Lead Scoring Case Study

Team Members:

1. Harsh Patel
2. Nikhil Jindal
3. Vaibbhav Nadkarnni

Problem Statement

- X Education company sells online courses to industry professionals.
- Company is facing challenges in lead conversion despite generating significant leads daily.
- The company's lead conversion rate stands at modest 30%, well below the desired target of 80%.
- X Education seeks a solution to identify 'Hot Leads' - those with the highest potential for conversion.
- The goal is to prioritize resources towards engaging with leads most likely to convert into paying customers.

Business Objective:

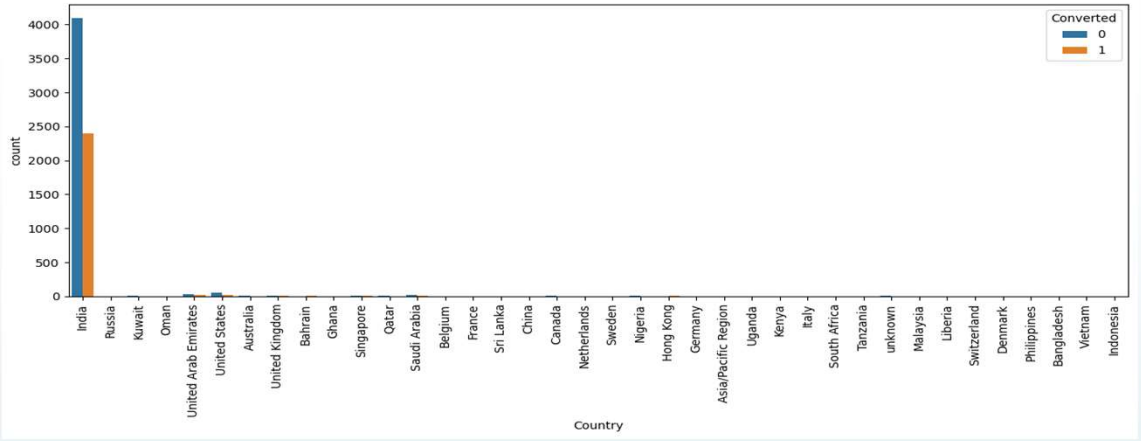
- To increase lead conversion rate from the current 30% to 80% by identifying and prioritizing the most promising leads.
- Build a model and assign lead score to each of the leads to find customer with higher lead score.
- Deployment of the model for future application.

Solution Methodology

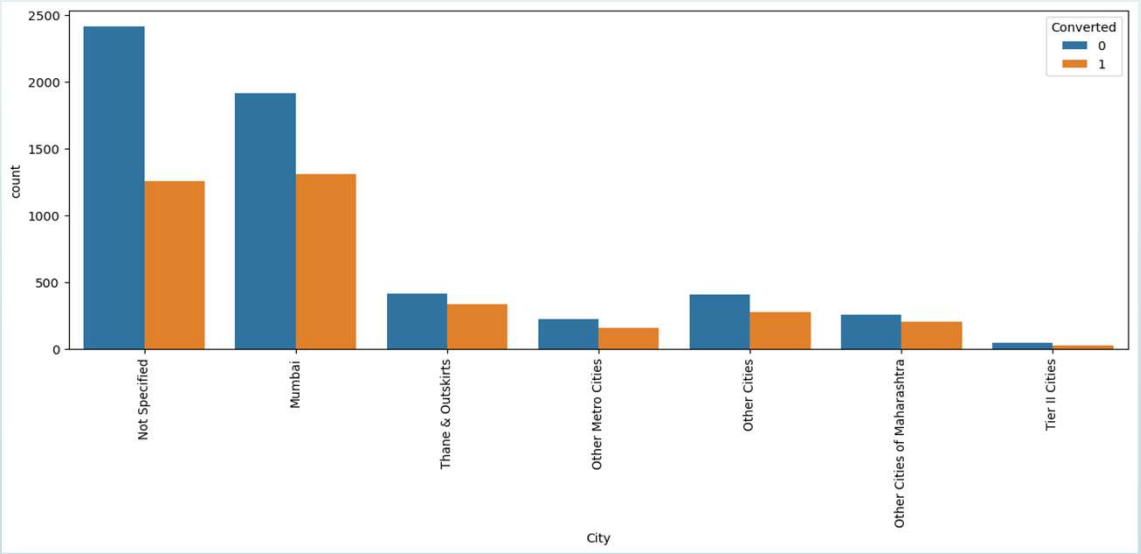
- Data Cleaning and Data Manipulation:
  - Check and handle duplicate values.
  - Check and handle NaN values and missing values.
  - Check and handle “Select” level values in categorical variables.
  - Drop columns with large number of missing values and not useful for analysis.
  - If required, impute missing values.
  - Check and handle outliers.
- Exploratory Data Analysis:
  - Univariate Analysis: value count, box plots, distribution of variables, etc.
  - Multivariate Analysis: correlation coefficient, heatmaps, pattern identification, etc.
- Dummy variable creation, binary variable encoding, and feature scaling.
- Building a logistic regression model and evaluation of the model.
- Building ROC curve.
- Model presentation.
- Conclusion and recommendation.

