Lead Scoring - Case Study

AUG, 2021

SAVAN AGARWAL (BATCH: C29)

Problem Statement

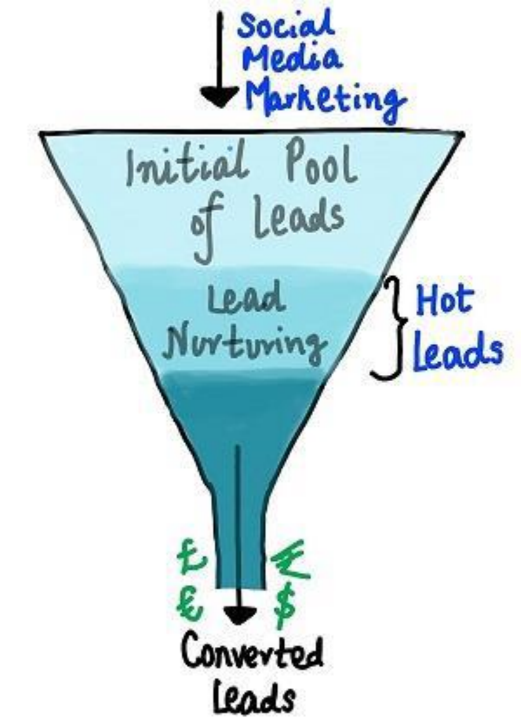
X Education sells online courses and needs a solution that can increase their customer lead conversions. The focus is selection of the most promising leads, i.e., the leads that are most likely to convert into paying customers.

The company needs a model wherein a lead score is assigned to each of the customer leads. The high score value would denote a high conversion chance of the prospect customer while lower score value means lower conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

