SME0822 - Análise Multivariada

Relatório Final

06 de Novembro de 2018

Prof. Dr. Osvaldo Anacleto

Alunos:

- David Roberto Cunha Nascimento - 9847515
- Lucas Yudi Sugi - 9293251
- Matheus Araujo Jorge - 9266705
- Thauan Leandro Gonçalves - 9293543
- Willian Ramos - 8936162

1. Introdução

Um modelo de negócios que representa grande parte do mercado há muito tempo é o chamado **B2B**, do inglês *Business to Business*. Nesse segmento, os clientes de uma determinada empresa também são uma empresa, a qual pode negociar diretamente com um consumidor final ou ser outra empresa B2B.

Para quem está tentando vender algum produto ou serviço a outra empresa, é importante saber por quanto tempo a empresa cliente continuará usufruindo do seu negócio. Por parte do fornecedor, ter esse tipo de informação pode levar a estratégias que visam tentar manter o cliente caso ele tenha o objetivo de abandonar sua empresa. Entretanto, muitos clientes não deixam claro isso ao ser contratado, seja por não querer firmar acordos longos de antemão ou pela falta de um projeto de longo prazo.

Atualmente, a coleta de dados é realizada pela maioria das empresas, inclusive as B2B, mesmo sem ter uma ideia precisa do porquê estar coletando aqueles dados específicos. O objetivo deste relatório é apresentar um modelo de classificação que preverá se dado cliente irá abandonar a empresa em um período de dois anos. Para tanto, utilizaremos uma base de dados contendo algumas informações de clientes de uma empresa do ramo B2B.

A Seção 3 descreve a base de dados utilizada em termos do significado de seus atributos enquanto a Seção 4 apresenta uma análise estatística dos dados, buscando informações relevantes ao processo de classificação. A Seção 5 propõe nossos modelos de classificação, indicando algoritmos, técnicas de validação e parâmetros utilizados. As Seções 6 e 7, apresentam, respectivamente, os resultados obtidos e uma breve discussão sobre os mesmos.

2. Objetivos

Nosso objetivo é, portanto, criar um classificador para a empresa capaz de dizer com a maior precisão possível se o seu cliente irá abandonar a companhia em dois anos ou não.

Para o nosso problema iremos considerar que 0 será negativo e 1 positivo, estas categorias estão sendo definidas para podermos calcular as métricas da matriz de confusão.

Assim, em nosso estudo será priorizada a sensibilidade, i.e, a probabilidade de um cliente positivo realmente ser positivo, pois isso é mais benéfico (manter cliente) para a empresa.

3. Descrição da base de dados

A base de dados é composta por 500 observações, ou seja, 500 clientes, da empresa analisada. Cada observação possui 9 atributos, descritos a seguir:

- Cliente: número de identificação do cliente;
- Duração: período de permanência do cliente com a empresa, em dias;
- censura: variável binária, onde o valor “1” indica que o cliente não abandonou a empresa em um período de 2 anos, enquanto o valor “0” indica que a empresa foi abandonada;
- valorGasto: valor médio mensal gasto pela empresa para reter o cliente;
- indB2B: variável binária, onde o valor “1” representa uma empresa do ramo B2B e o valor “0”, uma empresa de outro setor;
- receita: receita anual do cliente, em milhões;
- nEmpregados: quantidade de empregados do cliente;
- TotalProdutos: quantidade de tipos de produtos que a empresa adquiriu enquanto cliente;
- TotalFreq: quantidade total de aquisições do cliente.

Pela descrição dos atributos apresentada, observa-se que nossa variável resposta é a ‘censura’, enquanto as demais serão utilizadas para criar o modelo e dizer se uma empresa ainda será cliente em dois anos.

4. Análise exploratória

O principal objetivo da análise exploratória dos dados é tentar identificar informações a respeito da distribuição dos dados, assim como tentar compreender, através do uso de estatísticas, medidas e gráficos, o processo de geração dos dados. Isso é importante para que possamos decidir quais métodos de classificação deveremos utilizar e quais hipóteses e suposições podemos inferir sobre os resultados obtidos por tais métodos.

