Aprendizagem Automática 2022/2023

Course Assignment

Group number: 36

Student names and ids:

Luís Pires 54471,

Inês Martins 54462,

Miłosz Włoch 59437

How many hours each student contributed to the project:

Luís Pires: 8 hours

Inês Martins: 8 hours

Miłosz Włoch: 8 hours

—---

Dataset description:

This dataset contains the data used for the 2006 TED Talk "The Best Stats You-ve ever seen" by Hans Rosling.
It contains 3 datasets related to country population, fertility rate and life expectancy up to 2016.

Aims of the assignment:

- Make a prediction models for predicting these 3 variables

- Models should be **TESTED** for **10 countries (selected randomly)**

- Make predictions for 2017
- OPTIONAL: Make predictions for 2018
- Compare the above results with actual historical data from the World Bank Site (e.g. for life expectancy)

Feature selection:

Before we could start to clean the data, we had to do the feature selection. Each one of 3 datasets had similar structure. Additionally, it was not difficult to select proper attributes. In the end, in every dataset we left everything except “Country Code”, “Indicator Name” and “Indicator Code”. These ones would only be noise for our regression models, that is why we got rid of them. “Country Name” lets us know which record regards which country. Columns marked as the following years starting from 1960 and ending in 2016 are our training data for our models.

Data preprocessing:

Datasets had some missing values. That is why before we could train any of the models to be able to predict country population, fertility rate or life expectancy we had to handle missing values. As for replacing null values with some numerical ones, we made our minds to choose Simple Imputer.

Data validation:

All the models have been validated based on the .csv files that come from the World Bank. For every task we selected 10 random countries. To avoid selecting countries which have missing data, we deleted rows in the validation csv with missing data and verified if the generated countries exist in both training and validation datasets.

Model results:

- **A summary of the performed work**

  We decided to choose one type of Machine Learning model to do regression for one dataset. Because we had 3 datasets, we applied 3 different models, one for each dataset. This could be a problem because we are comparing the results of models with different datasets, but since the data behaves in a similar way, we didn't find it to be a problem.

  The three models chosen were linear regression, SVR and polynomial regression.

  To choose the best model for SVR we did a grid search over a few of their hyperparameters to find the best ones. In this case, the best parameter occurred to be C=100000 and gamma=0.01 for the default kernel (rbf). For the polynomial regression we found the best parameter to be degree= 3. The linear regression did not have parameters.

  For every country we presented some basic statistics such as: RVE, RMSE, Maximum Error and Mean Absolute Error to show how the models perform.

  To perform prediction for any of the models, it was necessary to take advantage of Machine Learning algorithms that allow us to extrapolate, not only interpolate values as we were expected to estimate values that were below the scope, i.e. values for 2017 and 2018.

  A table regarding country population using linear regression and relevant statistics.

Country	RVE	RMSE	ME	MAE	Truth / Predictions (2017)	Truth / Predictions (2018)
IDA & IBRD total	0.9987	47028050	49323734	46969064	6343292514/ 6298678119	6420905201 / 6371581467
IDA only	0.9727	80606388	85420395	80453836	1060401803 / 984914526	1084433733 / 999013338

Ethiopia	0.9475	14178096	14864677	14160604	106399926 / 92943395	109224410 / 94359733
Poland	0.7699	2518643	2593455	2593455	37974750 / 40416367	37974750 / 10568205
Bangladesh	0.9933	7050274	7290760	7046025	159685421 / 166486711	161376713 / 168667473
Cambodia	0.9271	1027034	1055287	1026634	16009413 / 15011431	16249795 / 15194508
South Sudan	0.8935	683371	721695	682231	10975924 / 10189076	10975924 / 10333160
Equatorial Guinea	0.8261	319798	335137	319411	1262008 / 958322	1308966 / 973829
Haiti	0.9921	239145	244868	239075	10982367 / 10878315	11123183 / 10878315

A table regarding fertility rate using Support Vector Regressor (C=100000, gamma=0.01) and relevant statistics.

