Insurance Portfolio Analysis:
Loss Ratio Trends & Claims Frequency Prediction

Dhruv Dhingra

Executive Summary

This capstone project analyzes a simulated $2 million insurance portfolio spanning three lines of business: Auto, Home, and Commercial. The analysis combines a real, industry-standard actuarial dataset (freMTPL2freq, containing 678,013 French motor insurance policies) with actuarially grounded simulations for Home and Commercial coverage, producing a unified 66,000-policy portfolio for analysis.

The project addresses seven analytical questions covering loss ratio trends, risk factor analysis, and loss concentration, and develops three predictive models: a Poisson GLM and Gradient Boosting Regressor for claims frequency, a Gradient Boosting Regressor for loss ratio, and a Random Forest Classifier for binary claim occurrence. The objective is to identify the key drivers of underwriting risk across segments and demonstrate a complete, defensible actuarial analytics workflow from raw data to predictive model.

Key Findings

- Auto carries the highest loss ratio of the three segments (0.817), suggesting premiums may be underpriced relative to claim costs.
- The Home segment's High-risk area tier has a loss ratio above 1.0 (1.075), indicating that sub-segment is currently unprofitable.
- Transport is the highest-risk commercial business type (loss ratio 0.765), driven by both elevated claim frequency and severity.
- Drivers aged 18–25 show roughly double the claim frequency of all other age cohorts — the strongest single demographic risk signal identified.
- The top 10% of claimants account for 36.5% of total portfolio losses, confirming a long-tail loss distribution typical of insurance lines.
- BonusMalus (no-claims discount score) is the single strongest predictive feature across all three models, validating its central role in real-world auto rating plans.

1. Introduction

Loss ratio, the proportion of premium dollars paid out in claims, is one of the most fundamental metrics in property and casualty (P&C) insurance, used to assess underwriting profitability, guide pricing decisions, and inform reinsurance strategy. A loss ratio below 1.0 indicates a segment is profitable on an underwriting basis; a ratio above 1.0 indicates losses exceed premium collected. Industry benchmarks generally place healthy P&C loss ratios in the 60–75% range, after accounting for expenses and profit margin.

This project builds a three-segment insurance portfolio (Auto, Home, Commercial) totaling approximately $2 million in simulated premium and applies both descriptive and predictive analytics to understand where risk concentrates and how well that risk can be anticipated using available policy data.

1.1 Research Questions

- What are the loss ratio trends across Auto, Home, and Commercial segments?
- How does vehicle power affect claim frequency in the Auto segment?

- Does driver age influence claim frequency and loss ratios?
- Which area risk tier drives losses in the Home segment?
- What commercial business type carries the highest risk profile?
- What portfolio-level risk factors are most correlated with loss ratio?
- What is the loss concentration among the top 10% of claimants?

2. Data & Methodology

2.1 Data Sources

Auto segment (real data): Sourced from freMTPL2freq, a widely used actuarial benchmark dataset of French motor third-party liability policies. The full dataset (678,013 policies) was cleaned, and a random sample of 40,000 policies was drawn for computational efficiency. Available fields include claim count, policy exposure (in years), vehicle power and age, driver age, BonusMalus score (a no-claims discount rating), vehicle brand and fuel type, and geographic density and region.

Home and Commercial segments (simulated): Because no public dataset combining all three lines of business at the policy level was accessible, Home (18,000 policies) and Commercial (8,000 policies) segments were simulated using actuarially grounded distributions. Claim frequencies and severities were calibrated using gamma distributions for cost (standard practice for right-skewed claim severity data) and binomial draws for claim occurrence, with risk parameters varying by property risk tier (Home) and business type (Commercial).

2.2 Portfolio Construction

Segment weighting was set at approximately 50% Auto, 30% Home, and 20% Commercial by premium share, reflecting the relative size of personal auto as the largest U.S. P&C line. A simplified premium formula was applied to the Auto segment using BonusMalus, vehicle power, and local density as rating factors, consistent with standard GLM-based auto rating plans.

