



Predição com aprendizagem supervisionada

Banco Pan

Controle do Documento

Histórico de revisões

Data	Autor	Versão	Resumo da atividade
08/08/2022	Amanda Fontes, Gabriel Rios, Izabella Almeida	1.2.1	Adição do tópico "Contexto da indústria" e das Personas
10/08/2022	Izabella Almeida	1.2.2	Adição do Value Proposition Canvas
11/08/2022	Amanda Fontes, Gabriel Rios, Izabella Almeida, Livia Coutinho, Pedro Baptista	1.2.3	Adição da matriz SWOT, Matriz de Riscos, tópicos de Compreensão dos Dados, Planejamento Geral da Solução e Introdução
18/08/2022	Amanda Fontes, Gabriel Rios, Livia Coutinho	2.1.1	Adição da nova Persona e das User Journey Mappings, adição de legenda para as imagens
24/08/2022	Amanda Fontes, Gabriel Rios, Izabella Almeida	2.2.1	Adição das descrições das features utilizadas no modelo preditivo e agregação de registros (seção de Preparação dos Dados)
25/08/2022	Amanda Fontes, Livia Coutinho	2.2.2	Refino das descrições das features e adição da documentação referente às manipulações necessárias nos dados
29/08/2022	Gabriel Rios	3.1.1	Correção no tópico 4.3
06/09/2022	Amanda Fontes	3.2.1	Redação inicial da metodologia
08/09/2022	Izabella Almeida	3.2.2	Preenchimento da modelagem na seção 4.4 e avaliação na seção 4.5
12/09/2022	Amanda Fontes	4.1.1	Revisão dos itens 4.4 e 4.5 e atualização das legendas das imagens do documento
19/09/2022	Izabella Almeida	4.2.1	Reconstrução do item 1.
20/09/2022	Izabella Almeida, Gabriel Rios, Livia Coutinho	4.2.2	Preenchimento dos itens 2.1 e 2.3, continuação do preenchimento da seção 4.4 do documento, adicionando o conceito de hiperparâmetros, gráficos e tabelas. Preenchimento da seção 4.5 com descrição dos experimentos realizados até o momento e as respectivas análises.

22/09/2022	Amanda Fontes, Izabella Almeida, Gabriel Rios	4.2.3	Adição dos processos referentes à normalização e padronização dos dados, atualização das legendas das imagens. Revisão do documento.
23/09/2022	Amanda Fontes, Gabriel Rios, Livia Coutinho	4.2.4	Preenchimento da seção 3 (Metodologia)
06/10/2022	Gabriel Rios, Amanda Fontes, Lívia Coutinho	5.2.2	Finalização do texto da seção 3 (metodologia CRISP-DM) e revisão

Sumário

1. Introdução	4
2. Objetivos e Justificativa	5
2.1. Objetivos	5
2.2. Proposta de Solução	5
2.3. Justificativa	5
3. Metodologia	6
3.1. CRISP-DM	6
3.2. Ferramentas	6
3.3. Principais técnicas empregadas	6
4. Desenvolvimento e Resultados	7
4.1. Compreensão do Problema	7
4.1.1. Contexto da indústria	7
4.1.2. Análise SWOT	7
4.1.3. Planejamento Geral da Solução	7
4.1.4. Value Proposition Canvas	7
4.1.5. Matriz de Riscos	7
4.1.6. Personas	8
4.1.7. Jornadas do Usuário	8
4.2. Compreensão dos Dados	9
4.3. Preparação dos Dados	10
4.4. Modelagem	11
4.5. Avaliação	12
5. Conclusões e Recomendações	13
6. Referências	14
Anexos	15

1. Introdução

A empresa parceira deste módulo é o Banco Pan, instituição cujo principal objetivo é oferecer produtos voltados para as classes C, D e E, a fim de transformar desafios em conquistas. A instituição, que é sediada em São Paulo, ultrapassou a marca de quinze milhões de clientes no terceiro trimestre de 2021.

O problema a ser enfrentado está relacionado ao atendimento da companhia que, atualmente, não é personalizado para os diferentes propósitos dos clientes. Eles, de modo geral, sentem-se insatisfeitos com o atendimento telefônico do banco, porque, muitas vezes, não têm suas necessidades supridas ou, quando têm, elas não são supridas de forma completa. De modo geral, as pessoas entram em contato com o banco a fim de fazerem reclamações sobre os serviços contratados, adquirirem novos produtos ou cadastrarem-se como novos clientes.

Diante disso, o modelo preditivo que está sendo construído tem o principal propósito de prever se uma ligação recebida está sendo efetuada por um cliente atritado, um cliente engajado ou um potencial novo cliente. Essa classificação permitirá que os atendentes tenham um preparo prévio diante da lida com o público, o que evita, portanto, maiores insatisfações por parte dos clientes.

Esse processo está sendo efetuado por seis estudantes de tecnologia do Instituto de Tecnologia e Liderança, matriculados no segundo período, sendo eles: Amanda Fontes, Gabriel Torres, Izabella Faria, Lívia Coutinho, Pedro Baptista e Vinícios Lugli.

2. Objetivos e Justificativa

2.1. Objetivos

O Banco Pan deseja, de modo geral, a partir da implementação desse modelo preditivo, melhorar a avaliação da instituição no mercado e em canais de reclamações. Especificamente, o Pan deseja aprimorar o relacionamento com os clientes do banco que já tiveram ou ainda têm atrito em suas interações com a empresa.

2.2. Proposta de Solução

A partir dos objetivos do parceiro de negócios, a solução criada visa mitigar as dificuldades enfrentadas pela empresa quando o assunto é o relacionamento com clientes do banco. Assim, será criada uma predição com aprendizagem supervisionada capaz de classificar o status do cliente a partir dos dados oferecidos pelo banco, fazendo com que ele se enquadre em três principais grupos: aqueles que têm potenciais atritos com o banco, aqueles que são novos clientes ou os que estão dispostos a adquirir novos produtos. O foco dessa implementação deve ser atender melhor os clientes com atritos, a fim de melhorar essa relação entre os indivíduos e a empresa e facilitar o trabalho dos atendentes que precisam lidar com essas pessoas.

2.3. Justificativa

A fim de atender aos desejos da empresa parceira, esse modelo foi desenvolvido para conseguir prever possíveis clientes conflitados, o que possibilitará um atendimento mais eficiente e direcionado aos problemas trazidos por eles. Diante desse objetivo, o projeto tem o potencial de melhorar o relacionamento dos usuários com o banco e, a longo prazo, será capaz de gerar mais lucros e crescimento de mercado. Isso será possível porque a instituição financeira conseguirá engajar mais os consumidores, aumentará o potencial de suas avaliações positivas e, por fim, devido ao sucesso, captará ainda mais pessoas que se filiarão a essa empresa.

3. Metodologia

3.1. CRISP-DM

De acordo com o IBM SPSS Modeler CRISP-DM Guide, o Cross-Industry Standard Process for Data Mining (CRISP-DM) constitui uma das mais importantes metodologias relacionadas ao processo de mineração de dados. É por meio do CRISP-DM que os dados de uma empresa podem ser transformados em informações capazes de guiar o gerenciamento do negócio. A metodologia é composta pelas seguintes etapas:

Business Understanding: antes de iniciar o processo de mineração de dados, é necessário refletir sobre o que o cliente espera obter como resultado. Para isso, é fundamental examinar as metas, riscos e recursos disponíveis para o desenvolvimento do produto.

Data Understanding: uma vez compreendidos os objetivos de negócios, é necessário explorar a base de dados disponível, a fim de bem compreender o conjunto de informações que será minerado. Constituindo uma fase do CRISP-DM que apresenta uma significativa demanda por tempo, exige que sejam analisados com precisão os atributos presentes e os valores preenchidos nos registros existentes.

Data Preparation: constitui a etapa de preparação dos dados disponíveis para que eles possam ser devidamente lidos e interpretados pelos processos de mineração aos quais serão submetidos. É quando ocorre a Feature Engineering, constituída pela seleção de atributos que serão utilizados, bem como os procedimento de limpeza de dados, agregação de registros, derivação de novos atributos e separação de conjuntos de dados para treinamento e teste.

Modeling: a modelagem de dados é a fase na qual os dados, já preparados, são submetidos a diferentes algoritmos – utilizando, a princípio, os parâmetros padronizados do modelo. Posteriormente aos testes realizados, ocorre, ainda, a aplicação de uma série de técnicas de manipulação de dados responsáveis pelo refinamento dos modelos construídos, a exemplo das ferramentas de ajustes de hiperparâmetros.

Evaluation: nessa fase, os modelos testados a partir da base de dados são avaliados por meio de métricas definidas na fase de Business Understanding. São os chamados “critérios de sucesso”, os quais serão capazes de indicar se os procedimentos até então realizados estão tecnicamente corretos e efetivos para os objetivos do cliente.

Deployment: a última etapa da metodologia CRISP-DM se refere à implementação do modelo final definido na fase de avaliação. É nessa fase que podem surgir novos insights para aprimorar o produto final. Além disso, é quando uma revisão do projeto é conduzida a fim de atestar que os objetivos foram alcançados ao final do processo.

3.2. Ferramentas

Durante a construção do modelo preditivo, foram usadas as seguintes ferramentas:

Google Colaboratory: produto do Google Research que é um serviço de nuvem gratuito que permite a escrita e execução de código pelo próprio navegador, com ênfase em machine learning, análise de dados e educação. É uma ferramenta que possibilita a agregação de código fonte e textos descritivos, resultando na criação dos notebooks (cadernos).

Google Drive: repositório do grupo com o intuito de armazenar a documentação e demais arquivos que foram necessários, como as dinâmicas feitas durante os encontros de instrução, em que tivemos que redigir textos em conjunto.

Google Docs: SaaS (Software as a service) que oferece a criação e edição de textos sem a instalação de programas no computador e facilitando o compartilhamento e edição simultânea de arquivos que foi utilizado pelo Banco Pan, por exemplo, durante a elaboração da documentação.

Google Sheets: Recebemos o banco de dados do Banco Pan pelo Google Sheets.

GitHub: plataforma de armazenamento dos repositórios, englobando do código fonte à documentação.

Jira Software: ferramenta utilizada para o monitoramento de atividades e acompanhamento do projeto.

Miro: lousa interativa de colaboração visual a qual empregamos na construção das personas.

Canva: plataforma de design gráfico empregada na criação dos mapas das jornadas do usuário e em alguns dos gráficos.

Link List: sistema que permite a agregação de diversos links em uma só página, facilitando o acesso de todos os membros do grupo às ferramentas utilizadas e às páginas de desenvolvimento, como o Google Colab.

Fun Retrospectives: software utilizado ao final de cada sprint para otimizar a dinâmica e engajamento da retrospectiva, etapa do método Scrum em que o grupo se reúne para discutir sobre o processo de desenvolvimento do modelo que está sendo construído, pautando-se majoritariamente na análise do que deu certo, do que não deu certo e do que precisa ser corrigido na próxima sprint melhorando continuamente a produtividade da equipe.

3.3. Principais técnicas empregadas

O desenvolvimento do projeto, realizado em Jupyter Notebook, contou com uma série de ferramentas aplicadas ao Google Colaboratory. Foram utilizados os pacotes Pandas e Numpy, que são de uso fundamental em Python e foram utilizados na análise exploratória dos dados, bem como em sua posterior limpeza e tratamento. A biblioteca SciKit-Learn também foi utilizada, ante a sua aplicabilidade em Machine Learning. As ferramentas que o compõem são essenciais para o desenvolvimento de um modelo preditivo.

Modelos utilizados:

- KNN (K-Nearest Neighbors)
- Gaussian Naive Bayes
- Gradient Boosting
- Decision Tree Classifier
- Random Forest
- Support-Vector Machine
- Regressão Logística

A utilização de vários modelos trouxe um maior tipo de amostragem para escolhermos o melhor deles, aumentando a chance de obtermos um bom modelo.

