LEAD SCORE CASE STUDY

Analysis Presentation

Sameer Kumar Ramakrishnagari (Sept 19, 2023)

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.
- The company requires to build a model wherein to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

Goal of the Case Study

- To Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
- A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

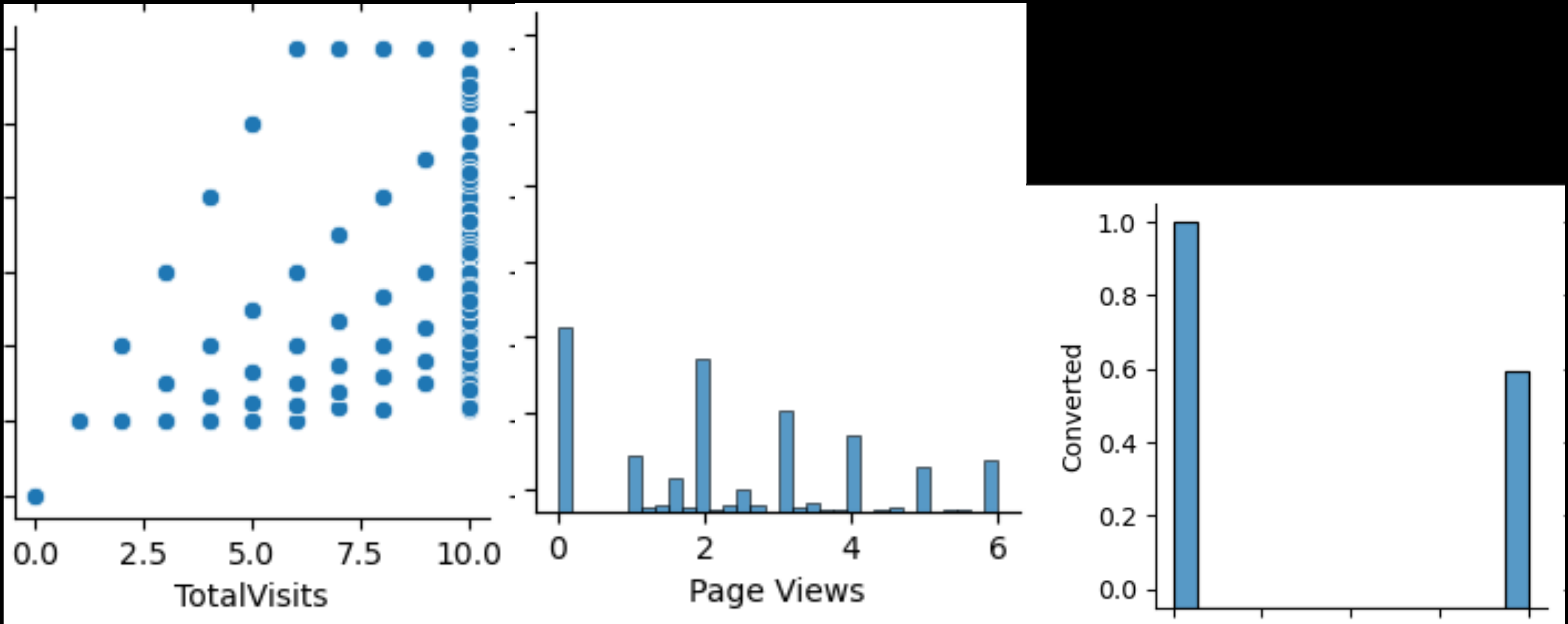
Data Importing and Cleaning

- The Leads data is read using Python pandas library and the data-frame is analyzed to look at it’s shape, datatype information of columns and descriptive statistics of the numeric columns.
- Then we proceeded with Data Cleaning where we handled the dropping of duplicate records, dropping the columns with more than 50% of null values.
- The Outliers for the numerical columns are identified bt looking at the box plot and handled by imputing the values above 95th percentile. For categorical variables we replaced them with the Mode of the respective column.
- Furthermore we dropped the columns with skewed values in terms of value counts and columns with just one constant variable in it.
- Finally the records are dropped for columns with less than 2% of null values.

