MESTRADO PROFISSIONAL
EM ECONOMIA

O USO DA APRENDIZAGEM DE MÁQUINA PARA A CRIAÇÃO DE MODELOS PREDITIVOS DE EVASÃO DE COTISTAS NO MERCADO DE FUNDOS DE INVESTIMENTOS

ERIC CARVALHAL XAVIER

Brasília-DF, 2021

ERIC CARVALHAL XAVIER

O USO DA APRENDIZAGEM DE MÁQUINA PARA A CRIAÇÃO DE MODELOS PREDITIVOS DE EVASÃO DE COTISTAS NO MERCADO DE FUNDOS DE INVESTIMENTOS

Dissertação apresentada ao programa do Mestrado Profissional em Economia Aplicada do Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa, como parte dos requisitos para a obtenção do Título de Mestre em Economia.

Orientador
Professor Dr. Alexandre Xavier Ywata de Carvalho

Brasília-DF 2021

ERIC CARVALHAL XAVIER

O USO DA APRENDIZAGEM DE MÁQUINA PARA A CRIAÇÃO DE MODELOS PREDITIVOS DE EVASÃO DE COTISTAS NO MERCADO DE FUNDOS DE INVESTIMENTOS

Dissertação apresentada ao programa do Mestrado Profissional em Economia Aplicada do Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa, como parte dos requisitos para a obtenção do Título de Mestre em Economia.

Aprovada em 22 / 12 / 2021

Banca Examinadora

Prof. Dr. Alexandre Xavier Ywata de Carvalho

Prof. Msc. Mathias Schneid Tessmann

Prof. Dr. Adonias Evaristo da Costa Filho

Prof. Dr. Daniel Oliveira Cajueiro

X3u Xavier, Eric Carvalhal
O uso da aprendizagem de máquina para a criação de modelos preditivos de evasão de cotistas no mercado de fundos de investimentos / Eric Carvalhal Xavier. – Brasília: IDP, 2021.

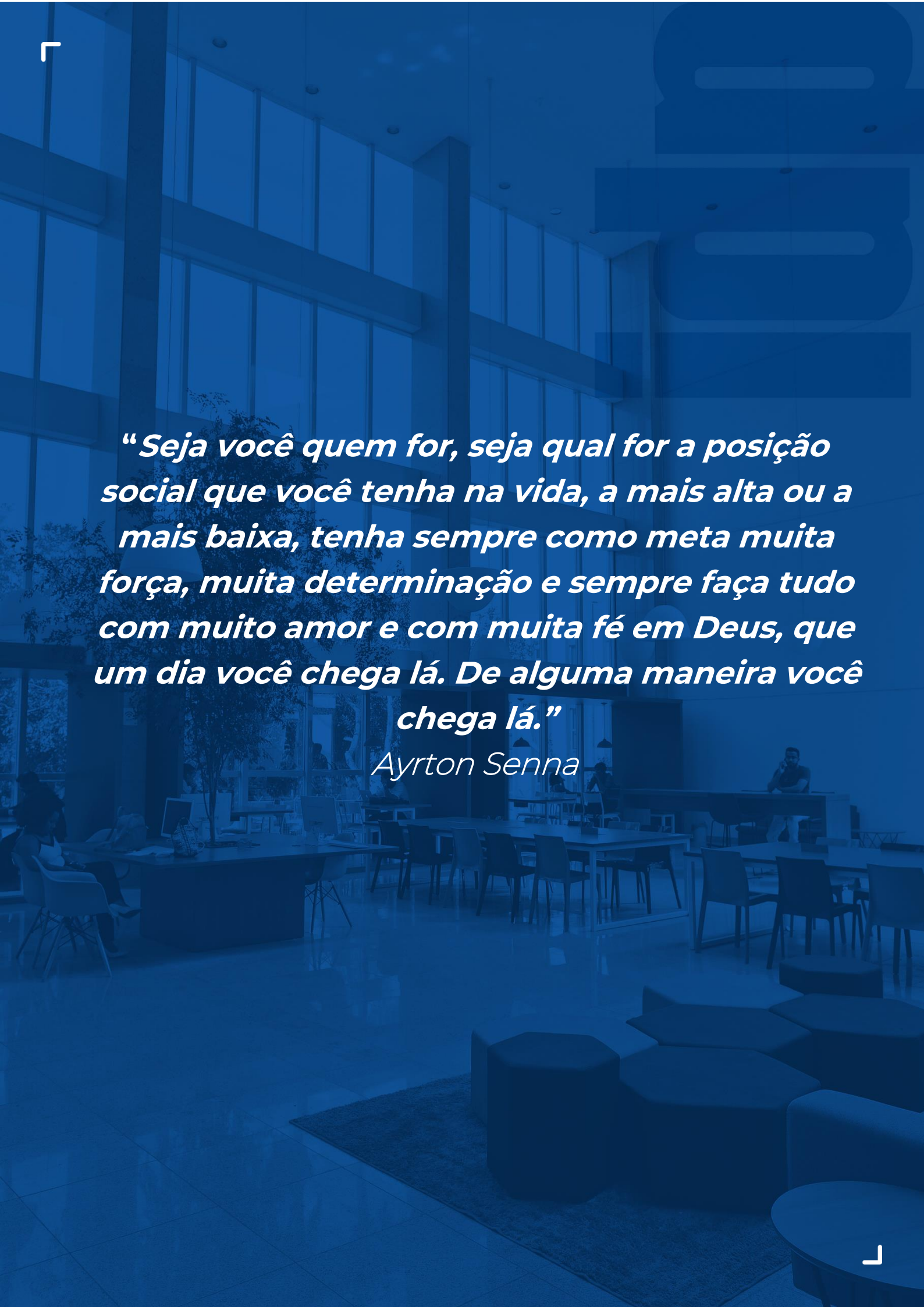
59 p.: il. Color.
Inclui bibliografia.

Trabalho de Conclusão de Curso (Dissertação) – Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa – IDP, Mestrado Profissional em Economia, Brasília, 2021.
Orientador: Prof. Dr. Alexandre Xavier Ywata de Carvalho.

1. Fundos de investimento. 2. Aprendizagem de máquina. 3. Evasão de clientes. 4. XGBoost. 5. Relacionamento com clientes. I. Título.

CDD: 341.6251

Ficha catalográfica elaborada pela Biblioteca Ministro Moreira Alves
Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa



“Seja você quem for, seja qual for a posição social que você tenha na vida, a mais alta ou a mais baixa, tenha sempre como meta muita força, muita determinação e sempre faça tudo com muito amor e com muita fé em Deus, que um dia você chega lá. De alguma maneira você chega lá.”

Ayrton Senna

AGRADECIMENTOS

O processo da busca pelo conhecimento requer disciplina e investimento de tempo. Agradeço a Deus por ter permitido que um sonho se tornasse realidade em minha vida. Para me tornar Mestre em Economia, Ele me capacitou, me orientou, me sustentou e me fez chegar até aqui. Obrigado, Senhor!

O tempo é a maior riqueza que temos, sem sabermos exatamente a medida total que nos resta. Com este tempo, que a cada segundo se esgota, nós seguimos o curso da vida com as nossas escolhas. Nós o usamos no trabalho, com a família, com os amigos, nos estudos, no lazer etc. Quero agradecer imensamente a minha esposa Rosany, ao meu filho Pedro e a minha mãe Mara, por terem compreendido o investimento que eu fiz do meu tempo neste projeto, em detrimento de investir com eles. Foram muitos dias e muitas noites e finais de semana dedicados a este projeto. Essa vitória é de vocês e para vocês, afinal, o que há de mais importante para mim são vocês. Espero que a realização deste sonho torne o desfrute do nosso futuro mais prazeroso e confortável! Muito obrigado! Amo vocês!

Por fim, agradeço ao Prof. Dr. Alexandre Xavier Ywata, brilhante professor e magnífico líder, pela orientação e oportunidade, ao Prof. Me. Mathias Tessmann, excelente professor, pela parceria de sempre, ao Luis Enciso, brilhante gestor de soluções de tecnologia, pela motivação e apoio, e ao Eduardo Jansen, formidável cientista de dados, por todo o compartilhamento do seu conhecimento e tempo. Sem os quais não seria possível a realização do trabalho. Muito obrigado, senhores!

RESUMO

O objetivo desse trabalho foi identificar a existência de variáveis comportamentais de cotistas de fundos de investimento que pudessem ajudar a explicar o encerramento de relacionamento com o banco e buscar modelos com a melhor performance preditora de evasão. A metodologia utilizou a pesquisa aplicada para gerar conhecimentos de aplicação prática com a abordagem quantitativa. O procedimento adotado foi por meio da pesquisa empírica a base de dados de uma instituição financeira do Brasil. Para garantir o atendimento à Lei 13.709 de 14 de agosto de 2018, todas as informações pessoais foram descaracterizadas e não disponibilizadas durante o processo de pesquisa, análise de dados e desenvolvimento dos modelos preditivos para a aprendizagem de máquinas. A população totalizou 2.907.270 clientes pessoas físicas dos segmentos de média a alta renda e amostra balanceada com a técnica *under sampling* de 200.000 clientes. As principais variáveis identificadas como importantes na predição da evasão foram a posse de produtos, saldo, perfil digital, dentre outras. Foi realizada a comparação de performance em relação a curva ROC para os modelos Regressão Logística, Árvore de Decisão, KNN, *Naive Bayes*, *Randon Forest*, *XGBoost*, LDA, MLP e SVC, e os resultados apontaram o *XGBoost* como o modelo de melhor performance preditiva com curva ROC de 0,7861 e matriz de confusão com acurácia de 71,16.

Palavras-chaves: Fundos de Investimento; Aprendizagem de Máquina; Evasão de Clientes; XGBoost; Relacionamento com Clientes.

ABSTRACT

The objective of this work was to identify the existence of behavioral variables of investment fund shareholders that could help explain the termination of the relationship with the bank and seek models with the best performance predictor of dropout. The methodology used applied research to generate knowledge of practical application with the quantitative approach. The procedure adopted was through empirical research on the database of a financial institution in Brazil. To ensure compliance with Law 13,709 of August 14, 2018, all personal information was de-characterized and not made available during the research process, data analysis and development of predictive models for machine learning. The population totaled 2,907,270 individual customers from the middle to high income segments and a balanced sample with the under sampling technique of 200,000 customers. The main variables identified as important in predicting dropout were product ownership, balance, digital profile, among others. A comparison of performance was performed in relation to the ROC curve for the Logistic Regression, Decision Tree, KNN, Naive Bayes, Random Forest, XGBoost, LDA, MLP and SVC models, and the results pointed to XGBoost as the model with the best predictive performance. with a ROC curve of 0.7861 and a confusion matrix with an accuracy of 71.16.

Keywords: Investment Funds; Machine Learning; Customer Evasion; XGBoost; Relationship with Customers..

