Machine Learning 1

Lead Scoring Case Study - Report

**A brief summary report in 500 words** explaining how we proceeded with the assignment and the learnings that we gathered **submitted by** Janarthanan Balasubramanian & Siva Prakash.

# 

# Solution Approach

**1. Reading and Understanding the Data:** There are 36 features and the target variable. Loaded the data into a pandas dataframe and checked its shape, type of variables, percentage of nulls and unique levels in categorical variables.

**2. Data Visualization:** To get more insights, we created visualizations:

* Categorical Variables (univariate count plot, bar plot segmented on target variable)
* Binary Variables (categorical variables with 2 Levels)
* Numeric Variables (histogram, box plot and correlation matrix)

**3. Data Clean Up:**

* Some categorical variables had a level ‘Select’ (customer did not select any value). The value is missing and hence updated with null values.
* **Dropped the columns** that cannot be used in the model:
  + Columns with more than 45% null values
  + Binary variables with no variance (will not contribute to the model)
  + Categorical variables that are skewed (will adversely affect the final model)
  + Tags: status update by the callers; the objective is to reduce the number of calls.
* **Data Imputation:** For the remaining columns we imputed the null values appropriately (with median for numeric columns; ‘Unknown’ for categorical variables).
* **Reduce the number of levels:** If there are levels with very less number of records, we marked them as ‘Others’ or ‘Unknown’ (later dropped while creating dummies)

**4. Data Preparation Steps:**

* Label encoding (mapping binary levels to 0 or 1)
* Creating dummies for categorical variables (only n-1 levels)
* Splitting the data into train and test set (70:30 ratio)
* Rescaling the numeric data using Standard Scaler.
* Check for any multi-collinearity with these new features.

**5. Building the Model:** We had around 70 features, which was reduced to 15 using RFE (recursive feature elimination). Out of 15, one more feature were removed by building the model and observing the p-value (> 0.05). The final model (with p-value < 0.05 and VIF < 5) had 14 features.

**6. Model Evaluation:**

* Predict the probability with the final model.
* Plot the **ROC curve** with different probabilities. The area under the curve > 0.8.
* Compute the **metrics** (accuracy, specificity, sensitivity) for different probabilities.
* Plot the metrics to find the intersecting point that can be chosen as **optimal cut off**.
* With the optimal cutoff (0.34), updated the **predicted value** (1 or 0) for target variable.
* Validate it with the **Precision vs. Recall** Tradeofff.

**7. Predicting for Test Data**

* With the final model predict the probabilities on test data.
* Based on cutofff value, update the predicted value for target as 0 or 1.
* Create the confusion matrix and compute the metrics to ensure that the model having more than 80% accuracy, sensitivity and specificity.

We assigned a **lead score** between 0 and 100 for each leads. The likelihood of conversion as a probability is difficult to interpret. To make it more interpretable and capable of answering business questions, we computed the **log odds** and **odds**.

**Learnings:** The lead source, lead origin, occupation of the customer and the time they spend on the website are significant features that need to be focused on to improve the conversion rate.