Implementing Nowcasting Techniques for Timelier Publications

Gergely Attila Kiss[1], Beáta Horváth[2]

[1]Hungarian Central Statistical Office, Hungary

[2] Hungarian National Bank, Hungary

Abstract

The aim of the project is to develop and implement new method to provide earlier estimates of the STS publications, namely Producers Prices in Industry (PPI), Producers Volume in Industry (PVI) and Services Producers Price (SPPI) indexes. These predictions could be disseminated as experimental statistics to be timelier in communicating important changes about the economy to the public and policymakers. The method is focused on applying Machine Learning techniques for estimating the named indicators. In the current state we have a base algorithm to predict the indicators. This uses mainly tree based regressors, but we are planning to extend the algorithm to use different types of recursive neural networks too, as they are the state-of-the-art technique. We are also testing different types of data sources to reach a consistent state of predictions. Our preliminary results for PVI shows that the algorithm works fairly well in presence of appropriate train data.

Keywords: Machine Learning, Nowcasting, short-term statistics (STS)

1. Introduction

Our goal with this project is to be able to provide timelier information to policymakers and other stakeholders about three key indicators of the economy Producers Prices in Industry index (PPI), Producers Volume in Industry index (PVI) and Services Producers Price index (SPPI). The current Short-Term Statistics (STS) publication dates at the Hungarian Central Statistical Office (HCSO) come with high latency (ranging between t+30 and t+90). Whereas, this project is to produce new estimates with using the available and new data sources to produce indices from in month (t+0 days) up to t+20 days for the above three. This could make stakeholders increase their agility, as the estimated indicators, would come significantly sooner.

For this we are developing a nowcasting algorithm that we could test on the European statistical Awards' Nowcasting competitions. The tool was used to nowcast monthly PPI index without any specifications for the needs of the countries. It produced good and consistent results for several of them. Unfortunately for Hungary it could not provide acceptable results and will need to be further tweaked for the country. The algorithm in development is easily applicable for other time series, as it could already produce PVI estimates in the last months of the competition and further modifications to reach SPPI

estimations can be easily made. Our idea was to use standard supervised machine learning models (random forest, ridge, lasso, etc.).

We plan to further enhance the algorithm's efficiency with domain expert knowledge to tailor for the Hungarian needs. Our expectation is that it will increase the efficiency of the algorithm while decrease its timeliness. The change in both is expected due to the reason that in the competition we had to provide in month (t+0) estimates and that created a strong constraint on the available data. Analog to the balance-variance trade-off, we plan to find the balance between timeliness and accuracy that provides the optimal solution.

2. Current state and results

One of the most important part of any data project is to acquire the proper data for the purpose of the project. In our case, we have to find sources that have an economically sound relationship to the target indicators. There are some usual suspects in the recent literature for nowcasting price indicators, especially CPI and inflation. These mainly mention energy related prices, most of the time oil and gasoline (Knotek and Saeed, 2023 and 2014), high frequency scanner data (Beck et al.,2024) and online observable prices (Aparicio and Bertolotto, 2020 and Macias et al., 2023).

2.1 Current data sources

Our current sources cover the World Data Bank monrhly prices form World Bank Prospects Group's pink sheet that covers monthly prices and price indexes for several aggregated topics including energy and raw materials. This source we plan to use for all indicator as it can help in providing a general economic state to the models. We extend this for PVI with two country specific sources: the published volume index of industry contracts and a new statistic under development at HCSO called Hungarian Truck Toll Mileage Index, that measures how much trucks travelled in the month in the country.