Ajustes de hiperparâmetros usados:

- Random Search
- Grid Search

Os benefícios envolvendo os ajustes de hiperparâmetros são os de trazer o melhor de cada modelo, e conseqüentemente, aumentar a acurácia, precisão e revocação de cada um.

Metodologias utilizadas:

- **Agile:** conjunto de técnicas utilizadas para gestão de projetos que oferece mais agilidade, prontidão para possíveis erros, eficiência e flexibilidade.
- **Kanban:** segundo o site “Kanban Guides”, Kanban é uma estratégia para otimizar o fluxo de valor através de um processo que utiliza um sistema visual, baseado em um sistema puxado (pull-based system). Além disso, é uma metodologia ágil.
- **Scrum:** segundo o site “Conta Azul”, scrum é um método de trabalho realizado a partir de pequenos ciclos de atividades dentro de um projeto. Cada ciclo de atividade é planejado previamente e se chama Sprint, composto por um período de tempo predefinido (no grupo “Panco”, este período foi de 15 dias) em que as tarefas devem ser realizadas pela equipe. Após o final de cada sprint, é feita uma validação do protótipo do projeto com o parceiro, onde recebemos feedbacks sobre o produto e perguntamos informações relevantes para o processo. Após, é feita a “Scrum Retrospective”, uma reunião em que é detalhada todos os erros, acertos e interações importantes do grupo durante a sprint, para análise do que pode ser corrigido e, no que deve ser mantido ou retirado em relação ao grupo.

4. Desenvolvimento e Resultados

4.1. Compreensão do Problema

4.1.1. Contexto da indústria

Análises preditivas são de fundamental importância para as instituições financeiras e possuem uma série de aplicações em Business Intelligence. No contexto em que se insere o nosso cliente — o Banco PAN — um modelo preditivo capaz de tornar mais eficiente o processo de atendimento ao cliente seria de grande utilidade para que o relacionamento entre os usuários e o banco fosse mais satisfatório.

Com a maior adesão da inteligência artificial em ferramentas de predição para grandes empresas nos últimos anos, notou-se um ganho de valor significativo entre as corporações. Nesse sentido, compreende-se que um atrativo de novos clientes se refere a um bom atendimento por parte da instituição, o qual pode ser potencializado pelo uso da IA, já consolidada como uma das principais tendências de mercado atualmente. Um exemplo da aplicação dos modelos preditivos para atendimento ao cliente, nesse contexto, são os chatbots utilizados por muitos bancos em plataformas digitais. Destacam-se, nesse cenário, os seguintes players de mercado:

Bank of America: um dos maiores bancos dos Estados Unidos, o Bank of America, lançou em 2018 uma assistente virtual que ajuda os clientes a tomarem decisões. Ela já atendeu a mais de 150 mil chamados desde seu lançamento, referentes a sugestões de investimentos, pagamento de contas e emissão de notificações.

Bradesco: constitui um dos cases mais conhecidos de bots que utilizam IA no Brasil. A Bia — assistente virtual do banco — nasceu como uma forma de automatização para o back-office, utilizada apenas por funcionários. Atualmente, atua sobre 91 serviços e produtos do banco.

Banco Original: o chatbot de IA foi criado com o objetivo de ampliar o nível de resolução dos atendimentos, a fim de aumentar a retenção de clientes. A IA atua na resolução de dúvidas e na efetivação de transações bancárias, sendo responsável por cerca de 70% de todos os atendimentos ao cliente. Desde sua implementação, a taxa de retenção de clientes subiu de 60% para 90%.

Royal Bank of Canada: nos últimos anos, vem utilizando um bot de inteligência artificial que aprende com as solicitações e atividades bancárias dos clientes. Conforme ele vai

cruzando os dados de atividades financeiras recentes, torna-se capaz de realizar análises preditivas e antecipar questões e problemas, oferecendo sugestões personalizadas.

AS CINCO FORÇAS DE PORTER

Michael E. Porter, no artigo “How Competitive Forces Shape Strategy”, ressalta a importância de se analisar não somente a estrutura interna do negócio, como também as forças competitivas que o cercam. Considerando-se as circunstâncias sob as quais o Banco PAN se encontra, visualizar a intensidade dessas forças é imprescindível para uma melhor compreensão de sua situação no mercado ante a solução a ser desenvolvida.

Quando se trata da rivalidade entre concorrentes, é fundamental pontuar que os clientes de um banco, naturalmente, terão preferência por manter uma conta bancária onde houver um melhor sistema de atendimento. Assim, considerando-se o potencial dos sistemas das demais instituições do mesmo segmento do Banco PAN, este encontra-se mais suscetível à força competitiva da concorrência.

Existem, ainda, outras forças competitivas que devem ser consideradas ao se realizar uma análise de negócios. Uma delas se refere ao poder de barganha dos fornecedores, que, nesse contexto, não é prioritário, visto que a equipe será responsável por fornecer ao cliente a solução esperada. O poder de barganha dos compradores, por outro lado, constitui uma ameaça significativa, visto que os clientes do banco, em busca de uma boa prestação de serviços – incluindo um atendimento eficiente –, podem pressioná-lo no sentido de considerar o encerramento de suas contas na instituição.

A ameaça de novos entrantes também constitui um fator de preocupação, haja vista a crescente tendência da adoção de Inteligência Artificial por parte das empresas a fim de oferecer uma boa experiência para o cliente. Por fim, a ameaça de produtos ou serviços substitutos, os quais poderiam configurar alternativas ao modelo preditivo que será construído, não possui relevância significativa, visto que a opção mais favorável na indústria, atualmente, contempla o uso de tecnologias como a que será entregue para o Banco PAN.

Espera-se que a solução desenvolvida acompanhe as atuais tendências do mercado, atingindo as expectativas do Banco PAN quanto ao objetivo de potencializar a qualidade do atendimento que propõe. Desse modo, há a possibilidade de que a proposta, além de beneficiar a instituição, a eleve ao patamar de player de mercado, assim como os outros casos mencionados.

4.1.2. Análise SWOT

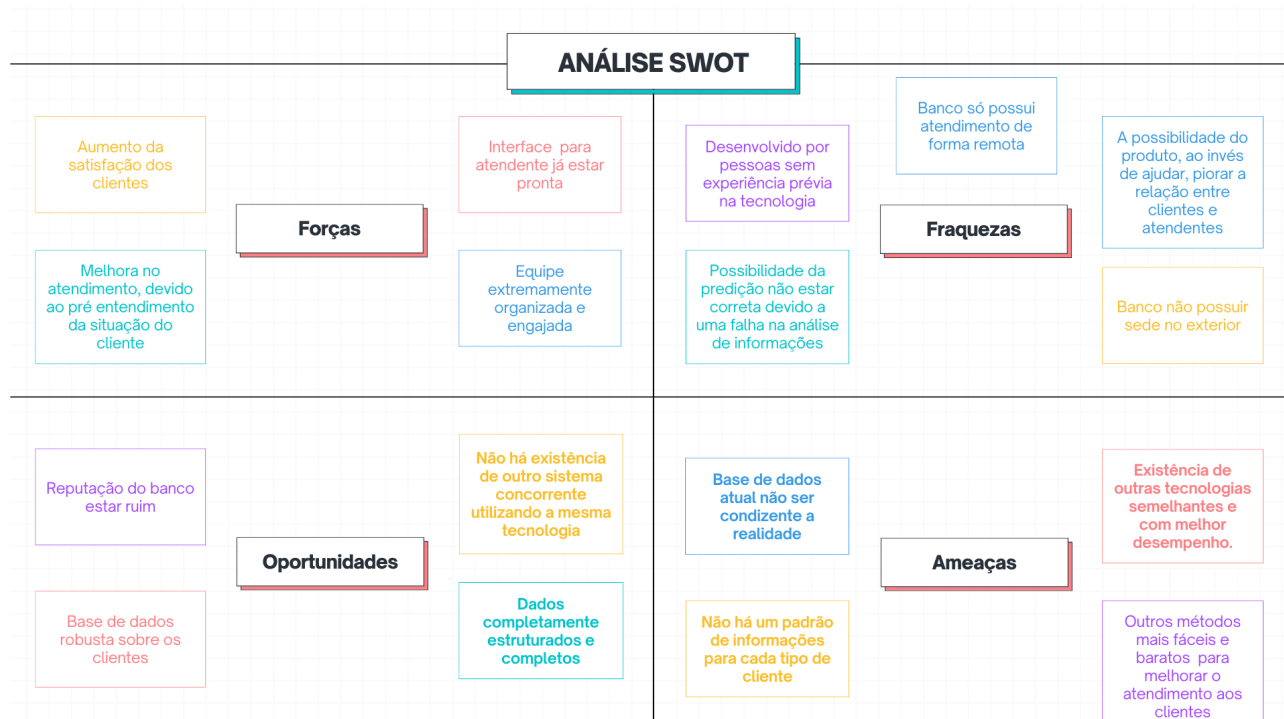


Figura 1: Análise SWOT

4.1.3. Planejamento Geral da Solução

a) Dados disponíveis:

Fonte: dados captados pelo Banco PAN

Conteúdo:

Dados	Descrição
“anomes”:	Mês e ano correspondentes aos valores dos atributos
“vlr_credito”:	Crédito no mercado
“vlr_saldo”:	Quanto o cliente tem de crédito no PAN (diferentemente do valor que possui na conta, este é quanto ele deve)
“num_atend_atrs”:	Número de atendimentos que o cliente tem e estão em atraso (atraso por parte do PAN em não cumprir os prazos)
“vlr_score”:	Score do cliente (de 0 a 1000) no mercado
“num_produtos”:	Quantidade de produtos que o cliente adquiriu no banco.
“num_atend”:	Número de atendimentos, dentro e fora do prazo (é considerado atendimento se gerou um protocolo)
“num_cpf”:	CPF do cliente

“qtd_oper”:	Quantidade de operações de um cliente referentes a um determinado produto
“qtd_reclm”:	Quantidade de reclamações por cliente
“qtd_restr”:	Restritivo de mercado, se tem alguma pendência ou não.
“vlr_renda”:	Valor de renda do mercado. É uma medida preditiva

Rating mensal de risco do cliente, que varia de AA (zero risco) a HH (prejuízo absoluto). É uma avaliação interna do banco, não tendo relação com a situação do cliente no mercado. Tem um viés de risco (entender quem é o cliente, como ele se comporta no mercado e no Banco PAN). É uma variável fundamental para medir o atrito dele com o banco.

b) Solução proposta:

A solução que estamos propondo pretende diminuir as dificuldades enfrentadas pela empresa quando o assunto é o relacionamento com clientes do banco, visto que, atualmente, existem muitas reclamações, em diferentes veículos de avaliação, a respeito de atendimento que não estão direcionados para o real problema do atendido.

c) Tipo de tarefa: (regressão ou classificação):

O tipo de tarefa que iremos fazer será de classificar, isso devido ao fato de que o objetivo da nossa IA será de analisar, levando em consideração o comportamento dos clientes do banco no passado, um cliente, assim, em que esse efetuar um contato com o banco seja por qualquer um dos meios hoje disponíveis por eles e assim com base em suas informações classificá-lo como cliente atritado, cliente que busca novos produtos ou cliente novo, passando tal informação para o atendente que efetuará o atendimento deste cliente, isso com a intenção de oferecer diferentes tipos de atendimento, sendo esse correspondente a situação de cada cliente.

d) Utilização da solução proposta:

A solução proposta deverá ser utilizada no sistema do Banco PAN pelos atendentes do canal de telefone, que irão conseguir visualizar qual categoria os clientes que estão em contato se consideram, sendo elas: cliente atritado, cliente novo, clientes que querem adquirir novos produtos.

e) Benefícios trazidos pela solução proposta:

A problemática trazida pelo parceiro indica a falta de informações sobre o cliente como um grande dificultador na comunicação. Isso ocorre porque, nas situações em que ele se encontra a buscar ajuda devido aos problemas que enfrentou diretamente com o banco, acaba sendo recebido com outras indicações de produtos, situação que afeta negativamente a efetividade do sistema de atendimento dos clientes e agrega reclamações constantes ao banco. Situado o ponto fraco da comunicação entre cliente e empresa, a nossa solução busca prever a personalidade do cliente em relação ao banco, indicando, assim, se busca novos produtos ou se está com problemas e dúvidas sobre funcionalidades da empresa. Logo, os

benefícios da aplicação serão reduzir as reclamações em aberto e melhorar a qualidade de atendimento ao cliente.

