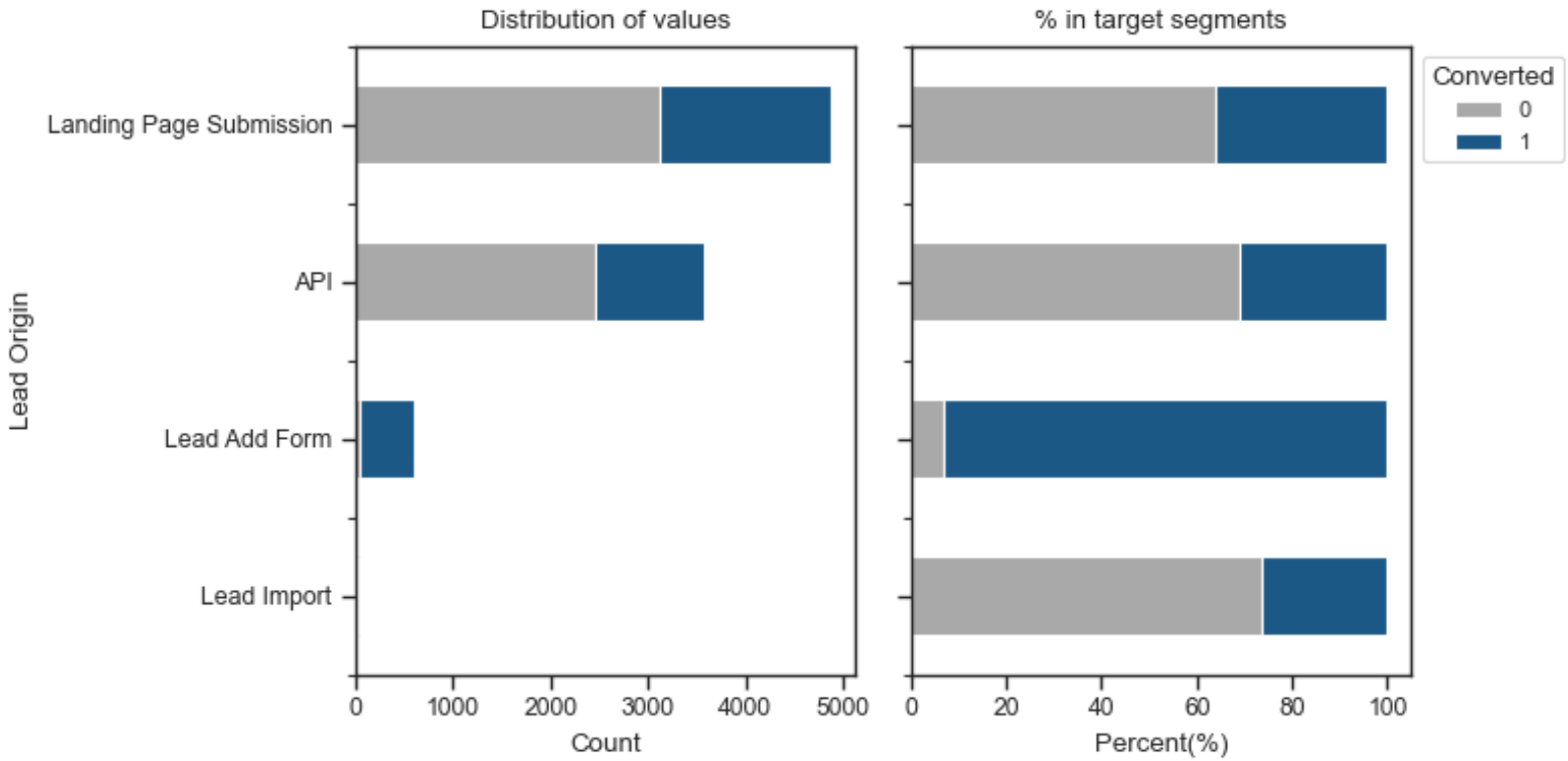


Target Variable:

- 37.9% of observations are converted leads
- The data is relatively balanced and good enough to build model

