technische universität
dortmund

Case Studies

Summer 2022

Project II
One-Quarter-Ahead Forecasts of US GDP Growth
A Machine Learning Approach

Jorge Eduardo Durán Vásquez
June 9, 2022

in collaboration with
Christopher Gerlach, Felix Fikowski, Sven Pappert

Lecturers: Prof. Dr. Matei Demetrescu, M.Sc. Karsten Reichold
M.Sc. Alexander Gerharz

Contents

1 Introduction 1

2 Data 2

3 Methods 2

3.1 Regression Trees 2

3.1.1 Regression Trees model 2

3.1.2 Pruning Regression Trees 3

3.1.3 Prediction in Regression Trees 5

3.2 Random Forest 5

3.2.1 Bootstrap Sampling 5

3.2.2 Bagging (Bootstrap aggregation) 5

3.2.3 Random Forest Algorithm 6

3.2.4 Prediction in Random Forest 6

3.2.5 Variable importance in Random Forest 7

3.3 Root mean square forecasting error 7

3.4 Software tools 7

4 Results 7

4.1 Analysis of the Data Set 7

4.2 Regression Tree of GDP growth 8

4.3 Regression Tree of GDP growth using all variables 9

4.4 Regression Trees one-quarter-ahead forecasts of GDP growth 10

4.5 Variable importance Random Forest 11

4.6 Random Forest one-quarter-ahead forecasts of GDP growth 12

4.7 Model Comparison 13

5 Conclusion 14

References 15

Appendix 16

A Additional figures 16

1 Introduction

Governments, business planers and investors take the real Gross Domestic Product (GDP) growth as an important macroeconomic indicator to assess the performance of a country's economy. The GDP is one of the most important indicators that reflects societies' material living standards (OECD, 2009) and is often used for defining strategies and monetary policies. For example, policymakers use historical macroeconomic indicators, among others GDP, to forecast and evaluate possible expansions or recessions in the economy (Kitchen J. Monaco R., 2003). These predictions allow governments and companies to develop strategies that respond to the corresponding economic situation.

The purpose of this report is to fit various Regression Tree (RT) models, as well as Random Forest (RF) models to forecast the US real GDP growth. The models use quarterly times series that are provided by the Federal Reserve Economic Data starting from the first quarter of 1959. We use time series of the unemployment rate, manufacturing industry capacity, inflation, Federal Funds rate, as well as stock market indexes and measures of liquidity in the economy (M1 real). We start fitting a RT model with lagged information of the GDP growth. Then, lags of the six additional macroeconomic indicators are added to create a new RT model to forecast GDP growth. In the same way, RF models are implemented, one using only lagged information of the GDP and another using lagged information of all variables. For all models we make one-quarter-ahead predictions with all information available until the forecasted value. After applying the methods outlined above and comparing the forecasting errors with the ones found for the autoregressive model AR(1) and vector autoregressive models VAR(1) and VAR(3) in previous work, we conclude that the model with the lowest forecasting error is the RF that uses lagged information of all variables. During the studied time period the RF model gives reasonable predictions of expansions and recessions of the economy, however it also fails to predict extreme observations, which diminishes its performance. For this model, using the permutation variable importance method we found that lag 7 and 8 of `FEDFUNDS` followed by lag 9 of `CUMFNS` and lag 2 of `UMRATESTx` are the variables that improve the forecasting error of RF.

This report is structured as follows. In section 2, the data set is described. Section 3 explains the used software tools and statistical methods, which include definition, algorithms and prediction in RTs and RF. In section 4, the results of the analysis are presented. Finally, section 5 concludes the report, discusses the current constraints of this analysis, and provides recommendations for future research.

2 Data

From a data set that is provided by Federal Reserve Bank of St. Louis in the US this report uses the time series `GDP growth`, unemployment rate `UNRATESTx`, manufacturing industry capacity `CUMFNS`, `inflation`, Federal Funds rate `FEDFUNDS`, as well as the stock market index `S&P 500 growth`, and the measure of liquidity in the economy `M1REAL growth`. To find detailed information about each of the variables, we invite the reader to take a look at the first report (Duran, J. 2022).

3 Methods

3.1 Regression Trees

The theory on Regression Trees (RTs) that will be explained in this report relies on the book of Hastie (2009), otherwise it will be specified. Linear models often fails predicting non-linear effects. In contrast, Tree-based methods can be used to identify and characterize these non-linear effects. These methods split the feature space into a set of hyper-rectangles, and then fit a simple model (like a constant) in each of them. We will describe one of the most popular methods for tree-based regression and classification called CART.

A binary tree structure starts with a so called Root Node. Then from the Root Node two branches, also called "edges", relate one node to another node (inner/daughter nodes). Edges represent the solution of an attribute test based on the parent node. Each inner node, including the root node, contain an attribute test. A leaf, which is the node without binary split is located at the end of the branch, represents a return value. To interpret the tree, one starts traversing the tree from the root node, makes the corresponding attribute tests in each inner node, and following the direction that the edges specify, you reach the leaf node, where the prediction is located.