Exploratory Data Analysis

- Since only country India has high values, other countries has negligible values, They can be dropped.

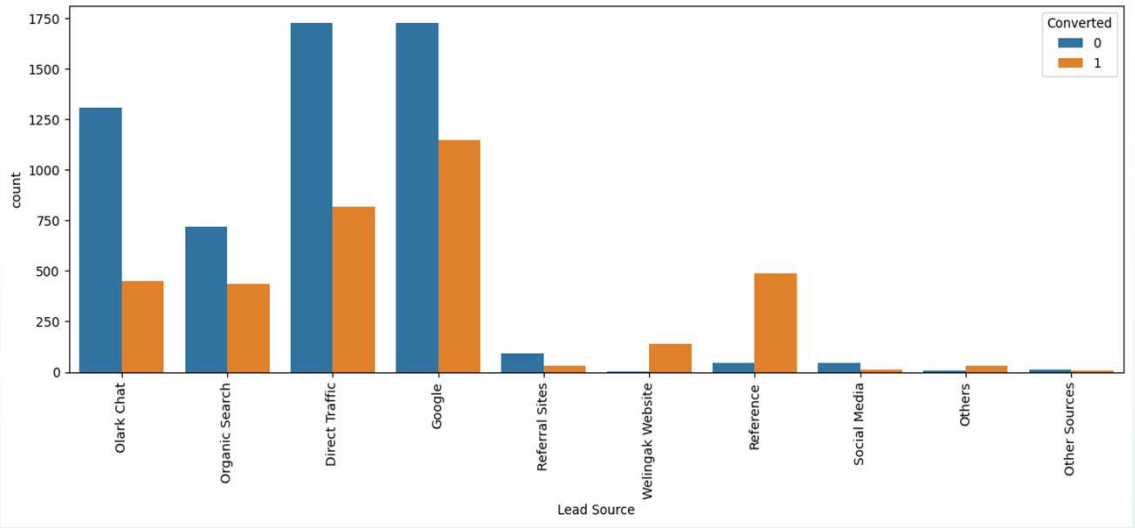


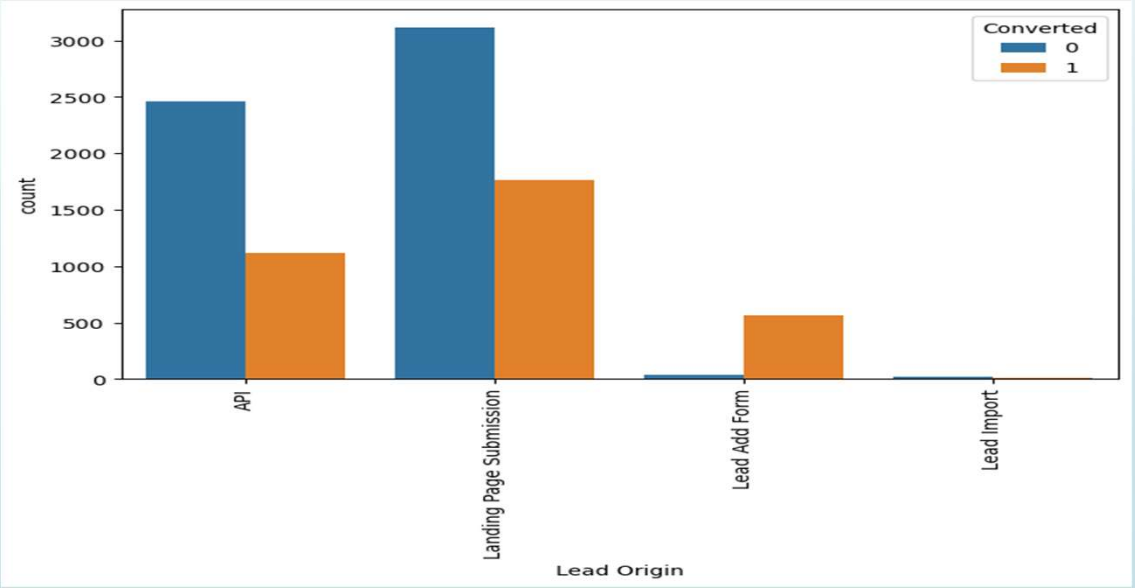
- Majority of the customer has not selected city but most of them converted to paid customers. Mumbai city also show high score of converted customers.



