Machine Learning – Course Project

2022/2023

Group 6 – Alexandre Sobreira (59451) André Dias (59452) Nihan Ahat (61010) Tiago Rodrigues (49593)

Hours contributed – Alexandre: 20h; André: 20h; Nihan: 20h; Tiago: 20h

Country population, fertility rate and life expectancy are valuable indicators that are interconnected in some way, given that they all indicate how populations evolve. Regarding these indicators, country population refers to the total number of residents regardless of legal status or citizenship (midyear estimates). Fertility rate to the number of children a woman gives birth to during her childbearing years. Life expectancy refers to the number of years a newborn would live if the patterns of mortality at the time of birth remain the same throughout his life.

Three comma-separated values (CSV) files were provided, each containing data from 1960 to 2016 for several countries or aggregates of countries (e.g., East Asia & Pacific). The main objective of this course project is to make a prediction model for each of the three indicators for 2017 and 2018 and validate each model with the real values for ten random countries.

Considering that the data's values evolve over time, the best type of model to use is the regression model. As such, the following algorithms were chosen: Linear Regression (Lr), Random Forests Regression (RFr) and Extreme Gradient Boost Regression (XGBr). Lr was chosen since it is a straightforward and optimal regression model typically used in predictive analysis. The RFr, since it introduces a greater tree diversity compared to decision trees, therefore trading higher bias for lower variance, yielding an overall better model. XGBr is considered a powerful approach for building supervised regression models, which easily handles structured data, is scalable and is regarded as a highly accurate implementation of gradient boosting.

Initially, the CSV files were read using pandas, where it was possible to visualise all the variables in each dataset: "Country Name", "Country Code", "Indicator Name", "Indicator Code", and years from 1960 to 2016 with the value of a specific indicator. The columns "Country Code", "Indicator Name", and "Indicator Code" were dropped since they are uninformative attributes.

The three data sets were in a wide format, meaning that each row related to one country/aggregate, and each column to each year (1960-2016), with information about an indicator according to each dataset.

Considering the nature of this project's goals, a dataset transformation was required. One possible transformation is to change the datasets from wide to long format. In the long format, each row corresponds to the value of an indicator in a given year for a specific country/aggregate. After this transformation, the datasets had with three columns: "Country Name", "Year", and the indicator column. For this procedure, the melt function was applied to all the datasets.

Since the "Country Name" was a categorical variable, it cannot be used as given. Either the variable had to be removed or One-Hot-Encoding (OHE) applied. With OHE, each categorical value is converted into a new categorical column and assigned a binary value of one or zero. In this case, one indicated the presence of a specific country and zero its absence. OHE was applied because the model obtained would not make sense given that only the variable "Year" would be present in the training data, with several rows for the same year.

For each dataset, it was observed that there were null values in specific years for given countries. The long format structure has one significant advantage when dealing with null values since null dropping will only affect the missing year for the specific country/aggregate. On the contrary, if a wide format was maintained, the complete information of the country/aggregate would be lost, resulting in data reduction. Considering the latter and that imputation for year values would lead to unnecessary bias, all the null values were removed.

In addition to OHE and dropping null values, other pre-processing steps were considered: removing the aggregates in the datasets; creating a Delta variable which consists of the difference between an indicator on two consecutive years; reducing the datasets to years from 2001 to 2016.

Before the three chosen regression algorithms could be applied, each data set was divided into X_train, containing the independent variables ("Year", "Country Name" encoding) and y_train, containing the dependent variable (country population or fertility rate or life expectancy). After this step, the model training and evaluation proceeded.

For model validation, simple cross-validation was used, and ten countries were randomly chosen from the datasets after the removal of all the aggregates. This was done to allow testing models without aggregates in the forthcoming work. Data for each indicator from 2017 and 2018 were downloaded from the World Bank Open Data. Aggregate dropping and Delta variable application were considered to be in accordance with what had been done for the training set. From each indicator CSV, two dataframes were created with "Country Name" and the corresponding value of the year for the ten randomly selected countries. For X_test, an array of arrays containing the year to predict (2017 or 2018) and one hot encoding of the respective country was created. For y_test, an array with the truth values of the corresponding indicator for each country and year was created.

The ratio of variance explained (RVE) was used to assess each model's quality. This ratio measures the proportion of variation (dispersion) explained by a given model. Table 1 shows the values of RVE obtained for the selected models for each indicator with the OHE method.

