Lead Score Case Study

Assignment by

- **Mazar Godhrawala**
- **Mayur Punamiya**
- **Mohammed Amanullah**

Problem Statement

An X Education need help to select the most promising leads, i.e., the leads that are most likely to convert into paying customers. The company requires us to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO has given a ballpark of the target lead conversion rate to be around 80%.

Business Objectives

- ✓ X education wants a model to assign a lead score to know most promising or hot leads.
- ✓ There are some more problems presented by the company which the model should be able to adjust to if the company's requirement changes.
- ✓ The model should be such that it can be used accurately when it is Deployed in future.

METHODOLOGY USED TO DERIVE CONCLUSION

Step-1:

Data Importing, Inspecting, Cleaning & Manipulation

a) Handling of Duplicate Data.
b) Handling NA or Missing Values.
c) Dropping of Unnecessary Columns (i.e., which are not taken for in Analysis)
d) Dropping of Columns having large number of missing values.
e) Imputation of Values where required.
f) Handling Outliers.

Step-2:

Data Analysis – Exploration

Univariate Analysis.

- Categorical Variables
- Numerical Variables

Bivariate data analysis:

- Correlation coefficients and pattern between the variables etc.

METHODOLOGY USED TO DERIVE CONCLUSION

Step-3:

Model Building Preparation & Validation

- Dummy Variables
- Test-Train Split
- Scaling

Step-4:

Model Evaluation

- Creating a data frame with the actual conversion flag and predicted probabilities
- Creating new column 'Predicted’
- Finding the Optimal Cutoff
- Precision-Recall View

Step-5:

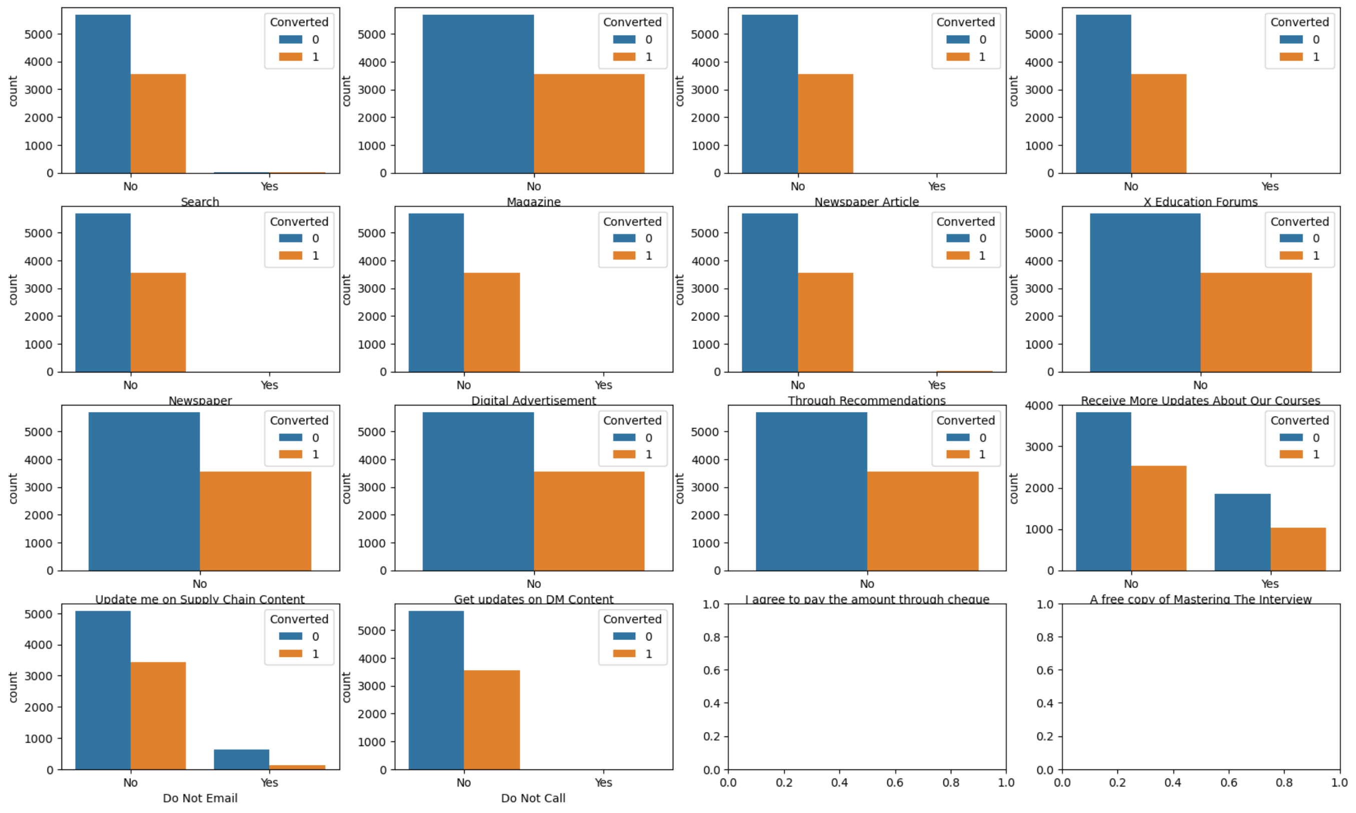
Making Predictions based on the Test Set

