SWEDISH COMPANIES DATASET

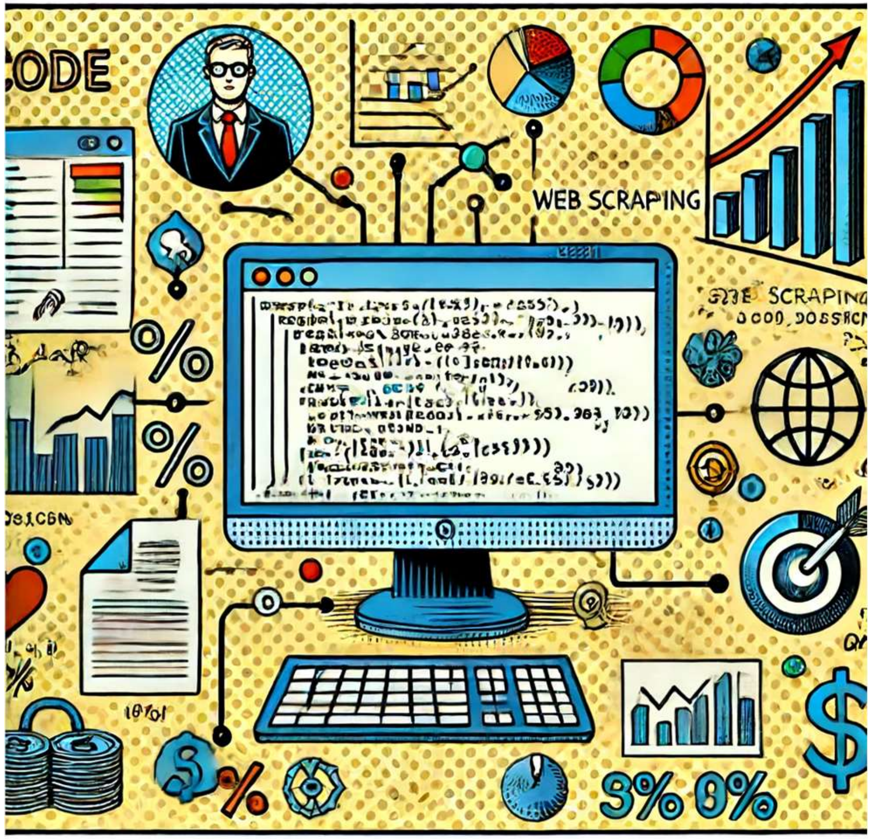
WITH A BUNCH OF MACHINE LEARNING TECHNIQUES

INTRODUCTION

- Once I completed a series of online courses about data science and machine learning, I was eager to put my knowledge into practice to show others what I am capable to achieve.
- I needed a dataset. Internet offers plenty of good ones and ready to use. However, a data scientist must be also able to collect and prepare data by himself/herself. So I thought what could be something interesting to build up. Being myself an entrepreneur and living in Sweden, I decided to create a dataset about Swedish companies and their balance sheet data.
- This project required a certain effort, during my free time: let's see what's happened!