Table 1 – *RVE scores obtained for the applied models for each indicator and year, only applying OHE.*

OHE		Linear Regression			Random Forests (Regression)			XGBoost (Regression)		
		Population	Life Expectancy	Fertility	Population	Life Expectancy	Fertility	Population	Life Expectancy	Fertility
RVE	2017	0,866	0,805	0,444	0,999	0,965	0,921	0,985	0,917	0,790
	2018	0,861	0,786	0,443	0,999	0,959	0,917	0,983	0,921	0,787

From the Table, it is possible to observe that RFr produces the best models for all the indicators. Another important conclusion is that the fertility rate indicator fits worst models than the others, which is especially evidenced in Lr.

The observed results could be considered a good enough indicator of the model's capability of producing accurate forecasts for the years to come. Still, other data pre-processing steps could be applied to further enhance the models.

As previously mentioned, elements that represent groups of countries (aggregates) were present in the datasets. Since the objective is to create a model for specific countries, it was considered that these instances could lead to bias in the models. With this in mind, the models were trained and validated with datasets only containing the countries and not the aggregates. The RVE values obtained when dropping the aggregates are present in Table 2.

Table 2 – *RVE scores obtained for the applied models for each indicator, year, applying OHE and aggregate removal.*

OHE; drop_regions		Linear Regression			Random Forests (Regression)			XGBoost (Regression)		
		Population	Life Expectancy	Fertility	Population	Life Expectancy	Fertility	Population	Life Expectancy	Fertility
RVE	2017	0,866	0,805	0,444	0,999	0,963	0,922	0,998	0,934	0,877
	2018	0,861	0,786	0,443	0,999	0,957	0,917	0,997	0,934	0,873

Doing this step did not significantly change the Lr and RFr performance. On the other hand, XGBr saw a slight increase in performance for country population and life expectancy indicators and a considerable increase for the fertility rate.

Following this, usage of the delta variable was evaluated, by applying the ".diff" function to the year columns of the original dataset. The results obtained can be seen in Table 3.

Table 3 – *RVE scores obtained for the applied models for each indicator and year, applying OHE, aggregate removal and delta variable.*

OHE; drop_regions; delta		Linear Regression			Random Forests (Regression)			XGBoost (Regression)		
		Population	Life Expectancy	Fertility	Population	Life Expectancy	Fertility	Population	Life Expectancy	Fertility
RVE	2017	0,998	-0,115	-0,091	0,999	0,318	-2,177	0,999	0,175	-1,290
	2018	0,992	0,061	-0,427	0,998	0,124	-4,466	0,998	0,159	-2,907

It is possible to see that the models obtained for the life expectancy and fertility indicators are worse than the ones previously obtained and, in some cases, the RVE becomes negative. When the error is greater than the total variance of the y, the RVE equation can compute a negative value if the value of the coefficient is greater than 1. This could be because the difference between sequent years is sometimes negative, meaning that the variable decreases compared to the previous year. This can influence the model's ability to predict accurately.

Despite this, the models obtained for the population indicator were excellent, especially Lr, which saw a significant improvement over the previously tested models.

This diverging behaviour between different indicators could be due to the higher values of the latter variable. Unlike population, whose values are in the order of hundreds of thousands or millions, the other indicators have much lower ranges, especially fertility, with values ranging from 1 to about 7, meaning a smaller variation. When the delta variable is applied, this may lead to an even reduced amount of information. Because of this, it was thought not to include the delta variable in the forthcoming work since it hinders the performance of the majority of the models.

A factor that could also influence the results is the presence of years before 2000, given that there was a bigger variability of the indicators before this time, due to several historical events, such as the post-World War 2 Boom. This older data can lead to a bigger dispersion of the indicators resulting in less accurate models. As such, models with only data from 2001 to 2016 were tested by selecting row where year was greater than 2000. The results obtained can be seen in Table 4.

Table 4 – *RVE scores obtained for the applied models for each indicator and year, applying OHE, aggregate removal and data from 2001 to 2016.*

OHE; drop_regions; less		Linear Regression			Random Forests (Regression)			XGBoost (Regression)		
		Population	Life Expectancy	Fertility	Population	Life Expectancy	Fertility	Population	Life Expectancy	Fertility
RVE	2017	0,988	0,965	0,902	0,999	0,965	0,922	0,999	0,953	0,872
	2018	0,986	0,948	0,896	0,998	0,958	0,918	0,998	0,952	0,865

From the presented results, it is possible to see a significant increase in the performance of the Lr models due to the removal of the years before 2001. XGBr and RFr were not as affected, possibly due to their ensembled nature, which uses sub-samples for various sub-models, likely reducing the influence of bias present in the data.

