CADERNOS DO IME – Série Estatística

Universidade do Estado do Rio de Janeiro - UERJ
ISSN on-line 2317-4536 / ISSN impresso 1413-9022 - v.56, p.39-65, 2024
DOI: 10.12957/cadest.2024.83639

PRECIFICAÇÃO DE SEGURO AUTOMÓVEL COM MACHINE LEARNING E MODELOS LINEARES GENERALIZADOS

Josemar C. Cabral
UFRJ – Universidade Federal do Rio de Janeiro
josemarcabral@hotmail.com

Eduardo Fraga L. de Melo
UERJ – Universidade do Estado do Rio de Janeiro
EMAp / FGV
Susep – Superintendência de Seguros Privados
eduardofraga@ime.uerj.br

Resumo

Neste trabalho, aplicamos modelos de Machine Learning (Árvores de Regressão, Random Forest, Boosting e XGBoost) para precificação ou tarifação de uma carteira de seguro automóvel e comparamos com modelos lineares generalizados (GLM) em dados de sinistros de seguro automóvel considerando as principais características do segurado e do veículo. Com base em critérios de avaliação de performance fora-da-amostra, os resultados indicaram que o XGBoost é o melhor método preditivo tanto para a frequência como para a severidade, apresentando ganhos na predição quando comparado ao GLM comumente utilizado em tarifação de seguros.

Palavras-chave: *Machine Learning, Seguro Automóvel, Tarifação, GLM.*

1. Introdução

Podemos afirmar que o processo de precificação e tarifação começou há séculos tendo em conta os relatos, segundo a história do seguro, de que os babilônios, gregos e chineses já utilizam mecanismos semelhantes aos contratos de seguros atuais para garantir as perdas no transporte de cargas. Assim, conforme o desenvolvimento e aperfeiçoamento verificado nos contratos de seguros desde o início do século XIV, também tem havido esta evolução no processo de precificação. Assim, é cada vez mais notório, nos dias de hoje, o uso de modelos matemáticos e ferramentas sofisticadas para o cálculo do prêmio de seguro, com o objetivo de que este seja suficiente para cobrir os custos com sinistros, os gastos com a operação e gerar margem de lucro.

Neste trabalho, aplicamos modelos de Machine Learning (Árvores de Regressão, Random Forest, Boosting e XGBoost) sobre uma base de dados de sinistros de seguro automóvel e comparamos os resultados com modelos lineares generalizados (GLM).

O nosso objetivo é comparar os resultados do processo de precificação utilizando algoritmos de Machine Learning (ML) e modelos lineares generalizados resultantes da manipulação dos dados de uma carteira de sinistros Auto. As variáveis de interesse são o número de ocorrências de sinistros e indenizações que podem ser explicadas em função da faixa etária e sexo do condutor, região, marca e ano de fabricação do veículo.

Assim, a partir do banco de dados disponibilizado publicamente pelo sistema AUTOSEG (SUSEP, 2023) informações obtidas através dos arquivos submetidos semestralmente pelas companhias de seguros sob supervisão da SUSEP (Superintendência de Seguros Privados), separamos, primeiramente, a base de dados em duas amostras sendo 60% para treino e 40% para teste ou previsão.

A base de dados contém as seguintes variáveis:

- Ano modelo: ano de fabricação do veículo;
- Categoria: a categoria tarifária;
- Região: contém o código e descrição das regiões de circulação;
- Sexo: descrição do sexo do condutor.
- Faixa etária: faixas etárias do condutor;
- Exposição: quantidade de veículos expostos;
- Freq ColisaoParcial: número de sinistros da cobertura colisão parcial;
- Ind ColisaoParcial: total de indenizações de sinistros da cobertura colisão parcial;

- IS Média: média das importâncias seguradas das apólices incluídas no agrupamento.

Este trabalho está organizado em 6 seções. Na seção 2, começamos por fazer uma revisão de literatura onde apresentamos as principais referências encontradas sobre trabalhos anteriores. Na seção 3, apresentamos a metodologia utilizada. Na seção 4, efetuamos análise preliminar dos dados descrevendo as classes de cada uma das variáveis da base de dados. Na seção 5, focamo-nos na análise dos modelos de frequência e severidade com a análise exploratória e a interpretação dos resultados produzidos pelo pacote computacional R (2024). Por fim, na seção 6, considerações finais são feitas.

2. Revisão de Literatura

Não obstante se verificar recentemente uma tendência maior para pesquisa na área de ML e GLM sobretudo no que diz respeito a precificação de produtos de seguros, há ainda uma escassez de trabalhos neste campo do saber, particularmente quanto à aplicação do ML no processo de tarifação de seguros. Todavia, algumas referências serão trazidas para este artigo, cada uma com a sua particularidade, abordando sobre o tema. Assim, esta seção visa apresentar um resumo da literatura disponível que, sobretudo, entendemos ser relevante para o tema em estudo.

No que concerne aos métodos de ML, a metodologia utilizada neste trabalho baseia-se no livro de Izbicki e dos Santos (2020) que aborda não somente os aspectos teóricos de cada um dos algoritmos de ML mais utilizados como também apresenta detalhes das bibliotecas e comandos do programa computacional R. No que diz respeito ao GLM, a metodologia utilizada segue Goldburd *et al*. (2016).

Em Izbicki e dos Santos (2020), o objetivo é de estreitar a relação e criar a ponte entre modelos de ML e a estatística, ou seja, a pesquisa visa principalmente analisar, discutir e trazer os pontos fortes de cada um dos métodos de ML e explorar a sua aplicação na estatística. Assim, neste trabalho utilizamos as metodologias descritas em Izbicki e dos Santos (2020) referentes aos métodos de ML para precificação de seguro automóvel e comparamos os resultados com o GLM, uma abordagem benchmark para seguros deste tipo.

Mendes *et al*. (2017) foca na comparação entre algoritmos de ML (Árvores de Regressão, Bagging, Random Forest, Boosting e Redes Neurais), modelos preditivos tradicionais (GLM e modelos aditivos generalizados) e técnicas de regressão penalizadas

(Lasso, regressão Ridge e redes elásticas) a partir de uma base de dados simulada de seguro de automóvel. Naquele artigo, os autores abordam sobre as vantagens e desvantagens de cada um dos procedimentos e concluem afirmando que nenhum método avaliado apresenta melhor desempenho relativamente aos restantes, tendo como base os indicadores em análise, apesar de que alguns métodos tenham apresentado um erro de predição menor. Isto significa que a base de dados em qualquer trabalho desempenha um papel primordial para a seleção do método mais preciso e confiável. Portanto, o campo de abrangência da pesquisa deste artigo torna-o útil e relevante para este trabalho.

Hanafy e Ming (2021) explora, a partir de uma base de dados enorme em seguros, vários métodos tais como Regressão Logística, XGBoost, Random Forest, Árvores de Regressão, Bayes Ingênuo e K-NN para prever ocorrência de sinistros e avaliar os resultados de cada um deles. Sendo que, os autores consideram o algoritmo Random Forest como o melhor modelo tendo em conta a sua robustez e capacidade preditiva.

Prieto (2005) é o outro trabalho que incluímos na revisão de literatura e que se aproxima ao objeto de pesquisa do nosso estudo, na medida em que aborda aspectos da precificação do seguro automóvel a partir de métodos de ML e aplicação do GLM. Em Prieto (2005), utiliza-se três procedimentos como a árvores de decisão, regressão logística e GLM para calcular a frequência e o custo médio com sinistros de colisão da base de dados de uma seguradora brasileira.

