Lisbon School of Economics & Management
Universidade de Lisboa

MESTRADO

ACTUARIAL SCIENCE

TRABALHO FINAL DE MESTRADO

RELATÓRIO DE ESTÁGIO

MODELOS DE DETEÇÃO DE COMPORTAMENTOS FRAUDULENTOS NA SUBSCRIÇÃO

VASCO JORGE ALEXANDRE

JUNHO 2026

Lisbon School of Economics & Management
Universidade de Lisboa

MESTRADO

ACTUARIAL SCIENCE

TRABALHO FINAL DE MESTRADO

RELATÓRIO DE ESTÁGIO

MODELOS DE DETEÇÃO DE COMPORTAMENTOS FRAUDULENTOS NA SUBSCRIÇÃO

VASCO JORGE ALEXANDRE

SUPERVISÃO

BEATRIZ HENRIQUES XAVIER

BEATRIZ DE ALMEIDA CURIOSO

TIAGO MARQUES FARDILHA

JUNHO 2026

Abreviaturas

ABI – *Association of British Insurers*

APS – Associação Portuguesa de Seguradores

GLM – *Generalized Linear Model* – Modelo Linear Generalizado

Lasso – *Least Absolute Shrinkage and Selection Operator*

RSS – *Residual Sum of Squares* – Soma dos Quadrados dos Resíduos

cp – *Complexity Parameter* – Parâmetro de Complexidade

XGBoost – *Extreme Gradient Boosting*

LightGBM – *Light Gradient Boosting Machine*

GOSS – *Gradient-based One-Side Sampling*

EFB – *Exclusive Feature Bundling*

TN – *True Negative* – Verdadeiro Negativo

FN – *False Negative* – Falso Negativo

FP – *False Positive* – Falso Positivo

TP – *True Positive* – Verdadeiro Positivo

MSE – *Mean Squared Error* – Erro Quadrático Médio

RMSE – *Root Mean Squared Error* – Raiz do Erro Quadrático Médio

MAE – *Mean Absolute Error* – Erro Absoluto Médio

SHAP – *Shapley Additive Explanations*

LIME – *Local Interpretable Model-agnostic Explanations*

Abstract

Fraud in the insurance sector is one of the main challenges faced by insurance companies, leading to significant financial losses that are ultimately reflected in the premiums paid by the entire community of policyholders. In motor insurance, beyond the fraud associated with the moment of the claim, fraudulent behavior at the underwriting stage has been increasing, particularly so-called early claims – claims reported shortly after the policy is issued and often linked to opportunistic fraud.

This report, resulting from an internship at Fidelidade – Companhia de Seguros, S.A., aims to identify, at the underwriting stage, the policies with the highest risk of early claims, using only the information available at that point. To this end, two complementary approaches were developed, classification and regression, and several linear and non-linear models were fitted: generalized linear models, decision trees, Random Forest, LightGBM and XGBoost. In the classification approach, the class imbalance was addressed through suitable metrics and the optimization of the decision threshold, and an economic analysis of the flagging strategies was also carried out.

The results show that more complex models generally achieve better predictive performance, although at the cost of lower interpretability, and that there is no universally superior model: the choice depends on the business objective, between maximizing claim detection and maximizing investigation efficiency.

Keywords: Insurance fraud; Early claims; Underwriting; Machine learning; Gradient boosting; Imbalanced data.

Resumo

A fraude no setor segurador representa um dos principais desafios para as companhias de seguros, traduzindo-se em perdas financeiras significativas que acabam por se refletir nos prémios pagos por toda a comunidade de segurados. No ramo automóvel, para além da fraude associada ao momento do sinistro, têm vindo a aumentar os comportamentos fraudulentos na fase de subscrição, com destaque para as *early claims* – sinistros reportados num curto período após a emissão da apólice e frequentemente associados a fraude oportunista.

O presente relatório, resultante de um estágio na Fidelidade – Companhia de Seguros, S.A., tem como objetivo identificar, no momento da subscrição, as apólices com maior risco de ocorrência de *early claims*, recorrendo apenas a informação disponível nessa fase. Para o efeito, foram desenvolvidas duas abordagens complementares, uma de classificação e uma de regressão, e ajustados diferentes modelos, lineares e não lineares: modelos lineares generalizados, árvores de decisão, Random Forest, LightGBM e XGBoost. Nos modelos de classificação, o desequilíbrio entre classes foi acautelado através de métricas adequadas e da otimização do *threshold* de decisão; tendo-se ainda realizado uma análise económica das estratégias de sinalização.

Os resultados mostram que os modelos mais complexos apresentam, em geral, melhor desempenho preditivo, embora à custa de menor interpretabilidade, e que não existe um modelo universalmente superior: a escolha depende do objetivo de negócio, entre maximizar a deteção de sinistros e maximizar a eficiência da investigação.

Palavras-chave: Fraude nos seguros; *Early claims*; Subscrição; *Machine learning*; Gradient boosting; Dados desequilibrados.

Índice

Índice de Figuras

Índice de Tabelas

Agradecimentos

A realização deste relatório de estágio não teria sido possível sem o apoio e o contributo de várias pessoas, às quais deixo o meu sincero agradecimento.

À Fidelidade – Companhia de Seguros, S.A., e em particular à equipa do Departamento de Pricing, Modelação e Estudos Técnicos, agradeço a oportunidade de realizar este estágio e a forma como fui acolhido. Deixo um agradecimento especial às minhas orientadoras, Beatriz Henriques e Beatriz Curioso, pela disponibilidade, pela partilha de conhecimento e pelo acompanhamento ao longo de todo o trabalho. Agradeço igualmente ao Fébio Levezinho, pelo apoio e pela constante partilha de informação, e ao João Pedro, por ter contribuído para o projeto, sem que nada o obrigasse a fazê-lo.

Ao meu orientador do ISEG, Tiago Fardilha, agradeço a orientação científica, o rigor e as sugestões que em muito enriqueceram este relatório.

A todos os meus professores, pelos conhecimentos transmitidos ao longo deste percurso e pela troca de ideias que tornou esta aprendizagem mais rica.

Por fim, deixo um agradecimento muito especial à minha família, pelo apoio incondicional e pelo incentivo constante, e à minha namorada, pelo apoio diário, por estar sempre ao meu lado e por compreender o tempo que esta etapa me exigiu. Obrigado por acreditarem sempre em mim – sem vocês, este caminho teria sido muito mais difícil.

A todos, o meu Muito Obrigado!

Nota Prévia

Este relatório de estágio de mestrado foi desenvolvido em estrita concordância das políticas e orientações de integridade académica estabelecidas pelo ISEG, Universidade de Lisboa. O trabalho aqui apresentado resulta da minha própria investigação, análise e redação, salvo indicação em contrário. Em prol da transparência, apresento a seguinte declaração sobre a utilização de ferramentas de inteligência artificial na elaboração deste relatório de estágio:

Declaro que foram utilizadas ferramentas de inteligência artificial durante o desenvolvimento deste relatório para auxiliar na tradução e na revisão bibliográfica. No entanto, toda a redação final, síntese e análise crítica são da minha autoria, pelo que a utilização de ferramentas de inteligência artificial não compromete a originalidade e a integridade deste relatório. Todas as fontes de informação, quer tradicionais, quer auxiliadas por inteligência artificial, foram devidamente citadas de acordo com as normas académicas. O uso ético da inteligência artificial na investigação e na escrita foi um princípio orientador ao longo da elaboração deste relatório.

1 Introdução

O presente relatório resulta de um estágio de quatro meses na empresa Fidelidade – Companhia de Seguros, S.A., no Departamento de Pricing, Modelação e Estudos Técnicos, enquadrado na Direção de Atuariado e Pricing.

A fraude no setor segurador constitui um dos principais desafios para as companhias de seguros, representando perdas financeiras significativas e aumento dos custos operacionais. Em 2024, o valor das fraudes detetadas pelas seguradoras em Portugal ascendeu a 87 milhões de euros, sendo o automóvel o segmento que continua com mais casos de fraude comprovada, com cerca de 40 milhões de euros em indemnizações indevidas (Castro & Panda, 2026).

No mesmo ano, a ABI (*Association of British Insurers*) identificou reclamações fraudulentas no valor de 1,16 mil milhões de libras no Reino Unido, sendo o ramo automóvel o que registou o maior valor, com 576 milhões de libras (ABI, 2025).

No ramo automóvel, para além da fraude associada ao momento do sinistro, tem-se verificado um aumento dos comportamentos fraudulentos ocorridos na fase de subscrição, destacando-se as chamadas *early claims*, isto é, sinistros reportados num curto período temporal após a emissão da apólice. Este tipo de ocorrência é frequentemente associado à fraude oportunista, na qual o tomador do seguro procura obter cobertura para um risco ou sinistro já existente e deliberadamente ocultado no momento da contratação.

Castro & Panda (2026) referem que, segundo a APS (Associação Portuguesa de Seguradores), os casos de fraude não detetada acabam por ser pagos pela comunidade de segurados e estão refletidos nos preços dos seguros. Dado este contexto, o principal objetivo deste trabalho não é punir os tomadores de seguros destas apólices fraudulentas, mas sim conseguir identificar corretamente estes casos no momento da subscrição. A deteção e prevenção de uma parte significativa destas situações poderá contribuir para a redução das taxas de risco e, consequentemente, para a diminuição dos prémios de todas as apólices.

Face à impossibilidade de modelar diretamente o comportamento fraudulento, o presente trabalho foca-se em modelar as *early claims* com base nos dados disponíveis no momento da subscrição, procurando identificar padrões que permitam antecipar situações de risco.

1.1 Organização do Texto

O presente trabalho encontra-se dividido em 6 capítulos.

No primeiro capítulo é feita uma pequena introdução e enquadramento do problema em estudo, apresentando breves informações e motivações sobre o assunto.

O segundo capítulo é dedicado à revisão de literatura, sendo abordados os principais conceitos relacionados com fraude no setor segurador, com um maior destaque para *early claims* nos seguros automóveis. São também abordados os diferentes modelos, softwares, métricas e abordagens utilizados durante o trabalho.

No terceiro capítulo são apresentados os dados utilizados, os procedimentos de preparação e tratamento da informação, assim como uma breve análise das variáveis a modelar.

O quarto capítulo incide sobre os diversos modelos e respetivos parâmetros, enquadrados no contexto do problema em análise, bem como o processo de seleção do *threshold* ótimo, as diferentes abordagens utilizadas para interpretar os modelos e, ainda, a forma de análise dos custos.

Por fim, no quinto e sexto capítulos são apresentados os resultados e as principais conclusões do trabalho, destacando-se os resultados obtidos, as limitações identificadas e sugestões para desenvolvimentos futuros.

2 Revisão de Literatura

2.1 Tipos de Fraude e Comportamentos Fraudulentos

A fraude no mundo dos sinistros automóveis é muito vasta, e não será possível abordar em detalhe todos os métodos e esquemas usados. Porém, pode ser dividida em duas grandes categorias: a fraude ocasional, que ocorre de forma esporádica de modo a beneficiar de uma situação pontual; e a fraude organizada, que envolve grupos que lucram com esquemas sistemáticos (Gonçalves, 2025).

Alguns tipos de fraude ocasional são:

- Inversão de responsabilidades – alteração ou falsificação das circunstâncias de um sinistro de forma a atribuir a culpa a outra pessoa ou entidade;
- Deturpação ou omissão de factos – no momento da comunicação do sinistro ou no momento da subscrição;
- Agravamento de prejuízos – quando o segurado exagera deliberadamente os danos sofridos ou inclui prejuízos inexistentes na reclamação apresentada.

