DS C64 - Lead Scoring Case Study Logistic Regression

Agenda

Problem Statement

Business Goal

Strategy

Exploratory Data Analysis

Build a model

Model Evaluation

Conclusion

Problem Statement

X-Education is an education company which provides online courses to industry professionals

Recently, although X Education gets a lot of leads, its lead conversion rate is very poor

X Education want to select the most promising leads who are most likely to convert into paying customers

Business Goal

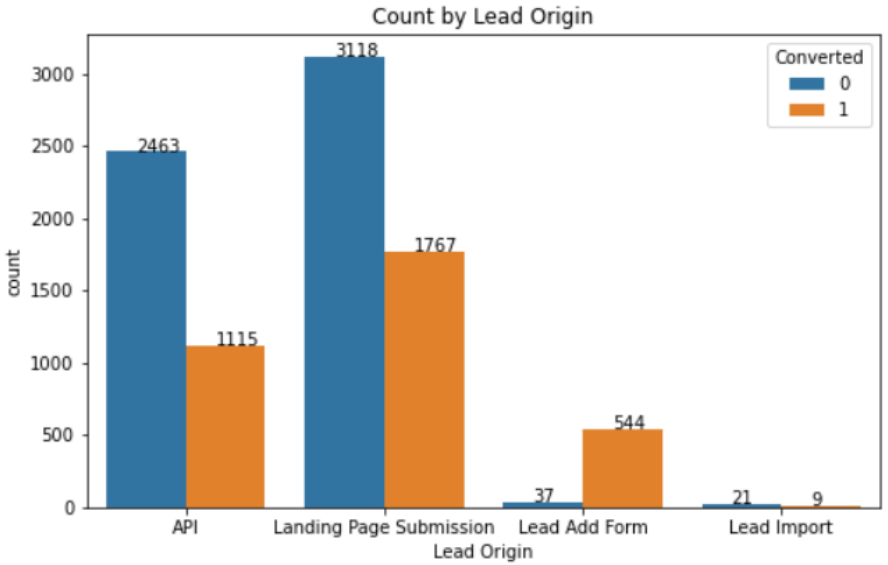
The company requires a model that predicts and assigns a lead score between 0 and 100 to each lead, such that customers with a higher lead score have a higher conversion chance and customers with a lower lead score have a lower conversion chance.

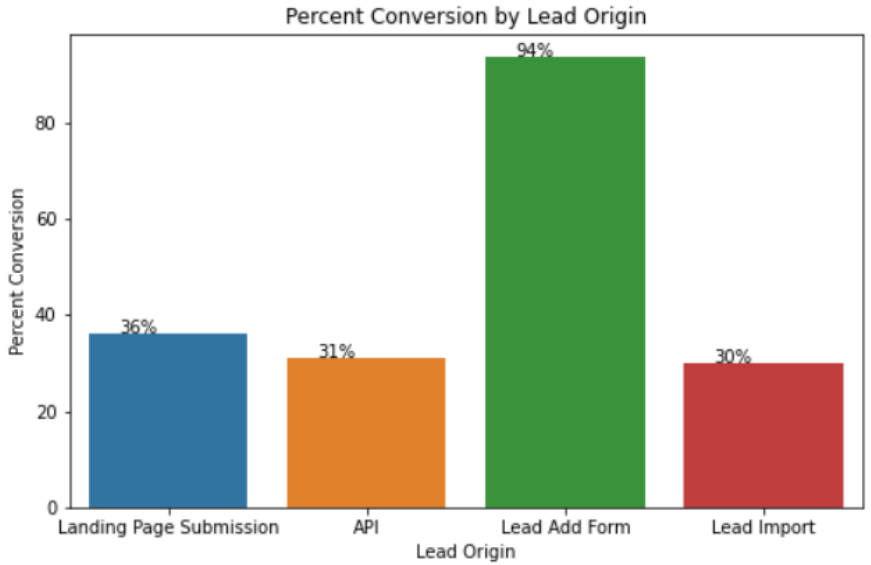
The target lead conversion rate is around 80% or more

Strategy

1. Data pre-processing
    - Data Understanding, Cleaning
    - Address Missing Values
2. Exploratory Data Analysis
    - Univariate Analysis for Categorical Variables & Numerical Variables
    - Bivariate Analysis for Numerical Variables
3. Build a logistic model
    - Refer slide 15 for details
4. Evaluate the model
    - Refer slide 16 for details