Com a base de dados de uma carteira de automóvel de contratos subscritos no período compreendido entre janeiro de 2004 a dezembro de 2004, segundo Prieto (2005), no que diz respeito ao algoritmo de árvores de decisão, as variáveis ano do veículo, bônus, tipo de veículo e a idade foram as mais significativas para estimar a frequência de sinistros. Relativamente ao custo médio com sinistros, as variáveis com maior destaque na explicação da variação da variável resposta foram a procedência do veículo, tipo do veículo, origem da apólice e fabricante do veículo, o que levou à conclusão de que a construção de árvores distintas para cada variável de interesse foi acertada.

Por outro lado, para ajuste do GLM, o autor utiliza o procedimento de Genmod (PROC GENMOD) do pacote SAS, que funciona com um parâmetro de escala associado ao parâmetro de dispersão ϕ da família exponencial. Jain (2018) aborda e compara três procedimentos de precificação de seguros a citar: GLM (frequência e severidade), Tweedie GLM e Redes Neurais Artificias. A comparação efetuada pelo autor é com base

nos critérios do teste MSE (Erro Quadrático Médio), AIC (Akaike Information Criterion), razão entre o prêmio de risco e a validação cruzada. O GLM apresentou o melhor resultado em termos de precisão seguido pelo Tweedie GLM e redes neurais respetivamente.

As referências apresentadas acima e a peculiaridade de cada uma delas é que permite o enriquecimento e crescimento da investigação no campo da precificação de seguros aplicando algoritmos de ML e modelos estatísticos considerados tradicionais, como é o caso do GLM. Este trabalho visa igualmente agregar mais opções a este campo de pesquisa ainda em desenvolvimento e, sobretudo, contribuir que o setor segurador possa acompanhar as novas tendências.

3. Metodologia

Antes de apresentarmos nesta seção os modelos utilizados no artigo, introduziremos os critérios de comparação de performance, que são o Erro Absoluto Médio (MAE) e a Raíz do Erro Quadrático Médio (RMSE).

O Erro Absoluto Médio (MAE) é uma medida utilizada para avaliar a precisão de um determinado modelo e é calculado da seguinte forma (1):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

Onde y_i, e $\hat{y}_i$ representam os valores observados e preditos respetivamente. Por sua vez, o RMSE é definido como (2):

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

3.1 Modelos de Machine Learning em Atuária

Desde suas origens, a inteligência artificial foi impulsionada pelo desejo de se compreender e revelar relações complexas nos dados, com o objetivo de desenvolver modelos capazes não apenas de realizar previsões precisas, mas também para extrair conhecimento de uma forma compreensível. Nessa busca, o campo de ML tem se diversificado consideravelmente, resultando em uma vasta gama de pesquisas que exploram diferentes aspectos e metodologias.

Dentro do espectro de métodos de ML, as técnicas baseadas em árvores de decisão destacam-se por sua eficácia e utilidade, oferecendo resultados tanto confiáveis quanto interpretáveis para uma ampla variedade de conjuntos de dados.

O desenvolvimento das árvores de decisão remonta a Morgan e Sonquist (1963), que introduziram o conceito através do método Detector de Interação Automático (DIA), destinado a lidar com efeitos não aditivos em análises de dados de pesquisa. Esse marco inicial foi seguido por significativas evoluções e a criação de programas computacionais dedicados à análise de dados, contribuições notáveis feitas por pesquisadores como, por exemplo, Messenger e Mandell (1972).

Entretanto, a evolução metodológica das árvores de decisão foi significativamente impulsionada por contribuições pioneiras de Moore (1987), Friedman *et al.* (1977), Friedman (2001) e Quinlan (1979, 1986). Estes pesquisadores enriqueceram substancialmente o campo de ML ao desenvolver algoritmos pioneiros para árvores de decisão. A escolha frequente das árvores de decisão e técnicas relacionadas advém de uma série de atributos vantajosos que as posicionam como ferramentas analíticas altamente eficientes e acessíveis:

- Natureza não-paramétrica: podem modelar relações complexas sem a necessidade de pressupostos iniciais sobre a distribuição dos dados;
- Flexibilidade com tipos de dados: habilidade para processar dados heterogêneos, sejam eles ordinais, categóricos ou uma combinação dos dois.
- Seleção intrínseca de variáveis: eficiência em identificar e utilizar apenas as variáveis mais relevantes, aumentando a robustez do modelo contra dados irrelevantes.
- Robustez contra outliers e erros: tolerância a anomalias nos dados, o que contribui para a construção de modelos razoavelmente estáveis.
- Alta interpretabilidade: facilidade de compreensão dos resultados por usuários com pouca ou nenhuma formação em estatística ou atuária, democratizando o acesso à análise de dados complexos.

E relevante enfatizar que as árvores de decisão constituem a base de fundação de´ uma variedade de algoritmos modernos, tais como Random Forest, Boosting (FREUND, 1995; FRIEDMAN, 2001) e XGBoost (CHEN & GUESTRIN, 2016), onde são usados como blocos para construções de modelos mais complexos. Para obter uma visão

completa das variadas técnicas em ML, é recomendável a leitura de Izbicki e dos Santos (2020).

Ainda neste contexto, a chamada aprendizagem atuarial representa um campo inovador que integra técnicas de ML e inteligência artificial `a ciência atuarial, aplicando algoritmos avançados e modelos computacionais para a análise de vastos volumes de dados atuariais, incluindo informações sobre seguros, previdência e riscos financeiros. Estas metodologias oferecem amplas aplicações no domínio atuarial, tais como:

- Precificação de seguros: emprego de algoritmos de aprendizagem de ML para estabelecer prêmios de seguros, levando em conta uma avaliação mais apurada dos riscos associados aos clientes.
- Análise de riscos e sinistros: utilização de algoritmos para prever riscos e identificar tendências que possam sinalizar fraudes ou padrões de sinistralidade.
- Gestão de investimentos e provisões: aplicação de modelos preditivos para aprimorar estratégias de investimento e a gestão de provisões técnicas.
- Personalização de planos: Desenvolvimento de planos de seguros personalizados, baseados em análises detalhadas das necessidades específicas de cada cliente.

Deste modo, a adoção de técnicas de ML está revolucionando o modo como os atuários conduzem suas análises, oferecendo insights mais aprofundados sobre riscos e padrões comportamentais, além de aprimorar significativamente as decisões estratégicas nos setores de seguros, saúde e previdência. Algumas referências notáveis neste campo emergente incluem Wüthrich e Merz (2023) e Denuit e Trufin (2019).

Dentro deste contexto, o ML surge como uma poderosa alternativa aos modelos lineares generalizados (GLM), ao acelerar a modelagem e previsão através da identificação de estruturas de dados complexos e frequentemente não lineares, sem a necessidade de suposições prévias sobre a relação entre covariáveis e a variável resposta, ou sobre as distribuições probabilísticas subjacentes. Esta abordagem permite a geração de modelos preditivos robustos, capazes de capturar a complexidade dos dados atuariais, pavimentando o caminho para inovações e eficiências sem precedentes na área.

De forma resumida, as abordagens de ML usadas neste artigo estão descritas a seguir.

