Lead Score Case Study

Submitted by:

Chetna Sahu

Sweta Mishra

Problem Statement

- X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.
- Once these people land on the websites, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals.
- 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%.

Business Objectives

- To help X education to select the most promising leads known as ‘hot leads’ who are most likely to convert into paid customers.
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads where the leads with higher lead score have a higher conversion chance and the leads with lower lead score have lower conversion chance.
- Identify the outliers, if any in the dataset and justify the same.
- Consider both technical and business aspects while building the model.
- Identify the driver variables and understand their significance which are strong indicators of lead conversion.
- Summarize the conversion predictions by using evaluation metrics like accuracy, sensitivity, specificity and precision.

Problem Solving methodology

- Data Sourcing, Cleaning and Preparation
  - Read the data from source
  - Convert data into clean format suitable for analysis
  - Outliers Treatment
  - Exploratory Data Analysis
  - Feature Standardization
- Feature Scaling and Splitting Train and Test Sets
  - Feature Scaling of Numeric data
  - Splitting data into Train and Test set

Problem Solving methodology

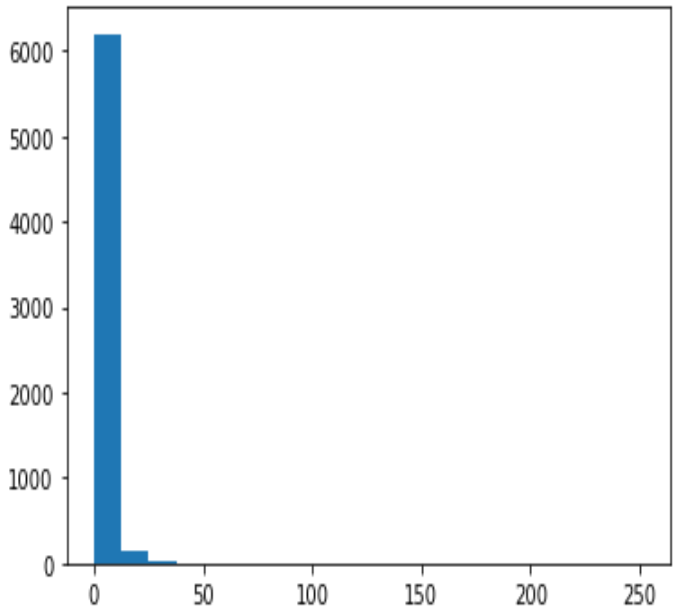
- Model Building
  - Feature selection using RFE
  - Determine the optimal model using Logistic Regression
  - Calculate various metrics like accuracy, sensitivity, specificity, precision & recall and then evaluate the model
- Results
  - Determine the lead score and check if target final predictions amount to 80% conversion rate.
  - Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

Exploratory Data Analysis

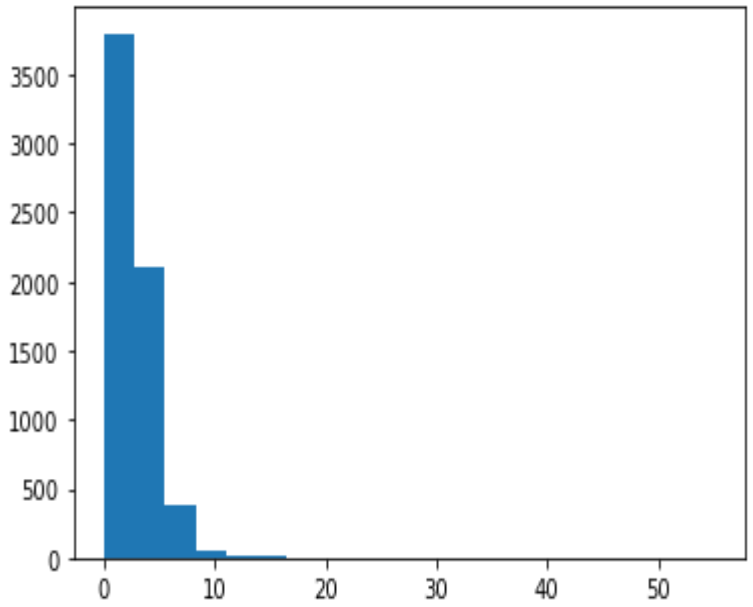
- **Numerical columns:**

As we can see from the below diagrams that it shows High peaks and skewed data. There might be a possibility of outliers. We need to treat those outliers.

