Technical University of Denmark

02450- Introduction to Machine Learning and Data Minning

Project 2 report - Supervised learning: Classification and regression

Written by

Group 127
Eva Kaštelan - 232469
Noa Margeta - 232470
Filip Penzar - 232452

November 10, 2023

Contribution

Section	Eva Kaštelan	Noa Margeta	Filip Penzar
Regression Part A	40%	30%	30%
Regression Part B	30%	30%	40%
Classification	30%	40%	30%
Discussion	30%	30%	40%
Exam problems	33%	33%	33%

Table 1: Contribution Percentage per Section

Contents

Regression

Part A

For the regression problem, the set goal is to predict the value of the Gross Tertiary Education Enrollment attribute. The dataset contains 16 (standardized) attributes, counting out the attribute that is predicted. The dataset is further altered with one-of-K encoding of each country's (observation's) continent. This is done by adding 6 columns for 6 continents respectively (all apart from Antarctica). Considering the transformation, it sums the total number of attributes up to 22. This amount can be considered large and more importantly irrelevant information for the regression. For this reason Sequential feature selection is implemented to shift through all the dataset's features and select those more often chosen by different models of the Sequential feature selection method. The method is two-leveled cross validation where the inner level performs the feature selection whose final choices of attributes for every fold of the outer loop are shown in Figure 1. Each inner loop iteration's performance and model choice is shown in figures 3-7 in the appendix.

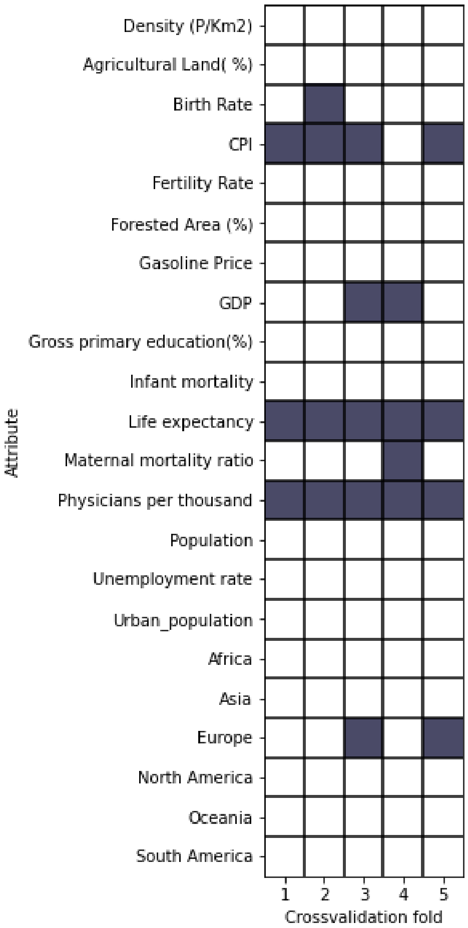

Figure 1: Sequential feature selection for linear regression

After multiple runs of the Sequential feature selection method the attributes most often found in the feature selection's final iteration were roughly estimated. These are: Birth rate, CPI, Fertility rate, GDP, Life expectancy, Maternal mortality ratio, and Physicians per thousand people. They are used for the following regression models, both in Part A and Part B.

As a complexity-controlling variable, the regularization parameter λ is introduced for the linear regression model. When λ is small, the optimal weights are large indicating high variance but low bias and making the model more prone to overfitting. On the other hand, when λ is set to larger values, the weights become smaller indicating lower variance but higher bias in the solutions. Exactly this property of the changing regularization constant is easily seen from the left panel of Figure 2 showing the mean coefficient values of the optimal weights for different values of λ. Finding the optimal λ (the one with the smallest generalization error) is done with 10-fold cross validation. The range of values considered for λ is $[10^{-3}, 10^{6}]$. Figure 2's right panel shows the value of the generalization error for the different values of λ.

