PPOC PROJECT REPORT

Understanding Various Development Indicators

By: Nikunj Kumar Jain (21129015)

Nitish Shankar(180491)

Kartikey(200493)

1. Introduction

Since the income inequality is an indicator of the economy of a nation and that has, as a consequence, a significant social impact as well, it becomes a matter of importance that this subject be properly investigated with mathematical rigour.

The Gini coefficient, is a measure of statistical dispersion intended to

represent
the income inequality or the wealth inequality within a nation or a social group.

In this project we tried to develop and test 2 regression models – Multiple Linear Regression and kNN based Regression for this index on a dataset comprising of the data from a set of OECD countries from 2010 to 2018.

2. Problem Definition and Algorithm

2.1 Task Definition

Figure out some of the major factors affecting the indicator, and train 2 machine learning models to estimate the indicator value ie the Gini Coefficient.

2.2 Algorithm Definition

Algorithms used are:-
>> Multiple Linear Regression (Parametric Approach)
>> kNN based Regression (Non-Parametric Approach)

3. Experimental Evaluation

3.1 Methodology

1)After reviewing reports by World bank UNDP and various economic policy institutions, we hypothesised that Income Inequality depends largely on 3 category of variables – Macroeconomic, Political Economy and Demographic.
2)We selected GDP growth rate, Inflation Rate and Unemployment rate as the Macroeconomic indicators; Corruption Perceptions Index as a

variable that affects the political economy and Secondary School Enrollment, Labor Force participation rate and population growth rate as the Demographic variables.

Each of these are standard variables who exact definition can is up for reference on the world bank website.

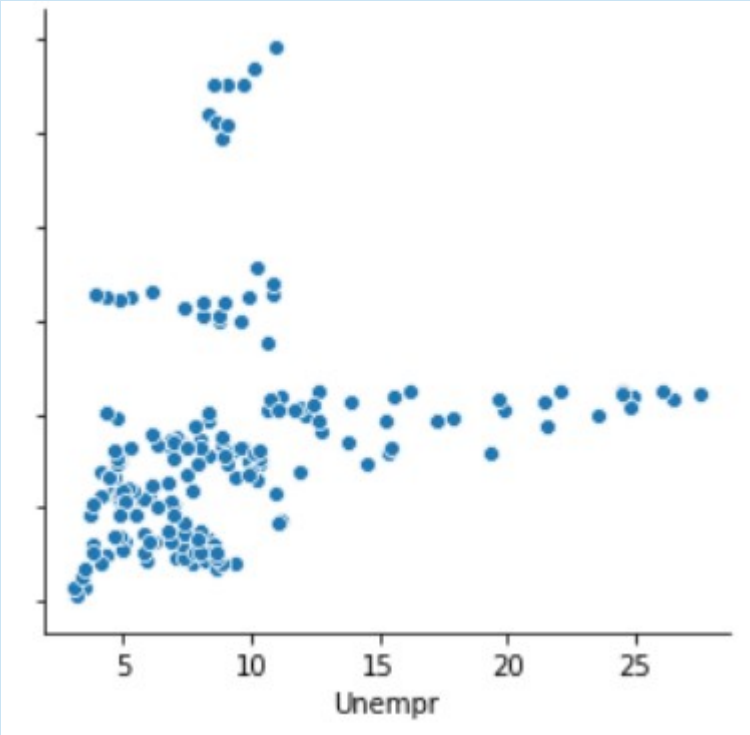
3)The data for all these independent variables and the dependent variable (Gini Coefficient) is compiled for a group of OECD countries - Austria, Belgium, Canada, Colombia, Denmark, Finland, France, Greece, Hungary Ireland Italy Netherlands Norway Poland, Portugal Spain, Sweden, Switzerland, Turkey, United Kingdom and United States for the years 2010-2018. (as much that was available)

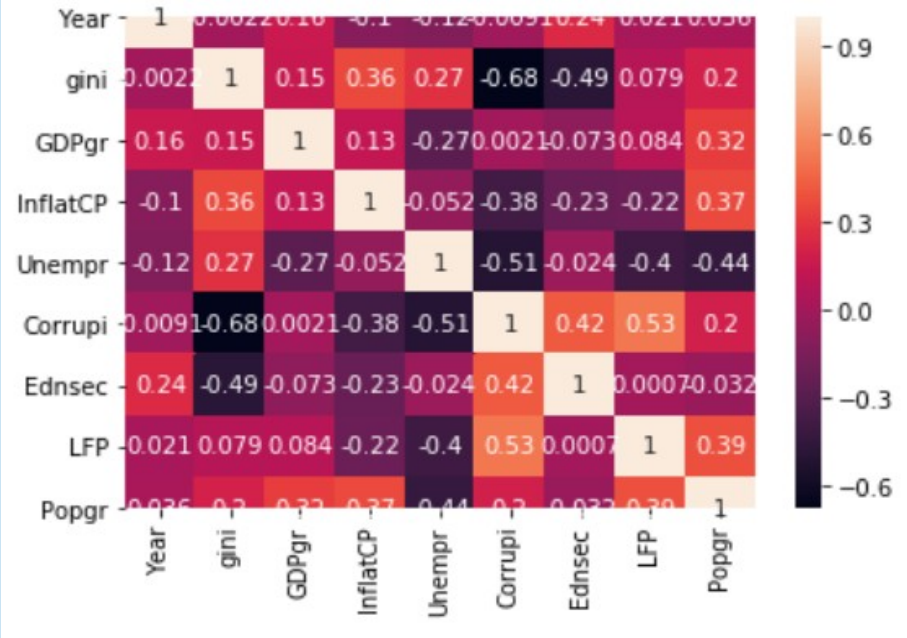
4)Data is then cleaned with missing values removed and checked for the duplicate values.

