**Modelagem estatística aplicada à retenção de clientes**

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

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

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

\*autor correspondente: nome@email.com

**Modelagem estatística aplicada à retenção de clientes**

**Resumo**

Tópico obrigatório para o depósito do TCC, porém opcional para a etapa dos Resultados preliminares.

**Palavras-chave:** (inserir até cinco palavras diferentes das contidas no título, separadas por ponto-e-vírgula).

**Introdução**

O título da seção “Introdução” deve ser grafado em negrito, alinhado à esquerda e com a primeira letra maiúscula. Nesta seção o autor deve contextualizar o assunto, estabelecendo uma relação com outros trabalhos publicados sobre o tema e apresentar a justificativa para a realização do trabalho, situando a importância do problema estudado e solucionado, apoiado em citações bibliográficas de fontes confiáveis. O último parágrafo da introdução deve discriminar o objetivo do trabalho, que deve ser enunciado de forma coerente e direta.

A **Introdução** deve ser redigida em no **máximo duas páginas** e não deve conter subtópicos, figuras ou tabelas. Ressalta-se que o TCC não deverá exceder o total de 30 páginas, incluindo o(s) Apêndice(s) e/ou Anexo(s), sendo este o máximo de páginas permitido pelas “Normas de Elaboração do Trabalho de Conclusão de Curso” do MBA USP ESALQ.

**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. Os dados, compostos originalmente por 7.043 observações e 53 variáveis, são oficialmente disponibilizados na plataforma IBM Cognos Analytics. No terceiro trimestre de 2019, de todos os clientes, 26.5% cancelaram os serviços da companhia; 6.4% contrataram o serviço recentemente; e os 67% restantes, já assinavam os serviços, e mantiveram o contrato ativo.

As variáveis censitárias, foram obtidas do American Community Survey, e são referentes ao período de 5 anos, de 2013 a 2017. A ACS é uma pesquisa anual que ocorre nos Estados Unidos da América, e abrange características sociais, econômicas, demográficas e habitacionais da população norte-americana (ACS, 2017). Neste trabalho, optou-se por consultar as estimativas de períodos de cinco anos, devido maior confiabilidade estatística para áreas geográficas menos populosas.

|  |  |
| --- | --- |
| **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 |

|  |  |
| --- | --- |
| **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 |

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) |

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) |

Os motivos mais frequentes que acarretaram a perda de clientes, estão relacionados à alguma empresa concorrente fornecendo melhores dispositivos e planos; e algum comportamento ou postura do profissional do suporte técnico.

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**

A picture containing map

Description automatically generated

A picture containing chart

Description automatically generated

Chart, bar chart

Description automatically generated

A picture containing chart

Description automatically generated

Chart

Description automatically generated

O título da seção Resultados Preliminares deve ser alinhado à esquerda, grafado em negrito com as primeiras letras das palavras em letras maiúsculas. É permitido que a seção seja dividida em subtópicos com formatação de acordo com a descrição no item 1.1 Formato e margens, apresentados na mesma ordem da seção Material e Métodos. Nesta seção devem ser apresentados os resultados parciais da pesquisa, ou seja, os resultados obtidos até o momento.

**Conclusão(ões) ou Considerações Finais**

Tópico obrigatório para o depósito do TCC, porém opcional para a etapa dos Resultados preliminares.

**Agradecimentos** (opcional, 1 parágrafo sucinto)

O título da seção Agradecimentos deve ser alinhado à esquerda e grafado em negrito, primeira letra da palavra grafada em letra maiúscula. Trata-se de seção opcional, de no máximo três linhas, na qual o autor agradece aqueles que contribuíram de maneira relevante para o desenvolvimento do trabalho e elaboração do TCC, mas que não tiveram o envolvimento intelectual necessário à atribuição de coautoria do mesmo, abstendo-se totalmente da menção ou citação de nomes de empresas, instituições ou pessoas que permitiram ou contribuíram com o desenvolvimento do trabalho, a menos que esteja documentalmente autorizado a fazê-lo.

**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

**Apêndice ou Anexo** (opcional)

Apêndices são textos e/ou documentos que foram elaborados pelo autor e que são importantes para complementar a argumentação do trabalho. Anexos são textos ou documentos que ilustram, mas que não foram elaborados pelos autores. Apêndices deverão seguir as mesmas normas de formatação do restante do texto, inclusive para figuras e tabelas.

O TCC deverá conter no máximo 30 páginas, incluindo o(s) Apêndice(s) e/ou Anexo(s).