UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
BACHARELADO EM ESTATÍSTICA

WILLIAM IRINEU ALVES DE LIMA

Prevendo o Churn de Clientes em Telecom com Regressão Logística

Goiânia
2024

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
BACHARELADO EM ESTATÍSTICA

WILLIAM IRINEU ALVES DE LIMA

Prevendo o Churn de Clientes em Telecom com Regressão Logística

Projeto de Análise de Dados apresentado ao Curso de Bacharelado em Estatística da Universidade Federal de Goiás como parte das exigências para aprovação no componente curricular Modelos Lineares Generalizados, visando o desenvolvimento de habilidades práticas na aplicação de técnicas estatísticas para a análise de dados.

Goiânia

2024

Sumário

1 Introdução

A rotatividade de clientes, ou “Churn”, é um fenômeno que ocorre quando uma determinada pessoa deixa de fazer negócios com uma empresa ou serviço que mantinha contrato. É um fenômeno que aflige diversos ramos, como por exemplo clientes de uma empresa telefônica, clientes que compram um produto mensalmente e a saída de funcionários de uma empresa.

Perder clientes implica perda de receita, custos adicionais com a reaquisição e um impacto negativo na reputação da marca. No entanto, o churn também apresenta uma oportunidade valiosa para a indústria, permitindo que as empresas compreendam melhor as necessidades de seus clientes e reavalie suas estratégias, a fim de oferecer serviços mais personalizados e eficientes (MATTISON, 2005).

Em 2023, segundo os dados coletados pela ANATEL, o setor de telecomunicações no Brasil atingiu um marco com 47 milhões de acessos de banda larga fixa, um crescimento de 4,1% em relação a 2022. A competição entre grandes e pequenos provedores, como Claro, Vivo, Tim, Oi e Algar, tem se intensificado, impulsionada pela crescente demanda por velocidade e conexão.

Considerando esses avanço temos que , perder clientes implica perda de receita, custos adicionais com a reaquisição e um impacto negativo na reputação da marca.

1.1 Justificativa para a Escolha do Tema e da Análise

A escolha do tema baseia-se em minha experiência de três anos em análise de dados, cursando Estatística pela UFG e atuando em uma empresa de telecomunicações, onde analiso rotatividade de clientes (churn) e desenvolvo estratégias de retenção em vendas de pacotes de internet. Além disso, tenho experiência com People Analytics, analisando turnover em RH, o que me oferece uma visão crítica sobre retenção de clientes e colaboradores. Esse trabalho permitirá aplicar e aprofundar técnicas de regressão logística em um contexto prático, identificando fatores de cancelamento e propondo ações para retenção, unindo teoria e prática para gerar insights estratégicos, além disso dada a relevância do churn no contexto de negócios e a necessidade de estratégias eficientes de retenção, a análise preditiva se mostra essencial a utilização de modelos estatísticos, como a regressão logística, pode fornecer insights valiosos sobre os fatores que levam os clientes a cancelar seus contratos.

1.2 Objetivo do Trabalho

Este trabalho busca desenvolver e ajustar um modelo de regressão logística que permita prever a probabilidade de churn entre clientes de uma empresa de telecomunicações. Com a análise, pretende-se identificar as principais variáveis associadas ao cancelamento e gerar "insights"práticos para a empresa, possibilitando que tome decisões informadas sobre estratégias de retenção de clientes, além disso pretende-se realizar a analise descritiva dos dados pra obter outros resultados que auxilie a tomada de descisão.

2 Metodologia

Neste estudo, aplicamos um modelo de regressão logística para prever a probabilidade de churn (cancelamento) de clientes em uma empresa de telecomunicações, utilizando variáveis demográficas, características dos serviços contratados e detalhes financeiros dos clientes. A metodologia seguiu as etapas descritas a seguir.

2.0.1 Seleção e Preparação dos Dados

Inicialmente, selecionamos um conjunto de dados contendo informações de 7043 clientes, incluindo a variável resposta binária (Churn: Sim ou Não) e diversas variáveis explicativas, como dados demográficos, características dos serviços e informações financeiras. O pré-processamento dos dados envolveu:

- Limpeza de valores ausentes e tratamento de dados inconsistentes.
- Transformação de variáveis categóricas em variáveis dummy, garantindo que o modelo pudesse utilizar informações de variáveis não numéricas.

2.0.2 Divisão dos Dados

Para assegurar uma avaliação robusta do modelo, dividimos o conjunto de dados em duas amostras: 70% dos dados foram reservados para o treinamento do modelo, enquanto os 30% restantes foram utilizados como amostra de teste, permitindo a validação do modelo em dados não vistos.

