UNIVERSIDADE FEDERAL DA PARAÍBA

CENTRO DE CIÊNCIAS HUMANAS, SOCIAIS E AGRÁRIAS

GRADUAÇÃO EM ADMINISTRAÇÃO

RETENÇÃO DE CLIENTES E CIÊNCIA DE DADOS:

Uma Análise de Modelos de Aprendizado de Máquina na Previsão de Churns em uma Associação de Investidores-Anjos da Cidade de São Paulo.

JUAN ALMEIDA FERNANDES

Bananeiras - PB

Outubro de 2024

JUAN ALMEIDA FERNANDES

RETENÇÃO DE CLIENTES E CIÊNCIA DE DADOS:
Uma Análise De Modelos de Aprendizado de Máquina na Previsão de Churns em uma Associação de Investidores-Anjos da Cidade de São Paulo.

Trabalho de conclusão de curso apresentado como parte dos requisitos necessários à obtenção do título de Bacharel em Administração, pelo Centro de Ciências Humanas, Sociais e Agrárias, da Universidade Federal da Paraíba / UFPB.

Docente Orientador: Prof. Dr. Gustavo Correia Xavier.

Bananeiras - PB
Outubro de 2024

Catalogação na publicação
Seção de Catalogação e Classificação

F363r Fernandes, Juan Almeida.
Retenção de Clientes e Ciência de Dados: uma análise de modelos de aprendizado de máquina na previsão de Churns em uma Associação de Investidores-Anjos da cidade de São Paulo. / Juan Almeida Fernandes. - Bananeiras, 2024.
36 f.

Orientação: Gustavo Correia Xavier.
TCC (Graduação) - UFPB/CCHSA.

1. Churn. 2. Ciência de Dados. 3. CRISP-DM. 4. Machine Learning. 5. XGBoost. I. Xavier, Gustavo Correia. II. Título.

UFPB/CCHSA-CHÃ CDU 658 (042)

Elaborado por BRUNA ISABELLE MEDEIROS DE MORAIS - CRB-15/813

Folha de aprovação

Trabalho apresentado à banca examinadora como requisito parcial para a Conclusão de Curso do Bacharelado em Administração.

Aluno: Juan Almeida Fernandes.

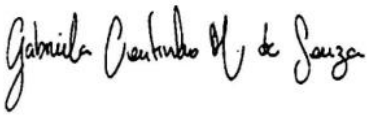
Trabalho: RETENÇÃO DE CLIENTES E CIÊNCIA DE DADOS: Uma Análise de Modelos de Aprendizado de Máquina na Previsão de Churns em uma Associação de Investidores-Anjos da Cidade de São Paulo.

Data de aprovação: 31/10/2024.

Banca examinadora

Prof. Dr. Gustavo Correia Xavier

Orientador(a)



Profª. Ma. Gabriela Coutinho Machado de Souza

Examinadora

“Toda pessoa deveria ser aplaudida de pé pelo menos uma vez na vida, porque todos nós vencemos o mundo.”

August Pullman

AGRADECIMENTOS

Agradeço acima de tudo a minha mãe, minha maior companheira de vida e quem está junto comigo nos momentos bons e ruins. Sei o quanto ela me apoiou para construir esse trabalho e estou hoje firme e forte muito por causa dela. Ela é minha inspiração e espero a orgulhar cada vez mais.

Agradeço a minha colega e amiga de curso, Sayonara, por ter topado ser minha dupla ao longo de todos esses anos. Fizemos trabalhos incríveis, nos divertimos muito, quebramos a cabeça e muito mais durante esses cinco longos anos. Compartilhar esses momentos juntos tornou toda a jornada mais leve e divertida.

Agradeço a todos os meus colegas que fizeram parte do meu círculo social durante essa etapa. Seja assistindo aulas, fazendo monitorias ou participando da Empresa Júnior, foi ótimo ter compartilhado algumas horas dos meus dias com cada um.

Agradeço ao meu orientador, Prof. Dr. Gustavo Correia Xavier, por todo o apoio prestado e por persistir em mim e no tema e aceitar o desafio de finalizar esse meu ciclo na graduação. Cada código criado e palavra dita fizeram melhorar imensamente o meu trabalho e chegar onde estou hoje, sou grato por todos os ensinamentos no decorrer desse 1 ano.

Agradeço também a todo time do GVAngels, por ter acreditado em mim e no meu projeto e permitir com que eu pudesse inovar na associação através deste trabalho.

Por fim, agradeço a todos os professores, alunos e monitores dos cursos e aulas de tecnologia e programação que fiz desde 2021, foi através deles que pude despertar a minha paixão pelo mundo *tech* e é por isso que estou hoje aqui.

RESUMO

O presente estudo tem como objetivo analisar como diferentes modelos de Aprendizado de Máquina contribuem para melhorar a previsão do ato de cancelamento (*churn*) do cliente na Associação de Investidores-Anjos da Fundação Getúlio Vargas (GVAngels). Para tanto, foi adotada uma abordagem quantitativa para investigar o desempenho dos modelos de Aprendizado de Máquina na previsão de *churns* da organização, sendo que o estudo empírico partiu do *framework* CRISP-DM, com o intuito de orientar o desenvolvimento dos processos de Ciência de Dados e a criação de uma inteligência de dados capaz de prever cancelamentos e apoiar positivamente a tomada de decisões estratégicas acerca de *churn* na empresa. Conforme o *framework* CRISP-DM, o cientista de dados passou por todas as etapas de compreensão do negócio, entendimento e preparação dos dados, para então descrever e explicar como diferentes fatores e variáveis podem contribuir com o conjunto de informações a serem incluídos no modelo preditivo. Em seguida, ocorreu a realização da modelagem e avaliação e, por fim, a implantação na rotina organizacional. Dessa forma, seguindo os passos do *framework*, foi inicialmente construída uma estrutura de dados robusta que reúne todas as informações de membros ativos e inativos do grupo, contendo aproximadamente 600 registros únicos de pessoas e mais de 3.000 linhas referentes às atividades de engajamento, que hoje está sendo plenamente aplicada na Gestão de Dados e Relacionamento da empresa. Essa base serviu como escopo à geração dos modelos preditivos de Regressão Logística, LASSO, Random Forest e XGBoost, que, por sua vez, foram comparados estatisticamente quanto ao seu poder preditivo e, graças a uma diferença significativa, foi declarada a vitória da técnica de XGBoost em relação às outras, por apresentar as métricas de *recall* e F1-Score superiores às demais, sendo elas as principais medidas consideradas dentro do problema de *churn*. Portanto, os resultados da pesquisa contribuem para a prática empresarial, ao oferecer uma solução eficaz que venha a minimizar o impacto dos *churns* na Associação de Investidores-Anjo da FGV, bem como à área acadêmica, uma vez que demonstram a aplicabilidade das técnicas de Data Science em um contexto real e complexo, explorando a eficiência desses modelos no *small data*, caso tipicamente encontrado em organizações semelhantes à associação em estudo.

Palavras-chave: Churn; Ciência de Dados; CRISP-DM; Machine Learning; XGBoost.

SUMÁRIO

1. INTRODUÇÃO

Adquirir novos clientes pode ser cinco vezes mais custoso do que reter os que já consomem dos bens e/ou serviços da empresa (Kotler; Keller, 2013). Tendo isso em vista, as grandes organizações têm alterado o seu foco estratégico para atuar em retenção de clientes em vez de mirar na aquisição de novos (Franceshi, 2019), a fim de construir uma vantagem competitiva sustentável à longo prazo (Neslin *et. al*, 2006). Contudo, gerenciar esse processo de cancelamento do consumidor, conhecido como *churn*, é um grande desafio presente para vários gestores em seu dia a dia, dado que a saída de clientes afeta consideravelmente a saúde financeira e a posição de marca da empresa, em razão da alta dinamicidade e competitividade no mercado (Araújo, 2022; Pimentel, 2019; Serpa, 2023). Diante disso, como maneira de contornar essa problemática, é essencial aos empresários que se conheçam os fatores que levam ao cancelamento do cliente, bem como identificar aqueles que tem maior propensão ao cancelamento, podendo, assim, tomar ações preventivas que evitem o ato do *churn*.

Dada a sua relevância prática, a temática da retenção de clientes tem motivado o surgimento de estudos que avaliam o desempenho das novas tecnologias, especificamente modelos preditivos, aplicadas na previsão de *churns* (Alves; Lima; Oliveira, 2022; Eidelwein, 2023; Serpa, 2023; Silveira, 2022). Pois, com o mercado tech em ascensão, surgiu o interesse de investigar como as inovações da área de Inteligência e Ciência de Dados podem contribuir na solução de árduos e constantes problemas na rotina de toda e qualquer empresa. Porém, apesar da crescente literatura na área, a maior parte dos estudos analisam casos de empresas de caráter exclusivamente comercial e de grande porte, carecendo de pesquisas que explorem cenários de organizações menores que, na maior parte das vezes, possuem pequenos volumes de dados (*small data*).

Nesse sentido, o presente trabalho tem como objetivo analisar e avaliar o cancelamento de clientes de uma renomada instituição brasileira de investidores-anjo, como forma de reduzir a saída de membros e, assim, intensificar o seu potencial de mercado (Serpa, 2023). Por se tratar de uma empresa real com uma base de usuários ativos e inativos ainda não estruturada e significativamente menor se comparada a outras pesquisas semelhantes, este estudo adiciona importância à literatura ao aplicar técnicas de análise de dados e modelos de Aprendizado de Máquina (AM), mediante a abordagem CRISP-DM, para compreender como modelos preditivos podem contribuir para o processo decisório na associação de investidores-anjos e, consequentemente, aumentar retenção de clientes associados na empresa. Dessa forma, este trabalho apresenta contribuições tanto de caráter teórico – em razão de incentivar novos estudos na área de dados– quanto prático, ao entregar um produto aplicável para futuras tomadas de decisão acerca do *churn* e à gestão estratégica da organização em estudo de forma tecnológica e inteligente.

Para se alcançar o objetivo do trabalho, é fundamental conhecer primeiro os conteúdos que são ligados ao tema de pesquisa, bem como revisitar a literatura envolvendo *churns* e modelos preditivos. Posteriormente, serão exibidos os métodos utilizados para o projeto, os resultados da aplicação dos modelos preditivos e, por fim, as considerações finais sobre o trabalho e sugestões para estudos futuros.

2. REFERENCIAL TEÓRICO

Primeiramente, é de fundamental importância esclarecer alguns aspectos e temáticas que irão nortear a compreensão acerca do problema de pesquisa como um todo. Dessa forma, nesse capítulo serão apresentados os conceitos a respeito de *churn*, Aprendizado de Máquina (*Machine Learning*) e do funcionamento do processo de análise de dados, bem como serão

revisitados os estudos já existentes sobre as temáticas nos últimos anos, a fim de se obter uma visão ampla de tudo que foi contribuído para a área de tecnologia e à problemática do *churn*.

2.1 Churn

Sendo um dos maiores desafios para lidar no mundo corporativo, o *churn* é marcado como o episódio em que um ou mais clientes deixam de usufruir um determinado bem e/ou serviço de uma firma, impactando diretamente na fonte de receita organizacional (Silveira, 2022). Tal comportamento do consumidor é motivado por diversas razões, sejam elas involuntárias – por decisão da instituição ofertante do produto – ou voluntárias, como o preço em comparação com a concorrência, qualidade do item e/ou serviço promovido, mudanças de residência, questões políticas e entre outras causas que, se não conversadas previamente com a empresa, resultam na conversão de *status* desse cliente de ativo para inativo (Eidelwein, 2023).