Data Visualization & Analysis

Univariate and Bivariate Analysis

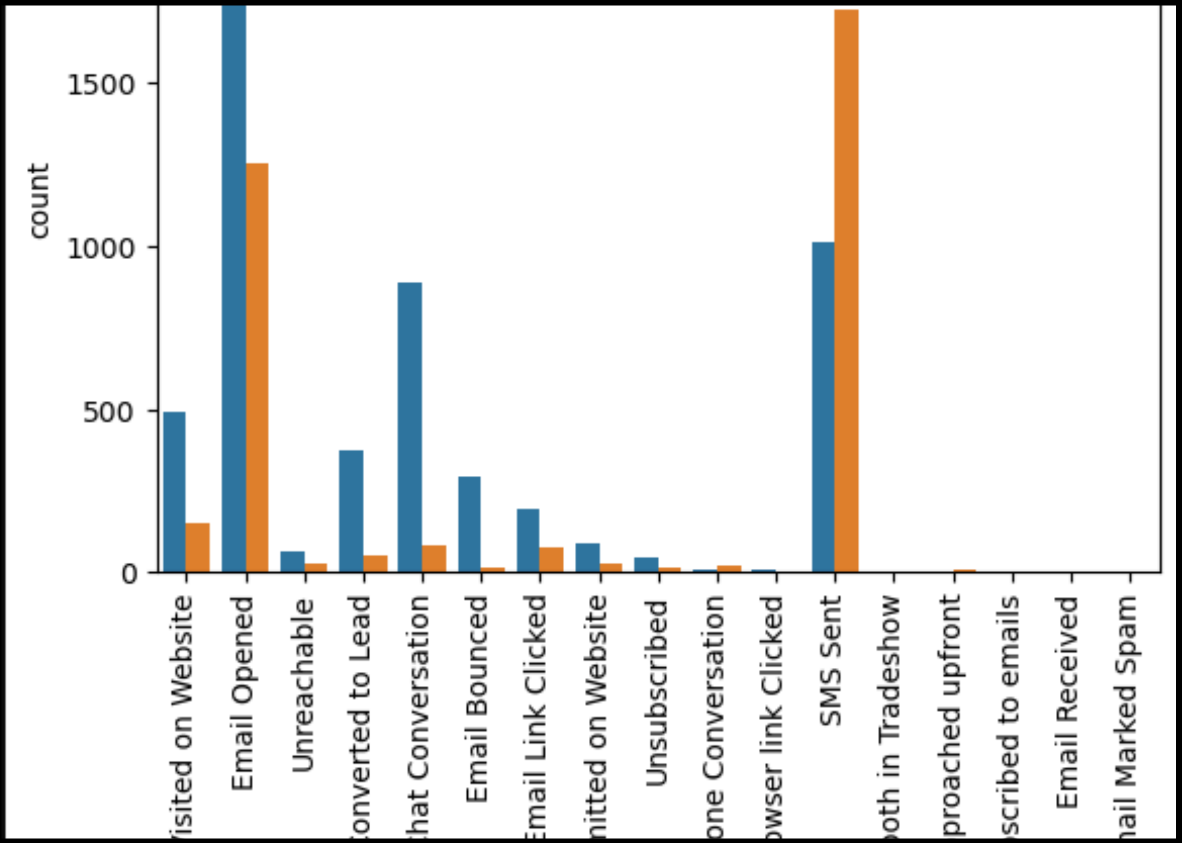
- We performed the Univariate and Bivariate analysis on the columns by plotting the scatter plot for numeric-numeric variables and count plot on categorical variables with target variable ‘Converted’.
- We observe from the Scatter Plot that the PageViews and TotalVisits are closely related which is expected. Some of the plots are shown below.

