UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ADMINISTRACAO
CURSO DE ADMINISTRACAO

RODRIGO EIDELWEIN

# MODELOS PREDITIVOS E PREDICAO DE CHURN: ANALYTICS ALIADO A ADMINISTRACAO

Porto Alegre, RS
2023

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ADMINISTRAÇÃO
CURSO DE ADMINISTRAÇÃO

RODRIGO EIDELWEIN

# MODELOS PREDITIVOS E PREDIÇÃO DE CHURN: ANALYTICS ALIADO À ADMINISTRAÇÃO

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Administração

Orientador: Prof. Dr. Pablo Cristini Guedes

Porto Alegre
agosto de 2023

“No one can tell what goes on in between the person you were and the person you become. No one can chart that blue and lonely section of hell. There are no maps of the change. You just come out the other side. Or you don’t.” — STEPHEN KING

## AGRADECIMENTOS

Ao prof. Pablo Guedes, que foi professor sem ter me dado uma aula sequer, que aceitou um projeto sem me conhecer e que me deu uma chance quando eu mesmo não o faria.

Ao prof. Vinícius Brei, por toda inspiração durante o curso, fazendo despertar a paixão pela pesquisa acadêmica e por nunca aceitar menos que o máximo que os alunos podem entregar. Suas aulas foram as mais exigentes e difíceis do curso, mas ao final de cada assignment havia uma sensação de realmente ter feito algo de se orgulhar.

Aos meus amigos da UFRGS, em especial André, Felipe, Aline e Fernanda, as amizades que me sustentaram ao longo de diversos anos, que dividiram inúmeros momentos de trabalho e de lazer e que levarei comigo o resto da minha vida.

Ao Luiz Kruel, líder, chefe, colega, amigo, companheiro de trincheira e meu maior apoiador.

A todos professores e colaboradores da UFRGS, suas contribuições são inumeráveis e eternas.

## RESUMO

Em um contexto global de transformação digital para todas facetas da vida, torna-se fundamental otimizar ao máximo todos os aspectos variáveis de uma organização, em particular o conhecimento adquirido e auxiliar ao processo de tomada de decisão. No momento em que é possível mensurar e quantificar as características particulares de cada cliente e cada produto e entender quais são os aspectos que possuem impacto positivo e negativo no relacionamento e na manutenção do cliente com a empresa, enxerga-se a disseminação da cultura analítica, em formas multivariadas, aparecer em toda gama de empresas. De pequenos projetos desenvolvidos internamente a grandes departamentos dedicados exclusivamente para a ciência de dados, ou até mesmo consultorias especializadas contratadas para atender demandas pontuais, vemos a mudança de decisões baseadas em conhecimento empírico para conhecimento estatístico impulsionado por dados. Este trabalho busca desenvolver um modelo de previsão de abandono de cliente (churn) em uma empresa facilitadora de meios de pagamentos, a empresa Adquirente. Através da criação de bases de dados, essas extraídas da mineração de milhões de blocos de transações individuais observadas durante o período de um ano, e do perfil cadastral histórico de todos os clientes, será feita a análise do perfil dos clientes, sua classificação, clusterização e detalhamento categórico de suas variáveis. Visando conceber um modelo de retenção de clientes, através do uso de diferentes algoritmos de Machine Learning, serão testadas e investigadas as diferentes variáveis observadas e buscar-se-á a escolha de um algoritmo como preditor de abandono de clientes, para a empresa poder traçar uma estratégia de manutenção dos clientes. Após a aplicação, ambos algoritmos tiveram uma alta aderência com o *dataset*, apresentando índices na faixa de 90% de acerto.
Palavras-chave: analytics, big data, churn, data science, machine learning, retenção de clientes



**Palavras-chave:** Analytics. churn. machine learning. preditivo. big data. Python. automação.

## ABSTRACT

In a global context of digital transformation of all aspects of life, it becomes essential to optimize all variable aspects of any given organization if it is to remain competitive, in particular the acquired and auxiliary knowledge to the decision making process. Once it's possible to measure and quantify the individual characteristics of every client and product, and to understand which aspects have positive and negative impacts on the clients' continued business relationship with the Company, we see the proliferation of analytical culture, in multivaried forms, arise across the board. From in-house low scale development to large scale departments or even outsourced consultants, the shift from empirical decision making gives way to data driven statistical knowledge. This paper aims to develop a client churn prediction model for an acquiring company in the financial services sector, Acquirer. Through the creation of databases and datasets, extracted from millions of individual transactions over the course of an year, as well as historical registry data of all their clients, this work will analyze the clients' profile, classify them, organize them in clusters and categorize their variables. Aiming to create a client retention model, utilizing different Machine Learning algorythms, the study of the clients data will enable the creation of a churn probability predictor, from which the Company can develop their entire client retention strategy. After running the script, both models showed a very high rate of performance with the current dataset, with ratings hitting 90% correct predictions.



**Keywords:** analytics, big data, churn, data science, machine learning, client retention.

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| ABECS | Associação Brasileira de Empresas de Cartões de Crédito e Serviços |
| BI | Business Intelligence |
| CAC | Custo de Aquisição de Cliente |
| IA | Inteligência Artificial |
| IoT | Internet of Things |
| KPI | Key Performance Indicator |
| ML | Machine Learning |
| NPS | Net Promoter Score |
| NSU | Número Sequencial Único |
| POS | Point of Sale |
| TI | Tecnologia da Informação |

## LISTA DE FIGURAS

**SUMÁRIO**

# 1 INTRODUÇÃO

Considerando a necessidade de uma organização manter suas vantagens competitivas e fontes estáveis de receitas, a manutenção da carteira de clientes é uma das prioridades de qualquer empresa no ramo de serviços e produtos. No caso de produtos de alta padronização (também chamado de *comoditizados*), a aquisição de clientes geralmente ocorre de duas maneiras: penetração em mercados já existentes ou através de aquisição de clientes de empresas concorrentes via ofertas mais atraentes.

Sendo assim, torna-se de grande importância para as empresas o conhecimento da sua carteira, a manutenção de um bom atendimento e fidelização de seus clientes, o entendimento da receita esperada, tanto global (orçamentária) quanto por cliente ou tipo de cliente, também conhecido como o CLV (REINARTZ; KUMAR, 2003), e acompanhamento da relação econômica do cliente com a empresa, buscando evitar ou dirimir a sua mortalidade ou perda de clientes, conhecido na indústria como churn.

Empresas que atuam direta ou indiretamente no setor tecnológico estão cada vez mais fazendo uso de data analytics e análise preditiva, e se tornando dependentes de automação e otimização de processos para poder lidar com o grande número de clientes e, principalmente, com o universo expansivo que é a geração de volume de dados destes mesmos clientes (CHEN; CHIANG, 2012). O valor destes dados é intangível, mas facilmente perceptível; mais e mais as organizações dependem de dados para a tomada de decisão de nível gerencial ao nível estratégico (LAUDON; LAUDON, 2016), mudando o paradigma do conhecimento empírico para o conhecimento analítico. Alinhada a esse processo, a mudança da empresa com respeito à retenção de clientes também acarreta numa mudança no modelo de negócios no que tange ao relacionamento com os clientes, consequência direta da necessidade de manter satisfeito e bem atendido o seu cliente como forma preventiva de retenção. Empresas da indústria de cartões, como a Cielo, já incorporaram índices de satisfação de clientes, como NPS, em seu modelo de remuneração variável dos colaboradores, para ressaltar a importância e incentivar o bom atendimento de clientes à todo quadro organizacional, por entender ser fundamental para sua estratégia de retenção de clientes (CIELO, 2022).

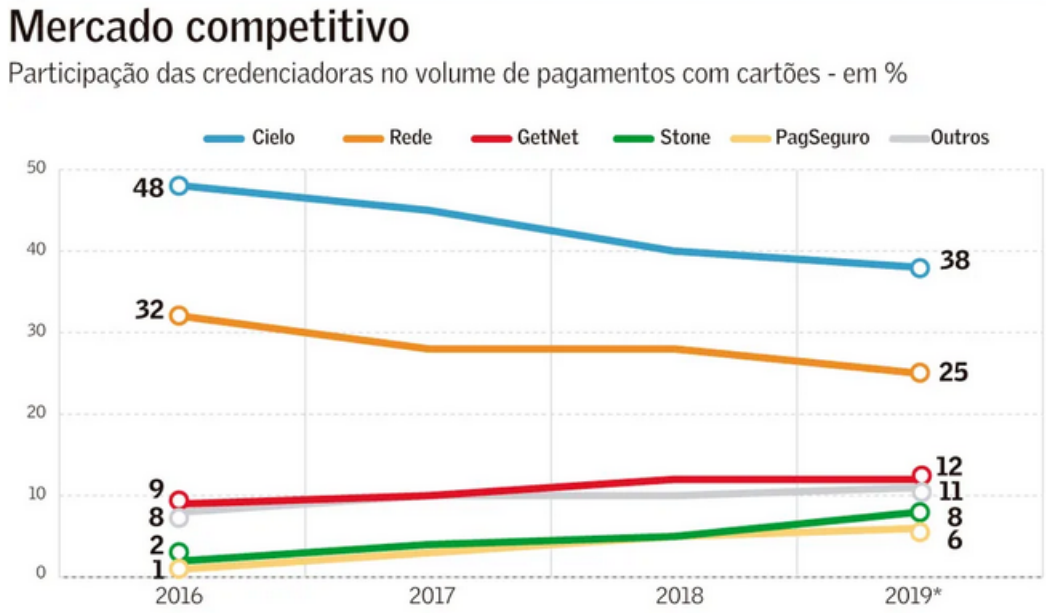
Para o setor de adquirência, mais coloquialmente conhecido como setor de cartões de crédito, mas que também envolve cartões de débito, vouchers, pagamentos eletrônicos, carteiras digitais, tokenização de serviços, crédito de giro, empréstimos caucionados a transações, para citar algumas outras ramificações, uma única transação já possui uma

miríade de pontos de dados únicos. Um único NSU (número sequencial único, ou tentativa de transação) irá registrar, por exemplo, o código do estabelecimento comercial, o horário da tentativa, o modo de captura da transação, o valor, o modelo de terminal utilização, a versão do software do terminal utilizado, o número do cartão do cliente, o emissor do cartão do cliente, o arranjo (também conhecido como bandeira) do pagamento, a modalidade de pagamento escolhida. Uma única transação registra mais de 40 variáveis para cada tentativa, com milhões de transações sendo processadas diariamente somente no Brasil. Sem ferramentas de Big Data, tais volumes não poderiam nem ser processados, muito menos estudados e transformados em valiosos *insights* (MCAFEE; BRYNJOLFSSON, 2012) que tem potencial para agregar valor às suas empresas (MANYIKA et al, 2011).

Como o mercado de adquirência é altamente competitivo, com empresas credenciadoras diariamente tomando ações para não só penetrar em segmentos não atendidos, mas ativamente buscar capturar clientes de seus pares, a utilização de dados verídicos inviabilizaria o estudo por abrir vantagens competitivas de empresas reais. A base de dados utilizada foi, então, modelada a partir de conceitos padronizados do segmento, que são publicados trimestralmente pelas empresas de capital aberto que atual no setor, mas os clientes e os valores foram gerados de modo aleatório, dentro de parâmetros pré-definidos. A empresa Cielo, líder de participação de mercado no Brasil, não somente em volume, mas em abrangência de atuação geográfica, foi a escolhida para ser o modelo para a criação da base de dados, mas também foram incorporadas algumas características de perfil das empresas Stone Co. e PagBank (anteriormente PagSeguro), que também possuem capital aberto e realizam divulgação de resultados e abertura de dados seletos ao mercado.

O setor de Meios de Pagamento foi por muito tempo caracterizado por uma estabilidade nas bases de clientes. Como praticamente todos os players do mercado estavam intimamente relacionados ou faziam parte de Instituições Bancárias, o relacionamento comercial com os seus respectivos clientes era uma extensão da relação comercial do cliente com o banco. Somente a partir de 2016/2017, com o surgimento de novos players independentes, organizados mais como fintechs que como bancos tradicionais, que o mercado foi profundamente abalado. Os grandes players incumbentes (Cielo, ligada ao Banco do Brasil e Bradesco, e Rede, ligada ao Itaú) começaram a rapidamente perder participação de mercado para os novos entrantes, marcando uma rápida mudança do status quo para um período de alta volatilidade, que só seria freado pelo esfriamento econômico que ocorreu em decorrência da pandemia da COVID-19.

Figura 1: Entrada de Novos Players Pressionando Incumbentes




Fonte: VALOR ECONÔMICO, 2019

Outro conceito importante no setor de Meios de Pagamento é o ARV, ou Antecipação de Recebíveis de Vendas. Estas operações são extremamente comuns e importantes no ciclo operacional no ecossistema de Meios de Pagamentos, sendo uma fonte de receitas financeiras para as empresas adquirentes e uma fonte barata de capital de giro para os credenciados, mais barata que outras fontes de crédito geralmente disponíveis. Como as operações de empréstimo estão atreladas a recebíveis de operações já concluídas, o risco de inadimplência cai consideravelmente, sendo substituído pelo risco de Chargeback, ou distrato das transações por vontade do cliente, geralmente por suspeita de fraude ou desacordo comercial. O tratamento do risco de Chargeback é feito pela própria adquirente em parceria com o Instituidor do Arranjo (bandeira). Algumas ofertas comerciais pressupõem a obrigatoriedade da contratação de ARV, que ocorre de forma automática, já creditando na conta de domicílio de pagamento escolhida pelo cliente em um ou dois dias úteis após a transação. Este é um grande diferencial competitivo, pois com a certeza da contratação da operação, a empresa Adquirente pode mexer com as margens e baixar taxas operacionais dos clientes, tornando ela como opção mais barata e mais atrativa frente à concorrência.

Como iremos ver, o churn é um fenômeno inevitável, pois não ocorre somente por vontade da empresa, havendo situações impossíveis de serem evitadas. Contudo, através de data analytics, pode ser profundamente estudado e, por aplicação correta de modelos preditivos estatísticos, mitigado, gerando um impacto econômico positivo para a

empresa. Com uma situação saudável de churn no modelo econômico, além das receitas não perdidas com os clientes atuais, a empresa pode se dedicar mais à prospecção de novos clientes para adição à carteira e o desenvolvimento de novos produtos e soluções para continuar servindo aos seus clientes.

Este estudo busca mergulhar na base de clientes da empresa Adquirente, criando classificações para primeiro poder compreender os perfis de clientes existentes em sua carteira atual, se estão condizentes com a avaliação atual da companhia, através de diversas variáveis cadastrais e operacionais dos seus mais de 300 mil clientes observados ao longo de um ano. Por se tratar de informações de alto valor estratégico, foi criada uma base de dados fictícia, mas cujos parâmetros estão todos em linha com o mercado, com distribuições de valores dentro de intervalos coerentes. Então, utilizando-se de técnicas e algoritmos de análise preditiva rodando em bibliotecas da linguagem *Python*, busca testar as variáveis em diferentes modelos escolhidos para verificar a criação de um índice preditivo de churn por cliente, que poderá ser então utilizado para viabilizar uma estratégia de retenção de clientes e de gestão do churn da empresa. Serão utilizadas duas técnicas distintas para também verificar qual modelo possui maior aderência aos preceitos elencados e que melhor poderá ser utilizado com os dados disponíveis.