2.2 The nowcasting algorithm

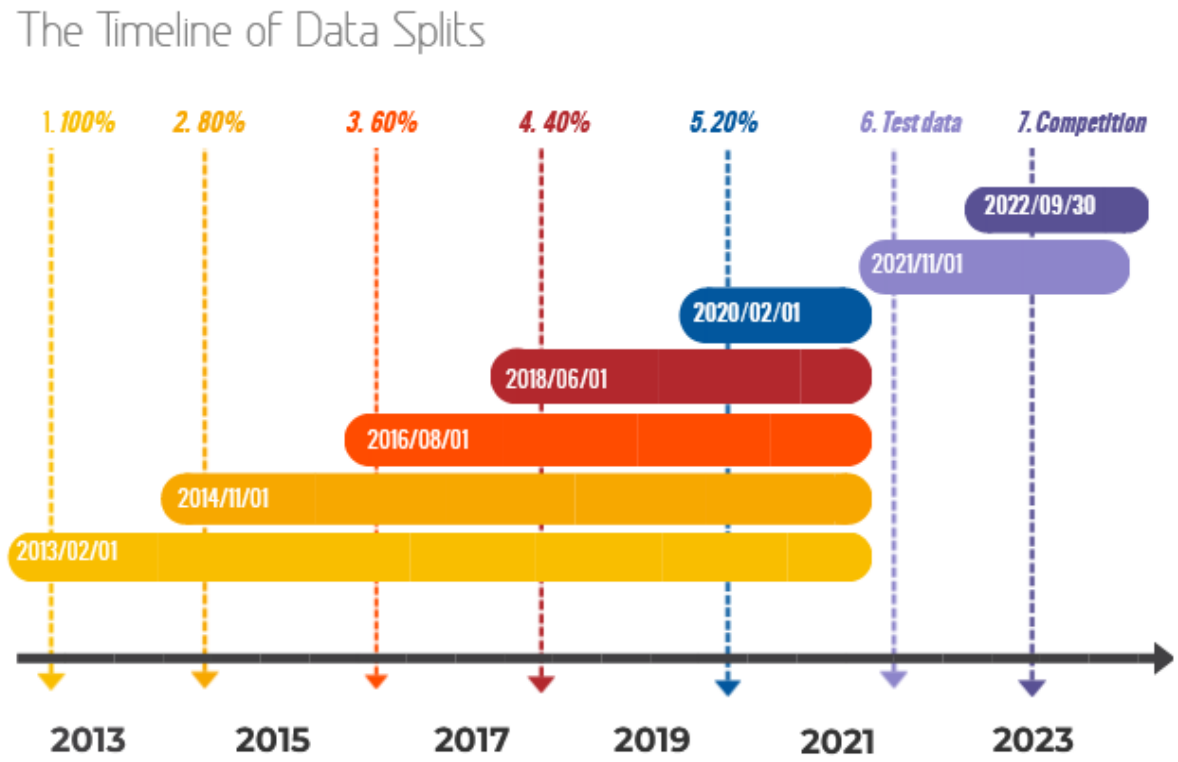
The current state of the algorithm works as following: use standard transformations of lagging and differencing the time series to create a large pool of possible explanatory variables. Then, it uses several different feature selection procedures to create a more concise pool of variables. After the pool is complete, it starts the hyperparameter tuning using different sample sizes, estimation window types and one-step ahead forecasts for cross validation.

We define our own version for splitting the time series as we think backwards approach is

rather plausible then a forward approach from the beginning of the time series due to the fact that we want to be more precise in the most recent periods and the more recent the data the more related it should be for the nowcast. This means we define the train and test sample cuts to be at the last test size portion of the whole event window. While the number of splits limits how much data should the training section contain from the cut backwards in time. In default, the first split contains only 20% of the train data, the second 40%, the third 60%, etc until it contains the whole event window. The Figure 1 illustrates how the different splits are constructed.

Then we do a feature selection that is taking all the data relevant for the given indicator. The feature selection methods are based on correlation, random forest regression, Ridge regression and LASSO. Each select the best 5-10 features on their own and then we use the union of all the sets of selected features. After selecting the features, the final train data is collected as following: take the union of all the features selected by any of the feature selection methods and create a train data for each cut of the train splits. In the end resulting in 5 different training data that all contain the same variables but their time horizon is different.

Figure 1: Illustration of data splits.



After acquiring the final version of test data and train data splits the hyperparameters are up to be selected. For creating the hyperparameter grid we defined a grid for each model separately and due to the feasible size of the defined hyperparameter space we decided to

iterate over all possible combination of hyperparameters for all models. For cross validation, we chose to define our own method due to the sequential characteristics of the timeseries and the specific need to create a one-step ahead forecast to evaluate properly. Therefore, we defined the cross-validation method to slice the data at each month of the training data set. Resulting each cross-validation split to contain the whole period of training data in one-month steps. Of course, we have to use a restriction in this case not to estimate any model on only one data point and added a minimal size for the first split, that is defined as how large ratio of the whole period should be used for the first estimation and forecast. We also experimented with two different event windows for estimation one that is a sliding window always including a fixed number of data points, and an expanding window that always expands with the latest observations. We found that the latter performed better in case of shorter time series.