2.0.3 Ajuste do Modelo:

Para esse trabalho, será utilizado a regressão logística com função de ligação logit, utilizamos a função logística para transformar o preditor linear em uma probabilidade. A função logística é representada como:

$$f(\eta) = \frac{1}{1 + e^{-\eta}}$$

onde $\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$ é o preditor linear que combina os coeficientes $\beta_0, \beta_1, \ldots, \beta_n$ e as variáveis independentes $X_1, X_2, \ldots, X_n$. Esta transformação nos permite modelar a probabilidade do evento de interesse (por exemplo, churn: sim ou não) como um valor contínuo entre 0 e 1, facilitando a interpretação do resultado em termos probabilísticos. Esses coeficientes indicam a força e a direção da relação entre cada variável independente e

a probabilidade de ocorrência do evento de interesse. Adicionalmente, a seleção das variáveis mais significativas foi realizada com o Critério de Informação de Akaike (AIC), permitindo um modelo mais conciso e eficiente.

2.0.4 Seleção de Variáveis

Para obter um modelo parcimonioso, ira ser aplicado o Critério de Informação de Akaike (AIC) durante a etapa de seleção de variáveis. Esse critério penaliza a complexidade excessiva do modelo, incentivando a escolha das variáveis que mais contribuem para a explicação da variável resposta, removendo aquelas que têm pouca ou nenhuma contribuição significativa.

2.0.5 Avaliação do Modelo

A qualidade do ajuste será verificada por meio de uma análise de resíduos, identificando potenciais outliers e pontos influentes que poderiam afetar o desempenho do modelo. Além disso, será verificado a multicolinearidade entre as variáveis preditoras foi analisada utilizando o Fator de Inflação da Variância (VIF). Variáveis com VIFs altos irão ser foram removidas para minimizar os problemas de redundância e melhorar a estabilidade das estimativas.

2.0.6 Métricas de Desempenho

Para avaliar a capacidade preditiva do modelo, utilizamos as seguintes métricas:

- **Curva ROC e AUC (Área sob a Curva):** A curva ROC foi construída para avaliar a discriminação do modelo entre as classes de churn. A AUC fornece uma medida de desempenho global, indicando a eficácia do modelo em classificar corretamente os clientes que cancelaram.
- **Acurácia, Sensibilidade e Especificidade:** Essas métricas foram calculadas para avaliar a performance do modelo em prever corretamente a classe da variável resposta.
- **Teste de Hosmer-Lemeshow:** Esse teste de bondade de ajuste foi aplicado para verificar se o modelo representa bem os dados observados, com hipótese nula indicando bom ajuste.

2.0.7 Interpretação dos Resultados

Os coeficientes estimados foram interpretados em termos de *odds ratios*, permitindo a compreensão do impacto de cada variável preditora na probabilidade de churn. Para isso, aplicamos a função exponencial aos coeficientes estimados (e^{β_i}), convertendo-os em *odds ratios*, o que facilita a interpretação dos efeitos das variáveis.

2.0.8 Dados para tomada de decisão

Variáveis com coeficientes significativos serão utilizadas para fornecer insights sobre os fatores mais influentes no comportamento de cancelamento dos clientes, destacando o quanto aumentam ou reduzem a probabilidade de churn.

3 Análise e Resultados

Cada linha representa um cliente, a base contém informações sobre 7043 clientes de uma empresa de telecomunicações na Califórnia no terceiro trimestre(CHURN-IBM,). Cada cliente é identificado por um ID único e descrito por uma série de variáveis, como:

- Dados demográficos (sexo, idoso, possui parceiro e dependentes); Informações sobre os serviços contratados (telefone, internet, segurança online, streaming, suporte técnico);
- Detalhes financeiros (método de pagamento, cobranças mensais e totais);
- A variável de interesse é Churn, que indica se o cliente cancelou o serviço ou não.

Observe a Figura 1, ela descreve o tipo de cada coluna e qual os valores unicos presente em cada uma.

Descrição das Variáveis	Tipo	Categorias
ID do Cliente	Texto	Código Único
Gênero	Categórica	Female, Male
Idoso	Binária	No, Yes
Parceiro	Binária	Yes, No
Dependentes	Binária	No, Yes
Tempo de Contrato	Inteira	Intervalo de 1 a 72 (meses)
Serviço de Telefone	Binária	No, Yes
Múltiplas Linhas	Categórica	No, Yes
Serviço de Internet	Categórica	DSL, Fiber optic, No
Segurança Online	Binária	No, Yes
Backup Online	Binária	Yes, No
Proteção de Dispositivo	Binária	No, Yes
Suporte Técnico	Binária	No, Yes
Streaming de TV	Binária	No, Yes
Streaming de Filmes	Binária	No, Yes
Contrato	Categórica	Month-to-month, One year,Two year
Faturamento Eletrônico	Binária	Yes, No
Método de Pagamento	Categórica	Electronic check,Mailed check, Bank transfer (automatic), Credit card (automatic)
Cancelamento	Binária	No, Yes
Cobrança Mensal	Decimal	Valor em $
Cobrança Total	Decimal	Valor em $
Cancelamento	Binária	No, Yes

Figura 1 – Tabela Descrição das variaveis, tipo e categoria

Fonte: Elaborado pelo autor

Considerando a linha de tempo de cada cliente, temos que cada cliente permanece um periodo com a empresa e deixa de fazer negocio, observe a Figura 2.

