Modelo multinível e o efeito empresa para a detecção de fraudes: uma comparação com modelos Logístico e XGBoost

Resumo

O combate às fraudes no setor financeiro apresenta-se como um grande desafio para as instituições bancárias. Sob a hipótese de que o contexto de cada companhia é relevante para detecção dessas infrações e consequentemente para a modelagem de um algoritmo de prevenção, este trabalho propôs-se a construir uma abordagem multinível que considera tais relações na sua formulação, a fim de compará-lo com outros modelos e técnicas tradicionalmente implementados no mercado. Para tal, foram considerados 26.434 processos de abertura de contas, em 39 organizações e desenvolvidos três modelos: Multinível, XGBoost e Regressão Logística Binária. Dentre os resultados obtidos, observou-se um melhor desempenho do modelo Multinível para as métricas AUC ROC e "log loss", para um nível de confiança de 95%, demonstrando a capacidade do modelo em capturar os efeitos aleatórios dos agrupamentos e, portanto, uma alternativa viável ao problema proposto.

Palavras-chave: Modelagem, Multinível; XGBoost; Regressão Logística; Fraude.

Multilevel model and the company effect for fraud detection: a comparison with Logistic and XGBoost models

Abstract

Combating fraud in the financial sector is a big challenge for banking institutions. From the hypothesis that the context of each company is relevant for these infractions' detection and for the modeling of a prevention algorithm, this work proposed to build a multilevel approach that considers such relationships to compare them with other models and techniques traditionally implemented in the market. To this end, three models were developed from a database with 26,434 account opening processes in 39 organizations: Multilevel, XGBoost, and Binary Logistic Regression. Among the results obtained, the multilevel model presented a superior result in the comparative AUC ROC (for a confidence level of 95%), demonstrating the potential to capture the random effects of clusters.

Keywords: Multilevel; Mixed Modeling; Logistic Regression; XGBoost; Fraud.

Introdução

Entre as diversas consequências do período pandêmico, encontra-se a aceleração dos processos de digitalização das mais diversas atividades, sejam elas corriqueiras ou pontuais, vividas por cada indivíduo em sua esfera pessoal e profissional. Sob a óptica macroeconômica, um dos setores impactados foi o financeiro, cuja transformação digital foi amplamente catalisada pelo distanciamento social e os hábitos de consumo virtuais intensificados. As pessoas passaram a utilizar com maior frequência outras modalidades de pagamento que não o papel moeda e o cartão de crédito físico, o que fez com que as organizações vissem suas transações digitais crescerem consideravelmente e novos hábitos se formarem dentro da sociedade (Mastercard New Payments Index, 2021). Soma-se ainda, em ambiente nacional, a implementação de uma nova modalidade de pagamento virtual em novembro de 2020: o PIX. Em seu primeiro ano, ela se tornou a principal modalidade de

transferência de dinheiro, seja em relações comerciais ou em troca de valores entre pessoas físicas (Banco Central do Brasil, 2022), reforçando ainda mais o quadro delineado anteriormente. Portanto, a sociedade brasileira avançou consideravelmente em seu processo de digitalização e bancarização simultaneamente.

Além das transformações citadas, destaca-se ainda a grande mudança que o setor vem passando com o crescimento das modalidades conhecidas como “as a service": BaaS (“Banking as a Service”), CaaS (“Credit as a Service”) e IaaS (“Investment as a Service”). Em linha geral, estas modalidades transformam os produtos das instituições financeiras em serviços, permitindo que seus clientes corporativos sejam os fornecedores das soluções para os clientes finais (pessoas físicas ou jurídicas). Desta forma, novas estruturas de mercado são constituídas e resultam em novas hierarquias nas relações entre instituições financeiras, empresas e clientes finais.

Por trás dessa realidade, estão as novas arquiteturas tecnológicas das empresas do setor. Os tradicionais bancos e as novas “fintechs encontram-se cada vez mais munidos de capacidade de armazenamento, processamento e disponibilização dos dados em ambiente “big data”. Este avanço, impulsionado pelo momento atual, também tem permitido que as empresas repensem seus modelos tradicionais de prevenção às fraudes e a adoção de novas soluções, inclusive de novas abordagens estatísticas e computacionais nesta frente. As fraudes bancárias, um dos principais temas do setor, estão longe de serem estanques, ao contrário, evoluem na mesma velocidade em que são combatidas, sendo tradicionalmente um dos principais desafios de qualquer empresa que atue nesta cadeia de processos.

Quando se atualiza o combate às fraudes bancárias dentro deste novo cenário, novos prós e contras podem ser enumerados. Por um lado, a grande e rápida digitalização pode favorecer os mais diversos golpes, sejam provenientes de mecânicas de engenharia social já conhecidas, sejam novas modalidades de “phishing” e “malware” tipicamente digitais, por exemplo. Por outro, a digitalização das operações bancárias habitualmente provê às instituições maior volume de dados, sejam os transacionais, que caracterizam as movimentações financeiras, sejam os cadastrais, que permitem maior conhecimento dos usuários em seu processo de “onboarding” e ao longo do seu relacionamento com as organizações.

Nesse contexto de novas hierarquias mercadológicas de um setor financeiro “as a service”, observa-se a oportunidade de experimentação da abordagem multinível para a construção de modelos de detecção à fraude. Para além disso, nota-se a oportunidade de uma abordagem comparativa entre ela e outros modelos mais utilizados no mercado, como a classe dos modelos lineares generalizados. Por fim, por se tratar de um estudo de caso, delimita-se o escopo desse trabalho ao problema das fraudes ideológicas. Dessa forma, este

estudo tem como objetivo desenvolver e comparar um modelo multinível com outros modelos para a prevenção às fraudes cadastrais em uma estrutura de mercado hierárquica.

Assim, busca-se apresentar uma aplicação prática da abordagem multinível, ampliando a discussão sobre sua pertinência no campo do risco e prevenção à fraude, fundamental para as instituições financeiras. Tal aplicação se dá considerando este novo contexto social e tecnológico, que se consolida sob o guarda-chuva de transformações intitulado de “novo normal”.

Material e Métodos

Para a melhor compreensão dos dados, faz-se necessário conhecer o modelo de atuação da instituição à qual os dados pertencem originalmente. Pode-se descrever de maneira sucinta a atuação da empresa como uma fornecedora de soluções tecnológicas que possibilitam outras empresas, sejam elas do segmento financeiro ou não, oferecer aos seus clientes serviços bancários, como emissão e processamento de cartões, abertura de contas digitais, adquirência e risco & “compliance”. Ou seja, uma empresa sob a modalidade BaaS. Desta forma, a base de dados para este projeto contempla empresas de segmentos e serviços distintos, criando condições potencialmente favoráveis para o desenvolvimento do modelo multinível proposto.

Portanto, este trabalho é um estudo de caso com o objetivo de comparar o modelo multinível com diferentes outros modelos para a prevenção a fraude, a partir de um conjunto de dados de uma instituição que atua em uma estrutura de mercado hierárquica, posicionada como fornecedora de tecnologia e serviços de BaaS para diversas companhias. A Figura 1 a seguir, mostra de maneira simplificada esse ecossistema e os aninhamentos presentes em cada nível hierárquico.

