LEAD SCORE CASE STUDY

Pramod Bhaskar
Karthik Ramadass
Vishwas Yadav

Business Objective

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Goals of the case study

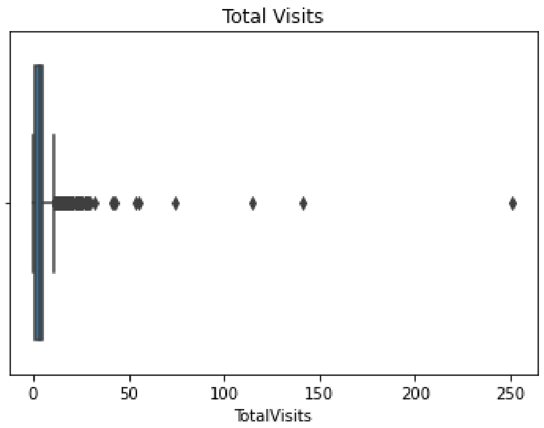
Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

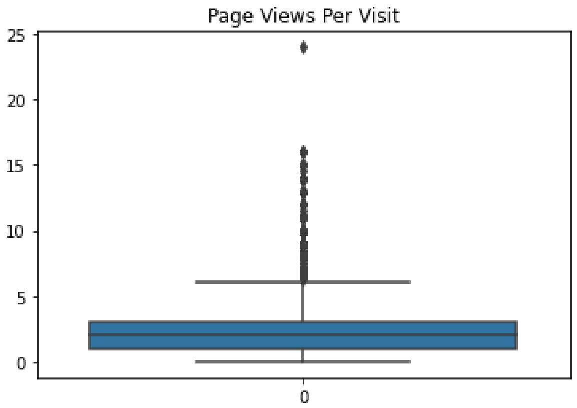
High level Approach

- ❑ Data Extraction and Cleansing
    - ▪ Handle Null values
    - ▪ Handle Outliers, Impute missing values
- ❑ Data Preparation
    - ▪ Convert Binary variable (Y/N) to 0’s and 1’s
    - ▪ Identify Correlation through heatmap and acting upon
    - ▪ Create Dummy variable for multiple categorical values
    - ▪ Dropping the extra dummy variables
- ❑ Test and Train Split
- ❑ Feature Scaling using Standard Scaler method
- ❑ Model Building using Stats model (Generalized Linear Model method)
- ❑ Feature selection using RFE
- ❑ Create multiple models to reduce the features using p-Value and VIF parameters
- ❑ Determine multiple metrics using Confusion Matrix
- ❑ Plotting the ROC curve
- ❑ Finding Optimal Cut-off Point
- ❑ Making predictions on the test set
- ❑ Validation of the Model for Train and Test data and arrive at the results

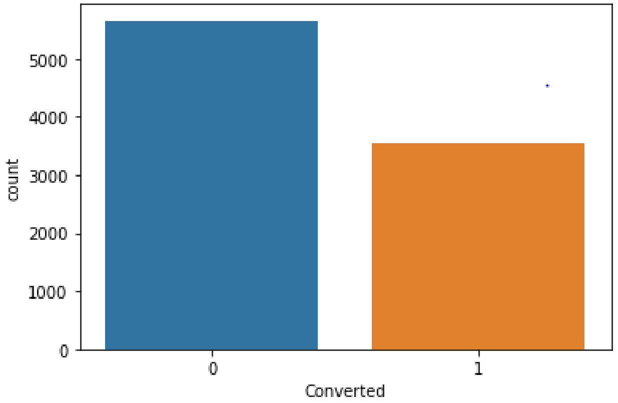
Exploratory Data Analysis of Outlier and Target variables

As per below screen shot, it's evident that the variables 'Total Visits' and 'Page Views Per Visit' are the two variables that have outliers. As the number of outlier records were low, we removed the outlier records