Alguns tipos de fraude organizada são:

- Furto de veículos – para posterior comercialização, quer na sua totalidade, quer apenas as peças, podendo simultaneamente receber uma indemnização da seguradora em virtude do roubo;
- Incêndio ou dano intencional – destruição ou danificação propositada com o objetivo de obter uma compensação financeira indevida.

Considerando que surgem regularmente novas formas e alternativas de obter dinheiro das seguradoras de forma ilegal, estas devem preparar-se e adaptar-se à constante evolução. Com o progresso da Inteligência Artificial ficará cada vez mais difícil distinguir um sinistro honesto de um simulado ou inventado através da manipulação de imagens (Gonçalves, 2025).

2.2 Modelos e *Machine Learning*

Durante o estudo e modelação foram testados vários modelos, explicados de seguida, com o objetivo de prever a ocorrência de sinistros nos primeiros dias da apólice (*early claims*). Para o efeito, foram adotadas duas abordagens.

Na primeira, formulou-se um problema de **classificação**, no qual a variável objetivo assume apenas dois valores:

- 0 – não ocorreu sinistro nos primeiros 90 dias da apólice;
- 1 – ocorreu sinistro nos primeiros 90 dias da apólice.

Na segunda, formulou-se um problema de **regressão**, no qual a variável objetivo representa o número de sinistros ocorridos nos primeiros 90 dias da apólice.

2.2.1 Modelo Linear Generalizado

Os modelos lineares generalizados (GLM) constituem uma extensão dos modelos de regressão linear clássicos e representam um vasto conjunto de métodos estatísticos com múltiplas aplicações. A presente secção segue Dobson & Barnett (2018), onde podem ser consultados mais detalhes.

O que distingue um GLM de um modelo linear clássico é, primeiramente, a distribuição da variável resposta, que deixa de estar restrita à distribuição Normal e passa a poder pertencer a qualquer membro da família exponencial de distribuições (Poisson, Normal, Binomial, Gamma, entre outras). Diz-se que uma distribuição pertence à família exponencial quando a sua função de densidade pode ser escrita na forma canónica

$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)],$$

onde a, b, c e d são funções conhecidas. Quando $a(y) = y$, diz-se que a distribuição está na forma canónica e a função $b(\theta)$ é então designada por parâmetro natural da distribuição.

A segunda distinção reside na relação entre o valor esperado da variável resposta e as variáveis explicativas, que é estabelecida através de uma função de ligação (*link function*). Ao contrário do modelo linear clássico, em que o valor esperado da resposta é modelado diretamente como uma combinação linear das variáveis explicativas, num GLM essa combinação linear é

relacionada com o valor esperado por intermédio de uma transformação adequada, o que permite acomodar respostas cujo domínio não coincide com a reta real.

Formalmente, segundo Dobson & Barnett (2018), um modelo linear generalizado é definido a partir de um conjunto de variáveis aleatórias independentes $Y_1, \dots, Y_N$, cada uma com distribuição pertencente à mesma família exponencial, satisfazendo as seguintes propriedades: a distribuição de cada Y_i assume a forma canónica e depende de um único parâmetro θ_i, tal que

$$f(y_i; \theta_i) = \exp[y_i b(\theta_i) + c(\theta_i) + d(y_i)],$$

e todas as distribuições dos Y_i são da mesma forma. Em geral, os parâmetros θ_i não são de interesse; o objetivo é antes a modelação de um conjunto reduzido de parâmetros $\beta_1, \dots, \beta_p$, com $p < N$. Designando $\mathrm{E}(Y_i) = \mu_i$, onde μ_i é função de θ_i, pressupõe-se a existência de uma transformação de μ_i tal que

$$g(\mu_i) = \boldsymbol{x_i^T \beta},$$

onde g é a função de ligação, uma função monótona e diferenciável; $\boldsymbol{x_i}$ é o vetor $p \times 1$ de variáveis explicativas da i-ésima observação; e $\boldsymbol{\beta}$ é o vetor $p \times 1$ de parâmetros.

É possível concluir que um GLM é composto por três elementos:

- Um conjunto de variáveis resposta $Y_1, \dots, Y_N$ que partilham a mesma distribuição da família exponencial – a componente aleatória;
- Um conjunto de parâmetros e de variáveis explicativas que formam o preditor linear $\eta_i = \boldsymbol{x_i^T \beta}$ – a componente sistemática;
- A função de ligação g que relaciona o valor esperado de cada resposta com a respetiva componente sistemática.

No presente trabalho foram ajustados dois modelos lineares generalizados distintos, correspondentes a um problema de classificação e a um de regressão, descritos nas subsecções seguintes.

2.2.1.1 Distribuição Binomial

O primeiro, de classificação, assenta na distribuição Binomial com função de ligação *logit*, dando origem a um modelo de regressão logística, adequado à modelação de uma resposta de

natureza binária. Sendo $Y \sim Bin(n, \pi)$, com n conhecido e π o parâmetro de interesse, a função de probabilidade é

$$f(y;\pi) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \qquad y = 0,1,\dots,n,$$

que pode ser reescrita na forma

$$f(y;\pi) = \exp\left[y \log \pi - y \log(1-\pi) + n \log(1-\pi) + \log \binom{n}{y}\right].$$

Esta expressão está na forma canónica, sendo o parâmetro natural

$$\mathrm{b}(\pi) = \log(\pi) - \log(1-\pi) = \log\left(\frac{\pi}{1-\pi}\right).$$

A função de ligação canónica é a função *logit*, definida por

$$g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{x_i^T \beta}.$$

Resolvendo esta equação em ordem a π, obtém-se a forma logística para a probabilidade,

$$\pi = \frac{e^{x_i^T \beta}}{1+e^{x_i^T \beta}},$$

que garante que o valor modelado se mantém no intervalo $[0,1]$, como exigido por uma probabilidade. No presente trabalho, dado que o objetivo é a modelação de uma única variável binária, considera-se a distribuição Binomial no seu caso particular de Bernoulli, $B(\pi)$, correspondente a $n = 1$.

2.2.1.2 Distribuição de Poisson

O segundo, de regressão, assenta na distribuição de Poisson com função de ligação logarítmica, apropriado à modelação de dados de contagem. Sendo $Y \sim Po(\theta)$, a função de probabilidade é

$$f(y;\theta) = \frac{\theta^y e^{-\theta}}{y!} = \exp[y \log \theta - \theta - \log y!], \qquad y = 0,1,2,\dots,$$

Esta expressão encontra-se na forma canónica, sendo o parâmetro natural dado por $\log(\theta)$. A função de ligação canónica associada é a logarítmica definida por

$$g(\mu) = \log \mu,$$

o que conduz ao modelo de regressão de Poisson, $\log \mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$. A escolha da ligação logarítmica é natural, uma vez que garante que o valor esperado μ_i permanece positivo.

2.2.1.3 Regularização: Ridge e Lasso

Em modelos lineares, a presença de muitas variáveis explicativas ou de forte correlação entre elas pode conduzir a sobreajustamento e instabilidade nas estimativas. Para contornar este problema, recorrem-se a técnicas de regularização, que introduzem uma penalização na função objetivo, reduzindo a magnitude dos coeficientes (James, Witten, Hastie, & Tibshirani, 2021).

Regularização Ridge

A regularização Ridge introduz uma penalização quadrática sobre os coeficientes, conduzindo ao seguinte problema de otimização:

$$\min_{\beta}\{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\},$$

onde $\lambda \geq 0$ é o parâmetro de regularização que controla a intensidade da penalização.

Este método tem como principal efeito a redução da magnitude dos coeficientes, aproximando-os de zero, mas sem os anular completamente.

Regularização Lasso

A regularização Lasso utiliza uma penalização baseada no valor absoluto dos coeficientes:

$$\min_{\beta}\{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|\},$$

onde $\lambda \geq 0$ é o parâmetro de regularização que controla a intensidade da penalização.

Ao contrário da regressão Ridge, usar regressão Lasso pode forçar alguns coeficientes a serem exatamente iguais a zero, realizando simultaneamente regularização e seleção de variáveis. Torna-se relevante em contextos com um grande número de variáveis explicativas.

Em muitos casos, opta-se por combinações dos dois métodos, como o Elastic Net, que integra ambas as penalizações e permite beneficiar das vantagens de cada abordagem.

2.2.2 Validação Cruzada (*Cross-Validation*)

A avaliação do desempenho de um modelo em dados não observados é essencial para garantir a sua capacidade de generalização (James et al. 2021). A validação cruzada consiste numa técnica de reamostragem que permite estimar de forma mais fiável o erro fora da amostra e apoiar a escolha de modelos e parâmetros.

No presente trabalho, para o modelo GLM, foi utilizada a abordagem de k-*fold cross-validation*, na qual o conjunto de dados de treino é dividido em k subconjuntos de dimensão semelhante. Em cada iteração, o modelo é ajustado com base em $k - 1$ subconjuntos, sendo o subconjunto restante utilizado para validação, repetindo-se o processo k vezes.

O erro final do modelo é obtido pela média dos erros de validação em todas as iterações, fornecendo assim uma estimativa mais estável e robusta do desempenho fora da amostra. Esta abordagem permite ainda selecionar o valor ótimo de parâmetros de afinação, como o parâmetro de regularização, assegurando um equilíbrio adequado entre a capacidade preditiva e a complexidade do modelo.

2.2.3 Árvore de Decisão

As árvores de decisão são modelos preditivos baseados numa estrutura hierárquica de divisões sucessivas, onde os dados são divididos em subconjuntos distintos cada vez mais homogéneos. A presente secção segue James et al. (2021). Trata-se de modelos simples e fundamentais, que servem de base aos restantes modelos abordados neste trabalho.

Formalmente, o modelo corresponde a uma partição do espaço das variáveis explicativas num número finito de sub-regiões disjuntas "retangulares", a cada uma das quais é atribuída uma previsão (James et al. 2021).

O modelo é representado sob a forma de árvore, partindo de um nó raiz que engloba a totalidade das observações. A árvore subdivide recursivamente o espaço das variáveis explicativas em regiões distintas, sendo cada divisão representada por um nó interno que dá origem a dois ramos. Os nós terminais, designados por folhas, contêm a previsão final do modelo: nas árvores de regressão, esta corresponde à média da variável resposta das observações de treino pertencentes a essa região; nas de classificação, à classe mais frequente entre essas observações.

É uma abordagem descendente (*top-down*) por começar no topo da árvore, onde todas as observações pertencem a uma única região, e localmente ótima (*greedy*), por escolher em cada passo a melhor divisão imediata, sem ponderar o efeito que essa escolha terá nas divisões subsequentes.

As divisões são formadas por uma variável explicativa X_j e um ponto de corte s que divide a região corrente nos dois semiespaços $\{X \mid X_j < s\}$ e $\{X \mid X_j \geq s\}$. Sendo a divisão escolhida a que mais reduz a soma dos quadrados dos resíduos,

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \widehat{y_{R_J}})^2,$$

onde $\widehat{y_{R_J}}$ denota a média da variável resposta das observações de treino contidas na região R_j. O processo é repetido recursivamente em cada região até se atingir um critério de paragem.

