Inflation Forecasting Using Machine Learning Methods

Ivan Baybuza, *Ludwig Maximilian University of Munich**
ibaybuza@nes.ru

Inflation forecasting is an important practical problem. This paper proposes a solution to this problem for Russia using several basic machine learning methods: LASSO, Ridge, Elastic Net, Random Forest, and Boosting. Despite the fact that these methods already existed in the early 2000s, for a long time they remained almost unnoticed in the professional literature related to the forecasting of inflation in general, and Russian inflation in particular. This paper is one of the first attempts to apply machine learning methods to the forecasting of inflation in Russia. The present empirical study demonstrates that the Random Forest model and the Boosting model are at least as good at inflation forecasting as more traditional models, such as Random Walk and autoregression. The main result of this paper is the confirmation of the possibility of more accurate forecasting of inflation in Russia using machine learning methods.

Keywords: *inflation forecast, machine learning, boosting, random forest*

JEL Codes: *C53, E37*

Citation: Baybuza, I. (2018). Inflation Forecasting Using Machine Learning Methods. *Russian Journal of Money and Finance*, 77(4), pp. 42–59.

doi: 10.31477/rjmf.201804.42

1. Introduction

It is difficult to overestimate the importance of inflation forecasting for rationally thinking and acting economic agents: numerous economic obligations, including wages and interest rates, are usually expressed in nominal prices. In practice, central banks implement monetary policy guided mainly by their expectations of how inflation will behave in the short or medium term, rather than by its current values, since the rate of inflation does not react immediately to the tightening or easing of monetary policy, but rather with a certain lag. Therefore, price forecasting is important both for households and businesses, and for official authorities.

* The author is a graduate of the HSE/NES Joint Bachelor's Program in Economics (BAE 2018) and a student on the Master's Programme in Economics of Ludwig Maximilian University of Munich, Germany.

One of the key tasks of central banks is to maintain price stability. Over the past three decades, many central banks have adopted inflation targeting policies to solve this problem.

Inflation targeting policy is based primarily on public confidence in monetary authorities. Central banks introduced the practice of publishing their own forecasts for inflation and other key macroeconomic variables in order to establish and maintain this confidence. As a result, the problem of the quality of these forecasts became more acute. The Bank of Russia completed the transition to an inflation targeting regime in December 2014, and in these circumstances, the forecasting of inflation in Russia is now more important than ever.

The task of inflation forecasting can reasonably be divided into two parts: short- and long-term forecasting. It is not easy to explicitly divide up forecast horizons, but it is intuitively clear that forecasting inflation for one quarter and for five years are different tasks. While it may be sufficient to use the idea of the neutrality of money (Romer, 2012, p. 514–515) for predicting inflation in the long term, inflation forecasting in the short term is a more challenging problem.

According to current research in this area (for example, Stock and Watson, 2007 and 2008; also Faust and Wright, 2013), it is known that the behaviour of inflation in the short term can be closely approximated using fairly simple models based only on a time series of inflation. At the same time, predictions involving other macroeconomic indicators are worse than single-factor models.

Inflation forecasting using other macroeconomic variables as predictors has two serious limitations: on the one hand, a rather large number of potentially informative predictors, and on the other hand, the limited duration of the time series available, which leads to the so-called 'curse of dimensionality' (Stock and Watson, 2011). The dynamics of inflation are influenced, to varying degrees, by many different macroeconomic factors, and the maximum length of the time series that can be used is about 700 time observations (US, monthly data).

The above limitations can potentially lead to the so-called 'overfitting problem', the result of a small number of time observations relative to the number of explanatory variables. Here, overfitting means adjusting the model to random patterns in the 'learning' sample which are absent in the general population. In other words, the model achieves a very low prediction error in sample, but gives a very inaccurate forecast when predicting out of sample. This is why, in practice, multivariate models often give less accurate inflation forecasts than univariate ones.

The overfitting problem can be solved by pre-selection of explanatory variables on the basis of theoretical concepts, for instance, by selecting rates of real activity. However, this approach has a number of potential vulnerabilities. First, the predictive power of explanatory variables can change over time, as well as depending on the forecast horizon. Quite often, variables that predict well one or two months ahead can give a very inaccurate inflation forecast six months or

a year ahead, and vice versa. Second, overfitting of explanatory variables for a particular data sample remains possible. The selection of explanatory variables itself can be viewed as a kind of initial hyperparameter, on which the quality of the forecast primarily depends.

For these reasons, pre-selection of parameters should be carried out in a subsample of data, but not solely on the basis of theoretical concepts, in order to improve the prediction results.

