University of Warsaw
Faculty of Economic Sciences

Working Papers
No. 38/2020 (344)

WE HAVE JUST EXPLAINED REAL CONVERGENCE FACTORS USING MACHINE LEARNING

Piotr Wójcik
Bartłomiej Wieczorek

Warsaw 2020

WORKING PAPERS 38/2020 (344)

We have just explained real convergence factors using machine learning

Piotr Wójcik[1]*, Bartłomiej Wieczorek[2]

[1] *Faculty of Economic Sciences, Data Science Lab WNE UW, University of Warsaw*
[2] *Data Science Lab WNE UW*

* *Corresponding author: pwojcik@wne.uw.edu.pl*

Abstract: There are several competing empirical approaches to identify factors of real economic convergence. However, all of the previous studies of cross-country convergence assume a linear model specification. This article uses a novel approach and shows the application of several machine learning tools to this topic discussing their advantages over the other methods, including possibility of identifying nonlinear relationships without any a priori assumptions about its shape. The results suggest that conditional convergence observed in earlier studies could have been a result of inappropriate model specification. We find that in a correct non-linear approach, initial GDP is not (strongly) correlated with growth. In addition, the tools of interpretable machine learning allow to discover the shape of relationship between the average growth and initial GDP. Based on these tools we prove the occurrence of convergence of clubs.

Keywords: cross-country convergence, conditional convergence, determinants, machine learning, non-linear

JEL codes: O47, C14, C52

Acknowledgements:

Research presented in this article was partially financed by the Polish National Science Center under contract number 2016/21/B/HS4/00670.

1. Introduction and literature overview

Seeking for factors of economic growth or income convergence between countries has been a topic of empirical research for decades. One of the most common methods used in empirical research is the analysis of conditional beta convergence (Barro and Sala-i Martin, 2007). It looks for a relationship between the average annual growth rate and initial income, often conditioned on some additional factors. The choice of said factors has a crucial impact on the inference about the occurrence of convergence (Durlauf, 2009). Conclusions regarding the importance of individual factors often vary between different empirical approaches.

Several approaches were proposed so far. Most of the approaches were linear ones (i.e. they seek for a linear relationship between growth and explanatory variables) (e.g. Sala-i Martin (1997) or Hendry and Krolzig (2004)), often incorporating Bayesian averaging with different priors in order to take into consideration the uncertainty of the correct form of the model (e.g. Sala-i Martin, Doppelhofer, and Miller (2004), Ley and Steel (2007), Doppelhofer and Weeks (2011)). Above mentioned approaches confirmed the existence of beta convergence – initial GDP was one of the most important factors of growth. However, Ciccone and Jarociński (2010) questioned those methods both on theoretical and empirical grounds – Bayesian Model Averaging require arbitrary assumptions and can lead to incorrect conclusions, and only the correct identification of growth factors would allow to form a credible recommendation for economic policy. Moreover, He and Xu (2019) show that identifying a variable to be a statistically relevant factor of growth in a linear specification might be a result of an inappropriately specified model. In the correct non-linear specification, it might not be correlated with growth.

We propose a novel approach. Its main purpose is to identify important determinants of growth and extend the previous research in two ways. First, the article attempts to identify potential non-linearities. Second, it uses methods that allow to select variables relevant for growth and explain the relationship in any specification. This is where commonly used machine learning algorithms come on the stage. We considered LASSO (Tibshirani, 1996), support vector regression (SVR), (Vapnik, 1995), random forests (Breiman, 2001), gradient boosting machines (GBM), (Friedman, 2001) and two instances of extreme gradient boosting (XGBoost) (Chen and Guestrin, 2016). All of the above-mentioned approaches have the advantage of dealing successfully with both linear and non-linear relationships. Leave-One-Out cross validation is used for tuning model parameters based on minimization of the Mean Absolute Error (MAE), as we are not using any specified priors. In addition to successful prediction, the ability to interpret what a model has learned is of equal importance.

The research hypothesis verified in the article states that machine learning tools allowing for non-linearity explain cross-country growth rates with higher accuracy than linear models. In addition, these tools still allow for model interpretation and measuring feature importance, thus they are helpful in formulating policy recommendations.

In this article, two datasets widely used before were considered, to make results comparable with other studies – first, the dataset used in Fernandez, Ley, and Steel (2001), second, the dataset introduced in Sala-i Martin, Doppelhofer, and Miller (2004).

This article is structured as follows. In the first section, the methods used in the empirical part are briefly described. In the second section, the results of the analysis are presented on the two above-mentioned datasets. The results are then compared with the findings of previous research. The last section summarizes the conclusions.

2. Methods

In this section, machine learning tools applied in the empirical part are briefly introduced in a non-technical, intuitive way. Before applying algorithms, all variables were standardized, as the literature suggests.