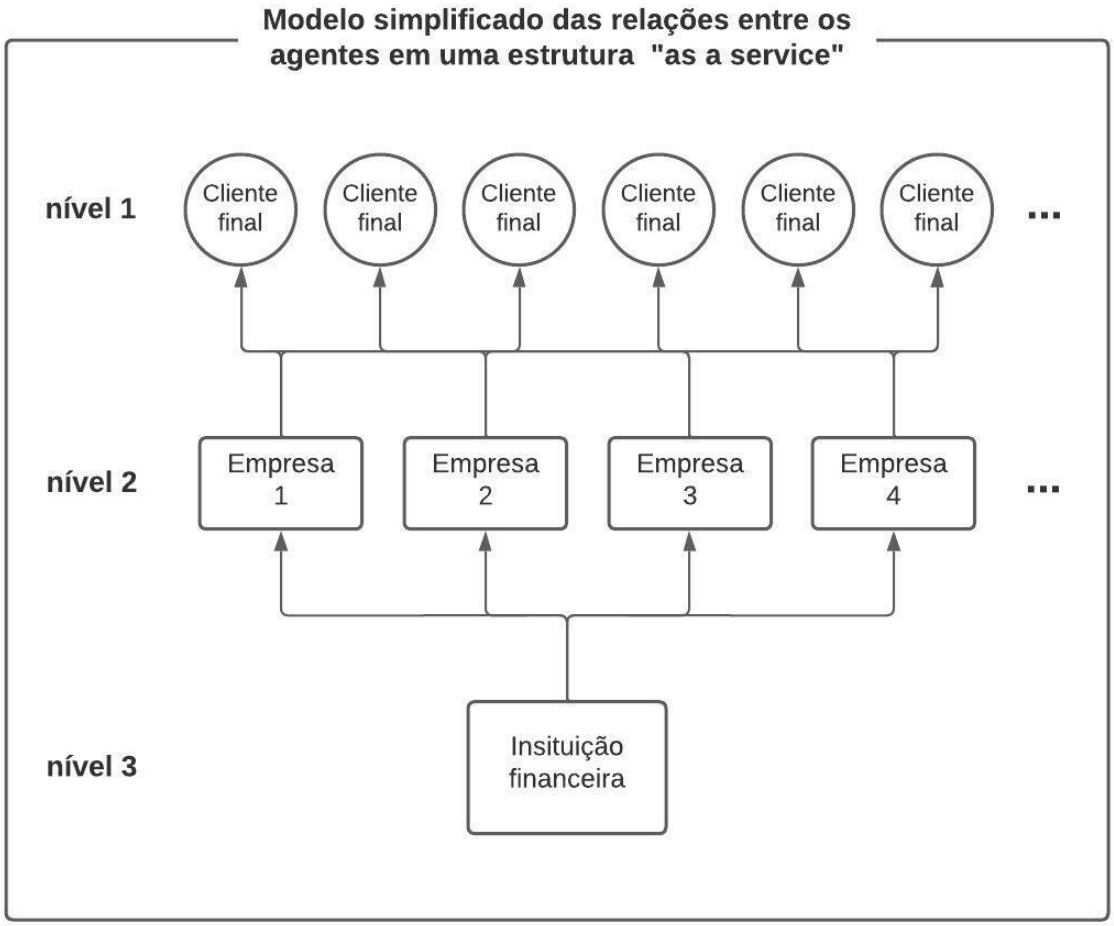

Figura 1. Modelo simplificado das relações entre agentes em uma estrutura “as a service”
Fonte: Dados originais da pesquisa

Sobre o conjunto de dados explorado, refere-se a uma base de processos de abertura de contas e as respectivas variáveis correspondentes à análise destas solicitações. Esta base conta ainda com uma avaliação binária (aprovação ou reprovação) definida por uma equipe de analistas. Mais detalhes sobre as variáveis e o fluxo operacional que as geram serão apresentados no decorrer deste trabalho. Desse conjunto, foram extraídas todas as requisições analisadas por algum profissional da empresa no período de 01/08/2022 a 20/09/2022, totalizando 26.434 observações.

Os dados de cada módulo de avaliação cadastral foram coletados diretamente do banco de dados da instituição (anonimizados), em um ambiente cloud, utilizando linguagens PySpark, Python e SQL pela plataforma Databricks. Uma vez tratados, foram concatenados constituindo uma única base, contendo as ocorrências, as variáveis explicativas e a variável resposta. Por fim, a amostra extraída foi compilada e armazenada em formato CSV para ser utilizada na exploração e desenvolvimento dos modelos.

Com os dados em mãos, foi adotada a metodologia CRISP-DM como biblioteca ferramental e processo de mineração de dados, adaptando-se cada etapa às idiossincrasias deste projeto. Conforme define Shearer (2000), o processo de mineração de dados pode ser dividido em seis etapas: problema de negócio, entendimento dos dados, preparação dos dados, modelagem, validação e implementação. Além da sequência de execução, o modelo prevê em diversos momentos o retorno a estágios anteriores, caso os resultados obtidos impliquem em uma revisão de uma etapa já superada, conforme a Figura 2 a seguir:

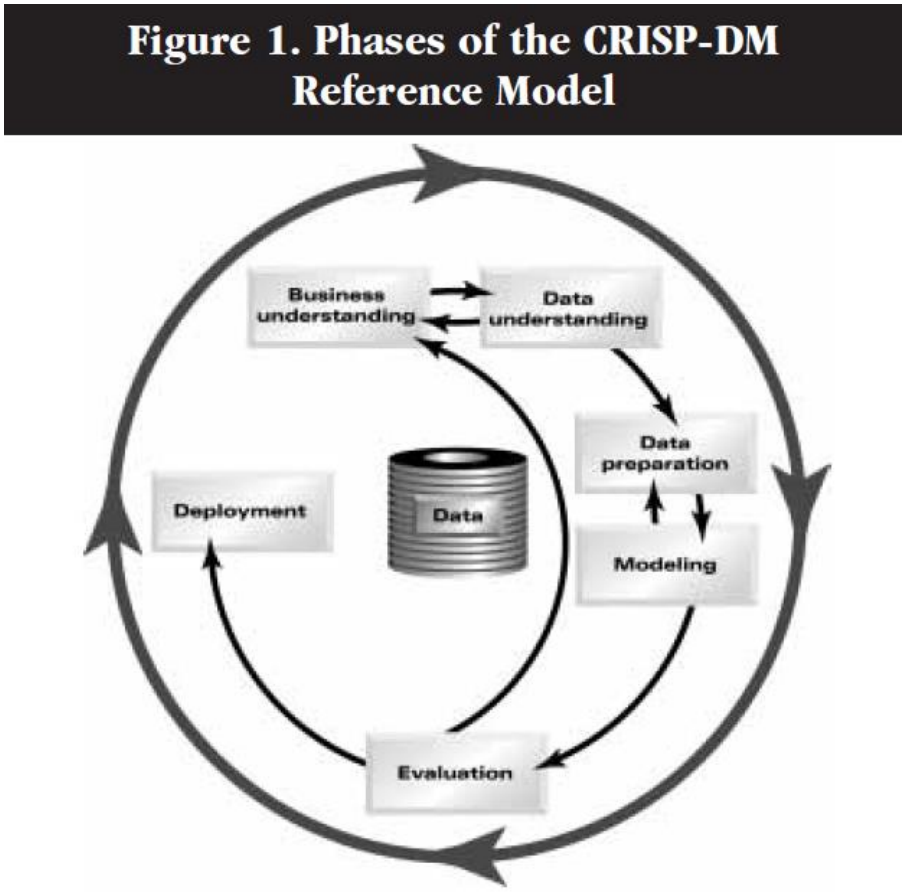

Figura 2. CRISP-DM
Fonte: Shearer C., 2000

Problema de negócio

O combate às fraudes transacionais é mais do que uma necessidade formal acompanhada pelo Banco Central, é uma atividade crítica para as empresas do universo financeiro, pois envolve a administração e o equilíbrio de diversos objetivos. Por um lado, uma atuação mais conservadora sobre as suspeitas de fraude ajuda a instituição a proteger-se de danos de magnitudes diversas, desde pequenas quantias até volumes consideráveis de

dinheiro. Por outro, adotar critérios mais conservadores e extrapolar os casos fraudulentos, impedindo a entrada de possíveis clientes idôneos ou transações regulares, pode implicar na perda de receita e impactar no crescimento da companhia. A implementação destas decisões sobre o que é ou não fraude foram realizadas por equipes de analistas ao longo do tempo, compondo um custo operacional considerável para as empresas e, às vezes, com critérios amplos e de desempenhos questionáveis.

Portanto, o problema de negócio está em encontrar soluções que acelerem o processo de abertura de contas, reduzindo custos e minimizando a exposição da companhia às fraudes financeiras.

Como alternativa a este desafio está a adoção de modelos de “Machine Learning”, que vem trazendo novas perspectivas para as conhecidas “torres de risco”. Quando tais modelos se mostram capazes de substituir as atividades manuais das equipes de maneira considerada satisfatória, são muitos os ganhos a serem colhidos pelas empresas. Pode-se citar: volume de transações analisadas (escalabilidade da operação), velocidade de aprendizagem, tempo de análise, padronização de critérios, redução de custos e caráter inovador atribuído à marca.

Entendimento dos dados

Para o melhor entendimento dos dados utilizados neste trabalho, é necessário inicialmente compreender o processo operacional de abertura de contas. Ao solicitar o cadastramento em uma determinada empresa, o cliente fornece uma série de dados, sejam eles cadastrais, fotos e/ou arquivos. Tais dados passam então a serem analisados em uma esteira, composta por diversas etapas de avaliação denominadas módulos. Os resultados destas análises são armazenados em bancos de dados separados, relacionados por uma chave primária que permite a formação de um processo único. Estas saídas (ou “outputs”) de cada módulo assumirão o papel de variáveis explicativas dos modelos.

Vale ressaltar que um determinado cadastro pode ser automaticamente reprovado por algum módulo. Caso isso não ocorra, ele passa para a análise seguinte, até chegar a penúltima etapa de avaliação, que tem como objetivo: aprovar, recusar, ou declarar-se neutra. Caso o “status” seja neutro, o processo é direcionado à mesa de operações para ser analisado por um agente humano. Esta mesa de operações é considerada o último módulo, sendo responsável por analisar os dados cadastrais e todas as saídas das etapas anteriores, retornando um entre dois resultados: aprovado ou recusado.

