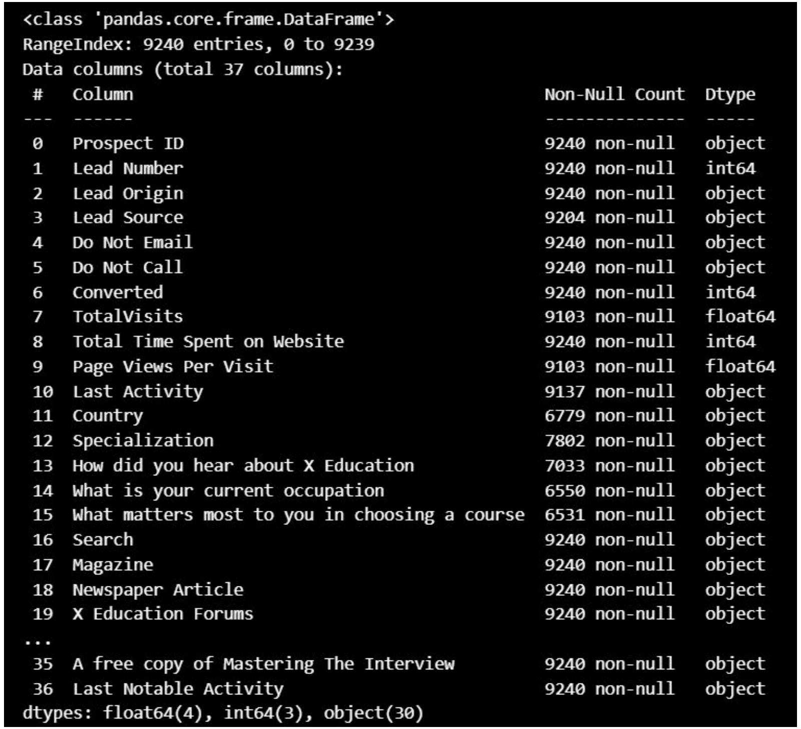
Logistic Regression Lead Score

Updated: 13 Mar 2024

Step 1: Importing and Data general overview

About the dataset

- The data set contains 9240 rows and 37 columns.
- Prospect ID: A unique ID with which the customer is identified.
- Converted: The target variable. Indicates whether a lead has been successfully converted or not.
- 35 independence variables
- Data types: float64(4), int64(3), object(30)



