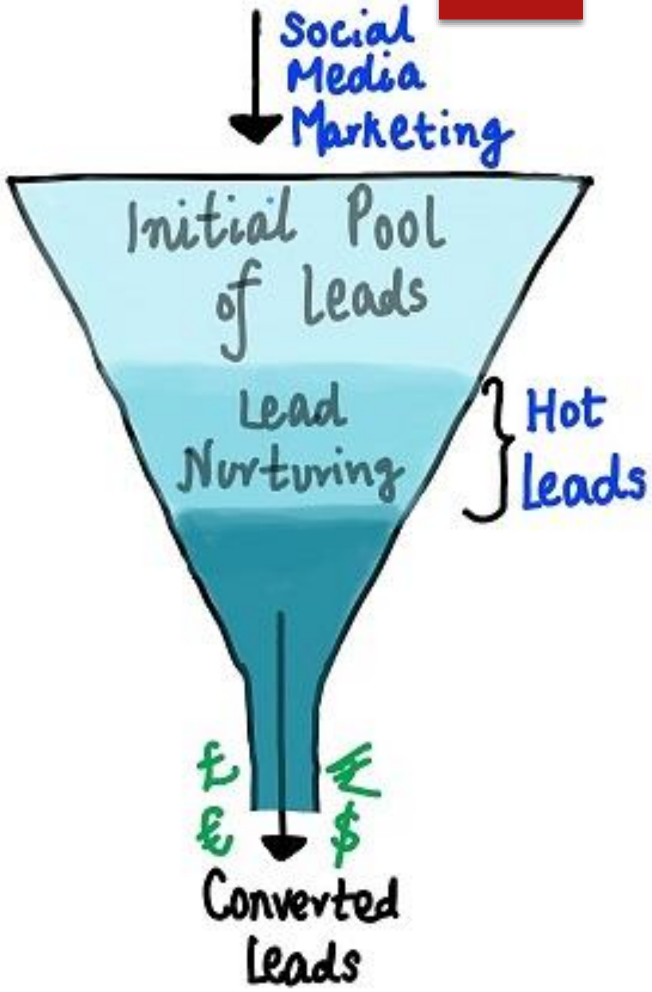
ONLINE LEAD SCORING CASE STUDY

BUSINESS PROBLEM

- An education company named X Education sells online courses to industry professionals.
- The company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- With the help of hot 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.

BUSINESS OBJECTIVE

- To build a machine learning model, mainly 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.



STEPS TO ATTAIN OBJECTIVE

Importing and Inspecting Dataset

Data Manipulation

Exploratory Data Analysis

Data Preparation

Model Building

Model Evaluation on Train Data Set

Model Evaluation on Test Data set

IMPORTING AND INSPECTING DATASET

Leads.csv data set was imported

Basic information about features data types, missing values was checked.

Shape of the data set was also checked.

The data set was having 37 columns and 9240 rows.

DATA MANIPULATION

Cleaning of Data

- Select values were present in many features, which were updated to null as per business perspective.
- Null percentage was checked in columns and rows.
- Number of unique values was checked for each feature.

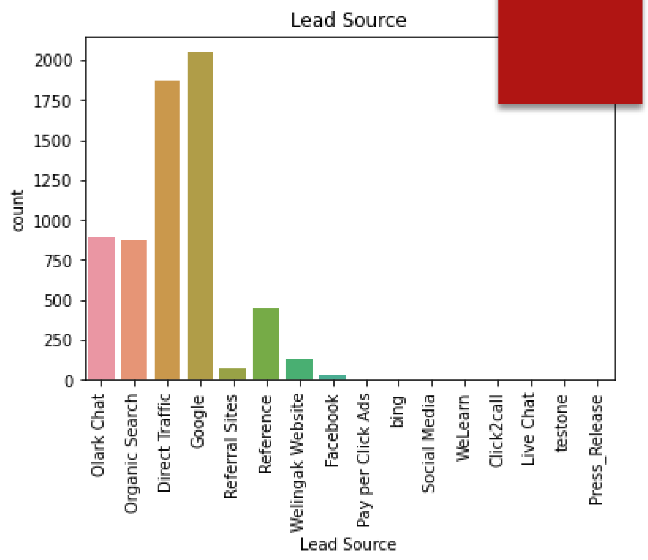
Missing Value Treatment

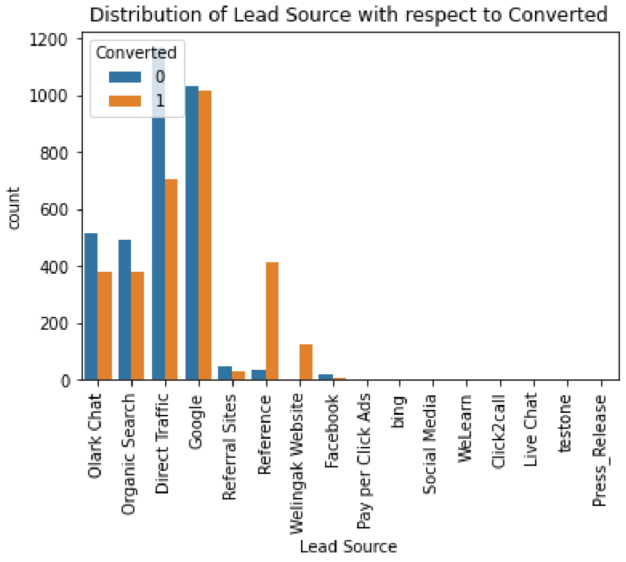
- Features with more than 35% missing values were dropped.
- Some rows were also dropped as imputation of these were creating a skewed data.
- Also, some missing values were imputed as in feature Country.