Exploratory Data Analysis Cont.

- Conversion rate from Google and Direct Traffic is high.
- Welingak Website, Organic Search, and Reference show low lead count but high conversion rate.
- Landing Page Submission and API has the highest number of leads and conversion rate.
- Lead Add Form has higher conversion rate but low lead count.

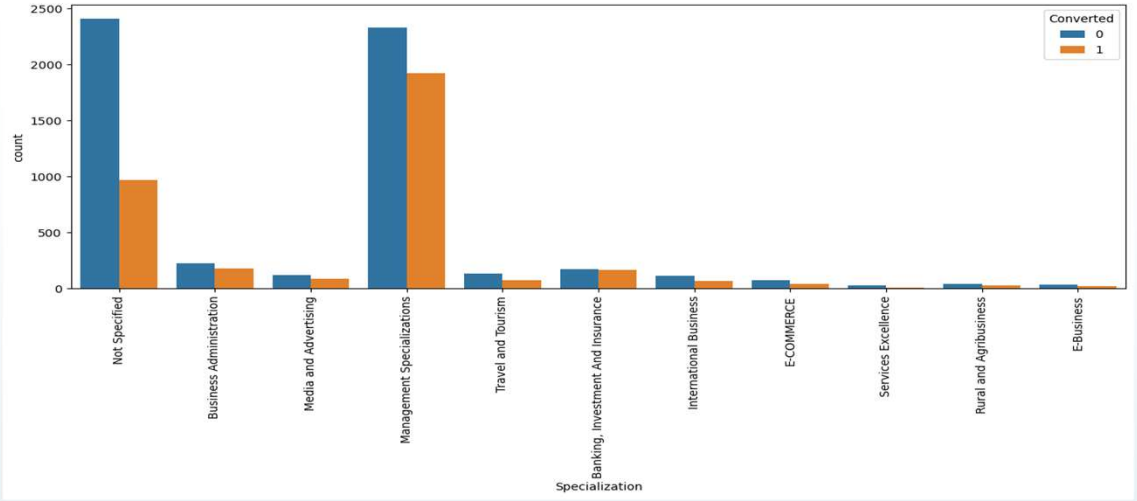


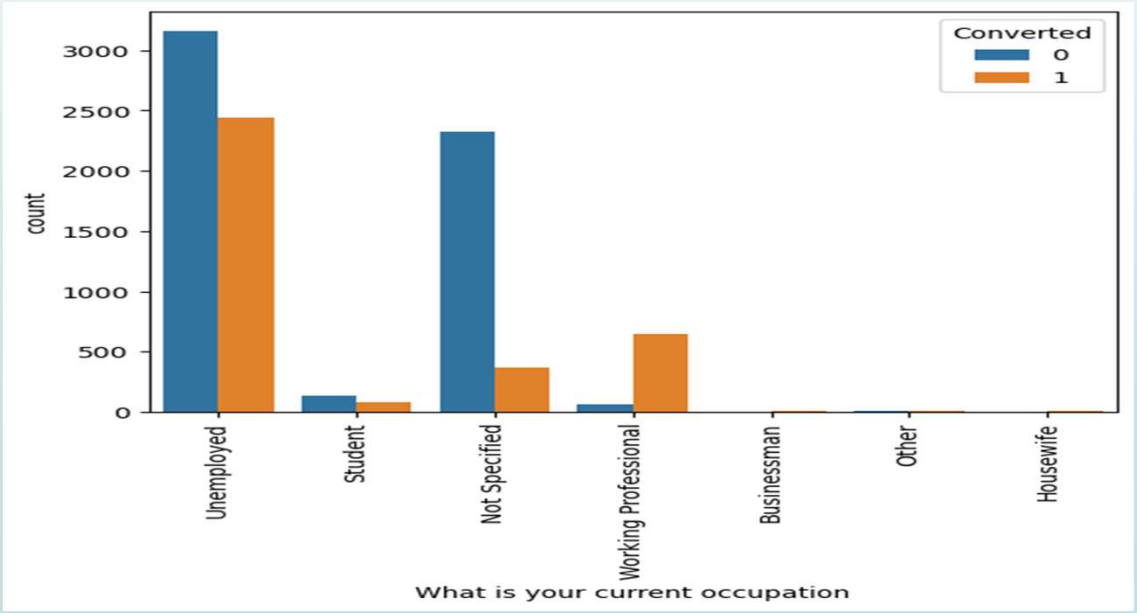


Exploratory Data Analysis Cont.

- After merging all specialization with management in them, we can observe that Management specialization has the highest paid customer conversion rate.

- Conversion rate of working profession is high but lead count is less. Whereas unemployed customers shows the highest conversion rate with more lead counts.