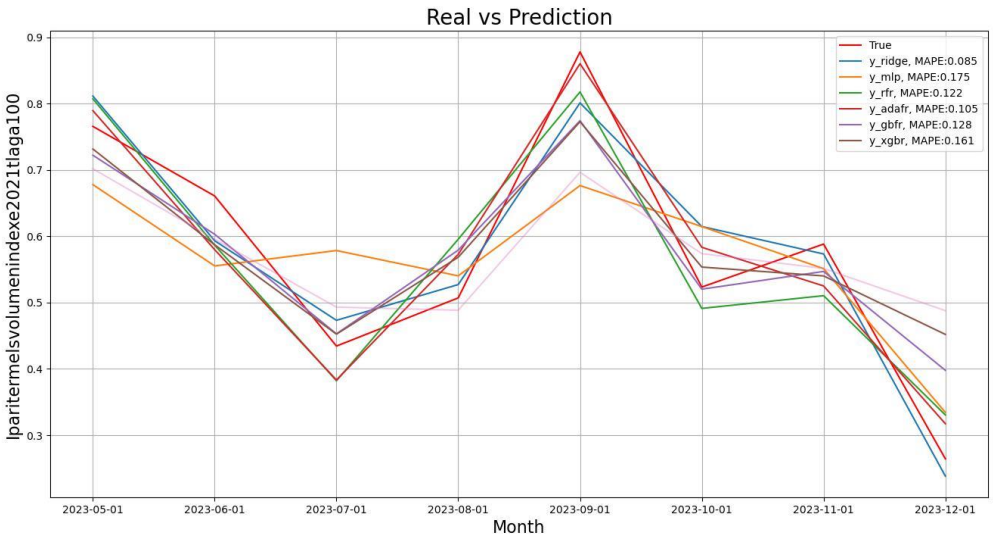
Then with these cross-validation splits we calculate for each model at each hyperparameter setting the predicted one-step ahead forecast for each cross-validation split. Creating a time series of one-step ahead forecast that we use to calculate the overall RMSE on the training period for the hyperparameter settings and then save the best hyperparameters for each model.

After finding the hyperparameters we proceed to prepare the forecasting. The train samples are extended with the test period and the models are estimated again in cross-validation splits in one-months steps to create a test period predicted time series that we can compare to the original test period values. A predicted series is created for each model and data size. Then we evaluate, all the predicted time series for the test period using MAPE. After selecting the best three and five predictions in MAPE order we create 2 ensemble predictors too that are the average of them to check if those perform better than the created predictions. This is done once in a seasonally unadjusted and once in the adjusted settings.

2.3 Result

Currently we have results only for the PVI the algorithm provides the estimates that are fairly close to the original time series although it leaves place for improvement. O As it can be seen on the figure most of our best predictors follow the ups and downs of the original PVI series although their precision should be better as the best performer only reaches a little under 10% MAPE.

Figure 2: PVI estimates and their MAPE



3. Conclusions

We are focusing first on creating a framework on how to use the nowcasting techniques and exploring data sources. The former is to make the algorithm modulary updatable later on and easy to implement to new time series. The latter is to be able to tailor the modelling and data soruces for the specific needs of the indicators. The base of the framework is fixed now and we will focus on extending our data sources to be able to provide a starting point or benchmark for further experimenting for each indicator

Our current result also shows that after finding economicly established data for the target variables we will still need to further improve our algorithm. We are planning to extend it with state-of-the-art tehniques such as Long-Short Term Memory Networks and other Recurrent Neural Networks. However, it must be noted that the next main steps are to include new data sources that can help in creating the first benchmark results.

Acknowledgment

We thank Gábor Lovics, Mária Pécs, Csanád Temesvári and Miklós Salánki for their helpful insights and comments.

References

Knotek, Edward S., II, and Saeed Zaman. 2023. “A Real-Time Assessment of Inflation Nowcasting at the Cleveland Fed.” Federal Reserve Bank of Cleveland, Economic Commentary 2023-06. https://doi.org/10.26509/frbc-ec-202306

Knotek, Edward S., II, and Saeed Zaman. 2014. “Nowcasting U.S. Headline and Core Inflation.” Federal Reserve Bank of Cleveland,Working Paper No. 14-03. https://doi.org/10.26509/frbc-wp-201403

Aparicio, Diego, and Manuel I. Bertolotto. “Forecasting Inflation with Online Prices.”*International Journal of Forecasting*, vol. 36, no. 2, Apr. 2020, pp. 232–47. *ScienceDirect*, https://doi.org/10.1016/j.ijforecast.2019.04.018.

Macias, Paweł, et al. “Nowcasting Food Inflation with a Massive Amount of Online Prices.”*International Journal of Forecasting*, vol. 39, no. 2, Apr. 2023, pp. 809–26.*ScienceDirect*, https://doi.org/10.1016/j.ijforecast.2022.02.007.

Beck, Günther W., et al. „Nowcasting consumer price inflation using high-frequency scanner data: evidence from Germany“ ECB Working paper No. 2930, Apr, 2024

The World Bank Prospects Group „World Bank Commodity Price Data (The Pink Sheet) Monthly Prices“, Apr, 2024, url: https://www.worldbank.org/en/research/commodity-markets