# Modelagem estatística aplicada na prevenção ao *churn*

Lucas Franz Monteiro[1]; Ana Julia Righetto[2]

[1] Risco e Modelagem. Rua Francisco Freire – Jardim Carlos Cooper; 08664-300 Suzano, São Paulo, Brasil
[2] Head in Statistics and Customer Experience, ALVAZ Agritech, Av. Ayrton Senna da Silva, 600, Londrina, Paraná, Brasil

# Modelagem estatística aplicada na prevenção ao *churn*

## Resumo

O advento do *big data*, caracterizado entre outas coisas pela alta disponibilidade de dados, trouxe o desafio de analisá-los em busca de padrões que auxiliem a tomada de decisão de indivíduos e corporações. Este trabalho tem como objetivo apresentar uma análise de dados voltada à prevenção ao *churn*, aplicando técnicas de aprendizado supervisionado, com destaque para o modelo de regressão logística binária, treinado de modo a classificar clientes com elevado e baixo risco de rescindirem o contrato de prestação de serviço. O melhor modelo de classificação obtido nos experimentos, seria capaz de indicar corretamente à companhia, 90% dos clientes perdidos no período estudado, oferecendo à esta, a possibilidade de desenvolver estratégias voltadas a este público, a fim de manter o vínculo entre empresa e cliente, ajudando desta forma a minimizar as perdas de receita decorrentes dos contratos encerrados, bem como na redução de custos com captação de novos consumidores.
**Palavras-chave:** *Churn*; análise supervisionada; regressão logística.

## Introdução

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

De acordo com o Instituto Brasileiro de Geografia e Estatística [IBGE] (2019b), 82,7% dos domicílios brasileiros, em 2019, tinham acesso à internet, tendo a região sudeste o maior índice, na qual 84,9% da população dispunha de acesso à internet. Ademais, 81% dos indivíduos com idade mínima de dez anos, possuía *smartphone* de uso pessoal, dos quais 91% acessavam a internet através do dispositivo.

No mesmo ano, a quantidade de empresas do setor de telecomunicações foi alavancada em 13%, com relação ao ano pregresso, contabilizando 11.043 companhias (IBGE, 2019a). Esse crescimento naturalmente favorece a competitividade no setor, e nesse cenário, para Ferreira (2012) é fundamental que as companhias fidelizem seus clientes, a fim de se manterem competitivas, e minimizarem o *churn*, evento que, de acordo com Glady e colaboradores (2009), é caracterizado pela perda de clientes para a concorrência.

Uma vez que as estratégias de retenção de clientes geram retornos sobre o investimento superiores às de captação de novos consumidores (Jahromi et al., 2014), o presente trabalho tem como objetivo aplicar técnicas de modelagem estatística na prevenção ao *churn*, auxiliando a companhia na tomada de decisões proativas para retenção de seus clientes, além de contribuir na elucidação das características que mais impactam a probabilidade de ocorrência do fenômeno.

## Material e Métodos

O *dataset* utilizado no trabalho é composto por 7.043 clientes de uma companhia de telecomunicações fictícia, que presta serviços de telefonia e internet, no estado da Califórnia. Contabilizando originalmente 53 variáveis, o conjunto de dados oficialmente disponibilizado na plataforma IBM Cognos Analytics, indica que no terceiro trimestre de 2019, o índice de *churn* da companhia foi igual a 26,5%. O restante da base de dados é constituído por novos clientes, que contrataram os serviços recentemente, e por consumidores que já assinavam algum serviço, e mantiveram o contrato ativo.

As variáveis originais foram renomeadas, com o objetivo de facilitar a manipulação dos dados, de modo que os espaços em branco foram substituídos pelo caractere "_"; as letras maiúsculas deram lugar a letras minúsculas; e as variáveis qualitativas binárias tiveram o prefixo "flg_" acrescido aos seus nomes. Seguindo estes critérios, as variáveis inicialmente identificadas por *"Number of Referrals"* e *"Unlimited Data"*, foram renomeadas para *"number_of_referrals"* e *"flg_unlimited_data"*, respectivamente. Deliberou-se também por adaptar o conteúdo das variáveis binárias, originalmente constituídas pelos valores "Yes", em casos de ocorrência do evento, e "No" em casos de não ocorrência do evento, para "1" e "0", respectivamente.

A fim da enriquecer os dados, foram coletadas e anexadas ao *dataset* variáveis censitárias da população norte-americana, disponibilizadas pela pesquisa anual *American Community Survey*, a qual de acordo com o United States Census Bureau [USCB] (2022) incorpora características sociais, econômicas, demográficas e habitacionais da nação. 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, aplicou-se a metodologia quantitativa aplicada descritiva.

Tabela 1. Variáveis do *dataset* original, selecionadas para utilização (continua)

| 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 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 |

Tabela 1. Variáveis do *dataset* original, selecionadas para utilização (conclusão)

| Variável | Descrição |
|---|---|
| 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: International Business Machines Corporation [IBM] (2019)

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: Dados 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: Dados originais da pesquisa

Tabela 4. Variáveis censitárias, referentes ao código postal 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: Dados originais da pesquisa

O conjunto de dados foi dividido em partições de treinamento e teste, reservando 70% dos dados para a primeira, e 30% para a segunda partição. Ao todo foram estimados quatro

modelos, sendo um modelo de regressão logística binária clássica, uma árvore de decisão, e duas *random forests*, os quais, com base no comportamento conjunto das variáveis preditoras, calcularam a probabilidade de *churn* de cada cliente. Os parâmetros do modelo de regressão logística foram estimados por máxima verossimilhança.

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 métricas como acurácia, sensitividade, especificidade, e pela área sob a curva ROC (*Receiver Operating Characteristic*).

A curva ROC, segundo James e colaboradores (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.

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

- caret (Kuhn, 2022) – Criação de amostras aleatórias estratificadas para treinamento e teste; construção de matrizes de confusão.
- fastDummies (Kaplan, 2020) – Construção de variáveis binárias, a partir de variáveis categóricas policotômicas.
- ggrepel (Slowikowski, 2021) – Inclusão de rótulos não sobrepostos, nos gráficos.
- glue (Hester e Bryan, 2022) – Operações com dados em formato de texto.
- MASS (Venables e Ripley, 2002) – Seleção de modelos de regressão logística, utilizando AIC.
- randomForest (Liaw e Wiener, 2002) – Treinar modelo *random forest*.
- readxl (Wickham e Bryan, 2019) – Leitura de *dataset* no formato xlsx.
- ROCR (Sing et al., 2005) – Calcular a área sob a curva ROC.
- rpart (Therneau e Atkinson, 2022) – Treinar árvore de decisão.
- rpart.plot (Milborrow, 2022) – Extrair sintaxe das folhas da árvore de decisão.
- stats (R Core Team, 2021) – Treinar modelo de regressão logística binária.
- tidycensus (Walker e Herman, 2022) – Obtenção dos dados censitários da *American Community Survey*.
- tidyverse (Wickham et al., 2019) – Manipulação e transformação de dados.
- tigris (Walker, 2022) – Obtenção de *shapefiles* do estado da Califórnia.
- zipcodeR (Rozzi, 2021) – Consulta das associações entre *zip codes* e condados do estado da Califórnia.

## Resultados e Discussão

Foi constatado 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, conforme apresentado na Figura 1. 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.