3.1.1 Regression Trees model

In RTs we have a continuous response Y and a vector of inputs X with size 1xp. Each observation is represented as (x_i, y_i) for $i = 1, 2, ..., N$, where N is the total number of observations, $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$ and p the number of attributes. The model in RTs makes M partitions of the sample space and assigns a constant c_m to each partition region (R_m), the equation of the model is given by:

$$f_{RT}(x_i) = \sum_{m=1}^{M} c_m I(x_i \in R_m) \tag{3.1}$$

Where the indicator function $I(x)$ gets the value of one when the observation x_i belongs to the partition region R_m. When the criterion of minimization of the equation (3.1) is the sum of squares errors, the estimator for the constant c_m is the average of y_i in the respective region R_m:

$$\hat{c}_m = ave(y_i|x_i \in R_m)$$

To determine the partition regions R_m, literature suggests to take all the data and for each variable x_{ij} in X determine the following half-planes:

$$R_1(j,s) = \{X|x_{ij} \leq s\} \text{ and } R_2(j,s) = \{X|x_{ij} > s\}$$

Where s represents the splitting point in the middle of two observations x_ij for the attribute j. To find the set of splitting points S for a given attribute j, the RT sorts the observations of that attribute x_{ij} and calculates the middle point for each pair of observations. The splitting point is found solving the following minimization problem for the splitting variable j and splitting point s:

$$\min_{j,s}\Big[\min_{c_1} \sum_{x_{ij} \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_{ij} \in R_2(j,s)} (y_i - c_2)^2\Big] \tag{3.2}$$

For any choice j and s, the inner minimization of equation (3.2) is solved by:

$$\hat{c}_1 = ave(y_i|x_{ij} \in R_1(j,s)) \text{ and } \hat{c}_2 = ave(y_i|x_{ij} \in R_2(j,s))$$

To determine the best splitting variable x_{ij} and splitting point s scanning through all the tuples (x_ij, s) is necessary. For each variable and splitting point in S a computation of the inner minimization using equation (3.2) is necessary. When the best split (x_ij, s) is found, we partition the data in two resulting regions and then repeat recursively the splitting process on each of the two regions. The procedure stops when the number of observations in a leaf node is between the parameters `minbucket` and n_{min}.

3.1.2 Pruning Regression Trees

The theory that will be presented in this section is based on Breiman L. et al. (2000). When the previous procedure of growing a RT is executed, a next step is necessary to avoid overfitting. It is the so-called pruning, which consists of evaluating top down the accuracy of each split in the tree. If the accuracy of the overall model is not improved by a given threshold, the respective branch is pruned and the analysis continues to the next branch.

To understand the concept of pruning, first denote the sub-tree T as a part of a bigger tree T_0. Now, the accuracy of a tree $R^2(T)$ is calculated using the following formula:

$$R^2(T) = 1 - RE(T)$$

Where $RE(T)$ corresponds to the relative mean square error of T, calculated as follows:

$$RE(T) = R(T)/R(\bar{y})$$

The term $R(T)$ corresponds to the mean square error calculated with:

$$R(T) = \frac{1}{N}\sum_{i=1}^{N}(y_i – \hat{f}_{RT}^{(T)}(x_i))^2$$

Where $\hat{f}_{RT}^{(T)}(x_i)$ is the predicted value of the regression tree T at time i. The value of $R(\bar{y})$ corresponds to the relative error of the baseline predictor assuming that nothing is known about the explanatory variables x_{ip}. This is the mean square error calculated with the mean of the response values y as a predictor. It is given by:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \text{ and } R(\bar{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i – \bar{y})^2$$

Knowing the previous concepts, we now state the condition that needs to be fulfilled in order to accept a new split in the RT, it is given by the following inequality:

$$(1 + \mathbf{cp})R^2(T) \leq R^2(T_0)$$

This can be interpreted as a condition that states that the tree T_0 must be at least $(1+cp)$ times more accurate than the old tree T.

To find an estimation of $R^2(T)$ the literature suggests to use v-fold cross-validation. The estimation is represented by $R^2_{CV}(T)$. To calculate the estimation, we denote the set of all observations (x_i, y_i) as learning sample L. Then the v-fold cross-validation estimate $R^2_{CV}(T)$ is calculated, first, by dividing the set L in equal size V subsets $L_1, \ldots, L_V$. Second, for each $v = 1, \ldots, V$ calculate $L – L_v$ and apply the same growing and pruning procedure to obtain the predictor tree $\hat{f}_{RT}^{T_v}$. Then, the cross-validated estimate is calculated as:

$$R_{CV}(T) = \frac{1}{N}\sum_{v}\sum_{(x_i,y_i)\in L_v}(y_i – \hat{f}_{RT}^{T_v}(x_i))^2$$

And the cross-validated relative mean square error and accuracy are given by:

$$RE_{CV}(T) = R_{CV}(T)/R(\bar{y}) \text{ and } R^2_{CV}(T) = 1 - RE_{CV}(T)$$

3.1.3 Prediction in Regression Trees

The expression used to calculate the prediction of RTs is given by:

$$\widehat{f}_{RT}(x_i) = \sum_{m=1}^{M} \widehat{c_m} I(x_i \in R_m) \tag{3.3}$$

The algorithm of RTs outlined above uses the following parameters that are suggested in the literature. First, a `minsplit` (n_{min}) of 20, this corresponds to the minimum number of observations in a node that are needed to execute a split. Second, `minbucket` is the upper integer of $\lceil$`minsplit/3`$\rceil$ which in this case is 7. Finally, the complexity parameter `cp` is equal to 0.01.

