UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DEPARTAMENTO DE ESTATÍSTICA

Trabalho de Conclusão de Curso

Comparação entre Regressão Logística Multinomial e *Extreme Gradient Boosting* para predição de canais de negociação em cobrança

Rafaela Vidal Galetto

Porto Alegre
2022

Rafaela Vidal Galetto

Comparação entre Regressão Logística Multinomial e *Extreme Gradient Boosting* para predição de canais de negociação em cobrança

Trabalho de conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientadora: Profa. Drª. Lisiane Priscila Roldão Selau.

Porto Alegre
2022

Rafaela Vidal Galetto

Comparação entre Regressão Logística Multinomial e *Extreme Gradient Boosting* para predição de canais de negociação em cobrança

Este Trabalho foi julgado adequado para obtenção dos créditos da disciplina Trabalho de Conclusão de Curso em Estatística e aprovado em sua forma final pelo Orientador(a) e pela Banca Examinadora.

Orientadora:

Profa. Drª. Lisiane Priscila Roldão Selau, UFRGS,

Doutora pela Universidade Federal do Rio Grande do Sul, Porto Alegre, RS.

Banca examinadora:

Profa. Drª. Márcia Helena Barbian, UFMG,

Doutora pela Universidade Federal de Minas Gerais, Belo Horizonte, MG.

Jefferson Ferreira Alves da Nobrega,

Bacharel em Estatística pela UFPR – Realize CFI.

Porto Alegre
2022

“But do you know the darker the night, the brighter the stars glow.”
(Katy Perry)

Agradecimentos

À UFRGS pela oportunidade de ensino e desenvolvimento profissional.

Aos professores do Instituto de Matemática e Estatística da UFRGS por todo o conhecimento e aprendizado oferecido nesses anos de graduação.

Aos meus pais por todo apoio de sempre. Não existe uma lembrança na minha memória em que vocês não estão ao meu lado, apoiando com muito amor todos os passos que dou nessa caminhada da vida. Sirlei e Inagé, vocês são a minha inspiração: sem vocês nada disso seria possível.

À minha irmã, Luísa, por vibrar comigo com as minhas conquistas, me dar força nas situações difíceis e sempre roubar as minhas melhores risadas. Como já te disse, no final de tudo sempre será nós duas, juntas.

Ao Guilherme, que acompanha toda essa jornada de perto. Obrigada por ter sido meu conforto nas horas difíceis, por torcer e comemorar comigo, pelas milhares de ajudas em programação e cálculo. Também te agradeço por sempre ser tão compreensivo, amoroso e meu maior parceiro em todos os âmbitos da vida. Obrigada por ter deixado os dias frios no Campus do Vale mais quentinhos e a fila gigantesca do restaurante universitário – RU – mais divertida. Teu apoio nesses 5 anos foi fundamental para minha força.

Aos amigos que conheci na estatística, em especial, aos seguintes: Rafael, Gustavo, Franciele, Juliana e Lincon. Vocês deixaram esse processo mais fácil pela companhia e parceria incríveis; obrigada por levarem essa amizade para além da graduação.

À minha orientadora Lisiane por apoiar a ideia do trabalho e, com muita paciência, me ajudar nessa etapa. Tuas orientações foram essenciais. Muito obrigada pela confiança, por ser um exemplo e por acreditar em mim em todo esse processo.

Para finalizar, ao meu professor de matemática do ensino médio, Dorval (Nico), por ter sido o responsável por me apresentar à estatística e descobrir com o que queria seguir como profissão. Sem a tua aula e teus ensinamentos talvez eu não estivesse aqui.

Resumo

Este trabalho tem como foco o estudo do perfil de clientes inadimplentes no momento de utilizar um canal de negociação para realizar um acordo de sua dívida em uma empresa financeira. Para realizar o estudo, foi proposta a inclusão de um novo modelo no ciclo de crédito: o modelo de propensão à utilização dos canais de negociação de dívidas disponíveis em um setor de cobrança. O banco de dados utilizado foi disponibilizado por uma empresa do mercado financeiro e é composto de oito meses de negociações realizadas por clientes que estavam inadimplentes na empresa estudada. Foram utilizados dois métodos para realizar a análise dos dados: Regressão Logística Multinomial e o *Extreme Gradient Boost*. Para escolher o melhor modelo, foi avaliada a matriz de confusão e calculados o *Recall*, a Precisão e o *F-score* para cada modelo e para cada canal de negociação. Os dois modelos ajustados apresentaram boa adequação aos dados fornecidos pela instituição financeira. Todavia, o *Extreme Gradient Boosting* teve métricas mais estáveis entre as amostras de treinamento e validação. Em média, ele demonstrou resultados melhores nas três métricas avaliadas e apresentou mais capacidade de identificar as observações em relação à classe minoritária. Dessa forma, o modelo ajustado utilizando o *Extreme Gradient Boosting* se mostrou como uma boa alternativa como ferramenta para avaliar a propensão à utilização de canais de negociação e poderá ajudar na elaboração de ações estratégicas na área de cobrança.

Palavras-chave: canais de negociação, cobrança, *Extreme Gradient Boosting*, Regressão Logística Multinomial.

Abstract

This work has focused on studying profiles of defaulter clients when using a negotiation channel to settle their debt in a financial company. To undertake this study, it was proposed to include a new model of the credit cycle, the model of propensity to use the debt negotiation channels available in a collection sector. The database used is available by a company from the financial market, and it was made of eight-month negotiations made with clients that were defaulters in the studied company. Two methods were used to analyze the data: The Multinomial Logistic Regression and the Extreme Gradient Boost. To select the best model, the confusion matrix was evaluated and Recall, Precision and F-score were calculated for each model and for each negotiation channel. These two adjusted models have presented great adequacy to the data given by the financial institution. However, Extreme Gradient Boosting had more stable metrics between the training and validation samples. In averaging numbers, it demonstrated better results in the three metrics evaluated, and it presented more capacity for identifying the observations related to the minority class. This way, the adjusted model using Extreme Gradient Boosting proved to be a good alternative as a tool to assess the propensity to use negotiation channels and could help in elaborating strategic actions in the collection area.

Keywords: negotiation channels in collection, Debt Collecting, Extreme Gradient Boosting, Multinomial Logistic Regression.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 Introdução

O ciclo de crédito é o processo realizado para ajudar a diminuir os riscos, perdas e prejuízos de um negócio. Ele existe em todas as empresas que concedem crédito ou pensam em viabilizar esse benefício, independente do tipo de mercado, desde o varejo até o financeiro. Após definido o produto e o perfil de cliente ideal, as primeiras etapas do ciclo são a concessão e manuntenção de crédito. Nessa fase inicial, utiliza-se de análises considerando diferentes tipos de dados e modelos estatísticos para melhorar a classificação de bons e maus clientes e, dessa forma, diminuir riscos (Tsai et al., 2014). Alguns modelos comuns existentes no ciclo de crédito são os modelos de aprovação de crédito - utilizado nas decisões para novos clientes - e modelos comportamentais, comumente chamados "*Behavioural Scoring*" (Moraes, 2012), utilizados para clientes que já possuem relacionamento com a empresa.

Mesmo que as análises e modelos, na concessão e manutenção de crédito, sejam acuradas, frequentemente haverá clientes inadimplentes ao final do ciclo de crédito. A forma mais tradicional e comum de cobrar esses clientes é voltada, por exemplo, à utilização da intervenção humana por meio de telefonemas ou, até mesmo, pelo envio de cartas via correio. De maneira geral, as empresas que necessitam desses serviços utilizam como recurso as assessorias de cobrança, ou seja, empresas especialistas em cobrar dívidas complexas (quando o atraso já é superior a 1 ano, por exemplo). Além disso, se faz a utilização de centrais internas das próprias empresas com operadores ligando ou recebendo ligações de clientes para realização de negociações, notificações físicas, telecobranças, etc.

Nos setores de cobrança, assim como nos de crédito, existem análises e modelos de perfil de cliente que buscam estimar a probabilidade de receber pagamentos dos clientes, como os modelos de "*Collection Score*" (Machado, 2015). Contudo, para executar o processo de cobrar efetivamente o cliente, em geral, se é baseado somente em uma variável: os dias em atraso.

Em paralelo, em decorrência da pandemia de COVID-19, iniciada em 2020, diversas empresas tiveram a necessidade de fechar seus estabelecimentos ou modificar, de forma abrupta, a carga horária dos funcionários. Os setores de cobrança foram diretamente impactados, pois as centrais de cobrança não possuiam estrutura para *home office*. Em muitos casos, a forma predominante de negociação

de dívidas era que o cliente fosse até o estabelecimento físico da empresa. Ademais, o setor da economia já passava por um processo de transformação digital (Machkour et al., 2020), e isso não era uma exceção entre as empresas de recuperação de crédito. Com as circunstâncias pandêmicas, essas mudanças foram mais impulsionadas e necessárias pela busca por serviços 100% digitais (Amankwah-Amoah, 2021), eficazes, de baixo custo e com rápida implantação.

Além disso, como já apontado por autores em estudos (Nunes, 2018), as pessoas nascidas após o ano de 2000 são jovens adultos chamados “nativos digitais”. Eles cresceram inseridos em uma era de transformações tecnológicas. Isso fortalece a ideia de que muitos processos do ciclo tradicional de crédito deveriam ser repensados. A cobrança clássica, em que é necessário falar por telefone com um operador, por exemplo, pode ser um mecanismo ultrapassado para esse público.

