



PANORAMA
analytics
Banco PAN



inteli
instituto
de tecnologia
e liderança

The inteli logo consists of a stylized 'i' shape made of red dots, with the word 'inteli' in a large, lowercase, white sans-serif font below it. To the right of 'inteli', the words 'instituto', 'de tecnologia', and 'e liderança' are stacked vertically in a smaller, white sans-serif font.

Controle do Documento

Histórico de revisões

Data	Autor	Versão	Resumo da atividade
08/08/2022	Antônio Ribeiro Cavalcante, Luiz Alencar, Alberto Miranda	1	Criação do documento Atualização da seção 4.1.1
11/08/2022	Henrique Lemos, Alberto Miranda	1.1	Atualização da seção 4.2
12/08/2022	Antônio Ribeiro, Gabriel Carneiro	1.2	Revisão dos tópico referentes a sprint 1

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

O Banco PAN é um banco brasileiro, com sede na cidade de São Paulo, sendo controlado pelo BTG Pactual. Atua nas áreas de cartões de crédito, crédito consignado, financiamento de veículos, investimentos de renda fixa e banco digital, principalmente para as classes C, D e E. A empresa foi fundada em 1969 e possui 378 funcionários.

A instituição financeira possui uma grande base de dados, com isso, deseja utilizá-la para conhecer melhor seus clientes, podendo oferecer atendimentos mais personalizados e resolver o máximo de atritos possíveis.

2. Objetivos e Justificativa

2.1. Objetivos

O objetivo principal com o desenvolvimento do projeto é maximizar a relação do cliente com o Banco Pan, eliminando possíveis atritos entre cliente e o tratante. Dentre as demais metas, estão: retenção e conversão de potenciais clientes.

2.2. Proposta de Solução

Como proposta de solução definimos que após a entrada e análise dos dados, haverá uma target de três colunas novas para o usuário. Estes três novos campos serão: clientes atritados e seus possíveis atritos, clientes sem atrito com o banco mas que possuem alto engajamento e um último campo para caso o indivíduo seja um possível cliente.

2.3. Justificativa

A solução foi desenvolvida visando a melhor experiência do usuário e eficácia dos resultados. Os três novos campos na tabela constituem o **MVP** (*Minimum viable product ou Produto com mínima viabilidade*), portanto minimiza eventuais problemas e conflitos que poderão ocorrer com o parceiro.

A primeira coluna visa entender qual parte dos clientes estão insatisfeitos com seus produtos e serviços e necessitam de uma maior atenção. A segunda, tem como propósito expandir o relacionamento da outra base de clientes que não estão atritados, de alguma forma, com o banco. Por fim, a última coluna busca o aumento da base de clientes.

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.

A metodologia utilizada foi a CRISP-DM constituída pelas etapas de entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e deploy.

3.1. CRISP-DM

A metodologia CRISP-DM possibilita uma visão geral dos processos de mineração de dados que ocorrem em um projeto de *data science*. Como explicitado anteriormente, é formada de 6 etapas.

1. **Entendimento do negócio:** Além de entender o funcionamento da empresa, utilizamos este processo para compreender as expectativas do cliente com o projeto e o que caracteriza o sucesso da solução.
2. **Entendimento dos dados:** Etapa de compreensão dos dados, seus significados e explorá-los ao máximo. Entender a origem das informações é essencial para que possamos posteriormente tratá-las da melhor forma possível.
3. **Preparação dos dados:** Selecionar os dados e fazer as manipulações necessárias. Remover valores em branco, agregar dados e/ou conjunto de dados.
4. **Modelagem:** Encontrar e executar os melhores modelos, e utilizar ferramentas para encontrar as melhores correlações entre features.
5. **Avaliação:** Avaliar os modelos gerados na etapa anterior utilizando ferramentas como F1 Score, Precisão e Revocação, Matrizes de confusão; e validar o melhor modelo
6. **Deploy:** Fazer um repositório, que possibilita ao usuário final, inserir os dados de forma eficiente para gerar um resultado.

3.2. Ferramentas

Google Colab: O Google Colaboratory, ou Google Colab, foi o ambiente utilizado para gerar os testes, manipulações e modelos.

Pandas: biblioteca de software para python, utilizada para manipulação de dados e análise.

Matplotlib: biblioteca para python, utilizada para visualização de dados e no processo de geração de gráficos.

ScikitLearn: ferramenta para python que permite a criação de modelos, separação de conjuntos de dados para teste e treino, gera matrizes de confusão entre outras técnicas.

Google Drive: plataforma em nuvem do google para armazenar e compartilhar arquivos.

GitHub: compartilhamento de código com a comunidade e principal método para análise do que foi produzido.

3.3. Principais técnicas empregadas

RandomForestClassifier: Um tipo de modelo que gera diferentes “Decision Trees” e combina todos os resultados para chegar em um resultado final.

F1_Score: Um modelo matemático que produz o cálculo da média harmônica entre a “precision” e o “recall”.

Matriz de confusão: Um tipo de gráfico que mostra a frequência de classificação para cada classe do modelo. Isto é, ele avalia entre: Verdadeiro positivo que é a previsão ocorreu de forma certa para o valor positivo. Falso positivo que é a previsão errada para um valor positivo. Falso verdadeiro que faz uma previsão certa de um valor negativo. Falso negativo prevê um valor errado para o valor negativo.

Curva ROC: Um modelo matemático que gera uma curva de probabilidade a partir da taxa de variação entre verdadeiros-positivos e falsos-negativos.

GridSearch e Random Search: São técnicas de ajuste de hiperparâmetros que visam à otimização do algoritmo que compõe o modelo preditivo. Busca de forma otimizada os valores que melhor se encaixam nos parâmetros de cada modelo.

Precisão e Revocação:

4. Desenvolvimento e Resultados

4.1. Compreensão do Problema

4.1.1. Contexto da indústria

5 Forças de Porter

- **I. Competição em um setor**
 - O banco PAN não é tão forte nesse ponto, visto que há vários players nesse mercado que competem bem pelo share.
 - Contudo, o diferencial do banco é sua priorização aos clientes de baixa renda.
- **II. Potenciais novos entrantes**
 - O banco PAN tem uma certa força nesse quesito visto que o setor bancário é uma área difícil de entrar e se estabelecer firmemente, ainda mais com o cenário de alta de juros.
- **III. Poder de barganha com os fornecedores**
 - Nesse ponto o poder de barganha do Banco Pan é de médio para baixo. Já que como todo banco o fornecimento de crédito depende da situação econômica do país e política de governo.
 - O banco também necessita de uma base de clientes cada vez maior para aumentar seu poder de barganha com seus fornecedores externos, como clínicas e farmácias (Saúde Pan) e outras plataformas como para o financiamento de carros.
- **IV. Poder de barganha com os clientes**
 - Novamente, o banco Pan não é tão forte nesse quesito, porque seus produtos não são tão específicos, e sendo assim, não pode barganhar muito com os clientes.
- **V. Ameaça de substitutos**
 - Assim como na primeira força, o banco PAN pode ser muito bem ameaçado por outras fintechs que oferecem os mesmo serviços, como Banco Inter, Nubank, Next, entre outros.

Principais players do mercado:

Alguns dos principais players da área de banking são o Bradesco, Caixa Econômica e BTG :

- 1) Bradesco: É um dos maiores bancos do Brasil, sendo extremamente consolidado no mercado e atuando nos grupos C, D e E com boa penetração devido uma longa relação com esse público.
- 2) Caixa Econômica: é o banco de referência já que por ser uma estatal tendem a realizar operações de crédito e financiamentos mais arriscados, visando estimular a economia. Além disso, a confiança que o cliente tem por seu banco estar relacionado a entidade que emite o capital é alta.
- 3) BTG: É o maior case da explosão de bancos digitais, sendo um dos bancos que mais cresce no Brasil. Apesar de visar um público alvo diferente do PAN visando as classes A,B e C,

Modelo de negócio:

O Banco Pan, assim como outras instituições financeiras, tem como *principal produto ofertado o crédito*. Logo seu modelo de negócio é diretamente impactado pela condição financeira do país e o preço do dinheiro no momento. O Banco Pan tem como público alvo pessoas de baixa renda, por isso há mais risco ao ceder crédito, fazendo com que seja necessário maior cuidado ao efetuar empréstimos.

Tendências de mercado:

Atualmente o mercado bancário passa por um forte processo de digitalização e desburocratização. Tarefas que antes envolviam estar presencialmente no banco e demoravam muito tempo e papelada tendem a ser feitas digitalmente de forma simples e rápida. Atualmente o banco Pan tem desempenho mediano na área de digitalização no setor bancário tendo uma página própria e um aplicativo mobile simplório.

4.1.2. Análise SWOT

Matriz Swot

Forças <ul style="list-style-type: none"> Produtos muito bem ranqueados para a classe C D e E . Comunicação muito bem efetiva com o público alvo. Ótima identificação com o cliente.. 	Fraquezas <ul style="list-style-type: none"> Atualmente não possui um atendimento personalizado. Poucos produtos.
Oportunidades <ul style="list-style-type: none"> Humanização dos clientes frente uma época de dificuldades. Open Banking. Faz parte do grupo BTG 	Ameaças <ul style="list-style-type: none"> Bancos digitais pioneiros com uma comunicação muito forte. Público alvo com menos oportunidade de educação financeira

4.1.3. Planejamento Geral da Solução

Foi nos fornecido um arquivo CSV composto por 12.5 milhões de linhas e um total de 16 colunas. Cada linha representa uma pessoa física relacionada a um hash de um CPF. Nas colunas temos os seguintes campos:

- **CPF:** hash de um CPF verdadeiro que não pode ser nulo.
- **Quantidade de produtos que o cliente possui:** número inteiro, que pode ser nulo, que representa os produtos que aquela pessoa tem do banco.
- **Tempo de relacionamento:** tempo que se passou desde a contratação do primeiro produto do banco, podendo ser nulo.
- **Número de atendimentos:** quantidade de atendimentos protocolados de um cliente, não necessariamente finalizados ou resolvidos. Esse Campo pode ser nulo.
- **Número de reclamações:** número de reclamações abertas por um cliente em órgãos externos que o banco tem acesso ou ao FAC do banco.
- **Quantidade de atendimentos atrasados:** número, que pode ser nulo, em que a resolução do problema do cliente passou do prazo determinado para tal.

- **Saldo de crédito do cliente no banco:** soma dos produtos de crédito contratados pelo cliente do banco Pan.
- **Grau de risco crédito cliente:** classificação interna do grau de risco de crédito de cada cliente.
- **Saldo de crédito do cliente no mercado:** soma dos produtos de crédito contratados pelo cliente em qualquer instituição financeira.
- **Quantidade de produto de mercado:** número inteiro, que pode ser nulo, que representa os produtos que aquela pessoa tem de instituições financeiras.
- **Índice restritivo:** número que indica quantos pagamentos atrasados o cliente possui no mercado.
- **Grau de risco crédito mercado:** classificação externa do grau de risco de crédito de cada cliente.
- **Valor renda presumida:** estimativa de renda de cada cliente.
- **Índice atritado:** indica se o cliente está ou não atritado, é discreto/binária.
- **Índice engajado:** indica se o cliente está ou não engajado, é discreto/binária.
- **Índice novo cliente:** indica se o cliente está propenso ou não a entrar no banco, é discreto/binária.

A tarefa da nossa solução é de classificação de clientes em três campos, clientes atritados e seus possíveis atritos, clientes não atritados que possuem um alto engajamento e que podem ter seu relacionamento com o banco maximizado, e por fim, possíveis novos clientes.

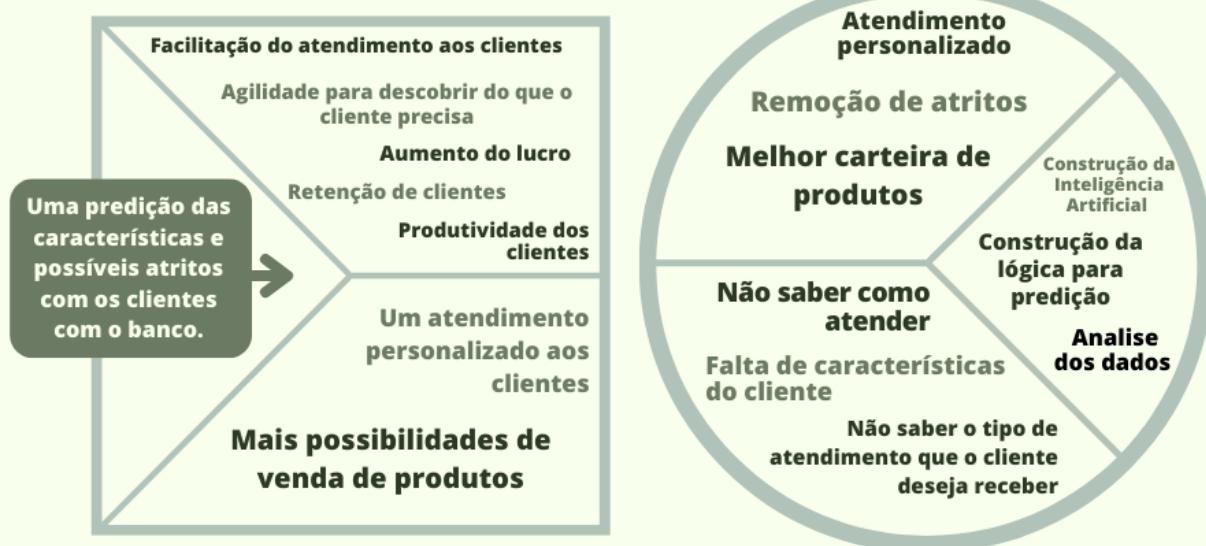
A solução será utilizada para aumentar a eficácia do atendimento do call center do Banco Pan, já permitirá aos atendentes identificarem clientes atritados, novos potenciais clientes.

