1 Introduction

1.1 Kaggle French Motor Third-Party Liability (MTPL) Dataset

The *freMTPL2freq* dataset, published on Kaggle as part of the French Motor Claims Datasets,[1] is a widely recognized resource in actuarial science and insurance analytics. This dataset contains information on motor third-party liability (MTPL) insurance policies, specifically focusing on claim frequency modeling.

Dataset Overview

- **Source:** Kaggle – French Motor Claims Datasets (*freMTPL2freq*)
- **Number of policies:** 677,991
- **Observation period:** One-year exposure per policy
- **Main objective:** Predict the frequency of insurance claims per policy

Data Structure and Key Features The dataset contains a variety of explanatory variables that describe the policyholders, the insured vehicles, and the insurance contracts. These features include:

- **Policyholder characteristics:** Age of the driver, gender, and driving experience.
- **Vehicle attributes:** Vehicle age, vehicle brand, and vehicle power.
- **Policy details:** Type of coverage, bonus-malus level (No-Claim Discount), and geographic area of residence.
- **Exposure:** The fraction of the year during which the policy was active.
- **Claim information:** Number of claims filed within the exposure period.

Relevance The *freMTPL2freq* dataset is particularly suitable for analyzing

Table 1: Description of the generated dataset variables

Variable	Meaning	Type
`postcode`	Hungarian postcode	Categorical
`licence_age`	Age of the driving licence (years)	Numeric
`n_drivers`	Number of drivers	Integer
`young_driver`	Age of the youngest driver	Numeric
`old_driver`	Age of the oldest driver	Numeric
`domestic`	Expected kilometres driven in Hungary next year	Numeric
`foreign`	Expected kilometres driven outside Hungary next year	Numeric
`is_financed`	Vehicle financed (1 = yes, 0 = no)	Binary (0/1)
`is_rh`	Right-hand steering wheel (1 = yes, 0 = no)	Binary (0/1)
`fuel`	Fuel type (e.g., petrol, diesel, electric)	Categorical
`odometer`	Current odometer reading (km)	Numeric
`daily_commute`	Daily commute time (minutes)	Numeric
`exam`	Recent technical exam within the last 180 days (1 = yes, 0 = no)	Binary (0/1)
`non_payment`	Termination of previous insurance due to non-payment (1 = yes, 0 = no)	Binary (0/1)
`casco`	Casco (comprehensive insurance coverage) (1 = yes, 0 = no)	Binary (0/1)
`is_retired`	Policyholder is retired (1 = yes, 0 = no)	Binary (0/1)
`is_disabled`	Policyholder is disabled (1 = yes, 0 = no)	Binary (0/1)
`seasonal_tyre`	Seasonal tyre usage (1 = yes, 0 = no)	Binary (0/1)

[1] https://www.kaggle.com/datasets/floser/french-motor-claims-datasets-fremtpl2freq

2 Exposure

General claim frequency modeling has traditionally assumed that the occurrence of claims follows a Poisson distribution, implying that claims are proportional to the time at risk and independent of the observation period. However, there is a noticeable seasonality effect: November shows the highest claim frequency, while January and February show the lowest. Furthermore, the impact of COVID-19 curfews has introduced anomalies that are difficult to explain. Post-COVID, there has been a consistent decline in claim frequency.

To address seasonality and the pre/post-COVID differences, we leverage country-wide claim count summaries available on a quarterly basis.

3 Postcode-Level Territorial Modelling

Since some postcodes have limited exposure due to biased data, we model the expected claim count for each postcode and interpret the deviation as the territorial impact.

Observations indicate a negative exponential relationship between the geographical distribution of claims and the travel distance (measured in hours under optimal traffic conditions). To incorporate this, we employ a distance-weighted Bayesian smoothing approach using the following function:

$$f_p = \frac{c_p + \alpha \sum_i c_i e^{-\beta \cdot d_{i,p}}}{e_p + \alpha \sum_i e_i e^{-\beta \cdot d_{i,p}}}$$

where:

- f_p = smoothed claim frequency for postal code p
- c_p = observed number of claims in postal code p
- e_p = exposure (weight) for postal code p
- $d_{i,p}$ = distance between postal codes i and p (in hours, optimal traffic conditions)
- α = smoothing parameter

This approach allows for a more stable estimation of claim frequencies for postcodes with limited data by borrowing strength from neighboring areas.