Em concomitância, mesmo com todas as mudanças ocasionadas pelas transformações digitais e a implantação de diversos novos canais de negociação, sabe-se que a cobrança tradicional, utilizando intervenção de uma pessoa, não será descontinuada por ainda existirem perfis com a necessidade de utilização desse recurso. Também, sabe-se que a cobrança realizada utilizando intervenção de uma pessoa é uma forma que necessita de mais investimentos para gerir. Logo, infere-se que a cobrança mais eficiente, ou seja, que gera menos custo e mais retorno financeiro, é aquela que identifica o canal de cobrança mais adequado a cada perfil de consumidor.

A partir disso, para melhorar a estratégia de cobrança, o objetivo desse trabalho foi desenvolver um modelo estatístico para estimar a probabilidade de que o cliente realize a negociação da dívida em cada um dos canais digitais e tradicionais disponíveis na gestão de cobrança.

Nas áreas de crédito e cobrança, não há um consenso sobre o método ideal para o desenvolvimento de modelos. Em geral, a Regressão Logística é uma das técnicas amplamente utilizadas tanto para modelos de concessão quanto para modelos comportamentais ou de propensão de pagamento de dívida desde a década de 1970 (Altman et al., 1997). Nesse trabalho, será utilizada a Regressão Logística Multinomial (RLM) como técnica base, pois a variável dependente é o canal de negociação e possui mais de duas categorias.

Além da Regressão Logística Multinomial, também existem técnicas de *Machine Learning* indicadas, tais como algoritmos de árvores de decisão, *k-Nearest*

Neighbors, Naive Bayes, Random Forest, Gradient Boosting, que não exigem o conhecimento das relações entre as variáveis explicativas e de resposta (Aniceto, 2016).

O método *Extreme Gradient Boosting* (*XGBoost*) está sendo muito utilizado em diversos desafios de *Machine Learning* e análise de dados. Seu sucesso se deve muito pela escalabilidade em diversos cenários, sendo até 10 vezes mais rápido do que outras soluções, possuir grande flexibilidade e obter bons resultados em diversos problemas e conjuntos de dados (Chen et al., 2016). Logo, nesse trabalho foi comparado o algoritmo *XGBoost* com a técnica base.

O conjunto de dados utilizado nesse trabalho é referente a clientes inadimplentes em uma empresa do ramo financeiro que concede crédito. A criação das variáveis e seleção da amostra foi realizada a partir do cruzamento de seis fontes de dados diferentes. Esses dados são de assuntos relacionados ao histórico de relacionamento do cliente com a empresa. Ao final da criação e coleta dos dados, foram geradas 270 variáveis explicativas e a amostra total foi composta por 1.735.834 clientes que realizaram uma negociação de dívida entre os meses de abril e dezembro de 2021.

Os subsecivos desse trabalho estão organizados da seguinte forma: na seção 2, apresenta-se o referencial teórico, relatando as abordagens estudadas. Em seguida, na seção 3, apresenta-se a metodologia com as etapas de desenvolvimento desse trabalho e os resultados na seção 4. Por fim, na seção 5, apresenta-se as considerações finais do estudo realizado.

2 Referencial Teórico

Nessa seção, está organizado o referêncial teórico. Inicialmente, foi abordado sobre o modelo de Regressão Logística Multinomial, sequencialmente o conceito de *Machine Learning*, o algoritmo *Gradient Boosting* e sua extensão - *Extreme Gradient Boosting*, o conceito de *Feature Selection* e o método de *Permutation Importance*. Por fim, os últimos tópicos abordados são sobre o conceito de *K-Fold Cross – Validation* e as métricas de avaliação de resultados dos modelos.

2.1 Regressão Logística Multinomial

Comumente o modelo de Regressão Logística é utlizado para estudar a relação entre uma variável dicotômica e um conjunto de covariáveis. Segundo Lopes (2004), o modelo logístico com variável resposta politômica é construído através do ajuste simultâneo de k – 1 modelos de Regressão Logística Binária. Portanto, são estimados k – 1 vetores de parâmetros β_i', correspondente a k – 1 categorias da variável resposta. Então, tem-se k – 1 comparações com a categoria de referência escolhida.

Seja $x = (x_0, x_1, ..., x_p)$ o vetor das p covariáveis do modelo, de dimensão $p + 1$, em que $x_0 = 1$, e uma variável aleatória Y de natureza nominal que pode assumir os níveis 0, 1, ... , q. Segundo Figueira (2006), uma abordagem análoga, ao que é feito na Regressão Logística para resposta binária, seria descrever a função *logit,* fazendo a comparação de Y = k com Y = 0 para $k \in \{1, \ldots, q\}$, utilizando o valor zero como categoria de referência. Logo, as funções *logit* são as seguintes:

$$g_k(x) = ln\left[\frac{P(Y = k|x)}{P(Y = 0|x)}\right] = \beta_{k0} + \beta_{k1}x_1 + \ldots + \beta_{kp}x_p \text{ , para } k \in \{1, \ldots, q\}$$

No caso da Regressão Logística Multinomial, os parâmetros são estimados utilizando o Método da Máxima Verossimilhança. Esses cálculos podem ser encontrados em Hosmer e Lemeshow (1989).

Uma expressão geral para as probabilidades condicionais de um modelo com q + 1 categorias é dada pela seguinte definição:

$$P(Y = k \mid x) = \frac{e^{g_k(x)}}{\sum_{k=0}^{q} e^{g_k(x)}} \text{ , para } g_0(x) = 0 \text{ e } k \in \{1, \ldots, q\}$$

2.2 *Machine Learning*

Segundo James et al. (2013), *Machine Learning* pode ser entendido como uma subcategoria da Inteligência Artificial que utiliza recursos capazes de analisar uma conjuntos de dados por meio de métodos estatísticos específicos, além de usar uma variedade de algoritmos para encontrar padrões no banco de dados. Com base nesses padrões, consegue-se fazer classificações ou predições.

Existem duas categorias de problemas em *Machine Learning* que são chamadas de aprendizado supervisionado e não supervisionado (James et al., 2013). No aprendizado supervisionado, se quer ajustar um modelo que relaciona a resposta aos preditores com o objetivo de prever a resposta para futuras observações (a previsão) ou a melhor compreensão da relação entre a resposta e os preditores (a inferência). No aprendizado não supervisionado, para cada observação no conjunto de dados, se observa variáveis, porém não há uma variável resposta para conduzir/ supervisionar a análise. Esse trabalho se encaixa na categoria de aprendizado supervisionado, uma vez que se quer predizer o canal de comunicação mais adequado para um determinado cliente realizar uma negociação de dívida.

2.3 *Gradient Boosting*

O G*radient Boosting* é uma técnica de *Machine Learning* para problemas de regressão e classificação que produz um modelo de previsão na forma de um *ensemble* de modelos de previsão fracos, geralmente árvores de decisão. Ele constrói o modelo em etapas, como outros métodos de *boosting*.

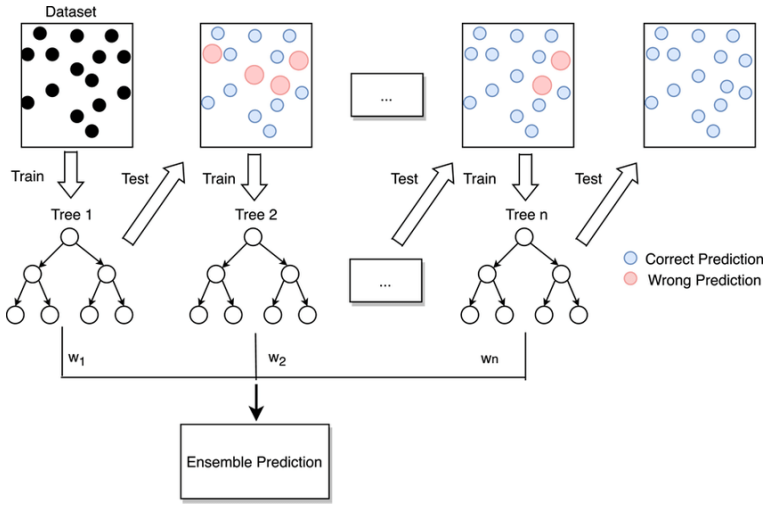

Figura 1: Estrutura do *Gradient Boosting Machine.*

Fonte: Journal of Advances in Modeling Earth Systems.

Dessa forma, também permite a otimização de uma função de perda diferenciável arbitrária (Friedman, 2002). A ideia do *Gradient Boosting* originou-se na observação de que o *boosting* pode ser interpretado como um algoritmo de otimização de uma função de custo (Breiman, 1996). Isso pode ser explicado mais facilmente no contexto da regressão por mínimos quadrados, em que o objetivo é ajustar um modelo F $\hat{y} = F(x)$ que busca minimizar o erro quadrático médio (EQM) $\frac{1}{n}\sum_i(\hat{y}_i - y_i)^2$, em que o índice *i* percorre algum conjunto de treinamento de tamanho *n* dos valores reais da variável de saída *y*, em que é tido o seguinte:

$\hat{y}_i$ = o valor previsto $F(x)$;

y_i = o valor real;

n = o número de amostras em y.

Na sequência, considera-se um algoritmo de *Gradient Boosting* com M etapas. Em cada etapa m (1 ≤ m ≤ M) *Gradient Boosting*, considere algum modelo F_m de menor complexidade (para um m baixo, tal modelo pode simplesmente retornar $\hat{y}_i = \bar{y}$, isto é, a média de y). Para melhorar F_m, o algoritmo deve adicionar algum novo estimador, $h_m(x)$. Observe abaixo:

$$h_m(x) = y - F_m(x)$$

Logo, o *Gradient Boosting* ajustará h ao resíduo $y - F_m(x)$. Como em outras variantes de *boosting*, cada F_{m+1} procura corrigir os erros de seu antecessor F_m.

