AI for finance: Project 3

Giovanni Scialla

April 2023

Scenario

Predicting the probability of bank failures at a macro level

Systematic risk: Bank failure is the root cause of virtually all crises

Idiosyncratic risk: Bank failures occur during all types of economic environments

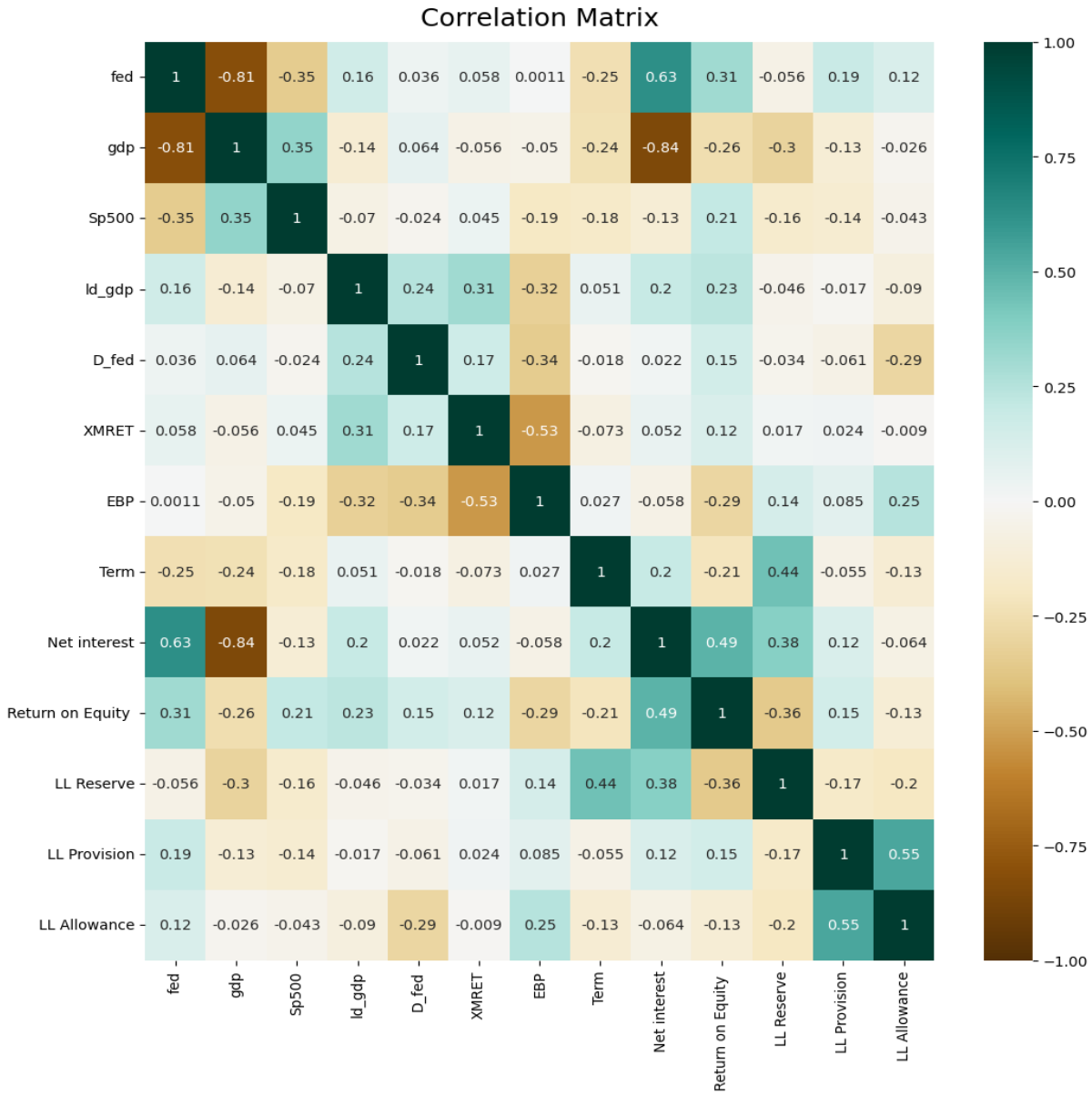
In our project we are looking for predicting the systematic risk that may trigger a contagion of bank failures.

