Lead Scoring Case Study

GROUP MEMBERS-

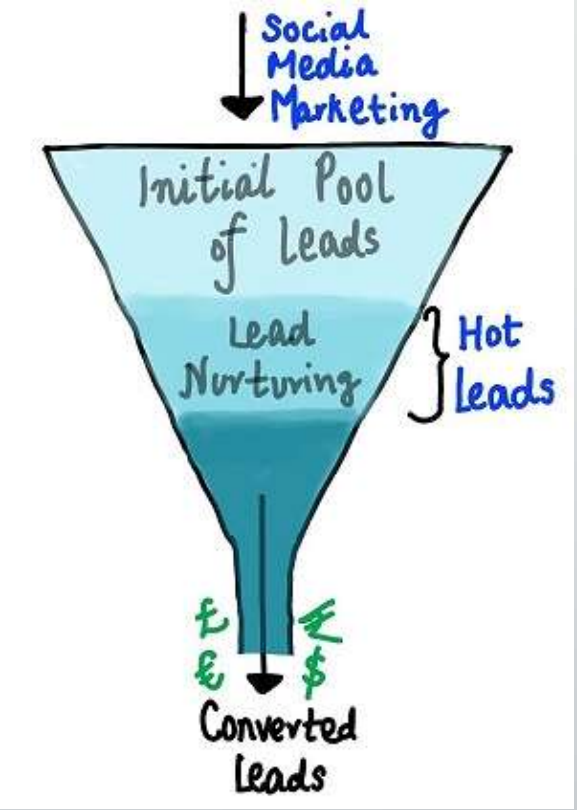
RAHUL PAL

ARTI DEVARSHE

TANMAY

- **Problem Statement**
- Education company named X Education sells online courses to industry professionals. Company got lots of leads, but lead conversion is poor, only about 30%. To make this conversion higher company wants to identify potential leads. If company identify potential leads sales team will focus more on communicating with the potential leads rather than call to everyone.

- **Solution**
- By using leads data company need to build logistic regression model so that by applying that model education company can achieve potential leads



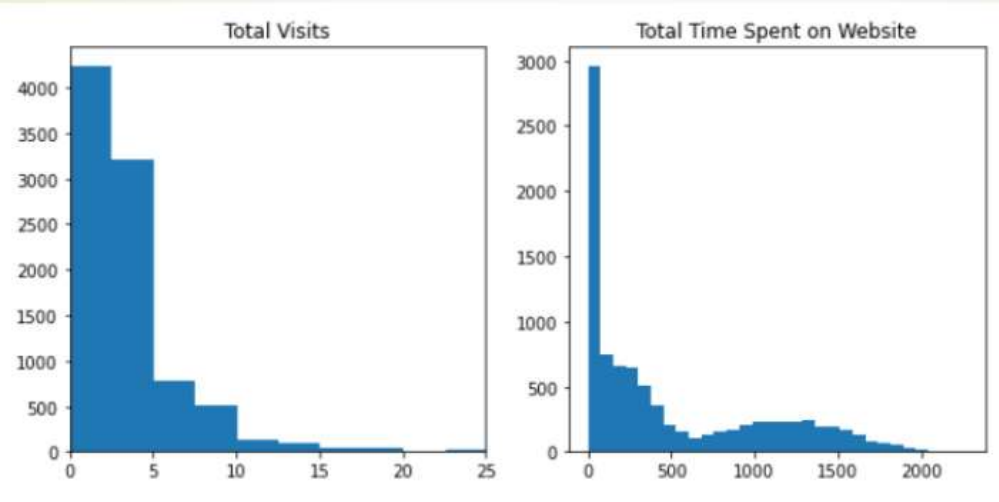
Data Processing Route

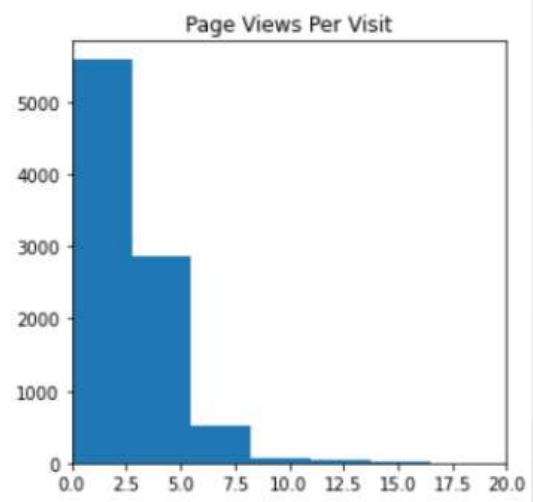
- **A.) Data cleaning**
- 1. Check duplicity in data and then remove it when required.
- 2. Check for NA values and missing values.
- 3. Drop columns, if it contains large number of missing values and not useful for the analysis.
- 4. Imputation of the values, if necessary.
- **B.) EDA**
- 1. Check and handle outliers when it required.
- 2. Univariate data analysis and Bivariate data analysis between the variables and Visualization.
- **C.)** Feature Scaling & Dummy Variables and encoding of the data.
- **D.)** Classification technique: Logistic regression used for the model making and prediction.
- **E.)**Validation of the model.
- **F.)** Model presentation.
- **G.)** Conclusions and recommendations.