2.1. LASSO (Least Absolute Shrinkage and Selector Operator)

LASSO (Tibshirani, 1996) is one of several regularization methods. It can be viewed as the extension of Ordinary Least Squares (OLS) model. It differs from OLS because of its cost function – it not only minimizes the sum of squared residuals, but also takes into account the sum of absolute values of the parameters of the linear model as an additional constraint. Adding such penalty in the optimization results in searching for parameters that fit the data well, but additionally are as small as possible. Parameters at less important variables will shrink towards zero, some of them will even be set to be equal to zero. At the expense of a certain bias (LASSO estimates are biased), LASSO often allows to obtain more precise forecasts on the test sample (Hofmarcher, Crespo Cuaresma, Grün, and Hornik, 2015).

What is crucial in the case of growth regressions, LASSO can be considered as variable selection method, which can be used even when the initial number of variables exceeds the number of observations. It is often used by researchers as a preliminary stage of analysis, combined with a subsequent model estimation on selected variables using OLS (Schneider and Wagner, 2008). No a priori assumptions or selection of a subset of variables are needed. One has only to determine the optimal weight for the additional constraint λ, which can be done via cross validation.

2.2. SVR (support vector regression)

Similarly to OLS, SVR (Vapnik, 1995) fits a hyperplane that is positioned as close to all data points as possible. However, while OLS minimizes the sum of squared errors, SVR tries to fit the errors within a specified distance from the hyperplane (Smola and Schölkopf, 2004). Moreover, the setup includes additional regularization hyperparameter C, which controls how much one wants to avoid misclassifying each observation. The most important advantage of SVR over OLS is the ability to model non-linear relationships between variables using selected kernel functions. SVR applies an implicit non-linear mapping into a higher dimensional feature

space, where it is more probable to find an appropriate hyperplane (Vapnik, 1995). Thus, one can think of SVR as a process of performing a linear regression in a more dimensional space. Two widely used types of kernels are radial basis function and polynomial kernel. We applied both kernels in the empirical part of the article.

2.3. Decision tree regression

Decision tree regressions are predictive models structured in a tree-like way. The model breaks data into smaller sub-datasets with respect to the values of explanatory variables. The process of such a break can be viewed as asking a series of questions whether observations satisfy specified conditions. Each question creates separate nodes, which narrows the possible output value. The whole process starts in the so-called root, where on has no boundaries for the variables, and stops in the leaf, where observations satisfying the set of conditions with respect to explanatory variables obtains their final prediction. The decision of how to make splits heavily affects tree's accuracy. Decision tree regressions often use Mean Squared Error (MSE) metric to decide whether to split a node in two sub-nodes. The decision to split the data has to take into consideration two factors – first, whether the decision to split is a correct one, second – with respect to which variable the split should be performed and what should be the optimal threshold value. Tree models require some stopping criterion to set – one can use the maximum depth of the tree.

2.4. Bagging (Bootstrap Aggregating) and random forests

Bagging is an ensemble technique. It combines multiple models – called weak learners – trained on different bootstrap subsamples of the original dataset. The process of sampling is done randomly with replacement. On each of the subsamples, only one model is trained. Then, the prediction from the bagging model is obtained by the majority voting from all models in case of classification problem, or by averaging the predicted values from all models in case of regression problems.

Bagging approach is therefore often used to decrease the variance of predictions. The weak learner is a simple predictive model, which predictions might not be strongly correlated with the real values. However, combining multiple weak learners can create a strong learner – the predictive model, which predictions might be correlated with the real values much stronger. The common examples of weak learners are decision trees or OLS models. We used both of those models as weak learners for selected models in the empirical part of the research.

Random forests are a particular case of bagging algorithms. They were first introduced by Breiman (2001). In simple words, they are a combination of tree models. Each tree is trained on a different bootstrap subsample of the original dataset, just like in bagging. In addition, at each split of each tree, only a random subset of all predictors is considered. This way, the trees are decorrelated, which is the main advantage of random forests over other bagging approaches with tree models used as weak learners.

Random forests are robust to the problem of multicollinearity and can be applied to a large number of potential predictors without initial selection. In addition, they are indifferent to non- linear interlinkages between the data. They require tuning of two parameters – the number of trees and the number of predictors considered at each split.

2.5. Boosting

Boosting is another type of ensemble. However, its' principle is quite different than the one from the bagging technique. While bagging averages the predictions from multiple weak learners, boosting approach combines them iteratively. One can imagine that during each step, weak learner is trained on the weighed sample. The weights are set with respect to the prediction error from the previous iteration – the higher the error, the higher the weight in the following step. Thus, this approach ensures that the model obtained after several iterations has the lowest possible prediction error.

There are multiple boosting methods considered in the literature. One of the most popular is gradient boosting. It became popular after Friedman (2001) described the algorithm of gradient descent in the function space, and then applied it to the cost functions of popular predictive models.

The parameters to optimize in the gradient boosting approach is the learning rate, which is a kind of shrinkage parameter – it shows how quickly the errors are corrected between the weak learners. Other hyperparameters are those that are needed by the selected weak learners. Nowadays, the extension of gradient boosting, called eXtreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016) is widely used. XGBoost extends gradient boosting by different penalization of trees, adding a shrinkage to the leaf nodes and the extra randomization parameter.

