LEAD SCORING

THE PROBLEM

An education company named X Education which sells online courses to industry professionals, wants to come with an efficient process to identify the most potential leads, also known as ‘Hot Leads’.

The company tracks various parameters when individuals visit their website and wishes to use this data in identifying patterns.

STEPS TAKEN

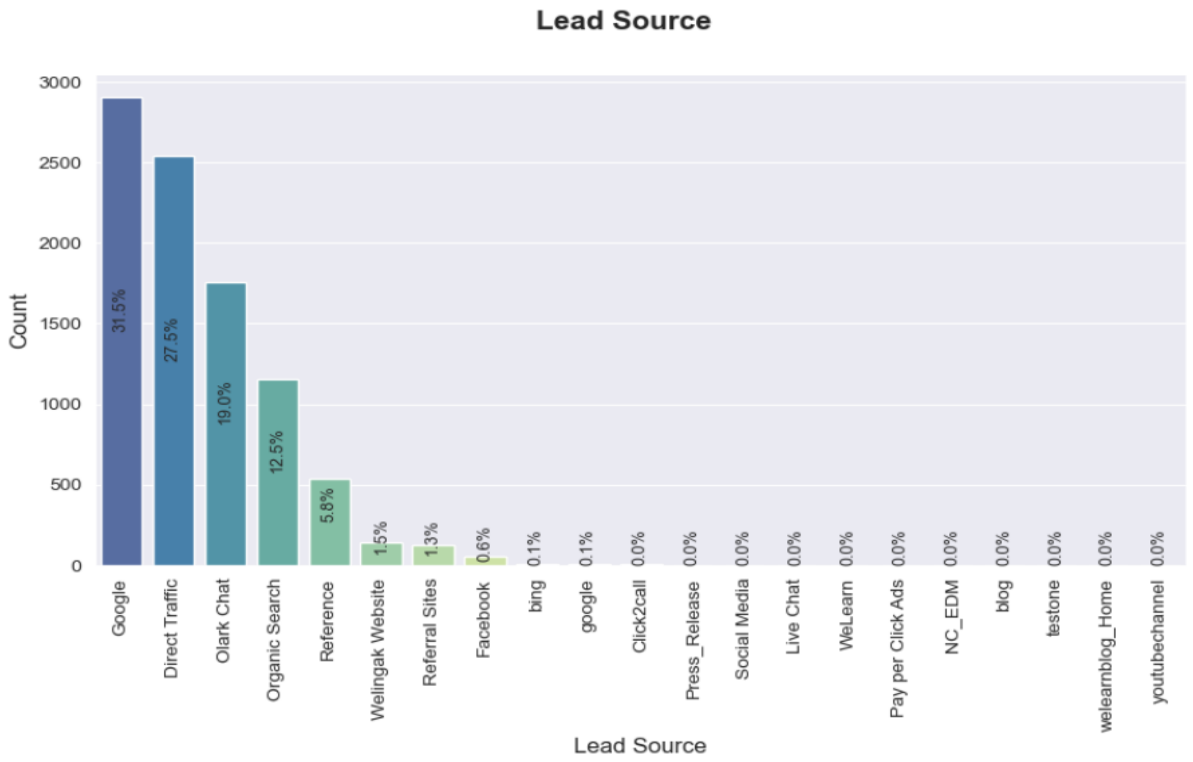
- Data Understanding
- Data Cleaning
- Exploratory Data Analysis – Univariate, Bivariate and Multivariate
- Data Preparation and Splitting into Train and Test set
- Building Models
- Model evaluation on Train data set
- Model evaluation on Test data set

