554.488/688 Financial Computing I
Fall 2018 - Final Project Assignment due Dec 11th

Loan Performance Prediction

Data for this project are available in this folder: data

To summarize what we did in the lectures, we started with data for 471,277 loans. We found 1,585 loans to remove from the data set because we viewed them as problematic from the point of view of computing the cash flow. [1] For the 469,692 loans that remained, we compute the associated sequence of payments (incoming cash) for each loan age.

The 469,692 loans have been randomly broken up into 3 equal sized groups (each group has 156,564 loans). You have been provided with acquisition data for groups 0,1 and 2, and performance data (cash flow) for groups 0 and 1. You do not have access to cash flow data for group 2 and for this assignment you should not make use of any performance data related to group 2.

The data that you have access to has been stored in two different ways, either as .csv and as hdf5 files:

- The files A0.csv and A0.h5 gives you Acquisition data frame for group 0, and the CASH_FLOW0.csv and CASH_FLOW0.h5 give you the CASH_FLOW data frames for group 0.
- The files A1.csv and A1.h5 gives you Acquisition data frame for group 1, and the CASH_FLOW1.csv and CASH_FLOW1.h5 give you the CASH_FLOW data frames for group 1.
- The files A2.csv and A2.h5 gives you Acquisition data frame for group 2.

In the hdf5 files, the stored data frame objects are called "AcquisitionData" and "CASH_FLOW".

Your job is to use the data provided to you (groups 0 and 1) to create five prediction models related to loan duration and five related to loan revenues.

Specifically, for each $m = 12, 24, 36, 48, 60$ you should

- create a model that predicts the chance $p(m)$ that a loan is still active for at least m months (i.e. there is data for loan age equal to m.), using data at time of acquisition, and
- create a model that predicts the cumulative revenue $r(m)$ for a at loan age m months [2]

[1] To be clear, we included lots of other cases where there were problems, and we made rather gross assumptions about how to calculate cash flows for all the remaining loans.

[2] Cumulative revenue is the sum of payments up to and including the loan age

Once you have created your 10 prediction models, you should use them on the group 2 data to make predictions.

Please submit

- narrative (see below) - you can submit a word file separately or incorporate your narrative into the jupyter notebook
- the jupyter notebook you used to produce the models, and apply them to the group 2 data.
- a .csv file called `predictions.csv` with one row for each loan in group 2 and the following 11 columns (and only these columns please!!!)
  - LID
  - p(12), p(24), p(36), p(48), p(60)
  - r(12), r(24), r(36), r(48), r(60)

Since the data you have been given is broken up into two pieces, you can build a predictor using the group 0 data and when you think you have a really good idea for how to make predictions, you can test your predictions on group 1 data. (You can do the reverse as well.)

Importantly, it would be wise to do some splitting of the groups of data you have into even smaller groups for developing a predictor because each time you test on the same data, you run the risk of over-fitting.

Your grade will be based on the following considerations:

- **Quality of the narrative.** Your notebook or accompanying word document should provide a clear explanation of things you tried and how you arrived at your final prediction approach.
- **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.** How well did your predictors perform? Specifically, for the probability estimates $\hat{p}$, when activity of the loan is coded as Y=0(inactive) and Y=1(active) at each loan age, the performance will be calculated as the average absolute difference $|\hat{p} - Y|$. For revenue estimates, the average absolute difference between observed and predicted revenue will be used as the performance criterion.
- **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?