3.2 Random Forest

The next chapter is mainly based on Hastie (2009), otherwise will be specified. RTs provide a model with good interpretability and a low bias. However, one of the major disadvantages of RTs is their high variance in prediction, this means that a few changes in the training dataset can generate a completely different tree, and therefore predictions. To mitigate this problem ensemble models are used. Improvements in classification accuracy and variance can be done from growing an ensemble of trees (Random Forest, RF) and using all of them in the estimation of the prediction. To do so, RF uses the concepts of Bootstrap Sampling and Bagging (Bootstrap aggregation).

3.2.1 Bootstrap Sampling

Suppose we have the training data set $L = (L_1, L_2, ..., L_N)$, where $L_i = (x_i, y_i)$ denotes an observation. The basic idea of Bootstrap Sampling is to randomly draw B datasets with replacement from the training data L, where each sample has the same size as the original training set. This procedure generates B different sub-sets B_b for $b = 1, ..., B$. The remaining observations in the set $L - B_b$ are called "Out of bag samples" (OOB).

3.2.2 Bagging (Bootstrap aggregation)

To understand the concept of bagging, suppose that we fit a model to each of the bootstrap samples B_b of our training data L. For a given input x_i we obtain the predictions $\hat{f}^b(x)$. Bootstrap aggregation or bagging averages the predictions made over the collection of bootstrap samples. It is a procedure that reduces the variance in prediction, and is defined by:

$$\hat{f}_{bag}(x_i) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x_i)$$

In the case of RF, $\hat{f}^b(x_i)$ denotes the b^{th} tree's prediction with the input vector x_i. Each bootstrap tree will typically involve different features than the original and might have a different number of terminal nodes. The bagged estimate is the average prediction at x_i from these B trees.

3.2.3 Random Forest Algorithm

The Algorithm followed for constructing a RF for regression analysis is as follows:

1. Create B trees, starting from $b = 1$ until B and for each of them:

    a. Draw a bootstrap sample B_b of size N from the training data.

    b. Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.

        i. Select k variables at random from the p variables.

        ii. Pick the best variable/split-point (x_{ik}, s) among the k variables.

        iii. Split the node into two daughter nodes.

2. Output the ensemble of trees $\{T_b\}_1^B$.

We have explained that a RF is constructed with random features, first in the bootstrap samplings and second in selecting at random, and at each node, a small group of input variables. Accordingly, the basic idea is to grow a tree each time using the methodology explained in section 3.1.1 and do not apply pruning (Breiman, L., 2001). The idea in RF, showed in the previous algorithm, is to keep low bias and to improve the variance reduction of bagging by reducing the correlation between the trees. This is achieved in the tree-growing process through random selection of the input variables, the bootstrap samples, and the idea of not pruning.

3.2.4 Prediction in Random Forest

To make a prediction at a new point x_i, we use:

$$\hat{f}_{RF}^{B}(x_i) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}_{RF}^{T_b}(x_i) \qquad (3.4)$$

In addition, the values that where use in the RF algorithm were for k the number $\lfloor \sqrt{p} \rfloor$ that in this case is eight, a minimum node size of five, as well as 500 trees for the value of B.

3.2.5 Variable importance in Random Forest

When the T_b-th tree is constructed, the prediction accuracy obtained with the OOB samples is recorded. Then the values for the jth variable are randomly permuted in the OOB samples, and the accuracy is again calculated. The decrease in accuracy as a result of this permutation is averaged over all trees. This rate is used as a measure of the importance of variable j in the RF (Breimann, 2001). Then the same procedure is repeated for all p variables in the input set X. The variable that has the highest contribution rate to the accuracy is the most important variable, sometimes this rate is given in percentage (Breimann, 2001).

3.3 Root mean square forecasting error

The measure of forecasting error for the N observations in the prediction of the response $f(x_i)$ is given by the root mean square forecasting error (RMSFE) (Fahrmeir et al, 2013). The predictions can be taken from equations 3.3 or 3.4, it is calculated with:

$$RMSFE = \sqrt{\frac{1}{N}\sum_{i=1}^{N-1}(f(x_{i+1}) - \hat{f}(x_{i+1}))^2}$$

3.4 Software tools

The Software used for this report is R in the version R 4.0.5 GUI 1.74 for MAC iOS (R Core Team, 2021). The packages used in this report are for RTs the so called “rpart” (Wright, M. & Ziegler, A., 2017) and for Random Forest ”Ranger” (Therneau, T. & Atkinson, B., 2019).

4 Results

4.1 Analysis of the Data Set

To obtain more information about the variables that are used in this report we invite to take a look at the report (Duran, J. 2022).

4.2 Regression Tree of GDP growth

We predict the `GDP growth` by fitting a RT with all the data between the second quartile of the year 1959 and the last quartile of the year 2021. The first observation, Q1 1959, has been used for data transformation purposes. The set of explanatory variables that are used to predict an observation are the observations between lag 1 and lag 10 of the forecasted observation. This means that the ten observations between Q2 1959 and Q3 1961 are used for constructing the first set of lag values for observation Q4 1961. That is the reason why we predict the values from the Q4 1961 until Q4 2021. The total number of tuples used to fit the RT model is 241. We do not predict observation with missing values in the explanatory variables. The resulting RT is depicted in Figure 1.

