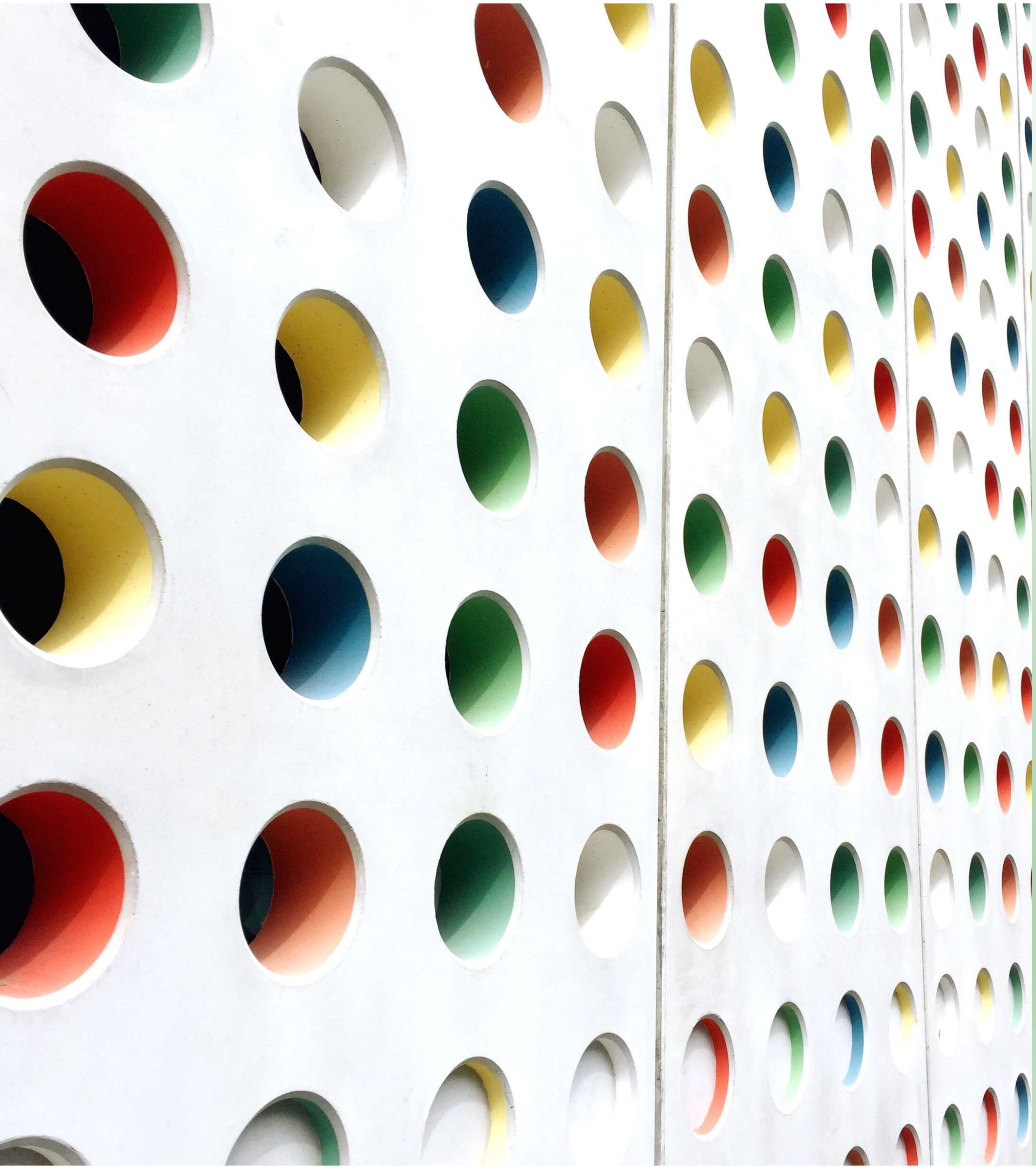


Lead Scoring Case Study

BY

BIJI KRISHNA , BAPPI BANIK, SANTANU BISWAS

Problem Statement

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.
- If they successfully identify this set of 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