3.2 Arvores de Regressão

Arvores de regressão ou de classificação oferecem métodos populares para medir importância de variáveis e separação binária. Para este método, referimos a Moore (1987) e Denuit e Trufin (2019). Este método seleciona características individuais para parcionamento do espaço de covariáveis, e a importância das variáveis é medida pela análise da contribuição de cada componente na queda total da função objetiva. A separação binária tem a vantagem que pode ser feita de modo aditivo.

3.2.1 Random Forest - RF

Arvores de regressão formam a base para diferentes métodos de funções de regressão que podem ser combinados. Desta forma, citamos Random Forest e Algoritmos de Boosting que essencialmente utilizam árvores de regressão. RF foram introduzidas por Breiman (2001).

Assim, uma segunda estratégia para lidar com a interação e os efeitos não lineares é empregar um modelo florestal aleatório. A base deste método está no desenvolvimento de árvores de regressão pelo algoritmo CART (MOORE, 1987) de árvores de classificação e regressão. Este algoritmo particiona sistematicamente o espaço variável do preditor através de um processo recursivo, buscando identificar divisões ótimas nos dados de treinamento, minimizando avidamente uma função de custo quadrática.

O processo começa extraindo repetidamente amostras com substituição do conjunto de dados de treinamento, gerando o que é conhecido como amostra bootstrap (EFRON & TIBSHIRANI, 1994). Esta amostra bootstrap serve como base para a construção de árvores de regressão altas individuais usando o algoritmo CART. Ao repetir este procedimento, é obtida uma coleção de árvores de regressão independentes.

Para aumentar ainda mais o poder preditivo, essas árvores individuais são agregadas usando uma técnica denominada bagging (BREIMAN, 1996). No bagging, múltiplas funções de regressão são treinadas a partir de amostras de bootstrap, e suas previsões são calculadas para criar uma função de regressão composta. Para otimizar este processo de agregação, é introduzido um refinamento. Em cada ponto de decisão dentro de cada árvore de regressão do conjunto, um subconjunto aleatório de preditores é selecionado sem substituição. Este aprimoramento dá origem ao que é comumente referido como Random Forest (BREIMAN, 2001).

Um exemplo de aplicação de árvores de regressão, Random Forest e Boosting em atuária e em mortalidade é encontrado em Levantesi e Pizzorusso (2019).

3.2.2 Boosting

Assim como Random Forest, o Boosting também consiste na agregação de diferentes estimadores da função de regressão. Contudo, esta combinação é feita de outra forma. Existem diversas variações e implementações de Boosting. O estimador é construído incrementalmente. Inicialmente, se atribuí o valor de 0. Este estimador possui alto viés, mas baixa variância (zero). A cada passo, se atualiza o valor do estimador de modo a diminuir o viés e aumentar a variância da nova função. Isto é feito adicionando-se ao estimador uma função que prevê os resíduos.

Uma forma de se fazer isso é com a utilização de uma árvore de regressão. E´ importante que essa árvore tenha profundidade pequena de modo a evitar o superajuste. Além disso, ao invés de simplesmente adicionar essa função por completo, adicionamos ela multiplicada por λ (chamado de learning rate): um fator entre 0 e 1 que tem por finalidade evitar o super-ajuste. Uma outra diferente implementação de Boosting foi feita neste artigo e tem acumulado fama por um bom desempenho, que é o XGBoost (CHEN & GUESTRIN, 2016).

3.2.3 XGBoost

XGBoost (também conhecido como eXtreme Gradient Boosting) é um esquema escalável de machine-learning para Boosting de árvores. As capacidades preditivas deste algoritmo têm sido amplamente reconhecidas em um número grande de desafios e competições de ML. Este método, referimo-nos a Chen e Guestrin (2016), enquadra-se na categoria de algoritmos de Boosting, que funcionam por meio da construção de um conjunto de modelos (geralmente árvores de decisão) de maneira sequencial. Cada novo modelo corrige os erros cometidos pelos modelos anteriores, ajustando os pesos das observações com base nos erros residuais.

O XGBoost combina as seguintes características principais: (i) o ajuste dos modelos é feito através da minimização do gradiente da função de perda, o que permite ajustar os parâmetros de maneira mais eficiente; (ii) o algoritmo inclui termos de regularização para penalizar a complexidade do modelo (normalmente a profundidade das árvores), o que ajuda a evitar o overfitting; (iii) um de seus grandes diferenciais é sua

implementação otimizada, com suporte a paralelismo, utilizando uma estrutura de memória eficiente para armazenar os dados, o que reduz a quantidade de computação necessária durante o treinamento; (iv) durante o processo de Boosting, o XGBoost ajusta gradualmente o aprendizado, controlado por uma taxa de aprendizado (learning rate), que regula o peso dado a cada novo modelo.

3.3 Modelos Lineares Generalizados

Nesta subseção apresentamos os modelos lineares generalizados aplicados para a frequência (número de sinistros) e para a severidade (valor dos sinistros).

3.3.1 Frequência

Para precificação utilizando GLM, a frequência é caracterizada como a contagem do número de sinistros. Assim, começamos por apresentar principais conceitos sobre a família de distribuição Poisson utilizada, neste trabalho, para o ajuste da frequência.

Dada uma variável resposta y, o GLM é dado por (3)

$$f(y) = c(y, \phi) exp\left\{ \frac{y\theta - a(\theta)}{\phi} \right\} \tag{3}$$

onde θ e ϕ são parâmetros canônico e de dispersão respetivamente.

E a função de ligação (4):

$$g(\mu) = X\beta, \tag{4}$$

onde X são as covariáveis (sexo do condutor, ano de fabricação, faixa etária do condutor, etc), μ é a média de y e β coeficiente do modelo.

Chamamos a expressão (3) de família exponencial.

A distribuição Poisson definida por (5):

$$f(y) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \quad y = 0, 1, \ldots \tag{5}$$

pertence à família exponencial e tem valor esperado $E(Y)$ igual a variância $Var(Y)$. Entretanto, quando acontece que:

$$E(Y) < Var(Y)$$

chamamos de sobredispersão. Para dados de contagem existe a família *quasipoisson* derivada da função quasi-verossimilhança que é dada por

$$Q(\lambda, \sigma^2, y) \propto \frac{1}{\sigma^2}\{y\log(\lambda) - \lambda\}$$

se $Var(Y) = \sigma^2\lambda$

Observação: Se $\sigma^2 = 1$, então $Q(\lambda,\sigma^2,y)$é proporcional a função log-verossimilhança da Poisson.

3.3.2 Severidade

A severidade é caracterizada como o montante ou valor do sinistro e este modelo está associado `a distribuições contínuas como verossimilhança.

A distribuição Gama utilizada no modelo para avaliar a severidade também faz parte da família exponencial e tem função densidade dada por (6):

$$f(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}, y > 0 \tag{6}$$

O modelo Gama tem a forma:

$$y \sim Ga(\mu,\phi)$$

$$g(\mu) = X\beta,$$

μ é a média de Y e ϕ é o parâmetro de dispersão. A função de ligação canônica é a função inversa $1/\mu$, isto é,

$$1/\mu = X\beta.$$

Usualmente considera-se a função de ligação log:

$$log(\mu) = X\beta$$

atendendo a dificuldade na interpretação dos parâmetros β.

4. Análise dos Dados

A partir da base de dados pública extraída do sistema AUTOSEG procedemos com alguns ajustes e mantemos apenas as variáveis que, do ponto de vista do mercado, são relevantes para os métodos aqui aplicados quando se trata de precificação do seguro

automóvel. Nesta seção, exploramos a base de dados descrevendo as variáveis nela contida e efetuamos a análise exploratória com intuito de encontrar, em uma primeira etapa, as covariáveis que apresentam mais relevância para a explicação da variável resposta.