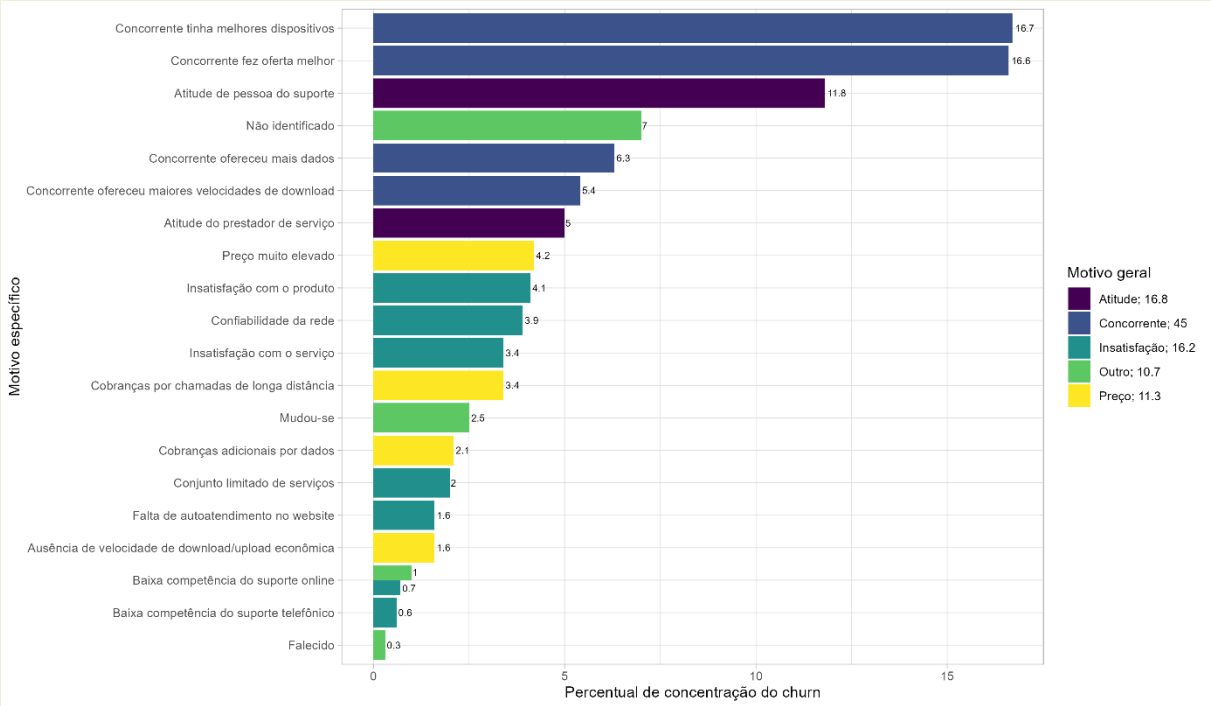


Figura 1. Distribuição dos motivos de *churn*
Fonte: Resultados originais da pesquisa

Para Mattison (2005), que segmenta o fenômeno do *churn* em duas grandes categorias, voluntário e involuntário, são participantes da primeira categoria os clientes que optam pela rescisão do contrato de serviço de forma deliberada ou não; e da segunda categoria, aqueles que têm o contrato rescindido por decisão e vontade da companhia, geralmente motivada por indícios de fraude, não pagamento ou não utilização do serviço.

Tendo em vista essa segmentação, entende-se que a totalidade dos eventos de *churn* presentes na base de dados em estudo é composta por casos voluntários, ou com um percentual de *churn* involuntário máximo equivalente ao da categoria genérica *"Other"*, isto é, 10,7%, constituída por clientes com motivo de cancelamento desconhecido, uma vez que os

demais grupos *"Attitude"*; *"Competitor"*; *"Dissatisfaction"*; e *"Price"*, correspondem a formas distintas de cancelamento voluntário.

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, segundo a qual 100% dos clientes com score de satisfação inferior a três, incorreram no fenômeno do *churn*, e nenhum dos clientes com score maior ou igual a quatro cancelou o serviço.

Tabela 5. Percentual de *chun*, por índice de satisfação do cliente

| satisfaction_score | Quantidade de 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% fizeram no máximo uma indicação. Verificou-se também expressiva diferença no tempo de relacionamento com a companhia entre os consumidores que incorreram no evento do *churn*, e os demais. Para o primeiro público, o tempo médio de relacionamento foi de 18 meses, e para o segundo, de 37 meses, o que revela uma associação negativa entre as variáveis *"tenure_in_months"* e *"flg_churn"*.

Ao fazer uma análise espacial dos dados, identificou-se que os condados com maior presença de clientes estão localizados ao sul do estado da Califórnia, 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, conforme ilustrado na Figura 2.

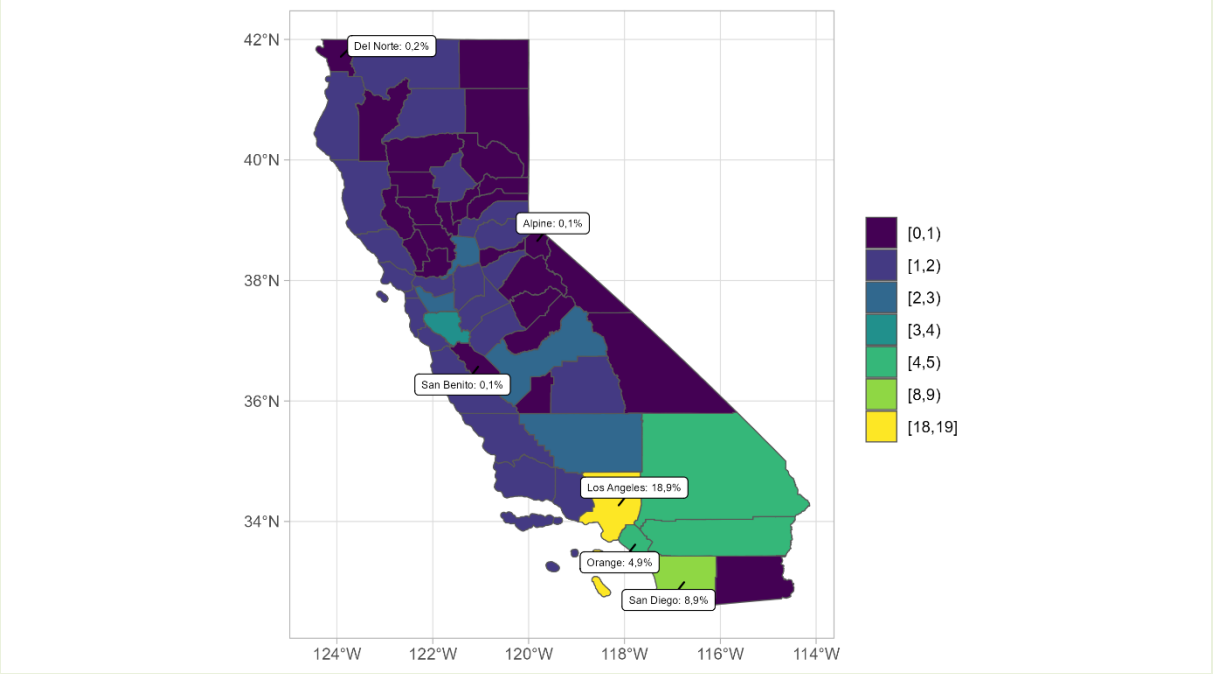


Figura 2. Distribuição de clientes, por condado, com destaque para os condados com maiores e menores representatividades
Fonte: Resultados originais da pesquisa

Salienta-se que os condados menos populosos do estado em que a companhia presta seus serviços, são os mais relevantes no aspecto do percentual da população que é ou foi cliente, como apresentado na Figura 3. 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. Em contrapartida, juntos, os três condados contabilizam apenas 84 clientes, ou 1,2% do total de 7.043 consumidores.