Na implementação utilizada neste trabalho, esses critérios foram definidos pelos seguintes parâmetros:

- Número mínimo de observações por folha: uma divisão só pode ocorrer se deixar no mínimo este número de observações em cada um dos dois nós resultantes;
- Profundidade máxima da árvore: número de divisões permitidas, dado que árvores demasiado complexas podem conduzir a sobreajustamento;
- Parâmetro de complexidade (*cp*): uma divisão só é mantida se reduzir o erro global da árvore em pelo menos este valor; divisões cujo ganho seja inferior a este limiar não são realizadas, o que funciona como mecanismo de poda e controlo da complexidade.

A principal vantagem destes modelos reside na sua interpretabilidade (a estrutura em árvore traduz-se numa sequência de regras de decisão facilmente compreensíveis e visualizáveis) e na capacidade de lidar de forma natural com variáveis explicativas quantitativas e qualitativas, sem necessidade de pré-processamento extensivo. Em contrapartida, quando usadas isoladamente, apresentam um poder preditivo geralmente inferior ao de outros métodos de aprendizagem supervisionada.

Além disso, sofrem de variância elevada: basta dividir o conjunto de treino em duas partes e ajustar uma árvore a cada parte para obtermos resultados bastante diferentes, porque a árvore tende a captar ruído aleatório dos dados e a sobreajustar (*overfitting*) (James et al., 2021). Estas limitações motivam os métodos abordados nas secções seguintes.

2.2.4 Random Forest

O modelo Random Forest surge precisamente para corrigir as fraquezas identificadas acima. Trata-se de um método de *ensemble* que agrega um conjunto de árvores de decisão, em que cada árvore funciona como *weak learner*, combinando duas ideias:

- *Bagging* (*Bootstrap Aggregating*), que por sua vez tem duas partes: na fase de *bootstrap*, o algoritmo gera repetidamente novos conjuntos de treino por amostragem aleatória com reposição a partir do conjunto original, e treina uma árvore separadamente em cada um. Na fase de agregação, combinam-se as previsões das várias árvores para obter o valor final: em problemas de regressão escolhe-se a média, em problemas de classificação a classe maioritária. Como a média de um conjunto de previsões reduz a variância, é esta agregação que torna o modelo mais estável do que uma única árvore.
- Seleção aleatória de variáveis: Em cada divisão a árvore só pode escolher entre um subconjunto aleatório de m variáveis, retirado do total de p variáveis explicativas. Quanto ao valor de m, Hastie, Tibshirani, & Friedman (2009) recomendam, como valores por defeito, $m \approx \lfloor \sqrt{p} \rfloor$ em classificação e $m \approx \left\lfloor \frac{p}{3} \right\rfloor$ em regressão, devendo, ainda assim, ser tratado como um parâmetro de afinação ajustado ao problema em concreto.

A seleção aleatória de variáveis resolve um problema do *bagging* simples: se existir um preditor muito forte, quase todas as árvores o usariam no topo, ficando muito parecidas e correlacionadas; ao restringir as variáveis candidatas, as árvores tornam-se menos correlacionadas, o que torna a sua média menos variável e, por isso, mais fiável, pois, como referido em Hastie et al. (2009), a redução da variância de um *ensemble* depende não só do número de modelos mas também da correlação entre os seus erros; ao serem menos correlacionadas, as árvores tendem a cometer erros diferentes que se anulam parcialmente na média, aumentando a estabilidade da previsão final.

O modelo Random Forest suporta, naturalmente, os mesmos parâmetros que uma árvore de decisão, aos quais acresce o número de árvores do conjunto, B, e o número de variáveis candidatas em cada divisão, m. Sendo a previsão final dada por:

$$f(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x),$$

onde $f_b(x)$ representa o valor previsto pela árvore b.

De notar que um maior número de árvores tende a melhorar a estabilidade e a exatidão das previsões, traduzindo-se naturalmente num maior custo computacional. Em contrapartida, o ganho em exatidão tem como custo a perda de interpretabilidade do modelo, já que deixa de existir uma única árvore que se possa visualizar.

2.2.5 Gradient Boosting Machines

O *boosting*, tal como o *bagging*, é uma abordagem geral aplicável a vários métodos de aprendizagem que consiste em transformar *weak learners* em modelos com boa capacidade de previsão. A presente secção segue James et al. (2021). Ao contrário do *bagging*, não constrói as árvores de forma independente, mas sequencialmente: cada árvore é construída usando informação das árvores já existentes.

O principal objetivo desta abordagem é treinar consecutivamente árvores de decisão, em que cada árvore seguinte se foca em corrigir os erros residuais (a diferença entre os valores observados e previstos) das árvores anteriores; deste modo, cada nova árvore depende fortemente das anteriores. Cada uma destas árvores pode ser pequena, com apenas alguns nós terminais, pois, ao ajustar pequenas árvores aos resíduos, o modelo melhora gradualmente nas regiões onde tem pior desempenho.

Alguns dos parâmetros de *boosting* aplicados a árvores são:

- *Learning rate* ou *shrinkage*: número positivo pequeno, tipicamente entre 0,01 e 0,001, que controla a velocidade de aprendizagem do modelo;
- O número de árvores: este método pode sobreajustar facilmente se este valor for demasiado grande;
- Número de divisões por árvore: controla a complexidade de cada árvore e, com ela, a ordem de interação do modelo.

O *gradient boosting* é um tipo específico de algoritmo de *boosting* que otimiza o modelo minimizando uma função de perda através de uma abordagem de descida de gradiente (Hastie et al. 2009). Trata-se de uma abordagem iterativa, que adiciona sequencialmente árvores de decisão ao modelo atual, cada uma ajustada para corrigir os erros do modelo anterior. Ao

contrário do *boosting* tradicional, o *gradient boosting* concentra-se nos pseudo-resíduos, isto é, nos gradientes negativos da função de perda avaliados nas previsões correntes.

O método foi formalizado por Friedman (2001) que enquadrou o *boosting* como um problema de otimização numérica num espaço de funções. A cada iteração m, calcula-se o pseudo-resíduo de cada observação como o gradiente negativo da função de perda L avaliado na aproximação corrente F_{m-1},

$$\tilde{y} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1(x)}},$$

ajustando-se de seguida uma árvore de decisão a estes pseudo-resíduos por mínimos quadrados, e a aproximação é atualizada segundo

$$F_m(x) = F_{m-1}(x) + \nu\rho_m h(x; a_m),$$

onde $h(x; a_m)$ é a árvore ajustada na iteração m, ρ_m o respetivo coeficiente obtido por pesquisa em linha e ν o parâmetro de *shrinkage* (taxa de aprendizagem), com $0 < \nu \leq 1$. Este parâmetro regula a contribuição de cada nova árvore, abrandando a aprendizagem; juntamente com o número de árvores (iterações) M, constitui uma das principais formas de regularização do método (Friedman, 2001). Friedman (2001) mostra empiricamente que valores menores de ν tendem a produzir melhor desempenho, exigindo, em contrapartida, um M maior, com retornos decrescentes a partir de certo ponto. A complexidade de cada árvore é controlada pelo número de nós terminais J, que determina a ordem máxima de interação que o modelo consegue captar.

A distinção entre os problemas de regressão e de classificação resolve-se, no *gradient boosting*, pela escolha da função de perda L a minimizar, mantendo-se inalterada toda a restante mecânica do algoritmo. Neste trabalho, tanto o XGBoost como o LightGBM – implementações deste método, descritas nas subsecções seguintes – foram ajustados a ambos os casos.

Para a classificação binária recorre-se à perda logística (log-verosimilhança Binomial),

$$L(y, F) = \log(1 + e^{-2yF}), \quad y \in \{-1, 1\},$$

em que $F(x)$ se relaciona com a probabilidade da classe através de *log-odds* (Friedman, 2001), reproduzindo a ligação *logit* do modelo Binomial.

Para regressão adota-se a função de perda de Poisson, correspondente ao simétrico da respetiva log-verosimilhança,

$$L(y, F) = e^F - yF,$$

onde $\mu = e^{F(x)}$, o que reproduz a ligação logarítmica do modelo de Poisson. Esta correspondência reflete, no contexto dos modelos baseados em árvores, o mesmo princípio que rege os modelos lineares generalizados, em que a natureza da variável resposta é acomodada pela distribuição e respetiva função de ligação.

2.2.5.1 Extreme Gradient Boosting

O modelo Extreme Gradient Boosting (XGBoost) é uma implementação escalável e eficiente de *gradient boosting* proposta por Chen & Guestrin (2016), amplamente adotada pela sua rapidez e desempenho preditivo, que introduz várias inovações.

A primeira, e mais distintiva, é a utilização de um objetivo regularizado em que à função de perda é adicionado um termo que penaliza a complexidade de cada árvore,

$$L = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \text{ com } \quad \Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2,$$

onde l é a função de perda (diferenciável e convexa) que mede o desvio entre a previsão $\hat{y}_i$ e o valor observado y_i; f_k representa a k-ésima árvore do modelo; T é o número de folhas dessa árvore e w o vetor dos pesos atribuídos às folhas. Este termo de regularização suaviza os valores aprendidos e ajuda a evitar o sobreajustamento (Chen & Guestrin, 2016).

A segunda é o recurso a uma aproximação de segunda ordem da função de perda, que utiliza não só o gradiente, mas também a segunda derivada para otimizar o objetivo, permitindo derivar uma pontuação de qualidade das divisões aplicável a uma vasta gama de funções de perda.

Para além destes, o XGBoost incorpora técnicas adicionais de regularização e eficiência: o *shrinkage* e a subamostragem de colunas, esta última herdada das Random Forests, que previne o sobreajustamento de forma ainda mais eficaz do que a subamostragem de linhas (Chen & Guestrin, 2016). No plano algorítmico, destaca-se um algoritmo que trata de forma unificada valores em falta e entradas nulas atribuindo a cada nó uma direção por defeito aprendida a partir dos dados, e um procedimento de quantis ponderados para a proposta eficiente de pontos de divisão candidatos. Estas otimizações, combinadas com a exploração eficiente de cache e de computação fora de memória, tornam o sistema altamente escalável.

2.2.5.2 Light Gradient Boosting Machine

O modelo Light Gradient Boosting Machine (LightGBM) é uma implementação de *gradient boosting* desenvolvida por Ke et al. (2017), concebida para ser altamente eficiente quando a dimensão das variáveis e o volume de dados são elevados. Os autores identificam como principal limitação das implementações convencionais o facto de, para cada variável, ser necessário percorrer todas as observações a fim de estimar o ganho de informação de todos os pontos de divisão possíveis, o que é computacionalmente dispendioso (Ke et al., 2017).

Para ultrapassar este problema, o LightGBM assenta num algoritmo baseado em histogramas, que agrupa os valores contínuos das variáveis num número reduzido de intervalos discretos, reduzindo o custo de procura das divisões. Sobre esta base, os autores propõem duas técnicas inovadoras.

A primeira, *Gradient-based One-Side Sampling* (GOSS), parte da observação de que as observações com gradientes maiores contribuem mais para o ganho de informação; assim, mantém todas as observações com gradientes elevados e seleciona aleatoriamente apenas uma fração das observações de gradiente pequeno, aplicando-lhes um fator de correção para não distorcer a distribuição dos dados (Ke et al., 2017).

A segunda, *Exclusive Feature Bundling* (EFB), explora a esparsidade do espaço de variáveis agrupando variáveis mutuamente exclusivas, que raramente assumem valores não nulos em simultâneo, numa única variável, reduzindo assim o número efetivo de variáveis sem perda apreciável de informação.