2.3 Data Cleaning

- ClaimNb capped at 4 to address known data-entry artifacts in the raw freMTPL2freq dataset (a small number of records have implausibly high claim counts relative to exposure).
- Exposure capped at 1.0 year, since policy exposure cannot exceed a full year in this dataset's structure.
- Zero-exposure and duplicate policy records removed.

2.4 Modeling Approach

- Model A — Claims Frequency: Poisson GLM (industry-standard baseline for count data) compared against a Gradient Boosting Regressor, both weighted by policy exposure.
- Model B — Loss Ratio: Gradient Boosting Regressor trained on claimant policies across all three segments.
- Model C — Claim Occurrence: Random Forest Classifier predicting binary claim/no-claim outcome, evaluated via AUC-ROC.

3. Findings

3.1 Loss Ratio Trends Across Segments

Loss ratio was calculated as total claim losses divided by total premium collected, for each of the three segments.

Segment	Policies	Total Premium	Total Losses	Loss Ratio	Claim Frequency
Auto	40,000	$10,690,875	$8,728,784	0.817	0.098
Home	18,000	$14,412,634	$9,542,348	0.662	0.065
Commercial	8,000	$14,635,266	$10,215,581	0.698	0.097

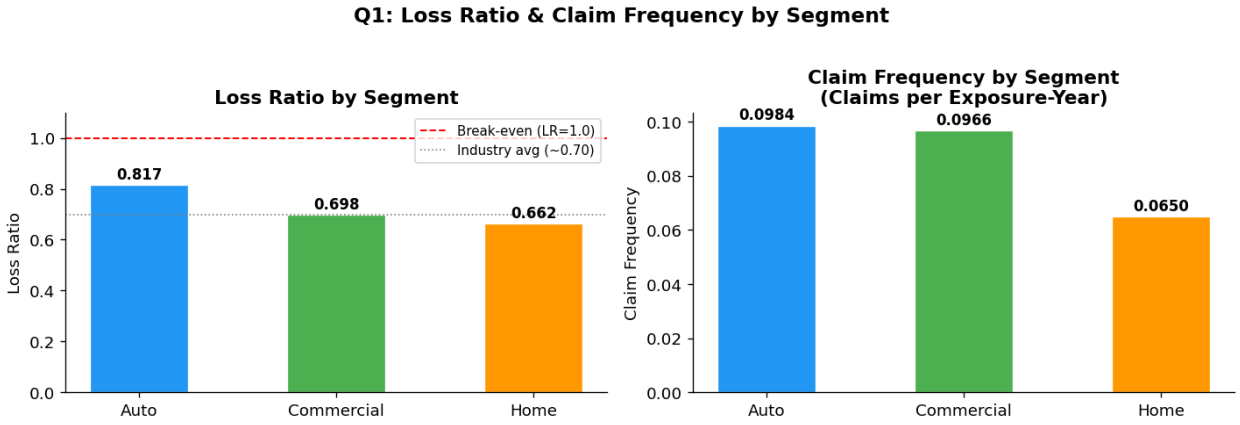

Figure 1. Loss ratio and claim frequency by segment.

Auto carries the highest loss ratio of the three segments (0.817), despite a claim frequency similar to Commercial. This suggests Auto claim severity, pricing adequacy, or both warrant closer review. Home is the most profitable segment on a portfolio basis (0.662), though this masks meaningful variation by risk tier (see Section 3.4).