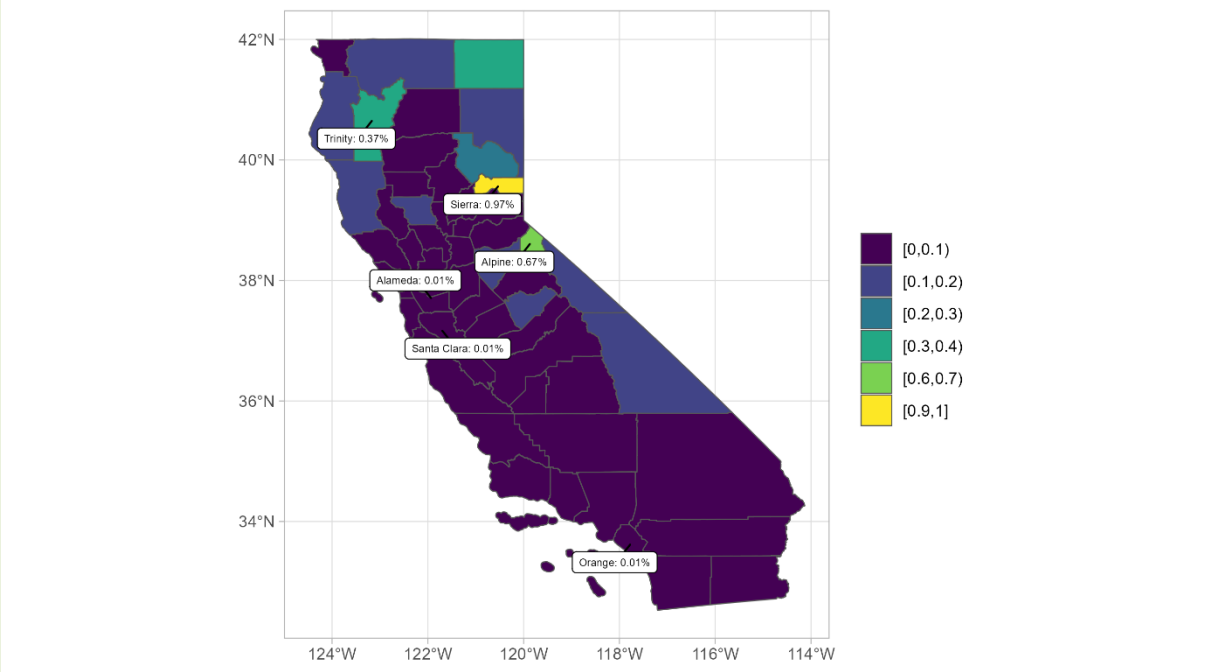


Figura 3. Proporção de habitantes, por condado, que foram ou são clientes da companhia, com destaque para os condados com maiores e menores representatividades
Fonte: Resultados originais da pesquisa

Del Norte, um dos condados com menor participação na carteira de clientes da companhia, foi a localidade com o maior índice de *churn*, como ilustrado na Figura *4*, enquanto San Diego, o segundo condado com a maior quantidade de clientes, foi também o segundo com maior percentual de perda de clientes, seguido por Stanislaus.

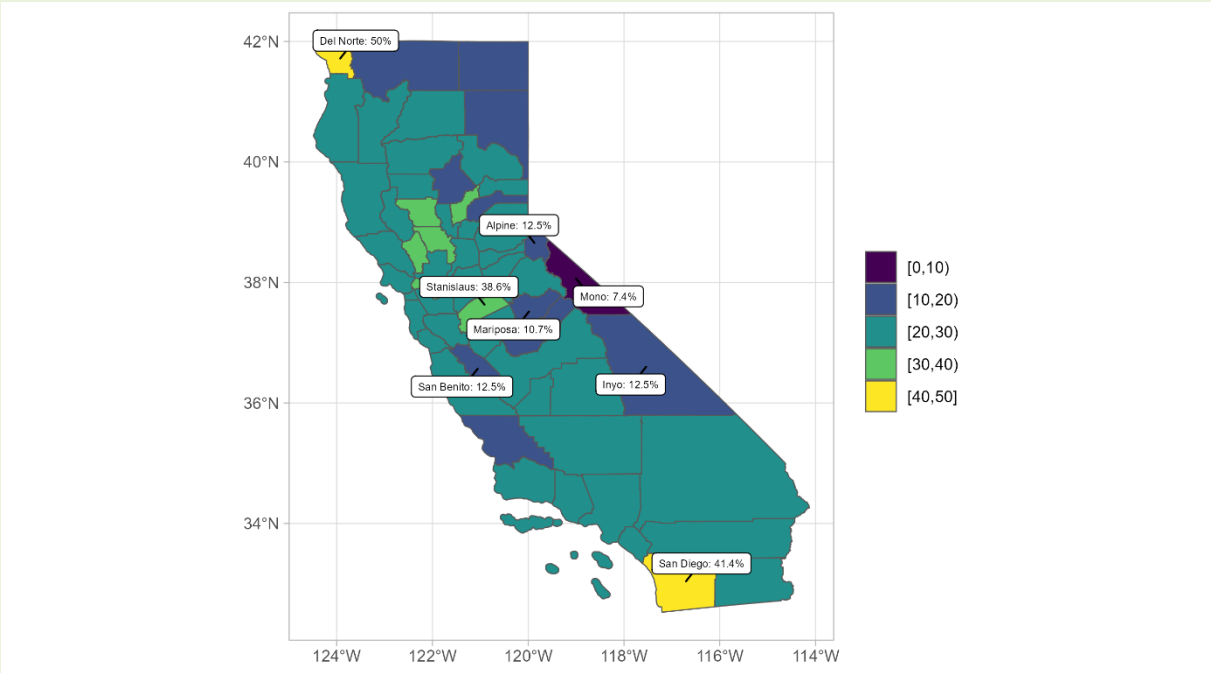


Figura 4. Índice de *churn* por condado
Fonte: Resultados originais da pesquisa

Conforme evidenciado na Figura *5*, 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 Califórnia.

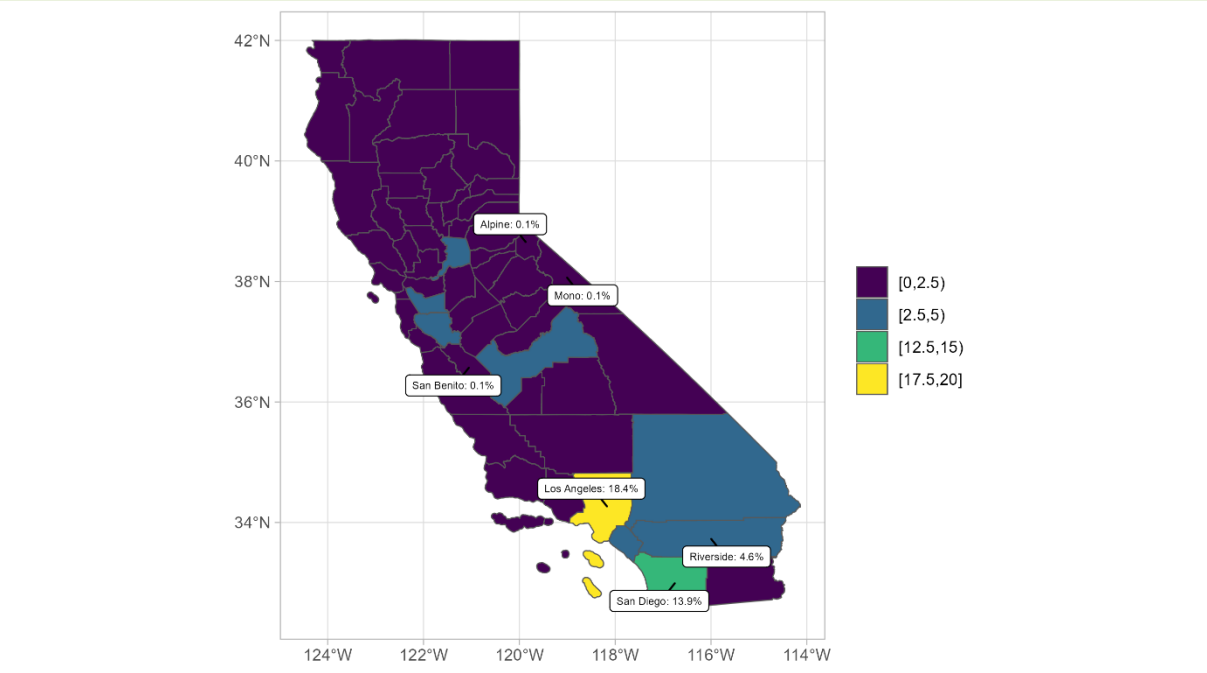


Figura 5. Distribuição do *churn*, por condado
Fonte: Resultados originais da pesquisa

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.

