|  |  |  |  |
| --- | --- | --- | --- |
|  | Ben Calderaio, Conrad Urffer, Clara Bucar, Tammy Lacher, Jeff Pinegar | |  |
| Assignment: | Project 4: Proposal |  |
| Due Date: | March 23, 2023 |  |

**Credit Default Risk Prediction**

# Construction of a model for predicting the risk of Credit Default?

## Scope

Base on a dataset of loan applications together with a classification of which of those loans eventually experienced at least one missed payment, a model will be constructed that can ideally classify future applications and risk or safe. Bankers we tend to be risk adverse so the emphasis will be on identifying those application that are risky (positive for risk). There will be less concern about rejecting application that are likely to be safe (false positives for risk). Still there must be balance we can not simple reject everyone and assume no risk of default.

## Questions

* Can we build a model, with accuracy greater than 75%, to predict which potential creditors are likely to pose a future risk of default?
* To achieve 75% accuracy (true positives for risk), does the model have an acceptable low level of false positives.

## Potential Models

The data set includes a target (label) value so this will be Supervised learning for classification with categorical and continuous independent variables. The potential options include:

* Logistic Regression
* K Nearest Neighbor Clustering
* ROC AUC (Receiver Operating Characteristic - Area under the Curve)
* SVM (Support Vector Machine)

Because of the large number of features the following methods and models may be requires.

* Decision Tree (Random Forest)
* Lasso Feature Selection
* Principal component Analysis (PCA)

## As time permits

The features set include some objectively financial variables (loan amount, real estate ownership, income, etc.) as well as some “social variables” (gender, number of children, education, etc.). If time permits, it may be interesting and possible to assess the relative weights of these features.

# Data Overview *(potential visuals and interactions)*

## Characteristics of the data set.

* 307,511 records
* 1 target (label)
* 120 potential features
  + 63 Categorical features
  + 57 Continuous Variables (floats or integers)

## Probable Required Data Transformation

* If all the features are retained, 12 will likely require binning.
* 116 features contain blanks to manage
* 10 continuous features will clearly need to be scaled because of the range.

Logo Source: https://www.affinityonefcu.org/primary-savings/