Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Antonio C. da Silva Júnior

Utilizando a Regressão Logística para Classificação de Churn em um Ambiente de Startup

Curitiba
2020

Antonio C. da Silva Júnior

Utilizando a Regressão Logística para Classificação de Churn em um Ambiente de Startup

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. WALMES M. ZEVIANI

Curitiba
2020

Utilizando a Regressão Logística para Classificação de Churn em Ambiente de Startup

Using Logistic Regression for Customer Churn Classification in Startup Environment

Antonio C. da Silva Júnior

Campus Santos, Universidade Paulista Av. Conselheiro Nébias 766, Boqueirão, 11045-002, Santos, SP, Brasil*

asdfasdf.
Palavras-chave: asdfasdf, asdfasdf, asdfasdf, asdfasdf, asdfasdf

asdfasdf.
Keywords: asdfasdf, asdfasdf, asdfasdf, asdfasdf, asdfasdf

1. Introdução

A importância do relacionamento de longo prazo entre cliente e empresa é um assunto vastamente discutido na literatura. Devido aos efeitos do aprendizado e à redução dos custos de manutenção, atender um cliente se torna menos dispendioso a cada ano adicional de relacionamento [1]. Por conta do aumento dos custos para atração de novos clientes em um mercado competitivo e à potencial redução dos custos associados aos relacionamentos de longo prazo, a retenção de clientes se torna essencial para a sobrevivência e o sucesso econômico das empresas do setor de serviços [2]. De acordo com [3], dependendo do estudo e do segmento no qual a empresa está inserida, o custo para adquirir um novo cliente pode ser de cinco a vinte e cinco vezes superior ao da manutenção de um cliente já existente.

O desenvolvimento de estratégias para retenção de clientes se tornou uma prática comum entre empresas de diversos segmentos, e em consequência, antever o abandono de clientes se tornou um anseio constante. Em um momento de generalizados esforços na direção da cultura orientada a dados, os modelos preditivos para detecção de abandono de clientes, predominantemente utilizados por grandes companhias no setor de telecomunicações, se tornaram ferramentas populares nas empresas, independentemente da magnitude e da área de atuação.

A literatura comprova que a modelagem preditiva de abandono de clientes, também conhecida como modelagem preditiva de churn, é um tema bastante explorado e que possibilita inúmeras maneiras de desfecho: Botelho e Tostes [4] ajustaram um modelo de regressão logística para predizer a probabilidade de churn em uma grande empresa de varejo; Vafeiadis et al. [5] tiveram sucesso, entre os métodos comparados, na classificação de churn através do SVM (kernel polinomial) com AdaBoost em uma empresa de telecomunicações; Baseados nos dados de avaliações online de clientes, Kumar e Yadav [6] propuseram um modelo preditivo de churn baseado em regras através de redes neurais artificiais e teoria dos conjuntos aproximados.

Com base nos dados de uma startup brasileira que tem como principal produto uma plataforma digital para conectar vendedores de diversos segmentos aos grandes marketplaces, a proposta deste artigo é apresentar um modelo preditivo que possibilite não só a classificação de vendedores propensos a abandonar a empresa, mas que também permita a interpretação dos motivos que possivelmente estejam impactando a predição. Diante da variedade de técnicas disponíveis e das particularidades de cada modelo de negócio, a escolha do algoritmo adequado se torna uma etapa crucial do processo de modelagem. Portanto, tendo como referência a abordagem de [7], que utilizaram uma modelagem híbrida multicritério considerando múltiplos decisores para a escolha de um modelo pre-

*juniorssz@gmail.com

ditivo de churn, o algoritmo escolhido para desenvolver o classificador proposto foi a regressão logística.

2. Materiais e métodos

2.1. Estruturação do conjunto de dados

Os dados utilizados neste trabalho referem-se a clientes de uma startup paranaense, anonimizados e com variáveis quantitativas padronizadas com média 0 e desvio padrão 1. Considerando que estes clientes contrataram uma plataforma digital que possibilita a venda de produtos nos principais marketplaces, neste trabalho eles serão chamados de vendedores. Devido às características da arquitetura do banco de dados e às particularidades do negócio da companhia, houve a necessidade de realizar um longo processo de data wrangling. Este processo inicia-se por um diagnóstico preliminar dos dados, ou seja, se estão no formato adequado, se respondem as perguntas que motivaram a análise e o que é necessário para colocá-los no formato ideal. Em seguida avalia-se a ocorrência de dados faltantes, valores inconsistentes e duplicatas e, por fim, realiza-se um processo de limpeza e transformação, de modo a se obter um conjunto de dados adequado para o estudo [8].