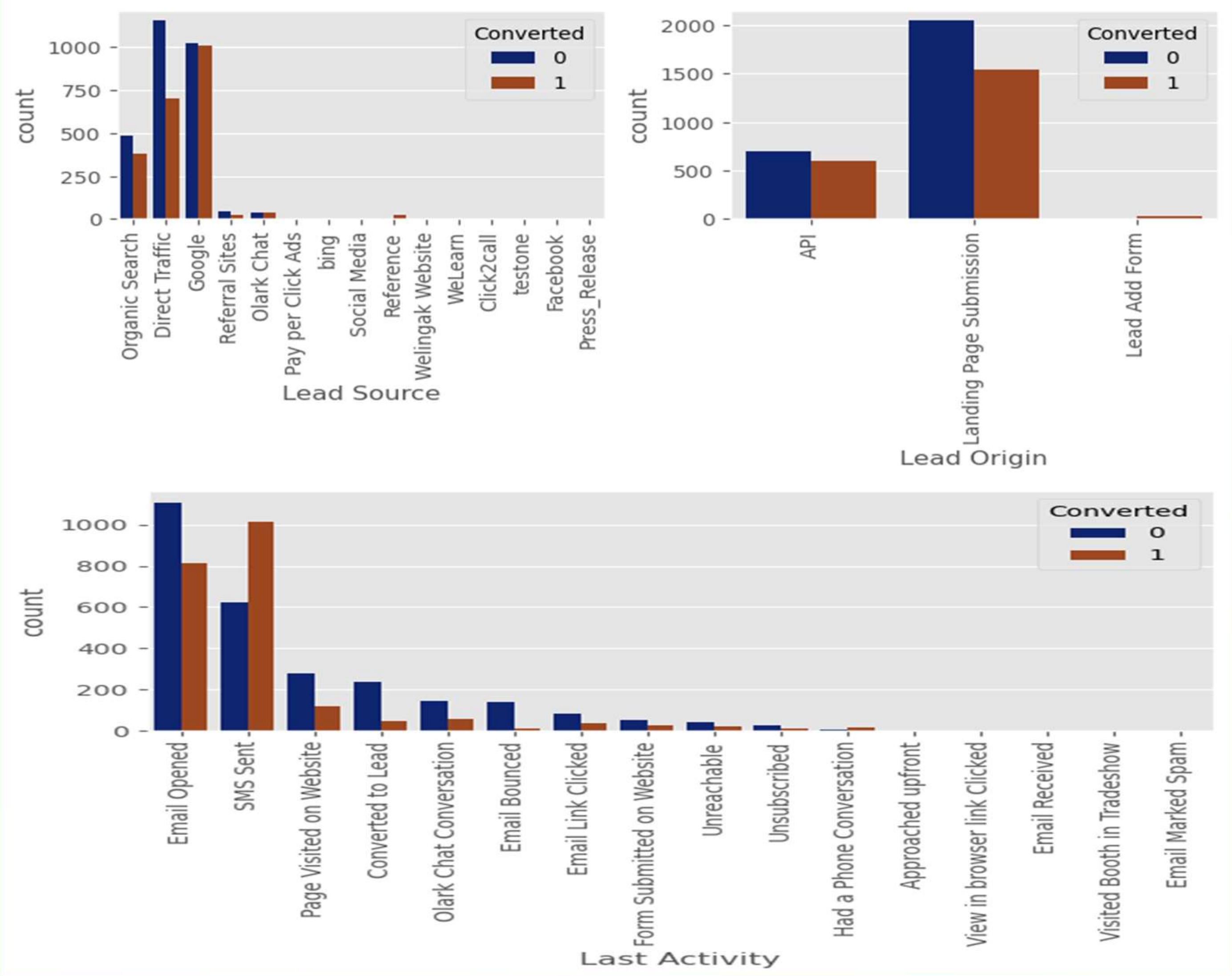
- X education n wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

Solution Methodology

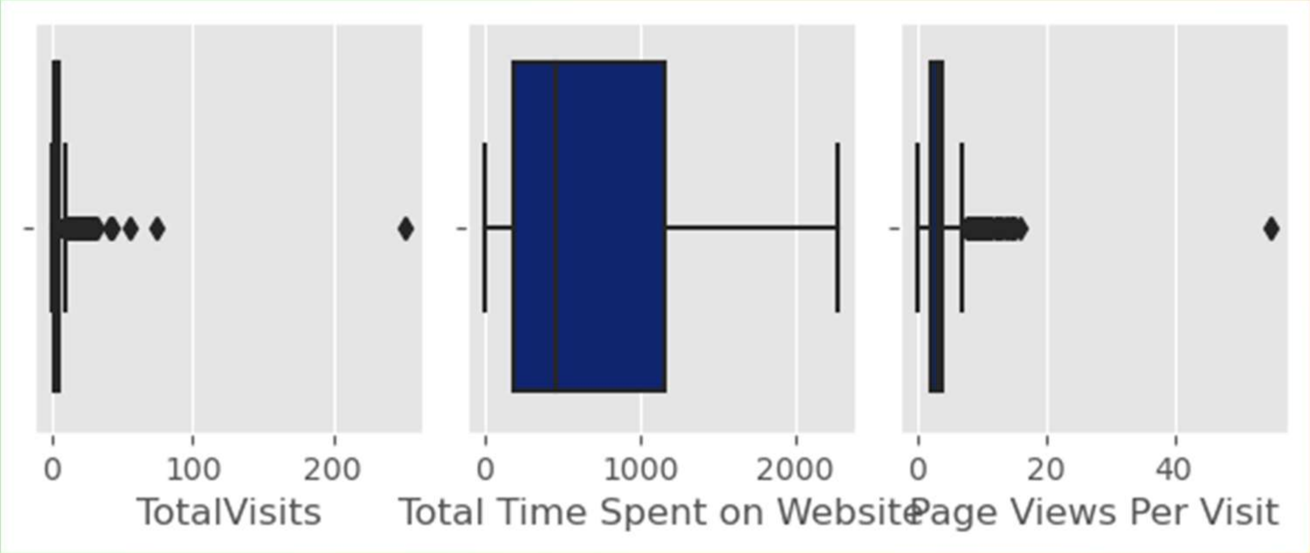
- Data cleaning and data manipulation.
  1. Check and handle duplicate data.
  2. Check and handle NA values and missing values.
  3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
  4. Imputation of the values, if necessary.
  5. Check and handle outliers in data.
- EDA
  1. Univariate data analysis: value count, distribution of variable etc.
  2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

