ML pilot study – coding textually described data on economic activity collected from Labour Force Survey

Organisation: Statistical Office of the Republic of Serbia (SORS)

Author(s): Nevena Pavlovic, Sinisa Cimbaljevic, Branko Josipovic, Dusica Zecevic

Date: 27 February 2020

Update: 30 July 2020

Version:

1. Background and why and how this study was initiated

(as many sentences as necessary, as few as possible)

Machine learning team in SORS was formed in 2019. Members of this team are SORS' employees from different departments: Department for development and international cooperation, Department for regional data centers and Department for quality, methodologies and standards. Members of the machine learning team at the Statistical Office of the Republic of Serbia (SORS) have decided to try some of the text classification techniques using several machine-learning algorithms. Text classification is very useful in statistical surveys and is applicable to almost all surveys. As it is a large area demanding to set various parameters, it was decided to use it for the classification of NACE activities. The task was to create a classifier that could classify textually described activities based on what the interviewer entered during the CAPI (CATI or PAPI) interview. The reasons for using ML algorithms are very practical and lie in the need to reduce the time required for manual classification. In SORS classification is done by the trained interviewer or an employee who has expertise in this area. Considering that, a large number of surveys include a section for text classification, it takes a lot of time to classify them. That was the main reason to start working with ML, based on textually described NACE activities. SORS is now developing algorithm in order to decrease time and costs, but at the same time with caution when it comes to accuracy and precision of the classified data. Intention is to use ML algorithm for classification of NACE activities obtained through next Population Census.

ML team has started to clean more than 60000 occupations and 60000 NACE activities conducted during previous LFS. All of these answers are in short narrative text format.

Approximately 20000 rows of textually described NACE activities were extracted and used to develop the ML algorithm.

2. Data

2.1 Input Data **(short description)**

From previously conducted surveys, a data set containing approximately 20,000 records was prepared. Data was divided into two parts, one used for training (80% of dataset rows) and another for testing (20% of dataset rows). This dataset contains three columns: NACE Activity code, Activity name, Interviewer description. More than 250 different types of economic activities were part of this data set. However, the set did not contain the same number of records for each activity code. This is one of the future tasks for ML team in SORS, so that the input set for each activity contains a sufficient number of record for learning. Example of few records of this set is shown in table below.

Activity code	Activity name	Interviewer description
293	Manufacture of parts and accessories for motor vehicles	Production of cables for the automotive industry
511	Passenger air transport	Private air passenger transport
352	Manufacture of gas; distribution of gaseous fuels through mains	Production and servicing of equipment and materials for gasification, pheromone

2.2 Data Preparation

(e. g.: Data Cleaning, Normalisation… – or: none)

Data used as training set was collected through Labour Force Survey (LFS). Interviewers have codded activities during the survey. After this phase, SORS coders have checked data, corrected errors and coded missing values. SORS employees had task to select part of this data so that the set contains data with different codes, after which data was merged. Data set of approximately 20000 rows was prepared as training set.

2.3 Feature Selection

(yes/no, if yes: how, why)

2.4 Output data

(short description)

The data output is stored in excel tables and then inserted into the SQL server for further analysis.

3. Machine Learning Solution

3.1 Models tried

(e. g.: Multi-Layer-Perceptron, Random Forest, SVM, ...)

Three different classifiers were tried: Random Forest, SVM and Logistic regression. All tree classifiers showed approximately similar results. Idea was to apply ML algorithm on 3-digit level of NACE classification (NACE class), and then to reduce results on two digit level.

3.2 Model(s) finally selected and the criterion

(i.e.: which model was why seen being the best?)

Although the results for all classifiers are similar , the best for now are the results from SVM.

3.3 Hardware used

(e.g.: Intel Core i5-6300U, 2.4GHz)

Intel(r) core(tm) i5-6500u cpu @ 3.20ghz, 8.00GB RAM

The plan is to use a virtual machine in the future to speed up the process.

3.4 Runtime to train the model

(e.g.: 2 hours for 500,000 training samples and 25 features)

Given that the volume of data is not large and the average RAM memory was used to run the algorithm, running process took only a few minutes. Around 20000 records and 263 NACE classes (3-digit level).

4. Results

(e. g. in terms of RMSE, MAE, distributional accuracy [*], F1 (micro or macro), recall, accuracy, (threshold,) ..., perhaps as a table for different situations (if available))

[*]: If used: How did you measure distributional accuracy? By proportions, moments, quantiles, correlations, ...?

Three results for three different classifiers.

