LEAD SCORING CASE STUDY

- PRANAV REDDY, MOHAMMED SAIF, MEGHANA REDDY

PROBLEM STATEMENT

The problem at hand is to identify and prioritize the leads that are most likely to convert into paying customers. This requires developing a model that assigns a lead score to each lead, where higher scores correspond to higher chances of conversion. X Education's current lead conversion rate is about 30%, and the CEO expects a target conversion rate of around 80% by focusing on high-potential leads. The data provided contains various attributes related to leads, such as source, time spent on the website, visits, and last activity, with a target variable indicating whether the lead was converted.

MODEL BUILDING APPROACH

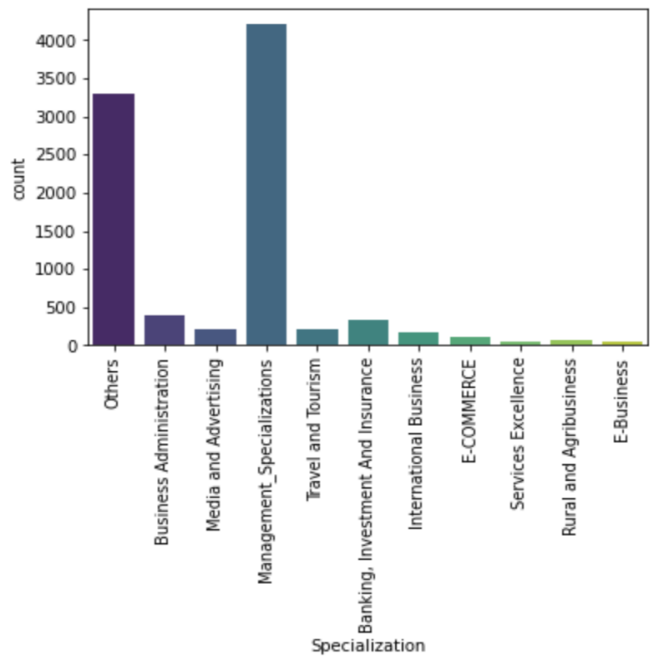
- Importing Data
- Data Inspection
- Data cleaning
- Exploratory Data Analysis
- Creating Dummy Variables for the Categorical Variables
- Model Building Using Logistic Regression
- Prediction On Test Dataset
- Assigning Lead Score with respect to Lead_Num_ID
- Finding out the Hot Leads which should be contacted
- Conclusions

Exploratory Data Analysis (EDA)

1. Univariate Analysis (Categorical)
2. Bivariate Analysis (Categorical)
3. Numerical Variable Analysis

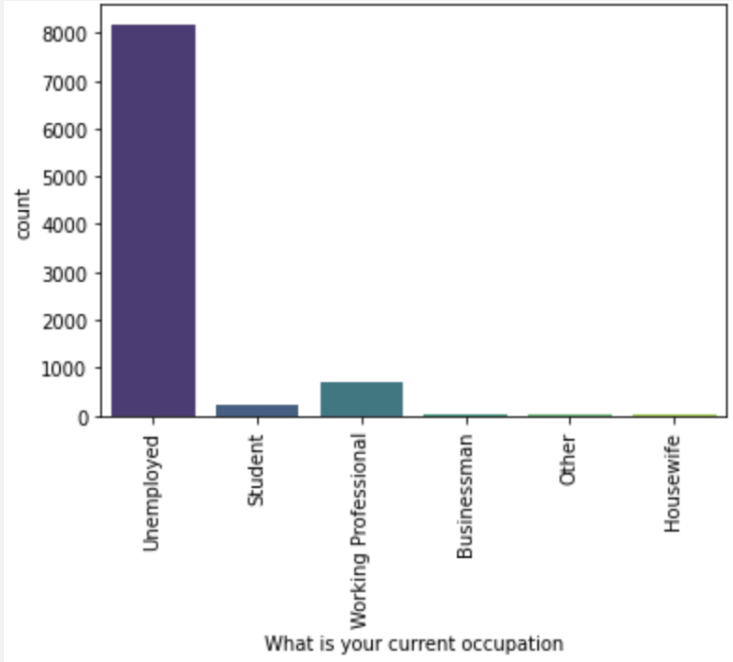
UNIVARIATE ANALYSIS (CATEGORICAL)

