**Project Progress Report**

**Credit Scoring: Classification of Customer Risk Level**

Tianxin Yin, Juehan Wang, Behzad Rezvani, Vanessa Vu

**Project Overview**

The main objective of this project is to create a classification model which positions a customer in either low-risk or high-risk group regarding their eligibility of a credit offer, based on financial parameters such as income or savings. Credit scoring is essential in the finance world when determining whether an organization should lend money to a customer [(Thomas, 2009)](https://paperpile.com/c/jYzUPZ/M5Xq). It also affects the amount of loan given, interest rates charged, and even the loan's credit limit. By using machine learning techniques, the project aims to automate customer classification into risk pools.

The goal of our model is to improve accuracy by minimizing false misclassifications, thereby enhancing the reliability of customers’ credit scoring. This model is expected to assist lenders in the evaluation of credit risks and improve decision-making. Our main goal is to achieve accurate predictions of a customer’s potential default, while enhancing our knowledge of how savings and income dependency contribute to credit risk.

**About the 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. Key variables include income, savings, the history of the customer’s borrowing and the amount of the loan requested. For the purpose of this project, the focus will be primarily on income and savings as the main independent variables. The dataset utilized for this project is a CSV file containing various attributes related to customer financial profiles. Key features include:

* **Income:** The annual income of the customer
* **Savings:** The amount saved by the customer
* **Checking:** The amount on customers’ checking account balance
* **Age:** The age of the customer
* **Credit History:** A binary indicator of past credit behavior i.e.; good or bad
* **Loan Amount:** This is the level of loan that a customer is seeking.
* **Risk Label:** The target variable indicating whether the customer is low-risk or high-risk

The dataset consists of roughly 20,000 entries which are quite representative of the diversity existent within the customer profile. As initial observations however showed, the dataset does contain some discrepancies and also categorical variables that need some preprocessing.

**Project Plan**

The project will be implemented using Python as the programming language. We will utilize scikit-learn and pandas for model design and data processing, while matplotlib will be employed for data visualization. Furthermore, metrics and confusion grids, including precision-recall and F1 score, will be used to evaluate the performance of our model. In order to ensure the scalability of the project, the system implementation base will either be a private computer or a Google Colab environment.

The data preprocessing step [(Hand & Henley, 2007)](https://paperpile.com/c/jYzUPZ/sA7Y) consists of removing noise from the dataset by filling in missing values, eliminating inconsistencies, and adjusting numerical features (income and savings) [(Han et al., 2011)](https://paperpile.com/c/jYzUPZ/0MHg). Next, for feature selection, the two key features that we will choose for our model are income and savings [(“Recent Developments in Consumer Credit Risk Assessment,” 2007)](https://paperpile.com/c/jYzUPZ/Ocoe). For model building, we will build a predictive model by employing a classifier, specifically a logistic regression model or decision tree [(Bensic et al., 2005)](https://paperpile.com/c/jYzUPZ/dCD7). Furthermore, the following metrics will be used to investigate the performance of the model: accuracy, precision, recall, and F1 score. Additionally, we will construct a confusion matrix to analyze the false positives and false negatives produced by the model. To improve generalization and avoid overfitting, the model will undergo internal validation through cross validation.

**Methodology**

The project follows the following key methodological steps:

* **Data preprocessing:**
  + **Cleaning:** Addressing missing values by imputing them using the mean for numerical variables and mode for categorical variables.
  + **Encoding:** Converting categorical variables into numerical format using one-hot encoding.
  + **Normalization:** Scaling the income and savings features to ensure that they contribute equally to model performance
* **Exploratory Data Analysis (EDA):**
  + Initial visualizations revealed trends in income and savings across different risk categories. Correlation analysis indicated a positive relationship between savings and low-risk classification.
* **Model Development**
  + We plan to experiment with several classification algorithms, including Logistic Regression, Decision Trees, and Random Forests. Each model's performance will be evaluated based on metrics such as accuracy, precision, recall, and F1 score.

**Objectives**

The primary objective of this project is to accurately classify customers as low-risk or high-risk based on their financial attributes. Additionally, we aim to identify the most significant factors contributing to credit risk classification, providing insight to the key factors behind credit scores. Our goal is to develop a robust model that can be effectively deployed in real-world applications for credit scoring and to enhance decision making processes in financial institutions.

**Scope**

The scope of this project includes several essential tasks. First, we will develop a classification model using the dataset to distinguish between low risk and high risk customers. After creating the model, we will conduct thorough testing to evaluate the model’s performance. Additionally, we will provide insights and visualizations that help explain the model’s predictions.

**Updates and Modifications**

Based on our progress and initial assessments, we decided to refine our feature selection by exploring additional features, including employment status and debt-to-income ratio. Furthermore, we adjusted our model to prioritize ensemble methods, such as Random Forests, due to their effectiveness in handling imbalanced datasets to improve accuracy.

**Progress**

As of now, the initial data cleaning and preprocessing steps have been completed. Furthermore, the EDA step is also complete. We documented key findings, highlighting trends and correlations within the data. For the model training portion, a Logistic Regression model has been developed, which achieved an accuracy of 48%. We are currently working on tuning hyperparameters and evaluating additional models to improve performance. Our preliminary results indicate that income and savings are the most significant features for classifying low-risk and high-risk customers.

**Next Steps**

Moving forward, our priorities will include conducting further model training with additional algorithms and hyper parameter tuning to refine performances. We will also perform validation using a separate test set to ensure the model’s robustness and reliability. We will then begin drafting the final report, incorporating insights and recommendations based on our findings.

**References**

[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.](http://paperpile.com/b/jYzUPZ/dCD7)

[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.](http://paperpile.com/b/jYzUPZ/sA7Y)

[Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Elsevier.](http://paperpile.com/b/jYzUPZ/0MHg)

[Recent developments in consumer credit risk assessment. (2007). *European Journal of Operational Research*, *183*(3), 1447–1465.](http://paperpile.com/b/jYzUPZ/Ocoe)

[Thomas, L. C. (2009). *Consumer Credit Models: Pricing, Profit and Portfolios*. OUP Oxford.](http://paperpile.com/b/jYzUPZ/M5Xq)