4 Custom Loss Function

To optimize profit, we introduce a custom loss function that adjusts the objective function based on predicted price (P), observed claim (A), and market price (M). This function prioritizes cases that have a more significant financial impact using the following formula:

$$L(P) = \max(M, A) - P + (M - A - (\max(M, A) - P)) \cdot \sigma(P, M - \epsilon, k)$$

where:

- k = sharpness parameter controlling the transition speed
- ϵ = small positive value to define the transition point
- $\sigma(P, M - \epsilon, k)$ = smooth step function defined as:

$$\sigma(P, M - \epsilon, k) = \frac{1}{1 + e^{-k(P-(M-\epsilon))}}$$

The step function σ mimics a soft thresholding behavior, enabling a smoother gradient than a hard cutoff and thus improving convergence during optimization and provides a convex objective function.

This loss function design aims to balance underpricing and overpricing risks, focusing the model's attention on cases with a higher impact on profitability. The smooth step function σ enables a gradual transition, making the optimization process more stable.

Recent research by Burka, Kovács, and Szepesváry (2021) demonstrates the potential of machine learning methods — such as random forests and neural networks — to outperform traditional GLMs in MTPL claim classification tasks. Their findings underscore the importance of incorporating complex interactions and nonlinearities in tariff models. Motivated by their methodology and evaluation framework, our approach also emphasizes model interpretability and economic relevance, especially under asymmetric financial risks.

Building on these insights, we constructed a baseline GLM model using manually engineered feature transformations. This included binning for categorical conversion and, where appropriate, optional polynomial terms — for instance, to capture nonlinear effects in variables such as the age of the driver's license.

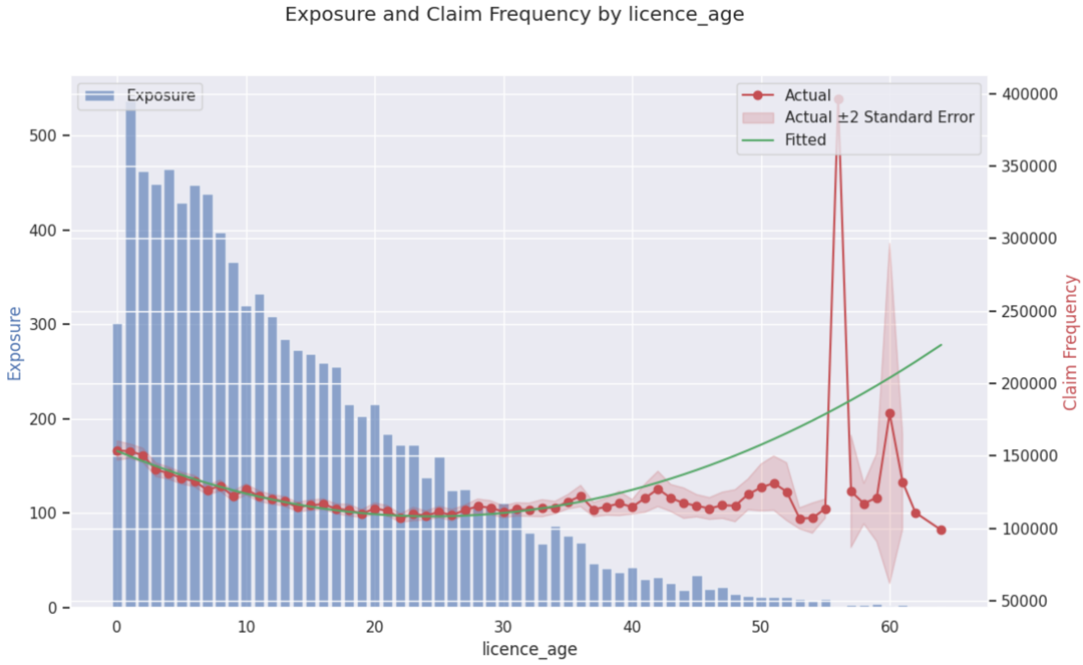

Figure 1: Your caption here

Where necessary, outlying high or low values were also manually grouped to ensure robustness and reduce sensitivity to extreme cases.