Importa notar uma distinção adicional face a outras implementações: enquanto o crescimento das árvores se faz tipicamente por níveis, o LightGBM adota um crescimento orientado pelas folhas (*leaf-wise*), expandindo a cada passo a folha que mais reduz a perda. A combinação de GOSS e EFB permite ao LightGBM acelerar o treino do *gradient boosting* convencional em mais de vinte vezes, mantendo uma exatidão praticamente equivalente (Ke et al., 2017).

2.3 Software Utilizado

No desenvolvimento do presente trabalho foram utilizadas as diversas ferramentas computacionais para tratamento, manipulação, análise, modelação e confirmação dos dados e resultados:

SAS Enterprise Guide

O SAS Enterprise Guide constitui uma das principais ferramentas usadas para armazenamento, tratamento e manipulação de dados (SAS Institute Inc, 2023). Permite cruzar diferentes tabelas e variáveis de modo a obter uma melhor base de dados final com todas as variáveis consideradas mais importantes. Desta forma, foi possível agregar a informação proveniente de diversas fontes, tratar valores em falta e construir a tabela analítica que serviu de base à modelação. Trata-se de uma ferramenta amplamente utilizada no setor segurador, oferecendo um ambiente robusto e válido para a gestão de grandes volumes de dados.

RStudio

O *RStudio* foi utilizado para análise, desenvolvimento e visualização estatística, sendo a principal ferramenta para resolução de erros, adaptação e tratamento de dados, implementação e interpretação dos modelos preditivos utilizados e posterior análise dos resultados. Trata-se de um ambiente de desenvolvimento integrado para a linguagem R (R Core Team, 2024), que disponibiliza um vasto ecossistema de pacotes para *machine learning*.

2.4 Métricas de Avaliação de Modelos

A avaliação do desempenho dos modelos constitui uma etapa fundamental no decorrer do trabalho. Como referido anteriormente, foram abordados dois tipos de problemas, um de classificação binária e outro de regressão, cada um exigindo um conjunto próprio de métricas, descritas em seguida.

2.4.1 Métricas de Classificação

No problema de classificação, e em particular no contexto das *early claims* e de comportamentos potencialmente fraudulentos, a escolha das métricas de avaliação assume particular relevância devido ao desequilíbrio existente entre classes. Nestes cenários, métricas tradicionais, como a *accuracy* podem conduzir a interpretações incorretas do desempenho do modelo, uma vez que um classificador que privilegie sistematicamente a classe maioritária pode atingir valores elevados sem identificar corretamente os casos de fraude, que constituem a classe de interesse (James et al., 2021).

Tratando-se de um problema de classificação binária, todas as métricas consideradas derivam da matriz de confusão (*confusion matrix*), que confronta os valores previstos pelo modelo com os valores reais observados (Sokolova & Lapalme, 2009):

		Valor Previsto	
		1	0
Real	1	Verdadeiro Positivo – (TP)	Falso Negativo – (FN)
	0	Falso Positivo – (FP)	Verdadeiro Negativo – (TN)

Tabela 1: Matriz de confusão para classificação binária

Os resultados da classificação podem ser repartidos por quatro grupos:

- TP (*True Positive*) – observações que o modelo classificou como 1 e que são efetivamente 1;
- FN (*False Negative*) – observações que o modelo classificou como 0, mas que são efetivamente 1 – Erro **Tipo II**;
- FP (*False Positive*) – observações que o modelo classificou como 1, mas que são efetivamente 0 – Erro **Tipo I**;
- TN (*True Negative*) – observações que o modelo classificou como 0 e que são efetivamente 0.

A partir destas quatro quantidades definem-se as métricas utilizadas neste trabalho.

A *accuracy* representa a proporção de observações corretamente classificadas pelo modelo relativamente ao total de observações analisadas, traduzindo a eficácia global do classificador:

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

A *precision* mede a proporção de observações classificadas pelo modelo como positivas que correspondem efetivamente a valores positivos, refletindo a fiabilidade das previsões positivas:

$$Precision = \frac{TP}{TP+FP}$$

O *recall* representa a proporção de casos positivos reais que foram corretamente identificados pelo modelo, traduzindo a sua capacidade de detetar a classe de interesse:

$$Recall = \frac{TP}{TP+FN}$$

O *F-score* (F_β) combina a *precision* e o *recall* numa única medida, através de uma média harmónica ponderada, permitindo atribuir diferentes níveis de importância a cada métrica:

$$F_\beta = \frac{(1+\beta^2) * Precision * Recall}{(\beta^2 * Precision) + Recall}$$

O parâmetro β controla o peso relativo atribuído a cada métrica:

- $\beta > 1$ atribui maior importância ao *recall*;
- $\beta < 1$ atribui maior importância à *precision*;
- $\beta = 1$ corresponde ao F_1 *score*, onde ambas têm igual importância.

No presente trabalho, optou-se por privilegiar a *precision* em relação ao *recall*, recorrendo a um valor de $\beta < 1$ no cálculo do *F-score*. Desta forma, garante-se que os casos sinalizados como fraudulentos correspondem, com elevada fiabilidade, a fraudes efetivas, reduzindo o número de falsos positivos encaminhados para análise.

2.4.2 Métricas de Regressão

No problema de regressão, a variável objetivo corresponde ao número de sinistros ocorridos nos primeiros dias da apólice, tratando-se, por isso, de uma variável de contagem. O desempenho dos modelos é avaliado com base na diferença entre os valores previstos $\hat{y}_i$ e os valores reais observados y_i.

A métrica de erro mais comum é o erro quadrático médio (MSE), que corresponde à média dos quadrados dos erros de previsão, penalizando de forma mais acentuada os erros de maior magnitude (James et al., 2021).

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$$

A raiz do erro quadrático médio (RMSE), corresponde à raiz quadrada do MSE, sendo expressa na mesma unidade da variável resposta, o que facilita a sua interpretação:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}$$

Por sua vez, o erro absoluto médio (MAE), por não elevar os erros ao quadrado é menos sensível à presença de valores extremos:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|$$

Recorreu-se também ao desvio de Poisson, que generaliza a soma dos quadrados dos resíduos aos modelos lineares generalizados e mede o desajustamento entre os valores observados e os valores ajustados pelo modelo (Dobson & Barnett, 2018).

Para a distribuição de Poisson, é dada por:

$$\text{D} = 2\sum_{i=1}^{n}\left[y_i \log\frac{y_i}{\widehat{y_i}} - (y_i - \widehat{y_i})\right].$$

A calibração de um modelo traduz o grau de concordância entre os valores previstos e os observados ao longo da escala de risco. Uma forma comum de avaliar consiste em ordenar as observações pelas previsões e agrupá-las em *bins* de igual dimensão, definidos por quantis. Para cada *bin*, calcula-se a média dos valores previstos, a média dos valores observados e a respetiva taxa de sinistralidade.

As médias dos valores previstos e observados são representadas graficamente, utilizando a bissetriz (reta $y = x$) como referência, por corresponder à calibração perfeita. Quanto mais próximos os pontos estiverem dessa reta, maior a concordância entre valores previstos e observados e, consequentemente, melhor a calibração do modelo. Esta análise baseia-se no conceito de autocalibração, particularmente relevante na avaliação de modelos de tarifação em seguros (Denuit, Charpentier, & Trufin, 2021).

3 Base de Dados

3.1 Construção

A base de dados deste trabalho foi construída no SAS Enterprise Guide, a partir da informação interna da Fidelidade. Consideraram-se todas as novas apólices iniciadas entre 1 de janeiro de 2024 e 30 de setembro de 2025, cruzando-as posteriormente com os respetivos sinistros ocorridos entre 1 de janeiro de 2024 e 31 de dezembro de 2025. A esta estrutura juntaram-se variáveis provenientes de outras tabelas, com o objetivo de reunir o maior número possível de informação relevante para os modelos.

Por uma questão de simplicidade e de aproximação à realidade, foram consideradas apenas as apólices com uma única unidade de risco. Os sinistros ocorridos em 2026 não foram considerados, evitando assim que os resultados fossem influenciados pelo comboio de tempestades registado nesse período.

3.2 Descrição

A base de dados é composta por cerca de 580.000 observações, referentes a aproximadamente 512.000 apólices distintas, que podem ou não registar a ocorrência de um sinistro nos primeiros dias de vigência. De notar que o número de observações é superior ao número de apólices, uma vez que cada acidente adicional numa apólice, no período em estudo, dá origem a uma nova linha na tabela.

As variáveis organizam-se nos seguintes grupos:

- **Apólice:** características e coberturas da apólice;
- **Cliente:** características do tomador do seguro e do condutor;
- **Veículo:** características do veículo segurado;
- **Comercial:** condições de distribuição e contratação da apólice;
- **Acidente:** ocorrência de acidente e, quando aplicável, custo associado e coberturas;
- **Histórico:** informação histórica do veículo e do cliente;
- **Derivadas:** variáveis transformadas ou calculadas a partir das anteriores.

Combinando variáveis numéricas e categóricas, a base resultante é sólida, representativa e diversificada, revelando-se adequada à análise do problema em estudo.

Para garantir confidencialidade, todas as variáveis estão sob nomes genéricos, sem referência a rótulos internos ou comerciais, e os dados foram anonimizados, mantendo-se a distribuição original.

3.3 Análise Exploratória dos Dados

Antes da seleção de variáveis e do ajuste dos modelos, a base de dados foi exportada para o RStudio para uma análise exploratória. Nesta fase examinaram-se o tipo de cada variável, a distribuição dos dados, a formatação, a presença de erros e a potencial relevância de cada variável para o estudo.

Todas as variáveis foram convertidas para os tipos de dados adequados. Em alguns casos, procedeu-se ainda à agregação de classes e à criação de uma categoria "*Desconhecido*", destinada a acolher as observações sem valor atribuído. Esta abordagem reduziu o número de categorias, facilitando a interpretação dos resultados e contribuindo para maior estabilidade dos modelos.

De seguida, construíram-se as variáveis alvo, distintas consoante a abordagem de modelação.

No problema de classificação, calculou-se, para cada apólice, a diferença entre a data do primeiro acidente e a data de início do risco e, sempre que esta foi inferior ou igual a um número de dias previamente definido, atribuiu-se à variável *sinistro* o valor 1. Nos restantes casos, (diferença superior a esse limiar ou ausência de registo de acidente) considerou-se que não houve *early claim*, fixando-se a variável em 0.

No problema de regressão, utilizou-se a diferença calculada anteriormente e contou-se o número de sinistros ocorridos, por cada apólice, antes do limiar de dias definido, registando-se esse valor na variável *nr_sin*.

