LEAD SCORING CASE STUDY

PRESENTED BY

- RUPINDER KAUR
- SEEMA R TIRKEY
- RUDRARUP GOSWAMI

PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- The company receives lot of leads but only 30% of them gets converted.
- X Education faced the task of elevating its lead conversion rate.
- The company wants to develop a model aimed at assigning lead scores, with the objective of identifying ‘Hot Leads’.
- The company can prioritize Hot Leads to improve the likelihood of conversion. The CEO set an ambitious target of achieving an 80% lead conversion rate.

BUSINESS OBJECTIVE

- X Education wants to identify most promising leads, known as Hot Leads
- Identify features, to increase the rate of conversion.
- Building of Model for future use.

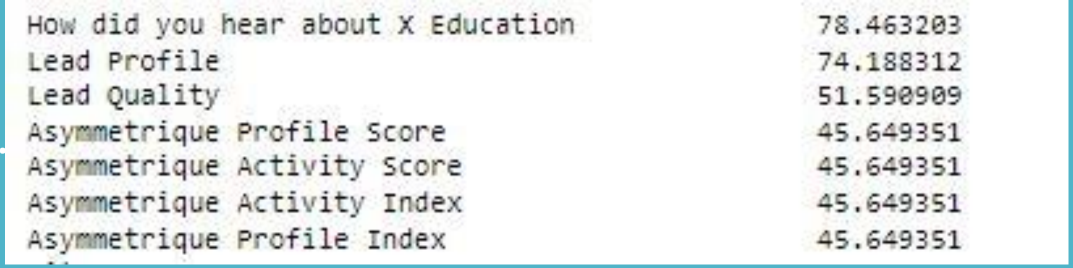
Solution Methodology

- Data Cleaning
  - Check and handle duplicate data
  - Check and remove columns with more than 40% null values
  - Imputation of values, if necessary
  - Removing unwanted columns
- EDA
  - Categorical and Numerical Analysis
  - Removing highly skewed and unique value columns
  - Comparing Variables with Target variable
  - Handling outliers

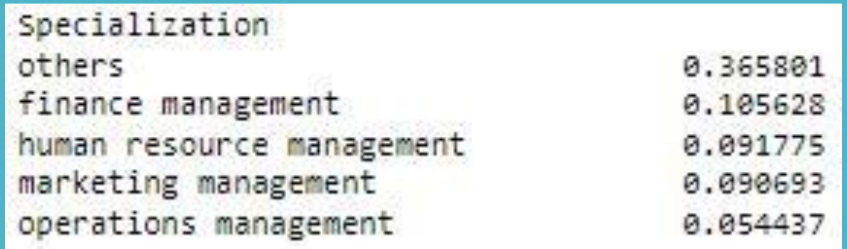
- Data Preparation
  - Dummy Variable
  - Splitting and Scaling
  - Feature selection with RFE
- Model Building using logistic regression
- Model Evaluation
  - Accuracy, Specificity and Sensitivity
  - Threshold Probability optimization with ROC Curve
- Model Creation on Test Data
- Conclusion and Recommendation

Data Cleaning

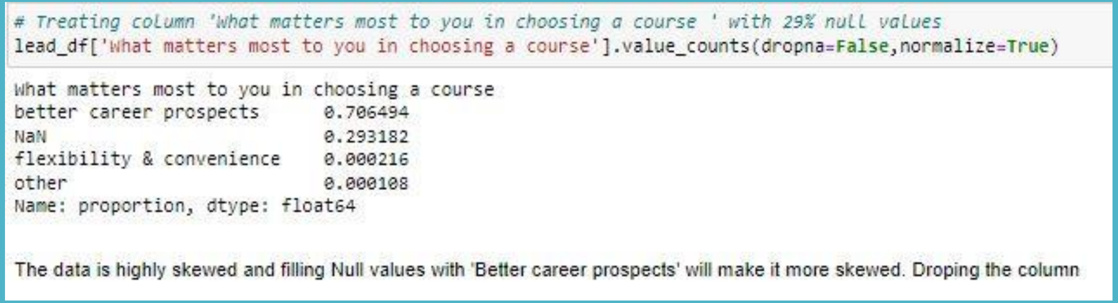
- Initial Data has 9240 rows and 37 Columns
- Dropped columns with over 40% null values.
- Imputing values in columns wherever required
- Dropping highly skewed columns



```
How did you hear about X Education              78.463203
Lead Profile                                    74.188312
Lead Quality                                    51.590909
Asymmetrique Profile Score                      45.649351
Asymmetrique Activity Score                     45.649351
Asymmetrique Activity Index                     45.649351
Asymmetrique Profile Index                      45.649351
```



```
Specialization
others                            0.365801
finance management                0.105628
human resource management         0.091775
marketing management              0.090693
operations management             0.054437
```



```
# Treating column 'What matters most to you in choosing a course ' with 29% null values
lead_df['What matters most to you in choosing a course'].value_counts(dropna=False,normalize=True)
```

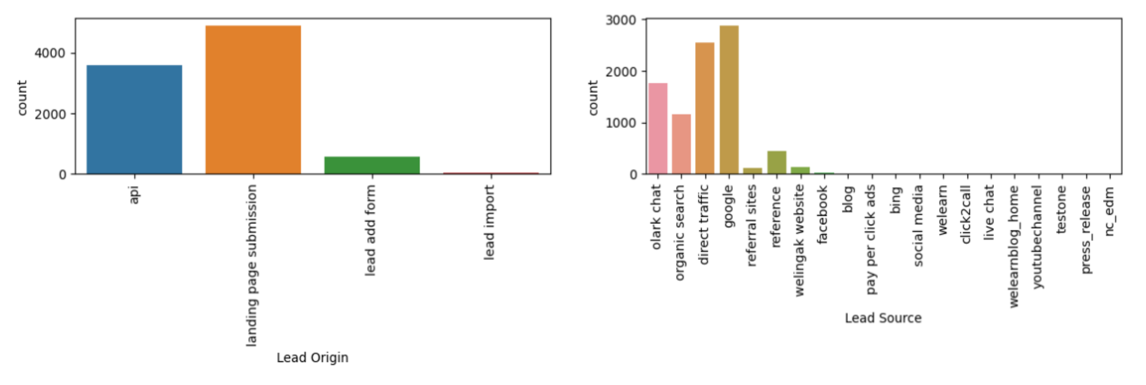
```
What matters most to you in choosing a course
better career prospects         0.706494
NaN                             0.293182
flexibility & convenience       0.000216
other                           0.000108
Name: proportion, dtype: float64
```