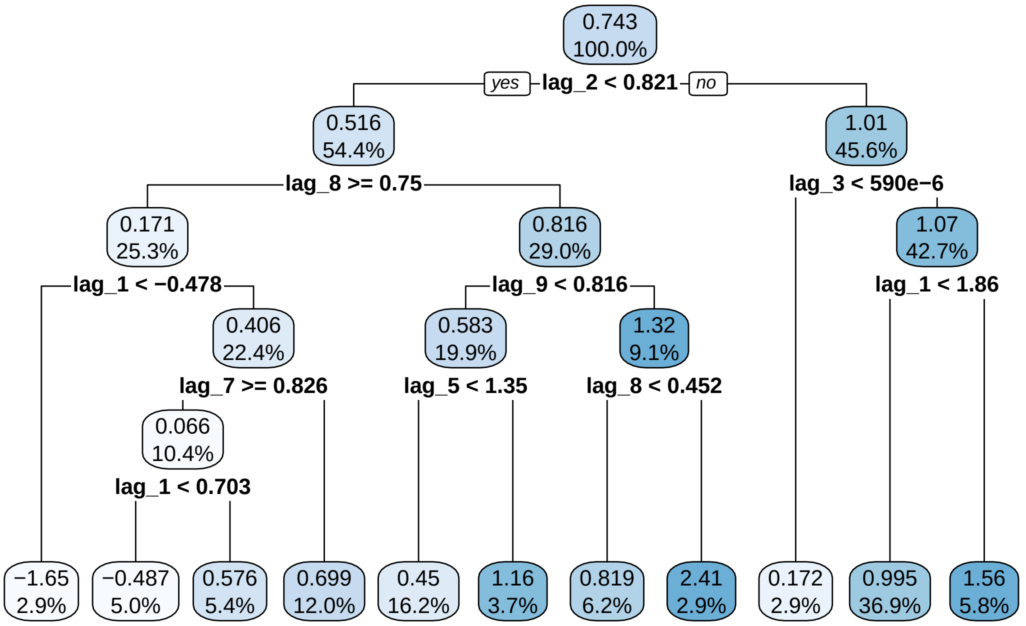

Figure 1: Regression Tree using only lagged information of `GDP growth`

In Figure 1 it is possible to see that the second lag of `GDP growth` is the variable that generates the highest improvement in the accuracy of the prediction. It is located at the root node and generates two partition regions that divide the data in two sets, one with 54% of the total observations and the other with 46%. The right branch of the tree shows an interaction between the lower lags of `GDP growth`, it computes an attribute test for lag 3 and lag 1 respectively. The left branch shows a relation between low and high order lags. We can find that the terminal node that is most to the left has 2.9% of the total observations, that is equivalent to 7 observations (`minbucket`), and isolate the negative extreme observations related with the COVID-19 pandemic. In the same way, the eighth terminal node, from the left to the right, isolate the extreme positive observation of GDP growth. Figure E.1 in the appendix depicts the prediction for the training data set which generates a RMSFE of 0.901 for the whole period, and shows signs of overfitting.

4.3 Regression Tree of GDP growth using all variables

The `GDP growth` is now predicted using lag 1 to lag 10 of all variables. As was explained in section 4.2, we first fit the model with the observations between Q2 1959 and Q4 2021 and then predict the values between the Q4 1961 until Q4 2021. Figure 2 illustrates the obtained regressing tree for this setup.

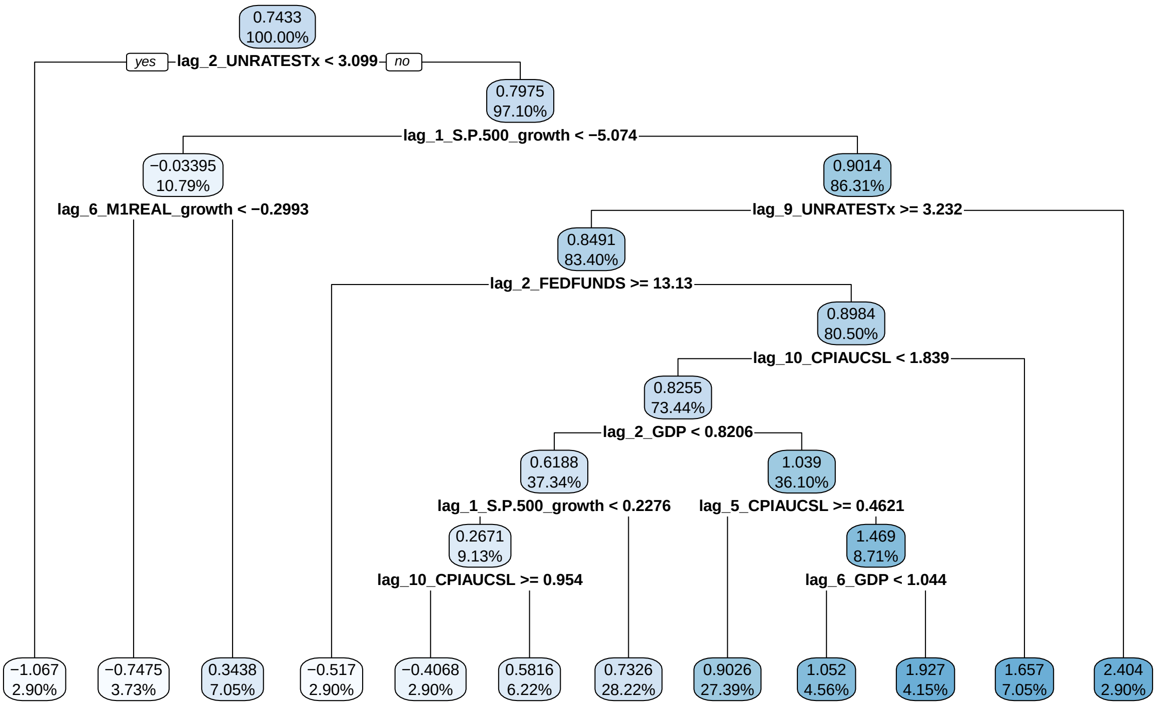

Figure 2: Regression Tree using lagged information of all variables

The first important fact that we can see in Figure 2 is that the RT mainly uses attribute tests of the new variables that were added. This indicates that the new variables add more information to the model and help to generate better partition regions. The previous tree uses lag 2 of the variable `UNRATESTx` in the root node to make the first split. If the condition is met, this split separates only 3% (7 observations) of the data. However, in this partition are observations that lie between the Q1 of 2019 and the Q2 2020. This means that the subset isolate the extreme negative value of `GDP growth` that occurs in 2020 due to COVID-19 Pandemic.

