Code Rush

PREPARED BY:

**Pralad Kadel**
Data Analyst Apprentice

Code Rush
Chakupat-10, Lalitpur
Nepal
coderush.com.np

# Default Prediction Model

# PROJECT PROPOSAL

PREPARED FOR:

**Shristi Shrestha**
Associate Project Manager

Code Rush
Chakupat-10, Lalitpur
Nepal
coderush.com.np

## Introduction

To every bank, it is very important that the person taking loans from them pay back on time. Thus, knowing which clients are likely to default in advance can be very beneficial to them. So, this project is an attempt to create a machine learning model that can predict beforehand if there's any chance of account default of a customer based on their income, loan amount, provided personal information and other factors. The model is intended to be used as a reference tool for the client and his financial institution to help make decisions on issuing loans, so that the risk can be lowered, and the profit can be maximized.

## Dataset

The data set that is being used for this project is a credit risk data set from the site Kaggle, with 12 features and around 32k records. Data includes loan details, e.g., amount, purpose, status, interest rate and loan grade and Customer details, e.g., age, income, employment, loan percent income, default history etc. Table below represents the data dictionary of final attributes that will be utilized in our model.

| **Feature Name** | **Description** |
|---|---|
| person_age | Age of customer |
| person_income | Annual income |
| person_home_ownership | Home ownership status |
| person_emp_length | Employment length (in years) |
| loan_intent | Intent of the loan |
| loan_grade | Grade of the loan |
| loan_amnt | Total loan amount |
| loan_int_rate | Interest rate |
| loan_status | Status of the loan (default or not) |
| loan_percent_income | Percent income |
| cb_person_default_on_file | Historical default |
| cb_person_cred_hist_length | Credit history length |

Dataset link: https://www.kaggle.com/datasets/laotse/credit-risk-dataset

## Related Work

Reddy and Kavitha [1] showed that using Neural Networks through attribute relevance analysis to build a prediction model increases the speed of Neural Network and feasible accuracy. A simple Neural Network model was used, and Info Gain algorithm was applied on attributes to eliminate less informative variables.

Shoumo et al. [2] focused on applying appropriate dimensionally reduction approach using Recursive Feature Elimination with Cross-Validation (RFECV) and Principal Component Analysis (PCA), noise handling, parameters tuning, using a grid search with cross-validation and on handling the imbalanced data problem. SVM and RFECV based models showed the ability to outperform other regression and tree-based models when prediction credit risk, where Support Vector Machine, Random Forest, Logistic Regression and Gradient Boosting algorithms were used in this paper.

The above research focused on using different classification models and a comparison between them, or added features selection technique to improve prediction, while in this project firstly we check if traditionally used 28/36 rule is sufficient enough to lend loans or prediction model are better than that. Then we will use logistic regression and gradient boost classification model incorporating multiple features from the datasets to cover more than one aspect that might impact prediction accuracy and performance to draw our inference.

## Methodology

Firstly, we will explore the data and then drop or replace the missing values in the datasets by their respective mean, median etc. Then we will analyze the correlation of different variables and visualize our prepared dataset using histogram, scatter plots, catplot etc. Then we will evaluate the accuracy and effectiveness of using the standard 28% mortgage rule and 36% debt to income ratio that is traditionally used by the lenders. After that, we will encode all the non-numeric data in the set to dummy variables and use logistic regression model and gradient boost classifier. Then we will compute and compare the accuracy scores, F1 scores etc. among them and see the effectiveness of modern machine learning algorithm over traditional manual methods.

## References


1. Reddy M. and Kavitha B., “Neural Networks for Prediction of Loan Default Using Attribute Relevance Analysis,” in Proceedings of International Conference on Signal Acquisition and Processing, Bangalore, pp. 274-277, 2010.

2. Shoumo S., Dhruba M., Hossain S., Ghani N., Arif H., and Islam H., “Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking,” in Proceedings of TENCON IEEE Region 10 Conference, Kochi, pp. 2023-2028, 2019.

3. A. Jeremy Mahoney (2020, Sep 9), Credit risk modeling with machine learning. *towardsdatascience* site, Last accessed 3rd Sep, 2022: *https://towardsdatascience.com/credit-risk-modeling-with-machine-learning-8c8a2657b4c4*