Para a presente análise, as variáveis respostas para a frequência e severidade são *Freq ColisaoParcial* e *Ind ColisaoParcial* respectivamente, e a variável Exposição utilizada como offset no GLM com a finalidade de corrigir o tamanho dos grupos que podem ser muito distintos. Sendo as demais variáveis utilizadas como preditoras para explicação da variável resposta.

A partir da base de dados contendo as variáveis descritas na nossa introdução, procedemos com análise dos dados visando encontrar, do ponto de vista da análise, o melhor conjunto de dados quer seja para treinamento como para o teste. Assim, após a leitura dos dados efetuou-se uma série de manipulações até chegar aos dados definitivos, isto é:

- Exclusão das variáveis *Freq RouboFurto*, *Ind RouboFurto*, *Freq ColisãoPerdaTotal*, *Ind ColisãoPerdatotal*, *Freq Incendio*, *Ind Incendio*, *Freq Outros* e *Ind Incendio* da base de dados;
- Filtragem dos dados, com intuito de reduzir o tamanho da base de dados e tornar a sua leitura menos exaustiva consideramos somente a categoria pick up;
- Agrupamento dos anos atribuindo como ano 2018 os anos inferiores a esta data. Considerando que este intervalo contém cerca de 138.632 observações.

A manipulação efetuada na base de dados reduziu o número de variáveis e, consequentemente, o tamanho da amostra para 81.406 observações, visando não somente permitir uma abordagem mais prática como também uma melhor gestão do tempo. Foi possível obter os conjuntos de dados para treinamento e teste na medida em que a amostra de treino representa 60% e a de teste 40% da base de dados respetivamente.

Apresentamos abaixo as variáveis que fazem parte da base de dados e que para a precificação de uma carteira de seguro de automóvel são geralmente consideradas pelo mercado segurador. Entretanto, vale realçar que nem todas as variáveis listadas foram utilizadas ou tomadas em conta para a explicação das variáveis de interesse.

- **Região**

A localização ou trajeto habitual do veículo é nos dias de hoje um fator de preenchimento obrigatório na subscrição de um seguro automóvel. Em países mais desenvolvidos, o prêmio do seguro de um determinado modelo de veículo pode estar sujeito a agravamento ou bonificação atendendo a localidade, procedimento justificado por razões como: tamanho da população, fator de desenvolvimento ou qualidade das estradas, índice de criminalidade etc. Não é difícil, por exemplo, constatar que uma localidade com estradas ou vias devidamente sinalizadas e com padrões sofisticados de segurança rodoviária apresentará menor risco de ocorrência de colisão do que regiões menos desenvolvidas. Analogamente, podemos afirmar fluxo de pessoas deve ser uma variável também a ser considerada no processo de precificação.

- **Ano de Fabricação**

Esta é uma variável categórica que se deve ter em conta no processo de tarifação pelo fato do ano de fabricação dizer muito sobre as condições técnicas do veículo. Assim, é razoável concluir que veículos mais antigos podem representar maior risco para a seguradora e, portanto, pode ser uma variável relevante para o nosso estudo. Sexo

A variável categórica sexo costuma ser relevante no processo de precificação em alguns países incluindo o Brasil, na medida em que o gênero do condutor pode impactar na maneira como o veículo é dirigido. Não obstante ser cada vez maior a lista de países, principalmente europeus, em que o sexo do condutor não representa um fator diferencial devido, sobretudo, a questões do âmbito da igualdade de gênero, é para a nossa análise uma característica significativa no mercado segurador brasileiro.

- **Faixa Etária**

Esta é uma das variáveis mais importantes quando se trata do seguro automóvel. E de amplo conhecimento que o desempenho do ser humano em muitas tarefas ou atividades está intrinsecamente relacionado ao fator idade, pois se com o tempo o indivíduo adquire maturidade e experiência permitindo-lhe aperfeiçoar `as suas habilidades e competências sobre determinada tarefa, é também verdade que ao passar dos anos perde-se ou reduz-se habilidades cognitivas e capacidades motoras.

Assim, do exposto acima e adicionado outros fatores associados a idade da pessoa, consideramos a variável faixa etária visando perceber o conjunto ou intervalo de idades em que o condutor está ou não mais propenso ao risco.

- **Exposição**

Esta é uma variável quantitativa que representa o tempo de vigência de cada apólice durante o período de observação, que para o nosso estudo é o período semestral. Assim, o número de expostos é definido como o melhor estimador para a quantidade de veículos segurados.

- **Importância Segurada Média**

Conforme o glossário disponibilizado pelo sistema AUTOSEG, a variável importância segurada média representa a média do valor seguro dos contratos, ponderada pela exposição de cada uma das apólices. Esta é uma variável quantitativa que nos permite ter visibilidade sobre o capital segurado médio dos riscos subscritos pela companhia de seguro.

- **Frequência**

A frequência é a variável que representa o número de sinistros ocorridos num período em estudo. E para a presente abordagem concentramo-nos na quantidade de sinistros referente a cobertura colisão parcial.

- **Indenização**

Esta é uma variável que representa o montante pago com sinistros e similarmente ao caso da frequência, a indenização aqui referida diz respeito a cobertura colisão parcial.

4.1 Análise Exploratória da Frequência

Nesta seção tratamos de explorar as variáveis da base de dados visando sobretudo ter visibilidade a partir dos gráficos sobre quais variáveis são potenciais candidatas a incluir nos modelos para explicação das variáveis respostas para ambos os casos frequência e severidade. De realçar que nesta análise consideramos simplesmente algumas variáveis categóricas existentes na nossa base de dados. No R utilizamos a função model.matrix para transformar as variáveis categóricas em colunas de 0 e 1. Sendo que o número 0 indica que determinado registro não tem a característica observada e 1 indica que tem aquela característica.

Assim, começamos por efetuar a análise da relação da covariável sexo com a variável resposta, que é a frequência de sinistros da cobertura colisão parcial, aqui representada por Freq ColisaoParcial. Cabe destacar que as tabelas 1, 2 e 3 apresentam frequências relativizadas pela exposição. A Tabela 1 ajuda-nos a verificar que a covariável sexo pode ser significativa para explicação da variável resposta pelo fato do

gênero masculino apresentar uma proporção observada maior na distribuição da frequência do que o sexo feminino.

Tabela 1: Frequência Relativa - Variável Sexo

Sexo	Proporção
F	0,43
M	0,57

Fonte: Autores (2024)

No que concerne a covariável ano de fabricação, Ano Modelo, observa-se pela Tabela 2 um comportamento distinto entre as diferentes classes da covariável ano de fabricação relativamente ao número de sinistros, o que demonstra que o ano de fabricação do veículo impacta no número de ocorrência de sinistros e, portanto, é significativa para explicação da variável de interesse. No processo de transformação das variáveis categóricas em fatores de 0 e 1 a classe ano modelo 2022 ficou sem representatividade.

Tabela 2: Frequência Relativa - Variável Ano de Fabricação

Modelo	Proporção
2018 ou anterior	0,76
2019	0,11
2020	0,09
2021	0,04

Fonte: Autores (2024)

A outra variável a considerar no preditor é a faixa etária do condutor e, portanto, a análise da frequência mostra-nos que há aparentemente uma relação entre o número de sinistros e a idade do condutor pela distribuição da frequência nas diferentes classes definidas.

