554.488/688 Computing for Applied Mathematics
Fall 2020 - Final Project Assignment

Loan Performance Prediction Exercise

The aim of this project is to develop prediction models for the length of time that FNMA holds a mortgage loan and for predicting foreclosure of a loan, based on information available to FNMA at the time the loan is put on their books. The data used is a portion of the single family loan portfolio for FNMA originating in the second quarter of 2000.

Data for this project are available in this folder: Data
at this url: `http://jesse.ams.jhu.edu/~dan/ComputingForAppliedMathematics`

You will find five files:

- TRAIN.csv a full data set with information for 200,000 loans (rows) for training with response variable values (NMONTHS, FORCLOSED) available
- TESTPARTIAL.csv a partial data with information for 71,197 loans (rows) with response variable values (NMONTHS, FORCLOSED) missing
- a glossary file from FNMA describing the variables
- a text file with 108 column names
- a text file with abbreviated 108 column names used in the data sets above (note: only 38 columns are actually used and provided)

The TRAINING and TESTPARTIAL data sets contain informatio for disjoint sets of loans.

In the training data, the response variables we wish to predict are:

- NMONTHS, the number of months until the mortgage is taken off the books due to foreclosure, prepayment, etc..
- FORCLOSED is a boolean variable that indicates whether the mortgage foreclosed (True) or not (False)

There are several variables available in the training data that could be used to predict the response variables.

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

Once you have arrived at what you consider to be your best predictor of NMONTHS, you should

- use your predictor to predict NMONTHS for the loans in the TESTPARTIAL data set where you are not given the luxury of ground truth
- give an estimate of the mean absolute difference between predicted value and true value for NMONTHS when your predictor is used

For the second item, it would be wise to set aside some data for estimating error.

Once you have arrived at what you consider to be your best predictor of FORCLOSURE, you should

- use your predictor to predict FORCLOSURE for the loans TESTPARTIAL data set where you are not given the luxury of ground truth
- try to make the false positive rate for your predictor as close as you can to 50% (here ”positive” means FORCLOSURE=True) and make an effort to get the true positive rate to be as high as possible
- give an estimate of the false and true positive rates when your predictor is used to predict foreclosure for the loans in the TESTPARTIAL dataset

Once you have created prediction models, you should use them on the TESTPARTIAL data and submit your predictions using the format procided below.

Please submit

- narrative (see below) - you can submit a word or pdf file separately or incorporate your narrative into the jupyter notebook
- the jupyter notebook you used to produce the models, and apply them to the date.
- a .csv file called `predictions.csv` with one row for each of the 71,197 loans and the following 3 columns (and only these columns please!!!)
  - LID
  - NMONTHS
  - FORCLOSURE

Your grade will be based on the following considerations:

- **Quality of the narrative.** Your notebook or accompanying word/pdf document should provide a clear explanation of things you tried and how you arrived at your final prediction approach. Make sure to explain all the methods you tried and how you picked variables to include in your model.

- **Effort.** How much effort went into your work? Did you stop after trying one approach or did you try several?

- **Creativity.** Did you do something novel?

- **Performance in prediction.** How well did your predictors perform?

- **Performance in error estimation.** Did you provide estimates of error rates (mean absolute error for NMONTHS, false and true positive rates for FORCLOSURE)? How well did your predictors perform?

- **Meeting the requirements.** Did you follow the instructions as stated, e.g. is the file name for your predictions correct? Is the file correctly formatted?