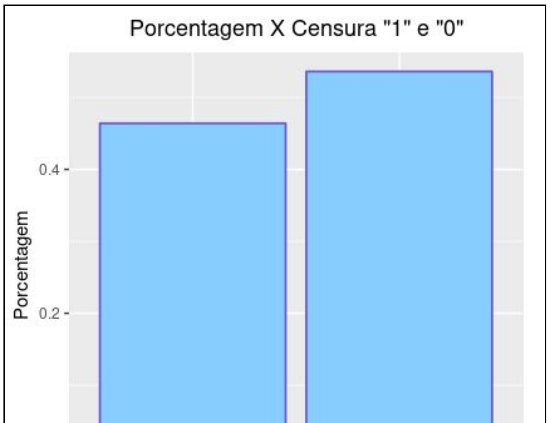

Figura 1: Esta figura representa a porcentagem de empresas que não permaneceram por mais de 2 anos (censura “0”) e as que permaneceram (censura “1”).

Analisando a distribuição das classes entre censura “1” e “0”, concluímos que não possuiremos muitos problemas devido à diferença entre os números de indivíduos pertencentes às duas classes. Isso é interessante pois não precisaremos utilizar técnicas para tratar este problema, como *downsampling* e *oversampling*.

A Figura 2 mostra a correlação entre os atributos numéricos.

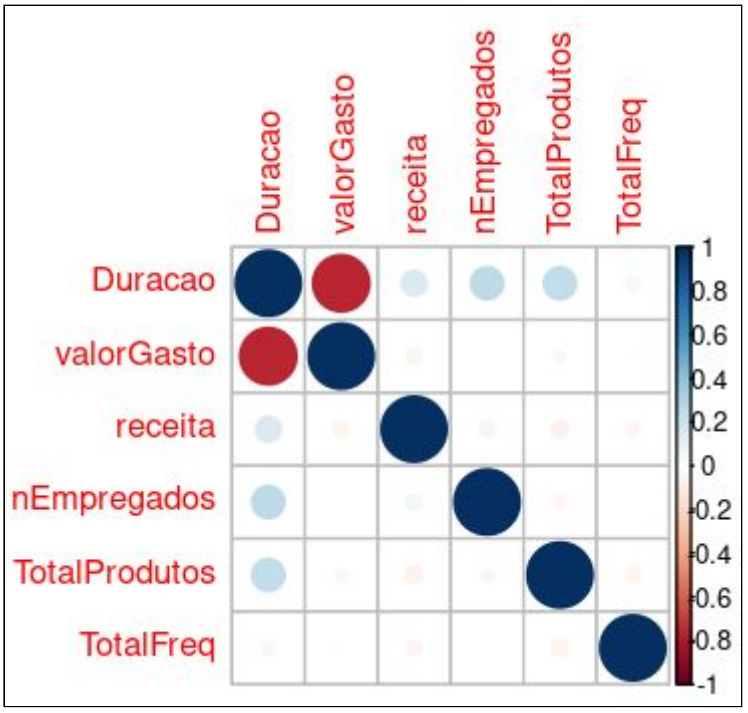

Figura 2: Esta figura mostra a correlação entre os atributos numéricos, tanto contínuos como discretos, sem levar em consideração o atributo *censura.*

Analisando o gráfico acima, podemos perceber uma relação inversamente proporcional forte entre a duração do vínculo com a empresa e o valor gasto com marketing. Isso pode nos indicar que: à medida que a empresa começa a ter um vínculo mais duradouro, é necessário gastar menos com Marketing. Em relação a outra correlação, temos uma relação diretamente proporcional e moderada entre a duração e o número de tipos de produtos diferentes comprados e entre o número total de compras. Esta pode nos indicar que a medida que a empresa mantém os vínculos, ela possui uma tendência a explorar mais os tipos de produtos e, claramente, realizar mais compras. Por fim, podemos concluir também que as outras correlações não se mostraram muito forte.

