

**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 |
| 26/08/2022 | Gabriel Carneiro | 1.3 | Documentação de todo o processo de feature engineering |
| 09/09/2022 | Gabriel Carneiro, Henrique Lemos, Luiz Alencar, Alberto Miranda | 1.4 | Documentação dos modelos gerados na sprint 3 |
| 23/09/2022 | Antônio Ribeiro, Gabriel Carneiro | 1.5 | Comparação dos modelos gerados |
| 04/10/2022 | Antônio Ribeiro, Gabriel Carneiro | 1.6 | Avaliação do modelo escolhido |
| 07/10/2022 | Gabriel Carneiro | 2.0 | Revisão final do documento.  Conclusão das seções 5 e 6 |

**Sumário**

[**1. Introdução**](#_2et92p0) **6**

[**2. Objetivos e Justificativa**](#_tyjcwt) **7**

[2.1. Objetivos](#_3dy6vkm) 7

[2.2. Proposta de Solução](#_1t3h5sf) 7

[2.3. Justificativa](#_4d34og8) 7

[**3. Metodologia**](#_2s8eyo1) **8**

[3.1. CRISP-DM](#_17dp8vu) 8

[3.2. Ferramentas](#_3rdcrjn) 9

[**3.3. Principais técnicas empregadas**](#_ca7ysvms7hyj) **9**

[**4. Desenvolvimento e Resultados**](#_lnxbz9) **10**

[4.1. Compreensão do Problema](#_35nkun2) 10

[4.1.1. Contexto da indústria](#_1ksv4uv) 10

[**4.1.2. Análise SWOT**](#_3ycc0pc5rwq4) **12**

[4.1.3. Planejamento Geral da Solução](#_2jxsxqh) 12

[4.1.4. Value Proposition Canvas](#_z337ya) 14

[4.1.5. Matriz de Riscos](#_3j2qqm3) 15

[**4.1.6. Personas**](#_1dw2arujn33f) **16**

[**4.1.7. Jornadas do Usuário**](#_o8fdhln6ttqp) **17**

[4.2. Compreensão dos Dados](#_2xcytpi) 18

[**4.3. Preparação dos Dados**](#_i0rc4xf3ced4) **36**

[**Número de CPF (num\_cpf\_hash)**](#_za7rwma1stgf) **38**

[**Valor de crédito no mercado (vlr\_credito)**](#_ootxhst32vt6) **38**

[**Valor do Crédito oferecido pelo banco (vlr\_saldo)**](#_7ch8uexa8i2t) **38**

[**Valor do Score (vlr\_score)**](#_5keeu9vposrz) **38**

[**Número de atendimentos (num\_atend)**](#_5keeu9vposrz) **38**

[**Número de atendimentos atrasados (num\_atend\_atrs)**](#_fh4qjxxr1u5i) **38**

[**Quantidade de Operações realizadas (qtd\_oper)**](#_b8y6ce9vdsn) **39**

[**Número de produtos (num\_produtos)**](#_b8y6ce9vdsn) **39**

[**Quantidade de reclamações abertas (qtd\_reclm)**](#_b8y6ce9vdsn) **39**

[**Quantidade de restritivos (qtd\_restr)**](#_b8y6ce9vdsn) **39**

[**Classificação do cliente com o banco (cod\_rating)**](#_b8y6ce9vdsn) **39**

[**Índice de Atrito (ind\_atrito)**](#_b8y6ce9vdsn) **39**

[**Índice de Engajamento (ind\_engaj)**](#_b8y6ce9vdsn) **39**

[4.4. Modelagem](#_3whwml4) 40

[**O modelo utilizado**](#_fgmur5x78h4n) **40**

[**Subdatasets**](#_yn748wbxgf27) **41**

[**Target: Índice de engajamento**](#_cs0pfzzhzuf1) **41**

[**1º Conjunto de Dados: Agregação somatória**](#_1s9nz13uspw4) **41**

[**Modelo 1: Random Forest Classifier**](#_1ny6s2rxli5g) **41**

[**2º Conjunto de Dados: Dataset a partir da safra do mês 11 (Novembro de 2021)**](#_biezqqc7v7m6) **44**

[**Modelo 1: Random Forest Classifier**](#_h7crg0kpt8eu) **44**

[**Modelo 2: SVM**](#_81n67ncigsir) **46**

[**Modelo 3: MLP Classifier**](#_q3ivhdo0ezj2) **46**

[**Modelo 4: GradientBoostingClassifier**](#_vg6rx19x6n8q) **47**

[**3º Conjunto de Dados: Dataset de clientes engajados (Utilizando o método da mediana para todas as colunas)**](#_ks7kvp9ijutf) **48**

[**Modelo 1: Random Forest Classifier**](#_375lvngl581y) **48**

[**Target: Índice de novo cliente**](#_63133h7xp4it) **49**

[**Modelo 1: Random Forest Classifier**](#_blqiiao196zw) **49**

[**Avaliação do modelo a partir da divisão de treino e teste**](#_4jv3ikwl5jts) **50**

[**Avaliação do modelo a partir de todo o conjunto de dados**](#_5q91app0624n) **51**

[**Modelo 2: Gradient Boosting Classifier**](#_oiv4yinrycgt) **51**

[**Avaliação do modelo a partir da divisão de treino e teste**](#_nxrcn336n00u) **51**

[**Avaliação do modelo a partir de todo o conjunto de dados**](#_za8e3ywgc5k) **52**

[**4.5. Avaliação**](#_l6s6ieqy5tf7) **53**

[**Escolha do modelo para o índice de engajamento**](#_e67uebq2rgyc) **53**

[**Escolha do modelo para o índice de novo cliente**](#_89dz8ejshowu) **54**

[**5. Conclusões e Recomendações**](#_ij5zpva40iqn) **55**

[**6. Referências**](#_1pxezwc) **56**

# 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

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**:Métricas utilizadas para avaliar acertos e erros de modelos preditivos.

