Lead Score Case Study

Group Members
1. Vishal C
2. Shashwat Avi
3. Thirumala Reddy

Problem Statement

- An Education company “X Education" sells online courses to industry professionals.
- X Education gets lots of leads and the lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only 30 of them are converted to customers.
- To make this process more efficient, the company needs to identify the most potential leads, also reffered as ‘Hot Leads’.
- If successfully identified this set of leads, the lead conversion rate would go up as the sales team would now be focusing more on communicating with the potential leads.

Business Objective:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use**.**

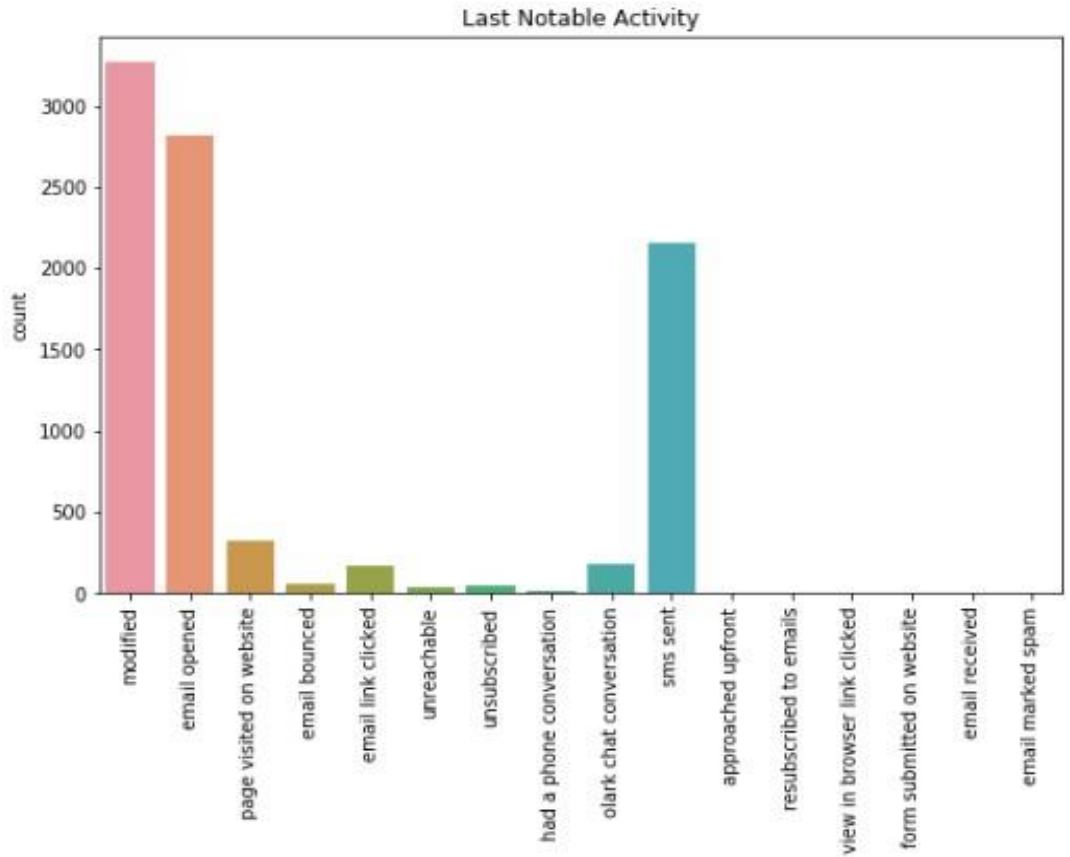
Solution Methodology

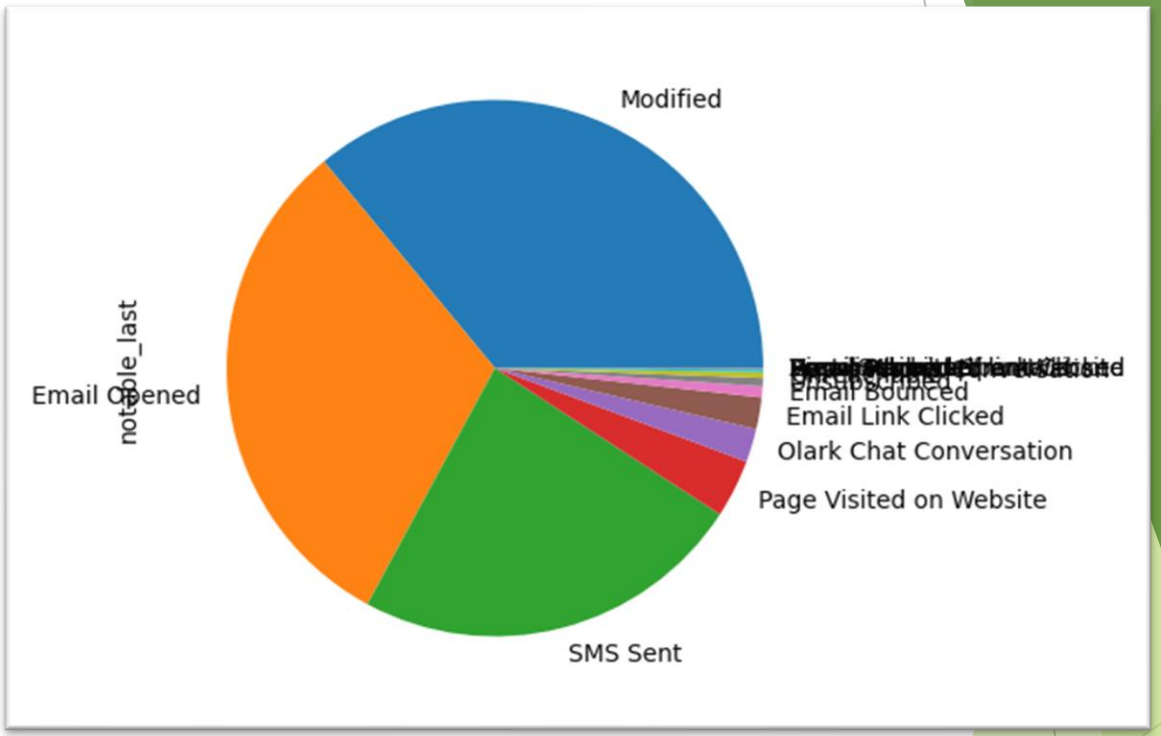
- Data cleaning and data manipulations done:-
  1. Checking and handling duplicate data
  2. Checking and handling NA values and missing values
  3. Dropping columns, if it contains large amount of missing values and if it is not useful for our analysis
  4. Imputation of values wherever necessary
  5. Checking and handling outliers in data
- EDA
  1. Univariate data analysis: value count, variable distribution etc.
  2. Bivariate data analysis: coefficients correlation and variable patterns etc.
- Feature Scaling & Dummy Variables & Data Encoding
- Classification technique: logistic regression for the model making and prediction
- Validation of the model
- Model presentation
- Conclusions and recommendations