A Figura 3 a seguir mostra os histogramas dos atributos numéricos da base, tanto os atributos discretos quanto os contínuos. A principal análise que podemos realizar sobre os atributos seriam a suposição de normalidade

multivariada. Porém, analisando os gráficos, concluímos que somente o histograma referente a receita das empresas possui um formato semelhante a normal univariada. O número de empregados, duração e valor gasto mostram claramente uma assimetria.

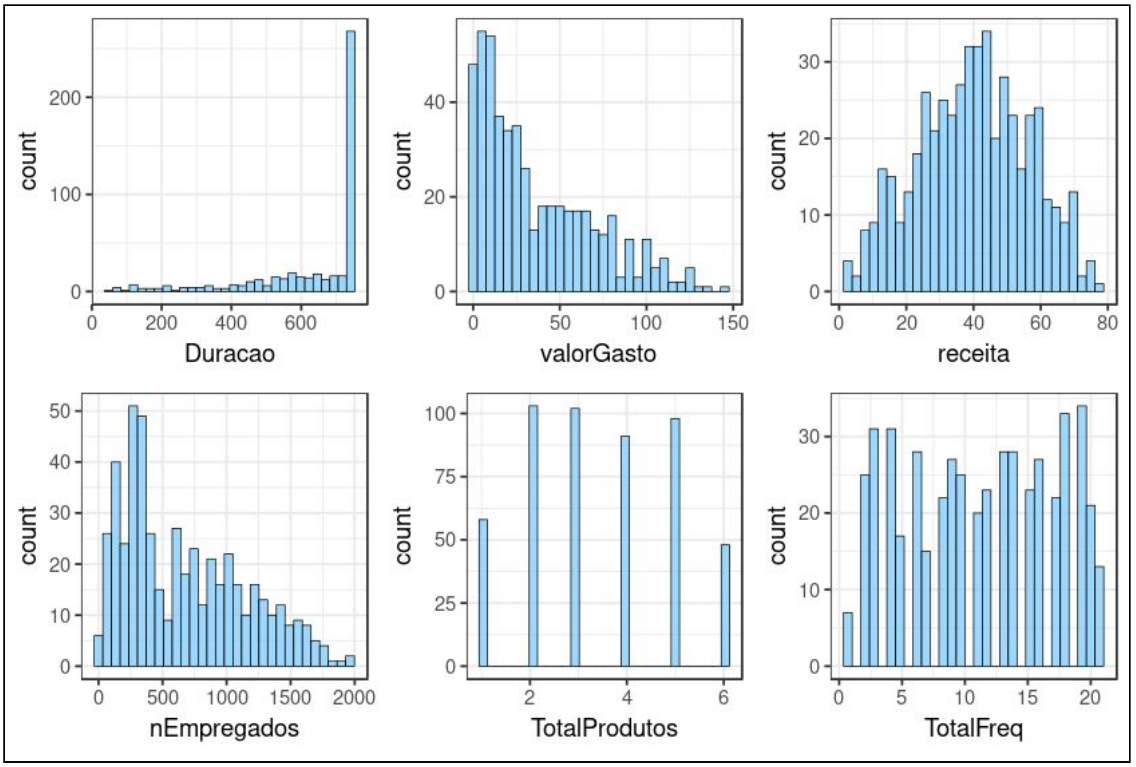

Figura 3: Os 6 gráficos acima mostram os histogramas dos atributos numéricos da base.

Para uma melhor verificação da suposição de normalidade dos dados, foi feito na figura abaixo gráficos, usando a função *qqnorm* no software R, esses gráficos avaliam as distribuições marginais dos vetores aleatórios a partir da amostra disponível dos seus componentes. Sendo assim comparam os quantis da amostra com o quantis esperados da situação onde a amostra segue uma distribuição normal,espera-se que não haja muitos desvios da reta nos pontos.

