**Modelagem estatística aplicada na prevenção ao *churn***

Lucas Franz Monteiro¹\*;Profa. Dra. Ana Julia Righetto2

1 Risco e Modelagem. Rua Francisco Freire – Jardim Carlos Cooper; 08664-300 Suzano, São Paulo, Brasil

2 Nome da Empresa ou Instituição (opcional). Titulação ou função ou departamento. Endereço completo (pessoal ou profissional) – Bairro; 00000-000 Cidade, Estado, País

**Modelagem estatística aplicada na prevenção ao *churn***

**Introdução**

O avanço tecnológico impulsionado pela globalização, e as políticas públicas de inclusão digital, têm como uma das principais favorecidas a disseminação da informação, viabilizada pragmaticamente, pelas companhias de telecomunicação.

De acordo com a Pesquisa Nacional por Amostra de Domicílios Contínua (2019), 82.7% dos domicílios brasileiros, em 2019, tinham acesso à internet, sendo a região sudeste a primeira do ranking das regiões com maior índice, na qual 84.9% da população dispunha de acesso à internet. No mesmo ano, 81% da população com idade mínima de dez anos, possuía *smartphone* de uso pessoal, sendo que 91% deste público, contava com acesso à internet através do dispositivo.

Conforme dados mais recentes da Pesquisa Anual de Serviços, desenvolvida pelo IBGE, no Brasil, em 2019, a quantidade de empresas do setor de telecomunicações foi alavancada em 13%, com relação ao ano anterior, contabilizando 11.043 companhias. Essa crescente traz consigo o aumento da concorrência entre as empresas do setor, na busca por prestar melhores serviços ao público, e da consequente ampliação do *market share*.

Tendo em vista esse cenário, segundo Ferreira (2012), é fundamental que as companhias fidelizem seus clientes, a fim de preservar a competitividade, e de acordo com Jahromi; Stakhovych e Ewing (2014), as estratégias de retenção geram retornos sobre o investimento superiores às de captação, raciocínio corroborado por Dhurup (2012), segundo o qual a obtenção de novos clientes requer maiores investimentos quando comparada à manutenção dos clientes atuais.

De acordo com o Cambridge Dictionary (2022), *churn* é o fenômeno no qual clientes param de comprar o produto ou serviço de uma companhia, principalmente para comprá-los de um competidor. Segundo Mattison, R. (2005), o *churn* é voluntário quando o cliente rescinde o contrato de serviço, de forma deliberada ou não; e involuntário quando é a empresa quem opta pela rescisão, geralmente por motivos de fraude, não pagamento ou não utilização do serviço.

Esse trabalho tem como objetivo aplicar técnicas de modelagem estatística na prevenção ao *churn*, a fim de permitir a detecção em tempo hábil, de clientes com elevado risco de cancelamento do serviço, auxiliando a companhia na tomada de decisões proativas para retenção de seus clientes, além de contribuir na elucidação das principais razões que levam ao cancelamento do serviço.

**Material e Métodos**

O *dataset* utilizado no trabalho é composto por clientes de uma companhia de telecomunicações fictícia, que fornece serviços de telefonia e internet, no estado da California. Composto originalmente por 7.043 observações e 53 variáveis, o conjunto de dados oficialmente disponibilizado na plataforma IBM Cognos Analytics, indica que no terceiro trimestre de 2019, o índice de *churn* foi de 26.5%. O restante da base de dados é formada por novos clientes, que contrataram os serviços recentemente, e por consumidores que já assinavam algum serviço, e mantiveram o contrato ativo.

A fim da enriquecer os dados, variáveis censitárias da população norte-americana, disponibilizadas pela pesquisa anual *American Community Survey*, que incorpora características sociais, econômicas, demográficas e habitacionais da nação (ACS, 2017), foram coletadas e anexadas ao *dataset*. Neste trabalho, optou-se por consultar as estimativas de cinco anos, que compreendem o período de 2013 a 2017, devido maior confiabilidade estatística para áreas geográficas menos populosas.

Neste trabalho, foi aplicada a metodologia: qualitativa, quantitativa, exploratória, descritiva, estudo de caso, etc.