LISTA DE QUADROS

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS E SIGLAS

ANBIMA	Associação Brasileira das Entidades dos Mercados Financeiro e de Capitais
AUC	Área embaixo da Curva (*Area Under the Curve*)
BACEN	Banco Central do Brasil
CVM	Comissão de Valores Mobiliários
ETF	*Exchange-Traded Fund*
FIP	Fundo de Investimento em Participações
FII	Fundo de Investimento Imobiliário
FIDC	Fundo de Investimento de Direito Creditórios
IDP	Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa
LGPD	Lei Geral de Proteção de Dados Pessoais

SUMÁRIO


1

1 INTRODUÇÃO

O presente estudo busca analisar o comportamento dos cotistas no mercado de fundos de investimento, para que seja possível identificar variáveis explicativas para a evasão. Desta forma, será possível identificar, previamente, os cotistas que encerrarão o relacionamento, e fomentar a sua fidelização por meio da oferta de produtos e serviços adequados.

Trata-se de trabalho de caráter inovador, pois não foi identificado outra pesquisa acadêmica sobre o mercado de fundos de investimentos com aplicação dos modelos de aprendizagem de máquina em base de dados real.

De Acordo com a base de dados histórica da ANBIMA disponível no Boletim de Fundos de Investimento[1], é possível observar que no período de dezembro de 2010 a dezembro de 2020, o patrimônio líquido total dos fundos de investimento das classes Renda Fixa, Ações, Multimercados, Cambial, Previdência, ETF, FIDC, FIP, FII, Off Shore, cresceu 262% no período, saindo de R$ 1,6 trilhão para R$ 6 trilhões.

No ano de 2020 as economias globais foram fortemente impactadas pelas ações de prevenção à pandemia. O desemprego e a queda das vendas e dos serviços foram as principais consequências que a população mundial teve de enfrentar em razão da pandemia da COVID-19.

Conforme dados da ANBIMA[2], no mês de abril de 2020, a indústria de fundos de investimento contabilizou a saída de recursos na ordem de R$ 91,1 bilhões, sendo o maior resgate mensal desde 2002 e o pior mês do ano em relação a captação líquida – R$ 80,7 bilhões negativos.

---

[1] Fonte: https://www.anbima.com.br/pt_br/informar/relatorios/fundos-de-investimento/boletim-de-fundos-de-investimentos/boletim-de-fundos-de-investimentos.htm

[2] Fonte: https://www.anbima.com.br/pt_br/informar/relatorios/fundos-de-investimento/boletim-de-fundos-de-investimentos/industria-de-fundos-de-investimento-sinaliza-resiliencia-em-um-ambiente-de-incerteza.htm

Na contramão dos principais indicadores econômicos que mostraram uma queda geral das atividades econômicas, o mercado de fundos de investimento registrou captação líquida positiva no valor de R$ 174,5 bilhões no consolidado do ano de 2020.

Considerando que o mercado de fundos de investimento mostrou resiliência em 2020, apresentando captação líquida positiva, mesmo em um contexto de pandemia e de crise de produção devido ao *lockdown*[3], este trabalho possui o objetivo de verificar a movimentação dos cotistas no período de dezembro de 2018 a dezembro de 2020, com o foco de analisar possíveis correlações entre resgates totais para verificar se esses movimentos nos saldos podem ser explicados estatisticamente por algumas variáveis comportamentais pertencente aos cotistas.

De acordo BACEN[4] o *Open Banking*, ou Sistema Financeiro Aberto, é o compartilhamento padronizado de dados, produtos e serviços por meio de abertura e integração de sistemas, com o uso de interface dedicada para essa finalidade, por instituições financeiras, instituições de pagamento e demais instituições autorizadas a funcionar. O BACEN assegura que a implementação será de forma segura, ágil e conveniente, com previsão dos dados transacionais referentes aos clientes estarem disponíveis ao mercado em 31/05/2022.

Em relação ao atendimento no que versa na Lei 13.709 de 14 de agosto 2018, que versa sobre a Proteção de dados, o compartilhamento dos dados pessoais de clientes ou de serviços do escopo do *Open Banking* depende de prévio consentimento por parte dos respectivos clientes se caracterizando por meio de manifestação livre, informada, prévia e inequívoca de vontade, feita por meio eletrônico, pela qual o cliente concorda com o compartilhamento de dados ou de serviços para finalidades determinadas.

O ponto de ligação com este trabalho, é que a competitividade provocada pelo *Open Banking* incentivará a inovação e o surgimento de novos modelos de negócio centrados em uma nova experiência para

---

[3] Conforme decretos dos poderes executivo federal, estadual e municipal, determinadas atividades econômicas tais como bares e restaurantes foram fechadas para conter a disseminação da pandemia.

[4] Consulta realizada em 27/03/2020 no site https://www.bcb.gov.br/acessoinformacao/perguntasfrequentes-respostas/openbanking

o cliente com segurança, agilidade e conveniência, favorecendo a inclusão financeira e a educação financeira.

O escopo do estudo foi na estratificação da base de dados de clientes do segmento de pessoa física de média e alta renda, no período de dezembro de 2018 a dezembro de 2020, para identificação de comportamentos destes clientes que pudessem evidenciar alguma relação de evasão. Foi realizada a avaliação de alguns métodos de modelos preditivos de aprendizagem de máquina para observar aquele ou aqueles que pudessem ser escolhidos devido a sua melhor performance preditiva. Por fim, foi observado que o método com melhor performance foi o XGB *Extreme Gradient Boosting* para a explanação dos resultados e conclusões.

Além dessa breve introdução, o trabalho é composto por mais quatro seções, sendo a seção dois que contém uma revisão de literatura acerca dos temas para introduzir e contextualizar os conceitos de fundos de investimentos, de gestão do relacionamento com clientes, de *Machine Learning* (aprendizagem de máquina), de *Churn* (evasão) e de trabalhos e pesquisas existentes que são correlatos ao tema. A seção três explicita a metodologia por meio da análise exploratória prévia da base dados, e com a exposição do resultado da avaliação dos modelos preditivos e as técnicas quantitativas utilizadas e a seção quatro apresenta os resultados com a evidência da matriz de confusão e a curva ROC. Por fim, a seção cinco traz a conclusão com a indicação da aplicação do algoritmo.


2

2 REVISÃO DA LITERATURA

Conforme informações disponibilizadas pela Comissão de Valores Mobiliários - CVM[5] os fundos de investimento são modalidades de investimentos coletivos. É uma estrutura formal que reúne recursos financeiros de diversos cotistas (investidores), para investimento conjunto e desta forma conseguirem vantagem competitiva. Esta modalidade de aplicação financeira é criada por um administrador, geralmente uma instituição financeira, que formalmente o constitui e define os seus objetivos e políticas de investimento.

Os investidores interessados aportam seus recursos nesses condomínios e conforme as disposições constantes no prospecto e na lâmina de informações, o gestor do fundo poderá investir em diversos ativos financeiros: CDB, ações, debêntures, moedas estrangeiras, investimentos estrangeiros, derivativos, entre outros.

De acordo com a CVM, se fosse permitido a todos os fundos investirem em qualquer um dos ativos citados no parágrafo anterior, seria difícil para os investidores que não possuem vasta experiência e conhecimento, compreenderem minimamente os riscos assumidos no momento do investimento. Por essa razão, foram criadas diferentes classes de fundos, considerando os tipos de ativos que compõem as carteiras e, consequentemente, os fatores de risco. Dentre as principais classes, temos Renda Fixa, Ações, Multimercados e Cambial.

Durante a etapa do levantamento bibliográfico, quando pesquisado o tema de diversificação no mercado de fundos, foi identificada uma vasta amostra de trabalhos e artigos com mergulho nos princípios práticos da Teoria de Portfólios por Markovitz (1952) para abordagens práticas e teóricas para explicar o risco de determinadas carteiras. Contudo, muito embora a Teoria seja relevante ao tema, o objetivo não é verificar os riscos das classes, mas se dedicar à análise da

[5] Consulta realizada em 28/03/2020 no site: https://www.investidor.gov.br/portaldoinvestidor/export/sites/portaldoinvestidor/publicacao/Cadernos/CVM-Caderno-3.pdf

relação do movimento de evasão de clientes em relação a quantidade de classes de fundos diferentes em seus portfólios.

O trabalho de Vidigal (2012) que versa sobre a "Diversificação de um portfólio de Fundos Multimercado: Uma análise empírica" concluiu, dentre outros aspectos, que quanto maior for o número de ativos em uma carteira, o risco desta carteira tende a diminuir. Da forma semelhante, pode ser possível verificar se o efeito da diversificação possui relação direta com a (i) manutenção, (ii) permanência e (iii) evolução das aplicações dos clientes no banco que possuírem a diversificação dentre as classes existentes.

De acordo com o trabalho de Veiga, Gibran e Bonsere (2019) que versa sobre "*Open Banking*: Expectativas e Desafios para o Mercado Financeiro no Brasil" o *Open Banking* traz um novo cenário para o sistema financeiro nacional como sendo de inclusão e competitividade com consumidor brasileiro aderente à realização de transações digitais.

É necessário trazer a definição adotada pelo Banco Central sobre o *Open Banking*:

> "O Open Banking, ou sistema financeiro aberto, é a possibilidade de clientes de produtos e serviços financeiros permitirem o compartilhamento de suas informações entre diferentes instituições autorizadas pelo Banco Central e a movimentação de suas contas bancárias a partir de diferentes plataformas e não apenas pelo aplicativo ou site do banco, de forma segura, ágil e conveniente. (BRASIL. BACEN..., 2021)."

O artigo de Guimarães (2021) aborda a concorrência bancária e o *open banking* no Brasil. De acordo com a pesquisa *Fintech Deep Dive* realizada em 2020 pela PwC6 em parceria com a ABFintechs, foi traçado um perfil do ecossistema de inovação e empreendedorismo no segmento de serviços financeiros do país com base em informações fornecidas por 148 fintechs de diferentes setores de atuação. Foi possível observar que o *Open Banking*, em conjunto com os novos meios de pagamento, como o Pix, gerará um ambiente propício para inovações. A pesquisa mostrou que das 148 empresas fintechs, 73% afirmaram que desenvolvem soluções para Pix e/ou *Open Banking* e

6 "PwC" refere-se à PricewaterhouseCoopers Brasil Ltda., firma membro da network da PricewaterhouseCoopers.

76% delas esperam colher benefícios das duas iniciativas logo no primeiro ano.

Conforme dados disponíveis no caderno do CADE - Conselho Administrativo de Defesa Econômica, sobre Mercado de Instrumentos de Pagamento publicado em outubro de 2019, Guimarães (2021) destacou que a partir da posse dos dados bancários, restam poucas barreiras para impedir a competição entre os gigantes da tecnologia e as instituições financeiras. O ponto de atenção para os novos entrantes seriam os custos irrecuperáveis relativos a investimentos iniciais em marketing, tecnologia e formação de uma rede de distribuição, além da incerteza sobre o peso da fidelidade dos clientes às marcas estabelecidas.