2.1.1. Definição da resposta e covariáveis de desempenho

Inicialmente foram definidos como churn ($Y = 1$) os vendedores que estiveram inativos por 30 dias corridos desde a data da última atividade e permaneceram no mesmo estado em definitivo, considerando como atividade o acesso à plataforma digital ou a ocorrência de uma venda online. Em seguida, em função da data de corte estabelecida conforme a Tabela 1, foram mantidos no conjunto de dados somente os vendedores com pelo menos 90 dias de histórico. O período de 90 dias, finalizado na data de corte, foi dividido igualmente em dois subperíodos, onde foram calculadas métricas como faturamento, ticket médio, quantidade de produtos publicados, quantidade de pedidos cancelados, número de dias em atividade e etc., em cada cada um dos subperíodos. Em seguida, através da Equação (1), foi calculado desempenho do vendedor em função de diversas métricas, onde $V1$ e $V2$ são os valores calculados para cada subperíodo.

As métricas desempenho, data à natureza da equação de origem, possuem o comportamento explicado pela Tabela 2. Ao término desta etapa foi obtido um

Tabela 1: Definição da data de corte

Vendedor	Data de corte
Definido como churn	Última atividade
Em atividade normal	Realização da análise

Tabela 2: Interpretação das métricas de desempenho

Valor	Desempenho
0,5	Mantido
> 0,5	Aumentado
< 0,5	Reduzido

conjunto de dados composto pela variável resposta (churn) e, como covariáveis, 10 métricas de desempenho, onde cada observação representa um vendedor.

$$Desempenho = \frac{V2}{V1 + V2} \tag{1}$$

2.1.2. Adição de outras covariáveis

Foram adicionadas covariáveis qualitativas que representam o estágio do vendedor, o plano contratado e a região de origem, bem como covariáveis quantitativas como o faturamento total, total de produtos publicados, quantidade total de pedidos e etc., resultando em um conjunto de dados com 26 covariáveis.

2.1.3. Criação de covariáveis binárias

Dada a necessidade de analisar o comportamento da variável resposta em função de uma covariável qualitativa com n categorias, deve-se criar $n-1$ covariáveis binárias (dummies), que assumem valores iguais a 0 ou 1, ficando por conta do pesquisador decidir qual das categorias será a referência (dummy = 0) [9]. Portanto, as covariáveis qualitativas adicionadas foram transformadas em binárias, resultando em um conjunto de dados composto por 35 variáveis e 11.131 observações.

2.2. Modelos de regressão logística binária

O objetivo da regressão logística binária é o estudo da probabilidade de ocorrência de um evento de interesse (Y), apresentado na forma dicotômica ($Y = 1$ se o evento de interesse ocorrer; $Y = 0$, caso contrário), em função de um vetor vetor de covariáveis $(X_1, ..., X_n)$. Sua definição ocorre através da Equação (2), onde β_j $(j = 0, 1, 2, ..., p)$ representam os parâmetros a serem

estimados, sendo β_0 o intercepto e os demais, parâmetros de cada covariável. E o subscrito i representa cada observação da amostra ($i = 1, 2, ..., n$) [9].

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} \quad (2)$$

A Equação (2), conhecida como logito, modela a log-chance de ocorrência do evento de interesse. Portanto, para obter uma expressão para a probabilidade de ocorrência do evento é necessário isolar matematicamente π_i, resultando na Equação (3).

$$\pi_i = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_{i1}+...+\beta_p X_{ip})}} \quad (3)$$

A estimação dos parâmetros β_i é realizada por máxima verossimilhança, método que consiste em encontrar, através da programação linear, os parâmetros que maximizam a função de verossimilhança representada através da Equação (4).

$$L = \prod_{i=1}^{n} \left[\pi^{Y_i}(1-\pi_i)^{1-Y_i}\right] \quad (4)$$

Entretanto, matematicamente é mais conveniente trabalhar com o logaritmo da função de verossimilhança, conhecido como função de log-verossimilhança [9, 4], representado através da Equação (5).