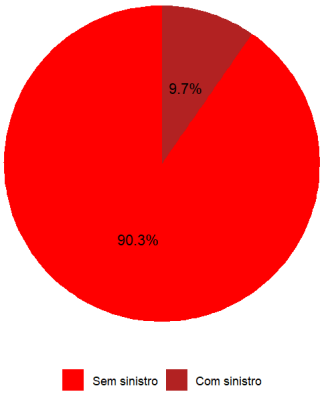

Figura 1: Distribuição variável objetivo sinistro

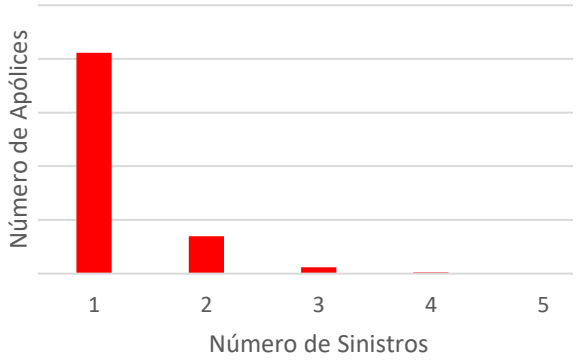

Figura 2: Distribuição variável objetivo nr_sin

A Figura 1 apresenta a distribuição da variável binária *sinistro*, enquanto a Figura 2 detalha a distribuição da contagem *nr_sin*. Como é possível observar, e como tem vindo a ser referido ao longo do trabalho, existe um desequilíbrio de classes, com apenas 9,7% das apólices a registarem pelo menos um sinistro nos primeiros dias de vigência. De notar ainda que, entre as apólices com sinistro, a grande maioria registou apenas uma ocorrência, diminuindo rapidamente a frequência para contagens superiores.

Foi também analisada a frequência de sinistros em função da duração da apólice (Figura 3).

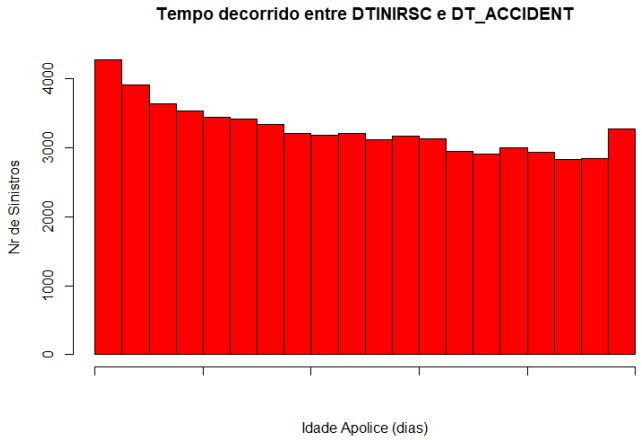

Figura 3: Relação entre número de sinistros e idade da apólice em dias

Observa-se uma cauda claramente decrescente: os sinistros concentram-se nos primeiros dias após o início do risco e tornam-se progressivamente menos frequentes à medida que a apólice envelhece. Este padrão sustenta o interesse no estudo das *early claims*, por ser precisamente neste período inicial que se espera maior risco de fraude na subscrição.

Recorrendo à análise gráfica univariada, analisou-se depois a maioria das variáveis explicativas face à variável-alvo *sinistro*. Tal permitiu identificar desde logo aquelas cujo comportamento diferia entre classes, fornecendo uma primeira perceção das que poderiam revelar-se mais importantes na construção dos modelos preditivos.

As figuras seguintes apresentam a taxa média da variável *sinistro* em cada categoria da variável em análise, estando ainda representada, a tracejado, a média global de *sinistro*.

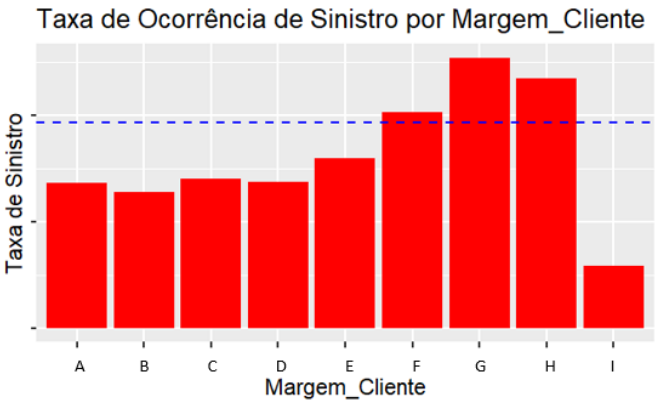

Figura 4: Taxa de Ocorrência de Sinistro por categoria da variável Margem_Cliente

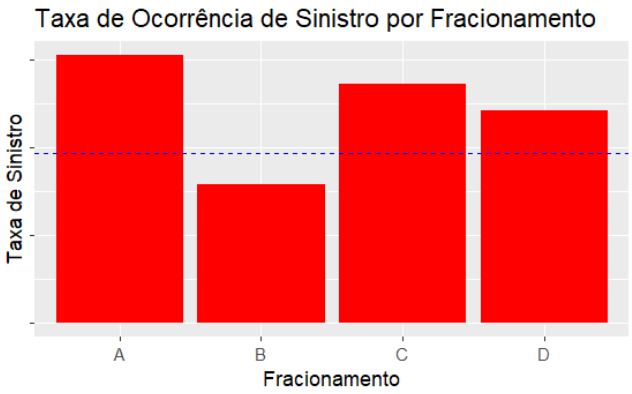

Figura 5: Taxa de Ocorrência de Sinistro por categoria da variável Fracionamento

Na variável *Margem_Cliente* (Figura 4), que representa o valor esperado da rentabilidade do cliente, dividido entre classes, a taxa de ocorrência de sinistro tende a aumentar à medida que a margem diminui: as categorias com margens inferiores (F, G e H) exibem taxas acima da média global, o que aponta para uma possível associação positiva entre margens reduzidas e a ocorrência de *early claims*. As categorias com margem superior (A, B, C e D) situam-se abaixo dessa média.

A variável *Fracionamento* (Figura 5), que representa a periodicidade de pagamento do prémio de seguro, revela um padrão semelhante: as categorias A, C e D apresentam taxas superiores à média global, com destaque para a categoria A, que atinge o valor mais elevado, ao passo que a categoria B regista uma taxa relativamente baixa. Tal sugere que um diferente fracionamento poderá estar associado a uma maior propensão a sinistros nos primeiros dias.

3.4 Seleção de Variáveis

Neste passo analisaram-se em detalhe, e removeram-se, todas as variáveis que, por diferentes motivos, não deveriam integrar a base usada na construção dos modelos.

Em primeiro lugar, excluíram-se as variáveis que continham informação indisponível no momento da subscrição da apólice, garantindo que os modelos se baseiam apenas em dados efetivamente conhecidos na fase de decisão.

Em seguida, removeram-se variáveis fortemente correlacionadas entre si, dado que a multicolinearidade pode distorcer os coeficientes em alguns modelos. Foram também removidas as variáveis redundantes. Optou-se, assim, por conservar apenas as variáveis de maior relevância, o que melhora a estabilidade dos modelos e a interpretação dos resultados.

Concluída a base de dados, subsistem ainda algumas variáveis numéricas com valores em falta. Estas serão removidas no modelo GLM, uma vez que este não admite observações com dados ausentes.

3.5 Separação entre Treino e Teste

Selecionadas as variáveis a considerar, a fase seguinte consistiu em dividir a base de dados em dois conjuntos disjuntos: treino e teste. O propósito desta separação é treinar e ajustar os modelos no conjunto de treino e avaliar o seu desempenho num conjunto distinto – o conjunto de teste. Trata-se de uma prática comum que previne o sobreajustamento e assegura um bom desempenho em dados nunca observados, garantindo a capacidade de generalização dos modelos.

Cerca de 80% da base foi destinada ao treino e os restantes 20% ao teste. A divisão respeitou a proporção média da variável objetivo em ambos os conjuntos, assegurando uma representação equilibrada das classes e, com ela, um ajuste e uma avaliação mais robustos, bem como maior consistência na comparação de resultados.

4 Metodologia

Após a construção da base de dados e a sua divisão em conjuntos de treino e teste, procedeu-se ao ajuste dos modelos considerados utilizando o conjunto de treino.

Foram desenvolvidas duas abordagens complementares ao problema. Na primeira, de classificação, cada modelo procura estimar a probabilidade de ocorrência de um sinistro nos primeiros dias da apólice, representada pela variável binária *sinistro*. Na segunda, de regressão, o alvo passa a ser o número de sinistros ocorridos nesse período, *nr_sin*, modelado como uma variável de contagem. Ambas as abordagens permitem, no final, atribuir a cada apólice uma medida de risco – uma probabilidade estimada, no caso da classificação, ou um número esperado de sinistros, no caso da regressão.

Por fim, foi avaliado o desempenho do modelo no conjunto de teste através das métricas descritas anteriormente.

4.1 Modelos de Classificação

4.1.1 Modelo Linear Generalizado

O primeiro passo para ajustar o modelo GLM foi remover as variáveis com valores em falta dos conjuntos de treino e teste. De seguida, todas as variáveis categóricas foram transformadas em variáveis binárias, convertendo os dados para uma estrutura matricial esparsa.

O modelo foi ajustado com o método *cv.glmnet*, recorrendo a uma regressão logística com distribuição Binomial. Foram usados três parâmetros para otimizar o compromisso entre a capacidade preditiva e a complexidade do modelo: *nlambda* $= 50$, *nfolds* $= 3$ e $\alpha = 0{,}5$.

O parâmetro λ controla a intensidade da penalização aplicada aos coeficientes do modelo, permitindo reduzir a complexidade e minimizar o risco de sobreajustamento. O modelo escolheu o melhor valor usando validação cruzada, entre 50 valores distintos.

O parâmetro α define o tipo de regularização utilizada: sendo regularização *Ridge* ($\alpha = 0$) e regularização *Lasso* ($\alpha = 1$). Valores intermédios, como o utilizado, representam uma combinação simultânea das propriedades de ambas as técnicas (*Elastic Net*).

4.1.2 Árvore de Decisão

O segundo modelo ajustado foi uma árvore de decisão, usando o conjunto de treino sem alterações e desenvolvido através da função *rpart*.

Foram definidos dois parâmetros: $cp = 0{,}0005$, que define o quanto uma nova divisão precisa de melhorar o modelo para ser executada, e $minbucket = 100$, que impõe um número mínimo de observações em cada nó terminal. O parâmetro *cp* controla a complexidade e o crescimento da árvore, valores mais baixos permitem mais divisões e captação de padrões mais específicos, dando origem a uma árvore mais complexa, embora mais ajustada aos dados, enquanto o *minbucket* limita esse crescimento, evitando folhas com poucas observações.

4.1.3 Random Forest

O modelo seguinte foi o Random Forest, no qual foi também utilizado o conjunto de treino sem alterações. Foi desenvolvido através da função *ranger*, com os parâmetros: $num.trees = 200$, $min.node.size = 10$, $importance = permutation$ e $probability = True$.

Foram utilizadas 200 árvores de decisão, permitindo aumentar a estabilidade e robustez das previsões, o número de observações mínimo em cada nó terminal da árvore foi fixado em 10, e a opção $probability = True$ permite obter diretamente probabilidades de ocorrência de sinistro. A importância das variáveis foi avaliada por permutação, medindo o impacto da alteração aleatória de cada variável no desempenho do modelo, permitindo identificar os fatores com maior influência e capacidade preditiva.

4.1.4 LightGBM

À semelhança do modelo GLM, foi necessário preparar a base de dados para garantir a compatibilidade com o algoritmo. Tanto no conjunto de treino como no conjunto de teste, as variáveis categóricas foram convertidas em variáveis numéricas através da transformação de fatores em números inteiros e os dados foram estruturados em forma de matriz. Após a transformação, foram criados os objetos no formato específico do LightGBM.

