UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Criação de score de risco para negociação da precificação de seguro de entregadores

Diogo Silva Panham

Dissertação de Mestrado do Programa de Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria (MECAI)

ICMC USP SÃO CARLOS

Diogo Silva Panham

Criação de score de risco para negociação da precificação de seguro de entregadores

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria. *VERSÃO REVISADA*

Área de Concentração: Matemática, Estatística e Computação

Orientador: Prof. Dr. Pedro Luiz Ramos

USP – São Carlos
Novembro de 2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

S586c

Silva Panham, Diogo
Criação de score de risco para negociação da precificação de seguro de entregadores / Diogo Silva Panham; orientador Pedro Luiz Ramos. -- São Carlos, 2023.
73 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. . I. Luiz Ramos, Pedro, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Diogo Silva Panham

Creation of risk score for insurance pricing negotiation of delivery drivers

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master – Professional Masters in Mathematics, Statistics and Computing Applied to Industry. *FINAL VERSION*

Concentration Area: Mathematics, Statistics and Computing

Advisor: Prof. Dr. Pedro Luiz Ramos

USP – São Carlos
November 2023

AGRADECIMENTOS

Os agradecimentos principais são direcionados ao meu Orientador e à Instituição de ensino por proprocionar o desenvolvimento e aprofundamento da dissertacão,

RESUMO

PANHAM, D. S. **Criação de score de risco para negociação da precificação de seguro de entregadores**. 2023. 73 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

O objetivo deste estudo consistiu na proposição de uma nova precificação para o seguro de entregadores em uma empresa líder no mercado de *food delivery* na América Latina, com a finalidade de reduzir os custos elevados relativos ao prêmio pago à seguradora. O objetivo principal foi desenvolver uma metodologia para propor uma nova precificação de seguro, visando a redução dos custos elevados relacionados ao prêmio pago à seguradora. Este estudo realizou uma análise detalhada, estimando a probabilidade de sinistros com base em variáveis como rotas de entrega, modais de transporte, e perfis detalhados dos entregadores. A metodologia adotada envolveu a aplicação de técnicas avançadas de *machine* learning e análise estatística. Foram empregados modelos como regressão logística e árvores de decisão, os quais foram fundamentais para construir um perfil de risco robusto e realizar uma classificação eficiente de risco. A análise foi sustentada por um conjunto de dados abrangente, incluindo dimensões como frequência de sinistros, características das rotas, e dados demográficos dos entregadores, fornecendo uma base sólida para a modelagem. Os resultados obtidos com a nova metodologia de precificação demonstraram uma redução significativa nos custos de seguro. Isso foi possível pela incorporação de uma avaliação de risco mais precisa e personalizada, que levou em conta o padrão de ocorrência dos sinistros e os comportamentos de risco dos entregadores. Além disso, a precificação proposta ofereceu uma flexibilidade maior na contratação do seguro, tornando-o mais acessível para os entregadores e mais vantajoso para a empresa contratante e a seguradora. A inovação deste estudo reside na utilização de uma abordagem de score de risco baseada na probabilidade de sinistros para a precificação de seguros no segmento de *food delivery*, evidenciando ser uma alternativa eficiente e viável para a gestão de custos elevados associados ao seguro.

Palavras-chave: Seguro, *Machine Learning*, Precificação.

ABSTRACT

PANHAM, D. S. **Creation of risk score for insurance pricing negotiation of delivery drivers**. 2023. 73 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

The objective of this study was to propose a new pricing model for delivery drivers' insurance for a leading food delivery company in Latin America. The aim was to reduce the high costs associated with the premiums paid to the insurance company. To achieve this objective, an analysis was conducted to develop a methodology for proposing a new insurance pricing strategy, targeting the reduction of the high costs associated with the premiums paid to the insurer. This study carried out a detailed analysis, estimating the probability of claims based on variables such as delivery routes, transportation modes, and the detailed profiles of delivery drivers. The adopted methodology involved the application of advanced machine learning and statistical analysis techniques. Models such as logistic regression and decision trees were used, which were fundamental in building a robust risk profile and performing efficient risk classification. The analysis was supported by a comprehensive dataset, including dimensions such as claim frequency, route characteristics, and demographic data of the delivery drivers, providing a solid basis for modeling. The results obtained with the new pricing methodology demonstrated a significant reduction in insurance costs. This was possible due to the incorporation of a more accurate and personalized risk assessment, which took into account the pattern of claims occurrence and the risk behaviors of the delivery drivers. Furthermore, the proposed pricing offered greater flexibility in insurance contracting, making it more accessible for the delivery drivers and more advantageous for both the contracting company and the insurer. The innovation of this study lies in the use of a risk score approach based on the probability of claims for insurance pricing in the food delivery segment, proving to be an efficient and viable alternative for managing the high costs associated with insurance.

Keywords: Insurance, Machine Learning, Pricing.

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

O Brasil é um país de dimensões continentais e com uma grande população; portanto, é factível supor que tenha uma grande quantidade de veículos rodando no país. Também é possível supor que, assim como o surgimento da pandemia de COVID-19, alguns modais de transporte tomaram maior participação de uso, como é o caso de motocicletas e bicicletas, através do uso de aplicativos de entrega como instrumento de trabalho. Tais modais têm desempenhado um grande papel para a obtenção de trabalho formal e informal. Uma pesquisa realizada pelo IPEA (Instituto de Pesquisa Econômica Aplicada), utilizando dados do IBGE (Instituto Brasileiro de Geografia e Estatística), revela um aumento significativo na adesão de trabalhadores aos aplicativos de entrega: em 2016, havia cerca de 30 mil trabalhadores nesse segmento; número que ascendeu para 278 mil em 2021, especificamente em aplicativos de *food delivery*. De acordo com PCdoB65 (2021), nos últimos 5 anos (dados coletados a partir de 2015), houve um crescimento de 979,8% dos brasileiros trabalhando para aplicativos que fazem algum tipo de entrega.

Desta forma, tendo este cenário de crescimento mencionado, muitas empresas de *food delivery* viram sua demanda aumentar exponencialmente após o surto de COVID-19, com aumento de 24% comparado ao início da pandemia em 2020, conforme mencionado por Kercher (2022). Já que o contato físico foi restringido, assim o trabalho dos entregadores de *food delivery* se tornou essencial neste período. Com base em um quadro muito maior de prestadores de serviço em sua base, empresas neste segmento decidiram por proporcionar a seus prestadores uma maior cobertura (serviço prestado por seguradoras de veículo) em vista de acidentes que são inevitáveis no processo operacional. De acordo com a Lei 14.297, torna-se uma obrigação legal para as empresas que operam aplicativos de delivery a contratação de seguros de vida para os entregadores, abrangendo acidentes ocorridos durante o expediente, inclusive na ausência de vínculo empregatício (BRASIL, 2022). Esta exigência legal coloca as empresas de *food delivery* frente ao desafio de equilibrar a necessidade de oferecer a cobertura de seguro obrigatória com a gestão eficiente dos custos associados. Portanto, a determinação de um preço de seguro adequado não é apenas uma questão de conformidade legal, mas também um elemento crucial na manutenção da

sustentabilidade financeira das operações de delivery, assegurando simultaneamente a proteção adequada dos entregadores.

Os métodos de aprendizado de máquina desempenharam um papel crucial no avanço da precificação de seguros. Aproveitando a sua capacidade de processar grandes quantidades de dados e identificar padrões intricados, estas técnicas provocaram uma transformação revolucionária dentro da indústria, facilitando uma análise mais precisa dos fatores de risco, levando a estratégias de precificação mais justas. Além disso, eles contribuem para a detecção de fraudes, reforçando assim a segurança para seguradoras e segurados. Denuit, Charpentier e Trufin (2021) desenvolveu o procedimento de autocalibração para corrigir o viés nos modelos de precificação de seguros que são treinados minimizando a *deviance*. A abordagem é usada para prever reivindicações em seguros usando redes neurais e modelos lineares generalizados. Campo e Antonio (2023) desenvolveram um modelo de precificação de seguro orientado a dados para fatores de risco estruturados hierarquicamente. Eles comparam o desempenho preditivo de três modelos e descobrem que a distribuição Tweedie é adequada para modelar e prever o custo de perda em um contrato. O impacto profundo desses métodos é inegável, pois eles melhoraram significativamente a precisão dos cálculos e o gerenciamento geral de riscos.

O seguro de automóveis para motocicletas é considerado, para as seguradoras, como alto risco, sendo visto por muitos como um risco excluído. Segundo Filho (2010), o risco excluído é: "diz-se risco excluído da cobertura e por ele o segurado não percebe qualquer contraprestação, nem o segurado tem, em caso de prejuízo resultante dele, qualquer expectativa de indenização secundária". Assim, portanto, devido ao risco alto, o preço praticado e a falta de cobertura tornam o custo desta operação cara, tanto para seguradoras como para empresas que contratam este serviço.

Com base nesta suposição, o estudo é referente a um case de uma empresa de *food delivery* que teve um crescimento considerável nos últimos anos e viu um aumento no número de entregadores na sua base de prestadores de serviço. Para tanto, a empresa fez a contratação de uma seguradora para realizar a cobertura de seus entregadores em todas as rotas em exercício da função. Devido à grande demanda, os custos com o seguro começaram a tornar-se grandes para a empresa; assim, fez-se necessário um estudo que possibilitasse a redução do custo com relação ao prêmio pago à seguradora. Como prêmio, referimos ao valor cobrado pela seguradora com a finalidade de cobrir o risco por ela assumido, assim como outras despesas, como afirmado por CANÔAS (2007). Para enfrentar este desafio, uma combinação de técnicas de aprendizado de máquina e modelagem estatística foi empregada para construir um escore de risco. As técnicas selecionadas foram escolhidas com base em sua robustez técnica e interpretabilidade, aspectos cruciais dada a necessidade de os resultados obtidos serem facilmente compreensíveis para diversos públicos.

O principal objetivo deste estudo é explorar estratégias para reduzir os custos do seguro, mantendo uma cobertura adequada para os motoristas de entrega de alimentos. Para atingir esse

objetivo, uma série de abordagens metodológicas foram empregadas. Inicialmente, a técnica de *Undersampling - Nearest neighbor rule undersampling* - foi aplicada para equilibrar o banco de dados de rotas, garantindo uma representação mais equitativa dos diferentes tipos de rotas na análise, uma metodologia corroborada por Smith *et al.* (2022). Posteriormente, uma árvore de decisão foi utilizada para criar o perfil de sinistros. A decisão de usar essa técnica foi motivada por sua facilidade de interpretação, tornando o modelo gerado mais acessível à área de negócios. A árvore de decisão facilitou a identificação e a visualização clara dos principais determinantes dos sinistros. Finalmente, a regressão logística foi empregada para calcular as probabilidades de ocorrência de sinistros com base nos perfis identificados. A seleção da regressão logística foi influenciada por sua facilidade de implementação e alta interpretabilidade. A combinação dessas técnicas resultou em um modelo de escore de risco transparente e compreensível, com suas entradas efetivamente utilizadas nas negociações. Este estudo fornece uma importante contribuição para o campo ao apresentar *insights* sobre o desenvolvimento de um modelo de escore de risco que auxilia na otimização dos custos do seguro para empresas de entrega de alimentos. Ao empregar técnicas analíticas avançadas, a pesquisa oferece recomendações valiosas e descobertas para a indústria, melhorando ainda mais a compreensão e o gerenciamento de prêmios de seguro no contexto de serviços de entrega de alimentos.

O restante desta dissertação é apresentado da seguinte forma: O capítulo 2 apresenta uma revisão da literatura, fornecendo alguns conceitos básicos de composição de prêmio de seguro e técnicas de *Machine Learning* e amostragem. Nos capítulos 3 e 4, será apresentado o desenvolvimento do estudo, com a aplicação das técnicas mencionadas anteriormente. No capítulo 5, o cálculo da precificação é realizado com base nos resultados das técnicas aplicadas no capítulo 4 e, finalmente, temos as conclusões

CAPÍTULO

2

REVISÃO BIBLIOGRÁFICA

Este capítulo de revisão bibliográfica visa fornecer um panorama abrangente dos estudos e teorias relacionados ao tema central da dissertação: a precificação de seguros no contexto de operações de entrega e a aplicação de métodos de aprendizado de máquina e modelagem estatística nesse domínio. A revisão explorará literatura pertinente sobre as práticas atuais de precificação de seguros, enfatizando como a inovação tecnológica, especialmente o aprendizado de máquina, vem transformando esta área. Além disso, serão discutidos modelos estatísticos e suas aplicações em cenários de risco e seguros, estabelecendo um alicerce teórico para os métodos empregados e os resultados alcançados neste estudo. Esta revisão não apenas contextualiza o trabalho realizado, mas também ilumina as lacunas existentes no conhecimento atual, justificando, assim, a necessidade e relevância da pesquisa conduzida.

2.1 Estimativa prêmio seguradora

As seguradoras entendem a operação de *delivery* feita por motos e motocicletas como de alto risco, considerando que boa parte dos acidentes que ocorrem no país são acidentes de motocicletas, assim como evidencia o artigo sobre indenização paga pelo seguro DPVAT (SCARAMUSSA; SÁ, 2020). Para o modelo da empresa em análise, a composição do prêmio levou em consideração os sinistros ocorridos de acordo com o relatório da DPVAT (Danos Pessoais Causados por Veículos Automotores de Via Terrestre), que trata-se de um seguro obrigatório no Brasil que tem o objetivo de indenizar vítimas de acidentes de trânsito, seja por morte, invalidez permanente ou despesas médicas. É válido para qualquer veículo automotor de via terrestre. Em posse destes dados públicos, a seguradora cruza as informações com os dados e relatórios enviados pela empresa de *food delivery*, com a especificidade de região, cidade, estado, etc. De maneira geral, a construção do prêmio cobrado pela seguradora pode ser feita de várias formas e conforme a necessidade de cada seguradora em questão; em muitas vezes, a informação sobre o cálculo não é divulgada pela seguradora. Contudo, para fins de exemplificação, usaremos

um modelo generalista. Podemos citar uma forma de composição, assim referenciada de acordo com Müller (2022), em que os mesmos citam uma abordagem atuarial para o cálculo do prêmio, a fim de cobrir os riscos suficientes para pagamento das indenizações e cobrir outros custos de operação; para tanto, será utilizado como referencial o Modelo de Risco anual, que consiste na precificação do seguro com base na soma dos sinistros em determinado ano (DPVAT), assim como uma margem de segurança para eventuais casos de sinistro que possam extrapolar uma média. Como não tem por finalidade do estudo a precificação do prêmio, será apresentado de maneira geral uma forma de cálculo do Prêmio puro:

$$P = \lambda E[X] + Z_{1-\alpha}\sqrt{\lambda E[X^2]} \tag{2.1}$$

Em que:

- P é o prêmio puro total,
- λ é o número de sinistros,
- $E[X]$ é a esperança de X,
- $Z_{1-\alpha}$ é o valor crítico (teste unicaudal), sendo α o nível de significância ,
- $E[x^2] = \frac{\sum_{i=1}^{N}(X_i * Xi)}{N}$.

Como X entendemos o valor da indenização paga para cada sinistro.Para o cálculo da esperança de X foi utilizado a fórmula:

$$\mathbf{E}[X] = \frac{\sum_{i=1}^{N} Xi}{N}$$

em que:

- Xi é o valor do i-énesimo sinistro,
- N é o número de sinistros,

No final para o valor individual do prêmio foi dividido pelo número de bilhetes emitidos do dpvat, que não equivale necessariamente ao valor arrecadado.

Na avaliação do prêmio de seguro, um aspecto crucial é a ligação entre o risco e o prêmio. A precificação deve refletir não apenas a probabilidade de ocorrência de sinistros, mas também o valor médio justo associado a cada rota. Isso implica na necessidade de distinguir entre rotas de baixo, médio e alto risco, ao invés de aplicar um valor uniforme de prêmio. A determinação de um prêmio adequado envolve a avaliação da probabilidade de sinistros em diferentes rotas e a estimativa do custo médio desses sinistros, permitindo que as seguradoras estabeleçam prêmios mais justos e personalizados, refletindo melhor o risco real associado a cada entrega.

Enquanto a avaliação precisa do prêmio de seguro requer uma análise detalhada do risco associado a cada rota, um aspecto igualmente crítico na gestão de uma seguradora é entender a dinâmica do capital ao longo do tempo. Esta compreensão é fundamentalmente enraizada na teoria da ruína, que fornece uma perspectiva mais ampla sobre a sustentabilidade financeira das operações da seguradora frente aos riscos assumidos.

A teoria da ruína é um conceito fundamental no estudo da sustentabilidade financeira de uma seguradora. Ela examina a trajetória do capital inicial, $U(t)$, de uma seguradora no ano t como um processo estocástico, que é impactado positivamente pelo acúmulo de prêmios e negativamente pelos pagamentos de sinistros. O foco central dessa teoria é a probabilidade de ruína, ou seja, a probabilidade de que o capital da seguradora em algum momento caia para zero ou abaixo, indicando insolvência. Este conceito é crucial para a gestão de riscos na indústria de seguros, uma vez que uma compreensão precisa da probabilidade de ruína ajuda as seguradoras a balancear adequadamente a relação entre a cobrança de prêmios e a cobertura de sinistros. Para uma análise aprofundada da teoria da ruína e sua aplicação em seguros, o trabalho de Embrechts et al. (EMBRECHTS; KLÜPPELBERG; MIKOSCH, 2013) oferece insights valiosos e uma exploração detalhada dos modelos matemáticos envolvidos.

2.2 Undersampling e oversampling

Gerenciar um grande conjunto de dados e empregá-los em modelos de classificação frequentemente envolve o desafio de lidar com dados desbalanceados. Esse é um problema comum em modelos de classificação, onde a distribuição desigual dos dados pode afetar significativamente a eficácia do modelo.Segundo He e Garcia (2009), o desbalanceamento de classes pode causar um viés nos modelos de classificação, com as técnicas de *machine learning* padrão tendendo sempre para a classe majoritária.

Para resolução deste problema, temos duas soluções: utilização de técnicas em nível de algoritmo e técnicas de nível de dados. De acordo com Chawla *et al.* (2002) as técnicas de nível de algoritmo têm por objetivo modificar o algoritmo para levar em consideraçào a distribuição dos dados ou impulsionar a identificação da classe minoritária. Contudo, as técnicas de nível de dados, como mencionado por He e Garcia (2009) com relação às técnicas de nível de dados, têm por objetivo modificar a base de dados antes do aprendizado realizado pelo algortimo; com base neste abordagem há duas categórias de técnica a de *Oversampling* e *Undersampling*.

