# Lead Scoring Case Study

## Problem Statement

- X Education is an online education company that provides courses to industry professionals. The company employs marketing channels, such as Google search engines and websites, to attract leads to its platform.

- Once on the platform, leads can browse through the available courses or fill out a form to learn more information. These leads can also come through referrals.

- The sales team at X Education works to convert the acquired leads into paying customers. However, despite many leads, the company's conversion rate is only around 30%. In order to increase the conversion rate, X Education needs a model to prioritize leads for the sales team to focus on. The CEO has set a target conversion rate of 80%.

## Objectives

- To achieve the CEO's goal of an 80% conversion rate, X Education needs a lead scoring model. The model will assign a score between 0 and 100 to each lead based on their likelihood of conversion. Leads with higher scores will be considered more promising and will be prioritized by the sales team. The model will take into account data such as browsing behaviour, form fill-ups, video views, and past referrals to calculate the score.

- The model will be built using logistic regression, and will provide a way for the sales team to focus their efforts on leads that are most likely to convert. By prioritizing the leads with higher scores, the company hopes to increase the conversion rate and achieve the CEO's target.

- A higher score would mean that the lead is a hot lead, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted

# Methodology

1. Data Understanding, Cleaning and Preparation

- Reading the data and looking for duplicates if any
- Looking at the info and statistical summary of the data
- Handling missing values and treating outliers
- Analysing Variables and the behaviour of variables with target variable using visualization

2. Train - Test split and scaling

3. Model Building

- Coarse Tuning using RFE
- Manual Fine Tuning using VIF and p-value

## Methodology

4. Model Evaluation

- Accuracy
- Sensitivity and Specificity
- Threshold determination using ROC
- Precision and Recall

5. Predictions on the Test set

## Data Manipulation

- The dataset contains 9240 rows and 37 columns with no observed duplicates.
- We dropped the 'Prospect ID' and 'Lead Number' columns as they only serve as unique identifiers. We also replaced 'Select' values with np. nan as they represent missing values.

- We dropped columns with more than 45% missing values. For variables with high percentages of missing values, we imputed them with 'Not Specified'. For variables with low percentages of missing values, we imputed them with the mode.

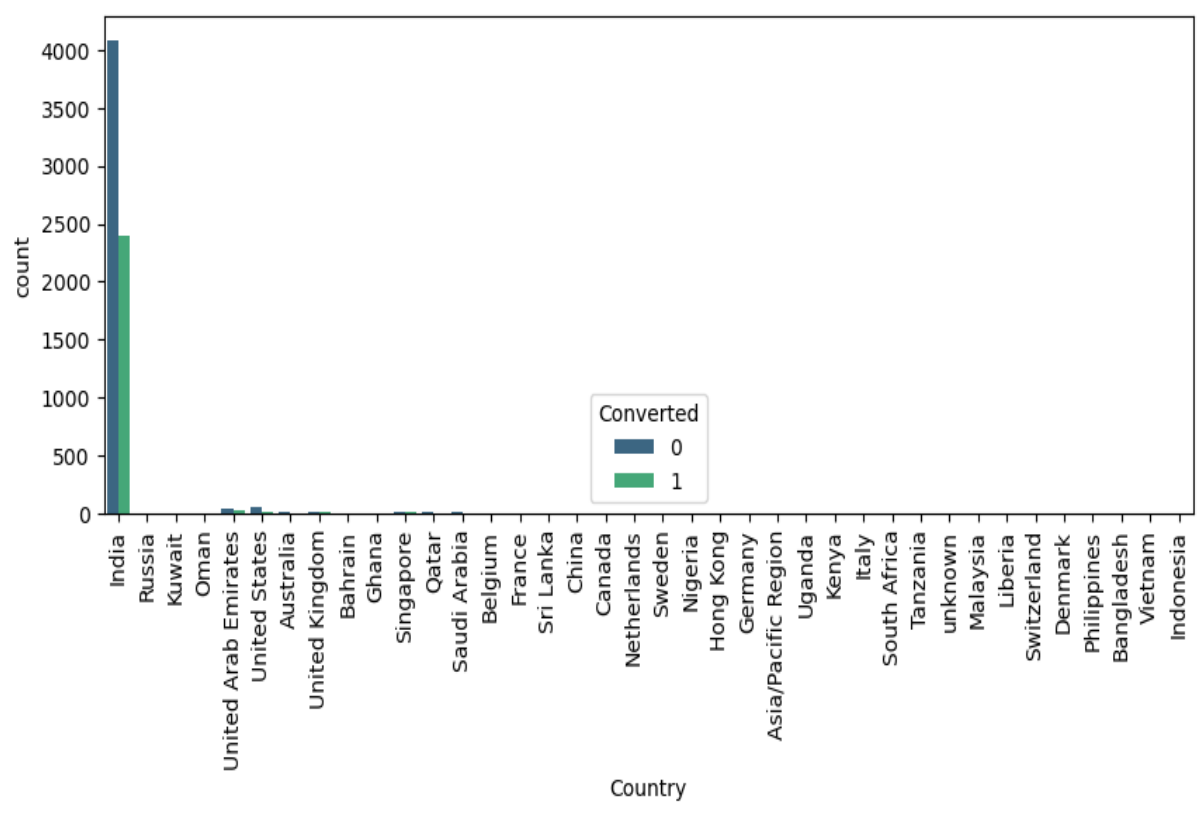
- To handle low-frequency values in a variable, we created a separate category called 'Others'. We also removed variables where most of the values were 'No' as they were not helpful in our analysis.