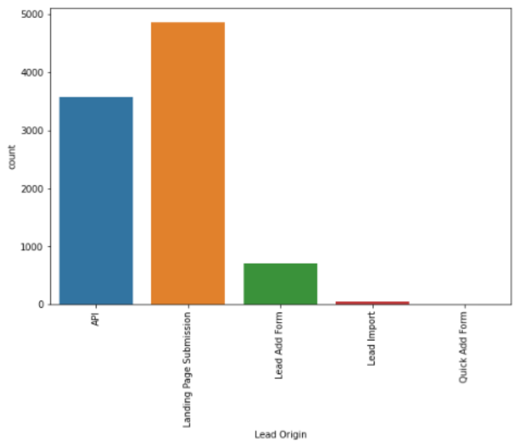


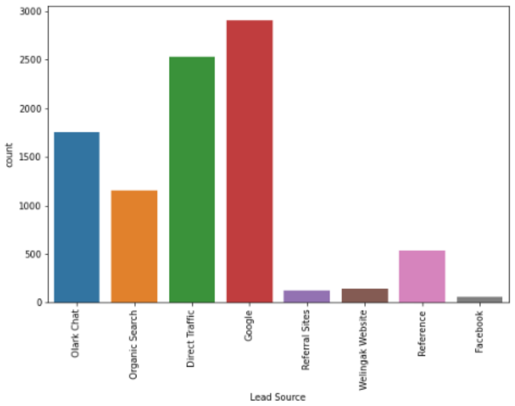


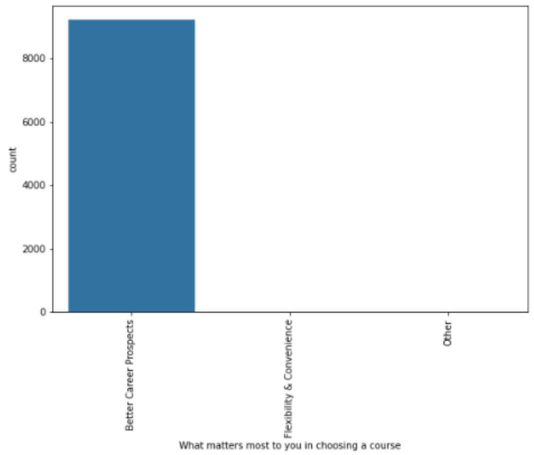
The below graph shows the data of historically, how many Customers have converted successfully

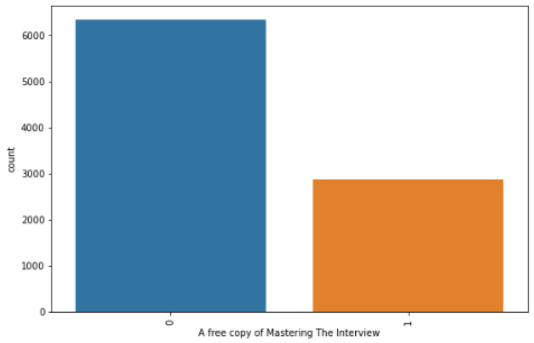


Exploratory Data Analysis of Other Categorical variables



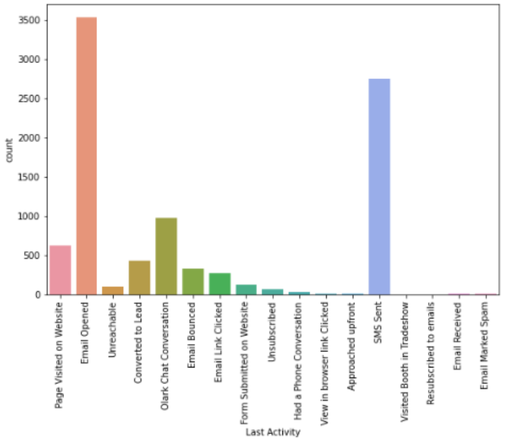


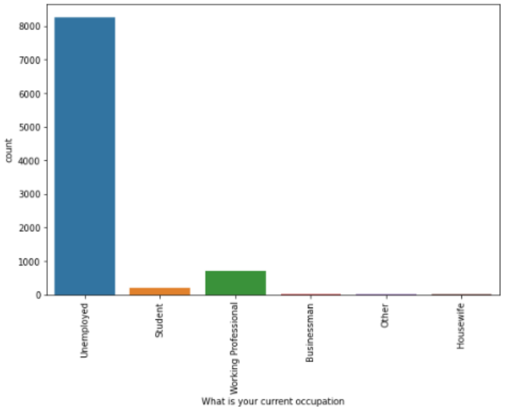


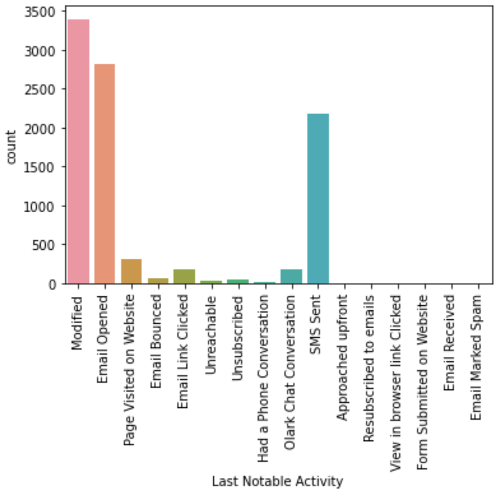


Plotting bar graph for these 7 categorical variables helped us in identifying the distribution of values. **Key observations are:**

