Previsão da satisfação dos clientes do banco Santander

Alberto Rodrigues Ferreira, Willian Miranda Araújo

Universidade de São Paulo
albertor@ime.usp.br
w.miranda00@usp.br

1 Introdução

Neste trabalho, elaboramos uma análise de dados completa com o objetivo de predizer se clientes serão satisfeitos ou insatisfeitos no inicio do seu relacionamento do banco Santander. Realizamos alguns pré-processamentos que diminuíram a quantidade variáveis, em seguida realizamos uma análise exploratória de algumas das principais variáveis que possivelmente irão influenciar nos modelos preditivos abordados, identificamos que há um grande problema de desbalanceamento dos dados e tentamos contornar isso optando por um algoritmo do tipo *UnderSampling*. Existe um grande problema bastante evidente que é a grande quantidade de variáveis, realizamos uma seleção de variáveis utilizando da técnica de informação mútua, que estima a dependência de duas variáveis aleatórias. Por fim, testamos 6 modelos preditivos com o método de otimização de hiperparâmetros RandomSearch e decidimos qual o melhor modelo com base na métrica F1 Score.

1.1 Descrição do problema

O problema consister em prever se um cliente é satisfeito ou insatisfeito com base em uma série de variáveis, logo é um problema de classificação. Isso é importante para o banco pois assim que identificado um cliente insatisfeito, poderá será tomada alguma providência imediata para que o cliente não deixe de ser cliente.

1.2 Descrição do conjunto de dados

Nesse trabalho, temos um conjunto de 76020 observações e 370 variáveis explicativas que são anônimas, ou seja, não sabemos o real significado das variáveis.

2 Pré-processamento

Nesta etapa, foi realizadas alguns procedimentos com os dados com o objetivo de melhorar a qualidade, custo computacional dos algoritmos e obter melhores

predições. Houve uma separação entre conjunto de treino(80%) e teste(20%) do conjunto de dados, no conjunto de teste serão realizados os mesmos processos que o conjunto de treino mas só será utilizado na verificação das métricas abordadas, é o conjunto que não faremos nenhuma intervenção de otimização.

2.1 Pré-processamento inicial

Inicialmente existem 370 variáveis explicativas, dessa forma muitas dessas variáveis não tem nenhuma informação sobre a classe, com isso foram retiradas algumas variáveis do conjunto de dados que se enquadram nesse padrão.
Foram realizadas inicialmente dois procedimentos:

1. Exclusão de variáveis com desvio padrão zero.
2. Exclusão de variáveis exatamente iguais.

A exclusão de variáveis com desvio padrão zero significa que os valores dessa variável são iguais em todas as observações e portanto não possuem nenhuma informação. A exclusão de variáveis iguais no conjunto de dados foi realizada pois fornecem a mesma informação, então somente uma dessas variáveis iguais continuaram na análise.
Com esse processo, foram excluídas 66 variáveis irrelevantes, restando 304 variáveis, já contribuindo significativamente para o desempenho da análise de dados.

2.2 Principais variáveis preditivas e criação de variáveis

Nesta etapa, estamos interessados em observar a relação de diferentes variáveis com a variável resposta. Neste ponto, existem 304 variáveis possível que podemos analisar, o que é não traz muita viabilidade. Então, utilizamos um atributo *feature_importance*(importância das variáveis) de três modelos de classificação AdaBoost diferentes, que diferem apenas por possuirem trêm combinações de hiperparâmetros diferentes. O cálculo da importância das variáveis desse algoritmo se dá por meio da exclusão de cada variável para o cálculo da impureza dos nós de cada árvore, quanto maior a redução maior é a importância da variável, isso é realizado para todas as variáveis.