2.6. Cross validation

Machine learning algorithms require selection of hyperparameter values, i.e. parameters that are not optimized in the model training procedure (e.g. penalty for too large parameters in the LASSO approach, the cost of incorrect classification in SVR, the number of trees in random forest, or the learning rate in gradient boosting approach). Hyperparameters can be chosen arbitrarily, but it's better to choose them consciously. In the empirical part of the article, we use a Leave-One-Out cross validation (LOOCV) procedure. For each combination of values of hyperparameters, each model is estimated n times on the sample without the 1st, 2nd, 3rd... observation, respectively. The single observation left aside is used as a test sample – for assessing the quality of prediction. Based on all predictions for a specific combination of hyperparameters, Mean Absolute Error (MAE) is calculated. Finally, we select and apply the model with the hyperparameters that minimize the MAE for prediction.

2.7. Variable importance and interpretable machine learning

Many machine learning algorithms have their own specific way to measure the importance of each feature. But the lack of model interpretability is the most important limitation of many machine learning tools. The quality of predictions is important in research, but it is even more important to understand the mechanism that drives the prediction of the particular phenomenon. In recent years, a field called Interpretable Machine Learning (IML) or eXplainable Artificial Intelligence (XAI) has been developing rapidly (Molnar, 2019). It offers additional tools to overcome the black-box dilemma and allow for easy comparability of variable importance across different models.

In the empirical part of this paper, we used the measure called model reliance (Fisher, Rudin, and Dominici, 2019), inspired by the permutation-based approach of Breiman (2001). It describes how much the model's performance relies on different covariates. In the permutation-based approach, to assess the importance of a selected feature, one calculates the error of the prediction from the model on the original dataset, e_{orig}, and on the artificial dataset, with the values of said feature randomly permutated, e_{perm}. The higher the ratio e_{perm} / e_{orig}, the more important the feature, as it describes how much the error arose when the feature became non-informative. Model reliance generalizes such approach by taking into consideration not one, but all permutations that permute values of the selected feature.

After identifying influential variables, one has to understand the relationship between these variables and the response from the model. On the level of each observation, ceteris paribus profile can be analyzed. In essence, they show how a conditional expectation of the dependent variable changes with the values of a particular explanatory variable, while all other variables are kept constant (Goldstein, Kapelner, Bleich, and Pitkin, 2013). Averaging ceteris paribus profiles over all observations shows how the expected model response behaves as a function of a selected feature. This procedure, called Partial Dependence Profile, was first introduced by Friedman (2001).

2.8. Methods used in the empirical part

In the empirical part of the research, we estimated several machine learning models on two datasets, mentioned in the introduction. Selected models are LASSO, OLS model with variables selected via LASSO, SVR models with radial basis function and polynomial kernels (individually), random forest, gradient boosting machine with decision trees selected as weak learners, and XGBoost models with two different weak learners – OLS models and decision trees. Hyperparameters of the models were selected via Leave-One-Out cross validation, with exception of the number of boosting rounds for boosting algorithms, which was arbitrarily set to 50. After the estimation process, we assessed the importance of features using model reliance approach. And finally, we analyzed the Partial Dependence Profiles for the initial level of GDP, to verify the possible occurrence of beta convergence.

3. Empirical results

The above-mentioned methods – LASSO, OLS, SVR, random forest and boosting algorithms – were applied on two widely studied datasets. The first of the datasets was used by Fernandez, Ley, and Steel (2001) and includes 41 explanatory variables for 72 countries. This dataset was referred to hereafter as FLS. The second one was used by Sala-i Martin, Doppelhofer, and Miller (2004) and includes 67 explanatory variables for 88 countries – referred to hereafter as SDM.

Schneider and Wagner (2008) applied LASSO type regression on these datasets and claim that estimation results were in line with the findings of the original paper. However, it is not confirmed below, especially for SDM data.

In each case, the results are reported and compared with the results of the original papers and some of their follow-ups. Variable importance ranking (based on the model reliance measure) is provided to show results in a consistent way and compare them with previous studies.

3.1. Analysis on FLS data

Tables 1 and 2 show the ranking of important variables identified originally in Fernandez, Ley, and Steel (2001), additionally compared with following Hendry and Krolzig (2004) and previous Sala-i Martin (1997). The convention is that the lower the ranking, the more important the variable. We show only the first most important 20 variables, according to Fernandez, Ley, and Steel (2001).

Based on Table 1, it appears that all approaches confirm the occurrence of conditional convergence, although, in ensemble-techniques (random forest and boosting algorithms), the initial level of GDP has a lower ranking. On the contrary, for LASSO, OLS based on LASSO and SVR models, initial GDP level is the most important factor. Most of the approaches also agree on the importance of *Fraction Confucian*, *Life expectancy* and *Equipment Investment* factors. However, machine learning tools do not confirm 5-10 out of 20 most important variables indicated by Fernandez, Ley, and Steel (2001).