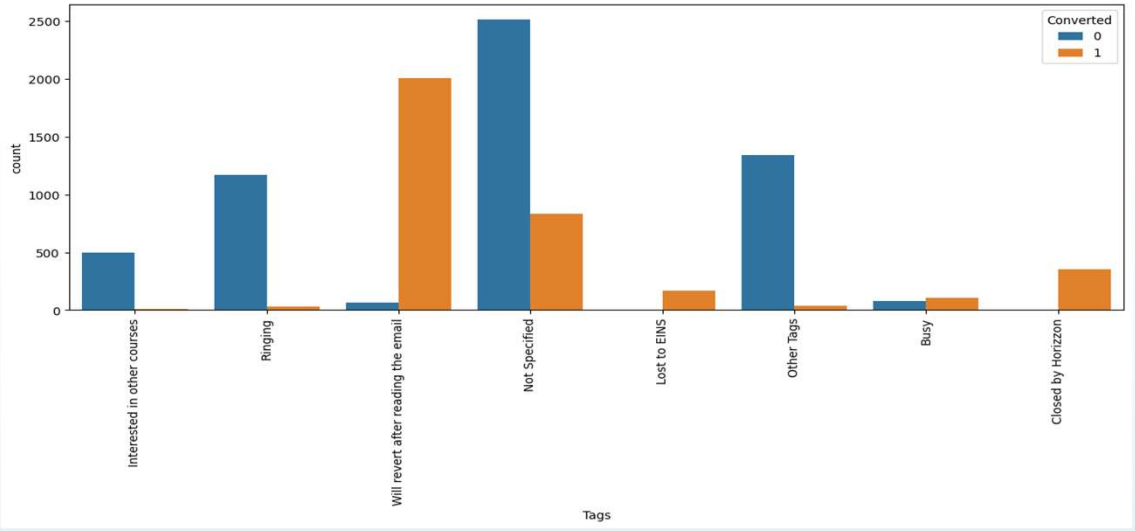


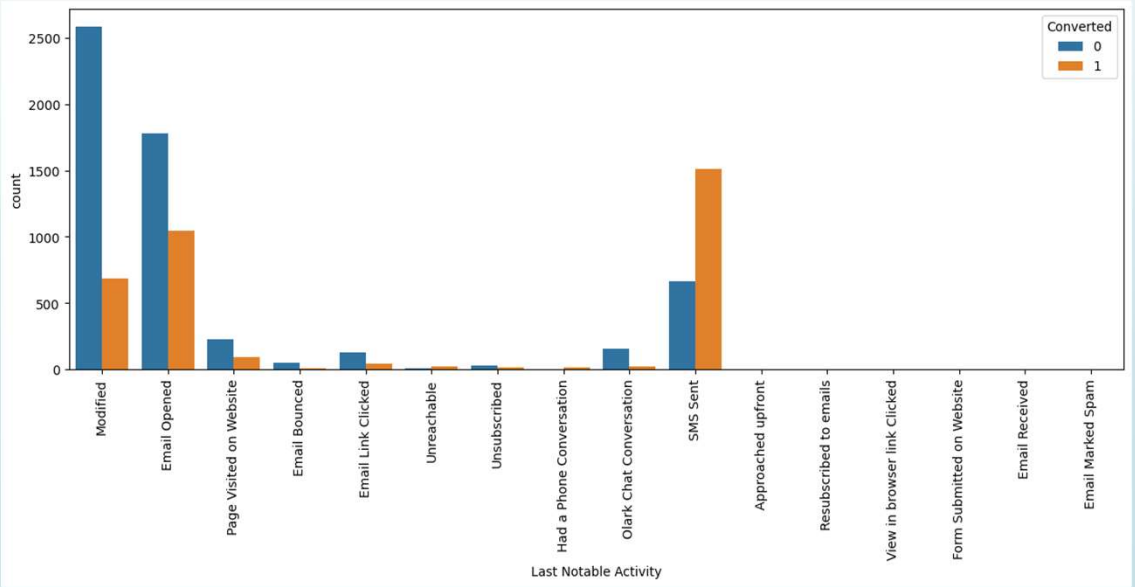


Exploratory Data Analysis Cont.

- Will revert after reading the email has the highest conversion rate but low lead counts. Whereas, not specified category has the highest lead count but moderate conversion rate.

