Bob Gaglardi School of Business and Economics,

Thompson Rivers University

MBA Project

Rezoning ICBC Driver Performance to Reflect Localized Risk

A Machine Learning Algorithm

Sakshay Kapur

Supervisor: Dr. Mohammad Mahbobi

August 2025

Tables of Content

List of Tables

List of Figures

Abstract

The current insurance premium zoning system in British Columbia, managed by the Insurance Corporation of British Columbia (ICBC), is based on fixed geographic boundaries that often fail to reflect actual crash risk patterns. This static approach can group together regions with significantly different risk profiles, leading to cross-subsidization and misaligned premium structures. This study proposes a data-driven alternative that leverages detailed territory level crash data to improve both fairness and accuracy in premium setting. Using a comprehensive dataset encompassing crash severity, frequency, and contextual variables, four premium generation methods such as structural, bimodal, normal, and uniform were tested.
To uncover regional risk patterns, clustering algorithms (K-Means, DBSCAN, and Hierarchical Clustering) were applied and compared against ICBC's existing territorial zones to identify spatial inconsistencies. In parallel, supervised machine learning models (Random Forest, XGBoost, LightGBM, and CatBoost) were used to predict premiums and evaluated using standard performance metrics.
The results show that clustering reliably identifies territorial misalignments, and only the structurally generated data preserves realistic pricing relationships. These findings suggest that a machine learning driven rezoning framework can enhance actuarial precision, promote equity, and support more transparent premium allocation across British Columbia.

Introduction

British Columbia's road safety landscape is complex and varied. It has the urban side of BC with Metro Vancouver to the remote highways of the North. Setting fair and accurate auto insurance premiums is essential for maintaining public confidence, financial sustainability, and regulatory compliance. In British Columbia, ICBC currently uses a territory-based zoning system to adjust premiums by geographic area. These zones are defined by municipal and regional boundaries, but they may not always reflect actual driving risk as indicated by crash statistics. As urban development and traffic patterns evolve, these static zones risk becoming misaligned with real world crash data. When areas with differing risk profiles are grouped together, it can lead to cross-subsidization where drivers in lower risk regions pay more to offset higher risk areas. This can reduce the fairness of the pricing system and weaken incentives for road safety improvements.

To help ICBC better understand and price that risk, this study analyzed ICBC's 14 territories using publicly available data. The focus was on zone definitions, premium levels, and underlying factors such as accident counts, violation tickets, vehicle types and features, and regional characteristics. The process began with raw crash reports and concluded with mapped territory clusters and premium prediction models. The project had two main goals: first, to examine how premium levels

are currently determined and how they relate to performance indicators and second, to explore potential re-zoning strategies that could better align insurance pricing with actual crash risk.

An insurance portfolio groups policyholders by shared risk characteristics and sets premiums accordingly. If everyone paid the same rate, low-risk drivers would defect to cheaper alternatives, leaving the insurer with disproportionately high-risk customers. To avoid this, insurers assign risk factors and classify policyholders into tiers where each tier's premium reflects its aggregate risk (Henckaerts et al., 2018). This stratification underpins our premium classification approach.

This study looks at ways to improve ICBC's current zoning system by using a more data informed approach to insurance pricing. It uses a detailed crash dataset to simulate different premium outcomes. Four types of Estimated Premiums were calculated. Those were Normal, Structural, Bimodal, and Uniform. Each based on different assumptions about how crash risks are spread across regions. These premiums combined base rates, crash severity factors, territory level risk indicators and fixed surcharges, and were summarized by territory to show how they vary across the province.

The goal was to see if alternative methods could highlight regional differences that might be overlooked under the current zoning setup. To compare the results, the study used bar charts, density plots, and sensitivity tests, including minor changes to key inputs (±10%) to check how stable the premium estimates were. Five machine learning models such as Linear Regression, Random Forest, XGBoost, LightGBM, and CatBoost were tested to see how well territory level data could predict premiums. These models offer a way to measure performance and support the broader aim of creating a fairer and more accurate insurance pricing system. Since risk does not always follow municipal boundaries, the study also used clustering methods such as K-Means, DBSCAN, and Hierarchical Clustering to find areas that stand out from their surroundings. These techniques helped identify naturally occurring groups of higher or lower risk, which could be useful for thinking about future zoning changes.

Research Questions

- How do performance parameters such as crash type, number of drivers, and infractions relate to insurance premium zoning?
- Can machine learning algorithms be effectively used to analyze and potentially rezone insurance premiums based on risk characteristics?

The structure of this paper is as follows. It begins by examining the limitations of ICBC's current territory-based zoning system for auto insurance pricing. It touches on the motivation behind to explore whether a more data informed approach could lead to fairer and more accurate premium assessments across British Columbia. The next section presents the core dataset used in the study, a detailed collection of crash records enriched with regional indicators. Using this data, four versions of Estimated Premiums: Normal, Structural, Bimodal, and Uniform were generated. Each version reflects a different assumption about how crash risk is distributed and incorporates base rates, severity multipliers, territory level risk factors, and fixed surcharges. These premiums are then aggregated by territory to reveal spatial variation. Following premium simulations, the report evaluates whether these alternative methods expose regional differences that may be hidden under the current zoning system. To support this analysis, visual tools such as bar charts and kernel density plots are used, along with sensitivity testing that applies ±10% changes to key inputs to assess the robustness of the premium estimates. The report explores the predictive power of territory level features using five supervised machine learning models: Linear Regression, Random Forest, XGBoost, LightGBM, and CatBoost. This helps to estimate premiums and supports the broader goal of building a more regionally responsive and actuarially sound pricing system. In the final analytical section, the report then shifts focus to spatial clustering. Recognizing that crash risk may not align neatly with municipal boundaries, it applies three clustering algorithms such as K-Means, DBSCAN, and Hierarchical Clustering to identify areas that behave differently from their neighbors. These techniques help uncover organically emerging high and low risk clusters, offering insights into potential rezoning opportunities.

The report concludes by summarizing key findings and discussing how data driven insights could inform future policy decisions around insurance zoning and premium setting. It emphasizes the potential for improved fairness, transparency, and alignment between pricing and actual crash risk.

Literature Review

Early developments in insurance analytics have predominantly relied on statistical and time series methodologies to model claim behavior and inform premium structures. Generalized Linear Models (GLMs) have been widely adopted to relate claim frequency to driver characteristics such as age and driving experience, with the objective of balancing profitability and market competitiveness (David, 2015). Generalized Additive Models (GAMs) extend the capabilities of GLMs by incorporating non-linear smooth functions, thereby capturing more intricate risk patterns in pricing models. For analyzing temporal trends, Autoregressive Integrated Moving Average (ARIMA) models have been employed to examine claim volumes and seasonal fluctuations in historical data, supporting actuarial adjustments to premium rates over time. Artificial Neural Networks (ANNs) represent a more flexible modeling approach, capable of capturing high dimensional interactions among risk factors. These models have demonstrated improved accuracy in predicting pure premiums and have outperformed traditional techniques such as ARIMA in claim-amount forecasting tasks (Selvakumar et al., 2021).

Despite these methodological advancements, several critical gaps remain in literature. One major limitation is the insufficient integration of individual performance indicators such as traffic violations, accident counts, number of drivers per vehicle, and driver age into the geospatial zoning frameworks that underpin premium setting. Henckaerts et al. (2018) underscore the challenges of incorporating continuous and spatial risk factors into pricing models but do not explicitly address how personal and behavioral metrics can inform territory-based rate classifications. Similarly, Eriksen and Jones (1972) acknowledge the complexities of fair premium setting in regulated markets, yet their work does not extend to the spatial mapping of performance parameters. Furthermore, current research often overlooks the issue of bias and transparency in machine learning models. Although complex models like ANNs and hybrid ARIMA-ANN frameworks offer strong predictive capabilities, their lack of interpretability complicates insurers' ability to justify pricing decisions, raising concerns about fairness and regulatory compliance.

Another notable gap lies in the limited use of longitudinal data to assess evolving risk profiles over time. Most studies rely on cross-sectional datasets, which fail to capture the dynamic nature of

individual risk across policy lifecycles. Incorporating time series analysis into predictive frameworks could significantly enhance the accuracy of indemnity forecasts and premium adjustments. Additionally, while Selvakumar et al. (2021) explored a range of models including linear regression, exponential smoothing, ARIMA, and ANN, they did not investigate more recent deep learning architectures such as convolutional neural networks (CNNs), transformers, or ensemble methods like stacking, boosting, and bagging, which may offer further improvements in predictive performance. The geographic scope of existing research also remains narrow. Many studies, such as Selvakumar et al. (2021), concentrate on individual national markets (e.g., India) and do not incorporate cross-country comparisons that reflect varying regulatory and economic contexts.

Contemporary literature further builds on these foundations by applying machine learning techniques directly to pure premium modeling. Kumar et al. (2024) conducted a comparative evaluation of four algorithms of GLM, AGLM, XGBoost, and neural networks, using a French automotive liability dataset. Through 5-fold cross validation, they assessed model performance with metrics such as MAE, RMSE, and decile chart analysis. Their findings revealed that while traditional GLMs remained competitive, XGBoost and neural networks outperformed in predictive accuracy and segmentation of risk across quantiles. However, they caution that complex models like neural networks may sacrifice interpretability, which is crucial for regulatory transparency and operational deployment (Kumar et al., 2024).

Feature selection, a crucial step in model development, has also seen advancements through sensitivity-based methods. Saranya and Pravin (2021) introduced a variance-based sensitivity analysis (VSA) technique that identifies optimal feature subsets for disease classification tasks. Their approach demonstrated superior accuracy and sensitivity compared to wrapper-based selection methods. Although applied in the healthcare domain, this method offers transferable potential to insurance pricing, particularly in identifying the most impactful predictors in large scale datasets (Saranya & Pravin, 2021).

In this study, feature selection serves a practical function during the data preparation phase, following the approach outlined by Saranya and Pravin (2021). As crash statistics are aggregated

by territory, variables such as crash frequency, severity, percentage of fatal incidents, and average driver age are emphasized as primary predictors for risk profiling. Meanwhile, variables deemed less impactful or potentially redundant are filtered out through a combination of statistical analysis and domain expertise. This target selection helps reduce noise in the clustering and classification stages, leading to more efficient computation and enhancing the clarity and consistency of the resulting territorial risk groupings.

In parallel, the detection of fraudulent claims remains a priority for insurers seeking to reduce financial loss and administrative overhead. Kowshalya and Nandhini (2018) addressed this issue using synthetic datasets and implemented Naïve Bayes, J48, and Random Forest classifiers. Their results showed that Random Forest excelled in classifying fraudulent claims with an accuracy exceeding 99%, while Naïve Bayes performed best in estimating premium percentage classifications. Their work highlights the importance of preprocessing, attribute selection, and algorithm choice in building reliable fraud detection systems, which are foundational for ensuring fair pricing and operational integrity (Kowshalya & Nandhini, 2018).

Taken together, these studies underscore a growing shift toward data-driven, machine learning based insurance modeling. Yet, challenges related to data transparency, model interpretability, integration of behavioral metrics, and longitudinal risk tracking remain areas for future exploration. Bridging these gaps will be essential to advancing fair, explainable, and adaptive premium zoning frameworks.

Methodology

In response to the existing limitations in literature, the present study utilizes a substantially larger dataset comprising over one million anonymized ICBC policy and crash records, enabling more sophisticated feature engineering such as territory level risk profiles and severity multipliers and more rigorous model validation. Crash records are assigned to ICBC's 14 existing territorial zones, and clustering algorithms including K-Means, DBSCAN, and hierarchical clustering are applied to identify spatial patterns (Henckaerts et al., 2018). These clustering results are then evaluated using a suite of regression models to validate and refine zoning strategies. By combining large-scale spatial analysis with detailed performance indicators and longitudinal data, this

research advances the methodology for premium classification and territorial rezoning, offering a data driven framework that improves existing actuarial and statistical approaches.

Addressing these gaps through large-scale, bias aware, longitudinal ML models that marry transparency with advanced algorithms can advance the science of premium zoning and support more equitable, data-driven insurance pricing strategies.

Building upon the foundation laid by previous research, this project seeks to address notable gaps by integrating machine learning with geospatial analysis to refine insurance premium zoning. Although earlier studies like Kumar et al., (2024) have highlighted the predictive capabilities of models like GLMs, XGBoost, and neural networks for estimating pure premiums, few have explored the combined use of territorial clustering and spatial segmentation alongside individual level risk factors such as crash severity, etc. Additionally, similar to Saranya and Pravin (2021) approach, sensitivity analysis is an important tool for guiding feature selection and has seen limited application in spatial premium modeling. To bridge these gaps, this research introduces a data driven zoning framework that aggregates crash data across geographic regions, calculates frequency and severity metrics, and assigns risk categories using clustering and classification techniques. These clusters are then mapped to existing ICBC territorial boundaries, with premium multipliers applied based on publicly available base rates. The goal is to evaluate the potential of machine learning driven rezoning to produce fairer, more actuarially sound premium structures. The following section details the methodology used to prepare the data, construct risk profiles, perform clustering, and assess pricing outcomes across territorial zones.

The publicly available ICBC crash, and contravention datasets were sourced and merged into a single master table comprising 1,048,575 incident records. During this consolidation, each file underwent an iterative process of cleaning, addressing missing values, standardizing field formats, and removing duplicates before being combined. The resulting dataset includes attributes such as Crash Breakdown, Year, Month, Day of Week, Time Category, Crash Severity, Crash Type, Total Crashes, Total Victims, geographic descriptors (Municipality Name, Region, Territory, Municipality Boundary, Street Full Name, Cross Street Full Name, Latitude, Longitude), and binary flags for animal involvement, cyclist involvement, motorcycle involvement, heavy vehicle

involvement, pedestrian involvement, parked vehicle involvement, intersection crashes, and mid-block crashes.

To prepare the data for premium zoning analysis, each incident was assigned to its corresponding ICBC territory based on municipality and region identifiers. The crash contraventions were run against each crash territory by checking “Municipality with Boundary” then “Municipality” then “Region”. If all three are empty, they were labeled as “Unknown”. The data were then aggregated by territory to calculate total incident counts, as well as subtotals for fatal crashes, casualty crashes, and property damage only crashes. Additional aggregations captured the number of crashes involving pedestrians, cyclists, motorcycles, and those occurring at intersections.

For each territory, the proportions of fatal, casualty, and intersection crashes were computed relative to total crash volumes. These proportions were translated into categorical risk zones: Low, Medium, High, and Very High, using predefined thresholds for both crash frequency and crash severity (fatality rate). Finally, by combining the frequency and severity classifications, an overall risk rating (ranging from Low Risk to Very High Risk) was assigned to each territory, providing a structured framework for subsequent premium zoning applications. In simple words, it calculates crash severity and crash frequency which helps to define overall risk category.

Once Risk profiles were created, they need to be linked to territory which reloads the full crash dataset, repeats the territory assignment, and then groups by territory to count crashes, fatalities, pedestrian and cyclist involvement. Then it converts those counts into percentage buckets (e.g. under 2% fatal = “Low Severity,” etc.) based on preset thresholds and using a lookup table that combines frequency and severity buckets, it assigns each territory a final “Risk Category.” And if the threshold choices are off, many territories may land in unexpected categories, so a few manual random checks were done before proceeding ahead.

The consolidated crash dataset was now annotated with risk profiles which were first classified by ICBC territory code to enable territory specific premium calculations. The Base Premium for each record was computed as the product of a Severity Factor and a Risk Factor. Subsequent variables required by ICBC’s Basic Insurance Tariff were then derived (ICBC, 2020, p. 4) as:

$$Premium = [BRP * \prod_{i=1}^{6} F_i] + LP + UDPP + UDAP \quad (1)$$

where:

BRP = Base Rate Premium

F_1 to F_6 are defined as:

- CDF = Combined Driver Factor.
- DDF = Disability Discount Factor.
- HVVCF = High Value Vehicle Charge Factor.
- ASTF = Advanced Safety Technology Factor.
- DF = Distance Factor.
- TF = Transition Factor; and

LP = Learner Premium.

UDPP = Unlisted Driver Protection Premium.

UDAP = Unlisted Driver Accident Premium

To maintain the confidentiality of ICBC's actual pricing data, in this study four data generation methods Structural, Normal, Bimodal, and Uniform were used to simulate premium values. Since real premium data includes sensitive and proprietary information, these datasets were designed to reflect the general structure and behavior of pricing patterns without disclosing any private or regulated details.

This allowed for meaningful analysis using clustering and regression techniques, while respecting privacy and confidentiality guidelines. By comparing these variants, the study aimed to isolate whether the observed patterns are driven by real structural signals in the data or are merely artifacts of randomness or noise, ensuring that findings are not overly dependent on a specific statistical distribution. Within each method, a few assumptions were made to keep things consistent. The Heavy Vehicle Crash Factor (HVVCF) was set using a simple flag, heavy vehicles got a multiplier of 2, while non-heavy ones stayed at 1. Since there wasn't enough data for Liability Penalty (LP) and Unlisted Driver Premium Penalty (UDPP), both were set to zero. Other adjustment factors like Driver Disqualification Factor (DDF), At Fault Severity Tuning Factor (ASTF), Driver Factor (DF), and Territory Factor (TF) were all defaulted to 1, meaning no extra tweaks were applied.