# 2 OBJETIVOS

## 2.1 Objetivo Geral

O objetivo geral do trabalho baseia-se na proposta de pesquisa, sendo essa: desenvolver um modelo de tratamento preditivo de abandono de clientes (churn) para uma empresa facilitadora de meios eletrônicos de pagamento (adquirente/credenciadora)

## 2.2 Objetivos Específicos

Para a compreensão completa do objetivo geral, são formados objetivos específicos ao estudo. Sendo assim, os primeiros objetivos específicos os seguintes:

- Organizar os dados operacionais da empresa em um dataset compatível com os modelos preditivos;
- Identificar e mensurar os clientes que abandonaram a empresa, mês a mês;
- Aplicar diferentes algoritmos estatísticos (Random Forest e XGBoost);
- Apurar qual modelo estatístico possui melhor índice de acerto (utilizando Métricas de Matriz de Confusão) para o dataset;
- Justificar a escolha por um modelo analítico específico para aplicação;

# 3 REVISÃO DA LITERATURA

Por meio da revisão teórica, busca-se contextualizar os temas abordados neste trabalho e fornecer subsídios para o entendimento dos dados de pesquisa discutidos. Para a melhor compreensão do estudo, neste capítulo há um breve panorama conceitual da literatura disponível sobre Churn, buscando definir e exemplificar a terminologia e suas subdivisões, quando possível focando tanto em exemplos abrangentes, como em específicos do ramo da adquirência. Também visa-se explorar os conceitos de Big Data, Machine Learning, análise preditiva e algoritmos utilizados na Academia e no mercado, todos fundamentais para o processo operacional de Gestão de Churn.

## 3.1 Churn

Em um cenário econômico altamente competitivo, como é o setor de adquirência, especialmente após o período não ironicamente denominado "Guerra das Maquininhas", torna-se imperativo o trabalho de gestão e manutenção da base ativa de clientes de qualquer participante do mercado. Uma falta de gestão torna a carteira de clientes altamente vulnerável a ataques por competidores, busca por parte dos clientes a alternativas mais compatíveis ao seu porte (ou ao menos mais baratas), perda de clientes por quebra de equipamentos e falta de atendimento operacional e logístico apropriado. Tal volatilidade gera pressão nas margens financeiras e impacta o resultado operacional da empresa, fazendo tornando necessário o esforço de captura de novos clientes para repor àqueles perdidos.

Com os altos custos de aquisição de clientes, via marketing, prospecção, parcerias com agentes vendedores externos, publicidade e promoções, o custo de manutenção de clientes existentes na base ativa de uma empresa é demonstradamente menor que o custo de aquisição de um novo cliente do mesmo porte (KEAVENEY, 1995). Para viabilizar tal ação em larga escala, para uma empresa de grande porte, são necessários sistemas de informações gerenciais (OLIVEIRA, 2012) e uma estratégia corporativa para manter ativo o relacionamento com clientes existentes, estudando a sua mortalidade natural e também os motivos que podem levar o cliente a abandonar a empresa em favor de empresas competidoras ou produtos alternativos. Este fenômeno é conhecido na indústria pelo termo *churn*.

Churn é o ato de um cliente encerrar seu vínculo com uma empresa com a qual possui algum relacionamento econômico, tanto em favor de uma concorrente quanto em

favor de um outro produto ou tecnologia. Também chamada de taxa de abandono, taxa de atrito ou taxa de mortalidade, a taxa de churn é o percentual da base de clientes de uma empresa que abandona os seus produtos/serviços ao longo de um período observado (LEJEUNE, 2011).

O churn é comumente dividido em duas grandes categorias: o churn voluntário e o churn involuntário. Apesar do nome "abandono de clientes", na literatura identificamos duas forças distintas para cada tipo de churn. No caso do churn involuntário, o rompimento da relação parte da empresa para o cliente (HADDEN et al., 2007), quando o cliente viola algum acordo comercial ou deixa de efetuar o pagamento pelos serviços. No ramo dos meios de pagamento, os casos mais comuns de churn involuntário são os de fraudes de pagamento, quando o cliente é conivente ou participa ativamente em práticas ilegais, como clonagem de cartões, falta de uso dos equipamentos, especialmente para clientes de menor porte e pessoas físicas, e ainda por falta de pagamento, seja de mensalidades, alugueis de equipamentos ou juros sobre o montante de crédito tomado, sendo este último menos comum por ainda ser uma prática recente ao mercado.

Já para o churn voluntário, a força motora advém dos clientes, que optam, por vontade própria, em cessar o seu relacionamento econômico com a empresa e não utilizar mais os seus serviços (HADDEN et al.,2007). Ele ainda é subdividido em churn deliberado e acidental. Quando o churn é chamado de deliberado, entende-se que o cliente tomou uma decisão em abandonar o fornecedor do serviço ou produto por questões econômicas quando vê vantagem em deixar de usufruir ou maior vantagem em outro fornecedor ou por questões de relacionamento com a empresa, como insatisfação com os serviços prestados e má percepção com o atendimento pós-venda. Já para o churn acidental, variáveis ambientais fora do controle imediato dos dois agentes são dadas como as forças motores, como a mudança de endereço do cliente para uma localidade não atendida pela empresa, precarização das condições econômicas do cliente (como perda de emprego) que o forçam a encerrar o produto.

Embora seja inevitável, em qualquer ramo, o eventual abandono de clientes, a principal consequência do churn, para a empresa, é a perda de receitas. Portanto, é um grande interesse para a organização ter uma estratégia para retenção ou reposição de clientes. As ações e estratégias empregadas por uma organização para o tratamento de churn são chamadas de Gestão de Churn (LU et al., 2014).

Em uma abordagem dirigida, os clientes que serão impactados são conhecidos pela companhia. Eles são previamente selecionados e especificamente impactados com

a ação da empresa, seja por qualquer canal de relacionamento escolhido, que geralmente é definido em escopo pelo porte do cliente. Clientes de maior porte tendem a receber um tratamento mais direto com um vendedor ou agente de relacionamento especializado, capaz de oferecer mais benefícios e uma maior margem de negociação de condições comerciais. (MALHOTRA; BAZERMAN, 2008). Para poder dirigir as ações em um nível de larga escala, quando estamos tratando de milhares de clientes potenciais para impacto, no entanto, é necessário ter um trabalho prévio de CRM (ULLAH et al., 2019) ou haver indícios que o cliente possui um perfil compatível com o perfil de cliente com risco de abandono identificado pela organização. Geralmente, envolve alguma oferta ou condição comercial diferenciada como incentivo para não se concretizar o abandono (BUREZ; POEL, 2008), perene ou por tempo limitado, com o *upsell* de uma oferta mais robusta. Entretanto, caso o cliente não seja avaliado de forma correta, a empresa corre o risco de baixar mensalidade ou taxas de um cliente que está atualmente satisfeito com o serviço, gerando uma perda de receitas desnecessárias que podem gerar pressão para aumento de receitas ou redução de custos em outras áreas da companhia.

Já as estratégias não dirigidas são feitas para um público mais amplo, de forma massificada, como e-mail marketing, SMS ou publicidade em veículos de mídia. Como riscos (TSAI; LU, 2009). Tais ações comumente são acompanhadas de uma campanha de mídia visando atrair o cliente com benefícios especiais ou novas condições comerciais mais vantajosas. Como são normalmente direcionadas a clientes de menor porte e em grande número, o que inviabiliza financeiramente um contato mais direto com um agente de vendas, podem fazer uso de vantagens financeiras, pois, havendo menor poder de barganha e normalmente fazendo uso de produtos padronizados *balcão*, estes clientes possuem maiores *spreads* e um espaço para comprimir margens enquanto ainda se mantém lucrativos (FAULKNER; MORGAN, 2016). Como risco, também há a possibilidade (inevitabilidade, diga-se de passagem) que serão impactados indevidamente clientes que não estavam com risco de churn, mas que podem sentir-se incomodados com o contato impessoal ou ainda que podem notar que suas condições atuais não são tão atraentes quanto à concorrência, gerando o efeito contrário ao desejado pela ação.

Tendo em vista que o custo de aquisição de novos clientes (CAC) é, em grande parte dos casos observados, maior que o custo de manutenção de clientes já existentes (VERBEKE et al., 2012; HADDEN et al.,2008), torna-se economicamente favorável para a organização ter um tratamento para amenizar a sua taxa de churn. A aquisição de novos clientes pode envolver custos operacionais, marketing e propaganda, validações de fichas

cadastrais, custos que podem exceder em ordens de magnitude o custo que a mesma organização teria em manter um cliente que já está em sua base atual (KEAVENEY, 1995). Uma forma de fazer esse tratamento interno é tentar prever quais clientes possuem uma maior propensão ao abandono e já realizar um trabalho de retenção antes que o churn seja concretizado (NESLIN et al., 2006).

Na literatura, há dois tipos de abordagens de clientes em gestão de churn: abordagens reativas e abordagens proativas (BUREZ; POEL, 2008). Uma abordagem reativa é considerada a prática mais comum: a organização aguarda o cliente cancelar o serviço para oferecer uma contraoferta para evitar o abandono. Esta é uma prática comum em empresas de telecomunicações e de TV por assinatura (COUSSMENT; DEN; POEL, 2009). O benefício desta forma é ausência de um processo dedicado para a retenção de clientes: a empresa é notificada do abandono pelo cliente e somente então toma alguma ação para o reter. Não há esforço ou custo interno, fora algum esforço de contato e um incentivo oferecido ao cliente para que permaneça em relação comercial. Em contrapartida, há um risco mais elevado de o cliente já ter contratado outro serviço substituto ou estar indisposto a aceitar a contraoferta, impossibilitando o processo de retenção.

A outra abordagem, proativa, também possui riscos e benefícios particulares. Como benefícios, ela já tende a selecionar clientes rentáveis para a empresa que possuem indícios de abandono, através de modelos preditivos de churn. Clientes pouco rentáveis ou com risco de não poder honrar seus pagamentos tendem a não ser alvo dessa ação. Contudo, para haver este tipo de ação é necessário que o modelo preditivo tenha acurácia e precisão: falsos positivos acarretam em desembolsos desnecessários para clientes que não iriam abandonar a empresa (BUREZ; POEL, 2008), como negociação de taxas mais baixas para clientes sem nenhuma tendência de migração para a concorrência. Ainda, há, no geral, um esforço da organização para desenvolver e operacionalizar este processo de avaliação e contato com os clientes, que também possui custos por si só.

## 3.2 Big Data e Algoritmos Preditivos

Quase todas ações humanas geram algum tipo de informação, algum tipo de dado que pode ser observado. Com a digitalização da sociedade, estes dados são passíveis de captura e registro, muitas vezes sem a ciência do próprio indivíduo autor da informação. Atualmente, a maior parte desses dados é gravada em algum modo e pode ser acessada eletronicamente. Fazer ligações telefônicas, pagar contas, utilizar um passe para andar

de ônibus e usar cartão de crédito criam pontos de dados que podem registrados, armazenados, categorizados, e, principalmente, analisados para entender os padrões de uso dos seus agentes. Para começar a tirar qualquer sentido do conjunto de gigantescas massas de dados gerados a cada instante, precisamos fazer uso do Big Data (MANYIKA et all, 2011).

Big Data é um conceito que possui uma definição abrangente, mas normalmente há o consenso de se tratar de gigantescos volumes de dados que estão além da capacidade humana de serem compreendidos por meio de ferramentas tradicionais, ou seja, sem ferramentas eletrônicas específicas para leitura, armazenamento e análise de dados.

O aumento no interesse em Big Data é facilmente evidenciado em companhias que trabalham em ambientes complexos e com serviços eletrônicos (WANG et al, 2016), mas suas aplicações vão muito além dos setores de serviço. Apesar do já popularizado uso em setores como finanças e telecomunicações, mais comumente associados com Big Data, o uso também serve a fins diversos como saúde, ao analisar dados de milhões de pacientes para estudar doenças, sociologia, estudando fenômenos sociais, muitas vezes ligados a mídias sociais, e mobilidade urbana, como o desenvolvimento de aplicações que permitem compartilhamento de veículos (WU et all, 2014).

Academicamente, há caracterização de Big Data nos chamados *5 Vs*: volume, variedade, velocidade, variabilidade e veracidade. Originalmente cunhado por LANEY (2001) como os 3 Vs, à medida que o campo de ciência de dados foi expandindo de popularidade e usuários, mais dois Vs foram adicionados, demonstrando a expansão no entendimento do conceito.

### 3.2.1 Os Cinco Vs de Big Data

Volume: Talvez a característica mais notável e reconhecida, a gigantesca quantidade de dados criados a cada instante é o principal ponto focal de Big Data. Trabalhar com Big Data envolve utilizar ferramentas específicas para ler e categorizar dados que não são possíveis de serem sequer mensurados de outra forma. A capacidade de armazenar e utilizar os grandes volumes de dados está entre os principais desafios das organizações que buscam trabalhar com Big Data, pois envolve grandes investimentos em processamento de dados e em profissionais capacitados para fazer uso das informações (MANYIKA et all, 2011).

Variedade: Representa a diversidade dos tipos de informações a serem estudadas.

Por isso entendemos a quantidade de diferentes variáveis que estão sendo observadas ao mesmo tempo, em paralelo, como também ao tipo e formato de informação em foco. Informações estruturadas e não-estruturadas são geradas simultaneamente e passíveis de estudo, mas necessitam de diferentes técnicas para mensuração (LANEY, 2001). Variáveis booleanas, numéricas, textuais, visuais, audíveis, sensoriais, discretas e contínuas, todas são passíveis de intersecção, mas dependem do correto tratamento para haver algum sentido nos dados.

Velocidade: Dados são gerados a cada fração de segundo, muitas vezes em tempo real, como pode ser evidenciado pela proliferação no uso de plataformas de streaming, como Twitch e Youtube (FORD et al, 2016) ou aplicativos de mobilidade urbana, como Google Maps ou até mesmo Uber. Velocidade também se refere à leitura destes dados, que precisa ser feita de maneira o mais próximo possível de instantânea, pois necessita acompanhar o ritmo de evolução da criação dos dados, possibilitando consultas que irão pesquisar os volumes imensos de dados (ZIKOPOULOS et al, 2012). Ainda, a velocidade se refere ao prazo de obsolescência dos dados: à medida que novos dados mais atuais já estão sendo gerados, a validade e utilidade dos dados históricos vai perdendo relevância.

Variabilidade: Algumas vezes confundida ou agrupada com a variedade, a variabilidade está ligada à ideia de diversidade interna das próprias variáveis, o quanto uma variável tem de diferença de outros pontos de entrada da mesma variável. Esta diversidade indica que não somente a quantidade de informações diferentes tem valor, mas também que a mudança das próprias características internar e divergentes da observação gera por sua vez novas análises de informação (LANEY, 2001).

Veracidade: Trabalhar com uma quantidade gigantesca de dados complexos com diferentes características e que são gerados a cada segundo requer um extremo cuidado quanto à autenticidade das informações que estão sendo observadas e das conclusões que estão sendo tiradas em cima destas. A certeza, tanto na confiabilidade dos dados observados, quanto a acurácia das análises geradas das consultas geradas com eles, especialmente no que envolve diferentes fontes e formatos de dados, é o quesito de muita preocupação na utilização de Big Data (ZIKOPOULOS; EATON, 2011).

### 3.2.2 Análise Descritiva

O primeiro passo de qualquer análise baseada em dados, ou ainda, o primeiro *momento* em que qualquer organização que trabalha com dados se encontra, é denominado

de Análise Descritiva. Nesse estágio, o foco está em mensurar, apurar e resumir os grandes volumes de informação a uma maneira em que possam ser compreendidos. Antes de qualquer grande empreitada de uso de dados em Machine Learning, um importante passo é ter a noção de quais dados estão disponíveis para serem utilizados e como eles descrevem um fenômeno já existente. A Análise Descritiva serve como alicerce para análises mais profundas, como a Análise Preditiva, além de propiciar entendimento sobre o universo dos dados que estão disponíveis. Somente após ter o entendimento (e comprovação) da existência do fenômeno a ser estudado, e a sua mensuração, é possível atuar de forma reativa para tomar alguma ação desejada, ou registrar os eventos para alguma ação futura.

Como o estudo será efetuado de forma multivariada, ou seja, diversas variáveis serão exploradas e testadas para verificar sua relação e como explicam o fenômeno de Churn, é importante compreender os tipos de dados e variáveis existentes.

#### *3.2.2.1 Tipos de Dados*

TESSAROLO e MAGALHÃES (2015) classificam os dados em três tipos principais: Dados Estruturados, Não Estruturados e Semi-Estruturados.