Em relação ao cenário internacional, o trabalho de Perez e Strohl (2019) mostra que o *Open Banking* já é uma prática adotada na Europa, sendo precursora na regulamentação deste assunto com a finalidade de trazer inovação por meio da eficiência no mercado financeiro, resultando um ambiente mais competitivo com a manutenção dos direitos dos consumidores.

O artigo de Brodsky e Oakes (2017) traz a importância do cuidado redobrado pelas instituições que possuem a maior quantidade de clientes pois o *Open Banking* visa beneficiar os usuários, gerar inovação e concorrência, estimulando o reposicionamento das instituições financeiras frente às novas tendências.

Tendo em vista os desafios do *Open Banking*, os métodos quantitavos são extremamente importantes para entender as relações que existem entre as variáveis que compõem o mundo das finanças, assim, segundo Gujarati (2011), a econometria pode ser compreendida literalmente como medição econômica. A econometria consiste na aplicação da estatística matemática a dados econômicos para dar suporte empírico aos modelos formulados pela economia matemática e obter resultados numéricos (TINTNER, 1968).

A teoria econômica faz declarações ou hipóteses principalmente de natureza qualitativa. A teoria microeconômica, por exemplo, possui como premissa que, tudo o mais constante, quando há uma redução no preço de um bem, deve resultar no aumento da quantidade demandada por este bem. Desta forma, a teoria econômica pressupõe

uma relação negativa ou inversa entre o preço e a quantidade demandada de um determinado bem. Contudo, a teoria em si não oferece nenhuma medida quantitativa da relação entre as duas variáveis (GUJARATI, 2011).

No trabalho de econometria, o executor da modelagem pode se deparar com dados provenientes de observações em contraste aos dados experimentais. Existem duas implicações relevantes para a modelagem empírica na econometria. A primeira diz que, quem modela deve dominar habilidades muito diferentes das necessárias à análise de dados experimentais. A segunda mostra a diferença entre quem coleta dados e quem os analisa, visto que o responsável pela modelagem deve estar bastante familiarizado com a natureza e a estrutura dos dados em questão. (SPANOS, 1999).

Por fim, Gujarati (2011) traz que, embora existam uma ampla diversidade de escolas de pensamento sobre a metodologia econométrica, a metodologia econométrica tradicional segue os seguintes passos: (i) Exposição da teoria ou hipótese; (ii) Especificação do modelo matemático da teoria; (iii) Especificação do modelo estatístico ou econométrico; (iv) Obtenção dos dados; (v) Estimação dos parâmetros do modelo econométrico; (vi) Testes de Hipóteses; (vii) Projeção ou previsão; (viii) Uso do modelo para fins de controle ou política.

Nos dias atuais observamos a aprendizagem de máquina, também conhecida no termo em inglês como *machine learning*, na vida das pessoas todos os dias, direta ou indiretamente. Quando acessamos as redes sociais, quando fazemos pesquisas na internet, quando utilizamos aplicativos, ou seja, em todas essas interações é possível desenvolver um padrão de estilo de vida e consumo para que haja o registro das preferências deste usuário.

De acordo com Borges (2020) os algoritmos de *Machine Learning* podem ser aplicados para resolução de diversos tipos de problemas, ou seja, são usados tanto para predizer um indicador num setor industrial quanto dentro da área da saúde, contribuindo assim para a eficácia dos processos, diagnósticos, tomadas de decisão entre outros.

Com a evolução da tecnologia e o uso de smartphones, o ser humano acabou por se tornar um grande fornecedor e consumidor de

dados. Uma das consequências deste processo é na evolução da experiência do usuário, como por exemplo, ao procurar por um tênis para correr em um site de pesquisa, e logo depois propagandas com modelos de tênis específicos começam a aparecer espontaneamente em aplicativos específicos ou em redes sociais. Esse encontro da necessidade do usuário e com a oferta de produtos ocorre devido ao registro do comportamento do usuário ao realizar suas pesquisas ou interações, contudo, esse comportamento pode evoluir de forma rápida e novas previsões em temas distintos podem ser criadas. Esse é exatamente o foco do processo de aprendizagem de máquinas, coletar e usar os padrões mais atuais dos usuários para entender o processo completo ou usar esses padrões para fazer previsões de um futuro próximo.

A dissertação de Borges (2020)[7] cita que, conforme Mohri, Rostamizadeh e Talwalkar (2018), o *Machine Learning* pode ser definido como um método computacional para predição de dados, sendo que a partir de uma base de dados, como por exemplo, uma série temporal que apresenta os dados ao longo do tempo, a "máquina" aprende sobre os padrões de comportamento destes dados através de algoritmos com o objetivo de realizar predições para um resultado futuro.

Ao se questionar sobre a importância do foco no cliente, Gupta e Lehmann (2005) evidenciam a necessidade da empresa se posicionar ao passo que suas ações estejam voltadas para os clientes. Estes precisam ser vistos como o principal ativo das empresas, e, por esta razão, todas as decisões estratégicas devem ser tomadas em convergência com a visão do cliente, mantendo o objetivo de identificar e dimensionar seus possíveis resultados.

É possível depreender no trabalho de Borges (2020) que, de acordo com Moschovakis (2001), um algoritmo é geralmente definido como uma série de modelos matemáticos computadorizados que por meio de um *input* inicial consegue realizar inúmeras operações com o objetivo de devolver um resultado ou um comportamento futuro, ou a solução de um problema. Sendo que esses resultados são definidos como *output*.

[7] Dissertação de Mirele Marques Borges com o tema: Machine Learning como ferramenta gerencial para predição de indicadores e detecção de anomalias. Submetida ao programa de Pós-Graduação Mestrado Profissional em Engenharia de Produção da Universidade Federal do Rio Grande do Sul em 2020.

No que tange ao ambiente de gerenciamento das informações e seus algoritmos, temos a contribuição de Payne (2006), que descreve o CRM - *Customer Relationship Management*[8], como um processo que auxilia de forma sólida a execução e manutenção da estratégia corporativa, visto que as informações disponíveis são apresentadas de maneira mais tempestiva do que indicadores financeiros tradicionais. Os motivos principais para a utilização e implantação de sistemas de CRM podem ser percebidos na melhora do relacionamento com os clientes, na minimização dos custos, na maximização da eficiência e produtividade. Os sistemas de CRM inverteram o eixo decisório das organizações para o foco no reconhecimento do protagonismo do cliente deixando de lado as tomadas de decisão com foco no produto.

É importante citar que dentre as estratégias utilizadas pelas empresas que atuam com foco no cliente destaca-se o Gerenciamento de Relacionamento com o Cliente. Esta estratégia não se limita apenas a um software para o gerenciamento de informações armazenadas em bancos de dados, mas, principalmente, na visão voltada para o cliente em detrimento do foco no produto. Dessa maneira a busca passa a ser no desenvolvimento do produto certo para o cliente.

De acordo com Brown (2001) o CRM é uma ferramenta de marketing que consiste na integração no contato com o cliente (*front office*) com as operações de retaguarda (*back office*), onde há uma máxima atenção voltada ao cliente favorecendo a aprendizagem aprender a cada interação.

É possível compreender com Tronchin (2002) que a meta final de um modelo de negócios empresarial que seja baseado em Gestão de Relacionamento com Clientes, é o crescimento da lucratividade da empresa por meio da retenção de seus melhores clientes. Essa retenção ou fidelização é alcançada pela gestão das informações comportamentais individuais.

Podemos identificar o sucesso de um sistema de gestão de relacionamento com cliente por meio da capacidade de encantar os clientes e de identificar aqueles que estão dispostos a encerrar o relacionamento. Nesse sentido o conceito de *Churn* conforme

[8] CRM - *Customer Relationship Management*, conhecido em português como Gerenciamento do relacionamento com o cliente.

Ghorbani e Taghiyareh (2009) é a característica dos clientes que têm a intenção de sair e se tornar cliente de uma empresa concorrente.

Os autores Nitzan e Libai (2011) enfatizam que devido às consequências que levam a uma melhor compreensão e previsão do *churn*, existe a necessidade de um CRM efetivo e sistêmico. De acordo com Veloso (2013) existe um desafio que cresce em ritmo acelerado nas organizações que é a questão do *churn* que pode representar significativo impacto nas receitas das organizações. É possível observar a atenção empregada pelos especialistas sobre a retenção de clientes, principalmente, conforme preconiza Nitzan; Libai (2011), em relação ao impacto da retenção no valor da vida útil do cliente e no resultado da organização.

O *churn* é relacionado à estratégia de marketing que identifica que um consumidor está migrando de uma empresa para outra. Na posição atual de cliente, ainda existe relacionamento, contudo a sua saída para o concorrente poderá ocorrer num futuro próximo conforme preconiza Glady, Baesens e Croux (2009).

Caso a organização queira impedir a evasão, é necessária uma ação de retenção que, de acordo com Ghorbani e Taghiyareh (2009) é necessária a gestão do *churn*, consistindo na prévia identificação dos clientes classificados em *churn* com a execução de ações de fidelização e marketing para mantê-los na condição de clientes. Por fim, os autores evidenciam um dos primeiros passos para a gestão do *churn* que consiste em investigar as suas causas.

Na etapa de revisão de literatura para identificar trabalhos semelhantes no Brasil, não foi possível identificar a existência de pesquisa sobre o tema de evasão de clientes (*Churn*), exclusivamente, no mercado de fundos de investimentos com a utilização de modelos preditivos com aprendizagem de máquinas, contudo foi possível identificar trabalhos em setores distintos com foco em modelos preditores.

A dissertação de Gauer (2016) buscou responder afirmativamente à questão de sua pesquisa sobre a possibilidade das variáveis relacionadas à saldo de operações bancárias serem preditoras de clientes evasores. Neste caso o modelo transformou a variável Média de Saldo Diários em uma escala logarítmica, a fim de avaliar a oscilação de

acordo com os períodos. Não houve a plotagem de indicador de performance, tal como a curva ROC por exemplo. O autor recomendou a aplicação deste modelo como um complemento aos indicadores já existentes na gestão de fidelização dos clientes.

Foi possível identificar na dissertação de Gauer (2016) a citação de outros trabalhos com foco em modelos preditivos com destaque para o trabalho de Botelho e Tostes (2010), com foco no mercado de cartão de crédito, e o trabalho de Pinho (2009), orientado para uma administradora de investimentos que disponibiliza a diversificação de investimentos por meio da compra e venda de ações, cotas em fundos de investimentos, derivativos e títulos do tesouro direto.

