Title

Predictive Modeling for Insurance Pricing: A Comparative Analysis of Actuarial Techniques and Machine Learning Algorithms

Author

Lyu, Ting

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Predictive Modeling for Insurance Pricing:

A Comparative Analysis of Actuarial Techniques

and Machine Learning Algorithms

A thesis submitted in partial satisfaction of the

requirements for the degree

Master of Science in Statistics

by

Ting Lyu

2024

ABSTRACT OF THE THESIS

Predictive Modeling for Insurance Pricing:

A Comparative Analysis of Actuarial Techniques

and Machine Learning Algorithms

by

Ting Lyu

Master of Science in Statistics

University of California, Los Angeles, 2024

Professor Yingnian Wu, Chair

This thesis examines insurance pricing with the goal of improving predictive accuracy through a comparative analysis of traditional actuarial techniques and modern machine learning algorithms. By utilizing real-world datasets from insurance companies, the research applies five distinct methodologies to analyze the key variables within the insurance dataset. The primary objective is to identify the most effective approaches in forecasting claim amounts. The findings of this study seek to advance predictive accuracy and provide substantial business value, thereby promoting innovation and excellence in risk management within the insurance industry.

The thesis of Ting Lyu is approved.

Arash A. Amini

Hongquan Xu

Yingnian Wu, Committee Chair

University of California, Los Angeles

2024

To my parents

for all the love and support

Table of Contents

List of Figures

ACKNOWLEDGMENTS

I am deeply grateful to my committee chair, Dr. Yingnian Wu, for his invaluable guidance and support throughout my research. I am also very thankful to Dr. Arash A. Amini and Dr. Hongquan Xu for their insightful inputs and encouragement. I appreciate the deep understanding of my field I gained from their courses in my first year and their continued support during my thesis work in the second year. Special thanks to the staff at the Department of Statistics and Data Science for their unwavering support.

1. Introduction

In the era of big data, various industries have been at the forefront of digital transformation and big data integration, using these technologies to enhance business processes (Le, 2023). In the insurance industry, the accuracy of pricing models is critical for objective risk evaluation, which influences both the profitability of insurers and the affordability for policyholders. Traditionally, insurance pricing has relied on actuarial techniques, utilizing statistical methods and historical data to estimate risk and to set premiums. While these methods are robust and well-established, they have their own limitations in capturing the dynamic nature of modern insurance data.

As the evolution of Machine Learning (ML) algorithms, it gives ways in predictive modeling for enhancing accuracy and efficiency. However, the integration of ML into insurance pricing presents the challenges such as data interpretation complexities, high dimensionality, heterogeneity, and regulatory compliance as well. According to *the Actuary Magazine*, "the application of actuarial big data is limited to traditional actuarial model construction and includes various machine learning algorithms, mapping knowledge domains and other modeling methods" (Le, 2023). It is essential to select suitable modeling techniques based on the scenario's objectives and to optimize these models through a series of test indicator and visual evaluations. In this study, five distinct methodologies have been selected to achieve models with the most accurate results.

The objective of this study is to conduct a comparative analysis of actuarial techniques and machine learning algorithms to identify the most effective models for insurance pricing. This

analysis is important, not only for the accuracy of the predictions, but also for the interpretability of the models, computational efficiency. By examining these dimensions, the research aims to provide a comprehensive evaluation of how these methodologies preform in practical situations.

The findings of this study hold significant implications for insurance companies seeking to optimize their pricing strategies in an increasingly competitive market. Enhanced predictive accuracy can lead to more precise premium setting, better risk management, and ultimately, improved profitability. Moreover, this research contributes to the existing body of knowledge by identifying gaps in the current literature and proposing potential directions for future research. Through rigorous analysis and evaluation, it aims to provide actionable insights that can help insurers adopt the most effective predictive modeling approaches, thereby enhancing their ability to manage risk and achieve financial stability.

2. Literature Review

This chapter introduces the evolution of predictive modeling in insurance pricing, contrasting traditional actuarial techniques with contemporary machine learning (ML) approaches. By exploring the strengths and limitations of both methodologies, this review establishes a critical foundation for understanding their application and efficacy in modern insurance pricing strategies.

2.1 Traditional Actuarial Techniques

In the insurance pricing, the most commonly used traditional actuarial techniques are Generalized Linear Models (GLMs) and Credibility Theory. GLMs extend the capabilities of linear regression models to accommodate a continuous response variable given continuous and/or categorical predictors. The general form is

$$y_i \sim N(x_i^T \beta, \sigma^2),$$

where x_i contains known covariates and β contains the coefficients to be estimated (PennState, 2024). To enhance the model's flexibility, the response variable y_i is assumed to follow an exponential family distribution, such as Poisson for count data and Gamma for continuous positive data. This framework enables the inclusion of multiple predictor variables and accounts for the inherent distribution of the response variable, making it robust for predicting insurance claims. The adaptability of GLMs is essential for handling the diverse and complex nature of insurance data.

On the other hand, Credibility Theory offers a distinct approach by integrating individual risk experience with aggregate data to adjust premiums. According to "Credibility Methods for Individual Life Insurance", the main idea can be formulated as follows:

Let $X = (X_1, X_2, \dots, X_n)$ denote independent losses, which X_j be the annual loss amount from policyholder j

$$\zeta = E(X_j),$$

$$\sigma^2 = var(X_j)$$

$$S = \sum_{j=1}^{n} X_j,$$

$$\bar{X} = \frac{S}{n}$$

Then,

$$E(S) = n\zeta,$$

$$E(X) = \zeta,$$

$$var(S) = n\sigma^2,$$

$$var(X) = \frac{\sigma^2}{n}$$

Credibility Estimate is $P = Z\bar{X} + (1 - Z)M$, which M be the estimate of the mean for this group, Z be credibility factor (Maxwell Gong ,Zhuangdi Li). This technique balances the trade-off between stability and responsiveness in insurance pricing by assigning a credibility factor that quantifies the degree to which individual experience should influence the overall premium. Despite their proven effectiveness, both GLMs and Credibility Theory are constrained by their reliance on assumptions of linearity, independence, and homoscedasticity. These assumptions may limit their adaptability to more complex, non-linear relationships in the data.