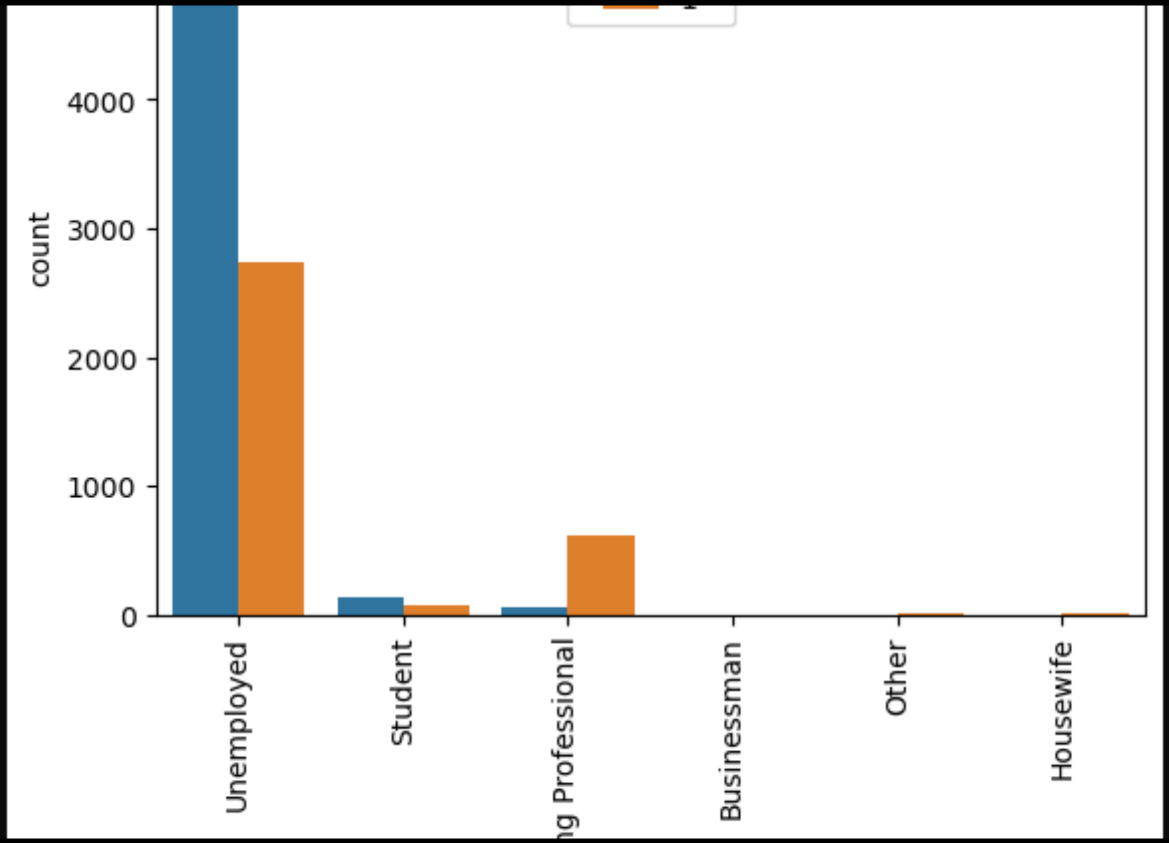


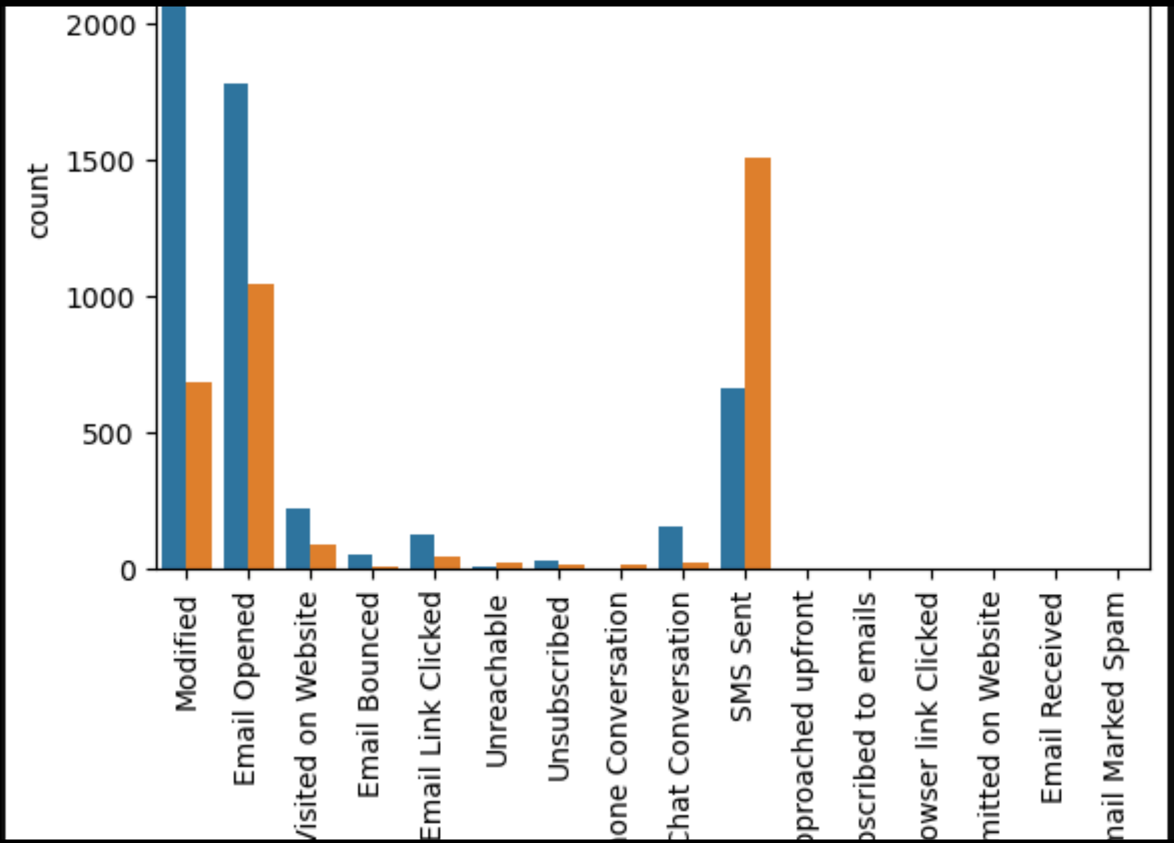
Bi-variate Analysis

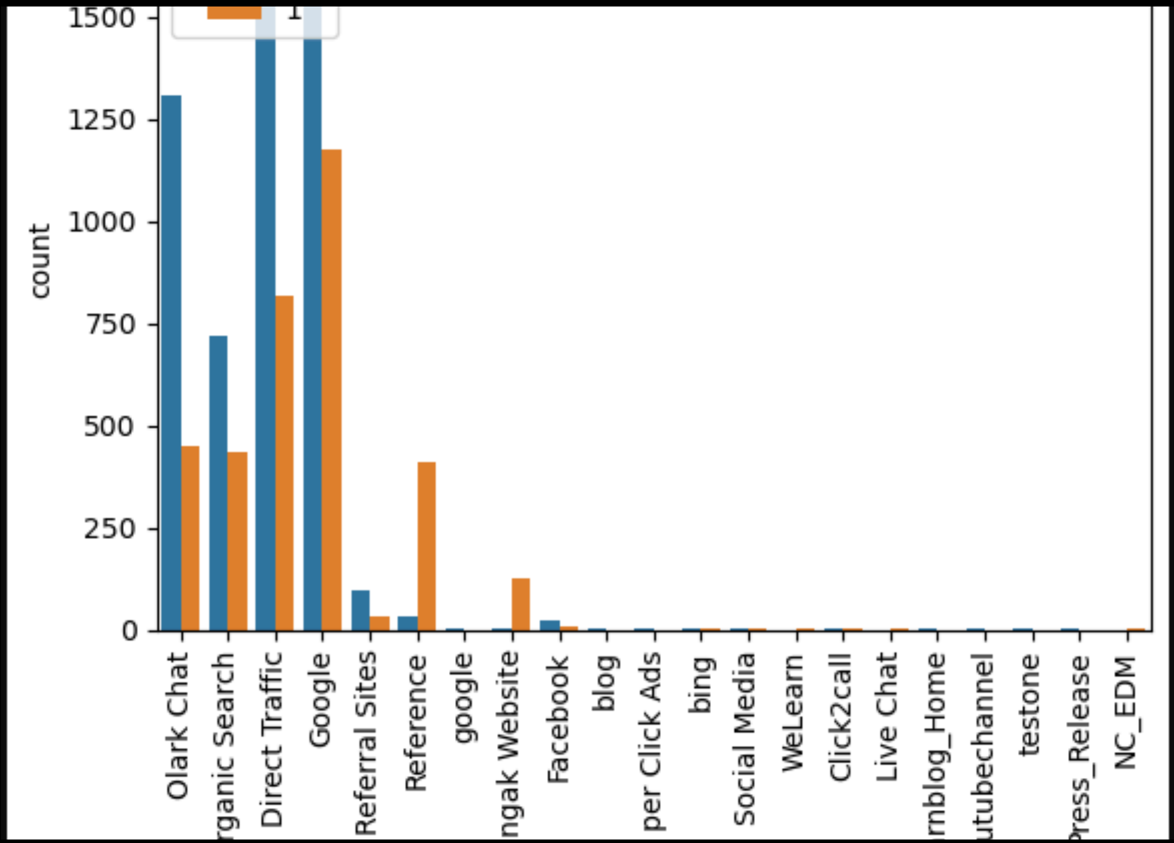
- From the Bi-variate analysis by plotting the bar plots for categorical variables with target variable Converted and aggregate function as count we derive some insights.
- We observe that many Unemployed, Working Professionals and Student's visit the website and the conversion rate is higher among the Working Professionals.
- Further from the plots we see that customers don't like be called and conversion rate is very high when the last activity after a phone conversation and SMS sent as the plots suggest.
- Specialization doesn't seem to play much role as customers from many backgrounds are interested in checking out the courses
- We also see that the NewsPaper Article, Digital Advertisement, Search, Through Recommendations, X Education Forums doesn't seem to influence the conversion rate much.
- Lead Source for conversion rate is pretty high among the Reference and Wellingak Website.
- Finally we dropped the columns that doesn't have much effect on the conversion rate before the model building step.