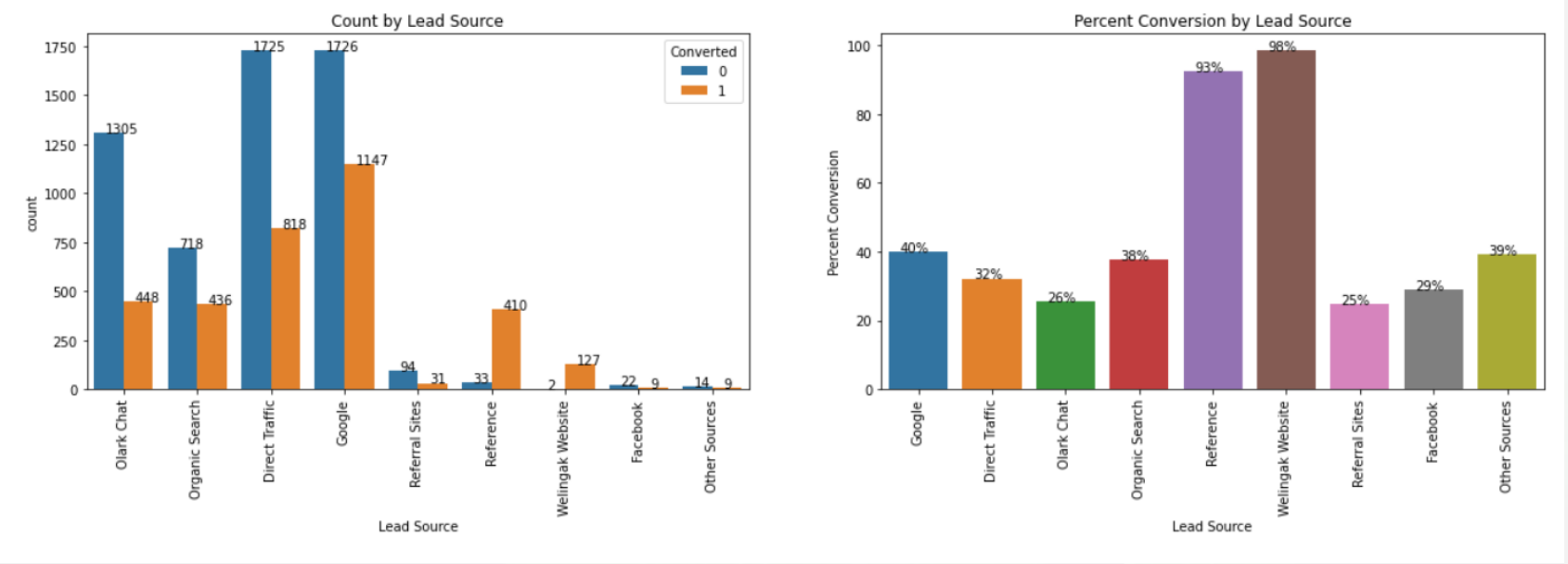
Lead Origin Variable





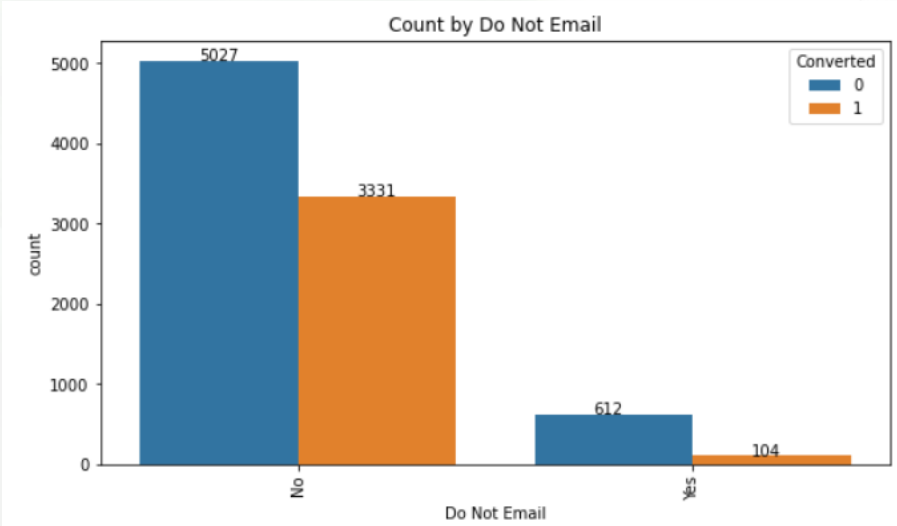
1.Most leads are generated from Landing Page Submission which counts at ~54% of total leads and its conversion rate is 36% which is quite low.
2. Second most leads are generated from API which counts at ~39% of total leads and its conversion rate is 31% which is quite low.
3. Highest lead conversion is from Lead Add Form which is 94% though lead generated from it is comparatively less i.e. 6%

