Relatório de Análise de Dados

Análise de Rotatividade de Clientes Bancários

Curso: Tecnologia em Análise e Desenvolvimento de Sistemas

Disciplina: Análise de Dados

Professor: Dr. Anisio Silva

Instituição: Instituto Federal de São Paulo (IFSP) – Câmpus Boituva

Autores: José Vinicius e Vitor Faustino

Repositório do GitHub:
https://github.com/vitao-bolado/Trabalho1_AnaliseDados/blob/main/Trabalho1.ipynb

1. Análise Inicial da Base de Dados

1.1. Descrição e Contexto

O presente trabalho tem como objetivo realizar uma análise completa da base de dados **"Base 06 – Rotatividade de Clientes Bancários"**. Nosso ponto de partida foi entender o contexto do problema: uma instituição bancária que deseja compreender os motivos que levam seus clientes a encerrar o relacionamento (churn). A base contém 10.002 registros e 14 variáveis com informações sobre os clientes.

O foco da nossa análise foi:

- Identificar os fatores que influenciam a saída dos clientes.
- Comparar o perfil dos clientes que saíram com os que permaneceram.
- Construir modelos preditivos simples para avaliar o churn.

1.2. Tratamento dos Dados

Para garantir a qualidade da análise, o primeiro passo foi o tratamento dos dados.

Variáveis Removidas: As colunas RowNumber, CustomerId e Surname foram removidas, pois são apenas identificadores e não possuem valor para a análise preditiva de comportamento.

Dados Nulos: Ao investigar a base, encontramos uma pequena quantidade de dados nulos:

Variável	Quantidade de Nulos
Geography	1
HasCrCard	1
Age	1
IsActiveMember	1

Como apenas 4 registros continham valores faltantes, optamos pela estratégia de exclusão dessas linhas (dropna()). Essa decisão foi tomada pois o volume de dados removidos é estatisticamente insignificante e não compromete a representatividade da amostra.

2. Análise Estatística e Visual

2.1. Análise Descritiva

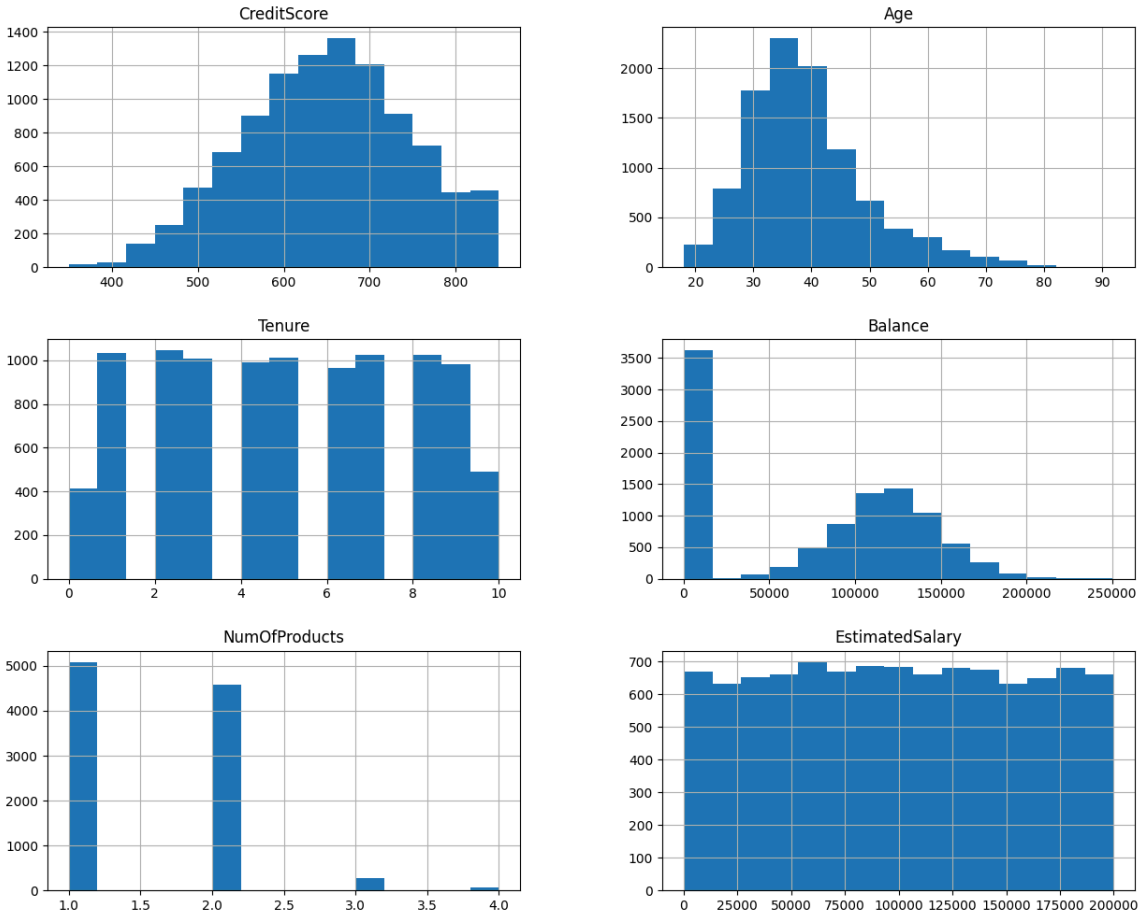
Com a base de dados limpa, partimos para a análise estatística descritiva das variáveis numéricas, que nos deu os seguintes insights:

	CreditScore	Age	Tenure	Balance	NumOfProducts	EstimatedSalary
count	9998.00	9998.00	9998.00	9998.00	9998.00	9998.00
mean	650.53	38.92	5.01	76481.49	1.53	100099.79

std	96.63	10.49	2.89	62393.19	0.58	57510.94
min	350.00	18.00	0.00	0.00	1.00	11.58
25%	584.00	32.00	3.00	0.00	1.00	50983.75
50%	652.00	37.00	5.00	97173.29	1.00	100218.21
75%	718.00	44.00	7.00	127641.42	2.00	149395.88
max	850.00	92.00	10.00	250898.09	4.00	199992.48

2.2. Distribuição das Variáveis e Outliers

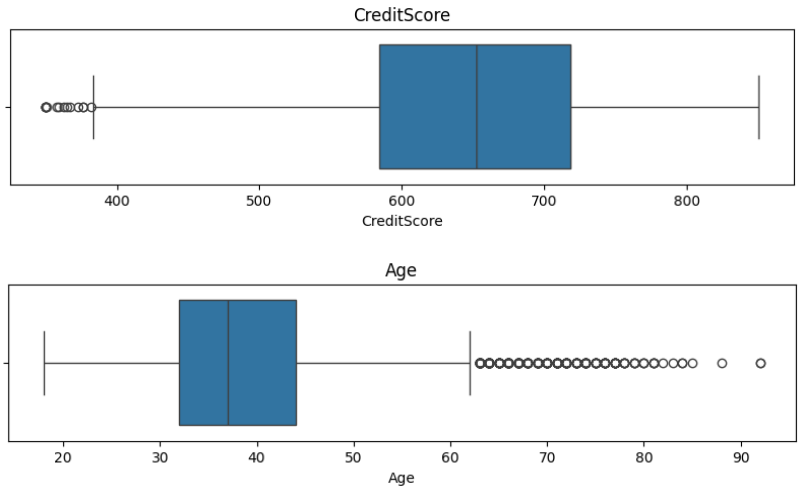
Na parte visual da análise, geramos **histogramas** para entender a distribuição dos dados e **boxplots** para investigar a presença de outliers.