Nossa solução traz uma maior retenção de clientes, uma maior captura de clientes e uma melhora do relacionamento de quem já é cliente PAN.

O critério de sucesso para a nossa IA será a capacidade de diferenciar entre novos clientes e clientes atritados, e como o atendimento pode se personalizar para encaixar com a situação do cliente.

4.1.4. Value Proposition Canvas

Value proposition canvas



Value proposition

Customer segments

4.1.5. Matriz de Riscos

Chance	Ameaça					Oportunidade				
Muito alta							Uma demora para a atualização dos dados		Quantidade significativa de dados para treinar a IA..	
Alta										
Alta			disparidade dos dados a serem gerados			expectativas muito altas para a entrega		Maior retenção de clientes	Maior obtenção de clientes novos.	
Médio		Impossibilidade de comunicar os dados desejados de forma binária	DataSet incompleto (muitos campos vazios)			falta comunicação entre os membros			Melhor atendimento de clientes pelos atendentes	Aumento do engajamento dos clientes
Médio		Subjetividade de analisar engajamento de cliente		Disparidade de perfil entre usuários			Tratamento do cliente personalizado para determinada situação.		implementação do projeto pelo cliente	
Baixa										
Muito Baixa				Bias inerentes do dataset						Maior vendas de produtos
	1	2	3	4	5	5	4	3	2	1

4.1.6. Personas

Cliente (*afetado pela solução, e afeta também pela autoria dos dados*):



Janete Cruz

Janete tem 36 anos, é autônoma e trabalha no setor de beleza. Seu filho está no 3º ano do Ensino Médio, é mãe solteira e cliente PAN há um bom tempo. Busca um aumento de crédito para ajudar seu filho a ingressar em uma faculdade e investir em seu negócio.

Atendente (*afetado pela solução visto que a utiliza para ter mais produtividade em seu trabalho*):



Anderson Ribeiro

Anderson tem 23 anos, trabalha no call center do Banco Pan há 8 meses. Atende muitos chamados por dia e seu principal estresse é ter que "adivinhar" qual é o "tipo" de cliente que está lidando.

4.1.7. Jornadas do Usuário

Mapeando uma possível jornada do usuário temos:

1. Ator: os vendedores e pessoal de marketing do Banco Pan

2. Cenário + Expectativas: um cenário possível é o vendedor procurando dados para contribuir no seu atendimento. Espera que a IA entregue o máximo de insights possíveis de clientes atritados, com um bom relacionamento ou novos clientes, personalizando assim seu atendimento. Dentre suas expectativas está: entregar um melhor atendimento ao cliente maximizando sua produtividade.

3. Fases da Jornada: após o cliente digitar o CPF pela ligação, o atendente se conecta com o requerente e puxa a ficha com os dados do cliente e coletar os insights disponíveis.

4. a) Ações: Após ser notificado, o vendedor abre a plataforma e puxa os dados do cliente. Em questão de alguns instantes o vendedor recebe as informações mais relevantes e filtradas em grau de prioridade para que ele tenha uma melhor visualização, e prossegue com o protocolo de atendimento.

b) Pensamentos: A princípio o usuário sente motivado com a plataforma, após a pesquisa o vendedor acaba pensando sobre a possível demora para que os dados sejam retornados. Uma vez com os dados do vendedor, a confiança dele será pautada de acordo com a qualidade de informação do cliente.

c) Emoções : Dentre os três pensamentos do usuário, temos: empolgado 😃, receio 😱 e pensativo 🤔

5. Oportunidades: Podemos diminuir a ansiedade do usuário com textos intuitivos e que tragam segurança a ele como ("Falta pouco para os resultados, fique tranquilo, estamos personalizando ao máximo para ajudá-lo") e otimizando ao máximo o tempo de resposta da plataforma e simplificar.

4.2. Compreensão dos Dados

1. Os dados fornecidos são de propriedade do Banco Pan, em formato CSV. O book que contém mais de 12 milhões de linhas possui as informações como o CPF (criptografado), crédito no mercado, crédito bancário disponível, número de atendimentos que o cliente não teve retorno (em 1 ano), score da pessoa no mercado, número de produtos que o cliente possui, quantas vezes ele ligou (apenas para central), operações de crédito, quantidades de reclamações, pendências no mercado e renda prevista pelo mercado.
 - a. **Não aplicável**
 - b. Os dados possuem uma validade pequena, visto que muito deles são atualizados correntemente
 - c. **Não aplicável**
 - d. O book já possui os dados sensíveis criptografados, a única precaução é não quebrar a criptografia.

2. Em uma análise inicial, percebemos que **apenas 5% dos clientes do banco Pan que foram atendidos tem reclamações** e que apenas **0,01% dos clientes possuem reclamações**.

Após isso iniciamos uma análise dos dados fornecidos pelo Banco Pan pela coluna crédito de mercado disponível por pessoa. Nela, notamos que 43% das pessoas não possuem nenhum dado nesse campo (*isso pode representar para nós um público que ainda não é cliente do banco*).

Abaixo segue uma tabela na qual substituímos os campos que não são números por 0 (*essa substituição foi efetuada para ter uma noção concreta em relação ao total de pessoas na tabela*).

index	credito
count	12505293.0
mean	18032.081787934952
std	51753.884498058964
min	0.0
25%	0.0
50%	287.4
75%	17015.05
max	10348109.079999998

Então chegamos ao seguintes dados: **43% das pessoas não possuem dados de crédito** (*podemos trabalhar com esse grupo pensando em potenciais clientes*), **7% das pessoas possuem um crédito de R\$0 - R\$287** (*podemos pensar esse grupo como pessoas que não possuem muitos produtos de crédito*), **25% das pessoas possuem um crédito de R\$287 - R\$17.015** (*grupos de pessoas que possuem crédito bacana, mas nem tantos produtos*) e **25% das pessoas possuem R\$17.015 - R\$10.348.109 de crédito** (*grupos de pessoas que possuem amplos produtos de créditos*).

Separamos dessa maneira pois acreditamos que esses quatro grupos podem ter problemas drasticamente diferentes um dos outros. Dentro desses grupos percebemos que existem aqueles que possuem dados de score e outros não. Dentro de cada novo grupo vamos analisar a quantidade de reclamações, o número de atendimentos, número de atendimentos atrasados e valor de saldo. Selecionamos esses campos pois queremos entender se pessoas com saldos menores possuem mais atendimentos com o banco, se mais atendimentos geram mais atendimentos atrasados e se mais atendimentos atrasados geram reclamações. Isso tudo verificando também se o banco a existência ou não de dado do score influencia em algo. Queremos também identificar a porcentagem de pessoas com ratings baixos e altos em cada grupo para compreender como a informação de crédito influencia na confiança no cliente.

Grupo de pessoas que não temos dados de crédito de mercado:

(Descrição geral): número de atendimentos e reclamações baixo.

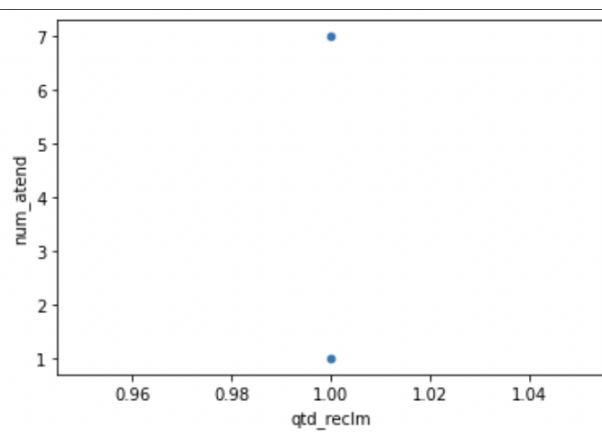
1 to 8 of 8 entries Filter ▾										
index	anomes	vir_credito	vir_saldo	num_atend_atrs	vir_score	num_produtos	num_atend	qtd_oper	qtd_reclm	qtd_restr
count	5472819.0	0.0	553460.0	299.0	2617340.0	158557.0	1625.0	0.0	103.0	5053165.0
mean	202138.19569274993	NaN	6576.011834853216	1.0836120401337792	373.9851280307488	2.24739368176744	1.3526153846153846	NaN	1.0	2.718075701070517
std	44.10152754212618	NaN	90833.14035254323	0.3613439631712712	154.37550917942534	1.436333845984527	0.6653600939086469	NaN	0.0	3.5785882141207517
min	202104.0	NaN	0.01	1.0	0.0	1.0	1.0	NaN	1.0	1.0
25%	202107.0	NaN	1037.91	1.0	268.0	1.0	1.0	NaN	1.0	1.0
50%	202110.0	NaN	2861.285	1.0	363.0	2.0	1.0	NaN	1.0	2.0
75%	202201.0	NaN	7334.129999999998	1.0	446.0	3.0	2.0	NaN	1.0	3.0
max	202204.0	NaN	32102768.81	4.0	1000.0	12.0	7.0	NaN	1.0	413.0

HH	0.735954
A	0.253164
AA	0.005122
B	0.001668
H	0.001200
C	0.001196
D	0.000629
E	0.000490
G	0.000293
F	0.000285

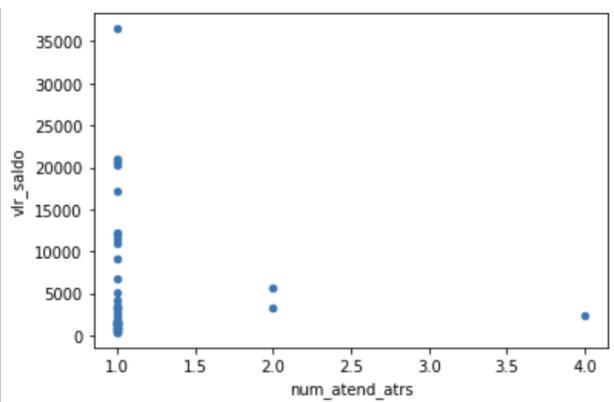
Dentro desse grupo temos o seguinte percentil de rating das pessoas, ou seja, a grande maioria das pessoas que o banco não tem dados de crédito ele classifica como um risco alto para a disponibilização de produtos.

(crédito e score são NaN)

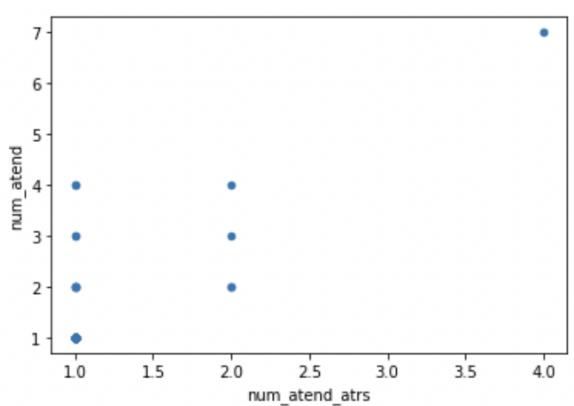
1 to 8 of 8 entries Filter ▾										
index	anomes	vir_credito	vir_saldo	num_atend_atrs	vir_score	num_produtos	num_atend	qtd_oper	qtd_reclm	qtd_restr
count	2855479.0	0.0	167211.0	51.0	0.0	22918.0	293.0	0.0	26.0	2768583.0
mean	202107.6409351986	NaN	6992.21370790427	1.1372549019607843	NaN	2.151060301946069	1.3139931740614335	NaN	1.0	2.6392576996969206
std	8.081408651977991	NaN	163586.5280164983	0.4906977824746006	NaN	1.259896101723838	0.7146045653959636	NaN	0.0	3.502599330905598
min	202104.0	NaN	0.01	1.0	NaN	1.0	1.0	NaN	1.0	1.0
25%	202105.0	NaN	1054.56	1.0	NaN	1.0	1.0	NaN	1.0	1.0
50%	202107.0	NaN	2716.099999999999	1.0	NaN	2.0	1.0	NaN	1.0	2.0
75%	202109.0	NaN	7137.929999999999	1.0	NaN	3.0	1.0	NaN	1.0	3.0
max	202204.0	NaN	32102768.81	4.0	NaN	11.0	7.0	NaN	1.0	413.0