Dados estruturados são dados de bancos relacionais, geralmente estruturados em matrizes (por isso assim denominados) em linhas e colunas, facilitando a sua utilização em bancos relacionais. Praticamente todas organizações trabalham, de alguma forma, com dados estruturados em planilhas (comumente Excel ou variantes open source) em diferentes estágios de sofisticação de análise, mas com uma lógica geral similar de disposição em linhas e colunas. Outros exemplos seriam sistemas de ERP, bancos de dados corporativos, como Bancos Oracle e Microsoft ODBC, ou planilhas de inserção manual de dados para fins diversos.

Dados não estruturados são documentos, imagens, áudios, vídeos, streaming, transcrições ou qualquer outro tipo de dados não codificados em registros com uma estrutura padrão. Geralmente são dados que demandam um processamento mais refinado e laboroso para o seu uso e de difícil relação com outras fontes de dados. A menos que a organização em questão tenha um processo específico para tratamento destes dados, tendem a serem utilizados de forma mais individual ou com tratamento manual por seus usuários.

Dados semi-estruturados são uma intersecção dos dois tipos acima; possuem apenas uma parte da sua informação codificada de forma estruturada, existindo em um meio termo relacional. Geralmente possuem uma estrutura (ou *framework*) estruturado, mas o conteúdo dentro desta estrutura está não estruturado. Como exemplos mais comuns,

podemos ver arquivos em formato XML ou RDF.

#### *3.2.2.2 Tipos de Variáveis*

Uma variável pode ser entendida como um conjunto de resultados possíveis para um fenômeno específico, podendo tomar a forma de valores, características ou atributos que representam um ponto específico deste fenômeno. Como este trabalho propõe analisar uma grande quantidade de variáveis (a sua base consolidada possui 66 campos variáveis únicos para cada cliente), e a relação e forma de utilização dos dados possui uma dependência intrínseca com a sua própria natureza, faz-se necessária uma breve explicação sobre os tipos de variáveis categorizadas pela literatura estatística.

Variáveis qualitativas se referem à uma característica não numérica que busca ser analisada em seu contexto ou como se relaciona com outras características. Elas não podem ser mensuradas em termos de quantidades, mas podem possuir uma ordem natural e compreensível ou podem tratar de características discretas que às distinguem de outras de sua classe. Como exemplos podemos ver a cidade sede de cada empresa, o seu segmento de atuação, se o cliente já acusou churn ou não na base de dados observada.

Variáveis quantitativas são as variáveis numéricas, elas podem ser medidas, pesadas, mensuradas e comparadas em ordens de grandeza. Para a base estudada, temos como principais exemplos o valor de TPV a cada mês, assim como o TPV total e o TPV médio mensal, identicamente para o ARV, nas mesmas aberturas. O churn é uma variável qualitativa, mas é apontado derivadamente de uma variável quantitativa. Para cada mês, foi apurado o valor transacionado de cada cliente. Para cada mês, foi verificado se o valor é positivo ou nulo, tornando-se uma variável qualitativa de S/N para aquele mês. Quando constatado que o cliente ficou três meses consecutivos com o indicador N, ele se torna um churn.

### 3.2.3 Machine Learning

Machine Learning é o campo da ciência de dados dentro da Inteligência Artificial (IA) que almeja deixar os computadores executarem tarefas de forma autônoma a medida que vão aprendendo comportamentos através de padrões identificados em os seus dados processados. Em seu cerne, procura o aperfeiçoamento de algoritmos ou automações diretamente pelas máquinas, com o mínimo de interação humana. O seu nome vem

do paralelo comumente associado ao aprendizado humano, por associação, contexto e estudo.

Machine Learning pode ser também visto como uma evolução do Data Analytics possibilitado pela expansão e disponibilidade de Big Data, permitindo que colossais volumes de dados sejam diretamente alimentados em modelos especificamente preparados para evoluir a aprender de forma autônoma. Além das mais conhecidas aplicações nos setores Financeiro e Logístico, Machine Learning tem sido amplamente utilizado no setor Médico, utilizando dados de milhares ou milhões de pacientes para tentar encontrar padrões em diagnósticos, tratamentos e prevenção de doenças.

Atualmente, a aplicação que mais deve ser conhecida e certamente entrou no *Zeitgeist* é o ChatGPT, ferramenta desenvolvida pela OpenAI, baseada em redes neurais, que aprende a executar tarefas conforme interage com novos usuários (QUINTANS-JÚNIOR et all, 2023). Esta tecnologia, um largo passo na evolução dos chatbots, atingiu mais de 100 milhões de usuários com apenas dois meses de lançamento (REUTERS, 2023). Contudo, suas limitações começam a ficar mais evidentes ao passo que vai aprendendo comportamentos incorretos com usuários, replicando-os a usuários futuros. A Universidade de Purdue conduziu um estudo utilizando perguntas do popular site Stack Overflow, e verificou que o ChatGPT errou 52% das perguntas relacionadas à programação (REGISTER, 2023).

### 3.2.4 Análise Preditiva

A análise preditiva consiste no estudo estatístico de características observáveis, notadamente históricas, para encontrar correlações de comportamento e tentar gerar informações sobre tendências de eventos futuros (KUHN; JOHNSON, 2013). O seu uso está bastante associado ao processo de tomada de decisão, para o qual serve em papel de suporte à gestão, analisado em conjunto com outras ferramentas de análise de dados.

O processo preditivo começa o delineamento do problema de pesquisa, a definição de qual é o fenômeno que se procura explicar ou quais resultados ou quesitos são desejados alcançar. Cada decisão baseada em dados é única, com suas próprias características e restrições (PROVOST; FAWCETT, 2013), também necessitando de diferentes modos de tratamento de pesquisa. Cada tipo de problema irá demandar um tipo de ação e um método de pesquisa particular, cabendo ao pesquisador procurar o método correto para o seu problema de pesquisa. No caso de procurar similaridades de características

entre um grupo de pessoas ou clientes, uma pesquisa de clusterização pode ajudar a identificar padrões e comportamentos similares em grupos (clusters) analíticos. Caso queira, no entanto, tentar estimar ações ou valores, como é o problema de pesquisa propostos por este trabalho ("qual é a probabilidade que este cliente irá desligar seu relacionamento econômico com a minha empresa?"), um método de clusterização não irá trazer insights sobre a chance de abandono, mas uma pesquisa de regressão pode conseguir identificar uma propensão ao churn, caso encontre variáveis colineares.

Para entender esses dados, primeiro é preciso entender qual é o problema de negócio que se busca estudar, pois simplesmente jogar dados em um modelo analítico não irá trazer previsões, apenas correlações (PROVOST; FAWCETT, 2013). Após ter um problema de pesquisa definido, é passado para o estágio de compreensão das informações que estão sendo escolhidas e suas intersecções. Diferentes bases de dados relacionais podem ter formatos e fontes de informações diferentes, mas também podem possuir pontos de convergência nos quais os dados podem ser conectados e conhecimento gerado a partir de sua exploração.

Com o avanço dos campos de Analytics e Machine Learning, que por sua vez possibilitaram a proliferação de campos como Business Analytics fora da Academia e de grandes corporações, a análise preditiva surge como uma ferramenta de apoio à tomada de decisão baseada em dados que busca gerar uma maior precisão e vantagens competitivas a seus usuários (PROVOST; FAWCETT, 2013).

#### *3.2.4.1 Aprendizagem Supervisionada*

A aprendizagem supervisionada se preocupa em estudar um fenômeno para se procurar uma resposta a um problema conhecido. O pesquisador já está ciente do seu questionamento e procura um resultado ainda desconhecido, mas o alvo já é conhecido. "Qual é a probabilidade de um cliente virar churn?". "Quais produtos adquiridos por um cliente o tornam mais provável de adquirir um outro produto no portfólio da empresa?". Utiliza-se de algoritmos que irão treinar com dados existentes para buscar padrões e relações que expliquem o fenômeno esperado. Uma condição importante é que precisa já haver dados passíveis de explicar tal fenômeno procurado. Aquisição e mineração de dados classificados/estruturados, ou a ação de classificação e estruturação dos mesmos, muitas vezes é um grande investimento (no múltiplo sentido da palavra) para os cientistas de dados (PROVOST; FAWCETT, 2013). Os dois grandes grupos de aprendizagem supervisionada são a Classificação, que busca um objetivo categórico (geralmente binário),

enquanto a Regressão busca um objetivo numérico, mensurável.

#### *3.2.4.2 Aprendizagem não Supervisionada*

A maior diferença da aprendizagem não supervisionada para a aprendizagem supervisionada é que, neste caso, o objetivo de pesquisa é conhecido, mas o objeto ainda não está completamente conhecido ou categorizado. Clustering, a identificação de grupos de características em comum dentro da base de dados, pode mostrar aglomerados de clientes que possuem similaridades de comportamento ou de natureza, mas essas similaridades não necessariamente irão trazer alguma informação que possa ser utilizada pelo pesquisador. A análise não supervisionada pode servir como um primeiro passo na exploração de novas bases de dados ou em busca de insights em bases conhecidas, que por sua vez podem gerar um problema de pesquisa definido que poderá ser pesquisado por análise supervisionada, utilizando-se de metodologias diferentes. Uma outra aplicação que vem ganhando importância é o aprendizado por contexto, quando pesquisadores utilizam modelos que vão aprender sobre linguagem e imagens em bases de treino como documentos ou vídeos (HOFFMAN, 2001).

#### *3.2.4.3 Mineração de Dados*

Uma distinção muito importante no processo de exploração de dados é a diferenciação de mineração de dados para o uso dos dados minerados. Como já dito anteriormente, o dado por si só não traz nenhuma informação, ele precisa ser corretamente laborado para agregar algum valor. É comum que pesquisadores e organizações confundam os dois, especialmente em estágios mais iniciais de data science, mas (PROVOST; FAWCETT, 2013) salientam a necessidade de ter os dois processos distintos.

> There is another important distinction pertaining to mining data: the difference between (1) mining the data to find patterns and build models, and (2) using the results of data mining. Students often confuse these two processes when studying data science, and managers sometimes confuse them when discussing business analytics. The use of data mining results should influence and inform the data mining process itself, but the two should be kept distinct. (PROVOST; FAWCETT, 2013, p.25)

### 3.2.5 Análise Prescritiva

A análise prescritiva é o passo complementar da predição: após o reconhecimento de padrões e o entendimento das probabilidades de ocorrência de um fenômeno, são fei-

tas simulações e recomendações de ações a serem tomadas para os cenários de maior probabilidade, risco ou mais desejados. Ela está adiante da análise preditiva em termos de sofisticação e apenas empresas que possuem departamentos avançados de Analytics conseguem trabalhar neste nível. Por sua complexidade, também está além do escopo deste trabalho.

> While descriptive analytics answers 'what happened?' and predictive analytics addresses 'what will happen?', prescriptive analytics tackles 'how to make it happen?' (FRAZZETTO et al., 2019)

Um exemplo deste tipo de análise é o sistema de logística da Amazon: observando os padrões de compra e demanda e suprimento das cadeias logísticas, tenta-se prever possíveis gargalos, remapear rotas de entrega para ganhos de eficiência, estabelecer bases avançadas em lugares estratégicos, sempre visando otimizar toda a cadeia. Outro exemplo que pode ser averiguado na literatura, é a abordagem proposta por SUSNJAK (2023), do Prescriptive Learning Analytics Framework (PLAF), um modelo de duas etapas. A primeira, preditiva, trabalha com dados já refinados e tratados, fazendo a análise exploratória dos dados. Uma segunda etapa, a então prescritiva, lidaria com cenários *what if?* (e se?), criando cenários, utilizando os dados verídicos, para simular melhorias possíveis nos resultados já observados in natura.

### 3.2.6 Algoritmos

Machine Learning utiliza-se de algoritmos para fazer o treino e a validação da aprendizagem desejada para um estudo específico. Algoritmos nada mais são que instruções programadas com um propósito e parâmetros específicos, visando aprender ou explicar algum problema. Com a alteração de alguns hiperparâmetros, podem ser adaptados para mudar como processam cada iteração, o que pode vir a melhorar ou degradar o resultado de sua aprendizagem. A calibragem dos hiperparâmetros faz parte das tarefas executadas pelo cientista de dados. Um dos maiores problemas com o uso de algoritmos está na qualidade dos dados utilizados (SOKKHEY, P., OKAZAKI, T, 2020), que requer um trabalho de curadoria e tratamento dos dados antes de sua aplicação própria.

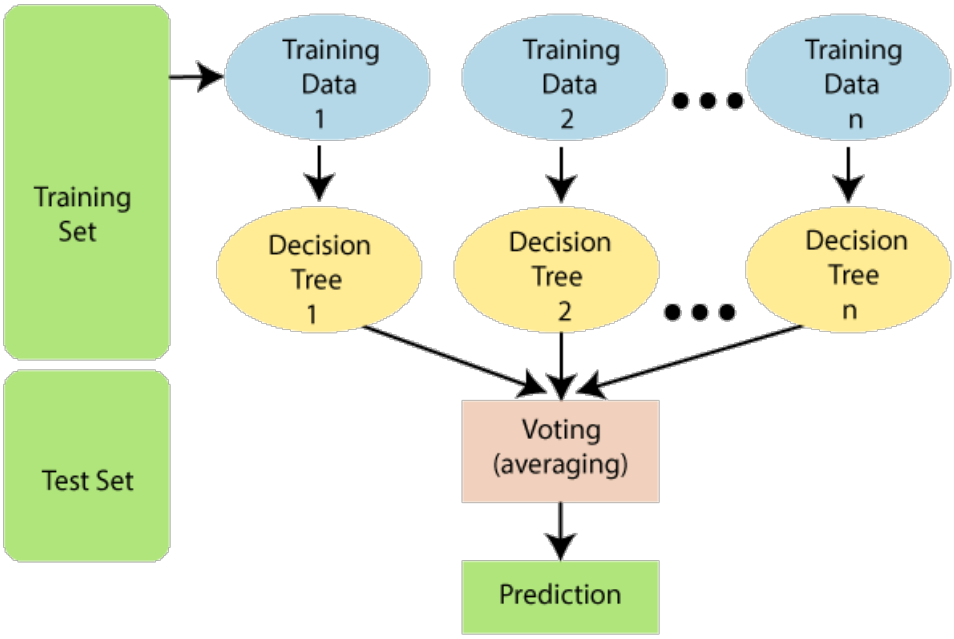
Atualmente, a grande parte dos algoritmos usados na pesquisa acadêmica podem ser utilizados como bibliotecas open source para as linguagens de programação mais populares, como Python e R. Para este estudo, utilizaremos dois algoritmos executados em Python, fazendo uso do aplicativo também open souce Jupyter Lab, que permite a criação de *notebooks* que facilitam a organização e execução de scripts.

*3.2.6.1 Random Forest*

Random Forest (Floresta Aleatória, por se tratar de diversas árvores) é um refinamento de árvores de decisão que utiliza amostras dos dados de treino para selecionar derivações de variáveis aleatórias no processo de indução das árvores de regressão (BREIMAN, 2001). O resultado final do modelo é uma média ponderada dos resultados de cada árvore de regressão, cada uma votando para a sua classe final. Em outras palavras, o modelo não utiliza todas as variáveis a cada iteração, mas sim uma seleção aleatória, permitindo uma maior dimensionalidade e fazendo com que o modelo consiga melhor se adaptar a dados omissos da base, uma ocorrência comum em bases cadastrais empresariais.

Essas duas características, a aleatoriedade dos dados utilizados e a ponderação dos resultados, contribuem para a estabilidade do modelo, ajudando a evitar problemas com *overfitting* (quando o modelo consegue prever com alta precisão os dados de treino, mas não consegue prever em novos dados).