The premium calculation followed a clear step-by-step process. First, the Base Premium was calculated by multiplying the base rate, severity factor, and territory risk factor. Then came the Crash Damage Factor (CDF), which varied depending on the crash type head on collisions which

could bump it up to multiple of 1.5, while smaller incidents had lower values. The Unlisted Driver Attribution Penalty (UDAP) was figured out by taking the probability that an unlisted driver caused the crash and multiplying it by 1,000.

Equation (1) pulls together crash severity, territory risk, and vehicle-related details to give a well-rounded premium estimate, which then feeds into the machine learning models for zoning analysis. To make the premium calculations work, the crash data was combined with a risk profile file created earlier and matched with ICBC territory codes, which were simplified into single-letter labels. For each code, the system looked up a base rate, applied a crash severity multiplier, and added a territory risk multiplier. Fixed surcharges like those for heavy vehicles, at fault drivers, and crash types were added in, and the final premium values were written out. These completed estimates were then ready for use in the next phase of analysis.

Data

The dataset was compiled from publicly available ICBC crash records across British Columbia, covering a five-year span from 2019 to 2023. The 1,048,575-incident records data include reported vehicle collisions aggregated at the territory level, aligned with ICBC's existing premium zoning regions. Each entry contains information such as crash severity, intersection involvement, and other territory level factors relevant to contravention. The dataset encompasses a wide range of driving environments, from urban areas like Vancouver, Surrey, Burnaby, Kelowna, and Victoria to more rural or remote regions such as Peace River North, Cariboo, and the Kootenays. Crash incidents were grouped to help identify temporal patterns, account for seasonal variation, and improve the reliability of model estimates. This territorial crash summary serves as the foundation for the estimated premium calculations, clustering, and spatial analysis carried out in the study.

To enable thorough analysis while maintaining the confidentiality of ICBC's actual pricing data, the study uses four premium generation methods such as Uniform, Bimodal, Normal, and Structural. Each method was designed to reflect a different statistical pattern. Uniform and Bimodal distributions provide baseline scenarios to test model behavior under simplified or randomized conditions. The Normal variant introduces a more realistic distribution, capturing typical variation in pricing without replicating proprietary structures. The Structural method, on

the other hand, incorporates crash severity, frequency, and territory level risk factors to more closely resemble how premiums might be calculated in practice. By comparing model and clustering results across all four approaches, the study assesses which methods best capture meaningful pricing relationships and whether the findings hold up under different data assumptions.

Uniform Data Generation

Using a uniform random selection approach for data generation means that every possible value within a defined range has an equal chance of being selected unless specific weights are manually applied. This method ensures a diverse spread of data points, but it doesn't necessarily mirror real world distributions. Instead of relying on complex statistical shapes like a normal curve or patterns drawn from actual datasets, values for each factor are simply chosen at random within a set range. For instance, the Disability Discount Factor (DDF) might fall anywhere between 0.8 and 1.2, while the Safety Tech Discount Factor (ASTF) ranges from 0.9 to 1.1. The Distance Factor (DF) is allowed to vary from 0.5 to 1.5, and the Transition Factor (TF) from 0.9 to 1.1. For the High Value Vehicle Crash Factor (HVVCF), heavy vehicles are assigned a value between 1.8 and 2.2, while non heavy vehicles fall between 0.8 and 1.2. The Crash Damage Factor (CDF) is adjusted by randomly shifting the original value up or down by as much as 10%, and the Unlisted Driver Attribution Penalty (UDAP) follows the same logic, but with a wider margin of up to 20%.

This kind of uniform data generation offers several advantages. First, it provides clarity and control, avoiding any hidden patterns that might be embedded in historical data. Second, the ranges are realistic wide enough to reflect meaningful changes, such as a discount increasing or decreasing by 20%. Third, the process is repeatable, i.e., by using the same random seed, the exact same set of values can be regenerated for consistency. Finally, it allows for fair comparisons. Since all four data generation methods uniform, bimodal, normal, and structural follow the same setup but differ in how they select values, it's possible to compare their outcomes side by side and understand how the choice of distribution influences premium predictions.

Structural data generation

Structural equation models reflect a causal relationship, not just a statistical correlation. These models are often used in settings where controlled experiments aren't possible, and they typically involve simultaneous relationships and/or measurement errors. These errors might come from basic inaccuracies in how something is measured, or from the fact that what we can observe doesn't exactly match the theoretical concepts we care about (Goldberger, 1972).

Importantly, the structural parameters in these models are not the same as the coefficients you may get from simply running regressions on observed data though the model still enforces certain limits on those regression coefficients. This creates subtle problems around identification (figuring out the true relationships), which requires sophisticated statistical techniques to handle properly (Goldberger, 1972).

Every factor in the pricing model is derived directly from specific fields in the dataset, with no randomness involved. This approach ensures the system remains fully explainable and produces consistent results every time it runs. The Disability Discount Factor (DDF) is fixed at 1.0, meaning disability-related discounts are not considered in this case. The At Fault Safety Tech Factor (ASTF) is calculated from a Tech Score ranging from 0 to 1. This score is converted into a discount using the formula: 1 minus 0.1 times the Tech Score, and then clamped between 0.9 and 1.0, so vehicles with more advanced technology receive a slightly better discount. The Distance Factor (DF) increases by 5% for each additional crash victim, with a maximum cap of 3.0, reflecting the idea that more victims typically lead to higher costs. The Transition Factor (TF) accounts for the year of the crash, adding 0.5% to the rate for every year after 2018 to capture gradual pricing drift over time. For vehicle classification, the High Value Vehicle Crash Factor (HVVCF) is set at 2.0 for heavy vehicles and 1.0 for all others, effectively doubling the cost for heavy vehicles. Crash severity influences the Combined Driver Factor (CDF), which is set at 1.0 for low severity, 1.1 for medium, and 1.2 for high severity crashes. The Unlisted Driver Attribution Penalty (UDAP) is calculated by multiplying the probability of the driver being at fault by $1,000, representing the added risk. The Unlisted Driver Premium Penalty (UDPP) is a flat $50 for crashes involving injury, and $0 for property only damage. Lastly, the Learner Penalty (LP) is a fixed $100 charge if the driver is a learner, and nothing if they are not. This rule-based structure makes the model transparent and easy to audit, with each factor clearly tied to a specific input. Thus, it helps to

achieve clarity, repeatability, and real-world logic. Any other factor such as victim count or year can also be pointed out and explained exactly why a premium changed. Appendix A.1 provides similar formulation.

Bimodal data generation

Instead of just picking numbers randomly, this method assumes there are two types of people or situations for each factor like careful drivers vs. riskier ones. So, two groups (called a "bimodal" mix) were mixed with different averages and spreads, which gives the data more shape and makes it feel more real. Each factor reflects patterns observed in the data, offering a more realistic and nuanced view of driver and vehicle behavior. The Disability Discount Factor (DDF) shows that about 70% of drivers are generally safer, with scores around 0.90, while the remaining 30% are riskier, closer to 1.10. This split suggests that most people drive cautiously, though not everyone fits that mold. The At Fault Safety Tech Factor (ASTF) captures a mix of vehicles, some equipped with advanced safety technology that earns them a slight discount, and others without those features, receiving no bonus. The Distance Factor (DF) is evenly split, with half of the crashes occurring in urban areas where trips are shorter, and the other half in rural regions where trips tend to be longer.

Crash year adjustments, represented by the Transition Factor (TF), are mostly neutral, but a few cases receive a small bump to account for pricing drift over time. For the High Value Vehicle Crash Factor (HVVCF), both heavy and non-heavy vehicles are distributed across slightly higher and lower cost ranges, reflecting variability in how vehicle type influences premiums. The Crash Damage Factor (CDF) follows expected patterns for about 70% of cases, based on crash type, while the remaining 30% are flagged as riskier, suggesting more severe or unusual incidents. Finally, the Unlisted Driver Attribution Penalty (UDAP) shows that most drivers incur a standard surcharge, but around 40% face steeper penalties, likely due to higher probabilities of being at fault. Together, these distributions paint a more human picture of risk, one that's not perfectly uniform but grounded in realistic variation. Because real people aren't all the same. There's usually one main group and a smaller different group. This way of generating numbers captures that. It's more realistic than just random guessing, but still more flexible than using set formulas. It helps models

learn patterns while still reflecting real-world variety and formulae can be referred to Appendix A.2.

Normal Distribution

This modeling approach relies on the principle that most values tend to cluster around a normal baseline, with only modest variation. Each factor is centered around 1.0 or its own base value and allowed to fluctuate slightly, following a bell-shaped curve characteristic of the normal distribution. This structure ensures consistency while still capturing realistic differences. For example, the Disability Discount Factor (DDF) hovers around 1.0, with shifts of about ±10% to reflect individual driver differences enough to introduce variation without destabilizing the model. The Safety Tech Discount Factor (ASTF) varies even less, typically within ±5%, since most vehicles today share similar safety technologies. In contrast, the Distance Factor (DF) shows greater variability, up to ±20%, acknowledging that driving habits, especially trip lengths can differ widely across individuals.

The Transition Factor (TF) introduces a gentle drift of around ±5%, simulating the passage of time and adding a dynamic element to the model across crash years. For vehicle classification, the High-Value Vehicle Crash Factor (HVVCF) starts at 1.0 for standard vehicles and 2.0 for heavy ones, with a touch of noise added to prevent overly rigid categorization. Crash severity is captured through the Crash Damage Factor (CDF), which is anchored to a baseline tied to crash type and allowed to vary by ±10% to reflect the inherent unpredictability of real-world incidents. Finally, the At-Fault Penalty (UDAP) begins from a fixed base and fluctuates by about 10%, offering a flexible way to represent varying degrees of responsibility in crash scenarios. These controlled variations help the model remain grounded while still allowing for meaningful distinctions across cases.

This approach is not only intuitive but also computationally efficient when the design matrix is well-balanced. As Healy and Westmacott (1956) noted, exact least squares computations are straightforward for special design matrices that incorporate balance and orthogonality. However, when data is missing or the design is unbalanced, computations become more complex, often

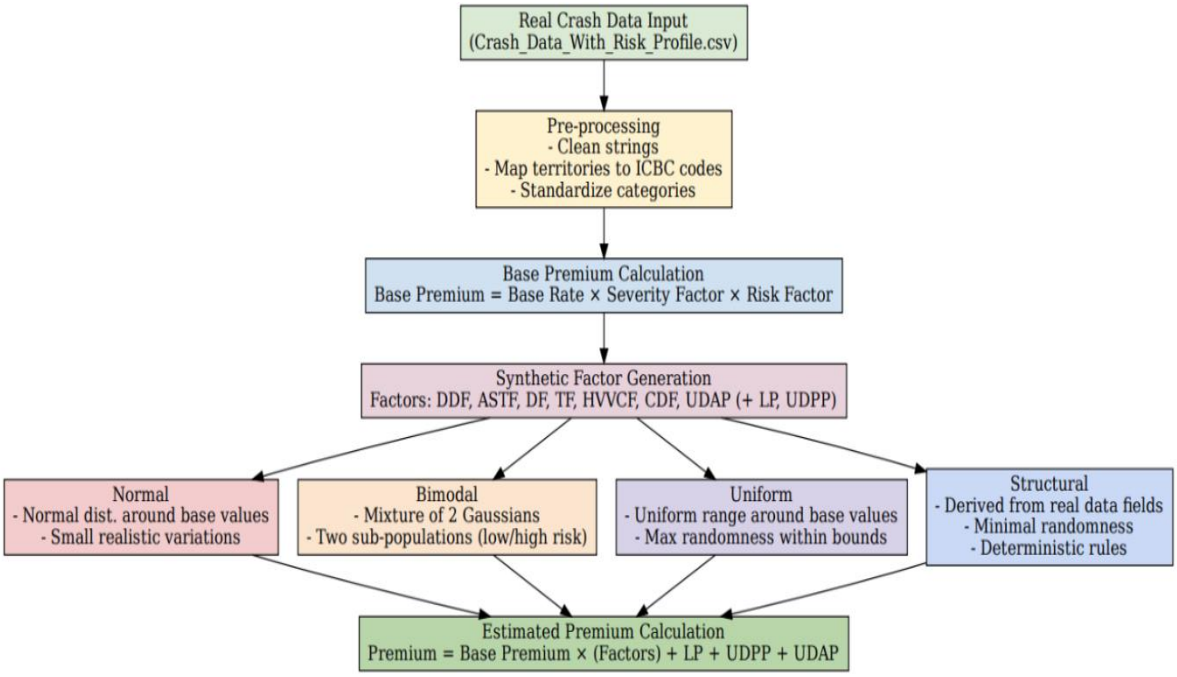
requiring the inversion of large matrices. In such cases, it is natural to fill in missing values with their expected values based on current parameters, then re estimate using a simple least-squares algorithm, iterating until the estimates stabilize. More broadly, it may even be possible to add hypothetical rows to the design matrix rows that never existed in the real world in a way that facilitates least-squares analysis.

This idea aligns well with the model's use of normal distributions to smooth out irregularities and maintain analytical tractability. It makes sense if you imagine each number being shaped by countless subtle influences like driving habits, location, and vehicle type while avoiding extreme outliers by keeping everything within sensible bounds and formulae can be referred to Appendix A.3.

Figure-1 illustrates the premium data generation process. Due to the confidentiality issue and lack of access to the ICBC premium data, real crash data are preprocessed and mapped to ICBC territories to estimate base premiums. Factor values are then generated using four distinct distributions as Normal, Bimodal, Uniform, and Structural and combined with fixed loadings to derive the estimated premiums.

Factor values such as the Disability Discount Factor (DDF), Advanced Safety Tech Factor (ASTF), Distance Factor (DF), Transition Factor (TF), High-Value Vehicle Crash Factor (HVVCF), Crash Damage Factor (CDF), and At Fault Penalty (UDAP) are generated based on the selected data generation methods. The estimated premiums are then computed by applying the generated factors to the base premium and adding fixed loadings for the Learner Premium (LP) and Unlisted Driver Protection Premium (UDPP). The result is a set of estimated premium datasets.

Figure-1: ICBC Zoning Premium Data Generation Process



Source: Authors

Table-1 presents statistical characteristics of four premium distributions of Bimodal, Normal, Structural, and Uniform. The Average Premium column shows the original mean values, while the Mean (μ), Standard Deviation (σ), Skewness, and Kurtosis offer deeper insight into each distribution's structure. Mean values are consistent with earlier averages. Normal and Uniform distributions display the greatest variability (σ = 667 and 660), while the Structural distribution is more tightly clustered (σ = 380), reflecting its design to generate realistic pricing. All distributions show positive skewness, indicating a tendency toward higher premium outliers, with the Normal distribution being the most skewed (1.33). Kurtosis values reveal that the Normal distribution has the sharpest peak and heaviest tails (3.79), whereas the Structural distribution is the flattest (0.22), making it most stable. These metrics collectively illustrate the unique behavior of each dataset.

Table-1: Descriptive Statistics of Estimated Premiums by Alternative Distribution

Distribution	Average Premium ($)	Mean (μ)	Standard Deviation(σ)	Skewness	Kurtosis
Bimodal	2,134.28	2128.43	615.52	0.73	0.72
Normal	2,128.15	2127.44	667.36	1.33	3.79
Structural	1,888.56	1886.55	380.59	0.74	0.22
Uniform	1,975.75	1969.88	660.08	0.79	0.76

Source: Authors' calculations

Kernel Density Estimation (KDE)

Kernel Density Estimation (KDE) is a widely used nonparametric technique for estimating the probability density function (PDF) of a continuous variable based on a finite sample. Unlike histogram-based methods, KDE provides a smooth and continuous density curve by assigning a kernel function to each data point and aggregating their contributions across the domain. This approach avoids the abrupt bin edges of histograms and offers greater flexibility in revealing the underlying distributional structure of the data. At its core, KDE evaluates the density at a point by summing the influence of all sample observations x_i, each weighted by a kernel function centered at that observation. The general form is given by equation (1):

$$f(x) = (1/n)\, \Sigma_{i=1}^{n}\, Kernel\,(x, x_i) \qquad (1)$$

Here, n is the number of data points, and the kernel function defines how each data point contributes to the estimate at location x. The kernel function must satisfy the condition:

$0 \leq \text{Kernel}\,(x, x_i) < \infty$, and all values of x and x_i must be within a bounded range.