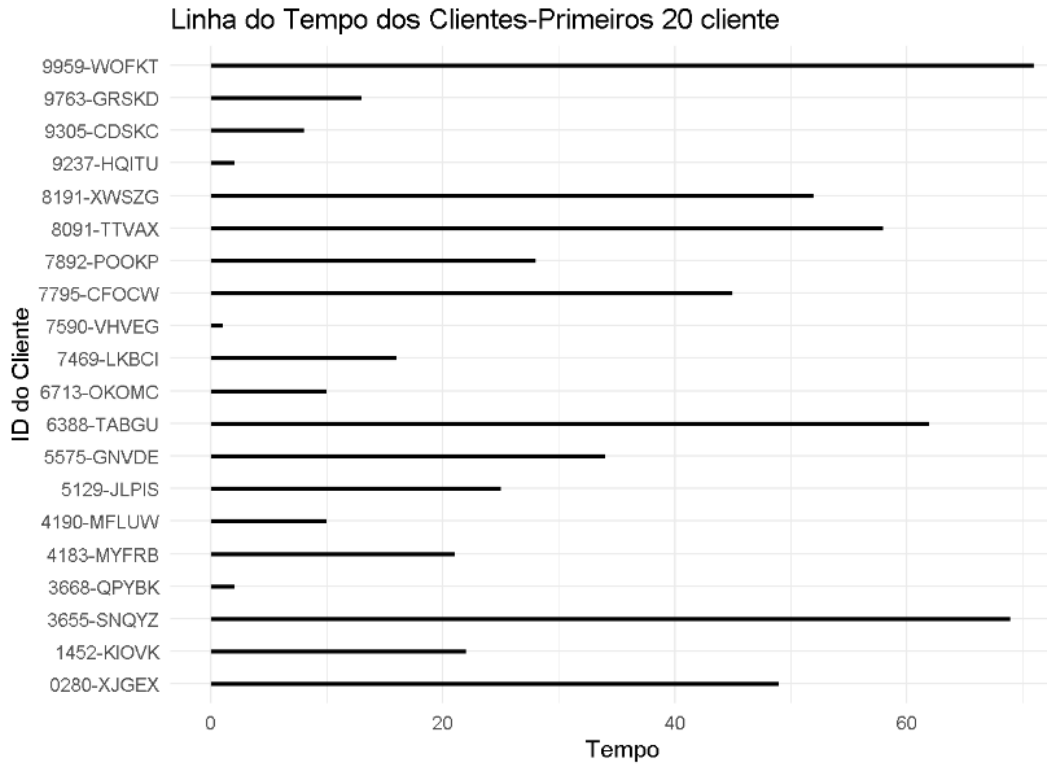


Figura 2 – Tabela Descrição das variaveis, tipo e categoria

Fonte: Elaborado pelo autor

3.1 Analise Descritiva

Após um tratamento no dados(codigo),realizei uma analise descritava inicial, visualizando a quantidade de cancelamento para cada coluna(Figura 3 e Figura 4). Alguns resultados que podemos observar dos graficos(Figura 3 e Figura 4) é que :

- cancelamento é maior para os clientes que sao idosos;
- Clientes com parceiros e dependentes a taxa é menor;
- Cliente com Serviço e múltiplas linhas são proporcionais;
- Cancelamento para sexo masculino e feminino são próximos;
- Temos de Cancelamento que cerca de 26 % da base de dados.

Outra observações encontradas relevantes para escolha de variaveis a serem colocadas no modelo, presentes na Figura 5 e Figura 6, são:

- Cliente que mais saíram no ultimo mes sao aqueles que nao possuem Segurança Online, Backup Online, Proteção de Dispositivo e Suporte Técnico;
- Para o tipo de serviço em que o cliente usa Fibra Ótica a taxa é maior;
- Clientes que possuem o contrato mensal cancelaram mais do que os outros tipos de contratos;
- Clientes que recebem o boleto de pagamento online tem a taxa maior do que os que não recebem;
- Cliente que usam o debito em conta tem a maior taxa;
- O tempo mediano para quem cancelou no ultimo mes é de 10 meses;
- Clientes que cancelaram possuem a mediana de pagamento de 75 mensalmente; Os clientes estão concentrado em duas faixa de tempo de contrato: até 1 ano e 5 a 6 anos;
- A maioria dos clientes estão com o contrato mensal, seguido por 2 anos e 1 ano.

Essas variáveis com destaque na analise descritiva são possíveis variáveis importante para predição de churn, a serem incluidas na modelagem.

