Machine learning approaches for predicting NACE activity codes of Hungarian business entities

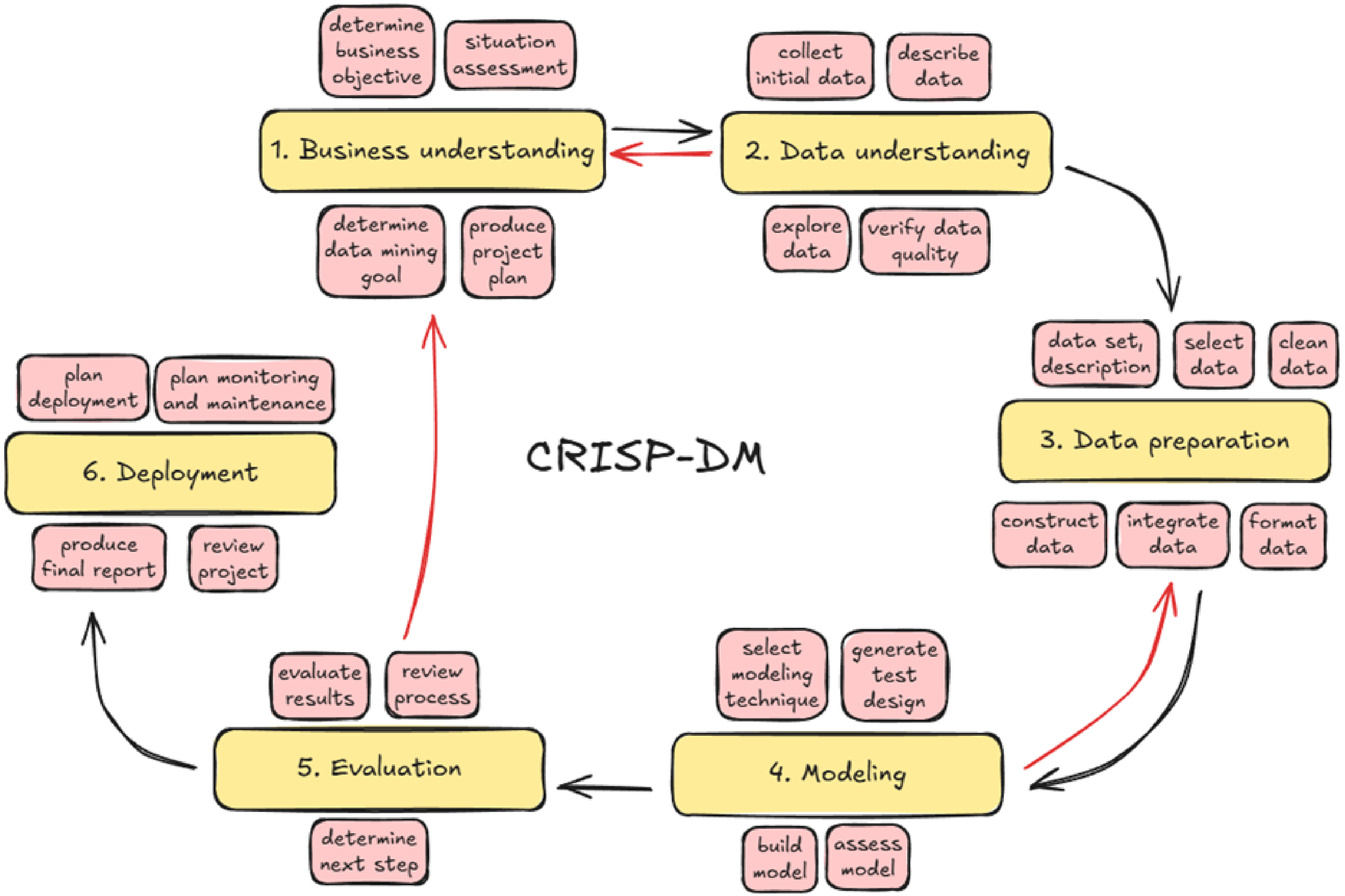
Gergo Bence Mamuzsics

National Tax and Customs Administration of Hungary
Data Science Department
Email: mamuzsics.gergo_bence@nav.gov.hu

December 3, 2025

Presentation overview

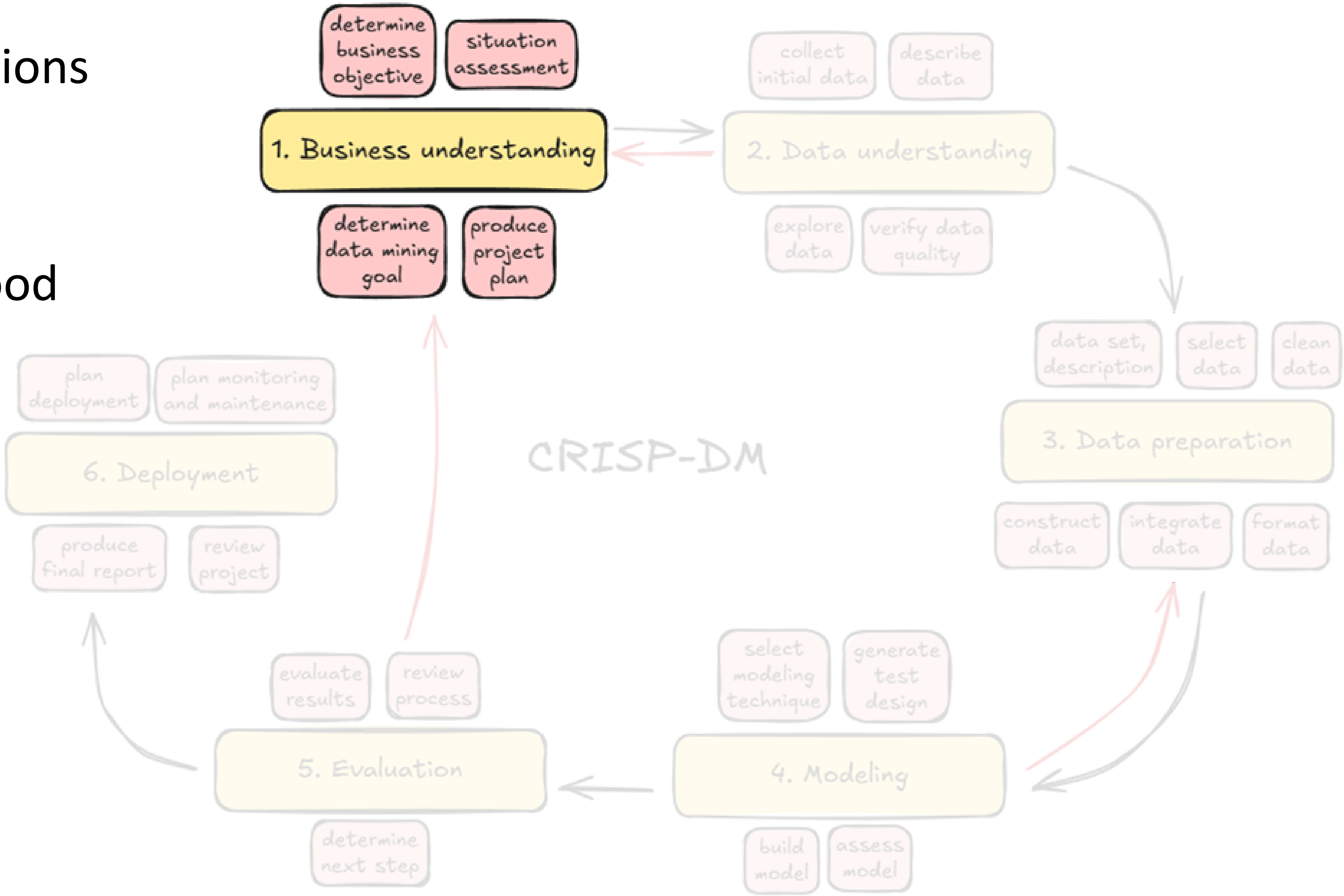
1. Introduction
   - Motivation
   - Data mining goal
2. Data understanding
   - data domains, EDA
3. Data preparation
   - handling missing data
   - transformations
4. Modelling
   - cross-validation, metrics
   - experiments: MNB, NCC, MLP
5. Evaluation
6. Deployment



CRoss Industry Standard Process for Data Mining

1. Introduction

- AI use in EU tax administrations
- Why activity–declaration mismatch matters
- Consequences of lacking good estimation tools
- Motivation for the project



