LDAT 2310 : Data science for finance and insurance.

Individual project

Download the databases “DBtrain.csv » that contains information about 70 000 insurance policies and the database “DBtest.csv”. The databases are about Motor insurances.

Explanatory variables in the databases are

Gender : 1 for male drivers and 2 for female drivers

DriverAge : driver’s age

CarAge : age of the vehicle

Area : 1) suburban, 2) urban, 3) countryside low altitude, 4) countryside high altitude (mountain regions)

Leasing : 1) Yes, 2) No

Power : 1) low horsepower, 2) normal horsepower 3) intermediate horsepower 4) high horsepower

Split : splitting of the premium 1) Monthly, 2) Quarterly, 3) Yearly

Contract : type of guarantees. 1) basic 2) intermediate 3) full

Exposure : Duration of the contract in years

NbClaims : number of claims

1) Use techniques seen in class (GLM, Regression trees, RF ,GBM) to propose a model for the claim frequency. Determine which factors discriminate best between claims frequency and severity.
2) Choose the best method and predict the claim frequency for the contract in the test database “DBtest”. You will complete the csv file. DBtest by adding one column reporting predicted claim frequencies.

The final report is limited to 15 pages, but appendix are allowed. A printed version must be handed before the 30th of November (12AM). A pdf version AND the file DBtest completed with forecasted claims frequency must be sent to your instructor: donatien.hainaut@uclouvain.be

WARNING: the database DBtest.csv must have the following format:

1) Semi-column (;) as separator
2) Comma (,) for decimal numbers