A técnica de *Oversampling* tem por finalidade replicar ou gerar dados da classe minoritária e a de *Undersampling* tem como finalidade remover manualmente dados da classe majoritária. Neste estudo foi utilizada a técnica de Undersampling para poder também minimizar ruídos que possam haver da classe majoritária. Conforme Seiffert *et al.* (2010), o Undersampling é uma técnica que tem como finalidade remover manualmente dados da classe majoritária, sendo um exemplo simples o *random undersampling*.É o algoritmo não heurístico que tem por finalidade

balancear a base de dados eliminando randomicamente dados da classe majoritária. Esta operação pode ocasionar riscos, pois no momento da eliminação dos dados podem estar sendo descartada informações importantes para classificação do modelo. Esta técnica é ideal para situações com um grande volume de dados, pois, nesses casos, ainda se mantêm informações relevantes.

2.3 Decision tree

Existem vários modelos de classificação, uma deles é a árvore de decisão, ela é uma metodologia não paramétrica, de acordo com Izbicki e Santos (2020) : "Uma árvore é construída por particionamentos recursivos no espaço das covariáveis. Cada particionamento recebe o nome de nó e cada resultado final recebe o nome de folha;

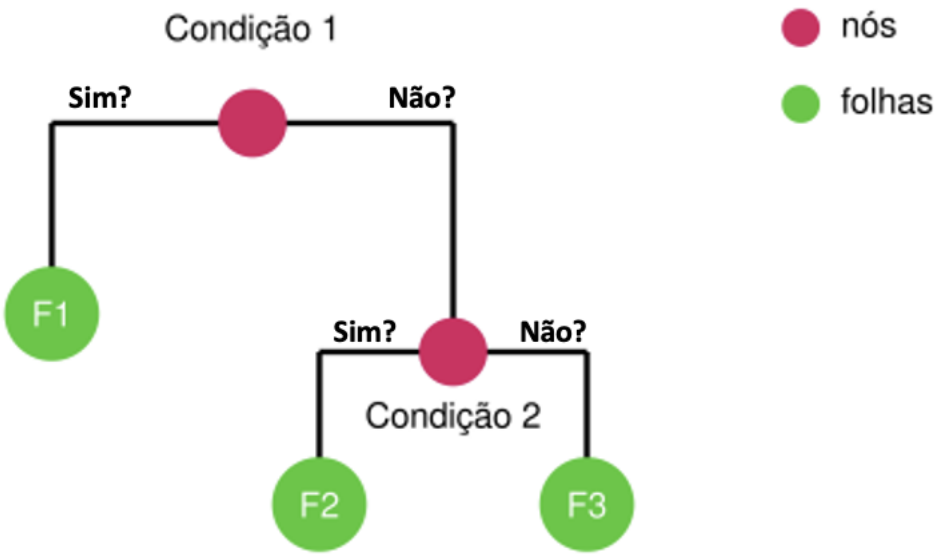

Figura 1 – Estrutura de uma árvore de decisão (Fonte: (IZBICKI; SANTOS, 2020))

De acordo com a figura 1 a estrutura da árvore ocorre da seguinte maneira: a primeira condição, se válida, vai para próxima ramificação à esquerda, caso não seja satisfeita esta condição dá-se prosseguimento pela direita, até ser atingida a folha. Esta estrutura é seguida nas árvores de decisão, sejam elas de regressão ou de classificação.

2.3.1 Árvore de classificação

De modo geral,a árvore de classificação tem construção análoga à árvore de regressão, com algumas ressalvas. Quando prevemos Y para uma observação com x variáveis, e estas se encontram na zona Rk, a previsão não se baseia mais na média amostral, mas pela moda deste

conjunto de treinamento:

$$g(x) = moda\{y_i : x_i \in R_k\}. \tag{2.2}$$

Na árvore de regressão a avaliação para encontrar a melhor partição em cada etapa no processo de distribuição dos dados nos nós da árvore é utilizado o erro quadrático, mas na árvore de classificação é utilizado o índice de Gini, conforme fórmula a seguir:

$$\sum_{R} \sum_{c \in C} \hat{P}_{R,c}(1 - \hat{P}_{R,c}), \tag{2.3}$$

Na fórmula o R representa umas das particões induzidas pela árvore e $\hat{P}_{R,c}$ é a proporção de observações da categoria c que pertencem a região ou partição dos dados R (BROWNLEE, 2018). O índice de Gini é mínimo quando as proporções $(\hat{P}_{R,c})$ são zero ou um; neste caso pode-se dizer que é uma árvore "pura", que significa quando uma folha tem somente observações de uma única classe (JAMES *et al.*, 2013). A árvore por assim dizer vai fazer divisões dos dados, no processo de montar sua estrutura, formando de uma certa forma padrões ou *cluster* destes dados (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

2.4 Regressão logística binária

O modelo de regressão logística binária tem por finalidade inferir a probabilidade de ocorrência de um evento definido por Y, sendo apresentado como uma variável dicotômica, assumindo portanto valor 1 quando ocorrência do evento e valor 0 quando não ocorrência do evento, isto com base no compartamento das variáveis explicativas. Este vetor com as váriveis explicativas e seus parâmteros é definido pela seguinte formula:

$$Z_i = \alpha + \beta_1.X_{1i} + \beta_2.X_{2i} + ... + \beta_k.X_{ki} \tag{2.4}$$

Em que:

- Z é o logito:,
- β é os parâmetros das variáveis explicativas,
- X_j são as variáveis explicativas ,
- i representa cada observação ,
- α é o intercepto.

Definir a probabilidade p_i de ocorrência de um evento para cada umas das observações com base na função do logito Z_i. Detalhes sobre a construção da fórmula de probabilidade não serão explorados; a discussão se concentrará na apresentação de sua forma final a seguir:

$$p_i = \frac{1}{1 + e^{-(\alpha+\beta_1.X_{1i}+\beta_2.X_{2i}+...+\beta_k.X_{ki})}} \tag{2.5}$$

De maneira geral, a regressão logística não estima os valores da variável dependente, mas sim a probabilidade de ocorrência do evento para cada uma das observações. Cabe ainda uma observação; para a estimação dos parâmetros no logito é utilizada a **estimação por máxima verossimilhança**. Sobre a regressão linear é utilizado o método de minimização da somatória dos quadrados dos resíduos. Conforme Czepiel (2002),"para variáveis categóricas, é inadequado usar a regressão linear porque os valores de resposta não são medidos em uma escala de razão e os termos de erro não são normalmente distribuídos". Para o método classificatório será usada a função de log-verossimilhança a partir da qual é elaborada a estimação de máxima versossimilhança conforme fórmula a seguir:

$$LL = \sum_{i=1}^{n} \left\{ \left[(Y_i) \cdot \ln \left(\frac{e^{z_i}}{1 + e^{z_i}} \right) \right] + \left[(1 - Y_i) \cdot \ln \left(\frac{1}{1 + e^{z_i}} \right) \right] \right\} = \max \tag{2.6}$$

É crucial notar que, apesar da regressão logística ser frequentemente usada em situações onde a variável resposta é binária, ela também pode ser aplicada em cenários onde a variável resposta é ordinal ou nominal com mais de duas categorias. Este método é uma alternativa robusta à regressão linear em cenários onde a relação entre as variáveis independentes e a variável dependente não é necessariamente linear e onde os erros não precisam ter uma distribuição normal. A estimação por máxima verossimilhança empregada na regressão logística pode ser computacionalmente intensiva (dependendo de fatores como tamanho do conjunto de dados, complexidade do modelo, etc.), mas oferece resultados confiáveis e interpretações intuitivas, tornando-a uma ferramenta poderosa na análise de dados categóricos (JR; LEMESHOW; STURDIVANT, 2013).

Concluindo a exposição do método teórico abordado nesta seção, é importante destacar a aplicação prática destes conceitos no contexto desta dissertação. Neste trabalho, a variável aleatória adotada será Y, que pode assumir os valores 0, 'Não sinistro' e 1, 'Sinistro'. Esta variável será utilizada para calcular a probabilidade de ocorrência de sinistros, relacionando-a à proporção do risco por tipo de rota. A adoção dessa abordagem é fundamental, pois evidencia que não é adequado aplicar uma taxa uniforme de seguro para todas as rotas. Ao contrário, a taxa deve refletir o risco específico associado a cada tipo de rota, assegurando uma precificação mais justa e alinhada com os princípios de equidade e eficiência na gestão de riscos. Dessa forma, a dissertação busca aliar teoria e prática para oferecer uma contribuição significativa à precificação de seguros no segmento de entregas, um mercado em constante crescimento e de grande relevância econômica e social.

2.5 Nearest neighbor rule undersampling

A técnica de subamostragem *Nearest neighbor rule undersampling* (NNRU) é uma abordagem amplamente utilizada para lidar com o desbalanceamento de dados em problemas de classificação. O desbalanceamento de classes ocorre quando a distribuição das classes é significativamente desigual, o que pode levar a um viés no desempenho do modelo de aprendizado de máquina, favorecendo a classe majoritária em detrimento da classe minoritária. Nesse contexto, a NNRU visa equilibrar as classes reduzindo a quantidade de amostras da classe majoritária sem descartar informações relevantes.

A NNRU é fundamentada na ideia de que as amostras da classe majoritária mais próximas à fronteira de decisão entre as classes são as mais críticas para a tarefa de classificação. Essas amostras são selecionadas com base nos vizinhos mais próximos pertencentes à classe minoritária. Essa abordagem permite que o modelo aprenda a distinguir melhor as características das classes, melhorando sua capacidade de generalização para a classe minoritária.

Vários estudos têm demonstrado a eficácia da NNRU em diversos domínios de aplicação. Em um estudo realizado por Zhang *et al.* (2018) sobre a classificação de estágios de sono em dados desbalanceados, a NNRU foi aplicada para melhorar o desempenho do modelo. Os resultados mostraram que a técnica de NNRU aumentou significativamente a acurácia e a sensibilidade para a classe minoritária, em comparação com outras técnicas de subamostragem.

Outro estudo realizado por Li *et al.* (2020) investigou a aplicação da NNRU na detecção de fraudes financeiras. Os autores utilizaram a técnica para equilibrar os dados desbalanceados e treinar um modelo de detecção de fraudes. Os resultados mostraram que a NNRU melhorou significativamente a precisão e a taxa de detecção de fraudes em comparação com outros métodos de subamostragem.

É importante destacar que a NNRU possui algumas limitações. Em conjuntos de dados com sobreposição significativa entre as classes ou com regiões de fronteira de decisão complexas, a NNRU pode não ser tão eficaz. Nesses casos, outras técnicas de subamostragem, como SMOTE (*Synthetic Minority Over-sampling Technique*), podem ser mais adequadas.

Em suma, a técnica de subamostragem NNRU é uma abordagem promissora para lidar com o desbalanceamento de dados em problemas de classificação. Sua fundamentação teórica na seleção de amostras próximas à fronteira de decisão e os resultados positivos encontrados em estudos empíricos reforçam sua eficácia. No entanto, é importante considerar as características específicas do conjunto de dados e avaliar outras técnicas de subamostragem, conforme necessário, para obter os melhores resultados em cada cenário.

Em relação a um exemplo prático em Python da biblioteca que implementa a técnica NNRU, uma opção seria utilizar o pacote imbalanced-learn conforme Figura 2, que fornece implementações de várias técnicas de balanceamento de classes:

```python
from imblearn.under_sampling import NeighbourhoodCleaningRule
from sklearn.datasets import make_classification

# Criação de um conjunto de dados de exemplo
X, y = make_classification(n_samples=1000, weights=[0.95, 0.05])

# Aplicação da NNRU
nnru = NeighbourhoodCleaningRule()
X_res, y_res = nnru.fit_resample(X, y)
```

Figura 2 – Uso da biblioteca no python da técnica NNRU

Nesse exemplo, utilizamos a função make_classification para gerar um conjunto de dados de exemplo com uma classe majoritária (95%) e uma classe minoritária (5%). Em seguida, aplicamos a NNRU utilizando a classe NeighbourhoodCleaningRule do pacote imbalanced-learn, que realiza a subamostragem com base na regra dos vizinhos mais próximos.

CAPÍTULO

3

METODOLOGIA

Este capítulo apresenta a metodologia adotada nesta pesquisa, delineando uma abordagem sistemática para a análise e precificação de seguros no contexto de entregas de *food delivery*. A metodologia empregada nesta pesquisa inicia com uma análise exploratória detalhada das rotas e informações dos entregadores, fornecendo uma compreensão aprofundada do conjunto de dados. Esta fase é seguida por testes de hipótese para as variáveis identificadas, o que ajuda a estabelecer a relevância e significância estatística de cada uma. Após esta etapa inicial, procedemos com o balanceamento dos dados, utilizando técnicas de amostragem e clusterização, para assegurar uma representação equitativa e melhorar a confiabilidade dos modelos preditivos. Com os dados balanceados, desenvolvemos modelos de *machine learning*, começando com a utilização da árvore de decisão para a geração de *clusters* e, em seguida, empregando a regressão logística para calcular a probabilidade de ocorrência de sinistros. Este processo resulta na definição de um escore de risco, que é um elemento chave na precificação de seguros para entregadores. Cada etapa da metodologia é interdependente e contribui de maneira significativa para a construção de um modelo robusto e confiável, capaz de alinhar as necessidades práticas do mercado de seguros com uma abordagem analítica rigorosa

Figura 3 – Resumo da metologia adotada

3.1 Análise exploratória dos dados

A análise exploratória dos dados é um momento crucial para o entendimento do problema do negócio. Nesta etapa é realizado o entendimento, a visualização, organização e resumo dos

dados.

3.1.1 Descrição das tabelas

Antes do início das análises foi realizado o levantamento dos *datasets*, relacionados a rotas e ao sinistros. Considera-se como rota o percurso que o entregador realiza desde o momento em que sai do restaurante para efetuar entregas e retorna ao estabelecimento. Nesse contexto, o trajeto de ida e volta ao restaurante é considerado uma rota.

Existem, por definição, diversos tipos de rotas que podem ser descritos da seguinte maneira:

- SPMD: Uma origem - Múltiplo Destinatários;
- MPMD: Múltiplas origens - Múltiplos destinatários;
- SPSD: Uma origem - Um destinátario;
- MPSD: Múltiplas origens - Um Destinatário

Para construção da base de dados utilizada foram utilizadas as seguintes tabelas, constituindo no total 51 variavéis que foram usadas na análise diretamente ou auxiliaram na construção de váriaveis derivativas:

tabela_logistica_curada.rotas

Essa base foi relacionada com a base de dados da seguradora e dos entregadores. Optou–se por utilizar das seguintes informações da tabela:

- id_entregador: id do entregador (Chave de relacionamento de tabelas);
- data_particao: Data de criação da rota (A seguradora usa como referência a data que a rota foi criada);
- created_timestamp: Data e horário da rota;
- total_distância_origem_ate_destino: soma das distâncias percorridas no dia de ida até destino da entrega;
- total_distancia_ate_origem: soma das distâncias percorridas no dia de volta até o restaurante;
- total_rota_duracao: soma do total de tempo das rotas no dia;
- id_rota: id da rota;
- entregador_lo: Contrato do entregador na Empresa (Ex: Autônom, Operador Nuvem);

- entregador_modal: Modal do entregador (tipo Veículo usado);
- regiao_logística: Região da rota;
- distancia_origem_ate_destino: distância percorrida da última rota do entregador no dia na ida para entrega do pedido;
- distancia_ate_origem: distância percorrida da última rota do entregador no dia na volta para restaurante;
- tipo_rota: Tipo de rota de entrega;
- rota_duracao: Tempo da última rota do dia;
- rank: Quantidade de entregas feito na empresa desde o início;
- rota_avanco: Quantos dias entre a rota atual e a próxima rota do entregador (Campo passível a atualização);
- rota_atraso: Quantos dias entre a rota anterior e a rota atual (Campo passível a atualização);
- rota_estado: Estado da rota (Se concluída ou não);
- cancelamento_rejeicao_tipo: Motivo de cancelamento da rota;
- rota_modelo: modelo de negócio (Logística ou Market Place);
- otimizador_busca: Definição da rota (Manual, ou pelo software Version1/Version2);
- multiplicador_frete: multiplicador dinâmico de preço aplicado a rota;
- taxa: É a taxa que o motorista recebe dependendo da rota e se houve uma promoção na data;
- max_rank: Campo calculado para o valor máximo com base na coluna de classificação;
- max_rota_atraso: Campo calculado para o valor máximo com base na coluna de rota_atraso

tabela_logistica_curada.monitoramento_entregador

Esta tabela contém dados sobre a posição dos motoristas que estão *online* (ou seja, 'disponíveis para receber pedidos') usando o Aplicativo de Motoristas. Foi utilizadas as seguintes informações:

- id_entregador: id do Entregador;
- evento_dia: dia em que a localização do motorista foi registrada;
- id_rota: id da rota;

- segmentacao: segmentação do driver no momento(profissional, casual, inativo..);
- tempo_minutos: tempo online do entregador no dia (Campo calculado). Subtraindo o tempo máximo e mínimo do entregador em um determinado dia para encontrar o tempo (em minutos);
- media: tempo médio do entregador online. Somando tempo_minutos e dividindo pela quantidade de dias online (campo calculado);
- cpf: CPF do entregador

sinistro

São informações dos sinistros que ocorreram e foram fornecidas pela seguradora. Essa tabela é importante, pois consta a informação *target* da análise (Se é ou não sinistro).

- Número do Sinistro:número de controle da seguradora que identifica o sinistro conforme ocorre e é informado pelo entregador ou família do entregador.
- Data Atualização: Data de atualização do processo, qualquer alteração no processo de análise essa informação é alterada;
- Data Ocorrência: Data em que ocorreu o sinistro, informada pelo entregador ou familiar;
- Data de Aviso: Data em que a segurador foi avisada do sinistro;
- Nome Sinistrado: Nome do Entregador que sofreu o acidente;
- CPF Sinistrado: CPF do Entregador que sofreu o acidente;
- Cobertura: Informações sobre a cobertura conforme o tipo do Sinistro;
- Valor Avisado: Valor do prêmio a ser pago ao Beneficiário;
- Motivo da Recusa: Motivos da recusa quando o sinistro foi indeferido;
- Data do Pagamento: Data de Pagamento do prêmio ao Beneficiário;
- Modal: Veículo usado pelo entregador nas entregas;
- Data de Nascimento: Data de Nascimento do Entregador;
- Cidade: Cidade em que houve o acidente;
- Tipos de Corrida: Se a corrida ocorreu no trajeto de entregas da Empresa, ou se foi no retorno para casa;
- Encerrado: Status da Análise do Sinistro

tabela_logistica_curada.promocao_rotas

Esta tabela contém informações sobre as promoções logísticas oferecidas aos motoristas para aumentar / manter o abastecimento.

- nome_promocao: nome da promoção (registrado na Frota);
- regiao_logistica: nome da região logística onde a promoção é aplicada;
- data: data da promoção

logistics_curated.driver_tracking

Tabela com informações de rastreamento da rota, via aplicativo do entregador.

