Relatório Técnico: Previsão de Churn em Beneficiários de Plano de Saúde

31 de julho de 2025

CIENTISTA DE DADOS JÚNIOR (INCLUSIVA PARA PERFIS DA DIVERSIDADE)

Rebeca Chuffi Saccochi

Introdução

O objetivo deste estudo é **identificar os fatores relacionados ao cancelamento de beneficiários de um plano de saúde** e **gerar insights** que ajudem a reduzir o **churn** - taxa de rotatividade, ou seja, perda de clientes de uma empresa durante um período específico. Utilizaremos modelos de Machine Learning treinados em um *dataset* fornecido pela Cassi.

Metodologia

A metodologia adotada foi baseada nas seguintes etapas:

1. **Exploração e Limpeza dos Dados**: Analisamos o dataset para entender as variáveis, tratar valores ausentes ou fora do domínio.
2. **Pré-processamento**: Foram realizadas transformações nas variáveis categóricas (com Label Encoding) e variáveis numéricas (com normalização). Além disso houve criações de colunas específicas que nos ajudasse a identificar os fatores explicativos de churn.
3. **Divisão do Dataset:** O dataset foi dividido em treinamento e teste, utilizando a função *train_test_split*.
4. **Modelagem:** Três modelos de machine learning foram testados:
    - Random Forest: Para prever o churn, com base em várias variáveis do dataset.
    - Naive Bayes: Considerando a distribuição probabilística das variáveis.
    - SVM (Support Vector Machine): Modelo para classificar os dados de maneira eficiente.
5. **Avaliação**: As métricas de precisão, recall e F1-score foram usadas para comparar o desempenho dos modelos. Além disso, curvas ROC também foram geradas, além de matrizes de confusão e correlação.

Além disso, as bibliotecas utilizadas foram:

- **Pandas**: Manipulação e análise de dados.
  NumPy: Operações matemáticas e álgebra linear.
  Scikit-learn: Modelos de aprendizado de máquina e avaliação (Random Forest, Naive Bayes, SVM).
  Matplotlib: Geração de gráficos.
  Seaborn: Visualizações avançadas.

Tratamento e limpeza de dados

Atributos

O dataset contém as seguintes variáveis (101.063 instâncias)

- ID_CLIENTE: identificação única do cliente (chave primária)
- CANCELADO: Verifica se o plano foi cancelado ou não.
- TITULARIDADE: se é titular ou dependente do plano
- FAIXA_RENDA: baixa, média ou alta renda
- TEMPO_DE_PLANO_MESES:
- IDADENAADESÃO: qual idade na época da adesão
- TEMPO_DE_PLANO_MESES: há quantos meses o tem o plano
- SEXO: sexo
- UF: localidade
- INADIMPLENTE: se está inadimplente (sim ou não)
- QTD_CONSULTAS_12M: quantas consultas nos últimos 12 meses
- QTD_INTERNACOES_12M: quantas internações nos últimos 12 meses
- VALOR_MENSALIDADE: valor da mensalidade paga

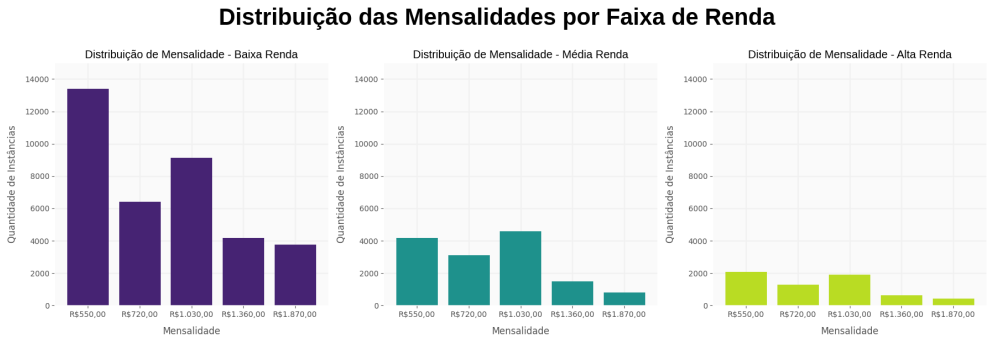
Tratamento e limpeza de dados

A parte de tratamento e limpeza de dados foi a mais delicada, pois foram necessárias várias transformações:

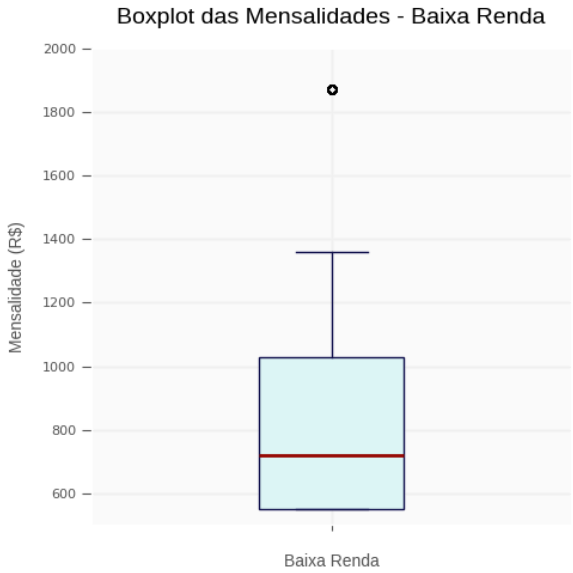
1. **ID_CLIENTE:** Verificamos a existência de várias duplicatas num campo que é uma chave primária. 6.565 instâncias foram deletadas
2. **SEXO:** 1855 instancias estavam com dados faltantes na coluna “SEXO”. Como faríamos análises relacionadas a esse atributo e apenas 1.96% do dataset total estava com dados ausentes, decidimos excluir tais instâncias. Poderíamos substituir os valores pela moda (Masculino), mas como as classes já estavam desbalanceadas, isso geraria um desbalanceamento maior.
3. **UF:** Assim como no atributo anterior, tínhamos classes desbalanceadas e uma porcentagem pequena do dataset total com dados faltantes, portanto decidimos excluir as 1878 linhas com valores faltantes.
4. **VALOR_MENSALIDADE:** Esse foi um atributo em que tivemos que fazer uma escolha diferente, pois 36.8% tinham valores faltantes, porcentagem que não pode ser excluída sem gerar danos ao modelo. Utilizamos Machine Learning para preencher os dados

faltantes dessa coluna. Antes de realizar o treinamento do modelo, analisamos alguns fatores:

- Criamos essa visualização para analisar a distribuição de Mensalidades por Faixa de Renda (note que apesar da mensalidade ser um atributo numérico, aqui trataremos como categórico, visto que temos apenas 5 possibilidades no total)



- Analisamos também o boxplot da distribuição da Baixa Renda para entender se precisaríamos mesmo de Machine Learning para fazer esse preenchimento.



