Lead Scoring – Case Study

ASHU VERMA
JUNAID SHAIKH
USHA ATHIMAKULA

Leads – Case Study

Background:

X Education sells online courses to industry professionals. Leads are acquired through marketing via different Search Engines, Past referrals

Problem Statement:

There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. The typical lead conversion rate at X education is around 30% which is very less

Goal:

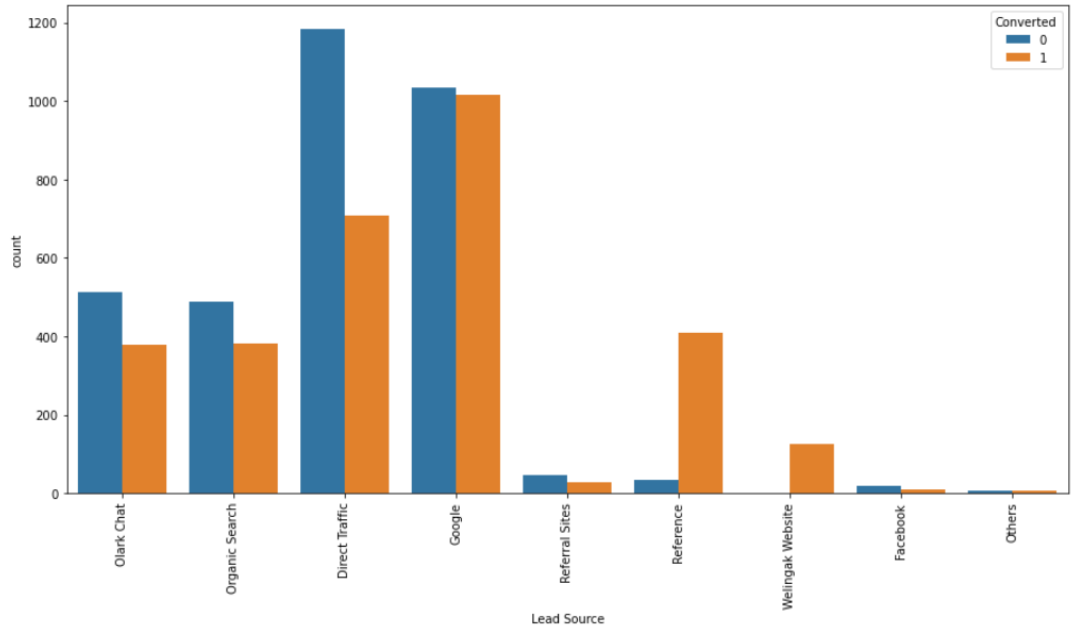
To identify the leads that are most likely to convert into paying customers.

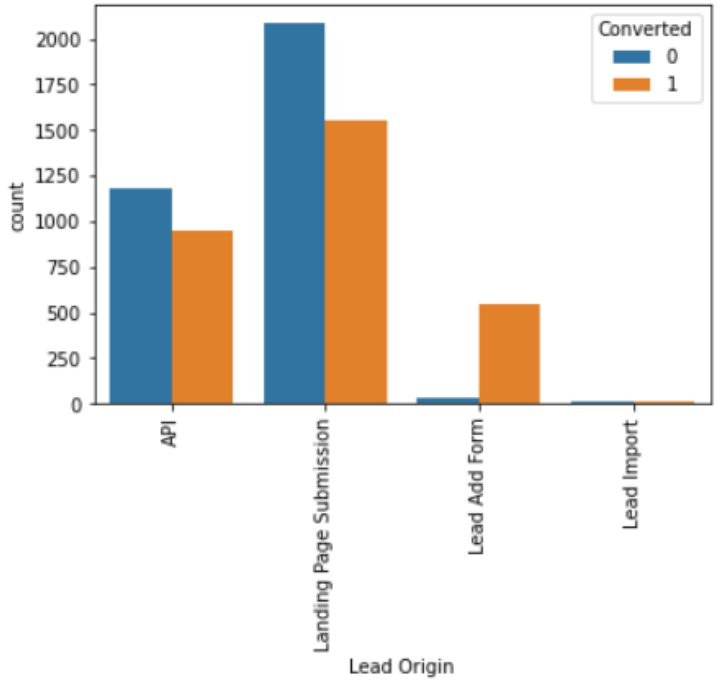
Approach

- Analyze the data provided
- Data Cleaning
  - Check for missing values, incorrect data formats
  - For missing values: impute the missing values or drop the rows/columns
  - Set a threshold for missing data in columns, Ex: Drop a column if there are more than 30% missing values/Nulls
- Exploratory Data Analysis (EDA)
  - For Categorical Columns, missing values can be replaced by mode
  - For Numerical Columns, missing values can be replaced by mean/median
  - Incorrect data types or data formats to be corrected
- Observe the outliers, analyze and remove them if necessary
- Analyze the Target columns of each Dataset using necessary visualizations
  - Converted is the target Column
    - Understand the effect of each categorical and numerical variable on the TARGET Column by performing exploratory data analysis (EDA)
    - Create dummy columns for all the categorical columns and scale the numerical columns
    - Build a logical regression model, train and test the model using hold-out method and finally evaluate the metrics

