Solution Description:

The project focused on binary classification, and the following steps were undertaken to achieve the desired outcome:

1. Preprocessing: The data underwent preprocessing steps, which included handling missing values and encoding boolean variables. For variables with less than 30% missing values, median imputation was performed. For variables with more than 60% missing values, Multiple Imputation by Chained Equations (MICE) was used. Boolean variables were encoded using binary encoder.
2. Univariate Analysis: Univariate analysis was conducted to gain insights into individual variables. This step helps in understanding the distribution and characteristics of each variable in the dataset.
3. Null Value Handling: Missing values were examined, and variables with missing values greater than 60% were imputed using MICE. This approach ensures that missing values are estimated using other variables in the dataset.
4. Train-Validation Split: The data was split into training and validation sets for logistic regression modeling. This split allows evaluating the performance of the model on unseen data.
5. Missing Value Imputation: Missing values in the dataset were imputed using appropriate techniques to ensure that the model's input data is complete.
6. Modeling: Three different models, namely CatBoost, XGBoost, and LightGBM, were employed for binary classification. GridSearchCV was utilized to tune the hyperparameters of these models, enabling optimal performance.
7. Model Evaluation: The models were evaluated based on the area under the receiver operating characteristic curve (AUC). The AUC is a common metric used to measure the performance of binary classification models. The mean AUC scores obtained were 0.76 for CatBoost, 0.76 for XGBoost, and 0.775 for LightGBM.
8. Feature Importance: The feature importance of the LightGBM model was plotted to identify the variables that had the most significant impact on the classification task.
9. ROC Curve: The ROC curve was plotted for the predictions obtained from the LightGBM model, as it yielded the highest AUC score. The ROC curve is a graphical representation of the true positive rate against the false positive rate and helps in visualizing the model's performance.
10. Missing Value Threshold: Attempted to remove columns with missing values exceeding 60% but observed no significant improvement in the results. This suggests that removing these columns did not contribute significantly to the model's performance.
11. Correlated Variable Removal: Correlated variables were removed to mitigate multicollinearity issues. Logistic regression models are sensitive to multicollinearity, which can affect the interpretability and stability of the model.
12. Outlier Treatment: Outliers were addressed using binning and winsorization methods. However, these techniques did not lead to a notable improvement in the results.