Lead Scoring Case Study - Summary Report

Problem Statement

To help X Education to select the most promising leads (Hot Leads), i.e. the leads that are most likely to convert into paying customers.

- To build a logistic regression model to assign a lead score value between 0 and 100 to each of the leads which can be used by the company to target potential leads.
- The objective is thus classified into the following sub-goals:
- Create a Logistic Regression model to predict the Lead Conversion probabilities for each lead.
- Decide on a probability threshold value above which a lead will be predicted as converted, whereas not converted if it is below it.
- Multiply the Lead Conversion probability to arrive at the Lead Score value for each lead.

Problem Solving Methodology

- We started by first importing all the required libraries that will be used as part of lead scoring assignment. Next, we inspected the data, its data types, shape of the data frame and described its various statistics.
- Cleaning and Treatment of the data for better analysis.
    - Remove columns which has only one unique value: We dropped the following columns as they have only one unique value and hence cannot be responsible in predicting a successful lead case –'Magazine', 'Receive More Updates About Our Courses' , 'Update me on Supply Chain Content' , 'Update me on Supply Chain Content' and 'I agree to pay the amount through cheque'.
    - Removing rows where a particular column has low missing values. 'Lead Source' is an important column for analysis. Hence all the rows that have null values for it were dropped.
    - Imputing NULL values with Median: The columns 'TotalVisits' and 'Page Views Per Visit' are continuous variables with outliers. Hence the null values for these columns were imputed with the column median values.
    - Removing columns which had all unique values.
    - Imputing NULL values with Mode. The columns 'Country' is a categorical variable with some null values. Also, majority of the records belong to Country 'India'. Thus, imputed the null values for this with mode (most occurring value). Then binned rest of category into 'Outside India'.
    - Binning of categorical variables to reduce the number of dummy values.
- After cleaning the data, we then moved towards Exploratory Data Analysis (EDA) to understand the data trends of each variable in our data set.
- Outlier Analysis on continues variables like 'TotalVisits', 'Total Time Spent on Website' and 'Page Views Per Visit'. Outlier treatment was performed using IQR method.
- We also converted the binary variables (yes/no) variables to (1/0) for our analysis.

- Dummy variables were created for all categorical variables and original variables were dropped.
- Splitting of data into Train-Test data sets.
- Feature scaling of continues variables using Standard Scalar method.
- Applying Recursive Feature Elimination (RFE) to identify the best performing subset of features for building the model. We filtered 15 most important features from our dataset.
- Building the model with features selected by RFE. Eliminate all features with high p-values and VIF values and finalize the model.
- Perform model evaluation with various metrics like sensitivity, specificity, precision, recall, etc.
- Decide on the probability threshold value based on Optimal cut-off point and predict the target variable value.
- Use the model for prediction on the test dataset and perform model evaluation.
- Lead Score was calculated for all the leads in the original dataset. The higher the value of the lead score the higher the chances of converting the lead and vice versa.
- Formula for Lead Score calculation is: Lead Score = 100 * Conversion Probability
- We then identified the top 3 variables that are most important for our lead conversion case study. Tags_Closedby Horizzon, Tags_Lostto EINS, Lead Source_WelingakWebsite.

Challenges in solving the Lead Score Case Study

Preparing the data for actual analysis was challenging some of these challenges and the solution approaches are mentioned below

- As the number of variables in the data set provided is 17 which is significantly high for understanding the data trends and performing EDA on each variable was challenging.
- Treating of null or missing data was a challenging task. Different strategies were used to treat the missing values.
    - Variables with 50% and more missing values were dropped.
    - Some continuous variables were imputed with median values.
    - Some categorical variables were imputed with mode values.
    - For some categorical variables we created new sub categories to identify the missing or unknown data.
- Performing Univariate and Bivariate Analysis on all both continuous and categorical variables.
    - Univariate analysis was performed on all the important variables after the unnecessary variables were removed.
    - Bivariate analysis was challenging as the heatmap plotted with so many variables was not at all readable. Hence, the heatmap was plotted after the RFE was done and top 15 variables were identified.
- Decision to remove unnecessary columns. Since the number of columns in the data set was significantly high, we had to remove some columns so that the decision making is possible. Various strategies were used to drop the columns

    - Columns with all unique values were dropped as they do not contribute much to our analysis.
    - Columns with only 1 value in the entire dataset were also removed as they do not contribute much to our analysis.
    - Some columns were highly correlated to each other like the 'Asymmetrique Activity Index', 'Asymmetrique Activity Score' and 'Asymmetrique Profile Index', 'Asymmetrique Profile Score'. We dropped 'Asymmetrique Activity Score' and 'Asymmetrique Profile Score' variables.
    - Variables with high p-value and high VIF were also dropped.
- Creating dummy variables. Since the original dataset had 17 columns and there are a lot of categorical variables with many subcategories. After creating the dummy variables, the total number of variables were more than 400. Processing 400 variables in our analysis would have been a challenging task. Also making business decisions would have been difficult.
    - To solve this challenge, we merged the subcategories in categorical variables from the data set to a new category. This reduced the number of dummy variables created.
- Dropping first variable while creating dummy variables. Initially, while creating the dummy variables, we had used the drop first feature, but this feature at times removed the most important subcategory. Hence, we manually dropped the variable which was least important for our analysis. In most of the cases we dropped the Unknown category which was created for the null or missing data.
- After identifying the top 15 variables using RFE and manual approach we removed variables which had p-value more than 0.5 and variables which had VIF value more than 5. We could only remove 2 variables out of the 15 identified using RFE. Although we wanted to reduce the number of variables further, we could not remove them as their p-value and VIF were in the acceptable range.