Previsão da satisfação dos clientes do banco Santander

Alberto Rodrigues Ferreira
Willian Miranda A. da Silva

Instituto de Matemática e Estatística
Universidade de São Paulo
Aprendizagem Estatística em Altas Dimensões

Descrição do problema

1. Identificação de clientes insatisfeitos do banco Santander no início de seu relacionamento.
2. Predição da variável de satisfação do cliente, 1 para clientes insatisfeitos e 0 para clientes satisfeitos.

Conjunto de dados

1. Existem 76020 observações e 371 variáveis anônimas.
2. Classe altamente desbalanceada, 73012 observações de clientes satisfeitos(96%) e 3008 observações clientes insatisfeitos(4%).

Objetivos

1. Maximizar a métrica F1 Score, consequentemente maximizar recall e precision da classe de clientes insatisfeitos.
2. Obter um bom modelo preditivo com as variáveis mais importantes.

Desafios

1. Lidar com uma grande quantidade de observações e variáveis.
2. Tentar resolver o problema de desbalanceamento.

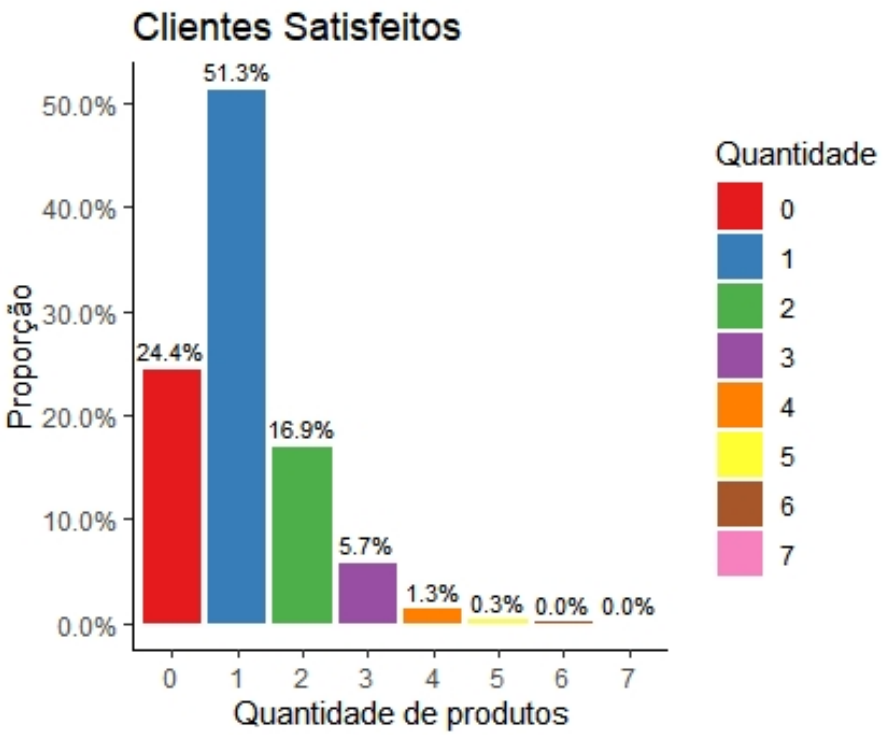
Pré-processamento inicial

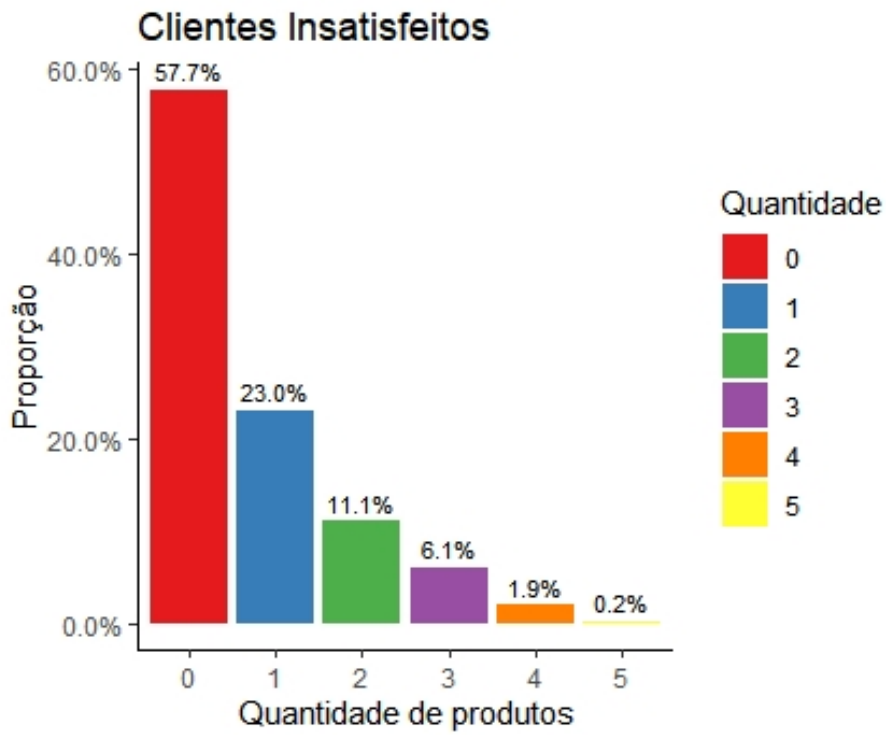
1. Exclusão de variáveis com desvio padrão zero.
2. Exclusão de variáveis exatamente iguais.
3. Análise exploratória de variáveis possivelmente relevantes para a predição.

Análise das principais variáveis de predição

1. Analisamos algumas variáveis que potencialmente possuam um bom poder preditivo.
2. Criamos novas variáveis para aumentar a separabilidade entre as classes.