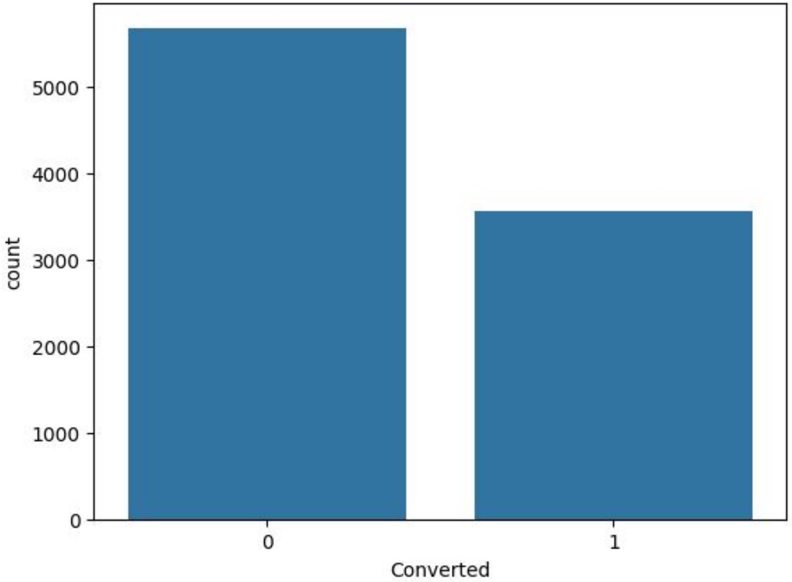
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                     Non-Null Count  Dtype
---  ------                                     --------------  -----
 0   Prospect ID                                9240 non-null   object
 1   Lead Number                                9240 non-null   int64
 2   Lead Origin                                9240 non-null   object
 3   Lead Source                                9204 non-null   object
 4   Do Not Email                               9240 non-null   object
 5   Do Not Call                                9240 non-null   object
 6   Converted                                  9240 non-null   int64
 7   TotalVisits                                9103 non-null   float64
 8   Total Time Spent on Website                9240 non-null   int64
 9   Page Views Per Visit                       9103 non-null   float64
 10  Last Activity                              9137 non-null   object
 11  Country                                    6779 non-null   object
 12  Specialization                             7802 non-null   object
 13  How did you hear about X Education         7033 non-null   object
 14  What is your current occupation            6550 non-null   object
 15  What matters most to you in choosing a course  6531 non-null   object
 16  Search                                     9240 non-null   object
 17  Magazine                                   9240 non-null   object
 18  Newspaper Article                          9240 non-null   object