$$logL = \sum_{i=1}^{n} \left\{\left[Y_i ln(\pi_i)\right] + \left[(1-Y_i) ln(1-\pi_i)\right]\right\} \quad (5)$$

2.3. Seleção de covariáveis através do algoritmo stepwise

A comparação entre dois modelos pode ser realizada através do Critério de Informação de Akaike (AIC), definido por $-2logL+2p$, onde $logL$ é a log-verossimilhança maximizada e p é o número de parâmetros do modelo, devendo-se selecionar o modelo que apresentar o menor valor de AIC. Entretanto, avaliar todas a regressões possíveis, mesmo para um número moderado de covariáveis, pode ser computacionalmente invivável, mas como alternativa é possível utilizar o algoritmo *stepwise*, que funciona da seguinte forma [10]:

1. Ajuste do modelo com todas as covariáveis;
2. Avaliação tanto a exclusão como a inclusão de cada covariável, através do cálculo do AIC;
3. Exclusão (ou inclusão) da covariável cuja exclusão (ou inclusão) resulta em menor AIC;
4. Repete-se os passos 1 a 3 para o novo modelo e o processo continua até que nenhuma exclusão (ou inclusão) que resulte em menor AIC.

2.4. Validação cruzada por K-fold

A validação cruzada por k-fold é uma técnica de reamostragem aplicada com o propósito de reduzir o vício e a variância do modelo, que funciona da seguinte forma [11]:

1. Separação aleatória dos dados de treino em k partições de tamanho aproximadamente igual (folds);
2. Isolamento de uma das partições e treino do modelo com os dados das demais partições concatenadas;
3. Validação do modelo com os dados da partição isolada, através de determinada métrica de avaliação;
4. Repete-se os passos 2 e 3 até que o modelo seja validado em todas as k partições;
5. A métricas de avaliação de cada iteração são resumidas, normalmente através da média aritmética.

2.5. Ajuste do modelo de regressão logística binária

O conjunto de dados foi separado aleatoriamente em duas partes, garantindo a proporção aproximada de 47,3% de ocorrência de churn ($Y = 1$) em ambas as amostras. A amostra maior, com 75% dos dados, foi utilizada para o ajuste de dois modelos à partir de todas as covariáveis disponíveis. Um deles, denominado *modelo completo*, foi ajustado da maneira tradicional, ao passo que o segundo modelo, denominado *modelo restrito*, teve o ajuste realizado através do algoritmo *stepwise*.

Através do teste da razão da verossimilhança, representado através da Equação (6), é possível verificar a qualidade do ajuste do *modelo completo*, ajustado com j covariáveis, em comparação com o *modelo restrito*, ajustado com $j-k$ covariáveis, sendo k o número de covariáveis removidas do ajuste [9]. Quando a estatística do TRV é inferior ao valor da distribuição do χ^2 com k graus de liberdade e 5% de significância, não rejeitamos a hipótese nula, ou seja, concluimos que a remoção de k covariáveis não afeta a qualidade do ajuste do modelo.

$$TRV = -2(logL_{completo} - logL_{restrito}) \quad (6)$$

Ao realizar o teste, foi constatado que a remoção das 14 covariáveis, no *modelo restrito*, não alterou a qualidade do ajuste, uma vez que a estatística do teste foi inferior ao valor da distribuição do χ^2 com 14 graus

Tabela 3: Covariáveis utilizadas pelo modelo

Covariável	Descrição	Medição
Métricas de desempenho		
x1	Desempenho do número de pedidos no período	percentual
x2	Desempenho do número de dias em atividade no período	percentual
x3	Desempenho do número de pedidos cancelados ou suspensos no período	percentual
x5	Desempenho da média de dias de atraso (postagens) no período	percentual
x6	Desempenho da média de dias de atraso (entregas) no período	percentual
x7	Desempenho do faturamento no período	percentual
x10	Desempenho do ticket médio no período	percentual
Qualitativas		
x12	Necessita de dias adicionais para postagem	1 = sim; 0 = não
x13	Está no estágio B	1 = sim; 0 = não
x14	Esta no estágio I	1 = sim; 0 = não
x15	Está no estágio R	1 = sim; 0 = não
x17	Possui o plano M	1 = sim; 0 = não
x21	Localizado na região Sudeste	1 = sim; 0 = não
Quantitativas		
x23	Valor do fluxo de caixa	contínua
x25	Valor total do faturamento	contínua
x30	Número total de produtos publicados	discreta
x32	Número de dias desde a contratação até a primeira venda	discreta
x33	Número de dias desde a contratação até estar pronto para operar	discreta
x34	Número de dias em atividade no período	discreta

de liberdade e 5% de significância. Portanto, optou-se pelo *modelo restrito* para a continuidade do estudo, uma vez que este possui complexidade inferior com relação ao *modelo completo*, sem perda de qualidade. A Tabela 3 exibe as 19 covariáveis selecionadas para o modelo.