Dessa forma, os modelos de “machine learning” seriam inseridos hipoteticamente imediatamente antes da mesa, conforme a Figura 3. Para isso, eles devem ser treinados tendo

como variável resposta o “status” da mesa de operação e como variáveis explicativas as múltiplas saídas de cada módulo.

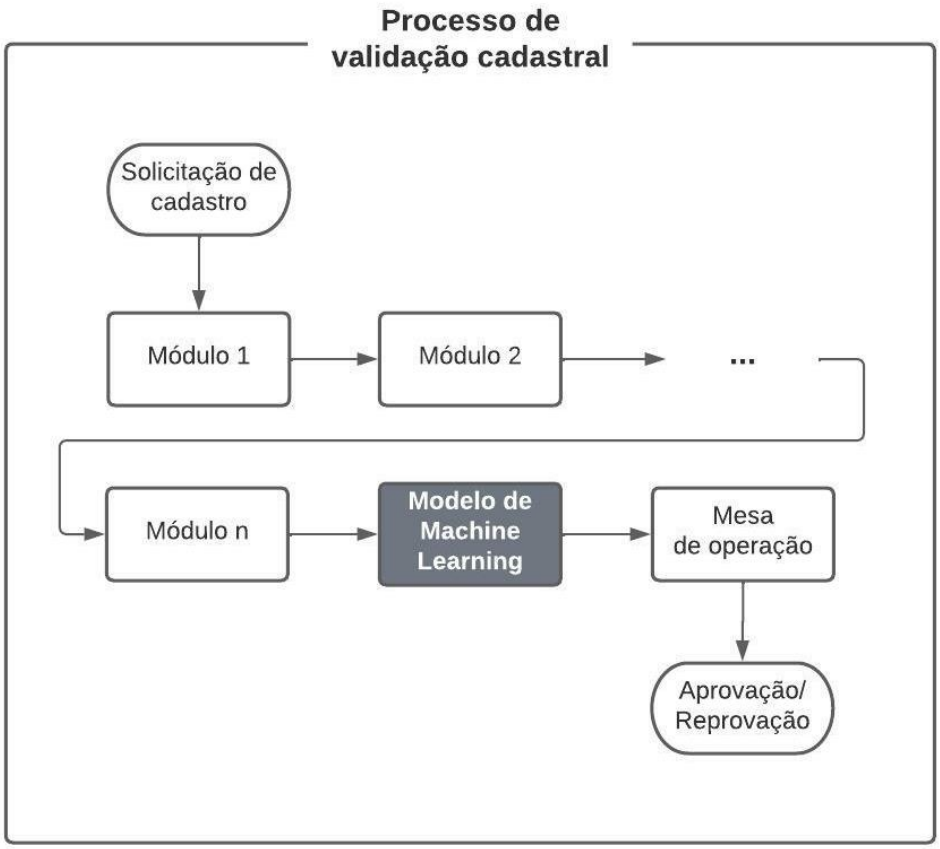

Figura 3. Processo de validação cadastral
Fonte: Dados originais da pesquisa

Para preservar o sigilo destes dados bem como o caráter estratégico desse processo, todas as ocorrências receberam novas chaves primárias aleatórias. Os módulos, por sua vez, foram renomeados de forma padrão, como “var1”, “var2” etc. A título de exemplificação, as variáveis podem ser uma classificação (como aprovação, neutralidade ou recusa), um valor numérico correspondente a um grau de similaridade, ou ainda um índice de aprovação.

Preparação dos dados

Com os dados coletados, foram necessárias diversas etapas de preparação até a constituição de uma base única de treinamento para o desenvolvimento dos modelos. A Tabela 1 consolida os tratamentos adotados, que foram construídos como funções (também chamados métodos na linguagem Python), a fim de otimizar o desenvolvimento e a implementação dos processos.

Tabela 1. Métodos de preparação e pré-processamento

ID	Método	Descrição
1	Ranking	Seleção da avaliação final em cada módulo.
2	Filtro	Filtro da base sem valores faltantes em determinadas variáveis que garantia que a requisição havia percorrido todos os módulos.
3	Imputação manual	Imputação de dados segundo regras de negócio para algumas métricas e dimensões.
4	Aplicação de prefixo	Criação de uma nomenclatura que identificasse cada módulo com a aplicação de prefixo por meio de dicionário.
5	Imputação iterativa	Utilizando a biblioteca Scikit-learn Iterative Imputer para a imputação de valores em variáveis numéricas utilizando um modelo de regressão linear.
6	Imputação simples	Utilizando a biblioteca Scikit-learn Simple Imputer para a imputação de valores em variáveis categóricas com a regra de aplicação de valores mais frequentes.
7	Padronização	Processo de padronização das variáveis numéricas.
8	“Dummy”	Transformação de variáveis categóricas em variáveis “dummies” (0/1). Para as bases dos modelos não hierárquicos, a variável que identifica o subemissor também foi inclusa no processo.
9	“Encode”	Transformação da variável “target” (resposta) em representação numérica (0, 1): 0 para não evento (aprovação) e 1 para evento (fraude/reprovação).
10	VIF	Exclusão de variáveis numéricas que apresentem fator de inflação da variância (VIF) maior que 5.
11	Seleção de “Features”	Seleção de variáveis por critérios próprios para cada modelo, colaborando com o tratamento de “overfitting”
12	Treino, teste, validação	A base foi dividida em três partes. A base de treino, composta de 60% das observações. A base de teste, 20% das observações e utilizada para a engenharia de hiperparâmetros. A base de validação, composta por outros 20% da base.
13	Modelagem	Construção dos modelos propostos: Multinível, Regressão Logística e XGBoost

Fonte: Dados originais da pesquisa

Para maior compreensão do processo de preparação dos dados, construiu-se o fluxo mostrado na Figura 4:

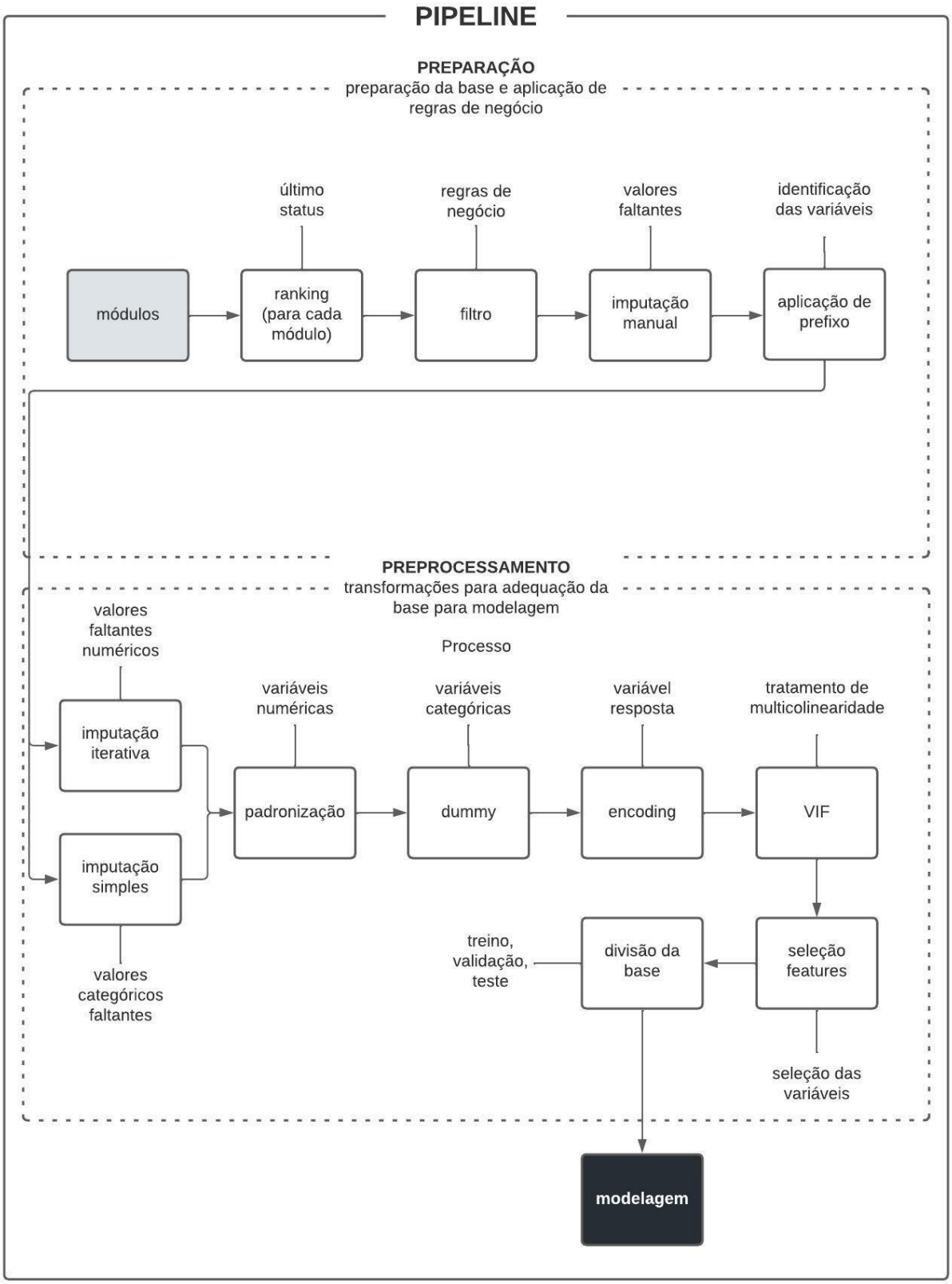

Figura 4. Fluxo de preparação e pré-processamento dos dados

Fonte: Dados originais da pesquisa

Ressalta-se ainda que o processo de treinamento, teste e validação foi escolhido a fim de proporcionar absoluta igualdade de condições entre o modelo hierárquico, o modelo de regressão logística e o modelo selecionado com o apoio da plataforma de AutoML ("Automated Machine Learning"). Esta plataforma utiliza tal técnica com a finalidade de preparar o modelo em uma base de treino, encontrar a melhor recomendação de hiperparâmetros em uma base de testes e apurar os resultados em uma base de validação. Por conseguinte, utilizamos a mesma técnica na construção dos três modelos.

Os processos de imputação foram aplicados para variáveis julgadas com baixo índice de valores faltantes. Tal julgamento foi feito a partir da observância da preservação das curvas de distribuição dos dados após as substituições. Observou-se também o impacto das imputações nas estatísticas de média, mediana, curtose e assimetria, a fim de trazerem a menor oscilação possível à estrutura das variáveis. Reforça-se ainda que, a fim de tornar este trabalho publicável, respeitando os sigilos necessários para tal, a base teve suas chaves-primárias e os nomes das métricas e dimensões dos módulos mascarados ao longo do processo de preparação.

Por se tratar de três abordagens diferentes, as técnicas na seleção de variáveis também foram aplicadas adequando-se a cada caso, como apresentado na tabela a seguir. Três técnicas foram experimentadas, sendo os resultados da AUC ROC da base de validação a métrica utilizada para definir qual seria adotada. As duas primeiras foram variações do estimador da biblioteca SelectFromModel da plataforma Scikit-learn entre XGBoostClassifier e LogisticRegression. Tais estimadores foram os de melhor desempenho para os modelos XGBoost e GLM respectivamente. A terceira técnica foi a seleção pelo teste chi-quadrado, com p-valor estipulado em 0,05, da biblioteca SelectFwe da plataforma Scikit-learn. O modelo "ensemble" GPBoost, por sua vez, apresentou resultado semelhante ao da biblioteca XGBoost, tendo melhor performance considerando as variáveis selecionadas com a biblioteca SelectFromModel e estimador XGBoostClassifier. Dessa forma, duas bases foram geradas: uma para a criação do modelo GLM Logístico Binário e outra para os modelos XGBoost e Multinível.

Tabela 2. Técnicas de seleção de variáveis utilizadas

Técnica	Biblioteca	Parâmetros
Seleção por modelo	SelectFromModel	estimator = XGBoostClassifier, threshold =mean
Seleção por modelo	SelectFromModel	estimator = LogisticRegression(max_iter=1000), threshold = mean
Teste chi-quadrado	SelectFwe	score_func = chi2, alpha=0.05

Fonte: Dados originais da pesquisa

Modelagem

Inicialmente, cabe destacar as três abordagens utilizadas para a construção dos modelos e a maneira pela qual foram selecionadas. A primeira delas, a multinível, deu-se principalmente pela aderência da estrutura de mercado ao modelo teórico, com agrupamentos por meio da hierarquia existente entre a entidade financeira (fornecedora dos dados), seus clientes diretos (as outras empresas, aqui chamadas de subemissores) e os clientes finais. Outra motivação deu-se pela percepção de ser oportuno o desenvolvimento de um estudo sobre tal técnica, tendo em vista o número ainda restrito de publicações sobre esta abordagem.

A segunda escolhida foi a Regressão Logística Binária (GLM). Isso se deu pela ampla utilização, seja acadêmica, ou no mercado desta abordagem clássica. Por não considerar a captura dos efeitos aleatórios de nível grupo em sua formulação, o seu contraponto com a abordagem multinível torna possível evidenciar o impacto de considerar-se os agrupamentos sociais sobre as métricas de avaliação no processo de modelagem.

Após um experimento de “AutoML” (procedimento padrão no processo de modelagem da companhia e amplamente difundido no mercado), o modelo com maior valor de área sob a curva ROC foi o XGBoost. Por esta razão, a terceira abordagem foi construída utilizando a biblioteca de mesmo nome. Com isso, formou-se o objeto de estudo deste projeto, com três abordagens diferentes, a partir de critérios próprios, unindo questões acadêmicas a soluções de mercado, sobre um problema real.

Vale destacar ainda que a biblioteca utilizada para o desenvolvimento do modelo multinível trata-se de uma combinação de três modelos (“ensemble”): Processo Gaussiano, Modelo de Efeitos Mistos (“Mixed Effects Model”) e Árvores de Decisão (LightGBM). Assim, por tratar-se de um modelo baseado em árvores de decisão, tem-se aqui um contraponto com o modelo XGBoost, que também parte deste método.

Como resultado, por um lado, desenvolveu-se um modelo de classificação binário, com abordagem multinível para classificação de solicitações de abertura de contas como aptas ou fraudulentas. Por outro, com o desenvolvimento de dois outros modelos, uma Regressão Logística e um XGBoost, foi possível estabelecer uma comparação entre as três soluções.

Validação

A etapa de validação deve adequar-se ao caso em questão, no que diz respeito ao problema de negócio, às características do conjunto de dados e à modelagem executada. Neste estudo comparativo, foram extraídas métricas de desempenho tais como Acurácia, Precisão, “Recall”, área sob a curva ROC (AUC ROC) e F1 Score. Visualizações também

foram utilizadas para este fim, como a Matriz de Confusão e o gráfico Precisão & Recall x “Threshold”. Porém, a título de comparação dos modelos, a métrica principal escolhida foi AUC ROC, mostrada também de forma gráfica.

Implementação

A etapa de implementação, ou “deploy”, corresponde à construção da esteira de produção (ou “pipeline”) e à implementação propriamente dita do modelo dentro de um processo produtivo. Por não ser parte do escopo deste projeto, esta etapa não foi desenvolvida; porém, espera-se que este trabalho possa colaborar com soluções semelhantes na organização estudada e ambientes nos quais os leitores atuem.

Resultados e Discussão

Nesta etapa de resultados e discussão, diversos gráficos e estatísticas serão apresentados. Contudo, a fim de permitir um conhecimento completo dos códigos implementados, das etapas de modelagem, das tabelas e dos gráficos, foi criado um repositório disponível no endereço: https://github.com/samuel-haddad/MBAUSP.

O primeiro passo após a preparação e o pré-processamento da base foi a exploração dos dados a fim de conhecer melhor os dois “datasets”. Os gráficos a seguir apresentam a distribuição das requisições pelos subemissores bem como a proporção de aprovados e recusados por cada um deles. Apesar da fraude ser um evento raro quando se observa a base completa de solicitações, na última etapa do processo (de avaliação humana) há um maior equilíbrio maior entre as duas classes, sendo 65% o percentual de aprovação e 35% o de reprovação, conforme a Figura 5. Isso se dá, pois a maior parte dos casos idôneos já foram aprovados. Portanto, restam apenas os casos aos quais os módulos não foram capazes de dar um parecer favorável, ou desfavorável.

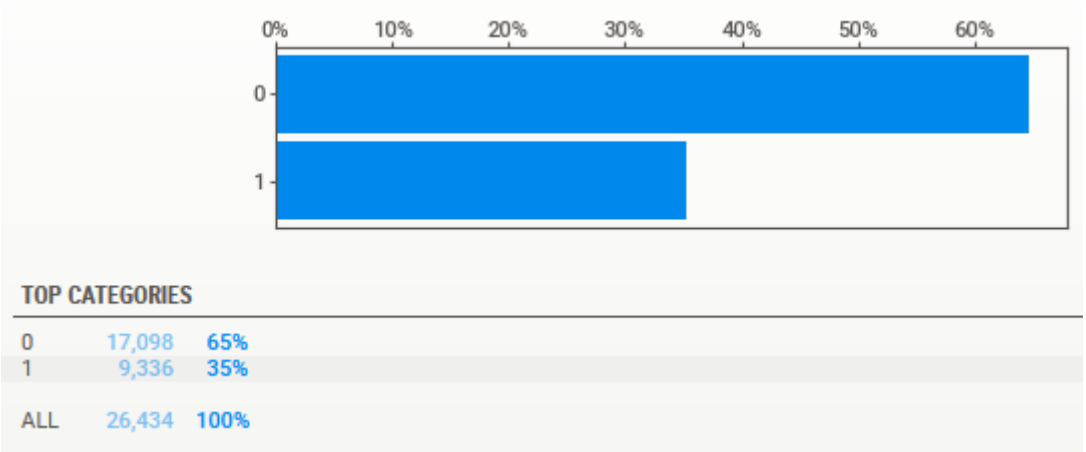

Figura 5. Distribuição da aprovação e reprovação da base
Fonte: Resultados originais da pesquisa

Cada subemissor está representado por um valor numérico na base de dados na variável “group”. Olhando a distribuição dos dados na Figura 6, é possível observar a concentração do volume de solicitações em alguns deles, como 234, 223, 211, 182, 279, 161 e 109. Partindo da hipótese de que as fraudes e inconsistências nestes grupos sejam em alguma medida caracterizadas por estes agrupamentos e o modelo multinível seja capaz de capturar seus feitos aleatórios, tal concentração reforçará a pertinência da construção do modelo multinível, apresentando um cenário adequado para essa abordagem.