 19  X Education Forums                         9240 non-null   object
...
 35  A free copy of Mastering The Interview     9240 non-null   object
 36  Last Notable Activity                      9240 non-null   object
dtypes: float64(4), int64(3), object(30)
```

Step 2: EDA and Cleansing Data - Converted variable analysis

The target variable in your dataset, which represents whether a lead has been successfully converted or not with 0 for "not converted" and 1 for "converted successfully," would be considered a categorical variable.

HOW TO DO

1/ Check unique value of target column
2/ Change Converted into category
3/ Show the distribution of unique value in Converted by countplot
4/ Count numbers of each unique values for the target column



RESULT:

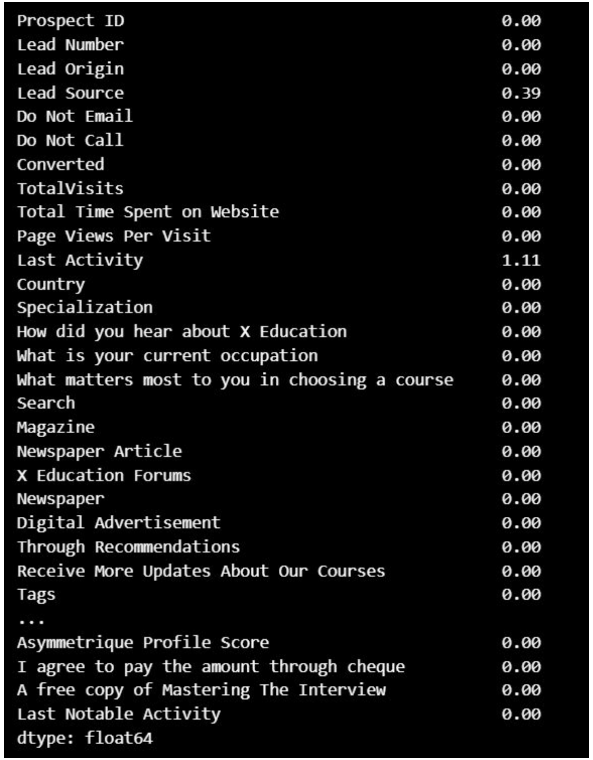
- Target var: should be category.
- There are 5679 customers with converted unsuccessfully with value 0 in target column and 3561 are converted successfully with value 1.
- % customer can be converted successfully in data frame: 38.5% in this dataset

Step 2: EDA and Cleansing Data - Missing value treatment

HOW TO DO

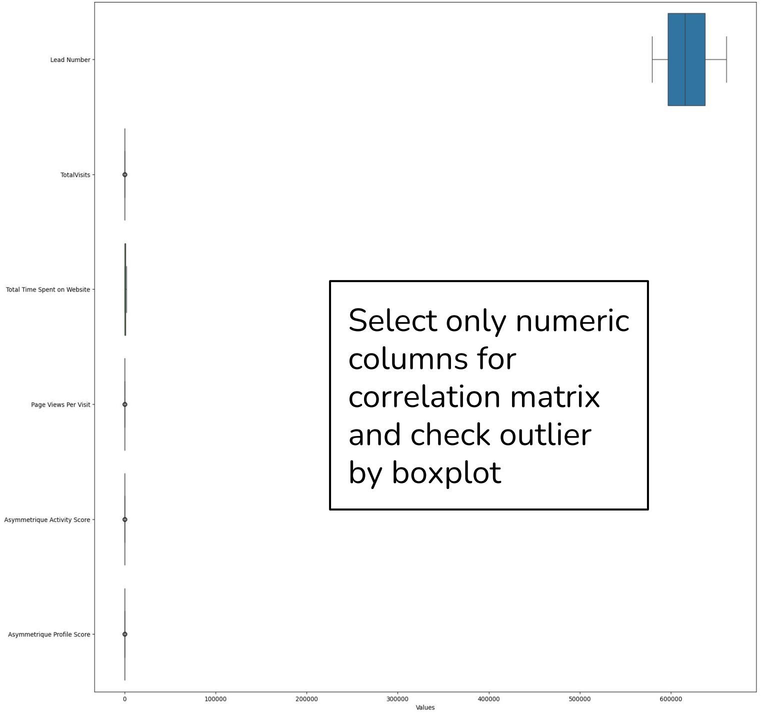
1/ Finding missing data and check % missing value in each column.
2/ Drop cols having missing values.
3/ Replace missing values with the mean value.
4/ Check unique value of some cols before making decision of treating missing values.
5/ Replace missing values in specified columns with "Other" and “Select”.
6/ Check % missing value in each column.

RESULT: No missing value in data lead

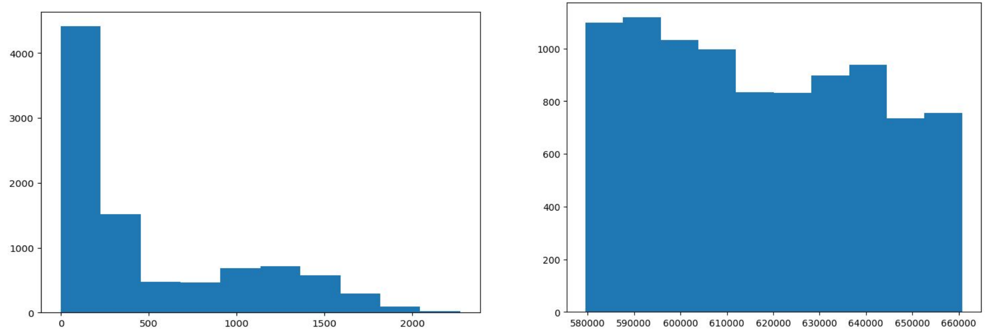