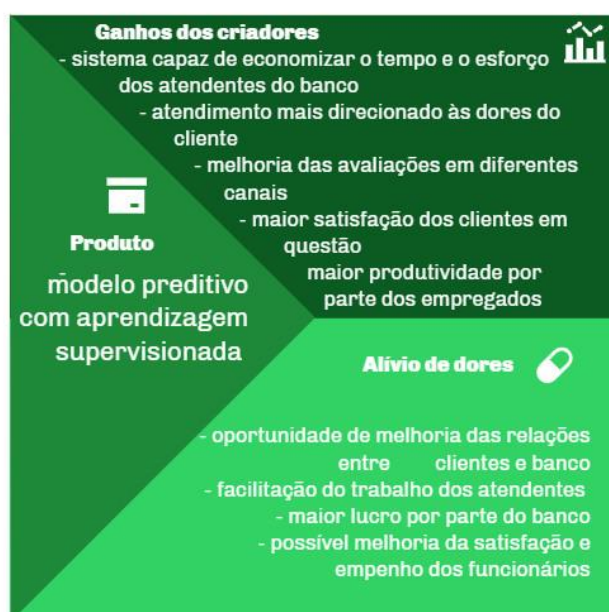
f) Critério de sucesso e a medida utilizada para o avaliar:

Contando que o principal objetivo do banco parceiro seja prever os futuros clientes que podem ter problemas com a empresa de acordo com os dados coletados do mesmo, o principal critério para declarar que o projeto teve êxito é uma boa taxa de acerto em previsões, já que é a medida principal onde podemos avaliar o nosso produto. Dentro das taxas de acerto, será possível analisar também quais os principais aspectos do cliente que alteram os resultados.

4.1.4. Value Proposition Canvas

Canvas da proposta de valor

Proposta de valor



Perfil do cliente



Figura 2: Canvas da proposta de valor

4.1.5. Matriz de Riscos

Matriz de Risco						
Probabilidade	Ameaça			Oportunidade		
Alta	Atraso na entrega dos entregáveis de cada sprints.				Facilitar o trabalho dos atendentes.	
Médio		Não cumprir o propósito de melhorar a relação entre banco e clientela.	Criar um sistema cuja atribuição seja errônea.	Melhorar o padrão de relacionamento do banco PAN com seus clientes.		
Baixa			Vazamento de dados.	O banco PAN colocar a solução desenvolvida em uso.		
	Baixo	Médio	Alta	Alta	Médio	Baixo
	Impacto					

Figura 3: Matriz de Riscos

4.1.6. Personas



NOME: Aparecida Silva de Jesus

IDADE: 62 anos

GÊNERO: Feminino

OCUPAÇÃO: Aposentada

"Sou uma mulher guerreira que faz de tudo pelos filhos". Aparecida é moradora de uma simples casa na região do Brás em São Paulo - SP. Ela é mãe de 3 filhos: Werlinson, José e Lara, que a ajudam financeiramente. Ela escolheu o Banco PAN para receber sua aposentadoria.

Considerações biográficas e comportamentais

Pouco
conhecimento
tecnológico

Gosta de
simplicidade

Não tem
muita
paciência
no dia a dia

Dores/Motivações atuais com o problema:

Escolheu o
PAN por
produtos
como o saúde
pan

Está enfrentando
alguns problemas
para concluir o
processo de
aposentadoria

Objetivos/necessidades específicas em relação ao problema:

Conseguir
ser uma
cliente nova
do banco

Conseguir
colocar a
aposentadoria
no banco

Obter o
Saúde
PAN

Figura 4: Persona de cliente do Banco PAN



NOME: Claudio Santos Bezerra

IDADE: 28 anos

GÊNERO: Masculino

OCUPAÇÃO: Estudante de mestrado/Professor

"Apenas a educação transforma". Claudio é morador de um apartamento na Consolação em São Paulo - SP. Ele está fazendo mestrado em "Estudos Literários" na USP. Ele está com um score muito baixo no SERASA, porém deseja um empréstimo consignado no Banco PAN. Atualmente, ganha dinheiro dando aula de português para uma escola de ensino médio

Considerações biográficas e comportamentais

Deseja investir em um comércio para ajudar na renda

É otimista e esperançoso

Pavio curto

Dores/Motivações atuais com o problema:

Tem insistido para o seu banco disponibilizar o empréstimo, apesar do score baixo.

Liga semanalmente para o SAC do banco PAN para fazer reclamações

Tem tido problemas para pagar o seu financiamento de veículos feito pelo PAN

Objetivos/necessidades específicas em relação ao problema:

Conseguir o empréstimo para o seu comércio

Resolver sua dívida com o financiamento

Figura 5: Persona de cliente do Banco PAN



NOME: Lorryne Stephane Soares da Silva

IDADE: 18

GÊNERO: Feminino

OCUPAÇÃO: Estudante/Atendente

"Meu sonho é ajudar o máximo de pessoas possível". Lorryne é estudante bolsista integral de enfermagem no Albert Einstein, para ajudar a mãe nos custos de casa, ela trabalha como atendente no SAC banco PAN no período vespertino.

Considerações biográficas e comportamentais

É altruísta e gosta de fazer trabalho voluntário

Sempre foi bastante dedicada nos estudos

Dores/Motivações atuais com o problema:

Tem dificuldade com tecnologia, principalmente em procurar dados do cliente

Por ser muito sensível, tem dificuldade em lidar com clientes atirados

As vezes se confunde em quando pode oferecer um produto pro cliente ou não

Objetivos/necessidades específicas em relação ao problema:

Se preparar psicologicamente para um possível cliente atirado

Ter informação sobre seu cliente de forma fácil e rápida

Figura 6: Persona de atendente do Banco PAN

4.1.7. Jornadas do Usuário

USER JOURNEY MAP



PERSONA: LORRAYNE STEPHANE SOARES DA SILVA

CENÁRIO:

A estudante ingressou recentemente no mercado de trabalho, atuando como atendente do Banco PAN. Contudo, ela fica muito desmotivada quando não consegue guiar o cliente a uma solução para sua solicitação.

EXPECTATIVAS:

Quer se preparar para atender um cliente atritado, tendo informação rápida sobre o cliente e saber para qual cliente ela deve oferecer novos produtos.

FASE 1	FASE 2	FASE 3
Lorryne é atendente em meio período e tem dificuldades para gerar um bom atendimento ao cliente que se irrita quando ela informa que a reclamação dele não consegue ser solucionada pelo setor dele e que ela repassará a ligação para o setor certo. O cliente reclama por ter que relatar o ocorrido novamente e a Lorryne, que apenas está fazendo o seu trabalho, se desgasta.	O Banco Pan implementa a A.I. Panco, em que o modelo prediz se trata-se de um cliente atritado, de um novo cliente ou de um cliente que deseja adquirir um novo produto. Logo, a atendente já atenderá a ligação sabendo qual o possível perfil do cliente, lidando melhor com clientes atritados e oferecendo novos produtos para os clientes certos.	Uma vez tendo oferecido um bom atendimento voltado às necessidades do cliente, o atendente ficará menos estressado e tende a ficar mais satisfeito com o seu trabalho, reduzindo o índice de demissões.

OPORTUNIDADES

- Aprimorar o atendimento ao cliente
- Adaptação a uma nova tecnologia
- Fortalecimento da imagem do banco
- Automação de parte do trabalho do atendente

RESPONSABILIDADES INTERNAS

Time de atendimento ao cliente:

Capacitação quanto ao uso da Inteligência Artificial implementada

Time de marketing:

Divulgação das medidas tomadas pelo banco para aprimorar o atendimento ao cliente, mostrando sua adaptação a novas tecnologias

Figura 7: User Journey Map de atendente do Banco PAN

USER JOURNEY MAP



PERSONA: CLÁUDIO SANTOS BEZERRA

CENÁRIO:

Devido ao seu baixo score no SERASA, viu o Banco PAN como única possibilidade para o empréstimo consignado que queria fazer. Contudo, enfrenta dificuldades para concretizar esse objetivo, necessitando corriqueiramente do suporte ao cliente da instituição.

EXPECTATIVAS

Conseguir um empréstimo para construir o seu empréstimo e deseja resolver a sua dívida com um financiamento.

FASE 1	FASE 2	FASE 3
Cláudio apresenta insatisfação com o serviço e decide contatar a central de atendimento do Banco PAN. É atendido, conta a história, mas é informado, no entanto que será repassado ao setor correspondente à sua reclamação, em que terá que contar, mais uma vez, o seu problema.	O Banco Pan implementa a I.A, em que o modelo prediz se trata-se de um cliente atritado, de um novo cliente ou de um cliente que deseja adquirir um novo produto. Logo, o cliente em ligação já é conectado para ser atendido por alguém do setor qualificado a solucionar sua demanda, sem ter que ser repassado diversas vezes.	Uma vez que recebe um bom atendimento voltado às suas necessidades, o cliente ficará satisfeito, dará um bom feedback pós atendimento, auxiliando a aumentar o posição do banco nos rankings avaliativos e estará mais propenso a manter sua conta ativa, a adquirir novos produtos e a indicar o Banco PAN aos seus conhecidos.

OPORTUNIDADES

- Aprimorar o atendimento ao cliente
- Adaptação a uma nova tecnologia
- Fortalecimento da imagem do banco
- Automatização de parte do trabalho do atendente

RESPONSABILIDADES INTERNAS

Time de atendimento ao cliente:

Capacitação quanto ao uso da Inteligência Artificial implementada

Time de marketing:

Divulgação das medidas tomadas pelo banco para aprimorar o atendimento ao cliente, mostrando sua adaptação a novas tecnologias

Figura 8: User Journey Map de cliente do Banco PAN

4.2. Compreensão dos Dados

a) Item cujo preenchimento é requerido caso haja mais de um conjunto de dados, descrevendo como serão agregados /mesclados. No entanto, não se aplica, visto que o conjunto de dados é único. A base de dados fornecida pelo Banco PAN consiste em uma tabela que une todos os campos que serão utilizados para análises.

b) Sobre a descrição dos riscos e contingências relacionados aos dados (qualidade, cobertura/diversidade e acesso), a manipulação dos dados que serão tratados para a construção do modelo preditivo é bastante delicada. Visando extrair da melhor forma possível as informações necessárias, prezando pelo cuidado e segurança do conjunto de materiais que foi confiado à equipe, é fundamental supervisionar os riscos e estabelecer contingências que procurem mitigar os seus impactos.

Os maiores riscos estão associados ao vazamento de dados e à insuficiência de dados para a construção de um modelo preditivo adequado. Todavia, cabe ressaltar que os dados foram fornecidos pelo cliente de forma anonimizada através do uso da tecnologia de criptografia hash para que, caso haja vazamento de tais dados, os prejuízos sejam mínimos.

A linha do tempo dos dados em que o modelo irá se basear compreende o período de tempo do mês de abril do ano 2021 até abril do ano seguinte, sendo os clientes identificados pelo campo CPF, que está mascarado e representa o número do documento do cliente.

Todos os riscos podem ter consequências, logo são importantes e devem ser gerenciados. Por isso, visando minimizar o risco de exposição de informações privadas e sigilosas dos usuários do Banco PAN, cabe ao grupo manipular os dados de maneira cautelosa, contemplando a restrição solicitada de não incluir a base de dados no repositório do GitHub, devendo esta ser administrada de forma independente.

A descrição de como será selecionado o subconjunto para análises iniciais, não se aplica, visto que não haverá subconjunto para análises iniciais dos dados.