Tabela 1: Variáveis do dataset original, selecionadas para utilização

|  |  |
| --- | --- |
| **Variável** | **Descrição** |
| customer\_id | Identificador único do cliente |
| gender | Sexo do cliente |
| age | Idade do cliente |
| flg\_married | Indica se o cliente é casado |
| number\_of\_dependents | Quantidade de dependentes que moram com o cliente |
| city | Cidade da residência principal do cliente |
| zip\_code | Cep da residência principal do cliente |
| latitude | Latitude da residência principal do cliente |
| longitude | Longitude da residência principal do cliente |
| number\_of\_referrals | Quantidade de indicações, feitas pelo cliente fez até o presente |
| tenure\_in\_months | Tempo de casa do cliente, ao final do trimestre |
| offer | Última oferta de marketing aceita pelo cliente, se aplicável |
| flg\_phone\_service | Indica se o cliente assina o serviço de telefonia residencial da companhia |
| avg\_monthly\_long\_distance\_charges | Valor mensal médio das cobranças de chamadas de longas distâncias, calculado até o final do trimestre |
| flg\_multiple\_lines | Indica se o cliente assina múltiplas linhas telefônicas da companhia |
| internet\_type | Tipo do serviço de internet assinado pelo cliente |
| avg\_monthly\_gb\_download | Volume mensal médio de download, em gigabytes, calculado até o final do trimestre |
| flg\_online\_security | Indica se o cliente assina um serviço adicional de segurança online, fornecido pela companhia |
| flg\_online\_backup | Indica se o cliente assina um serviço adicional de backup online, fornecido pela companhia |
| flg\_device\_protection\_plan | Indica se o cliente assina a um plano adicional de proteção do dispositivo, para seu equipamento de internet, fornecido pela companhia |
| flg\_premium\_tech\_support | Indica se o cliente assina um plano adicional de suporte técnico da companhia, com tempos reduzidos de espera |
| flg\_streaming\_tv | Indica se o cliente utiliza a internet para assistir programas de televisão de um fornecedor externo |
| flg\_streaming\_movies | Indica se o cliente utiliza a internet para assistir filmes de um fornecedor externo |
| flg\_streaming\_music | Indica se o cliente utiliza a internet para escutar música de um fornecedor externo |
| flg\_unlimited\_data | Indica se o cliente pagou uma taxa mensal adicional, para ter downloads/uploads ilimitados |
| contract | Tipo de contrato atual do cliente |
| flg\_paperless\_billing | Indica se o cliente optou por cobrança sem papel |
| payment\_method | Método de pagamento, utilizado pelo cliente, para pagar a fatura |
| monthly\_charge | Valor total da mensalidade atual do cliente, cobrada por todos os serviços utilizados |
| total\_charges | Cobranças totais do cliente, exceto valores adicionais, cobrados por utilização superior ao especificado no plano do cliente, calculadas até o final do trimestre |
| total\_refunds | Reembolsos totais do cliente, calculados até o final do trimestre |
| total\_extra\_data\_charges | Cobranças totais do cliente, por downloads de dados extras, acima do especificado em seu plano, ao final do trimestre |
| total\_long\_distance\_charges | Cobranças totais do cliente, por chamadas de longa distância, acima das especificadas em seu plano, ao final do trimestre |
| satisfaction\_score | Índice da satisfação geral do cliente com a companhia |
| customer\_status | Status do cliente ao final do trimestre |
| flg\_churn | Indica se a firma perdeu o cliente |
| cltv | Valor do tempo de vida do cliente (Customer Lifetime Value). Quanto maior o valor, mais valioso o cliente |
| churn\_category | Categoria de alto-nível, para o motivo da perda do cliente. Todos os clientes, ao deixarem a companhia, são questionados sobre o motivo da saída |
| churn\_reason | Motivo específico da perda do cliente |

Fonte: resultados originais da pesquisa

Tabela 2: Variáveis construídas através do processo de feature engineering

|  |  |
| --- | --- |
| **Variável** | **Descrição** |
| valor\_cobranca\_geral | Cobranças gerais do cliente, incluindo valores adicionais por utilização superior ao especificado em seu plano, ao final do trimestre |
| tx\_valores\_reembolsados | Percentual de valores reembolsados, em relação às cobranças gerais |
| tx\_concentracao\_cobranca\_mes\_q3 | Quanto dos valores cobrados até o final do trimestre, estão concentrados na mensalidade atual do cliente |
| valor\_cobrancas\_extras | Valores totais, cobrados por chamadas de longa distância e downloads de dados extras, acima do especificado no plano do cliente, ao final do trimestre |
| tx\_contrib\_cobrancas\_extras\_cobranca\_geral | Representatividade dos valores cobrados de forma adicional, em relação aos valores gerais, cobrados do cliente |
| qtd\_servicos\_principais | Quantidade de serviços principais assinados pelo cliente. Os serviços principais são telefonia e internet |
| qtd\_servicos\_adicionais | Quantidade de serviços adicionais assinados pelo cliente |
| qtd\_streamings | Quantidade de streamings utilizados pelo cliente |

