SAN project:
Review of Team A by Team G

review report

David Reinstein, Vojtěch Tilhon, Matěj Týfa, Gabriela Havranová

1. Problem summary

Team A consists of David Čech, Viktor Korladinov, Diana Korladinova a Tomáš Mlynář.

The main goal of their project is to assess whether the growing number of incoming Ukrainian refugees has affected the woman unemployment rate in the Czech Republic.

2. Approach summary

2.1. Dataset

No existing dataset was available. They combined publicly available data into their own complex dataset consisting of 117 predictors and 2478 data points.

The dataset consists of 3 parts: unemployment rate, number of Ukrainian refugees, and economics statistics

2.1.1. Dataset creation

Data was sourced from the Ministry of the Interior of the Czech Republic, the Ministry of Labour and Social Affairs of the Czech Republic, and the Czech Statistical Office. The data provided was not in a uniform format and therefore required pre-processing.

2.1.2. Dataset description

Dataset contains data from January 2009 to September 2023 with the exception of refugee-related columns that start in March 2022. Data that were not provided on a monthly basis were considered constant for the provided time period.

2.2. Preliminary analysis

The authors carried out a search for recurring patterns, seasonal spikes, and correlation between factors using time and seasonal plots. Multiple graphs were plotted in their analysis and are available at their GitHub repository[1]. Correlation analysis was performed using the Pearson, Spearman rank, and Kendall rank correlation methods. Autoregressive tendencies were explored using lag plots.

Patterns, spikes, and correlations were discovered, noted, and discussed in the paper.

2.3. Model

Two different models were created. First model integrated data on the refugee situation, while the other model does not take refugees into consideration. Those models were

[1] https://github.com/DianaKorladinova/SAN_Unemployment_Refugees/tree/main/graphs

compared to find out if the refugee data increased their effectiveness. If both of the models behave comparably well, then refugees have no discernible influence.

Due to the short timeframe of the invasion the models were trained twice:

A. expand data for the refugee model with 0 values outside of the invasion time
B. restrict the time interval to the duration of the invasion

Ridge, LASSO, and Huber regressions were implemented and applied onto standardized data (subtracting mean and dividing by standard deviation).

2.4. Evaluation

The authors used temporal cross-validation (training fold encompasses values that precede those in the testing fold) while predicting all 14 regions of Czech Republic at once.

3. Results summary

All three regression types performed comparably. Adding refugee-related information considerably improved all predictors. All metrics indicated that the LASSO (with economic and refugee predictors spanning over the last two years) model named B1_LASSO performed the best.

The authors included statistical hypotheses testing. Their null hypothesis was: “For both models, the metrics’ values on each fold of the cross-validation have the same distribution.", an alternative was: “For both models, the metrics’ values on each fold of the cross-validation do not have the same distribution.”

Paired t-test and Wilcoxon signed-rank test were performed on RMSE values from cross-validation of models B1 and B2. Both tests yielded p-values below the chosen significance level of 0.05, namely 0.012 and 0.009 respectively. The authors concluded that the null hypothesis can be rejected and that one of the models is significantly better than the other one.

This led to the conclusion that the claim proposing no significant improvement in predictions with the addition of refugee-related predictors should be rejected.

4. The judgment

The dataset was constructed using reasonable data from reputable sources. The process of data gathering, preprocessing, and analysis was thought out and properly executed. We admire the creation of custom dataset, which must’ve been really time consuming.

In the process of model creation, multiple risks were identified and their impact evaluated. Multiple approaches for modeling were considered and even implemented. Their hypothesis was properly tested and a conclusion was reached.

We were not able to find any significant statistical or technical errors in their work.

The written report is properly formatted, includes all of the important information and is very well written. It even includes proper references and citations. The work even conforms to the licenses under which the used data sets were distributed.

Great job :)

5. Critical remarks

The report was written so well we could only find few mistakes:

- In the conclusion, we would like to see more emphasis on answering the problem that you formulated in section 2. While the question was indeed answered, it should be made clearer in the text.
- The uchazeciOZamestnaniUoZ dataset column is described to represent the “the total number of female job applicants by the end of the month” while it contains the total number of all job applicants.
- Nitpicks: some typos or typographical errors, inconsistent mix of czech and english attribute names (“year” vs. “kraj”)

6. Questions

1. Did you consider using other regression models, e.g. SARIMA, since you worked with seasonal data?
2. Did you come across any obstacles, if so how did you overcome them?
3. Do you feel that you achieved what you set out to do?
4. If you were given more time (and perhaps even a budget) how would you extend this study?