No que se refere à descrição das restrições de segurança, apesar dos dados já terem sido disponibilizados após uma criptografia daqueles que são sensíveis, a base de dados não deve ser incluída no repositório do GitHub ou em qualquer via de acesso público, sendo gerenciada de maneira independente.

A abordagem do modelo preditivo a ser desenvolvido prioriza as predições relacionadas aos clientes atritados com o banco. Desse modo, para uma melhor visualização dos dados que embasariam esse objetivo, foram definidos os campos de maior relevância da base de dados utilizada pela equipe: quantidade de reclamações, número de atendimentos atrasados, número total de atendimentos, valor do Score, quantidade de restrições, rating, valor de crédito e valor de renda. Com base nesse conjunto de informações de cada cliente, será possível desenvolver um modelo capaz de indicar se existem atritos em sua relação com a instituição financeira. Alguns desses campos serão analisados em conjunto. São eles: rating e score, e valor de renda e valor de crédito, de modo que seja possível obter informações mais detalhadas através do cruzamento de dados.

Abaixo, encontram-se as visualizações gráficas realizadas a partir dos cálculos estatísticos dos campos estudados.

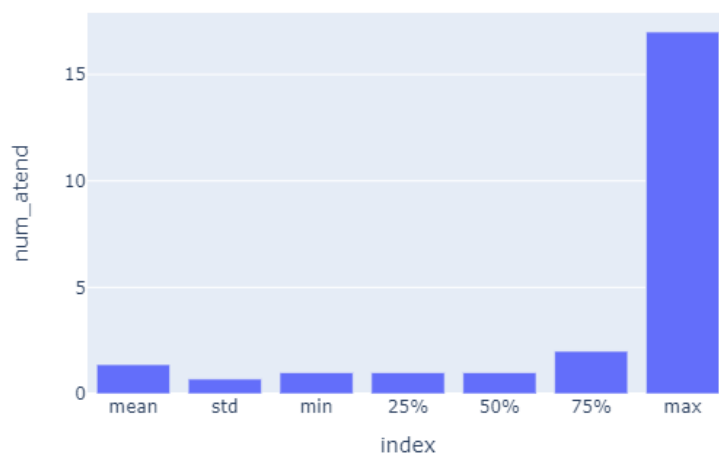


Figura 9: Número de atendimentos por cliente (média, desvio-padrão, mínimo, máximo e quantis)

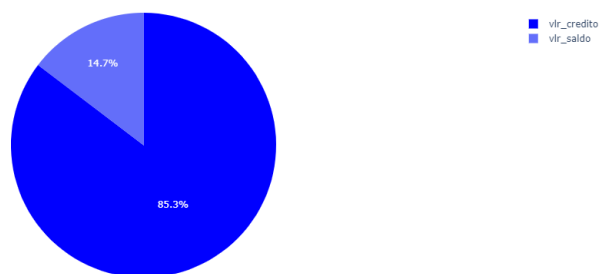


Figura 10: Comparação entre o valor de crédito e o valor de saldo)

Comparação entre o número de atendimentos e atendimentos atrasados:

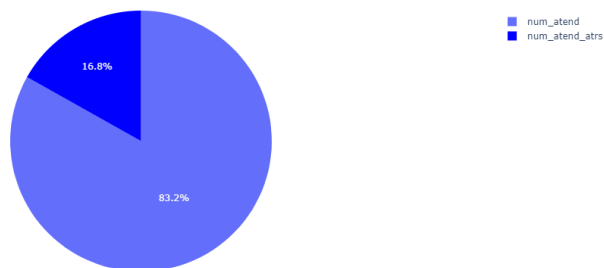


Figura 11: Comparação entre o número de atendimentos e o número de atendimentos atrasados

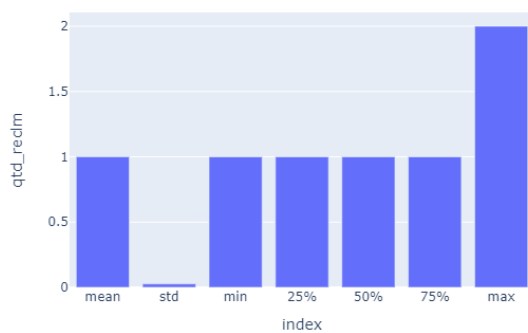


Figura 12: Quantidade de reclamações (média, desvio-padrão, mínimo, máximo e quantis)

A coluna que será usada como “target” é de natureza binária, sendo ela:

ind_atritado	Indicador de cliente atritado, 1 para Sim e 0 para Não
--------------	--

4.3. Preparação dos Dados

O primeiro passo no processo de preparação dos dados, foi a retirada dos indivíduos - representados pelas linhas - sem cod_rating, para facilitar o processo de aprendizagem do modelo preditivo. Tal escolha foi feita pelo fato de que o cod_rating é uma coluna do banco de dados que corresponde a um parâmetro de quantos produtos cada indivíduo possui no banco, logo, quem tem o cod_rating nulo não trata-se de um cliente do banco.

Após este passo inicial todos os campos nulos das colunas "num_atend", "num_atend_atrs", "qtd_reclm", "qtd_restr", "ind_atritado", "ind_engajado", "ind_novo_cliente", "vlr_score", "num_produtos" foram alterados para zero (0), o que deve-se ao fato de serem valores complexos de serem identificados durante o processo de aprendizagem do modelo preditivo.

E haja vista que o modelo preditivo não é capaz de ler strings, o cod_rating dos clientes foi convertido em números ordenados, conforme demonstrado abaixo:

A	AA	B	C	D	E	F	G	H	HH
0	1	2	3	4	5	6	7	8	9

Foram, também, retirados indivíduos - linhas - cujo valor de crédito e de saldo constam como nulos, pelo fato de estes valores estarem zerados serem um forte indicador de que estes indivíduos não possuem conta no Banco PAN.

Em seguida, foi verificado pelo grupo Panco que há uma expressiva quantia de indivíduos que apesar de possuírem informações bem concretas, não possuem sua renda registrada, por conseguinte, como essa coluna não é relevante no processo de aprendizagem do modelo preditivo, removemos esta coluna da tabela. O CPF hashado, por sua vez, estava sendo lido como string, o que era um obstáculo no processo de aprendizagem do modelo preditivo, também o removendo. Estes citados acima, foram os motivos que nos induziram a remover as colunas “vlr_renda” e “num_cpf_hash” da tabela utilizada.

Com o tratamento dos nulos finalizados, é possível comprovar que a quantidade de dados nulos na tabela é, agora, zero.

Com o intuito de que a base de dados não possua mais de um CPF por linha, foi decidido entre o grupo de que, inicialmente, faremos uso de uma safra aleatória. Escolhemos deste modo

para minimizar a chance de que haja um viés da escolha da safra, podendo ter uma quantidade variável de clientes atritados em cada período.

A agregação de registros se aplica no projeto em questão, contudo, por ora, ainda não foi realizada. Nesse momento, foi escolhida uma safra aleatória – a fim de evitar vieses durante o processo – que é parâmetro para a seleção de CPFs únicos. Para fazer isso, escolhemos um número aleatório que vai de um a doze (valores esses que representam os meses das safras disponibilizadas pelo banco) e o utilizamos para determinar qual safra vai ser escolhida. Logo após essa escolha, é feito um script que seleciona apenas as doze safras existentes e, a partir do número aleatório gerado anteriormente, escolhe uma dentre elas. Com a safra determinada, aplicamos uma máscara ao data frame previamente tratado que contém todos os dados dos clientes do banco a fim de obter apenas aqueles pertencentes à safra descrita.

Os valores ausentes, definidos na base de dados do cliente como “Not a Number” (NaN), indicam que não existe registro do atributo para um determinado cliente. Esses valores foram substituídos por zero para os seguintes campos:

- num_atend (número total de atendimentos)
- num_atend_atrs (número de atendimentos atrasados)
- qtd_reclm (quantidade de reclamações)
- qtd_restr (quantidade de restritivos no mercado)
- ind_atritado (índice de atrito do cliente)
- ind_engajado (índice de engajamento do cliente)
- ind_novo_cliente (índice que identifica um novo cliente)
- vlr_score (valor do score no Serasa)
- num_produtos (número de produtos adquiridos)

Posteriormente a esse processo de limpeza de dados ausentes/nulos, foram definidas as features a serem utilizadas na construção da lógica do modelo preditivo. Para essa definição, foram considerados os atributos da base de dados do cliente que possuem relevância ao se analisar o comportamento de um cliente que utiliza os serviços oferecidos pelo Banco PAN.

Safra

Esse atributo é fundamental para determinar as previsões relacionadas ao perfil do cliente que está entrando em contato com o banco. Ao analisar a safra, é possível observar o comportamento do cliente ao longo do tempo e estimar seu comportamento futuro enquanto cliente do banco.

Coluna correspondente: “anomes” (ano e mês)

Valor de crédito

A coluna em questão permite avaliar a situação do cliente no mercado ao declarar seu valor total de crédito. Portanto, foi selecionada enquanto determinante para as predições relacionadas aos possíveis atritos entre cliente e banco, bem como indicações de novos clientes ou clientes que possam ter a intenção de adquirir novos produtos.

Coluna correspondente: "vlr_credito" (valor de crédito no mercado)

Valor de saldo

Indica o saldo total do cliente no banco, o que pode indicar se existe atrito na relação que será predita pela Inteligência Artificial.

Coluna correspondente: "vlr_saldo" (valor de saldo)

Número de atendimentos

A quantidade de atendimentos configura um dos principais indicativos de atrito de um cliente com o Banco PAN.

Colunas correspondentes: "num_atend" (número total de atendimentos) e "num_atend_atrs" (número de atendimentos atrasados)

Valor do score

O valor do score no Serasa de um determinado cliente é capaz de identificar a necessidade do cliente de adquirir, por exemplo, certo valor de crédito no banco. Além disso, é possível, através desta métrica, testar hipóteses relacionadas ao nível de atrito entre o cliente e o banco, sendo o score inversamente proporcional ao índice de atrito.

Coluna correspondente: "vlr_score" (valor do score no Serasa)

Número de produtos

A quantidade de produtos adquirida por um cliente no Banco PAN configura um atributo capaz de determinar a intensidade da relação entre o cliente e o banco, visto que, quanto maior o número de serviços aderidos, menor pode ser o nível de atrito com a instituição, considerando a preferência pelo Banco PAN quanto a esses serviços.

Coluna correspondente: "num_produtos" (número de produtos adquiridos)

Quantidade de operações

O campo em questão também é determinante para que se saiba a intensidade da relação cliente-banco, haja vista que, se um cliente possui muitas operações referentes a um mesmo serviço, deve ser considerada a preferência pelo Banco PAN a outras instituições financeiras.

Coluna correspondente: "qtd_oper" (quantidade de operações referentes a um serviço do banco)

Quantidade de reclamações

Esse atributo define um dos principais atributos capazes de determinar se um cliente é atritado ou não com o banco, visto que, se um cliente possui muitas reclamações, ele certamente se sente insatisfeito com os serviços oferecidos.

Coluna correspondente: "qtd_reclm" (quantidade de reclamações do cliente)

Quantidade de restritivos no mercado

O campo é determinante para que a inteligência artificial a ser desenvolvida crie previsões precisas relacionadas ao índice de atrito do cliente com o banco, considerando suas restrições quanto a outras instituições financeiras.

Coluna correspondente: "qtd_restr" (quantidade de restritivos no mercado)

Rating do cliente

A métrica se refere ao risco que o cliente representa para o banco, e analisar essa classificação é de fundamental importância para que a IA possa prever se existem atritos nessa relação ou se é provável que o cliente adquira novos produtos na instituição.

Coluna correspondente: "cod_rating" (rating mensal de risco do cliente)

Métrica utilizada para agregação de registros: Moda

Índice de atrito

A métrica interna do banco avalia possíveis conflitos entre o Banco PAN e seus clientes. Esse índice será fundamental para que a IA possa classificar a intenção do cliente que busca atendimento da instituição, visto que, se esse índice for alto, pode indicar possíveis reclamações que virão do cliente.