EDA – INSIGHTS OF DATA

Lead Origin

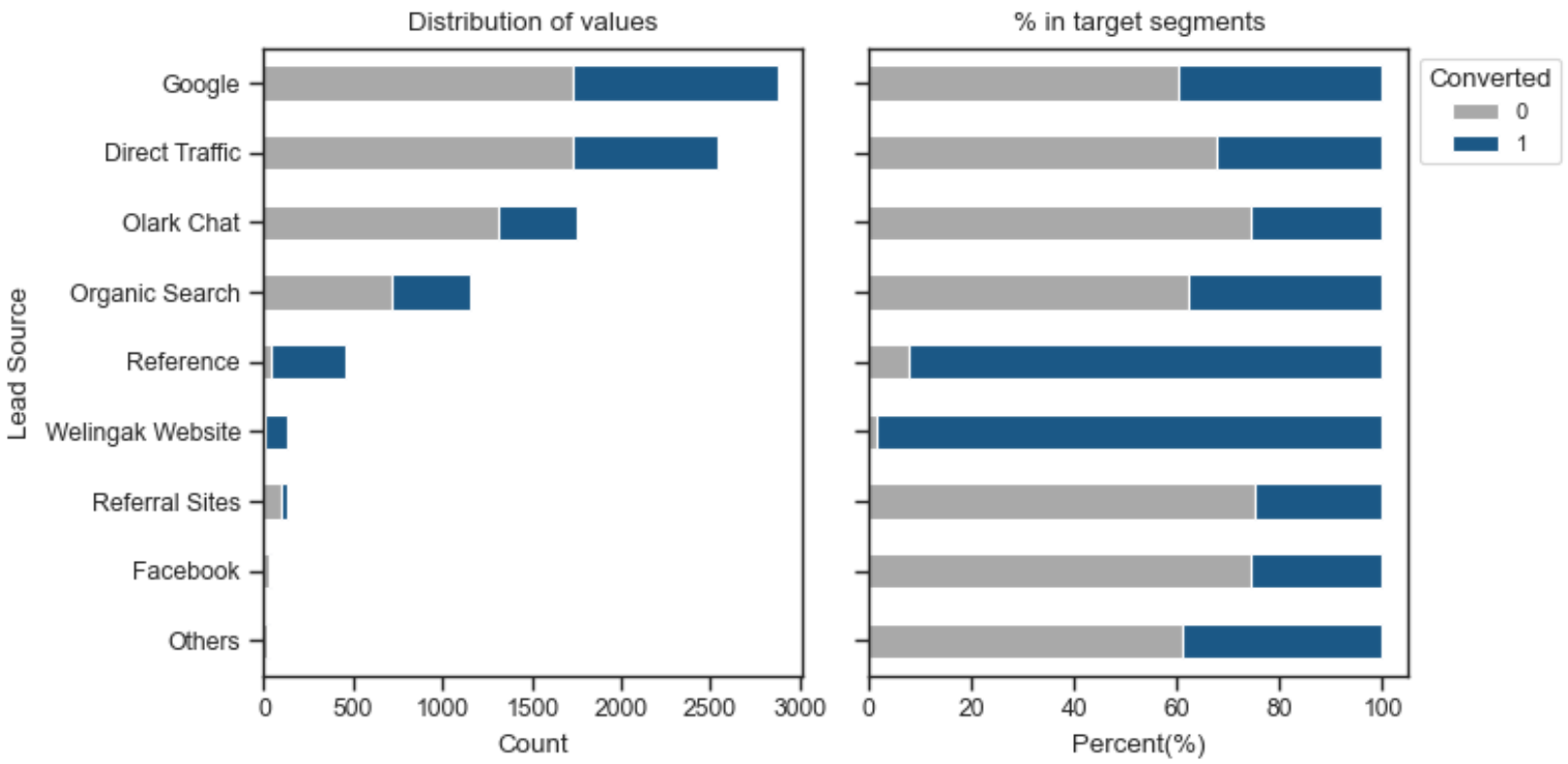


Inference:

- Most of leads come from Landing Page Submission and API
- Leads from Lead Add Form are much more likely to be hot leads