O trabalho de Pinho (2009) utilizou Algoritmos Genéticos como um diferencial competitivo na explicitação dos clientes a evadirem da administradora de investimentos. De acordo com o autor, os Algoritmos Genéticos foram desenvolvidos por John Holland na década de 60 e 70 e fornecem um mecanismo de busca adaptativa. De acordo com Larose (2006) quando da utilização dos algoritmos genéticos, a aptidão de um universo variado de potenciais soluções são testadas e comparadas, e as soluções consideradas com maior aptidão terão mais chances de cruzarem entre si as informações importantes para o caso em questão, criando soluções ainda mais propensas. Desta maneira, o autor da pesquisa conclui que:

> "A evolução de regras por algoritmos genéticos resultou no encontro de regras com alta acurácia e abrangência na solução do problema de evasão de clientes. Tais regras podem ser facilmente implementadas em sistemas inteligentes, bem como interpretadas por usuários envolvidos no processo de tomada de decisão de marketing. Uma vez objetivado reduzir as taxas de evasão de clientes, e aumentar a lucratividade futura da empresa pela permanência prolongada do status ativo do cliente, poderiam trabalhar ações específicas de marketing aos clientes classificados como futuros evasivos, reduzindo-se custos de alocação de marketing. Há de se reconhecer que a utilização de outras técnicas de inteligência computacional, ou estatística, poderiam produzir resultados melhores em termos de acurácia e abrangência, ficando este aspecto a ser investigado em passos futuros. (PINHO, 2009)"

O artigo de Botelho e Tostes (2010) buscou a modelagem da probabilidade de clientes encerrarem o relacionamento com uma empresa de cartões de crédito. Foi utilizado dados do histórico de

relacionamento para a descrição das possíveis variáveis que influenciam o abandono ou a permanência do cliente. O modelo de regressão logística foi utilizado em uma amostra de calibração de 70.000 clientes que possuíam cartão de crédito. De acordo com o autor *"o modelo foi validado em uma amostra de 30.000 clientes, usando-se o teste de KS [9](Kolmogorov-Smirnov) e a curva ROC [10](Receiver Operating Characteristic), que demonstraram a boa adequação do modelo à amostra de validação."*

O trabalho de pesquisa de Oliveira (2021) buscou desenvolver o aprendizado de máquinas para prever a evasão de clientes de um jornal de grande circulação regional. Dentre os modelos testados, o trabalho se propôs a testar uma metodologia que utiliza *Auto Machine Learning* com aprendizado de dados em *stream*, conforme o autor, essa metodologia é capaz de mitigar eventuais desvios de conceito que possam surgir nos modelos implementados em ambiente de produção. Os resultados apresentados foram satisfatórios, sendo o modelo Arvore de Decisão*: Extremely Fast* com a melhor performance apresentando precisão de 96,84 e AUC [11]de 0,9586.

Foi possível observar no artigo de Souza, Neto, Junior e Souki (2009) que a estratégia de incrementar o relacionamento com o cliente tende a aumentar a satisfação do cliente. A base teórica do estudo abordou indicadores de satisfação, lealdade, mediadores da mudança, custos de mudança e retenção de clientes. Por meio da utilização do método da regressão logística, os pesquisadores identificaram que as variáveis idade, contratos ativos de longo prazo, quantidade de

---

[9] O teste de Kolmogorov-Smirnov (KS) é uma estatística não paramétrica para testar se as funções de distribuição de probabilidades de dois grupos São iguais. O valor do KS do modelo é a maior diferença entre as distribuições acumuladas das probabilidades dos grupos de clientes ativos e não ativos para o uso do cartão de crédito. (Botelho e Tostes 2010).

[10] A curva ROC é uma métrica que permite estudar a variação para as medidas de sensibilidade e especificidade para avaliação da performance dos modelos. (FAWCETT, 2003).

[11] A *Area Under the Curve* agrega o desempenho de um modelo de classificação em um único número, o que facilita a comparação do desempenho geral de vários modelos de classificação. Um classificador aleatório tem uma AUC de 0,5 e um classificador perfeito possui uma AUC igual a 1. (Zhu, Baesens e Broucke, 2017).

produtos, dentre outras, são as principais para identificar os clientes propensos a encerrarem o relacionamento no mercado bancário.

Conforme o trabalho de pesquisa de Schneider (2016), foi possível compreender que a aplicação de métodos de *machine learning* para o gerenciamento do *Churn* cresceu nas últimas décadas pelo advento do crescimento da internet e suas tecnologias. O autor seguiu com a constatação que o ano 2000 foi um marco para a sua pesquisa, pois foi o período mais antigo que foi possível encontrar artigos sobre o tema. O objetivo do trabalho foi reunir a maior quantidade possível de artigos, pesquisas e técnicas de *machine learning* para servir como uma referência consolidada para futuras consultas como um repositório de trabalhos existentes em diversas áreas de aplicação: Telecomunicações, Financeiras, Jornais, Varejo etc.

De acordo com o trabalho de Lento (2017) foi possível compreender a importância do balanceamento das amostras por meio das técnicas *Under-sampling*[12] e *Over-sampling*[13] quando se tratar de classes desbalanceadas com baixa frequência. O autor explicita que, conforme apresentado por Drummond, Holte *et al.* (2003), tornar artificialmente as classes em tamanhos iguais por meio do balanceamento das amostras é geralmente mais eficaz, sendo observado no escopo do seu estudo resultados melhores com o método de *Random forest* com balanceamento *over-sampling.*

Em relação aos modelos preditivos sobre evasão de clientes em bancos, o trabalho de De Franceschi (2019), trouxe a análise em uma amostra de 2343 clientes, contemplando 118 variáveis em relação ao período de 17 meses anteriores a fevereiro de 2019. O autor informou que utilizou o software *RStudio* e que executou a análise com os métodos preditivos Regressão Logística, Regressão Logística com seleção *Stepwise* de variáveis independentes, Florestas Aleatórias (*Random Forests*) e Redes Neurais. Por fim, observamos que o modelo com maior performance foi regressão logística com curva ROC de 0,74 e acurácia média de 81.

No quadro a seguir temos o resumo dos estudos citados neste trabalho que utilizaram modelos preditivos de *churn.*

[12] *Under-sampling* consiste em eliminar observações da classe mais frequente (BUREZ; POEL, 2009).
[13] *Over-sampling* consiste em replicar a classe menos frequente (BUREZ; POEL, 2009).

Quadro 1 – Síntese dos estudos preditivos de *churn*		
Autor	**Objeto da Pesquisa**	**Técnicas**
Pinho (2009)	Análise RFV do cliente por algoritmos genéticos na otimização de estratégias de marketing.	Algoritmos genéticos; *Data mining.*
Souza, Neto, Junior, Souki (2009)	Identificando clientes propensos a encerrarem o relacionamento: um subsídio aos programas de relacionamento e à retenção de clientes no mercado bancário brasileiro.	Regressão logística.
Botelho e Tostes (2010)	Modelagem de probabilidade de *churn.*	Regressão logística.
Gauer (2016)	Modelagem de evasão de clientes bancários adimplentes: identificação de padrões pelo histórico de suas operações	Escala logarítmica; *Data mining.*
Schneider (2016)	Análise preditiva de churn com ênfase em técnicas de Machine learning: uma revisão.	Regressão logística; Redes Neurais*; Stepwise;* Árvore de Decisão; *Support Vector Machine; Random Forest*
Lento (2017)	Random Forest em dados desbalanceados: uma aplicação na modelagem de churn em seguro saúde	Regressão logística; *Akaike; Stepwise;* Árvore de Decisão; *Bootstrap e Bagging; Random Forest*
De Franceschi (2019)	Modelagens preditivas de churn: o caso do banco do brasil.	Regressão logística; *Stepwise*; *Random Forests*; Redes Neurais.
Oliveira (2021)	Algoritmos de Aprendizado de Máquina na Predição e Avaliação de Evasão de Clientes em Ambiente de Produção	Regressão logística; *Adaboost*; Árvore de Decisão *EF*.

Fonte: Elaborado pelo autor.


3

3 METODOLOGIA

3.1 DADOS

A base de dados analisada se refere a uma população de 2.907.270 clientes pessoas físicas do segmento de média e alta renda de uma instituição financeira sediada no Brasil. A base [14]foi descaracterizada para atender os requisitos constantes da Lei 13.709 de 14 de agosto de 2018 que versa sobre a Lei Geral de Proteção de Dados Pessoais - LGPD.

As informações descaracterizadas foram tratadas no ambiente de nuvem de acesso restrito do banco por meio dos aplicativos *Microsoft* SQL e *Python* com objetivo de garantir a preservação dos dados e o uso meramente estatístico para fins de estudo. É importante destacar que algumas constatações observadas neste trabalho por possuírem caráter confidencial estratégico, serão suprimidas deste trabalho e disponibilizadas em sua totalidade apenas em ambiente interno do banco.

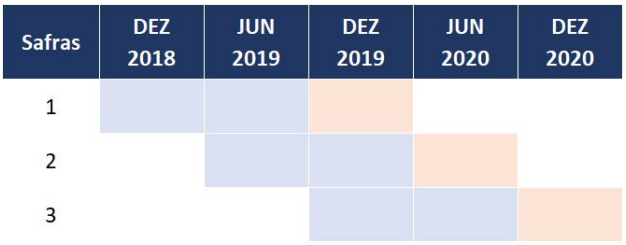
O tratamento foi realizado de forma a evidenciar as características e perfis dos clientes, com destaque para as variáveis de gênero, faixa etária, estado da federação e posse de mais de uma classe de fundos de investimento. No processo de modelagem houve o balanceamento da amostra *Under Sampling* com 200.000 observações.

Para que fosse possível verificar a evasão dos cotistas no período analisado, foi utilizada como premissa a análise do resgate total das posições individuais por meio de análises do comportamento a cada 180 dias, em média, conforme esquema da Figura 1. Nesse caso foi escolhida premissa pragmática de liquidação total da posição para buscar perseguir originalmente o cliente que realmente zerou sua posição e se evadiu (recurso não está mais no banco), e através deste

[14] Todas as informações pessoais tais como, CPF, Nome, Data de Nascimento, volume financeiro, Renda etc., não foram disponibilizadas, nem extraídas, nem salvas em dispositivos externos.

corte, buscar eliminar da análise as evoluções ou involuções de saldo que ocorrem normalmente no dia a dia dos cotistas.

Figura 1 - Esquema da avaliação do comportamento dos cotistas



Nota: Área rachurada azul avalia o comportamento dos cotistas: (i) Verifica a variação do saldo investido entre o 1º e 2º períodos; (ii) verifica no 2º período a posse de mais de uma classe de fundo. A área rachurada laranja verifica a evasão: (i) clientes que resgataram todo o saldo dos fundos; (ii) verifica se o cliente não possui mais nenhum fundo de investimento.

Fonte: Elaborado pelo autor.