Data source: Federal Reserve Bank website. (quarterly samples 1984-2020)

Both financial and bank balance sheet data are used in order to predict commercial bank failures.

Exploratory Data Analysis

By plotting the correlation matrix we can see how both FED and GDP variables show a multicollinearity behaviour with the Net Interest Margin. So we can safely drop them from our dataset



Predicting models over 6 forecast horizons

The bank failure prediction will be addressed ad a Multi-class classification problem with 5 classes: 'very low risk' - 'low risk' - 'medium risk' - 'high risk' - 'very high risk'

Models and hyperparameters used

Logistic Regression (LR) penalty: 'l2' , regularization: '0.1' , solver: 'lbfgs'

Multilayer Perceptron (MLP) layer size: '(8,5,5)' , learning rate: '0.1' , activation function: 'tanh'

Random forest (RF) estimators: '200' , loss function: 'gini' , max depth: '2'

Support Vector Machine(SVM) regularization: '1.0' , kernel: 'linear' , degree: '2'

K-Nearest Neighbours (KNN) k-neighbours: ’5’ , weight function: ’distance’ , p distance: ’Manhattan distance’

Decision Tree (DT) loss function: ’entropy’ , max depth: ’6’ , min split: ’2’

NOTE: 80% of data will be used for training instead of 70% to have at least one observation for the ’very high risk’ class (recession of 2008)

Auroc Values

models	1 Quarter	2 Quarter	3 Quarter	4 Quarter	5 Quarter	6 Quarter
Logistic Regression	0.916111	0.871581	0.837693	0.779132	0.698573	0.608502
MLP	0.894444	0.758918	0.695898	0.984839	0.630797	0.547265
Decision Tree	0.729167	0.565696	0.525862	0.504310	0.439655	0.439655
Random Forest	0.965000	0.947979	0.953032	0.956302	0.862961	0.863258
SVM	0.864167	0.829964	0.788050	0.701249	0.629013	0.443222
KNN	0.897778	0.877973	0.864893	0.890309	0.891795	0.896254

Auroc Gains

models	1 Quarter	2 Quarter	3 Quarter	4 Quarter	5 Quarter	6 Quarter
Logistic Regression	-0.048889	-0.076397	-0.115339	-0.205707	-0.193222	-0.287753
MLP	-0.070556	-0.189061	-0.257134	0.000000	-0.260999	-0.348989
Decision Tree	-0.235833	-0.382283	-0.427170	-0.480529	-0.452140	-0.456599
Random Forest	0.000000	0.000000	0.000000	-0.028537	-0.028835	-0.032996
SVM	-0.100833	-0.118014	-0.164982	-0.283591	-0.262782	-0.453032
KNN	-0.067222	-0.070006	-0.088139	-0.094530	0.000000	0.000000

Wald test statistic between models

models	Logistic Regression	MLP	Decision Tree	Random Forest	SVM	KNN
Logistic Regression	0.000000	0.292841	4.561755	-2.038338	1.818942	-0.923359
MLP	-0.292841	0.000000	1.804721	-1.473965	0.340198	-0.902881
Decision Tree	-4.561755	-1.804721	0.000000	-5.410466	-1.974568	-3.600187
Random Forest	2.038338	1.473965	5.410466	0.000000	2.010162	0.773316
SVM	-1.818942	-0.340198	1.974568	-2.010162	0.000000	-1.199718
KNN	0.923359	0.902881	3.600187	-0.773316	1.199718	0.000000