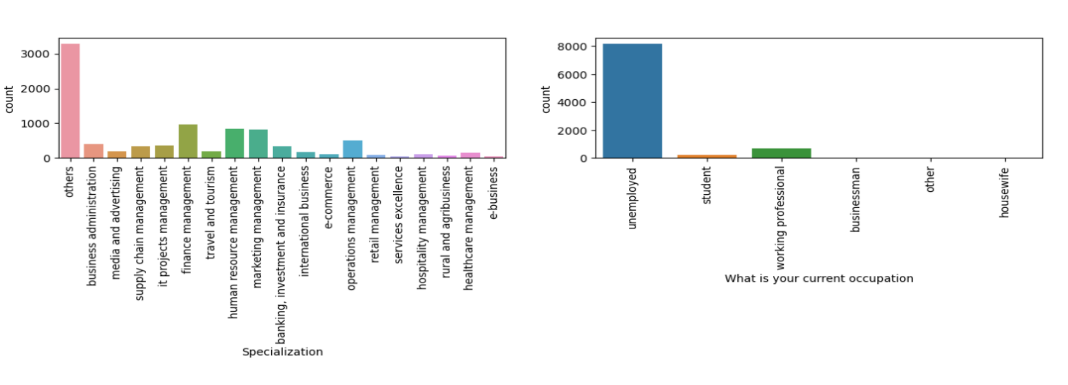
The data is highly skewed and filling Null values with 'Better career prospects' will make it more skewed. Droping the column

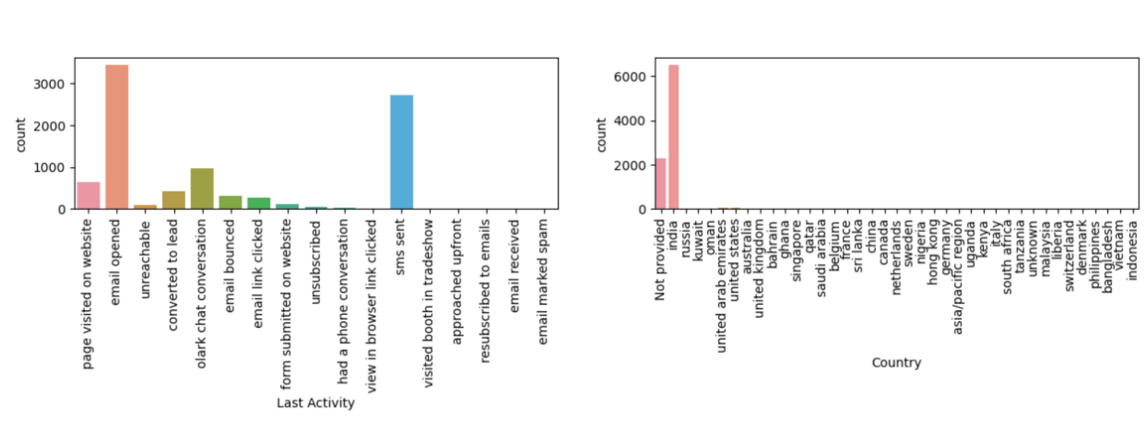
EDA

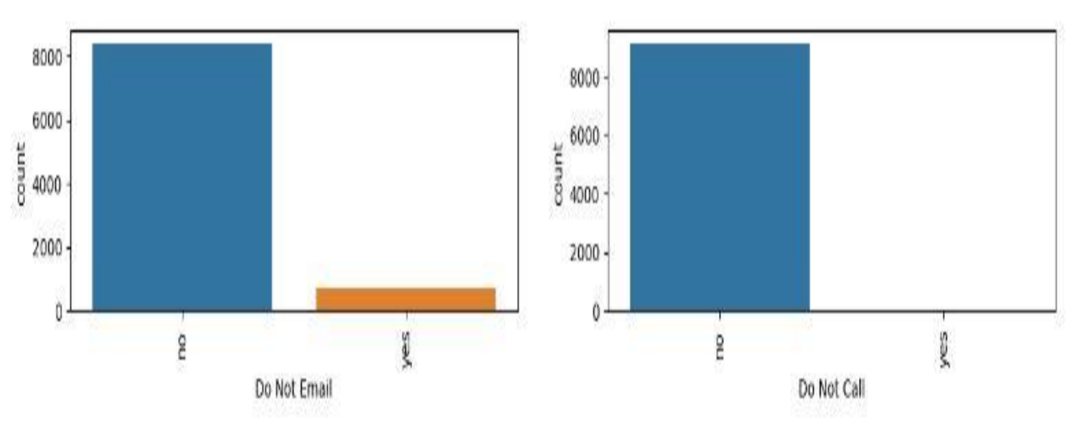
- Check data imbalance.
- Perform univariate and bivariate analysis for categorical and numerical variables.
- Comparing variables with target variable.

