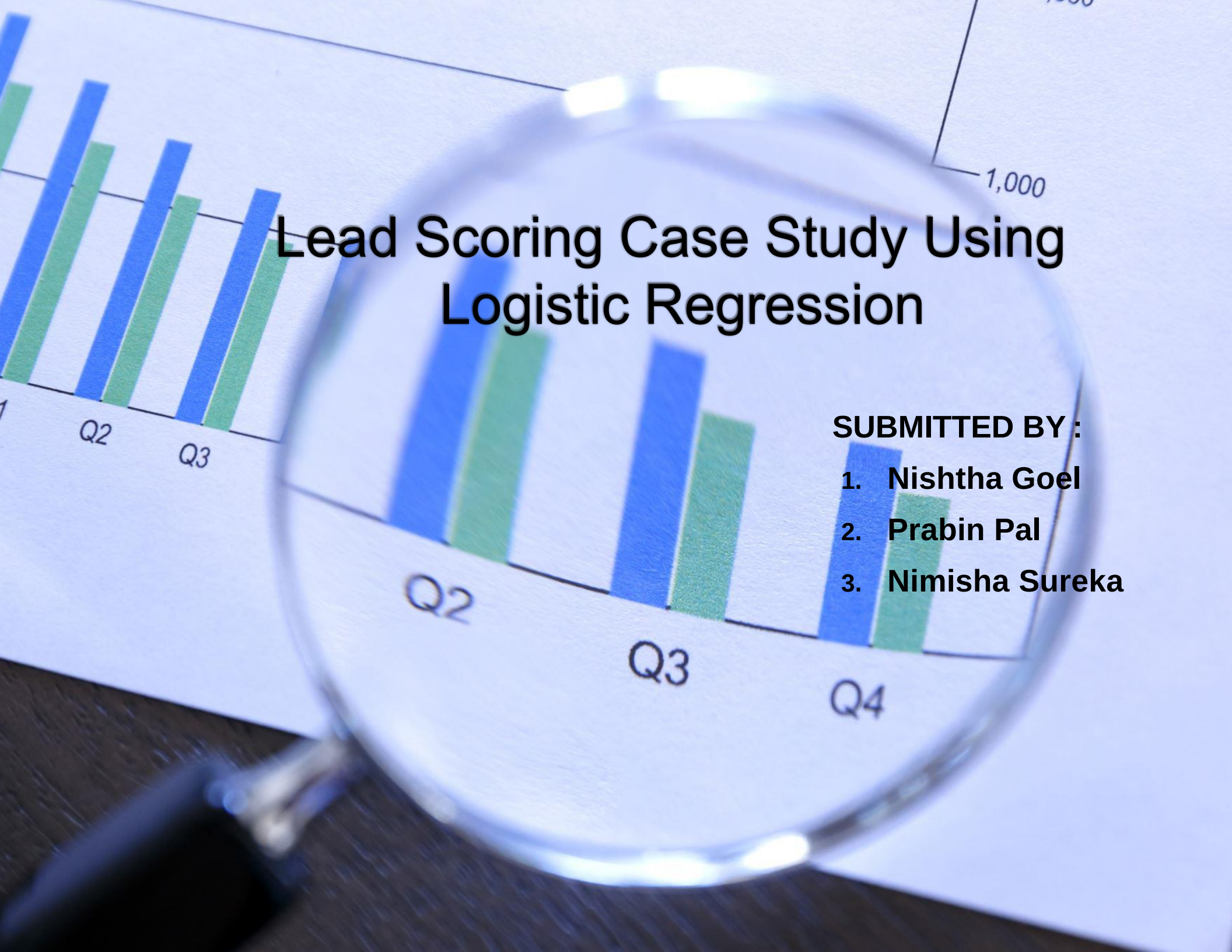



# Lead Scoring Case Study Using Logistic Regression

**SUBMITTED BY :**

1. **Nishtha Goel**
2. **Prabin Pal**
3. **Nimisha Sureka**

# Contents

- **Problem statement**
- **Problem approach**
- **EDA**
- **Correlations**
- **Model Evaluation**
- **Observations**
- **Conclusion**

# Problem Statement

- An education company named X Education sells online courses to industry professionals.
  On any given day, many professionals who are interested in the courses land on their website and browse for courses. They have process of form filling on their website after which the company that individual as a lead.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc.Through this process, some of the leads get converted while most do not.
- The typical lead conversion rate at X education is around **30%.** Now, this means if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as Hot Leads.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone

# Business Goals

- Lead X wants us to build a model to give every lead a lead score between 0 - 100 . So that they can identify the Hot leads and increase their conversion rate as well.