- segmentacao: classificação interna do entregador
- evento_timestamp: tempo de ocorrência do evento no banco de dados
- data: data da promoção

3.1.1.1 Premissas da seleção da base de dados

Foram removidos da base de sinistros os registros de sinistros indeferidos e de "Rotas para Casa". Essas informações foram excluídas porque inicialmente foi planejado criar um score com base nos sinistros e suas informações correlacionadas. A análise da seguradora identificou que os sinistros indeferidos não atendiam aos requisitos que a empresa de *delivery* e a seguradora caracterizavam como uma rota coberta. Com relação às "Rotas para Casa", elas ocorrem após o entregador encerrar o aplicativo e, portanto, não é possível rastrear informações importantes sobre a rota. Além disso, o número de rotas para casa na base de sinistros é pequeno em comparação com o total de rotas.

A partir de 28 de outubro de 2019, foram filtradas as tabelas utilizadas na base de sinistros. Esse filtro foi estabelecido por dois motivos: o primeiro é que essa foi a data em que as rotas começaram a ser cobertas; o segundo é que o filtro foi utilizado para reduzir o tamanho da base de rotas, que possuía milhões de registros, evitando problemas de desempenho.

3.1.2 Análise das Rotas

Entende-se como rota o caminho percorrido pelo entregador. Realizando as primeiras análises é possível perceber que o uso da moto corresponde aproximadamente 83% das rotas, seguido por entregas feitas por bicicletas, com 13% (Figura 4).

A análise dos dados revelou que a base está desbalanceada quando dividida entre sinistros e não sinistros, como mostrado na (Figura 5), a porcentagem de rotas que ocorreram sinistro

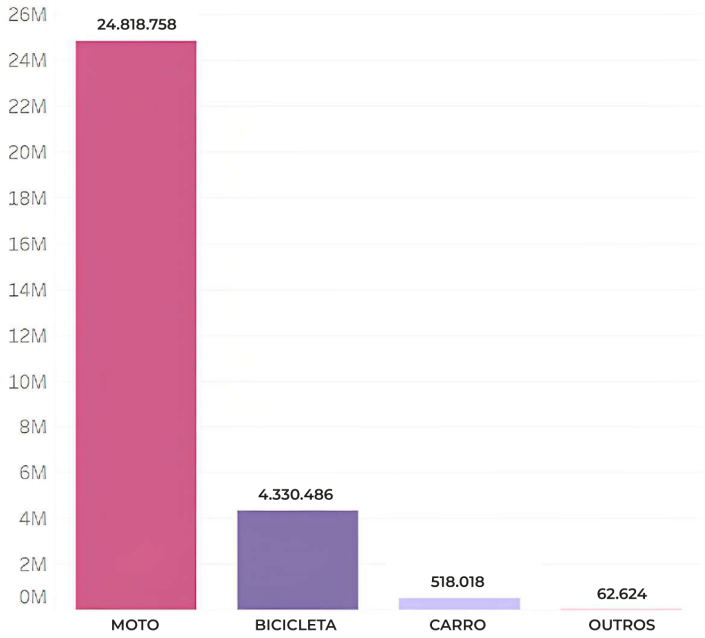

Figura 4 – Quantidade de rotas em um determinado período por modais de entrega

equivale a menos de 1% do total da base de dados. Este desbalanceamento é uma informação crucial, pois pode impactar a precisão dos modelos de previsão de sinistros. Modelos treinados em dados desbalanceados podem tender a prever melhor a categoria mais representada, neste caso, os 'não sinistros', levando a uma possível subestimação dos sinistros. Para um negócio, isso pode significar uma falha em identificar adequadamente os riscos, resultando em estratégias de mitigação de riscos inadequadas e potencial perda financeira.

	BICICLETA	MOTO	OUTROS
SEM SINISTRO	4.330.469	24.818.379	62.624
COM SINISTRO	17	379	

Figura 5 – Modais de Entrega na visão de sem Sinistro e com Sinistro

Além disso, foi observado que as rotas ocorrem predominantemente durante a noite e tarde (Figura 6). Essa tendência temporal tem implicações diretas para a gestão de riscos e para a otimização de operações. Por exemplo, pode ser necessário implementar medidas de segurança adicionais durante esses períodos, como iluminação melhorada ou monitoramento mais rigoroso. Do ponto de vista operacional, isso pode significar a realocação de recursos ou ajustes nos horários de trabalho para garantir uma operação mais segura e eficiente.

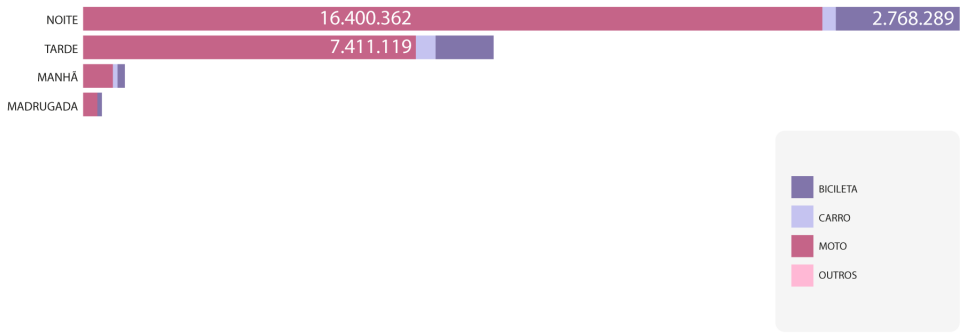

Figura 6 – Turno por Rota segregado por tipos de modais

Por fim, a constatação de que a informação é a mesma para rotas com sinistros (Figura 7) sugere que os sinistros estão distribuídos uniformemente ao longo das rotas, independentemente de características específicas da rota. Isso pode indicar que os fatores de risco para sinistros são mais relacionados a variáveis externas, como condições climáticas ou comportamentais dos condutores, do que às características intrínsecas das rotas. Para o negócio, isso implica na necessidade de focar as estratégias de prevenção de sinistros em fatores externos, como treinamento de motoristas para condições adversas ou investimento em tecnologias de monitoramento e prevenção de acidentes.

Essas observações, portanto, fornecem insights valiosos que podem ser utilizados para aprimorar a tomada de decisões, otimizar operações e melhorar estratégias de gestão de riscos, contribuindo assim para a eficiência operacional e a rentabilidade do negócio.

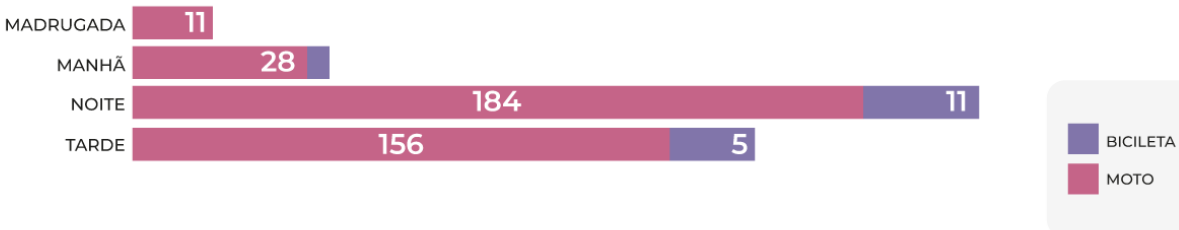

Figura 7 – Turno pelos modais de entregas na base de sinistro.

Cerca de 93% das rotas consistem em *SINGLE PICK SINGLE DROP* (SPSD) conforme visto na figura 8, o que significa que o entregador parte de uma origem para um destinatário.

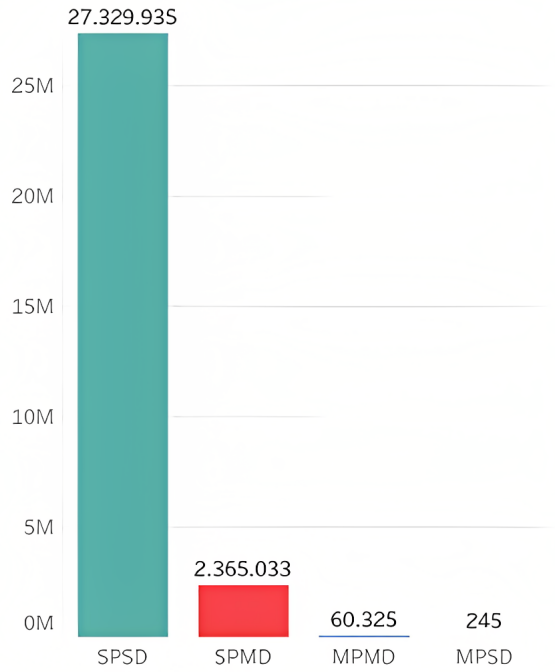

Figura 8 – Formas de Entrega - Siglas.

3.1.3 Análise das informações de entregadores

Com relação a faixa etária, está entre 18 e 30 anos. Entretanto, é importante destacar que a base de dados atual apresenta inconsistências, uma vez que 38% das informações sobre a idade dos entregadores não foram preenchidas. Há informações na base de dados de entregadores com idade superior a 80 anos, levantando dúvidas sobre a confiabilidade desse campo na tabela. Atualmente, o processo de armazenamento de informações dos entregadores é realizado por meio de uma tecnologia de OCR (Optical Character Recognition), que extrai informações da carteira de motorista do entregador através de uma foto. Entretanto, a qualidade da imagem pode influenciar na leitura, resultando em registros com erros devido à falha de leitura da imagem.

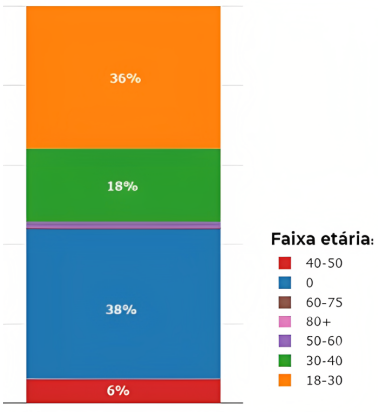

Figura 9 – Faixa de Idade Entregadores.

Devido à pouca informação disponível sobre o entregador na base de dados, que se limita às informações presentes apenas na carteira de motorista, e em cumprimento à LGPD (Lei Geral de Proteção de Dados), que mantém o sigilo das informações pessoais, foi decidido que o foco seria nas informações estatísticas apenas das rotas.

3.1.4 Teste de hipótese

O procedimento estatístico conhecido como teste de hipótese consiste em analisar uma suposição feita sobre um parâmetro populacional. Seu objetivo é avaliar a probabilidade dessa

hipótese ser verdadeira, utilizando dados de uma amostra. Essa amostra pode ser proveniente de uma população maior ou de um processo de geração de dados. Para realizar os testes com as variáveis qualitativas foi utilizado o Qui-Quadrado. O Teste Qui-Quadrado de Independência avalia á associação entre variáveis categóricas (ou seja, analisando sua independência ou relação), neste caso é um teste não paramétrico. Este teste utiliza uma tabela de contingência para analisar os dados. Uma tabela de contingência é um arranjo no qual os dados são classificados de acordo com duas variáveis categóricas. As categorias de uma variável aparecem nas linhas e as categorias da outra variável aparecem nas colunas. Cada variável deve ter duas ou mais categorias. Cada célula reflete a contagem total de casos para um par específico de categorias. O p-value para cada termo avalia a hipótese de que o coeficiente não tem diferença em relação a zero (ausência de impacto). Se o p-value for pequeno (< 0.05), a hipótese inicial pode ser descartada. Isso sugere que um termo com p-value reduzido é provavelmente relevante para o modelo, pois variações nesse termo influenciam a variável resposta. Contudo, um p-value elevado (> 0,05 não significativo) indica que as alterações na variável não afetam diretamente a resposta.

significance level: 0,05 -> 95%

- **regiao_logistica_**: 0,0026
- **entregador_modal**: 0,0002
- **entregador_lo**: 0,0042
- **turno**: 0,0016
- **segmentacao**: 0,0384
- **rota_estado**: ’1,2765e-20
- **rota_modelo**: 0,0270
- **tipo_rota**: 0,1526

Foi adotada a regressão logística para as variáveis quantitativas. A análise de regressão é empregada para estabelecer uma equação que descreva a relação estatística entre uma ou mais variáveis e a variável de resposta. No modelo adotado, foi considerado o resultado binário, que corresponde aos valores "sinistro"(1) ou "não sinistro"(0). Os campos utilizados foram (p-valor):

‘

- **tempo_minuto**: 0,0001
- **media**: 0,0002
- **nome_promocao**: 0,9241
- **multiplicador_frete**: 0,6940
- **distancia_origem_ate_destino**:0,7331
- **distancia_ate_origem**: 0,9630
- **fee**: 0,0063
- **total_distancia_ate_destino**: 0,0059
- **total_distancia_ate_origem**: 0,0438
- **total_rota_duracao**:0,2461
- **max_rank**:0,0035
- **max_rota_atraso**:0,0076

Com base nos ajustes da regressão logistica, do teste de hipótese e na análise descritiva foram identificadas aquelas que apresentaram significância estatística:As variáveis significativas foram: **tempo_minutos, taxa, media, total_distancia_origem_ate_destino, total_distancia_ate_origem, max_rank, max_rota_atraso, regiao, modal, entregador_lo, turno, segmentacao, rota_estado e rota_modelo**.

3.1.5 Análise de outliers

A base de dados das rotas é caracterizada por uma grande dispersão, o que pode afetar negativamente diferentes etapas do processo, tais como a clusterização, a amostragem e a modelagem. Para minimizar o impacto da alta variabilidade dos dados e reduzir possíveis erros, foi realizada uma análise dos valores discrepantes, utilizando a fórmula do **z-score**.

$$z = \frac{x - \mu}{\sigma}. \tag{3.1}$$

Para aplicar a fórmula do z-score, foi necessário centralizar a média dos dados em zero e definir o desvio padrão como 1. Dessa forma, qualquer valor acima ou abaixo desse limite (3 ou -3 desvios padrões da média) é considerado um outlier e pode ser identificado. Após a identificação desses valores discrepantes, eles foram removidos da base de dados para garantir que as análises subsequentes fossem mais precisas e confiáveis.

3.2 Amostragem

Durante o processo de análise exploratória, foi possível identificar um desequilíbrio nos dados, uma vez que a base de sinistros apresentava um número reduzido de registros, em comparação à base de rotas sem ocorrências de sinistros, que era bastante extensa. Dessa maneira, situações podem ocorrer em que o estudo com toda a população em análise (ou seja, a base de rotas sem sinistro) se torne inviável ou indesejável, tornando-se necessária a extração de um subconjunto representativo da população, conhecido como amostra. A amostragem é fundamental para assegurar a representatividade dos resultados obtidos, os quais, por meio de procedimentos estatísticos apropriados, podem ser utilizados para inferir, generalizar ou tirar conclusões acerca da população em questão.

Devido aos recursos computacionais limitados e ao desbalanceamento evidente dos dados, que implica em variabilidade nos eventos observados, foi utilizada uma técnica de ***Undersampling***.

O *Undersampling* é um método utilizado para lidar com desbalanceamento de classes em que a classe majoritária é reduzida para diminuir a disparidade entre as categorias. Devido às limitações da biblioteca Mlib do Spark, ferramenta de uso geral na empresa, foi necessário utilizar um conceito semelhante ao algoritmo **NNRU**.

O algoritmo NNRU é uma técnica de undersampling que tem como objetivo equilibrar classes desbalanceadas em um conjunto de dados. O algoritmo baseia-se no conceito de que exemplos similares tendem a pertencer à mesma classe.

O processo do NNRU envolve a identificação dos exemplos da classe majoritária que possuem um ou mais vizinhos próximos na classe minoritária, ou seja, exemplos da classe

minoritária que estão próximos aos exemplos da classe majoritária. Em seguida, os exemplos da classe majoritária são removidos até que o número de exemplos das duas classes seja equilibrado.

Com base na premissa do algoritmo NNRU, foi utilizado o método não supervisionado *Kmeans*. Esse algoritmo define centroides aleatórios e calcula a distância dos pontos da base de dados em relação a esses centroides. Os pontos são agrupados em *clusters* com base nas menores distâncias dos centroides. Como os grupos formados pela clusterização possuem similaridade, a redução de informações nesses grupos resulta em uma perda mínima de informações, uma vez que um grupo resume as informações necessárias que ele pode adicionar ao estudo.

Após a clusterização, foi utilizada outra técnica de amostragem chamada de "amostragem estratificada". Nesse tipo de amostragem, a população heterogênea é dividida em subpopulações ou estratos homogêneos (como os *clusters* formados na clusterização usando *Kmeans*). Em cada estrato, é retirada uma amostra. O número de estratos é definido inicialmente, e o tamanho de cada um é obtido. Em seguida, é especificado o número de elementos a serem retirados da subpopulação de cada estrato, podendo ser uma alocação uniforme ou proporcional.

3.2.1 Clusterização e balanceamento dos dados

Na base de dados disponível, é possível observar que existem mais 27 milhões de rotas sem sinistro e apenas 396 rotas com sinistros, o que indica uma problemática de desbalanceamento de classes. Essa situação é comum em problemas de classificação, nos quais as classes não são representadas igualmente. Embora a maioria dos conjuntos de dados de classificação apresente uma pequena diferença na quantidade de instâncias em cada classe, em alguns casos, o desequilíbrio é esperado.

Para solucionar esse problema de desbalanceamento, foi empregada a técnica de clusterização. A análise de *cluster*, ou *clustering*, tem como objetivo agrupar objetos de forma que os objetos em um mesmo grupo sejam mais similares entre si do que em outros grupos (clusters).

Na presente pesquisa, foi utilizada a técnica de clusterização k-means para lidar com o problema de dados desbalanceados identificado na base de dados. O *kmeans* é um algoritmo de clusterização amplamente utilizado, que visa agrupar um conjunto de objetos em k *clusters*, de forma que os objetos dentro de cada cluster sejam similares entre si e diferentes dos objetos em outros *clusters*.

A determinação do número ideal de *clusters* é uma etapa crítica na aplicação do k-means. Para isso, foi utilizado o método do cotovelo (*elbow*), que consiste em traçar um gráfico (Figura 10) da soma das distâncias quadráticas entre cada ponto e seu centróide mais próximo em relação ao número de *clusters*. O número de *clusters* é selecionado no ponto de inflexão da curva, onde a adição de mais *clusters* não resulta em uma redução significativa na soma das distâncias quadráticas.