Tabela 3: Frequência Relativa - Variável Faixa Etária

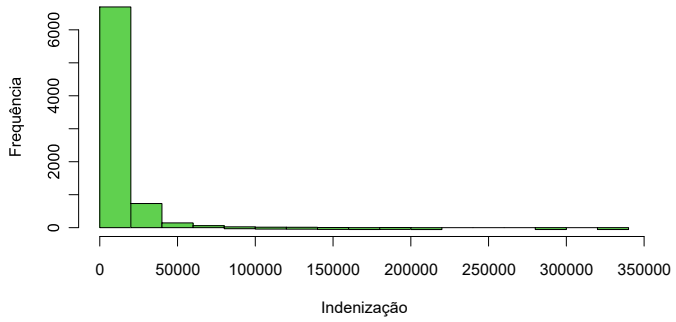
Faixa	Proporção
Abaixo de 26 anos	0,03
26 a 35 anos	0,18
36 a 45 anos	0,27
46 a 55 anos	0,25
Acima de 55 anos	0,27

Fonte: Autores (2024)

4.2 Análise Exploratória da Severidade

Analogamente ao caso da frequência, procedemos com análise exploratória dos dados de severidade a fim de examinar as variáveis a considerar no preditor dos nossos modelos. Primeiramente, tratando-se de uma variável resposta de caráter quantitativa começamos por observar o comportamento do histograma na escala real e logarítmica. Assim, a partir dos histogramas, podemos constatar a assimetria dos dados e no primeiro caso observamos que, quanto menor é o montante pago com sinistros, maior é a frequência (figura 1).

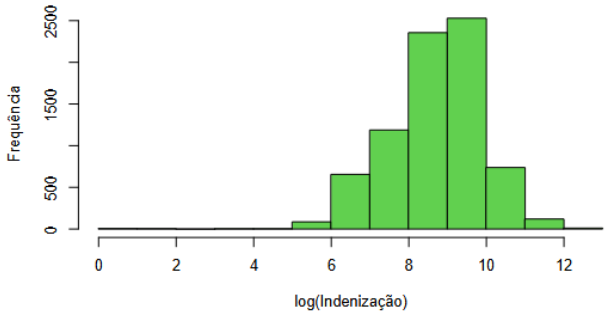
Figura 1: Histograma das Indenizações de Colisão Parcial



Fonte: Autores (2024)

No que concerne à escala do log das indenizações (figura 2) observamos um comportamento menos assimétrico dos dados.

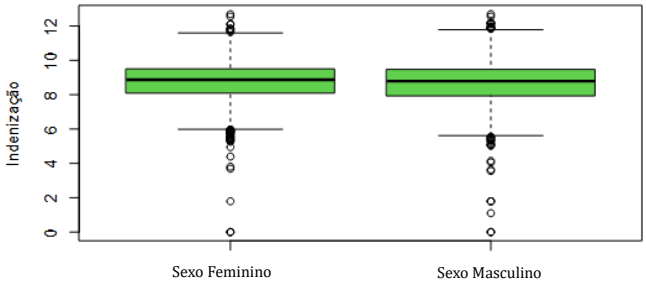
Figura 2: Histograma do Log das Indenizações de Colisão Parcial



Fonte: Autores (2024)

Apesar da Figura 3 não ilustrar com clareza a significância da covariável sexo para explicação da variável resposta que, no caso da severidade, trata-se das indenizações com sinistros e denotada por Ind ColisaoParcial, iremos considerá-la no nosso estudo. De realçar que os gráficos apresentados abaixo, referentes `a análise exploratória, utilizamos a escala logarítmica de modo a permitir melhor visibilidade quanto a distribuição.

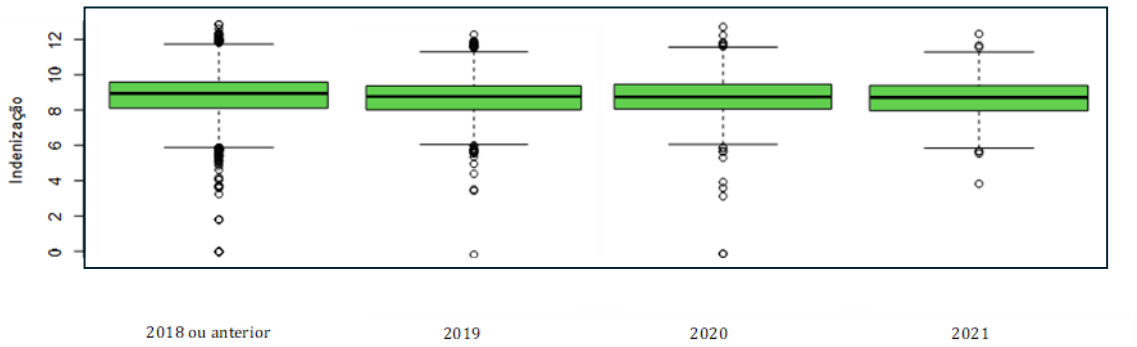
Figura 3: Indenizações x Sexo



Fonte: Autores (2024)

Outra variável a considerar no preditor é o ano de fabricação do veículo e apesar da figura 4 não ilustrar convincentemente sobre a significância de cada um dos fatores desta covariável em relação a variável resposta, iremos considerá-la inicialmente no modelo.

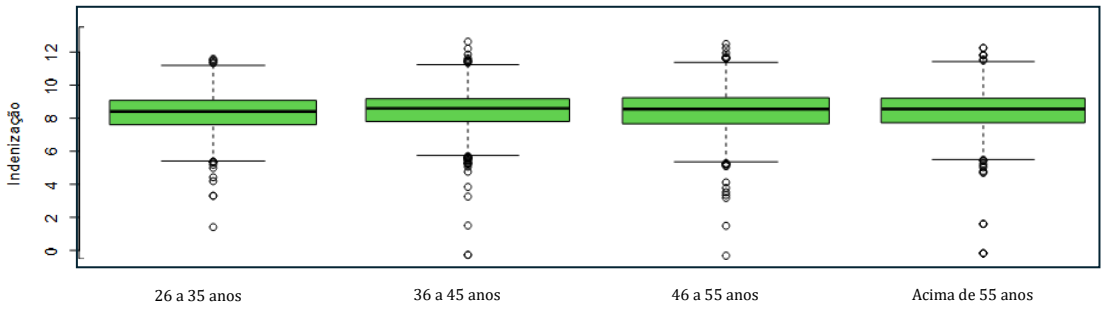
Figura 4: Indenizações x Ano de Fabricação



Fonte: Autores (2024)

A partir do conjunto de boxplots (figura 5), verificamos comportamentos distintos das classes da covariável faixa etária em relação aos montantes pagos com sinistros e, portanto, iremos considerá-la no preditor.