Step-6:

Deriving Conclusion & Recommendation based on Model.

Data Cleaning and Preparation

- Firstly, we dropped all the columns which had more than 30% values missing or NA.
- We Checked the other remaining columns and drop columns which are not required for our analysis

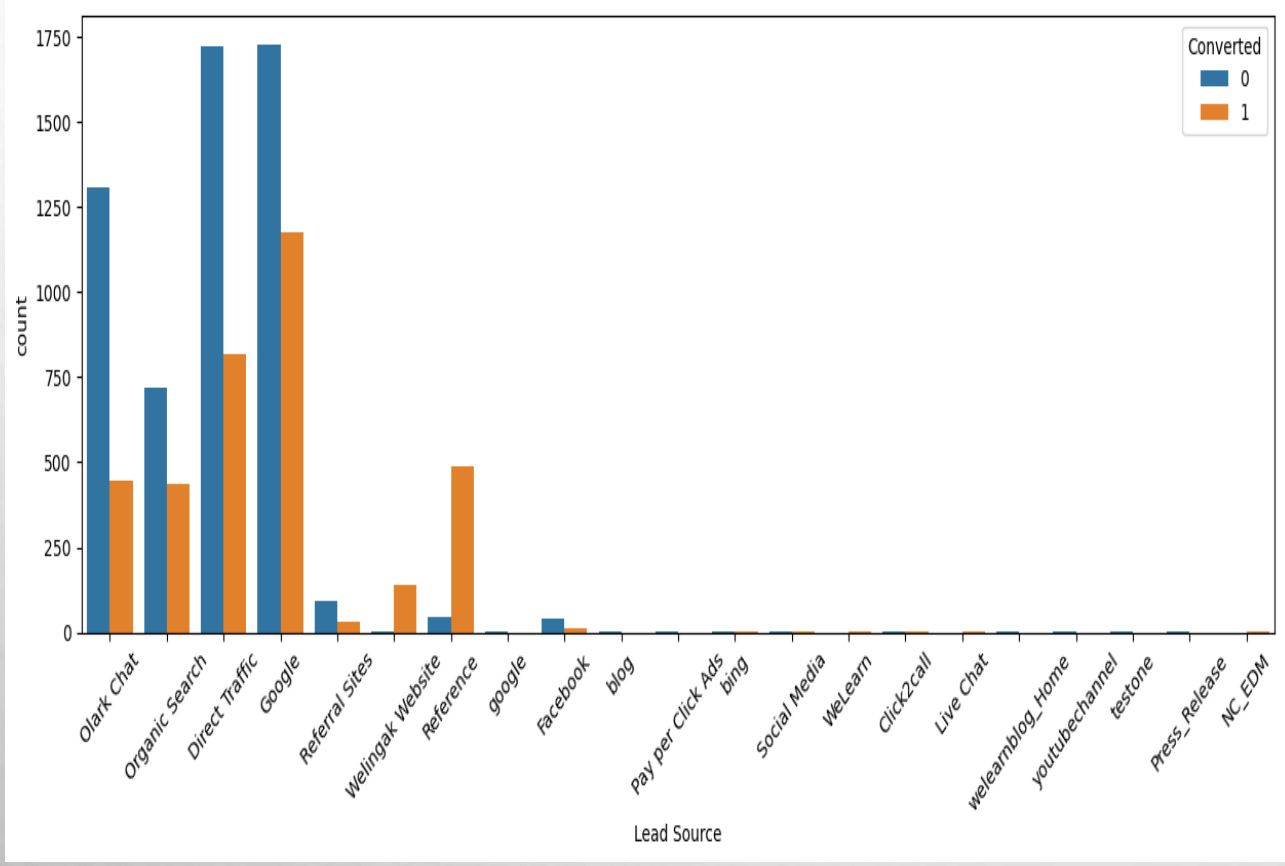


Since most of the above contain No as their value we can drop them.