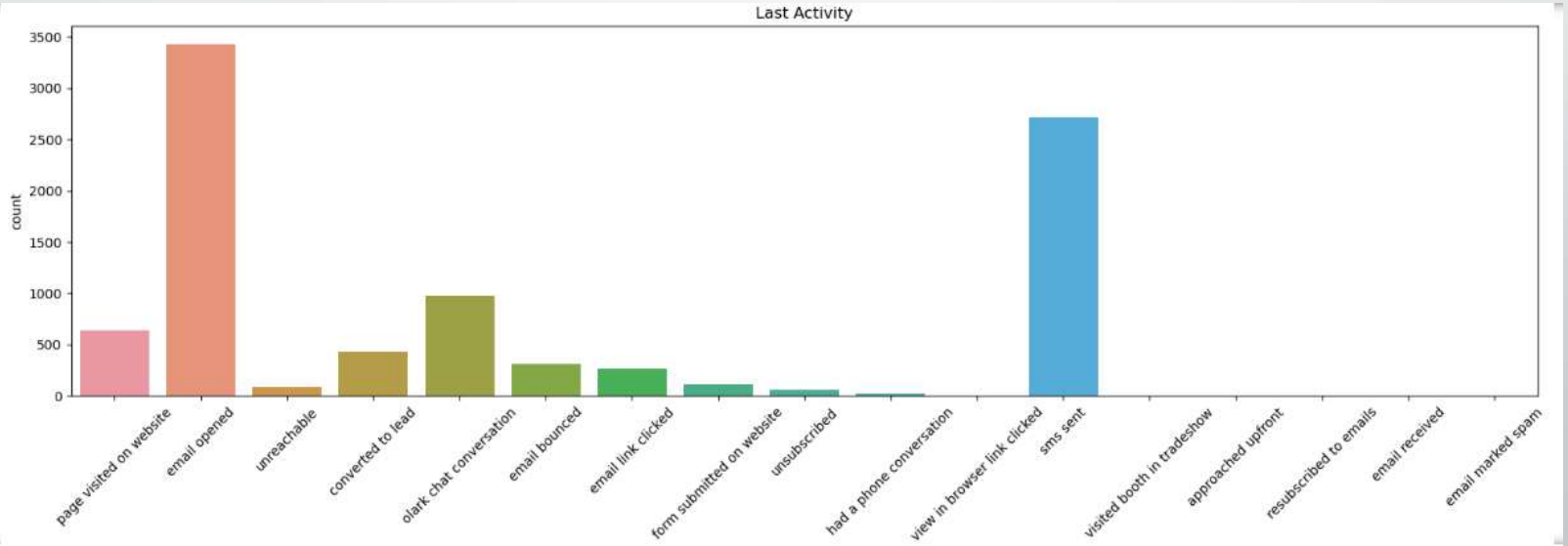
Data cleaning and manipulation

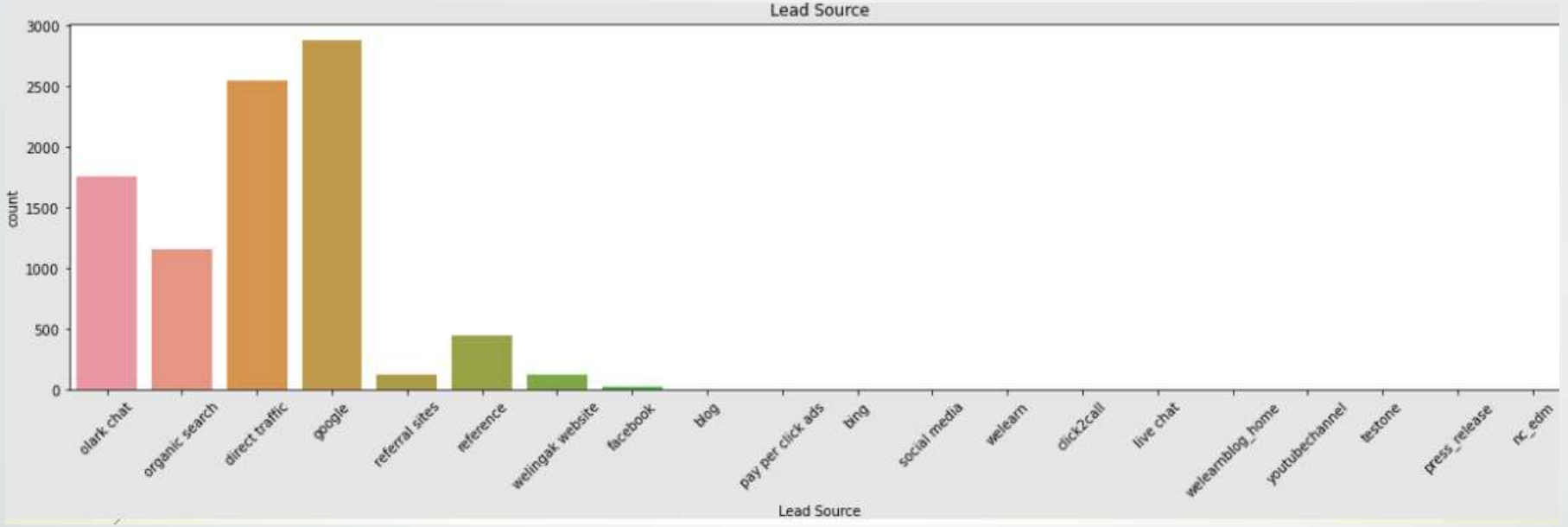
- We have seen that dataset has 9240 rows and 37 columns.
- Dropping several variables which contains unique values.
- Checking for duplicates shows no duplicates in dataset.
- Dropped columns which are not necessary for analysis.
- Dropping the columns ('Tags', 'Lead Quality', 'Asymmetrique Profile Index', 'Asymmetrique Activity Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score', 