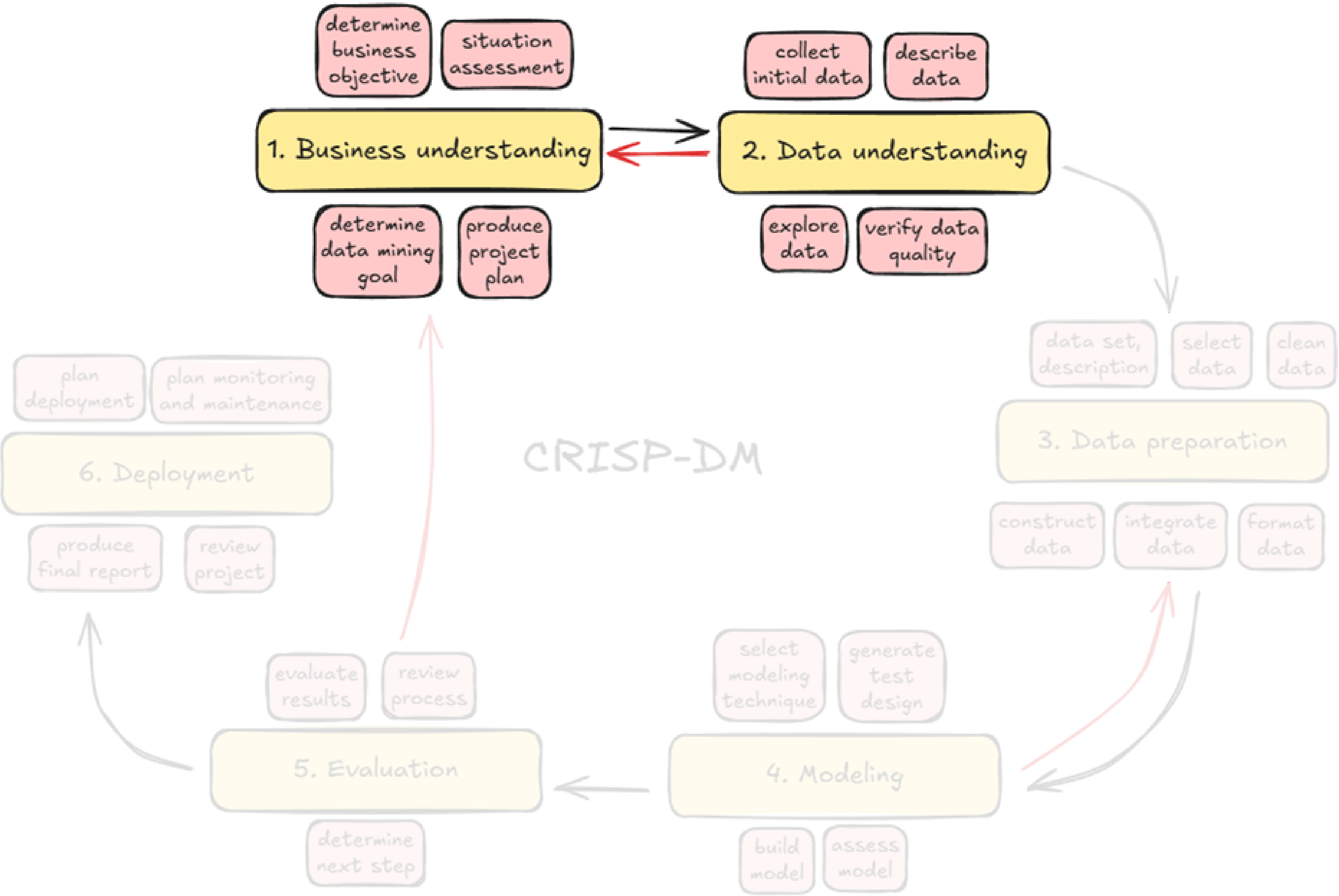
The NACE codes

- Statistical Classification of Economic Activities in the European Community
- updated in 2008 and 2025
- 4 levels of hierarchy:

(1) section	C=== *Manufacturing*
(2) group	10== *Manufacture of food products*
(3) division	101= *Processing and preserving of meat and production of meat products*
(4) class	1012 *Processing and preserving of poultry meat*

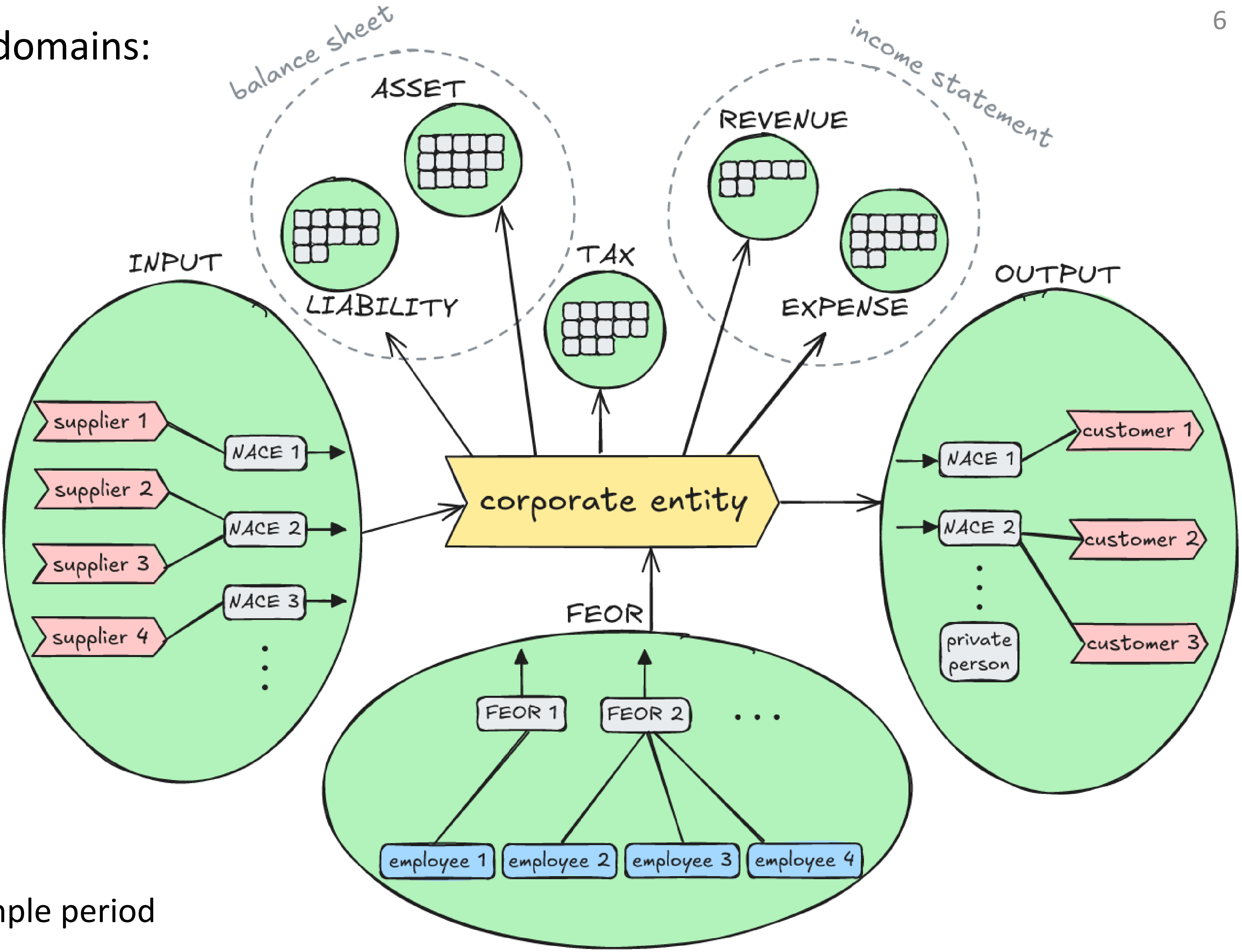
- main activity is mandatory at company registration
- importance:
  - ➢ statistical data collection
  - ➢ tax assesment
- **research goal**: building a sufficiently effective classification model for NACE classes

2. Data understanding



a) Overview of the data domains:

F - FEOR
I - INPUT
O - OUTPUT
A - ASSET
L - LIABILITY
R - REVENUE
E - EXPENSE
T - TAX



b) Target variable:

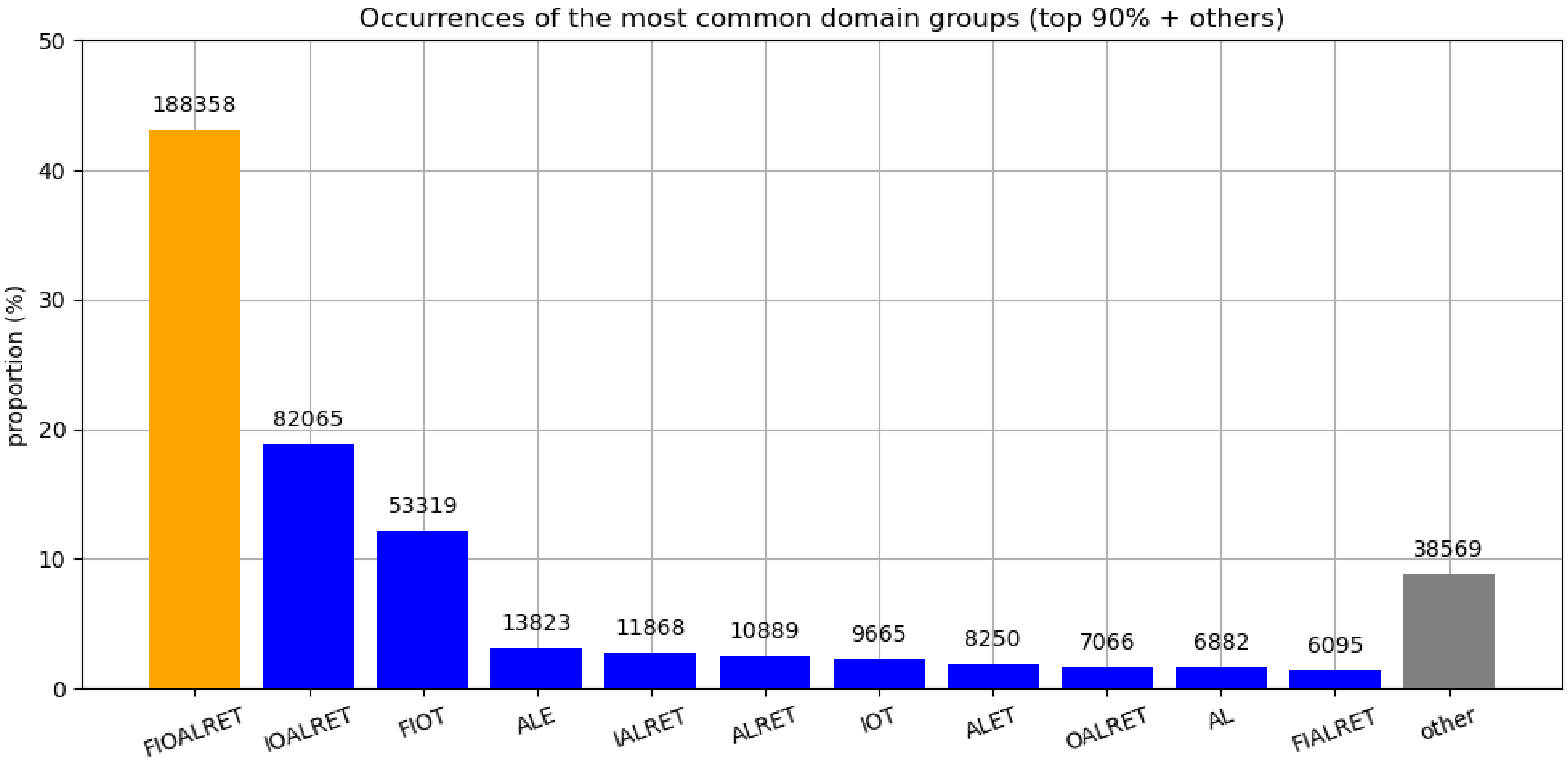
4-digit code of the main activity

- considering noise factors

c) Data exploration - example period

EDA – Data availability of the domains

Domain	Unique Taxpayers
FEOR	261,596
INPUT	371,002
OUTPUT	352,429
ASSET	360,135
LIABILITY	361,441
REVENUE	312,316
EXPENSE	349,918
TAX	405,272
Total (any domain)	**436,849**



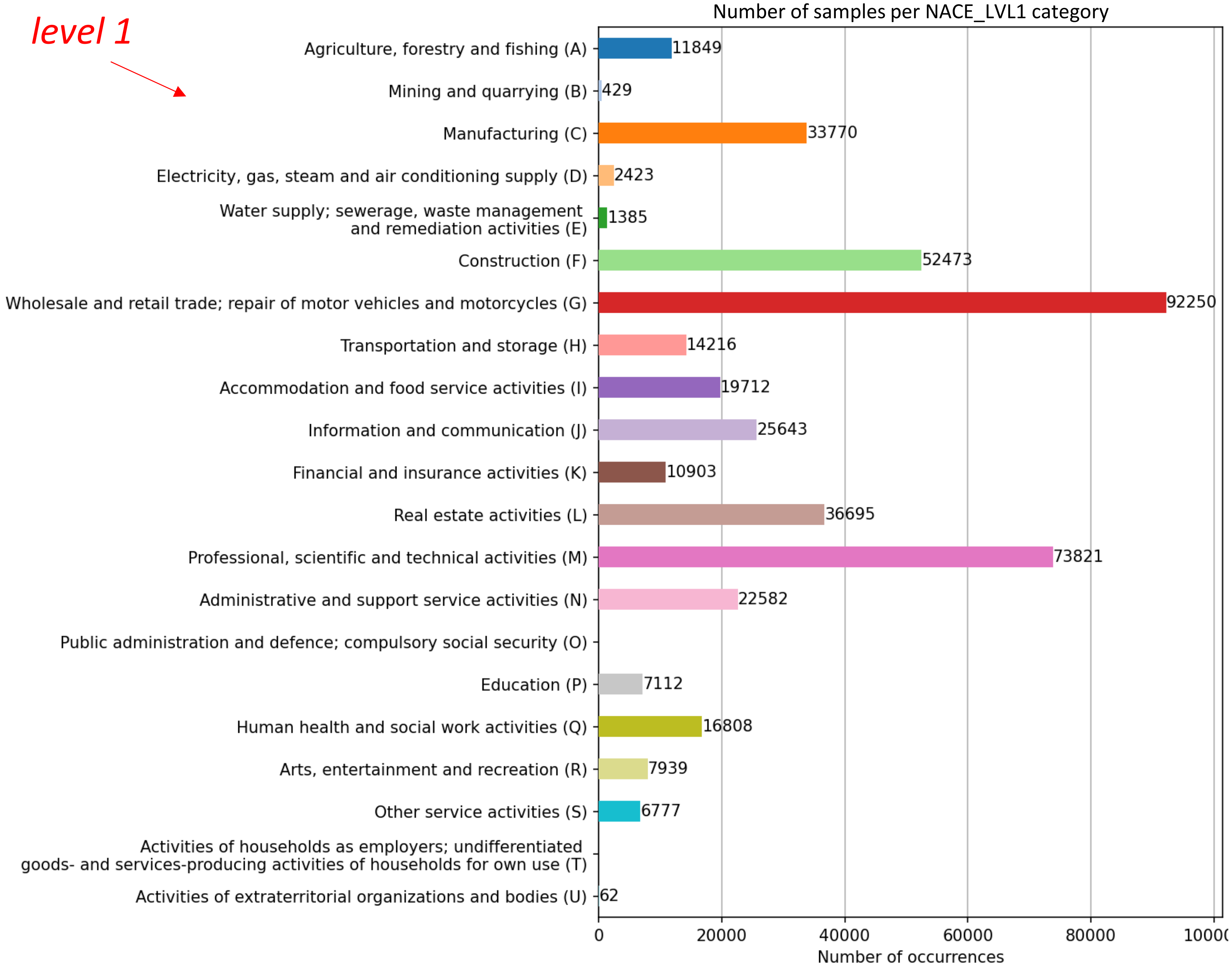
EDA – Domain features

- features are sparse in some domains
- high dimensionality of F,I,O
- orders of magnitude differences in values
- low to moderate correlations between the features