# 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



### 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.

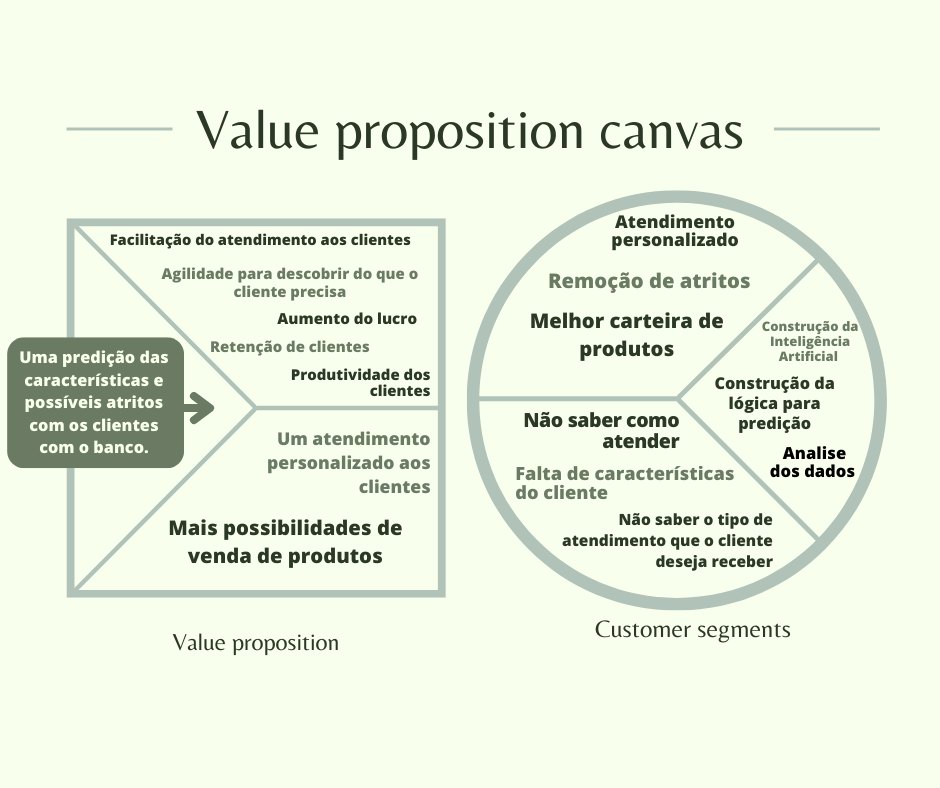
A tarefa da nossa solução é de classificação de clientes em três campos, clientes atritados e seus possíveis atritos, clientes não atritados que possuem um alto engajamento e que podem ter seu relacionamento com o banco maximizado, e por fim, possíveis novos clientes.

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



### 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 entres 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 3 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.
   1. ***Não aplicável***
   2. Os dados possuem uma validade pequena, visto que muito deles são atualizados recorrentemente
   3. ***Não aplicável***
   4. 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).*

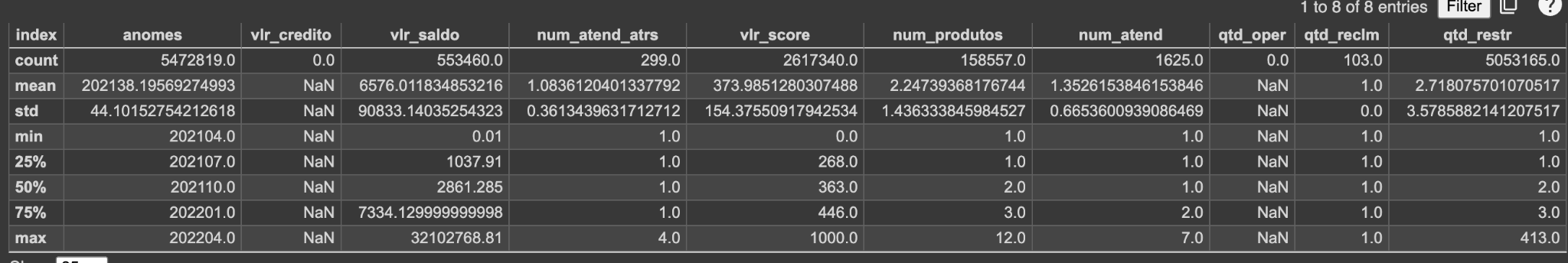


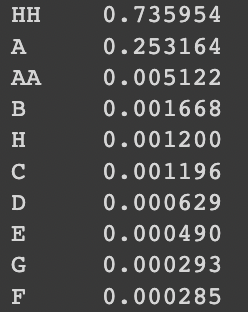
Então chegamos ao seguintes dados: **43% das pessoas não possuem dados de crédito** *(podemos trabalhar com esse grupo pensando em potenciais clientes)* , **7% das pessoas possuem um crédito de R$0 - R$287** *(podemos pensar esse grupo como pessoas que não possuem muitos produtos de crédito),* **25% 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.

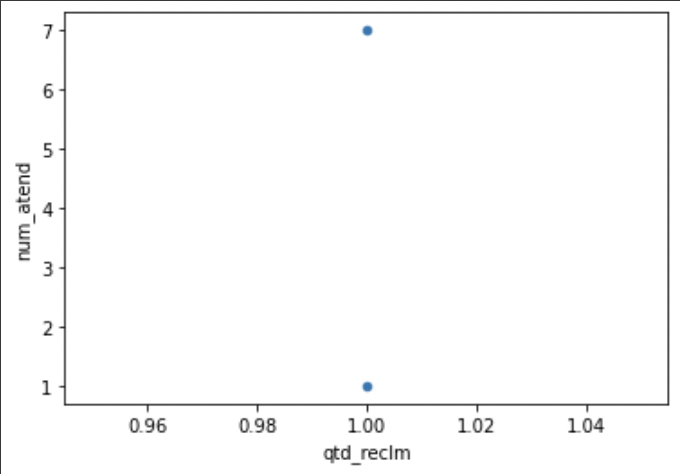
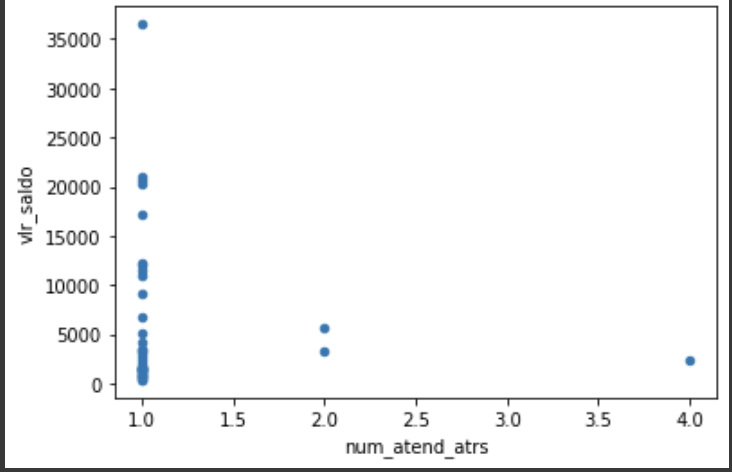
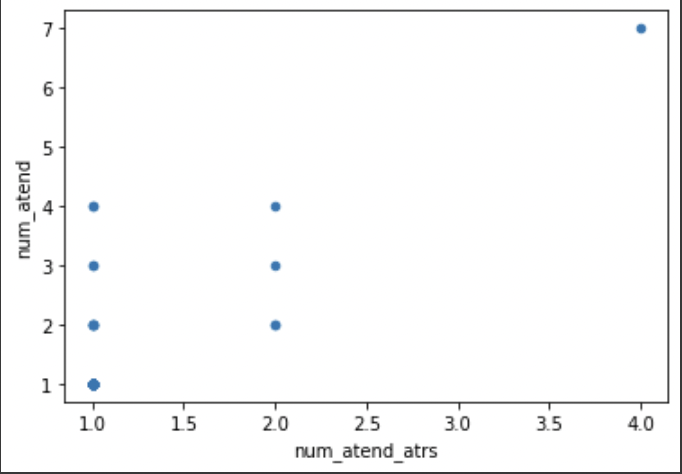
**Grupo de pessoas que não temos dados de crédito de mercado:**

