LEAD SCORING CASE STUDY – X EDUCATION

Identification of hot leads for X Education's sales team so that they may focus their sales efforts on the most promising leads and ultimately reach the 80% conversion rate target.

Team Members - Prashant Dhar Dwivedi & Rakshith P Shetty

INDEX

About X-Education

- X Education is an education organization that sells online courses to industry professionals.
- A lot of professionals interested in the courses visit their website and search for courses.
- The company promotes its courses on various websites and search engines such as Google.
- When these users arrive at the website, they may browse the courses, fill out a course registration form, or watch some videos.
- People are regarded as leads when they fill out a form with their email address or phone number. The company also gets leads through past referrals**.**
- Once these leads are obtained, members of the sales team begin making calls, composing emails, and so on.
- The normal lead conversion rate at X education is roughly 30% during this approach.

Problem Statement & Expected Outcome

Problem Statement

- X Education receives a large number of leads; yet, its lead conversion rate is quite low, hovering around 30%.
- To improve the efficiency of this process, the organization wishes to identify the most promising leads, referred to as 'Hot Leads.'
- If they are successful in identifying this group of leads, the lead conversion rate should increase because the sales staff will now be focusing on connecting with the potential leads rather than calling everyone.

Expected Outcome

- The company expects us to create a model in which you give a lead score to each lead so that customers with higher lead scores have a higher conversion chance and customers with lower lead scores have a lower conversion chance.
- The CEO, in particular, has stated that the target lead conversion rate should be about 80%.

Steps taken for analysis

Data Cleaning:
Load, Understand and clean the dataset

EDA:
Univariate and Bivariate Data understating and Visualization

Data Preparation:
Creating Dummy Variables, Train test Split, Feature scaling

Feature Selection:
Selection of top 30 features by RFE, Feature Elimination by VIF and P value

Model Evaluation:
Confusion matrix, ROC Curve, Accuracy, Specificity & Sensitivity

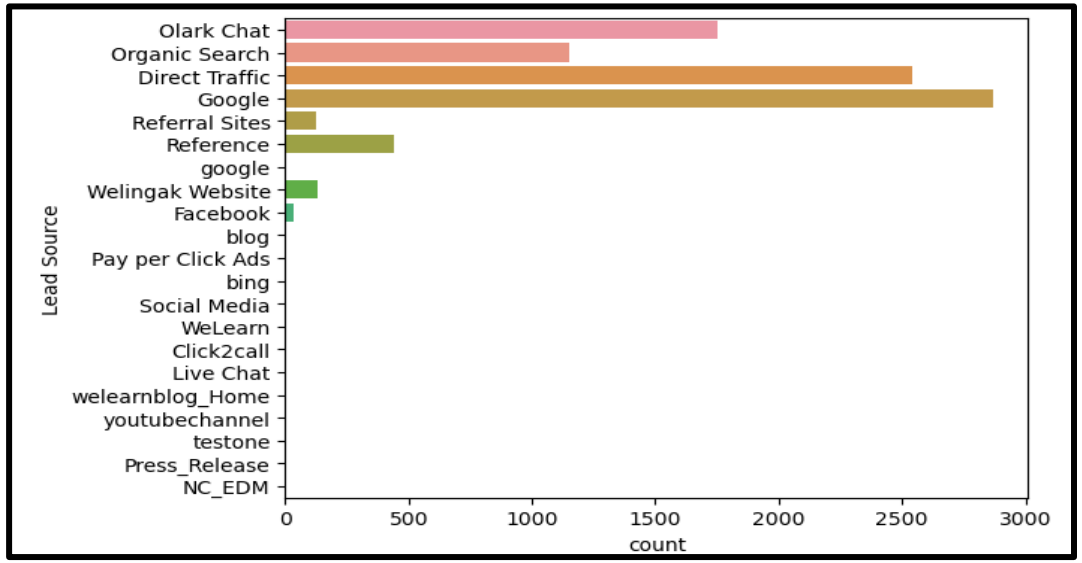
Prediction on Test set :
Compare the evaluation metrics of Train and test dataset

