Predição com aprendizagem supervisionada
Banco Pan

inteli
instituto
de tecnologia
e liderança

Controle do Documento

Histórico de revisões

Data	Autor	Versão	Resumo da atividade
08/08/2022	Amanda Fontes, Gabriel Rios, Izabella Almeida	1.1	Adição do tópico “Contexto da indústria” e das Personas
10/08/2022	Izabella Almeida	1.2	Adição do Value Proposition Canvas
11/08/2022	Amanda Fontes, Gabriel Rios, Izabella Almeida, Livia Coutinho, Pedro Baptista	1.3	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	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	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	2.3	Refino das descrições das features e adição da documentação referente às manipulações necessárias nos dados
29/08/2022	Gabriel Rios	2.4	Correção no tópico 4.3
06/09/2022	Amanda Fontes	3.1	Redação inicial da metodologia
08/09/2022	Izabella Almeida	3.2	Preenchimento da modelagem na seção 4.4 e avaliação na seção 4.5

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

Nosso parceiro de negócios é o Banco PAN, que tem sua sede em São Paulo (SP) e atua ajudando pessoas das classes C, D e E, transformando desafios em conquistas. Quando a questão é seu porte e posicionamento no mercado, o Banco PAN ultrapassou 15 milhões de clientes no 3º trimestre de 2021.

A problemática envolve o atendimento deles, que, atualmente, não é personalizado para os possíveis propósitos do cliente e do banco. Ele é feito sem muito preparo dos atendentes quanto à situação atual de seus clientes: se estão atritados com o banco, buscam novos produtos ou são potenciais clientes novos. Este processo é um dos responsáveis por um grande número de reclamações, as quais fizeram com que a instituição figurasse como a segunda maior no ranking de reclamações de bancos do Brasil.

2. Objetivos e Justificativa

2.1. Objetivos

Descreva resumidamente os objetivos gerais e específicos do seu parceiro de negócios

2.2. Proposta de Solução

A partir dos objetivos do nosso parceiro de negócios, a nossa solução visa mitigar as dificuldades enfrentadas pela empresa quando o assunto é o relacionamento com clientes do banco. Assim, criaremos 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 da nossa 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

Faça uma breve defesa de sua proposta de solução, escreva sobre seus potenciais, seus benefícios e como ela se diferencia.

3. Metodologia

Descreva as etapas metodológicas que foram utilizadas para o desenvolvimento, citando o referencial teórico. Você deve apenas enunciar os métodos, sem dizer ainda como ele foi aplicado e quais resultados obtidos.

3.1. CRISP-DM

Descreva brevemente a metodologia CRISP-DM e suas etapas de processo

Cross-Industry Standard Process for Data Mining (CRISP-DM) constitui uma das mais importantes metodologias relacionadas ao processo de mineração de dados.

Business Understanding

Data Understanding

Data Preparation

Modeling

Evaluation

Deployment

3.2. Ferramentas

Descreva brevemente as ferramentas utilizadas e seus papéis (Google Colaboratory)

3.3. Principais técnicas empregadas

Descreva brevemente as principais técnicas empregadas, algoritmos e seus benefícios

Algoritmos utilizados:

KNN (K-Nearest Neighbors)

Random Forest

4. Desenvolvimento e Resultados

De maneira geral, você deve descrever nesta seção a aplicação dos métodos aprendidos e os resultados obtidos por seu grupo em seu projeto

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 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