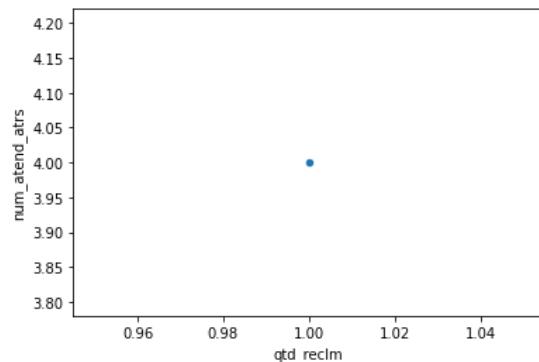
(gráfico de comparação entre números de atendimentos com quantidade de reclamações, na qual percebe-se que não há pontos de atritos no atendimento nessas pessoas, mas percebemos que só há reclamação quando há atendimento)



(gráfico que compara valor de saldo com atendimentos atrasados. Percebemos que quanto menor o saldo da pessoa maior o número de atendimentos atrasados)



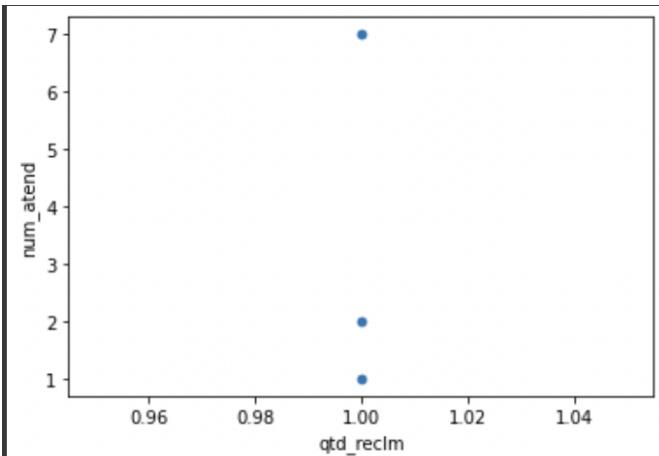
(gráfico que compara número de atendimentos com número de atendimentos atrasados. Não identificamos que quanto maior o número de atendimentos mais o número de atendimentos atrasados.)



(gráfico que compara quantidade de reclamações por número de atendimentos atrasados. Percebemos que só há reclamação quando há número de atendimentos atrasados)

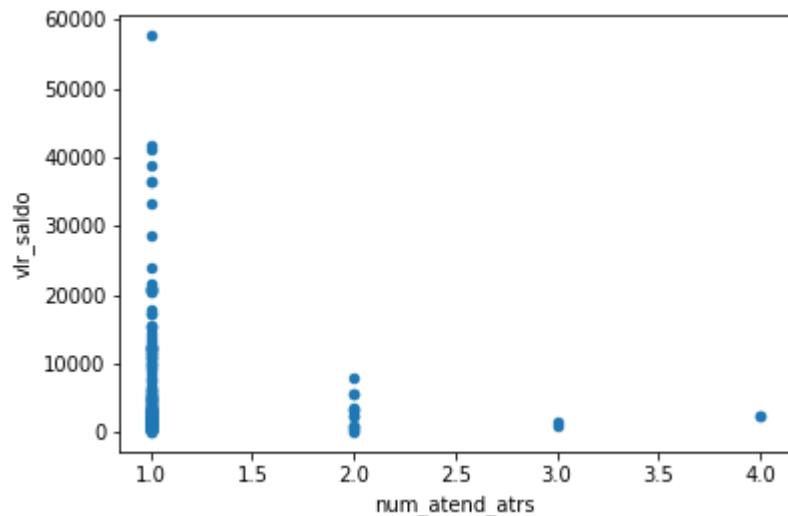
(crédito é NaN e score é diferente de NaN)

index	anomes	vlr_credito	vlr_saldo	num_atend_atrs	vlr_score	num_produtos	num_atend	qtd_oper	qtd_reclm	qtd_restr
count	5471257.0	0.0	552684.0	297.0	2615778.0	157835.0	1619.0	0.0	102.0	5052521.0
mean	202138.17738775568	NaN	6577.592470653285	1.0841750841750841	374.20845155819796	2.2441283618969177	1.3533045089561457	NaN	1.0	2.71800137000915
std	44.094508287116234	NaN	90896.40558786893	0.362497077787024	154.15076810307283	1.4324889423170653	0.6661095935755083	NaN	0.0	3.5785405018934258
min	202104.0	NaN	0.01	1.0	58.0	1.0	1.0	NaN	1.0	1.0
25%	202107.0	NaN	1038.02	1.0	268.0	1.0	1.0	NaN	1.0	1.0
50%	202110.0	NaN	2862.07	1.0	364.0	2.0	1.0	NaN	1.0	2.0
75%	202201.0	NaN	7335.53	1.0	446.0	3.0	2.0	NaN	1.0	3.0
max	202204.0	NaN	32102768.81	4.0	1000.0	12.0	7.0	NaN	1.0	413.0

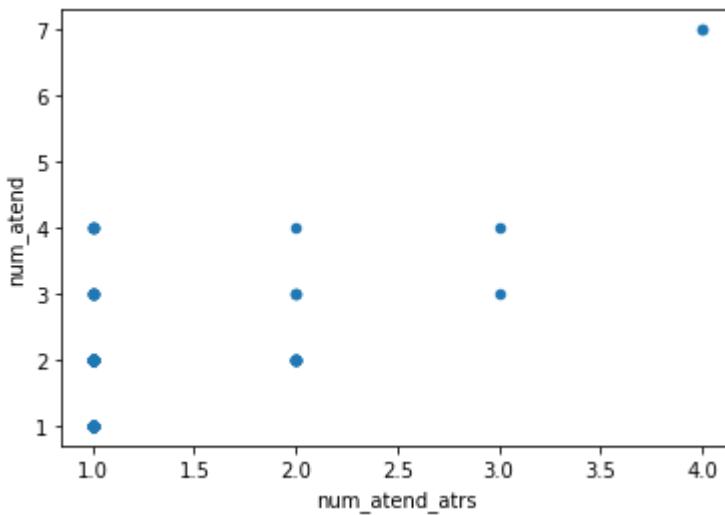


(gráfico

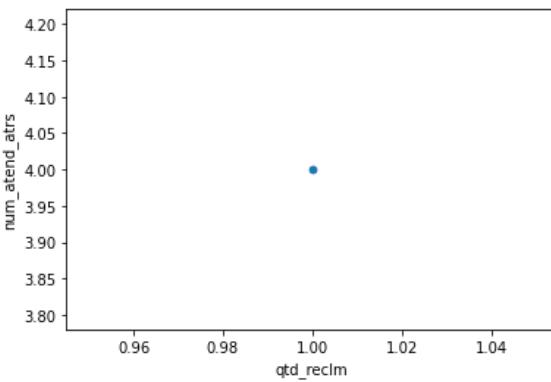
que compara quantidade de reclamações com números de atendimentos. Percebemos que não há relação direta entre números de atendimentos com quantidade de reclamações, mas que só há reclamação quando há atendimento)



(gráfico que compara números de atendimentos atrasados com valor de saldo.
 Percebemos que quanto menor o saldo da pessoa mais a quantidade de
 atendimentos atrasados.)



(gráfico que compara número de atendimentos atrasados com números de
 atendimentos. Percebemos que a quantidade de atendimentos influencia na
 quantidade de atendimentos atrasados)



(gráfico que compara
 quantidade de reclamações com número de atendimentos atrasados)

Grupo de pessoas que tem crédito de R\$0 - R\$287:

(Descrição geral): número de atendimentos e reclamações baixo.

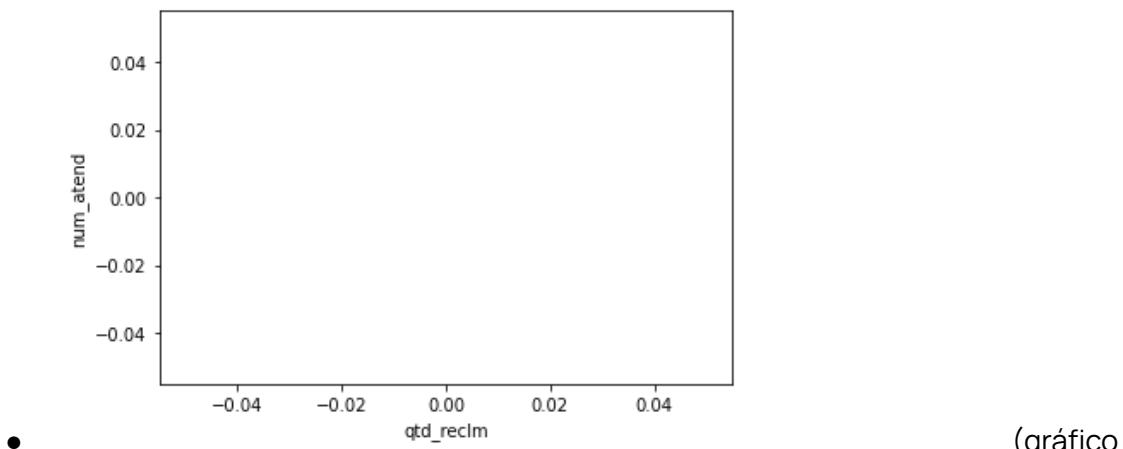
index	anomes	vir_credito	vir_saldo	num_atend_atrs	vir_score	num_produtos	num_atend	qtd_oper	qtd_reclm	qtd_restr
count	105975.0	105975.0	63132.0	80.0	85943.0	90478.0	322.0	105975.0	1.0	48756.0
mean	202131.57724935125	160.4169594715735	2157.323657158012	1.175	458.1851692400777	1.419870236521586	1.34472049689441	3.6301958008964377	1.0	2.8298876035769958
std	40.94847240226497	92.69948335464152	5354.587705736953	0.49746191254243755	218.0245310924172	0.7317946962602946	0.6478722058823735	2.539363742570723	NaN	3.166461927130489
min	202104.0	0.01	0.01	1.0	0.0	1.0	1.0	1.0	1.0	1.0
25%	202107.0	73.475	225.09	1.0	291.0	1.0	1.0	2.0	1.0	1.0
50%	202110.0	196.75	661.97	1.0	446.0	1.0	1.0	3.0	1.0	2.0
75%	202112.0	238.18	1741.159999999999	1.0	615.0	2.0	2.0	5.0	1.0	3.0
max	202204.0	287.4	175667.19	4.0	987.0	7.0	5.0	48.0	1.0	75.0

A	0.667633
HH	0.164101
H	0.075714
D	0.017012
G	0.015206
E	0.014858
C	0.014810
F	0.014050
AA	0.009250
B	0.007366

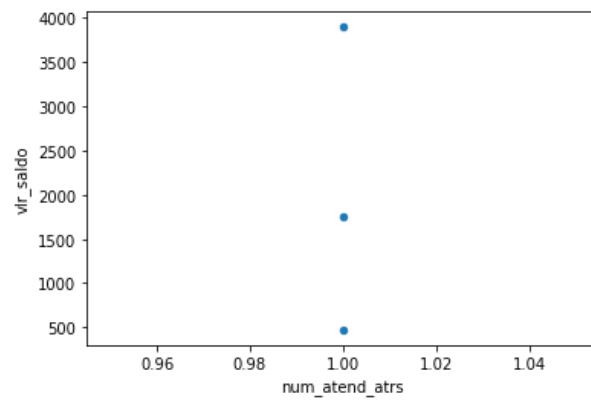
(Percentil de rating dos clientes desse grupo. Percebemos que quando o banco tem os dados de crédito do cliente, ele tende a confiar mais no cliente para disponibilizar novos produtos.)