- Next the columns with null or missing values, we imputed them with 0.0
- We Checked the other remaining columns and drop columns which are not required for our analysis

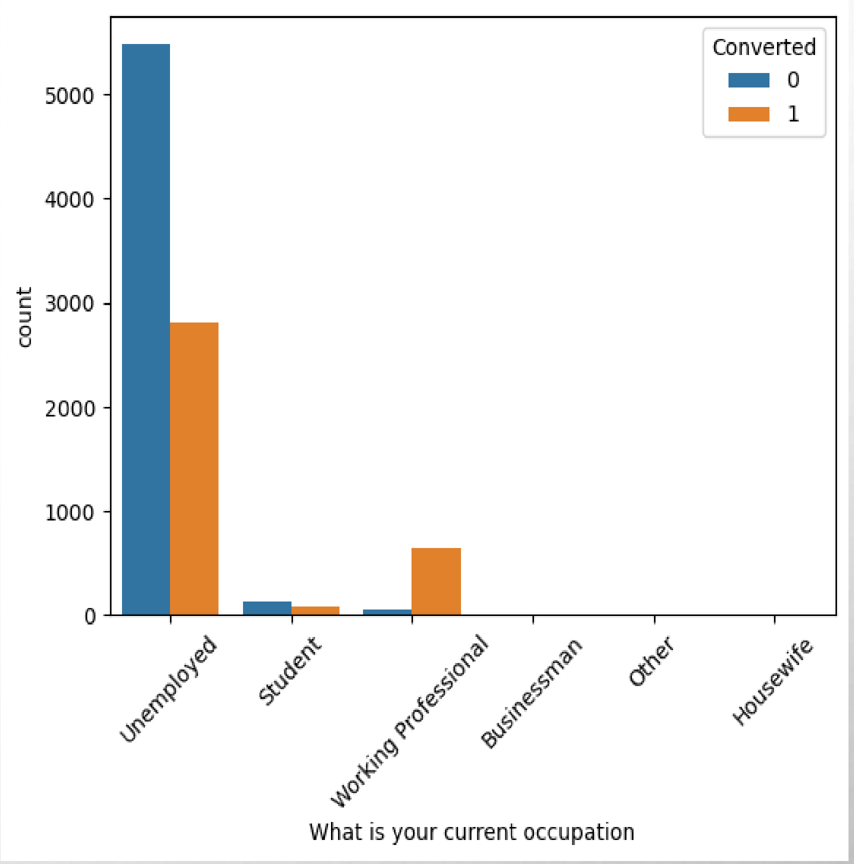
DATA EXPLORATION (EDA) FINDINGS

Lead Sources



What Chart Shows:
Most conversions are from 'Google' and 'Direct Traffic' for 'Lead Source'