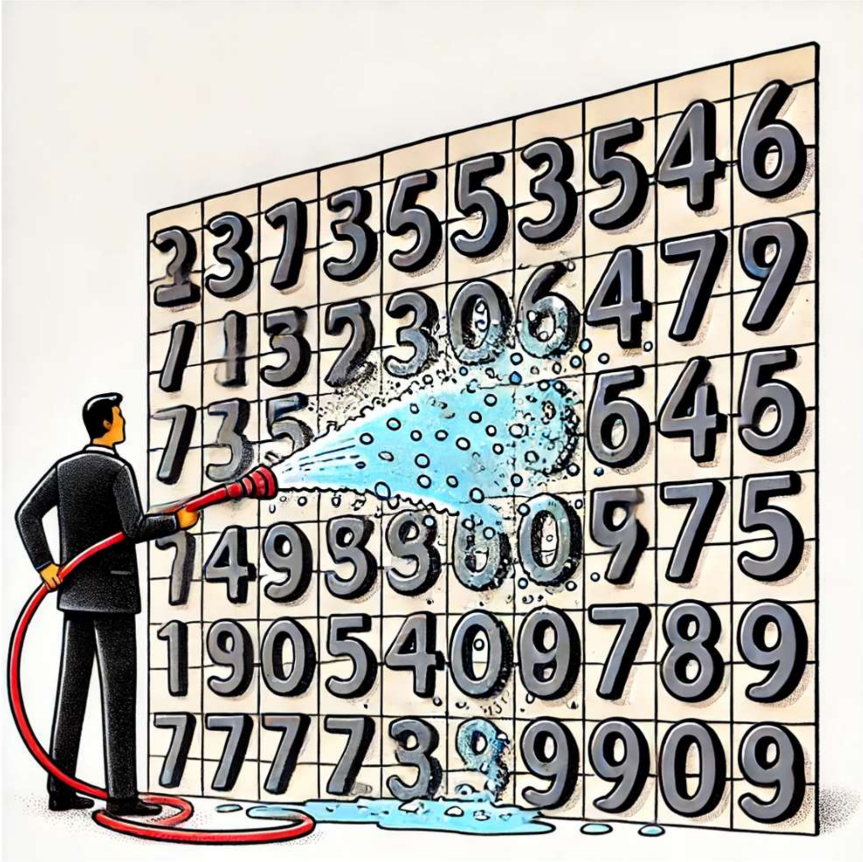
HOW THE DATASET WAS BUILT

- There are several websites that freely supply access to balance sheet data. However, they don't allow to download a complete database at once.
- Therefore I decided to create a Python script to do some web scraping on one of these sites, by using libraries such as Beautiful Soup. The process took about one week to run.
- The result is a comma-separated value file which contains (at the current date), among the other features:
  - 462,468 Swedish joint-stock companies.
  - Historical balance sheet data ranging from 2012 to 2022.
  - General data such as organization number, location, type, commercial category etc.
  - KPIs such as DuPont analysis, solvency, EBIT and so on.



DATA CLEANING – LEVEL 1

- The raw data collected by the first script is far from being suitable for machine learning. For a number of reasons, it contains missing entry points, invalid values, duplicates and more.
- So I created a second script that prepares the data. The cleaning process is performed in two steps.
- The first level drops duplicates (by organization number, an univocal identification number used in Sweden), removes irrelevant features, translates column names and activity types into English.
- Note: this first cleaning does not alter the raw values.