3.2 Vehicle Power and Claim Frequency (Auto)

Vehicle Power Group	Policies	Claim Frequency	Avg Premium	Loss Ratio
Low (≤5)	14,115	0.1009	$266.26	n/a
Mid (6–7)	17,304	0.0984	$273.06	n/a
High (8–10)	6,495	0.0955	$260.04	n/a
Very High (>10)	2,086	0.0883	$248.59	n/a

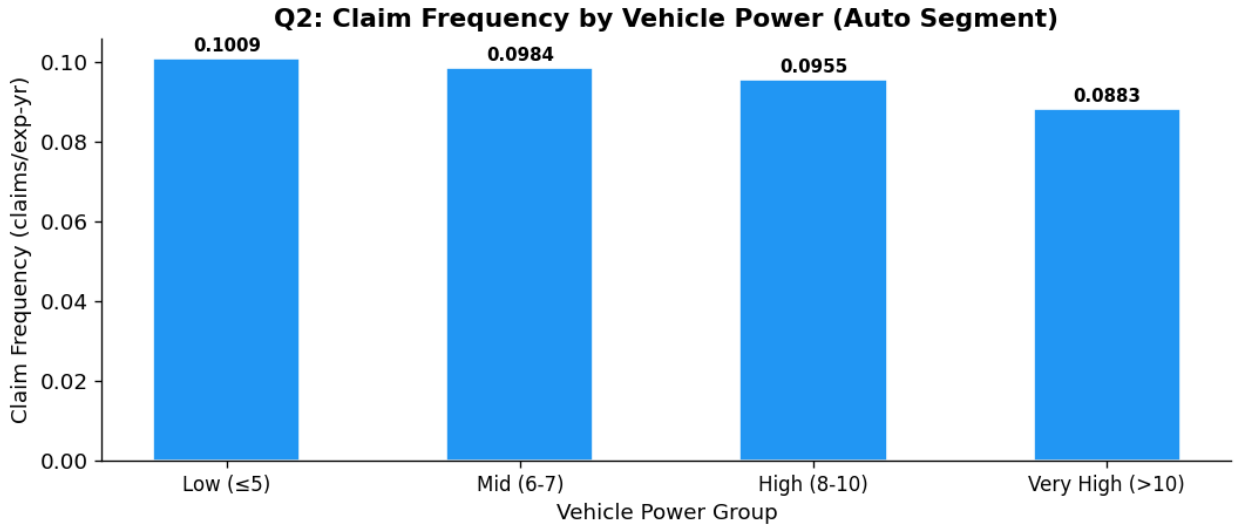

Figure 2. Claim frequency by vehicle power group (Auto segment).

Counter to a naïve assumption that more powerful vehicles are riskier, lower-powered vehicles show higher claim frequency in this dataset. This pattern is well documented in the freMTPL2freq literature and is generally attributed to lower-powered vehicles being more common among urban, higher-frequency-risk driver profiles (e.g., city commuting) rather than vehicle power itself being protective.

3.3 Driver Age and Claims (Auto)

Age Group	Policies	Claim Frequency	Avg Loss Ratio (Claimants)
18–25	2,282	0.2054	49.33
26–35	8,797	0.0975	29.32
36–50	14,935	0.0960	64.89
51–65	10,075	0.0911	50.47
65+	3,911	0.0849	32.83