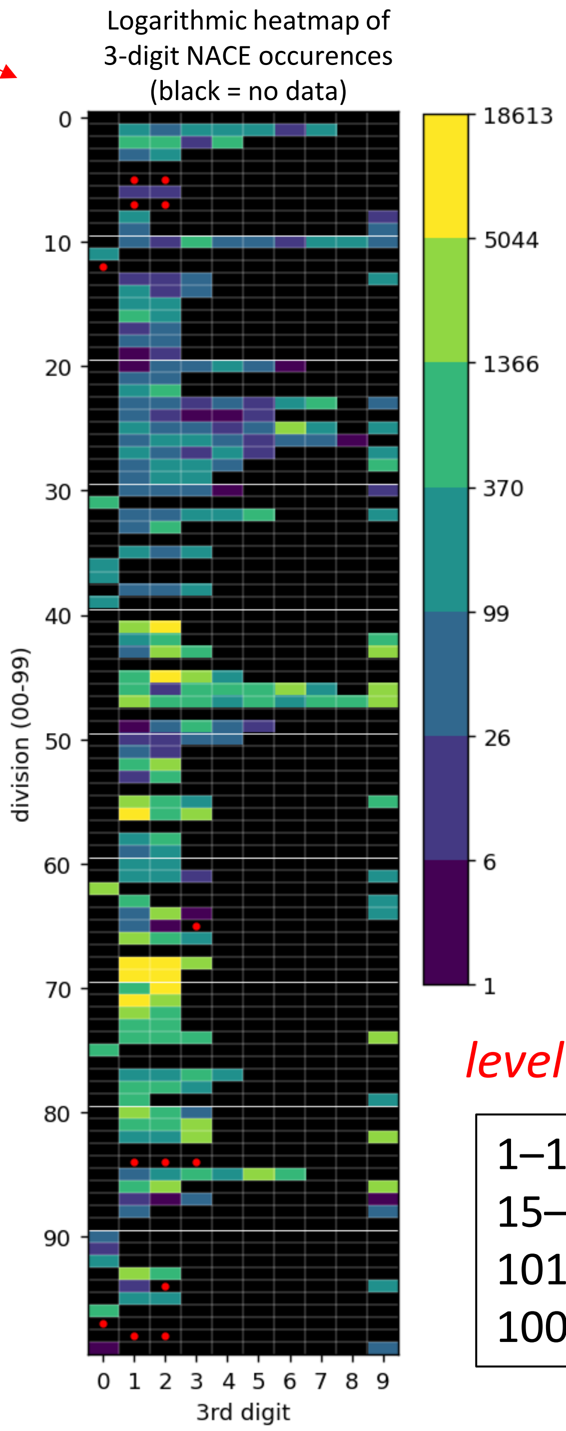
Domain	Non-Zero Proportion
FEOR	0.62%
INPUT	5.40%
OUTPUT	1.94%
ASSET	42.86%
LIABILITY	33.33%
REVENUE	14.29%
EXPENSE	66.67%
TAX	30.77%

EDA – NACE frequencies

level 1



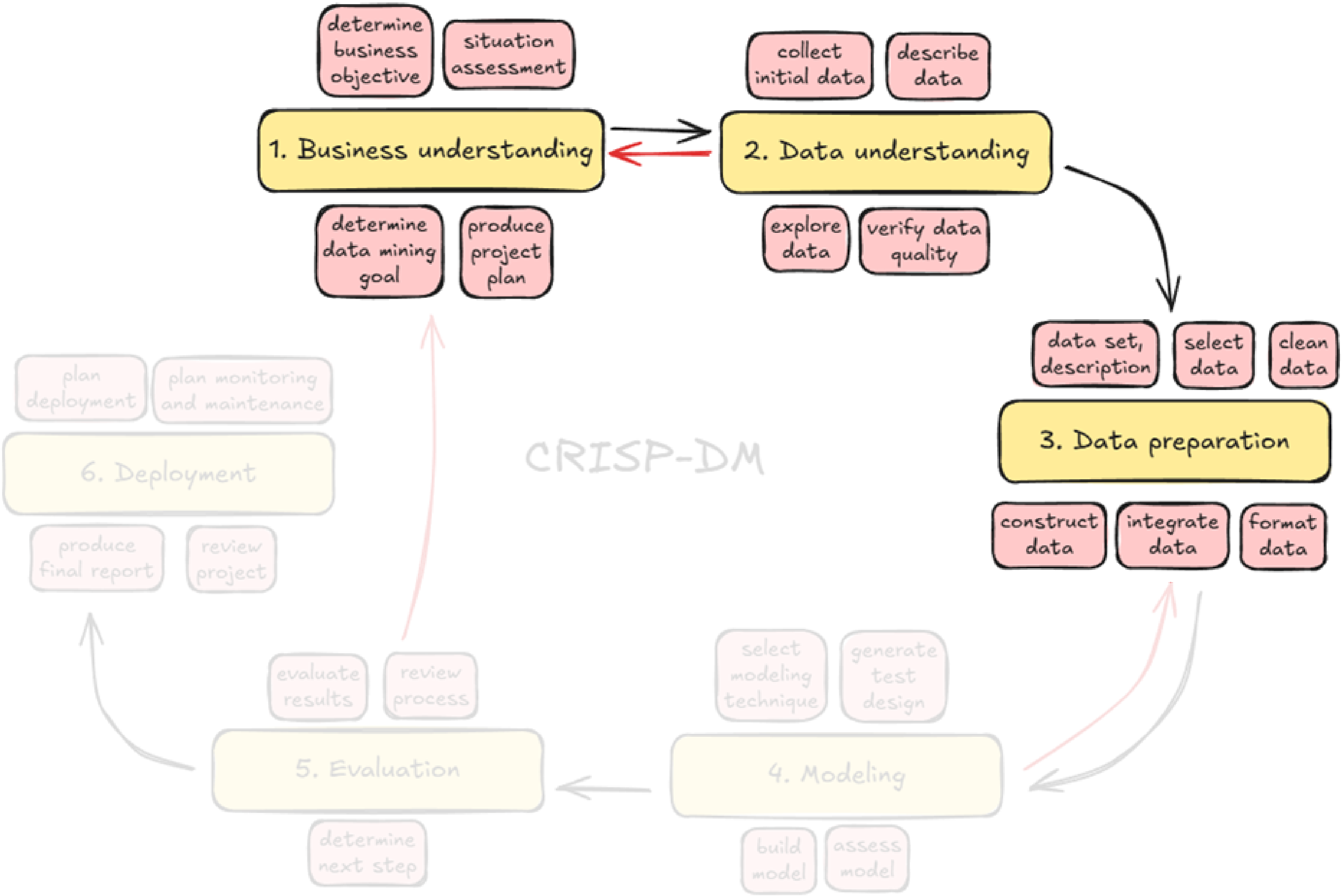
level 3

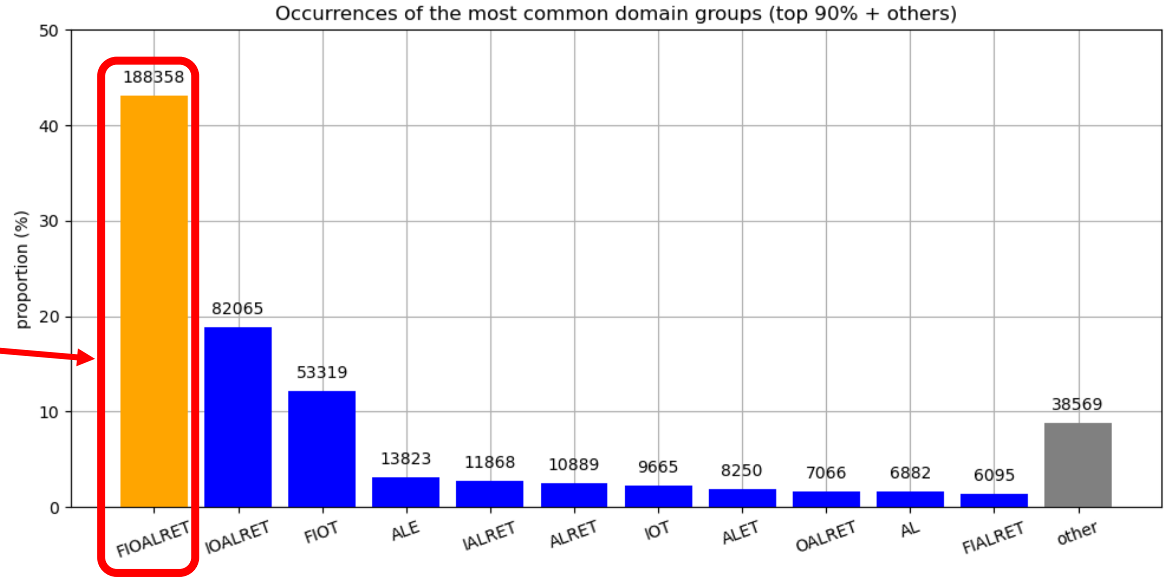


level 4

1–14: 15%
15–100: 28%
101–1000: 41%
1001–18,614: 16%

2. Data preparation





1) Handling missing data

- we want to compare per-domain model efficiencies on the same set of taxpayers
- for this purpose now we only use taxpayers having data in all the 8 domains

```
samples: 436,849 → 188,358
classes:     590 →     572
```

2) Handling rare categories

- only using NACE classes with at least 15 samples

```
samples: 188,358 → 187,610
classes:     572 →     457
```

