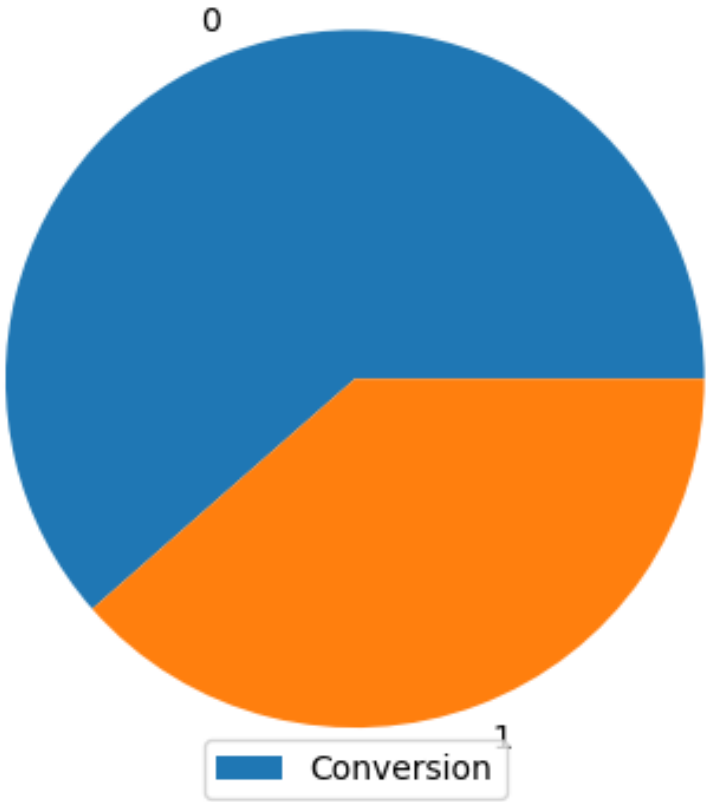
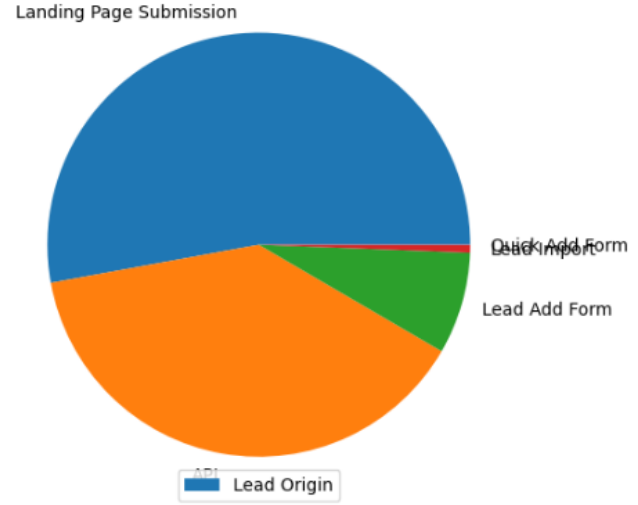
Lead Scoring Case Study : Presentation
Submitted by Anuj Sharma

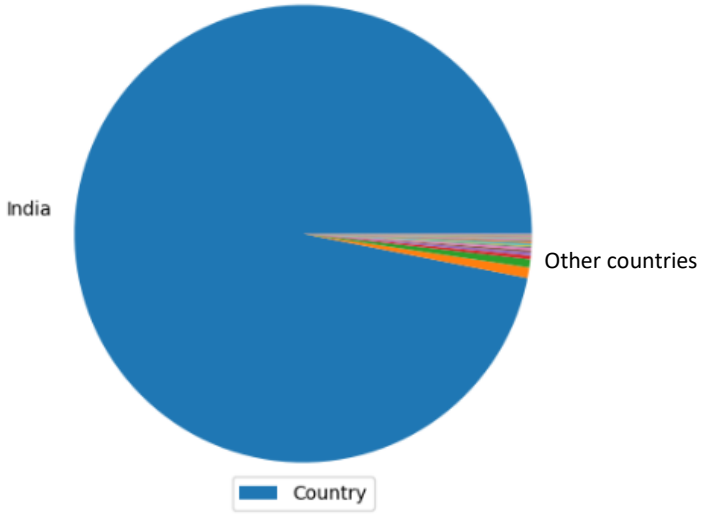
Leads Conversion Percentage

Conversion Percentage is : 39.0 %

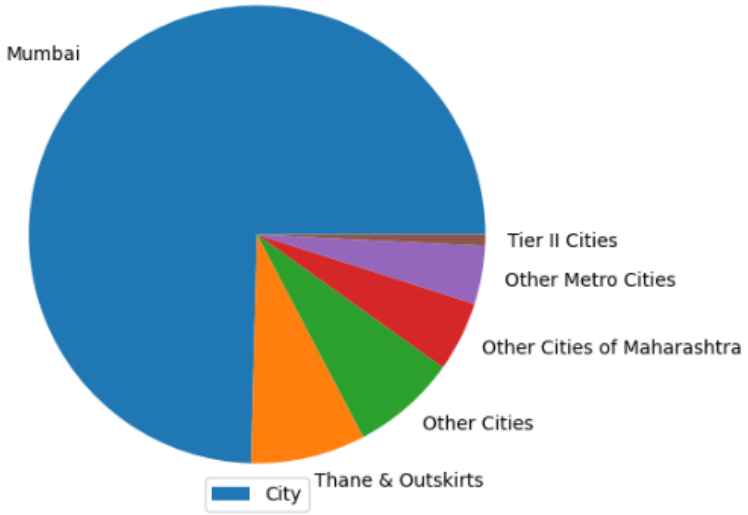


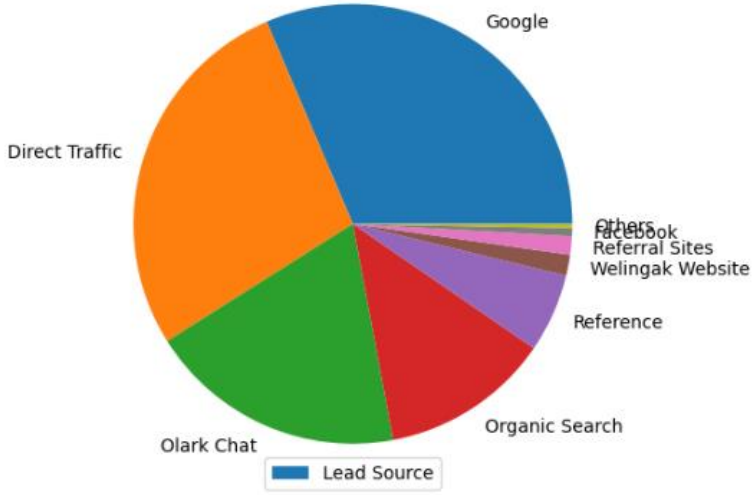
Lead Source & Origin
Leads Demographic Distribution