Univariate Category Analysis

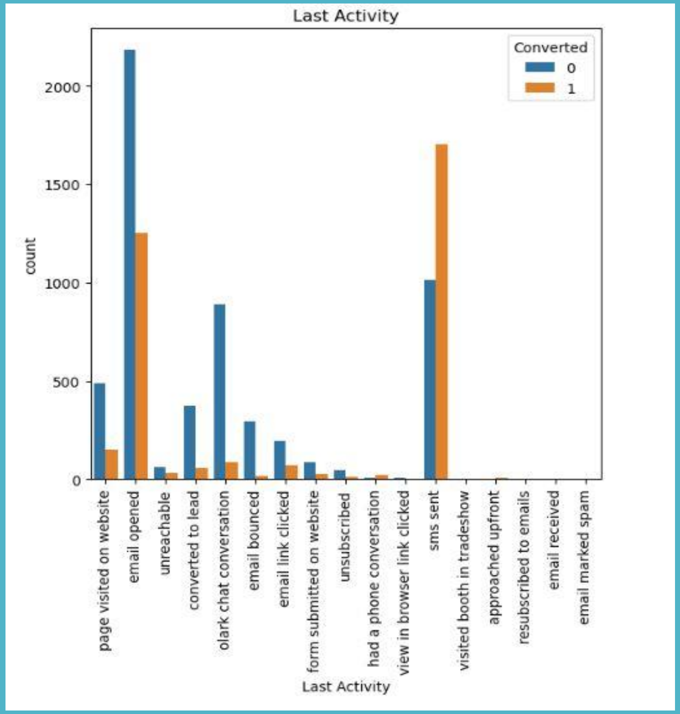


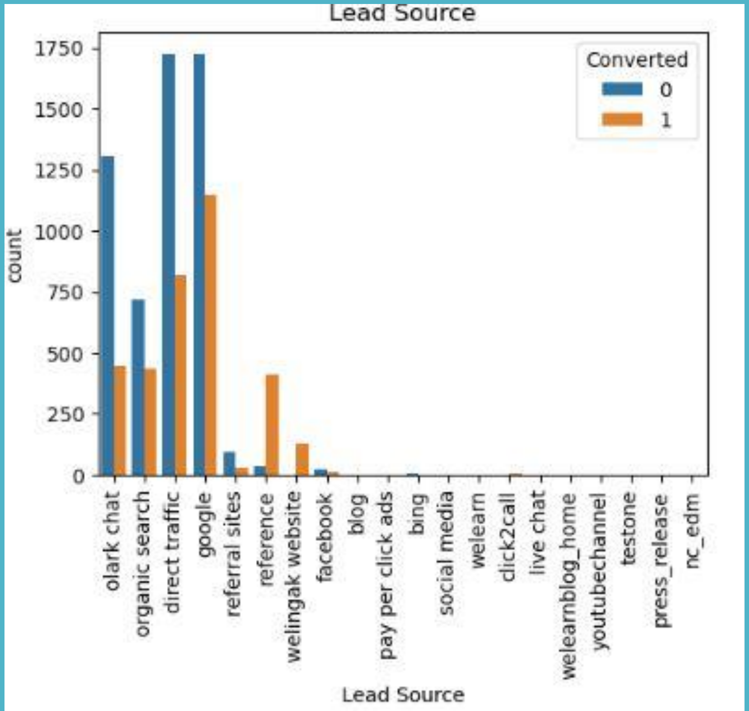


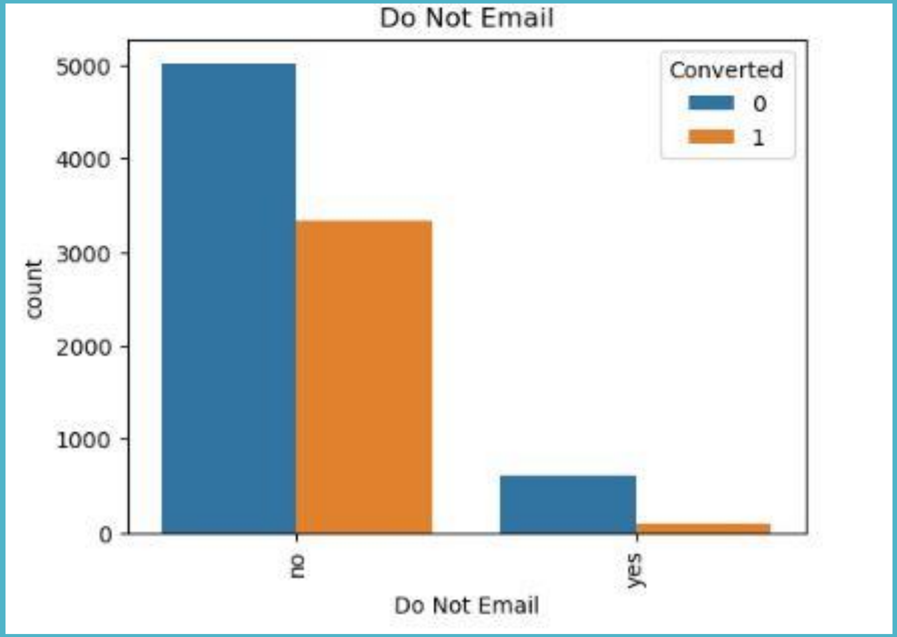




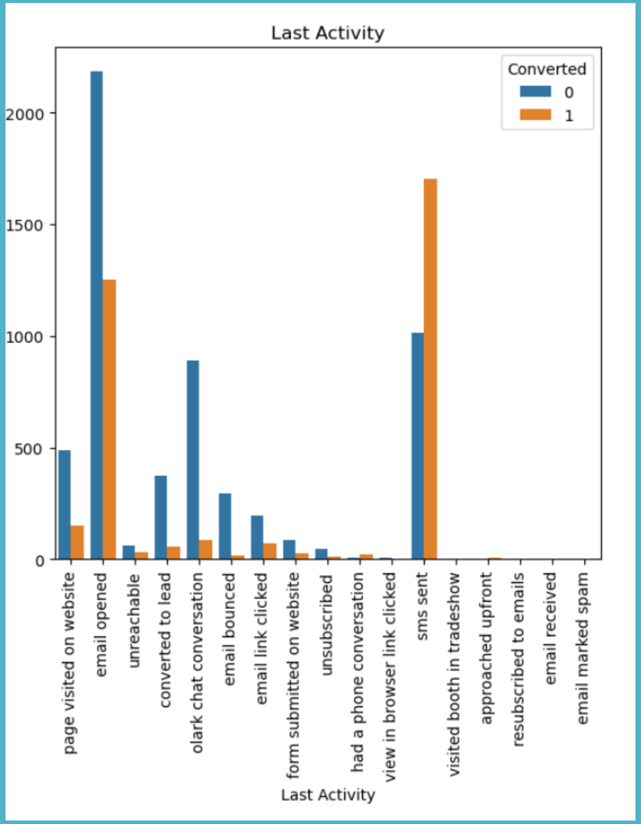
Bi-Variate Analysis with Target