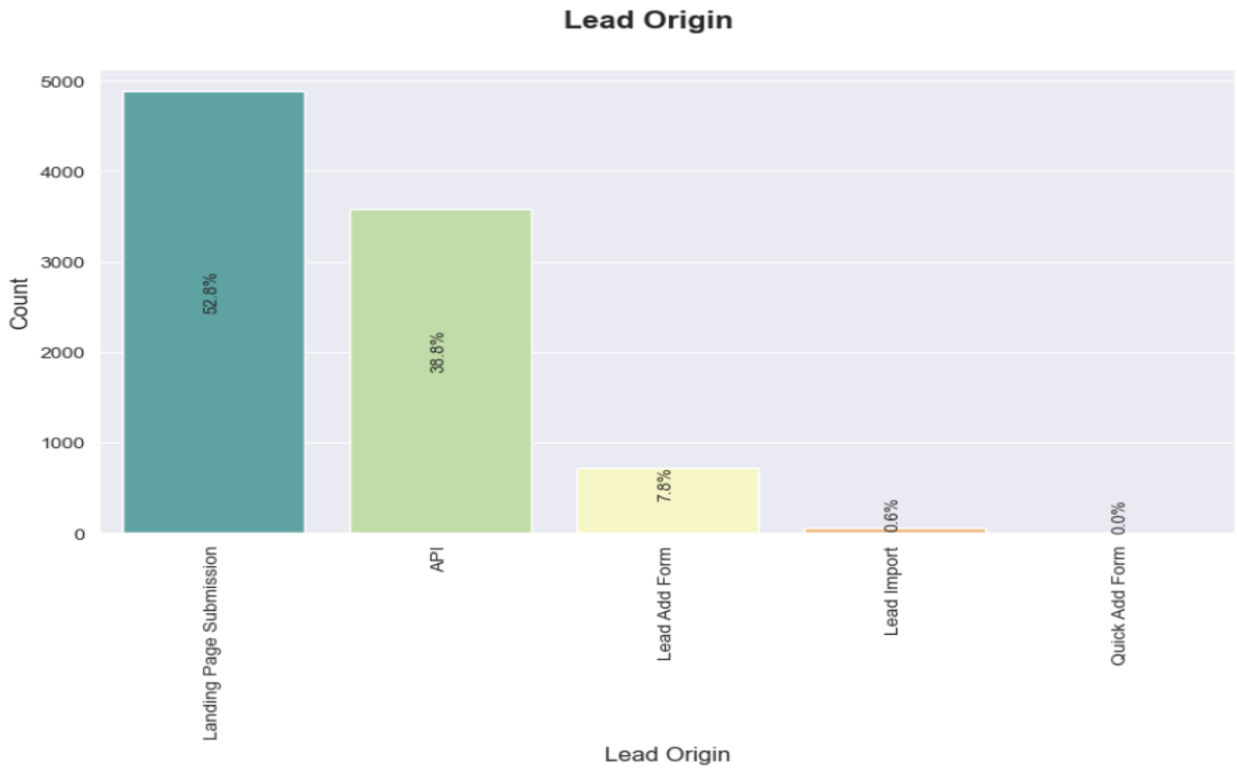
STEPS TAKEN –DATA CLEANING

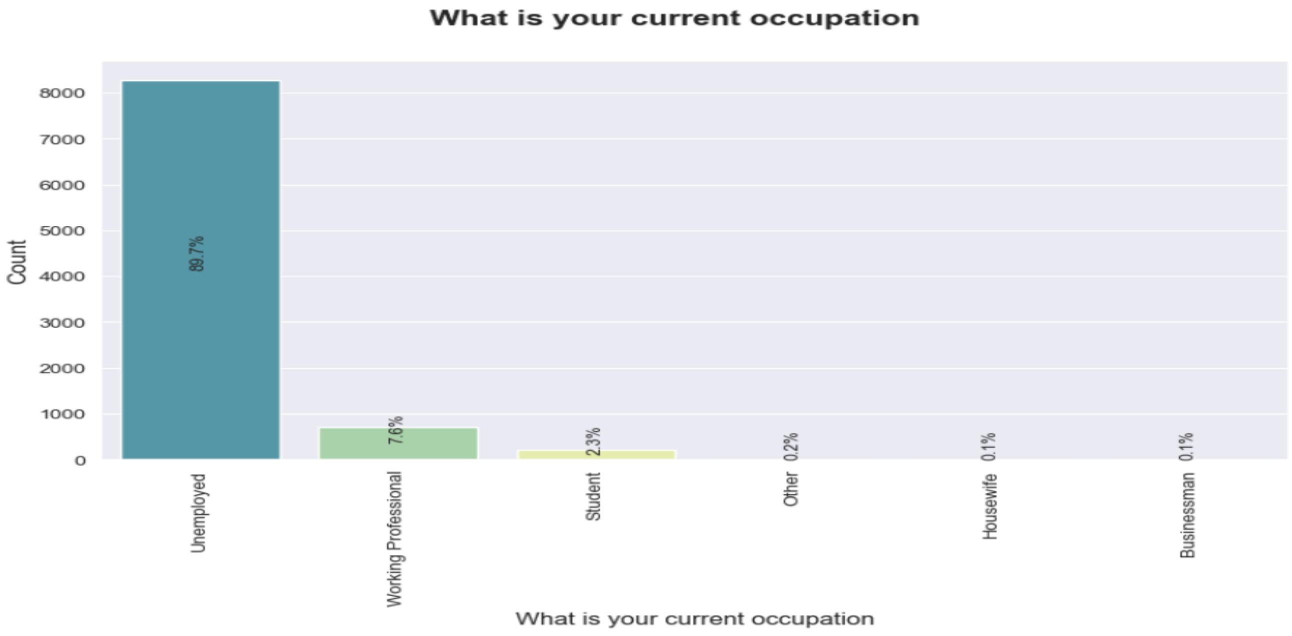
- Cells with value as ‘Select’ were marked as ‘Missing’
- Columns with more than 30% missing values were dropped
- No issues found with respect to data types
- Remaining missing values were imputed with mode of the column
- Columns having unique value for each row were dropped
- Columns having same value for all rows were dropped
- Outliers were capped to Q3 + 1.5 * IQR value

