Two applications of machine learning at Statistics Sweden

Jacob Kasche

Gustaf Strandell

Automatic coding of occupation in the Swedish occupational register

With Jens Malmros and Simon Godskesen

The occupational register contains the occupation for individuals who are employed in Sweden.

The occupation is coded according to a 4-digit code standard - SSYK

- 7111 – building carpenter
- In total 429 - classes

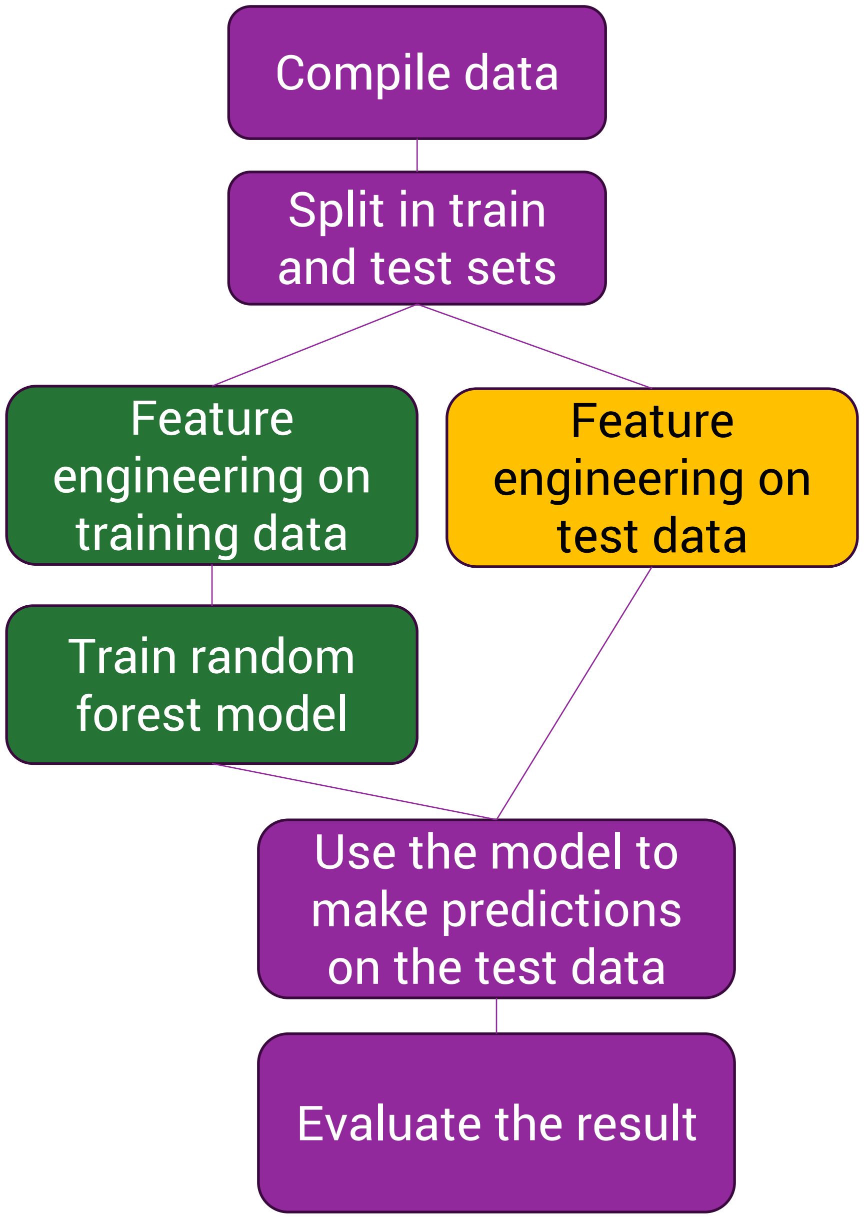
The classification task

Company: ACME AB

	Fill in:	Or:	
Name	**SSYK-code**	**Occupation/main duties**	**In english**
Reginald Dwight	2652		
Stefani Germanotta	2655		
James Osterberg		Frisör	Hairdresser
David Jones		Klipper, färgar, fönar	Cut, color, blow-dry
Vincent Furnier		Montör, VVS	Assembler, plumbing
Elizabeth Grant		Spadgubbe	Constructor
Alecia Moore		VD, ekonomi	GD, economics
Marshall Mathers III		Konsult	Consultant
Calvin Broadus Jr		Latmask	Slacker

Code SSYK by using information about the company, the individual and the provided text!

