Summary - Process & Learnings

We followed the steps of the CRISP-DM framework for solving the lead scoring case study, as summarised below.

Business Understanding

Process
1. Analysed the business problem and developed the current and required marketing-sales funnels
2. Identified that the goal is to build a logistic regression model to score all incoming leads and classify them as ‘Hot’ or ‘Cold’ before they are sent to the sales team

Learnings
1. How to convert a business problem to a data science problem
2. Role of data science in business

Data Understanding

Process
1. Reviewed the data dictionary
2. Loaded the data into a dataframe and inspected rows and columns

Learnings
1. Nature of data available in the real-world
2. Saw that the dataset contained variables that were
    - relevant to lead scoring
    - not relevant to lead scoring
    - not available for lead scoring

Data Cleaning & Preparation

Process
1. Removed all columns with more than 45% missing values, checked for duplicate rows and checked descriptive statistics of numerical variables
2. Printed and inspected unique values of categorical variables using DataFrame.value_counts(), which surfaced a number of anomalies
3. Fixed anomalies by (i) imputing ‘NaN’ for the value ‘Select’, (ii) dropping columns with more than 45% missing values, (iii) dropping columns with low variability, and (iv) imputing remaining missing values
4. Grouped levels under the category ‘Other’ in categorical variables with many levels and dropped variables like ‘Tags’
5. Treated outliers and performed univariate/bivariate analyses using data

visualisation

6. Encoded categorical variables with dummy variables, split the dataset into training and testing sets, and scaled numerical variables

Learnings

1. How to identify anomalies in real-world data sets
2. Saw that missing values can have placeholders such as 'Select', 'Missing', etc.
3. How to group multiple levels in categorical variables
4. Importance of scaling to ensure that no variable has an outsized impact on the model

Model Building

Process

1. Built a preliminary logistic regression model using all 32 predictor variables and then performed recursive feature elimination to select 20 predictors.
2. Then performed manual elimination by removing variables with large p-values and variance inflation factors, finally resulting in 17 predictors
3. Plotted an ROC curve (AUC score = 0.89) and identified the optimal probability threshold as 0.4 by the precision-recall view

Learnings

1. Importance of recursive feature elimination, which saves a lot of time by removing low impact variables
2. Area under the ROC curve as a good indicator of model robustness

Model Evaluation

Process

Calculated metrics such as accuracy, sensitivity, specificity, precision and recall for both training and test sets

Learnings

1. How to calculate important metrics and evaluate the model on the test set
2. Saw that our model had very similar values on both datasets, indicating robustness

Model Interpretation & Lead Score Assignment

Process

1. Derived the equation of logistic regression and interpreted some of the coefficients
2. Assigned lead scores to all of the 9240 leads and filtered hot leads (about 3603 i.e. almost 40%)

Learnings

1. How to interpret a logistic regression model using log-odds and coefficients of predictors
2. How to calculate logistic regression probabilities