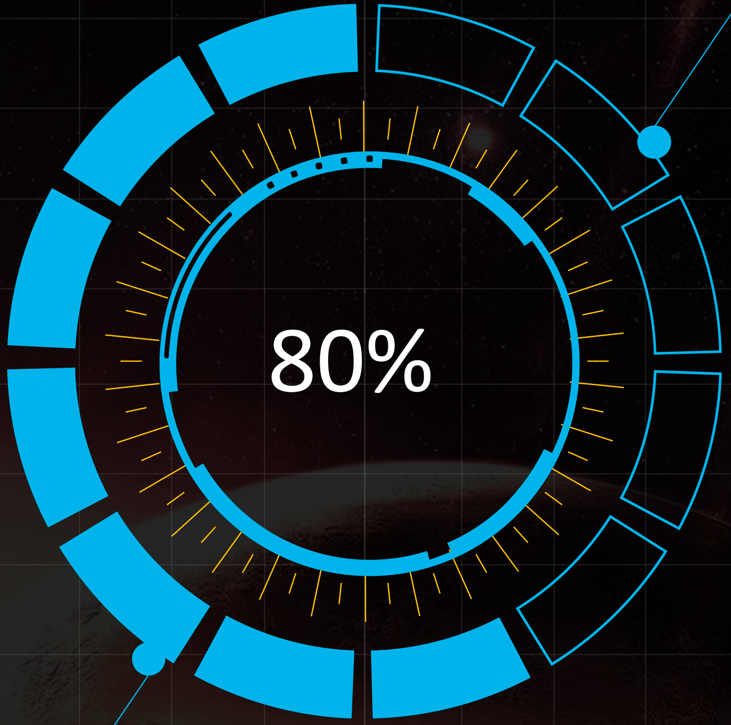
Logistic Regression

Lead Score Case Study

Problem Statement

A problem for the online course provider X Education is that their lead conversion rate is only 30%. The company's goal is to find "Hot Leads" with higher conversion potential in order to increase efficiency. The goal is to develop a predictive model that assigns scores to leads so that the sales team can get in touch with the most likely-to-convert leads first. Achieving an 80% lead conversion rate is the CEO's goal. Optimizing the lead conversion process and boosting the efficacy of the sales team's endeavors are the objectives.



