Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Análise Multivariada

Modelagem de perda de clientes em Marketing (Churn)

Aluno: Bruno Henrique Rasteiro NºUSP: 9292910
Aluno: Mayk Leandro Barbosa Xavier NºUSP: 9379041
Professor: Osvaldo Anacleto

São Carlos
2018

1 Introdução

Churn, numa definição mais generalista, é uma métrica que indica o número de clientes que deixam de se relacionar com a empresa após um contato inicial. Para calcular o churn, o que você precisa fazer é somar o número de clientes que cancelou seu produto/serviço no período analisado.

Churn = total de clientes cancelados

Para que uma empresa consiga fazer a expansão da sua base de clientes é preciso que a taxa de novos clientes (ou *prospects*) exceda o seu churn rate – taxa de clientes cancelados.

É de extrema importância para as agências de marketing entender o máximo possível sobre o churn rate e o que causa ele, sendo que o churn determina se a agência tem um número suficiente de clientes para se manter com sucesso.

No presente trabalho analisamos os dados de perda de clientes de uma empresa e buscamos encontrar relações significativas entre a perda de clientes e outras variáveis como investimento em marketing, receita, número de funcionários entre outras. Sabendo a relação das variáveis mais significativas, aplicamos diferentes técnicas de classificação para identificar os clientes que possivelmente não iriam permanecer por um período superior a 2 anos.

O objetivo principal dessa análise de classificação é desenvolver uma ferramenta que possa identificar possíveis clientes que vão ficar após o período de churn, usando somente os dados de um serviço inicial com várias empresas. Se for possível classificar estes clientes ideais com uma taxa de erro mínima, as agências que usarem essa ferramenta poderão definir quais clientes valem mais a pena e quais são um desperdício de dinheiro.

2 Metodologia

A análise dos dados foi dividia em 2 etapas. A primeira foi a análise descritiva e exploratória dos dados e a segunda a aplicação dos métodos de classificação.

2.1 Análise descritiva e exploratória dos dados

Com a análise descritiva buscamos respostas ou alguma informação sobre as seguintes questões

- Quais variáveis estão mais relacionadas e possuem mais relevância para a classificação?
- Qual a relação das principais variáveis com o churn?

Primeiramente foi feito um conjunto de plots que mostrava a diferença dos valores das variáveis por classe da variável *censura*. Foram feitos box-plots, histogramas e plots de densidade. Esse gráficos tem como objetivo buscar padrões de comportamento das variáveis dentro das classes de *censura* e saber se as variáveis apresentam padrões semelhantes a alguma distribuição conhecida.

Em seguida foi calculada a correlação entre as variáveis do nosso conjunto de dados e foi feita também um mapa de calor (*heatmap*) da correlação para facilitar a visualização da correlação entre as variáveis.

Na busca por uma medida para evidenciar nossas variáveis mais preditivas, usamos o discriminante linear de Fisher, ou critério de Fisher [Bishop, 2006]. O discriminante consiste em gerar uma pontuação para cada variável em relação a comparação de dois grupos, para exemplificarmos esse cálculo vamos supor uma variável v que desejamos discriminar em relação aos grupos 0 e 1 (censura igual a 0 e 1), sua média é calculada respectivamente em relação a cada grupo por $\mu_0(v)$ e $\mu_1(v)$, o mesmo para a variância dada por $\sigma_0^2(v)$ e $\sigma_1^2(v)$. Com isso podemos obter a pontuação $F(v)$ como mostra a Equação 1.

$$F(v) = \frac{(\mu_0(v) - \mu_1(v))^2}{\sigma_0^2(v) + \sigma_1^2(v)} \tag{1}$$

Aplicamos a análise de componentes principais (ACP) a fim de diminuir a dimensionalidade do problema para nos permitir identificar possíveis grupos e observar evidencias de que as variáveis destacadas na análise dos gráficos e no score de Fisher, são as mais representativas para prever nossa variável de interesse (*censura*). Os resultados da ACP é apresentada em um gráfico de barras que mostra a proporção da variância explicada por cada componente juntamente com a proporção acumulada.