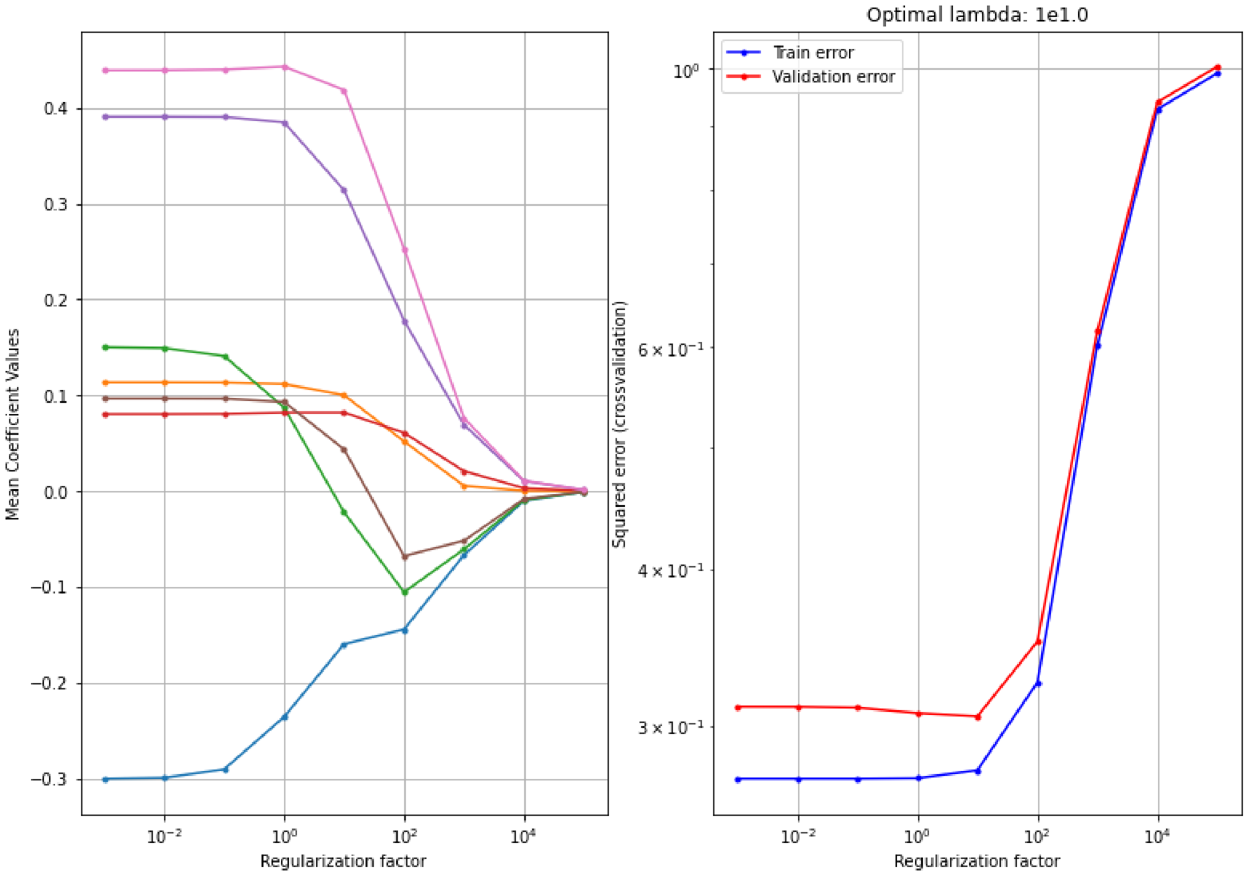

Figure 2: Choosing the regularization factor value

The lowest generalization error is with $\lambda = 10$, consequentially it is the optimal value of λ and the one used in the linear regression model.

The model learns the optimal values for the weights and bias during the training process to minimize the generalization error. What the linear model does when given an input x is, essentially, it computes a weighted sum of the input's features, and then the bias term is added to this sum. The result of the whole sum is then the wanted output and prediction ŷ.

$$\hat{y} = \sum_{i=1}^{n} w_i x_i + b$$

The equation of the linear regression model with the smallest generalization error determined previously in this section is as follows:

x_i	attribute
x_3	birth rate
x_4	CPI
x_5	fertility rate
x_8	GDP
x_{11}	life expectancy
x_{12}	maternal mortality
x_{13}	physicians per 1000

$$\begin{aligned} y = -0.000003 - 0.231 \times x_3 + 0.112 \times x_4 \\ +0.084 \times x_5 + 0.084 \times x_8 + 0.0383 \times x_{11} \\ +0.0091 \times x_{12} + 0.4445 \times x_{13} \end{aligned} \tag{1}$$

The given formula shows the effect of each individual attribute of the input (of the attributes chosen for the regression with the sequential feature selection) on the output. From it we can see that the attributes Physicians per 1000 people and Birth rate carry the largest (in magnitude) weights and therefore have the greatest influence over the output. This correlates with the correlation matrix (Figure 8 in appendix) where these two attributes are shown to correlate most with the tertiary enrollment attribute.

Part B

To investigate different machine learning methods and their effectiveness, two-layer cross validation was performed. Three models were used: Artificial Neural Network (ANN), regularized linear regression model and a baseline model. The baseline model simply outputted the mean of the training data regardless of the input. For the ANN, the complexity-controlling parameter used was h, the number of hidden units. h was chosen from the set $\{1, 10, 20\}$, to represent a low, medium and high complexity ANN. The transfer function used in the hidden layer was $tanh(x)$. There was no transfer function used in the output layer because this was a regression problem. For the regularized linear regression model, the range of values considered for λ was $[10^{-3}, 10^{6}]$ (with a 10 factor step).

For the two-level cross validation, $K_1 = K_2 = 10$ was used for the outer and inner loop folds respectively. Each model type was evaluated individually - for the ANN, h associated with the lowest mean of the mean squared error on the inner test folds was chosen for the outer fold. Same was true for the regularized linear regression but with respect to λ. The train/test splits D_i^{par}/D_i^{test} were re-used for all three methods to allow for a statistical comparison. The results can be seen in Table 2.

Outer fold	ANN		Linear regression		baseline
i	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	20	0.8108	10.0	0.6094	1.9051
2	20	0.3908	10.0	0.2133	1.0220
3	20	0.8707	1.0	0.5612	0.9818
4	20	0.8382	1.0	0.6094	0.7922
5	20	0.2454	10.0	0.2015	0.6968
6	20	0.2499	10.0	0.1689	0.9892
7	20	0.4913	1.0	0.2326	0.4361
8	20	0.2271	10.0	0.1383	0.9511
9	20	0.5132	10.0	0.1824	0.4987
10	10	0.4815	1.0	0.1856	1.2144

Table 2: Two-level cross validation for the regression problem

In 90% of the cases, ANN with $h = 20$ was chosen. It can therefore be assumed that the higher complexity ANN performed better then the lower complexity ones. The chosen λ^* for the linear regression model was 10 in 60% of the runs, and 1 in the other 40%. This confirms the assumption made in Part A for the optimal λ.

Furthermore, the models' performance (E_i^{test} from Table 2) were statistically evaluated to determine if there was a significant difference between them. Paired $t - test$ was used on the models pairwise (ANN vs. linear regression, ANN vs. baseline, linear regression vs. baseline) with $\alpha = 0.05$. The results of the tests can be seen in Table 3.

