

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 Gallo [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 prever 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 Silva Júnior, Almeida e Santos [7], que utilizaram uma modelagem híbrida multicritério considerando múltiplos decisores para a escolha de um modelo preditivo de churn, o algoritmo

\*juniorssz@gmail.com

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 variável 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 um dos subperíodos. Em seguida, através da Equação (1), foi calculado o desempenho do vendedor em função de diversas métricas, onde  $V1$  e  $V2$  são os valores calculados para cada subperíodo.

$$Desempenho = \frac{V2}{V1 + V2} \quad (1)$$

**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

As métricas desempenho, data à natureza da equação de origem, possuem o comportamento explicado pela Tabela 2. Ao término desta etapa foi obtido um conjunto de dados composto pela variável resposta (churn) e, como covariáveis, 9 métricas de desempenho, onde cada observação representa um vendedor.

#### 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 23 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 32 variáveis e 11.131 observações.

## 2.2. Modelo 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 de covariáveis ( $X_1, \dots, X_n$ ). Sua definição ocorre através da Equação (2), onde  $\beta_j$  ( $j = 0, 1, 2, \dots, p$ ) representa os parâmetros a serem esti-

mados, sendo  $\beta_0$  o intercepto e os demais, parâmetros de cada covariável. E o subscrito  $i$  representa cada observação da amostra ( $i = 1, 2, \dots, n$ ) [9].

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

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  $\pi_i$ , resultando na Equação (3).

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

A estimação dos parâmetros  $\beta_j$  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).

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

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).

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

### 2.3. Ajuste do modelo de regressão logística binária

A comparação entre dois modelos de regressão logística pode ser realizada através do Critério de Informação de Akaike (AIC), definido por  $-2\log L + 2p$ , onde  $\log L$  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 as combinações possíveis pode ser computacionalmente inviável, mesmo para um número moderado de covariáveis. Portanto, para ajudar a encontrar o melhor modelo com o menor número de covariáveis possível foi utilizado o algoritmo *stepwise* [10].

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 menor, com 25% dos dados, foi separada para a etapa de avaliação do poder preditivo do modelo, ao passo que a amostra maior foi utilizada

para o ajuste de dois modelos por validação cruzada K-fold com 5 folds [11], 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. Quando a estatística do TRV é inferior ao valor da distribuição do  $\chi^2$  com  $k$  graus de liberdade e 5% de significância, não rejeitamos a hipótese nula, ou seja, concluímos que a remoção de  $k$  covariáveis não afeta a qualidade do ajuste do modelo [9].

$$TRV = -2(\log L_{\text{completo}} - \log L_{\text{restrito}}) \quad (6)$$

Ao realizar o teste, foi constatado que a remoção das 17 covariáveis através do algoritmo *stepwise* não alterou a qualidade do ajuste, uma vez que a estatística do teste foi inferior ao valor da distribuição do  $\chi^2$  com 17 graus 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 14 covariáveis selecionadas para o modelo.

## 3. Resultados e discussões

### 3.1. Análise e interpretação das estimativas dos parâmetros

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  $\beta_j$  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  $\beta$  ser tão ou mais extremo que  $|z_{\hat{\beta}_j}|$ .