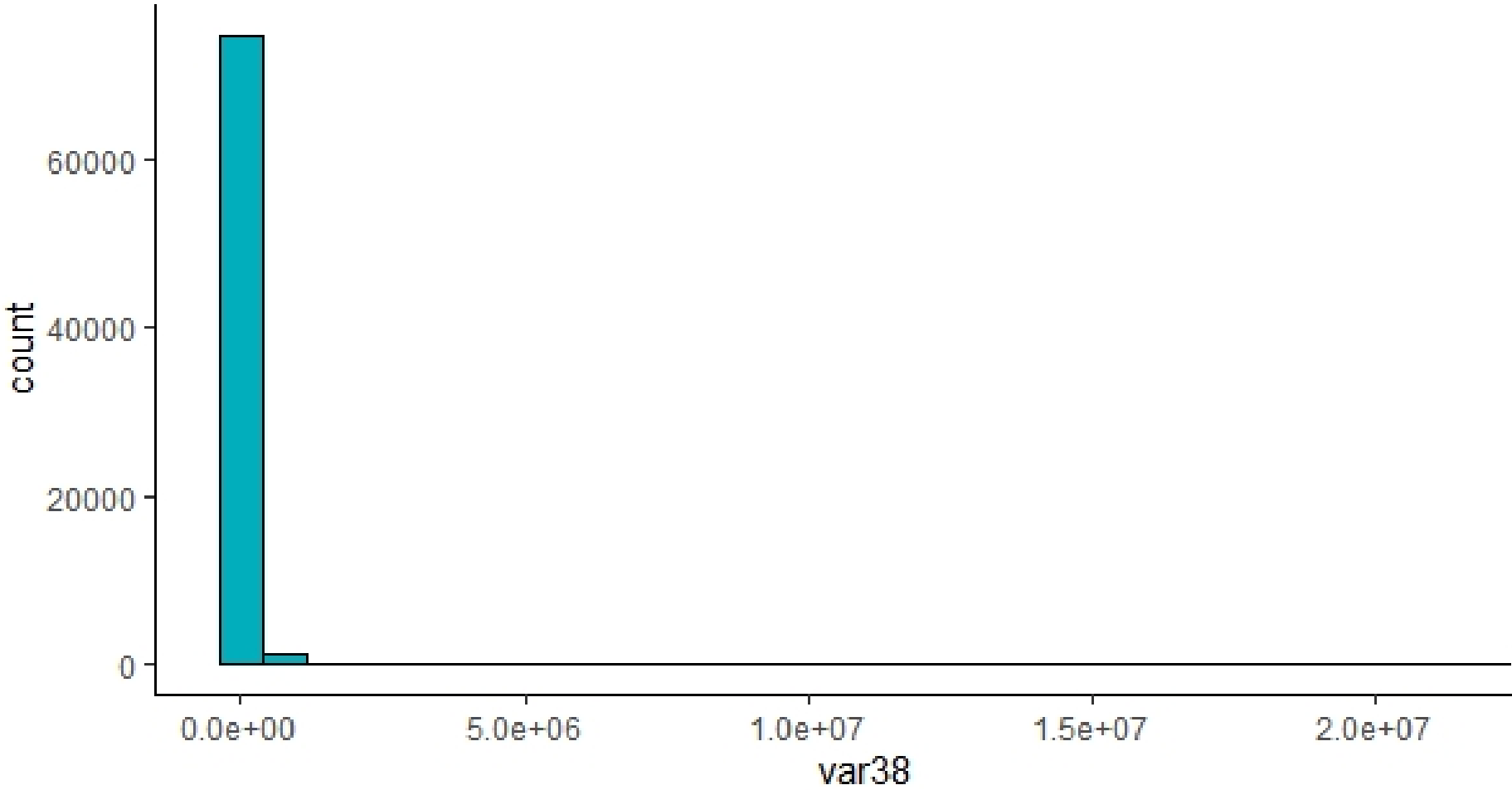
Número de produtos



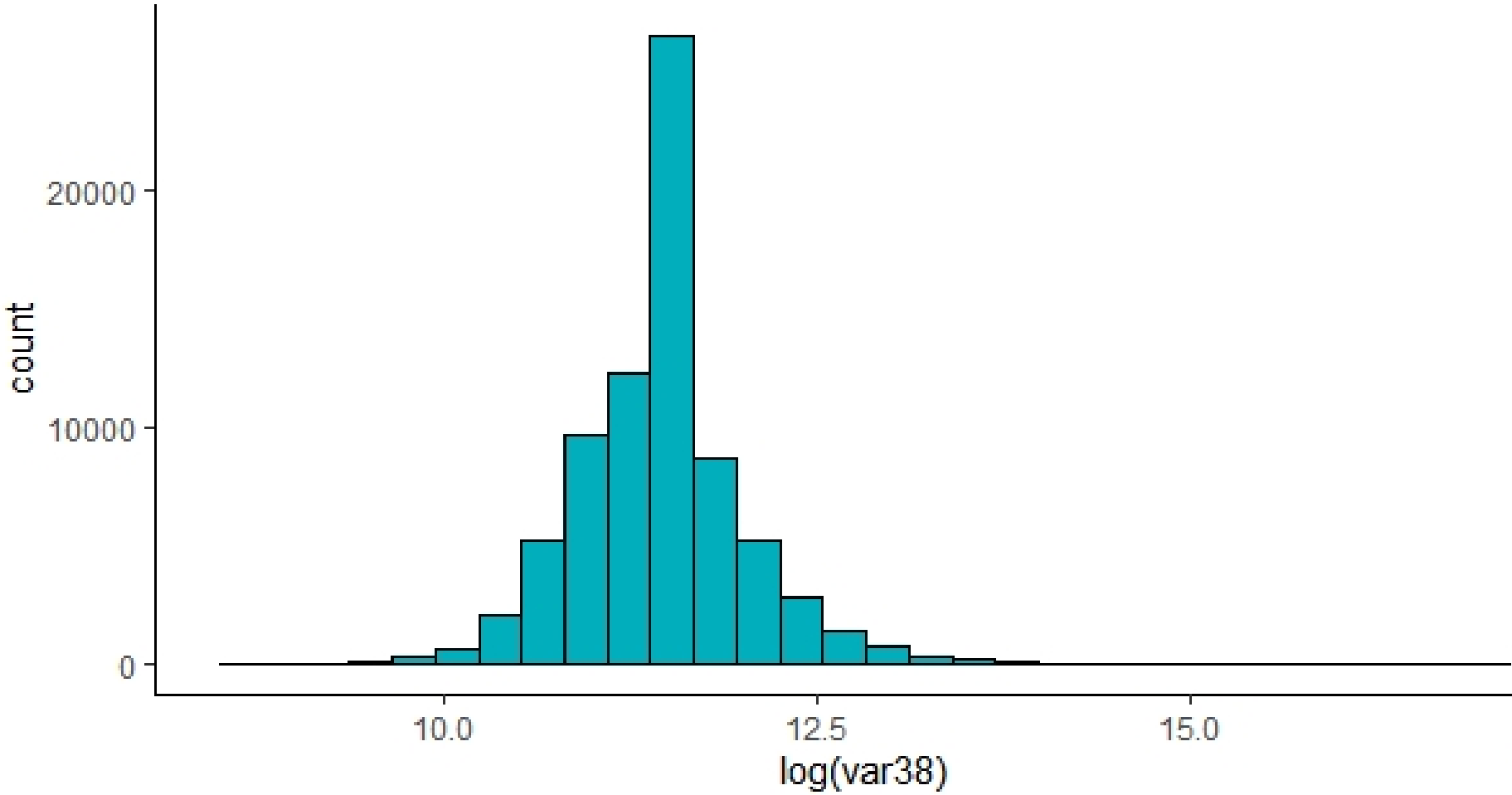


$$\text{quant\_produtos0}=\begin{cases}1 & \text{, se } x=0\\ 0 & \text{, se } x\neq 0\end{cases} \qquad \text{quant\_produtos1}=\begin{cases}1 & \text{, se } x=1\\ 0 & \text{, se } x\neq 1\end{cases}$$