Approaches to solving the overfitting problem have long been a preoccupation of the field of computer science known as machine learning (ML). Over the past few decades, many different ML models have been created: LASSO regression (Least Absolute Shrinkage and Selection Operator), Ridge regression, Principal Components Analysis, Decision Trees, Random Forest (RF), Boosting, Neural Networks, etc. Their application has led to significant breakthroughs in areas such as text categorisation (Sebastiani, 2002) and image recognition (Simonyan and Zisserman, 2014). However, many powerful ML methods such as the Random Forest model have started to be used to predict macroeconomic variables only relatively recently. In particular, works have begun to appear in which the authors use ML methods to predict inflation (Chakraborty and Joseph, 2017). The author of this paper has not come across any published works on this topic regarding the Russian economy.

In the light of the above, the main purpose of this paper is to test ML methods for forecasting inflation in Russia. For this task, the following popular ML methods were selected: LASSO and Ridge regressions, Elastic Net model, Random Forest model, and Boosting. Forecasts obtained using ML methods are compared with results obtained using traditional econometric methods: Random Walk, Autoregressive model of order 1 (AR(1)), and Autoregressive model of order p (AR(p)).

This paper has the following structure: Section 2 contains a review of literature on the subject; Section 3 describes data and methods used in the study; Section 4 presents the models used in this paper; Section 5 shows the results of the study; and Section 6 contains the conclusion to the study.

2. Literature review

This review presents the works that influenced the choice of methods used in this paper. Works by Stock and Watson (2008) and Faust and Wright (2013) are frequently cited in the field of inflation forecasting. The basic models used in this study (random walk, autoregressions of orders 1 and p) were chosen based on the results obtained in these works. In the next two works by Chakraborty and Joseph (2017) and Garcia et al. (2017), ML methods are used to predict inflation. Finally, the work by Andreev (2016) is important for understanding the actual methods used by the Bank of Russia to forecast Russian inflation.

A more detailed review of each of the works is provided below.

Stock and Watson (2008) compare many different models, dividing them into four main groups. The first group includes models based only on the use of a time series of inflation: Autoregression of Moving Average (ARMA), RW, as well as the authors' own model with an Unobserved Components and a Stochastic Volatility (UC-SV). In the second group, the authors include models in which the explanatory variables are indicators of economic activity, first of all, and then the unemployment rate and the output gap. The third group includes models in which predictions are based on expected inflation or forecasts such as surveys of professional forecasters (SPFs). The fourth group includes models in which the explanatory variables are variables other than indicators of economic activity, i.e. variables which are not used in the second group models. The authors model future inflation for four quarters, obtaining forecasts in pseudo-real time in a rolling window of 10 years. The quality criterion of the model is the Root Mean Squared Error (RMSE) relative to the error of the UC-SV model as a reference. The main conclusion of the work is that, as far as the quality of the forecast is concerned, models based on indicators of economic activity do not systematically improve on univariate models that take into account only the dynamics of inflation itself.

The work by Faust and Wright (2013) is a review and comparison of the best models for inflation forecasting at that time. The authors compared 17 models, including the AR model, the UC-SV model, the RW model in two variations (RW and its modification, RW-AO[2]), the Phillips Curve-based model, Structural Vector Autoregression (SVAR), Bayesian Model Averaging (BMA), the Dynamic Stochastic General Equilibrium model (DSGE), and more. The authors nowcast inflation in the current quarter (h=0) and forcast it up to eight quarters ahead, obtaining forecasts in pseudo-real time in an expanding window. The quality criterion for the model is the Root Mean Squared Error (RMSE) of the model relative to the error of the AR(1) model, which is selected as a reference. The authors conclude that the models based on other predictions (primarily surveys of professional forecasters) have the best predictive capacities. Moreover, the authors conclude that very simple methods, such as random walk, predict inflation surprisingly well. In general, the findings of their review correlate well with those of the previous work: both studies support the hypothesis that the best multi-factor models, both with activity rates and with other variables as predictors, systematically fail to surpass the best univariate models that use only a time series of inflation.

The work by Chakraborty and Joseph (2017) is a review of the best ML methods in terms of their practical application for solving several important problems faced by central banks. For our purpose, the forecasting of inflation in the UK, as carried out by the authors of the article, is of particular interest. The authors forecast inflation for the medium-term horizon of two years, using data

[2] A variant of the random walk model was proposed by Atkeson and Ohanian (2001).

for various macroeconomic variables (money supply, unemployment rate, bank rates, GDP, and other indicators) between the first quarter of 1988 and the fourth quarter of 2015. The models are trained on an initial 15-year window (the first quarter of 1990 to the fourth quarter of 2004) with further quarterly expansion to the fourth quarter of 2013; the quality measure is Mean Absolute Error (MAE) for the entire forecasting period, with the period before and after the crisis of 2008 taken separately. The authors apply the following models: Nearest Neighbours, Decision Tree, RF, Neural Networks, Support Vectors, Support Cectors combined with Neural Networks, Ridge regression, VAR(1), AR(p), and AR(1). AR(1) is used as the benchmark. As a result, the authors come to the conclusion that the model of Support Vectors combined with Neural Networks gives the highest-quality forecasts for the entire forecasting horizon, while the Random Forest model performs best in the post-crisis period. However, all other ML methods also show better results than traditional benchmarks in the form of VAR(1), AR(p), and AR(1).