Segundo Martins (2021), é natural que ocorram fragmentações naturais no relacionamento comercial que acarretam no cancelamento do serviço por um consumidor. Assim, é normal que as organizações já estejam preparadas para lidar com a mortalidade de clientes ao longo do tempo e criem uma meta estipulada de *churns* ideal para o período, que é acompanhada mensalmente nas reuniões estratégicas e se baseia na relação positiva entre clientes entrantes e inativos da empresa (Alves; Lima; Oliveira, 2022). Tendo isso em vista, são desenvolvidas ações de marketing e pós-vendas voltadas exclusivamente para trabalhar na retenção da base de membros ativos da instituição (Silveira, 2022), de maneira a monitorar os usuários e, caso seja identificado pela equipe ou sinalizado pelo próprio o interesse de romper o relacionamento com a firma, propor o oferecimento de outros serviços adicionais ou submeter uma contraoferta com valor mais baixo, a fim de manter o cliente ativo na organização. Para tanto, a empresa deve ter uma plataforma muito bem construída de Gestão do Relacionamento com o Cliente (*Customer Relationship Management)*, conhecida popularmente como CRM, que irá registrar todas as informações internas e externas referentes aos hábitos de consumo e atividades dos clientes em uma grande base de dados desde o seu primeiro contato com a equipe de vendas, permitindo que a instituição ofertante acompanhe o engajamento desde o início da relação, podendo personalizar toda a experiência e interação com o consumidor e, dessa forma, entender as motivações tanto para ele permanecer usufruindo seus produtos como se afastar (Serpa, 2023; Silveira, 2022).

Todavia, existem casos em que o gestor responsável não consegue identificar previamente o interesse de saída do cliente a tempo de propor uma medida corretiva, afinal – como sustenta Martins (2021) – nem todo *churn* é previsível, especialmente porque, em razões das limitações humanas de raciocínio e tempo, não é possível visualizar todas as perspectivas de maneira simultânea e, por consequência, pode se perder um detalhe que mudaria o rumo da história desse cancelamento. Assim, Burez (2007) orienta que estratégias proativas como o julgamento e o acompanhamento manual da base completa de compradores – ambas tarefas complexas, carregadas de vieses e extensas – devem ser desempenhadas aliadas a todo poder computacional que as tecnologias proporcionam como forma de driblar essas barreiras, dado que, num piscar de olhos, poucas linhas de código seriam capazes de organizar, analisar e entregar hipóteses extremamente úteis para a geração de *insights* que viriam a minimizar o impacto financeiro, gerencial e de marketing dos *churns* no mundo dos negócios.

Diante disso, para que as empresas estejam preparadas para acompanhar essa nova era tecnológica, é fundamental conhecer primeiro alguns termos e temáticas relativas ao vasto universo *tech*, como o estudo do *Data Science*, Inteligência Artificial e *Machine Learning*.

2.2 Ciência de Dados

Dentre as inúmeras áreas de estudo que surgiram com a avanço da tecnologia, a Ciência de Dados – também conhecida como *Data Science* – possui um lugar especial nos setores mais desenvolvidos de TI das grandes empresas, devido a sua grande habilidade de lidar com múltiplos micro e macro problemas gerenciais por meio da ciência (Batista, 2022).

De acordo com Eidelwein (2023), o aumento do interesse e implementação das máquinas nas corporações se deu de forma reativa ao alto crescimento tecnológico após a década de 50 e ganhou força com a famosa era do *Big Data*, em que pessoas e organizações tem sido inundadas com um gigantesco volume de dados gerados em escala global, que perpassam os limites de armazenamento e raciocínio humano (Eidelwein, 2023; Martins, 2021). Com essa conjuntura, surgiu uma grande necessidade para as empresas – sobretudo as que trabalham em ambientes altamente complexos e com serviços virtuais (Wang *et. al*, 2016) – de coletar, tratar e manipular todas essas novas informações ao mesmo tempo para conhecer cada vez mais o seu público-alvo e conquistar fatias de mercado, sendo necessário criar e aplicar constantemente novas ferramentas e tecnologias baseadas em dados, dando vida a área de *Data Science*.

Dessa forma, tendo como foco o entendimento do negócio como um todo, os profissionais de *Data Science* – intitulados como cientistas de dados – buscam gerar hipóteses e realizar inúmeros testes com os dados disponíveis, até comprovar um determinado resultado que venha a impactar diretamente nas decisões tomadas no ambiente corporativo. Assim como o próprio nome alude, eles devem agir como cientistas, aplicando um método científico rigoroso ao longo de toda jornada dos dados, fazendo, conforme Blei e Smyth (2017), forte uso do pensamento estatístico e computacional em suas operações. Para isso, eles geralmente seguem uma rotina pré-definida de atividades cíclicas, que, conforme ilustrado no fluxo abaixo, se iniciam na definição inicial do problema de pesquisa, passando pela geração de hipóteses, coleta de dados e indo até a experimentação e interpretação dos resultados obtidos.

Figura 1: Etapas da ciência de dados

Fonte: elaborado pelo autor, 2024.

Vale mencionar que, caso não se alcance resultados confiáveis e alinhados com o que era esperado, o processo pode ser reiniciado a qualquer momento. Pois, como também acontece na metodologia científica, busca-se sempre o máximo de confiabilidade e qualidade nos resultados e, para isso, o processo pode se repetir quantas vezes forem necessários até chegar à excelência.

2.2.1 Metodologia CRISP-DM

Ademais, como forma de padronizar as necessidades e processos adotados por todo cientista de dados em sua rotina de trabalho, foi construída e oficializada uma outra sequência de atividades, que – embora também se conserve o rigor científico – possui ações mais voltadas para boas práticas de extração, transformação e armazenamento de dados (ETL) em conjunto com variados modelos de inteligência e aprendizado de máquina no desenvolvimento de novos modelos matemáticos-estatísticos.

Conhecido pela sigla CRISP-DM, o Processo Padrão de Indústria Cruzada para Mineração de dados – *Cross-Industry Standard Process for Data Mining* – é uma das metodologias mais usadas em projetos de ciências de dados desde sua criação em 1996. A maior eficácia do *framework* se concentra em suas 6 etapas bem definidas e cíclicas, que permitem com o que o dono do projeto retorne quantas vezes forem necessárias até se alcançar um resultado satisfatório na última etapa, a implantação no contexto de negócios – fases essas que serão apresentadas a partir dos resultados do presente trabalho.

Figura 2: Framework CRISP-DM para Ciência de Dados

Fonte: Araújo, 2022.

Antes de iniciar o debate acerca do funcionamento do método de Aprendizado de Máquina na resolução de problemas, deve-se conhecer primeiro a sua tecnologia-mãe, a Inteligência Artificial (IA).

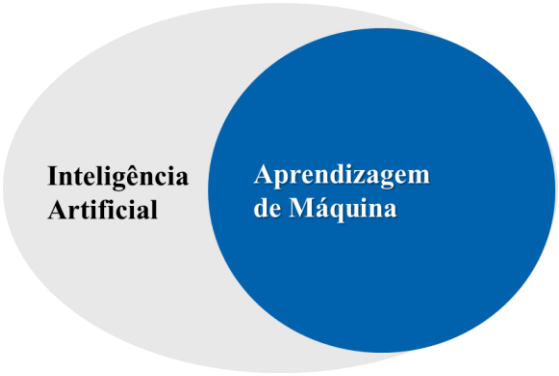
2.3 Inteligência Artificial

Segundo Anna Brown, da Software Analytics & Soluções - SAS (2023), a Inteligência Artificial é considerada uma ciência que – por meio da observação e análise de situações-exemplo – treina máquinas que sejam capazes de pensar e performar em níveis iguais ou superiores aos seus criadores, os humanos. De acordo com Frank Herbert (1965), jornalista e autor da renomada obra "Duna", isso só foi possível quando o indivíduo passou a compreender as próprias barreiras de raciocínio e, para não esquecer o que aprendeu, passou a transferir constantemente os seus conhecimentos e experiências às máquinas, sempre exigindo-as cada vez mais e questionando quais eram os seus verdadeiros limites computacionais (Martins, 2021).

Sendo cunhado em 1950 por John McCarthy, o termo "IA" iniciou como a ciência de fazer máquinas inteligentes (Amaro Junior, 2022) e hoje vem assumindo um significado ainda maior no universo *tech*, compreendendo todo e qualquer sistema computacional que objetive mimetizar uma ou mais habilidades humanas, como o raciocínio lógico. Tudo isso se dá por meio de um método de aprendizado em que a máquina irá primeiramente observar um determinado contexto e, mediante a geração de hipóteses, construir e implementar soluções estatístico-computacionais na situação-problema identificada e, caso não haja uma boa performance, irá reiniciar o ciclo até alcançar um nível ideal de acurácia e precisão no resultado.

Para fins de entendimento, pode-se imaginar um cenário de uma sala de aula, na qual o professor tem o objetivo de ensinar e validar os conhecimentos dos discentes – mediante a aplicação de provas – e caso o aluno não alcance o desempenho suficiente, ele deverá compreender os erros cometidos, estudar os conteúdos novamente com o docente e refazer o exame quantas vezes forem necessárias até conquistar uma nota satisfatória pra avançar na disciplina. Semelhante a essa conjuntura, as máquinas são capazes de aprender constantemente consigo mesmas, assumindo os papéis tanto de mestre quanto de estudante nesse processo próprio da IA intitulado como Aprendizado de Máquina (em inglês, *Machine Learning*).

Figura 3: Diagrama de Venn sobre a relação entre Inteligência Artificial e *Machine Learning*



Fonte: adaptado de Kumar, 2022.

Logo, conforme o diagrama apresentado na figura 3, pode-se afirmar que o Aprendizado de Máquina é caracterizado como um dos inúmeros métodos usados na grande área da Inteligência Artificial, ou seja, nem toda IA é considerada um modelo de *Machine Learning* (Oracle, 2024).

2.4 Aprendizado de Máquina

De forma concisa, um algoritmo de Aprendizado de Máquina é um dos métodos aplicados no processamento de IAs, em que – a partir dos dados fornecidos pelo usuário – o programa de computador executa uma sequência de ações e avalia repetidamente as consequências de cada alternativa gerada, finalizando somente ao obter um resultado que cumpra com os objetivos buscados pelo cientista de dados ao escrever o código (Batista, 2022). Logo, o "*learning*" (em português, aprendizado) entra justamente em todo o processo que se decorre da criação e seleção de futuros possíveis até se chegar a uma conclusão satisfatória, pois, assim como o ser humano, a máquina foi e vem sendo projetada para aprender mediante observação, estudo e experiências de tudo que acontece no ambiente que está inserida para se adaptar ao que esperam dela.

Para entender melhor acerca dessa curva de aprendizagem das IAs, Arthur L. Samuel (1959) – considerado um dos criadores do termo *Machine Learning* – realizou uma pesquisa sobre o desempenho dos computadores em um jogo de damas e concluiu que, em ritmo progressivo, a máquina foi treinando, identificando erros ao ser derrotada e se aprimorando ao longo do tempo, tornando-se quase invencível. Quem comprovou isso foi Robert Nealey, autoproclamado rei da dama na década de 50 que, em 1962, perdeu o seu título de invicto justamente para o computador IBM 7094, sendo um dos marcos mais impactantes na área de Inteligência Artificial (IBM, 2024). Desde então, saindo dos jogos e indo até o universo corporativo, os algoritmos de Aprendizado de Máquina passaram a ser cada vez mais aplicados

em prol dos dados organizacionais internos das empresas, pois, conforme comentado anteriormente, grande parte das instituições tiveram um aumento significativo em volume de informações que não conseguiu acompanhar o ritmo de trabalho humano na era do *Big Data*, sendo de extrema ajuda a implementação de tais tecnologias em conjunto com os colaboradores, a fim de gerenciar todas essas bases de dados e, com isso, criar valor às firmas.

Para isso, tanto as empresas que atuam diretamente no setor *tech* quanto as restantes fazem uso intenso de ferramentas de análise de dados, modelos preditivos, robôs para automação e otimização de tarefas e entre outros meios para lidar com esse fluxo desenfreado de dados e informações de clientes, fornecedores, concorrentes, *churns* e entre outros chegando a cada segundo. Assim, ao implementar máquinas nas operações, o gestor consegue analisar os aspectos estratégicos em cada processo e, com a grande quantidade de dados históricos disponíveis, fazer com que elas trabalhem para identificar os pontos críticos positivos e negativos na situação-problema, como, por exemplo, os consumidores que mais compram, os que tendem a serem inadimplentes, os mais fiéis à marca e, sendo o foco do presente estudo, os com potencial de cancelamento de um bem e/ou serviço (Oracle, 2024).