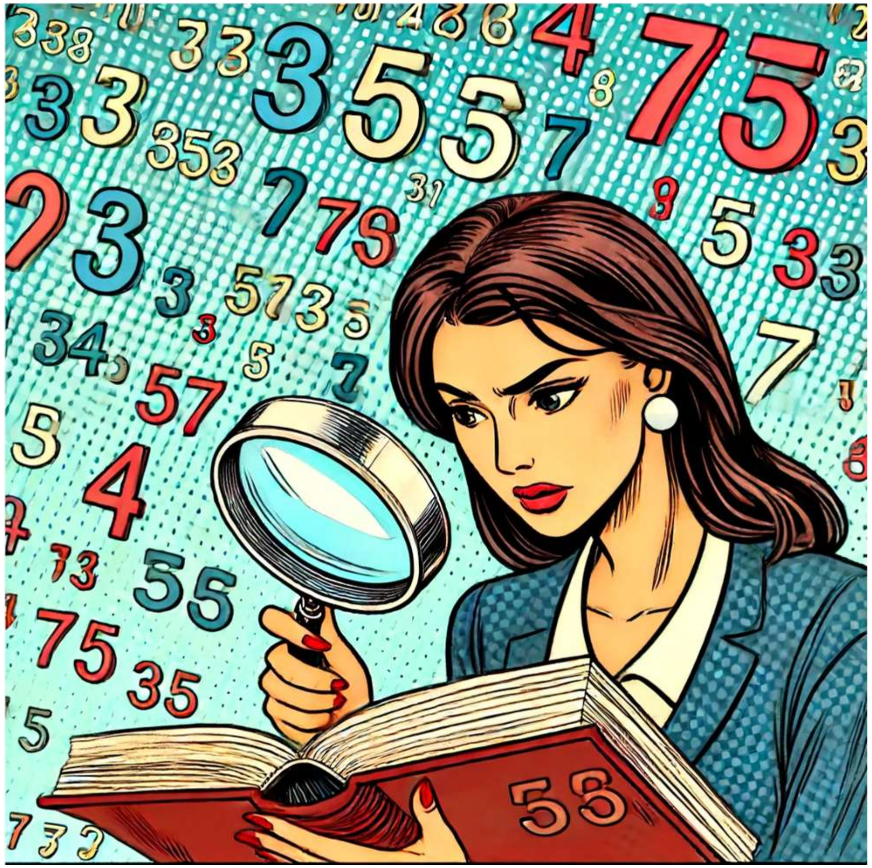


DATA CLEANING – LEVEL 2

- The second level is deeper: some features are replaced with more meaningful ones, some others (textual ones) are mapped to ordinals.
- Data points with too many missing values are deleted.
- Data points with a limited amount of missing or invalid values are instead processed. There is a series of 5 different criteria used to replace the bad data, depending by the availability of adjacent values and the median values against tailored subsets of the main dataset.

DATA ANALYSIS EXAMPLES

- The data is now clean and can be used.
- The dataset is huge but versatile. Rather than using it on its whole scale, its content can be filtered to create new, lighter datasets for specific purposes.
- It can be used for example for extracting trends grouped by commercial category or number of employees.
- The data may be also used to evaluate the state of wealth in a certain region.

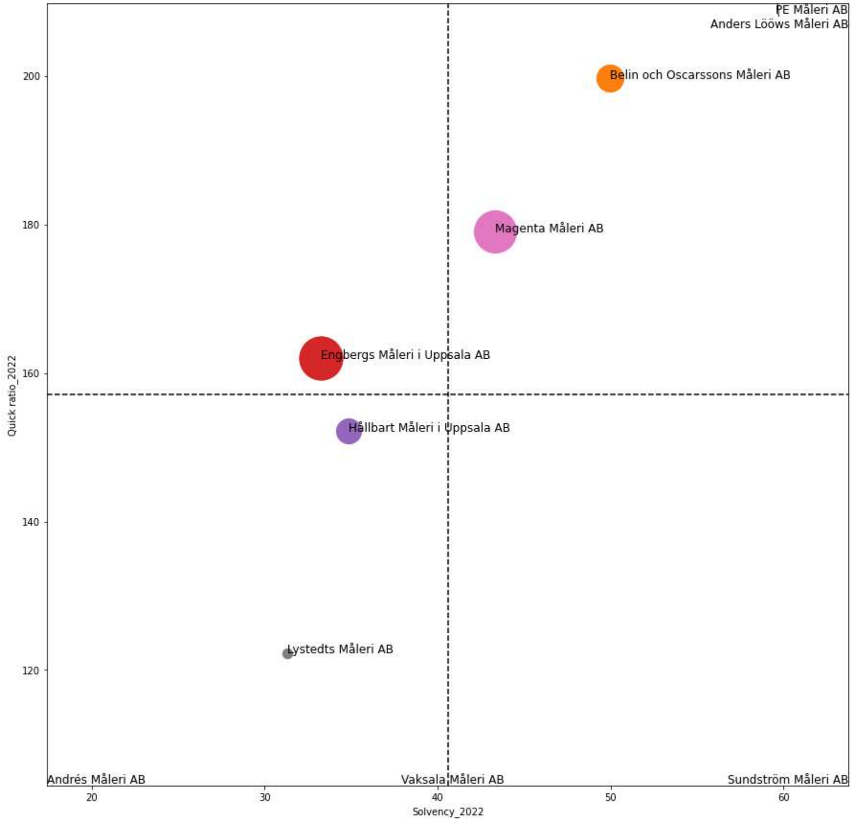


DATA ANALYSIS: A SIMPLE CASE STUDY

- Recently me and my partner decided to paint our home. It is a set of 4 buildings that would require us to spend most probably more time than a year of holidays. So we started to look for a good house painting company in Uppsala's county.
- I've got the feeling that the dataset could provide interesting insights about which painting companies better meet our needs.

DATA VISUALIZATION

- I plotted the companies as dots of different sizes (corresponding the number of employees) on a cartesian space of solvency and quick ratio and then split into 4 areas, according to the median values. In the lower left square there are, let's say, ”bad” companies, while in the upper right there are ”good” ones (so, high solvency and quick ratio).
- Is it more convenient to choose a company with low solvency and low quick ratio, which is probably facing a crisis and may likely accept a lower rate or a wealthy one? Better a small company or a big one?