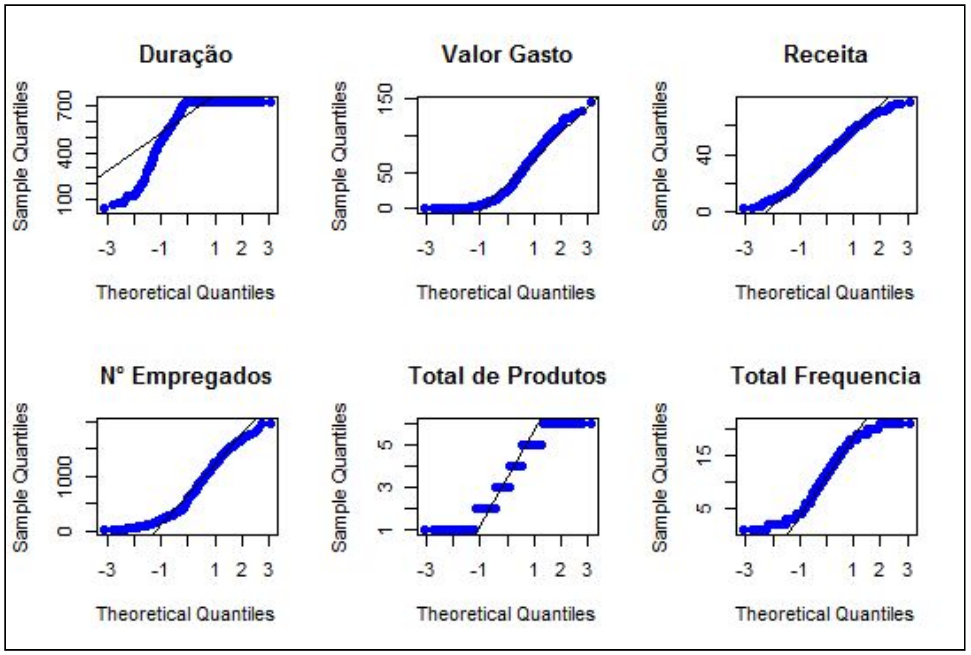

Figura 4: Gráficos do tipo *qqnorm* para as covariáveis da base.

Pode-se observar com o gráfico acima a covariável com maior problema (desvio da reta) foi a de duração, o que era esperado dado sua assimetria do seu histograma visto pela Figura 2 , em relação às outras covariáveis percebe-se poucos desvios da reta, desvios estes mais perceptíveis nas caldas. Além disso para suposição de normalidade multivariada foi feito o gráfico abaixo onde se consiste em avaliar a distância de Mahalanobis que para grandes amostras possui uma distribuição chi-quadrado.

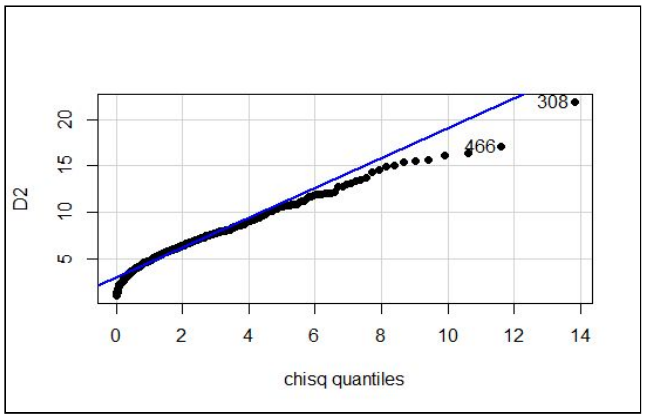

Figura 5: Gráfico da distância de Mahalanobis

Podemos perceber que nem todos os pontos estão em cima da reta, porém não contém uma grande desvio a ponto de desconsiderarmos a suposição de normalidade.

Analisando com maior detalhes algumas das covariáveis foi visto pela figura 2 que as covariáveis valor gasto e duração, apresentam uma correlação forte negativa, é de interesse para uma melhor analise, o gráfico de dispersão entre estas, a fim de obter uma melhor visualização numérica entre essas covariáveis.