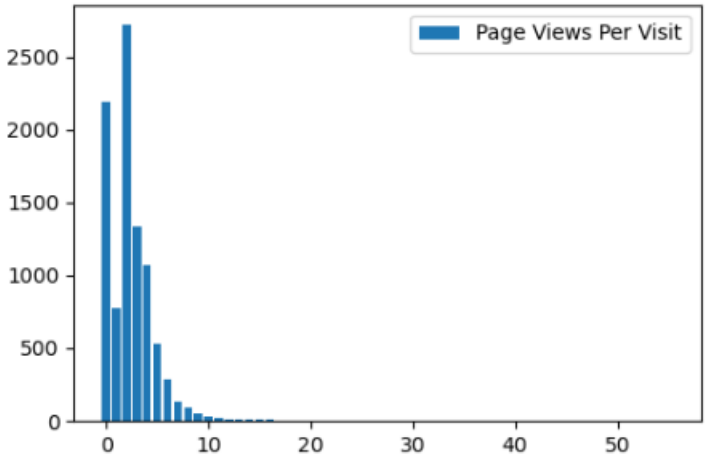
Maximum Leads in the Data Set are from Mumbai, India

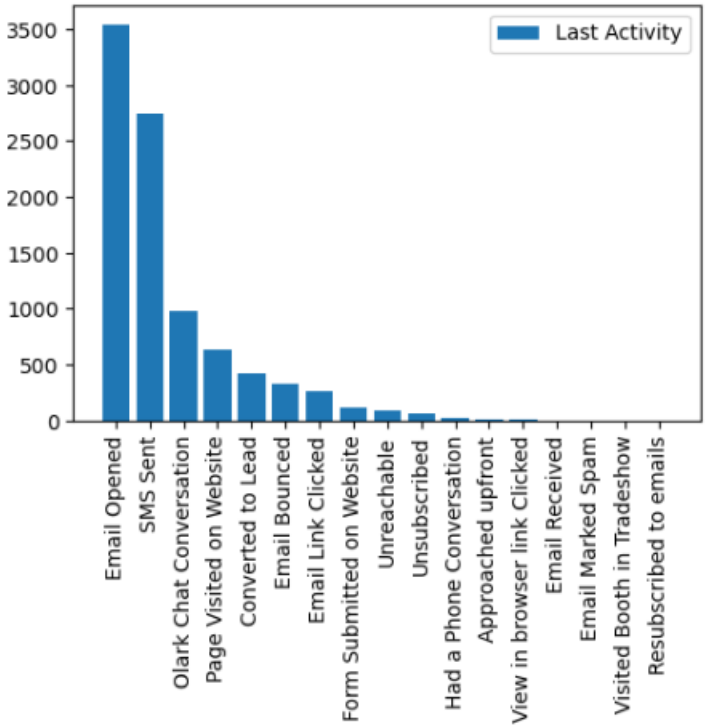


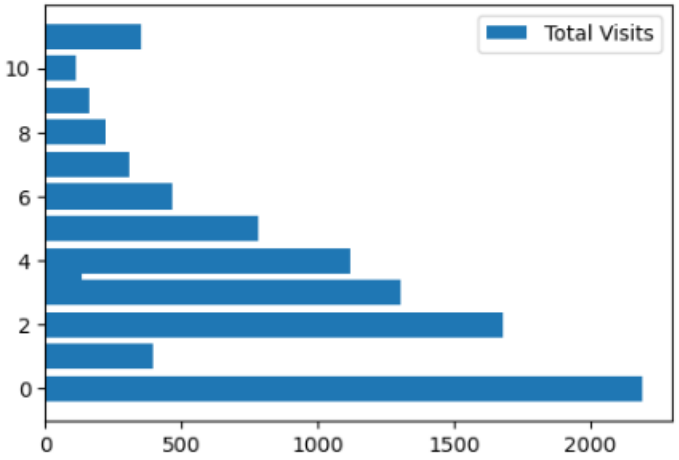


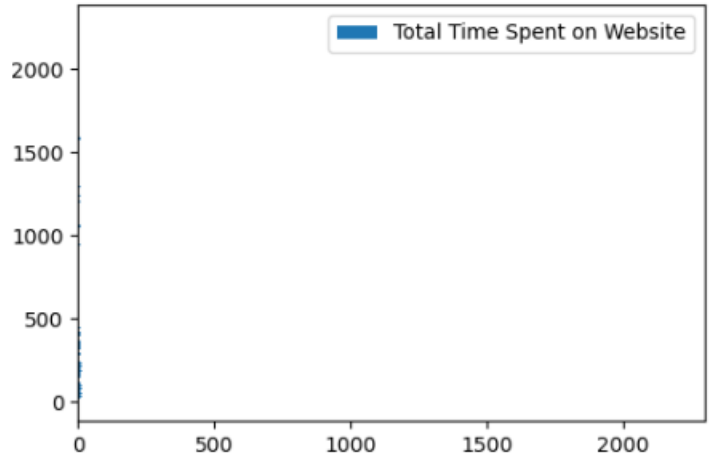
The Maximum Lead Sources are through Direct Search methods on Internet

