

Estudo Analítico sobre Desconexão e Retenção

Objetivo

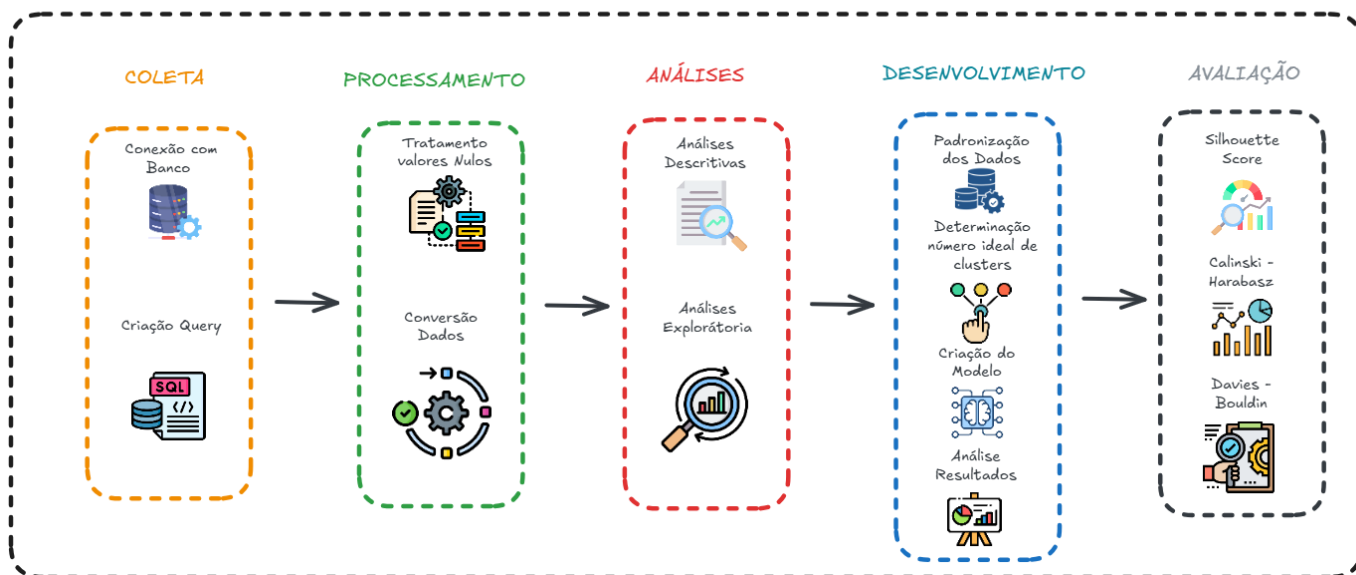
O objetivo desta análise é compreender os padrões de comportamentos dos clientes que solicitaram desconexão. Desta forma, através dos resultados encontrados vamos buscar identificar fatores determinantes que influenciam a decisão de cancelamento, bem como oportunidades para melhorar a retenção de clientes. O estudo visa fornecer insights valiosos que possam ser utilizados para desenvolver estratégias eficazes de retenção e melhorar a satisfação dos clientes.

Metodologia

Para alcançar o objetivo do estudo, foram aplicadas as seguinte Pipeline:

PIPELINE

Estudo Analítico sobre Retenção e Desconexão



• Coleta de Dados

Os dados foram coletados do banco de dados BDS mais especificamente da VIEW denominada como VW_ATENDIMENTO_ORDEM_DESCONEXAO_CONSOLIDADO. Certamente, nela temos armazenado as ordens de solicitação de desconexão dos clientes da Ultra Fibra. Portanto, a view é alimentada através da tabela TB_ATENDIMENTO_ORDEM_DESCONEXAO_CONSOLIDADO.

• Validação da Base de Dados

A aplicação de determinadas validações na base é crucial para garantirmos a qualidade dos dados antes de prosseguir com as análises e modelagens. Neste sentido, foi aplicado o **Identificação de Valores Nulos** e **Identificação de valores duplicados** e isso foi de extrema importância para compreendermos os padrões e comportamentos que temos na base VW_ATENDIMENTO_ORDEM_DESCONEXAO_CONSOLIDADO.

• Processamento de Dados

O processamento de dados é uma etapa crucial para garantir que os dados estejam no formato adequado para as análises e modelagens que vão ser futuramente realizadas. Certamente, as etapas aplicadas foram **Conversão do tipo de dados de determinadas colunas** e **Remoção dos Valores Nulos**. Portanto, a partir destas medidas tomadas tivemos como ganhos: melhora na qualidade dos dados e dados devidamente processados facilitando assim a criação das análises e a construção de modelos preditivos.

• **Análise Descritiva e Exploratória**

Nesta etapa, temos o desenvolvimento de dois tipos de análises as **descritivas e exploratórias**. Certamente, podemos concluir que as descritivas foram aplicadas para identificar padrões, desvios e anomalias nos dados. Enquanto as exploratórias permite irmos a um passo além, estabelecendo classificações, agrupamentos e correlações entre os dados. Com base, no que foi identificado nos dados que foi utilizado para o estudo, se tornou necessário dividir as análises em três grupos diferentes (Cancelamentos, Retenção e Operacional) para termos uma visão abrangente e detalhada e com diferentes aspectos que influenciam a decisão dos clientes de solicitar desconexão.

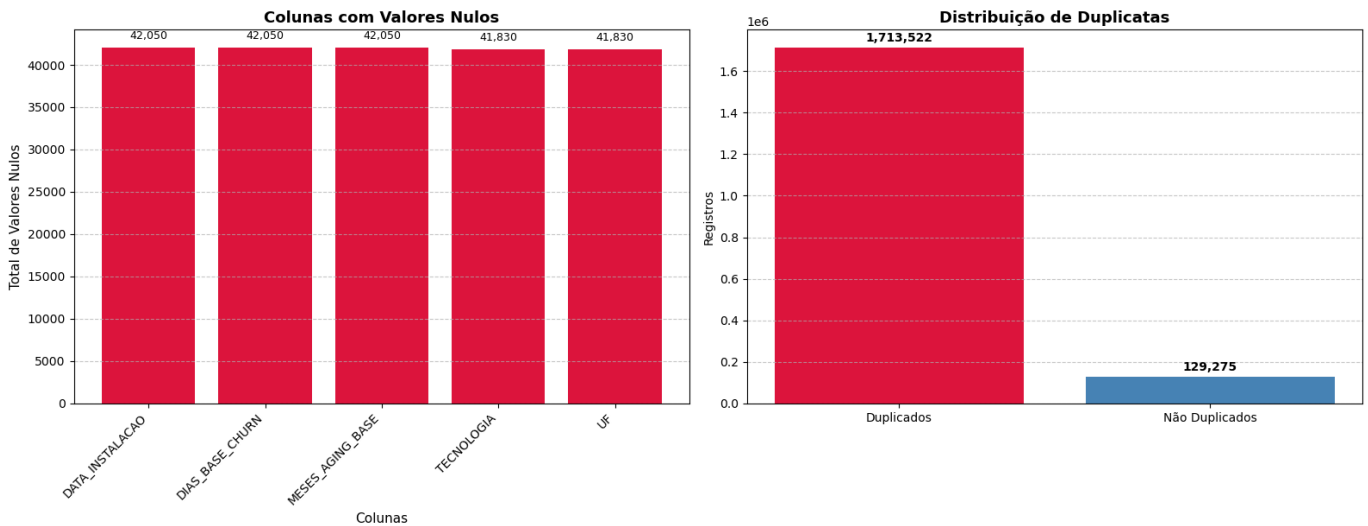
- 1. Cancelamento -> A análise de Cancelamento foca nos clientes que efetivamente solicitaram a desconexão. O objetivo é identificar padrões e características comuns entre esses clientes. Compreender esses fatores pode ajudar a empresa a identificar os principais motivos de insatisfação e desenvolver estratégias para mitigar esses problemas.
- 2. Retenção -> A análise de Retenção examina os clientes que foram retidos após solicitar a desconexão. Esta seção busca identificar as características e ações que contribuíram para a retenção desses clientes. Entender o que funcionou bem pode ajudar a empresa a replicar essas estratégias com outros clientes em risco de desconexão.
- 3. Operacional -> A análise Operacional explora os dados relacionados ao desempenho e à qualidade dos serviços oferecidos pela empresa e pela quantidade de protocolos impostas por cada cliente.

• **Modelagem e Aplicação de Modelos de Machine Learning**

Aplicação do Algoritmo K-means para realizar uma clusterização com o intuito de segmentar os clientes em grupos com características semelhantes. Desta forma, podemos identificar e analisar os clusters formados, destacando as principais características de cada grupo e os padrões de comportamento observados.

1. Coleta e Validação dos Dados

⚠ **Diagnóstico de Qualidade dos Dados**



A iniciativa de realizar uma validação nos dados é termos um diagnóstico inicial da qualidade dos dados utilizados no estudo, com foco na identificação de lacunas que possam comprometer a consistência das análises futuras. A avaliação contempla a presença de valores nulos e registros duplicados.

1.1 Colunas com Valores Nulos

Segundo o gráfico temos 5 colunas com volumes expressivos de dados ausentes:

Coluna	Quantidade de Valores Nulos
DATA_INSTALLACAO	42.050

Coluna	Quantidade de Valores Nulos
DIAS_BASE_CHURN	42.050
MESES_AGING_BASE	42.050
TECNOLOGIA	41.830
UF	41.830

Insights Observados:

- As colunas DATA_INSTALACAO, DIAS_BASE_CHURN e MESES_AGING_BASE apresentam exatamente o mesmo número de valores ausentes, sugerindo que os registros afetados podem estar concentrados em um mesmo subconjunto da base.
- As colunas TECNOLOGIA E UF também compartilham o mesmo volume de nulos, o que pode indicar uma origem comum de falha no preenchimento ou ausência de integração com os sistemas de origem desses dados.
- Certamente, a ausência desses dados podem estar ocorrendo por um problema sistêmico ou na importação dos dados para o banco em produção.