Fonte: resultados originais da pesquisa

Tabela 3: Variáveis censitárias, referentes ao condado da residência principal do cliente

|  |  |
| --- | --- |
| **Variável** | **Descrição** |
| county | Condado da residência principal do cliente |
| condado\_idade\_mediana\_habitantes | Idade mediana dos habitantes |
| condado\_indice\_gini\_desigualdade\_renda | Índice de gini, de desigualdade de renda |
| condado\_qtd\_habitantes | Quantidade de habitantes |
| condado\_renda\_familiar\_mediana | Renda familiar mediana |
| condado\_tx\_habitantes\_homens | Percentual da população composta por homens |
| condado\_tx\_habitantes\_menor\_18\_anos | Percentual da população composta por indivíduos menores de 18 anos |
| condado\_area\_terra\_m2 | Área territorial do condado, em metros quadrados |
| condado\_densidade\_populacional | Densidade populacional (número de habitantes / área) |

Fonte: resultados originais da pesquisa

Tabela 4: Variáveis censitárias, referentes ao cep da residência principal do cliente

|  |  |
| --- | --- |
| **Variável** | **Descrição** |
| zip\_code\_idade\_mediana\_habitantes | Idade mediana dos habitantes |
| zip\_code\_indice\_gini\_desigualdade\_renda | Índice de gini, de desigualdade de renda |
| zip\_code\_qtd\_habitantes | Quantidade de habitantes |
| zip\_code\_renda\_familiar\_mediana | Renda familiar mediana |
| zip\_code\_tx\_habitantes\_homens | Percentual da população composta por homens |
| zip\_code\_tx\_habitantes\_menor\_18\_anos | Percentual da população composta por indivíduos menores de 18 anos |
| zip\_code\_area\_terra\_m2 | Área territorial do cep, em metros quadrados |
| zip\_code\_densidade\_populacional | Densidade populacional (número de habitantes / área) |

Fonte: resultados originais da pesquisa

O conjunto de dados foi dividido em partições de treinamento e teste, e ao todo foram estimados dois modelos, sendo um modelo de regressão logística binária clássica, e um modelo de regressão logística binária multinível, os quais, com base no comportamento conjunto das variáveis preditoras, calcularam a probabilidade de *churn* de cada cliente. Os parâmetros do modelo clássico foram estimados por máxima verossimilhança, e os do modelo multinível, por máxima verossimilhança restrita, método que de acordo com Fávero e Belfiore (2017) gera estimações não viesadas da variância dos termos de erro.

A modelagem multinível foi aplicada com o intuito de permitir a identificação de heterogeneidades entre os clientes bem como entre os condados, possibilitando a especificação de componentes aleatórios em cada nível.

Nessa perspectiva, além dos parâmetros do modelo, também foram estimados os componentes de variância dos termos de erro do intercepto e do coeficiente angular, os quais tiveram suas significâncias estatísticas analisadas por meio de testes de razão de verossimilhança, de modo a identificar se a presença de níveis superiores gera aleatoreidades nos interceptos e nos declives, o que foi utilizado como indicador para definir se o modelo deveria ser estimado com interceptos aleatórios, com declives aleatórios, com ambos, ou se um modelo de regressão logística clássica era suficiente, caso identificada a ausência de aleatoreidades de interceptos e declive (Fávero e Belfiore, 2017).

Os modelos foram testados no conjunto de teste, a fim de avaliar a capacidade de generalização em dados não utilizados no treinamento, e tiveram suas performances comparadas por meio de testes da razão de verossimilhança. O desempenho de cada modelo também foi avaliado por meio de matrizes de confusão, e pela área sob a curva ROC (*Receiver Operating Characteristic*).

A curva ROC, segundo James et al. (2021), é traçada por um gráfico que apresenta para todos os pontos de corte, a interação entre os verdadeiros positivos (sensitividade), e os falsos positivos (1 – especificidade) do modelo, plotados respectivamente no eixo das ordenadas e abscissas. A área sob a curva ROC, cujo valor máximo é um, foi utilizada para comparar a performance preditiva dos modelos, uma vez que quanto maior a área, maior a capacidade preditiva.