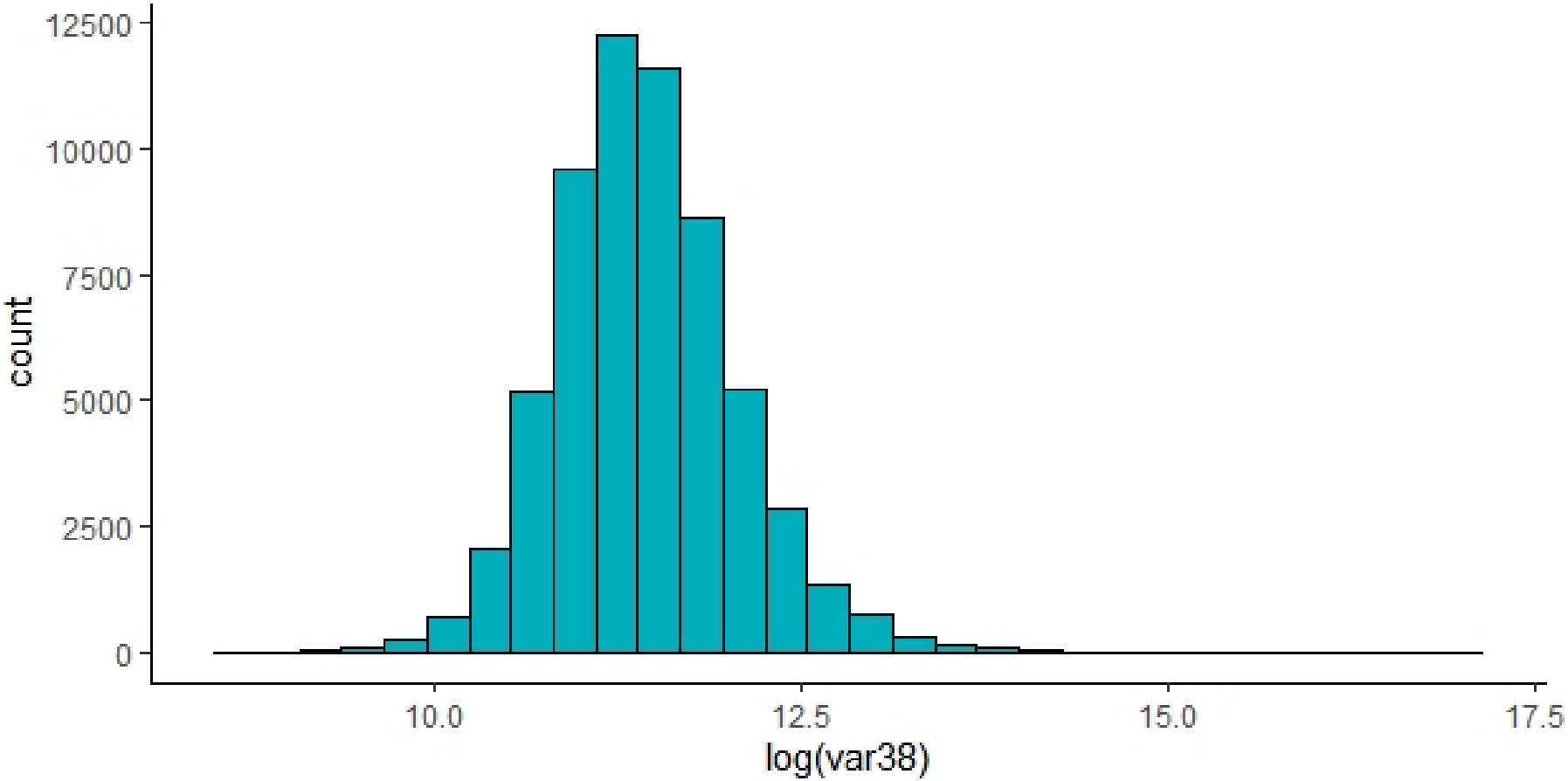
Var38


60000
40000
20000
0
count
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
var38