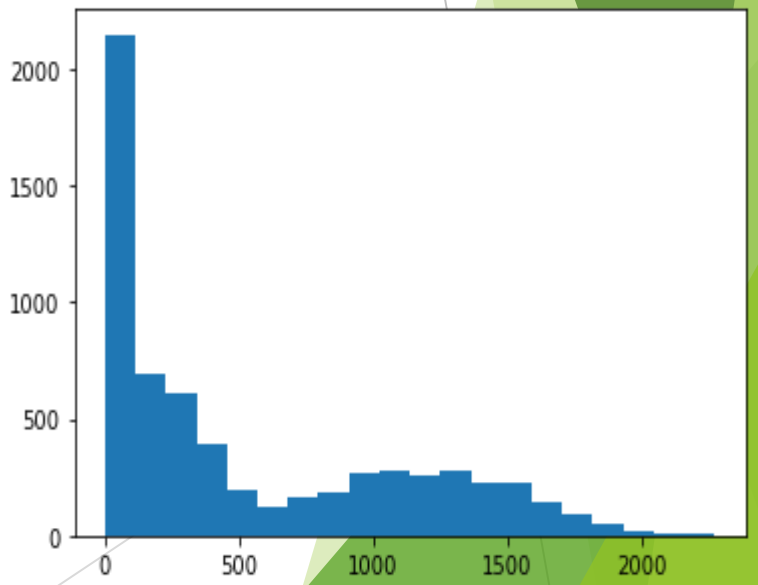
Total Visits



Page Views Per Visit



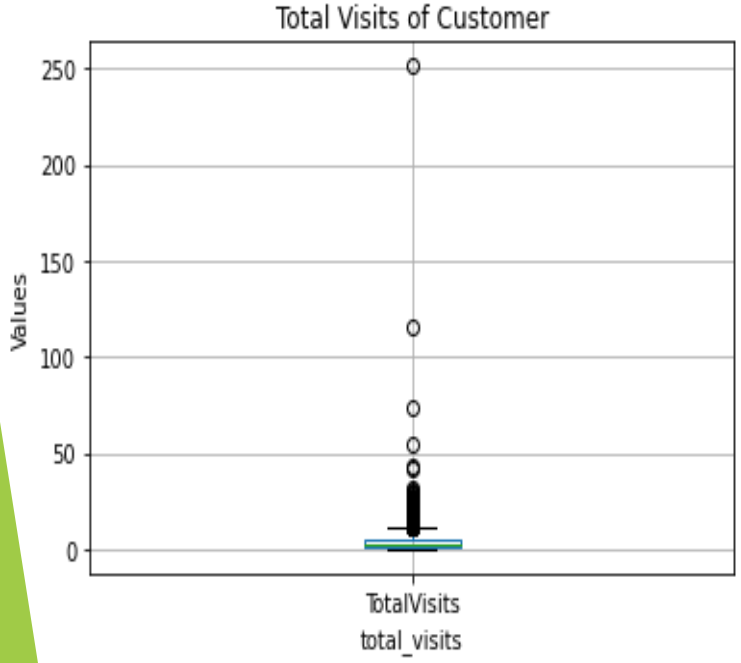
Total Time Spent on Website



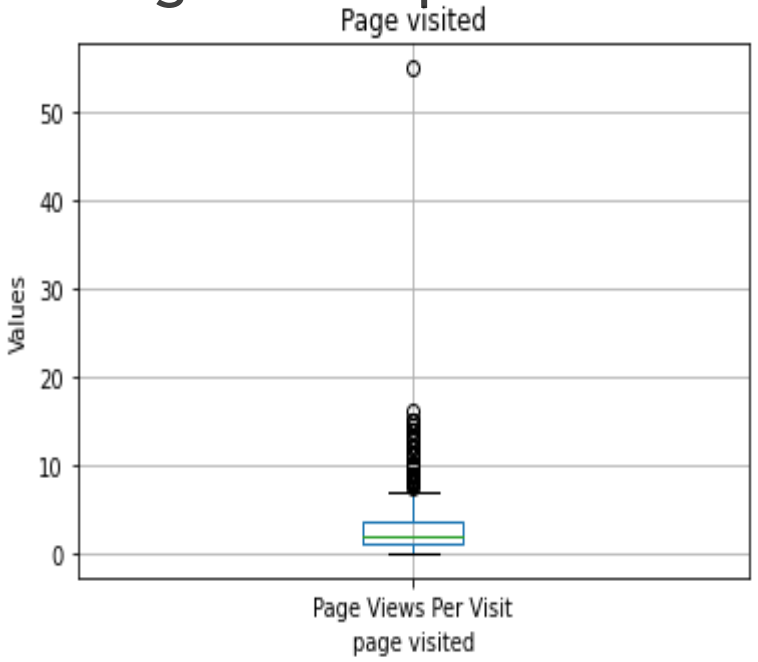
Check for the Outliers

As we can see that 'Total Visits' and 'Page Views Per Visit' variables have huge outliers. So we will use 1.5 IQR rule to remove outliers.

