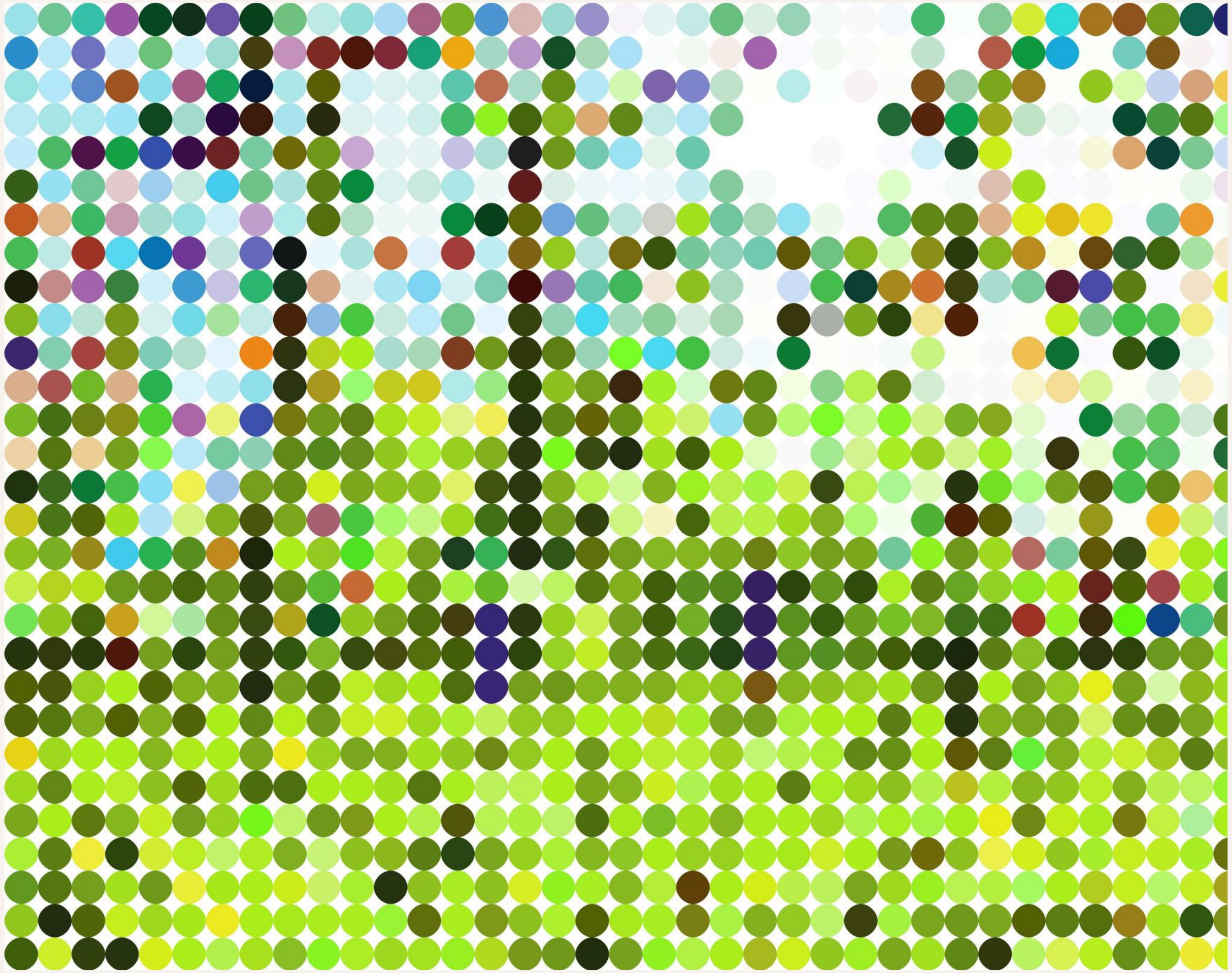
# Lead scoring case study

AORPITA CHOUDHURY

ARSALAN KHURSHEED

ARNAV SWAROOP

DS C52



# Business Understanding

- X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. The company also gets leads through past referrals.
- X Education wants to pursue the most promising leads that will turn into paying customers.
- 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%.
- Although X Education gets a lot of leads, its lead conversion rate is very poor.