Data Cleaning

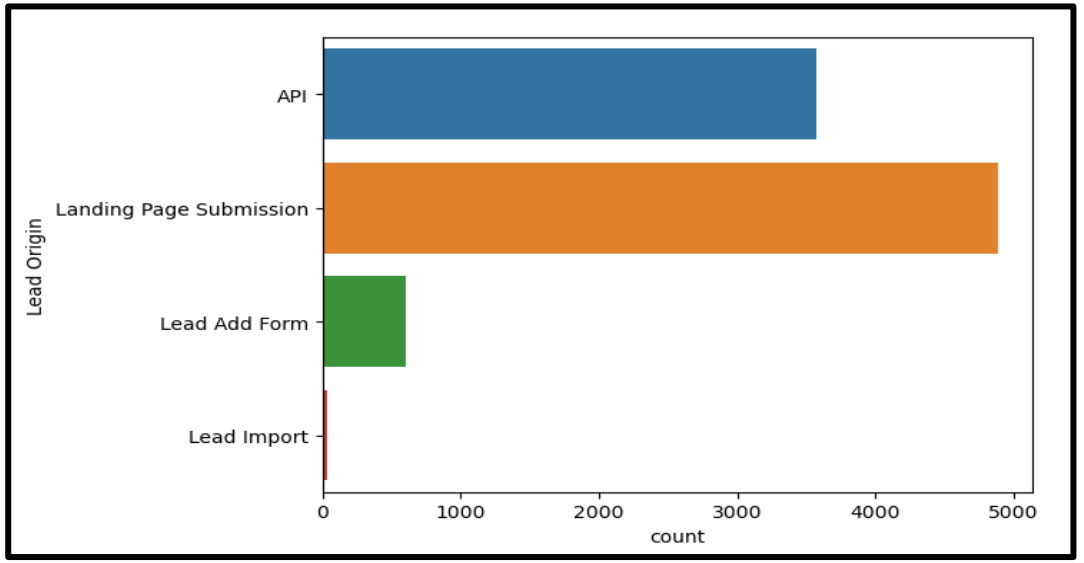
- Columns With more than 35% null Values Were dropped.
- Columns which is irrelevant or data imbalance with respect to analysis(Prospect ID, Lead Number, DM Content etc.) have been dropped.
- Categorical variables have a level named 'Select' that must be handled because it is equivalent to a null value. As a result of the columns with Select has a variable has been replaced with Mode of that specific Column.

EDA-Univariate Analysis

Count Plot of Lead Source



Count Plot of Lead Origin

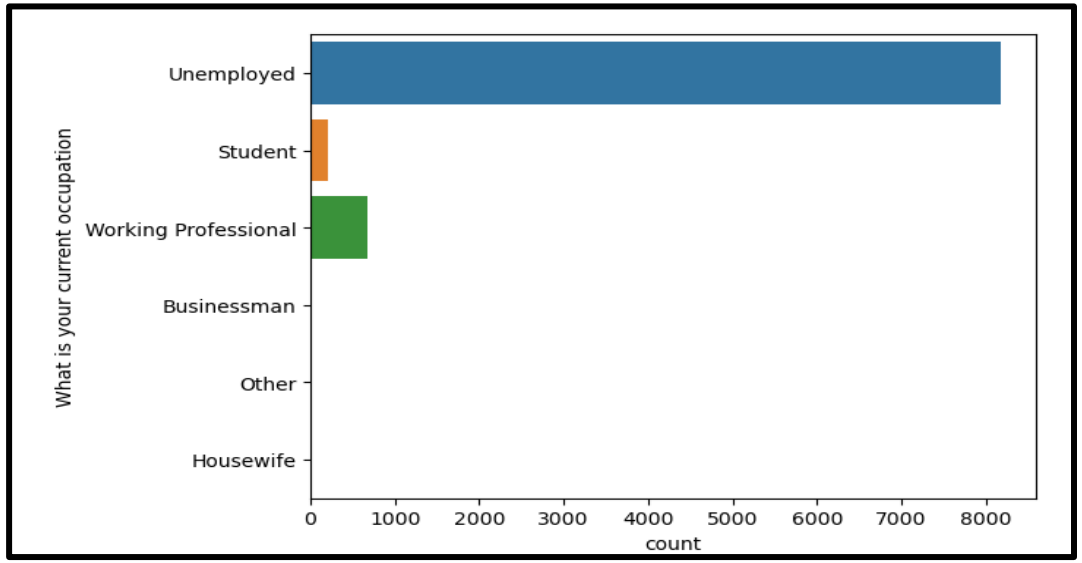


Inference:

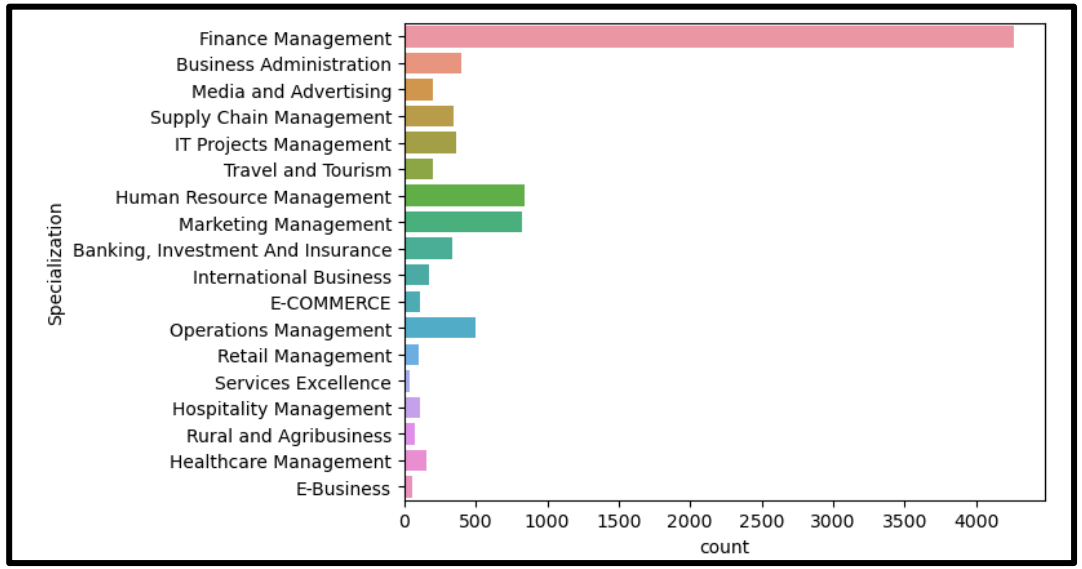
- Lead Source: The Majority of the Leads were identified through Google, Direct Traffic and Olark Chart.
- Lead Origin: The Majority of the Leads were originated through Landing page Submission and API.

EDA-Univariate Analysis

Count Plot of Current Occupation



Count Plot of Specialization

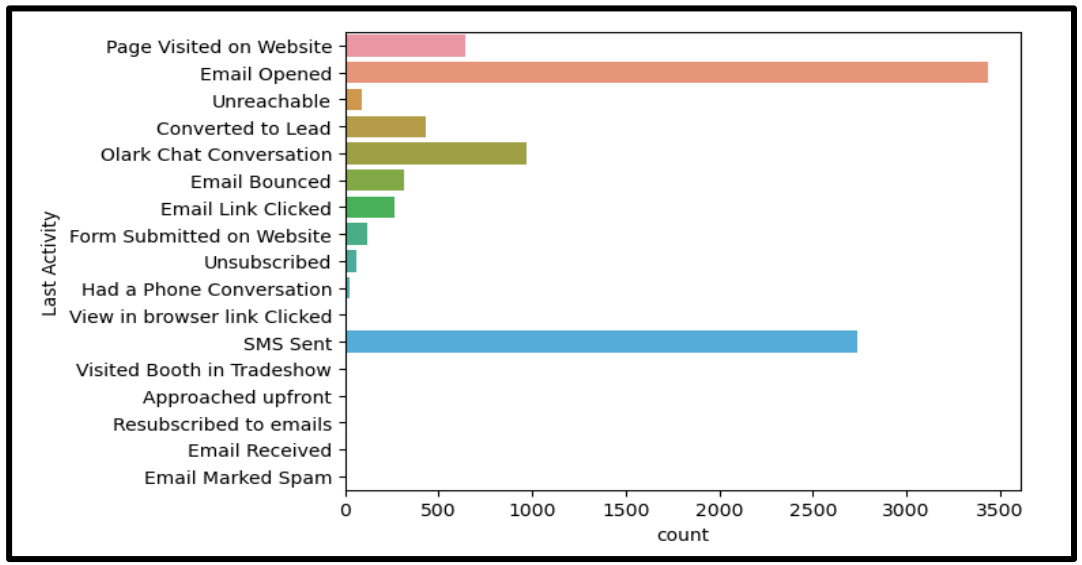


Inference:

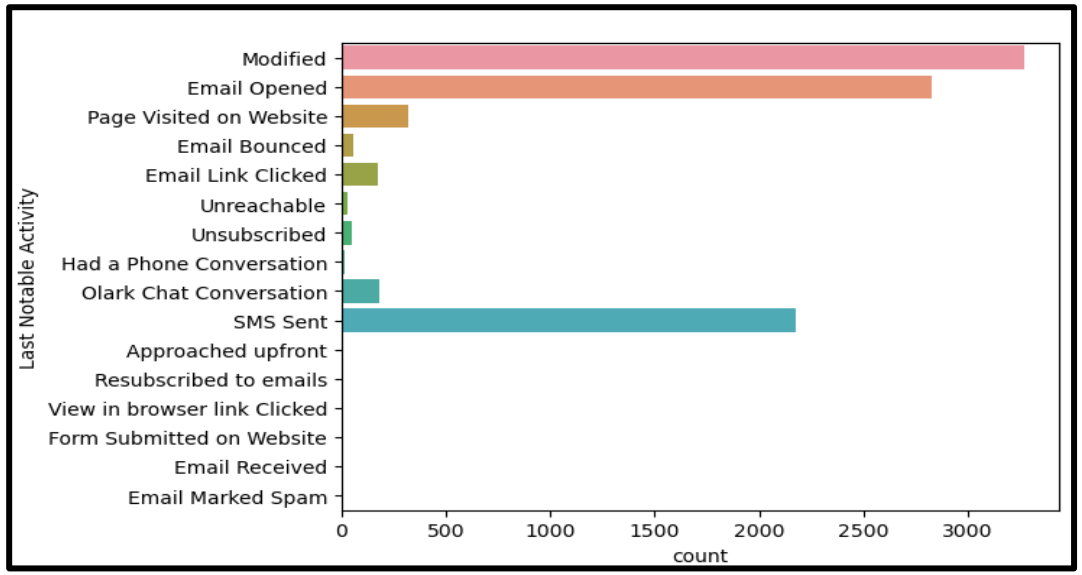
- Current Occupation: The Majority of the Leads were identified as Unemployed.
- Specialization : The Majority of the Leads has specialization as Finance Management.

EDA-Univariate Analysis

Count Plot of Last Activity



Count Plot of Last Notable Activity

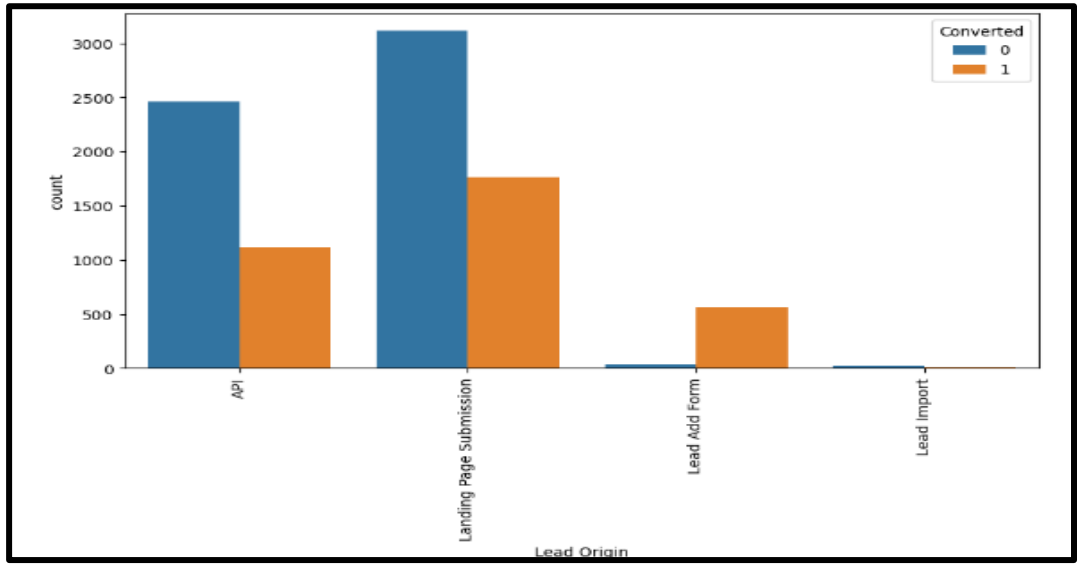


Inference:

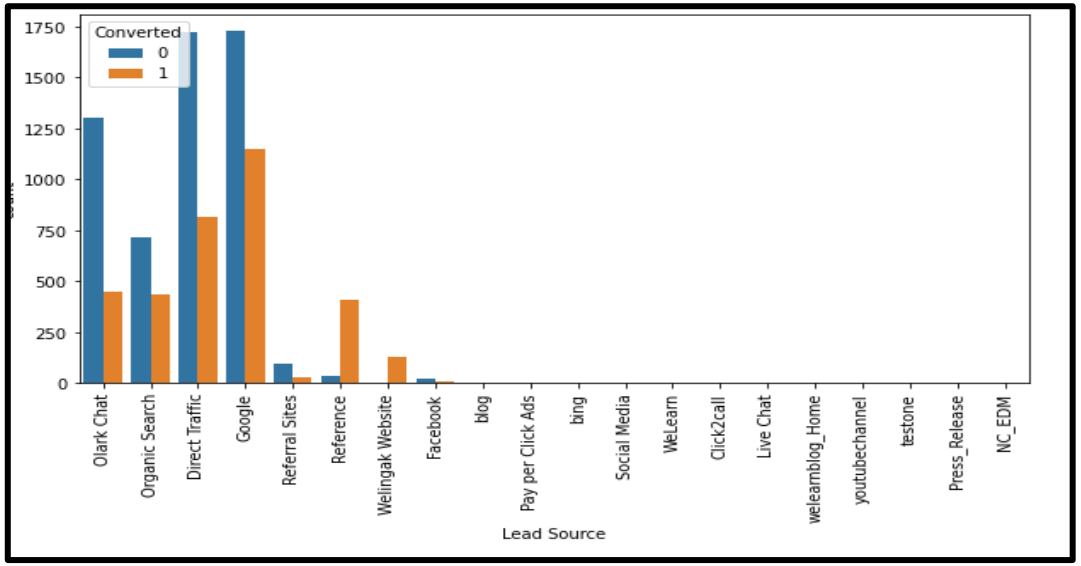
- Last Activity : The Majority of the Leads last activity's are Email Opened and SMS Sent.
- Last Notable Activity : The Majority of the Leads last notable activity's are Modified, Email Opened and SMS Sent.