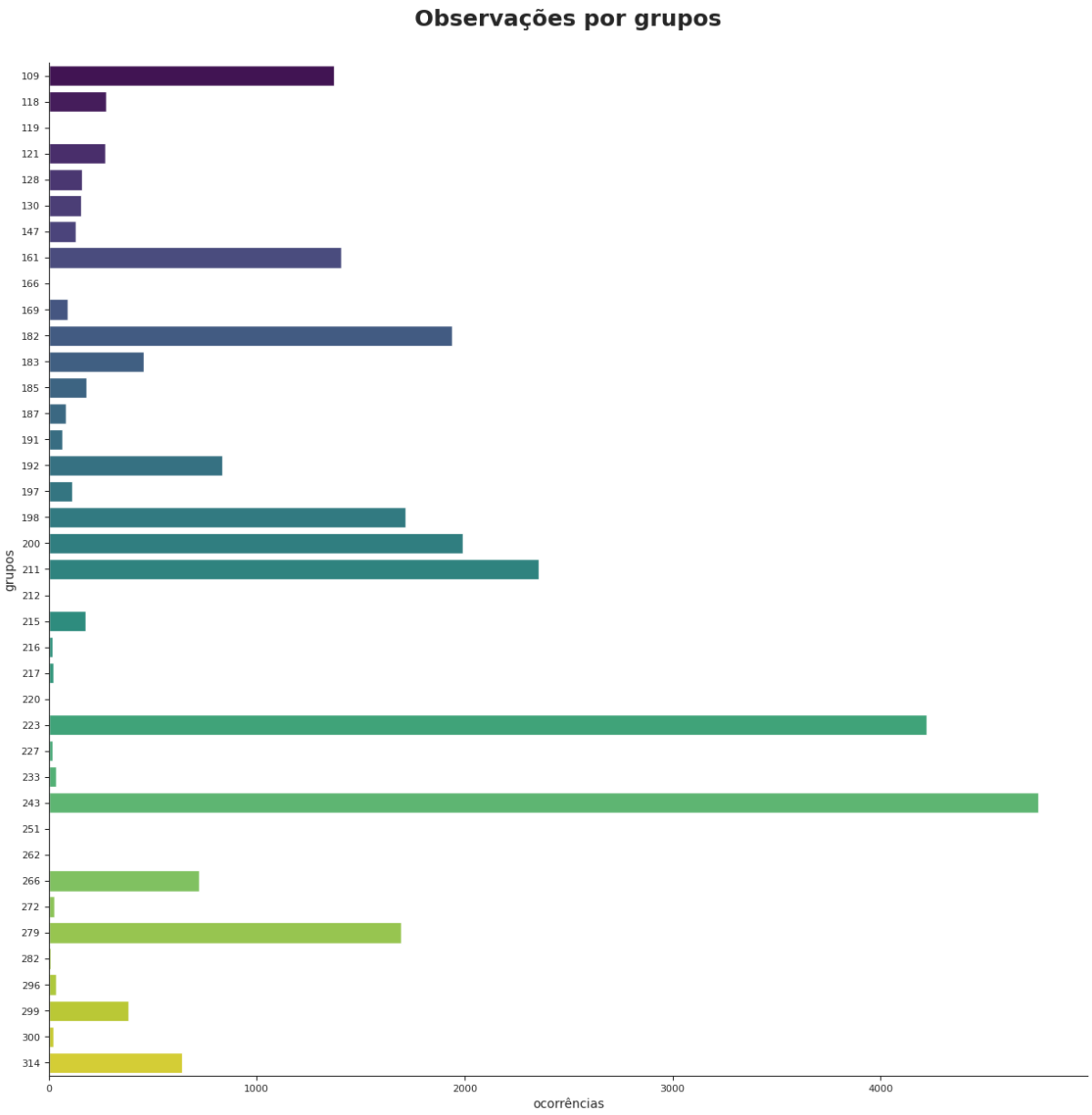

Figura 6. Distribuição das observações por grupo

Fonte: Dados originais da pesquisa

No que diz respeito às variáveis numéricas, foram criados gráficos de caixa a fim de compreender as suas dispersões. Todas elas estão contempladas na Figura 7, mesmo que algumas possam ter sido removidas no processo de seleção de “features” na geração das três bases descritas na etapa de preparação dos dados. Nos gráficos é possível observar a existência de outliers; porém, estes não receberam um tratamento diferenciado, devido ao fato de a própria fraude ser um comportamento anormal esperado. Assim, optou-se pela observação das métricas de desempenho ao final da geração dos modelos e recomenda-se a criação de métricas de monitoramento para o caso de serem colocados em produção.

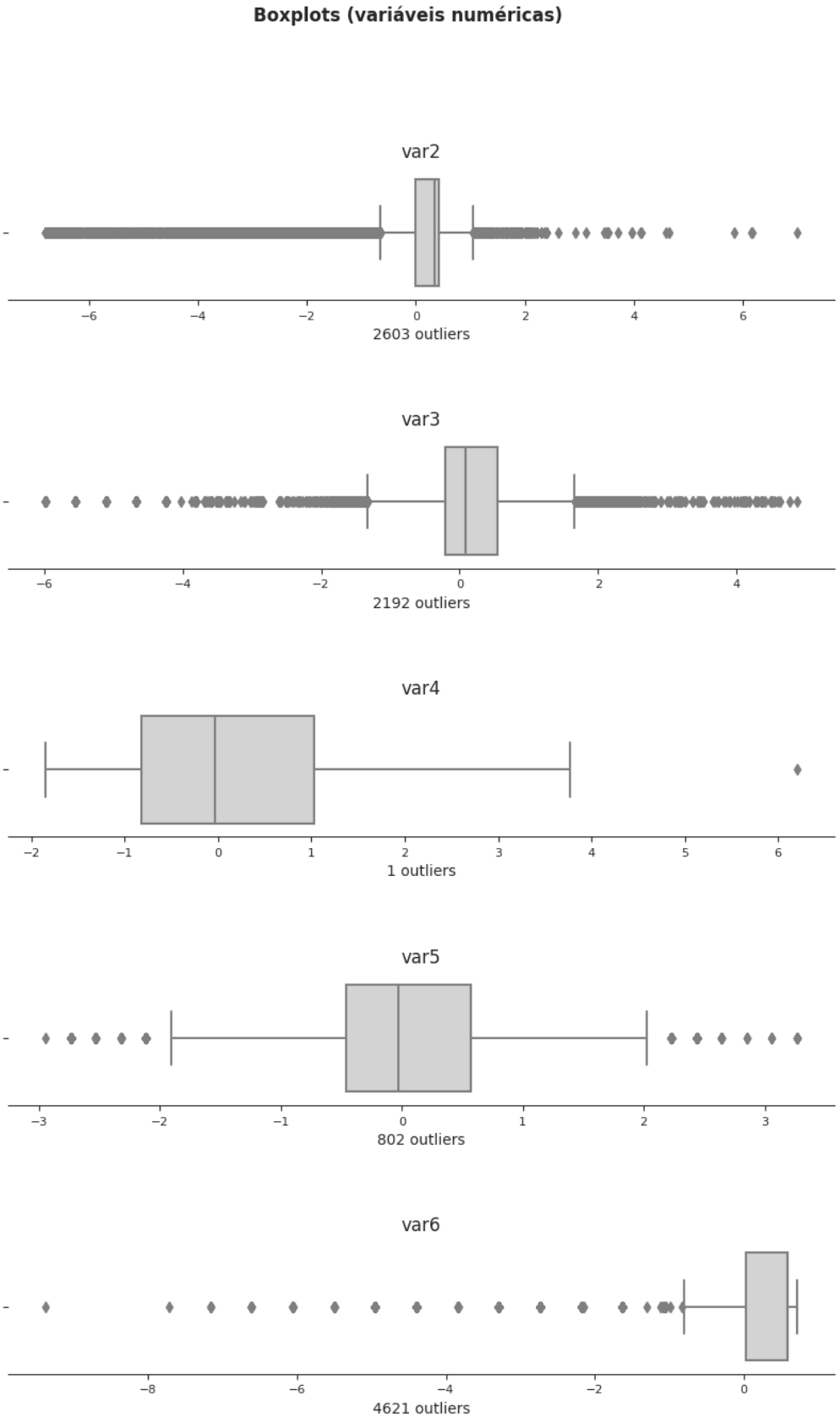

Figura 7. Diagrama de caixa das variáveis contínuas

Fonte: Dados originais da pesquisa

Por fim, foi gerada uma matriz de correlação e associação entre as variáveis; porém ela não foi explorada com maior detalhamento devido à grande quantidade de variáveis explicativas empregadas no modelo. Os gráficos apresentados podem ser observados no "notebook" de exploração dos dados no sítio https://github.com/samuel-haddad/MBAUSP/blob/main/eda_tcc.ipynb.

