ENSAE PARIS

Macroeconometrics and Machine learning
Report

January 21, 2024

Random Forest for Inflation Forecasting Understanding American inflation, its predictability and its fundamentals

Students :
Victor Francey
Rémi Hurstel

Professor:
Anna Simoni

Contents

1 Introduction

Following the unexpected and consequent increase in the level of prices in 2021 and beginning of 2022, the problematic of being able to forecast inflation and to understand the fundamentals of this economic phenomenon to better understand what drives its variability has become more prominent. Indeed, on a theoretical point of view, a significant number of variables look credible to explain at least partly inflation movements and surges. This is why we have decided to propose an exploration of variables of interest to better understand inflation using machine learning tools, and more precise random forests. While bearing interest to the United States instead of Sweden in our report, we found inspiration in the work realized by Malte Meuller in 2022, named *Inflation forecasting with Random Forest A Machine Learning approach to macroeconomic forecasting* [1].

Despite all its limitations, this method has the significant advantage to allow us to consider a lot of potential explanatory variables at the same time and hence to incorporate different types of data. To present briefly the dataset we use, we consider variables of interest from 1991 to nowadays, an interesting time frame as inflation has observed different upwards and downwards cycles during this period. Furthermore, this for instance allows us to train the model on some specific periods, use it to predict on other periods and then to observe the fit of such a method to consider the fact that inflation fundamentals have changed or not during this period, we will come back to this point later in this report. We use very different types of data, we use monetary data such as M1, we use a wide set of economic indicators such as consumer confidence or industrial production and we consider international factors such as trade, foreign inflation. The full set of considered variables is available in the appendix.

The report will follow this specific guideline. Firstly, we will show the quality of the goodness of fit of the model. Then, we will analyse the relative significance and importance of the different variables in order to give some economic meaning and substance to this statistical procedure. Later on, we will present the results we obtained by sequentially training the data on a specific period and use to predict inflation on the last part of the time frame in order to test the predictability of inflation, in particular during periods with significant shocks such as during the pandemic. We will also propose alternative ways to train the data that may lead to better outcomes once using the training to forecast on periods where we did not train the data on.

2 Goodness of fit and economic analysis

2.1 Data

All data we use come from the FRED website. The inflation data we choose as forecast target is the Consumer Price Index (CPI) in the U.S. We choose 30 inputs for our model to predict the CPI. The list of the data we used is described in the Table 1 in the appendix.
To make each of the 31 time series stationary, we compute the annual change for each of these series:

$$\tilde{x_t} = \frac{x_t - x_{t-12}}{x_{t-12}}$$

The data we used as inputs for the Random Forest spans from January 1993 to December 2023 (372 monthly observations).

[1] https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=9073925&fileOId=9073927

2.2 Random Forest: Fit and Forecast

Before fitting the model, we concatenate input data from lag -1 to lag -12. This results in 360 columns ($30 * 12 = 360$) and 360 observations ($372 - 12 = 360$) in the input dataset for the Random Forest. The Random Forest we use has 200 trees ($ntree = 200$) and uses 10 predictors in each split ($mtry = 10$). This Figure 1 shows that overall, we get a goof fit which is not a

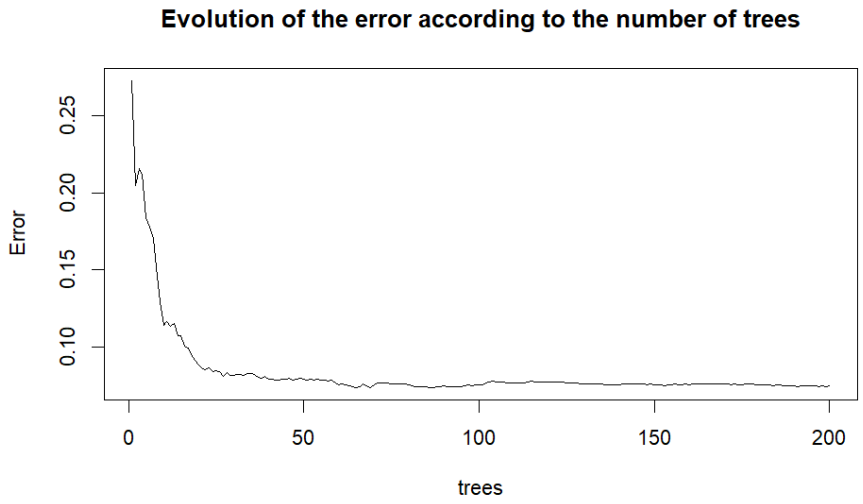

Figure 1

significant result in itself. Indeed, we took a large set of variables to build our dataset which comprises among other different indicators of inflation itself, a weak fit of data would then have been a surprise and would have been difficult to understand.

