Lead Scoring Case Study Summary

Problem Statement:

X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines.
X Education needs help in selecting the leads that are most likely to convert into paying customers. The typical lead conversion rate is 30% and the CEO has given target to enhance lead conversion rate to around 80%.

Our Goals of the Case Study:

- To 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.
- To adjust to if the company's requirement changes in the future so you will need to handle these as well.

Solution Approach:

The given problem is the classification problem, hence we decided to do logistic Regression to calculate the Lead rate.

The steps followed for the solution are as follows:

1. **UNDERSTANDING AND INSPECTING DATA:**
   Here we took Look and feel of the data, we observed following things:
   - Number of rows and columns
   - Column wise information
   - Statistical Summary of the dataset

2. **CHECKING DATA QUALITY AND DATA CLEANING:**
   - Data types of each columns
   - Checked first few rows how data looks & how the data is spread.
   - Checked and dropped duplicates. Also cross checked that after drop the shape of dataset is same.
   - For categorical variables, imputed select entry with null value.
   - Checked & Treated Null/Missing values. Dropped columns having missing values greater than 30%.
   - Analysed columns with missing values less than 30% and took actions like imputing, replacing etc.
   - Verified imputation by plotting graphs.
   - Checked data balance and found, that columns 'Do Not Call','Search','Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper, 'Digital Advertisement', 'Through Recommendations', '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' are having highly imbalance data or skewed

data and these variables will not contribute significantly to analysis, so it was better to drop these columns.

3. EXPLORATORY DATA ANALYSIS (EDA):

- Univariate Analysis:
  - ✓ 'Converted' is our target variable.
  - ✓ Calculated Conversion Rate (Resulted in 38.5%)
  - ✓ Checked ratio of imbalance (Resulted in 0.63)
  - ✓ We observed from value count and count plot that data is properly balanced with respect to ratio.
- Analysed categorical variables:
  - ✓ We observed & calculated from the plot that:
    a. Conversion rate for 'API' is ~ 31% and for 'Landing Page Submission' is ~36%.
    b. For 'Lead Add Form' number of conversion is more than unsuccessful conversion.
    c. Count of 'Lead Import' is lesser.
  - ✓ Clubbed lower frequency values together under a common label 'Others'
- Analysed categorical variables using box plot:
  - ✓ Analysed & Treated outliers.
- Bivariate Analysis:
  - ✓ Heat map to understand the attributes correlation.
  - ✓ Observed that 'TotalVisits' and 'Page Views per Visit' are highly correlated with correlation of 0.72 and 'Total Time Spent on Website' has correlation of 0.36 with target variable 'Converted'.
  - ✓ Plotted "Total Time Spent on Website" and “Total Visits” vs Converted variable to check data distribution.

4. DATA PREPARATION:

- Created Dummy Variables
- Dropped repeated columns for which dummy variables were created.

5. TEST-TRAIN SPLIT:

- Putting feature variable to X and y.
- Split the data into train and test of 70:30 ratio.

6. FEATURE SCALING:

- Used 'StandardScaler' here for scaling.
- Created 'scaler' object for 'StandardScaler'.
- Checked the correlation matrix.
- OBSERVATION: The heatmap clearly shows which all variable are multicollinear in nature, and which variable have high collinearity with the target variable.
  We will refer this map for building the logistic model so as to validate different correlated values along with VIF & p-value, for identifying the correct variable to select/eliminate from the model.
  Hence from above heatmap we can see that:
  a. 'Lead Source_Facebook' and 'Lead Origin_Lead Import' having higher correlation of 0.98.

b. 'Do Not Email' and 'Last Activity_Email Bounced' having higher correlation.
c. 'Lead Origin_Lead Add Form' ,'Lead Source_Welingak Website', 'Last Activity_SMS Sent' and 'What is your current Occupation_Working Professionals' having positive correlation with our target variable 'Converted'.
d. 'Lead Origin_Lead Add Form' and 'Lead Source_Referance' having higher correlation of 0.85.
e. 'TotalVisits' and 'Page Views Per Visit' having correlation of 0.72.

7. MODEL BUILDING:

- Model 1: Started with all the columns selected by RFE.
  - ✓ Dropped 'What is your current occupation_Housewife' because of insignificant variable p-value=0.999(p>0.05)
- Model 2:
  - ✓ Dropped 'Last Activity_Others' because of p-value=0.01
- Model 3:
  - ✓ We observed that for 'logm3' P-values of variables are significant and VIF values are below 3.
  - ✓ So not need to drop any more variables and we proceed with making predictions using this model only considering model 'logm3' as final model.

