Data Science Lab: Process and methods
Politecnico di Torino

Project Assignment
September Call, A.Y. 2022/2023

Last update: September 8, 2023

1 Project dates

Start date: August 31, 2023 at 23:59 (CET)
Due date: September 17, 2023 at 23:59 (CET)

Due date is a **strict deadline**.

2 Problem description

Social scientists frequently employ census data to examine the levels of inequality pertaining to income, employment, education, housing, and various other facets of life. These significant analyses should naturally contribute to discussions about bias in categorization situations within these fields. Census data can thus be utilized for practical investigations into the fairness of algorithms.

In this project, you are required to predict the commute time to work for each individual using comprehensive census data. The study's findings hold potential implications for understanding socioeconomic disparities and informing evidence-based policy decisions.

2.1 Dataset

> **Warning:** For this project, you are not allowed to use external datasets other than the one provided. Adoption of external resources, including pre-trained models, will result in failure of the exam.

The dataset contains 130,801 instances describing the travel time to work based on an individual's personal information. The travel time is measured in minutes. Several demographics and other metadata characterize each record. The following is a short description of each of them:

- COW: Class of worker
- SCHL: Educational attainment
- MAR: Marital status
- OCCP: Occupation
- POBP: Place of birth
- WKHP: Usual hours worked per week in the past 12 months
- SEX: Gender of the individual

- RAC1P: Recoded detailed race code
- MIG: Mobility status
- HICOV: Health insurance coverage recode
- LANP: Language spoken at home
- PAOC: Presence and age of own children
- PINCP: Total person's income
- PUBCOV: Public health coverage recode
- VPS: Veteran period of service
- DEAR: Hearing difficulty
- MIL: Military service
- MIGSP: Migration recode - State or foreign country code
- FER: Gave birth to child within the past 12 months
- ENG: Ability to speak English
- JWAP: Time of arrival at work - hour and minute
- JWDP: Time of departure for work - hour and minute
- OC: Own child
- FDEYEP: Vision difficulty allocation flag
- JWMNP: Commute time to work

The dataset is located at: this URL.

Within the archive, you will find the following elements:

- **development.csv** (development set): a comma-separated values file containing the records from the development set. This portion does have the `JWMNP` column, which you should use to train and validate your models.
- **evaluation.csv** (evaluation set): a comma-separated values file containing the records corresponding to the evaluation set. This portion does not have the `JWMNP` column.
- **sample_submission.csv**: a sample submission file.
- **data_dictionary.txt**: a list of available features and the encoding used (name and values).

2.2 Task

You are required to build a regression pipeline to predict the commute time to work (in minutes) of each individual in the Evaluation Set.

2.3 Evaluation metric

Your submissions will be evaluated through R^2 Score (coefficient of determination).

3 Submit your result

Submission file To get your results evaluated, you have to upload a result file on our submission competition platform, Data Science Lab Environment (DSLE). The submission file must be a CSV file formatted as follows:

```
Id,Predicted
104642,21.0
104643,40.0
104644,68.0
104645,67.0
...
```

The submission file must contain a header line and a row for each record in the Evaluation collection. Each row must have two fields:

- the Id of the corresponding record in the Evaluation set. It corresponds to the column Id in the evaluation CSV file.
- the Predicted label for the corresponding record.

You can find a sample submission file in the project material (see 2.1).

Submission platform The submission platform is the same one you used during the course laboratories. Therefore, you have to use your personal key. If you do not have the key or have problems using it, please send an email to lorenzo.vaiani@polito.it. Please refer to the guide on the course website to go through the submission procedure.

You can find the DSLE platform at http://trinidad.polito.it:8888

4 Upload the report and the software

The report and the software have to be submitted by the due date reported in Section 1. This is a strict deadline.

Submission All the required files (i.e., for the report and the software) must be included in a **single ZIP archive**. The archive must be uploaded to the "Portale della Didattica", under the *Homework* section. Please use as description: **report_exam_september_2023**.

> **Info:** A ZIP archive is a ZIP archive, not a RAR, a 7z, or a tarball archive, nor any of those renamed with a trailing .zip extension.

Formatting rules The formatting rules for both the report and the software are described in the document with exam rules. You can find it on the course website.