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

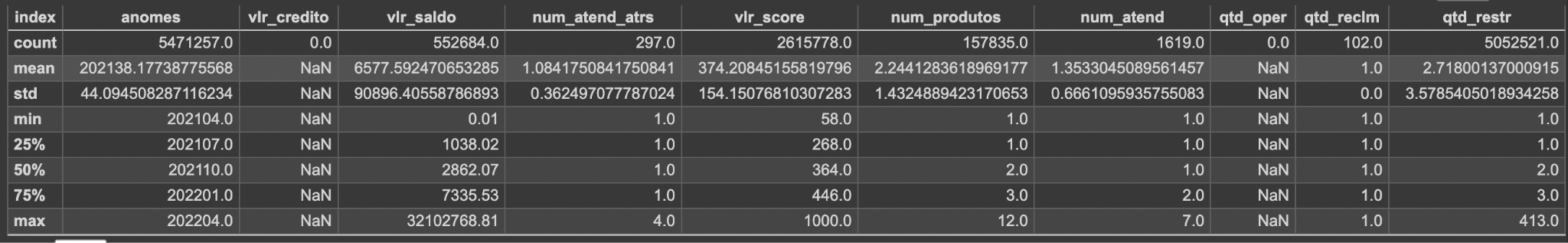


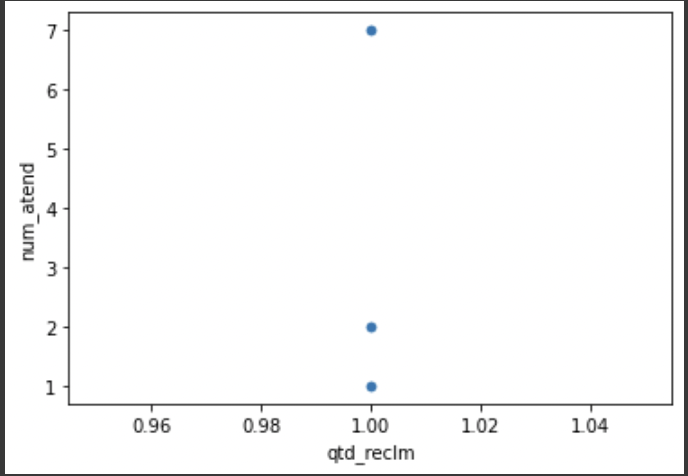
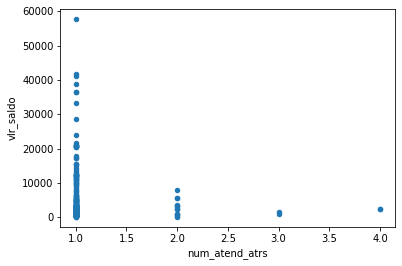
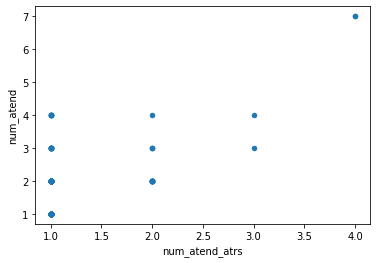
Dentro desse grupo temos o seguinte percentil de rating das pessoas, ou seja, a grande maioria das pessoas que o banco não tem dados de crédito ele classifica como um risco alto para a disponibilização de produtos.

(crédito e score são NaN)

*  (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)

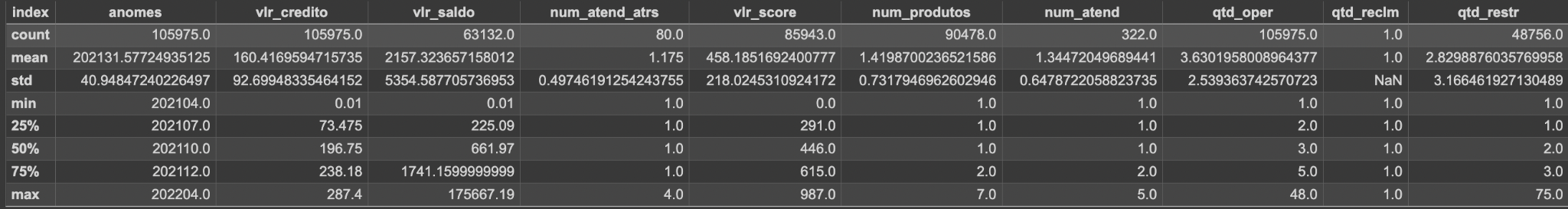
(crédito é NaN e score é diferente de NaN)

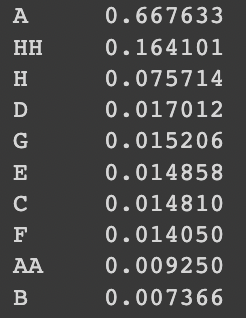


*  (gráfico que compara quantidade de reclamações com números de atendimentos. Percebemos que não há relação direta entre números de atendimentos com quantidade de reclamações, mas que só há reclamação quando há atendimento)
*  (gráfico que compara números de atendimentos atrasados com valor de saldo. 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)

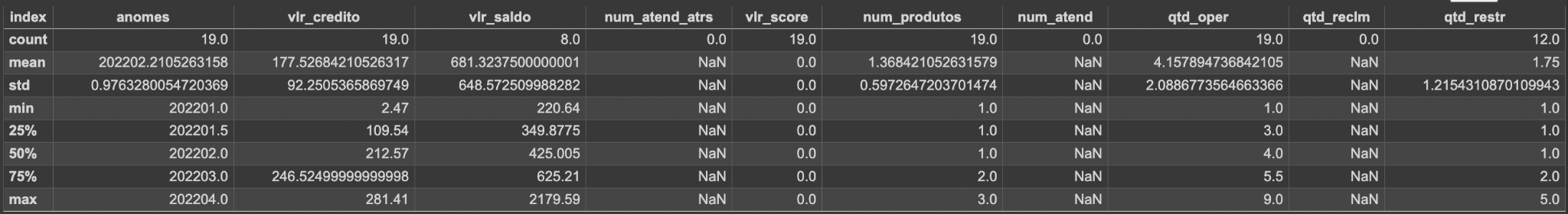
**Grupo de pessoas que tem crédito de R$0 - R$287:**

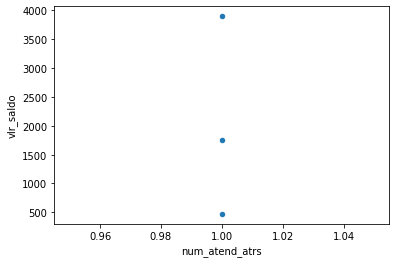
(Descrição geral): número de atendimentos e reclamações baixo.