O modelo foi configurado como um problema de classificação binária, através da função *lgb.train* e ajustado da seguinte forma:

- *objective* = *binary* – função objetivo de classificação binária;
- *metric* =*average_precision* – métrica de avaliação, própria para classes desequilibradas;
- *learning_rate* = 0,03 – controla a contribuição incremental de cada árvore;
- *num_leaves* = 64 – define a complexidade das árvores;
- *max_depth* = −1 – não impõe restrições à profundidade das árvores;
- *min_data_in_leaf* = 20 – assegura um número mínimo de observações por folha;
- *feature_fraction* = 0,8 – apenas 80% das variáveis são utilizadas em cada iteração;
- *bagging_fraction* = 0,8 e *bagging_freq* = 5 – introduzem amostras aleatórias de observações de forma periódica;
- *nrounds* = 2000 – define o máximo de iterações durante o ajuste.

4.1.5 XGBoost

Por fim, foi ajustado o modelo XGBoost. Uma vez que o XGBoost, tal como o LightGBM, requer inputs numéricos, todas as variáveis categóricas foram convertidas em numéricas e os dados estruturados em forma de matriz. Foram em seguida criados os objetos do tipo *xgb.DMatrix*, desenhados para otimizar o desempenho computacional do XGBoost.

O modelo foi configurado como um problema de classificação binária, através da função *xgb.train*, e ajustado da seguinte forma:

- *objective* = *binary:logistic* – função objetivo de classificação binária;
- *eval_metric* = *aucpr* – área sob a curva precision-recall como métrica de avaliação;
- *eta* = 0,03 – controla a contribuição de cada árvore para o modelo final;
- *max_depth* = 4 – define a profundidade máxima das árvores usadas;
- *min_child_weight* = 10 – controla o peso mínimo necessário em cada nó;
- *gamma* = 2 – ganho mínimo para realizar uma divisão, penalizando a complexidade;
- *subsample* = 0,8 – percentagem de observações utilizadas em cada iteração;
- *colsample_bytree* = 0,8 – percentagem de variáveis consideradas em cada árvore;
- *nrounds* = 500 – define o máximo de iterações durante o ajuste.

4.1.6 *Thresholds*

Após o ajuste dos modelos de classificação, os valores de probabilidade obtidos das observações pertencentes ao conjunto de treino foram convertidos em classes binárias através da aplicação de *thresholds*, da seguinte forma:

$$f(x) = \begin{cases} 1, & x \geq t \\ 0, & x < t \end{cases}$$

em que t é o *threshold* escolhido, x é a probabilidade prevista pelo modelo e $f(x)$ é a classe final atribuída.

Para a escolha do *threshold* ótimo, foram aplicados múltiplos valores consecutivos de t ao conjunto de treino, avaliando-se qual destes maximizava a métrica F_β. Este procedimento foi repetido para diferentes valores de β.

Optou-se por privilegiar valores de $\beta < 1$, dando maior peso à *precision* relativamente ao *recall*. Esta decisão prende-se com o facto de, do ponto de vista operacional, ser preferível deixar passar alguns casos potenciais do que sinalizar um número excessivo de apólices para análise manual. Assim, considera-se mais aceitável a ocorrência de erros do Tipo II do que de erros do Tipo I.

O *threshold* selecionado, que varia consoante o modelo e o valor de β, foi depois usado na classificação do conjunto de teste, sendo os resultados utilizados na avaliação do desempenho do modelo.

4.2 Modelos de Regressão

4.2.1 Modelo Linear Generalizado

O modelo de regressão de base foi um GLM com distribuição de Poisson, ajustado com a função *glm* e *family* $=$ *poisson* sobre o conjunto de treino após remoção das variáveis com valores em falta. Ao contrário da versão de classificação, não foi aplicada regularização, sendo todos os coeficientes estimados por máxima verosimilhança.

4.2.2 Árvore de Decisão

Foi ajustada uma árvore de regressão com a função *rpart*, usando $cp = 0{,}00005$ e $minbucket = 50$. À semelhança da árvore de classificação, o *cp* controla o crescimento da árvore e o *minbucket* impõe um número mínimo de observações por folhas.

4.2.3 Random Forest

O Random Forest de regressão foi desenvolvido com a função *ranger*, utilizando $num.trees = 200$, $min.node.size = 10$ e $importance = permutation$. As 200 árvores aumentam a estabilidade das previsões, o número mínimo de observações por nó terminal foi fixado em 10 e a importância por permutação novamente usada para identificar as variáveis mais relevantes.

4.2.4 LightGBM

Após a preparação dos dados em formato matricial e a criação dos objetos *lgb.Dataset*, o LightGBM foi ajustado com a função *lgb.train* para um problema de regressão de contagem, utilizando $objective = poisson$, $metric = rmse$ e $nrounds = 2000$.

4.2.5 XGBoost

Por fim, depois de convertidas as variáveis categóricas em numéricas e criados os objetos *xgb.DMatrix*, o XGBoost foi ajustado com a função *xgb.train*, utilizando $objective = count{:}poisson$ e $nrounds = 1000$, ajustado numa lógica de regressão de contagem.

4.2.6 Seleção de Apólices

Ao contrário dos modelos de classificação, que produzem uma classe binária a partir de um *threshold*, os modelos de regressão produzem um número esperado de sinistros por apólice. A sinalização de apólices para revisão foi, por isso, feita por ordenação: foram selecionadas as apólices com previsões acima de quantis elevados da distribuição das previsões (0,925, 0,95 e 0,975), correspondendo às apólices de risco mais alto segundo cada modelo.

Para avaliar a calibração dos modelos desenvolvidos, aplicou-se o procedimento descrito anteriormente: as apólices do conjunto de teste foram ordenadas segundo as previsões e agrupadas em *bins* definidos por quantis. Em cada *bin*, a média prevista foi representada no eixo das abcissas e a média observada no eixo das ordenadas, avaliando-se a concordância entre o risco previsto e o observado pela proximidade dos pontos à bissetriz. Adicionalmente, foi analisada a taxa de sinistralidade de cada grupo.

4.3 Avaliação e Custos

Por fim, os resultados obtidos no conjunto de teste foram comparados com os valores reais observados e avaliados através das métricas previamente definidas (de classificação ou de regressão, consoante a abordagem). Estes resultados foram posteriormente agregados em tabelas, de forma a facilitar a análise e a comparação entre modelos.

Adicionalmente, procedeu-se a uma análise preliminar dos custos associados às decisões tomadas. Para tal, todos os prémios foram convertidos para períodos temporais comuns, sendo o cálculo efetuado de forma proporcional. Esta conversão foi feita em duas perspetivas:

- Considerando apenas o horizonte de observação dos sinistros ou até à data de cancelamento, consoante o que ocorresse primeiro – avaliando o impacto imediato da estratégia sobre as *early claims*;
- Considerando todo o período em observação, que varia entre apólices – averiguando se as apólices sinalizadas correspondem, de facto, a clientes com sinistralidade elevada a longo prazo.

Em ambos os casos, foi efetuada a soma dos prémios que seriam potencialmente perdidos em resultado da aplicação da estratégia adotada, comparando-se esse valor com os custos que seriam evitados através da redução de sinistros.

4.4 Abordagens de Interpretabilidade

4.4.1 Importância das Variáveis

De forma a compreender as escolhas e os cálculos dos modelos, procedeu-se à avaliação da importância das variáveis explicativas. Esta análise permitiu identificar os atributos com maior contributo para o poder preditivo dos modelos, fornecendo uma compreensão adicional dos padrões subjacentes aos dados. Adicionalmente, serviu de apoio à seleção de variáveis ao longo do trabalho, orientando a inclusão e a remoção de variáveis explicativas da base de dados.

4.4.2 Árvores de Decisão

Uma vez que os modelos de *machine learning* apresentaram melhor desempenho preditivo, e a importância das variáveis não é, por si só, suficiente para os interpretar, procedeu-se a uma análise adicional com o objetivo de aumentar a interpretabilidade dos resultados. Para tal, foram ajustadas árvores de decisão aos valores previstos pelos modelos, como modelos auxiliares, com o intuito de identificar nós ou regras associadas a uma maior taxa média de ocorrência de sinistro.

Esta abordagem permitiu aproximar o comportamento dos modelos mais complexos através de estruturas mais interpretáveis, facilitando a identificação de padrões relevantes. Adicionalmente, contribui para uma comunicação mais clara dos resultados e para a justificação das decisões de seleção de apólices para revisão, tornando mais transparentes os critérios subjacentes às decisões do modelo.

5 Interpretação e Resultados

5.1 Interpretação dos Modelos

A análise da importância das variáveis revela uma assinalável consistência entre os diferentes modelos desenvolvidos, tanto na classificação como na regressão.

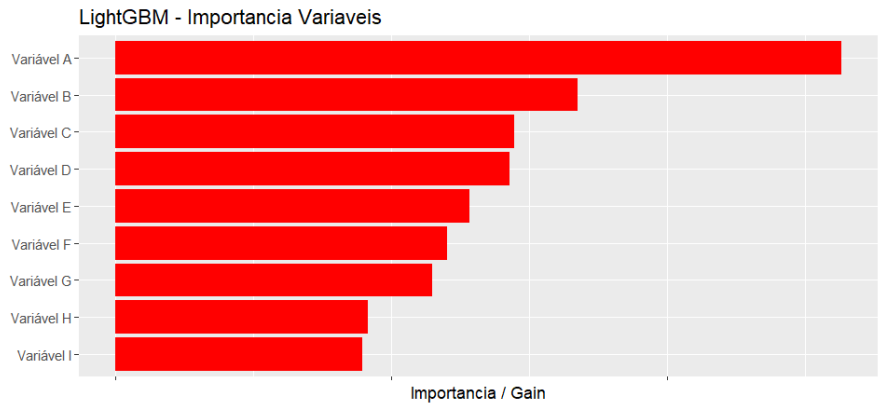

Figura 6: Gráfico de importância das variáveis do modelo LightGBM

A Variável A, evidenciada na Figura 6, destaca-se sistematicamente como a variável de maior importância. Para além desta, um conjunto de variáveis surge de forma recorrente entre as mais relevantes nos vários modelos, nomeadamente as Variáveis B, C e D e, em menor grau, a Variável F.

No que respeita às árvores de decisão ajustadas aos resultados dos modelos, esta abordagem não produziu resultados particularmente vantajosos. As árvores obtidas revelaram-se muito semelhantes entre si, apresentando sensivelmente as mesmas variáveis e pontos de corte, independentemente do modelo. Ao comparar as previsões destas árvores com as dos modelos originais e com os valores reais, concluiu-se que o ganho em interpretabilidade não compensava a perda significativa de capacidade preditiva, pelo que a análise não foi aprofundada.

Da análise dos gráficos de valores previstos *versus* observados (Figuras 7 a 11, em anexo), destaca-se o bom alinhamento global de todos os modelos com a reta de calibração, indicando que as previsões acompanham adequadamente a sinistralidade observada. O GLM e o LightGBM são os modelos que apresentam o ajustamento mais consistente, com os pontos a distribuírem-se de forma muito próxima da diagonal ao longo de toda a gama de risco.