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

**Tabela 3:** Covariáveis utilizadas pelo modelo

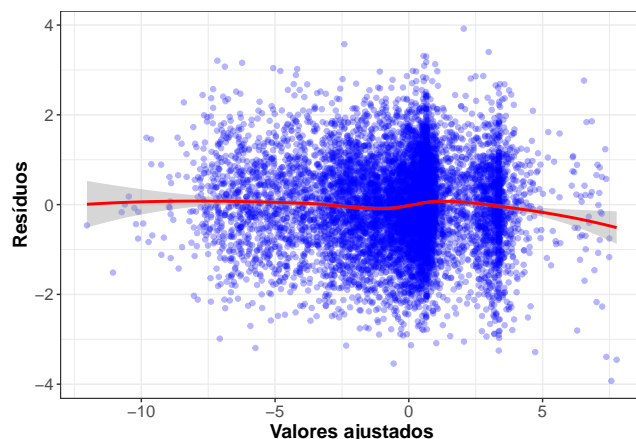
Covariável	Descrição	Medição
<b>Métricas de desempenho</b>		
X2	Desempenho do número de pedidos cancelados ou suspensos no período	percentual
X4	Desempenho da média de dias de atraso (postagens) no período	percentual
X5	Desempenho da média de dias de atraso (entregas) no período	percentual
X6	Desempenho do faturamento no período	percentual
X9	Desempenho do ticket médio no período	percentual
<b>Qualitativas</b>		
X11	Necessita de dias adicionais para postagem	1 = sim; 0 = não
X13	Esta no estágio I	1 = sim; 0 = não
X14	Está no estágio R	1 = sim; 0 = não
X20	Localizado na região Sudeste	1 = sim; 0 = não
<b>Quantitativas</b>		
X22	Valor total do faturamento	contínua
X27	Número total de produtos publicados	discreta
X29	Número de dias desde a contratação até a primeira venda	discreta
X30	Número de dias desde a contratação até estar pronto para operar	discreta
X31	Número de dias em atividade no período	discreta

**Tabela 4:** Estimativas dos parâmetros do modelo

Covariável	Estimativa	Erro padrão	Wald	P-valor
Intercepto	1.5513	0.1316	11.7902	0.0000
X2	0.3693	0.0387	9.5437	0.0000
X4	-0.3558	0.0444	-8.0085	0.0000
X5	-0.2567	0.0487	-5.2749	0.0000
X6	-0.8296	0.1052	-7.8829	0.0000
X9	0.4532	0.1006	4.5028	0.0000
X11	-0.3950	0.0697	-5.6703	0.0000
X13	-2.6981	0.1237	-21.8040	0.0000
X14	-2.2880	0.1246	-18.3608	0.0000
X20	0.1598	0.0628	2.5453	0.0109
X22	-0.1925	0.0629	-3.0595	0.0022
X27	-0.1717	0.0748	-2.2949	0.0217
X29	-0.4516	0.0321	-14.0894	0.0000
X30	0.4201	0.0457	9.1935	0.0000
X31	-1.6460	0.0863	-19.0675	0.0000

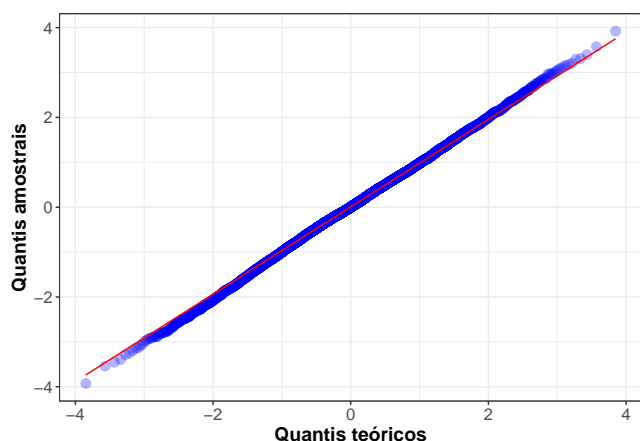
Através da Tabela 4 observa-se que todas as estimativas apresentaram significância ao nível de 5%, não sendo necessária qualquer intervenção adicional no modelo final obtido através do algoritmo *stepwise*.

Com a intenção de generalizar o método de análise dos resíduos da regressão linear para todos os modelos lineares generalizados, Dunn e Smyth [12] propuseram os resíduos quantílicos aleatorizados, definidos por  $r_i = \phi^{-1}(u_i)$ , onde  $\phi^{-1}$  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

**Figura 1:** Gráficos dos resíduos versus valores ajustados

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, mesmo com uma leve fuga nas caldas. Portanto, pode-se considerar que o modelo foi bem ajustado.