At the end of this step, 69% rows were left.

EXPLORATORY DATA ANALYSIS

- Outlier Detection
  - Outliers were checked in the numerical continuous variables, I f any they were capped as per the quartiles obtained.
- Univariate Analysis
  - In this all the features were plotted to analyze their distribution.
  - Some features which were having highly skewed data were dropped
- Bivariate Analysis
  - Relation between target and other variables were plotted to analyze their distribution.





DATA PREPARATION

Dummy Variable

- Dummy variables were created for the categorical columns

Splitting dataset in Train and Test data

- Dataset was split into train and test with 7:3 ratio

Scaling of Features

- The numerical features were scaled to have a common range so that their values do not influence their relevance in the model.

MODEL BUILDING

Feature Selection through RFE

- Through RFE, best 15 features were selected

Manually removing features from model

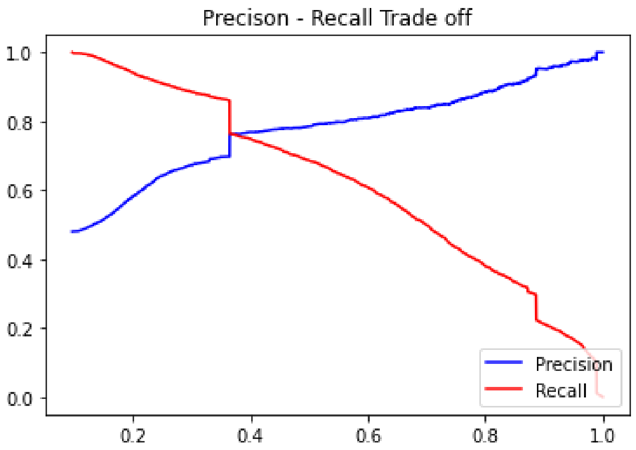
- After RFE, models were built and analyzed iteratively and variables were removed, until the p-value and VIF of every feature was not obtained within the threshold limit.

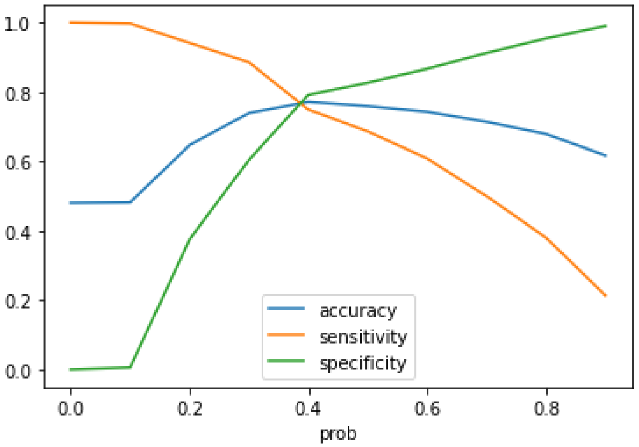
Threshold

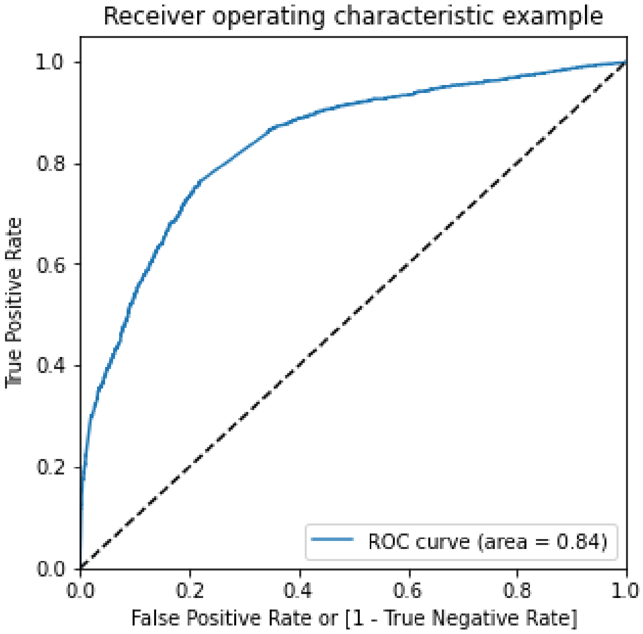
- VIF < 5
- p-value < 0.05

ROC CURVE & THRESHOLD DETECTION

- ▶Area under the ROC curve: 84%
- ▶Precision Recall trade off suggests threshold to be in between 0.35 to 0.4
- ▶Accuracy, sensitivity and specificity trade of also suggests threshold to be between 0.35 to 0.4
- ▶Threshold chosen as 0.37







MODEL EVALUATION ON TRAIN DATA SET

- Confusion Matrix :->
- Accuracy Score : 77%
- Sensitivity Score : 76%
- Sepcificity Score : 78%
- Precision Score : 76%
- Recall Score : 76%

Actual v / Predicted >	Not converted	Converted
Not converted	1811	508
Converted	508	1633

MODEL EVALUATION ON TEST DATA SET

- Confusion Matrix :->
- Accuracy Score : 77%
- Sensitivity Score : 76%
- Sepcificity Score : 79%
- Precision Score : 77%
- Recall Score : 76%

Actual v / Predicted >	Not converted	Converted
Not converted	780	208
Converted	220	704

RESULTS – SUGGESTION TO X COMPANY

The company should focus on those leads which have :

- High number of visits
- Spending much time on website
- Leads having Source:
  - Welingak Website
  - Reference
  - Olark chat
  - Google

Also, the company should avoid leads which have :

- Current occupation:
  - Unemployed
  - Student

Thank you