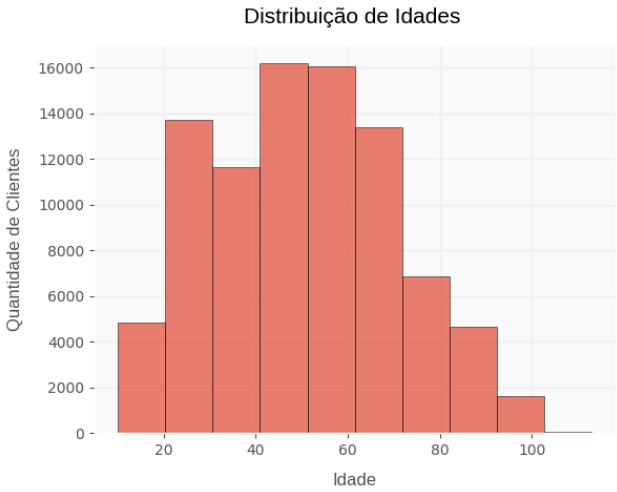
- Verificamos que 21455 tinham dados faltantes em Baixa Renda, 8192 em Média Renda e 3743 em Baixa Renda, ou seja, seria necessário um estudo mais procurando para preencher esses dados. Mediana e Média não seriam uma possibilidade, visto que dentro de cada classe tínhamos uma heterogeneidade de valores de mensalidades.

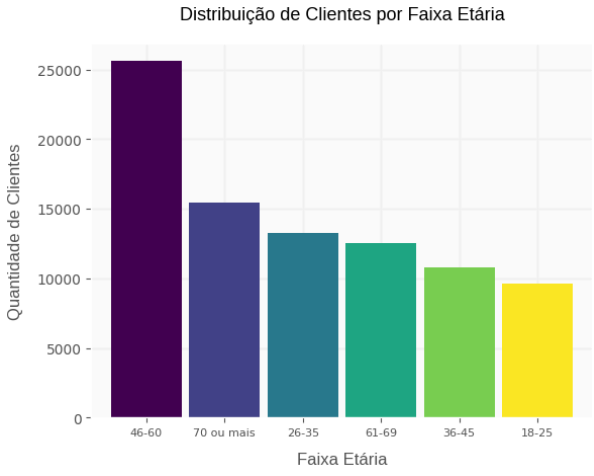
- Tentamos utilizar o atributo IDADE NA ADESÃO para tentar cruzá-lo com as Faixas de Renda, mas não foi suficiente. Precisamos utilizar Machine Learning.
- Inicialmente tentamos Random Forest, mas o mesmo estava apresentando overfitting (as métricas de avaliação estavam quase perfeitas) - isso é comum em alguns conjuntos de dados grandes. Tentamos usar validação cruzada, mas o problema permaneceu. Tentamos também normalizar os dados, mas não teve efeito diferente.
- Optamos por usar um momento um pouco Naive Bayes como segunda opção, e conseguimos um modelo que funcionou bem para o nosso propósito (F1-score ponderada de 0.8406)

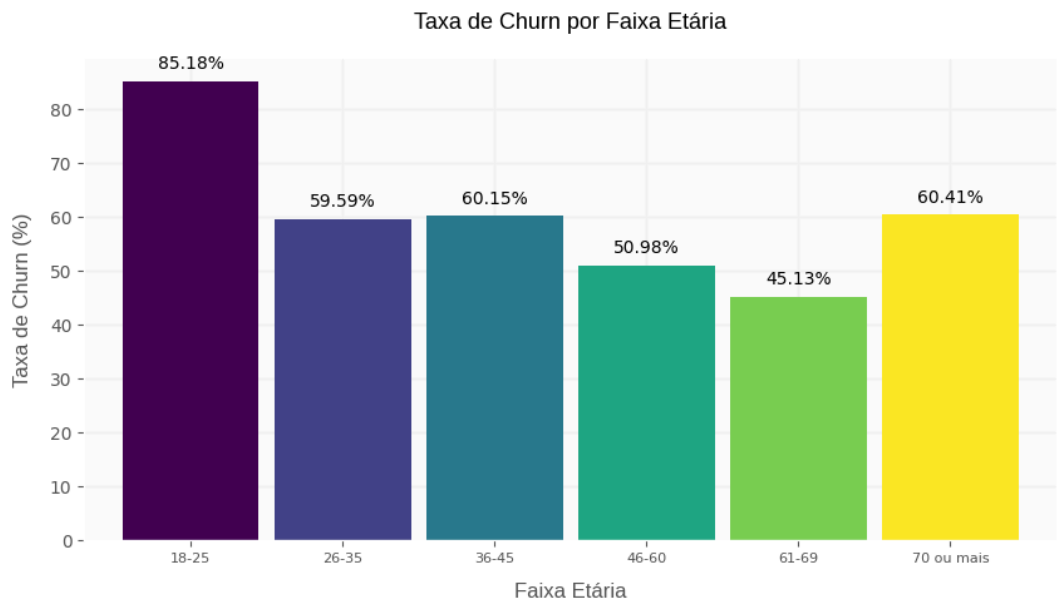
5. **IDADE e TITULARIDADE:** Criamos uma outra coluna de IDADE ATUAL considerando a idade adesão e tempo de plano. Além disso, excluímos todos as idades negativas ou maiores que 116 (dado da pessoa mais velha do mundo. Sabemos também que não é possível ser titular com menos de 18 anos, então assumimos que os clientes com menos de 18 anos eram DEPENDENTES.

