Prediction of Air Pollution

Data Mining I (CC4018) - 19/20: Practical Assignment

Rita P. Ribeiro

November, 8th, 2019

Description

Just recently, New Delhi appeared on the news for a bad reason: its air pollution levels have soared to hazardous levels. Experts say that:

> *"(...) the deterioration of air quality this year is particularly concerning, (...) A combination of human and environmental factors, including agricultural crop burning to clear fields and fumes from passenger and freight vehicles combined to create a 'perfect storm of pollution'. (...) The weather — slowing winds and stagnant air — also allowed for a build-up in pollution"*
>
> *(In TIME Magazine, November 6th, 2019)*

Figure 1: Extract from a news article from TIME Magazine - November, 6th, 2019

This practical assignment aims to predict air pollution in Beijing, China, using the data set Beijing Multi-Site Air-Quality Data Data Set.

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.

A few years ago, China established the Air Quality Index (AQI) based on the level of five atmospheric pollutants, namely sulfur dioxide (SO2), nitrogen dioxide (NO2), suspended particulates (PM10), carbon monoxide (CO), and ozone (O3) measured at the monitoring stations throughout each city. An individual score is assigned to the level of each pollutant, and the final AQI is the highest of those five scores. The pollutants can be measured quite differently. SO2, NO2 and PM10 are measured as an average per day. CO and O3 are more harmful and are measured as an average per hour. The final AQI value is calculated per day and has the interpretation shown in Table 1.

Your task is to predict the AQI or Air Pollution Level for a given day. You can start to focus on the data set from one of the monitoring sites, and then, if you have the time, extend your study to the other monitoring sites.

AQI	Air Pollution Level	Health Implications
0 - 50	Excellent	No health implications
51 -100	Good	No health implications
101-150	Slightly Polluted	Slight irritations may occur, individuals with breathing or heart problems should reduce outdoor exercise.
151-200	Lightly Polluted	Slight irritations may occur, individuals with breathing or heart problems should reduce outdoor exercise.
201-250	Moderately Polluted	Healthy people will be noticeably affected. People with breathing or heart problems will experience reduced endurance in activities. These individuals and elders should remain indoors and restrict activities.
251-300	Heavily Polluted	Healthy people will be noticeably affected. People with breathing or heart problems will experience reduced endurance in activities. These individuals and elders should remain indoors and restrict activities.
300+	Severely Polluted	Healthy people will experience reduced endurance in activities. There may be strong irritations and symptoms and may trigger other illnesses. Elders and the sick should remain indoors and avoid exercise. Healthy individuals should avoid out door activities.

Table 1: China AQI levels and Health Implications (Daily Targets).

Tasks

Using the above data set, you have a set of main tasks to accomplish as described next. Still, you are free to include other tasks to increase the value of your assignment.

Task 1: Data importation, clean-up and pre-processing

In this part of your work you should focus on importing the provided data into an appropriate R format so that your posterior analysis is made simpler. You should also check if it is necessary to carry out any data clean-up and/or pre-processing steps.

Task 2: Data exploratory analysis

This part involves summarising and visualising the data in forms that you think are useful. Try to think about interesting questions that could be interesting to check with the available data, and provide answers either using textual summaries or data visualisation.

Task 3: Predictive modelling

You should define a predictive task that can help to predict the air pollution, through the AQI value or Air Pollution Level, given its feature values. After defining the task, you should use your available data to select and obtain a good model for this task. Justify your suggested model.

Tools

In your work, you should use R programming language. You can find material for dynamic reporting in R with markdown here.

Deliverables

The practical assignment is **mandatory** and should be performed by groups of, **preferably, three students**.

Until the next **November, 22nd, 2019** you should inform me of the group constitution: full names and student numbers.

Your assignment should be sent to me by email with the subject "[DMI] Group X", where X is the number that I will assign to your group, and including the following items in a compressed file:

- a final report [1] in PDF generated dynamically with the identification of group members, and with a structure similar to the following:
  - introduction;
  - problem definition;
  - data pre-processing;
  - exploratory data analysis;
  - predictive modelling: experimental setup and obtained results;
  - conclusions, shortcomings and future work;
  - appendix (optional).
- the source of a ready to execute dynamic report that produces your final report with all the code that is necessary to run to obtain the results you present;
- any complementary files needed to execute your report (e.g. data files, data objects).

Important Notes

- Any data pre-processing steps must be presented and justified.
- All the algorithms and parameters used should be indicated.
- Organization of discourse and presentation, clarity of language and ideas are rewarded.
- Long sequences of poorly formatted output dumps are penalized.
- The report should also refer to any source you used and make explicit which part of the work it influenced.
- It is important that your code does not rely on an absolute path, so that it can be run on any computer.
- The R code should not appear on the report, just the output of it.
- If your report contains code that takes too much time to run you should include it with the `eval=FALSE` code chunk option so that the code is not executed. If the result of your code is needed, you may run it locally on your computer, save it in a binary file and load it in your report.

Deadline

The deadline for submitting the practical assignment is **December, 23rd, 2019**.

[1] Please keep your report under 4-5 pages. You can freely attach appendices with additional information you consider relevant, but your work should be perfectly understood by the report alone.