1.2 Distribuição de Duplicatas

Tipo de Registro	Quantidade
Registros Duplicados	1.713.522
Registros Únicos	129.275

Insights Observados:

- A base apresenta uma proporção extremamente elevada de registros duplicados (aproximadamente 93%). Isto acontece pelo fato que um cliente pode ter mais de um protocolo e por conta disso gera mais registro de um cliente na base.
- Desta forma, esta duplicidade não atinge a confiabilidade das análises subsequentes.

2. Processamento

O processamento de dados é uma etapa crucial para garantir que os dados estejam no formato adequado para as análises e modelagens que vão ser futuramente realizadas. Neste sentido, foi aplicada duas etapas fundamentais para o tratamento dos dados:

- **Conversão dos Dados**

Foi realizada a conversão explícita dos tipos de dados das colunas, com o objetivo de garantir que cada variável estivesse representada de forma adequada ao seu conteúdo e uso analítico.

Observações:

1. As conversões realizadas foram de Varchar para Inteiro e Date para Date Time.
2. A aplicação da tipagem correta melhora a performance de algoritmos de machine learning, e evita erros em operações matemáticas, agrupamentos e visualizações.

- **Tratativa dos Valores Nulos**

Foi realizada a limpeza dos dados por meio da remoção de registros com valores ausentes em colunas consideradas críticas para a análise.

Cr terios Utilizados:

- Remo  o de registros com valores nulos simultaneamente nas colunas: TA_INSTALACAO, DIAS_BASE_CHURN, MESES_AGING_BASE, TECNOLOGIA e UF.
- A decis o foi baseada na impossibilidade de imputa  o confi vel para essas vari veis, dada sua import ncia anal tica.

Observa  es:

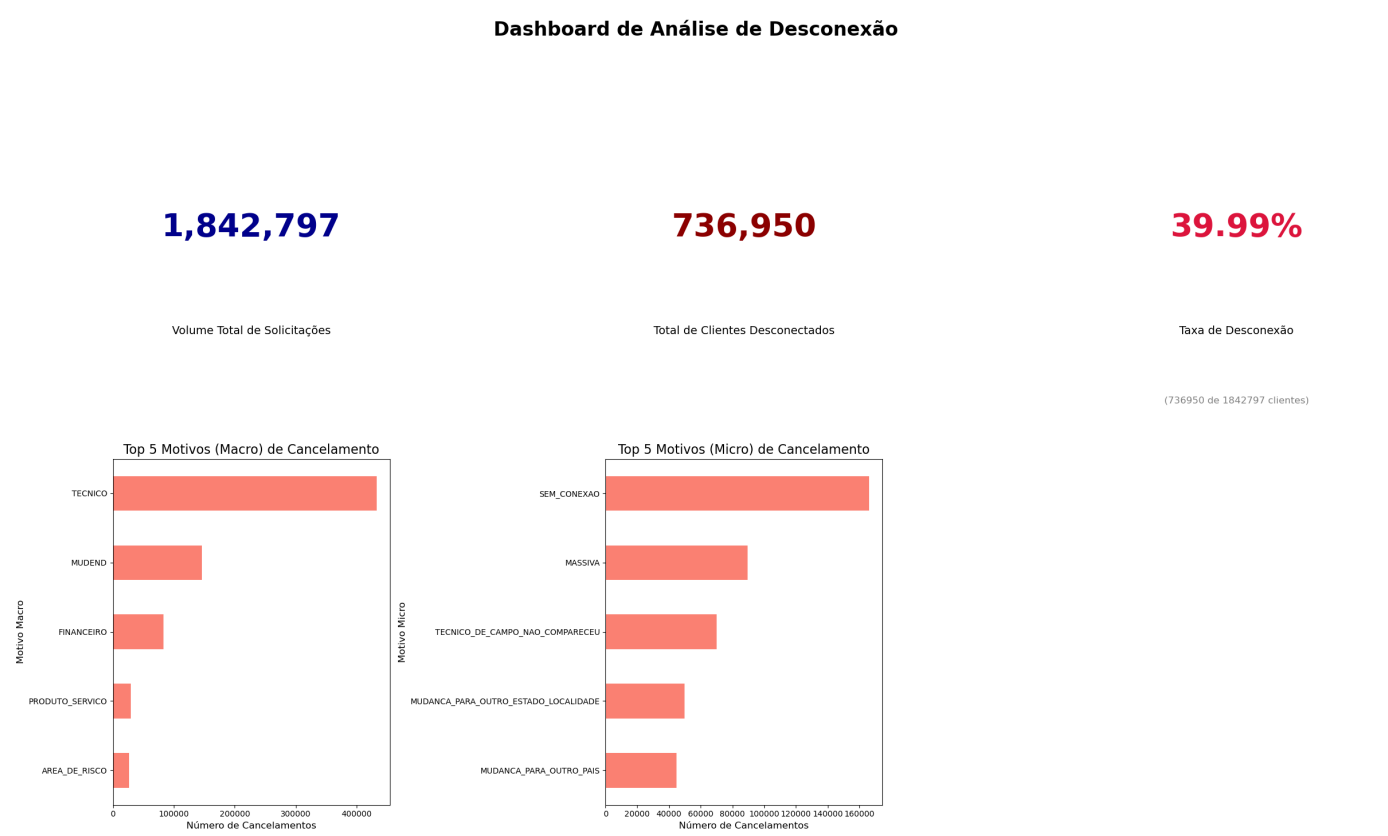
- A remo  o foi preferida   imputa  o para evitar introdu  o de vi s em vari veis-chave.
- A base resultante apresenta maior consist ncia e est  pronta para an lises explorat rias e modelagens supervisionadas.

3. An lises Descritivas e Explor rias

3.1 An lises referente ao Cancelamento de Clientes

A se  o a seguir tem como finalidade analisar os dados relacionados   **desconex o de clientes**, identificando o volume de solicita  es, a taxa de cancelamento, os principais motivos macros e micros que levam o cliente a se desconectar da base e principalmente a correla  o com as demais features presentes na base.

Indicadores Principais



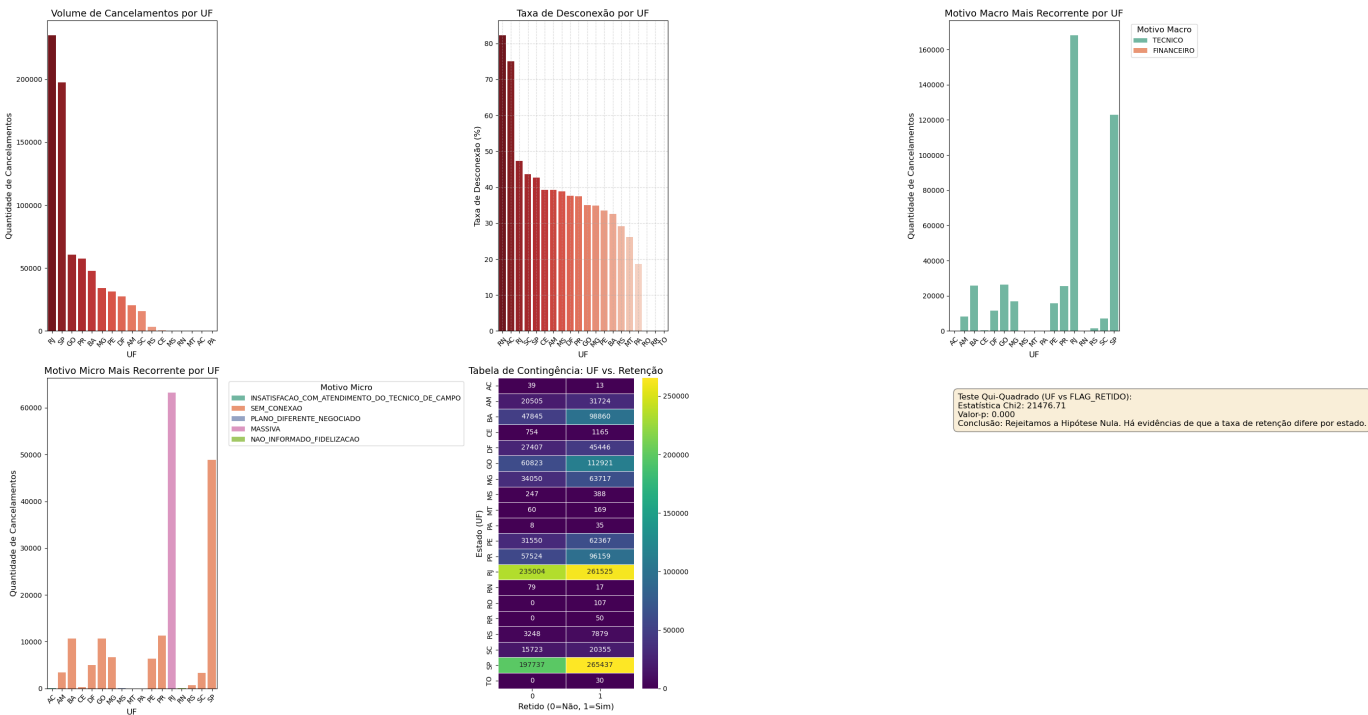
- O motivo **TÉCNICO** lidera os cancelamentos, sugerindo falhas na entrega do serviço ou no suporte técnico.
- **MUDANÇA DE ENDEREÇO** e **ÁREA DE RISCO** indicam fatores externos à empresa, mas que podem ser mitigados com estratégias de retenção geográfica ou planos de migração.
- **FINANCEIRO E PRODUTO_SERVIÇO** apontam para oportunidades que temos do nosso lado como TIM para revisão na precificação e qualidade desses serviços na área da ultra fibra.
- **TÉCNICO DE CAMPO NÃO COMPARECEU** indicam falhas de planejamento na expansão dos serviços ou comunicação com o cliente.

Conclusão:

A análise inicial sobre os indicadores principais envolvendo a **Desconexão** revela uma taxa de cancelamento elevada, com predominância de causas técnicas e estruturais. A concentração de motivos relacionados à infraestrutura, cobertura e suporte técnico indica que há oportunidades claras de melhorias operacional que podem impactar diretamente a retenção de clientes. Além disso, fatores externos como mudança de endereço ou país também aparecem com frequência, reforçando a importância de estratégias de monitoramento e antecipação de churn.

