Statistical Office of the Republic of Serbia

ML pilot study in Serbia
-Classification of NACE activities-

A short overview of our study

March 2020

Background

- ML team in SORS (Statistical Office of the Republic of Serbia)
- Text classification because it is very useful in statistical surveys and is applicable to almost all surveys
- Task: to create algorithm that could classify textually described activities based on what the interviewer entered during the interview
- Reasons for using ML algorithms: to reduce time required for classification

Data

- We started with cleaning 60000 rows gained through LFS
- Approximately 20000 rows of textually described NACE activities were extracted
- Dataset contains information on NACE activity code, activity name, interviewer description
- Data output is stored in excel tables → SQL

Machine Learning solution and results

- Python programming language (environment Pyzo)
- Three different classifiers were tried: Random Forest, SVM and Logistic regression
- Similar results were gained:
    - *Three digit level ≈ 60% of accuracy*
    - *Two digit level ≈ 70% of accuracy*
- Goal: to get higher accuracy

Benefits

- Contribution to the statistical office in process of modernization and reducing costs
- Gained knowledge can be applied practically
- Cooperation among colleagues in the SORS
- Cooperation of SORS with other countries that are part of the ML project
- Improvement of employees' capabilities

Next steps

- Further work in order to get higher accuracy (main requirement for using the ML algorithm in production)
- To include the employees of the SORS regional offices in the ML team
- To include University professors, faculties, researchers that are dedicated to science in ML team
- Further collaboration with other countries that are part of ML project
- To use ML algorithm for different surveys – challenge CENSUS 2021

Thank you for your attention!

nevena.pavlovic@stat.gov.rs