

Finding the correct ETIM-class

INTERMEDIATE DATA SCIENCE WITH PYTHON

SPRINGBOARD CAPSTONE PROJECT

SPRING 2019

Pål Kristian Granholt



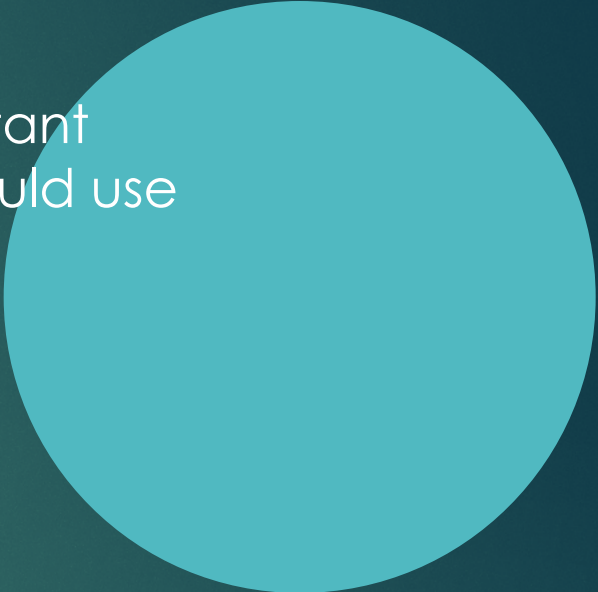
The Client

2

- ▶ Norwegian Electrical Trade Organization (EFO)
- ▶ Works closely with suppliers and wholesalers on product data
- ▶ Product database with 250,000 active products

The Assignment

3

- ▶ European Technical Information Model (ETIM) is an important classification model that all products in the database should use
 - ▶ Not always easy to classify manually
 - ▶ Create a classification model for ETIM-classes
- 

The Data

- ▶ After wrangling and cleaning, about 202,000 products remain
- ▶ About 1,800 ETIM-classes
- ▶ Heavily skewed data in most features



The Models

5

- ▶ Bag-of-words and multinomial Naïve Bayes
- ▶ Hyperparameters:
 - ▶ Norwegian stop-words
 - ▶ Ngram from 1 to 2 words
 - ▶ Laplace-smoothing with very small alpha
 - ▶ Minimum term to increase model speed



Results

6

- ▶ Model performs well considering the skewed data
- ▶ Many of the misclassifications are almost correct (i.e. closely related product, but with different ETIM-class)

| Results from the classification reports | | | |
|---|-----------|--------|----------|
| | Precision | Recall | f1-score |
| Baseline model | 0.64 | 0.64 | 0.64 |
| Extended model | 0.84 | 0.84 | 0.84 |

Recommendations

7

- ▶ Model performs well, should be usable with minor tweaks to presentation:
 - ▶ Mainly inform of inaccuracies, and show how confident the model is in its classification.
 - ▶ Show several of the top predictions if uncertain.
 - ▶ Potentially show similar products to alleviate misclassifications to similar products
- ▶ More data important for improving the model, should update the data regularly