Regressão Logística

Conforme Fávero e Belfiore (2019) definem, o objetivo de uma regressão logística é estimar a probabilidade de ocorrência de um evento baseado no método da verossimilhança. Mais especificamente, neste projeto o modelo implementado é uma Regressão Logística Binária, pela variável resposta possuir apenas duas opções de classificação: fraude e não fraude. Para sua implementação, a biblioteca utilizada foi a "Logistic Regression" da Scikit-learn, que possui como diferencial, para algumas abordagens clássicas, a aproximação utilizada na definição das variáveis preditoras e seus respectivos parâmetros. No lugar da aplicação do procedimento "stepwise", o algoritmo implementa procedimentos de regularização, uma técnica comum em aprendizado de máquinas, conforme apresentado na documentação do modelo disponível em https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression.

No que diz respeito à construção do modelo, foi utilizada como técnica de otimização de hiperparâmetros um processo iterativo de "grid search". Os principais resultados estão na Figura 8 e o código completo pode ser visitado na página https://github.com/samuel-haddad/MBAUSP/blob/main/logreg_model.ipynb.

	amostra	acuracia	precisao	recall	auc_roc	f1_score	logloss(normalizada)
0	teste	76.30	71.41	54.67	80.25	61.93	0.502163
1	validacao	76.96	72.46	54.59	81.09	62.27	0.492416

Figura 8. Métricas de desempenho – GLM Logística Binária
Fonte: Dados originais da pesquisa

Dentre os modelos construídos, a regressão logística foi a que apresentou o menor percentual abaixo da curva ROC. Em comparação ao modelo GPBoost, que será apresentado ao final, todas as métricas foram inferiores. Reforça-se, novamente, que as métricas de comparação dizem respeito à amostra de validação. Nas figuras 9, 10 e 11 estão o

desempenho do modelo em três representações gráficas: Matriz de confusão, AUC ROC e Precisão & “Recall” x “Threshhold”.

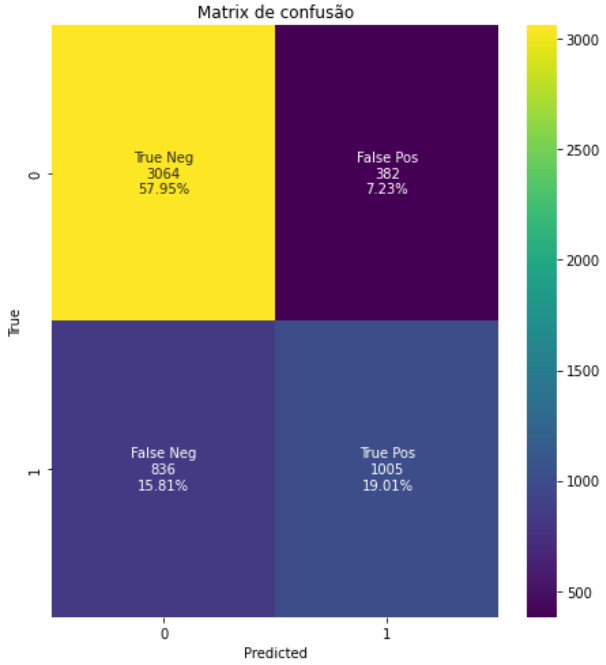

Figura 9. Matrix de Confusão

Fonte: Dados originais da pesquisa

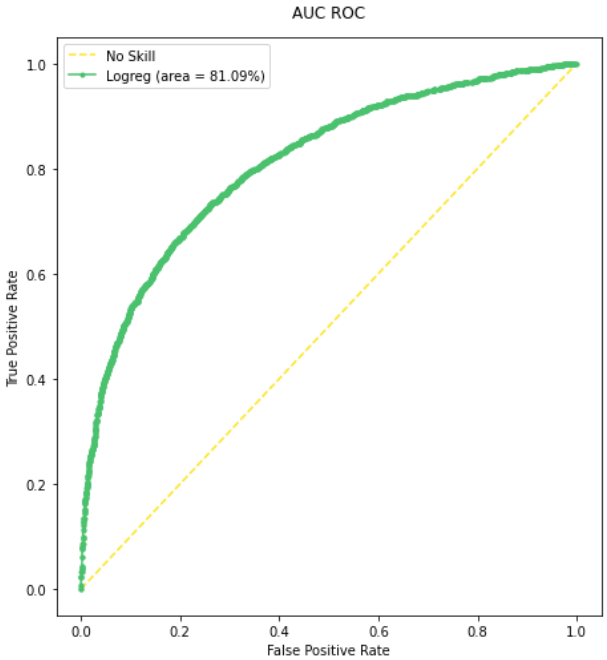

Figura 10. AUC ROC

Fonte: Dados originais da pesquisa

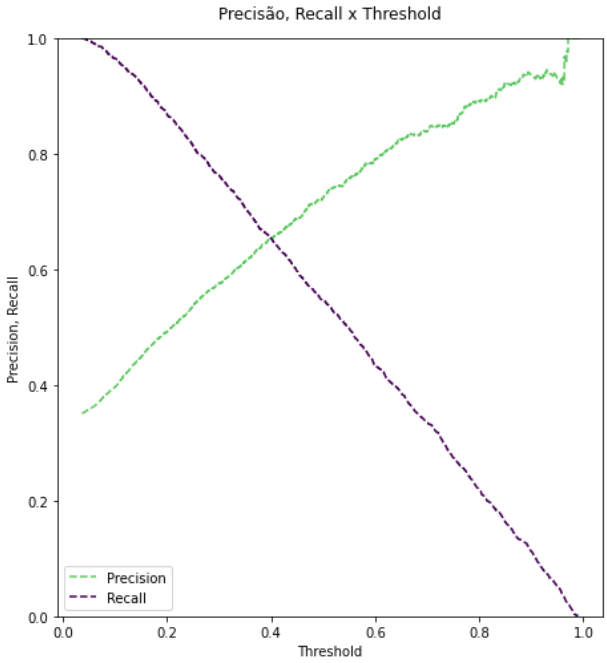

Figura 11. Precisão & “Recall” x “Threshold”

Fonte: Dados originais da pesquisa

XGBoost

O “Extreme Gradient Boost” é um algoritmo baseado em árvore de decisão, que utiliza uma estrutura otimizada de “gradient boosting”, e ganhou notoriedade por sua ampla adoção nas competições de “machine learning” ao redor do mundo, conforme o levantamento feito no repositório Distributed (Deep) Machine Learning Community, (2022). Originalmente, o algoritmo nasceu de um projeto de pesquisa de Tianqi Chen e popularizou-se dentro da comunidade de cientista de dados, sendo implementado posteriormente como pacote para as mais diferentes linguagens (XGBoost developers, 2022). Atualmente, é uma biblioteca de código aberto e conta com a contribuição da comunidade para o seu contínuo desenvolvimento.

A biblioteca em questão foi a de melhor desempenho em um experimento de AutoML desenvolvido na plataforma Databricks, segundo a métrica comparativa adotada, AUC ROC. No experimento, diversos pacotes e parâmetros foram testados, como LightGBM, Logistic Regression e a própria XGBoost, com o objetivo de desenvolvimento de um modelo de classificação binária. Uma vez extraídos os hiperparâmetros, o modelo foi gerado novamente, conforme o código que está disponível em: https://github.com/samuel-haddad/MBAUSP/blob/main/xgboost_model.ipynb. A tabela com as métricas de desempenho pode ser observada na Figura 12.

	amostra	acuracia	precisao	recall	auc_roc	f1_score	logloss(normalizada)
0	teste	77.57	73.28	57.24	81.41	64.28	0.488186
1	validacao	78.14	74.38	56.76	82.33	64.39	0.477099

Figura 12. Métricas de desempenho – XGBoost
Fonte: Dados originais da pesquisa

As métricas de validação, em linhas gerais, foram superiores às de teste, que apresentam configuração padrão da biblioteca. Já ao se observar o gráfico de Precisão, “Recall” x “Threshold” da Figura 15, comparativamente ao GLM percebe-se que há uma oscilação vertical menor da curva de precisão quanto maior o ponto de corte. Isso indica uma taxa de erros menor nesta faixa. Este é um ponto interessante a se considerar, pois um alto número de falsos negativos significa um maior número de fraudadores que conseguem abrir contas.