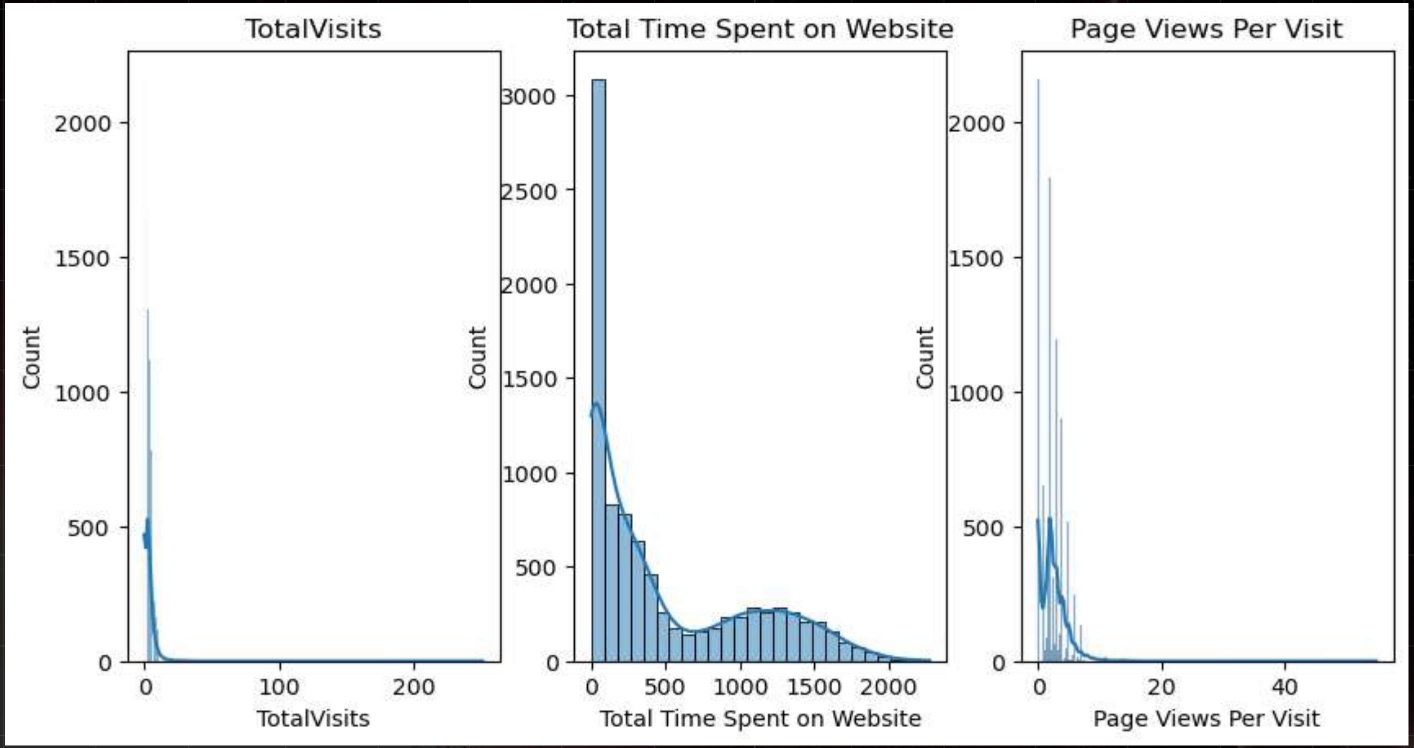
Objective

Create a logistic regression model that will allow the business to target potential leads by giving each lead a score between 0 and 100. In contrast, a lower number would indicate that the lead is chilly and unlikely to convert, but a higher score would indicate that the lead is hot and likely to convert.

Data Cleaning

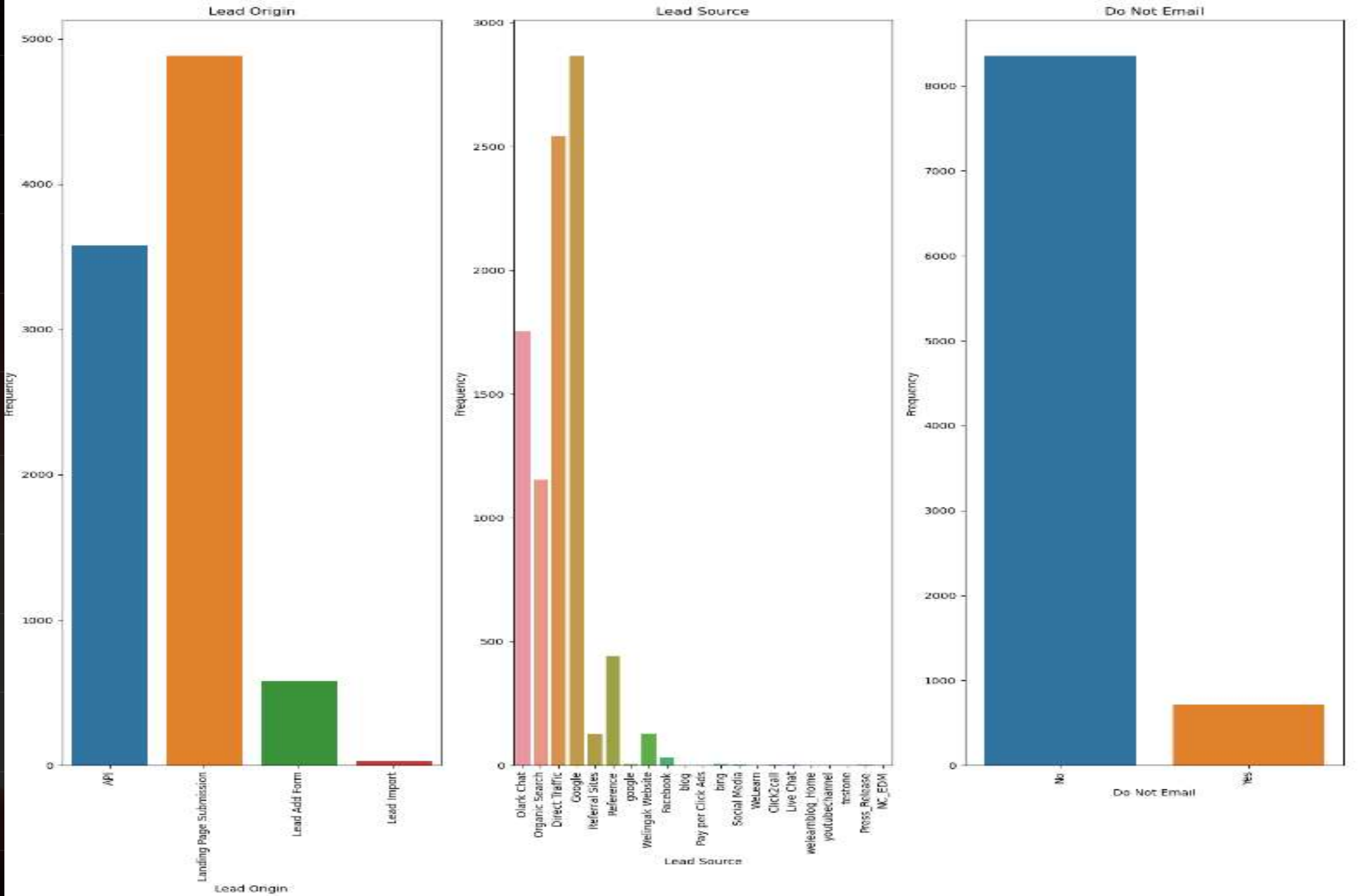
- The columns **Magazine**,R**eceive More Updates About Our Courses**,**Update me on Supply Chain Content**,**Get updates on DM Content**,**I agree to pay the amount through cheque** containing only one unique value and were dropped.
- All the **select** values were handled by replacing them will null values.
- The columns **Prospect ID**,**Lead Number** were not required for the analysis and were dropped.
- Another columns like **Asymmetrique Activity Index**,**Asymmetrique Profile Index**,**Asymmetrique Activity Score**,**Asymmetrique Profile Score**,**How did you hear about X Education**,**Tags**,**Lead Quality**,**Lead Profile** and **City** had large amount of null values and were dropped.
- Columns like **Specialization**,**Country**,**What is your current occupation** and **What matters most to you in choosing a course** were looking important and null values in them were imputed with **not given** value.
- The remaining null values were very less in percentage so we removed all the rows with null values and there were no not null data in the columns of the dataset.
- At last,we checked the values in each column and the columns **Do Not Call**,**Search**,**Newspaper Article**,**X Education Forums**,**Newspaper**,**Digital Advertisement**,**Through Recommendations** and **Country** containing heavily imbalanced data were removed.