Specialization



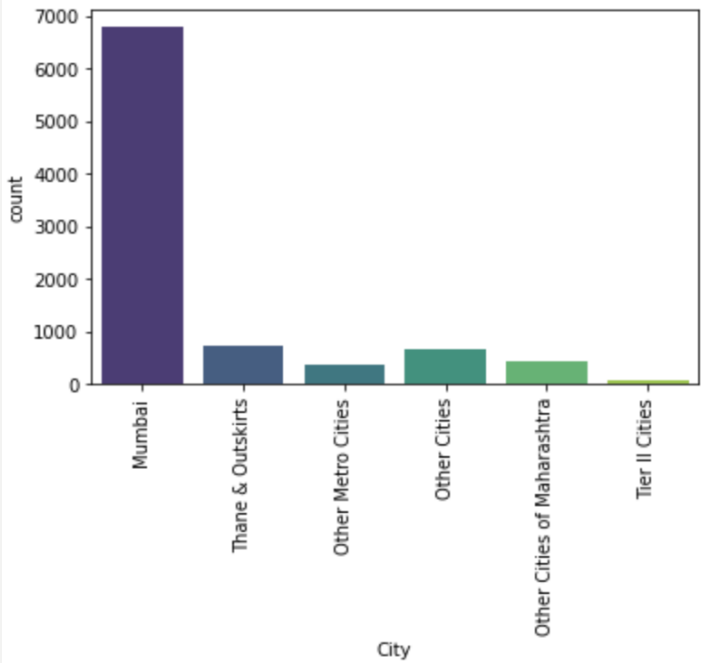
A large number of customers have their specialization listed as "Others" or "Management Specializations."

What is your current occupation



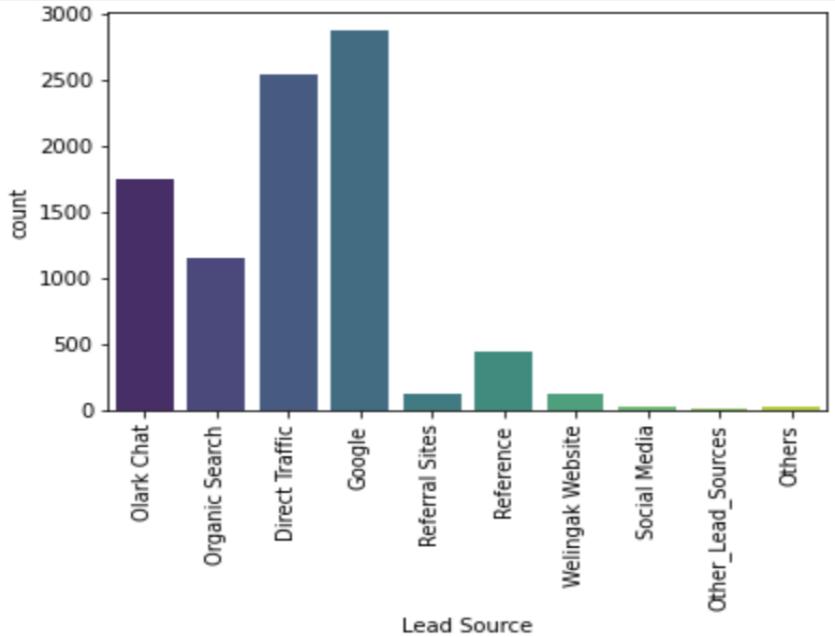
Most customers are categorized as unemployed.