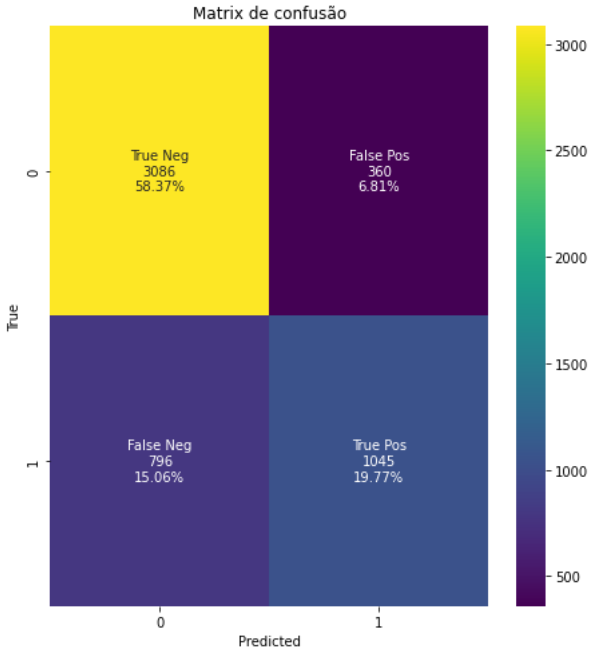

Figura 13. Matrix de Confusão

Fonte: Dados originais da pesquisa

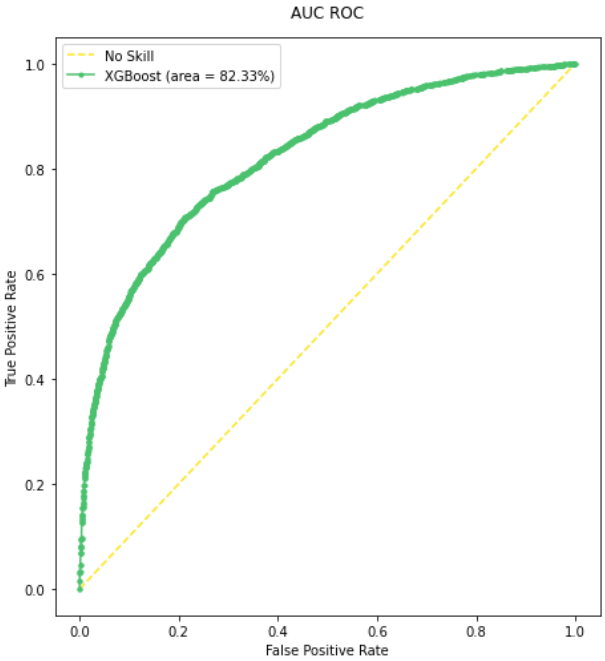

Figura 14. AUC ROC

Fonte: Dados originais da pesquisa

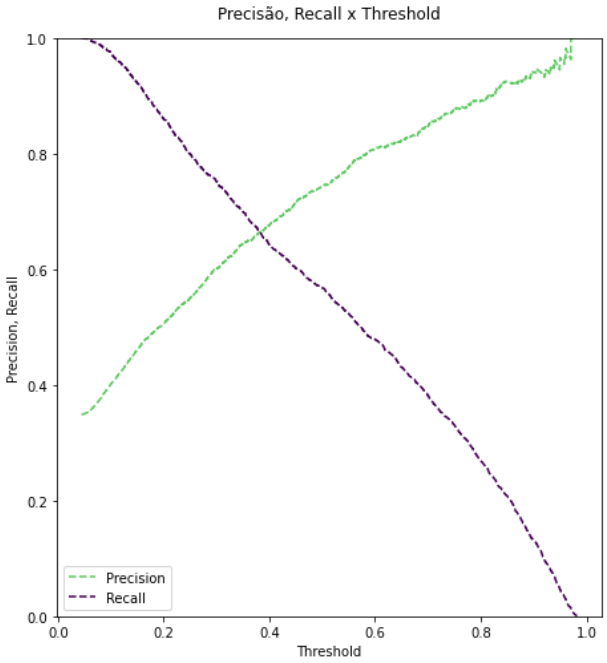

Figura 15. Precisão & “Recall” x “Threshold”
Fonte: Dados originais da pesquisa

GPBoost (Multinível)

O interesse por uma abordagem multinível para a construção deste modelo de “machine learning” se deu, inicialmente, pela percepção dos agrupamentos estruturais em decorrência da presença dos subemissores. Como apontam Fávero e Belfiore (2019), uma das principais vantagem destes modelos sobre os de regressão tradicionais, por exemplo OLS, está por considerar os agrupamentos naturais dos dados, ou seja, permite identificar e analisar as diferenças entre indivíduos e entre grupos, podendo definir componentes aleatórios em cada nível da análise. Desta maneira, espera-se levar para a construção do algoritmo uma representação mais fiel da realidade, em que os indivíduos não se comportam de maneira absolutamente independente, mas pertencem e compartilham de um grupo e suas características. Neste estudo de caso, dois níveis são considerados: o do processo de cadastro (indivíduo) e o do subemissor (grupo).

Na prática e de forma sucinta, o GPBoost é um algoritmo de reforço que aprende iterativamente os parâmetros de (co)variância e adiciona uma árvore ao conjunto de árvores usando um gradiente e/ou uma etapa de “Newton boosting”. A principal diferença para os algoritmos de boosting existentes é que: primeiro, ele considera a dependência entre os dados devido ao agrupamento e, segundo, aprende os parâmetros de (co)variância dos efeitos aleatórios (Sigrist, 2020).

Neste estudo, a engenharia de hiperparâmetros foi feita a partir de processos de “grid Search” e “cross-validation” e os resultados dos gráficos referem-se à amostra de validação. O código completo com a implementação do modelo pode ser consultado na página: https://github.com/samuel-haddad/MBAUSP/blob/main/gpboost_model.ipynb.

Por limitação na disponibilidade dos dados, foram consideradas na análise somente as variáveis preditoras dos módulos de avaliação de cada processo e a variável grupo. Porém, acredita-se que a inclusão de variáveis referentes aos subemissores, ou seja, que caracterizem os comportamentos compartilhados dos grupos, podem favorecer a captura dos efeitos aleatórios pelo modelo multinível e, consequentemente, sua performance.

A abordagem multinível apresentou as melhores métricas de desempenho. O modelo desenvolvido obteve os melhores resultados nos dois principais indicadores comparativos, AUC ROC e “log loss”, com maior área sob a curva e menor valor na respectiva função de custo. Estas métricas podem ser analisadas na Figura 16.

	amostra	acuracia	precisao	recall	auc_roc	f1_score	logloss(normalizada)
0	teste	76.30	71.11	55.02	80.27	62.04	0.505169
1	validacao	78.19	73.48	60.00	83.03	66.06	0.473676

Figura 16. Métricas de desempenho – GPBoost
Fonte: Dados originais da pesquisa

Ao observar as curvas de sensibilidade e especificidade em todos os modelos, nota-se que as probabilidades intermediárias, principalmente entre 20% e 80% apresentam-se verticalmente centralizadas nos gráficos. Devido às curvas não evoluírem tangenciando a borda superior da visualização, mas apresentarem um crescimento quase linear dos indicadores na faixa intermediária, fica evidente uma dificuldade maior dos modelos em explicar esta faixa de corte. Se por um lado isso é, em alguma medida, um comportamento esperado, por outro pode ser um bom objeto de estudo sobre pontos de corte para aprovação e reprovação do modelo.

As Figuras 17, 18 e 19 apresentam graficamente os resultados do modelo GPBoost.