In Garcia et al. (2017), inflation in Brazil is predicted by ML methods in 12 different periods. The first forecast period is five days before the release of inflation statistics, and the twelfth forecast period is 11 months and five days before the release of inflation statistics. In their work, the authors use various macroeconomic variables (a total of 59 series), reflecting the state of the financial market, the labour market, the balance of payments, and the public debt of the country. They also use professional forecasts of Brazilian inflation collected during the FOCUS survey conducted by the Central Bank of Brazil. In the work, a time series from January 2003 to December 2015 is used. Models are trained on a nine-year rolling window; the RMSE and the MAE serve as criteria for the quality of the forecast. The authors apply the following models: RW, AR(p), where the number of lags is chosen based on the Bayesian information criterion, the dynamic factor model using principal component analysis, three varieties of LASSO regression, RF, Complete Subset Regression (CSR), and two models based only on surveys of professional forecasters. As a result, the authors come to the conclusion that inflation for five days and for one month and five days is best predicted by the LASSO model and the model that uses forecasts obtained from the surveys. However, the authors note that the LASSO model actually cuts off all variables except survey forecasts; if survey forecasts are excluded from the model, the LASSO result is significantly worse. From the third forecast horizon (two months and five days) to the last (11 months and five days), the best results are given by the CSR model. In addition, the RF model and the dynamic factor model perform better than the AR and RW models over all forecasting horizons.

Andreev (2016) describes a combined forecast method for forecasting inflation in Russia. The main goal of the algorithm presented in the work is to combine different models without preliminary screening of potential predictors. The following models are included in the algorithm: Vector Autoregression,

Bayesian Vector Autoregression (BVAR), Ordinary Least Squares regression (OLS regression), RW, Linear Threshold Autoregression (LTAR), and the Unobserved Components (UC) model. Multivariate methods, such as VAR, BVAR, and OLS, are trained not on all variables, but only on some combination of them. In other words, *n* subsets of the initial data array are created, whith only some of the variables included. After training, they are aggregated on each sub-sample of the base model. Thus, combined forecasts using the VAR model, the BVAR model, and the OLS model are constructed. Finally, multivariate and univariate models are aggregated in a combined forecast. This algorithm makes use of the different strengths of particular models, using all variables.

In our study, we used the RF and boosting methods, which belong to the class of so-called 'ensemble' methods. The methods of this class are based on the general idea of using multiple training algorithms and subsequently combining them so that the final forecast is more accurate than any individual forecast. We expect that results of our study may be of practical value to the Bank of Russia, by expanding the range of potential methods for inflation forecasting.

3. Data description and methodology

The main sample consists of 92 macroeconomic series, excluding the time series of price level (93 series if this is included). The data reflect the state of business activity, industrial production, the financial market, the employment rate, the balance of payments, and the prices of the main export goods of the Russian economy. The consumer price index (CPI) was chosen as a measure of inflation. Explanatory variables are taken for the period from February 2002 to June 2016 (173 observations). CPIs are used for the period from February 2002 to January 2018 (192 observations). The data represent either the monthly rate of change of the relevant variables or their monthly values in levels. All series are adjusted for seasonal and calendar factors, rendered stationary and standardised. Table 1 of the Appendix gives all the variables and the means by which they were transformed.[3]

Researchers working with this data may come up against the so-called 'ragged edge problem' associated with uneven disclosure of data on various indicators by statistical agencies. Inflation data are published by the Federal State Statistics Service (Rosstat) at the beginning of the month following the reporting month, while other key macroeconomic variables are published at the end of the month. This problem may prevent real-time forecasting. However, when forecasting in pseudo-real time, when all data are known, the problem is not so pressing. Therefore, no adjustments are made for the ragged edge problem in this work.

[3] The author of this paper expresses his deep gratitude to Professor Konstantin Styrin of the NES for providing the data.

After adjusting for seasonal and calendar factors, the first difference in the logarithms of price levels was taken to obtain the monthly inflation:

$$\pi_t = \log(CPI_t) - \log(CPI_{t-1}).$$

Figure 2 (see Appendix) shows trends in monthly inflation over time; in principle, it shows that the first difference of logarithms may be sufficient to render the series stationary. We conducted the extended Dickey-Fuller test, which confirmed the hypothesis that the price level in Russia is an integrated process of order 1.

The quality of forecasts is estimated using the root mean squared error (RMSE) of the forecast:

$$RMSE = \sqrt{\frac{1}{T - T_0 - 1} * \sum_{t=T_o}^{T} (\pi_t - \hat{\pi}_t)^2}.$$

The forecasting is performed in pseudo-real time out of sample, i.e., for dates outside the limits of the estimation sample, with a rolling 10-year window. Predictions are obtained out of sample, because when training with a large number of predictors, the high accuracy of forecasting in the training sample likely indicates not the high quality of the model, but its overfitting. Inflation is predicted 1, 2, 3,..., 24 months ahead, following the last available monthly inflation value in pseudo-real time. In addition, average inflation values are calculated on the horizon of one month, two months, one quarter, half a year, one year, a year and a half, and two years. The models are compared in two ways: according to individual monthly inflation forecasts and according to average values on the above mentioned time horizons.