City



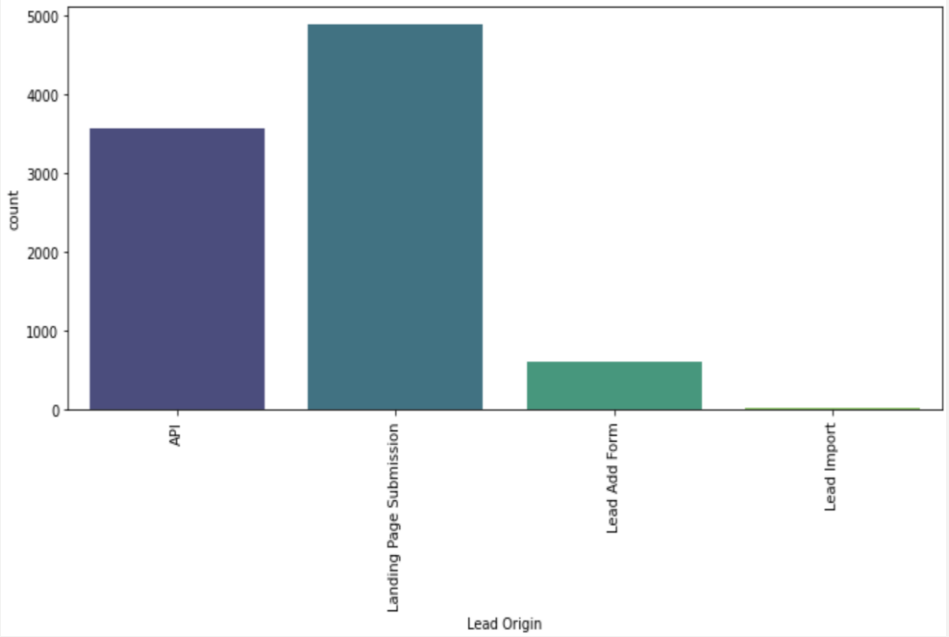
Majority of the customers are from Mumbai city.

Lead Source



Google has the highest number of leads as the source, which is logical since most people primarily use Google as their search engine.

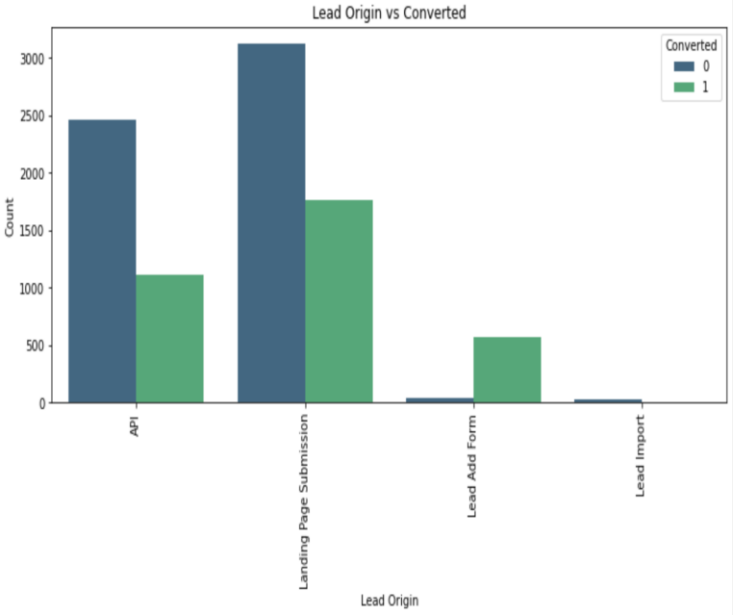
Lead Origin



The majority of leads were generated through "Landing Page Submission" and "API" origins.

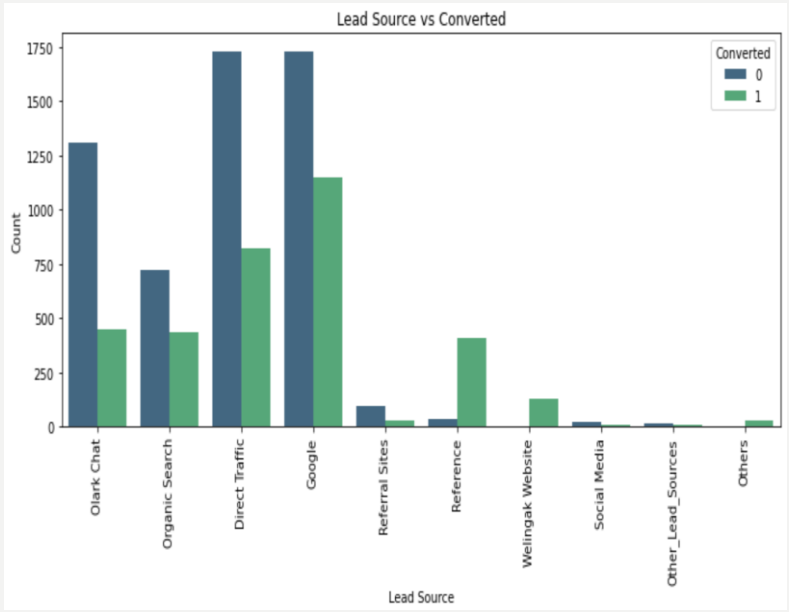
BIVARIATE ANALYSIS (CATEGORICAL)