Method A	Method B	Conf. int.	p-value
ANN	Lin. reg.	(0.1286, 0.2747)	1.512e-4
ANN	baseline	(-0.7305, -0.1432)	8.325e-3
Lin. reg.	baseline	(-0.9046, -0.3724)	4.176e-4

Table 3: Statistical evaluation of the regression models

The z values used in the tests were computed as $E_i^A - E_i^B$ where E_i^A corresponds to E_i^{test} of Model A and E_i^B corresponds to E_i^{test} of Model B. Statistically significant difference can be seen in the performance of the ANN and linear regression. Linear regression model outperformed ANN - the small p-value and a positive confidence interval not containing 0 both indicate this. There is also a statistically significant difference in the performance of both ANN and linear regression model compared to the baseline model. In both cases, the trained models did better than the baseline - indicated by the small p-value and a negative confidence interval not containing 0.

A possible explanation of why the linear regression outperformed ANN is that the ANN did not have enough data to train on since the dataset contained only 159 observations.

Providing additional training time only resulted in over-fitting and poor performance on the test dataset. Linear regression model on the other hand, performed well due to the strong linear correlation between Gross Teritary Education Enrollment and other attributes as can be seen in the Correlation Matrix (Fig. 8). As was expected, both models were better than the baseline. This goes to show that even the relatively simple models are able to significantly reduce the error and improve the overall accuracy of predictions.

Classification

For the classification problem, the goal is to create a model that can predict the continent a country is on, based on all of the datasets' features. The features are: GDP, Life Expectancy, Physicians per thousand people, Gross Primary Education Enrollment, Gross Tertiary Education Enrollment, Infant Mortality, CPI, Agricultural Land, Density, Birth rate, Forested area, Gasoline price, Maternal mortality ratio, Unemployment rate, Urban population, Population, Physicians per thousand people. It is a multi-class classification problem and there are 6 classes: Africa, Asia, Europe, North America, Oceania and South America.

Three different methods are used for the classification problem: a baseline, Logistic regression and K-nearest neighbours classification. The baseline is a model that computes the largest class on the training set and predicts all test data to belong to that class. For the Logistic regression and K-nearest neighbours method complexity-controlling parameters are introduced: λ (regularization constant) and k (the number of neighbours) respectively. The range of values considered for λ is $[10^{-3}, 10^{6}]$ and the range of values considered for k is $[1, 9]$. As in the regression Part B, two-level cross validation is used for comparing the models' error rates and the values of the complexity-controlling parameters (inner level) and then evaluating the best chosen model's performance (outer level). The results are shown in Table 4.

Outer fold	KNN		Logistic regression		baseline
i	k_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	3	0.1875	0.01	0.25	0.75
2	5	0.4375	10.0	0.375	0.6875
3	2	0.3125	0.001	0.4375	0.875
4	3	0.3125	0.001	0.375	0.875
5	1	0.5	0.1	0.4375	0.75
6	6	0.3125	10.0	0.25	0.6875
7	8	0.5625	0.1	0.5	0.8125
8	6	0.375	0.001	0.5	0.8125
9	9	0.3125	1.0	0.375	0.8125
10	2	0.4667	100.0	0.4	0.4667

Table 4: Two-level cross validation for classification problem

From the error values computed as: $E = \frac{\text{Number of misclassified observations}}{N^{test}}$, it can approximately be seen that the KNN and Logistic regression models have similar results and that their error rates are notably lower than those of the baseline model. The accuracies of the trained models that are displayed in Table 5 also coincide with these hypotheses.

Model	Accuracy
KNN	64.1509%
Log. reg.	60.3774%
baseline	22.0126%

Table 5: Classification model accuracies

In the following part the hypotheses are tested using a statistical method and are shown to be true.

McNemera's test is used to compare the models. The results are given in Table 6.

Method A	Method B	Conf. int.	p-value
KNN	Log. reg.	(-0.0316, 0.1069)	0.3771
KNN	baseline	(0.3319, 0.5069)	9.9951e-16
Log. reg.	baseline	(0.2937, 0.4701)	1.1671e-13

Table 6: Statistical evaluation of the classification models

With the test results of KNN and Logistic regression a statistically significant inference cannot be made as the p-value is very high. However, from the test results of comparing KNN and Logistic regression to the baseline it can be concluded that KNN and Logistic regression models are statistically significantly better than the baseline model because the confidence intervals are far away from 0 and the p-values are very small.

The logistic regression model is a multinomial model, meaning it decides the output out of more than two different classes. To do this, the model learns the optimal values for the weights and bias during the training process for each of the possible classes, i.e., a separate linear combination of the input features for each single category. When given an input, the linear combination for each of those classes is individually computed. Then the softmax function is applied to each of the linear combinations to obtain the probabilities for the classes, i.e., the probability distribution over all classes.

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad for\ i = 1, 2, \ldots, K$$

Finally, the output is equal to the class with the highest probability of the calculated distribution. The linear combinations for the different classes of our model are shown in the appendix 2-7. The relevancy of these are incomparable with the regression model in Part A of the report since for the classification all the attributes were used for the prediction whereas with the sequential feature selection only some of the attribute were chosen for the regression model. Another difference is in the goals of the two tasks. It's essential to consider these distinctions when evaluating the results and their implications.

Discussion

Within the scope of the regression problem, linear regression turned out to be the best fitting model. This was due to the fact that Gross Tertiary Education was highly linearly correlated to some of the other attributes, as seen in the Correlation matrix (Fig. 8) and further confirmed with the Sequential fetaure selection (Fig. 1). If the correlation between the targeted variable and some of the other variables was not so strongly linear in nature, but rather, for example, exponential, our guess is that ANNs would perform much better compared to linear regression. This is because ANNs have a non-linear transfer function that sets them apart from linear regression and allows the model to infer non-linear relations between variables. An ANN without a non-linear transfer function can be shown to be the same as a linear regression model.

KNN and logistic linear regression performed equally well in the classification problem. As was discussed in Report 1, observations corresponding to the same continent clustered together in the high dimensional space spanned by the attributes. Because of this, KNN achieved reasonably high accuracy - if data was not nicely clustered classification with KNN would be much harder. Similar line of reasoning can be made for the accuracy of the logistic regression model.