In practice, symmetric kernel functions are most used. A symmetric kernel function can be written as equation (2):

$$Kernel\,(x, x_i) = (1/h) \cdot K\,((x - x_i)/h) \qquad (2)$$

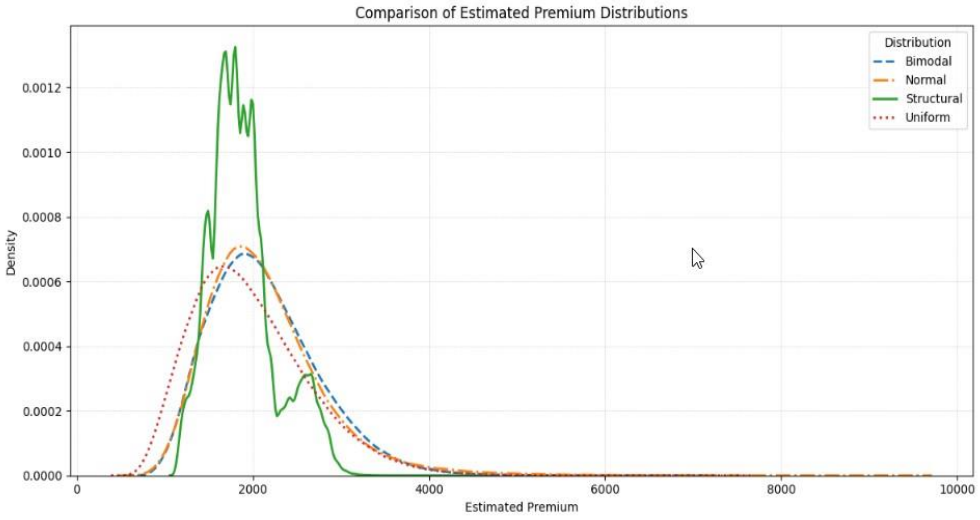
where K is the kernel shape (e.g., Gaussian, uniform, triangular), and h is the bandwidth (or smoothing parameter). Substituting this into the original formula gives the standard symmetric KDE form as equation (3):

$$f(x) = (1 / (n \cdot h)) \Sigma_{i=1}^{n} K ((x - x_i)/h) \quad (3)$$

The bandwidth h plays a critical role in shaping the final estimate. A smaller bandwidth results in a highly detailed curve with sharp local variations, while a larger bandwidth produces a smoother and more generalized density function that may obscure important local features. Bandwidth selection is therefore crucial and can be determined through methods such as cross-validation or rule of thumb estimators like Silverman's rule (Węglarczyk, 2018).

Overall, KDE offers a powerful, flexible approach to analyzing the shape of empirical data distributions, making it particularly useful in domains such as insurance, where underlying risk patterns may not follow standard parametric forms. KDE was plotted to help visualize how premium values are distributed. Each distribution is shown separately, making it easy to compare their shapes. Figure-2 shows four smooth distributions, each representing a different premium distribution, helping to understand how premiums vary under different scenarios.

Figure-2: Comparison of Estimated Premium Distributions



Source: Authors' calculations

Similarly, all four distribution datasets again were leveraged, and each row was tagged with its distribution name (Bimodal, etc.) and to find the one Estimated Premium column in each dataset and copied its values into a common Estimated Premium column to merge all four into one Data Frame. Each premium can be adjusted by increasing the crash-type factor (CDF) by 10% to assess how sensitive the premium estimates are to changes in that parameter as in Figure-2. This adjustment is applied using the formula (CDF*1.10)/CDF) which effectively scales the premium based on the increased severity factor. It is important to note that this formula assumes the CDF value is never zero. If CDF were zero, the calculation would result in a division by zero error or produce undefined values (NaNs), which could disrupt the analysis. Therefore, ensuring that CDF is always positive and properly defined is essential for maintaining the integrity of the sensitivity testing. The data is grouped by distribution type, which helps compare different premium distributions.

Premium Values

The key statistics for the Final Premium values have been calculated to provide a comprehensive overview of data distribution. These include the count, which represents the total number of records, the mean which indicates the average premium and the standard deviation which reflects how much the premium values deviate from the average.

Table- 2 below further illustrates the spread of the data by presenting the minimum and maximum premium values, along with the 25th, 50th (median), and 75th percentiles. These percentile values offer insight into how the premiums are distributed across the dataset.

Table-2: Premium Descriptive Statistics by Types of Datasets

Distribution	Count	Mean	Std	Min	25%	50%	75%	Max
Bimodal	710,210	2,134.27	624.92	667.72	1,676.59	2,055.63	2,507.66	7,064.41
Normal	731,342	2,128.15	667.28	611.04	1,664.42	2,028.34	2,470.99	9,544.02
Structural	804,741	1,888.55	384.60	1126.39	1,628.33	1,820.99	2,060.41	4,604.17
Uniform	709,817	1,975.75	669.59	532.70	1,481.90	1,879.88	2,372.43	7,530.30

Source: Authors' calculations

After generating four datasets using bimodal, normal, uniform, and structural approaches, the four labeled sets were merged into a single comprehensive dataset.

Machine Learning Models

Machine learning models were subsequently trained and evaluated using the combined dataset, with their performance assessed through Root Mean Squared Error (RMSE), MAE, R^2 and 90% QAE metrics. To explore a range of predictive capabilities, several algorithms were implemented, including Linear Regression, Random Forest, XGBoost, LightGBM, and Cat Boost. These models were selected to provide a diverse comparison between traditional and ensemble-based approaches. Among popular machine learning algorithms, Random Forest stands out for its ensemble approach, where multiple decision trees are built using bootstrapped samples and predictions are made by averaging their outputs. Each tree considers a random subset of features at each split, and growth continues until a predefined stopping criterion is met, such as a limit on splits or node size (Zimmerman et al., 2018). Building on the idea of iterative improvement, XGBoost takes a different route by constructing boosted trees sequentially, with each new tree learning from the errors of the previous ones. Its use of regularization and focus on scalability make it a widely adopted choice for large scale tasks (Karakilic, Hatas, & Pacal, 2025). In a similar approach, LightGBM offers its own innovations, introduced by Microsoft in 2017, including techniques like gradient-based one-sided sampling and exclusive feature bundling, which help speed up training while preserving accuracy (Ke et al., 2017). Lastly, Cat Boost brings unique strengths to the table, especially in handling categorical variables and high-dimensional data. Its consistent performance across diverse applications from precipitation forecasting to biomass estimation has made it a reliable option in many domains (Luo et al., 2021; Qian et al., 2021; Samat et al., 2022).

Finally, the compiled dataset was subjected to each model, and results were compared based on RMSE, MAE, R^2 and 90% QAE to identify the most accurate predictive approach. Pandas and numpy were imported for core data wrangling and matplotlib for visualization, and key components from scikit-learn for splitting the dataset and evaluating performance. For modeling, a solid lineup was used whereas, linear regression, random forest, Boost, Light, and Cat Boost were needed to be installed in the base environment. The merged dataset was loaded. The goal was to work exclusively with complete records to maintain modeling integrity as over filtering for completeness can introduce bias or shrink the sample size, which may undermine statistical power.

For the feature set, things were kept clean by grabbing all numeric columns, then excluding the premium columns from that list. What's left forms the input features. Because it is easy to accidentally include numeric IDs or indices, which could mislead the model if not handled properly. Therefore, the five models, each one mapped to a short name, each with a fixed random seed for reproducibility, were instantiated. And default hyperparameters were used so it runs quickly.

The following steps will run a consistent workflow using the same seed for fairness:
- Split the data 80/20 into training and testing sets.
- Fit each model on the training portion.
- Predict on the testing set.
- Calculate RMSE, MAE, R^2 and 90% QAE (Quantile Absolute Error at the 90th Percentile.
- Store the results (model name, target column, RMSE, MAE, R^2 and 90% QAE)

Results were shaped into a Data Frame, then pivoted it for cleaner comparison. As shown in Table -3,4,5 and 6, rows as models, columns as different target generation methods, and values.

Table-3,4,5 and 6 describing the Model Performance Metrics by Premium Type presents a comparison of five machine learning models of CatBoost, LightGBM, Linear Regression, Random Forest, and XGBoost evaluated across four premium generation methods such as Bimodal, Normal, Structural, and Uniform. The models were assessed using the four key metrics:
Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (4)$$

Where:

$\hat{y}_i$ = predicted value

y_i = actual value

For a sample of n observations y (y_i, i = 1, 2, . . ., n) and n corresponding model predictions y^.

The RMSE as shown in equation (4) has been used as a standard statistical metric to measure model performance in many areas including meteorology, air quality, and climate research studies (Hodson, 2022). The second accuracy measurement is Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \qquad (5)$$

MAE is a useful measure widely used in model evaluation. For a sample of m observations y (y_i, i = 1, 2, . . ., n) and n corresponding model predictions $\hat{y}_i$ (Hodson, 2022). In this study, Coefficient of Determination (R^2) also used to measure the goodness of it of the estimated models:

$$R^2 = 1 - \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n}(y_i - \bar{y}_i)^2} \qquad (6)$$

where $\bar{y}_i$ is the mean of the actual values & R^2 represents the proportion of variance in the dependent variable explained by the model. This metric is commonly employed in evaluating regression models and is noted for offering greater insight than many alternative measures (Chicco et al., 2021). The next accuracy measurement is the 90th Percentile Absolute Error (90% QAE):

$$\text{QAE}_{90} = Q_{0.90}(|\hat{y}_i - y_i|) \qquad (7)$$

where $Q_{0.90}$ represents the 90th percentile of absolute errors. This metric is a quantile-based dispersion measure, conceptually similar to the Quantile Absolute Deviation (QAD) framework (Akinshin, 2022).

While the machine learning models in this study produce continuous outputs in the form of predicted pure premiums, classification-based metrics were also employed to evaluate their effectiveness in identifying high-risk territories. This dual evaluation framework allows for a comprehensive assessment of both numerical prediction accuracy using RMSE, MAE, and R^2

and categorical classification accuracy, which is critical for premium zoning and decision-making.

To facilitate classification analysis, continuous predictions were converted into discrete risk categories for Multi-class Classification, for AUC, Accuracy, Precision, Recall and F_1 Score. Territories were divided into Low, Medium, and High-risk groups by segmenting the actual premium distribution into equal quantile ranges. Predicted premiums were discretized using the same quantile-based boundaries to ensure consistent categorization.

The Receiver Operating Characteristic (ROC) curve is a a way to visualizes the performance of a binary classifier at various threshold settings. ROC plots the True Positive Rate also known as Sensitivity or Recall against the False Positive Rate.
The AUC is the area under the ROC curve. It provides a single numerical value that indicates the overall ability to distinguish between different clusters. For instance,
While AUC = 1 indicates a perfect classifier, AUC = 0.5 shows that the classifier performs no better than random guessing. Other performance accuracy measures were also calculated. The most basic metric represents the proportion of correct predictions out of all predictions. It's calculated as (True Positives + True Negatives) / (Total Predictions).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.
Measures the accuracy of positive predictions. It indicates the proportion of correctly predicted positive instances out of all instances predicted as positive. Calculated as True Positives / (True Positives + False Positives).
as in equation (10) below:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

Sokolova and Lapalme (2009) describe Recall as the average ability of a classifier to correctly identify class labels, measured separately for each class. In simple words, Recall measures the ability of the model to identify all relevant positive instances. It indicates the proportion of correctly predicted positive instances out of all actual positive instances. Calculated as True Positives / (True Positives + False Negatives).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

The last accuracy measurement is F1-score as in equation (12) below:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

As Sharma (2023) describes F_1 score as the harmonic means of precision and recall, providing a balanced measure that considers both false positives and false negatives. It's particularly useful for imbalanced datasets where high precision or recall alone might be misleading.

Together, these metrics offer a balanced view of prediction accuracy, model fit, and sensitivity to larger errors.

Among the four premium types, the Structural dataset consistently leads to the strongest model performance. It yields lower RMSE and MAE values and significantly higher R^2 scores across all models, indicating a clearer and more learnable structure. CatBoost performs best under this dataset, achieving an RMSE of 158.31 CAD, an MAE of 131.38 CAD, and an R^2 of 0.85, suggesting it explains 85% of the variance in premium predictions. Comparable results from XGBoost and LightGBM further reinforce the effectiveness of structure preserving data.

On the other hand, models trained on Bimodal and Normal distributions show weaker performance, with higher error rates and R^2 values typically ranging from 0.20 to 0.24. These results imply that these distributions lack the underlying structure needed for accurate prediction. While Random Forest performs reasonably well on Structural data, it tends to lag slightly behind the gradient boosting models overall.

The 90% QAE metric adds another layer of insight by highlighting worst case prediction errors.

For example, under the Structural dataset, 90% of Cat Boost's absolute errors fall below 248.54 CAD, whereas in the Uniform dataset, this threshold increases sharply to 987.81 CAD, indicating greater volatility and less reliability.

The results suggest that Structural data offers meaningful patterns that all models, especially tree-based ones like CatBoost and XGBoost can effectively learn. Gradient boosting models consistently outperform Linear Regression and show greater robustness across different data types. Additionally, the 90% QAE metric proves valuable in assessing prediction reliability in higher risk scenarios, which is particularly relevant for insurance pricing applications.

Table-3: Model Performance Metrics by Premium Type- RMSE

Model	Bimodal	Normal	Structural	Uniform
CatBoost	552.08	625.02	158.31	620.76
LightGBM	551.86	625.27	159.64	620.82
Linear Regression	573.27	644.85	233.07	637.18
Random Forest	598.71	672.48	168.28	672.94
XGBoost	552.61	625.39	158.46	621.68

Source: Authors' calculations

Table-4: Model Performance Metrics by Premium Type, MAE

Model	Bimodal	Normal	Structural	Uniform
CatBoost	438.75	456.32	131.38	496.23
LightGBM	438.70	456.31	132.72	496.38
Linear Regression	453.86	471.80	184.66	507.58
Random Forest	472.18	495.79	137.39	533.58
XGBoost	439.04	456.72	131.44	496.90

Source: Authors' calculations

Table-5: Model Performance Metrics by Premium Type, R^2

Model	Bimodal	Normal	Structural	Uniform
CatBoost	0.24	0.20	0.85	0.18
LightGBM	0.24	0.20	0.84	0.18
Linear Regression	0.18	0.15	0.67	0.14
Random Forest	0.11	0.08	0.83	0.04
XGBoost	0.24	0.20	0.85	0.18

Note. Values represent accuracy scores across different data distributions.

Source: Authors' calculations

Table-6: Model Performance Metrics by Premium Type, 90% QAE

Model	Bimodal	Normal	Structural	Uniform
CatBoost	883.10	916.33	248.54	987.81
LightGBM	881.92	914.72	249.30	988.63
Linear Regression	914.82	945.50	358.36	1010.06
Random Forest	963.86	1021.55	272.26	1083.75
XGBoost	884.00	917.75	249.06	990.07

Note. Values represent accuracy scores across different data distributions.

Source: Authors' calculations

Table-7*:* Model Performance Metrics by Premium Type, AUC

Model	Bimodal	Normal	Structural	Uniform
CatBoost	0.7229	0.7352	0.9200	0.6859
LightGBM	0.7237	0.7356	0.9211	0.6859
Linear Regression	0.7067	0.7151	0.8761	0.6732
Random Forest	0.6877	0.6962	0.9115	0.6483
XGBoost	0.7214	0.7329	0.9216	0.6821

Source: Authors' calculations

Table-8: Model Performance Metrics by Premium Type, Accuracy

Model	Bimodal	Normal	Structural	Uniform
CatBoost	0.654	06579	0.8171	0.6256
LightGBM	0.6553	0.659	0.8177	0.6246
Linear Regression	0.6413	0.6388	0.7862	0.619
Random Forest	0.6304	0.6373	0.8169	0.6026
XGBoost	0.6525	0.656	0.8213	0.6217

Source: Authors' calculations

Table-9: Model Performance Metrics by Premium Type, Precision

Model	Bimodal	Normal	Structural	Uniform
CatBoost	0.6408	0.6301	0.951	0.6103
LightGBM	0.6439	0.631	0.949	0.6086
Linear Regression	0.6294	0.6117	0.7465	0.6036
Random Forest	0.6207	0.6194	0.8874	0.5902
XGBoost	0.6403	0.6285	0.9336	0.6025

Source: Authors' calculations

Table-10: Model Performance Metrics by Premium Type, Recall

Model	Bimodal	Normal	Structural	Uniform
CatBoost	0.6986	0.7636	0.6678	0.6873
LightGBM	0.6927	0.7648	0.6708	0.6905
Linear Regression	0.6849	0.7586	0.8645	0.6848
Random Forest	0.6677	0.7111	0.7251	0.6619
XGBoost	0.6934	0.7619	0.6911	0.7074

Source: Authors' calculations

Table-11: Model Performance Metrics by Premium Type, F1 Score

Model	Bimodal	Normal	Structural	Uniform
CatBoost	0.6684	0.6905	0.7847	0.6465
LightGBM	0.6675	0.6915	0.786	0.647
Linear Regression	0.656	0.6773	0.8016	0.6416
Random Forest	0.6433	0.6621	0.798	0.624
XGBoost	0.6658	0.6888	0.7942	0.6507

Source: Authors' calculations

Table-5 presents a comparison of model performance across four dataset types: structural, bimodal, normal, and uniform. Models trained on structural data consistently delivered the highest predictive accuracy, with R^2 values ranging from approximately 0.67 to 0.85. In contrast, models using bimodal and normal datasets showed significantly weaker performance, with R^2 values between 0.11 and 0.24. Uniform data produced results that were nearly random, with some R^2 values approaching 0.04.

