



TURING LAB EMPRESA BANCO PAN

**Grupo: Felipe Campos, Henrique Marlon,
João Carazzato, Julia Togni, Melyssa
Rojas, Mike Mouadeb**

Controle do Documento

Histórico de revisões

Data	Autor	Versão	Resumo da atividade
08/08/2022	Julia Togni	1.1	Criação do documento
10/08/2022	Julia Togni	1.2	Adição dos critérios avaliados na sprint 1
12/08/2022	Henrique Marlon	1.3	Artefato 1 (proposta de solução e justificativa) e 2 (Compreensão dos Dados)
12/08/2022	Julia Togni	1.4	Atualização das Forças de Porter
24/08/2022	João Carazzato Mike Mouadeb	1.5	Adição dos processos de preparação dos dados
25/08/2022	Julia Togni Mike Mouadeb	1.6	Formatação da seção 4.3 do documento
08/09/2022	Felipe Campos Julia Togni	1.7	Preenchimento dos tópicos 4.4 e 4.5
21/09/2022	Mike Mouadeb Melyssa Rojas	4.4	Complemento do tópico 4.4

Sumário

1. Introdução	5
2. Objetivos e Justificativa	6
2.1. Objetivos	6
2.2. Proposta de Solução	6
2.3. Justificativa	6
3. Metodologia	7
3.1. CRISP-DM	7
3.2. Ferramentas	7
3.3. Principais técnicas empregadas	7
4. Desenvolvimento e Resultados	8
4.1. Compreensão do Problema	8
4.1.1. Contexto da indústria	8
4.1.2. Análise SWOT	11
4.1.3. Planejamento Geral da Solução	12
4.1.4. Value Proposition Canvas	13
4.1.5. Matriz de Riscos	14
4.1.6. Personas	15
4.1.7. Jornadas do Usuário	17
Do usuário(a) do modelo	17
- Da pessoa afetada pelo modelo	18
4.2. Compreensão dos Dados	19
4.3. Preparação dos Dados	22
4.4. Modelagem	23
4.5. Avaliação	28
4.5.1 Análise dos Resultados Finais	29
5. Conclusões e Recomendações	30
6. Referências	31

Anexos

32

Link oficial do collab:

32

1. Introdução

Nesse módulo o parceiro de negócio é o Banco Pan. É um banco brasileiro, com sede na cidade de São Paulo, controlado pelo BTG Pactual. Atua nas áreas de cartões de crédito, crédito consignado, financiamento de veículos, investimentos de renda fixa e banco digital. Atualmente composto por 378 funcionários, e com uma cartela de clientes que ultrapassa os 19 milhões. Focados na atuação junto às pessoas físicas, especificamente das classes C, D e E.

Devido à seu grande público, tem enfrentado problemas de relacionamento, sendo o banco com uma das maiores taxas de queixas nos principais canais de reclamação, como: BACEN e PROCON. Com isso, trouxe a proposta de os alunos dos Inteli desenvolverem um mecanismo capaz de prever se o cliente é uma pessoa com atrito ou não, para eles reduzirem essas queixas e assim melhorar o relacionamento do banco com seus clientes.

2. Objetivos e Justificativa

2.1. Objetivos

Objetivo geral:

O atendimento não é personalizado para os possíveis propósitos do cliente e do banco. Atualmente o processo é manual, dificultando ao atendente ser mais efetivo em abordagens para oferecer mais serviços e produtos. Dessa forma, desejam automatizar essa função a fim de aprimorar o atendimento dos clientes.

Objetivos específicos:

- Diminuir o número de reclamações nos sites como: BACEN e PROCON
- Melhorar o relacionamento dos clientes com o banco

2.2. Proposta de Solução

A nossa proposta é um modelo que, com base em dados simples, retorne um indicador de atrito com o banco. Facilitando o seu atendimento, graduando a atenção dada ao cliente apontado.

2.3. Justificativa

Porque isso irá auxiliar o banco no relacionamento com os seus clientes, parâmetro atualmente em estado crítico, conforme dados do Bacen, os quais apontam ele como o 3 pior banco, sendo o relacionamento um fator que corrobora esse argumento.

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

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

- No tratamento dos dados, desbalanceados, foi aplicada a aplicou significância nos dados, significância de 10%, eliminando os outliers e garantindo maior confiabilidade/verossimilhança dos dados; Para criar a safra selecionou somente os cliente atritados e criou-se duas novas safras atribuindo respectivamente 30% dos dados como sendo cliente atritados e 540% da segunda sendo clientes atritados, criando a proporção de 60/40 e 70/30 para que fosse possível elaborar os testes

4. Desenvolvimento e Resultados

4.1. Compreensão do Problema

4.1.1. Contexto da indústria

Contexto da indústria (principais players, modelos de negócio, tendências)

4.1.1.1 Principais players:

- Banco Caixa Econômica Federal
- Bradesco (BIA)
- Inter (BABI)
- Jira (novo banco digital, direcionado para as classes C, D e E)

4.1.1.2 Modelo de negócio:

Como?

- **Parcerias Principais**
 - Fornecedores de serviços
 - Stakeholders
- **Atividades Principais**
 - Atendimento ao cliente
 - Empréstimo
 - CDB
 - Cartões de crédito/débito
- **Recursos Principais**
 - Parcerias para fornecer serviços
 - Patrocinadores
 - Liquid pool

O que?

- **Proposta de Valor**
 - Serviços de automóveis
 - Serviços de saúde
 - Serviços de alimentos
 - Atendimento ao cliente

- Ajuda ao cliente
- FAQ

Para quem?

- **Relacionamento com os clientes**
 - Central de atendimento
 - Via telefone
 - Via e-mail
 - Redes Sociais
 - Vídeos
- **Segmento de Clientes**
 - Todo o Brasil
 - Focado nas classes C, D e E
- **Canais**
 - Sites
 - Marketing
 - Redes Sociais
 - Recomendações
 - Telefone
 - Centrais de Atendimento

Quanto?

- **Estrutura de Custos**
 - Atendentes
 - Investimentos (Novos serviços, novas tecnologias e entre outros)
 - Fornecimento de Serviços
- **Receitas**
 - Empréstimos
 - Clientes depositando dinheiro
 - Vendas de serviços

4.1.1.3 Tendências

Últimos encontros entre executivos de grandes bancos têm revelado o interesse e a necessidade dos bancos de desconstruir seu modelo atual de se comunicar com o cliente, para evoluir para um modo mais personalizado, por meio de IA. Esta tecnologia tem o desafio hoje de


interpretar os milhões de dados que as empresas coletam e transformá-los em insights úteis sobre os clientes, o mercado, seus serviços e assim gerar mais receitas ou diminuir os custos.