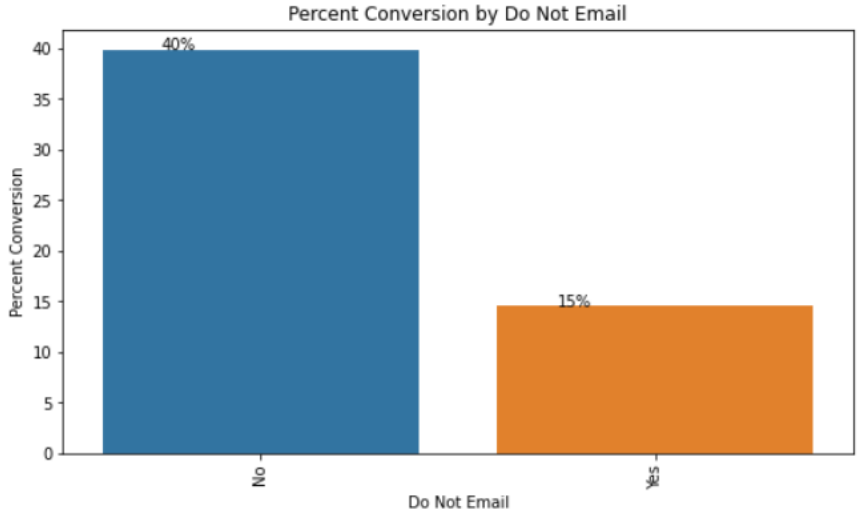
Lead Source Variable



1.Most leads are generated from Google which counts at ~32% of total leads and its conversion rate is 40% which is quite low.
2.Next most leads are generated from Direct Traffic which counts at ~28% of total leads and its conversion rate is 32% which is quite low.
3.The lead generation from sources like "Olark Chat", "Organic Search","Referral Sites" are moderate and also conversion is in general range of 25% to 40%
4.Converstion rate is maximum for Welingak Website which is 98% though the lead generated from this source is less i.e.~2%.
5.Next highest Converstion rate is for Reference source which is 93% and the lead generated from this source is ~5%.