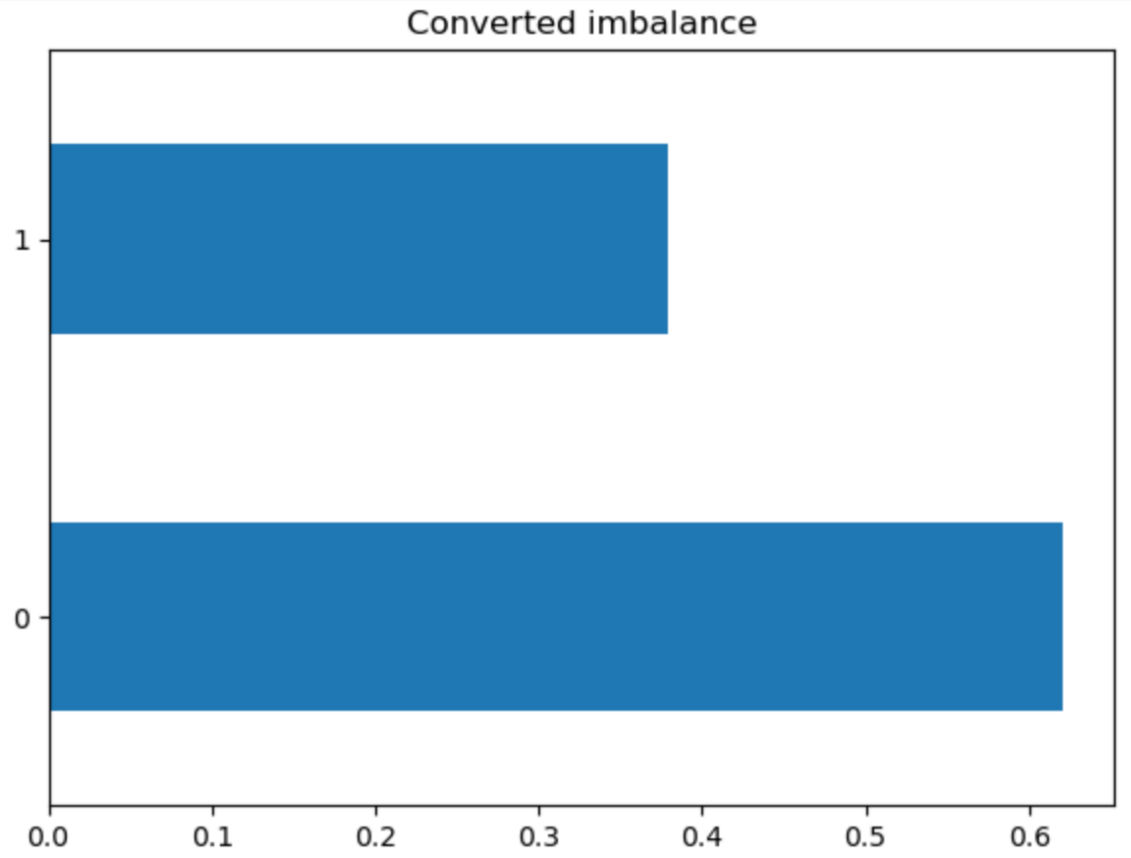
# Objectives

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- Target lead conversion rate to be 80% or more.

