Revista Brasileira de
Risco e Seguro

# *Actuarial Learning* e GLM: Comparação de Ajustes para Seguros Residenciais

**Ludmila de Melo Souza**
Departamento de Ciências Contábeis e Atuariais (CCA), Universidade de Brasília (UnB).
ludmilamelo@unb.br
https://orcid.org/0000-0003-2340-2929

**Eduardo Fraga L. de Melo**
Instituto de Matemática e Estatística (IME), Universidade do Estado do Rio de Janeiro (UERJ). Escola de Matemática Aplicada – EMAp/FGV.
eduardoflm@yahoo.com.br
https://orcid.org/000-0002-0138-1875

## *Resumo*

A literatura atuarial cada vez mais apresenta metodologias e modelos para a precificação de seguros. Essa evolução tem aprimorado significativamente a eficiência dos procedimentos estatísticos utilizados por atuários para a análise de risco, elevando a qualidade e a exatidão desses processos dentro das corporações seguradoras. Historicamente, o setor de seguros tem se apoiado em Modelos Lineares Generalizados (GLM) para auxílio na precificação de seguros devido à sua flexibilidade e capacidade de modelar a relação entre variáveis preditoras e a resposta de interesse, geralmente a frequência e severidade de sinistros (McCullagh & Nelder, 2019). No entanto, a utilização de técnicas de aprendizado de máquina na precificação e análise de seguros oferece uma abordagem mais flexível e potencialmente mais precisa para identificar padrões complexos e não lineares nos dados, e sua aplicação no contexto atuarial fez com que desenvolvesse um novo campo de estudos, os estudos em *Actuarial Learning* (Tzougas; Kutzkov, 2023). Tal estudo visa preencher a lacuna apontada na literatura sobre a necessidade de expandir o uso de técnicas de *actuarial learning* na ciência atuarial e comparar sua eficácia na precificação com os modelos GLM comumente usados pela área. Para isso, foram utilizados dados reais de apólices de seguro residencial compreensivo no Brasil entre 2019 e 2020. Os resultados demonstram que os métodos de *Actuarial Learning*, especialmente o XGBoost, fornecem previsões substancialmente mais precisas quando comparados ao GLM tradicional. Essa melhoria na precisão pode levar a uma estimativa de risco mais acurada, resultando em precificação mais fidedigna e políticas de subscrição mais eficazes, o que pode, por sua vez, melhorar a lucratividade e a satisfação dos clientes das seguradoras.

## *Palavras-chave*

*Actuarial Learning*. GLM. Seguros.

## *Sumário*

## *Abstract*

**Actuarial Learning and GLM: Comparison of Adjustments for Home Insurance**

**Ludmila de Melo Souza**
Department of Accounting and Actuarial Sciences (CCA), University of Brasília (UnB).
ludmilamelo@unb.br
https://orcid.org/0000-0003-2340-2929

**Eduardo Fraga L. de Melo**
Institute of Mathematics and Statistics (IME), Rio de Janeiro State University (UERJ). School of Applied Mathematics – EMAp/FGV.
eduardoflm@yahoo.com.br
https://orcid.org/000-0002-0138-1875

## *Summary*

The actuarial literature has increasingly presented methodologies and models for use in insurance pricing. This evolution has significantly improved the efficiency of the statistical procedures used by actuaries for risk analysis, raising the quality and accuracy of these processes within insurance corporations. Historically, the insurance sector has relied on Generalized Linear Models (GLM) for pricing due to their flexibility and ability to model the relationship between predictor variables and the response of interest, usually the frequency and severity of claims (McCullagh & Nelder, 2019). However, the use of machine learning techniques in pricing and insurance analysis offers a more flexible and potentially more accurate approach to identify complex non-linear patterns in the data, and its application in the actuarial context has led to the development of a new field of study, Actuarial Learning studies (Tzougas & Kutzkov, 2023). This study aims to fill the gap pointed out in the literature on the need to expand the use of actuarial learning techniques in actuarial science and compare their efficiency in pricing with the GLM models commonly used by the area. To this end, real data from comprehensive residential insurance policies in Brazil between 2019 and 2020 were used. The results demonstrate that Actuarial Learning methods, especially XGBoost, provide substantially more accurate forecasts when compared to traditional GLM. This improvement in precision can lead to a more accurate risk estimate, resulting in fairer pricing and more effective underwriting policies, which can, in turn, improve the profitability and customer satisfaction of insurers.

## *Keywords*

Actuarial Learning. GLM. Insurance.

## *Contents*

## *Sinopsis*

**Aprendizaje Actuarial y GLM: Comparación de ajustes para seguros de hogar**

**Ludmila de Melo Souza**
Departamento de Ciencias Contables y Actuariales (CCA), Universidad de Brasília (UnB).
ludmilamelo@unb.br
https://orcid.org/0000-0003-2340-2929

**Eduardo Fraga L. de Melo**
Instituto de Matemática y Estadística (IME), Universidad del Estado de Río de Janeiro (UERJ). Escuela de Matemática Aplicada – EMAp/FGV.
eduardoflm@yahoo.com.br
https://orcid.org/000-0002-0138-1875

## *Resumen*

La literatura actuarial está presentando cada vez más metodologías y modelos para su uso en la tarificación de seguros. Esta evolución ha mejorado significativamente la eficiencia de los procedimientos estadísticos utilizados por actuarios para el análisis de riesgo, elevando la calidad y la exactitud de estos procesos dentro de las corporaciones aseguradoras. Históricamente, el sector de seguros ha dependido de los Modelos Lineales Generalizados (GLM) para la asistencia en la tarificación de seguros debido a su flexibilidad y capacidad de modelar la relación entre variables predictoras y la respuesta de interés, generalmente la frecuencia y severidad de los siniestros (McCullagh & Nelder, 2019). Sin embargo, el uso de técnicas de aprendizaje automático en la tarificación y análisis de seguros ofrece un enfoque más flexible y potencialmente más preciso para identificar patrones complejos y no lineales en los datos, y su aplicación en el contexto actuarial ha dado lugar al desarrollo de un nuevo campo de estudios, el aprendizaje actuarial (Tzougas & Kutzkov, 2023). Este estudio tiene como objetivo llenar el vacío señalado en la literatura sobre la necesidad de expandir el uso de técnicas de aprendizaje actuarial en la ciencia actuarial y comparar su eficacia en la tarificación con los modelos GLM comúnmente usados por el área. Para ello, se utilizaron datos reales de pólizas de seguro de hogar comprensivo en Brasil entre 2019 y 2020. Los resultados demuestran que los métodos de Aprendizaje Actuarial, especialmente el XGBoost, proporcionan predicciones sustancialmente más precisas en comparación con el GLM tradicional. Esta mejora en la precisión puede llevar a una estimación de riesgo más acurada, resultando en tarificaciones más fidedignas y políticas de suscripción más eficaces, lo que a su vez, puede mejorar la rentabilidad y la satisfacción del cliente de las aseguradoras.

## *Palabras clave*

Aprendizaje Actuarial. GLM. Seguros.

## *Síntesis*

## 1. Introdução

A precificação de seguros visa avaliar com a maior exatidão possível o prêmio puro de cada segurado e é a referência para avaliação interna de risco, pois incorpora grande parte das informações disponíveis dos contratos subscritos. Quando a apólice de seguro é vendida, a seguradora não conhece os custos finais desse serviço, mas confia em dados históricos e modelos atuariais para prever um preço sustentável para o seu produto, como afirmam Denuit; Hainaut e Trufin (2019). Nos últimos anos, o setor de seguros tem enfrentado desafios devido às mudanças nos padrões de risco, aos avanços tecnológicos e às expectativas crescentes dos consumidores. Isso porque a precisão na precificação de seguros não impacta apenas a sustentabilidade financeira das seguradoras, mas também a acessibilidade e a percepção de valor pelos segurados.

A literatura atuarial, cada vez mais, tem apresentado metodologias e modelos para utilização em precificação de seguros. Essa evolução tem aprimorado significativamente a eficiência dos procedimentos estatísticos utilizados por atuários para a análise de risco, elevando a qualidade e a exatidão desses processos dentro das corporações seguradoras. A modelagem atuarial do seguro não vida, também conhecido como seguro geral, engloba uma série de técnicas e enfrenta diversos desafios para precificar riscos de maneira eficaz e garantir a solvência das companhias seguradoras. As técnicas empregadas e os desafios enfrentados são cruciais para entender a complexidade e a importância da modelagem atuarial nesse segmento. Dentre elas, podem ser citadas:

1. **Técnicas de precificação baseadas em experiência**: Estas técnicas são fundamentadas no uso de dados históricos para prever a probabilidade e o custo de sinistros futuros. O trabalho de Bailey e Simon (1960), que explora a aplicação de regressão linear e análise de variância em dados de seguro, ilustra a relevância destes métodos estatísticos no entendimento das tendências e variações em sinistros.

2. **Modelos Lineares Generalizados (GLM)**: A introdução dos GLMs, por Nelder e Wedderburn (1972), revolucionou a modelagem estatística ao permitir o ajuste de distribuições de sinistros que desviam da normalidade, uma característica comum em dados de seguros não vida. Esses modelos são essenciais para modelar a frequência e a severidade dos sinistros, servindo como uma base sólida para a precificação de riscos (MCCULLAGH & NELDER, 2019).