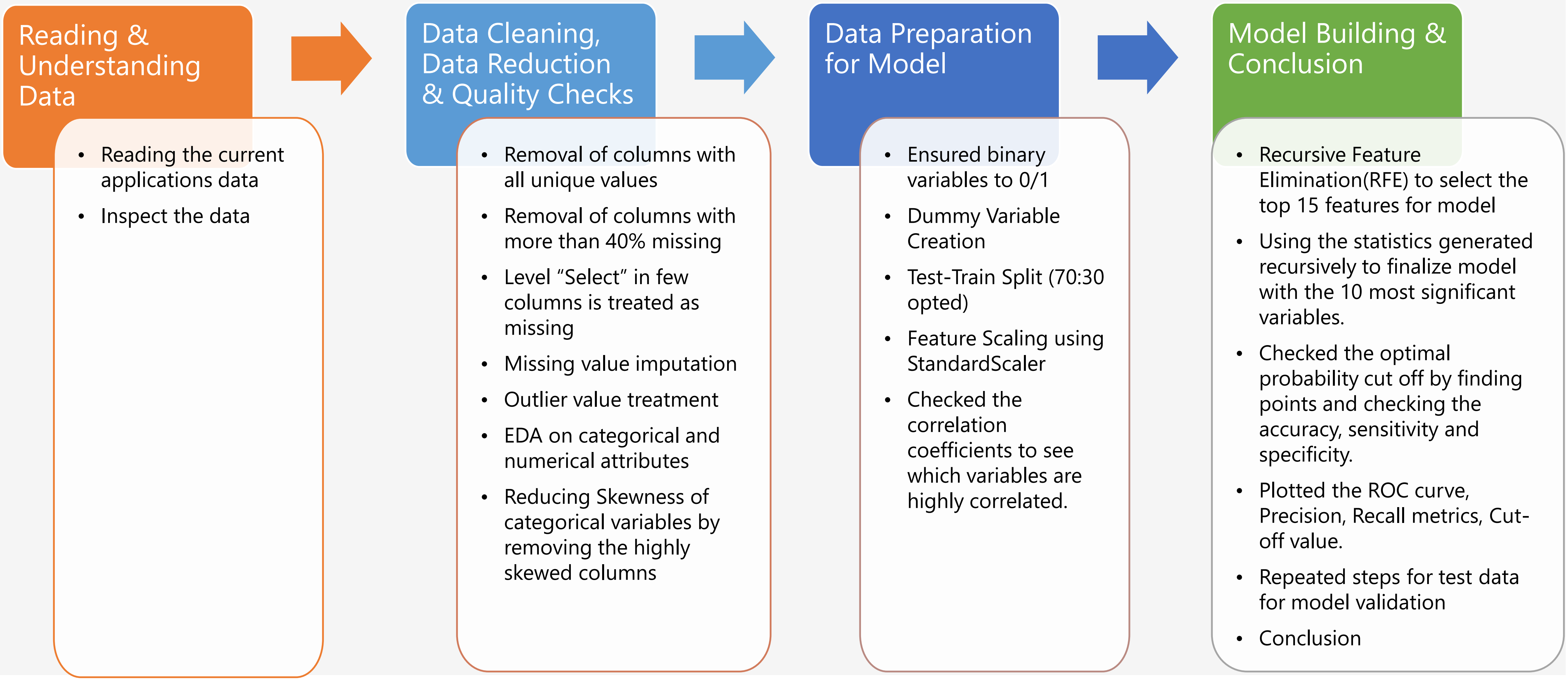
Objective

- To help the company in selecting the most potential leads, also known as 'Hot Leads' whose lead conversion rate is around 80%.
- To build a model wherein a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- Help the sales team to divert their focus on potential leads & avoid them from making useless phone calls.



Lead Conversion Process

High Level Approach for Overall Case Study



Reading & Understanding Data

- Reading the current applications data
- Inspect the data

Data Cleaning, Data Reduction & Quality Checks

- Removal of columns with all unique values
- Removal of columns with more than 40% missing
- Level “Select” in few columns is treated as missing
- Missing value imputation
- Outlier value treatment
- EDA on categorical and numerical attributes
- Reducing Skewness of categorical variables by removing the highly skewed columns

Data Preparation for Model

- Ensured binary variables to 0/1
- Dummy Variable Creation
- Test-Train Split (70:30 opted)
- Feature Scaling using StandardScaler
- Checked the correlation coefficients to see which variables are highly correlated.

Model Building & Conclusion

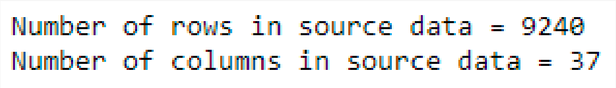
- Recursive Feature Elimination(RFE) to select the top 15 features for model
- Using the statistics generated recursively to finalize model with the 10 most significant variables.
- Checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity.
- Plotted the ROC curve, Precision, Recall metrics, Cut-off value.
- Repeated steps for test data for model validation
- Conclusion

Data Fetching, Inspecting & Preparation

Reading & Inspecting Data

Leads.csv

This data contains all the information of the leads both from original source and sales team along with a TARGET variable(Converted) with 1 and 0. 1 denotes “Converted” and 0 means “Not Converted”



```
Number of rows in source data = 9240
Number of columns in source data = 37
```

Data Cleaning

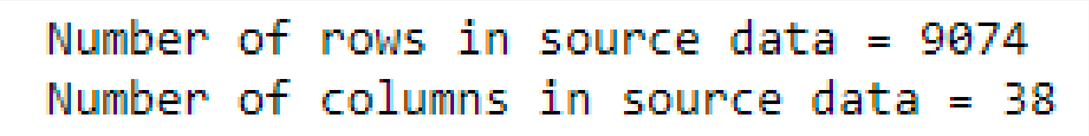
We found "Prospect ID“ & "Lead Number“ columns to have all unique values. These variables are not useful in the model. Hence we can drop them.

We can also drop >=40% missing values.