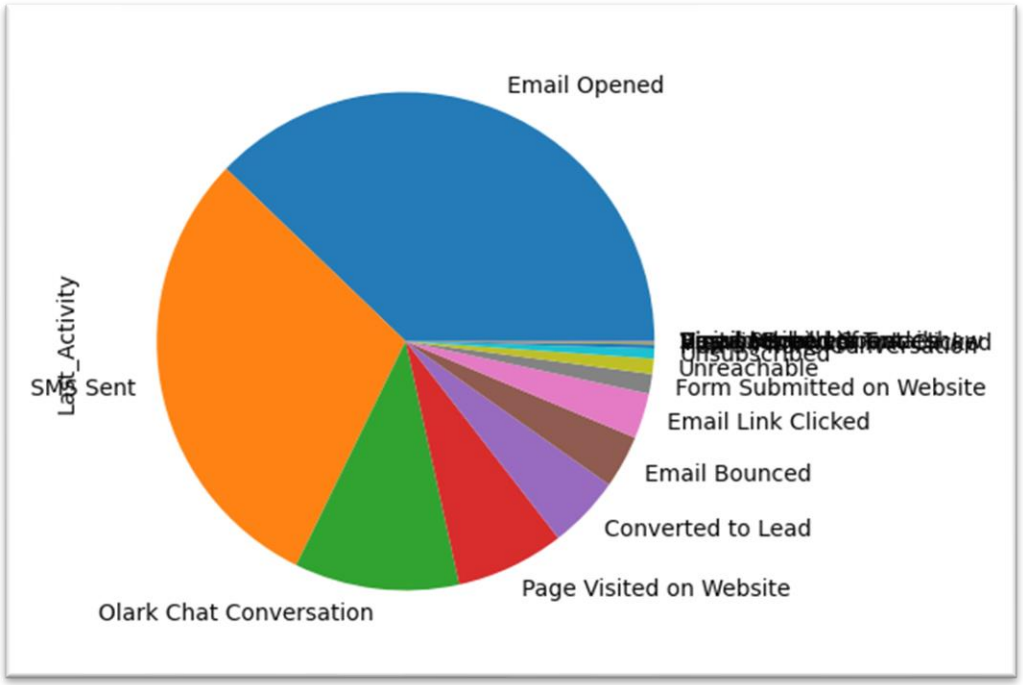
Data Manipulation

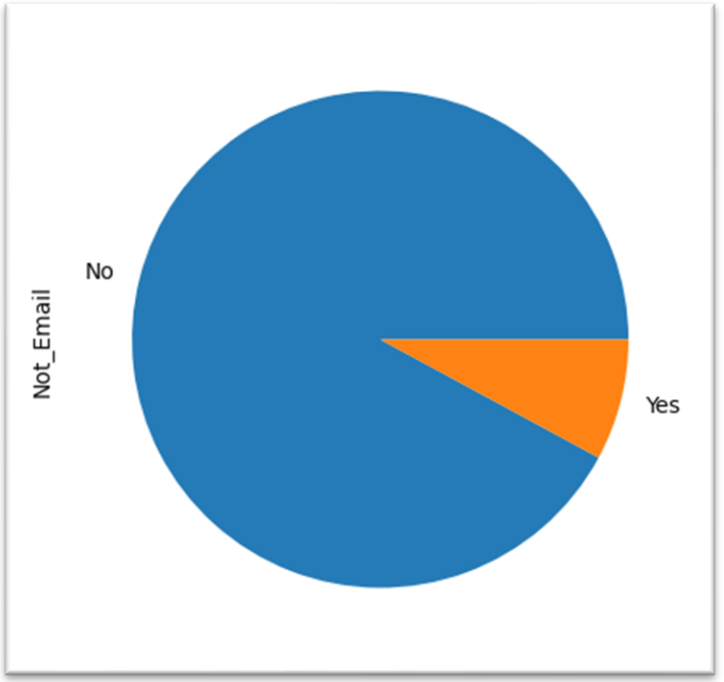
- Total Number of Rows =37, Total Number of Columns =9240.
- All Single value features i.e., “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”, “Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- We remove the “Prospect ID” and “Lead Number” as it is not necessary for our analysis.
- On checking the value counts for some object type variables, we notice that some of the features do not have enough variance, so we have dropped them, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- Dropping 8 columns having more than 40% as missing value such as ‘How did you hear about X Education’, ‘Activity Index’, ‘Profile Index', ‘Activity_Score', ‘Profile_Score', ‘Lead_Quality', ‘Lead_Profile’ and 'City’.
- Further we drop columns like “Country”, “Better Career Prospects”, ”Do Not Email” and “Do Not call” due to high percentage (above 95%) data of one distinct value and we don’t want our analysis to be biased and the contribution of this variable to the variability of the model would be negligible.