Transformação logarítmica-Var38

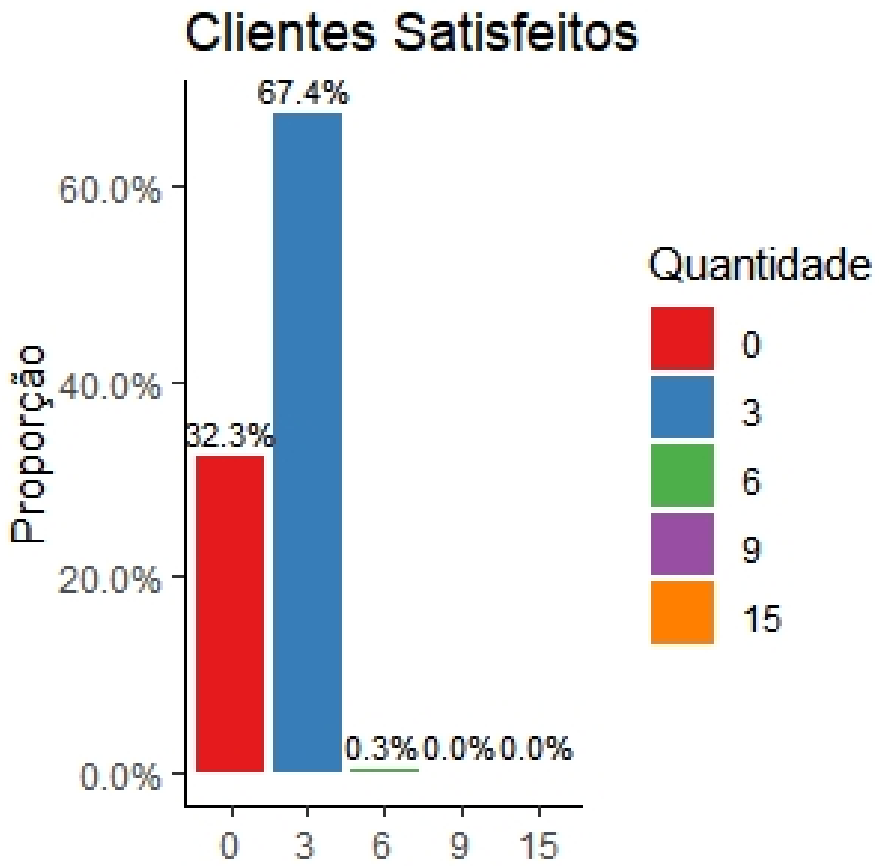


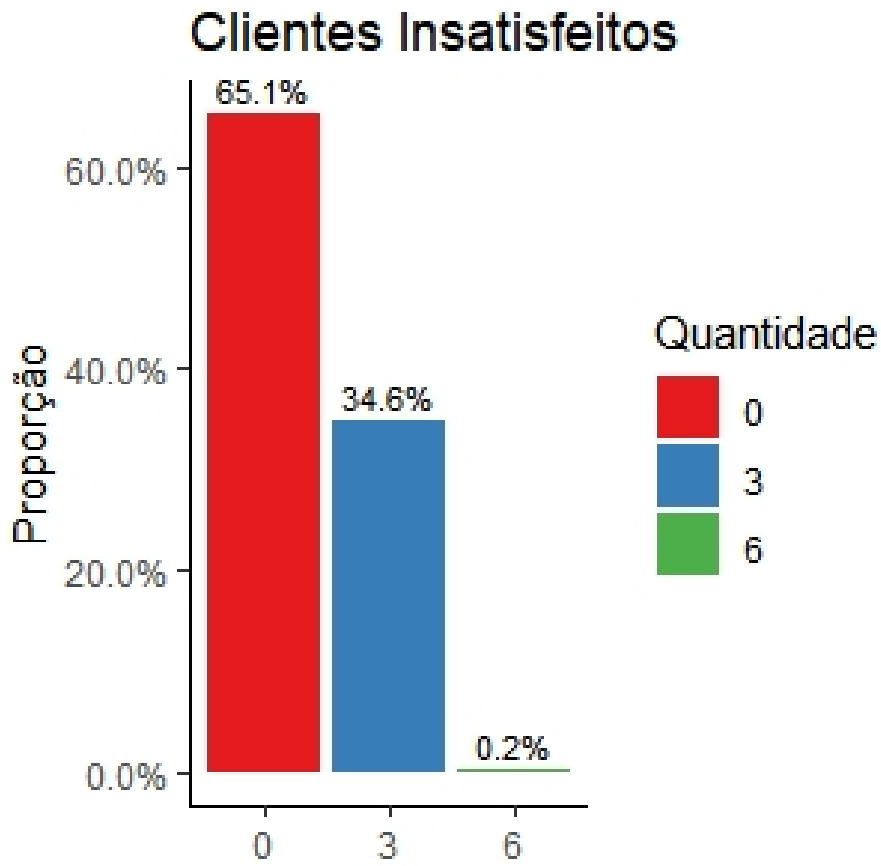
Transformação logarítmica-Var38



$$\text{var38\_normal}=\begin{cases} log(var38) & \text{, se } x \neq moda \\ 0 & \text{, se } x = moda \end{cases} \qquad \text{var38\_normal\_dummy}=\begin{cases} 1 & \text{, se } x \neq moda \\ 0 & \text{, se } x = moda \end{cases}$$