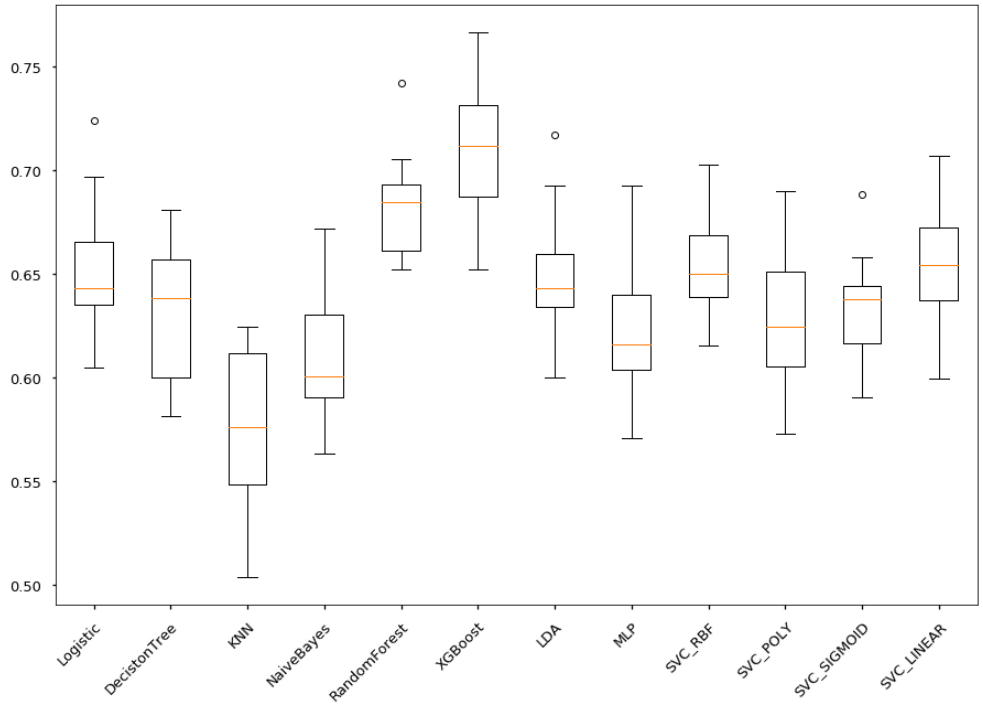
Conforme entendimento presente no trabalho de Borges (2020), a avaliação de modelos de *machine learning* por métricas parametrizadas é uma maneira capaz de indicar a performance comparativa de modelos preditivos. Corroborando com este pensamento, foi observado nos estudos constantes do quadro 1 do capítulo anterior, que os modelos preditivos utilizados pelos pesquisadores resultaram em performances diferentes para cada um dos casos, ou seja, não é porque um modelo teve performance melhor em um estudo que este modelo sempre será o melhor para outros casos. Dessa forma, cada estudo específico que contenha métricas diferentes avaliam comportamentos diferentes induzidos pelo algoritmo de classificação.

De acordo com Hossin e Sulaiman (2015) as métricas de avaliação podem ser observadas em três tipos de categorias, tais como as probabilísticas, as *threshold* e métricas de ranqueamento, no qual cada um desses tipos avaliam o classificador com objetivos distintos. Por fim, os autores destacam que esses tipos de métricas são um método de grupo escalar, em que todo o desempenho é medido e descrito por meio de um único valor, permitindo uma melhor análise comparativa, entretanto, é importante ressaltar que alguns finos detalhes de seus comportamentos fiquem sombreados. Essa visão é complementada ao longo deste trabalho, conforme Gama *et al.* (2014), tendo em vista a importância de avaliar a precisão da detecção de desvios de conceito separadamente, além da avaliação do desempenho da estratégia do aprendizado da máquina.

Para evidenciar a importância da análise do máximo de fatores que possam ser incorporados no algoritmo, é possível observar com Lavesson e Davidsson (2008) que o foco em apenas um critério de avaliação é uma das práticas mais observadas, sendo importante perceber que ao aplicar um algoritmo de *machine learning* para explicar um problema do mundo real, é recorrente a existência de outros fatores relevantes que devem ser considerados.

O gráfico 1 traz a comparação dos modelos preditivos no pré processamento da base para indicar os métodos com as melhores performances, sendo observado o *XGBoost* em destaque.

Gráfico 1 – Comparação dos Algoritmos pela Curva ROC



Fonte: Elaborado pelo autor.

3.2 TÉCNICAS QUANTITATIVAS

O objetivo deste trabalho é identificar a existência de variáveis comportamentais de cotistas de fundos de investimento para obter o modelo preditivo de melhor performance para avaliar o *churn.* Foi exposto na figura 2[15] que, dentre todos os modelos avaliados, o *XGBoost* foi o modelo com a melhor performance da Curva ROC, seguido pelos

[15] Figura 2 – Comparação dos Algoritmos pela Curva ROC.

modelos *Random Forest*, *SVC* Linear e Regressão logística. Neste tópico abordaremos os conceitos teóricos destas técnicas.

3.2.1 *Extreme Gradient Boosting* – XGBoost

O trabalho de Guilhon (2020) utilizou o algoritmo XGBoost para atribuir perfis de risco de sucesso ou fracasso das transferências voluntárias do Governo Federal para os municípios. O autor cita que o algoritmo XGBoost foi proposto originalmente por Chen e Guestrin (2016) por meio da percepção de sucesso nas aplicações de aprendizagem de máquina a dois fatores, sendo o uso de modelos efetivos que conseguem capturar as dependências complexas dos dados e de sistemas de aprendizagem com a capacidade de escala a partir de extensos conjuntos de dados.

De acordo com Medeiros (2021) o XGBoost é nome de uma biblioteca que implementa um modelo de classificação e regressão baseado em árvores de decisão. Na sua dissertação de mestrado Junior (2021) traz que, de acordo com Panesar (2019), o XGBoost tem como principais características, a regularização e distribuição no processamento computacional.

Conforme as obras dos autores Breiman *et al.* (1984) e Metha, Agrawal e Rissanem (1996), as árvores de decisão são utilizadas para aprendizagem computacional por possuírem simples entendimento e apresentação do modelo proposto, o que facilita a compreensão do usuário. Tendo em vista que não são paramétricos, seus algoritmos não precisam de informações prévias acerca da distribuição dos dados, o que torna o modelo adequado para geração de conhecimento.

É possível concluir, segundo Junior (2021), que tais características fazem com que as partes do treinamento e da validação dos dados sejam eficientes e velozes gerando um aperfeiçoamento da precisão do modelo conforme mais árvores vão sendo criadas. Esse processo resulta na otimização da velocidade de processamento dos dados, garantindo uma boa performance e bom desempenho.

Em sua tese de mestrado, Medeiros (2021), aborda que, de acordo com Taconeli (2008) a árvore de decisão presente no XGBoost é um modelo voltado para classificação ou regressão, sendo formado em

sucessivas partições binárias de uma amostra que busca a constituição de subamostras internamente homogêneas. Ainda, de acordo com Taconeli (2008), a classificação destas subamostras é baseada em uma medida descritiva e na predição de novos elementos.

Os principais modelos de *boosting* [16] conforme o estudo de Guilhon (2020) são descritos abaixo:

- *Adaboost* (FREUND; SCHAPIRE, 1995)
- *Gradient boosting machine* (FRIEDMAN; HASTIE; TIBSHIRANI, 2001)
- *Stochastic gradiente boosting* (FRIEDMAN, 2002)
- *XGBoost* (CHEN; GUESTRIN, 2016).

O trabalho de Guilhon (2020) mostrou as melhorias que Friedman, Hastie e Tibshirani (2000) trouxeram para o *XGBoost* em relação às técnicas de *gradient tree boosting* na função objetivo regularizada, onde para um determinado conjunto de dados com *n* exemplos e *m* características, $D = \{(x_i,\ y_i)\}$ $(|D| = n,\ x_i \in \mathbb{R}^m, y_i \in \mathbb{R})$, um arranjo de árvores de decisão utiliza *K* funções aditivas para a previsão de um resultado:

$$\hat{y_i} = \phi(x_i) = \sum_{k=1}^{K} f_k\,(x_i), f_k \in F$$

onde $F = \{f(x) = w_{q(x)}\}$ $(q : \mathbb{R}^m \to T, w \in \mathbb{R}^t)$ é o espaço de árvores de regressão.

Nesta exposição, a variável *q* é descrita como a estrutura de cada árvore responsável por mapear um exemplo para o índice de folha correspondente. A variável *T* é o número de folhas da árvore, que também pode ser entendida como a complexidade do modelo. Cada f_k corresponde a uma estrutura de árvore independente *q* com peso *w* em suas folhas.

Tendo em vista que as árvores de regressão possuem um *score* específico em cada folha, a variável w_i serve para representar o *score* da

[16] Técnica de arranjo (*ensemble*) para criar uma coleção de árvores. Os modelos são aprendidos sequencialmente com os algoritmos de aprendizagem (*learners*) iniciais e demais, com o objetivo de se ajustar aos dados e minimizar os erros de seus antecessores. (FREUND; SCHAPIRE, 1995).

i-ésima folha. A previsão é a soma dos scores de cada folha, classificadas conforme as regras de *q*. (GUILHON, 2020).

3.2.2 Random Forests

O trabalho de De Franceschi (2019) mostra que o método de *Random forests* utiliza muitas árvores de decisão em seu modelo. Ele realiza classificações por meio de um vetor de entrada, que é tratado em cada uma das árvores da floresta. De acordo com Zimmermann *et al.* (2018), cada uma das árvores fornecem uma classificação, e o conjunto de árvores escolhem a classificação com o maior número de votos.

O objetivo do algoritmo, conforme demonstrado por Lento (2017), é aperfeiçoar a redução da variância através do *bagging*[17] por meio da redução da correlação entre as árvores, sem necessariamente elevar de forma considerável o valor da variância. Além disso, de acordo com Friedman, Hastie e Tibshirani (2001), existe uma facilidade de paralelização do processo visto a independência do cálculo de cada uma das árvores.

É possível visualizar o modelo matemático do *Random Forests* por meio do trabalho de Schneider (2016), no qual um conjunto de treinamento $X = x_1, \ldots, x_n$ com respostas $Y = y_1, \ldots, y_n$, realiza a agregação *bagging* de forma repetida *B vezes*. Esse processo seleciona uma amostra de forma aleatória com a substituição do conjunto de treinamento e realiza o ajuste das árvores com estas amostras.

$$b = 1, \ldots, B$$

Para a amostra (substituição) com *n* exemplos de treinamento de *X*, *Y*, são denominados X_b, Y_b. Do treinamento de uma árvore de decisão, dado a função f_b em X_b, Y_b. Por fim, Schneider (2016) mostra que as previsões para novas amostras *x*, pós treinamento, podem ser

[17] Criação de inúmeros subconjuntos aleatórios dos dados de treinamento (*bootstraps*). Cada subconjunto é utilizado para o treinamento de uma árvore de decisão. O resultado é obtenção de um conjunto de árvores diferentes. A média de todas as previsões das diferentes árvores é utilizada como resultado, o que torna a decisão mais robusta. (BREIMAN, 1996).

realizadas por meio da média das previsões de todas as árvores de regressões individuais sobre *x* conforme:

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b(x)$$

Por fim, Lento (2017) mostra que o método pode ser utilizado tanto para classificação quanto para predição. Quando for utilizada para classificação, a *random forest* reúne o voto de cada uma das árvores e classifica pela maioria, e do caso de uso para regressão, o resultado se caracteriza pela média das previsões de cada árvore.

