**LDAT 2310 : Deep learning for insurance & finance.**

**Individual project**

Download the dataset “Car Insurance Claim.xlsx» that contains information about 8423 insurance policies. This dataset comes from SAS Enterprise Miner. The explanatory variables in the databases are:

|  |  |
| --- | --- |
| Driving Children | # of children as passengers |
| DOB | Date of birth |
| Age | Age |
| Home Children | # of children at home |
| Years on Job | Years on job |
| Income | Income in $ |
| Single Parent? | whether a single parent |
| Home Value | Home value in $ |
| Marital Status | whether married |
| Gender | Gender |
| Education | 5 levels |
| Occupation | 8 levels |
| Travel Time | time to travel from home to work |
| Car Use | whether the car is for commercial use |
| Vehicle Value | Car value in $ |
| Time In Force | Years as costumer |
| Car Type | 6 categories |
| Red Car? | Whether the car is red |
| Total Claims (5 Years) | Total claims, years t-6 to t-1 (\*) |
| Claims Frequency (5 Years) | # of claims (\*) |
| License Revoked | Whether the license was revoked once |
| Vehicle Points | motor vehicle record point |
| City Population | whether the driver lives in urban area |
| Car Age | Age of car |
| Claims Flag (Crash) | Binary claim flag, year t |
| Claims Amount | Claim amount, year t |
| (\*) Warning : 5 years if 'Time in Force' is >= 5yrs | |

The exposure is 1 year (contract with a duration lesser than 1 year are discarded).

**1) Supervised learning (10 pts)**

Fit a few neural network models (minimum 4) for predicting claim frequencies (to be programmed in Keras) and select the best architecture (criterion: deviance).

* + 1.1 Compare the predictions with the observed claims frequencies for relevant subgroups (global, education, gender, city, …). Eventually adjust the bias. **(5 pts)**
  + 1.2 Determine which factors discriminate best between claims frequency using the model interpretation tools seen in class (PDP, ICE). **(3 pts)**
  + 1.3 Use LIME and Shapley’s value to analyze two policies. **(2 pts)**

**2) Unsupervised learning (10 pts)**

Convert the variable “Age” in 10 categories.

* Fit a variational auto-encoders to the following categorical variables (select the right loss function)

|  |  |
| --- | --- |
| Age | Age (categorized) |
| Single Parent? | whether a single parent |
| Marital Status | whether married |
| Gender | Gender |
| Education | 5 levels |
| Occupation | 8 levels |
| Car Use | whether the car is for commercial use |
| Car Type | 6 categories |
| City Population | whether the driver lives in urban area |

* Check the quality of the reconstruction for at least 4 different auto-encoders **(5 pts)**
* Identify clusters of insured with the K-means algorithm **(3 pts)**
* Select the optimal number of cluster according to the deviance **(2 pts)**

The final report is limited to 20 pages, but appendix are allowed (but limited to 10 pages). You must upload the following elements on the Moodle platform before the 30th of November (12AM):

1. A pdf version of your report
2. A zip archive with the code