Figura 2: Estrutura do Random Forest



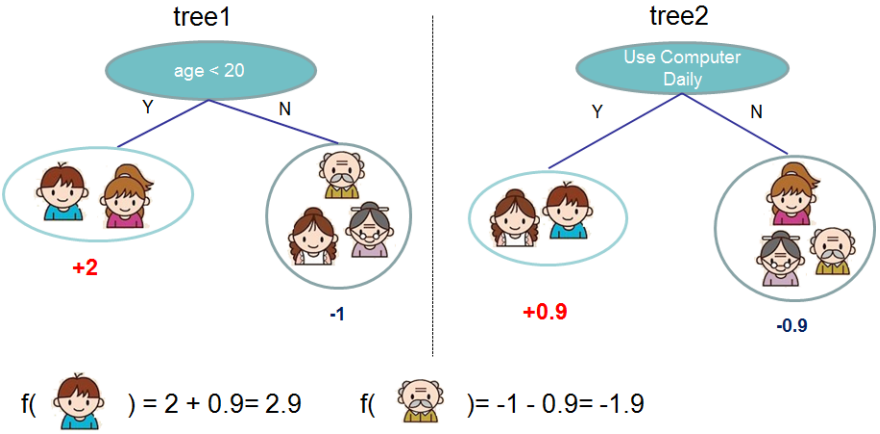

Fonte: javatpoint.com

O seu uso já disseminado no setor bancário, com diversos casos de uso disponíveis, além da facilidade do uso de sua biblioteca, o tornam um modelo atraente para a pesquisa proposta, sendo este escolhido como concorrente do XGBoost.

*3.2.6.2 XGBoost*

Extreme Gradient Boost é um algoritmo que faz uso de árvores de decisão usando também o algoritmo Gradient Descent, que por sua vez vai alcançando níveis ótimos em equações diferenciais (CHEN; GUESTRIN, 2016). A cada árvore de regressão, é mantido um score contínuo para cada folha, sendo este utilizado ao final para ponderar o resultado do modelo, assim evitando problemas de *overfitting*. Essa ponderação dos resultados, combinada com as características intrínsecas do desenvolvimento do algoritmo e a sua biblioteca open source, fizeram o XGBoost ser rapidamente disseminado e adotado como um modelo popular para predição de dados, ganhando várias competições do Kaggle (NIELSEN, 2016).

Figura 3: Estrutura do XGBoost




Fonte: xgboost.readthedocs.io

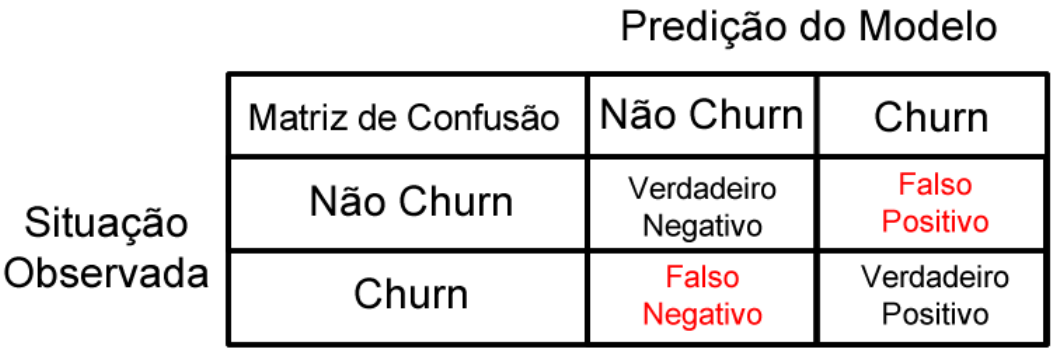
### 3.2.7 Avaliando os Modelos Preditivos

Como qualquer trabalho estatístico, os algoritmos de Machine Learning possuem métricas de avaliação da aderência do algoritmo aos dados em que está sendo aplicado. A avaliação das métricas de resultado é tão importante quanto a escolha dos modelos a serem executados. A principal destas é a Matriz de Confusão, que avalia os acertos do modelo em comparação aos erros tipo 1 e 2 (falsos positivos e falsos negativos, respectivamente) e as suas métricas derivadas. Estes serão os parâmetros de avaliação para a determinação do modelo escolhido a ser proposto por este trabalho.

*3.2.7.1 Matriz de Confusão e Métricas Derivadas*

O propósito principal de uma Matriz de Confusão é apurar como o modelo identificou a sua classificação prevista contra os dados já conhecidos, para mensurar o quanto está acertando e quanto está errando. Cabe então ao pesquisador decidir se o modelo está aceitável ou se necessita de alterações. Os quatro campos principais da tabela são:

Figura 4: Exemplo de Matriz de Confusão




Fonte: elaborado pelo autor

Verdadeiro Negativo: São as observações que o modelo classificou como negativas (neste caso, Churn = N) que realmente podem ser observadas como negativas na base de dados, com o modelo então acertando o resultado.

Falso Positivo: Também conhecido como Erro tipo 1 em Estatística, são observações que o modelo concluiu que eram positivas, mas que na verdade são observadas como negativas na base.

Falso Negativo: Também conhecido como Erro tipo 2 em Estatística, são observações que o modelo concluiu que eram negativas, mas que na verdade são observadas como positivas na base.

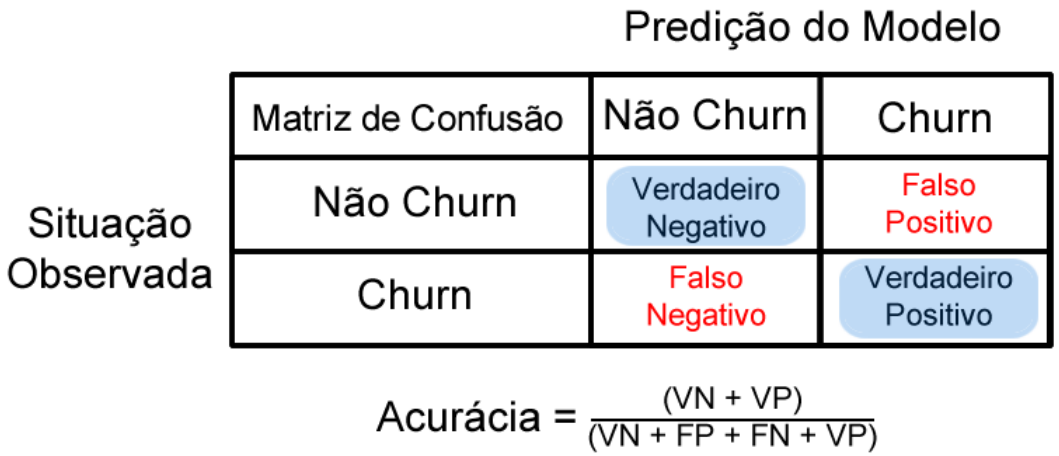
Verdadeiro Positivo: São observações que o modelo classificou como positivas (Churn = S) e, de fato, podem ser observadas como positivas na base de dados.

A partir destes campos, são verificados alguns indicadores de performance, a fim de efetivamente medir o resultado do modelo.

O primeiro deles, a Acurácia, testa quantas observações o modelo realmente acertou sobre o total de observações. O cuidado ao utilizar essa métrica é avaliar se o modelo não está sofrendo de *overfitting*, com o resultado de treino de validação positiva não balanceado entre verdadeiros positivos e verdadeiros negativos, dando uma falsa impressão de um bom resultado quando na verdade o modelo está sofrendo de falta de precisão.

A segunda métrica é a precisão, que apura a quantidade de apurações realmente

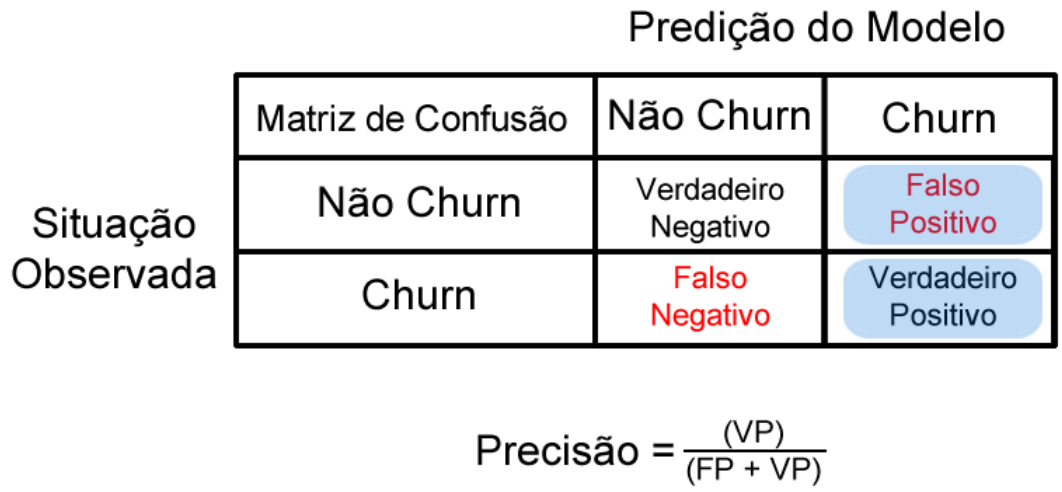
Figura 5: Exemplo de Matriz de Confusão




Fonte: elaborado pelo autor

verídicas sobre o total de predições verídicas.

Figura 6: Precisão



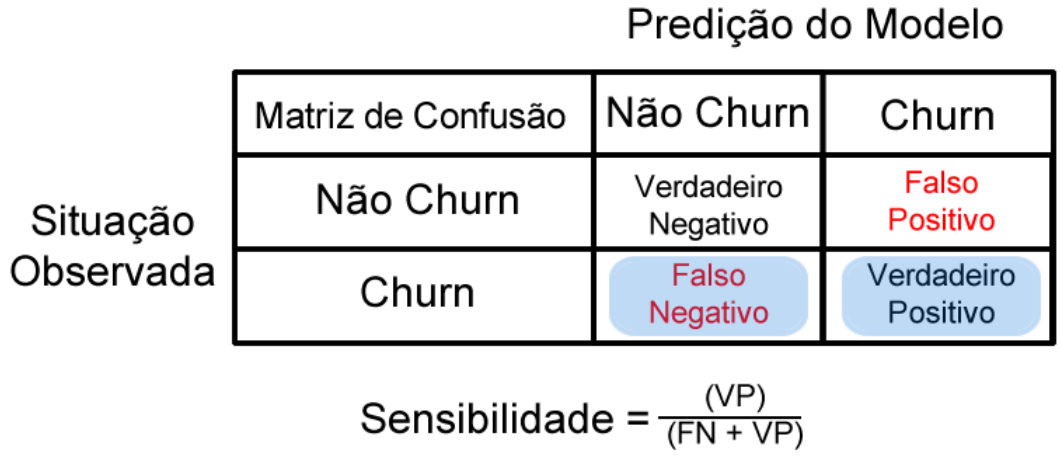

Fonte: elaborado pelo autor

A terceira métrica é o Recall, ou Sensibilidade, que mede a quantidade de resultados positivos sobre o total de observações verídicas

A Especificidade, é similiar à Sensibilidade, porém apura o total de observações negativas corretas sobre o total de observações negativas.

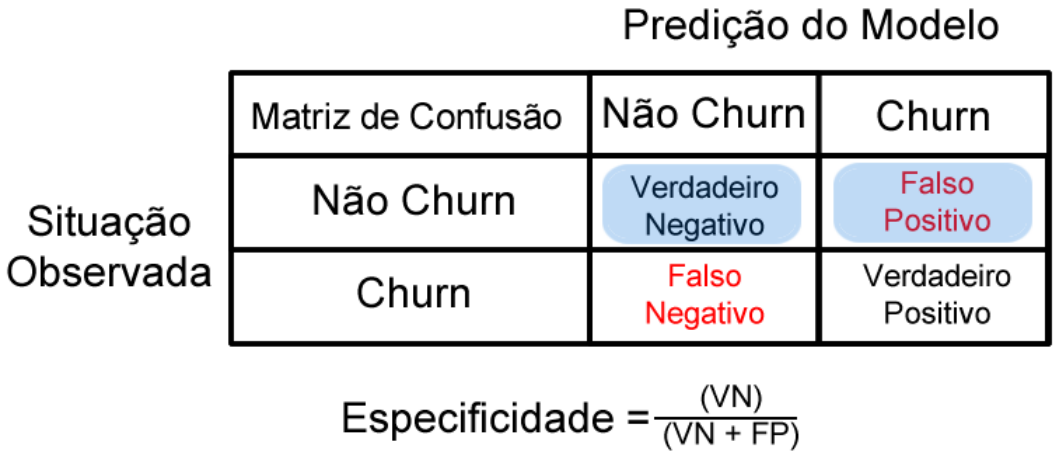
Em outros termos, a medida da Sensibilidade irá indicar quanto da base o modelo realmente consegue predizer o Churn, ao passo que a Especificidade irá medir quantos clientes o modelo irá acertar que não irão dar Churn. O risco de uma Sensibilidade baixa é a perda do cliente por não identificar corretamente que ele seria um desertor, já o risco envolvido em uma Especificidade baixa é que a empresa irá incorretamente agir sobre um

Figura 7: Sensibilidade




Fonte: elaborado pelo autor

Figura 8: Especificidade




Fonte: elaborado pelo autor

cliente que não iria abandonar a empresa, gastando esforços de maneira possivelmente desnecessária.

O próximo indicador é o F1-Score, que calcula uma média harmônica entre a Precisão e a Sensibilidade

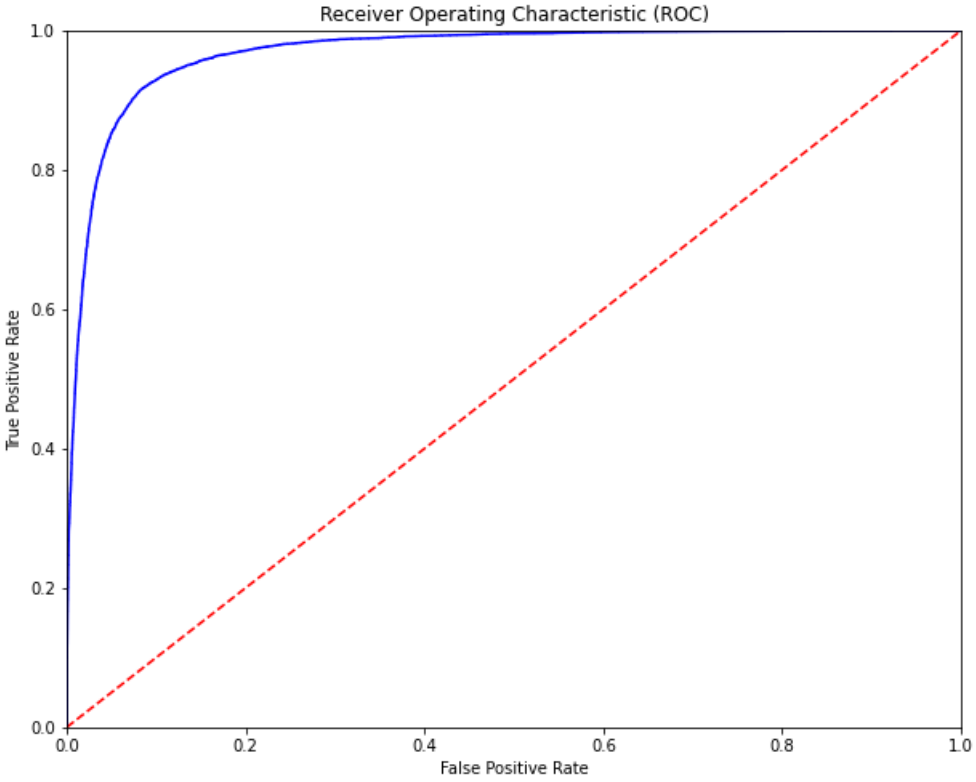
Por último, as curvas ROC (Receiver Operating Characteristic) e AUC (Area Under the Curve) plotam o Recall (Verdadeiro Positivo) contra o Falso Positivo em diferentes limites da amostra, permitindo encontrar pontos ideais para a utilização do modelo, maximizando a taxa de acerto. A Área Abaixo da Curva age em conjunto, como um valor absoluto de resultado do modelo. Quanto mais perto está de 1, melhor o resultado.