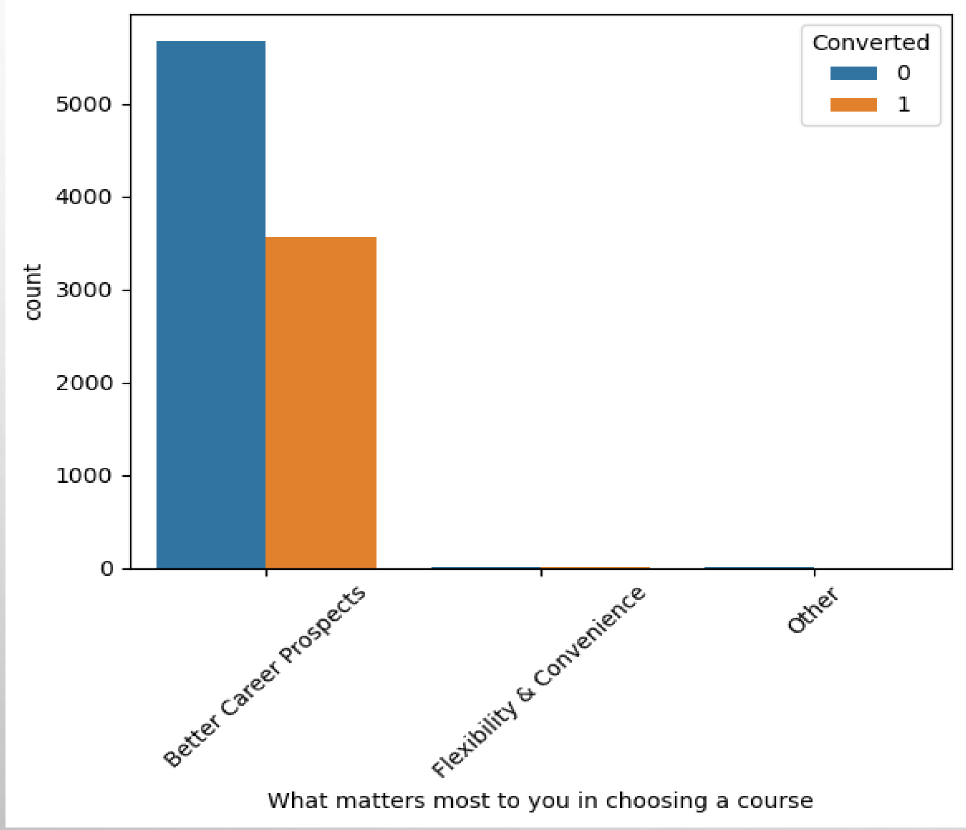
Current Occupation



What Chart Shows:
Most leads are from 'Unemployed' but conversion is low. 'Working Professional' has high conversion rate

DATA EXPLORATION (EDA) FINDINGS

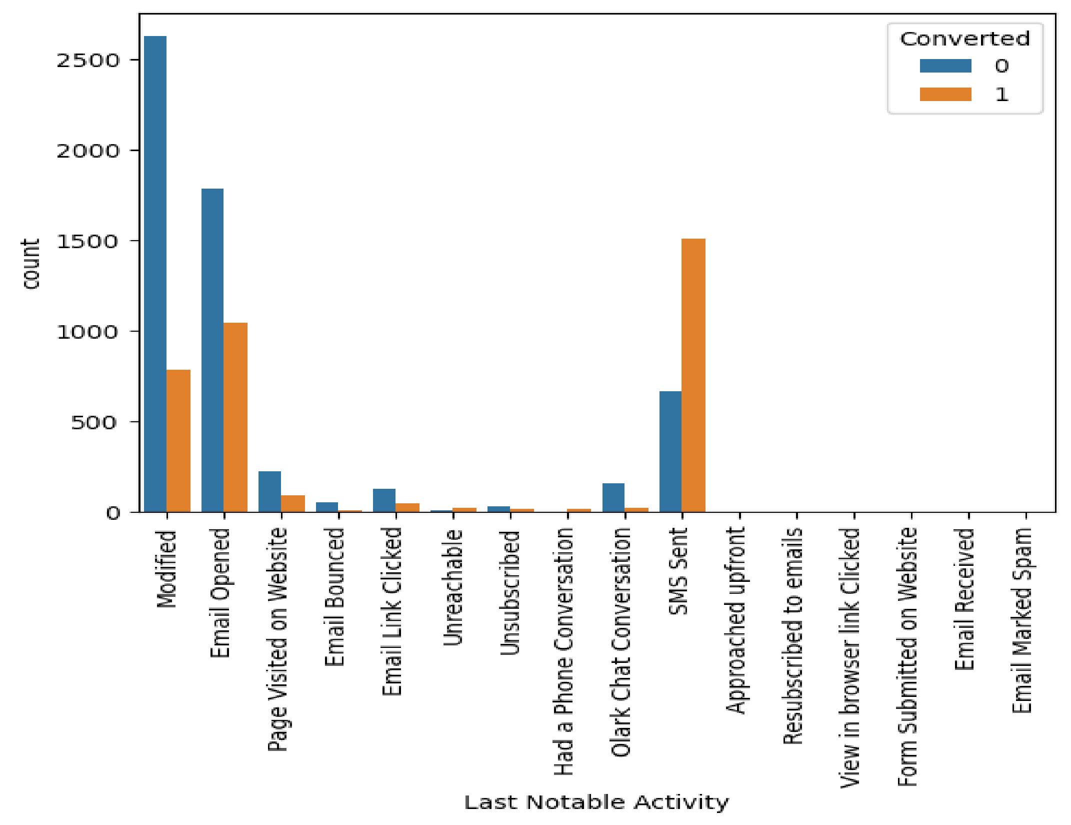
What matters most to you in choosing a course