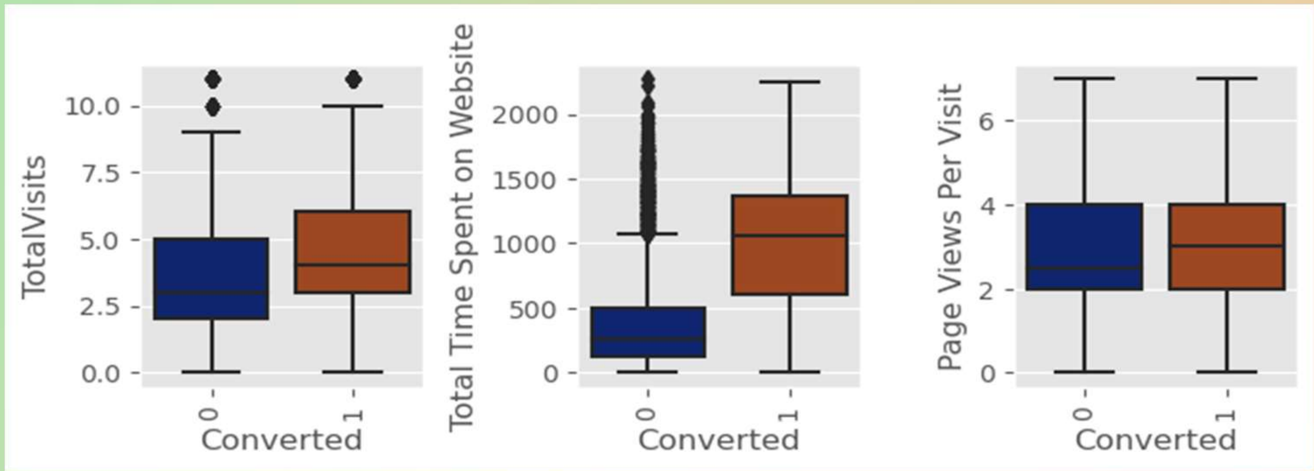
DataManipulation

- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”
- Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- Dropping the columns having more than 30% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.



Lead Source:-
Google Search and Direct Traffic has the highest lead counts followed by Olark Chat and Organic Search
The Conversion rate is higher for Reference and Welingak Website
Lead Origin:-
API and Landing page Submission has the maximum count of leads with low conversion rate
Lead Add Forms have low count of leads but high conversion rate
Lead Import and low count as well as conversion
To improve the overall conversion rate we need to increase the conversion rate of Google Search, Direct Traffic, API, and Landing page Submission.
Also we need to increase the lead count for Reference, Welingak Website and Lead Add Forms
•Count of leads in maximum for the Last Activity Email Opened
•Coversion is maximum for the Last Activity SMS Sent
We should target to call more people with Last Activity as Email Opened and increase their convesrion rate and also increase the count of users with last activity as SMS Sent