Tabela 1: Importância das variáveis do modelo 1

Variáveis	Importância
var38	0.2915
var15	0.1234
saldo_medio_var5_hace3	0.0672
saldo_medio_var5_ult3	0.0528
saldo_medio_var5_hace2	0.0343
num_var45_ult3	0.0311

Tabela 2: Importância das variáveis do modelo 2

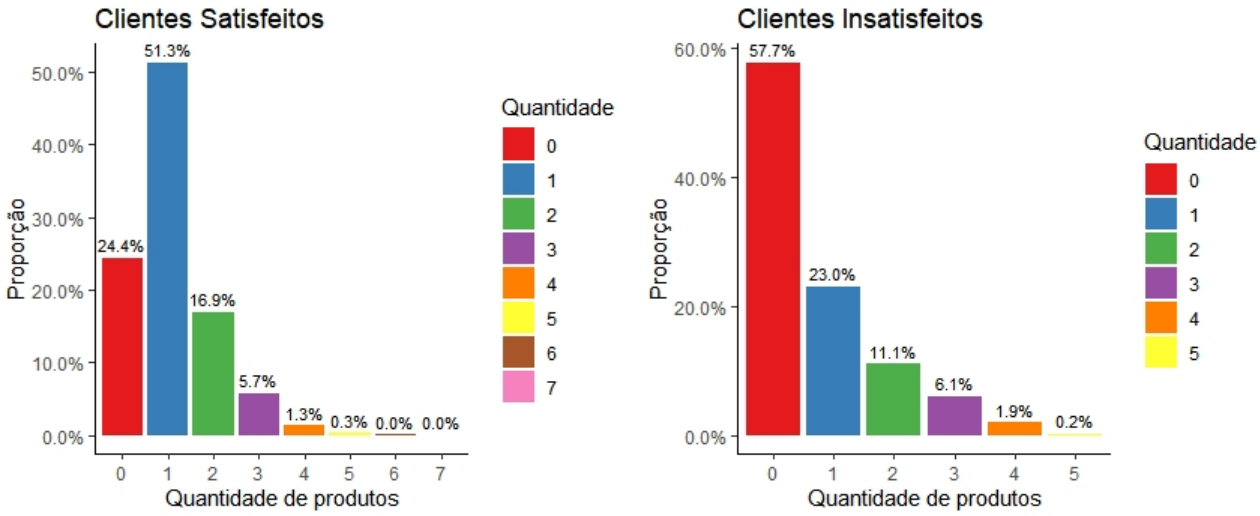
Variáveis	Importância
var38	0.1273
var15	0.0855
saldo_medio_var5_ult3	0.0545
saldo_medio_var5_hace3	0.0527
saldo_medio_var5_hace2	0.0436
saldo_var30	0.0378

Tabela 3: Importância das variáveis do modelo 3

Variáveis	Importância
var15	0.0799
var38	0.0739
saldo_medio_var5_hace3	0.0568
saldo_var30	0.0543
imp_trans_var37_ult1	0.0468
saldo_medio_var13_corto_hace2	0.0462

Logo, exploramos algumas relações de algumas dessas variáveis apresentadas como relevantes e algumas outras que achamos interessantes abordar.

Número de produtos



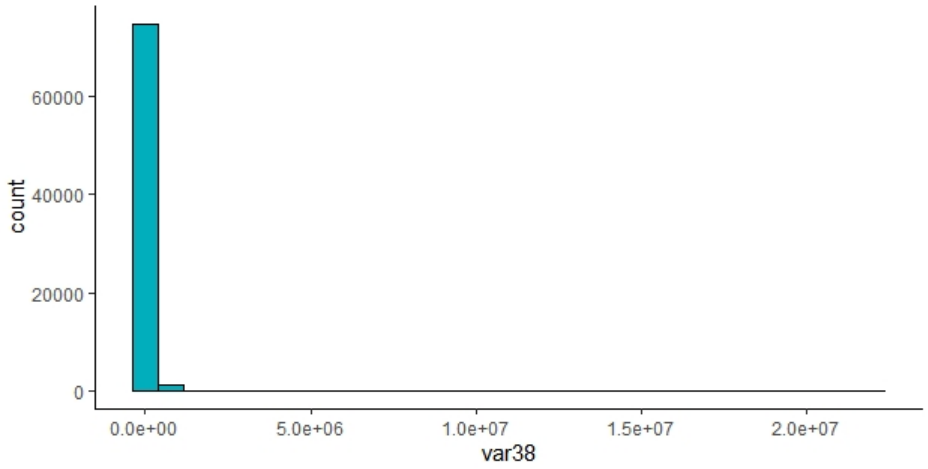
Essa variável nos diz respeito ao número de produtos dos clientes, podemos perceber que clientes satisfeitos têm sua maioria com essa valor igual a um(51,3%), enquanto os clientes insatisfeitos possuem sua grande maioria como valor zero(57,7%), isso faz sentido pois clientes satisfeitos tendem a ter mais serviços do banco. Com base nisso, criamos duas variáveis que verificam os valores dessa variável para cada observação.