3. **Técnicas de segmentação de riscos**: A capacidade de segmentar riscos é fundamental na construção de uma carteira de seguros equilibrada. Breiman *et al.* (1984) destacam o uso de técnicas como análise de *cluster* e árvores de decisão para identificar grupos homogêneos de riscos, permitindo uma precificação mais direcionada e personalizada.

4. **Modelagem de dependência e riscos catastróficos**: Modelos de dependência, como Cópulas, são ferramentas valiosas no estudo de correlações entre diferentes tipos de riscos. Embrechts, McNeil e Staumann (2002) enfatizam a importância desses modelos na compreensão de eventos de baixa frequência e alta severidade, como desastres, por exemplo.

Historicamente, o setor de seguros tem se apoiado em Modelos Lineares Generalizados (GLM) para auxílio na precificação devido à sua flexibilidade e capacidade de modelar a relação entre variáveis preditoras e a resposta de interesse, geralmente a frequência e severidade de sinistros (McCullagh; Nelder, 2019). Na abordagem GLM, muitas vezes, a dependência entre a frequência dos sinistros e a severidade é introduzida tratando a frequência dos sinistros como uma covariável no modelo de regressão, segundo Saputri e Devianto (2020).

Podem ser apontadas duas boas características dos modelos GLM. Em primeiro lugar, a regressão já não se restringe à distribuição normal dos dados, a qual, para análise de dados de seguros, não é conveniente, mas estende-se às distribuições de outras famílias, o que permite a modelagem apropriada de contagens de frequência, dados assimétricos ou de variáveis binárias, por exemplo. Em segundo lugar, o GLM modela o efeito aditivo de variáveis explicativas em uma transformação da média linearmente relacionada com as variáveis explicativas (Antonio & Beirlant, 2007; Saputri & Devianto, 2020).

No entanto, por mais que esses modelos sejam simples, facilmente explicados e possuam propriedades estatísticas interessantes, eles são sempre bastante restritos para captarem efeitos complexos, de acordo com Blier-Wong *et al*. (2020). Nesse sentido, o advento e a evolução das técnicas de *Machine Learning* nos últimos anos têm oferecido novas oportunidades para a precificação de seguros. Com o advento do *big data* na última década, a inovação tem se concentrado no uso de métodos de aprendizagem atuarial para impulsionar os modelos atuariais clássicos. Para Tzougas e Kutzkov (2023), a utilização de técnicas de aprendizado de máquina na precificação e análise de seguros oferece uma abordagem mais flexível e potencialmente mais precisa para identificar padrões complexos e não lineares nos dados, e sua aplicação no contexto atuarial fez com que se desenvolvesse um novo campo de estudos, os estudos em *Actuarial Learning*.

Desse modo, métodos como árvores de decisão, florestas aleatórias, *boosting* e redes neurais artificiais têm sido explorados para melhorar a acurácia das previsões de sinistros, capturando complexidades e interações entre variáveis que os métodos tradicionais podem não detectar, nas palavras de Guelman (2012). Segundo Blier-Wong *et al*. (2020), uma vantagem significativa dos modelos recentes de *actuarial learning* é que eles aprendem transformações não lineares e interações entre variáveis dos dados sem especificá-las manualmente. Isso é realizado implicitamente com modelos baseados em árvores e explicitamente com redes neurais. Adicionalmente, os autores apontam outra vantagem na existência de muitos modelos para diferentes tipos de formatos de recursos (Blier-Wong *et al*., 2020).

O período entre 2019 e 2020 foi particularmente significativo para o mercado de seguros, marcado por eventos globais disruptivos, como a pandemia da COVID-19, que alteraram padrões de sinistralidade e destacaram a necessidade de métodos de precificação mais dinâmicos e adaptáveis. Essa conjuntura fornece um contexto valioso para investigar como os GLMs e as técnicas de *Actuarial Learning* se comparam na precificação de seguros nesse período. Vale salientar que a utilização de *actuarial learning* está se expandindo rapidamente e mostra grande promessa para uso em ciência atuarial (Guelman, 2012). Destaca-se que introdução do aprendizado de máquina na ciência atuarial é recente, mas ainda não organizada, na visão de Blier-Wong *et al*. (2020) e não há muitas orientações para a escolha das várias técnicas de *machine learning* disponíveis, conforme apontam Balona e Richman (2020).

## 1.1 Objetivos

Este estudo visa preencher uma lacuna na literatura sobre a necessidade de expandir o uso de técnicas de *machine learning* na ciência atuarial e comparar sua eficácia na precificação com os modelos GLM comumente usados pela área. Para isso, serão utilizados dados reais de apólices de seguro residencial compreensivo no Brasil, de uma mistura de carteiras de seguradoras, no período de 2019 a 2020.

O seguro residencial compreensivo desempenha um papel essencial ao proporcionar proteção financeira contra uma variedade de riscos, incluindo, mas não se limitando a, danos por fenômenos naturais, roubo e responsabilidade civil. Esse tipo de seguro é crucial para a segurança financeira dos proprietários de imóveis, tornando a precisão em sua precificação uma questão de grande importância tanto para os segurados quanto para as seguradoras. Análises incorretas podem levar a prêmios inadequados, sub ou sobre seguro, e, em última análise, a uma carteira de seguros insustentável. Adicionalmente, o seguro residencial compreensivo, cobrindo uma ampla gama de riscos para proprietários de imóveis, emerge como um produto crítico nesse contexto, exigindo métodos de precificação precisos e adaptativos.

Nessa feita, este trabalho contribui para o entendimento de como as seguradoras podem melhorar suas estratégias de precificação frente a mudanças rápidas e significativas no mercado e na sociedade. Adicionalmente, também figuram como objetivos específicos deste estudo:

1. Analisar a relação e o poder preditivo existente entre as variáveis independentes e a variável resposta do modelo GLM para frequência e severidade de sinistros.

2. Analisar a qualidade e a adequação do ajuste estatístico, diagnosticando e validando o modelo estimado.

3. Comparar a qualidade dos modelos obtidos por meio do GLM com os obtidos por meio de técnicas de *actuarial learning*.

O presente trabalho é composto por esta introdução e mais quatro seções. Na segunda seção, são tecidas algumas considerações preliminares acerca dos aspectos básicos da tarifação dos seguros privados não vida, além de uma breve revisão bibliográfica sobre a utilização dos GLM e técnicas de *actuarial learning.* Já a terceira seção versa os procedimentos de pesquisa realizados no trabalho, desde a coleta de dados até as técnicas utilizadas, e, na quarta seção, são apresentados os resultados alcançados em linha com os objetivos estabelecidos. Por fim, na quinta seção, são tecidas as considerações finais, além de lançadas as possíveis perspectivas para pesquisas e trabalhos futuros.

## 2. Modelagem Atuarial de Seguros

### 2.1 Generalized Linear Models (GLM)

Os GLMs foram originalmente introduzidos na prática atuarial como um método para melhorar a precisão dos preços de seguro automóvel e, posteriormente, sua utilização foi rapidamente ampliada à maioria das linhas de negócio. Hoje, os GLMs são aplicados rotineiramente à subscrição, precificação ou gestão de sinistros, frequência de sinistros, graduação de taxas de mortalidade e morbidade, modelagem de reserva de perda, entre outros (Denuit *et al*., 2019; Mccullagh & Nelder, 2019). Esses modelos baseiam-se em uma suposição distributiva, numa pontuação linear envolvendo as características disponíveis e em uma ligação assumida entre a resposta esperada e a pontuação. Uma vez que o atuário tenha selecionado esses componentes de acordo com os dados, a inferência estatística é conduzida com a ajuda do princípio da máxima verossimilhança (Denuit *et al*., 2019).

A abordagem GLM é um exemplo de modelagem de regressão. Um modelo de regressão visa explicar algumas características de uma resposta, com a ajuda de características que atuam como variáveis explicativas. Nas aplicações relacionadas a seguros, os atuários geralmente tentam explicar o valor médio da resposta que entra no cálculo puro do prêmio: probabilidade de ocorrência do sinistro ou número esperado de sinistros (frequência de sinistros) e valores de sinistros esperados correspondentes (severidade), por exemplo. Logo, em atuária, os GLMs separam as características sistemáticas dos dados das variações aleatórias que devem ser compensadas pelo seguro. Portanto, um modelo GLM para uma variável resposta $Yi$ é definindo como:

$$score_i = Y_i = x_i^T \beta = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} \, , i = 1, 2, \ldots, n$$

Em que:

$\beta = (\beta_0, \beta_1, \ldots. \beta_p)$ é um vetor de dimensão p + 1 contendo os coeficientes de regressão desconhecidos.

$\beta_j$ são parâmetros a serem estimados a partir dos dados.

