Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Bundesamt für Statistik BFS
Office fédéral de la statistique OFS
Ufficio federale di statistica UST
Federal Statistical Office FSO

Automatic Classification with NOGAuto: Exploring Rules for Use in Statistical Production

Athanassia Chalimourda, Mathias Constantin

Data Science, AI and statistical methods
Interoperability and Registers

use of R in official statistics, 27-29 November 2024

Introduction

The classification of the economic activities according to the *Nomenclature générale des activités économiques* (**NOGA**), the Swiss **NACE**, is performed manually by coding experts.

NOGAuto is an assistance system for automatic classification and interaction with the coding expert originally written in **R**.

- Automatic Classification of Economic Activities with NOGAuto
- Global Performance Measures
- Precision by Class in Decision Making
- R and Python Packages
- Conclusions

NOGAuto, Shiny App, Methods (Business Registers Data):
Lorenz Helbling, Mathias Constantin,
Cindia Duc Sfez, Daniele Marx
Methodological Support (Statistical Methods):
Athanassia Chalimourda, Daniel Assoulin

Automatic Classification with NOGAuto

The *Nomenclature générale des activités économiques* (**NOGA**) has hierarchical categories – Example:

*The operation of a drugstore and the marketing of all drugstore, herbal, dietetic and cosmetic products, medicines and health products (NOGA – Code: **477501**)*

- Sector (3 classes): **3** – *Services*
- Section (21): **G** – *Trade; Maintenance and repair in motor vehicles*
- Two digits (Division, 88): **47** – *Retail trade (excluding trade in motor vehicles)*
- Four digits (615): **4775** – *Retail trade in cosmetic and body care products*
- Six Digits (794): **477501** – *Drugstore*

- An activity description is turned into a vector (word2Vec)
- Supervised machine leaning with a gradient boosting machine (GBM) which associates a NOGA-Code to an activity description
- The predicted code is assigned to a description with a prediction probability approximated by the number of GBM-trees that voted for that code

Global Performance Measures

Accuracy, Balanced Accuracy, Cohen's Kappa

Global Performance Measures, like Accuracy, Balanced Accuracy and Cohen´s Kappa evaluate changes in the development of the automatic classification.

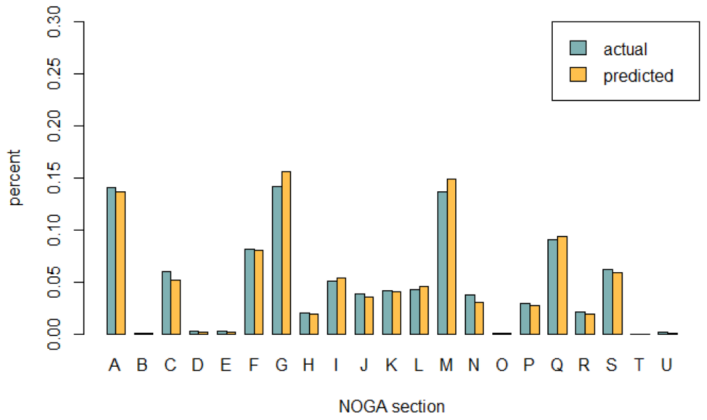
- Accuracy: Overall percentage of elements for which the predicted and the actual class are the same
- Balanced Accuracy: Mean value of the agreement percentages by class (with respect to the actual classes)
- Cohen's Kappa: The accuracy is corrected for the class agreement (actual versus predicted) expected by chance
- Comparison between the distributions of the actual and the predicted classes

Global Performance Measures

Accuracy, Balanced Accuracy and Cohen's Kappa for NOGA Section

We passed from a training set of 25000 elements and a test set of 6400 elements to a training set of 80000 elements and test set of 60000 elements.