2.4 *Extreme Gradient Boosting*

O algoritmo *Extreme Gradient Boosting* é um método baseado em árvores de decisão e pode ser utilizado tanto para regressão quanto para classificação (Wang et al., 2018). Ampliando o conceito do *Gradient Boosting*, o *Extreme Gradient Boosting* tem, como principal melhoria, a normalização da função de perda que busca mitigar a variância do modelo: é um método de *ensemble* baseado na construção de árvores de decisão com profundidade reduzida (Chen et al., 2016). Prossegue a ideia de treinamento aditivo, em que as árvores já construídas nas

iterações do modelo não sofrem alteração, e cada nova árvore é construída com base no aprendizado resultante do resíduo da árvore anterior.

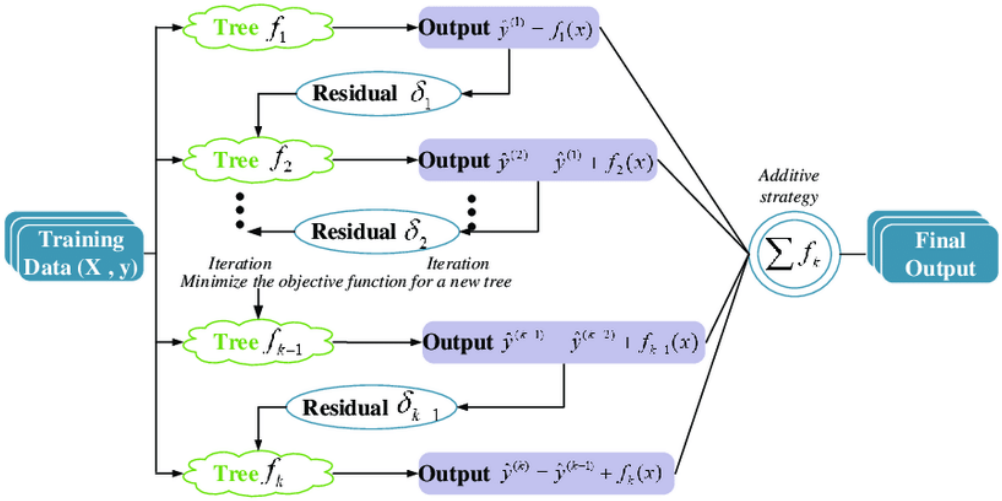

Figura 2: Estrutura do *XGBoost.*

Fonte: Cheng et al., 2020

Dessa maneira, temos que $\hat{y}_i^{(k)} = \hat{y}_i^{(k-1)} + f_k(x_i)$ é a predição realizada na k-ésima iteração e em toda iteração o *XGBoost* otimiza o modelo, diminuindo o erro da predição. Ao final, a predição é dada pela soma ponderada da predição individual de cada árvore, como segue abaixo:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),$$

Assim, f_k pertence ao espaço de funções *F* que contém todas as árvores e *k* representa o número de árvores. Para aprender a função f_k de cada árvore, o *XGBoost* estabelece uma função objetivo com regularização:

$$L(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k),$$

Sendo assim, ϕ são todos os parâmetros de aprendizagem para o *XGBoost*; $l(y_i, \hat{y}_i)$ é a função de perda, representando o erro entre o predito e o observado: quanto menor l for, melhor é a performance do algoritmo; $\Omega(f_k)$ é o termo de regularização que penaliza a complexidade e previne o *overfitting*.

Em outras palavras, no *XGBoost* cada uma das árvores de é construída para minimizar uma função perda pré-definida, mas em cada estimação é colocado mais

peso nos casos preditos erroneamente pelas árvores já desenvolvidas. O modelo final é determinado coletivamente pelos resultados de todas as árvores desenvolvidas.

Também reduz as complexidades de modelagem e, então, a probabilidade de *overfitting* do modelo. O tradicional *Gradient Boosting* lida apenas com a primeira derivada na aprendizagem, porém o *Extreme Gradient Boosting* melhora a função de perda com a expansão de Taylor. Enquanto o nível de complexidade aumenta para o aprendizado das árvores, a normalização evita os problemas associados ao excesso de *overfiting* (Chen et al., 2016).

2.5 Seleção de variáveis

Como o objetivo deste trabalho é ajustar modelos que discriminem tipos de canais de negociação em cobrança, um conjunto expressivo de variáveis foi criado, como será observado na seção 4.

Além de exigir alto poder computacional, utilizar muitas variáveis preditoras pode levar a alguns problemas, tal como *overfitting.* Isto posto, seleção de variáveis consiste em selecionar um subconjunto de variáveis relevantes para a construção do modelo (Li et al., 2017). Existem vários métodos de seleção de variáveis disponíveis; para este trabalho, foi utilizado o método de *Permutation Importance* (PIMP).

O PIMP pode ser aplicado utilizando qualquer método em que seja possível acessar a importância das variáveis. Importância de variáveis se refere a técnicas que calculam uma pontuação para todas as variáveis explicativas utilizadas no desenvolvimento de um modelo. Essa pontuação é chamada de importância da variável. Uma pontuação mais alta significa que a variável explicativa específica terá um efeito maior no modelo que está sendo usado para prever uma determinada variável resposta.

O método de *Permutation Importance* ajusta *s* vezes o modelo para a variável resposta e avalia, em cada vez, a importância das variáveis explicativas. Ao final, tem-se um vetor de medidas de importância de tamanho *s* para cada variável às quais se assume uma distribuição de probabilidade, podendo escolher entre as distribuições Normal, Gama e Log-normal. Com isso, pode-se calcular os estimadores de máxima verossimilhança dos parâmetros da distribuição selecionada e a probabilidade de ser observada uma importância igual a *v* ou superior, usando a variável resposta verdadeira (Altmann et al., 2010).

2.6 Validação Cruzada *K-Fold*

O objetivo dos modelos de *Machine Learning* é estimar a função que melhor ajusta aos dados de entrada para obter previsões corretas de forma generalizada. Quando é discutido sobre a previsão de um modelo, é importante entender os erros de predição: o viés e a variância. O viés é a diferença entre a previsão média do nosso modelo e o valor correto do que estamos tentando prever. A variância é o erro devido à variabilidade de uma previsão do modelo para um determinado ponto de dados.

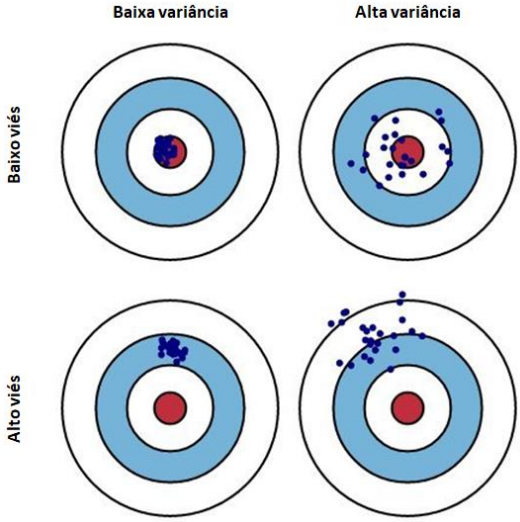

Figura 3: Ilustração gráfica de Viés e Variância.

Fonte: http://scott.fortmann-roe.com/docs/BiasVariance.html

Ao aumentar a variância reduziremos o viés. O contrário também é verdadeiro, reduzindo a variância é aumentado o viés em relação à complexidade do modelo. Na medida em que mais parâmetros são adicionados a um modelo, a complexidade dele aumenta e a variância se torna a principal preocupação, enquanto o viés diminui constantemente, por exemplo. Portanto, se quer encontrar um equilíbrio entre esses dois erros. Para reduzir a variação nas previsões do modelos técnicas de reamostragem, aumentar o conjunto de treinamento e testar diferentes combinações de variáveis explicativas e parametrização do modelo costumam ser ferramentas importantes.

Validação cruzada é uma técnica para avaliar a capacidade de generalização de um modelo a partir de um conjunto de dados. Ela é amplamente empregada em modelos de *Machine Learning*. O método Validação Cruzada *K-Fold* é uma técnica que consiste em dividir, em tamanhos parecidos, e de forma aleatória, o conjunto de

treinamento em *k* novos bancos de dados, os *folds*. Em seguida, um desses *k folds* é separado como conjunto de validação e o modelo é ajustado nos demais *k – 1 folds*. Esse processo é realizado *k* vezes. Em cada vez, um *fold* diferente é separado como conjunto de validação, resultando, por fim, *k* estimativas de erros e acurácias. De maneira geral, os valores de *k* mais populares são 5 e 10, sendo definido com base no poder computacional da máquina trabalhada (James et al., 2013).

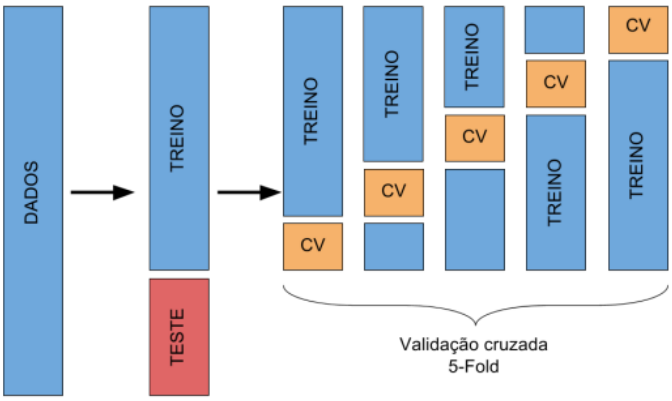

Figura 4: Funcionamento da validação cruzada com 5 – *folds*.

Fonte: Curso de Machine Learning para Ciência de Dados da UFPR.