Denuit *et al*. (2019) enfatizam o significado da linearidade da função. Os autores reiteram que a palavra “linear” em GLM significa que as variáveis explicativas são combinadas linearmente para chegar à previsão de (uma função da) média. Logo, a linearidade nos GLMs refere-se à linearidade nos coeficientes $\beta_j$, não nas características. Nos modelos GLM, em vez de igualar a resposta média à pontuação como no modelo de regressão linear normal, há uma transformação diferenciável contínua um para um que mapeia a média para a pontuação linear, chamada função de ligação do GLM. Especificamente, a média:

$$\mu_i = \mu(x_i)$$

da resposta $y_i$, $i = 1,2,....,n)$ está vinculada à pontuação envolvendo variáveis explicativas com o auxílio de uma transformação linearizante suave e invertível *g*, que é dada por:

$$g(\mu_i) = score_i$$

É importante salientar neste ponto que não transforma a resposta $Y_i$, mas sim seu valor esperado μi. Para proporções binomiais, a resposta média corresponde à probabilidade de sucesso de modo que μi está confinado ao intervalo unitário [0, 1]. As funções de distribuição são, portanto, candidatas naturais para transformar a pontuação de valor real na resposta média. As funções logit, probit e complementar log-log link mapeiam o intervalo unitário [0, 1] para toda a linha real, removendo a restrição na resposta média. A função de ligação logit é frequentemente usada para probabilidades (μi) em aplicações atuariais:

$$ln\frac{\mu_i}{1-\mu_i} = score_i \quad \longleftrightarrow \quad \mu_i = \frac{\exp(score_i)}{1+score_i} = = \frac{1}{1+\exp(-score_i)}$$

O GLM típico para modelar seguros é um GLM Binomial com função de ligação logit, também conhecido como modelo de regressão logística. Se a probabilidade de sucesso for próxima de 0 (de modo que a resposta seja geralmente 0), então também poderá ser possível usar um GLM de Poisson multiplicativo como uma aproximação, dado que a saída de um GLM multiplicativo é mais fácil de comunicar para um público não técnico. Isso ocorre porque a massa de probabilidade está quase inteiramente concentrada em {0,1} quando o parâmetro de Poisson é pequeno o suficiente (Denuit *et al*., 2019). A expectativa transformada da resposta é modelada, assim temos:

$$g(E[y_i] = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

Uma vez especificado o GLM em termos de função de ligação e distribuição, ou seja, uma vez selecionada uma função de variância pelo atuário, os parâmetros estimados são obtidos pelo método de máxima verossimilhança. Isso fornece ao atuário não apenas estimativas dos coeficientes de regressão $\beta_j$, mas também erros padrão estimados para grandes amostras.

A função de verossimilhança é o produto das probabilidades de observação do valor de cada resposta, sendo a função de densidade de probabilidade usada no lugar da probabilidade para respostas contínuas (Denuit *et al*., 2019). Assim, as estimativas de máxima verossimilhança dos parâmetros $\beta_0$, $\beta_1$,..., $\beta_p$ maximizam a seguinte função:

$$L(\beta) = \sum_{i=1}^{n} \ln f_\theta\,(y_i) = \sum_{i=1}^{n} \frac{Yi\ \theta i - a(\theta i)}{\phi / vi} + \text{constante em relação a } \beta$$

Em que $\theta_i$ envolve o $score_i$. Para encontrar a estimativa de máxima verossimilhança para β, L é diferenciado em relação a $\beta_j$ e igualado a 0.

## 2.2 *Actuarial Learning* (AL)

O aprendizado de máquina nasceu na década de 60 como um campo da inteligência artificial que tinha o objetivo de aprender padrões com base em dados, desde os anos 90, essa área expandiu seus horizontes e começou a se estabelecer como um campo próprio (Izbicki & Santos, 2020). As aplicações de aprendizado de máquina, com o passar do tempo, passaram a ter interseções com as aplicações estatísticas, apesar de se tratar de um campo bastante interdisciplinar, conforme afirmam Izbicki e Santos (2020).

A utilização de técnicas de aprendizado de máquina na precificação e análise de seguros oferece uma abordagem mais flexível e potencialmente mais precisa para identificar padrões complexos e não lineares nos dados. Além disso, sua aplicação no contexto atuarial fez com que desenvolvesse um novo contexto de estudos, os estudos em *Actuarial Learning* (AL). Os métodos de AL que têm sido usados até agora para abordar com eficiência a regressão alternativa e problemas de classificação em seguros incluem, por exemplo, XGBoost, floresta aleatória (RF), árvores de decisão (DTs), redes neurais (NNs), entre outros (Tzougas & Kutzkov, 2023).

A **árvore de regressão** é uma metodologia não paramétrica que possui como *output* resultados de fácil interpretação e compreensão. Trata-se de uma metodologia construída por "particionamentos recursivos no espaço das covariáveis" e apresentam a vantagem de lidar trivialmente com covariáveis discretas (Izbicki; Santos, 2020, p. 77). Conforme Izbicki e Santos (2020, p. 77):

> A utilização da árvore para prever uma nova observação é feita do seguinte modo: começando pelo topo, verificamos se a condição descrita no topo (primeiro nó) é satisfeita. Caso seja, seguimos a esquerda. Caso contrário, seguimos a direita. Assim prosseguimos até atingir uma folha.

A literatura explica que a árvore de regressão estabelece uma partição do espaço das covariáveis em regiões distintas e disjuntas: $R_1$, $R_2$, ..., Rj. A variável resposta ou dependente (Y), nesse sentido, é dada por:

$$g(x) = \frac{1}{|\{i: xi \in Rk\}|}$$

Portanto, para prever o valor da covariável X, observa-se a região a que X pertence e então, calcula-se a média dos valores da variável Y das amostras do conjunto de treinamento pertencente àquela região (Izbicki & Santos, 2020). A criação da estrutura de uma árvore de regressão é feita por meio de duas etapas: (i) a criação de uma árvore completa e complexa e (ii) a poda dessa árvore, com a finalidade de evitar o super ajuste (Izbicki & Santos, 2020). Em que pese as metodologias de árvore de regressão terem como principal vantagem a facilidade interpretativa de seus resultados, alguns autores como Breiman (2001) alertam para o seu baixo valor preditivo, quando comparadas às demais metodologias.

Por isso, Izbicki e Santos (2020) defendem que a metodologia de ***random forest*** tem o potencial de mitigar esse efeito por meio da combinação de diversas árvores que fazem a predição do comportamento de uma mesma variável explicativa. A técnica de *random forest* tenta diminuir essa correlação modificando o mecanismo de criação das árvores para que essas se tornem diferentes umas das outras. Portanto, ao invés de escolher qual das d covariáveis será utilizada em cada um dos nós da árvore, em cada passo só é permitido que seja escolhida uma dentre as m < d covariáveis. Dessa maneira, as covariáveis que entram no modelo são escolhidas de forma aleatória dentre àquelas originais e para cada nó, um novo conjunto de covariáveis é selecionado (Izbicki & Santos, 2020).

De forma semelhante à *random forest*, a técnica ***boosting*** agrega diferentes estimadores em função da regressão. No entanto, diferentemente da *random forest*, no *boosting* o estimador em função de x é construído de forma incremental. Destaca-se que esse estimador possui alto viés, mas baixa variância (a saber, zero). Assim, a cada passo, o valor de g é atualizado de modo a diminuir o viés e aumentar a variância da nova função. Isso é feito adicionando-se a g uma função que prevê os resíduos ri = Yi − g(xi). Adicionalmente, a função g é adicionada multiplicando-se seu valor por um fator multiplicador λ chamado de *learning rate*, o qual varia entre 0 e 1 e possui o objetivo de evitar supera ajuste (Izbicki & Santos, 2020). Dessa maneira, conforme Izbicki e Santos (2020), a versão *boosting* consiste em:

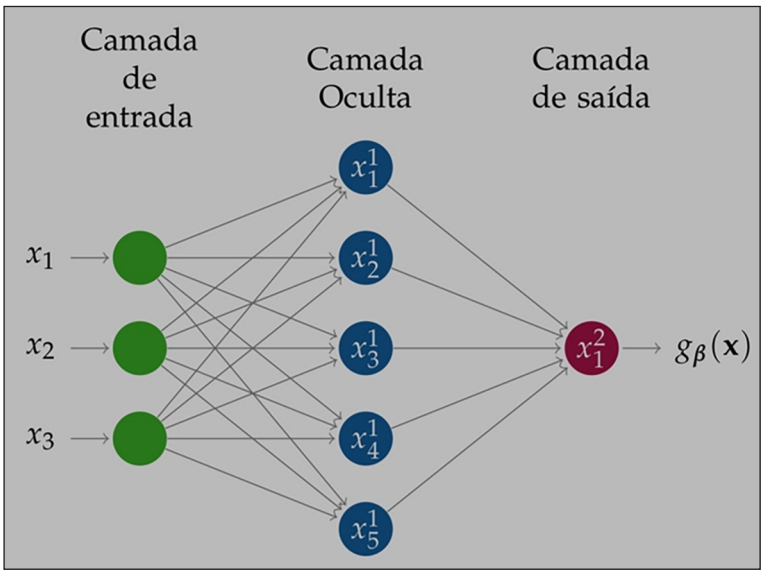
1. Define g(x) ≡ 0 e ri = yi $\forall_i$ = 1,.. .,n.
2. Para b = 1,.. .,B:
    - Ajusta uma árvore com p folhas para (x1, r1),.. .,(xn,rn). Seja g b (x) sua respectiva função de predição.
    - Atualiza g e os resíduos: g(x) ← g(x) + λg b (x) e ri ← Yi − g(x).
3. Retorna o modelo final g(x)

De tal modo, os *tuning parameters* do modelo *boosting* são B, p e λ, em que λ é pequeno e B ≈ 1000 e p é da ordem de 2 ou 4. Os autores reiteram que não necessariamente o *boosting* é feito com árvores, mas em geral são utilizados estimadores "fracos" como, por exemplo, uma regressão com poucos parâmetros.

