SUMMARY

Data Cleaning:

- Null values above 40% in a column were removed.
- Value counts in categorical columns were examined to determine the best course of action: if imputation results in skew, the column was discarded; otherwise, a new category called "others" was created; high frequency values were imputed; and irrelevant columns were removed.
- Numerical categorical data were imputed using the mode, and columns containing a single unique customer response were excluded.
- Additional tasks included handling outliers, resolving inaccurate data, organizing low frequency values, and mapping binary category values.

EDA:

- Data imbalance was examined; only 38.5% of leads were converted.
- Analyzed numerical and categorical variables using univariate and bivariate methods. "Lead Origin," "Current occupation," "Lead Source," and so on offer insightful information about the impact on the target variable.
- Website time has a beneficial effect on lead conversion.

Data Preparation:

- generated (one-hot encoded) dummy features for categorical variables
- Divided the train and test sets into a 70:30 ratio
- Used standardization for feature scaling
- Removed a few columns that had strong correlations with one another

Model Building:

- Reduced variables from 48 to 15 using RFE. Dataframe will become easier to handle as a result.
- To construct models, the manual feature reduction approach involved removing variables whose p-value was more than 0.05.
- Before the final Model 4, which was stable with p-values < 0.05, a total of three models were constructed. With VIF < 5, there is no indication of multicollinearity.
- With 12 variables, the final model, logm4, was chosen, and it was utilized to make predictions for both the train and test sets.

Model Evaluation:

- A confusion matrix was created, and based on the accuracy, sensitivity, and specificity plots, a cutoff point of 0.345 was chosen. This cutoff resulted in approximately 80% accuracy, specificity, and precision. In contrast, the precise recall view provided almost 75% fewer performance metrics.
- The CEO requested that we increase conversion rate to 80% in order to solve the business challenge; nevertheless, metrics decreased when we adopted a precision-recall viewpoint. Thus, the sensitivity-specificity view will be our best cut-off for our final forecasts.
- A cutoff of 0.345 was used to assign the lead score to the training set.

Making predictions on test data:

- Making Test Predictions: Using the finished model, scale and make predictions.
- The evaluation metrics for both the test and the train are very near to 80%.
- A score for lead was assigned.
  Top 3 characteristics are:
  Webpage of Welingak, the major source
  Lead Reference/Source
  Present Profession: Working Professional

Recommendations:

- On the Welingak website, more money can be spent on advertising and other things.
- Rewards/discounts for supplying references that result in leads; promote additional reference provision.
- Targeting working professionals is recommended because they have a greater conversion rate and can afford to pay higher costs.