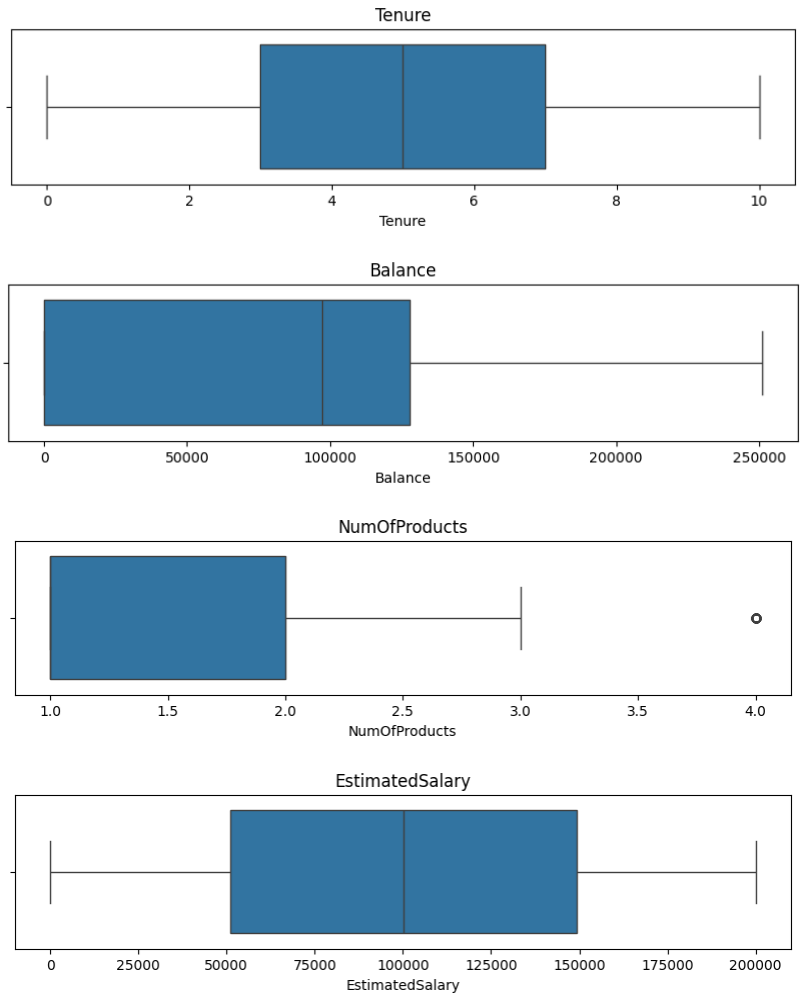


Observamos que:

- A variável **Age** se destacou por ter uma concentração maior de clientes na faixa dos 30 a 45 anos.
- A variável **Balance** chamou a atenção por sua distribuição com dois picos: um grande número de clientes com saldo zerado e outro grupo com saldo elevado.

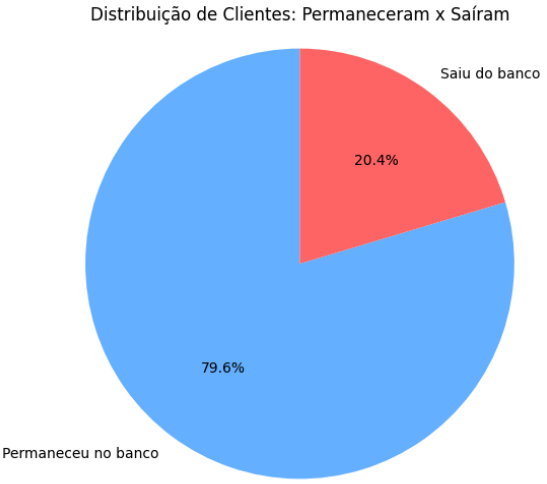
Os boxplots confirmaram a presença de alguns outliers, principalmente na variável Age, indicando clientes com idade mais avançada que o padrão.





