QRM Course, Assignment 4

The goal of this assignment is to build a credit scoring model (or, rather, default prediction model) on the basis of logistic regression and machine learning methods.

Task number one is to find a suitable dataset for which the model will be built.

The following sources can be used:

Lending Club

Fannie Mae and Freddy Mac (mortgage companies in the US)

Kaggle has dozens if not hundreds of loan datasets: of personal loans, car loans, credit card loans, mortgages etc.

The kind of dataset you are looking for is: it has to contain (lots of) records of loan holders with some of them defaulted, and it has to have individual loan or loan holder characteristics which will serve as explanatory variables in your logistic regression.

The more interesting dataset you find and use, the higher your assignment grade will be.

Most of the datasets you find online are quite big, so if it has millions of records, that is maybe too much for your assignment: in that case, take a part of it, or perform under-sampling of non-defaulted loans so that you get more balanced dataset.

First step is to clean and preprocess your dataset.

Then fit logistic regression and choose the best model by stepwise selection.

Report all the diagnostics for your model, including the correct residuals and their diagnostic.

Interpret the values of significant coefficients in terms of odds ratios of default.

Train at least one but better two ML methods on your dataset, it can be SVM, tree-based method or GB. Find important features by means of SHAP values, compare performance on training and test sets, tune important parameters on a validation subset, and perform the usual model diagnostics, as you did for logistic regression.

Write a concise report summarizing your data processing steps, model selection steps, diagnostics and interpretation of the final model with your recommendation about which model is in your opinion the best. The most important thing in your report (apart from actual results being correct of course) is your use of visual tools, i.e., graphs of different varieties illustrating your data and findings.

Feel free to use any software, so Matlab, Python, R or any other package that have good logistic regression and ML capabilities.