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

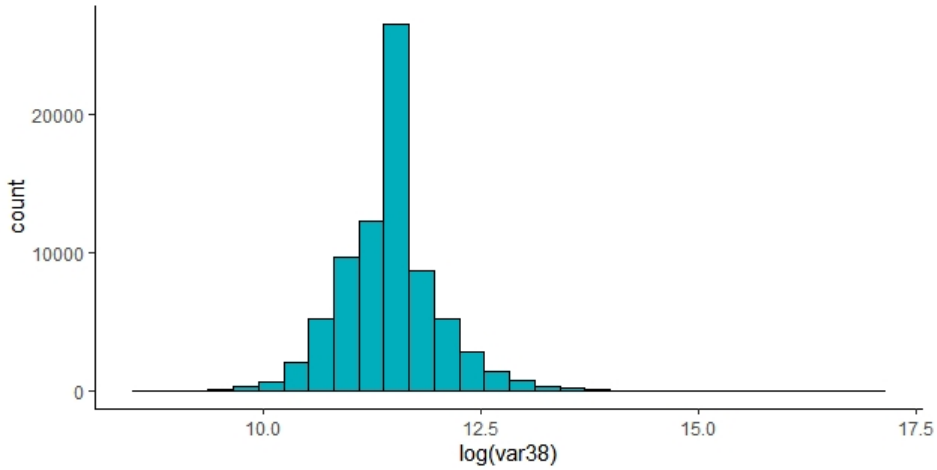
$$\text{quant\_produtos1}=\begin{cases}1 & \text{, se } x=1\\ 0 & \text{, se } x\neq 1\end{cases}$$

Var38

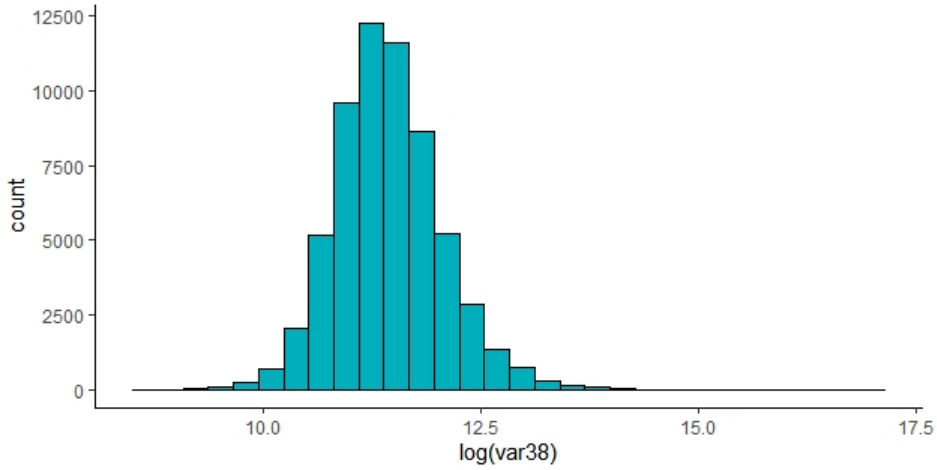
A variável var38, apesar de não ter significado para o problema, apresentou em algumas ocasiões muito importante para a performance dos modelos, então fizemos algumas análises relacionadas a ele.



Não temos muita informação útil nesse histograma, então tentamos algumas transformações e a logarítmica apresentou um bom resultado.