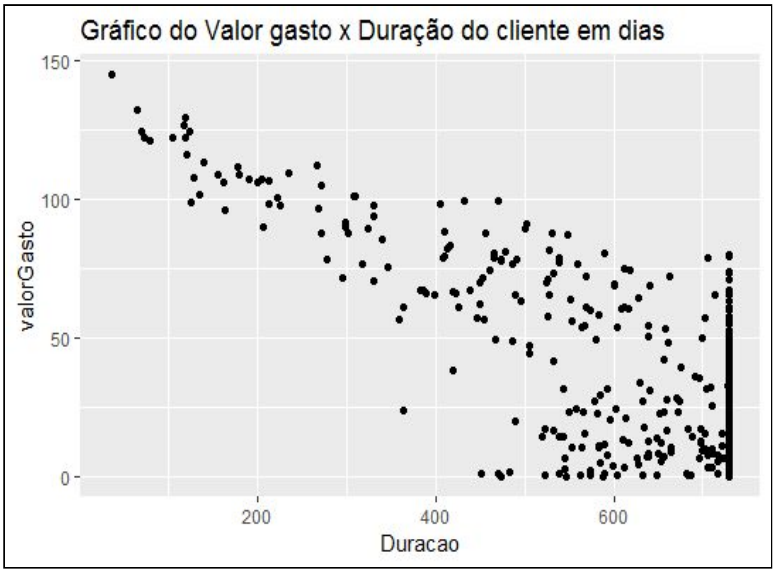

Figura 6: Gráfico de dispersão entre as covariáveis valor gasto e duração.

A alta correlação negativa é melhor explicada pelo figura 6, a dispersão dos pontos em clientes com duração de 600 a mais dias se concentra em valores entre R$0,00 e aproximadamente R$60,00, já em relação entre clientes com duração de 0 a 200 dias o valor gasto é maior, entre aproximadamente R$60,00 e R$150,00. Assim é de interesse apontar que a empresa está tendo um gasto maior com clientes que os abandonam em menos de 2 anos.

Dado que o interesse é analisar os fatores que fazem os clientes ficarem ou não por 2 anos com a empresa, onde temos a variável censura como resposta, descrita na seção 3, segue abaixo os boxplot das covariáveis da base com a variável resposta:

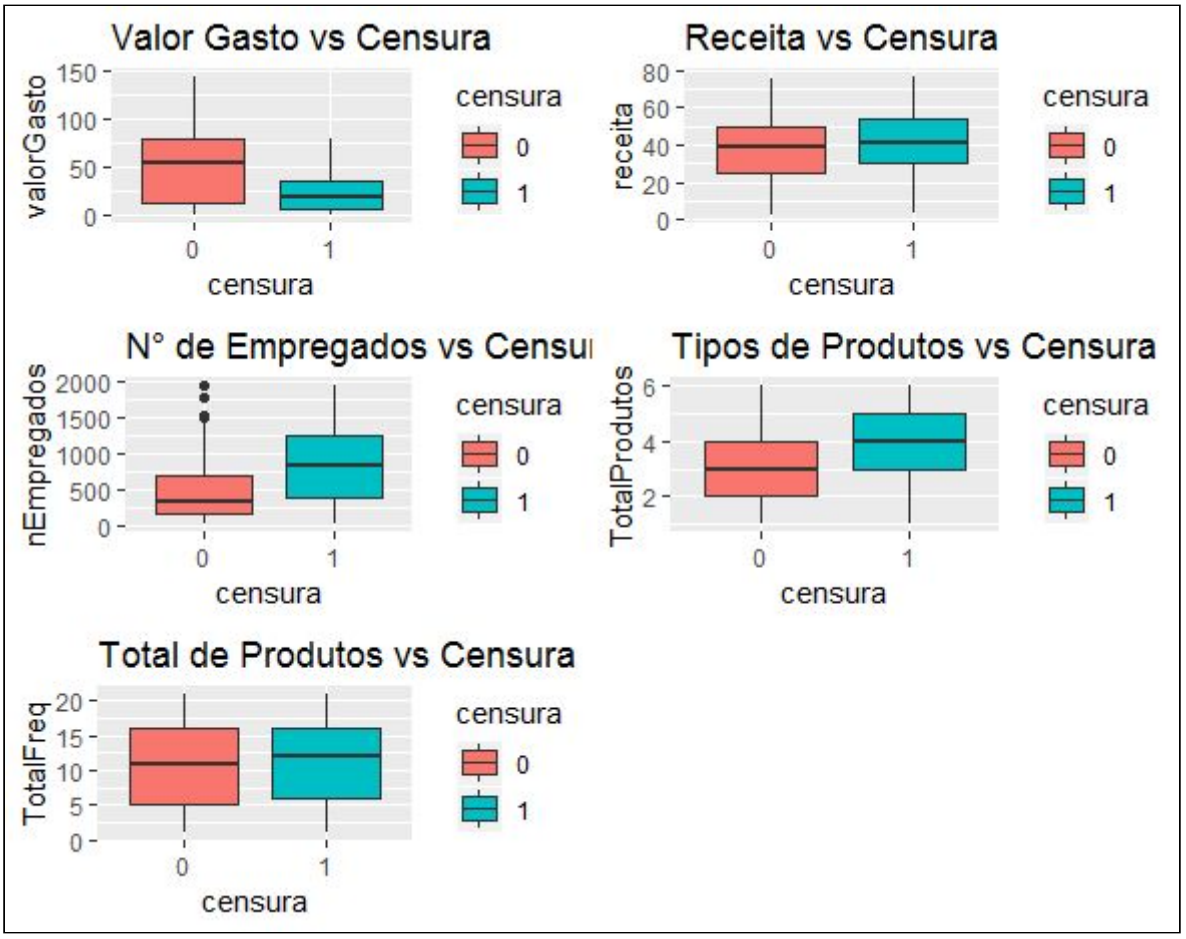

Figura 7: Boxplot das covariáveis da base com a variável censura.

Como era de se esperar o primeiro boxplot (valor gasto vs censura) mostra a discrepância em média, entre os clientes do tipo “0” e ”1”;visto com maior detalhe acima, o segundo boxplot (Receita vs censura), não demonstra uma diferença significativa, pode-se observar pelo seus valores medianos (linha preta horizontal do boxplot) que não existe um deslocamento vertical muito grande entre eles. Já para o terceiro boxplot (N° de Empregados vs

censura), pode se perceber que clientes do tipo “0”, (abandonou a empresa em um período de 2 anos), possui em média um número de empregados menor em relação a clientes do tipo “1”, porém contendo alguns *outliers (valores discrepantes)*, assim se pode ter como base que o número de empregados tem indícios de ser uma covariável significante para as classificações. O boxplot (Tipos de produto vs censura) nos mostra que clientes do tipo “1” tendem a consumir mais tipos de produtos em relação aos clientes do tipo “0”, já o boxplot (Total de produtos vs censura) nos dá uma relação em média semelhante entre os clientes dos dois tipos,tendo uma diferença maior na mediana, porém não tão significante visualmente.

Com a figura 7 descrita acima, já se pode ter a priori alguma das características dos clientes que abandonaram a empresa no período de 2 anos em relação aos que ficaram. Feito essa análise, podemos ter interesse em saber mais detalhes sobre os boxplots onde se apresentou uma maior diferença entre esses dois tipos de clientes, com isso foi feita a figura abaixo, onde mostra com maior riqueza a frequência da quantidade de diferentes produtos vendidos para estes clientes:

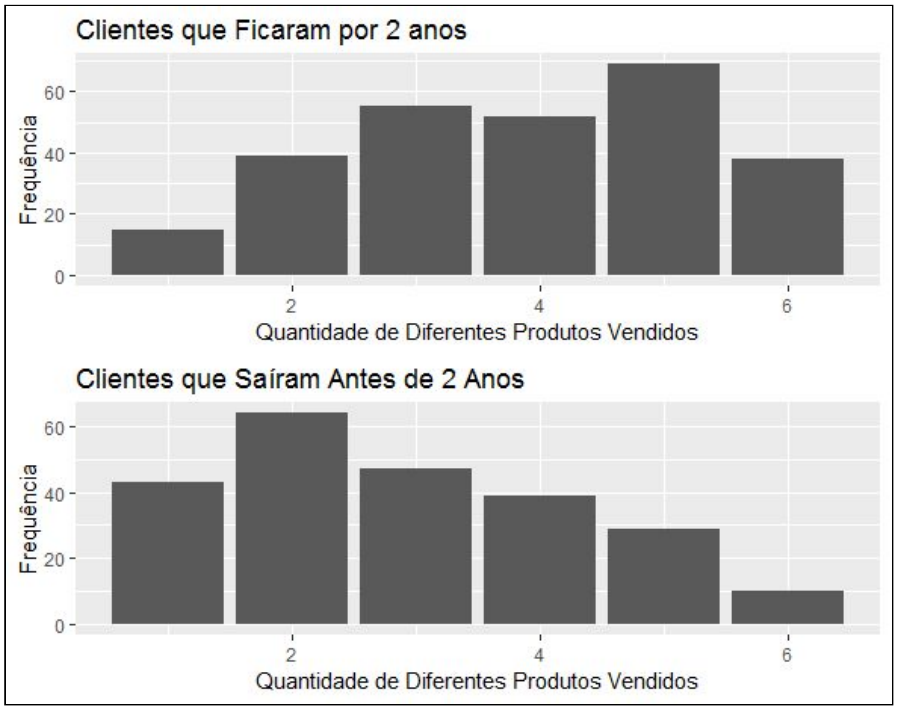

Figura 8: Histogramas da quantidade dos diferentes produtos vendidos para clientes dos dois tipos :“0” e “1”.

5. Modelo de classificação

Para poder resolver o problema de classificação os seguintes modelos foram utilizados:

- Regressão logística.
- Análise discriminante linear (LDA).
- Análise discriminante quadrático (QDA).
- Knn.
- Árvore de decisão (DTREE).

Tais modelos foram avaliados quanto acurácia, sensibilidade, especificidade (métricas obtidas a partir da matriz de confusão) e *Area Under Curve* (AUC) definida como a área abaixo da curva ROC.

Para que houvesse uma maior precisão nas medidas nós utilizamos um processo parecido com o k-fold. Ao invés de dividir o dataset em vários subconjuntos de uma só vez, nós realizamos várias iterações (50 no total) e para cada uma delas dividimos o dataset em treino (60%) e teste (40%). Assim, a partir de 50 valores de cada métrica foi extraído sua média.

Isso foi necessário para que não houvesse um overfit nos classificadores assim como uma maior precisão nas medidas.

6. Modelagem

Iremos descrever aqui como foi realizado a modelagem dos classificadores. Para que houvesse uma melhora na precisão das métricas nós aplicamos a padronização (média 0 e variância 1) nos dados. Após isso iniciamos um processo de seleção das variáveis.

Nessa etapa nós utilizamos as considerações e insights obtidos a partir da análise exploratória em conjunto com vários métodos e testes.

Assim, inicialmente verificamos que a variável Duração causava muitos problemas nos classificadores em principal na regressão logística (separação perfeita) e no QDA. Por causa disso decidiu-se descartá-la.

Após isso, nós utilizamos métodos de seleção de variáveis como forward, backward, stepwise e lasso na regressão logística para se ter uma idéia de quais atributos aparentavam não interferir na classificação. Segundo os modelos gerados por tais métodos e pela análise exploratória verificamos que as seguintes variáveis pareciam ter um efeito bem pequeno na variável resposta:

- indB2B.
- TotalFreq.
- receita.

Sabendo disso, nós decidimos tirar uma por uma nos modelos propostos na Seção 5 verificando se havia um efeito muito grande nas métricas. O que foi observado ao realizar isso é que a ausência das três

variáveis não afetam as medidas calculadas, i.e, retirando-as dos modelos não houve piora e nem melhora nas médias das métricas.