In both problems, cross validation played a major role in choosing the best model parameters for the task. It also allowed for the estimation of the generalization error and model comparison. With the data obtained from two-level cross validation, it was possible to use statistical methods and compare different models.

Unfortunately, we did not find any relevant articles connected to classification and regression problems and our dataset. The reason is probably the specific nature of the classification and the regression problem we solved. The closest paper we found was *Predicting and explaining corruption across countries: A machine learning approach, M. Lima and D. Delen* but it used a different dataset and its goals were different - the prediction of corruption across countries. The paper successfully used ANN, Random Forest and SVD for the classification problem. It also used K-fold cross validation.

To sum up, non-trivial models outperformed the baseline model in both regression and classification. Because of this, we conclude this project was successful in developing machine learning models for classification and regression problems set out in Report 1.

Exam Problems

Question 1

Option C: Setting the threshold value to 0.8 shows that B is incorrect as it has TPR: 0.5 and FPR: 0. Next setting the threshold value to 0.75 the values for A and B are TPR: 0.5 FPR: 0.25 which is incorrect, leaving only C. Checking all other threshold values for C we confirm it indeed is the correct answer.

Question 2

Option C: The split is made according to x7 = 2 and x7 != 2. This leaves 1 observation for x7 = 2 and 134 obs. for x7 != 2. By using the formula for impurity gain and the classification error we get: 98/135 - (134/135)*(97/134) = 1/135 = 0.0074.

Question 3

Option A: The network has 7 input neurons, 10 hidden neurons and 4 output neurons. The network thus contains (7+1) * 10 + (10 + 1) * 4 = 80 + 44 = 124 parameters to train.

Question 4

Option D: Just by looking at the Congestion level 4, it is clear that it is classified when b1 > -0.16 on the C split. This is only true for the C option.

Question 5

Option C: The formula for a model is: K1*(n*K2*(train + test) + train + test) where n is either the number of different ANNs or different lambda values. In our case n = 5 for both types of models. We get the following formula: 105 * (train + test) for a model type. For both models it is 105 * ($train_ANN$ + $test_ANN$ + $train_LLR$ + $test_LLR$) = 105 *(20 + 5 + 8 + 1) = 3570 ms.

Question 6

Option B: If 1 > $e^{\hat{y}_k}$ for k in {1,2,3} then y=4 would be assigned. This means $\hat{y}_k$ must be < 0 for all k. By plugging in the numbers, we see this happen only in option B.