4.1.1.4 5 Forças de Porter:

1. Ameaça de produtos e serviços substitutos:
 - Bancos já mais influentes no mercado
 - Fintech's emergentes no ramo
2. Poder de barganha dos fornecedores
 - Suspensão do serviço por parte do BTG, quem administra o banco
 - Mudanças na produção. Alteração da carteira de produtos anteriormente disponíveis
 - Os parceiros externos, referente ao produto de saúde deles, por exemplo, farmácias e clínicas parceiras não mais atenderem clientes
3. Poder de barganha dos compradores
 - Portabilidade bancária dos clientes
 - Não se interessarem pelos produtos oferecidos pelo Pan e procurarem em concorrentes
 - Desistência do serviço previamente contratado pelos clientes
4. Ameaça de novos entrantes
 - Menos espaço no mercado, por conta do Banco Pan focar nas classes mais baixas da economia brasileira
 - Obsolescência do serviço, o Banco Pan não se atualizar frente às mudanças micro e macro da economia brasileira
5. Rivalidade de concorrentes
 - Concorrentes disputando mesmo marketshare, como o Banco Caixa Econômica que atua nas classes mais baixas e é mais consolidado no território brasileiro, no âmbito de crédito, assim como o Pan
 - Concorrentes maiores no mercado, como o Banco Bradesco, que também possui uma taxa das classes mais baixas
 - Concorrentes com mais marketing, os dois bancos supracitados são mais conhecidos por grande parte da população brasileira
 - Concorrentes com melhores relacionamentos com o cliente, como o Banco Pan possui altas taxas de reclamações, isso pode tornar os outros bancos mais atrativos

4.1.2. Análise SWOT



Problemas para visualizar a imagem:  MatrizSWOT.jpg

4.1.3. Planejamento Geral da Solução

a) quais os dados disponíveis :

Os stakeholders disponibilizaram a base de dados da empresa, onde segmentam os clientes de acordo com as suas características financeiras e com base no seu histórico de relacionamento com o banco.

b) qual a solução proposta:

A nossa solução propõe uma segmentação dos usuários, metrificando o engajamento, relacionamento e, e posteriormente, a sua propensão a produtos do banco, o que melhoraria suas experiências com o Banco Pan. A partir disso, buscamos fazer um modelo de machine learning, que, quando exposto a uma base de dados supervisionada, poderá classificar, clientes e não clientes a partir de seus dados, em um gráfico setorizado, identificando se possuem atritos com o banco ou não a fim de mitigar as reclamações recorrentes da empresa. Em síntese, visamos facilitar e melhorar o atendimento.

c) qual o tipo de tarefa (regressão ou classificação):

Classificação

d) como a solução proposta deverá ser utilizada:

Ao realizar um atendimento, o funcionário do Banco PAN, tendo a sua disposição informações sobre o seu perfil, conseguirá, por exemplo, avaliar se vale ofertar um novo produto para ele ou se aquele cliente está tendo problemas com o banco.

e) quais os benefícios trazidos pela solução proposta:

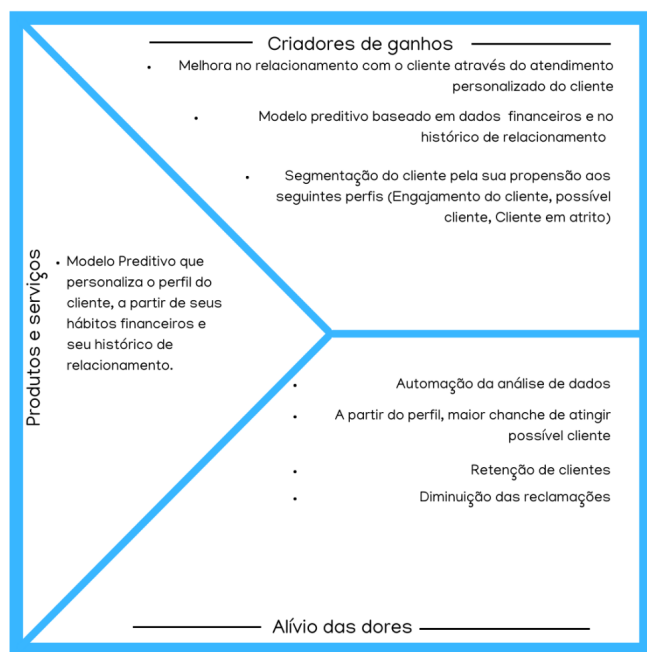
Personalização do atendimento ao cliente, consequentemente melhorando o seu relacionamento com o banco. Além de mitigar a imagem negativa frente ao mercado, conforme o ranking BACEN.

f) qual será o critério de sucesso e qual medida será utilizada para o avaliar:

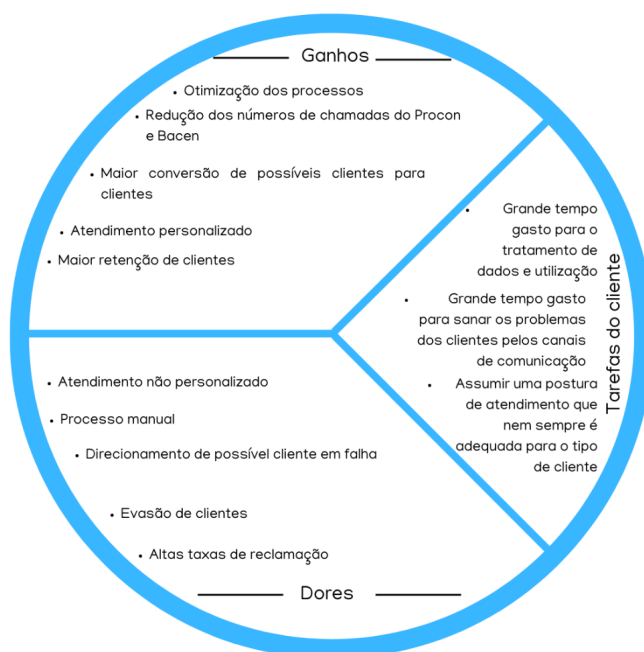
O modelo atingirá sucesso caso consiga identificar e classificar o usuário nas classes apontadas. O benchmark com a possibilidade de identificar randomicamente um usuário da mesma classe, que o modelo identificou.

