Richard Lowther and Gabriel Gonzales

# Project Plan

## 1. Introduction to the problem

Loan companies use different attributes of a customer to determine whether or not they are approved to receive a loan. These companies use industry-standard variables, including the customer's debt-to-income ratio (DTI ratio), income, credit score, employment history, etc. These requests are then either approved or denied, which thus necessitates a binary classification method. If a loan company were to provide too many "bad" loans quickly, it could jeopardize the company and possibly cause bankruptcy.

In the case of banks, having more debts than assets due to poorly placed loans renders the bank insolvent and incapable of repaying its depositors and tarnishing its reputation. Additionally, banks providing bad loans in excess strongly impact the economy and are a proven marker to indicate when economic crises occur. According to Fredriksson and Frykström[1], "When there are large volumes of bad loans, they have normally been preceded by sharp credit growth, resulting in higher loan-to-value ratios among corporations and households."

Because of the importance of providing proper loans, mathematical models have been developed to more closely predict whether or not the customer meets the adequate requirements to receive a loan. Logistic regression is a machine learning algorithm that, in this scenario, can be given the customer's credentials and will provide a probability of the loan being approved or rejected.

## 2. Related work

Multiple machine learning models can provide similar answers to the given data. Logistic regression is seen as a simpler machine learning model compared to what Lin[2] used in their research, such as XGBoost, an enhanced version of a Gradient Boosting Decision Tree, Random Forest, and AdaBoost. These models are more advanced than logistic regression due to their implementation of decision trees to more accurately provide the probability of loan approval. However, these models require more time and development due to their complex nature.

Hai and Ngoc[3] used logistic regression in their study of 1,000 loan applicants and achieved an accuracy of 76.80%. Similarly, Lin found that standard logistic regression had an accuracy of 81.06% as compared to the previously mentioned methods that reached accuracies in the lower 90% range. These accuracies were based on a database that had 32,581 entries. The data used in this project will contain 252,000 entries, which should produce higher accuracies than what the previous authors had achieved.

## 3. Proposed methodology/models

This project will focus only on logistic regression as its machine learning model and will develop the weight values of each variable provided in the data set. The specific attributes per individual in this data set are their income, age, years of professional experience, marital status, house and car ownership, profession, city and state, duration of employment in their current job, duration of residence in their current home, and whether or not they were flagged as a risky

applicant. Using some or all of these variables, the model will be able to predict the probability that the applicant received their loan accurately.

## 4. Experiment setups

The data set that will be utilized to perform this logistic regression model experiment will be taken from the "Loan Approval Dataset"[4]. The data will be cleaned, and the categorical variables, such as house and car ownership and years of professional experience, will then be converted into numerical representations through label encoding. Statistical analysis will be conducted using chi-squared tests and mosaic plots to determine the most significant predictors from the data set, and drop the least significant predictors from the model entirely.

The data will need to be validated, and the group will perform this through accuracy, precision, recall, Area Under the Receiver Operating Characteristic Curve (AUC-ROC), and F1-Score. For reference, the AUC-ROC is on a scale of 0 to 1.0, where 0.5 indicates random guessing, and 1 indicates perfect performance for distinguishing positive and negative instances. The team will also ensure that the logistics regression model will be generalized to fit new data using the K-fold cross-validation method. Upon validation, the model will need to be optimized, and that will be performed by adjusting the regularization parameters and addressing any dataset imbalances that may be present.

## 5. Expected results

The logistic regression model is expected to effectively classify loan applicants as either good or bad credit risks based on selected financial and demographic variables while achieving an accuracy between 70% and 85%. It is anticipated that not all variables within the dataset will be significant predictors of credit risk, and some will be able to be omitted without affecting the model's performance. The team expects that the most meaningful predictors for getting approved for a loan will be years of professional experience, house and car ownership, profession, and duration of employment in their current job. Some of the less meaningful predictors are hypothesized to be city, state, and age. The model is expected to yield high precision and recall for approved applicants in order to minimize false results, such as declining a loan for someone who should be approved. It is also expected that the AUC-ROC will be above 0.7, indicating that the model is at an acceptable level and is able to distinguish between high-risk and low-risk applicants accurately.

## 6. Team roles and responsibilities

As this is a team of two, the workload will be split down the middle as evenly as possible. Both team members will contribute to writing the report equally and plan to work on writing the code together to ensure that one team member is not left behind when doing this project. Setting up an online call and sharing each other's screens will reduce any holes in one's knowledge and reduce the downtime required to bring the other team member up to speed.

References

1) Fredriksson, O., & Frykström, N. (2019, November 2). *How bad loans affect banks and financial stability*. Sveriges Riksbank. https://www.riksbank.se/en-gb/press-and-published/notices-and-press-releases/notices/2019/how-bad-loans-affect-banks-and-financial-stability/

2) Lin, J. (2024). Research on loan default prediction based on logistic regression, randomforest, xgboost and adaboost. *SHS Web of Conferences*, *181*, 02008. https://doi.org/10.1051/shsconf/202418102008

3) Hai, H. T., & Ngoc, D. T. H. (2020, May). *Application Of Logistic Regression Model In Consumer Loans Credit Scoring*. International Journal of Advanced Research and Publications.

4) Sharma, R. (2024, April 20). Loan approval dataset. Kaggle. https://www.kaggle.com/datasets/rohit265/loan-approval-dataset?resource=download