2.5 Revisão de literatura

Conforme salienta Martins (2021, p. 20) em seu estudo, "ainda que o aprendizado de máquina seja uma excelente ferramenta para conquista de conhecimento, não existe algoritmo que descreva, com qualidade, todas as adversidades". Isso se deve ao fato de que, como existem muitas variáveis envolvidas, diferentes modelos de *Machine Learning* podem atender circunstâncias distintas de problemas. Dessa forma, para entender como alguns autores refletiram sobre o *churn*, foi usado como arcabouço teórico dessa pesquisa um conjunto de artigos acadêmicos que trabalharam variados algoritmos de Aprendizado de Máquina e em contextos distintos de negócios, como mostrado abaixo no quadro 1:

Quadro 1: Relação entre trabalhos publicados e algoritmos de Aprendizado de Máquina utilizados em diferentes contextos de negócios

Algoritmos trabalhados	Contexto de negócio	Métricas de avaliação	Referências bibliográficas	Língua
Regressão Logística, K-Means e KNN	Evasão de clientes em uma Fintech Brasileira	Acurácia	(Alves; Lima; Oliveira, 2022)	Português
Regressão de Cox	Previsão de Churns de Clientes de Seguros de Vida do Banco do Brasil	Estimador de Kaplan-Meier	(Araújo, 2022)	Português
Random Forest	Análise de Churn de clientes em uma operadora de telecomunicações	Acurácia	(Batista, 2022)	Português
Random Forest e XGBoost	Previsão de abandono de clientes em uma empresa facilitadora de pagamentos	Acurácia, Precisão, Recall, Especificidade e F1-Score	(Eidelwein, 2023)	Português
Weighted Random Forest	Evasão de funcionários de uma empresa de Telecomunicações	Área sobre a curva ROC; F1-Score e Recall	(Gao; Wen; Zhang, 2019)	Inglês

CatBoost, SVM, Regressão Logística, Árvores de decisão, Random Forest e XGBoost	Avaliação da produtividade de funcionários	Coeficiente de correlação de Mathews (MCC)	(Jain; Tomar; Jana, 2021)	Inglês
SVM, Árvores de decisão, Random Forest, Naive Bayes, Regressão Logística e KNN	Evasão de médicos com depressão de clínicas e hospitais	Acurácia	(Joseph *et. al*, 2021)	Inglês
Random Forest	Previsão de *churn* de uma instituição varejista	Acurácia, Precisão, Recall e F1-Score	(Martins, 2021)	Português
Similarity Forest, Regressão Logística e Random Forest	Retenção de clientes em uma empresa de telecomuniçãoes	Área sobre a curva ROC e teste de Kruskal-Walls	(Óskarsdóttir *et. al*, 2018)	Inglês
Redes Neurais	Análise de sentimentos e predição de *churn* em CRM	Acurácia e Taxa de Falsos Positivos	(Pimentel, 2019)	Português
Regressão Logística, Árvores de decisão e Redes Neurais	Previsão de risco de inadimplência de alunos do Ensino Superior em uma IES privado do Rio de Janeiro	Acurácia, Precisão, Recall e Especificidade	(Saadia *et. al*, 2022)	Português
Random Forest	Previsão de abandono de clientes em uma instituição financeira alemã	Acurácia, Precisão, Recall, Especificidade e F1-Score	(Serpa, 2022)	Português
Regressão Logística, Árvores de decisão, Random Forest e Redes Neurais	Predição de churns em uma empresa de telecomunicações	Acurácia, Precisão, Recall e F1-Score	(Silveira, 2022)	Português
Random Forest, Gradient Boosting, Redes Neurais Profundas	Avaliação da satisfação do empregado em uma indústria de bens de consumo	Acurácia	(Srivastava; Eachempati, 2021)	Inglês
Redes Neurais Profundas, GCN e LSTM	Evasão de colaboradores em empresas	Erro médio absoluto	(Teng *et. al*, 2021)	Inglês
K-Means, LOF e CBLOF	Churn de clientes de uma instituição bancária	Precisão, Recall e F1-Score	(Ullah *et. al*, 2019)	Inglês
Regressão Logística, Árvores de decisão, Redes Neurais, SVM e Random Forest	Análise de dados pessoais e profissionais de empregados	Área sobre a curva ROC	(Yuan, 2021)	Inglês

Fonte: elaborado pelo autor, 2024.

Diante do quadro, percebe-se que parte dos trabalhos publicados aplicaram diferentes modelos de *Machine Learning* na análise preditiva do *churn*, sendo Regressão Logística e Random Forest os mais utilizados, fato esse que teve influência na escolha de modelos para a presente pesquisa.

Figura 4: Nuvens de palavras com algoritmos de Aprendizado de Máquina mais aplicados.

Fonte: elaborado pelo autor, 2024.

Destaca-se que é de responsabilidade direta do cientista entender bem a problemática do negócio e aplicar o melhor sistema de aprendizagem para o contexto em questão. Em contrapartida, nenhum desses artigos trabalhou sob o cenário de associação de investidores e, ainda mais, com *small data*, sendo esse um estudo pioneiro à área, tendo, assim, grande relevância teórico-prática para o incentivo de mais propostas de pesquisa na esfera *tech*.

3. MÉTODO

Para alcançar o propósito do presente trabalho, foi adotada uma abordagem quantitativa com o objetivo principal de – mediante a aplicação de modelos estatísticos de *Machine Learning* advindos da Ciência de Dados – descrever, explicar e analisar como diferentes fatores e variáveis influenciariam no ato de *churn* do cliente na empresa em estudo, a Associação de Investidores-Anjos da Fundação Getúlio Vargas.

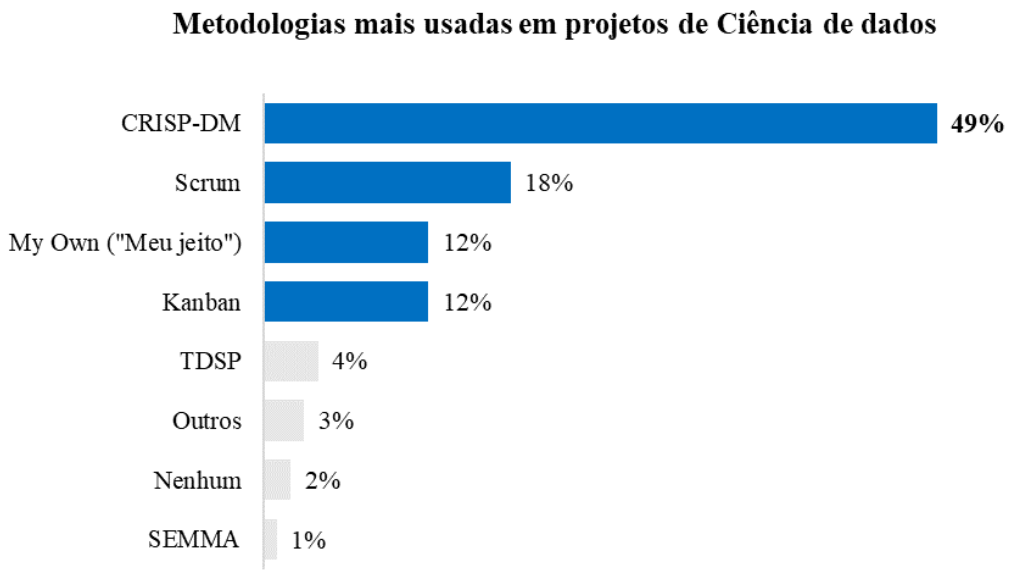
Criada em 2017 por ex-alunos da FGV, a organização tem sede em São Paulo-SP e possui hoje mais de 300 associados ativos na comunidade que, juntos, já movimentaram mais de 50 milhões de reais na economia nacional e internacional, fazendo jus ao título de instituição referência em Investimento-anjo na América Latina (CB Insights, 2022). Para fins de compreensão, vale destacar que o perfil de clientes do grupo é de alta renda, formado principalmente por pessoas que estejam ou já estiveram ocupando um cargo de alta posição em uma empresa (como fundador, sócio, *c-level* ou diretoria) e que estejam interessados em investimento-anjo, fortalecimento de comunidade, *networking* e/ou educação-anjo.

Primeiramente, para fundamentar o estudo, foi realizada uma consulta de caráter exploratório por meio da leitura de artigos acadêmicos, livros, documentos, teses, notícias e sites da internet acerca do uso de ferramentas e modelos de *Machine Learning* em prol da resolução de problemas organizacionais em outros contextos de análise. Assim, foi desenvolvida uma importante base teórica que, a partir da visão de outros autores, norteou a execução das etapas aplicadas da pesquisa, como a coleta de dados necessários para a criação da base de dados-mãe que seria empregada nos modelos de Aprendizado de Máquina.

A partir desse momento se adotou a abordagem aplicada do *framework* CRISP-DM, devido a sua grande popularidade e utilização na construção de projetos de ciência de dados por especialistas e profissionais da área. Tal escolha se deu em concordância com a pesquisa

levantada em 2024 pelo grupo Data Science Process Alliance, que chegou à conclusão de que o *framework* tem cerca de metade de aprovação dos cientistas de dados e profissionais de TI em geral, ganhando inclusive de modelos bem conhecidos do mundo corporativo, como o *Scrum* e o *Kanbam*. Os resultados estão ilustrados na figura 5 abaixo:

Figura 5: Resultados da pesquisa do grupo Data Science Process Alliance



Fonte: adaptado de datascience-pm.com, 2024.

Dessa forma, os resultados do presente trabalho estarão de acordo os processos adotados no *framework* CRISP-DM, sendo apresentado não só como se iniciou o projeto, como também todos os seus avanços até o momento. Para adeptos do modelo, a repetição é regra (Provost; Fawcett, 2016) e fundamental para se alcançar o objetivo detalhado na primeira etapa da metodologia, a compreensão do negócio.

4. RESULTADOS

4.1 Compreensão do negócio

Primeiramente, a pergunta a ser respondida por um cientista de dados ao iniciar um projeto é justamente qual problema vai ser resolvido e sob que contexto (Provost; Fawcett, 2016). Para tanto, o profissional responsável deve conhecer bem não só o motivo de ter sido chamado para cumprir essa tarefa, mas também o cenário envolvido, ou seja, a empresa em questão.

Para esse estudo, será considerada a Associação de Investidores Anjos da Fundação Getúlio Vargas – GVAngels – que tem enfrentado uma dificuldade comum de empresas cujo produto principal é um serviço e que será levado em conta no desenvolvimento do projeto, o cancelamento de assinatura de clientes. Considera-se como *churn* aquele membro que, decorridos 12 meses de associação (tempo de filiação anual), tomou a decisão de descontinuar com a instituição. Vale mencionar que tal cancelamento só pode ser feito em intervalos de 1 ano, pois, como o plano se trata de um valor fixo que é pago logo antes da entrega do serviço, o cliente só pode deixar de consumir no fim do compromisso financeiro acordado entre ele e a empresa.

Além do contexto, entender os objetivos esperados pelos *stakeholders* que estejam direta ou indiretamente ligados à proposta é de suma relevância, para alinhar as expectativas e construir um plano estrutural que esteja de acordo com os critérios de performance, recursos disponíveis e resultados esperados do projeto final (Provost; Fawcett, 2016). Para assimilar esses fatores, o profissional de *Data Science* deve manter uma comunicação efetiva com todos os colaboradores de diferentes áreas da empresa, indo desde o topo estratégico – representado pela figura do gestor – até a base operacional, a fim de evitar o surgimento de ruídos que prejudiquem o esclarecimento acerca da verdadeira dimensão da situação-problema (Provost; Fawcett, 2016).