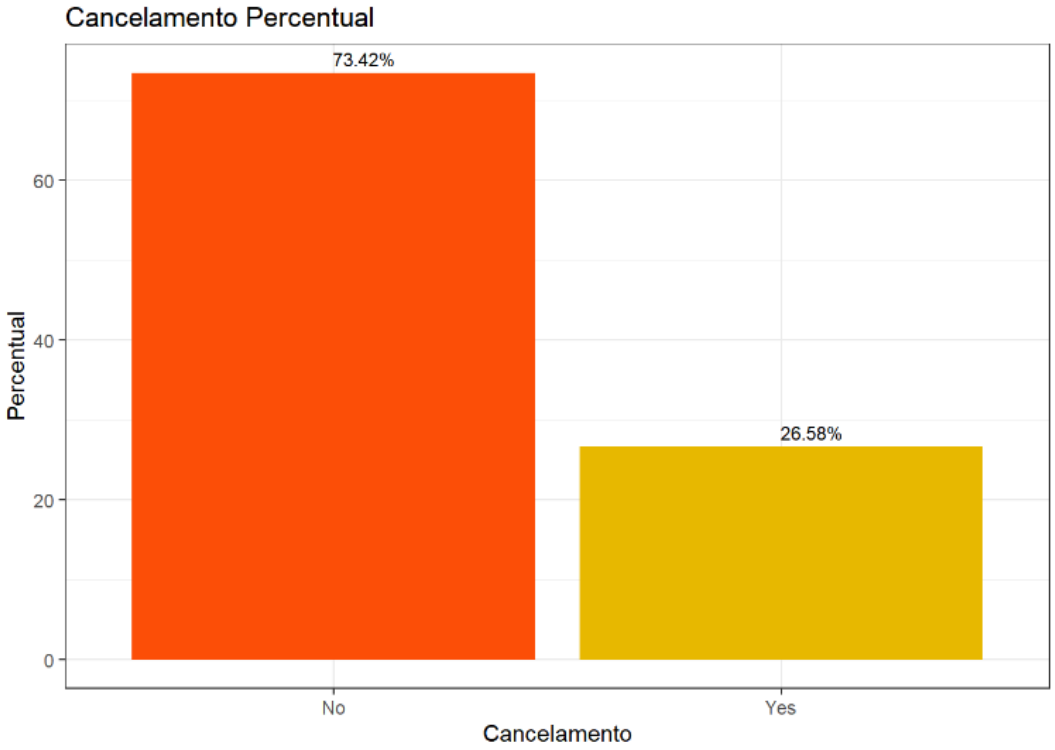


Figura 3 – Tabela Descrição das variaveis, tipo e categoria

Fonte: Elaborado pelo autor

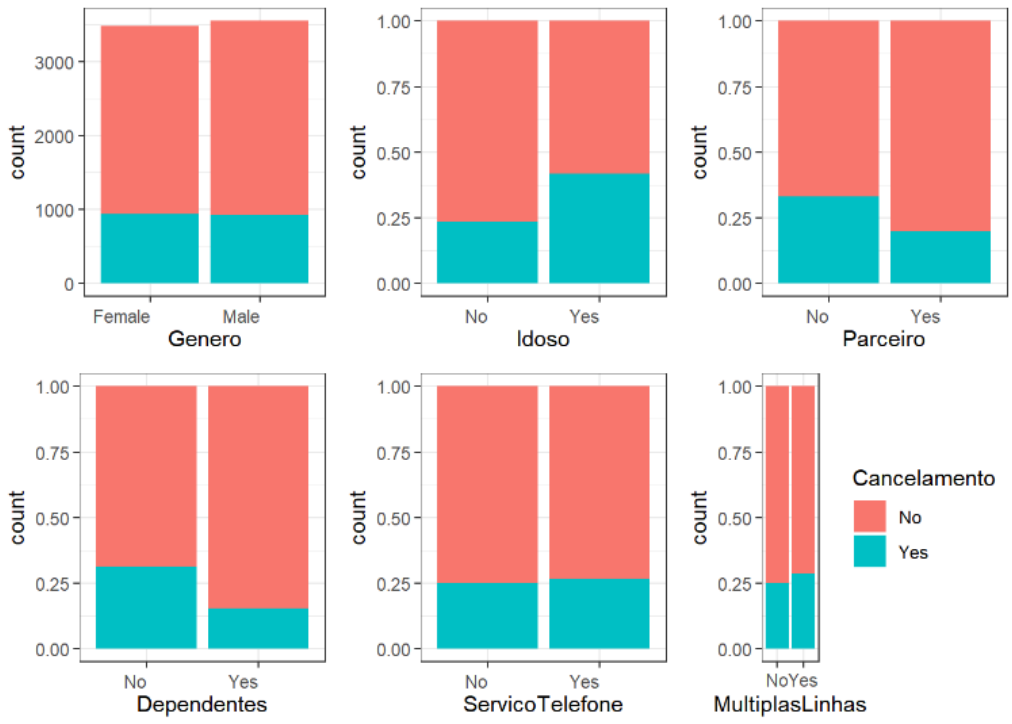


Figura 4 – Tabela Descrição das variaveis, tipo e categoria

Fonte: Elaborado pelo autor

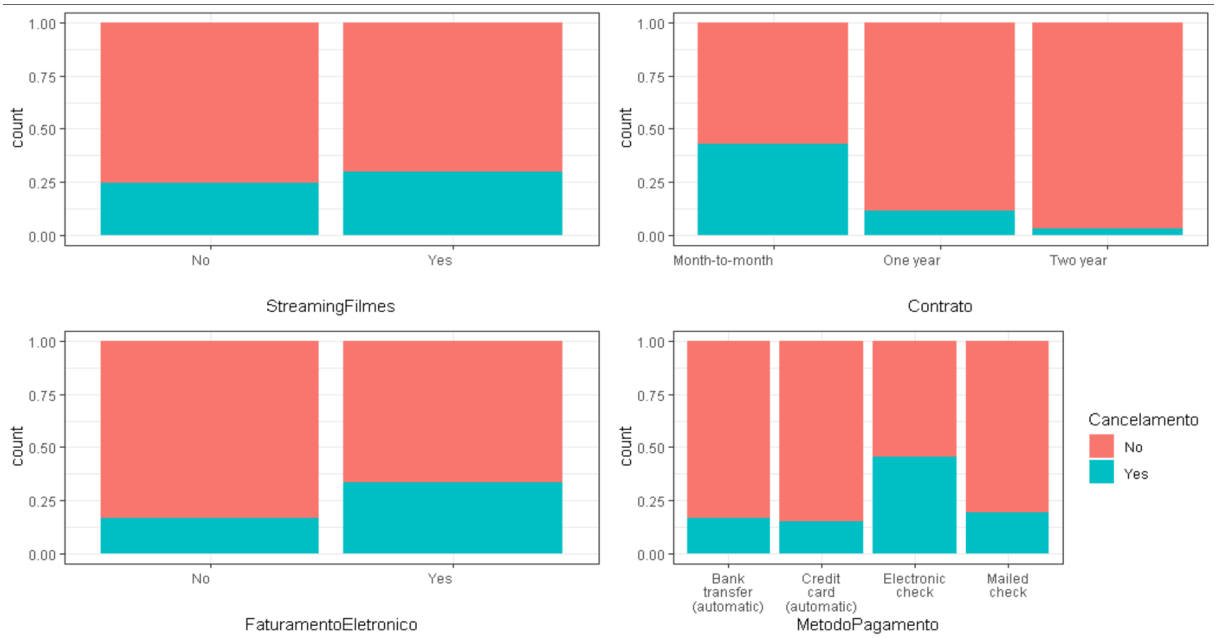


Figura 5 – Tabela Descrição das variaveis, tipo e categoria

Fonte: Elaborado pelo autor

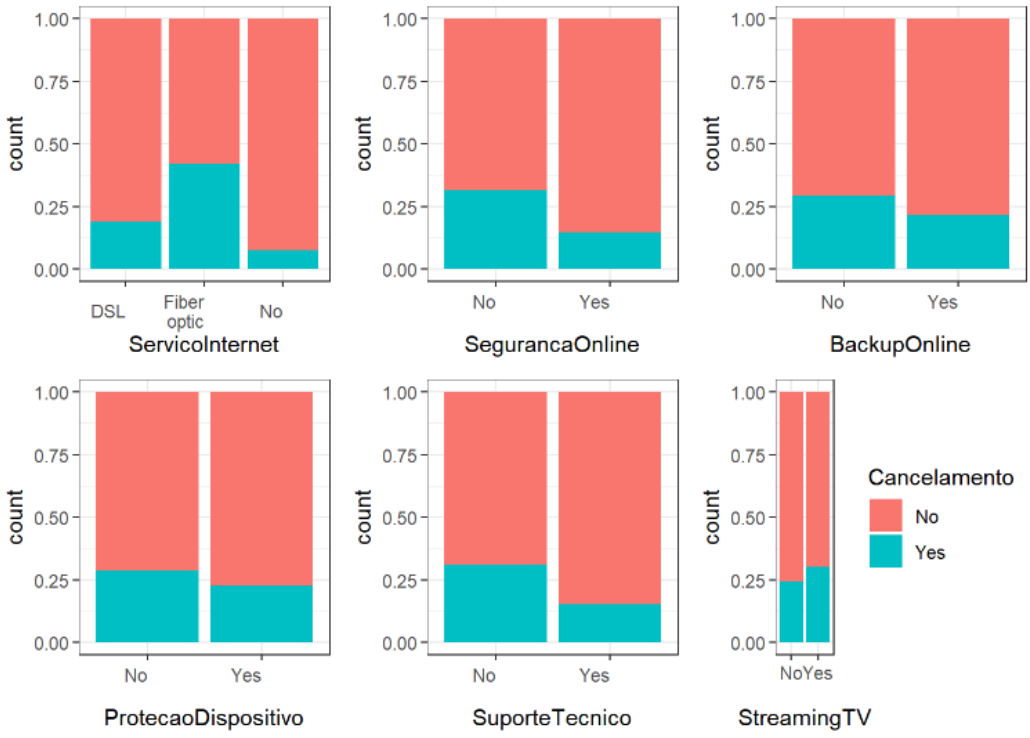


Figura 6 – Tabela Descrição das variaveis, tipo e categoria

Fonte: Elaborado pelo autor

3.2 Criação do Modelo

Inicialmente, separamos os dados em conjuntos de treino e teste, sendo 70% para treino e 30% para teste, para avaliar o desempenho do modelo. Em seguida, criamos um modelo geral, incluindo todas as variáveis disponíveis como potenciais preditores do cancelamento. Para refinar o modelo, utilizamos o Critério de Informação de Akaike (AIC) com a função stepAIC, que seleciona automaticamente as variáveis mais relevantes, removendo as que pouco contribuem para o ajuste. Esse processo resulta em um modelo mais enxuto e eficiente, mantendo as variáveis essenciais. Observe a Figura 7 abaixo:

Figura 7 – Tabela Descrição das variaveis, tipo e categoria

Fonte: Elaborado pelo autor

Por fim chegamos ao modelo 2(Figura 8 e Figura 9) que mantém apenas as variáveis mais significativas, proporcionando um equilíbrio entre simplicidade e precisão preditiva, as variáveis selecionadas para prever o churn dos clientes incluem:

- **Tempo:** Refere-se ao período que o cliente permanece com o serviço. Observamos que clientes com maior tempo de contrato tendem a ter uma menor probabilidade de cancelamento, sugerindo uma fidelidade maior com o passar do tempo.

- **Cobrança Mensal:** Representa o valor cobrado mensalmente. Clientes com cobranças mensais mais altas podem estar mais propensos a cancelar, especialmente se perceberem o custo como elevado em relação ao benefício.

- **Idoso:** Indica se o cliente é idoso ou não. Esse fator pode refletir diferenças de comportamento, pois clientes idosos podem ter maior fidelidade e padrões de uso distintos em comparação com clientes mais jovens.

- **Características dos Serviços Contratados:** *ServicoTelefone, MultiplasLinhas, ServicoInternet, BackupOnline, ProtecaoDispositivo, StreamingTV, StreamingFilmes*: Esses serviços refletem o perfil de uso dos clientes. A ausência de certos serviços, como Proteção de Dispositivo e Backup Online, mostrou-se associada a uma maior taxa de cancelamento, indicando que a falta de suporte e segurança pode impactar a satisfação.

- **Contrato:** Refere-se ao tipo de contrato do cliente (mensal ou anual). Clientes com contratos mensais apresentaram uma taxa de churn maior, o que sugere que contratos de longo prazo podem ser uma estratégia para reduzir o cancelamento.

- **Faturamento Eletrônico e Método de Pagamento:** Essas variáveis representam as formas de faturamento e o método de pagamento. Clientes que optam por pagamento online ou em débito automático exibem diferentes perfis de cancelamento, possivelmente devido à conveniência ou comprometimento financeiro.

- **Categoria_Tempo:** Uma categorização do tempo de contrato, que ajuda a segmentar os clientes de acordo com a duração de sua relação com a empresa, facilitando a análise de padrões de cancelamento por período de fidelização.

Esse modelo final identifica as variáveis mais impactantes no churn, oferecendo insights claros sobre os fatores que levam clientes a cancelar o serviço.

Resumo Modelo 2

```
summary(model_2)

##
## Call:
## glm(formula = Cancelamento ~ Tempo + CobrancaMensal + Idoso +
##     ServicoTelefone + MultiplasLinhas + ServicoInternet + BackupOnline +
##     ProtecaoDispositivo + StreamingTV + StreamingFilmes + Contrato +
##     FaturamentoEletronico + MetodoPagamento + Categoria_Tempo,
##     family = binomial(link = "logit"), data = train)
```

Figura 8 – Tabela Descrição das variaveis, tipo e categoria

Fonte: Elaborado pelo autor

Variável	Estimativa	Pr(>|z|)
(Intercepto)	-0,46557547	0,00661 **
Tempo	-0,03670718	< 2e-16 ***
Idoso (Sim)	0,40057739	5,06e-05 ***
Serviço Telefone (Sim)	-0,39326101	0,00838 **
Múltiplas Linhas (Sim)	0,24896363	0,00814 **
Serviço Internet (Fibra Óptica)	0,94164496	< 2e-16 ***
Serviço Internet (Sem Internet)	-0,74263758	2,97e-06 ***
Backup Online (Sim)	-0,11692944	0,19595
Proteção de Dispositivo (Sim)	-0,01721924	0,8534
Streaming TV (Sim)	0,24968637	0,00867 **
Streaming Filmes (Sim)	0,21495696	0,02278 *
Contrato (1 ano)	-0,73288236	5,00e-09 ***
Contrato (2 anos)	-1,68905812	1,41e-14 ***
Faturamento Eletrônico (Sim)	0,38170335	1,63e-05 ***
Método de Pagamento (Cartão Crédito Automático)	-0,03091178	0,8203
Método de Pagamento (Cheque Eletrônico)	0,33622045	0,00276 **
Método de Pagamento (Cheque Enviado)	0,0195429	0,88415