Coluna correspondente: "ind_atrito" (índice de atrito do cliente)

Índice de engajamento

É importante analisar esse campo para que a inteligência artificial possa prever o quanto engajado o cliente é com a instituição. Se o índice em questão for alto, é alta a probabilidade de que o cliente entre em contato com o banco buscando adquirir novos produtos ou serviços.

Coluna correspondente: "ind_engaj" (índice de engajamento do cliente)

Índice de identificação de novo cliente

Esse índice identifica potenciais clientes do Banco PAN. Dessa maneira, serão devidamente identificados ao entrarem em contato com o atendimento da instituição, de modo que a intenção de abrir uma conta no banco seja predita pela IA.

Coluna correspondente: "ind_novo_cliente" (índice que classifica um novo cliente)

Por fim, foram definidos os campos que não serão utilizados na construção da lógica do modelo preditivo, os quais encontram-se dispostos abaixo.

Número do CPF

Não será utilizado porque os clientes não serão avaliados um a um. O modelo preditivo irá considerar um conjunto de clientes. Além disso, deve ser mantida a anonimização dos clientes, visto que seus comportamentos individuais não podem ser expostos no presente projeto. Por fim, destaca-se que os valores contidos nesse campo são *strings*, ou seja, não são aceitos pelo modelo.

Campo correspondente: "num_cpf"

Valor da renda

Não será utilizado porque a maioria dos valores para esse campo são nulos (NaN), ou seja, não existe uma quantidade suficiente de registros para que o atributo em questão exerça influência, adequadamente, sobre as predições realizadas.

Campo correspondente: "vlr_renda"

NORMALIZAÇÃO E PADRONIZAÇÃO

Após a definição das features, analisamos que havia muito mais registros de clientes não-atritados do que atritados. Por esse motivo, fizemos um processo de balanceamento nos dados, a fim de torná-los proporcionais e evitar vieses. Neste processo, multiplicamos a quantidade de linhas de clientes atritados por 1.4 e, posteriormente, concatenamos essa quantidade com o número ideal de linhas de clientes não conflitados, criando um novo dataframe de 60% de clientes não-atritados e 40% de conflitados com o banco.

Uma vez realizado esse processo, foi necessário realizar a normalização dos dados. Para isso, foram removidos os outliers do dataframe, ou seja, os valores que mais destoam dos outros registros e que poderiam enviesar resultados de importantes medidas estatísticas. Assim, para as features utilizadas, não foram considerados os 2,5% maiores valores, tampouco os 2,5% menores valores.

4.4. Modelagem

A fim de compor um modelo eficiente e funcional, realizamos testes de modelagem com 7 diferentes modelos algorítmicos, sendo eles: o modelo KNN (K-Nearest Neighbor), o SVM (Support-Vector Machines), o Random Forest, o Regressão Logística, o Gaussian NB, o Decision Tree e o Gradient Boosting.

Em todos eles, foram utilizadas duas técnicas de ajuste de hiperparâmetros, isto é, o processo de localizar a configuração de hiperparâmetros que resulta no melhor desempenho de um modelo. Para fins de esclarecimento, hiperparâmetro é uma variável cujo valor é usado para controlar o processo de aprendizado de máquina.

Entre estes algoritmos está o 'Grid Search', que faz uma pesquisa dentro do espaço de valores especificados por nós e testa cada uma das combinações entre eles. Já o método 'Random Search' faz uma quantidade específica de combinações destes valores, de acordo com o número estabelecido por nós e, durante as iterações, como é testado uma quantidade limite, diferente do 'Grid Search' que testa todos, este método busca e escolhe os valores de forma randômica.

Posteriormente à confecção de todos esses modelos, foram analisados os resultados apresentados por cada um deles. Para melhor visualizar a eficiência dos modelos testados, foram selecionadas as métricas: precisão, revocação e acurácia. Além disso, foi produzida uma matriz de confusão para todos os casos, de modo que fosse indicada a proporção entre os falsos positivos e os falsos negativos apresentados pelos modelos preditivos.

- **KNN (K-Nearest Neighbor):**

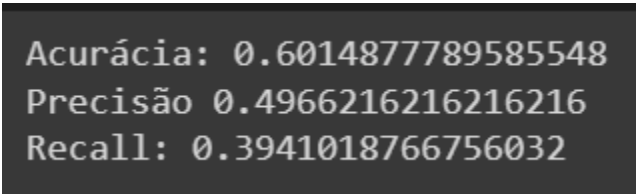
Segundo o autor Ramesh Sharda, o modelo KNN é conhecido como uma "avaliação preguiçosa", visto que não necessita, tampouco realiza, um trabalho prévio de indução de um modelo. Basicamente, esse método separa um grupo de testes e outro de estoque, com o qual o grupo de testes é comparado. Nesse processo, os elementos do grupo de testes são classificados a partir da sua semelhança com os elementos presentes do grupo de estoque.

Após o ajuste de hiperparâmetros, verificamos que a melhor combinação foi tanto a do Grid Search quanto a do Random Search, com um score de 0.62, sendo eles:

Grid Search: {'algorithm': 'auto', 'weights': 'uniform', 'leaf_size': 10, 'metric': 'manhattan', 'n_neighbors': 5,}

Random Search: {'algorithm': 'kd_tree', 'weights': 'uniform', 'leaf_size': 20, 'metric': 'manhattan', 'n_neighbors': 5,}

Durante a realização dos experimentos, foi percebido que o modelo apresenta uma acurácia de 0.6, demonstrando-se mediano para a predição de clientes atritados. Por terem sido encontrados, posteriormente, resultados mais satisfatórios, esse modelo foi descartado para utilização.



Acurácia: 0.6014877789585548
Precisão 0.4966216216216216
Recall: 0.3941018766756032

Figura 13: Acurácia, precisão e recall do modelo KNN

Abaixo, é possível visualizar a matriz de confusão gerada por esse modelo:

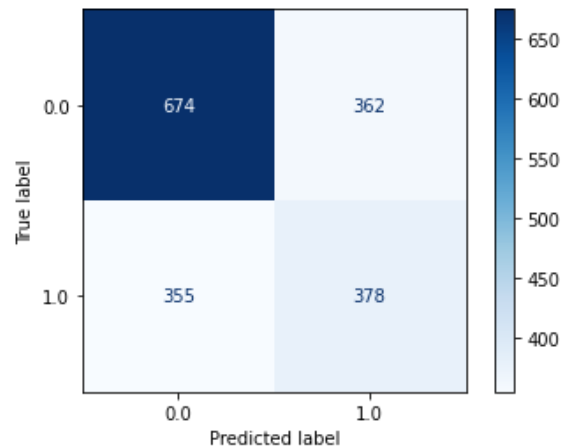


Figura 14: Matriz de confusão do modelo KNN

A partir dela, consegue-se perceber que o modelo prediz que 362 pessoas que não têm atrito, na verdade, o possuem e 355 pessoas que possuem atrito, na verdade não o têm.

- **SVM (Support-Vector Machines):**

De acordo com o livro “Machine Learning na prática: algoritmos em python”, escrito por Fernando Anselmo, Support-Vector Machine pode ser entendido como um algoritmo supervisionado capaz de classificar, regredir e encontrar outliers. O principal objetivo do SVM é encontrar um hiperplano ótimo que é capaz de separar, de modo linear, os pontos de dados em dois componentes, a fim de maximizar a margem.

Nos experimentos realizados, ótimas métricas de acurácia, precisão e revocação foram encontradas, como é possível perceber na imagem a seguir.

Acurácia: 0.9759717314487633
 Precisão: 0.9910873440285205
 Recall: 0.9504273504273504

Figura 15: Acurácia, precisão e recall do modelo SVM

Contudo, ao ser observada, a matriz de confusão demonstra que o modelo não possui bons resultados, tendo em vista que erra muito mais do que acerta: 769 pessoas foram classificadas como não atritadas quando, na verdade, possuíam atrito com o banco e 112 pessoas foram classificadas como atritadas quando não eram. Por isso, esse modelo foi descartado no processo de escolha.

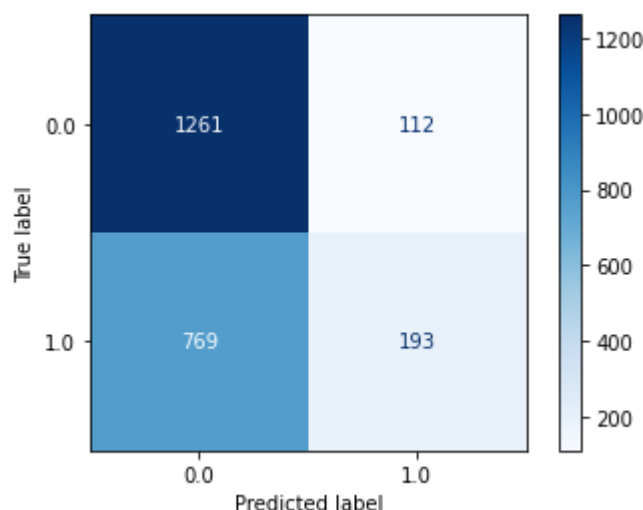


Figura 16: Matriz de confusão do modelo SVM

- **Random Forest:**

Random forest, de acordo com Tony Yiu, criador do artigo “Understanding Random Forest”, consiste em um grande número de árvores de decisão individuais que operam como um conjunto, gerando, assim, uma floresta. Cada uma das árvores presentes na “floresta” aleatória possui um palpite a respeito da previsão da classe e aquela que possui mais “votos”, ou seja, a que aparece de forma mais recorrente nos palpites, é escolhida como a previsão do modelo.

“Um grande número de modelos relativamente não correlacionados (árvores) operando como um comitê superará qualquer um dos modelos constituintes individuais.”

Esse modelo é capaz de permitir que as árvores em conjunto consigam superar cada um dos seus erros individuais, atribuindo maior eficácia ao modelo.

Após o ajuste de hiperparâmetros, verificamos que a melhor combinação do Grid Search quanto a do Random Search, dão o mesmo score de 0.98, sendo eles:

Grid Search: {'n_estimators': 10, 'max_features': 'auto', 'max_depth': 7, 'criterion': 'gini'}

Random Search: {'n_estimators': 10, 'max_features': 'auto', 'max_depth': 9, 'criterion': 'entropy'}

Porém, apesar do mesmo score, escolhemos os hiperparâmetros do 'Random Search' por, após ter testado em nosso modelo, este ter obtido maiores métricas em relação ao 'Grid Search'.

No teste final, esse modelo atingiu ótimas métricas, como é possível visualizar abaixo:

```
Acurácia: 0.9925611052072264
Precisão 1.0
Recall: 0.9812332439678284
```

Figura 17: Acurácia, precisão e recall do modelo Random Forest

Além das métricas exemplares, esse modelo apresentou uma matriz de confusão muito boa, com apenas 43 erros:

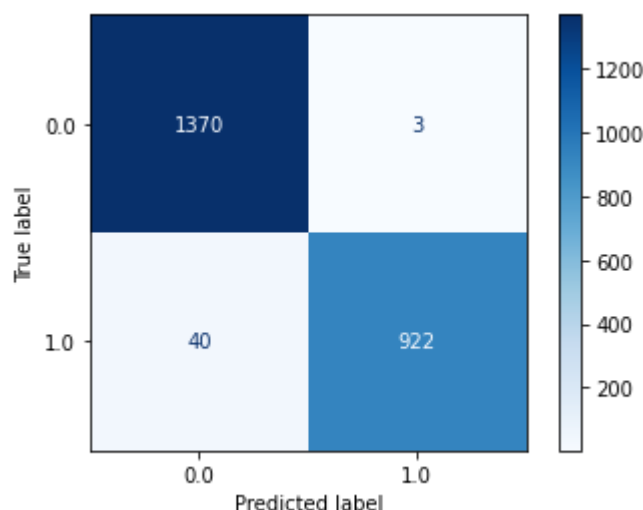


Figura 18: Matriz de confusão do modelo Random Forest

Devido a isso, por ora, esse modelo permanece como uma opção viável para a escolha do modelo de precisão mais adequado, no experimento em questão.

