China Contamination

Data Science: Capstone Proyect

Jorge Haces

16/6/2020

Contents

1 Overview **2**
1.1 Introduction 2
1.2 Project Description 2
1.3 DataSet 2

2 Methods and Analysis **3**
2.1 Data Stage 3

1 Overview

This is the final project required for **Data Science: Capstone** course offered by *edX HarvardX for Professional Certificate Program in Data Science*. The theme of this proyect is the Contamination, specifically in China and the aim is to predict the air quality in the fastest growing country nowadays.

1.1 Introduction

Contanimation is defined as the presence of materials in the air that cause serious harm or discomfort to people. The contamination has increased since the Industrial Revolution began, in the second half of the 18th century, with production processes in factories, the development of transportation and the use of fuels.

According to the World Health Organization (WHO), the state of the current atmosphere causes, by the simple act of breathing, the death of around seven million people a year (fine particle respiration).

The most common air pollutant gases are carbon monoxide, sulfur dioxide, chlorofluorocarbons, and nitrogen oxides. Photochemicals such as ozone and smog are increased in the air by nitrogen oxides and hydrocarbons reacting with sunlight.

Contaminants are classified into:

- Primaries are those that are emitted directly into the atmosphere such as sulfur dioxide, carbon monoxide
- Secondary are those that are formed by atmospheric chemical processes that act on primary contaminants such as sulfuric acid, which is formed by the oxidation of sulfur dioxide, nitrogen dioxide that is formed by oxidizing the primary pollutant nitric oxide and ozone that is formed from oxygen. [1]

1.2 Project Description

An analysis of the data will be carried out based on the following models: k Nearest Neighbors, Logistic Regression, Support Vector Machines (SVM), Random Forests and Neural Network to help us predict if pollution will grow even more (2.3% in 2018 almost at double compared to 2010). To answer the question, will china be able to comply with the Paris agreement signed in 2015? Whose goal is to reduce the global temperature to 2°C in 2050.

For this we will divide the data into two: training data and test data. Later, we will train the different models in the first set and then will be evaluated in the second set. Finally, we will use the **Root-Mean-Square-Error (RMSE)** and the **"overall accuracy"** to rate the performance of each model and thus identify the best for this project.

1.3 DataSet

The Dataset used in this project is *Beijing Multi-Site Air-Quality Data Data Set*, available at the UCI Machine Learning Repository [2].

This data set includes hourly air pollutants data from 12 nationally-controlled air-quality monitoring sites. The air-quality data are from the Beijing Municipal Environmental Monitoring Center. The meteorological data in each air-quality site are matched with the nearest weather station from the China Meteorological Administration. The time period is from March 1st, 2013 to February 28th, 2017. Missing data are denoted as NA.

The Attribute Information is the following:

- No: row number
- year: year of data in this row
- month: month of data in this row
- day: day of data in this row
- hour: hour of data in this row
- PM2.5: PM2.5 concentration (ug/m^3)

- PM10: PM10 concentration (ug/m^3)
- SO2: SO2 concentration (ug/m^3)
- NO2: NO2 concentration (ug/m^3)
- CO: CO concentration (ug/m^3)
- O3: O3 concentration (ug/m^3)
- TEMP: temperature (degree Celsius)
- PRES: pressure (hPa)
- DEWP: dew point temperature (degree Celsius)
- RAIN: precipitation (mm)
- wd: wind direction
- WSPM: wind speed (m/s)
- station: name of the air-quality monitoring site

The data is contained in a Zip file named *PRSA2017_Data_20130301-20170228.zip* containing 12 files (one for each municipality), as follow:

- PRSA_Data_Aotizhongxin_20130301-20170228.csv
- PRSA_Data_Changping_20130301-20170228.csv
- PRSA_Data_Dingling_20130301-20170228.csv
- PRSA_Data_Dongsi_20130301-20170228.csv
- PRSA_Data_Guanyuan_20130301-20170228.csv
- PRSA_Data_Gucheng_20130301-20170228.csv
- PRSA_Data_Huairou_20130301-20170228.csv
- PRSA_Data_Nongzhanguan_20130301-20170228.csv
- PRSA_Data_Shunyi_20130301-20170228.csv
- PRSA_Data_Tiantan_20130301-20170228.csv
- PRSA_Data_Wanliu_20130301-20170228.csv
- PRSA_Data_Wanshouxigong_20130301-20170228.csv

2 Methods and Analysis

2.1 Data Stage

```
## 'data.frame':    420768 obs. of  18 variables:
##  $ No     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ year   : int  2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...
##  $ month  : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ day    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ hour   : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ PM2.5  : num  9 11 8 8 8 10 8 8 3 3 ...
##  $ PM10   : num  9 11 8 8 8 10 8 8 6 6 ...
##  $ SO2    : num  6 7 NA 3 3 4 6 8 9 10 ...
##  $ NO2    : num  17 14 16 16 NA 8 13 20 23 18 ...
##  $ CO     : int  200 200 200 NA 300 200 300 300 300 300 ...
##  $ O3     : num  62 66 59 NA 36 64 61 54 50 56 ...
##  $ TEMP   : num  0.3 -0.1 -0.6 -0.7 -0.9 -1.6 -2.4 -0.8 0.4 1.5 ...
##  $ PRES   : num  1022 1022 1023 1024 1024 ...
##  $ DEWP   : num  -19 -19.3 -19.7 -20.9 -21.7 -21.1 -20.3 -19.9 -19.4 -19.7 ...
##  $ RAIN   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ wd     : Factor w/ 16 levels "E","ENE","ESE",..: 15 15 15 8 15 5 5 6 5 2 ...
##  $ WSPM   : num  2 4.4 4.7 2.6 2.5 2 2.3 2 2.7 2.9 ...
##  $ station: Factor w/ 12 levels "Aotizhongxin",..: 12 12 12 12 12 12 12 12 12 12 ...
```