These outcomes underscore a key insight: only the structural dataset effectively captures the underlying pricing logic required for actuarially sound predictions. The other synthetic distributions lack the complexity and realism needed to support meaningful premium modeling.

Table-6,7,8,9,10, & 11 reveals clear variations in model performance across the four data generation methods. On the structural dataset, all models delivered significantly stronger results, with AUC scores ranging from 0.87 to 0.92. LightGBM and XGBoost slightly outperformed the others, each achieving AUCs above 0.68 and balanced precision, recall, and F1-scores near 0.64, indicating their superior ability to distinguish between low-, medium-, and high-risk categories when the data contains strong structural patterns. Linear Regression also performed well, posting the highest recall (0.8016), though its F1-score was marginally lower than those of the tree-based models.

In contrast, the bimodal and normal datasets offered only moderate classification power. AUC values fell within the 0.68–0.90 range. Again, LightGBM and XGBoost led the pack, consistently outperforming Random Forest and Linear Regression with slightly better precision and recall. The top models recorded F1-scores around 0.79–0.80, reflecting a modest but reliable ability to manage false positives and false negatives.

The uniform dataset posed the greatest challenge. AUC values declined further to 0.68–0.69. Precision and recall deteriorated across all models, and F1-scores dropped below 0.42, with Linear Regression performing the worst (F1 = 0.65). These results underscore the difficulty of classifying risk when the data lacks inherent structure.

Therefore, model performance is highly sensitive to the nature of data distribution. Structural datasets consistently produced the strongest results, enabling LightGBM and XGBoost to excel in risk classification. Meanwhile, bimodal and normal datasets allowed for only moderate separation, and the uniform dataset resulted in the weakest performance across all metrics. Overall, gradient boosting methods demonstrated greater robustness than Random Forest and Linear Regression, but their effectiveness was closely tied to the presence of meaningful structure in the data.

Table-12: Threshold values for each ML model

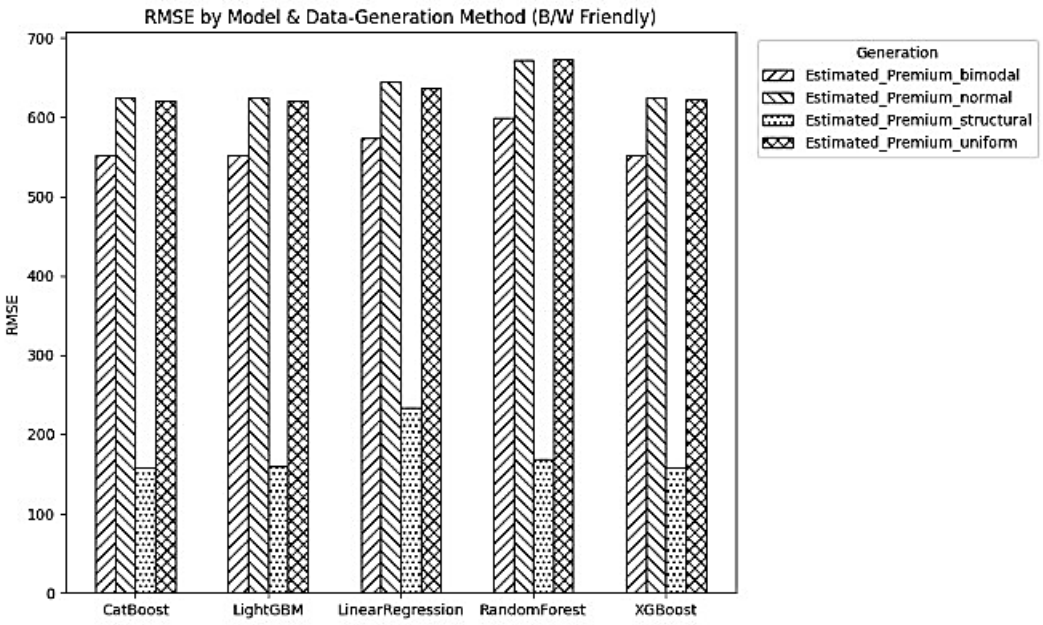
Model	MC_q1 Bimodal	MC_q1 Normal	MC_q1 Structural	MC_q1 Uniform	MC_q2 Bimodal	MC_q2 Normal	MC_q2 Structural	MC_q2 Uniform
CatBoost	1964.28	1948.14	1826.80	175.27	2513.62	2480.31	2166.63	2356
LightGBM	1964.28	1948.14	1826.80	1752.27	2513.62	2480.31	2166.63	2356
Linear Regression	1964.28	1948.14	1826.80	1752.27	2513.62	2480.31	2166.63	2356
Random Forest	1964.28	1948.14	1826.80	1752.27	2513.62	2480.31	2166.63	2356
XGBoost	1964.28	1948.14	1826.80	1752.27	2513.62	2480.31	2166.63	2356

Source: Authors' calculations

MC_q1 and MC_q2 in Table-12 represent the two threshold values used to divide predicted premium amounts into three distinct risk categories—low, medium, and high—for multi-class evaluation. These thresholds are derived exclusively from the training data to avoid any influence from the test set. By default, MC_q1 corresponds to the value below which one-third of the training premiums fall, while MC_q2 marks the point below which two-thirds of the premiums fall. Predictions below MC_q1 are classified as low risk, those between MC_q1 and MC_q2 as medium risk, and those above MC_q2 as high risk. This categorization enables the conversion of continuous premium predictions into discrete classes, allowing the evaluation of a model's ability to classify territories by risk level in addition to predicting precise premium values.

Figure-3 below shows a visualized presentation with two bar plots one for RMSE and one for R^2 . Bar charts deliver clarity fast, short bars in the RMSE plot mean lower error (better) and tall bars in the R^2 plot mean stronger fit.

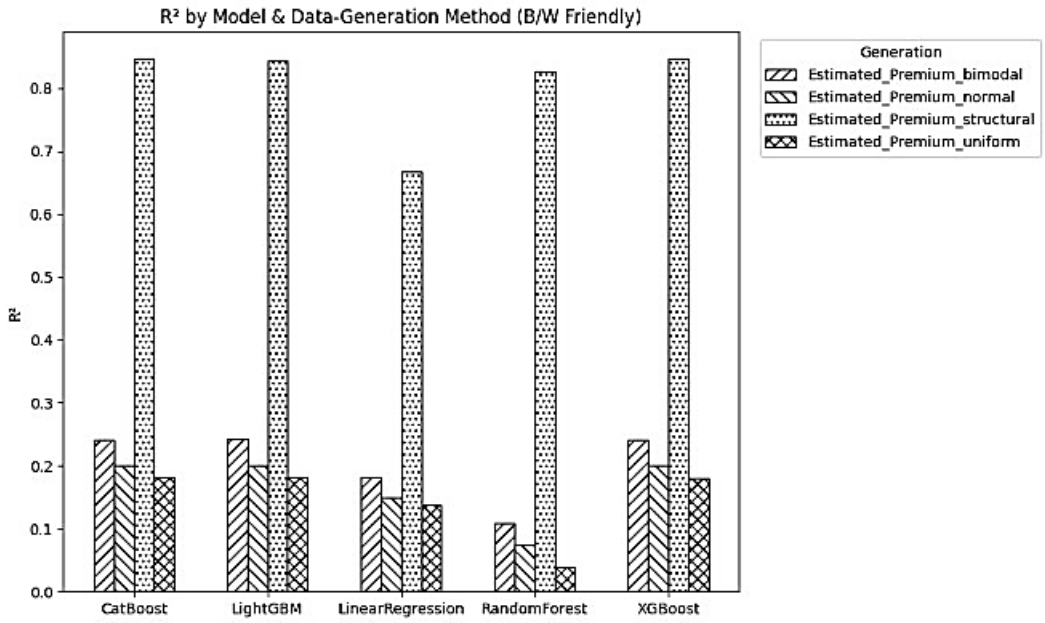
Figure-3: RMSE by Regression Model and Dataset



Source: Authors' calculations

Lower value of the Root Mean Squared Error (RMSE) indicates predictions closer to actual premiums. Across all five algorithms: CatBoost, LightGBM, Linear Regression, Random Forest, and XGBoost, the structural dataset consistently yielded the lowest RMSE values, ranging from 150 to 230. This demonstrates that structural data generation provides a clear, learnable signal. In contrast, the remaining three datasets produced substantially higher errors (RMSE between 550 and 720), rendering them unsuitable for precise premium estimation. Among these weaker options, the bimodal dataset achieved the best performance (approximately 580–630), followed by the normal distribution dataset (approximately 600–660), with the uniform dataset performing worst (approximately 650–720). These results indicate that only the structural dataset supports meaningful predictive modeling, while the other approaches fail to deliver sufficiently accurate signals for premium prediction.

Figure-4: R^2 by Regression Model and Dataset



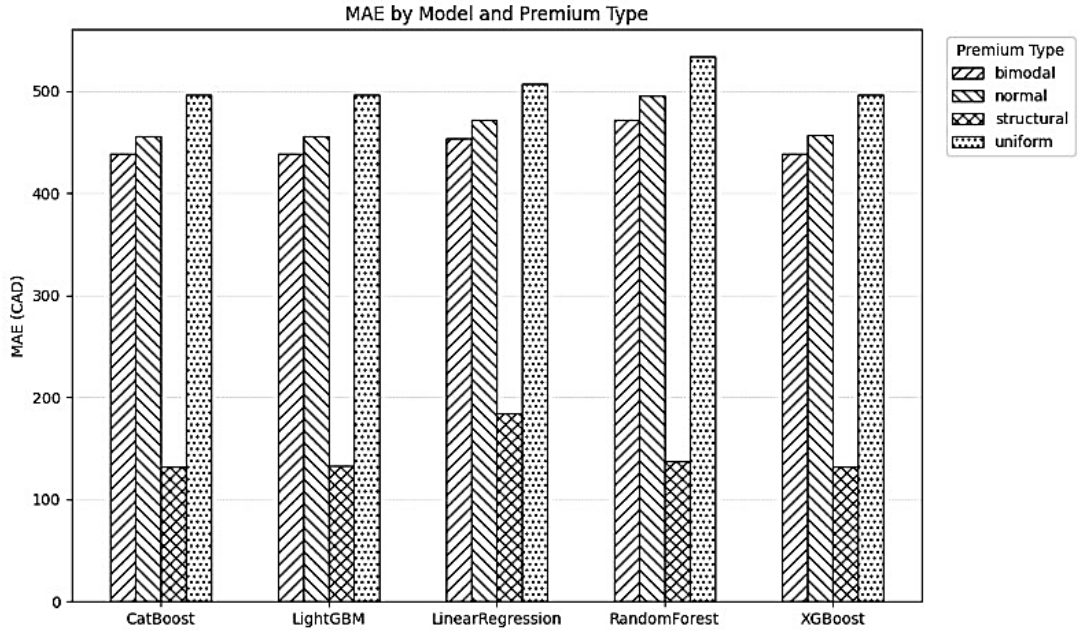
Source: Authors' calculations

Regression Model explanatory power was assessed using the coefficient of determination (R^2) as in figure 4, where higher values indicate a greater proportion of premium variance captured. Across all five algorithms: CatBoost, LightGBM, Linear Regression, Random Forest, and XGBoost the structural dataset consistently achieved R^2 values between 0.85 and 0.88, explaining roughly 85–88% of the variance in premiums.

By contrast, the remaining datasets exhibited minimal explanatory capacity. For bimodal data, the models achieved R^2 values between 0.20 and 0.25, meaning they were able to explain only about one fifth of the variation in the premiums. Predictions on normal and uniform data were similarly limited, with R^2 values hovering around 0.15 to 0.22. Notably, when using Random Forest on uniform data, the R^2 dropped below zero—signaling that the model performed worse than simply predicting the average premium for all cases, a clear indication of poor predictive power in that scenario.

These results confirm that only the structural data encodes the true pricing relationships, whereas the bimodal, normal, and uniform approaches fail to provide a meaningful signal for the models to learn from.

Figure-5: MAE by Regression Model and Premium Type Dataset

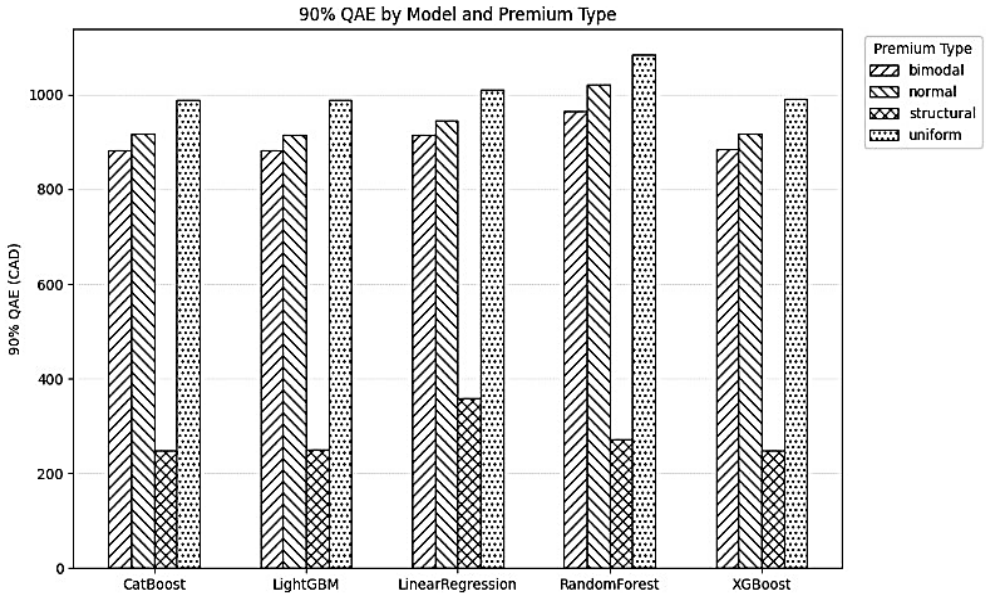


Source: Authors' calculations

Figure-5 presents the Mean Absolute Error (MAE) for five machine learning models of CatBoost, LightGBM, Linear Regression, Random Forest, and XGBoost evaluated across four premium distributions such as Bimodal, Normal, Structural, and Uniform. Results show that the Structural dataset generally leads to lower MAE values across all models, suggesting that it contains more consistent patterns that models can learn effectively. For instance, CatBoost and XGBoost report MAEs around 130–135 CAD for Structural premiums, which are notably lower than those observed for other distributions. In comparison, Bimodal, Normal, and Uniform datasets tend to produce higher MAEs, typically in the range of 440 to 500 CAD indicating that these distributions may be more difficult to model accurately.

Overall, tree based boosting models such as CatBoost, XGBoost, and LightGBM perform better than Linear Regression and Random Forest across most premium types, with particularly strong results on the Structural dataset. Random Forest shows relatively higher MAE even on Structural data, suggesting it may be less well suited for this specific modeling task. The chart highlights how structured input data can improve model performance and how MAE serves as a useful metric for comparing predictive accuracy across different modeling approaches.

Figure-6: 90% QAE by Regression Model and Dataset



Source: Authors' calculations

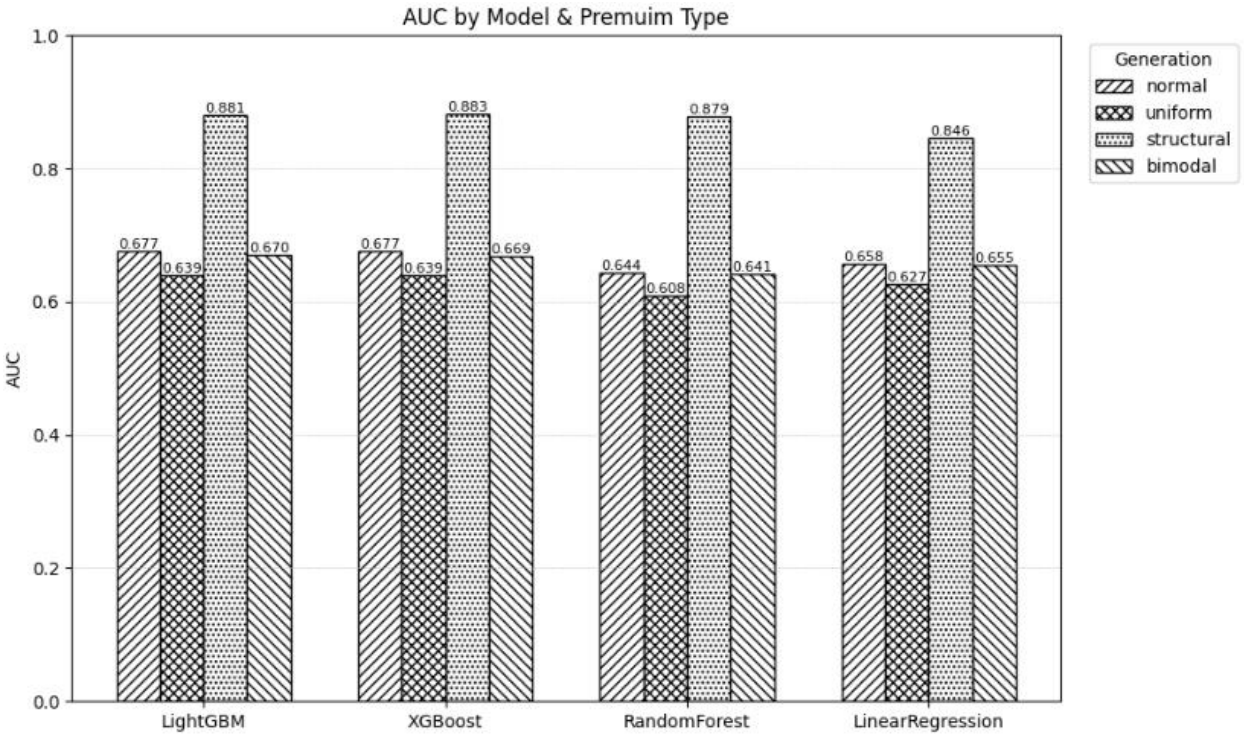
Figure-6 illustrates the 90th Percentile of Absolute Error (90% QAE) for five machine learning models across four premium generation methods. The y-axis represents the error threshold in Canadian dollars (CAD) below which 90% of each model's absolute prediction errors fall. The Structural dataset consistently results in the lowest 90% QAE values across all models, suggesting that its underlying structure supports more reliable predictions, even in higher error scenarios. CatBoost and XGBoost maintain 90% QAE values below 250 CAD for Structural premiums, which is lower than those observed for other distributions. In contrast, Bimodal, Normal, and Uniform datasets tend to produce much higher 90% QAEs typically ranging from 850 to 1,000

CAD indicating that models face greater difficulty managing outlier errors in these less structured datasets.