2.7 Métricas de avaliação dos resultados

Para avaliar o desempenho dos modelos, será utilizada a matriz de confusão. A matriz de confusão é uma tabela cruzada entre as classificações preditas e os dados reais da variável resposta, calculada, em geral, a partir dos dados de teste. Na diagonal principal, encontram-se os acertos do modelo (Naser et al., 2020). Com base na matriz de confusão, é possível calcular estatísticas que resumem a performance para um determinado classificador com base, também, nas predições realizadas no conjunto de teste.

Obs\Pred	0	1	Correta (%)	Geral (%)
0	370	0	100	
1	104	0	0	78,06

Figura 5: Exemplo de uma matriz de confusão de ordem 2.

Fonte: Figueira (2006).

Uma das métricas utilizadas para problemas de classificação múltipla é a de Precisão para cada categoria da variável resposta. Essa métrica expressa a

proporção de observações que o modelo prediz como de certa categoria e, de fato, é daquela categoria.

Outra métrica utilizada para esse tipo de desfecho é o *Recall* para cada categoria da variável resposta - é uma métrica que avalia a proporção de observações positivas que foram classificadas corretamente como positivas. Basicamente, ele mede a habilidade do modelo de encontrar as observações de certa categoria.

Tendo calculadas as métricas de Precisão e Recall, será calculada a métrica *F-score* para cada canal de negociação. A métrica será dada pela fórmula abaixo:

$$F - score = \frac{2 * precisão * recall}{(precisão + recall)}$$

O *F-score* é a média harmônica entre a Precisão e o *Recall.*

Ainda sobre a matriz de confusão, será avaliada a Acurácia do modelo. Ela é dada pela proporção de classificações corretas sobre o total da amostra (Naser et al., 2020).

Em vista disso, foi utilizado o *Population Stability Index* (*PSI*) para comparar a distribuição das variáveis nos meses de desenvolvimento com relação aos meses de teste e validação e, ademais, a distribuição da predição dos canais de negociação. O PSI busca avaliar o quanto a distribuição no tempo *n* varia em relação à distribuição no tempo zero pela aplicação da função abaixo:

$$PSI = \sum_{i=0}^{n} \left((F_{i,t} - F_{i,t+1}) * \ln\left(\frac{F_{i,t}}{F_{i,t+1}}\right) \right),$$

Em que há o seguinte:

$F_{i,t}$ = frequência relativa dos registros da variável na amostra no tempo *t*;

$F_{i,t+1}$ = frequência relativa dos registros da variável na amostra no tempo *t* + *1*;

Assim sendo, conforme Yurdakul (2018), um PSI inferior a 10% não apresenta diferenças relevantes. Um PSI superior a 10% e inferior a 25% apresenta diferença relevante e precisa ser observado com cautela. Por último, um PSI superior a 25% possui muita diferença em relação à amostra inicial.

Essa métrica é utilizada pelas empresas para fazer o monitoramento da estabilidade das métricas de avaliação de resultados e, também, o comportamento das variáveis do modelo ao longo dos meses antes e após iniciar a utilização em produção. Ela ajuda a diagnosticar quando o modelo desenvolvido passa a apresentar potenciais problemas de *performance* devido a mudanças na população. Para avaliação das variáveis, em geral, se cria categorias com base na distribuição dos valores na amostra e, então, se calcula o PSI para a amostra de desenvolvimento e compara-se com o PSI da mesma variável em outros períodos, conforme exemplificado na Figura (6).

Variável	Categoria	Desenvolvimento	abr/21	mai/21	jun/21	jul/21	set/21	out/21	nov/21	dez/21
Idade	Entre 18 e 30 anos	34,7%	34,2%	35,3%	35,2%	31,5%	36,8%	36,7%	35,7%	36,0%
	Entre 31 e 40 anos	31,9%	30,4%	29,9%	31,0%	35,0%	32,1%	31,7%	31,6%	31,2%
	Mais do que 40 anos	33,4%	35,4%	34,7%	33,8%	33,5%	31,2%	31,6%	32,6%	32,8%
	PSI	**0,0%**	**0,2%**	**0,2%**	**0,0%**	**0,6%**	**0,3%**	**0,2%**	**0,0%**	**0,1%**

Figura 6: Exemplo da avaliação da estabilidade da variável idade.

Para todas as métricas, após finalizado o desenvolvimento dos modelos, é necessário realizar a comparação dos resultados entre as amostras de treinamento, teste e validação. Essa comparação é relevante para que seja avaliada a estabilidade das métricas quando o modelo é aplicado em uma amostra diferente. Espera-se, então, que os resultados das métricas sejam próximos entre as amostras, demonstrando que o modelo ajustado é capaz de extrapolar o aprendizado para dados diferentes dos que foram utilizados na amostra de treinamento.

3 Metodologia

Segundo (Gerhardt et al., 2009), o trabalho desenvolvido é uma pesquisa aplicada, visto que se propõe a resolver um problema real do setor de cobrança de uma empresa.

3.1 Etapas do desenvolvimento do trabalho

Esse trabalho teve o desenvolvimento baseado nas etapas de criação de modelos de *Credit Scoring* sugeridas por Selau (2008), conforme a Figura (7).

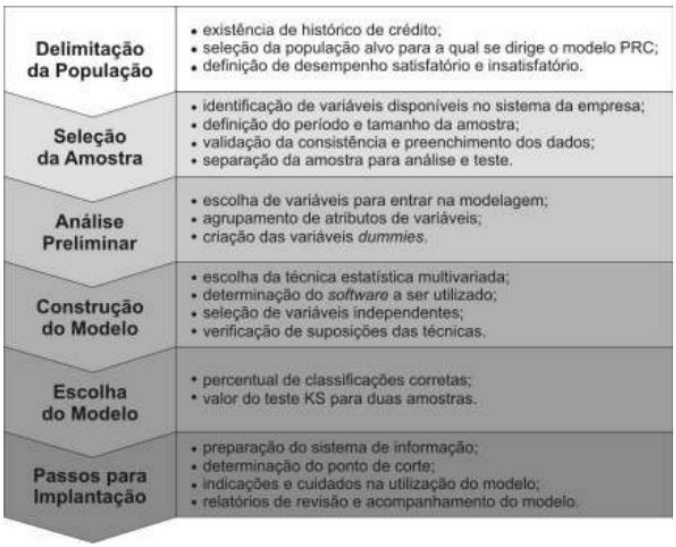

Figura 7: Etapas para desenvolvimento de modelos de *Credit Scoring*.

Fonte: Selau (2008).

Portanto, a seguir, são apresentadas as etapas do desenvolvimento deste trabalho:

1. Revisão bibliográfica;

Foram revisados conteúdos referentes aos pontos apresentados na seção 2: Regressão Logística Multinomial, *Machine Learning*, *Gradient Boosting*, *Extreme Gradient Boosting*, *Feature Selection*, *Permutation Importance*, *K-Fold Cross – Validation* e as métricas de avaliação de resultados dos modelos.

2. Delimitação da população e coleta do banco de dados:

Nesta etapa, é definido a população alvo para o estudo. Além desse fato, é realizada a avaliação do histórico de dados disponível para desenvolvimento do trabalho e fontes de dados para a criação das variáveis. As origens dos dados podem conter informações do tipo situacional e transacional. Situacional são

informações que não apresentam variações significativas ao longo do tempo e seus melhores exemplos são informações de cadastro. Já as informações transacionais são aquelas que possuem uma janela de tempo específica, podendo ser uma transação com intervalo diário (compras) ou mesmo com intervalos maiores (como faturas, as quais possuem intervalo mensal). Para tornar os dados transacionais em uma variável que possa ser utilizada na modelagem, foram selecionadas as janelas de tempo trimestral e semestral para sumarizar as informações (quantidade de compras realizadas no último trimestre, por exemplo).

Para algumas fontes de dados e janelas de tempo, foram aplicadas operações matemáticas, variando de acordo com a natureza do dado apresentado. As principais operações foram as seguintes: contagem, média, mínimo e máximo (como valor máximo do limite utilizado no último semestre).

3. Pré-processamento:

a) Seleção da amostra:

No desenvolvimento de modelos, separa-se o conjunto de dados em partes, sendo, no mínimo, uma para desenvolvimento e outra para testar o desempenho do modelo em uma amostra diferente. O objetivo deste processo é comparar as métricas de avaliação dos modelos entre as duas amostras para verificar se os resultados do modelo ajustado mantém-se estáveis quando o modelo é aplicado a um novo conjunto de dados. Neste estudo, utlizou-se a separação em amostras de treinamento, teste e validação.

b) Análise da variável resposta:

Esta etapa consiste em avaliar a distribuição da variável resposta no banco de dados. O objetivo desta avaliação é, dessa forma, verificar se existe desbalanceamento da variável resposta. Ao desenvolver um modelo sem considerar a desproporcionalidade da variável resposta nos dados, é possível acabar no caso em que os parâmetros do algoritmo não diferenciarão a classe minoritária das demais categorias, acreditando que estão agregando resultado devido à aparente alta acurácia (Chawla, 2002).

c) Seleção de variáveis:

Devido ao elevado número de variáveis que podem ser criadas para o desenvolvimento de modelos, uma etapa bastante importante é a seleção de

variáveis. Além da utilização de um método para realizar a seleção de variáveis, nessa etapa, foi feita uma análise preliminar com o objetivo de excluir variáveis que não seriam úteis para os modelos. Ao final deste processo, foi feita a normalização das variáveis quantitativas.

4. Ajustes dos modelos: desenvolvimento prático do modelo de Regressão Logística Multinomial e do *XGBoost.*

5. Resultados: comparação das métricas de avaliação de resultados entre a RLM e o *XGBoost.*

6. Discussão dos resultados.

3.2 Banco de dados

Para realizar o trabalho proposto, foi utilizado um conjunto de dados de uma empresa financeira que tem como produto a concessão de crédito. O conjunto de dados utilizado não é aberto ao público para garantir o sigilo das informações dos clientes da empresa estudada. A coleta e organização deste banco de dados foi parte deste trabalho e as informações utilizadas serão descritas na seção Resultados.