Traversing the tree to the right, it is possible to see that the not fulfilment of the two conditions in the variables lag 1 `S&P 500` and lag 9 `UNRATESTx` also isolate the extreme positive observation that occurs in `GDP growth` because of the pandemic. The predictions of the RT can be seen in the Figure E.2 in the Appendix. In the fourth level of the tree still 83% of the data is concentrated. To define the remaining prediction regions, seven attribute tests need to be executed. In the sixth level of the tree, it is possible to find the same attribute test that is in the root node of RT depicted in Figure 1. It divides the remaining data until that point in two groups of comparable size.

The RT algorithm tries to reduce the sum of square error by grouping or isolating observations that present similar values of the response variable y. In general, most of the terminal nodes contain small groups (of size `minbuket`) of observations that have similar `GDP growth` values, some of these groups can be related with economic events that have been isolated for the RT algorithm. In contrast, as a result of the pruning process and to avoid overfitting there are two predicting nodes that concentrate 55% of the observations.

4.4 Regression Trees one-quarter-ahead forecasts of GDP growth

We forecast the `GDP growth` by applying RT models in the data between Q4 1961 and Q4 2021. Two type of models are fitted. The first one uses ten lags of `GDP growth` and each time all the information before the forecasted value is used in the model. The second model uses lag 1 until lag 10 of all seven variables described in the data section and is fitted each time with all available information until the forecasted time point. The forecasting results of both models can be seen in Figure 3.

Both models start predicting from Q1 1962 and have the same predicted values until Q2 1968. This is because the number of observations in the starting models do not exceed the `minsplit` parameter. For this observations the predicted value is the mean of the observations that are located in the root node until the `minsplit` value is exceeded. The RT models analyzed in this report and the models in (Duran, J. 2022) require diverse setting parameters as well as different number of pre-sample observations. This situation must be handled carefully for a direct comparison of the models. We first use the overall RMSFE as a comparison criteria. Given the same training dataset for all models, the RMSFE for the RT using only information of GDP growth is 1.26, while for the RT using all variables is 1.23. From (Duran, J. 2022) the RMSFE value for AR(1) is 1.20, and 1.98, respectively 3.87, for VAR(1) and VAR(3). In addition, the RT models used in this report

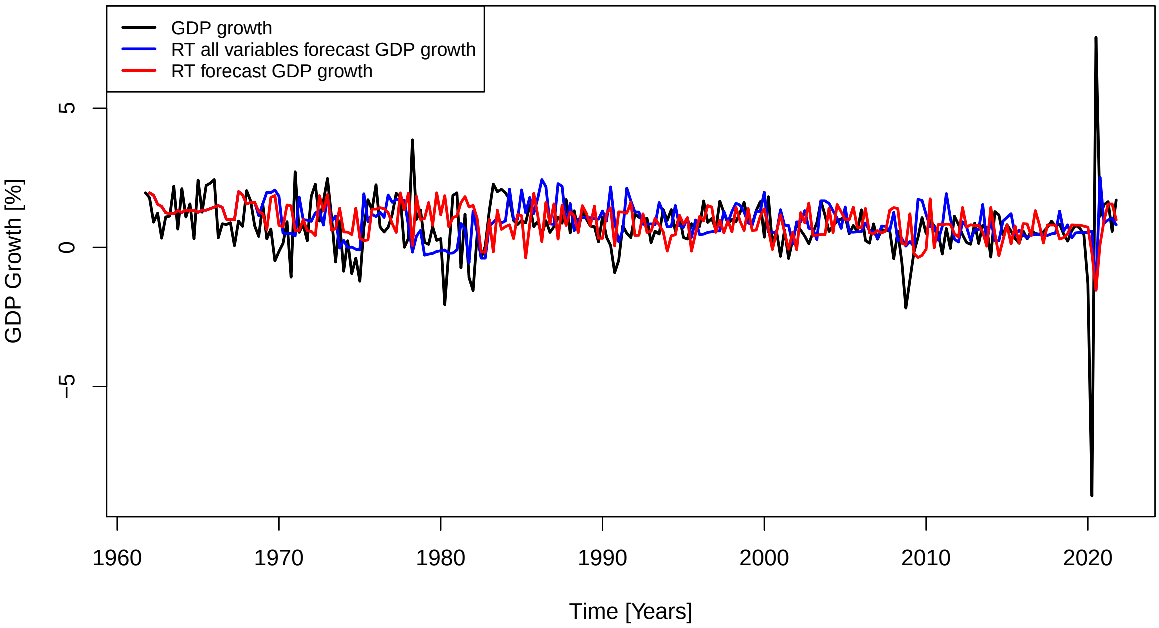

Figure 3: One-quarter-ahead forecasts Regression Trees

require ten pre-sample observations to construct the attribute space, which means that the prediction starts from the twelfth observation. AR(1) only requires three pre-sample observations for parameter estimation and the prediction starts from the fourth. For the VAR(1) and VAR(3) models the number of required pre-sample observations are eight and 25, respectively and the predictions are the observations ninth and 26, respectively.