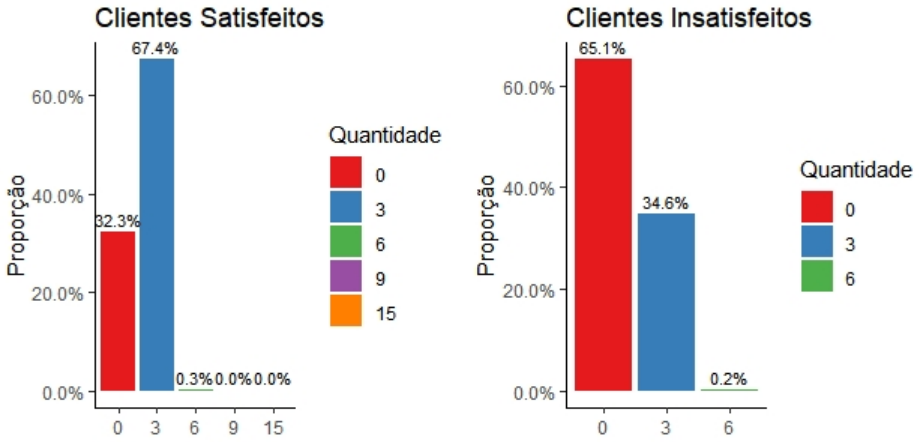
Dessa forma, visualmente aparenta ter aproximadamente uma distribuição normal Realizamos a exclusão da moda dessa variável e então notamos uma melhor aproximação para a distribuição normal que possivelmente possa influenciar na predição.



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

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

num_var5



Com esta variável, podemos notar que grande maioria das observações dos clientes insatisfeitos possuem valor 0 e dos clientes satisfeitos. Com isso, foram criadas duas variáveis *dummy*, que retratam essas observações.

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

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

Idade

Figura 1: Idade

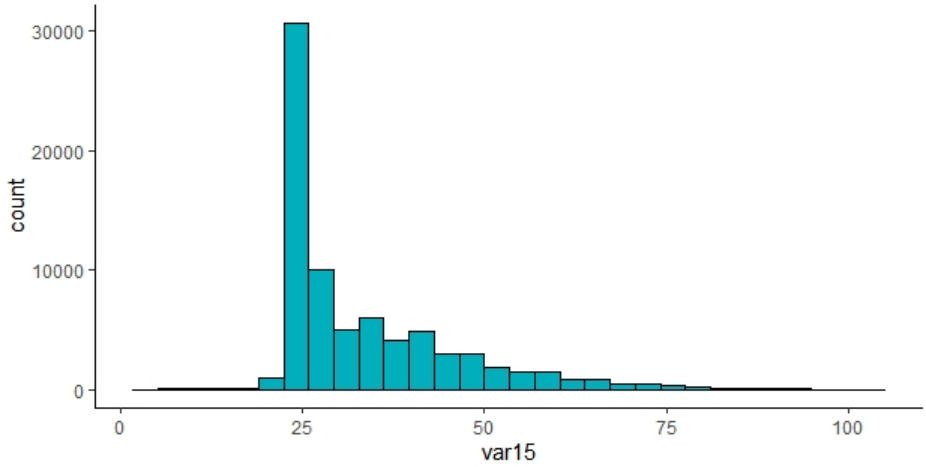
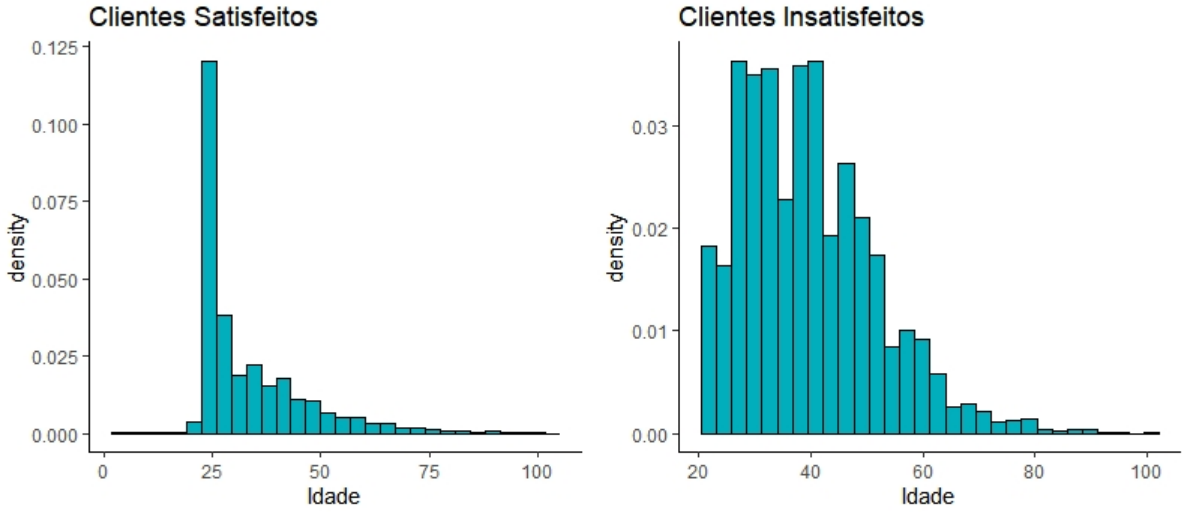