# Leads data analysis

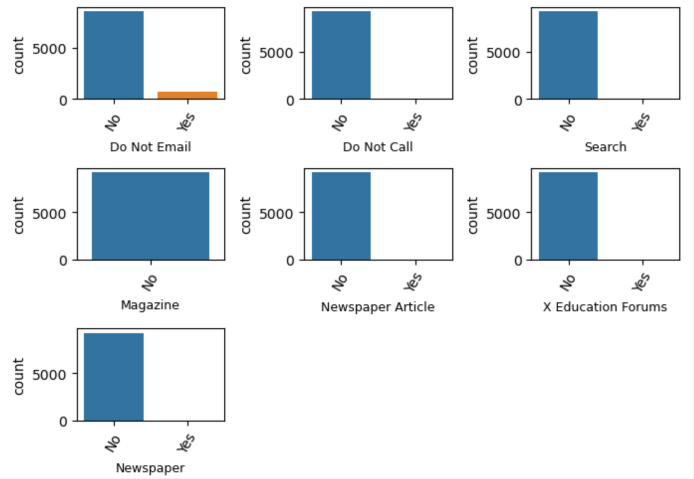
- Looking for missing values in all columns
  - Drop columns with more than 40% missing values as they will not add value to our analysis
- Based on understanding of data we should drop one of the 2 unique columns
- We may drop the miscellaneous columns which we feel will not add any value to our analysis
- After analysis of a few data points visually we see that there are many “Select” values which seem to be missing values. We can replace these with NaN
- We can now do outlier analysis on our dataframe
- We should now proceed with EDA
- Post this we can start to build our model

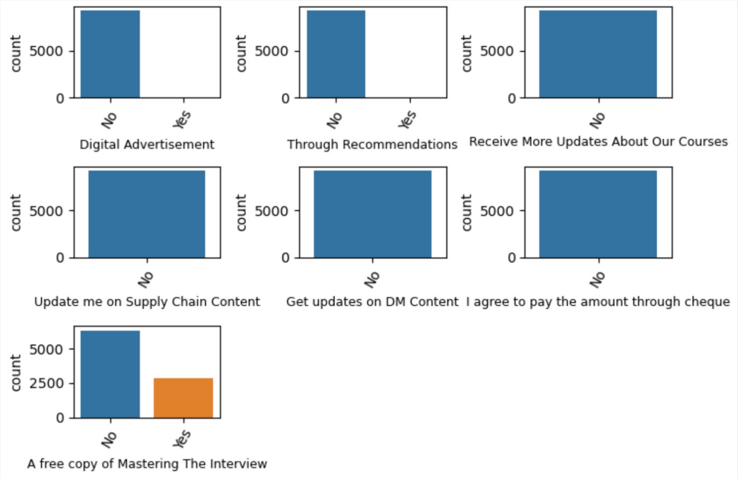
# Data imbalance check



- Approx. 38% of the data is converted into a lead
- This means that the dataset is not very imbalanced and thus we need not divide our dataset