Lead Origin Vs Converted



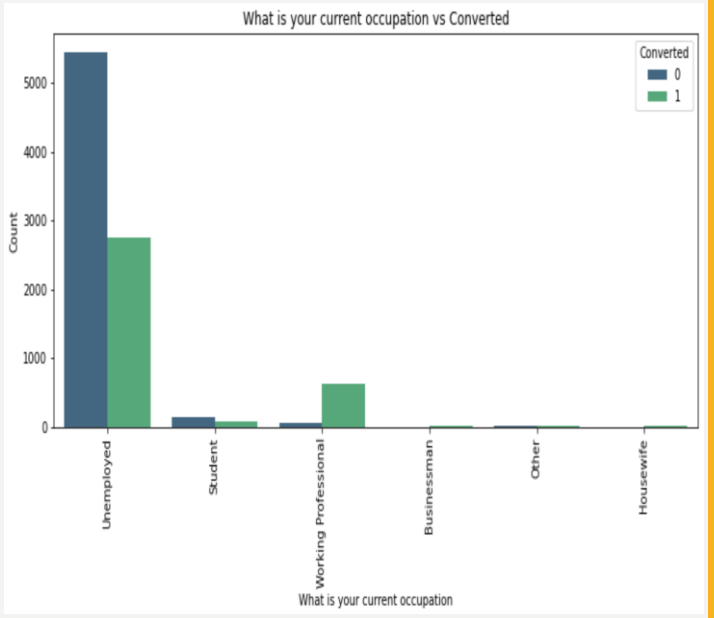
Leads originating from the Lead Add Form have a higher conversion rate compared to those from API and Landing Page Submission.

Lead Source Vs Converted



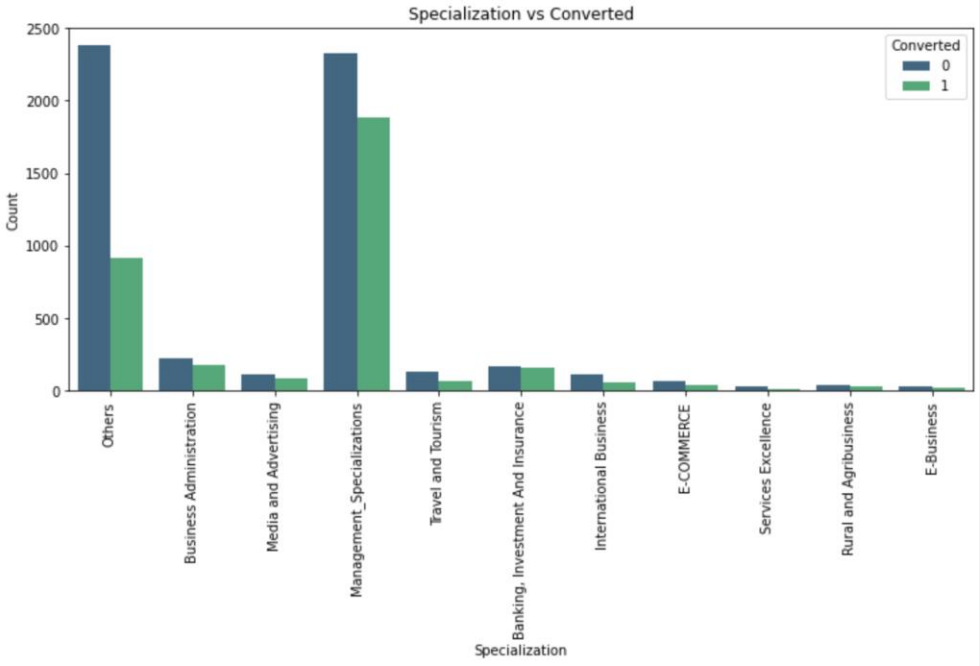
The conversion rate is highest for leads from the Reference Source, while those from Google, Direct Traffic, and Olark Chat have relatively lower conversion rates.