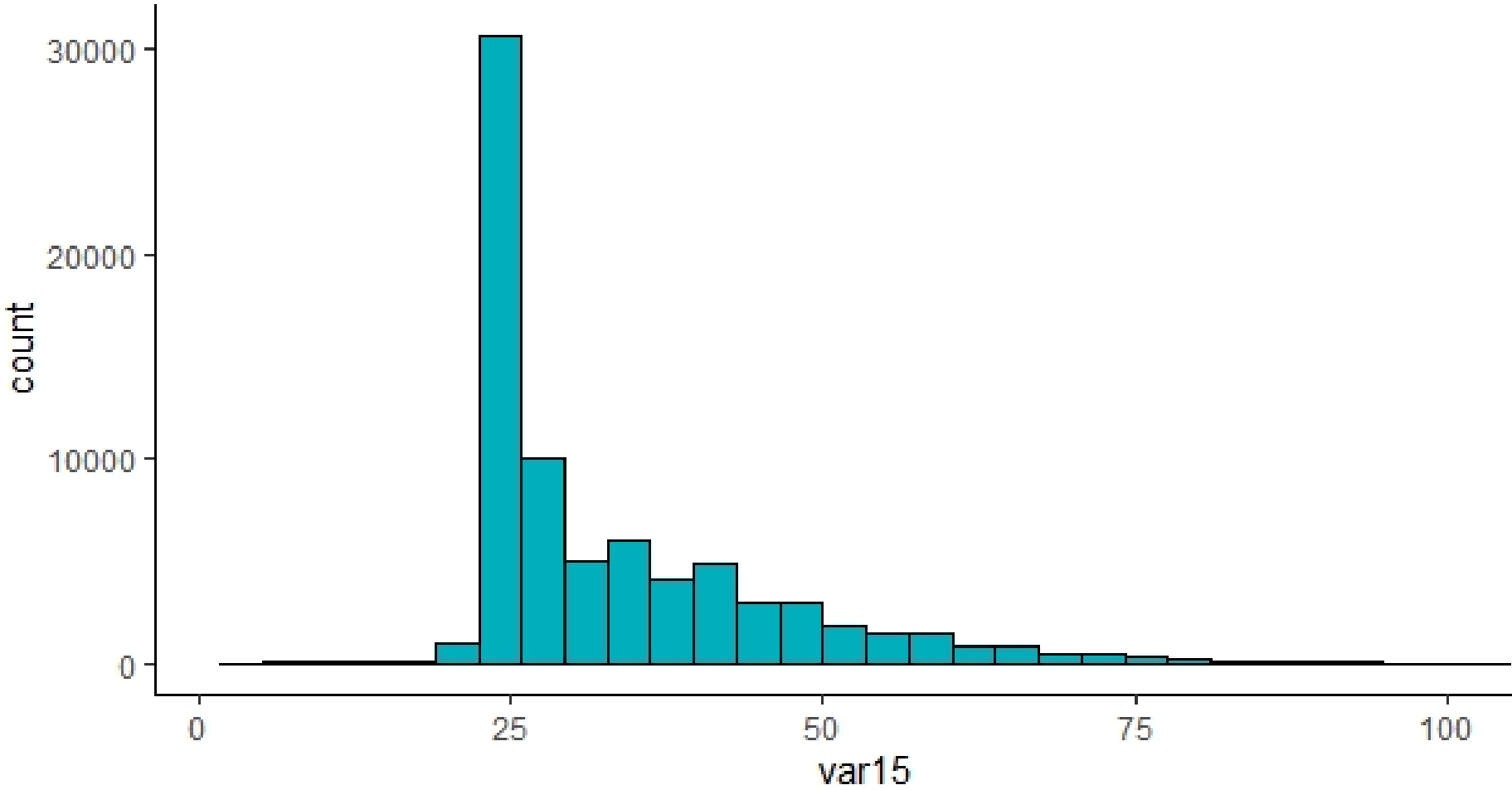
num_var5



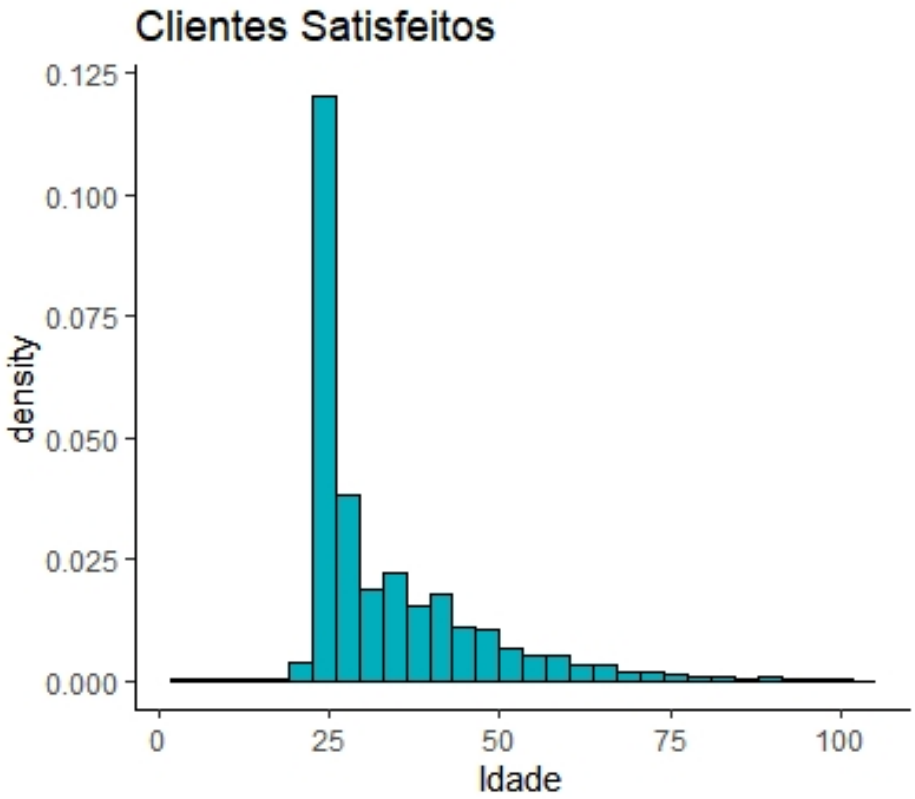


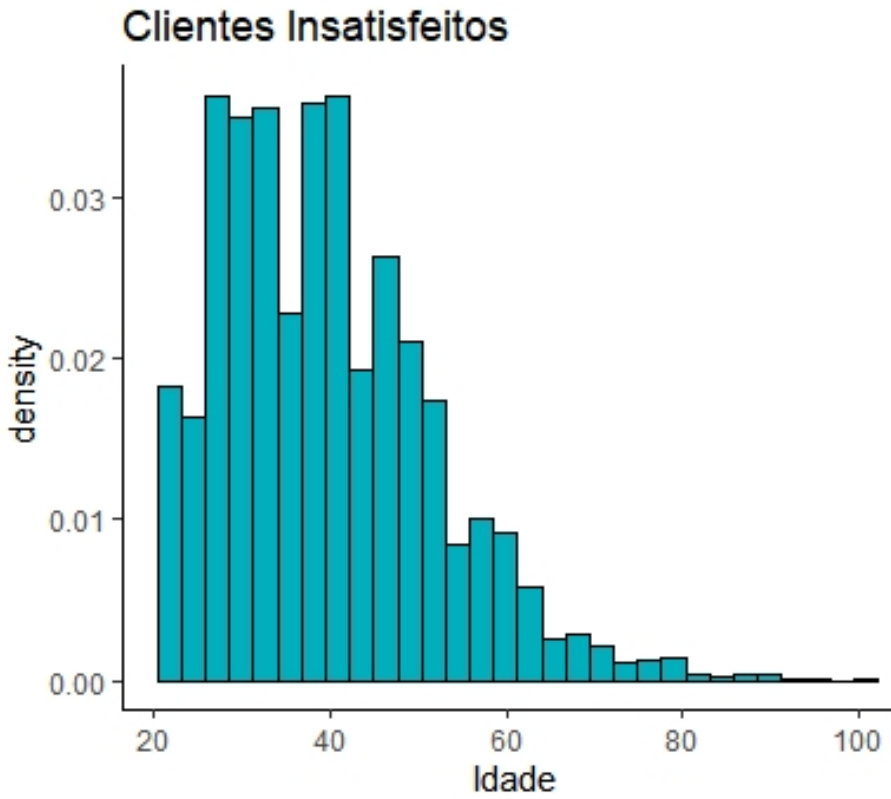
$$\text{num\_var5\_6}=\begin{cases}1 & \text{, se } x=6\\ 0 & \text{, se } x\neq 6\end{cases} \qquad \text{num\_var5\_0}=\begin{cases}1 & \text{, se } x=0\\ 0 & \text{, se } x\neq 0\end{cases}$$