O método do cotovelo é uma técnica comum para determinar o número ótimo de *clusters*

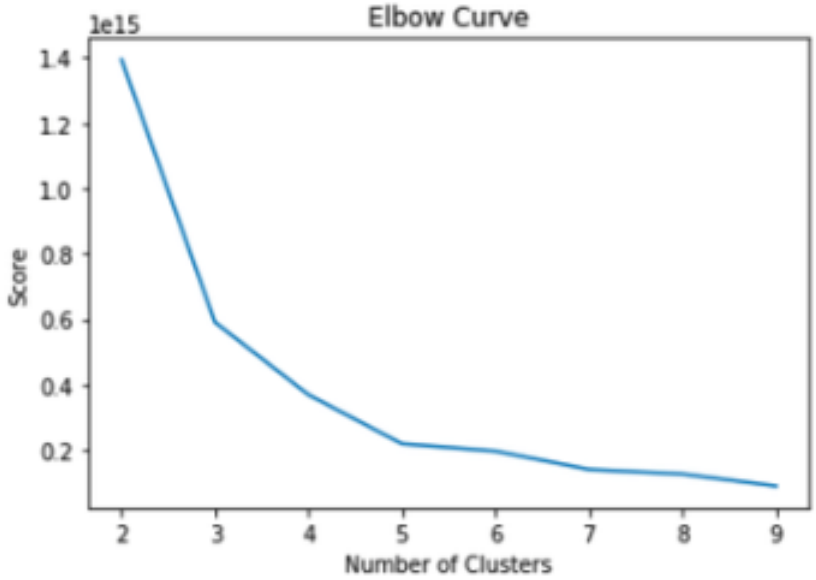

Figura 10 – Gráfico de saída do método do cotovelo

em um conjunto de dados durante a aplicação do algoritmo de agrupamento *kmeans*.

Especificamente, este método envolve a execução do algoritmo de K-means para diferentes valores de k, enquanto se mede a soma dos erros quadráticos para cada valor de k. É importante notar que, para garantir a eficácia do algoritmo, apenas as variáveis quantitativas foram consideradas em um conjunto de dados contendo mais de 27 milhões de rotas sem sinistro.

Com base nesses resultados, foram definidos 5 clusters para a análise (Tabela 1). Essa abordagem proporcionou uma visão mais clara e concisa dos padrões presentes nos dados, permitindo que sejam tomadas decisões mais precisas e informadas.

Tabela 1 – Resultado da Clusterização

Cluster	**Frequência**
2	8118906
0	8609076
1	7647725
4	1132151
3	4247284

O resultado do *kmeans* foi utilizado para gerar uma amostra estratificada da base de dados. A estratificação é uma técnica estatística utilizada para dividir uma população heterogênea em subpopulações ou estratos homogêneos. Neste caso, a estratificação foi realizada utilizando o método *kmeans*. A partir da estratificação, uma amostra é retirada de cada estrato. A fórmula utilizada para a obtenção da amostra estratificada é:

$$n_i = \frac{N_i}{N} n \tag{3.2}$$

Em que:

- n_i: subestrato
- N_i: estrato
- N: total da população

Para a seleção das amostras no substrato, foi empregado um método de amostragem aleatória simples com o objetivo de obter uma quantidade de registros próxima ao número de registros contidos na base de sinistros fornecida pela seguradora. Quando algoritmos de *machine learning* e *AI* (*Artificial Intelligence*) são utilizados em métodos de classificação, é comum buscar uma base de treinamento balanceada, com a mesma proporção de amostras para cada classe. Em bases de dados com classes dicotômicas, essa proporção seria de 50% para cada classe. Para obter essa base de treinamento balanceada, é necessário ajustar o tamanho da amostra da classe majoritária (rotas sem sinistros) para o tamanho da classe minoritária (rotas com sinistros). A partir da aplicação da fórmula mencionada anteriormente, foi possível reduzir a base de dados e chegar ao tamanho de amostra ideal (no caso optou-se pelo quantidade equivalente a base de rotas com sinistro) para o uso dos algoritmos de *Decision Tree* e *Logistic Regression*.

CAPÍTULO

4

MODELAGEM

4.1 Multicolinearidade nos dados

Antes de iniciar o modelo com a árvore de decisão, realizou-se um teste de Multicolinearidade, que é um desafio frequentemente encontrado em regressões, onde as variáveis independentes mostram correlações lineares precisas ou quase precisas. Para isso, construiu-se uma matriz de correlação das variáveis quantitativas (Figura 11). É possível observar uma correlação significativa entre a variável "total_route_duration"e as variáveis "total_distance"e "média"e também entre "Temp_minutos"e "média". Na primeira relação, optou-se por considerar a variável "total_distance", enquanto na segunda relação, optou-se por considerar a variável "media". Foi adotado um valor acima de 0,5 como critério de confirmação da relação entre as variáveis citadas. Essas variáveis possuem a mesma fonte de origem, ou seja, representam a mesma informação.

4.2 Desenvolvimento da modelagem

Este algoritmo da Árvore de decisão foi empregado para gerar perfis de sinistros e validar a consistência dos dados. Além disso, o algoritmo oferece um método de validação das variáveis utilizadas, conhecido como "feature importance".

Com base na amostra gerada pelo procedimento de balanceamento mencionado, foram selecionadas as colunas (*features*, variáveis) a serem utilizadas no modelo de árvore de decisão. A construção dos nós da árvore é determinada por alguns métodos:

- Entropia: Através da entropia, o algoritmo analisa a distribuição dos dados nas variáveis preditoras em relação à variação da variável alvo. Uma entropia elevada indica maior dispersão dos dados, enquanto uma baixa sinaliza uma organização mais definida desses dados quando analisados em relação à variável alvo.

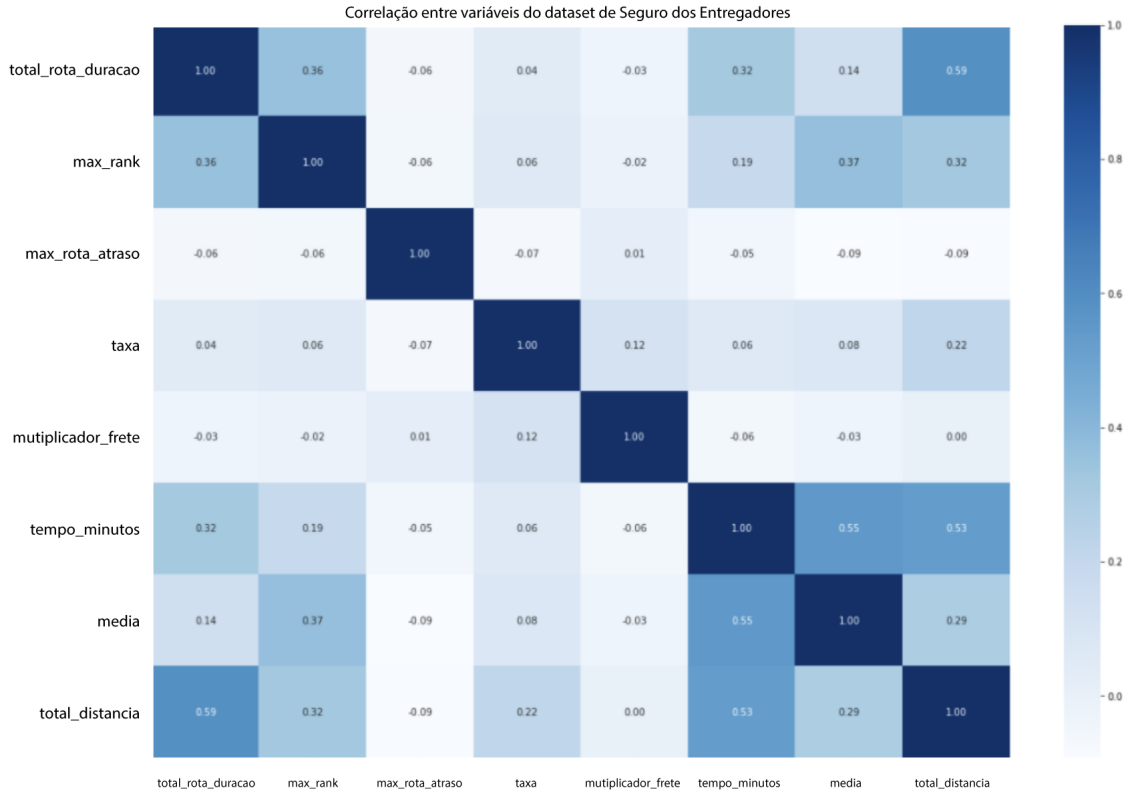

Figura 11 – Matriz de Correlação das variáveis

- Índice de Gini: Assim como a entropia, o cálculo do índice de Gini também verifica a distribuição dos dados nas variáveis preditoras em relação à variação da variável alvo, porém utiliza um método diferente.

- Regressão: Nos problemas de regressão, o objetivo é prever um valor, e não uma classe. Para isso, a árvore utiliza os conceitos de média e desvio padrão, permitindo a obtenção de um resultado final numérico.

Esses conceitos são aplicados tanto na construção da árvore quanto na função de *Feature Importance*. Essa função, como o próprio nome sugere, resume de forma concisa quais *features* utilizadas no modelo são mais relevantes e resumem de maneira eficaz o cenário que está sendo modelado, sem perda de informação. Para o modelo em questão, foram apresentados os seguintes resultados da *feature importance*:

A importância das variavéis (ou *feature importance*) em uma árvore de decisão é calculada com base na contribuição de cada variavél para a melhoria do modelo. Em termos técnicos, para cada nó da árvore que faz uma divisão baseada em uma variável específica, o grau de redução da impureza (geralmente medido por Gini ou entropia para problemas de classificação) é considerado. Esta redução de impureza é acumulada para cada característica ao longo de todas as árvores na floresta (em modelos de Random Forest) ou em uma única árvore. A importância

Tabela 2 – Visão Geral - *Features Importance* no modelo desenvolvido

	Variável	**Importância**
19	Motocicleta	35,049406
8	Noite	24,224920
4	media	20,413633
6	Madrugada	9,054848
5	total_distancia	5,430839
1	max_rank	3,626361
0	total_rota_duracao	2,199993
13	Inativo	0,000000
18	Carro	0,000000
17	Bicicleta	0,000000
16	sem_segmentacao	0,000000
15	Super	0,000000
14	Professional	0,000000
10	Casual	0,000000
12	Churn Passivo	0,000000
11	Churn	0,000000
9	Tarde	0,000000
7	Manhã	0,000000
3	multiplicados_frete	0,000000
2	max_rota_atraso	0,000000
20	Moto com Reboque	0,000000

final de uma variável é então normalizada, de forma que a soma de todas as importâncias das variavéis seja igual a 1. Este método fornece uma visão intuitiva sobre quais características são mais importantes na previsão do modelo.

É possível observar na Tabela 2 que algumas *features*, de acordo com o modelo, não demonstraram ganho de informação suficiente. Com os parâmetros ajustados, como o nível de profundidade da árvore de decisão, as *features* "Motocicleta", "Noite", "média", "Madrugada", "total_distancia", "max_rank"e "total_rota_duracao"resumiram o cenário de previsão de sinistros de forma mais aderente.

Após o treinamento do modelo, foi gerada a matriz de confusão (base de teste) (Figura 12). Vale ressaltar que a matriz de confusão é uma ferramenta utilizada em modelos de classificação. Trata-se de uma tabela que mostra as frequências de classificação para cada classe do modelo. Considerando o exemplo mencionado acima, ela nos fornecerá as seguintes frequências:

- Positivo Verdadeiro (TP - *True Positive*): refere-se à situação em que a classe alvo, dentro do conjunto real, foi corretamente identificada pelo modelo.
- Positivo Falso (FP - *False Positive*): acontece quando a classe que estamos tentando identificar, no conjunto real, é erroneamente reconhecida pelo modelo.

- Negativo Verdadeiro (TN - *True Negative*): é quando a classe que não é nosso objetivo principal, no conjunto real, é corretamente classificada pelo modelo.
- Negativo Falso (FN - *False Negative*): ocorre quando a classe que não estamos tentando identificar, dentro do conjunto verdadeiro, é erroneamente reconhecida pelo modelo.

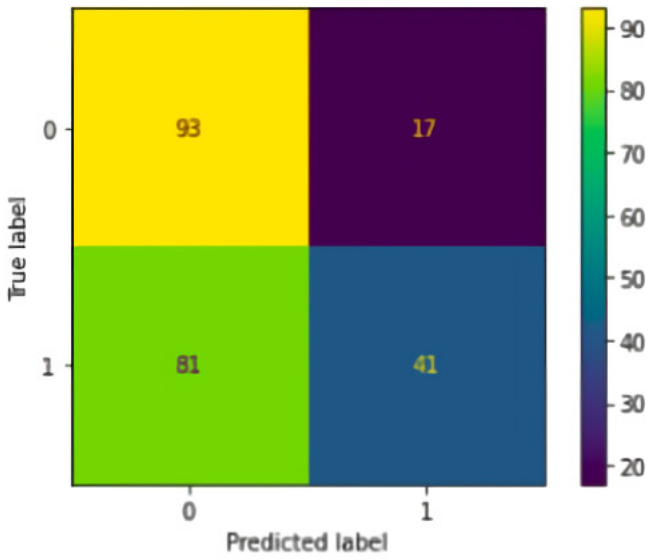

Figura 12 – Matriz de Confusão

Foram avaliadas as seguintes métricas para validar a aderência do modelo: Métricas de avaliação do modelo:

- A acurácia informa, em geral, o quão é preciso o modelo..
- A precisão ou VPP (Valor Preditivo Positivo) pode ser interpretada como: dos casos classificados como certos, quantos eram realmente certos?
- O *recall* ou Sensibilidade indica com que frequência o classificador encontra exemplos de uma classe. Se um exemplo pertence àquela classe, o *recall* nos informa com que frequência ele é classificado corretamente.
- O F1 escore combina o *recall* com a precisão de modo a fornecer um único número representativo.

A métricas desta Árvore de decisão referentes a matriz de confusão são as seguintes:

E a seguinte Arvore de Decisão foi gerada:

A árvore de decisão em questão (Figura 13) foi gerada em meio a diversos outros cenários analisados, sendo testado por diferentes parâmetros na árvore (profundidade da árvore, critério:

Tabela 3 – Métricas de avaliação do modelo para a Árvore de Decisão (Output Python)

	precision	recall	f1-score	support
0	0,53	0,85	0,65	110
1	0,71	0,34	0,46	122
accuracy			0,58	232
macro avg	0,62	0,59	0,56	232
weighted avg	0,63	0,58	0,55	232

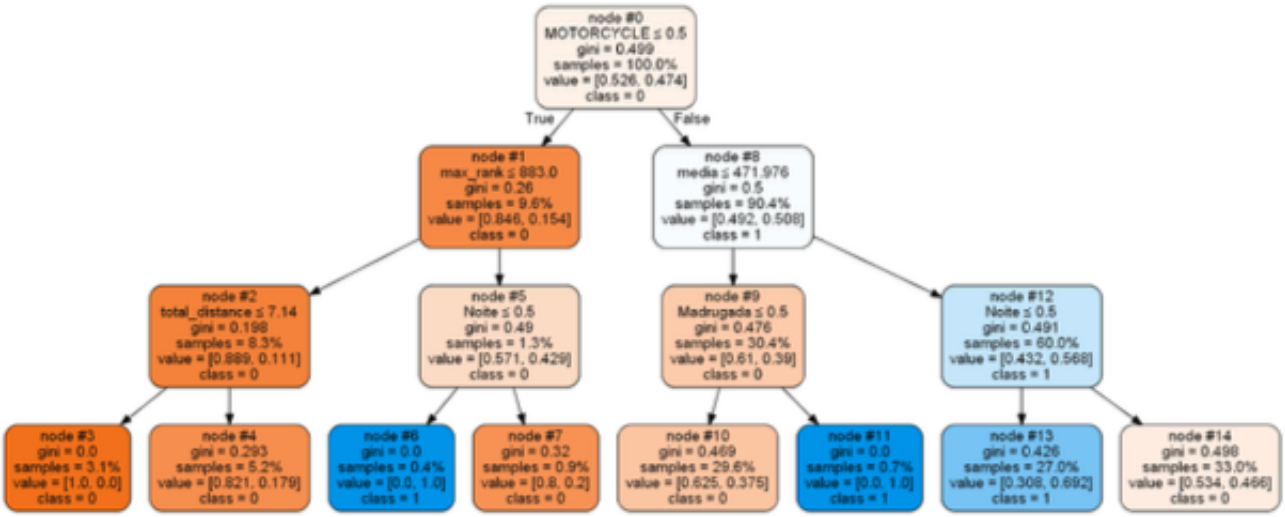

Figura 13 – Arvore De Decisão Seguro de Entregadores

Gini ou Entropia, etc),variavéis (como por exemplo: de localidade, operador, etc). A validação da árvore foi realizada junto à área de negócio, e uma vez que apresentou uma lógica que refletia a realidade do empreendimento, esta árvore acima foi a escolhida, apesar de algumas métricas de validação do modelo,presentes na tabela 3, poderem ser aprimoradas . A avaliação criteriosa da árvore levou em consideração sua capacidade de capturar as características e o comportamento dos dados relevantes para a tomada de decisões no contexto da área de negócio, especificamente da área financeira e operacional. Essa validação junto à área de negócio é essencial para garantir que o modelo seja capaz de fornecer *insights* acurados e úteis para o melhor direcionamento das estratégias e ações em relação as negociações com a seguradora parceira.

4.2.1 Generalizando o modelo

Considerando a limitação no tamanho da base de dados de sinistros, o modelo foi adaptado e aplicado a um conjunto de rotas sem registros de sinistros. Esta generalização implica na aplicação do modelo a toda a base de dados, sem a imposição de filtros específicos, simulando sua utilização em condições operacionais cotidianas. O objetivo desta estratégia é avaliar o desempenho do modelo na classificação das rotas sob uma perspectiva prática, antecipando seu funcionamento em um ambiente de produção. Os resultados obtidos neste contexto foram os

seguintes:

Tabela 4 – Resultado das métricas para a generalização do modelo:(0 (Não sinistro) vs 1 (Sinistro))

	precision	recall	f1-score	support
0	1,00	0,91	0,95	480
1	0,00	0,00	0,00	0
accuracy			0,91	480
macro avg	0,50	0,46	0,48	480
weighted avg	1,00	0,91	0,95	480

Observou-se uma redução no F1-escore ao generalizar o modelo, indicando que as métricas de avaliação podem não ser ótimas. No entanto, é importante ressaltar que, mesmo diante dessa redução, a área de negócio validou o modelo e considerou que o cenário apresentado condiz com suas observações e experiência operacional. Essa validação foi embasada na percepção de que a árvore de decisão utilizada na generalização captura de forma significativa os padrões e características relevantes do contexto operacional.

Vale destacar que, embora o F1-score tenha sido impactado, não foi um valor significativo. Essa medida de acurácia, apesar de modesta, indica que o modelo ainda possui uma capacidade razoável de realizar previsões corretas (Tabela 4), considerando o conjunto de dados disponível.

A seguir, apresenta-se a Matriz de Confusão in Figura 14, que é uma ferramenta que permite visualizar as frequências de classificação para cada classe do modelo, fornecendo informações adicionais sobre o desempenho do modelo na tarefa de classificação.

