LEAD SCORE CASE STUDY

PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- In order to increase the lead conversion rate, the company first should 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 Objective:--

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

Solution Approach

Data Cleaning and Manipulation

1. Check and handle duplicate data.
2. Check and handle NA values and missing values.
3. Drop columns, if it contains a large number of missing values and are not useful for the analysis.
4. Imputation of the values, if necessary.
5. Check and handle outliers in data.
6. To deal with columns having outliers will create bins for them.

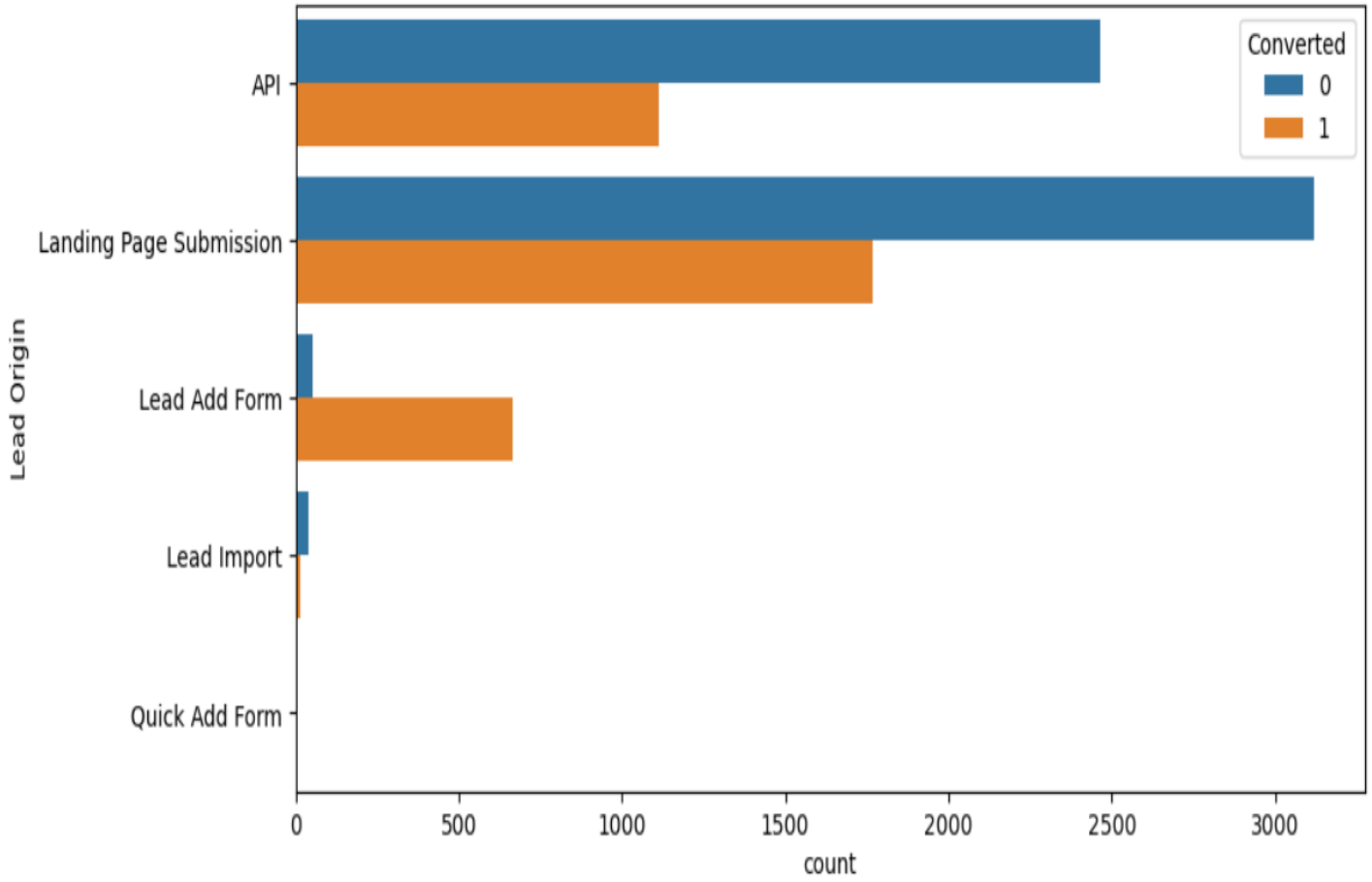
EDA

1. Univariate data analysis: value count, distribution of variables, etc.
2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.
3. Feature Scaling & Dummy variables and encoding of the data.
4. Classification technique: logistic regression is used for model making and prediction.
5. Validation of the model.
6. Model presentation.
7. Conclusions and recommendations.