- The CEO want to achieve a lead conversion rate of 80%.

- They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full man power and after achieving target what should be the approaches.

# Solution Approach

- Importing the data and inspecting the data frame
- Data preparation and Cleaning
- EDA
- Dummy variable creation
- Test-Train split
- Feature scaling
- Correlations
- Model Building (RFE  VIF and p- Values)
- Model Evaluation
- Making predictions on test set

# Importing the data and inspecting the  data frame

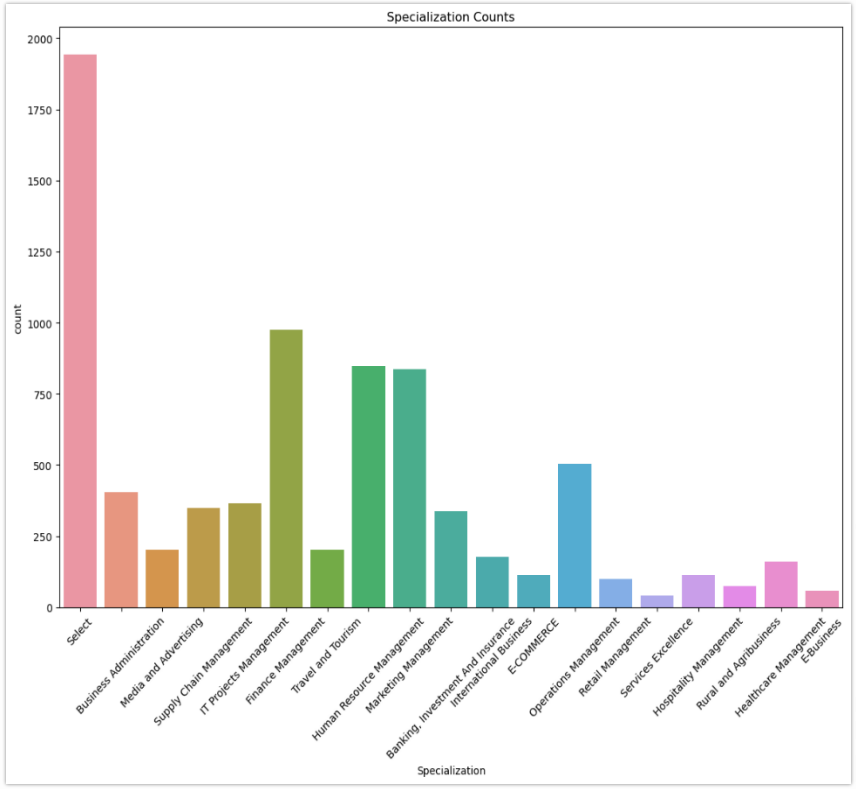
- The Leads dataset was downloaded from portal along with leads data Dictionary
- The variables were carefully studied by referencing the dictionary and the CSV file.
- The data was imported into the python notebook for data cleaning, EDA and modelling

# EDA – Data Cleaning

- The variables were checked for null values and organised in a descending order to get a better view.
- The variables which had null values greater than 35% were discarded as not much inference could be obtained from them.
- The “City” and “Country” columns were also dropped as these variables did not impact the analysis for Leads conversion.
- Three of the variables had value “Select” which means no value was selected when data was collected ad thus we can treat it as Null value and drop them