Actual and predicted classes, n = 60000



Global Measures	NOGA Section
Accuracy	0.78
Balanced Accuracy	0.63
Kappa	0.75

- Training and test sets are simple random samples from a 200000 data base of activity descriptions originally in french.
- The activity descriptions are of different quality, some are a few phrases long and informative, others consist only of a one or two words.
- This performance represents a baseline, which can be improved.

Performance Measures by Class

Precision

Precision (or positive predictive value) measures the proportion of correctly predicted elements (true positives) out of all elements predicted in a given class (true positives *and* false positives).

High precision means the model rarely predicts a class incorrectly.

Can we find subsets of higher precision within a predicted class using the prediction probabilities of the elements?

A threshold on the prediction probabilities of a predicted class would indicate a subset of higher precision within that class.

The threshold can subsequently be used to decide which elements could be classified automatically and which should be classified by the coding expert.

Performance Measures by Class

Results on *precision by class*

Class	Frequency	Percent	Precision
A. Agriculture, Forestry, Fishing	8390	14.0	97.1
B. Mining, Quarrying	23	0.0	75.0
C. Manufacturing	3560	6.0	66.4
D. Electricity, Gas, Steam, Air Supply	131	0.2	85.5
E. Water Supply	142	0.2	67.7
F. Construction	4838	8.1	80.2
G. Wholesale and Retail Trade	8456	14.2	73.0
H. Transportation and Storage	1210	2.0	83.0
I. Accomodation and Food	3003	5.0	80.6
J. Information and Communication	2298	3.8	71.0
K. Finance and Insurance	2467	4.1	78.8
L. Real estate	2552	4.3	72.2
M. Profesional, Scientific, Technical Activities	8127	13.6	71.5
N. Administrative and Support Services	2230	3.7	74.1
O. Public Administration	51	0.1	35.6
P. Education	1763	3.0	75.1
Q. Human Health, Social Work	5421	9.1	86.1
R. Arts, Entrtainment, Rectreation	1285	2.2	58.6
S. Other Service Activities	3685	6.2	75.6
T. Households as Employers	5	0.0	100.0
U. Extraterritorial Organisations	91	0.2	14.3
Total	59728	100	

Some classes are too sparse for reliable results. The classes for which upsampling is not possible, should be combined in one class and be left to coding experts.

Detailed results on precision will be presented for the following classes:

- *A. Agriculture (97.1%)*: high precision in various settings
- *R. Arts, Entertainment, Recreation (58.6%)*: low precision
- *C. Manufacturing (66.4%)* and *G. Wholesale and Retail Trade (73%)* are often mixed up in automatic classification

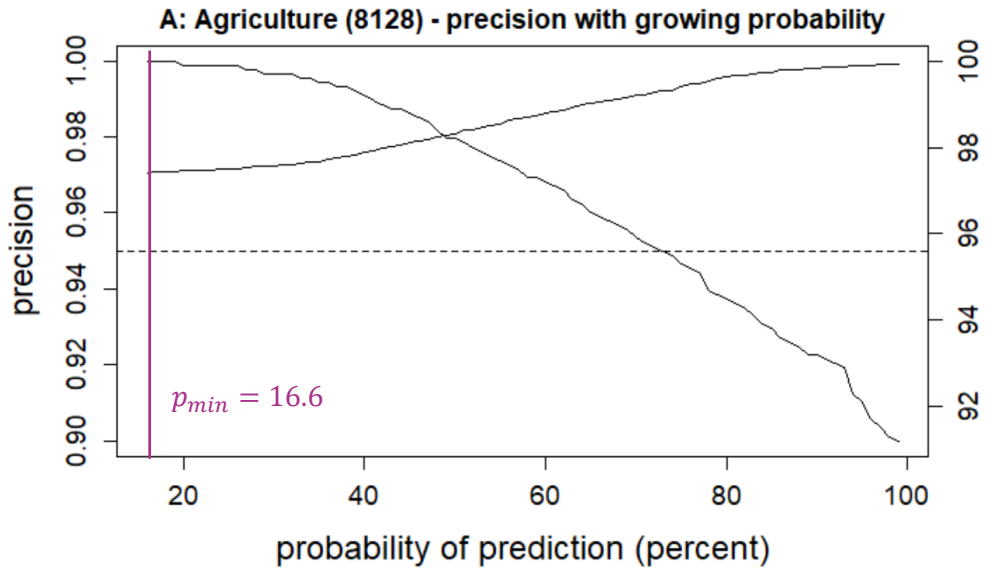
The names of the classes are abbreviated