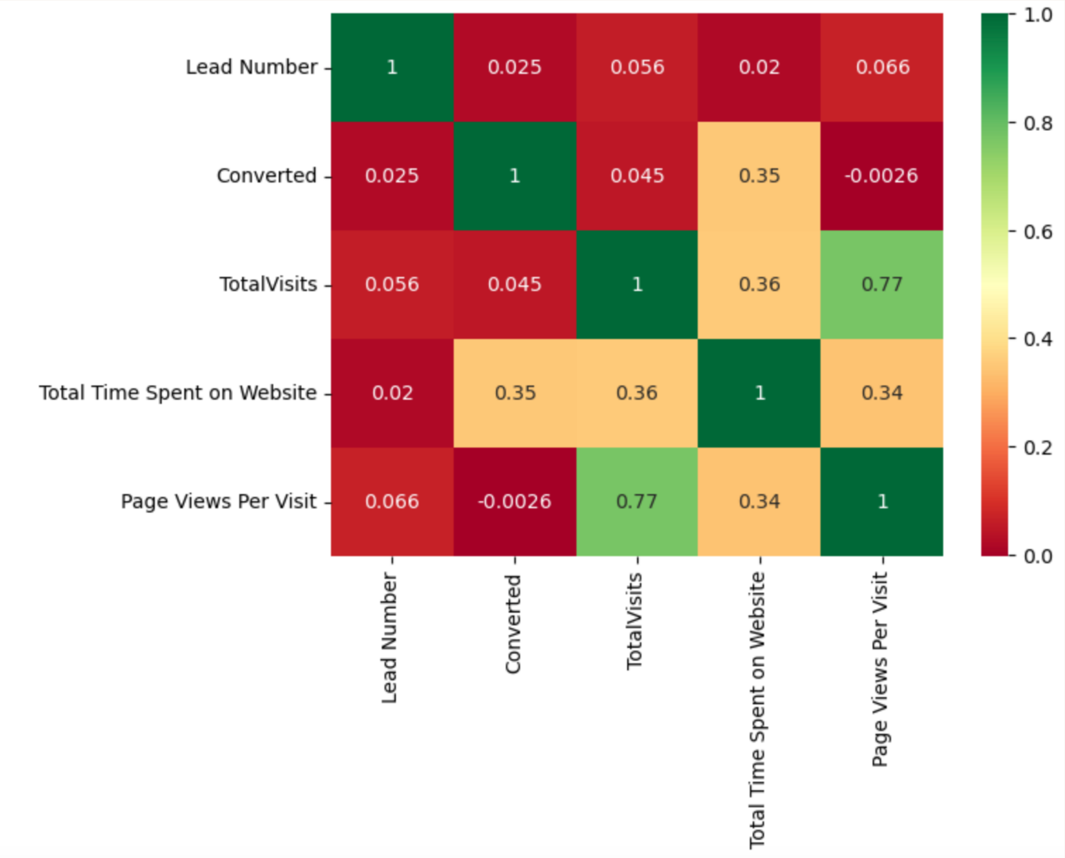
# Dropping miscellaneous columns





- We can drop the columns shown here since they mostly contain No as the answer and will not add any value to our model
- We can keep the columns ‘Do Not Email’ & ‘A free copy of Mastering The Interview’

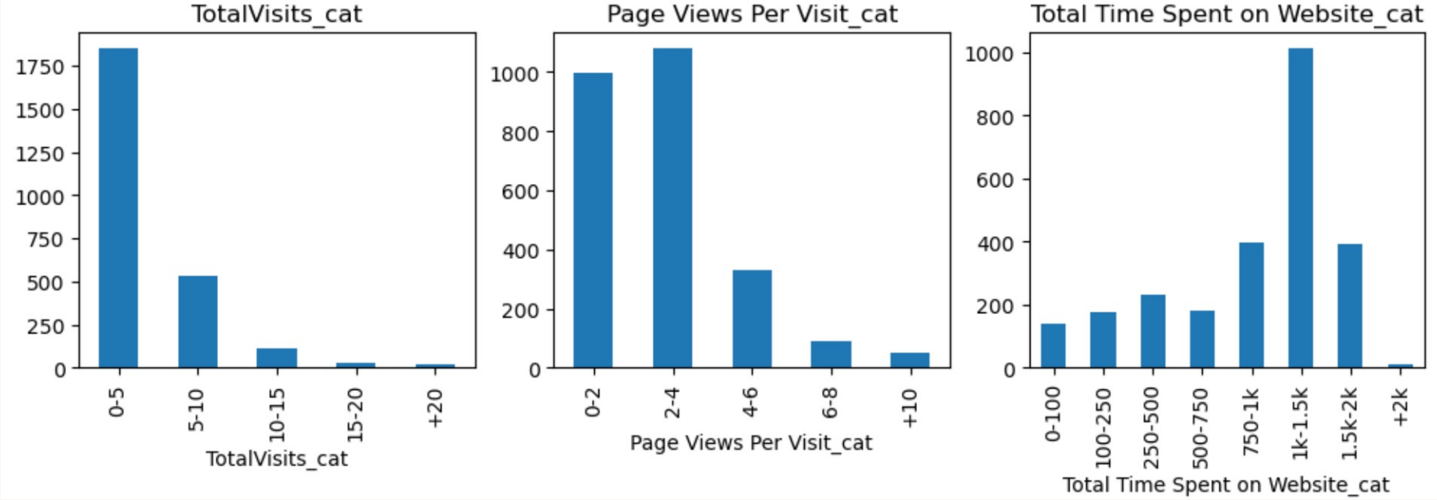
# Correlation matrix of numerical columns



| | Lead Number | Converted | TotalVisits | Total Time Spent on Website | Page Views Per Visit |
|---|---|---|---|---|---|
| Lead Number | 1 | 0.025 | 0.056 | 0.02 | 0.066 |
| Converted | 0.025 | 1 | 0.045 | 0.35 | -0.0026 |
| TotalVisits | 0.056 | 0.045 | 1 | 0.36 | 0.77 |
| Total Time Spent on Website | 0.02 | 0.35 | 0.36 | 1 | 0.34 |
| Page Views Per Visit | 0.066 | -0.0026 | 0.77 | 0.34 | 1 |

