AML&PM – Group Work

The dataset we use is called «Lending Club Loan Data» and can be downloaded from *kaggle* (https://www.kaggle.com/wendykan/lending-club-loan-data). Here you can also find some analyses and predictions done on this dataset, in Python notebook.

The data set describes data for all personal loans issued through the https://www.lendingclub.com/ website that operates an online credit marketplace for the years 2007-2015, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. It includes 887k observations and 75 variables.

The assignment is divided into two main parts:

- **Part 1:** We aim at finding a regression for estimating the interest rate applied to a particular lending request.
- **Part 2:** We would like to find a classification model for the default status.

Document in a text document or any other approach you deem useful (such as R notebook) the following:

- All the operations you apply;
- An explanation / justification of why you apply these operations;
- An interpretation of your results (all of them, including partial results) .

Each group will use only a part of the total dataset, obtained by subsetting it using the following line of code (supposing that your group ID is t)

```
Your_dataset <- lending_dataset[which(lending_dataset$id%%8+1==t),]
```

Part 1:

Preparatory tasks:

- Create a copy of your dataset, eliminating the entries that have an "na" in the interest rate variable `int_rate`. (Interest rate is used as output variable).
- Apply the "*validation set approach*" to reserve a meaningful amount of data for the test phase.
- Using one of the approaches for *model selection* discussed in class, reduce the number of predictors. For interpretability reasons, start with approaches that conserve the original predictor space. If any useful significant subset is possible, use a *base transformation*.
- Compute the *correlation matrix* for the selected set of predictors and the output variable, if useful, also using *graphical representation*.

Main task:

- Compare *three different methods to perform regression*, using the *cross-validation* method to compute the best parameters. Consider using some *regularization* for the parameters shrinkage. Test the train error rate, the CV error rate and the test error.

Part 2:

Preparatory tasks:

- Our goal in the second part of the assignment is to predict if a new customer will be able to fully pay back their loans using a classification method. Thus, we concentrate on the "concluded lends" in the data set, i.e., on all lends whose `loan_status` is not `Current`. To this end, filter out all observations with `loan_status == Current`.
- For the remaining observations, check if the `loan_status` is "`Fully Paid`". If not, change the value of `loan_status` to "`DEFAULTED`".
- Create a validation set.

Main tasks:

- Use *Principal Component Analysis* for *base transformation* and then compare it with the *Partial Least Squares Regression* result. Select the best base with cross validation, using the better of the two approaches.
- Perform the classification using *KNN*, *Logistic Regression*, *Decision tree* and *Random forest*.
- Compare the respective train and test error performances to select one of these approaches.
- Perform the prediction on the validation set and compute the *confusion matrix*.
- Conceptually compare your approach with a solution existing for this problem. (Default prediction is a very well-known problem in literature).

Other useful resources:

- https://nycdatascience.com/blog/r/p2p-loan-data-analysis-using-lending-club-data/
- http://blog.yhat.com/posts/machine-learning-for-predicting-bad-loans.html
- https://medium.com/@jiaminhan/peer-to-peer-loan-default-prediction-using-lending-club-data-3f75886cb1e
- https://www.datasciencecentral.com/profiles/blogs/analysis-of-lending-club-s-data