GRANT journal
ISSN 1805-062X, 1805-0638 (online)
Vol. 11, issue 02
EUROPEAN GRANT PROJECTS | RESULTS | RESEARCH & DEVELOPMENT | SCIENCE

Economic Forecasting using Artificial Intelligence

Ádám Csápai[1]
Anna Kalafutová[2]

[1]Department of Operations Research and Econometrics, Faculty of Economic Informatics, University of Economics in Bratislava; Dolnozemská cesta 1, 852 35 Bratislava; adam.csapai@euba.sk
[2]Department of Public Administration and Regional Development, Faculty of National Economy, University of Economics in Bratislava; Dolnozemská cesta 1, 852 35 Bratislava; adam.csapai@euba.sk

Grant: I-22-110-00
Name of the Grant: Projects of young teachers, researchers and doctorate students at the University of Economics in Bratislava
Subject: AH - Economics

Abstrakt In this paper we compare the prediction capabilities of various artificial intelligence methods. It has been shown that with the help of these methods we can more accurately predict economic variables when we have a large number of datapoints. We forecast interest rates, inflation and wages in Slovakia. We use 37 times series for forecasting. We use a simple linear regression model as a benchmark. Next, we forecast using selected ensemble machine learning techniques. These techniques are bagging, random forests and boosting. We do short term forecasts and compare the RMSE of the models. Our findings are in line with the existing literature, based on which artificial intelligence methods can increase forecast accuracy if a relatively large dataset is available.

Key words Prediction, artificial intelligence, Slovakia

1. INTRODUCTION

The basis for making economic and political decisions is data that helps us monitor macroeconomic conditions. Methods for tracking economic conditions using big data have evolved over time, and so econometric techniques have advanced in emulating, explaining, and automating the best practices of forecasters in investment markets, central banks, and other market monitoring tasks. Forecasting is mainly used to evaluate the state of the economy, for example the development of GDP. Forecasting models are used, for example, to monitor the state of the economy and the subsequent adoption of measures by the Central Bank. In this paper we show different forecasting methods based on big data and compare their performance to a simple linear regression model based on RMSE. In the second section we present the models, the third section contains the results while the fourth section concludes.

Many authors have written in detail about big data processing using predictive tools (Friedman, et al., 2001; James et al., 2013). The use of big data creates problems in the field of modelling, namely distortion of the results or the creation of false positive results. The advantage of using big data is that they are not subject to subsequent revisions and provide the earliest possible information about the state of the economy (Baldacci et al., 2016).

In this paper, we specifically focus on bagging, random forests and boosting techniques. By using bagging, we average the values across models and improve the estimation performance. Boosting produces an iterative estimator given a misclassified observation (Varian, 2014) and provides a sketch of solutions that can correspond to the soft-thresholding estimator in linear models (Kapetanios and Papailias, 2018). At the same time, it is consistent with a linear model where variables can grow rapidly and the model gradually selects the best ones that suit it (Kapatenios and Papailis, 2018). Boosting estimates parameters impartially, but with the impossibility of using missing values (Holmes, Ward and Scheuerell, 2020). Random forests produce out-of-sample matches, and there are many variations in their use. They, for example, create simple summaries of the relationships in the data (Varian, 2014).

2. METHODOLOGY AND METHODS

In this chapter we introduce the applied methods, namely the linear regression as the benchmark and regularized least squares and ensemble machine learning methods. The section is based on Maehashi and Shintani, who provide a comprehensive overview (Maehashi and Shintani, 2020).

We evaluate model performance using the Root-mean-square error metric. It is a measure of differences between predicted and observed values of the same dataset. It represents the quadratic average of these differences. We write the RMSE as

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}} \qquad (1)$$

where i denotes variable i, N the number of data points, x_i represents the actual observations while $\hat{x}_i$ stands for the estimated time series.

Our first model is a simple linear regression model, which we write as

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\top}\boldsymbol{\beta} + \varepsilon_i,\ i = 1, \dots, n, \qquad (2)$$

we assume that the reader is aware of the properties of this model so we do not provide in depth description in this paper and continue with describing more advanced methods.