	Country	Year	gini	GDPgr	InflatCP	Unempr	Corrupi	Ednsec	LFP	Popgr
0	Austria	2010	30.3	1.837094	1.813534	4.82	79.00000	98.807213	60.400002	0.240394
1	Austria	2011	30.8	2.922797	3.286579	4.56	77.86903	98.061470	60.490002	0.337081
2	Austria	2012	30.5	0.680446	2.485676	4.87	69.00000	97.843323	60.759998	0.455937
3	Austria	2013	30.8	0.025505	2.000156	5.33	69.00000	99.197380	60.919998	0.589387
4	Austria	2014	30.5	0.661273	1.605812	5.62	72.00000	99.663078	60.689999	0.781542
5	Austria	2015	30.5	1.014502	0.896563	5.72	76.00000	100.320999	60.720001	1.120993
6	Austria	2016	30.8	1.989437	0.891592	6.01	75.00000	100.963020	61.209999	1.081396
7	Austria	2017	29.7	2.258572	2.081269	5.50	75.00000	100.455208	61.220001	0.694621
8	Austria	2018	30.8	2.501595	1.998380	4.85	76.00000	99.957840	61.369999	0.487072
9	Belgium	2010	28.4	2.864293	2.189299	8.29	71.00000	156.861526	54.070000	0.913639

5)Now exploratory data analysis is done to summarize the resulting data frame and see the correlation between the variables.

	Year	gini	GDPgr	InflatCP	Unempr	Corrupi	Ednsec	LFP	Popgr
count	191.000000	191.000000	191.000000	191.000000	191.000000	191.000000	191.000000	191.000000	191.000000
mean	2013.895288	33.194241	2.052353	1.865153	9.037958	69.269100	113.052099	59.733173	0.523811
std	2.533882	5.778483	2.924694	2.213068	4.979078	17.307807	17.651489	5.461378	0.576021
min	2010.000000	25.300000	-10.149315	-1.735888	3.120000	33.889670	84.283028	48.130001	-1.853715
25%	2012.000000	28.650000	1.070043	0.547608	5.770000	58.000000	100.469463	55.904999	0.179853
50%	2014.000000	32.500000	1.920446	1.515678	7.950000	74.000000	106.424942	60.509998	0.512923
75%	2016.000000	35.200000	2.913904	2.450838	10.210000	84.000000	123.012417	63.809999	0.925330
max	2018.000000	54.600000	25.176245	16.332464	27.469999	94.039270	163.934723	69.480003	1.702644





6)Now first, the multiple linear regression model is run by splitting the data 80-20 into training and testing set respectively. Dummy variable were introduced after one-hot coding the different countries.

7)Likewise, after that kNN model is run on NORMALIZED DATA by using Min-Max Scaler. The optimal k-Value is determined by looking at the RMSE data. The value for which it comes out to be minimum is selected, which was 2 in our case.

3.2 Results

The obtained R squared values are 98.51% and 99.06% respectively for the Multiple Linear Regression and the kNN based Regression Models, which implies that both are reasonably accurate in predicting the Gini Coefficient values.

Present the quantitative results of your experiments. Graphical data presentation such as graphs and histograms are frequently better than tables. What are the basic differences revealed in the data. Are they statistically significant?

3.3 Discussion

Given that the obtained R squared values for both the models come very close to 100%, it can be concluded that our hypothesis vis-à-vis the raw data, was well supported with the models.

Is your hypothesis supported? What conclusions do the results support about the strengths and weaknesses of your method compared to other methods? How can the results be explained in terms of the underlying properties of the algorithm and/or the data.

4. Future Work

Model shall be run on a bigger and much more exhaustive dataset and over a longer span of time to check for its correctness.

5.Bibliography:

>https://www.un.org/en/un75/inequality-bridging-divide

>https://www.transparency.org/en/cpi/2021?gclid=Cj0KCQjwzqSWBhDPARIsAK38LY87cCROWezK5bqcaBhSAy6DUYC6pbeIevFYUX8uPmpV-7zhB3ikumYaAsALEALw_wcB

>https://www.imf.org/external/pubs/ft/sdn/2015/sdn1513.pdf

>https://www.adb.org/sites/default/files/publication/234271/adbi-wp696.pdf