EDA – INSIGHTS OF DATA

Lead Source

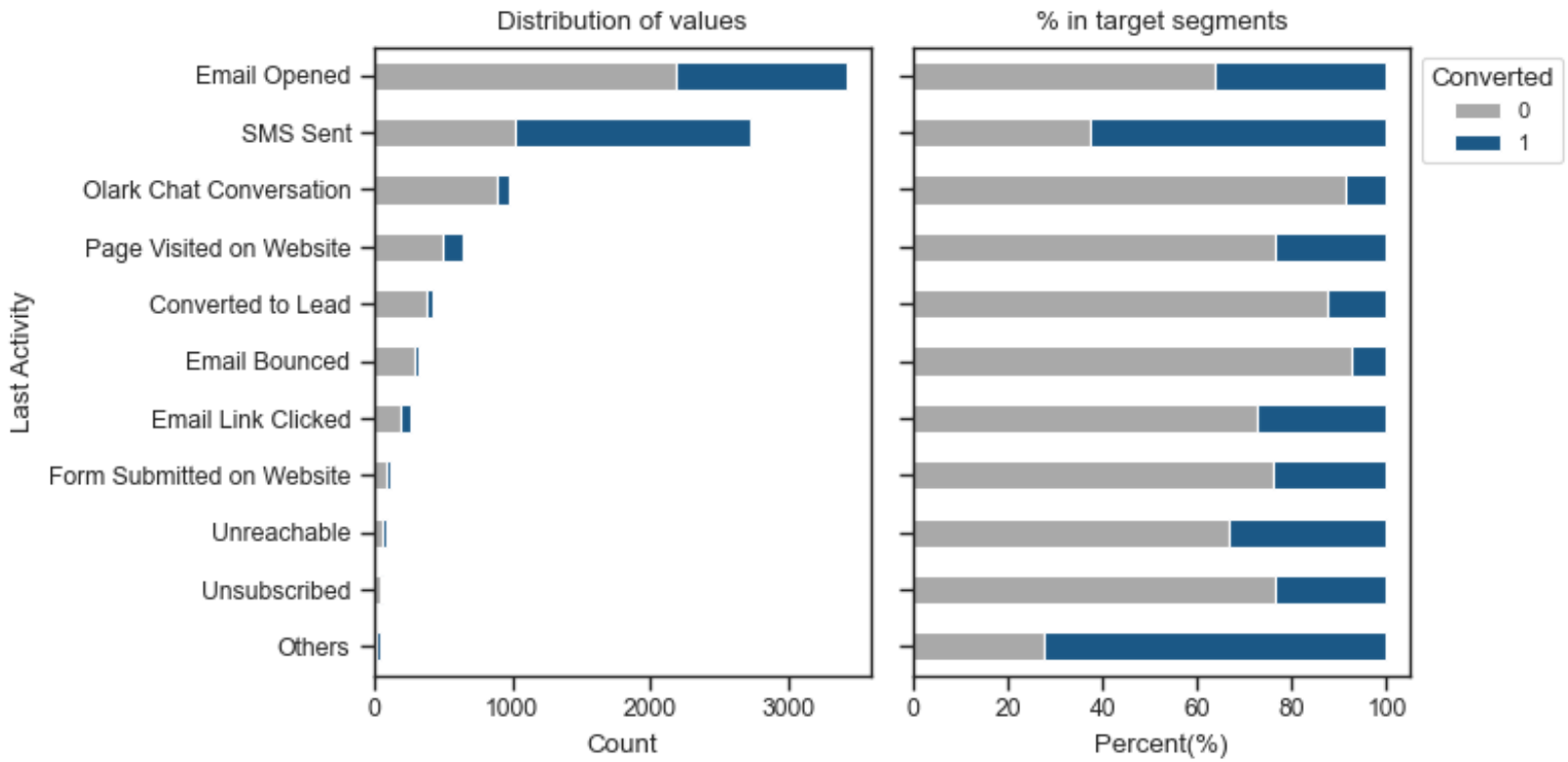


Inference:

- Most of leads come from Google and Direct Traffic
- Leads from Welingak Website and Reference are much more likely to be hot leads