Ensemble machine learning methods. Our methods employ decision (or regression) trees, as one of our aims is to account for possible nonlinearities in the data. These trees detection groups of similar observations by generating nodes step by step within the tree. Assume that y_{t+h} is our target variable representing the selected macroeconomic time series we want to forecast. Initially, each observation of y_{t+h} is sorted into nodes based on some predictor variable $X_i = (x_1, x_{2x}, \dots, x_N)$. Selected nodes are assigned the value of the sample mean y_{i+h} conditional on a selected predictor. If there are nodes left with no assigned values, they are divided by using the remaining predictors. The process ends when we assign values to all the nodes.

Based on this we can write a regression tree with M terminal nodes as

$$y_{t+h} = \sum_{m=1}^{M} \theta_m 1_{|x_i \in h_m]} + \varepsilon_{i+h_n}, \qquad (3)$$

where$1_{|x_i \in h_m]}$ represents the indicator function, R_m is a portion of the space of X_b and θ_m gives us the sample mean of y_{th} conditional on $X_t \in R_m$. The aim of the estimation is to select the tree structure which minimizes $\sum_{i=1}^{T} \varepsilon_{i+h}^2$. To find this tree we select sorting variables from X_t and pick splitting values at each node. We use the algorithm which selects the optimal values for sorting and splitting, respectively (Breiman et al., 1984). Regression trees do well if nonlinearity and variable interactions are present, but their out-of-sample forecast performance is generally suboptimal, because they are sensitive to changes in the data. To solve this problem, we not use regression trees themselves but they serve as the basis of our ensemble machine learning methods, namely bagging, boosting and random forests.

Bagging. Bagging stands for the bootstrap aggregating procedure. The capability of the method to improve forecast accuracy has been empirically shown (Breiman, 1996). Bagging also reduces forecast errors for i.i.d. data (Bühlmann and Yu, 2002). The same has been shown for time series data as well (Inoue and Kilian, 2008). In bagging, we generate bootstrap samples of $X_t = (x_i, x_2, \dots, x_{\mathrm{N}})'$ and y_{t+h} B times and then a regression tree computes the forecast $\hat{y}_{t+\mathrm{h}}^{(b)}$ for each bootstrap sample $X_t^{(b)}$ and $y_{t+\mathrm{h}}^{(b)}$. In the last step we average the forecasts of each boostrap sample $B^{-1}\sum_{b=1}^{B} \hat{y}_{t+\mathrm{h}}^{(b)}$, which diminishes the overfitting and large volatility problem of individual forecasts. In our application below, we set the number of bootstrap samples at $B = 10$.

Random forests. Random forests are a derivative of bagging (Breiman, 2001). Bagging forecasts are stable only if regression trees of different bootstrap samples are not highly correlated. If they are, averaging might not be sufficient to reduce forecast variance, since individual regression trees in bootstrap samples are similar.
A dropout procedure has been proposed for decorrelating regression trees of individual samples (Hastie et al., 2009). More precisely, the set of predictors $X_i = \left(x_{16}, x_{21}, \dots, x_{N_i}\right)'$ is reduced by randomly drawing subsets $X_i^* = (x_{it}^*, x_2^*, \dots, x_{kt}^*)'$ where $k < N$. For each X_i^* bagging is employed as a forecast method as $B^{-1}\sum_{b=1}^{E} \hat{y}_{t+h}^{(b)}$ where $\hat{y}_{t+h}^{(b)}$ is computed using a bootstrap sample $X_t^{*(b)}$ and $y_{t+b}^{(b)}$. This procedure is repeated for multiple subsets and the forecast average is calculated for each. The correlation is reduced because subsampling results in differently structured regression trees, which should lead to stable forecast. In this paper the subset of predictor variables is set at $k = N/2$.

Boosting. Boosting was introduced boosting as an alternative solution to the overfitting problem (Schapire, 1990).
Assume that $\sum_{m=1}^{M} \theta_m 1_{[X_t \in R_m|}$ gives us a simple regression tree with initial value of $f_0(X_i) = \eta \sum_m^M \theta_m 1_{[X_t \in R_m|}$. $\eta \in (0,1)$ represents the learning rate set at $\eta = 0.1$. In boosting the depth of regression trees should be shallow, which implies that each base learner $f_s(X_t)$, for $s = 0,1, \dots S$, is a weak learner. In the actual stage the algorithm employs information of forecast errors from previous trees and searches for a new algorithm with L_2 loss function based on this information. This produces model updates at s-th stage using

$$f_s(X_t) = f_{s-1}(X_t) + \eta \sum_{m=1}^{M_s} \theta_{sm} \mathbf{1}_{[X_t \in R_{sm}]}, \qquad (4)$$

where $\sum_{m=1}^{M_s} \theta_{sm} \mathbf{1}_{[X_t \in R_{sm}]}$ is estimated for the residual from $(s-1)$-th stage, $y_{t+h} - f_{s-1}(X_t)$. The model is updated until s reaches a set limit on boosting stages.