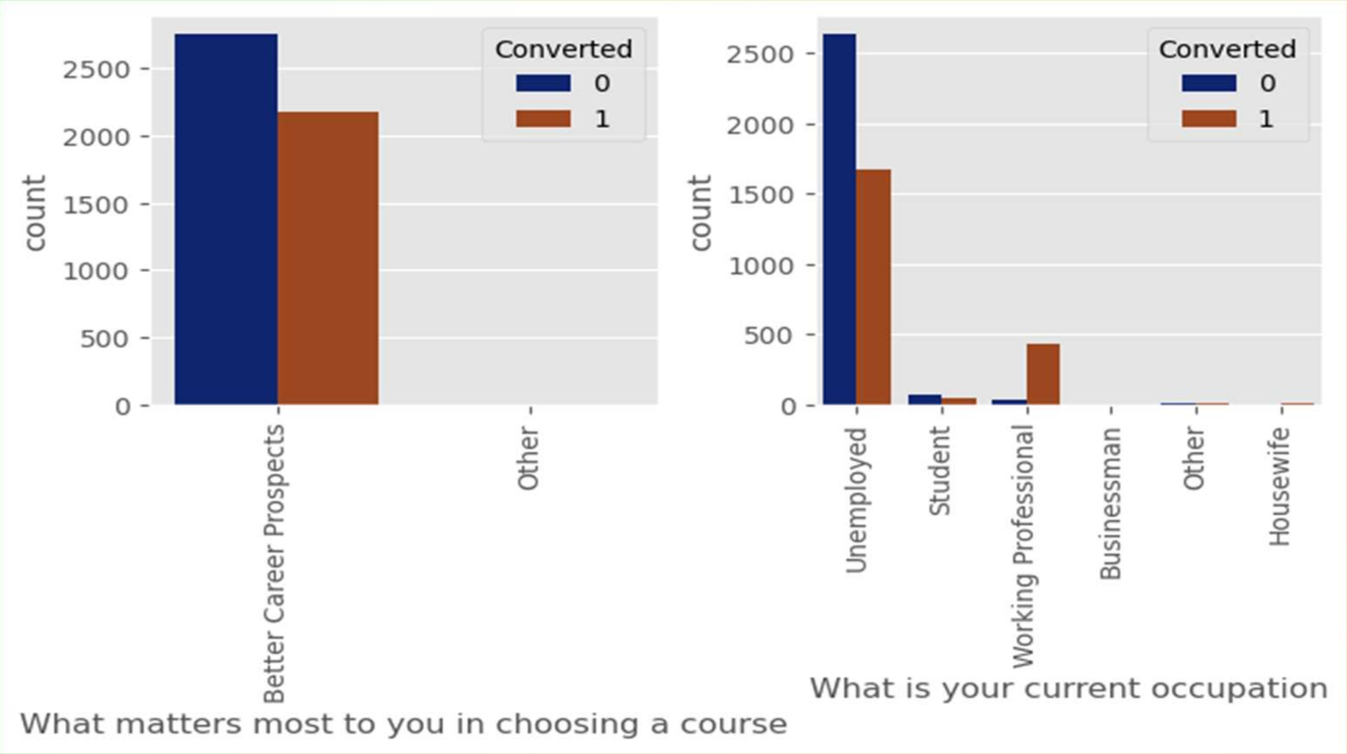


High **TotalVisits** and **Page Views Per Visit** create more conversion

Users with **more Time Spent on website** are likely to be converted

The website can be made more informative and appealing to increase the time user spends on website.

As the data showed many outliers we took 95th Percentile plotted again.



People who are looking for better career prospects have the highest number of leads and conversion
Working Professionals have the highest conversion
Unemployed users create more leads
To increase the overall conversion rate, we need to increase the number of leads for Working Professionals by reaching out to them more and also increasing the conversion rate of leads of Unemployed users.