EDA – INSIGHTS OF DATA

Last Activity

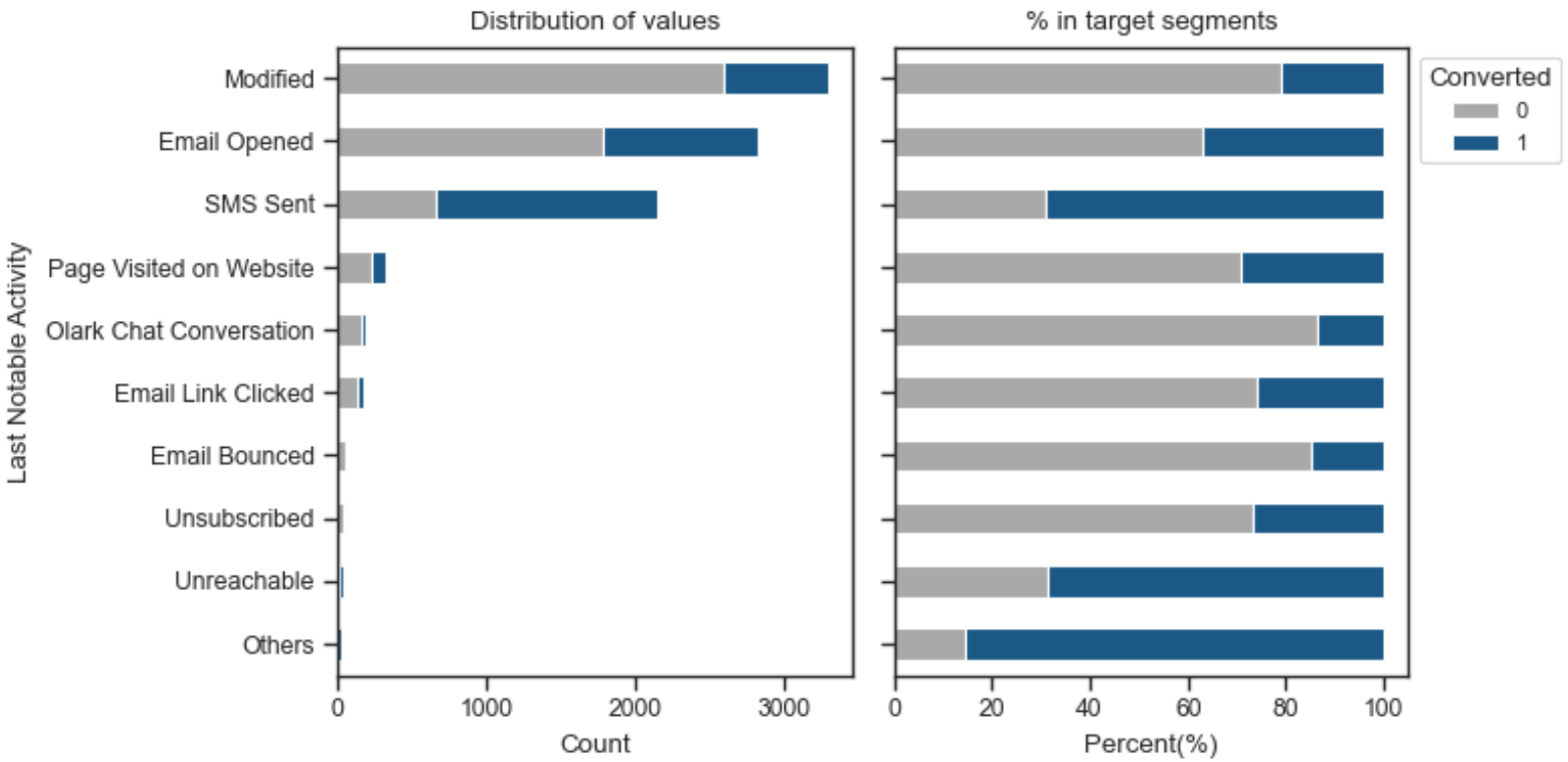


Inference:

- Most of leads have opened email as last activity.
- But Leads with last activity is SMS Sent are more likely to be hot leads