- Checking and removing the fields that have all unique values
- Checking removing the columns with >=40% of Missing values
- Few columns with a level “Select” means that customer did not fill info and is treated as missing
- Treatment of Missing values
  - Imputation with mode value on categorical variables
  - Removal of rows for <1.5% missing values in numeric columns
- Treatment of Outlier values
  - Imputation with 5th Percentile value for data in till 5th Percentile
  - Imputation with 95th Percentile value for data in above 95th Percentile

Data Reduction – Removing columns not used in Model

- Identifying the columns with high number of categories but less percentage of data in them. This data is classified as “Other”.
- Identifying and dropping columns with only 1 level. Not a value-add to model.
- Identifying and dropping columns which are Highly Skewed Categorical Variables
- Numerical column analysis
- Checking the imbalance in the target variable. Calculated the conversion rate. Conversion Rate is 37.86 %
- Analyzed the numeric attributes with respect to target variable “Converted”
- Inference: From the chart above, the leads spending more time on website are more likely to convert, thus website should be made more engaging to increase conversion rate.
- Removing "Sales team generated" variables (To avoid overfitting)
- Percentage of rows retained in data cleaning process = 98.2%

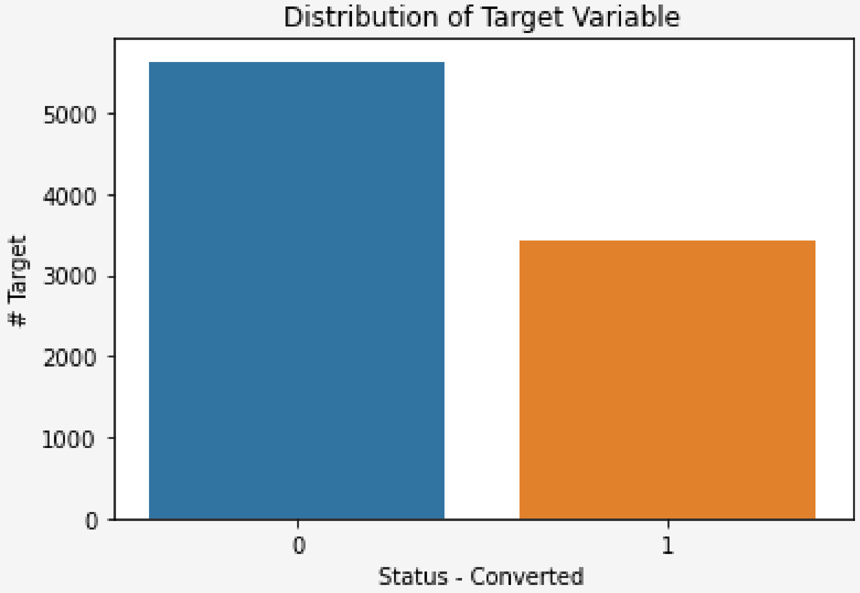


```
Number of rows in source data = 9074
Number of columns in source data = 38
```

Data Analysis

Target variable analysis

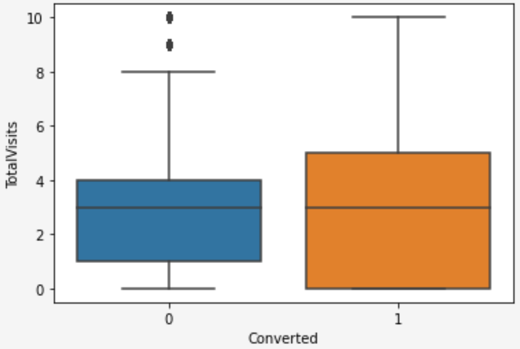
- We have total 9074 entries of unique customers and we needs to identify out of these which have the highest probability of getting converted.
- Decision Criteria:
  - Potential Leads can be bifurcated on the basis of Leads Score (which is probability of getting converted).
  - Out of 9074 entries, we see that the Conversion Rate is 37.86 % which means that around 37% of leads are converted and 73% of leads are not converted.
- Task at hand:
  - Identify solution so that the lead conversion rate could be increased.