Nesse sentido, o próximo passo do projeto se deu em conhecer quais informações registradas sobre clientes ativos e inativos a instituição continha, sendo em um primeiro momento mediante conversas rotineiras com a equipe da associação e logo após partindo para a parte analítica, mergulhando, por fim, nos dados do GVAngels.

4.2 Entendimento dos dados

Tendo conhecimento do problema de negócio a ser desvendado, o direcionamento do cientista é extrair dados, o seu principal insumo, para começar a arquitetar a solução (Provost; Fawcett, 2016). A forma como essa matéria-prima é extraída irá depender da maneira que a gestão da informação é promovida na empresa, sendo de modo interno ou terceirizado, centralizado em uma área de Tecnologia ou segmentado por departamentos, de acesso público ou privado para o cientista de dados e entre outras considerações.

No caso em específico desse projeto, ao ser iniciado o trabalho de entendimento dos dados só estava disponibilizada para acesso uma única base intitulada “bot”, que continha informações sociodemográficas e de interesse dos membros do grupo, sendo eles ativos ou cancelamentos. Cada registro era efetuado na etapa de cadastro do associado em uma plataforma externa à empresa e redirecionada via API (*Application Programming Interface*) para uma planilha no Google Sheets de modo automático.

Essas informações cadastradas somavam cerca de 440 registros e estavam divididas em 54 colunas. Para facilitar a identificação, pode-se as classificar conforme as seguintes categorias no quadro 2:

Quadro 2: Categorias de dados presentes na base “bot”

Categorias	Exemplos de informações
Sociodemográficas	Gênero, nascimento, endereço de nascimento, escolaridade, nacionalidade, estado civil, residência, CEP
Pessoais	Nome, Certidão de Pessoa Física – CPF, Registro Geral – RG, e-mail, telefone, LinkedIn, foto
Profissionais	Empresa, setor, cargo atual, posição hierárquica e áreas de expertise
Específicas do grupo	Objetivos esperados na comunidade, interesses de investimento, nível de conhecimento sobre investimento-anjo, quanto espera investir de capital, NPS, Termos de comunicação

Fonte: elaborado pelo autor, 2024.

Contudo, embora apresentasse um número razoável de fatores para o avanço do trabalho, muitos dos dados não faziam sentido para a criação de um modelo preditivo eficaz ou apresentavam um grande número de informações nulas que poderiam prejudicar a performance do algoritmo de Aprendizado de Máquina a ser desenvolvido, sendo essencial os remover à melhoria dos resultados esperados (Goldschmidt; Passos, 2015; Pimentel, 2019; Martins, 2021). Outrossim, com o fim de não ferir os direitos dos usuários assegurados pela Lei Geral de Proteção de Dados – LGPD (Brasil, 2018), quaisquer elementos que permitissem identificar, direta ou indiretamente, um membro que estivesse vivo não deveriam ser constados no modelo final. Dessa forma, cortando as variáveis que entrariam nesses critérios de desqualificação, restavam menos de 40 colunas para o avanço do projeto.

Segundo Provost e Fawcett (2016), o pensador analítico de dados precisa sempre ter o questionamento se os dados disponíveis hoje têm valor suficiente para justificar o investimento ou modelo de máquina a ser desenvolvido. Estando na figura do principal responsável, o cientista de dados deve refletir e entender se a resposta dessa pergunta seria “Sim” ou “Não”, o que foi realizado e concretizado na segunda opção no presente estudo. Tal percepção se deu mediante o conhecimento da jornada do cliente dentro do grupo ao longo do tempo, pois, embora as informações obtidas relativas ao perfil sociodemográfico, profissionais e específicas de interesse detivessem seu peso, o que mais importava era o caminho percorrido pelo membro dentro da associação ou, em outras palavras, o engajamento.

Acerca disso, Provost e Fawcett (2016, p. 12) defendem que:

> Os dados sociodemográficos fornecem uma capacidade substancial para modelar os tipos de consumidores mais propensos a comprar um produto ou outro. No entanto, dados sociodemográficos têm seu limite; depois de certo volume, nenhuma vantagem adicional é conferida. Em contrapartida, dados detalhados sobre transações individuais dos clientes melhoram substancialmente o desempenho.

Logo, tendo maior conhecimento sobre tudo que cada associado realiza enquanto membro ativo, foi possível coletar mais dados que estavam divididos por área na instituição que viriam a acrescentar e melhorar a performance do modelo preditivo a ser criado (Martens; Provost, 2011), sendo elas Comercial, Marketing, Comunidade e *Dealflow*. Cada uma dessas áreas apresenta informações relevantes para a construção da jornada de experiência do membro, que seriam, por sua vez, importantes para compreender a correlação entre eventos que antes não se imaginava terem associação. Dentre as possibilidades de engajamento do membro para com o grupo que poderiam ser consideradas são:

Figura 6: Nuvens de palavras com atividades de engajamento dos membros da Associação

Fonte: elaborado pelo autor, 2024.

Cada uma dessas atividades representa caminhos em que os clientes da empresa podem seguir ao se tornarem membros oficiais, podendo ir por um eixo mais voltado para investimentos (com aportes financeiros, tokenização e liderança de rodadas de negócios de startups), conhecimento (participando de *masterclasses*, oficinas de inovação para alunos da Fundação Getúlio Vargas e outras instituições parceiras, mentoria de startups e virar aluno do curso próprio da empresa sobre Investimento-anjo) ou se direcionando mais para a geração de networking, ao criar conexões de valor com pessoas de mesmo interesse profissional.

Assim, foi iniciado um processo contínuo e iterativo que envolvia conversar com os líderes da área, compreender como cada um registrava as informações relevantes para o trabalho, entender se esses dados poderiam ser centralizados em uma única base e, em caso de dúvidas, marcar novamente um alinhamento com o departamento para esclarecimento e continuidade do projeto. Essa fase representou uma das principais adversidades para o desenvolvimento do trabalho como um todo, devido aos custos variáveis dos dados (Provost; Fawcett, 2016), pois, uma vez que eles estavam espalhados pelas áreas da empresa, cada líder registrava suas informações conforme fosse melhor para a própria rotina, gerando planilhas desagregadas e não padronizadas, sendo um verdadeiro desafio coletar esses dados e os normalizar para aplicação completa no projeto. Dessa forma, foi necessário encontrar a melhor forma de limpar e combinar os registros dos membros, para garantir que existisse não só um registro único por cliente (Provost; Fawcett, 2016), mas também uma base que considerasse as interações entre as áreas como um todo.

Seguindo esse caminho, foi constituída uma estrutura robusta de dados composta por mais de 600 registros de clientes ativos e inativos da instituição entre 2017 e 2024, mediante a integração de nove diferentes planilhas de Excel e CSV que estavam espalhadas pelas áreas, processos e plataformas da organização.

Intitulada pelo cientista de dados responsável como “GV Aplanilha”, a nova base gerada reúne 5 planilhas que refletem e agregam as principais necessidades de cada área em um único arquivo em Excel, sendo elas:

Quadro 3: Organização das planilhas na GV Aplanilha

Planilhas	Quantidade de Registros	Descrição
Planilha-mestre	606	Contém todas as informações relativas a pagamentos, como dados para contato, data de renovação, último valor a ser pago e entre outros
Atividades	3.628	Desenvolvida a partir do que foi informado por cada uma das áreas, registrando em um só lugar todas as participações em eventos, investimentos e indicações de membros ativos e inativos do grupo
Bot	441	Reúne as informações preenchidas pelo recém-associado na etapa de cadastro
Datas	104	Lista os eventos e suas respectivas datas para conexão com a planilha “Atividades”
Câmbio	72	Registra os investimentos realizados e, em caso internacional, também realiza o câmbio da moeda para o padrão brasileiro (R$)

Fonte: elaborado pelo autor, 2024.

Dentre as planilhas compiladas, destaca-se a “Planilha-Mestre”, por recuperar mais de 200 informações perdidas sobre ex-clientes da associação, em razão de não haver uma arquitetura de dados apropriada até o presente projeto, e a “Atividades”, com 3.628 registros de participações e engajamento dos membros que foram reunidas em um só lugar, permitindo ter uma visão completa da jornada da experiência do cliente.

Após a fase de coleta de dados, que teve duração aproximada de sete meses em decorrência das inúmeras vezes que o processo teve que ter sido reiniciado para melhorar a qualidade dos dados e refletir com maior exatidão a necessidade real de negócios (Provost; Fawcett, 2016), a próxima tarefa a ser realizada é a Análise Exploratória dos Dados, conhecida popularmente na área de TI como EDA.

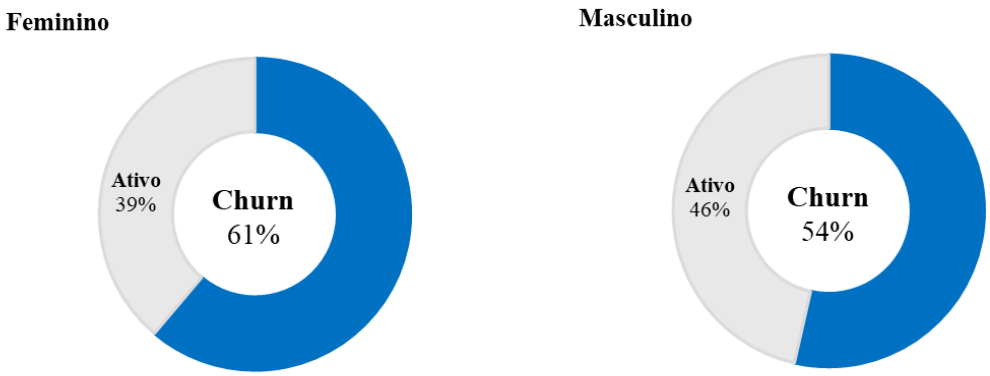
4.2.1 Análise Exploratória dos Dados (EDA)

Como afirma Eidelwein (2023), a EDA é um dos passos mais importantes e que não pode ser pulado na construção de qualquer modelo de Aprendizado de Máquina, pois é por meio dessa atividade investigativa que o cientista de dados consegue analisar as informações disponíveis e verificar se elas já conseguem por si só descrever o fenômeno estudado, no caso, o problema de negócios. Assim, é possível sinalizar os pontos fortes e limitações da base de dados, permitindo com o que dono do projeto atue de maneira proativa e questione se é necessário ou não mudar as direções de resposta e esforços para o melhor encaminhamento da solução (Provost; Fawcett, 2016).

Além disso, é mediante a Análise Exploratória que o profissional responsável compreende com maior profundidade os dados coletados e descobre o que pode ser digno de nota (Knaflic, 2019), ou seja, qual informação mais se destaca na base de dados e que seria, por sua vez, mais interessante para transmitir ao modelo preditivo a ser criado. Tal investigação pode ser iniciada através do estudo de variáveis mais gerais a todas as bases de dados que envolvem pessoas, como a informação de gênero, localização e idade, e entender como elas se relacionam com o fenômeno estudado.

No caso do projeto em execução, pode-se exemplificar com a observação da relação do gênero com a chance de cancelamento do cliente conforme mostrado na figura 7, em que o cancelamento é demonstrado quando a variável “y” é 1 (cor azul). Mediante a análise gráfica, foi possível perceber que dentre as mulheres que estão ou estiveram no grupo, 52 das 85 totais (61,18%) já saíram da associação. Porém, quando considerada a base inteira, essas mesmas 52 empresárias representam somente 15,7% dos *churns*, ou seja, devido a sua minoria no grupo, o número impacta mais quando calculado dentro do público em específico, no caso, o feminino.

