Lead Scoring Case Study – Methodology

Submitted by

Parijaat Sunil
Pratikshit Gaur
Shreya Pattan

Solution Methodology

We have arrived at our proposed solution using the below steps:

- **Understanding the data:** shape, data types, number of missing values
    - Available data had 37 columns and 9240 rows initially

- **Data cleaning including**
    - Dropped columns which had unique values and single value
    - Replaced 'Select' values with NaN values
    - Dropped columns with more than 40% of missing data
    - Dropped Country column as it was extremely skewed
    - Imputed values in 'What is your current occupation', 'Specialization', and 'City'
    - Dropped Tags column as the data had many ambiguous values
    - Standardizing columns having binary Yes/No data with 1/0
    - Data available after data cleaning: 37 columns and 9204 rows

- **EDA**
    - Univariate data analysis
        - Bar Graph – Converted variable
        - Box Plot – Total time spent on website
    - Bivariate data analysis
        - Comparison against converted variable
            - Bar Graph
            - Box Plot
    - Multivariate data analysis
        - Heatmap

- **Data preparation**
    - Creation of dummy variables for categorical columns
    - Creation of train test split (used 75% of data for train and 25% of data for test)
    - Usage of standard scaler to standardize the numerical data columns
    - Total size of the data after data preparation is 96 columns and 9204 rows

- **Creation of Model**
    - RFE using 15 variables
    - Manual model building – 3 Iterations

  - VIF Analysis – All columns had a VIF value lesser than 5
  - Both the p-values and VIFs seem to be decent enough for all the variables
  - We can use this model to make our predictions using this final set of features.

- **Evaluation of Model**
  - Accuracy, Sensitivity and Specificity
  - ROC curve
  - Finding the optimal cutoff point
  - Precision and recall tradeoff analysis
  - Selected 0.43 as the optimal cutoff for conversion probability
  - Parameters of model on train set are
    - Accuracy: 81%
    - Sensitivity: 75%
    - Specificity: 75%

- **Final prediction on test set**
  - Parameters of model on test set are
    - Accuracy: 81%
    - Sensitivity: 78%
    - Specificity: 84%

- **Calculation of lead scores and listing of final factors**
  - Multiplication of probability value by 100 to calculate the lead score

- **Conclusion**
  - The below columns are used to predict if the lead is likely to be converted with approximately 81% accuracy. The below list is in descending order of correlation

  - Feature Correlation:
    - Lead Origin_Lead Add Form
    - Lead Source_Welingak Website
    - What is your current occupation_Working Professional
    - Last Notable Activity_Unreachable
    - Last Notable Activity_Olark Chat Conversation
    - Do Not Email
    - Last Activity_SMS Sent
    - Total Time Spent on Website
    - Last Notable Activity_Modified
    - Lead Origin_Landing Page Submission
    - City_Other Metro Cities
    - City_Thane & Outskirts