3.2.3 Support Vector Classification - Linear

O trabalho de Schneider (2016) introduz o modelo *Support Vector Machine – SVM*, que se trata de modelos de aprendizagem com algoritmos que estudam os dados para obter classificações lineares e não lineares. Esta técnica, conforme Araujo (2015), foi desenvolvida pelo russo Vladmir Vapnik para resolver problemas de classificação, ademais, de acordo com Chamasemani e Singh (2011), também pode ser aplicada para estudos de regressão. Dessa forma, de acordo com Gunn (1998), quando se tratar da apresentação de problemas de classificação, o modelo é conceituado como *Support Vector Classification – SVC* e quando se tratar de problemas de regressão de vetores, o modelo é conceituado como *Support Vector Regression – SVR.*

A Classificação de Vetores de Suporte, do inglês *Support Vector Classification*, - *SVC* é abordada no trabalho de Araujo (2015), como a técnica capaz de classificar objetos de classes diferentes por meio da aprendizagem da máquina. O autor completa que, conforme Gunn (1998), a *SVC* Linear busca encontrar um classificador que possa distinguir as classes de forma assertiva, ou seja, não permite que nenhum objeto das classes seja classificado de forma errada. Este classificador, demonstrado por Chamasemani e Singh (2011), é chamado de hiperplano, podendo ser demonstrado por meio da equação:

$$f(x) = w.x + b = 0$$

Em continuidade à demonstração, Araujo (2015), destaca *w* como o vetor que representa o conjunto de entradas, conhecido como vetor de peso; *x* como os pontos sobre o hiperplano e *b* como a distância entre o hiperplano e o ponto de origem. Conforme Lorena e Carvalho (2007), o hiperplano realizará a separação dos dados em duas regiões, sendo $f(x) > 0$ ou $f(x) < 0$. Para a obtenção da classificação é usada a função:

$$h(x) = sgn(f(x)) = \begin{cases} -1 \; se \; w.x + b < 0 \\ +1 \; se \; w.x + b > 0 \end{cases}$$

Onde *-1* é uma classe de objetos, e *+1* os objetos de outra classe.

3.2.4 Regressão Logística

De acordo com Cabral (2013) os modelos de regressão logística são uma das ferramentas estatísticas mais usadas na análise estatística de dados quando a necessidade é modelar relações entre variáveis. Um dos principais objetivos destes modelos é identificar a relação entre uma ou mais variáveis explicativas (ou independentes) e uma variável resposta (ou dependente).

Esta ideia é ratificada no trabalho de Paula (2004) e de Dastile, Celik e Potsane (2020) quando observamos que a regressão logística é um dos principais modelos executados na modelagem de dados, tendo em vista a simplicidade e facilidade na interpretação dos parâmetros.

Um dos casos particulares dos modelos lineares generalizados são os modelos onde a variável resposta apresenta apenas duas categorias ou que de alguma forma foi dicotomizada assumindo valores 0 ou 1 sendo o modelo de regressão logística o mais popular desses modelos. (CABRAL, 2013).

De acordo com De Franceschi (2019), a variável dependente da regressão logística é o logaritmo natural dos *odds* de ocorrência de um fenômeno.[18] O autor da pesquisa demonstra que o modelo *logit* é aderente, por exemplo, para definir a relação entre a probabilidade de

[18] Probabilidade de ocorrência de um evento dividida pela probabilidade de não ocorrência, (DE FRANCESCHI 2019).

um cliente evadir e um vetor de características do cliente $[x_1, \dots, x_k]$, definido pela função a seguir:

$$log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

A variável dependente, conforme demonstração de De Franceschi (2019), para ser entendida como evento de probabilidade, deve ser utilizada na equação a seguir:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}}{[1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}]}$$


4

4 RESULTADOS

O método *XGBoost* foi selecionado para a presente pesquisa por ter demonstrado performance superior em relação aos demais métodos, desta forma evidenciaremos neste capítulo alguns dos principais resultados com base neste método de aprendizagem de máquina. Considerando que os resultados obtidos demonstraram aptidão para serem utilizados pelo banco, algumas informações serão suprimidas por serem confidenciais e estratégicas.

4.1 RESULTADOS[19] DO TREINAMENTO DA BASE

O treinamento do algoritmo observou o desempenho de uma série de variáveis explicativas de *churn.* Para cada uma das variáveis o algoritmo verificou na amostra a relação entre aquela variável e a evasão, e assim, quantificou a quantidade e percentual de evasão para o treinamento conforme exemplo teórico da Tabela 1.

Tabela 1 – Exemplo teórico da visão geral de *churn* por variável explicativa

Segmento	Qtd. Clientes	Qtd. Evasão	% Evasão
M	*R*	*R'*	*a %*
N	*S*	*S'*	*b %*
O	*T*	*T'*	*c %*
P	*U*	*U'*	*d %*
Q	*V*	*V'*	*e %*

Fonte: Elaborado pelo autor.

O tratamento da tabela 1 foi aplicado na base histórica de 2.907.270 representada pela amostra de 200.000 clientes. Foi utilizada a função de seleção de variáveis[20] para remover e identificar as características que são relevantes para o treinamento do modelo, como por exemplo, a exclusão de variáveis com característica colineares, ou seja, altamente correlacionadas com outras variáveis, sendo possível a

19 Tendo em vista o caráter estratégico e confidencial destas informações, nem todas as variáveis foram citadas.

20 Ferramenta para aprendizado de máquina em Python. Disponível em https://towardsdatascience.com/a-feature-selection-tool-for-machine-learning-in-python-b64dd23710f0.

remoção de uma dessas variáveis colineares para uma melhor eficiência do modelo.

Após a realização desta seleção, restaram 38 variáveis para o treinamento do modelo em relação a evasão, das quais citamos algumas[21]:

- Gênero;
- Cesta de Serviços;
- Faixa Etária;
- Unidade da Federação;
- Tempo de Relacionamento;
- Volume de Investido;
- Volume de crédito;
- Quantidade de produtos contratados;
- Perfil Digital;
- Segmentação;
- Tipo de fundos investidos;
- Outras.[22]

4.2 MATRIZ DE CONFUSÃO E CURVA ROC - *EXTREME GRADIENTE BOOSTING (XGBOOST)*

De acordo com Junior (2021) o aprendizado de máquina possui uma grande diversidade de algoritmos que são utilizados na classificação ou previsão de resultados que, por sua vez, são baseados em variáveis para que possam explicar a variável resposta.

Após o tratamento dos dados e a compreensão de que o movimento de evasão pode ser explicado pelas variáveis demonstradas neste capítulo, evidenciaremos os resultados da capacidade preditiva com a matriz de confusão e Curva ROC.

É possível compreender, conforme Tharwat (2018), que a matriz de confusão é utilizada para demonstrar a eficiência de modelos preditivos usados para classificação. Cada linha da matriz significa uma ocorrência da verdadeira classe, ao tempo que cada coluna indica uma ocorrência específica de uma classe preditiva do modelo.

[21]Dados descaracterizados conforme Lei 13.709 de 14 de agosto de 2018 que versa sobre a proteção de dados pessoais.
[22] Foi gerada análise de score *churn* para 38 variáveis explicativas no total.

Para uma melhor visualização do conteúdo da matriz de confusão, a tabela 2 mostrará a composição das métricas Verdadeiro Positivo (TP), Falso Positivo (FP), Verdadeiro Negativo (TN) e Falso Negativo (FN).

Tabela 2 – Demonstração teórica dos resultados da matriz de confusão.		
	Previsão Negativa	**Previsão Positiva**
Realidade Negativo	*TN (Verdadeiro Negativo)*	*FP (Falso Positivo)*
Realidade Positivo	*FN (Falso Negativo)*	*TP (Verdadeiro Positivo)*

Fonte: Elaborado pelo autor.

O resultado verdadeiro negativo (TN – *True Negative*) se refere aos resultados que o modelo previu como negativo e que na realidade foi negativo, ou seja, o modelo acertou. Em relação ao resultado falso positivo (FP – *False Positive*) o modelo previu que seria positivo, quando na realidade foi negativo, evidenciando um erro. O resultado falso negativo (FN – *False Negative*) se refere aos resultados que o modelo previu como negativo, contudo, na realidade foi positivo, resultando em erro do modelo. E o resultado verdadeiro positivo (TP – *True positive*) se refere aos resultados que o modelo previu como positivo e que na realidade foi positivo, ou seja, caracterizando o acerto do modelo.

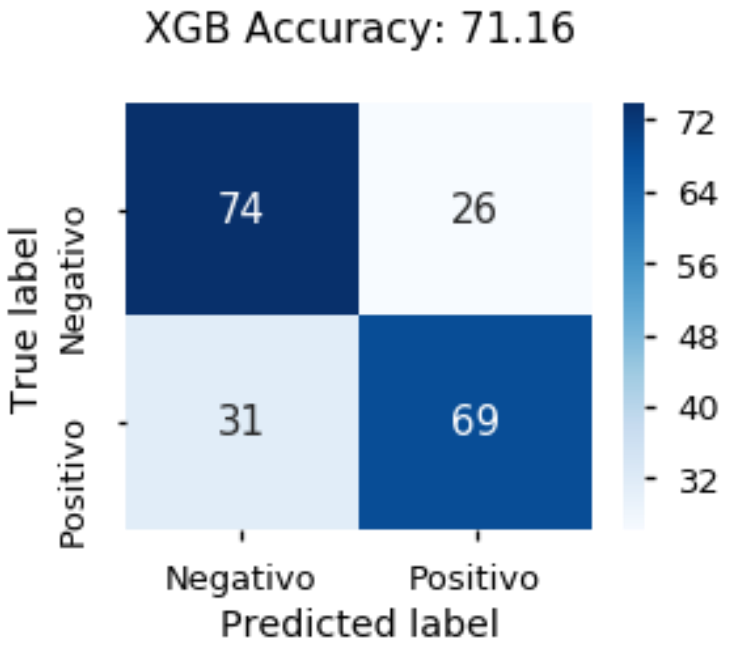
É possível extrair métricas para a avaliação dos resultados da Tabela 2 conforme Tharwat (2018), Hossin e Sulaiman (2015) e Luque *et al.* (2019), descritos na Tabela 3:

Tabela 3 – Descrição das métricas para avaliação dos resultados da matriz de confusão	
Métricas	**Descrição**
Precisão	A precisão é usada para medir a performance do modelo em relação aos casos reais positivos que são previstos corretamente a partir do total de previsões positivas, inclusive as previsões falsas positivas. $$\text{Precisão} = \frac{\text{TP}}{\text{FP} + \text{TP}}$$
Recall	O *recall* é usado para evidenciar os casos reais, ou seja, a quantidade de casos verdadeiros positivos em relação aos casos reais verdadeiros que foram previstos como falsos. $$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
F1 *Score*	Essa é uma média harmônica entre precisão e recall, pois combina ambas para trazer um número único que indique a qualidade geral do modelo. É uma métrica que observa tanto a precisão do modelo quanto a quantidade de casos reais que não são previstos corretamente. $$\text{F1 Score} = \frac{2 * \text{Precisão} * Recall}{\text{Precisão} + Recall}$$

Fonte: Elaborado pelo autor.