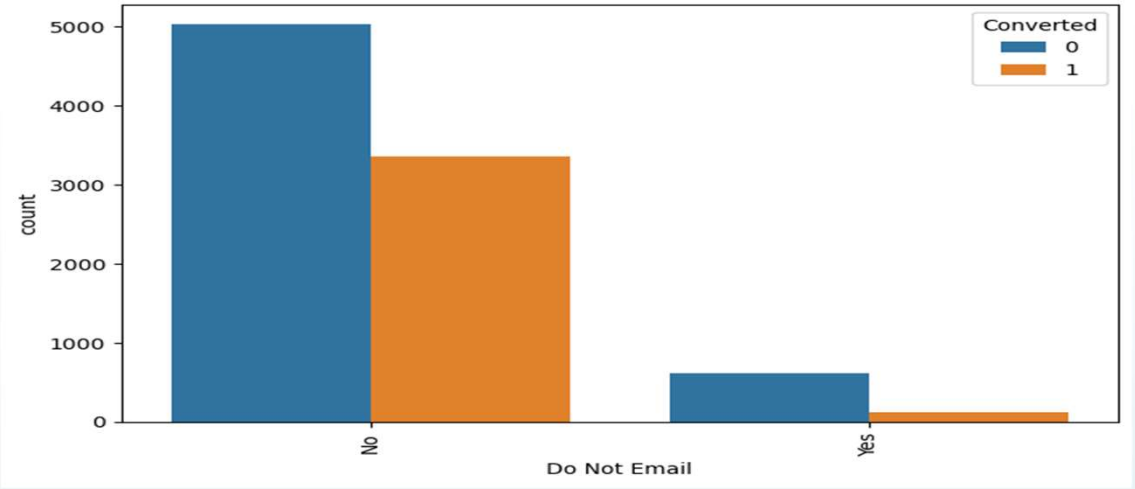
- SMS sent has the highest conversion rate but low lead counts. Email opened has higher lead counts and good conversion rate as well. It is better to focus on these two categories.