The results of LASSO and SVR models seem very consistent with Hendry and Krolzig (2004) – among top 13 variables in their model, 12 are also the most important in LASSO, and among their top 11 variables, 10 are also the most important in both SVR approaches.

Besides that, most of the machine learning algorithms indicated *Non-Equipment Investment* as an important convergence factor, which is in contrary to original articles. There are also important growth determinants, missed by original approaches, but confirmed by other methods (including Hendry and Krolzig (2004)) – *Size labor force*, *Ethnolinguistic Fractionalization* and *Higher education enrollment*.

We can also spot some similarities between ensemble models. Random forest and three boosting approaches acted quite similarly – all of those models neglected *Sub Saharan dummy*, *Rule of Law*, *Latin American dummy* and *Fraction Hindu* factors, which were high in rankings for other models. Similarly, for all of the above-mentioned models, *Number of Years open economy* was listed as an important variable, which is in line with Fernandez, Ley, and Steel (2001) and Sala-i Martin (1997), but was neglected by LASSO, OLS and SVR models.

Table 1: Rank of importance of growth determinants for FLS data

Variable	FLS (2001)	S (1997)	HK (2004)	LASSO	OLS (LASSO)	SVR (poly)	SVR (radial)
GDP level in 1960	1	1	1	1	1	1	1
Fraction Confucian	2	1	11	5	7	5	3
Life expectancy	3	7	2	2	3	4	2
Equipment investment	4	1	8	8	10	8	6
Sub Saharan dummy	5	10	4	4	5	3	4
Fraction Muslim	6	1	–	17	20	29	29
Rule of Law	7	1	10	11	13	10	10
Number of Years open economy	8	1	–	35	31	35	27
Degree of Capitalism	9	17	–	18	25	19	17
Fraction Protestant	10	22	–	16	23	11	11
Fraction GDP in mining	11	13	16	12	17	13	12
Non-Equipment Investment	12	19	–	15	19	17	13
Latin American dummy	13	8	7	9	8	9	9
Primary School Enrollment, 1960	14	15	12	13	12	12	15
Fraction Buddhist	15	23	–	22	27	21	14
Black Market Premium	16	30	–	19	21	18	21
Fraction Catholic	17	24	–	–	–	38	40
Civil Liberties	18	10	–	20	14	27	33
Fraction Hindu	19	35	3	3	2	2	5
Primary exports, 1970	20	16	–	28	26	31	30
...							
Size labor force	25	28	5	6	4	6	7
...							
Ethnolinguistic fractionalization	28	36	13	10	11	14	16
...							
SD of black market premium	30	14	–	33	36	30	22
...							
Higher education enrollment	34	39	6	7	6	17	8
...							
Public Education share	40	–	–	26	24	22	20

FLS (2001) is a reference to Fernandez, Ley, and Steel (2001), S (1997) is a reference to Sala-i Martin (1997), HK (2004) is a reference to Hendry and Krolzig (2004), OLS (LASSO) indicates OLS model with variables selected using LASSO approach, SVR (poly) indicates SVR model with polynomial kernel, SVR (radial) indicates SVR model with radial basis function kernel.

Table 1: Rank of importance of growth determinants for FLS data (cont'd)

Variable	FLS (2001)	S (1997)	HK (2004)	Random Forest	GBM	XGBoost (trees)	XGBoost (OLS)
GDP level in 1960	1	1	1	10	7	7	13
Fraction Confucian	2	1	11	5	15	2	6
Life expectancy	3	7	2	7	2	4	21
Equipment investment	4	1	8	1	1	1	2
Sub Saharan dummy	5	10	4	32	40	41	41
Fraction Muslim	6	1	–	22	25	5	18
Rule of Law	7	1	10	27	41	28	31
Number of Years open economy	8	1	–	4	6	9	10
Degree of Capitalism	9	17	–	30	22	34	24
Fraction Protestant	10	22	–	14	9	10	7
Fraction GDP in mining	11	13	16	25	17	22	23
Non-Equipment Investment	12	19	–	3	4	3	3
Latin American dummy	13	8	7	26	36	39	39
Primary School Enrollment, 1960	14	15	12	20	19	13	22
Fraction Buddhist	15	23	–	2	3	6	1
Black Market Premium	16	30	–	29	27	27	33
Fraction Catholic	17	24	–	12	18	23	25
Civil Liberties	18	10	–	21	30	21	15
Fraction Hindu	19	35	3	34	39	30	30
Primary exports, 1970	20	16	–	13	16	19	5
...							
Size labor force	25	28	5	9	10	8	8
...							
Ethnolinguistic fractionalization	28	36	13	19	12	17	14
...							
SD of black market premium	30	14	–	8	8	11	9
...							
Higher education enrollment	34	39	6	7	13	12	29
...							
Public Education share	40	–	–	11	11	15	12

FLS (2001) is a reference to Fernandez, Ley, and Steel (2001), S (1997) is a reference to Sala-i Martin (1997), HK (2004) is a reference to Hendry and Krolzig (2004), GBM indicates gradient boosting model, XGBoost (trees) indicates XGBoost model with Decision Trees used as the weak learners, XGBoost (OLS) indicates XGBoost model with OLS models used as weak learners.