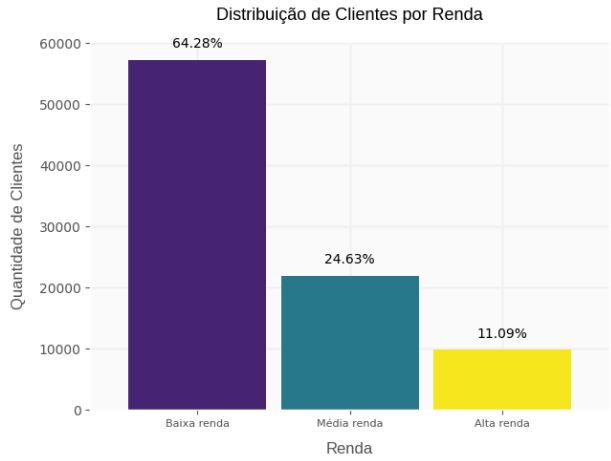
Comparação entre grupos

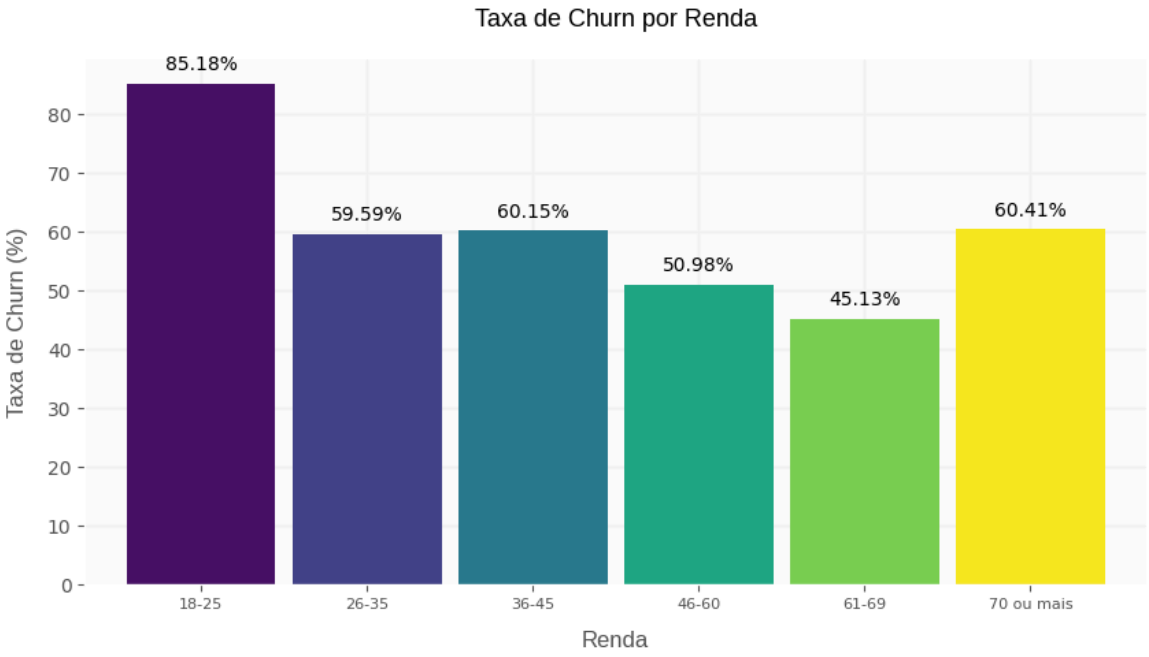
Nessa etapa, criamos visualizações que nos ajudassem a avaliar a distribuição e a relação de algumas divisões de grupos com as idades. Além disso, precisamos criar classes (por exemplo, para o atributo idade) para poder comparar a taxa de churn também em atributos numéricos. Seguem algumas visualizações:

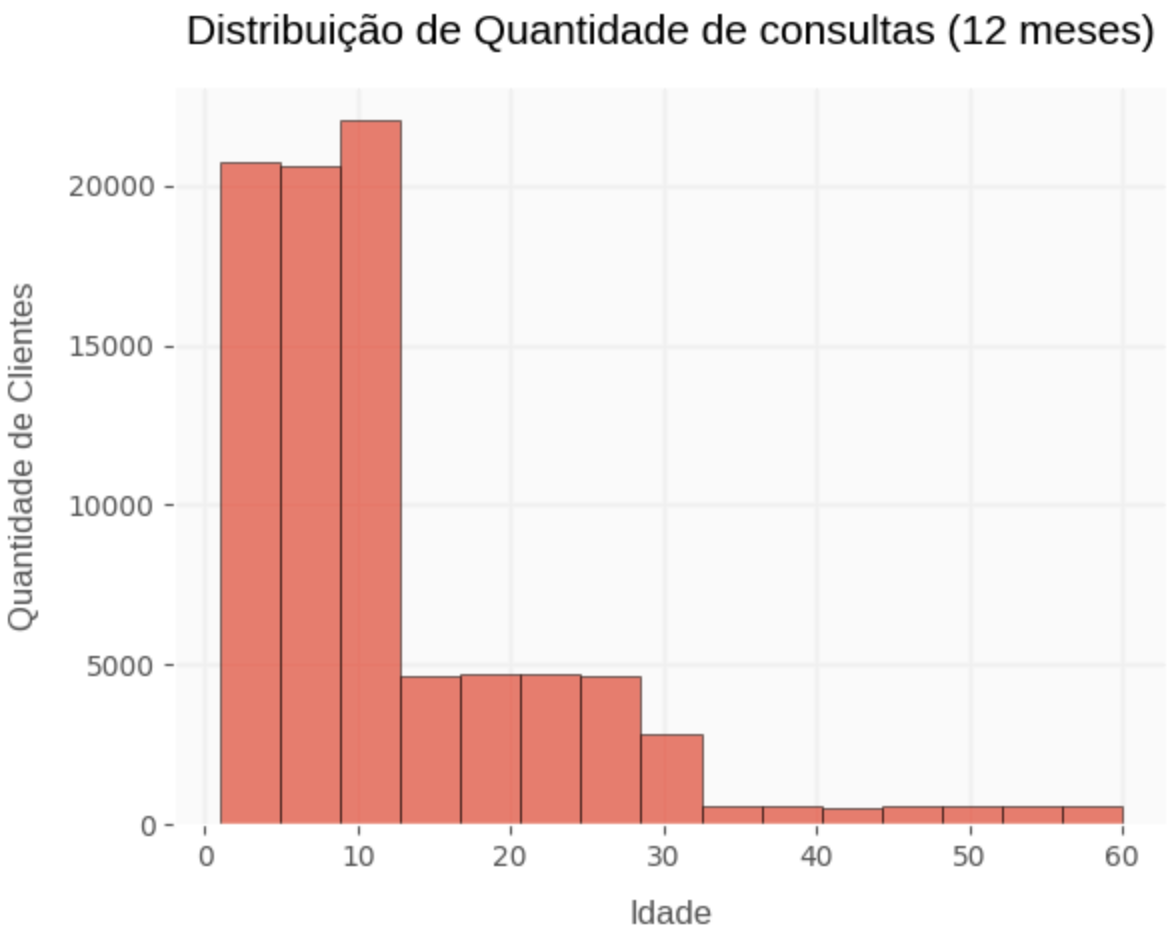


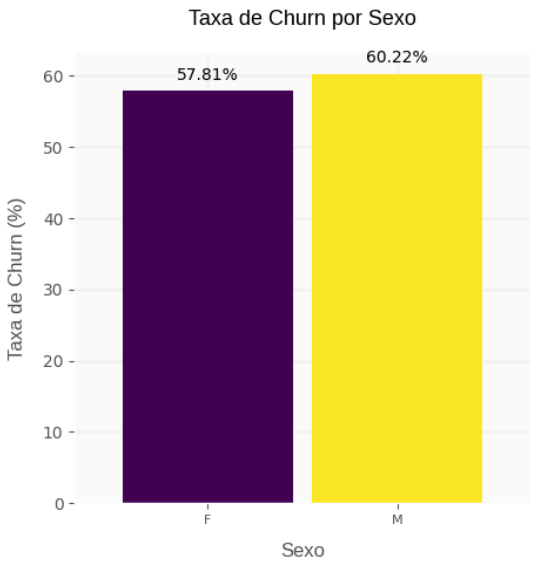


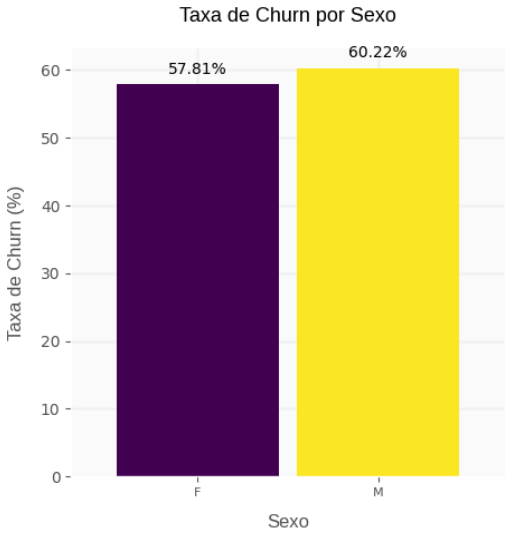

Taxa de Churn por Faixa Etária
85.18%
59.59%
60.15%
50.98%
45.13%
60.41%
Taxa de Churn (%)
0
10
20
30
40
50
60
70
80
18-25
26-35
36-45
46-60
61-69
70 ou mais
Faixa Etária