STEPS TAKEN – EXPLORATORY DATA ANALYSIS

- Univariate Analysis
  - Bar plots were plotted for Categorical variables
  - Box plots were plotted for Numerical variables
  - Inferences:
    - Dominant Lead Sources: Google, Direct Traffic, Olark Chat
    - Dominant Lead Origin: Landing Page Submission, API, Lead Add Form
    - Dominant Country: India
    - Dominant Occupation: Unemployed, Working Professional


Lead Source

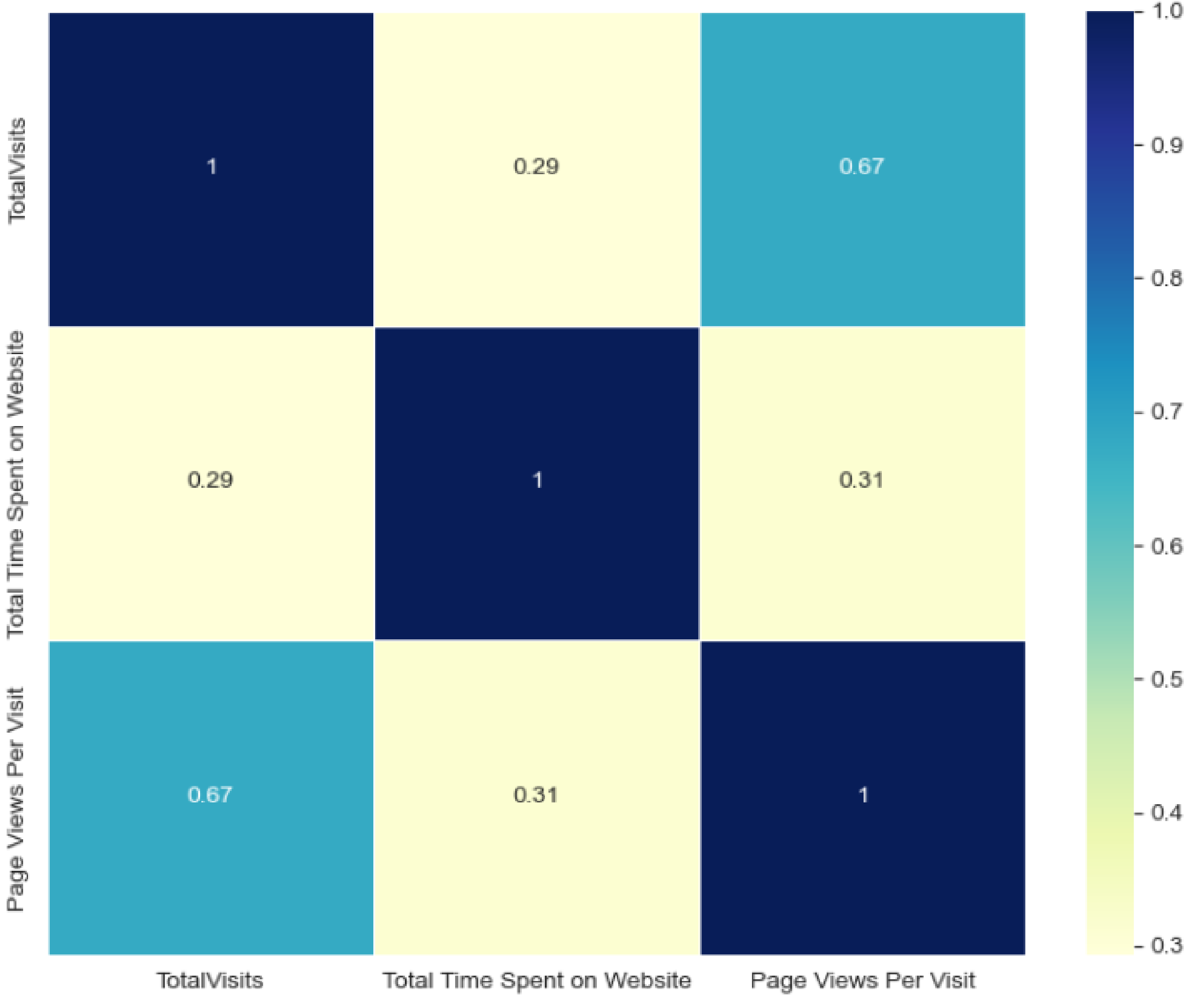

Lead Origin

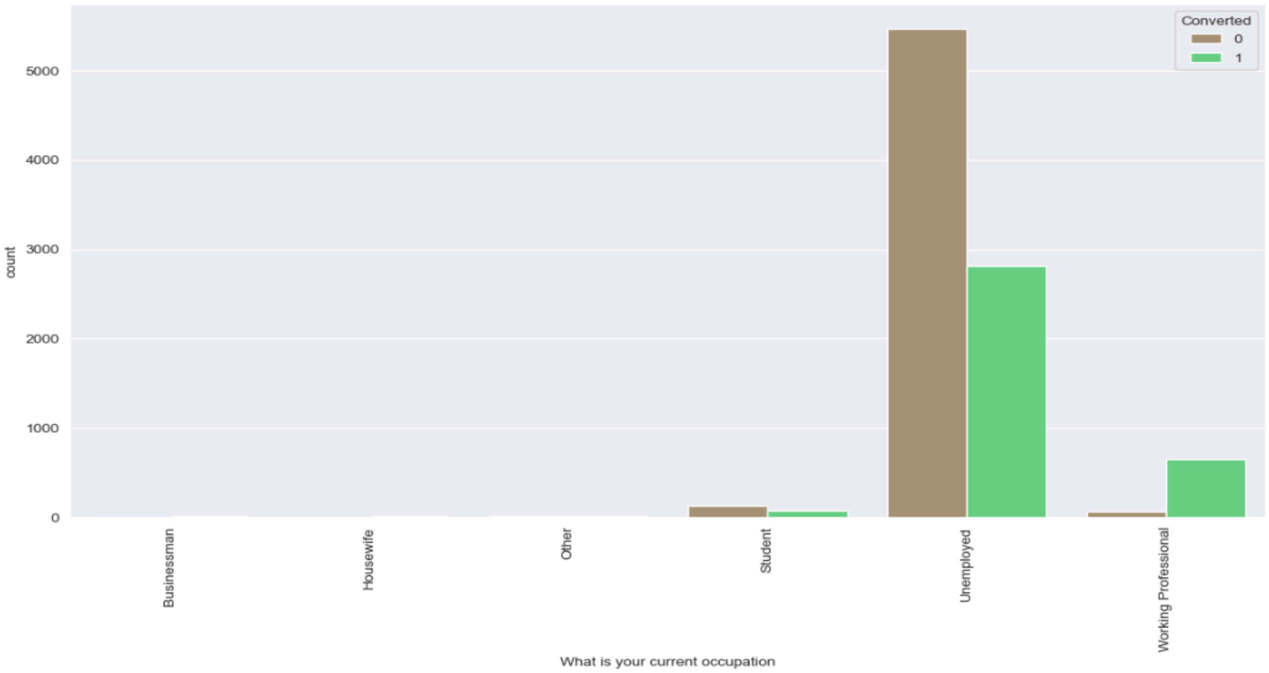

What is your current occupation

STEPS TAKEN – EXPLORATORY DATA ANALYSIS

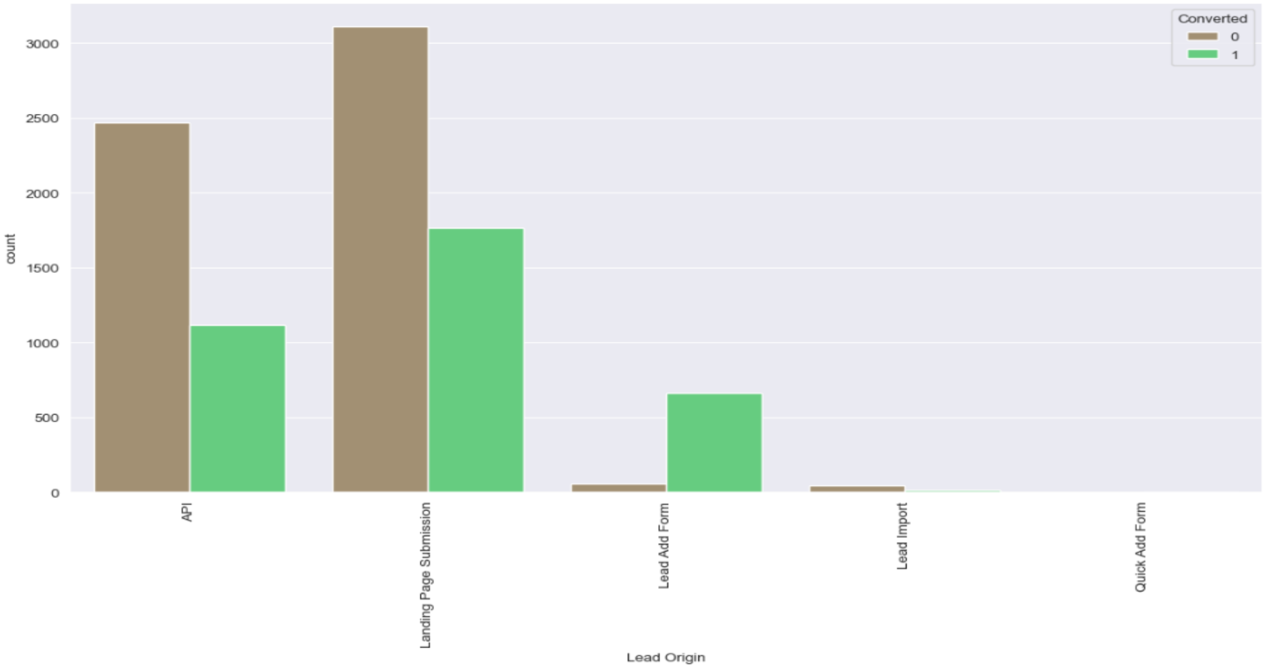
- Bivariate Analysis
    - Bar plots were plotted for Categorical variables against target variable
    - Inferences:
        - Last Notable Activity / Last Activity = SMS Sent, has higher conversion rate
        - Lead Source = 'Reference' and 'Welingak Wesite' has higher conversion rate
        - Lead Origin = 'Lead Add Form' has higher conversion rate
        - Occupation = 'Working Professional' has higher conversion rate
    - Numerical variables were plotted against each other
    - Inference:
        - TotalVisits had linear relationship with Page Views per Visit, Same was evident with correlation matrix as well.