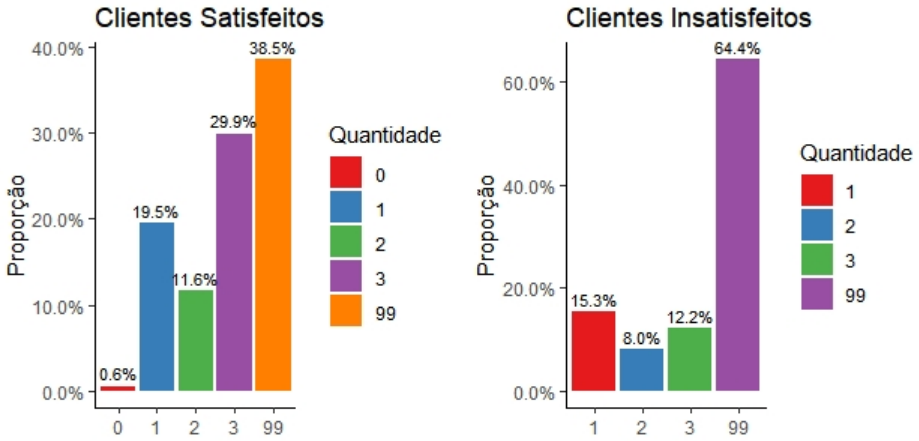
Figura 2: Idade por classe



Podemos perceber que a forma da variável idade por classe são diferentes e também que Clientes Satisfeitos possuem observações com idades menores que 21 anos e quanto aos Clientes Insatisfeitos isso não acontece, logo criamos uma variável dummy que associa a clientes com idades menores que 21, que só acontece nos clientes satisfeitos.

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

var36



Maior parte das observações dos clientes insatisfeitos possuem valor 99 e também observamos que o valor 0 só acontece quando o cliente é satisfeitos, então foram criadas variáveis com base nessas características.

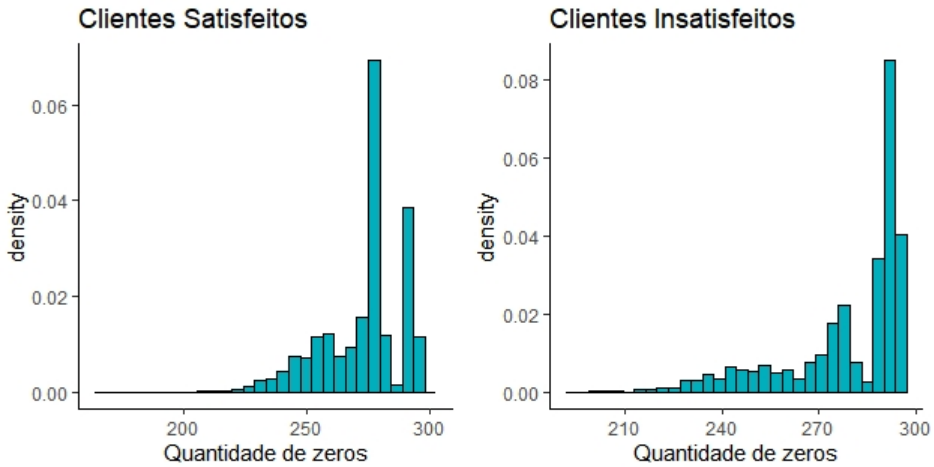
$$\text{var36\_99}=\begin{cases}1 & \text{, se } x = 99\\ 0 & \text{, se } x \neq 99\end{cases}$$

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

Quantidade de zeros

Foi criada uma variável que testamos nos modelos preditivos e aparentou ter bons resultados, inclusive tendo a maior dependência entre a variável resposta. Essa variável simplesmente é a quantidade de zeros em todas as variáveis.