Note: The missing value treatments are not executed for this EDA analysis to avoid overkill of analytics information. We generally do the treatment for Machine Learning model but can avoid in this EDA.

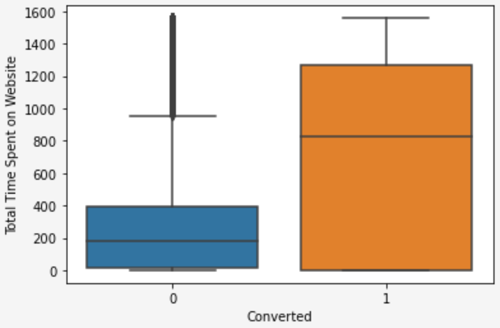
Numerical Column Analysis

- Visualizing 'TotalVisits' w.r.t target variable 'Converted'



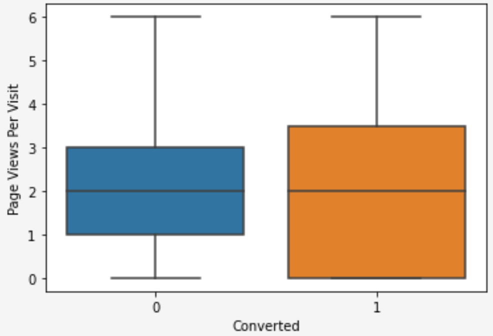
Inference: Median for converted and not converted leads is almost same. Nothing conclusive can be said on the basis of `TotalVisits`

- Visualizing 'Total Time Spent on Website' w.r.t target variable 'Converted'



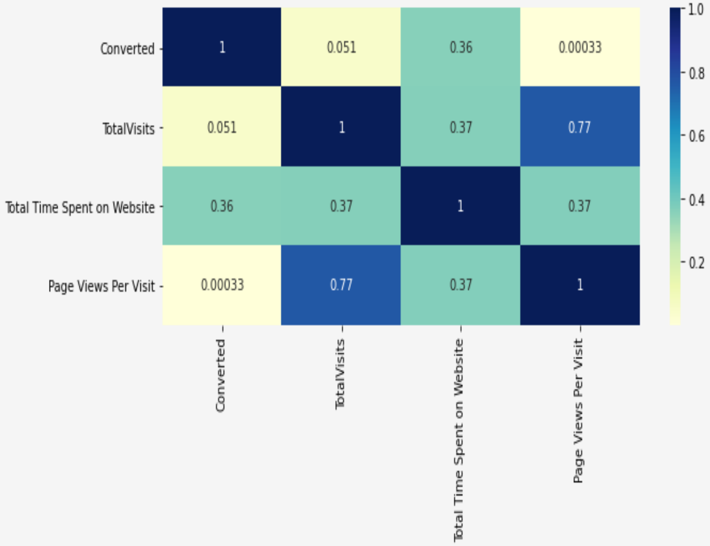
Inference: From the chart above, the leads spending more time on website are more likely to convert, thus website should be made more engaging to increase conversion rate.

- Visualizing 'Page Views Per Visit' w.r.t target variable 'Converted'