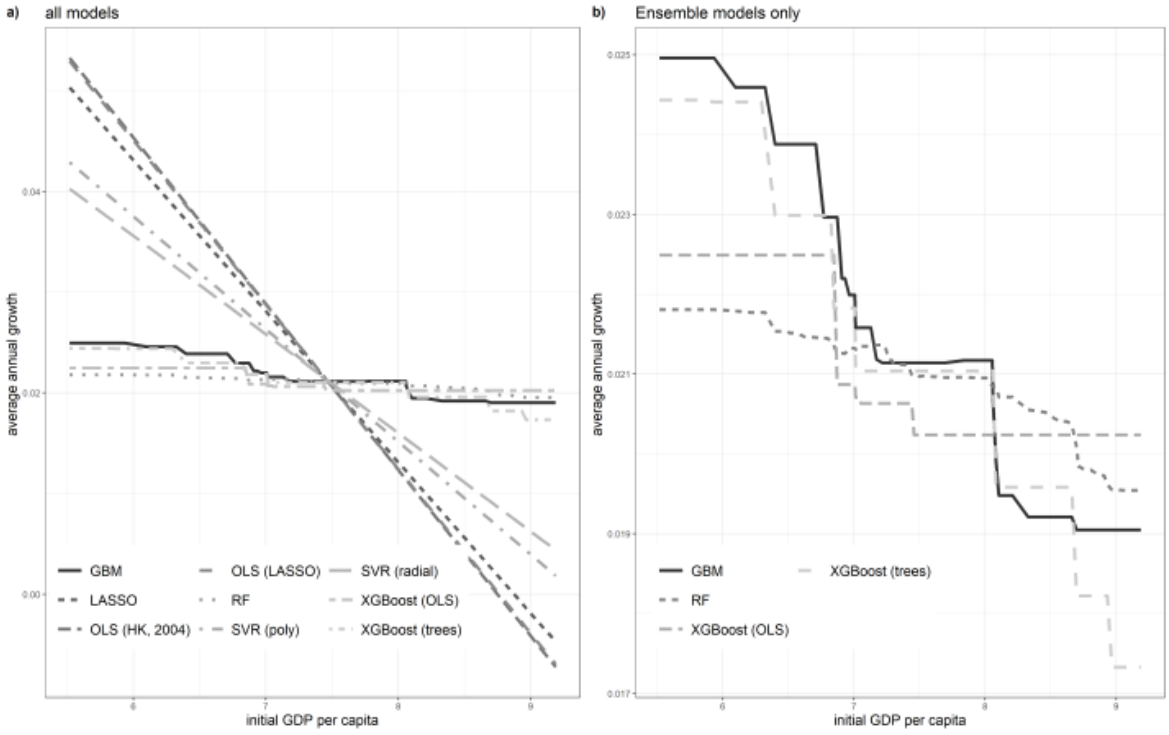
Based on the estimation results of all machine learning models and replicated Hendry and Krolzig (2004) results, Partial Dependence Profiles for initial income were calculated. We plotted them on Figure 1a).

Profiles show that the relationship between the expected growth rate and initial GDP seems to be linear in the case OLS, LASSO and SVR models (as expected), but it turned out to be non-linear (but partially linear) for the ensemble algorithms (Figure 1b)). The relationship is negative, but, again, much flatter for the ensemble algorithms.

If we look closely at ensemble algorithms, we could spot some contrary conclusions. For instance, if we take into consideration GBM or XGBoost with Decision Trees learners, we can see that the strongest convergence occurs for the poorest countries, it does not occur for countries with middle-ranged initial income, and then it occurs again for some of the richest.

However, for the XGBoost model with OLS learners, it appears that countries have similar growth pace for a wide range of initial GDP, and the only downward trend appears for the low-to-middle income countries. In the case of random forest model, the downward trend is consistent, but rather flat, compared to other ensemble methods.

Figure 1: Partial Dependence Profiles for initial income for models estimated on FLS data



In the end, for all estimated models, fit to data measures were compared – see Table 2. All models explain more than 90% of the variability of growth on the whole sample. Moreover, each of machine learning algorithms is better than a linear model in terms of every considered measure. Boosting approaches explain the relationship the best among all models, with the lowest prediction errors, which was to be expected looking at their specification. Although the best model in terms of R2 or the metrics is the XGBoost with OLS learners, we can say that such results confirm our research hypothesis – models that allow for non-linearities explain more of the variability of growth and have lower prediction errors compared to the linear models like LASSO or OLS.

