# Lead Scoring Assignment

By Manisha & Niranjan

# *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 markets its courses on several websites and search engines like Google.

Once these people land on the website, 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 Objective**

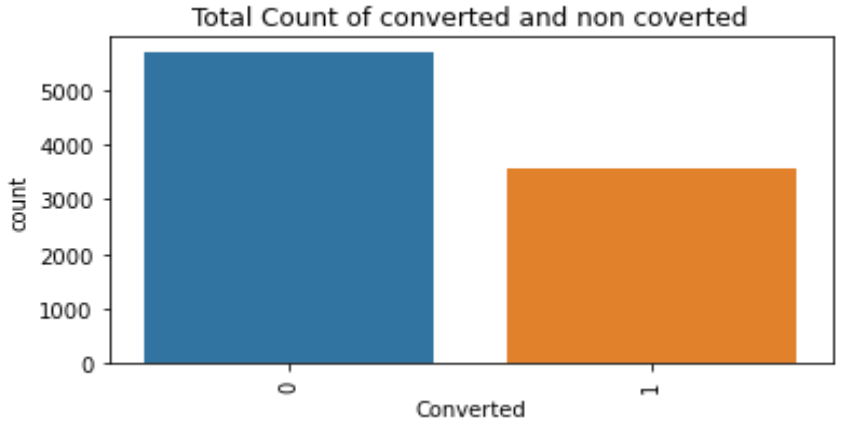
X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.

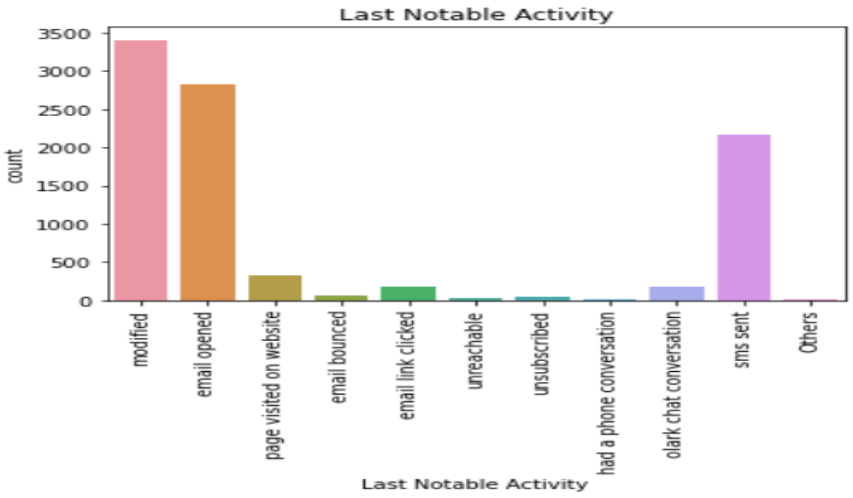
To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