Na modelagem multinível, foram adotados dois níveis, a saber, cliente (nível 1), e condado (nível 2). Este tipo de modelo, de acordo com Fávero e Belfiore (2017), é denominado HLM2, onde “HLM” é acrônimo de *Hierarchical Linear Model*, e o numeral “2” indica que o modelo será aplicado a dados aninhados em dois níveis.

A linguagem de programação R v. 4.1.1 (R Core Team, 2021), foi utilizada para desenvolver o trabalho, com o auxílio dos pacotes:

* tidyverse (Wickham et al., 2019) – Manipulação e transformação de dados.
* caret (Kuhn, 2021) – Criação de amostras aleatórias estratificadas para treinamento e teste; construção de matrizes de confusão.
* stats (R Core Team, 2021) – Treinar modelo de regressão logística binária; criar grupos com k-means.
* glmmTMB (Brooks et al., 2017) – Modelagem multinível.
* pROC (Robin et al., 2011) – Calcular a área sob a curva ROC.
* lmtest (Zeileis e Hothorn, 2002) – Comparar os log-likelihoods dos modelos, através de teste da razão de verossimilhança.
* sf (Pebesma, 2018)
* tigris (Walker, 2022)
* stringr (Wickham, 2019)
* glue (Hester e Bryan, 2022)
* readxl (Wickham e Bryan, 2019)
* ggpubr (Kassambara, 2020)
* ggrepel (Slowikowski, 2021)
* tidycensus (Walker e Herman, 2022)
* zipcodeR (Rozzi, 2021)
* buildmer (Voeten, 2022)

**Resultados Preliminares**

Na Figura 1, verifica-se que os motivos mais comuns que acarretaram a perda de clientes, estão relacionados à alguma empresa concorrente oferecendo melhores dispositivos e planos, bem como aspectos relacionados ao comportamento ou postura do profissional do suporte técnico. Curiosamente, apenas 11.3% dos clientes cancelaram o serviço por razões vinculadas ao preço praticado pela companhia, e eventuais cobranças por utilização extra de serviços.

Chart, bar chart

Description automatically generated

Figura 1: Distribuição do motivo de churn

Fonte: Resultados originais da pesquisa

A satisfação dos clientes, que para Caldeira (2006) resulta da discrepância entre a percepção emocional do serviço, e a expectativa gerada pelas propagandas de marketing, tem, de acordo com Anderson e Sullivan (1993), e Fornell (1992), uma relação diretamente proporcional com a retenção, de modo que quanto maior a satisfação, maior a retenção. Esse fenômeno é salientado na distribuição da variável “satisfaction\_score”, conforme apresentado na Tabela 5 e na Figura 2.

Tabela 5: Percentual de churn, por índice de satisfação do cliente

|  |  |  |
| --- | --- | --- |
| **satisfaction\_score** | **Qtd clientes** | **% Churn** |
| 1 | 922 | 100 |
| 2 | 518 | 100 |
| 3 | 2.665 | 16,1 |
| 4 | 1.789 | 0 |
| 5 | 1.149 | 0 |

Fonte: Resultados originais da pesquisa

Dos clientes perdidos, 65% não apresentavam uma das principais características de fidelidade, a saber, a indicação do produto ou serviço para outras pessoas, e 90% tinham feito no máximo uma indicação.

Chart, box and whisker chart

Description automatically generated

Figura 2: Distribuição de variáveis de interesse

Fonte: Resultados originais da pesquisa

Conforme ilustrado na Figura 2, os condados com maior presença de clientes estão localizados ao sul do estado da California, a saber Los Angeles, o qual concentra 18.9% de todos os clientes, seguido por San Diego e Orange, os quais, juntos, não alcançam a mesma relevância de Los Angeles, congregando 13.8% dos clientes.

A picture containing chart

Description automatically generated

Figura 3: Distribuição de clientes, por condado, com destaque para os condados com maiores e menores representatividades

Fonte: Resultados originais da pesquisa

Salienta-se que os condados menos populosos do estado, são os mais relevantes no aspecto do percentual da população que é ou foi cliente da companhia, como apresentado na Figura 3. Porém, juntos, os três condados contabilizam apenas 84 clientes, ou 1.2% do total de 7.043 consumidores. Sierra, Alpine e Trinity estão entre os cinco condados com menor número de habitantes da Califórnia, e dos aproximadamente 2.885 habitantes de Sierra, 0.97% são ou já foram clientes da companhia em algum momento.

