Predictive Modeling of Technical Reserves in Industrial Risk Insurance;Team 26.

Team 26/Mathurance hackathon

1 Background of the Problem

In the industrial risk insurance sector, accurately estimating technical reserves is critical for financial stability and risk management. These reserves cover both reported but unpaid claims and late-emerging claims. Traditional methods, such as Run-Off Triangles, struggle with:

- The increasing complexity of claim settlements.
- Variability in settlement amounts and timelines.
- The growing volume of data.

Given that a significant portion of our dataset contained missing values—66 % of the upper triangle was missing and 75% of the overall data had missing values—predicting technical reserves using traditional statistical methods alone proved inadequate. Machine learning approaches also faced challenges due to the missing data structure. Thus, a Monte Carlo simulation using a Pareto distribution was implemented to generate realistic estimates before applying machine learning models.

2 Methodology Used

2.1 Data Preparation

1. **Data Cleaning**: Checked for and removed duplicate records.
2. **Feature Engineering**: Created a *Development Year* feature to track claim progression.
3. **Data Aggregation**: Extracted *Year of Sinister* (from *Date de Survenance*), grouped data by *Year of Sinister* and *Development Year*, and computed *Total Règlement* (sum of claim settlements for each group).

2.2 Data Reshaping for Run-Off Triangles

1. Transformed data into a matrix where:
   - Rows represent the *Year of Sinister*.
   - Columns represent the *Year of Development*.
   - Intersection values represent *Total Règlement*.
2. Created sub-triangles based on the four insurance sub-branches.

2.3 Handling Missing Data

1. Identified missing values: 66% in the upper triangle, 75% overall.
2. Applied Monte Carlo simulation using a Pareto distribution:
    - Chosen because it models rare, high-intensity insurance claims(a characteristic of insurance claims).
    - Distribution parameters were selected based on historical data trends.
    - Missing values were filled to create a well-simulated dataset.

2.4 Reformatting for Machine Learning

1. Converted structured Run-Off Triangle back into tabular format:
    - **Features (X)**: *Year of Sinister*, *Year of Development*.
    - **Target (y)**: *Total Règlement*.
2. This format allowed for efficient machine learning predictions.

2.5 Model Selection and Training

1. Machine Learning models were applied per sub-branch:
    - **Random Forest Regression** for three sub-branches (sufficient data available).
    - **Linear Regression** for the fourth sub-branch (only 20 observations, making it the optimal choice).

3 Understanding the Context: Insurance Perspective on Claims Development and Settlement

In insurance, claims development and settlement analysis are crucial for risk assessment, pricing strategies, and reserve allocation. The analysis focuses on the lifecycle of claims, from the year of occurrence (sinistre year) to settlement, ensuring insurers maintain adequate reserves to cover future liabilities. The primary objectives behind our study are:

- **Understanding Claim Evolution**: Analyzing how claims evolve over time helps insurers predict future liabilities and adjust their financial reserves accordingly.
- **Risk Exposure Evaluation**: Identifying the most common sub-branches (types of insurance claims) allows for targeted risk mitigation strategies.
- **Reserving and Financial Stability**: By examining settlement trends, insurers can determine whether they have sufficient reserves to meet future claim obligations.

4 Development Year Distribution (Claims Maturity Analysis)

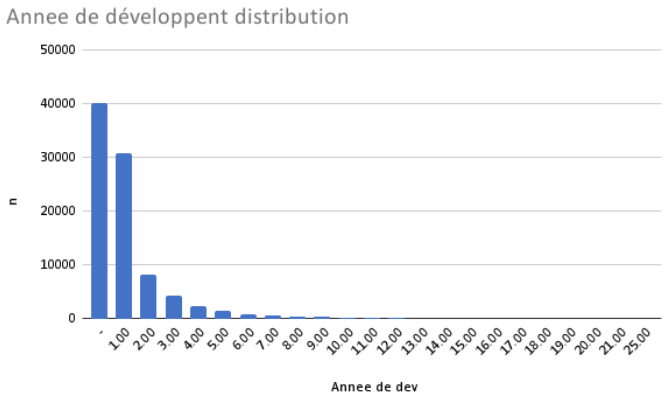

Figure 1: Development Year Distribution (Claims Maturity Analysis)

4.1 Observations

- Most claims are resolved within the first few years (1-3 years), with a steep decline in claim frequency beyond year 5.
- A long tail exists for certain claims, suggesting prolonged settlement for complex cases (e.g., liability cases or catastrophic losses).