3.3 Implementação

A implementação do *Extreme Gradient Boosting* deu-se por meio da função *XGBClassifier(),* disponível no pacote *sklearn*. A parametrização do modelo foi a padrão da função implementada, com exceção do parâmetro *max_depth,* que foi definido incrementando 1, iniciando com o número 3, até que as métricas avaliadas se estabilizassem e nenhuma melhora fosse identificada, finalizando em 8. Também, foi indicado o número 3 para o parâmetro *num_class.*

Para a Regressão Logística Multinomial, criou-se um modelo por meio da função *LogisticRegression()*, também do pacote *sklearn*.

Para definição dos pesos de cada classe, utilizados para tratar o problema de desbalanceamento dos dados do algoritmo *XGBoost*, foi utilizada a função *compute_sample_weight()*, disponível, também, no pacote *sklearn*.

Para o balanceamento dos dados na RLM foi utilizada a função *RandomUnderSampler(),* também sendo parte do pacote *sklearn*.

Para definição da validação cruzada, foi utilizada a função *RepeatedKFold()* do pacote *scikit learn* e parametrização dada por *n_splits* igual a 10, n_*repeats* igual a 5, *random_state* igual a 7.

Os resultados da construção dos modelos de propensão à utilização de canais de negociação em cobrança com utilização das técnicas de Regressão Logística Multinomial e o *Extreme Gradient Boosting* são apresentados na seção Resultados. Todas as técnicas de modelagem utilizadas nesse trabalho foram construídas utilizando a linguagem de programação *Python* na versão 3.8 com apoio da interface do *Jupyter Notebook*. A *sintaxe* utilizada está disponível na seção Anexos.

4 Resultados

Nesta seção, são apresentados os resultados para a coleta e organização do banco de dados e o ajuste dos modelos de Regressão Logística Multinomial e *XGBoost*. Ao final da seção, é feita a comparação dos resultados obtidos com a aplicação das duas abordagens.

4.1 Delimitação da população e coleta do banco de dados

Conforme descrito anteriormente, para a realização do trabalho, foi utilizado o conjunto de dados de uma empresa financeira que concede crédito.

A população alvo são os clientes que possuem crédito e estão inadimplentes. A amostra é constituída por clientes que realizaram alguma negociação com a indicação do canal utilizado, sendo essa a variável resposta.

Na Tabela (1), são apresentadas as fontes de dados disponibilizadas acerca da relação do cliente inadimplente com a empresa e a quantidade de variáveis criadas. O período utilizado para desenvolvimento dos modelos compreende negociações de dívidas realizadas de abril de 2021 a dezembro de 2021, uma vez que esse seria o período em que todos os canais de negociação disponíveis já haviam sido implementados e estruturados. Também, foi observado diferenças nos dados ao comparar meses do ano de 2021 com meses do ano de 2020 devido à pandemia de COVID-19. Portanto, é preferido utilizar dados de meses mais próximos ao período atual. Foi aconselhado remover a amostra do mês de agosto de 2021 por problemas internos da empresa estudada.

Tabela 1: Assunto das variáveis explicativas e quantas variáveis foram criadas.

Assunto	Variáveis criadas
Acionamentos de cobrança	13
Interações com canais digitais de cobrança	46
Histórico de acordos	14
Histórico de compras no *e-commerce* da empresa	7
Estatísticas do histórico de dívidas	168
Dados cadastrais	22
Total de varáveis	270

4.2 Seleção da amostra

O conjunto de dados foi separado em treinamento, teste e validação segundo a Tabela (2). A amostra dos meses de abril de 2021 até outubro de 2021 foi separada de forma aleatória em 70% para treinamento e 30% para teste. Para o conjunto de treinamento, foi utilizada a validação cruzada 10-*folds*. Os meses mais recentes (de novembro de 2021 e dezembro de 2021), com o objetivo de simular o resultado do modelo em produção, foram utilizados apenas para validação.

Tabela 2: Organização do banco de dados em treinamento, teste e validação.

Período	Amostra	Quantidade de clientes
abr/21 a out/21	70% treinamento	938.071
	30% teste	402.031
nov/21 a dez/21	validação	395.732
Base total		1.735.834

4.3 A variável resposta

A variável resposta - canal de cobrança utilizado para negociação da dívida - é uma variável qualitativa nominal, inicialmente, com cinco níveis. Ao avaliar o desbalanceamento apresentado na Figura (8), foi discutido com a equipe de cobrança da empresa estudada a possibilidade de alguns canais serem agrupados. Dessa maneira, o canal 1 e o canal 5 foram agrupados devido à similaridade no acesso dos clientes e no uso estratégico da área de cobrança; além disso, os canais 2 e 4 foram agrupados após uma análise descritiva das variáveis e concordância da área de cobrança da empresa estudada. Logo, uma nova distribuição de canais de negociação foi identificada para a variável resposta final, conforme mostra a Figura (9).

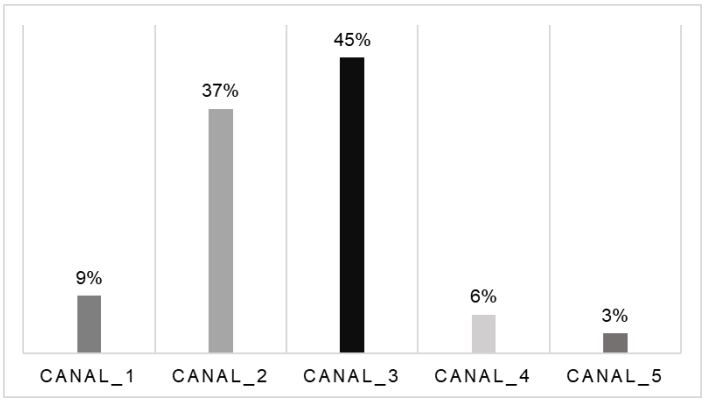

Figura 8: Gráfico da distribuição original de cada canal de cobrança no total de acordos.

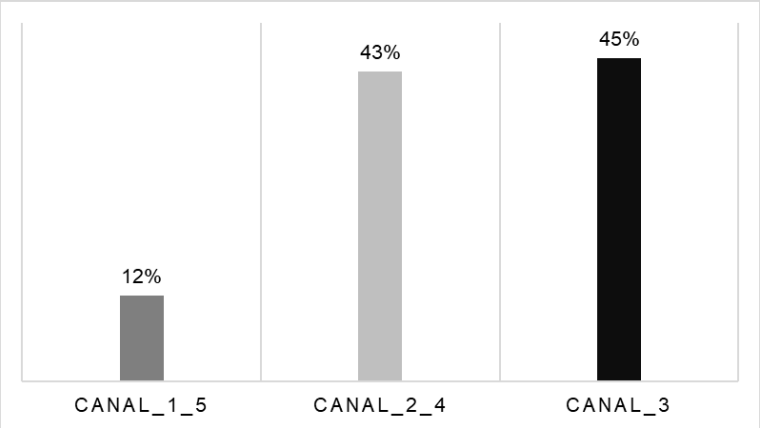

Figura 9: Gráfico da distribuição após o agrupamento dos canais de cobrança no total de acordos.

Apesar de que se tenha agrupado canais de cobrança semelhantes com o objetivo de minimizar o desbalanceamento dos dados, ainda é tido uma classe com representatividade menor em comparações com as demais. Em vista disso, foram testados três métodos de balanceamento de dados com cada técnica de modelagem.

O primeiro método foi o de estimação de pesos para cada classe, que consiste em ajustar pesos inversamente proporcionais às frequências das classes da variável resposta. Observe na fórmula abaixo:

$$PESO = \frac{Número\ de\ observações\ na\ amostra}{Número\ de\ classes * Número\ de\ observações\ na\ classe}$$

Esses pesos são importantes de serem ponderados para que a otimização do algoritmo aconteça em relação ao peso individual. Dessa forma, cada ponto é representado de forma igual. Na regressão logística, os coeficientes são ajustados através de um algoritmo de minimização para o logarítimo da verossimilhança negativa. Além disso, a ponderação é aplicada no cálculo da verossimilhança negativa ao ajustar o modelo para que valores de peso menores resultem em um valor de erro menor e, por sua vez, menos atualização nos coeficientes do modelo. Um valor de peso maior resulta em um cálculo de erro maior e, por conseguinte, mais atualização nos coeficientes do modelo.

Os outros métodos testados foram o *OverSampling,* que consiste em aumentar a quantidade de observações da classe minoritária, reamostrando com reposição, de forma aleatória, observações já existentes no conjunto de dados, e o método de *Undersampling,* que consiste em deixar todas as classes com a mesma

quantidade de observações da classe minoritária, reamostrando sem reposição, também, de forma aleatória (Maione, 2020).

Os métodos que apresentaram melhor performance e menor diferença entre treinamento e validação para a Regressão Logística Multinomial e o *XGBoost* foram *UnderSampling* e o método de estimação de pesos, respectivamente. Portanto, foram os métodos de balanceamento dos modelos finais. Os métodos de balanceamento de dados foram aplicados após a seleção de variáveis.

4.4 Limpeza do banco de dados e seleção das variáveis

Nessa etapa, foi feita a análise das variáveis que irão entrar na modelagem. Todas as variáveis explicativas que foram identificadas como quantidades de algum indicador (quantidade de acordos feitos utilizando o canal 1, por exemplo) tiveram as observações *missings* substituídas pelo valor zero.

Devido à quantidade expressiva de variáveis explicativas disponibilizadas, foi feita uma análise preliminar conforme a Tabela (3) para definir as melhores variáveis para fazerem parte do desenvolvimento do modelo.