2.2 Evolution of Machine Learning in insurance

Compared to traditional methods, the integration of machine learning (ML) has catalyzed significant advancements in insurance pricing methodologies. ML techniques such as

Decision Trees, Random Forests, Gradient Boosting Machines, and Neural Networks have significantly enhanced the ability to identify patterns, assess risk, and improve pricing accuracy. These methods excel in handling large, complex datasets and capturing non-linear relationships, which are often prevalent in insurance data.

Decision Tree learning is a widely used machine learning technique known for its robustness to noisy data and its capability to learn disjunctive expressions. This method approximates discrete-valued functions by recursively partitioning the data into subsets based on feature values, thereby facilitating efficient algorithms to identify an optimal tree structure (M.Mirchell, 1997). A fundamental algorithm underpinning many contemporary decisions tree induction methods is Hunt's algorithm, which constructs the tree in a recursive manner by partitioning the training records into increasingly homogeneous subsets (Kumar, 2006).

Let D_t be the set of training records associated with node t and

Let $y = \{y_1, y_2, \dots, y_c\}$ be the class labels.

Hunt's algorithm operates as follows:

Step 1 (Homogeneous Node Condition):

If all the records in D_t belong to the same class y_t, then t is designed as a lead node and is labeled as y_t.

Step 2 (Heterogeneous Node Condition):

If D_t contains records from multiple classes, an attribute test condition is selected to partition the records into smaller, more homogeneous subsets. For each outcome of the test condition, a child node is created, and the records in D_t are distributed among the children based on these outcomes. The algorithm is then recursively applied to each child node.

This approach, detailed by Kumar (2006), forms the foundational principle of Hunt's algorithm in decision tree learning, and it will be employed in the subsequent case study to illustrate its application and effectiveness.

Random Forests address the limitations of individual trees by building multiple decision trees and aggregating their predictions, thereby enhancing model stability and predictive accuracy. Leo Breiman and Adele Cutler extended the algorithm by integrating the "bagging" method with the random selection of features. Unlike single decision tree, Random Forests do not require cross-validation or a separate test to obtain an unbiased estimate of the test error. Each tree in the forest is trained on a bootstrap sample of the original data, with approximately one-third of the cases left out of the sample. These out-of-bag cases, which are not included in the bootstrap sample, provide an unbiased estimate of the classification error as more trees are added to the forest. Once each tree is built, all data are passed through the tree to compute proximities for each pair of cases. Proximities are increased by one for cases that occupy the same terminal node. At the end of the run, these proximities are normalized by dividing by the number of trees (Leo Breiman, Adele Cutler, 2002). This method enhances the robustness and accuracy of the model by leveraging multiple perspectives from various decision trees. The principal advantage of Random Forests over Decision Trees lies in their ability to reduce overfitting and improve predictive performance by averaging multiple trees trained on different subsets of the data.

Gradient Boosting Machines (GBMs) advance the concept of ensemble learning by iteratively combining weak learners (e.g. Decision Trees) to construct a robust predictive model. The fundamental principle of boosting involves sequentially adding new models to the ensemble. During each iteration, a new weak learner is trained with a focus on the errors of the ensemble accumulated thus far. This iterative process allows GBMs to progressively

improve the model's accuracy by addressing the errors in previous iterations. The basic algorithm of GBMs was derived by Freidman (2001).

Let the dataset $(x, y)_{i=1}^{N}$, where $x = (x_1, \dots, x_d)$ be the explanatory variables, y be the response variable. The goal is to reconstruct the unknown functional dependence $x \xrightarrow{f} y$ with the estimate $\widehat{f(x)}$ using a loss function $\psi(y, f)$

$$\widehat{f(x)} = arg \min_{f(x)} \psi(y, f(x))$$

Friedman's Gradient Boost Algorithm is as follows:

Step 1: Initialize $\widehat{f_0}$ with a constant

Step 2: **for** t=1 to number of iterations M:

1. Compute the negative gradient $g_t(x)$
2. Fit a new base-learner function $h(x, \theta_t)$
3. Find the best gradient descent step-size ρ_t

$$\rho_t = arg \min_{\rho} \sum_{i=1}^{N} \psi[y_i, \widehat{f_{t-1}}(x_i) + \rho h(x_i, \theta_t)]$$

4. Update the function estimate

$$\widehat{f_t} \leftarrow \widehat{f_{t-1}} + \rho_t h(x_i, \theta_t)$$

Step 3: end **for**

This algorithm (Alexey Natekin, Alois Knoll, 2013) optimizes the correlation between the overall error and the new base learner at each iteration. Its principal advantage is its iterative approach, which focuses on correcting errors immediately to enhance predictive accuracy.

This algorithm will be used in the subsequent case study to illustrate its efficacy in reducing discrepancies.

2.3 Potential Limitations between Traditional and Modern Techniques

While both traditional actuarial techniques and machine learning algorithms provide valuable tools for predictive modeling in insurance pricing, each has inherent strengths and limitations. Actuarial methods like GLMs and Credibility Theory are grounded in well-established statistical principles, making them highly interpretable and relatively easy to implement. However, their reliance on predefined assumptions and linear relationships can constrain their ability to adapt to new and emerging patterns in data. These methods are also limited by their dependence on historical data, which may not fully capture the dynamic and evolving nature of risk factors in the insurance domain.

In contrast, while ML algorithms are good at handling the complex and large dataset and capturing non-linear patterns, they have issues with its interpretability and transparency which posing challenges for regulatory compliance. Additionally, the implementation and maintenance of ML models demands substantial computational resources and expertise background, which can be a big challenge for some insurance companies. These potential shortcomings also show the necessity for further research that not only applies within different kinds of datasets but also addresses the practicalities of model deployment in the insurance industry, includes balancing accuracy, interpretability, and operational feasibility in selecting appropriate predictive modeling techniques for insurance pricing. By providing a comprehensive comparative analysis of traditional actuarial methods and contemporary

machine learning algorithms, this research aims to fill existing gaps and contribute to a more complete understanding of their respective strengths and weaknesses.