(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.)

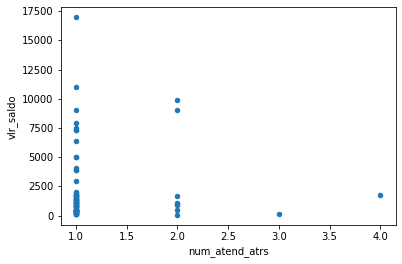
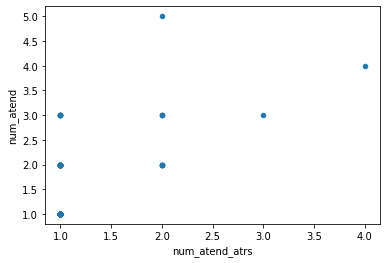
(valor de score sendo NaN)



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

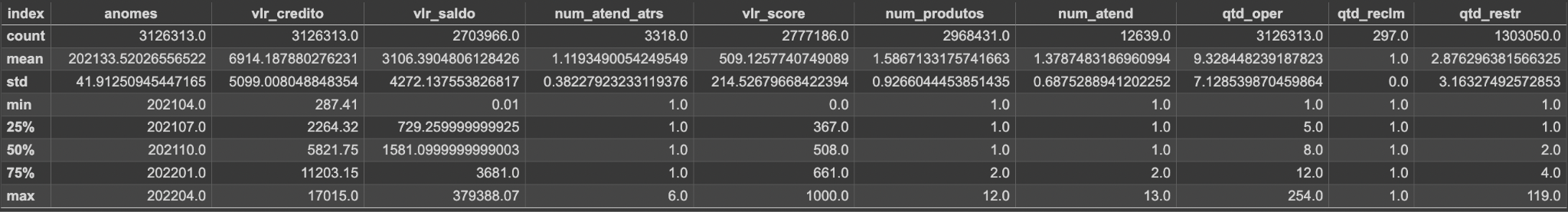
(valor de score não sendo NaN)

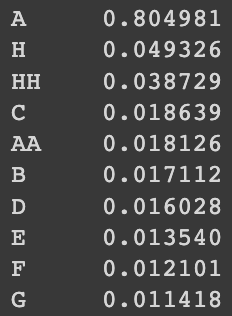


*  (gráfico que compara quantidade de reclamações com número de atendimentos)
*  (gráfico que compara número de atendimentos atrasados com valor de saldo)
*  (gráfico que compara número de atendimentos atrasados com número de atendimentos. Percebemos que quando há mais atendimentos há maior número de atendimentos atrasados)
*  (gráfico que compara quantidade de reclamações com número de atendimentos atrasados)

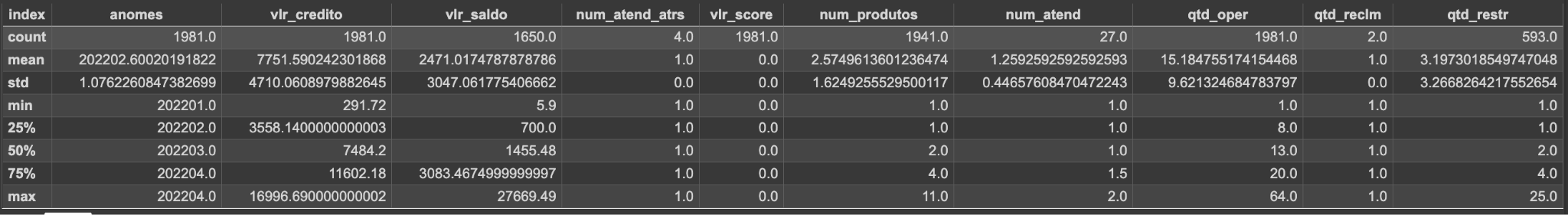
**Grupo de pessoas com crédito de R$287 - R$17.015:**

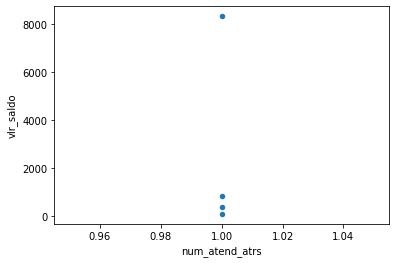
(Descrição geral): número de atendimentos alto e reclamações baixo.



(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)

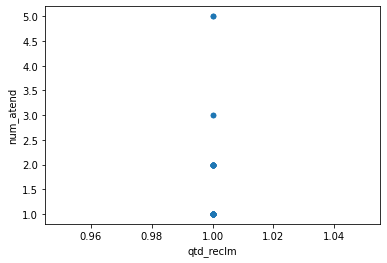
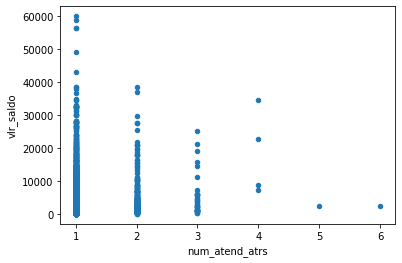
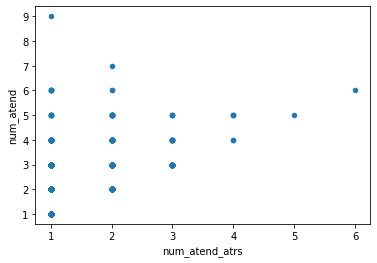
(campo score é NaN)



*  (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)

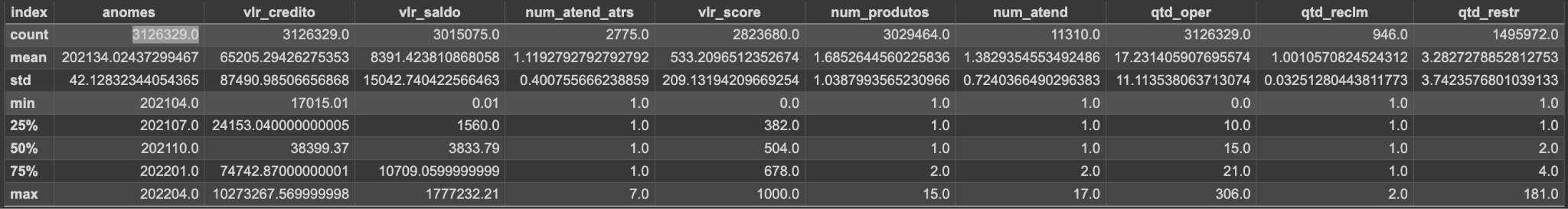
(campo score não é NaN)