Leads’ Time Spent on Website

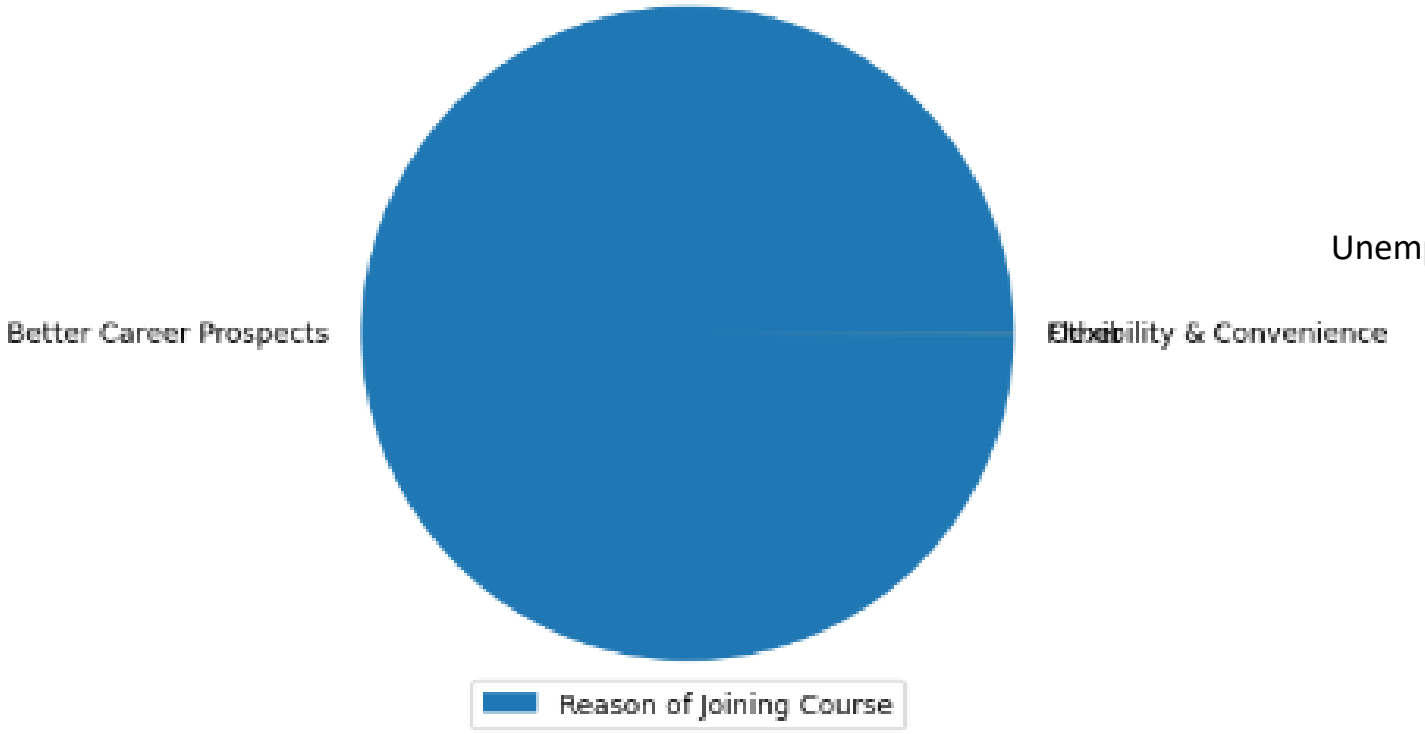


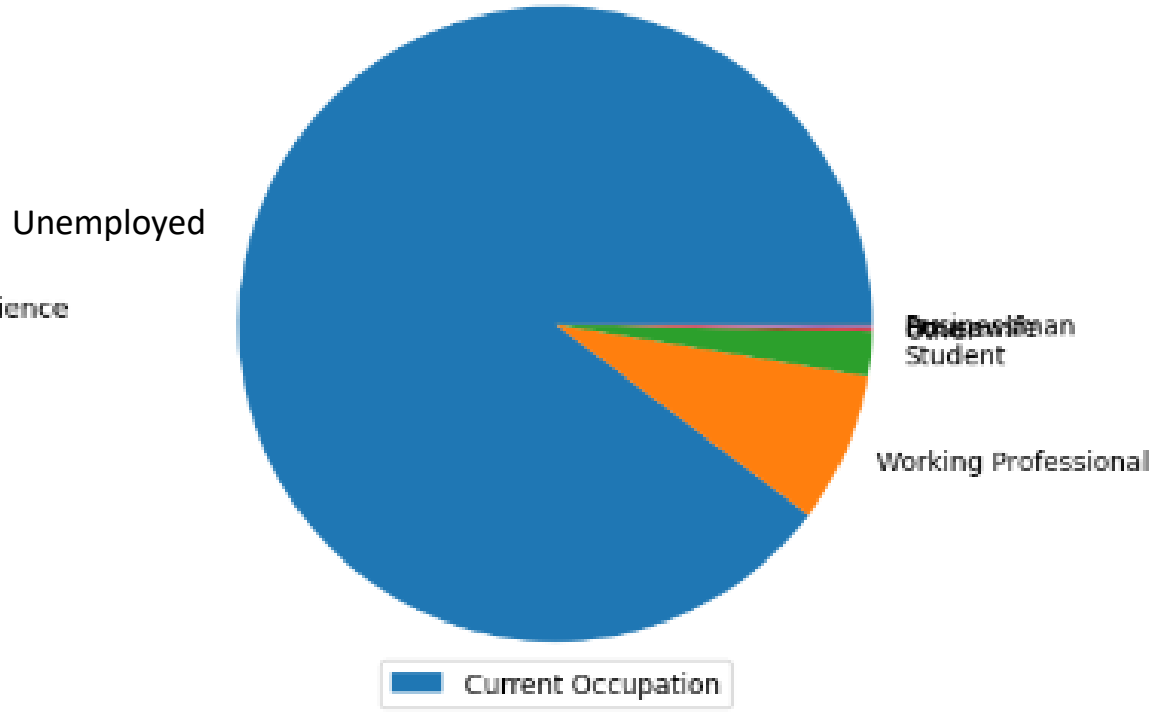




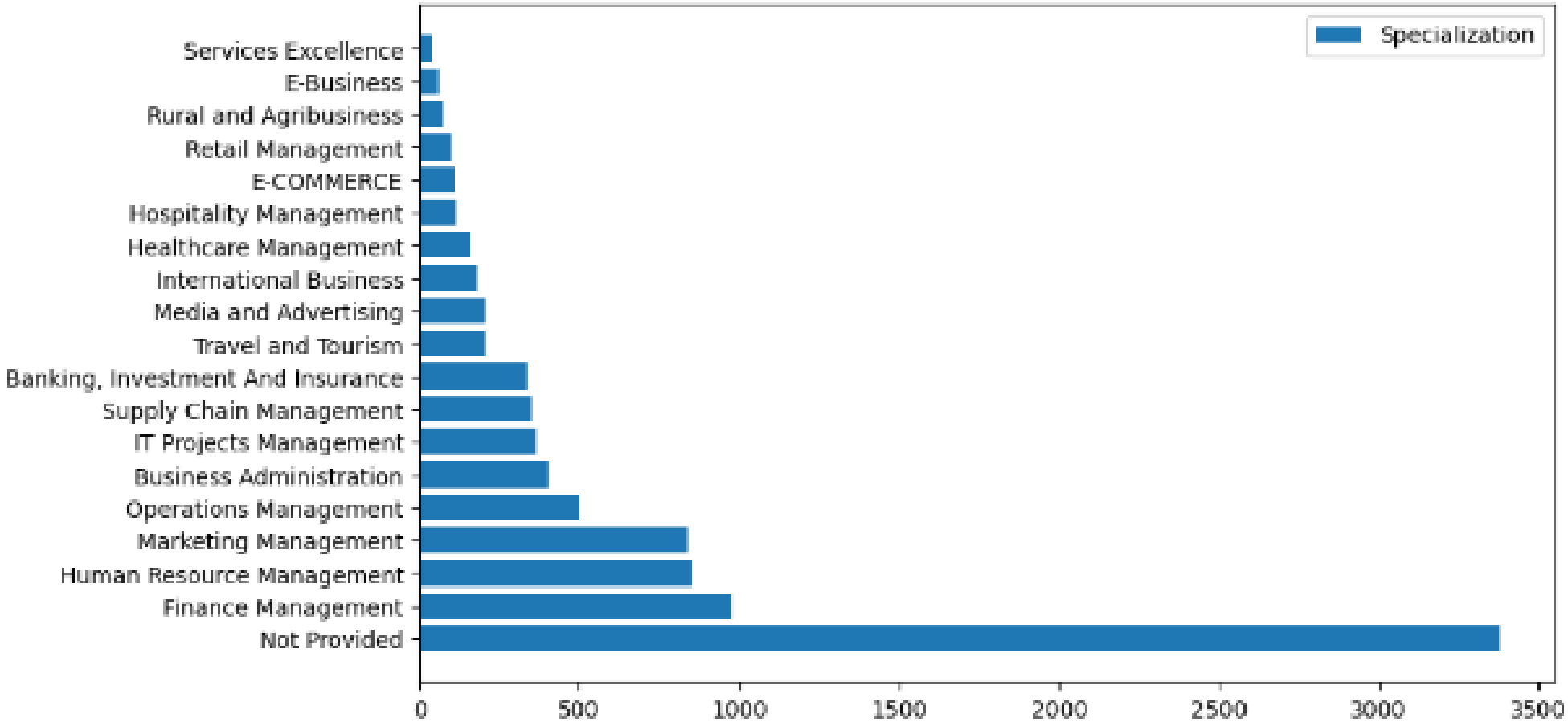


Leads’ Reason of Joining the Course vs Current Occupation