Figura 9 – Tabela Descrição das variaveis, tipo e categoria

Fonte: Elaborado pelo autor

3.3 Aperfeiçoando o Modelo

Para melhorar e validar o modelo, foi realizada uma análise de resíduos e verificação de multicolinearidade, um fenômeno onde variáveis preditoras estão altamente correlacionadas, o que prejudica a precisão das estimativas. Para isso, utilizamos o VIF (Variance Inflation Factor), que indica o grau de correlação de cada variável com as demais no modelo. Valores de VIF acima de 5 ou 10 sugerem alta multicolinearidade.Ao analisar as variáveis, identifiquei que Tempo, Cobrança Mensal e Cobrança Total tinham VIFs altos, indicando redundância. Assim, removi Cobrança Mensal e Cobrança Total, mantendo apenas Tempo, o que satifez a não multicolinearidade. Observe a estrutura (Figura 10.

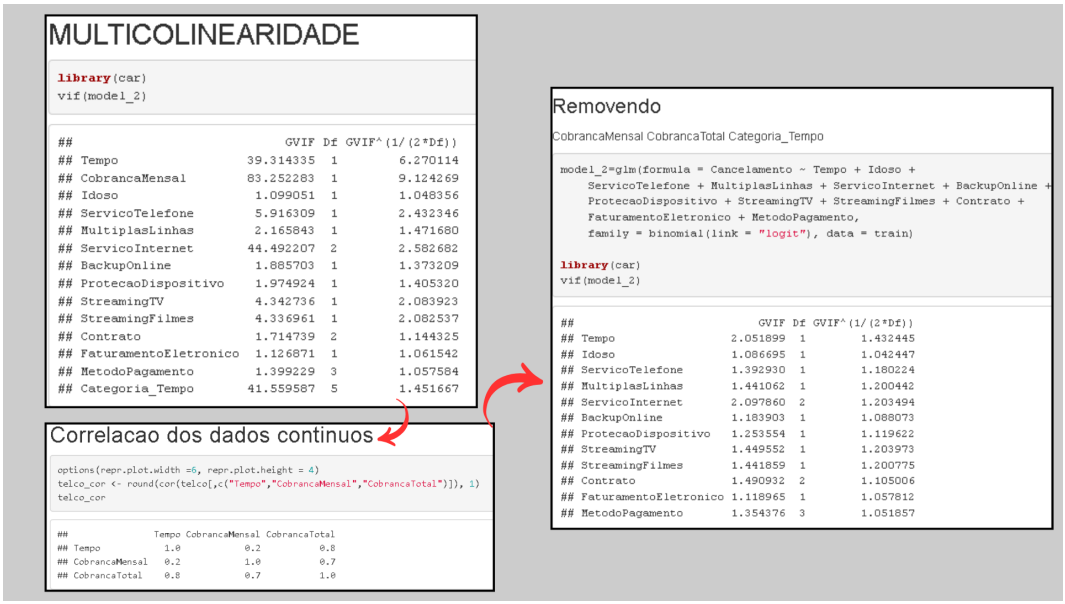


Figura 10 – Tabela Descrição das variaveis, tipo e categoria

Fonte: Elaborado pelo autor

Para dar continuidade ao aperfeiçoamento do modelo, realizamos uma análise de resíduos para verificar o ajuste e a adequação do modelo de regressão logística. Os gráficos apresentados oferecem uma visão detalhada de possíveis problemas de ajuste e pontos influentes. Observe Figura 11

- Gráficos de Resíduos de Deviance, Pearson e Resíduos Padronizados: Estes gráficos mostram a distribuição dos resíduos. A ausência de um padrão claro sugere que o modelo está capturando bem as relações nas variáveis, mas é importante observar que alguns pontos estão distantes da linha central, indicando outliers potenciais.

- Distância de Cook: Esse gráfico ajuda a identificar observações com grande influência no ajuste do modelo. Algumas observações têm valores de distância de Cook elevados, o que sugere que elas podem estar influenciando excessivamente os coeficientes do modelo.

- Valores de Leverage (hii): Este gráfico identifica pontos com alto leverage, ou seja, observações que se afastam dos demais e podem ter influência no ajuste. A maioria dos dados está dentro dos limites, mas alguns pontos com valores elevados podem estar afetando o ajuste.

- Q-Q Plot dos Resíduos: O gráfico Q-Q dos resíduos avalia a normalidade dos resíduos padronizados. A proximidade dos pontos com a linha indica que os resíduos estão aproximadamente normais, o que é um bom sinal para a adequação do modelo.