Figura 7: Distribuição de *churns* por gênero no GVAngels

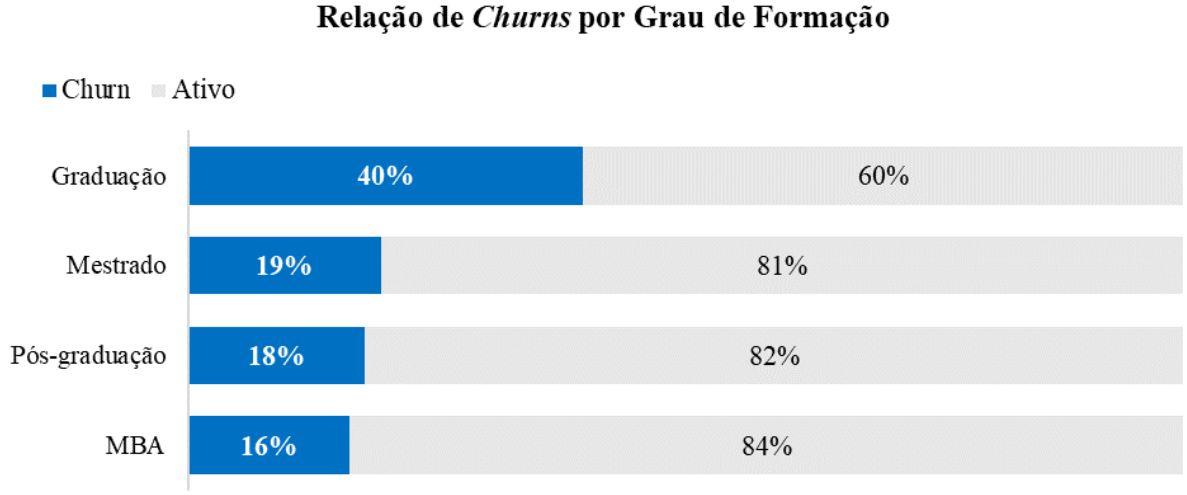


Fonte: elaborado pelo autor, 2024.

Em um primeiro momento essa análise pode ser considerada desnecessária, mas tudo dependerá do que se espera alcançar no problema de negócios. Por exemplo, caso a empresa em questão esteja com uma preocupação em cima do engajamento do público feminino, esses gráficos poderão nortear melhor as conversas sobre o tema e fazer com que as tomadas de decisão sejam orientadas por dados e não por simples achismos (Provost; Fawcett, 2016).

Ademais, a EDA permite com que o cientista de dados possa trabalhar com as informações disponíveis e imaginar ações potenciais a serem tomadas no futuro, como compactar ou somar variáveis à base. Para ilustrar essa situação, segue a figura 8 que representa a distribuição de *churns* por grau de formação no grupo.

Figura 8: Distribuição de *churns* por grau de formação no GVAngels



Fonte: elaborado pelo autor, 2024.

Durante o processo investigativo da escolaridade, observou-se que qualquer grau de formação acima da graduação agia positivamente na minimização do cancelamento de assinaturas e, com base nisso, foi possível criar na fase seguinte um código que identifique todo caso de cliente com formação nesses quesitos e marque no modelo final, sendo mais prático e chegando a mesma resposta em menor tempo. Para tanto, o cientista de dados deve possuir ciência e tecnologias substanciais para aplicar devidamente a essa base, o que já ultrapassa a fase de entendimento de dados, chegando, assim, a preparação dos dados para o modelo final.

4.3 Preparação dos dados

Relembrando, a área de Data Science é composta por "princípios fundamentais que norteiam a extração de conhecimento a partir de dados" (Provost; Fawcett, 2016, p. 2) e o profissional desse campo busca justamente colher conhecimentos por meio de tudo que é obtido. Para tanto, o dado não pode ser trabalhado da forma pura que ele chega ao cientista, pois – por si só – ele não traz nenhuma informação e ainda pode vim cheio de ruídos e inconsistências, sendo essencial o tratar e laborar devidamente para agregar valor à resolução do problema de negócios (Eidelwein, 2023).

Para tanto, o responsável pelo projeto pode fazer uso de poderosas ferramentas e tecnologias analíticas para o apoiar nessa etapa, que vão desde planilhas no Excel até códigos de programação em Python. A escolha de quais recursos computacionais irá utilizar irá

depender muito das competências do cientista e, sobretudo, do que melhor responde as necessidades dele no período de tempo do projeto. Afinal, ele deve compreender se vale, dentro da relação esforço *versus* tempo, considerar ferramentas mais acessíveis e fáceis de se manipular em curto prazo ou aprender novas técnicas que possam melhorar a solução à longo prazo.

De qualquer maneira, os dados devem ser "manipulados e convertidos em formas para que rendam melhores resultados" (Provost; Fawcett, 2016, p. 31) e, para isso, foi utilizado o conhecimento do cientista responsável em linguagem de programação Python. Ela vem se tornando popular nos últimos 20 anos e, conforme mencionado por Wes McKinney (2023), passou a deixar de ser somente uma linguagem de computação científica inovadora usada por poucos, para ser a mais utilizada no mundo empresarial e acadêmico para diversos fins, que vão desde o desenvolvimento de *softwares* até a criação de modelos de *Machine Learning*, como é o caso do presente projeto. Assim sendo, a linguagem Python foi aplicada em todo o processo de preparação dos dados, dividido em quatro etapas.

4.3.1 Integração

Para iniciar, foi necessário realizar a integração entre as diferentes planilhas na base principal, pois, assim como um dado por si só não carrega muita informação (Eidelwein, 2023), uma só planilha também tem sua eficácia reduzida se não usada com o apoio de outras. Assim, a associação das diferentes tabelas foi realizada por meio da coluna "ID_Associado", que foi designada como chave-primária para configurar a junção dos dados a partir das informações de cada membro ativo e inativo, sendo impossível sem ela. Para conhecimento, uma chave-primária tem como propósito "fornecer uma identidade exclusiva para cada registro" (Nield, 2023, p. 93), evitando duplicatas no modelo final e mantendo a integridade dos dados.

4.3.2 Limpeza

A seguir ocorreu a limpeza das informações, sendo essa uma das partes mais fundamentais para a melhoria do modelo de Machine Learning. Tal etapa inclui a eliminação manual de atributos/colunas que não sejam necessárias para o modelo e também a identificação e preenchimento de valores que estejam nulos (Goldschimdt; Passos, 2015; Pimentel, 2019; Martins, 2021). No primeiro caso, notou-se que muitas colunas repetiam informações e, para evitar a sobrecarga do modelo e com o intuito de construir uma base compacta (Batista, 2022), elas foram eliminadas. Além disso, quando se trata de valores ausentes, múltiplas linhas e células da tabela estavam em branco, o que poderia prejudicar a eficiência do modelo de Machine Learning. Isto posto, foi imprescindível verificar a ocorrência desses casos e entender qual a justificativa para o valor está nulo, isso porque, a título de exemplo, caso um cliente não tenha realizado nenhum investimento até o momento de recolhimento da base, esse valor seria reconhecido como em branco, mas não necessariamente seria preciso o eliminar, pois, ainda assim, é uma informação crucial saber que o membro não realizou nenhum investimento.

Nesse sentido, em casos semelhantes ao que foi ilustrado, a técnica mais prudente a ser tomada pelo cientista de dados é preencher os valores nulos do atributo afetado por um número ou texto que reflita o que seria esperado estar na coluna, nesse caso, um zero. Vale ressaltar que, em outras situações, o caminho seguido pelo profissional pode ser diferente (como colocar a média ou mediana dos valores, preencher com o último número, com o que mais aparece), tudo dependerá da interpretação do cientista de dados acerca da informação prevista na coluna.

4.3.3 Engenharia de Variáveis (*Feature Enginnering*)

Após a limpeza, o cientista de dados pode perceber através da Análise Exploratória realizada na etapa de Entendimento dos dados que ainda existem variáveis que expliquem melhor o problema de negócios, mas que não podem ser encontradas de forma natural na base de dados, devendo ser criadas a partir das informações já coletadas. Diante disso, o processo que se segue é chamado de Engenharia de Variáveis ou *Feature Enginnering*, no qual novos valores são construídos tendo como base uma ou mais colunas existentes. Pode-se citar, por exemplo, a estimativa de idade através da data de nascimento com uma simples subtração da data de hoje menos o dia que o cliente nasceu. Outrossim, é possível ir além e agrupar esses membros por faixa etária, permitindo uma análise mais objetiva, processo esse efetuado e apresentado na figura 9.

Figura 9 – Demonstração da criação das colunas “Idade” e “Faixa Etária”

```
# Calcula a idade
df_final['idade'] = (pd.to_datetime('today') - df_final['nascimento']).dt.days // 365.25

# Calcula a faixa etária
def faixa_etaria(idade):
  if idade < 30:
    return "Menor que 30"
  elif 30 <= idade < 40:
    return "Entre 30 e 40 anos"
  elif 40 <= idade < 50:
    return "Entre 40 e 50 anos"
  elif 50 <= idade < 60:
    return "Entre 50 e 60 anos"
  elif idade > 70:
    return "Acima de 70"

df_final['faixa_etaria'] = df_final['idade'].apply(faixa_etaria)
```

Fonte: elaborado pelo autor, 2024.

4.3.4 Seleção e tratamento dos dados

Em seguida ocorreu a seleção e transformação das variáveis, para serem apropriadas a um modelo de *Machine Learning*, mediante a codificação dos atributos (Goldschimdt; Passos, 2015; Martins, 2021). Esse processo acontece em razão dos requisitos que determinados algoritmos de Aprendizado de Máquina exigem para serem executados (Provost; Fawcett, 2016), como, a título de exemplo, que todas as variáveis sejam numéricas em caso de modelos de regressão logística. Para tanto, todos os atributos passaram por um processo de *One-Hot Encoding*, onde as variáveis categóricas foram convertidas em números (em sua maioria de ordem binária) para que a máquina possa os armazenar e compreender com maior eficiência, aumentando a performance dos modelos de Machine Learning a serem criados.

Portanto, a etapa de preparação dos dados engloba várias subtarefas que têm como propósito substancial tratar os dados antes de sua aplicação propriamente dita no modelo final de Aprendizado de Máquina (Eidelwein, 2023; Martins, 2021). Essa fase pode consumir um tempo considerável do cientista de dados responsável, dado que ocorre um processo iterativo e

cíclico de seleção, exploração, avaliação e eliminação de variáveis até se encontrar o melhor conjunto de elementos que respondam as necessidades da situação-problema. Decerto, Foster Provost e Tom Fawcett (2016), autores da obra "Data Science para Negócios: o que você precisa saber sobre Mineração de dados e Pensamento analítico de dados", afirmam que é nesse momento que a criatividade humana, o bom senso e o conhecimento de negócios entram em jogo, pois, logo em seguida vem a etapa mais experimental da metodologia CRISP-DM, a modelagem.

4.4 Modelagem

Sendo uma das fases mais dinâmicas para o cientista de dados, a modelagem é o momento em que o dono do projeto irá escolher e aplicar diferentes técnicas computacionais, matemáticas e estatísticas para desenvolver a solução voltada ao problema de negócios. Essa resolução se dá através de um modelo, que nada mais é do que uma "representação simplificada da realidade" (Provost; Fawcett, 2016, p. 46).

Para facilitar o entendimento, pode-se imaginar um engenheiro cartográfico que tem como missão produzir um mapa. Ao desenvolver, ele precisará conhecer o objetivo a ser atingido, coletar dados, escolher quais deles melhor se encaixam ao propósito do mapa a ser construído e compreender a forma de dispor todo esse conhecimento, escolhendo a melhor técnica de desenho para isso. Da mesma forma se comporta um modelo preditivo em ciência de dados, ele irá se ajustar ao problema de negócios e o profissional responsável irá – a depender das necessidades – abstrair ou destacar detalhes que sejam relevantes ao modelo final (Provost; Fawcett, 2016).

Além disso, o cientista de dados deve se considerar dentro de um ambiente de laboratório, no qual o computador será a sala e todo o conhecimento e compostos químicos à disposição serão a sua base de dados. Nesse espaço deverá ter uma área de treinamento, em que ele irá testar diferentes substâncias e dosagens, e uma de testes, em que os experimentos serão realizados e analisados em tempo real. No entanto, para que se obtenha bons resultados na segunda etapa, boa parte dos recursos devem ser direcionados para o treinamento, a fim de realizar testes mais eficazes e menos custosos à longo prazo. De tal forma ocorre na modelagem, no qual os dados são divididos em bases de treino e teste e também categorizados quanto as variáveis explicativas do modelo (x) e o valor que se busca prever, a variável alvo (y).