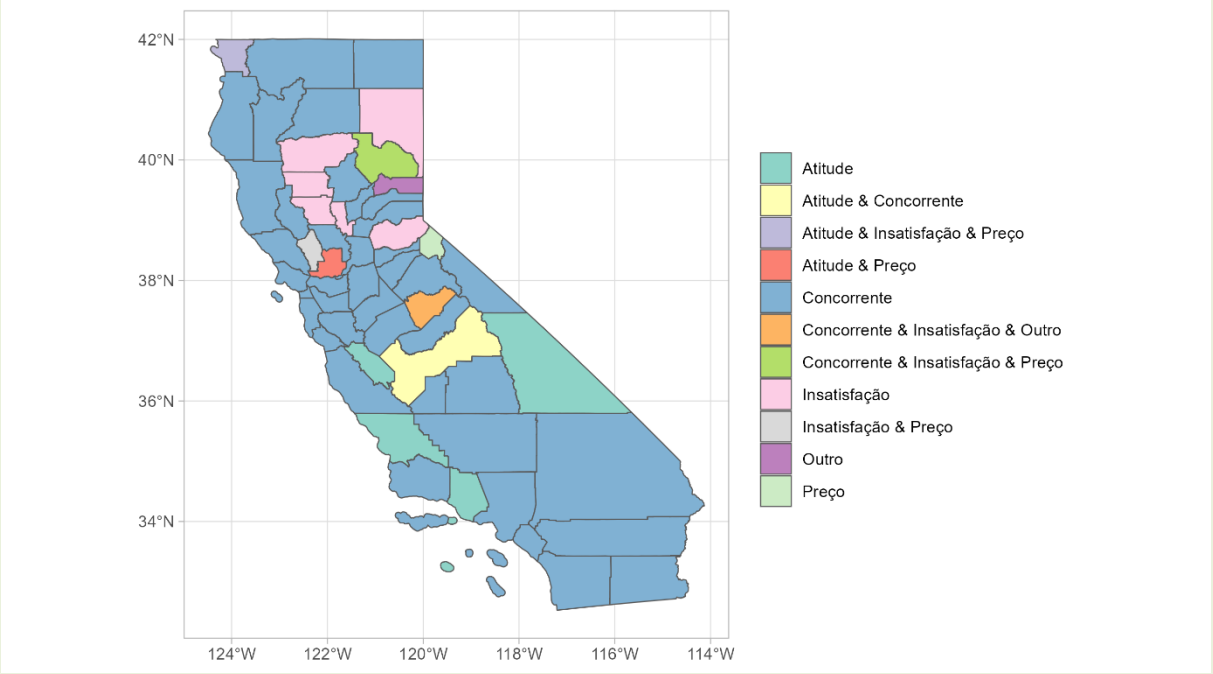


Figura 6. Motivos mais frequentes de *churn*, por condado
Fonte: Resultados originais da pesquisa

Terminada a análise espacial, calculou-se as correlações entre as variáveis dependentes numéricas, onde foi identificado que as três maiores correlações positivas acontecem com quatro variáveis relacionadas ao valor das cobranças, das quais duas foram construídas no processo de *feature engineering*. Conforme apresentado na Tabela 6, detectou-se correlação perfeita entre os valores extras cobrados por chamadas de longa distância e downloads de dados, e o valor das cobranças adicionais por chamadas de longa distância, uma vez que o segundo valor está contido no primeiro.

Tabela 6. Dez maiores correlações, em valor absoluto, entre variáveis preditoras numéricas

| Variável 1 | Variável 2 | Correlação |
|---|---|---|
| valor_cobrancas_extras | total_long_distance_charges | 1,000 |
| valor_cobranca_geral | total_charges | 0,972 |
| valor_cobranca_geral | tenure_in_months | 0,853 |
| qtd_servicos_principais | monthly_charge | 0,836 |
| total_charges | tenure_in_months | 0,826 |
| valor_cobrancas_extras | valor_cobranca_geral | 0,780 |
| valor_cobranca_geral | total_long_distance_charges | 0,779 |
| tx_contrib_cobrancas_extras_cobranca_geral | avg_monthly_long_distance_charges | 0,745 |
| condado_tx_habitantes_menor_18_anos | condado_idade_mediana_habitantes | -0,723 |
| qtd_streamings | monthly_charge | 0,695 |

Fonte: Resultados originais da pesquisa

Tanto as cobranças totais como o tempo de casa do cliente, possuem elevada correlação positiva com o valor das cobranças gerais, representado no *dataset* pela variável “valor_cobranca_geral”, apontando que os valores cobrados, incluindo ou não despesas

extras, seguem a mesma direção, de modo que o movimento ascendente de um está relacionado ao movimento ascendente do outro, e vice-versa. De modo semelhante, o prolongamento do tempo de relacionamento do cliente com a companhia está vinculado a maiores cobranças gerais. Ambas as correlações são superiores a 0.8, e figuram no ranking das três correlações de maior magnitude.

Segundo apresentado na Tabela 6, das dez maiores correlações em valor absoluto, apenas uma é negativa, e ocorre entre o percentual de pessoas com idade inferior a 18 anos, e a idade mediana dos habitantes do condado, sinalizando que nos condados, conforme a proporção de indivíduos menores de idade aumenta, a idade mediana dos habitantes naturalmente reduz, e vice-versa.

Ao anexar variáveis geográficas do *American Community Survey* ao *dataset*, composto por clientes com residência em 1.626 zonas de informação postal distintas, passou-se a ter colunas com valores faltantes, segundo apresentado na Tabela 7.

Tabela 7. Número de observações para as quais dados geográficos de cep estão faltando

| Variável | Quantidade de observações com dados faltantes |
|---|---|
| zip_code_renda_familiar_mediana | 428 |
| zip_code_indice_gini_desigualdade_renda | 104 |
| zip_code_idade_mediana_habitantes | 84 |
| zip_code_tx_habitantes_homens | 20 |
| zip_code_tx_habitantes_menor_18_anos | 20 |

Fonte: Resultados originais da pesquisa

Devido ao impacto da ausência de conteúdo em alguns algoritmos de modelagem, que na etapa de treinamento eventualmente omitem observações com valores faltantes, e a acentuada granularidade da variável *"zip_code"*, que poderia acarretar o sobreajuste do modelo, optou-se por não utilizar na modelagem estatística as variáveis relacionadas a cep.

Seguindo o mesmo raciocínio, as colunas pertinentes às cidades nas quais estão localizadas as residências dos clientes também foram excluídas da modelagem. Mantiveram-se, entretanto, as variáveis com dados dos 58 condados nos quais a firma presta serviço, em virtude de o condado ser a unidade geográfica com menor granularidade disponível no *dataset*.

Além destas, foram removidas do processo de modelagem outras seis variáveis, sendo elas *"customer_id"*; *"latitude"*; *"longitude"*; *"customer_status"*; *"churn_category"* e *"churn_reason"*, as quais, de acordo com o conhecimento prévio do negócio, são incapazes de contribuir para explicar o fenômeno em análise, ou são desdobramentos da variável resposta, como é o caso das três últimas.