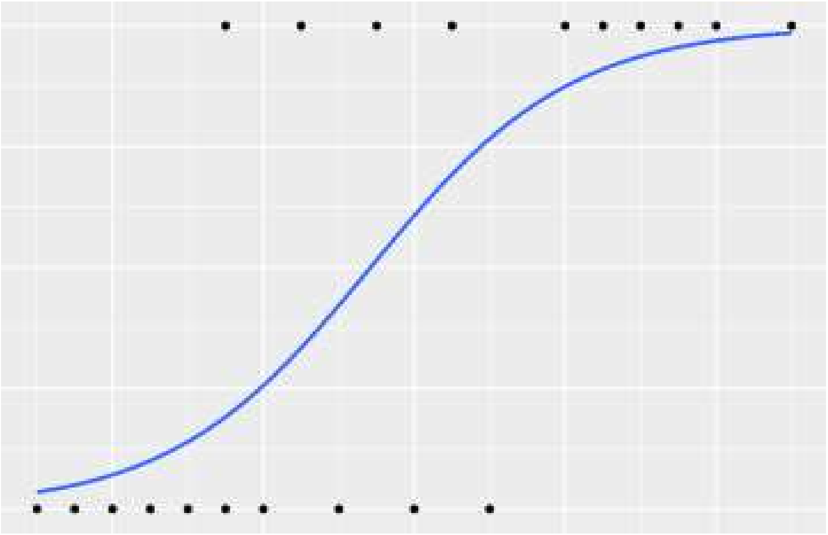


CLASSIFICATION

- I was wondering which companies would likely pay the dividends. Having such information for the past years and lots of parameters, it may be possible to make an accurate prediction.
- Let's say that my goal is to outperform my friend Giovanni. Giovanni is a guy known to not being very smart. If he would have to predict which companies pay the dividends, he would tend to apply the simplest possible rule, shaking his hands: "*If they paid the last year, then they will do it this years as well!*"
- By simulating such a rudimental model in Excel, it turns out that he gets a F1-score of 70%. That will be my threshold to discriminate between a good classification model and a bad one!

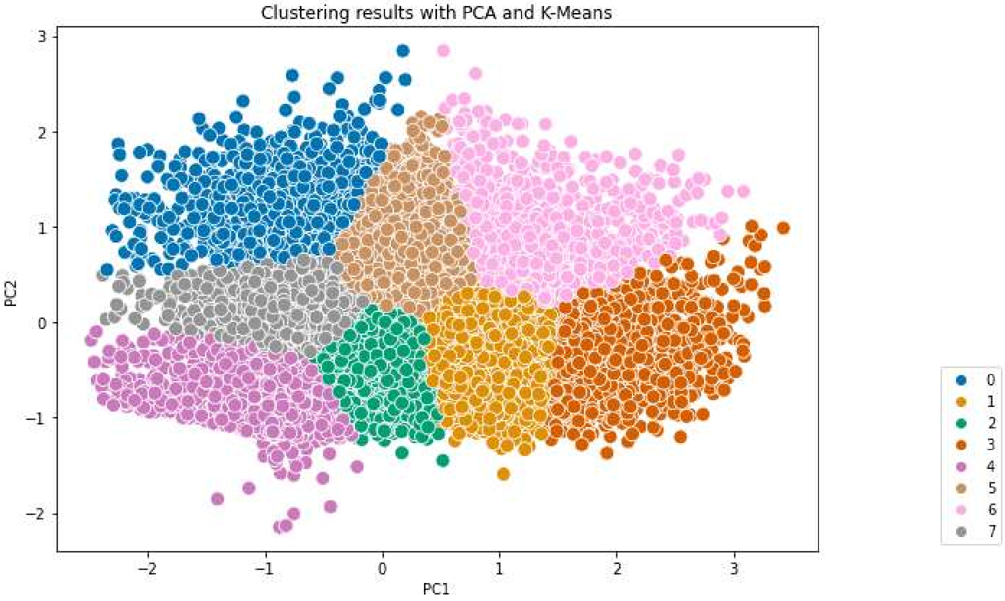
CLASSIFICATION: LOGISTIC REGRESSION

- Since my purpose is a binary classification, I chose to build a model with the logistic regression, because of its good compromise between lightness and performance.
- The model is basically made of the following parts:
  - Read the data.
  - Feature engineering, creation, reduction, selection combination and scaling.
  - Balance and weigh the classes.
  - Double search of optimal parameters: first a random search across logistic regression parameters; then a refined grid search around the previous result.
  - Training and test are performed on a reasonable subset of the dataset; the model built on this subset is then used to predict the whole dataset.
  - The final result is a fair F1-score of 85%, so the model works considerably better that Giovanni!