3. Data Description

3.1 Description of Data Collection Process and Data Sources

The dataset utilized in this study is sourced from an insurance company's comprehensive historical database. This dataset includes detailed records pertaining to insurance policies, encompassing a wide array of variables essential for predictive modeling. These records offer valuable insights into policyholder demographics, claims history, and financial aspects, which are crucial for accurate insurance pricing. The original dataset comprises 30 columns, providing general information such as age, gender, occupation, income, and location, as well as professional insurance information like coverage amount, premium amount, deductible, policy type, and customer preferences among 13,000 individuals.

For a more focused comparative analysis, the dataset has been distilled into seven key variables, which will form the basis for subsequent analysis. The initial exploratory data analysis (EDA) involved data cleaning and preprocessing steps, such as checking for missing values and converting categorical variables to factors. This streamlined dataset will facilitate the evaluation and comparison of different predictive modeling techniques.

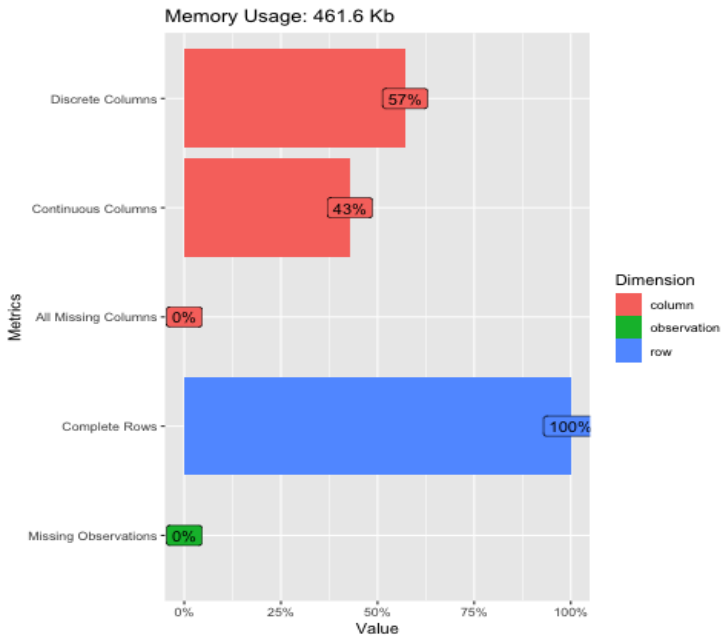

Figure 1: Data Mining and Data Processing

3.2 Explanation of the Dataset Used for Analysis

The dataset consists of historical insurance policy data, providing a robust foundation for predictive modeling. This curated dataset captures essential aspects of policyholder profiles and their claims history, offering valuable insights for analyzing and comparing the performance of traditional actuarial techniques and machine learning algorithms. The dataset includes the following seven key variables:

1. Claim Amount: The claim amount is a critical variable for analyzing claims history and predicting future claims, serving as the primary response variable in the predictive models.
2. Age: The age of the policyholder is a critical factor in risk assessment and premium calculation, as different age groups exhibit varying risk profiles.

3. Income: The income level of the policyholder provides insights into their economic status and potential risk exposure, impacting their insurance needs and premium affordability.

The following two figures present the histogram of these three numeric variables and their corresponding correlation matrix. In the correlation matrix, a value close to 1indicates a strong positive relationship, while a value close to -1 indicates a strong negative relationship. It concludes that the relationship of three numeric variables tend to be neutral. A value near 0 signifies a neutral or weak relationship. The analysis of the correlation matrix reveals that the relationships among the three numeric variables tend to be neutral.

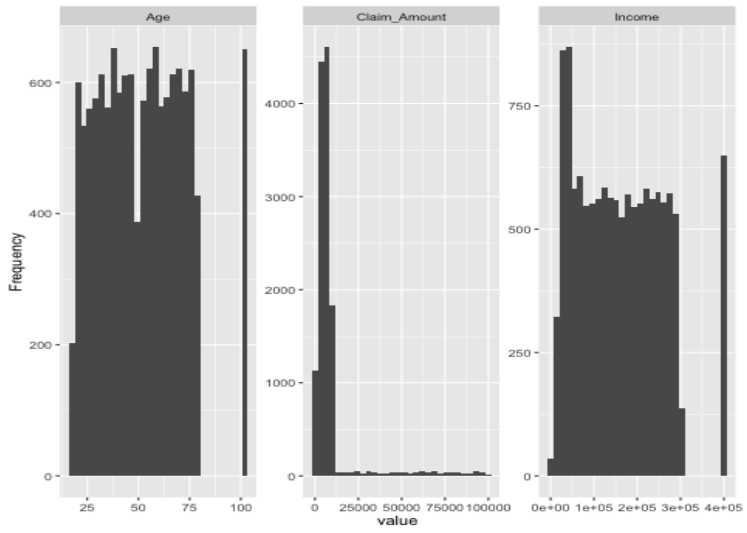

Figure 2 the Histogram of Three Numeric Variables (Age, Income, Claim Amount)