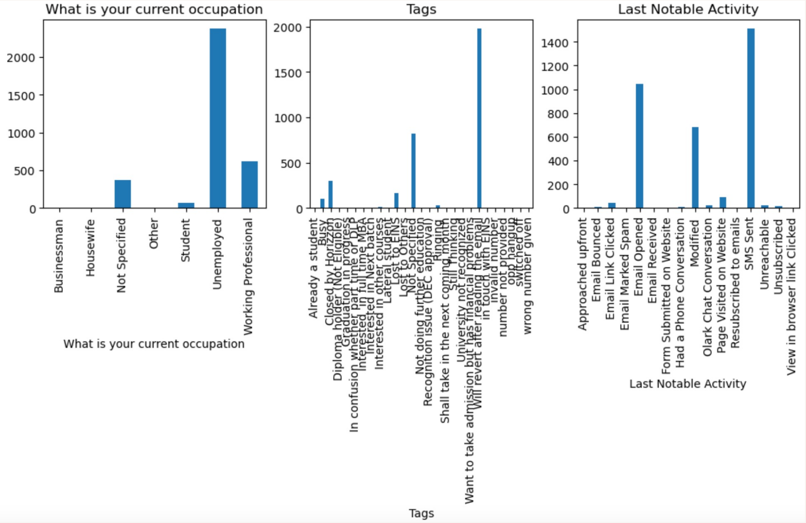
- We do not observe a high correlation here except in between Total visits & Page views per visit

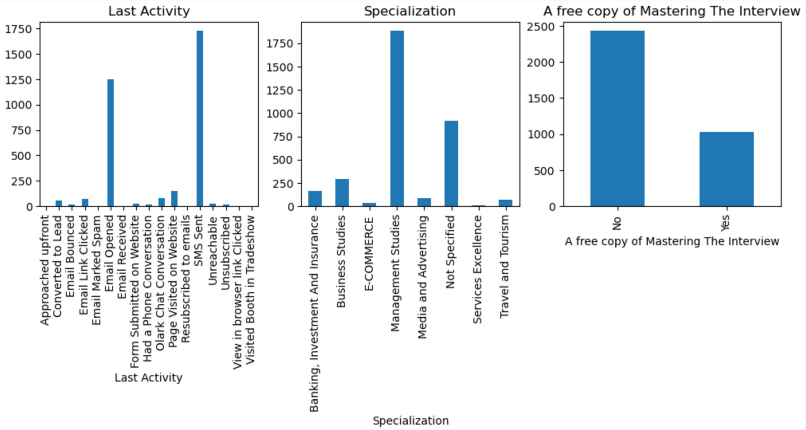
# Binning the numerical variables

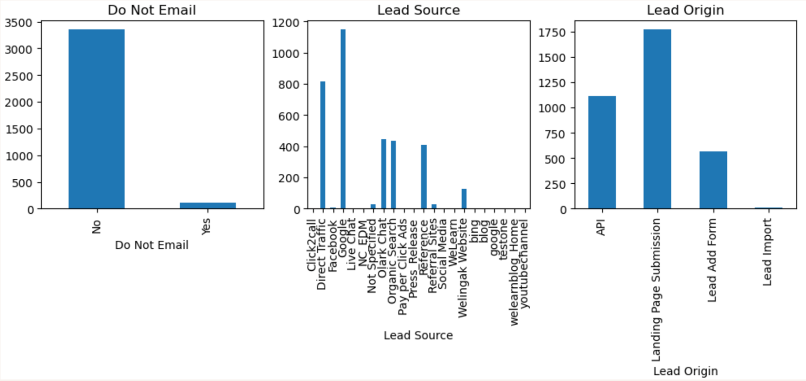


- We can see here that the conversions for people who frequently visit the website and spend more than average time on it is higher