Notamos que as observações dos clientes insatisfeitos possuem uma tendência



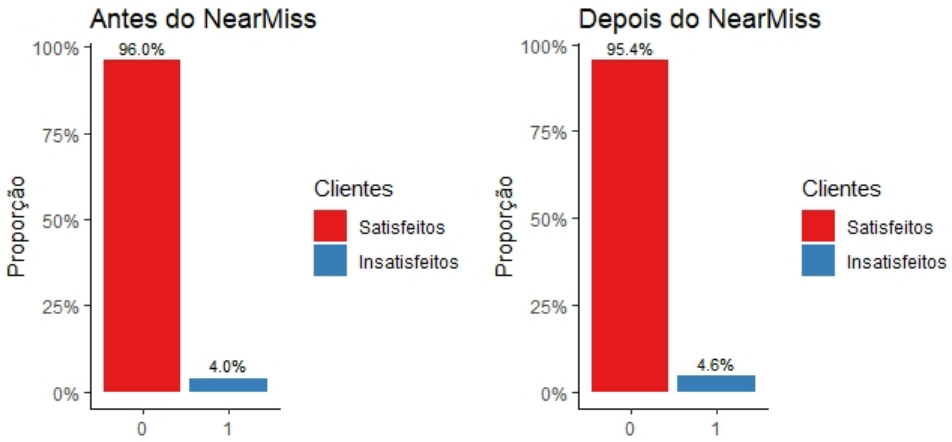
maior a ter mais zeros e com isso pode-se ter uma maior separabilidade entre as classes.

3 Balanceamento de dados

As classes do nosso problema são altamente desbalanceadas, então tentamos contornar essa problema atráves de um balanceamento manual, que exclui observações da classe majoritária de acordo um critério. Os procedimento desse algoritmo estão abordados a seguir:

3.1 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 50000 observações da classe majoritária e nenhuma observação foi excluída da classe minotária.



	Clientes Satisfeitos	Clientes Insatisfeitos
Antes NearMiss	56126	2406
Depois NearMiss	50000	2406

4 Seleção de Variáveis

Um dos principais desafios deste trabalho é lidar com uma grande quantidade de variáveis, com isso elaboramos um processo de seleção de variáveis com base na dependência de cada variável explicativa e a variável resposta. Os procedimentos básicos que realizamos são dados a seguir:

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.

4.1 Métricas utilizadas

- Recall = $\frac{VP}{VP + FN}$
  Intuitivamente, pode ser vista como a proporção de uma determinada classe do conjunto de dados inteiro classificada corretamente.

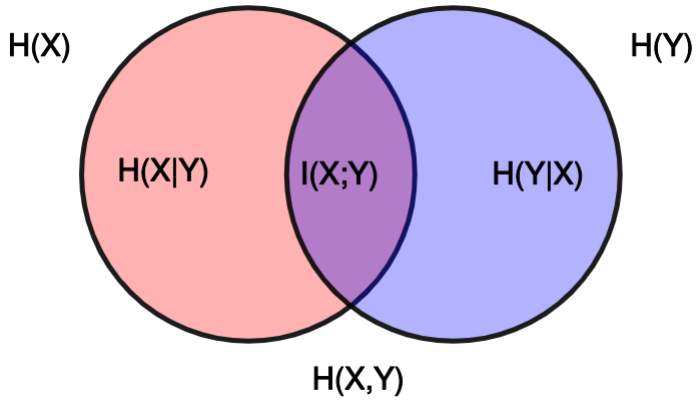
- Precision = $\frac{VP}{VP + FP}$
  Podemos interpretar a fórmula acima como a proporção das observações classificadas em determinada classe (por exemplo, clientes insatisfeitos), terem sido classificadas corretamente.