Figura 5: Indenizações x Faixa etária 26 a 35 anos



Fonte: Autores (2024

5. Resultados

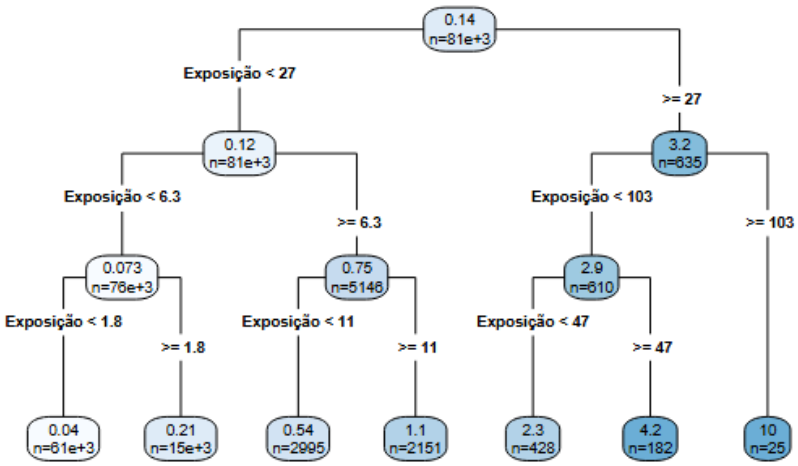
5.1 Frequência

Nesta subseção, apresentamos os principais resultados do pacote computacional R resultantes da aplicação dos modelos ou métodos utilizados.

5.1.1 Arvores de Regressão

A Figura 6 ilustra uma árvore de regressão em que a melhor variável explicativa é a exposição e podemos interpretá-la da seguinte forma: uma amostra de 81 mil observações a frequência média está em 0,14, se a exposição é menor que 27 mantém-se o número de observações com uma frequência média de 0,12. E se for maior que 27, o número de observações desce para 635 e com a frequência média de 3,2 e assim sucessivamente até atingir uma folha. A figura abaixo é obtida através da função rpart.plot da biblioteca com o mesmo nome. O método de separação usado na árvore de regressão foi ANOVA.

Figura 6: Ajuste Árvore de Regressão



Fonte: Autores (2024)

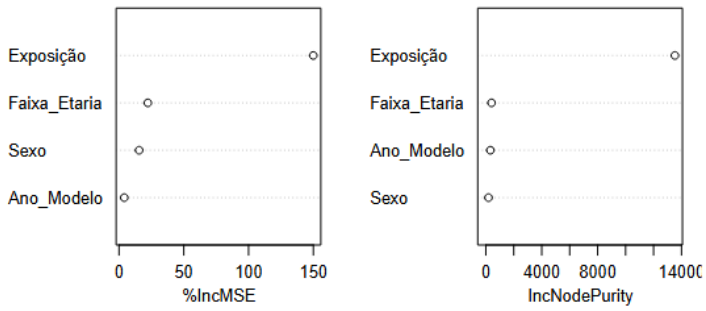
Para este método, o MAE e o RMSE estimados foram de 0,1849 e 0,4402 respetivamente.

5.1.2 Random Forest

O gráfico abaixo (figura 7) ilustra o grau de importância de cada covariável na predição da variável resposta segundo os critérios de diminuição média na precisão (%IncMSE) e diminuição média de Gini (IncNodePurity). Este modelo costuma ser facilmente interpretável e tem a particularidade de lidar com covariáveis discretas. A

partir do gráfico, verificamos, no que diz respeito ao critério %IncMSE, que a variável exposição apresenta um grau de importância maior do que as restantes variáveis, seguida pela faixa etária, sexo e a variável ano de fabricação. No que concerne ao critério IncNodePurity, a exposição mantém-se como a variável mais importante para o modelo, seguida pela faixa etária, ano de fabricação e o sexo. A função varImpPlot da biblioteca Random Forest do pacote computacional R permite-nos obter este gráfico com os dois critérios indicados acima. Para Random Forest foi usado o algoritmo de Breiman (baseado no código original de Breiman e Cutler, ver Breiman (2001)). Utilizamos número de variáveis aleatoriamente amostradas como candidatos em cada divisão como sendo a raiz quadrada do número de regressores.

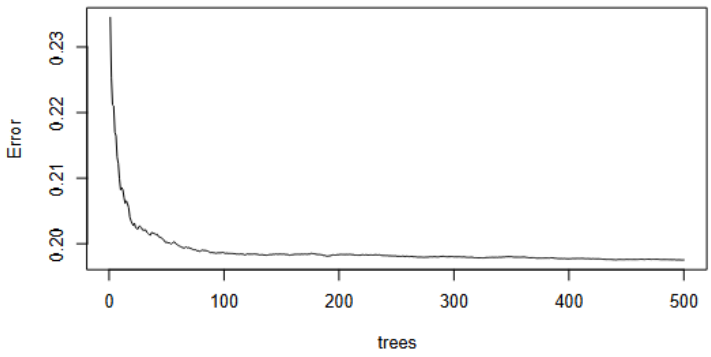
Figura 7: Importância das Variáveis ajuste Random Forest – Frequência



Fonte: Autores (2024)

A Figura 8 ilustra o comportamento decrescente do erro à medida que se registra o surgimento de mais árvores.

Figura 8: Erro em função do número de árvores - ajuste Random Forest



Fonte: Autores (2024)

O MAE e o RMSE para este método foram de 0,1832 e 0,4299 respetivamente.

5.1.3 Boosting

Analogamente aos algoritmos já abordados acima, o Boosting também agrega diferentes estimadores da função de regressão (Izbicki e dos Santos, 2020). No **R** utilizamos a função *bst*, pertencente a biblioteca com o mesmo nome, para implementação do Boosting e no que diz respeito a eficácia do modelo, consideramos ser bastante razoável na medida em que apresenta um MAE de 0,1805 e RMSE de 0,4301, que é ligeiramente inferior aos de Árvores de Regressão e Random Forest. Os parâmetros do melhor ajuste Boosting foram: (i) algoritmo de aprendizagem de árvore; e (ii) profundidade máxima das árvores de 6. O custo por falso positivo foi de 0,20.

5.1.4 XGBoost

O XGBoost é considerado como uma variação do Boosting e no **R** a sua implementação é feita pelo pacote com o mesmo nome (Izbicki e dos Santos, 2020). O MAE e RMSE encontrados foram de 0,1765 e 0,4277 respetivamente e, portanto, apresenta-se como o melhor método, no que diz respeito aos dois critérios, quanto a precisão para predição da frequência de sinistros. Os parâmetros do melhor ajuste XGBoost foram: (i) função objetiva Tweedie; (ii) taxa de aprendizado de 0,10 que previne *overfitting* (quanto menor essa taxa, mais robusto contra o *overfitting*, mas com lentidão maior no cálculo); (iii) redução de perda mínima para executar uma partição adicional na árvore de 0,50; (iv) profundidade máxima de uma árvore de 4; (v) regularização L2 (e L1) nos pesos igual a 1 (e 0).

5.1.5 GLM - Frequência

Diferentes ajustes de modelos GLM foram feitos tendo em conta a significância das covariáveis incluídas no preditor. De realçar que para a análise do número de sinistros tomamos a variável exposição como offset.

No primeiro ajuste efetuado, incluímos no preditor apenas uma covariável que é a variável sexo que mostrou ser significativa pois tem um p-valor inferior ao nível de significância de 5%, com AIC, para o modelo, de 50.909.

No segundo modelo, acrescemos a covariável faixa etária e verificamos que ambas são significativas para explicação da frequência de sinistros. O AIC encontrado para este ajuste é de 50.775.

No terceiro modelo, adicionamos a covariável ano de fabricação no preditor com exceção da classe de modelos referentes ao ano de 2022 pelo fato deste não apresentar

quaisquer registros que impactam sobre a nossa análise. As covariáveis consideradas neste ajuste apresentaram p-valores muito próximos de zero e consequentemente são todas significativas, com AIC de 50.640.