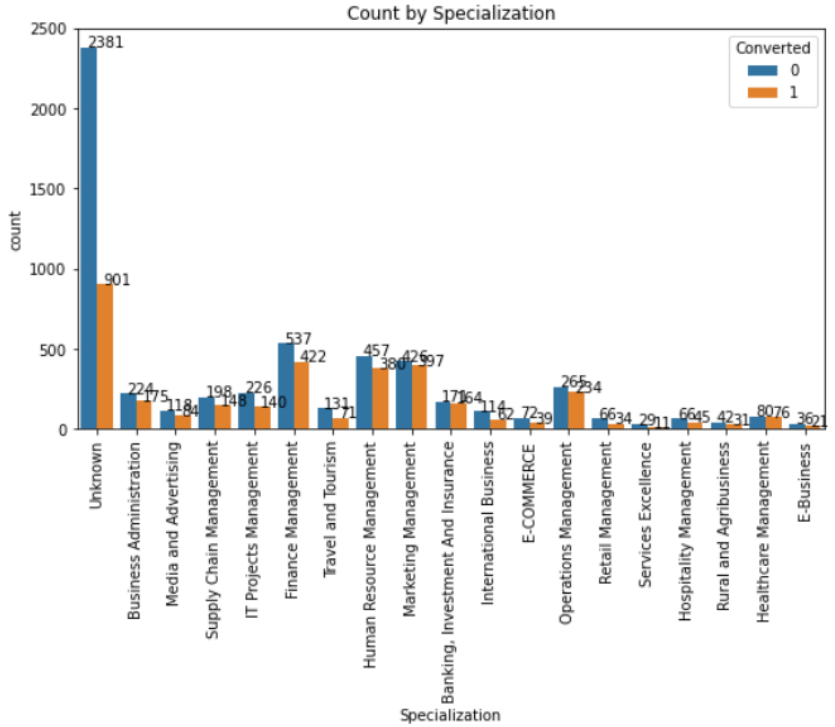
Do Not Email Variable

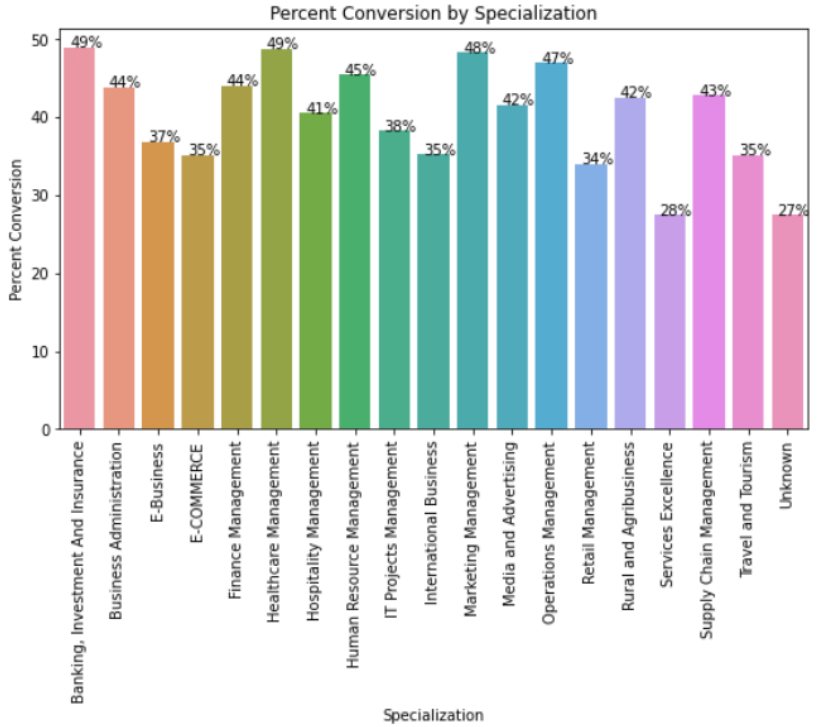




1. Most of the customers has opted for sending email which stands at ~92% and conversion rate for these customer is higher which is 40%
2. Those who has opted to not to send email has lower conversion rate i.e. 15%.

