* Briefly outline your problem again and what data you would use.

Problem: Detecting (financially) vulnerable customers using unsupervised ML (clustering).

Data - ‘vulnerability’ indicators

- Financial: number of payments, payments mean, payments standard deviation, has/no debt, was on ‘emergency use’ flag

- Demographic/descriptive: tenure, website engagement, ‘contacted customer support’ flag

- Consumption data: average consumption, median daily consumption

- Are the assumptions of a probabilistic setting, stationarity and a priori knowledge satisfied? Justify why for each one.

* probabilistic setting:

Given the initial dataset is extracted from the entire customer population, we can assume the samples are independent as it would entail unique identifiers. However, even after normalisation, chances that the identically-distributed requirement will be satisfied are very low

* stationarity:

Same data collection principles are applied cross-organisation and no changes would impact the input data (assuming transformation methods are be replicated)

* a priori knowledge:

The problem framing is a result of industry need combined with feature discovery (EDA), and the feature pre-processing

* If they aren’t justified, how could you change your data or problem so that they are?

The identically distributed assumption might not be satisfied in the population sampling phase. This can be controlled by forcing specific sampling methods (e.g. 10% of customers from each tenure bracket) or specifying that the samples need to have to same statistics (mean, standard deviation)

If not satisfied, the stationarity assumption can be addressed in the pre-processing step using drift detection methods.