After executing RT for predicting `GDP growth`, we can see that the RMSFE of the one-quarter-ahead RT using the first 10 lags of all variables is lower than of the tree that only uses the lagged information of GDP. However, this is not enough to beat the accuracy of AR(1), which has a better performance in terms of RMSFE and uses less pre-sample observations.

4.5 Variable importance Random Forest

The method described in section 3.9 is used to evaluate the importance of the variables in a RF model that uses observations between lag 1 and lag 10 of all seven variables. Figure 4 illustrates the first five and the last five important variables.

The variable that generates most improvement in the sum of squared errors along the 500 random trees is the lag 7 of the variable `FEDFUNDS` followed by the lag 8 of the same variable. In contrast, the variable that deteriorates the sum of squared errors the most is the lag 4 of the variable `UNRATESTx`, followed by lag 3 of the same variable. Comparing

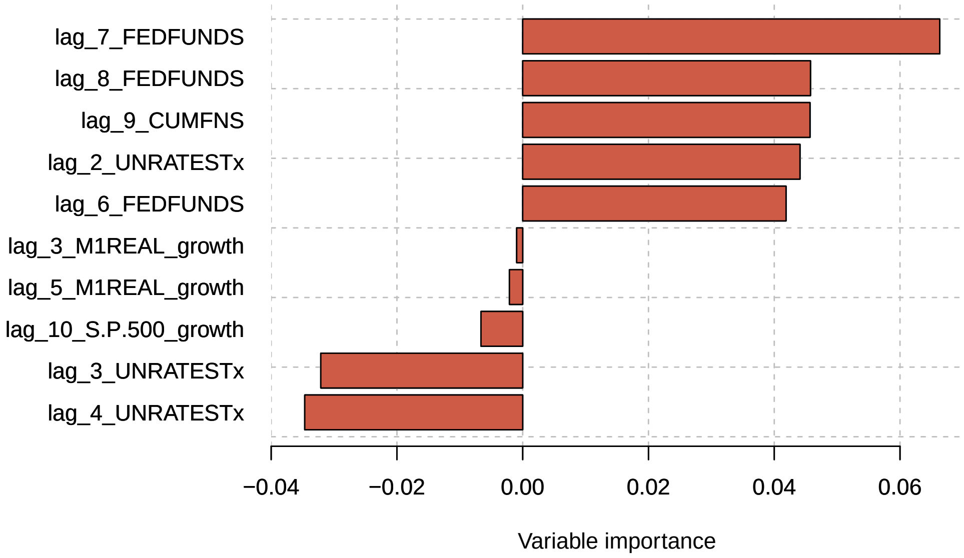

Figure 4: Variable Importance measure Random Forest

this result with the variables that are used in the RT of all variables (Figure 2), we can see that the variable that is in the root node is at the fourth position of the RF variable important measure. Also, it is possible to see that in the top 5 of important variables for RF there is no other variable that is used in the tree of section 4.2. This can be explained by the randomness presented in the RF algorithm, the random selection of the bagging samples as well as the randomness in the selection of the set of variables that are used for splitting. The presence of outliers in `GDP growth` as well as in other variables impacts the measure of importance used in RF.

4.6 Random Forest one-quarter-ahead forecasts of GDP growth

In this section of the report we forecast the `GDP growth` by applying the RF model in the data between Q4 1961 and Q4 2021. The algorithm uses the observations between lag 1 and lag 10 of all seven variables taking the forecasted time point as a reference. A RF algorithm is run each time with all available information until the forecasted period. The official `GDP growth` as well as the forecasting result of one-quarter ahead RF can be seen in Figure 5. There we can see that the one-quarter-ahead RF forecasted values fit appropriately contractions and expansion of the economy, however the complete effects in `GDP growth` for economic events are not well predicted. In addition, we can see that the model tries to fit the contraction effect of the COVID-19 pandemic, but it has problems to forecast the recovery of the economy that follows. The RMSFE of the one-quarter-ahead RF for the forecasted values that are between Q1 1962 and Q4 2021 is 1.20.

4.7 Model Comparison

As we have discussed in section 4.3, the one-quarter-ahead AR(1) model, with 1.20, has a lower RMSFE than the VAR(1), VAR(3) and the one-quarter-ahead RTs fitted with only the information of the GDP and with the information of all variables, respectively. The RMSFE value for the one-quarter-ahead RF fitted for all variables is 1.20. With rounding to the second decimal point, this value is equal to the RMSFE for AR(1). This indicates that the RF has a comparable error in prediction as AR(1). It is important to be aware that the RMSFE of RF can increase or decrease depending on the random bagging samples and the random splitting variables that are used in the fitting process. A comparison of these two models can be seen in Figure 5.