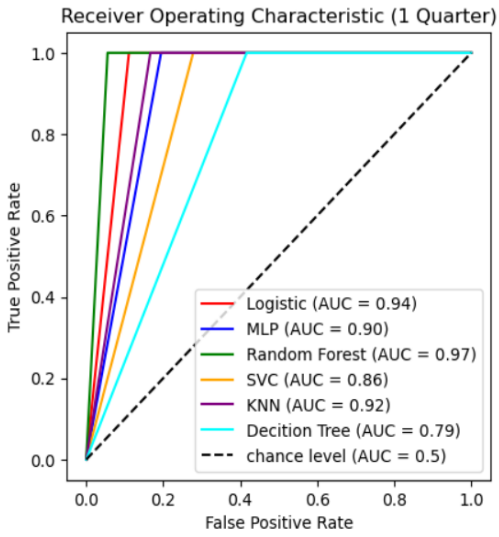
As we can see from our results , the model that better classifies the bank risk failures over the 5 'bucket risk' classes is the Random Forest Classifier. By using the Wald test we can see how the score against all other models never gets negative. On the other hand , with huge forecast horizons other models such as MLP and KNN may perform better in terms of Auroc values.

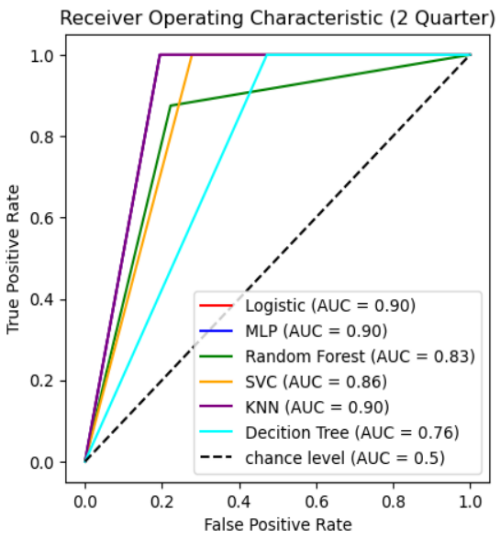
Comparing the ROC curve for 'high risk' failures

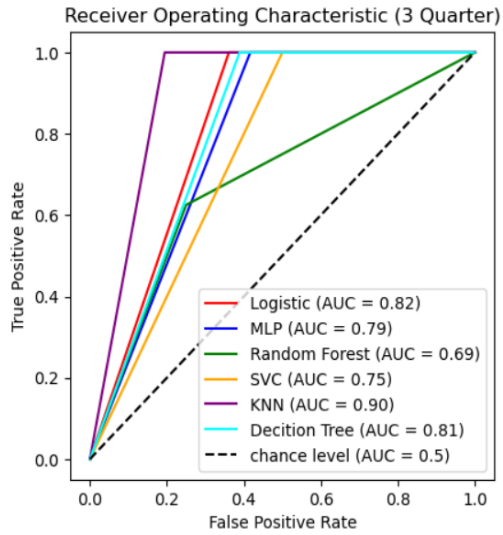
Let's focus our predictions only for the high risk failures. In this case we are treating the problem as a binary classification: 'risk' vs 'no risk'

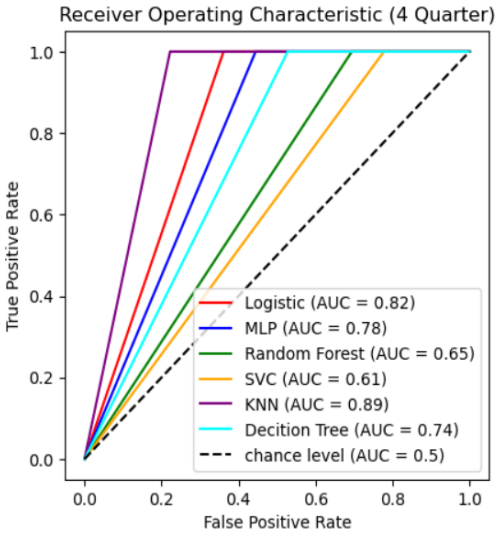
Threshold (above 80 percentile dataset) Risk failure rate > 0.1957

