Case Summary

Starting with hot leads, or leads with a high likelihood of conversion, is an effective strategy to work on the leads. This will increase conversion rates and make efficient use of time. While nurturing time for leads with low scores (cold leads) should be lowered, nurturing time for hot leads can be enhanced.

A logistic regression model can be used to distinguish between hot and cold leads. We will create a logistic regression model using the meta data offered for each lead in order to assign a lead score to each lead.

I. Data Analysis

A. The data contains columns with higher missing percentages. Additionally, some columns have "Select" as their default value. We will first treat these as missing values and treat them similarly to other missing values.
B. Categorical columns with less than a five percent missing value will have a mode imputed to them.
C. Quantitative variables having a lower percentage of missing values will have the median imputed. According to statistical research, there is no discernible difference between the median and mean for these columns, so imputing using the median shouldn't cause any problems.
D. Categorical columns that have a missing value percentage of more than 70% will be removed.
E. Since imputed values can cause the data to be inflated, additional missing values will be recognised as such.

II. Data Preparation

A. There are outliers in the dataset, according to the boxplot and the descriptive statistics. Data loss from removing the outliers is about 9%.
B. We won't eliminate the outliers since doing so will enable us to assign a lead score to every lead. The final analysis of the model does show that the metrics (Accuracy, Sensitivity, and Specificity) are good, thus we will leave outliers in.
C. Few categorical variables/levels are crucial for lead conversion, according to a quick bivariate study. This will serve as our conclusion.
D. We will use the following approach to transform categorical data since logistic regression utilizes numerical data.
    1. Dummy variables: Dummy variables will be used to address categorical variables with low/moderate levels.
    2. Label Encoding: In order to encode higher level variables, we shall utilize label encoding. To prevent a sharp increase in dataframe size, this is done.

E. Columns with zero variance, or columns with a single constant value, will be removed because they don't provide any data or dimensions to the model.
F. A quick heatmap shows that several factors are correlated. VIF will also be utilized while creating the model.

III. Model Building

A. We'll use both RFE and PCA to compare the two methods to see which one produces a better model because the data frame is so large.
B. We will split train and test datasets from the data. On the basis of the test dataset, we will make predictions using the training dataset.
C. To guarantee that the data is on the same scale and computationally efficient, numerical data will be scaled using a standard scaler.
D. The functions below were developed to carry out repetitious chores:
    1. Createmodel - prints the model summary, VIF, and return model after receiving a dataframe as input.
    2. Conf Scores - accuracy, sensitivity, and specificity are returned after receiving a confusion matrix as input.
    3. Calctrainseult - returns confusion metrics and scores after taking cutoff as an input (accuracy, sensitivity, specificity)
E. RFE:
    1. To begin modeling, we will utilize RFE to choose the top 20 variables.
    2. The parameters used to fine-tune the model (i.e. dropping variables)
        a) High p-value (variable not significant).
        b) High VIF (high collinearity with another variable).
    3. ROC and AUC confirms that we have a decent model.
    4. To determine the best cutoff value, we shall utilize the approach below:
        a) Plot specificity, sensitivity, and accuracy
        b) Plot recall, precision
        c) Finding the best cutoff is crucial since sensitivity and specificity must be balanced.

IV. Predications

A. To make predictions on the test dataset, use model6 and the best cutoff.

V. Use PCA

A. To determine whether this produces a better model, we will utilize PCA.
B. PCA helps reduce dimensionality and addresses the multicollinearity problem.

C. Making predictions using a model created using PCA produces respectable results, but the challenge is low.
   1. Score lower for models created without PCA.
   2. Define unique variables or circumstances that contributed to a high score.

VI. Model Selection and Lead Score

A. For the final forecast, use the model created using the rfe approach.

B. This model provides the best results and is simple to suggest and comprehend.

C. Each lead will be given a lead score based on the chance that the model anticipated (Lead Score = Predicted Probability * 100).

D. Create a dataframe and visualize conversion Vs cutoff.

VII. Conclusion

A. Top three features that influence choice:
   1. Tags
   2. Lead Quality
   3. Asymmetrique Profile Index

B. Top three categories that influence choice:
   1. Lead Origin ==> Landing Page Submission
   2. Lead Origin ==> Lead Add Form
   3. Lead Source ==> Olark Chat

VIII. Learnings

A. Before developing a model, EDA is a crucial step. EDA's key insights aid in handling data appropriately.

B. Cleansing data facilitates the development of effective models. To guarantee that the quality of the data is not affected, procedures including missing value imputation, scaling, and outlier treatment must be carried out.
   1. Missing value - Columns with a greater proportion of missing values can be eliminated, but columns with a lower percentage can be imputed.
   2. Outlier treatment - An outlier might affect the model, making it less useful. The treatment of outliers should thus be done with caution. It is important to take precautions to prevent a major loss of data from happening as a result.
   3. Scaling - To make sure that quantitative columns are on the same scale, they should be scaled.

C. RFE is an effective approach for locating critical characteristics so that model construction may begin. On the other side, by creating additional principal components, PCA aids in dimensionality reduction.
D. Functions that carry out recurring tasks can aid in the development of modular code. This aids the code's ability to be reused.
E. Knowing how to balance sensitivity and specificity is essential for choosing the best cutoff for the mode.
F. Confusion metrics are a helpful way to gauge how well a model is working. One may obtain accuracy, sensitivity, and specificity using confusion measurements.