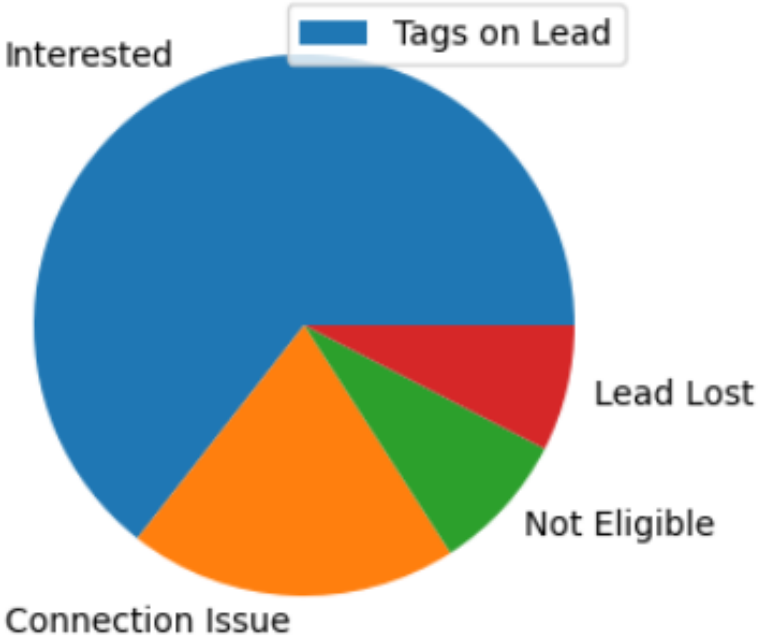




Specialization Distribution



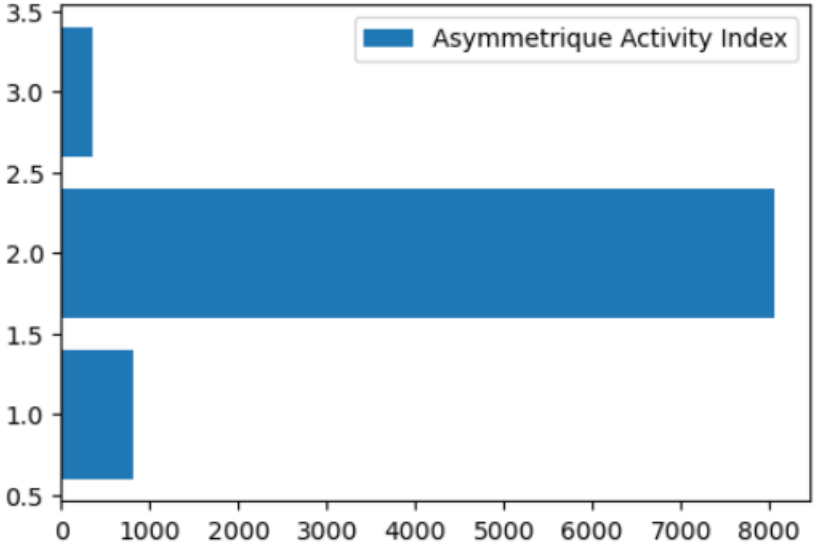
Leads Quality, Tags & Profile – Most Relevant Features in the Logistics Regression Model