A Tabela 4 apresenta as estimativas dos parâmetros do modelo para cada covariável utilizada. Através da estatística z de Wald, definida pela Equação (7), onde $\hat{\beta}_j$ é a estimativa de um particular parâmetro β_j do modelo e $ep(\hat{\beta}_j)$ é o seu erro padrão, é possível obter a significância estatística de cada estimativa. Calculadas as estatísticas z de Wald, através da distribuição normal padrão a um determinado nível de significância, obtemos os respectivos valores críticos e verificamos se estes rejeitam ou não a hipótese nula do teste z de Wald ($H_0 : \hat{\beta}_j = 0$) [9]. Em outras palavras, o p-valor do teste z de Wald indica a probabilidade de β ser tão ou mais extremo que $|z_{\hat{\beta}_j}|$.

$$z_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{ep(\hat{\beta}_j)} \tag{7}$$

Tabela 4: Estimativas dos parâmetros do modelo

Covariável	Estimativa	Erro padrão	Wald	P-valor
Intercepto	2.0202	0.2877	7.0209	0.0000
x1	0.9518	0.4203	2.2645	0.0235
x2	-0.8450	0.3671	-2.3020	0.0213
x3	0.3707	0.0394	9.4203	0.0000
x5	-0.3570	0.0452	-7.8964	0.0000
x6	-0.2565	0.0492	-5.2152	0.0000
x7	-1.0116	0.2294	-4.4100	0.0000
x10	0.5565	0.1113	4.9997	0.0000
x12	-0.4137	0.0705	-5.8713	0.0000
x13	-0.5755	0.3053	-1.8854	0.0594
x14	-2.9920	0.2846	-10.5136	0.0000
x15	-2.7657	0.2839	-9.7422	0.0000
x17	-4.7352	0.6830	-6.9334	0.0000
x21	0.1617	0.0633	2.5519	0.0107
x23	0.5037	0.0562	8.9657	0.0000
x25	-0.2098	0.0662	-3.1708	0.0015
x30	-0.2306	0.0810	-2.8470	0.0044
x32	-0.4785	0.0324	-14.7557	0.0000
x33	0.4674	0.0461	10.1424	0.0000
x34	-1.6710	0.0867	-19.2738	0.0000

Através da Tabela 4 observa-se que somente a estimativa do parâmetro da covariável X13 não foi significativa ao nível de 5%, entretanto, optou-se por mantê-la no modelo, uma vez que através do teste da razão da verossimilhança foi constatado que a retirada da mesma afeta a qualidade do ajuste.

Com a intenção de generalizar o método de análise dos resíduos da regressão linear para todos os modelos