```
Prospect ID                                      0.00
Lead Number                                      0.00
Lead Origin                                      0.00
Lead Source                                      0.39
Do Not Email                                     0.00
Do Not Call                                      0.00
Converted                                        0.00
TotalVisits                                      0.00
Total Time Spent on Website                      0.00
Page Views Per Visit                             0.00
Last Activity                                    1.11
Country                                          0.00
Specialization                                   0.00
How did you hear about X Education               0.00
What is your current occupation                  0.00
What matters most to you in choosing a course    0.00
Search                                           0.00
Magazine                                         0.00
Newspaper Article                                0.00
X Education Forums                               0.00
Newspaper                                        0.00
Digital Advertisement                            0.00
Through Recommendations                          0.00
Receive More Updates About Our Courses           0.00
Tags                                             0.00
...
Asymmetrique Profile Score                       0.00
I agree to pay the amount through cheque         0.00
A free copy of Mastering The Interview           0.00
Last Notable Activity                            0.00
dtype: float64
```

Step 2: EDA and Cleansing Data - Check outliers



Select only numeric columns for correlation matrix and check outlier by boxplot

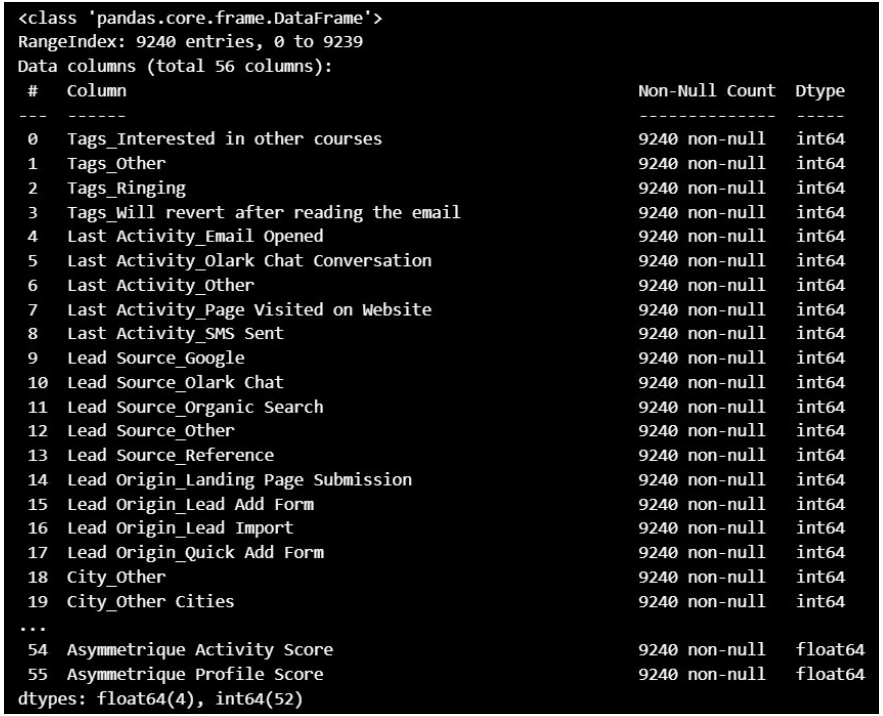


Hist for some num col having outliers

Step 3: Data preparation

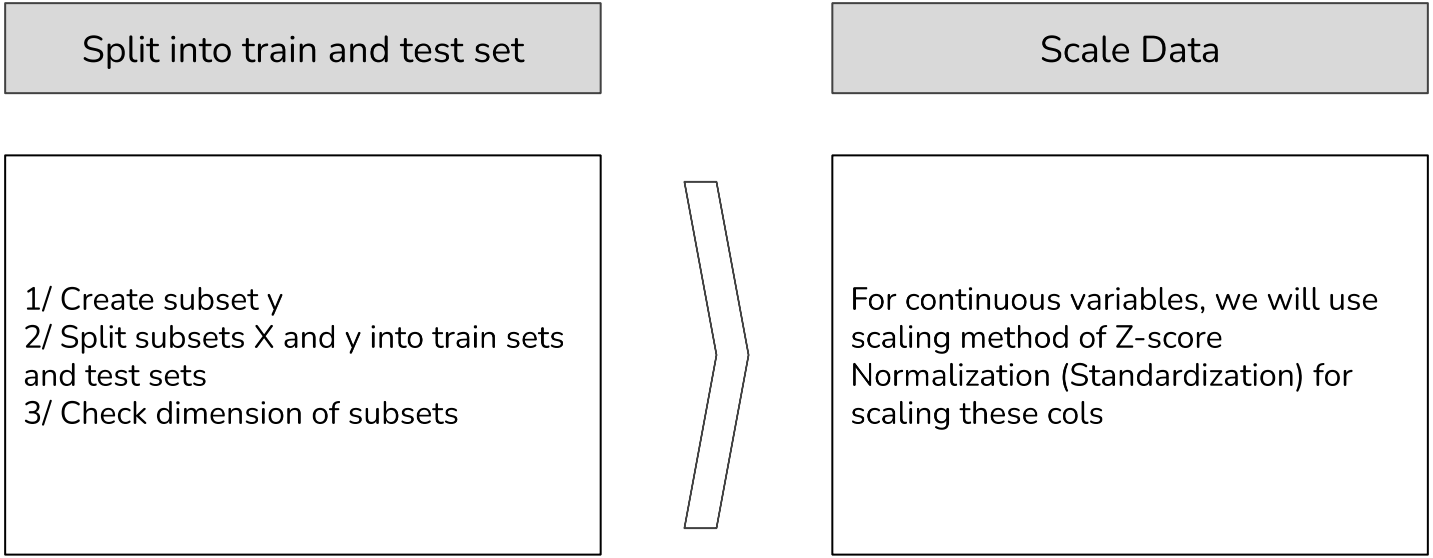
HOW TO DO

1/ Transform binary variables by one hot encoding(Yes/No >> 0/1)
2/ Transform remaining category variables into numeric by dummies



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 56 columns):
 #   Column                                       Non-Null Count  Dtype  
---  ------                                       --------------  -----  
 0   Tags_Interested in other courses             9240 non-null   int64  
 1   Tags_Other                                   9240 non-null   int64  
 2   Tags_Ringing                                 9240 non-null   int64  
 3   Tags_Will revert after reading the email     9240 non-null   int64  
 4   Last Activity_Email Opened                   9240 non-null   int64  
 5   Last Activity_Olark Chat Conversation        9240 non-null   int64  
 6   Last Activity_Other                          9240 non-null   int64  
 7   Last Activity_Page Visited on Website        9240 non-null   int64  
 8   Last Activity_SMS Sent                       9240 non-null   int64  
 9   Lead Source_Google                           9240 non-null   int64  
 10  Lead Source_Olark Chat                       9240 non-null   int64  
 11  Lead Source_Organic Search                   9240 non-null   int64  
 12  Lead Source_Other                            9240 non-null   int64  
 13  Lead Source_Reference                        9240 non-null   int64  
 14  Lead Origin_Landing Page Submission          9240 non-null   int64  
 15  Lead Origin_Lead Add Form                    9240 non-null   int64  
 16  Lead Origin_Lead Import                      9240 non-null   int64  
 17  Lead Origin_Quick Add Form                   9240 non-null   int64  
 18  City_Other                                   9240 non-null   int64  
 19  City_Other Cities                            9240 non-null   int64  
...
 54  Asymmetrique Activity Score                  9240 non-null   float64
 55  Asymmetrique Profile Score                   9240 non-null   float64
dtypes: float64(4), int64(52)
```

