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.

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, mortgages etc.

I also published a dataset on Canvas so that you can see what kind of dataset you are looking for: it has to contain 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. If you are struggling with finding a right dataset, feel free to use the one on Canvas but the more interesting dataset you find and use, the higher your assignment grade will be.

First step is to clean and preprocess your dataset. Most of the datasets you find online are quite big, so if it has hundreds of thousands of records, that is maybe too much for your assignment: take just a part of it, or perform undersampling of non-defaulted loans so that you get more balanced 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.

Write a concise report summarizing your data processing steps, model selection steps, diagnostics and interpretation of the final model.

Feel free to use any software, so Matlab, Python, SPSS, SAS, R or any other packages that have good logistic regression capabilities.