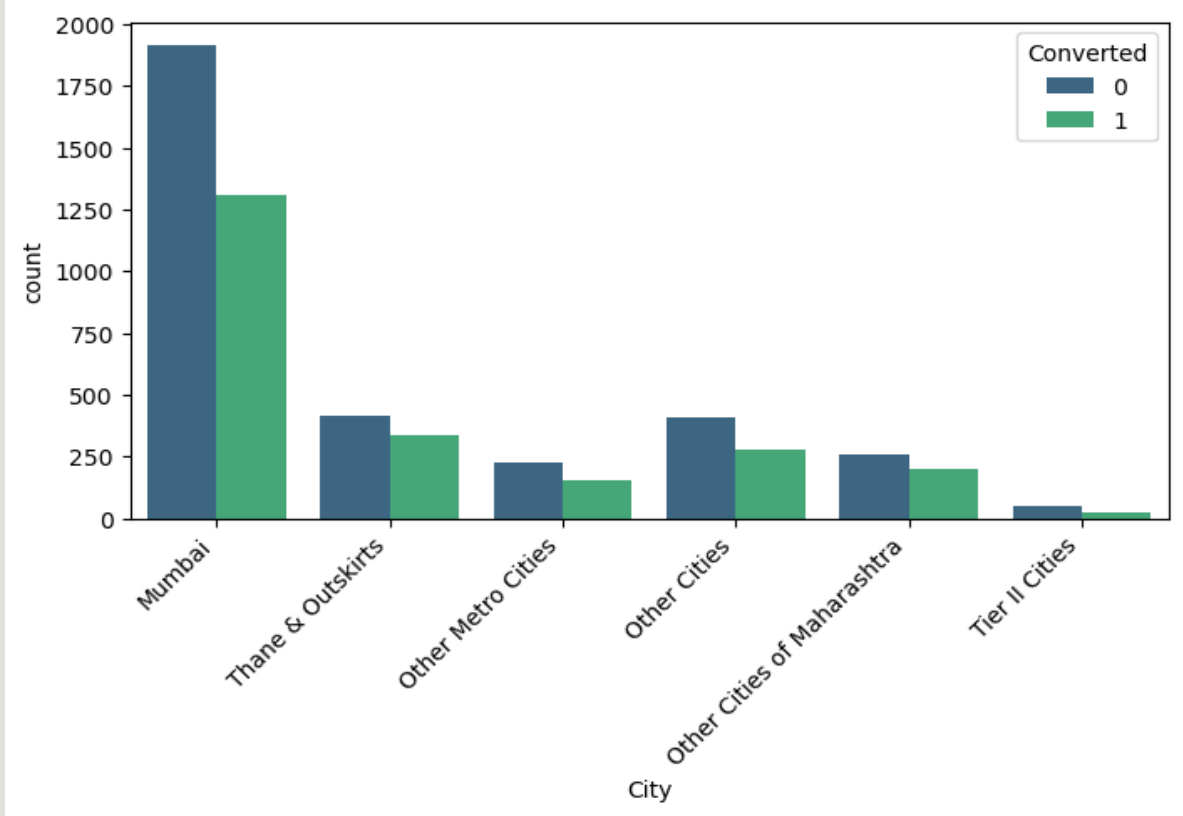
- For numerical columns, we handled outliers using the capping method. The conversion rate was 38%.

- Overall, we have taken the necessary steps to ensure data quality and improve the accuracy of our analysis.

# EDA

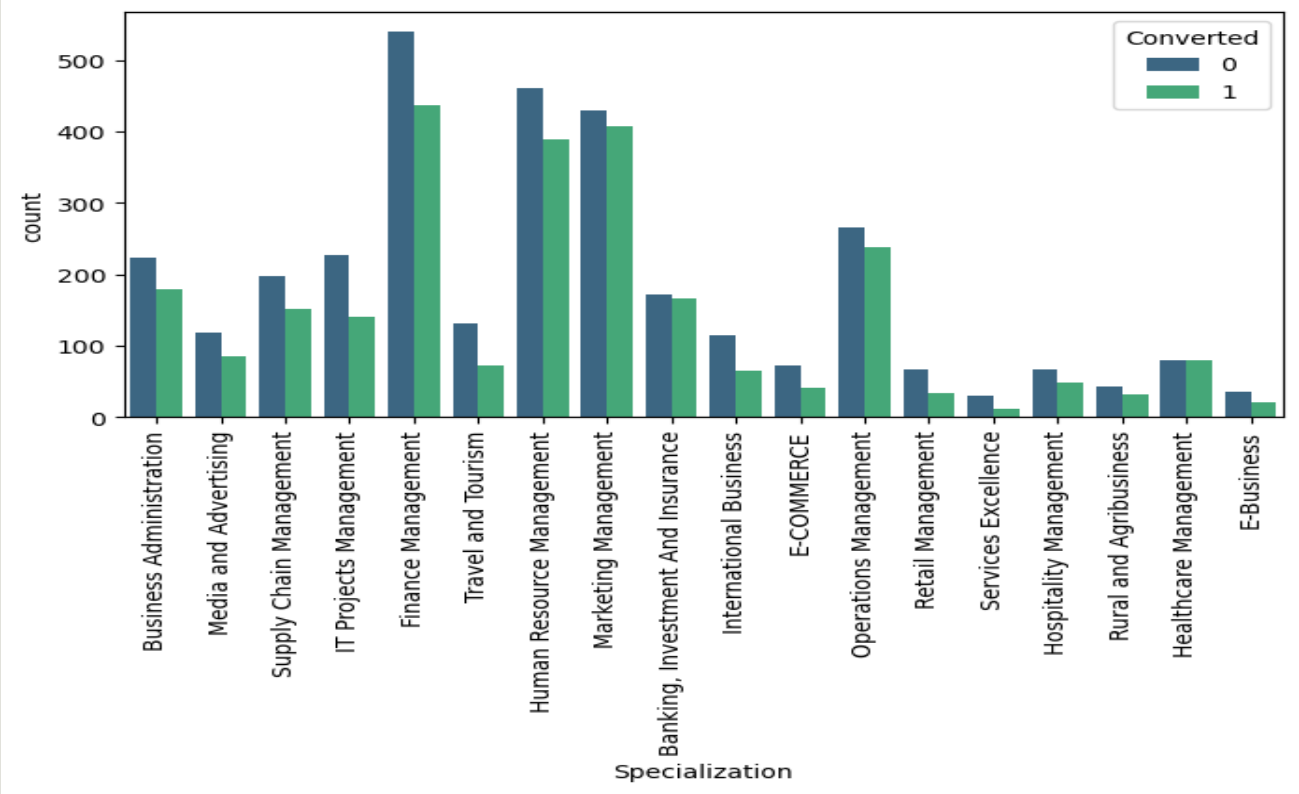
Analyzing ‘Country’ and ‘City’ variables