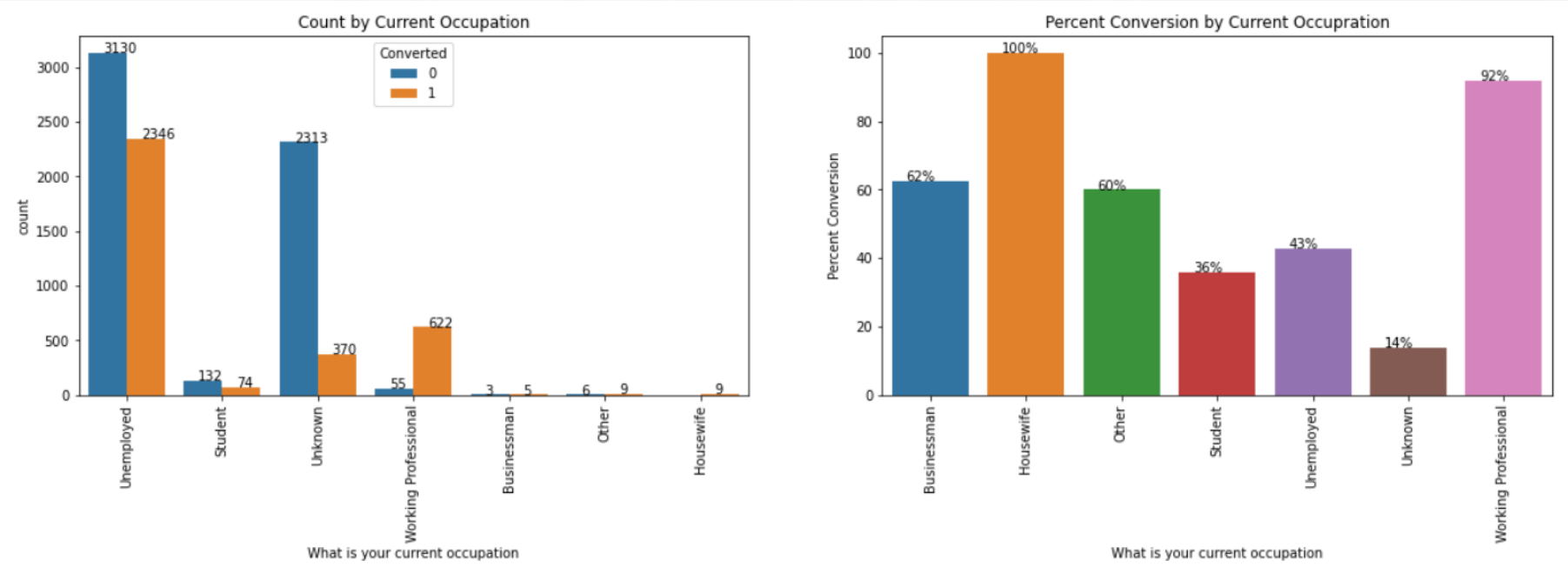
Specialization Variable





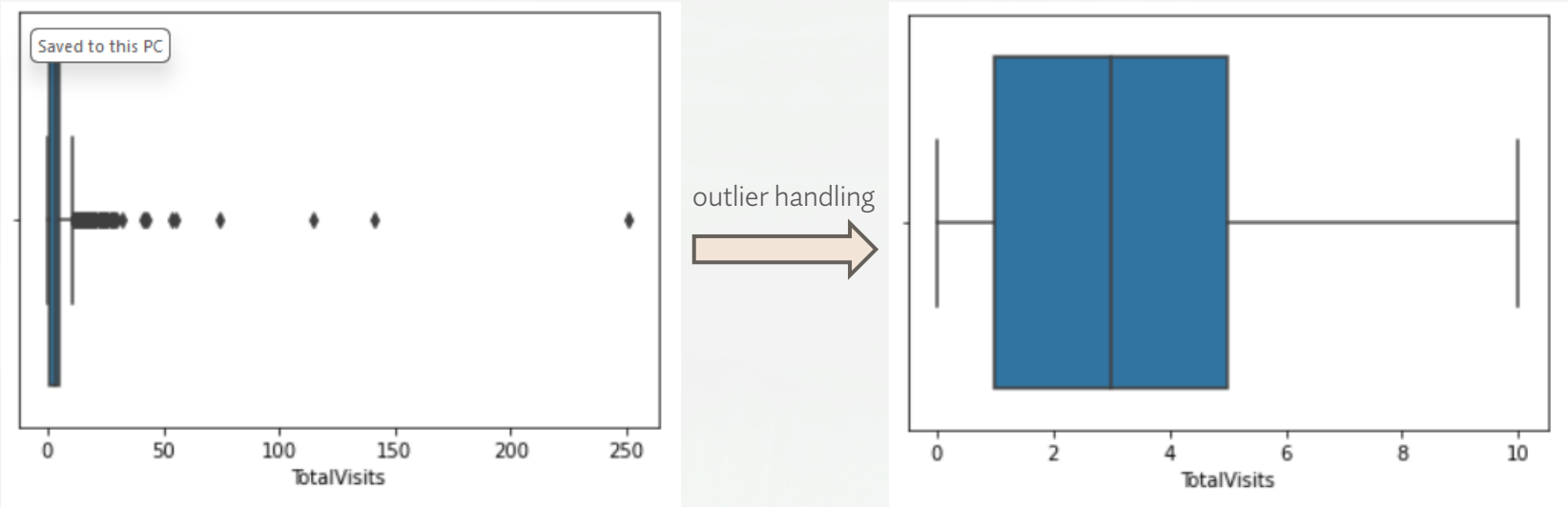
1.Majority of the leads are not having any specialization specified and its conversion rate is 27%.
2.Conversion rate for specialization like "Banking, Investment and Insurance","Healthcare Management", "Marketing Management", "Operations Management", "Human Resource Management", "Finance Management","Business Management" is well above 44%.

What is your current occupation



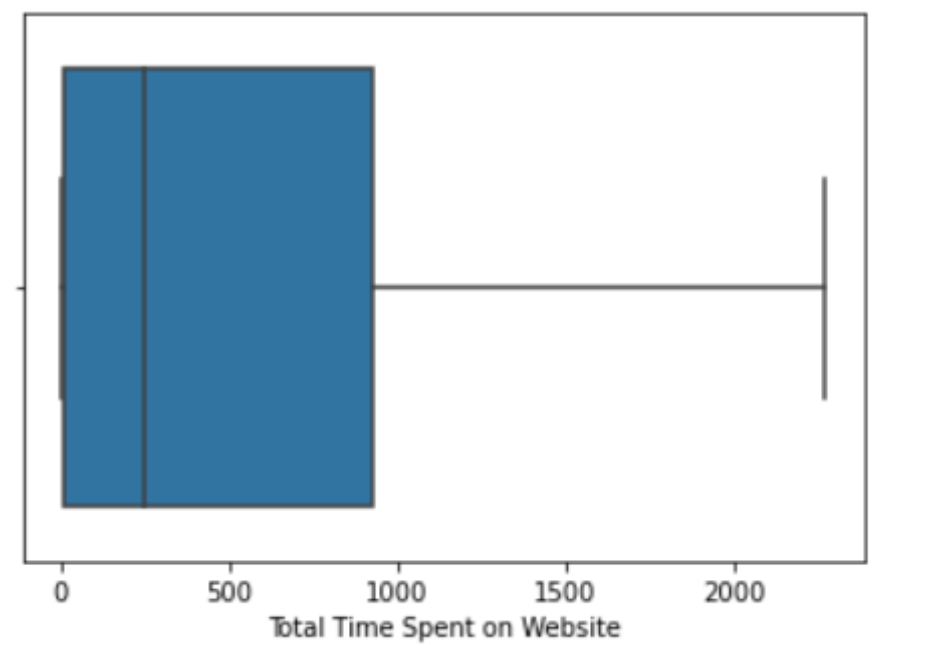
1. Most highest number of leads are generated by customers who are Unemployed however their rate of conversion is less 43%.
2. Conversion rate for Working Professional is high i.e 92% though the lead generated by them is less comapred to Unemployed.