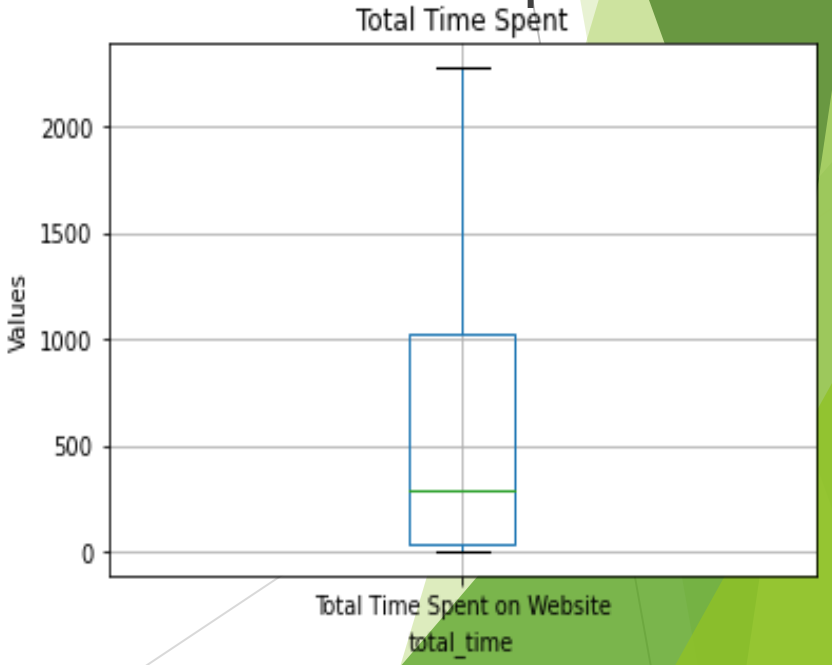
Total Visits



Page views per visit



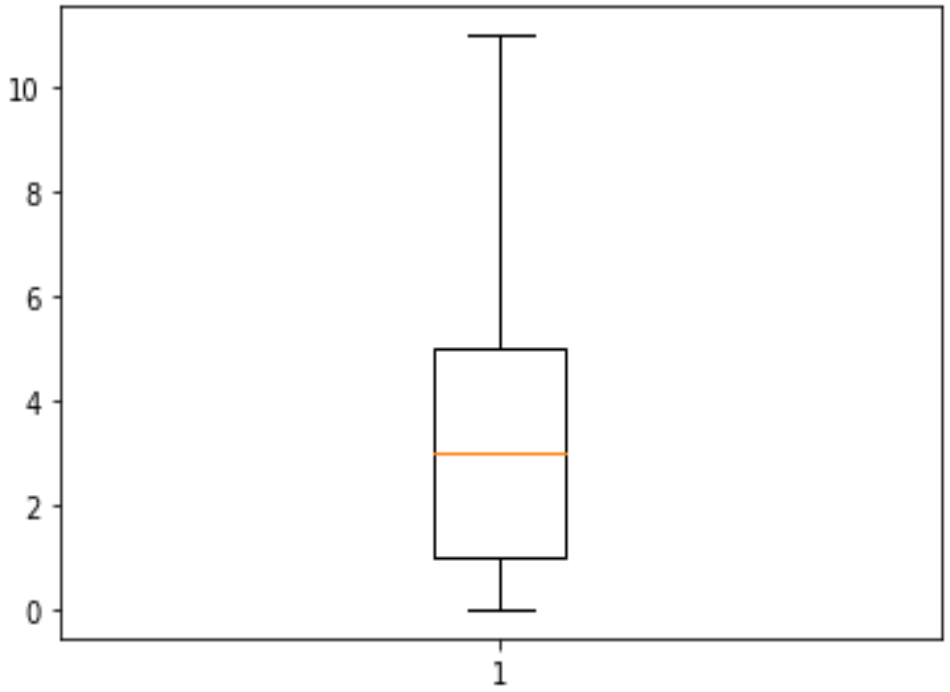
Total time spent



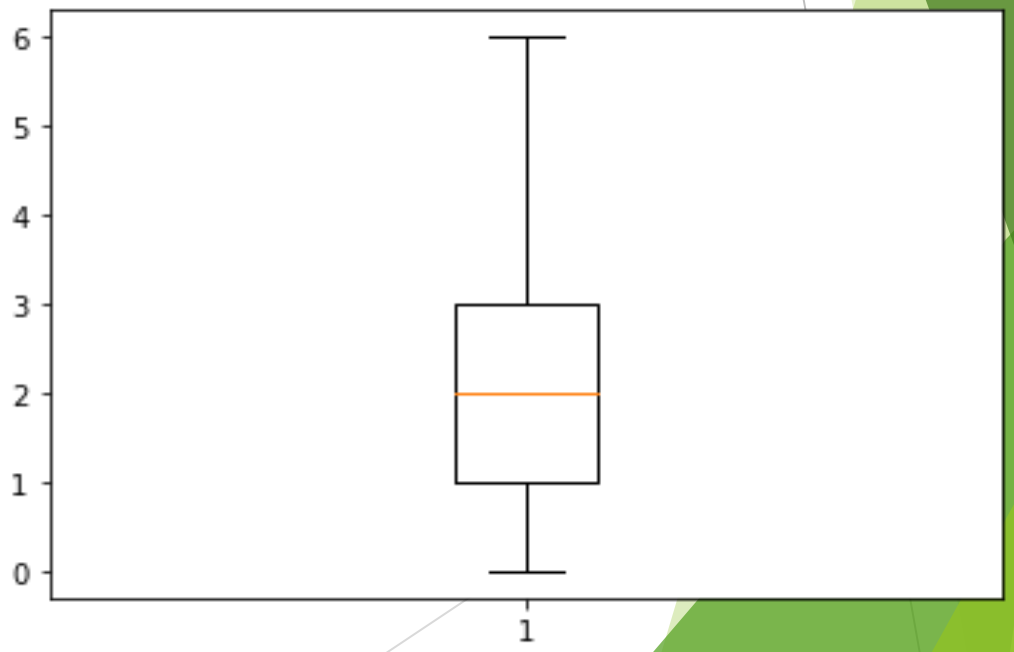
After performing outlier treatment

Looking at both the box plots and the statistics, there are upper bound outliers in both total visits and page views per visit columns. We can also see that the data can be capped at 99 percentile.

