Certificación Internacional Data Analytics y Big Data

Prof. Gustavo A Rojas

por: Federico Ruilova A

San José, Costa Rica 31 de Marzo , 2020

Task 4 : Complete a Data Science Capstone Project

-

SUMMARY OF DATA SCIENCE PROJECT PROPOSAL

PROBLEM

The world is currently facing a pandemic that affects humans health. The virus know as covi-19 is disrupting the traditional economic standards and models . All markets are being affected, even automated prediction systems failed to predict a recovery since we are living through the results of very complex problem.

This capstone project attempted to answer to an hypotheses based on 2 questions :

1. ¿Can we predict Fatalities caused by this pandemic based on ; a) world economic index and, b) world happiness index?

2. Is there any correlation between this 2 data sets and the summary of reported cases on covid19 and the fatalities caused?

Methodology

1. Problem Framing (based on available data sources)

2.Collect the raw data from different sources (3 data sets were used)

3.Prepared the data, preparing data and merging.

4.Exploratory Data Analysis and continuos fixing

5. Performed in-depth analysis and Model testing

6.Results and Lessons Learned

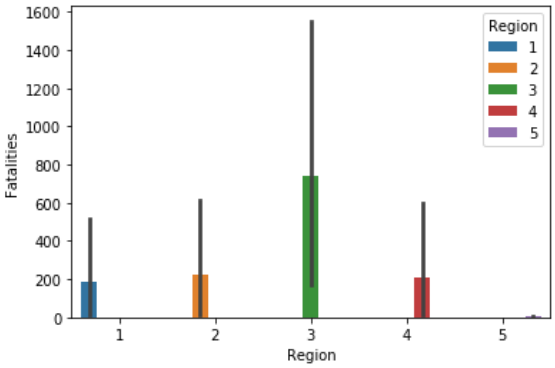
DATA SOURCES

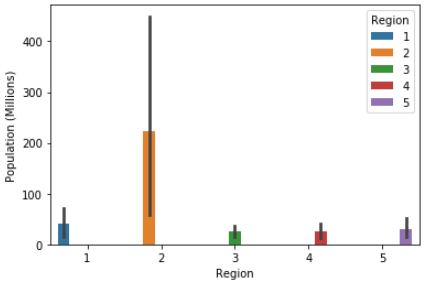
Data Sets	Source	Content	Reasons
The Economic Freedom Index 2020	https://www.kaggle.com/lwhelan/code-to-create-an-enriched-dataset/data	This file contains the Happiness Score for 153 countries along with the factors used to explain the score.The Happiness Score is a national average of the responses to the main life evaluation question asked in the Gallup World Poll (GWP), which uses the Cantril Ladder.The Happiness Score is explained by the following factors: - GDP per capita - Healthy Life Expectancy - Social support - Freedom to make life choices - Generosity - Corruption Perception - Residual error	Updated global , with social aspects and not just economic. Data Set Appendix https://happiness-report.s3.amazonaws.com/2020/WHR20_Ch2_Statistical_Appendix.pdf
The Economic Freedom Index	https://www.heritage.org/index	The data (last updated 26/02/2019) is presented in CSV format as follows: CountryID, Country Name, WEBNAME, Region, World Rank, Region Rank, 2019 Score, Property Rights, Judical Effectiveness, Government Integrity, Tax Burden, Gov't Spending, Fiscal Health, Business Freedom, Labor Freedom, Monetary Freedom, Trade Freedom, Investment Freedom, Financial Freedom, Tariff Rate (%), Income Tax Rate (%), Corporate Tax Rate (%), Tax Burden % of GDP, Gov't Expenditure % of GDP , Country, Population (Millions), "GDP (Billions, PPP)", GDP Growth Rate (%), 5 Year GDP Growth Rate (%), GDP per Capita (PPP), Unemployment (%), Inflation (%), FDI Inflow (Millions), Public Debt (% of GDP)	Picked to look for correlation in economic systems, and summarized complex systems
COVID-19 WORLD DATA SET	https://www.kaggle.com/c/covid19-global-forecasting-week-2/data	Cumulative number of confirmed COVID19 cases in various locations across the world, as well as the number of resulting fatalities, by date and by country .	Real world data collected until 31st of March 2020.