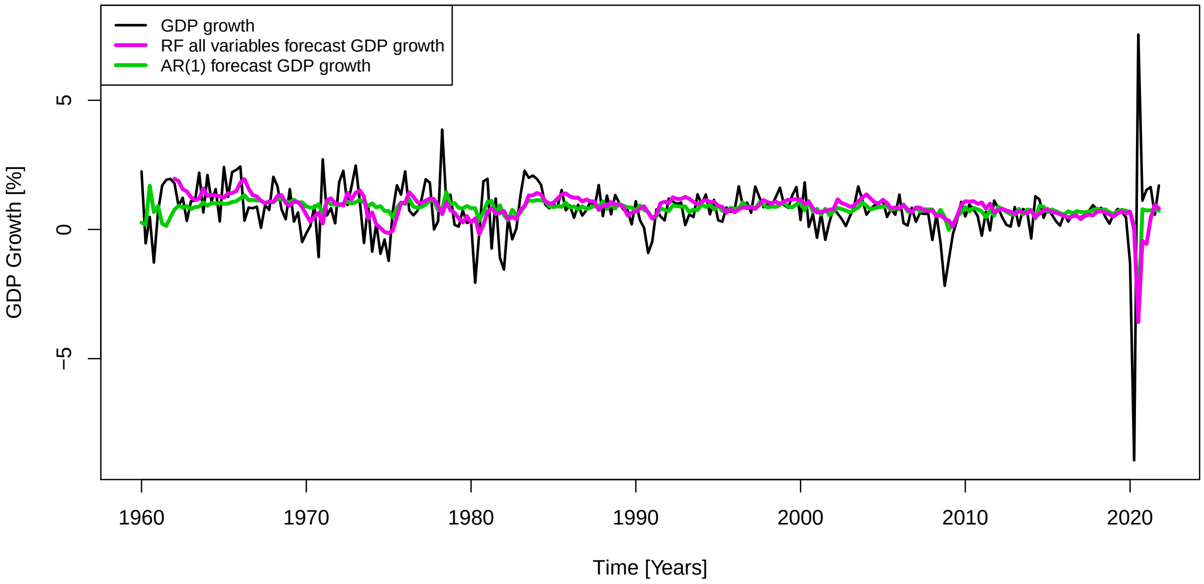

Figure 5: Random Forest vs. AR(1) forecast

In Figure 5, it is possible to see that the AR(1) model starts predicting earlier than the RF model because AR(1) uses just 3 pre-sample observations, while RF needs at least 10. In general, we can see that the RF model predicts expansions or recessions in the economy better than AR(1). AR(1) predictions have a low variability over the whole time period. Both models fail in predicting the extreme recovery value in the economy in 2020 after the COVID-19 pandemic. If we only consider the observations between Q1 1962, when the RF algorithm starts predicting, and Q1 2020, before the COVID-19 pandemic, the values for RMSFE are 0.76 for RF and 0.78 for AR(1). Based on this comparison we can conclude that the RF model provides more accurate predictions than AR(1).

5 Conclusion

This report presents the results of forecasting the US real GDP growth by fitting Regression Tree (RT) models as well as Random Forest (RF) models. The used macroeconomic data is provided by the Federal Reserve Economic Data. To construct the training data set we used seven variables with 252 observations. The variables that are used in the models are unemployment rate, manufacturing industry capacity, inflation, Federal Funds rate, S&P 500 growth and measures of the variation of liquidity in the economy like the growth in M1 real.

The tree structures of the RT model based on lagged information of the `GDP growth` and of the RT model using lags of all seven macroeconomic indicators generate a good fitting on the training data and are useful to identify important variables to predict `GDP growth`. According to these structures, the lag 2 of variable `GDP growth` and the lag 2 of `UMRATESTx` are the variables that provide most information to increase the prediction accuracy of the model. The one quarter ahead prediction error, measured by the root mean square forecasting error (RMSFE), of the RT model using all variables is lower than the error obtained with RT model using only lagged information of GDP. In the same way, using permutation variable importance for a RF using all variables and all observations, we found that lag 7 and 8 of `FEDFUNDS` followed by lag 9 of `CUMFNS` and lag 2 of `UMRATESTx` are variables that lower the forecasting error of RF. Furthermore, two RF models are implemented to predict one quarter ahead observations. The first model uses only lagged information of GDP and the other lagged information of all variables. The second RF model using all variables gives reasonable and consistent predictions of expansions and recessions in the economy, however it also fails to predict extreme observations. After applying the methods outlined above and comparing the forecasting errors RMSFE with the ones founded for the autoregressive model AR(1) and vector autoregressive models VAR(1) and VAR(3) in previous work, we conclude that the model with the lowest forecasting error is the RF that uses lagged information of all variables.

Lastly, further research on exploring different lag-order sets in RTs and RF models is recommended. Also, hyperparameter tuning to select the best parameter values for RTs and RF could be done. In addition, future work could implement methods of outlier treatment for the variables `M1REAL` and Unemployment Rate. Finally, data inconsistency of the `M1REAL` variable, due to the change of its definition in 2020, could be addressed.

References

Breiman L., Friedman J. H., Olshen R. A., & Stone, C. J. (2000) Classification and Regression Trees, Wadsworth, chapter 8 pp 216-233.

Duran, J (2022). One-Quarter-Ahead Forecasts of US GDP Growth A Vector Autoregression Approach. Case Studies-TU Dortmund, Dortmund, Germany.

Fahrmeir, L., Kneib, T., Lang, S. & Marx, B. (2013). Regression Models, Methods and Applications. Berlin: Springer-Verlag.

Hastie, T. and Tibshirani, R. & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer, chapter 9.2 pp 305-307.