EDA-Bivariate Analysis

Distribution of Lead Origin



Distribution of Lead Source

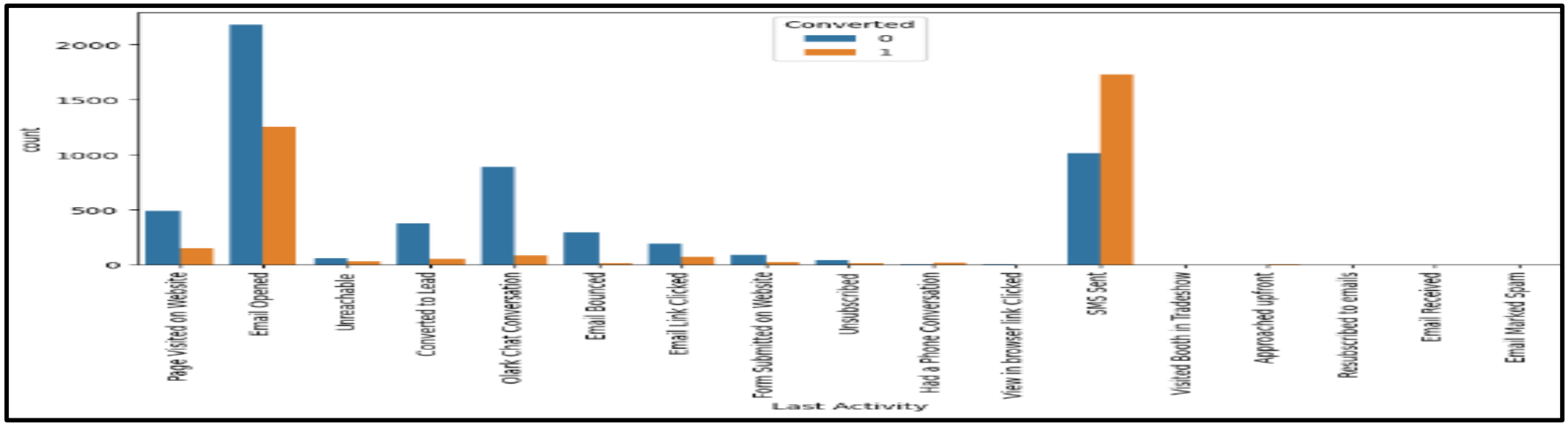


Inference:

- Lead Origin : Lead Source API and Landing Page Submission have higher lead conversion with lead conversion rates of 31% and 36% respectively.
- Lead Source: Lead conversion are good lead source of Google and Direct Traffic with Lead conversation rate of 32% and 40%.

EDA-Bivariate Analysis

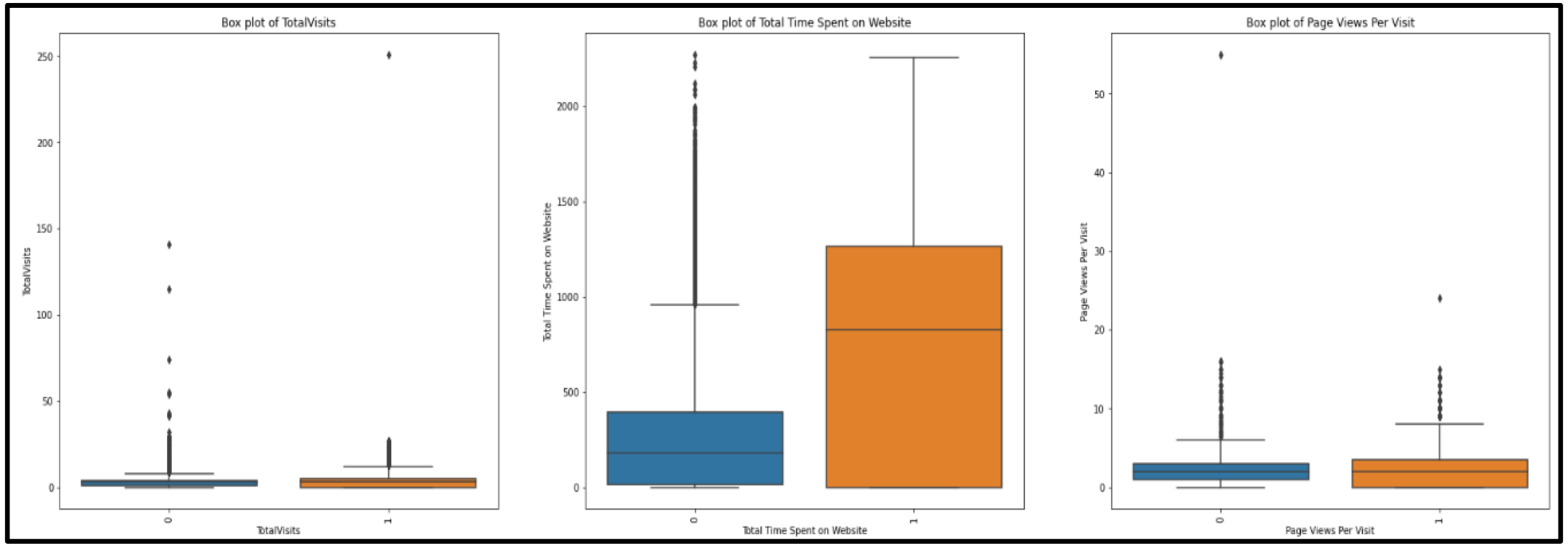
Distribution of Last Activity



Inference:

- Last Activity : Last lead activity of SMS Sent and Email Opened Submission have higher lead conversion with lead conversion rates of 63% and 36% respectively.

