Lead Score Case Study

A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.

Solution Summary

1. **Reading and Understanding Data**
    - Import and read the dataset
    - Load the data with 9240 rows and 37 columns
    - Inspect statistical aspects
    - Verify data type of each column

2. **Data Cleaning and Data Manipulation**
    - Check if 'Prospect ID' and 'Lead Number' has any duplicates or not
    - Check number of missing values in each column
    - Dropped column which has more than 40% missing value but kept 'Lead Quality' feature as its important for our analysis.
    - Find columns with 'Select' value and replace them with null value.
    - Find columns more than 15% missing value and impute them as 'Not Specified'
    - Removed rows less than 15% missing values
    - Checked values count of each object type column and eliminated irrelevant columns which has only one value.

3. **Explanatory Data Analysis**
    - Checked Lead Conversion rate which stood as 37.85%.
    - Performed Univariate, Bivariate and Multivariate Analysis using numerical and categorical variable and observed some meaningful insights.

4. **Data Transformation**
    - Identified and handled outliers
    - Converted categorical variables values (Yes/No) to binary value (1/0)
    - Created dummy variable for the categorical features
    - Removed all the repeated and redundant variables
    - Converted all numerical and categorical features into integer type
    - Drop all numerical null values after dummy variable creation
    - Split the data into 70% train and 30% test dataset
    - Used Standard Scaling to scale all the variables
    - Plotted heatmap to check correlation between all the variables

5. **Model Building**
    - We build a logistic regression model with the help of RFE with 15 features.
    - Performed X y Train splitting and selected features through feature ranking using RFE
    - Manually eliminated independent variables with high p value (>0.05) and checked VIF and eliminated high VIF value (>5).
    - Evaluated final model on the basis of significant variables and no multicollinearity.

6. **Making Prediction and Model Evaluation**
    - Train Dataset
        - Predicted train data with Actual and Predicted Converted variable and make the probability of predicted value.
        - Calculated and plotted confusion matrix, accuracy, specificity, sensitivity and ROC curve to find the optimal cut off probability.
        - Checked Precision and Recall and plotted a tradeoff between them.

    - Test Dataset
        - Transformed numeric features of the test data with Standard Scaler.
        - Performed X y Test splitting and dropped features to align with train data.
        - Evaluated the model with accuracy, sensitivity and specificity.
7. **Conclusion**
    - Train data is very close range to test data. We can conclude that model is accurate and Stable. It can change with the company's requirement in future.
    - As per Lead data, high lead score are hot leads and low lead score are cold leads.