2.2 Métodos de classificação

Foram aplicados os seguintes métodos de calssificação.

1. K-vizinhos mais próximo (KNN).
2. Árvore de decisão.
3. Análise discriminante linear(LDA).
4. Análise discriminante quadrática(QDA).

Todos os métodos foram aplicados duas vezes, uma primeira com todas as variáveis da base e uma segunda vez com as variáveis mais representativas, que foram selecionadas a partir da análise descritiva. Para avaliar os classificadores usamos a acurácia média obtida com validação cruzada, juntamente com a curva ROC e a área abaixo da curva (AUC).

A validação cruzada foi feita executando os classificadores em 5 amostras de treino e teste colhidas aleatoriamente, a amostra de treino é constituída de 80% da base enquanto que a de teste representa os 20% restantes. Para cada execução foi calculada a acurácia do classificador e então é extraída uma média e desvio padrão das 5 execuções como forma de avaliação do classificador. Este processo foi aplicada para os 4 classificadores.

A curva ROC é calculada através da função *roc_curve* da biblioteca *sklearn*, a função requer como parâmetro os rótulos verdadeiros das instâncias e a probabilidade da instância ser da classe positiva (no nosso caso 1). Com os parâmetros em mãos, a função obtém valores de *threshold* que dizem qual o limiar da probabilidade para classificar a instância como sendo da classe verdadeira. Por exemplo, o primeiro *threshold* é 0, isso quer dizer que todas as instâncias que tiverem probabilidade menor que 0 serão rotuladas como sendo da classe verdadeira, isso implica que todas serão rotuladas como sendo da classe falsa e o resultado disso é que o classificador acerta todos os negativos (TFP = 0) e erra todos os positivos (TVP = 0). Algo similar ocorre quando temos um *threshold* igual a 1, entretanto nesse caso as duas taxas serão iguais a 1, visto que todas as instâncias serão rotuladas como sendo da classe verdadeira. Entre zero e um estão

os outros valores de *threshold* que vão dizendo as taxas de verdadeiro positivo e falso negativo, ou seja, os pontos no grafíco da curva.

A AUC como o próprio nome diz é a área abaixo da curva, nesse caso a curva ROC, seu valor é interessante pois nos fornece uma medida quantitativa da curva ROC e consequentemente da performance do classificador, sua interpretação é direta, quanto maior seu valor melhor foi o desempenho obtido na classificação.

Particularmente para o *KNN* foi feito um processo a mais para determinar o melhor k para o método. Para isso aplicamos o método para 100 *K's* diferentes (k de 1 a 100). Para cada k é calculado o erro médio quadrático com o mesmo sistema de validação cruzada e é tomado como melhor k o que possui menor erro.

3 Resultados

3.1 Análise descritiva e exploratória dos dados

O principal resultado obtido na análise descritiva e exploratória foi que as variáveis *valorGasto*, *nEmpregados* e *TotalProdutos* são as que possuem um maior potencial preditivo da *censura*, a partir disso resolvemos testar na etapa seguinte (Seção 3.2) o desempenho dos classficadores com todas as variáveis, e com somente as três para ver os resultados dessa redução de dimensionalidade.

3.1.1 Dispersão por classe e correlação

Analisando a distribuição das variáveis na Figura 1, é notável que nenhuma delas exceto a variável receita (que está próxima de uma normal), apresenta um padrão semelhante ao de alguma distribuição conhecida. A variável *nEmpregados* nos mostra que empresas com menos funcionários (aproximadamente < 750) tem uma tendência a deixar de ser cliente antes de completar 2 anos. Além disso, analisando a variável *TotalProdutos* notamos que, em média, as empresas que saíram em menos de 2 anos compraram menos produtos.

É interessante notar também que o valor médio mensal gasto com campanhas de marketing para reter um cliente (*valorGasto*), é menor para a maioria dos clientes que ficaram mais de 2 anos. Observando a correlação na Figura 2 entre a variável *Duracao* e *valorGasto*, vemos que a correlação é forte e negativa, o que implica que o valor gasto com campanhas para reter clientes que ficaram mais de dois anos diminui com o tempo. Isso nos leva a hipótese de que clientes mais antigos provavelmente já conhecem a qualidade dos produtos e serviços da empresa e devido a isso não é necessário um investimento em marketing muito grande para mantê-los. As demais variável possuem uma correlação baixíssima.