What is your current occupation Vs Converted



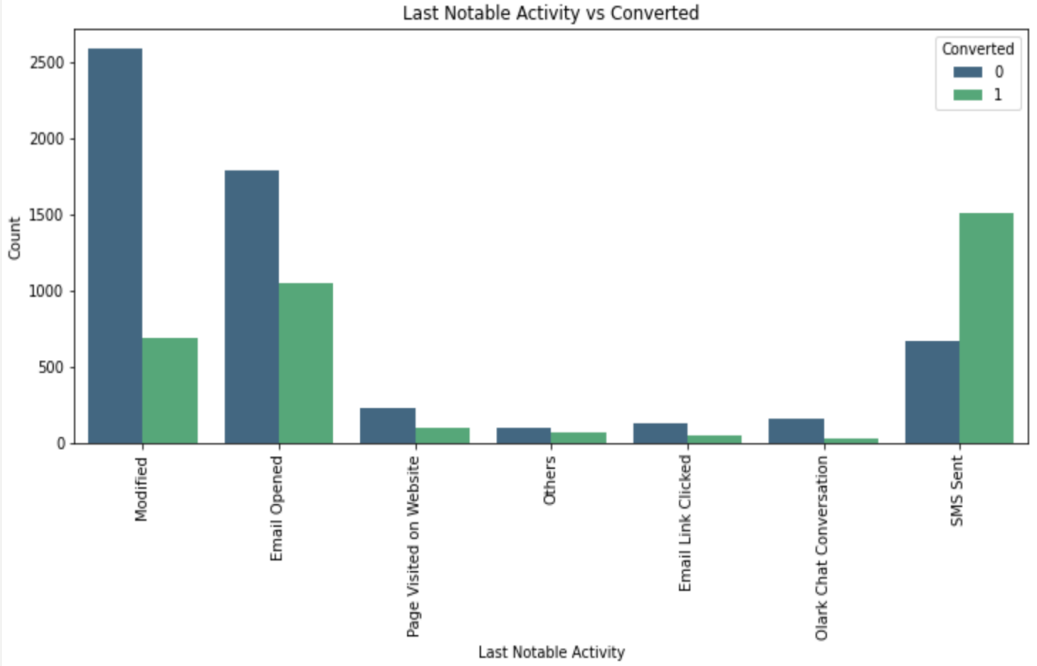
Working professionals have a higher conversion rate than unemployed individuals, likely because they are more aware of current market demands.