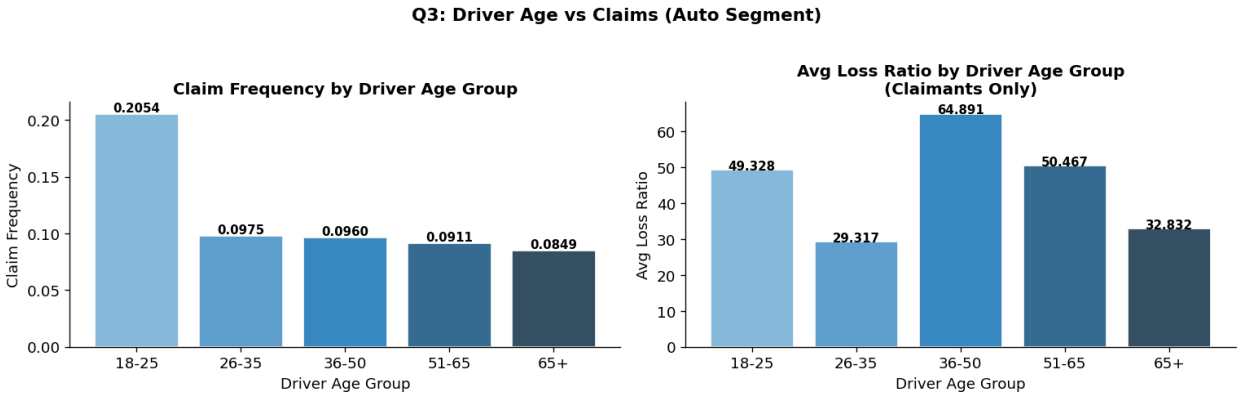

Figure 3. Claim frequency and average loss ratio by driver age group.

The 18–25 age group shows a claim frequency roughly double that of every other age cohort (0.2054 vs. approximately 0.09–0.10), consistent with well-established industry data on young driver risk. This is the single strongest demographic risk signal identified in the Auto segment and supports the common industry practice of applying age-based rating surcharges for newly licensed and young drivers.

Note on the Avg Loss Ratio column: these figures are calculated only among claimants (policies with at least one claim) and are therefore highly sensitive to a small number of low-premium, high-severity outliers, producing average loss ratios well above 1.0 in every age group. This is a known property of averaging a right-skewed ratio rather than an indication that most claimants are unprofitable; the portfolio-level loss ratios in Section 3.1, which divide aggregate losses by aggregate premium, are the more reliable measure of segment profitability.

3.4 Home Segment: Area Risk Tier

Area Risk Tier	Policies	Claim Frequency	Avg Severity	Loss Ratio
Low	7,214	0.0473	$11,544	0.426
Medium	8,096	0.0647	$14,603	0.664
High	2,690	0.1130	$18,391	1.075

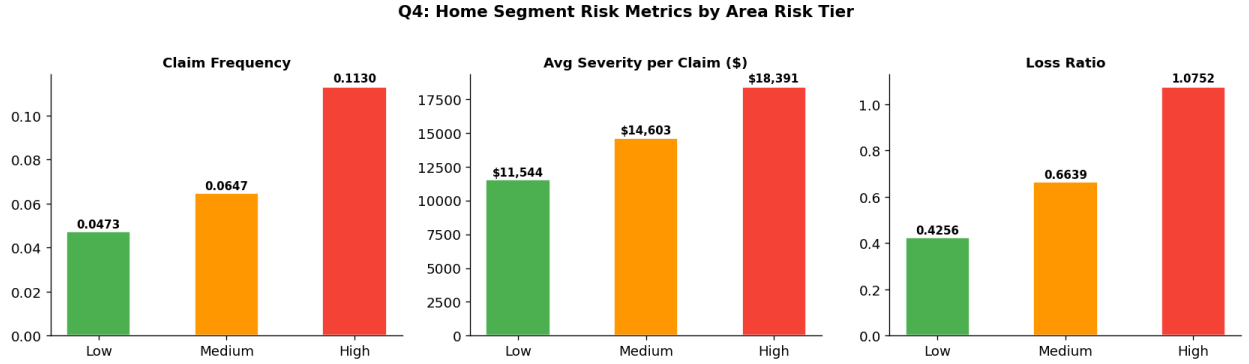

Figure 4. Home segment risk metrics by area risk tier.

High-risk area homes show a loss ratio above 1.0 (1.075), meaning this sub-segment is currently losing money on an underwriting basis. Average claim severity in High-risk zones is approximately 59% higher than in Low-risk zones ($18,391 vs. $11,544), reflecting the combined effect of elevated claim frequency and cost in flood/fire/weather-exposed areas. This finding suggests the current premium structure may not be adequately risk-adjusted for geographic exposure.