Inicialmente treinou-se o modelo clássico de regressão logística com todas as variáveis manualmente escolhidas, e posteriormente aplicou-se o procedimento de seleção de variáveis, com diferentes critérios, mantendo no modelo, apenas variáveis que contribuíssem para a redução do AIC (Akaike Information Criterion). A finalidade desse procedimento foi obter um modelo com maior capacidade preditiva, visto que de acordo com James e colaboradores (2021), excluir variáveis que não ajudam a prever o evento de interesse, reduz a taxa de erro nos dados de teste.

Na Figura 7, observa-se que na estratégia regressiva de seleção das variáveis, o modelo iniciou completo, com todas as variáveis disponíveis para treinamento, e a quantidade de variáveis reduziu a cada uma das 66 iterações, tendo iniciado com um total de 115, e terminado com 50 variáveis dependentes. Nas abordagens progressiva e bidirecional, o modelo iniciou vazio, sem nenhuma variável preditora, e o número de variáveis dependentes ampliou-se de igual modo até a vigésima-quinta iteração.

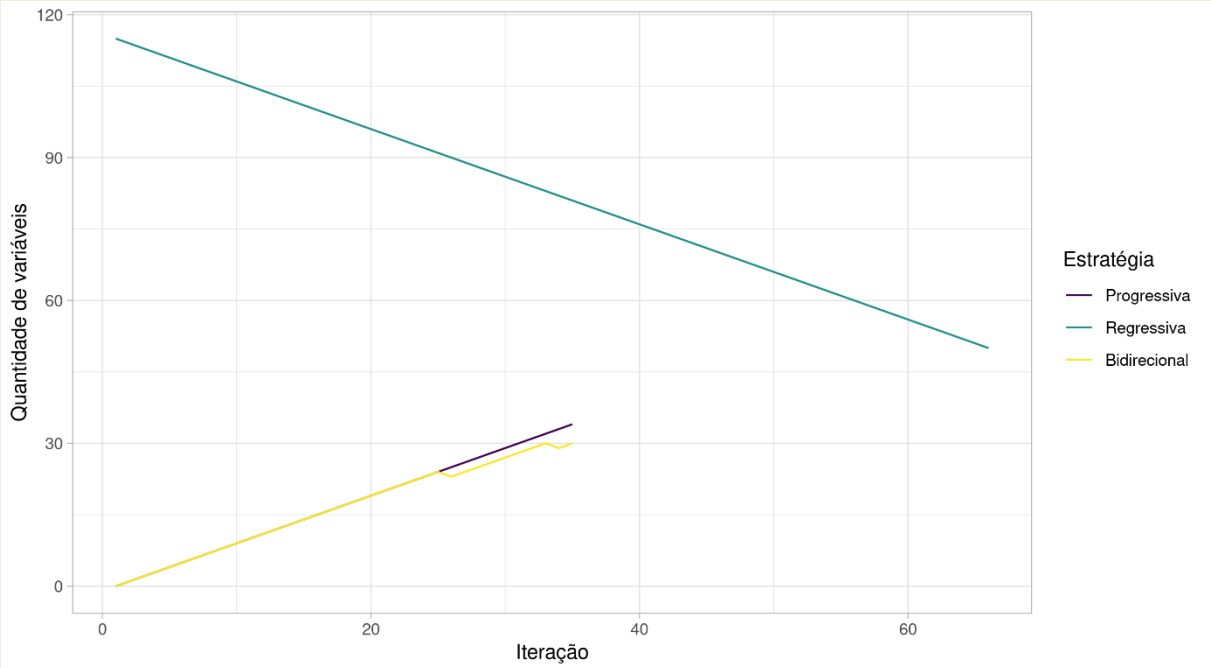


Figura 7. Quantidade de variáveis selecionadas, em cada iteração do procedimento de seleção gradual, para diferentes estratégias
Fonte: Resultados originais da pesquisa

Na vigésima-sexta iteração, a estratégia progressiva anexou uma nova variável, e de modo antagônico, o critério bidirecional excluiu uma variável do modelo. Outra ruptura de padrão deu-se na trigésima-quarta iteração, na qual uma nova variável foi excluída pelo critério de seleção bidirecional. Ressalta-se que na estratégia progressiva, o número de variáveis aumentou a cada iteração.

A cada iteração, o AIC do modelo foi reduzido, o qual atingiu um valor mínimo ao final do processo. Conforme ilustrado na Figura 8, a desigualdade entre os AICs dos modelos estimados pelas estratégias progressiva e bidirecional, manteve-se mínima durante todo o processo.

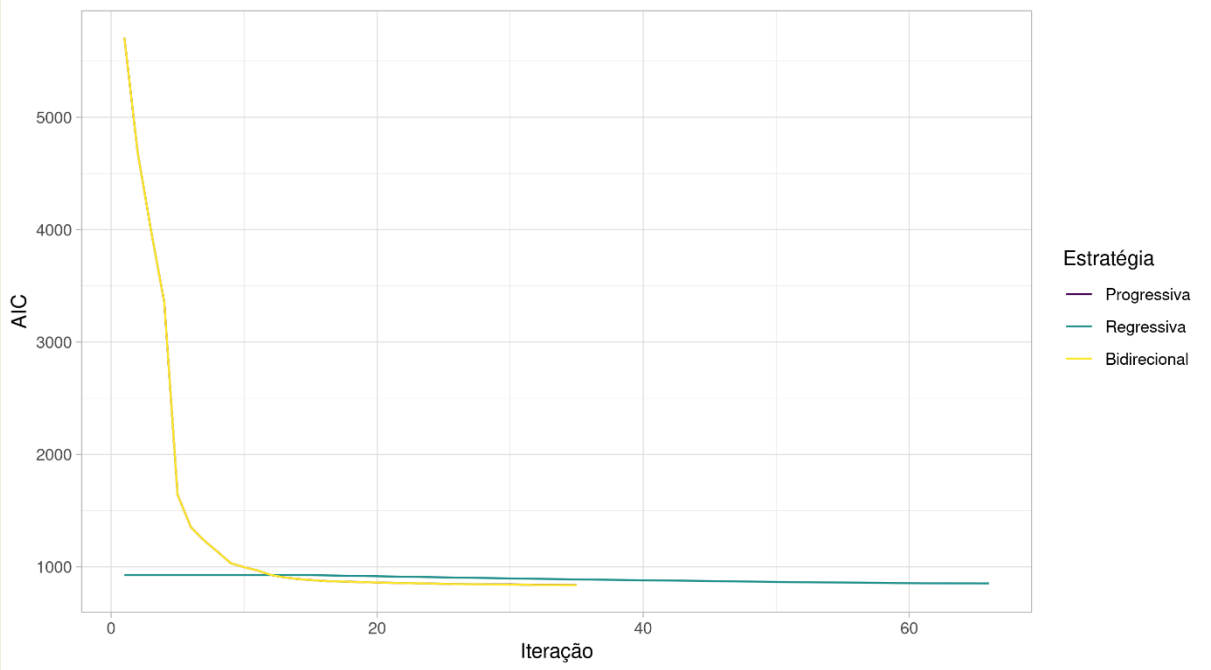


Figura 8. Redução do AIC, a cada iteração do procedimento de seleção gradual de variáveis, para diferentes estratégias
Fonte: Resultados originais da pesquisa

De acordo com a Tabela 8, das três estratégias, a bidirecional resultou no modelo mais performático, alcançando um AIC de 836,78, com um total de 30 variáveis preditoras, ao final de 35 iterações. A quantidade de iterações na estratégia progressiva foi idêntica, a qual estimou um modelo final com quatro variáveis preditoras a mais, e AIC superior em 1.93 unidades ao do melhor modelo. Por sua vez, a seleção regressiva resultou no modelo menos performático, com o maior número de variáveis dependentes, 20 a mais que o melhor modelo, e AIC igual a 852,35.