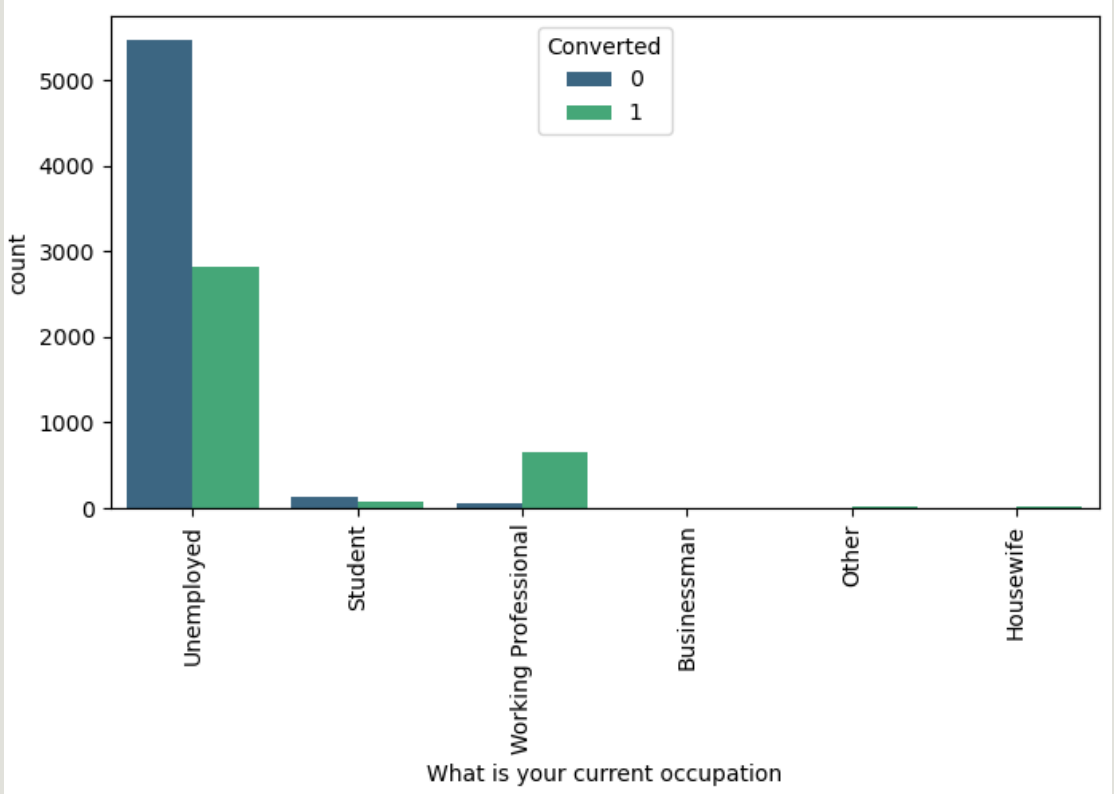



- We observed that the majority of visitors/leads in the dataset were from India, with Mumbai being the most common city. As a result, we decided to drop the 'country' variable from our analysis as we deemed it insignificant for our purposes.

# Analysing 'Specialization' and ‘What is your current occupation’ variables

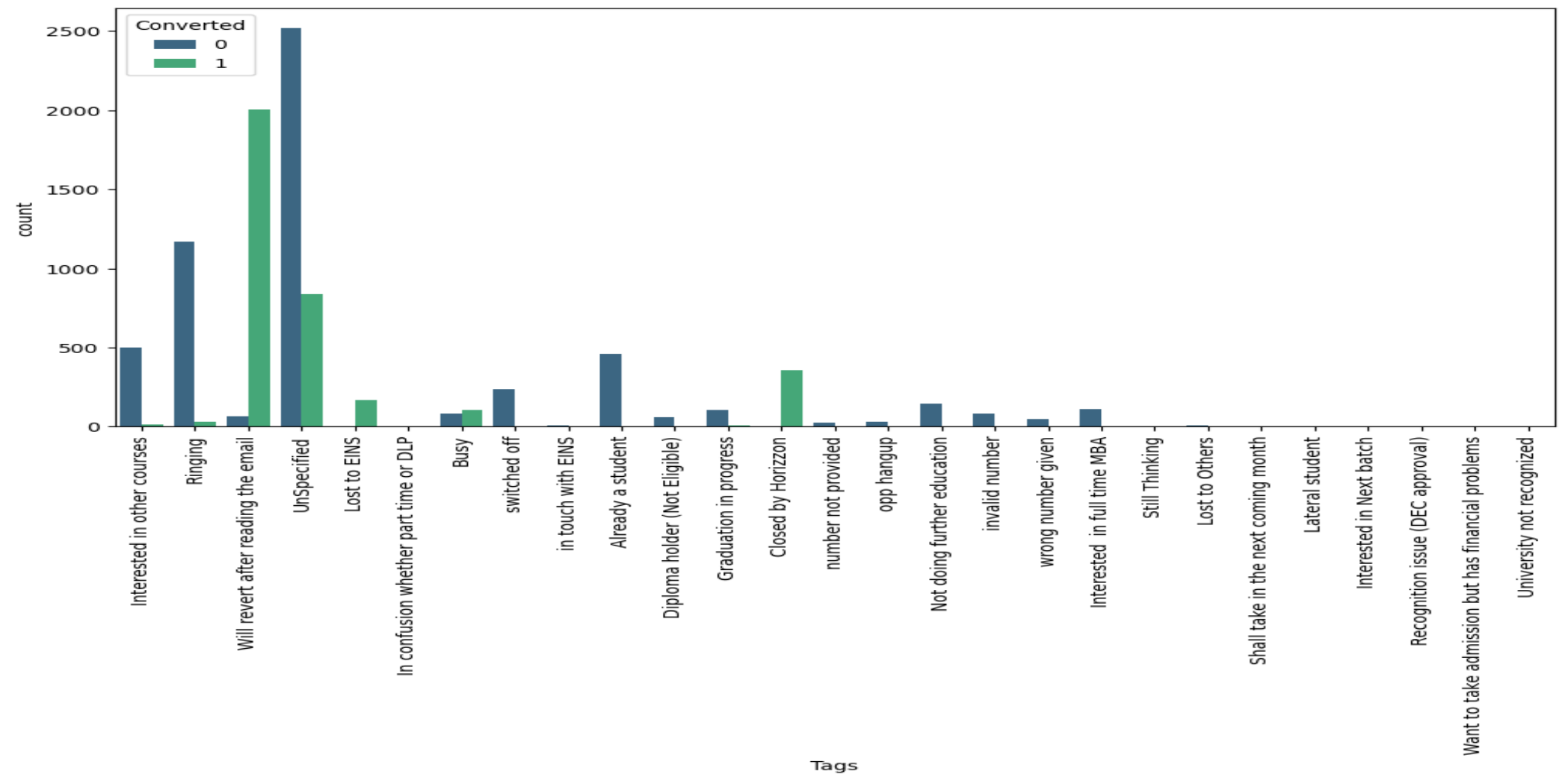



- There are multiple categories in ‘Specialization’ which is related to the 'management' profile. Hence, clubbed those into a single category called management