4.1.4. Value Proposition Canvas

Value proposition canvas



Value proposition



Customer segments

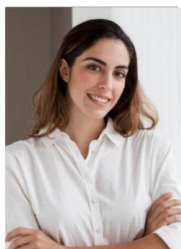
Para melhor visualização acessar: [poster Value proposition canvas simple white.png](#)

4.1.5. Matriz de Riscos

	Ameaças					Oportunidades				
90%						Poucos bancos possuem IA desse tipo	Clientes serem atendidos mais facilmente conforme suas necessidades	Atendentes terem uma facilidade ao atender		
70%				Os dados apresentados serem ineficazes para o algoritmo		A base de dados disposta contribui para um alto grau de aprimoramento do modelo.	Diminuição do CAC (custo de aquisição do cliente)			
50%				Clientes sendo definidos de forma errada pela IA		Melhorar o NPS do banco frente aos clientes				
30%			Condicionalmento dos usuários para utilizarem o modelo proposto		Atendentes não se adaptarem a IA					
10%					IA ser discriminatória					
	Muito Baixo	Baixo	Moderado	Alto	Muito Alto	Muito Alto	Alto	Moderado	Baixo	Muito Baixo

4.1.6. Personas

4.1.6.1 Persona que utiliza o modelo



NOME: Juliana dos Santos

IDADE: 32

OCUPAÇÃO: Atendente de telemarketing

" O domínio das tecnologias nos torna mais preparados para o futuro "

Biografia:

Natural de São José dos Campos

Fez faculdade de contabilidade, mas não atua na área

Entrou na área de atendimento do Banco Pan durante a Pandemia

Características (personalidade, conhecimentos, interesses, habilidades):

Apesar de ter formação, quando entrou na área de contabilidade não se identificou

Adora estar com sua família nas horas vagas

Busca sempre se atualizar sobre tecnologias por conta do seu emprego

Ela é muito habilidosa com o público, sendo então sempre escolhida para gerenciar clientes atritados

Preza por um bom atendimento, mas tem dificuldade com as tecnologias disponibilizadas pelo banco onde trabalha

Motivações com o problema:

Juliana sente-se desafiada diariamente por conta de receber diversas ligações

Ela tenta manejar os conflitos da melhor forma possível, mas o banco não tem um suporte efetivo para os atendentes

Tem dificuldade de analisar e entender o problema do cliente

Dores com o problema:

A falta de informações à deixa desconfortável para falar com o cliente

Ao ter que lidar com clientes atritados, ela se sente ineficiente por conta de não conseguir solucionar o problema por ligação

A comunicação entre os canais é inexistente, deixando o cliente mais atritado

A ineficiência da comunicação faz com que o cliente tenha que refazer suas queixas

Para melhor visualização acessar/ o segundo template da persona está disponível no link também: https://miro.com/app/board/uXjvOhTQ1nE=?share_link_id=273601996726

4.1.6.2 Persona afetada pelo modelo



NOME: Patrick Martins

IDADE: 26

OCUPAÇÃO: Advogado

"Detesto perder tempo com burocracias de banco"

Biografia:

Morador de
Tocantins,
Casado,
cristão

Fez direito na
Universidade
Estadual do
Tocantins

Usuário de
diversos
bancos
digitais

Características (personalidade, conhecimentos, interesses, habilidades):

Patrick
adora jogar
futebol

É apaixonado
por carros e
adoraria ter
uma coleção

Tem
dificuldade
em gerenciar
suas finanças

Está
conectado
em todas as
redes sociais

Gosta muito de
games, filmes
de ficção e
desenhos
animados.

Motivações com o problema:

Patrick quer
conseguir pagar
suas contas no
app do seu banco,
sem burocracias

Ele acredita que
banco digital
deveria facilitar
com sua vida
financeira

Já teve muitos
problemas
com seu banco
digital

Hoje em dia tem
dúvidas se seu
banco digital é o
ideal, tem em
mente trocá-lo

Dores com o problema:

Ele efetuou o
pagamento e o
banco não registrou,
e então aparece que
a conta está em
atraso

Quando tem
problemas com
o app, o banco
não o responde
para ajudá-lo

Mesmo pagando as
contas certinhas,
não melhoram seu
crédito, seu limite
de cartão
permanece o
mesmo

Mesmo tendo pago a
fatura, não aparece no
aplicativo e os
telefonistas do banco
ligam diariamente
cobrando

Para melhor visualização acessar:

https://miro.com/app/board/uXjvOhTQ1nE=/?share_link_id=273601996726

4.1.7. Jornadas do Usuário

- Do usuário(a) do modelo

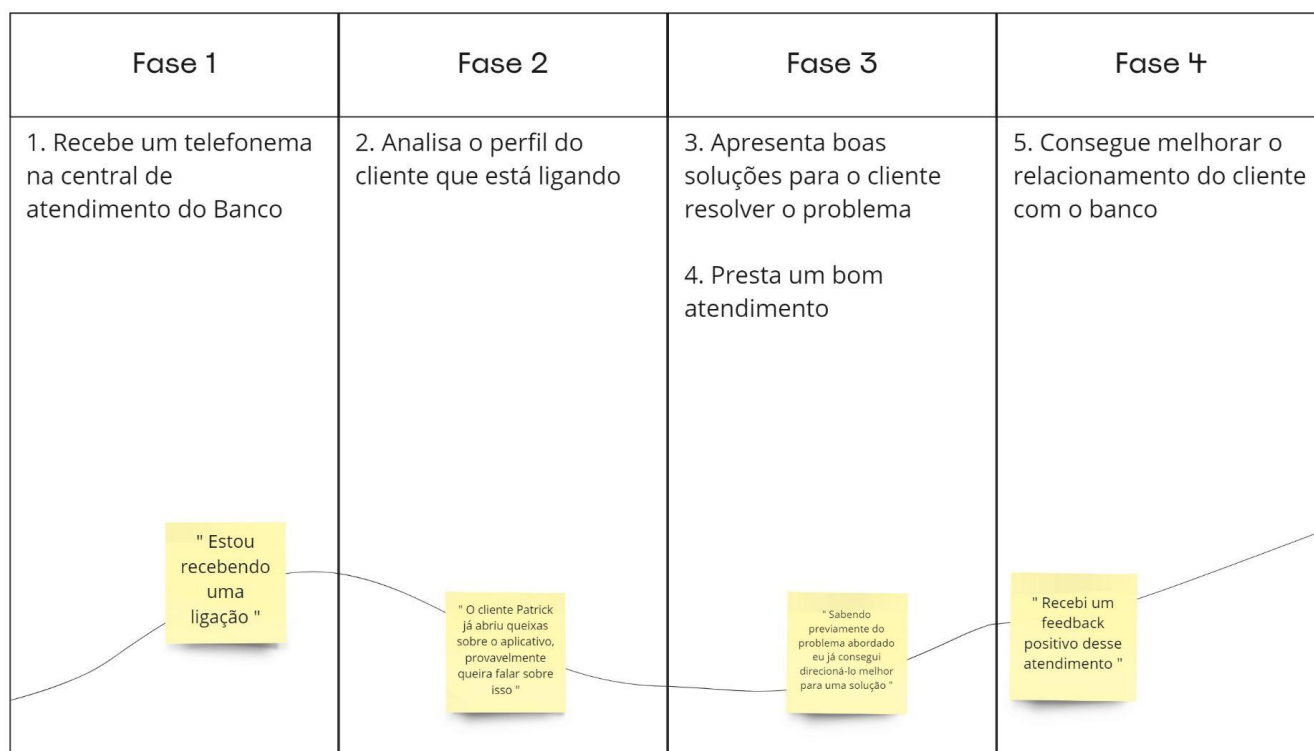


Juliana dos Santos

Cenário: Atendendo clientes do banco, podendo tratar sobre queixas ou dúvidas de novos produtos

EXPECTATIVAS:

- atendimento rápido
- resolução de problemas
- apaziguar a relação do cliente com a empresa



- (para melhor visualização:

https://miro.com/app/board/uXjVOgVUEhl=/?share_link_id=627045217191)

- Da pessoa afetada pelo modelo

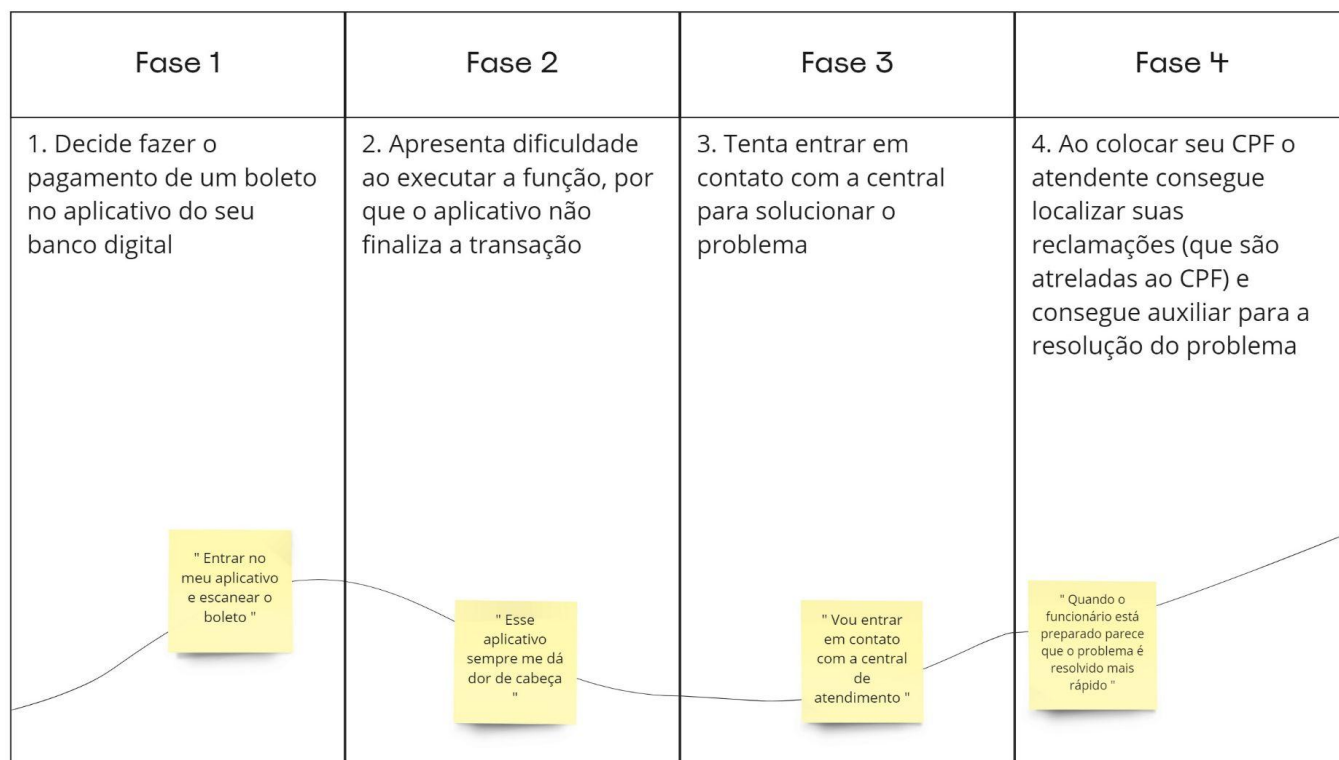


Patrick Martins

Genário: Problema com o aplicativo do banco

EXPECTATIVAS:

- Ligar para o banco
- Solucionar o problema
- Obter sucesso na tarefa



4.2. Compreensão dos Dados

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

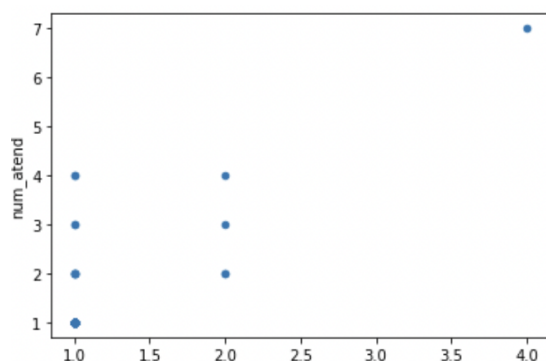
1.1 Os dados que utilizaremos foram disponibilizados pelo cliente através de um book de variáveis anonimizado de sua Base de dados, no formato **CSV** contendo desde informações de saldo e crédito até o número de reclamações e atendimentos realizados para tal cliente com a divulgação proibida. Diante do documento, se vê que possuímos mais de doze milhões de linhas de dados, mas com o empecilho de possuir diversas colunas vazias. Contudo, possuindo uma grande diversidade de dados com 14 colunas cria-se a possibilidade de gerar e identificar o perfil de usuários atritados com o Banco(cliente).