Através da Equação (8), obtida a partir da Equação (2), é possível modelar a chance de ocorrência do evento de interesse para uma particular observação e, em con-



**Figura 2:** Gráfico quantil-quantil

sequência, avaliar o quanto a chance de ocorrência do evento de interesse se altera em média, em função de uma particular estimativa. Adicionalmente podemos dizer que o aumento de  $k$  unidades em uma particular covariável, mantidas as demais condições constantes, multiplica a chance de ocorrência do evento de interesse por  $e^{k \cdot \hat{\beta}_j}$ , onde  $\hat{\beta}_j$  representa a estimativa do parâmetro desta particular covariável [9].

$$\frac{\pi_i}{1 - \pi_i} = e^{\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}} \quad (8)$$

As Tabelas 5, 6 e 7 representam as estimativas dos parâmetros do modelo e estão ordenadas decrescentemente em função da variação absoluta da chance de ocorrência de churn. Suas interpretações podem ser realizadas da seguinte forma, começando pela Tabela 5 que representa as covariáveis quantitativas: a chance de churn fica multiplicada por  $e^{-1,6460} = 0,1928$  para 1 unidade a mais na covariável X31, mantidas as demais condições constantes. Em outras palavras, 1 unidade a mais na covariável X31 impacta na redução da chance de churn em 81%, fixadas as demais covariáveis. Como as covariáveis quantitativas foram padronizadas com média 0 e desvio padrão 1, é importante ressaltar que 1 unidade na covariável X31 não representa 1 dia de atividade, assim sendo, neste estudo a melhor forma de interpretar as estimativas dos parâmetros das covariáveis quantitativas é assimilando que estimativas menores que zero reduzem em média a chance de churn, à medida que o valor de suas respectivas covariáveis elevam, ao passo que a estimativa maior que zero aumenta em média a chance de churn, à medida que o valor de sua respectiva covariável também aumenta, fixadas as demais covariáveis. Quanto às covariáveis qualitativas representadas na Tabela 6, podemos interpretá-las

**Tabela 5:** Chances de ocorrência de churn para covariáveis quantitativas

Covariável	Estimativa	Chance	Variação (%)
X31	-1.6460	0.1928	-81
X30	0.4201	1.5221	52
X29	-0.4516	0.6366	-36
X22	-0.1925	0.8249	-18
X27	-0.1717	0.8422	-16

**Tabela 6:** Chances de ocorrência de churn para covariáveis qualitativas

Covariável	Estimativa	Chance	Variação (%)
X13	-2.6981	0.0673	-93
X14	-2.2880	0.1015	-90
X11	-0.3950	0.6737	-33
X20	0.1598	1.1733	17

**Tabela 7:** Chances de ocorrência de churn para covariáveis de desempenho

Covariável	Estimativa	Chance (0,5)	Variação (%)
X6	-0.8296	0.6605	-34
X9	0.4532	1.2543	25
X2	0.3693	1.2028	20
X4	-0.3558	0.8370	-16
X5	-0.2567	0.8795	-12

conforme o exemplo: a chance de churn para os vendedores que estão no estágio I (X13) é 93% menor com relação aos vendedores que não estão no mesmo estágio, mantidas as demais condições constantes. Por fim, com relação às covariáveis de desempenho (Tabela 7), dadas as suas características representadas na Tabela 2, suas interpretações podem ser realizadas de acordo com o exemplo: a chance de churn de um vendedor que melhorou o seu desempenho nos subperíodos avaliados com relação ao faturamento ( $X6 > 0,5$ ), é em média mais de 34% menor, fixadas as demais covariáveis. Em contrapartida, a chance de churn de um vendedor que piorou o seu desempenho nos subperíodos avaliados com relação ao faturamento ( $X6 < 0,5$ ), é também menor em média, entretanto, em um valor percentual inferior a 34.

### 3.2. Avaliação do poder preditivo do modelo

Para possibilitar a avaliação do poder preditivo do modelo na amostra de validação, é necessário antes definir o valor de *cutoff*, ou seja, um ponto de corte de