Among the models, Random Forest shows the highest 90% QAE values, especially on Uniform and Normal data, where errors exceed 1,000 CAD. This contributes to increased volatility and reduced reliability in tail cases. Gradient boosting models such as CatBoost, XGBoost, and LightGBM generally perform better than Linear Regression and Random Forest in limiting large errors, particularly when trained on structured data.

Overall, the 90% QAE metric provides valuable insight into model robustness, where large prediction errors can carry significant financial implications. The results suggest that structured data supports more stable predictions, and that tree-based ensemble methods, especially CatBoost and XGBoost are well suited for managing risk in high-error scenarios.

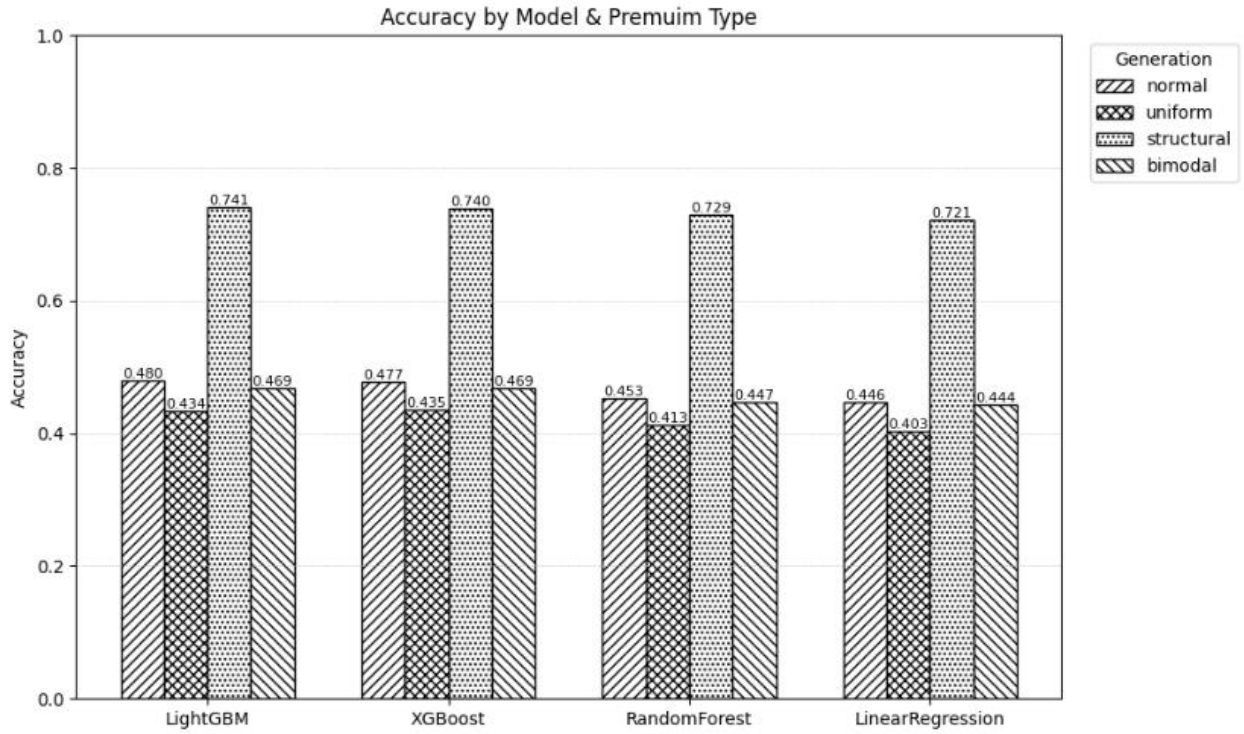
Figure-7*: AUC by Regression Model and Premium Type Dataset*



Source: Authors' calculations

The AUC results clearly demonstrate that the structural dataset consistently enabled the highest class separability across all models, with scores exceeding 0.84 and peaking at 0.883 for XGBoost. In contrast, the normal, bimodal, and uniform datasets yielded significantly lower AUCs, clustering between 0.63 and 0.68. These findings suggest that when data possesses strong inherent structure, models can effectively differentiate between risk categories. However, in less structured distributions, their discriminative power diminishes considerably.

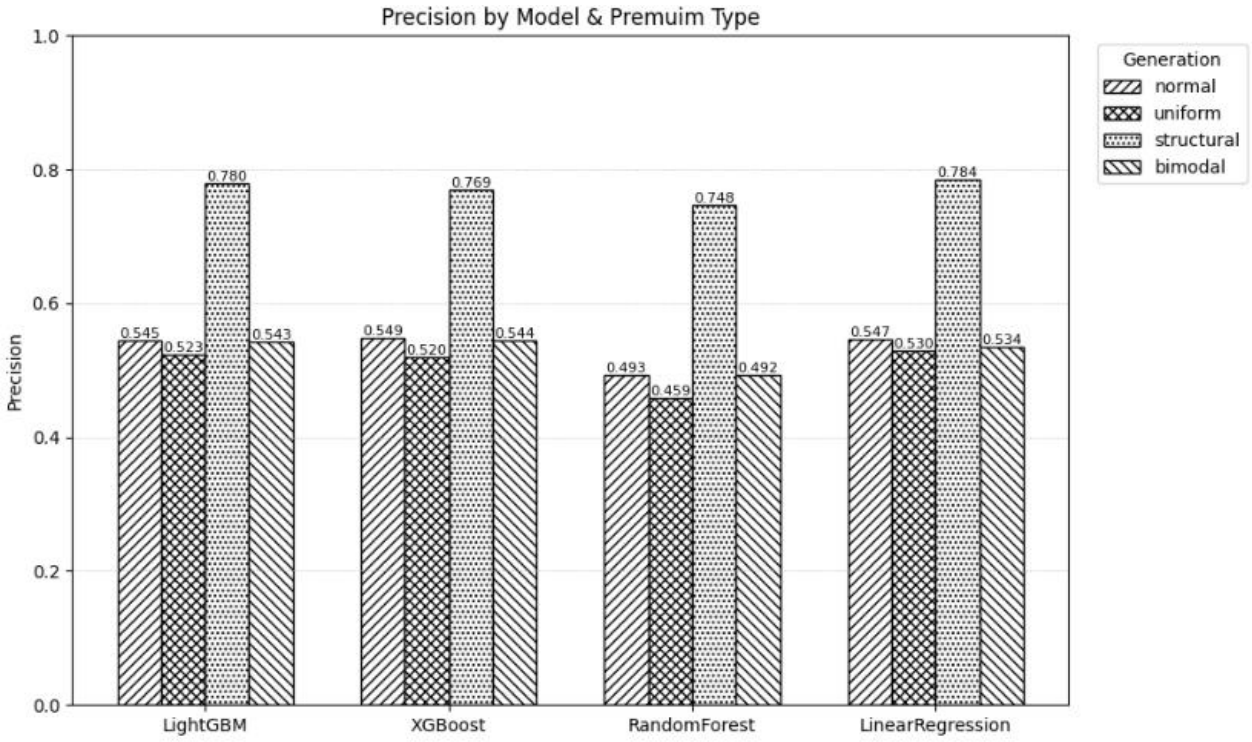
Figure-8*: Accuracy by Regression Model and Premium Type Database*



Source: Authors' calculations

Accuracy trends mirrored the AUC results. The structural dataset again delivered superior performance, with all models achieving accuracies above 0.72, and LightGBM reaching a high of 0.741. Conversely, accuracy dropped to approximately 0.43–0.48 for the normal, uniform, and bimodal datasets, indicating limited classification success under weaker data signals. Tree-based models slightly outperformed Linear Regression in these more challenging scenarios, though the margins were modest.

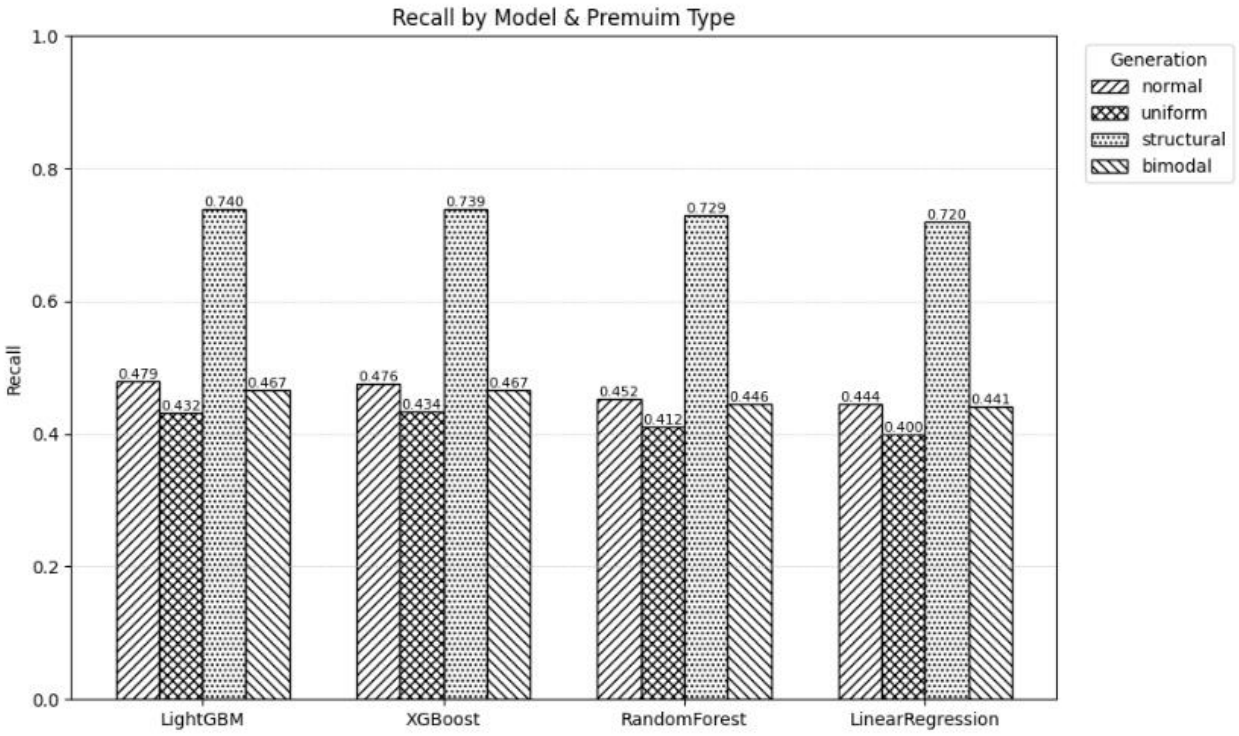
Figure-9*: Precision by Regression Model and Premium Type Database*



Source: Authors' calculations

Precision scores further reinforce the advantage of structured data. The structural dataset enabled the highest precision, ranging from 0.748 with RandomForest to 0.784 with Linear Regression. This means that under structured conditions, models were usually correct when flagging high-risk territories. For the other datasets, precision hovered around 0.52–0.55 for boosting models and dipped to ~0.49 for RandomForest, reflecting a higher incidence of false positives.

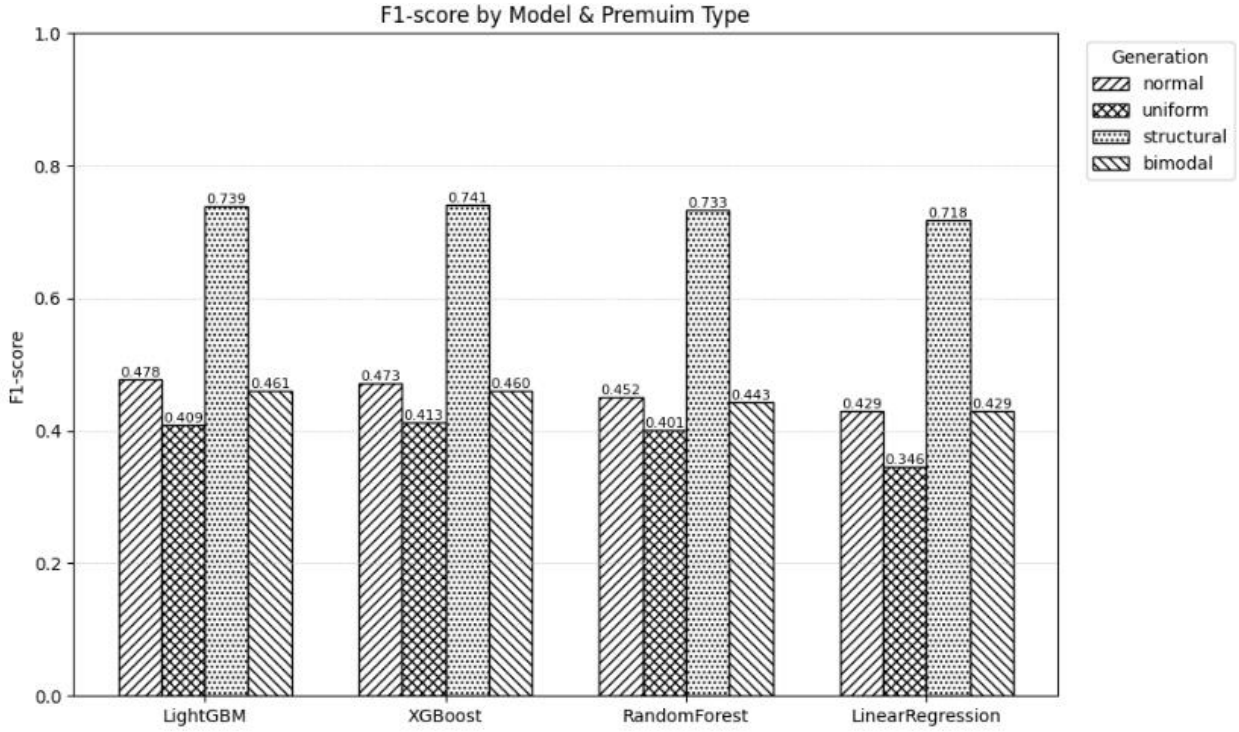
Figure-10*: Recall by Regression Model and Premium Type Database*



Source: Authors' calculations

Recall was also strongest with the structural dataset, reaching 0.740 for both LightGBM and XGBoost. These models successfully identified most true high-risk cases. In contrast, recall dropped to ~0.43–0.47 for the normal, bimodal, and uniform datasets, indicating that many true positives were missed when the data lacked structure.

Figure-11*: F1- Score by Regression Model and Premium Type Database*



Source: Authors' calculations

The F1-score, which harmonizes precision and recall, offers the most comprehensive view of model performance. The structural dataset again led the way, with scores between 0.72 and 0.74, reflecting a strong balance between detecting true risks and minimizing false alarms. On the other datasets, F1-scores fell sharply to the 0.41–0.48 range, with Linear Regression performing worst on uniform data (0.346). These results highlight the critical role of data structure in achieving reliable predictive outcomes.

