Candidate evaluation test

For each of the following problems, please send us your written solution, including an attachment with any code written and a short explanation of the models and/or the validation approach followed.

Feel free to use any software or programming language, as long as they are easily obtainable for us.

Problem 1

The scenario is the following: you are building a MTPL Frequency model on your portfolio. You've already come up with a baseline prediction, based on the data that was already available in your insurance portfolio.

You then discover an external data provider, that claims it can enrich your portfolio with new data, which will greatly enhance the predictive power of your existing model. Your task is to assess whether this claim is true.

Database:

Files *train.csv* and *test.csv* are two non-overlapping portions of your insurance portfolio, for the experience years ranging from 2017 to 2019. The following variables are included:

- *id*: unique row ID
- *expy:* policy exposure, measured in policy years
- *base_pred* : the expected annual frequency for the policy
- *var1.bin – var40.bin:* the variables provided by the external provider and attached to the database. You can assume these variables to be all binary (1=TRUE, 0=FALSE)
- *var1.cat – var3.cat*: categorical variables available in your database;
- *var1.num – var2.num*: numerical variables available in your database;
- *period*: calendar year of each observation.

Additionally, the file *train.csv* also contains

- *n_claim*: MTPL claim count

Note: each record represents a policy year exposure of a policyholder. The exposure is measured in policy years. You can assume that the base_pred already takes into account some information regarding the policyholder (such as its BM class, age, etc.), as well as vehicle information (Engine power, make/model, bodywork, fuel type…) and territorial type information (zip code, population density, urbanization,….). For the sake of the exercise, it does not matter what the additional variables refer to.

Questions:

1- Using the file *train.csv*, build a frequency model for the yearly expected claim frequency (n_claim/expy), using both the baseline prediction and the additional variables. The model should be built with the objective of improving upon the predictive power of the baseline model. **Describe the approaches used, with particular attention to procedures like – but not limited to – handling missing values, model structure design, model validation strategy, and validation metrics used. Quantify the improvement of predictive power, with KPIs and goodness of fit metrics.**

2- Make a prediction of the expected frequency on the file test.csv, using the model built on Question 1, and call it "expected_freq". **Send back a csv file, containing the columns: ID, expected_freq**.

3- Answer the following:

   a. What family of models did you use for Question 1?
   b. What other choices did you have?
   c. Which criteria did you use to select the family you used?

Problem 2

Consider the file *Tarif_Structure.xlsx*. Each sheet colored in blue contains a table used for determining the tariff of a motor insurance. The green sheet named "Calculator example" has an example of "standard" rating structure, for a particular policyholder whose data are reproduced in the light-yellow box on the top-left. The calculator shows the final TPL standard premium. You can assume that the coefficients given are technically and commercially sound, and they are made so that the TPL standard premium yields a sustainable and profitable tariff, and still sell competitively.

Questions:

1. Assume you want to launch a PAYD (Pay as You Drive) product. You are given the necessary adjustment and price-per-Km (*ppk*) factors (see cells with prefix PAYD). Build a proposal for the PAYD premium, considering a proper amount of km as baseline and taking into account both adjustment coefficients and all the traditional rating factors.
Comment on the telematics proposition chosen if necessary.
Hint: Try multiplying the PAYD adjustment and km factors.

2. Create a new sheet and simulate an insurance portfolio generating a minimum set of profiles by properly varying the rating factors. Based on the obtained portfolio, compare standard vs PAYD premium. Justify and comment the obtained results

Problem 3

An actuary produced two GLM models to assess the expected cost of a book of policies. In the table below we sampled a few specific profiles, reporting the two predictions and the observed loss cost.

Profile id	*Actual Loss Cost*	*Model 1*	*Model 2*
1	337	497	433
2	653	499	550
3	385	468	427
4	632	513	602
5	331	485	397
6	558	501	517
7	349	585	419
8	446	480	452
9	667	532	662
10	487	491	485

Discuss which of the two models you would use.