Cancer forecast through machine learning techniques

Non scientific project within the context of IBM data science certification

Jean-Francois Moy – April 2020

Introduction

1.1 Background

- The share of the population is growing in developed countries, and developing countries will probably follow.
- Countries need to be ready to face this increasing trend as it takes time to structure a health care system.
- Forecasting the share of population with cancer could help seizing the future needs.

1.2 Problem

- Causes of cancer are not 100% know
- Three main factors: genetics, randomness and environment.
- Only data about people's environment will be considered.

1.3 Interest

- Try to forecast healthcare needs to try to save some lives or at least, extend life expectancy.

Data Acquisition and use

- A list of 14 features has been created.
- Features have been selected regarding scientific publication
- Data sources = data sets from Our World in Data and UN websites
- Data has been cleaned to keep only complete sets.
- The result is a data set of 21 metrics for 117 countries, from 1990 to 2012.

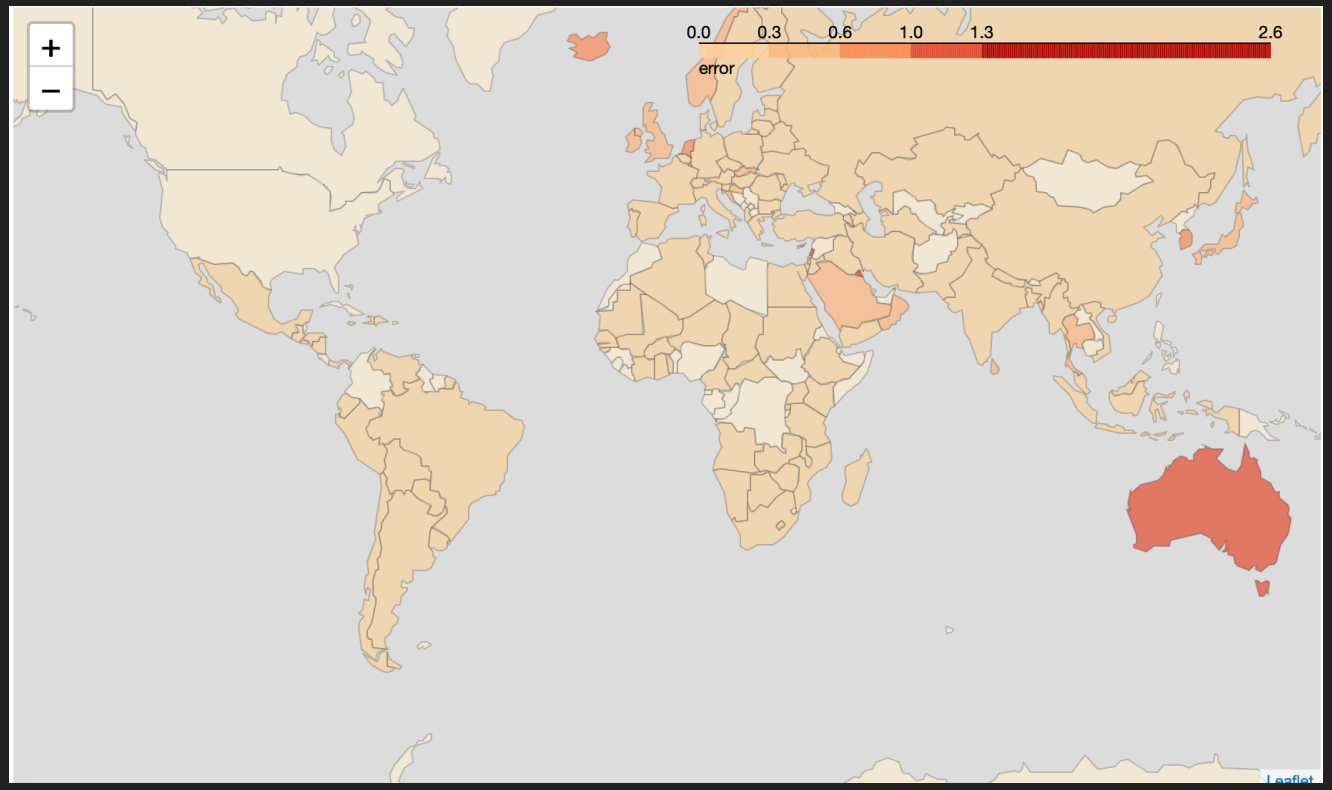
Methodology

- Plotted the data to have an overview
- Selected features have been used to feed algorithms.
- 4 algorithms have been considered to build a model.
- LASSO regression have been chosen: least absolute shrinkage and selection operator, LassoCV from sklearn library
- Chosen because it performs both variable regularization and selection and thus produces a model with fewer features.

Results

- The regression produced a 4 features model
- The performance of the model is given by a R2 of 0.86 which is satisfactory.
- Also, we can see on the graph on the right that the error between prediction and the actual values are generally acceptable

	Coefficients Metamodel
over65_percent	0.069711
mean_school	0.008057
GDP_per_capita	0.000016
meat	0.001868



Conclusion and future directions

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.

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.