Existem diferentes implementações e variações de *boosting* (Izbicki & Santos, 2020). Nessa linha, Chen e Guestrin (2016) descreveram em seu artigo um sistema escalonável de aumento de árvore ponta a ponta chamado ***XGBoost***, que é usado amplamente por cientistas de dados "*para alcançar resultados de última geração em muitos desafios de aprendizado de máquina*" (Chen & Guestrin, 2016, p. 785). Os autores demonstraram que a utilização da técnica XGBoost é eficiente para modelar uma ampla gama de situações, tais como predição de vendas, previsão do comportamento do cliente, detecção de movimento, previsão da taxa de evasão do curso, entre outras. Nessa linha, a literatura afirma que o fator mais importante por trás do sucesso do XGBoost é sua escalabilidade em todos os cenários. O sistema funciona dez vezes mais rápido do que as soluções populares existentes em uma única máquina e pode ser dimensionada para bilhões de exemplos em sistemas distribuídos ou configurações de memória limitada. A escalabilidade do XGBoost se deve para vários sistemas importantes e otimizações algorítmicas (Chen & Guestrin, 2016; Izbicki & Santos, 2020).

Por fim, um conceito bastante antigo em inteligência artificial pode ser utilizado, trata-se do conceito de **Redes Neurais Artificiais**. Matematicamente, no contexto de regressão, trata-se de um estimador não linear de $r(x)$, que pode ser representado graficamente por uma estrutura como a da Figura 1 (Izbicki & Santos, 2020):

**Figura 1** – Representação gráfica de redes neurais naturais




Fonte: Izbicki & Santos, 2020.

Os nós do lado esquerdo da figura 1 representam as entradas da rede, isto é, cada uma das covariáveis de análise. Os nós da segunda camada são os chamados de nós da camada oculta da rede. Cada flecha representa um peso (parâmetro) β. Cada nó nessa camada representa uma transformação dos nós das variáveis da camada anterior. De forma geral, uma rede neural pode ter múltiplas camadas ocultas e um número distinto de neurônios em cada camada. Destaca-se que as escolhas são feitas pelo usuário, e a propagação das covariáveis de entrada na rede neural é feita sequencialmente nessas camadas. Assim, se f(z) = z e não há nenhuma camada oculta e, portanto, a rede neural representa uma regressão linear usual (Izbicki & Santos, 2020). Tzougas e Kutzkov (2023), ao compararem modelos GLM e de AL, reafirmam que o reforço da rede neural permite explorar as faltas interações de tipo não linear que não podem ser capturadas por GLM, como regressão logística.

## 3. Dados e Modelos

Este estudo visa preencher a lacuna apontada na literatura sobre a necessidade de expandir o uso de técnicas de *actuarial learning* e comparar sua eficácia na precificação de seguros com os modelos GLM comumente usados na área. Para isso, serão utilizados dados reais de seguro residencial compreensivo de 2019. No quadro 1, são apresentadas as variáveis que constavam na base de contratos.

**Quadro 1** – Variáveis disponíveis na base de Apólices

| Variável | Tipo | Descrição |
|---|---|---|
| Cobertura | Categórica | 220: Danos elétricos |
| | | 270: Incêndio/Raio/Explosão/Demolição/Desentulho |
| | | 520: Vendaval/Furacão/Ciclone/Tornado/Granizo/Queda de aeronaves/Impacto de veículos/Tremor de terra/Terremoto |
| UF | Categórica | Estado da federação onde se localiza a residência segurada contemplando os 26 estados mais o Distrito Federal |
| Início da Vigência | Data | Data de início da vigência da Apólice |
| Fim da Vigência | Data | Data de término da vigência da Apólice |
| Exposição | Numérica | Diferença entre a data de fim e a data de início da vigência, esse resultado dividido por 365 dias |
| Classe | Categórica | 01: Casa Habitual |
| | | 02: Casa Veraneio |
| | | 03: Apartamento Habitual |
| | | 04: Apartamento Veraneio |

Fonte: Elaborado pelos autores.

No quadro 2, são apresentadas as variáveis que constavam na base de sinistros ocorridos.

**Quadro 2** – Variáveis disponíveis na base de sinistros

| Variável | Tipo | Descrição |
|---|---|---|
| Cobertura | Categórica | 220: Danos elétricos |
| | | 270: Incêndio/Raio/Explosão/Demolição/Desentulho |
| | | 520: Vendaval/Furacão/Ciclone/Tornado/Granizo/Queda de aeronaves/Impacto de veículos/Tremor de terra/Terremoto |
| UF | Categórica | Estado da federação onde se localiza a residência segurada contemplando os 26 estados mais o Distrito Federal |
| Classe | Categórica | 01: Casa Habitual |
| | | 02: Casa Veraneio |
| | | 03: Apartamento Habitual |
| | | 04: Apartamento Veraneio |
| Indenização ou severidade | Numérica | Valor da indenização, em R$, paga pela seguradora |
| Ocorrência do sinistro | Data | Data da ocorrência do sinistro nos anos de 2019 e 2020 |
| Ano | Ano | Ano da ocorrência do sinistro – podendo ser 2019 e 2020 |

Fonte: Elaborado pelos autores.

Após a obtenção dos dados, juntaram-se os dois bancos de dados. Feito isso, as variáveis categóricas foram transformadas em variáveis indicadoras. Em seguida, os dados foram agrupados e foi calculada a existência de sinistros para as covariáveis indicadoras Classe, Cobertura e UF. Em virtude de os valores se repetirem para os vários sinistros, foram criadas duas variáveis: ***expos* – que é a média da exposição – e *inden* – que é a média das indenizações por apólice**. Após esse passo, a base de dados foi divida em amostra de treino e em amostra de teste, para que se pudesse avaliar a qualidade do ajuste dos modelos.

Para os modelos GLM, foram calculados o Erro Absoluto Médio (MAE) e a Raiz do Erro Quadrático Médio (RMSE) para a amostra de teste. O MAE é uma medida da diferença entre os valores preditos e os valores observados. Ele indica quão grande é o erro em média, com todas as diferenças sendo tratadas igualmente, independentemente da direção.

O RMSE é a raiz quadrada da média dos quadrados dos erros. O erro é a diferença entre os valores observados e os valores preditos. O RMSE é uma medida comumente usada para avaliar a precisão de modelos de predição, pois dá mais peso a grandes erros. Ambas as métricas, MAE e RMSE, são usadas para quantificar a precisão das predições. Dessa maneira, valores mais baixos indicam melhor ajuste do modelo aos dados. A escolha entre MAE e RMSE depende do contexto específico e se grandes erros são consideravelmente mais problemáticos do que pequenos erros, o que faz do RMSE uma escolha mais adequada para esse caso.

Feito isso, foram realizados modelos GLM e modelos Actuarial Learning. Dessa forma, foram utilizados os modelos GLM (Modelo 1 e Modelo 2), abordados a seguir.

### *3.1 Modelos lineares generalizados (GLM)*

**Modelo 1: Frequência de Sinistros**

$$P\,(Y = 1) = \frac{1}{1 + e^{-(\beta_1 X_1 + \beta_2 X_2 + \ldots . \beta_n X_n)}}$$

Onde:

- P(Y=1) é a probabilidade da variável resposta ser 1 (ou a categoria de interesse em um contexto binário), no caso do estudo, a frequência de sinistros;
- X1, X2, …, Xn são as variáveis preditoras incluídas no modelo, representada pelas variáveis indicadoras de Classe, Cobertura e UF;
- β1, β2 …, βn são os coeficientes do modelo para cada variável preditora, que quantificam a relação entre cada preditor.

**Modelo 2: Severidade**

$$\log(\mu i) = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_n X_{ni}$$

$$l(\beta) = \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i)(\log 1 - p_i)]$$

Onde:

- $\mu i$ é o valor esperado da variável resposta **indenização** para a observação i;
- X1, X2, …, Xn são as variáveis preditoras incluídas no modelo, representada pelas variáveis indicadoras de Classe, Cobertura e UF;

- β1, β2 …, βn são os coeficientes do modelo para cada variável preditora, que quantificam a relação entre cada preditor, assumindo que o modelo não inclui o intercepto;

- $l(\beta)$ é a função de verossimilhança logarítmica para um modelo de regressão logística. Essa função é usada para estimar os parâmetros do modelo (como os pesos associados às variáveis independentes) através da maximização da verossimilhança. A função de verossimilhança é maximizada em relação aos parâmetros do modelo para encontrar a estimativa que torna os dados observados mais prováveis.