Specialization Vs Converted



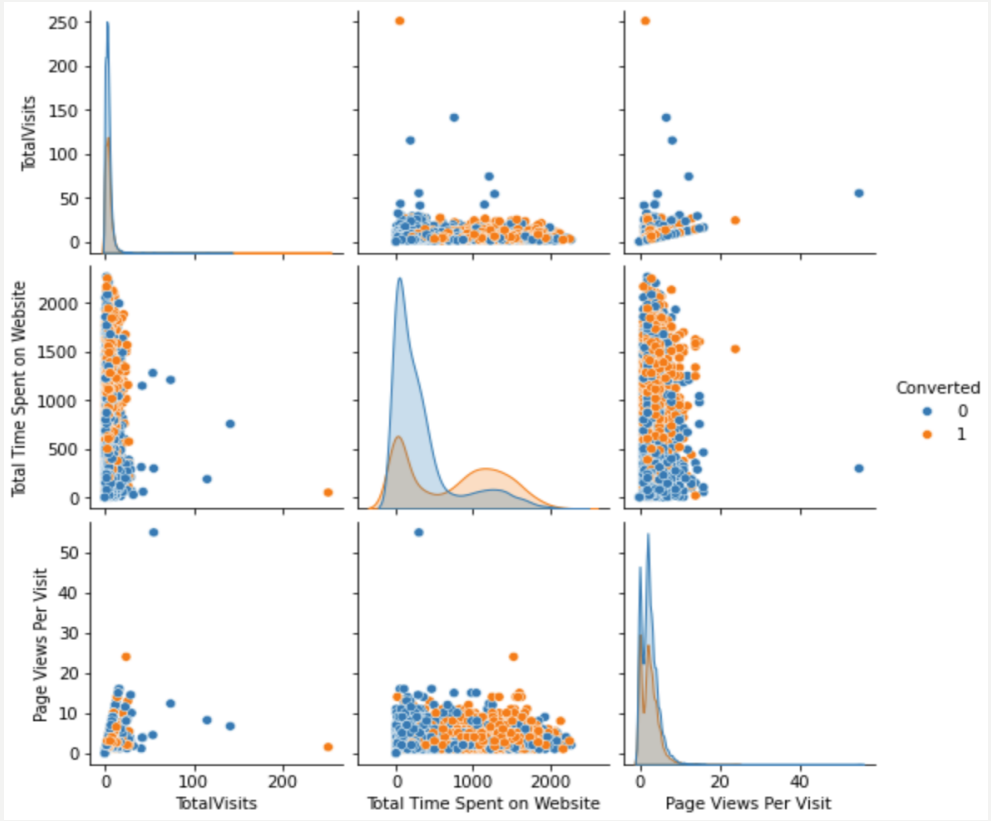
Google has the highest number of leads as the source, which is logical since most people primarily use Google as their search engine.