Em seguida, ele deve conhecer e selecionar as melhores técnicas computacionais que irão o ajudar em seu principal objetivo, que é prever um valor desconhecido a partir de um conjunto finito de dados históricos. Contudo, isso não é uma tarefa fácil, devido à vasta oferta de algoritmos e ferramentas matemáticas-estatísticas para aplicação, sendo uma das principais qualidades do cientista de dados ter disposição e curiosidade para conhecer cada uma delas. Serão utilizados nesse projeto sob a problemática de predição do *churn* os algoritmos de Regressão Logística – com e sem regularização LASSO – e Florestas Aleatórias (*Random Forest*) – modelos esses bastante trabalhados por grande parte da literatura sobre a área de Data Science – e, a fim de se buscar diferentes perspectivas e promover mais disrupções no campo de ciência de dados, também foi utilizada a técnica de *boosting*, o XGBoost.

4.4.1 Regressão Logística

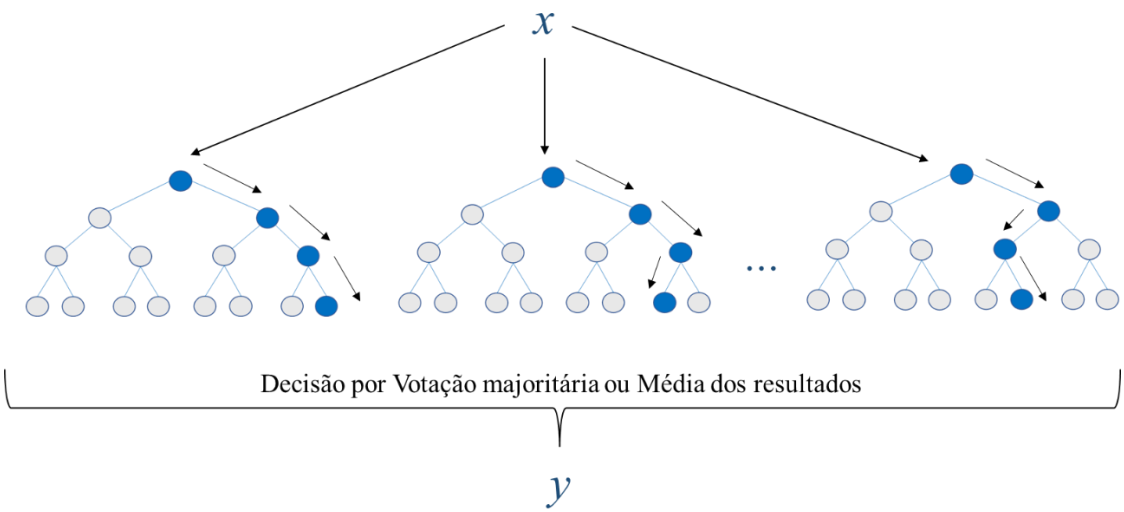
Para conhecimento, a técnica de regressão logística fornece a probabilidade de um evento binário de interesse ocorrer ou não, baseada em uma função matemática que mapeia as variáveis de resultado (independentes) e as classifica – de acordo com um limiar de probabilidade – em um valor 0 e 1. Para isso, ela considera todos os aspectos conhecidos da base de dados, que foram previamente extraídos e organizados de maneira quantitativa para

aplicação computacional, como participações em eventos, quantidade de investimentos, data de entrada e todos os demais da planilha. Todavia, a regressão logística em seu modelo mais simples e puro considera todos os atributos disponíveis, o que pode acarretar em erros de generalização como *overfitting*, problema em que o algoritmo se ajusta demais aos dados de treinamento fornecidos e perde, por consequência, a capacidade de generalizar para novos dados, afetando significativamente o poder preditivo. Para preveni-los, os cientistas de dados empregam algumas técnicas de regularização como o LASSO, no qual são identificadas as variáveis menos importantes e – mediante um critério de penalização – reduzidas para zero, resultando em um modelo fácil de interpretar, mais parcimonioso e com melhor capacidade de generalização, pois apenas os atributos mais relevantes são incluídos na modelagem final.

4.4.2 Florestas Aleatórias (Random Forest)

Já o algoritmo de Florestas Aleatórias é um "método de aprendizagem de máquina combinado" (Silveira, 2022, p. 22) composto por uma série de árvores simples que, através de múltiplas regras e decisões paralelas, resultam em uma previsão mais precisa e robusta acerca do problema de negócios trabalhado. Tal estimativa se dá através de uma média geral ou votação majoritária, em que a própria inteligência de máquina irá definir, a partir das respostas obtidas pelas árvores geradas aleatoriamente, o resultado do modelo.

Figura 10 – Esquematização do modelo de Random Forest



Fonte: adaptado de Martins, 2021.

Dada sua versatilidade e capacidade de trabalhar com grande volume de dados e variáveis de diferentes tipos, a técnica de Random Forest tem sido uma escolha frequente para lidar com problemas de classificação e regressão no mundo dos negócios.

4.4.3 XGBoost

Complementando o Random Forest, a técnica de XGBoost oferece uma solução alternativa poderosa na construção de modelos preditivos. Baseando-se também em árvores de decisão para gerar resultados, o XGBoost emprega uma abordagem de *boosting*, construindo modelos de árvores sequencialmente e – diferente dos demais algoritmos – aprendendo de

maneira contínua e dinâmica com os erros anteriores, corrigindo-os e passando para a geração de uma nova árvore de decisão. Assim, tendo a visão orientada por aprendizagem de tentativa e erro, esse algoritmo vem ganhando força devido à capacidade de capturar relações mais profundas entre as variáveis e gerar por sua vez, resultados cada vez mais precisos e eficientes na construção de modelos preditivos complexos.

Logo, foram desenvolvidos 4 algoritmos de *Machine Learning* considerando, para isso, diferentes hiperparâmetros – configurações externas definidas pelo programador que controlam a estrutura e o comportamento de um modelo de aprendizado da máquina – e métodos da biblioteca *scikit-learn*, que é amplamente utilizada por programadores, especialmente no campo da Aprendizado de Máquina (McKinney, 2023). Dentre os métodos utilizados, destacam-se o Grid e o Random Search, que permitiram identificar a melhor combinação de variáveis nos modelos de Random Forest e XGBoost.

Posteriormente, foram executados os testes de acordo com os 25% da base de dados separada aleatoriamente para essa etapa. Por fim, conforme apresentando na figura 11, foi desenvolvido um código em Python – através do método predict_proba – que estimasse a probabilidade de um cliente ativo se tornar um *churn* e apresentasse os 10 membros com maiores chances de cancelamento da assinatura dentro da Associação de Investidores-anjos.

Figura 11 – Demonstração do código de previsão da probabilidade de *churn*

```
# Selecionando clientes ativos
clientes_ativos = df_quant[df_quant['y'] == 0]

# Removendo a coluna 'ID_Associado' antes da previsão
X_clientes_ativos = clientes_ativos.drop(['ID_Associado', 'y'], axis=1)

# Prever a probabilidade de churn para clientes ativos
clientes_ativos['probabilidade_churn'] = rf_model_grid.predict_proba(X_clientes_ativos)[:, 1]

clientes_ativos = clientes_ativos.sort_values(by=['probabilidade_churn'], ascending=False)

# Exibir os 10 clientes com maior probabilidade de churn
clientes_em_risco = clientes_ativos[['ID_Associado', 'probabilidade_churn']].head(10)
```

Fonte: elaborado pelo autor, 2024.

4.5 Avaliação

Na penúltima etapa do CRISP-DM, quatro modelos de *Machine Learning* foram avaliados de acordo com os princípios da Ciência de Dados. Tal processo investigativo é fundamental para apurar se os algoritmos utilizados foram devidamente construídos – no que tange à confiabilidade e validade do modelo (Provost; Fawcett, 2016; Martins, 2019) – e se respondem as principais perguntas geradas ainda na fase de compreensão de negócios. Para tanto, o cientista de dados pode fazer uso das métricas de avalição que levam em conta a matriz de confusão, uma estrutura matricial que contabiliza os casos de verdadeiro positivo (VP), verdadeiro negativo (VN), falso positivo (FP) e falso negativo (FN), permitindo mensurar e examinar o quanto o modelo está errando e acertando de acordo com o problema de negócios estudado, no caso, o *churn* (Silveira, 2022; Eidelwein, 2023).

Em síntese, os casos mais importantes para avaliar no contexto aplicado são os de verdadeiros positivos (VP), pois eles representam os *churns* reais, ou seja, que o modelo previu como cancelamento e realmente efetuou o ato. Ao identificar quem são esses casos, a empresa pode agir de maneira proativa e enviar ofertas especiais para eles, a fim de captar o seu interesse e aumentar o tempo de participação no grupo. Em seguida, elenca-se os falsos negativos (FN) – o modelo previu como não *churn*, mas cancelou – e os falsos positivos (FP), que foram previstos como cancelamentos, mas não se tornou um fato. Por fim, os casos de verdadeiros negativos (VF) representam aqueles que não foram previstos e realmente não converteram no cancelamento da assinatura.

Figura 12 – Matriz de confusão do *Churn*

Previsão	Realidade	
	Verdadeiro	Falso
Positivo	**Churn real** (o modelo previu e acertou o churn)	**Churn equivocado** (o modelo previu, mas não era churn)
Negativo	**Membro ativo** (o modelo não previu como churn e acertou)	**Churn não previsto** (o modelo não previu, mas era churn)

Fonte: elaborado pelo autor, 2024.

Levando em conta os modelos construídos e as matrizes de confusão descritas abaixo na figura 13, todos apresentaram semelhanças quanto boa parte das relações de previsão e realidade, com exceção do modelo de Regressão Logística, que previu mais FP do que os restantes somados. Do ponto de vista corporativo, concentrar esforços em clientes que o modelo previu como *churn*, mas na verdade não cancelariam com ou sem qualquer plano de oferta é custoso, pois, embora seja de menor gravidade, o vendedor ou responsável pode oferecer descontos e bônus na tentativa de reter o membro que, talvez, nem seriam necessários para essa pessoa em específico.

Figura 13 – Matrizes de confusão dos modelos aplicados

Regressão Logística Simples

Previsão	Realidade	
	Verdadeiro	Falso
Positivo	51	26
Negativo	38	21

LASSO

Previsão	Realidade	
	Verdadeiro	Falso
Positivo	59	7
Negativo	57	13

Random Forest

Previsão	Realidade	
	Verdadeiro	Falso
Positivo	60	6
Negativo	58	12

XGBoost

Previsão	Realidade	
	Verdadeiro	Falso
Positivo	63	7
Negativo	57	9

Fonte: elaborado pelo autor, 2024.

Em contrapartida, todos os outros modelos – com exceção do algoritmo de Regressão Logística – apresentaram um número de falsos negativos superior ao de falsos positivos, sendo um dado também preocupante, pois eles não previram o *churn* e, dependendo do cliente, pode haver uma grande perda de receita por parte da instituição sem oportunidade de intervir, sendo tarefa do cientista de dados trabalhar continuamente para reduzir esse número e aumentar a performance do algoritmo de Aprendizado de Máquina. Contudo, antes de considerar esses valores como verdade absoluta, ele deve compreender o quão confiáveis esses números são, tarefa essa que pode ser facilitada através do uso de métricas que nasceram da matriz de confusão, como a acurácia, precisão, *recall* e F1-score .