EDA-Bivariate Analysis



Inference:

- Past Leads who spends more time on the Website have a higher chance of getting successfully converted than those who spends less time as seen in the box-plot

Data Preparation

- Creation of Binary Mapping for the column “A free copy of Mastering The Interview”.
- Creation dummy features for columns with categorical variables.
- Concatenation of dummy variable Dataframe and Existing dataset dataframe.
- Dropping all the columns with categorical variables as dummy variable Dataframe will help us in identifying the categorical variables.
- Splitting Train & Test Sets at 70:30 ratio.
- Feature scale all the numerical variable using Min Max Scaler.

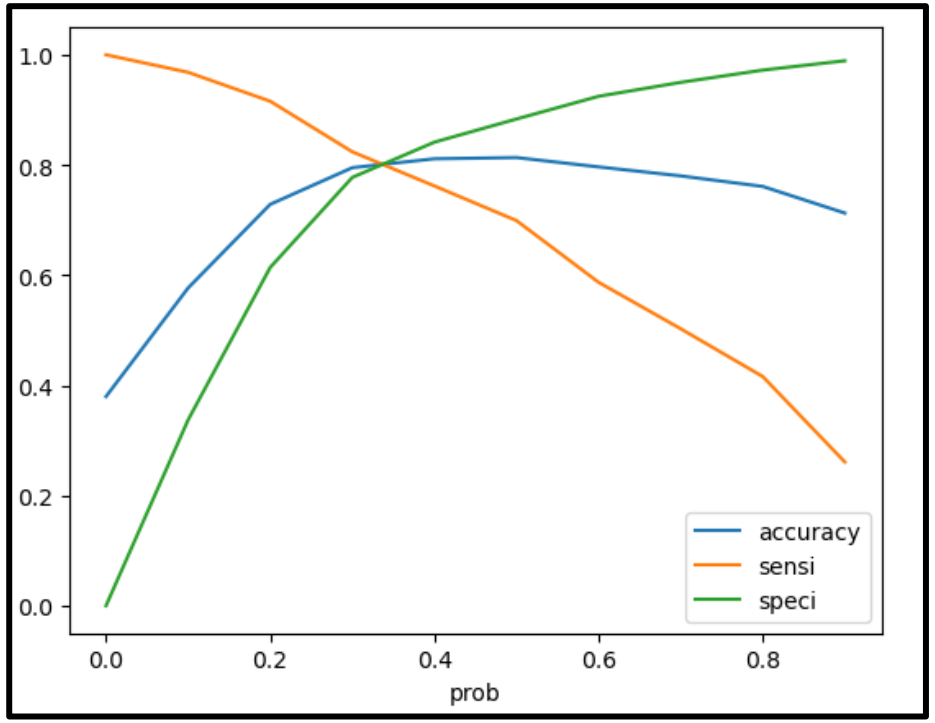
Feature Selection

- Importing and Performing Recursive Feature Elimination (RFE) to select 30 top Features from the Existing dataframe.
- The model can then be fine-tuned by manually removing columns with a higher Variance Inflation Factor (VIF) or a higher P value for the feature.
- After Manually eliminating the column, our final model is saved in logm7 model which has 15 features included n the model.

Model Evaluation- 0.5 Cutoff

Actual/Predicted	Not_Converted	Converted
Not_Converted	3492	461
Converted	728	1691