- a. Por enquanto, encontramos as seguintes possíveis mesclas: num_atend X num_atend_atrs, qtd_atend X num_atend_atrs, num_atend_atrs X vlr_saldo.
 - b. Os dados pertencem a um recorte temporal recente.
 - c. [null]
 - d. O book possui os dados anonimizados.
2. Primeiramente, percebemos que apenas dentre os todos os atendimentos apenas 3,7% aprox. são reclamações, contrastando com o ranking Bacen, no qual o Banco Pan se posiciona em 3º lugar. Também percebemos que em média um cliente abre mais de um atendimento no Banco Pan(aprox. 1,37 atendimentos/cliente). Tudo isso, conforme a análise das colunas num_atend e qtd_reclm da tabela a seguir:

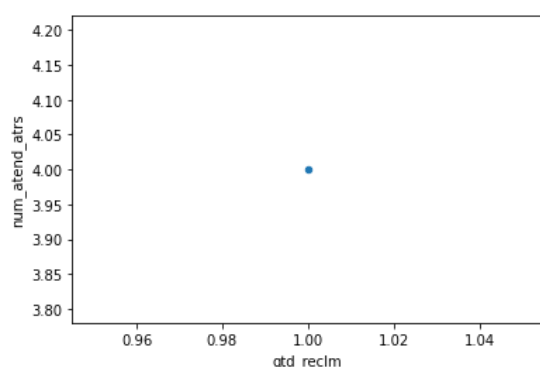
index	anomes	vlr_credito	vlr_saldo	num_atend_atrs	vlr_score	num_produtos	num_atend	qtd_oper	qtd_reclm	qtd_restr
count	12505293.0	7032474.0	6600003.0	6604.0	8738902.0	6688796.0	26545.0	7032474.0	1364.0	8330173.0
mean	202135.33767701403	32065.026640424185	5864.500675229832	1.117655966081163	466.78216920157706	1.6193715580502082	1.3735166680059897	11.980313044882925	1.000733137829912	2.857061431977463
std	42.85810319160321	65672.94291319748	28558.148290187528	0.38907955984007286	207.45916856812772	0.9849523710378423	0.6974935271539803	10.274287798464782	0.027076518053694085	3.5611560558544166
min	202104.0	0.0	0.01	1.0	0.0	1.0	1.0	0.0	1.0	1.0
25%	202107.0	2974.3225	994.69	1.0	329.0	1.0	1.0	5.0	1.0	1.0
50%	202110.0	14245.005	2358.23999999999	1.0	429.0	1.0	1.0	10.0	1.0	2.0
75%	202201.0	33959.935	6748.949999999995	1.0	580.0	2.0	2.0	16.0	1.0	3.0
max	202204.0	10348109.079999998	32102768.81	7.0	1000.0	15.0	17.0	306.0	2.0	413.0

Link: [Captura de tela 2022-08-12 010731.pdf - Google Drive](#)

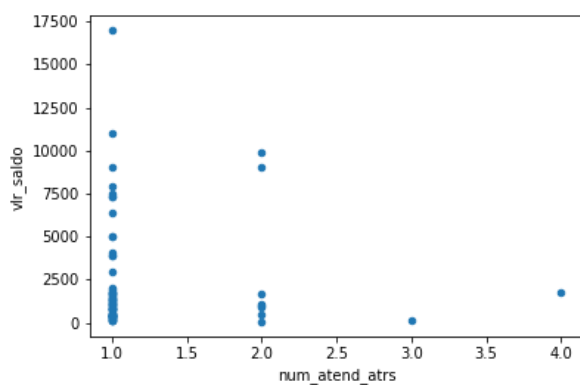
- Também podemos inferir os seguintes gráficos:



(gráfico de número de atendimentos X número de atendimentos atrasados.)



(gráfico de quantidade de reclamações X número de atendimentos atrasados. Observa-se uma correlação entre número de atendimentos atrasados e reclamações)



(gráfico de número de atendimentos atrasados X valor do saldo. Observa-se que o número de atendimentos atrasados se concentra na faixa de saldo até 2500 reais.)

3. A nossa predição possui natureza contínua e tem como objetivo o resultado da coluna `ind_atritado` (coluna que responde se o cliente é atritado ou não é) o target de nossa predição é:

- Coluna `num_atend` (Para se obter a quantidade de atendimentos por cliente).
- Coluna `qtd_reclm` (Para se obter a quantidade de reclamações).
- Coluna `num_atend_atrs` (Para se obter o número de atendimentos atrasados).

4.3. Preparação dos Dados

Para descrever quais manipulações foram necessárias nos registros, selecionamos o mês 11 do ano de 2021, pois este período apresentou o maior número de clientes com índice de atrito. Dessa forma nosso algoritmo terá mais informações para ser treinado. Por conta do número de atritos ser uma porcentagem muito pequena em relação ao mês escolhido, decidimos criar uma safra sumarizada, onde 40% dessa safra indica pessoas atritadas, e 60% indicam pessoas sem atrito. Em suma, decidimos pela safra artificial, para nosso algoritmo lidar com uma proporção de dados melhor, e aumentar a verossimilhança do resultado.

Para a elaboração do nosso algoritmo, não foi necessária a agregação de registros e/ou derivação de novos atributos. Porém, a tabela disponibilizada tinha muitos campos em brancos, sendo necessário o tratamento. Dessa forma, removemos as linhas que possuíam valores ausentes/em branco das colunas `vlr_saldo` e `cod_rating`, e após isso pegamos todas as linhas que usaremos e substituímos os espaços em branco restantes pelo número 0, a fim de preencher todos os espaços que eram necessários para o entendimento da máquina.