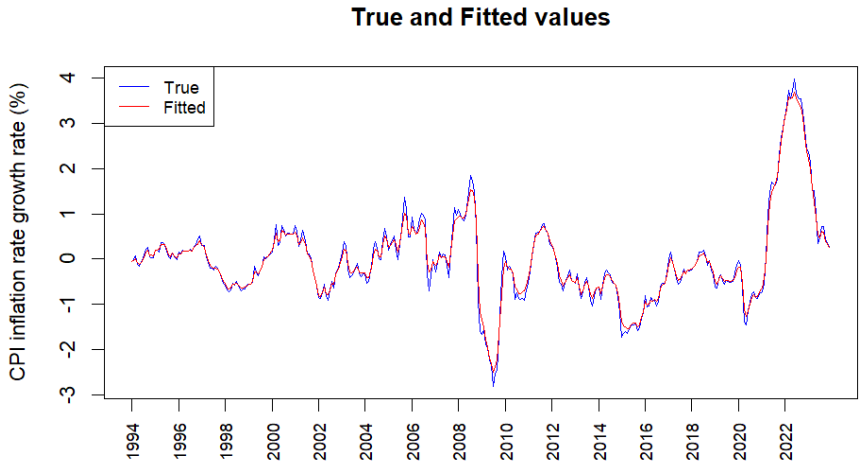

Figure 2

The Figure 2 is just another way to visualize and show the previous result which is that we get a good fit overall with our set of variables. One can observe that the main errors of prediction come from a difficulty to predict the maxima reached by inflation during its periods with high volatility.

2.3 Importance of the variables

To analyse the influence of each variable on inflation, we choose to observe the mean importance of each of the 30 variables in our random forest model.

Without any surprise once again, one can observe that a set of lagging inflation indicators

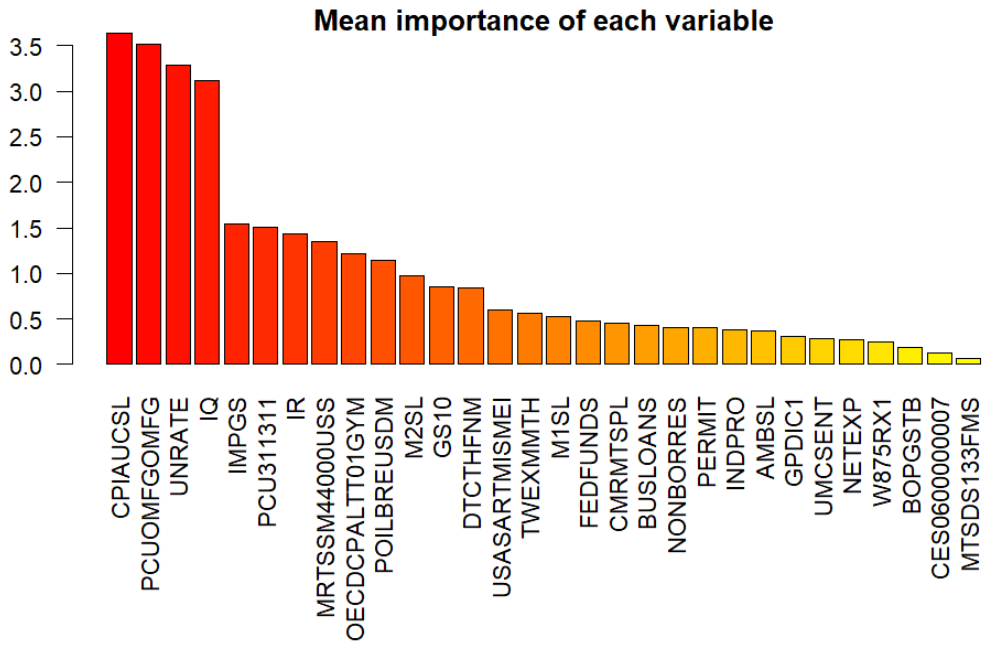

Figure 3

emerge as the main significant predictors of future inflation (Figure 3). What is more interesting is to look at non-inflation indicators emerging as variables with a powerful forecasting power. This set of indicators is mainly comprised by the unemployment rate (UNRATE), revenue index (IR), retail trade and brent price. What is interesting is that the first three of these variables can be interpreted as demand-side variables. Indeed, the situation of the worker, its professional situation (employed or unemployed), its revenues and its willingness to consume (retail trades) appears as a major economic indicator of forthcoming inflation level.