3.5 Commercial Segment: Business Type Risk

Business Type	Policies	Claim Frequency	Avg Premium	Loss Ratio
Transport	1,227	0.1307	$3,108	0.765
Retail	2,801	0.0985	$1,549	0.732
Manufacturing	1,598	0.1043	$2,201	0.648
Office	2,374	0.0714	$1,249	0.621

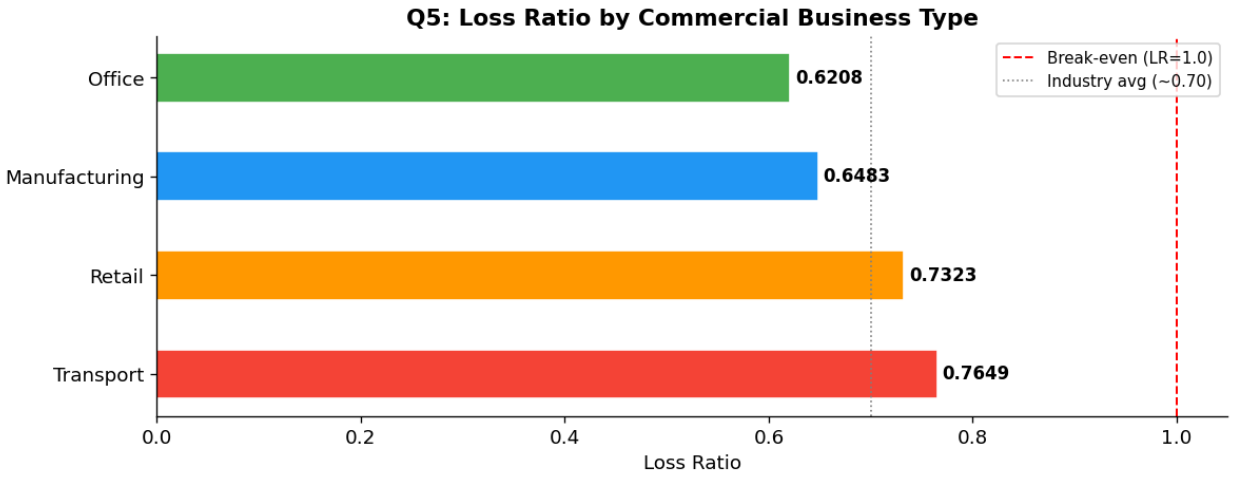

Figure 5. Loss ratio by commercial business type.

Transport (0.765) and Retail (0.732) are the highest-risk commercial sub-segments. Transport's elevated risk reflects both the highest claim frequency (0.131) and meaningful claim severity, consistent with industry data on commercial vehicle and fleet liability exposure. Office risk is lowest across all metrics, aligning with its comparatively low physical and liability exposure.