EDA – INSIGHTS OF DATA

Last Notable Activity

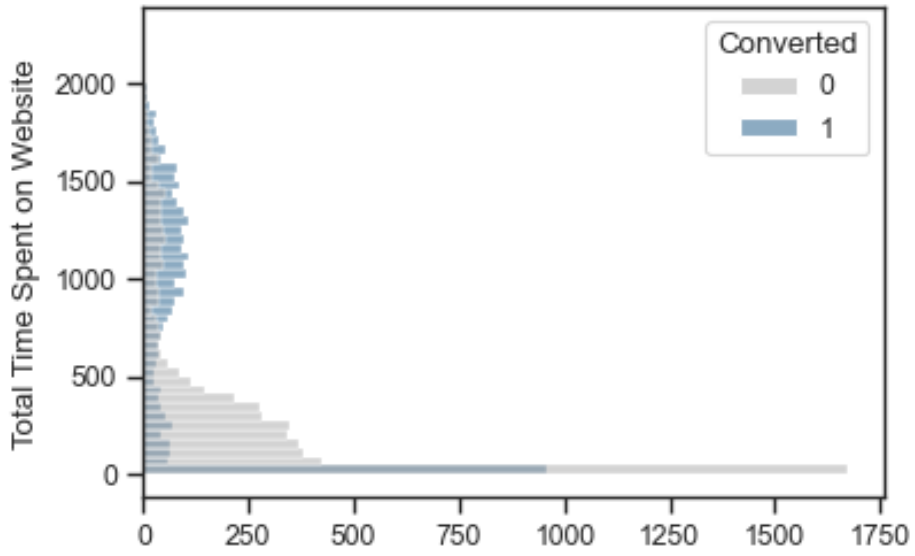


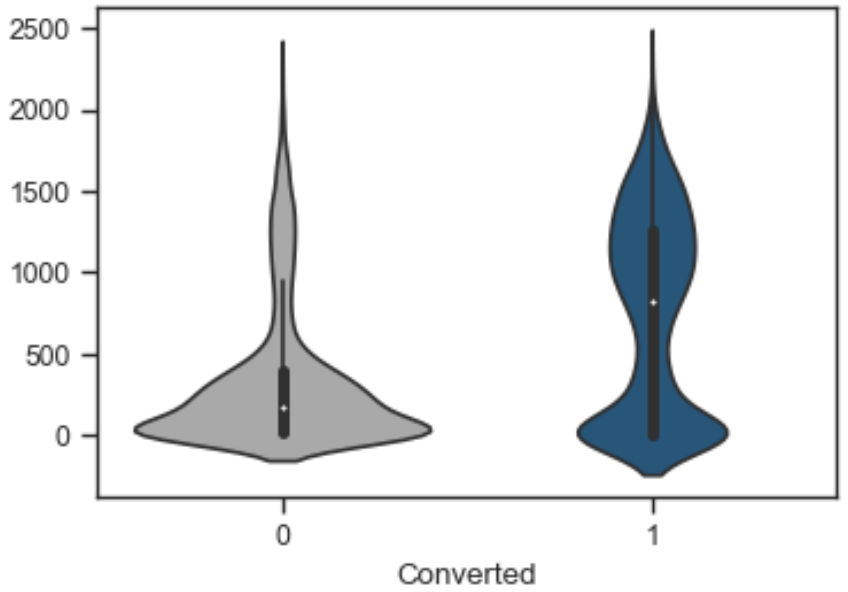
Inference:

- Most of leads have Email Opened and Modified as last notable activity.
- But leads with last activity is SMS Sent are more likely to be hot leads

EDA – INSIGHTS OF DATA

Total Time Spent on Website





Inference:

- Leads who spent more time on Website tend to be hot leads

3. DATA PREPARATION FOR MODEL BUILDING

- Treating categorical variables
- Splitting the Data into Train and Test sets.
- Scaling the Data

3. DATA PREPARATION FOR MODEL BUILDING

Treating Categorical Variables

- Mapping values of binary variable with 0 for No and 1 for Yes.
- Creating dummy variables from other categorical variables
- Drop categorical variables

3. DATA PREPARATION FOR MODEL BUILDING

Splitting and Scaling Data

- Splitting the Data into Train and Test sets, separating feature variables and target variable data
- Scaling: using Standardization technique

4. BUILDING MODEL AND MAKING PREDICTIONS

- Building logistic regression model
- Making predictions.

4. BUILDING MODEL AND MAKING PREDICTIONS

Building Model

Final Model

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0376	0.125	-0.300	0.764	-0.283	0.208
Do Not Email	-1.1631	0.165	-7.038	0.000	-1.487	-0.839
Total Time Spent on Website	1.0901	0.040	27.322	0.000	1.012	1.168
Lead Origin_Landing Page Submission	-1.1896	0.127	-9.350	0.000	-1.439	-0.940
Lead Source_Olark Chat	1.0180	0.122	8.373	0.000	0.780	1.256
Lead Source_Reference	3.3436	0.236	14.144	0.000	2.880	3.807
Lead Source_Welingak Website	6.3075	1.014	6.223	0.000	4.321	8.294
Last Activity_Olark Chat Conversation	-0.9113	0.169	-5.393	0.000	-1.243	-0.580
Last Activity_SMS Sent	1.2599	0.075	16.909	0.000	1.114	1.406
Specialization_Others	-1.1813	0.123	-9.580	0.000	-1.423	-0.940
What is your current occupation_Working Professional	2.5389	0.192	13.213	0.000	2.162	2.916
Last Notable Activity_Modified	-0.9082	0.081	-11.210	0.000	-1.067	-0.749
Last Notable Activity_Others	2.5237	0.827	3.051	0.002	0.902	4.145