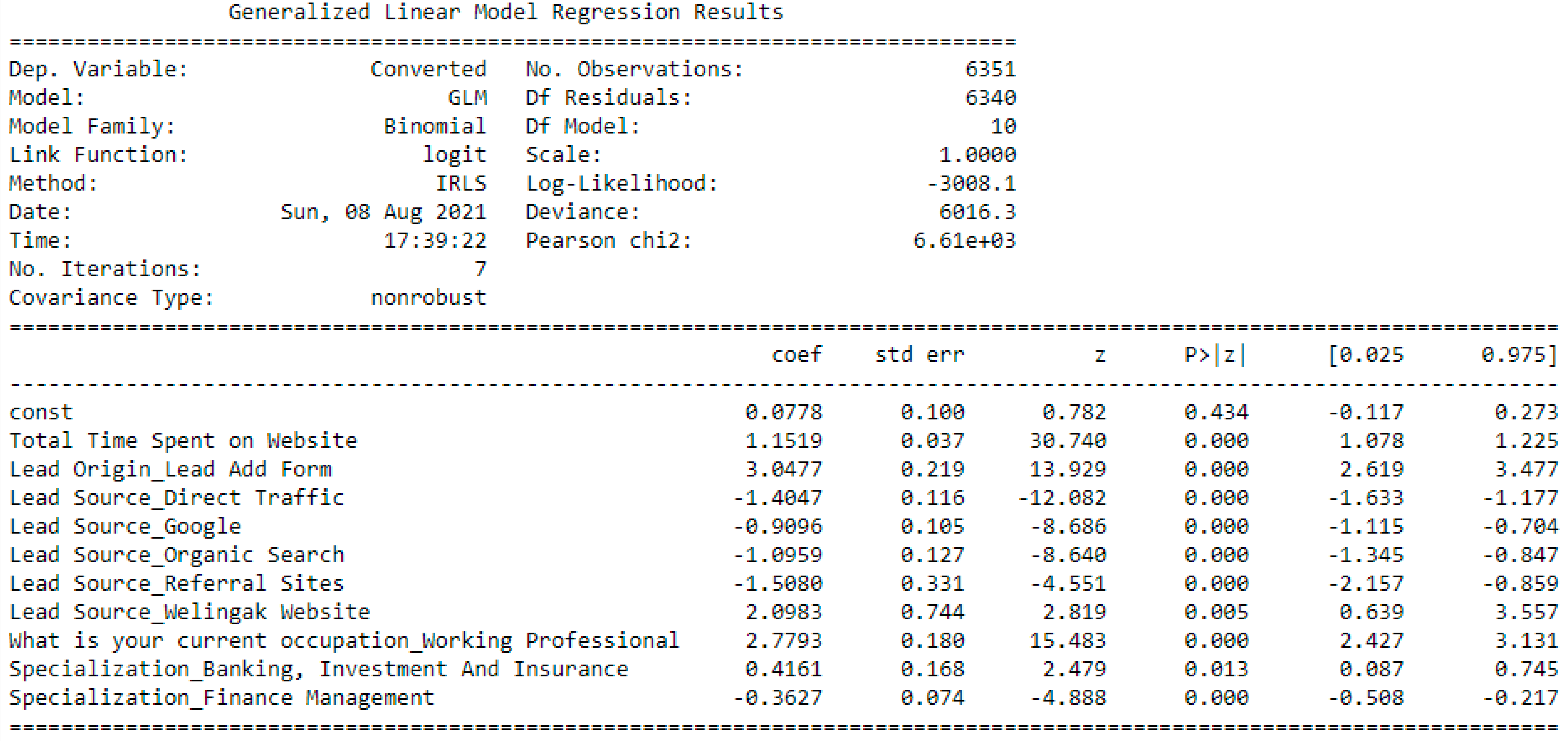
Inference: Median for converted and not converted leads is almost same. Nothing conclusive can be said on the basis of `TotalVisits`