Desconexão por UF

Análise de Cancelamentos por UF: Insights Chave



Análise:

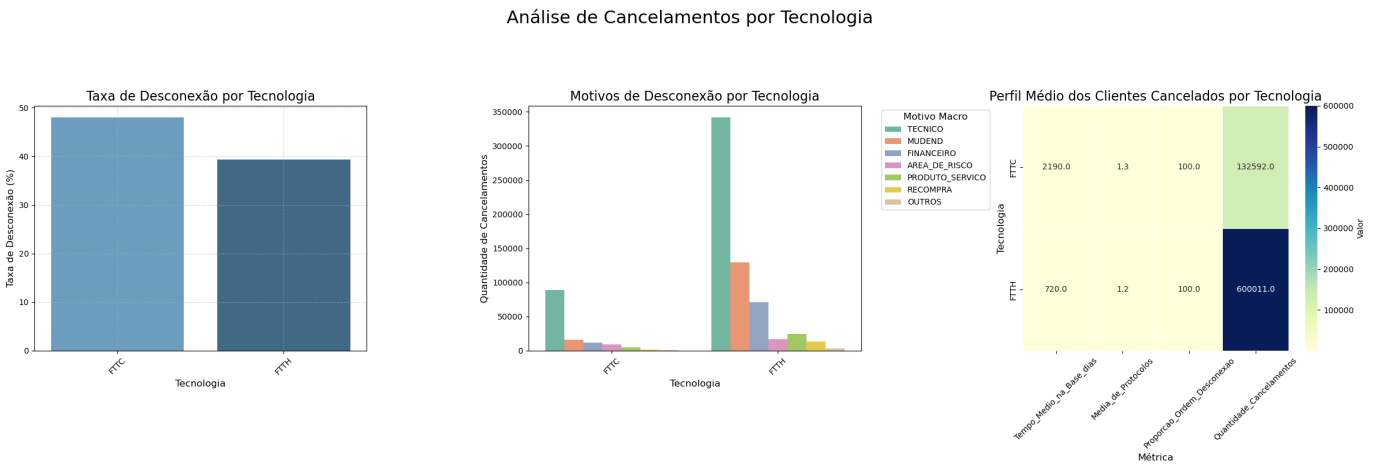
- Os Top 5 Estados com maior volume de cancelamento, são: SP, RJ, MG, BA, PR. Desta forma, podemos observar uma forte concentração dos cancelamentos nas regiões Sudeste e Sul, e isso se torna muito condizente quando entendemos que nessas UFs se concentra a maior base dos clientes.
- Estados da Região Norte como RN, AC e RJ e SC apresentam as maiores taxas de desconexão e podemos interpretar isso associada a problemas estruturais, como limitação de infraestrutura e dificuldade de manutenção.
- Certamente, podemos identificar que o motivo agrupado TÉCNICO predominam entre as UFs. Isto demonstra que a maior dos cancelamentos ocorridos nos estados, esteja relacionado a problemas de serviço, atendimento técnico e infraestrutura técnica.

- Quando olhamos Motivos Micros envolvendo Cancelamento, encontrei alguns padrões. Pois, a maior dos estados há uma predominância quando se trata de cliente SEM_CONEXÃO mas apenas no estado do RJ temos uma relevância maior de MASSIVA. Isso demonstra que na maioria dos estados os nossos clientes sofrem de problemas de sinal ou uma certa ausência de conexão e isso é um fator primordial para eles realizarem uma solicitação de desconexão, mas no estado do Rio de Janeiro isso não acaba sendo seguido e o seu motivo micro em destaque se torna MASSIVA que está relacionado com os cancelamentos em massa por eventos externos.
- A partir da aplicação do teste estatístico denominado como Qui-Quadrado podemos concluir que existe sim uma diferença significativa nas taxas de retenção e cancelamento entre as UFs já que SP, RJ e MG possuem altos volumes de retidos enquanto estados da região Norte possuem menor propensão à retenção. Portanto, podemos definir que devemos realizar a criação de diferentes planos de ações para os estados já que em alguns podemos priorizar medidas envolvendo questão operacionais e comerciais e outros questões técnicas, infraestruturas e gestão da rede.

Conclusão:

- A análise indica que os cancelamentos variam significativamente por região, tanto em volume, taxas e os motivos relacionados. Problemas técnicos e operacionais predominam na maioria dos estados enquanto razões financeiras e comerciais ganham relevância na região norte do país. Portanto, a estratégia de retenção e prevenção ao churn devem ser regionalizada e aplicada de forma individual para maximizar seus resultados.

Desconexão por Tecnologia



🔍 Análise:

- A tecnologia FTTC (Fiber to the Cabinet) possui uma taxa de desconexão significativamente maior que a FTTH. Desta forma, isso indica que clientes com FTTC são mais propensos a cancelar, sinalizando possíveis problemas com infraestrutura e técnicos e com isso gerando uma menor satisfação dos clientes.
- Os motivos macros TÉCNICO E FINANCEIRO se apresentam com uma ampla margem em ambas tecnologias. Isto demonstra que à necessidade de melhorar processos operacionais e financeiros.
- Em relação à análise de perfil médios dos clientes cancelados por tecnologia, é possível observar que clientes FTTC possuem um tempo médio na base 3x maior do que clientes FTTH. Isso pode indicar que são clientes mais antigos e que há uma certa possível resistência à migração para tecnologias mais novas (FTTH).
- A quantidade de cancelamentos em FTTH é muito maior, o que está alinhado ao maior crescimento da base recente dessa tecnologia.

🎯 Pontos de Atenção:

- Alta taxa de desconexão na FTTC (48%) e isto demanda ações imediatas, como:
 - Revisão da infraestrutura

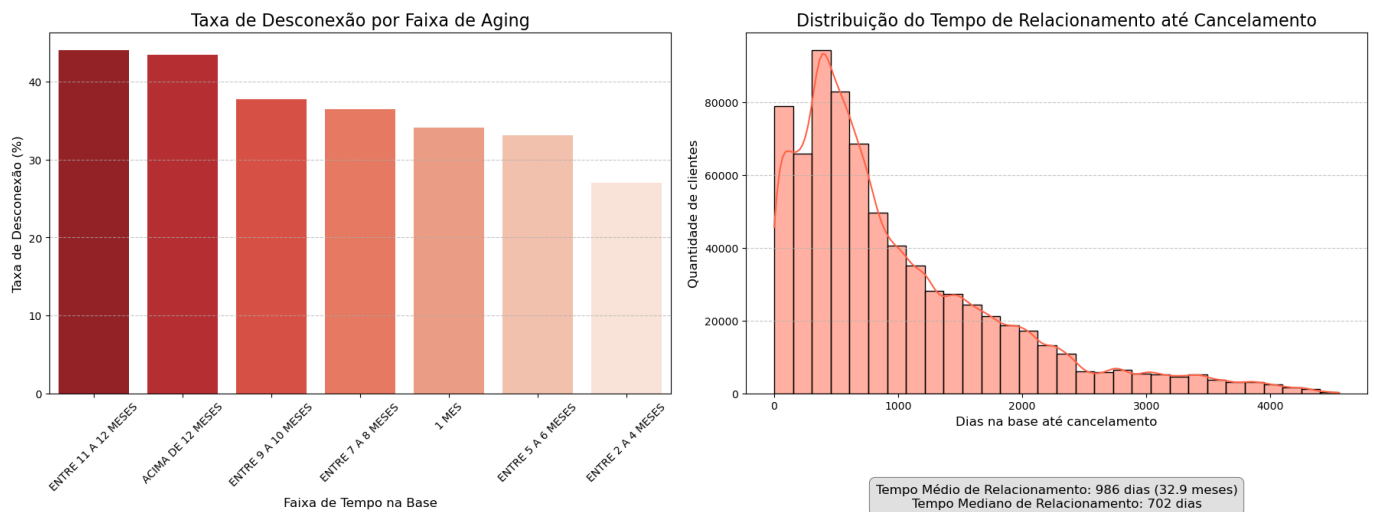
- Plano de Ação envolvendo a migração para FTTH.
- Retenção focada nesses clientes.
- Embora FTTH tenha taxa menor, o volume absoluto de cancelamentos é muito alto, indicando necessidades de melhorias na:
 - Experiência na instalação
 - Atendimento técnico
- Motivos técnicos seguem dominando. Isso reforça que, independente da tecnologia, há desafios operacionais e de qualidade na entrega do serviço.

✓ Conclusão:

A análise evidencia que, embora a tecnologia FTTH apresente uma taxa de desconexão menor que a FTTC, ela concentra o maior volume absoluto de cancelamentos. Os motivos técnicos são predominantes em ambas as tecnologias, reforçando a necessidade de melhorias operacionais. A priorização de estratégias para migração, qualidade de serviço e fidelização é essencial para reduzir os índices de cancelamento.