Bi-variate Bar Plots









Data Preparation

Feature Scaling

- We scaled the numerical features using StandardScaler and created a binary map for the columns with ‘Yes’ or ‘No’ entries.

Creating Dummies

- Using create_dummies from pandas library we create the dummy variables for columns with multiple levels and dropping the first columns to avoid redundancy.

Test-Train Split

- We split the data in Training and Test datasets with 70/30 split and extract the X (independent) and y (Target) variables for Test and Train.

Model Building

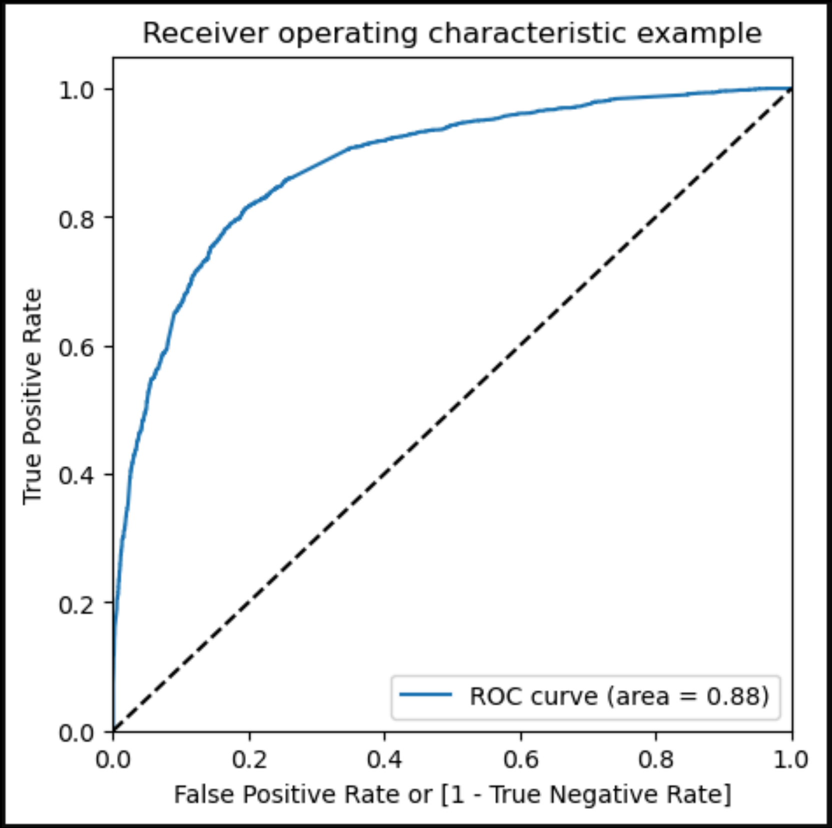
Use RFE for Feature Selection

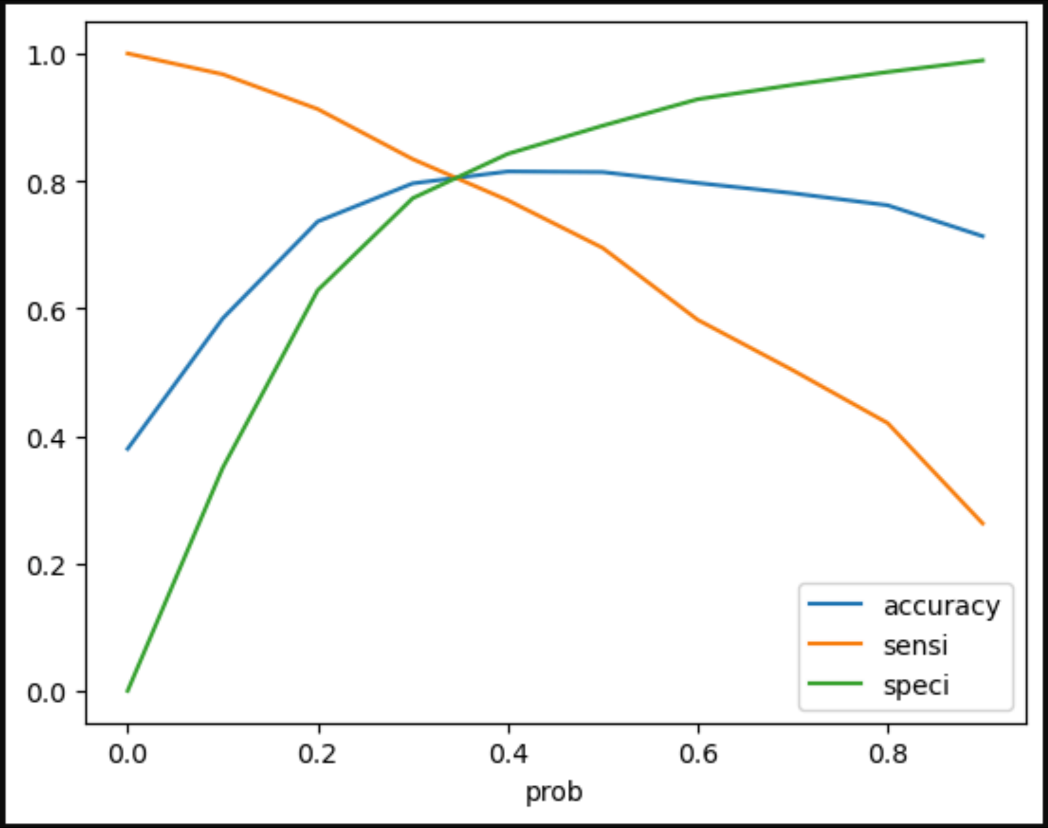
- We select 20 features by performing RFE on the Train dataset. Columns that support are extracted into a list.
- Correlation between these features is examined using a heat map.
- We used GLM from stats models to build the first model and generate model summary. After checking for the features with higher significance we drop them and proceed with building next model with reduced features.
- We look at the VIF's for every model along the way to see if all the features are below 5.
- Iterating though we end with our Third model whose summary is line with significance and VIF. We then made predictions on the third model. Cut-off is chosen randomly to be 0.5 and a data frame is constructed to hold the actual and predicted probabilities.
- We look at the confusion matrix and calculated the model accuracy which comes out to be 81.38% on the Training dataset.