3) Treating high dimensionality

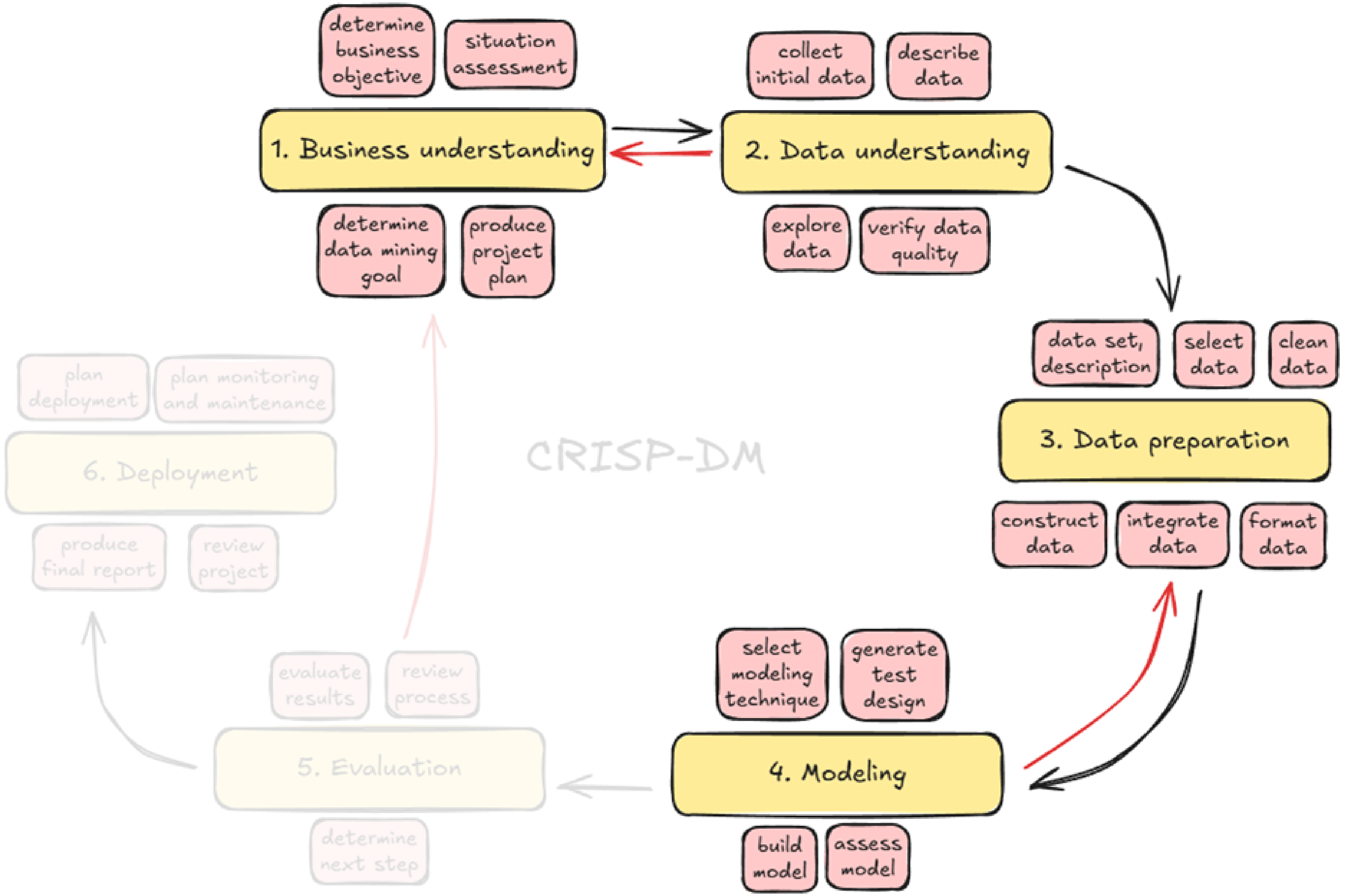
- feature selection/extraction trials:
  - aggregating hierarchical domains (F,I,O)
  - PCA, t-SNE
- modelling choices:
  - cosine distance for NCC (Nearest Centroid Classif.)
  - tolerant models, e.g. MNB (Multinom. Naive Bayes)

4) Handling the order-of-magnitude differences

- log(x+1)
- L1 normalization
  - a) + Centered Log-Ratio transformation (CLR)
  - b) + Feature Standardization (FS)

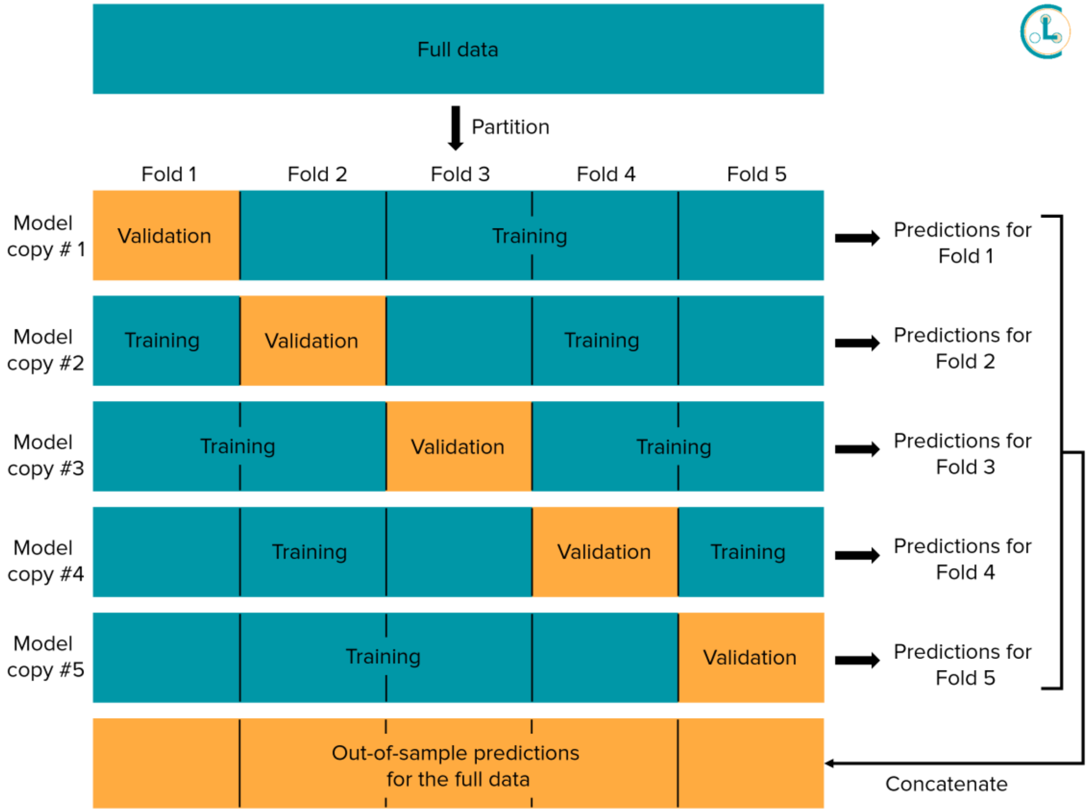
3. Modelling

- Train-validation strategy
- Evaluation metrics
- Classification models
  1. Multinomial Naive Bayes
  2. Nearest Centroid Classifier
  3. Multi-Layer Perceptron
- Results
  - domains separately
  - multi-domain performances



- Train-validation strategy:
  - **stratified 5-fold cross-validation** (fixed folds)
  - no separate test set at this stage
    - more samples remain
    - only limited hyperparam trials accepted
    - *final tests are conducted on a different period*
- Evaluation metrics:
  - „**characteristic rank quartile**” (CRQ)
    - definition: 3rd quartile of intra-class rank medians
    - focuses on balanced class-level prediction quality
    - uses only the ranks, not the specific predicted values
    - easy to interpret (e.g. CRQ = 6.0)
  - balanced categorical cross-entropy loss (CCE)
    - differentiable
    - aligns sufficently well with CRQ
    - outlook: hierarchical variant could be implemented