All RMSE values are considered in relation to the corresponding value of the AR(1) bencmark model: if the RMSE value of any model on a certain horizon is less (more) than one, then this model predicts inflation in such a month or over such a horizon better (worse) than the benchmark model.

It is important to note that no multivariate models, except for the AR(1) and LASSO combination, employ inflation lags in training or predict inflation using them. This is done deliberately to determine the strength of ML models in comparison with single-factor standards.

4. Models

This section describes the models used in the work. To begin, two traditional econometric models based only on past trends, the random walk model and the AR model, are briefly described. The following is a description of the ML models (LASSO, Ridge, Elastic Net, Random Forest, and Boosting) used in this paper, as well as their various specifications.

4.1. Random walk (RW)

This is the simplest model considered in this paper, but it has a good predictive capacity, which in practice is not so easy to improve on by including additional predictors. Mathematically, the model has the following form:

$$\pi_{t+h} = \pi_t + \varepsilon_{t+h},$$

where $h = 1, 2, \ldots, 24$, ε_t is an unexpected fluctuation in inflation.

4.2. Autoregression (AR)

This paper uses a recursive version of the autoregression model. This means that, in order to predict for h periods in advance, the missing inflation values between the time t and the time $t + h$ are consistently predicted. According to Faust and Wright (2013), the iteration method produces more accurate forecasts than the direct method (when the inflation forecast is immediately constructed at the time $t + h$). The number of lags p in the AR(p) model is selected using the Bayesian Information Criterion (BIC). Mathematically, the autoregressive model has the following form:

$$\pi_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j \pi_{t-j} + \varepsilon_t,$$

where ε_t is an unexpected fluctuation in inflation.

4.3. Models with regularisation

A typical feature of the so-called 'overfit' model (i.e. where the model is adjusted for random patterns in the 'training' sample which are absent in the general population) is that attributes (explanatory variables in terms of the regression analysis, or predictors) can be assigned large coefficients in the resulting solution. If the number of attributes is larger than the number of time observations, or if there are correlated attributes, the problem of minimisation of the root mean squared error can have an infinite number of solutions. In such a situation, attempting to achieve a perfect approximation of noisy data can lead to a dramatic change in the coefficient values and to their abnormal increase. That is why it is reasonable to control the size of the coefficient vector, taking it into account when constructing the quality function. For this purpose, an additional term, a regulariser or penalty term based on the norm of the coefficient vector, is added to the basic function in the minimisation problem.

Mathematically, the regression with regularisation has the following form:

$$Q_R(x) = Q(x) + \alpha R(x),$$

where $Q(x)$ is a certain quality function (in this paper, RMSE), $R(x)$ is the penalty for the norm of the regression coefficient vector, and α is the hyperparameter responsible for the ratio of accuracy to the size of the parameter vector.

The hyperparameter α is determined out-of-sample (i.e. outside the sample used for the estimation of model parameters) through cross-validation: an optimal value of the hyperparameter α is a value which allows us to obtain the most accurate forecast out-of-sample, i.e. which minimises $Q(x)$ (see for example Murphy, 2012, p. 206).

There are many different regularisers, but the most common one looks like this:

$$R(x) = \gamma \sum_{i=1}^{q} |x_i| + (1-\gamma) \sum_{i=1}^{q} x_i^2,$$

where $\gamma \in [0; 1]$, q is the dimensionality of the parameter vector x.

A regulariser with γ equal to 1 is called a L_1-regulariser, and the OLS model together with this regulariser is called a LASSO model (or LASSO regression). A regulariser with γ equal to 0 is called a L_2-regulariser, and the OLS model together with this regulariser is called a ridge regression model. When $\gamma \in (0;1)$, a mixed regulariser is obtained, and the OLS model including it is called an elastic net model.

Each of the regularisation methods has its advantages and disadvantages. An important advantage of L_2-regularisation is the presence of a clear analytical solution when used together with a quality function that is differentiable in a closed form. For the root-mean-square quality function, the analytical solution looks like this:

$$x = (X^T X + \alpha I)^{-1} X^T y,$$

where X is a matrix, each column of which is a time series of observations of an attribute (predictor), I is a unity matrix of size $q \times q$, and y is a vector of responses (values of the indicator, which the model seeks to explain using a combination of attributes).

Moreover, as can be seen from the solution, the addition of L_2-regularisation ensures a positive determinant of the matrix, making it invertible. Thanks to this, the problem will always have a single solution, which is very convenient in practice.