Desconexão por Aging

Análise de Desconexão por Tempo na Base

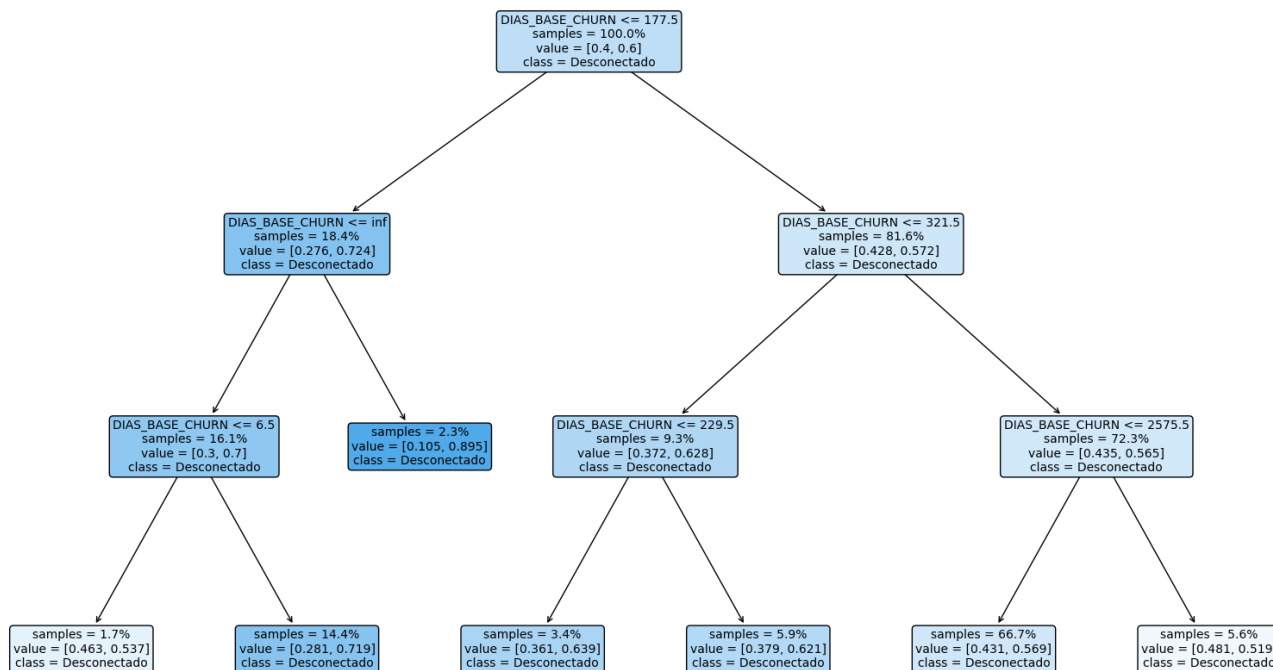


🔍 Análises:

- Clientes com poucos dias na base apresentam uma taxa de desconexão significativa, indicando problemas na primeira experiência do cliente.
- A maior taxa de desconexão ocorre nos clientes com 11 a 12 meses e acima de 12 meses, isto sugere uma relação com o término de contratos de fidelização ou um aumento de preço após o primeiro ano.
- A faixa com menor taxa de desconexão está entre 2 a 4 meses, mostrando que, nesse período, o cliente está mais propenso a permanecer e ainda está testando os serviços.
- De acordo, com a distribuição do tempo de relacionamento até o cancelamento tivemos como resultado que o tempo médio de um cliente até o cancelamento é de 986 dias. Desta forma, podemos concluir que a maioria dos cancelamentos ocorre entre 500 a 1000 dias de relacionamento que aproximadamente é de 1,5 a 3 anos. Após esse período, observa-se uma redução nos cancelamentos, mas o risco ainda permanece.

🧠 Aplicação do Modelo de Machine Learning DecisionTree (Árvore de Decisão):

Árvore de Decisão - Impacto do Tempo na Base na Desconexão



O modelo de Classificação Árvore de Decisão foi aplicado com o intuito de entender como o tempo de permanência do cliente na base (aging) impacta diretamente no comportamento de desconexão.

Desta forma, o modelo permite de forma visual, identificar os principais pontos de corte no tempo da base que estão associados a uma maior ou menor probabilidade de churn.

• Funcionamento do Modelo

- A variável preditora ou seja aquela utilizada para análise foi: DIAS_BASE_CHURN.
- O modelo particiona os dados em diferentes grupos, de acordo com faixas de tempo que melhor separam os clientes "Desconectados" e "Ativos".
- A cada divisão ele cria um nó. Portanto, ele busca maximizar a pureza dos grupos encontrados, isto é, ele identifica onde a proporção de desconectados é significativamente maior ou menor.

• Principais Insights Extráídos

1. Cancelamento Imediato

- Alta taxa de cancelamento muito precoce.
- Forte indicativo de problemas na ativação, instalação ou primeira experiência do cliente.

2. Pico de Desconexão no Ciclo de 11 a 12 meses

- A árvore mostra que o risco volta a aumentar fortemente próximo aos 321 dias até 2725 dias, evidenciando problemas associados ao término de contratos, aumento de preços ou insatisfação acumulada.
- Portanto, esse comportamento reflete claramente nos dados descritivos e nas taxas observadas nas faixas de aging.

3. Desconexão de clientes com Longo Prazo na Base - Acima de 7 anos.

- Em cliente com tempo de base mais longe ocorre uma leve tendência de estabilização, mas ainda assim com uma taxa de desconexão relevante.
- Isto sugere que, mesmo clientes muito antigos, se não forem bem assistidos, podem abandonar o serviço.

- **Contribuições do Modelo de Classificação**

- A árvore de decisão permitiu uma segmentação clara dos clientes com base no tempo de relacionamento, apontando exatamente os períodos críticos onde devemos atuar.
- Demonstrou estatisticamente os comportamentos já observados nas análises exploratórias e descritivas, comprovando que aging é, de fato, uma das principais variáveis de risco de churn.
- Facilitou a identificação de ações direcionadas, tanto para o onboarding quanto para retenção no fim do primeiro ano.

- **Recomendações Estratégicas**

- Ativação e um Onboarding Eficiente: Garantir uma excelente experiência nos primeiros dias para reduzir o cancelamento precoce.
- Relacionamento Contínuo: Monitorar a jornada de clientes de médio e longo prazo, com programas de fidelização e até mesmo acompanhamento consultivo.
- Monitoramento Preditivo: Implementar modelos preditivos, para podermos detectar clientes em risco antes da desconexão ocorrer.

- ☑ **Conclusão:**

A análise comprova que o churn não é um evento aleatório, mas altamente associado ao tempo de vida do cliente na base. A partir dos insights, torna-se necessário a empresa investir na experiência inicial do cliente, atuar de forma proativa na retenção e no desenvolvimento de estratégias comerciais e operacionais extremamente eficazes para retenção.

3.2 Análises referente a Retenção de Clientes

O objetivo desta análise é compreender o comportamento dos clientes no processo de retenção, identificando o volume de retenção, a taxa de retenção, os principais motivos macros e micros que levam o cliente a ser retido da base e principalmente a correlação com as demais features presentes na base.

Indicadores Principais

Dashboard de Análise de Retenção

1,842,797

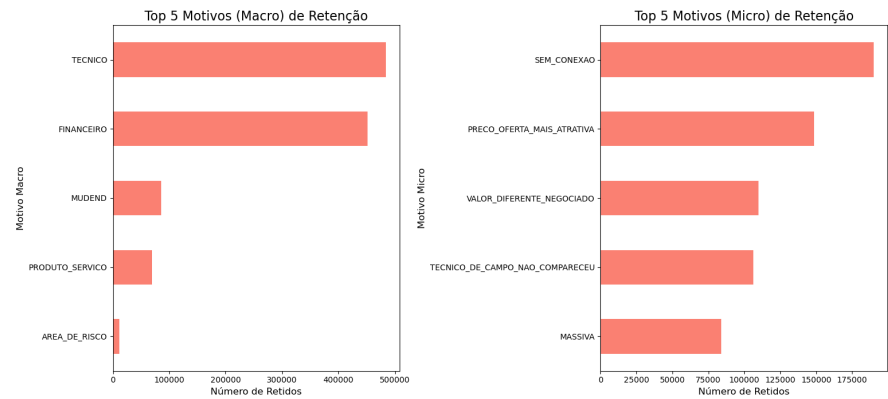
Volume Total de Solicitações

1,105,847

Total de Clientes retidos

60.01%

Taxa de Retenção



🔍 Análise:

- A taxa de retenção de 60% indica que a cada 10 clientes que tentam cancelar, 6 são convencidos a permanecer na base.
- As causas técnicas e financeiras são predominantes, somando cerca de 82% dos casos de retenção, indicando onde estão os maiores pontos de atuação para estratégias de fidelização.
- A maior concentração de retenção ocorre quando o cliente tem problemas operacionais, especialmente "SEM CONE~XÃO".
- Problemas financeiros como preço mais atrativo ou valores diferentes do negociado são altamente sensíveis e possuem uma grande taxa de sucesso na retenção, sugerindo uma margem pequena de oportunidades nas ofertas e negociações.
- Questões operacionais, como falta de comparecimento de técnicos, ainda aparecem como dores relevantes para o cliente.

🎯 Pontos de Atenção:

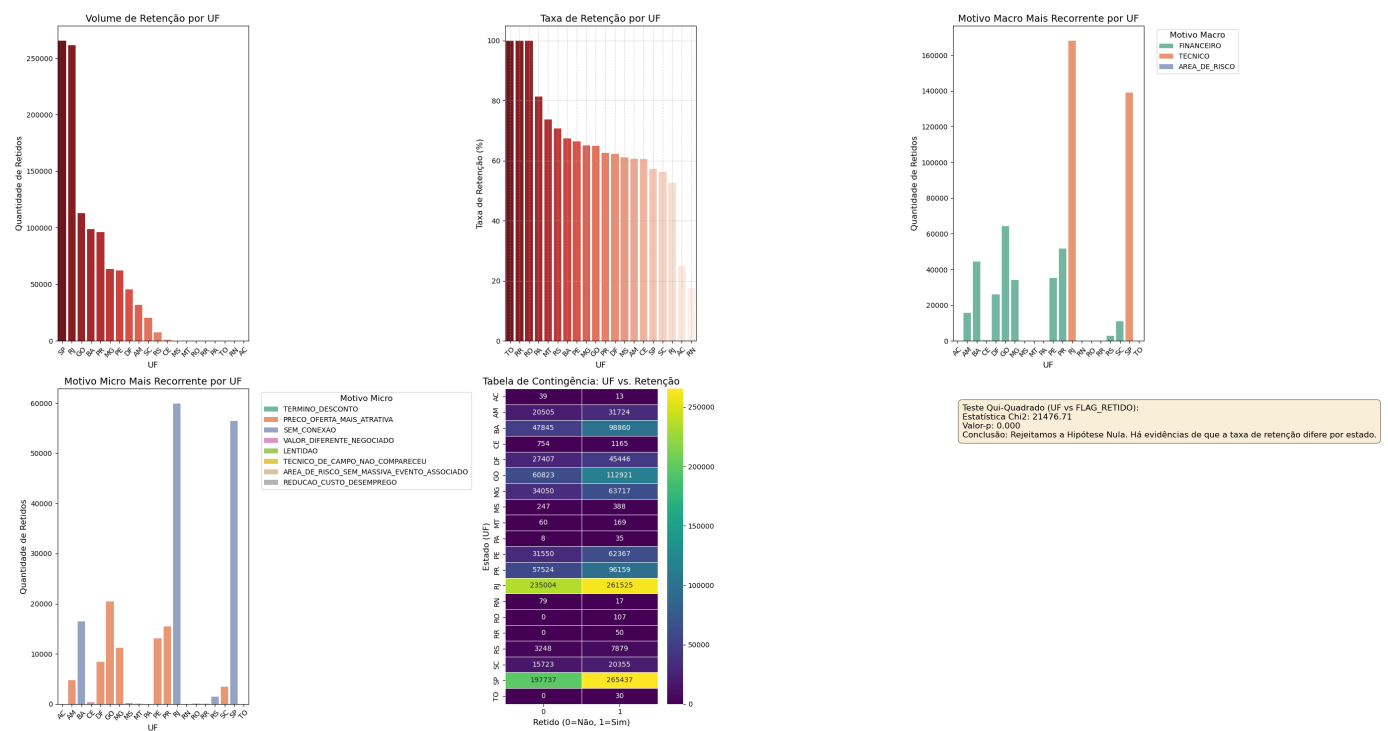
- Alta efetividade na retenção quando as causas são técnicas resolvíveis ou financeiras que permitem negociação.
- Processos comerciais e de relacionamento na venda estão conseguindo mitigar uma parte significativa dos cancelamentos.
- Divergências no valor contratado vs. valor cobrado são recorrentes e precisam ser endereçadas via revisão de processos comerciais e transparência na jornada do cliente.