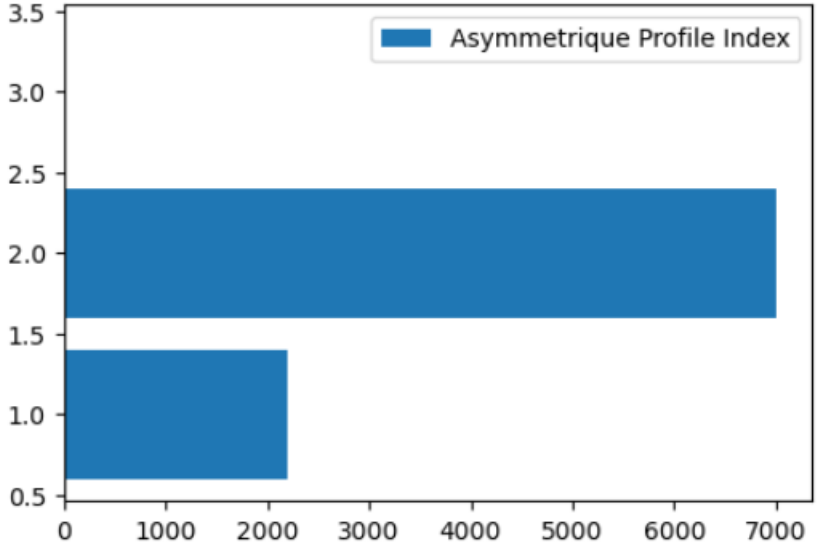


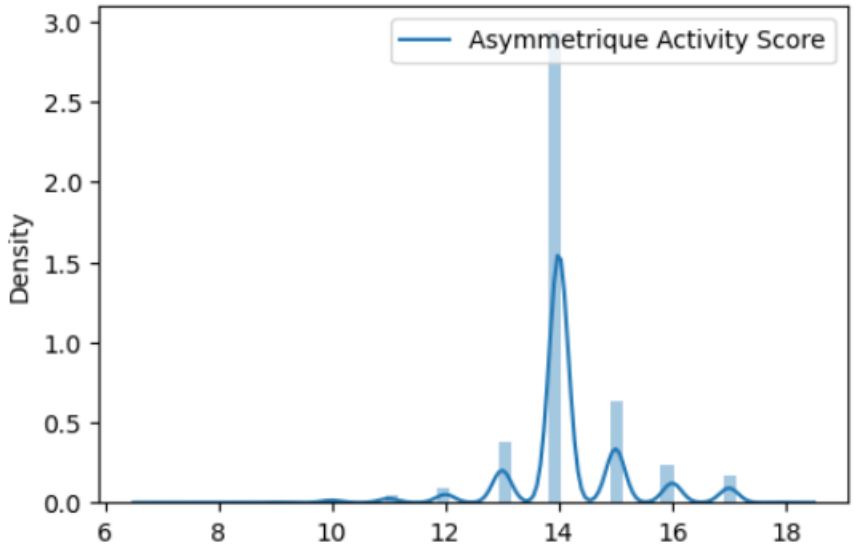
Leads Generated through Advertisement Program
-Need to investigate, as hardly leads are generated through advertisements

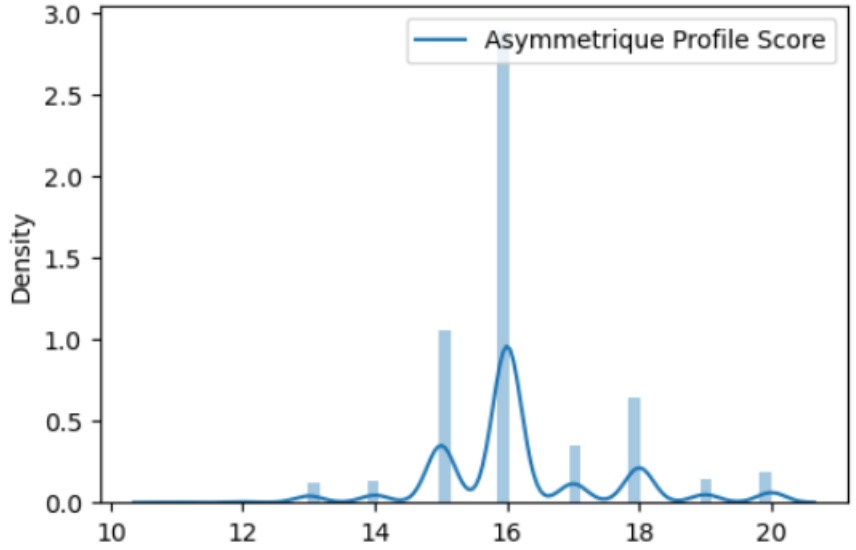

```
0    9221
1      19
Name: Advertisement, dtype: int64
```