Idade



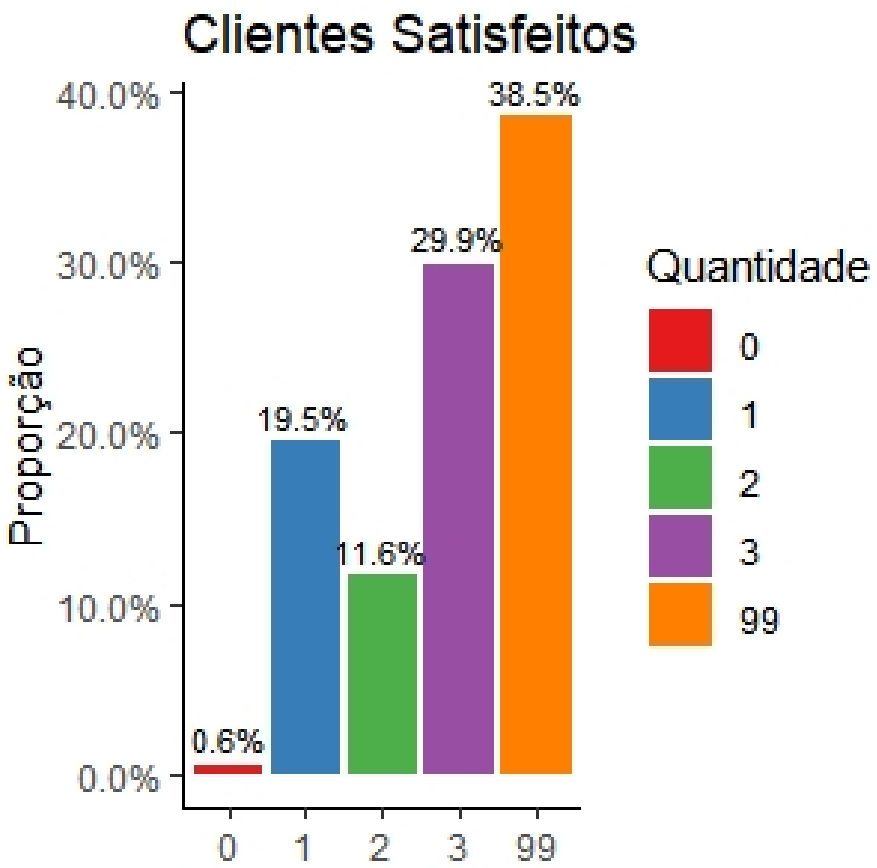
Idade por classe

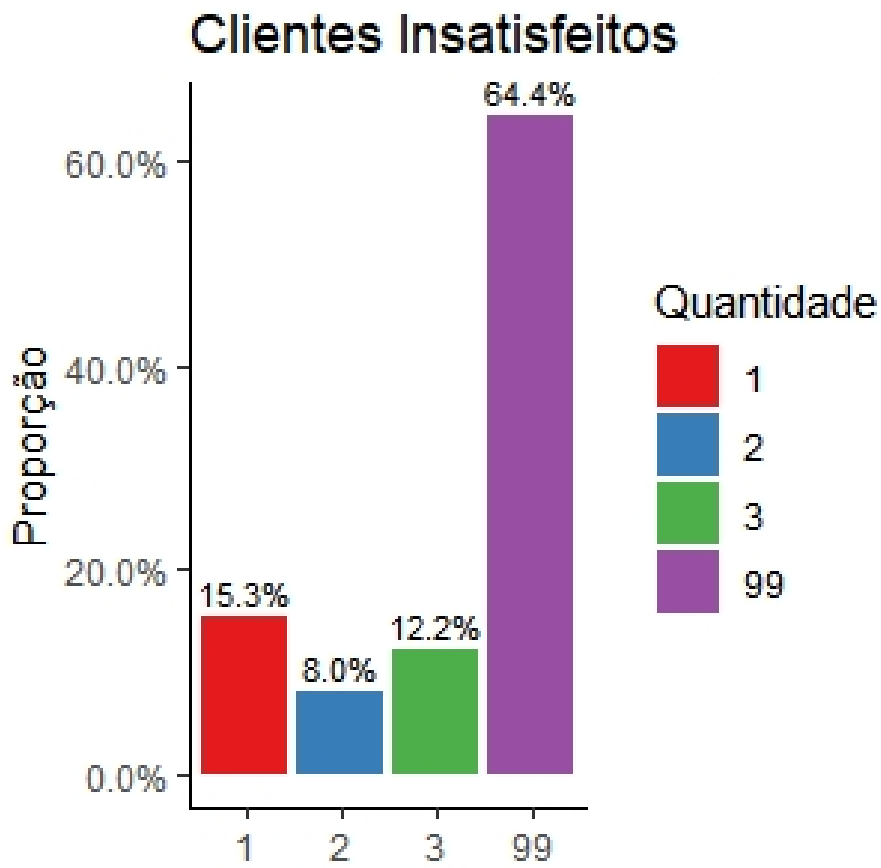




$$\text{idade\_menor} = \begin{cases} 1 & \text{, se } x \leq 21 \\ 0 & \text{, se } x > 21 \end{cases}$$