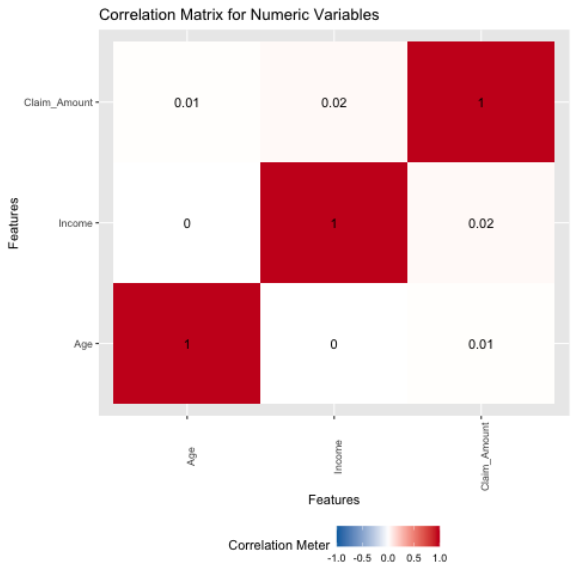

Figure 3 Correlation Matrix for Three Numeric Variables

4. Gender: Gender influences risk profiles and insurance pricing, with statistical differences observed in claims frequency and severity between male and female policyholders.
5. Marital Status: Marital status affects risk assessment, as married individuals might exhibit different risk behaviors compared to single individuals, influencing their insurance requirements.
6. Education: Education level correlates with risk behaviors and insurance requirements, as higher education levels often associate with lower risk profiles.
7. Occupation: The policyholder's occupation affects risk assessment based on occupational hazards and income stability, playing a significant role in determining premium rates.

The following two figures illustrate the histograms of four categorical variables and their corresponding correlation matrix. In the correlation matrix, dark blue boxes indicate strong negative relationships, while light blue boxes represent negative relationships between the variables.

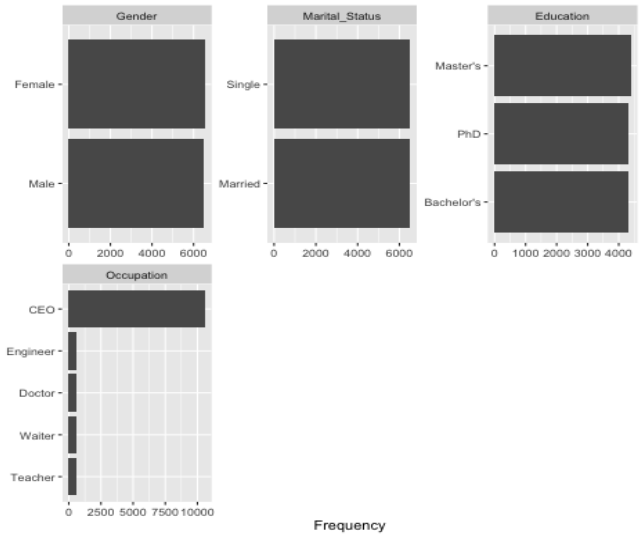

Figure 4 the Histograms of Four Categorical Variables (Gender, Education, Occupation, Marital Status)

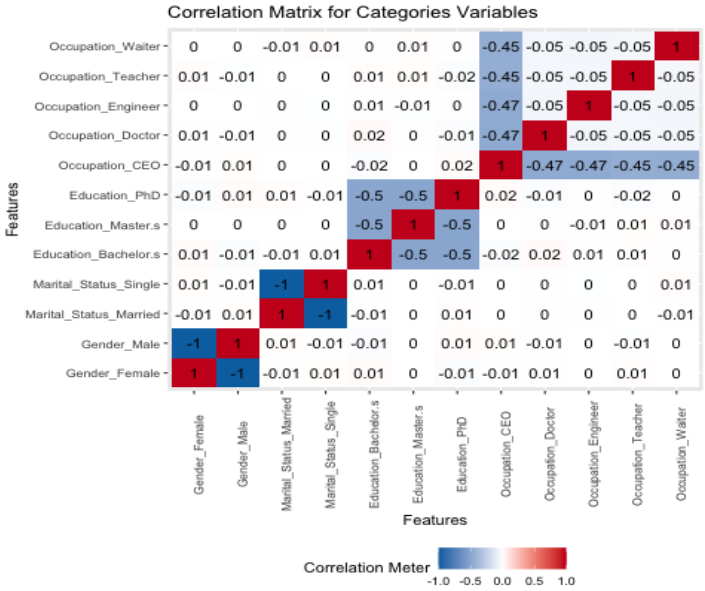

Figure 5 Correlation Matrix for Four Categories Variables

3.3 Evaluation Metrics

In the comparative analysis of predictive modeling techniques for insurance pricing, the efficacy of both traditional actuarial methods and modern machine learning algorithms will be evaluated using a set of robust evaluation metrics. These metrics (Williams, 2019) serve as objective measures to assess the performance and predictive accuracy of each approach.

1. Deviance: Deviance measures the goodness-of-fit of GLMs by comparing the fitted model to a saturated model that perfectly predicts the response variable. Lower deviance values indicate better model fit, as they represent smaller discrepancies between observed and predicted values.

$$Deviance = 2 \cdot (loglikelihood\ of\ saturated\ model - loglikelihood\ of\ fitted\ model)$$

2. Akaike Information Criterion (AIC): AIC provides a measure of the relative quality of statistical models for a given set of data. It balances the goodness-of-fit of the model with its complexity, penalizing models with excessive parameters.

$$AIC = -2 \cdot loglikelihood + 2 \cdot (\#\ of\ parameters\ in\ Model)$$

3. Bayesian Information Criterion (BIC): Similar to AIC, BIC also assesses model fit and complexity, with a preference for simpler models. It penalizes model complexity more heavily than AIC, often resulting in the selection of a more conservative model.

$$BIC = -2 \cdot loglikelihood + (\#\ of\ parameters\ in\ Model) \cdot \log(\#\ of\ Recoeds\ in\ Dataset)$$

4. Mean Absolute Error (MAE): MAE measures the average magnitude of errors in predictions, providing a straightforward assessment of predictive accuracy. It is calculated as the average absolute distance between predicted and actual values.

$$MAE = \frac{1}{N}\sum_{j=1}^{N}|y_j - \hat{y}_j|$$

5. Root Mean Squared Error (RMSE): RMSE quantifies the average magnitude of prediction errors, placing greater emphasis on larger errors due to its quadratic nature. It is calculated as the square root of the average squared differences between predicted and actual values.

$$RMSE = \sqrt{\frac{\sum_{j=1}^{N}\left(y_j - \hat{y_j}\right)^2}{N}}$$

6. R-squared (the coefficient of determination): R-squared is used to evaluate the performance of linear regression model. It indicates the proportion of variance in the dependent variable that is predictable from the independent variables. It reflects the model's explanatory power, with higher values indicating better model fit.