Tabela 3: Análises preliminares feitas para remoção de variáveis e quantas foram removidas em cada etapa.

Análise	Descrição	Número de variáveis removidas
1. Sem variabilidade	Remoção das variáveis com desvio padrão igual a zero	7
2. Sem informação	Remoção das variáveis com 97% das observações sem informação	2
3. Correlação baixa	Avaliou-se a correlação entre as variáveis explicativas e a quantidade de vezes que cada uma se repetia com uma correlação acima ou igual a 70%. Entre cada par, excluiu-se a variável com maior número de repetições entre os pares.	167
Total restante		94

A maioria das variáveis removidas na análise 3 são relacionadas ao assunto “Estatísticas do histórico de dívidas”, o que era esperado dado que as variáveis são médias, somas, mínimos e máximos relacionados ao mesmo indicador.

Após a análise preliminar, ainda restaram 94 variáveis úteis para compor os modelos. Dado isso, foi utilizado o método de permutação de importâncias, valendo-se do *XGBoost* como algoritmo para realizar a seleção de variáveis, resultando em 26 variáveis ao final.

Foi realizada a normalização dos dados para fazer a seleção de variáveis e desenvolvimento dos modelos. Trata-se por normalização de dados o processo que padroniza as informações de diferentes variáveis. Por exemplo, os valores de uma fatura são expressos em reais com grandezas diferentes de valores de idade. Nesse trabalho, optou-se por padronizar os dados pela média, em que retira-se a diferença do valor pela média e divide-se pelo desvio padrão. Observe a fórmula a seguir:

$$\dot{x}_n = \left(\frac{x_n - \mu x}{\sigma x}\right)$$

4.5 Regressão Logística Multinomial

Conforme descrito neste trabalho, a Regressão Logística Multinomial foi escolhida como técnica base para comparação, pois é um método bastante utilizado no ciclo de crédito em geral. Foi feito um primeiro ajuste do modelo com as 26 variáveis selecionadas no método de permutação de importâncias; removeu-se as 4 variáveis que tiveram o p-valor maior do que o nível de significância de 5%, e novamente foi feito o ajuste do modelo com as 22 variáveis restantes.

Nas tabelas abaixo, avaliou-se a matriz de confusão e as métricas Precisão, *Recall* e *F-score* na amostra de validação.

Tabela 4: Matriz de confusão para a predição da amostra de validação - modelo RLM.

Resposta Predita	**Resposta observada**			
	Canal_1_ 5	**Canal_2_4**	**Canal_3**	**Total**
Canal_1_ 5	5.932	3.248	3.767	12.947
Canal_2_4	26.362	101.497	62.365	190.224
Canal_3	17.237	47.258	128.066	192.561
Total	**49.531**	**152.003**	**194.198**	**395.732**

Tabela 5: Métricas de avaliação para a predição da amostra de validação - modelo RLM.

Canal de negociação	Precisão	*Recall*	*F-score*
Canal_1_5	46%	12%	19%
Canal_2_4	53%	67%	59%
Canal_3	67%	66%	66%
Média	**55%**	**48%**	**48%**

Para o modelo utilizando a Regressão Logística Multinomial, na amostra de validação, tem-se que, para os 395.732 clientes, o modelo previu corretamente 235.495 (59,5%).

Conforme a Tabela (7), avaliando o PSI, o modelo manteve a distribuição da predição dos canais de negociação estável em comparação com a amostra de treinamento. Contudo, percebe-se que a distribuição da predição dos canais de negociação segue o mesmo padrão da distribuição da variável resposta na base de dados, conforme mostra a Figura 6.

Avaliando o *Recall* não se obteve resultados estáveis entre os canais. Em média, o modelo identificou o canal de negociação de 48% das observações de forma correta. De todos os clientes que utilizaram o canal_1_5 (classe minoritária), o modelo classificou corretamente 12%. De todos os clientes que utilizaram o canal_2_4, o modelo classificou corretamente 67%. De todos os clientes que utilizaram o canal_3 (classe majoritária), o modelo classificou corretamente 66%. O modelo apresenta resultados piores na classe minoritária.

Avaliando a precisão, novamente não se observa estabilidade devido ao baixo acerto para o canal_1_5. Todavia, assim como o *Recall*, observa-se uma melhora com relação aos demais canais. De todas as classificações para o canal_1_5 (classe minoritária), 46% eram realmente o canal_1_5. De todas as classificações para o canal_2_4, 53% eram realmente o canal_2_4. De todas as classificações para o canal_3 (classe majoritária), 67% eram realmente o canal_3.

Comparando as amostras de validação e treinamento, para alguns canais, observa-se estabilidade nas métricas e, em outros casos, apesar de esperada, existe uma diferença. Com relação às três métricas apresentadas na Tabela (6), o canal Canal_1_5 apresentou pequenas diferenças entre as amostras, permanecendo estável. O Canal_2_4 não ficou estável com relação à métrica de

Precisão, apresentando uma perda de performance e, no caso do *Recall*, a amostra de avalição teve um resultado mais alto em comparação à amostra de validação. Por fim, conforme a Tabela (6), o Canal_3 apresentou um resultado de Precisão ainda melhor na amostra de validação e estabilidade nas demais métricas.

Em relação a amostra de teste, todas as métricas tiveram resultados semelhantes a amostra de treinamento.

Tabela 6: Comparação das métricas de avaliação entre as amostras de treinamento, teste e validação - modelo RLM.

Canal de negociação	**Precisão**			***Recall***			***F-score***		
	Treino	**Teste**	**Validação**	**Treino**	**Teste**	**Validação**	**Treino**	**Teste**	**Validação**
Canal_1_5	47%	47%	46%	10%	11%	12%	17%	16%	19%
Canal_2_4	60%	60%	53%	61%	61%	67%	61%	60%	59%
Canal_3	59%	59%	67%	71%	70%	66%	64%	63%	66%

Tabela 7: Distribuição da predição entre os canais nas amostras de treinamento e validação - modelo RLM.

Canal de negociação	**Treinamento**	**Validação**
Canal_1_5	3%	3%
Canal_2_4	46%	48%
Canal _3	52%	49%
PSI	**0,0%**	**0,5%**

4.6 *Extreme Gradient Boosting*

Foi feito o ajuste do *XGBoost* para comparar com os resultados obtidos na RLM. Para o *XGBoost*, também foram utilizadas as 26 variáveis selecionadas no método de Permutação de importâncias. Foram consideradas as mesmas métricas de avaliação anteriores para avaliar os resultados deste modelo.

Tabela 8: Matriz de confusão para a predição da amostra de validação - modelo *XGBoost.*

Resposta Predita	**Resposta observada**			
	Canal_1_5	**Canal_2_4**	**Canal_3**	**Total**
Canal_1_5	33.967	45.369	41.264	120.600
Canal_2_4	9.897	81.983	37.930	129.810
Canal_3	5.667	24.651	115.004	145.322
Total	**49.531**	**152.003**	**194.198**	**395.732**

Tabela 9: Métricas de avaliação para a predição da amostra de validação - modelo *XGBoost*.

Canal de negociação	Precisão	*Recall*	*F-score*
Canal_1_5	28%	70%	40%
Canal_2_4	63%	55%	58%
Canal_3	79%	60%	68%
Média	**57%**	**61%**	**55%**

Para o modelo *XGBoost*, na amostra de validação, tem-se que, para os 395.732 clientes, o modelo previu corretamente 230.954 (58,36%). O modelo não apresenta seguir o comportamento de acertar mais a classe majoritária. Além disso, o modelo manteve a distribuição da predição dos canais de negociação estável, como apresentado na Tabela (11). A distribuição dos clientes fica, em média, 33% para cada canal de negociação predito.

Avaliando o *Recall*, percebe-se uma métrica estável entre os canais. Em média, o modelo identificou o canal de negociação de 61% das observações de forma correta. De todos os clientes que utilizaram o Canal_1_5 (classe minoritária), o modelo classificou corretamente 70%. De todos os clientes que usufruíram o Canal_2_4, o modelo classificou corretamente 55%. De todos os clientes que usaram o Canal_3 (classe majoritária), o modelo classificou corretamente 60%.

Avaliando a Precisão, não se observa a mesma estabilidade devido ao baixo acerto para o Canal_1_5, mas nota-se uma melhora com relação aos demais canais. De todas as classificações para o Canal_1_5 (classe minoritária), 28% eram realmente o Canal_1_5. De todas as classificações para o Canal_2_4, 63% eram realmente o Canal_2_4. De todas as classificações para o Canal_3 (classe majoritária), 79% eram realmente o Canal_3.

As métricas de avaliação de resultados para a amostra de validação ficaram estáveis em comparação com a amostra de treinamento, conforme a tabela (10). Em relação a amostra de teste, todas as métricas tiveram resultados semelhantes à amostra de treinamento.

Tabela 10: Comparação das métricas de avaliação entre as amostras de treinamento, teste e validação - modelo *XGBoost*.

Canal de negociação	Precisão			*Recall*			*F-score*		
	Treino	Teste	Validação	Treino	Teste	Validação	Treino	Teste	Validação
Canal_1_5	29%	28%	28%	67%	68%	70%	40%	40%	40%
Canal_2_4	70%	68%	63%	51%	50%	55%	59%	59%	58%
Canal_3	71%	70%	79%	64%	65%	60%	66%	68%	68%

Tabela 11: Distribuição da predição entre os canais nas amostras de treinamento e validação - modelo *XGBoost*.