'Lead Profile', 'How did you hear about X Education') having more than 35% as missing.

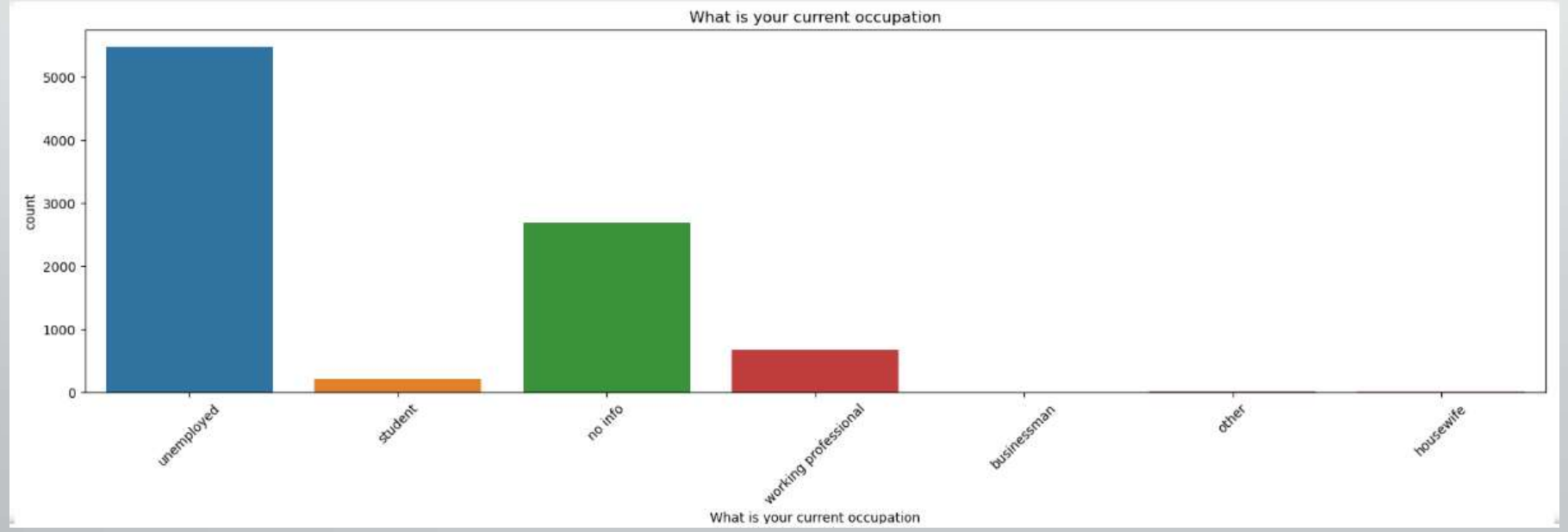
EDA – (Univariate and Bivariate Analysis) with Categorical Variables








Lead Source
count
3000
2500
2000
1500
1000
500
0
olark chat
organic search
direct traffic
google
referral sites
reference
welingak website
facebook
blog
pay per click ads
bing
social media
welearn
click2call
live chat
welearnblog_home
youtubechannel
testone
press_release
nc_edm
Lead Source