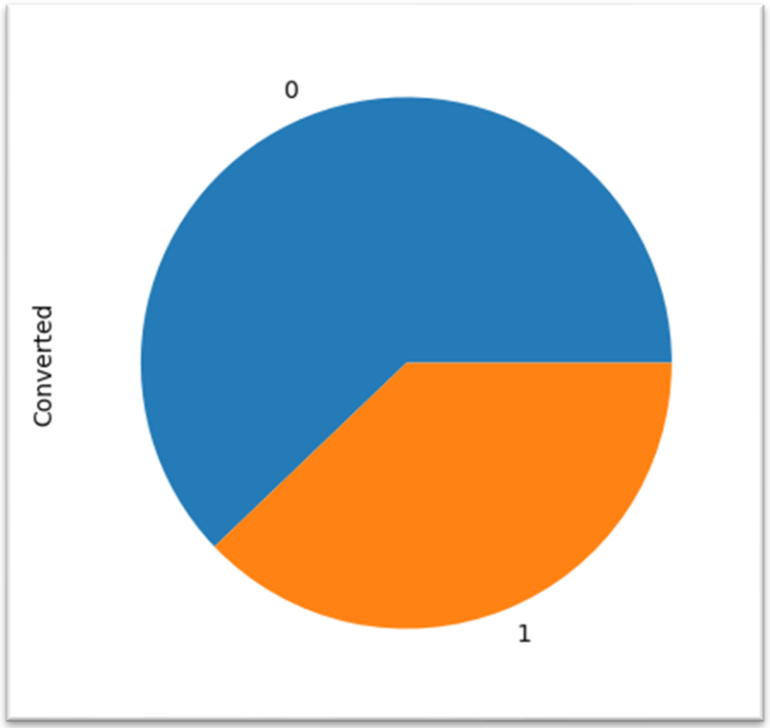
EDA

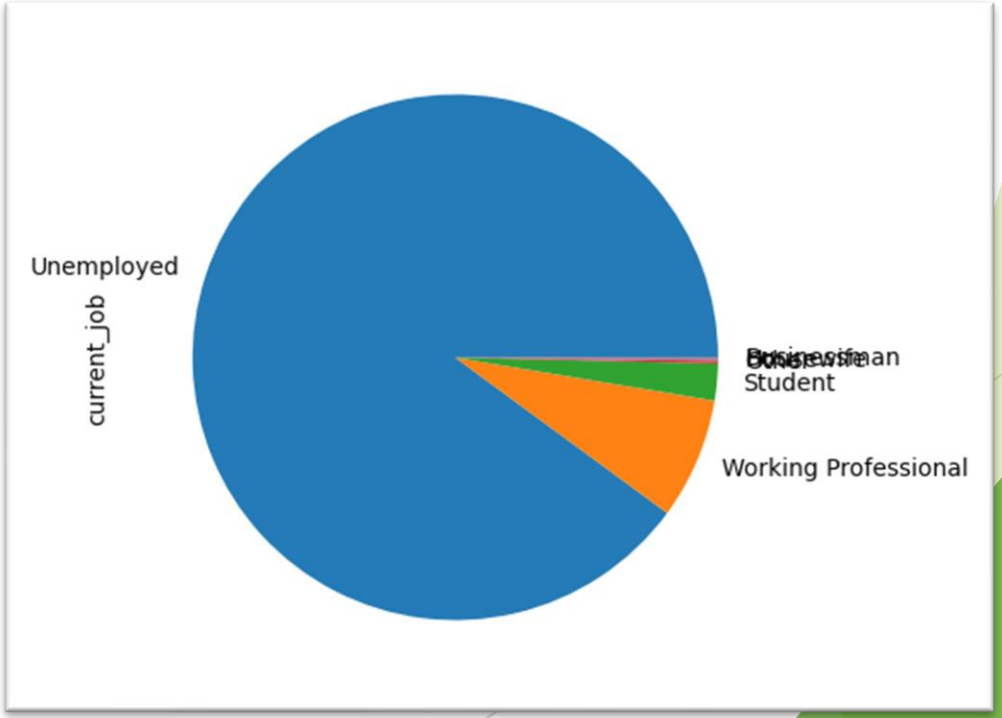

Last Notable Activity
count
3000
2500
2000
1500
1000
500
0
modified
email opened
page visited on website
email bounced
email link clicked
unreachable
unsubscribed
had a phone conversation
olark chat conversation
sms sent
approached upfront
resubscribed to emails
view in browser link clicked
form submitted on website
email received
email marked spam