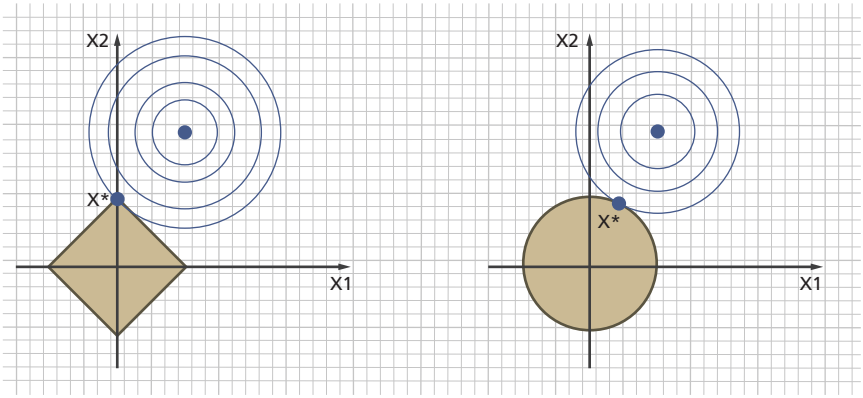
The L_1-regulariser, in contrast to L_2, does not provide a unique solution, but has another important property. Thanks to its mathematical conditionality, the use of this regulariser leads to the zeroing of some coefficients in the final solution. When minimising the convex RMSE function, the unconditional minimisation problem $Q(x) + \alpha||x||_1$ can be reduced to a conditional minimisation problem in the form:

$$\begin{cases} \hat{x} = arg \min\limits_{x} Q(x) \\ ||\hat{x}||_1 < S \end{cases}, \text{ for some } S$$

The solution of this system is the intersection of the admissible set $||\hat{x}||_1 < S$ with the line of the level nearest to the absolute minimum.

In a two-dimensional case, it will look like this:

Figure 1. Regularisation



Note: L_1-regularisation (left), L_2-regularisation (right)

Quite often, the solution of such a system will be the intersection of the admissible set with the level line at the vertex of the admissible set, which will mean the zeroing of one of the components (see the graph on the left of Fig. 1). Thus, the L_1-regulariser screens certain attributes, which allows the most important attributes of the object to be focused on. Because of this property, the LASSO model is well suited for pre-screening of variables before using other models. This paper provides an example of such a combinaton: pre-selection of parameters using the LASSO for later use in the OLS regression together with the first inflation lag. In other words, attributes pre-selected by the LASSO are incorporated into the AR(1) model.

Both regularisers have their advantages and disadvantages, so in practice a mixed regulariser known as elastic net regression is often used. In this model, the parameter γ is the hyperparameter, the value of which needs to be determined. Often it is selected using cross-validation, but this optimisation problem already has a hyperparameter (α), which is found using cross-validation. In this case, it is necessary either to allocate a part of the sample for the cross-validation, which is inadvisable due to the limited sample size of available time series, or to use the same learning sample, which can also lead to overfitting of the model. In this paper, the parameter is selected ad hoc and is defined as 0.5. This value is chosen because it is equidistant from the extreme cases (L_1- and L_2-regularisers). Moreover, this value for γ is also used in the work by Chakraborty and Joseph (2017).

4.4. Random forest (RF)

The Random Forest model is based on bootstrap union of so-called decision trees and was first proposed by Breiman (2001). A binary decision tree is used as the basic algorithm for the Random Forest. A binary tree is a graph consisting of

'parent' or 'root' nodes (interim nodes) and 'leaf' terminal nodes. A decision tree is constructed in stages. The first stage is the optimal division of the entire sample X into two sub-samples: $X_1(i,p) = \{x|x_i \leq p\}$ and $X_2(i,p) = \{x|x_i > p\}$ according to the specified quality function $Q(X, i, p)$. Next, each of the sub-samples is iteratively broken down using the same principle. The breakdown stops when a stopping criterion is fulfilled.

After that, n leaf nodes are created, each of which corresponds to a certain preserved sub-sample (which could contain only a single point). If the regression analysis problem is solved (as in the case of modelling and forecasting of inflation), each leaf node is assigned the average value of the explanatory variable across the points (observations) in the corresponding sub-sample. The resulting tree is a connected graph of root nodes, each of which contains a threshold predicate that breaks down the sub-sample into two parts, and leaf nodes, each of which contains the predicted values of the explanatory variable. Usually, the quality function is specified in the following form:

$$Q(X, i, p) = H(X) - \frac{|X_1|}{|X|} H(X_1) - \frac{|X_2|}{|X|} H(X_2),$$

where $H(X)$ is an informativeness criterion.