Exploratory Data Analysis



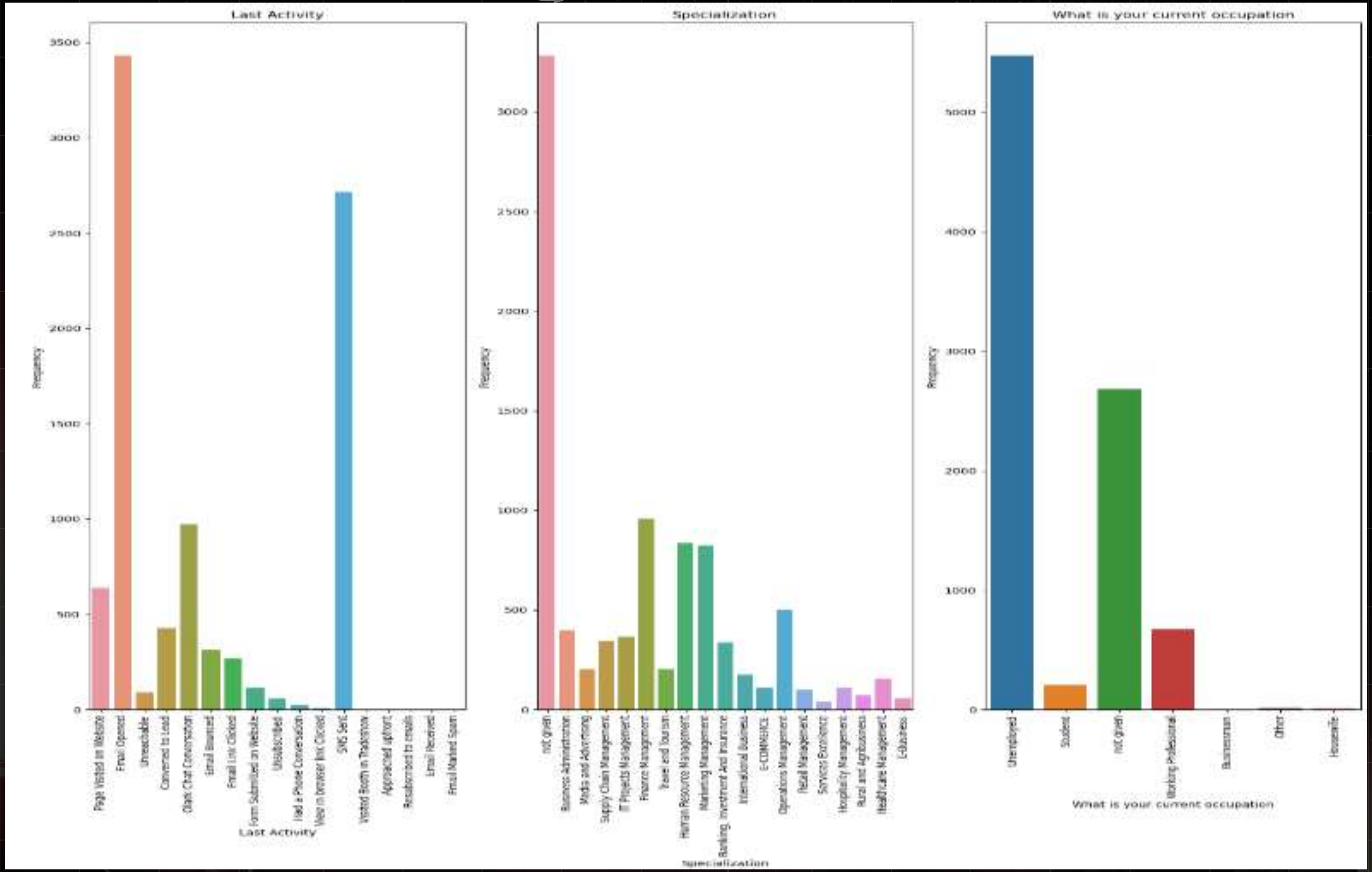
Univariate analysis of numerical columns