- F1 Score = $\frac{2 * Recall * Precision}{Recall + Precision}$
  É uma média harmônica entre a Recall e a Precision

Sendo VP(Verdadeiro Positivo), FN(Falso Negativo) e FP(Falso Positivo) valores oriundos da tabela de confusão, que auxilia a encontrar a acurácia de uma classificação.

4.2 Informação Mútua

A IM entre duas variáveis aleatórias, é uma medida de dependência mútua entre as mesmas, i.e., mede a informação compartilhada por X e Y.
Dessa forma, se X e Y são independentes, então a informação mutual entre elas é zero, pois conhecer X não implica em nenhuma informação adicional sobre Y, e vice versa. Por outro lado, quando X é uma função determinística de Y e Y é uma função determinística de X, então toda a informação passada por X é compartilhada com Y.
Portanto, sabendo o valor de X determinamos Y, e vice versa. Além disso, A informação mutual é mais geral do que o coeficiente de correlação e determina o quão diferente é a distribuição do par (X,Y) com relação ao produto das distribuições marginais de X e Y.



Do diagrama anterior, temos que os H's são as entropias das respectivas variáveis e a informação mutual seria o $I(X;Y)$ em violeta.
Em suma, entropia de uma v.a. é o nível médio de informação inerente a ela.
A fórmula a seguir explicita a forma de calcular a informação mútua entre duas variáveis aleatórias.

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

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

4.3 Processo de Seleção de Variáveis

O procedimento para selecionar variáveis relevantes é por meio da métrica informação mútua entre cada variável explicativa e a variável resposta. Maiores valores de $IM(X,Y)$ faz com que a respectiva variável seja possivelmente mais importante para os modelos preditivos. Foi realizada uma validação cruzada de 3 folds com o modelo *AdaBoost* com o acréscimo das variáveis, e para cada acréscimo são calculadas algumas medidas como recall, precision e f1 score, com isso determinamos o número de variáveis a serem utilizadas nos modelos preditivos.

Figura 3: Recall e Precision

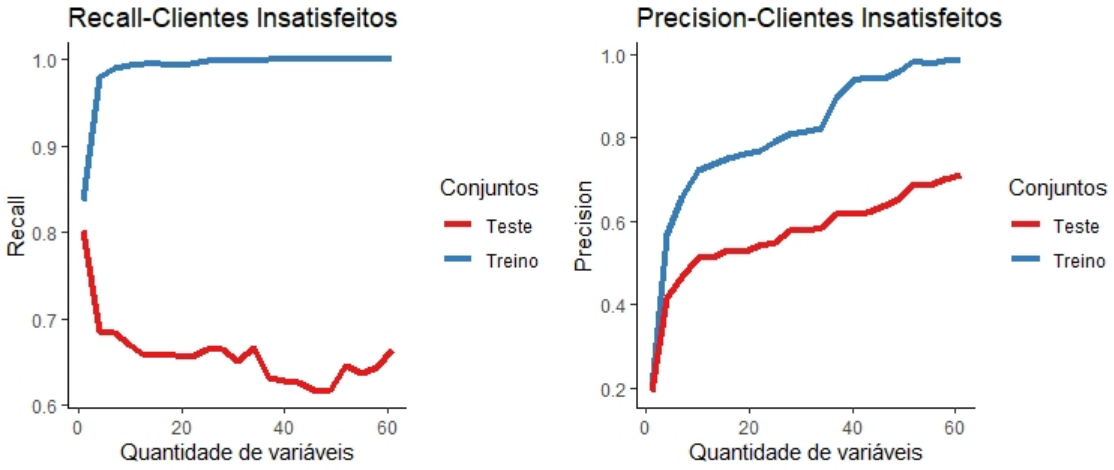
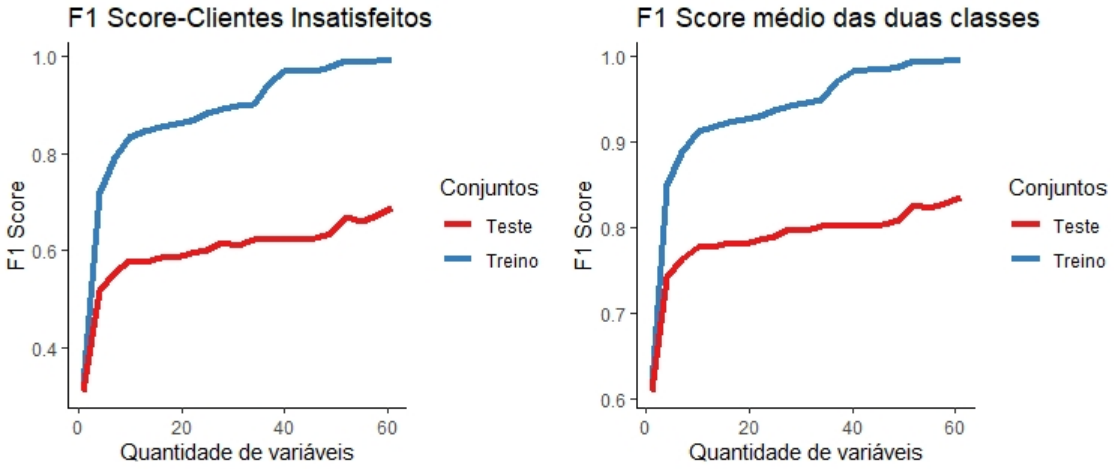