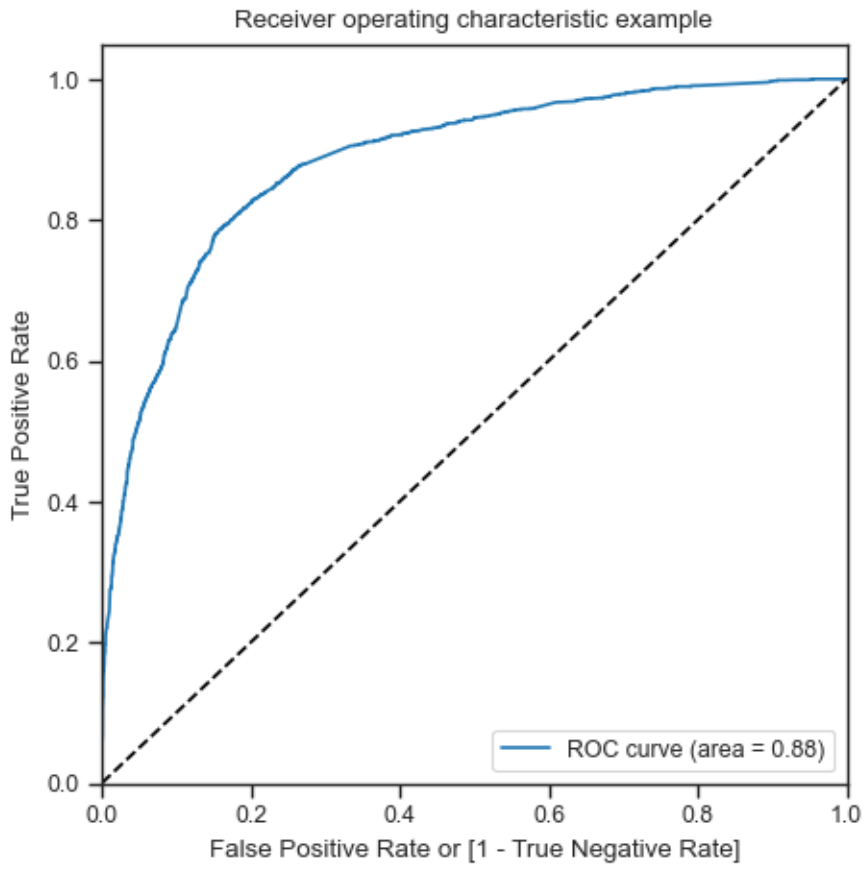
VIFs

	Features	VIF
8	Specialization_Others	2.16
3	Lead Source_Olark Chat	2.07
10	Last Notable Activity_Modified	1.77
2	Lead Origin_Landing Page Submission	1.69
6	Last Activity_Olark Chat Conversation	1.61
7	Last Activity_SMS Sent	1.54
1	Total Time Spent on Website	1.29
4	Lead Source_Reference	1.22
9	What is your current occupation_Working Professional	1.18
0	Do Not Email	1.13
5	Lead Source_Welingak Website	1.09
11	Last Notable Activity_Others	1.00