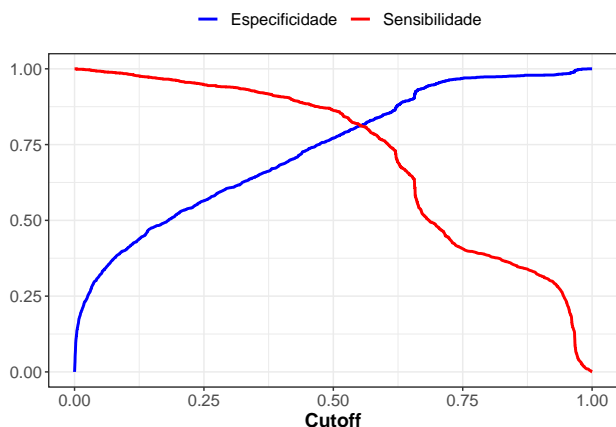


Figura 3: Curva de Sensibilidade

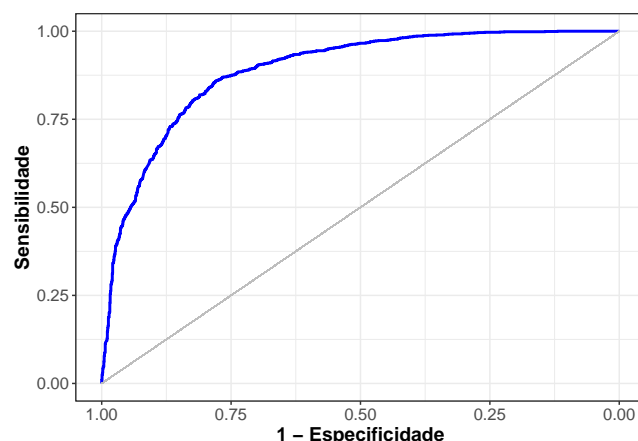


Figura 4: Curva ROC

modo que as observações com probabilidade de ocorrência de churn superior ao *cutoff* sejam classificadas como churn ( $Y = 1$ ) e, caso contrário, classificadas como não churn ( $Y = 0$ ). A escolha do *cutoff* foi realizada através da análise da curva de Sensibilidade, que é um gráfico que apresenta os valores da sensibilidade e da especificidade em função dos valores de *cutoff*, e da curva ROC (*Receiver Operating Characteristic*), gráfico que apresenta a variação da sensibilidade em função de  $(1 - \text{especificidade})$ , que mostra o comportamento do *trade off* entre a sensibilidade e a especificidade em função da alteração do *cutoff* [9]. Através do cálculo da área sob a curva ROC (*AUC - Area Under the Curve*) é possível avaliar a eficiência global do modelo, sendo  $AUC = 1$  o melhor valor possível. O modelo em estudo apresentou  $AUC = 0,8894$ , o que indica que o modelo possui uma boa eficiência global. Analisadas as curvas de Sensibilidade (Figura 3) e ROC (Figura 4) e considerando os requisitos do negócio, optou-se por um valor de *cutoff* que garantisse um pouco mais de sensibilidade com relação à especificidade. Portanto, para continuidade do estudo foram consideradas como churn ( $Y=1$ ) as observações com probabilidade de ocorrência de churn superior a 0,54.

Definido o valor de *cutoff*, através do cruzamento dos valores preditos pelo modelo e os valores observados, foi construída a matriz de confusão (Tabela 8), que apresenta em sua diagonal principal o número de classificações corretas e, na diagonal secundária, o número de classificações incorretas. A partir da matriz de confusão foram calculadas as seguintes métricas de avaliação: sensibilidade (taxa de classificação correta entre as observações com a ocorrência de churn), especificidade (taxa de classificação correta entre as observações sem a ocorrência de churn) e acurácia

Tabela 8: Matriz de confusão

Predito	Observado	
	0	1
0	Verdadeiro negativo (VN)	Falso negativo (FN)
1	Falso positivo (FP)	Verdadeiro positivo (VP)

(taxa global de classificações corretas), representadas pelas equações (9), (10) e (11), respectivamente.

$$\text{Sensibilidade} = \frac{VP}{VP + FN} \quad (9)$$

$$\text{Especificidade} = \frac{VN}{VN + FP} \quad (10)$$

$$\text{Acurácia} = \frac{VN + VP}{VN + VP + FN + FP} \quad (11)$$

Com a sensibilidade de 0,8217, especificidade de 0,8029 e acurácia de 0,8118 (com intervalo de confiança de 0,7968 a 0,8261), o poder preditivo do modelo foi considerado adequado para o negócio.

#### 4. Conclusões

asdfsdf

#### 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.