Sensitivity Analysis

After generating premium datasets using four distribution methods of Bimodal, Normal, Uniform, and Structural, a sensitivity analysis was carried out to examine how changes in input parameters affect model predictions. The ma

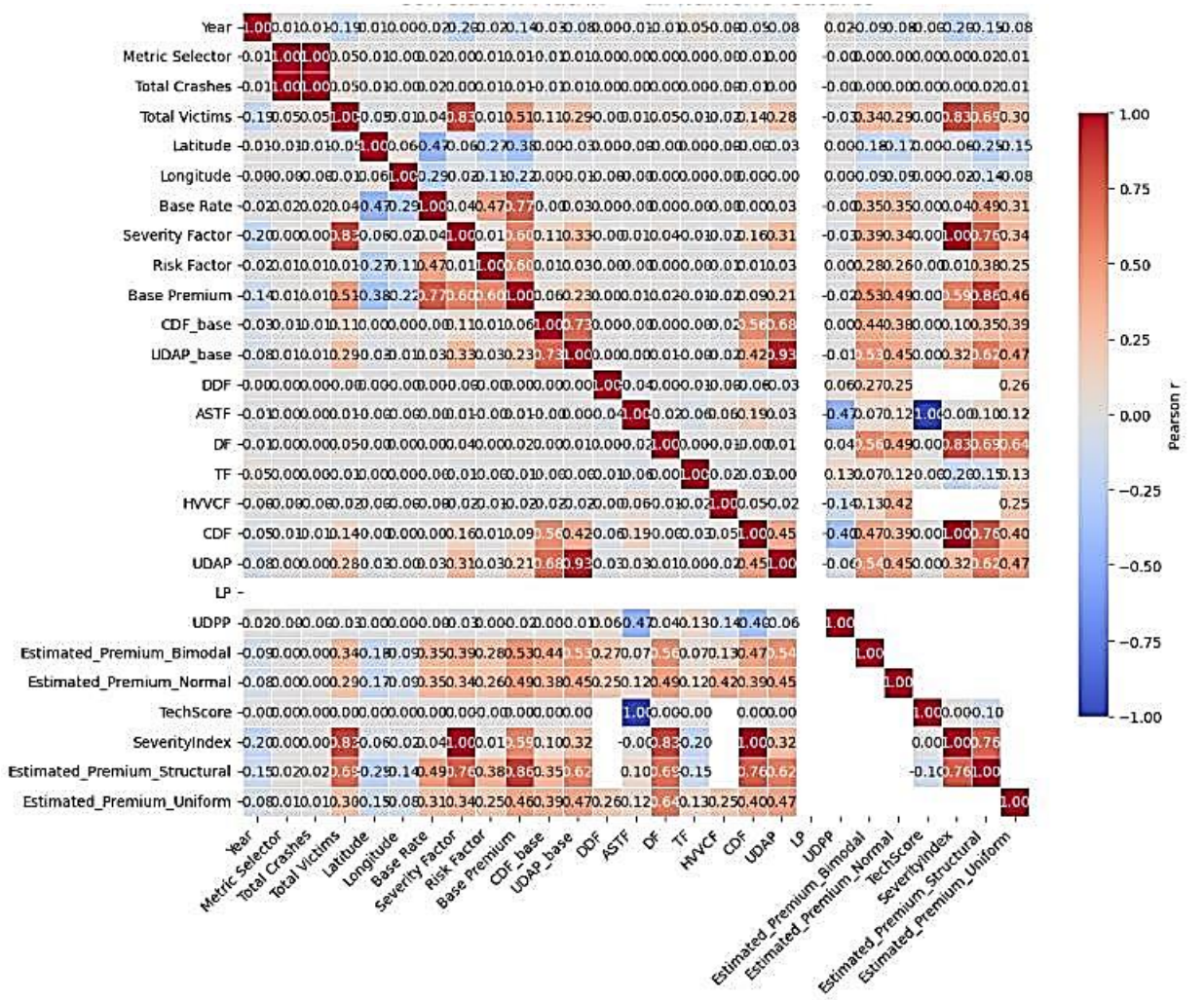
in goal was to understand how small adjustments, particularly to the Combined Driver Factor (CDF), influence final premium estimates. To reflect real world variability, selected rating factors were modified by ±10%, allowing the study to observe how the model responds to controlled changes. This approach is consistent with practices in optimization research, where slight shifts in cost coefficients help assess model stability (Andersen et al., 2025). As noted by Zheng et al.

(2025), sensitivity analysis is a useful tool for identifying which inputs have the greatest impact and for evaluating the reliability of pricing models.

In addition to these tests, analysis was used to explore linear relationships among features across the four datasets. Numeric variables were filtered and combined into a single Data Frame with distribution labels. A heatmap created with Seaborn helped visualize these relationships, showing which features tend to move together and offering insight into possible multicollinearity or structural differences. While correlation analysis does not capture non-linear effects or sensitivity directly, it provided helpful context before conducting more targeted perturbation tests.

Overall, sensitivity analysis supports model validation and helps improve transparency in decision making, especially in areas like insurance pricing. By adjusting key parameters, analysts can identify points where the model's behavior changes noticeably, revealing both stable and sensitive regions. Combining sensitivity testing with exploratory tools like correlation mapping offers a more complete view of model performance and reliability under varying conditions.

Figure-12: Sensitivity Analysis Visualization



Source: Authors' calculations

The influence of each component on the overall premium was quantified by examining its coefficient in the premium formula. The coefficients were then ranked to determine which factors exerted the greatest upward or downward pressure on cost.

Table-13: Factor impact on Sensitivity Analysis

Factor	Coefficient	Impact Description
UDAP (Unlisted Driver Accident Premium)	+0.65	Largest positive effect: adding accident coverage for unlisted drivers raises premiums
Base Rate Premium	+0.46	Reflects base cost based on vehicle type and location
DDF (Driver Discount Factor)	–0.63	Largest discount: safe driving or low mileage significantly reduces premiums
CDF (Combined Driver Factor)	+0.38	Moderate increase: linked to recent violations, tickets, or claims
TF (Transition Factor)	+0.25	Smaller impact: captures rate increases in high traffic or busy areas
ASTF (Advanced Safety Tech Factor)	+0.08	Minor effect: extra safety gear has limited pricing influence
UDPP (Unlisted Driver Protection Premium)	+0.08	Minor effect: basic add-ons offer minimal premium changes

Source: Authors' calculations

Table-7 reveals a clear hierarchy in terms of pricing sensitivity, with some factors exerting strong upward or downward pressure on premiums, while others have only marginal effects. The Unlisted Driver Accident Premium (UDAP) emerged as the most influential factor, with a coefficient of +0.65. This indicates that adding accident coverage for drivers not listed on the policy significantly increases the premium. It reflects the insurer's heightened risk exposure when covering individuals whose driving history may be unknown or unverified.

Next in line is the Base Rate Premium, which showed a coefficient of +0.46. This component captures the foundational cost of coverage, determined primarily by vehicle type and geographic location. It serves as the starting point for premium calculations before additional risk factors are applied.

On the discount side, the Driver Discount Factor (DDF) stood out with a substantial negative coefficient of –0.63. This factor rewards safe driving behavior and low-mileage usage, offering significant reductions in premium for qualifying drivers. It underscores the insurer's incentive to promote responsible driving habits.

The Combined Driver Factor (CDF) contributed a moderate increase to premiums, with a coefficient of +0.38. This factor is closely tied to driver history, particularly recent violations, tickets, or claims. It reflects the elevated risk associated with drivers who have demonstrated problematic behavior on the road. The Transition Factor (TF) had a smaller but still notable impact, with a coefficient of +0.25. This factor captures rate adjustments in high-traffic or densely populated areas, where accident frequency and claim severity tend to be higher.

Finally, both the Advanced Safety Tech Factor (ASTF) and the Unlisted Driver Protection Premium (UDPP) showed minimal influence on pricing, each with coefficients around +0.08. These results suggest that while safety enhancements and basic add-ons are considered in pricing, their effect on the overall premium is relatively minor.

Therefore, the premium structure is most sensitive to unlisted driver coverage (UDAP) and base rate settings, while safe driving behavior offers the most substantial discounts. Other factors, such as traffic density and driver history, play a moderate role, and technological or protection add-ons contribute only marginally to pricing adjustments.

Clustering Approach

Clustering techniques, including K-means, DBSCAN, and hierarchical clustering, were employed to explore the development of new insurance zones. This analysis aligns with the second research question, which examines how algorithmic methods can be applied to assess and reconfigure premium zones based on underlying risk factors.

K-Means

K-Means is an unsupervised learning algorithm used to uncover natural groupings in data. It works by initially placing *k* randomly chosen cluster centers (centroids), assigning each data point to the nearest one, then recalculating the centroids as the mean of their assigned points. This cycle continues iteratively until the centroids stabilize and no longer move (Masruroh et al., 2023). To ensure spatial contiguity, these cluster labels needed to be merged with the geographic polygons and apply a graph-based or region-growing method. BC Data Catalogue was checked but doesn't seem like there is ICBC territory BC Polygon Layer. Therefore, to get going, as a workaround was decided, to get the BC Polygon Layer from BC Data catalogue and then merge labels back into geodata frame. So, territories geojson file (map of territories) was needed to create the Territory clusters file. Once BC polygon layer was ready, it needed to be reaggregated into the 14 ICBC territories and then merged to get ICBC territory BC Polygon layer file. K-Means initially applied followed by the elbow and silhouette plots to visually determine an optimal K.

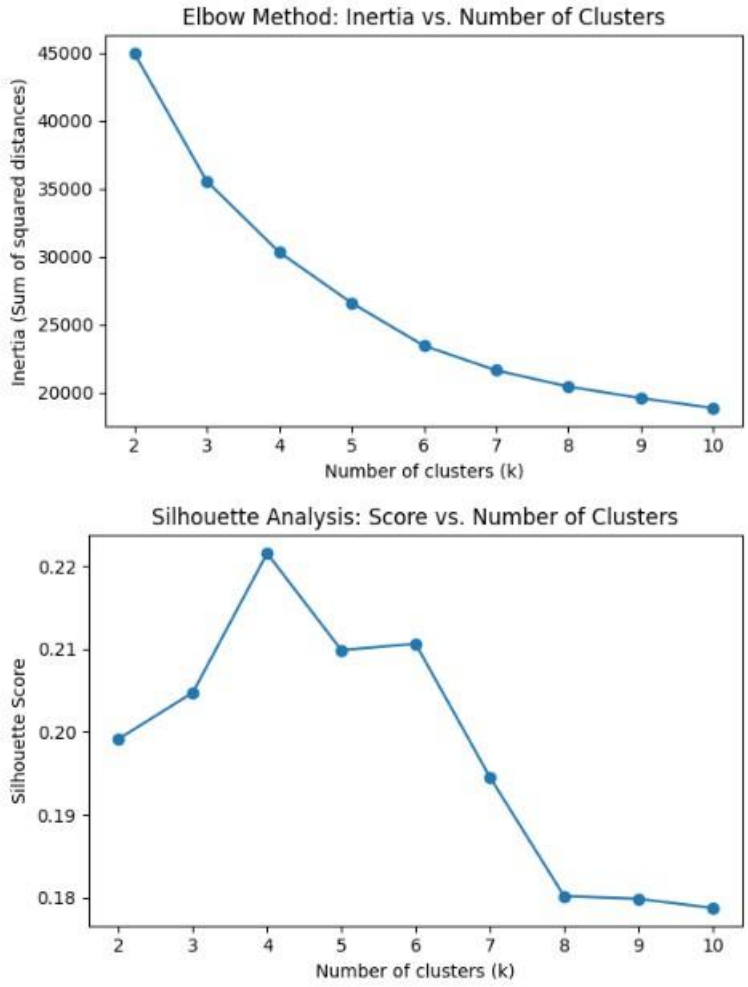
The elbow method analyzes how the number of clusters affects the variance captured within a dataset. It runs K-means clustering across a range of K values, calculating the inertia (or within cluster sum of squares) for each configuration. The goal is to identify a point—often resembling an "elbow" in the plot—where adding more clusters yields diminishing returns in explained variance (Kaur & Saini, 2022).

Elbow Method was chosen because it offers a clear, intuitive way to identify the point at which adding more clusters no longer improves model insight. Beyond this threshold, further segmentation only adds complexity, splitting regions into smaller groups without yielding meaningful gains. By stopping at the optimal number of clusters, it helps maintain a streamlined model while still capturing the most relevant regional crash risk patterns.

Silhouette analysis evaluates how effectively data points are grouped by measuring two aspects: how similar a point is to others in its assigned cluster, and how dissimilar it is from points in neighboring clusters. For each number of clusters (*K*), an average silhouette score is computed across all data points to assess clustering quality (Kumar, Rani, Pippal, & Agrawal, 2025).

Figure-13 presents the result of the Elbow method and Silhouette Analysis.

Figure-13: Elbow Method and Silhouette Analysis



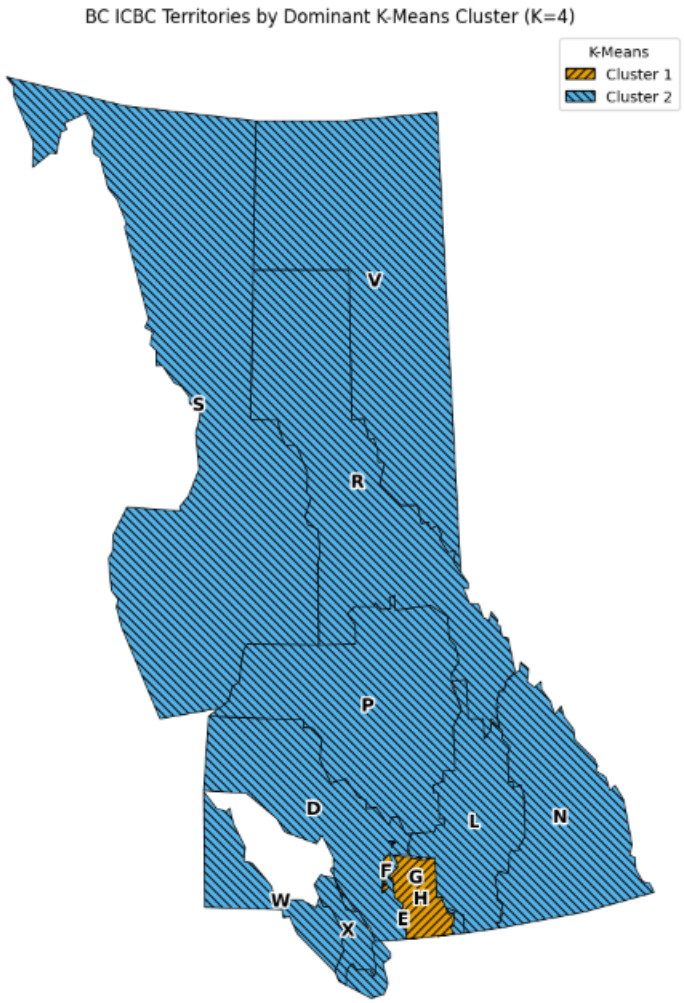
Source: Authors' calculations

Elbow Method for each k (2 through 10), it plots the total "inertia," i.e. the sum of squared distances of each point to its cluster center. And lower inertia means points are closer to their cluster center

(tighter clusters). Therefore, as you increase k, inertia always drops, but after a certain point the improvement is flattening out. Thus, around k=4, you see a noticeable “bend” (or elbow). Beyond 4, adding more clusters only marginally reduces inertia, so 4 gives us most of the benefit without over-splitting. Silhouette Analysis (bottom plot) for each k, it computes the average silhouette score, a number between –1 and +1 that measures how well each point fits its own cluster versus the nearest other cluster. Higher silhouette means clearer, more distinct clusters. Thus, the silhouette score is highest at k=4, clearly indicating that’s where clusters are most internally cohesive and well separated.

Guided by the elbow method and silhouette analysis, the K-means algorithm groups customers into *K* clusters according to similarities in their search behavior. Each customer is assigned to a cluster that best reflects their interaction patterns (Kumar et al., 2025). And since both plots show k=4, 4 clusters seem the sweet spot for the data. Therefore, after selecting K, the model was re-fitted on the final dataset and assigned the resulting labels back to the territory index and merged these cluster labels with the geographic polygons.

Figure-14: BC ICBC Territories by Dominant KMeans_4 Cluster



Source: Authors' calculations

This is a choropleth of BC's 14 ICBC territories, each shaded by its dominant K-Means cluster (k = 4, but only clusters 1 and 2 ever "won"). Here's what it tells us:

Cluster 1: ○ Territories F, G and H ended up in this group:

- F – Squamish / Whistler Area
- G – Pemberton Area / Hope Area
- H – Fraser Valley

○ These three coastal/mountain-corridor zones share similar premium-profile characteristics (higher estimated premiums or different crash patterns), which set them apart in our four-cluster K-Means model.

Cluster 2:

Every other territory fell into this second cluster

- D (Lower Mainland), E (Maple Ridge / Pitt Meadows), L (Thompson / Okanagan), N (Kootenays), P (Cariboo), R (Prince George), S (Northern Coast), V (Peace River Area), W & X & Y (Vancouver Island regions), and even Z (outside BC).

The key insight from this map is that the southern mountain corridor territories specifically regions F, G, and H exhibit characteristic crash metrics that differ noticeably from both the urban Lower Mainland areas (D and E) and the more rural or coastal zones. These distinctions are significant enough to justify grouping F/G/H into their own cluster.