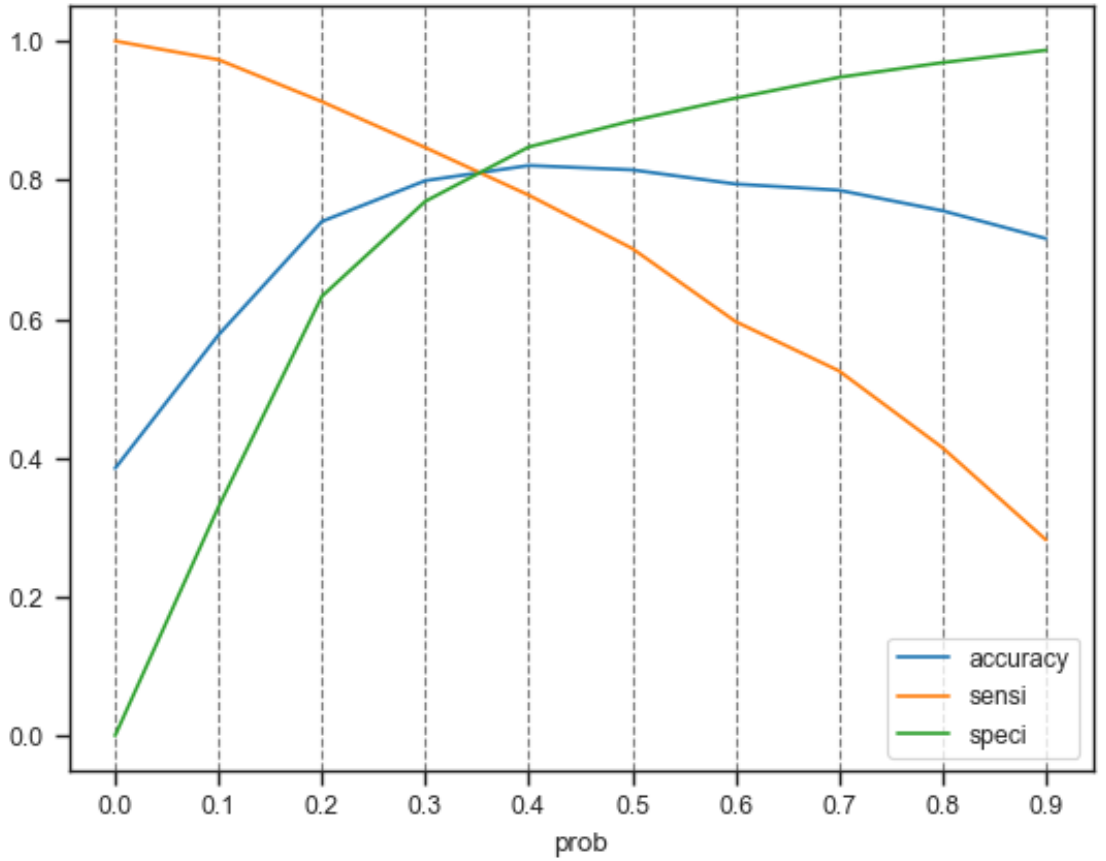
4. BUILDING MODEL AND MAKING PREDICTIONS