Total Visits

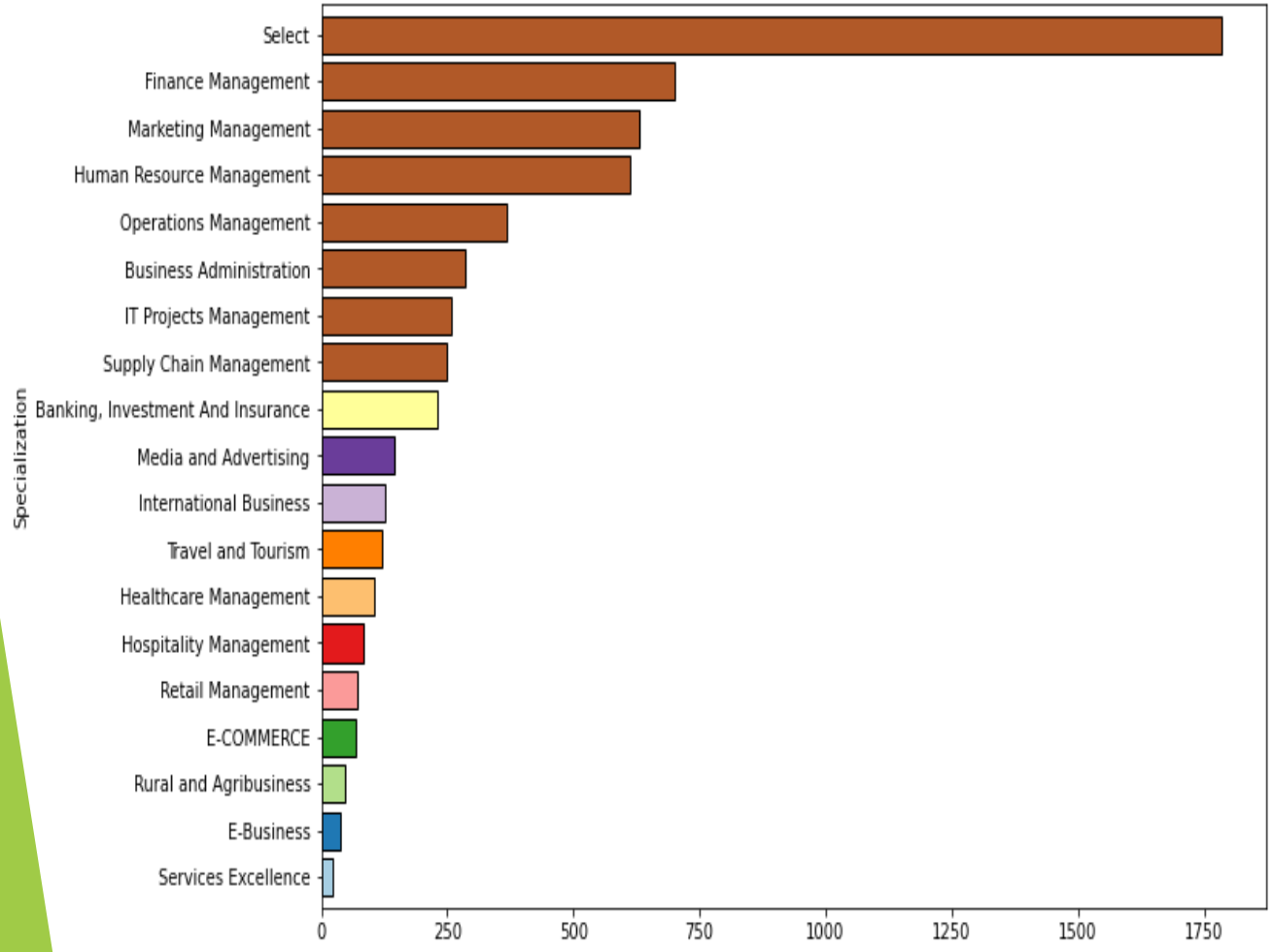


Page views per visit



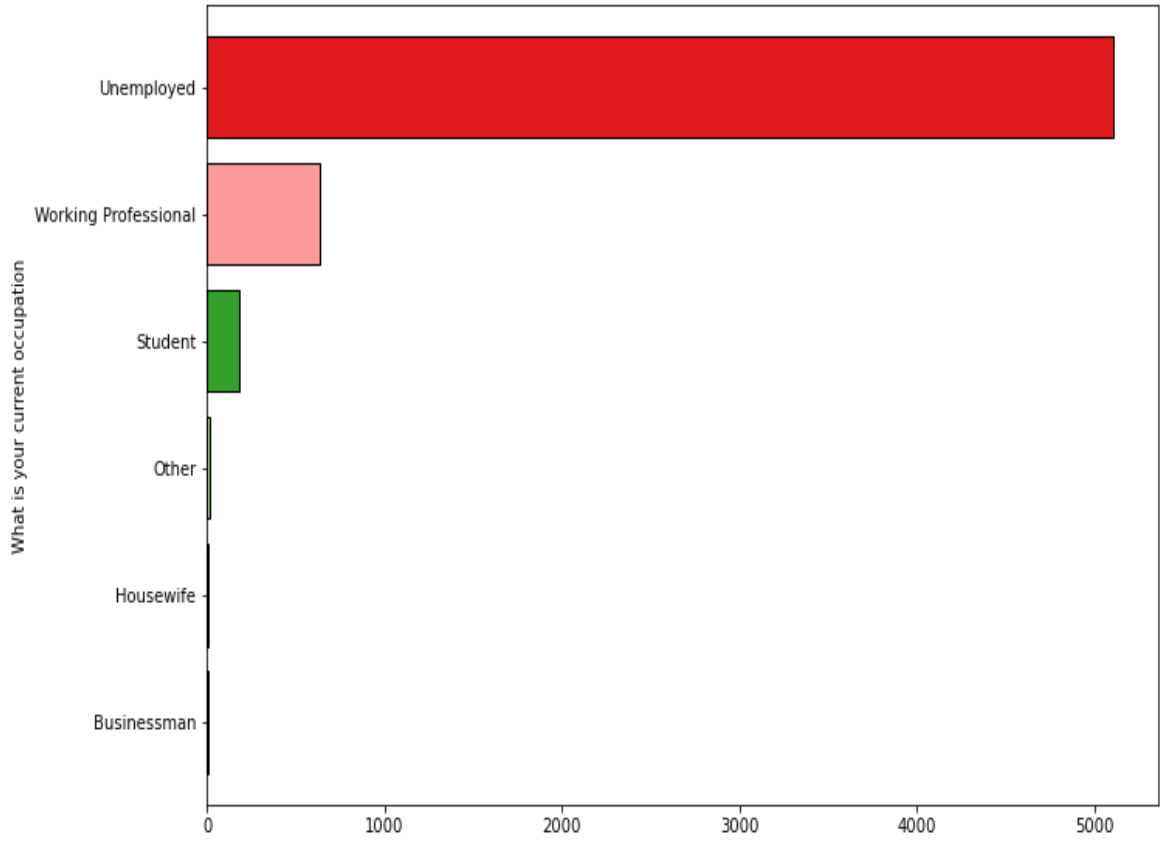
Categorical columns

- Specialization



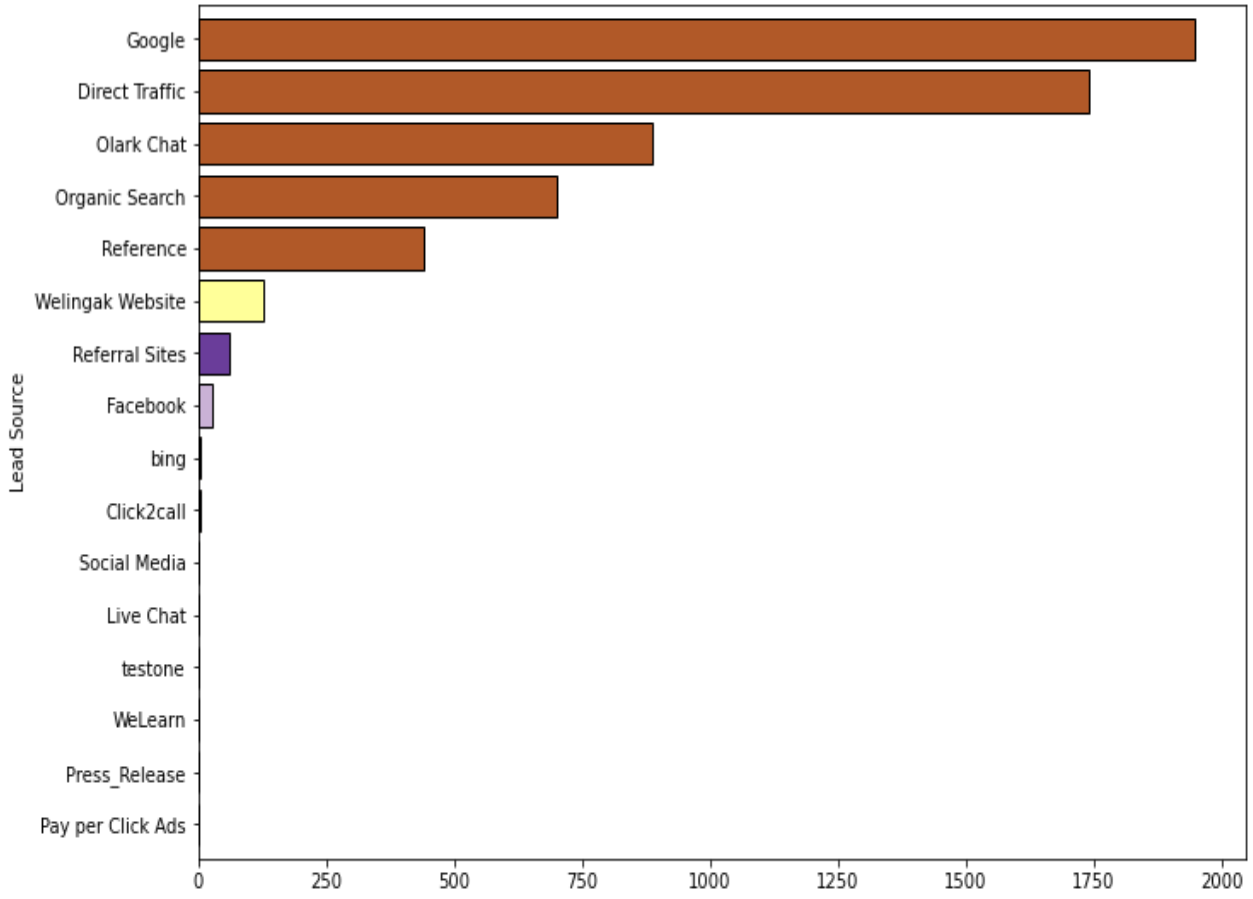
- The select category is of no use it needs to be removed.
- As we can look, most of the specialization taken are Finance management then comes the Marketing Management and then Human Resource Management.