TotalVisits



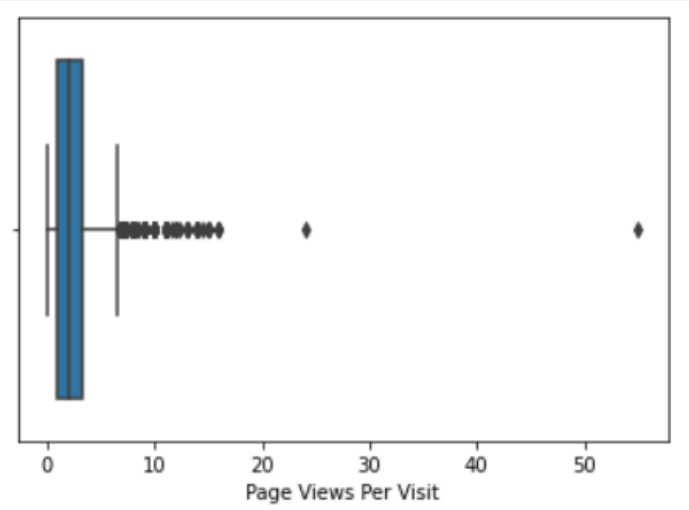
1. This variable 'TotalVisits' has higher values which is true for more visit leads however, it will affect ML model.
2. Capping the outliers to 95 percentile value
3. After handling outlier, The median of total visit is 3 times

Total Time Spent on Website

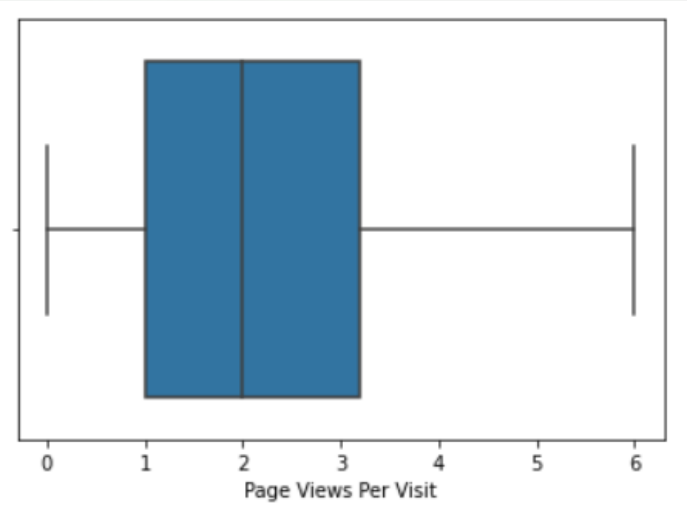


1. The values for Q1, median, and Q3 are 0, 250, and 900, respectively.
2. The maximum is greater than 2000 (in time units).