8. Finding the Optimal Cutoff Point:

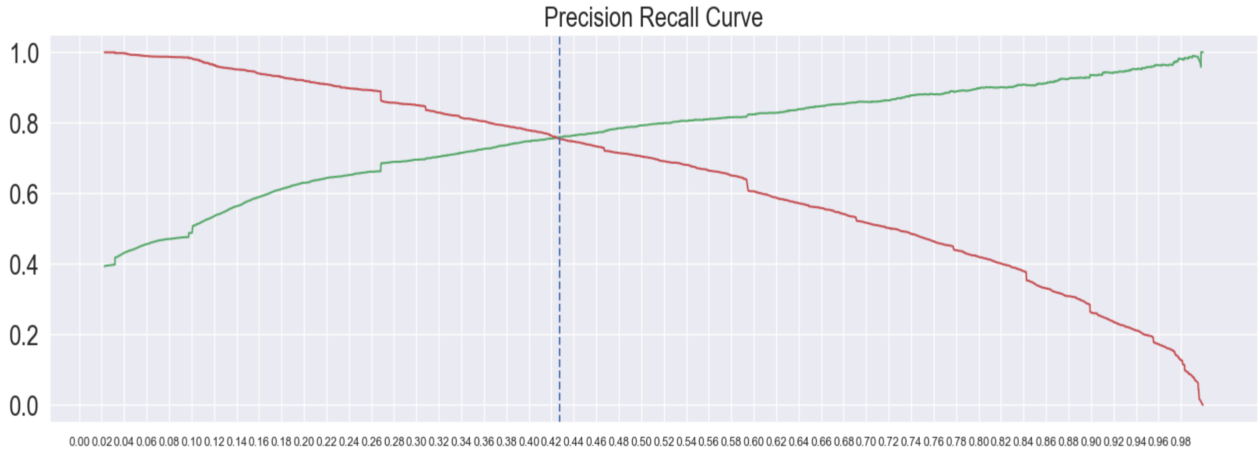
- Created columns with different probability cutoffs.
- Plotted accuracy, sensitivity and specificity for various probabilities and found that 0.358 is optimal cutoff point to take.
- Assigned the lead score to the leads based

9. MODEL EVALUATION using confusion matrix:

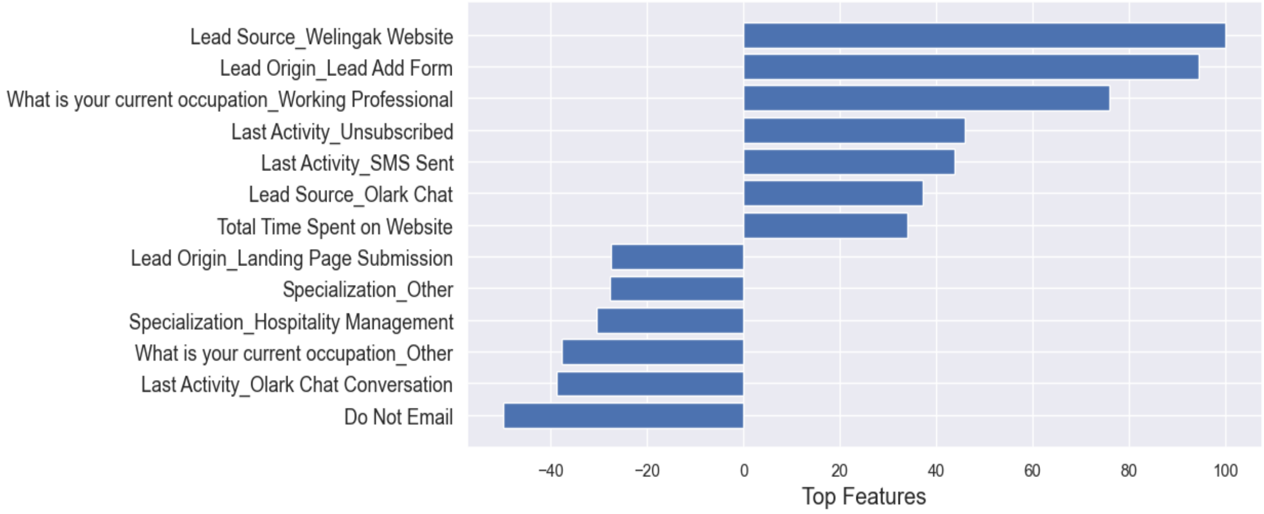
Observed following values for the Train Data:
Accuracy : 80%
Sensitivity : 80%
Specificity : 81%
Precision: 72%
Recall: 80%

10. Plotting the ROC Curve:

- From 'precision_recall_curve' we saw that cutoff point is 0.427.



- Determined top feature based on final model
- Got a relative coeffient value for all the features wrt the feature with the highest coefficient
- Plot the feature variables based on their relative importance



11. FINAL MODEL LINE EQUATION:

Converted = 0.261843 + 3.15 X Lead Source_Welingak Website + 2.98 X Lead Origin_Lead Add Form + 2.39 X What is your current occupation_Working Professional + 1.45 X Last Activity_Unsubscribed + 1.38 X Last Activity_SMS Sent + 1.17 X Lead Source_Olark Chat + 1.07 X Total Time Spent on Website - 0.87 X Lead Origin_Landing Page Submission - 0.87 X Specialization_Other - 0.96 X Specialization_Hospitality Management - 1.19 X What is your current occupation_Other - 1.22 X Last Activity_Olark Chat Conversation.

12. FINAL OBSERVATION:

- Evaluation Metrics for the train Dataset:
  A. Accuracy :0.80
  B. Sensitivity:~0.80
  C. Specificity:0.81
  D. Precision: 0.72
  E. Recall: 0.80

- Evaluation Metrics for the test Dataset:
  A. Accuracy : 0.80
  B. Sensitivity: ~ 0.80
  C. Specificity: 0.80
  D. Precision: 0.72
  E. Recall: 0.80

13. RECOMMENDATION:

To improve the potential lead conversion rate X-Education will have to mainly focus important features responsible for good conversion rate are :

- Lead Source_Welingak Website : As conversion rate is higher for those leads who got to know about course from 'Welingak Website',so company can focus on this website to get more number of potential leads.
- What is your current occupation_Working Professional : The lead whose occupation is 'Working Professional' having higher lead conversion rate ,company should focus on working professionals nad try to get more number of leads.
- Lead Origin_Lead Add Form: Leads who have engaged through 'Lead Add Form' having higher conversion rate so company can focus on it to get more number of leads cause have a higher chances of getting converted.
- Last Activity_SMS Sent: Lead whose last activity is sms sent can be potential lead for company.
- Total Time Spent on website: Leads spending more time on website can be our potential lead.