Com tudo isso obtemos fortes indícios de que as variáveis que possuem maior correlação com a *censura* e portanto um maior poder preditivo são *valorGasto*, *nEmpregados* e *TotalProdutos*.

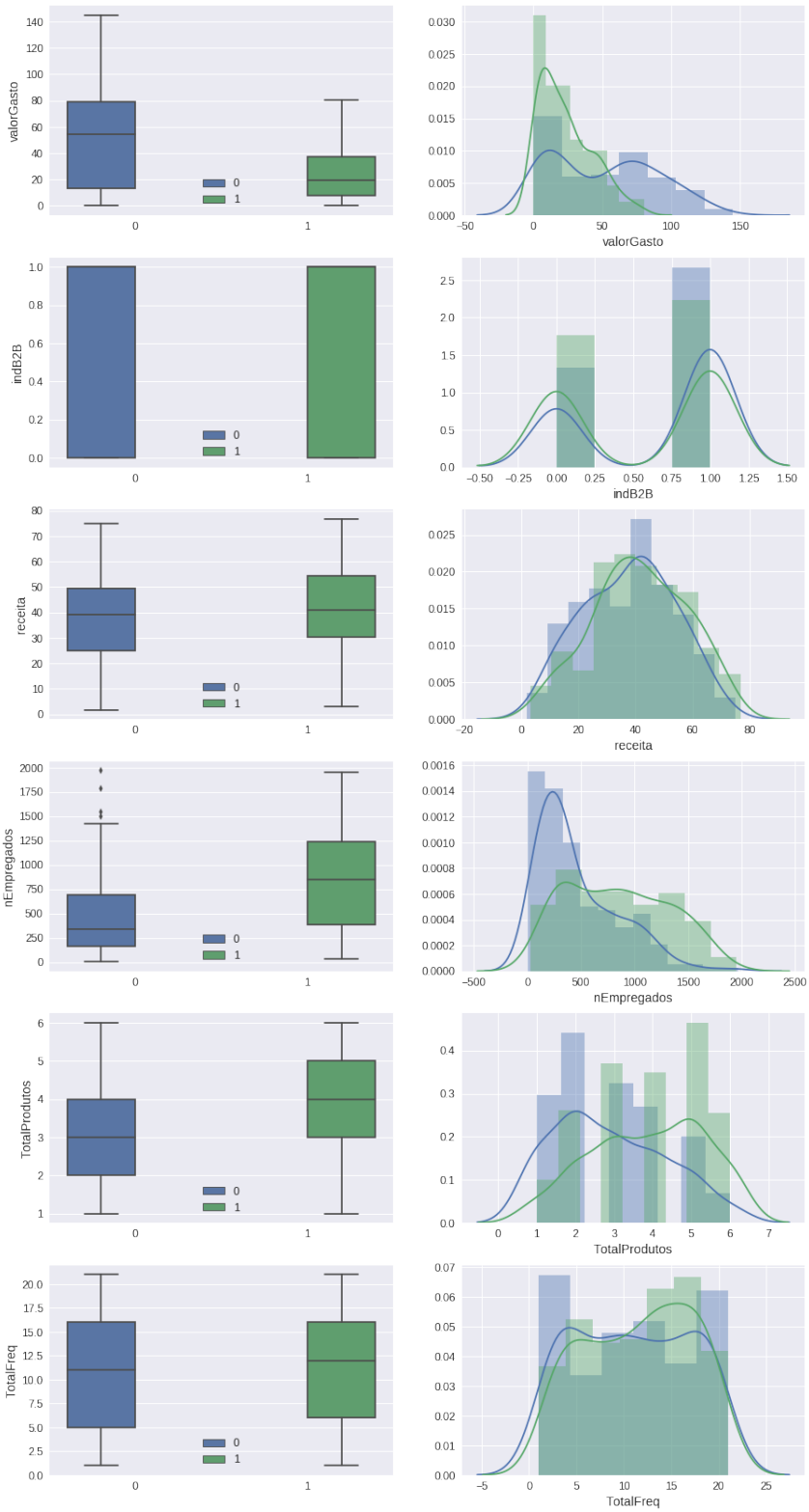

Figura 1: Dispersão por classe

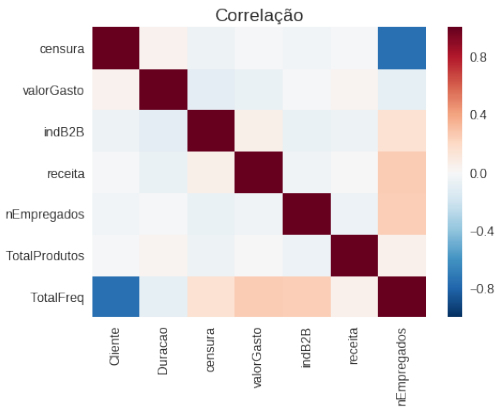

Figura 2: Mapa de calor da matriz de correlação

3.1.2 Discriminante linear de Fisher

O resultado do discriminante linear de Fisher observado na Imagem 3 evidencia a hipótese levantada na análise das distribuição por classe de que as variáveis mais preditivas são *valorGasto*, *nEmpregados* e *TotalProdutos*, e portanto tem uma importância maior para classificação de um cliente.

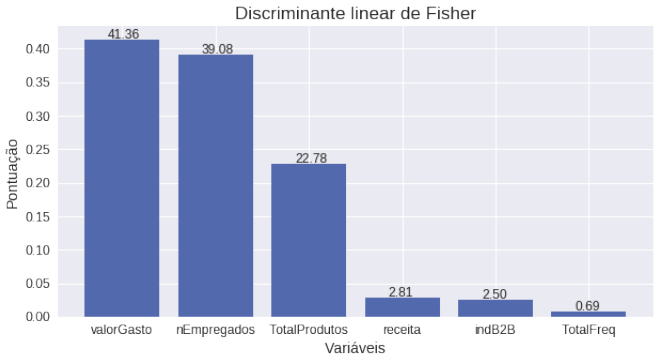

Figura 3: Pontuação do discriminante linear Fisher para cada variável

3.1.3 Análise de componentes principais (ACP)