3 Inflation predictability: what can we learn by training the model in different ways?

In this new section, we perform a new method by training the random forest model only on a part of the data set in order to observe its performance on test data.

3.1 Training on the past to forecast future

The first way we split our data set in train and test data is to put the data of the period starting in January 1994 and ending in December 2013 (240 observations) into the train data set. So, the test data set includes the data from January 2014 to December 2023. As one can observe, if trained only from 1994 to 2013, the model will be much less able to forecast inflation on the test period (2014-2023) (Figure 4). Moreover, when we compute the MSE (mean square error), on the training and test data in order to compare the accuracy of the forecast, we get a MSE of 0.01 on training data but the MSE is equal to 1.41 on test data. It means that we have a real drop in prediction accuracy between training and test data. This is a significant matter of

concern that we would not be able to forecast the high variability of inflation.

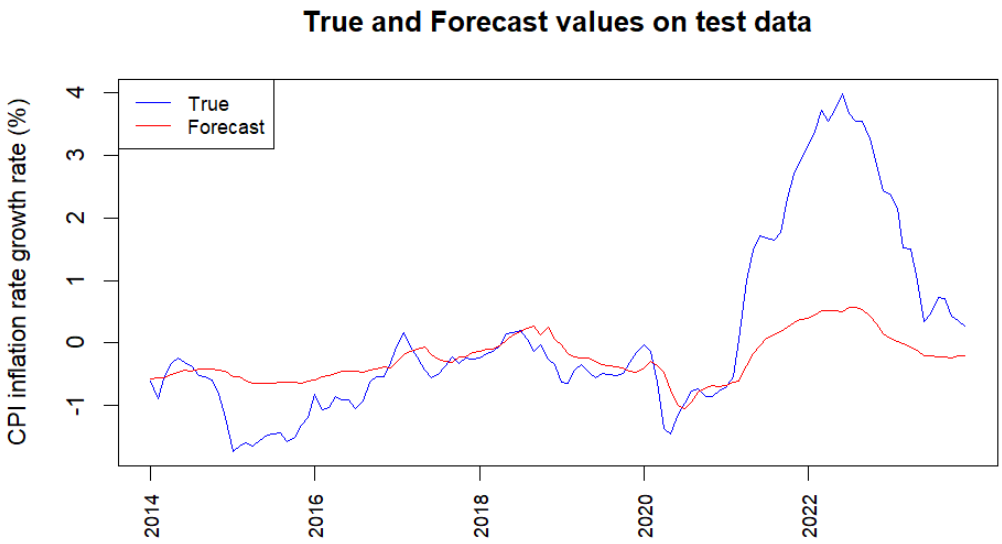

Figure 4

Furthermore, it may indicate that variables leading to inflation volatility may differ between periods and may cause the failure of this procedure. If variables that matter the most between 1994 and 2014 and from 2014 are different, this means that catching the wide set of variables that matter to understand inflation may be really complicated and that inflation has many potential sources.