Heatmap of Continous Numerical Columns

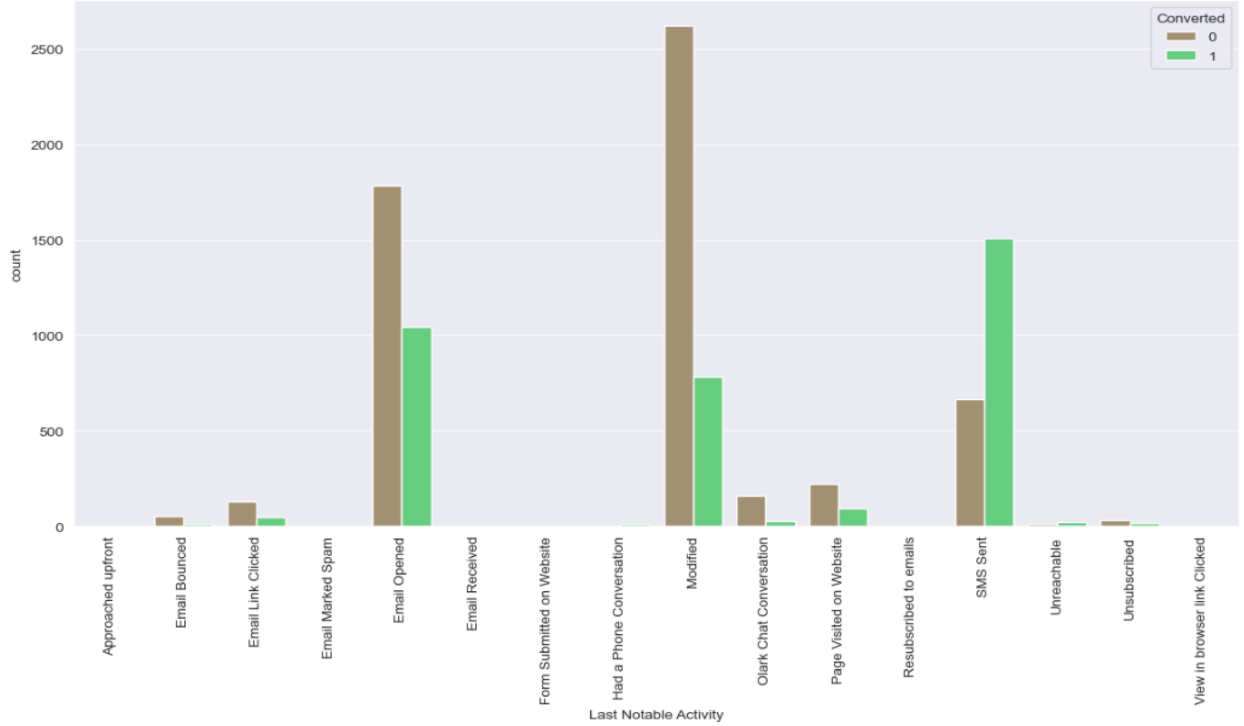




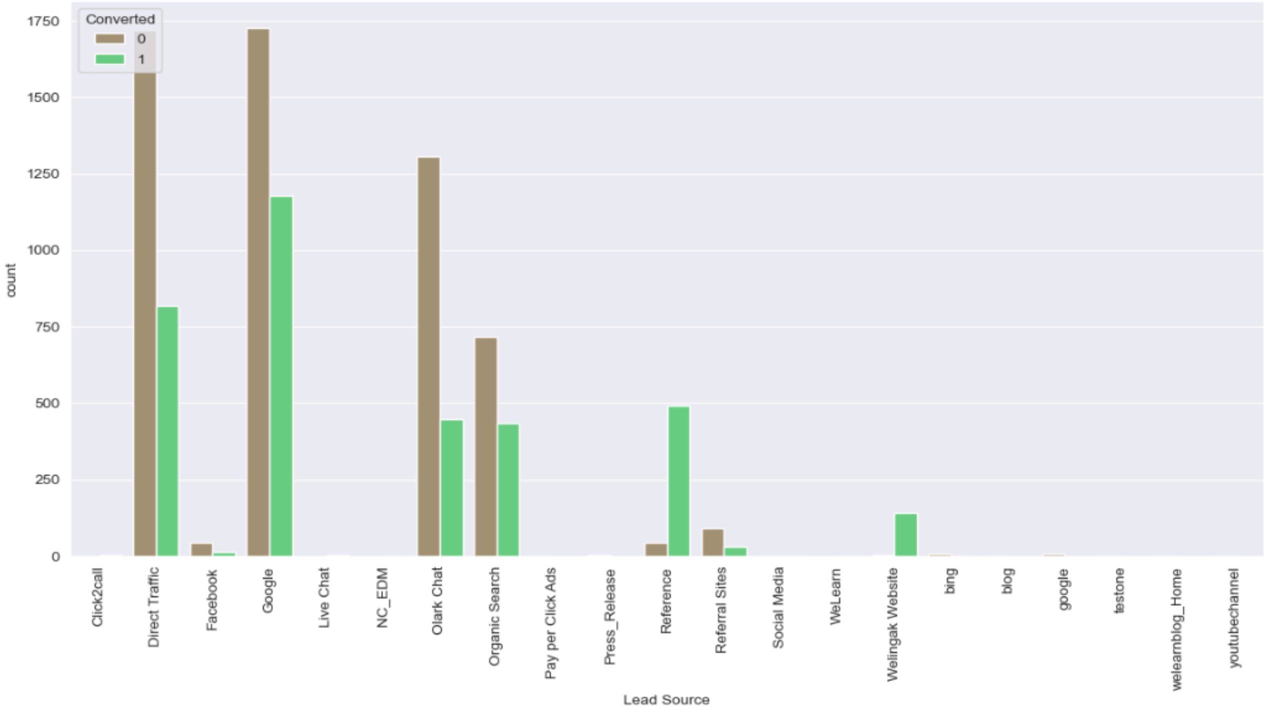
Converted
0
1
count
0
1000
2000
3000
4000
5000
Businessman
Housewife
Other
Student
Unemployed
Working Professional
What is your current occupation



Converted
0
1
count
0
500
1000
1500
2000
2500
3000
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Origin



Converted
0
1
count
0
500
1000
1500
2000
2500
Approached upfront
Email Bounced
Email Link Clicked
Email Marked Spam
Email Opened
Email Received
Form Submitted on Website
Had a Phone Conversation
Modified
Olark Chat Conversation
Page Visited on Website
Resubscribed to emails
SMS Sent
Unreachable
Unsubscribed
View in browser link Clicked
Last Notable Activity



Converted
0
1
count
0
250
500
750
1000
1250
1500
1750
Click2call
Direct Traffic
Facebook
Google
Live Chat
NC_EDM
Olark Chat
Organic Search
Pay per Click Ads
Press_Release
Reference
Referral Sites
Social Media
WeLearn
Welingak Website
bing
blog
google
testone
welearnblog_Home
youtubechannel
Lead Source