# Specialization

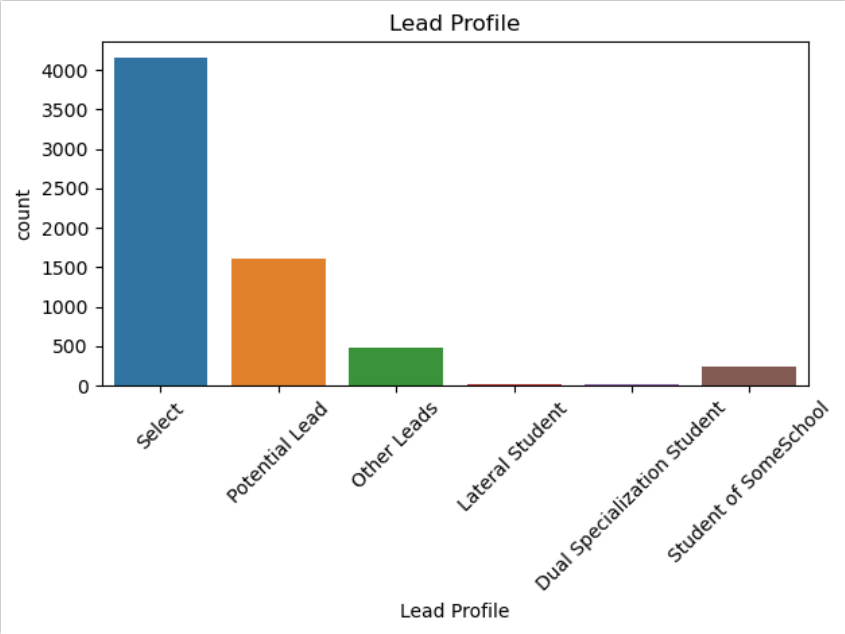
## Specialization Count Chart




- Leads from HR, IT Projects and travel and Tourism specializations are high probability to convert
- The Select(Null value) is less than 35% of the total count and hence Variable was not dropped.

# Lead Profile

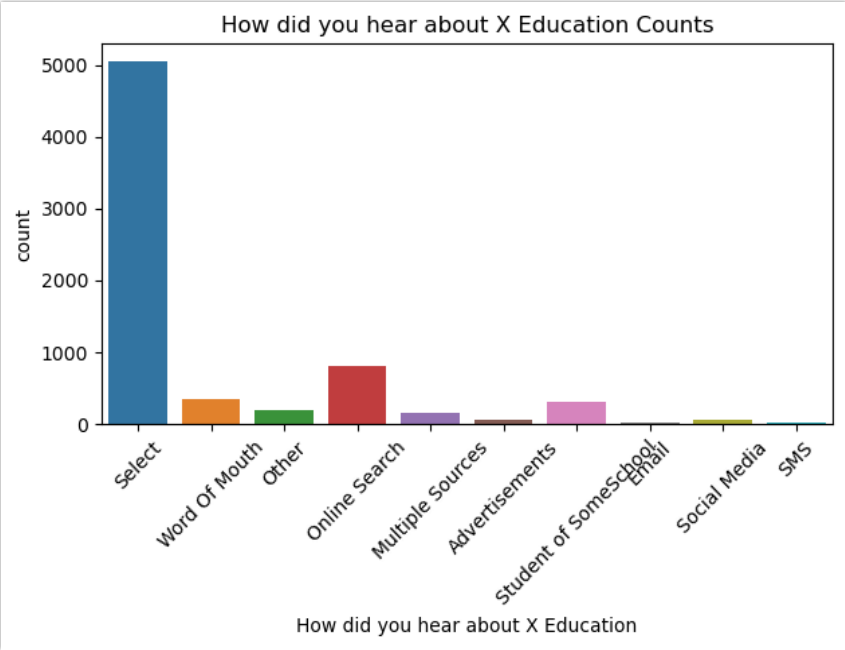
## Lead Profile Count Chart




- The Select(Null value) is more than 45% of the total count and hence Variable was dropped