3. DATA

The dataset consists of 37 Slovakian macroeconomic time series from and spans the period from November 2008 to December 2019, meaning that $N = 37, T = 137, N \times T = 5069$. The target variables are the interest rate, inflation and wage rate. We do short term one-period-ahead forecasts as a preliminary analysis.
Firstly, we clean the data of all NA values and then split it to two subsamples, which we use for the predictions. The first subsample consists of 70% of all observations of all the predictors and serves as the training data set. The second subsample consist of 30% of all the observations and serves as the testing data set. After we split the data in half, we start estimating the aforementioned models.

4. RESULTS

In this section we present our results. To save space, we only present one figure instead of three figures for each individual forecasted variable. We begin our analysis with estimating a simple linear regression model, which serves as our benchmark. Figure 1 compares the actual and predicted values of the linear regression model. According to Figure 1 the prediction of the simple linear regression model is fairly accurate. As the measurement of accuracy we use the Root Mean Squared Error (RMSE), as stated in the methodology section. The RMSE of this model is 0.04494223 for interest rates, 0.04444563 for inflation and 0.03226302 for wages. All of these values can be considered fairly low

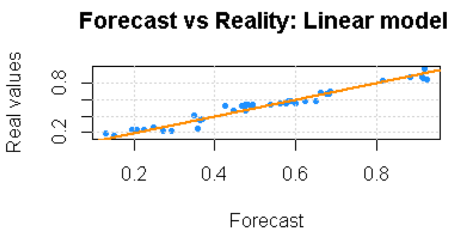

Figure 1: Forecast vs Reality: Linear model

Our next model is just a single regression tree, which is inappropriate on its own, but can showcase if and how much improvement bagging, random forests and boosting provides compared to it. The RMSE of the single tree model is 0.05345150 for the interest rate, 0.04017693 for inflation and 0.06363820 for the wage.

We employ bagging, described above, as the first ensemble machine learning method. We use 13 bootstrap samples and 500 trees for each forecasted variable, which allows us to accomodate the relatively large dataset. Figure 2 gives us the prediction results. We see that Figure 2 is almost identical to Figure 1 and we cannot really make any distinction based only on these graphs. To better encapsulate the model performance, we calculate the RMSE for each variable. The interest rate has an RMSE of 0.02687112, inflation 0.02437744 and wage 0.02547582. The improvement is smaller in the case of wage, but bagging is almost two times more accurate in the case of the interest rate and inflation. In addition, we see that the model benefitted from setting the number of trees higher instead of just one, which is illustrated on Figure 3. The increasing number of trees results in an exponential decrease in RMSE.

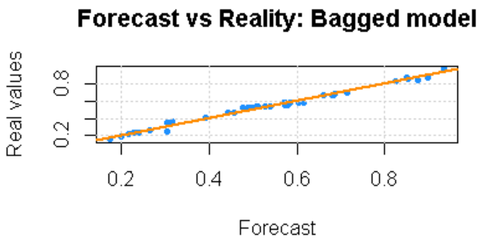

Figure 2: Forecast vs Reality: Linear model

Our next method of choice is the random forest. We are interested in confirming whether all ensemble machine learning methods overperform or it is just bagging. The description of random forests is provided in chapter 2. The number of variables randomly sampled as candidates at each fit is 4, while the number of trees remains the same. The figure comparing predicted and actual values is very similar to the above figures so we leave it out to save space. The RMSE of the interest rate is 0.03074335, of inflation is 0.02752882 and it is 0.02765973 for wage. The last of this class of methods left is boosting.