# EDA on categorical columns

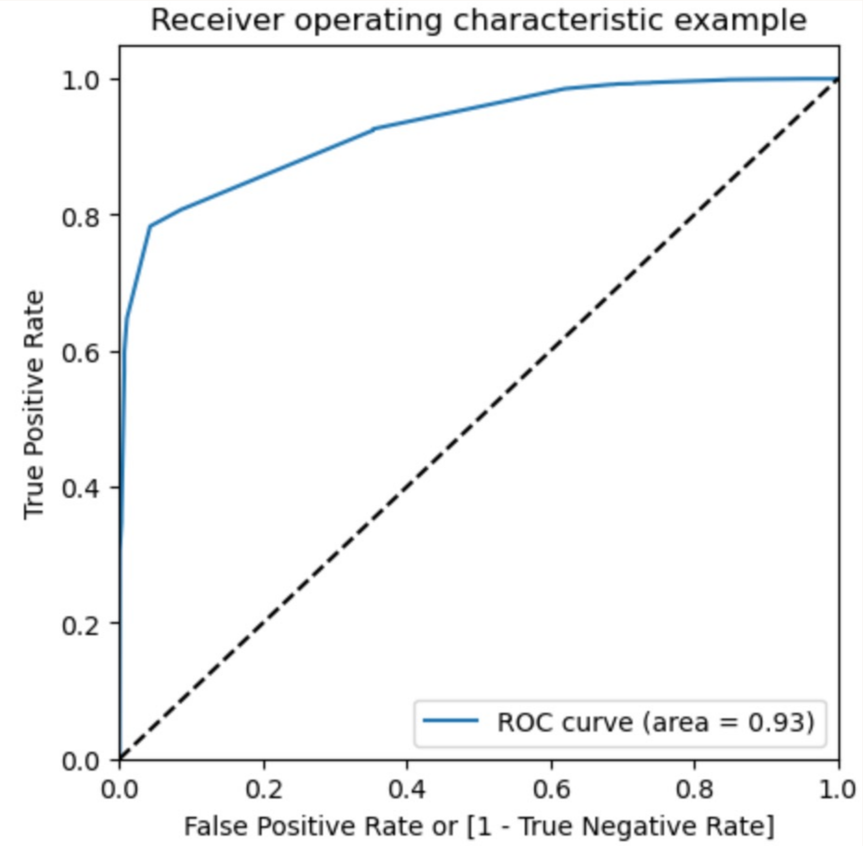


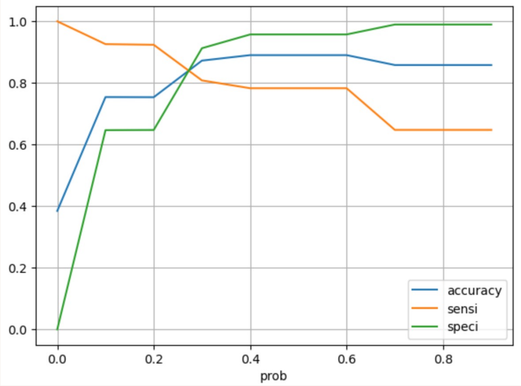


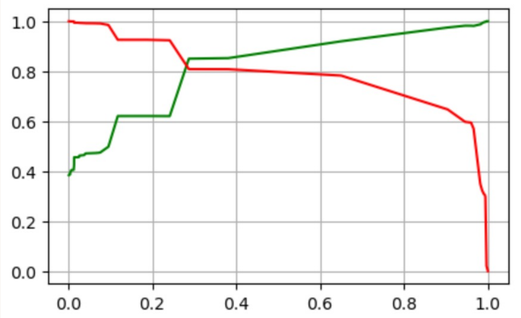


- We see the high variability in the converted leads for these categorical variables based on the values present
- This will further help in identifying the important columns when we create the dummy variables