Appendix

Figures

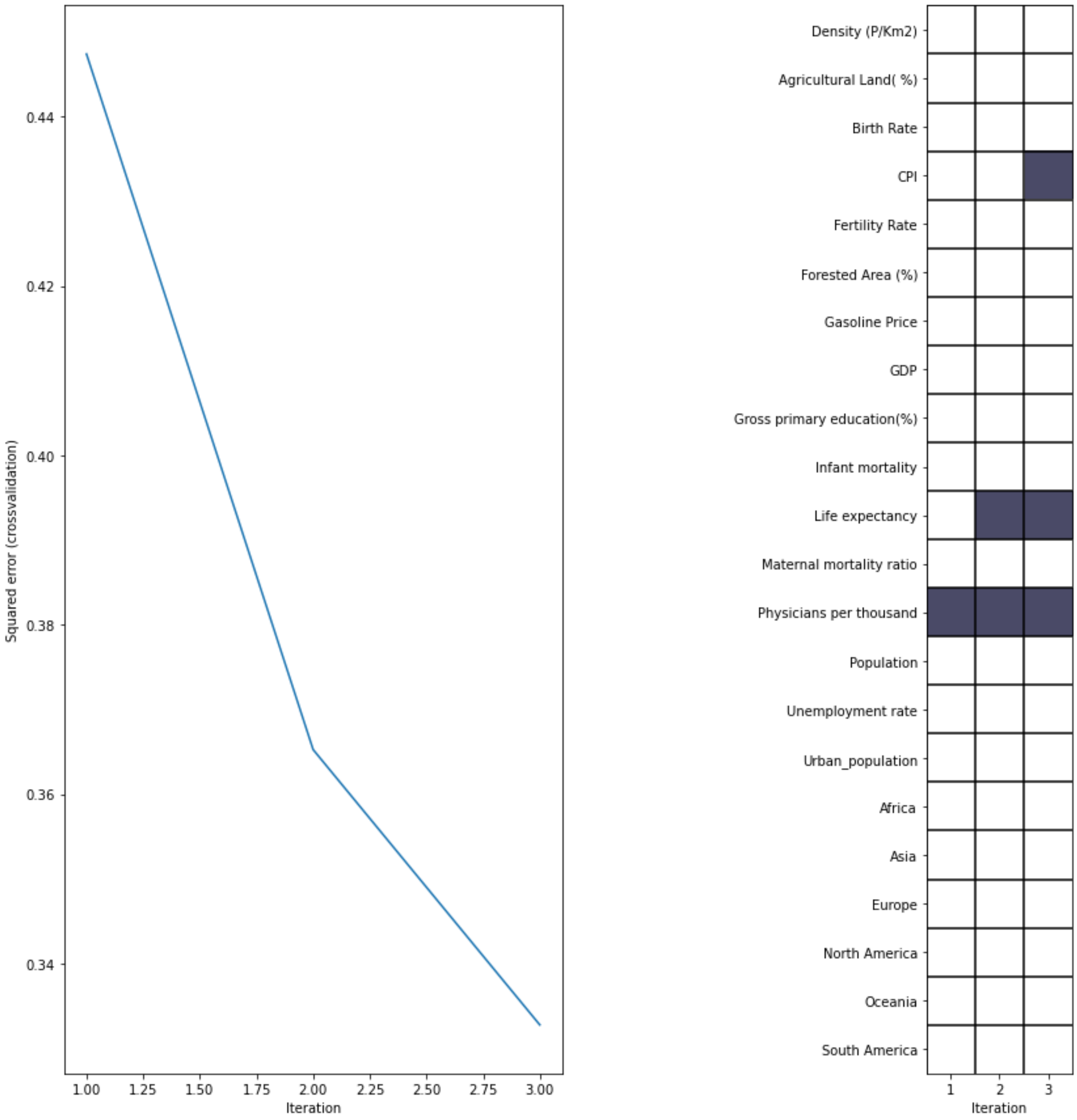

Figure 3: 1st model of sequential feature selection

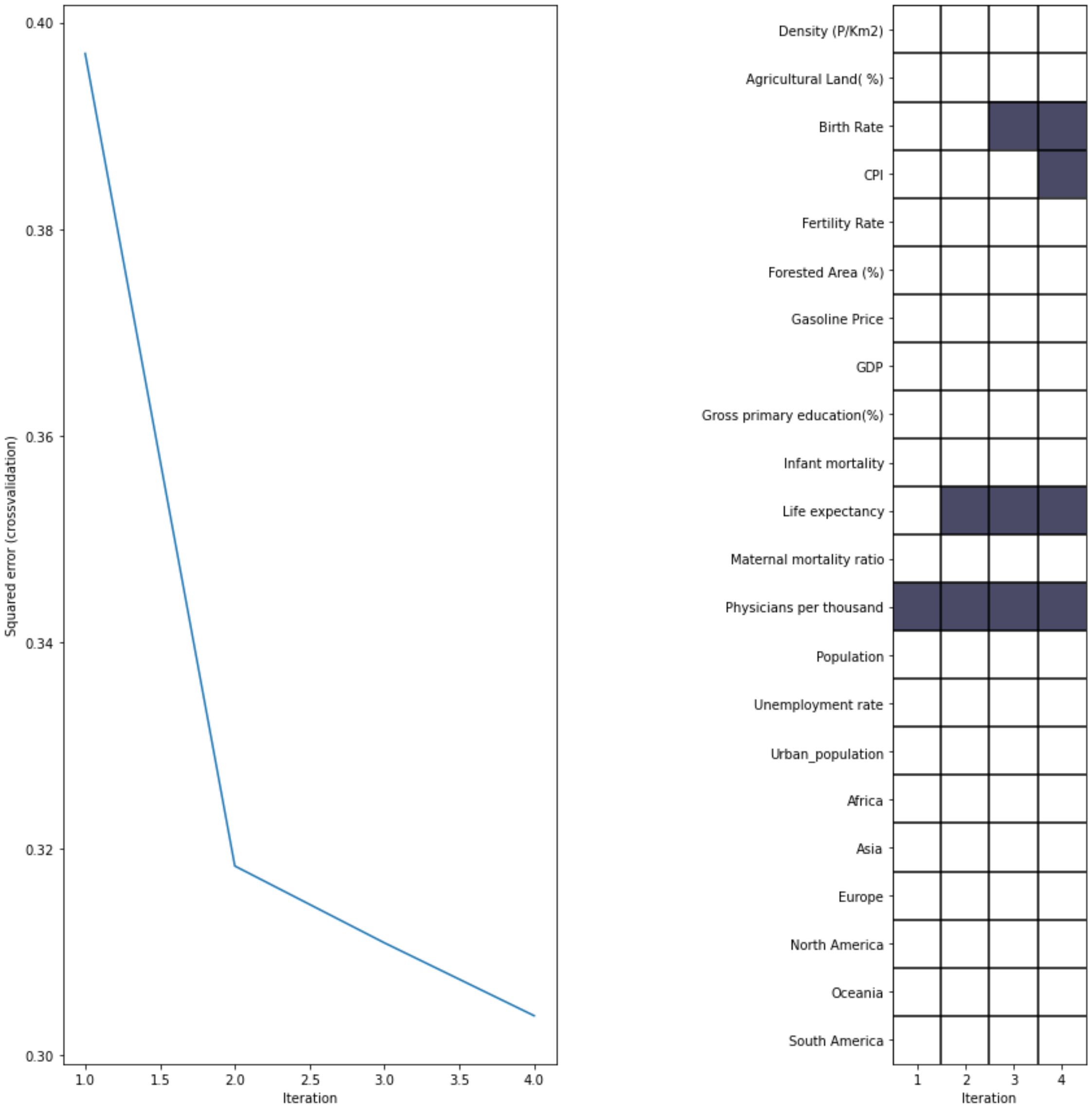

Figure 4: 2nd model of sequential feature selection

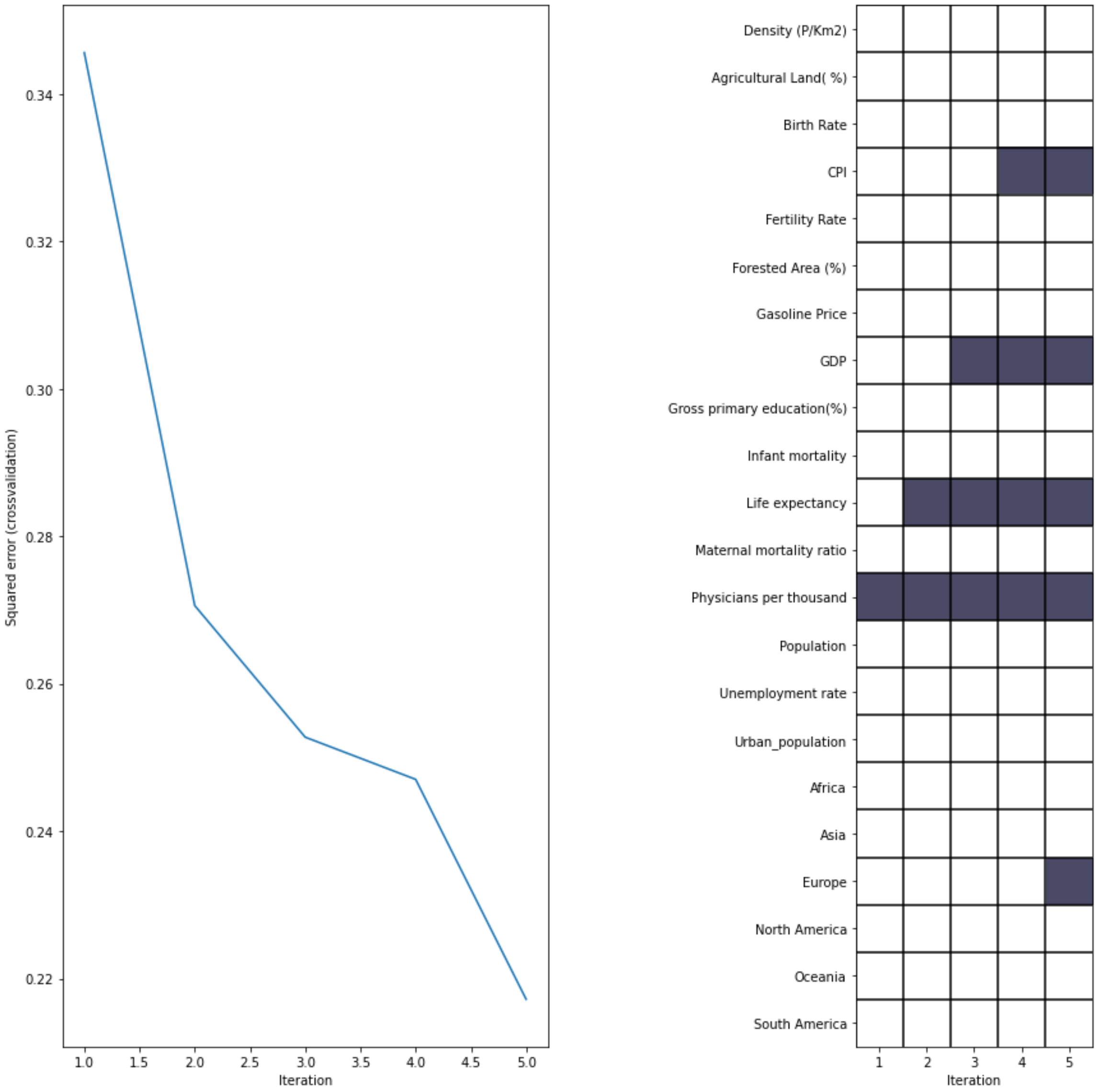

Figure 5: 3rd model of sequential feature selection

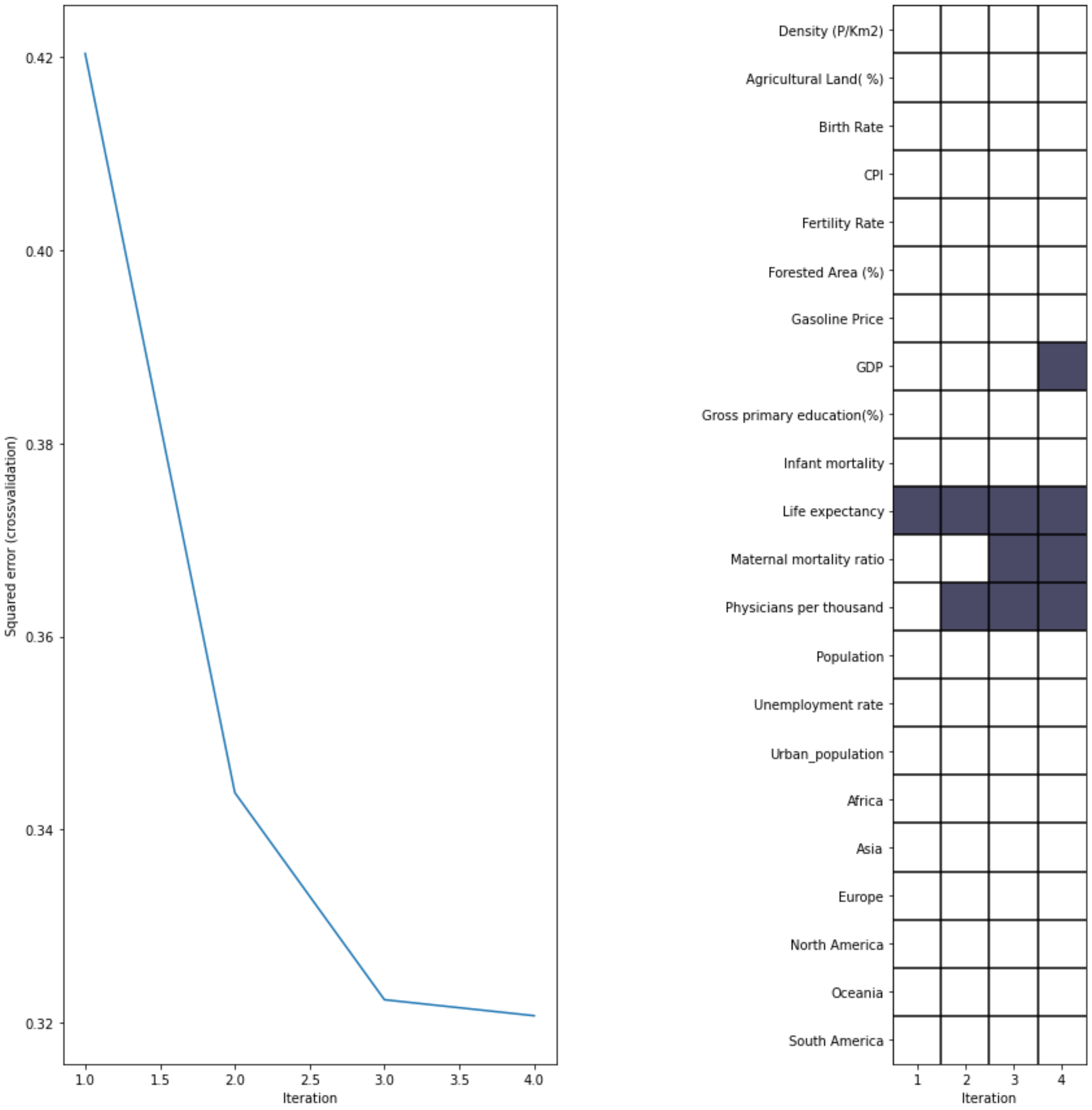

Figure 6: 4th model of sequential feature selection

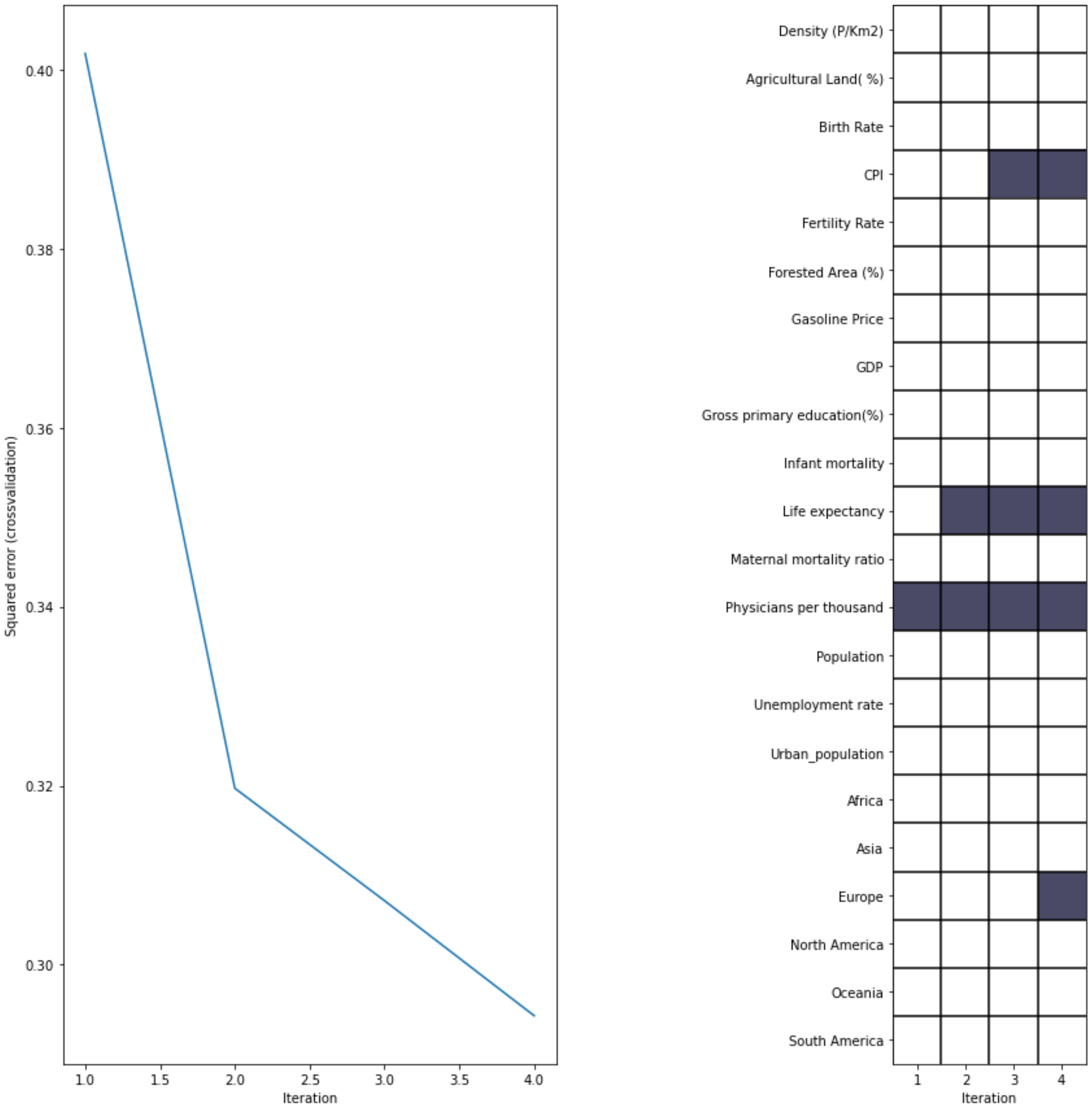

Figure 7: 5th model of sequential feature selection

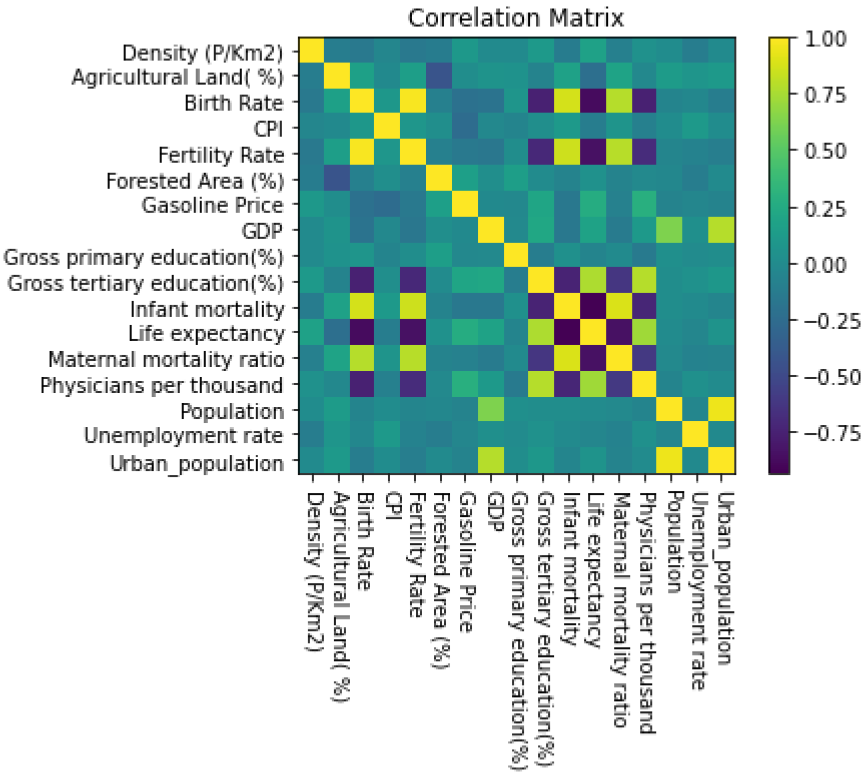

Figure 8: Correlation matrix

$$
\begin{aligned}
y_{Africa} =& 1.8829 - 0.0009 \times x_1 + 1.7773 \times x_2 - 0.4338 \times x_3 \\
&+ 0.7653 \times x_4 + 0.8862 \times x_5 + 1.6322 \times x_6 \\
&- 0.5831 \times x_7 - 0.4761 \times x_8 - 1.1277 \times x_9 \\
&+ 0.1707 \times x_{10} + 1.1365 \times x_{11} + 0.0359 \times x_{12} \\