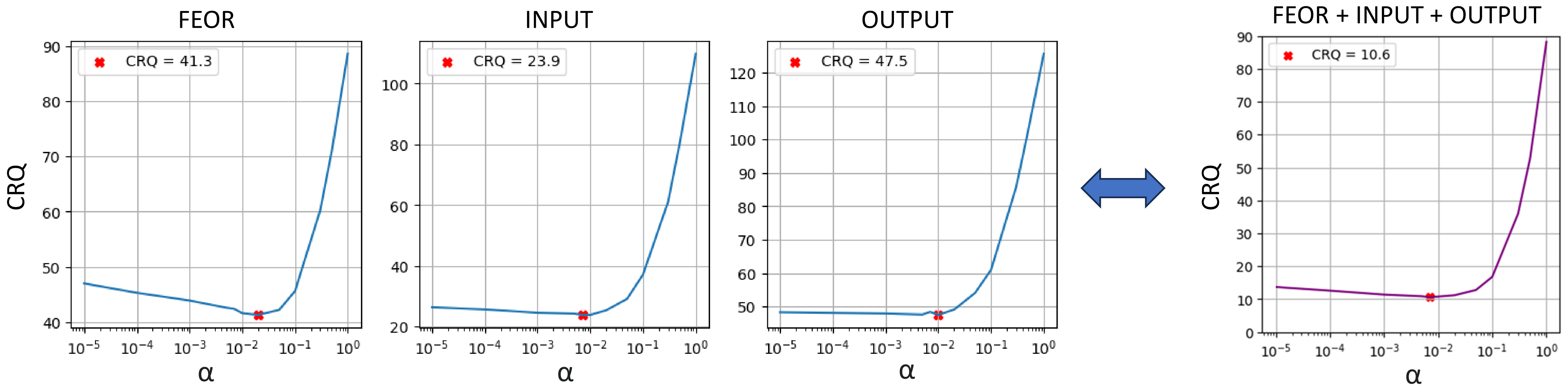
https://docs.cleanlab.ai/v2.7.0/tutorials/pred_probs_cross_val.html



1. Multinomial Naive Bayes (MNB) CRQ: 9.0

- assumes independent features
- simple, fast and scalable
- works well with high-dimensional sparse datasets
- 1 tunable parameter: α > 0
- tuning e.g.:

Combined domains are much more powerful than separate ones!



2. Nearest Centroid Classifier (NCC) CRQ: 11.0

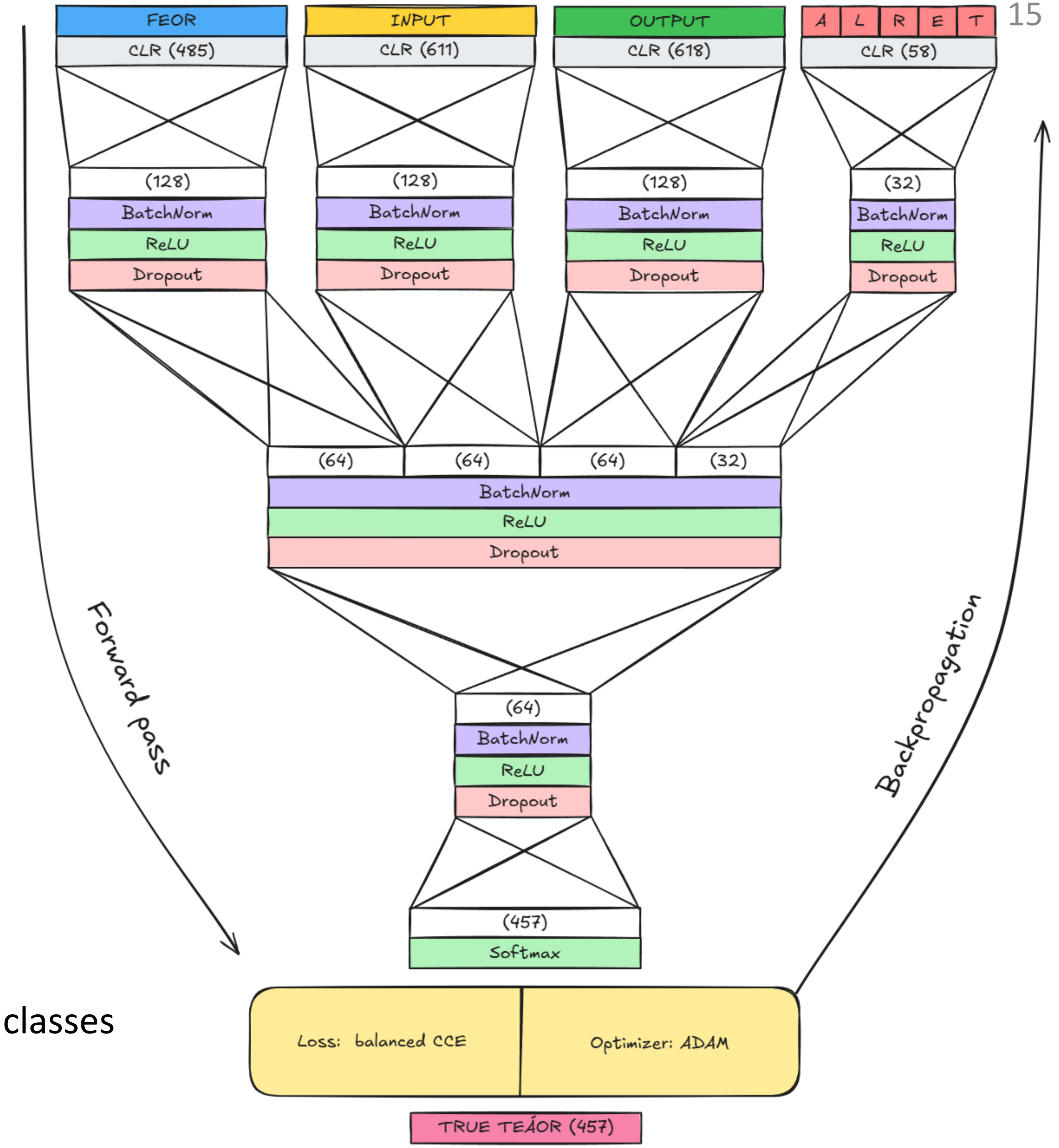
- Assumption:
  - samples of each class are centered around a single point (the class centroid) in the feature space
  - points belonging to a class are closer to that class's centroid than to any other
- easy to interpret, fast, scalable
- Steps:
  1. calculating class centroids (train data)
  2. calculating distances (test data vs. centroids)

gridsearch results

		F	I	O	A	L	R	E	T
preproc	L1								
	L1 + CLR		X	X	X	X	X		
	L1 + FS	X						X	
	L1 + CLR + FS								X
centroid	Euclidean		X	X	X	X		X	
	directional	X					X		X
distance	Euclidean				X	X		X	
	cosine	X	X	X			X		X
CRQ		36	15	27	117	170	188	109	148

3. MultiLayer Perceptron (MLP) CRQ: 8.0

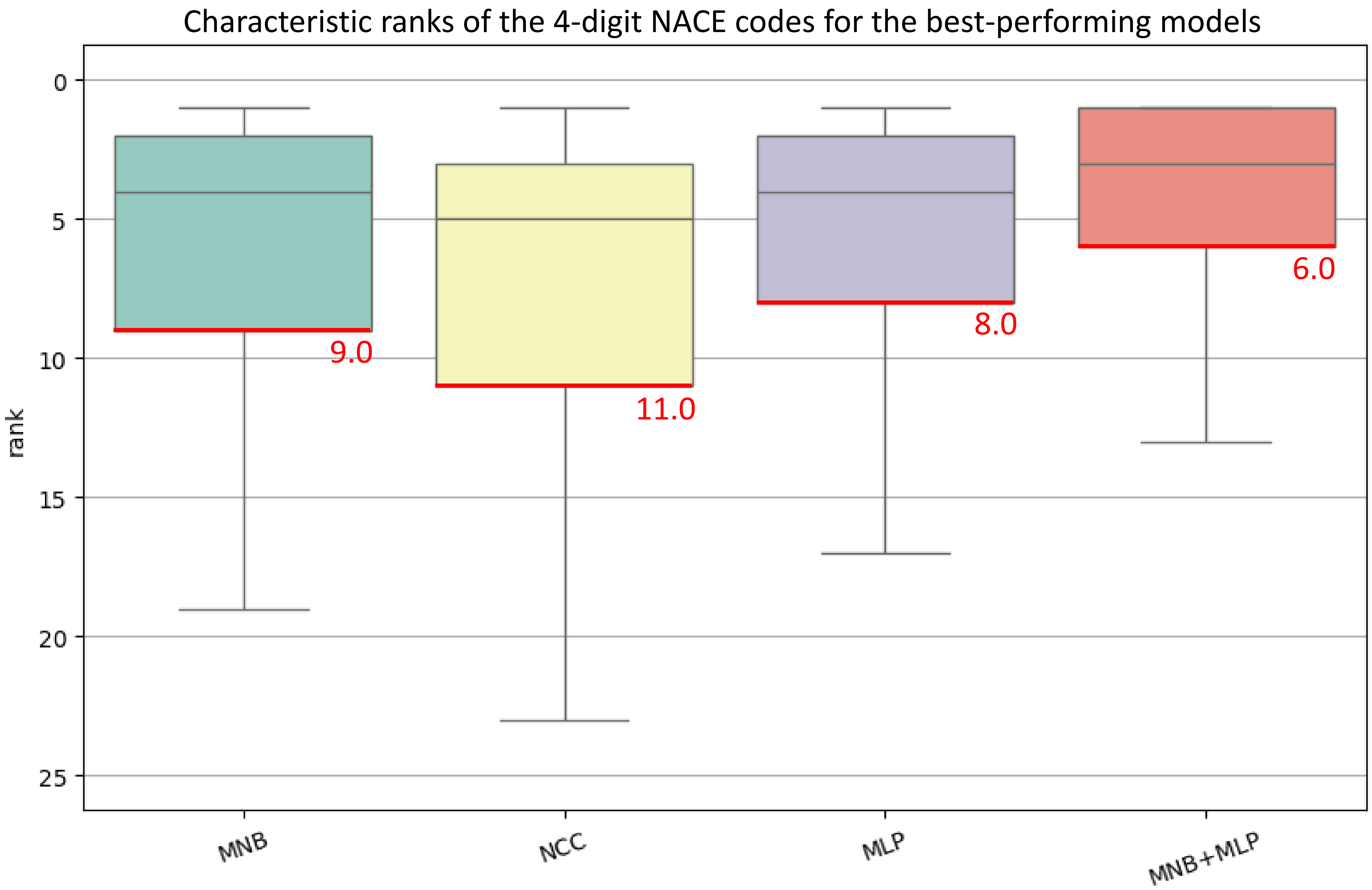
- Training setup
  - Domains trained individually first
  - Combined training leveraged MLP flexibility
- Best-performing architecture
  1. Separate preprocessing for each domain
  2. Branches: F, I, O, ALRET (merged)
  3. Two dense layers per domain branch
  4. Embeddings concatenated
  5. Two additional dense layers applied
  6. Softmax outputs NACE probabilities
- Key components
  - BatchNorm + ADAM → faster convergence
  - ReLU → nonlinear relationships
  - Dropout → reduced overfitting
  - Balanced categorical cross-entropy → equality for rare classes