Tabela 8. Resumo das reduções de AIC, no treinamento dos modelos de regressão logística, com diferentes estratégias de seleção

| Estratégia | Iteração | AIC | Quantidade de variáveis |
|---|---|---|---|
| Progressiva | 1 | 5709,08 | 0 |
| Progressiva | 35 | 838,71 | 34 |
| Regressiva | 1 | 926,07 | 115 |
| Regressiva | 66 | 852,35 | 50 |
| Bidirecional | 1 | 5709,08 | 0 |
| Bidirecional | 35 | 836,78 | 30 |

Fonte: Resultados originais da pesquisa

A capacidade preditiva do melhor modelo foi avaliada para diferentes pontos de corte, e verificou-se, como ilustrado na Figura 9, a existência de uma relação diretamente proporcional entre o aumento do ponto de corte e a queda da sensitividade, de modo que quanto maior a probabilidade mínima aceita para que um cliente seja classificado como perdido para a companhia, menor o número de clientes que incidem no evento e são assim classificados pelo modelo.

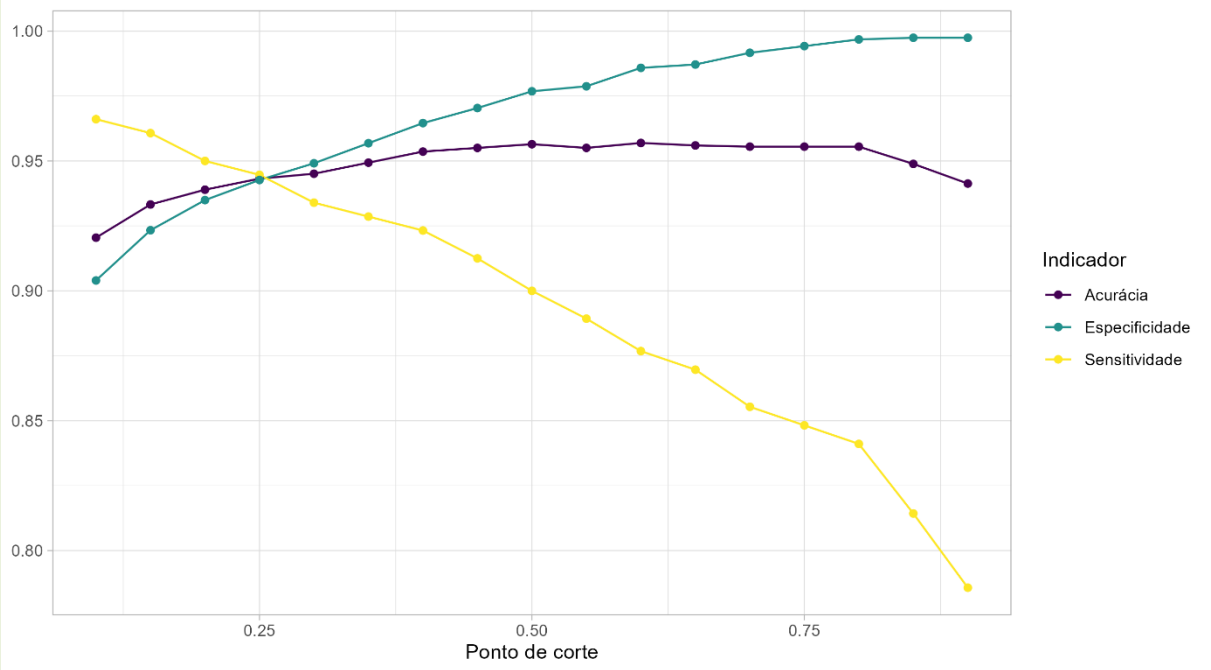


Figura 9. Indicadores da capacidade preditiva do modelo de regressão logística com seleção bidirecional de variáveis, para diferentes pontos de corte
Fonte: Resultados originais da pesquisa

Por exemplo, ao adotar um ponto de corte de 0,95, apenas clientes com probabilidade maior ou igual a 95% de incidirem no evento de *churn*, serão classificados como perdidos. Contudo, de todos os clientes efetivamente perdidos, apenas 78,6% atendem a esse critério, de modo que os demais 21,4%, deixam de ser classificados como incidentes no evento.

Observou-se que a diferença entre acurácia, especificidade e sensitividade foi reduzida ao definir o ponto de corte em 0,25. O corte tradicional de 0,5 é suficiente para que 90% dos consumidores efetivamente perdidos, e 97% dos clientes que continuam utilizando os serviços da companhia sejam corretamente classificados como incidentes e não incidentes no evento de *churn*, respectivamente.

Dos coeficientes estimados para o melhor modelo de regressão logística, os coeficientes associados às variáveis *"satisfaction_score_3"*, *"satisfaction_score_4"* e *"satisfaction_score_5"* são negativos, e o p-valor vinculado a cada uma das três variáveis não é estatisticamente significante, conforme apresentado na Tabela 9. A negatividade destes

coeficientes indica que clientes com índices de satisfação iguais a três, quatro ou cinco possuem probabilidade de *churn* inferior aos demais, o que intuitivamente é o comportamento esperado, uma vez que índices maiores indicam maior satisfação do cliente com a companhia.

Tabela 9. Coeficientes estimados para cada uma das variáveis preditoras do modelo de regressão logística clássica

| Variável preditora | Coeficiente | Erro padrão | Valor z | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 34,794 | 1.380,479 | 0,025 | 0,980 |
| satisfaction_score_4 | -60,216 | 1.747,677 | -0,034 | 0,973 |
| satisfaction_score_5 | -60,679 | 1.970,931 | -0,031 | 0,975 |
| satisfaction_score_3 | -38,961 | 1.380,479 | -0,028 | 0,977 |
| flg_online_security1 | -3,701 | 0,451 | -8,214 | 0,000 |
| number_of_referrals | -0,672 | 0,110 | -6,115 | 0,000 |
| monthly_charge | 0,077 | 0,012 | 6,242 | 0,000 |
| tx_concentracao_cobranca_mes_q3 | 2,176 | 0,364 | 5,976 | 0,000 |
| contract_Two_Year | -2,364 | 0,425 | -5,558 | 0,000 |
| flg_married1 | 1,832 | 0,286 | 6,396 | 0,000 |
| number_of_dependents | -0,964 | 0,191 | -5,048 | 0,000 |
| county_San_Diego_County | 1,312 | 0,290 | 4,524 | 0,000 |
| flg_premium_tech_support1 | -0,933 | 0,249 | -3,753 | 0,000 |
| flg_phone_service1 | -2,376 | 0,573 | -4,149 | 0,000 |
| contract_One_Year | -0,851 | 0,279 | -3,051 | 0,002 |
| offer_Offer_E | 0,611 | 0,275 | 2,223 | 0,026 |
| county_Mendocino_County | -18,072 | 2.180,543 | -0,008 | 0,993 |
| offer_Offer_A | 1,477 | 0,537 | 2,748 | 0,006 |
| county_Lake_County | -16,676 | 1.029,119 | -0,016 | 0,987 |
| age | 0,014 | 0,006 | 2,441 | 0,015 |
| county_Nevada_County | -18,672 | 2.308,246 | -0,008 | 0,994 |
| tx_contrib_cobrancas_extras_cobranca_geral | 5,731 | 2,018 | 2,840 | 0,005 |
| county_Fresno_County | 1,038 | 0,542 | 1,914 | 0,056 |
| county_El_Dorado_County | 1,444 | 0,828 | 1,743 | 0,081 |
| county_Tulare_County | -1,463 | 0,806 | -1,816 | 0,069 |
| total_charges | 0,000 | 0,000 | -2,647 | 0,008 |
| county_San_Mateo_County | -1,743 | 1,120 | -1,556 | 0,120 |
| avg_monthly_long_distance_charges | -0,050 | 0,019 | -2,626 | 0,009 |
| valor_cobrancas_extras | 0,001 | 0,000 | 2,098 | 0,036 |
| internet_type_Fiber_Optic | -0,792 | 0,371 | -2,135 | 0,033 |
| flg_online_backup1 | -0,395 | 0,231 | -1,711 | 0,087 |