- **Regressão logística:**

A regressão logística é um dos algoritmos de classificação mais conhecidos e mais utilizados, principalmente quando não é possível fazer o uso da regressão linear. Esse método não leva em consideração outliers que não fornecem novas informações ao modelo.

Após o ajuste de hiperparâmetros, verificamos que a melhor combinação do Grid Search quanto a do Random Search, dão o mesmo score de 0.7, sendo eles:

Grid Search: {'solver': 'newton-cg', 'penalty': 'none', 'multi_class': 'multinomial', 'fit_intercept': True, 'dual': False}

Random Search: {'solver': 'newton-cg', 'penalty': 'none', 'multi_class': 'multinomial', 'fit_intercept': True, 'dual': False}

Apesar de sua simplicidade a aparente eficácia para a classificação de modelos, o algoritmo de regressão logística apresentou métricas desfavoráveis para os dados em questão, tendo uma acurácia de 0.69.

Acurácia: 0.6907545164718385
 Precisão 0.9880952380952381
 Recall: 0.2225201072386059

Figura 19: Acurácia, precisão e recall do modelo Regressão Logística.

Corroborando as métricas acima apresentadas, a matriz de confusão desse modelo demonstra um número elevado de erros, tanto de precisão, quanto de revocação. Além disso, esse modelo acerta muito pouco quando é preciso classificar corretamente um cliente atritado, sendo esse o alvo da predição em questão.

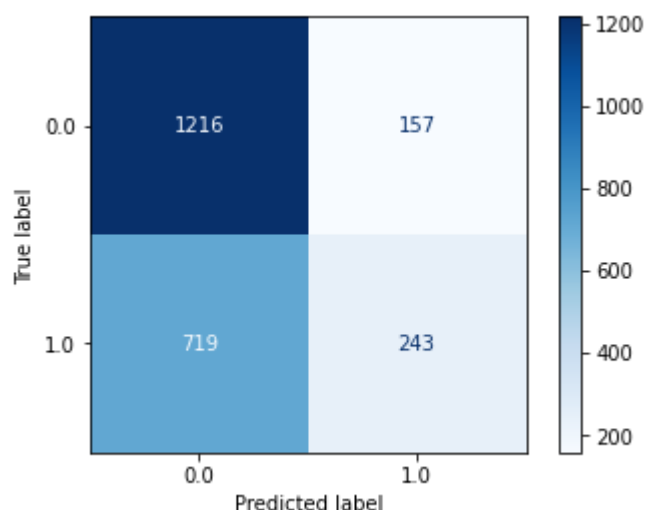


Figura 20: Matriz de confusão do modelo Regressão Logística

Levando em consideração esses resultados, o modelo de regressão logística não será considerado para a realização da predição de clientes atritados.

- **Gaussian NB:**

O modelo Gaussian Naive Bayes é a extensão do teorema de Bayes, sendo usado para muitas funções de classificação. A regra de Bayes fornece uma fórmula para a probabilidade de ocorrência do evento Y dada a condição X. Quando possuímos recursos independentes, essa regra é estendida ao modelo Naive Bayes.

Após o ajuste de hiperparâmetros, verificamos que a melhor combinação do Grid Search é superior a do Random Search, sendo o primeiro com score de 0.88 e o segundo de 0.85. Sendo eles:

Grid Search: {'var_smoothing': 1e-09}

Random Search: {'var_smoothing': 1.873817422860387e-09}

Para os dados em questão, o modelo Gaussian NB apresentou métricas pouco satisfatórias, apesar de medianas. Isso ocorreu porque, em outros modelos algorítmicos, foi possível obter métricas melhores.

```
Acurácia: 0.8862911795961743
Precisão 0.9925925925925926
Recall: 0.7184986595174263
```

Figura 21: Acurácia, precisão e recall do modelo Gaussian NB

Além disso, a matriz de confusão foi capaz de demonstrar o alto número de erros presentes na classificação de clientes. O principal tipo de erro encontrado foi o de revocação, demonstrando que o modelo classifica muito mais clientes atritados como não atritados do que o contrário.

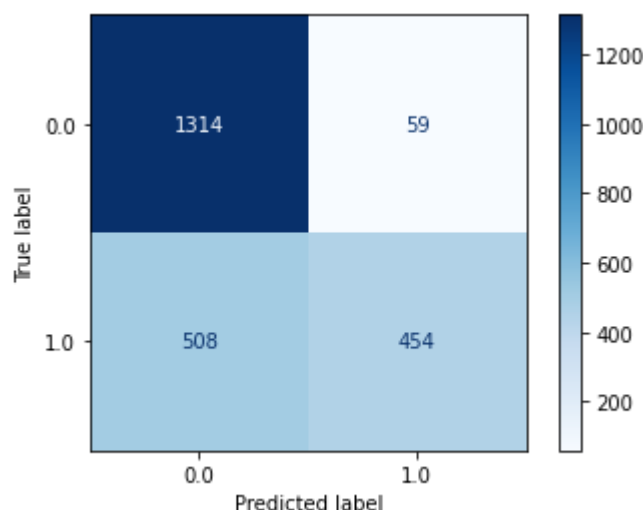


Figura 22: Matriz de confusão do modelo Gaussian NB

Diante desses resultados, foi decidido que esse modelo, assim como outros anteriores, seria descartado das opções para a escolha do modelo de precisão.

- **Decision Tree Classifier:**

O modelo de árvore de decisão, de acordo com a definição da International Business Machine, possui uma estrutura hierárquica em árvore, que consiste em um nó raiz, ramos, nós internos e nós folhas. O seu aprendizado emprega uma estratégia de dividir e conquistar, conduzindo uma grande busca para identificar os pontos de divisão ideais dentro da árvore. Esse processo é repetido de maneira constante, de cima para baixo, até que a maioria dos registros tenham sido classificados especificamente.

Após o ajuste de hiperparâmetros, verificamos que a melhor combinação do Grid Search quanto a do Random Search, dão o mesmo score de 0.98, sendo eles:

Grid Search: {'criterion': 'gini', 'max_depth': 3, 'max_features': None, 'splitter': 'best'}

Random Search: {'criterion': 'gini', 'max_depth': 9, 'max_features': None, 'splitter': 'random'}

Porém, apesar do mesmo score, escolhemos os hiperparâmetros do 'Grid Search' por, após ter testado em nosso modelo, este ter obtido maiores métricas em relação ao 'Random Search'.

Esse modelo, para os dados dispostos, apresentou métricas satisfatórias, as quais podem ser utilizadas, futuramente, como parâmetro para a escolha deste algoritmo como sendo o utilizado no modelo que será criado.

Acurácia: 0.958303886925795
 Precisão 0.9503424657534246
 Recall: 0.9487179487179487

Figura 23: Acurácia, precisão e recall do modelo Decision Tree Classifier

Além dos números anteriormente apresentados, esse modelo gerou uma matriz de confusão muito relevante, com um total de 69 erros, sendo eles baixos para os casos de

peças avariadas sendo classificadas como não avariadas, quando comparado a outros modelos.

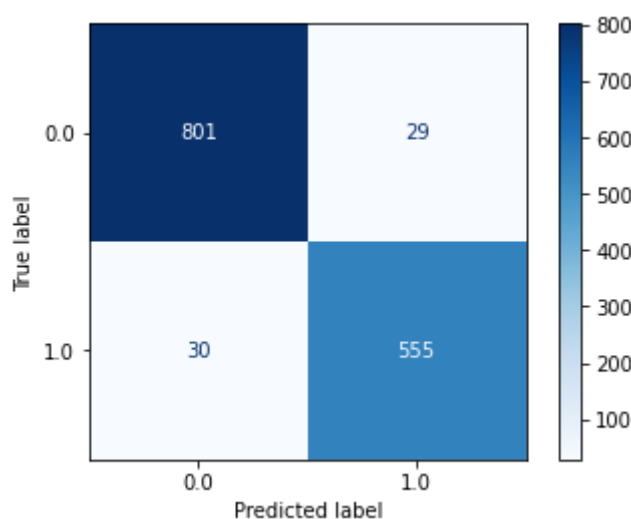


Figura 24: Matriz de confusão do modelo Decision Tree Classifier

Logo, esse algoritmo será considerado no momento de escolha daquele que será utilizado no modelo de predição solicitado pelo banco Pan.

- **Gradient Boosting:**

A partir de definições presentes na biblioteca Scikit Learn, o algoritmo Gradient Boosting constrói um modelo aditivo de forma progressiva, permitindo a otimização de funções de perda diferenciáveis arbitrárias. Logo, em outras palavras, a aprendizagem desse modelo envolve a construção de um modelo forte usando uma coleção de modelos “mais fracos”, sendo, portanto, um algoritmo tido como reforço.

Após o ajuste de hiperparâmetros, verificamos que a melhor combinação do Grid Search quanto a do Random Search, dão o mesmo score de 0.98, sendo eles:

Grid Search: {'n_estimators': 10, 'max_depth': 3, 'criterion': 'friedman_mse'}

Random Search: {'n_estimators': 10, 'max_depth': 3, 'criterion': 'friedman_mse'}

Esse modelo demonstrou ótimo desempenho para o conjunto de dados disposto. Dentre todos os resultados, o melhor deles foi encontrado com a utilização do learning rate de 0.5. Com base nessas métricas, esse é o modelo favorito no processo de escolha.

```
Learning rate: 0.5
Acurácia (treinamento): 0.986
Acurácia (teste): 0.993
Acurácia (final): 0.993
Precisão: 0.997
Recall: 0.984
```

Figura 25: Acurácia, precisão e recall do modelo Gradient Boosting

Além das métricas exemplares, esse modelo apresenta uma matriz de confusão com ótimos resultados, tendo o menor índice de erros dentre todos os modelos analisados.

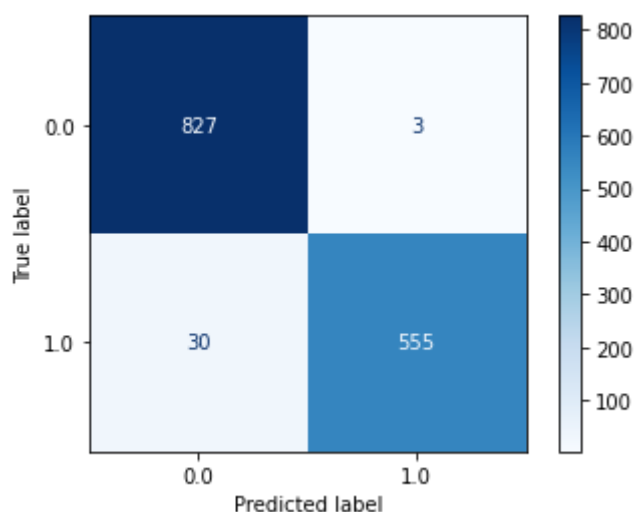


Figura 26: Matriz de confusão do modelo Gradient Boosting

Resumindo, os scores de cada modelo ficaram desta forma:

ACURÁCIA DOS MODELOS

	Default (sem modificação de hiperparâmetro)	Random Search (usando modificação de maior score de hiperparâmetro)	Grid Search (usando modificação de maior score de hiperparâmetro)
Gradient Boosting	99.1	99.3	99.3
Random Forest	98.9	99.2	99.1
Gaussian NB	88.8	85.7	88.8
Support-Vector Machine	60.3	60	60
Decision Tree	97.4	99	99.2
Logistic Regression	62.7	69	69
K-Nearest Neighbor	59	60	60

