LEAD SCORING CASE STUDY – METHODOLOGY

Submitted by

Kumar Kanishka

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 with unique values and those with a single value.
  - Replaced 'Select' values with 'NaN'.
  - Removed columns with more than 40% missing data.
  - Excluded the highly skewed Country column.
  - Imputed missing values in 'What is your current occupation', 'Specialization', and 'City'.
  - Removed the Tags column due to ambiguous data.
  - Standardized columns with binary Yes/No values to 1/0.
  - Post data cleaning, dataset comprises 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**
  - Created dummy variables for categorical columns.
  - Split the data into training and testing sets (75% train, 25% test).
  - Applied standard scaling to numerical data columns.
  - After data preparation, the dataset comprises 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: 80%
    - Sensitivity: 74%
    - Specificity: 72%

- **Final prediction on test set**
  - Parameters of model on test set are
    - Accuracy: 80%
    - Sensitivity: 74%
    - Specificity: 83%

- **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 80% accuracy.

  - Feature Correlation:
    - Do Not Email -1.418059
    - Total Time Spent on Website 0.963550
    - Lead Source_Direct Traffic -0.559833
    - Lead Source_Welingak Website 2.755532
    - Last Activity_Converted to Lead -1.341549
    - Last Activity_None -1.453436
    - Last Activity_Olark Chat Conversation -1.100634
    - Lead Origin_Lead Add Form 3.439238
    - What is your current occupation_Working Professional 2.826707
    - Last Notable Activity_SMS Sent 1.483721
    - Last Notable Activity_Unreachable 1.326286