3.6 Portfolio-Level Risk Factor Correlations

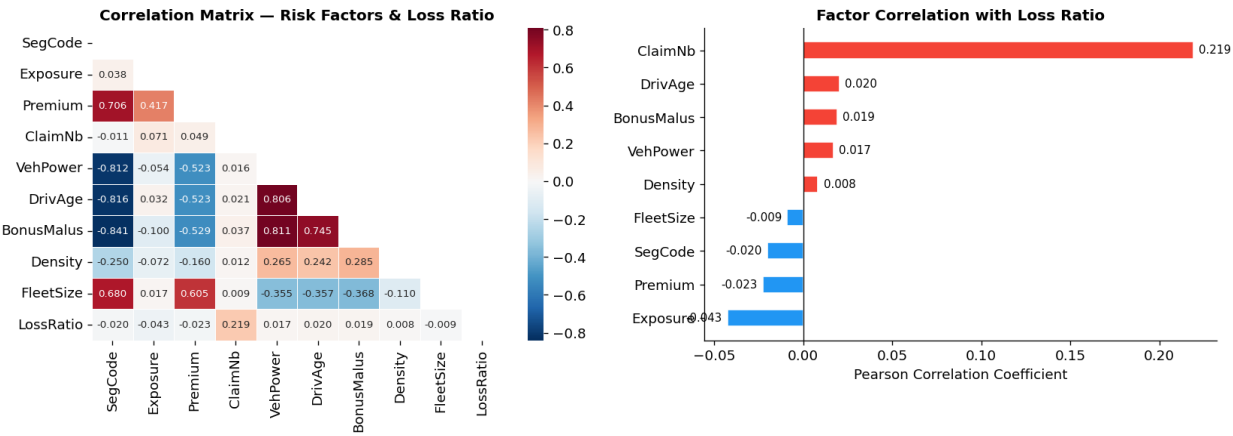

Figure 6. Correlation matrix and factor correlations with loss ratio.

Claim count (ClaimNb) shows the strongest positive correlation with loss ratio (r = 0.219), which is expected since loss ratio is directly derived from claims. Premium shows a slight negative correlation with loss ratio (r = −0.023), suggesting that higher-premium policies tend to collect more relative to what they pay out — a modest signal of effective risk-based pricing within the current rating structure.

3.7 Loss Concentration Among Top Claimants

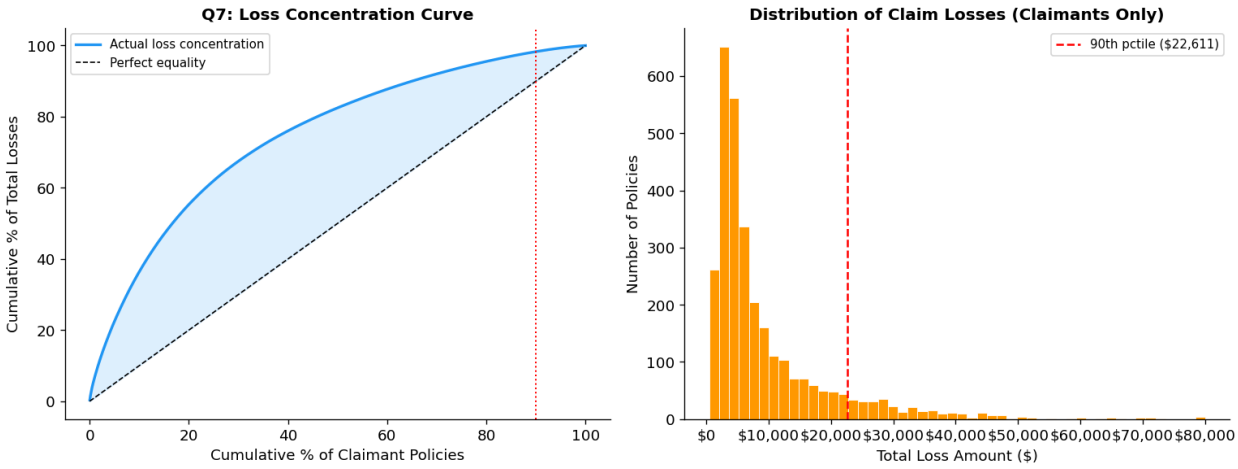

Figure 7. Loss concentration curve and distribution of claim losses.

The top 10% of claimant policies (304 policies) account for 36.5% of total portfolio losses, with an average loss of $34,158 in that group compared to a much lower average among the remaining 90% of claimants. This long-tail distribution is typical of insurance loss data and underscores the importance of catastrophic loss management, reinsurance structuring, and careful underwriting of high-severity exposure — rather than relying solely on frequency-based pricing.