Developing the model



142 627 manually coded posts

Use all 29 383 posts coded 2019 as test

Only industry (NACE) and text as features

Feature engineering of texts in three steps

1. Preprocess:

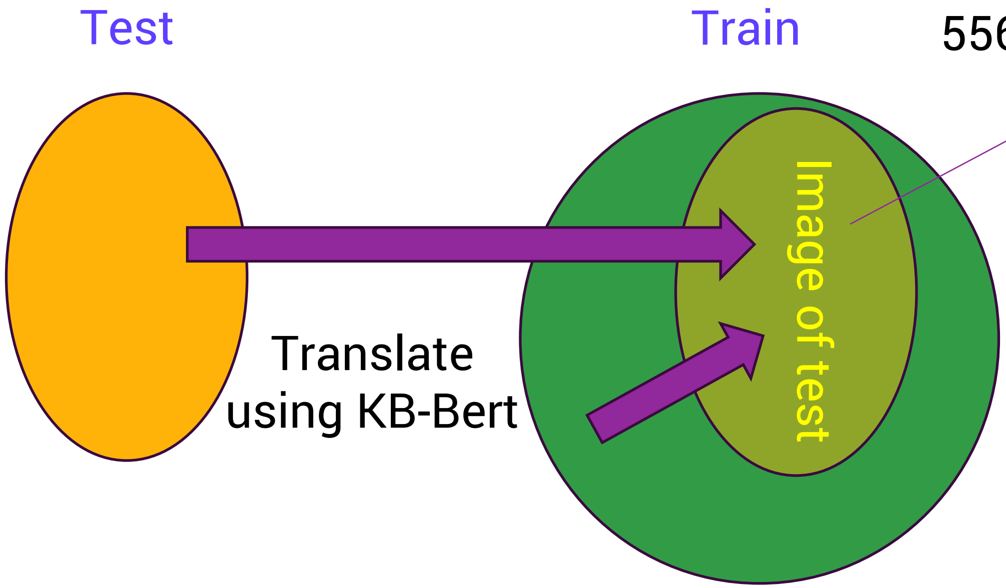
Han är Buss sjafför!!! → busschaufför

lower case letters only
remove special characters
remove stopwords
correct spelling
correct contractions

Reduces the number of unique texts from 38 000 to 28 000

Feature engineering of texts in three steps

2. Translation: 26% of the posts in test has a text which is not in train



5567 unique texts

Examples

original (test)	**translated (train)**	**similarity**
ittekniker servicedesk	itsupport tekniker	0.87
jobbar hönseri	djurskötare fjäderfä	0.86
special forces	extratjänst	0.58
gudomliga interventioner	gudtjänstbesök	0.58

Every post in train and test will have a text from the ”image of test” and a cosine similarity

3. Impact Encoding: Texts in the image of test are replaced with their coding history, from 5567 texts to 1766

Results

The model codes almost 70% of the posts in test with an accuracy of almost 90%.

Low education occupations: 80% of the posts with an accuracy of over 90%
Managers: 30% of the posts with an accuracy of about 70%

Most common errors among the coded posts

Manual coding	Model coding	# posts
Other drivers of motor vehicles and bicycles	Lorry drivers etc	43
Business salespeople	Shop salepeople, specialist trade	40
Cashiers, etc.	Shop salespeople, specialist trade	30
Accountants	Financial assistants	29
Financial assistants	Accountants	29

Automatic re-coding of NACE codes at Statistics Sweden

Jacob Kasche, Statistics Sweden

SCB

Sveriges officiella statistik

What is NACE?