Assim, sabendo que elas não afetam a variável resposta também decidimos retirá-las dado que procuramos, também, um modelo que seja parcimonioso (menos covariáveis) mas que ainda explique bem os dados.

7. Resultados

Após realizar o processo descrito nas Seções 5 e 6 nós obtivemos os seguintes resultados conforme a tabela abaixo:

	Acurácia	Sensibilidade	Especificidade	AUC
Regressão Logística	0.808	0.830	0.783	89.398
Knn	0.830	0.889	0.761	91.081
LDA	0.808	0.833	0.779	89.414
QDA	0.834	0.849	0.816	91.789
DTREE	0.776	0.831	0.712	91.081

Tabela 1: Resultado das médias (50 valores) das métricas.

Pela tabela podemos concluir que existem dois bons classificadores que poderíamos ser considerados no problema que seriam o Knn e o QDA. Ambos serão explicados na próxima seção.

8. Discussão

Os classificadores Knn e QDA podem ser considerados as melhores opções quando comparado aos outros porque:

- Knn: Possui a maior sensibilidade dentre todos os modelos.
- QDA: Possui o maior equilíbrio das métricas dentro todos os modelos.

Conforme definimos na Seção 2 devemos priorizar a sensibilidade, logo, o Knn apresenta como a melhor opção para o nosso problema dado que alcança quase 90% de acerto. Fato este que seria muito interessante para a empresa saber com 90% de precisão os clientes que não irão abandonar após 2 anos.

Salientamos, contudo, que o QDA apresenta-se como um modelo bem equilibrado na sensibilidade e especificidade. Caso a empresa almeja este equilíbrio, tal modelo também é viável.

Desse modo, tendo essas duas opções, que são razoáveis, não cabe a nós definir um modelo único que seria perfeito. Na verdade isso cabe aos tomadores de decisão da empresa a qual estaríamos prestando consultoria. Eles decidiram o que seria melhor (mais sensibilidade vs equilíbrio).

Além disso, iremos fazer uma pequena análise nos modelos Knn e QDA ao atribuir um peso a um custo falso negativo (Cn) em relação a um custo falso positivo (Cp). Na tabela abaixo você pode ver qual o limiar necessário dado o custo associado:

	Custo	Sensibilidade	Especificidade	Limiar
Knn	0.5	86.194	83.190	0.413
	2	93.284	75.431	0.543
QDA	0.5	83.528	85.345	0.451
	2	95.896	69.397	0.703

Tabela 2: Custos relativos (Cn = x * Cp).

Nota-se que com esses custos pode-se atribuir pesos diferentes a um falso negativo/positivo de modo que satisfaça o problema do cliente. Com eles podemos achar o limiar necessário para que haja uma boa classificação.

Interessante verificar que para um custo de 0.5 a sensibilidade e especificidade tendem a ser similares (equilíbrio) mas quando aumentamos esse valor para 2 a sensibilidade é priorizada, já que, esse peso nos indica que desejamos menos falsos negativos.

9. Conclusão

Após todas as etapas de desenvolvimento notou-se que existem três variáveis significativas para a censura, i.e, afetam na sua resposta (cliente abandona ou não a empresa). Tais atributos seriam o número de empregados, valor gasto e total produtos.

Sabendo pela Seção 4 que existe uma relação inversamente proporcional entre duração e valor gasto, e diretamente proporcional entre duração e número de empregados/total produtos, podemos verificar que caso o seu cliente esteja seguindo essas relações lineares ao longo do tempo então é possível que ele não abandone-o.

Entretanto, caso o tempo passe e o número de empregados/total produtos não aumenta mas o valor gasto sim, podemos pensar que irá existir uma insatisfação muito grande do mesmo, logo, ele possivelmente abandonará.

Por fim, concluímos que existem dois bons classificadores para esse problema utilizando as três variáveis ditas anteriormente na qual sua escolha dependerá dos clientes a qual estaríamos prestando serviço, visto que o QDA apresentou-se como um modelo mais equilibrado e Knn com maior sensibilidade.