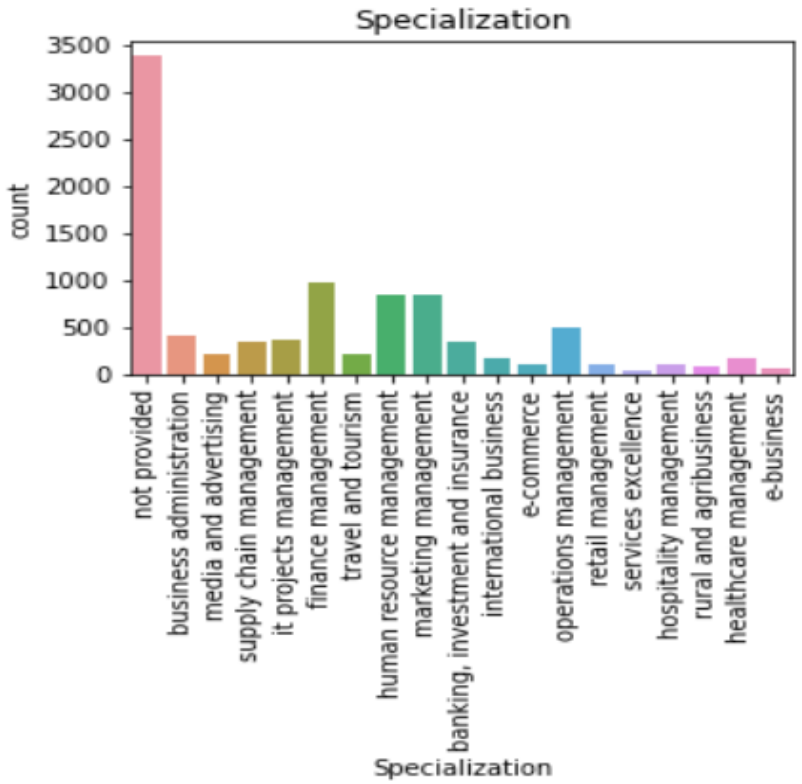
# *Approach*

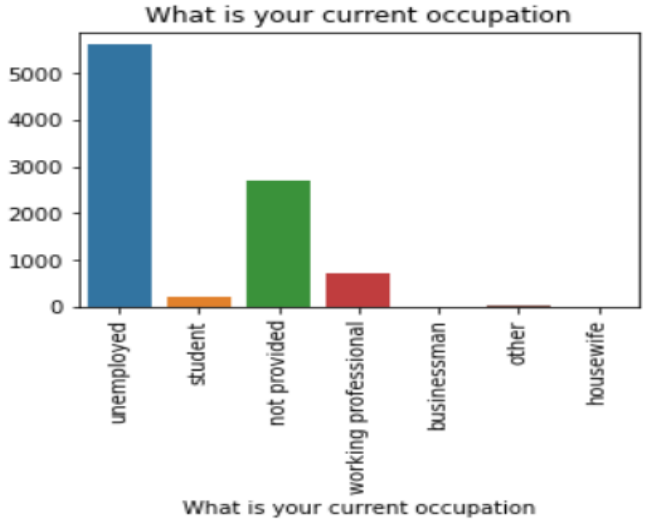
- Import/Load the data.
- Check the structure/metadata of the data.
- Clean and prepare the data for further analysis.
- Scaling features.
- Prepare the data for model building.
- Build logistic regression model.
- Assign the lead score of each leads.
- Test the model on train set.
- Evaluate model by different measures.
- Test the model on test set.
- Measure the accuracy of the model.