### *3.2 Actuarial Learning*

**Modelo 3: Árvores de regressão**

De maneira simples, uma árvore de decisão de regressão é um modelo que prediz um valor de resposta contínuo (variável dependente) com base em várias variáveis independentes. A fórmula para a árvore de decisão é mais um conjunto de regras condicionais do que uma fórmula algébrica tradicional. Neste trabalho, utilizamos árvores de regressão para modelar a relação entre a frequência de sinistros e a severidade e as covariáveis (estado da federação, classe de seguro, cobertura e exposição da apólice), já apresentadas nos quadros 1 e 2.

**Modelo 4: *Random Forest***

Devido à sua natureza de conjunto e ao fato de que cada árvore de decisão dentro da floresta é construída com diferentes amostras e divisões dos dados, não há uma equação algébrica única e simples que possa representar um modelo de floresta aleatória como um todo. No entanto, cada árvore individual dentro da floresta aleatória pode ser expressa como uma série de decisões baseadas em divisões nos valores das variáveis preditoras. Matematicamente, é possível expressar as previsões de cada árvore como uma função de classificação f que produz um valor para uma observação dada x, mas isso seria uma representação muito complexa e não prática para entender intuitivamente ou usar em cálculos manuais. Neste trabalho, utilizamos *random forest* para modelar a relação entre a frequência de sinistros e as covariáveis (estado da federação, classe de seguro, cobertura e exposição da apólice) e a severidade, já apresentadas nos quadros 1 e 2.

**Modelo 5: *Boosting***

O b*oosting* é um método de aprendizado de máquina que cria um modelo forte a partir de um número de modelos fracos. No contexto da regressão, esses modelos fracos são frequentemente árvores de decisão. O modelo de boosting ajusta repetidamente as árvores de decisão aos resíduos (erros) da iteração anterior para melhorar a previsão. Assim como a floresta aleatória, não há uma fórmula algébrica simples para a representação de um modelo de boosting, pois ele é construído iterativamente e pode envolver várias árvores de decisão. No entanto, de forma genérica, o modelo final de boosting pode ser descrito como uma soma ponderada das árvores de decisão. O modelo *boosting* neste trabalho utilizou todas as covariáveis do banco de dados como preditoras (Classe, Cobertura, Exposição, UF) para prever a frequência de sinistros e a severidade. Ele usou árvores de decisão com uma profundidade máxima de 6 como aprendizes básicos, realizou até 1000 iterações de boosting com uma taxa de aprendizagem de 0.005 e ajustou os valores de resposta centralizando-os.

**Modelo 6: XGBoost**

Suponha um conjunto de dados D = {($xi$, $yi$)}, onde $xi$ representa as características (ou variáveis independentes) de cada observação $i$, e $yi$ é a variável dependente (ou rótulo) correspondente. O objetivo do XGBoost é construir um modelo $f(x)$ que preveja $y$ com base em $x$. Para isso, ele cria uma soma de árvores de decisão em que $k(x)$ é o número total de árvores, e cada $fk(x)$ é uma árvore de decisão:

$$f(x) = \sum_{k=1}^{k} fk(x)$$

A função objetivo do XGBoost que o algoritmo tenta minimizar é composta por uma função de perda $L$ e um termo de regularização $\omega$:

$$Obj = \sum_{i} L(yi, xi) + \sum_{k} \omega(fk)$$

Em que, $yi$ é o valor observado, $y = f(xi)$ é a previsão do modelo, e ($fk$) é a $k$-ésima árvore.

O termo de regularização $\omega$ é usado para controlar a complexidade do modelo. Durante o treinamento, o XGBoost atualiza os pesos das folhas de cada árvore para minimizar a função objetivo. Ele usa o gradiente da função de perda para encontrar a direção que reduzirá o erro, ajustando-se pelos parâmetros de regularização e taxa de aprendizado.

O modelo XGBoost utilizou todas as covariáveis do banco de dados como preditoras (Classe, Cobertura, Exposição, UF) para prever a frequência de sinistros e severidade. Ele foi configurado para usar árvores de decisão com boosting, especificamente com a função objetivo Tweedie, que é adequada para os dados com muitos zeros e valores positivos. O modelo inclui uma série de parâmetros de regularização e controle de complexidade para otimizar o desempenho e evitar o *overfitting*.

**Modelo 7: Rede Neural Artificial**

Para o caso dessa rede, se x = (x1, x2, x3) é o vetor de entrada, então um dado neurônio j da camada oculta é calculado com:

$$x_j^1 = f\,(\beta_{0,j}^0 + \sum_{i=1}^{3} \beta_i^0, x_1^0\,)$$

em que:

$x_1^0 = xi$ para $i$ = 1,2,3;

f é uma função definida pelo usuário, chamada de função de ativação;

o índice superescrito nessa equação denota a camada da rede.

Calculados os valores de $x_1^0$ para todo neurônio j da camada oculta, pode-se então calcular a saída do modelo.

O modelo de Rede Neural utilizado neste trabalho foi realizado preparando um tensor a partir de um *dataframe* que contém suas variáveis preditoras para treinamento de um modelo de *actuarial learning* (Classe, Cobertura, Exposição, UF), excluindo a variável de resposta frequência de sinistros em um modelo e severidade em outro, por serem variáveis respostas dos modelos. O modelo de rede neural desenvolvido consistiu em uma arquitetura sequencial com quatro camadas, onde a primeira camada possui 64 neurônios, a segunda, 32, a terceira, 16, e a camada de saída possui um único neurônio. A função de ativação 'sigmoide' foi utilizada na camada de saída, com funções de ativação lineares nas camadas anteriores por padrão. A função de perda *'mean squared error'* foi empregada para compilação do modelo.

## 4. Resultados

### 4.1 Modelos lineares generalizados (GLM)

#### *4.1.1 Frequência de sinistros*

No quadro 3, são apresentados os resultados obtidos para o modelo GLM de existência de sinistros para amostra de treino. Analisando os p-valores do Modelo 1, relacionado à existência de sinistros, é possível verificar que algumas variáveis como Casa Veraneio (classe 2), Apartamento Habitual (classe 3) e Apartamento Veraneio (classe 4) são altamente significativas a possuírem sinistros, quando comparadas com a Casa Habitual (classe 1). Isso sugere que essas classes têm um efeito estatisticamente significativo sobre a variável resposta existência de sinistros.

A maioria das variáveis relacionadas ao estado da federação não parece ser significativa, exceto Pará e Pernambuco, que têm p-valores mais baixos, embora ainda acima do nível de significância convencional. Alguns coeficientes têm valores extremamente altos ou baixos (por exemplo, Apartamento Habitual e estado do Piauí), o que pode indicar sobre ajuste, categorias com poucas observações ou outras questões com os dados ou o ajuste do modelo. Verificou-se também que as estimativas de desvio padrão para algumas variáveis categóricas são muito altas, o que sugere grande incerteza nessas estimativas.

A *deviance* nula e a *deviance* residual mostram quão bem o modelo com os preditores se ajusta aos dados comparados ao modelo mais simples possível, o modelo nulo, que inclui apenas o intercepto. Quanto mais baixa a *deviance residual,* em comparação com a *deviance nula*, melhor o modelo com preditores se ajusta aos dados. O Critério de Informação de Akaike (AIC) é uma medida de qualidade do modelo que penaliza a complexidade. Um AIC menor indica um modelo melhor, equilibrando ajuste e simplicidade. O número de iterações do *scoring* de Fisher indica quantas iterações foram necessárias para o algoritmo convergir para uma solução. O modelo precisou de 12 iterações, o que é um pouco alto, mas não necessariamente um problema.

**Quadro 3** – GLM para Existência de Sinistros

| Variável | Estimativa | Desvio-Padrão | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| Intercepto | -0.79065 | 0.95431 | -0.828 | 0.407389 | |
| CLASSE2 | -0.74662 | 0.17599 | -4.243 | 2.21e-05 | *** |
| CLASSE3 | -133.998 | 0.06279 | -21.339 | < 2e-16 | *** |
| CLASSE4 | -122.588 | 0.27140 | -4.517 | 6.27e-06 | *** |
| COBERTURA270 | -0.07400 | 0.24054 | -0.308 | 0.758364 | |
| COBERTURA520 | 0.61069 | 0.18099 | 3.374 | 0.000741 | *** |
| UFAL | -140.886 | 143.398 | -0.982 | 0.325864 | |
| UFAM | 134.828 | 112.672 | 1.197 | 0.231447 | |
| UFBA | -125.594 | 113.482 | -1.107 | 0.268408 | |
| UFCE | -1.259.283 | 13.248.530 | -0.095 | 0.924274 | |

