Mortgage loan default prediction

Problem statement/Context
Predict the default risk of current loans based on historical mortgage loan data.

Criteria for success
TBD (Model performance metrics - ROC, Confusion matrix, F2-score, Recall?)

Scope of Solution space
We are using U.S. mortgages lent after 1999.

Constraints within solution space
Not considering all of the assets of the borrower.

Stakeholders to provide key insight
Banks and lending companies that provide mortgages to borrowers and homebuyers

Data sources
https://freddiemac.embs.com/FLoan/Data/downloadQ.php

The source data is a single family loan level dataset from Freddie Mac and has credit performance data on all mortgages that the company purchased or guaranteed from 1999 to 2023.

The dataset covers approximately 52.6 million mortgages originated between January 1, 1999 and March 31, 2023.

The source data will have the loans provided, the attributes of these loans and the delinquency status for these loans. The loan and borrower attributes would include information such as credit score, first payment date, first time homebuyer flag, property state and zipcode, property type, loan to value, Debt to income ratio, loan term, interest rate, etc. The input features to train the model will be the loan and borrower attributes. The target will be the delinquency status of the loan.

Approach to solving problem
Perform EDA and feature engineering on the loan default dataset. Setup training dataset with the relevant loan, borrower attributes and the default status as the target. Evaluate different classification models on the training dataset and identify the best model to use for predicting the default status.
Since the loan dataset is huge (52.6 million), initial data exploration, EDA, feature engineering, model evaluation will be performed on a smaller dataset (2000 Q1). The 2000 Q1 dataset has 117k loans. After the initial exploration of 2000 Q1 data, additional loan data for other years will be included in the training dataset if needed. Additionally, loan dataset for some years will be set aside to be used as a test dataset.

Deliverables

Github repository containing the code. Project report and slide deck will also be provided.