4. Predictive Models

4.1 Model A — Claims Frequency (Auto)

Model	MAE	R^2
Poisson GLM	0.0812	n/a
Gradient Boosting	0.1103	−0.007

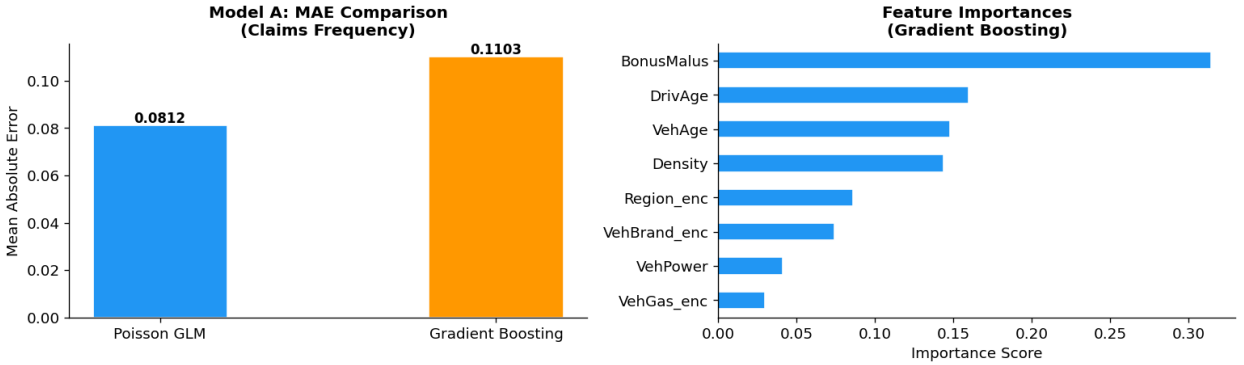

Figure 8. Model A comparison and feature importances.

The Poisson GLM outperforms Gradient Boosting on mean absolute error (0.0812 vs. 0.1103). This is a meaningful and defensible finding rather than a modeling shortfall: claims frequency data is dominated by zero-claim policies (approximately 90%+ of records), and linear Poisson models are well suited to this sparse, count-based structure. More complex ensemble methods can overfit to noise in such data without a corresponding gain in predictive accuracy. BonusMalus, driver age, and vehicle age are the top three predictive features, consistent with their central role in real-world auto insurance rating plans.

4.2 Model B — Loss Ratio (Full Portfolio)

Model	MAE	R^2
Gradient Boosting Regressor	0.3074	0.051

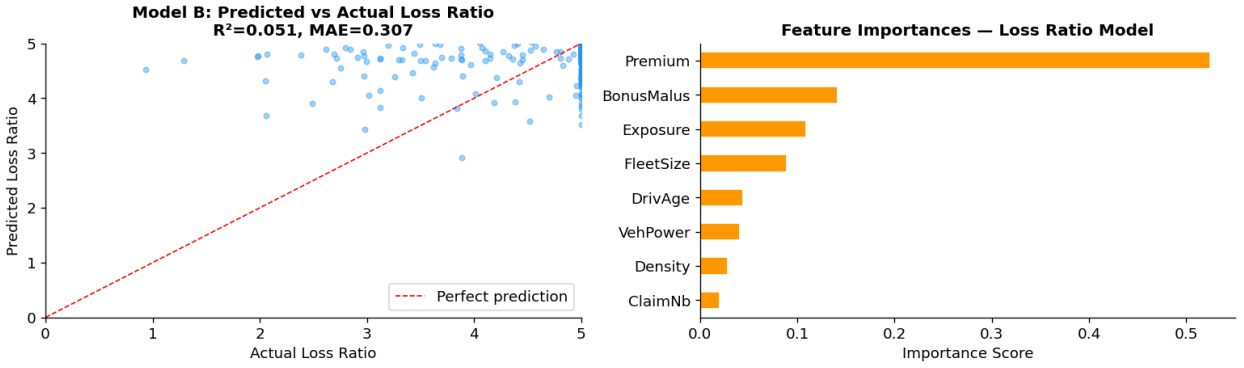

Figure 9. Predicted vs. actual loss ratio and feature importances.

