HEALTHCARE PREDICTION CANCER RELATED USING DATA SCIENCE TECHNIQUES

Non scientific project within the context of IBM data science certification

Jean-Francois Moy

March 2020

1. Introduction

1.1 Background

Nowadays, the number of people knowing someone who have or have had a cancer has increased. The share of population having cancer in developed country has reached an historical high when this rate remains quite low in developing countries. Therefore, we can expect an increase of this disease in developing countries when these ones will know a development.

Even if miracle solution still doesn't exist, we have a quite good knowledge of what it takes to take care of patients. In order to being able to diagnose and heal people correctly, each country must have the necessary resources.

But building a hospital takes time. Training doctors and nurses takes times as well. Increasing the resources is not possible overnight. This should be considered on a long-term strategy.

The goal of this project is to find a model able to predict the share of a population having cancer. This will allow the concerned authorities to forecast their needs.

1.2 Problem

First of all, it has to be mentioned that this article is not a scientific publication. It is not claiming to explain scientifically some root cause of cancer. The idea is to draw a model that seem reasonable to help countries structuring their healthcare system.

It is generally admitted that cancer is due to three main factors: genetics, randomness and environment. As you can imagine, it would be difficult to get some data about genetic material for ethic reasons, and way more difficult about randomness. Thus, we will mainly consider data about people's environment. Once again, this is not a scientific publication, and the goal is only to produce a support to help structuring health care systems.

Given the nature of the problem, an issue that we could face would be to find a model composed by features that are difficult to forecast and this unusable.

1.3 Interest

The interest of this project is idealist but laudable. This is about trying to forecast healthcare needs to try to save some lives or at least, extend life expectancy. It could interest developing countries who need a support to structure the future of their healthcare system.

2. Data acquisition, cleaning and use

As mentioned in the presentation, we will focus on environmental indicators. A list of features has been created, including some suspected factors causing cancer. This list is also including indicators more related to the country people live in and that doesn't seem to be directly correlated to the disease.

Feature Name	Description
over65_percent	Share of the population over 65 years-old
HDI	Human Development Indicator
life_expectancy	Life expectancy of an individual
urban_percentage	Share of the population living in cities
mean_school	Mean number of years of schooling
GDP_per_capita	Gross Domestic Product per person
pesticide	Quantity of pesticides used on the country per ha
cig_prevalence	Share of the population smoking
equator_dist_abs	Absolute distance of the country from the equator
fat	Quantity of fat eaten per person per year
meat	Quantity of meat eaten per person per year
protein	Quantity of protein eaten per person per year
milk	Quantity of milk eaten per person per year
calories	Quantity of calories eaten per person per year

2.1 Data selection/acquisition and cleaning

Most of the data that have been used in this project are coming from 'Our world in Data' website (https://ourworldindata.org) . As mentioned in its 'About' page, this website gives free access to data provided by 'researchers at the University of Oxford, who are the scientific editors of the website content; and the non-profit organization Global Change Data Lab, who publishes and maintains the website and the data tools that make our work possible'. Some other data are coming from the United Nation website (https://www.un.org).

The first step has been to determine the kind of data that could be relevant. After reading different articl es about cancer, the decision has been made to use the features presented in the table above. Also, it's b een decided to gather as much features as possible to feed to model, assuming that the algorithms are a ble to select, or give a consistent weight to each feature. 21 data sets have been gathered, each represe nting a single feature.

The second step has been to build only one data frame from the different data sets, each feature coming from a distinct file. The material was quite clean and uniform, but values were missing for some dates or some countries.

These features have been collected for 117 countries and from 1990 to 2012. This period has been chosen because it allows us to have a continuous dataset for a large number of countries. It has been difficult to collect data before 1990 as developing countries generally have few data before 1990. It has also been difficult to collect data after 2012 because some datasets are not available or free yet.

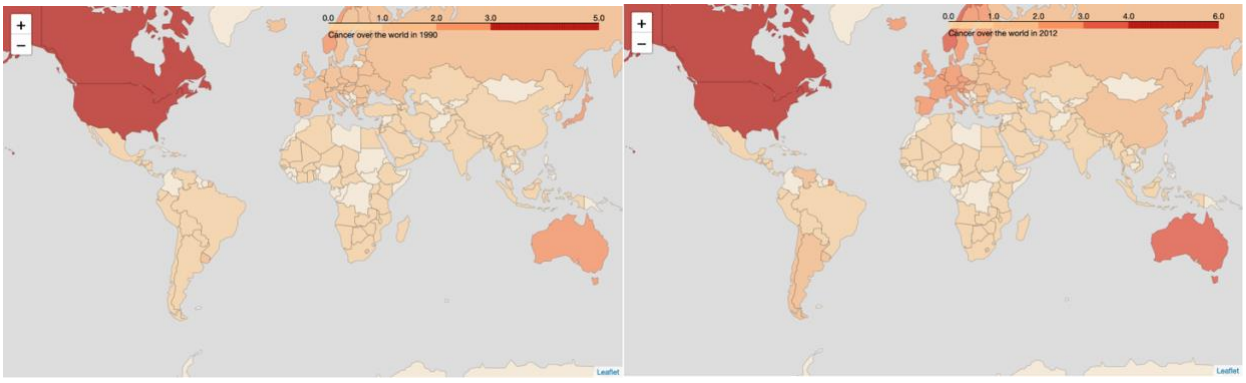
Finally, the dataset has been saved into one file to easily process it under Python.