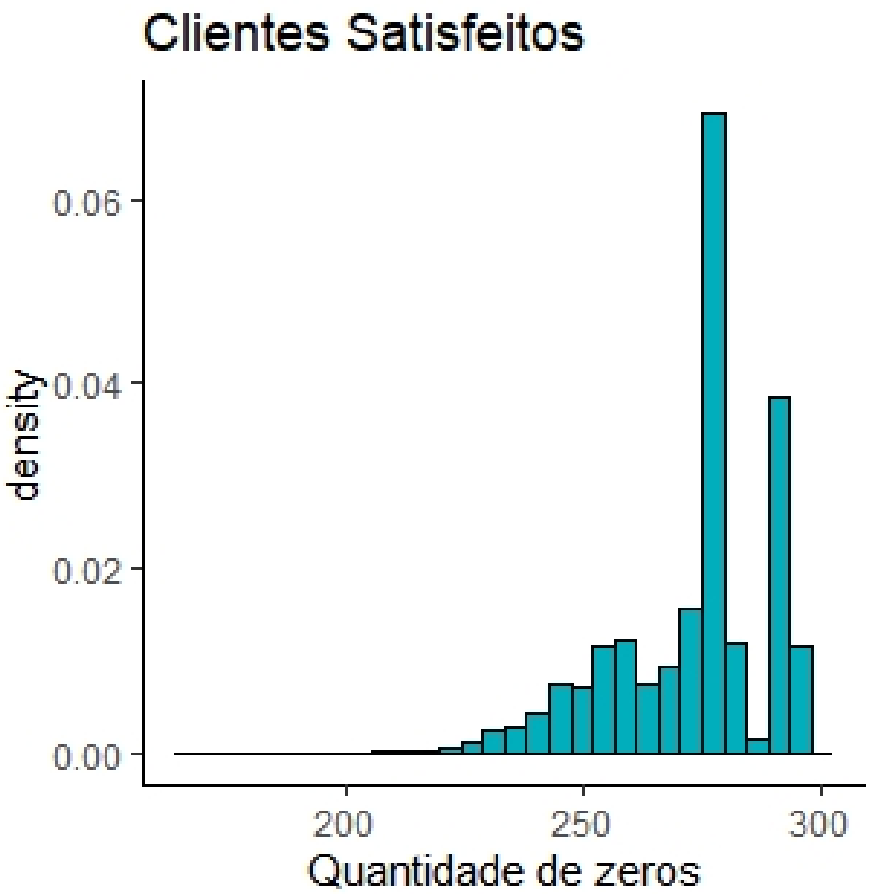
var36

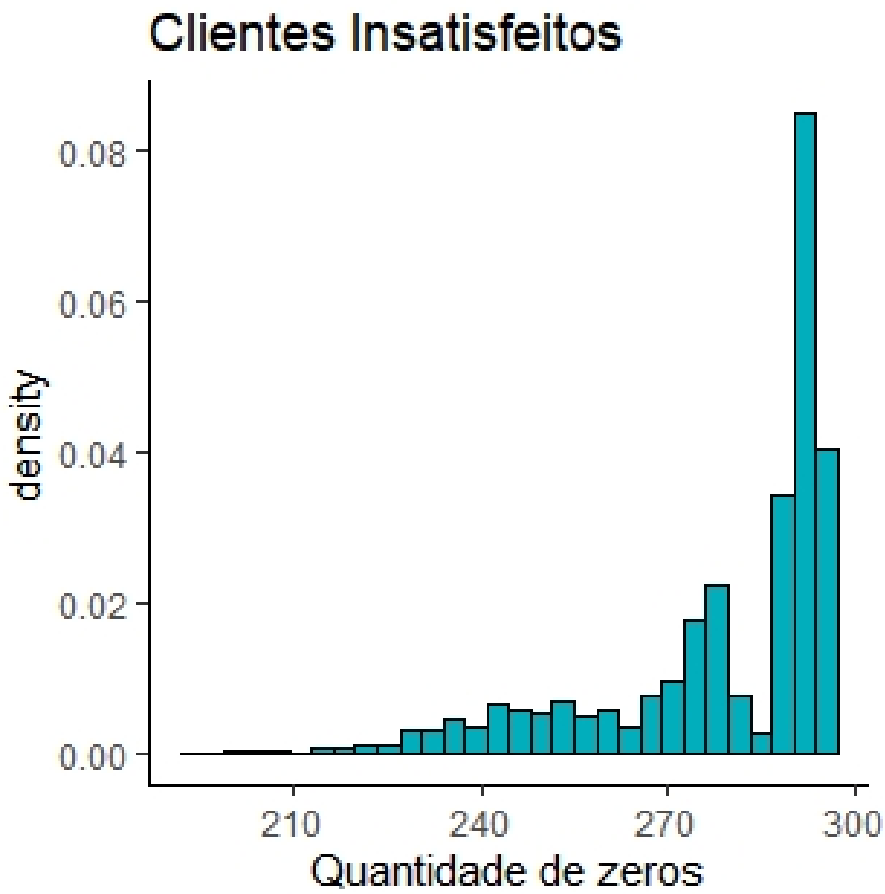




$$\text{var36\_99}=\begin{cases}1 & \text{, se } x = 99\\ 0 & \text{, se } x \neq 99\end{cases} \qquad \text{var36\_0}=\begin{cases}1 & \text{, se } x = 0\\ 0 & \text{, se } x \neq 0\end{cases}$$

Quantidade de zeros





Problema de desbalanceamento

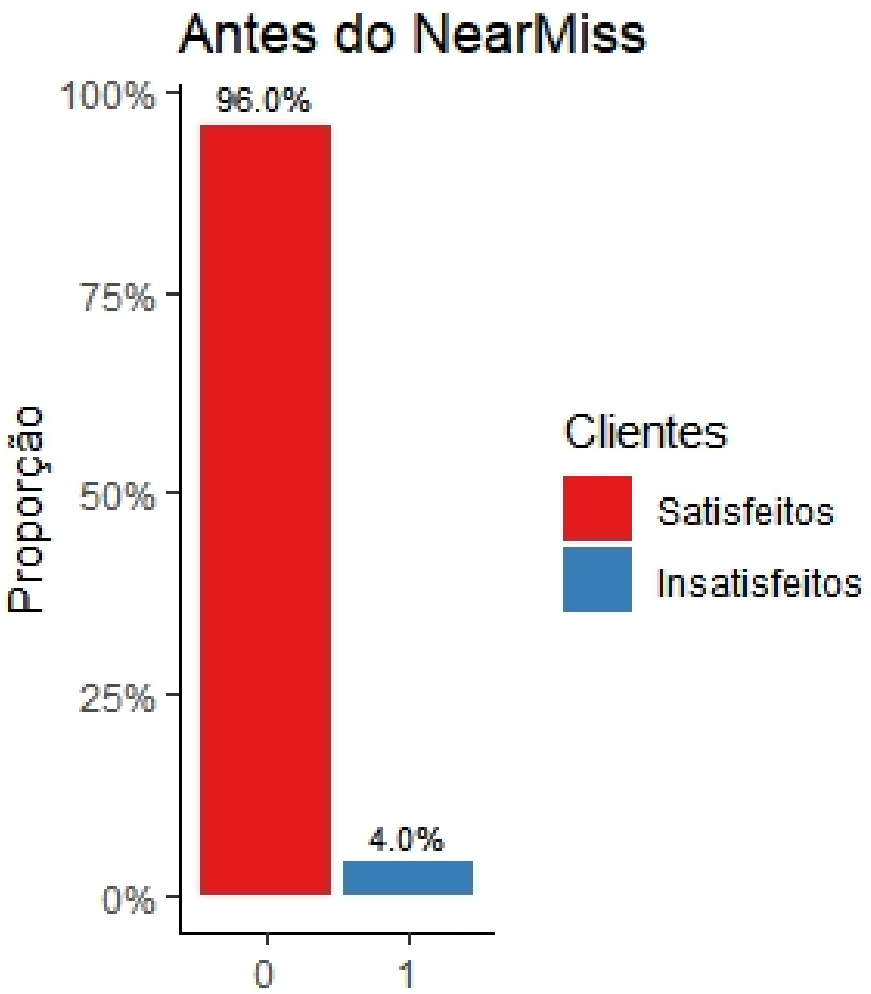
- Tentamos resolver este problema realizando o balanceamento manualmente através do algoritmo NearMiss.

