Grand Project

Project Title: Credit Risk Assessment for Loan Approval

Problem Statement:

A financial institution wants to automate its loan approval process by developing a machine learning model that can accurately assess the credit risk associated with loan applicants. The goal is to minimize the risk of default while maximizing the number of approved loans, thereby improving efficiency and reducing manual intervention in the lending process.

Dataset:

Dataset Link: https://www.kaggle.com/datasets/nanditapore/credit-risk-analysis

The dataset you'll be working with contains various features related to loan applicants, including demographic information, credit history, employment details, financial indicators, and loan characteristics. The dataset also includes a binary target variable indicating whether a loan was repaid or defaulted.

Questions to Explore:

1. What are the key factors influencing credit risk in loan approval decisions?
2. How do demographic variables (e.g., age, income, education) affect the likelihood of default?
3. What role does credit history (e.g., credit score, previous defaults) play in assessing credit risk?
4. Are there specific employment or financial indicators (e.g., debt-to-income ratio, length of employment) that correlate with loan repayment?
5. Can we build a predictive model that accurately predicts the likelihood of default for loan applicants?

6. Which machine learning algorithms are most suitable for this binary classification task?
7. How can we handle imbalanced classes (i.e., more instances of non-defaults than defaults) in the target variable?
8. What preprocessing steps are necessary for data cleaning, feature engineering, and feature selection?
9. What evaluation metrics (e.g., accuracy, precision, recall, F1-score, ROC-AUC) are appropriate for assessing the model's performance?
10. How can the model results be interpreted and translated into actionable decisions for loan approval?

Project Workflow:

1. **Data Preprocessing:**
   - Handle missing values, encode categorical variables, and scale numerical features.
   - Split the dataset into training and testing sets.

2. **Exploratory Data Analysis (EDA):**
   - Explore the distribution of features and the target variable.
   - Identify correlations between features and credit risk.
   - Visualize relationships using plots and charts.

3. **Feature Engineering:**
   - Create new features or transform existing ones to improve predictive performance.
   - Consider techniques such as binning, one-hot encoding, or feature scaling.

4. **Model Selection and Training:**
   - Experiment with different machine learning algorithms such as logistic regression, decision trees, random forests, gradient boosting, etc.
   - Train multiple models using the training data and evaluate their performance using cross-validation.

5. **Hyperparameter Tuning:**
   - Use techniques like grid search or random search to optimize the hyperparameters of the best-performing models.

6. **Model Evaluation:**
   - Evaluate the trained models on the testing set using appropriate evaluation metrics.
   - Compare the performance of different models and select the best one for deployment.

7. **Interpretability and Actionable Insights:**
   - Interpret the model results to understand which features are most influential in predicting credit risk.
   - Establish decision thresholds for loan approval based on model predictions and risk tolerance.

8. **Model Deployment:**
   - Deploy the trained model into production, either as an API or integrated into the company's loan approval system.
   - Monitor the model's performance in real-time and update it as needed.

Deadline : 16th Feb 2024 , Friday 11:59 AM