# EDA- Leads data analysis



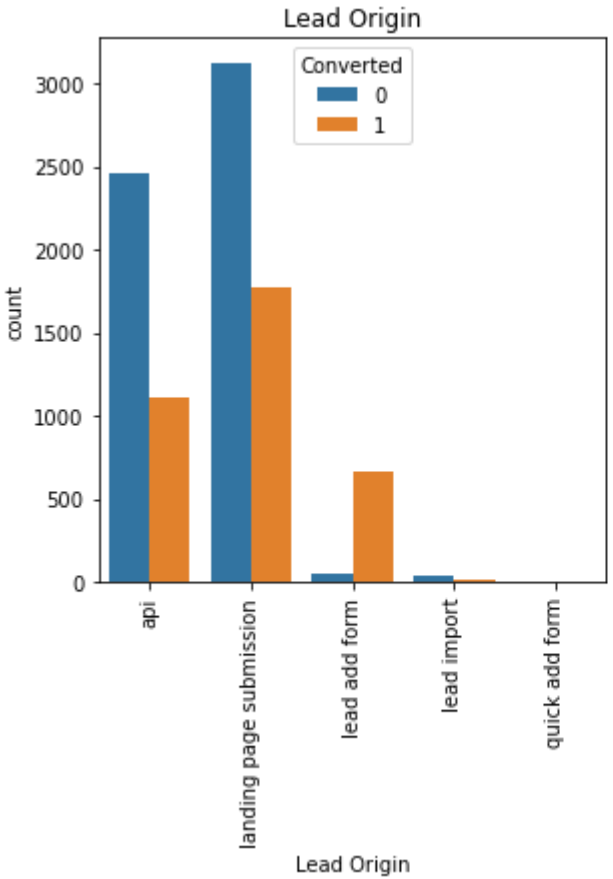



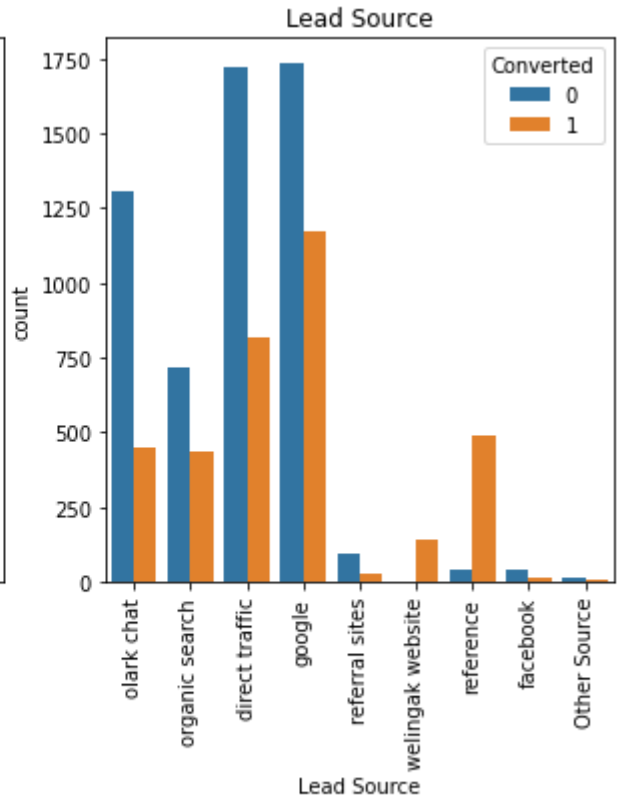



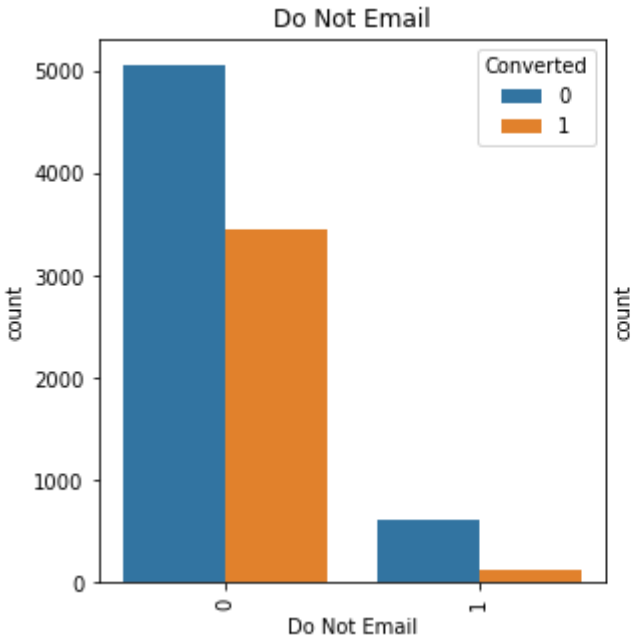

Comments

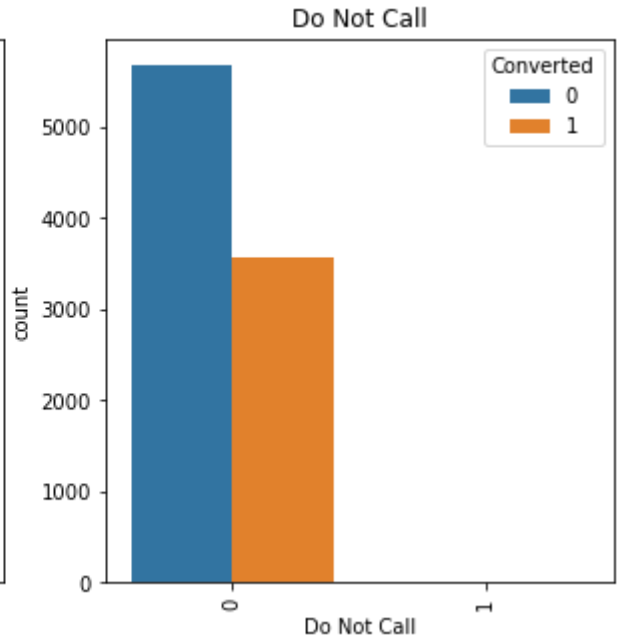
- 38% customer got converted as per leads data
- Highest leads have came from Finance management, unemployed and who's last notable activity was modified.