Page Views Per Visit

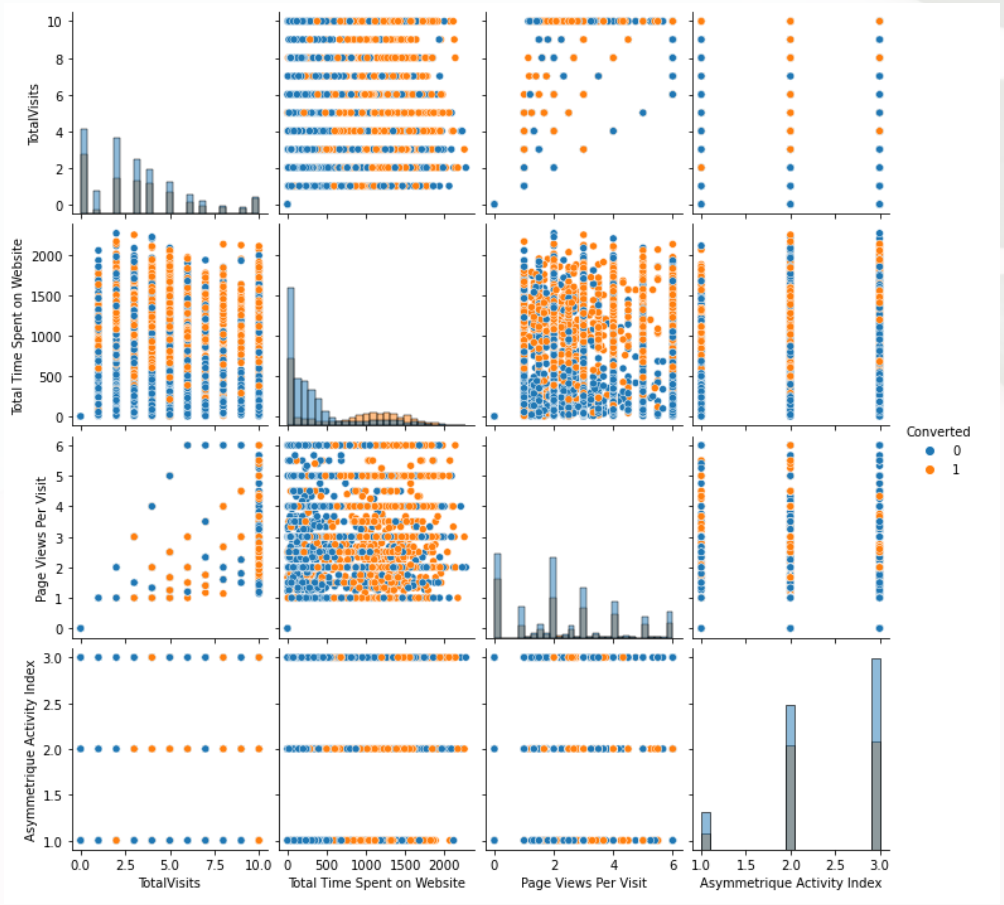


outlier handling



1. This variable 'Page Views Per Visit' has higher values which is true for more Page views visit leads however, it will affect ML model.
2. Capping the outliers to 95 percentile value.
3. After handling outlier, The median of page views per visit is 2 pages

Numerical variables pair plot



More the time spent by customer on the website, more is the conversion rate

Build a logistic model

- Convert a categorical variable into dummy variables
- Separate features and the target variable (Converted)
- Split the data into training and testing sets
- Standardize numerical columns
- RFE for Feature Selection
- Build a logistic model using statsmodels
- VIF for Detecting Multicollinearity
- Drop features based on VIF and P-value
- Rebuild the model

Evaluate the model

- Confusion matrix
- Plotting ROC curve
- Calculate accuracy, sensitivity and specificity for each probability cut-offs
- Choose Optimal Probability Cut-off
- Making Predictions on the test set
- Compare the metrics between train set and test set
- Assign lead score for each of customers

Confusion Matrix

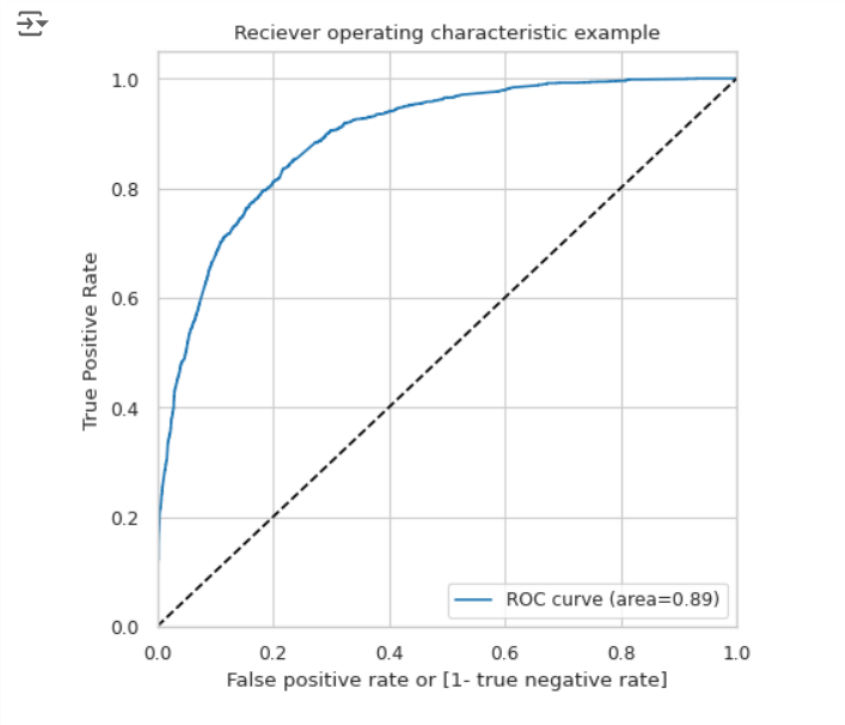
Cut-off = 0.5 (Default)