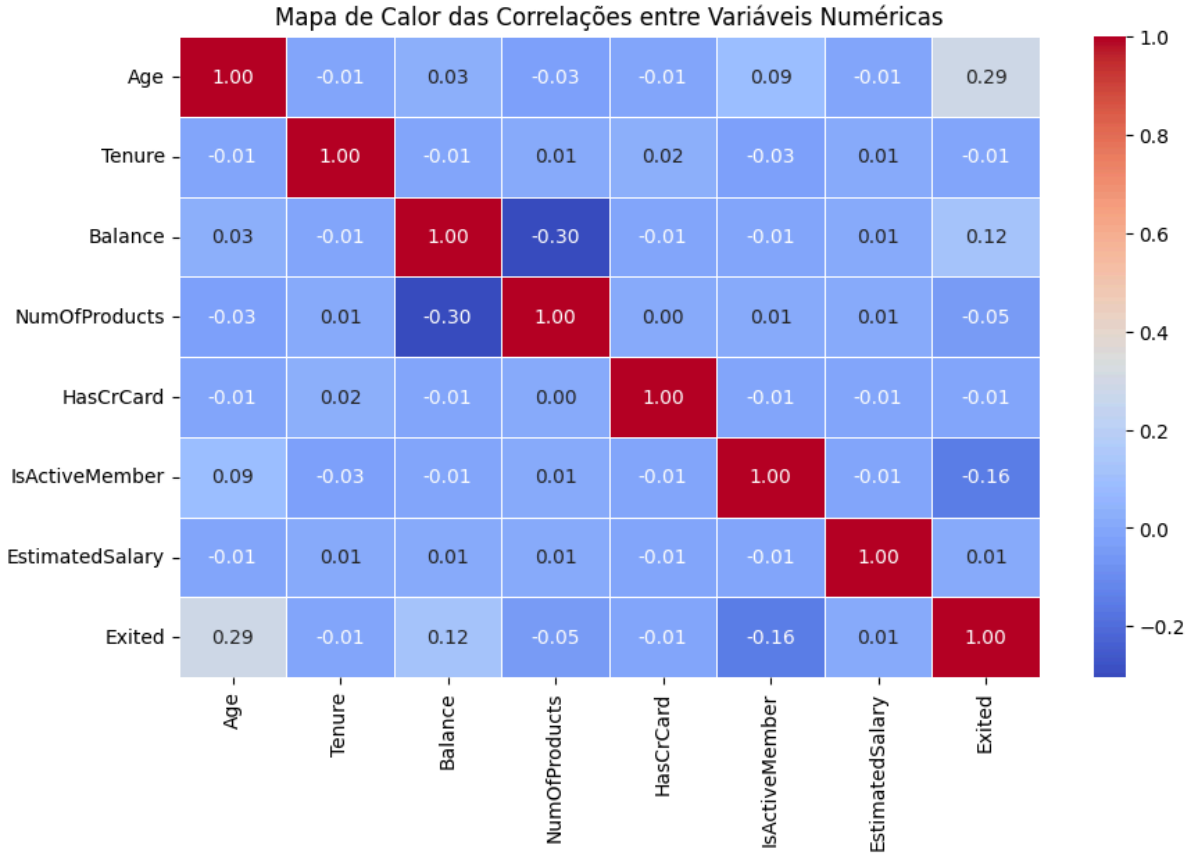
2.3. Proporção Geral de Churn

Antes de aprofundar nas hipóteses, foi importante visualizar a proporção geral de clientes que deixaram o banco. O gráfico de pizza abaixo mostra que **20.4%** dos clientes da base analisada encerraram seus serviços, um número significativo que justifica a investigação.



2.4. Mapa de Correlação

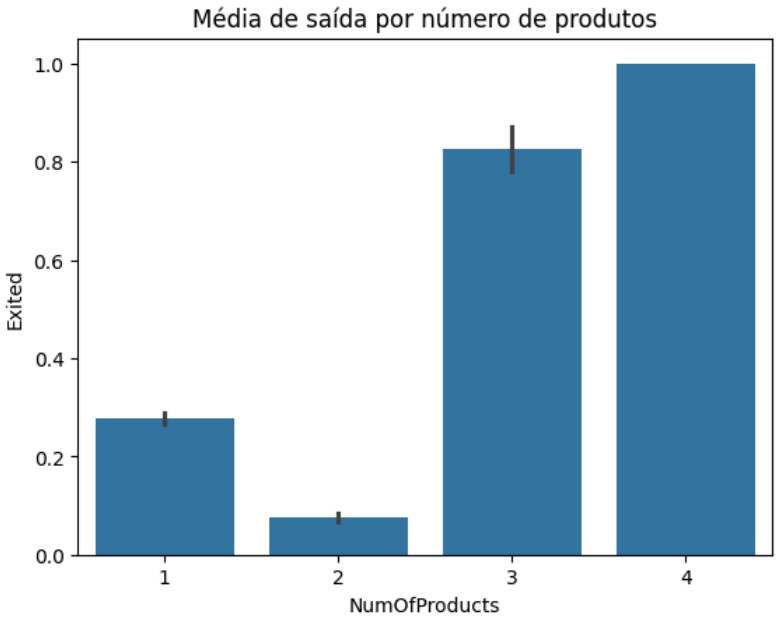
O mapa de calor nos ajudou a visualizar as correlações entre as variáveis numéricas. Um dos achados mais interessantes foi que a variável **Age** apresentou a maior correlação positiva (0.29) com Exited, enquanto **IsActiveMember** teve a correlação negativa mais forte (-0.16).