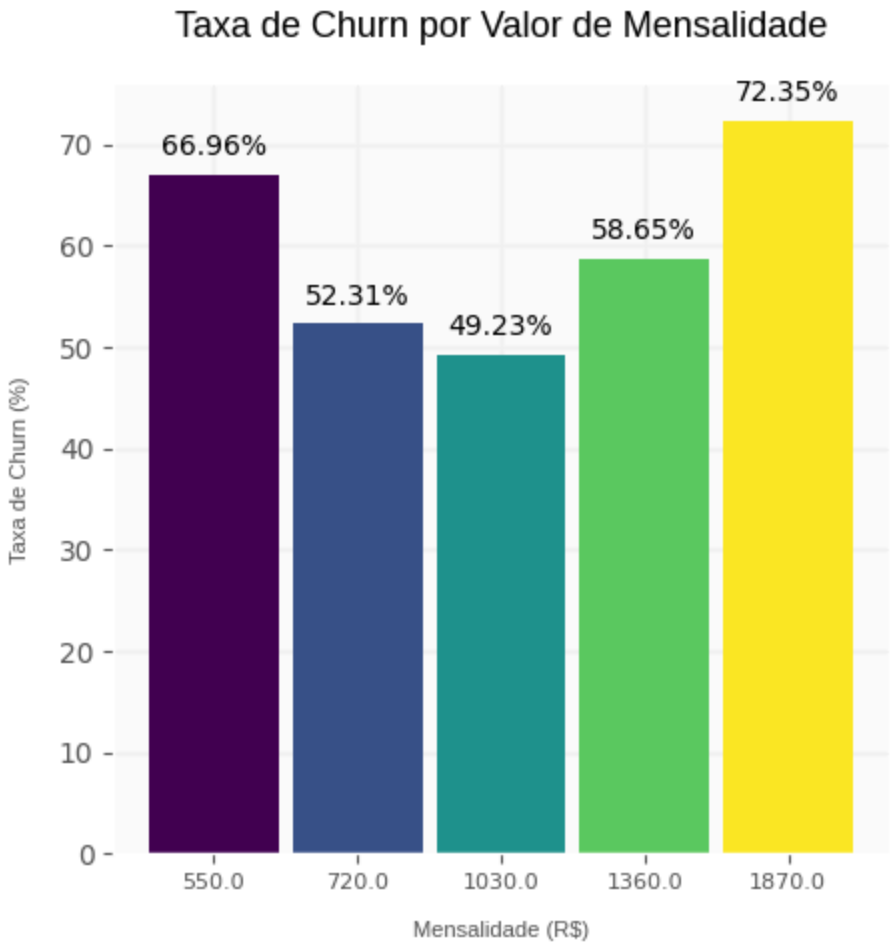

Distribuição de Clientes por Renda
64.28%
24.63%
11.09%
Quantidade de Clientes
0
10000
20000
30000
40000
50000
60000
Baixa renda
Média renda
Alta renda
Renda

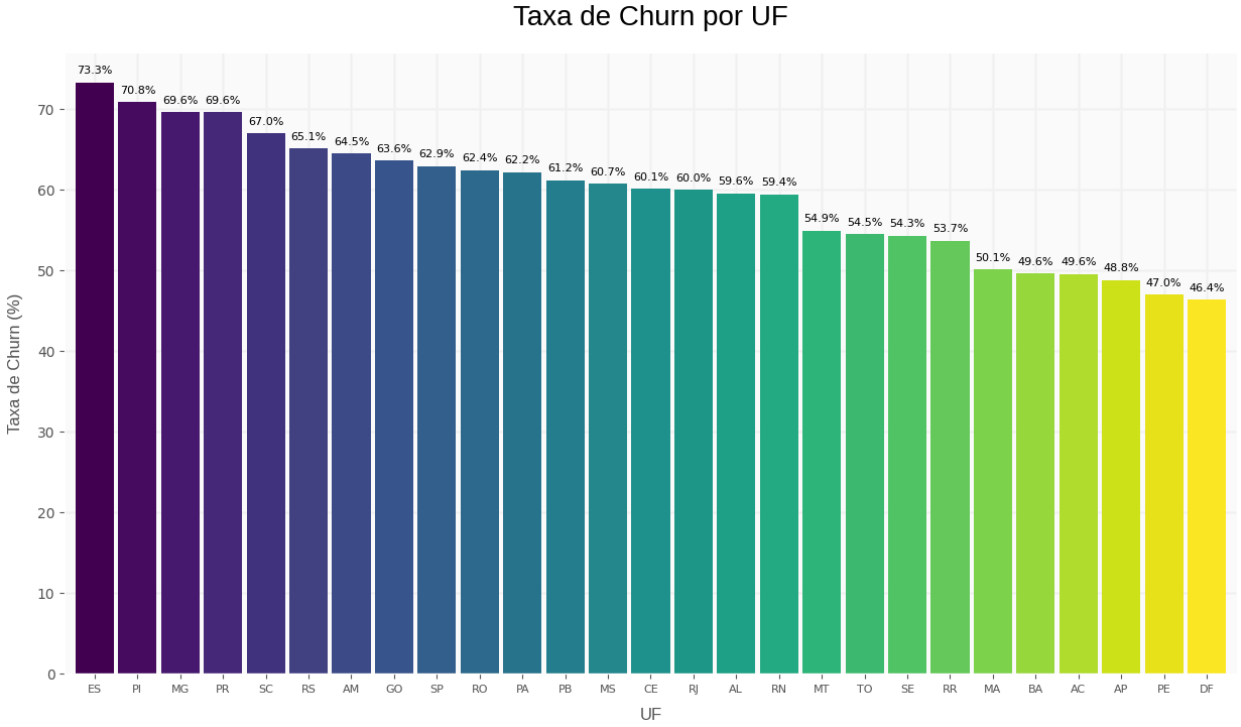

Taxa de Churn por Renda
85.18%
59.59%
60.15%
50.98%
45.13%
60.41%
Taxa de Churn (%)
0
10
20
30
40
50
60
70
80
18-25
26-35
36-45
46-60
61-69
70 ou mais
Renda


Distribuição de Quantidade de consultas (12 meses)
Quantidade de Clientes
20000
15000
10000
5000
0
0
10
20
30
40
50
60
Idade