- Two digit level:

Random Forest ≈ 69% of accuracy
SVM ≈ 75% of accuracy
Logistic regression ≈ 69% of accuracy

- Three digit level:

  Random Forest ≈ 55% of accuracy
  SVM ≈ 63% of accuracy
  Logistic regression ≈ 63% of accuracy

5. Code/programming language

(e.g. the Python code is stored in GitHub)

To develop the classifier, the Python programming language was used with popular free data science libraries such as Sci-Kit Learn and Pandas. The environment in which members of the SORS machine learning team were developing code for the classifier is called Pyzo. It is completely free and open-source, it supports all versions of Python and has integrated shell.

6. Evolution of this study inside the organisation

(e. g.: Collaboration within the organisation? Has this study advanced ML within the organisation?)

Persons involved in the ML group work in different SORS departments and are engaged in different types of jobs, in order to develop different ideas and to contribute to the project through different points of view. Memebrs of team are on different positions in SORS (from junior adviser to deputy directors) and in their daily work they are engaged in various jobs (IT stuff such as developers and system engineers, Business Register specialist, etc.). In the near future, plan is to involve colleagues from regional offices in the ML team.

7. Is it a proof of concept or is it already used in production?

(If it is a proof of concept: Was it successful? How will its results prospectively be used in the future?)

Considering that the precision achieved is not sufficient for production, the idea is for the ML group to continue working on the algorithm until higher accuracy is reached. Accuracy that must be reached for a production is 90 percent of accuracy. Standard has been set by

SORS management. If the algorithm provides the required accuracy, the classification could be completely switched to machines. Otherwise, the algorithm can help in the work or can be used for classification some part of the material in order to make work less costly and more precise.

The idea is to use the ML algorithm in almost all surveys to classify NACE activities, and the biggest challenge and eventual goal is to use this algorithm to classify the text obtained through the 2021 Census.

7.1 What is now doable which was not doable before?

Working on the development of algorithms and machine learning is still ongoing. We learned how important is the use of machine learning in modernization of statistics and how useful results of ML can be. ML is a very important part for the modernization of the statistical system; it presents the techniques that are needed for further development, for reducing costs and getting results in a much faster way.

7.2 Is there already a roadmap/service journey available how to implement this?

(as many sentences as necessary, as few as possible)

As mentioned above, ML team in SORS is working on increasing accuracy. Higher accuracy is the main requirement for using the ML algorithm in production. The precision for wider use of the algorithm in SORS is over 90%%..

7.3 Who are the stakeholders?

A large number of departments /groups in SORS would benefit from using the ML algorithm for classifying NACE activities. The data obtained through ML algorithm would be available much faster and the costs would be reduced. Example that can explain this further: for coding 5000 record, four coders must spend few weeks, and for the same amount of records computer (ML algorithm) need just few minutes.

7.4 Robustness

After reaching the required accuracy, the plan is to test the algorithm on a larger number of data, in order to further verify the accuracy and to test how algorithm works on huge data sets (it is planned to use the input data from the 2011 Census)

7.5 Fall Back

ML algorithm in SORS is developed as a tool that can help in the further work. The risks of using ML algorithm are minimized, but there is always a risk of not using the ML.
There is understanding of SORS Managmet, that ML is very important for modernization of statistics, but as we are producers of official statistics, certain standards must be met, such as accuracy.

8. Conclusions and lessons learned

(e.g.: ML can be used for editing but one has to have the following points in mind ...)

Working on the ml project helped SORS in the modernization process. The acquired knowledge through the project can be applied practically. It also had a positive impact on cooperation among colleagues in the SORS as well as in the cooperation of SORS with other countries that are part of the ML project. The development of the ML algorithm represents a significant contribution to the statistical office, as well as the improvement of employees' capabilities.

9. Potential organisation risk if ML solution not implemented

(as many sentences as necessary, as few as possible)

It is important to raise awareness of new techniques and how new techniques can be helpful in further work. The idea is to present to colleagues in SORS (including management) why ML is significant and what are the benefits of using it.

10. Has there been collaboration with other NSIs, universities, etc?

(yes/no, if yes: which ones?)

Idea is to include University professors, faculties, researchers that are dedicated to science in ML group, as well as to involve other institutions that are producers of statistics. In addition, the idea is as a first step, to include the employees of the SORS regional offices in the ML team.

11. Next Steps

(as many sentences as necessary, as few as possible)

ML team in SORS will try to work with TensorFlow (free and open source software library for data science). The goal is to get higher accuracy in order to use algorithm in production.