These evaluation metrics collectively provide a comprehensive framework for assessing the performance of predictive models. By employing these metrics, the study aims to identify the most effective techniques for insurance pricing, considering both accuracy and model complexity. The insights gained from this analysis will inform the selection and implementation of predictive modeling approaches in the insurance industry.

4. Methodology

The predictive modeling approach in this study involves developing separate models using traditional actuarial techniques and modern machine learning algorithms. Actuarial techniques are well-established in the insurance industry due to their interpretability and robustness, whereas ML algorithms are recognized for their superior accuracy and the ability to handle complex and large datasets. This dual approach aims to provide a comprehensive analysis of the strengths and weaknesses of each methodology in the context of insurance pricing.

4.1 Generalized Linear Models (GLMs)

Generalized Linear Models (GLMs) are fundamental in actuarial science, particularly for predicting insurance claims and pricing. Their adaptability to various types of response distributions, especially those from the exponential family, makes GLMs highly valuable. In this study, a GLM is employed to model the relationship between several predictor variables (Age, Gender, Income, Education, Occupation, and Marital Status) and the insurance claim amounts.

The analysis begins by splitting the dataset into training and test sets to facilitate model validation and performance evaluation. The training set is used to fit the GLM, while the test set is reserved for assessing the model's predictive performance. Then, the GLM is specified with a Gaussian distribution and a log-link function to address the positive skewness observed in the Claim Amounts. This transformation stabilizes the variance and makes the distribution more symmetric, thereby improving model performance. The model formula is expressed as:

$$\log(ClaimAmount) = \beta_0 + \beta_1 \cdot Age + \beta_2 \cdot Gender + \beta_3 \cdot Income + \beta_4 \cdot Education + \beta_5 \cdot Occupation + \beta_6 \cdot MaritalStatus$$

```
##
## Call:
## glm(formula = Claim_Amount ~ Age + Gender + Income + Education +
##     Occupation + Marital_Status, family = gaussian(link = "log"),
##     data = trainData)
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           8.463e+00  4.967e-02 170.371   <2e-16 ***
## Age                   2.063e-04  5.719e-04   0.361   0.7184
## GenderMale           -1.159e-02  2.390e-02  -0.485   0.6276
## Income                1.464e-06  1.030e-07  14.213   <2e-16 ***
## EducationMaster's     4.740e-02  2.832e-02   1.674   0.0942 .
## EducationPhD         -5.179e-02  3.018e-02  -1.716   0.0862 .
## OccupationDoctor      1.376e+00  3.574e-02  38.506   <2e-16 ***
## OccupationEngineer    1.312e+00  3.622e-02  36.236   <2e-16 ***
## OccupationTeacher     1.278e+00  3.842e-02  33.273   <2e-16 ***
## OccupationWaiter      1.357e+00  3.641e-02  37.281   <2e-16 ***
## Marital_StatusSingle  4.161e-03  2.387e-02   0.174   0.8616
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 189327735)
##
##     Null deviance: 2.4064e+12  on 10399  degrees of freedom
## Residual deviance: 1.9669e+12  on 10389  degrees of freedom
## AIC: 227740
##
## Number of Fisher Scoring iterations: 8
```

Figure 6 the summary() Output of Generalized Linear Model

Upon fitting the GLM to the training data, the summary output provides insights into the significance of each predictor variable. The summary results typically include coefficients, standard errors, z-values, and p-values for each predictor. The Significant variables (at the 0.05 level) include Income, and specific Occupation (Doctor, Engineer, Teacher, and Waiter). These results suggest that Income and certain Occupations strongly influence the Claim Amount. Variables such as Age, Gender, Education, and Marital Status do not show significant effects at the conventional 0.05 level, although Education (Master's, PhD) is borderline significant. The intercept, representing the baseline log of the claim amount, shows a significant positive effect. This underscores the significance of Income and certain occupational categories as predictors of insurance claims.

The model's performance is summarized by the following metrics: the null deviance represents the fit of a model with only the intercept, while the residual deviance represents the fit of the specified model. In this case, the null deviance is 2.4064e+12 with 10399 degrees of freedom and the residual deviance is the deviance of the current model which is 1.9669e+12. Lower deviance values indicate better model fit, as the discrepancy between observed and predicted values is reduced. Additional performance metrics include the Root Mean Square Error (RMSE), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). The RMSE of 13104.4 on the test set indicates the magnitude of prediction errors, with lower values implying better predictive accuracy. AIC and BIC, with values of 227740.3 and 227827.3 respectively, serve as the measures of model fit, penalizing for model complexity. Lower values of AIC and BIC indicate a more parsimonious model that balances fit and complexity. While the model exhibits reasonable fit, there is potential for improvement. Future work could explore alternative modeling approaches, including machine learning algorithms, to enhance predictive accuracy and compare their performance against traditional GLMs.

4.2 Credibility Theory

Credibility theory is another fundamental actuarial tool used to refine premium estimates by integrating individual risk data with aggregate data. This approach balances the specific experience of policyholders with the overall portfolio experience, thereby adjusting premiums based on the credibility assigned to individual risks. In this study, the Bühlmann-Straub credibility model is applied to handle heterogeneous risk groups.