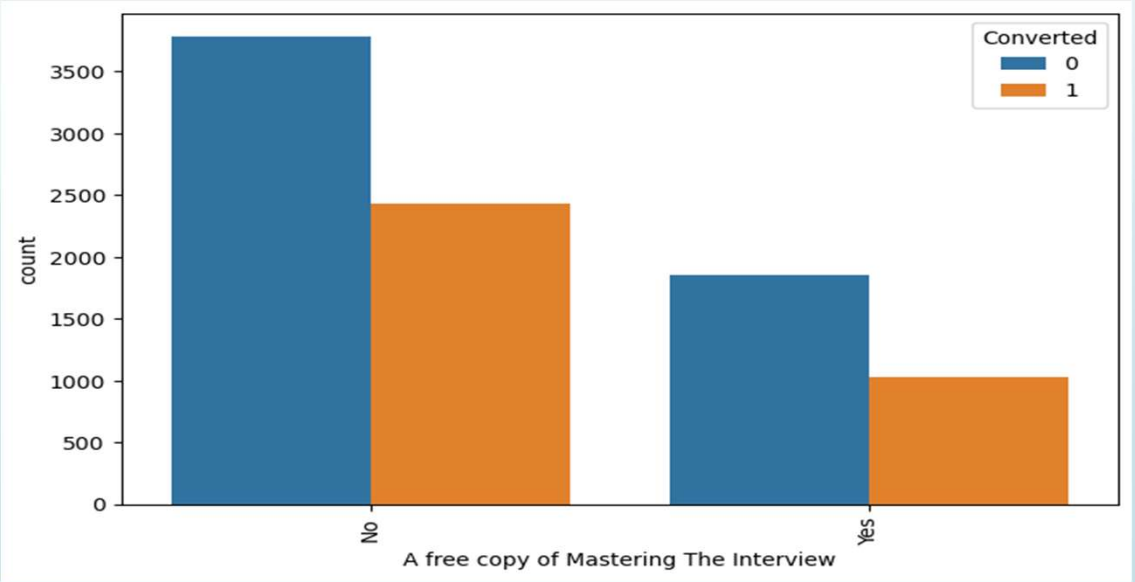


Exploratory Data Analysis Cont.

- Customer who opted to receive emails has the highest lead counts and higher conversion rate.

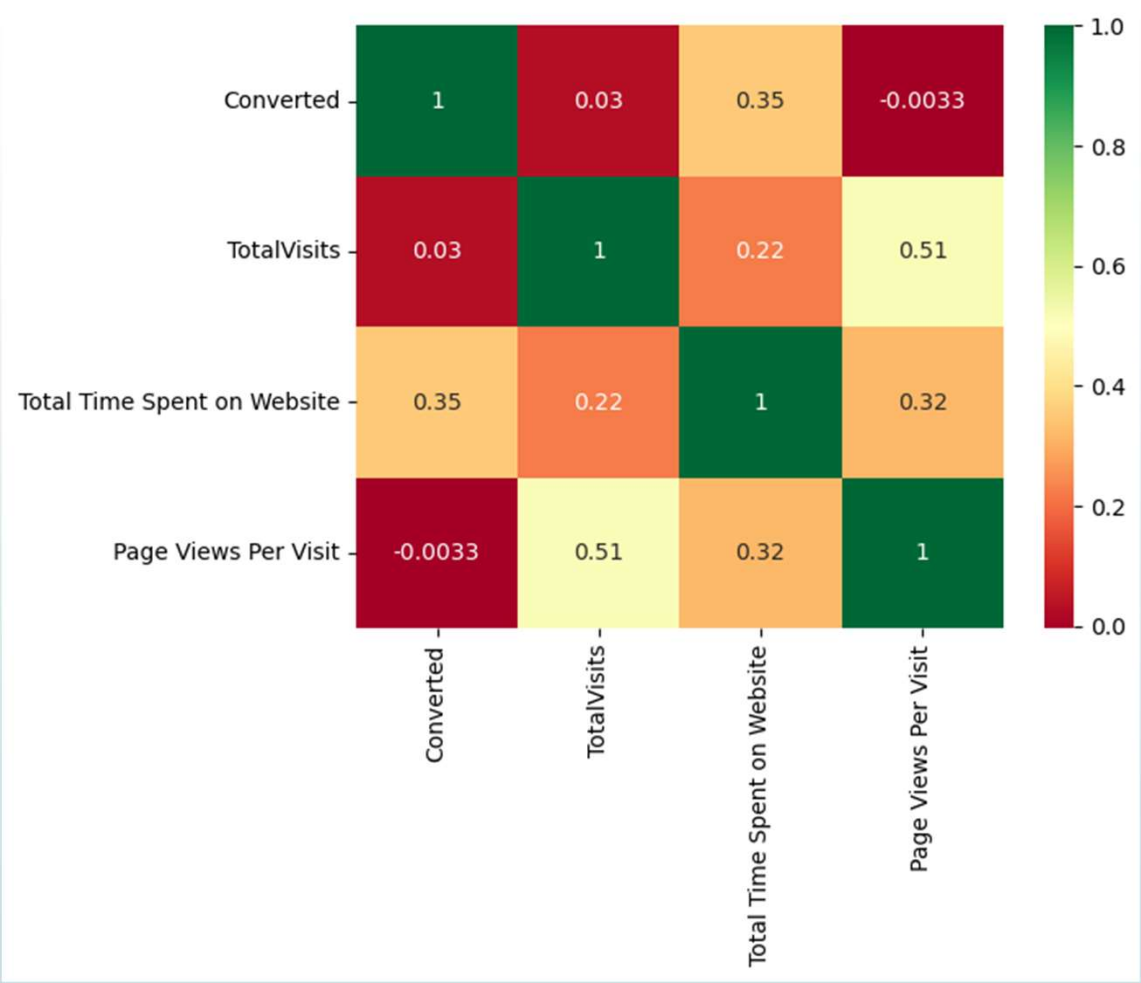


- Customers who did not opt to receive a free copy of Mastering the Interview has the highest lead count and higher conversion rate.