Exploratory Data Analysis



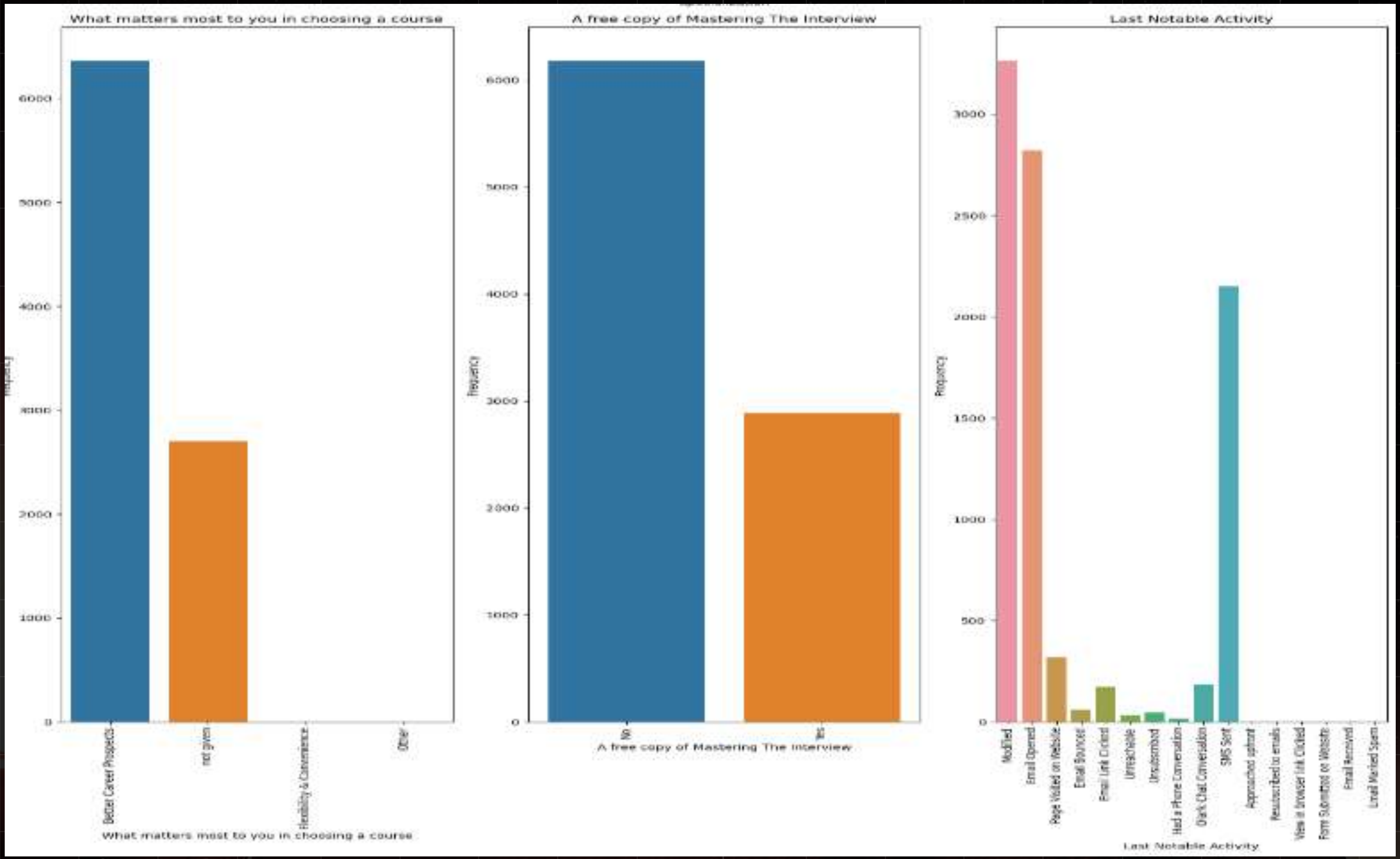
Univariate analysis of categorical columns