✅ Conclusão:

A análise mostra que a retenção possui um papel fundamental na contenção do churn, com uma taxa de 60 de sucesso. Adotar ações preventivas na operação, melhorar os processos comerciais e fortalecer as ofertas de valor para o cliente são fatores decisivos para aumentar ainda mais essa taxa de retenção e consequentemente, proteger a base de clientes e a receita recorrente da empresa.

Retenção por UF

Análise de Retenção por UF: Insights Chave



Análise:

- Os estados mais populosos naturalmente apresentam maior volume de solicitações e, consequentemente de retenções. Isto foi identificado, a partir de percebemos que RJ lidera o volume absoluto de retenções, seguido por SP e MG, refletindo tanto a dimensão da base de clientes quanto a efetividade dos processos de retenção nesses locais.
- As taxas de retenção são inversamente proporcionais ao tamanho da base em muitos casos - Estados com menor base de clientes demonstram maior capacidade de retenção, o que pode estar associado a um atendimento mais personalizado já que a quantidade de clientes é menor.
- O motivo FINANCEIRO é predominante na maioria dos estados, isto evidencia uma maior sensibilidade na precificação e concorrência nesses mercados.
- O motivo TECNICO é predominante nos Estados de RJ e SP, evidenciando que problemas na qualidade do serviço e suporte tecnico são fatores criticos nesses dois estados.
- As dores dos clientes se distribuem entre problemas técnicos (instabilidade, desconexão e suporte falha) e questões financeiras (preço e divergência de valores).
- Existe dependência estatística significatva entre as UFs e a taxa de Retenção. Desta forma, a retenção não ocorre de forma uniforme no Brasil, havendo diferenças estatisticamente comprovada por região.

Pontos de Atenção:

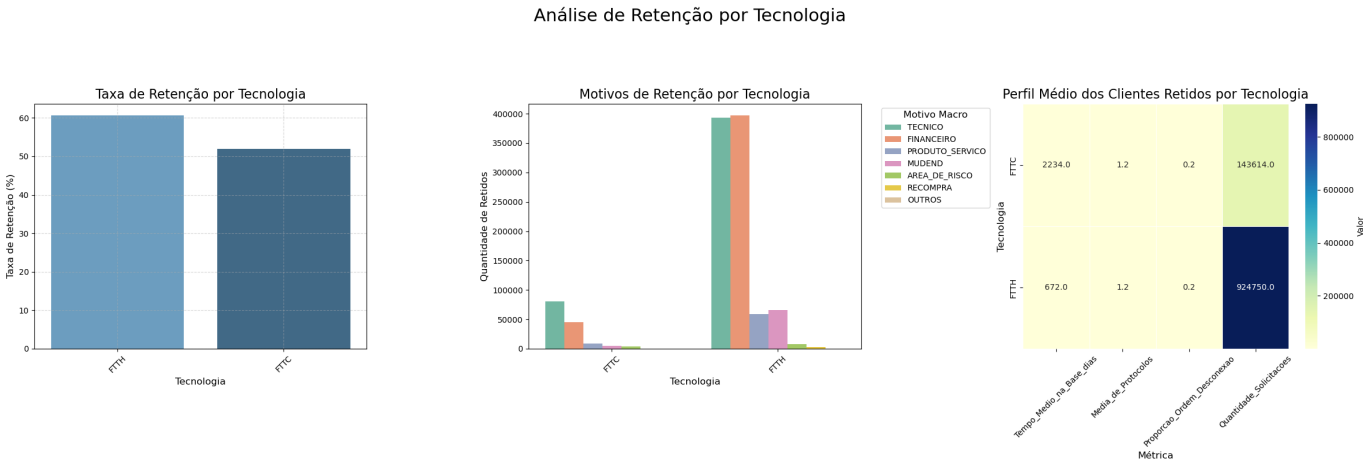
- Região Sudeste:
 - Altíssima pressão de mercado. Possuem clientes mais sensíveis tanto em relação à qualidade do serviço quanto aos preços e contratos. No entanto, as estratégias referente aos problemas técnicos, operacionais e comerciais precisam ser altamente eficiente e competitivos.
- Região Norte:
 - Altas taxas de retenção, geralmente associadas a bases menores e atendimentos mais personalizado ao cliente.

Conclusão:

A retenção não é homogênea entre os estados brasileiros. Ocorre variações claramente associadas a fatores como o perfil da base, condições operacionais e técnicas e até mesmo a sensibilidade dos clientes em relação aos preços.

Neste sentido, a atuação assertiva, direcionada e regionalizada é essencial para melhorar os índices e mitigar o churn.

Retenção por Tecnologia



Análise

- Clientes de FTTH são mais propensos a permanecer, devido à maior estabilidade e qualidade da conexão via fibra ótica.
- A tecnologia FTTC, oferece uma experiência inferior ao cliente e isso acaba impactando diretamente na satisfação e a retenção.
- Os volumes referente FINANCEIRO E TÉCNICO são equivalentes, e isso demonstra que tanto o preço quanto os problemas operacionais impactam nas decisões dos clientes.
- Os clientes FTTC possuem tempo médio altíssimo de 6anos, e isso é reflexo de uma base legada e com um taxa de retenção menor e isto acontece por um desgaste acumulado desses clientes por estarem usando uma tecnologia defasada.
- A base FTTH é mais recente, com um tempo médio de 22 meses, apresentando uma retenção mais eficiente, refletindo a aceitação de um serviço tecnicamente superior.

Principais Insights

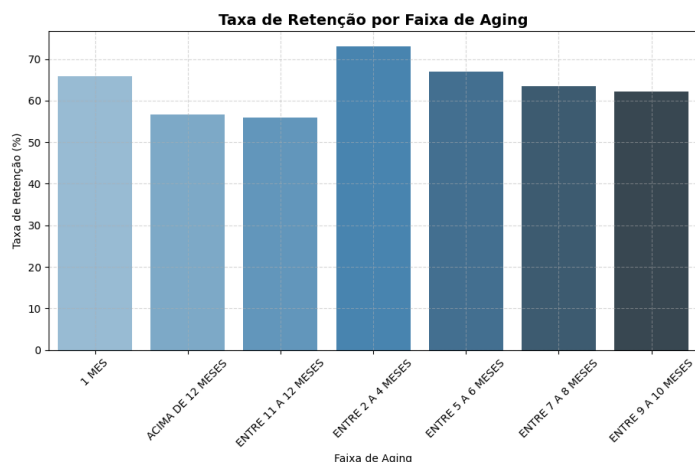
- A FTTH possui melhor retenção, apesar da base ser mais nova.
- FTTHC possui uma retenção mais desafiadora, especialmente por problemas técnicos.
- Os dois fatores principais de retenção são técnicos e financeiros, porém com pesos diferentes em cada tecnologia.

Conclusão

A tecnologia impacta diretamente na retenção de clientes. Já que FTTH apresenta melhores indicadores tanto de taxa de retenção quanto de crescimento sustentável e a FTTC, embora composta por clientes antigos, sofre com desafios operacionais que impactam negativamente a retenção. Portanto, a recomendação é ua estratégia de migração de clientes que são FTTC para FTTH complementando com ações de fidelizações para garantir a sustentabilidade da retenção no médio e longo prazo.

Retenção por Aging

Análise de Retenção por faixa de aging e Comportamento dos Clientes Retidos com uma Nova Tentativa de Cancelamento



Análise de Nova Tentativa de Cancelamento

0.2 dias

Tempo médio até nova tentativa

Tempo mediano: 0.0 dias
Total de clientes analisados: 147,712

🔍 Análise

- Clientes muito recentes (2 a 4 meses) possuem a maior taxa de retenção cerca de 73%, indicando alta eficácia das ofertas de retenção neste estágio inicial do cliente na base.
- Clientes com 11 a 12 meses e acima de 12 meses possuem a menor taxa de retenção (56% - 57%), evidenciando desgaste no relacionamento e maior resistência às ofertas de retenção.
- Há um comportamento interessante nos clientes de 1 mês (66%), que são menos fiéis que os de 2-4 meses, o que pode indicar uma fase de experimentação do serviço.
- A partir do card a direita do gráfico, o objetivo era responder a seguinte pergunta "Quanto tempo os clientes demoram, em média, para tentar cancelar novamente?" Desta forma, para podermos analisar isso e chegar nos melhores resultados foi considerado apenas aqueles clientes que foram retidos e que possuem +1 de um protocolo registrado. Portanto, ao interpretar os resultados podemos observar que o tempo médio para uma nova solicitação de desconexão é de 2 dias e isso nos mostra que uma parte relevante dos clientes que aceita retenção, podem estar aceitando apenas para obterem um benefício temporário ou evitar burocracias momentâneas. No entanto, isso também pode estar associado a ofertas de retenção que não são suficientemente robustas ou atrativas para sustentar o relacionamento no curto prazo.