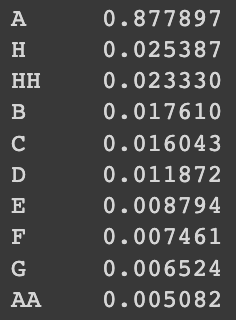


*  (gráfico que compara quantidade de reclamações com números de atendimentos)
*  (gráfico que compara número de atendimentos atrasados com valor de saldo. Quanto menor o saldo, maior o número de atendimentos atrasados.)
*  (gráfico que compara número de atendimentos atrasados com número de atendimentos)
*  (gráfico que compra a quantidade de reclamações com a quantidade de atendimentos atrasados)

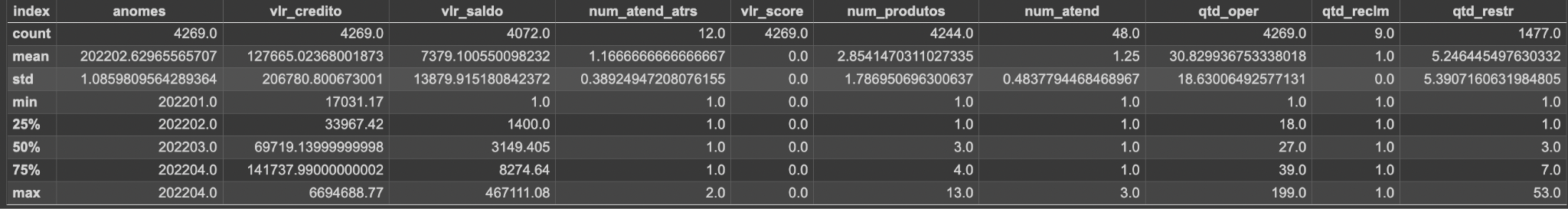
**Grupo de pessoas com crédito de R$17.015 - R$10.348.109:**

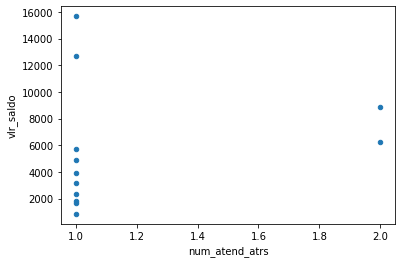
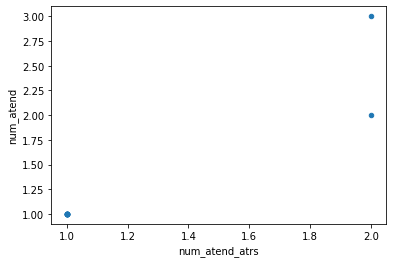
(Descrição geral): número de atendimentos e reclamações alto.



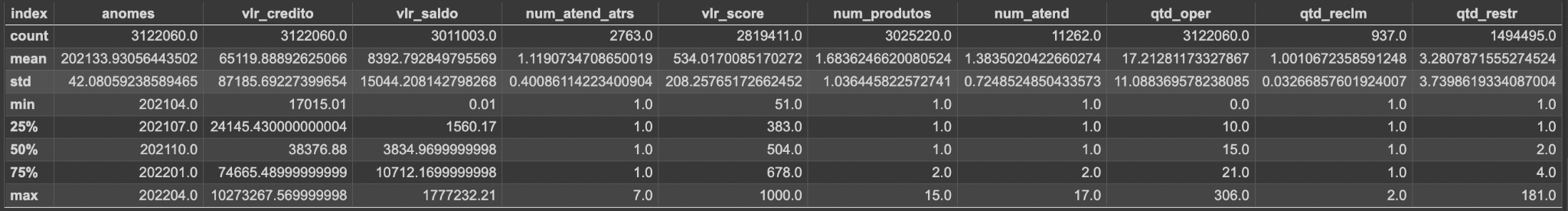
(percentil de cada "rating" desse grupo. Comprovando a hipótese novamente que quando o banco sabe a quantia de crédito do cliente ele confia mais nele)

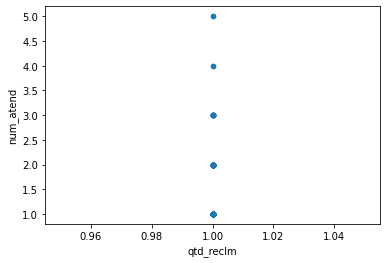
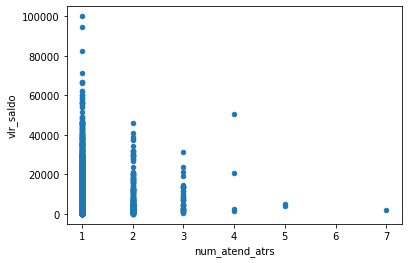
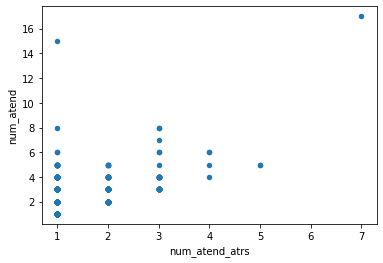
(score é NaN)



*  (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)



*  (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:**

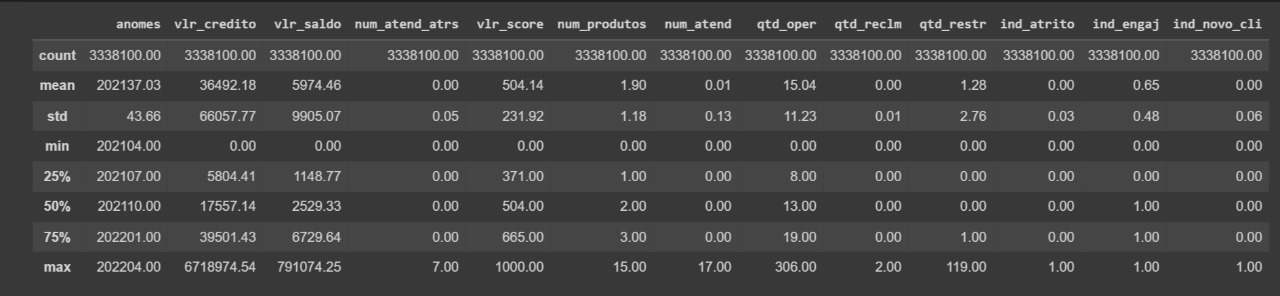
Percebemos que quando o banco sabe o score da pessoa normalmente essa pessoa tem um maior número de atendimentos, um maior número de atendimentos atrasados e um maior número de reclamações (hipótese: essas pessoas interagem mais com o banco). Também percebemos que quando uma pessoa reclama, ela reclama 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 divido em três subdatasets, foram eles:

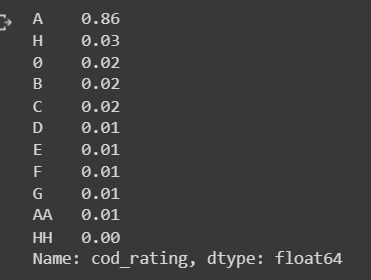
1. Clientes que em algum momento já foram engajados com o Banco
2. Clientes que em algum momento já tiveram um atrito
3. Possíveis novos clientes

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

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

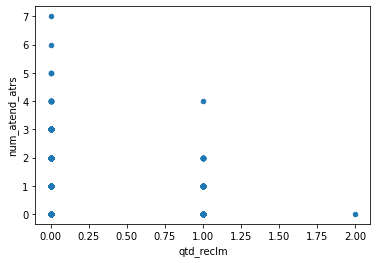


Em relação ao dataset original, os clientes que são ou já foram engajados com o banco representam cerca de 27%.

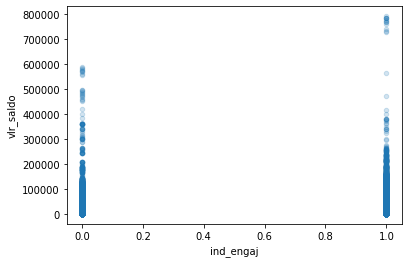
Verificando o **cod\_rating** dos clientes neste novo dataframe, temos a seguinte frequência: 

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 agrega 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 enconding.

| **Nome da feature** | **Método Utilizado** | **Motivo** |
| --- | --- | --- |
| 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 Enconding 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 Enconding 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]](#footnote-0) |

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.

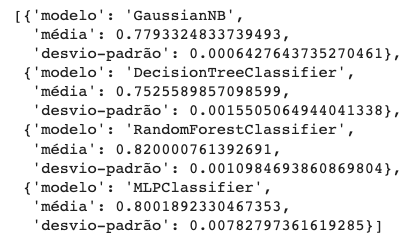
| **Nome da feature/coluna** | **Motivo de uso** |
| --- | --- |
| **Número de CPF** *(num\_cpf\_hash)* | Em uma primeira análise, o número de cpf nos permite, e garante, a manipulação e agregação dos dados em uma só linha. *\*\*Durante a modelagem esta coluna será removida visando a maior eficiência na predição.* |
| 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.  *\*\*Apesar de terem sido selecionadas por uma mesma lógica, será também explorado e utilizado as características individuais destes indicadores*. |
| Valor do Crédito oferecido pelo banco *(vlr\_saldo)* |
| 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 quão engajado. |
| Número de atendimentos atrasados *(num\_atend\_atrs)* |

| Quantidade de Operações realizadas *(qtd\_oper)* | 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 *(num\_produtos)* | 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 *(qtd\_reclm)* | 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 *(qtd\_restr)* | 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 *(cod\_rating)* |
| Índice de Atrito *(ind\_atrito)* | 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.   Também se trata de uma variável resposta para um dos modelos propostos. |
| Índice de Engajamento *(ind\_engaj)* |

## 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]](#footnote-1)** com um *cross validation* (método utilizado para avaliação de desempenho que consiste no particionamento dos dados)separando nosso dataset em 6 partes.



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

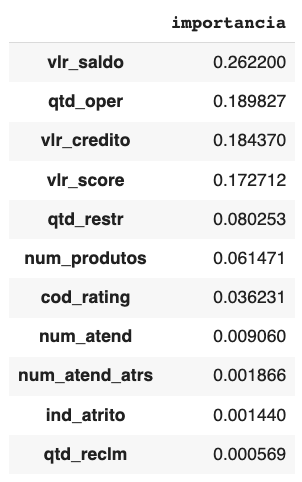
### Subdatasets

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. 

Antes de um modelo inicial foi realizado uma *feature* *importance* com o algoritmo RFC (Random Forest Classifier) para validar a importância de cada feature para nosso modelo. *(Imagem ao lado)*

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%.

Logo, a partir dessa base estatística seguimos nosso modelo utilizando as 6 primeiras colunas como variáveis preditoras para o modelo.

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 Search[[3]](#footnote-2) 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

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.

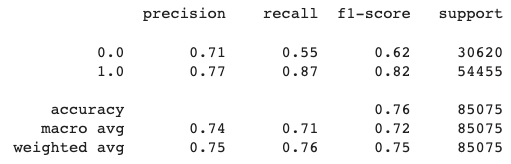
**Precision x Recall Plot**

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



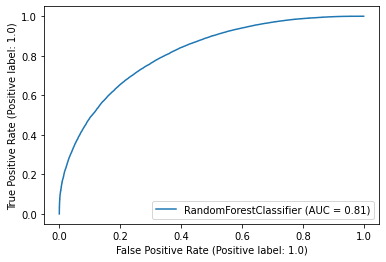
**Classification Report**

Relatório visando analisar a qualidade do algoritmo aplicado.

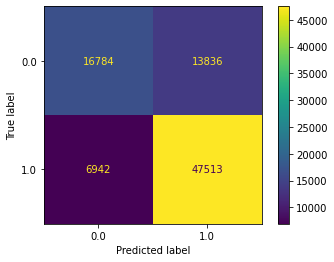
****

**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

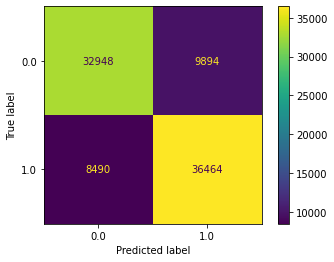
Inicialmente, tratamos os dados criando um data frame com somente clientes do mês 11.Foi feito um *label encoding* na coluna **cod\_rating** já que ela é um dado categórico. Substituímos valores em branco por zero nas colunas: **num\_atend\_atrs, num\_atend, qtd\_oper, qtd\_reclm, qtd\_restr,ind\_engaj.**

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.

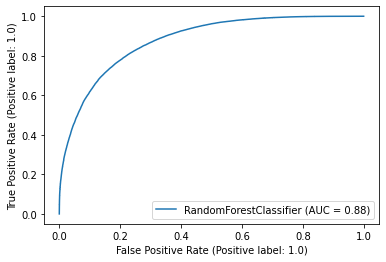
O modelo utilizado foi o RandomForestClassfier e uma cross validation de métrica de scoring f1,que é a média harmônica da acurácia e recall fazendo que o modelo seja balanceado tentando ter proporções similares de verdadeiros positivos e verdadeiros negativos, retornou aproximadamente 0.80.