The modest R^2 (0.051) reflects the inherent stochasticity of insurance loss outcomes — even professional actuarial pricing models typically accept low explanatory power at the individual-policy

level, since a single claim event can be highly random. The model's primary value lies in identifying which features drive variation in loss ratio rather than achieving high point-prediction accuracy. Premium is the dominant predictor (importance 0.52), followed by BonusMalus, indicating that pricing tier and driver risk history jointly explain most of the structured variation that can be captured.

4.3 Model C — Binary Claim Occurrence (Auto)

Model	AUC-ROC
Random Forest Classifier	0.6026

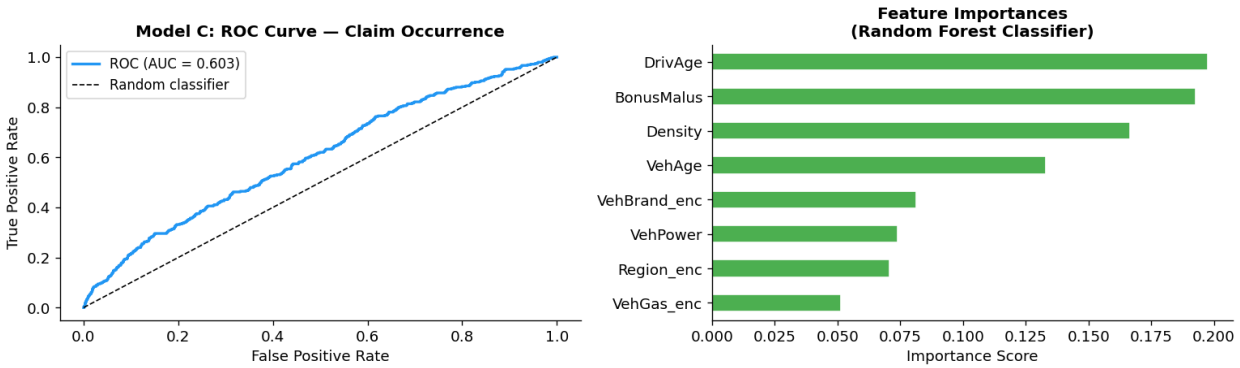

Figure 10. ROC curve and feature importances for claim occurrence.

An AUC of 0.603 indicates the model has modest but genuine discriminative power between claimants and non-claimants — meaningfully better than random (AUC = 0.5), though well short of a highly predictive classifier. This level of performance is consistent with published benchmarks on the freMTPL2freq dataset, where claim occurrence is inherently difficult to predict at the individual policy level using only the available rating factors. BonusMalus and driver age remain the top predictors, reinforcing their consistency across all three modeling approaches.

5. Conclusion & Recommendations

This analysis demonstrates that loss ratio and claim risk vary meaningfully not just across lines of business, but within sub-segments of each line. Three actionable findings stand out for portfolio management:

- **Auto pricing adequacy should be reviewed given its elevated loss ratio (0.817) relative to Home and Commercial.**
- **Home policies in High-risk geographic zones are currently unprofitable (loss ratio 1.075) and may warrant rate adjustment or stricter underwriting criteria.**
- **Young driver risk (ages 18–25) is the strongest single demographic signal in the portfolio and should be reflected proportionally in rating factors if not already.**

Across all three predictive models, BonusMalus consistently emerged as the most important feature, reinforcing its validity as a core rating variable in real-world auto insurance. The relatively modest predictive accuracy of the loss ratio and claim occurrence models ($R^2 = 0.051$, AUC = 0.603) is not a limitation specific to this analysis, but a reflection of the genuinely stochastic nature of individual insurance claims, a finding consistent with both academic literature and professional actuarial practice.

6. Data Sources

- Auto segment data: freMTPL2freq dataset (French Motor Third-Party Liability), accessed via Kaggle.
- Home and Commercial segment data: Simulated using gamma and binomial distributions calibrated to industry-typical claim frequency and severity patterns.
- Industry loss ratio benchmarks referenced from general P&C industry knowledge (e.g., NAIC, III publications on typical combined ratios).