Step 4 - 5: Split into train - test set and Scale Data



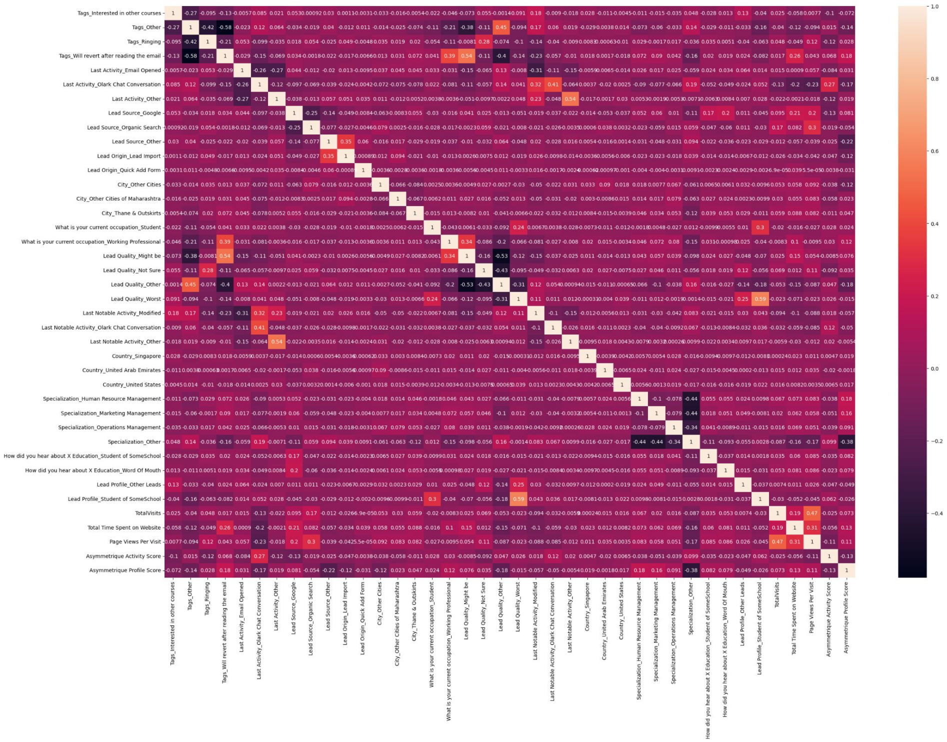
Split into train and test set

1/ Create subset y
2/ Split subsets X and y into train sets and test sets
3/ Check dimension of subsets

Scale Data

For continuous variables, we will use scaling method of Z-score Normalization (Standardization) for scaling these cols