(valor de score sendo NaN)

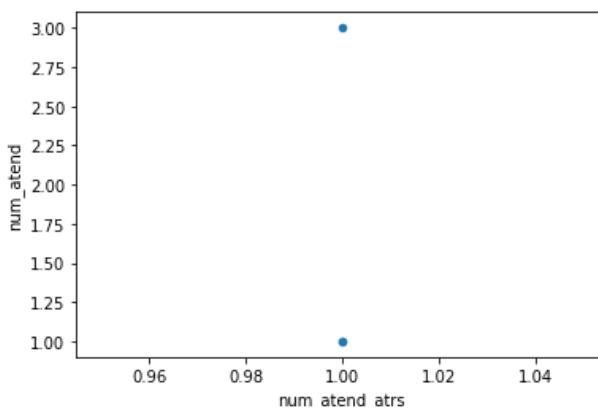
index	anomes	vir_credito	vir_saldo	num_atend_atrs	vir_score	num_produtos	num_atend	qtd_oper	qtd_reclm	qtd_restr
count	19.0	19.0	8.0	0.0	19.0	19.0	0.0	19.0	0.0	12.0
mean	202202.2105263158	177.52684210526317	681.3237500000001	NaN	0.0	1.368421052631579	NaN	4.157894736842105	NaN	1.75
std	0.9763280054720369	92.2505365869749	648.5725059988282	NaN	0.0	0.5972647203701474	NaN	2.0886773564663366	NaN	1.2154310870109943
min	202201.0	2.47	220.64	NaN	0.0	1.0	NaN	1.0	NaN	1.0
25%	202201.5	109.54	349.8775	NaN	0.0	1.0	NaN	3.0	NaN	1.0
50%	202202.0	212.57	425.005	NaN	0.0	1.0	NaN	4.0	NaN	1.0
75%	202203.0	246.5249999999998	625.21	NaN	0.0	2.0	NaN	5.5	NaN	2.0
max	202204.0	281.41	2179.59	NaN	0.0	3.0	NaN	9.0	NaN	5.0



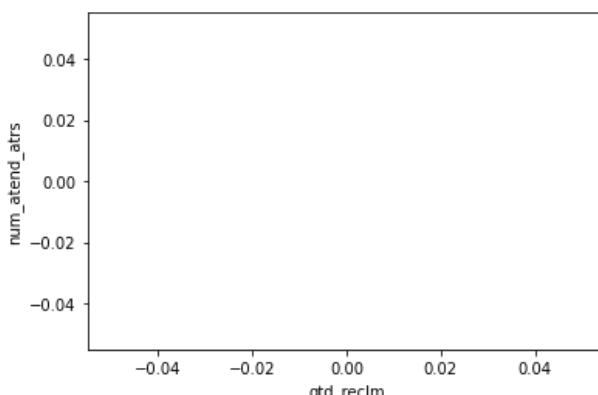
que compara quantidade de reclamações com número de atendimentos)



● (gráfico que compara número de atendimentos atrasados com valor de saldo)



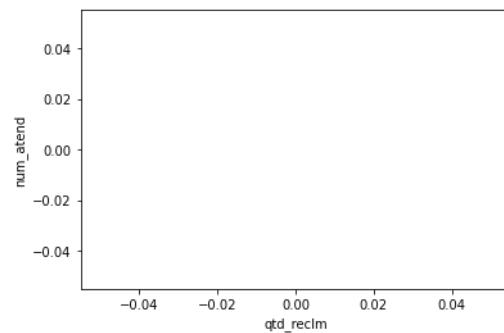
● (gráfico que compara número de atendimentos atrasados com número de atendimentos)



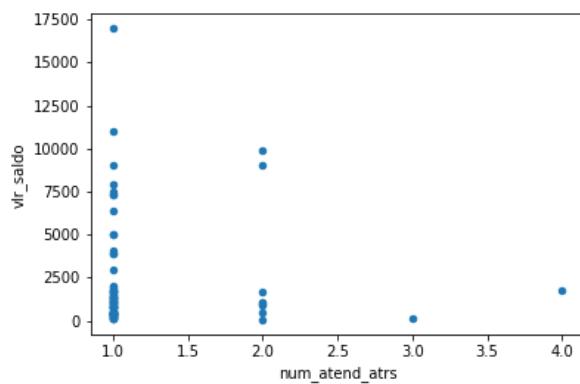
● (gráfico que compara quantidade de reclamações com quantidade de atendimentos atrasados)

(valor de score não sendo NaN)

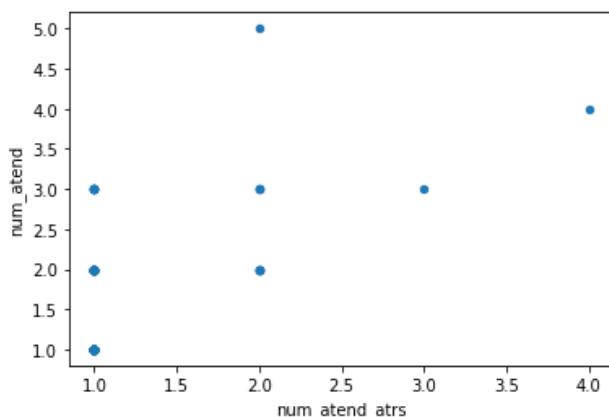
index	anomes	vir_credito	vir_saldo	num_atend_atrs	vir_score	num_produtos	num_atend	qtd_oper	qtd_reclm	qtd_restr
count	105956.0	105956.0	63124.0	80.0	85924.0	90459.0	322.0	105956.0	1.0	48744.0
mean	202131.564583412	160.4138913322511	2157.510717535321	1.175	458.28648573157676	1.419880829989277	1.34472049689441	3.6301011740722564	1.0	2.8301534547841785
std	40.941215255062815	92.69971368341983	5354.896866926304	0.49746191254243755	217.94213659034756	0.7318226698603035	0.6478722058823735	2.539435660680094	NaN	3.166753700575058
min	202104.0	0.01	0.01	1.0	70.0	1.0	1.0	1.0	1.0	1.0
25%	202107.0	73.4675	225.09	1.0	291.0	1.0	1.0	2.0	1.0	1.0
50%	202110.0	196.735	662.025	1.0	446.0	1.0	1.0	3.0	1.0	2.0
75%	202112.0	238.18	1741.159999999999	1.0	615.0	2.0	2.0	5.0	1.0	3.0
max	202204.0	287.4	175667.19	4.0	987.0	7.0	5.0	48.0	1.0	75.0



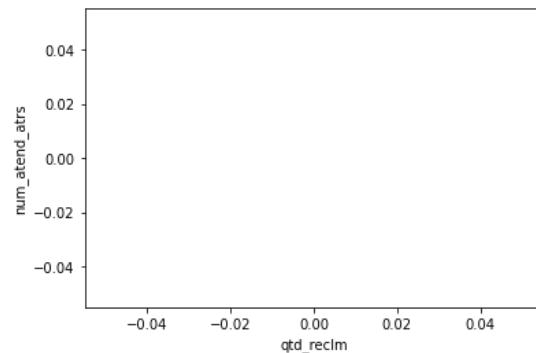
● (gráfico que compara quantidade de reclamações com número de atendimentos)



● (gráfico que compara número de atendimentos atrasados com valor de saldo)



● (gráfico que compara número de atendimentos atrasados com número de atendimentos. Percebemos que quando há mais atendimentos há maior número de atendimentos atrasados)



(gráfico que compara quantidade de reclamações com número de atendimentos atrasados)

Grupo de pessoas com crédito de R\$287 - R\$17.015:

(Descrição geral): número de atendimentos alto e reclamações baixo.

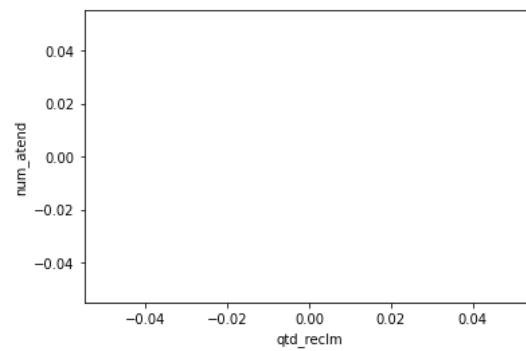
index	anomes	vlr_credito	vlr_saldo	num_atend_atrs	vlr_score	num_produtos	num_atend	qtd_oper	qtd_reclm	qtd_restr
count	3126313.0	3126313.0	2703966.0	3318.0	2777186.0	2968431.0	12639.0	3126313.0	297.0	1303050.0
mean	202133.52026556522	6914.187880276231	3106.3904806128426	1.1193490054249549	509.1257740749089	1.5867133175741663	1.3787483186960994	9.328448239187823	1.0	2.876296381566325
std	41.91250945447165	5099.008048848354	4272.137553826817	0.38227923233119376	214.52679668422394	0.9266044453851435	0.6875288941202252	7.128539870459864	0.0	3.16327492572853
min	202104.0	287.41	0.01	1.0	0.0	1.0	1.0	1.0	1.0	1.0
25%	202107.0	2264.32	729.259999999925	1.0	367.0	1.0	1.0	5.0	1.0	1.0
50%	202110.0	5821.75	1581.0999999999003	1.0	508.0	1.0	1.0	8.0	1.0	2.0
75%	202201.0	11203.15	3681.0	1.0	661.0	2.0	2.0	12.0	1.0	4.0
max	202204.0	17015.0	379388.07	6.0	1000.0	12.0	13.0	254.0	1.0	119.0

A	0.804981
H	0.049326
HH	0.038729
C	0.018639
AA	0.018126
B	0.017112
D	0.016028
E	0.013540
F	0.012101
G	0.011418

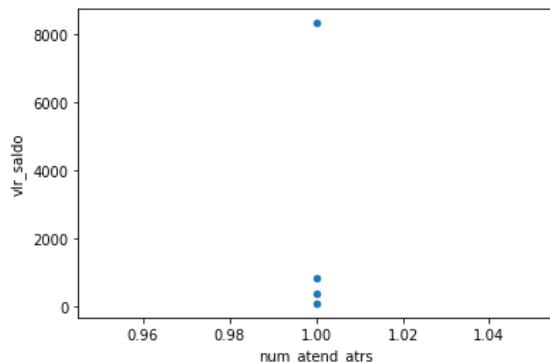
(percentil de rating dos clientes desse grupo. Comprovando a hipótese novamente que quando o banco tem o valor do crédito da pessoa ele confia mais nela)

(campo score é NaN)

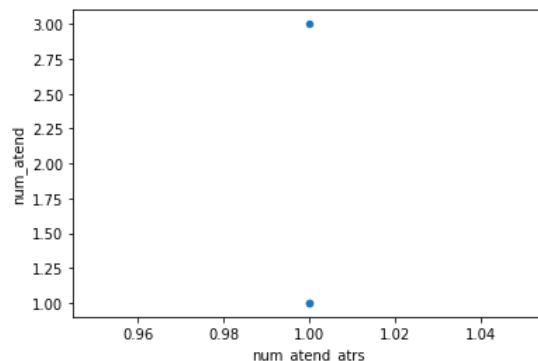
index	anomes	vlr_credito	vlr_saldo	num_atend_atrs	vlr_score	num_produtos	num_atend	qtd_oper	qtd_reclm	qtd_restr
count	1981.0	1981.0	1650.0	4.0	1981.0	1941.0	27.0	1981.0	2.0	593.0
mean	202202.60020191822	7751.590242301868	2471.0174787878786	1.0	0.0	2.5749613601236474	1.2592592592593	15.184755174154468	1.0	3.1973018549747048
std	1.0762260847382699	4710.0608979882645	3047.061775406662	0.0	0.0	1.6249255529500117	0.44657608470472243	9.621324684783797	0.0	3.2668264217552654
min	202201.0	291.72	5.9	1.0	0.0	1.0	1.0	1.0	1.0	1.0
25%	202202.0	3558.140000000003	700.0	1.0	0.0	1.0	1.0	8.0	1.0	1.0
50%	202203.0	7484.2	1455.48	1.0	0.0	2.0	1.0	13.0	1.0	2.0
75%	202204.0	11602.18	3083.467499999997	1.0	0.0	4.0	1.5	20.0	1.0	4.0
max	202204.0	16996.69000000002	27669.49	1.0	0.0	11.0	2.0	64.0	1.0	25.0



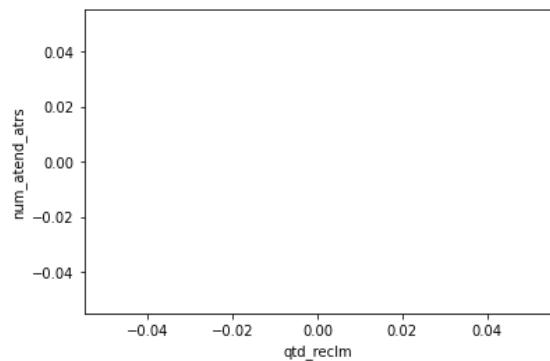
(gráfico que compara quantidade de reclamações com números de atendimentos)



(gráfico que compara número de atendimentos com valor de saldo)



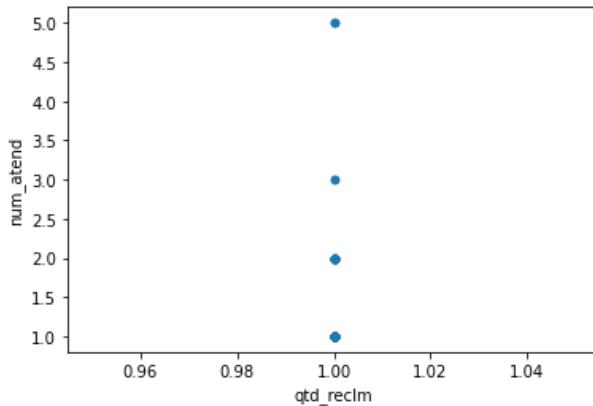
(gráfico que compara número de atendimentos atrasados com número de atendimentos)