Chart

Description automatically generated

Figura 4: Proporção de habitantes, por condado, que foram ou são clientes da companhia, com destaque para os condados com maiores e menores representatividades

Fonte: Resultados originais da pesquisa

Del Norte, um dos condados com menor participação na carteira de clientes da companhia, foi o condado com a taxa mais elevada de cancelamento dos serviços, enquanto San Diego, o segundo condado com a maior quantidade de clientes, foi também o segundo com maior índice de *churn*, seguido pelo condado de Stanislaus, como ilustrado na Figura 4.

A picture containing diagram

Description automatically generated

Figura 5: Índice de churn por condado

Fonte: Resultados originais da pesquisa

Na Figura 5, evidenciou-se que mais de 30% dos clientes que cancelaram o serviço, tem a residência principal localizada nos condados de Los Angeles e San Diego, situados ao sul do estado da California.

Diagram

Description automatically generated

Figura 6: Distribuição do churn, por condado

Fonte: Resultados originais da pesquisa

Detectou-se também a existência de um cinturão geográfico, localizado majoritariamente ao norte, em que a razão mais recorrente de *churn* está relacionada à insatisfação com o serviço prestado pela companhia, conforme apresentado na Figura 6.

Chart

Description automatically generated

Figura 7: Motivos mais frequentes de churn, por condado

Fonte: Resultados originais da pesquisa

**Referências**

Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10 (1), 439-446, <https://doi.org/10.32614/RJ-2018-00>

Kyle Walker (2022). tigris: Load Census TIGER/Line Shapefiles. R package version 1.6. <https://CRAN.R-project.org/package=tigris>

Hadley Wickham (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. <https://CRAN.R-project.org/package=stringr>

Jim Hester and Jennifer Bryan (2022). glue: Interpreted String Literals. R package version 1.6.2. <https://CRAN.R-project.org/package=glue>

Hadley Wickham and Jennifer Bryan (2019). readxl: Read Excel Files. R package version 1.3.1. <https://CRAN.R-project.org/package=readxl>

Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. <https://CRAN.R-project.org/package=ggpubr>

Kamil Slowikowski (2021). ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. R package version 0.9.1. <https://CRAN.R-project.org/package=ggrepel>

Kyle Walker and Matt Herman (2022). tidycensus: Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames. R package version 1.2. <https://CRAN.R-project.org/package=tidycensus>

G.C. Rozzi, zipcodeR: Advancing the analysis of spatial data at the ZIP code level in R, Softw. Impacts. (2021) 100099.

Cesko C. Voeten (2022). buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression. R package version 2.4. <https://CRAN.R-project.org/package=buildmer>

<https://www.census.gov/data/developers/data-sets/acs-5year.2017.html>

Mattison, R. 2005. The Telco Churn Management Handbook

Tamaddoni Jahromi, A., Stakhovych, S., & Ewing, M. (2014). Managing B2B customer churn, retention and profitability. Industrial Marketing Management, 43(7), 1258–1268. doi:10.1016/j.indmarman.2014.06.016

Anderson, E. W. & Sullivan, M. W. (1993). The Antecedents and Consequences of Customer Satisfaction. Marketing Science, 12 (Spring), 125-143

Fornell, C. (1992). A national customer satisfaction barometer: The Swedish experience. Journal of Marketing, 56 (1), 6–21.

Caldeira, S. (2006). Retenção de Clientes. In A. Correia, A. Sacavém, C. Colaço (Eds.), Manual de Fitness & Marketing (pp. 165-184). Lisboa. Visão e Contextos.

Um estudo sobre fidelização e retenção de clientes na área do fitness. Célia Marina Costa Ferreira. 2012

Dhurup, J. (2012). Establishing and maintaining customer relationships in commercial health and fitness centers in South Africa. *International Journal of Trade, Economics and Finance*, 3 (1).

**Apêndice ou Anexo**

Chart, line chart

Description automatically generated

Figura 8: Churn trimestral, por operadora de celular, no Brasil

Fonte: Teleco 2022.

Chart, bar chart

Description automatically generated

Figura 9: Market share, por operadora de celular, no Brasil

Fonte: Teleco, 2022.