Leads Profile Scoring and Activity Matrix

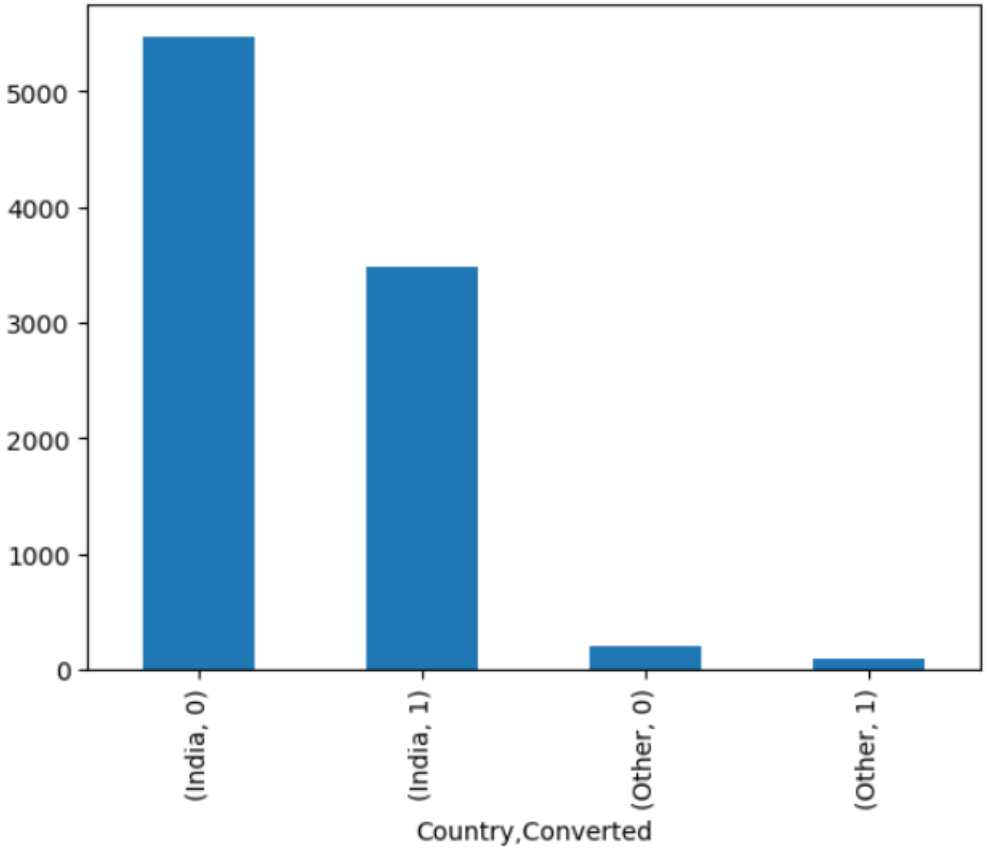


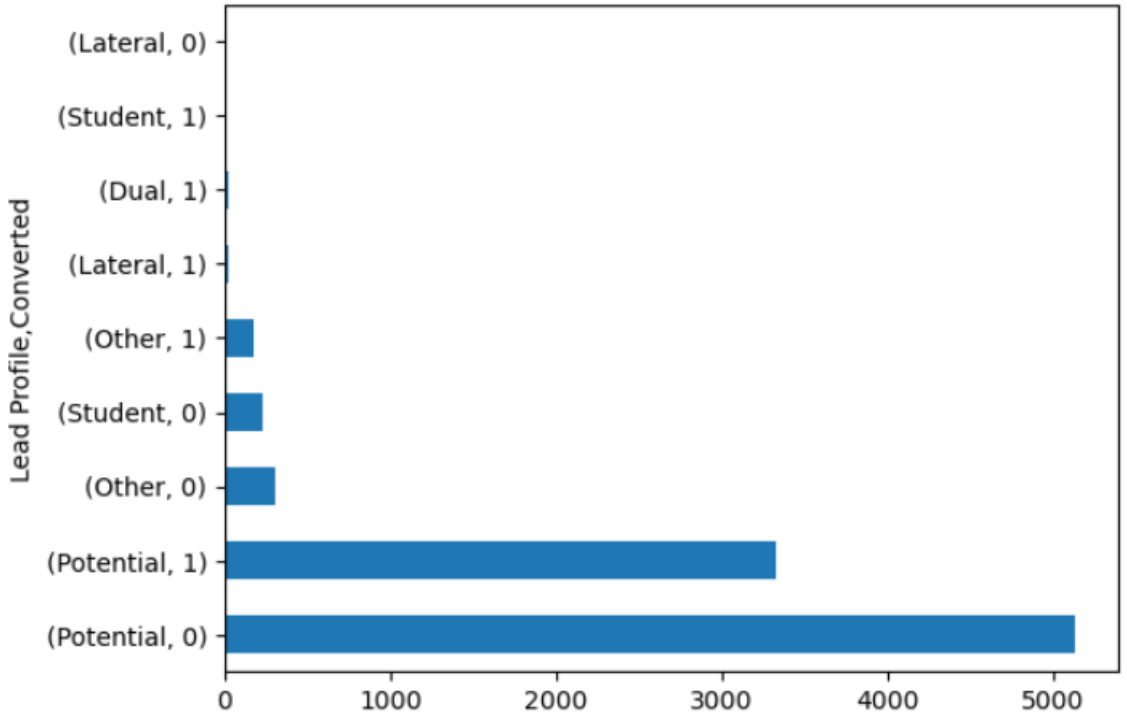




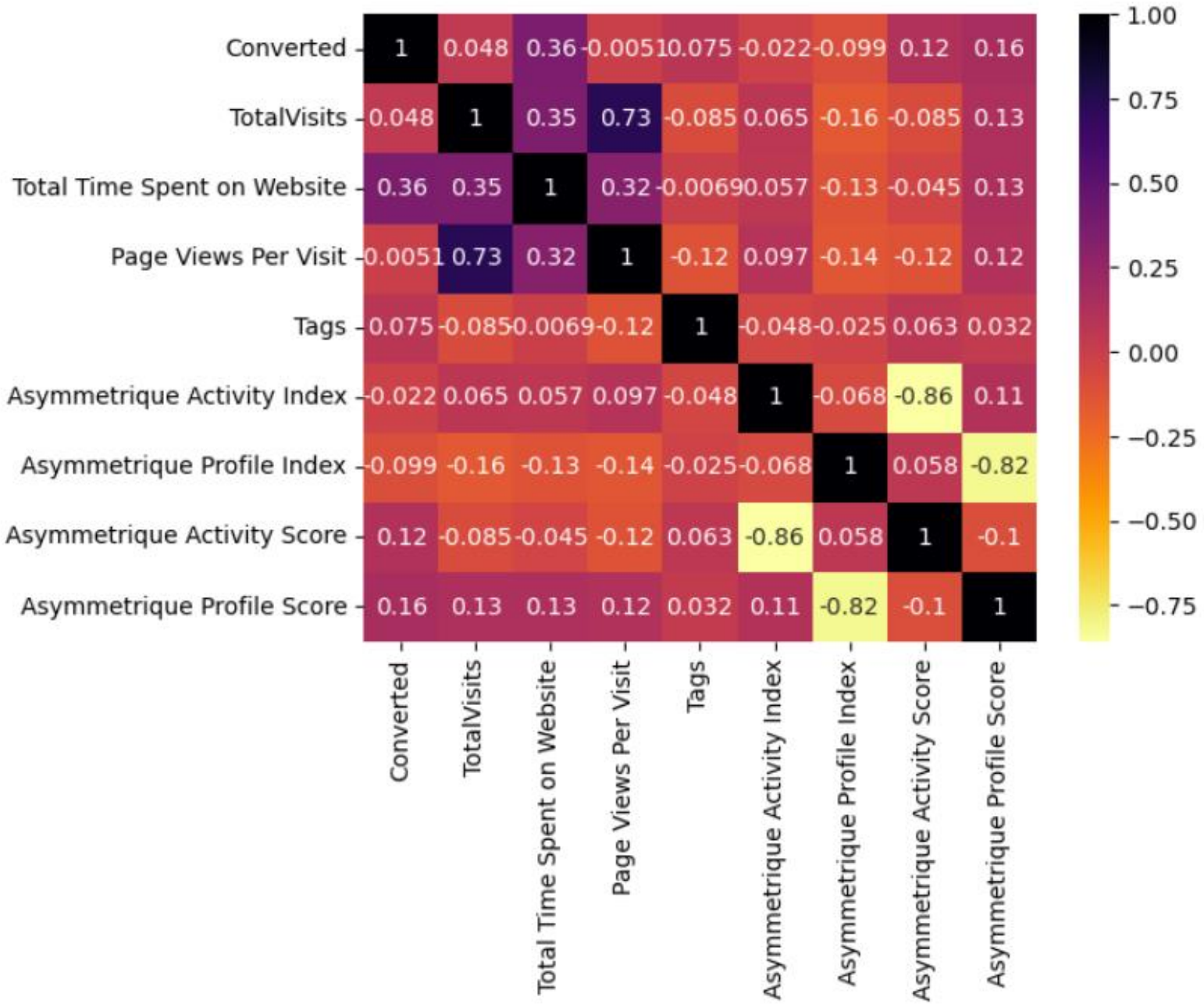


Bivariate Analysis- Lead Conversion vs Country and Lead Profile





Multivariate Analysis-
Leads Demographic Distribution



Variable Modification- Data Preparation for Model Training

- Removing noncontributing variables
    - Lead Number
    - Prospect ID
    - A free copy of Mastering The Interview
    - I agree to pay the amount through cheque’,
- Dummy Variable Creation
    - Actual Attributes in the original dataset : 37
    - Attributes after dataset Modification : 81

Train- Test Split

- Train Dataset size- 70%
- Test Dataset size- 30%
- Ramdon State- 100

Feature Selection- Model Training

- RFE to find the top 30 most contributing features : Used 30 to get the top 40% features, post RFE a backward approach is used to remove not required features.