Tested Loss Functions

We experimented with the following loss functions involving the predicted price P, the observed claim amount A, and the market benchmark price M:

- **Blended Loss Function:**

  This loss interpolates between squared error to the true claim A and to the market price M, using a mixing parameter $\alpha \in [0, 1]$:

$$L_{\text{blend}}(P; A, M, \alpha) = (1 - \alpha)(P - A)^2 + \alpha(P - M)^2$$

- **Maximum-Based Relative Loss:**

  This function penalizes the squared deviation from the higher of the true and market values:

$$L_{\max}(P; A, M) = (P - \max(A, M))^2$$

- **Smooth Margin-Penalized Loss:**

  This function uses a sigmoid weighting centered at the market price M, smoothly interpolating between penalizing underpricing and overpricing:

$$L_{\text{smooth}}(P; A, M, k) = \frac{1}{1 + e^{-k(M-P)}} \cdot (M - P)^2 + \left(1 - \frac{1}{1 + e^{-k(M-P)}}\right) \cdot (P - A)^2$$

where $k > 0$ controls the sharpness of the transition between the two penalty regimes.

These loss functions exhibit distinct behaviors under different pricing scenarios. The figures below illustrate their shapes when applied to (i) a profit-making policy, (ii) a loss-making policy, and (iii) a properly priced policy. Each plot compares the three losses as a function of the predicted price P, holding A and M fixed.