Modified
notable_last
Email Opened
Email Bounced
Email Link Clicked
Olark Chat Conversation
Page Visited on Website
SMS Sent

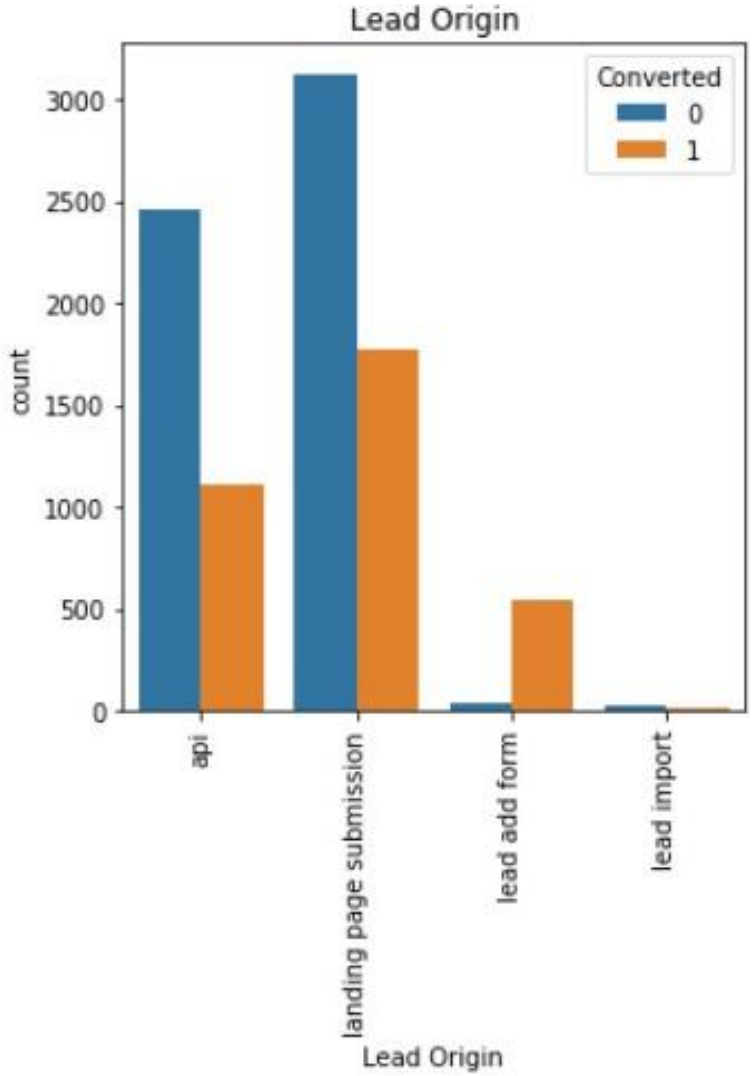

Email Opened
Last_Activity
SMS Sent
Unreachable
Form Submitted on Website
Email Link Clicked
Email Bounced
Converted to Lead
Page Visited on Website
Olark Chat Conversation