Devido a baixa correlação entre as variáveis era de se esperar que a ACP não iria resultar em uma redução considerável da dimensão. A variância explicada ficou bem distribuída entre os componentes (cerca de 20% para cada), dessa forma concluímos que não compensa reduzir a dimensionalidade com as componentes da ACP e fazer uma análise com essas componentes. A Figura 4 exibe esse resultado.

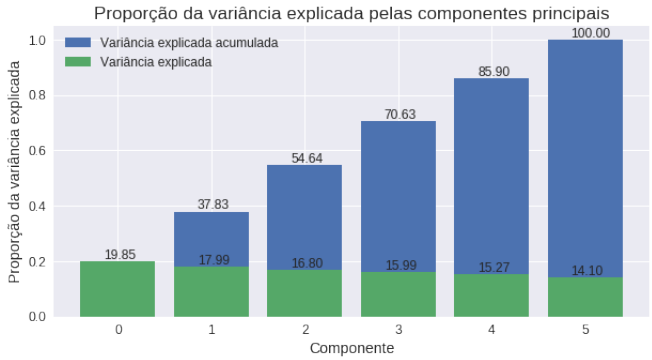

Figura 4: Proporção da variância explicada pelas componentes principais

3.2 Classificação

Depois de feita a análise exploratória dos dados e tendo adquirido um melhor conhecimento sobre a dependência entre as variáveis e a influência delas na censura, escolhemos o melhor valor de K a ser usado no KNN e avaliamos os 4 classificadores citados na seção 2.2 em duas situações, a primeira com todas as variáveis da base e a segunda usando somente as 3 variáveis mais preditivas segundo a análise descritiva.