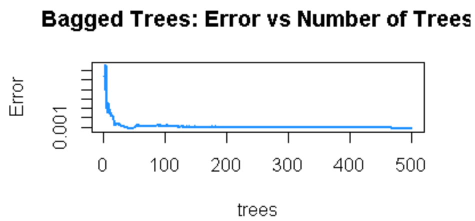

Figure 3: Bagged Trees: Error vs Number of Trees

Based on the method's description we set the number of trees higher to 5000. On Figure 4 we see none of the predictors had zero influence on the outcome. In the case of inflation, for example, most of the predictors, however, had quite small influence, while others, such as interest rates had higher than average. The RMSE of this boosted model for the interest rate 0.03494127, for inflation is 0.02157115 and for the wage is 0.02689550. Table 1 provides a concise overview of all the results and it is clear that all of the machine learning methods outperform the benchmark model for all of the variables.

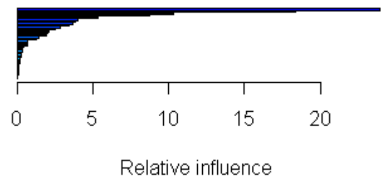

Figure 4: Relative influence of all variables.

Table 1: RMSE of ensemble machine learning methods

Model/RMSE	Interest rate	Inflation	Wage
Linear Model	0.04494223	0.04444563	0.03226302
Single Tree	0.05345150	0.04017693	0.06363820
Bagging	0.02687112	0.02437744	0.02547582
Random Forest	0.03074335	0.02752882	0.02765973
Boosting	0.03494127	0.02157115	0.02689550

5. CONCLUSION

In this paper we do short term predictions of the interest rate, inflation and wage in Slovakia using a relatively large dataset. To fully utilize big data, we employ machine learning methods capable of dealing with this kind of dataset. These methods are expected to be more accurate than our general benchmark model. Our preliminary results are appealing. Ensemble machine learning methods are almost two times more accurate than our benchmark if measured by RMSE in a short horizon. This suggests the presence of nonlinearities and variable interactions in the data. Moving on we plan to enlarge the dataset, apply the methods to industrial production and test the model performance on multiple forecast horizons for all of the Visegrad Four countries.

Sources

1. BALDACCI, E. et al. Big Data and Macoeconomic Nowcasting: from data access to modelling. 2016 edition. Luxembourg: Publication Office of European Union, 2016. 13 s. ISBN 978-92-79-63963-0
2. BREIMAN, L. Bagging predictors. In: Machine Learning, 24 (2), 1996, pp. 123–140.
3. BREIMAN, L. Random forest. In: Machine Learning 45 (1), 2001, pp. 5–32.
4. BREIMAN, L. FRIEDMAN, J., STONE, C.J., OLSHEN, R.A.. Classification and Regression Trees. In: Chapman and Hall/CRC, New York, NY, 1984.
5. BÜHLMANN, P., YU, B. Analyzing bagging. In: Annals of Statistics 30 (4), 2002, pp. 927–961.
6. FRIEDMAN, J., HASTIE, T., TIBSHIRANI, R. The elements of statistical learning. In Springer, 10 (1), 2001, pp. 1-764.
7. HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. The Elements of Statistical Learning. 2nd ed. New York: Springer, 2009.
8. HOLMES, E. E., WARD, E. J., SCHEUERELL, M. D., Analysis of multivariate time-series using the MARSS package. Version 3.11.3. Nortwest Fisheries Science Center, NOAA, 2020. 333 s.
9. INOUE, A., KILIAN, L. How useful is bagging in forecasting economic time series? A case study of U.S. consumer price inflation. In: Journal of American Statistical Association 103 (482), 2008, pp. 511–522.
10. JAMES, G., WITTEN, D., HASTIE, T., TIBSHIRANI, R. An introduction to statistical learning. In Springer 112, 2013, pp. 18.

11. KAPETANIOS, G., PAPAILIAS, F., Big Data & Macroeconomic Nowcasting: Methodological Review, ESCOE, 2018, pp 1-74.
12. MAEHASHI, K., SHINTANI, M. Macroeconomic forecasting using factor models and machine learning: an application to Japan. In: Journal of the Japanese and International Economies (58), 2020, pp. 1-17.
13. SCHAPIRE, R.E. The strength of weak learnability. In: Machine Learning 5 (2), 1990, pp. 197–227.
14. VARIAN, H.R. Big Data: New Tricks for Econometrics. In Journal of Economic Perspectives, 28 (2), 2014, pp 3-28