Figura 9: F1-Score

$$\text{F1-Score} = \frac{2 * (\text{Precisão} * \text{Sensibilidade})}{(\text{Precisão} + \text{Sensibilidade})}$$

Fonte: elaborado pelo autor

Figura 10: Curva ROC




Fonte: elaborado pelo autor

# 4 PROCEDIMENTOS METODOLÓGICOS

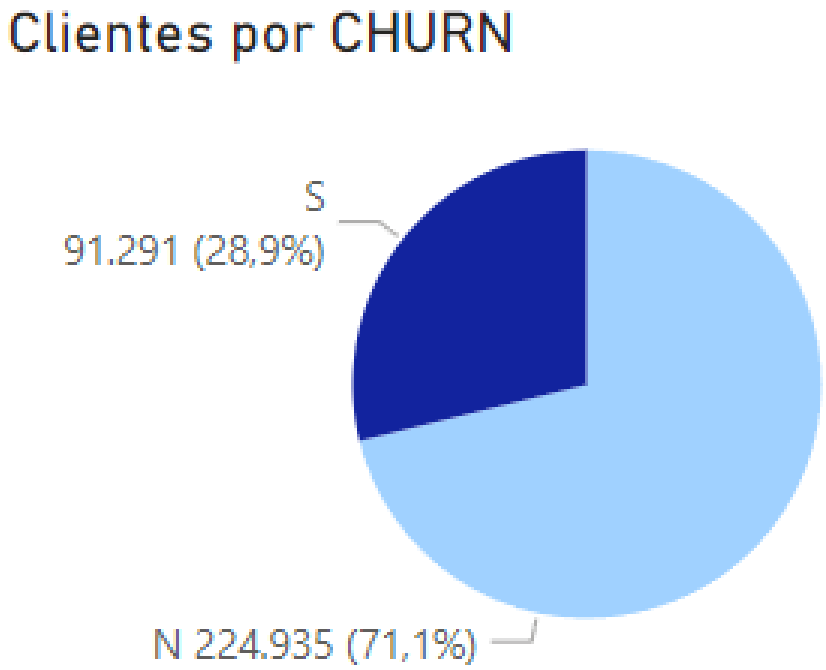
O problema de pesquisa proposto é a criação de um modelo de predição de churn, um indicador estatístico de o quão provável, baseado em características similares já observadas na base de clientes estudada, um cliente está de encerrar seu relacionamento comercial com a empresa Adquirente. Para tanto, as bases de dados (*datasets*) serão separadas, aleatoriamente, entre uma amostra de estudo (ou treino) de 70% dos clientes e uma amostra de aplicação (ou teste) de 30% dos clientes. Após a separação, serão aplicados dois algoritmos preditivos, XGBoost e Random Forest, escolhidos por serem já aplicados em outros estudos do setor, além de reconhecidamente apresentarem bons resultados quando utilizados em bases de dados dos setores Financeiro/Bancário.

Seguindo o modelo já utilizado de Churn Institucional da empresa Adquirente, padrão no setor de Meios de Pagamentos, no qual considera como churn o cliente que fica 90 dias sem nenhuma transação em seu POS, foi criado um marcador discreto de *IND-CHURN*, sendo esta uma variável textual S ou N, à qual é aplicada ao cliente após registrar 3 meses consecutivos sem valores transacionados. Internamente já são realizadas ações de *pré-churn* dentro da empresa, vendo recortes de tempo mais pontuais, como 10 ou 15 dias sem transações, mas dada a necessidade de haver informações mais padronizadas, e tendo conhecimento que há diferentes níveis de sazonalidade e frequência de transações pelos clientes, opta-se por fazer uso de um indicador institucional de prazo mais elevado.

Para poder montar um dataset compatível com a aplicação de algorítmos preditivos, foram geradas 3 bases de dados originais. Elas foram estruturadas de forma a deixar uma entrada de dados (chave) correspondente para cada cliente, a fim de que nenhum cliente apareça duas vezes com IDs diferentes. Para fins deste estudo, matrizes e filiais foram tratadas como entidades diferentes, sem chave de identificação ligando os dois entes. Para todos os parâmetros, foram inseridos limites máximos e mínimos de cada intervalo, a fim de possibilitar uma distribuição aleatória com o menor viés possível, mas ainda coerente com a realidade do mercado.

A primeira base, Cadastral, foi montada a fim de agregar o maior número possível de variáveis cadastrais que caracterizam os clientes (passados e presentes) da empresa Adquirente. Ela é composta de dados de natureza jurídica, localização geográfica, perfil de atuação de mercado, forma de captura dos clientes, data e tempo de credenciamento dos clientes, condições comerciais praticadas, quantidades de equipamentos que possui em aluguel, o tipo predominante de equipamento, o qual concentra o maior volume de

Figura 11: Quantidade Total de Clientes Churn




Fonte: Elaborado pelo Autor

transações e se ainda possui pelo menos um terminal apto a transacionar.

Para a distribuição geográfica, foram utilizados dados abertos do IBGE, com a população de cada Estado atuando como ponderador da quantidade de clientes a serem lá distribuídos. Para a separação entre entidades físicas e jurídicas, foi utilizada uma distribuição média do setor. Os dados de condição comercial (IND_OFERTA) indicam se o cliente já foi reenquadrado em uma nova oferta comercial, sendo esta o esforço de recálculo de preços sobre a base de clientes atual com novas condições de mercado, a fim de se tornar mais competitivo, além de quanto tempo está na base de dados já com a nova condição. As quantidades de equipamentos seguem uma lógica de média do mercado em relação ao porte dos clientes.

A segunda base de dados, Transacional, foi construída da mesma forma, mantendo a mesma chave individual para cada cliente, permitindo o cruzamento (e posterior agregação) de dados. Ela contém o volume transacionado (TPV) de cada mês, juntamente com um indicador de Ativo S/N para cada mês, a data da última transação, quantos meses esteve ativo, o TPV Médio de cada cliente e um indicador de Churn apurado, constantando que o cliente fechou 3 ou mais meses sem transações. Este indicador de Churn será a variável dependente do estudo, pois ela está definindo quais clientes concretamente deixaram de transacionar e acusaram Churn, o fenômeno que busca-se prever e mitigar.

A terceira e última base de dados, Financeira, contém os dados das operações de crédito de Antecipações de Recebíveis de Vendas (ARV). Para ARV, também foi elencado o valor total por mês, sendo apuradas variáveis como a existência de operações de ARV e , em caso positivo, a quantidade de meses com operações de ARV, valor médio contratado por mês, percentual contratado sobre o montante transacionado pelo cliente.

Com os dados arranjados em três bases estruturadas, foi feita a consolidação dos dados em um único dataset, para ser utilizado na aplicação dos algoritmos. O dataset foi agregado em 47 variáveis independentes e uma variável dependente (CHURN) para cada um dos 316.226 clientes. A base possui 29% dos clientes como CHURN = S, ficando compatível com a taxa média de churn do setor de 2,5% ao mês.

Dentre todas as variáveis observadas entre os três datasets, almeja-se identificar quais delas tem alguma relevância na decisão do cliente em realizar o abandono da relação comercial com a empresa Adquirente. Para tanto, será feito o uso de duas ferramentas de análise multivariada: XGBoost e Random Forest. As duas ferramentas serão estudadas por meio da linguagem *Python*, utilizando bibliotecas open source que fazem uso dos algoritmos mencionados.

Figura 12: Resumo dos Dados Qualitativos

| IND_PF | Clientes | %GT Clientes | Qtde Churn | Tx_Churn | TPV | %GT TPV | ARV |
|---|---|---|---|---|---|---|---|
| PJ | 268.664 | 84,96% | 67.586 | 2,96% | $148.453.225.905 | 99,68% | $27.000.779.009 |
| PF | 47.562 | 15,04% | 23.705 | 7,12% | $471.822.825 | 0,32% | $3.640.650 |
| **Total** | **316.226** | **100,00%** | **91.291** | **3,49%** | **$148.925.048.730** | **100,00%** | **$27.004.419.658** |

| NOME_CANAL | Clientes | %GT Clientes | Qtde Churn | Tx_Churn | TPV | %GT TPV | ARV |
|---|---|---|---|---|---|---|---|
| Agência | 157.723 | 49,88% | 39.775 | 2,97% | $86.867.353.608 | 58,33% | $15.567.096.636 |
| Hunter | 110.941 | 35,08% | 27.811 | 2,95% | $61.585.872.298 | 41,35% | $11.433.682.373 |
| Internet | 47.562 | 15,04% | 23.705 | 7,12% | $471.822.825 | 0,32% | $3.640.650 |
| **Total** | **316.226** | **100,00%** | **91.291** | **3,49%** | **$148.925.048.730** | **100,00%** | **$27.004.419.658** |

| NOME_PORTE | Clientes | %GT Clientes | Qtde Churn | Tx_Churn | TPV | %GT TPV | ARV |
|---|---|---|---|---|---|---|---|
| Microempresa (ME) | 91.468 | 28,92% | 23.172 | 2,99% | $54.671.975.507 | 36,71% | $4.958.050.968 |
| Empresa de Médio Porte | 86.143 | 27,24% | 21.684 | 2,97% | $45.875.694.883 | 30,80% | $10.918.106.897 |
| Empresa de Pequeno Porte (EPP) | 74.891 | 23,68% | 18.620 | 2,92% | $39.494.393.580 | 26,52% | $7.344.383.144 |
| Microempreendedor Individual (MEI) | 47.562 | 15,04% | 23.705 | 7,12% | $471.822.825 | 0,32% | $3.640.650 |
| Grande Empresa | 16.162 | 5,11% | 4.110 | 3,01% | $8.411.161.936 | 5,65% | $3.780.237.999 |
| **Total** | **316.226** | **100,00%** | **91.291** | **3,49%** | **$148.925.048.730** | **100,00%** | **$27.004.419.658** |

| NOME_SEGMENTO | Clientes | %GT Clientes | Qtde Churn | Tx_Churn | TPV | %GT TPV | ARV |
|---|---|---|---|---|---|---|---|
| COMÉRCIO E SERVIÇOS DIVERSOS | 66.288 | 20,96% | 28.441 | 5,78% | $11.116.483.990 | 7,46% | $1.349.986.983 |
| AUTOMOTIVO | 41.193 | 13,03% | 10.348 | 2,96% | $22.853.868.051 | 15,35% | $4.566.807.376 |
| FAST FOOD | 30.666 | 9,70% | 7.736 | 2,97% | $16.902.630.872 | 11,35% | $2.951.104.731 |
| CASA E CONSTRUÇÃO | 30.512 | 9,65% | 7.720 | 2,98% | $17.069.482.506 | 11,46% | $3.115.569.476 |
| ALIMENTAÇÃO | 30.491 | 9,64% | 7.627 | 2,94% | $16.824.871.555 | 11,30% | $3.045.436.869 |
| EDUCAÇÃO | 30.157 | 9,54% | 7.554 | 2,95% | $16.678.960.320 | 11,20% | $2.933.497.388 |
| COMÉRCIO E SERVIÇOS ESPECIAIS | 28.599 | 9,04% | 7.251 | 2,99% | $15.711.488.342 | 10,55% | $2.558.549.422 |
| ATACADO | 23.564 | 7,45% | 5.996 | 3,01% | $12.489.891.857 | 8,39% | $2.869.459.945 |
| FARMÁCIAS E DROGARIAS | 19.871 | 6,28% | 4.943 | 2,93% | $11.338.260.486 | 7,61% | $1.474.722.384 |
| POSTOS DE COMBUSTÍVEL | 14.885 | 4,71% | 3.675 | 2,90% | $7.939.110.750 | 5,33% | $2.139.285.084 |
| **Total** | **316.226** | **100,00%** | **91.291** | **3,49%** | **$148.925.048.730** | **100,00%** | **$27.004.419.658** |

Fonte: Elaborado pelo Autor

Figura 13: Resumo dos Dados Geográficos

| ESTADO | Clientes | %GT Clientes | Qtde Churn | Tx_Churn | TPV | %GT TPV | ARV |
|---|---|---|---|---|---|---|---|
| Minas Gerais | 52.339 | 16,55% | 15.120 | 3,49% | $24.692.794.606 | 16,58% | $4.482.013.266 |
| Rio Grande do Sul | 37.713 | 11,93% | 10.953 | 3,52% | $17.836.659.458 | 11,98% | $3.286.463.918 |
| São Paulo | 31.171 | 9,86% | 8.994 | 3,49% | $14.617.641.274 | 9,82% | $2.705.688.148 |
| Ceará | 25.622 | 8,10% | 7.377 | 3,49% | $11.912.775.284 | 8,00% | $2.166.077.365 |
| Bahia | 25.425 | 8,04% | 7.379 | 3,52% | $11.808.700.553 | 7,93% | $2.105.166.057 |
| Paraná | 23.219 | 7,34% | 6.785 | 3,54% | $10.760.711.838 | 7,23% | $1.972.670.120 |
| Santa Catarina | 14.032 | 4,44% | 4.007 | 3,45% | $6.702.932.900 | 4,50% | $1.180.306.704 |
| Pernambuco | 11.805 | 3,73% | 3.448 | 3,54% | $5.474.398.976 | 3,68% | $1.014.677.664 |
| Goiás | 9.743 | 3,08% | 2.763 | 3,41% | $4.629.319.764 | 3,11% | $819.468.868 |
| Paraíba | 8.904 | 2,82% | 2.556 | 3,47% | $4.274.097.584 | 2,87% | $753.185.001 |
| Rio de Janeiro | 8.808 | 2,79% | 2.482 | 3,39% | $4.116.811.086 | 2,76% | $747.488.849 |
| Espírito Santo | 8.401 | 2,66% | 2.470 | 3,58% | $3.984.957.660 | 2,68% | $665.763.259 |
| Pará | 7.678 | 2,43% | 2.216 | 3,51% | $3.590.822.256 | 2,41% | $676.164.060 |
| Mato Grosso | 7.676 | 2,43% | 2.216 | 3,48% | $3.621.661.648 | 2,43% | $704.656.642 |
| Maranhão | 7.164 | 2,27% | 1.981 | 3,33% | $3.430.567.877 | 2,30% | $577.054.990 |
| Piauí | 6.683 | 2,11% | 1.921 | 3,48% | $3.301.722.177 | 2,22% | $570.197.927 |
| Rio Grande do Norte | 5.774 | 1,83% | 1.688 | 3,55% | $2.742.916.885 | 1,84% | $546.544.175 |
| Mato Grosso do Sul | 4.892 | 1,55% | 1.399 | 3,45% | $2.358.990.158 | 1,58% | $418.262.076 |
| Tocantins | 4.757 | 1,50% | 1.329 | 3,37% | $2.334.560.071 | 1,57% | $424.185.798 |
| Alagoas | 3.504 | 1,11% | 1.059 | 3,69% | $1.640.612.364 | 1,10% | $256.319.523 |
| Rondônia | 3.065 | 0,97% | 835 | 3,27% | $1.510.375.159 | 1,01% | $253.509.569 |
| Amazonas | 3.012 | 0,95% | 875 | 3,52% | $1.361.681.728 | 0,91% | $243.495.037 |
| Sergipe | 2.536 | 0,80% | 767 | 3,70% | $1.127.704.385 | 0,76% | $233.720.345 |
| Amapá | 1.041 | 0,33% | 310 | 3,58% | $480.087.037 | 0,32% | $79.361.288 |
| Acre | 751 | 0,24% | 229 | 3,74% | $350.759.315 | 0,24% | $70.565.193 |
| Roraima | 483 | 0,15% | 127 | 3,18% | $248.240.172 | 0,17% | $48.726.872 |
| Distrito Federal | 28 | 0,01% | 5 | 1,98% | $12.546.514 | 0,01% | $2.686.944 |
| **Total** | **316.226** | **100,00%** | **91.291** | **3,49%** | **$148.925.048.730** | **100,00%** | **$27.004.419.658** |

Fonte: Elaborado pelo Autor

# 5 ANÁLISE DOS DADOS

Nesta seção, explicar-se-á como será feita a predição dos dados através da aplicação dos algoritmos selecionados, visando encontrar qual deles possui maior aderência ao dataset criado até aqui, assim como mensurar quantitativamente os seus resultados. Como já mencionado anteriormente, o objetivo deste trabalho é fazer a predição de churn dos clientes da empresa Adquirente, por meio de uma abordagem supervisionada. Como sugerido por (PROVOST; FAWCETT, 2013), a escolha do modelo utilizado (regressão de árvores) e os algoritmos utilizados foi feita a partir do entendimento que é uma metodologia já praticada e validada por outros pesquisadores no setor, tratando-se então de uma abordagem coerente.