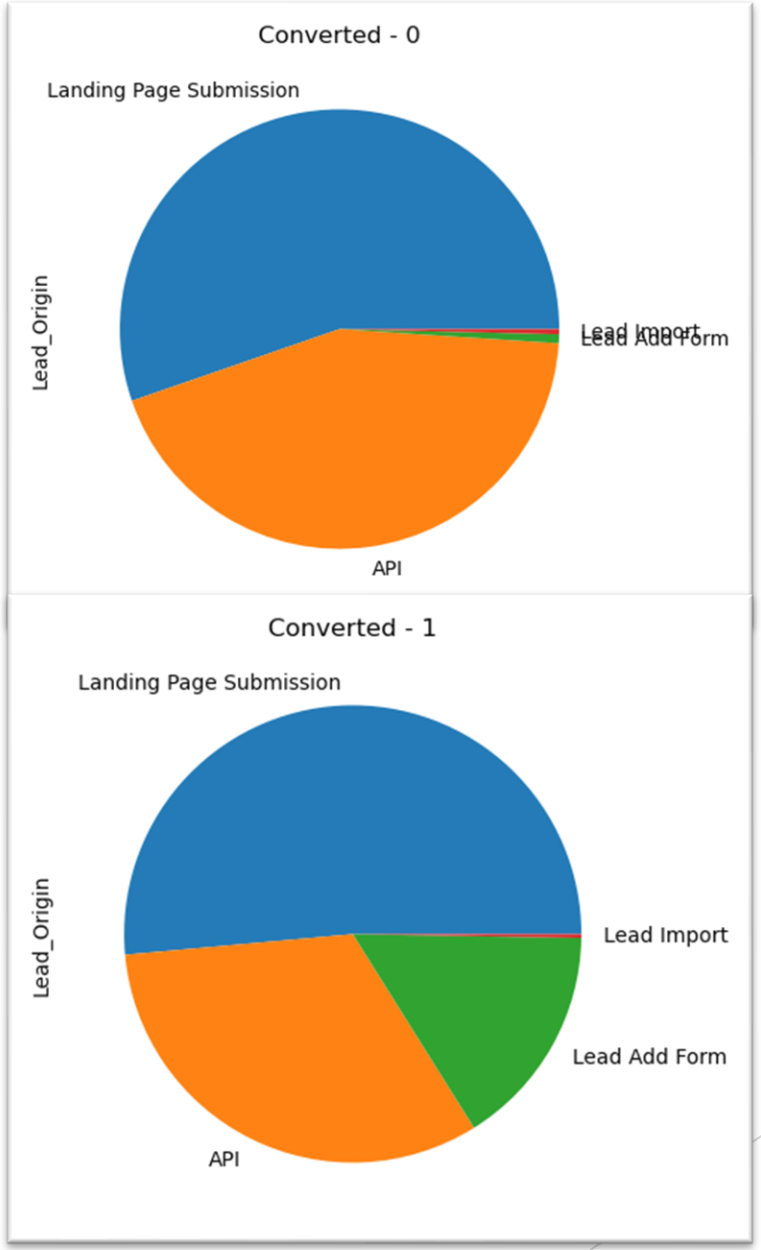

Not_Email
No
Yes

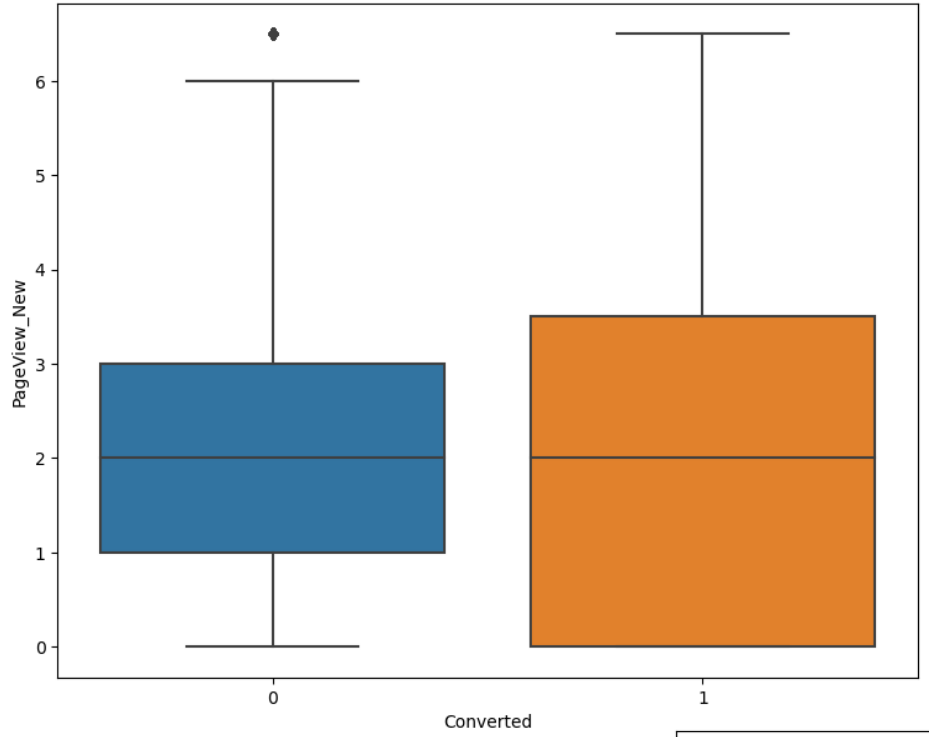

Converted
0
1

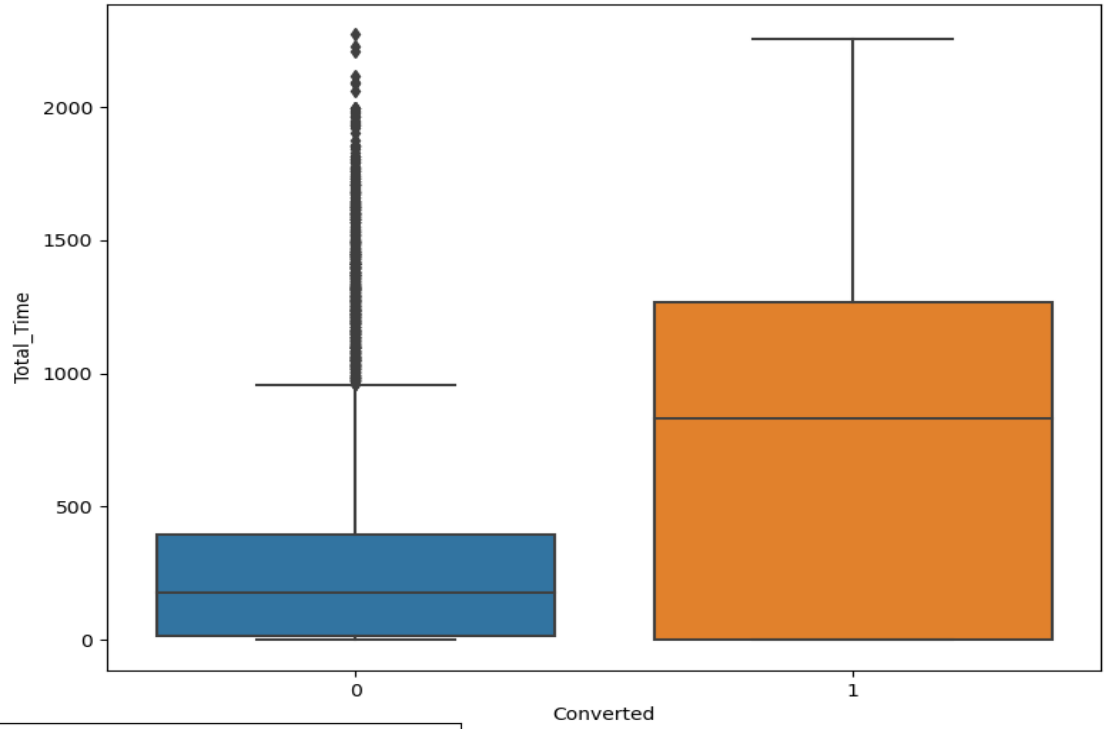

current_job
Unemployed
Student
Working Professional

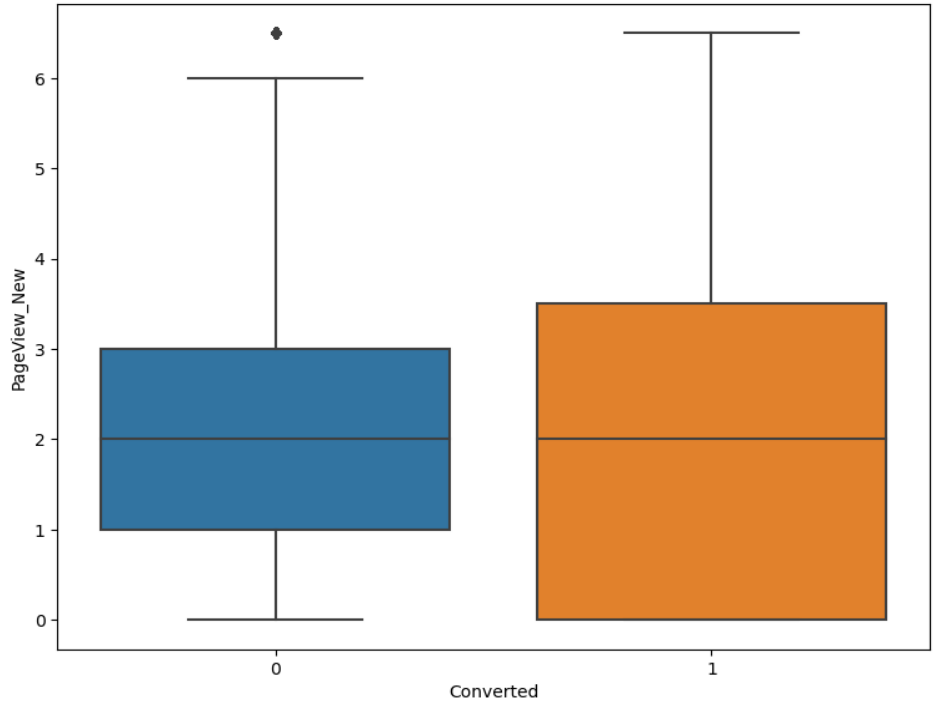
