Case Study: "Underwriter for a Day"

Introduction:

It is December 2024. You are the Chief Underwriter at the SafeRoads auto insurance company tasked with setting premiums for the next calendar year. Your primary objective is to ensure that premiums are accurately priced to cover anticipated losses while maintaining financial stability. The previous risk classification was lacking in rigour and was hurting the competitive position of your insurance company.

Your mission encompasses four key aspects:

1. **Risk Group Assignment Algorithm:** Develop an algorithm for categorizing new applicants into specific risk groups based on their individual characteristics.
2. **Predicting Total Losses:** Based on historical claims data and utilizing probability distributions, you need to model the total losses expected for the upcoming year.
3. **Premium Determination:** Assign appropriate premiums to each risk group, ensuring that the likelihood of not being able to cover claims next year does not exceed 0.5%.
4. **End-of-Year Assessment (evaluation):** After having collected data for the following year (2025), assess the quality of your underwriting algorithm.

To achieve this, you have at your disposition historical data on insurance claims filed in the previous year ("claim_data_groupX_2024.csv"), which includes essential variables for risk assessment:

- IDpol: a unique identifier for an insurance policy
- ClaimNb: the number of claims filed for a particular insurance policy
- Fraction of the calendar year that was covered
- Area: categorical variable indicating the area code
- VehPower: the power of the car (ordered categorical variable)
- VehAge: the vehicle age, in years
- DrivAge: The driver's age
- BonusMalus: Bonus/malus, between 50 and 350: <100 means bonus, >100 means malus
- VehBrand: The car brand (unknown categories)
- VehGas: The car fuel: Diesel or regular.
- Density: The density of inhabitants (number of inhabitants per km2) in the city the driver of the car lives in.
- Region: The policy region

Approach:

It may be useful to start your analysis by doing a descriptive analysis of the data – what is the distribution of individual claims, and what are total claims by subgroups (e.g. by driver's age, bonus/malus, etc.)? This will help you get a feeling for what the data looks like before applying more sophisticated statistical techniques.

1. **Risk Group Assignment Algorithm:** Develop an algorithm that groups insured based on their observable characteristics such as age, accident history, vehicle type, and other relevant factors that emerge from your analysis. This algorithm is an automatic decision rule, which sorts every new applicant with a certain set of characteristics into a single risk class. Referring back to the lecture notes for criteria for good risk classes, justify your risk categorization system.

2. **Predicting Total Losses:** Estimate the distribution of total losses for the next year. This involves estimating parameters for the frequency and severity distribution based on historical claims data and simulating the probability distribution for total losses next year. Using the distributions for frequency and severity seen in class, match the first two moments of the empirical distribution. Choose one distribution for severity and frequency, or get bonus points for checking which distribution has the best goodness-of-fit!

3. **Premium Determination:** Calculate premiums for each risk group by incorporating expected losses, while ensuring that the probability of observing claims larger than premiums is less than 0.5%.

4. **(after you receive the second data set on Oct 31) Evaluation:** At the end of the next year, you will receive realized claims data. Assess the performance of your risk categorization and premium determination strategies by comparing them to realized losses which are recorded in "claim_data_groupX_2025". These are new policyholders which you classified in January 2025 based on their characteristics and your algorithm developed in task #1.

    a. Calculate the actual premiums charged for each policy based on your algorithm and pricing strategy from steps 2 and 3 and compare them to claims. Don't adjust the algorithm you already developed based on this new data! This is an out-of-sample test for your algorithm developed on the 2024 data.
    b. Compute the loss ratio for each risk category. Is the loss ratio stable across your different risk categories, or does it fluctuate significantly, i.e. are there specific risk categories that significantly outperform or underperform your predictions? What does the stability (or lack thereof) in the loss ratio suggest about the quality of your risk categorization and premium calculation?

Deliverables:

By 10/29/2024: First (preliminary, not graded) submission
10/31/2024: In-class discussion that day about what approach you used, what worked, what didn't, and what open questions you still have. The goal is to learn from what the other groups did and get ideas for improving your group's underwriting. Make sure you come to class prepared to explain the approach you've taken – even if you're not the one who did most of the analysis. At the end of class, you will receive the second data set with realized claims for the next year.

By 11/14/2024: Final submission: You will submit the final report including a section on adjustments you did after an exchange in class. This submission will be graded and counts for 20% of your final grade. Some notes on your submission:

1. I strongly suggest that you conduct your analysis in a Jupyter Notebook and submit a thoroughly annotated pdf-version of your Notebook as well as the Notebook file.
2. Alternatively, write your analysis in a separate document and submit the document as well as the code. Do not submit the code only.
3. Document all the steps you have taken and problems you have encountered (whether solved or not). There is no one right way of doing this analysis; I want to know how you approached the problem and what insights you're drawing from the process.
4. Specify which team member did which part.