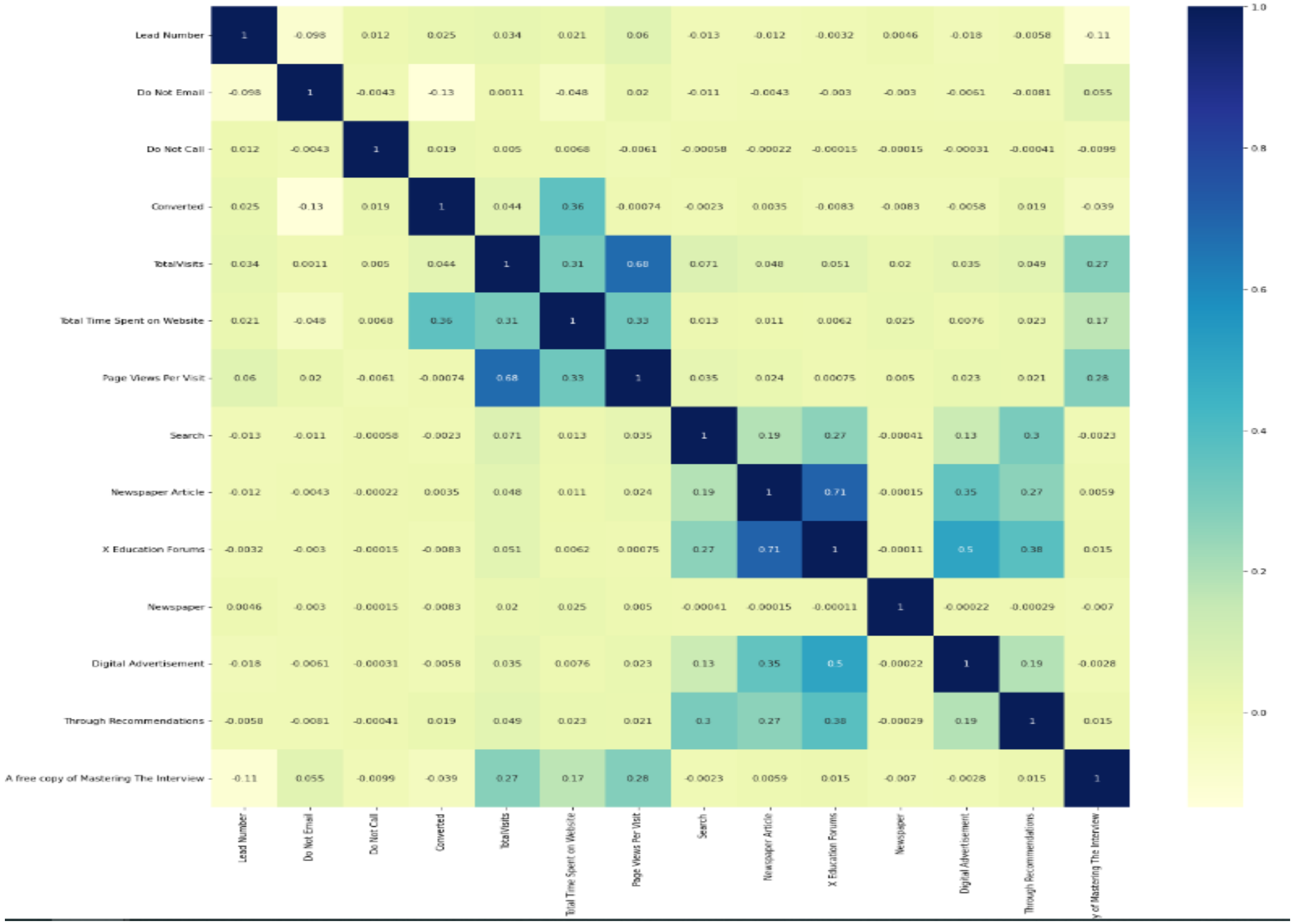
- Lead Origin mainly has two values API and landing Page Submission
- Better career prospects is what customers are looking for
- Most common Last Activity of customers are 'Email Opened' and 'SMS Sent'
- Most of the customers are Unemployed
- Last notable activity are 'Modified', 'Email Opened' and 'SMS Sent'