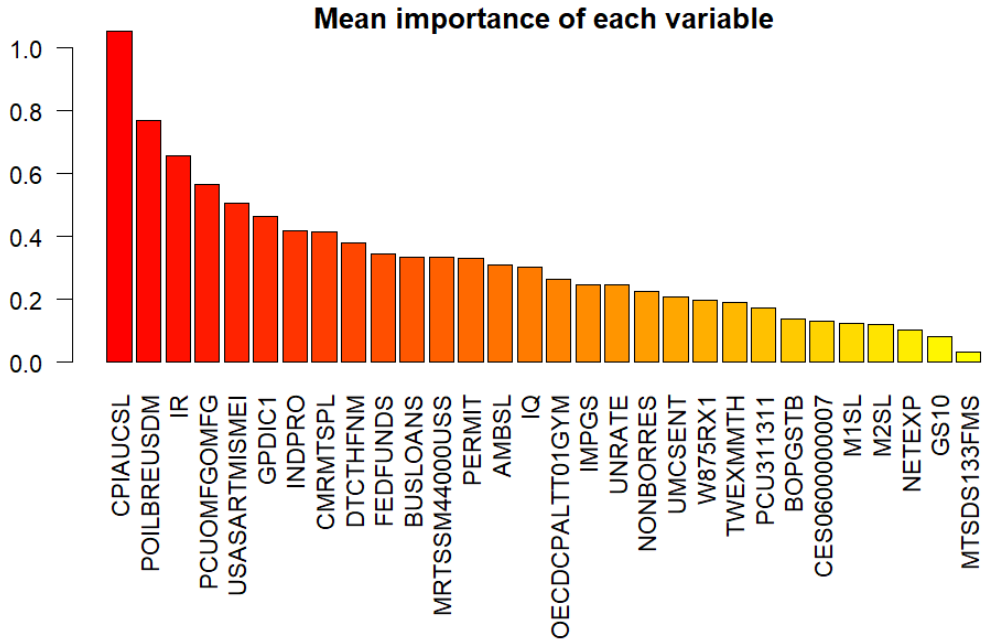

Figure 5

This figure confirms our intuition that the main drivers of inflation are not always the same at

any period and that some specific inflationary periods may have specific origins and explanations. One can remark it by comparing this graph to the previous one and remarking that variables do not have the same rank and the same mean importance, which means that one has to be aware of this when choosing the period over which he trains its random forest model. A bad choice may lead to non relevant forecasts. For instance, oil price plays a much more significant role in this example than in the previous one. At the opposite, some variables such as IQ (export prices) appear to play a less significant role.

3.2 Training with random observations

We lastly propose a last way to consider the fact that inflation fundamentals are too different through to be able to forecast inflation at all. What we propose is to still train the model on two thirds of the available data (240 of our 360 observations) and to use it to forecast inflation on the remaining 120 periods. The difference with the previous model is that we randomly select the observations we will use as train data. One can observe that this time, the model is really good

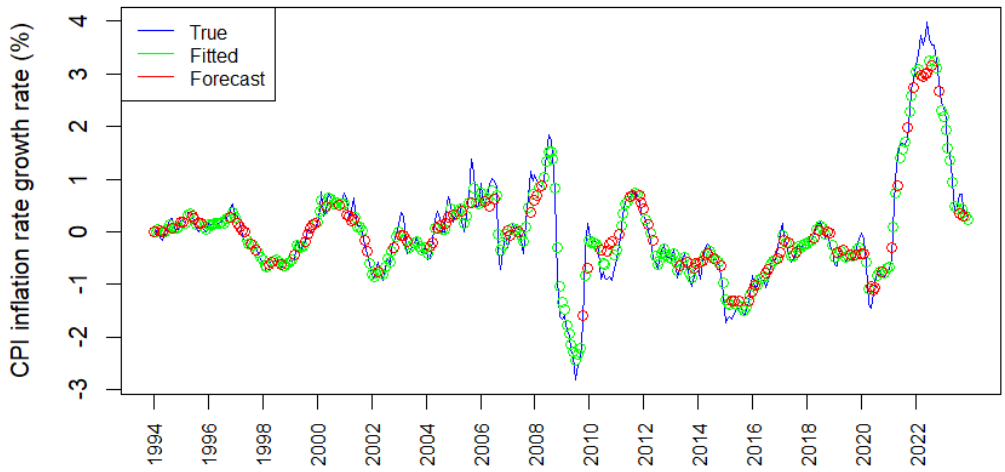

Figure 6

at predicting inflation and offers a really decent fit (figure 6). Indeed, this time the difference of the MSE between training and test data is not so high: on train, we get a MSE of 0.020 and on test a MSE of 0.081. It means that when we randomly choose the observations that we will use to train our random forest, it will not overfit contrary at the the case where we train only with the past observations.

The main takeaway is probably that indeed, there are many potential sources to inflation surges and that being able to train the model on these different periods is necessary in order to build a credible and useful model. In addition, this shall reassure us about the ability to forecast a model when there is no "crisis". This was already the case in the previous situation but the model never creates "fake" inflation that does not appear in the data, which would be a significant source of concern.

