IA|BE Data Science Certificate - Module 1

Third edition October 2022

Instructions for the Assignment

Please provide your answers in Dutch, French or English. You should demonstrate in your solution that you master the methods that have been discussed in Module 1. You are not supposed to use any approaches or methods that have not been covered in the Module 1 sessions. You must use `Python` for your calculations and graphics, in line with the computer labs of Module 1. Occasionally, and if justified, you can call a specific `R` library from `Python`.

Success!

Deliverables for the Assignment

Please hand in on or before February 1 via email (to Katrien Antonio and Gerda Elsen):

1. A notebook (.ipynb) or link to a Google Colab that documents your modelling steps and provides guidance for the reader, including some discussion of your findings and obtained insights. Your code should be well-organized and easy to read.

Please mention the names of your team members on your submission. It is allowed to work in teams (with two participants maximum); it suffices to submit one solution per team.

Assignment Questions

You analyze the data set (in `.csv`) that is available on the home page of Module 1. This data set contains observations on the variables listed in the table printed below. Your report should document the following steps:

1. An exploratory data analysis.
2. The construction of a (technical) tariff structure for a car insurance product. Hereto you analyze the frequency and the severity information in the data set with (at least) one of the predictive methods/algorithms discussed in the Module 1 lectures (e.g. LM, GLM, GAM, regularized GLM). You discuss the essential insights obtained with these models for frequency and severity. You define some risk profiles and illustrate the predictions obtained with your models for these risk profiles.

There is no need to answer the above questions separately (question by question) in your report. A well structured text that covers the above items is preferred. Be creative and rigorous! While the data set lists the 4–digit postal code of the policyholder, we do not expect a detailed analysis of the spatial heterogeneity (as this goes beyond the methods explained in Module 1).

Katrien Antonio, Jonas Crevecoeur, Roel Henckaerts and Michael Lecuivre.
Version December 2022.

ageph	age of the policyholder
CODPOSS	postal code in Belgium
duree	exposure, fraction of the year the insured is covered
lnexpo	log of exposure
nbrtotc	total number of claims during period of exposure
nbrtotan	rescaled number of claims for exposure equal to one
chargtot	total claim amount
agecar	age of the car: $0-1$, $2-5$, $6-10$, >10
sexp	sex of the policyholder: male or female
fuelc	type of fuel: petrol or gasoil
split	split of the premium: monthly, once, twice, three times per year
usec	use of the car: private or professional
fleetc	car belonging to a fleet: yes or no
sportc	sport car: yes or no
coverp	coverage: MTPL, MTPL+, MTPL+++
powerc	power of the car: <66, 66-110, >110