Para melhor visualização das features selecionadas, montamos uma tabela com cada uma e a explicação do motivo da seleção:

Campo	Descrição
<code>vlr_score</code>	Conseguimos identificar quem possui um bom relacionamento com o mercado, mas não confia no banco para colocar saldo
<code>vlr_saldo</code>	Quantidade depositada no banco; serve para a gente ver quem é ativo no banco e também conseguimos identificar quem são os clientes do banco
<code>num_atendimento_atr</code>	Iremos correlacionar com a quantidade de reclamações; será um identificador de insatisfação do cliente com o banco
<code>num_atendimento</code>	Quantidade de atendimentos realizados; Conseguimos correlacionar com o número de atendimentos atrasados.
<code>num_reclamacao</code>	A quantidade de reclamações realizadas ajuda a prever o perfil do atritado.
<code>qtd_oper</code>	Quantidade de operações por cliente ; a maioria das pessoas com atrito tem mais de dez operações com o banco.
<code>cod_rating</code>	Classificação do cliente diante do banco e suas interações; Definir qual é a participação do cliente no sistema do banco, para entender se é uma pessoa participativa ou não.

4.4. Modelagem

Nesta Sprint 3 nos aprofundamos no estágio de modelagem da CRISP-DM. Diversos testes foram executados durante essa fase para testar a eficiência do nosso modelo. Estes testes foram feitos sobre a mesma base de dados, porém um na proporção de 30% atritados para 70% não atritados, enquanto o outro foi na de 40% atritados para 60% não atritados. Foi feito os mesmo testes em ambos, extraíndo diferentes resultados para maior riqueza de detalhes do modelo.

Estes foram os testes feitos e suas respectivas descrições e resultados:

Observação: Os códigos feitos são os mesmo para cada tópico, com exceção do nome das variáveis, pois extraem dados de bases de dados com proporções diferentes. Por isso, foi colocado apenas um print de código, para otimização de espaço.

1. Gradient Boost

O Gradient Boost é um método de machine learning por regressão e classificação. Sua ênfase está em criar um modelo composto por vários métodos de predições fracos, mas que juntos conseguem aumentar sua eficiência. Este processo se dá por meio de um encadeamento dessas predições, de modo que uma prevê o erro da outra.

Aqui estão os códigos utilizados no modelo:

```
[ ] from sklearn.ensemble import GradientBoostingClassifier

# Variável que será utilizada para pegar apenas partes da amostra aos poucos para o treinamento do modelo
lr_list = [0.05, 0.075, 0.1, 0.25, 0.5, 0.75, 1]

# Loop que passa por todas as proporções do conjunto de treinamento para o treinamento do modelo
for learning_rate in lr_list:
    gb_clf = GradientBoostingClassifier(n_estimators=20, learning_rate=learning_rate, max_features=2, max_depth=2, random_state=0)
    gb_clf.fit(x_train4060, y_train4060)

    print("Learning rate: ", learning_rate)
    print("Accuracy score (training): {0:.3f}".format(gb_clf.score(x_train4060, y_train4060)))
    print("Accuracy score (validation): {0:.3f}".format(gb_clf.score(x_test4060, y_test4060)))

[ ] # Teste do modelo
y_pred4060gb = gb_clf.predict(x_test4060)

[ ] # Avaliação
import matplotlib.pyplot as plt
from sklearn.metrics import plot_confusion_matrix

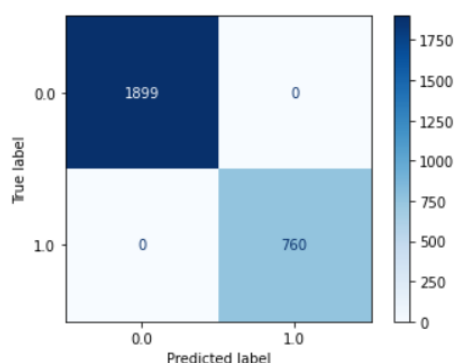
[ ] # Aplicação da matriz de confusão
_ = plot_confusion_matrix(gb_clf, x_test4060, y_pred4060gb, cmap='Blues')
```

E estes foram os resultados, respectivamente, do modelo de 40/60 e 30/70 e suas matrizes de confusão:

```

Learning rate: 0.05
Accuracy score (training): 0.726
Accuracy score (validation): 0.723
Learning rate: 0.075
Accuracy score (training): 0.726
Accuracy score (validation): 0.724
Learning rate: 0.1
Accuracy score (training): 0.726
Accuracy score (validation): 0.724
Learning rate: 0.25
Accuracy score (training): 0.729
Accuracy score (validation): 0.725
Learning rate: 0.5
Accuracy score (training): 0.736
Accuracy score (validation): 0.734
Learning rate: 0.75
Accuracy score (training): 0.744
Accuracy score (validation): 0.718
Learning rate: 1
Accuracy score (training): 0.742
Accuracy score (validation): 0.715

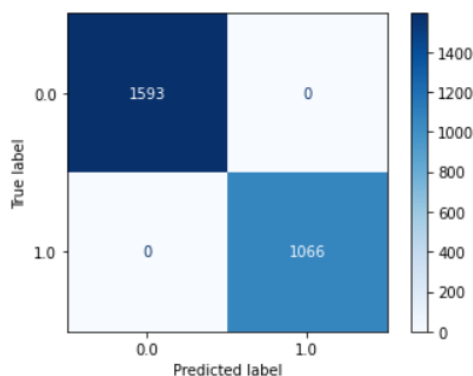
```



```

Learning rate: 0.05
Accuracy score (training): 0.778
Accuracy score (validation): 0.771
Learning rate: 0.075
Accuracy score (training): 0.792
Accuracy score (validation): 0.790
Learning rate: 0.1
Accuracy score (training): 0.792
Accuracy score (validation): 0.791
Learning rate: 0.25
Accuracy score (training): 0.792
Accuracy score (validation): 0.790
Learning rate: 0.5
Accuracy score (training): 0.794
Accuracy score (validation): 0.791
Learning rate: 0.75
Accuracy score (training): 0.791
Accuracy score (validation): 0.789
Learning rate: 1
Accuracy score (training): 0.792
Accuracy score (validation): 0.784

```



2. Árvore de Decisão

A árvore de decisão é um algoritmo de classificação capaz de dividir os grupos, neste caso, entre atritados e não atritados, por meio de um sistema de nós, que se baseiam em cada uma dos dados para tomar decisões.