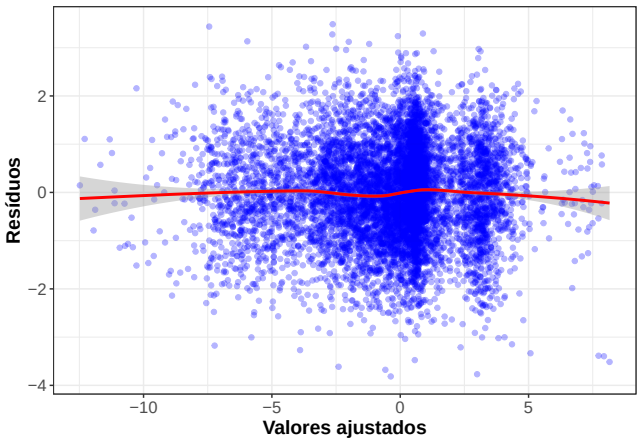

Figura 1: Gráficos dos resíduos versus valores ajustados

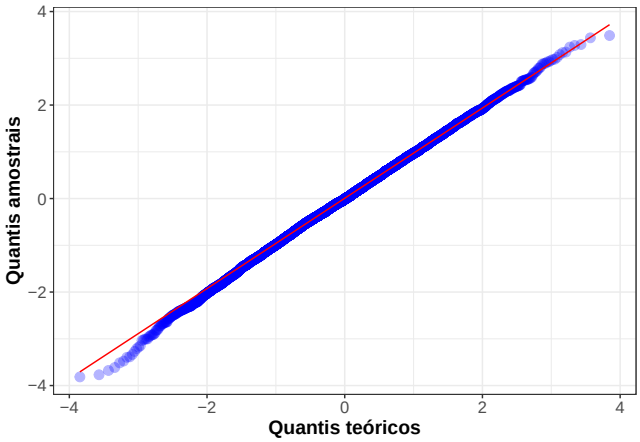

Figura 2: Gráfico quantil-quantil

lineares generalizados, Dunn e Smyth [12] propuseram os resíduos quantílicos aleatorizados, definidos por $r_i = \phi^{-1}(u_i)$, onde ϕ^{-1} é a inversa da função de distribuição acumulada da normal padrão e $u_i = F(y_i; \mu_i, \phi)$, com distribuição uniforme entre 0 e 1, é calculado com base na distribuição acumulada do modelo proposto. Caso o modelo logístico esteja bem ajustado, espera-se que os resíduos quantílicos aleatorizados se apresentem normalmente distribuídos e com variância constante [13]. A análise da qualidade do ajuste, através dos resíduos, foi realizada de forma gráfica. Ao comparar os resíduos com os valores ajustados (Figura 1) é possível observar que estes apresentam variabilidade aproximadamente constante e estão centrados predominantemente em 0, entre -2 e 2. No gráfico quantil-quantil [14] (Figura 2) nota-se que os resíduos estão, de forma razoável, aderentes à distribuição normal. Portanto, pode-se considerar que o modelo está bem ajustado.

2.5.1. Interpretação das estimativas dos parâmetros

asdfasdf

2.5.2. Avaliação do poder preditivo do modelo

asdfasdf

3. Conclusões

asdfasdf

Referências

[1] Jaishankar Ganesh, Mark J. Arnold, and Kristy E. Reynolds. Understanding the customer base of service providers: An examination of the differences between switchers and stayers. *Journal of Marketing*, 64(3):65–87, July 2000.

[2] Thorsten Hennig-Thurau. Customer orientation of service employees. *International Journal of Service Industry Management*, 15(5):460–478, December 2004.

[3] Amy Gallo. The value of keeping the right customers, aug 2014.

[4] Delane Botelho and Frederico Damian Tostes. Modelagem de probabilidade de churn. *Revista de Administração de Empresas*, 50(4):396–410, December 2010.

[5] T. Vafeiadis, K.I. Diamantaras, G. Sarigiannidis, and K.Ch. Chatzisavvas. A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, 55:1–9, June 2015.

[6] Harvendra Kumar and Rakesh Kumar Yadav. Rule-based customer churn prediction model using artificial neural network based and rough set theory. In *Advances in Intelligent Systems and Computing*, pages 97–108. Springer Singapore, 2020.

[7] Antonio Carlos da Silva Júnior, Isaque David Pereira de Almeida, and Marcos dos Santos. Ordenação de algoritmos para modelagem preditiva de churn: analisando o problema a partir dos métodos sapevo-m e vikor. In *Congresso Internacional de Administração*, oct 2020.

[8] Sean Kandel, Jeffrey Heer, Catherine Plaisant, Jessie Kennedy, Frank van Ham, Nathalie Henry Riche, Chris Weaver, Bongshin Lee, Dominique Brodbeck, and Paolo Buono. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4):271–288, September 2011.

[9] Luiz Paulo Fávero and Patrícia Belfiore. *Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®*. Elsevier Brasil, 2017.

[10] Cesar Augusto Taconeli. *Data Science and Big Data - Modelos Lineares*. Universidade Federal do Paraná, aug 2019.

[11] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[12] Peter K. Dunn and Gordon K. Smyth. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244, September 1996.

[13] Rafaela Kienen and Cesar Taconeli. Modelos generalizados aditivos para locação, escala e forma numa análise de custos de procedimentos hospitalares de uma operadora de planos de saúde. 33:330–342, 09 2015.

[14] Martin B. Wilk and Ram Gnanadesikan. Probability plotting methods for the analysis for the analysis of data. *Biometrika*, 55(1):1–17, 1968.