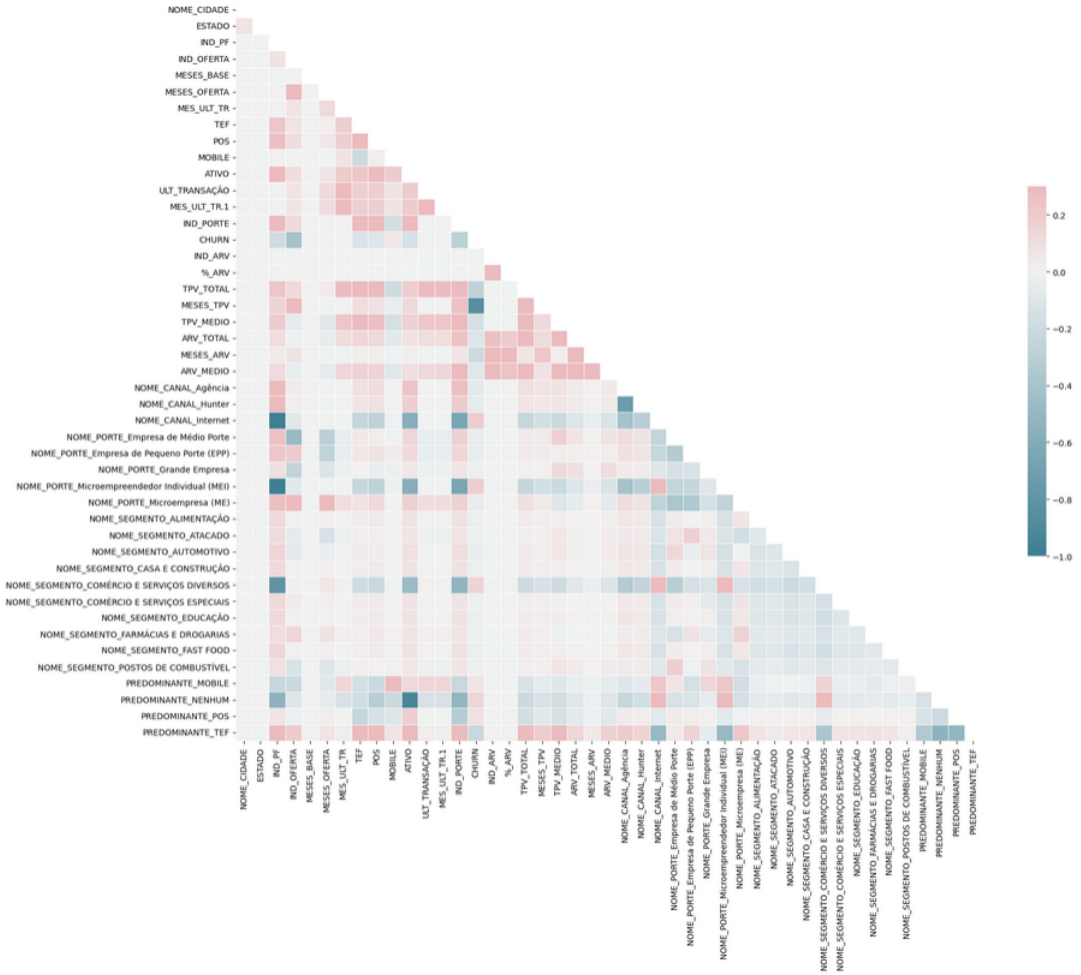
## 5.1 Preparação da Base para Treino

Apesar de, conceitualmente, o algoritmo XGBoost ser internamente balanceado para lidar com variáveis correlacionais (NIELSEN, 2016; CHEN; GUESTRIN, 2016), estudos testando bases de dados com variáveis correlacionais contra bases sem variáveis correlacionais, como ilustrado por VIKTESH (2021), sugerem que há sim um efeito no modelo sendo testado. Desta forma, para evitar problemas de *overfitting*, foram testadas as relações de correlação entre as variáveis da base e removidas as variáveis acima de 80%.

Uma vez normalizados os dados, faz-se a escolha da variável dependente, que para este estudo já foi apontada como a variável CHURN, que faz a devida marcação de abandono. Ela é então separada das demais variáveis e utilizada como parâmetro de correlação.

A próxima etapa é separar a partição dos dados que será direcionada para treino e a partição que servirá de teste. A base foi então repartida em 70% destinada para treino e aprendizagem do modelo e 30% destinada para teste e validação dos resultados. A base será idêntica para rodar ambos algoritmos preditivos, um pré-requisito para a validação do teste, evitando perda de informação devido a diferenças nas amostras.

Figura 14: Mapa de Calor de Correlação Entre as Variáveis



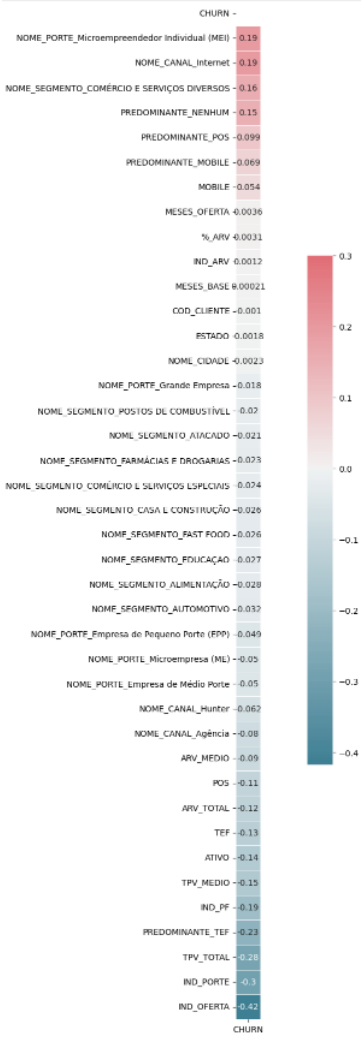

Fonte: Elaborado pelo Autor

### 5.1.1 Aplicação do Algoritmo XGBoost

O primeiro algoritmo a ser testado foi o XGBoost. Os hiperparâmetros foram ajustados tendo em mente uma taxa de aprendizagem mais conservadora, buscando evitar *overfitting* devido ao algoritmo corrigir e se adaptar rapidamente durante a criação de novas árvores, assim pesando o encaixe do modelo para o seu aprendizado a cada iteração. Um valor de 0,05 foi escolhido para a *learning_rate*. Para profundidade e o peso de cada nodo, foram testados valores de 5 a 20, sendo escolhido 15 pelo melhor resultado apresentado.

Os resultados do modelo foram, no geral, muito bons. O modelo conseguiu melhor prever os casos de Churn negativo que Churn positivo, o que não é exatamente ideal, o elevado peso de Falsos Negativos parece compensar a quantidade de Verdadeiros Positivos, mas a baixa quantidade de Falsos Negativos e Falsos Positivos confirma o bom desempenho.

Figura 15: Mapa de Calor de Correlação Entre as Variáveis e CHURN




Fonte: Elaborado pelo Autor

Figura 16: Separação da Base para Treino e Teste

```
#train and test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=56)
```

Fonte: Elaborado pelo Autor

Os indicadores de resultado tiveram números muito positivos.

A Acurácia ficou em 92%, indicando um grande índice de acerto das previsões.

A Precisão do modelo atingiu 93% para Churn N e 87% para Churn S, indicando que o algoritmo está corretamente prevendo os casos de abandono e não abandono para grande parte da base de dados.

A Sensibilidade teve o pior resultado de todo modelo, mas ainda consideravelmente positivo. Em 83%, há indícios de *overfitting* do modelo, mas o resultado ainda está aceitável. O risco de ter este indicador mais baixo é justamente o mais grave de todos, ter uma previsão incorreta de não churn para um cliente desertor.

A Especificidade do modelo teve o melhor resultado de todos, ficando em 95%,

Figura 17: Hiperparâmetros para o Algoritmo XGBoost

```
%%time
xgb_model = xgb.XGBClassifier(gamma = 0.001, learning_rate = 0.05, max_depth = 15, min_child_weight = 15, # scale_pos_weight=2,
                              n_estimators = 800, n_jobs=8, eval_metric='error').fit(X_train, y_train)

print('Accuracy of XGB classifier on training set: {:.2f}'
      .format(xgb_model.score(X_train, y_train)))
print('Accuracy of XGB classifier on test set: {:.2f}'
      .format(xgb_model.score(X_test, y_test))) #[X_train.columns]
```

```
Accuracy of XGB classifier on training set: 0.95
Accuracy of XGB classifier on test set: 0.92
CPU times: total: 14min 19s
Wall time: 1min 47s
```

Fonte: Elaborado pelo Autor

Figura 18: Resultados de Classificação do Algoritmo XGBoost

```
print(classification_report(y_test, y_pred))
```

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.93 | 0.95 | 0.94 | 67619 |
| 1 | 0.87 | 0.83 | 0.85 | 27249 |
| accuracy | | | 0.92 | 94868 |
| macro avg | 0.90 | 0.89 | 0.90 | 94868 |
| weighted avg | 0.91 | 0.92 | 0.92 | 94868 |

```
print('Área:', roc_auc_score(y_test, y_pred))
```

```
Área: 0.8908311388028118
```

Fonte: Elaborado pelo Autor

também reforçando os indícios de *overfitting* quando comparada à Sensibilidade.

O F1-Score teve um reflexo da discrepância da Sensibilidade, terminando em 85%.

Por último, o ROC AUC ficou em 0,89, reforçando o bom resultado do modelo.

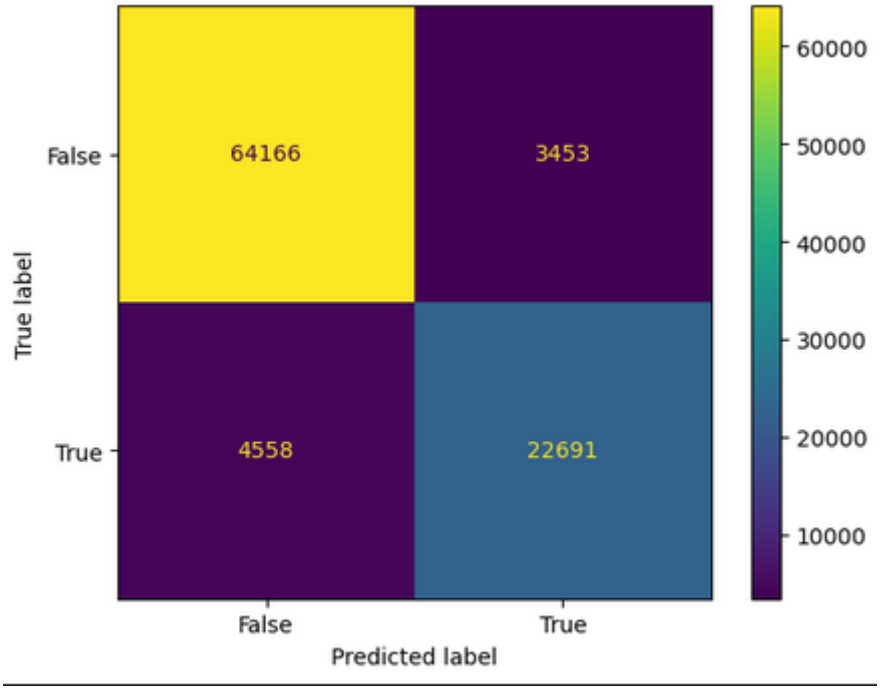
### 5.1.2 Aplicação do Algoritmo Random Forest

Como a proposta é fazer uma comparação justa entre os dois algoritmos, os hiperparâmetros do Random Forest foram colocados na maneira mais idêntica o possível aos do XGBoost.

Para profundidade de aprendizado do algoritmo foi deixada em 15, para ficar idêntica à sua contrapartida. O número de estimators também foi deixado em 800, com também 8 jobs paralelos.

Da mesma maneira que o XGBoost, os resultados apresentados pelo Random Forest foram muito bons. Contudo, logo de primeira vista, é possível notar um aumento nos

Figura 19: Matriz de Confusão do Algoritmo XGBoost




Fonte: Elaborado pelo Autor

Figura 20: Hiperparâmetros para o Algoritmo Random Forest

```
%%time
rfc = RandomForestClassifier(n_estimators= 800, max_depth=15, random_state=0, n_jobs=8)
rfc.fit(X_train, y_train)

CPU times: total: 3min 52s
Wall time: 29.5 s
RandomForestClassifier
RandomForestClassifier(max_depth=15, n_estimators=800, n_jobs=8, random_state=0)
```

Fonte: Elaborado pelo Autor

Falsos Negativos, o que é algo preocupante.

Assim como seu par, o modelo conseguiu melhor prever os casos de Churn negativo que Churn positivo, tendo a razão VP/VN um pouco abaixo, mas de maneira quase insignificante (0,353 para XGB contra 0,346 para RF). O desempenho geral ainda continua em um patamar muito bom.

A Acurácia ficou em 93%, também indicando um grande índice de acerto das previsões,

A Precisão do modelo atingiu 93% para Churn N e 91% para Churn S, indicando que o algoritmo está corretamente prevendo os casos de abandono e não abandono para maior parte da base de dados, ficando levemente acima do XGB.

A Sensibilidade também teve o pior resultado de todo modelo, mas em linha com o XGB. Em 82% de recall para Churn S comparado a 97% de recall para Churn N, realmente há fortes indícios de *overfitting* do modelo, mas o resultado continua em alta

Figura 21: Resultados de Classificação do Algoritmo Random Forest

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.93 | 0.97 | 0.95 | 157316 |
| 1 | 0.91 | 0.82 | 0.87 | 64042 |
| accuracy | | | 0.93 | 221358 |
| macro avg | 0.92 | 0.90 | 0.91 | 221358 |
| weighted avg | 0.93 | 0.93 | 0.92 | 221358 |

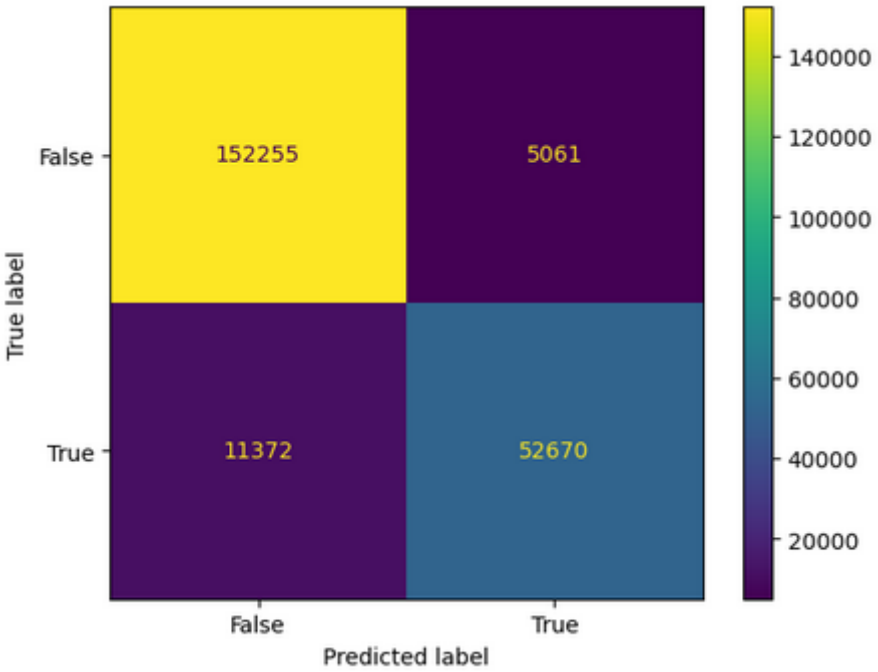
Fonte: Elaborado pelo Autor

performance. Novamente, o pior resultado ficou justamente no índice de maior risco, o de não acerto para um cliente desertor.

