Project Overview

Predicting Loan Default

In this project, your ultimate goal is to predict whether a loan applicant will default on their loan based on historical loan and applicant data. Loan default is a significant risk for financial institutions, and accurately predicting the likelihood of default can help lenders make more informed decisions about granting loans. By applying data mining techniques, you will develop predictive models that help to assess an applicant’s risk of default based on various factors.

You will complete the project over the semester in FIVE parts as follows:

- Part 1: Data Cleaning and Preparation for Modeling
- Part 2: Decision Tree Model
- Part 3: Logistic Regression Model
- Part 4: Neural Network Model
- Part 5: Champion Model Competition

Variables

The response variable is `loan_default`. You have the variables below to create the best possible model:

Variable	Description
loan_default	Whether the loan has defaulted. Values: "Yes", "No".
loan_amnt	The listed amount of the loan applied for by the borrower.
term	The number of payments on the loan (36 or 60 months).
int_rate	Interest rate on the loan.
installment	The monthly payment owed by the borrower if the loan originates.
grade	Loan grade assigned by Lending Club.
sub_grade	Loan subgrade assigned by Lending Club.

Variable	Description
emp_title	Job title supplied by the borrower.
emp_length	Employment length in years ("< 1 year" to "10+ years").
home_ownership	Home ownership status: RENT, OWN, MORTGAGE, OTHER.
annual_inc	Self-reported annual income of the borrower.
verification_status	Whether income was verified: Verified, Not Verified, etc.
issue_d	The month when the loan was funded.
purpose	Category provided by the borrower for the loan request.
title	Loan title provided by the borrower.
dti	Debt-to-income ratio.
earliest_cr_line	Month of the borrower’s earliest reported credit line.
open_acc	Number of open credit lines.
pub_rec	Number of derogatory public records.
revol_bal	Total revolving balance.
revol_util	Revolving line utilization rate (credit used / total credit).
total_acc	Total number of credit lines in the borrower’s credit file.
initial_list_status	Initial listing status of the loan: w (listed as a whole loan) or f (listed as a fractional loan).
application_type	Type of application: individual, joint, direct pay.
mort_acc	Number of mortgage accounts.
pub_rec_bankruptcies	Number of public record bankruptcies.
address	Borrower’s address.
delinq_2yrs	Number of 30+ days delinquency incidents in past 2 years.
fico_range_low	Lower end of FICO score range.
fico_range_high	Upper end of FICO score range.
inq_last_6mths	Number of credit inquiries in the past 6 months.

Variable	Description
mths_since_last_delinq	Months since last delinquency (NA if never).
last_credit_pull_d	Most recent date credit was pulled.
acc_now_delinq	Number of accounts currently delinquent.
hardship_flag	Whether borrower is under hardship plan (Y/N).
debt_settlement_flag	Whether borrower is in a debt settlement program (Y/N).

Dataset

You will be given THREE datasets to complete the project.

Part 1: The first dataset is given below and will be used to complete Part 1 of the project. Once you have preprocessed this dataset, you will use the preprocessed data to train your predictive models in Parts 2-4.

- **train.csv (https://miamioh.instructure.com/courses/239903/files/36479087?wrap=1) (https://miamioh.instructure.com/courses/239903/files/36479087/download?download_frd=1)**

Part 2-4: You will be given a holdout sample when you begin Part 2 of the project. You will repeat the preprocessing steps you completed in Part 1 on the holdout sample. In each assignment for Parts 2-4, you will train a different predictive model using the preprocessed training sample from Part 1. You will evaluate the model performance on the preprocessed holdout sample from Part 2-4.

- holdout data to be posted later

Part 5: You will be given the final dataset to score your project once you have completed the first four parts of the project. The final scoring dataset will be used for the competition.

- scoring data to be posted later

Important Information

- **You will be assigned a partner, and all work will be completed and turned in with your partner.**
- **It is very important to take each part of the assignment seriously, especially Part 1.**

- **The real goal of this assignment is for you to learn how to prepare and model data. If you choose to use an AI tool as an assistant, your guiding principle should be that you are a trained professional who will vet the code, verify the results, and guide the analysis. Ultimately, you are responsible for the integrity of your work.**