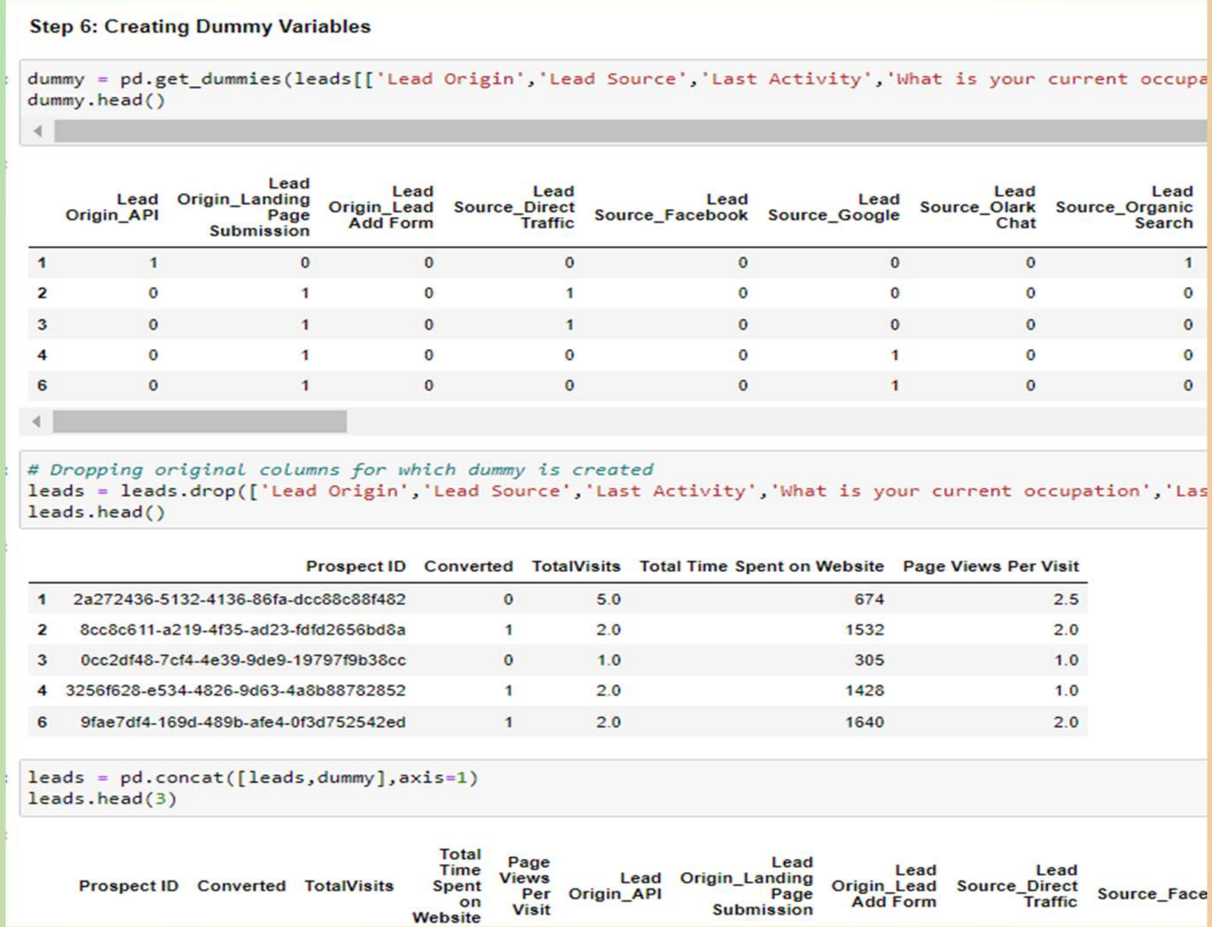
Data Conversion

- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 4925
- Total Columns for Analysis: 49

Model Building

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and VIF value is greater than 5
- Predictions on test data set
- Overall accuracy 79%

Dummy Variable Creation



Step 6: Creating Dummy Variables

```
dummy = pd.get_dummies(leads[['Lead Origin','Lead Source','Last Activity','What is your current occupa
dummy.head()
```

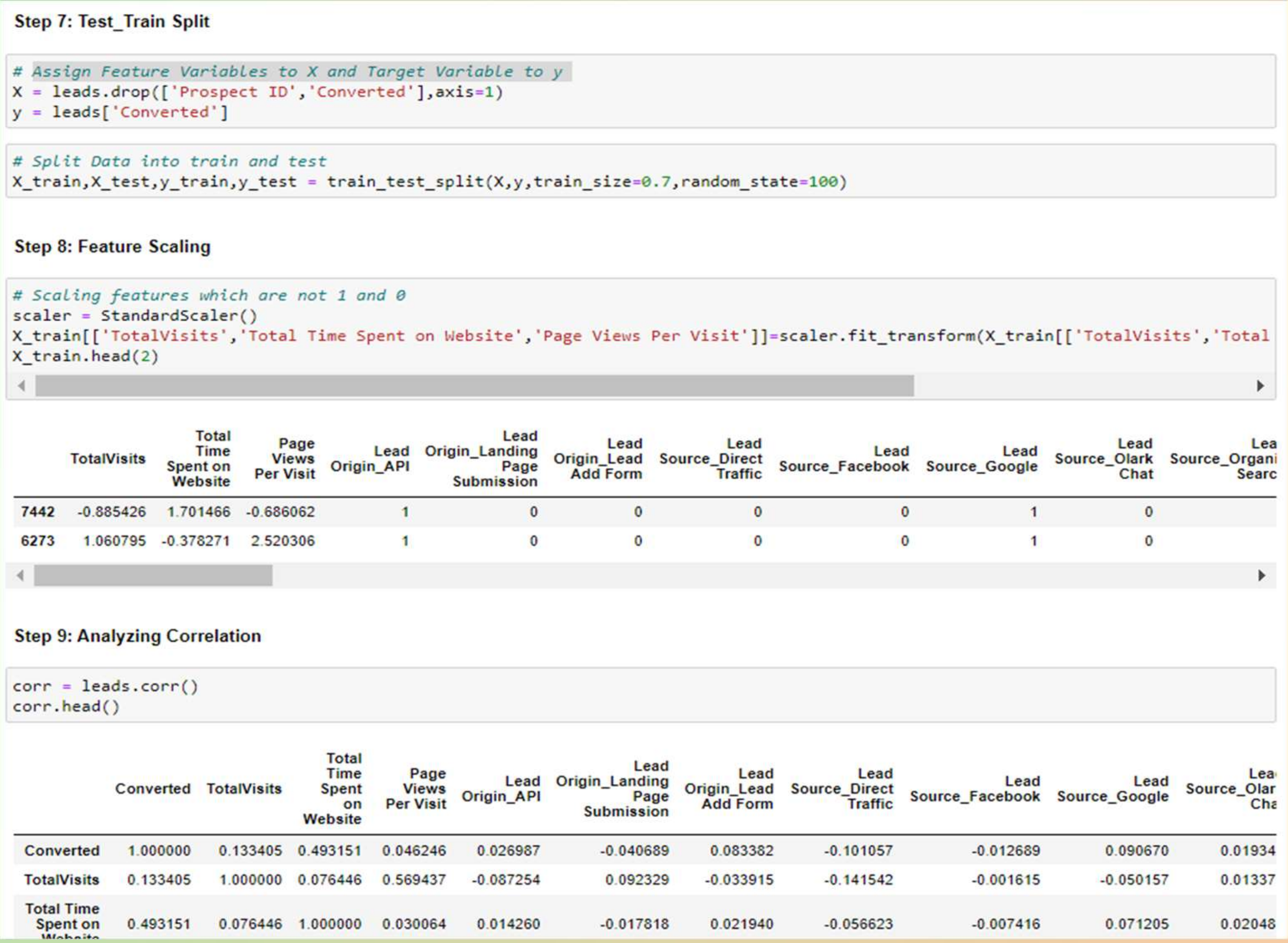
	Lead Origin_API	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Source_Direct Traffic	Lead Source_Facebook	Lead Source_Google	Lead Source_Olark Chat	Lead Source_Organic Search
1	1	0	0	0	0	0	0	1
2	0	1	0	1	0	0	0	0
3	0	1	0	1	0	0	0	0
4	0	1	0	0	0	1	0	0
6	0	1	0	0	0	1	0	0

```
# Dropping original columns for which dummy is created
leads = leads.drop(['Lead Origin','Lead Source','Last Activity','What is your current occupation','Las
leads.head()
```