Em todas as execuções para obter o melhor valor de K, observamos que ele sempre estava entre 1 e 20 conforme mostra a Figura 5, em virtude disso temos a Figura 6 que exibe o erro para os 20 primeiros K's. Nessa execuão obtemos que o melhor valor para K com todas as variáveis foi 8 e com 3 variáveis foi 5, entretanto salientamos que o melhor K varia conforme a execução, mas ele sempre ficou no intervalo de 1 a 20. Outro ponto que observamos foi que o valor de K tende a ser maior quando utilizamos as 6 variáveis.

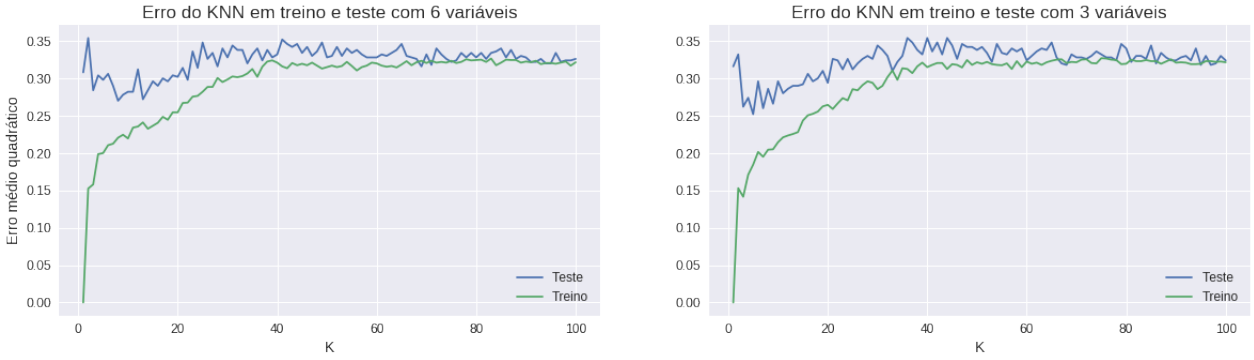

Figura 5: Comparação do erro do KNN em treino e teste com 6 e 3 variáveis

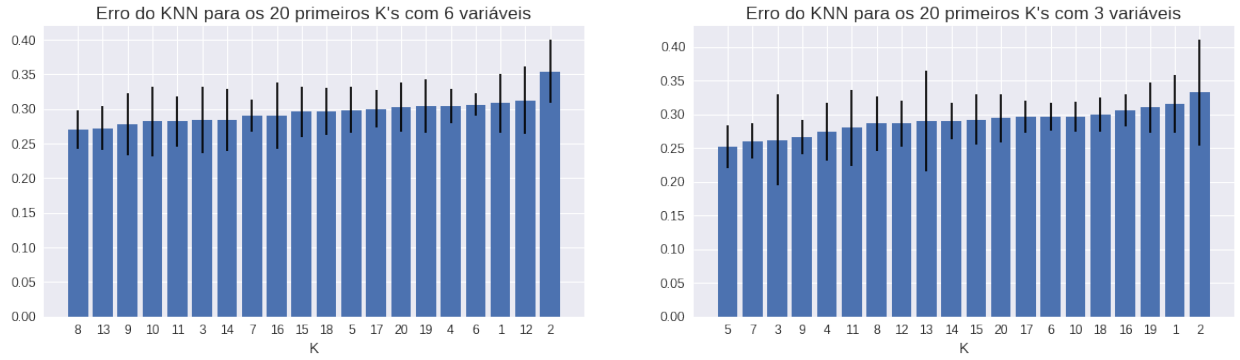

Figura 6: Comparação do erro dos 20 primeiros K's com 6 e 3 variáveis

Dentre os quatro classificadores podemos destacar a performance do QDA e do LDA com uma acurácia média de 84.4% e 81% respectivamente, em seguida temos a árvore de decisão com 76.8% e por último o KNN com 69.6%. Esses valores são referentes ao treino com todas as variáveis da base, observamos pelas Figuras 8 e 7 que não há uma discrepância significativa da acurácia ou da área abaixo da curva ROC quando utilizamos todas as variáveis ou somente as três mais preditivas.