3. Formulação e Teste de Hipóteses

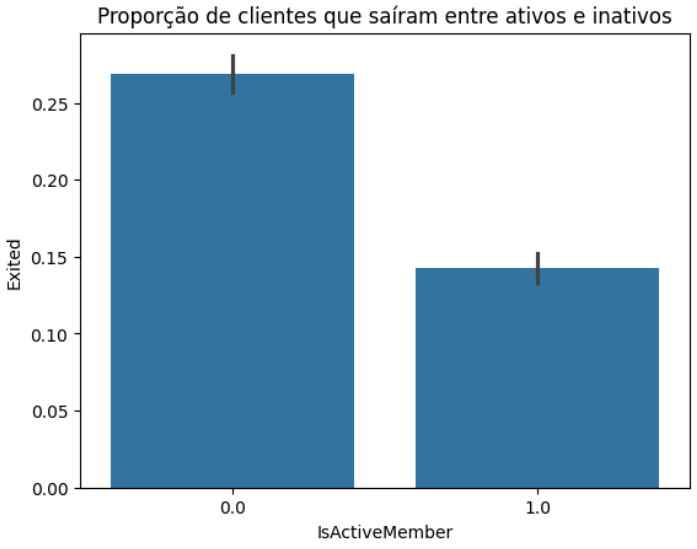
A análise exploratória nos permitiu levantar algumas hipóteses sobre o comportamento dos clientes. Formulamos e testamos três hipóteses principais:

Hipótese 1: Clientes com apenas 1 produto bancário saem mais do que os demais.



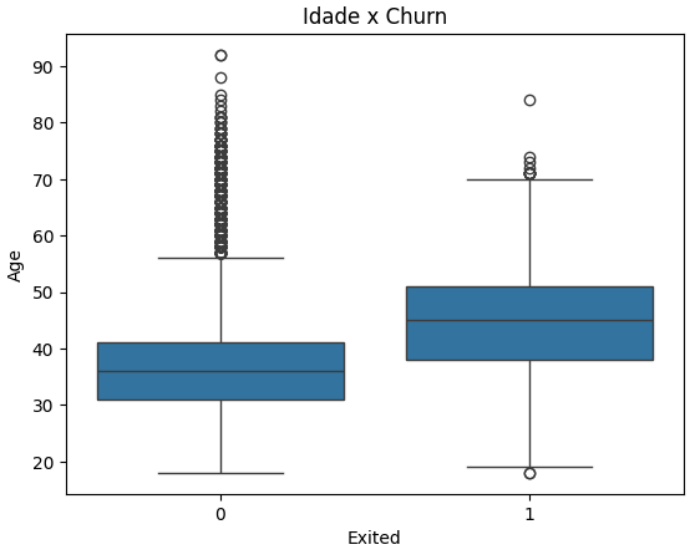
- **Análise Visual:** O gráfico de barras comparativo mostrou que a média de saída para clientes com 3 ou 4 produtos é drasticamente maior.
- **Teste Estatístico (Qui-Quadrado):** O teste Qui-Quadrado resultou em um **valor-p extremamente baixo (2.41e-76)**.
- **Conclusão:** Com isso, a hipótese é confirmada. Existe uma associação estatisticamente significativa entre o número de produtos e a saída do cliente.

Hipótese 2: Clientes ativos (IsActiveMember = 1) permanecem mais no banco.



- **Análise Visual:** A proporção de saída entre clientes inativos foi quase o dobro da observada em clientes ativos.
- **Teste Estatístico (Qui-Quadrado):** O teste nos deu um **valor-p de 9.83e-55**, um resultado muito expressivo.
- **Conclusão:** Este resultado confirma nossa hipótese: o engajamento do cliente tem uma forte associação com sua permanência.

Hipótese 3: Idade elevada está associada à maior chance de saída.



- **Análise Visual:** O boxplot revelou que a mediana da idade dos clientes que saíram é visivelmente maior.
- **Teste Estatístico (Teste t de Student):** O teste de diferença de médias de idade resultou em um **valor-p de 9.15e-187**.
- **Conclusão:** A diferença é estatisticamente significativa, validando a hipótese de que a idade é um fator relevante.

4. Análise Preditiva

Na etapa de modelagem, o objetivo foi investigar se era possível prever o churn. Para isso, implementamos dois modelos de regressão linear.

