ABSTRACT

Problem Statement:

To build a Machine Learning based Risk Prediction Model to predict the likelihood of a customer defaulting on a credit card.

Approach:

To build the risk prediction model, I will begin by exploring the dataset to identify key insights. This will be followed by data preprocessing, where I will handle missing values and identify and treat outliers to ensure data quality. Since the dataset does not contain any categorical features, encoding will not be necessary. I will then apply feature engineering and feature reduction techniques to retain the most important features and normalize the data. After preprocessing, I will split the dataset into training and testing sets. To address any class imbalance, appropriate techniques will be applied to ensure balanced model performance. I plan to build and train multiple machine learning models using the training data. These models will be evaluated and validated using the testing data based on cross-validation and metrics such as Accuracy, Precision, Recall, Confusion Matrix, Log-Loss, F1-Score, and AUC-ROC. The model that performs the best according to these metrics will be selected as the final model for predicting the likelihood of credit card default.

Planned ML Algorithms:

1) Logistic Regression
2) Random Forest
3) Support Vector Machines (SVM)

Dataset: Credit Risk Prediction.csv