The model's performance can also be enhanced by adding more data, in this case from the World Bank Open Data. Two variables were chosen to be added to all the datasets, the income of a country ("Income"), and their urban population ("urban_pop"). The results obtained by introducing these variables are presented in Table 5.

Table 5 – *RVE scores obtained for the applied models for each indicator and year, applying OHE, aggregate removal, data from 2001 to 2016 and with the introduction of the "Income" and "urban_pop" variables.*

OHE; drop_regions; less			Linear Regression			Random Forests (Regression)			XGBoost (Regression)		
			Population	Life Expectancy	Fertility	Population	Life Expectancy	Fertility	Population	Life Expectancy	Fertility
RVE	Urban Pop	2017	0,995	0,965	0,901	0,999	0,964	0,923	0,999	0,942	0,873
		2018	0,994	0,947	0,895	0,998	0,953	0,917	0,999	0,937	0,868
	Uban Pop + Income	2017	0,995	0,974	0,883	0,999	0,940	0,923	0,999	0,942	0,875
		2018	0,995	0,957	0,864	0,999	0,867	0,918	0,999	0,927	0,890

From Table 5, it is possible to see there were some performance improvements and declines depending on the model and indicator. The best overall model for the life expectancy indicator was obtained using the two extra variables with Lr. For fertility, the best model was also obtained using the two extra variables with RFr. For country population, no improvements were seen from the models obtained at this phase.

For the best models for each indicator (life expectancy – Lr with OHE, aggregates removal, data from 2001 to 2016 and the two extra variables; fertility rate – RFr with OHE, aggregates removal, data from 2001 to 2016 and the two extra variables) hyperparameter tunning was performed. However, this was not considered for the best model of country population since the RVE score of the model (RFr with OHE) was already extremely high (0.999).

The results obtained from hyperparameter tuning can be seen in Table 6.

Table 6 – *RVE scores obtained for the applied models for each indicator and year, applying OHE, aggregate removal, data from 2001 to 2016 and with the introduction of the "Income" and "urban_pop" variables after applying GridSearchCV on the hyperparameters displayed. The hyperparameters that produced the best model are highlighted.*

OHE; drop_regions; less; with "Income" and "urban_pop"	Hyperparameters	RVE	
		2017	2018
Linear Regression for Life Expectancy	fit_intercept:[**True**, False], copy_X:[**True**, False]	0,974	0,957
Random Forests for Fertility Rate	n_estimators: [50, 100, **150**], criterion: [squared_error, friedman_mse, **poisson**], max_depth: [5, 10, **None**]	0,927	0,932

From the Table, it can be concluded that the Lr model for life expectancy did not see increased performance after model tunning, since the best parameters were the default ones.

The RFr model for fertility rate improved after tunning with the following parameters: n_estimators: 150, criterion: poisson, max_depth: None.

The truth-predict plots for the best model obtained for the prediction of 2017 for each of the indicators is presented in Figure 1.

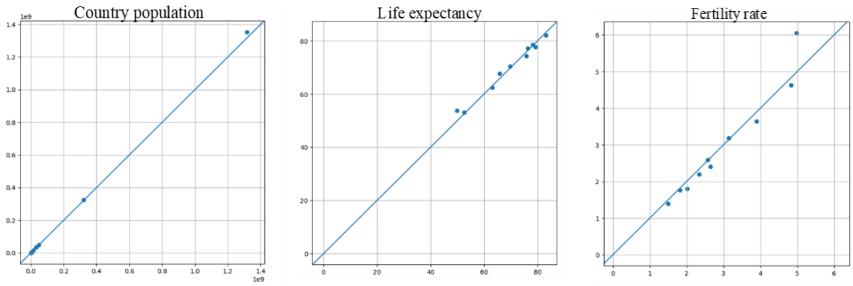

***Figure 1** – Truth-predict plots for the best models of each of the indicators.*

The truth-predict plots are indicative of the model's performance. A bigger distance from each point to the 45º line corresponds to a higher error, indicating worse models. As seen from Figure 1, the indicator country population has the best predictions (points closer to the 45º line), while the fertility rate has a bigger dispersion around the 45º line. This corroborates the RVE scores obtained for each model. The truth and predicted values of the models can be seen below:

Table 7 – Truth and predicted values for the best model of each indicator.