In contrast, the remaining regions across the province display broadly similar patterns in average premiums and crash metrics and thus have been grouped together into Cluster 2. This visualization effectively highlights where the outlier premium-risk profiles are concentrated, while also illustrating that the rest of the province behaves relatively uniformly within the KMeans clustering framework.

Density-Based Spatial Clustering of Applications with Noise

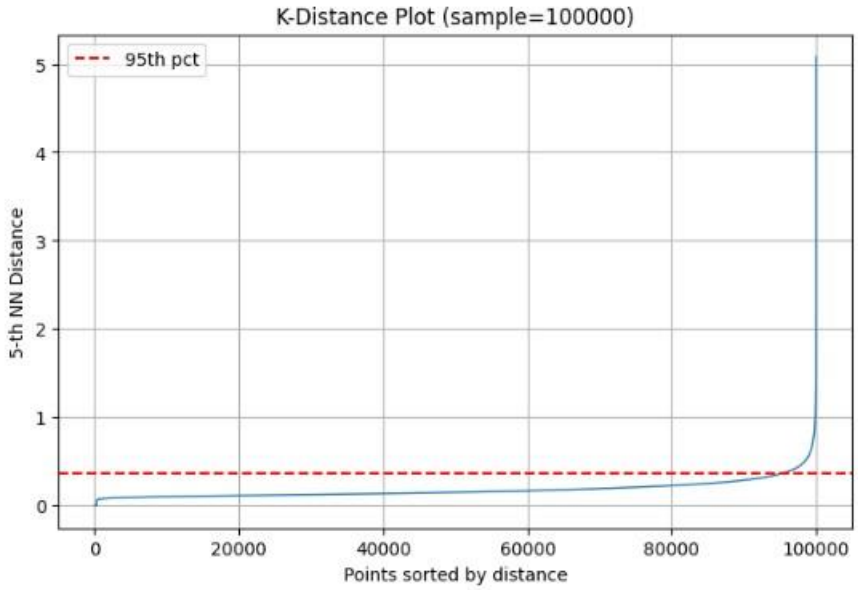
DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a robust clustering algorithm originally (Ester, Kriegel, Sander, & Xu, 1996), designed to handle large spatial datasets. It operates by identifying regions of high data density and forming clusters accordingly, while separating sparser areas that may contain anomalies. Through this process, dense clusters of typical data are filtered out, leaving behind a focused set of data points more likely to represent irregular or anomalous behavior ideal for deeper inspection.

The DBSCAN clustering algorithm identifies data points based on density within the feature space and plays a central role in the pruning process. By detecting dense regions and isolating noise, DBSCAN helps filter out irrelevant or misleading anomaly scores, enhancing the precision of anomaly detection (Firdaus & Suryani, 2024).

In our case, K-Distance plot with same = 100,000 was ran as this process began by randomly sampling up to 10,000 rows to ensure computational efficiency. For each sample point, 5 distances representing the distance to its fifth nearest neighbor were calculated as shown in

Figure 15. These distances were then sorted and plotted, providing a visual representation of their distribution. A dashed line was added at the 95th percentile to serve as a reference point, helping to identify potential outliers or threshold values for clustering or anomaly detection. re-9:

Figure-15: K- Distance Plot (Sample = 100000)



Source: Authors' calculations

To determine an appropriate value for ε in the DBSCAN clustering algorithm, the elbow of the k-distance curve (or alternatively, the 95th percentile line) was used as a guide as in Figure-15. The curve, generated using k=5, showed that distances remained relatively flat until around 0.35, after which they began to rise sharply, indicating a natural inflection point. Based on this observation, the ε value was set to 0.35, which also corresponds to the 95th percentile of the distance distribution. The minimum number of samples required to form a dense region was set to 5.

Now need to run DBSCAN per ICBC territory and Merge into the custom ICBC BC GeoJSON and plot the choropleth. The Python script kept on crashing, so it was divided into two chunks. Chunk one would deal with DBSCAN, and second chunk would plot the BC Map. But then even

with Chunk 1, it ran into RAM issues as it's very tough to run DBSCAN on millions of crash level rows, and since in the end only a territory-level cluster map was needed.

Based on the structure of the data and the clustering objective, the following approach was adopted. First, the dataset was streamed in manageable chunks to compute the mean values of the four premium features for each of ICBC's 14 territories, resulting in a compact summary with one line per territory. Next, DBSCAN was applied to these 14 mean vectors using the previously determined parameters. This clustering step assigned each territory to a distinct group based on its average premium characteristics. Finally, the resulting clusters were mapped onto the BC GeoJSON file, following the same geographic visualization method used in earlier steps.

In chunk 1 script, it never holds all 3.5 M records in memory at once only 500 k at a time then reduces to 14 lines and runs DBSCAN instantly. But on running, it turns out that using $\varepsilon = 0.35$ (picked from the crash-level k-distance plot) on the 14 territory-mean vectors is far too small every territory comes back as "noise" (label –1) as shown in Figure-16. That's because the scale and spacing of the 14 aggregated points are very different from the 3 million+ crash level points we sampled earlier.

Figure-16: DBSCAN Noise Territory Cluster (-1)

	Territory Code	Estimated_Premium_Normal	Estimated_Premium_Structural	\
0	D	1701.891061	1507.859635	
1	E	1870.613193	1677.494184	
2	F	2228.405564	1979.265394	
3	G	2365.596755	2078.041623	
4	H	2172.402263	1914.307999	
5	L	1869.833749	1668.678350	
6	N	1972.046627	1764.327627	
7	P	1924.921753	1715.906334	
8	R	1755.371817	1537.345913	
9	S	1818.359920	1569.977278	
10	V	1684.890382	1504.379741	
11	W	1992.934685	1785.143021	
12	X	1937.608653	1763.729844	
13	Z	1945.759725	1728.188996	

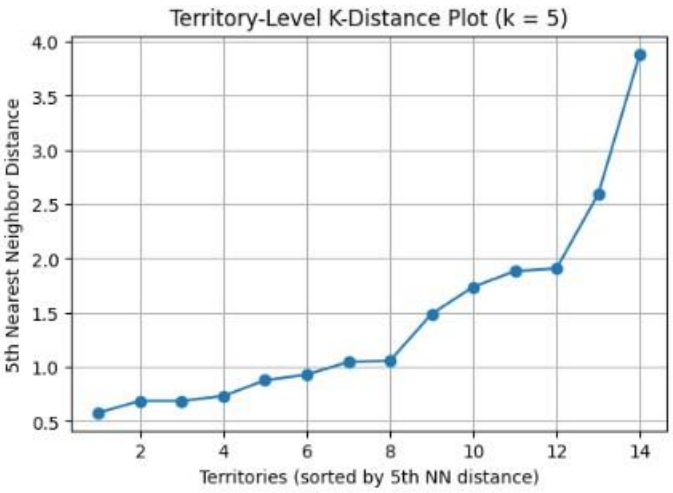
	Estimated_Premium_Bimodal	Estimated_Premium_Uniform	DBSCAN_Cluster
0	1698.938116	1566.285092	-1
1	1895.994264	1754.046053	-1
2	2228.394421	2063.761718	-1
3	2351.631377	2178.018156	-1
4	2126.622113	1977.423335	-1
5	1853.666890	1715.848233	-1
6	1988.246087	1826.244765	-1
7	1896.727301	1751.860796	-1
8	1699.079765	1575.114377	-1
9	1780.957543	1615.798478	-1
10	1676.888140	1543.882658	-1
11	2020.472535	1868.370286	-1
12	1983.775366	1829.920011	-1
13	1938.568763	1789.736458	-1

Source: Authors' calculations

Therefore, we need to pick ε for the territory-means themselves. To do that, we need to:

1. Compute a k-distance plot on the 14×4 table of (normal/structural/bimodal/uniform) means.
2. Choose ε at the "elbow" of that small plot (or at a high percentile).
3. Rerun DBSCAN with that new ε and see which territories group together. After re-running it, new Territory Level K Distance Plot was plotted as shown in figure 12 below.

Figure-17: Territory-Level K-Distance Plot



Source: Authors' calculations

Clearly the "elbow" in the territory-level plot appears right around the jump from ~1.1 up to ~1.5 (that big jump at the 8–9th territory). Picking ε = 1.5 seemed reasonable (with min samples=5 as before) and rerun DBSCAN now on the 14 mean-vectors. And results were better than last time, as shown below, from the 14-territory run with ε=1.5 and min samples=5, we have:

- Cluster 0: 12 of the 14 territories (D, E, H, L, N, P, R, S, V, W, X, Z)
- Noise (–1): Territories F and G

In simple words, the DBSCAN algorithm sees F (Squamish/Whistler) and G (Pemberton/Hope) as "outliers" relative to the rest of BC when looking at those four mean premium values much like they stood out in K-Means, but here they're treated as noise rather than forming their own cluster. All the other territories fall into a single dense group (cluster 0).

Figure-18: DBSCAN updated Territory Cluster

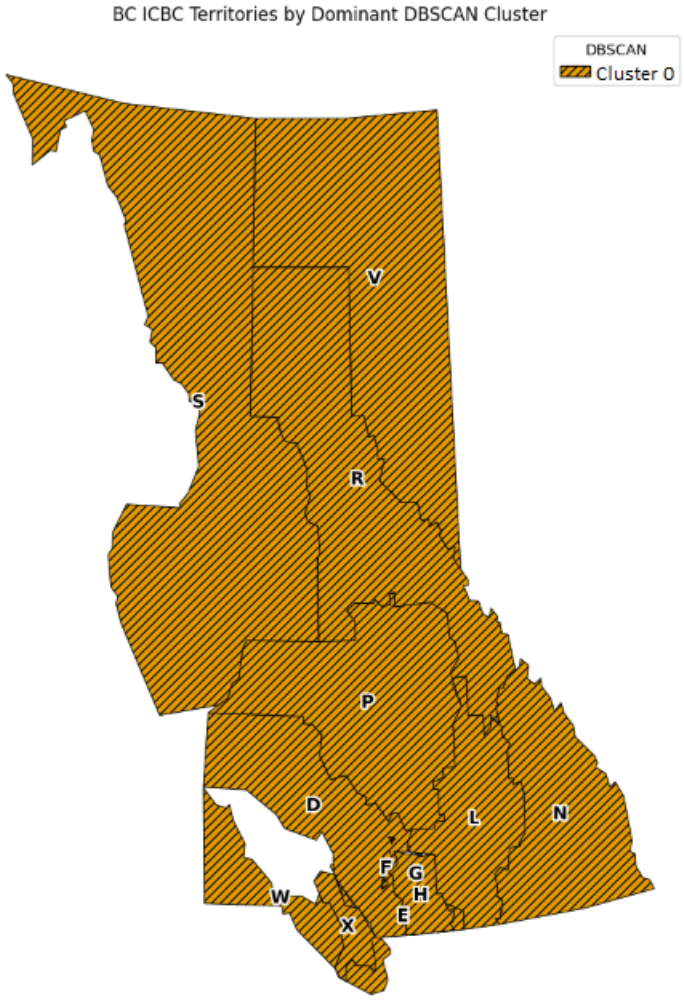
	Territory Code	Estimated_Premium_Normal	Estimated_Premium_Structural
0	D	1701.891061	1507.859635
1	E	1870.613193	1677.494184
2	F	2228.405564	1979.265394
3	G	2365.596755	2078.041623
4	H	2172.402263	1914.307999
5	L	1869.833749	1668.678350
6	N	1972.046627	1764.327627
7	P	1924.921753	1715.906334
8	R	1755.371817	1537.345913
9	S	1818.359920	1569.977278
10	V	1684.890382	1504.379741
11	W	1992.934685	1785.143021
12	X	1937.608653	1763.729844
13	Z	1945.759725	1728.188996

	Estimated_Premium_Bimodal	Estimated_Premium_Uniform	DBSCAN_Cluster
0	1698.938116	1566.285092	0
1	1895.994264	1754.046053	0
2	2228.394421	2063.761718	-1
3	2351.631377	2178.018156	-1
4	2126.622113	1977.423335	0
5	1853.666890	1715.848233	0
6	1988.246087	1826.244765	0
7	1896.727301	1751.860796	0
8	1699.079765	1575.114377	0
9	1780.957543	1615.798478	0
10	1676.888140	1543.882658	0
11	2020.472535	1868.370286	0
12	1983.775366	1829.920011	0
13	1938.568763	1789.736458	0

Source: Authors' calculations

To visualize the DBSCAN results, the 14-row territory mean vectors each assigned a DBSCAN cluster label were merged with the simplified set of 14 ICBC territorial polygons. This combined dataset was then used to generate a choropleth map, illustrating in Figure 19 the spatial distribution of clusters across British Columbia.

Figure-19: BC ICBC Territories by DBSCAN



Source: Authors' calculations

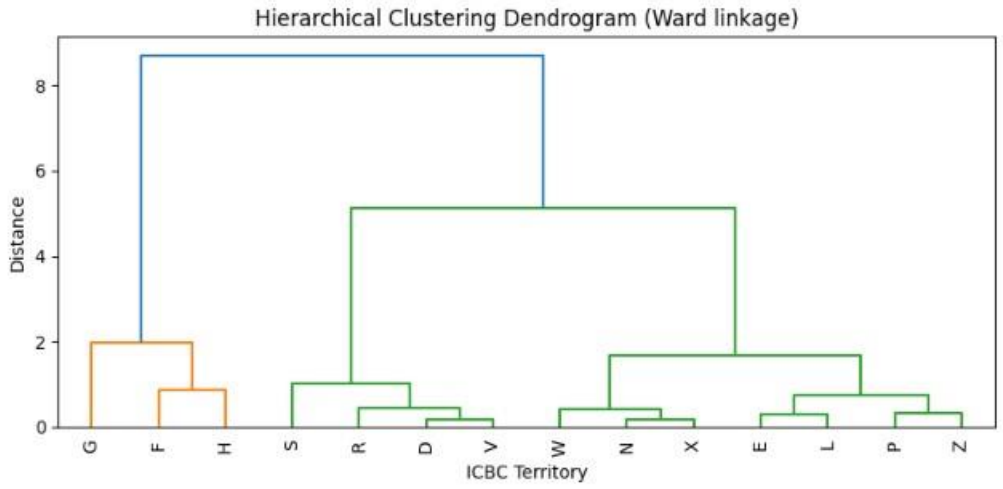
DBSCAN algorithm sees F (Squamish/Whistler) and G (Pemberton/Hope) as "outliers" relative to the rest of BC therefore considering them as noise. And rest as cluster in same pattern.

Hierarchical Clustering

Hierarchical clustering is an unsupervised machine learning method for customer segmentation involves constructing a dendrogram. It is a branching diagram that reveals how data points are related. This visual structure emerges from hierarchical clustering and helps uncover natural groupings within the dataset (Roux, 2018).

As Pramono, Surjandari, & Laoh, 2019, describes that agglomerative hierarchical clustering is applied to the prepared dataset to evaluate its results alongside those from K-means. Using Ward's minimum variance method which combines Ward's linkage with Euclidean distance, the algorithm constructs clusters by iteratively merging groups that result in the smallest increase in overall variance. This approach ensures that each resulting cluster maintains high internal consistency, with its members being closely related.