Utilizando o GridSearch para hiper parametrizar o modelo, obtivemos os seguintes resultados:

* bootstrap=True
* max\_depth=None
* min\_samples\_leaf=5
* min\_samples\_split=2
* n\_estimators=120
* n\_jobs=-1
* random\_state=42
* criterion='entropy'



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 precisão é de 0.75 e acurácia de 0.66



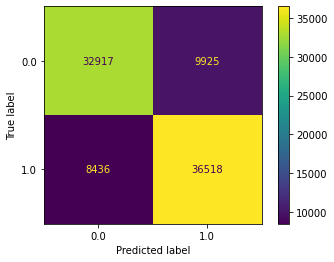
Acima temos a curva roc do modelo que confirma ainda há espaço significativo de melhoria da precisão’

Devido a valores baixos, foi realizado um balanceamento dos dados, a fim de que o número de pessoas engajadas e não engajadas tivesse uma menor disparidade. Após realizado este balanceamentos, os resultados foram os seguintes:

**Precisão:** 0.7862

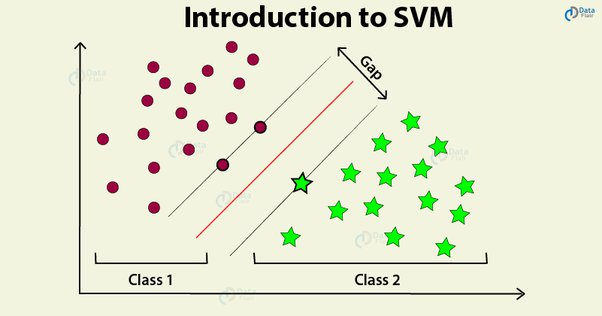
**Revocação:** 0.8123

**F1 Score:** 0.7991



##### 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:



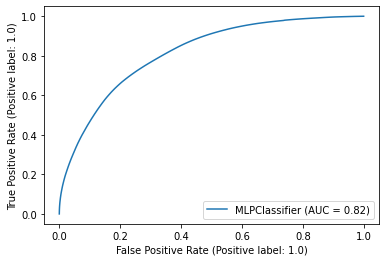
Utilizar o SVM é vantajoso, visto que o modelo trabalha muito bem com dados dissociados de maneira irregular.

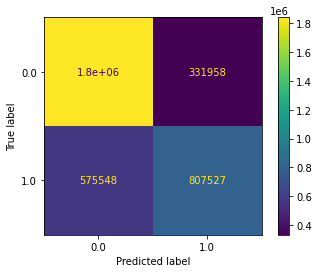
Devido ao uso de processamento, não foi possível obter resultados válidos,com este modelo, mesmo com todas as otimizações possíveis.

##### 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:

 **Curva ROC**

**Matriz de confusão**

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

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

##### Modelo 4: GradientBoostingClassifier

Trata-se de um modelo que utiliza a técnica de boosting para regressão e classificação. O intuito é agregar diferentes modelos fracos para gerar um preditivo melhor. No entanto, o modelo não rodou por exceder o nível computacional do Google Colab, mesmo com todas as otimizações feitas.

#### 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.

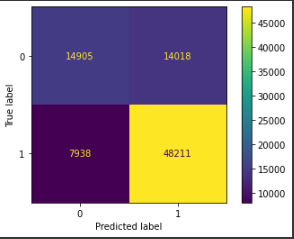
Feito isso, percebemos que mais de 90% da importância estava alocada em 6 colunas.

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.

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



### 

### 

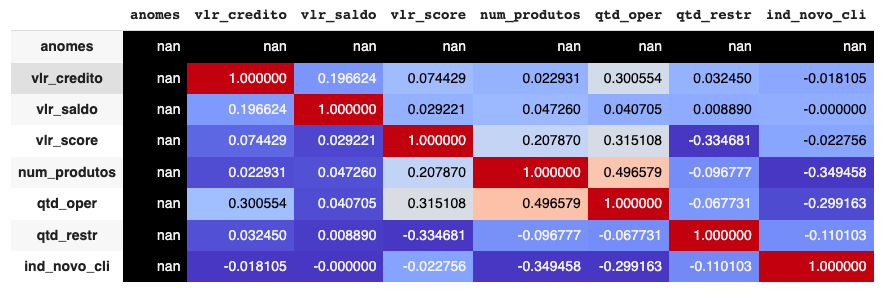
### 

### Target: Índice de novo cliente

##### Modelo 1: Random Forest Classifier

Utilizamos das análises anteriores como base para a modelagem de uma solução mais otimizada e que maximize os resultados para a predição de um potencial novo cliente. Para tanto, iniciamos uma *feature importance* para encontrar as correlações entre as colunas.





Contudo, algumas dessas colunas são específicas para clientes como: **vlr\_saldo, num\_produtos, qtd\_oper;** logo, não cabe utilizá-las, uma vez que necessitamos predizer se a amostra representa um potencial novo cliente.

Utilizando as features **vlr\_credito, vlr\_score, qtd\_restr** e o algoritmo Random Forest obtivemos os seguintes resultados:

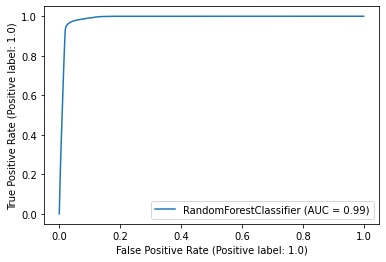
#### 

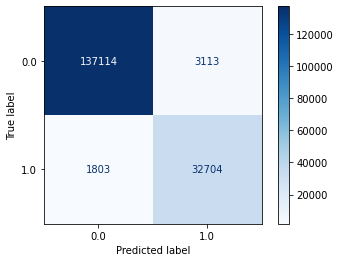
###### Avaliação do modelo a partir da divisão de treino e teste

**Revocação:** 0.947

**Precisão:** 0.913

**F1 Score:** 0.930

**Curva ROC**

**Matriz de confusão**

**Acurácia de treino:** 0.972

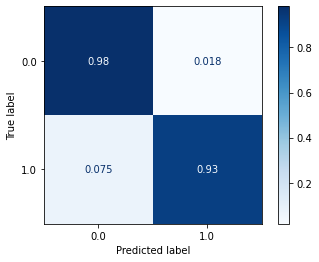
**Acurácia de teste:** 0.971

###### Avaliação do modelo a partir de todo o conjunto de dados

**Revocação:** 0.925

**Precisão:** 0.834

**Acurácia:** 0.976



##### Modelo 2: Gradient Boosting Classifier

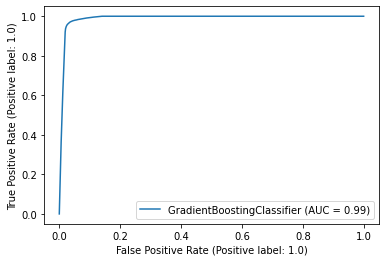
Assim como explicado anteriormente na seção "Modelo 4: GradientBoostingClassifier", utilizamos este algoritmo para classificar o índice de engajamento, obtendo os seguintes resultados:

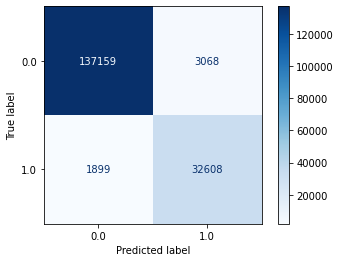
###### Avaliação do modelo a partir da divisão de treino e teste

**Revocação:** 0.944

**Precisão:** 0.914

**F1 Score:** 0.929

**Curva ROC**

**Matriz de confusão**

###### Avaliação do modelo a partir de todo o conjunto de dados

**Revocação:** 0.921

**Precisão:** 0.323

## 

## 

## 

## 

## 

## 4.5. Avaliação

A partir dos modelos gerados e suas respectivas avaliações na seção anterior, optamos pelo modelo que mais se encaixaria com as regras de negócio. A métrica utilizada para avaliar o modelo foi a Revocação (Recall, em inglês). A revocação é uma métrica de avaliação de modelos baseada nas classificações que o modelo acertou, podendo ser resumida pela seguinte fórmula:

Portanto, infere-se que, a revocação é a divisão entre os valores que o modelo acertou — classificando-o como positivo — sobre tudo que ele deveria ter classificado como positivo — soma dos acertos e erros.

O intuito do Banco PAN com o modelo é a maximização da relação entre banco e cliente, como foi explicitado na introdução deste documento. Logo, indicar que um cliente é engajado ou atritado, é menos prejudicial do que classificá-lo como não atritado ou não engajado. O principal motivo está na utilização final do modelo. Para melhorar o atendimento ao cliente, a abordagem irá se adaptar às três maneiras de classificação: engajado, atritado, potencial novo cliente; sendo melhor apresentar um atendimento mais cauteloso por um falso positivo, do que não apresentar este atendimento devido a um falso negativo.

### Escolha do modelo para o índice de engajamento

Baseado na análise de modelos no início da seção 4.4, e comparando os modelos documentados, o algoritmo que apresentou maior desempenho foi o *RandomForestClassifier*. Ao comparar as três diferentes maneiras de agregação, observa-se que a segunda maneira (anomes 11) apresenta melhor desempenho em relação aos outros conjuntos de dados.

Apesar de apresentar um número menor para a métrica de revocação, o modelo apresenta uma menor disparidade entre os valores de precisão e revocação, cerca de uma diferença 70% menor. Grandes diferenças entre revocação e precisão podem apresentar um possível indício de underfitting[[4]](#footnote-3) do modelo, contudo, analisando as diferenças de acurácia do treinamento e teste das demais modelagens, é possível observar um possível overfitting. Veja na comparação a seguir:

\****Modelagem utilizando o terceiro conjunto de dados.***

Como é possível observar, há uma grande diferença entre os valores, principalmente, ao compará-los com a modelagem escolhida:

***\*Modelagem utilizando o segundo conjunto de dados***

### Escolha do modelo para o índice de novo cliente

Por realizar um extenso processo de hiper parametrização e os estudos de modelos gerados anteriormente, otimizamos o tempo reutilizando essas métricas e avaliando sempre buscando o melhor resultado e que mais se encaixasse com as regras de negócio explicitadas anteriormente.

Utilizando a métrica de Revocação, pode-se observar que o algoritmo RandomForestClassifier traz os melhores resultados. Contudo, vale ressaltar que a grande disparidade entre as métricas de Precisão e Revocação do modelo GradientBoosting também traz indícios de underfitting, reforçando a escolha pelo primeiro modelo.

# 

# 

# 

# 5. Conclusões e Recomendações

Em síntese, a modelagem apresentada, utilizando o algoritmo Random Forest e a maneira de agregação, satisfaz as necessidades do Banco PAN, classificando o engajamento dos clientes e potenciais novos clientes. Contudo, compreendemos que este é apenas um passo a um caminho extenso para implementação do projeto. Para tanto, nós, como integrantes do Banco PAN, nos colocamos à disposição de qualquer dúvida acerca do projeto: Ariel Kisilrvzky, [Alberto Miranda,](https://www.linkedin.com/in/alberto-da-rocha-miranda-angrysine/) [Antônio Ribeiro](https://www.linkedin.com/in/antonioribeiro893/), [Gabriel Carneiro](https://www.linkedin.com/in/gabecarneiro/), [Luiz Alencar](https://www.linkedin.com/in/luiz-k-alencar/) e [Henrique Lemos](https://www.linkedin.com/in/henriquelfmatias/).

Recomendamos que para uma melhor aplicabilidade, o sistema seja integrado a uma API, estabelecendo conexões às atuais soluções do Banco PAN. É de grande importância que a equipe que utilizará o modelo seja devidamente treinada para utilizar ao máximo a capacidade da Inteligência Artificial de maneira responsável e ética, respeitando a LGPD e todas suas orientações.

# 6. Referências

Alguns dos conteúdos que utilizamos como base no desenvolvimento do projeto.

BERTAN, E. O que é Precisão e Revocação. Disponível em: <https://medium.com/computando-arte/o-que-%C3%A9-precis%C3%A3o-e-revoca%C3%A7%C3%A3o-b0b991b67cde>. Acesso em: 4 out. 2022.

‌AZANK, F. Dados Desbalanceados — O que são e como evitá-los. Disponível em: <https://medium.com/turing-talks/dados-desbalanceados-o-que-s%C3%A3o-e-como-evit%C3%A1-los-43df4f49732b>. Acesso em: 4 out. 2022.

Precision-Recall. Disponível em: <https://scikit-learn.org/stable/auto\_examples/model\_selection/plot\_precision\_recall.html#:~:text=A%20system%20with%20high%20precision>. Acesso em: 4 out. 2022.

HARRISON, M. Machine Learning Pocket Reference. [s.l.] “O’Reilly Media, Inc.”, 2019.

SILVA, L. A. Introdução à Mineração de Dados - Com Aplicações em R: com Aplicações em RL. Acesso em: 6 out. 2022.

‌Entenda o que é Underfitting e Overfitting (Machine Learning). Disponível em: <https://didatica.tech/underfitting-e-overfitting/>. Acesso em: 7 out. 2022.

‌

‌

‌

‌

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. [↑](#footnote-ref-0)
2. O F1-Score é uma média harmônica unindo as duas métricas de precisão e revocação. [↑](#footnote-ref-1)
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> [↑](#footnote-ref-2)
4. Em machine learning, overfitting é quando o modelo apenas "decora" os dados, não aprende de fato. Underfitting é quando o modelo não consegue predizer com eficácia pois não conseguiu aprender com os dados apresentados. [↑](#footnote-ref-3)