Bühlmann’s method for updating the predicted loss measure is based on a linear predictor that uses past observations. The Bühlmann-Straub model calculates credibility-adjusted

premiums by incorporating both the group-specific and overall means and variances of claim amounts (Tse, 2009). Initially, the data is grouped by a categorical variable, such as Occupation, to identify different risk groups. For each group, the mean and variance of the claim amounts are calculated, as well as the overall mean and variance across all groups. The model's primary objective is to derive a credibility factor (Z) for individual risks, effectively balancing specific data with collective data. The factor Z is derived as follows:

$$Z = \frac{n}{n + \frac{\sigma_b^2}{\sigma_w^2}}$$

where,

- n is the number of claims for an individual policyholder
- σ_b^2 is the variance between different groups
- σ_w^2 is the variance within groups

Using the formula, the credibility factor for each group is computed. subsequently, the credibility-adjusted premium for a policyholder is calculated as:

$$\hat{P} = Z \cdot Sample\ Mean + (1 - Z) \cdot Overall\ Mean$$

where the individual mean is the mean claim amount for the policyholder's group, and the overall mean is the mean claim amount across all groups.

The model's performance is evaluated using mean squared error (MSE), which quantifies the discrepancy between predicted premiums and actual claim amounts. The high MSE has the value of 3.54×10^8 suggests significant discrepancies between the predicted premiums and the actual claim amounts. Several factors may contribute to this result. The Bühlmann-Straub

model assumes constant variance within each group and between groups. If this assumption does not hold, the model's accuracy may be compromised. Additionally, this model may not fully capture the complexity and variability inherent in large insurance datasets. The linear adjustment method used in the Bühlmann-Straub model might oversimplify the relationships within the data. Non-linear relationships, common in insurance data, are not captured by this approach, leading to potential inaccuracies. Overall, these findings suggest a need for more sophisticated modeling approaches that can handle the complexity and non-linearity inherent in insurance data.

4.3 Decision Trees

Decisions trees are a powerful non-parametric supervised learning method used for classification and regression. This model operates by recursively splitting the dataset into subsets based on feature values, creating a tree-like model of decisions. Each split aims to maximize the homogeneity of the resulting subsets, thereby improving the predictive power of the model. Decision trees derive simple decision rules from the data features, making them effective in predicting the target variable's value.

In this study, the decision tree model below is utilized to predict insurance claim amounts using various policyholder attributes such as Age, Gender, Income, Education, Occupation, and Marital Status. The root node of the decision tree represents the entire dataset, initially showing an overall claim frequency of 9.1% and an exposure proportion of 100%. This indicates that, on average, 9.1% of the policies result in claims, and all data points are considered at this level.

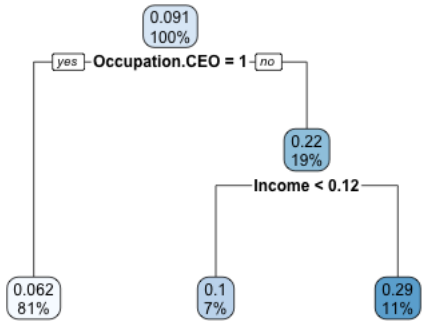

Figure 7 Decision Tree Model for Insurance data

The first split occurs based on the feature "Occupation.CEO=1," dividing the dataset into two groups. The left Node (Occupation = CEO) captures 81% of the policies, with a claim frequency of 6.2%. The right Node (Occupation ≠ CEO) shows the remaining 19% of policies, which have a higher claim frequency and undergo further splits based on additional features. One such feature is "Income < 0.12," where the second split increases the claim frequency from 22% to 29% for the subset of policies with income below 0.12. This splitting process continues, dividing the data into increasingly homogeneous groups until a stopping criterion is met, such as a maximum tree depth, minimum number of samples per leaf, or a threshold on impurity reduction.

The performance of the Decision Tree model is evaluated using three key metrics. The MAE of 5831.846 indicates, on average, the model's predictions deviate from the actual claim amounts by approximately $5831.85. Lower MAE values suggest better predictive accuracy, implying that the model's predictions are relatively close to the actual claim amounts. The MSE value of 1.60×10^8 reflects the average squared differences between the predicted and

actual claim amounts. Since MSE penalizes larger errors more heavily than MAE, the relatively high value suggests substantial discrepancies and variability in the model's predictions. Lower MSE values indicate a better fit of the model to the data, highlighting the need for potential model improvements or adjustments. R-squared value of 0.2221111 indicates that approximately 22.21% of the variance in the claim amounts is explained by the model. R-squared values range from 0 to 1, with higher values signifying a better fit of the model to the data. In this case, the R-squared value suggests that the model explains a relatively low proportion of the variance in the claim amounts. While the model offers a clear and interpretable structure for understanding the decision-making process, the high MAE and MSE values indicate substantial deviations from the actual claim amounts. Additionally, the low R-squared value highlights the need for more sophisticated models or further tuning to better capture the complexity of the data.

4.4 Random Forests

Random Forest is an ensemble learning technique that enhances predictive accuracy by constructing multiple decision trees and aggregating their results. This method is particularly advantageous for handling high-dimensional data and capturing complex feature interactions. By averaging the predictions of individual trees, Random Forest mitigates the overfitting tendencies commonly observed in single decision trees, leading to more robust and stable predictions.

A variable importance Plot from a random forest model tells the features that contribute most significantly to the model's predictive power. X-axis represents the important score, which can be measured by the mean decrease in accuracy or the mean decrease in impurity (Gini Importance). Y-axis lists the features in descending order of importance. Each point on the

plot represents the importance score of a feature. For instance, the highest points (Occupation and Income) indicate a higher importance for predicting claim amounts.