Correlation between the numerical variables



Model Metrics & Conclusion

Factors Responsible in Driving Leads



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6351
Model:                            GLM   Df Residuals:                     6340
Model Family:                Binomial   Df Model:                           10
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -3008.1
Date:                Sun, 08 Aug 2021   Deviance:                       6016.3
Time:                        17:39:22   Pearson chi2:                 6.61e+03
No. Iterations:                     7
Covariance Type:            nonrobust
========================================================================================================================
                                                           coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------------------
const                                                    0.0778      0.100      0.782      0.434      -0.117       0.273
Total Time Spent on Website                              1.1519      0.037     30.740      0.000       1.078       1.225
Lead Origin_Lead Add Form                                3.0477      0.219     13.929      0.000       2.619       3.477
Lead Source_Direct Traffic                              -1.4047      0.116    -12.082      0.000      -1.633      -1.177
Lead Source_Google                                      -0.9096      0.105     -8.686      0.000      -1.115      -0.704
Lead Source_Organic Search                              -1.0959      0.127     -8.640      0.000      -1.345      -0.847
Lead Source_Referral Sites                              -1.5080      0.331     -4.551      0.000      -2.157      -0.859
Lead Source_Welingak Website                             2.0983      0.744      2.819      0.005       0.639       3.557
What is your current occupation_Working Professional     2.7793      0.180     15.483      0.000       2.427       3.131
Specialization_Banking, Investment And Insurance         0.4161      0.168      2.479      0.013       0.087       0.745
Specialization_Finance Management                       -0.3627      0.074     -4.888      0.000      -0.508      -0.217
========================================================================================================================
```

Note: The missing value treatments are not executed for this EDA analysis to avoid overkill of analytics information. We generally do the treatment for Machine Learning model but can avoid in this EDA.

Factors Responsible in Driving Leads

Below features are most important ones which are responsible for leads conversion

- Total Time Spent on Website
- Lead Origin_Lead Add Form
- Lead Source_Direct Traffic
- Lead Source_Google
- Lead Source_Organic Search
- Lead Source_Referral Sites
- Lead Source_Welingak Website
- What is your current occupation_Working Professional
- Specialization_Banking, Investment And Insurance
- Specialization_Finance Management
- 'Total Time Spent on Website'

Model Metrics

Below are model metrics on Train data

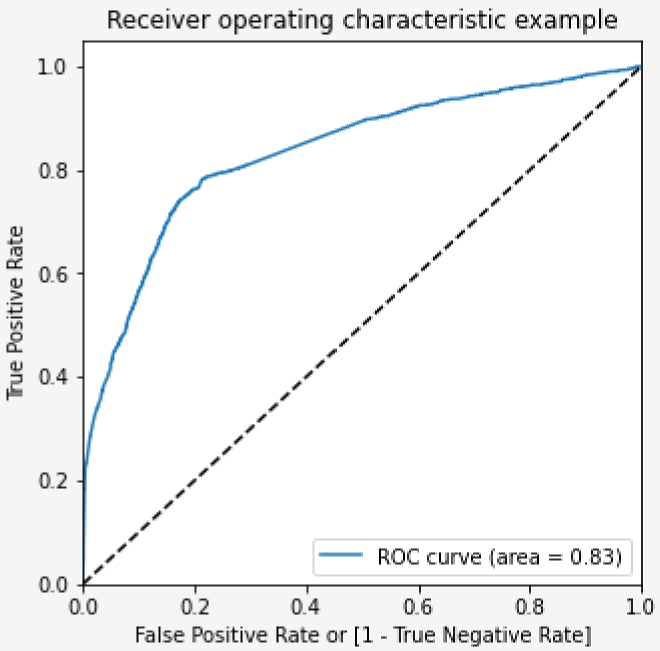
Train Data:

- Accuracy: 78.7%
- Sensitivity: 75.9%
- Specificity: 80.5%
- Precision: 71%
- Recall: 76%

Confusion Matrix

		Predicted	
		Not Converted	Converted
Actual	Not Converted	3144	761
	Converted	589	1857

ROC Curve



Model Metrics

Below are model metrics on Test data

Test Data:

- Accuracy: 79.4%
- Sensitivity: 75%
- Specificity: 82%
- Precision: 70%
- Recall: 75%

Confusion Matrix

		Predicted	
		Not Converted	Converted
Actual	Not Converted	1422	312
	Converted	248	741

Conclusion

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 79%, 75% and 82% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 76% Hence overall this model seems to be okay.

Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are:

- Lead Origin_Lead Add Form
- What is your current occupation_Working Professional
- Lead Source_Welingak Website