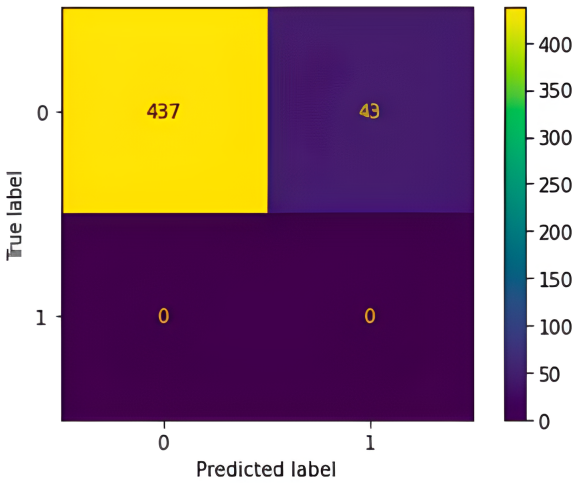

Figura 14 – Matriz de Confusão do Modelo Generalizado

4.2.2 Regressão logistica aplicadas aos clusters

Após a geração da árvore de decisão, os nós da árvore foram utilizados como possíveis grupos para definir perfis de entregadores e rotas. Esses nós foram aplicados à base de dados e posteriormente utilizados em um modelo de regressão logística.

O objetivo desse procedimento foi gerar probabilidades de ocorrência para cada um dos *clusters* identificados, utilizando a regressão logística. Através dessas probabilidades, foi possível obter um escore de risco para as rotas e entregadores.

A regressão logística, nesse contexto, permitiu realizar uma análise mais detalhada das características e dos padrões presentes nos *clusters* definidos pela árvore de decisão. Com base nessas análises, foi possível atribuir um escore de risco a cada rota e entregador, proporcionando uma medida quantitativa para avaliar o nível de risco associado a cada um deles.

Dessa forma, o uso combinado da árvore de decisão e da regressão logística permitiu uma abordagem mais sofisticada para a definição dos perfis e a avaliação do risco relacionado às rotas e entregadores, proporcionando informações valiosas para a tomada de decisões e o planejamento estratégico no contexto logístico. Como resultado é possível observar a matriz de confusão (Figura 15) usando mapa de calor:

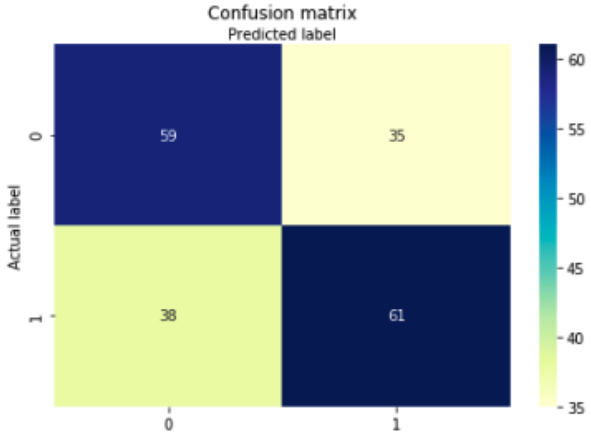

Figura 15 – Matriz de Confusão do Modelo de Regressão Logistica

As métricas de avaliação da matriz de confusão foram (base de teste):

- **Accuracy**: 0,621761
- **Precision**: 0,635416
- **F1-score**: 0,625641

A Curva Característica de Operação do Receptor (ROC) é uma representação gráfica (Figura 16) demonstrando a eficácia de um classificador binário conforme seu ponto de corte se altera. No modelo a curva de ROC obtida foi uma área sob a curva ROC (AUC) de 0,65 indica

um desempenho moderado do modelo na diferenciação entre as classes positivas e negativas, conforme resultado a seguir:

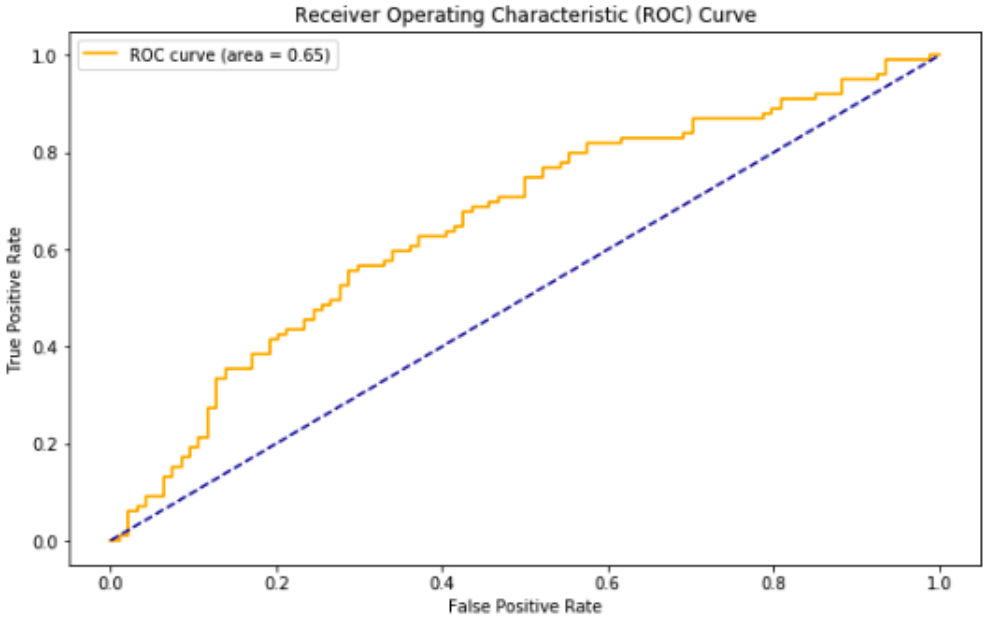

Figura 16 – Curva ROC do Modelo de Regressão Logística

Com o intuito de obter uma compreensão mais aprofundada dos resultados do modelo, procedeu-se à geração de um intervalo de confiança, considerando a agrupação por nó (folha) da árvore. Para realizar essa avaliação do modelo, foram geradas 20 amostras distintas, utilizando sementes (seeds) diferentes. O modelo foi aplicado a cada uma dessas amostras, agrupando os resultados por semente e nó, a fim de analisar as médias dos escores para cada folha e calcular o intervalo de confiança correspondente.

A geração de múltiplas amostras com diferentes sementes é uma prática comum para avaliar a estabilidade e a robustez do modelo. Ao realizar essa abordagem, é possível observar a variação dos resultados em diferentes conjuntos de dados de treinamento, o que proporciona uma visão mais abrangente do desempenho do modelo e permite identificar possíveis variações causadas pela aleatoriedade da seleção dos dados de treinamento.

Ao agrupar os resultados por semente e nó, é possível calcular as médias dos escores para cada folha e, posteriormente, estimar o intervalo de confiança correspondente. Esse intervalo de confiança fornece uma medida estatística que indica a precisão e a confiabilidade das médias calculadas, levando em consideração a variabilidade presente nas amostras.

Dessa forma, a geração das 20 amostras e o cálculo do intervalo de confiança, conforme tabela 5, são etapas fundamentais para a construção do escore e para a análise estatística do desempenho do modelo, permitindo uma avaliação mais robusta e confiável de suas capacidades preditivas.

Tabela 5 – Amostras dos nós (*Clusters*) - Resultado de Classificação

seed	nós	media_score	desvio_padrão	min	max
8	node 10	0,49	0,11	0,26	0,71
8	node 10	0,58	0,14	0,29	0,86
8	node 3	0,16	0,08	−0,00102	0,32
27	node 10	0,48	0,12	0,24	0,72
27	node 11	0,64	0,11	0,40	0,86
27	node 3	0,13	0,06	0,0011	0,26
32	node 10	0,48	0,11	0,25	0,70
32	node 11	0,66	0,07	0,51	0,79
32	node 3	0,15	0,08	−0,012	0,31
59	node 10	0,50	0,12	0,25	0,73
59	node 11	0,63	0,14	0,35	0,91
59	node 3	0,15	0,07	0,009	0,29
93	node 10	0,49	0,11	0,26	0,72
93	node 11	0,57	0,10	0,36	0,76
93	node 3	0,15	0,07	0,004	0,28
404	node 10	0,50	0,11	0,27	0,72
404	node 11	0,71	0,07	0,56	0,85

4.2.3 Implicações das variáveis do modelo de regressão logística na probabilidade de sinistros

Neste modelo, exploramos várias variáveis quantitativas e qualitativas para entender melhor suas implicações na probabilidade de sinistros em rotas de transporte. As variáveis como "Time minute", "Average time", "Fee", "Total distance to destination", "Total distance to origin", "Max rank"e "Max route lag"mostraram significância estatística, indicando uma forte relação com a ocorrência de sinistros.

Time Minute e Average Time: Observamos que tempos mais longos, tanto em minutos quanto na média, podem ser indicativos de rotas mais complexas ou sujeitas a congestionamentos frequentes. Essas condições podem elevar o risco de sinistros, seja por aumentar a exposição a potenciais perigos ou por induzir fadiga ao motorista.

Fee: A taxa associada à rota pode refletir sua complexidade ou risco. Rotas com taxas mais altas podem ser aquelas que atravessam áreas com maior probabilidade de acidentes ou desafios logísticos, sugerindo uma correlação direta entre a taxa cobrada e a probabilidade de sinistros.

Total Distance to Destination e Total Distance to Origin: As distâncias maiores, tanto para o destino quanto a partir da origem, sugerem um tempo prolongado na estrada. Isso aumenta não apenas a probabilidade de exposição a condições variadas e potencialmente perigosas de trânsito, mas também amplifica a chance de fadiga do motorista, elevando assim o risco de sinistros.

Max Rank e Max Route Lag: Estes indicadores podem ser interpretados como reflexos da experiência e eficiência do motorista. Valores mais elevados nestas variáveis podem indicar

motoristas com menor experiência ou eficiência operacional, o que pode contribuir para um aumento na probabilidade de ocorrência de sinistros.

Além dessas variáveis quantitativas, as variáveis qualitativas, incluindo região, modal, tipo de trabalhador, turno, segmentação, estado da rota e modelo da rota, também demonstraram impacto significativo na ocorrência de sinistros. Estes fatores podem estar relacionados com aspectos culturais, operacionais e ambientais específicos das rotas, que necessitam de uma análise mais aprofundada para entender completamente suas implicações.

Em suma, a análise das variáveis em nosso modelo de regressão logística oferece insights valiosos sobre os fatores que influenciam a probabilidade de sinistros em rotas de transporte. Essas descobertas são cruciais para o desenvolvimento de estratégias eficazes de gestão de riscos e segurança nas operações logísticas.

CAPÍTULO

5

RESULTADOS

5.1 Análise de resultados

Com base nos resultados obtidos, surgem duas informações relevantes:

- Cada rota foi classificada em nós (*nodes*), seguindo a estrutura da árvore de decisão;
- Para cada nó, foi atribuída uma probabilidade de ocorrência de sinistro, variando de 0 a 1, em que 1 representa uma chance de sinistro de 100%. Essa medida é denominada de escore.

Foram realizadas 10 amostras, cada uma contendo 400.000 rotas, com o propósito de aplicar o modelo de precificação. A primeira análise realizada sobre essas amostras foi o cálculo do desvio padrão, considerando uma matriz de nós por sementes. Concluiu-se que as variações entre os nós e as sementes eram mínimas, o que indicou que as amostras poderiam ser consideradas relativamente homogêneas.

A partir desse ponto, uma das 10 amostras selecionadas foi explorada para compreender de que forma os nós poderiam ser agrupados em classificações de risco mais relevantes para o contexto do negócio (Tabela 6).

Como resultado dessa exploração, foram identificadas três classificações de risco: alto, médio e baixo. Essas classificações permitiram uma abordagem mais tangível e adequada às necessidades da empresa.

Uma ressalva importante a ser destacada é que, em relação ao negócio, não se acredita que a proporção de sinistros possa chegar a 11% (Tabela 7). No entanto, conforme indicado pelo modelo, as corridas classificadas como representando 11% do total possuem um potencial maior de ocorrência de acidentes. Por outro lado, as corridas classificadas como 3, 4, 6 e 7 apresentam uma probabilidade de risco extremamente baixa.

Tabela 6 – Uma tabela resumida da definição de risco de acordo com a análise da amostra

Nós	Media Score	Número de Rotas	Coeficiente Gini	Total de Participação
node 10	0,44	129734	0,53	32%
node 11	0,61	1829	0,00	0,46%
node 13	0,64	42621	0,57	11%
node 14	0,47	162960	0,50	41%
node 3	0,14	11572	0,00	3%
node 4	0,14	44106	0,71	11%
node 6	0,23	1445	0,00	0%
node 7	0,14	6160	0,68	2%
TOTAL	**0,35**	**400427**	**0,37**	**100%**

Tabela 7 – Definição dos três níves de classificação de Risco

Nível de Risco	% da base	Probabilidade de Sinistro
Baixa probabilidade	16%	20%
Média probabilidade	73%	45%
Alta probabilidade	11%	63%

Embora o valor de 11% tenha servido como referência para as corridas com alta probabilidade de risco, ele ainda não pode ser considerado como a segmentação final de nossa base a ser apresentada à seguradora. É necessário avaliar se as demais amostras (Tabela 8) utilizadas no experimento apresentaram um comportamento semelhante em relação aos casos críticos (**alto risco**).

Tabela 8 – Comparação de risco entre as amostras

Amostras	Alto Risco	Baixo Risco	Médio Risco
164	11,00%	15,80%	73,20%
223	18,14%	16,71%	65,15%
251	14,89%	16,47%	68,64%
292	15,50%	16,69%	67,81%
410	17,50%	17,11%	65,39%
522	15,73%	16,66%	67,62%
588	15,99%	16,73%	67,29%
590	17,70%	16,74%	65,55%
791	18,66%	16,88%	64,46%
887	17,03%	16,58%	66,39%

Diante dos resultados obtidos, foi decidido trabalhar com o valor médio das rotas classificadas como alto e baixo risco. Para as rotas de médio risco, calculou-se a diferença em relação a 100% de modo a obter compatibilidade entre os valores.

Essa segmentação desempenha um papel fundamental no estudo, pois evidencia a necessidade de dividir os valores pagos de acordo com a criticidade e o grau de risco de cada rota. Isso

permitirá uma redução significativa nos custos, em comparação com o cenário atual, no qual se assume o máximo de risco - e, consequentemente, o preço mais elevado - para todas as rotas.

5.1.1 Definição de precificação rota

Definidas as classes de risco e a participação de cada uma delas nas amostras, tornou-se necessário estabelecer o preço considerado justo para cada classe, de modo que os preços, ponderados pela quantidade de rotas, pudessem determinar um preço único a ser utilizado na negociação com a seguradora.

Com o intuito de gerar dados para a precificação, foram explorados os registros de sinistros ocorridos entre janeiro de 2020 e fevereiro de 2021, auge da pandemia de covid-19. Os principais resultados encontrados foram os seguintes:

O valor médio pago por rota nos últimos meses foi de USD 0,08. O custo médio dos sinistros foi de USD 0,05. Esse valor foi obtido por meio da razão entre a soma dos sinistros indenizados e o total de sinistros em aberto em um determinado mês, e o número total de rotas realizadas no mesmo período. Para obter o custo médio histórico dos sinistros, esses valores mensais foram aplicados utilizando uma média ponderada.

É importante ressaltar que esse valor engloba tanto as indenizações já efetuadas quanto os sinistros em aberto. Portanto, é provável que esse preço seja ainda mais baixo, uma vez que apenas uma parte dos sinistros em aberto será convertida em indenização. Isso reforça a hipótese de que o produto está sendo precificado em excesso, resultando em um "*spread*"de USD 0,04.

Com base nessas conclusões, foi proposta a seguinte estrutura de preços:

Alto risco: para as rotas de alto risco, sugere-se manter o pagamento no valor de USD 0,08, que corresponde à precificação atualmente adotada para todas as rotas.

Baixo risco: para as rotas com baixa probabilidade de acidentes, propõe-se o pagamento do valor do custo atual das rotas, acrescido dos sinistros em aberto, totalizando USD 0,05.

Médio risco: foi sugerido um valor intermediário de USD 0,06 para a precificação das rotas com probabilidade média de risco.

Tabela 9 – Estrutura de classificação para os diferentes tipos de Risco

Perfil de Risco	**%**	**Preço (USD)**
Baixo Risco	17%	0,05
Médio Risco	67%	0,06
Alto Risco	16%	0,08
Média Ponderada		0,06

Por fim, o preço único final (Tabela 9) foi calculado com base na estrutura de precificação previamente mencionada e na segmentação da amostra, resultando em um valor final de USD

0,06 por rota.Um ponto essencial sobre precificação que foi mencionado é o seguinte: uma anonimização de dados foi aplicada aos resultados para salvaguardar a confidencialidade dos valores e negociações. Essa etapa foi tomada para garantir a proteção dos dados sensíveis durante o processo de avaliação, evitando que informações específicas sejam comprometidas ou utilizadas indevidamente. Isso demonstra um compromisso com a privacidade e a integridade do processo de negociação

CAPÍTULO

6

CONCLUSÃO

No cenário apresentado desta empresa de entrega global, diante de um contexto de crescimento exponencial devido à pandemia de Covid-19, o número de rotas de entrega aumentou substancialmente, como foi demonstrado pela análise de dados realizada. Esse aumento no volume de rotas resultou em um consequente aumento no valor de pagamento à seguradora, tornando necessária a adoção de um novo modelo de negócio.

Ao longo de todo o desenvolvimento do processo de coleta, análise, modelagem e consolidação dos resultados, a equipe de negócios desempenhou um papel ativo, validando as respostas e *insights* encontrados. É importante ressaltar que existem oportunidades de otimização em vários níveis do processo, como, por exemplo, na construção da árvore de decisão, por meio da utilização de outras variáveis ou da melhoria das métricas de avaliação do modelo. No entanto, como mencionado anteriormente, o modelo alcançado atende aos requisitos de negócios, conforme verificado por meio da construção do escore, que foi realizado integralmente pela equipe financeira.

Este estudo proporcionou insights valiosos sobre a precificação de seguros em uma empresa de entrega global, especialmente durante a pandemia de Covid-19. No entanto, é essencial reconhecer certas limitações. Primeiramente, o estudo focou em um contexto específico - um aumento substancial no número de rotas de entrega durante um período extraordinário. Isso pode limitar a generalização dos resultados para períodos normais ou para outras empresas em diferentes contextos. Além disso, o modelo dependeu significativamente de dados históricos, o que pode não refletir totalmente mudanças futuras nas tendências de entrega ou nos padrões de sinistros. Outra limitação importante é a potencial imprecisão nos dados, como foi observado com a idade dos entregadores. Isso destaca a importância de verificar e validar continuamente a qualidade dos dados utilizados em modelos preditivos

Para futuras pesquisas, seria benéfico expandir o escopo do estudo para incluir dados de múltiplas empresas de entrega em diversos contextos geográficos e econômicos. Isso ajudaria a

validar a eficácia do modelo em diferentes cenários e aumentar sua aplicabilidade. Além disso, uma análise mais aprofundada dos impactos de variáveis qualitativas, como comportamento do cliente e condições de trânsito, poderia enriquecer o modelo.Por fim, estudos longitudinais que acompanham as tendências de sinistros e padrões de entrega ao longo do tempo forneceriam uma compreensão mais rica das dinâmicas em jogo e ajudariam a refinar ainda mais os modelos de precificação de seguros.

