From manual to machine: challenges in machine learning for COICOP coding

Susie Jentoft, Statistics Norway

Boriska Toth, Statistics Norway

Daniel Müller, Norwegian University of Life Sciences

Machine learning for COICOP coding: agenda

- Problem statement: COICOP coding for the Household Budget Survey
- Human-in-the-loop model
- Data
- Implementation of maching learning
- Results

Machine learning for COICOP coding

- Household Budget Survey 2022
- All goods and services classified into COICOP groups (*Classification of Individual Consumption According to Purpose*)

FLOURS

01.1.1.2

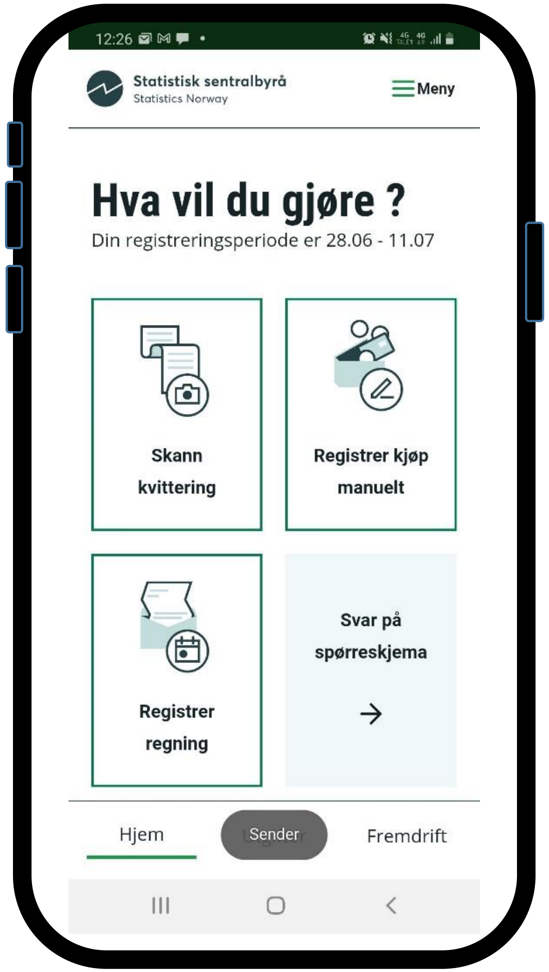
RICE

01.1.1.1

Norway’s Household Budget Survey 2022

- Innovative modernized survey: survey using a phone app, and big data sources
- More frequent publishing of HBS statistics
- Fine-grained information on purchasing habits in specific groups

- Users can scan in receipts or manually code them
- ~360,000 items from scanned receipts



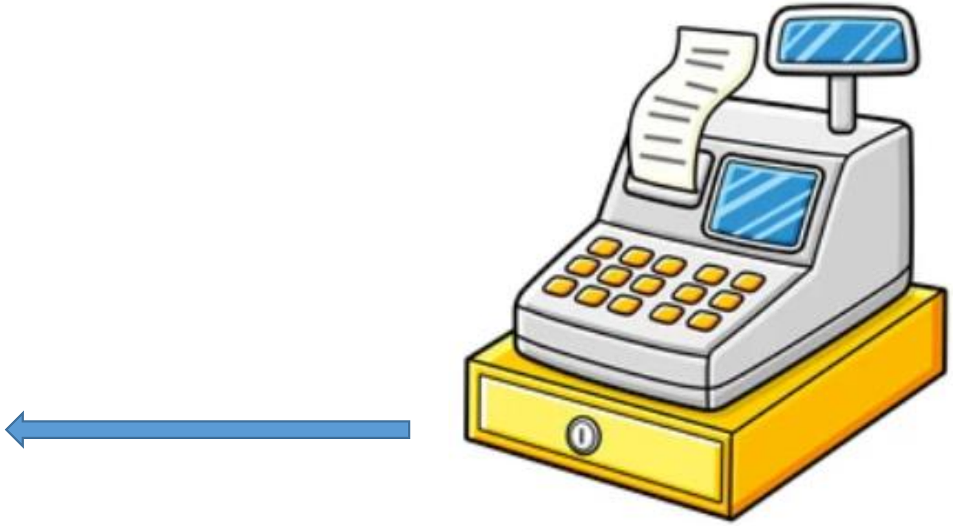
Hva vil du gjøre ?