Step 6: Correlation between each variables in lead



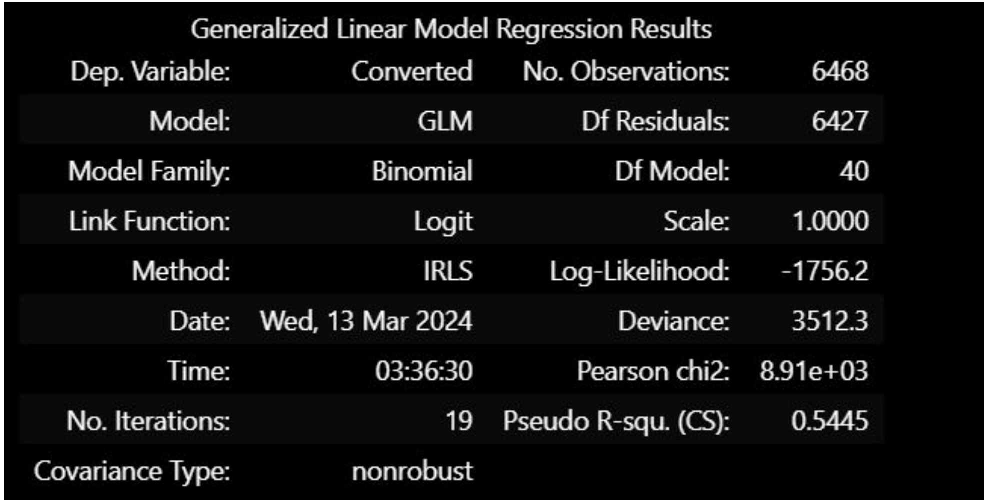
HOW TO DO

1/ Find the variables having high correlation with one another >= 70%
2/ Remove high correlation attributes