	Code	Year	dized_neoplasms_percent	65andover_percent	gram_per_day_per_capita	HDI	Life_expectanc
9	AFG	1998	0.48850854	2.34152651	76.30136985	0.339	54.90
10	AFG	1999	0.494118479	2.319288177	70.16438355	0.343	55.37
11	AFG	2000	0.502297156	2.291944549	67.78082191	0.345	55.84
12	AFG	2001	0.508535544	2.271866861	76.41095889	0.347	56.30

3. Methodology

First of all, it's been necessary to plot the collected data to confirm the consistency between what one can read or hear, and what the data is showing.

3.1 Exploratory Data Analysis



These two maps are showing the evolution of the prevalence of cancer between 1990 and 2012. We can see that the share is more important in developed country and this trend is growing over the years. Also, we can see that the prevalence seems to be increasing in all countries.

This is what we were expecting, and it consolidated our hypothesis. Therefore, the result we could expect, is a model including features related to the standard of living, or the level of development of a country.

3.2 Methodology – Use of the data

Data has been used to feed algorithms to find a model able to forecast the share of population with cancer. As the impact of each feature is so far unknown, different algorithms have been selected to maximize the quality of the output model.

At the very beginning of the study, 4 algorithms have been considered to draw models:

- SVR: Support Vector Regression, SVR from sklearn python library
- Non-linear regression, Model from lmfit library
- OLS regression: Optimum Least Square, OLS from statsmodels library
- LASSO regression: least absolute shrinkage and selection operator, LassoCV from sklearn library

After a study of each algorithm features, it's been decided to work with the LASSO algorithm because it performs both variable regularization and selection and thus produces a model with fewer features.

4. Results

4.1.1 Solution to the problems

The output of the LASSO regression is a list of coefficients for the features identified as correlated with the percentage of people having cancer. Here are the coefficients found:

The model has selected 4 features and assigned a value to it.

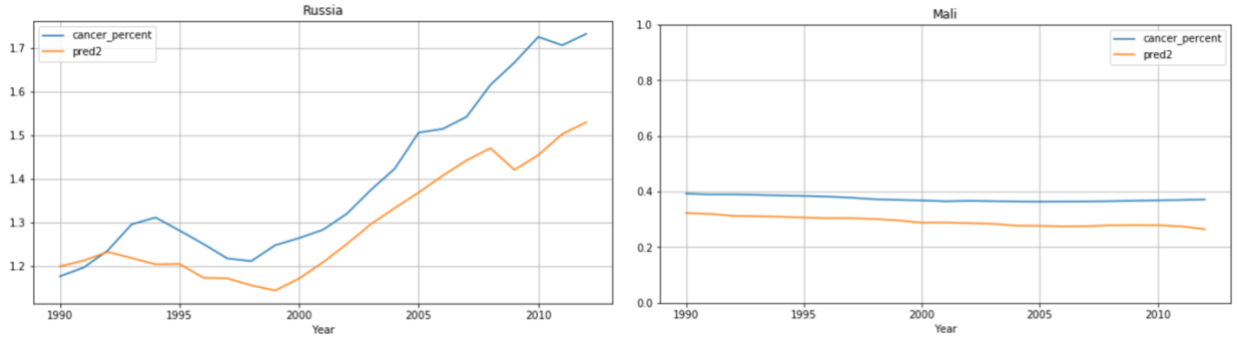
	Coefficients Metamodel
over65_percent	
mean_school	
GDP_per_capita	
meat	

You can contact me if you want the values found by the model.

To make a prediction, it's simply necessary to multiply the coefficient by the value of the feature. Then sum all the products.

For instance, if we apply the coefficient of the model to the projected values, the share of the population with cancer in France would be 2.74% in 2040, against 2.12%.

Here are two examples of the performance of the model:



We previously observed that the most developed countries were the ones with the higher level of development. Among the selected features, 3 out of 4 are HDI components, validating the assumption about the link between development and cancer share of the population.

Also, with a simple model, made of only 4 components, the R2 is 0.86447, which is satisfactory.

It's to be mentioned the US and Canada have been removed from the data base as they appear as outliers and are decreasing the quality of the model. The R2 of the model with these countries was 0.76232.

5. Conclusions

The LASSO methodology allowed us to create a model able to predict the share of a population having cancer by introducing only 4 features and a high confidence.

With simple calculations, it's possible for a country to determine the number of health care professionals they will need in the future.

6. Future directions

The problem we could face here is that the model has been built on past data, and is dealing with human body, culture and genetics. We could wonder if the trends we identified with this model will still be applicable in the future.