Taxa de Churn por Sexo
57.81%
60.22%
Taxa de Churn (%)
60
50
40
30
20
10
0
F
M
Sexo

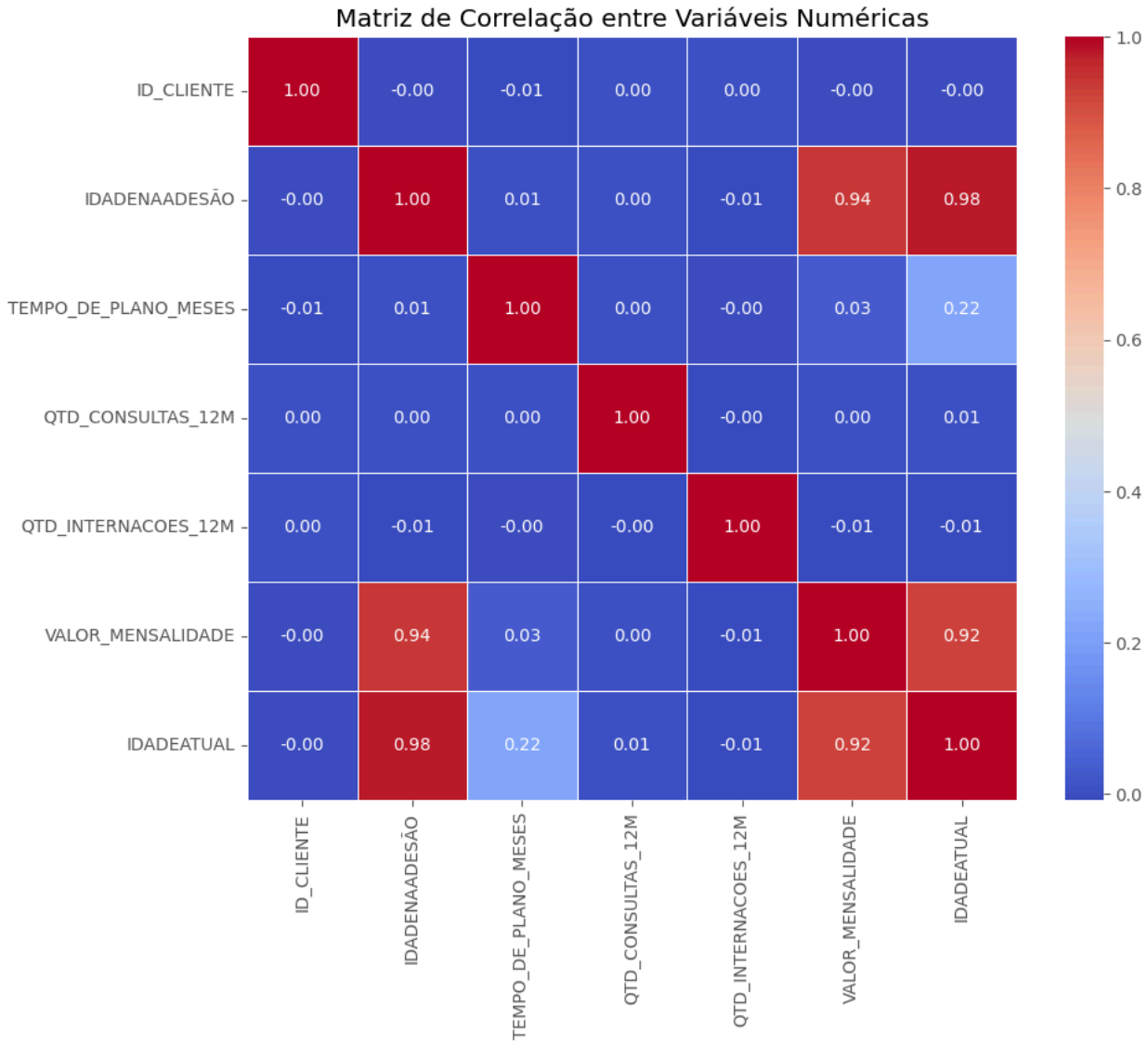

Taxa de Churn por Sexo
57.81%
60.22%
Taxa de Churn (%)
60
50
40
30
20
10
0
F
M
Sexo


Taxa de Churn por Valor de Mensalidade
Taxa de Churn (%)
Mensalidade (R$)
66.96%
52.31%
49.23%
58.65%
72.35%
550.0
720.0
1030.0
1360.0
1870.0


Taxa de Churn por UF
Taxa de Churn (%)
UF
ES 73.3%
PI 70.8%
MG 69.6%
PR 69.6%
SC 67.0%
RS 65.1%
AM 64.5%
GO 63.6%
SP 62.9%
RO 62.4%
PA 62.2%
PB 61.2%
MS 60.7%
CE 60.1%
RJ 60.0%
AL 59.6%
RN 59.4%
MT 54.9%
TO 54.5%
SE 54.3%
RR 53.7%
MA 50.1%
BA 49.6%
AC 49.6%
AP 48.8%
PE 47.0%
DF 46.4%

Análise de Relevância das variáveis

Utilizamos a matriz de correlação para entender melhor como nossos atributos se relacionam:



Através da matriz de correlação, notamos que:

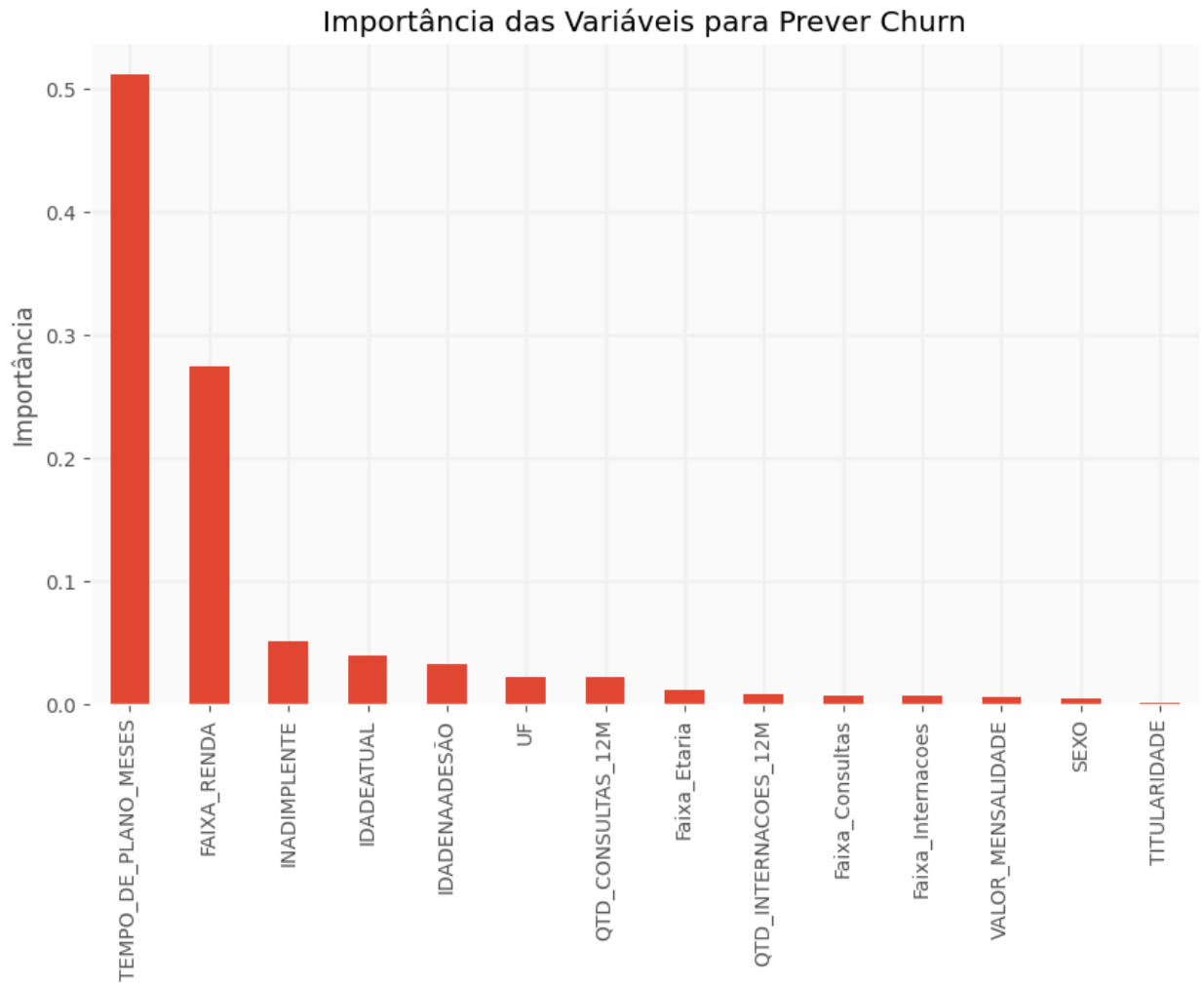
- IDADADEATUAL e IDADENAADESÃO tem correlação positiva forte, o que faz sentido pois é função da outra;
- VALOR_MENSALIDADE e IDADENAADESÃO tem correlação positiva forte, o que pode indicar que dependendo de qual idade a pessoa inicia o contrato do plano, os valores podem diferir de alguém que tem a mesma idade e tem o plano há mais tempo.

- QTD_CONSULTAS_12M e QTD_INTERNACOES_12M: tem uma correlação baixa, o que sugere que os números de consultas não tem relação direta com o número de internações

Dadas essas observações, podemos:

- Considerar apenas uma das variáveis (já tínhamos chegado à essa conclusão em seções anteriores) entre IDADADEATUAL e IDADENAADESÃO.
- Considerando a seleção de variáveis, podemos notar que, dado o noxxo contexto de taxa de churn, as variáveis VALOR_MENSALIDADE e IDADENAADESÃO tem relevância teórica, então possivelmente vamos utilizá-l**as.**

Calculamos também a Importância das variáveis para entender quais delas contrinuiriam mais para o modelo preditivo de Churn



Através do Gráfico de Importâncias das variáveis, notamos que:

- TEMPO_DE_PLANO_MESES é a variável mais importante para prever o churn, com uma importância de aproximadamente 0.5.
- FAIXA_RENDA é a segunda variável mais importante, com uma importância de cerca de 0.3.
- As variáveis como SEXO e UF têm uma importância muito baixa, sugerindo que esses fatores não contribuem significativamente para o modelo, ou a relação deles com o churn não é tão forte.
- O modelo está identificando que quanto mais tempo o cliente permanece no plano, maior a probabilidade de ele ter churn, talvez por uma relação de insatisfação com o tempo de uso.

A **faixa de renda** também parece ser um fator importante, possivelmente porque diferentes grupos de renda podem ter comportamentos distintos em relação ao churn.

Modelo

Consideramos três possibilidades de algoritmos: Random Forest, Naive Bayes e SVM.