# EDA- Leads data analysis



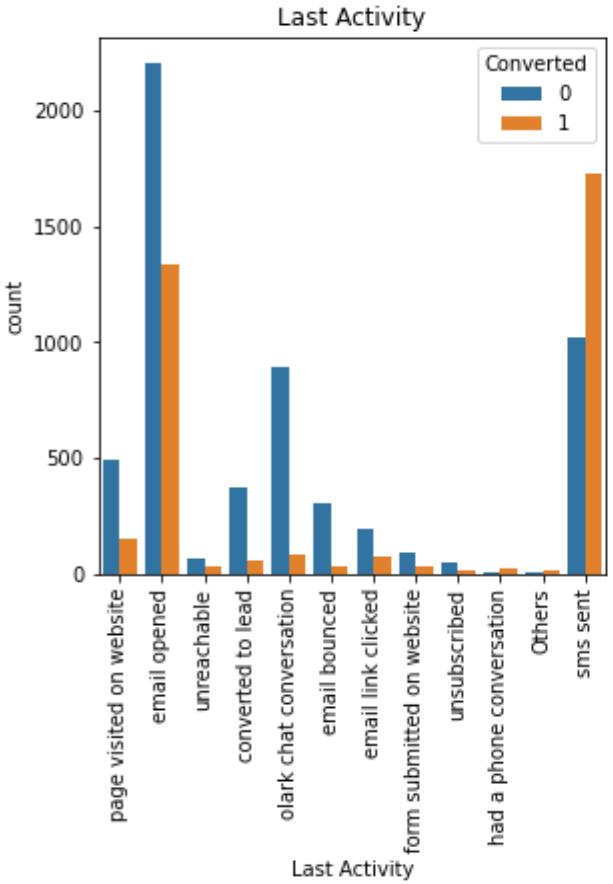



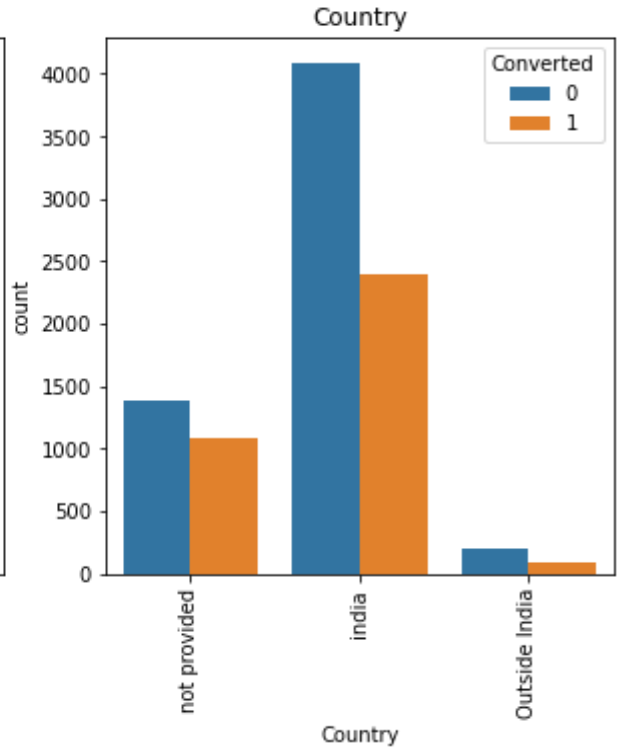



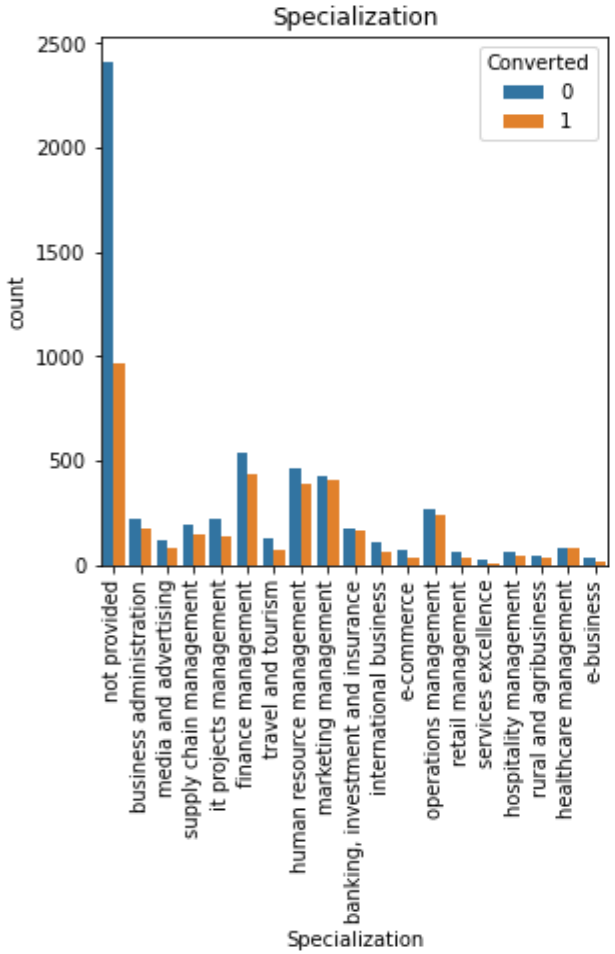

Comments

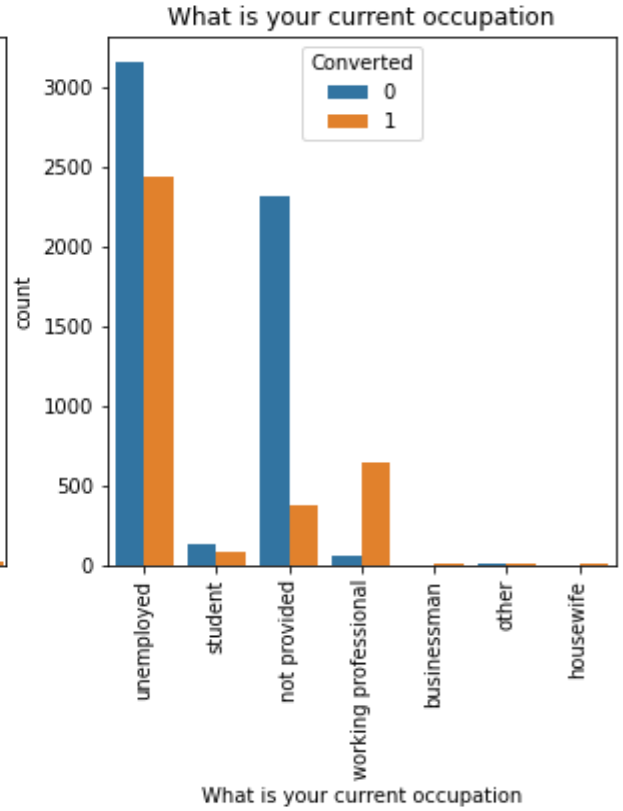
- Lead Origin- Landing page submission has high conversion compared to others.
- Lead Source- Google searches has high conversion compared to others sources, it has high conversion rate.
- Do not email- Do not email has high conversion compared to others sources.
- Do not call- most leads prefer not to inform through call