	Prospect ID	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
1	2a272436-5132-4136-86fa-dcc88c88f482	0	5.0	674	2.5
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	1	2.0	1532	2.0
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	0	1.0	305	1.0
4	3256f628-e534-4826-9d63-4a8b88782852	1	2.0	1428	1.0
6	9fae7df4-169d-489b-afe4-0f3d752542ed	1	2.0	1640	2.0

```
leads = pd.concat([leads,dummy],axis=1)
leads.head(3)
```

	Prospect ID	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_API	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Source_Direct Traffic	Source_Face



Step 7: Test_Train Split

```
# Assign Feature Variables to X and Target Variable to y
X = leads.drop(['Prospect ID','Converted'],axis=1)
y = leads['Converted']
```

```
# Split Data into train and test
X_train,X_test,y_train,y_test = train_test_split(X,y,train_size=0.7,random_state=100)
```

Step 8: Feature Scaling

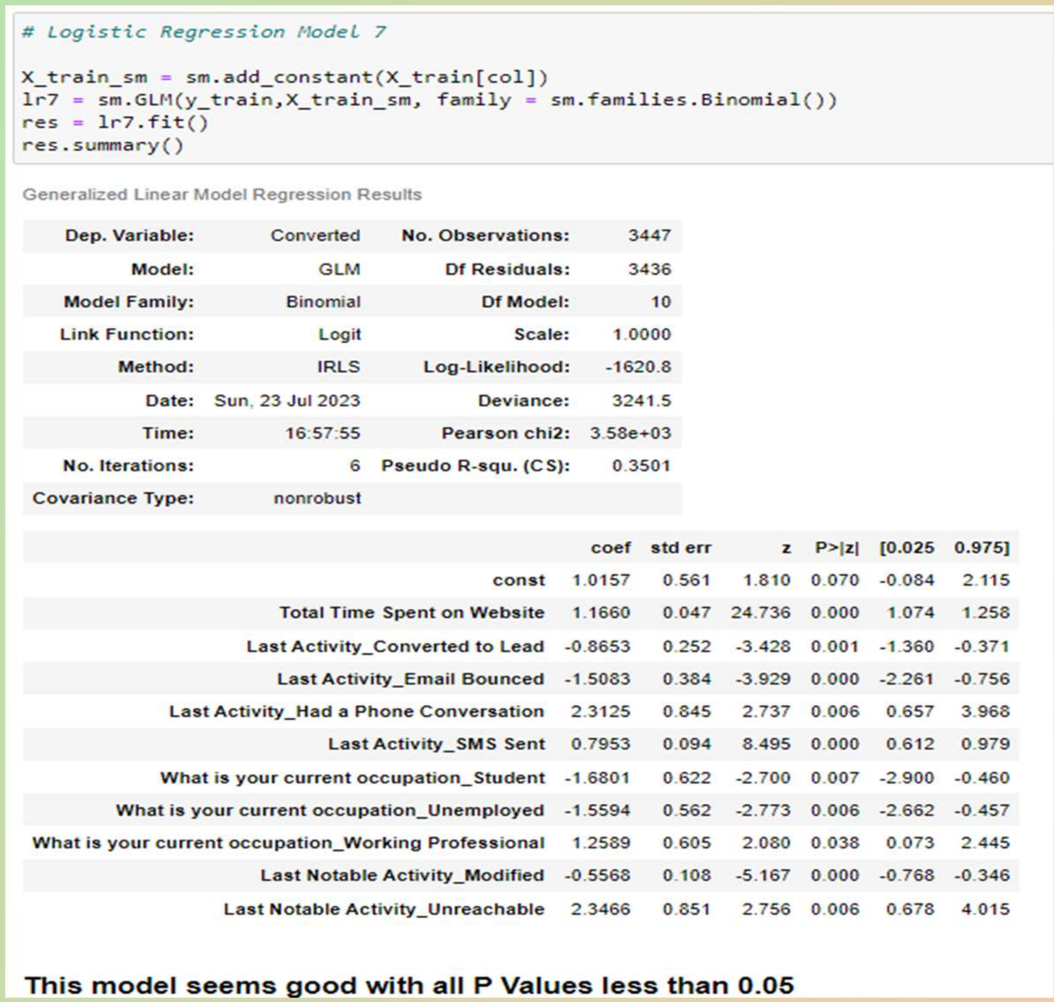
```
# Scaling features which are not 1 and 0
scaler = StandardScaler()
X_train[['TotalVisits','Total Time Spent on Website','Page Views Per Visit']]=scaler.fit_transform(X_train[['TotalVisits','Total
X_train.head(2)
```