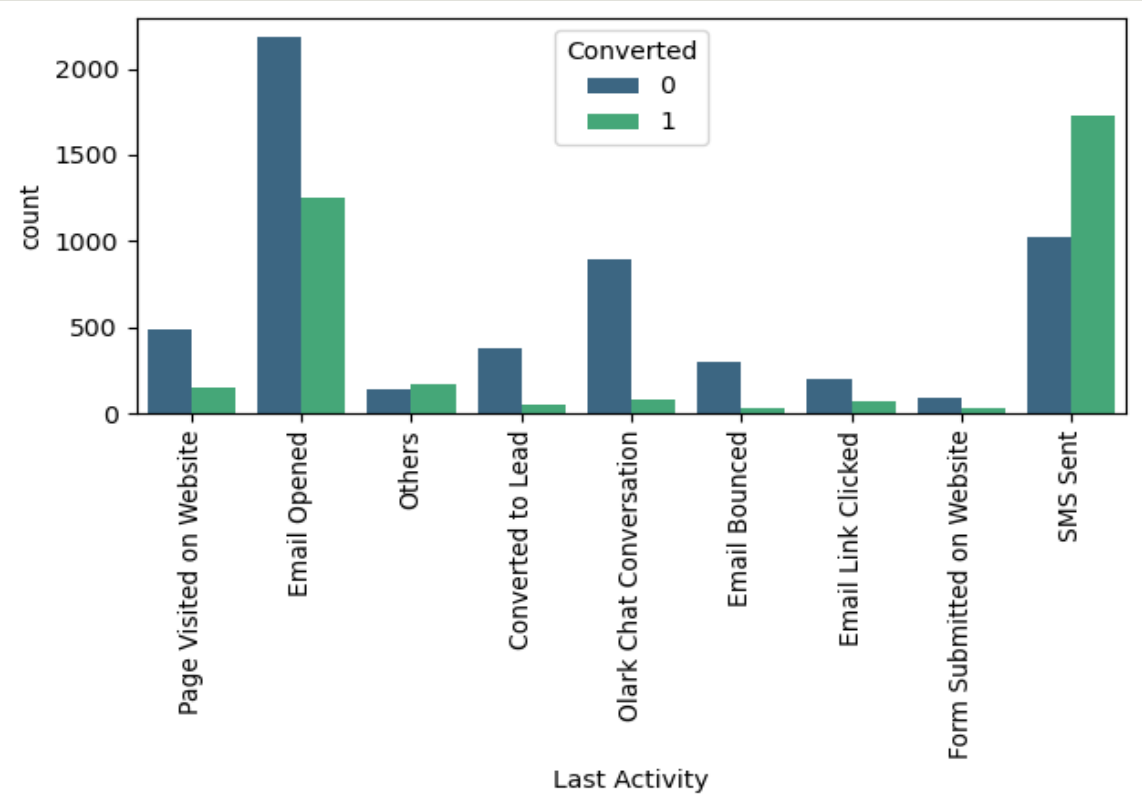
- Working Professionals are more likely to get converted into learners

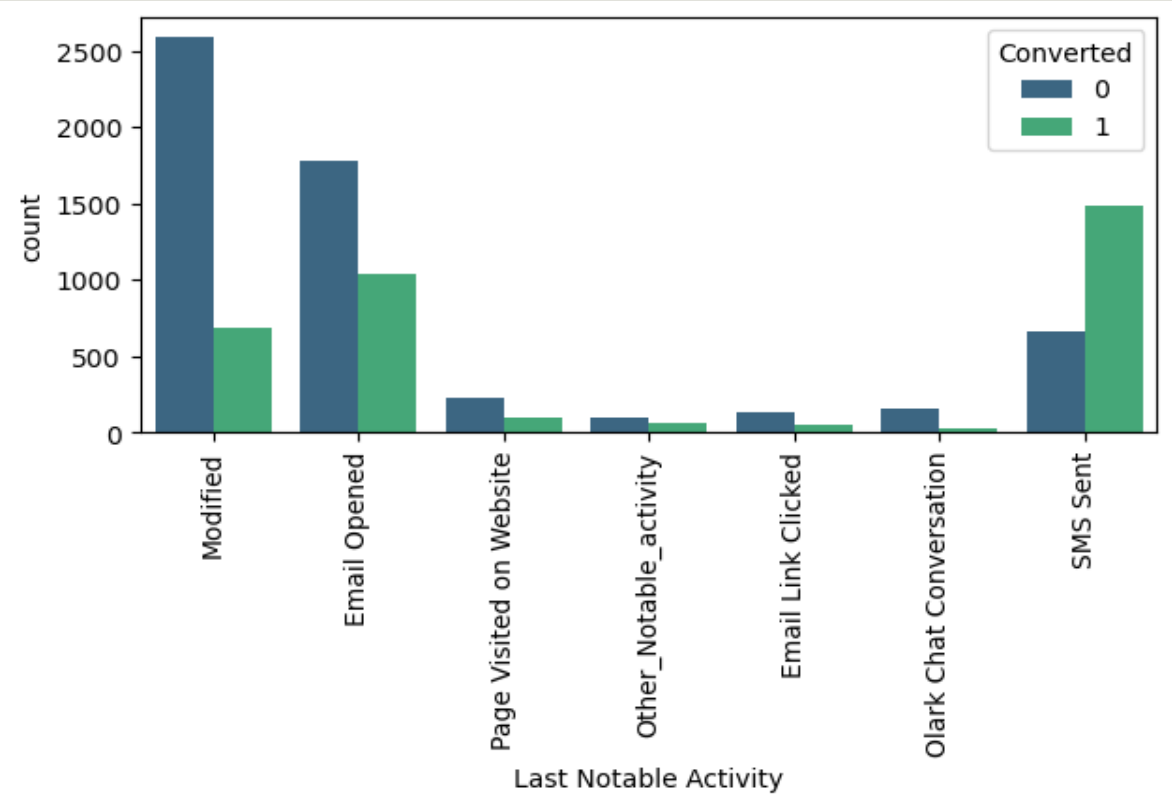
# Analysing 'Tags' variable




- leads that are tagged as 'Will revert after reading the email' and ‘Closed by Horizzon' are likely to get converted.