Significativamente, os resultados do projeto foram apresentados à seguradora e facilitaram efetivamente a negociação de um novo contrato de precificação. A implantação do modelo proposto resultou em economias substanciais de custos, superando milhões de dólares em relação aos pagamentos anteriores. Esse resultado destaca a eficácia do novo modelo e representa o ponto culminante deste estudo, fornecendo uma solução robusta e eficiente para a empresa de entrega global.

Para concluir, este estudo expande a literatura acadêmica ao introduzir uma abordagem que aprimora a precisão do modelo, levando em consideração os aspectos pragmáticos do setor. A fusão de metodologias avançadas com conhecimento empírico ilumina o valor de estreitar a lacuna entre a compreensão teórica e a aplicação prática. As descobertas e *insights* elucidados neste trabalho servem como um recurso valioso para futuras pesquisas e empreendimentos práticos no campo da precificação de seguros e otimização de negócios.

REFERÊNCIAS

BRASIL. Lei nº 14.297, de 5 de janeiro de 2022. **Diário Oficial [da] República Federativa do Brasil**, Brasília, DF, 2022. ISSN 1677-7042. Disponível em: <https://legislacao.presidencia.gov.br/atos/?tipo=LEI&numero=14297&ano=2022&ato=7aaITQE1kMZpWT8b7>. Citado na página 13.

BROWNLEE, J. **A Simple Guide to Classification and Regression Trees (CART)**. 2018. <https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/>. Citado na página 21.

CAMPO, B. D.; ANTONIO, K. Insurance pricing with hierarchically structured data an illustration with a workers' compensation insurance portfolio. **Scandinavian Actuarial Journal**, Taylor & Francis, p. 1–32, 2023. Citado na página 14.

CANÔAS, V. d. L. Análise do cálculo da provisão de prêmios não ganhos nas sociedades seguradoras. 2007. Citado na página 14.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: Synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, AI Access Foundation, v. 16, p. 321–357, 2002. Citado na página 19.

CZEPIEL, S. A. Maximum likelihood estimation of logistic regression models: theory and implementation. **Available at czep. net/stat/mlelr. pdf**, v. 83, 2002. Citado na página 22.

DENUIT, M.; CHARPENTIER, A.; TRUFIN, J. Autocalibration and tweedie-dominance for insurance pricing with machine learning. **Insurance: Mathematics and Economics**, Elsevier, v. 101, p. 485–497, 2021. Citado na página 14.

EMBRECHTS, P.; KLÜPPELBERG, C.; MIKOSCH, T. **Modelling of Risk Processes**. [S.l.]: Springer, 2013. Citado na página 19.

FILHO, N. N. M. Exoneração de responsabilidade do segurador: Estudo tópico de direito securitário. In: JUNIOR, R. M. d. A. N. N. N. (Ed.). **Responsabilidade Civil - Direito de Obrigações e Direito Negocial**. [S.l.]: Editora Revista dos Tirbuinais, 2010. cap. 5, p. 43. Citado na página 14.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. [S.l.]: Springer Science & Business Media, 2009. Citado na página 21.

HE, H.; GARCIA, E. A. A comparison of resampling methods for imbalanced classification on medical datasets. **IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)**, IEEE, v. 39, n. 6, p. 1633–1649, 2009. Citado na página 19.

IZBICKI, R.; SANTOS, T. M. dos. **Aprendizado de máquina: uma abordagem estatística**. [S.l.: s.n.], 2020. ISBN 978-65-00-02410-4. Citado na página 20.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning: with Applications in R**. [S.l.]: Springer Science & Business Media, 2013. Citado na página 21.

JR, D. W. H.; LEMESHOW, S.; STURDIVANT, R. X. **Applied Logistic Regression**. [S.l.]: John Wiley & Sons, 2013. Citado na página 22.

KERCHER, S. **Gasto com delivery sobe 24consumo do pós-pandemia**. 2022. CNN Brasil Business. Disponível em: <https://pcdob.org.br/noticias/numero-de-entregadores-por-aplicativo-cresce-979-em-cinco-anos/>. Citado na página 13.

LI, R.; LI, Y.; CHEN, H.; HE, X.; SHANG, Z. Application of nearest neighbor rule undersampling in financial fraud detection. **Mathematics**, Multidisciplinary Digital Publishing Institute, v. 8, n. 4, p. 482, 2020. Citado na página 23.

MÜLLER, B. D. Prêmio puro dpvat 2021 após a resolução cnsp 399. 2022. Citado na página 18.

PCDOB65. **https://pcdob.org.br/noticias/numero-de-entregadores-por-aplicativo-cresce-979-em-cinco-anos/**. 2021. Disponível em: <https://pcdob.org.br/noticias/numero-de-entregadores-por-aplicativo-cresce-979-em-cinco-anos/>. Citado na página 13.

SCARAMUSSA, F. S.; SÁ, E. C. Indenizações pagas pelo seguro dpvat: perfil epidemiológico dos acidentes envolvendo motocicletas no brasil, contextualização das internações hospitalares e ônus ao sistema único de saúde (sus), no período de 2015 a 2018. **Saúde Ética & Justiça**, v. 25, n. 1, p. 10–14, 2020. Citado na página 17.

SEIFFERT, C.; KHOSHGOFTAAR, T. M.; HULSE, J. V.; NAPOLITANO, A. Random undersampling for imbalanced data classification. **Journal of Intelligent Systems**, Walter de Gruyter, v. 19, n. 2, p. 159–177, 2010. Citado na página 19.

SMITH, J. *et al.* Effective use of nearest neighbor rule undersampling in risk analysis. **Journal of Data Science**, v. 50, p. 123–139, 2022. Citado na página 15.

ZHANG, H.; DING, L.; WANG, H.; LI, X. Nearest neighbor-based undersampling for imbalanced classification and its application to sleep stage classification. **Biomedical Signal Processing and Control**, Elsevier, v. 39, p. 413–421, 2018. Citado na página 23.

APÊNDICE

A

REDUCING DELIVERY INSURANCE COSTS THROUGH RISK SCORE MODEL FOR FOOD DELIVERY COMPANY

Reducing Delivery Insurance Costs through Risk Score Model for Food Delivery Company

Diogo Silva Panham[1,*+] **and Pedro L. Ramos**[2,*+]

[1]Institute of Mathematical Science and Computing, University of São Paulo, São Carlos, Brazil
[2]Faculty of Mathematics, Pontificia Universidad Católica de Chile, Santiago, Chile
[*]pedro.ramos@mat.uc.cl
[+]these authors contributed equally to this work

ABSTRACT

In this paper, we propose a novel pricing model for delivery insurance in a food delivery company in Latin America, with the aim of reducing the high costs associated with the premium paid to the insurer. To achieve this goal, a thorough analysis was conducted to estimate the probability of losses based on delivery routes, transportation modes, and delivery drivers' profiles. A large amount of data was collected and used as a database, and various statistical and machine learning models were employed to construct a comprehensive risk profile and perform risk classification. Based on the risk classification and the estimated probability associated with it, a new pricing model for delivery insurance was developed using advanced mathematical algorithms and machine learning techniques. This new pricing model took into account the pattern of loss occurrence and high and low-risk behaviors, resulting in a significant reduction of insurance costs for both the contracting company and the insurer. The proposed pricing model also allowed for greater flexibility in insurance contracting, making it more accessible and appealing to delivery drivers. The use of estimated loss probabilities and a risk score for the pricing of delivery insurance proved to be a highly effective and efficient alternative for reducing the high costs associated with insurance, while also improving the profitability and competitiveness of the food delivery company in Latin America.

Introduction

Given its continental dimensions and large population, Brazil stands as a fertile ground for a substantial vehicular presence within its borders. In recent times, the emergence of the COVID-19 pandemic has further accelerated the utilization of specific transportation modes, notably motorcycles and bicycles, propelled by the widespread adoption of delivery apps as indispensable tools for work. Consequently, these alternative delivery modalities have played a pivotal role in facilitating both formal and informal employment opportunities. A research study conducted by the Institute of Applied Economic Research (IPEA), utilizing data from the Brazilian Institute of Geography and Statistics (IBGE), sheds light on the pronounced growth in the number of workers embracing delivery apps. According to PCdoB65[1], in the last 5 years (data collected from 2015), shows that there was a growth of 979.8% Brazilians working for apps that make some kind of delivery.

Thus, having this growth scenario mentioned, many food delivery companies saw their demand increase exponentially after the outbreak of COVID-19, with a 24% increase compared to the beginning of the pandemic , as reported by Sofia Kercher[2]. During a period characterized by restrictions on physical contact, the role of food delivery drivers has become crucial in ensuring the provision of essential services. With a significant expansion in the number of service providers in this field, companies operating in the food delivery sector have recognized the necessity of enhancing the coverage provided to their drivers. This strategic decision is driven by the inevitability of accidents occurring during the operational process. By proactively addressing these risks, companies aim to prioritize the safety and well-being of their drivers while ensuring the continuity of their delivery services. Furthermore, determining the adequate insurance pricing for food delivery drivers is of paramount importance, as it enables companies to strike a balance between providing adequate coverage and managing costs effectively.

Machine learning methods have played a crucial role in the advancement of insurance pricing. Leveraging their ability to process vast amounts of data and identify intricate patterns, these techniques have brought about a revolutionary transformation within the industry. They facilitate a more precise analysis of risk factors, leading to fairer pricing strategies. Moreover, they contribute to fraud detection, thereby bolstering security for insurers and policyholders alike. Denuit et al.[3] developed the autocalibration procedure to correct the bias in insurance pricing models that are trained by minimizing deviance. The approach is used to predict claims in insurance using neural networks and generalized linear models. Campos and Antonio[4] developed a data-driven insurance pricing model for hierarchically structured risk factors. They compare the predictive performance of three models and find that the Tweedie distribution is well suited to model and predict the loss cost on a contract. The

profound impact of these methods is undeniable, as they have significantly improved the accuracy of calculations and the overall management of risks (see for instance,[4–6]).

Auto insurance for motorcycles is considered by insurers to be high risk, and is seen by many as an excluded risk. According to Filho[7], excluded risk is: "a risk is said to be excluded from coverage and for which the insured does not receive any consideration, nor does the insured have, in the event of loss resulting therefrom, any expectation of secondary indemnity". Therefore, due to the high risk, the price charged and the lack of coverage make the cost of this operation expensive, both for insurers and for companies that contract this service.

This paper presents a case study focused on a food delivery company that has undergone significant expansion in recent years, leading to a rise in the number of delivery drivers within its service provider base. To ensure the safety and security of its drivers, the company engaged an insurance provider to extend coverage across all routes where they operate. However, due to the surge in demand, the associated insurance costs escalated, necessitating a comprehensive study aimed at reducing the costs relative to the premiums paid to the insurance company. As premium we refer to the amount charged by the insurance company to cover the assumed risk, along with other associated expenses,as stated by Canôas[8]. To address this challenge, a combination of machine learning techniques and statistical modeling was employed to construct a risk score. The selected techniques were chosen based on their technical robustness and interpretability, crucial aspects given the requirement for the obtained results to be easily comprehensible to diverse audiences.

The primary objective of this study is to explore strategies for reducing insurance costs while maintaining adequate coverage for the food delivery drivers. To achieve this, a series of methodological approaches were employed. Initially, the Nearest Neighbor Rule Undersampling technique was applied to balance the route database, ensuring a more equitable representation of different route types in the analysis, a methodology corroborated by Smith et al.[9]. Subsequently, a decision tree was utilized to create the claims profile. The decision to use this technique was driven by its ease of interpretation, making the generated model more accessible to the business area. The decision tree facilitated the identification and clear visualization of the primary determinants of claims. Finally, logistic regression was employed to calculate the probabilities of claim occurrence based on the identified profiles. The selection of logistic regression was influenced by its ease of implementation and high interpretability. The combination of these techniques resulted in a transparent and understandable risk score model, with its inputs effectively utilized in negotiations. This study provides an important contribute to the field by presenting insights into the development of a risk score model that aids in optimizing insurance costs for food delivery companies. By employing advanced analytical techniques, the research offers valuable recommendations and findings for the industry, further enhancing the understanding and management of insurance premiums in the context of food delivery services.

The remainder of this paper is presented as follows: Section 2 presents a literature review, providing some basic concepts of insurance premium composition and Machine Learning and sampling techniques. In Sections 3 and 4, the development of the study will be presented, with the application of the techniques mentioned earlier. In Section 5, the pricing calculation is performed based on the results of the techniques applied in Section 4 and finally, we have the conclusions concerning the study and the results presented in Sections 4 and 5.

Background

Insurers perceive the delivery operation carried out by motorcycles as high-risk since a large portion of accidents that occur in the country involve motorcycles, as evidenced by the article on compensation paid by the "Danos Pessoais Causados por veiculos automotores de via terrestre" (DPVAT) (Personal Damages caused by Land Motor Vehicles) insurance , according to Scaramussa and Sá[10]. For the model of the company under analysis, the premium composition considered the claims occurred according to the DPVAT report, where the insurer cross-references this public data with the data and reports sent by the food delivery company, considering the specificity of region, city, state, etc.

In general, the construction of the premium charged by the insurer can be done in several ways, according to each insurer's needs. Often, the calculation information is not disclosed by the insurer. However, for exemplification purposes, we will use a generalized model. One possible way to calculate the premium,as referenced by Müller[11], is through an actuarial approach that covers the risks sufficient to pay indemnities and cover other operational costs. For this purpose, the annual risk model will be used as a reference, which prices insurance based on the sum of claims in a given year (DPVAT) and a safety margin for potential claims that may exceed the average. Since the pricing of the premium is not the study's purpose, a general way of calculating the pure premium is presented:

$$P = \lambda E[X] + Z_{1-\alpha}\sqrt{\lambda E[x^2]}$$

where P is the total pure premium, λ is the number of claims, $E[X]$ is the expectation of X, $Z_{1-\alpha}$ is the value obtained from a standard normal assuming a α significance level, $E[x^2] = \frac{\sum_{i=1}^{N}(X_i \times Xi)}{N}$. To calculate the expectation of X, we used the standard

formula:

$$\mathbf{E}[X] = \frac{\sum_{i=1}^{N} Xi}{N}$$

where Xi is the value of the i-th claim and N is the number of claims.

Finally, the individual premium value was divided by the number of DPVAT policies issued, which does not necessarily correspond to the amount collected.

Undersampling and Oversampling

When dealing with a large data set and using it in classification models, we can often face the challenge of working with unbalanced data, which is a common problem in classification models. According to He and Garcia[12], class imbalance can bias classification models, with standard machine learning techniques always tending towards the majority class.

Imagine that we have a data set that has two classes used for prediction, and that the amount of data related to one class is much larger in relation to the other class, in this scenario we will have an imbalance between the classes, a proper imbalance. This class imbalance can cause bias in classification models, with standard machine learning techniques always tending towards the majority class.

To resolve this problem, we have two solutions: the use of algorithm-level techniques and data-level techniques. According to Chawla et al.[13], algorithm-level techniques aim to modify the algorithm to take into account the distribution of data or to boost the identification of the minority class. However, data-level techniques, , as mentioned by He and Garcia[12] in relation to data-level techniques, aim to modify the database before the learning carried out by the algorithm, based on this approach there are two categories of technique: Oversampling and Undersampling.

The Oversampling technique aims to replicate or generate data from the minority class and the Undersampling technique aims to manually remove data from the majority class. In this study, the Undersampling technique was used to also minimize noise that may exist in the majority class. As Seiffert et al.[14] point out, Undersampling is a technique that aims to manually remove data from the majority class, a simple example being random undersampling.

A simple technique of undersampling is random undersampling. It is a non-heuristic algorithm that aims to balance the database by randomly eliminating data from the majority class. This operation can pose risks as important information for model classification may be discarded during the data elimination. This technique is ideal when we have a large volume of data, as in this case relevant information would still be retained.

Decision Tree

There are several classification models, one of which is the decision tree. It is a non-parametric methodology, according to Izbicki and dos Santos[15]: "A tree is constructed by recursively partitioning the covariate space. Each partition is called a node, and each final outcome is called a leaf. The structure of the tree works as follows: if the first condition is satisfied, it proceeds to the next left branch; otherwise, it continues to the right branch until a leaf is reached. This general structure is followed in both regression and classification decision trees.

In general, the construction of a classification tree is analogous to a regression tree, with some differences in the construction process. In the case of predicting Y for an observation with x covariates, where they fall into a region R_k, the prediction is no longer given by the sample mean but by the mode of the training set:

$$g(x) = \text{mode} y_i : x_i \in R_k.$$

In a regression tree, the evaluation to find the best partition at each step of the data distribution process in the tree's nodes uses the mean squared error. However, in a classification tree, the Gini index is used, as shown in the following formula:

$$\sum_{R} \sum_{c \in C} \hat{P}R, c(1 - \hat{P}R, c). \tag{1}$$

In the formula, R represents one of the partitions induced by the tree and $\hat{P}R, c$ is the proportion of observations in category c that belong to the region or partition of the data R, as Brownlee pointed out[16]. The Gini index is minimum when the proportions $(\hat{P}R, c)$ are zero or one, in this case we can say it is a "pure" tree, which means when a leaf has only observations of a single class.This concept is well explained by James et al.[17]. The tree, as it were, will make divisions of the data, in the process of building its structure, forming patterns or clusters of these data in a certain way. This principle is well described by Hastie et al.[18].

Binary Logistic Regression

The purpose of the binary logistic regression model is to infer the probability of the occurrence of an event defined by Y, which is presented as a dichotomous variable. It takes the value 1 when the event occurs and 0 when the event does not occur, based on the behavior of the explanatory variables. This vector with the explanatory variables and their parameters is defined by the following formula:

$$Z_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} \tag{2}$$

where Z is the logit, β is the parameters of the explanatory variables, X_j are the explanatory variables and i represent each observation, α is the constant term.

We need to define the probability p_i of the event occurring for each observation based on the logit function Z_i. The details of the probability formula construction will not be presented as they are standard in statistical books proceeding with its final form:

$$p_i = \frac{1}{1 + e^{-(\alpha+\beta_1 X_{1i}+\beta_2 X_{2i}+...+\beta_k X_{ki})}}. \tag{3}$$

In general, logistic regression does not estimate the values of the dependent variable, but rather the probability of the occurrence of the event for each observation. It is worth noting that for the estimation of parameters in the logit, **maximum likelihood estimation** is used. When discussing linear regression, the method of minimizing the sum of squares of residuals is used. For categorical variables, it is inappropriate to use linear regression as the response values are not measured on a ratio scale and error terms are not normally distributed. Additionally, the likelihood function can be used to achieve the maximum likelihood estimators. The log-likelihood function is formulated as follows:

$$\log L = \sum_{i=1}^{n} \left\{ \left[Y_i \cdot \ln\left(\frac{e^z}{1+e^z} \right) \right] + \left[(1 - Y_i) \cdot \ln\left(\frac{1}{1+e^z} \right) \right] \right\} = \text{maximize} \tag{4}$$

Nearest Neighbor Rule Undersampling

The Nearest Neighbor Rule Undersampling (NNRU) technique is a widely used approach to deal with data imbalance in classification problems. Class imbalance occurs when the distribution of classes is significantly unequal, which can lead to a bias in the performance of the machine learning model, favoring the majority class at the expense of the minority class. In this context, NNRU aims to balance classes by reducing the number of samples from the majority class without discarding relevant information. NNRU is based on the idea that samples from the majority class closest to the decision boundary between classes are most critical for the classification task. These samples are selected based on the nearest neighbors belonging to the minority class. This approach allows the model to better distinguish class characteristics, improving its generalization ability for the minority class.