NACE is the european classification of economic activity for statistical purposes

The current Swedish version of NACE consists of 821 codes

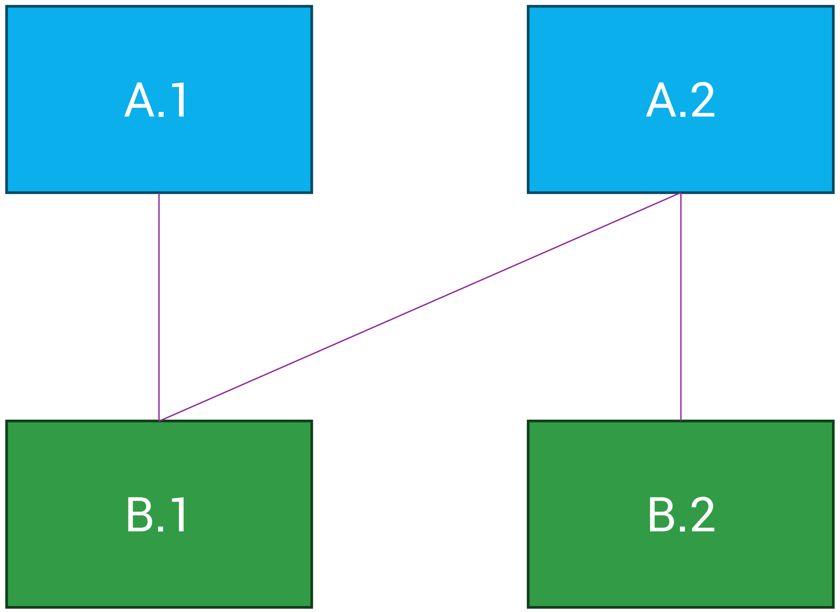
Every unit in the business register should have a NACE code, e.g.,

- enterprises, legal units, local units (workplaces)

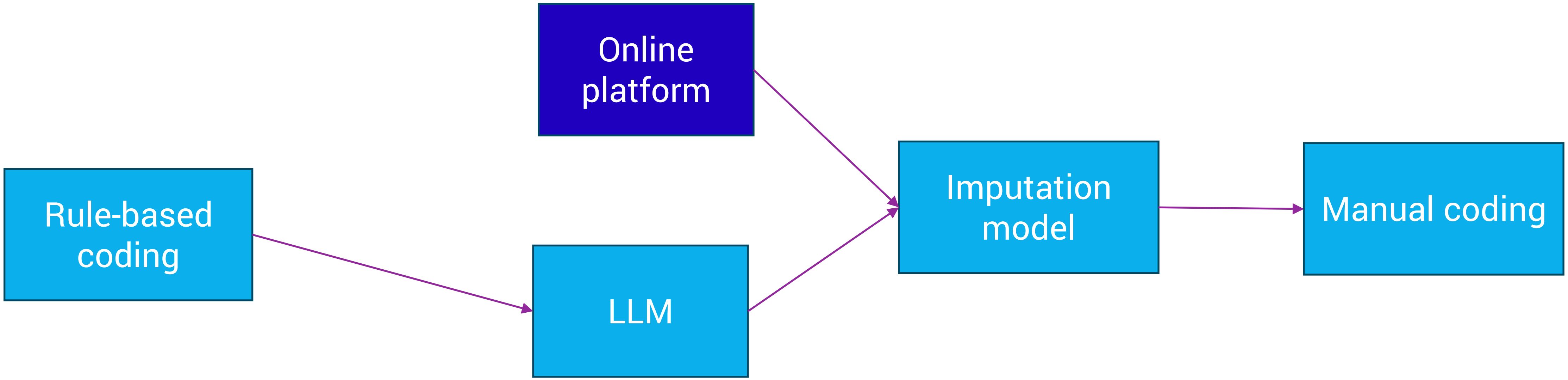
SCB

NACE-revision

- In 2025, Statistics Sweden will implement a new version of NACE codes, NACE Rev. 2.1.
- Eurostat provides a mapping from NACE Rev. 2.0 to NACE Rev 2.1
- Re-coding is problematic for one-to-many cases i.e., A.2 ⇒ (B.1, B.2)



Coding Pipeline



Why not use a single model?