No que concerne ao critério AIC, verificamos que o terceiro modelo tem o menor AIC e, portanto, é o melhor ajuste para explicar a frequência de sinistros da cobertura colisão parcial. No quesito da análise da capacidade preditiva deste modelo, obtemos o MAE e RMSE na ordem de 3,1464 e 3,4476 respetivamente.

5.2 Severidade

Apresentamos os principais resultados do pacote computacional **R** resultantes da aplicação dos métodos de ML e GLM para a severidade.

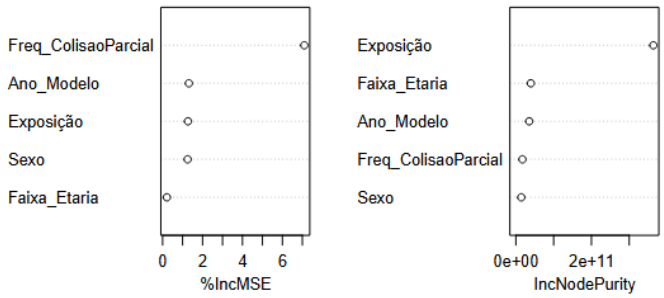
5.2.1 Arvores de Regressão

A estrutura dos modelos de ML no caso da severidade assemelha-se ao da frequência sendo a única diferença a variável resposta, que neste caso é o valor das indenizações. Assim, diferente do que verificamos no caso da frequência, não obtivemos, de acordo com o resultado, qualquer variável que explicasse bem o montante pago em indenizações com sinistros. Portanto, o modelo não apresentou resultados desejados. MAE e RMSE encontrados foram de 8.775,28 e 14.769,7. O método de separação usado na árvore de regressão foi Anova.

5.2.2 Random Forest

Destacamos o grau de importância das variáveis de acordo aos critérios indicados anteriormente. A frequência da cobertura colisão parcial tem o maior grau de importância na explicação da variável resposta pelo critério da diminuição média na precisão (%IncMSE). Por outro lado, no que diz respeito a diminuição média de Gini (IncNodePurity), a exposição representa a covariável de maior relevância seguida pela faixa etária. Para Random Forest para severidade (figura 9) foi usado o algoritmo de Breiman (baseado no código original de Breiman e Cutler, ver BREIMAN (2001)). Utilizamos número de variáveis aleatoriamente amostradas como candidatos em cada divisão como sendo a raiz quadrada do número de regressores.

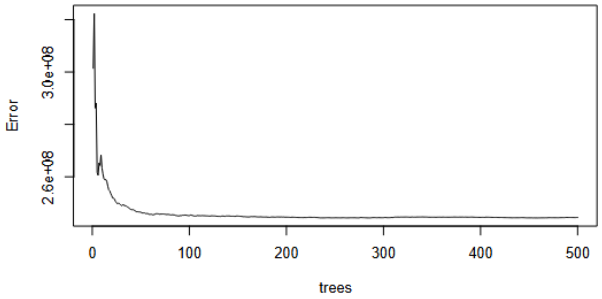
Figura 9: Importância das variáveis ajuste Random Forest – Severidade



Fonte: Autores (2024)

Similarmente ao que se observou no caso da frequência, aqui também o erro decresce à medida que aumenta o número de árvores, conforme ilustrado na Figura 10.

Figura 10: Ajuste Random Forest – Severidade



Fonte: Autores (2024)

O Random Forest estima, para o caso da severidade, um MAE na ordem de 8.895,93 e RMSE de 14.923,99.

5.2.3 Boosting

No caso do Boosting, apresentamos a relação entre os valores estimados e os observados visando averiguar a capacidade preditiva do modelo. No entanto, o MAE e RMSE encontrados foram de 8.848,31 e 14.972,62 respetivamente. Os parâmetros do melhor ajuste foram: (i) algoritmo de aprendizagem de árvore; e (ii) profundidade máxima das árvores de 6. O custo por falso positivo foi de 0,20.

5.2.4 XGBoost

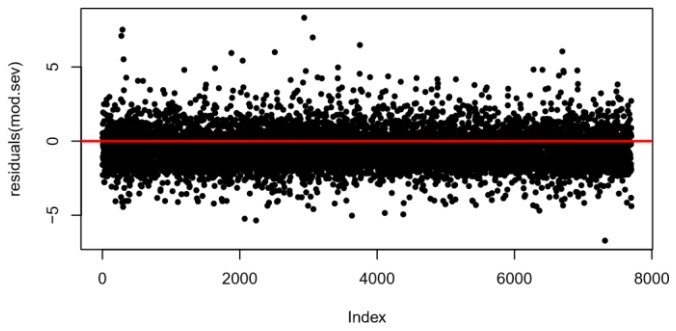
Igualmente ao resultado produzido para a frequência em que este método registou o menor erro absoluto médio, para a severidade confirma uma vez mais a capacidade preditiva deste modelo. O MAE estimado é de cerca de 8.744,48. Contudo, no que diz respeito ao RMSE, o XGBoost não apresentou uma capacidade preditiva melhor que os métodos anteriores.

Os parâmetros do melhor ajuste XGBoost foram: (i) função objetiva Tweedie; (ii) taxa de aprendizado de $0,10$ que previne *overfitting* (quanto menor essa taxa, mais robusto contra o *overfitting*, mas lentidão maior no cálculo); (iii) redução de perda mínima para executar uma partição adicional na árvore de $0,50$; (iv) profundidade máxima de uma árvore de 4; (v) regularização L2 (e L1) nos pesos igual a 1 (e 0).

5.2.5 GLM

No primeiro modelo, consideramos no preditor as variáveis categóricas sexo, ano de fabricação e faixa etária. E tendo em conta um nível de significância de 5% verificamos que as classes de perfis Ano de Fabricação 2019 e Ano de Fabricação 2020 são as únicas significativas. A Figura 11 representa os resíduos do modelo em questão.

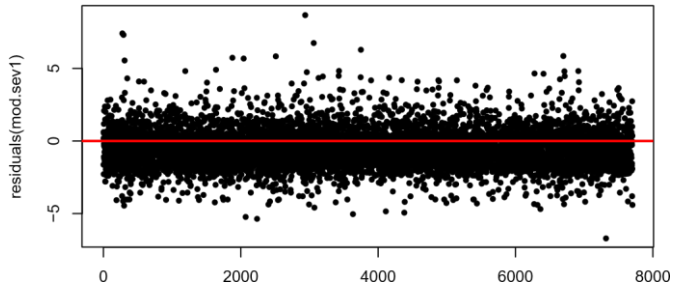
Figura 11: Resíduos - Modelo 1



Fonte: Autores (2024)

No segundo modelo (figuira 12), mantemos no preditor apenas as classes de perfis que mostraram serem significativas no ajuste anterior e, por conseguinte, obtivemos um ajuste com p-valores inferiores a 5%. Comparado ao AIC do ajuste anterior, este é ligeiramente superior.