Country	RVE	RMSE	ME	MAE	Truth / Predictions (2017)	Truth / Predictions (2018)
IDA only	0.9925245403591457	0.25033911296718914	0.29968416565145395	0.2440874326051592	3.96861812356812 / 4.157108823126984	3.91585153154177 /4.215535697193224
Puerto Rico	0.9879975360141445	0.3806402778220639	0.4304816600574368	0.37684186851169765	1.101 /1.4242020769659585	1.035 /1.4654816600574367
Honduras	0.997779967519219	0.2177017304900494	0.27716209921331325	0.20560577951359793	2.496 /2.6300494598138826	2.46 /2.737162099213313
Burundi	0.9768215589902687	0.3627726686694508	0.4167655037281408	0.35797764914447505	5.502 /5.801189794560809	5.41 /5.826765503728141
Iraq	0.9754591088058537	0.8250163049636068	0.8947882413779209	0.8217795537667625	3.762 /4.510770866155604	3.672 /4.566788241377921
Finland	0.9293375975008906	0.30772019040238985	0.34728832578393476	0.30476838828220343	1.49 /1.752248450780472	1.41 /1.7572883257839347
Romania	0.9138485275515666	0.05138885603780254	0.06239034787460174	0.049830312398244536	1.71 /1.7723903478746017	1.76 /1.7972702769218873
Bahrain	0.9975825407049537	0.28013129998910163	0.3384159462645302	0.2721906631595832	2.01 /2.215965380054636	1.987 /2.3254159462645303
Lao PDR	0.9964272326168312	0.21001386820141837	0.2627762961183677	0.20059845775390595	2.709 /2.8474206193894442	2.667 /2.9297762961183675
Middle East & North Africa	0.9975993145594753	0.17063060082257664	0.2034743716811267	0.16659816013003437	1.87517978097801 /2.004901729556952	1.87531275614857 /2.0787871278296968

A table regarding life expectancy using polynomial regression and relevant statistics.

Country	RVE	RMSE	ME	MAE	Truth /	Truth / Predictions

					Predictions (2017)	(2018)
India	0.9993978783941702	0.25476009215331563	0.32438666882552525	0.24058130893087792	69.165 / 69.322	69.416 / 69.740
Heavily indebted poor countries (HIPC)	0.9932189459597622	1.6796305711735826	1.959755899424664	1.6510102999071385	62.748671772017 / 64.09093647240661	63.1349911561301 / 65.09474705555476
Burundi	0.9786789553027837	2.36490201878733	2.36490201878733	2.3609311639936656	60.898 / 58.400081255007535	61.247 / 59.02305641700514
Iceland	0.9758192593865594	0.3623193780102142	0.3846931358028485	0.3615815041805419	82.6609756097561 / 83.04566874555894	82.8609756097561 / 83.19944548231433
Fiji	0.9798648531261887	3.212519175747963	3.2494457215219796	3.2123044659060014	67.252 /70.42716321029002	67.341 /70.59044572152197
Russian Federation	0.7039125173173002	0.5815809334890581	0.799772466451131	0.4958507661305447	72.4514634146342 /72.64339248044416	72.6621951219512 /73.46196758840233
Bahrain	0.9983284228275133	0.875287731498314	0.9847547688409719	0.8673824747419019	77.032 /77.78201018064283	77.163 /78.14775476884097
Lebanon	0.9855059289905224	2.3643330383854972	2.552163377738907	2.35619783148222	78.833 /80.99323228522553	78.875 /81.4271633777389
Seychelles	0.962876302963659	0.7372531935918525	0.877908495499284	0.7201842321276217	74.3 /73.42209150450071	72.8414634146341 /73.40392338339007
Europe & Central Asia (IDA & IBRD countries)	0.9613481556547468	0.685538637422562	0.8774601612809647	0.6448793664524572	73.9036024121306 /74.31590098375455	74.0456465503097 /74.92310671159066

- ## Discussion and conclusion

In this Course Project we can see the different behaviors from the different regression models. For the linear model the RVE changes considerably depending on the country. This happens due to the shape of the data being linear or not. Because of this we can conclude that the linear regression model may not be the best choice for this type of problem.
For the SVR model we noticed a curious behavior: all predictions were higher than the true values. This model is more robust than the linear model which can be seen from the similar RVE values.
For the polynomial regression, we can see better results than linear regression, but with specific examples it falls short of SVR. This happens because the degree chosen is 3, which gives the model more flexibility than the linear regression but not as much as the SVR.