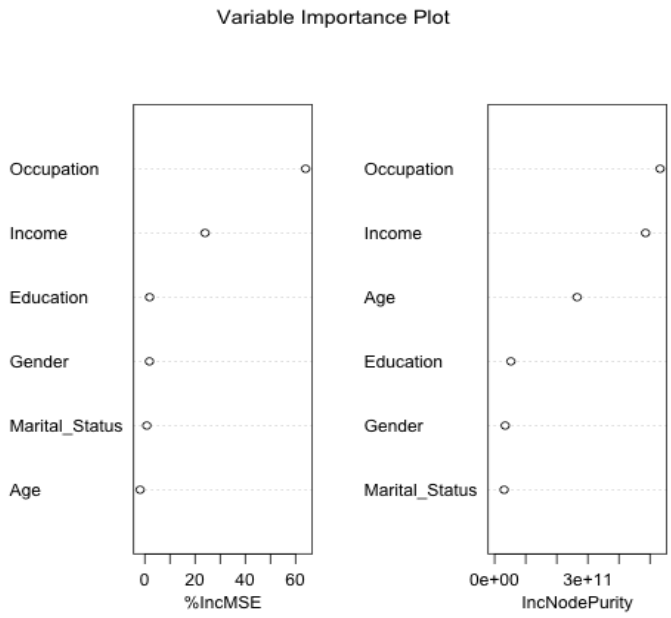

Figure 8 Variance Importance Plot by Random Forests

The evaluation metrics for the Random Forest model reveal an MAE of 5955.522, indicating the average absolute difference between predicted and actual claim amounts. The MSE of 1.61×10^8 shows a similar level of variability as observed with the Decision Tree model. The R-squared of 0.214387 suggests that 21.44% of the variance in claim amounts is explained by the model, which is slightly lower than that of the Decision Tree Model. The Random Forest model demonstrates competitive performance in predicting insurance claim amounts. While it exhibits slightly higher MAE and MSE values compared to the Decision Tree model, it provides a marginally higher R-squared value, indicating a slightly better fit to the data. The Random Forest model offers several advantages over single decision trees. Specially, by averaging the predictions of multiple trees, Random Forest reduces the overfitting tendencies of individual decision trees, resulting in more stable and generalizable

predictions. Additionally, the algorithm is well-suited for datasets with many features, as the random feature selection at each split helps to capture complex interactions between variables.

4.5 Gradient Boosting

Gradient Boosting is an advanced ensemble learning technique that combines multiple weak learners (e.g. decision trees) to create a strong predictive model. Unlike traditional ensemble methods that build trees independently, Gradient Boosting constructs trees sequentially, where each new tree focuses on correcting the errors made by the previous ones (Natekin). This iterative process allows the model to progressively improve its performance, making it highly effective for complex, high-dimensional datasets.

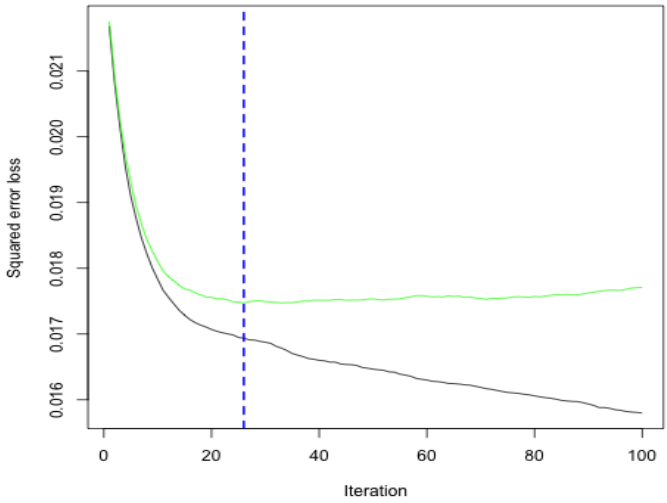

Figure 9 Out-of-Bag error Plot by Gradient Boosting Algorithm

This graph above illustrates the evolution of performance metrics as the Gradient Boosting algorithm incorporates a progressively larger number of base learners. In this classification task, the black line is the training Bernoulli Deviance, while the green line is the testing Bernoulli Deviance. Smaller deviance values correspond to better performance. The

algorithm employs cross-validation ("Method=cv") to evaluate the performance of the ensemble of learners. The blue dashed line indicates the optimal number of iterations according to the chosen metric and validation procedure, which is informative for future modeling.

In the context of insurance pricing, Gradient Boosting captures the nuanced interactions between policyholder characteristics (such as Age, Gender, Income, Education, Occupation, and Marital Status) and the resultant claim amounts. This allows insurance companies to set premiums more precisely and manage risk more effectively. Gradient Boosting model demonstrates an MAE value of 5851.337, indicating the average deviation of predictions from actual claim amounts. The MSE of 1.59×10^8 reflects the average squared differences. An R-squared value of 0.2231369 suggests that approximately 22.31% of the variance in claim amounts is explained by the model. The Gradient Boosting model exhibits a robust capability for predicting insurance claim amounts by iteratively refining its predictions to minimize errors. Its ability to capture complex interactions among variables and improve accuracy through successive iterations makes it a powerful tool for enhancing predictive performance in insurance pricing.

5. Comparative Analysis

When evaluating the five predictive methods based on their performance metrics, Gradient Boosting emerges as the most effective model for insurance pricing. Despite its relatively high MSE, Gradient Boosting's MAE and R-squared values indicate a strong ability to capture the variance in claim amounts and minimize prediction errors. This model's iterative refinement process, which focuses on poorly predicted instances, enables it to effectively capture complex relationships within the insurance data.

GLMs provide a solid baseline with interpretable results and a reasonable fit. However, they fall short in predictive accuracy compared to ensemble methods. The high RMSE of 13104.4 indicates substantial prediction errors, suggesting that while GLMs can identify significant predictors such as Income and specific Occupations, they struggle with the dataset's complexity. Credibility Theory, particularly through the application of the Bühlmann-Straub model, shows significant discrepancies between predicted and actual claim amounts, as evidenced by a high MSE of 3.54×10^8. The model has a very high MSE because some selected variables contain outliers in the large dataset, or important features that have a strong relationship with the target variable are missing. The model's assumptions of constant variance within and between groups do not hold well for the complex and variable nature of the insurance data used in this study. Consequently, both GLMs and Credibility Theory fall short in providing accurate predictive modeling for insurance pricing in this scenario. While valuable in certain contexts, these methods do not effectively capture the intricate patterns and relationships present in the insurance claims data analyzed here.

The comparative analysis of Decision Trees, Random Forests, and Gradient Boosting models reveals distinct strengths and weaknesses in their predictive capabilities for insurance pricing.