Exploratory Data Analysis



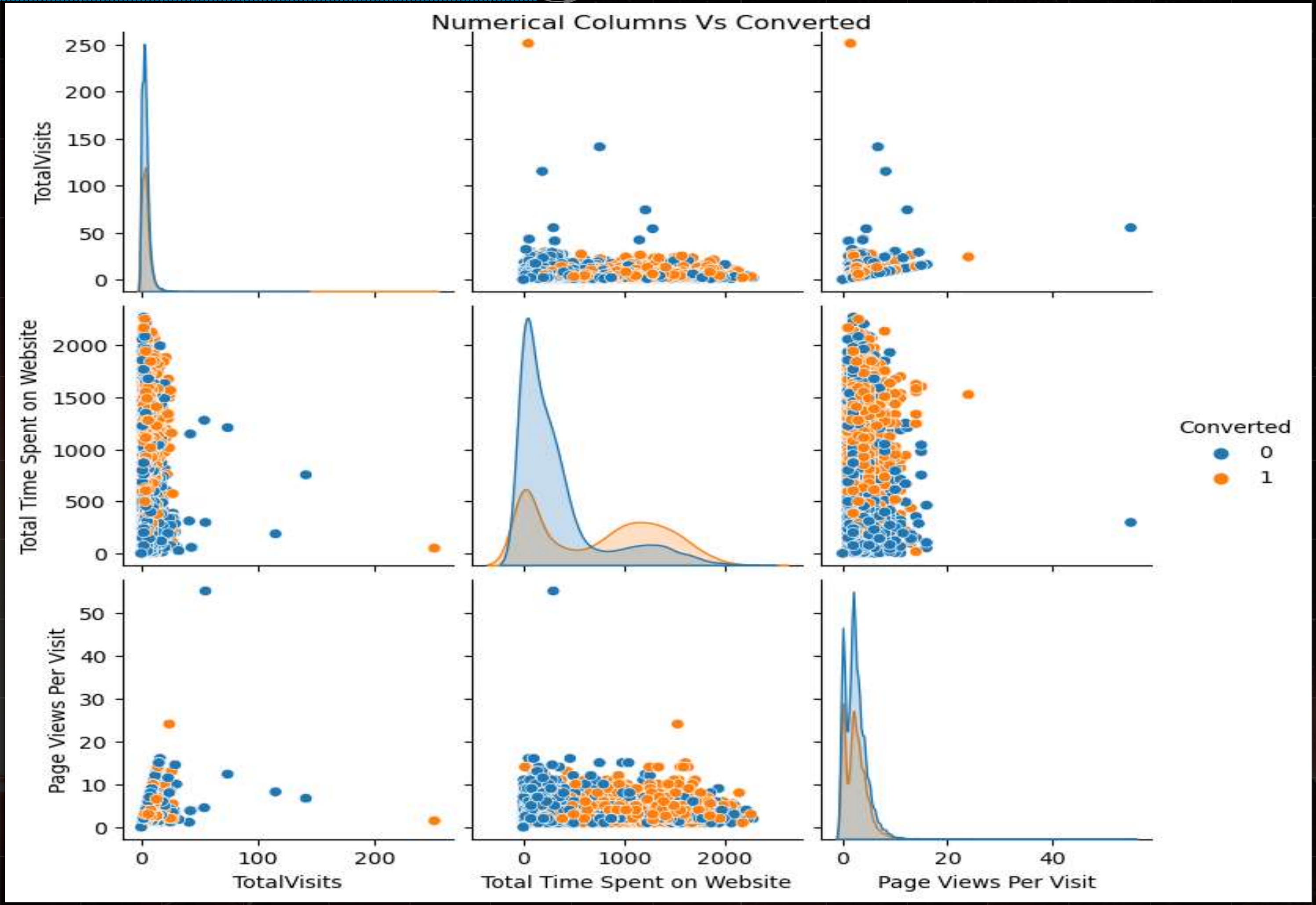
Univariate analysis of categorical columns

Exploratory Data Analysis



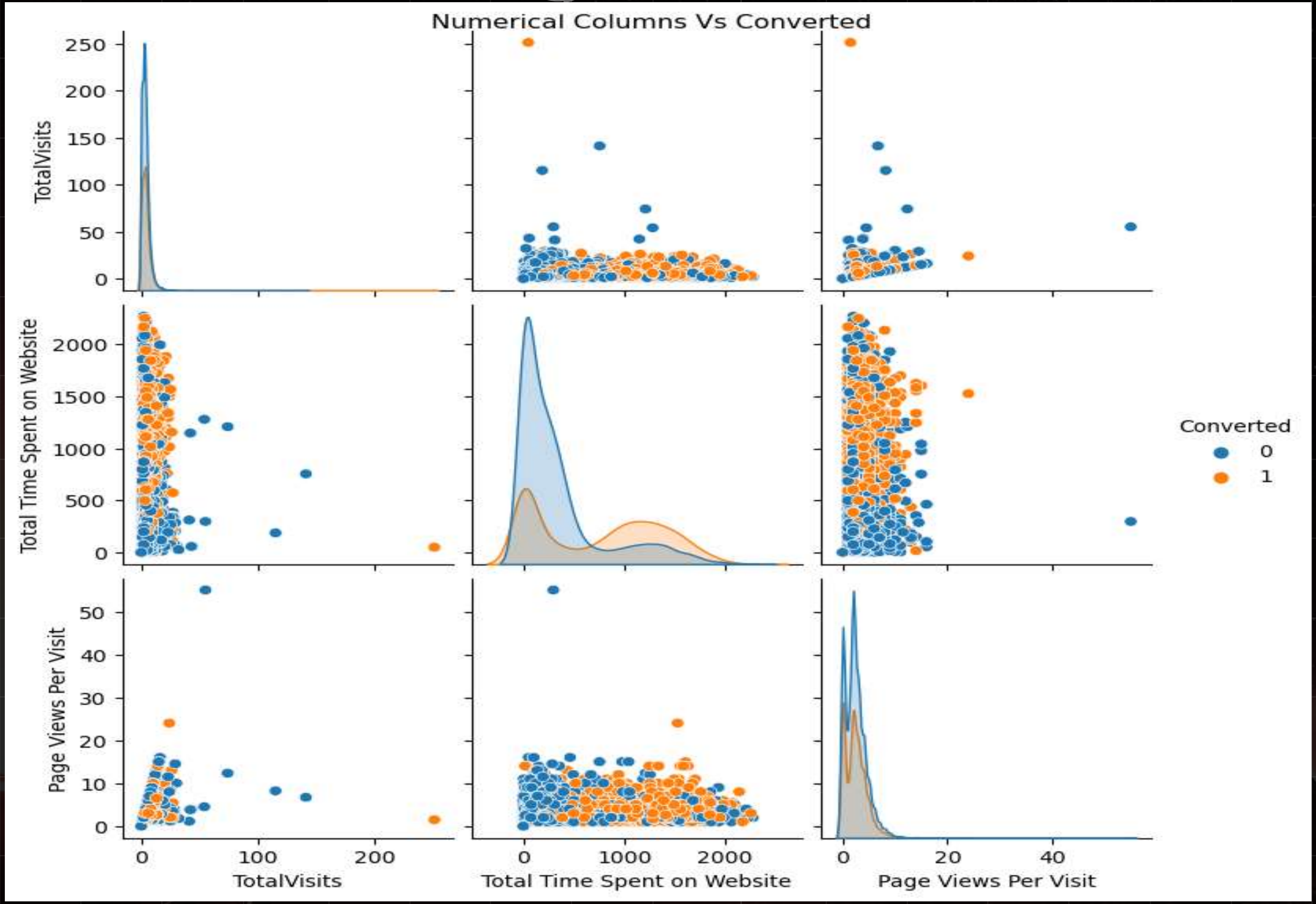
Univariate analysis of categorical columns