The informativeness criterion shows how homogeneous the objects (observations) in the sub-sample are in terms of the explanatory variable. The main idea here is to maximise this homogeneity, and to break the sample down into two parts, in each of which the spread of values of the explanatory variable is minimal. Therefore, for the regression analysis problem, the quadratic deviation is used as a loss function and the following informativeness criterion is minimised:

$$H(X) = \min_c \frac{1}{|X|} \sum (y_i - c)^2.$$

As we know, the minimum value of this kind of function is achieved when c is equal to the average value of the target (explained) variable.

$$H(X) = \frac{1}{|X|} \sum (y_i - \frac{1}{|X|} \sum y_j)^2.$$

In other words, the main goal of the algorithm at each stage is to minimise the sum of the weighted average variance within each of the two sub-samples resulting from the breakdown. Using the constructed tree, we can get predictions for values of the target variable with the new values of the explanatory variables.

The main advantage of the decision tree model is that the trees allow us to simply create effective (in terms of minimising the variance of the target variable) nonlinear dependencies. However, there is a serious drawback: the overfitting problem. For any given sample, it is possible to create a tree of such depth that

it will make no error at all. The RF model is designed to compensate for this drawback and reduce the variance of the base model. To do this, on the basis of the actual sample *X, N* artificial sub-samples of the original sample length are generated using the bootstrap. Also, an artificial sub-sample does not include all the attributes, only a random set. Randomisation therefore occurs in two directions.

Next, for each resulting artificial sample $\tilde{X}_n$, a decision tree $t_n(x)$ is built. The tree is built in such a way that in each leaf node there are at least l observations (in this paper the value l is equal to five). The final output of algorithm is the average across all constructed individual decision trees:

$$T_N(x) = \frac{1}{N}\sum_{i=1}^{N} t_i(x).$$

The number of decision trees in a RF is an important hyperparameter: the more trees there are in the forest, the more reliable the result is. At the same time, the more trees there are in the forest, the longer the operation time of the algorithm. Usually, the number of trees is chosen so that the resulting output of the RF stops changing. In this paper, 200 decision trees in each RF sufficed.

When working with time series, the drawback to using a regular bootstrap is loss of information extracted from data. This problem can be solved in two ways: by abandoning the bootstrap entirely and randomising samples only by attributes, or by using the block bootstrap (see for example Efron and Tibshirani, 1994), in which not individual points (observations) but entire blocks of a certain length (in this paper, blocks of length 10 are used) are randomly selected. Therefore, in this paper we consider two specifications of the model: without any bootstrap and with the block bootstrap.

As described in Section 3, in the current study of inflation forecasting using ML methods, data were made stationary. In traditional econometric models, this is done to avoid the so-called 'spurious' regression problem. However, the results of many ML algorithms, including the RF model and Boosting, are not subject to this problem. This paper therefore presents the results of the RF model using both data in a stationary form and data which have not been transformed. It is important to note that, in all specifications, the RF models were trained and used to obtain inflation forecasts (stationary series), but not forecasts of accumulated CPI values (non-stationary series). Accordingly, all RMSE calculations and their comparisons are made in the same way for all models.

4.5. Boosting

The gradient Boosting model was first proposed by Friedman (2000). The idea behind the gradient boosting algorithm is similar to the idea behind the RF model: both algorithms are ensemble methods. The base Boosting model can

represent any collection of models, but often, as with the RF model, a decision tree is selected. The main difference between the Boosting and the RF is that the base models are not trained independently, but rather taking into account the results of operation of the model on the previous iterations. The algorithm's operation can be described as follows:

1) The first base model is trained on the whole sample:

$$b_1(x) = arg \min_b \sum_{i=1}^{l} (b(x_i) - y_i)^2.$$

2) After the first step, the ensemble Boosting algorithm results in the first trained base model:

$$B_1(x) = b_1(x).$$

3) Next, residuals are calculated that are equal to the difference between a true value and a predicted value based on the first Boosting model:

$$e_i^1 = y_i - B_1(x_i).$$

4) The following model is trained on these residuals:

$$b_2(x) = arg \min_b \sum_{i=1}^{l} (b(x_i) - e_i^1)^2.$$

5) We add a new model to the algorithm obtained in the previous step with a certain coefficient $\gamma \in (0; 1]$. This technique is called 'step reduction'. In this paper, the coefficient γ is equal to 0.2. This technique helps to improve the model's operation and avoid overfitting. A new model is obtained:

$$B_2(x) = B_1(x) + \gamma b_2(x).$$

6) Then, the algorithm is built up iterativelyuntil the end. As a result of operation of the algorithm, the final model is:

$$B_N(x) = \sum_{i=1}^{N} \gamma^{i-1} b_i(x).$$

The algorithm terminates once all training cycles are completed. The number of training cycles is an important hyperparameter. Usually, in the training sample, errors tend to vanish as the number of iterations is increased. However, out of sample, too many iterations can actually increase errors as that the model begins to adjust for noise. In this paper, the number of iterations is 100.

As already mentioned above, both untransformed (non-stationary) and transformed (stationary) data are used for the RF and Boosting models.