Aplicando os conceitos demonstrados na Tabela 2 e Tabela 3, é possível observar a performance na prática do modelo no Gráfico 2. Observamos que quando o modelo deveria indicar "positivo" para o *churn* de algum cliente, ele acertou 69% das predições, e da mesma forma, quando o modelo deveria reportar "negativo" para o *churn* de algum cliente, o modelo acertou 74%, totalizando uma acurácia total de 71,16%.

Gráfico 2 – Matriz de confusão XGBoost



Fonte: Elaborado pelo autor.

Na Tabela 4 é possível observar que tanto a precisão quanto o *recall* estão com performance satisfatória e equilibrada, pois, de acordo com Tharwat (2018) o F1 *Score* mostra o desbalanceamento entre a precisão do modelo e o *recall*.

Tabela 4 – Resultados da matriz de confusão: Precisão, *Recall* e F1 *Score*.

	Precisão	*Recall*	F1 Score
Classe Negativa	*0,70*	*0,74*	*0,72*
Classe Positiva	*0,72*	*0,69*	*0,70*

Fonte: Elaborado pelo autor.

O trabalho de Fawcett (2006) mostra outra maneira de avaliar a performance dos modelos preditivos por meio da utilização da Curva ROC[23]. É possível avaliar a capacidade do modelo em separar os casos

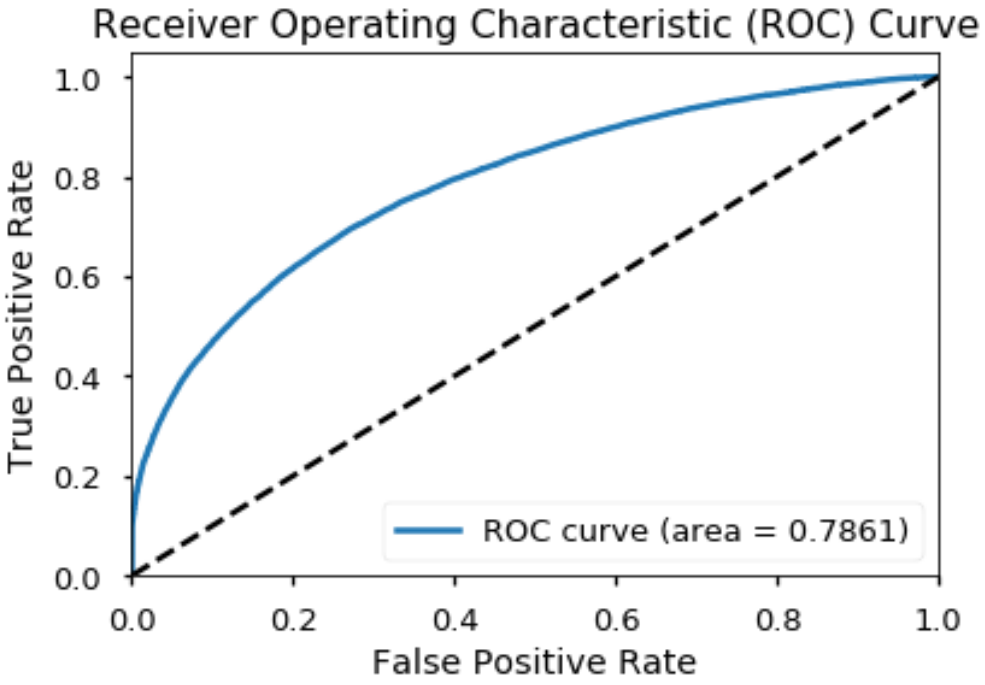
[23] *ROC - Receiver Operating Characteristic.*

positivos e negativos. Outro conceito complementar, é do autor Prati (2008) que mostra os gráficos ROC como um método gráfico para avaliação, organização e seleção de sistemas de diagnósticos ou predição. O gráfico é baseado na probabilidade de detecção da taxa de verdadeiros positivos e na probabilidade de falsos alarmes.

A dissertação de Oliveira (2021) ressalta que antes de comparar visualmente as curvas, a área sobre a curva[24] carrega a performance de um modelo de classificação em um único número, o que simplifica a comparação do desempenho geral de vários modelos de classificação.

É possível analisar no Gráfico 3 que a AUC resulta em 0,786. A AUC fornece uma estimativa da probabilidade de classificação correta da previsão do modelo (acurácia do teste). No caso, uma AUC de 0,786 reflete uma chance de classificação correta de aproximadamente 79% do modelo. De forma geral, um modelo cujas previsões são aleatórias tem uma AUC de 0,500, enquanto um modelo teoricamente perfeito tem uma AUC de 1,00.

Gráfico 3 – Curva ROC XGBoost



Fonte: Elaborado pelo autor.

[24] *AUC – Area Under the Curve*. É a área localizada embaixo da curva ROC.


5

5 CONCLUSÃO

Este trabalho buscou evidenciar matematicamente que os cotistas dos segmentos de média a alta renda possuem características específicas e que devem ser levadas em conta para sua permanência nas instituições financeiras. Dessa forma, o trabalho contribui com ações de relacionamento com o cliente e com o desenvolvimento ou aperfeiçoamento de produtos, visto que cada vez mais a competitividade do mercado, o acesso à informação e o lançamento de novos produtos, tornarão o processo de retenção e fidelização um grande desafio para as instituições.

O contexto que este trabalho se insere não podia ser mais apropriado, pois no exato mês da conclusão e apresentação, dezembro de 2021, o Banco Central iniciou a quarta fase do *Open Banking*, que, dentre outros produtos, disponibilizará as informações dos correntistas das instituições financeiras em relação às suas carteiras pessoais de investimentos. Esse movimento está previsto para finalizar em maio de 2022, o que resultará numa maior competição entre os bancos para manter os clientes, visto que as informações serão públicas mediante a autorização do próprio cliente.

Observamos que existem características que, quando presentes ou ausentes, quantificam em percentual relevante a evasão de clientes. Para exemplificar, ao analisar uma das variáveis explicativas com *score* relevante para o *churn*, observamos que a sua ausência no rol de características dos clientes, resultou numa evasão de, aproximadamente, *2,3 vezes* maior quando comparados com clientes que possuem a característica testada, o que destaca a importância do aproveitamento dos resultados de *churn* deste estudo.

Por meio do treinamento do modelo foi possível entender o comportamento da base histórica e aplicar o algoritmo de *churn* na base atual de clientes que, no caso, foi aplicado na base com posição de novembro de 2021. Como resultado da aplicação do algoritmo, foi identificada quantidade relevante de clientes com *status* de *churn*. Através da identificação dos clientes propensos a evadir, foi possível realizar uma simulação teórica dos reflexos financeiros caso o *churn* de

fato ocorra. Os valores financeiros simulados demonstraram grande impacto no resultado do banco, evidenciando a importância de sua aplicação imediata.

Sobre as possíveis ações a serem implementadas imediatamente, destacamos o trabalho direcionado de contato com cada um dos clientes identificados, de tal forma que seja realizada a oferta de produtos e serviços com o objetivo do cliente não permanecer com a característica de potencial *churn*. Outra ação em destaque, é a criação de estratégia de prospecção de novos clientes com base nas variáveis identificadas, por se tratar de variáveis que identificam clientes promissores para aceitar a oferta, desta maneira, quando os dados do *open banking* estiverem disponíveis, o banco já estará com estratégia pronta para atuação.

Desta forma, considerando o relevante impacto financeiro potencial deste trabalho, e que em 2022 o *open banking* será uma realidade no sistema financeiro nacional por meio do compartilhamento das informações pessoais pelos próprios clientes, o mercado se tornará mais competitivo e a capacidade preditiva de evasão em conjunto com as estratégias de relacionamento poderão se tornar questões de sobrevivência neste mercado.

REFERÊNCIAS

REFERÊNCIAS

ARAUJO, Kevin Martins. **Utilização do algoritmo de máquina de vetores de suporte (SVM) para predição de dados climáticos.** Trabalho de conclusão de curso de Graduação em Ciência da Computação. 96 f. Centro Universitário Luterano de Palmas do Tocantins. Palmas, 2015.

ASSOCIAÇÃO BRASILEIRA DE FINTECHS (ABFINTECHS). **Pesquisa Fintech Deep Dive 2020**. São Paulo: PricewaterhouseCoopers, 2020. Disponível em: https://www.pwc.com.br/pt/estudos/setores-atividade/financeiro/2021/pesquisa-fintech-deep-dive-2020.html. Acesso em: 12 dez. 2021.

BOTELHO, Delane; TOSTES, Frederico D. **Modelagem de probabilidade de *churn***. RAE – Revista de Administração de Empresas, v.50, n.4, 2010.

BORGES, Mirele Marques. ***Machine Learning* como ferramenta gerencial para predição de indicadores e detecção de anomalias**. Dissertação de Mestrado Profissional em Engenharia de Produção da Universidade Federal do Rio Grande do Sul. Porto Alegre, UFRGS, 2020.

BRASIL. Banco Central do Brasil (BACEN): ***Open Banking***. Banco Central do Brasil, Brasília. Disponível em: https://www.bcb.gov.br/estabilidadefinanceira/openbanking. Acesso em: 12 dez. 2021.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD).** Brasília: Presidência da República, 2018. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 20 mai. 2021.

BRASIL. Conselho Administrativo de Defesa Econômica (Cade). **Cadernos do Cade: Mercado de instrumentos de pagamento**. Brasília: Cade, 2019. Disponível em: https://cdn.cade.gov.br/Portal/Not%C3%ADcias/2019/Cade%20divulga%20estudo%20sobre%20mercado%20de%20instrumentos%20de%20pagamento__Cadernodeinstrumentosdepagamento_27nov2019.pdf. Acesso em: 10 out. 2021.

BREIMAN, L; FRIEDMAN, J.H; OLSHEN, R. A; STONE, C.J. ***Classification and regression trees.*** [S.l.]: Taylor & Francis Group, 1984.

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

BRODSKY, Laura; OAKES, Liz. ***Data sharing and Open Banking.*** McKinsey & Company, 2017. Disponível em: https://www.mckinsey.com/industries/financial-services/our-insights/data-sharing-and-open-banking#. Acesso em: 21/11/2021.

BROWN, Stanley A. CRM – ***Customer Relationship Management*****: uma ferramenta estratégica para o mundo e-business**. São Paulo: Makron Books do Brasil, 2001. (original inglês,2000).

BUREZ, J; POEL, D. Van den. ***Crm at a pay-tv company: Using analytical models to reduce customer attrition by targeted marketing for subscription services.*** *Expert Systems with Applications,* Elsevier, v. 32, n. 2, p.277-288, 2007

CABRAL, Cleidy Isolete Silva. **Aplicação do Modelo de Regressão Logística num Estudo de Mercado.** Tese de Mestrado em Matemática Aplicada à Economia e à Gestão. Universidade de Lisboa: 2013.