Data Manipulation:

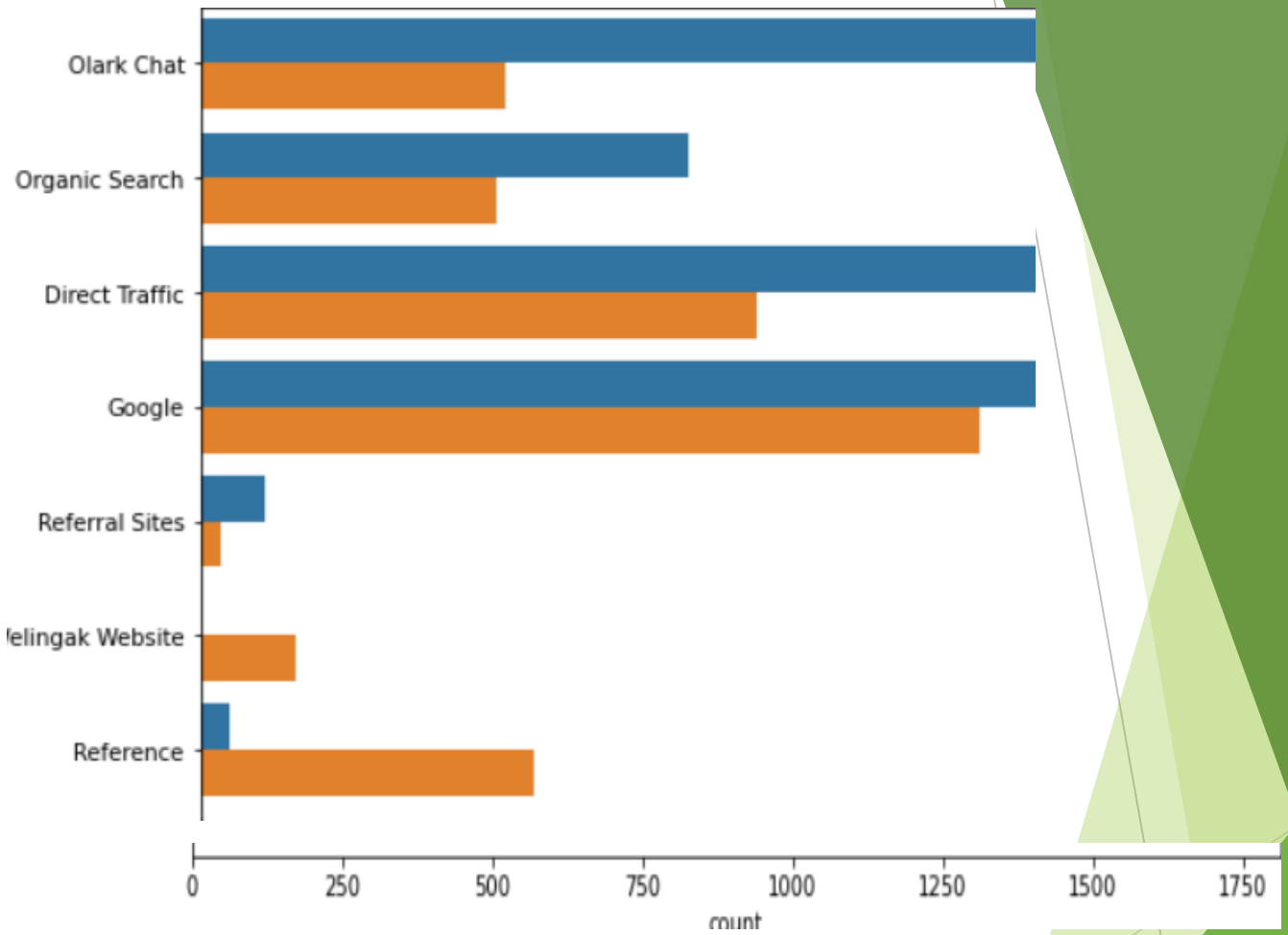
- Checking the value counts of each column and dropping unnecessary columns which are Prospect ID, Lead Number, Country, Receive More Updates About Our Courses, Update me on Supply Chain Content, Get updates on DM Content, I agree to pay the amount through cheque -column only has 'No' doesn't makes sense to keep it, Magazine
- Transforming columns to the yes/no category Do Not Email, Do Not Call, Search, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, a free copy of Mastering The Interview
- Remove columns having more than 40% null values
- Imputing missing values as per column data available
- Data is skewed in Lead source , we are going to replace these labels (Facebook, bing, Click2call, Live Chat,Press_Release, Social Media, testone, WeLearn, blog, Pay per Click Ads, welearnblog_Home, youtubechannel, NC_EDM) in one label as 'Others'.