&- 0.0358 \times x_{13} + 1.8228 \times x_{14} - 1.5197 \times x_{15} \\
&+ 0.9021 \times x_{16} + 1.0413 \times x_{17} - 0.7031 \times x_{18}
\end{aligned} \tag{2}
$$

$$
\begin{aligned}
y_{Asia} =& 3.5111 - 0.0015 \times x_1 + 2.2502 \times x_2 - 0.5621 \times x_3 \\
&+ 0.5081 \times x_4 + 0.7078 \times x_5 + 0.0869 \times x_6 \\
&- 0.6955 \times x_7 - 1.0411 \times x_8 - 0.8437 \times x_9 \\
&- 0.5939 \times x_{10} - 0.5126 \times x_{11} + 1.8231 \times x_{12} \\
&+ 0.5939 \times x_{13} + 0.1642 \times x_{14} + 0.4674 \times x_{15} \\
&+ 1.3461 \times x_{16} + 0.0205 \times x_{17} - 0.3889 \times x_{18}
\end{aligned} \tag{3}
$$

$$\begin{aligned}y_{Europe} =& -5.7246 - 0.0004 \times x_1 + 0.1772 \times x_2 + 0.3949 \times x_3 \\ & - 3.7698 \times x_4 - 1.2124 \times x_5 - 2.2579 \times x_6 \\ & - 0.1884 \times x_7 + 1.0343 \times x_8 - 0.6347 \times x_9 \\ & - 0.8231 \times x_{10} - 0.6846 \times x_{11} - 2.8313 \times x_{12} \\ & - 2.1019 \times x_{13} - 2.3629 \times x_{14} + 1.2620 \times x_{15} \\ & + 0.7640 \times x_{16} - 0.6175 \times x_{17} + 0.7903 \times x_{18}\end{aligned} \tag{4}$$

$$\begin{aligned}y_{N.America} =& 1.7886 - 0.0007 \times x_1 + 1.1630 \times x_2 - 0.1860 \times x_3 \\ & - 1.8023 \times x_4 - 0.6176 \times x_5 - 2.2239 \times x_6 \\ & - 0.2831 \times x_7 - 4.8182 \times x_8 + 1.6256 \times x_9 \\ & - 0.6012 \times x_{10} - 1.5982 \times x_{11} - 1.3601 \times x_{12} \\ & + 1.9336 \times x_{13} + 1.8098 \times x_{14} + 0.3650 \times x_{15} \\ & - 1.3345 \times x_{16} + 0.3595 \times x_{17} + 0.1559 \times x_{18}\end{aligned} \tag{5}$$

$$\begin{aligned}y_{Oceania} =& -1.9522 + 0.0032 \times x_1 - 2.2784 \times x_2 + 0.2710 \times x_3 \\ & - 0.4555 \times x_4 - 1.4281 \times x_5 + 3.2834 \times x_6 \\ & + 0.7347 \times x_7 + 0.7570 \times x_8 + 1.3726 \times x_9 \\ & + 0.0074 \times x_{10} + 1.2315 \times x_{11} - 0.2870 \times x_{12} \\ & - 0.8882 \times x_{13} - 3.2740 \times x_{14} - 0.8456 \times x_{15} \\ & - 1.2650 \times x_{16} - 1.0840 \times x_{17} - 1.4824 \times x_{18}\end{aligned} \tag{6}$$

$$\begin{aligned}y_{S.America} =& 0.4942 - 0.0001 \times x_1 - 3.0902 \times x_2 + 0.5149 \times x_3 \\ & + 1.1496 \times x_4 + 1.6012 \times x_5 - 0.5207 \times x_6 \\ & + 0.9765 \times x_7 + 0.1016 \times x_8 - 2.0795 \times x_9 \\ & + 0.1068 \times x_{10} + 0.4275 \times x_{11} - 0.0941 \times x_{12} \\ & + 0.8204 \times x_{13} + 1.8401 \times x_{14} + 0.2709 \times x_{15} \\ & - 0.4134 \times x_{16} + 0.2790 \times x_{17} + 1.6281 \times x_{18}\end{aligned} \tag{7}$$