Figura 12: Resíduos - Modelo 2

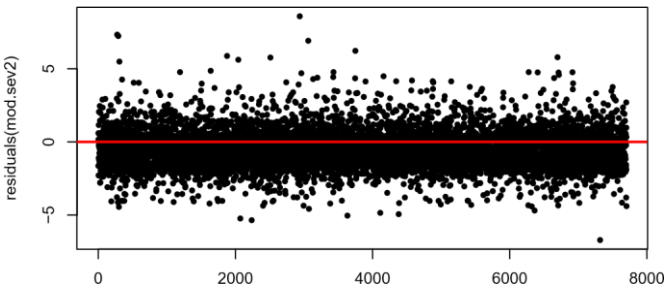


Fonte: Autores (2024)

No terceiro modelo (figura 13), adicionamos no modelo a classe de perfil Faixa Etária maior que 55 anos e, com exceção a esta variável, as outras duas mostraram, uma

vez mais, serem significativas para explicação das indenizações com sinistros. Este modelo apresenta, em relação aos anteriores, um melhor ajuste porque tem um AIC que é inferior aos dos modelos 1 e 2 de severidade, respetivamente. Sobre a análise da capacidade preditiva deste modelo, que é o melhor ajuste, o MAE e RMSE estimados foram de 10.691,82 e 18.332,79 respetivamente.

Figura 13: Resíduos - Modelo 3



Fonte: Autores (2024)

5.3 Resumo dos resultados

Nas Tabela 4 e 5, estão as métricas de performance para cada um dos modelos utilizados. Os menores MAEs e RMSEs estão em negrito para facilitar a identificação.

Tabela 4: Comparação pelo critério MAE

	Critério MAE	
Modelo	Frequência	Severidade
Arvores de Regressão	0,1849	8.775,28
Random Forest	0,1832	8.895,93
Boosting	0,1805	8.848,31
XGBoost	**0,1765**	**8.744,48**
GLM	3,1464	10.691,82

Fonte: Autores (2024)

Tabela 5: Comparação pelo critério RMSE

	Critério RMSE	
Modelo	Frequência	Severidade
Arvores de Regressão	0,4402	14.769,70
Random Forest	0,4299	14.923,99
Boosting	0,4301	14.972,62
XGBoost	**0,4277**	**14.637,11**
GLM	3,4476	18.332,79

Fonte: Autores (2024)

6. Considerações Finais

E cada vez mais notável a presença da inteligência artificial (IA) em diferentes setores de atuação e em nossa vida. O mercado de seguros, como um setor em expansão e relevante para a sustentabilidade da economia mundial, não é exceção à chegada de modelos de IA. E sendo o ML uma área de estudo pertencente à IA, ela também se apresenta como um instrumento inovador. Assim, o objetivo deste trabalho visou aplicar métodos de ML para tarifação de seguro automóvel e compará-los com o benchmark GLM, que é considerado um método tradicional no processo de precificação. Pelo que, apresentamos em forma de resumo, nas Tabelas 4 e 5, o desempenho de cada um dos modelos estudados, quer na abordagem da frequência como da severidade, com base nos critérios do erro absoluto médio e da raiz do erro quadrático médio. Tanto pelo critério MAE, quanto pelo RMSE, os resultados mostram que o XGBoost é o melhor modelo, tanto para frequência como para severidade.

Em geral, as variáveis da base de dados com maior impacto nos diferentes algoritmos e modelos aplicados foram o sexo, a faixa etária e o ano de fabricação do veículo. Em suma, constatou-se a partir do estudo efetuado e da revisão de literatura que há um leque considerável de métodos de ML sujeitos a estudo e discussão quanto as suas vantagens e desvantagens bem como a sua aplicabilidade no setor de seguros.

Recomendamos para futuros trabalhos que sejam exploradas outras classes de algoritmos de ML em produtos de seguros visando compará-los aos métodos estatísticos tradicionais em termos de precisão e confiabilidade.

Referências

BREIMAN, L. Bagging predictors. **Machine Learning**, 24:123–140. 1996.

BREIMAN, L. Random Forest. **Machine Learning**, 45:5–32. 2001.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. **In**: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785–794. 2016.

DENUIT, M.; TRUFIN, J. **Effective statistical learning methods for actuaries**. 2019.

EFRON, B.; TIBSHIRANI, R. J. **An introduction to the bootstrap**. Chapman and Hall/CRC. 1994.

Ekman, P.; Friesen, W. V.Manual for the facial action coding system. **Environmental Psychology & Nonverbal Behavior**. 1978.

FREUND, Y. Boosting a weak learning algorithm by majority. **Information and Computation**, 121(2):256–285. 1995.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **Annals of Statistics**, p.1189–1232. 2001.

FRIEDMAN, J. H. et al. A recursive partitioning decision rule for nonparametric classification. **IEEE Trans. Computers**, 26(4):404–408. 1977.

GOLDBURD, M.; KHARE, A.; TEVET, D.; GULLER, D. Generalized linear models for insurance rating. **Casualty Actuarial Society, CAS Monographs Series**, 5:77. 2016.

HANAFY, M.; MING, R. Machine learning approaches for auto insurance big data. **Risks**, 9(2):42. 2021.

IZBICKI, R.; DOS SANTOS, T. M. Aprendizado de máquina: uma abordagem estatística. Rafael Izbicki, 2020.

JAIN, N. Towards machine learning: Alternative methods for insurance pricing–poisson-gamma glm's, tweedie **GLM's and Artificial Neural Networks**. 2018.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R.; et al. **An introduction to statistical learning**, volume 112. Springer. 2013.

LEVANTESI, S.; PIZZORUSSO, V. Application of machine learning to mortality modeling and forecasting. **Risks**, 7(1):26. 2019.

MENDES, A.; DE VALERIOLA, S.; MAHY, S.; MARÉCHAL, X. **Machine learning applications to non-life pricing frequency modelling: An educational case study**,(2017) 1–25. 2017.

MESSENGER, R.; MANDELL, L. A modal search technique for predictive nominal scale multivariate analysis. **Journal of the American Statistical Association**, 67(340):768–772. 1972.

MOORE, D. H. **Classification and regression trees**, by leo breiman, jerome h. friedman, richard a. olshen, and charles j. stone. brooks/cole publishing, monterey, 1984, 358 pages, 27.95. 1987.

MORGAN, J. N.; SONQUIST, J. A. Problems in the analysis of survey data, and a proposal. **Journal of the American statistical association**, 58(302):415–434. 1963.

PRIETO, F. V. Precificação de seguros de automóvel. **Tese de Doutorado**, Universidade de São Paulo. 2005.

QUINLAN, J. R. Induction over large data bases, volume 79. **Computer Science Department,** Stanford University. 1979.

QUINLAN, J. R. **Induction of decision trees**. Machine learning, 1:81–106. 1986.

R. A better way to deploy R & Python. Disponível em: https://posit.co/. Acesso em: 2024.

SUSEP. AUTOSEG - Sistema de Estatísticas de Automóveis da SUSEP. Disponível em: https://www2.susep.gov.br/menuestatistica/Autoseg/principal.aspx. Acesso em: 2023.

WÜTHRICH, M. V.; MERZ, M. Statistical foundations of actuarial learning and its applications. **Springer Nature**. 2023.

MOTOR INSURANCE PRICING WITH MACHINE LEARNING AND GENERALIZED LINEAR MODELS

Abstract

In this work, we apply Machine Learning models (Regression Trees, Random Forests, Boosting, and XGBoost) for pricing or rate making of an motor insurance portfolio and compare the results with the ones obtained by a traditional generalized linear model (GLM) based on real claims data, considering the main characteristics of the policyholder and the vehicle. Based on out-of-sample performance evaluation criteria, results indicated that XGBoost is the best predictive method for both claims frequency and severity, showing improvements in prediction ability when compared to the commonly used GLM for insurance pricing.

Keywords: *Machine Learning; Motor Insurance; Pricing; GLM.*