Model Metrics

ROC & AOC

- We plot the ROC (Receiver Operating Characteristic Curve) to check if the model is optimal.
- We then plotted the Accuracy, Sensitivity and Specificity curve for different predicted probabilities where cut-off's ranging from (0-0.9) using confusion matrix.
- The plot clearly shows that all the three metrics Accuracy, Sensitivity and Specificity coverage at probability 0.3 so we choose it to be the optimal cutoff point.
- The we use the calculated probabilities from the model and apply the new cut-off 0.3 on it. The final predicted probabilities are saved to the result data frame.
- We look at all the metrics from the confusion matrix especially the sensitivity (83.42%)and the accuracy(79.61) of the model on Training Dataset.





Testing the Model

- We use the Testing dataset and use the model to predict the probabilities.

Model Metrics on Test Data

- Model Accuracy - 80.04%
- Model Sensitivity - 84.74%
- Model F-Score - 0.76

KPI’s

List of Features that contribute most to the Lead Conversion rate.

From the model few of the the identified KPI’s are given below

1. Lead Origin_Lead Add Form
2. Lead Source_Welingak Website
3. Last Activity_Had a Phone Conversation
4. Last Activity_Olark Chat Conversation
5. Occupation_Working Professional
6. Occupation_Student
7. Lead Origin_Lead Import
8. Last Notable Activity_SMS Sent
9. Do Not Email
10. Total Time Spent on Website