4.1. Regressão Linear Simples

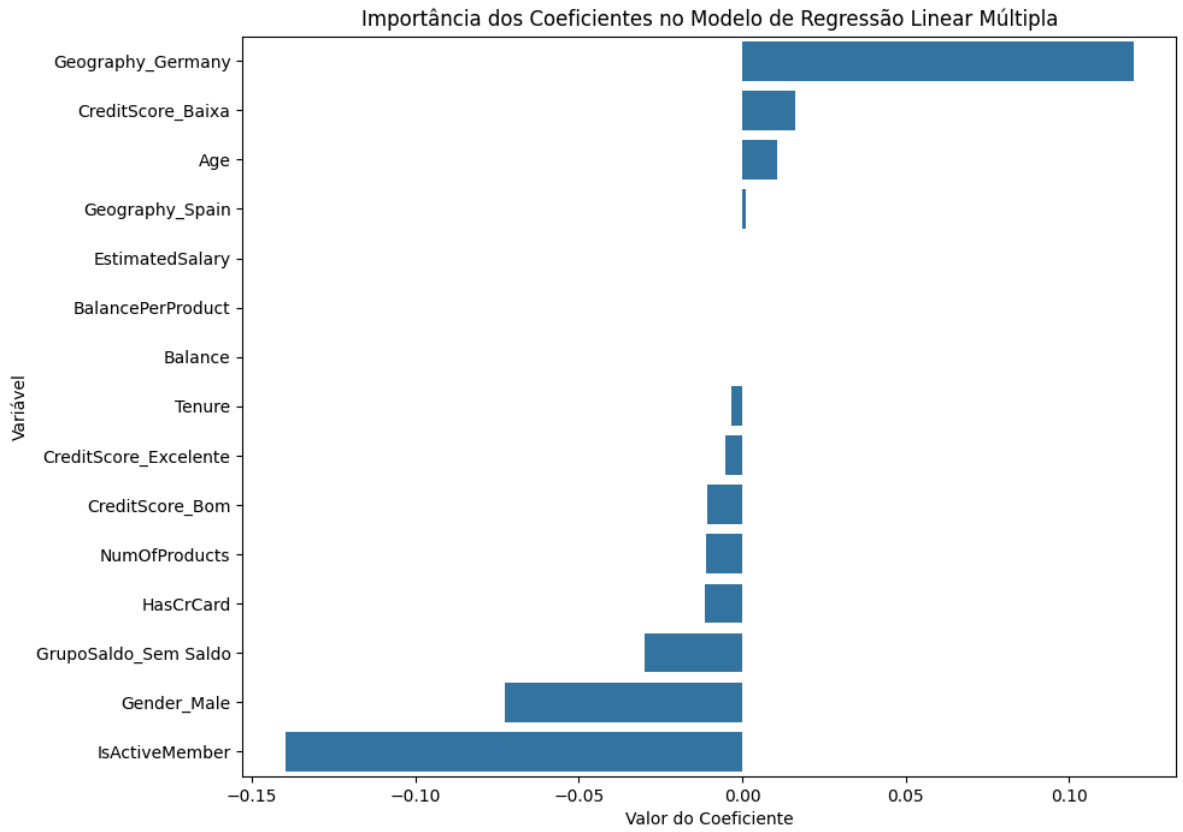
Usando apenas a Age para prever Exited, o modelo obteve um **R^2 de 0.090**, indicando que a idade sozinha explica apenas 9% da variabilidade no churn.

4.2. Regressão Linear Múltipla

Incluindo todas as variáveis, o modelo teve um desempenho melhor, com **R^2 de 0.152**. Apesar da melhora, o valor ainda é baixo, sugerindo que o churn é um fenômeno complexo.

4.3. Análise do Modelo Preditivo

Para entender quais fatores o modelo considerou mais importantes, plotamos os seus coeficientes.



O gráfico confirma que Age tem o maior impacto positivo na previsão de churn, enquanto IsActiveMember tem o maior impacto negativo.

Nota Metodológica: É importante notar que a regressão linear não é a ferramenta ideal para um problema de classificação (saída "sim" ou "não"). Em uma análise mais aprofundada, modelos como a Regressão Logística seriam mais apropriados.

5. Conclusão e Recomendações

5.1. Conclusões Gerais

Após todas as etapas de análise, foi possível concluir que:

1. **Idade, status de atividade e número de produtos** são os fatores mais influentes para prever o churn.
2. Clientes com **idade mais avançada e inativos** compõem o perfil de maior risco de saída.
3. **Pontuação de crédito e salário estimado** não mostraram correlação forte com a decisão de churn.

5.2. Recomendações Estratégicas

Com base nessas conclusões, elaboramos as seguintes recomendações:

1. **Programa de Retenção para Clientes de Maior Idade:** Desenvolver estratégias específicas para clientes mais velhos, como atendimento personalizado ou produtos adaptados.

2. **Campanhas de Engajamento para Clientes Inativos:** Criar um programa para reativar clientes, oferecendo incentivos para aumentar a frequência de uso dos serviços.
3. **Estratégia de Cross-Selling:** Focar em clientes com apenas um produto para incentivá-los a adquirir novos serviços, fortalecendo o relacionamento com o banco.
4. **Aprimoramento da Coleta de Dados:** Investir na coleta de dados qualitativos (como pesquisas de satisfação) para permitir análises futuras mais robustas.