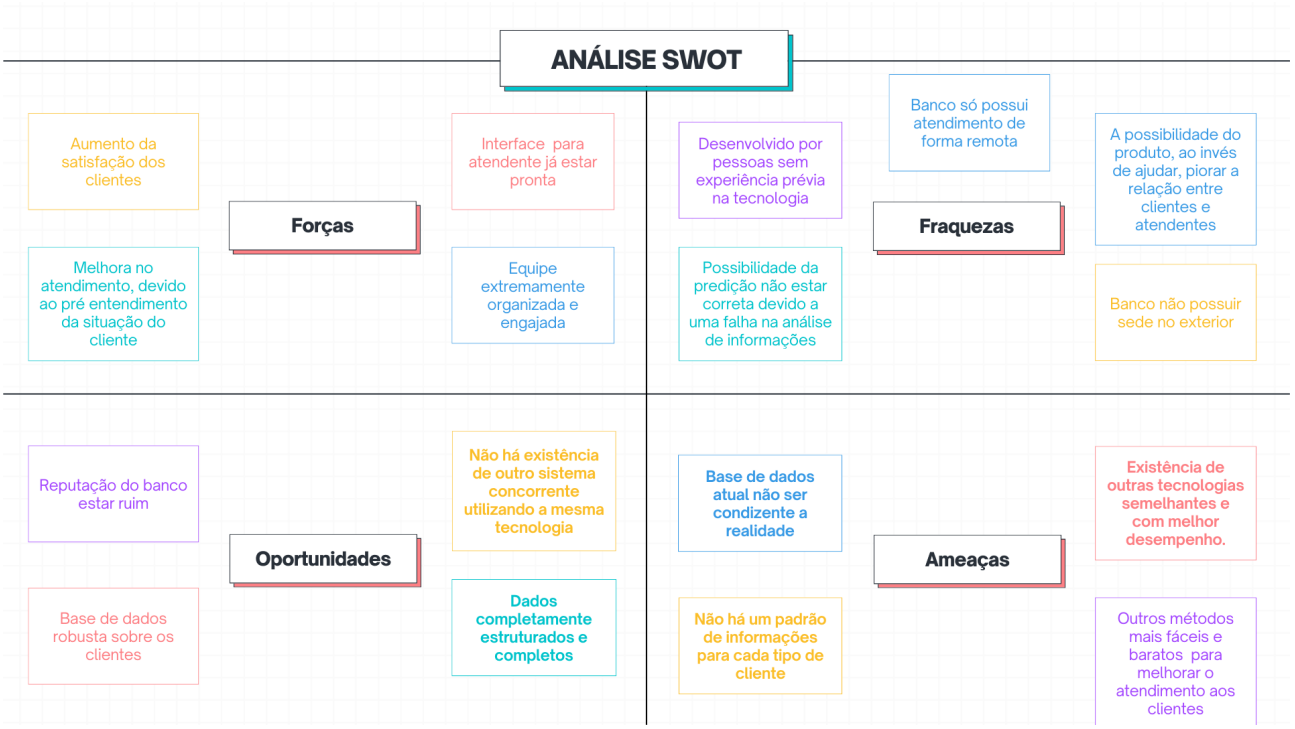


Figura 1: Análise SWOT

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

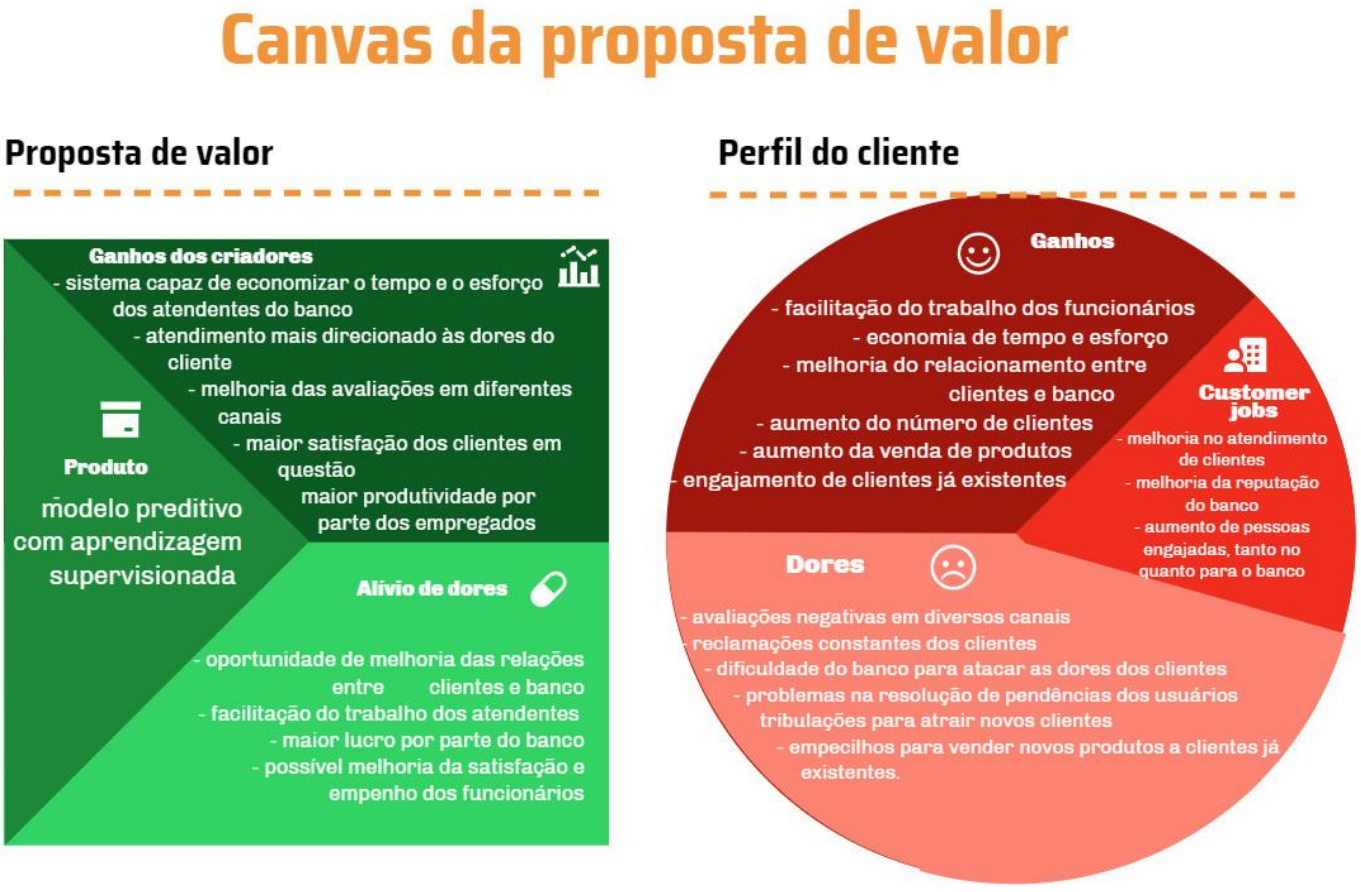


Figura 2: Canvas da proposta de valor

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 Iara, 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: Lorrayne Stephane Soares da Silva

IDADE: 18

GÊNERO: Feminino

OCUPAÇÃO: Estudante/Atendente

"Meu sonho é ajudar o máximo de pessoas possível". Lorrayne 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 atritados
- As vezes se confunde em quando pode oferecer um produto pro cliente ou não

Objetivos/necessidades específicas em relação ao problema:

- Se preparar psicológicamente para um possível cliente atritado
- 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
Lorrayne 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 ocorido novamente e a Lorrayne, 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 atendará 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
- 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 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

1. Descreva os dados a serem utilizados (disponibilizados pelo cliente e outros se tiverem sido incluídos), detalhando a fonte, o formato (CSV, XLSX, banco de dados, etc.), o conteúdo e o tamanho.

a) Se houver mais de um conjunto de dados, descrição de como serão agregados/mesclados.

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) Descrição dos riscos e contingências relacionados a esses 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. Se aplicável: descrição de como será selecionado o subconjunto para análises iniciais:

Não se aplica, visto que não haverá subconjunto para análises iniciais dos dados.

b. Se houver: descrição das restrições de segurança.

