Summary of a Lead Scoring Case Study

Problem Overview:

The aim of the case study is to construct a machine learning model for an education company X education. The objective was to predict and assign lead scores based on historical data, specifically utilizing logistic regression to classify leads as converted or not, with a predicted probability of conversion.

Approach

1. Exploratory Data Analysis and Cleaning

- Statistical and visual exploration of the dataset identified outliers, data distribution, and feature redundancies
- Through a series of plot between variables and converted variables. redundant variables were eliminated
- Columns with string values like "Select" (interpreted as null values) were addressed.
- In 2 variables, categories with very low representations were clubbed into others
- Columns with over 40% null values were removed.
- Outliers from two numerical variables were statistically removed.
- Rows with missing values were dropped, resulting in 98.2% of the original dataset for modeling.

2. Data Preparation

- Categorical variables were transformed into numerical data using dummy variables.
- The dataset was split into training and testing sets (70:30 ratio).
- MinMax scaling standardized data points to avoid bias from variables in higher scales.

3. Model Building

- An initial model was constructed using 15 variables through Recursive Feature Elimination (RFE) technique.
- Insignificant variables were removed, and Variance Inflation Factors (VIFs) were checked for multicollinearity.
- Optimal cutoff probability (0.34) was determined based on accuracy, sensitivity, and specificity.
- Model evaluation included checking performance measures such as ROC curve, accuracy, sensitivity, specificity, Recall, and Precision.
- Predicted values were calculated using the model, assigning lead scores as 100 multiplied by the predicted log odds.

Conclusions and Recommendations for Company Strategy

1. The model evaluation indicated acceptable accuracy, precision, and recall parameters, aligning with business needs.

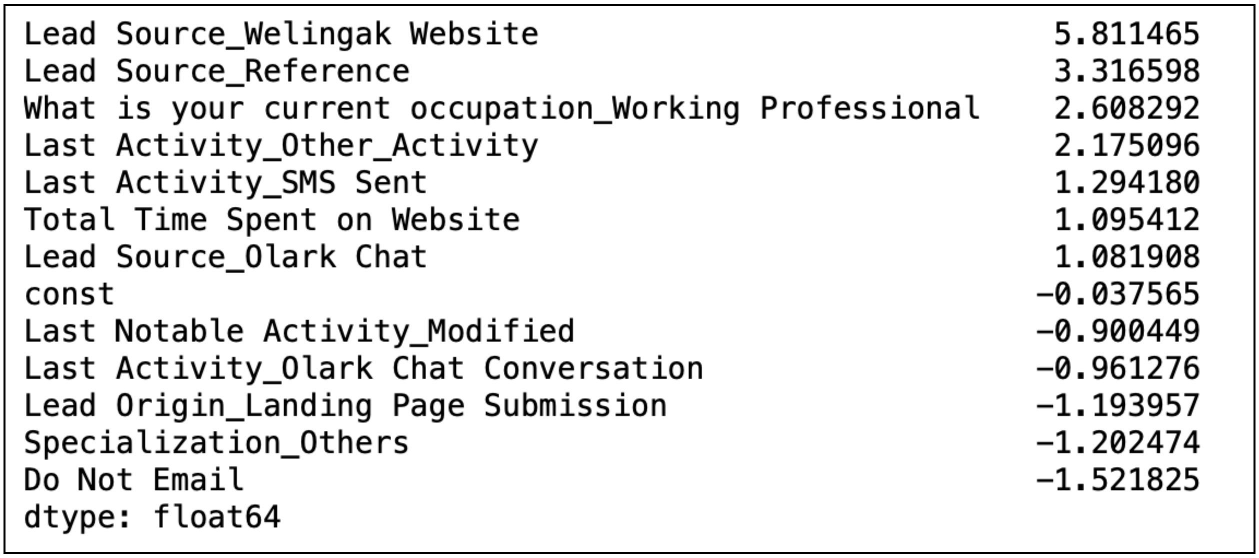
2. Top Features for Conversion Rate

- Lead Sourced from Welingkar Website, references and Olark Chat
- Whether person is working professional or not
- Last activity being SMS sent or other category

3. Recommendations

- Increasing the marketing budget for Lead Source_Welingak Website and Reference is advised, as these significantly influence lead scores
- Caution against investing excessive time in factors with minimal or negative impact on lead scores

Value of Coefficients of variables



```
Lead Source_Welingak Website                          5.811465
Lead Source_Reference                                 3.316598
What is your current occupation_Working Professional  2.608292
Last Activity_Other_Activity                          2.175096
Last Activity_SMS Sent                                1.294180
Total Time Spent on Website                           1.095412
Lead Source_Olark Chat                                1.081908
const                                                -0.037565
Last Notable Activity_Modified                       -0.900449
Last Activity_Olark Chat Conversation                -0.961276
Lead Origin_Landing Page Submission                  -1.193957
Specialization_Others                                -1.202474
Do Not Email                                         -1.521825
dtype: float64
```