Lead Scoring Case Study Summary

-Praveen Sankadal

1. **Data reading and understanding**
    a. Loading data and going through the columns.
    b. Checked for shape, description and indo of the data

2. **Data Cleaning**
    a. Checked for duplicates.
    b. Replaced 'Select' labels with null values.
    c. Dropped columns with over 50% null values.

3. **Data visualization**
    a. Created a function for plotting graphs.
    b. Created plots of categorical columns to geta clear understanding of data distribution.
    c. Imputed missing values with mode, replaces values with small occurrence with new category as “Others” or most repeated value.
    d. Checking for outliners by plotting a boxplot for numerical columns and handling outliners with quintiles of 95%.
    e. Mentioned observations in a well commented formate

4. **Data Preparation**
    a. Converted multicategory labels to binary variables (0 and 1).
    b. Created dummy variables for some categorical features.
    c. Split dataset into training and testing sets.
    d. Applied StandardScaler for data scaling.
    e. Utilized Recursive Feature Elimination (RFE) for model building with 15 variables.

5. **Model Building**
    a. Created function for training function called `model_build`(). Created function to calculate VIF values `calculate_vif()`. Created a common Function to calculate accuracy, sensitive, specisity.
    b. Trained the model checked p-value , Variance Inflation Factor (VIF) scores (<5.0) for multicollinearity.
    c. Removed insignificant variables based on P-Value scores.
    d. Determined optimal probability cutoff for the final model.
    e. Used Principal Component Analysis (PCA)for model selection just to compare the best feature selected between RFE and PCA. Selected RFE for better results.

6. **Model evalulation**
    a. Evaluated model performance metrics (accuracy, sensitivity, specificity).
    b. Plotted roc curve, plotted confusion matrix to calculate accuracy
    c. Plotted roc curve, plotted confusion matrix to calculate accuracy
    d. Predicted outcomes on the test set and evaluated model accuracy and sensitivity.
    e. Assigned lead scores to indicate potential hot leads.

7. **Conclusion**
    a. Test set demonstrates acceptable accuracy and sensitivity.
    b. The model exhibits stability and adaptability to changing business requirements.
    c. Identified top features contributing to a good conversion rate:
        i. Tags_Lost to EINS
        ii. Tags_Closed by Horizzon
        iii. Tags_Will revert after reading the email

Key Learnings

- Achieved model stability and adaptability in dynamic business environments.
- Identified critical features influencing lead conversion for strategic focus.
- Successfully implemented data cleaning, transformation, preparation, and model evaluation techniques.
- Utilized advanced methodologies such as RFE and probability cutoff optimization.
- Provided actionable insights to guide decision-making in lead management.