| Variável | Estimativa | Desvio-Padrão | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| UFDF | -0.54226 | 113.960 | -0.476 | 0.634189 |
| UFES | 0.37028 | 102.144 | 0.363 | 0.716971 |
| UFGO | 0.16845 | 0.94632 | 0.178 | 0.858717 |
| UFMA | -1.276.078 | 34.865.547 | -0.037 | 0.970804 |
| UFMG | 0.16245 | 0.95101 | 0.171 | 0.864364 |
| UFMS | 0.07270 | 0.94932 | 0.077 | 0.938958 |
| UFMT | 0.75405 | 0.94989 | 0.794 | 0.427294 |
| UFPA | 154.850 | 128.154 | 1.208 | 0.226927 |
| UFPB | 0.54295 | 134.330 | 0.404 | 0.686073 |
| UFPE | -168.566 | 140.145 | -1.203 | 0.229056 |
| UFPI | -1.338.611 | 53.541.199 | -0.025 | 0.980054 |
| UFPR | 100.152 | 0.93764 | 1.068 | 0.285461 |
| UFRJ | 0.24999 | 0.97829 | 0.256 | 0.798305 |
| UFRN | -0.66859 | 150.535 | -0.444 | 0.656940 |
| UFRO | 0.72154 | 114.512 | 0.630 | 0.528630 |
| UFRS | 108.363 | 0.93789 | 1.155 | 0.247928 |
| UFSC | 108.391 | 0.93802 | 1.156 | 0.247877 |
| UFSE | -1.276.078 | 34.865.547 | -0.037 | 0.970804 |
| UFSP | 0.41299 | 0.93863 | 0.440 | 0.659942 |
| UFTO | 0.74244 | 111.458 | 0.666 | 0.505334 |
| *Significância* | *0 '***' 0.001 '**' 0.01 '*' 0.05* | | | |
| *Null deviance:* | *15173 on 10945 degrees of freedom* | | | |
| *Residual deviance:* | *13626 on 10915 degrees of freedom* | | | |
| *AIC:* | *13686* | | | |
| *Number of Fisher Scoring iterations:* | *12* | | | |

Fonte: Elaborado pelos autores.

Ao aplicar o modelo na amostra de teste, foram obtidos os seguintes Erro Absoluto Médio (MAE) – 0,5138 e Raiz do Erro Quadrático Médio (RMSE) – 0,9211. Em média, a predição do modelo desvia de aproximadamente 0,514 unidades da verdadeira frequência observada nos dados de teste. Quanto menor o MAE, mais próximo as predições estão dos valores reais. Um MAE de 0,514 pode ser considerado bom ou ruim dependendo do contexto dos dados e da variabilidade inerente da variável existência de sinistros. O RMSE é geralmente mais sensível a erros maiores porque os erros são elevados ao quadrado antes de fazer a média. O resultado indica que, em média, o modelo tem um desvio de aproximadamente 0,921 unidades do valor real da frequência nos dados de teste. Comparado ao MAE, o RMSE dá uma ideia da variação das predições e é mais influenciado por predições muito distantes dos valores reais. MAE e RMSE são medidas de desempenho do modelo que têm valores não negativos, em que valores mais baixos indicam melhores predições.

### *4.1.2 Severidade*

No quadro 5 são apresentadas as estatísticas descritivas da média de indenização:

**Quadro 5** – Estatísticas descritivas da Severidade

| | |
|---|---|
| Mínimo | 20 |
| 1º Quartil | 846,10 |
| Mediana | 1.635,80 |
| Média | 2.863,40 |
| 3º Quartil | 3.228,00 |
| Máximo | 1.999.024,00 |

Fonte: Elaborado pelos autores.

A distribuição dos dados de severidade é altamente assimétrica, com uma cauda longa à direita. Isso é indicado pela diferença entre a média e a mediana, com a média sendo muito mais alta. A presença de um valor máximo extremamente alto (um outlier) tem um impacto significativo no cálculo da média. O valor máximo é muito mais alto do que o 3º quartil, indicando que os valores máximos são *outliers*. Esses resultados sugerem que a variável Severidade apresenta uma distribuição com forte assimetria positiva e é influenciada por valores extremamente altos. No quadro 6, são apresentados os resultados obtidos para o modelo GLM para severidade. Analisando os p-valores do Modelo 2, com relação à severidade, é possível verificar que a variável Casa Veraneio (classe 2) é significativa para explicar o valor da severidade, no entanto, com uma relação inversa.

Esses resultados indicam que o valor médio da severidade é menor quando o sinistro ocorre em Casa de Veraneio quando comparado com a severidade de Casa Habitual (classe 1). A variável de cobertura vinculada à fenômenos naturais (cobertura 520) possui p-valores mais baixos em um nível de significância de 5%. Os coeficientes representam a mudança logarítmica na média da variável de resposta. Por exemplo, o coeficiente para fenômenos naturais (cobertura 520) sugere que, mantendo tudo o mais constante, um aumento de uma unidade nessa variável está associado a um aumento de 12,6% na média esperada da severidade (pois *exp(0,126) ≈ 1,13*, o que implica um aumento de aproximadamente 13,4% na severidade).

No modelo de severidade não retornou como significativa a localização do sinistro como explicativa da severidade. Logo, em relação à variável estado da federação, nenhuma delas é estatisticamente significativa no modelo. A falta de significância para muitas variáveis sugere que o modelo pode não ser o melhor ajuste para esses dados ou que essas variáveis podem não ter um forte poder preditivo para a severidade dos sinistros. Aplicando-se o modelo de severidade na amostra de teste, foram obtidos Erro Absoluto Médio (MAE) em 275,52 e Raiz do Erro Quadrático Médio (RMSE) em 2344,19 para severidade.

O fato de o RMSE ser substancialmente maior que o MAE pode indicar que há grandes erros de previsão no conjunto de dados, ou seja, pode haver *outliers* ou previsões significativamente diferentes dos valores reais. Essas métricas por si só não dizem se o modelo é bom ou não; isso depende do contexto dos dados e da aplicação. Por exemplo, em algumas situações, um MAE de 2755,52 pode ser excelente, enquanto, em outras, pode ser inadequado. Isso normalmente é avaliado em comparação com um *benchmark* ou a variação da própria variável de resposta.

**Quadro 6** – GLM para Severidade

| Variável | Estimativa | Desvio-Padrão | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| Intercepto | 4,46E-01 | 7,28E-01 | 0.612 | 0.541 | |
| CLASSE2 | -1,90E-01 | 3,48E-02 | -5.449 | 5.25e-08 | *** |
| CLASSE3 | 3,80E-02 | 3,44E-02 | 1.106 | 0.269 | |
| CLASSE4 | 2,16E-01 | 2,21E-01 | 0.978 | 0.328 | |
| COBERTURA270 | 6,86E-02 | 7,52E-02 | 0.912 | 0.362 | |
| COBERTURA520 | 1,26E-01 | 4,89E-02 | 2.575 | 0.010 | * |
| UFAL | 1,08E-01 | 1,01E+00 | 0.107 | 0.915 | |
| UFAM | -3,06E-01 | 7,42E-01 | -0.412 | 0.681 | |
| UFBA | 6,75E-02 | 9,36E-01 | 0.072 | 0.943 | |
| UFCE | 1,75E-01 | 1,23E+00 | 0.142 | 0.887 | |

| Variável | Estimativa | Desvio-Padrão | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| UFDF | 4,38E-02 | 8,85E-01 | 0.049 | 0.961 |
| UFES | 5,14E-01 | 9,57E-01 | 0.537 | 0.592 |
| UFGO | 1,51E-02 | 7,33E-01 | 0.021 | 0.984 |
| UFMA | -5,11E-02 | 1,18E+00 | -0.043 | 0.966 |
| UFMG | 9,39E-02 | 7,40E-01 | 0.127 | 0.899 |
| UFMS | -3,16E-02 | 7,34E-01 | -0.043 | 0.966 |
| UFMT | -2,98E-01 | 7,28E-01 | -0.409 | 0.682 |
| UFPA | 3,59E-01 | 1,11E+00 | 0.323 | 0.747 |
| UFPB | -4,10E-01 | 7,51E-01 | -0.545 | 0.586 |
| UFPE | 4,40E-01 | 1,96E+00 | 0.225 | 0.822 |
| UFPR | -2,38E-01 | 7,26E-01 | -0.327 | 0.743 |
| UFRJ | -1,19E-01 | 7,50E-01 | -0.159 | 0.874 |
| UFRN | 4,95E+00 | 1,00E+01 | 0.494 | 0.621 |
| UFRO | -1,24E-01 | 8,13E-01 | -0.152 | 0.879 |
| UFRS | -2,50E-01 | 7,26E-01 | -0.344 | 0.731 |
| UFSC | -1,73E-01 | 7,27E-01 | -0.239 | 0.811 |
| UFSE | NA | NA | NA | NA |
| UFSP | -9,38E-02 | 7,27E-01 | -0.129 | 0.897 |
| UFTO | -2,95E-01 | 7,55E-01 | -0.391 | 0.696 |
| *Significância* | *0 '***' 0.001 '**' 0.01 '*' 0.05* | | | |
| *Null deviance:* | *6830 degrees of freedom* | | | |
| *Residual deviance:* | *7139.5 on 6802 degrees of freedom* | | | |
| *AIC:* | *121977* | | | |
| *Number of Fisher Scoring iterations:* | *9* | | | |

Fonte: Elaborado pelos autores.