	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_API	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Source_Direct Traffic	Lead Source_Facebook	Lead Source_Google	Lead Source_Olark Chat	Lea Source_Organi Searc
7442	-0.885426	1.701466	-0.686062	1	0	0	0	0	1	0	
6273	1.060795	-0.378271	2.520306	1	0	0	0	0	1	0	

Step 9: Analyzing Correlation

```
corr = leads.corr()
corr.head()
```

	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_API	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Source_Direct Traffic	Lead Source_Facebook	Lead Source_Google	Lea Source_Olar Cha
Converted	1.000000	0.133405	0.493151	0.046246	0.026987	-0.040689	0.083382	-0.101057	-0.012689	0.090670	0.01934
TotalVisits	0.133405	1.000000	0.076446	0.569437	-0.087254	0.092329	-0.033915	-0.141542	-0.001615	-0.050157	0.01337
Total Time Spent on Website	0.493151	0.076446	1.000000	0.030064	0.014260	-0.017818	0.021940	-0.056623	-0.007416	0.071205	0.02048

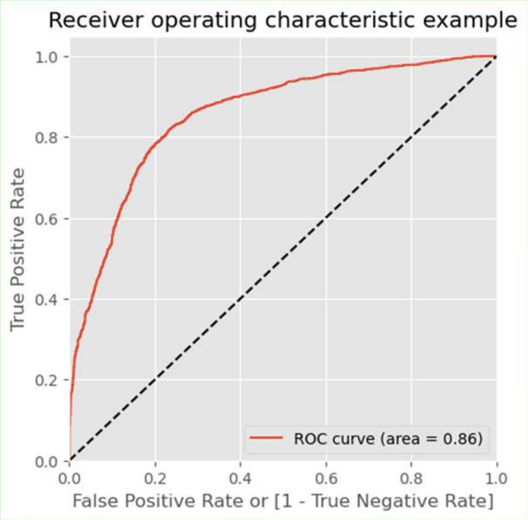


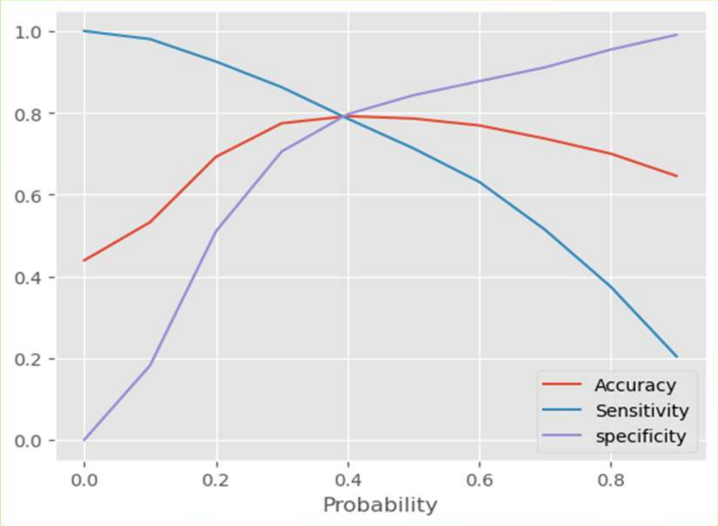
```
# Logistic Regression Model 7

X_train_sm = sm.add_constant(X_train[col])
lr7 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
res = lr7.fit()
res.summary()
```