Din registreringsperiode er 28.06 - 11.07

Norway’s Household Budget Survey 2022

- Innovative modernized survey: survey using a phone app, and big data sources
- More frequent publishing of HBS statistics
- Fine-grained information on purchasing habits in specific groups

- Purchase transaction information for all purchases made by debit card in 2022 from all of Norway’s major supermarket chains and retailer stores (~400,000 unique items)



Data source	Number of goods
Consumer price index	54,000
Other purchase transactions	3000
Manually coded by respondent	13,000
Manually labelled from scanned receipts	3000
Norwegian translation of UNSD reference	2500
Dictionary of search terms for survey app	2400
TVINN customs declarations	1,500,000

Purchase transactions data

Scanned receipts data

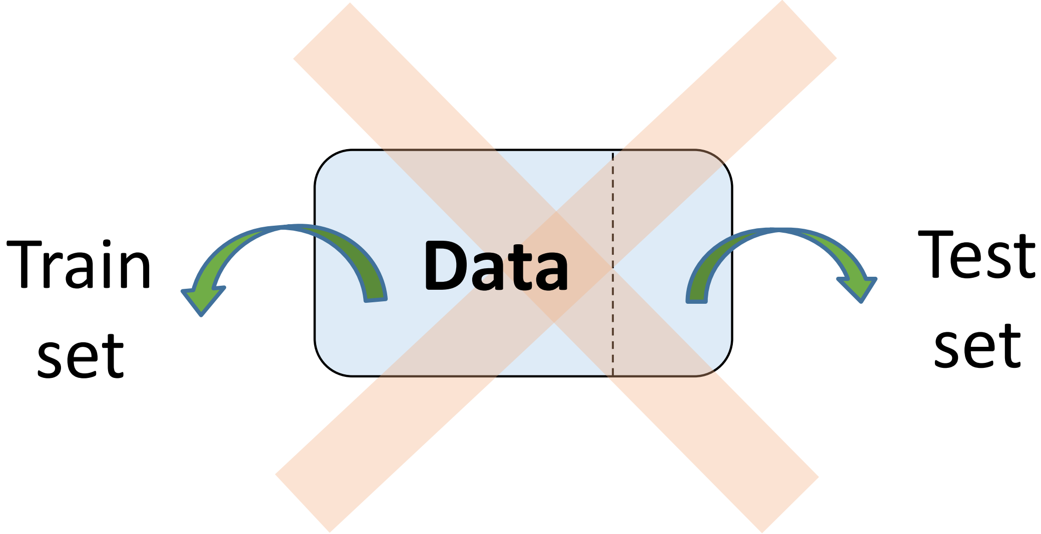
Dictionaries of keywords

Imports data

Data source	Number of goods
Consumer price index	54,000
Other purchase transactions	3000
Manually coded by respondent	13,000
Manually labelled from scanned receipts	3000
Norwegian translation of UNSD reference	2500
Dictionary of search terms for survey app	2400
TVINN customs declarations	1,500,000

Automating coding at NSI’s

- Evolving data between training and prediction
- Need for human labelling (new items, quality control)
- Many common items where lookup tables and rule-based methods perform the best