Forecast horizons: 4 Quarters





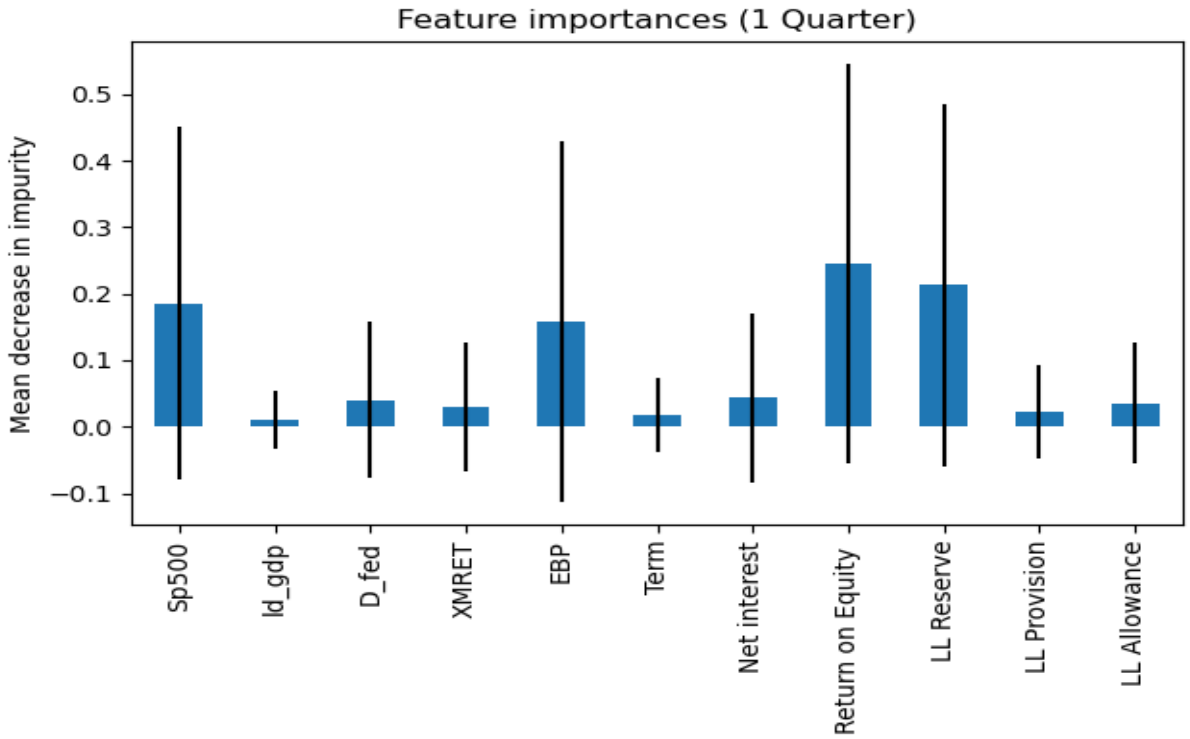


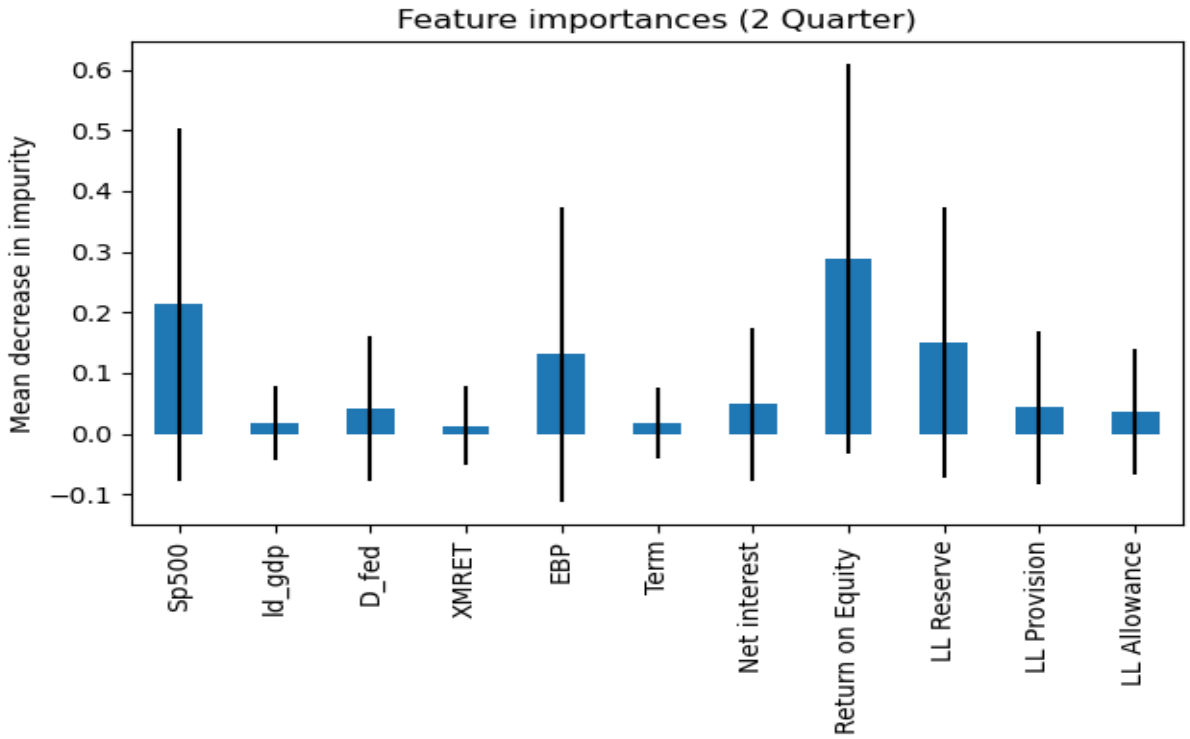