Metrics	Score
Accuracy	0.81
Sensitivity	0.70
Specificity	0.88

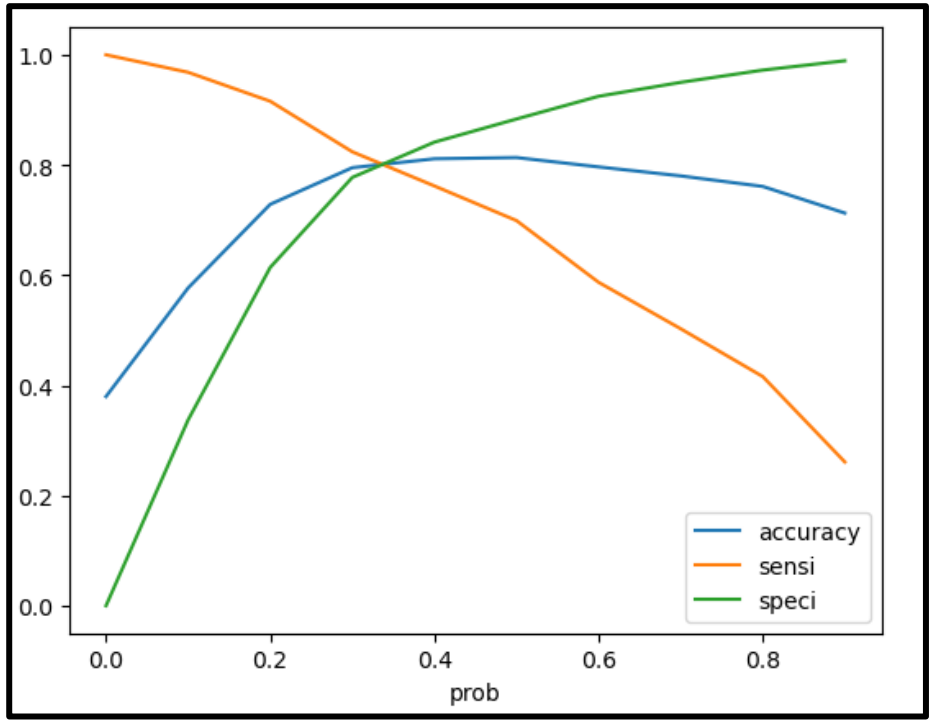


- Based on the above Plot we have considered lead score of 0.32 as the cutoff . Anyone with the lead score of above 0.32 will considered as Converted.

Model Evaluation – 0.32 Cutoff

Actual/Predicted	Not_Converted	Converted
Not_Converted	3114	839
Converted	449	1970

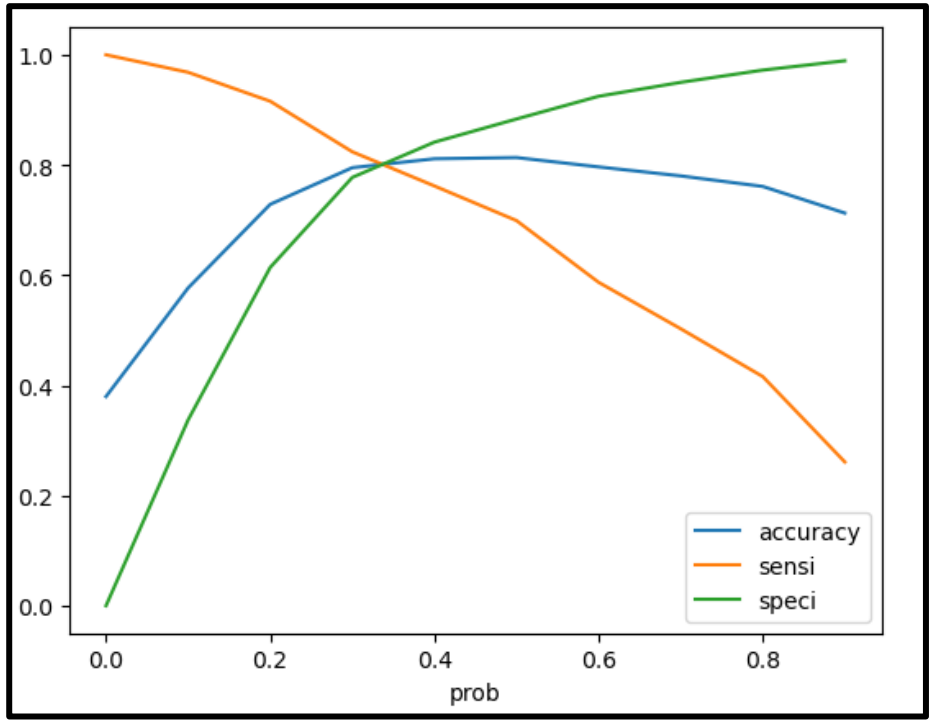
Metrics	Score
Accuracy	0.80
Sensitivity	0.81
Specificity	0.79
Precision	0.79
Recall	0.70



Model Evaluation – Test Dataset

Actual/Predicted	Not_Converted	Converted
Not_Converted	1319	370
Converted	181	861

Metrics	Score
Accuracy	0.80
Sensitivity	0.83
Specificity	0.78
Precision	0.70
Recall	0.82



Recommendation

- X Education should target Working professional has they tend to have higher lead conversion rate.
- More Lead engagement should be done on Welingak Website as it has higher lead conversion rate.
- Lead originated from Lead add form are most likely to convert as a customer