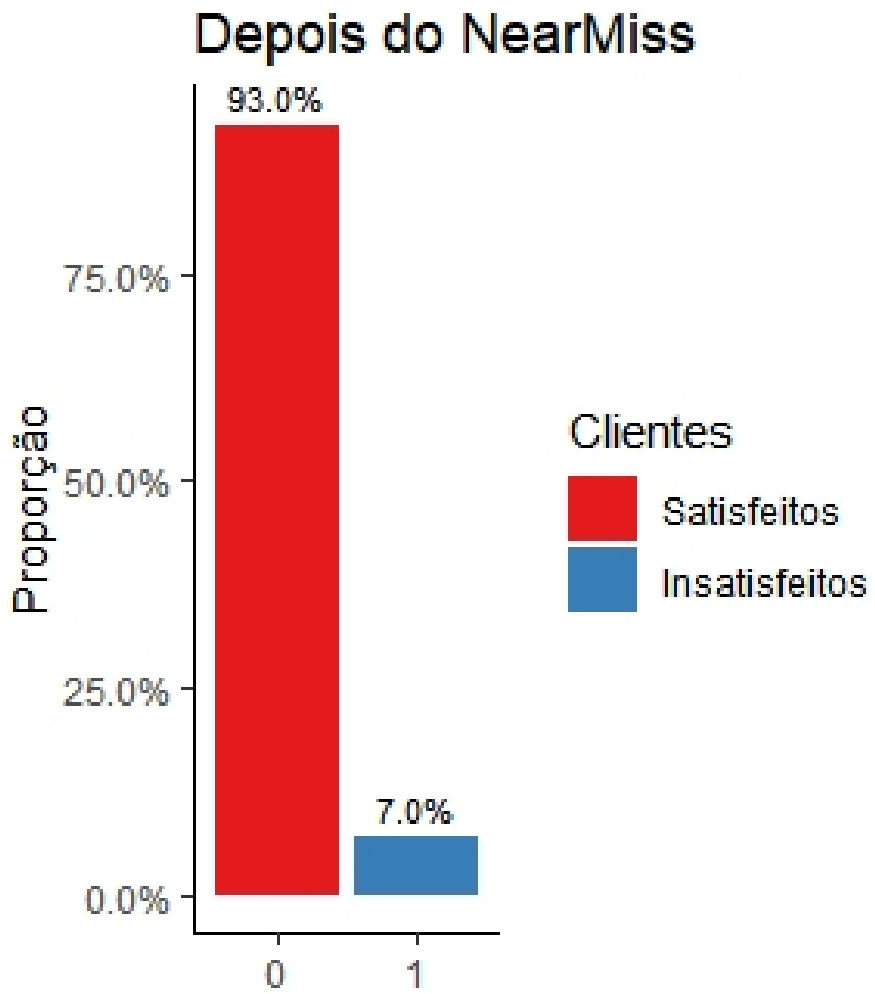
- Outra abordagem de balanceamento foi realizado pelos algoritmos de predição penalizando observações da classe majoritária.

NearMiss

- É um algoritmo que tenta balancear as classes através da técnica UnderSampling.
- São selecionadas um subconjunto de observações da classe majoritária que possuem as menores distâncias médias em relação a observações da classe minotária.
- Foram selecionados quarenta mil observações da classe majoritária e nenhuma observação foi excluída da classe minotária.

Proporção de classes





Proporção de classes

	Clientes Satisfeitos	Clientes Insatisfeitos
Antes NearMiss	73012	3008
Depois NearMiss	40000	3008

Seleção de variáveis

1. Utilizamos o método de informação mútua para seleção de variáveis.
2. Área sob a curva roc tende a não ser uma métrica adequada para dados desbalanceados.
3. Conforme as métricas: Recall, Precision, F1 e F1 médio foi escolhida a quantidade de variáveis a ser usada nos modelos preditivos.

Métricas utilizadas

- Recall = $\frac{VP}{VP + FN}$
- Precision = $\frac{VP}{VP + FP}$
- F1 Score = $\frac{2 * Recall * Precision}{Recall + Precision}$

Informação mútua

É uma medida que fornece o quão distante a distribuição conjunta de duas variáveis aleatórias é do produto das distribuições marginais.

$$IM(X,Y) = \mathbb{E}\left[log\left(\frac{\mathbb{P}(X,Y)}{\mathbb{P}(X)\mathbb{P}(Y)}\right)\right] = \mathbb{E}\left[log\left(\frac{\mathbb{P}(X/Y)}{\mathbb{P}(X)}\right)\right] = \mathbb{E}\left[log\left(\frac{\mathbb{P}(Y/X)}{\mathbb{P}(Y)}\right)\right]$$

Maiores valores de $IM(X,Y)$ sugerem que a variável explicativa fornece mais informação sobre a variável resposta.

Processo de seleção de variáveis

1. Foi realizada uma validação cruzada com o modelo AdaBoost com o acréscimo de variáveis de acordo a informação mútua.
2. A quantidade de variáveis escolhidas foram 31.
3. Algumas variáveis selecionadas são: Idade, quant_produto1, quant_zero, var36, ind_var5, num_var4, num_var5_0, num_var5, var15.

Modelo preditivos

Modelos preditivos utilizados:

1. Análise discriminante linear
2. Análise discriminante quadrático
3. Floresta aleatória
4. AdaBoost de árvores
5. GradientBoosting
6. Regressão Logística

Em alguns algoritmos foram utilizados hiperparâmetros que ajudam no balanceamento de classes.

AdaBoost

1. Boosting que combina vários algoritmos menores.
2. Utilizamos como algoritmo base árvore de decisão.
3. Observações que são classificadas erroneamente são penalizadas.

Avaliação da performance

- A métrica a ser maximizada é a F1 Score da classe minoritária.
- Foi realizada uma otimização de hiperparâmetros com o método Random Search com 20 iterações e validação cruzada de 3 folds.

Resultados

Tabela: Métricas dos modelos preditivos

Modelos	F1-Score	Desvio Padrão
Análise Discriminante Linear	0.3120	0.0209
Análise Discriminante Quadrático	0.3296	0.0162
Floresta Aleatória	0.4411	0.0333
AdaBoost	0.8732	0.0111
GradientBoosting	0.8889	0.0082
Regressão Logística	0.3425	0.0076

Avaliação detalhada

Métricas	AdaBoost	GradientBoosting
Recall-Clientes Satisfeitos	0.9912	0.9978
Recall-Clientes Insatisfeitos	0.8551	0.8145
Precision-Clientes Satisfeitos	0.9891	0.9862
Precision-Clientes Insatisfeitos	0.8799	0.9654
F1 Score-Clientes Satisfeitos	0.9902	0.9920
F1 Score-Clientes Insatisfeitos	0.8672	0.8834
F1 Score Médio	0.9287	0.9377

Referências

1. ROC Curve and Imbalanced Classification: `https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/`
2. Undersampling Algorithms for Imbalanced Classification: `https://machinelearningmastery.com/undersampling-algorithms-for-imbalanced-classification/`
3. Mutual Information: `https://en.wikipedia.org/wiki/Mutual_information#Motivation`
4. FRIEDMAN, Jerome; HASTIE, Trevor; TIBSHIRANI, Robert. The elements of statistical learning. New York: Springer series in statistics, 2001.