**PROJECT PROGRESS REPORT**

**CREDIT SCORING: CLASSIFICATION OF CUSTOMER RISK LEVEL**

Tianxin Yin, Juehan Wang, Behzad Rezvani, Vanessa Vu

**Project Overview**

This project aims to create a classification model that categorizes customers into **low-risk** or **high-risk** groups for credit offers. The model leverages financial features such as income and savings to automate credit risk evaluation, minimizing misclassification and enhancing decision-making in lending process.

**Dataset**

The dataset used for this project comes from an open-source credit scoring dataset from Kaggle.co, which provides information about the financial and personal data of the customers. The dataset, sourced from Kaggle, contains approximately 20,000 customer entries with features such as:

* **Income**: Annual income of customers.
* **Savings**: Total amount saved.
* **Credit History**: Binary indicator of credit behavior (good/bad).
* **Loan Amount**: Amount requested.
* **Risk Label**: Target variable indicating credit risk classification.

The dataset required preprocessing to address missing values, class imbalance, and outliers. Initial analysis revealed that **income** and **savings** are key predictors of credit risk.

**Objectives**

1. Develop a robust machine learning model to classify customers as **low-risk** or **high-risk**.
2. Identify significant features influencing credit risk to provide actionable insights.
3. Optimize performance metrics (e.g., accuracy, recall) for real-world applicability

**Methodology**

**1. Data Preprocessing**

* **Cleaning**. Imputed missing values using statistical methods. Detected and capped outliers in numerical fields like income and savings.
* **Encoding**. Converted categorical variables (e.g., Credit History) into numerical form using label encoding.
* **Normalization**. Standardized numerical features to ensure consistent scaling for better model performance.
* **Balancing**. Used **SMOTE (Synthetic Minority Oversampling Technique)** to address class imbalance by generating synthetic samples for the minority (high-risk) class.

**2. Exploratory Data Analysis (EDA)**

* **Correlation Analysis**. Highlighted that **income** and **savings** strongly correlate with low-risk classification.
* **Visualization**. Scatter plots revealed trends and distributions across risk groups.

**3. Model Development**

* **Model Selection**. Created baseline models using **Logistic Regression** and **Decision Trees**. Advanced to ensemble methods like **Random Forests** and **Gradient Boosting** for improved performance.
* **Hyper parameter Tuning**. Conducted grid search to optimize parameters (e.g., tree depth, number of estimators for Random Forest).
* **Cross-Validation**. Applied 5-fold cross-validation to improve reliability and reduce overfitting.

**4. Evaluation Metrics**

* **Accuracy**: Measures overall correctness of predictions.
* **Precision & Recall**: Prioritized minimizing false negatives to avoid misclassifying high-risk customers.
* **F1 Score**: Balanced precision and recall for a comprehensive evaluation.
* **Confusion Matrix**: Analyzed false positives and negatives for targeted improvements.

**Progress**

1. **Preprocessing & EDA**:
   * Completed data cleaning, encoding, normalization, and balancing.
   * EDA confirmed **income** and **savings** as the most influential features.
2. **Model Training**:
   * Logistic Regression achieved 48% accuracy, providing a baseline.
   * Random Forests improved accuracy to **78%** after hyper parameter tuning.

**References**

1. Bensic, M., Sarlija, N., & Zekic-Susac, M. (2005). *Modelling small-business credit scoring by using logistic regression, neural networks and decision trees*. Intelligent Systems in Accounting, Finance & Management, 13(3), 133–150.
2. Hand, D. J., & Henley, W. E. (2007). *Statistical Classification Methods in Consumer Credit Scoring: A Review*. Journal of the Royal Statistical Society. Series A, 160(3), 523–541.
3. Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Elsevier.

Thomas, L. C. (2009). *Consumer Credit Models: Pricing, Profit and Portfolios*. OUP Oxford