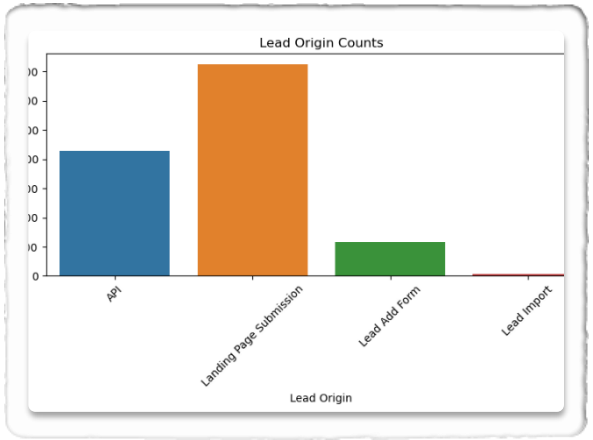
# How did you hear about X Education

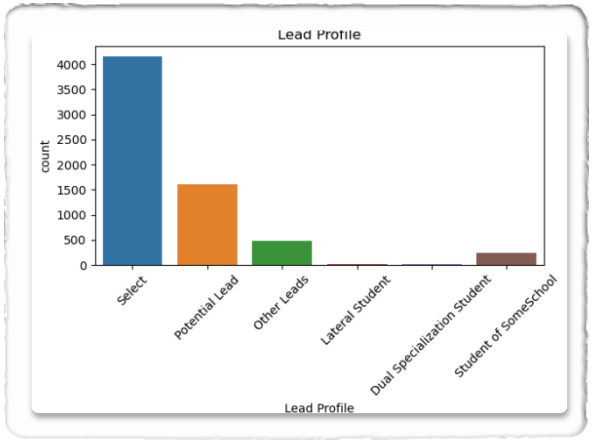
## How did you hear about X Education  Count Chart

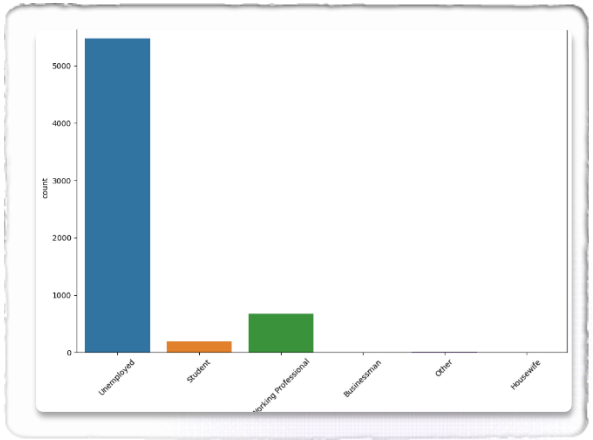



- The Select(Null value) is more than 50% of the total count and hence Variable was dropped.

# Lead Source & Lead Origin




## Lead Origin

Most of the Lead Origin Came from Landing Page Submissions and API

## Lead Source

Most of the lead source was direct traffic and Google search

## Occupation Status

Most of the leads which visited were Unemployed.

# Dummy variable creation

- 'Lead Origin', 'Lead Source', 'Do Not Email', 'Last Activity', 'Specialization', 'What is your current occupation', 'A free copy of Mastering The Interview', 'Last Notable Activity' were the variables with Data type as object and hence were categorical columns for which Dummies were created.
- For Variable Specialization the Select Value dummy column was dropped as it is null value.
- The variables for which dummies were created was then dropped.