Table 2: Measure of models' fit for FLS data

model	RMSE	MAE	R^2
OLS (HK, 2004)	0,0055	0,0041	90,72%
LASSO	0,0040	0,0031	95,21%
OLS (LASSO)	0,0037	0,0029	95,90%
SVR (poly)	0,0033	0,0024	96,72%
SVR (radial)	0,0035	0,0025	96,21%
Random forest	0,0045	0,0034	93,78%
GBM	0,0018	0,0012	98,97%
XGBoost (trees)	0,0014	0,0011	99,40%
XGBoost (OLS)	0,0006	0,0004	99,90%

3.2. Analysis on SDM data

The differences in conclusions between machine learning models and original studies are more striking in the case of the second dataset. Table 3 shows the ranking of important factors identified originally in Sala-i-Martin, Doppelhofer, and Miller (2004) and in a follow-up study by Doppelhofer and Weeks (2011), who used a "robust" version of Bayesian Model Averaging technique, which was the main approach of Sala-i-Martin, Doppelhofer, and Miller (2004), however, they obtained identical results.

Table 3: Rank of importance of growth determinants for SDM data

Variable	SDM (2004)	DW (2011)	LASSO	OLS (LASSO)	SVR (poly)	SVR (radial)
East Asian Dummy	1	1	1	2	1	2
Primary schooling 1960	2	2	3	4	3	1
Investment Price	3	3	4	3	2	4
GDP in 1960 (log)	4	4	–	–	21	28
Fraction of tropical area	5	5	5	1	13	18
Population density coastal 1960's	6	6	9	7	9	9
Malaria prevalence in 1960's	7	7	11	12	10	12
Life expectancy in 1960	8	8	–	–	11	10
Fraction Confucian	9	9	2	5	4	3
Latin America Dummy	11	11	–	–	23	34
African dummy	11	11	–	–	8	7
Fraction GDP in mining	12	12	–	–	26	31
Spanish colony	13	13	–	–	42	35
Years open	14	14	–	–	29	29
Fraction Muslim	15	15	–	–	41	66
Fraction Buddhist	16	16	7	6	5	5
Ethnolinguistic fractionalization	17	17	–	–	16	16
Government consumption share 1960's	18	18	12	8	18	15
Population Density 1960	19	19	–	–	47	45
Real exchange rate distortions	20	20	10	9	6	6

SDM (2004) is a reference to Sala-i Martin, Doppelhofer, and Miller (2004), DW (2011) is a reference to Doppelhofer and Weeks (2011), other abbreviations are consistent with those in Table 1.

Table 3: Rank of importance of growth determinants for SDM data (cont'd)

Variable	SDM (2004)	DW (2011)	Random Forest	GBM	XGBoost (trees)	XGBoost (OLS)
East Asian Dummy	1	1	3	1	4	5
Primary schooling 1960	2	2	4	16	2	18
Investment Price	3	3	9	7	6	8
GDP in 1960 (log)	4	4	38	17	26	33
Fraction of tropical area	5	5	22	36	59	49
Population density coastal 1960's	6	6	13	23	16	34
Malaria prevalence in 1960's	7	7	1	3	1	1
Life expectancy in 1960	8	8	2	6	5	7
Fraction Confucian	9	9	10	25	8	16
Latin America Dummy	11	11	52	53	30	51
African dummy	11	11	14	48	60	61
Fraction GDP in mining	12	12	42	42	27	50
Spanish colony	13	13	48	43	31	59
Years open	14	14	17	49	22	42
Fraction Muslim	15	15	47	51	25	35
Fraction Buddhist	16	16	5	2	3	4
Ethnolinguistic fractionalization	17	17	25	8	17	12
Government consumption share 1960's	18	18	36	44	50	53
Population Density 1960	19	19	12	5	9	6
Real exchange rate distortions	20	20	8	9	12	22

SDM (2004) is a reference to Sala-i Martin, Doppelhofer, and Miller (2004), DW (2011) is a reference to Doppelhofer and Weeks (2011), other abbreviations are consistent with those in Table 2.

Some of the conclusions based on the ranking are in line with Sala-i-Martin, Doppelhofer, and Miller (2004). *East Asian dummy* and *Investment price* – 1st and 3rd, respectively, most important factor in Sala-i-Martin, Doppelhofer, and Miller (2004) – were confirmed as important by all machine learning tools. Moreover, 2nd most important factor, *Primary schooling in 1960* was low in ranking just for GBM and XGBoost with OLS learners. Several other variables seem to have a strong impact on the growth rate, according to most machine learning algorithms: *Malaria prevalence in 1960s*, *Fraction Buddhist*, *Fraction Confucian* (which is consistent with earlier analysis on FLS data, however, it is low in ranking for GBM), *Life expectancy in 1960* (it is again in line with the analysis on FLS data, but this time, it was excluded by LASSO model).