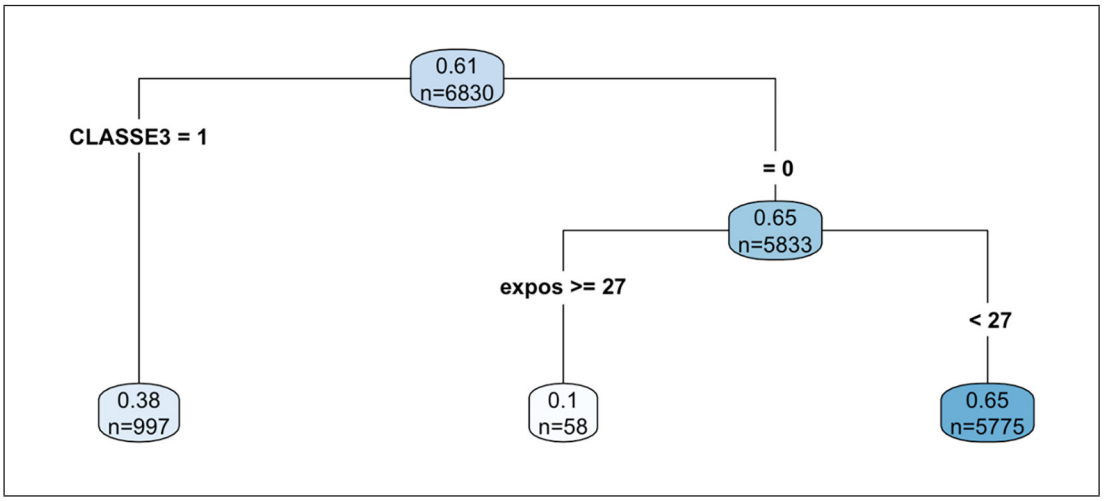
### 4.2 Actuarial Learning

#### *4.2.1 Árvore de Regressão*

- **Frequência**

De forma a alcançar o objetivo da pesquisa, foram aplicados ao banco de dados de seguro residencial compreensivo metodologias de *Actuarial Learning*. A figura 2 evidencia o resultado obtido por meio da técnica de Árvore de Regressão para frequência de sinistros:

**Figura 2** – Árvore de Regressão para amostra de treino (frequência)




Fonte: Elaborado pelos autores.

Na figura 1, verifica-se que o valor de 0,61 é a média da variável de resposta frequência de sinistros para todo o conjunto de dados (n=6830). Esse é o valor previsto se não aplicarmos nenhuma divisão adicional. A primeira divisão é baseada na variável Apartamento Habitual (classe 3). A árvore divide o conjunto de dados em registros onde classe 3 é igual a 1 e registros onde classe 3 é igual a 0. Portanto, nos casos em que o sinistro ocorreu em um Apartamento Habitual (classe 3), a média da frequência de sinistros é 0,38 (n=997). Para outras classes de sinistro, a árvore não fez nenhuma divisão adicional. Já para a segunda divisão, a árvore verifica a condição baseada na variável média da exposição (expos). Dessa forma, a divisão foi realizada quando a média da exposição é maior que 27. Assim, se a exposição é maior que 27, a frequência de sinistros é 0,1 (n=58), já se a média da exposição é menor que 27, o valor previsto de frequência de sinistros é 0,65.

As divisões (ou “splits”) na árvore são escolhidas para minimizar a soma dos quadrados dos resíduos dentro de cada grupo após a divisão. A árvore mostrada é uma versão podada do modelo completo, onde o corte foi escolhido para minimizar o erro de validação cruzada. Aplicando-se a metodologia de árvore de regressão de frequência de sinistros na amostra de teste, foram obtidos os seguintes Erro Absoluto Médio (MAE) em 0,4405 e Erro Quadrático Médio (RMSE) em 0,4683.

O RMSE é semelhante ao MAE, mas eleva os erros ao quadrado antes de calcular a média e, finalmente, tira a raiz quadrada do resultado. Isso dá mais peso aos erros maiores, tornando o RMSE mais sensível a *outliers* do que o MAE. Um RMSE de 0,4683 indica que, em média, o quadrado dos erros entre as previsões do modelo e os valores reais têm uma média de aproximadamente 0,4683 unidades. Como o RMSE é medido nas mesmas unidades da variável de interesse, neste caso, a frequência de sinistros, um RMSE de cerca de 0,47 sugere que o modelo tem uma precisão razoável, mas, assim como o MAE, indica que há uma variação significativa nas previsões do modelo em comparação com os valores reais.

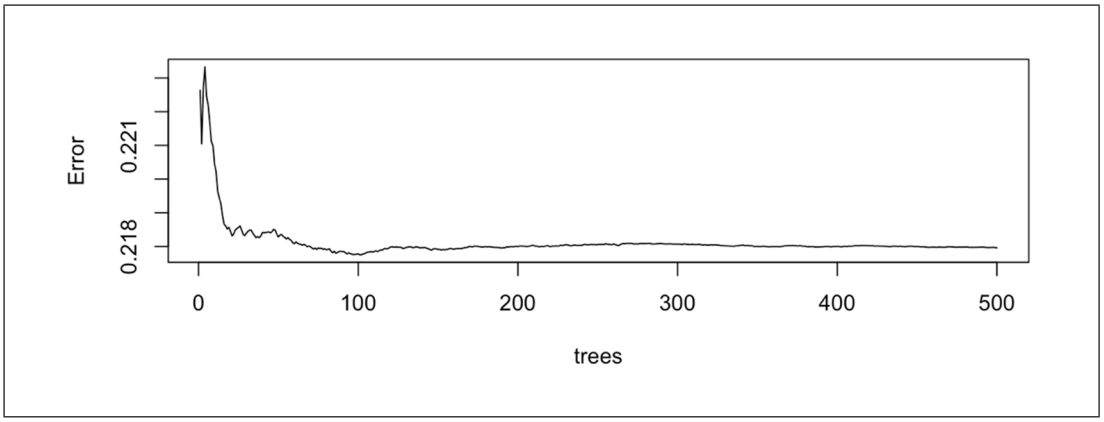
- **Severidade**
  Essa seção evidencia o resultado obtido por meio da técnica de Árvore de Regressão para severidade. Aplicando-se a metodologia de árvore de regressão para severidade na amostra de teste, foram obtidos Erro Absoluto Médio (MAE) em 210,04 e Erro Quadrático Médio (RMSE) em 1501,75.

  O MAE mais baixo em relação ao RMSE sugere que, enquanto a maioria das previsões do modelo pode estar relativamente próxima dos valores reais, há provavelmente alguns casos extremos onde o modelo erra por uma margem significativa, o que aumenta o RMSE. Na prática, se o custo de previsões erradas é muito alto, como no caso de seguros, um RMSE alto é motivo de preocupação e o modelo pode precisar de melhorias para minimizar esses erros grandes.

### *4.2.2 Random Forest*

- **Frequência**
  Os dados de seguro residencial também foram utilizados no modelo de *random forest*. Os resultados obtidos foram apresentados na figura 3. Nesta figura, o eixo y representa o erro (geralmente o erro quadrado médio – MSE – durante o treinamento do modelo) e o eixo x representa o número de árvores construídas no modelo. Conforme evidenciado na figura 3, à medida que mais árvores são adicionadas, o erro diminui rapidamente, o que é esperado porque a adição de árvores aumenta a precisão do modelo. Dessa forma, depois de um certo número de árvores, o erro se estabiliza e há pouca mudança, mesmo com mais árvores sendo adicionadas. Isso sugere que a adição de mais árvores não está melhorando significativamente o desempenho do modelo.

**Figura 3** – Random Forest para amostra de treino




Fonte: Elaborado pelos autores.

O ponto onde o erro se estabiliza indica quantas árvores são necessárias para alcançar um bom desempenho com o modelo. Neste caso, parece que depois de aproximadamente 100 árvores, o erro se estabiliza, indicando que mais árvores além desse ponto podem não ser necessárias. O objetivo é ter um modelo que tenha um erro baixo e estável, o que sugere que o modelo obtido com os dados de seguro residencial utilizados é capaz de fazer previsões precisas com uma variância baixa. Se o erro é baixo e consistente ao longo de várias árvores, isso também pode indicar que o modelo é robusto e não está sofrendo de sobreajuste aos dados de treinamento. Aplicando-se a metodologia de *random forest* na amostra de teste, foram obtidos Erro Absoluto Médio (MAE) em 0,4323 e Erro Quadrático Médio (RMSE) para frequência de sinistros em 0,4565.

- **Severidade**
  Ao aplicarmos a metodologia para severidade na amostra de teste, obtivemos Erro Absoluto Médio (MAE) em 232,44 e Erro Quadrático Médio (RMSE) em 2099,60. A diferença substancial entre o MAE e o RMSE sugere que o modelo tem uma distribuição de erros com *outliers* significativos ou erros esporádicos muito grandes, já que esses erros inflam o RMSE muito mais do que o MAE. Em modelos destinados à previsão de severidade de sinistros em apólices de seguro, é essencial que as métricas de Erro Absoluto Médio (MAE) e Erro Quadrático Médio (RMSE) mantenham-se minimizadas. A acurácia nestas previsões é crítica, visto que imprecisões podem resultar em consequências econômicas de grande escala. Um valor elevado de RMSE, em particular, pode ser indicativo de falhas do modelo em gerar estimativas precisas para casos de sinistros de alta severidade ou de ocorrência infrequente. Tal condição ressalta a necessidade de revisão e potencial aprimoramento do modelo analítico em uso.