Occupation



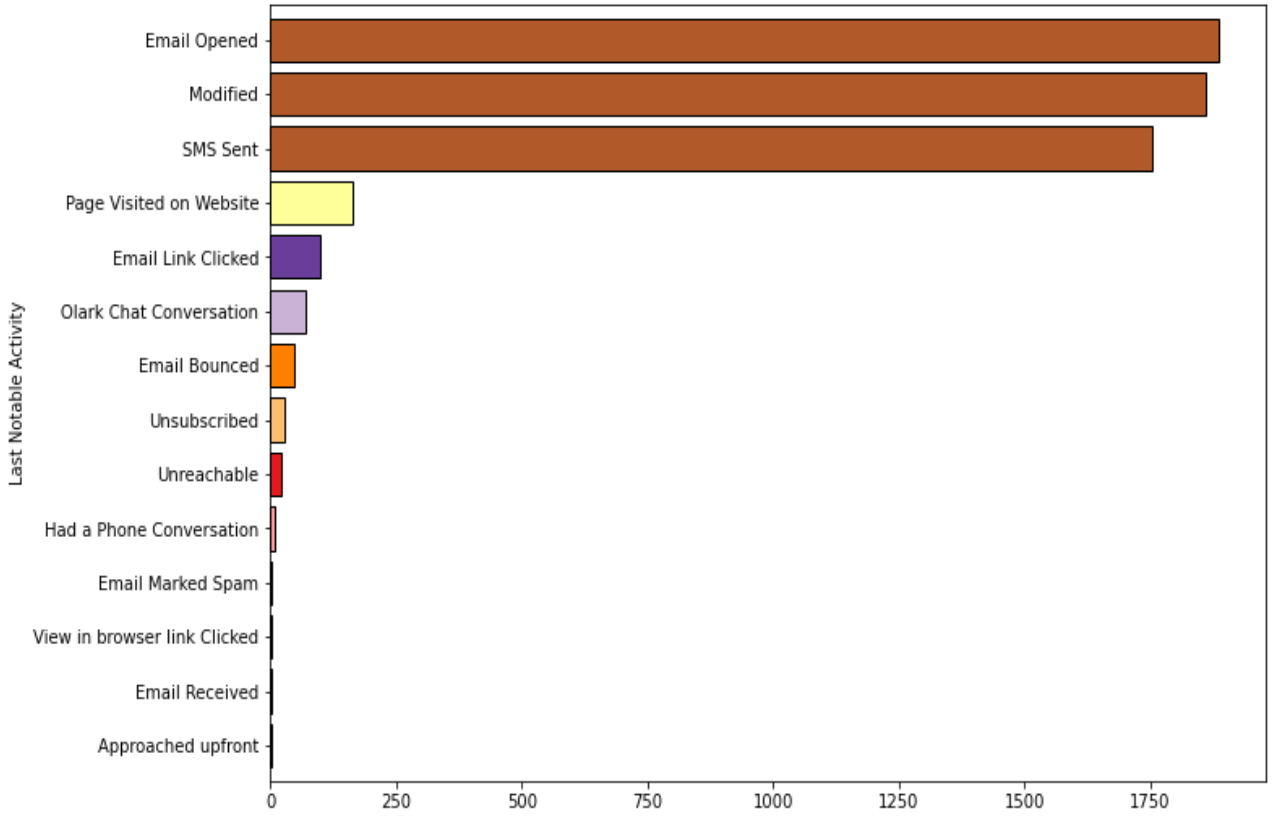
- As we can see Unemployed users are the most significant leads.

