AI For Finance – Project 3

# Predict bank failures

What to do ?  
In your project, you do one of the three options:

1. You find other AI algorithms that are better than what I have. Naturally, you have to replicate the results I have first.
2. Replicate the results for banks in any country, including Italy.
3. Find other variables that you think would make the current model/algorithms better.

Ok 😊

## Goal

Predict systematic bank failure and answer the following:

* Whar is the forecast horizion ? Given in next slides (?) = 4 ????
* How well can we predict bank failures ? -> Compute AUROC
* What variables have leading information about systematic bank failures ? -> Probably the ones in the given excel, maybe exists other variables (more information during lesson ??)

## Data

The data is inside an excel file called: ‘project3\_data.xlsx’

### Data preparation

Define systemtic failures and the corresponding frequency distribution of failed to active bank in %. There is a column in the excel source file called “Failed\_to\_Active\_banks”, calculate mean and standard deviation of that column. It’s a frequency distribution so use this 5 threshold:

1. < 0.0652%
2. 0.0652-0.1956%
3. 0.1957-0.3260%
4. 0.3260-0.4564%
5. >=0.4564%

For each threshold specify:

* midpoint (mean value of interval)
* Frequency: count of occurency of samples in that interval
* Relative Frequency
* Cumulative frequency (the sum of cum freq for each interval is equal to 100%)

### Predictor variables

These are the variables of the excel file that should be used as predictor variables.

* XMRET: Stock market excess returns
* EBP: Excess Bond Premium, corporate bond credit spreads
* TERMSPREAD: Treasury term spread, the difference between the yield on 3-month Treasury bills and 10-year Treasury bonds.
* DELTAFED: is the first difference of the Federal funds rate;
* ROE: is bank returns on equity;
* LLRR: bank loan loss reserve to total loans ratio;
* DELTANIM: is the first differences of bank net interest margin;
* PLL & ALL: are provision for loan losses, allowance for loan losses to total loans ratios, respectively;
* DELTAGDP: is the log difference of real GDP.

## Project

Now run Probit, logit, PNN, MLP, Random Forest, RNN, SVM, Tree Ensemble models. Our dependent variable is a binary one; it takes the value of “1” if the failure rate is higher than some threshold, “0” otherwise. Basically, 5 different thresholds will have 5 dependent variable.

As evaluation techniques we use AUROC. Consider, binary predictions are rarely “1” or “0”, that’s why use AUROC for classification. In addition to AUROC, find statistical difference between the AUROC values of each couple of two models. If we have two predictions using model 1 and 2, the accuracy of one model over the other is evaluated by computing a Wald type of test statistics:

W\_AUROC = (AUTOC1 – AUROC2) / SE((AUROC1-AUROC2))

Print all the results inside one or more dataframe.

### Caveats

1. Remember, you cannot have one lag of a predictor, and two-lags of another predictor in the same model JUST TO GET A BETTER FITNESS.
2. While normalizing the data is not required, you will most likely get a better predictor when the predictor data is not out of control.
3. After normalizing, you may want to report the correlation matrix. If there is multicollinearity (that is, two variables have same information), one can be dropped.

First, standardize data and then compute correlation matrix in order to understande which variables can be used together.

## Tests

1. Use 2/3 of the data for training.
2. Use the rest for validation.
3. For evaluation, compute AUROC values of each model at each forecast horizon.
4. Next, calculate the AUROC gains and their statistical significance relative to the highest AUROC at each forecast horizon.

Print Evaluation test results for one dataset at a time.

Then, Plot the ROC of each model at forecast horizon of 4 quarters (You can choose > 80 percentile dataset for this. The likely reason is that when the failures are very high, most likely economic contraction will happen) -> percentile mean threshold ? …

## Project report

1. Write the goals precisely.
2. Write about the data and preparing the data
3. For each algorithm, briefly mention the “hyperparameters”, and their sensitivity to prediction.
4. Conclude with the best prediction model at each forecast horizon.
5. In best models at each forecast horizon, what variable(s) has(have) the maximum contribution (as per AI algorithms or the logit model) to predictions?