Insights

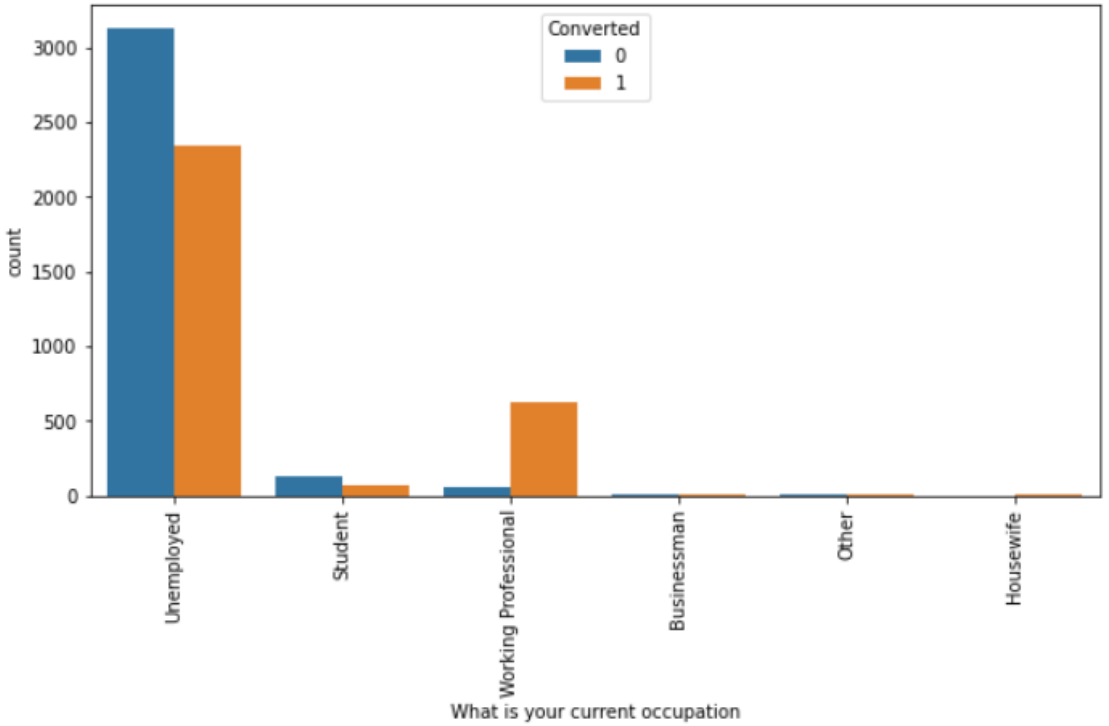
- It can be observed that Most Lead Origins are through Landing Page Submission followed by API. Conversions are also the most for the leads coming in through Landing Page Submission
- Most of the Lead sources are through Direct Traffic and Google.





Insights(Cont..)

- It can be observed that Most Leads are from Unemployed followed by working Professionals. However higher percentage of conversion is seen in working professionals



Training Results

Accuracy : 78.8

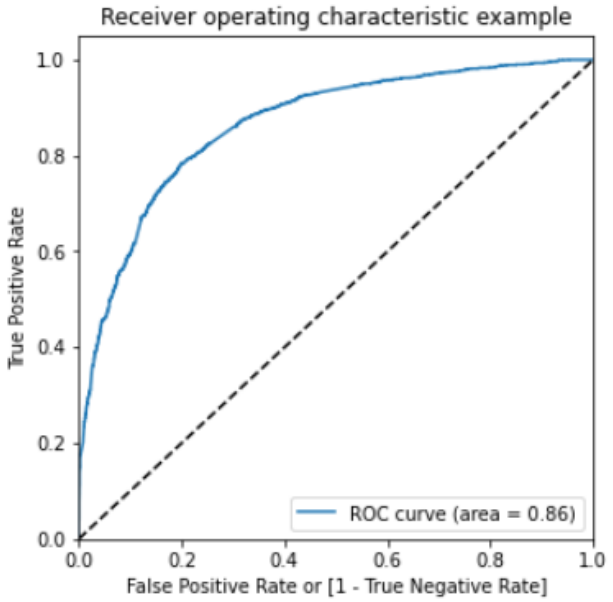
Sensitivity: 73.7

Specificity: 83.5

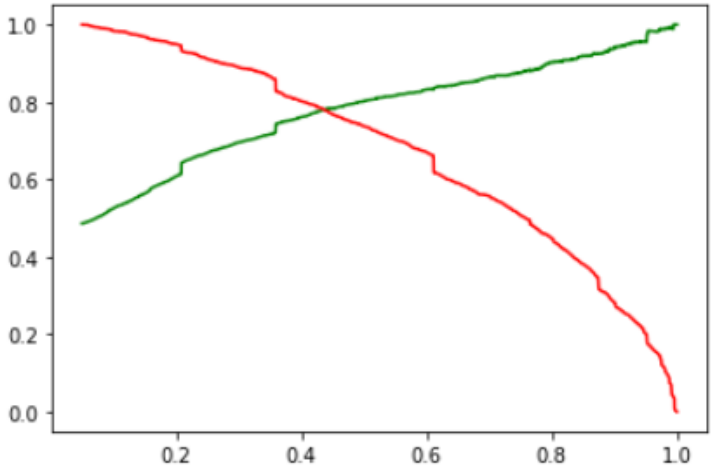
Precision: 80.1

Recall: 73.7

ROC Curve



Precision Recall Curve:



Test Results

Accuracy : 78.9

Sensitivity: 79.06

Specificity: 78.9

Precision: 78.3

Recall: 79.06

- We built a model with 78.9 % accuracy which is decent.
- We found that the three important features to be considered are: Lead Source, Total Time Spent on Website, Current occupation

Conclusion

- The model accuracy was around 79% where the top three variables influencing the leads conversion were:
  - Lead Source
  - Total Time Spent on Website
  - Current occupation

- In case if there are ample resources for callings (having interns), then over predictions can be made by increasing the false positives
  - With cutoff of around 0.2, False positives increased from 206 to 533

- On contrary, if there is limited resources for callings, then prediction power can be reduced by decreasing the false positives
  - With cutoff of around 0.8, False positives dropped from 206 to 47