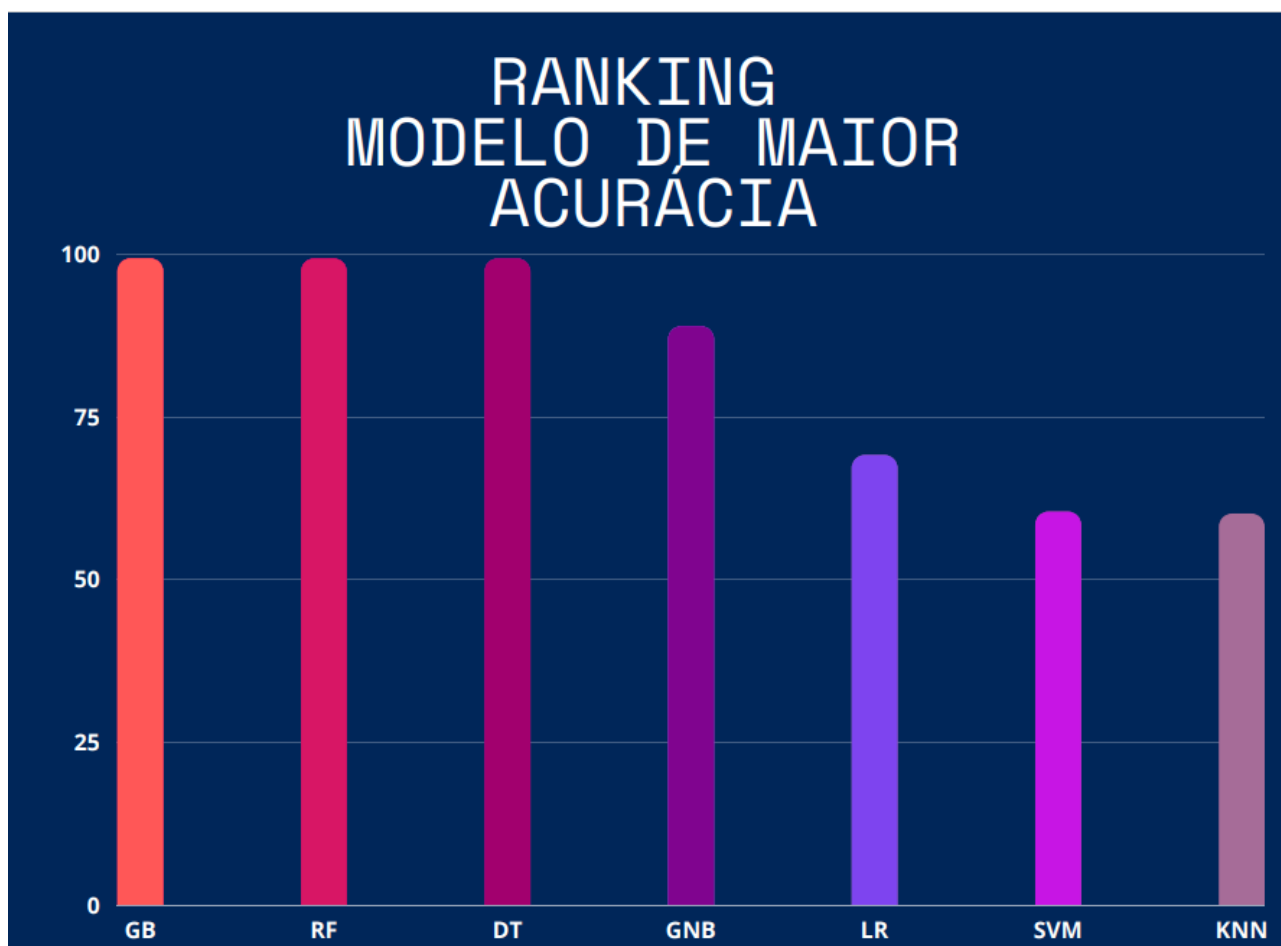


Figura 27: Ranking dos algoritmos testados por acurácia

Legenda do gráfico:

GB - Gradient Boosting, RF - Random Forest, DT - Decision Tree, GNB - Gaussian Naive Bayes
LR - Regressão logística, SVM: Support-Vector Machine, KNN - K-Nearest Neighbor.

Interpretação do gráfico:

No gráfico, é possível ver que os modelos de Gradient Boosting, Random Forest e Decision Tree são os maiores. Além disso, eles estão em valores muito próximos. Já Regressão Logística, SVM e KNN foram os modelos que foram piores em acurácia, onde SVM e KNN foram os últimos com valores muito próximos. Gaussian Naive Bayes é o modelo que ficou como mediano.

PRECISÃO DOS MODELOS

	Default	Random Search	Grid Search
Gradient Boosting	99.7	99.7	99.7
Random Forest	100	100	99.7
Gaussian NB	99.2	98	98.7

Support-Vector Machine	60	0	0
Decision Tree	95.3	99.1	99.7
Logistic Regression	63.6	98.8	98.8
K-Nearest Neighbor	47.3	49.6	49.6

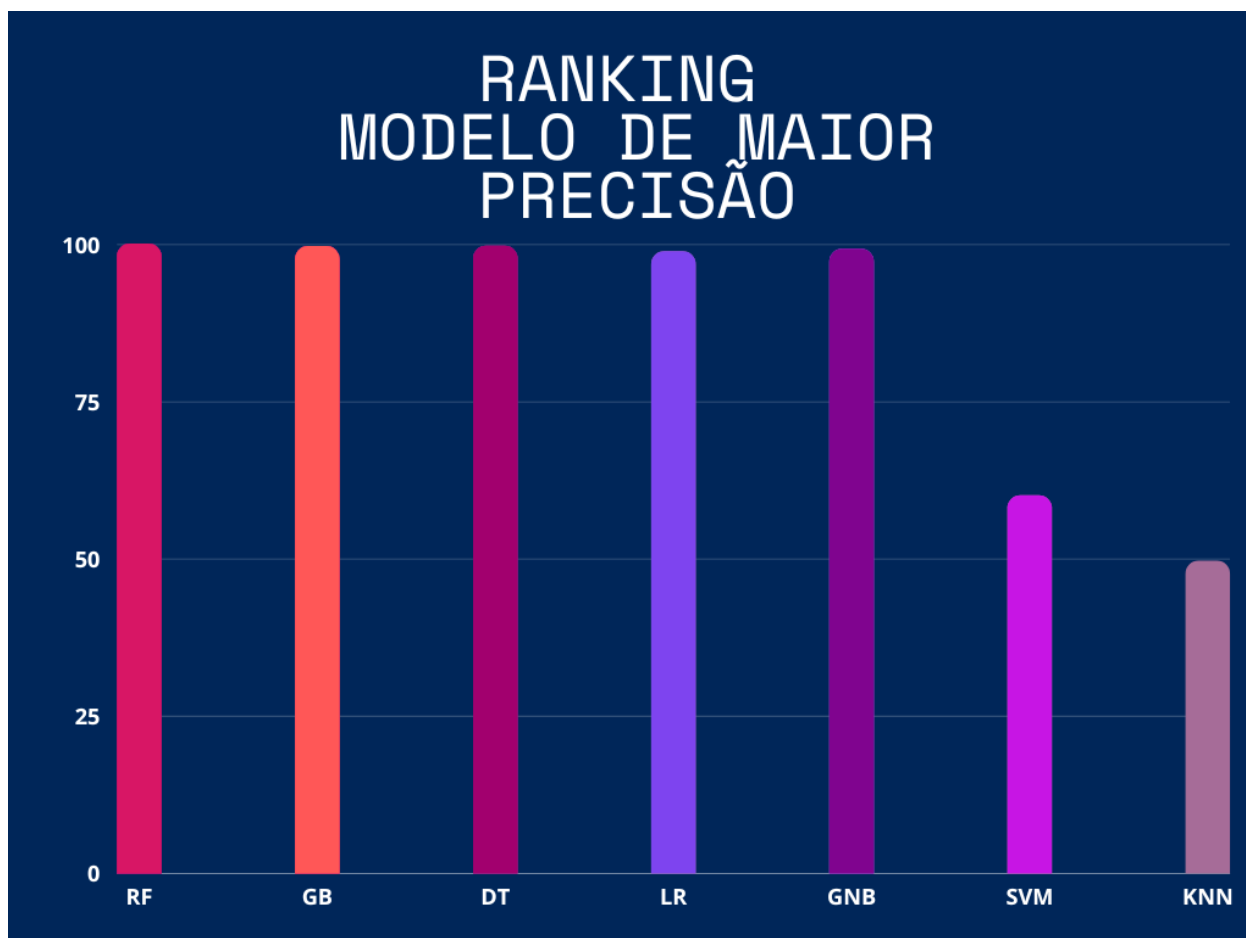


Figura 28: Ranking dos algoritmos testados por precisão

Legenda do gráfico:

GB - Gradient Boosting , RF - Random Forest, DT - Decision Tree, GNB - Gaussian Naive Bayes
 LR - Regressão logística, SVM: Support-Vector Machine, KNN - K-Nearest Neighbor.

Interpretação do gráfico:

No gráfico, é possível ver que os modelos de Random Forest, Gradient Boosting, Decision Tree, Regressão Logística, Gaussian Naive Bayes foram os maiores e tiveram valores muito próximos. Já os de SVM e KNN foram os que tiveram piores desempenhos.

RECALL DOS MODELOS

MODELO	Default	Random Search	Grid Search
Gradient Boosting	98.1	98.4	98.4
Random Forest	97.2	98.1	98.1
Gaussian NB	71.8	64	71.6
Support-Vector Machine	1.9	0	0
Decision Tree	98.3	98.3	98.3
Logistic Regression	15.8	22.2	22.2
K-Nearest Neighbor	43.5	39.4	39.4



Legenda do gráfico:

GB - Gradient Boosting , RF - Random Forest, DT - Decision Tree, GNB - Gaussian Naive Bayes
LR - Regressão logística, SVM: Support-Vector Machine, KNN - K-Nearest Neighbor.

Interpretação do gráfico:

No gráfico, é possível ver que os modelos de Gradient Boosting, Decision Tree e Random Forest, já os de Regressão Logística, Gaussian Naive Bayes foram os maiores e tiveram valores muito próximos. Já os de SVM e KNN foram os que tiveram piores desempenho.

Concluindo, os métodos de ajuste de hiperparâmetros que apresentaram as melhores performances (acurácia, precisão e recall) de cada modelo foram:

K-Nearest Neighbor: Random Search ou Grid Search

Logistic Regression: Random Search ou Grid Search

Gaussian NB: Grid Search

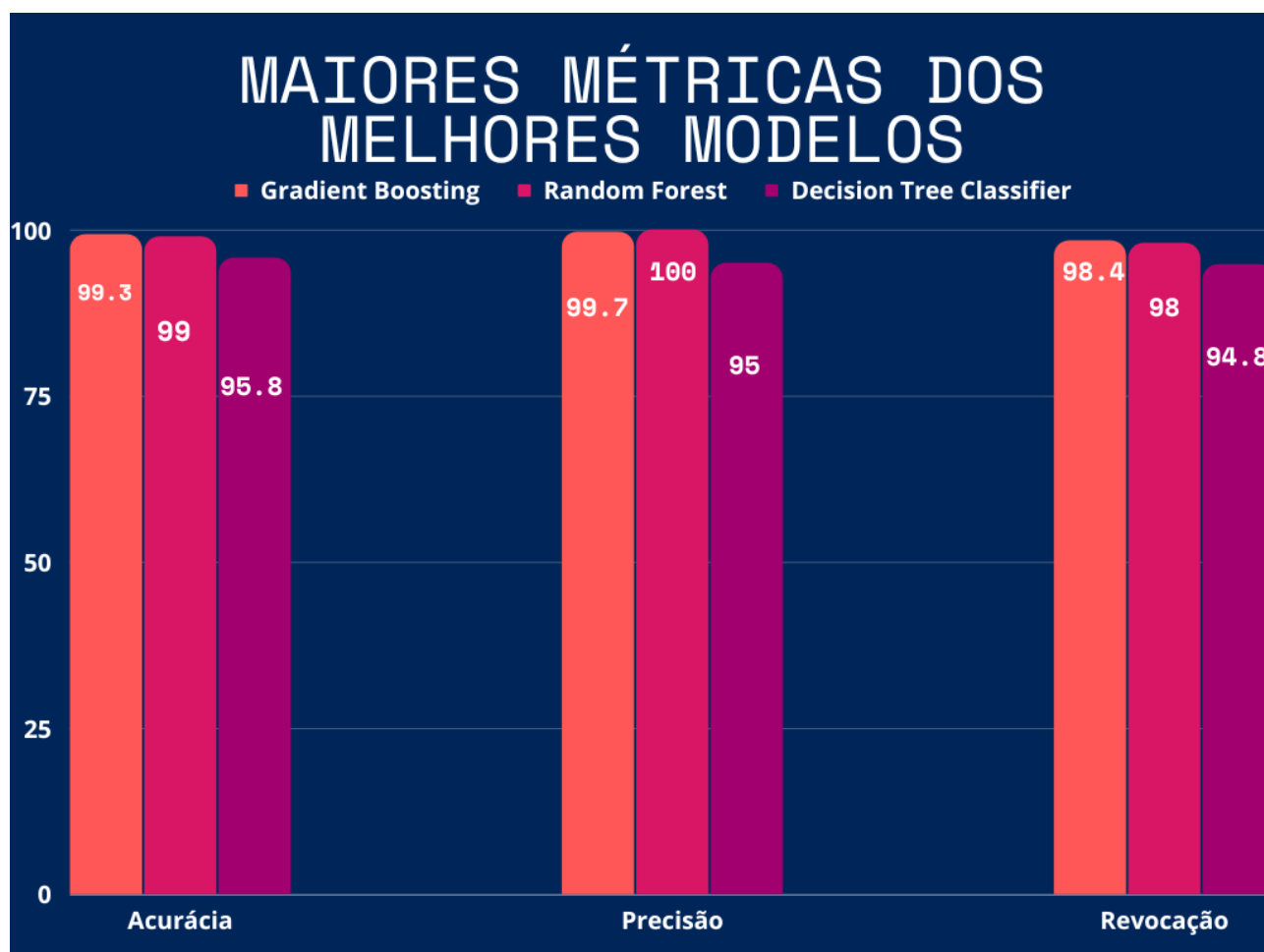
Decision Tree Classifier: Grid Search

Support-Vector Machine: Default

Random Forest: Random Search

Gradient Boosting: Random Search ou Grid Search

Tabela final de desempenho dos modelos:



4.5. Avaliação

Quando foi recebida pelo grupo a proposta do Banco Pan de criar um modelo preditivo de aprendizagem supervisionada que tivesse como impacto a redução das dificuldades pelas quais a empresa passa no que se refere ao atendimento para o cliente, foi decidido que trabalharíamos na criação de um modelo preditivo de classificação, não de regressão. Esta escolha baseou-se no fato que este tipo de modelo preditivo baseia-se em prever a categoria de uma observação previamente fornecida, procurando-se a estimação de um classificador que gere como a sua saída a classificação qualitativa de um não observado com base em dados de entrada, que englobam observações com classificações já definidas.

Ou seja, no contexto da problemática trazida pelo Banco Pan, o modelo se pauta no comportamento prévio dos clientes do banco, de modo que seja possível indicar se ele trata-se, ou não, de um cliente possivelmente atritado com o banco, sendo esta informação exibida no monitor do atendente responsável pelo caso do cliente que está em ligação com o banco.

A partir dos resultados que foram obtidos, como consequência da execução dos sete modelos gerados para a base de dados do Banco Pan, conforme supracitado no item 4.4, foi concluído pelo grupo, através das análises feitas sob apoio das métricas do biblioteca do SciKit Learn, que entre os sete modelos testados, três se destacavam na qualidade de seus resultados: Random Forest, Decision Tree e Gradient Boosting.

Com esta conclusão prévia realizada, o grupo analisou qual destes três modelos apresentaria os melhores resultados para a entrega de um modelo preditivo com a maior eficácia possível em seus resultados, utilizando métricas propícias para algoritmos de classificação. Foi observado pelo grupo que o Gradient Boosting, apresentou os melhores resultados quando o Learning Rate esteve manualmente definido em 0.5, com uma porcentagem de acertos alta e uma pequena quantia de falsos negativos. Ademais, a matriz de confusão gerada a partir desse modelo demonstra excelentes resultados, sendo o menor índice de erros dentre os sete modelos analisados.

As métricas extraídas de cada algoritmo testado foram precisão, revocação e acurácia. Além disso, foram também analisadas as matrizes de verificação para averiguação da proporção de erros e acertos gerados pelo modelo. Para este resultado, no entanto, foi priorizada a métrica de revocação (recall) que é relevante na análise de situações em que os falsos negativos são considerados mais prejudiciais do que os falsos positivos, haja vista que no caso do problema proposto pelo time de tecnologia do Banco Pan, é mais adequado que o algoritmo selecionado apresente casos de falso positivo em detrimento de um modelo com alta incidência de falsos negativos.

O Gradient Boosting trata-se de um algoritmo que faz parte do grupo de classificadores Ensemble que constrói um modelo preditivo progressivamente e permite a otimização de funções de perda diferenciáveis arbitrarias, ou seja, a aprendizagem do Gradient Boosting envolve a construção de um modelo forte usando uma associação de modelos resultantes de

preditores fracos, com o intuito de gerar um melhor modelo preditivo, sendo, por conseguinte, um algoritmo tido como reforço.

Além disso, caso o modelo apresentasse uma considerável amostra de resultados falsos negativos, o atendente do Banco Pan receberia em seu monitor, antes do atendimento, a informação de que o cliente não possui atritos, sendo que o cliente trataria-se de um atritado. Logo, o atendente efetuariam um tipo de atendimento não ideal ao cliente e o modelo preditivo solicitado pelo Stakeholder, Banco Pan, não estaria exercendo o seu propósito, de auxiliar na melhoria dos índices de qualidade do atendimento do Banco Pan.

```
-----
Learning rate: 0.5
Acurácia (treinamento): 0.984
Acurácia (teste): 0.991
Acurácia (final): 0.991
Precisão: 0.997
Recall: 0.981
-----
```

Figura 30: Métricas extraídas para o modelo Gradient Boosting

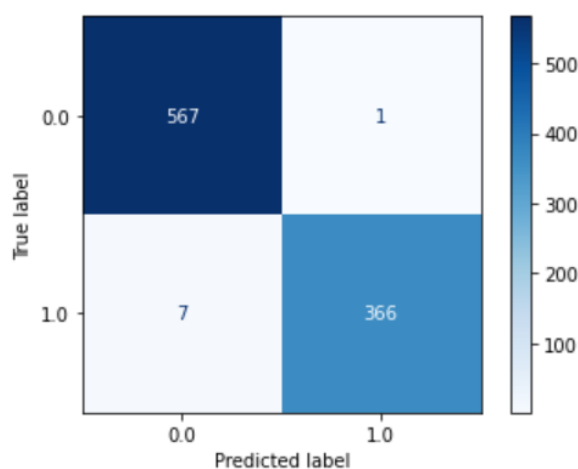


Figura 31: Matriz de confusão do modelo Gradient Boosting

5. Conclusões e Recomendações

Em conclusão, é possível afirmar que o modelo final de Gradient Boosting do projeto cumpre seu objetivo de mostrar se o cliente do Banco PAN é atritado ou não com a instituição. Portanto, aqui estão algumas recomendações para a companhia:

- Implementar o modelo no sistema da empresa;
- Inserir uma aplicação visual de fácil entendimento aos atendentes;
- Fornecer orientação aos atendentes para situações nas quais o modelo encarar o cliente como atritado e não atritado;
- Caso algum código e/ou documentação deste projeto apresente erros ou inconsistências, a empresa poderá entrar em contato com os membros do grupo responsáveis pelo desenvolvimento do projeto: Amanda Ribeiro, Gabriel Rios, Izabella Almeida, Lívia Coutinho, Pedro Baptista e Vinícios Lugli;
- Tratar os dados e informações de maneira ética;
- Trabalhar para gerar uma melhoria no bem estar dos clientes e atendentes do banco, minimizando qualquer tipo de atrito e reclamações, usando nosso modelo como um dos meios para que seja atingido esse objetivo.
- Acompanhar o comportamento do modelo preditivo por meio da métrica de reclamações com o banco, verificando se está fazendo diferença e de qual forma.
- Conforme o aumento da base de dados relativa aos clientes, cabe também à empresa desenvolver e aprimorar ainda mais o modelo entregue, buscando o constante aumento de sua acurácia, precisão e revocação;

O grupo que desenvolveu o modelo coloca-se na responsabilidade de:

- Fornecer possíveis esclarecimentos ao Banco PAN.
- Trabalhar com os dados de forma ética e mantendo em sigilo as informações de conteúdo restrito.

- Documentar e comentar o código, a fim de agilizar o entendimento dos programadores e profissionais de Data Science do Banco PAN.

6. Referências

ANSELMO, Fernando. **Machine Learning na Prática** : Modelos em Python. 1. ed. [S.l.]: -, 2020. p. 1-103.

CACAU, Camila. CHATBOT para bancos: 11 cases globais para conhecer. **TIVIT Labs**, 12 mar. 2021. Disponível em: <https://labs.tivit.com/ivirtualemployee/cases-chatbot-para-bancos/>. Acesso em: 08 ago. 2022.

CONTA AZUL. **Metodologia Scrum: o que é, métodos ágeis e guia prático**. Disponível em: <https://blog.contaazul.com/metodologia-scrum>. Acesso em: 23 set. 2022.

FACELI, Katti; LORENA, Ana C.; GAMA, João; AL, et. Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina. Grupo GEN, 2021. E-book. ISBN 9788521637509. Disponível em: <https://integrada.minhabiblioteca.com.br/#/books/9788521637509/>. Acesso em: 14 set. 2022.

IBM Corporation. (2011). IBM SPSS Modeler CRISP-DM Guide.

INTELLIGENCE BUSINESS MACHINE. **O que é uma árvore de decisão?**. Disponível em: <https://www.ibm.com/topics/decision-trees>. Acesso em: 11 set. 2022.

MEDIUM. **Uma breve introdução ao algoritmo de Machine Learning Gradient Boosting utilizando a biblioteca Scikit-Learn**. Disponível em: <https://medium.com/equals-lab/uma-breve-introdu%C3%A7%C3%A3o-ao-algoritmo-de-machine-learning-gradient-boosting-utilizando-a-biblioteca-311285783099>. Acesso em: 20 set. 2022.

RANKING DE RECLAMAÇÕES. **Reclame Aqui**. Disponível em: <https://www.reclameaqui.com.br/ranking/>. Acesso em 03 out. 2022.

SCIKIT LEARN. **GradientBoostingClassifier**. Disponível em:

<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html>. Acesso em: 11 set. 2022.

STRATEGY, How Competitive Forces Shape. by Michael E. Porter. **Harvard Business Review**, 1979

TOWARDS DATA SCIENCE. **Understanding Random Forest**. Disponível em:

<https://towardsdatascience.com/understanding-random-forest-58381e0602d2#:~:text=The%20random%20forest%20is%20a,that%20of%20any%20individual%20tree>. Acesso em: 10 set. 2022.

TURBAN, R. S. . D. D. . E. **Business intelligence e análise de dados para gestão do negócio**. 4. ed. Porto Alegre: Bookman, 2019. p. 1-566.

UMA VISÃO GERAL SOBRE MACHINE LEARNING. **Stat Place**. Disponível em:

<https://statplace.com.br/blog/uma-visao-geral-sobre-machine-learning/> . Acesso em 05 out. 2022

UPGRADE. **Gaussian Naive Bayes: What You Need to Know?**. Disponível em:

<https://www.upgrad.com/blog/gaussian-naive-bayes/>. Acesso em: 10 set. 2022.

Anexos

Utilize esta seção para anexar materiais como manuais de usuário, documentos complementares que ficaram grandes e não couberam no corpo do texto etc.