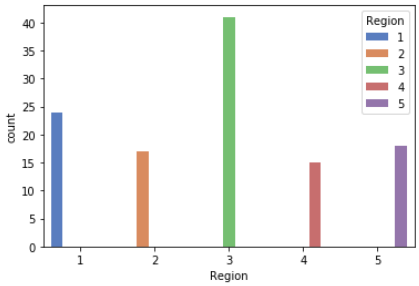
SUMMARY OF EXPLORATORY DATA ANALISYS

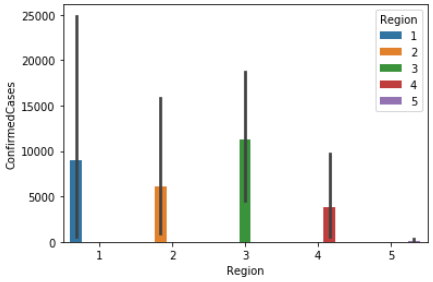
Regions:

1.Americas
2.Asia-Pacific
3.Europe
4.Middle East and North Africa
5.Sub-Saharan Africa

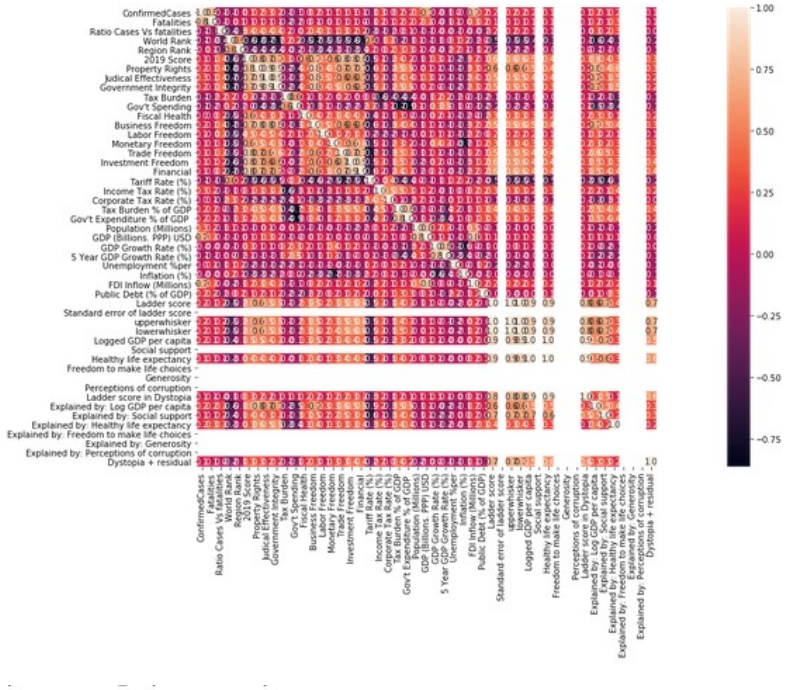


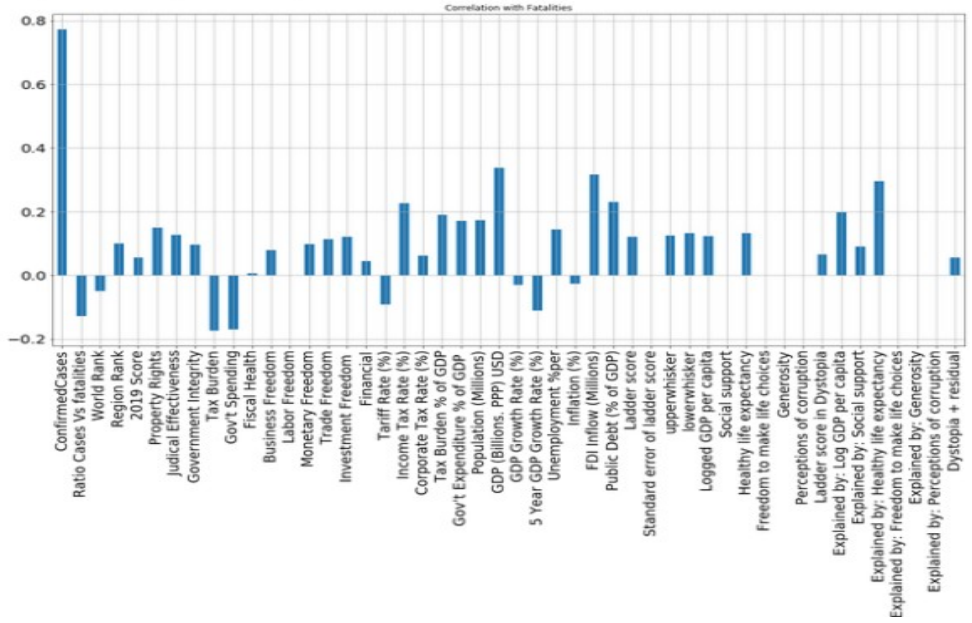


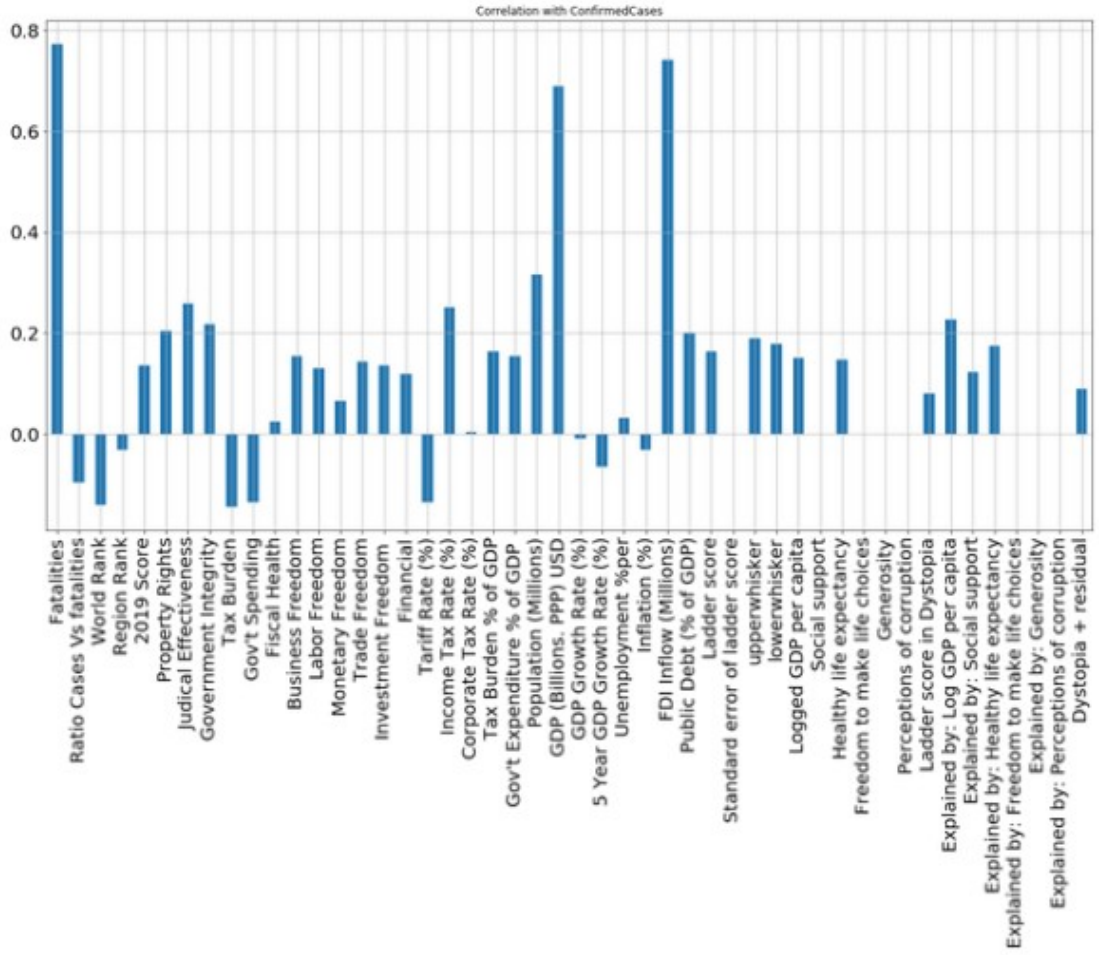




Efforts made to correlate Data

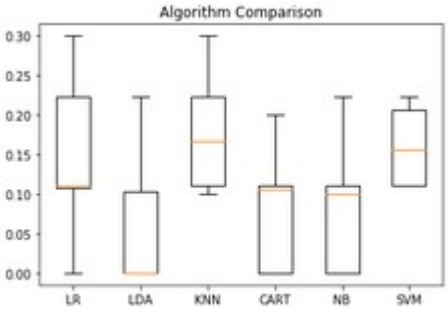


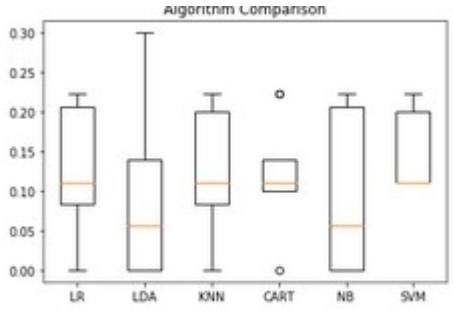


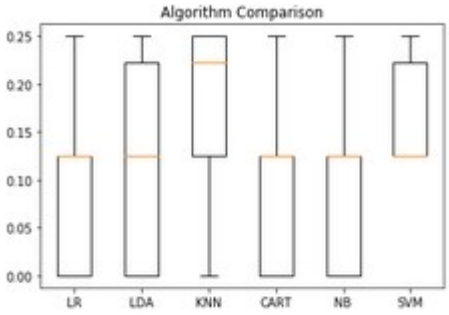

Correlation with ConfirmedCases
0.8
0.6
0.4
0.2
0.0
Fatalities
Ratio Cases Vs fatalities
World Rank
Region Rank
2019 Score
Property Rights
Judical Effectiveness
Government Integrity
Tax Burden
Gov't Spending
Fiscal Health
Business Freedom
Labor Freedom
Monetary Freedom
Trade Freedom
Investment Freedom
Financial
Tariff Rate (%)
Income Tax Rate (%)
Corporate Tax Rate (%)
Tax Burden % of GDP
Gov't Expenditure % of GDP
Population (Millions)
GDP (Billions, PPP) USD
GDP Growth Rate (%)
5 Year GDP Growth Rate (%)
Unemployment %per
Inflation (%)
FDI Inflow (Millions)
Public Debt (% of GDP)
Ladder score
Standard error of ladder score
upperwhisker
lowerwhisker
Logged GDP per capita
Social support
Healthy life expectancy
Freedom to make life choices
Generosity
Perceptions of corruption
Ladder score in Dystopia
Explained by: Log GDP per capita
Explained by: Social support
Explained by: Healthy life expectancy
Explained by: Freedom to make life choices
