554.488/688 Computing for Applied Mathematics
Fall 2022 - Final Project Assignment

Fannie Mae Loan Performance Prediction Project

Project Aim

The aim of this project is use data collected on a large number of loans in Oct of 2021, to develop prediction models for the number of months payments are made on mortgage loans and for predicting foreclosure of a loan based on information available to FNMA at the time the loan is put on their books.

Some background

FNMA aka Fannie Mae (look it up in Wikipedia) was put in place in order to ensure liquidity in the US mortgage loan markets. When a mortgage holder secures a mortgage from a bank, the bank will sell that mortgage to FNMA giving them the capital to enable them to make additional future loans. The FNMA bundles the mortgages they acqure into what are called mortgage-backed securities (MBS's) and sells them to investors while insuring the underlying mortgages against losses of principal. The investors recieve the bundle of monthy payments associated with the underlying mortgages. When a holder of a mortgages forcloseses/defaults, or sells their house, or refinances their mortgage while most of the principle is transferred to the holder of the MBS, but this means that their future cash flow might not be as expected (interest rates may have gone down so any future bond investments bring lower returns). So the investor, in pricing the value of their asset, would like to be able to predict outcomes such as foreclosure or when the loan will be settled.

Hopefully this brief description gives you an understanding as to why one woud be interested in, for a given loan, being able to determine how likely it is to foreclose, or determine its duration i.e. how many months of payments can be expected before monthly payments cease to occur.

Your personal dataset

- An email will be sent to every student in the class with urls for two comma delimited files: a training dataset and a test dataset.
- Each student will have their own unique set of data. Each data have been drawn from different populations - using results for someone elses dataset will likely lead to poor performance.
- You should not share these datasets with any other students in the class.
- You should not collaborate with other students in the class
- Any evidence of data sharing or collaboration will be viewed as an ethics violation and subject to the rules and regulations of the university.

Training set

The training set is a comma delimited file consisting of information for exactly 250,000 mortgage loans with 30 variables (LOAN_ID, 27 predictor variables, and 2 response variables):

- The LOAN_ID variable is a unique 12 character identifier for a mortgage loan.
- Predidctor variables (as well as the others) are described in the Appendix. These variables provide information about the mortgage known to Fannie Mae when the mortgage was acquired by them.
- The response variables are variables that ultimately become known by the time the data on loan performance was collected in Oct 2021.
  - NMONTHS variable is the number of months of mortgage payments made on the loan up until the date when data was collected.
  - FORECLOSURE variable is 1 if the loan foreclosed, and 0 otherwise as of the date when the data was collected.

Test set

The test set is also a comma delimited file consisting of information for 100,000 mortgage loans (drawn at random from the same loan population as your training set) with only the LOAN_ID and the 27 predictor variables. I have the ground truth, i.e. the NMONTHS and FORECLOSURE variables for the loans in your test set. Once I have your predictions I will be able to determine the quality of those predictions.

Your task

Your task is to use the training data to build a predictors of each of the two response variables NMONTHS, FORECLOSURE.

For FORECLOSURE, I am asking you to **pick 1,000 loans you think are most likely to foreclose.**

Reading the data to create a data frame

The two files you are provided with are comma delimited with all data represented as a string (each column/field of fixed size). To ease the process of reading these files to produce data frames, a jupyter notebook called "FunctionToReadData" has been provided. This function does the conversions of the fields for you so to create the data frames (either training or testing) you simpy give a command like:

```
df=read_data(fileid)
```

where fileid is the identifier of the downloaded file.

Some recommendations

- You can use any method you wish to build your prediction model, but I recommend that you
  - use regression for NMONTHS
  - logistic regression for FORECLOSURE
- get started early!!! don't put this off!!!
- don't assume prediction rates will be low - do the best you can
- it is not just important to get good predictions - it is also important to be able to quantify how well your predictions are likely to perform i.e. do a good job in estimating your error rates
- since you only have ground truth in the training set, it is recommended that you separate that dataset into a training set and a test set so that your error estimates are not underestimated due to over-fitting.
- try various choices of sets of variables to use as predictors and compare performance on test data

I will ask you to provide a couple of summary bits of information about the variables in your training set.

Deliverables

The submissions for this assignment should be done in Canvas in 3 separate parts:

(1) A question about your personal filecode (it will be in an email) Some summary statistics for your training dataset and predictions as to your performance in the predictions you make.

(2) All code you used to do your work in a single jupyter notebook.

(3) Your predictions in a .csv file.

Part 1 Questions

Here are the questions in Part 1 of the assignment:

- What is your personal 4 character file code?

- How many loans in your training set are foreclosures? (Answer should be an integer)
- What is the mean value of NMONTHS in your training set? (Answer should be a float rounded to 3 decimal places e.g. 17.524
- When your foreclosure predictions are compared to the true values for all 100,000 loans in the test set, what do you estimate to be the number of those 1,000 predicted foreclosures will actually be foreclosed in the test set?
- When your predictions of NMONTHS are compared to the true values for each of the 100,000 loans in the test set, what do you estimate to be the mean absolute error i.e.

$$\frac{\sum_{i=0}^{99,999} |NMONTHS_i - \hat{NMONTHS_i}|}{100,000}$$

where $\hat{NMONTHS_i}$ denotes your prediction for $NMONTHS_i$, the value of $NMONTHS$ for the i-th loan?

