Solution Description:

The project focused on binary classification, and the following steps were taken to build the solution:

1. Preprocessing: Initially, the data underwent preprocessing steps to prepare it for analysis. This may have involved handling missing values, outliers, and standardizing or normalizing the variables.
2. Univariate Analysis: A univariate analysis was performed to gain insights into the individual variables. This step helps in understanding the distribution and characteristics of each variable.
3. Handling Missing Values: The dataset was checked for missing values, and variables with null values less than 30% were processed using median imputation. For variables with null values greater than 60%, multiple imputations using the MICE method were applied.
4. Binary Encoding: Variables containing boolean values were encoded using binary encoding. This encoding technique transforms boolean variables into a binary representation, making them suitable for modeling.
5. Train-Validation Split: The data was split into a training set and a validation set to assess the performance of the models. This step helps in estimating the model's performance on unseen data.
6. Missing Value Imputation for Logistic Regression: Before applying logistic regression, missing values were imputed using an appropriate method. This ensures that the logistic regression model can handle complete data.
7. Model Selection and Evaluation: Three models, namely CatBoost, XGBoost, and LightGBM, were used for binary classification. GridSearchCV, a technique for hyperparameter tuning, was applied to each model to optimize their performance. The evaluation metric used was the area under the receiver operating characteristic curve (AUC).
   * CatBoost: Mean AUC score obtained was 0.761.
   * XGBoost: Mean AUC score obtained was 0.768.
   * LightGBM: Mean AUC score obtained was 0.775.
8. Feature Importance: For the LightGBM model, a feature importance plot was generated to identify the most influential variables in predicting the binary outcome. This plot helps in understanding the relative importance of each feature.
9. ROC Curve: The ROC curve was plotted for the LightGBM model, which yielded the highest AUC score. The ROC curve illustrates the trade-off between true positive rate and false positive rate, providing insights into the model's performance at different classification thresholds.
10. Removing Highly Missing Variables: Additionally, an attempt was made to remove variables with missing values exceeding 60%. However, this step did not yield a significant difference in the model's performance.
11. Removing Correlated Variables: Correlated variables were identified and removed before applying logistic regression. This step helps in eliminating multicollinearity, improving the interpretability of the model, and reducing overfitting.