5. Results

5.1. Overall results

The main results of this study are presented in Tables 2–5 of the Appendix. The Tables show the relative RMSE values of all specifications of each model for all forecast horizons (Table 2 – from the 1st to the 6th month, Table 3 – from the 7th to the 12th month, Table 4 – from the 13th to 18th month, Table 5 – from the 19th to the 24th month). Also, Table 6 of the Appendix gives the relative RMSE values for the average inflation over the horizon of 1, 2, 3, 6, 12, 18, and 24 months. We can draw the following main conclusions from the results:

1) The use of ML methods can improve the quality of forecasting of Russian inflation compared to reference models (benchmarks) that use only lags of inflation as predictors. However, significant disadvantage of these models in comparison with classical econometric models is the loss of interpretability in the classical sense.
2) The ensemble methods (RF and Boosting) predict average inflation better than the base model from the second month onwards.
3) Among all three specifications of the RF model, the specification with untransformed data gave the best result when forecasting both inflation in individual months and average inflation over entire forecast horizon. Comparing the results of the two specifications of the Boosting model, with stationary and non-stationary data, leads to the same conclusion.
4) Relatively speaking, the regularised models provide less accurate forecasts over all forecasting horizons. The AR(1) model combined with LASSO gave results worse than the base AR(1) model, except for forecasts one month ahead.
5) The AR(1) model combined with LASSO gave the highest quality results when forecasting inflation over the horizon of one month. The models that use only lags of inflation as predictors (RW and AR) gave the same quality of forecasts over this horizon. The remaining methods were less accurate with respect to forecasts than the base model.

A more detailed analysis of the results is provided below.

5.2. Results of using RF and Boosting models

As mentioned above, the ensemble models that use untransformed data gave better results than similar models that use transformed data. This is not surprising, since the application of the RF and Boosting models does not require prior transformation of data to a stationary form. In addition, it is worth noting that the specification of the RF model that uses the block bootstrap gave slightly worse results than the model without the block bootstrap when forecasting both inflation for specific months and average inflation over entire forecast horizon. This result probably indicates that data within time series have a high degree of serial correlation and that their random mixing with replacements negatively

affects the result of the model's operation, even when a special block bootstrap is applied.

The RF model, which uses untransformed data for forecasting inflation for individual months, performed better than the base model. With the exception of the inflation forecast one month ahead, the RF model generally predicts inflation for individual months more accurately. At the same time, the RF model predicts average inflation much better: when forecasting average inflation two years ahead, the error is 60% smaller than the error obtained using the base AR (1) model.

To understand this significant difference, let us compare monthly and accumulated inflation values over the same horizon, as predicted by both models. Figure 3 of the Appendix shows the dynamics of monthly and accumulated inflation (actual and forecast) over the forecast horizon of 12 months (the justification for taking this horizon being that RMSEs of monthly inflation forecasts 12 months ahead are almost identical: the value of the relevant RMSE is equal to 0.9827). The chart clearly shows a sharp short-term surge of inflation, associated, for the most part, with the fall of the rouble at the end of 2014. The AR(1) model represents this surge with some lag when forecasting each subsequent monthly figure for inflation over the entire horizon, as a result of which, around a year later, high average accumulated inflation appears, which is very different from the actual case. Here, the RF model, which is not trained on the inflation lag, shows higher resistance to such shocks. It is this very shock resistance which explains the differences between the model's predictions for average inflation.

The Boosting model using untransformed data is similar to the corresponding RF model. Figure 4 of the Appendix shows changes in inflation over 13 months plus the average cumulative inflation (since the relative RMSE value at 13 months is also approximately equal to 1) for the Boosting model and the AR(1) model. The graph shows that the Boosting model forecast is also more resistant to shocks than the AR(1) model forecast. The explanation given above concerning the RF model also goes, more or less, for the difference. It is worth noting that the Boosting model reacts to shock more sensitively than the RF model. This is related to the fact that the Boosting model learns from past forecast errors, which makes it more adaptable to a specific training sample.

The RF and Boosting model algorithms have several common features: both algorithms are ensemble methods and are based on the decision tree model. Therefore, it is logical that the results of their application should be similar. This paper draws the following conclusion: we observe a similarity in forecasting using separate features of the RF model and the Boosting model. Such a correspondence may indicate the reliability of the results obtained.

5.3. Results for regularisation models

All regularisation models showed lower prediction accuracy than the base model. This was primarily because these models are, to some extent, unstable.

Figures 5–10 of the Appendix show a change in the number of explanatory variables left by the LASSO model when forecasting inflation 1, 2, 6, 12, 18 and 24 months ahead. The number of explanatory variables picked at each forecast horizon is quite volatile. We see that bursts occur at certain moments when the model selects an abnormally large number of explanatory variables. On the other hand, from the forecast horizon of one month onward, for some dates, all variables are cut off. At the same time, the average number of predictors remaining in the model plummets as the forecast horizon changes from one to two months. This instability may indicate that, among the initial set of variables, there are no explanatory variables that either individually or as part of a small group of variables could effectively predict the dynamics of inflation.

