SAN Final Assignment - Work Plan

December 08, 2023

David Čech
cechdav@fel.cvut.cz

Viktor Korladinov
korlavik@fel.cvut.cz

Diana Korladinova
korladia@fel.cvut.cz

Tomáš Mlynář
mlynatom@fel.cvut.cz

Contents

1. Topic 2
1.1. Introduction 2
1.2. Data Sources 2
2. Work Plan 2
2.1. Main Question 2
2.2. Specific Questions 2
2.3. Datasets 2
2.4. Outline 3
2.5. Risks and Limitations 3

1. Topic

1.1. Introduction

In the past two years, a significant number of Ukrainian refugees have taken shelter in the Czech Republic due to the ongoing war in their country. Since this phenomenon has had a notable impact on the Czech Republic's economy, among other things, the main goal of this project is to assess whether the growing number of incoming evacuees has affected the unemployment rate in Czechia and in what way. The idea was inspired by the observations of a Labor Office employee stating that the ratio of unemployed women has been steadily increasing, and their hypothesis is that one of the main causes is the influx of Ukrainian workers.

1.2. Data Sources

For this project, we are mainly interested in three types of data: the unemployment rate in recent years, the number of Ukrainian refugees, and economic statistics (in case the increasing unemployment rate is only affected by the current economy). With that in mind, we chose the majority of our datasets from the Ministry of the Interior of the Czech Republic[1], the Ministry of Labour and Social Affairs of the Czech Republic[2][3], and the Czech Statistical Office as our data sources.

2. Work Plan

2.1. Main Question

The project's main goal is to gauge whether the growing number of incoming Ukrainian refugees directly affects the growing unemployment rate of women in the Czech Republic or whether it is a result of the country's changing economic situation.

2.2. Specific Questions

- What data will be used, and what is the structure of the dataset?
- What predictors are significant for this particular task?
- What is the influence of Ukranian refugees on women's unemployment rate?
- How about men's unemployment rate?

2.3. Datasets

Given the objective of the task, the dataset will be comprised of the following statistics for the duration of the Russian-Ukrainian war (03.2022-09.2023[4]):

- Temporal data: *month*, *year*
- Geographical: *kraj*
  - We will work with statistics for the entire Czech Republic specified for particular regions (kraje).
- Unemployment monthly statistics (Dataset Link)
  - Predictors: *uchazeciOZamestnaniUoZ*, *noveHlaseniUchazeci*, *noveHlasenaAUvolnenaVPM*, *obsazenaAZrusenaVPM*, *absolventiSkolAMladistvi*
  - Dependent variable: *uchazeciOZamestnaniUoZZeny* (ratio)
- Ukrainian Refugees Statistics (Dataset Link)
  - Predictors: *z_do_65* (ratio with *celkem*), *kraj*, *m_do_65* (ratio with *celkem*)
- Monthly Development of Difference in Inflation w.r.t Last Year: (Dataset Link)
  - Predictors: *monthly_inflation_rate_wrt_last_year* (2nd table)
- HDP

[1] https://www.mvcr.cz/clanek/statistika-v-souvislosti-s-valkou-na-ukrajine-archiv.aspx
[2] https://data.mpsv.cz/web/data/otevrena-data4
[3] https://data.mpsv.cz/web/data/otevrena-data16
[4] This end date is needed because of dataset availability.

  - not available (Eurostat and ČSÚ for NUTS3 available only till 2021), data for Czechia available only till 2022
- Criminality Rate (Dataset Link)
  - monthly counts of various types of crimes per region (for years 2022, 2023), .xlsx, parsing needed
  - Predictors: *break_in_thefts*, *general_thefts*
- Salaries (Dataset Link)
  - average salaries per quarter for every region w.r.t to the number of employed people (2nd column)
  - Predictors: *avg_monthly_salary* (per quarters - imputation)
- Minimum Wage - same for the whole year (Dataset Link)
  - Predictors: *monthly_min_wage*
- Energy
  - Cost Electricity (1MWh) (Dataset Link) and (Dataset Link)
    - Monthly costs of 1 MWh, probably necessary to web scrape the data
  - Cost of Natural Gas [1MWh] (Dataset Link) and (Dataset Link)
    - Monthly costs of 1 MWh, probably necessary to web scrape the data, same website as before
  - Cost of 1l of Gasoline (Dataset Link) and (Dataset Link)
    - Monthly costs, probably necessary to web scrape the data, same website as before
  - Predictors: *electricity_cost*, *nat_gas_cost*, *gasoline_cost*
- Zahraniční obchod (obchodní bilance) - (Dataset Link)
  - Predictors: *import_export_balance*
- REER (Reálný efektivní kurz koruny) (Definition) - probably a good measure of the Czech Crown exchange rate against foreign currencies (Dataset Link)
  - Predictors: *reer* (per quarters - imputation)

The dataset will be processed into a CSV file that will be available on the project's GitHub page.

2.4. Outline

- **Data Preparation**
  - unify format (.csv), data imputation (quarters to months, LAU1 to NUTS3 (okresy to kraje)), visualization, standardization
- **Feature Selection**
  - some of the predictors can be colinear - we would like to weed those out
- **Dimension Reduction & Visualization**
  - compare PCA, ISOMAP, t-SNE reductions
- **Linear Regression**
  - try to predict the women's unemployment **ratio** (the results could be outside of percentage interval - we will apply normalization)
- **Logistic Regression**
  - try to predict the women's unemployment **ratio** (percentage) - stay in the interval <0, 1>
- **Poisson Regression**
  - try to predict the women's unemployment **count** (count data)

2.5. Risks and Limitations

- Some datasets (HDP) are not publicly available
- Significant risk of confounders - the unemployment rate is influenced by numerous predictors (we cannot possibly address all of them)
- Short time interval (for the data) - we cannot observe the error distribution
- Model assumptions may not always be met
- At the beginning of the time period, some outliers can occur (use robust regression)