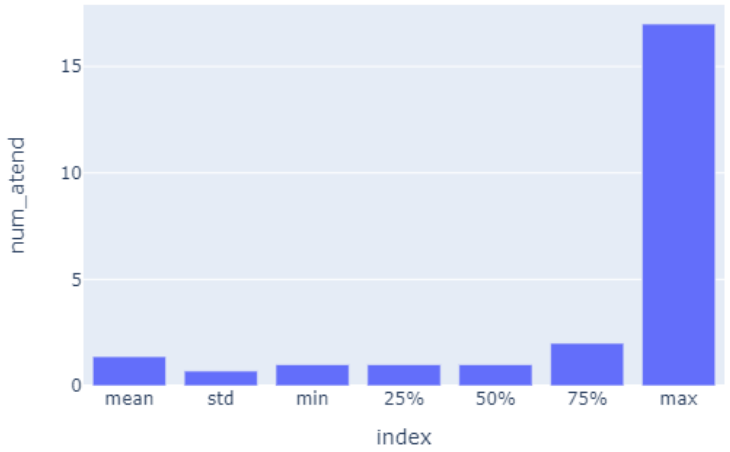
Apesar dos dados já terem sido disponibilizados após uma criptografia daqueles que são sensíveis, a base de dados não deve ser incluída no repositório do GitHub ou em qualquer via de acesso público, sendo gerenciada de maneira independente.

2. Descrição estatística básica dos dados, principalmente dos atributos de interesse, com inclusão de visualizações gráficas e como essas análises embasam suas hipóteses.

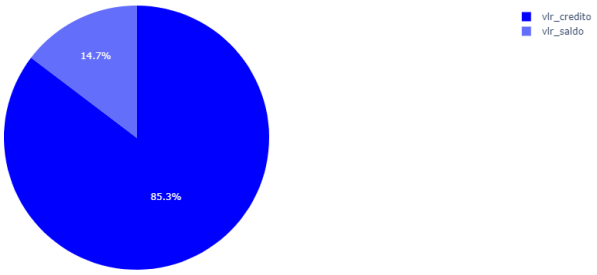
A abordagem do modelo preditivo a ser desenvolvido prioriza as predições relacionadas aos clientes atritados com o banco. Desse modo, para uma melhor visualização dos dados que embasariam esse objetivo, foram definidos os campos de maior relevância da base de dados utilizada pela equipe: quantidade de 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.

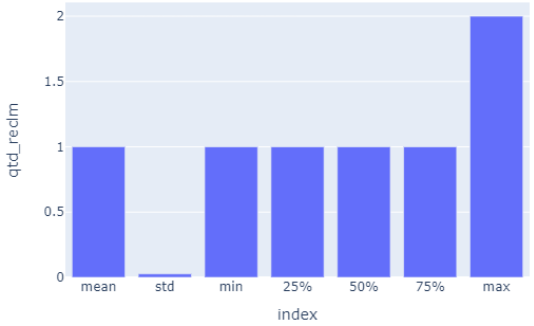


Número de atendimentos por cliente: média, desvio-padrão, mínimo, máximo e quantis



Comparação entre o valor de crédito e o valor de saldo)

Comparação entre o número de atendimentos e o número de atendimentos atrasados



Quantidade de reclamações: média, desvio-padrão, mínimo, máximo e quantis

3. Descrição da predição desejada ("target"), identificando sua natureza (binária, contínua, etc.)

As colunas que serão usadas como "target" serão de natureza binária, sendo elas:

ind_atritado	Indicador de cliente atritado, 1 para Sim e 0 para Não
ind_engajado	Indicador de cliente engajado, 1 para Sim e 0 para Não
ind_novo_cliente	Indicador de potencial novo cliente, 1 para Sim e 0 para Não

4.3. Preparação dos Dados

O primeiro passo 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 e HH foram transformados em 0, 1, 2, 3, 4, 5, 6, 7, 8 e 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 hasheado, 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 predições 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)
Métrica utilizada para agregação de registros: média

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)
Métrica utilizada para agregação de registros: média

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)
Métricas utilizadas para agregação de registros: O maior número.