Exploratory Data Analysis of Numerical variables



By plotting heat map for all the numerical variables, below is some key observations:

- ‘Total Visits’ and ‘Page Views Per Visit’ are two variables that are correlated but as the correlation co-efficient is 0.68 and not very strong, we will retain both the features
- Similarly, ‘X Education Forums’ and ‘Newspaper Article’ are correlated, but as the correlation co-efficient is 0.71, we will retain both the features

Model Building Summary

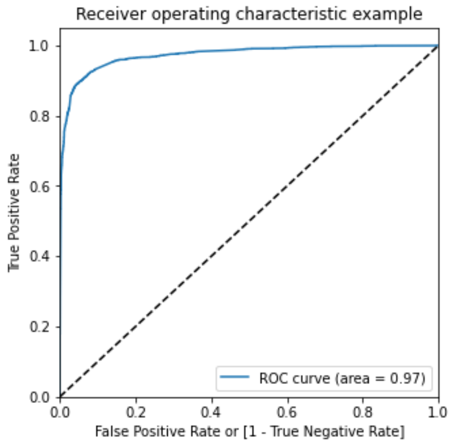
❑ Below are the steps followed to arrive at the final model:

- Used 70:30 split to create Train and Test data
- Standard Scaler method was used for scaling of all the variables
- As the no. of variables were 113, hence used RFE method to reduce the independent variables to 20
- Used GLM (Generalized Linear Regression Model) available within stats models library to build the models
- Initially, used the approach of p-value to eliminate the independent variables. The p-value greater than 5% was considered for elimination
- For all variables that have p-value of less than 5%, VIF (Variance Inflation Factor) method was used to check the multicollinearity. Any variable that had a VIF value of 5 was considered for elimination
- Using Confusion Matrix, derived some of the important values like Accuracy Score, Sensitivity and Specificity. After the final model was developed, compared the values of three parameters for both the Train and Test models
- ROC (Receiver Operating Curve) was plotted, and optimal cut-off value was determined by plotting graph for Accuracy Score, Sensitivity and Specificity for various probabilities. The graphs are shown in the next slide
- Based on the above cut-off, predictions were made to the test data set

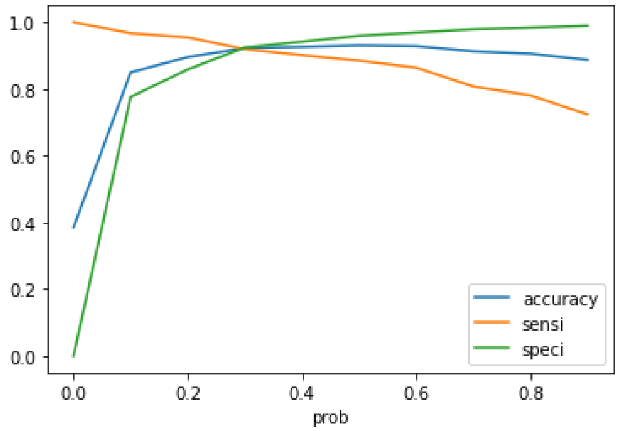
ROC and Optimal Cut-off graph

An ROC curve demonstrates several things:

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity)
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test



Determining Optimal Cut-off value: By plotting the graphs for Accuracy Score, Sensitivity and Specificity against Probability, the below graph helps in finding the cut-off value for the predictions of target variable



Conclusion

- Using cut-off value of 0.29, the target variable values for the train and test data sets were identified
- Below is a screen shot of the values of Accuracy Score, Sensitivity and Specificity for the Train and Test data sets. As the values are almost similar, we can consider the final model developed to predict the Lead conversions for the Education company with good accuracy

Metrics Comparison Train vs Test Data

Metrics	Train Data	Test Data
Accuracy	92.22%	92.00%
Sensitivity	91.97%	91.30%
Specificity	92.37%	92.44%

THANK YOU