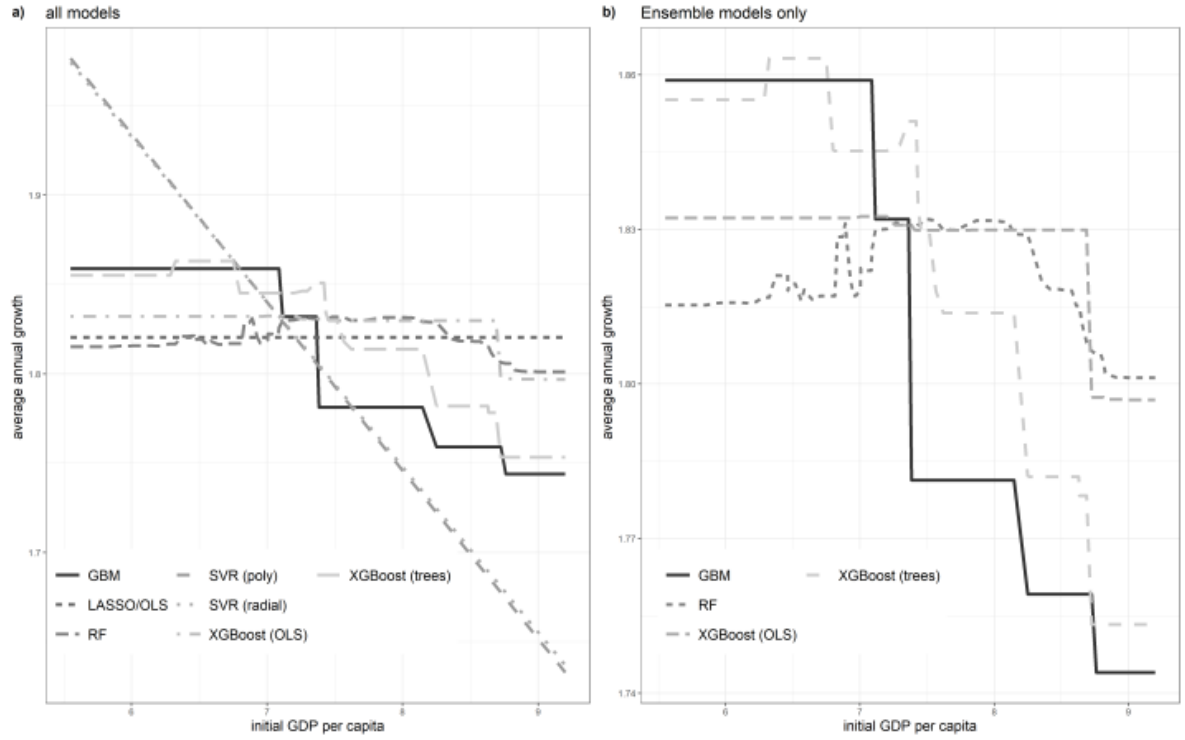
We can also spot some inconsistencies. For instance, *Population density coastal 1960's* was high in the rankings for LASSO and SVR methods, but much lower for the ensemble models. Moreover, *Population Density 1960* was considered as important for ensemble models, especially for GBM, but it was neglected by SVR models and excluded from the analysis by LASSO.

And finally, the most striking result is the ranking of GDP in 1960 (log). Initial GDP was one of the most important variables for models estimated on FLS data. It is also the key factor for the convergence analysis. However, it was excluded from the analysis by LASSO, and in the case of other machine learning approaches, it is in a very far position in the importance ranking. This might suggest the lack of conditional convergence, which was

observed in earlier studies. Such a conclusion is in line with He and Xu (2019), who suggested that such significance might have been a result of inappropriate model specification. In a correct, non-linear specification, initial GDP can be not (strongly) correlated with growth.

Partial Dependence Profiles for initial income for SDM dataset are plotted on Figure 2a). We discussed that we cannot draw any conclusion about the relation for LASSO model. Only SVR models show a linear, negative relationship between growth rate and initial GDP, which was to be expected.

Figure 2: Partial Dependence Profiles for initial income for models estimated on SDM data



In the case of ensemble models (Figure 2b), one can see some interesting phenomenon. If we look at the GBM model and XGBoost model with Decision Trees learners, we can spot that there are plenty of intervals for initial GDP with the same annual growth rate. However, countries that belong to the "poorer" interval generally grow faster than those which belong to the "richer" interval. We can say that countries with the initial GDP in the given interval, belong to the same convergence club. In the case of XGBoost with OLS learners, we can see only 3 such intervals, with 2 among them that do not differ significantly. In the case of random forest, the relation is not consistent – we can spot the convergence only among the richest countries.

In the end, we again show the measures of fit to data for all models (Table 4). Here, only ensemble models explain more than 90% of the variability of growth. Again, the best model is the XGBoost with OLS learners. However, again, we see the confirmation that models that allow for non-linearities perform better than their linear counterparts (LASSO, OLS or SVR).