We consider that the combined LASSO and AR(1) model confirms this hypothesis. The LASSO enables preliminary selection of variables with subsequent addition of the first inflation lag. With regard to regressions, the coefficient vector is of particular interest. For each attribute, the average absolute value of the coefficient vector was calculated (for many variables, this value was 0, since the characteristic could have been eliminated by the LASSO before) when forecasting inflation for each forecast horizon. In other words, for example, a coefficient matrix was used for developing 53 one-month-ahead forecast options. Some of the values in this matrix are equal to 0 because the LASSO model had already eliminated some of the variables. This is followed by a review of the absolute values of non-zero coefficients. Then, the average absolute value of the coefficient is calculated for each explanatory variable. As a result of data standardization, the coefficient values show the relative strength with which the predictors account for future inflation. This operation is performed for each forecast horizon separately.

Table 7 of the Appendix presents, in descending order, the five largest average absolute values of the coefficient for each forecast horizon of 1, 2, 3, 6, 12, 18, and 24 months. The table does not present the results for all months, since these seven periods are sufficient to identify the general pattern. We see that for all forecasting months, on average, it is the first lag of inflation that remains the main explanatory factor; the coefficients of all the other variables turn out to be significantly less. At all forecasting horizons from the sixth month onward, the values for the indicators obtained are significantly less than the coefficient of the first inflation lag. It should be noted that expansion of the forecast horizon gradually identifies indicators that do in fact account for inflation. Thus, when inflation is forecast for the 24th month ahead, two explanatory variables survive: the first inflation lag and loans to individuals over one year, while all other variables are almost completely zeroed out.

The quality of the Ridge model is similar to that of the LASSO model, even slightly worse. This makes sense: overfitting may be a problem for the LASSO model, but it is even more of an issue for the Ridge model. It is worth recalling here that, due to the peculiarities of the regulariser, the LASSO model has a strong

ability to cut off some of the variables in the regression, something the Ridge model is not capable of. Therefore, the forecast quality of the Ridge model is even lower than that of the LASSO model. The results of the elastic net model are almost identical to the results of the LASSO.

6. Conclusion

This paper aimed to prove the viability of ML methods for forecasting Russian inflation, compared to traditional methods. As the results demonstrate, this conjecture has been confirmed. Not all methods performed equally well in solving this problem: the regularisation models showed lower forecasting quality compared to the base model.

Both ensemble methods (RF and Boosting) showed results comparable to the basic AR(1) model in predicting monthly inflation. At the same time, they showed significantly better results when forecasting average inflation over a horizon of more than two months. We can therefore conclude that the RF and Boosting models show promise when applied to the task of forecasting Russian inflation.

This paper also addressed the issue of data transformation. According to our results, the RF and Boosting models perform better with untransformed rather than transformed data. This conclusion could be of use in further research using ML methods, since data transformation is a standard preparatory element of almost any empirical macroeconomic research using time series.

In addition to the models used in this article, there exist a number of other nonlinear ML algorithms, such as neural networks. Researchers would be well advised to also test these algorithms in future work on forecasting Russian inflation.

Appendix is available at
www.cbr.ru/eng/money-and-finance;
dx.doi.org/10.31477/rjmf.201804.42

7. References

Andreyev, A. (2016). *Integrated Inflation Forecasting at the Bank of Russia.* Series of Economic Research Reports, 14.

Atkeson, A., and Ohanian L. E. (2001). Are Phillips Curves Useful for Forecasting Inflation? *Federal Reserve Bank of Minneapolis Quarterly Review*, 25(1), pp. 2–11.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), pp. 5–32.

Chakraborty, C. and Joseph, A. (2017). *Machine Learning at Central Banks.* Bank of England Working Papers, N 674.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap.* CRC press.

Faust, J. and Wright, J. H. (2013). Forecasting Inflation. *Handbook in Economic Forecasting*, 2(1), pp. 2–56.

Friedman, J. H. (2000). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5), pp. 1189–1232.

Garcia, M. G., Medeiros, M. C. and Vasconcelos, G. F. (2017). Real-Time Inflation Forecasting with High-Dimensional Models: The Case of Brazil. *International Journal of Forecasting*, 33(3), pp. 679–693.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective.* MIT Press

Romer, D. (2012). *Advanced Macroeconomics*, 4th ed. McGraw-Hill Irwin

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM computing surveys* (CSUR), 34(1), pp. 1–47.

Simonyan, K. and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv*:1409.1556.

Stock, J. H. and Watson, M. W. (2007). Why Has U.S. Inflation Become Harder to Forecast? *Journal of Money, Credit and Banking*, 39(1), pp. 3–33.

Stock, J. H. and Watson, M. W. (2008). *Phillips Curve Inflation Forecasts.* NBER Working Paper, N 14322.

Stock, J. H. and Watson, M. W. (2011). Dynamic Factor Models. In: M. P. Clements and D. F. Hendry, eds. *The Oxford Handbook of Economic Forecasting.* Oxford University Press.