# EDA- Leads data analysis



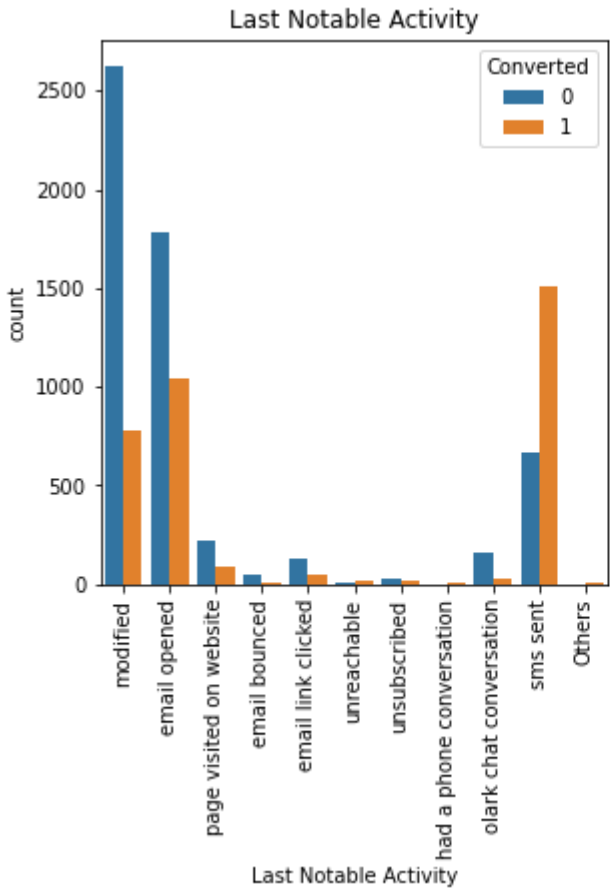



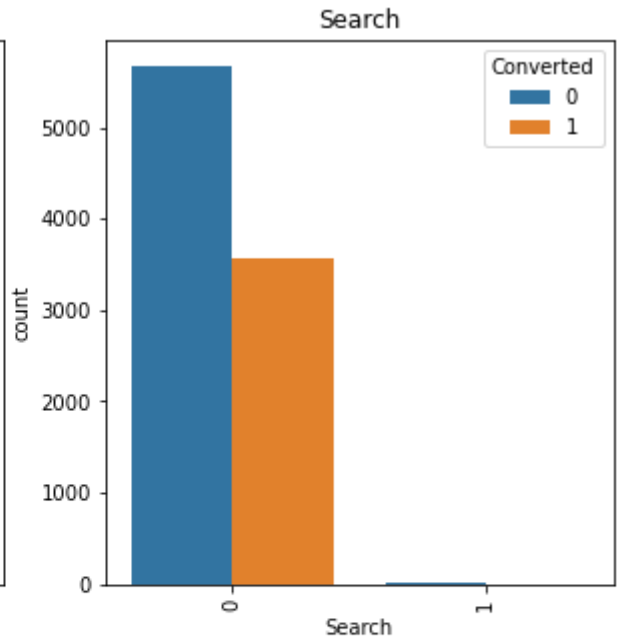



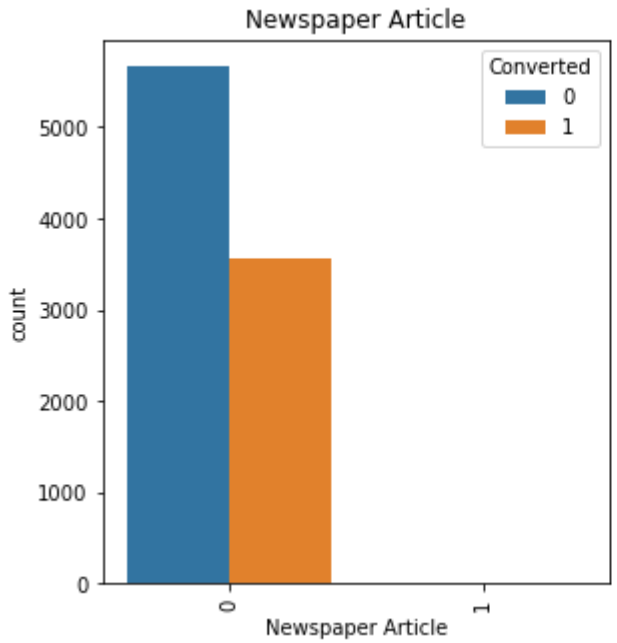

Comments

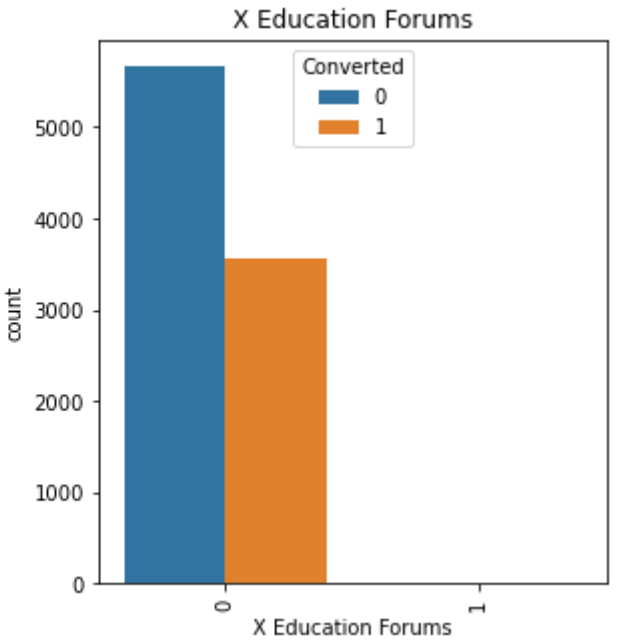
- Last Activity- SMS has showing highest confirmation leads, email also has high conversion.
- Country- India has high conversion compared to others country, it has high conversion rate.
- Specialization- Finance, human resource and marketing management has high conversion compared to others specialization.
- What is your current occupation- unemployed has high conversion compared to others occupation, it has high conversion rate.