Exploratory Data Analysis Cont.

- There is no correlation found between the variables using heatmap.



Data Conversion

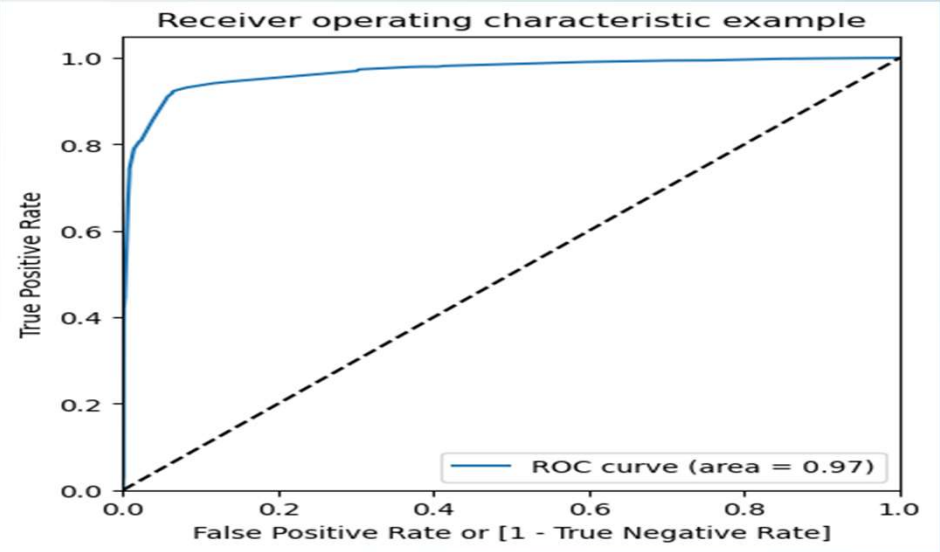
- Numeric Variables are normalized
- Created a function to remove top and bottom 1% of records for handling outliers
- Created a function for dummy variables of object data type variables
- Total rows for analysis: 8953
- Total columns for analysis: 64

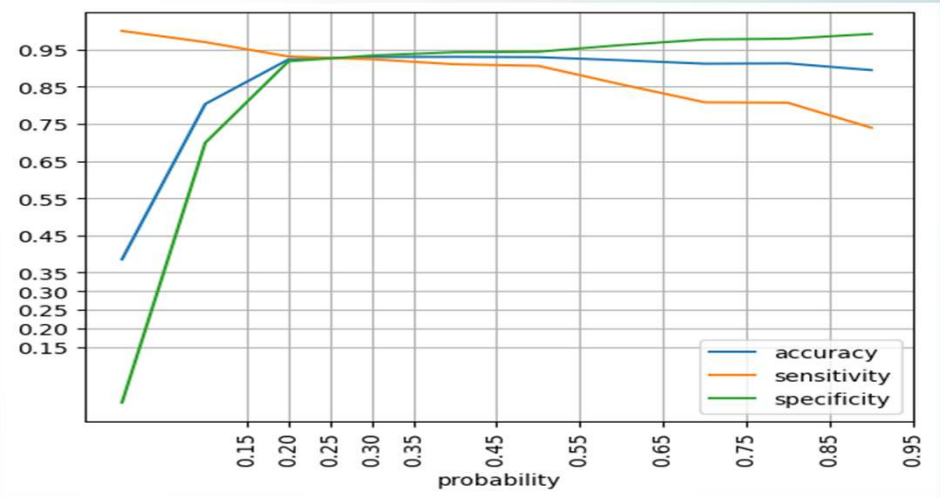
Model Building

- First step is to put “Converted” column to y variable and rest of the columns to X variable.
- Perform a train-test split with 70:30 ratio. Which means 70% data will be used to train the model and 30% of the data for testing the model.
- Utilizing recursive feature elimination (RFE) for feature selection
- Running RFE for 15 variables to select as output
- Building models by removing the variables with p value greater than 0.05 and variance inflation factor (VIF) is greater than 5
- Prediction on train dataset
- Model Accuracy of 93.11%, Sensitivity of 94.18%, and Specificity of 92.50%
- Model Precision score is 87.71% and Recall Score is 93.12%