Bin	Desvio Média 2 GLM	Desvio Média 2 Árvore Decisão	Desvio Média 2 Random Forest	Desvio Média 2 LightGBM	Desvio Média 2 XGBoost
15	-7,23%	2,92%	32,68%	-1,28%	21,57%
16	6,63%	19,03%	2,19%	45,77%	-5,85%
17	36,04%	11,52%	24,13%	30,13%	0,31%
18	43,99%	28,80%	30,48%	59,60%	27,56%
19	71,65%	53,15%	35,54%	42,93%	67,89%
20	56,19%	94,24%	125,17%	152,14%	139,68%

Tabela 2: Análise económica por bins nos modelos de regressão

Adicionalmente, analisou-se, para os *bins* mais elevados, o desvio face à taxa média de sinistralidade nos primeiros dias de vigência da apólice (Desvio Média 2), conforme apresentado na Tabela 2. Observa-se que todos os modelos, ao preverem valores mais elevados, tendem a acompanhar o aumento da sinistralidade, evidenciando boa capacidade de discriminação nos níveis de maior risco, com destaque para os modelos baseados em árvores, que apresentam os maiores desvios face à média global.

5.2 Resultados

5.2.1 Classificação

Modelo	β	t	Accuracy	Precision	Recall	Percentagem Analisada	Percentagem Correta	Desvio Média 1	Desvio Média 2
Sinistro			1	1	1	9,6%	9,6%	305,43%	1710,03%
Árvore Decisão	0,5	0,21	0,8458	0,1771	0,1660	9,00%	1,59%	60,39%	100,68%
Árvore Decisão	1	0,15	0,7529	0,1475	0,3290	21,43%	3,16%	29,90%	50,71%
GLM	0,5	0,16	0,8381	0,1963	0,2214	10,83%	2,13%	40,87%	55,40%
GLM	1	0,12	0,7228	0,1595	0,4420	26,61%	4,25%	25,79%	31,43%
LightGBM	0,3	0,28	0,8986	0,2780	0,0348	1,2%	0,33%	234,63%	292,84%
LightGBM	0,4	0,25	0,8937	0,2617	0,0586	2,15%	0,56%	164,79%	234,30%
LightGBM	0,5	0,23	0,8876	0,2476	0,0837	3,25%	0,8%	147,24%	197,24%
LightGBM	0,7	0,21	0,8784	0,2332	0,1161	4,78%	1,12%	121,93%	169,88%
LightGBM	1	0,18	0,8556	0,2116	0,1847	8,38%	1,77%	97,82%	137,85%
Random Forest	0,3	0,32	0,9004	0,2616	0,0201	0,74%	0,19%	154,93%	231,38%
Random Forest	0,5	0,30	0,8981	0,2368	0,0274	1,11%	0,26%	136,24%	209,03%
Random Forest	1	0,28	0,8948	0,2370	0,0431	1,75%	0,41%	105,61%	163,53%
XGBoost	0,3	0,22	0,8876	0,2517	0,0862	3,29%	0,83%	135,72%	211,30%
XGBoost	0,4	0,19	0,8695	0,2288	0,1514	6,35%	1,45%	123,07%	170,05%
XGBoost	0,5	0,18	0,8605	0,2209	0,1792	7,79%	1,72%	107,88%	156,20%
XGBoost	1	0,13	0,7725	0,1780	0,3784	20,42%	3,63%	52,66%	77,61%

Tabela 3: Métricas de desempenho e análise económica modelos de classificação binária

A Tabela 3 resume as métricas obtidas pelos diferentes modelos, seguidas da análise dos custos associados às decisões de cada um. A primeira linha (*sinistro*) representa o cenário ideal, em que todas as observações do conjunto de teste com *early claim* são corretamente sinalizadas. A coluna β corresponde ao valor de β utilizado no cálculo do *threshold* ótimo, apresentado na coluna t, que serviu para separar as observações entre as duas classes. Seguem-se as métricas usuais, a percentagem de apólices sinalizadas (Percentagem Analisada), a percentagem corretamente sinalizada (Percentagem Correta) e, por fim, a diferença face à sinistralidade média sobre todo o período observado (Desvio Média 1) e a diferença face à sinistralidade média sobre os primeiros dias de vigência da apólice (Desvio Média 2).

Da leitura dos resultados sobressai um compromisso claro entre a abrangência da sinalização e a sua precisão. Os modelos configurados com *thresholds* mais elevados, em particular o LightGBM e o Random Forest, ambos com $\beta = 0{,}3$, apresentam a maior *Precision* assim como a maior concentração de sinistralidade. Estes modelos sinalizam um número muito reduzido de apólices, mas com elevada fiabilidade, sendo por isso adequados a uma estratégia de investigação seletiva, em que se pretende minimizar os falsos positivos.

No extremo oposto, os modelos GLM e XGBoost, com $\beta = 1$, maximizam o *Recall*, captando uma fração muito superior dos sinistros, embora apresentem uma menor *Precision*. Estas configurações ajustam-se a uma estratégia de cobertura ampla, em que o objetivo é detetar o maior número possível de sinistros, aceitando um volume mais elevado de verificações.

O LightGBM evidencia-se como o modelo mais equilibrado: mantém valores de *Precision* e de concentração de sinistralidade elevados de forma consistente ao longo dos vários *thresholds*, o que lhe confere flexibilidade na adaptação ao nível de cobertura pretendido.

5.2.2 Regressão

A Tabela 4 resume as métricas obtidas pelos modelos de regressão assim como o valor absoluto da diferença entre a média observada e a média prevista. Os valores de RMSE e MAE são muito próximos entre os cinco modelos, o que indica capacidades preditivas semelhantes em termos de erro médio. A diferença mais relevante surge no desvio de Poisson, em que o Random Forest apresenta claramente o menor valor, bastante abaixo dos restantes modelos.

Modelo	Média Observada – Média Prevista	RMSE	MAE	Desvio de Poisson
GLM	0,0027	0,3891	0,2061	0,1295
Árvore Decisão	0,0028	0,3908	0,2058	0,1372
Random Forest	0,0096	0,3900	0,2146	0,0975
LightGBM	0,0060	0,3878	0,2014	0,1412
XGBoost	0,0097	0,3905	0,1985	0,1699

Tabela 4: Métricas de desempenho dos modelos de regressão

Na vertente económica, à semelhança dos resultados em classificação, observa-se na Tabela 5 que, à medida que se reduz a percentagem de apólices selecionadas, aumenta a sinistralidade no grupo sinalizado. Os modelos de *ensemble* destacam-se neste aspeto: o LightGBM apresenta o Desvio Média 2 mais elevado nos quantis 0,95 e 0,975, sendo ultrapassado pelo XGBoost no quantil 0,925; o XGBoost e o Random Forest disputam consistentemente a segunda posição. Estas apólices tendem, além disso, a manter uma sinistralidade elevada quando analisadas a longo prazo. O GLM, embora apresente grande proximidade à média em todas as análises, é o que revela a menor concentração de sinistralidade.

Modelo	quantil	Média Observada – Média Prevista	Percentagem Selecionada	Desvio Média 1	Desvio Média 2
nr_sin		0	9,83	294,19%	1641,25%
Árvore Decisão	0,925	0,0607	7,47	61,62%	80,22%
Árvore Decisão	0,95	0,1045	4,98	52,82%	94,20%
Árvore Decisão	0,975	0,1811	2,37	70,65%	105,12%
LightGBM	0,925	0,0159	7,50	88,34%	116,30%
LightGBM	0,95	0,0255	5,00	111,83%	152,09%
LightGBM	0,975	0,0567	2,50	141,06%	221,26%
GLM	0,925	0,0129	7,50	51,32%	65,36%
GLM	0,95	0,0096	5,00	42,73%	56,15%
GLM	0,975	0,0196	2,50	34,45%	66,29%
Random Forest	0,925	0,0590	7,50	83,55%	95,36%
Random Forest	0,95	0,0725	5,00	85,12%	125,10%
Random Forest	0,975	0,0836	2,50	109,65%	187,74%
XGBoost	0,925	0,0478	7,50	85,63%	133,00%
XGBoost	0,95	0,0647	5,00	110,60%	139,66%
XGBoost	0,975	0,1269	2,50	127,98%	157,63%

Tabela 5: Métricas de desempenho e análise económica modelos de regressão

6 Conclusões e Trabalho Futuro

6.1 Conclusão

Neste trabalho foram desenvolvidos e avaliados diferentes modelos de aprendizagem automática com o objetivo de identificar apólices com maior probabilidade de ocorrência de sinistro nos primeiros dias de vigência. Foram testados modelos lineares e não lineares, incluindo GLM, árvores de decisão, Random Forest, LightGBM e XGBoost.

Os resultados demonstraram que os modelos do tipo *black-box* apresentaram, em geral, melhor desempenho preditivo. Contudo, essa vantagem é contrabalançada por uma menor interpretabilidade, já que a sua complexidade dificulta a compreensão das relações entre as variáveis e as previsões produzidas.

No conjunto, os resultados evidenciam que não existe um modelo universalmente superior em todas as métricas. Na classificação, os modelos de *boosting* destacam-se pela precisão e pela concentração de risco. Na regressão, o GLM sobressai na qualidade do ajustamento e os modelos de *boosting* na eficácia económica. A escolha final depende, assim, do objetivo de negócio – maximizar a deteção de sinistros ou maximizar a eficiência da investigação.

6.2 Dificuldades e Limitações

Uma das principais dificuldades do trabalho esteve relacionada com os modelos de classificação, devido ao desequilíbrio entre as classes, uma vez que a ocorrência do evento de interesse é significativamente menos frequente do que os casos negativos. Esta característica condiciona o processo de aprendizagem dos modelos, favorecendo a classe maioritária e dificultando a correta identificação da classe minoritária, que é precisamente a mais relevante para o problema em estudo. Em consequência, foi necessário recorrer a métricas mais adequadas a cenários de desequilíbrio, bem como ao ajuste do *threshold* de decisão, de forma a controlar o compromisso entre *precision* e *recall*.

Outra dificuldade resultou da grande variabilidade das apólices com sinistro: não foi possível encontrar um padrão claro que as distinguisse, uma vez que a ocorrência de sinistros depende de múltiplos fatores diferenciadores.

Uma terceira dificuldade prendeu-se com o equilíbrio entre desempenho preditivo e interpretabilidade dos modelos. Embora os modelos mais complexos tenham apresentado melhores resultados, a sua natureza *black-box* dificultou a explicação direta das decisões.

A principal limitação do trabalho decorre da existência de regras e procedimentos já em vigor para a identificação e gestão das situações em análise. Essas regras refletem conhecimento operacional acumulado e políticas previamente definidas, o que condiciona a distribuição dos dados iniciais e, consequentemente os resultados dos modelos. Assim, mesmo quando os modelos apresentam capacidade preditiva relevante, a sua integração prática encontra limitações, dado ser necessário garantir consistência com os processos e as regras atualmente implementadas.

Por fim, a análise de custos foi realizada de forma simplificada, assumindo conversões proporcionais dos prémios e determinados pressupostos sobre o comportamento das apólices, o que pode não refletir totalmente a realidade operacional.

6.3 Sugestões de Melhoria e Trabalho Futuro

Como melhoria, seria relevante explorar métodos mais avançados de interpretabilidade, como SHAP values ou LIME, permitindo uma explicação mais detalhada das previsões dos modelos e ajudariam a mitigar a desvantagem associada à sua natureza *black-box*.

Outra possível extensão passaria pela otimização conjunta do *threshold* e da função de custo, integrando diretamente a componente económica na fase de modelação, em vez de a analisar apenas posteriormente.

Por fim, numa perspetiva futura, seria relevante avaliar o desempenho dos modelos sobre dados de 2026 ou de anos seguintes, de modo a confirmar a sua robustez e capacidade preditiva em novos períodos.