Figura 4: F1 Score



Com o acréscimo das variáveis, podemos observar que a precision da classe Clientes Insatisfeitos sempre aumenta e a recall dos clientes insatisfeitos aparenta se estabilizar entre 0.6 e 0.7. Já a métrica F1 Score aparenta aumentar a medida que mais variáveis são incluídas no modelo. Com isso, selecionamos 31 variáveis, tendo em vista que precisamos estabelecer um meio termo entre quantidade de variáveis e qualidade das métricas abordadas.

5 Resultados

Utilizamos 6 modelos de classificação diferentes e para cada um deles realizamos uma otimização de hiperparâmetos pelo método *Random Search*, que realiza

buscas aleatórias nos espaços dos hiperparâmetros fornecidos.
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. Esses hiperparâmetros penalizamos observações da classe majoritária, assim observações da classe minotária(Clientes Insatisfeitos) possuem mais importância na construção do algoritmo e consequentemente tende a fornecer melhores resultados.
No inicio da análise de dados separamos um conjunto de teste que são 20% dos dados no qual não realizamos nenhuma otimização com os hiperparâmetros, apenas estamos verificando as métricas nesse conjunto para ver o quão bom nosso modelo estão prevendo observações que nunca viram. Optamos por analisar primeiramente a métrica recall do clientes insatisfeitos, que é a proporção das observações insatisfeitos que estão sendo previstas corretamente. Pela tabela, notamos uma superioridade dos algoritmos Floresta Aleatória, AdaBoost e Regressão Logística e para esses algoritmos serão mostrados outras métricas com mais detalhes.

Tabela 4: Métricas dos modelos preditivos

Modelos	Recall
Análise Discriminante Linear	0.0465
Análise Discriminante Quadrático	0.1578
Floresta Aleatória	0.7973
AdaBoost	0.7409
GradientBoosting	0.5648
Regressão Logística	0.7492

Métricas	Floresta Aleatória	AdaBoost	Regressão Logística
Recall-Clientes Satisfeitos	0.7165	0.8671	0.6452
Recall-Clientes Insatisfeitos	0.7691	0.7375	0.7492
Precision-Clientes Satisfeitos	0.9869	0.9877	0.9842
Precision-Clientes Insatisfeitos	0.1006	0.1862	0.08
F1 Score-Clientes Satisfeitos	0.8303	0.9235	0.7794
F1 Score-Clientes Insatisfeitos	0.1779	0.2974	0.1446
F1 Score Médio	0.5041	0.6104	0.462

O melhor modelo é com base na F1 Score do clientes insatisfeitos, que nesse caso obtém valor máximo no modelo AdaBoost que obteve 0.7375 de recall do clientes insatisfeitos e 0.1862 de precision dos clientes insatisfeitos.

6 Códigos

Os códigos realizados neste trabalho estão disponível no GitHub em:
https://github.com/AlbertoRodrigues/trabalho_aplicacao_aprendizagem_estatistica

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.