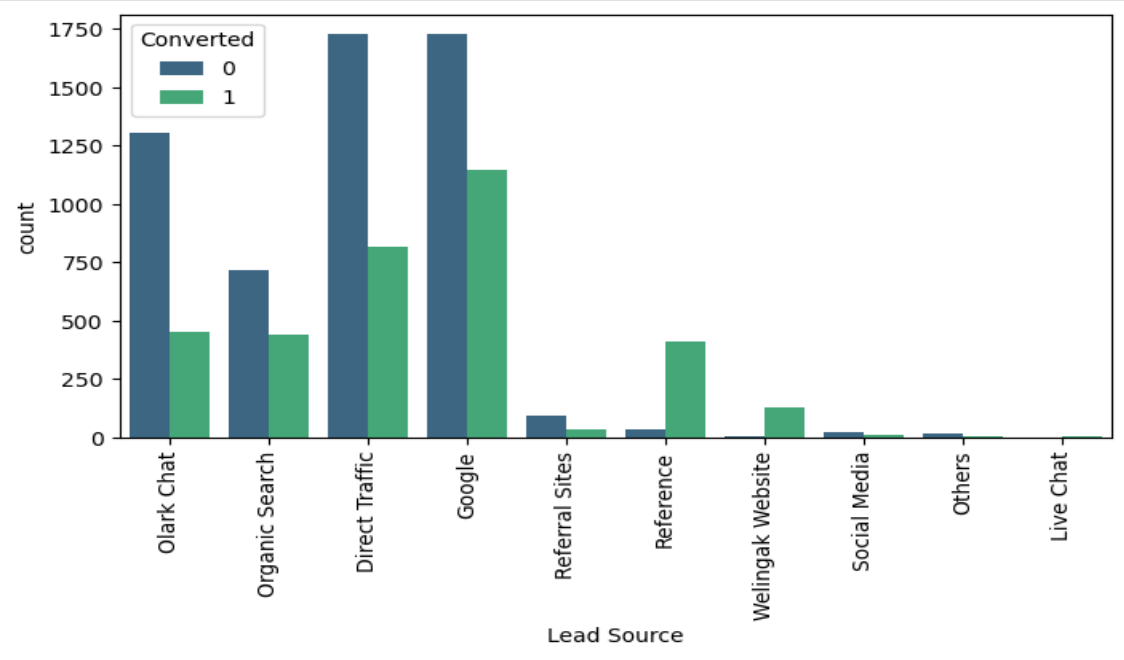
# Analyzing 'Last Activity' and 'Last Notable Activity' variables

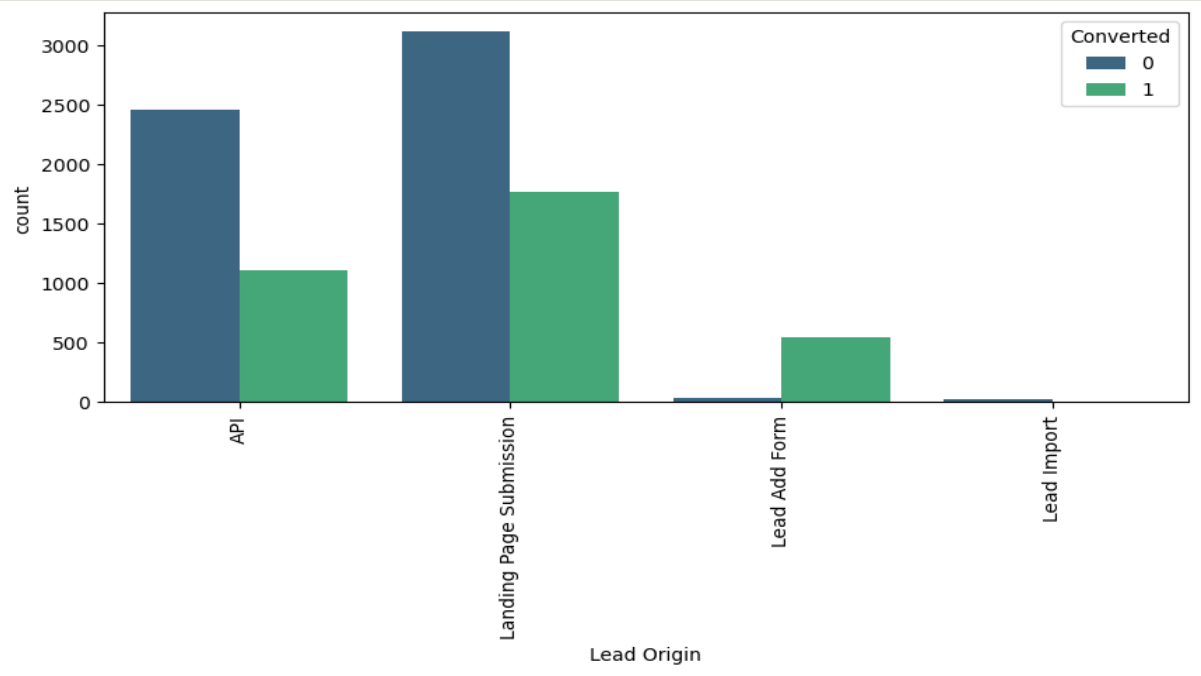



- Leads who are mapped as ‘SMS Sent’ are more likely to get converted whereas the percentage conversion of ‘Olark Chart Conversation’ is the least
- Leads who are mapped as ‘Email Opened’ also contributed significantly towards the conversion rate.