Kitchen J. Monaco R. (2003), REAL-TIME FORECASTING IN PRACTICE. Department of the Treasury. United States of America. URL: https://mpra.ub.uni-muenchen.de/21068/.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

OECD (2009). Gross Domestic Product (GDP), Organisation for Economic Co-operation and Development Publishing,, URL: https://www.oecd.org/berlin/44681640.pdf

Therneau, T. & Atkinson, B. (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. https://CRAN.R-project.org/package=rpart

Wright, M. & Ziegler, A.(2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, 77(1), 1-17. doi:10.18637/jss.v077.i01

Appendix

A Additional figures

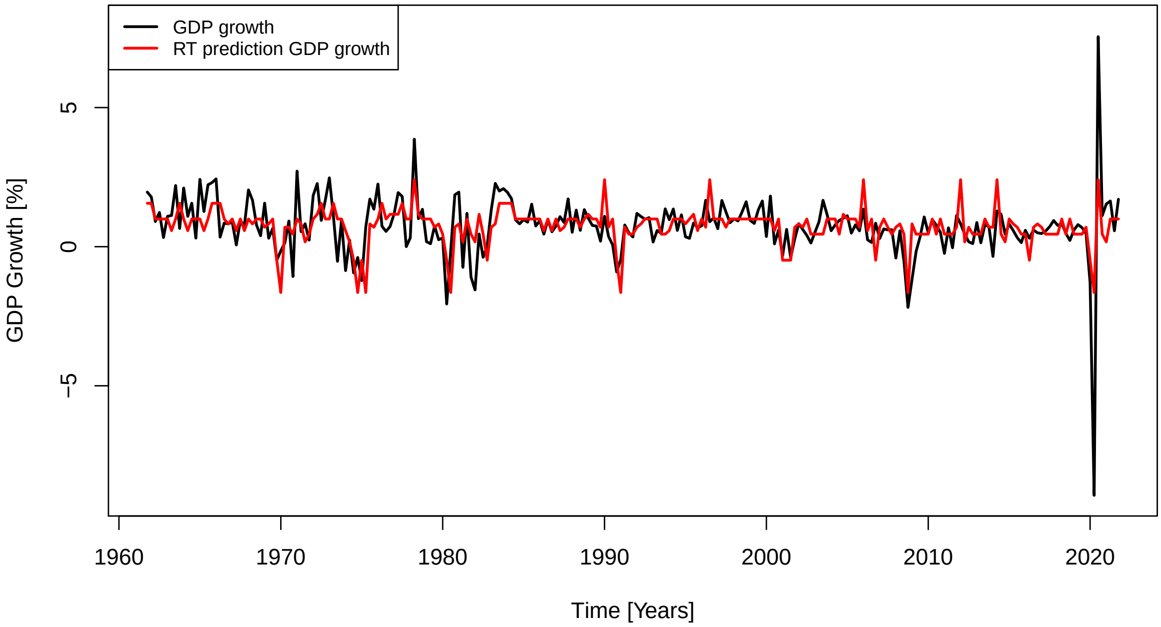

Figure E.1: Prediction Regression Tree `GDP growth`

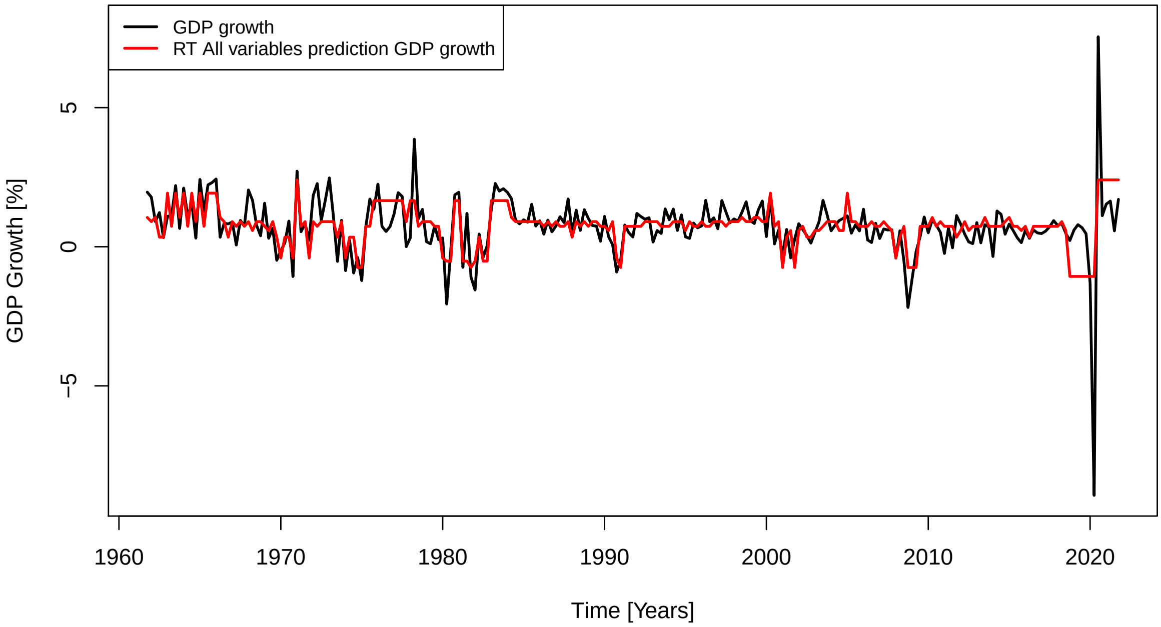

Figure E.2: Prediction Regression Tree all variables