Decision Trees provide a straightforward approach with moderate accuracy, as evidenced by their MAE and MSE values. However, the model's relatively low R-squared value indicates a limited ability to explain the variance in claim amounts.

Random Forests improve upon Decision Trees by reducing overfitting and offering more stable predictions. Despite slightly higher MAE and MSE values compared to Decision Trees, the Random Forest model's marginally higher R-squared value suggests a better fit to the data. This indicates that Random Forests can capture more complex interactions among variables, though not as effectively as Gradient Boosting. Gradient Boosting stands out as the most effective model among the three, exhibiting the lowest MSE and the highest R-squared value. This suggests superior accuracy and explanatory power. Gradient Boosting's iterative process, which focuses on correcting the errors of its predecessors, allows it to capture intricate relationships in the data, and to make it suitable for the complex nature of insurance pricing.

In conclusion, while all three machine learning models demonstrate predictive capability, Gradient Boosting emerges as the most robust and accurate method for predicting insurance claim amounts in this case. This analysis underscores the potential of machine learning algorithms in enhancing insurance pricing strategies by leveraging the complex interactions inherent in policyholder data. The limitations of GLMs and Credibility Theory highlight the importance of exploring advanced machine learning techniques to achieve higher predictive accuracy in insurance pricing.

6. Discussion

This study highlights the potential of machine learning algorithms, particularly Gradient Boosting, in enhancing the accuracy and reliability of insurance pricing models. By addressing the limitations of traditional methods and exploring advanced techniques, future research can further improve predictive capabilities, ultimately benefiting both insurers and policyholders through more precise and fair pricing strategies. The findings from this study open several ideas for future research in the field of insurance pricing:

6.1 Exploring Advanced Ensemble Methods

In this study, the superior predictive performance of machine learning (ML) techniques, particularly Gradient Boosting, has been established in the domain of insurance prediction. It is prudent to explore more advanced variants within the Gradient Boosting framework, such as Extreme Gradient Boosting (XGBoost). XGBoost is a decision tree ensemble method rooted in gradient boosting principles. It iteratively builds an additive model to minimize a loss function, similar to gradient boosting. Moreover, XGBoost integrates randomization techniques to reduce overfitting and enhance training speed. These techniques include random subsampling for training individual trees and column subsampling at tree and tree node levels (Bentéjac, 2019). By leveraging these advanced techniques, XGBoost could offer deeper insights into the intricate relationships within insurance data and potentially improve predictive accuracy.

6.2 Hyperparameter Optimization of Machine Learning Algorithm

While the present study provides an initial comparative analysis of various machine learning models, future research should prioritize fine-tuning hyperparameters for improved predictive

performance. With only seven variables considered here, constructing an effective machine learning model is a complex and time-consuming process. It entails selecting the appropriate algorithm and optimizing the model architecture through hyperparameter tuning. As more insurance variables are included, the complexity of this task increases. Techniques such as grid search or Bayesian optimization offer systematic approaches to identify optimal parameters for each model, potentially enhancing predictive accuracy (Li Yang, Abdallah Shami, 2020). These methods enable an exhaustive exploration of hyperparameter combinations, ensuring the selection of parameters that maximize model performance. By focusing on hyperparameter optimization, future studies can expedite model development and fully exploit predictive capabilities of machine learning algorithms in insurance prediction tasks.

6.3 Ethical and Regulatory

As certain aspects of Artificial Intelligence (AI) advances ahead of insurance practices, future research must address the ethical and regulatory implications of using advanced predictive models in insurance pricing. Adherence to the guidelines outlined in "Regulatory of Artificial Intelligence in Insurance: Balancing Consumer Protection and Innovation" by The Geneva Association is paramount. Ensuring fairness, transparency, and compliance with regulatory standards is essential for maintaining the integrity of the whole industry. This involves mitigating biases in predictive models, guaranteeing that AI-driven decisions are transparent, and upholding rigorous consumer protection standards (Noordhoek, 2023).

Bibliography

Alexey Natekin, Alois Knoll. (2013, December 4). Gradient boosting machines, a tutorial. p. doi: 10.3389/fnbot.2013.00021.

Bentéjac, C. (2019). A Comparative Analysis of XGBoost. doi.org/10.48550/arXiv.1911.01914 .

Kumar, V. (2006). *Introduction to Data Mining: Chapter 4 Classification: Basic Concepts, Decision Trees, and Model Evaluation.* ISBN:03214220527.

Le, H. (2023). Application and Practice of Actuarial Big Data: Part 2. *The Actuary Magazine*.

Leo Breiman, Adele Cutler. (2002). *Random Forests*. Retrieved from https://www.stat.berkeley.edu/~breiman/RandomForests/cc_copyright.htm

Li Yang, Abdallah Shami. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. doi.org/10.1016/j.neucom.2020.07.061.

M.Mirchell, T. (1997). *Machine Learning: Chapter 3 Decision Tree Learning.* ISBN:007428077.

Maxwell Gong ,Zhuangdi Li. (2018, December 11). Credibility Methods for Individual Life Insurance. pp. Risks 2018, 6(4), 144; doi.org/10.3390/risks6040144.

Noordhoek, D. (2023). *REGULATION OF ARTIFICIAL INTELLIGENCE IN INSURANCE: Balancing consumer protection and innovation.* . The Geneva Association. .

PennState. (2024). *STAT 504 Analysis of Discrete Data 6.1 - Introduction to GLMs*. Retrieved from https://online.stat.psu.edu/stat504/lesson/6/6.1

Tse, Y.-K. (2009). *Nonlife Actuarial Models Theory, Methods and Evaluation: 7 - Bühlmann credibility.* doi.org/10.1017/CBO9780511812156.012: Cambridge University Press.

Williams, B. (2019). How to Pick a Better Model . *casact.org*.

Yikai (Maxwell) Gong ,Zhuangdi Li . (2018). *Credibility Methods for Individual Life Insurance.*