LEAD SCORING CASE STUDY

SUMMARY

This is a brief summary of how we proceeded with the assignment and the insights that we gathered:

Import Libraries:

Imported the necessary libraries, like numpy & pandas for data manipulation, scikit-learn for modeling, and matplotlib & seaborn for visualization.

Loading and understanding the Data:

Loaded the data into Python with pandas

Got an overview of the data by examining its structure, size, and data types.

Computed basic statistics like mean, median, standard deviation, and percentiles.

Treating missing values and outliers:

Checked for missing values and treated the same by imputing, deleting, etc.

Identified and handled outliers by capping the data at 95 percentiles for analysis.

Data Visualization:

Created visualizations such as count plots and box plots to explore the distribution of data and relationships between variables.

Explored the relationships between variables using pair plots and heatmaps.

Data pre-processing:

Once missing values and outliers are dealt with, univariate and bivariate analysis was done on the data.

Columns which did not provide any useful insights, columns with unique values like Prospect ID, Lead number and columns which were highly skewed like City, Country, etc. were dropped.

Train-Test Split:

Split the data into training (70%) and testing (30%) sets to evaluate the model's performance.

Scaled the features using StandardScaler().

Created dummy variables for Categorical variables.

Model Training:

Using Recursive Feature Elimination (RFE) technique we selected 15 features.

Manual Feature Elimination was done using a logistic regression model and fit it to the training data. The model was trained until P-values and VIFs were satisfactory.

Model Evaluation:

Evaluated the model's performance on the test data by calculating accuracy, generating a classification report, and creating a confusion matrix. The values are as follows:

- Accuracy : 80.36%
- Sensitivity :80.29%
- Specificity : 80.41%

Model Interpretation:

Interpret the model coefficients and feature importance to understand which features are influential in making predictions.

Use the Model for Predictions:

Once the model's performance becomes satisfactory, we used it to make predictions on Test data. The values are as follows:

- Accuracy: 81.08%
- Sensitivity: 75.12%
- Specificity: 84.48%

Insights gathered from this Logistic regression model:

The important features that contribute to the probability of a lead being converted are:

- Lead Origin: Lead Add Form
- What is your current occupation: Working Professional
- Lead Source_Welingak Website

These features can be used to target marketing campaigns more effectively. For example, if a company is trying to sell a product or service to working professionals, they may want to focus their marketing efforts on leads that have those characteristics.

- The Sales team should focus on leads originating from ‘Add Forms’.
- They should focus on leads sourced from Welingak Website.
- They should focus on targeting Working Professionals
- Students can be approached, but they will have a lower probability of converting since they are already studying and would not be willing to enroll into a course specially designed for working

professionals, so early in the tenure. However, this can also be a motivating factor to ensure industry readiness by the time they complete their education.

- They should focus on targeting the people who spend more Time Spent on the Website.
- They should not focus on unemployed leads. They might not have a budget to spend on the course
- Do not focus on leads who are on ‘Do not email’ lists as they have a very low probability of converting.

This information can be used to improve the conversion rate of a company's marketing campaigns.