A Especificidade do modelo teve o melhor resultado de todos os indicadores apurados para os dois modelos, ficando em 97%. A maior discrepância em relação à Sensibilidade conta como um ponto negativo para o modelo.

Por fim, o F1-Score ficou em linha com a maioria dos indicadores, levemente abaixo pela ponderação negativa da Sensibilidade

Figura 22: Matriz de Confusão do Algoritmo Random Forest




Fonte: Elaborado pelo Autor

Pela natureza do Algoritmo, não há cálculo de ROC AUC para Random Forest. O ROC AUC avalia o ponto de corte do modelo para mudar o indicador para True ou False dependendo do hiperparâmetro escolhido, mas no caso do Random Forest, o valor sempre

retornará como True.

# 6 CONCLUSÃO

A avaliação do trabalho será feita contra os objetivos propostos em seu início.

- Organizar os dados operacionais da empresa em um dataset compatível com os modelos preditivos;

Este item foi cumprido na elaboração do Dataset Analítico, que foi agregado a partir dos 3 datasets originais, Cadastral, Transacional e Financeiro.

- Identificar e mensurar os clientes que abandonaram a empresa, mês a mês;

Este item também foi cumprido durante a elaboração do Dataset Analítico. Apurando o valor de TPV de cada cliente a cada mês, criou-se uma variável dicotômica de S/N para considerar o cliente como um ativo. Após 3 meses seguidos com o indicador N, o cliente é registrado como um Churn, confirmando a sua deserção.

- Aplicar diferentes algoritmos estatísticos (Random Forest e XGBoost);

Ambos algoritmos foram aplicados com sucesso.

- Apurar qual modelo estatístico possui melhor índice de acerto (utilizando Métricas de Matriz de Confusão) para o dataset;

Os resultados foram acima do esperado e ambos tiveram boa e semelhante performance, mas com algumas leves variações. Ambos tiveram altos índices em todas as métricas, o que vai em linha com serem modelos preditivos de uma natureza similar, mas também fazendo com que a escolha por um modelo em particular tenha que levar em consideração aspectos mais específicos da própria área de atuação.

- Justificar a escolha por um modelo analítico específico para aplicação;

Ambos algoritmos preditivos tiveram um desempenho excelente, atingindo diversas métricas acima de 90% de acerto. Ambos, também, apresentaram como menor indicador a Sensibilidade, tendo um índice de acerto de aproximadamente 0,82, ainda considerado elevado. Tendo em vista que ambos tiveram um desempenho extremamente semelhante, ambos poderiam ser indicados para o uso, sem muita distinção de resultado. Contudo, como já mencionado anteriormente, o ponto principal para a análise de ambos algoritmos foi justamente o ponto de risco mais crítico, que é o erro tipo 2, o Falso Negativo. Esses são os clientes que efetivamente desertaram a empresa, mas não foram corretamente avaliados pelo algoritmo.

Figura 23: Indicadores de Resultados dos Modelos

| Predição do Modelo | XGBoost | Random Forest |
|---|---|---|
| Acurácia | 92% | 93% |
| Precisão | 93% | 93% |
| Sensibilidade | 83% | 82% |
| Especificidade | 95% | 97% |
| F1-Score | 85% | 87% |

Fonte: Elaborado pelo Autor

Este erro implica em perda de receitas financeiras devido ao abandono dos clientes que não receberam a avaliação correta do modelo preditivo. Sendo assim, o modelo que mostrou menor indício de *overfitting* e melhor Sensibilidade foi o XGBoost, este o escolhido como o vencedor e o indicado como auxiliar no planejamento da ação de retenção de clientes da Empresa Adquirente.

Em caso de uma aplicação real do algoritmo, seria altamente recomendado o uso de uma matriz de custos para se tomar a decisão de nível de aplicação de uma ação de prevenção de churn. Por um lado, há o risco de se comprometer com altos valores monetários gastos em clientes que não estariam de fato abandonando a empresa, o que seria evidenciado pela Precisão do modelo. Uma ação de retenção mal planejada poderia acarretar em altos custos e perda de margens financeiras com clientes que não estariam em situação de abandono. O risco oposto seria a confirmação da deserção, a perda de clientes pelo seu valor completo, e não apenas um pedaço de sua margem de contribuição. A avaliação da taxa de churn, juntamente com os resultados da avaliação dos modelos preditivos, podem guiar a decisão da administração para diferentes lados, não há apenas uma decisão correta aqui. Como sempre, cabe ao Administrador a avaliação da situação com todas as ferramentas à sua disposição, a análise preditiva é apenas mais uma delas.

## 6.1 Considerações finais

A escolha do tema, que não está nas áreas clássicas da Administração, mas que no entendimento do autor deveria estar cada vez mais sendo inserida tanto no currículo do curso como no cotidiano de cada Administrador, foi feita como um desafio pessoal. Apesar de já trabalhar com Data Analytics há quatro anos, inclusive auxiliando a criar a área dentro da organização, Machine Learning ainda é algo deveras arcano e fora do cotidiano do trabalho, que tem maior relação com automação de processos para auxílio das atividades exercidas pela empresa. No entanto, por acreditar piamente que Big Data e IA serão os grandes drivers de impulsão ao processo decisório e, portanto, de necessitarem fazer parte do domínio das ferramentas do Administrador, a escolha do tema foi um desafio pessoal que abracei a duras penas, mas que trouxe grande crescimento profissional e acadêmico. Assim como Cálculo, uma vez passado o choque inicial de uma fronteira nova de estatística e matemática, a compreensão da linguagem de programação Python e SQL, assim como os princípios de ML e a aplicação dos algoritmos preditivos, são surpreendentemente acessíveis para àqueles que tem interesse no assunto. Não entrando desmerecendo toda programação por trás da IA e Matemática que está envolvida na criação dos algoritmos, a utilização destas ferramentas pelos analistas que vêm da área de negócios (sem um background das ciências computacionais) está completamente ao alcance e deve ser um grande norteador das organizações, mesmo não as de grande porte.

## 6.2 Limitações da Pesquisa

A maior limitação, por motivos óbvios, foi a necessidade de gerar uma base de dados fictícia (mas dentro de limites coerentes com o setor de Meios de Pagamentos) para utilizar os modelos preditivos. O autor teve um receio inicial que o modelo não seria aderente aos dados reais, com as distribuições variáveis "contaminando"o que seriam os dados verídicos dos limites estipulados, mas este foi um caso que não se confirmou.

Um grande problema é a impossibilidade de atualizar o modelo para dados de novos clientes que ingressaram após 202x, a base foi imaginada olhando para um período fechado de um ano, a inserção de mais clientes seria possível, mas os dados dos clientes atuais ficariam congelados. Contudo, o verdadeiro teste para os modelos seria pegar os indicadores de churn e acompanhar os clientes para avaliar se vão efetivamente confirmar o abandono ou a permanência na base em uma janela de tempo próxima.

## 6.3 Sugestões para Pesquisas Futuras

O maior problema encontrado nos modelos foi o *overfitting*, que causou um peso maior para os Verdadeiros Negativos e, assim, trouxe o resultado completo do modelo para níveis mais elevados que a Sensibilidade apontou. Como a Sensibilidade é, no entendimento do autor, o fator de maior risco para toda análise, por medir a quantidade que o algoritmo conseguiu prever de maneira correta o Churn S, portanto sendo um erro equivalente a perda de um cliente, uma administração de Gestão de Churn demandaria uma melhora nestes índices. Tais resultados podem ser alcançados por meio de reavaliação das variáveis utilizadas, assim como recalibragem dos hiperparâmetros de pesquisa. Para a avaliação dos parâmetros, recomenda-se o uso do Akaike Information Criteria, que avaliaria se os modelos estão com excesso de parâmetros.

Uma outra possibilidade é testar mais modelos preditivos para comparar a aderência dos modelos atuais, como Regressão Logística ou Redes Bayesianas. Incluir uma variável temporal, para tentar identificar não somente se o cliente irá confirmar o churn ou não, mas uma janela de tempo para o fato se concluir. Evoluir a análise para modelos mais complexos, como Redes Neurais e Deep Learning, também é uma avenida interessante, que demanda maior estudo e poder computacional.

# 7 BIBLIOGRAFIA


BREIMAN, L. **Random Forests**. Machine Learning, 45(1), 5-32, 2001.

BUREZ, J.; POEL, D.. **Separating financial from commercial customer churn: a modeling step towards resolving the conflict between the sales and credit department.** Expert Systems with Applications, v. 35, n. 1-2, p. 497-514, 2008.

**CHATGPT'S ODDS OF GETTING CODE QUESTIONS CORRECT ARE WORSE THAN A COIN FLIP**. The Register, 2023. Disponível em: <https://www.theregister.com/2023/0 Acesso em: 12 de agosto de 2023.

**CHATGPT SETS RECORD FOR FASTEST-GROWING USER BASE - ANALYST NOTE**. Reuters, 2023. Disponível em: <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>. Acesso em: 12 de agosto de 2023.

CHEN, H.; CHIANG, R. H. **Business Intelligence and Analytics: From Big Data to Big Impact.** MIS Quarterly, 36(4), 1165-1188, 2012.

CHEN, H.; GUESTRIN, C.**XGBoost: A Scalable Tree Boosting System.** Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. Disponível em: <https://doi.org/10.1145/2939672.2939785>. Acesso em: 07 de julho de 2022.

COUSSEMENT, K; VAN DEN POEL, D. **Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers.** Expert Systems with Applications, Volume 36, Issue 3, Part 2, Pages 6127-6134, 2009.

**DEMONSTRAÇÕES FINANCEIRAS 4T2022.** Cielo, 2023. Disponível em: <https://ri.cielo.com.br/informacoes-financeiras/central-de-resultados/>. Acesso em: 25 de março de 2023.

FAULKNER, P.; MORGAN, G. **Small businesses and power in retail negotiations.** Journal of Small Business and Enterprise Development, 23(2), 424-444, 2016.

FRAZETTO, D. et all. **Prescriptive analytics: a survey of emerging trends and technologies**. The VLDB Journal, 28(4), 575–595, 2019.

**GUERRA DAS 'MAQUININHAS' DE CARTÃO CONTINUA ACIRRADA EM 2019.** Valor Econômico, 2019. Disponível em: <https://valor.globo.com/financas/noticia/2019/01/06/g das-maquininhas-de-cartao-continua-acirrada-em-2019.ghtml>. Acesso em: 25 de março de 2023.

HADDEN, J. et all. **Computer assisted customer churn management: State-of-the-art and future trends.** Computers & Operations Research. 34. 2902-2917, 2007.

HADDEN, J. et all. **Churn prediction: Does technology matter?.** International Journal of Industrial and Manufacturing Engineering, 2(4), 524-536, 2008.

HOFFMAN, T. **Unsupervised Learning by Probabilistic Latent Semantic Analysis** Machine Learning, 42, 177–196, 2001.

**INTRODUCTION TO BOOSTED TREES**. dmlc XGBoost, 2023. Disponível em: <https://xgboost.readthedocs.io/en/latest/tutorials/model.html>. Acesso em: 12 de agosto de 2023.

KEAVENEY, S. **Customer switching behavior in service industries: An exploratory study.** Journal of Marketing, v. 59, n. 2, p. 71-82, 1995.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling.** Springer, 2013.

LANEY, D. **3D Data Management: Controlling Data Volume, Velocity, and Variety.** Gartner Research, 2001

LAUDON, K. C.; LAUDON, J. P. **Management information systems: Managing the digital firm.** Pearson, 2016.

LEJEUNE, M. **Measuring the impact of data mining on churn management.** Internet Research, v. 11, n. 5, p. 375 – 387, 2011.

LU, L.C. et all. **Consumer Attitudes toward Blogger's Sponsored Recommendations and Purchase Intention: The Effect of Sponsorship Type, Product Type, and Brand Awareness.** Computers in Human Behavior, 34, 258-266, 2014.

MALHOTRA, D.; BAZERMAN, M. H. **When More Can Be Less: The Impact of Base Value Neglect on the Asymmetry of Value Gains and Losses**. Organizational Behavior and Human Decision Processes, 105(2), 97-105, 2008.

MANYKIKA et all. **Big Data: The Next Frontier for Innovation, Competition, and Productivity**. McKinsey Global Institute, 2011.

MCAFEE, A.; BRYNJOLFSSON, E. **Big data: The management revolution.** Harvard Business Review, 90(10), 60-68, 2012.

NESLIN, S. et all. **Challenges and Opportunities in Multichannel Customer Management.** Journal of Service Research - J SERV RES. 9. 95-112, 2006.

NIELSEN, D. **Tree Boosting With XGBoost: Why Does XGBoost Win "Every"Machine Learning Competition?.** Norwegian University of Science and Technology, 2016.

OLIVEIRA, D.P.R. **Sistemas de Informações Gerenciais: estratégicas, táticas, operacionais**. Editora Atlas S.A. São Paulo, 2012.

PROVOST, F.; FAWCETT, T. **Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking.** O'Reilly Media, 2013.

QUINTANS-JÚNIOR et all, 2023. **ChatGPT: the new panacea of the academic world**. Revista da Sociedade Brasileira de Medicina Tropical. Disponível em: <https://www.scielo.br/j/rsbmt/a/ZmBhrHSXWwYc6nK8VJbh8pm/?format=pdf&lang=en>. Acesso em: 12 de agosto de 2023.

**RANDOM FOREST ALGORITHM**. Java T Point, 2023. Disponível em: <https://www.javatpoint learning-random-forest-algorithm>. Acesso em: 12 de agosto de 2023.

REINARTZ, W.; KUMAR, V. **The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration.** Journal of Marketing, 67(1), 77-99, 2003.

SOKKHEY, P., OKAZAKI, T,. **Hybrid Machine Learning Algorithms for Predicting Academic Performance** International Journal of Advanced Computer Science and Applications, Vol. 11, No. 1, 2020.

SUSNJAK, T. **Beyond Predictive Learning Analytics Modelling and onto Explainable Artificial Intelligence with Prescriptive Analytics and ChatGPT** Int J Artif Intell Educ, 2023. Disponível em: <https://doi.org/10.1007/s40593-023-00336-3>. Acesso em: 03 de setembro de 2023.

TESSAROLO, P. R. e MAGALHÃES, W. B. **A ERA DO BIG DATA NO CONTEÚDO DIGITAL: OS DADOS ESTRUTURADOS E NÃO ESTRUTURADOS.** Universidade Paranaense – Unipar, Paranavaí/PR, 2015.

TSAI, C.; LU, Y. **Customer churn prediction by hybrid neural networks.** Expert Systems with Applications, v. 36, n. 10, p. 12547-12553, 2009.

WANG, G. et all. **Big data analytics in logistics and supply chain management: Certain investigations for research and applications.** International Journal of Production Economics, Volume 176, Pages 98-110, 2016.

WIRTH, R.; HIPP, J. **CRISP-DM: Towards a standard process model for data mining.** Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (Vol. 1, pp. 29-39), 2000.

WU, X. et all. **Data Mining with Big Data.** IEEE Transactions on Knowledge and Data Engineering, 26, 97-107, 2014.

VERBEKE, W. et al. **Building comprehensible customer churn prediction models with advanced rule induction techniques.** Expert Systems with Applications, v. 38, n. 3, p. 2354-2364, 2011

ZIKOPOULOS, P.; EATON, C. **Understanding Big Data: Analytics for Enter-**

**prise Class Hadoop and Streaming Data.** McGraw-Hill Osborne Media, 2011.

ZIKOPOULOS, P. et al. **Understanding Big Data – Analytics for Enterprise Class Hadoop and Streaming Data**. McGraw-Hill, New York, 2012.