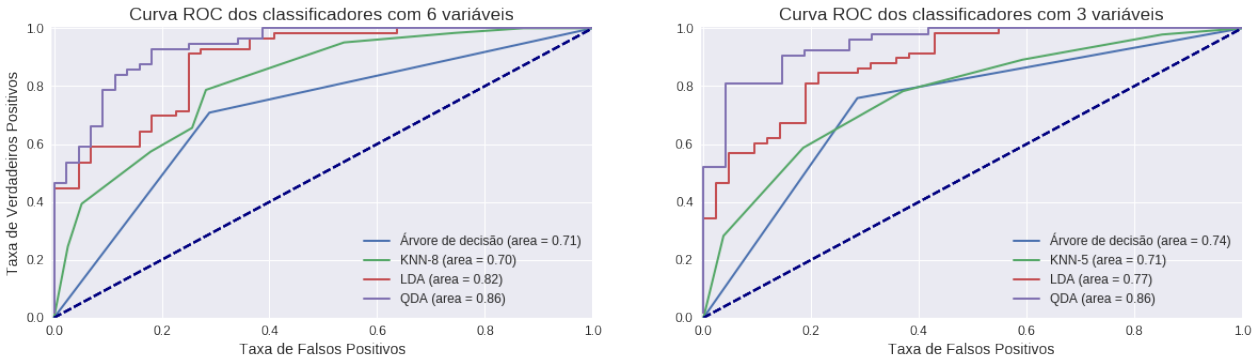

Figura 7: Comparação das curvas ROC dos classificadores com 6 e 3 variáveis

4 Conclusão

Podemos concluir que, para fins de classificação, vale a pena utilizar somente as variáveis *valorGasto*, *nEmpregados* e *TotalProdutos* para classificar os clientes de acordo com a variável *censura*. Isso é possível pois, a precisão dos classificadores usando somente essas 3 variáveis é muito próxima da precisão com as 6 variáveis, como mostra a Figura 8.

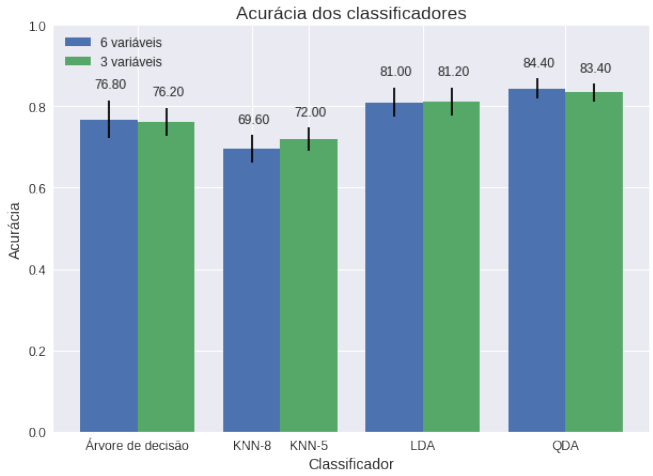

Figura 8: Comparação da acurácia dos classificadores com 6 e 3 variáveis

A partir da análise dos resultados obtidos, é possível concluir algumas práticas que seriam de maior interesse para a retenção do maior número de clientes possíveis no período de 2 anos.

A primeira delas diz respeito a variedade de produtos que os clientes compraram, ou seja, há indícios de que, quanto maior a variedade de produtos adquiridos pelo cliente, maior a chance do cliente manter o vínculo com a empresa.

A segunda prática é relacionada com o número de empregados, aconselha-se que o foco de prospects de novos clientes deve ser nas empresas com um maior número de empregados, como pode ser observado no gráfico do boxplot do número de empregados da 1.

A última prática é de certa forma inconclusiva, pois atráves da análise observamos que quanto menor o valor gasto, maior a chance de se manter o vínculo com o cliente. Porém, essa variável não nos informa a dispersão deste valor gasto ao longo do tempo (só temos a média), e sabemos que é de extrema importância decidir o quanto e quando investir em marketing no cliente. Portanto, concluímos que o valor gasto médio mensal é uma variável influente, porém não podemos aconselhar uma prática concreta sobre ela.

Referências

[Bishop, 2006] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.