DAM Assignment 2 – Loan Default Model

Background

This task is to predict which loan customers will default on their repayments. All customer records provided relate to personal loans that have issued. There are a large number of predictors for the loan status target (either 'Fully Paid' or 'Charged Off'). The type of data provided and the modelling problem are commonly seen in the financial lending industry.

The data set contains a total of 42 columns and 39,786 rows. A description of all the columns and their values is provided in the file 'data_dictionary.csv'. The column 'loan_status' is the modelling target, and the other 41 columns are candidate predictors. Variations of the data set are publicly available, and a kaggle competition also uses this data:

https://www.kaggle.com/wendykan/lending-club-loan-data

You are free to explore previous work that has been completed on this data. However, caution is advised in using other people's solutions since the data provided in this assignment has been modified.

There are two key deliverables for this assignment, Part A and Part B.

Part A – Modelling

Work in groups to create a model to predict which customers are likely to have loan status of 'Charged Off'. The data and submission process are managed via a kaggle competition. There will also be a live leaderboard. The link to the competition is here:

https://kaggle.com/c/dam-spring2017-assignment2

There are two data sets, one training and one validation:

- training.csv
- validation.csv

You should train and test your model on the training data set, and run probability predictions on the validation data set. The performance of your model will be evaluated using the AUC measure (Area Under the ROC Curve) for binary classification models on the validation data set. One part of the validation data is public and will be made visible once you submit your predictions, but a second part is private and will be withheld until the assignment is finished.

Your team is allowed up to 100 submissions (you shouldn't need all of those), and up to 10 submissions per day.

Part B – Management Presentation

You are required to individually submit a management presentation on your approach. You should discuss the following, in line with CRISP-DM:

- The business problem
- The available data
- Your data preparation process
- Details of your model training, including the assumptions you made with a rationale
- Your evaluation methodology, including the model's performance as both a classifier (class predictions) and probability predictions
- Preliminary results (kaggle public evaluation measures)
- Any particular insights you discovered about the data
- Consideration of ethical issues

Your presentation should be short and concise, no more than 10 slides.