# 8 APÊNDICE

Figura 24: Resumo dos Campos do Dataset

| NOME_COLUNA | BASE ORIGEM | TIPO | SUBTIPO | FORMATO | DESCRIÇÃO |
|---|---|---|---|---|---|
| COD_CLIENTE | Cadastral | Qualitativa | Categórica | Texto curto | Código identificador do cliente (CPF ou CNPJ, mascarado) |
| ID | Cadastral | Qualitativa | Categórica | Texto curto | Chave identificadora do cliente (interna) |
| NOME_CIDADE | Cadastral | Qualitativa | Categórica | Texto curto | Cidade sede/moradia do cliente |
| ESTADO | Cadastral | Qualitativa | Categórica | Texto curto | Estado sede/moradia do cliente |
| IND_PF | Cadastral | Qualitativa | Categórica | Texto curto | Indicador de Pessoa Fiscal (PF/PJ) |
| DATA_CADASTRO | Cadastral | Quantitativa | Contínua | Data | Data de credenciamento como cliente à empresa Adquirente |
| NOME_CANAL | Cadastral | Qualitativa | Categórica | Texto curto | Canal de aquisição de cliente |
| IND_OFERTA | Cadastral | Qualitativa | Categórica | Texto curto | Indicador que cliente aderiu à uma oferta comercial |
| DATA_OFERTA | Cadastral | Quantitativa | Contínua | Data | Data em que cliente aderiu à oferta (NULL se fora de oferta) |
| NOME_PORTE | Cadastral | Qualitativa | Categórica | Texto curto | Porte de cliente em relação à Receita Federal |
| NOME_SEGMENTO | Cadastral | Qualitativa | Categórica | Texto curto | Segmento de mercado de atuação do cliente |
| MESES_BASE | Cadastral | Quantitativa | Discreta | Integer | Quantidade de meses que cliente está credenciado à empresa |
| MESES_OFERTA | Cadastral | Quantitativa | Discreta | Integer | Quantidade de meses que cliente aderiu à oferta comercial |
| TEF | Cadastral | Quantitativa | Contínua | Integer | Quantidade de terminais tipo TEF do cliente |
| POS | Cadastral | Quantitativa | Contínua | Integer | Quantidade de terminais tipo TEF do cliente |
| MOBILE | Cadastral | Quantitativa | Contínua | Integer | Quantidade de terminais tipo TEF do cliente |
| PREDOMINANTE | Transacional | Qualitativa | Categórica | Texto curto | Tipo predominante de terminal pelo volume transacionado (TPV) |
| ATIVO | Cadastral | Qualitativa | Categórica | Texto curto | Indicador se cliente ainda possui terminais ativos |
| ULT_TRANSAÇÃO | Transacional | Qualitativa | Categórica | Integer | Quantidade de dias da última transação considerando a data final do dataset (31/12/202x) |
| MES_ULT_TR | Transacional | Qualitativa | Categórica | Data | Mês da última transação considerando a data final do dataset (31/12/202x) |
| ATIVO_JAN | Transacional | Qualitativa | Categórica | Texto curto | Indicador que cliente teve transações no mês em questão |
| ATIVO_FEV | Transacional | Qualitativa | Categórica | Texto curto | Indicador que cliente teve transações no mês em questão |
| ATIVO_MAR | Transacional | Qualitativa | Categórica | Texto curto | Indicador que cliente teve transações no mês em questão |
| ATIVO_ABR | Transacional | Qualitativa | Categórica | Texto curto | Indicador que cliente teve transações no mês em questão |
| ATIVO_MAI | Transacional | Qualitativa | Categórica | Texto curto | Indicador que cliente teve transações no mês em questão |
| ATIVO_JUN | Transacional | Qualitativa | Categórica | Texto curto | Indicador que cliente teve transações no mês em questão |
| ATIVO_JUL | Transacional | Qualitativa | Categórica | Texto curto | Indicador que cliente teve transações no mês em questão |
| ATIVO_AGO | Transacional | Qualitativa | Categórica | Texto curto | Indicador que cliente teve transações no mês em questão |
| ATIVO_SET | Transacional | Qualitativa | Categórica | Texto curto | Indicador que cliente teve transações no mês em questão |
| ATIVO_OUT | Transacional | Qualitativa | Categórica | Texto curto | Indicador que cliente teve transações no mês em questão |
| ATIVO_NOV | Transacional | Qualitativa | Categórica | Texto curto | Indicador que cliente teve transações no mês em questão |
| ATIVO_DEZ | Transacional | Qualitativa | Categórica | Texto curto | Indicador que cliente teve transações no mês em questão |

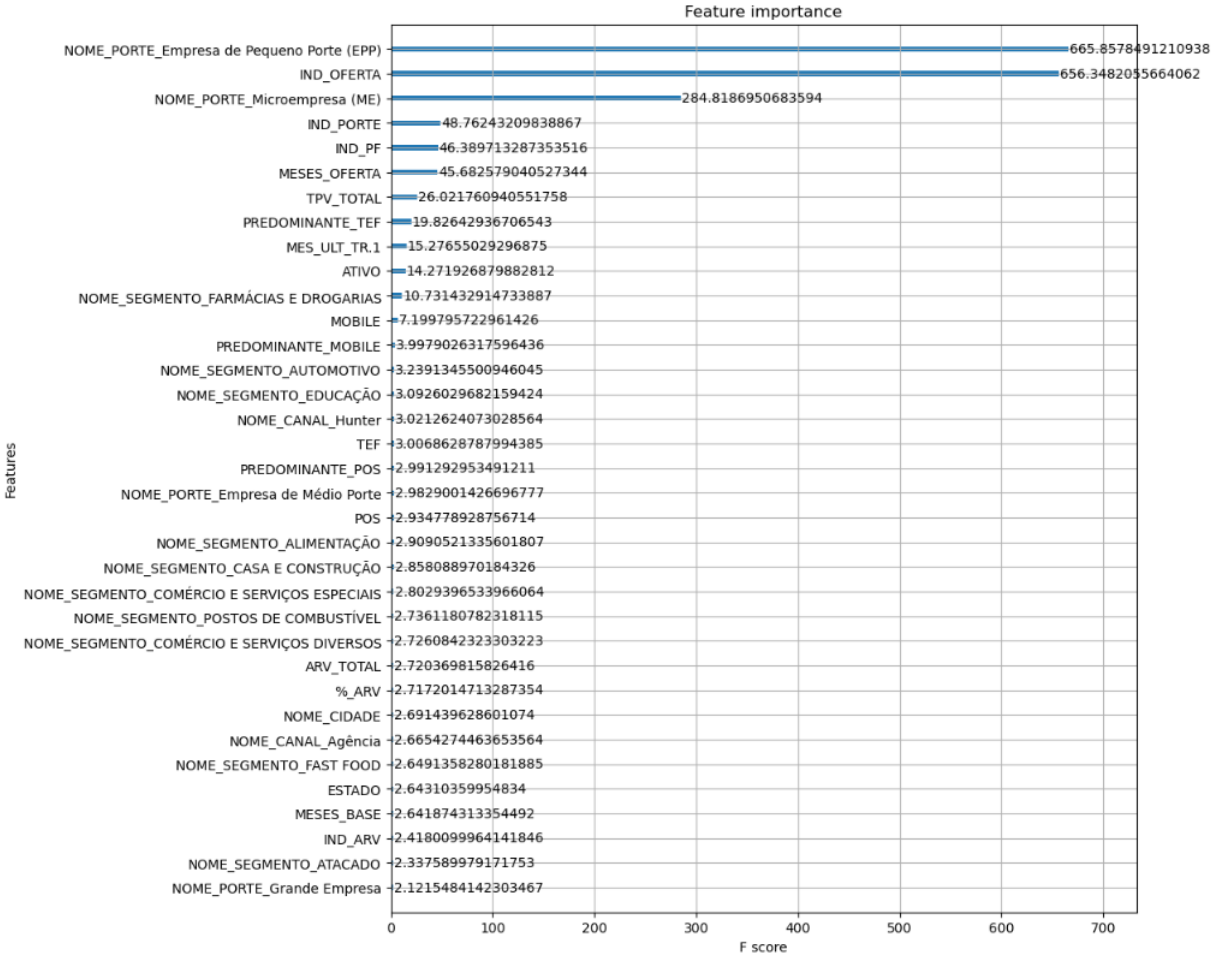
Fonte: Elaborado pelo Autor

Figura 25: Resumo dos Campos do Dataset - Continuação

| | | | | | |
|---|---|---|---|---|---|
| CHURN | Transacional | Qualitativa | Categórica | Texto curto | Indicador que cliente foi apurado como churn |
| TPV_JAN | Transacional | Quantitativa | Contínua | Decimal | Valor de volume transacionado (TPV) no mês em questão |
| TPV_FEV | Transacional | Quantitativa | Contínua | Decimal | Valor de volume transacionado (TPV) no mês em questão |
| TPV_MAR | Transacional | Quantitativa | Contínua | Decimal | Valor de volume transacionado (TPV) no mês em questão |
| TPV_ABR | Transacional | Quantitativa | Contínua | Decimal | Valor de volume transacionado (TPV) no mês em questão |
| TPV_MAI | Transacional | Quantitativa | Contínua | Decimal | Valor de volume transacionado (TPV) no mês em questão |
| TPV_JUN | Transacional | Quantitativa | Contínua | Decimal | Valor de volume transacionado (TPV) no mês em questão |
| TPV_JUL | Transacional | Quantitativa | Contínua | Decimal | Valor de volume transacionado (TPV) no mês em questão |
| TPV_AGO | Transacional | Quantitativa | Contínua | Decimal | Valor de volume transacionado (TPV) no mês em questão |
| TPV_SET | Transacional | Quantitativa | Contínua | Decimal | Valor de volume transacionado (TPV) no mês em questão |
| TPV_OUT | Transacional | Quantitativa | Contínua | Decimal | Valor de volume transacionado (TPV) no mês em questão |
| TPV_NOV | Transacional | Quantitativa | Contínua | Decimal | Valor de volume transacionado (TPV) no mês em questão |
| TPV_DEZ | Transacional | Quantitativa | Contínua | Decimal | Valor de volume transacionado (TPV) no mês em questão |
| TPV_TOTAL | Transacional | Quantitativa | Contínua | Decimal | Volume total de TPV do ano apurado |
| MESES_TPV | Transacional | Quantitativa | Discreta | Integer | Quantidade de meses que cliente teve volume transacionado |
| TPV_MEDIO | Transacional | Quantitativa | Contínua | Decimal | Valor médio mensal de TPV (considerando apenas meses com TPV > 0) |
| IND_ARV | Financeira | Qualitativa | Categórica | Texto curto | Indicador que cliente contratou operação de crédito de Antecipação de Recebíveis de Vendas |
| %_ARV | Financeira | Quantitativa | Contínua | Decimal | Percentual do valor passível de antecipar (operações de crédito) que cliente contratou |
| ARV_JAN | Financeira | Quantitativa | Contínua | Decimal | Valor antecipado no mês em questão |
| ARV_FEV | Financeira | Quantitativa | Contínua | Decimal | Valor antecipado no mês em questão |
| ARV_MAR | Financeira | Quantitativa | Contínua | Decimal | Valor antecipado no mês em questão |
| ARV_ABR | Financeira | Quantitativa | Contínua | Decimal | Valor antecipado no mês em questão |
| ARV_MAI | Financeira | Quantitativa | Contínua | Decimal | Valor antecipado no mês em questão |
| ARV_JUN | Financeira | Quantitativa | Contínua | Decimal | Valor antecipado no mês em questão |
| ARV_JUL | Financeira | Quantitativa | Contínua | Decimal | Valor antecipado no mês em questão |
| ARV_AGO | Financeira | Quantitativa | Contínua | Decimal | Valor antecipado no mês em questão |
| ARV_SET | Financeira | Quantitativa | Contínua | Decimal | Valor antecipado no mês em questão |
| ARV_OUT | Financeira | Quantitativa | Contínua | Decimal | Valor antecipado no mês em questão |
| ARV_NOV | Financeira | Quantitativa | Contínua | Decimal | Valor antecipado no mês em questão |
| ARV_DEZ | Financeira | Quantitativa | Contínua | Decimal | Valor antecipado no mês em questão |
| ARV_TOTAL | Financeira | Quantitativa | Contínua | Decimal | Valor antecipado total do ano apurado |
| MESES_ARV | Financeira | Quantitativa | Discreta | Integer | Quantidade de meses que cliente contratou operações de antecipação |
| ARV_MEDIO | Financeira | Quantitativa | Contínua | Número | Valor médio mensal de ARV (considerando apenas meses com contratação de valor) |

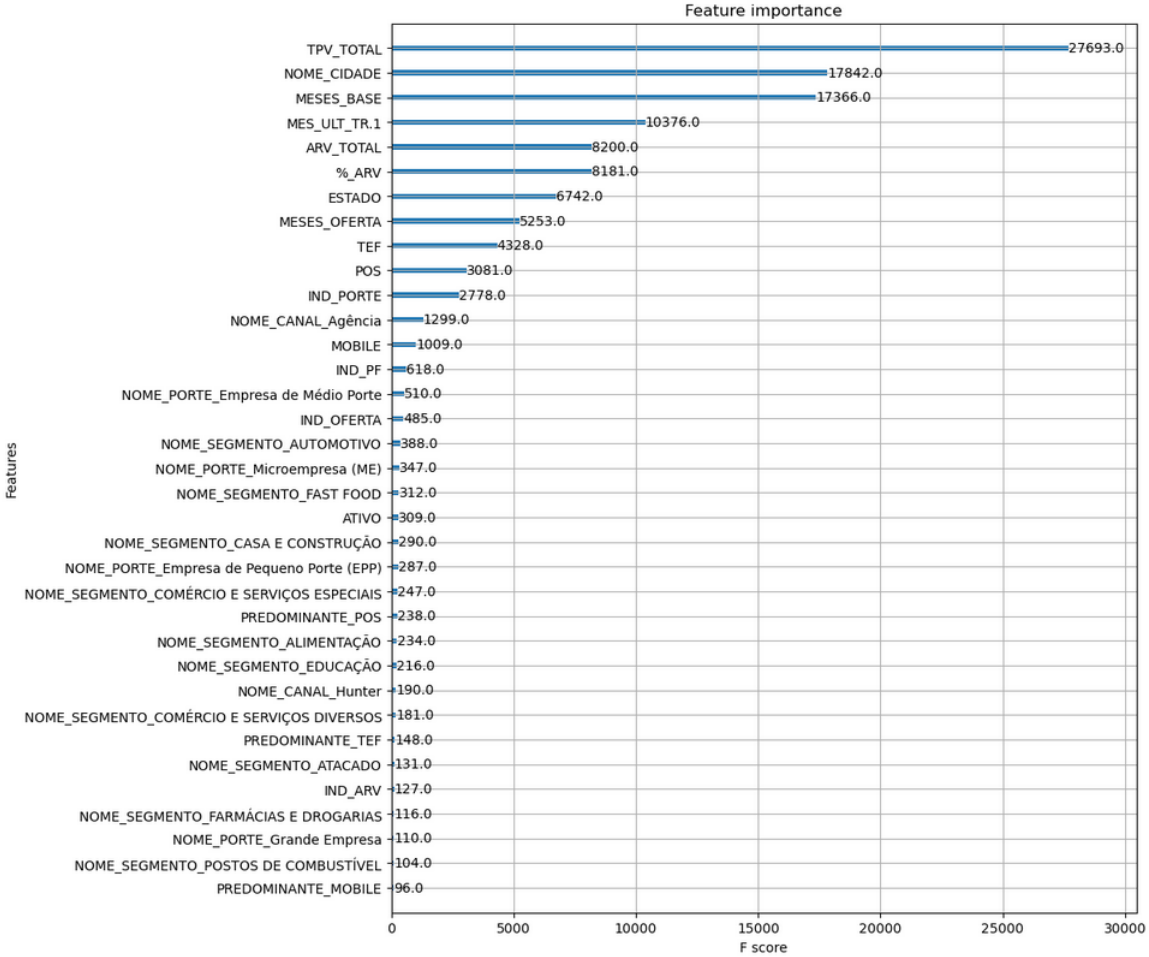
Fonte: Elaborado pelo Autor

Figura 26: XGBoost - Feature Gain




Fonte: Elaborado pelo Autor

Figura 27: XGBoost - Feature Weight




Fonte: Elaborado pelo Autor