Exploratory Data Analysis



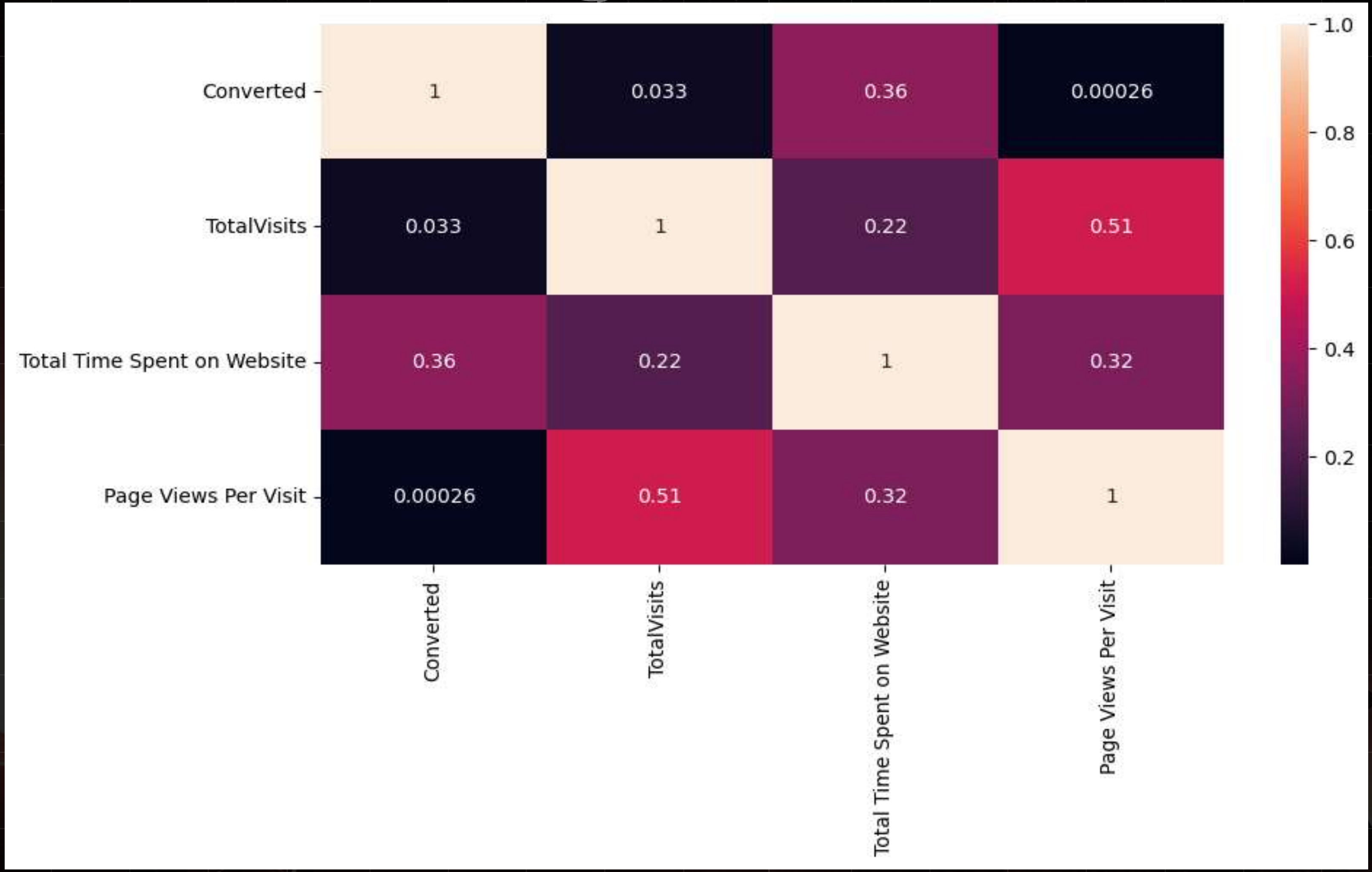
Bivariate analysis of numerical columns

Exploratory Data Analysis



Bivariate analysis of numerical columns

Exploratory Data Analysis



Correlation between numerical columns

Dummy Variable Creation, Train Test Split and Scaling

- Using one-hot encoding method, dummy variables were created for all the categorical columns in the dataset.
- The first level of each categorical dummy variable created were dropped to avoid multicollinearity.
- The dataset was divided into train and test dataset with 70% being the train and 30% to be the test dataset.
- The numerical columns in the dataset were scaled using the Standard Scaler method.

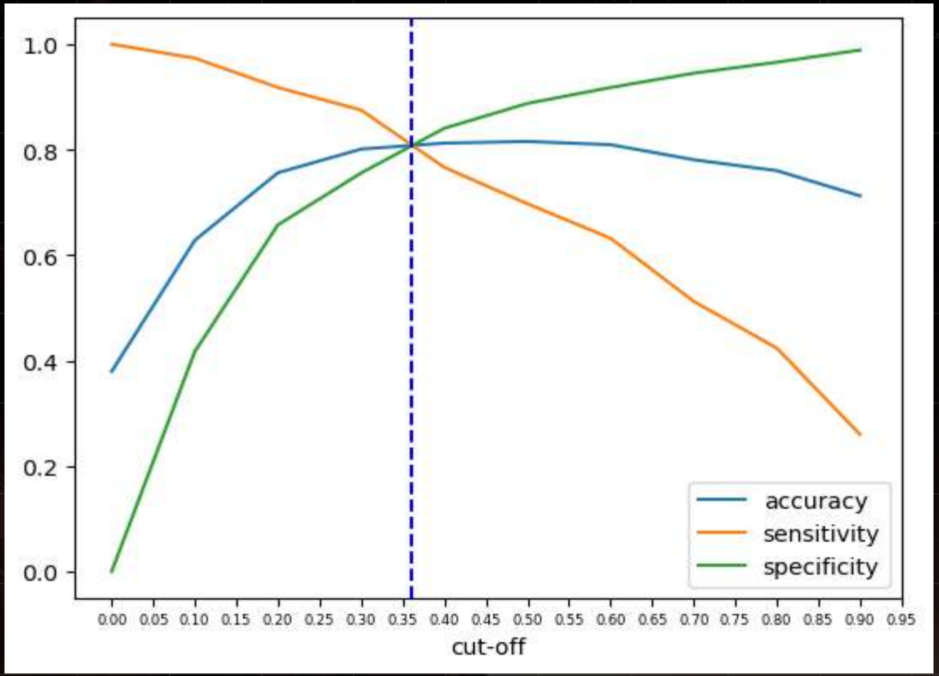
RFE and Model Building

- Using RFE, the top 15 columns were selected to be used for model building.
- The first model was built using all the 15 selected columns.