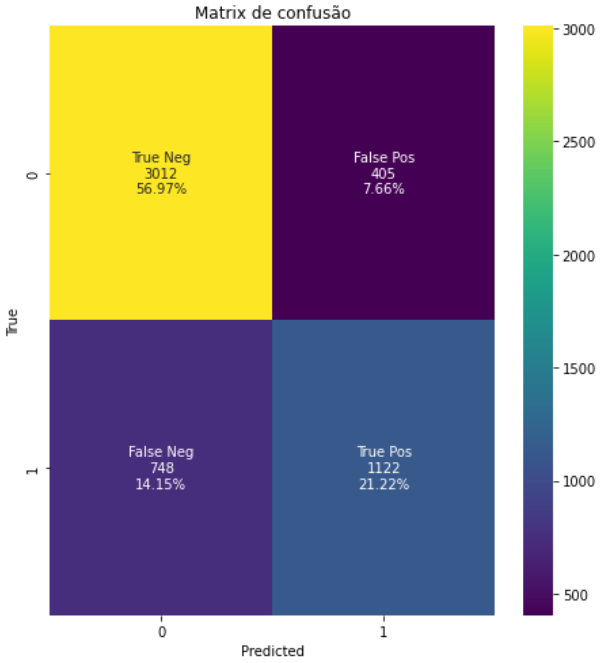

Figura 17. Matrix de Confusão

Fonte: Dados originais da pesquisa

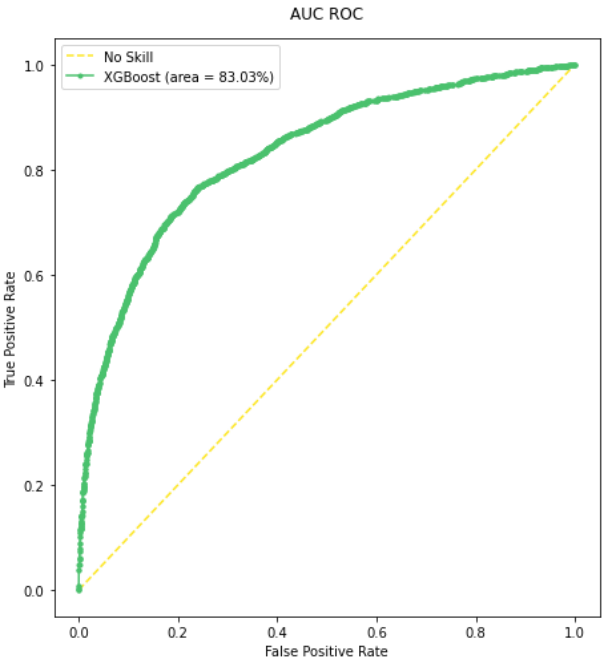

Figura 18. AUC ROC

Fonte: Dados originais da pesquisa

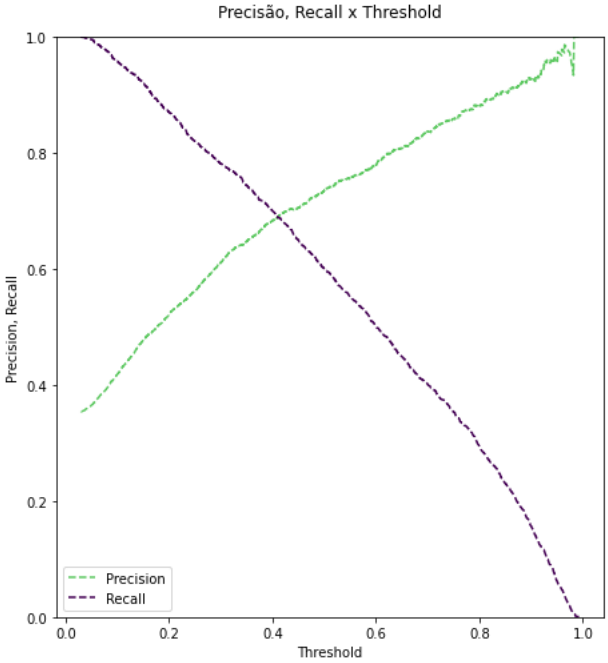

Figura 19. Precisão & “Recall” x “Threshold”

Fonte: Dados originais da pesquisa

Teste DeLong

Para comparar as áreas abaixo das curvas ROCs e verificar se os valores encontrados são estatisticamente significantes, foi utilizado o Teste DeLong. Este teste é uma abordagem não paramétrica, que permite a comparação das áreas, a partir da construção de uma matriz de covariâncias estimada usando a estatística U, ou estatística de Wilcoxon-Mann-Whitney (DeLong et al., 1988).

Duas a duas, as áreas foram comparadas utilizando a biblioteca Worc da Biomedical Imaging Group Rotterdam (2020). Como resultado, para um nível de confiança superior a 95%, pode-se afirmar que os valores encontrados para as AUC ROCs são estatisticamente diferentes entre si, portanto, com mesmo nível de confiança é possível afirmar que o modelo multinível se mostrou superior aos demais. Os resultados estão apresentados na Figura 20 e o código pode ser acessado em: https://github.com/samuel-haddad/MBAUSP/blob/main/delong_test.ipynb.

	logreg x xgboost	logreg x gpboost	gpboost x xgboost
p_value	0.000003	3.984302e-198	3.155797e-216

Figura 20. Teste DeLong

Fonte: Dados originais da pesquisa

Conclusões

Observa-se que os três modelos, Regressão Logística, XGBoost e Multinível, apresentaram métricas de desempenho relativamente próximas, tendo o modelo multinível, desenvolvido com o pacote GPBoost, a maior área sob a curva ROC (AUC ROC) na base de validação (83,03%). Dessa forma, mostrou-se possível o desenvolvimento de uma solução automatizada de aprendizado de máquina para a detecção de fraudes ideológicas, sendo a abordagem multinível aquela com maior capacidade preditiva em um contexto de agrupamento dos dados. Sugere-se ainda que um estudo de pontos de corte seja desenvolvido a fim de encontrar a maior adequação do modelo à realidade, tendo em vista que os erros (Falsos Positivos e Falsos Negativos) podem ter impactos diferentes no negócio e serem preferidos ou preteridos em alguma medida. Outro foco possível de evolução se dá na inclusão ou desenvolvimento de novas variáveis explicativas, referentes aos grupos, sob a hipótese de elas mostrarem-se relevantes nas caracterizações dos aninhamentos. Assim, acredita-se que este estudo cumpre seu papel em apresentar uma solução viável, capaz de contribuir com os objetivos da companhia. Traz ainda um olhar comparativo entre o uso de um modelo multinível e outros mais amplamente difundidos, utilizando diferentes técnicas de desenvolvimento, desde processos automatizados (AutoML), grades de otimização de hiperparâmetros (grid search) e técnicas de boosting, para a redução de viés e variância em modelos “ensemble”.

Agradecimento

À minha esposa, minhas filhas, meus irmãos, minha mãe e ao meu pai a gratidão de uma vida privilegiada de exemplos e amor. Aos amigos, agradeço por cada contribuição na construção da minha personalidade. Aos profissionais que fazem e fizeram parte da minha história, obrigado pela companhia em desafios e aprendizados. Ao meu orientador e aos mestres de uma vida, obrigado por me mostrarem um novo horizonte a cada contato. À Dock e aos meus colegas atuais, obrigado por me permitirem realizar este trabalho e tantos outros com prazer e mente aberta.

Referencias

Anderson, J. A. 1982. Handbook of Statistics. Elsevier, Volume 2: 169-191.

Banco Central do Brasil. 2022. Disponível em <https://www.bcb.gov.br/estatisticas/spbadendos >. Acesso em: 08 dezembro 2022.

DeLong E. R.; DeLong D. M.; Clarke-Pearson D. L. 1988. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach, International Biometric Society, Volume 44: 837-845.

Fávero, L. P.; Belfiore, P. 2019. Data science for business and decision making. 1ed. Cambridge: Academic Press. Cambridge, MA, USA.

Distributed (Deep) Machine Learning Community. Disponível em <https://github.com/dmlc/xgboost/tree/master/demo#machine-learning-challenge-winning-solutions>. Acesso em 12 dezembro 2022.

Mastercard. 2021. Mastercard New Payments Index: Consumer Appetite for Digital Payments Takes Off. Disponível em <https://investor.mastercard.com/investor-news/investor-news-details/2021/Mastercard-New-Payments-Index-Consumer-Appetite-for-Digital-Payments-Takes-Off/default.aspx>. Acesso em 22 novembro 2021.

Shearer C., 2000. The CRISP-DM model: the new blueprint for data mining, Journal of Data Warehousing, Volume 5: 13-22.

Sigrist F., 2020 Gaussian Process Boosting, arXiv:2004.02653,

Sigrist F., 2021 Latent Gaussian Model Boosting, arXiv:2105.08966

Sigrist F. 2020. Tree-Boosted Mixed Effects Models. Disponível em <https://towardsdatascience.com/tree-boosted-mixed-effects-models-4df610b624cb>. Acesso em 10 janeiro 2022.

XGBoost developers. 2022. XGBoost Documentation. Disponível em <https://xgboost.readthedocs.io/en/stable/>. Acesso em 06 junho 2022.

Wikipedia. 2022. XGBoost. Disponível em <https://en.wikipedia.org/wiki/XGBoost >. Acesso em 06 junho 2022.

Biomedical Imaging Group Rotterdam. 2020. WORC: Workflow for Optimal Radiomics Classification Documentation. Disponível em <https://worc.readthedocs.io/en/latest/autogen/WORC.statistics.html#module-WORC.statistics.delong t>. Acesso em 13 dezembro 2022.