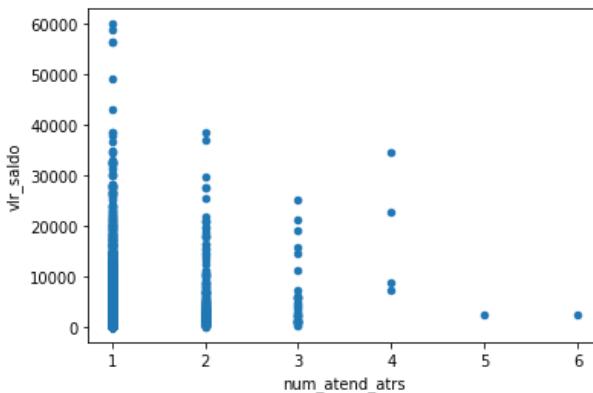
(gráfico que compara quantidade de reclamações com quantidade de atendimentos atrasados)

(campo score não é NaN)

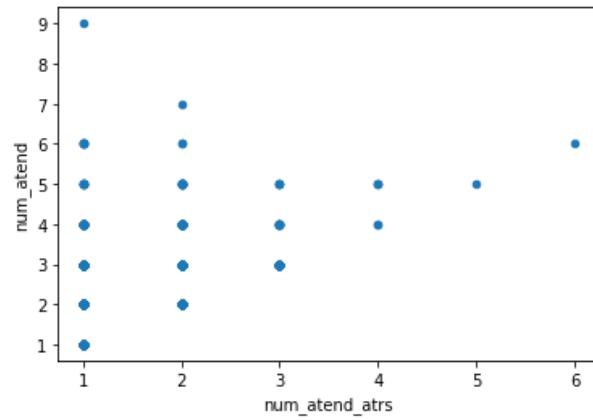
index	anomes	vir_credito	vir_saldo	num_atend_atrs	vir_score	num_produtos	num_atend	qtd_oper	qtd_reclm	qtd_restr
count	3124332.0	3124332.0	2702316.0	3314.0	2775205.0	2966490.0	12612.0	3124332.0	295.0	1302457.0
mean	202133.47646504917	6913.65692067296	3106.7784313384495	1.1194930597465298	509.4891988159433	1.586066983539469	1.3790041230574057	9.324735015356882	1.0	2.876150229911621
std	41.8866313774615	5099.20322851694	4272.749811219766	0.3824874298187135	214.17151247580347	0.9256302774873131	0.6879434710172982	7.125157998135824	0.0	3.163220883367518
min	202104.0	287.41	0.01	1.0	1.0	1.0	1.0	1.0	1.0	1.0
25%	202107.0	2263.52	729.289999999999	1.0	367.0	1.0	1.0	5.0	1.0	1.0
50%	202110.0	5820.64	1581.3000000000002	1.0	508.0	1.0	1.0	8.0	1.0	2.0
75%	202201.0	11202.8025	3681.909999999999	1.0	661.0	2.0	2.0	12.0	1.0	4.0
max	202204.0	17015.0	379388.07	6.0	1000.0	12.0	13.0	254.0	1.0	119.0



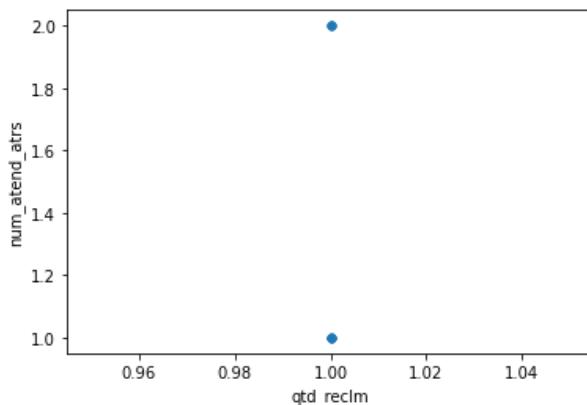
- (gráfico que compara quantidade de reclamações com números de atendimentos)



- (gráfico que compara número de atendimentos atrasados com valor de saldo. Quanto menor o saldo, maior o número de atendimentos atrasados.)



(gráfico que compara número de atendimentos atrasados com número de atendimentos)



(gráfico que compra a quantidade de reclamações com a quantidade de atendimentos atrasados)

Grupo de pessoas com crédito de R\$17.015 - R\$10.348.109:

(Descrição geral): número de atendimentos e reclamações alto.

index	anomes	vlr_credito	vlr_saldo	num_atend_atrs	vlr_score	num_produtos	num_atend	qtd_oper	qtd_reclm	qtd_restr
count	3126329.0	3126329.0	3015075.0	2775.0	2823680.0	3029464.0	11310.0	3126329.0	946.0	1495972.0
mean	202134.02437299467	65205.29426275353	8391.423810868058	1.1192792792792792	533.2096512352674	1.6852644560225836	1.3829354553492486	17.231405907695574	1.0010570824524312	3.2827278852812753
std	42.12832344054365	87490.98506856868	15042.740422566463	0.400755666238859	209.13194209669254	1.0387993565230966	0.7240366490296383	11.113538063713074	0.03251280443811773	3.7423576801039133
min	202104.0	17015.01	0.01	1.0	0.0	1.0	1.0	0.0	1.0	1.0
25%	202107.0	24153.04000000005	1560.0	1.0	382.0	1.0	1.0	10.0	1.0	1.0
50%	202110.0	38399.37	3833.79	1.0	504.0	1.0	1.0	15.0	1.0	2.0
75%	202201.0	74742.8700000001	10709.0599999999	1.0	678.0	2.0	2.0	21.0	1.0	4.0
max	202204.0	10273267.56999998	1777232.21	7.0	1000.0	15.0	17.0	306.0	2.0	181.0

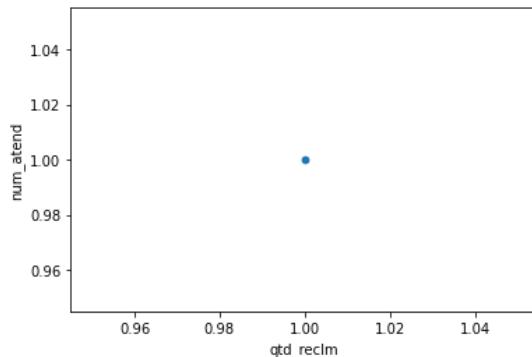
A	0.877897
H	0.025387
HH	0.023330
B	0.017610
C	0.016043
D	0.011872
E	0.008794
F	0.007461
G	0.006524
AA	0.005082

(percentil de cada "rating" desse grupo.

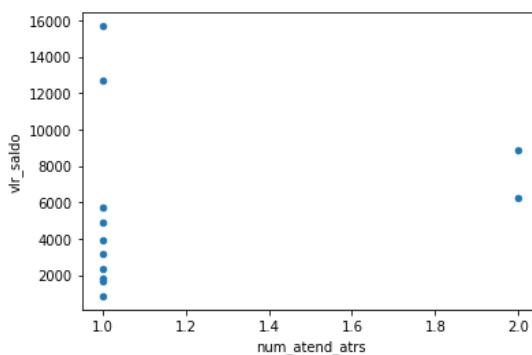
Comprovando a hipótese novamente que quando o banco sabe a quantia de crédito do cliente ele confia mais nele)

(score é NaN)

index	anomes	vlr_credito	vlr_saldo	num_atend_atrs	vlr_score	num_produtos	num_atend	qtd_oper	qtd_reclm	qtd_restr
count	4269.0	4269.0	4072.0	12.0	4269.0	4244.0	48.0	4269.0	9.0	1477.0
mean	202202.62965565707	127665.02368001873	7379.100550098232	1.1666666666666667	0.0	2.8541470311027335	1.25	30.829936753338018	1.0	5.246445497630332
std	1.0859809564289364	206780.800673001	13879.915180842372	0.38924947208076155	0.0	1.786950696300637	0.4837794468468967	18.63006492577131	0.0	5.3907160531984805
min	202201.0	17031.17	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0
25%	202202.0	33967.42	1400.0	1.0	0.0	1.0	1.0	18.0	1.0	1.0
50%	202203.0	69719.13999999998	3149.405	1.0	0.0	3.0	1.0	27.0	1.0	3.0
75%	202204.0	141737.99000000002	8274.64	1.0	0.0	4.0	1.0	39.0	1.0	7.0
max	202204.0	6694688.77	467111.08	2.0	0.0	13.0	3.0	199.0	1.0	53.0

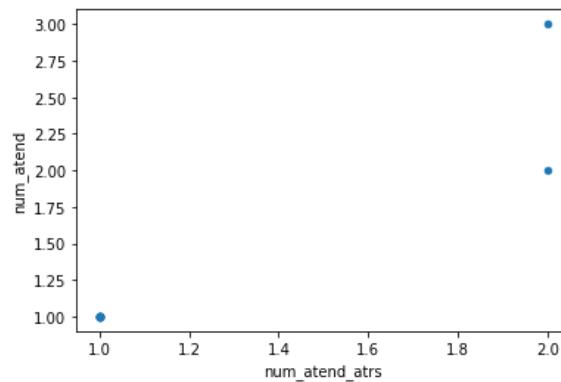


(gráfico que compara quantidade de reclamações com número de atendimentos)

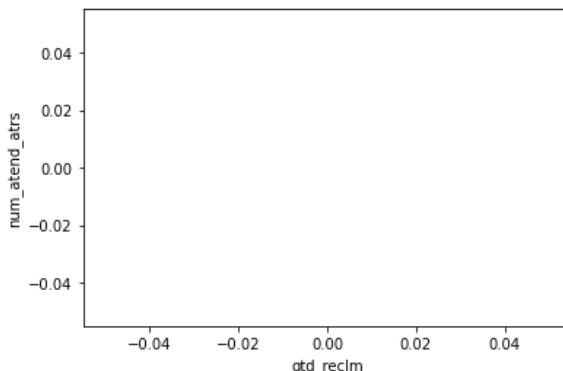


(gráfico que compara número de atendimentos atrasados com valor de saldo.

Quanto menor o saldo, maior o número de atendimentos atrasados)



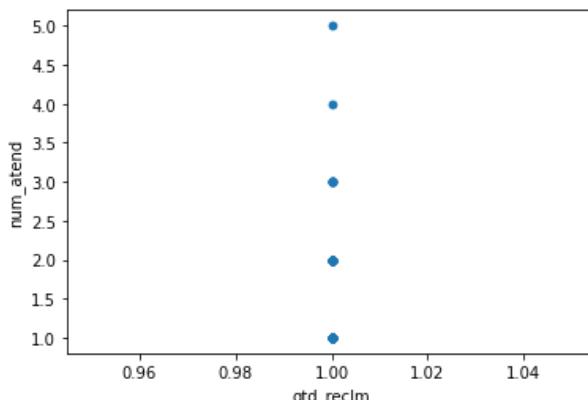
(gráfico que compara número de atendimentos atrasados com número de atendimentos)



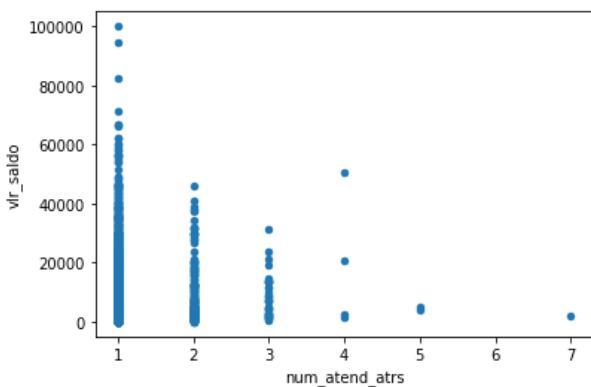
(gráfico que compara a quantidade de reclamações com a quantidade de atendimentos atrasados)

(score não é NaN)

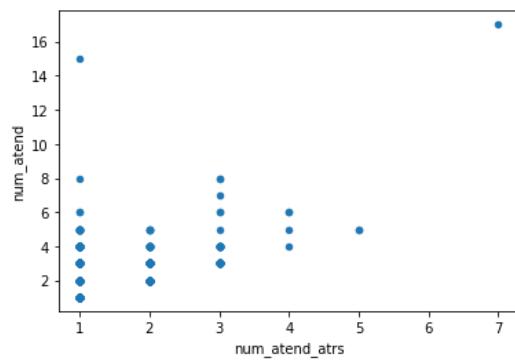
index	anomes	vlr_credito	vlr_saldo	num_atend_atrs	vlr_score	num_produtos	num_atend	qtd_oper	qtd_reclm	qtd_restr
count	3122060.0	3122060.0	3011003.0	2783.0	2819411.0	3025220.0	11262.0	3122060.0	937.0	1494495.0
mean	202133.93056443502	65119.88892625068	8392.792849795569	1.1190734708650019	534.0170085170272	1.6836246620080524	1.3835020422660274	17.21281173327867	1.0010672358591248	3.2807871555274524
std	42.08059238589465	87185.69227399654	15044.208142798268	0.4008611422340094	208.25765172662452	1.036445822572741	0.7248524850433573	11.088369578238085	0.03266857601924007	3.7398619334087004
min	202104.0	17015.01	0.01	1.0	51.0	1.0	1.0	0.0	1.0	1.0
25%	202107.0	24145.43000000004	1560.17	1.0	383.0	1.0	1.0	10.0	1.0	1.0
50%	202110.0	38376.88	3834.96999999998	1.0	504.0	1.0	1.0	15.0	1.0	2.0
75%	202201.0	74665.48999999999	10712.1699999998	1.0	678.0	2.0	2.0	21.0	1.0	4.0
max	202204.0	10273267.569999998	1777232.21	7.0	1000.0	15.0	17.0	306.0	2.0	181.0