🎯 Principais Insights

- Alta Retenção no início do ciclo (até 4 meses) -> Ofertas iniciais de retenção são eficientes.
- Retenção cai após 9 meses -> Necessário rever estratégias para clientes de longo prazo, talvez incluindo benefícios de fidelidade, upgrades ou atendimento mais personalizado.
- O comportamento de "tentativa imediata de novo cancelamento" é um sinal claro de que a percepção de valor da retenção precisa ser revista.

✅ Conclusões

- A retenção é mais efetiva em estágios iniciais do ciclo do cliente.
- Clientes antigos têm menor propensão à retenção, demonstrando a necessidade de estratégias robustas de relacionamento de longo prazo.
- O comportamento de tentativas imediatas após a retenção indica que há uma parte da base que não vê valor real na retenção oferecida, ou está explorando sistematicamente o processo para obter vantagens.

Essa análise reforça a necessidade de uma abordagem mais inteligente, baseada em dados, na definição das políticas de retenção e no design das ofertas, levando em consideração tanto o tempo de relacionamento (aging) quanto o histórico comportamental dos clientes.

4. Modelagem dos Dados com Aplicação de Técnicas de Machine Learning

Objetivo

O principal objetivo da modelagem foi a aplicação de **técnicas de aprendizado não supervisionado**, mais especificamente de **clusterização**, com o intuito em identificar padrões ocultos e agrupamentos dentro da base de clientes, de forma a compreender melhor os perfis e os comportamentos existentes.

Essa segmentação permite gerar insights sobre grupos de clientes com características semelhantes, possibilitando ações mais direcionadas, como retenção, campanhas personalizadas, melhorias operacionais e estratégias de negócio.

Escolha do modelo

O estudo aplicado neste projeto consiste em compreender e segmentar a base de clientes de acordo com seus comportamento, características contratuais, operacionais e padrões de retenção e desconexão.

Diferente de modelos supervisionados, onde existe uma variável - alvo conhecida (ex: prever se um cliente vai cancelar ou não), este cenário aplicado neste projeto não possui uma resposta pré - definida. O objetivo é descobrir grupos latentes dentro dos dados, ou seja, padrões ocultos que ajudam a entender como os clientes se agrupam naturalmente segundo suas semelhanças.

Certamente, o modelo escolhido foi o K-Means Clustering, um dos algoritmos mais populares e eficientes para tarefas de clusterização.

O K-Means é particularmente adequado para este projeto pelos seguintes motivos

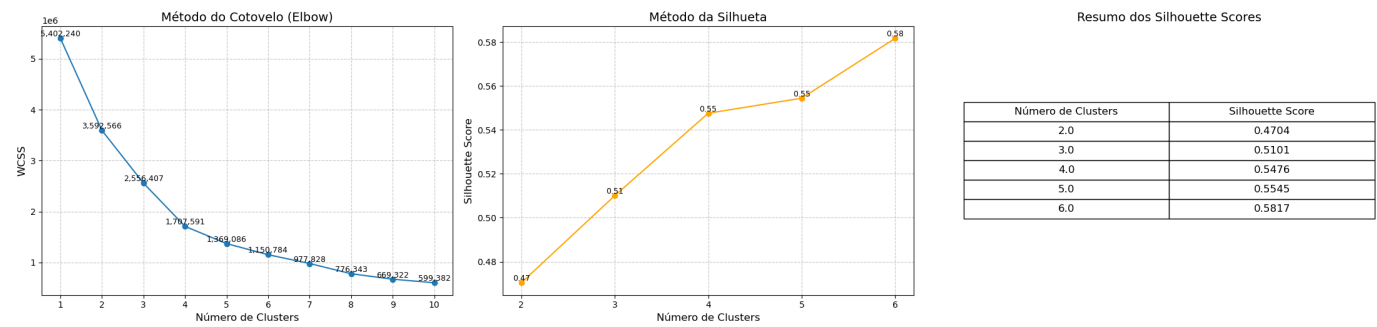
- **Simplicidade e Eficiência:** É um algoritmo leve, de rápida execução e com excelente escalabilidade, o que é fundamental dada a grande quantidade de registros na base de clientes.
- **Foco em Similaridade:** A definição dos clusters por meio de métodos como silhueta e cotovelo ajudam a criar grupos de clientes com comportamento operacionais e contratuais muito semelhantes - exatamente o tipo de agrupamento que este projeto busca identificar.
- **Aptidão para Dados Numéricos:** Como o algoritmo depende da aplicação de cálculos de distância envolvendo a distância euclidiana entre os diferentes clusters. É fundamental que a maior parte das variáveis - chave do modelo sejam numéricas, que a partir disso o desempenho do modelo seja mais eficiente e regular.

Pipeline de Desenvolvimento do Modelo

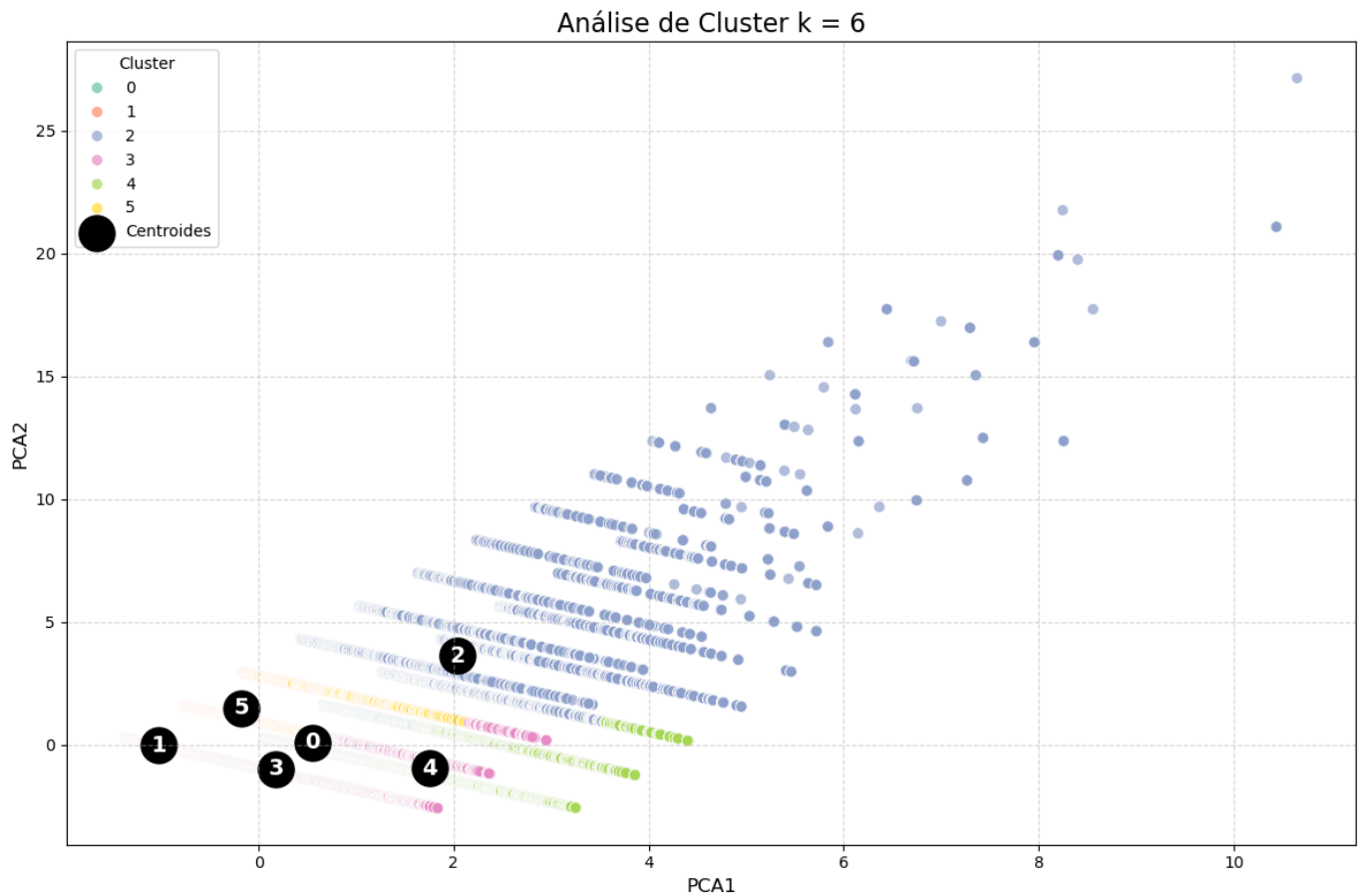
- **Definição do Número de Clusters**

Para determinarmos o valor ideal de clusters, foi aplicado dois métodos principais, denominados como Método do Cotovelo e Método da Silhueta:

Análise para Determinação do Número Ótimo de Clusters



- **Método do Cotovelo:** Foi utilizado com o objetivo de identificar o ponto no qual o aumento no número de clusters deixa de trazer uma redução significativa. Desta forma, a partir da análise desse gráfico, busca - se "cotovelo", ou seja, o ponto onde a curva começa a se estabilizar, indicando que o número de clusters além desse valor traz pouca melhoria na compactação dos grupos. Portanto, foi identificado que os clusters foca em torno de 6 grupos, sugerindo que esse valor oferece um equilíbrio adequado entre complexidade e coesão dos grupos.
- **Método da Silhueta:** A análise da silhueta foi aplicada para avaliar a qualidade dos agrupamentos formados. Desta forma, o método calcula um índice para cada ponto que vai variar entre -1 e 1, os valores que se aproximam de 1 indicam que o ponto está bem alocado no seu cluster, os valores próximos de 0 indicam que o ponto está na fronteira entre dois clusters e valores negativos (<0) sugerem que o ponto foi mal classificado. O gráfico acima, demonstra que o **Silhouette Score** aumentou progressivamente até K = 6, indicando uma separação consistente entre os grupos formados.
- **Informações referente ao treinamento do modelo**
 - Algoritmo: K - Means Clustering
 - Número de Clusters: 6
 - Variáveis utilizadas:
 - DIAS_BASE_CHURN -> Quantidade de dias que o cliente permaneceu na base até a data da solicitação
 - FLAG_ORDEM_DESCONEXAO -> Se já existe uma ordem de desconexão
 - QUANTIDADE_PROTOCOLOS -> Quantidade de Protocolos abertos pra aquele cliente
- **Análise dos Resultados da Clusterização**