Aqui estão os códigos utilizados no modelo:


```
[ ] from sklearn import tree
    from sklearn.tree import DecisionTreeClassifier

[ ] # treinamento do modelo pelo algoritmo da árvore de decisão
    clf = tree.DecisionTreeClassifier()
    clf = clf.fit(x_train4060, y_train4060)

[ ] # teste com o modelo
    y_pred4060dt = clf.predict(x_test4060)

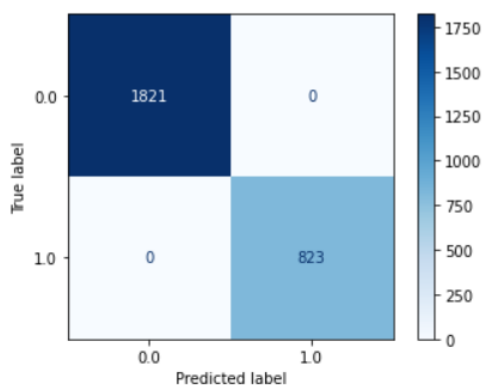
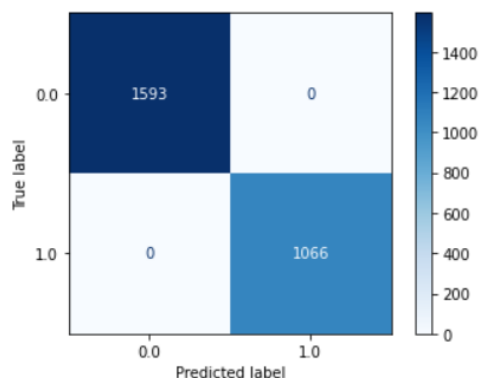
[ ] # Avaliação
    from sklearn.metrics import accuracy_score
    import matplotlib.pyplot as plt
    from sklearn.metrics import plot_confusion_matrix
    from sklearn.metrics import classification_report

[ ] # Aplicação da matriz de confusão
    _ = plot_confusion_matrix(clf, x_test4060, y_pred4060dt, cmap='Blues')

[ ] # Aplicação da métrica da acurácia
    accuracy_score(y_test4060, y_pred4060dt)

[ ] print(classification_report(y_test4060, y_pred4060dt))
```

E estes foram os resultados, respectivamente, do modelo de 40/60 e 30/70:



3. Random Forest

A Random Forest é um algoritmo de classificação, que seleciona uma amostra de dados do conjunto de treinamento e diversas variáveis aleatórias em que estas serão submetidas a cálculos para a criação de um novo nó, e nesse nó o processo dito se repete.

Aqui estão os códigos utilizados no modelo:

```
[ ] from sklearn.ensemble import RandomForestClassifier

[ ] # treinamento do modelo pelo algoritmo random forest
    rfc = RandomForestClassifier()
    rfc.fit(x_train4060, y_train4060)

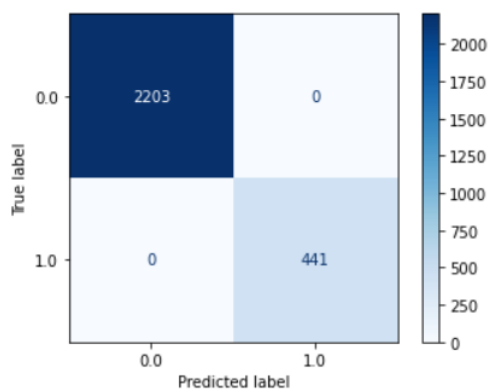
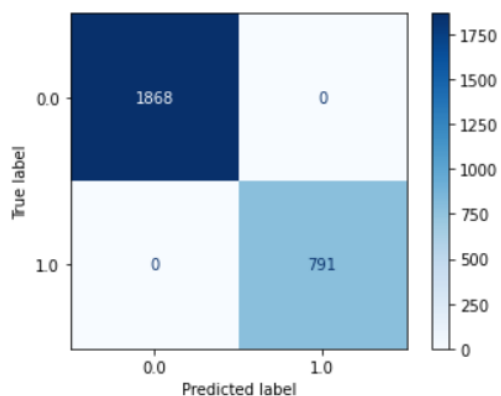
    RandomForestClassifier()

[ ] # teste com o modelo
    y_pred4060rf=rfc.predict(x_test4060)

[ ] # Avaliação
    from sklearn.metrics import accuracy_score
    import matplotlib.pyplot as plt
    from sklearn.metrics import plot_confusion_matrix
    from sklearn.metrics import classification_report

[ ] # Aplicação da matriz de confusão
    _ = plot_confusion_matrix(rfc, x_test4060, y_pred4060rf, cmap='Blues')
```

E estes foram os resultados, respectivamente, do modelo de 40/60 e 30/70:



Nessa sprint 4, além dos experimentos acima, utilizamos os mesmos algoritmos descritos, porém, usufruímos de um novo conjunto de dados, no caso, um conjunto com 60% dos clientes sendo atritados e 40% não atritados. O motivo da escolha deste conjunto para o experimento foi para tentar atingir uma maior quantidade de precisão, levando em consideração uma menor quantidade de falso negativo. Abaixo está a relação do conjunto referente aos algoritmos:

Gradient boost:

```
▼ Gradient Boost

[51] from sklearn.ensemble import GradientBoostingClassifier

[52] # Variável que será utilizada para pegar apenas partes da amostra aos poucos para o treinamento do modelo
lr_list = [0.05, 0.075, 0.1, 0.25, 0.5, 0.75, 1]

# Loop que passa por todas as proporções do conjunto de treinamento para o treinamento do modelo
for learning_rate in lr_list:
    gb_clf = GradientBoostingClassifier(n_estimators=20, learning_rate=learning_rate, max_features=2, max_depth=2, random
    gb_clf.fit(x_train6040, y_train6040)

    print("Learning rate: ", learning_rate)
    print("Accuracy score (training): {0:.3f}".format(gb_clf.score(x_train6040, y_train6040)))
    print("Accuracy score (validation): {0:.3f}".format(gb_clf.score(x_test6040, y_test6040)))

[53] # Teste do modelo
y_pred6040gb = gb_clf.predict(x_test6040)
```

Árvore de decisão:

```
▼ Árvore de decisão

[57] from sklearn import tree
    from sklearn.tree import DecisionTreeClassifier

[58] # treinamento do modelo pelo algoritmo da árvore de decisão
    clf = tree.DecisionTreeClassifier()
    clf = clf.fit(x_train6040, y_train6040)

# teste com o modelo
y_pred6040dt = clf.predict(x_test6040)
```

Random Forest:

▼ Random Forest

```
[63] from sklearn.ensemble import RandomForestClassifier

[64] # treinamento do modelo pelo algoritmo random forest
rfc = RandomForestClassifier()
rfc.fit(x_train6040, y_train6040)

RandomForestClassifier()

[65] # teste com o modelo
y_pred6040rf=rfc.predict(x_test6040)
```

Por fim, a escolha desses algoritmos foi baseada na sua flexibilidade e seu funcionamento ser análogo à condicionais, o que daria mais sentido à nossa lógica.