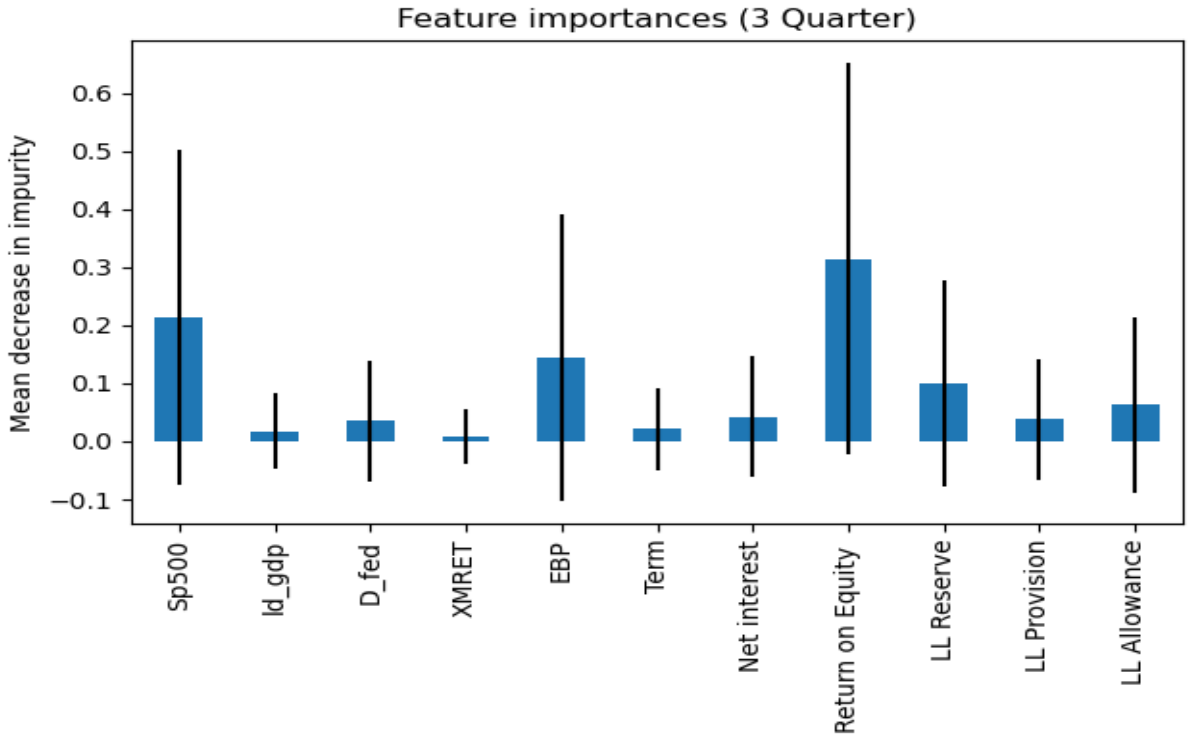
Even in this case , the Random Forest Classifier tends to lose performance in increasing the forecast horizons