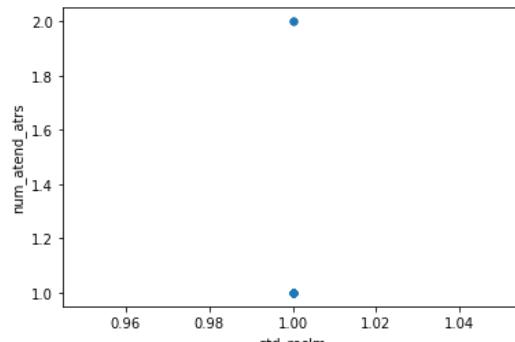
(gráfico que compara quantidade de reclamações com número de atendimentos)



(gráfico que compara número de atendimentos atrasados com valor de saldo.
Quanto menor seu saldo, maior o número de atendimentos atrasados)



(gráfico que compara número de atendimentos atrasados com número de atendimentos)



(gráfico que compara a quantidade de reclamações com a quantidade de atendimentos atrasados)

Conclusões preliminares:

Percebemos que quando o banco sabe o score da pessoa normalmente essa pessoa tem um maior número de atendimentos, um maior número de atendimentos atrasados e um maior número de reclamações (hipótese: essas pessoas interagem mais com o banco). Também percebemos que quando uma pessoa reclama, ela reclama somente uma vez (hipótese: a pessoa tenta vários contatos com o banco e oficializa a reclamação apenas em último caso). Percebemos que as pessoas que possuem crédito acima de R\$287, que representam 50% dos clientes, possuem quase que a totalidade de atendimentos do banco

(hipótese: banco oferece mais produtos para essas pessoas logo elas interagem mais com o banco). Percebemos que quanto maior o crédito da pessoa maior o número de atendimentos e reclamações. Percebemos que não há relação direta com maior número de atendimentos com atendimentos atrasados (hipótese: atendimentos atrasados acontecem em detrimento da complexidade dos problemas), mas identificamos que só há reclamação quando há atendimento atrasado.

Em uma segunda análise o dataset foi dividido em três subdatasets, foram eles:

1. Clientes que em algum momento já foram engajados com o Banco
2. Clientes que em algum momento já tiveram um atrito
3. Possíveis novos clientes

Esta análise estatística irá focar no primeiro subconjunto dos dados, de clientes engajados.

Aplicando uma descrição geral aos dados, temos o seguinte resultado:

	anomes	vlr_credito	vlr saldo	num_atend_atrs	vlr_score	num_produtos	num_atend	qtd_oper	qtd_reclm	qtd_restr	ind_atrito	ind_engaj	ind_novo_cli
count	3338100.00	3338100.00	3338100.00	3338100.00	3338100.00	3338100.00	3338100.00	3338100.00	3338100.00	3338100.00	3338100.00	3338100.00	3338100.00
mean	202137.03	36492.18	5974.46	0.00	504.14	1.90	0.01	15.04	0.00	1.28	0.00	0.65	0.00
std	43.66	68057.77	9905.07	0.05	231.92	1.18	0.13	11.23	0.01	2.76	0.03	0.48	0.06
min	202104.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	202107.00	5804.41	1148.77	0.00	371.00	1.00	0.00	8.00	0.00	0.00	0.00	0.00	0.00
50%	202110.00	17557.14	2529.33	0.00	504.00	2.00	0.00	13.00	0.00	0.00	0.00	1.00	0.00
75%	202201.00	39501.43	6729.64	0.00	665.00	3.00	0.00	19.00	0.00	1.00	0.00	1.00	0.00
max	202204.00	6718974.54	791074.25	7.00	1000.00	15.00	17.00	306.00	2.00	119.00	1.00	1.00	1.00

Em relação ao dataset original, os clientes que são ou já foram engajados com o banco representam cerca de 27%.

Verificando o **cod_rating** dos clientes neste novo dataframe, temos a seguinte frequência:

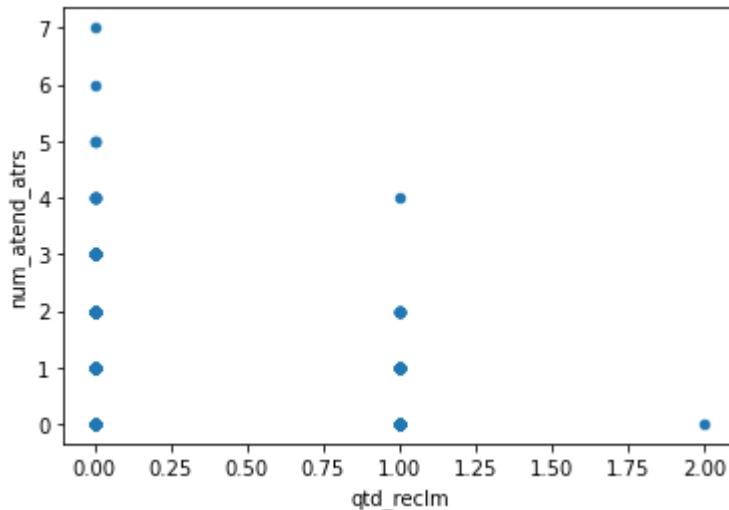
```

A    0.86
H    0.03
0    0.02
B    0.02
C    0.02
D    0.01
E    0.01
F    0.01
G    0.01
AA   0.01
HH   0.00
Name: cod_rating, dtype: float64

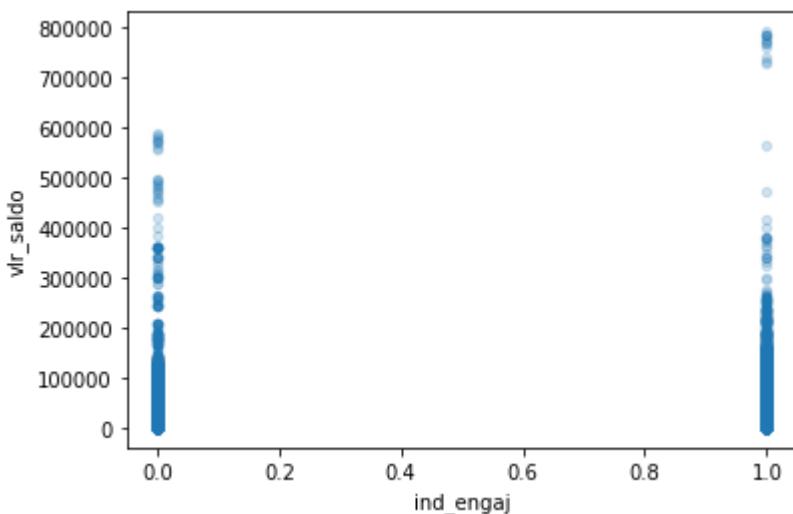
```

Podemos estabelecer uma relação entre o rating e o engajamento que ele possui, visto que 86% dos clientes engajados possuem classificação A em comparação aos 74% que são HH (*Análise anterior, em relação ao dataset por completo*).

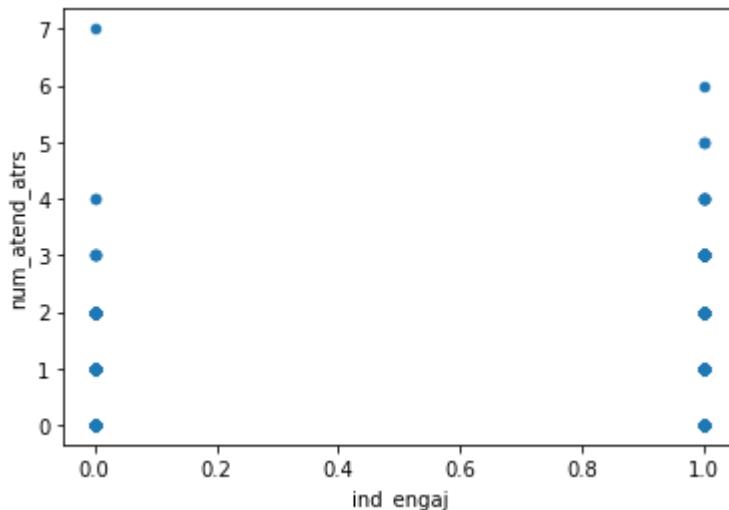
Podemos re-aplicar alguns dos gráficos da primeira análise a partir desse novo subdataset, como a comparação entre quantidade de reclamações e atendimentos em atraso:



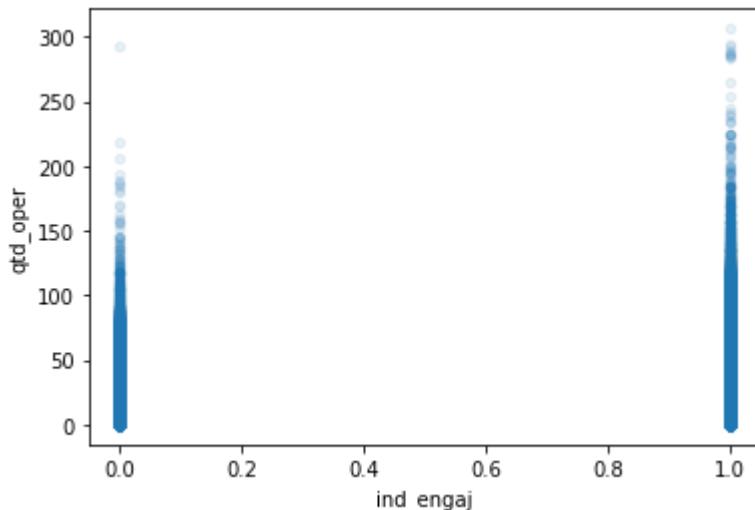
Relação de valor de saldo com o índice de engajamento:



Relação entre número de atendimentos atrasados e engajamento:



Relação entre quantidade de operações e engajamento:



3. Desejamos obter três novas colunas. A primeira, indicará se o cliente possui atrito, ou não, com o banco; a segunda, indicará o engajamento do cliente, como o número de produtos, quantidade de operações, e a terceira será um indicador se o usuário é um possível novo cliente. Todos estes novos campos, são de natureza **discreta**, pois apenas assumirão dois tipos de valores: SIM ou NÃO.

4.3. Preparação dos Dados

A priori, focaremos na descrição do subdataset que tem como parâmetro o engajamento do cliente, como já citado na seção anterior. Para esta divisão, foi utilizada a coluna **ind_engaj**. Foi gerada uma lista apenas com os CPFs cujos valores já se igualaram a 1 em algum momento dentre as safras. A partir dessa lista, filtramos o dataset original com os dados dos clientes que estivessem na lista.

Três colunas foram removidas, visto que não seriam necessárias para nosso modelo. Foram elas: **vlr_renda**, **ind_novo_cli** e **anomes**. O valor de renda foi removido, pois é um valor preditivo de mercado e não necessariamente agrupa e nem traz novos insights para o modelo. O índice de novos clientes foi removido, uma vez que não há sentido em manter uma coluna que indique possíveis novos clientes se o subconjunto foi gerado apenas com clientes. A coluna **anomes**, foi removida por se tratar de um registro de datas que não são mais relevantes para o modelo.

Para o aprendizado de máquina é necessário unirmos as informações contidas nestas safras para a obtenção de um modelo preditivo eficiente. Contudo, visando abranger diferentes cenários optamos por dividir esse conjunto de dados de 3 maneiras de agregação:

1. Foi utilizado o somatório visando manter o registro histórico do cliente. Este método foi aplicado nas seguintes features: qtd_oper, qtd_reclm, num_atend_atrs, num_atend e qtd_restritivo.
2. Uma maneira de agregação foi a utilização apenas da safra mês 11. O motivo do mês 11 em específico é dado a quantidade de informações presentes nesta safra. Além de ter sido uma recomendação do próprio parceiro.
3. Outra maneira de agregação foi a utilização da mediana, invés do somatório como está descrito na tabela.

Logo, dado as diferentes dimensões e validades das colunas, foram utilizadas 4 formas de agregação: mediana, valor máximo, média e label encoding.

<u>Nome da feature</u>	<u>Método Utilizado</u>	<u>Motivo</u>
vlr_credito	Mediana	Para estas features, foi utilizado o método da mediana, visando a eliminação de grandes desvios nos dados e outliers.
vlr_saldo		
num_produtos		
ind_engaj		
qtd_oper		
qtd_reclm		

num_atend_atrs		
num_atend		
qtd_restritivo		
ind_atrito	Valor Máximo	Este índice aceita apenas os valores 0 ou 1. Ao registrar o maior valor, temos o registro histórico se em algum momento este índice foi verdadeiro para o cliente.
vlr_score	Média	Utilizando a média para o score cobrimos eventuais variações.
cod_rating	Label Encoding e Mediana	Dado que o rating apresenta uma diferença de relevância entre AA e HH, utilizamos este método por ser mais eficiente. Após o Label Encoding calculamos a mediana dos números visando registrar seu valor. Utilizamos a mediana para evitar grandes diferenças entre os valores. Devido a existência de um valor médio entre 0 e 1, houve uma manipulação para os valores iguais a 0.5, que foram arredondados para 1 ¹

¹ Esta decisão foi tomada pois o número de clientes com o **cod_rating** igual a 0.5, abrangem clientes que foram 50% do tempo engajados com o banco. Esta parcela representa apenas de 1% e ao arredondarmos para 1 teríamos uma maior amostra (totalizando cerca de 65% de clientes engajados), trazendo benefícios para o modelo e seus resultados.