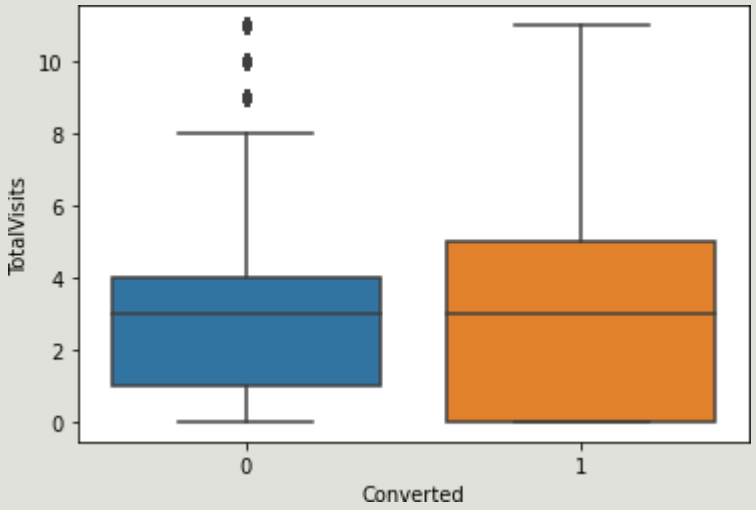
# Analyzing 'Lead Source' and 'Lead Origin' variables

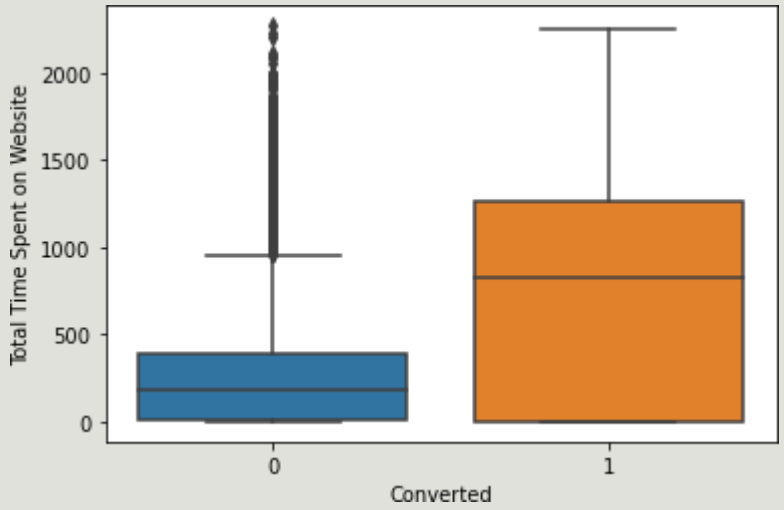


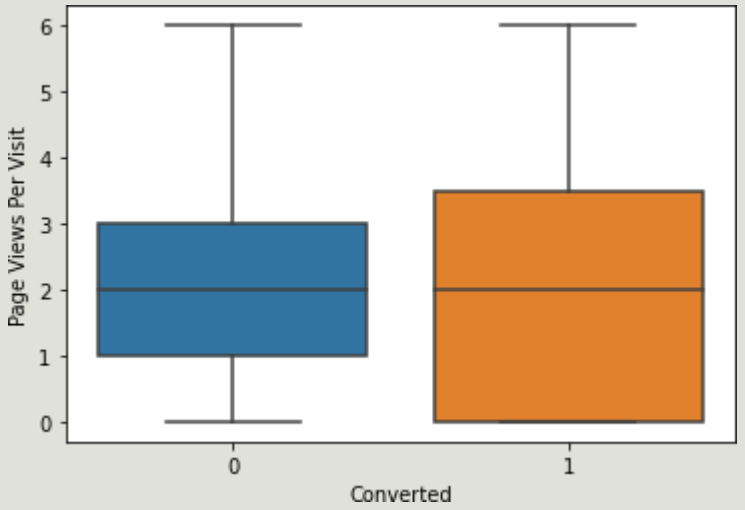




1. Tags whose lead source is from the ‘Welingak Website’ are more likely to get converted along with the leads who are through ‘Reference’. Tags from ‘Google’ are contributed majorly towards the conversion rate
2. Tags whose Lead Origin is ‘Lead Add Form’ are more likely to get converted. Moreover, ‘Landing
3. Page Submission’ also contributed majorly towards the conversion rate