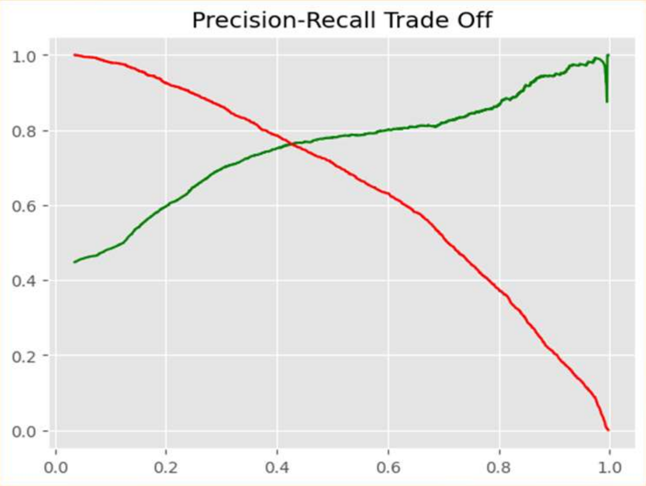
```
Generalized Linear Model Regression Results

   Dep. Variable:        Converted   No. Observations:      3447
           Model:              GLM       Df Residuals:      3436
    Model Family:         Binomial           Df Model:        10
   Link Function:            Logit              Scale:    1.0000
          Method:             IRLS     Log-Likelihood:   -1620.8
            Date: Sun, 23 Jul 2023           Deviance:    3241.5
            Time:         16:57:55       Pearson chi2:  3.58e+03
  No. Iterations:                6 Pseudo R-squ. (CS):    0.3501
 Covariance Type:        nonrobust

                                                 coef  std err       z  P>|z|  [0.025  0.975]
                                      const    1.0157    0.561   1.810  0.070  -0.084   2.115
                 Total Time Spent on Website   1.1660    0.047  24.736  0.000   1.074   1.258
             Last Activity_Converted to Lead  -0.8653    0.252  -3.428  0.001  -1.360  -0.371
               Last Activity_Email Bounced   -1.5083    0.384  -3.929  0.000  -2.261  -0.756
      Last Activity_Had a Phone Conversation   2.3125    0.845   2.737  0.006   0.657   3.968
                     Last Activity_SMS Sent    0.7953    0.094   8.495  0.000   0.612   0.979
       What is your current occupation_Student  -1.6801    0.622  -2.700  0.007  -2.900  -0.460
    What is your current occupation_Unemployed  -1.5594    0.562  -2.773  0.006  -2.662  -0.457
What is your current occupation_Working Professional  1.2589  0.605  2.080  0.038  0.073  2.445
             Last Notable Activity_Modified   -0.5568    0.108  -5.167  0.000  -0.768  -0.346
          Last Notable Activity_Unreachable    2.3466    0.851   2.756  0.006   0.678   4.015
```

This model seems good with all P Values less than 0.05

ROC Curve







- **Finding Optimal Cut-off Point**
- Optimal cut-off probability is that probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut-off is at 0.38.

Conclusion

► The Logistic regression model predicts the probability of the target variable having a certain value. The cut-off value is used to obtain the predicted value of the target variable.

► Optimum cut-off value is selected at 0.38, any lead with a probability greater than 0.38 can be considered as Hot Lead and any lead with a less than 0.38 probability can be considered as Cold Lead.

►Final model has 10 features:- 'Total Time Spent on Website', 'Last Activity Converted to Lead', 'Last Activity Email Bounced', 'Last Activity_Had a Phone Conversation', 'Last Activity_SMS Sent', 'What is your current occupation Student', 'What is your current occupation_Unemployed', 'What is your current occupation_Working Professional', 'Last Notable Activity Modified', 'Last Notable Activity Unreachable’

►Top 3 Factors are

1. Last Notable Activity_Unreachable with coefficient 2.346566
2. Last Activity_Had a Phone Conversation with coefficient 2.312539
3. What is your current occupation_Working Professional with coefficient 1.258886

► Train Accuracy Score: 0.789962286045837 Sensitivity : 0.799074686054197 Specificity : 0.7828335056876939 Precision Score : 0.7421731123388582 Recall Score : 0.799074686054197

► Test Accuracy Score: 0.7821380243572396 Sensitivity : 0.7954887218045112 Specificity : 0.7712177121771218 Precision Score : 0.7398601398601399 Recall Score : 0.7954887218045112

► Accuracy, Sensitivity and Specificity values of test set are around 78%, 79% and 77% which are approximately closer to the respective values calculated using the trained set.

► Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 0.799074686054197 (train) / 0.7954887218045112 (test)

► Hence overall this model seems to be good.

► The final model has Precision of 0.74, this means 74% of predicted hot leads are True Hot Leads.

► Final Prediction conversion on both train and test set is around 80%+ which is in line with the target

Thank You