Per-domain CRQ results:

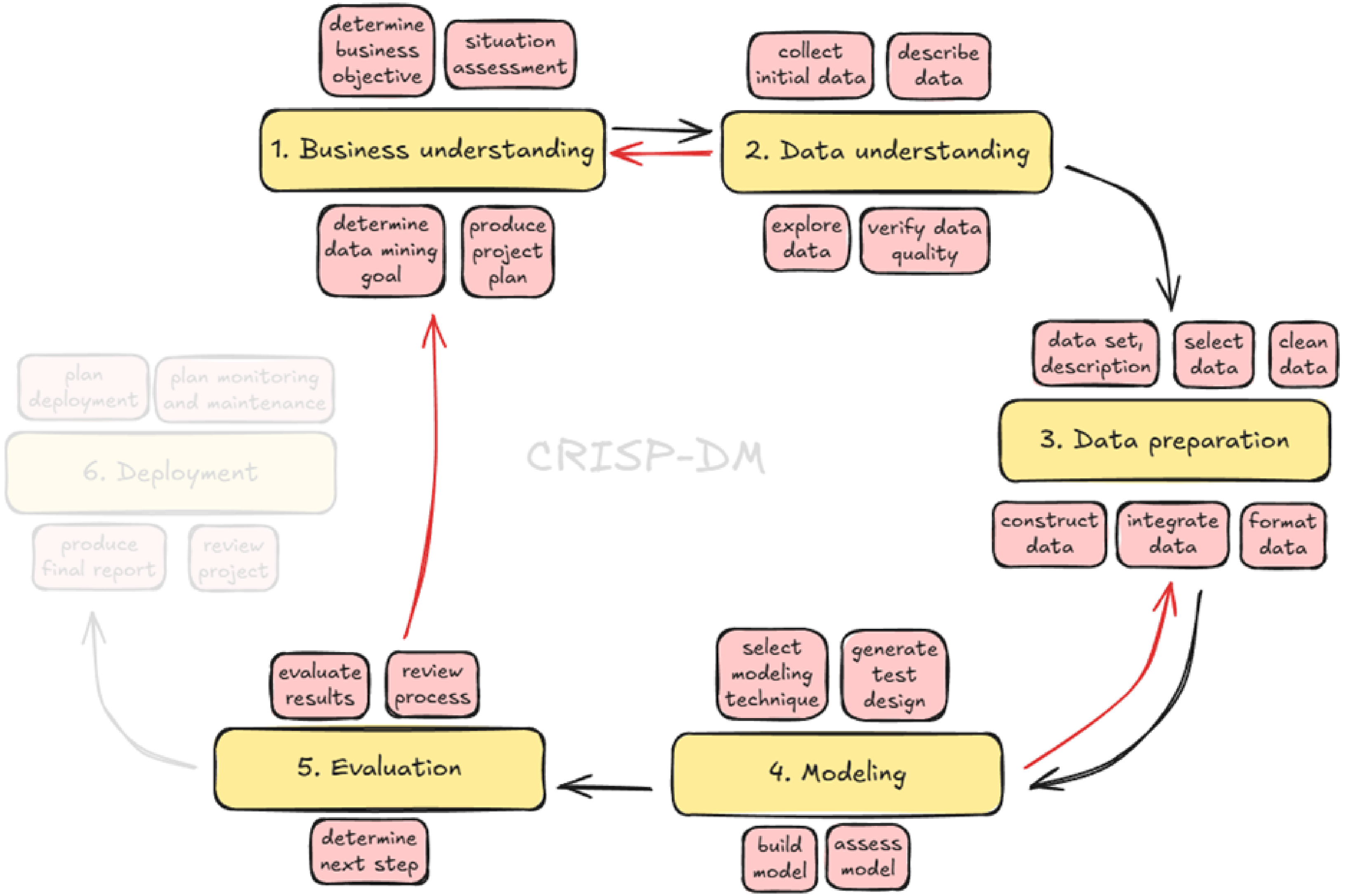
Model	**F**	**I**	**O**	**A**	**L**	**R**	**E**	**T**
MLP	34.0	18.0	25.0	106.0	160.5	189.5	88.0	126.0
NCC	36.0	15.0	27.5	117.0	170.0	188.0	109.0	148.0
MNB	41.3	23.9	47.5	132.0	180.0	298.0	126.0	188.0

Multi-domain model results:



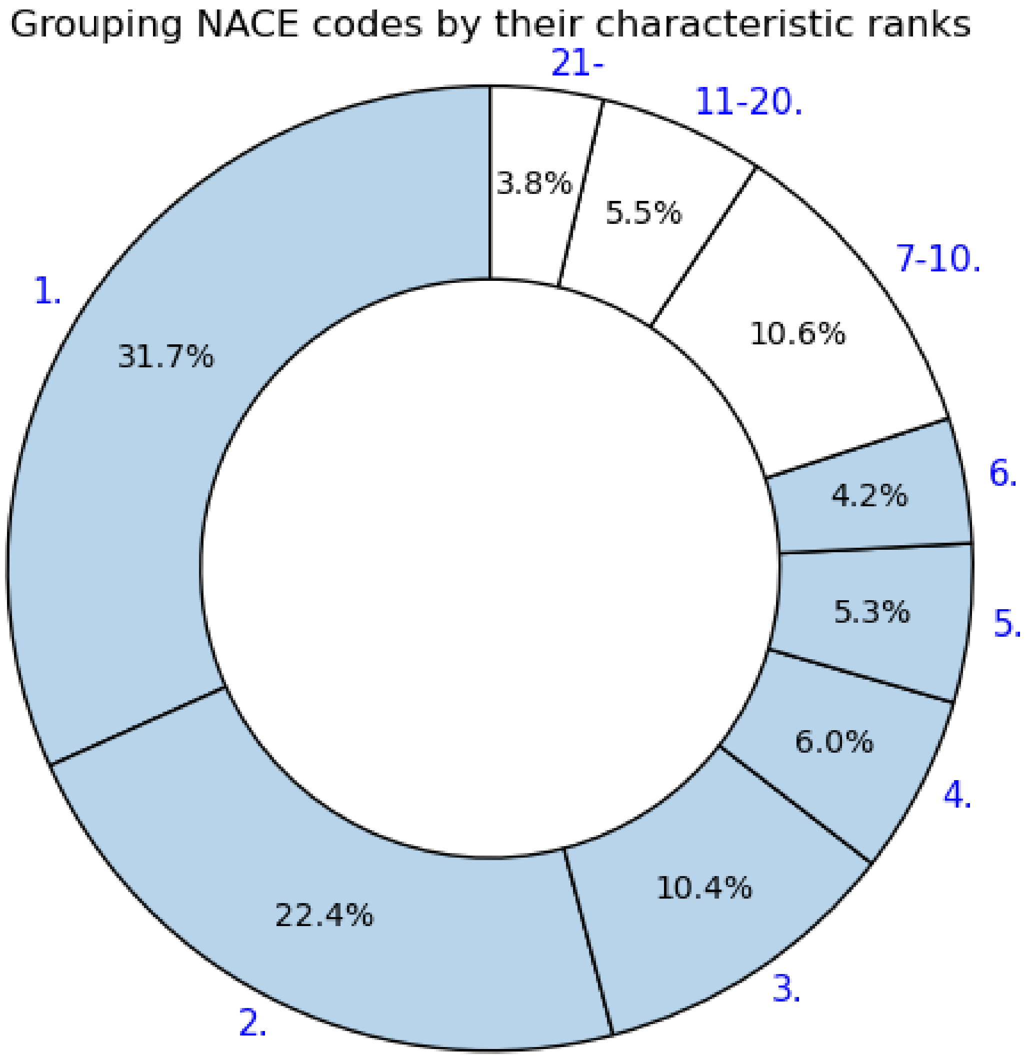
4. Evaluation

- Upgraded MLP
- Evaluation period:
  2024-01-01 – 2025-07-04

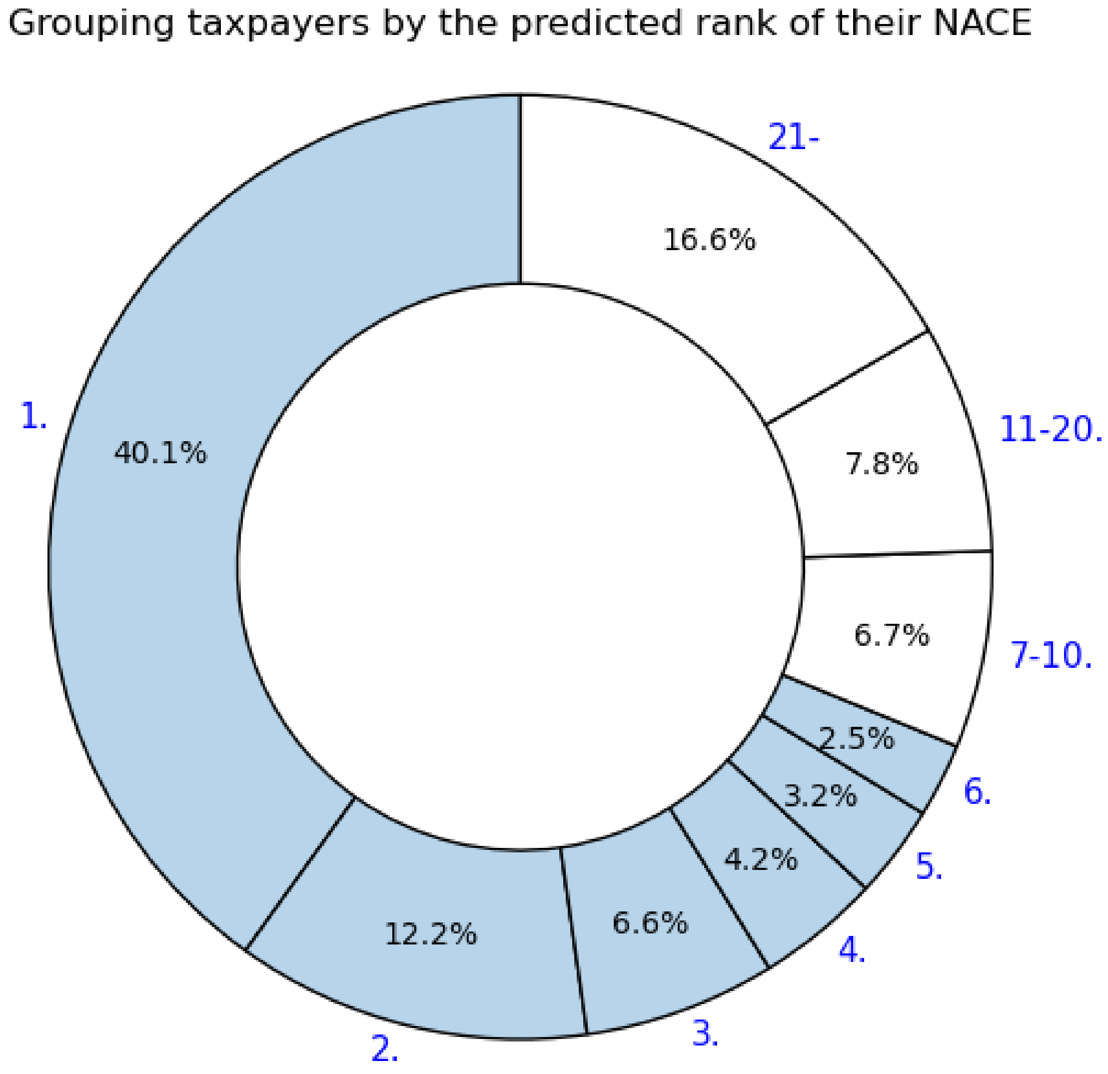


TEÁOR-focused ranks

*This is what we optimized for!

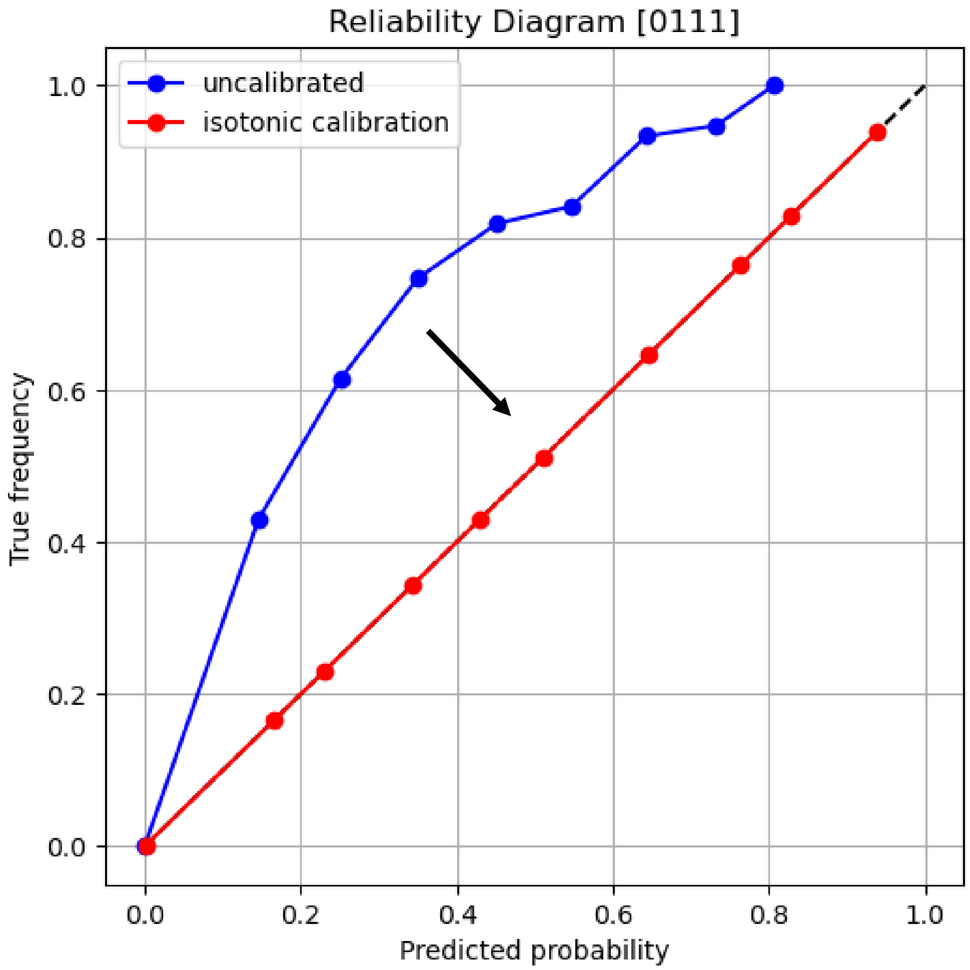


Taxpayer-focused ranks



Possible next steps:

- making probabilities more meaningful



- enrichment of rare NACE classes
- filtering outliers
- smarter ensembles
- predictive modelling for 2- and 3-digit NACE
- use of new domains
  (e.g. text content of invoices)
- modelling for individual entrepreneurs

7. Deployment

- personalized suggestions in data requests
- data reconciliation procedures
- detection of anomalies
- usage in risk-scoring models