Lead Source



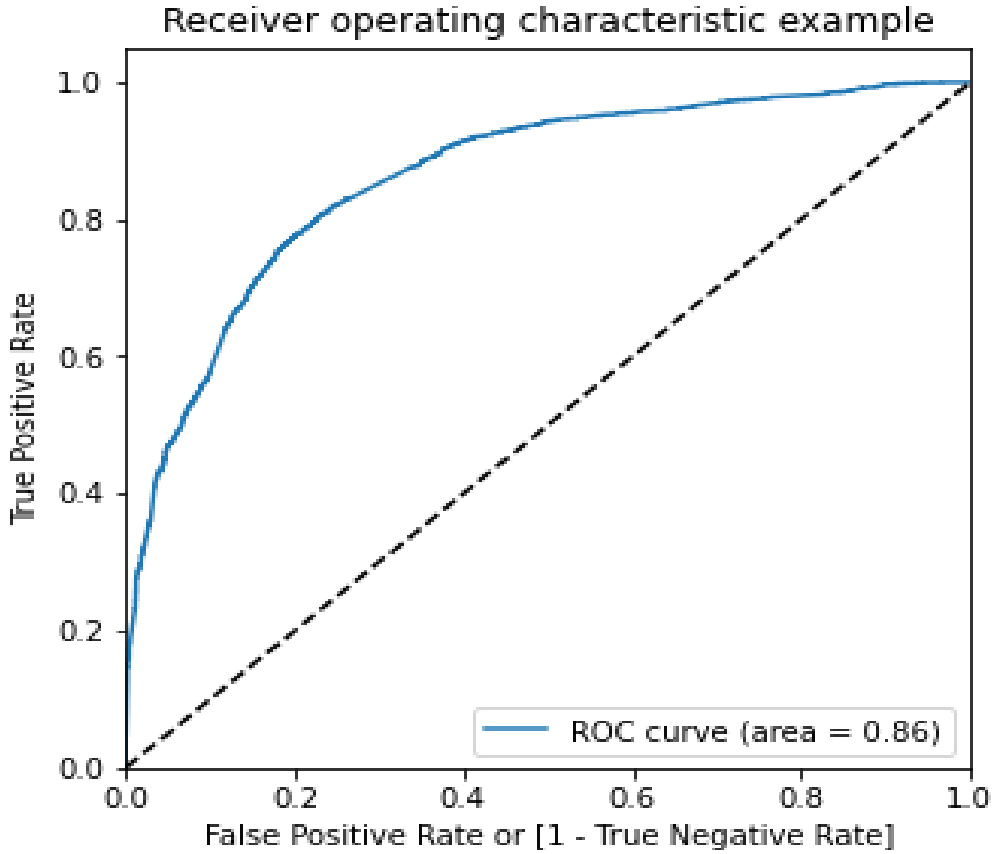
- Google has the highest percentage of lead source.

Last Notable Activity



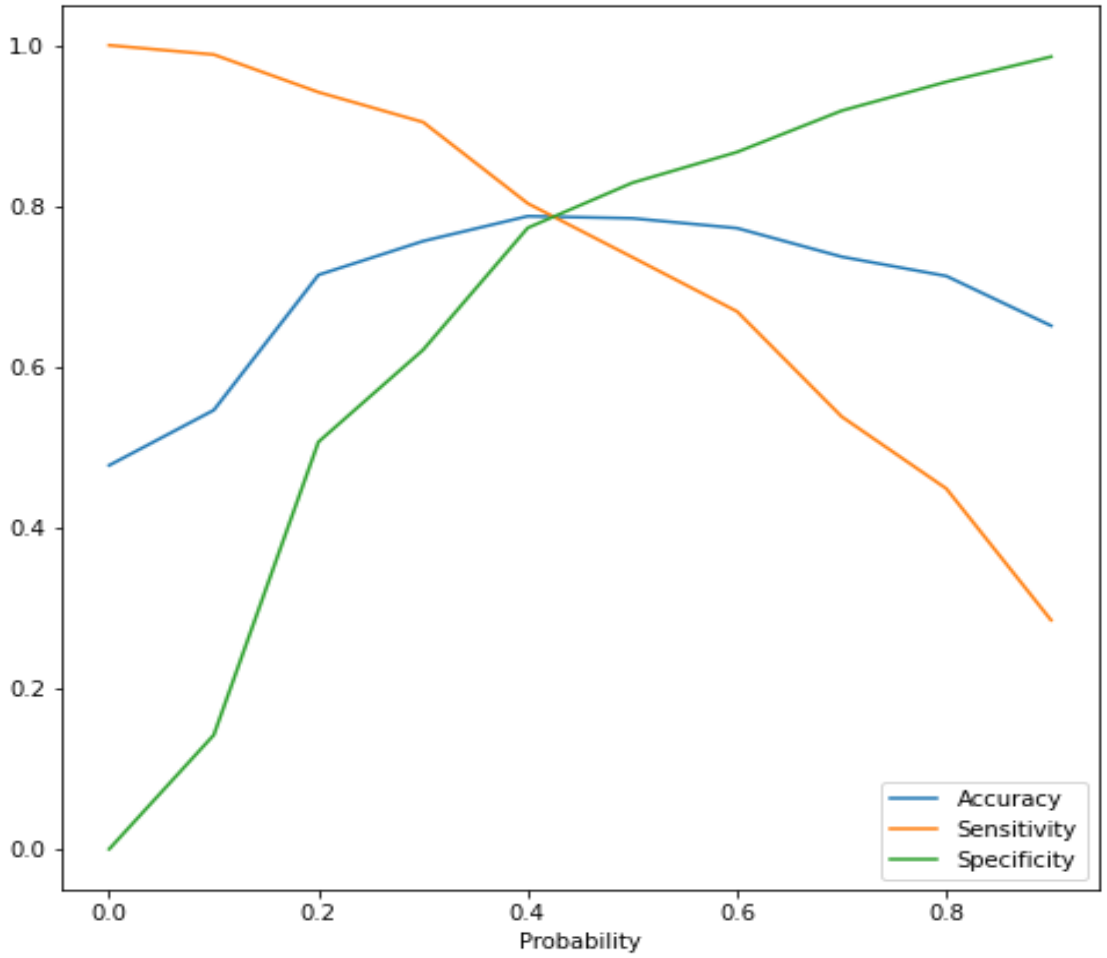
- Opened Email has the highest percentage of Last Notable Activity.