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)
Métrica utilizada para agregação de registros: Média

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)
Métrica utilizada para agregação de registros: O maior número.

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)
Métrica utilizada para agregação de registros: Mediana

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)

Métrica utilizada para agregação de registros: O maior número

Quantidade de restritivos no mercado

O campo é determinante para que a inteligência artificial a ser desenvolvida crie predições 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)
Métrica utilizada para agregação de registros: O maior

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 predizer 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 predizer o quão 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“

4.4. Modelagem

Para a Sprint 3, você deve descrever aqui os experimentos realizados com os modelos (treinamentos e testes) até o momento. Não deixe de usar equações, tabelas e gráficos de visualização de dados para melhor ilustrar seus experimentos e resultados.

Para a Sprint 4, você deve realizar a descrição final dos experimentos realizados (treinamentos e testes), comparando modelos. Não deixe de usar equações, tabelas e gráficos de visualização de dados para melhor ilustrar seus experimentos e resultados.

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 Classifier e o Gradient Boosting.

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.

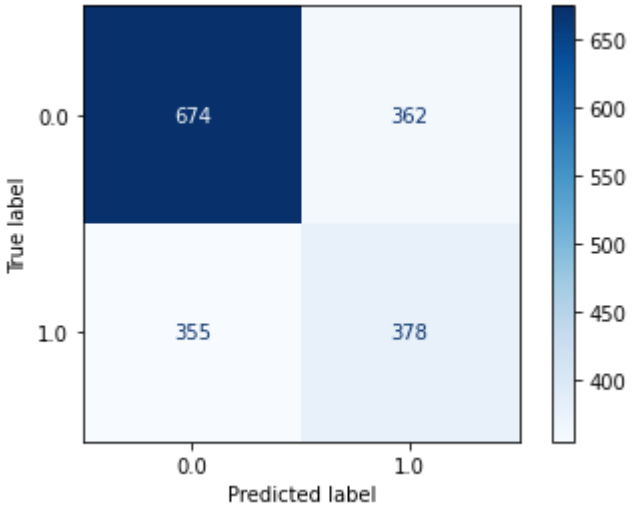
- **KNN (K-Nearest Neighbor):**

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.

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: 0.5946862634256642
Precisão 0.5108108108108108
Recall: 0.5156889495225102
```

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



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.

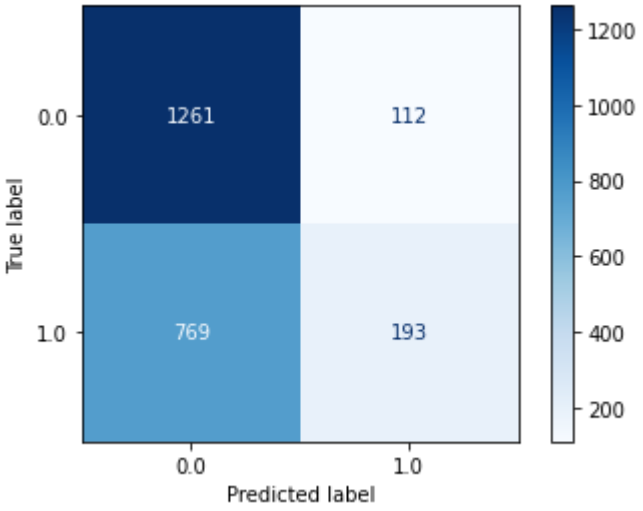
- **SVM (support vector machines):**

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
```

Contudo, ao ser observada, a matriz de precisão 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 e 112 pessoas foram classificadas como atritadas quando não eram. Por isso, esse modelo foi descartado no processo de escolha.



- **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.