Iniciando com a mais relevante para o estudo de retenção de *churns*, a sensibilidade – conhecida popularmente como *recall* – mede a porcentagem de VP que o modelo classificou corretamente, ou seja, quem realmente estava previsto de cancelar e aconteceu de fato (Serpa, 2023). Dessa forma, quanto maior for o seu resultado, maior será a proporção de membros que realmente cancelaram e que, por sua vez, foram corretamente identificados pelos modelos, oferecendo para a empresa uma oportunidade de evitar que esse cenário ocorra de fato. Enquanto isso, a precisão avalia a relação entre o número de classificações corretas de *churns* (VP) sobre o montante de previsões positivas (VP + FP), demonstrando, dessa forma, o quão preciso é o modelo (Batista, 2022; Serpa, 2023). Contudo, ao contrário do que se espera, a precisão tem um papel menos crítico quanto ao *churn*, uma vez que um modelo com alta precisão não necessariamente identificará todos os clientes que irão cancelar, podendo, inclusive, diminuir o valor de *recall*.

Um bom equilíbrio entre essas duas métricas é a avaliação por meio do F1-Score, que se dá mediante o cálculo da média harmônica entre precisão e *recall*. Sua utilidade é observada quando se busca um bom desempenho tanto na identificação de VP quanto na minimização de FP e FN. Quanto maior o seu resultado, mais equilibrados são os valores de precisão e recall, proporcionando uma maior confiabilidade ao modelo que o apresenta mais próximo de 1.

E embora seja uma das métricas mais comumente utilizadas em projetos de *Machine Learning*, a medida de acurácia não é recomendada para modelos preditivos de cancelamento de clientes. Pois, por se tratar de um problema de negócios geralmente desbalanceado na grande maioria dos projetos – onde uma classe pode ser maior que a outra –, ela pode não ser tão precisa quanto as outras, gerando múltiplos alarmes falsos que irão prejudicar a empresa como um todo (Provost; Fawcett, 2016).

Tendo conhecimento das métricas e de como as avaliar sob o contexto do problema de negócios, o próximo passo é analisar seus desempenhos frente aos modelos de *Machine Learning* em estudo. Para isso, foi considerado o algoritmo de Regressão Logística como ponto de partida devido a sua simplicidade e rapidez de treinamento, o que facilita a interpretação inicial dos resultados. Além dele, foram selecionadas as técnicas computacionais de LASSO – devido a sua capacidade de realizar seleção de *features* e gerar modelos mais parcimoniosos, facilitando a interpretação dos resultados e identificando os fatores mais relevantes para a decisão de *churn* –, *Random Forest* – pela sua ampla utilização em estudos recentes sobre análise preditiva de *churns* e capacidade de lidar com diferentes tipos de dados – e XGBoost, por sua alta performance e inovação pela forma como a tecnologia lida com os próprios erros do modelo. Vale destacar que a escolha das técnicas de Aprendizado de Máquina foi baseada não só em suas características, mas também visualizando a complexidade do problema de *churn* e os dados e recursos computacionais disponíveis no momento.

Portanto, conforme a tabela 1 abaixo, notou-se que a análise comparativa dos quatro modelos selecionados revelou particularidades interessantes em cada uma das métricas. Começando pela Regressão Logística, ficou claro que a simplicidade do modelo não seria capaz de suprir as necessidades do problema de *churn*, pois, como observado, todas as medidas de avaliação foram muito inferiores a qualquer modelo de Aprendizado de Máquina proposto

adiante. Nesse sentido, a técnica de LASSO já teve uma melhora significativa média de 0,18 se comparado a Regressão Logística, comprovando que para resolver o problema de negócios é necessário recorrer a técnicas mais complexas e que permitam extrair o máximo de informações para predizer o *churn*.

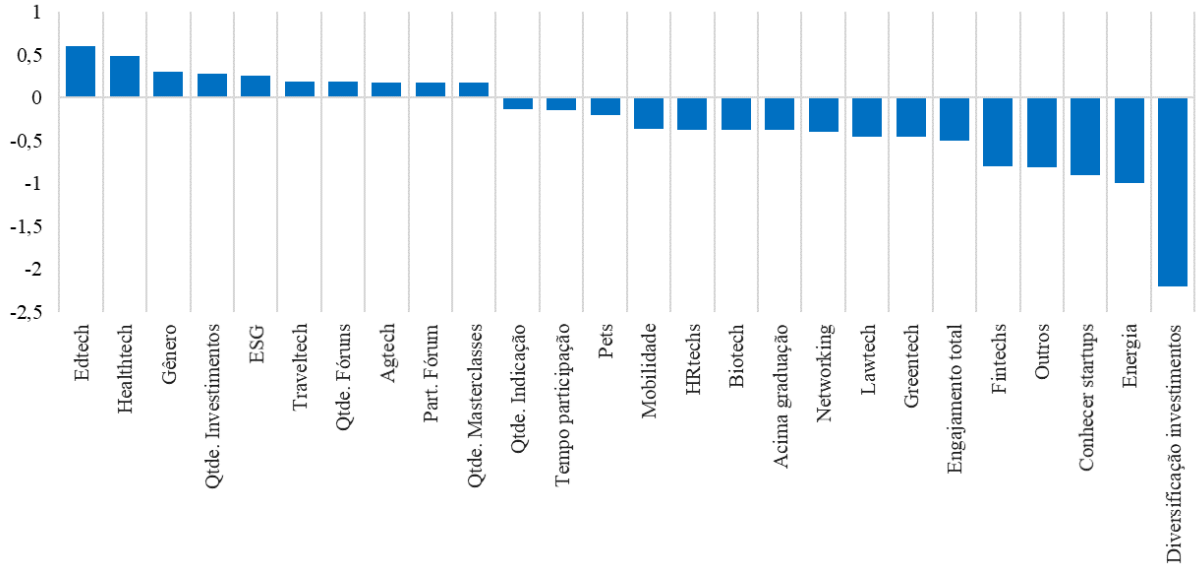
Tabela 1 – Métricas de avaliação por modelo de Aprendizado de Máquina

Métrica	Regressão Logística	LASSO	Random Forest	XGBoost
Recall	0.71	0.82	0.83	**0.88**
Precisão	0.66	0.89	**0.91**	0.90
F1-Score	0.68	0.86	0.87	**0.89**
Acurácia	0.65	0.85	0.87	**0.88**

Fonte: elaborado pelo autor, 2024.

Além disso, conforme demonstrado na figura 14, o algoritmo de LASSO possibilitou verificar padrões mais profundos nos dados e investigar como as *features* se comportam proporcionalmente com a tendência de cancelamento, como, por exemplo, o tempo de participação do membro do grupo que está em uma relação inversamente proporcional, ou seja, quanto menor o tempo que o associado fica, maior a chance de cancelar o serviço.

Figura 14 – Valores dos coeficientes na técnica de LASSO



Fonte: elaborado pelo autor, 2024.

Em contrapartida, o Random Forest se mostrou eficiente na precisão dos valores com 0.91. Mas, como já mencionado anteriormente, essa métrica em específico tem uma significância relativamente menor na problemática de *churn*. Ao contrário do que ocorreu com o modelo de XGBoost, que se revelou como sendo um dos algoritmos mais eficientes em compreender e capturar plenamente os casos de cancelamentos de clientes na associação, exibindo um dos melhores desempenhos nas métricas de *recall* e F1-Score, ambas intrinsecamente envolvidas no contexto do problema de negócios. Tal predominância do XGBoost sob os demais métodos se dá porque ele é capaz de aprender com os dados e os erros cometidos durante o processo de modelagem, permitindo com que o algoritmo se instrua constantemente com novas informações e, conforme a noção de Muoio (1997), obtenha sucesso em sua performance mais cedo.

Para corroborar tais fatos, foi analisada a significância estatística do modelo de XGBoost em comparação a Regressão Logística, com o intuito de investigar se os resultados alcançados pela técnica de *boosting* foram realmente causados pela superioridade do algoritmo ou simplesmente ao acaso. Em resumo, ela verifica se o modelo escolhido realmente pode ser implementado com confiabilidade ou se é melhor manter um baixo custo computacional com uma técnica mais simples como a regressão logística. De resultado, comprovou-se que o modelo de *boosting* prevê 4,9% mais os *churns* do que o outro, ao possuir significância estatística ao nível de 1% e estatística-t de 61.471. Além do mais, a diferença média de 0.049 entre os modelos é estatisticamente significante, ou seja, o modelo complexo de XGBoost está realmente performando melhor do que o modelo base de Regressão Logística. Portanto, pode-se constatar que os resultados superiores alcançados pelo XGBoost não podem ser atribuídos ao acaso e sim graças à capacidade preditiva do modelo.

Destarte, diante das observações feitas, é inegável que o XGBoost apresenta um desempenho altamente superior na previsão do *churn*, com resultados significativamente melhores do que todos os demais modelos comparados. Logo, ao adotar o XGBoost em suas operações, o time da Associação de Investidores-Anjo da FGV poderá identificar de forma mais precisa os clientes em risco de cancelamento e, com isso, tomar ações preventivas baseadas nesses dados que venham a reduzir a taxa de *churn*.

4.6 Implantação

Para finalizar os resultados, indo além de avaliar previamente as métricas de sucesso, o projeto de dados foi testado em um ambiente controlado de laboratório, pois, conforme afirmam Provost e Fawcett (2016), isso confere segurança, agilidade e facilidade ao dono do projeto antes de avançar para um contexto de uso real, onde riscos financeiros e operacionais estão oficialmente em jogo. Como a base de dados utilizada para o estudo foi coletada cerca de 2 semanas antes do desenvolvimento do projeto de Aprendizado de Máquina, foi possível observar em um curto espaço de tempo os resultados que os modelos de LASSO, *Random Forest* e XGBoost apontaram como maiores probabilidades de cancelamento – ou seja, os clientes com risco de converterem em *churn* – e verificar como eles se comportaram na realidade, investigando, assim, o potencial de acertos e pontos de atenção de cada método na prática.

Ao todo, foram 19 clientes únicos que os modelos identificaram como mais propensos a cancelar o serviço e, dentre eles, 2 realmente converteram em *churn*, sendo que ambos foram detectados nos modelos de Random Forest e XGBoost. Além deles, foram previstos outros 5 nomes com alta probabilidade de cancelamento nos próximos 90 dias, o que permitirá a empresa trabalhar melhor esses casos e evitar o aumento da taxa de *churn*. Fora isso, um outro caso intrigante foi observado um dia após esse sistema de probabilidade ficar pronto para testes, no qual uma das clientes que apareceu como risco de cancelamento em todos os modelos foi

mencionada durante uma reunião de equipe diária da empresa, falando em como essa pessoa estava pouca engajada com o grupo e, inclusive, verbalizou a vontade de cancelar o serviço poucos dias antes.

Nesse momento que se percebeu e comprovou com fatos como os dados e a realidade da associação estão intrinsecamente interligados, bastando apenas que o profissional adequado com as tecnologias corretas em mãos lidasse com o que hoje é considerado um dos ativos estratégicos mais valiosos das empresas, os dados (Provost; Fawcett, 2016).

5. CONSIDERAÇÕES FINAIS

Em suma, o presente trabalho teve como principal objetivo compreender o uso das novas tecnologias inovadoras de Ciência de Dados e modelos preditivos na área de Gestão dos Clientes e aplicá-las dentro do contexto de negócios da Associação de Investidores-anjos da Fundação Getúlio Vargas, como forma de reduzir o *churn* de clientes e respaldar as tomadas de decisões estratégicas orientadas por dados dentro da organização daqui em diante. Para tal estudo, foi empregado como metodologia o modelo CRISP-DM, que norteou todo o processo de compreensão do negócio, entendimento, preparação dos dados, modelagem, avaliação e implantação presente e futura no cenário organizacional.

