Context: The energy industry is under significant pressure - given the current macroeconomic context and global instability. This is transposed in price increases, which are then reflected across most economies and industries. The sharp price increase in both electricity and gas put a big financial pressure, particularly on the households with an already existing financial burden.

Problem: Detecting (financial) vulnerable customers using clustering (unsupervised).

Usability: Having such information can help companies tailor marketing approaches, or campaign-specific products that help the households (e.g. fixed price quotas, distribution of support funds); as well as communicate useful information (e.g. introducing smart meters, energy-saving tips).

1. What data would you use?

Primarily, I’d use customer data: demographic and transactional. Depending on what is available features such as: payment, usage, billing/debt information, tenure.

1. What are your key input and output variables? (Review section 2.4.)

Input: Both numerical (number of payments, total consumption, debt value, household size, tenure) and categorical data (e.g dwelling type)

Output: Since I see this as an unsupervised ML problem, the output variable would be the cluster number. Output is a categorical data type, but the problem is not a classification problem as the classes are not pre-defined.

1. What type of machine learning problem is this? (Review section 2.5.)

This is a clustering method (unsupervised machine learning). We try to understand who might be the vulnerable customers without any prior knowledge (pre-existing flag). We don’t have a dependent variable.

1. What steps would you take to solve this problem through machine learning? (Review section 2.6.)

I’d follow the steps outlined in the section 2.6, although the implementation wouldn’t be so linear. I would repeat the feature engineering a couple of times, until I’d get a meaningful cluster interpretation/feature summary. Data wouldn’t need to be partitioned into training and testing as this is an unsupervised problem.

1. Define the purpose of the ML project

This would start with the need for identifying the vulnerable customers, and use this information for decision making. I see this more as an ongoing (batch processed) model, because classes might change, and statuses can be tracked with time.

1. Obtain the data set for the analysis

Using the internal customer database, having both transactional and demographic. External household information could also be obtained, although data is sensitive and prone to error.

1. Explore, clean and pre-process the data

I’d perform EDA and inform business on the main findings. After this, I’d decide to deal with missing data or any outliers. Then, depending on the data type, I’d transform and normalise/ standardise the dataset.

1. Dimension reduction and feature engineering

After performing EDA, running correlation or feature importance - to understand the important variables; perform dimensionality reduction - to get an idea of what features might not be needed (e.g. maybe average consumption is not a good predictor of financial vulnerability, but rather whether customer has ever been on credit) .

1. Determine the ML task at hand

Because there are no pre-defined classes and the output variable is a cluster number, the problem is an unsupervised machine learning problem.

1. Choose the ML technique(s)

I’d choose to use several clustering methods, starting with the most popular: k-means, k-medoids, DBSCAN.

1. Use the ML technique(s)

I’d use the above techniques and compare the outputs for each one of them. I’d pick the one giving the best results (clusters that make most sense and with minimum distance within clusters, maximum distance between clusters).

1. Interpret the results

Compare the algorithms against each other and validate with the business if they make sense.

1. Deploy the ML technique (optional)

I’d rather investigate best methods to deploy the selected model as well as tracking results.

1. What might cause missing data in your data set? Which approach outlined in the lecture materials do you think would be most suitable for dealing with missing data, and why?

For the initial model iteration, I’d choose to inspect missing data along with the subject matter experts as of why data might be missing. Checking for any patterns/distributions of the observations where missing data occurs. Is it a system/data quality issue or there are other causes? For this reason, I do not see imputation a good method as it can introduce bias. I’d remove the records where missing data occurs.