What Chart Shows:

Highest leads as from "Better Career Prospects" and since almost all values belong to this category, this column can be dropped.

Last Notable Activity

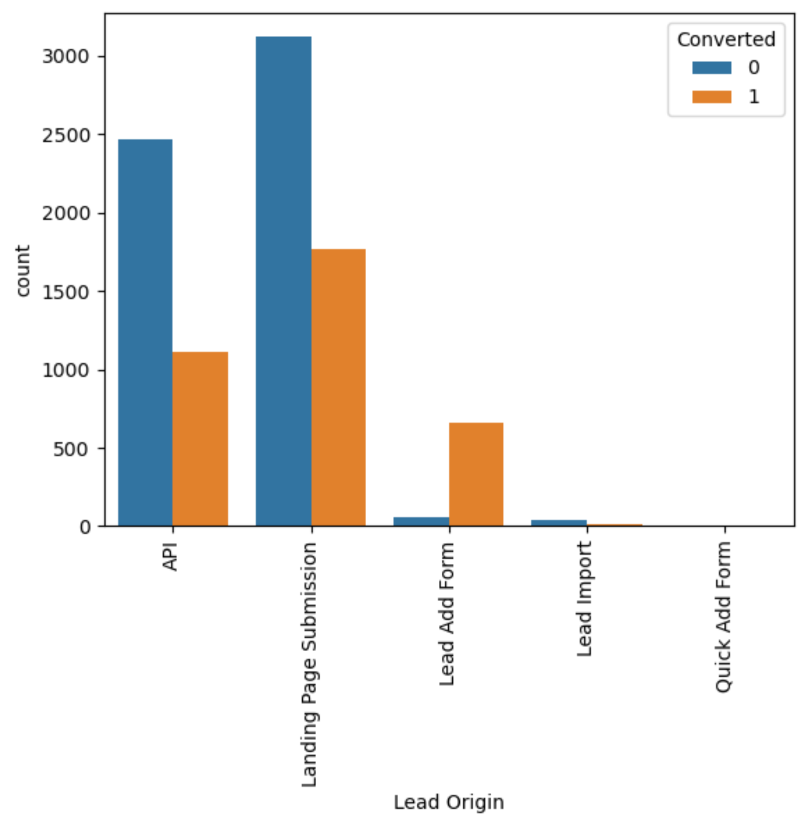


What Chart Shows:

Email Opened' has very low conversion but 'SMS sent' has very good conversion.