Various studies have demonstrated the effectiveness of NNRU in different application domains. In a study conducted by Zhang, Ding, Wang, and Li[19] on the classification of sleep stages in imbalanced data, NNRU was applied to improve the model's performance. The results showed that the NNRU technique significantly increased accuracy and sensitivity for the minority class, compared to other undersampling techniques.

Another study conducted byAnother study conducted by Li, Li, Chen, He, and Shang[20] investigated the application of NNRU in detecting financial fraud. The authors used the technique to balance the imbalanced data and train a fraud detection model. The results showed that NNRU significantly improved the accuracy and fraud detection rate compared to other undersampling methods. It is important to note that NNRU has some limitations. In datasets with significant overlap between classes or complex decision boundary regions, NNRU may not be as effective. In these cases, other undersampling techniques, such as SMOTE (Synthetic Minority Over-sampling Technique), may be more appropriate.

In summary, the Nearest Neighbor Rule Undersampling technique is a promising approach for dealing with data imbalance in classification problems. Its theoretical foundation in selecting samples close to the decision boundary and positive findings in empirical studies reinforce its effectiveness. However, it's important to consider the specific characteristics of the dataset and evaluate other undersampling techniques, as needed, to achieve the best results in each scenario.

In terms of a practical example in Python from the library that implements the NNRU technique, one option would be to use the imbalanced-learn package, which provides implementations of various class balancing techniques. Here we use the make_classification function to generate a sample dataset with a majority class (95%) and a minority class (5%). We then apply NNRU using the NeighbourhoodCleaningRule class from the imbalanced-learn package, which performs undersampling based on the nearest neighbor rule.

Methodology

Route Analysis

Routes are understood as the path taken by the delivery person. Upon conducting initial analysis, it is evident that motorcycle usage accounts for approximately 83% of the routes, followed by bicycle deliveries at 13%. Another relevant finding observed during the analysis was the data imbalance when separating it into insurance claims and non-claims. The total number of claim records was 396, while the non-claim records reached millions. Additionally, it is worth noting that the majority of routes occur in the afternoon and evening.

Analysis of Delivery Person Information

The age range falls between 18 and 30 years. However, it is important to highlight that the current database presents inconsistencies, as 38% of the information regarding the age of the delivery persons is missing. There are entries in the database with delivery persons above the age of 80, raising doubts about the reliability of this field in the table. Currently, the process of storing delivery person information is done through Optical Character Recognition (OCR) technology, which extracts information from the driver's license through a photo. However, the quality of the image can impact the reading process, resulting in records with errors due to image reading failures. Due to the limited information available about the delivery person in the database, which is restricted to the information present only in the driver's license, and in compliance with the LGPD (General Data Protection Law), which ensures the confidentiality of personal information, it was decided to focus solely on the statistical information related to the routes

Hypothesis Testing

The statistical procedure known as hypothesis testing involves analyzing an assumption made about a population parameter. Its objective is to evaluate the probability of this hypothesis being true, using data from a sample. This sample can come from a larger population or a data-generating process. For testing qualitative variables, the Chi-Square test was used. The Chi-Square Test for Independence examines the relationship between categorical variables, assessing if they are interdependent or associated. This non-parametric method employs a contingency matrix for data examination. Within this matrix, data is segmented based on two distinct categorical variables: one's categories are displayed row-wise, and the other's column-wise. Both variables should encompass two or more categories. Each matrix cell denotes the cumulative number of instances for a designated category combination. The p-value associated with each term probes the hypothesis that the coefficient remains neutral (void of impact). If the p-value is below 0.05, it suggests dismissing the base hypothesis. Consequently, a term having a diminutive p-value may hold importance for the model, given its fluctuations influence the outcome variable. Conversely, an elevated p-value indicates a lack of direct correlation between the variable changes and the outcome alterations.

significance level: 0,05 -> 95%

- **Region_**: 0.0026
- **Modal**: 0.0002
- **Worker_type**: 0.0042
- **shift**: 0.0016
- **segmentation**: 0.0384
- **route_state**: '1.2765e-20
- **route_model**: 0.0270
- **route_type**: 0.1526

Logistic regression was adopted for the quantitative variables. Regression analysis is used to establish an equation that describes the statistical relationship between one or more variables and the response variable. In the adopted model, the binary outcome was considered, which corresponds to the values "claim" (1) or "no claim" (0). The fields used were:

‘

- **time_minute**: 0.0001
- **average_time**: 0.0002
- **promotion**: 0.9241
- **freight_multiplier**: 0.6940
- **distance_origin_to_destination**:0.7331
- **distance_to_origin**: 0.9630
- **fee**: 0.0063
- **total_distance_to_destination**: 0.0059
- **total_distance_to_origin**: 0.0438
- **total_route_duration**:0.2461
- **max_rank**:0.0035
- **max_route_lag**:0.0076

After the analysis of qualitative variables, those that showed statistical significance were identified: **time_minute, fee, average_time, total_distance_origin_to_destination, total_distance_to_origin, max_rank, max_route_lag, region, modal, worker_type, shift, segmentation, route_state e route_model**.

Outlier Analysis

The route database is characterized by a large dispersion, which can negatively affect different stages of the process, such as clustering, sampling, and modeling. To minimize the impact of the high data variability and reduce possible errors, an analysis of outlier values was performed, using the z-score formula

$$z = \frac{x - \mu}{\sigma} \tag{5}$$

To apply the z-score formula, it was necessary to center the data mean at zero and set the standard deviation to 1. In this way, any value above or below this limit is considered an outlier and can be identified. After identifying these outlier values, they were removed from the database to ensure that subsequent analyses would be more accurate and reliable.

Sampling

During the exploratory analysis process, it was possible to identify an imbalance in the data, as the claims database had a reduced number of records compared to the routes database without claims occurrences, which was quite extensive. Therefore, situations may occur where studying the entire population in analysis (i.e., the database of routes without claims) becomes unfeasible or undesirable, making it necessary to extract a representative subset of the population, known as a sample. Sampling is crucial to ensure the representativeness of the results obtained, which, through appropriate statistical procedures, can be used to infer, generalize or draw conclusions about the population in question. Due to limited computational resources and the evident imbalance in the data, which implies variability in the observed events, an **Undersampling** technique was used.

Undersampling is a method used to deal with class imbalance in which the majority class is reduced to lessen the disparity between categories. Due to limitations in Spark's Mlib library, a tool commonly used in the company, it was necessary to use a concept similar to the **Nearest Neighbor Rule Undersampling** algorithm. The Nearest Neighbor Rule Undersampling (NNRU) algorithm is an undersampling technique that aims to balance imbalanced classes in a dataset. The algorithm is based on the concept that similar examples tend to belong to the same class. The NNRU process involves identifying examples from the majority class that have one or more close neighbors in the minority class, i.e., examples from the minority class that are close to examples from the majority class. Then, the examples from the majority class are removed until the number of examples in both classes is balanced.

Based on the NNRU algorithm's premise, the unsupervised Kmeans method was used. This algorithm sets random centroids and calculates the distance of data points in relation to these centroids. Points are grouped into clusters based on the shortest distances to the centroids. As the groups formed by clustering have similarities, reducing information in these groups results in minimal information loss, as a group summarizes the necessary information it can add to the study.

After clustering, another sampling technique called "stratified sampling" was used. In this type of sampling, the heterogeneous population is divided into homogeneous subpopulations or strata (like the clusters formed in the clustering using Kmeans). In each stratum, a sample is drawn. The number of strata is initially defined, and the size of each one is obtained. Next, the number of elements to be drawn from the subpopulation of each stratum is specified, which can be uniform or proportional allocation.

Clustering and data balancing

In the available data set, there are about 27 million routes without claims and only 396 routes with claims, indicating a problematic class imbalance. This situation is common in classification problems, in which the classes are not equally represented. Although most classification data sets have a small difference in the number of instances in each class, in some cases, imbalance is expected.

To solve this imbalance problem, a clustering technique was employed. Cluster analysis, or clustering, aims to group objects so that the objects in the same group are more similar to each other than in other groups (clusters). Here, the k-means clustering technique was used to handle the unbalanced data problem identified in the database. K-means is a widely used clustering algorithm that aims to group a set of objects into k clusters, so that objects within each cluster are similar to each other and different from objects in other clusters.

Determining the ideal number of clusters is a critical step in applying k-means. For this, the elbow method was used, which consists of plotting a graph (Fig. 1) of the sum of the squared distances between each point and its closest centroid against the number of clusters. The number of clusters is selected at the inflection point of the curve, where the addition of more clusters does not result in a significant reduction in the sum of squared distances.

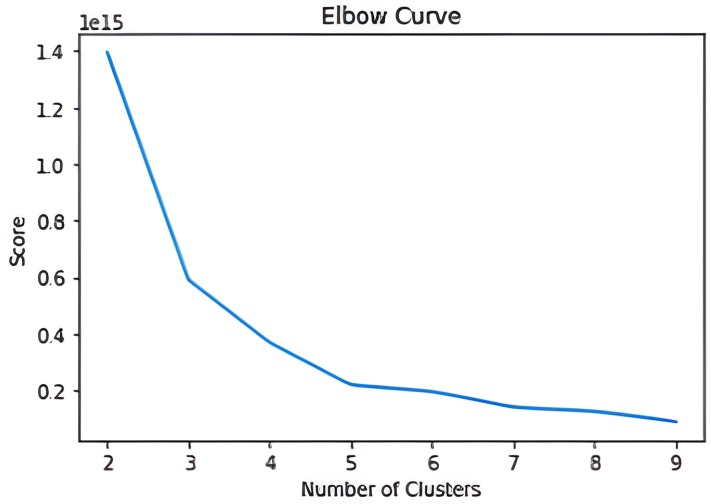

Figure 1. Graph Illustrating the Optimal Number of Clusters Using the Elbow Method

The elbow method is a common technique for determining the optimal number of clusters in a data set during the application of the K-means clustering algorithm. Specifically, this method involves running the K-means algorithm for different values of k, while measuring the sum of squared errors for each value of k. It is important to note that to ensure the effectiveness of the algorithm, only quantitative variables were considered in a data set containing more than 27 million routes without claims.

Based on these results, 5 clusters were defined for the analysis (Table 1). This approach provided a clearer and more concise view of the patterns present in the data, allowing for more precise and informed decisions to be made.

Table 1. Result of the clustering

Cluster	Count
2	8118906
0	8609076
1	7647725
4	1132151
3	4247284

The result of K-means was used to generate a stratified sample of the database. Stratification is a statistical technique used to divide a heterogeneous population into subpopulations or homogeneous strata. In this case, stratification was performed using the K-means method. From the stratification, a sample is drawn from each stratum. The formula used to obtain the stratified sample is:

$$n_i = \frac{N_i}{N} n \tag{6}$$

where n_i is the sub-stratum, N_i the stratum and N the total population.

For the selection of samples in the substratum, a simple random sampling method was used with the aim of obtaining a quantity of records close to the number of records contained in the claims database provided by the insurer. When machine learning and AI algorithms are used in classification methods, it is common to seek a balanced training base, with the same proportion of samples for each class. In databases with dichotomous classes, this proportion would be 50% for each class. To achieve this balanced training base, it is necessary to adjust the sample size of the majority class (routes without incidents) to the size of the minority class (routes with incidents). From the application of the previously mentioned formula, it was possible to reduce the data base and arrive at the ideal sample size for the use of Decision Tree and Logistic Regression algorithms.

Modeling

Before starting the model with the decision tree, a Multicollinearity test was carried out, which consists of a common problem in regressions, in which the independent variables present exact or approximately exact linear relations. For this, a correlation

matrix of the quantitative variables was constructed. A significant correlation was observed between the variable "total route duration" and the variables "total distance" and "average_time", and also between "time_minutes" and "average_time". In the first relationship, it was decided to consider the variable "total distance", while in the second relationship, it was decided to consider the variable "average_time".

Model development

The Decision Tree Algorithm was employed to generate claim profiles and validate the consistency of the data. Furthermore, the algorithm offers a method of validation for the variables used, known as "feature importance."

Based on the sample produced by the previously mentioned balancing procedure, the columns (features, variables) to be used in the decision tree model were selected. The construction of the tree nodes is determined by several methods:

- Entropy: Through entropy, the algorithm analyzes the distribution of the data in the predictor variables concerning the variation of the target variable. The higher the entropy, the greater the disorder of the data; the lower it is, the greater the order of this data when analyzed concerning the target variable.
- GINI Index: Like entropy, the GINI index calculation also checks the distribution of the data in the predictor variables concerning the variation of the target variable, but it uses a different method.
- Regression: In regression problems, the goal is to predict a value, not a class. For this, the tree uses the concepts of mean and standard deviation, allowing for the obtaining of a numerical final result.

These concepts are applied both in the construction of the tree and in the function of Feature Importance. This function, as the name suggests, succinctly summarizes which features used in the model are most relevant and effectively summarize the scenario being modeled, without loss of information. For the model in question, the following results of the feature importance were presented:

Table 2. Overview of Feature Importance in the Developed Model

	Feature	Importance
19	MOTORCYCLE	35.049406
8	Night	24.224920
4	average_time	20.413633
6	Dawn	9.054848
5	total_distance	5.430839
1	max_rank	3.626361
0	total_route_duration	2.199993
13	Inactive	0.000000
18	CAR	0.000000
17	BICYCLE	0.000000
16	without_segmentation	0.000000
15	Super	0.000000
14	Professional	0.000000
10	Casual	0.000000
12	Passive Churn	0.000000
11	Churn	0.000000
9	Afternoon	0.000000
7	Morning	0.000000
3	freight_multiplier	0.000000
2	max_route_lag	0.000000
20	Motorcycle with trailer	0.000000

It is possible to observe in Table 2 that some features, according to the model, did not demonstrate sufficient information gain. With adjusted parameters, such as the decision tree's depth level, the features "Motorcycle", "Night", "Average", "Dawn", "total_distance", "max_rank", and "total_route_duration" summarized the claims prediction scenario more adherently.

After training the model, a confusion matrix was generated (Fig. 2). It is worth mentioning that a confusion matrix is a tool used in classification models that shows the classification frequencies for each class of the model. Considering the mentioned example, it will provide us with the following frequencies:

- True Positive (TP): occurs when, in the real set, the class we are looking for was predicted correctly.
- False Positive (FP): occurs when, in the real set, the class we are seeking to predict was predicted incorrectly.
- True Negative (TN): occurs when, in the real set, the class we are not looking to predict was predicted correctly.
- False Negative (FN): occurs when, in the real set, the class we are not seeking to predict was predicted incorrectly.

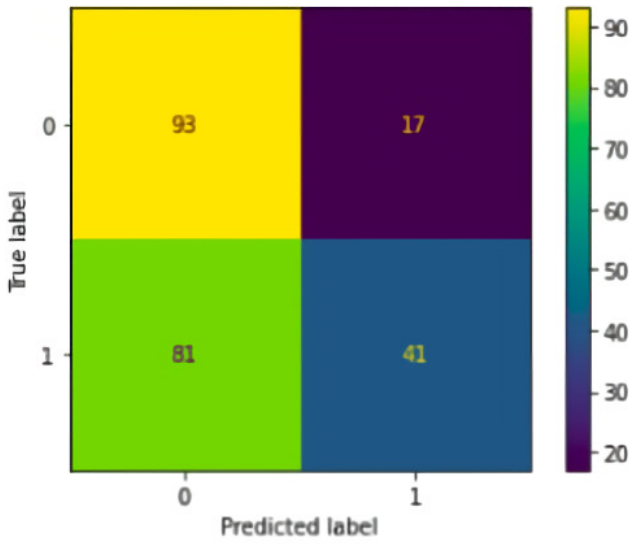

Figure 2. Confusion Matrix of the Classification Model's Performance

The following metrics were evaluated to validate the adherence of the model:

- Accuracy generally tells us how precise the model is.
- Precision can be interpreted as: of the cases classified as correct, how many were indeed correct?
- Recall indicates how often the classifier finds examples of a class. If an example belongs to that class, recall tells us how often it is correctly classified.
- The F1 score combines recall with precision to provide a single representative number.

The metrics of this Decision Tree related to the confusion matrix are as follows:

Table 3. Model Evaluation Metrics for the Decision Tree

	precision	recall	f1-score	support
0	0.53	0.85	0.65	110
1	0.71	0.34	0.46	122
accuracy			0.58	232
macro avg	0.62	0.59	0.56	232
weighted avg	0.63	0.58	0.55	232

From the adjusted results the following decision tree was generated:

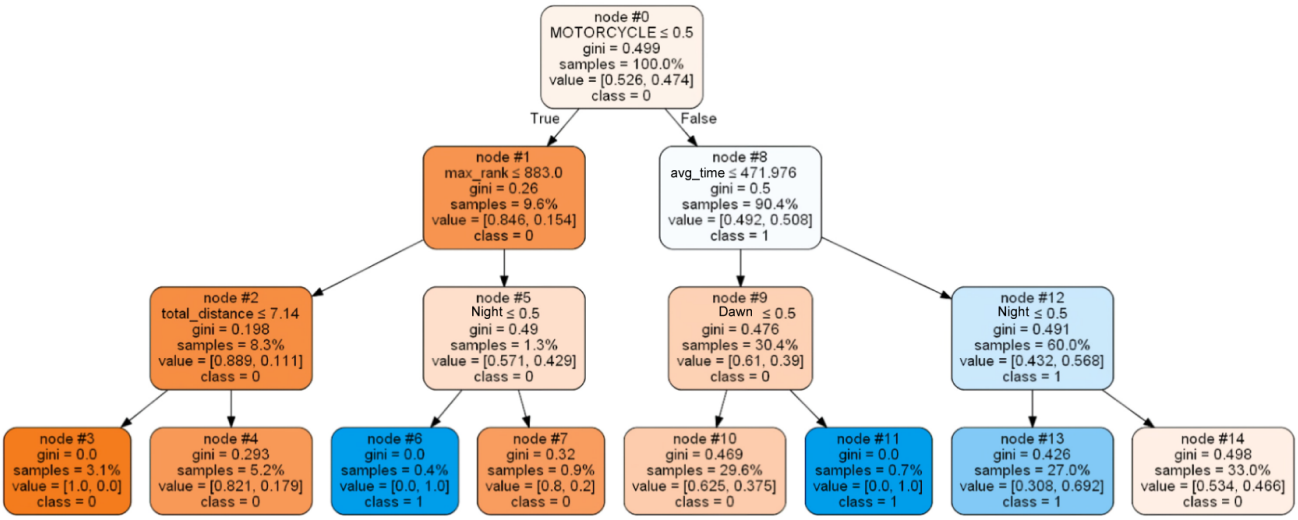

Figure 3. Insurance Decision Tree for Delivery Drivers

The decision tree in question (Fig. 3) was generated amid various other scenarios analyzed. Its validation was carried out together with the business area, and once it presented a logic that reflects the reality of the business, this tree was chosen, despite some model validation metrics that could be improved as presented in Table 3. The careful evaluation of the tree took into consideration its ability to capture the features and behavior of data relevant for decision-making in the context of the business area, specifically financial and operational. This validation with the business area is essential to ensure that the model can provide accurate and useful insights for better steering of strategies and actions in relation to negotiations with the partner insurance company.