# Analyzing the behaviour of numerical variables with the target variable



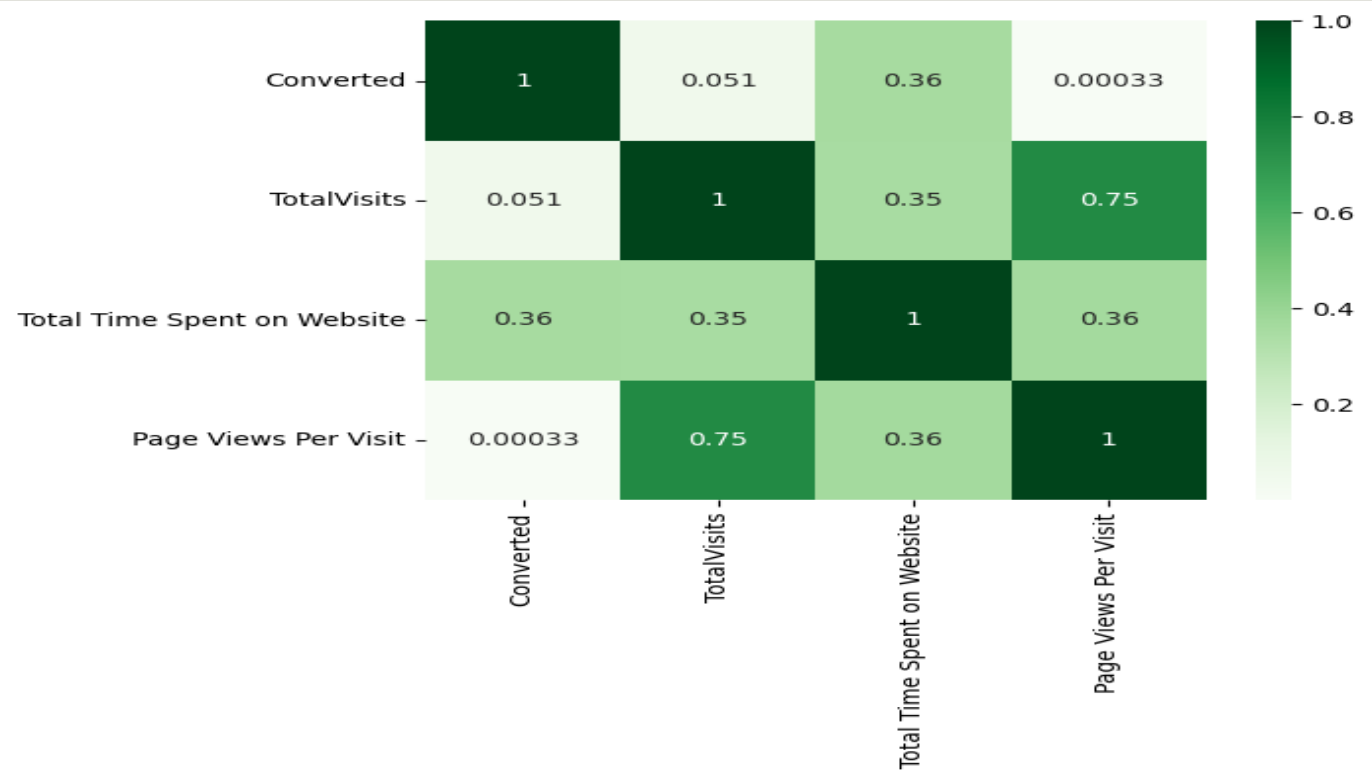




- Even though the median is the same, the upper limit is quite high for the converted visitors. Therefore, it means Visitors who visit the page frequently are likely to be get converted as learners.

- If the visitor spends more amount of time on the website, more likely to him/her become a learner.

- Although the median is the same for both of the ‘Page Views per Visit, nothing can be concluded over here.

# Looking at the correlation of numerical variables




- 'TotalVisits' variable is highly correlated with the ‘Pages Views Per Visit’ variable which may be affected in Our analysis
- Target variable (‘Converted’) has a good correlation with the ‘Total Time Spent on Website’ variable Whereas least correlated with the 'TotalVisits' variable.

# Preparing the data for modelling

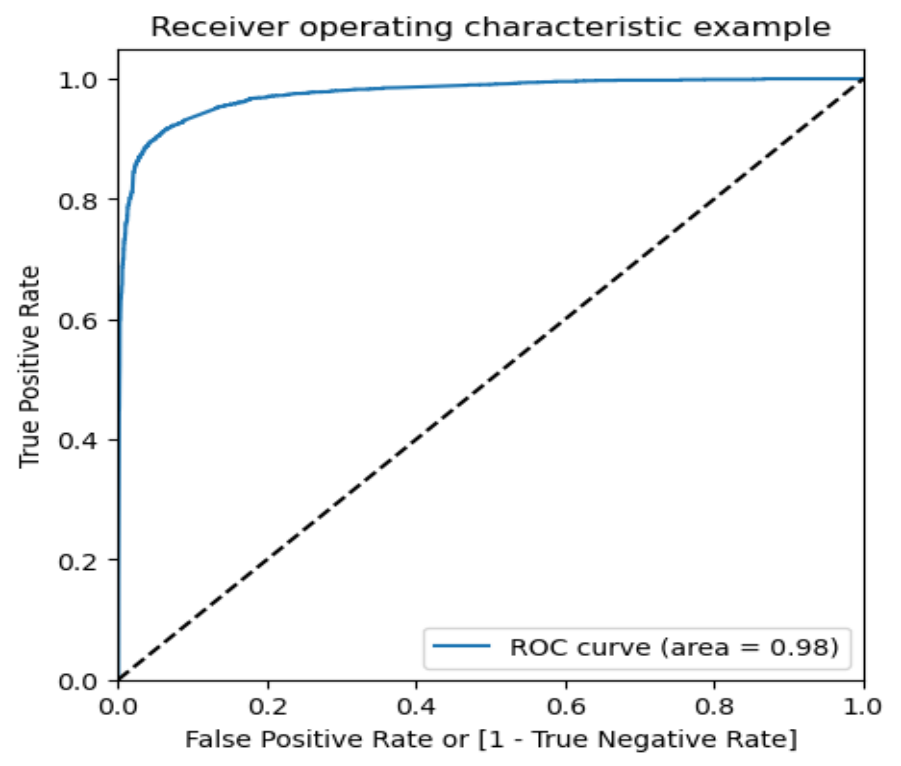
- Dummy variables were created for the categorical variables
- Dataset was then split into a train set and a test set with a 70:30 ratio keeping the random state at 100
- Numerical features were standardized