- Upon analysing the p-value and VIF, the column **What is your current occupation_Housewife** was dropped for the model building and the model was re-built with the remaining 14 columns.
- The updated model, with the 14 columns had p-values as well as VIF in acceptable range and the model was finalised.

Prediction on Train Data and Optimal cut-off

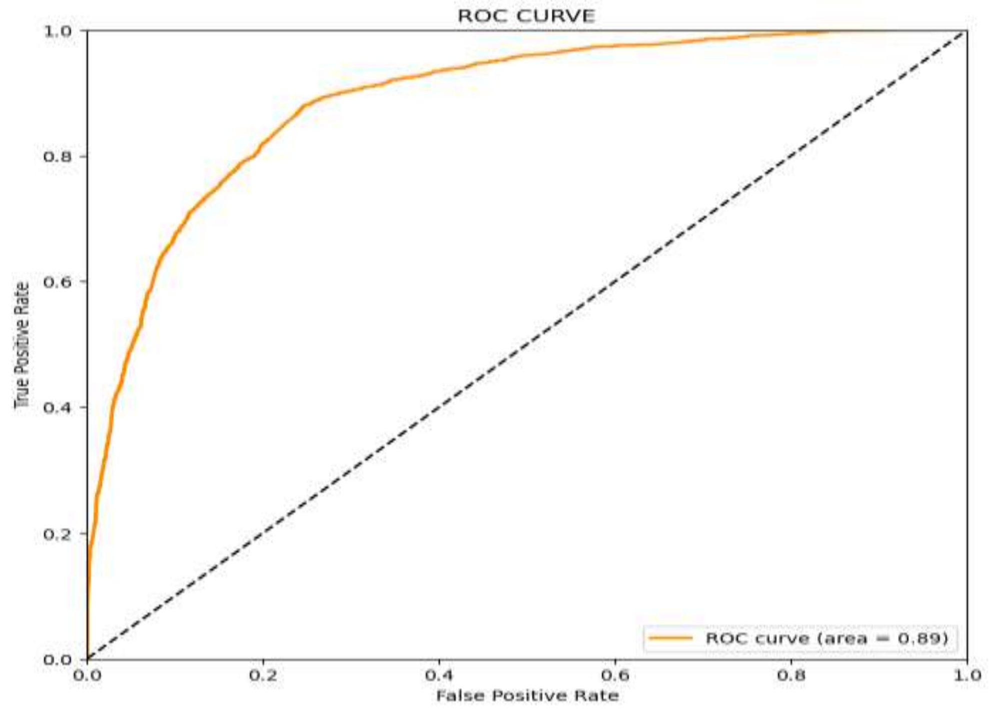
- Prediction was made on the train dataset and conversion probability was calcualted.
- A new dataframe was created with the actual conversion and conversion probability and accuracy, sensitivity and specificity was calculated for each probability cut-off ranging from 0 to 0.9 with a split of 0.1.
- Using the metrics calculated, a curve was made to find the optimal cut-off.
- The intersection of all the three was the point of optimal cut-off.
- Predicted column was created in the dataframe based on conversion probability and optimal cut-off.
- Confusion matrix was created from the updated dataframe.