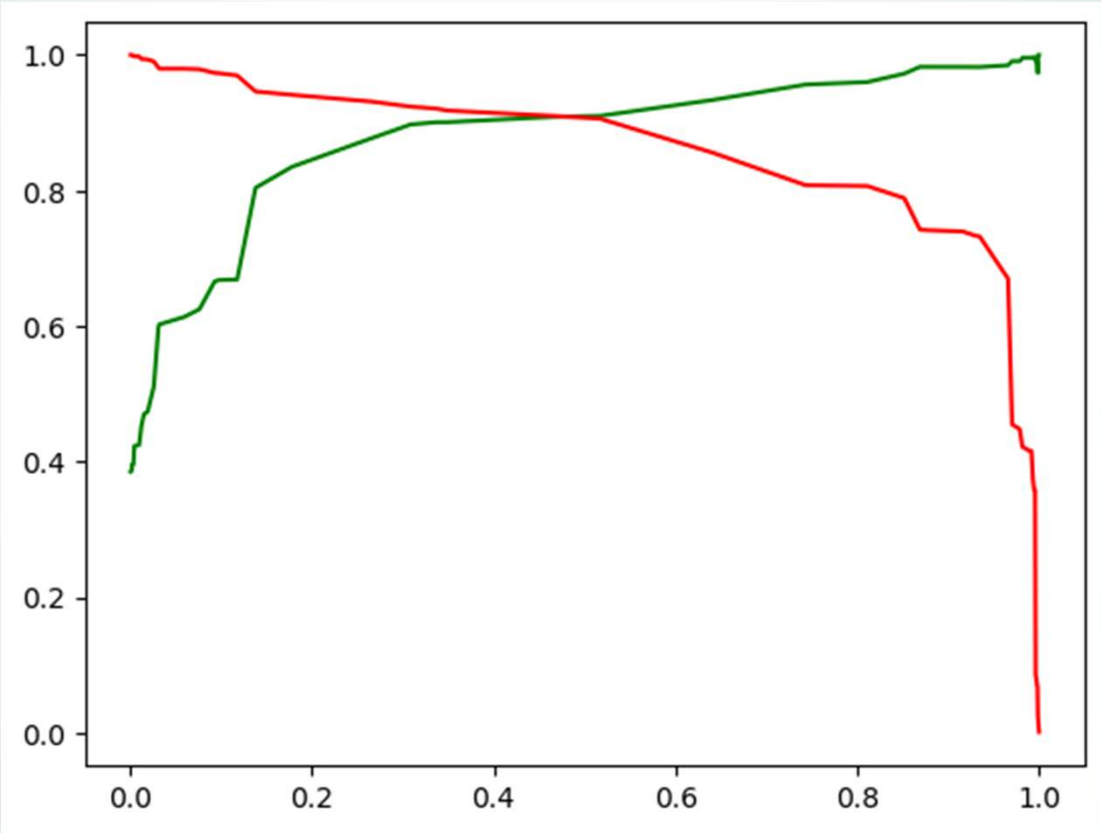
ROC Curve and Probability Curve

- The area under ROC curve is 0.97
- Finding optimal Cutoff Point
- Probability where we get balanced sensitivity and specificity
- From the second graph, it is visible that optimal cut off value is 0.25





Precision-Recall Trade Off



Conclusion

- It can be concluded that the variables that are essential for finding potential buyers are:
  - **Last Notable Activity:** When leads modify their applications, it shows their interest in buying the courses.
  - **Lead Origin:** Leads generated from Lead Add Form has higher conversion rate.
  - **Tags:** Leads that revert after the email are bound to be potential buyers.
  - **Current Occupation:** Leads who are working professionals or unemployed are proven to be potential buyers.
  - **Lead Source:** Leads from Welingak Website, Olark Chat, Reference, Google, and Direct Traffic needs to be targeted on priority.

Thank You!