Application test – Data Scientist

The present dataset describes a set of “open items”, in which every row can be assigned to exactly one independent individual/company. An “open item” is an invoice which has by been issued by a vendor for a debtor. Every open item contains several information:

- Invoice amount
- Due date at the end of the month?
- Debtor’s industry
- Debtor’s registered office
- How many open items did the debtor issued before? How many where paid on time?
- Was this open item paid on time? How long was the delay?

Exercise

Construct a forecast model based on this dataset which predicts the payment delay of open items.

An open item is settled on time or with a delay of 1-30 / 31-60 / 61-90 / >90 days respectively. Measure the quality of your model / models. You can choose any arbitrary approach. Make sure that you provide a piece of code which represents your solution (Python, any library or framework can be used). The final valuation focuses on your functional approach as well as on your code quality. Since you will present your solution from your own workstation, it is ok to use your favourite data science tool, as long as code is written in Python.

The dataset includes 40000 rows and 164 columns. Every attribute (excl. target variable) is standardised and processed by means of the One-Hot-Encoding-Procedure. Every row references an open item originating from one vendor. The intended debtor changes within the dataset. The semantic of the attributes is outlined in the following:

- Amount → Amount of the invoice
- Payment_period → Period from invoicing to payment target
- Due_date_month-end → Was the invoice due to month-end?
- Payment_delay → Delay of the invoice settlement
- N_items → Number of previous invoices
- N_items_late → Number of previous invoices paid late

  …
- Amount_euro_recently → Amount of the recently settled invoice

  …
- Amount_POOL → For what amount has the customer recently purchased from all creditors?
- Amount_L → For what amount has the customer recently purchased from only this creditor?
- Amount_PoL → For what amount has the customer recently purchased from all creditors except the current one?

  …
- Unemployment_rate → Unemployment rate when the invoice was due at the specific region
- Offsetting_account_XYZ → Invoice was posted on offsetting account XYZ?

- Due_date _month_2.0 → Invoice was due to February?
- Sector_ABC → Debtor operates in sector ABC?
- Headquarter_Hessen → Debtor's headquarter is in Hessen?

Background

Financial risks related to the extension of supplier credits can be reduced from the creditor's viewpoint if payment default or delay is predictable at an early stage. Using data-mining-methods to build forecast models can improve the liquidity, planning ability and profit of companies in a significant way.