Exploratory Data Analysis (EDA)

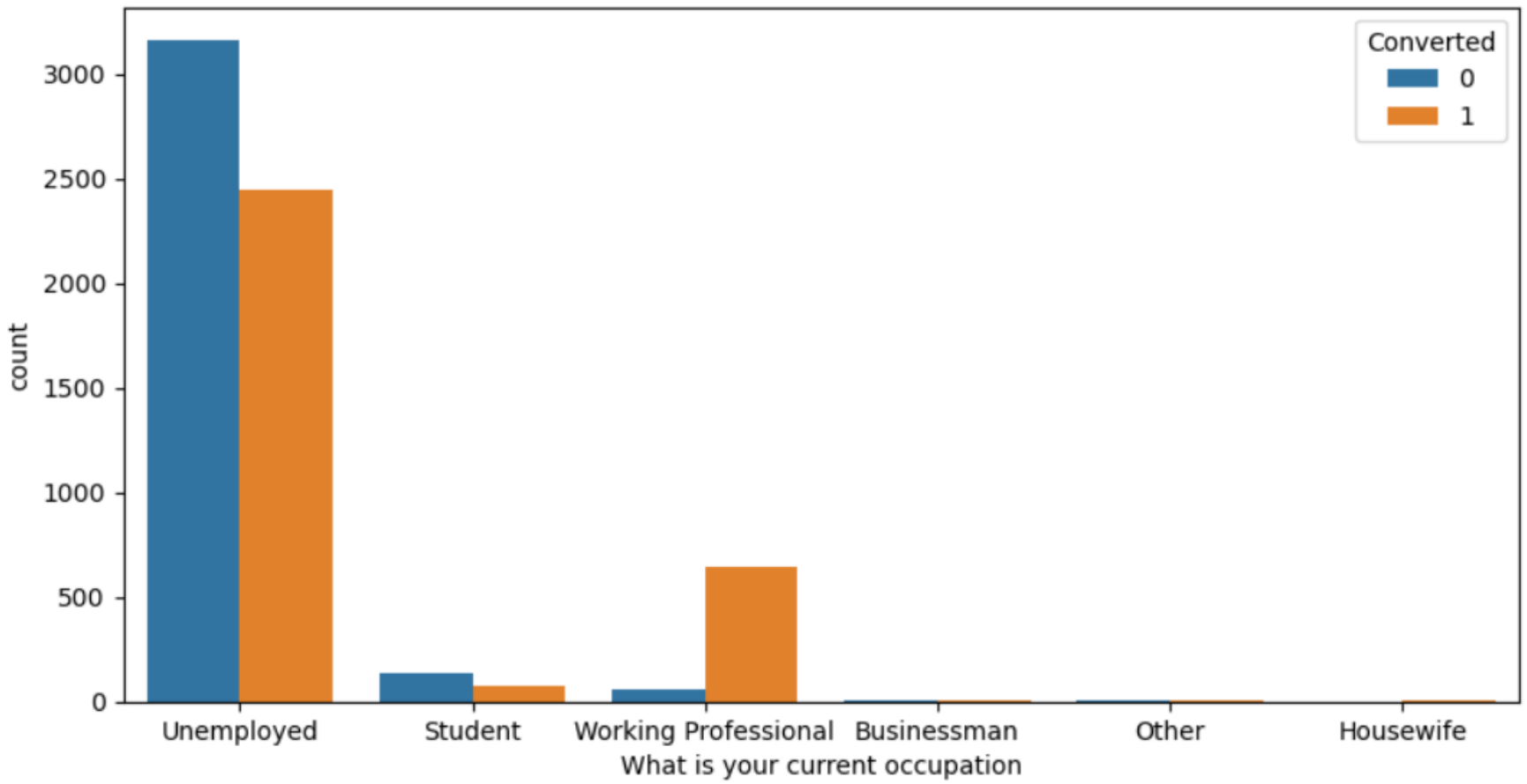


- Maximum lead conversion happened from Landing Page Submission.

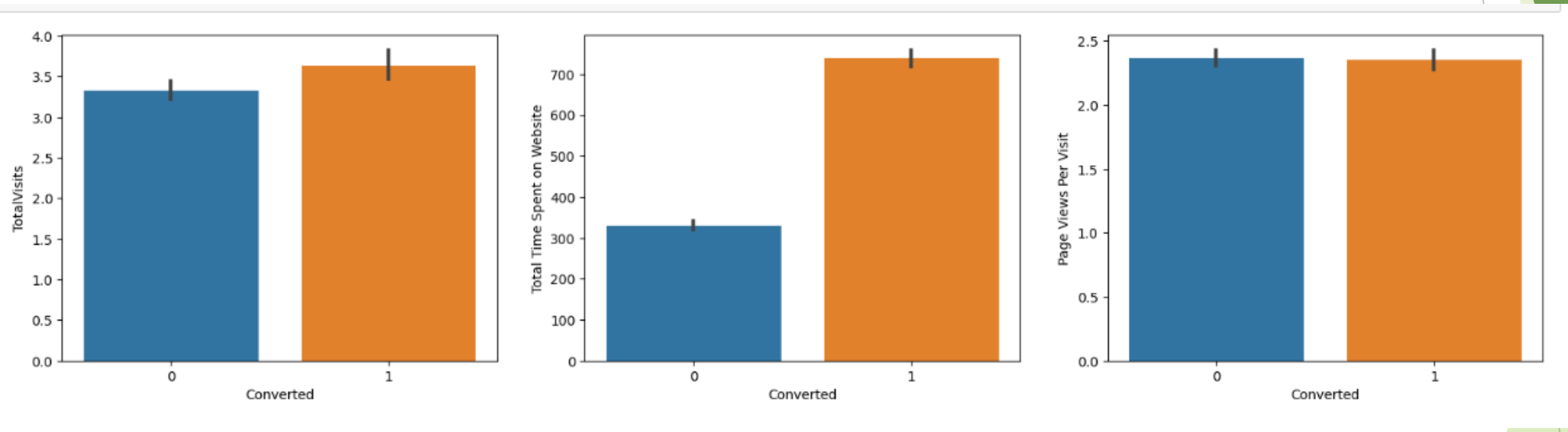
Major lead conversion in the lead source is from 'Google'



- Major lead conversion is from the Unemployed Group



- Major lead conversion from TotalVisits, Total Time Spent on Website, Page Views Per Visit



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.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p-value is greater than 0.05 and vi value is greater than 5
- Predictions on test data set
- **Test Data Set metrics**:

  Sensitivity: 84.57

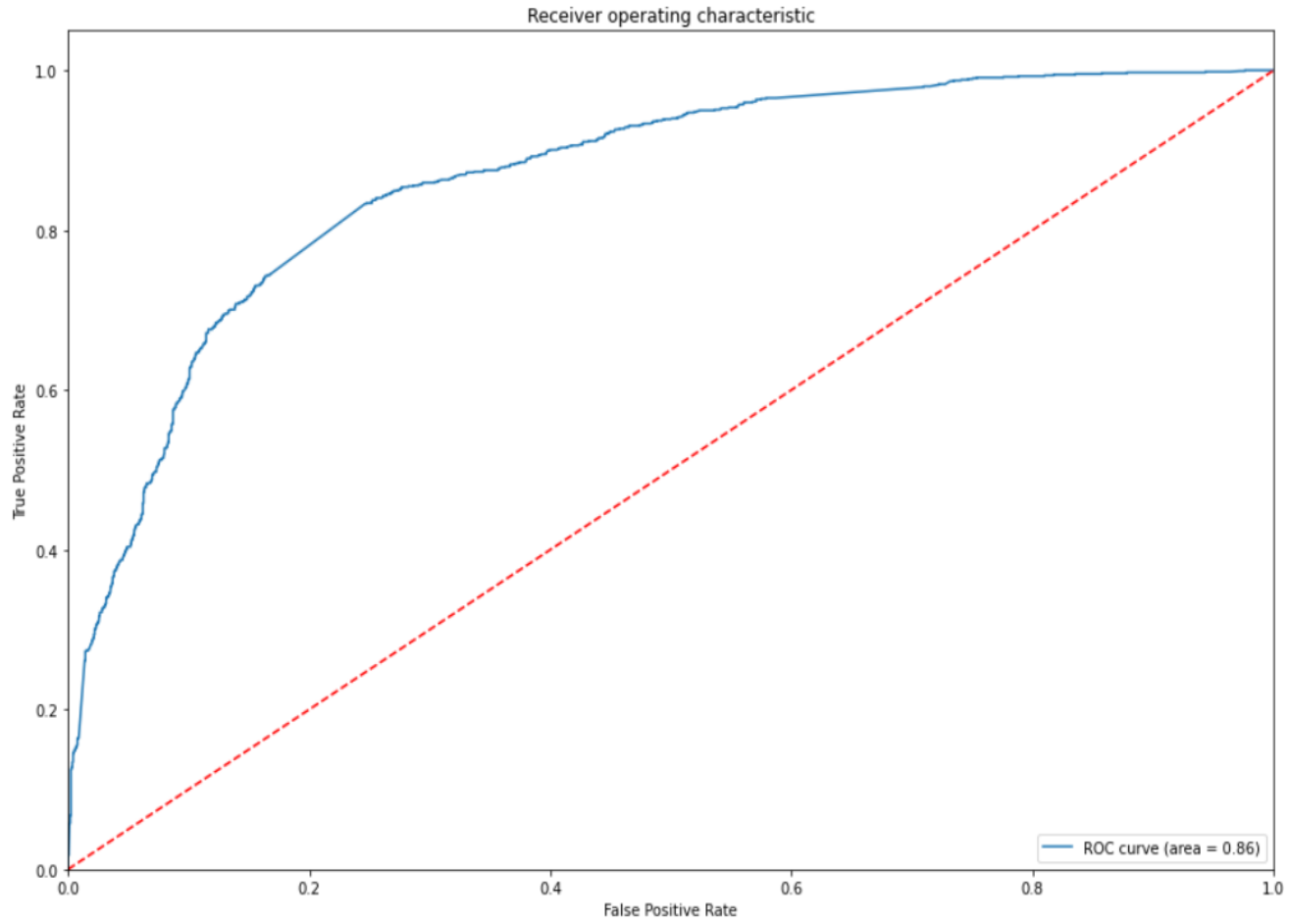
  Specificity: 73.23

  Precision: 67.35

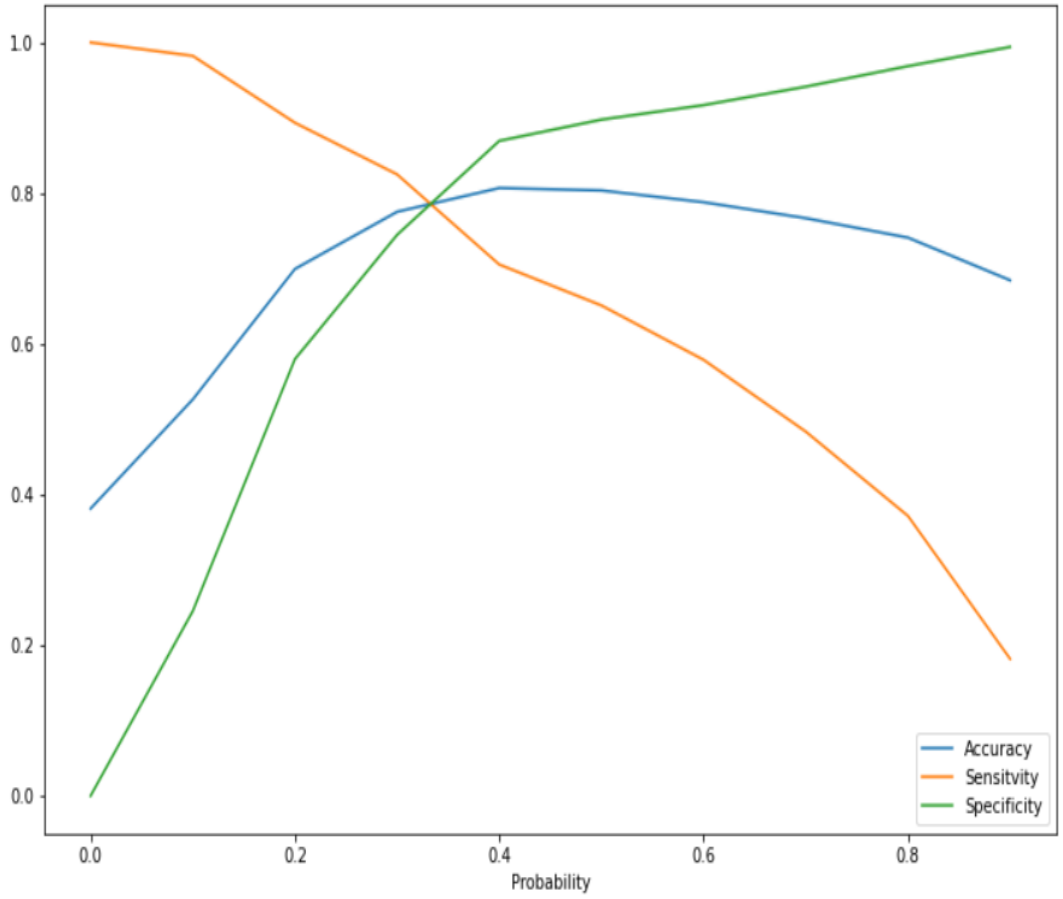
  Recall: 84.57

  Accuracy: 77.71

ROC Curve


Receiver operating characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.86)

Optimal Cut-off



- Optimal cut-off probability is that Probability where we get balanced sensitivity and specificity.
- From the graph it is visible that the optimal cut off is at 0.3.

Prediction on Test Data Set

- Before predicting on the test set, we need to standardize the test set and need to have exact same columns present in our final train dataset.
- After doing the above step, we started predicting the test set, and the new prediction values were saved in a new data frame.
- After this we did model evaluation i.e. finding the accuracy, precision, and recall.
- The sensitivity score we found is 84.57%, Specificity is 73.23%, precision 67.35%, recall 84.57% and accuracy is 77.71% approximately.
- This shows that our test prediction is having accuracy, precision, and recall scores in an acceptable range.
- This also shows that our model is stable with good accuracy and recall/sensitivity.
- Lead score is created on test dataset to identify hot leads – high the lead score higher the chance of conversion, low the lead score lower the chance of getting converted.

Conclusion

- The Sensitivity and Specificity, Accuracy, Precision and Recall score we got from test set are almost accurate.
- We have high recall score than precision score which is a sign of good model.
- In business terms, this model has an ability to adjust with the company's requirements in coming future.
- This concludes that the model is in stable state.
- **Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :**
    1. Lead Origin_Lead Add Form
    2. Total Time Spent on Website
    3. What is your current occupation_Working Professional