Para o modelo preditivo, é preciso substituir os valores nulos por numéricos. Todos os valores em branco (NaN, Null, etc) no dataset foram preenchidos com zero. Substituímos por zero, pois compreendemos que é o valor menos impactante para o resultado final, visto que por não terem sido preenchidos representam campos inexistentes ou de fato iguais a zero.

No entanto, devido a um erro de mineração dos dados, alguns clientes possuem valor nulo para **num_produtos**, mesmo possuindo valores de saldo e classificação de risco no sistema. O mesmo ocorre para **cod_rating**. Para estas features, foi decidido junto ao cliente que seriam classificados como outliers e deveriam ser removidos da base de dados.

Em um momento inicial, todas as features do dataset original foram utilizadas, exceto a **vlr_renda**, **ind_novo_cli** e **anomes** como citado anteriormente. De maneira geral, todos as features selecionadas permitem estabelecer um paralelo com o engajamento, e compreender como determinado índice o afeta. A seguir, está uma relação das colunas utilizadas e o que estão agregando ao projeto.

Nome da feature/coluna	Motivo de uso
Número de CPF (num_cpf_hash)	Em uma primeira análise, o número de cpf nos permite, e garante, a manipulação e agregação dos dados em uma só linha. <i>**Durante a modelagem esta coluna será removida visando a maior eficiência na predição.</i>
Valor de crédito no mercado (vlr_credito)	Ambas as colunas estão relacionadas a crédito do cliente e sua classificação como pagador, seja no mercado ou dentro do Banco Pan. Estas features nos permitirão compreender a relação entre engajamento e os serviços de crédito como um todo.
Valor do Crédito oferecido pelo banco (vlr_saldo)	<i>**Apesar de terem sido selecionadas por uma mesma lógica, será também explorado e utilizado as características individuais destes indicadores.</i>
Valor do Score (vlr_score)	
Número de atendimentos (num_atend)	Estes índices funcionam como complemento um do outro, e são um dos fortes indicadores de atrito do cliente com os serviços prestados. Permite mapear e estabelecer correlações entre o atendimento prestado e o quanto engajado.
Número de atendimentos atrasados (num_atend_atrs)	

Quantidade de Operações realizadas <i>(qtd_oper)</i>	Ao se tratar da quantidade de operações que o cliente efetua em sua conta, torna-se um dos mais fortes índices de engajamento que temos. A partir dele, podemos traçar diferentes paralelos entre engajamento e quantidade de operações. Permite também, indicar eventuais melhorias de atendimento e relação com o banco, a partir das outras features citadas.
Número de produtos (<i>num_produtos</i>)	Similar ao motivo anterior, a partir deste índice é possível estabelecer uma relação entre uma determinada faixa quantitativa de produtos e como isto espelha o engajamento do cliente.
Quantidade de reclamações abertas <i>(qtd_reclm)</i>	Maior indicador de atrito possível, uma vez que se trata de reclamações abertas em plataformas como o Bacen e Procon. Essencial para reforçar o entendimento da relação atrito e engajamento.
Quantidade de restritivos (<i>qtd_restr</i>)	Ambas features nos permitem compreender como o engajamento é afetado a partir do risco do cliente para o banco (cod_rating) e de eventuais dívidas de crédito que o cliente tenha (qtd_restr).
Classificação do cliente com o banco <i>(cod_rating)</i>	Como reforçado durante todo a seção, procuramos compreender da maneira mais exata possível como os demais fatores refletem o engajamento do cliente com o banco. Portanto, é de suma importância utilizarmos também os índices provenientes do dataset para entender se há atrito explícito (ind_atrito) e se o cliente é classificado como engajado.
Índice de Engajamento (<i>ind_engaj</i>)	Também se trata de uma variável resposta para um dos modelos propostos.

4.4. Modelagem

O modelo utilizado

Inicialmente foi efetuado uma avaliação genérica para cada um dos sub datasets com quatro algoritmos de machine learning : **GaussianDB**, **DecisionTreeClassifier**, **RandomForestClassifier** e **MLPClassifier**, utilizando o algoritmo de avaliação **F1 Score**² com um *cross validation* (método utilizado para avaliação de desempenho que consiste no particionamento dos dados) separando nosso dataset em 6 partes.

```
[{'modelo': 'GaussianNB',
'média': 0.7793324833739493,
'desvio-padrão': 0.0006427643735270461},
{'modelo': 'DecisionTreeClassifier',
'média': 0.7525589857098599,
'desvio-padrão': 0.0015505064944041338},
{'modelo': 'RandomForestClassifier',
'média': 0.820000761392691,
'desvio-padrão': 0.0010984693860869804},
{'modelo': 'MLPClassifier',
'média': 0.8001892330467353,
'desvio-padrão': 0.00782797361619285}]
```

Com base na média dos *cross validation*, escolhemos o **RandomForestClassifier**. Em síntese este modelo consiste em estabelecer um conjunto de "n" árvores de decisão a fim de reduzir o viés das árvores, a partir da compilação de diferentes árvores de decisões para um mesmo resultado. Esse modelo tem como base o *DecisionTreeClassifier*, que são justamente as "n" árvores de decisões geradas pelo modelo anteriormente citado.

Esse então consiste na elaboração de perguntas para "n" conjuntos de dados separados pelo próprio algoritmo, e que as respostas a essas perguntas cria nós (conjunto de dados para responder uma pergunta) e ramos (caminho a ser seguido após a resposta). A partir desses nós e ramos que a classificação é efetuada ao final de toda a análise do conjunto de dados. A preparação dessas perguntas podem seguir alguns padrões, o utilizado em nosso modelo é o *GINI*.

Essa análise consiste no cálculo do índice GINI, que verificará a oscilação dos dados nas variáveis preditoras em relação a variação da variável target. A partir disso a variável preditora que possuir o menor índice que acabou de ser calculado, será escolhida para o nó principal da árvore, pois um baixo valor do índice indica uma melhor consistência na distribuição dos dados. E assim uma árvore é criada no modelo de RandomForest, e seguirá para todas as outras que

² O F1-Score é uma média harmônica unindo as duas métricas de precisão e revocação.

serão geradas. Quando for necessário a predição, o resultado de várias árvores servirão como base para que seja obtido uma predição de classificação.

Subdatasets

O algoritmo foi aplicado aos três sub datasets descritos na seção 4.3.
Preparação dos dados.

Target: Índice de engajamento

1º Conjunto de Dados: Agregação somatória

Modelo 1: Random Forest Classifier

O primeiro passo foi a divisão do dataset nas variáveis resposta e nas variáveis preditoras.

Como explicitado anteriormente, para este modelo, foi definida a coluna de índice de engajamento como a variável resposta/dependente, a todas as demais colunas como variáveis independentes/preditoras.

	importancia	
vlr_saldo	0.262200	Antes de um modelo inicial foi realizado uma <i>feature importance</i> com o algoritmo RFC (Random Forest Classifier) para validar a importância de cada feature para nosso modelo. (<i>Imagen ao lado</i>)
qtd_oper	0.189827	
vlr_credito	0.184370	
vlr_score	0.172712	Aproximadamente 90% do nosso modelo é respondido a partir das colunas referentes a crédito, operação e restritivos. A quantidade de produtos também exerce uma significativa influência, com cerca de 6,14%.
qtd_restr	0.080253	
num_produtos	0.061471	
cod_rating	0.036231	
num_atend	0.009060	
num_atend_atrs	0.001866	
ind_atrito	0.001440	
qtd_reclm	0.000569	

Antes de aplicarmos o modelo, testamos quais seriam os melhores parâmetros para serem adicionados visando a melhor modelagem possível. Portanto, foram realizadas uma série de testes utilizando Grid Search³ obtendo o seguinte resultado:

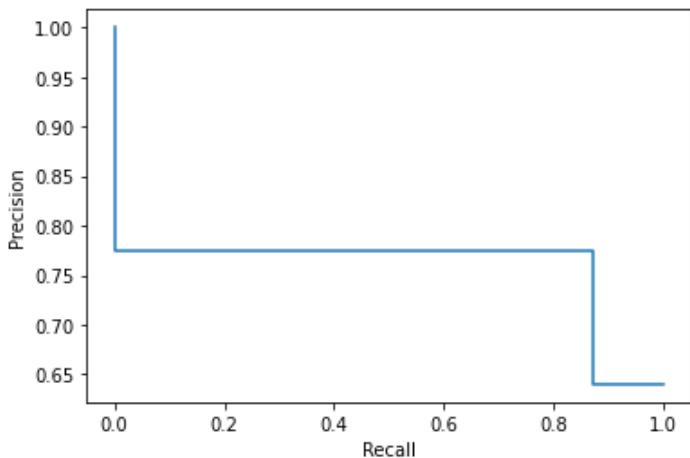
- bootstrap: True
- max_depth=None
- min_samples_leaf=8
- min_samples_split=2
- n_estimators=80
- n_jobs=-1
- random_state=42

³ Para mais detalhes dos hiperparâmetros citados, segue a referência da documentação do ScikitLearn.
 <<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>>

Foi alocado 33% do conjunto de dados para a divisão de treino.
Utilizando o F1-Score para calcular a performance do modelo obtemos uma média igual a: 0.82. A seguir há uma análise dos resultados utilizando: Precision x Recall, Curva ROC e Matriz de Confusão.

Precision x Recall Plot

Maneira de visualizar a relação e valores entre os avaliadores Precisão e Revocação



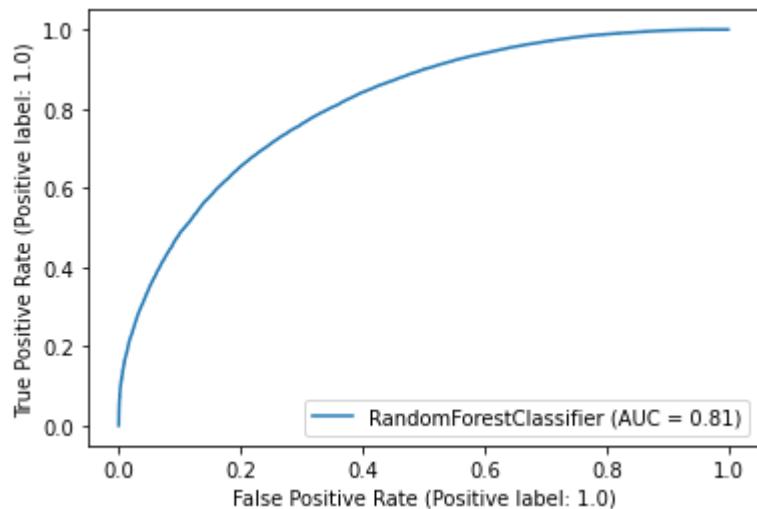
Classification Report

Relatório visando analisar a qualidade do algoritmo aplicado.

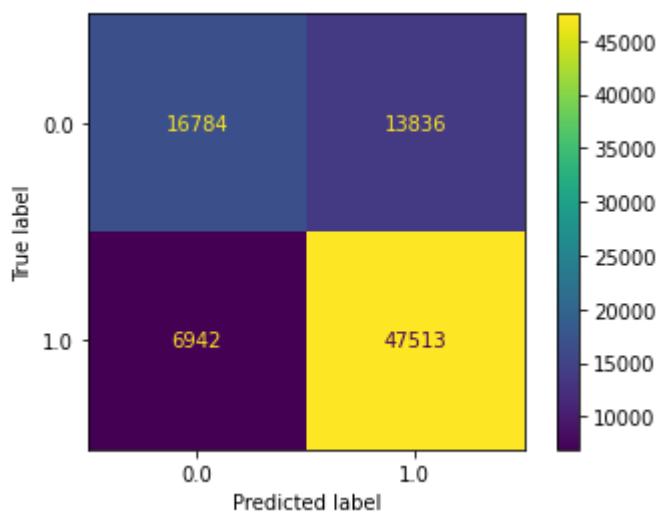
	precision	recall	f1-score	support
0.0	0.71	0.55	0.62	30620
1.0	0.77	0.87	0.82	54455
accuracy			0.76	85075
macro avg	0.74	0.71	0.72	85075
weighted avg	0.75	0.76	0.75	85075

Curva ROC

Relação entre os falsos negativos e falsos positivos do teste.