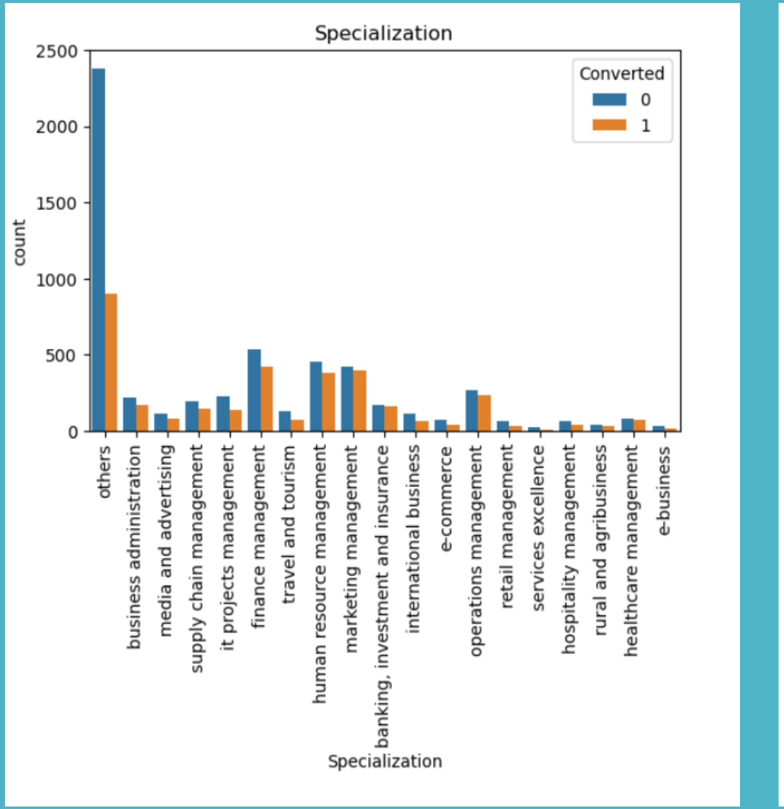


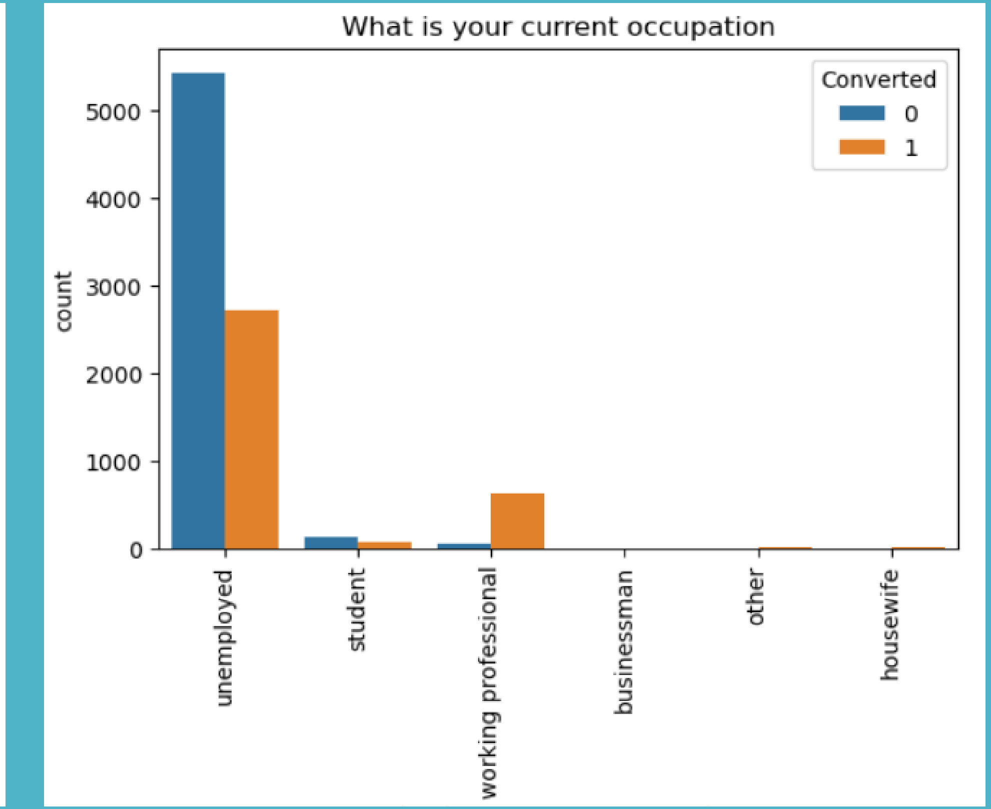




Bi-Variate Analysis with Target





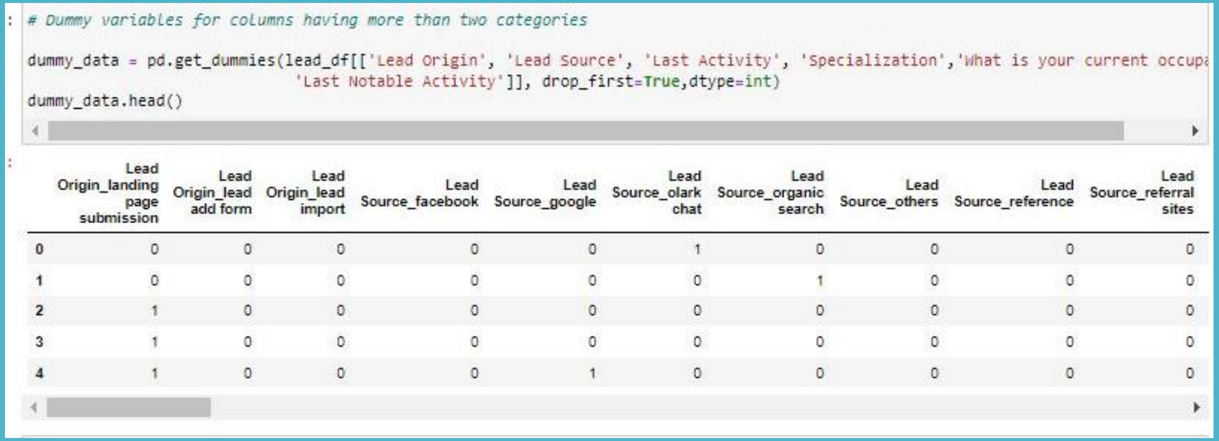


Data Preparation

- Created dummy features for categorical variables.