# EDA- Leads data analysis



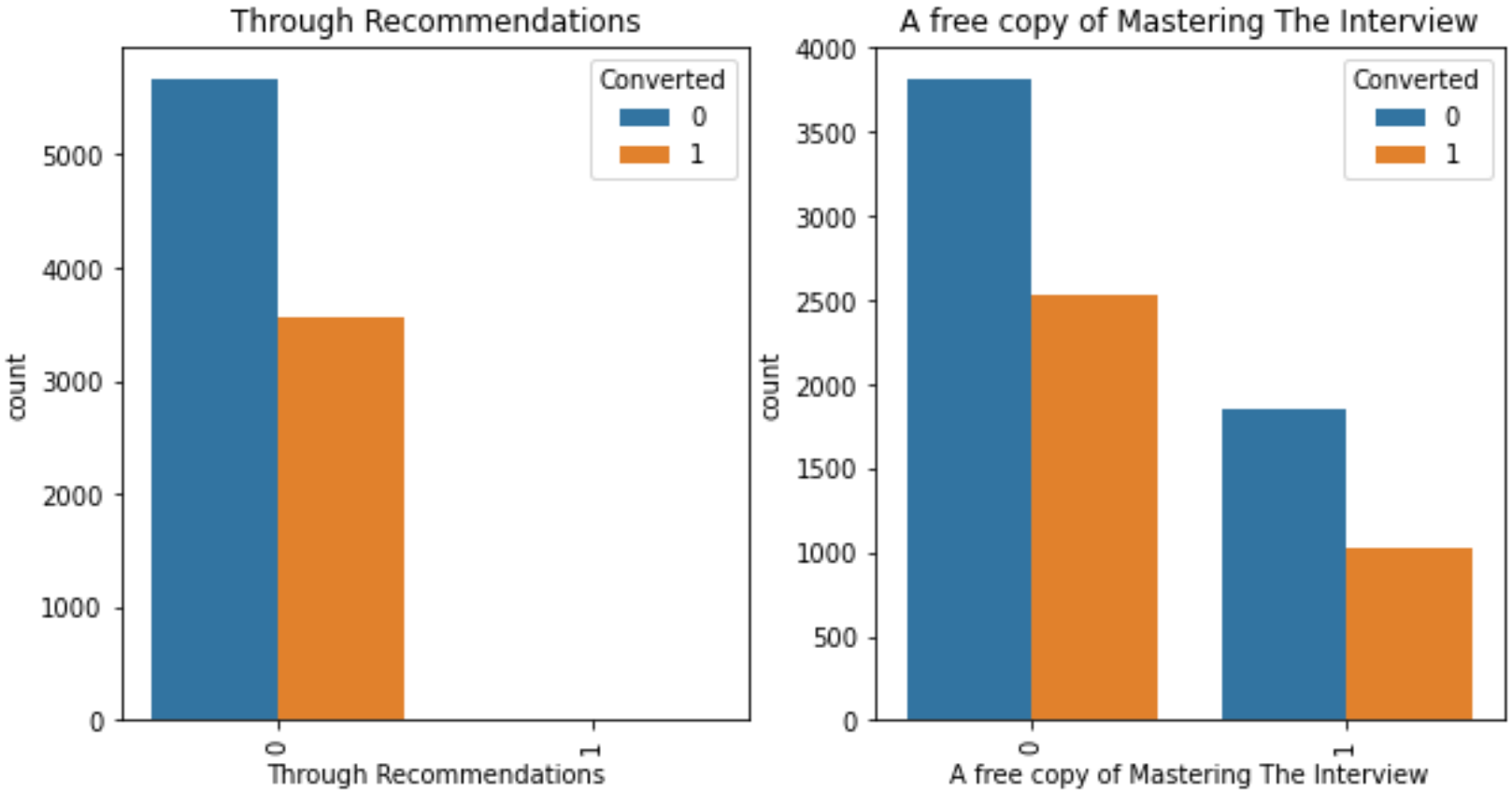




Comments

- Last Notable Activity – SMS sent has showing highest confirmation leads, email opened also has high conversion leads.
- Search- In this graph shows Searches are not good source of leads.
- Newspaper Article- In this graph shows Newspaper Article are not good source of leads.
- X Education forums- In this graph shows X Education Forums are not good source of leads.