- High quality demands both regarding accuracy and explainability
- Textual data of low quality
- No labels for training

Case Study: Real Estates

Case Study: Classes

- Dwellings
- Industrial premises
- Tenant-owners’ premises
- Other premises

Case Study: Rule-based

- The variable *legal form* includes a class of Tenant-owners’ Associations and therefore we can make the coding rule:
  - *If (legal form* = Tenant-owners’ Associations) ⇒ *NACE* =Tenant-owners’ premises

- Findings rules may be time consuming but using the ones that we can find with ease may induce very good results

Case Study: LLM

Step 1:

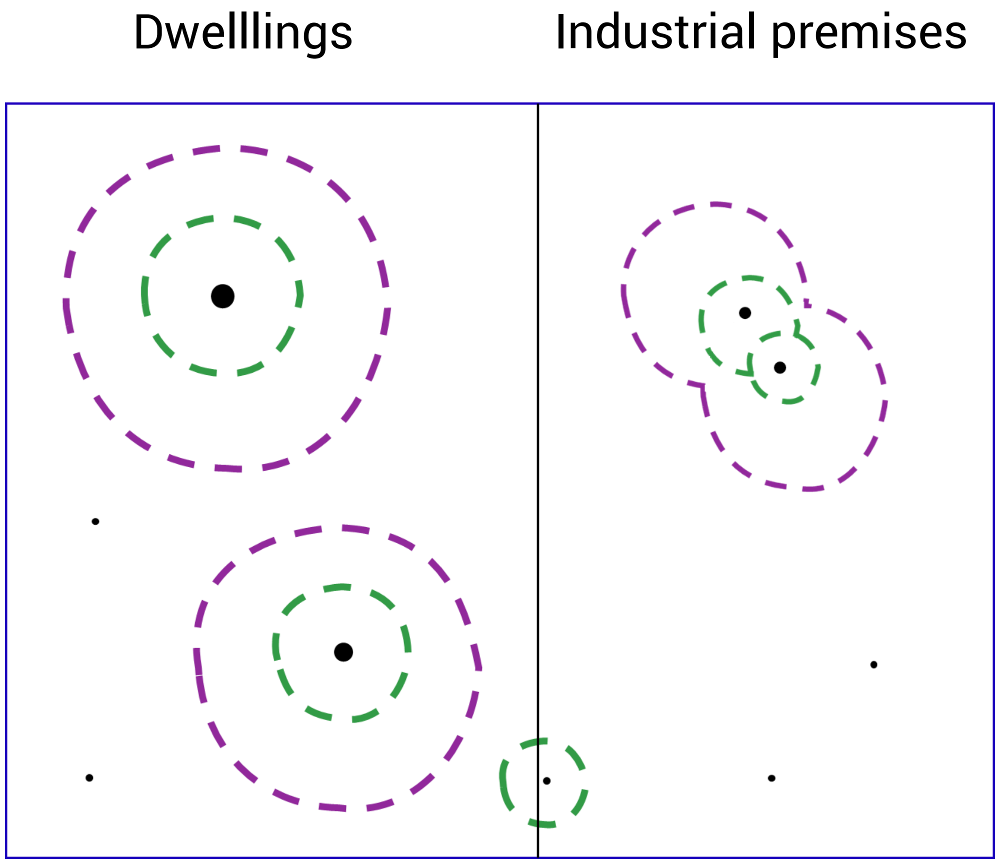
- Using the LLM to find similar words more automatic
    - > Dwellings
        - > Residence
            - > Homes

Step 2:

- Tagging a description of the legal units activity
    - *Leasing of homes for people in Örebro*

Step 3:

- Create decision rule between the tag and the NACE-code

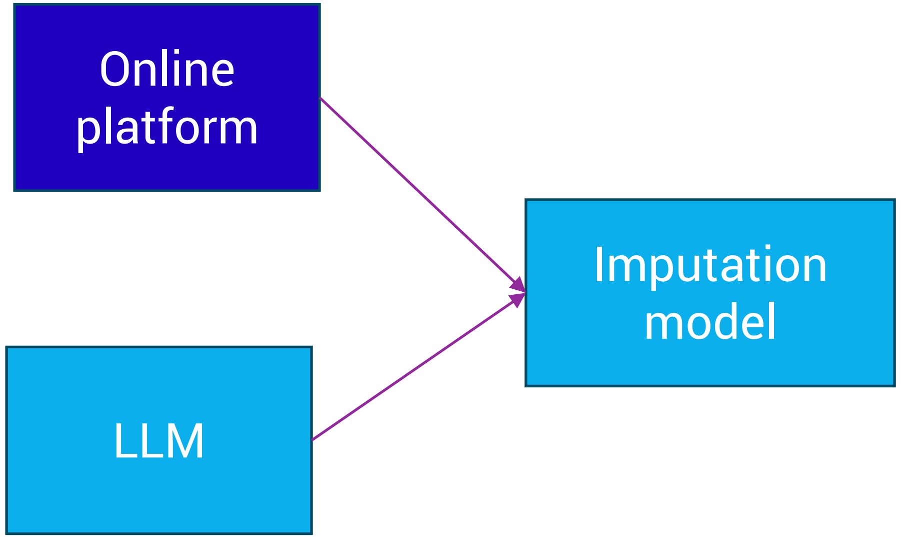


Case Study: Imputation

The remaining units suggests to code with help from other variables, for example:

- Region, revenue, number of employees

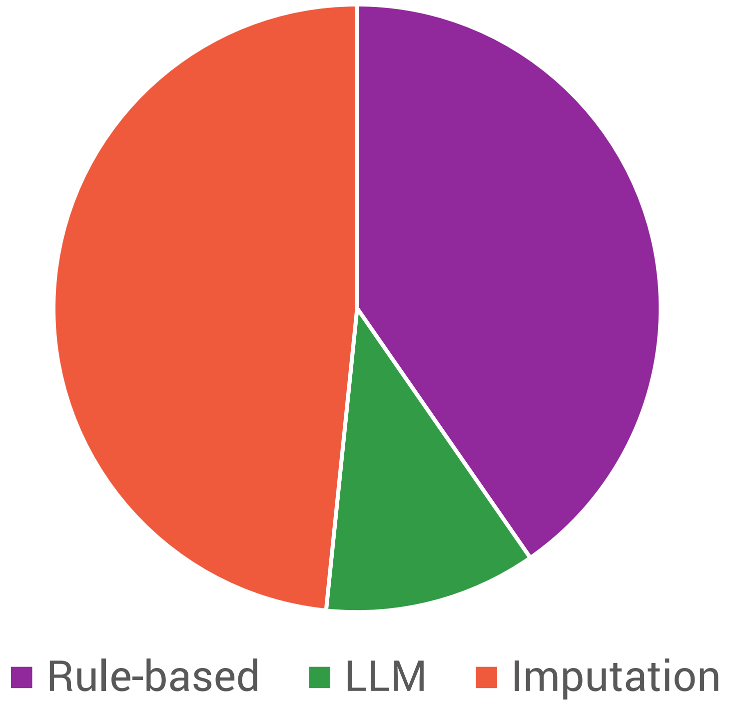
Suggest using a ML-model, for example Random Forest, to find more complex *hidden* rules in data



Case Study: Results

Method	Estimated accuracy	True accuracy
Rule-based	1	0.99
LLM	0.69	0.7
Imputation	0.58	0.44
Total	0.76	0.69

Case Study: Distribution of re-coded units



Case Study: Quality improvements

Rule-based	LLM	Imputation
• Finding more rules	• Selecting more accurate keywords	• Increase the usage of the online platform • Modelling e.g., selection of model and variables

SCB

Questions