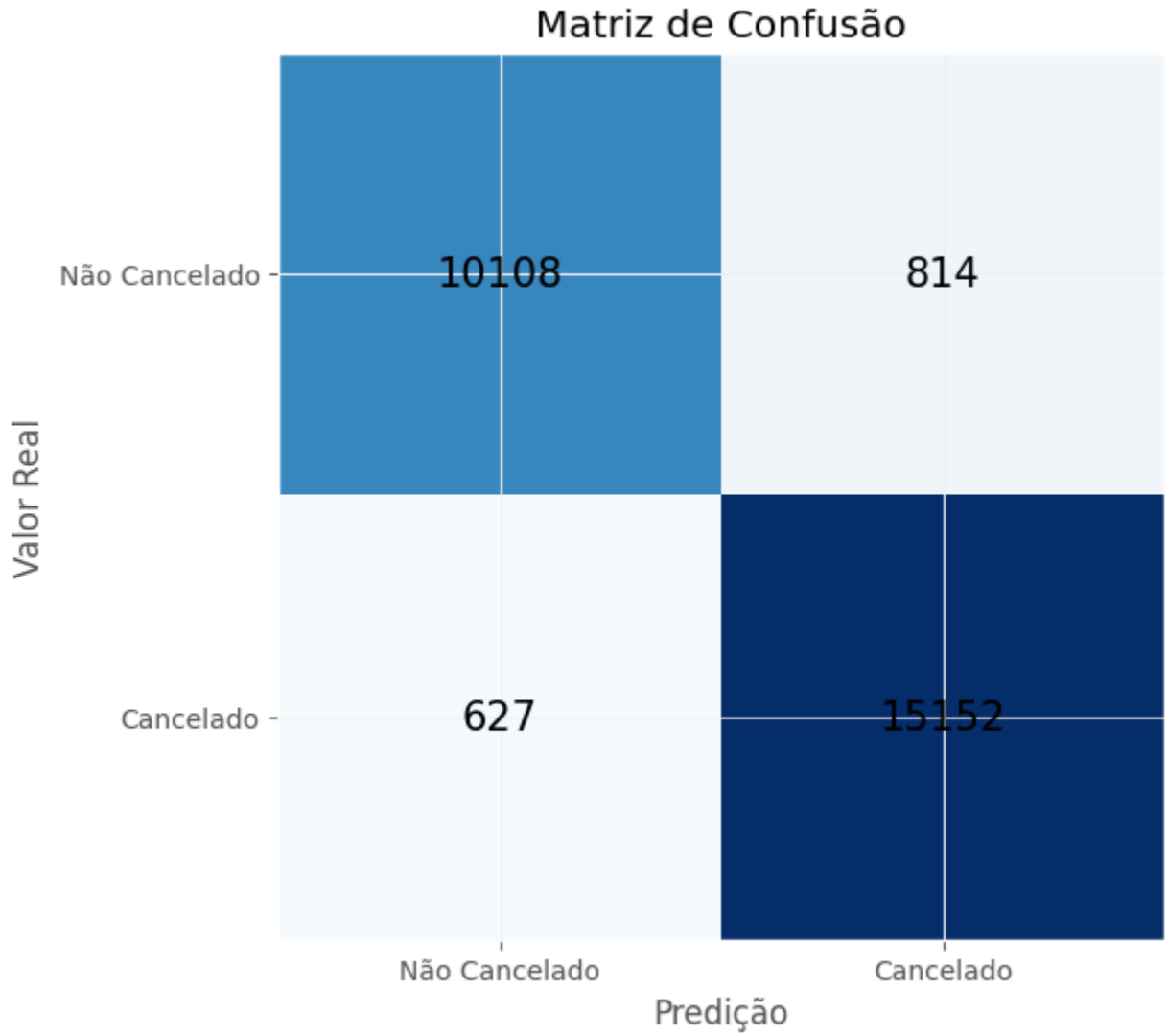
Modelo 1 - Random Forest

Consideramos o modelo Random Forest com os seguintes parâmetros e considerando as variáveis explicativas TEMPO_DE_PLANO_MESES e FAIXA_RENDA:

```
rf = RandomForestClassifier(n_estimators=100, random_state=42)
```

Relatório de Avaliação:

	precision	recall	f1-score	support
0	0.94	0.93	0.93	10922
1	0.95	0.96	0.95	15779
accuracy			0.95	26701
macro avg	0.95	0.94	0.94	26701
weighted avg	0.95	0.95	0.95	26701



Nesse caso, como queremos capturar tanto clientes que vão cancelar (o melhor possível) como minimizar os clientes classificados como "CANCELADOS", porém com uma predição errada. Logo focaremos no F1-score, precisão e recall como métricas de avaliação.

Estudamos também um modelo normalizando os atributos, mas obtivemos valores similares.

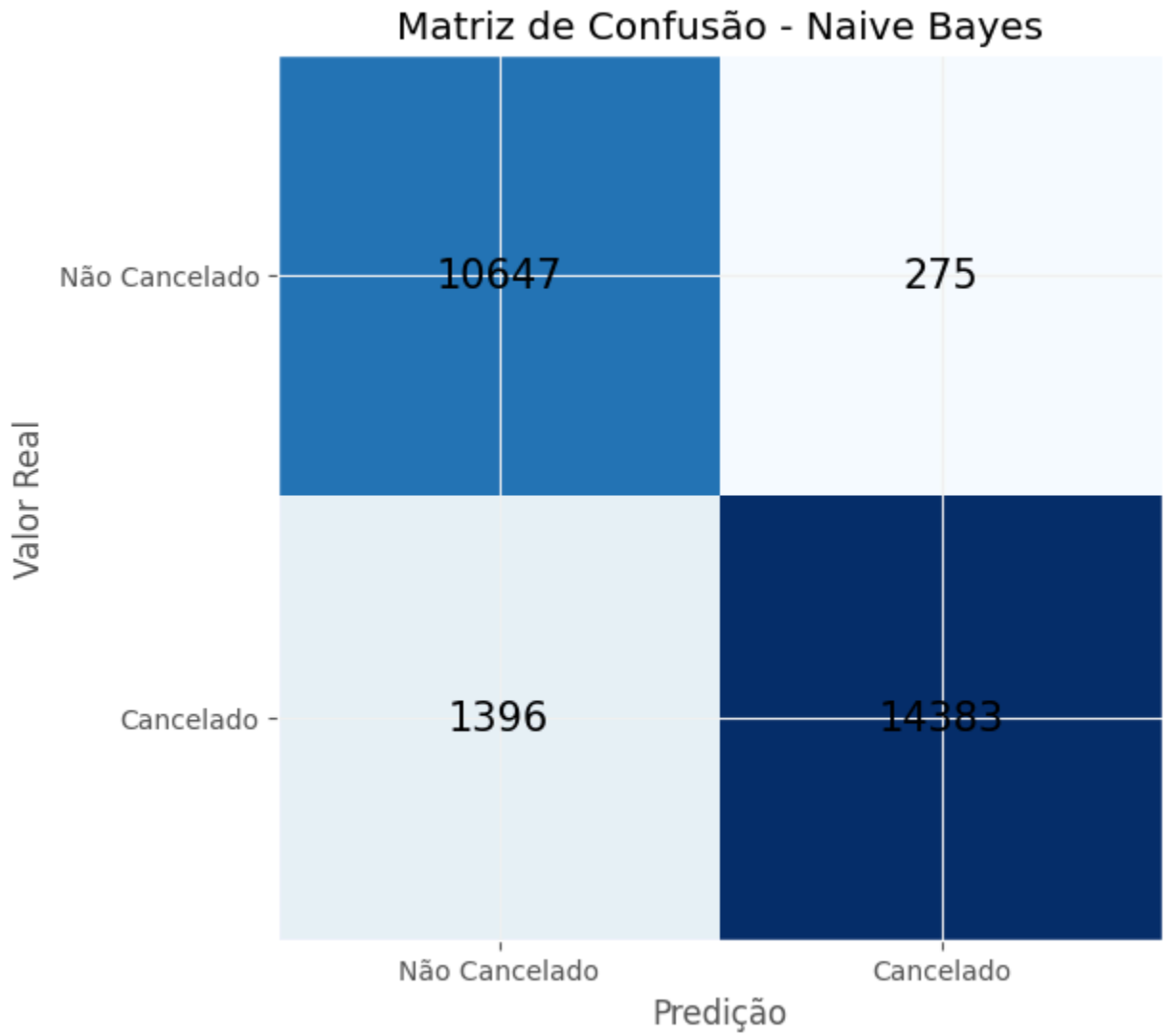
Modelo 2 - Naive Bayes

Testamos também o algoritmo Naive Bayes com as mesmas variáveis explicativas:

```
nb = GaussianNB()
```

Relatório de Avaliação:

	precision	recall	f1-score	support
0	0.88	0.97	0.93	10922
1	0.98	0.91	0.95	15779
accuracy			0.94	26701
macro avg	0.93	0.94	0.94	26701
weighted avg	0.94	0.94	0.94	26701