Receiver Operating Characteristic curve(ROC)



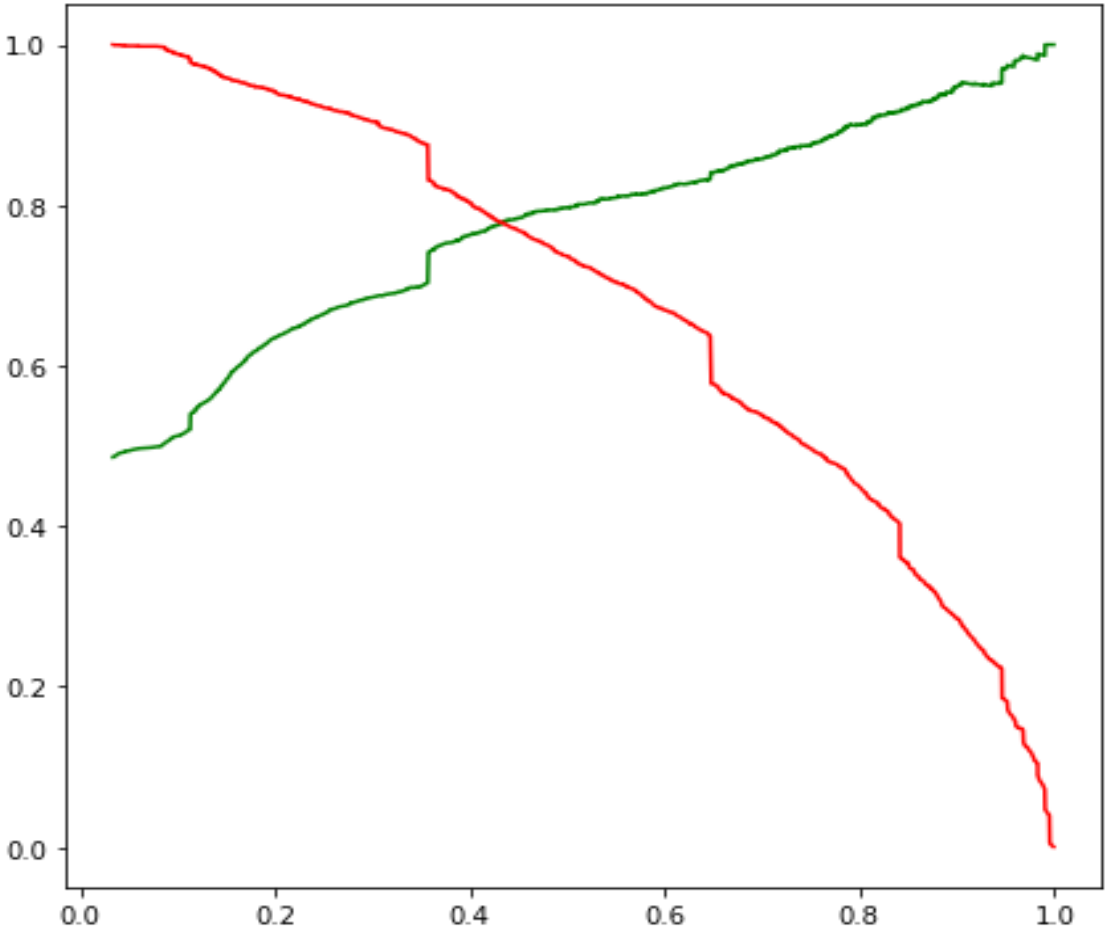
- ROC curve shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
- The area under the curve of the ROC is 0.86 which is quite good. So we seem to have a good model. Let's also check the sensitivity and specificity tradeoff to find the optimal cutoff point.

Model Evaluation: On Train Set



- As we can see that around 0.42, you get the optimal values of the three metrics. So we will choose 0.42 as our cutoff now.
- True negative(TN)= 1803
- True positive(TP)= 1463
- False negative(FN)= 524
- False positive(FP)= 692
- Accuracy= (TP+TN)/(TN+TP+FN+FP)= 0.78~78%
- Specificity= TN/(TN+FP) = 0.82~82%
- Sensitivity= TP/(TP+FN)= 0.73~73%

Precision and Recall



- The graph depicts an optimum cut off of 0.42 based on Precision and Recall.
- Precision= TP/(TP+FP)=0.79~79%
- Recall= TP/(TP+FN)=0.73~73%

Model Evaluation on Test Set

- Confusion metrics

  631 287

  111 756

- True negative= 631
- True positive= 756
- False negative= 111
- False positive= 287
- Sensitivity= 0.87~87%
- Specificity= 0.68~68%
- Precision= 0.72~72%
- Recall= 0.87~87%
- Accuracy= 0.77~77%

Conclusion

- We checked both Sensitivity-Specificity and Precision-Recall metrics, we have considered the optimum cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity, Specificity values for test set are 77%, 87% and 68% respectively and that of train set are 78%,73% and 82% respectively.
- The top 3 variables that contributes for lead getting converted in the model are: a) Lead Source Welingak Website b) Last Notable Activity Unreachable c) What is your current occupation Working Professional.
- Overall our model is good.