DATA EXPLORATION (EDA) FINDINGS

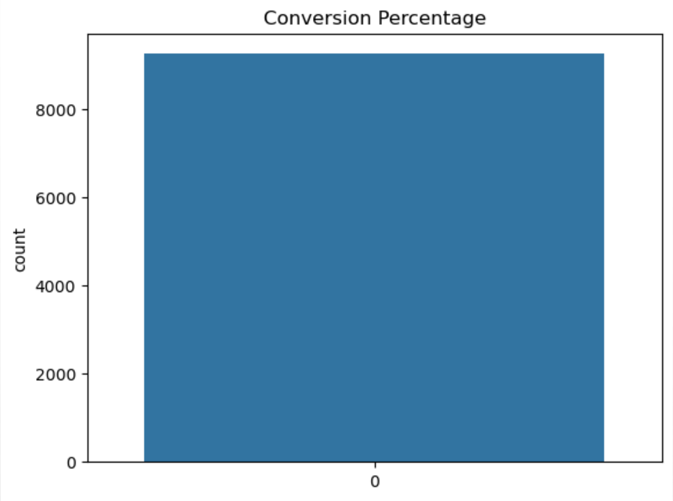
What matters most to you in choosing a course

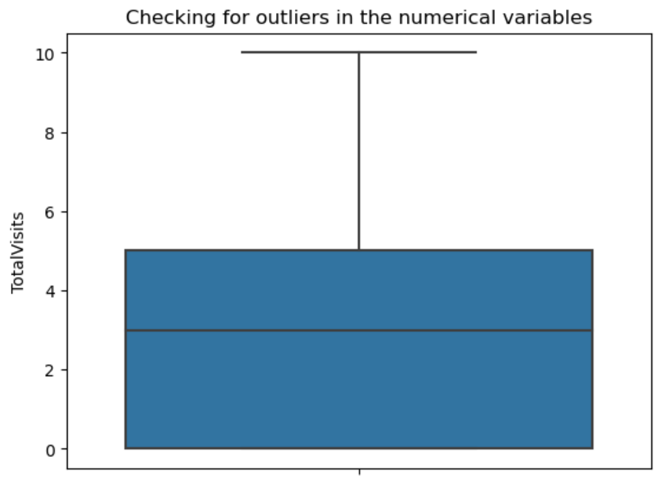


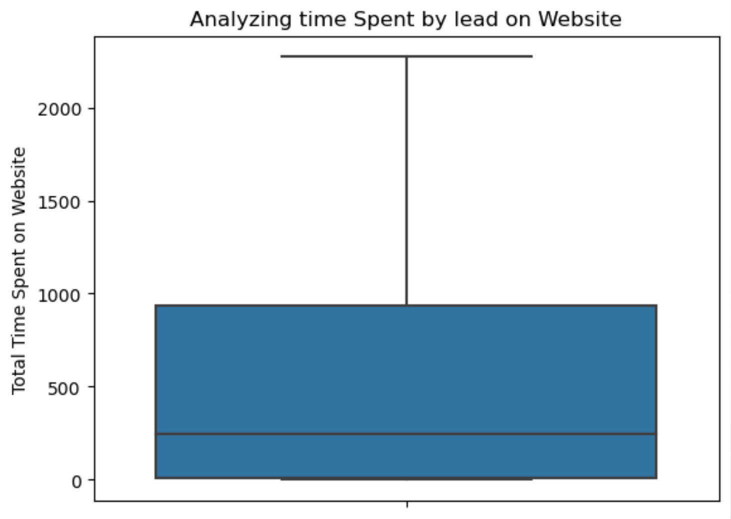
What Chart Shows:

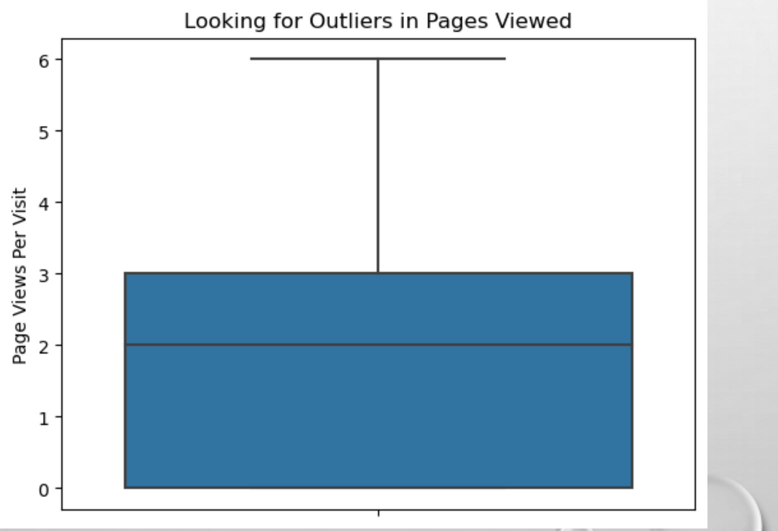
Lead Add Form' has good conversion while 'Landing Page Submission' generated most leads.

DATA EXPLORATION (EDA) FINDINGS









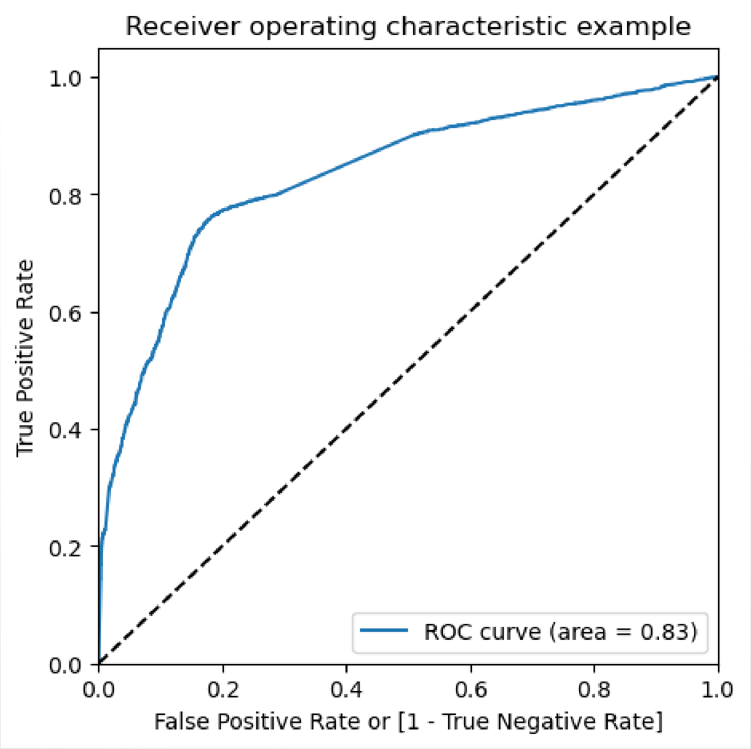
DATA CONVERSION

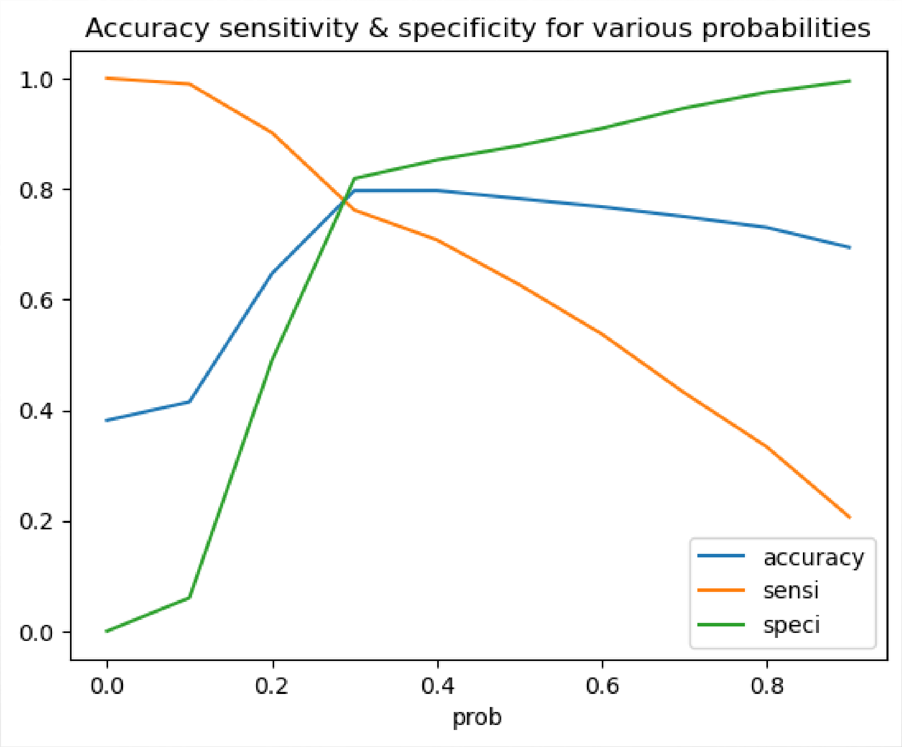
- ✓Numerical Variable normalized
- ✓Outliers handled
- ✓Dummy Variable Created
- ✓Test-Train Split
- ✓Feature Scaling
- ✓Corelations searched and found.