STEPS TAKEN – DATA PREPARATION

- Mapped Yes/No fields to 1/0
- Created Dummy variables for Categorical variables
- Data was divided into 70:30 ratio for train and test
- Numerical features were scaled in train data set
  - MinMax Scaler was used

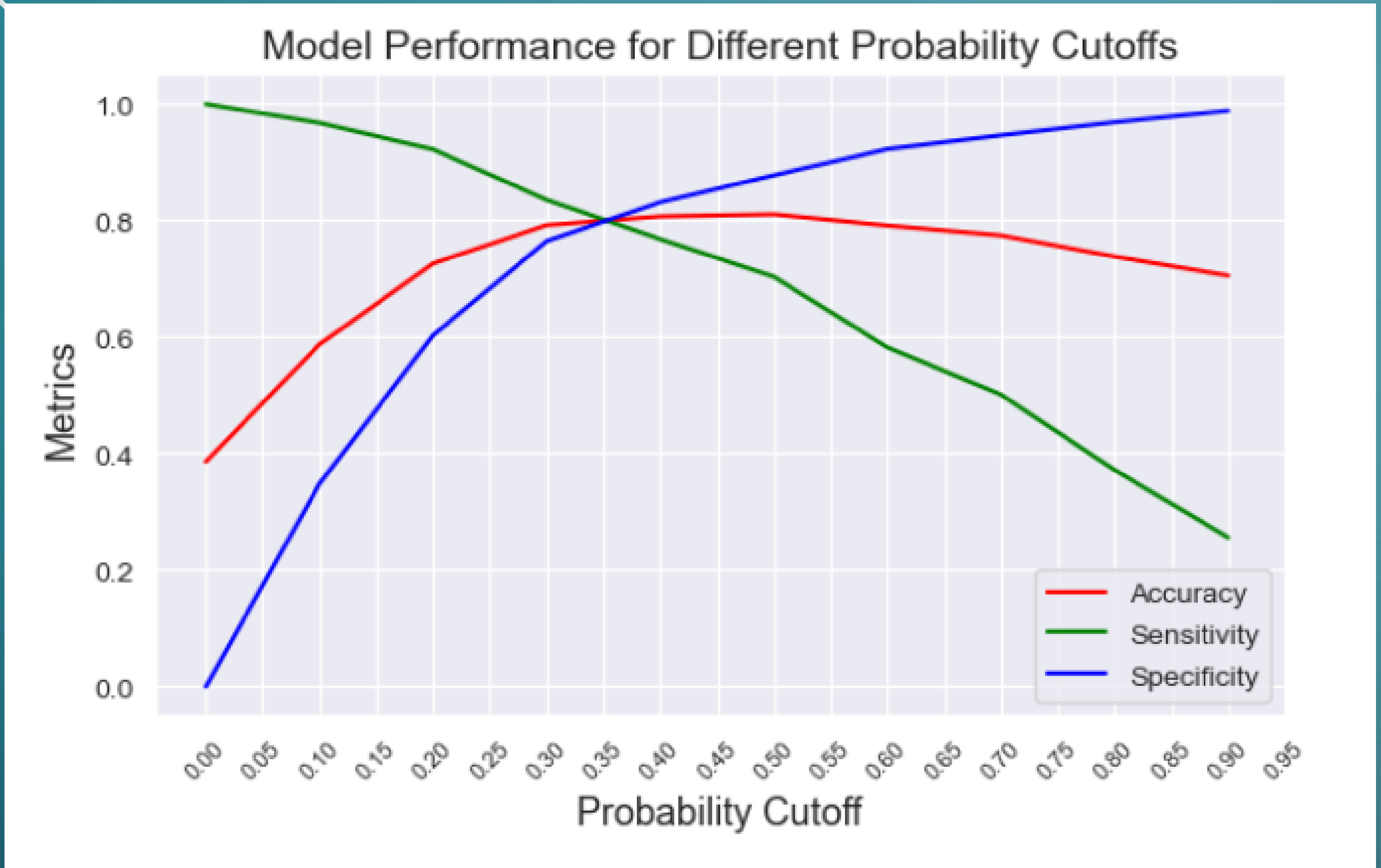
STEPS TAKEN – BUILDING MODELS

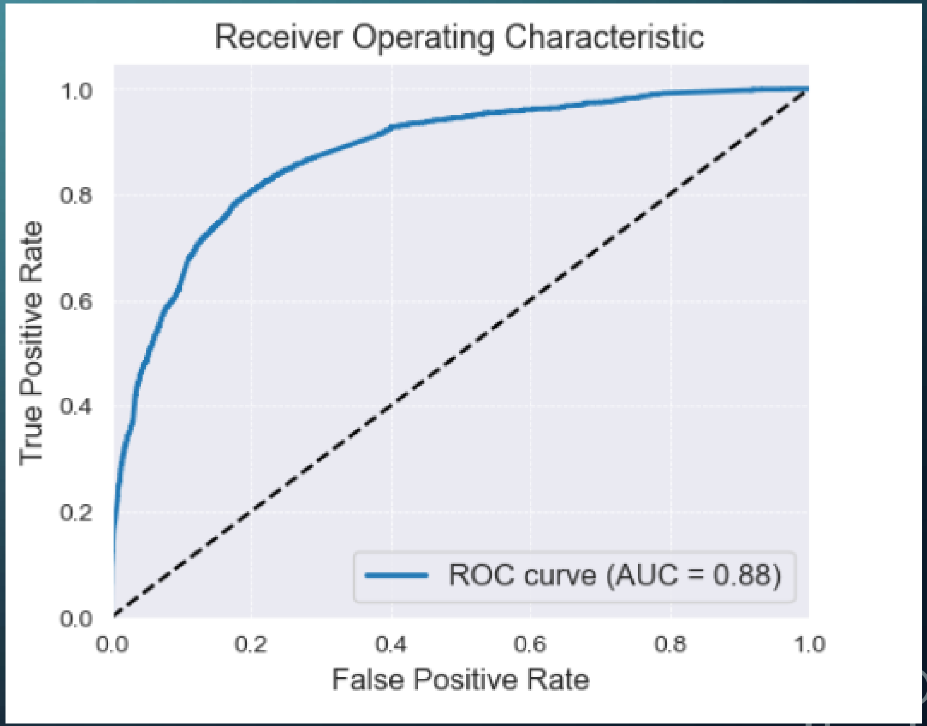
- Top 15 features were selected using RFE

  'TotalVisits', 'Total Time Spent on Website’, 'Lead Origin_Lead Add Form’, 'Lead Source_Olark Chat’, 'Lead Source_Welingak Website', 'Do Not Email_Yes’, 'Last Activity_Had a Phone Conversation’, ‘Last Activity_Olark Chat Conversation', 'Last Activity_SMS Sent’, 'Country_Italy', 'Country_Saudi Arabia’, 'What is your current occupation_Housewife’, 'What is your current occupation_Working Professional’, 'Last Notable Activity_Had a Phone Conversation’, 'Last Notable Activity_Unreachable’
- Model was built, Insignificant features were dropped one at a time based on their p-value. VIF was in acceptable range for all features
  - Features dropped:
    - Country_Italy, Country_Saudi Arabia, What is your current occupation_Housewife,
    - Last Notable Activity_Had a Phone Conversation

STEPS TAKEN – MODEL EVALUATION ON TRAINING DATA

- ROC Curve was plot. AUC = 0.88
- Plot of Accuracy, Sensitivity and Specificity was created for different thresholds