CHAMASEMANI, Fereshteh Falah; SINGH, Yashwant Prasad. ***Multi-class Support Vector Machine (SVM) classifiers-An Application in Hypothyroid detection and Classification.*** In: Sixth International Conference on Bio-Inspired Computing: Theories and Applications, 2011.

CHEN, Tianqi; GUESTRIN, Carlos. ***Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining***. 2016.

COELHO, Ricardo. **Repensando Banco de Varejo**. 3 Ed.a. São Paulo: Ed. Ápice, 2005.

DASTILE, X; CELIK, T; POTSANE, M. ***Statistical and machine learning models in credit scoring. A systematic literature survey***. *Applied Soft Computing*, v. 91, p. 106263, 2020.

DE FRANCESCHI, Pietro Reinheimer. **Modelagens Preditivas de *Churn*: O caso do Banco do Brasil.** Dissertação de Mestrado em Administração. Universidade do Vale do Rio dos Sinos. Porto Alegre, 2019.

DRUMMOND, C; HOLTE, R.C. *et al*. **C4.5, *class imbalance, and cost sensitivity: why under-sampling beats over-sampling*.** In: CITESEER. Workshop on learning from imbalanced datasets II. S.1, 2003. V.11.

FAWCETT, Tom. ***ROC Graphs: Notes and Practical Considerations for Data Mining Researchers*.** Intelligent Enterprise Technologies Laboratory. HP Laboratories. Palo Alto. 2003.

FAWCETT, Tom. ***An introduction to roc analysis.*** *Pattern recognition letter*s, Elsevier, v. 27, n. 8, p. 861-874, 2006.

FREUND, Y; SCHAPIRE, R.E. ***A decision-theoretic generalization of on-line learning and an application to boosting*.** Proc. 2nd European Conf. on Computational Learning Theory, p. 23-37, 1995.

FRIEDMAN, J; HASTIE, T; TIBSHIRANI, R. ***Greedy function approximation: a gradient boosting machine*.** Annals of Statistics, v. 29, p. 1189-1232, 2001.

FRIEDMAN, J.H. ***Stochastic gradient boosting***. Computational Statistics and Data Analysis, v. 38, p. 367-378, 2002.

GAMA, J. *et al*. ***A survey on concept drift adaptation*.** ACM computing surveys (CSUR), ACM New York, NY, USA, v. 46, n. 4, p. 1–37, 2014.

GAUER, Jefferson José Cerutti. **Modelagem de evasão de clientes bancários adimplentes: identificação de padrões pelo histórico de suas operações.** 91f. Dissertação de Mestrado em Gestão do Conhecimento e Tecnologia da Informação. Universidade Católica de Brasília, Brasília. 2016.

GHORBANI, A; TAGHIYAREH, F. CMF: ***A framework to improve the management of customer churn***. 2009 IEEE *Asia-Pacific Services Computing Conference* (APSCC), p. 457, 2009.

GIL, Antonio Carlos. **Métodos e Técnicas de Pesquisa Social**. São Paulo: Atlas, 2006.

GLADY, N.; BAESENS, B.; CROUX, C. ***Modeling churn using customer lifetime value.*** European Journal of Operational Research, v. 197, n. 1, p.402–411, ago. 2009.

GUILHON, Daniel Moreira. **Classificação de risco em transferências voluntárias federais utilizando XGBoost**. Dissertação de Mestrado em Ciência da Computação. Universidade Federal do Maranhão, São Luis. 2020.

GUIMARAES, Olavo. **Concorrência bancária e o *Open Banking* no Brasil.** p. 125-147 RDC, Vol. 9, no 1. Junho 2021

GUJARATI, Damodar N; PORTER, Dawn C. **Econometria Básica.** Porto Alegre: AMGH, 2011.

GUPTA, S.; LEHMANN, D.R. ***Managing Customers as Investments: The Strategic Value of Customers in the Long Run.*** NJ, Upple Saddle River: Wharton School Publishing, Pearson Education, Inc.. 2005.

HOSSIN, M.; SULAIMAN, M. ***A review on evaluation metrics for data classification evaluations.*** International Journal of Data Mining & Knowledge Management Process, Academy & Industry Research Collaboration Center (AIRCC), v. 5, n. 2, p. 1, 2015.

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

JUNIOR, Gerson Ratcow. **Modelagem Preditiva do Preço de Aluguel de Apartamentos por Bairros na Cidade de São Paulo.** Osasco: UNIFESP, 2021.

KOEHRSEN, Will. ***A feature selection tool for machine learning in Python.*** Towards data Science. Disponível em: https://towardsdatascience.com/a-feature-selection-tool-for-machine-learning-in-python-b64dd23710f0. Acesso em: 01/12/2021.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade. **Fundamentos de Metodologia Científica**. São Paulo: Atlas, 2003.

LAROSE, Daniel. ***Data Mining, Methods and Models.*** John Wiley & Sons, New Jersey, Canada, 2006.

LAVESSON, N.; DAVIDSSON, P. ***Generic methods for multi- criteria evaluation.*** In: SIAM. Proceedings of the 2008 SIAM International Conference on Data Mining. [S.l.], 2008. p. 541–546.

LENTO, Gabriel Carneiro. ***Random Forest* em dados desbalanceados: uma aplicação na modelagem de churn em seguro saúde.** 41 f. Dissertação de Mestrado em Matemática. Fundação Getúlio Vargas, Rio de Janeiro. 2017.

LORENA, Ana Carolina; CARVALHO, André C. P. L. F. de. **Uma Introdução às *Support Vector Machines*.** Revista de Informática Teórica e Aplicada (RITA) V.14, n.2, p.43-67, 2007.

LUQUE, A. *et al*. ***The impact of class imbalance in Classification performance metrics based on the binary confusion matrix.*** *Pattern Recognition*, [s.i.], v. 91, p. 216-231, 2019

MEDEIROS, Augusto Santana Veras de. **Estudo sobre o uso de análise técnica e *XGBOOST* em operações de *Day-trade*.** Natal: UFRN, 2021.

MEHTA, M. AGRAWAL, R. RISSANEN, J. ***Sliq: A fast scalable classifier for data mining.*** IN: SPRINGER. *International conference on extending database technology*. [S.l.], p. 18-32, 1996.

MOSCHOVAKIS, Yiannis N. ***What is an algorithm? In: Mathematics unlimited-2001 and beyond***. Springer, Berlin, Heidelberg, 2001. p. 919-936.

MOHRI, Mehryar; ROSTAMIZADEH, Afshin; TALWALKAR, Ameet. ***Foundations of machine learning***. MIT press: 2018.

NITZAN, I; LIBAI, B. ***Social Effects on Customer Retention. Journal of Marketing***. v. 75, n. 6, p. 24-38, Nov. 2011.

OLIVEIRA, Breno. **Algoritmos de Aprendizado de Máquina na Predição e Avaliação de Evasão de Clientes em Ambiente de**

Produção. 82 f. Dissertação de Mestrado em Ciência da Computação. Universidade Federal de Goiás. Goiania, 2021.

PANESAR, A. ***Machine Learning and AI for Healthcare: Big Data for Improved Health Outcomes***. Coventry: Apress, 2019.

PAYNE, A. ***Handbook of CRM: achieving excellence in customer management.*** Oxford: Elsevier Butterworth-Heinemann, 2006.

PAULA, G. A. **Modelos de regressão: com apoio computacional.** [S.l.]: IME-USP São Paulo,
2004.

PEREZ, Rafaella Di Palermo; STROHL, Juliana. ***Open Banking*: contexto cultural e experiência internacional**. In: Pedro Eroles (Coord.). Fintechs, Bancos Digitais e Meios de Pagamento. São Paulo: Quartier Latin, 2019.

PRATI, Ronaldo Cristiano; BATISTA, Gustavo Enrique de Almeida Prado Alves; MONARD, Maria Carolina. **Curvas ROC para avaliação de classificadores**. IEEE Latin America Transactions, Vol. 6, No. 2, junho 2008.

R, D. W. H; LEMESHOW, S; STURDIVANT, R. X. ***Applied logistic regression***. [S.l.]: John Wiley & Sons, 2013.

SCHNEIDER, Pedro Henrique. **Análise preditiva de churn com ênfase em técnicas de *Machine learning*: uma revisão.** 82 f. Dissertação de Mestrado em Matemática. Fundação Getúlio Vargas, Rio de Janeiro. 2016.

SOLTAU, Samuel Bueno. **Detecção de periodicidade em dados multi frequência de núcleos ativos de galáxias com aprendizagem de máquina (XGBoost)**. 152 f. Tese de doutorado em Engenharia de Materiais e Nanotecnologia. Universidade Presbiteriana Mackenzie, São Paulo, 2019.

SPANOS, Aris. ***Probability. Theory and statistical inference: econometric modfling with observational data.*** Reino Unido: *Cambridge University Press*, 1999. p. 21.

SOUZA, Jeislan Carlos de; NETO, Mário Teixeira Reis; JUNIOR, André Luiz Moura; SOUKI, Gustavo. **Identificando clientes propensos a encerrarem o relacionamento: um subsídio aos programas de relacionamento e à retenção de clientes no mercado bancário brasileiro.** Revista Gestão e Planejamento, Universidade Salvador, v. 10, nº 2, p. 123–140, 2009.

TACONELI, C. A. **Árvores de classificação multivariadas fundamentadas em coeficientes de dissimilaridade e entropia. 2008**. 100 f. Tese de Doutorado (Doutorado em Agronomia). Universidade de São Paulo. São Paulo, 2008.

TINTNER, Gerhard. ***Methodology of mathematical economics and econometrics.*** Chicago: The University of Chicago Press, 1968. p. 74.

THARWAT, A. ***Classification assessment methods***. *Applied Computing and Informatics*, 2018.

TRONCHIN, Valsoir. **CRM não é um bom serviço ao cliente.** Disponível em: www.1to1.com.br (Peppers and Rogers Group, Marketing 1 to 1, Inc.). Acesso em: 03 abr. 2010.

VEIGA, Fabio da Silva; GIBRAN, Sandro Mansur; BONSERE, Silvana Fatima Mezaroba. ***Open Banking*: Expectativas e Desafios para o Mercado Financeiro no Brasil.** Curitiba: Revista Unicuritiba, 2019.

VELOSO, F. J. M. **Um modelo para previsão de *churn* na área do retalho.** Universidade do Minho: Braga, 2013.

VIDIGAL, Roberto. **Diversificação de um Portfólio de Fundos Multimercado: Uma Análise Empírica.** São Paulo: Insper, 2012

ZHU, B.; BAESENS, B.; BROUCKE, S. K. Vanden. ***An empirical comparison of techniques for the class imbalance problem in churn prediction.*** *Information sciences*, Elsevier, v. 408, p. 84–99, 2017.

ZIMMERMAN, N. *et al.* ***A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring.*** *Atmospheric Measurement Techniques*, [S.l.], v. 11, n. 1, p. 291-313, 2018.


idp
A ESCOLHA QUE
TRANSFORMA
O SEU CONHECIMENTO