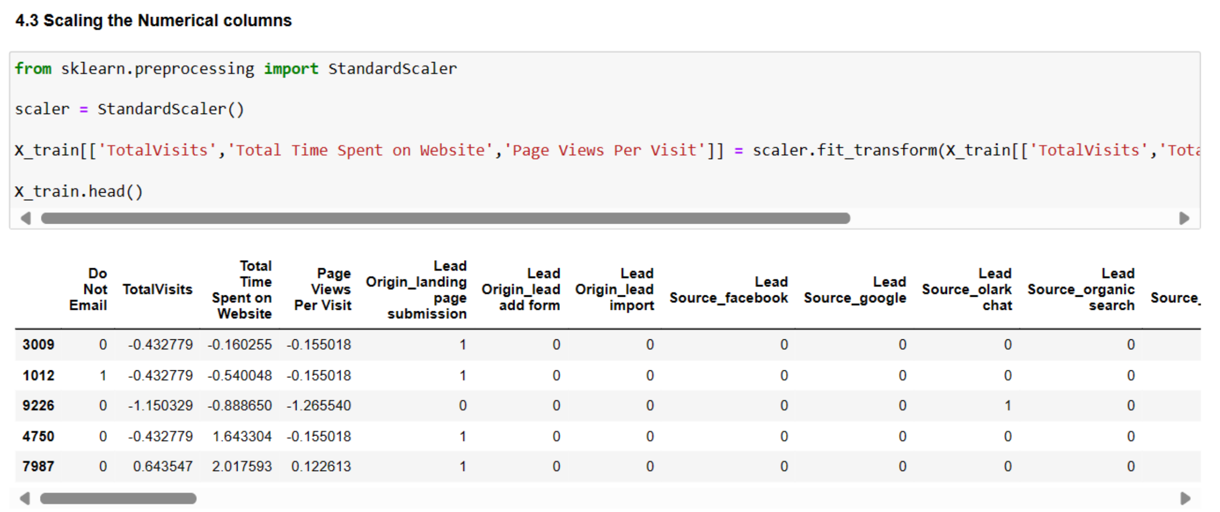
- Split the data into train and test sets.

- Scaling the Numerical Columns



```
# Dummy variables for columns having more than two categories

dummy_data = pd.get_dummies(lead_df[['Lead Origin', 'Lead Source', 'Last Activity', 'Specialization','What is your current occupa
                            'Last Notable Activity']], drop_first=True,dtype=int)
dummy_data.head()
```

	Lead Origin_landing page submission	Lead Origin_lead add form	Lead Origin_lead import	Lead Source_facebook	Lead Source_google	Lead Source_olark chat	Lead Source_organic search	Lead Source_others	Lead Source_reference	Lead Source_referral sites
0	0	0	0	0	0	1	0	0	0	0
1	0	0	0	0	0	0	1	0	0	0
2	1	0	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	0	0
4	1	0	0	0	1	0	0	0	0	0



4.3 Scaling the Numerical columns

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

X_train[['TotalVisits','Total Time Spent on Website','Page Views Per Visit']] = scaler.fit_transform(X_train[['TotalVisits','Tota

