Lending Club loan portfolio construction

The due date for the presentation slides is **Sunday, 31.3.2019, at 23:59.**

The due date for the project is **Saturday, 20.04.2019, at 23:59**.

Lending club is the world's largest online peer-to-peer lending platform, where borrowers can apply for loans and investors can select which loans they want to invest in (see https://www.lendingclub.com/ for more information). The borrowers provide detailed information about themselves and the purpose of the loans in their applications. Based on this information, Lending Club calculates a credit score for each loan. These scores are used to determine the interest rates for the loans, based on which investors make money.

Lending Club publishes almost all data on the loans including margin (i.e., interest rate), status of payments, and borrower details. This enables the investors to apply data science methods to construct portfolios that maximize expected returns. For the purpose of this project, the original raw data have been filtered. There are three data files available on the course website:

- `LC_TrainingData.csv`: The "Training dataset" has 103 546 rows (each describing one loan) and 21 columns for different loan attributes.
- `LC_TestData.csv`: The "Test dataset" has 96 780 rows and 22 columns. The extra column compared to the training dataset is the loan's interest rate (i.e., loan margin).
- `Column_descriptions.xlsx`: This file contains descriptions of the columns in the data files.

In this project, the objective is to construct optimal portfolios from the loans in the Test dataset. To do this, you must first build a model that predicts the default probability of any loan based on its attributes. Using the estimated default probabilities, you may then calculate the expected returns for all loan applications, and finally apply optimization tools to construct an optimal portfolio given relevant constraints.

In particular, the project is split into seven sub-tasks that should be completed. (Remember to also look at the last page!)

1. Prepare the data by e.g. checking for missing values, outliers and other potential issues that might influence fitting the data for next states. Justify your preprocessing choices. (20 pts)

2. Using the Training dataset, apply two different methods to build a model that predicts the probability of default (PD) of a loan based on its attributes. Applied methods could be, e.g., logistic Lasso-regression and Support Vector Machine (SVM).

   The "loan status" column tells whether the loan payments are in time or not. The loan status may have many different values. For simplicity, assume that the loan is in default if the payments are late in any way (grace period, late, charged off, or default). The loan is not in default if the loan status is either Current or Fully Paid. Assume also that the loan status for each loan describes the situation exactly 1 year after loan issuance. Hence the estimated PD is the 1-year probability of default. (20 pts)

HINT: You can also apply Cross Validation in your training set to see how the model generalizes to new data.

3. Next, apply the developed models to estimate the default probability of each loan in the Test dataset. Assume that the loss given default (LGD) is 40% for all loans (which is inline with the estimates published by Lending Club). Use PD and LGD to calculate the expected relative loss (ERL) for all loans: ERL = PD * LGD. Using the loan margin for each loan in the Test dataset, calculate the Expected Rate of Return = Loan Margin – Expected Relative Loss. (20 pts)

4. Next, build an optimal portfolio from the loans in the Test dataset. In particular, formulate and solve two optimization models that maximize the expected return (in USD) of the loan portfolio such that each model corresponds to one of the PD-models built in Step 1. Use the following constraints:

   (a) The budget should not exceed 5 million USD,
   (b) The amount invested in a single loan can be anything between zero and the loan amount,
   (c) For diversification, a maximum of 500 000 USD can be invested in any single US state.

   (20 pts)

5. Finally, calculate the realized return of both portfolios using the Test dataset. Compare the results to the average return of all loans in the Test dataset. In these calculations you can again assume an LGD of 40% for all loans. Also in the Test dataset, the loan status for each loan describes the situation exactly 1 year after loan issuance. (5 pts)

6. Sharpe Ratio Maximization: Above, the optimization objective was to maximize the expected return of the portfolio. In many applications, portfolio optimization has two objectives: maximize return and minimize risk. One way to consider these two aspects is to maximize the *Sharpe ratio* of the portfolio. It is given as

$$S = \frac{\mathbb{E}[R - R_f]}{\sqrt{\mathrm{Var}[R]}},$$

   where R is the expected return of the portfolio and R_f is the risk-free rate. Assume in this case that $R_f = 0$. Also assume that the returns from different loans are uncorrelated. Hence, it holds that

$$\mathrm{Var}\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} \mathrm{Var}(X_i)\,.$$

   Note that the variance of the return of a single loan can be calculated from the PD, LGD, Loan margin, and Loan Amount. Just apply the definition of variance for a general discrete distribution.

   The Sharpe ratio objective function is not linear and not necessarily concave either, whereby the problem may be difficult to solve. The following link describes how the Sharpe ratio maximization can be reduced to a convex quadratic problem: http://people.stat.sc.edu/sshen/events/backtesting/reference/maximizing%20the%20sharpe%20ratio.pdf.

The majority of points in this last bit will be given based on problem formulation and discussion on its characteristics: e.g. "what might be the difficulties rising from such formulation?" (5 pts)

HINT: Convex quadratic optimization problems can be solved e.g. with Gurobi.

7. Provide an informative and clear final analysis of the obtained results together with a summary and critical assessment of the applied methods. Remember, a good report is not necessarily a long one. (5 pts)