Step 7: Model Building

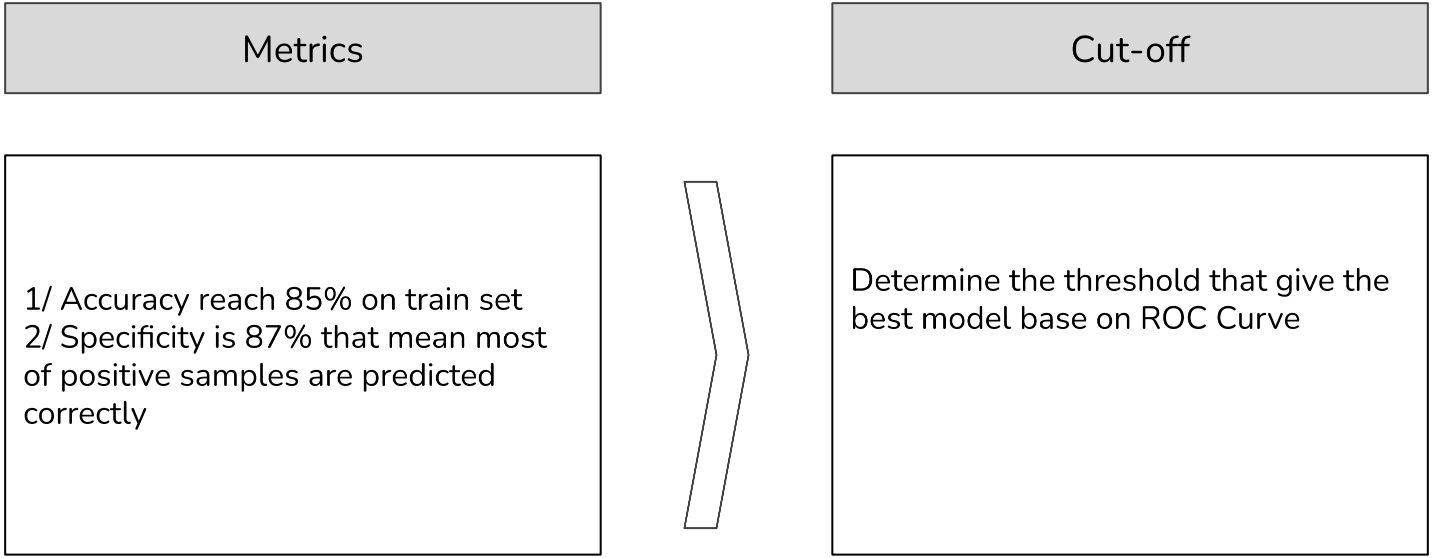
HOW TO DO

1/ Train Model on X_train
2/ First model
3/ Using REF to choose the best attributes



```
                 Generalized Linear Model Regression Results
   Dep. Variable:         Converted    No. Observations:       6468
           Model:               GLM        Df Residuals:       6427
    Model Family:          Binomial            Df Model:         40
   Link Function:             Logit               Scale:     1.0000
          Method:              IRLS      Log-Likelihood:    -1756.2
            Date:  Wed, 13 Mar 2024            Deviance:     3512.3
            Time:          03:36:30       Pearson chi2:    8.91e+03
  No. Iterations:                19  Pseudo R-squ. (CS):     0.5445
 Covariance Type:         nonrobust
```

Step 8: Data Evaluation



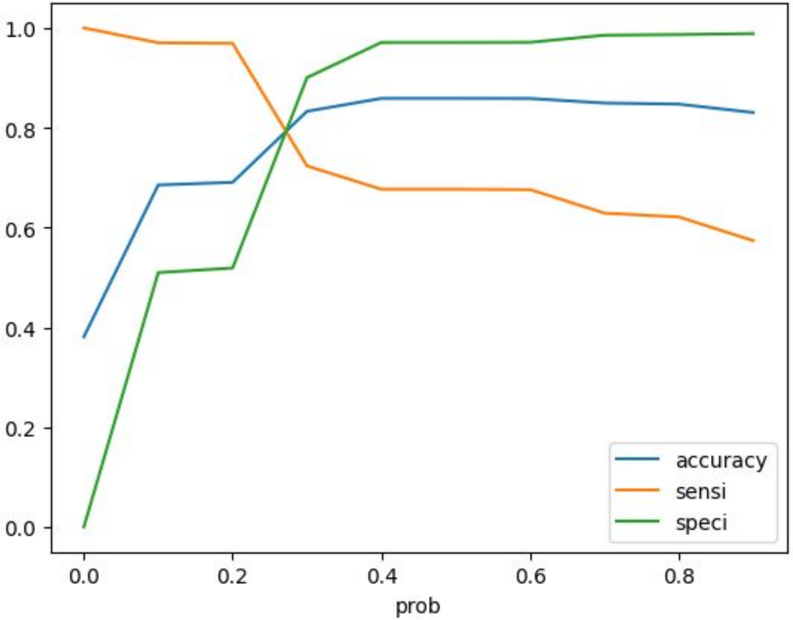
Metrics

1/ Accuracy reach 85% on train set
2/ Specificity is 87% that mean most of positive samples are predicted correctly

Cut-off

Determine the threshold that give the best model base on ROC Curve

Step 8: Data Evaluation



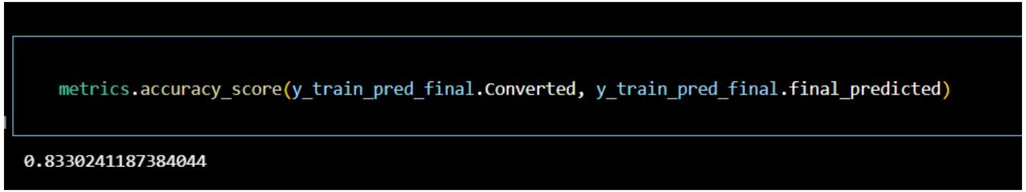
Cut-off

As shown in the illustration, the cut-off is approximate 0.3

Step 8: Data Evaluation

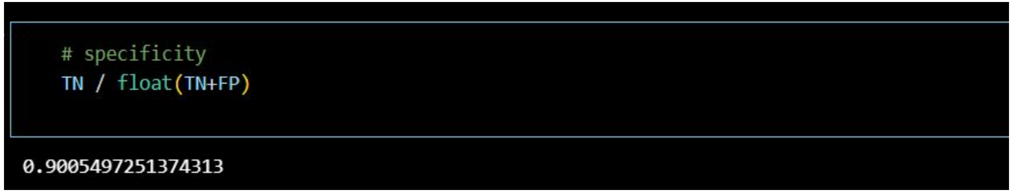
Test Set with Cut-off 0.3

1/ Accuracy reach 84% on test set
2/ Specificity is 90% that mean most of positive samples are predicted correctly



```
metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.final_predicted)

0.8330241187384044
```



