MS&E 246: Financial Risk Analytics

Course project: Design, implement and test a model of small-business loan charge-off ("default") based on a data set of roughly 150,000 equipment loans backed by the US Small Business Administration (SBA) between 1990 and 2014, which will be provided as an Excel spreadsheet. Variable descriptions are provided in a separate document.[1] Make sure you understand the nature of the loans (see the SBA website for more details on the loan program, which is called "504"), the structure of the data set and the variables provided. Detailed instructions follow.

- Clean the data set (eg remove loans labeled Exempt and Canceled, identify bad values). NB: loans which have a CHGOFF status but a gross charge-off amount of zero should not be dropped (treat these as defaults; the SBA probably made a full recovery due to selling collateral).
- Think about the treatment of missing data (for example, the InitialInterestRate is missing for many loans in the sample). One approach is to introduce dummy variables for missingness (as discussed in class).
- Think about the construction of training, validation and testing sets (eg. Random splitting vs. splitting by calendar date)
- Explore the loan data set to inform model building.
- Justify your model of default timing and the choice of predictor variables. Explore predictor variables beyond those provided in the SBA data set, in particular time-varying risk factors at the regional (i.e. zip-code) level. Also consider linear and nonlinear model alternatives.
- Select and implement an appropriate method for fitting the model parameters.
- Rigorously test predictive performance (in- and out-of-sample) of your model alternatives using a Receiver-Operating-Characteristic (ROC) curve or other appropriate metrics.
- Explain the fitting results and the fitted model. Which variables are important and why?
- Next develop, fit, and evaluate a model for the loss at default (constructed in addition to the model of default timing).
- Then, estimate the distribution of total loss on a portfolio of 1,000 loans randomly selected from the test set, over five and ten year periods.
    - Think of this as a hypothetical portfolio that is to be evaluated as of the current date. Use the time-independent loan/borrower attributes (eg. Loan term) from these loans along with current values of the macro-economic and any other time-varying variables (if you use any) when evaluating your fitted model.
- Measure the risk in terms of the VaR and the Average VaR (also known as expected shortfall) at the 95% and 99% levels. Include confidence bands for your estimates.
- Finally, estimate the distributions for the five and ten year losses of an investor who has bought a [5%,15%] tranche backed by the chosen portfolio. Also consider a [15%, 100%] senior tranche. Interpret and compare the distributions from a risk management perspective.

[1] It appears that the description in the variables PDF is wrong for ChargeOffDate. It contains the description for GrossChargeOffAmount, we believe.

- Owning a [x%, y%] tranche means that the first x% of losses due to default in the portfolio can occur before the investor in this tranche experiences losses, and any following losses affect this tranche entirely until losses in the portfolio reach y% (at which point the investor is "wiped out": the entire investment is written off). So e.g. a 10% realized loss rate on the portfolio is a 33% loss for the [5%, 15%] tranche, a 15% rate is 100% loss for the tranche, etc.

Write up a final report, detailing your models, statistical estimation approaches, tests, and results. We'd like to see sufficient details enabling us to replicate the results.

Please form a team of up to 4 students by Tuesday 1/16/24 10PM and send an email with the team composition (including names and email addresses) to the TA (one email per team).

The teams will **present their results in the last class on 3/15/24**. Mark your calendar!