- Dummy Variable Creation
    - Actual Attributes in the original dataset : 37
    - Attributes after dataset Modification : 81

Model Training- Logistics Regression using Statsmodel Library

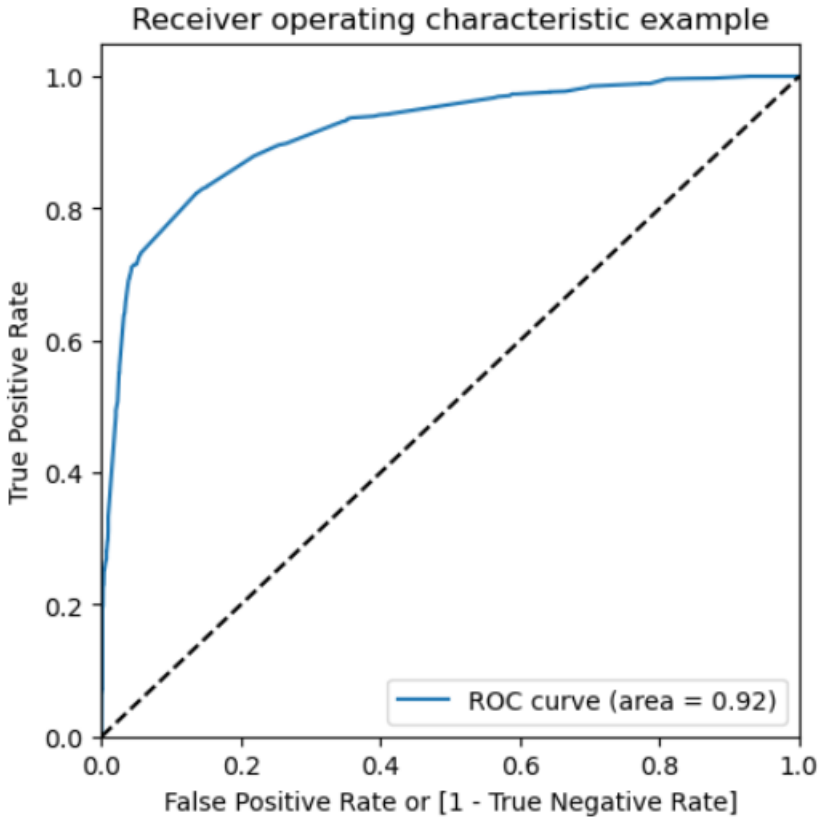
- With a total of 15 iterations, the final model is selected using backward approach on selecting the most optimum features using two methodologies
    - p-value
    - VIF
- In the final model- 15 features are selected with minimum multicollinearity and maximum contribution to the dependent variable.
- Then prediction are made from the achieved variables using Log Odds method from the GLM model.
- The predictions are between 0 and 1

Cut-Off Selection

- Initially a cut off 0.5 is selected and the following parameters are observed
    - Accuracy : 86.0%
    - Sensitivity : 72.6 %
    - Specificity : 94.6%
    - False Positive Rate : 05.3%
    - Positive Predicted Rate : 89.3%
    - Negative Predicted Value : 84.9%