Bibliografia

ABI. (17 de Novembro de 2025). *Fraudulent insurance claims continue to top £1 billion*. Obtido em 20 de Junho de 2026, de Association of British Insurers: https://www.abi.org.uk/news/news-articles/2025/11/fraudulent-insurance-claims-continue-to-top-1-billion/

Castro, C., & Panda, A. (03 de Junho de 2026). *Seguradoras detetaram fraudes de 87 milhões de euros em falsos acidentes só num ano*. Obtido em 20 de Junho de 2026, de Jornal de Notícias: https://www.jn.pt/justica/artigo/seguradoras-detetaram-fraudes-de-87-milhoes-de-euros-em-falsos-acidentes-so-num-ano/18091378

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 785-794).

Denuit, M., Charpentier, A., & Trufin, J. (2021). Autocalibration and Tweedie-dominance for insurance pricing with machine learning. *Insurance: Mathematics and Economics*.

Dobson, A. J., & Barnett, A. G. (2018). *An Introduction to Generalized Linear Models* (4th ed.). Boca Raton: Chapman & Hall/CRC.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 1189-1232.

Gonçalves, F. P. (10 de Março de 2025). *Fraude nos seguros: Da simulação de acidentes à manipulação de dados*. Obtido em 20 de Junho de 2026, de ECO: https://eco.sapo.pt/2025/03/10/fraude-nos-seguros-da-simulacao-de-acidentes-a-manipulacao-de-dados/

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York: Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R* (2nd ed.). New York: Springer.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., . . . Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, (pp. 3146-3154).

R Core Team. (2024). R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. Obtido de https://www.r-project.org/

SAS Institute Inc. (2023). SAS Enterprise Guide. Cary: SAS Institute Inc.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 427-437.

Anexos

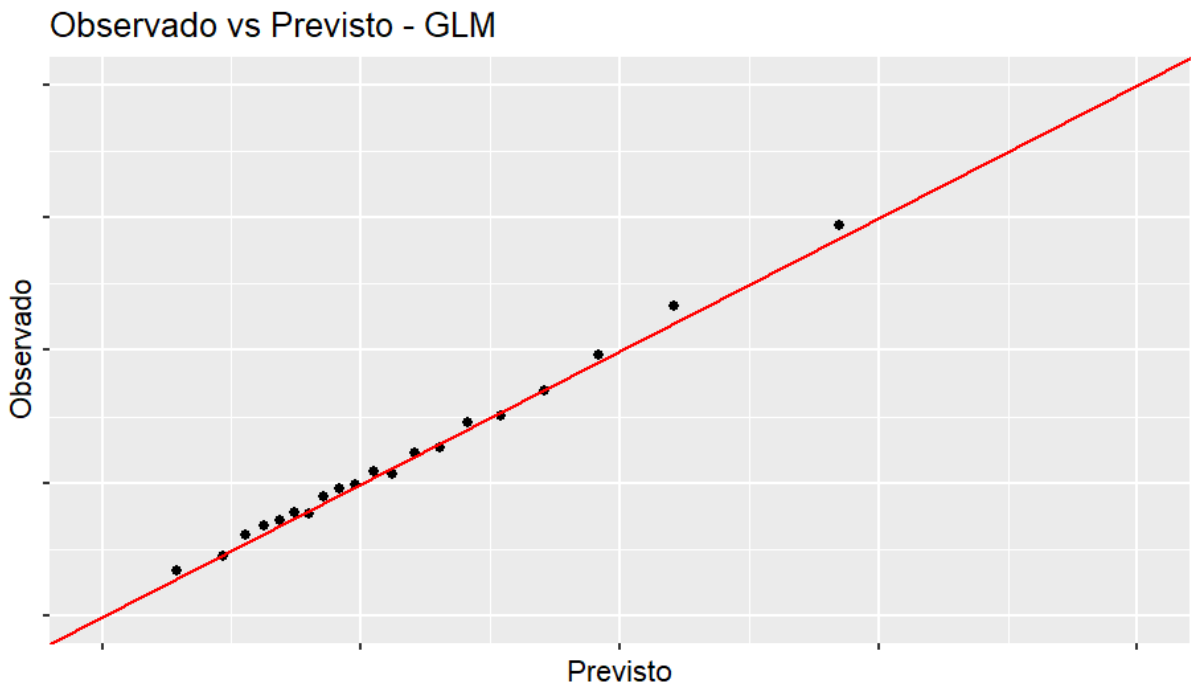

Figura 7: Gráfico médias Previstas vs Observadas - GLM

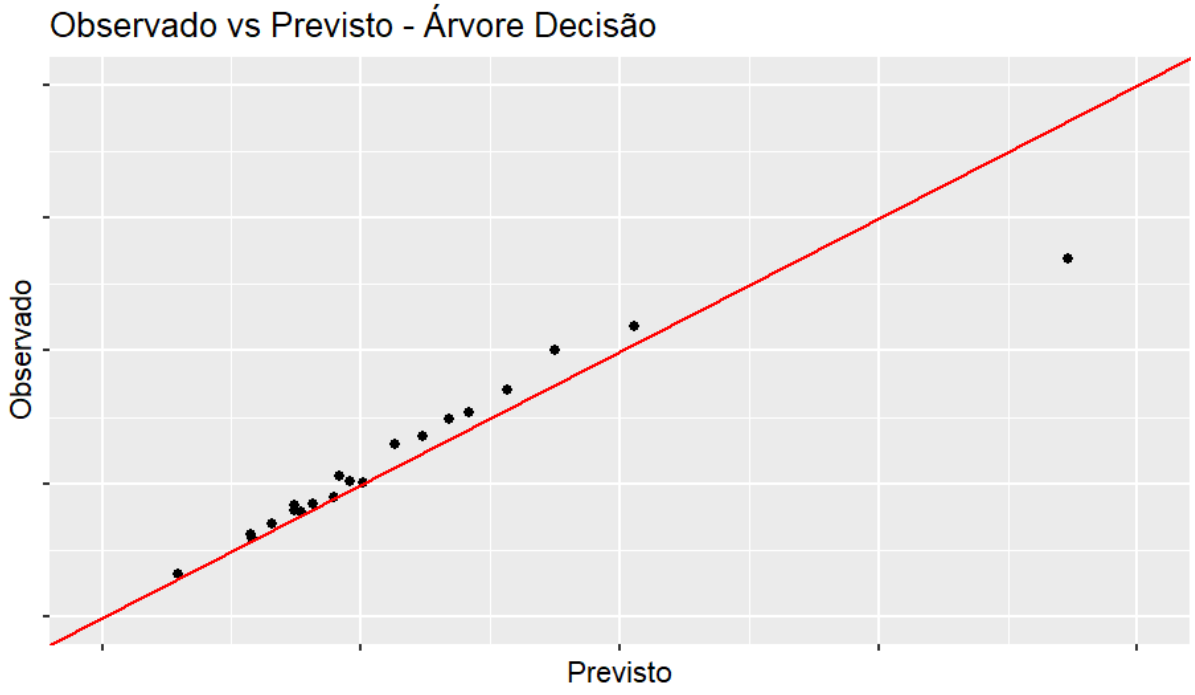

Figura 8: Gráfico médias Previstas vs Observadas - Árvore Decisão

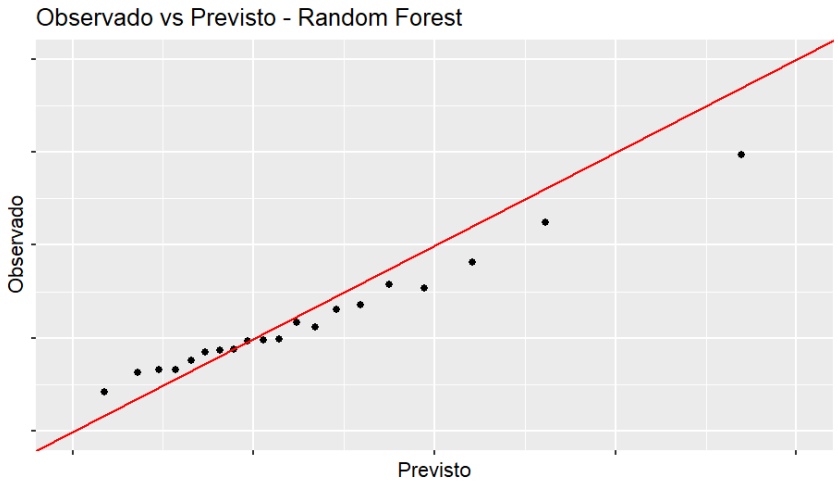

Figura 9: Gráfico médias Previstas vs Observadas - Random Forest

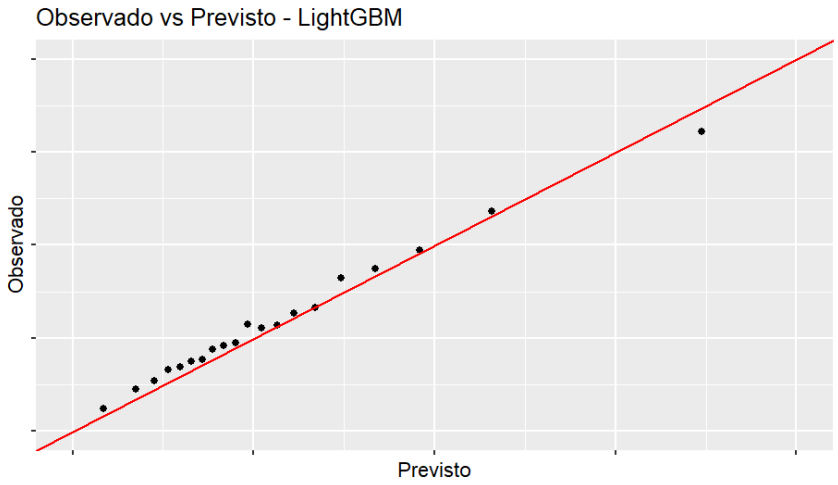

Figura 10: Gráfico médias Previstas vs Observadas - LightGBM

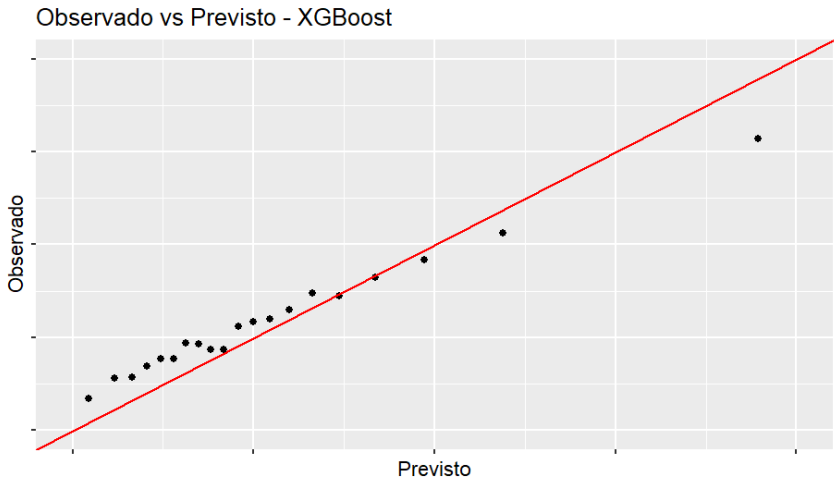

Figura 11: Gráfico médias Previstas vs Observadas – XGBoost