Fonte: Resultados originais da pesquisa

De acordo com James e colaboradores (2021), a acurácia dos coeficientes estimados é medida por seus respectivos erros padrão, e valores absolutos elevados da estatística z servem de evidência contrária à hipótese nula, segundo a qual a variável dependente associada ao coeficiente não é útil para prever o evento de interesse, e, portanto, o coeficiente é igual a zero. Esse comportamento pode ser observado na Tabela 9, na qual valores

absolutos pequenos da estatística z ocorrem em coeficientes cujo p-valor não é estatisticamente significante, como acontece com as variáveis de satisfação do cliente, e algumas referentes ao condado onde está localizada a residência do cliente *"county_Mendocino_County"*, *"county_Lake_County"*, relacionadas aos condados de Mendocino e Lake, respectivamente.

Tratando as variáveis preditoras como grupos, pensou-se em aplicar a regressão logística multinível, a qual possibilita modelar cenários nos quais a chance de ocorrência do evento, bem como o efeito das variáveis preditoras, varia entre os grupos (Sommet e Morselli, 2017). Inicialmente quis-se identificar quais eram as variáveis dependentes categóricas com maiores coeficientes de correlação intraclasse ajustado, a fim de serem posteriormente utilizadas como variáveis de nível dois, ou efeitos aleatórios, e verificou-se, conforme apresentado na Tabela 10, que das dezoito variáveis selecionadas, quatro possuem coeficiente maior ou igual a 0,18, sendo *"satisfaction_score"*, a variável com maior coeficiente de correlação intraclasse.

Tabela 10. Coeficientes de correlação intraclasse, calculados para variáveis dependentes categóricas

| Efeito aleatório | Coeficiente de correlação intraclasse ajustado |
|---|---|
| satisfaction_score | 0,990 |
| contract | 0,381 |
| flg_internet_service | 0,191 |
| offer | 0,180 |
| internet_type | 0,155 |
| payment_method | 0,085 |
| flg_online_security | 0,068 |
| flg_paperless_billing | 0,064 |
| flg_premium_tech_support | 0,061 |
| flg_unlimited_data | 0,056 |
| flg_married | 0,035 |
| county | 0,013 |
| flg_online_backup | 0,012 |
| flg_device_protection_plan | 0,007 |
| flg_streaming_tv | 0,006 |
| flg_streaming_movies | 0,006 |
| flg_streaming_music | 0,003 |
| flg_multiple_lines | 0,002 |

Fonte: Resultados originais da pesquisa

Entretanto, encontrou-se a ressalva pontuada por Sommet e Morselli (2017), segundo os quais a regressão multinível é aplicada em cenários em que os dados estão aninhados,

sendo que uma das diferenças fundamentais entre variáveis que representam níveis e variáveis preditoras, é que os níveis não possuem significado intrínseco.

Desse modo, no *dataset* utilizado no estudo, revelou-se que apenas a variável *"county"*, representando o condado da residência do cliente, satisfaz esse critério, a qual possui um coeficiente de correlação intraclasse igual a 0,013, revelando que a chance de *churn* varia pouco entre os condados, de sorte que apenas 1,3% da chance é explicada pelas diferenças existentes entre os condados. Consequentemente, concluiu-se que um modelo de regressão logística clássico, de um nível, basta para o presente estudo.

Treinaram-se também modelos de árvore de decisão e *random forest*, a fim de terem a performance comparada com a do modelo de regressão logística. O modelo *random forest*, inicialmente foi treinado com mil árvores, sem a variável *"county"*, devido a incapacidade do algoritmo utilizado em trabalhar com fatores cuja quantidade de níveis ultrapassa 53. Posteriormente transformou-se a coluna *"county"* em 57 variáveis binárias, a fim de que fosse possível utilizá-la no treinamento.

O modelo de árvore de decisão foi capaz de detectar a completa ausência do evento de *churn* em clientes com índice de satisfação igual a quatro ou cinco, e identificou que todos os clientes com score igual a um ou dois foram perdidos pela companhia, de modo que estas duas características constituem duas folhas da árvore. Para os demais clientes, com score de satisfação igual a três, uma das folhas indica que clientes que assinam o serviço de segurança online, e cuja representatividade do valor da mensalidade atual, no montante cobrado ao longo do trimestre é inferior a 19%, tem elevado índice de *churn*, sendo que 14,93% dos clientes da base de treinamento detêm essas características, dos quais 82% foram perdidos pela companhia.

Os três principais indicadores dos modelos estão ilustrados na Figura 10, para diferentes pontos de corte, na qual evidencia-se que o comportamento dos modelos de regressão logística e árvore de decisão é bastante semelhante entre si, os quais, para pontos de cortes inferiores ou iguais a 0,25, possuem maior acurácia e especificidade que os modelos *random forest*. Verificou-se também que a acurácia e principalmente a especificidade destes últimos sofre grande deterioração para cortes superiores a 0,5.

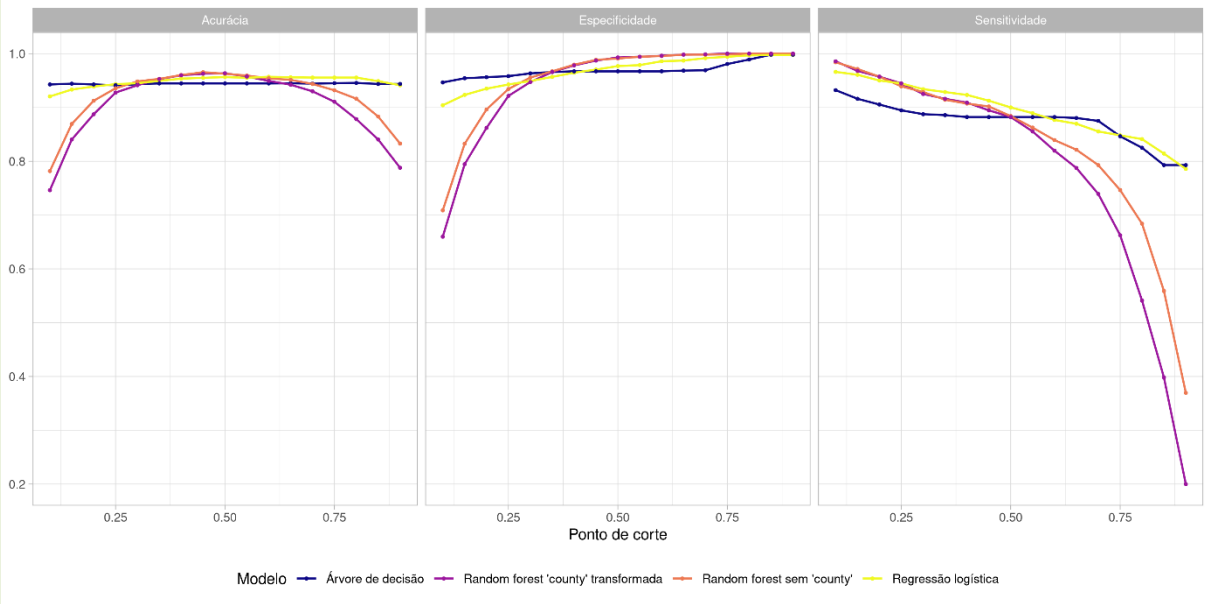


Figura 10. Indicadores da capacidade preditiva dos diversos modelos estimados, para diferentes pontos de corte
Fonte: Resultados originais da pesquisa

Por fim, foram calculadas as áreas sob as curvas ROC, dos quatro modelos, como indicador geral da performance, as quais estão sumarizadas na Tabela 11, na qual revela-se que o modelo de regressão logística treinado com seleção gradual de variáveis em abordagem bidirecional, possui maior capacidade preditiva que os demais modelos. Evidencia-se também que a diferença máxima entre as áreas sob as curvas de todos os modelos, é inferior a 0,01, de modo que, por esta métrica, não são percebidas grandes diferenças de performance entre os modelos.

