Insurance Data Science
- Assignment -

2023/2024

From 03/06/2024 to 30/06/2021

DATA:

File autodata.txt

File claimsdata.txt

PROBLEMS:

Consider you are working for the IMS Insurance Company.
You are asked to evaluate last year's claims data of the Automobile Insurance portfolio.

PART I

As a first task, you are asked to answer the following questions on a small report:

1. Perform a descriptive statistical data analysis of the *Number of Claims* of the Third Party Liability on Automobile Insurance. Comment on the features observed and highlight values or patterns that you think are important to characterize the phenomenom.

2. Perform a descriptive statistical data analysis of *Claims Severity* of the Third Party Liability on Automobile Insurance. Comment on the features observed and highlight values or patterns that you think are important to characterize the phenomenom.

3. Fit distributions to the Number of Claims and Claim Severity.

   - For the Number of Claims, remove the highest outlier from data. Refer that fact on your report.

   - For Claims Severity, choose an upper bound that allows you to fit a distribution of the Exponential Family. Refer, on your report, the upper bound considered and the number of claims removed from data and give a comment on the choice of the upper bound.

   - What is the mean value and standard deviation of the claims removed from data in question 3? Plot the removed data in a histogram and a boxplot. Comment. Give your opinion on how should the insurer include that data on the final premium structure.

You are now asked to propose a Pricing Structure for the Automobile Insurance portfolio.
Using all the results from Part I, you are asked to answer the following questions:

1. Fit a GLM to the Number of Claims data and estimate the claim frequency for each risk profile in your portfolio.

   - Detail and justify your model assumptions and choices.
   - Improve your model, using adequate statistical tests.
   - Evaluate and comment on the quality of the model.
   - Identify the Standard Insured characteristics and the correspondent claim frequency estimate.
   - Include any comments you consider relevant.

2. Fit a GLM to the Claim Costs of "common" claims.

   - Be clear about your definition of "common" claim.
   - Detail and justify your model assumptions and choices.
   - Improve your model, using adequate statistical tests.
   - Evaluate the quality of the model.
   - Include any comments you consider important.

3. Propose a Pricing Structure to the "common" claims. Identify the highest and lowest insured's risk profile and the correspondent premiums to be charged. Compare with the premium of the standard insured.

4. Propose a model that allow you to include the large claims in the Pricing Structure. We may choose some Machine Learning model that, accurately predicts the probability of reporting a large claim. Justify your choice.

TOOLS:

Software R Project / Python

Microsoft Word / Latex

DELIVERABLES:

- Report (around 12 pages) with answers and comments to the insurer.

- R/Python files developed to complete the report.