Generalizing the Model

Due to the restriction of the size of the claims database, the model was generalized and applied to the database of routes without claims. In this context, the following results were obtained:

Table 4. Result of Metrics for the Generalization of the Model

	precision	recall	f1-score	support
0	1.00	0.91	0.95	480
1	0.00	0.00	0.00	0
accuracy			0.91	480
macro avg	0.50	0.46	0.48	480
weighted avg	1.00	0.91	0.95	480

There was a reduction in the F1-score when generalizing the model, indicating that the evaluation metrics may not be optimal. However, it is important to note that, despite this reduction, the business area validated the model and considered that the presented scenario is consistent with their observations and operational experience. This validation was based on the

perception that the decision tree used in the generalization significantly captures the patterns and relevant characteristics of the operational context.

It is worth noting that, although the F1-score was impacted, the accuracy remains at a level of 50%. This measure of accuracy, although modest, indicates that the model still has a reasonable capacity to make correct predictions (table 4), considering the available data set.

Following, the Confusion Matrix is presented in Figure 4, which is a tool that allows you to visualize the classification frequencies for each class of the model, providing additional information about the performance of the model in the classification task.

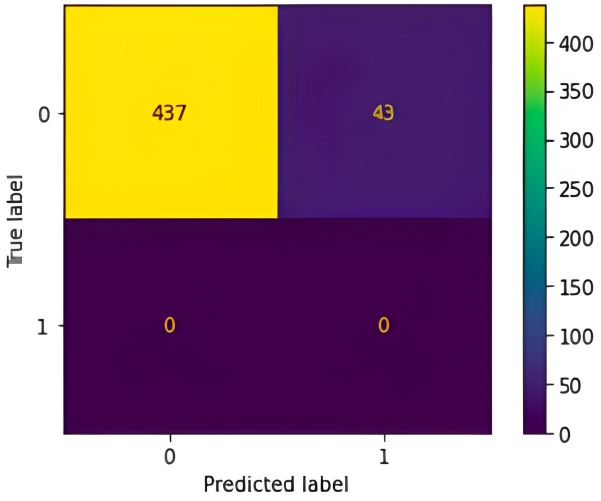

Figure 4. Confusion Matrix of the Generalized Model

Logistic Regression applied to clusters

After generating the decision tree, the nodes of the tree were used as possible groups to define profiles of deliverers and routes. These nodes were applied to the database and later used in a logistic regression model.

The purpose of this procedure was to generate occurrence probabilities for each of the identified clusters, using logistic regression. Through these probabilities, it was possible to obtain a risk score for the routes and deliverers.

Logistic regression, in this context, allowed for a more detailed analysis of the characteristics and patterns present in the clusters defined by the decision tree. Based on these analyses, it was possible to assign a risk score to each route and courier, providing a quantitative measure to evaluate the level of risk associated with each one of them.

Thus, the combined use of the decision tree and logistic regression allowed for a more sophisticated approach to defining profiles and assessing risk related to routes and deliverers, providing valuable information for decision making and strategic planning in the logistics context.As a result, it is possible to observe the confusion matrix in Figure 5 using a heat map:

The evaluation metrics of the confusion matrix were:

- **Accuracy**: 0.621761
- **Precision**: 0.635416
- **F1-score**: 0.625641

The Receiver Operating Characteristic (ROC) curve is a graphical (Fig. 6) representation that illustrates the performance of a binary classifier system as its discrimination threshold varies. In the model, the obtained ROC curve was:

In order to gain a deeper understanding of the model's results, we proceeded to generate a confidence interval, considering the grouping by node (leaf) of the tree. To evaluate the model, 20 distinct samples were generated, using different seeds. The model was applied to each of these samples, grouping the results by seed and node, in order to analyze the means of the scores for each leaf and calculate the corresponding confidence interval.

The generation of multiple samples with different seeds is a common practice to evaluate the model's stability and robustness. By conducting this approach, it is possible to observe the variation of the results in different training data sets, which provides a

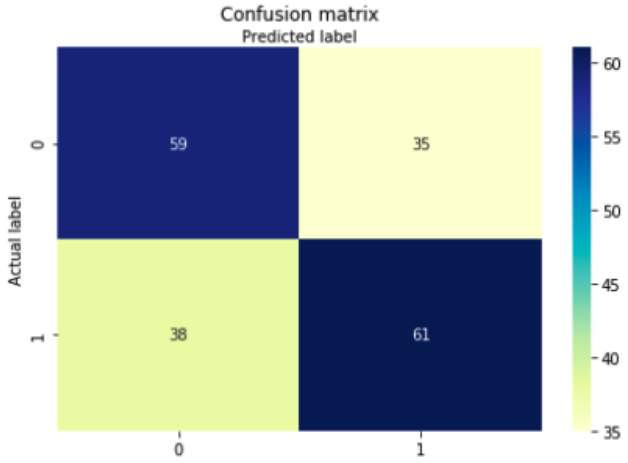

Figure 5. Samples of Nodes (Clusters) - Classification Result

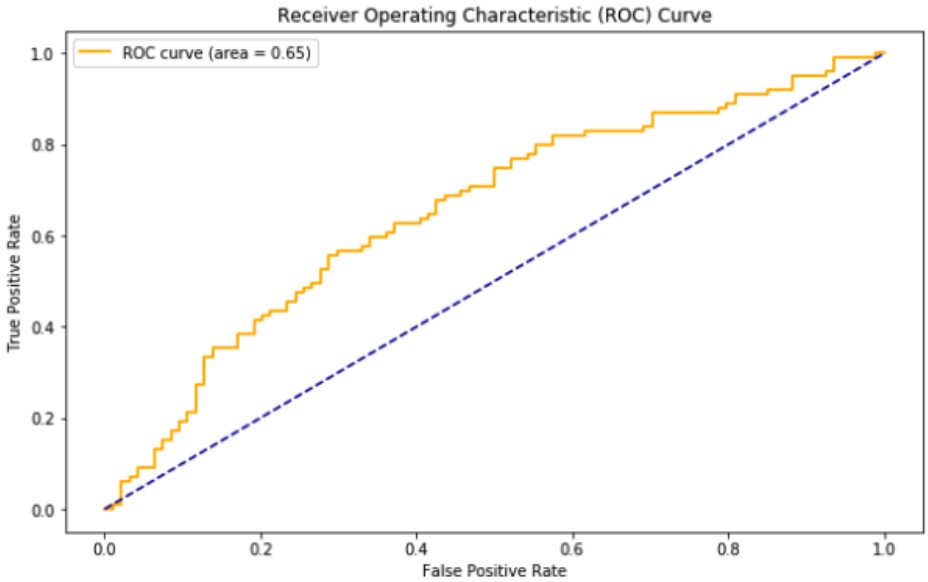

Figure 6. ROC Curve of the Logistic Regression Model

broader view of the model's performance and allows to identify possible variations caused by the randomness of the selection of training data.

By grouping the results by seed and node, it is possible to calculate the means of the scores for each leaf and subsequently estimate the corresponding confidence interval. This confidence interval provides a statistical measure that indicates the accuracy and reliability of the calculated means, taking into account the variability present in the samples (table 5).

Thus, the generation of the 20 samples and the calculation of the confidence interval are fundamental steps for the construction of the score and for the statistical analysis of the model's performance, allowing a more robust and reliable evaluation of its predictive capabilities.

Results

Based on the obtained results, two relevant pieces of information emerge: First, each route was classified into nodes, following the structure of the decision tree. Secondly, for each node, a probability of incident occurrence was assigned, ranging from 0 to 1, where 1 represents a 100% chance of an incident. This measure is referred to as a score.

Here, 10 samples were taken, each containing 400,000 routes, with the aim of applying the pricing model. The first analysis

Table 5. Samples of nodes (Clusters) - Classification Result

seed	nodes	average_score	standard_deviation	min	max
8	node 10	0.49	0.11	0.26	0.71
8	node 10	0.58	0.14	0.29	0.86
8	node 3	0.16	0.08	−0.00102	0.32
27	node 10	0.48	0.12	0.24	0.72
27	node 11	0.64	0.11	0.40	0.86
27	node 3	0.13	0.06	0.0011	0.26
32	node 10	0.48	0.11	0.25	0.70
32	node 11	0.66	0.07	0.51	0.79
32	node 3	0.15	0.08	−0.012	0.31
59	node 10	0.50	0.12	0.25	0.73
59	node 11	0.63	0.14	0.35	0.91
59	node 3	0.15	0.07	0.009	0.29
93	node 10	0.49	0.11	0.26	0.72
93	node 11	0.57	0.10	0.36	0.76
93	node 3	0.15	0.07	0.004	0.28
404	node 10	0.50	0.11	0.27	0.72
404	node 11	0.71	0.07	0.56	0.85

carried out on these samples was the calculation of the standard deviation, considering a matrix of nodes by seeds (table 6). It was concluded that the variations between the nodes and the seeds were minimal, which indicated that the samples could be considered relatively homogeneous.

From this point, one of the 10 selected samples was explored to understand how the nodes could be grouped into more relevant risk classifications for the business context. As a result of this exploration, three risk classifications were identified as can be seen in Table 7: high, medium, and low. These classifications allowed a more tangible approach and more suitable to the company's needs.

Table 6. A summary table of risk definition according to sample analysis

Node	Average Score	Number of routes	Gini Coefficient	Participation of total
node 10	0.44	129734	0.53	32%
node 11	0.61	1829	0.00	0.46%
node 13	0.64	42621	0.57	11%
node 14	0.47	162960	0.50	41%
node 3	0.14	11572	0.00	3%
node 4	0.14	44106	0.71	11%
node 6	0.23	1445	0.00	0%
node 7	0.14	6160	0.68	2%
TOTAL	**0.35**	**400427**	**0.37**	**100%**

Table 7. Definition of Three Level Risk Classification

Risk level	% of base	Probability of a claim
Low Probability	16%	20%
medium Probability	73%	45%
high Probability	11%	63%

An important caveat to be highlighted is that, in relation to the business, it is not believed that the proportion of incidents can reach 11% (table 8). However, as indicated by the model, the runs classified as representing 11% of the total have a higher potential for incident occurrence. On the other hand, the runs classified as 3, 4, 6, and 7 have an extremely low risk probability.

Although the value of 11% served as a reference for runs with a high risk probability, it cannot yet be considered as the final segmentation of our base to be presented to the insurer. It is necessary to assess whether the other samples (table 8) used in the experiment showed a similar behavior regarding critical cases (**high risk**).

Table 8. Risk comparison with the other samples

Samples	High risk	Low Risk	Medium Risk
164	11.00%	15.80%	73.20%
223	18.14%	16.71%	65.15%
251	14.89%	16.47%	68.64%
292	15.50%	16.69%	67.81%
410	17.50%	17.11%	65.39%
522	15.73%	16.66%	67.62%
588	15.99%	16.73%	67.29%
590	17.70%	16.74%	65.55%
791	18.66%	16.88%	64.46%
887	17.03%	16.58%	66.39%

Given the results obtained, it was decided to work with the average value of the routes classified as high and low risk. For medium-risk routes, the difference to 100% was calculated to achieve compatibility between values.

This segmentation plays a fundamental role in the study, as it highlights the need to divide the paid values according to the criticality and degree of risk of each route. This will allow a significant reduction in costs, compared to the current scenario, in which the maximum risk - and consequently the highest price - is assumed for all routes.

Route Pricing Definition

Once the risk classes and their respective participation in the samples were defined, determining a fair price for each class became necessary to facilitate negotiations with the insurer. The prices, weighted by the number of routes, would then be used to establish a single price.

To gather data for pricing, incident records from January 2020 to February 2021 were examined, yielding the following key findings:

- The average payment per route in recent months was USD 0.08.
- The average cost of claims was calculated as USD 0.05. This value was obtained by dividing the sum of compensated claims and the total open claims in a given month by the total number of routes carried out during the same period. To determine the historical average cost of claims, these monthly values were weighted accordingly.

It is worth noting that this value incorporates both compensated claims and open claims. Consequently, the actual cost is likely lower since only a portion of open claims will be converted into compensation. This suggests that the product may be overpriced, resulting in a USD 0.04 spread (difference between the average cost and the average value of the claim).

Based on these findings, the proposed pricing structure is as follows:

- High risk: For routes with a high risk, the payment should remain at the current rate of USD 0.08, which is the pricing currently applied to all routes.
- Low risk: For routes with a low probability of accidents, it is suggested to pay the current cost of the routes plus the open claims, totaling USD 0.05.
- Medium risk: Routes with a medium probability of risk should be priced at an intermediate value of USD 0.06.

Finally, the final single price was calculated by applying the aforementioned pricing structure to the sample segmentation, resulting in a final value of USD 0.06 per route (table 9).An essential point about pricing that was mentioned is this: a transformation was applied to the results to safeguard the confidentiality of the values and negotiations. This step was taken to ensure the protection of sensitive data during the evaluation process, preventing specific information from being compromised or misused. This demonstrates a commitment to privacy and the integrity of the negotiation process.

Table 9. Pricing Structure for Different Risk Levels

Profile Risk	%	Price (USD)
Low Risk	17%	0.05
Medium risk	67%	0.06
High risk	16%	0.08
Weighted Average		0.06

Conclusions

The study presented herein explored the challenges a global delivery company faced during a period of exponential growth prompted by the Covid-19 pandemic. The data analysis uncovered a significant surge in delivery routes, instigating elevated insurance payments and necessitating the embrace of a new business model.

While existing research provides insights into insurance pricing, our study approaches the topic with a unique and practical perspective by integrating machine learning and statistical methods with the real-world knowledge of industry professionals. This hybrid approach not only meets stringent technical precision and accuracy benchmarks but also guarantees the relevance and applicability of our model.

Throughout the process of data acquisition, analysis, modeling, and consolidation of results, the active participation of the business team played a pivotal role in validating our findings and insights. There remains scope for further enhancements, such as the incorporation of additional variables or fine-tuning of model evaluation metrics, but the model developed aligns neatly with business requirements, as demonstrated by its implementation and score construction executed by the finance team.

Significantly, the project's results were presented to the insurer and effectively facilitated the negotiation of a new pricing contract. The deployment of the proposed model led to substantial cost savings, exceeding millions of dollars relative to previous payments. This outcome underscores the efficacy of the new model and signifies the apex of this study, supplying a robust and efficient solution for the global delivery company.

To conclude, this study expands the academic literature by introducing an approach that enhances model accuracy while taking into account the industry's pragmatic aspects. The fusion of advanced methodologies with empirical knowledge illuminates the value of narrowing the divide between theoretical understanding and practical application. The findings and insights elucidated in this paper serve as a valuable resource for future research and practical endeavors in the realm of insurance pricing and business optimization.

Data Availability

The data that support the findings of this study are available from Latin American delivery food company but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Latin American delivery food company .

References

1. PCdoB65. https://pcdob.org.br/noticias/numero-de-entregadores-por-aplicativo-cresce-979-em-cinco-anos/ (2021).
2. Kercher, S. Gasto com delivery sobe 24de consumo do pós-pandemia (2022). CNN Brasil Business.
3. Denuit, M., Charpentier, A. & Trufin, J. Autocalibration and tweedie-dominance for insurance pricing with machine learning. *Insur. Math. Econ.* **101**, 485–497 (2021).
4. Campo, B. D. & Antonio, K. Insurance pricing with hierarchically structured data an illustration with a workers' compensation insurance portfolio. *Scand. Actuar. J.* 1–32 (2023).
5. Wuthrich, M. V. & Buser, C. Data analytics for non-life insurance pricing. *Swiss Finance Inst. Res. Pap.* (2023).
6. Maillart, A. Toward an explainable machine learning model for claim frequency: a use case in car insurance pricing with telematics data. *Eur. Actuar. J.* 1–39 (2021).
7. Filho, N. N. M. Exoneração de responsabilidade do segurador: Estudo tópico de direito securitário. In Nelson Nery Junior, R. M. d. A. N. (ed.) *Responsabilidade Civil - Direito de Obrigações e Direito Negocial*, chap. 5, 43 (Editora Revista dos Tirbuinais, 2010).
8. CANÔAS, V. d. L. Análise do cálculo da provisão de prêmios não ganhos nas sociedades seguradoras. *Insper* (2007).

9. Smith, J. *et al.* Effective use of nearest neighbor rule undersampling in risk analysis. *J. Data Sci.* **50**, 123–139 (2022).

10. Scaramussa, F. S. & Sá, E. C. Indenizações pagas pelo seguro dpvat: perfil epidemiológico dos acidentes envolvendo motocicletas no brasil, contextualização das internações hospitalares e ônus ao sistema único de saúde (sus), no período de 2015 a 2018. *Saúde Ética & Justiça* **25**, 10–14 (2020).

11. Müller, B. D. Prêmio puro dpvat 2021 após a resolução cnsp 399. *Lume* (2022).

12. He, H. & Garcia, E. A. A comparison of resampling methods for imbalanced classification on medical datasets. *IEEE Transactions on Syst. Man, Cybern. Part B (Cybernetics)* **39**, 1633–1649, DOI: 10.1109/TSMCB.2008.2006160 (2009).

13. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357, DOI: 10.1613/jair.953 (2002).

14. Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J. & Napolitano, A. Random undersampling for imbalanced data classification. *J. Intell. Syst.* **19**, 159–177, DOI: 10.1515/jisys.2010.015 (2010).

15. Izbicki, R. & dos Santos, T. M. *Aprendizado de máquina: uma abordagem estatística* (2020).

16. Brownlee, J. A simple guide to classification and regression trees (cart). https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/ (2018).

17. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning: with Applications in R* (Springer Science & Business Media, 2013).

18. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Science & Business Media, 2009).

19. Zhang, H., Ding, L., Wang, H. & Li, X. Nearest neighbor-based undersampling for imbalanced classification and its application to sleep stage classification. *Biomed. Signal Process. Control.* **39**, 413–421 (2018).

20. Li, R., Li, Y., Chen, H., He, X. & Shang, Z. Application of nearest neighbor rule undersampling in financial fraud detection. *Mathematics* **8**, 482 (2020).


ICMC
USP
SÃO CARLOS