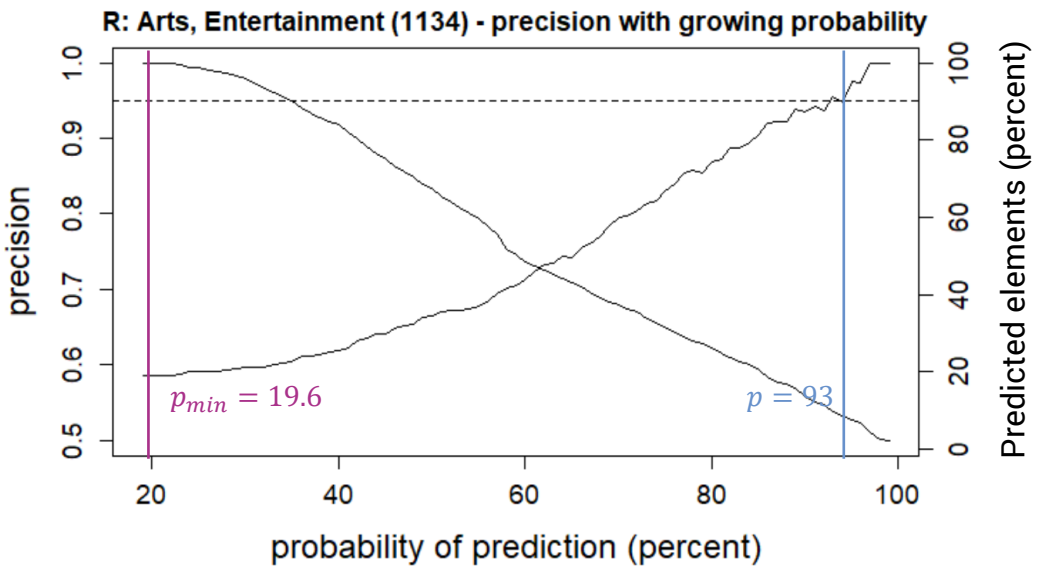
Performance Measures by Class

Precision for A: Agriculture (97.1%), R: Arts, Entertainment, Recreation (58.6%)

We use precision by class and an element's prediction probability to find subsets of higher precision in the predicted classes.

For which prediction probability threshold we achieve precision beyond 0.95?