Canal de negociação	Treinamento	Validação
Canal_1_5	33%	33%
Canal_2_4	28%	30%
Canal_3	40%	37%
PSI	**0,0%**	**0,5%**

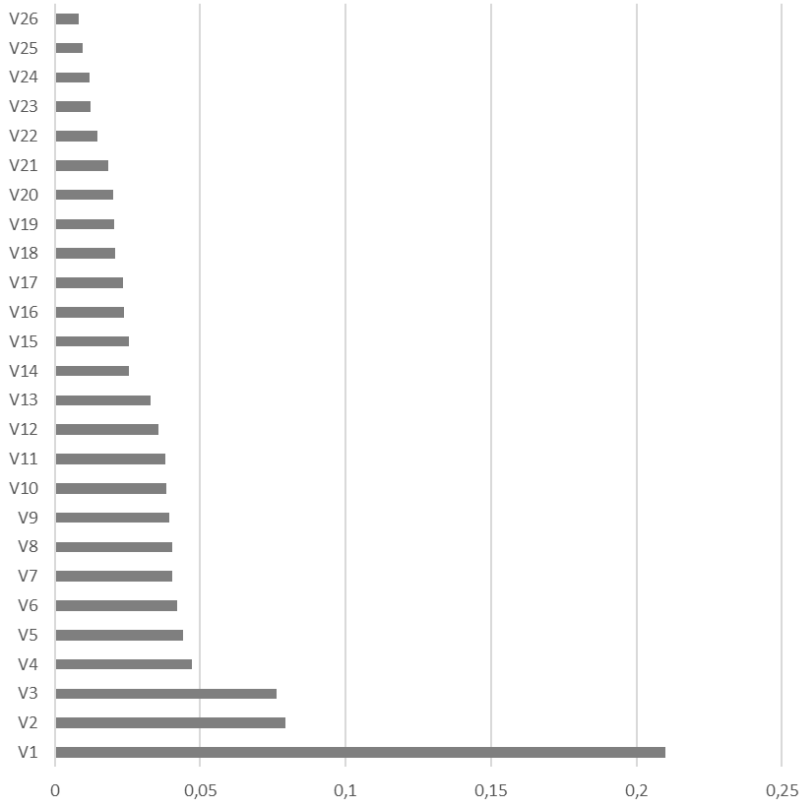

Figura 10: Importância das variáveis – modelo *XGBoost.*

Um ponto de observação para o modelo ajustado com a utilização do algoritmo *XGBoost* poderia ser o acúmulo de importância na variável mais importante do modelo e a diferença para as demais variáveis, apresentado na Figura (10). Como o intuito do desenvolvimento desse modelo é ser aplicado mensalmente

na base de dados dos clientes inadimplentes na empresa estudada, caso a variável em questão apresente alguma instabilidade, o desempenho do modelo poderia ser comprometido. Por conseguinte, foi avaliada a estabilidade das variáveis utilizando a métrica PSI. Como referência para o cálculo do PSI, foi utilizada a distribuição das variáveis na amostra de treinamento. Conforme a Tabela (12), nenhuma variável apresentou a necessidade de atenção, uma vez que todos os resultados de PSI estão abaixo de 10%.

Tabela 12: PSI em relação aos meses de treinamento para a distribuição das variáveis.

Variável	abr/21	mai/21	jun/21	jul/21	set/21	out/21	nov/21	dez/21
V1	0,0%	0,8%	0,2%	2,7%	0,1%	0,2%	0,1%	1,1%
V2	0,0%	0,0%	0,1%	0,1%	0,0%	0,0%	0,0%	0,1%
V3	0,2%	0,1%	0,0%	0,1%	0,6%	0,2%	0,4%	0,5%
V4	0,0%	0,1%	0,0%	1,4%	0,2%	0,8%	1,3%	2,9%
V5	0,1%	0,0%	0,1%	0,3%	0,0%	0,3%	0,5%	1,3%
V6	0,0%	0,1%	0,0%	0,1%	0,0%	0,1%	0,2%	0,3%
V7	0,0%	0,1%	0,1%	0,8%	0,0%	0,1%	0,2%	0,4%
V8	0,2%	0,2%	0,0%	0,6%	0,3%	0,2%	0,0%	0,1%
V9	0,7%	0,1%	0,2%	1,8%	1,4%	0,5%	0,0%	0,5%
V10	0,0%	0,4%	0,2%	0,3%	0,1%	0,0%	0,2%	1,1%
V11	2,6%	0,8%	0,0%	1,0%	4,1%	2,9%	3,4%	6,9%
V12	0,1%	1,0%	0,3%	3,8%	0,1%	0,4%	0,2%	0,7%
V13	0,0%	0,0%	0,0%	0,1%	0,0%	0,0%	0,0%	0,0%
V14	1,2%	0,2%	1,0%	1,4%	0,4%	0,2%	0,1%	0,0%
V15	0,0%	0,4%	0,2%	0,3%	0,1%	0,0%	0,2%	1,1%
V16	0,2%	0,0%	0,4%	0,2%	0,1%	0,0%	0,1%	0,0%
V17	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
V18	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
V19	0,0%	0,0%	0,0%	0,4%	0,2%	0,3%	0,3%	0,8%
V20	0,0%	0,0%	0,0%	0,0%	0,1%	0,0%	0,0%	0,0%
V21	0,0%	0,1%	0,0%	0,2%	0,0%	0,0%	0,0%	0,1%
V22	0,3%	0,0%	0,0%	0,8%	0,4%	1,1%	1,7%	3,0%
V23	0,1%	0,2%	0,0%	0,6%	0,0%	0,0%	0,0%	1,2%
V24	0,2%	0,4%	0,1%	2,2%	0,1%	1,2%	2,2%	6,1%
V25	0,2%	0,2%	0,6%	0,1%	0,3%	0,7%	0,0%	0,1%
V26	0,0%	0,0%	0,0%	0,3%	0,2%	0,2%	0,1%	0,1%

4.7 Comparação dos resultados e discussão

Na Tabela (13), são apresentados os resultados das métricas de avaliação em relação à amostra de validação para os dois modelos ajustados.

Tabela 13: Comparação das métricas de avaliação entre os modelos desenvolvidos.

Canal de negociação	**Precisão**		***Recall***		***F-score***	
	RLM	**XGBoost**	**RLM**	**XGBoost**	**RLM**	**XGBoost**
Canal_1_5	46%	28%	12%	70%	19%	40%
Canal_2_4	53%	63%	67%	55%	59%	58%
Canal_3	67%	79%	66%	60%	66%	68%
Média	55%	57%	48%	62%	48%	55%

Conforme as Tabelas (6) e (10), o *XGBoost* apresentou métricas mais estáveis em relação à amostra de treinamento e validação quando comparado a

RLM. Analisando o percentual de acertos geral das classificações, é tido que a RLM (59,5%) apresentou um resultado superior em comparação o *XGBoost* (58,36%).

Avaliando a métrica de Precisão, o modelo *XGBoost* apresentou resultados melhores para o Canal_2_4 (63%) e para o Canal_3 (67%) em comparação aos mesmos canais quando avaliado o resultado para a RLM (53% e 67%, respectivamente). Para o Canal_1_5, a RLM apresentou uma precisão melhor (46%) quando comparada a precisão do mesmo canal no *XGBoost* (28%).

Avaliando a métrica de *Recall*, o modelo de RLM apresentou melhor resultado para o Canal_2_4 (67%) em comparação ao mesmo canal no *XGBoost* (55%). Para o Canal_3, os dois modelos tiveram resultados próximos, sendo maior na RLM (66%) em comparação ao *XGBoost* (60%). Para o Canal_1_5, o *XGBoost* apresentou resultado superior (70%) em comparação ao mesmo canal quando avaliado o resultado para a RLM (12%).

O modelo utilizando *XGBoost* apresentou resultados mais positivos de precisão para dois canais. Apesar da RLM apresentar resultados melhores para 2 canais avaliando o *Recall,* o *XGBoost* garantiu que a classe minoritária também fosse bem discriminada, deixando os resultados mais parcimoniosos entre os tipos de canais.

A Precisão e o *Recall* são duas métricas que descrevem tipos de acertos importantes. Em vista disso, a métrica *F-Score* é calculada a partir de uma média harmônica da Precisão e do *Recall* do modelo. Logo, o *XGBoost* também apresenta resultados melhores para o *F-Score*, principalmente, avaliando o resultado para a classe minoritária. Em média, avaliando Precisão, *Recall* e *F-score*, os resultados do *XGBoost* foram superiores em relação à RLM.

Na literatura, estudos comparam a utilização de Regressão Logística e *Extreme Gradient Boosting*, em diferentes desfechos no ciclo de crédito, para resposta binária, e observam resultados melhores com o *XGBoost*, como Chang et al. (2018) e Carmona et al. (2018).

Os dois métodos tiveram bons resultados de classificação para os tipos de canais de negociação de cobrança. Além do mais, os dois modelos apresentaram piores resultados quando avaliamos a Precisão com relação ao canal com menor representatividade na base de dados.

Para o *Recall,* apenas a RLM apresentou pior resultado de classificação em relação ao canal com menor representatividade na base de dados.

A pior performance de classificação do canal minoritário corrobora com os resultados encontrados por Wang et al. (2018) em um estudo que compara a RLM e o *XGBoost* para a classificação de escolha de meios de locomoção em viagens.

5 Considerações Finais

Técnicas estatísticas trazem ganhos financeiros significativos para as empresas que fazem gestão de crédito. Esse trabalho teve como objetivo propor que mais um modelo fosse inserido no ciclo de crédito: um modelo de propensão à utilização de canais de negociação de dívidas.

Estudos demonstram a importância de ser entendido o perfil do cliente inadimplente dado às mudanças comportamentais ocasionadas por circunstâncias relacionadas à pandemia de COVID-19 pela digitalização dos meios de comunicação e pelas novas gerações tomadoras de crédito. Entender o perfil comportamental do cliente pode possibilitar uma melhor gestão dos recursos e ações estratégicas pelas áreas de negócio.