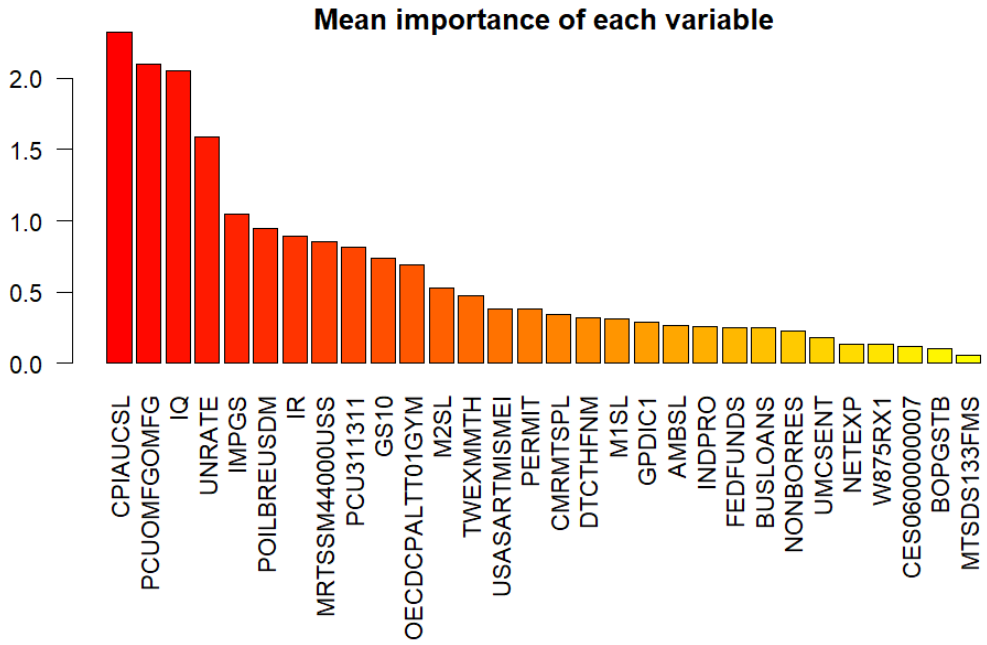

Figure 7

This last graph simply confirms again that the training period is crucial and that this choice has huge impacts and consequences on the forecasts which will be made out of the model. One can hope that by randomizing the choice of the periods, all inflationary periods will be partially included and then the different potential sources of inflation will be considered by the model leading to less risks of huge mistakes of prediction. One can actually remark that both unemployment rate and oil prices play here a significant role. The basic model probably underestimated the oil significance as it does not always play a role but can have a huge effect in some specific crisis, i.e. the recent and ongoing war in Ukraine. Similarly, the unemployment rate which did not appear as having a high explanatory power with the previous graph when training on 1994-2014 is again taken into consideration (is ranked 4) such as in the basic model when we trained on the whole period.

4 Conclusion

This work should be seen as an introduction to inflation forecasting. It confirms that inflation has persistence and that lagging inflation measures will be of great importance when trying to forecast forthcoming inflation. Furthermore, we can confirm that inflation has many potential sources and can appear through very different channels. Hence, particular attention shall be brought to the choice of the data and the dates on which the model is trained. For instance, taking into account different periods with high inflation is crucial to give the relevant level of significance to some specific variables that can play a significant role at some points and be almost useless at some other moments, leading to the risk of either ignoring them either according them too much importance. The last method that we presented, that is less data intensive, is worth being looked at by allowing to mix periods of interest without loosing a lot in goodness of fit. Finally, further interest could be brought to the choice of periods of interests, for instance by looking at the quality of forecasts either by training the model only on high inflation periods either by ignoring these periods and analyzing the difference.

