CQF Module 4.6: Decision Trees & Ensemble Models Exercises

Panos Parpas

The aim of this exercise is to encourage you to use the methods we saw in class on a more realistic application than the Asterix dataset. We will use data provided by the company LendingClub. The data has been used on several research papers see e.g. Jagtiani and Lemieux (2019); Malekipirbazari and Aksakalli (2015). The exercise is based on material from Hull (2020). This data was made available from LendingClub when its main business was peer-to-peer lending. In 2020 the company has abandoned the peer-to-peer lending platform.

We will use the dataset to develop a classification algorithm to determine whether borrowers should be accepted or rejected for a loan.

The full data is in the following file: `l e ndingclubFullDataSet.xlsx`. Y ou can read the meaning of each column by examining the file: `lendingclubdatadictionary.xlsx`.

We will focus on loans that have been completed either as "Fully Paid" or "Charged Off" in Column 'O'. The problem is to develop a classification algorithm to predict which loans will be "Fully Paid" and which will be "Charged Off".

Out of the 25000 loans there are 12,290 loans that have already been classified. We will concentrate on this dataset. To make the exercise a bit simpler the data (of 12,290 loans) is already split into a train/test/validate datasets in the following files:

- `lendingclubtraindata.xlsx`
- `lendingclubtestdata.xlsx`
- `lendingclubvaldata.xlsx`

To be concrete we associate a value of 1 when the loan was paid and 0 when it was not (see column E in the excel files above`loan_status`). We also dropped most of the columns from the dataset and just focus on the ones below.

- Home ownership
- Income per annum (USD)
- Dept to income ratio (in %)

Since home ownership is categorical we used 0 if the borrower was renting and 1 if they owned a home.

For the exercises below you can use the provided Python script to perform the calculations.

Exercise 1 Using the data from the training data set calculate the initial entropy of the dataset using the empirical probability that the loan is paid of.

Exercise 2 Calculate the information gain (i.e. reduction in expected entropy) if the home ownership feature is used to build a simple classification tree.

Exercise 3 Use the `DecisionTreeClassifier` from the Python library `sklearn` to produce a decision tree. Experiment with different parameters. Use the parameters:
`criterion='entropy',max_depth=4,min_samples_split=1000`
`,min_samples_leaf=200,random_state=0`,
to produce an initial tree. Experiment with the different techniques we discussed in the lecture to improve the initial tree. You could use other features from the dataset, and/or other algorithms. You can also develop a regression tree by trying to predict the level of interest to be applied on a loan.

References

Hull, J. C. (2020). Machine learning in business: An introduction to the world of data science.

Jagtiani, J. and Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: evidence from the lendingclub consumer platform. *Financial Management*, 48(4):1009–1029.

Malekipirbazari, M. and Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications*, 42(10):4621–4631.