Optimal Cut-off: 0.36
Accuracy: 0.80
Sensitivity: 0.79
Specificity: 0.81
Precision: 0.72
Recall: 0.79



ROC Curve

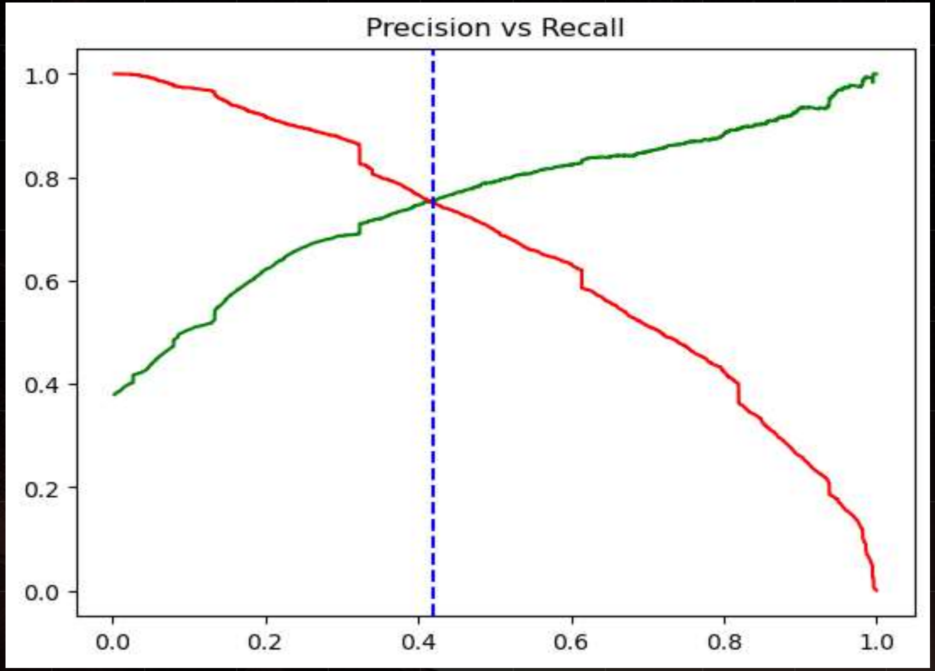
- Based on the true positive rate, false positive rate and threshold, an ROC curve was created and the area under the curve, AUC, came to be 0.89 which depicts that the model has a very good discriminatory power.



Precision Recall Trade-off

- A curve of precision and recall was made to do a trade between precision and recall.
- The intersection of the recall and precision was used as the updated optimal cut-off probability and used to update the dataframe containing the actual and conversion probability and the optimal cut-off was used to update the predicted value.
- Confusion matrix was created from the updated dataframe.

Updated Optimal Cut-off: 0.42
Accuracy: 0.81
Sensitivity: 0.75
Specificity: 0.85
Precision: 0.75
Recall: 0.75



Prediction on Test Dataset and Lead Score Assignment

- Prediction was made on the test dataset and the dataframe with the actual, conversion probabilty and predicted value was created using the optimal cut-off.
- Confusion matrix was created and the accuracy, sensitivity, specificity, precision and recall was calculated on the test dataset as well.
- The conversion probability was multiplied by 100 and was assigned to a new column calling **Lead Score.**

Updated Optimal Cut-off: 0.42
Accuracy: 0.82
Sensitivity: 0.76
Specificity: 0.86
Precision: 0.76
Recall: 0.76

Conclusion

The variables on which lead conversion depends the most is:

- When the LEAD ORIGIN is LEAD ADD FORM.
- When the CURRENT OCCUPATION is WORKING PROFESSIONAL.
- When the LEAD SOURCE is from Welingak Website and OLARK chat.
- Total Time Spent on Website.

- X education should take care of these columns which could lead to potential conversions.

Thank you