Categorical Variable Relation

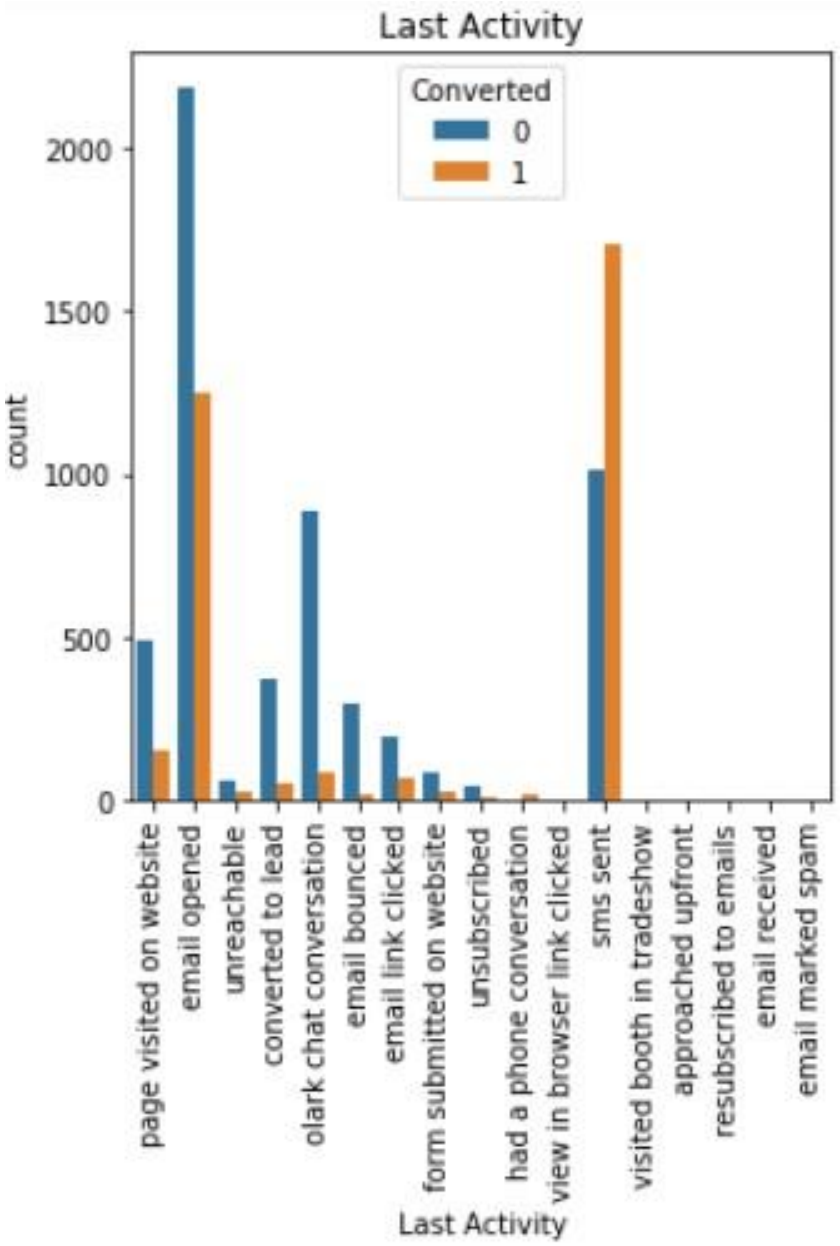


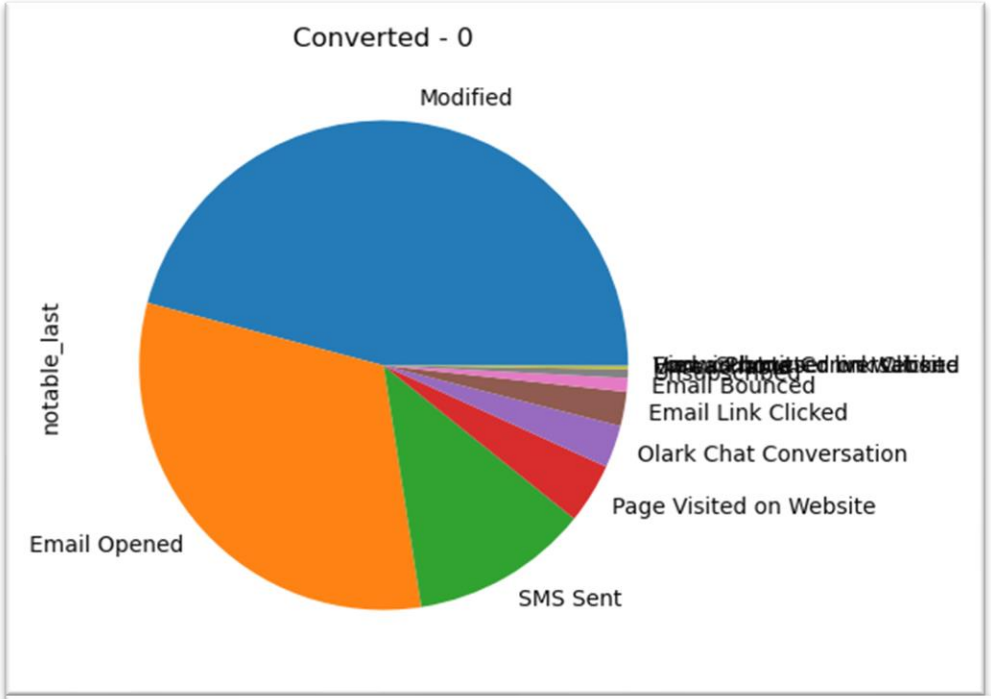


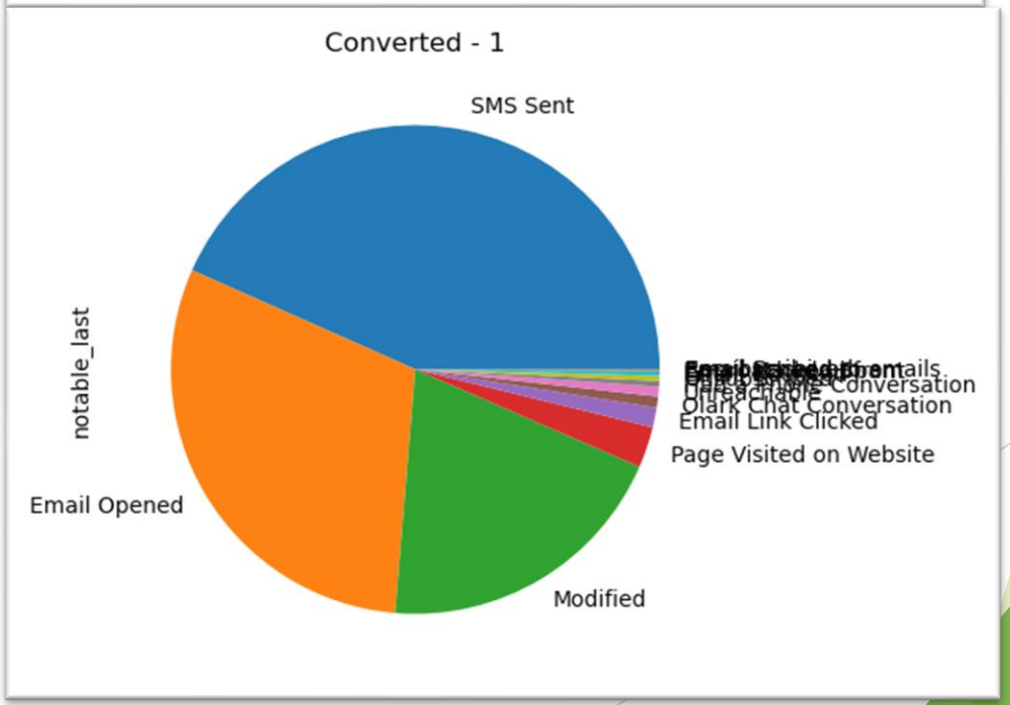

PageView_New
0
1
2
3
4
5
6
Converted
0
1


Total_Time
0
500
1000
1500
2000
Converted
0
1


PageView_New
0
1
2
3
4
5
6
Converted
0
1


Last Activity
Converted
0
1
2000
1500
1000
500
0
count
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
sms sent
visited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam
Last Activity