Matriz De Confusão



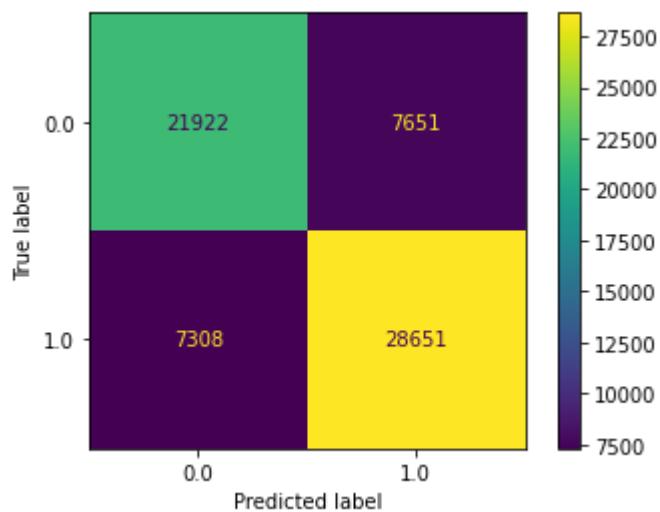
2º Conjunto de Dados: Dataset a partir da safra do mês 11 (Novembro de 2021)

Modelo 1: Random Forest Classifier

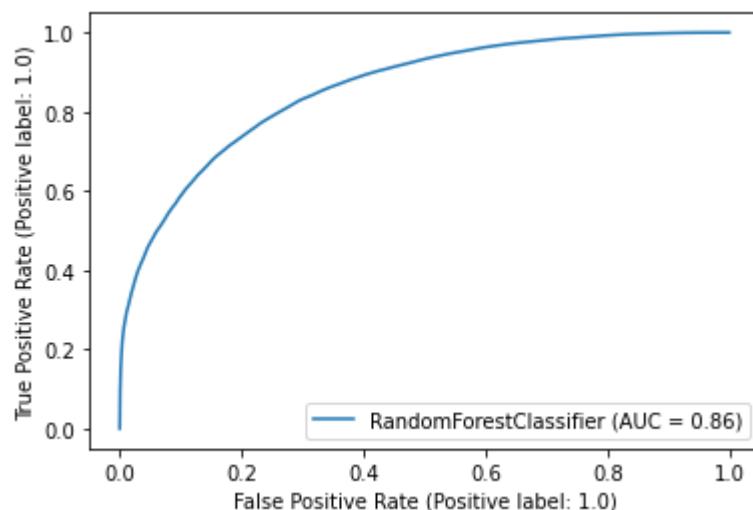
Inicialmente, tratamos os dados criando um data frame com somente clientes do mês 11. Foi feito um *label encoding* na coluna **cod_rating** já que ela é um dado categórico. Substituímos valores em branco por zero nas colunas: **num_atend_atrs**, **num_atend**, **qtd_oper**, **qtd_reclm**, **qtd_restr**, **ind_engaj**.

Além disso, removemos as linhas que tinham valor na nas colunas: **num_produtos**, **vlr_credito**, **vlr saldo**, **vlr_score**. Para finalizar tiramos as colunas **ind_atritado** e **ind_novo_cliente** por serem colunas de resposta.

O modelo utilizado foi o RandomForest classifier e uma cross validation de métrica de scoring f1, que é a média harmônica da acurácia e recall fazendo que o modelo seja balanceado tentando ter proporções similares de verdadeiros positivos e verdadeiros negativos, retornou: [0.79394022 0.794261 0.78859268 0.79290508 0.79443635].



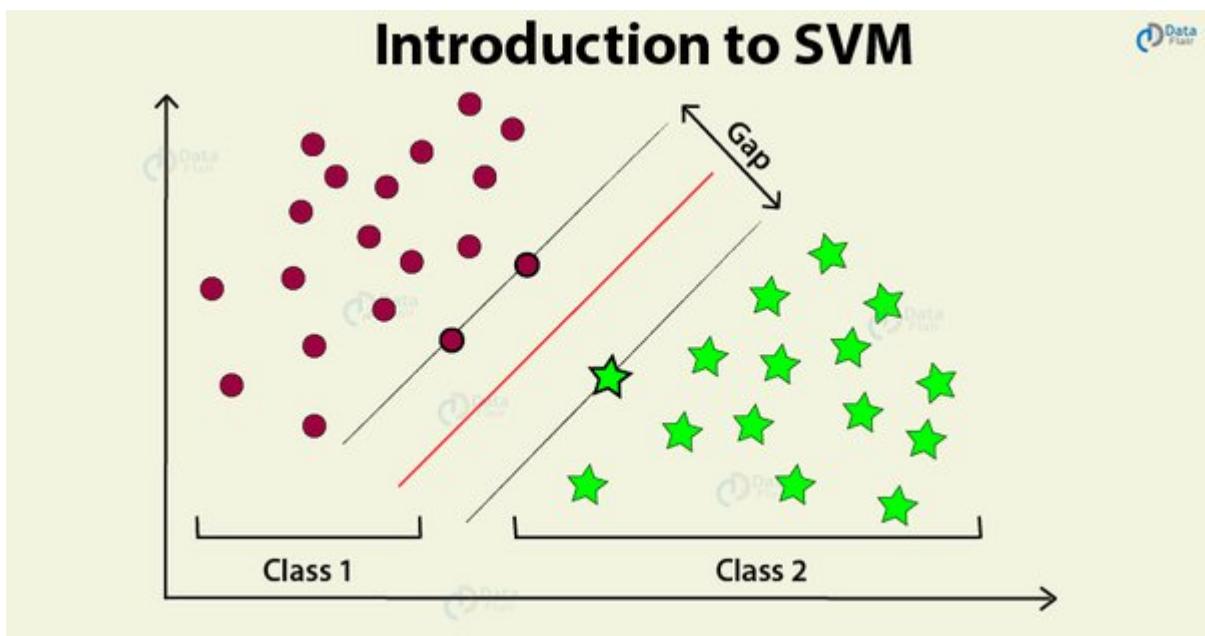
Podemos observar pela matriz de confusão acima que o modelo tende a ter falsos positivos e falsos negativos em proporção similar. Reforçando essas informações, a acurácia é 0.78844 o recall foi de 0.79484, ou seja o modelo tem um desempenho de quase 80 por cento.



Acima temos a curva roc do modelo que confirma ainda há espaço significativo de melhoria da precisão'

Modelo 2: SVM

Trata-se de um modelo supervisionado de classificação e regressão, onde é mapeado o conjunto de dados visando traçar uma divisão onde o dado pode ser categorizado; como na imagem abaixo:



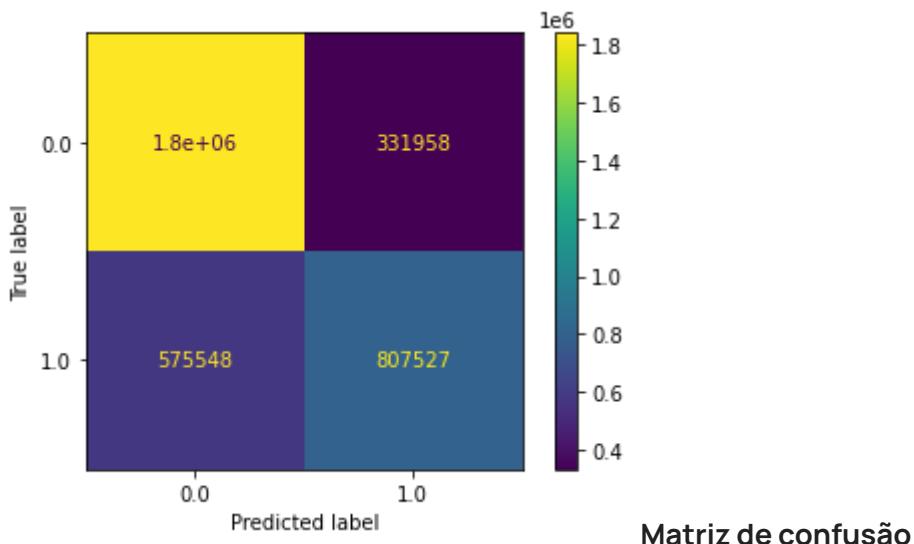
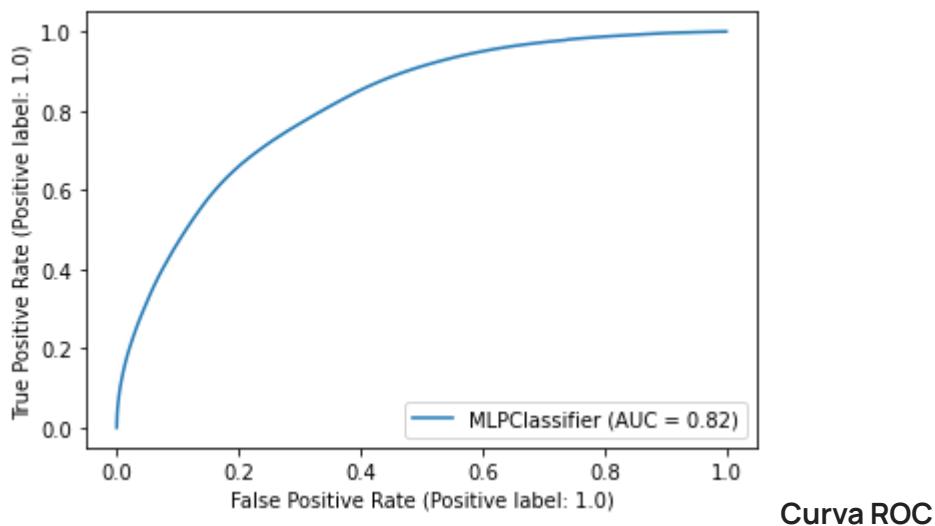
Utilizar o SVM é vantajoso, visto que o modelo trabalha muito bem com dados dissociados de maneira irregular.

Devido ao uso de processamento, não foi possível obter resultados válidos, com este modelo, mesmo com todas as otimizações possíveis.

Modelo 3: MLP Classifier

O Multi-Layer Perceptron Classifier, ou MLP Classifier, é um algoritmo de machine learning baseado em rede neural. (Explicar o funcionamento do MLP).

Os resultados obtidos com este modelo foram os seguintes:



Avaliando o modelo utilizando precisão, revocação e f1 score temos os seguintes resultados:

- Precisão: 0.58
- Revocação: 0.70
- F1 Score: 0.63

3º Conjunto de Dados: Dataset de clientes engajados (Utilizando o método da mediana para todas as colunas)

Modelo 1: Random Forest Classifier

A primeira coisa feita foi a implementação de um recurso importante com algoritmo de Random Forest Classifier para definir quais delas têm uma influência maior sobre o modelo.

Feito isso, percebemos que mais de 90% da importância estava alocada em 6 colunas.

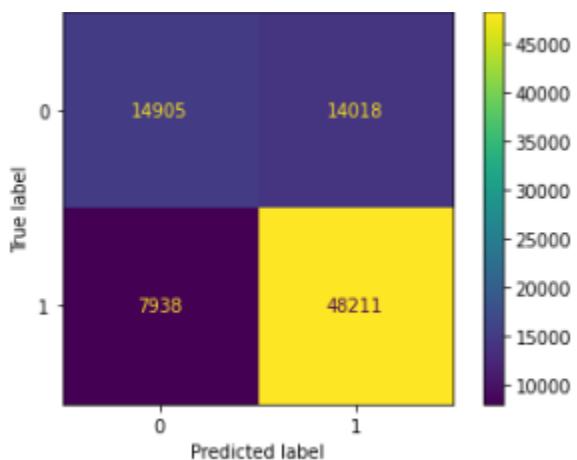
Logo, executamos nosso modelo apenas com as seis primeiras colunas para o modelo.

Alocamos 33% do conjunto de dados para a divisão de treino. Depois disso utilizamos F1-Score para calcular a performance do modelo obtemos uma média igual a: 0.814.

Em relação ao Recall tivemos uma média de 0.85 e Precisão de 0.77.

	importancia
vlr_saldo	0.290174
vlr_credito	0.218380
vlr_score	0.206166
qtd_oper	0.131967
num_produtos	0.056805
restr	0.033759
cod_rating	0.032269
qtd_restr	0.028842
ind_atrito	0.001593
num_atend	0.000036
num_atend_atrs	0.000007
qtd_reclm	0.000003

Abaixo, segue uma análise dos resultados utilizando a Matriz de Confusão.



4.5. Avaliação

A partir dos modelos gerados e suas respectivas avaliações na seção anterior, concluímos que o modelo utilizando o algoritmo RandomForestClassifier, gerado a partir do segundo conjunto de dados apresenta maior eficiência para o índice de engajamento.

[Para a próxima sprint, preencher essa seção a partir das validações com o parceiro]

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.

Backlog: ToDo List e Perguntas

ToDo

- Normalizar os dados
- Adicionar gráficos
- Check da influência das colunas X nas colunas Y com SqLearn
- Modelagem dos Dados
- Teste dos modelos preditivos
-

Perguntas

- Número de produtos como 0, mesmo possuindo saldo
R.: Deletar esse individuos
- Preenchimento dos dados nulos por 0
- Escolha das colunas

Anexos

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