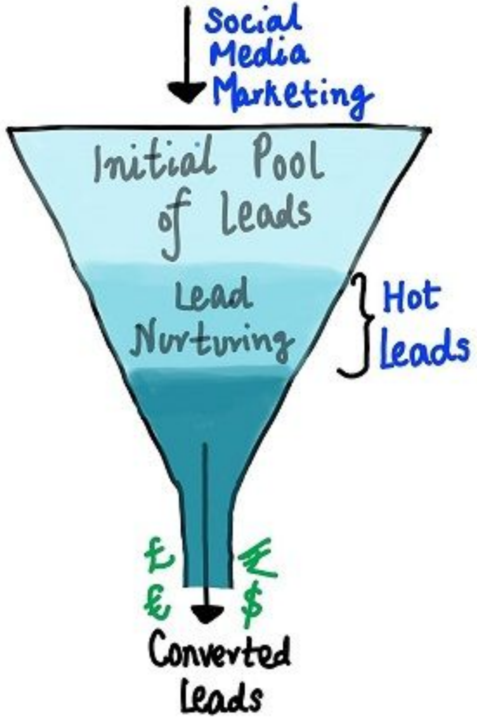
```
# specificity
TN / float(TN+FP)

0.9005497251374313
```

Business Case Study

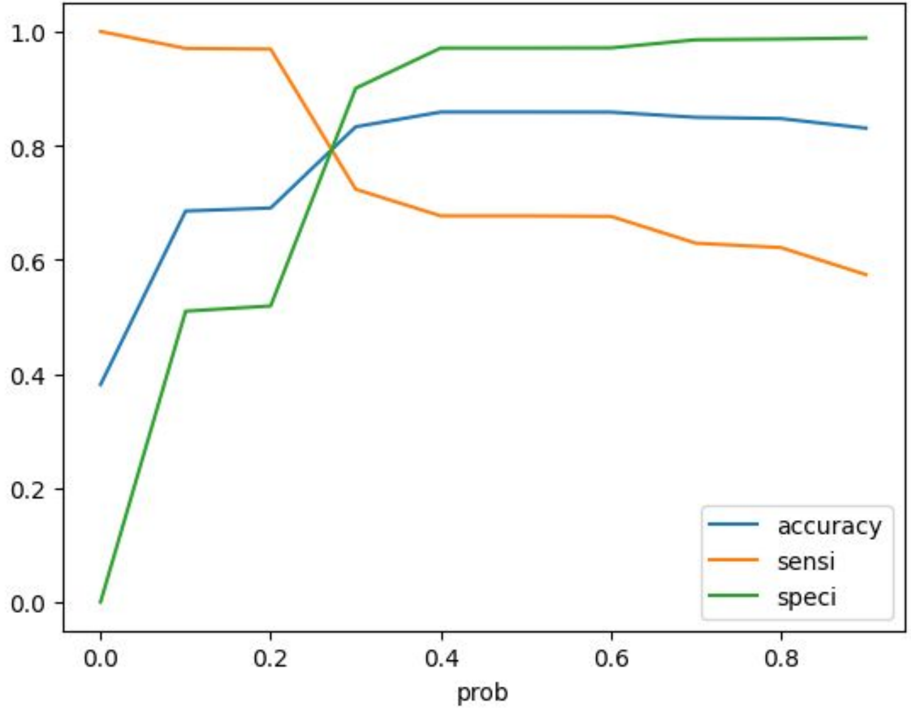
A user, becoming a lead when they have accessed to X Education's system: website, ads, surveys,...

From these customers' behaviors, the list of leads is built and the staffs will contact to them and persuade they become the real customer. This process could take much effort than it's necessary.



Business Case Study

Base on existing data, we try to build a machine learning model to handle this. The accuracy of model is about 85%, an considerate number to believe in practice. In fact, in some case, accuracy is not the most important metric we need to count. Specificity and Sensitivity are also crucial in special situation. The result of model return a probability for each input sample. Our mission is to choose an appropriate threshold for each condition.


prob
accuracy
sensi
speci