Converted - 0
Modified
notable_last
Email Bounced
Email Link Clicked
Olark Chat Conversation
Page Visited on Website
Email Opened
SMS Sent


Converted - 1
SMS Sent
notable_last
Olark Chat Conversation
Email Link Clicked
Page Visited on Website
Email Opened
Modified

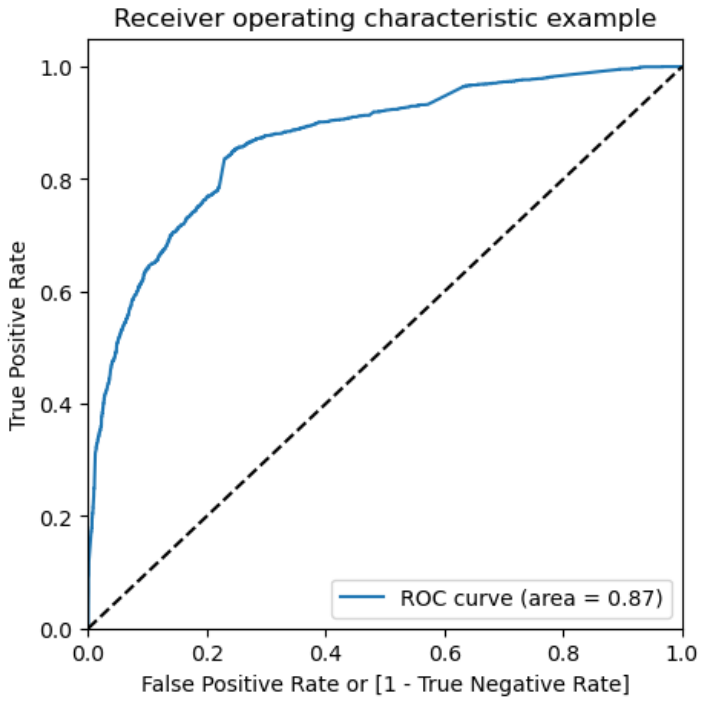
Data Conversion

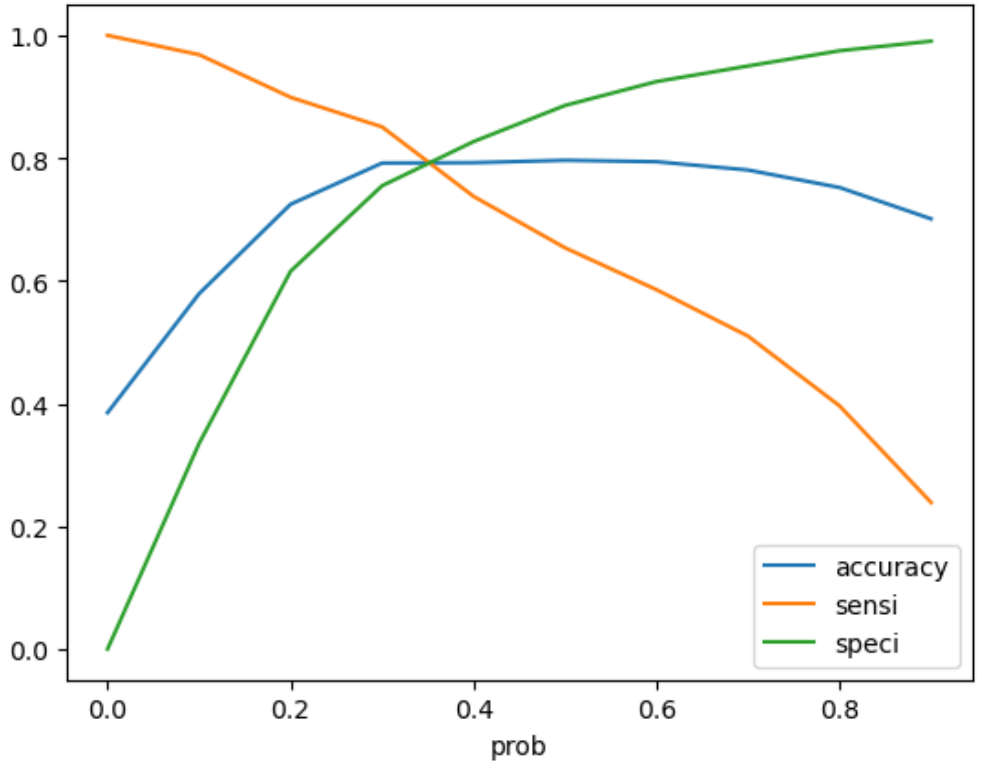
- Numerical Variables are Normalized using Standard Scaler
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 9074
- Total Columns including dummy for Analysis: 64 predictor variable and 1 target variable

Model Building

- Data is split into Training and Testing Sets
- Performed a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5
- Dropping variables with p-value > 0.5
- Validating using Predictions on test data set
- Overall accuracy 79%, sensitivity 83%, specificity 75%

ROC Curve





- **Finding Optimal Cut off Point,** Probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.3.

Conclusion

- We conclude that for high probability of conversion, the factors to focus on are :

a. Total Time spent on website
b. Whether they submitted a form on website (Lead Origin -Add form)
c. Whether an SMS was sent to the lead
d. Whether the lead is a working professional (Current job - working professional)
e. Whether the company had phone conversation as the most recent notable activity (Last Activity – had a phone conversation)