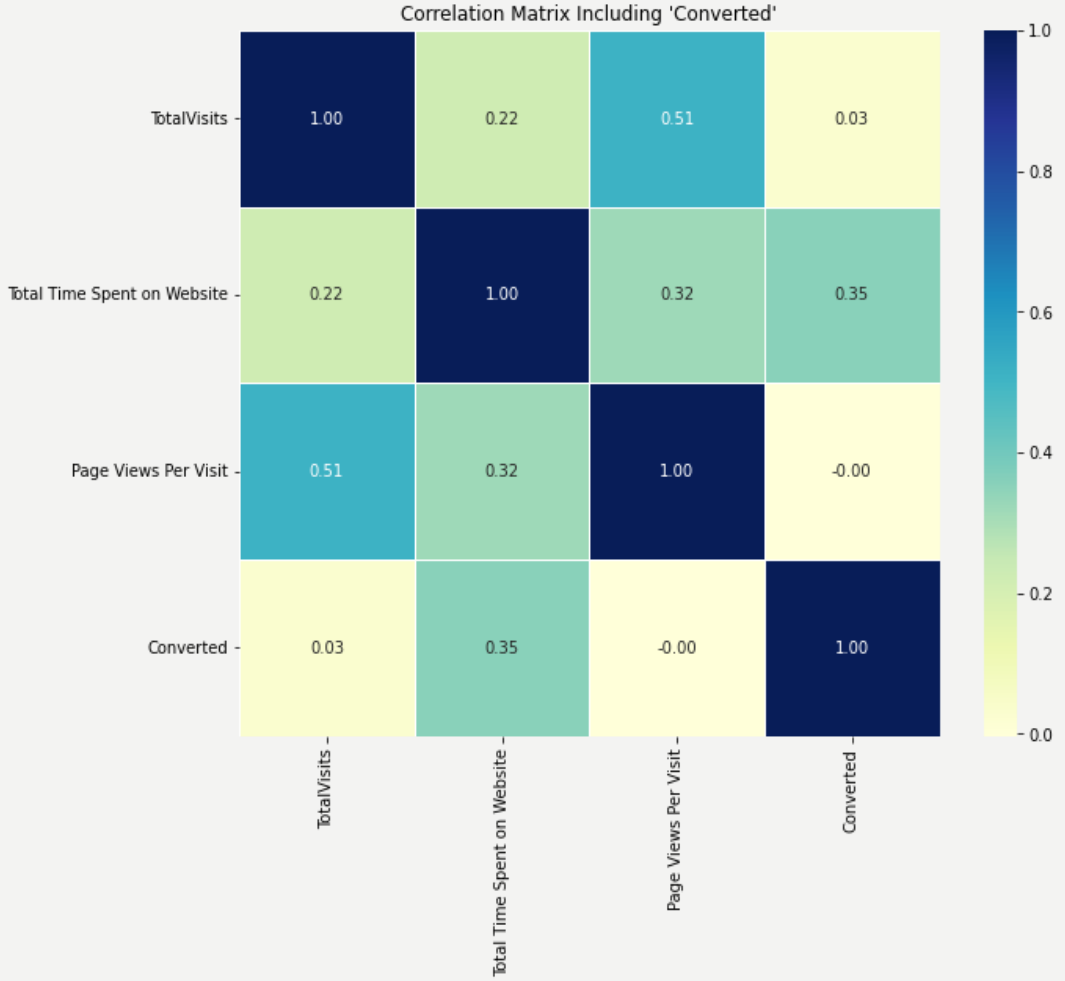
Last Notable Activity Vs Converted



SMS Sent has a higher conversion rate compared to Email Opened.

NUMERICAL VARIABLE ANALYSIS

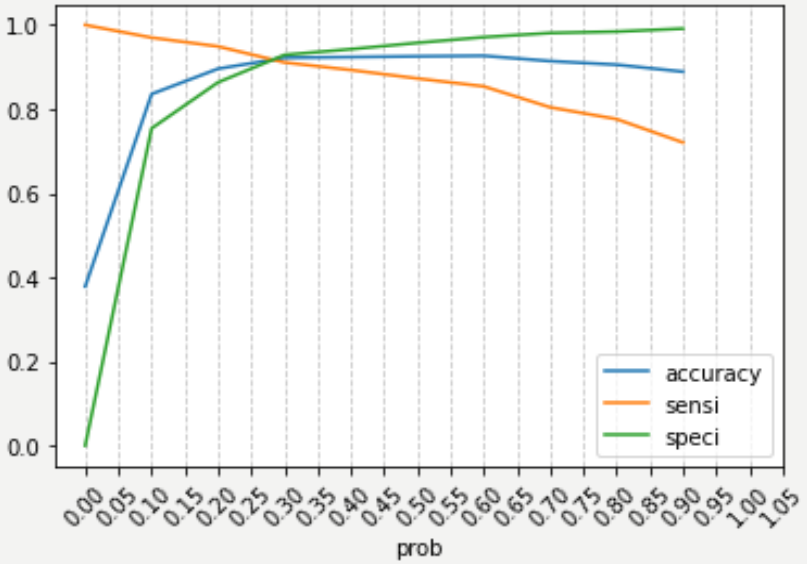


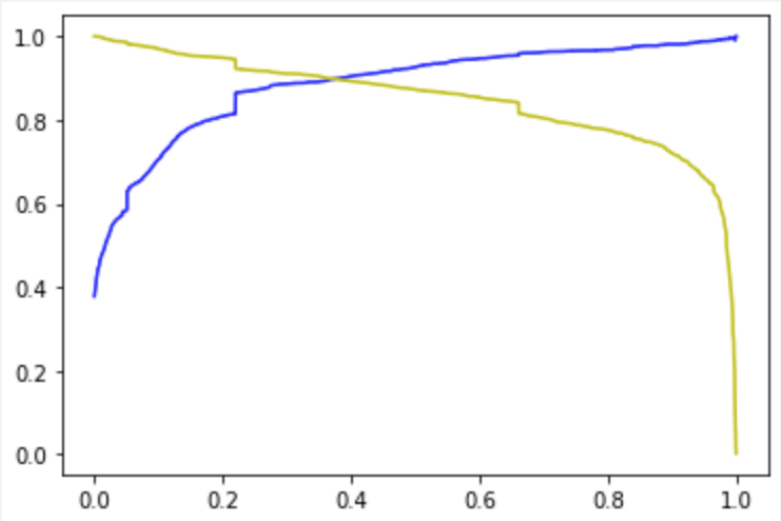


- **Total Time Spent on Website is the most significant factor influencing conversion, with higher times correlating positively.**
- **Moderate engagement in Page Views Per Visit also increases conversion likelihood.**
- **Extremely high TotalVisits or Page Views Per Visit do not contribute to conversions and may indicate inefficiency in user engagement or browsing behavior.**

WE CAN SEE THAT 'TOTAL VISITS' AND 'PAGE VIEWS PER VISIT' HAVE THE MOST CORRELATION WITH EACH OTHER.