MODEL BUILDING

- ✓Splitting the Data into Training and Testing Sets.
- ✓The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- ✓Generalized Linear Model Regression Results.
- ✓Feature Selection Using RFE.
- ✓Building Model.
- ✓Assessing the model.
- ✓Predictions made based on test data set.

PLOTTING ROC CURVE

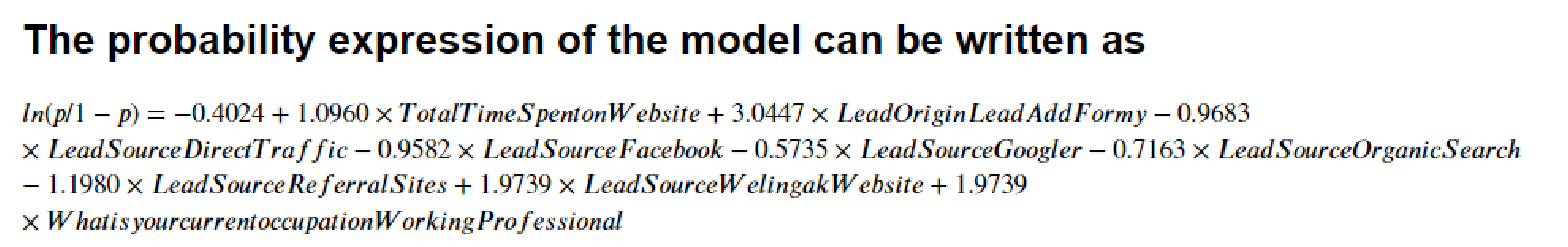




- Since we know that the perfect ROC Curve should be a value close to 1. We are getting a value of 0.83 indicating a good predictive model.
- From the curve above, we see that 0.25 is the optimum point to take it as a cutoff probability.

CONCLUSIONS & PREDICTIONS

- The Accuracy, Precision and Recall score we got from the test data are in the acceptable region.
- Accuracy, Sensitivity and Specificity values of test set are around 76%, 76% and 77% which are approximately closer to the respective values calculated using trained set.
- Also, the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is close to 80% (i.e., 78%)
- Hence overall this model seems to be good.
- A customer Lead sourced by "Welingak Website" is a Hot Lead.
- A customer who is currently "Working Professional" or "Unemployed" is a Hot Lead.
- Total Time Spent on Website gets high conversion



The probability expression of the model can be written as

$$ln(p/1-p) = -0.4024 + 1.0960 \times TotalTimeSpentonWebsite + 3.0447 \times LeadOriginLeadAddFormy - 0.9683 \times LeadSourceDirectTraffic - 0.9582 \times LeadSourceFacebook - 0.5735 \times LeadSourceGoogler - 0.7163 \times LeadSourceOrganicSearch - 1.1980 \times LeadSourceReferralSites + 1.9739 \times LeadSourceWelingakWebsite + 1.9739 \times WhatisyourcurrentoccupationWorkingProfessional$$