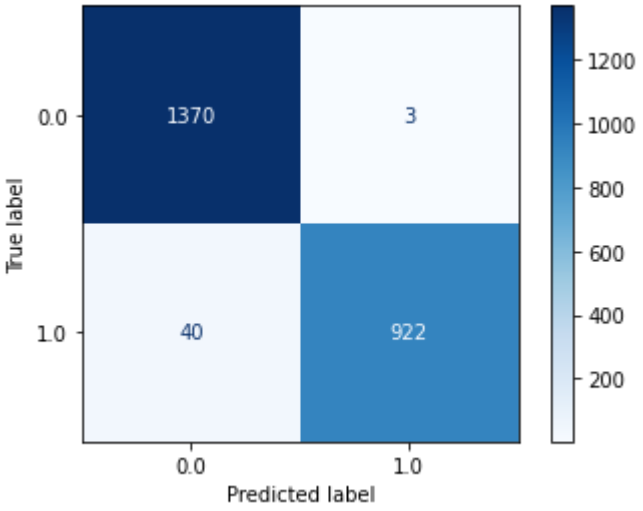
“Um grande número de modelos relativamente não correlacionados (árvores) operando como um comitê superará qualquer um dos modelos constituintes individuais.”

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

Nos testes realizados, esse modelo atingiu ótimas métricas, como é possível visualizar abaixo:

```
Acurácia: 0.9815845824411135
Precisão 0.9967567567567568
Recall: 0.9584199584199584
```

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



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.

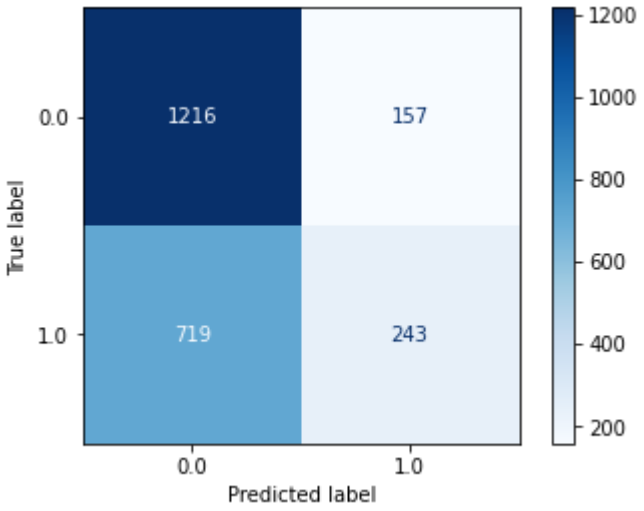
- **Regressão logística:**

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.

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 0.62.

```
Acurácia: 0.6248394004282656
Precisão 0.6075
Recall: 0.2525987525987526
```

Corroborando com 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.



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

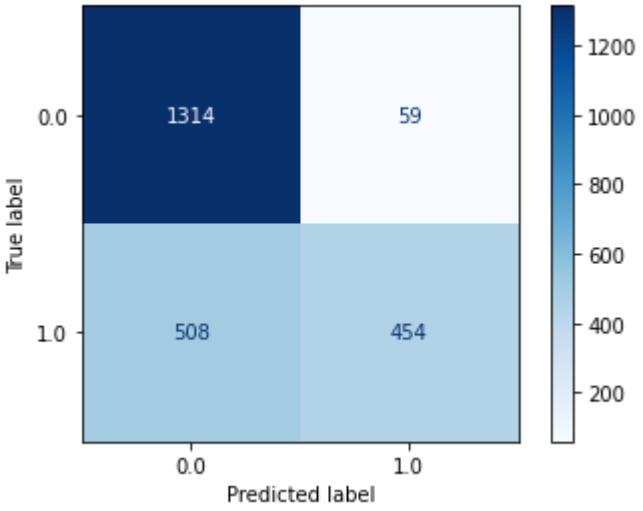
- **Gaussian NB:**

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.

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.

```
R ao quadrado: -0.0023614011232362397
Acurácia: 0.7571734475374733
Precisão 0.884990253411306
Recall: 0.47193347193347196
```

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.



