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 first quarter of 2000

Data for this project are available in the zip file in email

You will find four files:

- TRAIN.csv a full data set with information for 164,575 loans (rows) for training with response variable values (NMONTHS, FORECLOSED) available
- TEST.csv a partial data with information for 82,288 loans (rows) with response variable values (NMONTHS, FORCLOSED) missing
- glossary/pdf from FNMA describing 108 variables - we are only using a subset of 32 of these variables
- VariablesUsed.csv, which tells you which variables of the 108 in the glossary are actually available as predictors (see the Boolean column called "used") and also gives an abbreviated name for each variable in the glossary file. The abbreviated names are the ones that used in the datasets provided

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

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

- NMONTHS, the number of months until the mortgage is taken off the books due to foreclosure, prepayment, etc..
- FORECLOSED 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. **You should use regression for NMONTHS and logistic regression for FORECLOSURE.**

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 TEST.csv data set where you are not given the luxury of ground truth

Once you have arrived at what you consider to be your best predictor of FORECLOSURE, you should provide a list of the 1,000 loans in the test dataset that you consider most likely to foreclose.

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

Please submit the following items in Canvas in separate parts:

- part 1 - the jupyter notebook you used to do all of your NMONTHS predictions
- part 2 - a csv file called **PREDS.csv** with column headings: LID, NMONTHS and a row for each loan in the test dataset giving the loan id (**which is a string!!!**) to your file should have 82,289 rows (one for the column heading) and your prediction of NMONTHS for that loan (**this can be an int or a float**)
- part 3 - the jupyter notebook you used to get your FORECLOSURE predictions
- part 4 - a text file called **FORECLOSURES.txt** with 1,000 rows (no header) each row containing an LID (not quoted please).

Your grade will be in two parts.

- For NMONTH predictions, getting the mean absolute deviation between true NMONTHS (which I have access to) and your predicted NMONTHS as small as possible
- For FORECLOSURES, maximizing the number of correctly identified foreclosures out of the 1,000 you provide. (I have access to which loans actually foreclosed.)