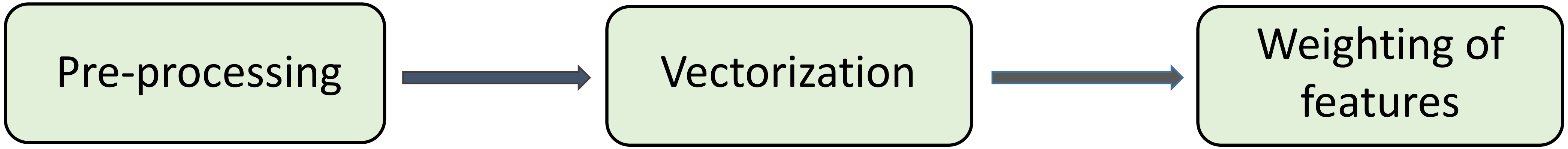


Human-in-the-loop paradigm

Define:

1. A target source of data from a specified time period that needs to be coded (i.e. transaction records and survey receipts for items purchased in 2022)
2. A human-in-the-loop procedure for how items in the target data source will be coded (machine learning, manual, lookup table, …)
3. Measures for evaluating that capture both the performance of automated coding (accuracy, F1) and the burden of human labelling

Feature generation for machine learning



Pre-processing

Vectorization

Weighting of features

1. Pre-processing (special characters, stopwords, stemming)
2. Vectorization: word grams and 2- or 3-character grams

Varenavn	Nirus	Tikka	Masala	Saritas	Curry	Paste	Masalamgic
Nirus Tikka masala	1	1	1	0	0	0	0
TIKKA MASALA Saritas	0	1	1	1	0	0	0
Masalamagic Curry Paste	0	0	0	0	1	1	1

3. Weighting: assign higher weight to rare features

Predicting COICOP – purchase transactions

Algorithm selection

Algorithm	Accuracy
SVM	0.83
Random Forest	0.76
Logistic regression	0.77
XGBoost	0.73

Training on transactions data sources, testing on a hold-out test set

Feature selection

Features	Accuracy
Product name	0.83
Product name + Group	0.89
Product name + Ingredients	0.84
Product name + Group + Ingredients	0.89
Product name + Group + Price	0.90

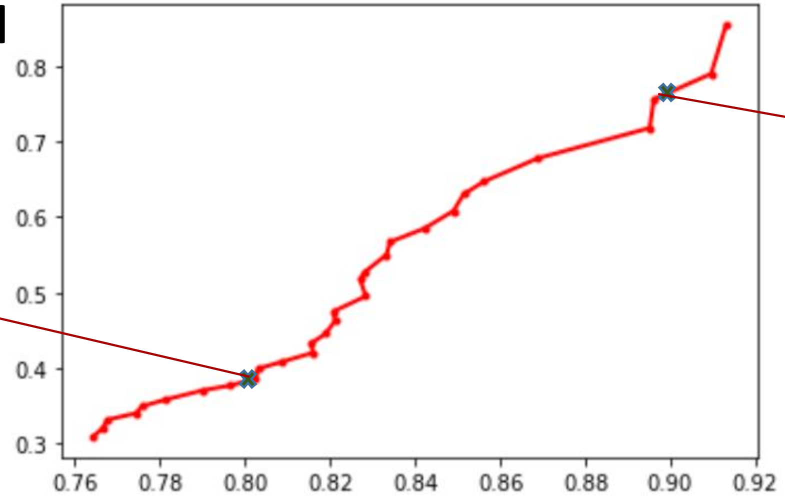
Predicting COICOP – survey data

Accuracy when training on 6 sources, testing on a random sample of 1000 manually labelled items from scanned survey receipts

Logistic regression	0.59	Random Forest	0.58

Varying the threshold for prediction probability

% manual labelling



Accuracy

80% accuracy, 62% labelled

90% accuracy, 25% labelled

Conclusions

- Machine learning is crucial for the modern, big data-based Household Budget Survey 2022
- Manual labelling is still needed, so we define a human-in-the-loop paradigm
- Good performance on transaction data; classifier needs improvement on survey data
- Future work: rule-based methods, improving the balance and representation of COICOP codes in the training data