4.2 Insurance Analysis

- **Short-Tailed vs. Long-Tailed Claims**: Short-tailed claims (e.g., property damage) are typically settled quickly, whereas long-tailed claims (e.g., liability and bodily injury) take years due to legal processes and medical evaluations.
- **Reserving Implications**: Insurers must maintain accurate reserves for long-tailed claims to ensure financial stability. The declining trend in claim frequency suggests that most liabilities are cleared early, reducing the risk of outstanding claims affecting future financials.

5 Sub-Branch Count (Risk Distribution by Insurance Type)

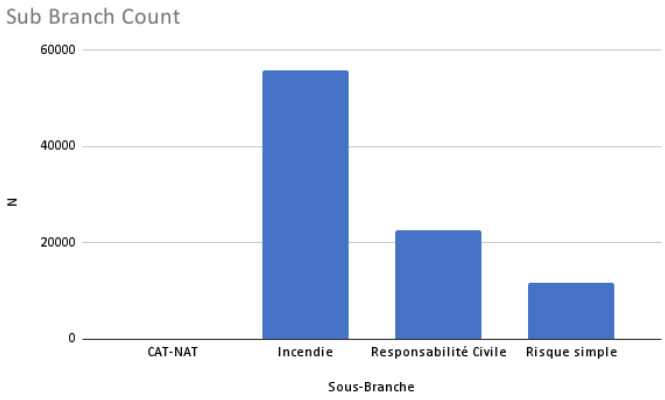

Figure 2: Sub-Branch Count (Risk Distribution by Insurance Type)

5.1 Observations

- The highest number of claims are concentrated in the "Incendie" (Fire) sub-branch, followed by "Responsabilité Civile" (Civil Liability) and "Risque Simple" (Simple Risk).
- "CAT-NAT" (Natural Catastrophes) has the lowest number of claims, possibly due to stricter underwriting policies or lower exposure.

5.2 Insurance Analysis

- **Risk Concentration**: The high number of fire-related claims suggests that insurers face significant exposure to property damage risks. This could indicate the need for higher reinsurance coverage or stricter underwriting criteria for fire insurance policies.
- **Civil Liability Exposure**: The notable presence of liability claims highlights legal and compensatory risks, which often lead to longer settlement periods.
- **Catastrophic Risk Management**: While natural catastrophe claims are lower in number, they typically involve large payouts. Insurers must ensure proper catastrophe modeling and reinsurance coverage to handle extreme but rare events.

6 Settlements Distribution Per Year (Claims Payout Trend Analysis)

6.1 Observations

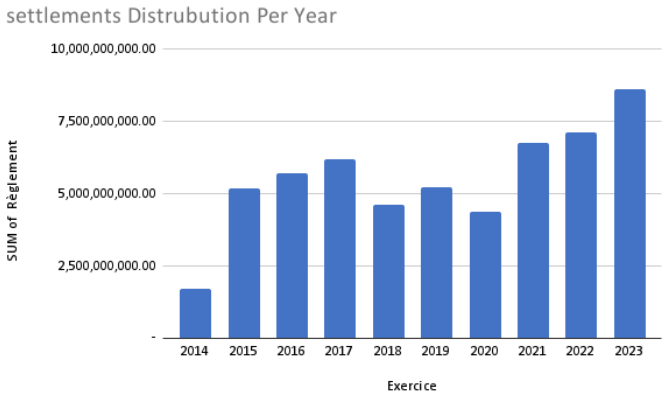

Figure 3: Settlements Distribution Per Year (Claims Payout Trend Analysis)

- Settlement amounts have been increasing over the years, peaking in 2023.
- A dip was observed around 2018-2020, possibly linked to economic factors or operational inefficiencies.
- The sharp rise post-2020 suggests an increase in claims frequency, larger claim amounts, or inflationary effects on claim payouts.