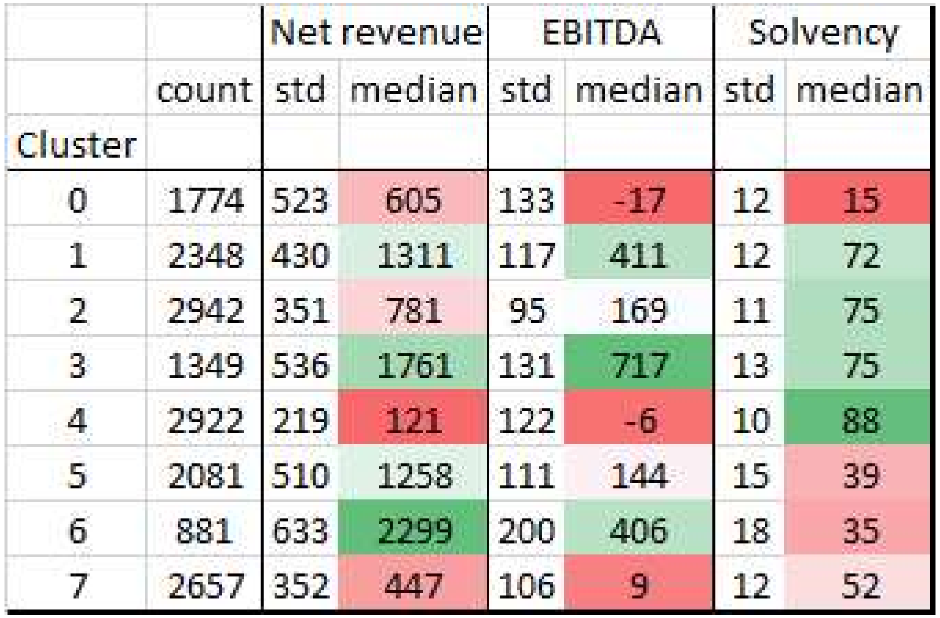
CLUSTERING

- Another thing I wanted to do is a basic clustering of the companies, considering net revenue, EBITDA and solvency.
- For this example, I limited the clustering to a subset containing only 20,000 randomly chosen active companies with 1 employee.
- K-means was used but, before applying the algorithm, som pre-processing actions were taken:
  - IQR technique to remove outliers (so keeping values between 25% and 75% quantiles).
  - Scaling with RobustScaler.
  - PCA to reduce the dimensionality.
- The results were evaluated with the Silhouette score. The best result was obtained with 8 clusters, with a score of 69%.



CLUSTERING: INTERPRETATION

- 8 clusters may appear excessive for such a limited task against EBITDA, solvency and net revenue.
- However, the clustering itself isn't necessarily meant to provide human-readable information, rather than it can be use as a middle step in a bigger classification model. By creating a feature with cluster number, a classification model may improve its performance.
- A brief look at the statistics of the clusters reveals that, for example, the cluster 4 is made by quite homogenous companies (since all the three standard deviations are relative low) with general low performance but a very high solvency.



		Net revenue		EBITDA		Solvency	
	count	std	median	std	median	std	median
Cluster							
0	1774	523	605	133	-17	12	15
1	2348	430	1311	117	411	12	72
2	2942	351	781	95	169	11	75
3	1349	536	1761	131	717	13	75
4	2922	219	121	122	-6	10	88
5	2081	510	1258	111	144	15	39
6	881	633	2299	200	406	18	35
7	2657	352	447	106	9	12	52

CONCLUSION

- That's all for now! ☺
- So, did you wonder which house painting company we chose? Well, I won't reveal it but I encourage you to use the dataset and get your own conclusions instead.
- However, this is the result...