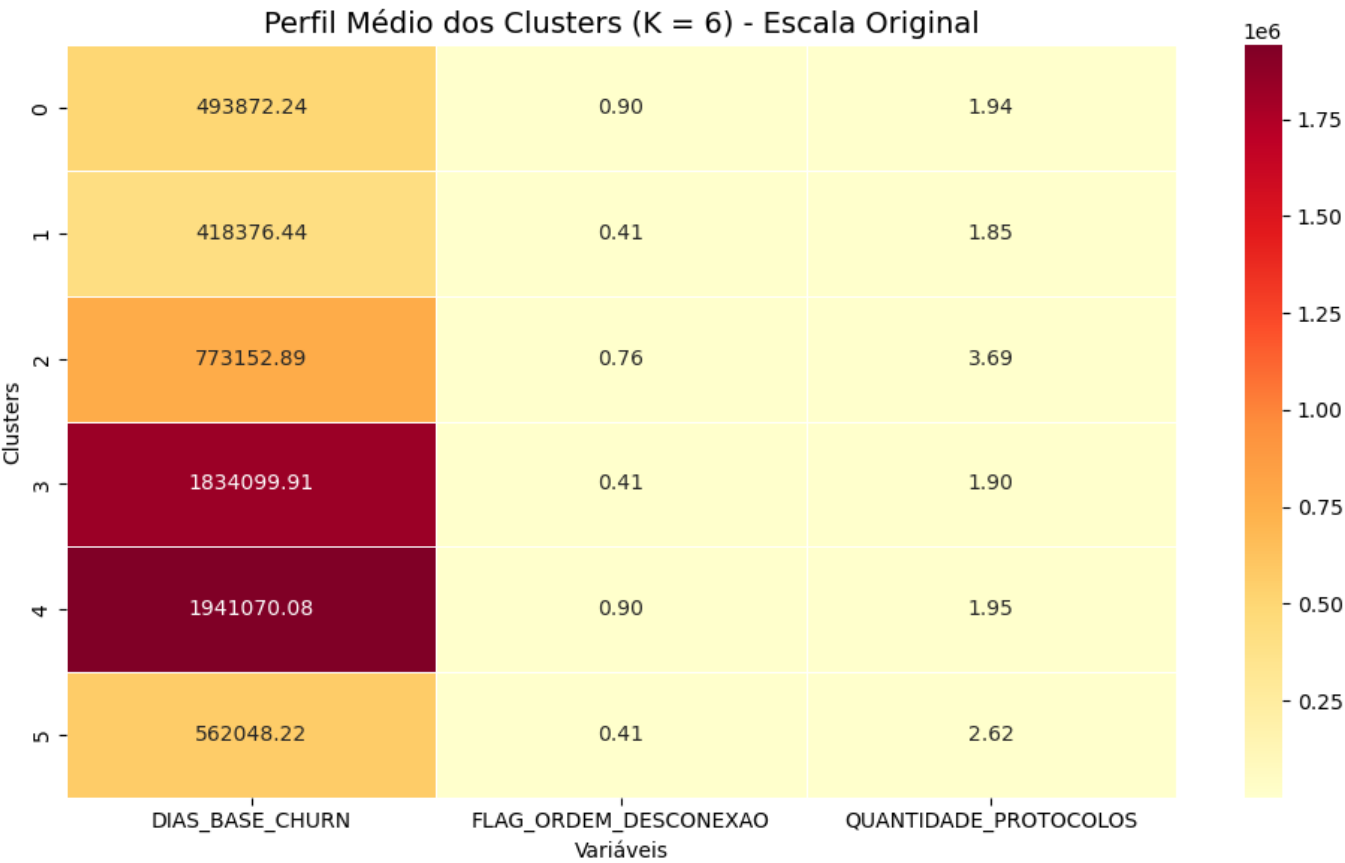
O objetivo desta etapa é analisar a distribuição dos grupos formados pelo algoritmo de clusterização (K - Means com $k = 6$) e entender os padrões encontrados na base de dados.

A representação gráfica foi realizada utilizando a técnica PCA (Análise de Componentes Principais), que reduz a dimensionalidade dos dados em dois componentes principais (PCA1 e PCA2) - permitindo visualizar graficamente os agrupamentos. No gráfico acima:

- Cada ponto representa um grupo de clientes.
- As cores indicam o cluster ao qual o cliente pertence (clusters 0 a 5).
- Os centróides dos clusters estão destacados em preto, representando a média dos atributos dos clientes de cada grupo.


Como interpretação dos resultados, temos:

- Observa -se que os clusters estão relativamente bem separados no espaço, com alguns agrupamentos mais compactos e outros mais dispersos.
- O **Cluster 2** aparenta ser o grupo com maior dispersão, indicando alta variabilidade dentro desse segmento.
- Clusters como 1,3 e 5 são mais compactos, indicando maior similaridade entre os clientes desses grupos.
- **Análise da Caracterização detalhada dos clusters**



Ao analisarmos os perfil dos Clusters, temos como resultados:

Cluster	DIAS_BASE_CHURN	FLAG_ORDEM_DESCONECAO	QUANTIDADE_PROTOCOLOS	Perfil
0	493.872	0.90	1.94	Clientes Novos com Alta Desconexão Tempo médio de base relativamente curto, alta incidência de ordens de desconexão e volume médio de interações/protocolos.
1	418.376	0.41	1.85	Clientes Novos e Estáveis Menor tempo de base, baixa taxa de desconexão e baixo volume de protocolos. Clientes mais tranquilos.

Cluster	DIAS_BASE_CHURN	FLAG_ORDEM_DESCONECAO	QUANTIDADE_PROTOCOLOS	 Perfil
2	773.152	0.76	3.69	Cientes Ativos e Problemáticos Tempo médio de base intermediário, alta proporção de desconexão e maior volume de protocolos (clientes mais demandantes).
3	1.834.099	0.41	1.90	Cientes Muito Antigos e Estáveis Longo tempo na base, baixa taxa de desconexão e volume médio de protocolos. Clientes fidelizados e de baixo risco.
4	1.941.070	0.90	1.95	Cientes Muito Antigos com Alto Risco Apesar do longo tempo de casa, possuem alta incidência de desconexão . Provavelmente estão insatisfeitos ou com risco iminente de churn.
5	562.048	0.41	2.62	Cientes Recorrentes e Moderadamente Demandantes Tempo de base intermediário, baixa desconexão, porém um número relativamente alto de protocolos.

Abaixo, estão os Insights Estratégicos por Cluster:

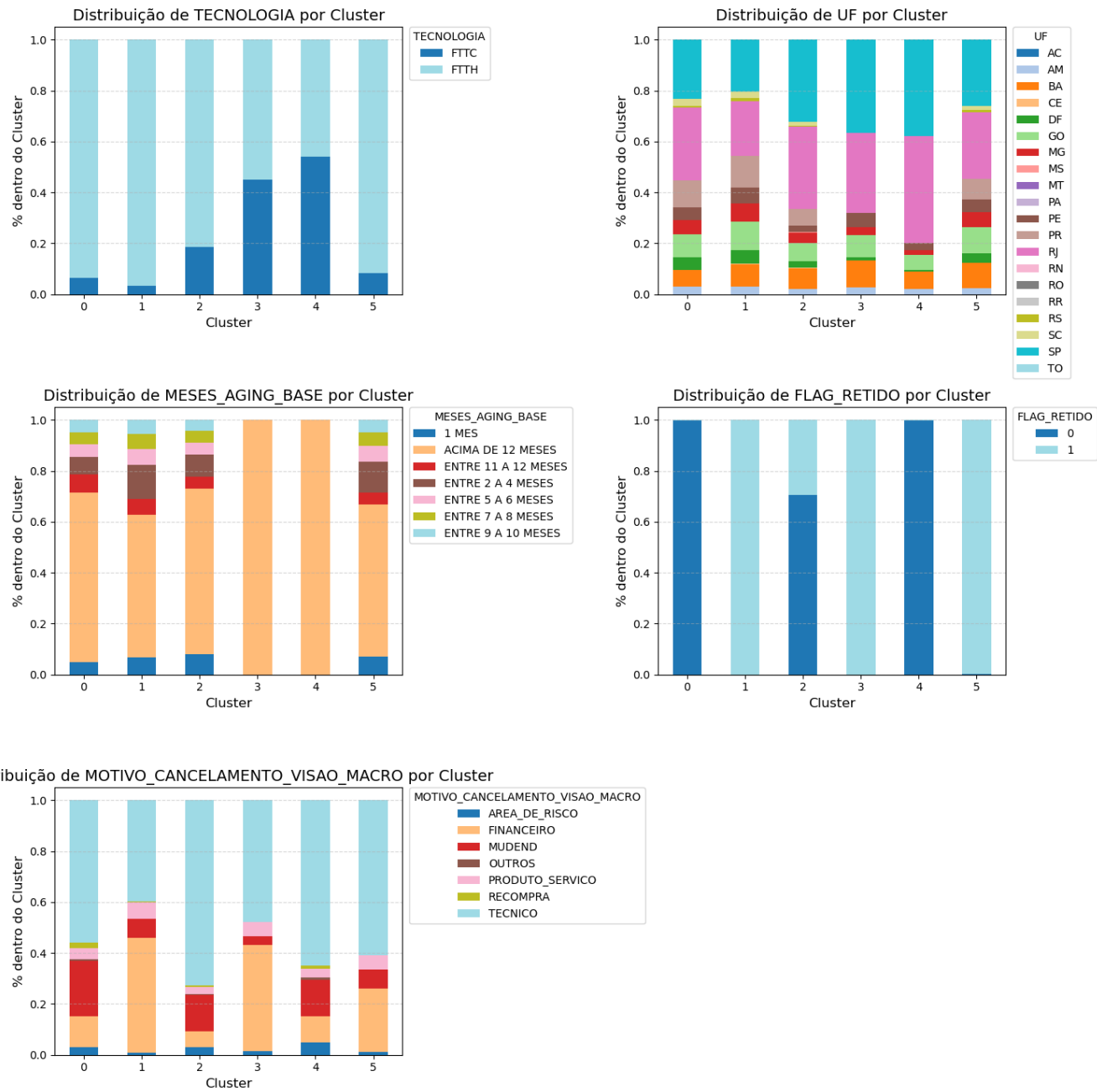
- **Cluster 2:** É o grupo mais crítico em termos de atendimento e insatisfação, pois além de apresentar alta taxa de desconexão, possui o maior volume de protocolo abertos.
- **Cluster 4:** Apesar de serem clientes muito antigos, apresentam alta propensão à desconexão. Pode ser sinal de desgaste com o serviço ou surgimento de ofertas mais atrativas no mercado. Ação prioritária de retenção aqui é essencial.
- **Cluster 3:** É o melhor grupo em termos de fidelização, com tempo médio muito alto, baixa desconexão e pouco acionamento. Desta forma, podemos concluir que são clientes leais e satisfeitos.