# *EDA- Leads data analysis*



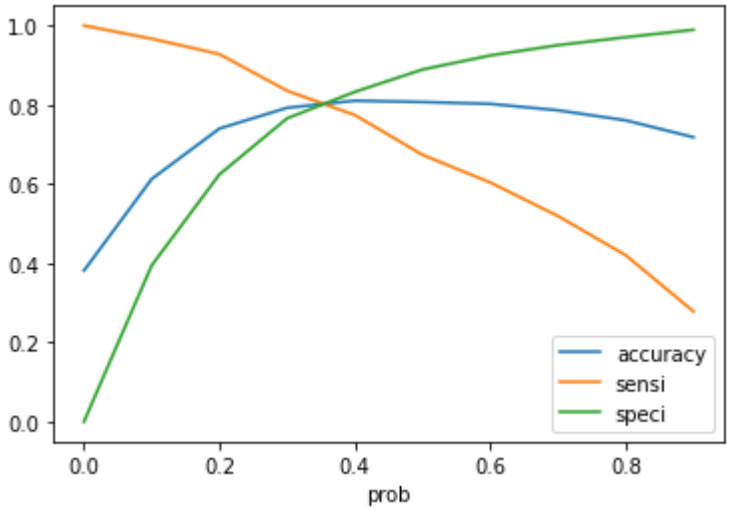

Comments

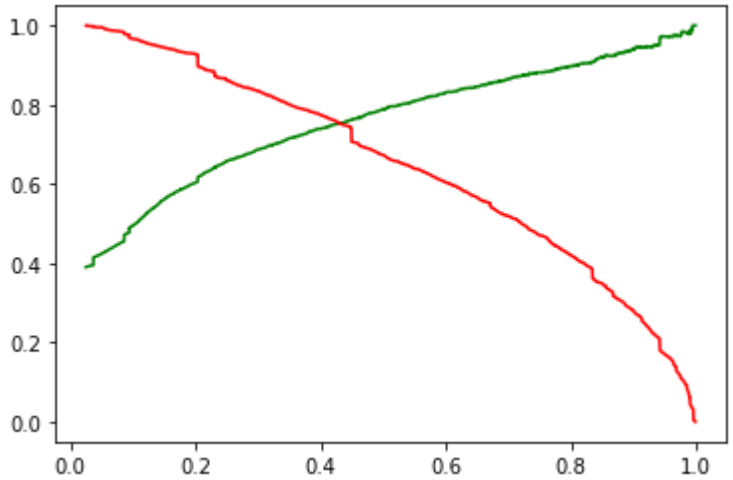
- Through Recommendations- In this graph shows Newspaper Article are not good source of leads.
- A free copy of Mastering The Interview- In this graph shows Newspaper Article are not good source of leads.

# *Model Building Approach*

- Splitting data into train and test sets.
- Scale variable in train set.
- Build the first model.
- Use RFE to eliminate less relevant variables.
- Build the next model.
- Eliminate variables based on high p- values
- Check VIF value for all the existing columns.
- Predict using train set
- Evaluate accuracy and other metrics.
- Predict using test set.
- Precision and recall analysis on test  predictions

# *Model Evaluation (Train set)*

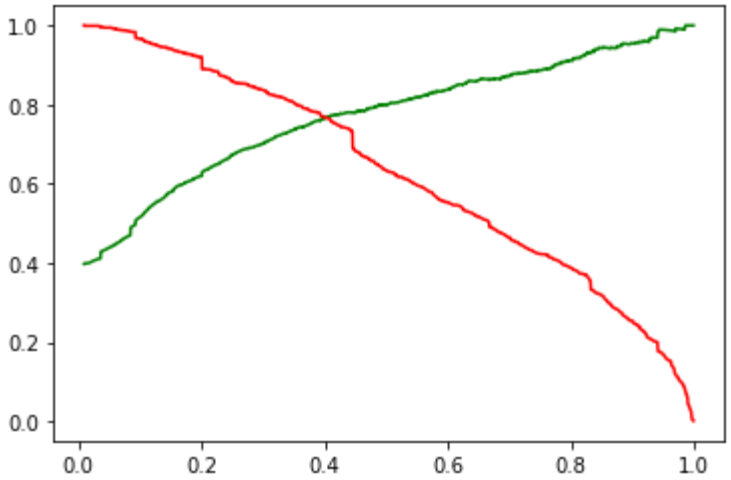



Comments

- Accuracy : 79%
- Sensitivity : 84%
- Specificity : 77%
- Precision: 68%
- Recall: 83%

# *Model Evaluation (Test set)*




Comments

- Accuracy : 80%
- Sensitivity : 83%
- Specificity : 77%
- Precision: 70%
- Recall: 83%

# *Summary*

## EDA:

- People spending higher than average time are the promising leads, so targeting them and approach them can be helpful to conversions.
- SMS can have a high impact on lead conversion.
- Landing page submission can help find out more leads.
- Finance, Marketing and Human resource management has high conversion rates, people from this specialization can be a hot leads.
- An alert massages or information has seen to have high conversion rates.

## Logistic regression model:

- The model shows high close to 79% accuracy
- The threshold has been selected from Accuracy, Sensitivity, Specificity measures and Precision, Recall curves
- The model shows 83% Sensitivity and 77% Specificity
- The model has find correct promising leads and leads that have chances of getting converted
- Overall the model proves to be accurate