MECON 6102

Problem Set 2 – Risk Analysis

Version: 2024/03/03

Due Date: 2024/04/25

Feel free to ask about algorithm-related problems. Please learn to debug your code using Google, GitHub, Stackoverflow, and the package manual.

This problem set aims to provide some experience applying econometric methods in risk analysis. The data set “credit_risk.csv” is available to you online. Your main task is establishing models that predict the default label using all other covariates.

A table of variable explanations is here:

Variable Name	Note	Explanations
age	Age of borrower	Age in number of years
edu	Education level	0: below high school, 1: high school, 2: college, 3: master, 4: above master
gender	Gender	0: female, 1: male
housing	Housing ownership	0: not own, 1: own
income	Income	Monthly income-level
job_occupation	Job type	0: unemployed/temporarily employed, 1: employed, 2: manager/senior worker
past_bad_credit	Historical default label	0: non-default, 1:default
married	Marital status	0: unmarried, 1: married
default_label	Default indicator	0: non-default, 1:default

1. Simple Logistic Model

Run logistic regression: regress default label on income and past bad credit. Summarize your result. Obtain prediction values in the regression above. Compute and plot the ROC curve. Compute AUC value. Explain your main results.

2. Full Logistic Model

Now you are free to use any variables and transformations you wish. Try to obtain a model with as high AUC as possible. Report the result from logistic regression, AUC value, and the ROC plot as your main result (similar to what you did in the previous simple task).

Hint: You might want to try to use interaction terms or non-linear transformations (polynomials, log transformations, dummy variables, etc.).

3. SVM/Random Forest

You might wonder whether non-linearity can help. Try SVM or Random Forests method. You can select either one. Then, report the key parameters of your model, the AUC value, and the ROC plot as your main result.

4. Out-of-Sample Test

You might worry that your results may be over-fitting. Consequently, you decide to split the sample into two parts, use one part to train a model, and test the model in the test sample.

Take the first 10,000 samples as your training sample, and use the rest as the test. Run a logistic regression/SVM/Random Forests in the training sample based on your model in Section 2 and report your model. Then, apply the model to the test sample to create an out-of-sample prediction score for each observation in the test sample. Report out-of-sample AUC and ROC.

You may also try to resemble different models to pursue a higher out-of-sample AUC.

Scores of this assignment is heavily related to your out-of-sample AUC score for this question and the description of your optimization logic.

5. Deliverable

The final report should contain results generated by your program. It would help if you properly visualize them and provide interpretations of the results, for example, explaining why a factor can predict default or what your logic is in pursuing higher AUC.

You should submit a ZIP file containing ONE code file (.py or .ipynb) and ONE analysis report (.pdf).