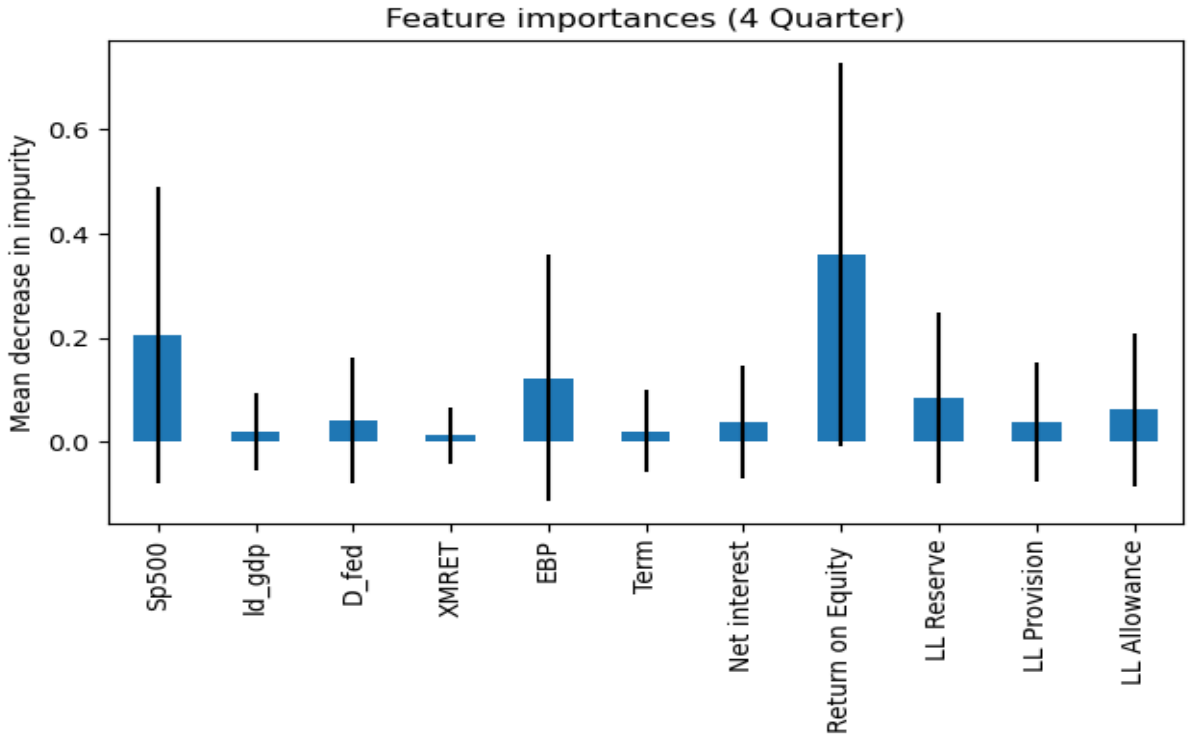
Variable contributions and ranking

Model: Random Forest









The variable that has contributed the most in the classification of bank risk is the Return on Equity (ROE) , followed by the SP500, Loan Loss Reserve (variable related to banks) and EBP. We can also see how the Loan Loss Reserve becomes less and less relevant for increasing horizons. The ROE as a key factor in the prediction of bank failure rate is no surprise, since it carries a lot of informations about all the other variables in the dataset, this can be seen from the initial correlation matrix, in which the ROE has a nice amount of correlation with all the other variables.