6.2 Insurance Analysis

- **Claims Inflation**: The rising trend in settlements could be due to higher repair/replacement costs, inflation, or an increase in insured values. This necessitates periodic premium adjustments to maintain profitability.

- **Underwriting Cycle Effects**: The dip between 2018-2020 may indicate a period of strict underwriting, where fewer risky policies were issued, or economic slowdowns led to fewer claims being reported.
- **Regulatory and Market Impacts**: Changes in government policies, legal frameworks, or court rulings can affect claim sizes, especially in liability insurance. The post-2020 increase may be due to relaxed underwriting, higher insurance penetration, or increased claims severity.

7 Sinistre Year Distribution (Claims Occurrence Trend Analysis)

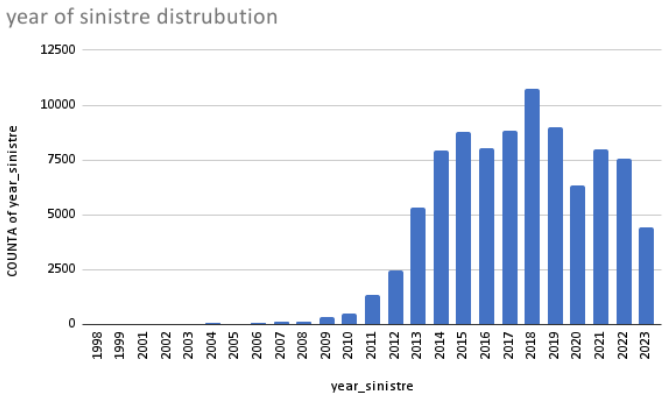

Figure 4: Sinistre Year Distribution (Claims Occurrence Trend Analysis)

7.1 Observations

- The number of reported claims increased steadily from 2005, peaking around 2017-2018 before slightly declining in recent years.
- The decline post-2019 suggests changes in either risk exposure, claim reporting behavior, or external factors like economic downturns affecting insurance demand.

7.2 Insurance Analysis

- **Insurance Market Growth**: The increase in claims from 2005 to 2018 could reflect growing insurance penetration, regulatory changes requiring more mandatory coverage, or an increase in insurable assets.
- **Improved Risk Management**: The slight decline after 2019 may indicate that businesses and individuals are adopting better risk management strategies, leading to fewer claims.
- **Pandemic and Economic Effects**: COVID-19 and economic fluctuations post-2019 could have affected claims reporting patterns, either reducing mobility-related risks (e.g., fewer car accidents) or increasing business-related claims due to financial strain.

8 Workflow Steps

1. Data Cleaning: Remove duplicates.
2. Feature Engineering: Create *Development Year*.
3. Data Aggregation: Compute *Total Règlement*.
4. Reshape into Run-Off Triangle.
5. Handle missing data using Monte Carlo simulation.
6. Reformat for Machine Learning.
7. Model selection and training.
8. Prediction and evaluation.

9 Results of the Model

- Successfully predicted technical reserves.
- Monte Carlo simulation completed missing values for accurate modeling.
- Random Forest Regression was the best fit for three sub-branches.
- Linear Regression worked best for the fourth sub-branch (20 observations).
- Predictions aligned well with historical claim trends.

category <chr>	Model <chr>	mse <dbl>
Responsabilité civile	Random Forest	4.847064e+14
Incendie	Random Forest	3.125033e+17
Risque simple	Random Forest	2.419168e+16
CAT-NAT	Linear regression	1.170308e+04

Figure 5: Results of Our Model

10 Conclusion

This study presented a structured approach to estimating technical reserves by integrating actuarial methods, Monte Carlo simulation, and Machine Learning models. By transforming raw data into a *Run-Off Triangle*, simulating missing values using a Pareto distribution, and applying predictive models, we developed a robust and interpretable solution for insurance companies.