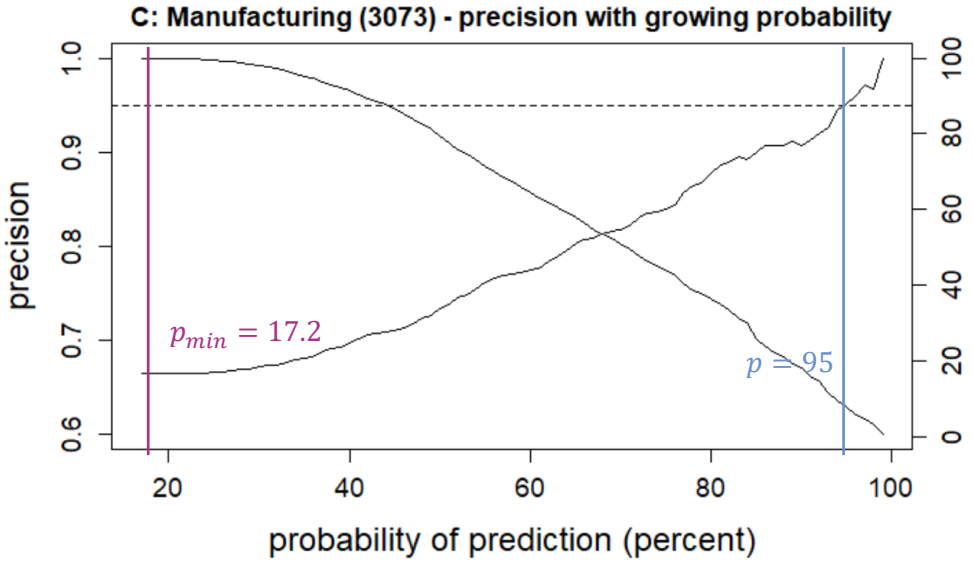
Performance Measures by Class

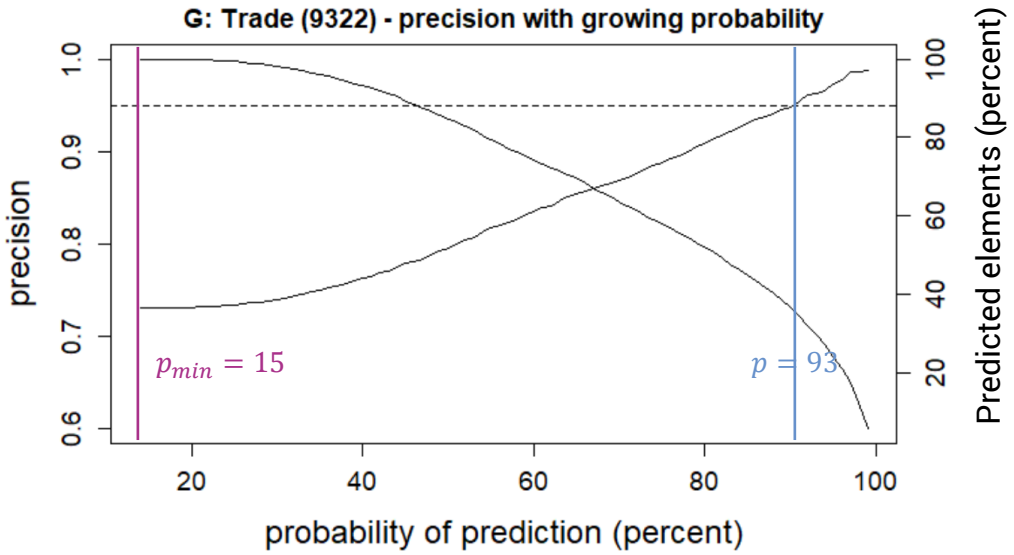
Precision for C: Manufacturing (66.4%) and G: Trade (73%)

For which prediction probability threshold we achieve precision beyond 0.95?

The lower is the overall precision of the predicted class, more false positives have high prediction probabilities.

C: Manufacturing and G: Trade are often mixed up by the automatic classification





Conclusions

- Global performance measures assure overall quality. They evaluate changes in the NOGAuto pipeline.
- Even in classes of overall weak precision, subsets of high precision can be found for higher prediction probabilities.
- Combination of performance measures by class (precision) and prediction probabilities can be used to delimitate where NOGAuto performs best, leaving the remaining elements to the expert.

Next steps include:

- Improve global performance by improving data quality, starting classification a level higher (economy sectors) and balancing the classes
- Improve and adjust the probability thresholds for the expert's intervention

R – Packages

Some of the R - Packages used:

- Dplyr
- Tokenizers
- Tidyverse
- Text2vec
- Xgboost
- Caret
- Shiny
- Flexdashboard

Some of the Python Modules used:

- sklearn
- Nltk (Natural Language Toolkit)
- word2Vec
- Xgboost

Thank you for your attention!