X_train.head()
```

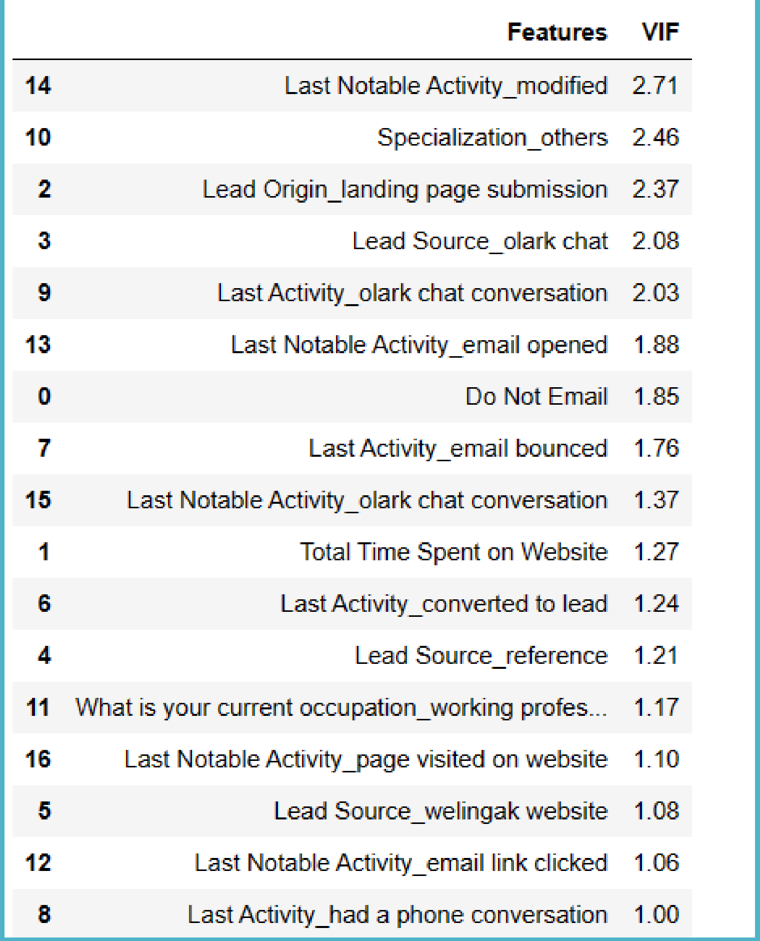
	Do Not Email	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_landing page submission	Lead Origin_lead add form	Lead Origin_lead import	Lead Source_facebook	Lead Source_google	Lead Source_olark chat	Lead Source_organic search	Source_
3009	0	-0.432779	-0.160255	-0.155018	1	0	0	0	0	0	0	
1012	1	-0.432779	-0.540048	-0.155018	1	0	0	0	0	0	0	
9226	0	-1.150329	-0.888650	-1.265540	0	0	0	0	0	1	0	
4750	0	-0.432779	1.643304	-0.155018	1	0	0	0	0	0	0	
7987	0	0.643547	2.017593	0.122613	1	0	0	0	0	0	0	

Feature Selection and Model Building

- Reduced the number of variables using recursive feature elimination (RFE). Selected 20 features with RFE and then removed columns with high p-value and VIF.
- Built four models before arriving at the final model.
- Ensured stability with p-values < 0.05 and no multicollinearity with VIF < 5.
- Final Model has 17 columns.

Final Features

These are the list of final features with their VIF values.



	Features	VIF
14	Last Notable Activity_modified	2.71