# Model evaluation (Train set)

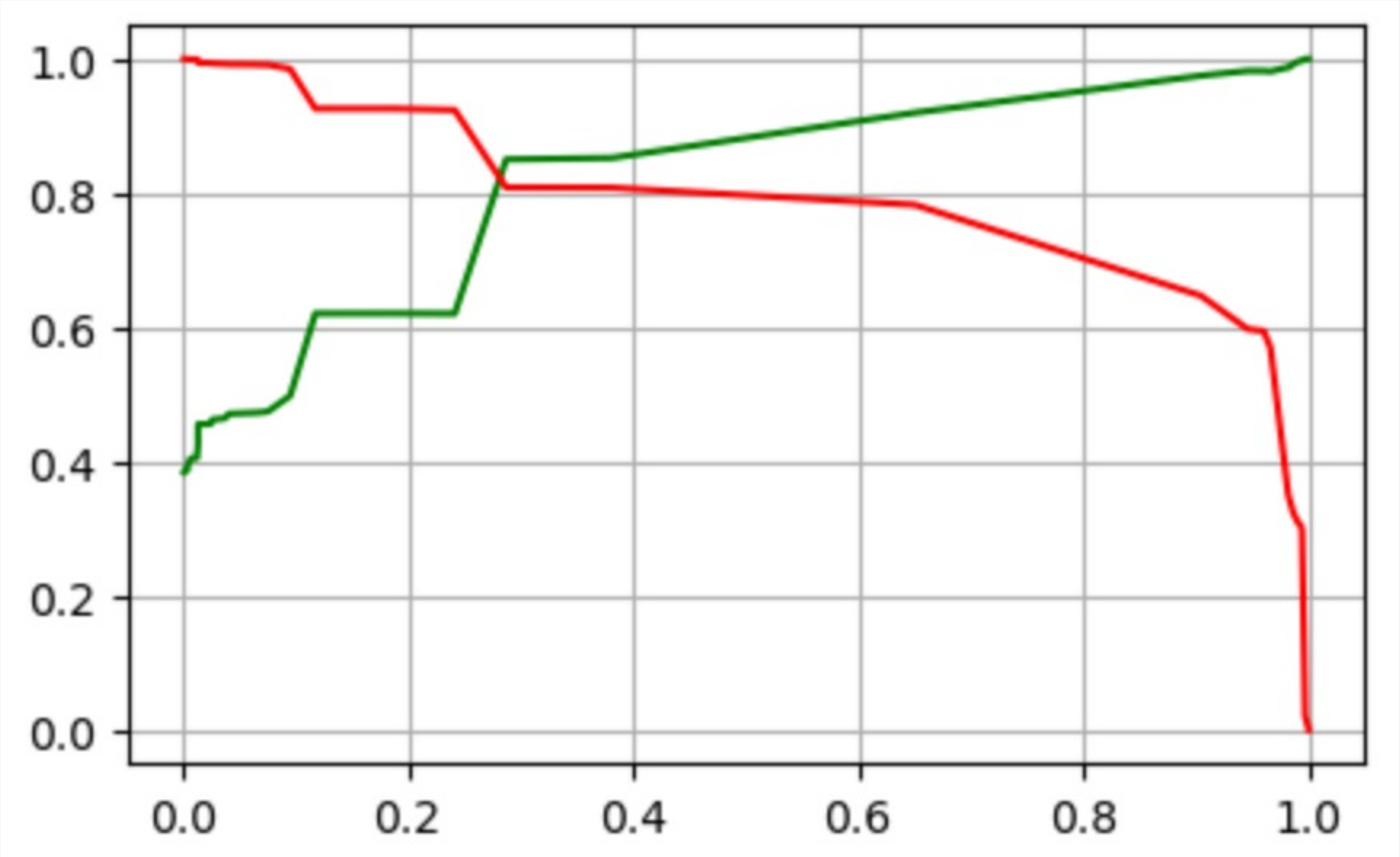






- Accuracy: 87.17%
- Sensitivity: 80.86%
- Specificity: 91.11%
- Precision: 91.98%
- Recall: 78.27%

# Model evaluation (Test set)



- Accuracy: 87.71%
- Sensitivity: 82.38%
- Specificity: 90.81%
- Precision: 83.83%
- Recall: 82.38%

- Test set threshold is se to 0.27

# Recommendations

- People that visit the website frequently and spend higher that average time on it are promising leads
- Lead sources which are from references, direct traffic or Olark chat should be pursued
- SMS messages have an impact on lead conversion
- People referring should be awarded so that possible leads can increase
- Working professionals who have a management specialization can be good leads
- We may want to target students and unemployed since the courses will empower them for better jobs and career prospects