  - Optimal Threshold came out to be 0.35
- Following metrics were calculated at Optimal Threshold:
  - Overall Model Accuracy: 80.19%
  - Sensitivity: 0.8
  - Specificity: 0.8
  - Positive Rate: 0.72
  - Negative Rate: 0.87


Model Performance for Different Probability Cutoffs
Metrics
1.0
0.8
0.6
0.4
0.2
0.0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
Probability Cutoff
Accuracy
Sensitivity
Specificity


Receiver Operating Characteristic
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve (AUC = 0.88)

STEPS TAKEN – MODEL EVALUATION ON TEST DATA

- Following metrics were calculated:
  - Overall Model Accuracy: 79.82%
  - Sensitivity: 0.8
  - Specificity: 0.8
  - Positive Rate: 0.71
  - Negative Rate: 0.87

FINAL RESULT

- The resulting model can predict at 80% accuracy.
- Top 3 contributing Factors came out to be:
  - ‘Total Time Spent on Website’
  - ‘Lead Origin_Lead Add Form’
  - Occupation_Working Professional

RECOMMENDATIONS

- Business should first target Working professionals whose Lead Originated from Lead Add Form, in descending order of Total Time Spent on Website.
- Business can make changes on the website to make content more engaging. As people spend more time on a website, they instinctively start to trust it more.
- Business should promote Lead Add Form more, as the leads generated from that have proved to be more reliable.