10	Specialization_others	2.46
2	Lead Origin_landing page submission	2.37
3	Lead Source_olark chat	2.08
9	Last Activity_olark chat conversation	2.03
13	Last Notable Activity_email opened	1.88
0	Do Not Email	1.85
7	Last Activity_email bounced	1.76
15	Last Notable Activity_olark chat conversation	1.37
1	Total Time Spent on Website	1.27
6	Last Activity_converted to lead	1.24
4	Lead Source_reference	1.21
11	What is your current occupation_working profes...	1.17
16	Last Notable Activity_page visited on website	1.10
5	Lead Source_welingak website	1.08
12	Last Notable Activity_email link clicked	1.06
8	Last Activity_had a phone conversation	1.00

ROC Curve

Area under ROC Curve is 0.89.

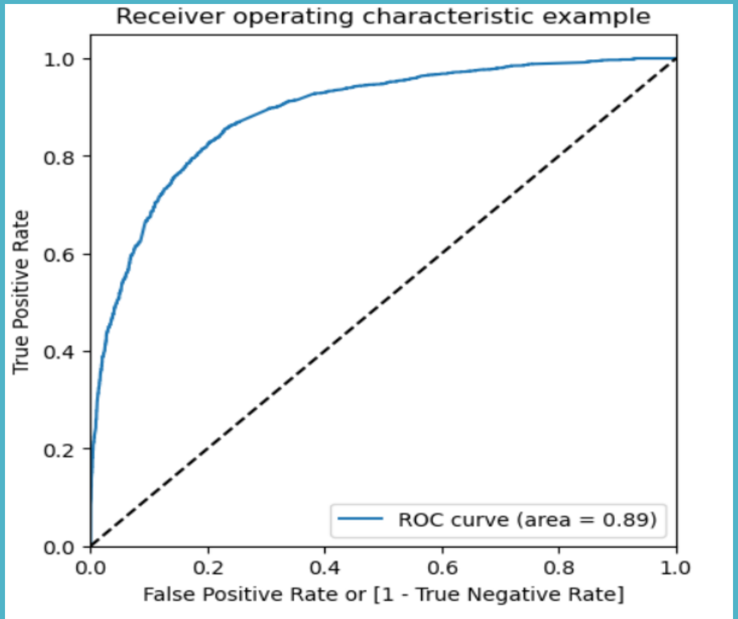
Optimal Probability: Cut off between Accuracy, Sensitivity, Specificity shows 0.4 is optimal probability.

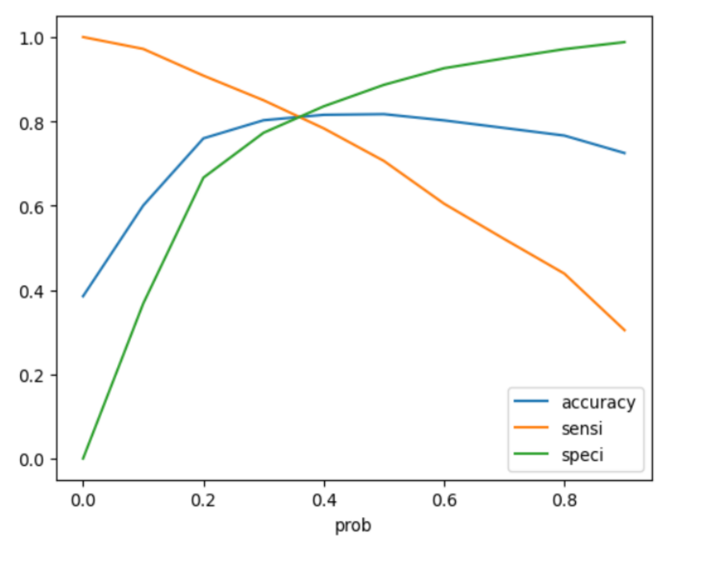
New values are:

Accuracy: 81.3

Sensitivity:79.9

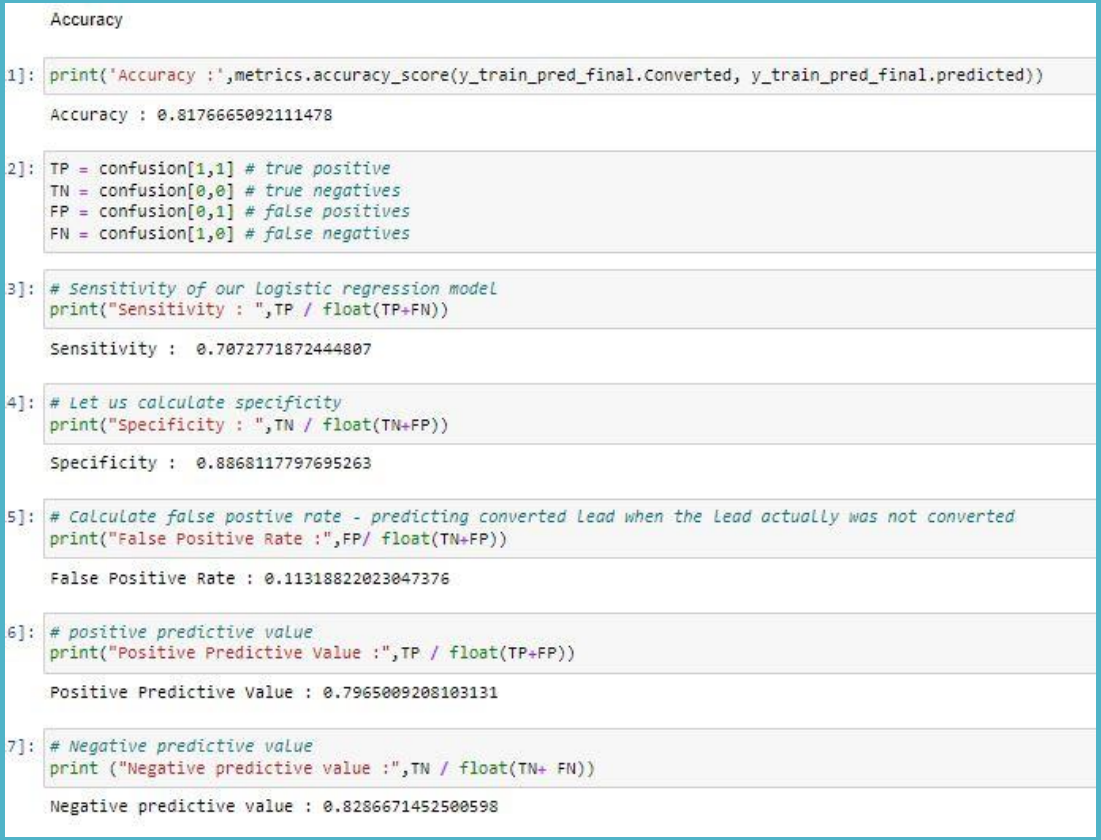
Specificity: 82.2





Model Evaluation

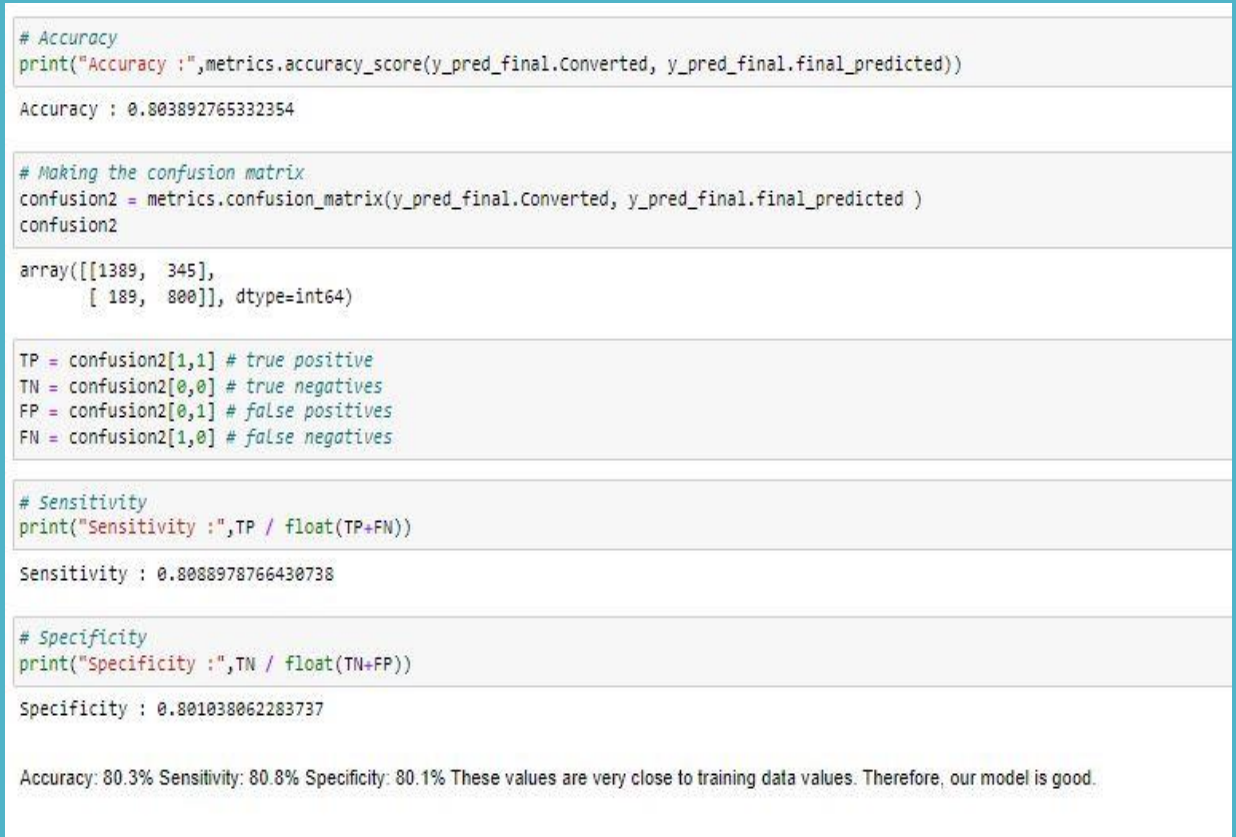
TEST MODEL



Accuracy