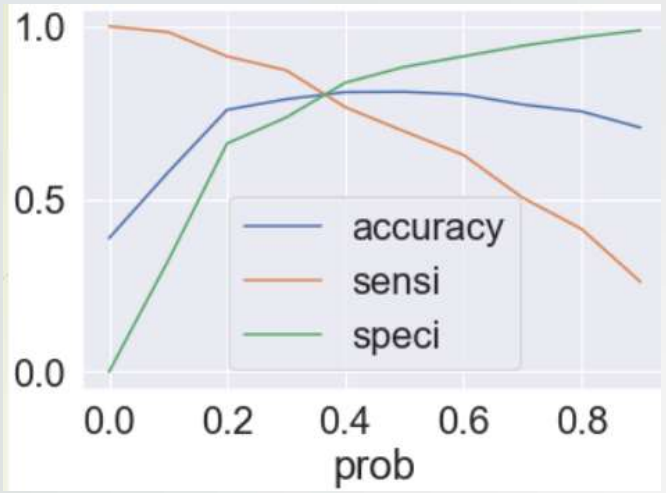

What is your current occupation
count
5000
4000
3000
2000
1000
0
unemployed
student
no info
working professional
businessman
other
housewife
What is your current occupation

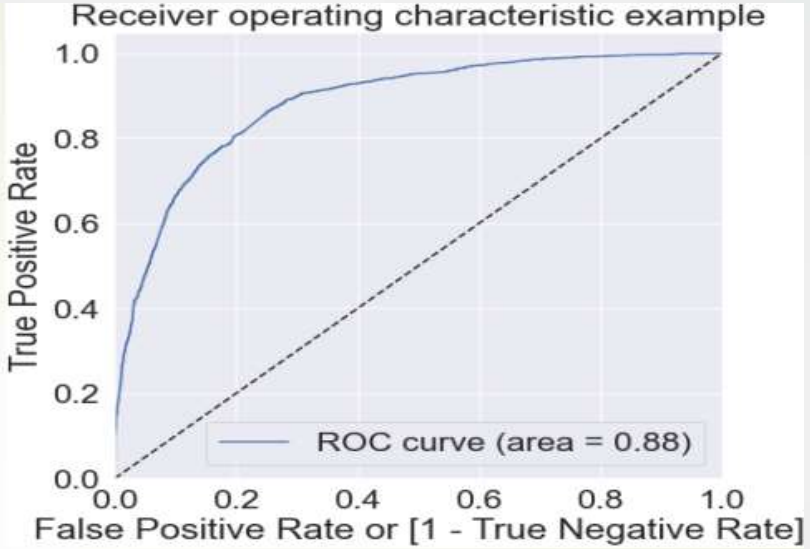
Data Conversion

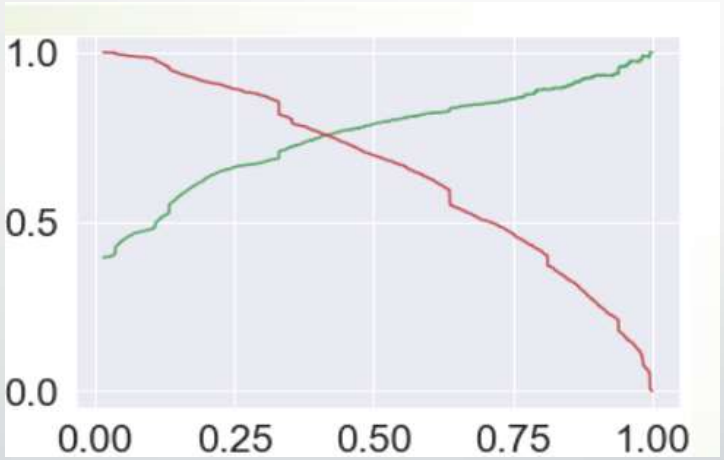
- Numerical variables are Normalized.
- Dummy variables are created for object type variables.
- We have 9074 rows and 81 columns for analysis.

Model Building

- Splitting data in train and test set, 70% train and 30% test.
- Using RFE for feature selection & running RFE with 15 variables as output.
- Building Model by removing the variable whose p- value>0.05 and VIF >5.
- Predictions on test data set.
- Overall accuracy attain on cut off value 0.35 is 81% approximately.







- We have balanced sensitivity and specificity at optimal cut off value 0.35.
- We have balanced precision and recall value at cut off value 0.41

Conclusion

We have found that the variables that mattered the most in the potential buyers are:

1. The total time spend on the Website
2. Total number of visits
3. When the lead source was Welingak website, Olark chat, Organic search and Google.
4. When the last activity was:
    a) SMS
    b) Olark chat conversation
5. When the lead origin was 'Lead add form'.
6. When their current occupation is working professional or unemployed.

So, The X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.