Part 3 Specifications

Once you have used your training set to come up with what you consider to be your best prediction model of NMONTHS, and your prediction model for FORECLOSURE, you are required to submit in Canvas single comma delimited file with the following **exact** specifications:

- The file should have a header with column names LOAN_ID, NMONTHS, FORECLOSURE.
- There should be exactly 100,000 lines after the header with the LOAN_ID, and your predicted values of NMONTHS and FORECLOSURE for every one of the loans in your test file
- The NMONTHS column should consist of a number (int or floating point) - your predictions can have fractional parts e.g. 15.542 is an allowable prediction for NMONTHS for a loan
- The FORECLOSURE column should consist of only 0's and 1's where 0 means you predict non-foreclosure and 1 means you predict foreclosure.
- Since you are being asked to find the 1,000 loans most likely to foreclose, your FORECLOSURE column should contain exactly 1,000 1's and exactly 99,000 0's.

It is important that your comma follows the specifications exactly. In order to ensure that you do this properly you will be provided in Canvas with a jupyter notebook that you can test your submission on for errors.

Evaluation

Your grade will be based on

- Quality of your NMONTH predictions as measured by the mean absolute deviation between true and predicted values in the test set.
- Quality of your FORECLOSURE predictions as measured by how many correct FORECLOSURES did you identify?
- How close was your estimated mean absolute deviation for your NMONTHS predictions to the mean absolute actually obtained?
- For FORECLOSURE, how close is your estimate of the number of FORECLOSURES identified to the actual number of FORECLOSURES correctly identified.

Appendix: Description of variables

For each variable, the number of characters used in the field for that variables is indicated in parentheses.

LOAN_ID (12): A unique identifier for the mortgage.

MONTHLY_REPORTING_PERIOD (6): The month and year that pertains to the servicer's cut-off period for mortgage loan information.

CHANNEL (1): The origination channel used by the party that delivered the loan to the issuer.

ORIGINAL_INTEREST_RATE (6): The original interest rate on a mortgage loan as identified in the original mortgage note.

CURRENT_INTEREST_RATE (6): The rate of interest in effect for the periodic installment due

ORIGINAL_UPB (10): The dollar amount of the loan as stated on the note at the time the loan was originated.

ORIGINAL_LOAN_TERM (3): Original Loan Term (Months)

ORIGINATION_DATE (6): For fixed-rate, adjustable-rate and Interest-only mortgages, the number of months in which regularly scheduled borrower payments are due at the time the loan was originated.

FIRST_PAYMENT_DATE (6): The date of the first scheduled mortgage loan payment to be made by the borrower under the terms of the mortgage loan documents.

LOAN_AGE (3): The number of calendar months since the mortgage loan's origination date. For purposes of calculating this data element, origination means the date on which the first full month of interest begins to accrue.

REM_MONTHS_LEGAL_MATURITY (3): The number of calendar months remaining until the

mortgage loan is due to be paid in full based on the maturity date as defined in the mortgage documents.

REM_MONTHS_MATURITY (3): The number of calendar months remaining until the outstanding unpaid principal balance of the mortgage loan amortizes to a zero balance, taking into account any additional prepayments, which could lead to the loan paying off earlier than its maturity date.

MATURITY_DATE (6): The month and year in which a mortgage loan is scheduled to be paid in full as defined in the mortgage loan documents.

LTV (2): The ratio, expressed as a percentage, obtained by dividing the amount of the loan at origination by the value of the property.

CLTV (3): The ratio, expressed as a percentage, obtained by dividing the amount of all known outstanding loans at origination by the value of the property.

NUMBER_OF_BORROWERS (2): The number of individuals obligated to repay the mortgage loan.

DTI (2): The ratio obtained by dividing the total monthly debt expense by the total monthly income of the borrower at the time the loan was originated.

B_CREDIT_SCORE_O (3): A numerical value used by the financial services industry to evaluate the quality of borrower's credit.

CB_CREDIT_SCORE_O (3): A numerical value used by the financial services industry to evaluate the quality of co-borrower's credit.

FIRST_TIME_HOME_BUYER_IND (1): An indicator that denotes if the borrower or co-borrower qualifies as a first-time homebuyer.

LOAN_PURPOSE (1): A character that denotes whether the mortgage loan is either a refinance mortgage or a purchase money mortgage. Purpose may be the purchase of a new property or refinance of an existing lien (with cash out or with no cash out).

Cash-Out Refinance = C
Refinance = R
Purchase = P
Refinance-Not Specified = U

PROPERTY_TYPE (2): Code for type of property
CO = condominium
CP = co-operative
PU = Planned Urban Development

MH = manufactured home
SF = single-family home

NUMBER_OF_UNITS (1): The number of units comprising the related mortgaged property (one, two, three, or four).

OCCUPANCY_STATUS (1): The classification describing the property occupancy status at the time the loan was originated.

PROPERTY_STATE (2): A two-letter abbreviation indicating the state or territory within which the property securing the mortgage loan is located.

MSA (5): The numeric Metropolitan Statistical Area Code for the property securing the mortgage loan. MSAs are established by the US Office of Management and Budget. An area usually qualifies as an MSA if it is defined by the Bureau of the Census as an urbanized area and has a population of 50,000 or more in a total metropolitan area of at least 100,000. An MSA may consist of one or more counties.

ZIP_CODE_SHORT (3): Limited to the first three digits of the code designated by the U.S. Postal Service where the subject property is located.

MORTGAGE_INSURANCE_PERCENTAGE (5): The original percentage of mortgage insurance coverage obtained for an insured conventional mortgage loan and used following the occurrence of an event of default to calculate the insurance benefit, as defined by the underlying master primary insurance policy.

FORECLOSURE (1): Indicator of foreclosure (0=non-foreclosure, 1=foreclosure)

NMONTHS (3): Number of months of mortgage payments made up until date the data was collected.