Country	Country Population				Life Expectancy				Fertility			
	2017		2018		2017		2018		2017		2018	
	Predicted	Truth	Predicted	Truth	Predicted	Truth	Predicted	Truth	Predicted	Truth	Predicted	Truth
United States	322068004	325122128	322068004	326838199	78.22	78.54	77.91	78.64	1.82	1.77	1.82	1.73
Uruguay	3436253	3422200	3436253	3427042	79.20	77.63	79.55	77.61	2.00	1.80	2.01	1.66
Central African Republic	4574948	4996741	4574948	5094780	49.93	53.72	50.27	54.37	4.89	6.05	4.88	6.04
Seychelles	93680	95843	93680	96762	75.83	74.30	76.17	72.84	2.44	2.41	2.42	2.41
Senegal	15173337	15157793	15173337	15574909	65.84	67.75	66.19	68.10	4.89	4.62	4.82	4.58
Saudi Arabia	31816823	34193122	31816823	35018133	76.39	77.16	76.67	77.21	2.58	2.58	2.59	2.55
India	1316729080	1354195680	1316729080	1369003306	69.64	70.47	70.17	70.71	2.34	2.20	2.34	2.18
Lesotho	2187609	2170617	2187609	2198017	52.60	53.06	52.94	53.73	3.16	3.19	3.19	3.14
Luxembourg	573032	596336	573032	607950	83.11	82.10	83.45	82.30	1.49	1.39	1.49	1.38
Kenya	47866042	48948137	47866042	49953304	63.19	62.48	63.53	62.68	3.69	3.64	3.39	3.58

From the table presented, it is possible to see that, overall, the predictions are close to the truth values. Despite this, a critical problem was assessed. The predictions for 2017 and 2018 were the same for the best model of country population. After further investigation, it was concluded that unlike Lr, for RFr and XGBr with no extra variables, the predictions for different years were always the same, indicating that the year was not taken into consideration by these models. As such, a new approach to structure the data that did not need "Country Name" was tried, which consisted in a wide format where each column corresponded to t – x, with t being the year to predict and x the number of years to consider before t, which would function as training for the model, e.g., for x = 3, three columns exist (t – 1, t – 2, t – 3). For 2017, X_train included all sets of three years that did not contain nulls, while X_test only considered 2014, 2015 and 2016. y_train consisted of the fourth year (t) of each set of three years. From this data structure, it was possible to obtain RFr and XGBr models that predicted accurately and according to the year. Table 8 presents the obtained truth-predicted values and best model's RVE considering t – 3. The predictions for 2018 were obtained using the predictions from 2017 for its X_test data.

Table 8 – Truth-predict values for the default best model of each indicator using t – 3 data structure.

Country	Country Population				Life Expectancy				Fertility			
	2017		2018		2017		2018		2017		2018	
	Predicted	Truth	Predicted	Truth	Predicted	Truth	Predicted	Truth	Predicted	Truth	Predicted	Truth
United States	327319054	325122128	331941134	326838199	79.02	78.54	79.30	78.64	1.78	1.77	1.77	1.73
Uruguay	3458333	3422200	3481581	3427042	77.72	77.63	78.00	77.61	1.99	1.80	1.98	1.66
Central African Republic	4645564	4996741	4702741	5094780	52.87	53.72	53.47	54.37	4.81	6.05	4.75	6.04
Seychelles	96061	95843	96710	96762	74.57	74.30	74.80	72.84	2.66	2.41	2.71	2.41
Senegal	15802058	15157793	16192996	15574909	67.47	67.75	67.95	68.10	4.72	4.62	4.67	4.58
Saudi Arabia	32908580	34193122	33576027	35018133	74.81	77.16	74.96	77.21	2.49	2.58	2.42	2.55
India	1340383120	1354195680	1362859880	1369003306	68.77	70.47	68.99	70.71	2.29	2.20	2.27	2.18
Lesotho	2234620	2170617	2267443	2198017	54.69	53.06	55.28	53.73	3.04	3.19	2.99	3.14
Luxembourg	595464	596336	608786	607950	82.56	82.10	82.80	82.30	1.47	1.39	1.48	1.38
Kenya	49530549	48948137	50472198	49953304	67.27	62.48	67.53	62.68	3.80	3.64	3.71	3.58
Best model	Default RFr RVE: 0.999		Default RFr RVE: 0.999		Default XGBr RVE: 0.912		Default XGBr RVE: 0.901		Default XGBr RVE: 0.964		Default XGBr RVE: 0.959	

Due to time and space limitations, further evaluation of the capability of this approach was not achieved. Testing the variation of x in t – x, usage of delta, and even introduction of extra variables would be desirable. The code for this approach will be sent in a separate file.