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

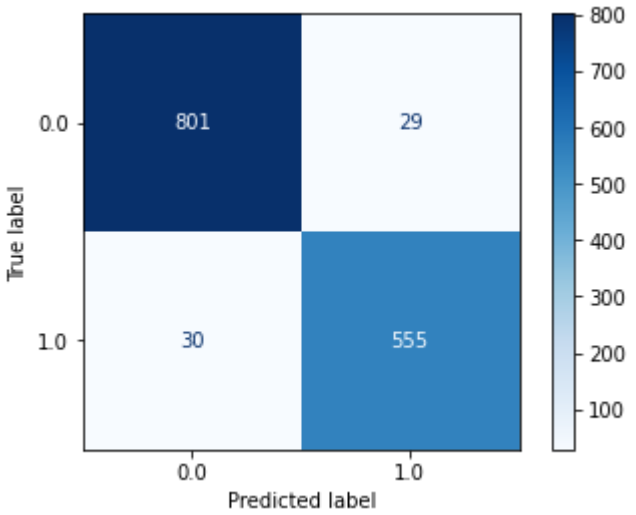
- **Decision Tree Classifier:**

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.

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.

```
R ao quadrado: 0.8280609617959016
Acurácia: 0.958303886925795
Precisão 0.9503424657534246
Recall: 0.9487179487179487
```

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 pessoas atritadas sendo classificadas como não atritadas, quando comparado a outros modelos.



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

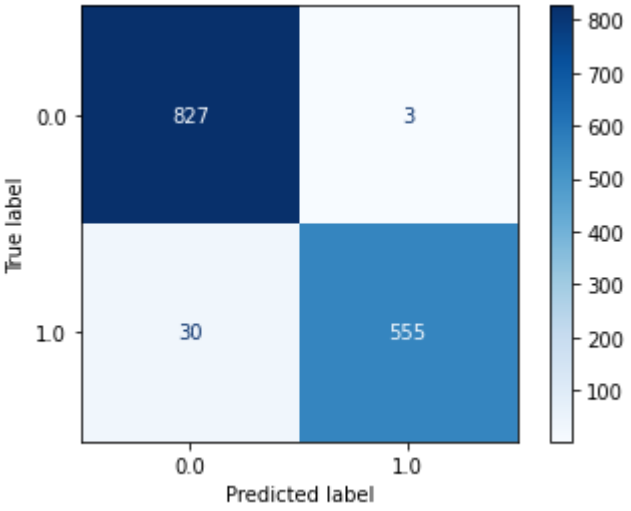
- **Gradient Boosting:**

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.

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): 0.977
Acurácia (final): 0.977
Precisão: 0.995
Recall: 0.949
```

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.


827
3
30
555
0.0
1.0
True label
0.0
1.0
Predicted label
800
700
600
500
400
300
200
100

4.5. Avaliação

Nesta seção, descreva a solução final de modelo preditivo, e justifique a escolha. Alinhe sua justificativa com a seção 4.1, resgatando o entendimento do negócio e explicando de que formas seu modelo atende os requisitos. Não deixe de usar equações, tabelas e gráficos de visualização de dados para melhor ilustrar seus argumentos.

5. Conclusões e Recomendações

Escreva, de forma resumida, sobre os principais resultados do seu projeto e faça recomendações formais ao seu parceiro de negócios em relação ao uso desse modelo. Você pode aproveitar este espaço para comentar sobre possíveis materiais extras, como um manual de usuário mais detalhado na seção “Anexos”.

Não se esqueça também das pessoas que serão potencialmente afetadas pelas decisões do modelo preditivo, e elabore recomendações que ajudem seu parceiro a tratá-las de maneira estratégica e ética.

6. Referências

Nesta seção você deve incluir as principais referências de seu projeto, para que seu parceiro possa consultar caso ele se interessar em aprofundar.

Utilize a norma ABNT NBR 6023 para regras específicas de referências. Um exemplo de referência de livro:

LUCK, Heloisa. **Liderança em gestão escolar**. 4. ed. Petrópolis: Vozes, 2010.

SOBRENOME, Nome. **Título do livro**: subtítulo do livro. Edição. Cidade de publicação: Nome da editora, Ano de publicação.

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.

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

Anexos

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