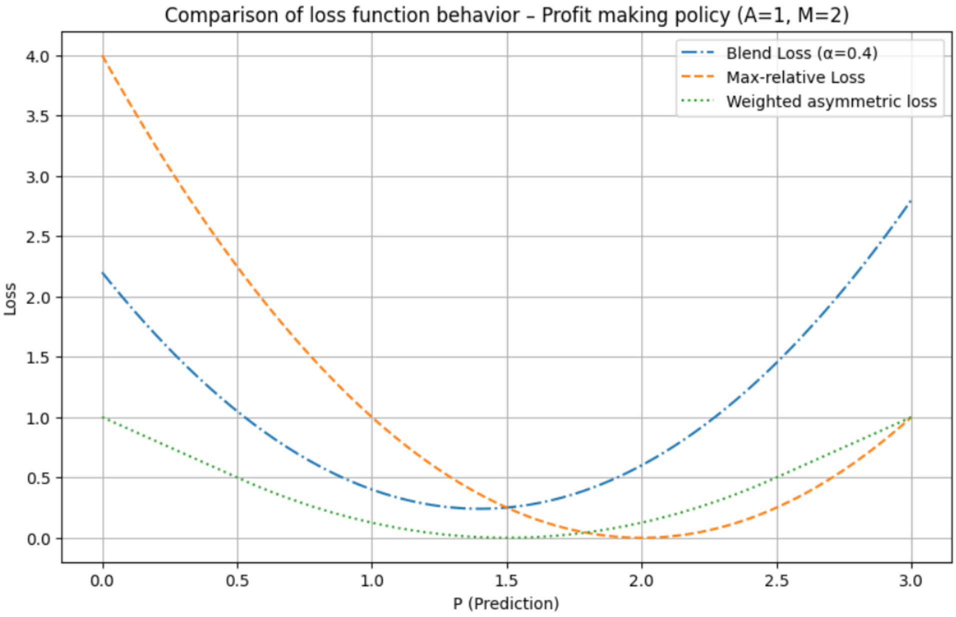

Figure 2: Loss function behavior – profit-making policy $(A = 1, M = 2)$

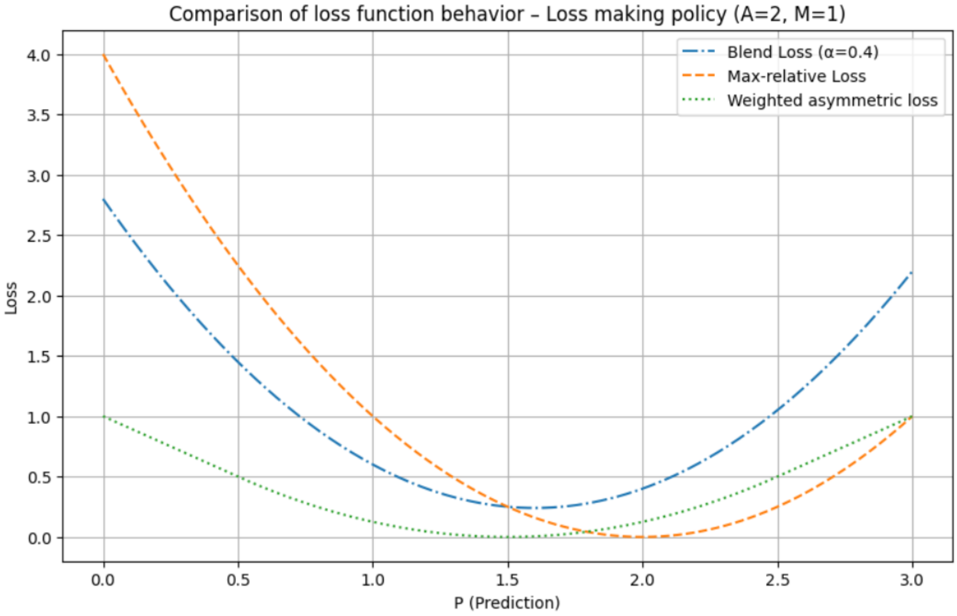

Figure 3: Loss function behavior – loss-making policy $(A = 2, M = 1)$

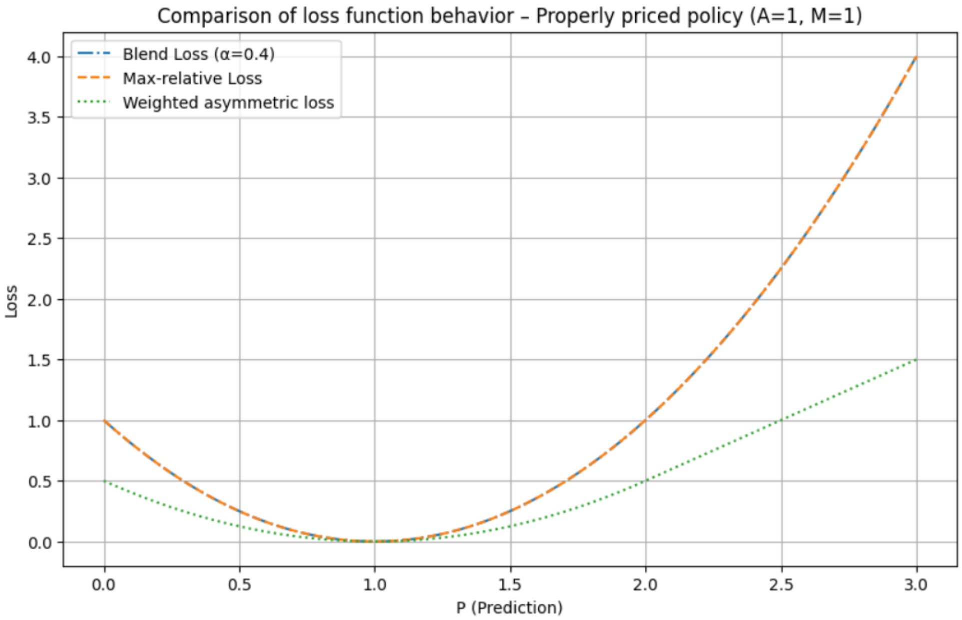

Figure 4: Loss function behavior – properly priced policy $(A = 1, M = 1)$