Hiperparâmetros:

Para nossas escolhas de hiperparâmetros, utilizamos a matriz de correlação, e a partir dessa selecionamos as features, que são nossos hiperparâmetros, que utilizamos em todos os experimentos.

[illegible]

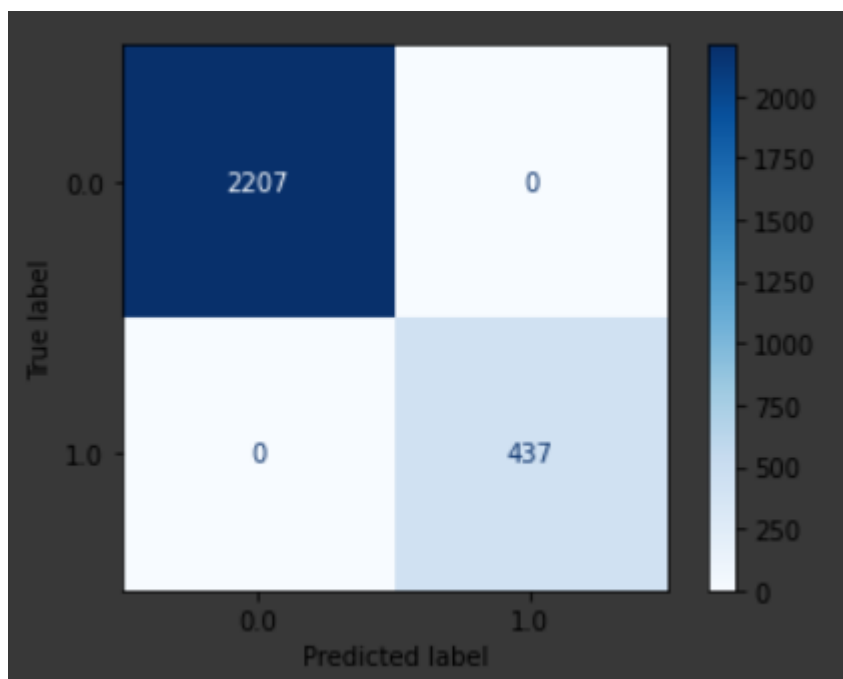
Os hiperparâmetros escolhidos foram: Valor de crédito, valor de saldo, número de atendimentos atrasados, número de produtos, número de atendimentos, quantidade de operações, quantidade de operações e índice de engajamento. Ademais, o motivo da escolha dessas features, é que as outras colunas não selecionadas, possuíam taxas de correlação, referente a nossa coluna de índice de atrito, extremamente baixa, no caso: valor de score e quantidade de restrição.

O Modelo Escolhido:

O modelo escolhido, por enquanto, é o modelo gerado pelo algoritmo random forest com proporção de 70% para o conjunto de treinamento, pois ele possui acurácia de 0.79, ou seja, dentre todos os modelos testados ele possui um dos maiores níveis de acurácia. Além de que, o algoritmo para gerar o modelo não gastava tantos recursos comparados ao XGBoost.

Portanto, a escolha se baseou mais na performance em questões da acurácia, além da matriz de confusão, que, no caso desse modelo escolhido, só teve quantidade numérica diferente de zero na diagonal positiva.

Matriz de confusão:



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.

4.5. Avaliação

Sabendo que, para o treinamento, quanto mais complexo o modelo sua capacidade de generalização é prejudicada assim como se utilizarmos modelos onde na separação dos grupos esses fiquem sem características em comum se os dois grupos forem muito diferentes, o modelo não será capaz de generalizar o conhecimento aprendido com os dados de treino (MÜLLER & GUIDO, 2017). Sob essa perspectiva, foram estruturadas duas safras onde ambas possuem os mesmos dados, mas em proporções diferentes, simulando dois cenários, que mais se aproximam da realidade, onde a empresa tem mais clientes com atrito.

Decidimos aplicar três métricas para testarmos o modelo construído, sendo elas: gradient boost, decision tree (árvore de decisão) e random forest (floresta aleatória). Nesses algoritmos analisamos a acurácia e a precisão para então decidir o mais apropriado para o modelo. Tendo dois modelos para teste, sendo esses uma safra com de distribuição 60/40 e outra com distribuição 70/30, rodamos as três métricas para ambas das safras sumarizadas, apresentando os seguintes resultados:

Para a safra de proporção 70/30 obtivemos:

Gradient Boost		Decision tree	Random forest
Acurácia	0.78	0.70	0.79
Precisão	0.9	0.53	0.82

Para a safra de proporção 60/40 obtivemos:

Gradient Boost		Decision tree	Random forest
Acurácia	0.7	0.67	0.72
Precisão	0.7	0.62	0.75

Realizamos os experimentos de validação cruzada para todos os conjuntos de dados, calculado as métricas de avaliação para todos os casos. As medidas de avaliação de acurácia e

precisão para todos os experimentos. Por fim, a melhor métrica de avaliação dentro dos valores obtidos foi destacada em vermelho.

4.5.1 Análise dos Resultados Finais

Com esses experimentos, as métricas de avaliação se comportaram de maneira melhor do que a esperada. Podemos concluir a partir das métricas que o melhor algoritmo, baseado na taxa de acurácia e precisão, para a safra 70/30, seria o gradient boost, onde as taxas de acurácia e precisão são maiores que as demais, e na safra 60/40, o algoritmo Random Forest nos trazendo resultados mais satisfatórios ao realizar testes.

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.

BANCÁRIOS DE ALAGOAS. **Estudo mostra que bancos não entendem necessidades das classes C, D e E**. Disponível em: <https://bancariosal.org.br/noticia/27665/estudo-mostra-que-bancos-nao-entendem-necessidades-das-classes-c-d-e-e>. Acesso em: 4 ago. 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.

Link oficial do collab:

https://colab.research.google.com/drive/1PMTL1TLI0VIG_myZbu0U0O5kQ6eYEj4T?usp=sharing