# Test-Train split and Scaling

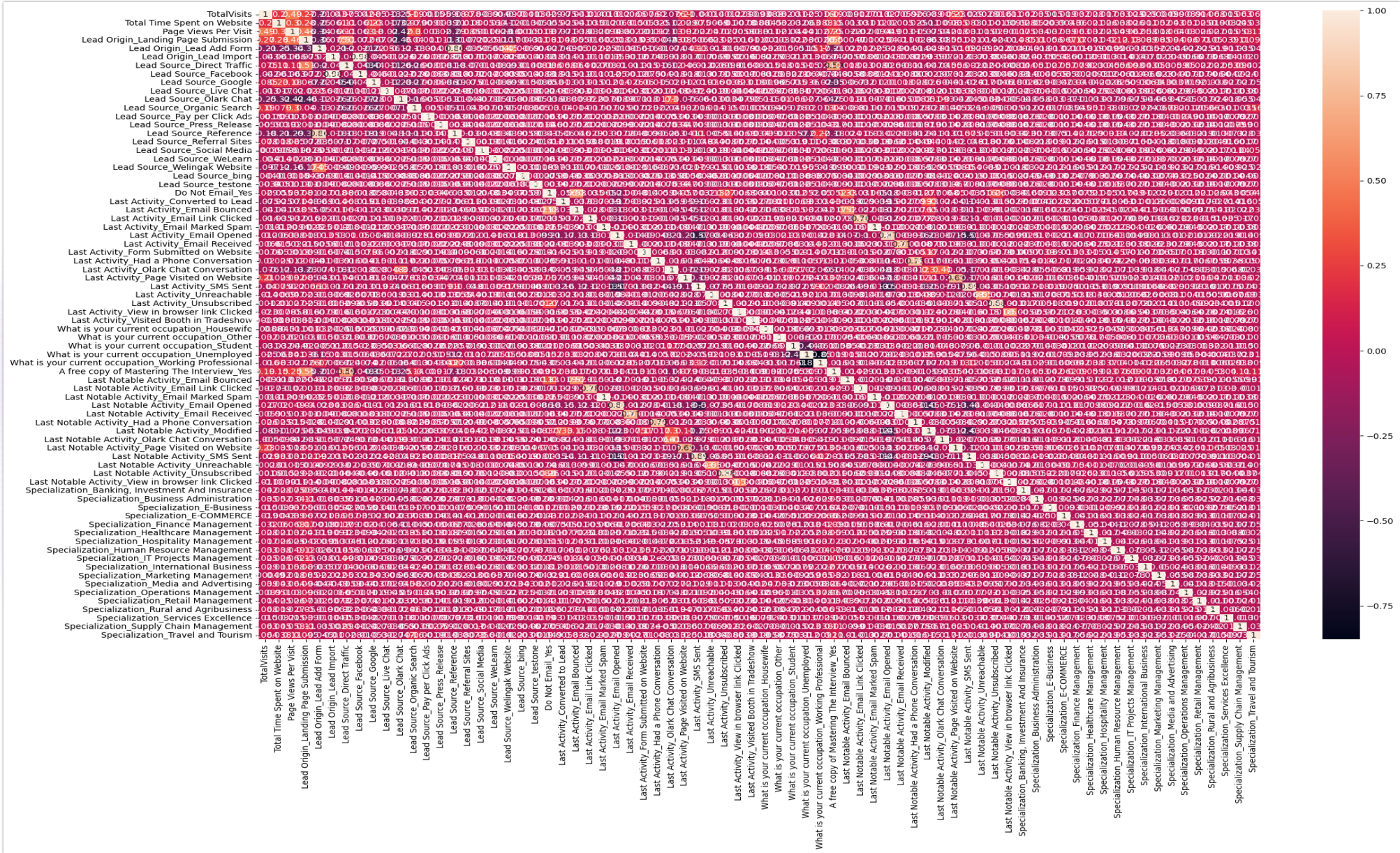
Test-Train Split:

- The data set was split with a 70% Train size and 30% test size with random state of 100.

Scaling:

- Min-Max scalar was used on 3 numerical variables, 'TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website'.