Tabela 11. Área sob a curva ROC, de diversos modelos estimados

| Modelo | AUC |
|---|---|
| Regressão logística com seleção bidirecional de variáveis | 0,9885 |
| Árvore de decisão | 0,9854 |
| Random forest sem a variável categórica "county" | 0,9836 |
| Random forest com a variável "county" transformada | 0,9808 |

Fonte: Resultados originais da pesquisa

## Conclusão

O modelo de prevenção ao *churn* apresenta às companhias a oportunidade de minimizar as perdas de clientes, além de auxiliar na identificação das principais características distintivas entre os clientes que continuam a utilizar os serviços prestados pela companhia, e os demais, que optaram por romper o relacionamento de prestação de serviços.

Os experimentos realizados com o *dataset* composto por 7.043 observações demonstraram que o índice de satisfação do cliente é a variável com maior capacidade de separação entre os clientes perdidos e os demais, possibilitando separar perfeitamente 62, 16% dos consumidores como incidentes ou não no evento estudado. O modelo com maior capacidade preditiva foi o de regressão logística, com seleção gradual de variáveis, em abordagem bidirecional. Ainda assim, a performance dos demais modelos avaliados, não revelaram grandes diferenças ao comparar as áreas sob a curva ROC, permitindo concluir que a utilização do modelo traria benefícios reais à companhia, em suas ações de retenção de clientes, e de melhoria na prestação de serviços.

## Agradecimento

Agradeço e dedico o presente trabalho à minha mãe.

## Referências

Anderson, E. W.; Sullivan, M. W. 1993. The Antecedents and Consequences of Customer Satisfaction. Marketing Science 12: 125-143.

Caldeira, S. 2006. Retenção de Clientes. p. 165-184. In: Correia, A.; Sacavém, A.; Colaço, C. Manual de Fitness & Marketing. Visão e Contextos, Lisboa, Lisboa, Portugal.

Ferreira, C. M. C. 2012. Um estudo sobre fidelização e retenção de clientes na área do fitness. Dissertação. Instituto Politécnico de Castelo Branco, Castelo Branco, Castelo Branco, Portugal.

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

Glady, N.; Baesens, B.; Croux, C. 2009. Modeling churn using customer lifetime value. European Journal of Operational Research 197: 402-411.

Hester, J.; Bryan, J. 2022. glue: Interpreted String Literals. R package version 1.6.2. Disponível em: <https://cran.r-project.org/package=glue>. Acesso em: 22 maio 2022.

Instituto Brasileiro de Geografia e Estatística [IBGE]. 2019. Pesquisa Anual de Serviços. Disponível em: <https://ftp.ibge.gov.br/Comercio_e_Servicos/Pesquisa_Anual_de_Servicos/pas2019/xlsx/tabelas_2019_xlsx.zip>. Acesso em: 08 maio 2022.

Instituto Brasileiro de Geografia e Estatística [IBGE]. 2019. Pesquisa Nacional por Amostra de Domicílios Contínua. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv101794_informativo.pdf>. Acesso em: 09 maio 2022.

International Business Machines Corporation [IBM]. 2019. Telco customer churn (11.1.3+). Disponível em: <https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113>. Acesso em: 24 jul. 2022.

Jahromi, A. T.; Stakhovych, S.; Ewing, M. 2014. Managing B2B customer churn, retention and profitability. Industrial Marketing Management 43: 1258-1268.

James, G.; Witten, D.; Hastie, T.; Tibshirani, R. 2021. An Introduction to Statistical Learning with applications in R. 2ed. Springer. New York, New York, USA. Disponível em: <https://web.stanford.edu/~hastie/ISLR2/ISLRv2_website.pdf>. Acesso em: 01 out. 2021.

Kaplan, J. 2020. fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables. R package version 1.6.3. Disponível em: <https://cran.r-project.org/package=fastDummies>. Acesso em: 12 set. 2022.

Kuhn, M. 2022. caret: Classification and Regression Training. R package version 6.0-91. Disponível em: <https://cran.r-project.org/package=caret>. Acesso em: 22 maio 2022.

Liaw, A; Wiener, M. 2002. Classification and Regression by randomForest. R News 2(3): 18-22.

Mattison, R. 2005. The Telco Churn Management Handbook. XiT Press, Oakwood Hills, IL, USA.

Milborrow, S. 2022. rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'. R package version 3.1.1. Disponível em: <https://cran.r-project.org/package=rpart.plot>. Acesso em: 12 set. 2022.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: <https://www.R-project.org/>. Acesso em: 19 out. 2021.

Rozzi, G. C. 2021. zipcodeR: Advancing the analysis of spatial data at the ZIP code level in R. Software Impacts 9: 100099.

Sing, T.; Sander, O.; Beerenwinkel, N.; Lengauer, T. 2005. ROCR: visualizing classifier performance in R. Bioinformatics 21(20): 7881.

Slowikowski, K. 2021. ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. R package version 0.9.1. Disponível em: <https://cran.r-project.org/package=ggrepel>. Acesso em: 22 maio 2022.

Sommet, N.; Morselli, D. 2017. Keep Calm and Learn Multilevel Logistic Modeling: A Simplified Three-Step Procedure Using Stata, R, Mplus, and SPSS. International Review of Social Psychology, 30(1): 203-218.

Therneau, T.; Atkinson, B. 2022. rpart: Recursive Partitioning and Regression Trees. R package version 4.1.16. Disponível em: <https://cran.r-project.org/package=rpart>. Acesso em: 12 set. 2022.

United States Census Bureau [USCB]. 2022. American Community Survey 5-Year Data (2009-2020). Disponível em: <https://www.census.gov/data/developers/data-sets/acs-5year.2017.html>. Acesso em: 22 maio 2022.

Venables, W.; Ripley, B. 2002. Modern Applied Statistics with S. 4ed. Springer. New York, New York, USA.

Walker, K. 2022. tigris: Load Census TIGER/Line Shapefiles. R package version 1.6. Disponível em: < https://cran.r-project.org/package=tigris>. Acesso em: 22 maio 2022.

Walker, K.; Herman, M. 2022. tidycensus: Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames. R package version 1.2. Disponível em: <https://cran.r-project.org/package=tidycensus>. Acesso em: 22 maio 2022.

Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L. D.; François, R.; Grolemund, G.; Hayes, A.; Henry, L.; Hester, J.; Kuhn, M.; Pedersen, T. L.; Miller, E.; Bache, S. M.; Müller, K.; Ooms, J.; Robinson, D.; Seidel, D. P.; Spinu, V.; Takahashi, K.; Vaughan, D.; Wilke, C.; Woo, K.; Yutani, H. 2019. Welcome to the tidyverse. Journal of Open Source Software 4: 1686.

Wickham, H.; Bryan, J. 2019. readxl: Read Excel Files. R package version 1.3.1. Disponível em: <https://cran.r-project.org/package=readxl>. Acesso em: 22 maio 2022.