Figure-20: Hierarchical Clustering Dendrogram



Source: Authors' calculations

Figure-20 describes that the dendrogram has two clear jumps:

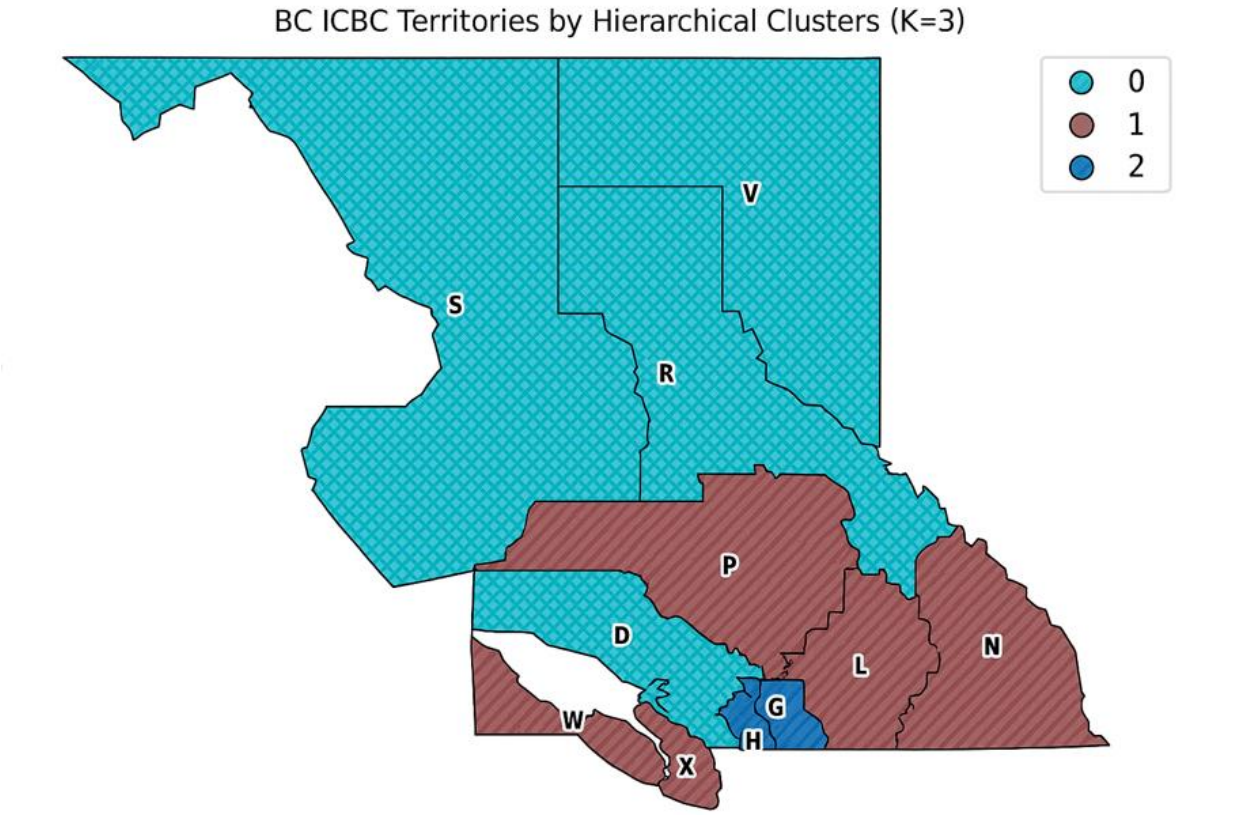
1. Largest jump at ~9 – if you cut here, you get 2 clusters:
    - Cluster A: G, F, H, S, R
    - Cluster B: D, E, L, N, P, V, W,

    X, Z
2. Next jump at ~5 – if you cut here, you get 3 clusters:
    - Cluster 1: G, F, H, S
    - Cluster 2: R, D, V, W, Z
    - Cluster 3: E, L, P, N, X

We select K=3 as a middle group for the big interior/north-coast block. Ward's agglomerative clustering was run on the territory-level means to create 3 clusters by joining those labels back to

the custom ICBC BC GeoJSON and rendering a clean choropleth with territory codes at their centroids

Figure-21 : BC ICBC Territories by Hierarchical Clusters



Source: Authors' calculations

Figure-21 clearly described that among the clustering methods evaluated, hierarchical clustering has clearly yielded the most favorable results to date. While alternative techniques have offered varying levels of performance, hierarchical clustering has consistently demonstrated superior effectiveness and interpretability in this analysis as it was able to dissect the data into three effective clusters.

Conclusion

The central idea of this paper was to examine whether machine learning algorithms could be used to enhance insurance premium zoning by leveraging regional crash data and reconstructing

risk-based pricing schemes. Through the generation of estimated premiums using four data distributions, combined with clustering techniques, regression analysis, and sensitivity analysis, this study explored patterns in geographic risk and assessed how stable premium predictions remain when inputs are varied. The results suggest that using a structured algorithm, risk informed data can lead to more accurate and reliable pricing outcomes, offering support for updating traditional territory-based insurance zoning with more data-driven methods.

The findings of this study carry several implications for ICBC's current premium policy approach, suggesting opportunities for refinement and modernization. At present, ICBC relies on fixed geographic boundaries and standardized rating tables that may not fully capture the diversity of risk profiles across British Columbia. The clustering analysis reveals that certain regions such as Squamish/Whistler and Pemberton/Hope exhibit crash severity and premium patterns that diverge from provincial norms. This suggests that the existing territorial zoning may benefit from re-evaluation, with the potential to redraw boundaries consistent with the results of such machine learning algorithm so that premiums more accurately reflect localized risk.

The sensitivity analysis further indicates that some rating factors, particularly the crash-type factor (CDF) and the unlisted-driver penalty (UDAP) have a disproportionate influence on premium outcomes. These findings imply that ICBC could achieve greater precision and fairness by focusing its calibration efforts on these high impact variables. Conversely, factors that showed minimal influence, such as ASTF and DDF, might be simplified or removed, reducing complexity in the rating structure without significantly affecting pricing accuracy. Another implication arises from the identification of outlier territories through DBSCAN clustering. These regions do not conform to broader patterns and may not be well served by a uniform pricing approach. ICBC could consider developing tailored strategies for these areas, such as customized renewal programs or targeted communications, to better address their distinct risk characteristics and avoid adverse selection.

The study also highlights the potential value of machine learning in premium setting. The strong performance of gradient boosted tree models suggests that nonlinear interactions among risk factors can offer predictive insights beyond traditional factor tables. This points to the possibility

of piloting an enhanced ML pricing engine that updates regularly with new claims data, allowing ICBC to respond more dynamically to changes in driving behavior, infrastructure, and emerging risks.

Finally, the study underscores the importance of transparency in building public trust. By sharing the rationale behind clustering methods, sensitivity results, and model performance, ICBC could help policyholders better understand how their premiums are determined. This could improve confidence in the system and clarify which factors are within a driver's control versus those shaped by external conditions.

Therefore, ICBC's current premium policy approach could be strengthened by adopting more granular, data-driven methods. These changes have the potential to improve fairness, responsiveness, and transparency, while better aligning premiums with actual risk across the province.

Thus, in conclusion, a robust end to end pipeline was developed to understand and model contraventions and evaluating the zones across British Columbia. From cleaning crash records and assigning territories to estimating premiums using four approaches, a foundation for data driven insurance modeling was built. Our clustering methods revealed consistent geographic patterns, particularly highlighting territories F and G as outliers. Sensitivity analysis and feature correlations deepened our understanding, while benchmarking regression models confirmed that tree-based techniques, especially XGBoost are best suited for these engineered features. And hierarchical Clustering had best results with three clusters.

Future Scope and Limitations

The implementation of advanced machine learning models presents a promising direction for improving ICBC's current premium setting scheme. These models particularly gradient boosted tree algorithms such as XGBoost, LightGBM, and CatBoost offer the ability to capture complex, non-linear relationships among crash severity, territorial risk, vehicle type, and other variables. This enables more accurate risk stratification within existing zones, allowing for personalized pure premiums that reduce cross subsidization between low and high-risk drivers. Clustering techniques, including K-Means and hierarchical clustering, reveal opportunities to redraw territorial boundaries based on actual claims experience. This could lead to more homogeneous risk pools

and enhance the accuracy of base rates. Additionally, DBSCAN analysis identifies outlier regions that do not conform to broader patterns. These territories could be treated with bespoke rating strategies or targeted mitigation programs, rather than being forced into generalized tables.

The sensitivity analysis highlights key rating factors such as the crash-type factor (CDF) and unlisted-driver penalty (UDAP) that have the greatest impact on premium outcomes. Focusing calibration efforts on these high leverage variables could improve pricing responsiveness while simplifying less influential components of the rating structure.

A key advantage of the machine learning approach is its capacity for continuous model updating. Unlike static tables revised annually, the models can be retrained on fresh claims data at regular intervals, allowing ICBC to respond more dynamically to emerging trends such as changes in driving behavior, infrastructure developments, or economic shifts. Additionally, tree-based models offer transparency through feature importance and partial dependence analysis, helping ICBC explain rate decisions to policyholders and regulators.

However, several limitations must be acknowledged. The current models rely on aggregated territorial data, which limits spatial resolution and may obscure localized risk patterns. Future iterations could benefit from finer geographic granularity, rigorous threshold validation, and the inclusion of contextual variables such as weather conditions, time-of-day effects, and traffic density. The absence of driver specific crash data also necessitated certain methodological compromises, constraining the ability to model individual-level risk with precision. Operationally, integrating machine learning into ICBC's pricing workflow will require investment in infrastructure, data governance, and staff training. Regulatory considerations around algorithmic fairness and transparency must also be addressed to ensure compliance and public trust. Moreover, the effectiveness of these models depends on the quality and timeliness of input data, underscoring the importance of robust data pipelines and validation protocols. Despite these challenges, the approach outlined in this study establishes a replicable framework for data-driven risk pricing and road safety evaluation across diverse geographic contexts. With continued refinement and careful implementation, it can help achieve a pathway towards a more adaptive, equitable, and transparent premium-setting system.

References

1. Akinshin, A. (2022). Quantile absolute deviation: A generalization of the median absolute deviation. *arXiv*. https://arxiv.org/pdf/2208.13459
2. Andersen, K. A., Boomsma, T. K., Efkes, B., & Forget, N. (2025). Sensitivity analysis of the cost coefficients in multiobjective integer linear optimization. *Management Science, 71*(2), 1120–1137. https://doi.org/10.1287/mnsc.2021.01406
3. Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R^2 is more informative than SMAPE, MAE, MAPE, MSE, and RMSE in regression analysis evaluation. *PeerJ Computer Science, 7*, e623. https://doi.org/10.7717/peerj-cs.623
4. David, M. (2015). Auto insurance premium calculation using generalized linear models. *Procedia Economics and Finance, 20*, 147–156.
5. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010
6. Firdaus, G. M., & Suryani, V. (2024). Enhanced pruning process with DBSCAN for attack detection. In *2024 1st International Conference on Cyber Security and Computing (CyberComp)* (pp. 113–118). IEEE. https://doi.org/10.1109/CyberComp60759.2024.10913747
7. Insurance Corporation of British Columbia. (2020). *Basic tariff*. https://www.icbc.com/assets/pa/6PyY5DEcoIT7z2ujjlwb3d/basic-tariff.pdf
8. Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica, 40*(6), 979–1001. https://doi.org/10.2307/1913851
9. Haji Mohammad, F. (2023). *Insurance premium calculation using machine learning methodologies* (Master's thesis, Carleton University). Carleton University Institutional Repository.
10. Healy, M., & Westmacott, M. (1956). Missing values in experiments analysed on automatic computers. *Applied Statistics, 5*(3), 203–206.
11. Henckaerts, R., Antonio, K., Clijsters, M., & Verbelen, R. (2018). A data driven binning strategy for the construction of insurance tariff classes. *Scandinavian Actuarial Journal, 2018*(8), 681–705. https://doi.org/10.1080/03461238.2018.1429300
12. Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *GEOSCIENTIFIC MODEL DEVELOPMENT*, *15*(14), 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022
13. Karakilic, M., Hatas, H., & Pacal, I. (2025). Open-circuit fault detection in T-type MLI using XGBoost: A machine learning-based approach. In *2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICHORA65333.2025.11017205
14. Kaur, B., & Saini, J. R. (2022). A strategy to identify loyalty using elbow curve method for customer segmentation. In *2022 IEEE Pune Section International Conference (PuneCon)* (pp. 1–7). IEEE. https://doi.org/10.1109/PuneCon55413.2022.10014742
15. Ke, G. L., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 3149–3157). Curran Associates, Inc.
16. Kowshalya, G., & Nandhini, M. (2018). Predicting fraudulent claims in automobile insurance. 2018 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT), 1338–1343. IEEE. https://ieeexplore.ieee.org/document/8473034

17. Kumar, R., Upadhyay, S., Rakhra, M., Mrsic, L., Prashar, D., & Khan, A. A. (2024). A machine learning and ratemaking evaluation of four auto insurance pure premium modeling algorithms. 2024 28th International Computer Science and Engineering Conference (ICSEC), 1–8. IEEE. https://doi.org/10.1109/ICSEC62781.2024.10770731
18. Kumar, S., Rani, R., Pippal, S. K., & Agrawal, R. (2025). Customer segmentation in ecommerce: K-means vs hierarchical clustering. *Telkomnika, 23*(1), 119–128. https://doi.org/10.12928/TELKOMNIKA.v23i1.26384
19. Ling, C. X., Huang, J., & Zhang, H. (2003). AUC: A better measure than accuracy in comparing learning algorithms. In *Proceedings of the 16th Canadian Conference on Artificial Intelligence* (pp. 329–341). https://doi.org/10.1007/3-540-44886-1_25
20. Luo, M., Zhang, R., Yan, X., Ma, J., Huang, J., & Chen, X. (2021). Combination of feature selection and CatBoost for prediction: The first application to the estimation of aboveground biomass. *Forests, 12*(2), 216. https://www.mdpi.com/1999-4907/12/2/216
21. Masruroh, S. U., Fadilah, A. T., Hulliyah, K., Ramadhan, A. F., Putri, R. A., & Irfan, M. A. (2023). Implementation of the K-means clustering algorithm for targeting ads. Case study: IBM Watson Analytics car insurance customer data. In *2023 11th International Conference on Cyber and IT Service Management (CITSM)* (pp. 1–3). IEEE. https://doi.org/10.1109/CITSM60085.2023.10455449
22. Pramono, P. P., Surjandari, I., & Laoh, E. (2019). Estimating customer segmentation based on customer lifetime value using two-stage clustering method. In *2019 16th International Conference on Service Systems and Service Management (ICSSSM)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICSSSM.2019.8887704
23. Roux, M. (2018). A comparative study of divisive and agglomerative hierarchical clustering algorithms. *Journal of Classification, 35*(3), 345–366. https://doi.org/10.1007/s00357-018-9259-9
24. Qian, Q., Jia, X., Lin, H., & Zhang, R. (2021). Seasonal forecast of nonmonsoonal winter precipitation over the Eurasian continent using machine-learning models. *Journal of Climate, 34*(17), 7113–7129. https://doi.org/10.1175/JCLI-D-21-0113.1
25. Samat, A., Li, E., Du, P., Liu, S., Miao, Z., & Zhang, W. (2022). CatBoost for RS image classification with pseudo label support from neighbor patches-based clustering. *IEEE Geoscience and Remote Sensing Letters, 19*, Article 8004105. https://doi.org/10.1109/LGRS.2020.3038771
26. Saranya, G., & Pravin, A. (2021). An efficient feature selection approach using sensitivity analysis for machine learning based heart disease classification. 2021 11th International Conference on Communication Systems and Network Technologies (CSNT), 539–542. IEEE. https://doi.org/10.1109/CSNT51715.2021.9509673
27. Selvakumar, V., Satpathi, D. K., Kumar, P. T. V. P., & Haragopal, V. V. (2021).Predictive modeling of insurance claims using machine learning approach for different types of motor vehicles. *Universal Journal of Accounting and Finance*, 9(1), 1–14. https://doi.org/10.13189/ujaf.2021.090101
28. Sharma, N. (2023, June 6). *Understanding and applying F1 score: AI evaluation essentials with hands-on coding example. Arize AI*. https://arize.com/blog-course/f1-score/
29. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002
30. Węglarczyk, S. (2018). Kernel density estimation and its application. *ITM Web of Conferences, 18*, 00037. https://doi.org/10.1051/itmconf/20182300037

31. Zheng, B., Shui, Q., Luo, Z., Hu, P., Yang, Y., Lei, J., & Yin, G. (2025). Optimization of engine piston performance based on multi-method coupling: Sensitivity analysis, response surface model, and application of genetic algorithm. *Materials, 18*(13), 3043. https://doi.org/10.3390/ma18133043
32. Zimmerman, N., Presto, A. A., Kumar, S. P. N., Gu, J., Hauryliuk, A., Robinson, E. S., Robinson, A. L., & Subramanian, R. (2018). A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. *Atmospheric Measurement Techniques, 11*(1), 291–313. https://doi.org/10.5194/amt-11291-2018

Appendix

A.1.

```
# — Finally: structural estimated premium —
factors = ["CDF","DDF","HVVCF","ASTF","DF","TF"]
df["Estimated_Premium_Structural"] = (
    df["Base Premium"] *
    df[factors].prod(axis=1)  # multiply all factors
    + df["LP"] + df["UDPP"] + df["UDAP"]
)

# Preview
print(df[["Base Premium"] + factors + ["LP","UDPP","UDAP","Estimated_Premium_Structural"]].head())


#output
df.to_csv("Crash_Data_With_Estimated_Premium_structural_data.csv", index=False)
print("Fixed and saved: Crash_Data_With_Estimated_Premium_structural_data.csv")
```

A.2.

```
# — 6) Final estimated premium —
df["Estimated_Premium_Bimodal"] = (
    df["Base Premium"]
    * df["CDF"] * df["DDF"] * df["HVVCF"]
    * df["ASTF"] * df["DF"] * df["TF"]
) + df["LP"] + df["UDPP"] + df["UDAP"]

# — 7) Preview
print(df.head(5))

#output
df.to_csv("Crash_Data_With_Estimated_Premium_bimodal_data.csv", index=False)
print("Fixed and saved: Crash_Data_With_Estimated_Premium_bimodal_data.csv")
```

A.3.

```
# 7) Compute final "normal-dist" premium
df["Estimated_Premium_Normal"] = (
    df["Base Premium"]
    * df["CDF"] * df["DDF"] * df["HVVCF"]
    * df["ASTF"] * df["DF"] * df["TF"]
) + df["LP"] + df["UDPP"] + df["UDAP"]

# 8) Preview
print(df.head(5))

#output
df.to_csv("Crash_Data_With_Estimated_Premium_normal_data.csv", index=False)
print("Fixed and saved: Crash_Data_With_Estimated_Premium_normal_data.csv")
```