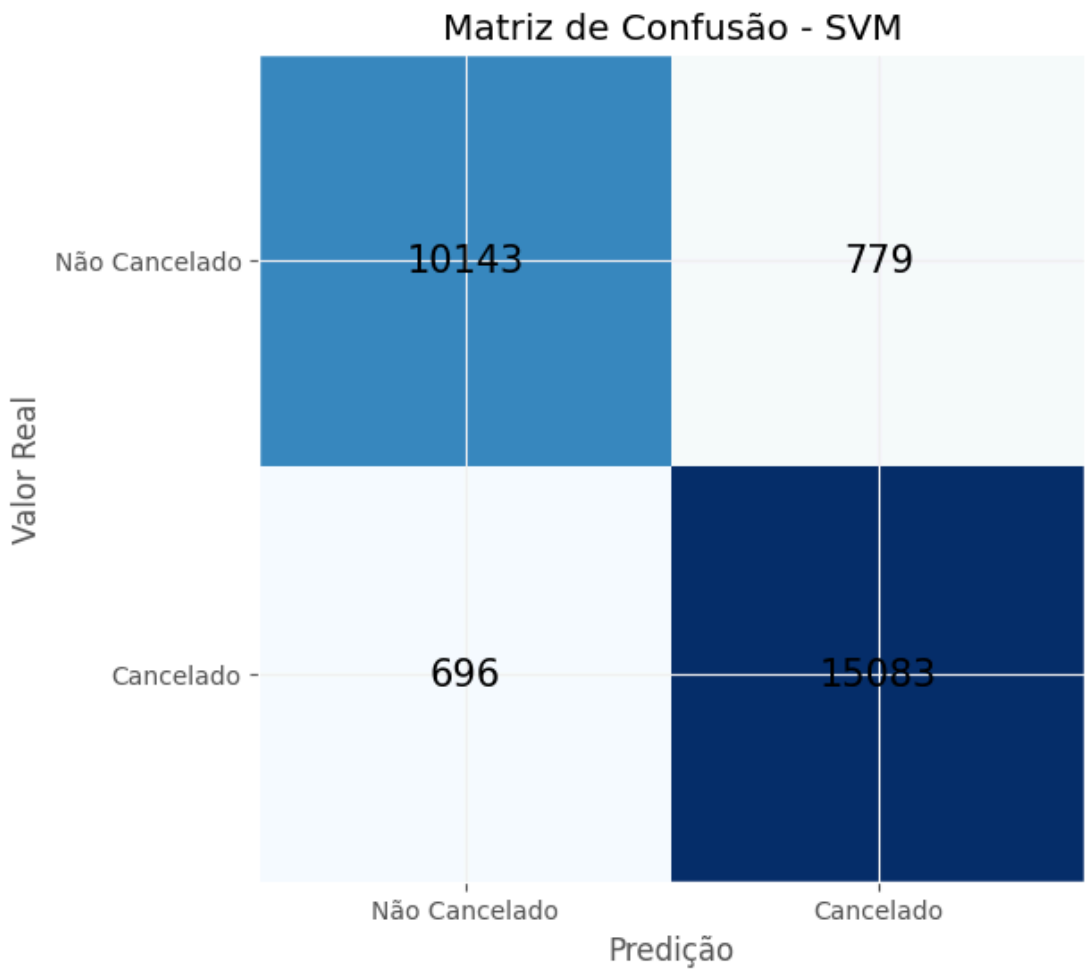


Modelo 3 - SVM

SMV é um algoritmo que funciona bem com grandes volumes de dados, apesar de ser um pouco mais lento. Além disso, é necessário normalizar os dados.

Relatório de Avaliação:

	precision	recall	f1-score	support
0	0.94	0.93	0.93	10922
1	0.95	0.96	0.95	15779
accuracy			0.94	26701
macro avg	0.94	0.94	0.94	26701
weighted avg	0.94	0.94	0.94	26701



Resultados:

F1-Score: Essa métrica é a média harmônica entre precisão e recall, oferecendo um equilíbrio entre as duas. É especialmente útil quando temos uma distribuição desigual de classes (como no caso de churn, onde as classes "Cancelado" e "Não Cancelado" podem ser desbalanceadas).

Precisão: Indica a proporção de previsões positivas corretas em relação ao total de previsões positivas feitas pelo modelo.

Recall: Mede a proporção de reais positivos que foram corretamente identificados pelo modelo.

Modelo 1:

F1-Score: 0.9546 | Precisão: 0.9490 | Recall: 0.9603

F1-Score no Treinamento: 0.9554| F1-Score no Teste: 0.9546

Modelo 2:

F1-Score: 0.9460 | Precisão: 0.9807 | Recall: 0.9136

F1-Score no Treinamento: 0.9554| F1-Score no Teste: 0.9451

Modelo 2:

F1-Score: 0.9534 | Precisão: 0.9509 | Recall: 0.9559

F1-Score no Treinamento: .9554| F1-Score no Teste: 0.9524

Conclusões:

Random Forest:

O Random Forest tem o F1-Score mais alto de 0.9546, o que indica um bom equilíbrio entre precisão e recall. O recall de 0.9603 é bastante alto, significando que o modelo consegue identificar a maior parte dos clientes que cancelam. A precisão também é boa (0.9490), mas não é a melhor.

Naive Bayes:

O Naive Bayes tem a precisão mais alta (0.9807), o que significa que ele faz menos previsões falsas positivas. No entanto, o recall (0.9136) é o mais baixo entre os modelos, indicando que ele está perdendo uma quantidade considerável de clientes que realmente cancelaram (falsos negativos). Isso pode ser um problema, já que perder clientes cancelados é mais crítico do que errar previsões de clientes que não cancelam.

SVM:

O SVM tem o F1-Score de 0.9534, que é muito próximo ao do Random Forest. Ele também apresenta um bom recall de 0.9559, e uma precisão de 0.9509, que é bastante equilibrada.

Portanto, após a análise dos três modelos acima, vamos usar o que utiliza Random Forest com quatro variáveis explicativas (adicionamos Idade atual e VALOR_MENSALIDADE nas duas iniciais)

O modelo escolhido (abaixo) apresenta o melhor equilíbrio entre precisão e recall, com um F1-Score mais alto do que os outros modelos. Além disso, tem uma taxa de recall muito alta, o que é importante no caso de previsão de churn, já que você quer minimizar a quantidade de clientes que cancelam e não são identificados pelo modelo.

Embora o modelo que utiliza Naive Bayes tenha uma precisão maior, a baixa taxa de recall é preocupante, pois ele perde clientes que realmente cancelam (FN).

Portanto, o modelo rf_churn é a melhor opção capaz de estimar quais beneficiários têm maior risco de evasão nos próximos 12 meses (ou seja, os planos cancelados).

Além disso, apesar de termos usado 4 variáveis explicativas, vimos que duas delas tem um peso grande na explicação da variável "CANCELADO": TEMPO_DE_PLANO_MESES', 'FAIXA_RENDA'.

Vimos anteriormente que quanto maior o tempo de plano, maior a tendência ao churn, ou seja, faz sentido tentar entender por que as pessoas costumam ficar insatisfeitas a medida que o tempo passa. Seria porque outros planos são mais atrativos no sentido de atualização de profissionais/hospitais? Talvez faça sentido implementar ações para tentar recuperar essa parcela possivelmente insatisfeita da população. Uma outra possível ideia é rodas pesquisas de satisfação com brindes em troca da participação.

Outro fator importante na taxa de Churn é a Faixa de Renda, sendo que a renda mais baixa tem uma porcentagem maior de Churn quando analisamos os dados do banco disponível. É possível que os reajustes de mensalidade de tempos em tempos estejam fazendo com que o pagamento mensal se torne difícil (esses ajustes acompanham os ajustes de salário mínimo?). Será que faria sentido a criação de mais uma categoria de plano para pessoas de rendas mais baixas? Quais seriam as opções que tornariam isso viável?

Observações: Não tive muito tempo para testar outros modelos e parâmetros, mas seria interessante fazê-lo num contexto de mais tempo e mais entendimento do cenário.