Making Predictions

ROC Curve

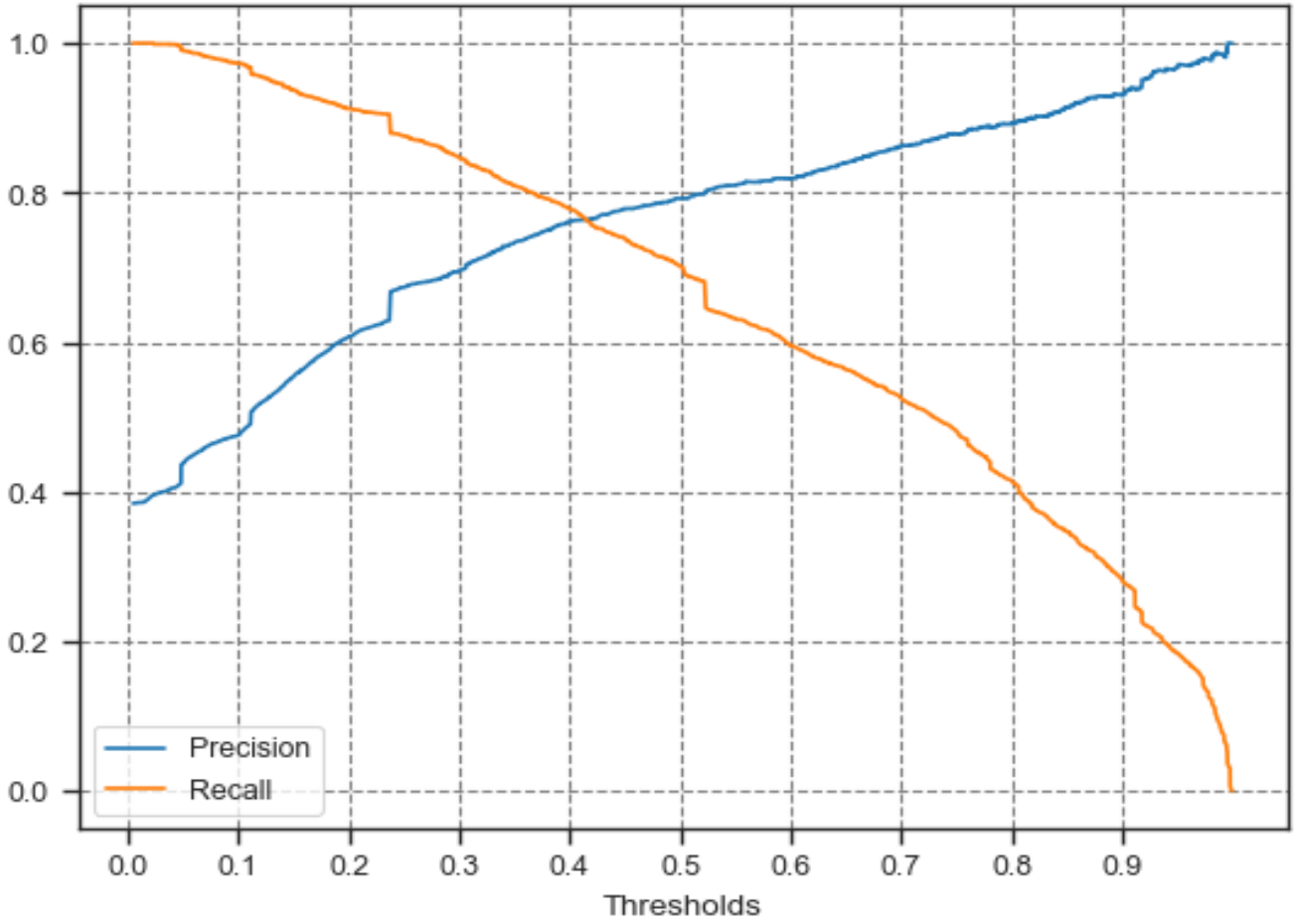


Optimal Cut – off Point: 0.35



4. BUILDING MODEL AND MAKING PREDICTIONS

Making Predictions



Precision and Recall Tradeoff:

- Precision increases, Recall decrease
- The balance point: 0.42
- With Precision around 80%, cut-off point should be 0.52, and Recall is around 68%

Lead Scoring:

- Lead Score = Probability * 100
- A higher score indicated a higher likelihood of conversion

5. MODEL EVALUATION AND INTERPRETATION

- Model Evaluation
- Model Interpretation.

5. MODEL EVALUATION AND INTERPRETATION

Model Evaluation

Model Performance on train set:

```
Sensitivity :  0.8094655242758058
Specificity :  0.8166284111196124
False Positive Rate : 0.18337158888038765
Positive Predictive Value : 0.733999260081391
Negative predictive value : 0.8727173616789315
```

Model Performance on test set:

```
Sensitivity :  0.7952047952047953
Specificity :  0.8184971098265896
False Positive Rate : 0.1815028901734104
Positive Predictive Value : 0.7171171171171171
Negative predictive value : 0.8735348550277606
```

With optimal cut-off point:

- The model performance is about the same on train and test datasets
- The model is good in accuracy
- The model is stable

5. MODEL EVALUATION AND INTERPRETATION

Model Interpretation

Feature coefs

Lead Source_Welingak Website	6.307545
Lead Source_Reference	3.343620
What is your current occupation_Working Professional	2.538926
Last Notable Activity_Others	2.523744
Last Activity_SMS Sent	1.259898
Total Time Spent on Website	1.090053
Lead Source_Olark Chat	1.018029
const	-0.037606
Last Notable Activity_Modified	-0.908189
Last Activity_Olark Chat Conversation	-0.911327
Do Not Email	-1.163114
Specialization_Others	-1.181291
Lead Origin_Landing Page Submission	-1.189630

5. MODEL EVALUATION AND INTERPRETATION

Model Interpretation

Model Interpretation:

- **Lead Source_Welingak Website, Lead Source_Reference** and **What is your current occupation_Working Professional** are 3 most positively significant features
- **Do Not Email, Specialization_Others** and **Lead Origin_Landing Page Submission** are 3 most negatively significant features

CONCLUSION

Significant insights:

- We highly recommend that the company should contact leads who come from Welingak Website and Reference because they are much more likely to be hot leads.
- We highly recommend that the company should **not** call leads who don't want to **receive email,** have **Others Specialization** or come from **Landing Page Submission** because they are not likely to be hot leads
- Since CEO wants the lead conversion rate is around 80%, the **Precision** should be around 80%, hence the cut-off point should be 0.52. That means sale team should contact leads who have lead score 52 at least (lead score = probability * 100).
- Because Leads with higher scores were indeed more likely to convert, the sales team to focus on high-priority leads, optimizing their communication efforts