assignment-03

October 6, 2020

1 EPA-1316 Introduction to *Urban* Data Science

1.1 Assignment 3: Prediction/Inference

TU Delft Q1 2020 Instructor: Trivik Verma **TAs:** Aarthi Meenakshi Sundaram, Jelle Egbers, Tess Kim, Lotte Lourens, Amir Ebrahimi Fard, Giulia Reggiani, Bramka Jafino, Talia Kaufmann
Computational Urban Science & Policy Lab

1.2 1. Introduction

Note: If you have not gone through **labs and homeworks 06-07**, kindly do so before starting this assignment, as those will help you with all the necessary knowledge for this assignment. This assignment will be useful for you when explaining your hypothesis with the use of evidence from exploratory data analysis.

1.1 Submission Please submit the results by Brightspace under **Assignment 03**, using a single file as example,

`firstname_secondname_thirdname_lastname_03.html`

If your file is not named in lowercase letters as mentioned above, your assignment will not be read by the script that works to compile 200 assignments and you will miss out on the grades. I don't want that, so be exceptionally careful that you name it properly. Don't worry if you spelled your name incorrectly. I want to avoid a situation where I have 200 assignments all called assignment_02.html

Please **do not** submit any data or files other than the `html file`.

Don't worry about your submission *rendering without the images* **after** you submitted the file on brightspace. That is a brigthspace related issue of viewing your own submission but when I download all assignments as a batch file, I get all your images and code as you intended to submit. So make sure that your html shows everything you want us to see **before you submit**.

1.2 How do you convert to HTML? There are 2 ways,

1. from a running notebook, you can convert it into html by clicking on the file tab on the main menu of Jupyter Lab

- File → Export Notebooks as… → Export Notebook to HTML

2. go to terminal or command line and type
   - `jupyter nbconvert --to html <notebook_name>.ipynb`

1.3 Learning Objectives This assignment is designed to support one learning objective. After completing this laboratory you will be able to:

- Explain the outcome of your regression model and relate them to your hypothesis.

1.4 Tasks This assignment requires you to go through the following tasks in inference or prediction.

0. Use Assignment 2 as a base for this assignment.
1. Build a regression or clustering model to explain your hypothesis (prediction or inference). Train and evaluate the model.
2. Use EDA, plot results, and make refinements to your models. (assignment 2 can be used here as a basis for your models - save time and remember the process is not linear!)
3. Perform error analyses, noting the best features, worst features, and any trends in misclassifications. Report your results and provide in-depth analysis of the performance using quick notes in markdown cells.
4. You should designate reasonable sub-sections of your data to serve as the training set, validation set, and test set. After training your model, provide to your model each observation from the testing set and try to accurately predict or classify the type of variable.
5. Use markdown cells to describe your explanations/choices/methods/analysis.
6. Discuss model weaknesses and future ideas for improvements.

A hint: you can use variations in the year of the data to use extra data in your observations. Does your model train on year 2016 and predict or classify correctly for 2017?

"Okay, now I know the tasks but what is the context?" Read on..

`Problem Statement`: While we all aspire to minimise crime, poverty, segregation or other inequitable and undesirable outcomes in society, it is necessary to first understand the severity of the issue and some of the most contributing factors. It may come as a surprise, but a lot of these outcomes are ubiquitous in many parts of The Hague. Pinpointing the exact causes of such observations is impossible, as these are highly nuanced and complex issues. However, factors such as gross income, economic disparity, liveability and government infrastructure/support are often strong indicators. For example, it is widely believed that there's an increase in crime in regions that have a large variance in citizens' income levels.

In The Hague, many neighbourhoods are quite poor, do not have access to public transport services or enough amenities and resources. A lot of these lacking infrastructure services result in complex socio-economic outcomes like gentrification, segregation, lack of jobs, excessive reliance on personal vehicle etc.

Consider your hypothesis from **assignment 02**. You are tasked with inspecting factors that may be correlated and/or contributing to a variable of your interest or the types thereof (e.g., crime, poverty, liveability, and their brackets or types). You should explore the same datasets from assignment 02 (they are mentioned again for clarity, below). Perform exploratory data analysis, iteratively

refine your questions and chosen datasets, and produce plots to understand and communicate your findings. You may have done most of the exploration in assignment 02, so choosing the same hypothesis will save you time. You are free to come up with a different hypothesis, but that would mean more time to go through data cleaning and EDA.

`Example` If I look at crime, I'd start by looking at surface-level factors that may be correlated to crime, the conditions of a neighbourhood: presence of shops, schools, amenities. Are certain crimes more likely to occur where there is a dearth of amenities? Is there any correlation at all with busy areas and the locations of crimes? Or with rich areas and crime? How should I measure the presence of amentities or the richness of a neighbourhood?

`Project Goal`: Improving socioeconomic issues is a complex, difficult, and slow process. You may not even have variables in your hypothesis that logically completely correlate with your variable of choice. But try! This is only practice. We want students to explore data and make statistical predictions in an area that is vastly affected by many confounding, real-world complex factors. The goal is for students to illustrate learned skills that align with our course goals. The primary goal of this project is to explain the outcomes of an event you observed and hypothesized in the cijfers data from The Hague. Any piece of information from Den Haag Cijfers or CBS can be used to perform this analysis, but the suggested feature set should include the variables on the demographic information of The Hague (see below for datasets).

`What we expect`: We expect you to think about your models in this assignment. What are the overall strengths and weaknesses of your models? Continue to improve your model to the best of your abilities. Lastly, comment on any additional ideas you have for a model that may be able to perform better, even if these ideas are not possible by any model that we have learned about in the course – for example, explain what type of information you wish your model could use, and how do you wish it could leverage that information. This doesn't have to be explained in mathematical terms or theory, but the goal is to think about current limitations of models and how you wish to improve upon them.

Remember to always document your code!

1.3 2. Download the Data

For this assignment I am providing you with the shapefiles of The Hague. Mikhail Sirenko has prepared these files with love and care so that you can connect it with either Den Haag Cijfers or CBS datasets without having to clean badly collected data.

Note: For data from CBS, data is only complete upto 2017. You will have to subset the data on municpality using the variable name `gm_naam = 's-Gravenhage` and then subset on neighbourhood resolution using variable name `recs = Buurt` to get the data that can match the shapefiles we have provided.

So after you unzip, we'll work with the file `neighborhoods.xxx`, which is in one of many geographic formats. Put the data in a convenient location on your computer or laptop, ideally in a folder called **data** which is next to this **jupyter notebook**. I recommend taking a look at the file with format `.json` in a text editor like *atom* for any system or notepad++ for windows. These will also make your life easy for everything else on your computer. Make sure you've set your working directory in the correct manner – okay?

It's a big file and it may take a while to load onto your laptop and into Python (running on the jupyter labs environment).

So, to summarise, you will use at least two datasets.

1. Download Shapefiles provided with the assignment
2. Get a second dataset of your choice from The Hague city region using the links above (curate them as you like)

More Data Sources You can find more data sources on Cities and Population, Climate indicators and Land-use in the following links in case you are attempting the **[Optional]** exercise.

- http://citypopulation.de/
- https://www.census.gov/programs-surveys/geography.html
- https://www.eea.europa.eu/data-and-maps
- http://download.geofabrik.de/

In case you get more data as shapefiles, and want to play with projections, a nice guide for it is here

1.4 3. Start your analysis

[1]:
```
# your code here
# use many cells if you like to structure your code well
```