ROC Curve to find Sensitivity and Specificity Trade-Off

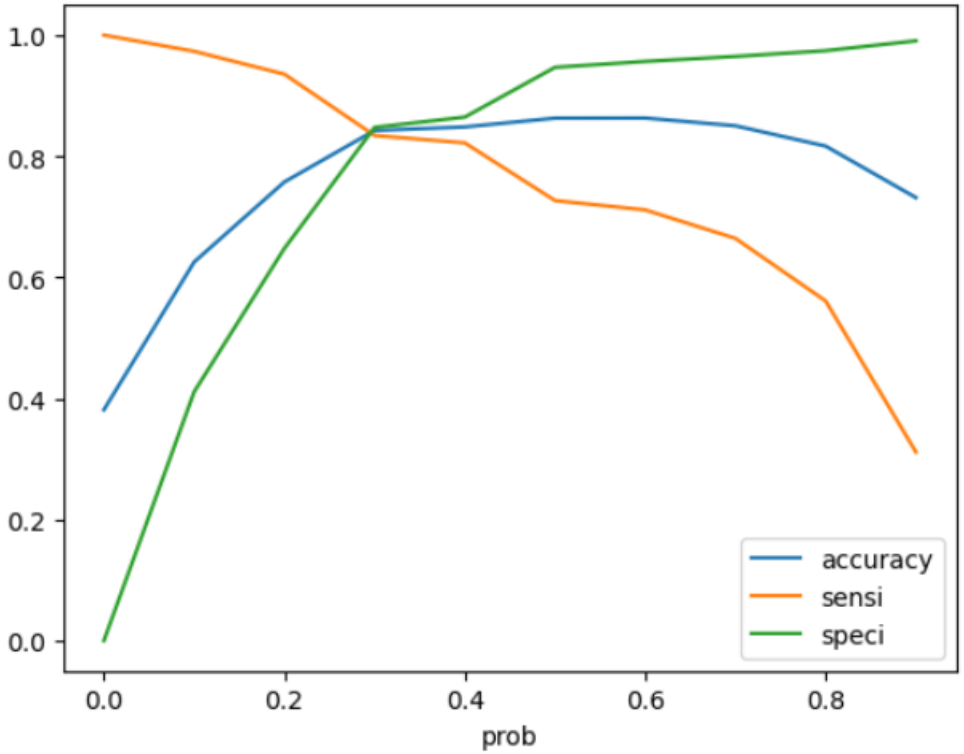
- In the achieved results, the sensitivity of the model is low and the specificity is higher, however the required is higher sensitivity and lower specificity
- The agenda of the model is to find the potential students who may opt for the course, so we do not want to miss even a single student who may opt in.



Cut-Off Selection

- Accuracy, sensitivity, and specificity are calculated at various cut offs and the following graph is achieved.
- Through graph a cut-off of 0.28 is selected with the following observations

- Accuracy	:	82.0%
- Sensitivity	:	87.9 %
- Specificity	:	78.1%
- False Positive Rate	:	21.8%
- Positive Predicted Rate	:	71.3%
- Negative Predicted Value	:	91.3%



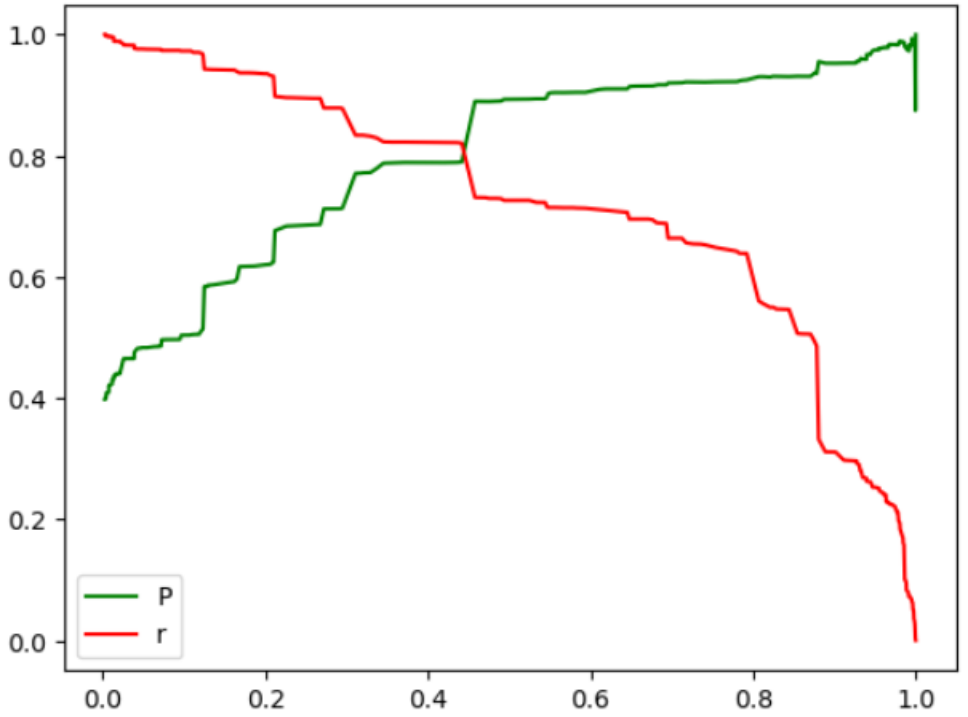
Sensitivity of the trained model is now higher and can be used to predict values on test dataset.

Precision vs Recall Trade Off

- The achieved values of Precision and Recall are following
    - Precision : 71.3%
    - Recall : 87.9 %

- Higher recall value was needed and there is already higher recall.

The cut-off suggested from Precision Recall graph is 0.42. Using this cut-off To predict test dataset values.



Predictions on Test Data Set

- Test Dataset is prepared by only keeping the required features achieved from model training and feature elimination iteration.
- Predictions are mode on Test data set and the following values are observed

- Cut-off	:	0.42
- Accuracy	:	84.0%
- Sensitivity	:	82.6 %
- Specificity	:	85.3%
- False Positive Rate	:	14.6%
- Positive Predicted Rate	:	78.7%
- Negative Predicted Value	:	88.2%

The results look good and model is set to be delivered.