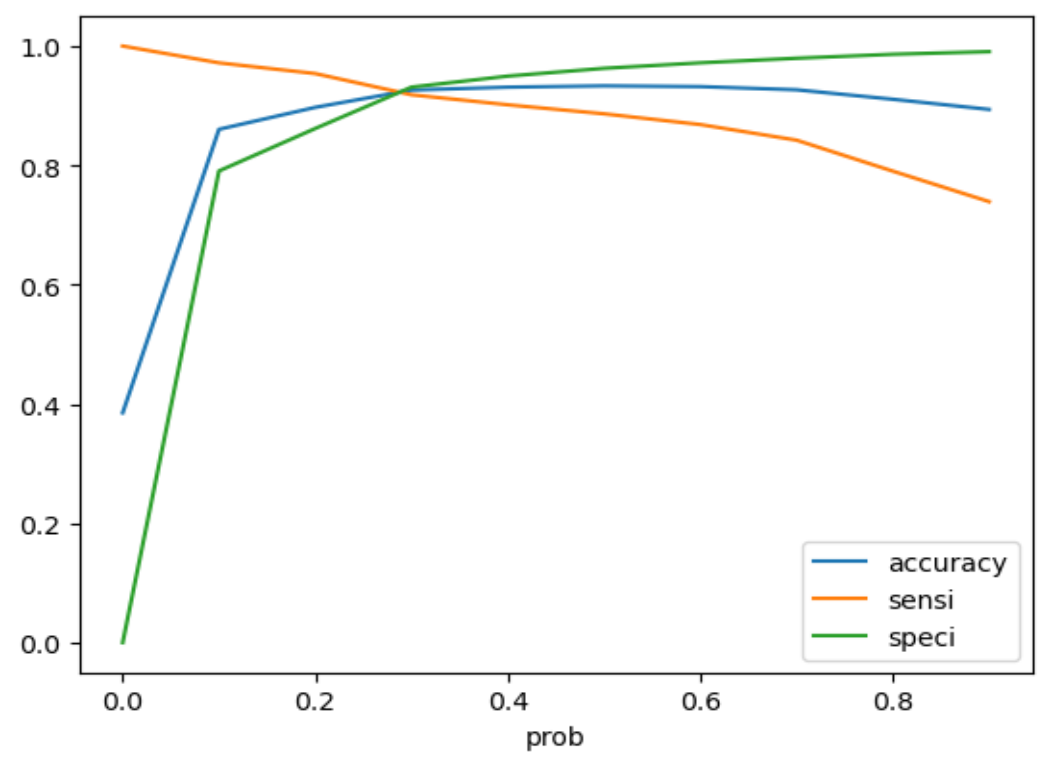
# Model Building

- Coarse tuning using RFE
  - The top 15 variables were selected using the Recursive Frequency Elimination approach

- Fine-tuning using VIF and p-value
  - Built our first model using the variables selected by the RFE method
  - Analyzed the model w.r.t the parameters such as p-value and VIF
  - Dropped ‘Last Notable Activity_SMS Sent’ as VIF > 5
  - Built our second model using the remaining variables
  - This model was considered the optimal model where the model was confident about all coefficient values obtained and also all the variables of the model had low VIF values which indicates that the variables are not much correlated with themselves (except the target variable)

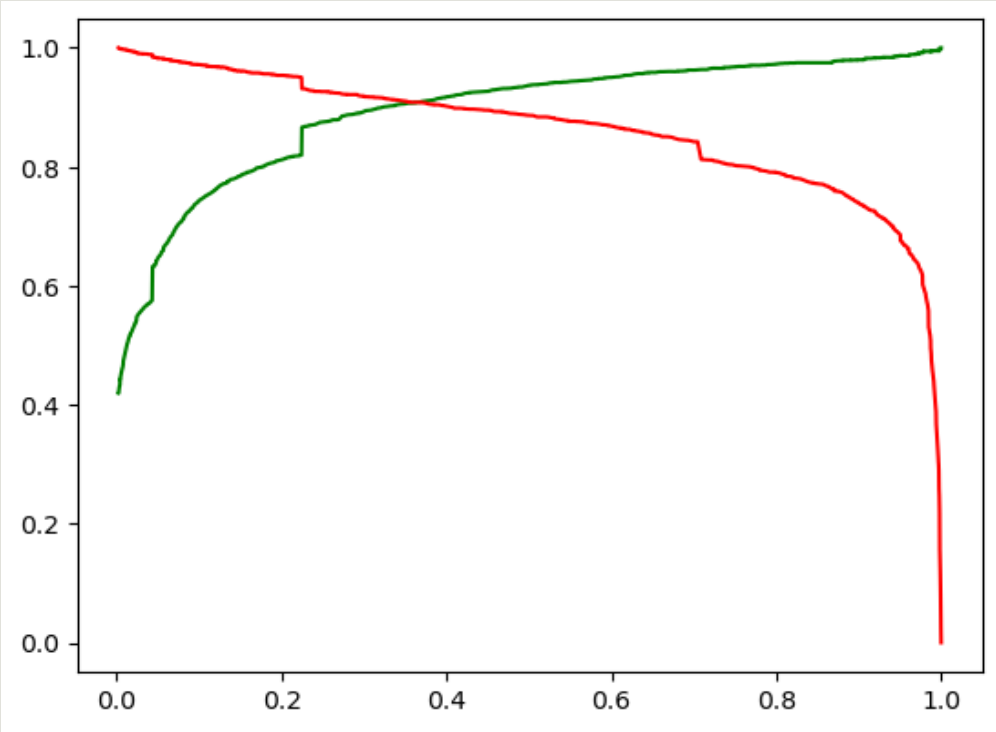
# Threshold determination using ROC Curve




- Optimal cut-off is the one which has balanced values of accuracy, sensitivity and specificity.
- Therefore, 0.3 was considered the optimal cut-off point

# Precision-Recall curve




- Green line represents precision and the red line represents recall
- For low threshold, precision is low and recall is high and for high threshold, high precision and low recall

# Result Outcomes

Following are the results obtained on the train set for the ideal cutoff of 0.3

- Accuracy - 92.6%
- Sensitivity - 91.8%
- Specificity - 93.1%
- Precision - 89.3%
- Recall - 91.8%
- F1_score - 90.5%

Following are the results obtained on the test set for the ideal cutoff of 0.3

- Accuracy - 91.4%
- Sensitivity - 89.8%
- Specificity - 92.2%
- Precision - 86.8%
- Recall - 89.8%
- F1_score - 88.3%

# Summary

Top three variables which contribute most towards the probability of a lead getting converted are :

1. Tags_Closed by Horizzon (7.05)
2. Tags_Lost to EINS (6.31)
3. Tags_Will revert after reading the email (5.01)

Some of the other important variables are

1. Lead Source_Welingak Website (3.80)
2. Last Activity_SMS Sent (2.11)
3. Lead Origin_Lead Add Form (1.28)

# Conclusion

The company should concentrate on the following important elements to improve the lead conversion rate, according to the findings of the logistic regression model:

1. Tags_Closed by Horizzon: Leads that have been assigned Tags as ‘closed by horizon’ are the ones with the highest conversion rates.

2. Tags_Lost: Leads that have been tagged as ‘Lost ‘also contribute to the conversion to a considerable extent.

3. Tags_Will revert after reading the email: Leads that have been tagged as ‘will revert after reading the mail’ also have a significant correlation with the conversion.

4. Other factors that have a good effect on conversion rates include the overall time spent on the website, lead origin from landing page submission, lead source from the business website, Olark Chat, and Last Activity from SMS Sent.

## Conclusion

5. The company should prioritize leads whose lead source is from the “Welingak Website” and generate as many leads as possible through this region.

6. In order to enhance the visitor experience, lengthen the time spent on the platform, and increase lead conversion rates, it is also advisable to ask the development team to improve the user interface of the company's app.

7. These actions will help the company increase lead conversion rates and boost the return on investment from its marketing initiatives.

# THANK YOU