Explained by: Generosity
Explained by: Perceptions of corruption
Dystopia + residual

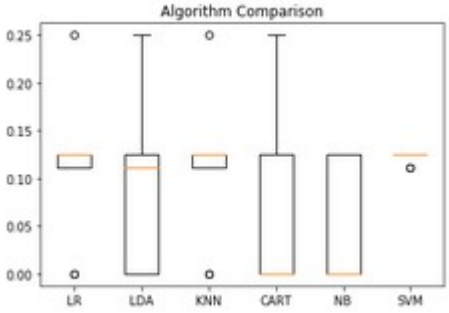
Efforts made to correlate Data

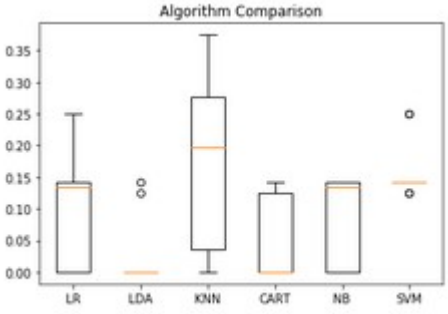
6 based models were used for the investigation, with more than 5 variations in the features selected, and the tunning ,a minimum of 30 different variations were tested ._











LESSONS LEARNED & REPORT

SUMMARY OF FINDINGS AND POTENTIAL OF THE PROPOSED HYPOTHESIS

1. There is correlation between the 2 indexes and the accumulated behavior of the pandemic until march 31st 2020.

2. A higher correlation was found between economical benchmarks and indexes than those found in the happines index, where economic factors take place along with perception, longevity and other features, this does not mean that there is no correlation at all between the happiness index and and the pandemic.

3. The matrix analyzed was too small to achieve high performance and accuracy on the tested models, but this opens a door to continue further research and hypothesis, since there is correlation and the date or data series were ignored and instead a summary was used, the correlated featured on this experiment could be use for continuous forecasting, prevention or even the creation of recovery models .

4. Higher GDP, means in most of the cases higher populations and bigger metropolis with a vast amount of factors and data behid that its worth the shot to explore in order to propose recovery models for this highly correlated economies and the fatalities that they suffered.

LESSONS LEARNED

1. Working with real world data requires and set of skills to zoom in and zoom out in different parts of the process, it was very a very enriching experience to attempt the experiment , analysis and hypothesis testing .

2. Data is everything , looking for data sources for real data was a challenge and opened a whole new set of skills to understand the importance of having robust and solid infrastructure with automated systems in order to create a efficient data acquisition process .

3. Correlation does not mean that the problem is answered but instead an invitation to iterate the process of exploration and clustering and take further .

4. Assumptions are fine, but theories must be proven with explratory data analysis , and questions must be answer with data .

RECOMENDATIONS

1. To have a more accurate models there are several experiments that could be conducted, e.g this experiment could be conducted on monthly, daily or weekly basis with real data coming in, at least from the pandemic, then I could help us understand how to react to recovery if its considered to be conducted several times and compared with new editions of the other 2 indexes.

2. In order to have more observations for fatality predictions, the analysis could be taken with clustering, using time as variable .