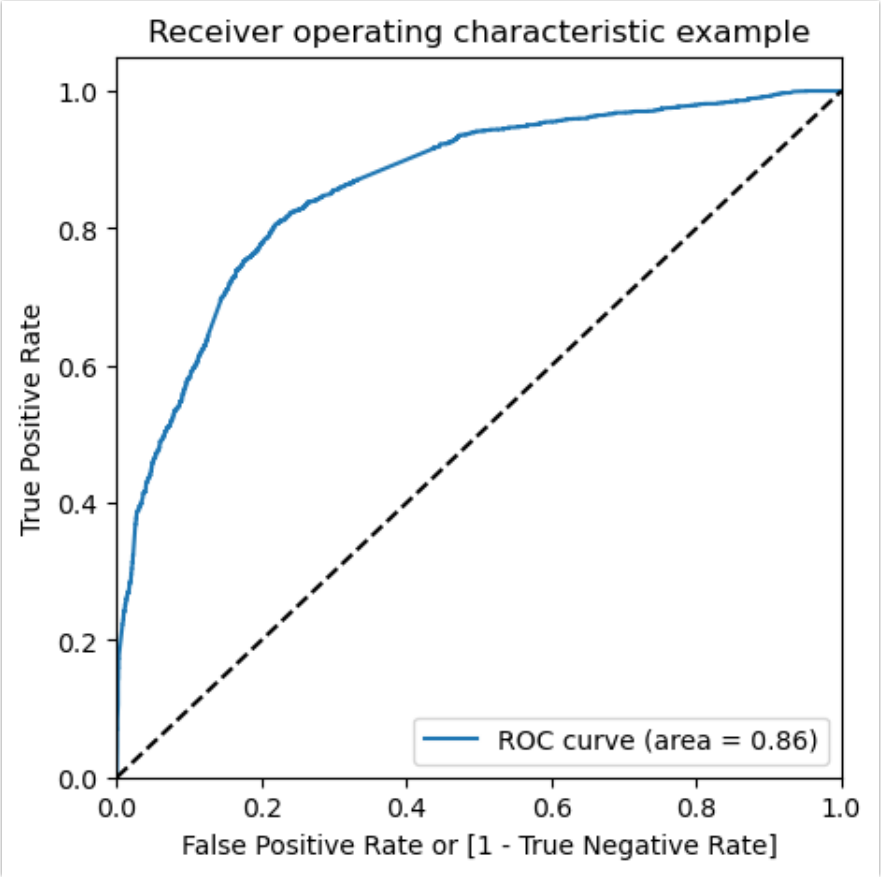
# Correlation

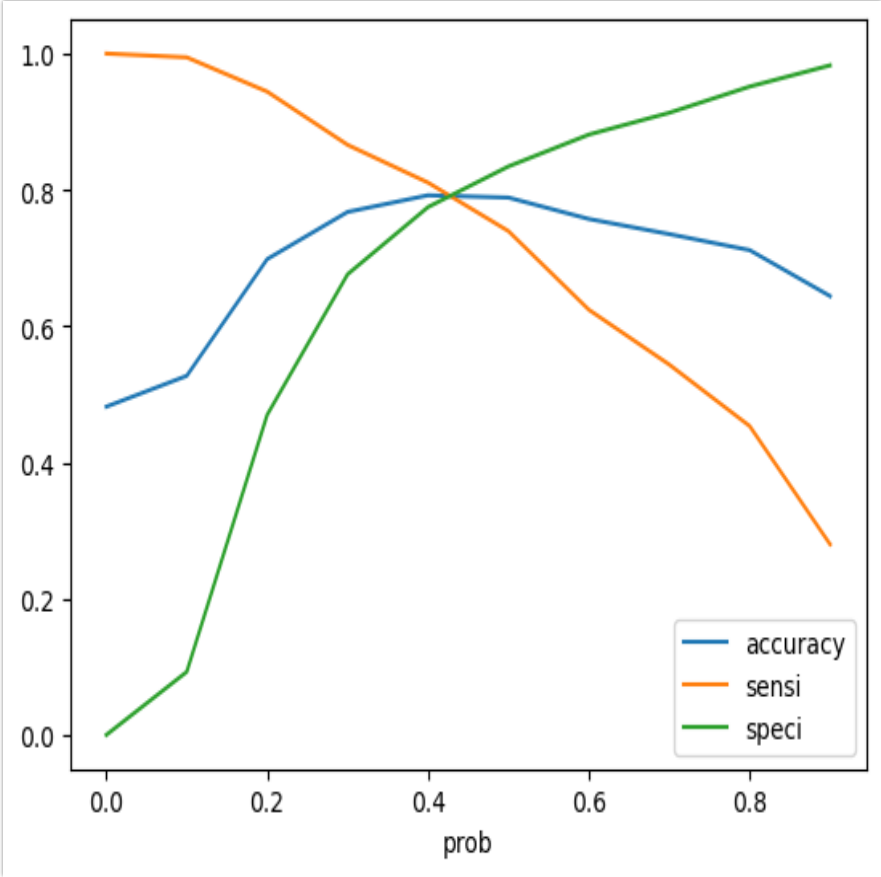


# Model Building and Evaluation

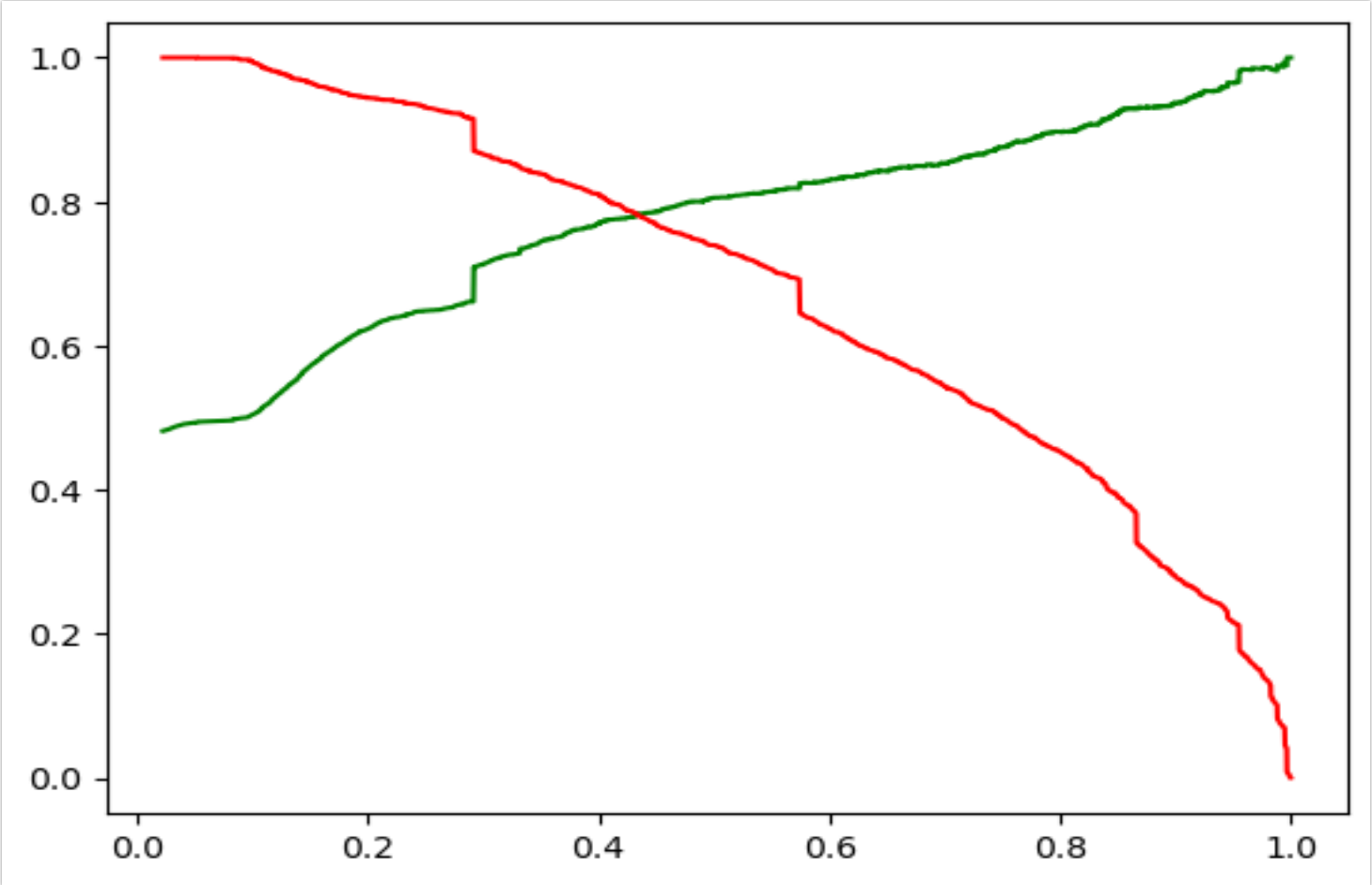
- For Model Building we used Logistic Regression and Ran RFE with 15 Features by fitting the Training and Test Data sets
- We did multiple iterations of Model builds until P- Values and VIF were within accepted range. The dataset (both training and test) were refit and high P values and VIFs columns were dropped and finally we checked the accuracy, sensitivity and specificity for the model.
- Confusion Matrices were also created and ROC curves were drawn to get optimal cutoff value.
- 0.42 is the tradeoff between Precision and Recall , Hence this value was chosen to consider any prospect lead with conversion probability higher than 42 % to be a hot lead

# Model Evaluation



# Precision and Recall Tradeoff Curve



# Observations

Train data

1. Evaluation Metrics for Train Data Set.

- Accuracy :0.79
- Sensitivity:0.74
- Specificity:0.83
- Precision: 0.78
- Recall: 0.74

Test Data

1. Evaluation Metrics for Test Data Set.

- Accuracy :0.79
- Sensitivity:0.78
- Specificity:0.79
- Precision: 0.78
- Recall: 0.77

# Conclusion

1. The model trained to predict conversions from leads is 80% accurate. This means the business team can use our model predictions to improve their conversion rate from 30% to 80%.
2. Phone call and text message are the most effective way to reach out to leads. This will lead to better conversion rate for business teams.
3. The model can also be optimised for zero false positive and we will still end up predicting 42% potential leads. For the business team, If mis-classification cost is high the model is still capable of capturing reasonably strong number of leads (42%)