Dentre os principais resultados do trabalho, evidencia-se a criação de modelos de Aprendizado de Máquina voltados para a análise preditiva de *churns* dentro da associação, ganhando destaque o algoritmo de XGBoost dentre os demais, por apresentar valores de *recall* e F1-Score – métricas demonstradas como mais indicadas para predição de cancelamento de clientes – superiores às técnicas de Regressão Logística, LASSO e Random Forest. Ademais, deve-se enfatizar também que o processo de execução da metodologia CRISP-DM proporcionou a criação de uma estrutura de dados robusta que hoje está sendo utilizada como principal centro de informações da empresa e gerenciada integralmente pelo dono desse projeto, promovendo, assim, uma cultura mais orientada por dados e que já está sendo implantada nas práticas da organização. Além das implicações práticas mencionadas, a pesquisa sobre a temática ampliou as possibilidades acadêmicas de estudo a respeito do emprego da Ciência de Dados e suas técnicas e métodos tanto na Gestão de Clientes, como no cenário corporativo como um todo – uma vez que foram utilizados quatro diferentes modelos de *Machine Learning* e técnicas estatísticas distintas simultaneamente, permitindo que a teoria acadêmica se una a prática empresarial com foco em resolver problemas reais e, conjuntamente, promover inovações em ambos os contextos.

Quanto as limitações encontradas nesse trabalho, ressalta-se que o pequeno tamanho da amostra coletada de clientes ativos e inativos balizou fortemente a pesquisa, visto que implicou em complicações ao desenvolver os modelos e encontrar a melhor forma de balancear os dados. Outrossim, em razão da limitação de tempo, pontua-se a exploração de somente quatro modelos de *Machine Learning* para aplicação, o que – embora se assemelhe a maioria dos estudos – pode representar uma oportunidade para ser aplicada por outros pesquisadores ou até mesmo pelo cientista de dados responsável em sua área na empresa, que pode, mais adiante, mesclar mais de um modelo preditivo e criar um sistema próprio, proporcionando uma inovação que vai além do nível de complexidade adotado nesse estudo. Logo, futuras investigações podem não só ampliar os modelos de pesquisa, como também configurar a base conforme técnicas para balanceamento, amostragem e revisão estatística de nível mais complexo, gerando, por sua vez, modelos mais intricados e eficientes de acordo com o problema de negócio.

Portanto, fica claro que o presente estudo impactou positivamente na transformação ativa da Gestão de Relacionamento e Decisões de Negócios na Associação de Investidores-Anjos da Fundação Getúlio Vargas. A elaboração tanto da base de dados intitulada “GV Aplanilha” como do modelo XGBoost para predição de *churns* representam um marco crucial

na área de Gestão de Relacionamento do Associado na empresa, permitindo com que ela caminhe rumo a uma cultura orientada por dados, em que as decisões são fundamentadas a partir de fatos. Destaca-se como próximos passos a implantação do modelo preditivo, o monitoramento contínuo do seu desempenho e a exploração constante de outras técnicas de *Machine Learning*, pois, afinal, novas tecnologias surgirão a todo momento e é essencial que o cientista de dados se mantenha atualizado sobre cada transformação, a fim de melhorar continuamente os processos e conferir, por meio da inovação, uma vantagem competitiva sustentável no mercado.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ALVES, George Victor de Souza; LIMA, Lucas Azevedo Rêgo; OLIVEIRA, Lucas Matheus da Silva. Abordagem analítica para predição e prevenção do Churn. **Revista de Engenharia e Pesquisa Aplicada**, [S.L.], v. 7, n. 3, p. 64-72, 30 nov. 2022.

AMARO JUNIOR, Edson *et al*. Artificial intelligence and Big Data in neurology. **Arquivos de Neuro-Psiquiatria**, [S.L.], v. 80, n. 51, p. 342-347, maio 2022.

ARAÚJO, José Maria Amorim. **Análise de sobrevivência e previsão de churn de clientes de seguros de vida do Banco do Brasil**. 2022. 70 f. Dissertação (Mestrado) - Curso de Computação Aplicada, Departamento de Ciência da Computação, Universidade de Brasília, Brasília, 2022.

BATISTA, Igor de Carvalho de Brito. **Análise de influência de características no modelo de predição de churn**. 2022. 51 f. TCC (Graduação) - Curso de Engenharia de Computação, Centro de Tecnologia, Universidade Federal do Rio Grande do Norte, Natal, 2022.

BLEI, David M.; SMYTH, Padhraic. Science and data science. **Proceedings Of The National Academy Of Sciences**, [S.L.], v. 114, n. 33, p. 8689-8692, 7 ago. 2017.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. **Lei Geral de Proteção de Dados Pessoais (LGPD)**. Brasília, DF: Presidência da República, 2024. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 02 de outubro de 2024.

BRYNJOLFSSON, Erik; HITT, Lorin M.; KIM, Heekyung Hellen. Strength in Numbers: how does data-driven decisionmaking affect firm performance?. **Ssrn Electronic Journal**, [S.L.], v. 1, n. 1, p. 1-33, abr. 2011.

BUREZ, Jonathan *et al*. CRM at a pay-TV company: using analytical models to reduce customer attrition by targeted marketing for subscription services. **Expert Systems With Applications**, [S.L.], v. 32, n. 2, p. 277-288, fev. 2007.

EIDELWEIN, Rodrigo. **Modelos preditivos e predição de churn**: analytics aliado à administração. 2023. 54 f. TCC (Graduação) - Curso de Administração, Escola de Administração, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2023.

FRANCESCHI, Pietro Reinheimer de. **Modelagens preditivas de Churn**: o caso do banco do brasil. 2019. 121 f. Dissertação (Mestrado) - Curso de Administração, Unidade Acadêmica de Pesquisa e Pós-Graduação, Universidade do Vale do Rio dos Sinos, Porto Alegre, 2019.

GAO, Xiang; WEN, Junhao; ZHANG, Cheng. An Improved Random Forest Algorithm for Predicting Employee Turnover. **Mathematical Problems In Engineering**, [S.L.], v. 2019, n. 1, p. 1-12, jan. 2019.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel Lopes. **Data Mining**: um guia prático. [S. L.]: Elsevier, 2015. 296 p.

HERBERT, Frank. **Duna**. 2. ed. [S. L.]: Aleph, 2017. 680 p.

JAIN, Nishant; TOMAR, Abhinav; JANA, Prasanta K.. A novel scheme for employee churn problem using multi-attribute decision making approach and machine learning. **Journal Of Intelligent Information Systems**, [S.L.], v. 56, n. 2, p. 279-302, 29 set. 2020.

JOSEPH, Richard *et al*. Employee Attrition Using Machine Learning And Depression Analysis. **2021 5Th International Conference On Intelligent Computing And Control Systems (Iciccs)**, [S.L.], v. 1, n. 1, p. 1000-1005, 6 maio 2021.

KOTLER, Philip; KELLER, Kevin Lane. **Administração de Marketing**. 14. ed. São Paulo: Pearson, 2013.

MACHINE Learning: O que é e qual sua importância. **SAS Insights**, 2023. Disponível em: https://www.sas.com/pt_br/insights/analytics/*machine-*

learning.html#:~:text=O%20aprendizado%20de%20m%C3%A1quina%20(em,o%20m%C3%ADnimo%20de%20interven%C3%A7%C3%A3o%20humana. Acesso em: 02 de mai. de 2024.

MARTENS, David; PROVOST, Foster. Pseudo-social network targeting from consumer transaction data. **New York University.** Nova Iorque, p. 1-31. set. 2011.

MARTINS, Lucas Gomes. **Aplicação de um modelo de aprendizagem de máquina para predição de churn**: um estudo de caso. 2021. 34 f. TCC (Graduação) - Curso de Engenharia Metalúrgica, Departamento de Engenharia Metalúrgica e de Materiais, Universidade Federal do Ceará, Fortaleza, 2021.

MCKINNEY, Wes. **Python para Análise de Dados**: tratamento de dados com pandas, numpy e ipython. 3. ed. São Paulo: Novatec, 2023. 620 p.

MUOIO, Anna. **They Have a Better Idea … Do You**. Fast Company, 31 de agosto de 1997. Disponível em: https://www.fastcompany.com/29116/they-have-better-idea-do-you. Acesso em: 11 de outubro de 2024.

NESLIN, Scott A. *et al*. Defection Detection: measuring and understanding the predictive accuracy of customer churn models. **Journal Of Marketing Research**, [S.L.], v. 43, n. 2, p. 204-211, maio 2006.

NIELD, Thomas. **Introdução à Linguagem SQL**: abordagem prática para iniciantes. São Paulo: Novatec, 2016. 141 p.

O que é Machine Learning. **IBM**, 2024. Disponível em: https://www.ibm.com/br-pt/topics/machine-learning. Acesso em: 09 de mai. de 2024.

O que é Machine Learning. **Oracle**, 2024. Disponível em: https://www.oracle.com/br/artificial-intelligence/machine-learning/what-is-machine-learning/. Acesso em: 09 de mai. de 2024.

ÓSKARSDÓTTIR, María et al. Time series for early churn detection: using similarity based classification for dynamic networks. **Expert Systems With Applications**, [S.L.], v. 106, p. 55-65, set. 2018.

PIMENTEL, Thiago Paiva. **Predição de churn baseada em detecção de padrões sequenciais e análise de sentimentos sobre as interações de clientes no CRM**. 2019. 66 f. Dissertação (Mestrado) - Curso de Ciências em Sistemas e Computação, Departamento de Ciência e Tecnologia, Instituto Militar de Engenharia, Rio de Janeiro, 2019.

PROVOST, Foster; FAWCETT, Tom. **Data Science Para Negócios**: o que você precisa saber sobre mineração de dados e pensamento analítico de dados livro. Rio de Janeiro: Alta Books, 2016. 384 p.

SAADIA, Giovanna Niskier *et al*. Machine Learning na Previsão do Risco de Inadimplência de Alunos do Ensino Superior. In: ENCONTRO DA ANPAD, 46., 2022, [S. L.]. **Anpad.com.br.** [S. L.]: Anpad, 2022. p. 1-19.

SALTZ, Jeff. **Metodologias mais usadas em projetos de Ciência de Dados**. Disponível em: https://www.datascience-pm.com/crisp-dm-still-most-popular/. Acesso em: 11 de out. de 2024.

SAMUEL, A. L.. Some Studies in Machine Learning Using the Game of Checkers. **Ibm Journal Of Research And Development**, [S.L.], v. 3, n. 3, p. 210-229, jul. 1959.

SERPA, Maria Luiza Rabelo. **Random Forest aplicado na análise de churn**: comparação do ajuste com dados completos versus ajuste em estratos definidos por covariável categórica. 2023. 33 f. Monografia (Especialização) - Curso de Estatística, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, Belo Horizonte, 2023.

SILVEIRA, Caio Cesar Vieira Trinta da. **Revisão e Aplicação de Métodos de Aprendizado de Máquina para a Predição de Churn**. 2022. 45 f. Monografia (Especialização) - Curso de Administração, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2022.

SRIVASTAVA, Praveen Ranjan; EACHEMPATI, Prajwal. Intelligent Employee Retention System for Attrition Rate Analysis and Churn Prediction. **Journal Of Global Information Management**, [S.L.], v. 29, n. 6, p. 1-29, 25 jun. 2021.

TENG, Mingfei *et al*. Exploiting the Contagious Effect for Employee Turnover Prediction. **Proceedings Of The Aaai Conference On Artificial Intelligence**, [S.L.], v. 33, n. 01, p. 1166-1173, 17 jul. 2019.

THE top 50 Angel Investing Groups. **CB Insights**, Nova Iorque, 22 de dez. de 2022. Disponível em: https://www.cbinsights.com/research/report/top-group-angel-investors-2022/. Acesso em: 03 de abr. de 2024.

ULLAH, Irfan *et al*. Churn Prediction in Banking System using K-Means, LOF, and CBLOF. **2019 International Conference On Electrical, Communication, And Computer Engineering (Icecce)**, [S.L.], v. 35, n. 18, p. 1-6, jul. 2019.

WANG, Gang *et al*. Big data analytics in logistics and supply chain management: certain investigations for research and applications. **International Journal Of Production Economics**, [S.L.], v. 176, p. 98-110, jun. 2016.

YUAN, Jia. Research on Employee Turnover Prediction Based on Machine Learning Algorithms. **2021 4Th International Conference On Artificial Intelligence And Big Data (Icaibd)**, [S.L.], v. 3, p. 114-120, 28 maio 2021.