Esse trabalho considerou a Regressão Logística como técnica base por ser usualmente aplicada nos modelos do ciclo de crédito das empresas financeiras. A Regressão Logística Multinomial serviu de referência para realizar a comparação com o algoritmo *Extreme Gradient Boosting*.

Os dois modelos ajustados nesse trabalho apresentaram resultados satisfatórios para a discriminação dos canais de negociação com maior representatividade na base de dados. Para o canal minoritário, o *XGboost* apresentou bons resultados a respeito da discriminação também. Em contraposição, a RLM não apresentou resultados satisfatórios com respeito ao canal minoritário.

As variáveis mais importantes para o ajuste do modelo de *XGBoost* dizem respeito à busca e à interação do cliente, já inadimplente, com os canais de comunicação digitais disponíveis pela empresa. Além disso, outro grupo de variáveis que se apresentaram como mais importantes na discriminação dos canais de negociação são relacionadas à utilização deles, pelo cliente, para tentar negociar dívidas anteriormente.

Em concordância com estudos que apontam uma mudança de perfil de cliente com relação a gerações, variáveis como idade e tempo de relacionamento do cliente com a empresa também se mostraram importantes para a predição do desfecho final.

Devido às limitações de falta de histórico disponível, não foi possível utilizar todas as fontes de dados existentes em relação aos meios de acionamento dos clientes inadimplentes. Também não foi possível avaliar a interação dos clientes aos

estímulos para canais digitais enquanto o cliente estava adimplente. Possivelmente, com mais informações, poderiam ter sido encontrados resultados mais expressivos.

A inclusão de um modelo que identifique qual é o canal mais provável que um cliente realize uma negociação de dívida pode ser muito importante para complementar a cadeia de decisões do ciclo de crédito. Combinado a resultados de outros modelos, como modelos de *Collection Score,* em que o desfecho é probabilidade de pagamento, o modelo proposto pode ajudar na elaboração de estratégias diferentes e dinâmicas.

Para o problema que foi identificado nesse trabalho, concluiu-se que o *XGBoost* se caracteriza como uma boa alternativa ao modelo de Regressão Logística Multinomial. Em média, todas as métricas avaliadas (Precisão, *Recall*, *F-score*) foram superiores no ajuste do modelo de *XGBoost* em comparação a RLM. Com o modelo de *XGBoost,* foi possível identificar, em média, 61% dos canais de negociação de forma equilibrada entre todos os canais, inclusive o canal minoritário. Dessa forma, combinado aos modelos de probabilidade de pagamento, pode ser possível otimizar os recursos de cobrança, sabendo qual cliente é mais provável de realizar um pagamento e por onde é mais provável que seja feita essa negociação.

Foi produzida a apresentação dos resultados obtidos para a equipe de negócio da empresa estudada. Os resultados encontrados satisfizeram à necessidade de negócio no que diz respeito a entender o canal mais adequado para estimular o cliente a fechar uma negociação. O primeiro teste, em produção, já está sendo estruturado.

Referências

Altman, E. I., Saunders, A. (1997). Credit risk measurement: Developments over the last 20 years. Journal of banking & finance, 21(11-12):1721–1742.

Altmann, A., Tolosi, L., Sander, O., Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. Bioinformatics, Volume 26, Issue 10, 15 May 2010, Pages 1340–1347.

Amankwah-Amoah, J., Khan, Z., Wood, G., Knight, G. (2021). COVID-19 and digitalization: The great acceleration. Journal of Business Research 136; 602–611.

Aniceto, M. C. (2016). Estudo comparativo entre técnicas de aprendizado de máquina para estimação de risco de crédito. Master's thesis, Universidade de Brasília.

Breiman, L. (1996). Bagging Predictors. Kluwer Academic Publishers. Boston. Manufactured in The Netherlands 24; 123 – 140.

Carmona, P., Climent, F., Momparler, A. (2019). Predicting failure in the U.S. banking sector: An extreme gradient boosting approach. International Review of Economics and Finance, 61; 304–323.

Chang, Y. C., Chang, K. H., Wu, G. J. (2018). Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. Applied Soft Computing Journal, 73; 914 – 920.

Chawla, N., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research (Vol. 16).

Chen, T., Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. in :Proceeding of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco (2016) pp. 785-794

Cheng, F., Yang, C., Zhou, C., Lan, L., Zhu, H., Li, Y. (2020). Simultaneous Determination of Metal Ions in Zinc Sulfate Solution Using UV–Vis Spectrometry and SPSE-XGBoost Method. Sensors, 20, 4936.

Figueira, C. V. (2006). Modelos de regressão Logística. Master's thesis, Universidade Federal do Rio Grande do Sul.

Friedman, J. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis 38; 367–378.

Gerhardt, T. E. e Silveira, D. T. (2009). Métodos de pesquisa. Editora da UFRGS.

Hosmer, D. W., Lemeshow, S. (1989). Applied Logistic Regression. Series in Probability and a Mathematical Statistics. Nova York: John Wiley, 1989.

James, G., Witten, D., Hastie, T., e Tibshirani, R. (2013). An introduction to statistical learning, volume 112. Springer

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino R. P., Tang, J., Liu, H. (2017). Feature Selection: A Data Perspective. ACM Comput. Surv. 50, 6, Article 94 (December 2017), 45 pages.

Lopes, L. S. L. (2004). Definição de um modelo de cobrança (Collection Score) utilizando regressão logística multinomial. Trabalho de conclusão de curso, Universidade Federal do Rio Grande do Sul.

Machado, A. R. (2015). Collection Scoring via Regressão Logística e Modelo de Riscos Proporcionais de Cox. Master's thesis, Universidade de Brasília.

Machkour, B., Abriane, A. (2020). Industry 4.0 and its implications for the financial sector. Procedia Computer Science 177; 496–502.

Maione, C. (2020). Balanceamento de dados com base em *oversampling* em dados transformados. Tese de Doutorado, Universidade Federal de Goiás.

Moraes, L. G. (2012). Uma abordagem alternativa de Behavioral Scoring usando modelagem híbrida de dois estágios com regressão logística e redes neurais. Trabalho de conclusão de curso, Universidade Federal do Rio Grande do Sul.

Naser, M.Z., Alavi A. (2020). Insights into Performance Fitness and Error Metrics for Machine Learning. https://arxiv.org/abs/2006.00887

Nunes, R. E. (2018). Gestão do ciclo de crédito dos cartões Private Label e os nativos digitais. Master's thesis, Universidade Federal do Rio Grande do Sul.

Selau, L. P. R. (2008). Construção de modelos de previsão de risco de crédito. Master's thesis, Universidade Federal do Rio Grande do Sul.

Tsai, C. F., Hsu, Y. F., & Yen, D. C. (2014). A comparative study of classifier ensembles for bankruptcy prediction. Applied Soft Computing Journal, 24, 977–984.

Wang, F., Ross, C. L. (2018). Machine Learning Travel Mode Choices: Comparing the Performance of an Extreme Gradient Boosting Model with a Multinomial Logit Model. Transportation Research Record, Vol. 2672(47) 35 – 45.

Yurdakul, B. (2018). Statistical properties of population stability index. Western Michigan University.

6 Anexo I – Sintaxe utilizada

```
###Regressão Logística
## Separação dos dados em x e y
x_lr, y_lr = tmp[numericas], tmp['RESPOSTA']

## Separação em dados de treino e teste
x_train_lr, x_test_lr, y_train_lr, y_test_lr = train_test_split(x_lr, y_lr, train_size = 2/3, random_state =
42)

## Balanceamento de classes
rus = RandomUnderSampler()
x_train_lr, y_train_lr = rus.fit_resample(x_lr, y_lr)

## Salvar lista de índices
idx_treino_lr = list(x_train_lr.index)
idx_teste_lr = list(x_test_lr.index)

## Normalização dos dados
scaler_lr = StandardScaler().fit(x_train_lr)

x_train_lr = pd.DataFrame(scaler_lr.transform(x_train_lr))
x_train_lr.columns = numericas

x_test_lr = pd.DataFrame(scaler_lr.transform(x_test_lr))
x_test_lr.columns = numericas


lr_model_001 = LogisticRegression(random_state=0, multi_class='multinomial', penalty='none',
solver='newton-cg').fit(x_train_lr, y_train_lr)
preds = lr_model_001.predict(x_test_lr)

###XGBOOST
## Segmentação dos dados em x, y
x, y = df[(df.AMOSTRA == 'DEV')][features], df[(df.AMOSTRA == 'DEV')]['RESPOSTA']

## Separação em dados de treino e teste
x_train, x_test, y_train, y_test = train_test_split(x, y, train_size = 2/3, random_state = 42)

## Salvar lista de índices
idx_treino = list(x_train.index)
idx_teste = list(x_test.index)

## Normalização dos dados
scaler = StandardScaler().fit(x_train)

x_train = pd.DataFrame(scaler.transform(x_train))
x_train.columns = features

x_test = pd.DataFrame(scaler.transform(x_test))
x_test.columns = features

sample_weight = compute_sample_weight('balanced', y_train)

k_fold = RepeatedKFold(n_splits = 10, n_repeats = 5, random_state = 7)
ml_model_001 = GridSearchCV(
   estimator = XGBClassifier(
      num_class = 3,
      max_depth = 8,
      metric='multiclass',
```

```
        objective = 'multi:softmax',
        eval_metric='mlogloss'
    ),
    param_grid = parametros, scoring = 'f1', n_jobs = -1, cv = k_fold, verbose = 1)

eval_set = [(x_train, y_train), (x_test, y_test)]

ml_model_001.fit(x_train, y_train, eval_set = eval_set, sample_weight = sample_weight,
early_stopping_rounds = 20, verbose = 0)
```