True Negative 3458	**False Positive** 447
False Negative 716	**True Positive** 1730

Cut-off = 0.35

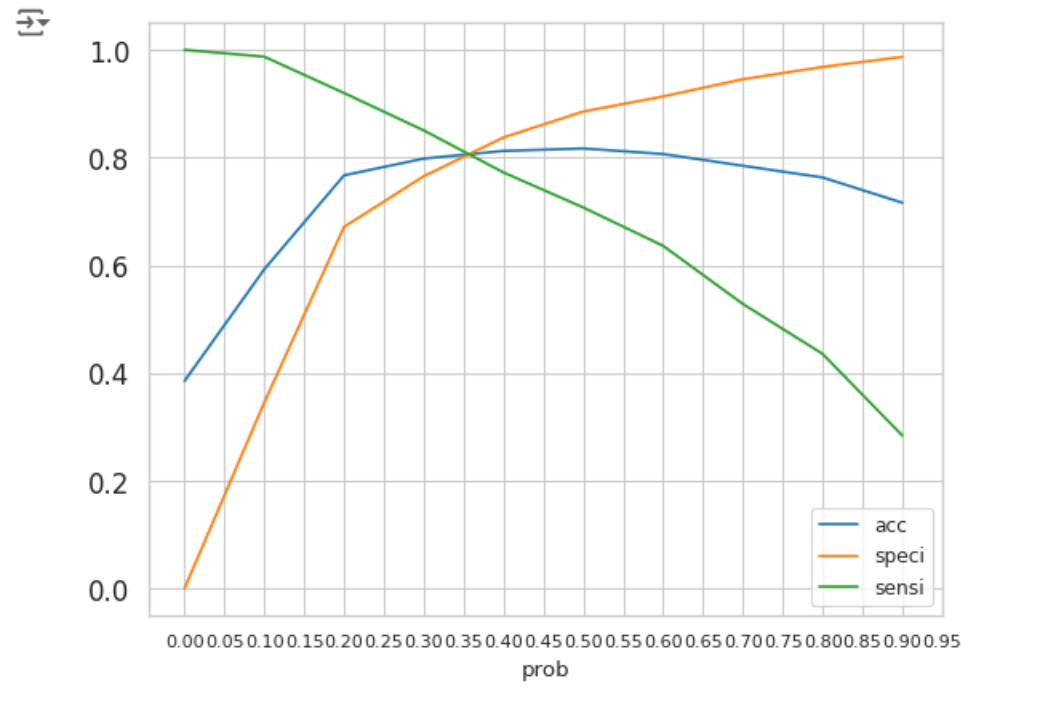
True Negative 3146	**False Positive** 759
False Negative 482	**True Positive** 1964

ROC curve



A ROC AUC of 0.89 suggests that the model has very high discriminative ability.

Optimal Probability Cut-off



Optimal Probability Cut-off = 0.35

Model Coefficients

```
const                                              -0.114779
Total Time Spent on Website                         1.125375
Lead Origin_Landing Page Submission                -1.125498
Lead Origin_Lead Add Form                           1.786744
Lead Source_Welingak Website                        2.532839
Do Not Email_Yes                                   -1.786724
Last Activity_Had a Phone Conversation              2.710007
Last Activity_Olark Chat Conversation              -1.304764
Last Activity_SMS Sent                              1.340702
Last Activity_Unsubscribed                          1.524911
Country_Unknown                                     1.332275
Specialization_Unknown                             -1.094328
What is your current occupation_Unknown            -1.193705
What is your current occupation_Working Professional    2.331500
dtype: float64
```

Top three variables in your model which contribute most towards the probability of a lead getting converted

1. Last Activity_Had a Phone Conversation (~2.7)
2. Lead Source_Welingak Website (~2.5)
3. What is your current occupation_Working Professional (~2.3)

Model Performance Metrics

1. Train Data:
   1.1. Accuracy : 80%
   1.2. Sensitivity : 80%
   1.3. Specificity : 81%
   1.4. Precision : 72%
   1.5. Recall : 80%

2. Test Data:
   2.1. Accuracy : 81%
   2.2. Sensitivity : 80%
   2.3. Specificity : 81%
   2.4. Precision : 71%
   2.5. Recall : 80%

The performance metrics for Train and Test data is close to each other. Thus the model generalizes well on Test/ unknown data

Thank you