#### *4.2.3 Boosting*

Os dados de seguro residencial ainda foram utilizados no modelo *boosting*. Ao aplicarmos a metodologia *boosting* na amostra de teste, obtivemos Erro Absoluto Médio (MAE) em 0,4258 e Erro Quadrático Médio (RMSE) para frequência em 0,4569. O RMSE maior que o MAE indica que existem erros de previsão mais significativos impactando o modelo. Ainda assim, o RMSE é relativamente baixo, o que pode indicar que o modelo, em geral, faz previsões confiáveis. A proximidade entre o MAE e o RMSE sugere que o modelo não tem muitos extremos, já que erros grandes aumentariam o RMSE de forma mais significativa comparado ao MAE. Portanto, o modelo parece ser consistente em suas previsões com erros moderadamente pequenos.

Para aplicação do modelo Boost para severidade para amostra de teste, foram obtidos Erro Absoluto Médio (MAE) em 39,35 e Erro Quadrático Médio (RMSE) para severidade em 766,68.

#### *4.2.4 XGBoost*

Os dados de seguro ainda foram utilizados para o modelo XGBoost. Com a metodologia XGBoost na amostra de teste, foram obtidos Erro Absoluto Médio (MAE) em 0,4210 e Erro Quadrático Médio (RMSE) para frequência de sinistros em 0,4553. Como resultado da aplicação do modelo XGBoost de severidade para amostra de teste, foram obtidos Erro Absoluto Médio (MAE) em 11,39 e Erro Quadrático Médio (RMSE) para severidade em 165,15.

#### *4.2.5 Redes Neurais Artificiais*

Os dados de seguro residencial compreensivo também foram utilizados para o modelo AL de redes neurais artificias. Aplicando-se a metodologia de Redes Neurais Artificiais na amostra de teste para frequência de sinistros, foram obtidos Erro Absoluto Médio (MAE) em 0,4306 e Erro Quadrático Médio (RMSE) em 0,4584. Já na amostra de teste para severidade, foram obtidos Erro Absoluto Médio (MAE) em 10,56 e Erro Quadrático Médio (RMSE) em 108,43.

**Resumo dos resultados**

**Quadro 7** – Análise comparativa dos modelos de frequência de sinistros

| | Método | MAE | RSME |
|---|---|---|---|
| GLM | GLM (frequência de sinistros) | 0,5138 | 0,9211 |
| *Actuarial Learning* | Árvores de regressão | 0,4405 | 0,4683 |
| | Random forest | 0,4323 | 0,4565 |
| | Boosting | 0,4258 | 0,4569 |
| | XGBoost | **0.4210** | **0,4553** |
| | Redes Neurais | 0,4306 | 0,4585 |

Fonte: Elaborado pelos autores.

**Quadro 8** – Análise comparativa dos modelos de severidade

| | Método | MAE | RSME |
|---|---|---|---|
| GLM | GLM (frequência de sinistros) | 275,52 | 2344,19 |
| *Actuarial Learning* | Árvores de regressão | 210,04 | 1501,75 |
| | Random forest | 232,44 | 2099,60 |
| | Boosting | 39,35 | 766,68 |
| | XGBoost | 11,39 | 165,15 |
| | Redes Neurais | **10,56** | **108,43** |

Fonte: Elaborado pelos autores.

## 5. Considerações Finais

Este estudo abordou como as técnicas avançadas de *Actuarial Learning* se comparam com os Modelos Lineares Generalizados (GLM), que são a base tradicional para análise de risco no setor de seguros. De forma específica, foi analisada para a amostra selecionada a precisão das previsões em relação à frequência e severidade dos sinistros, levando em conta variáveis como o estado da federação, a classe de seguro, a cobertura e a exposição da apólice. Por meio de uma análise comparativa clara, os Quadros 7 (frequência de sinistros) e 8 (severidade) ilustram as métricas de erro Médio Absoluto (MAE) e Raiz do Erro Quadrático Médio (RSME) para cada um dos métodos utilizados. De forma didática, essas métricas ajudaram a compreender o quão distantes as previsões dos modelos estão dos valores reais, com valores menores indicando previsões mais precisas.

Os resultados mostram que os métodos de *Actuarial Learning*, especialmente o XGBoost e Redes Neurais, fornecem previsões mais precisas quando comparados ao GLM tradicional para frequência de sinistros. Essa melhoria na precisão pode levar a uma estimativa de risco mais acurada, resultando em precificação mais justa e políticas de subscrição mais eficazes, o que pode, por sua vez, melhorar a lucratividade e a satisfação do cliente das seguradoras.

Na previsão de severidade, a diferença de desempenho entre os modelos é ainda mais acentuada. Novamente, métodos de aprendizado de máquina como XGBoost e Redes Neurais têm um desempenho melhor do que os modelos GLM e métodos de aprendizado atuarial tradicionais. O modelo de *Boosting* apresenta um MAE muito inferior em comparação com as Árvores de Regressão e *Random Forest*, sugerindo que os ajustes iterativos na previsão podem capturar nuances nos dados que outros modelos não captam. XGBoost e Redes Neurais apresentam um desempenho superior, indicando que esses modelos podem ser especialmente adequados para capturar a complexidade e as peculiaridades dos dados de seguros. Esses resultados sugerem um potencial considerável para o uso de técnicas de *actuarial learning* no setor de seguros. O uso de modelos mais avançados pode levar a uma melhor compreensão do risco e a utilização de práticas de precificação mais precisas.

É importante salientar que, apesar de os resultados serem interessantes do ponto de vista da divulgação da melhoria de ajuste obtida para seguros residenciais compreensivos, há algumas limitações da pesquisa que podem afetar a generalização e aplicabilidade dos resultados:

i) os dados abrangem um período específico, que pode não capturar variações cíclicas ou tendências de longo prazo no setor de seguros;

ii) a análise se concentra no mercado brasileiro e os resultados podem não ser diretamente aplicáveis a outras regiões devido a diferenças em fatores de risco, comportamento de sinistros ou práticas de seguros;

iii) a inclusão de variáveis externas adicionais, como dados socioeconômicos ou ambientais, poderia potencialmente melhorar a precisão do modelo, mas esses dados não estavam disponíveis para este estudo;

iv) embora os modelos de *Actuarial Learning* tenham apresentado uma melhor *performance* em comparação com o GLM, eles são intrinsecamente mais complexos, o que pode dificultar a interpretação e a implementação em ambientes operacionais;

v) eventuais mudanças no comportamento do consumidor ou na demanda de seguros residenciais não são contempladas pelos dados;

vi) mesmo com os argumentos de qualidade apresentados no capítulo 2, sabe-se que modelos mais complexos podem correr o risco de sobreajuste aos dados específicos de treinamento, reduzindo sua capacidade de generalizar para novos dados.

Reconhecer essas limitações é fundamental não só para contextualizar os resultados encontrados, mas também para orientar novas pesquisas. Neste sentido, sugere-se, em futuras pesquisas: (i) explorar como as previsões aprimoradas afetam a segmentação de clientes e a personalização de produtos de seguro; e (ii) avaliar como o uso de conjuntos de dados maiores e mais complexos pode melhorar ainda mais a precisão dos modelos.

## 6. Referências Bibliográficas

Antonio, K.; Beirlant, J. (2007). Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics*, v. 40, n. 1. (pp. 58-76).

Balona, C.; Richman, R. (2020). The Actuary and IBNR Techniques: A Machine Learning Approach. *SSRN Electronic Journal*.

Blier-Wong, C. *et al*. (2020). Machine Learning in P& C Insurance: A Review for Pricing and Reserving. *Risks*, v. 9, n. 1, p. 4.

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, v. 16, n. 3.

Chen, T.; Guestrin, C. (2016). XGBoost. Proceedings of the 22nd ACM SIGKDD. *International Conference on Knowledge Discovery and Data Mining. Anais*. ACM.

Denuit, M.; Hainaut, D.; Trufin, J. (2019). *Effective Statistical Learning Methods for Actuaries I*. Cham: Springer International Publishing.

Guelman, L. (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, v. 39, n. 3. (pp. 3659-3667).

Izbicki, R.; Santos, T. M. (2020). *Aprendizado de máquina: uma abordagem estatística*. (1. ed.). Câmara Brasileira do Livro.

Mccullagh, P.; Nelder, J. A. (2019). *Generalized Linear Models*. [s.l.] Routledge.

Saputri, U.; Devianto, D. (2020). *The model of life insurance claims with actuarial smoothing approach by using GLM Poisson regression*.

Tzougas, G.; Kutzkov, K. (2023). Enhancing Logistic Regression Using Neural Networks for Classification in Actuarial Learning. *Algorithms*, v. 16, n. 2, p. 99.