KU LEUVEN

LEUVEN STATISTICS RESEARCH CENTRE

Modern Data Analytics

Exam Project GOZ39a

August

Summary

There is no exam for this course, instead a project has to be handed in. The project involves a data analytics assignment. The main difference with the June-presentation is the fact that this is an individual project.

Errors to Avoid

The main errors we have seen during the June-presentations were:

- Missing requirements file or incomplete requirements file
- Code that did not work at all
- Reference to local hard drives
- If then / For loops applied on DataFrames
- Weak report_write-up

During the presentation you will be asked questions on your code, on python and on machine learning concepts explained during the lectures. Make sure you understand (and are able to explain) every single line of code you write !

The teaching team would be sad, if the only thing you achieved in our course, is producing a vanilla Jupyter Notebook where you make use of some standard charts. With this course, you should have advanced much further on the learning curve.

What do we expect from each you ?

- In a nutshell, **we want you to think as a data-scientist throughout the whole production pipeline:** retrieving & pre-processing data, exception handling, building a model, hyperparameter optimization, etc...
- We expect that you bring the topics explained during the course into practice. You should be able to bring value to the data. You can use techniques that were not covered during the course and can bring other python packages into the project.
- Make sure you start from the same python environment, used in the course. Of course you can update packages, install new ones,...
- Make sure that you understand the underlying mathematics in the approach that you use (supervised, unsupervised, nlp, AI,..). A data-scientist is much more than an expert in Sklearn, NLTK, Pytorch,etc...

What do you need to hand in ?

The deliverables shall consist of:

- a program (python only !)
- a report
- a presentation

Deliverable

The project has to result in **Three** deliverables

1. Your Python Code shared on a GitHub account
   This should be either Jupyter Notebook(s) or an App
2. Report (pdf) of maximum 3 pages
3. You will be invited for a presentation on-campus or on-line (depends on COVID-19)

Presentation

August 27 (Exact schedule will be made available before that date).
If there are a lot of students, a second date will be scheduled.

Delivery Date

1 week before the exam dates.

Delivery Mechanism

You will be assigned an S3-bucket on AWS. There is no need to have an AWS account yourself.
This bucket has three folders "data","code" and "report".

Data

In this folder you save all the data you have used (or links to the data) to solve your project

Code

If you are not able to deliver your python code in a GitHub folder, you can use this folder in the AWS S3-bucket.

Report

In this S3-bucket you can drop your report (pdf)

Grading

Based on the presentation and the way you answer questions.
The presentation shall be no longer than 10 minutes. The questions will take 5 minutes,

Grading Criteria

Below is in bullet-point format a non-exhaustive list of the criteria that we will take into account when we evaluate your work.

Modeling

- Are you able to reach out to different data-sets outside the assigned dataset ?
- Visualisation
- Code: Style & Organisation of your Python Code.
- Does the code actually work ?
  We should be able to clone your code on github and run it on our computer. Make sure that you use a requirements.txt file to specify the python packages you require.

- Delivery App
  If you deliver an App, the code should be on S3. The app should be deployed.

Content of the report

- Your pipeline : from retrieving data to the actual model
- Introduction and problem statement
- Research method & scientific character of the work done
- Argumentation
- Results: discussion / interpretation
- General conclusion
- Coherence / logical composition
- Originality & creativity
- References

Presentation

- Presentation: used language
- Presentation: content / accessibility
- Presentation: form / composition / timing
- Understanding underlying mathematics
- Answering questions on Python and ML Code

Failure

There are two ways to obtain a "no-pass" result:

- Your project did not receive a pass grade
- You did not hand in a report in time.

PROJECT AUGUST

KAYA IDENTITY

Introduction

The Kaya Identity makes a link between the CO2 emissions and:

- ***Population***
- ***GDP per capita***
  Describes how much an average person of a given country actually produces (=economic output)
- ***Energy Intensity***
  This describes the amount of energy required for a unit of economic output
- ***Carbon Footprint of the Energy***
  This number decreases if clean sources of energy are used.

Proposed Research Question

In your project, you have to investigate the link between the CO2 emissions of a given country and its population. Is there a link? How does population growth relates to CO2 emissions ? Do you see differences across different countries and regions. Is there a trend ?

Initial Data Set

For CO2 emissions, we attached an Excel file to this project ("CO2Highlights"). You can easily find data for population growth across the internet. There are many reliable sources.