```
1]: print('Accuracy :',metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.predicted))

    Accuracy : 0.8176665092111478

2]: TP = confusion[1,1] # true positive
    TN = confusion[0,0] # true negatives
    FP = confusion[0,1] # false positives
    FN = confusion[1,0] # false negatives

3]: # Sensitivity of our logistic regression model
    print("Sensitivity : ",TP / float(TP+FN))

    Sensitivity :  0.7072771872444807

4]: # Let us calculate specificity
    print("Specificity : ",TN / float(TN+FP))

    Specificity :  0.8868117797695263

5]: # Calculate false postive rate - predicting converted lead when the lead actually was not converted
    print("False Positive Rate :",FP/ float(TN+FP))

    False Positive Rate : 0.11318822023047376

6]: # positive predictive value
    print("Positive Predictive Value :",TP / float(TP+FP))

    Positive Predictive Value : 0.7965009208103131

7]: # Negative predictive value
    print ("Negative predictive value :",TN / float(TN+ FN))

    Negative predictive value : 0.8286671452500598
```

TRAIN MODEL

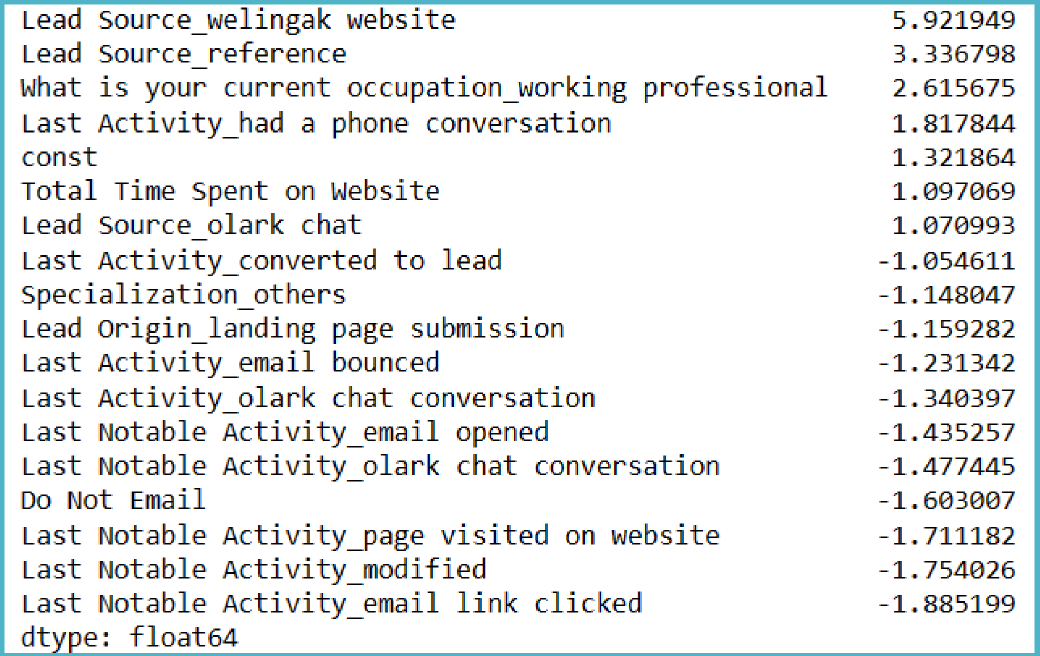


```
# Accuracy
print("Accuracy :",metrics.accuracy_score(y_pred_final.Converted, y_pred_final.final_predicted))

Accuracy : 0.803892765332354

# Making the confusion matrix
confusion2 = metrics.confusion_matrix(y_pred_final.Converted, y_pred_final.final_predicted )
confusion2

array([[1389,  345],
       [ 189,  800]], dtype=int64)

TP = confusion2[1,1] # true positive
TN = confusion2[0,0] # true negatives
FP = confusion2[0,1] # false positives
FN = confusion2[1,0] # false negatives

# Sensitivity
print("Sensitivity :",TP / float(TP+FN))

Sensitivity : 0.8088978766430738

# Specificity
print("Specificity :",TN / float(TN+FP))

Specificity : 0.801038062283737
```

Accuracy: 80.3% Sensitivity: 80.8% Specificity: 80.1% These values are very close to training data values. Therefore, our model is good.

Final Features and Hot Leads

- Lead Score is assigned to each lead and leads with lead score more than 80, are called Hot Leads.
- ID of Hot Leads are provided to the Company. They can be contacted as they have high probability of converting. This will help to increase the conversion rate.
- The final features and their relevance is shown.



```
Lead Source_welingak website                          5.921949
Lead Source_reference                                 3.336798
What is your current occupation_working professional  2.615675
Last Activity_had a phone conversation                1.817844
const                                                 1.321864
Total Time Spent on Website                           1.097069
Lead Source_olark chat                                1.070993
Last Activity_converted to lead                      -1.054611
Specialization_others                                -1.148047
Lead Origin_landing page submission                  -1.159282
Last Activity_email bounced                          -1.231342
Last Activity_olark chat conversation                -1.340397
Last Notable Activity_email opened                   -1.435257
Last Notable Activity_olark chat conversation        -1.477445
Do Not Email                                         -1.603007
Last Notable Activity_page visited on website        -1.711182
Last Notable Activity_modified                       -1.754026
Last Notable Activity_email link clicked             -1.885199
dtype: float64
```

Recommendations

Do’s

- Company should make call to leads coming from **"Lead Source_welingak website'** and **'reference'**, they are most likely to be converted.
- The company should contact **'working professionals'** as they are more likely to be converted.
- Leads who have 'Last Activity' as **'had a phone conversation 'are** more likely to be converted.
- The company should consider **'total time spent on the website'** as important feature and connect with leads spending more time on website. The company can make the website more engaging so that leads can spend more time on the website.

Recommendations

Don’ts

- The company should not make calls to leads whose 'Last Activity' was **'Olark Chat Conversation"** and **'email bounced'**. The company should not call to leads who have clicked **'Do Not Email'.**

- The company should not contact leads whose 'last Notable Activity is **'page visited on website', 'modified', 'email link clicked'**. They have very low chances of converting.

SUMMARY

Based on the analysis, it's recommended to focus more on Website advertising references that convert to leads, and target working professionals due to their higher conversion rate. To maximize lead conversion, prioritize hot leads identified by the model, implement personalized outreach, and increase contact attempts across various channels.

THANK YOU