Table 4: Measure of models' fit for SDM data

model	RMSE	MAE	R^2
LASSO	1,0600	0,8245	68,59%
OLS (LASSO)	0,9867	0,7613	72,78%
SVR (poly)	0,9313	0,6576	75,75%
SVR (radial)	0,9312	0,6573	75,76%
Random forest	0,5052	0,3690	92,87%
GBM	0,4065	0,3119	95,38%
XGBoost (trees)	0,0670	0,0529	99,87%
XGBoost (OLS)	0,0020	0,0013	99,99%

4. Conclusions

The main purpose of the article was to identify the important factors of economic growth by applying machine learning tools. We applied models that allow identifying non-linearities in the data, namely support vector regression, random forests and boosting algorithms. The models were estimated without any prior assumptions with the use of Leave-One-Out cross validation procedure. Moreover, we used the model reliance measure, which allows to easily assess the importance of features for any model type in a consistent and comparable way. To estimate our models, we used two common datasets – FLS data introduced in Fernandez, Ley, and Steel (2001) and SDM data introduced in Sala-i Martin, Doppelhofer, and Miller (2004). Machine learning tools confirmed the importance of several growth factors, such as life expectancy, investment in the equipment and its price. They also pointed at some factors that were low in the rankings of previous studies using purely linear approach, i.e. ethnolinguistic fractionalization, which measures the ethnic and linguistic diversity in the country. The most striking result from our analysis was the difference between the conclusion about the importance of initial GDP when allowing for nonlinearity of its relationship with the growth rate. For FLS data, this factor was one of the most important, which was consistent with the previous studies. In turn, it dropped down the ranking for SDM data, and was even excluded in the case of LASSO approach. This suggests that when using a simplified linear approach one can incorrectly conclude about the occurrence of conditional convergence, while when correctly identifying the non-linear relationship, cross-country convergence is not observed.

When analyzing Partial Dependence Profiles for initial GDP, we could also identify convergence clubs – groups of countries similar in terms of initial GDP per capita for which convergence is observed. We also showed that models that allow for non-linearities generally have higher predictive power and explained more variability of growth rates.

References

Barro, R., and X. Sala-i Martin (2007): Economic Growth. Prentice Hall of India Private Limited.

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

Chen, T., and C. Guestrin (2016): "XGBoost," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Ciccone, A., and M. Jarociński (2010): "Determinants of Economic Growth: Will Data Tell?,"
American Economic Journal: Macroeconomics, 2(4), 222–46.

Doppelhofer, G., and M. Weeks (2011): "Robust Growth Determinants," CESifo Working Paper Series 3354, CESifo.

Durlauf, S. (2009): "The Rise and Fall of Cross-Country Growth Regressions," History of Political Economy, 41, 315–333.

Fernandez, C., E. Ley, and M. Steel (2001): "Model uncertainty in cross-country growth regressions," Journal of Applied Econometrics, 16(5), 563–576.

Fisher, A., C. Rudin, and F. Dominici (2019): "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously," Journal of Machine Learning Research, 20(177), 1–81.

Friedman, J. H. (2001): "Greedy function approximation: A gradient boosting machine.," Ann.
Statist., 29(5), 1189–1232.

Goldstein, A., A. Kapelner, J. Bleich, and E. Pitkin (2013): "Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation," Journal of Computational and Graphical Statistics, 24.

He, Q., and B. Xu (2019): "Determinants of economic growth: A varying-coefficient path identification approach," Journal of Business Research, 101, 811–818.

Hendry, D. F., and H.-M. Krolzig (2004): "We Ran One Regression," Oxford Bulletin of Economics and Statistics, 66(5), 799–810.

Hofmarcher, P., J. Crespo Cuaresma, B. Grün, and K. Hornik (2015): "Last Night a Shrinkage Saved My Life: Economic Growth, Model Uncertainty and Correlated Regressors," Journal of Forecasting, 34(2), 133–144.

Ley, E., and M. F. Steel (2007): "Jointness in Bayesian variable selection with applications to growth regression," Journal of Macroeconomics, 29(3), 476 – 493, Special Issue on the Empirics of Growth Nonlinearities.

Molnar, C. (2019): Interpretable Machine Learning.

Sala-i Martin, X. (1997): "I Just Ran Two Million Regressions," American Economic Review, 87(2), 178–183.

Sala-i Martin, X., G. Doppelhofer, and R. I. Miller (2004): "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach," American Economic Review, 94(4), 813–835.

Schneider, U., and M. Wagner (2008): "Catching Growth Determinants with the Adaptive LASSO," Economics Series 232, Institute for Advanced Studies.

Smola, A. J., and B. Schölkopf (2004): "A Tutorial on Support Vector Regression," Statistics and Computing, 14(3), 199–222.

Tibshirani, R. (1996): “Regression Shrinkage and Selection via the Lasso,” Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267–288.

Vapnik, V. N. (1995): The Nature of Statistical Learning Theory. Springer-Verlag, Berlin, Heidelberg.

University of Warsaw
Faculty of Economic Sciences
44/50 Długa St.
00-241 Warsaw
www.wne.uw.edu.pl