ROC CURVE & OPTIMAL CUTOFF

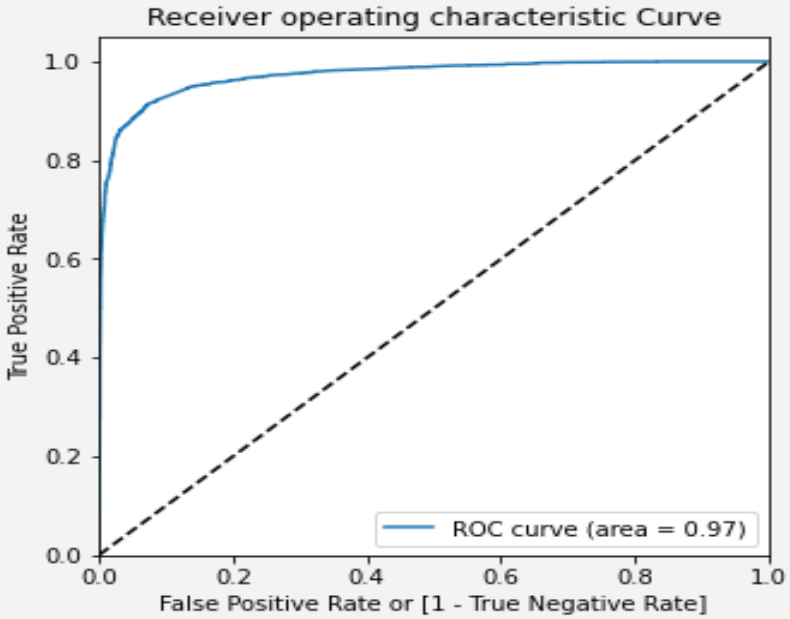




Recall Vs Precision

- **ROC-AUC Score:** 0.97
- **Optimal Cutoff:** 0.27 (Intersection point of Accuracy, Sensitivity and Specificity curves)

- ROC curve should be a value closer to 1 for a good model. We have got a value of 0.97 which is extremely good



MODEL EVALUATION

Training Data Confusion Matrix			
		Predicted	
Actual		Negative	Positive
	Negative	3621 (TN)	308 (FP)
	Positive	199 (FN)	2191 (TP)

Testing Data Confusion Matrix			
		Predicted	
Actual		Negative	Positive
	Negative	1546 (TN)	125 (FP)
	Positive	74 (FN)	964 (TP)

- CEO of X-Education has given a target of at least 80% and our model has achieved accuracy of ~92%.

Evaluation Metrics	Train Dataset	Test Dataset
Accuracy	91.98	92.65
Specificity	92.16	92.52
Precision	87.68	88.52
Recall	91.67	92.87

BUSINESS IMPLICATIONS & RECOMMENDATIONS

- Actionable Insights: Focus on high-scoring leads to maximize conversions.
- Allocate sales resources efficiently.
- Impact: Improved conversion rates and reduced time spent on low-priority leads.
- Use lead scoring in CRM systems.
- Regularly update the model with new data for better accuracy.
- Train sales teams to prioritize based on lead scores.

Conclusions:

According to final model, the variables that are important for verifying the Hot Leads are:

Tags_Closed by Horizzon	7.204061
Tags_Lost to EINS	5.711259
Tags_Will revert after reading the email	4.575609
Lead Source_Welingak Website	4.081473
Lead Origin_Lead Add Form	2.146431
Last Activity_SMS Sent	1.934491
Lead Source_Olark Chat	1.284689
Total Time Spent on Website	1.069427

- Tags_Closed by Horizzon
- Tags_Lost to EINS
- Tags_Will revert after reading the email
- Lead Source_Welingak Website
- Lead Origin_Lead Add Form
- Last Activity_SMS Sent
- Lead Source_Olark Chat
- Total Time Spent on Website

THANK YOU