5 Appendix

fred code	Description
CPIAUCSL	Consumer Price Index for All Urban Consumers: All Items in U.S. City Average
UNRATE	Unemployment Rate
FEDFUNDS	Federal Funds Effective Rate
M1SL	M1
MRTSSM44000USS	Retail Sales: Retail Trade
UMCSENT	University of Michigan: Consumer Sentiment
PCUOMFGOMFG	Producer Price Index by Industry: Total Manufacturing Industries
W875RX1	Real personal income excluding current transfer receipts
M2SL	M2
DTCTHFNM	Total Consumer Loans and Leases Owned and Securitized by Finance Companies, Level
AMBSL	St. Louis Adjusted Monetary Base (DISCONTINUED)
BUSLOANS	Commercial and Industrial Loans, All Commercial Banks
PCU311311	Producer Price Index by Industry: Food Manufacturing
BOPGSTB	Trade Balance: Goods and Services, Balance of Payments Basis
TWEXMMTH	Nominal Major Currencies U.S. Dollar Index (Goods Only) (DISCONTINUED)
USASARTMISMEI	Sales: Retail Trade: Total Retail Trade: Volume for United States
IR	Import Price Index (End Use): All Commodities
IQ	Export Price Index (End Use): All Commodities
IMPGS	Imports of Goods and Services
NETEXP	Net Exports of Goods and Services
OECDCPALTT01GYM	Consumer Price Index: All Items: Total
CMRMTSPL	Real Manufacturing and Trade Industries Sales
INDPRO	Industrial Production: Total Index
PERMIT	New Privately-Owned Housing Units Authorized in Permit-Issuing Places: Total Units
GPDIC1	Real Gross Private Domestic Investment
MTSDS133FMS	Federal Surplus or Deficit
POILBREUSDM	Global price of Brent Crude
GS10	Market Yield on U.S. Treasury Securities at 10-Year Constant Maturity, Quoted on an Investment Basis
NONBORRES	Reserves of Depository Institutions, Nonborrowed
CES0600000007	Average Weekly Hours of Production and Nonsupervisory Employees, Goods-Producing

Table 1: List of the data

Model	Training	Test
Training on past	0.011	1.414
Training on random observations	0.020	0.081

Table 2: MSE on training and test data according to model

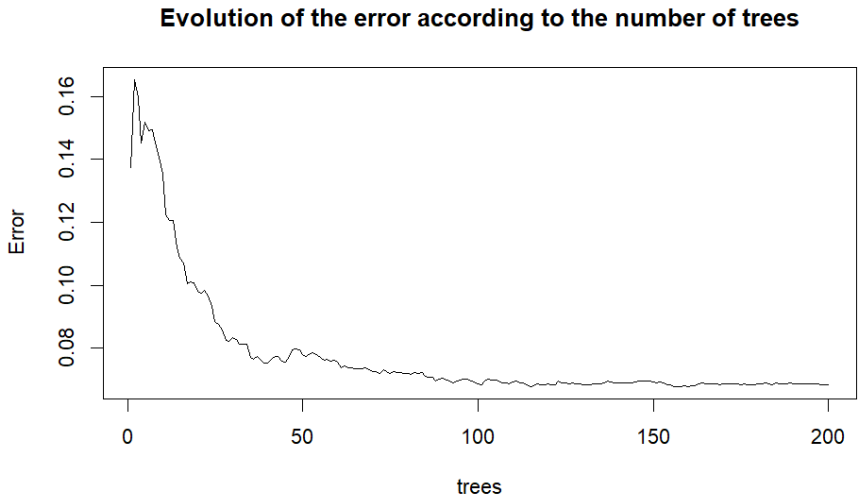

Figure 8: Error: model trained only with past observations

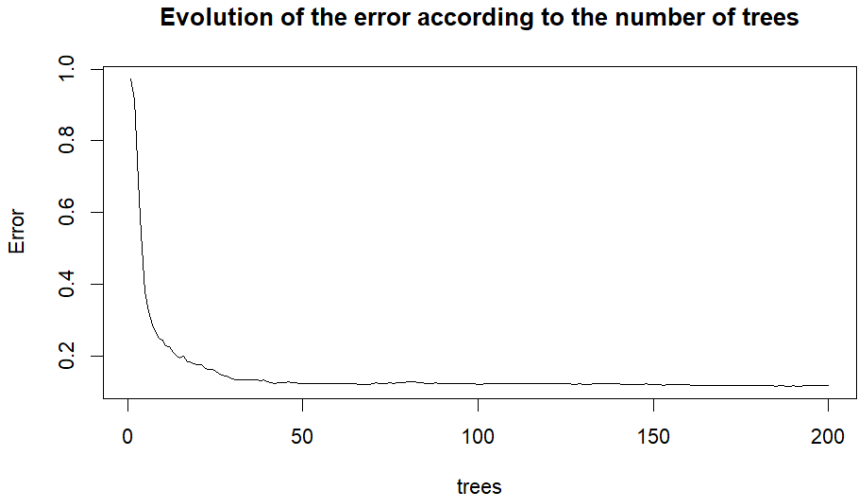

Figure 9: Error: model trained with random observations