- **Cluster 0:** Clientes novos, mas com comportamento de alta desconexão logo no início. Indicando falhas na jornada inicial, onboarding ou primeiros atendimentos.
- **Cluster 1:** Perfil ideal de clientes recém - chegados e satisfeitos, com baixa desconexão e baixa demanda.
- **Cluster 5:** Clientes de médio tempo na base, com volume moderado de protocolos, merecem atenção para entender se estão migrando para um comportamento mais crítico (como o cluster 2) ou não.

Como conclusões e recomendações, temos definido:

- Realizar ações de retenção imediatas nos cluster 2 e 4, com foco em resolver problemas técnicos, revisar ofertas e melhorar experiência.
- Aplicar programas de fidelização e benefícios aos clientes do cluster 3, pois são os mais rentáveis e estáveis.
- Melhorar o processo de onboarding e ativação para os clientes do cluster 0, buscando reduzir a desconexão no início da jornada.
- Acompanhar os clusters 1 e 5, com monitoramento para evitar que migrem para clusters problemáticos.
- **Análise do comportamento dos Cluster com as demais features da base**

Distribuição das Features por Cluster



1. Distribuição por Tecnologia

- O **Cluster 3** se destaca por ter uma presença relevante de clientes na tecnologia FTTC, enquanto os demais clusters são majoritariamente FTTH.
- Portanto, isso sugere que o cluster 3 representa um perfil típico de clientes legados ou em regiões onde FTTC ainda é predominante.

2. Distribuição por UF

- A distribuição por estado é bastante equilibrada entre os clusters, refletindo a distribuição da base.
- Desta forma, não há uma concentração exagerada em algum estado específico.

3. Distribuição por Tempo de Base (MES_AGING_BASE)

- Clusters 3 e 4 possuem uma maior concentração de clientes "ACIMA DE 12 MESES", confirmando que são compostos por clientes muito antigos.
- Clusters 0,1 e 2 concentram clientes nas faixas "ENTRE 2 A 4 MESES", mostrando que são formados predominantemente por **Clientes mais novos ou de médio prazo**

4. Distribuição por Motivo Macro de Cancelamento

- Clusters 2 e 5 possuem maior proporção de cancelamentos por motivo Técnico, mostrando que esse grupo enfrenta mais desafios operacionais.
- Cluster 4, apesar de ser composto por clientes antigos, possui elevada incidência de motivos financeiros, indicando que mesmo clientes longevos podem estar sensíveis a preço ou questões econômicas.
- Cluster 1 é caracterizado por uma maior proporção de "Outros", possivelmente ligado a motivos diversos, como indisponibilidade ou motivos pessoais.
- O motivo "Área de Risco" aparece de maneira residual, sem forte concentração em nenhum cluster.

5. Distribuição por Status de Retenção ou Desconexão

- A retenção é relativamente alta e homogênea em todos os clusters.
- Contudo, pequenas variações indicam que alguns clusters estão mais associados a clientes com sucesso na retenção, enquanto outros podem ter mais casos de insucesso.

- **Ações Estratégicas Sugeridas para cada Cluster:**

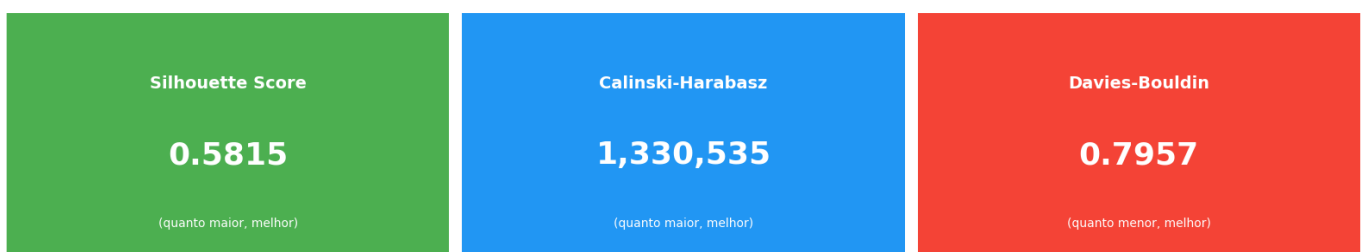
- Cluster 0: Focar na melhoria do onboarding e acompanhamento dos primeiros meses para reduzir desconexões precoces.
- Cluster 1: Manter ações de relacionamento e satisfação, pois são clientes saudáveis.
- Cluster 2: Visando melhorar o atendimento técnico desses grupos, devemos endereçar esses clientes para melhorar suas infraestrutura.
- Cluster 3: Avaliar necessidade de migração de clientes de FTTC para FTTH e oferecer upgrades.
- Cluster 4: Desenvolver ofertas financeiras, descontos ou pacotes que mitiguem risco de churn por preço.
- Cluster 5: Melhorar processos de atendimento técnico e acompanhamento de performance.

- **Avaliação da Performance do Modelo**

A avaliação de modelos de clusterização, diferentemente de modelos supervisionados, não se baseia em métricas de acurácia, precisão ou recall, e isso acontece porque não há rótulos pré - definidos em modelos não supervisionados. No entanto, utilizam-se métricas que avaliam coerência interna, separação entre grupos e qualidade estrutural dos clusters.

Foram utilizadas três métricas principais para avaliar a qualidade do modelo de clusterização KMeans:

Métricas de Avaliação do Modelo de Clusterização



1. Silhouette Score

- Resultado: 0.5815
- Interpretação: O Silhouette Score mede o quão bem cada ponto está relacionado com seu próprio cluster comparado ao clusters vizinhos. Portanto, o valor varia entre -1 e +1.
 - Próximo de +1: Boa separação entre clusters e alta coesão interna.
 - Próximo de 0: Clusters sobrepostos ou mal definidos.
 - Próximo de -1: Indica possível alocação incorreta dos pontos.
- Conclusão: O valor de 0.5815 é considerado bom em problemas de negócio complexos, como comportamento de clientes. Isso indica que os clusters são bem definidos e têm uma separação clara, especialmente considerando o volume e a natureza heterogênea dos dados.

2. Calinski-Harabasz Index (CH Index)

- Resultado: 1.330.535
- Interpretação: O CH Index avalia a dispersão entre os clusters (Separação) e a dispersão dentro dos clusters (Coerência Interna).
 - Quanto maior, melhor — indica clusters bem separados e compactos.
- Conclusão: Um valor extremamente alto como o resultado que foi gerado reforça que os clusters estão bem definidos, com baixa dispersão interna e boa separabilidade entre os grupos.

3. Davies - Bouldin Index (DBI)

- Resultado: 0.7957
- Interpretação: O DBI avalia a média da similaridade entre cada cluster e aquele que mais se assemelha a ele. Este índice considera tanto o tamanho dos clusters quanto a distância entre eles.
 - Quanto menor, melhor.
 - Valores próximos de 0 indicam boa separação entre clusters e baixa sobreposição.

✓ Conclusão da Avaliação:

- O modelo atende aos critérios de qualidade para ser implementado como parte da análise comportamental dos clientes.
- As métricas reforçam que o agrupamento tem valor prático e pode ser utilizado com segurança para estratégias de retenção, identificação de perfis de risco, personalização de ofertas e ações comerciais.
- A escolha de 6 clusters se demonstrou tecnicamente e visualmente a mais adequada, balanceando performance, granularidade dos perfis e capacidade interpretativa dos resultados.

✓ Conclusão Final do Projeto

Este projeto entregou uma análise robusta, profunda e orientada a dados, que proporciona uma visão estratégica dos clientes, seus comportamentos e seus riscos.

Através de uma combinação de análises descritivas, modelagem preditiva e segmentação via clusterização, foi possível transformar dados operacionais em informação acionável, fortalecendo a capacidade da empresa em tomar decisões mais inteligentes, personalizar suas estratégias e, principalmente, melhorar a experiência do cliente e reduzir o churn.

O trabalho realizado não só entrega valor imediato, como também constrói uma base sólida para futuros avanços em análise preditiva, inteligência de dados e melhoria contínua dos processos de negócio.

Tivemos como principais insights gerados:

- Existem perfis de clientes mais sensíveis a problemas técnicos, enquanto outros são mais sensíveis a questões financeiras ou operacionais.
- Clientes com poucos meses de contrato possuem risco elevado de desconexão, especialmente quando há falhas na experiência inicial (onboarding).
- Clientes antigos podem desenvolver resistência à retenção, principalmente quando enfrentam problemas recorrentes ou não percebem evolução no serviço.
- A retenção tem impacto direto sobre o ciclo de vida dos clientes, mas deve ser acompanhada de melhorias no serviço, suporte e comunicação.

E como próximos passos:

- Evoluir o modelo de clusterização incorporando variáveis comportamentais adicionais, como consumo de banda, uso de serviços e satisfação do cliente (NPS).
- Explorar modelos supervisionados para previsão de churn, utilizando os perfis dos clusters como features.
- Construção de dashboards dinâmicos que acompanhem em tempo real os indicadores de risco de desconexão, aging e retenção.