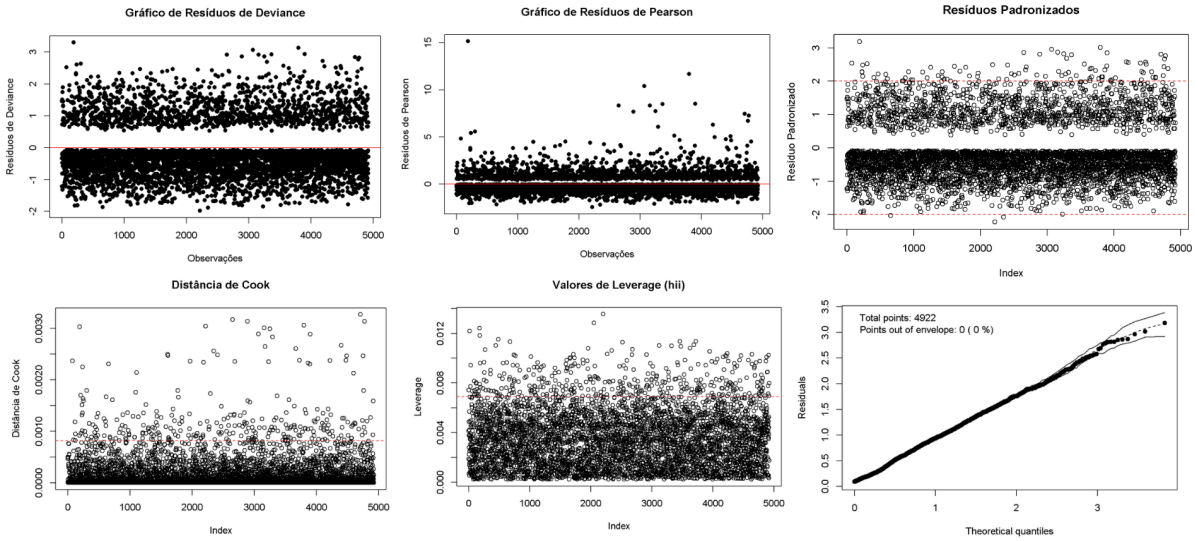


Figura 11 – Tabela Descrição das variaveis, tipo e categoria

Fonte: Elaborado pelo autor

Esses gráficos, em conjunto, nos permitem avaliar a estabilidade e a robustez do modelo. Eles indicam que, em geral, o ajuste está adequado, mas que existem algumas observações influentes e outliers que poderiam ser investigados mais a fundo para melhorar a precisão e robustez do modelo.

Para dar continuidade na validação do modelo, podemos utilizar o teste de Hosmer-Lemeshow, que é um teste estatístico de bondade de ajuste aplicado em modelos de regressão logística, iremos utilizar para verificar se os pontos da analise anterior estão atrapalhando ou não a modelagem, esse teste possui como hipóteses:

- H0 (hipótese nula): Não há diferença significativa entre os valores previstos pelo modelo e os valores observados. Ou seja, o modelo se ajusta bem aos dados.

- H1 (hipótese alternativa): Existe uma diferença significativa entre os valores previstos e observados. Portanto, o modelo não se ajusta bem aos dados.

```
#install.packages("glmtoolbox")

library(glmtoolbox)
hltest(model_2)
```

```
##
##     The Hosmer-Lemeshow goodness-of-fit test
##
##  Group Size Observed   Expected
##      1  493        3   3.531079
##      2  492       11  10.052221
##      3  492       21  22.105637
##      4  492       48  44.635475
##      5  492       81  75.215443
##      6  492      110 113.364747
##      7  492      167 163.482258
##      8  493      203 224.019817
##      9  492      295 291.886191
##     10  492      369 359.707133
##
##           Statistic =  5.86621
## degrees of freedom =  8
##             p-value =  0.66222
```

Figura 12 – Tabela Descrição das variaveis, tipo e categoria

Fonte: Elaborado pelo autor

O valor p do teste de Hosmer-Lemeshow (Figura 12) é maior que 0,05 (0.6622), o que significa que não há evidências estatísticas para rejeitar a hipótese nula de que o modelo se ajusta bem aos dados. Em outras palavras, o modelo apresenta um ajuste aceitável em relação à variável resposta (Figura 12).

3.4 Métricas de Desempenho

Finalizado a validação do modelo podemos agora avaliar a capacidade preditiva do modelo, o qual é feito utilizando os dados de teste, anteriomente usamos os dados de treino para treinar o modelo. Utilizando a curva Roc e o AUC (2.0.6) iremos definir um ponto de corte (ou cutoff) o qual refere-se ao valor limite da probabilidade a partir do qual uma observação será classificada em uma das duas classes. Para um modelo de regressão logística, como o utilizado aqui, a saída são probabilidades que indicam a chance deu um cliente cancelar o serviço. Normalmente, se a probabilidade calculada for maior que o ponto de corte, o cliente será classificado como "Sim"(irá cancelar). Caso contrário, será classificado como "Não"(não irá cancelar).

```
final_model=model_2
actual_churn <- validation$Cancelamento

pred <- predict(final_model, type = "response", newdata = validation[,-24])
```

```
validation$prob <- pred

# Using probability cutoff of 50%.

cutoff_churn <- factor(ifelse(pred >=0.50, "Yes", "No"))
conf_final <- caret::confusionMatrix(cutoff_churn, actual_churn, positive = "Yes")
accuracy <- conf_final$overall[1]
sensitivity <- conf_final$byClass[1]
specificity <- conf_final$byClass[2]
```

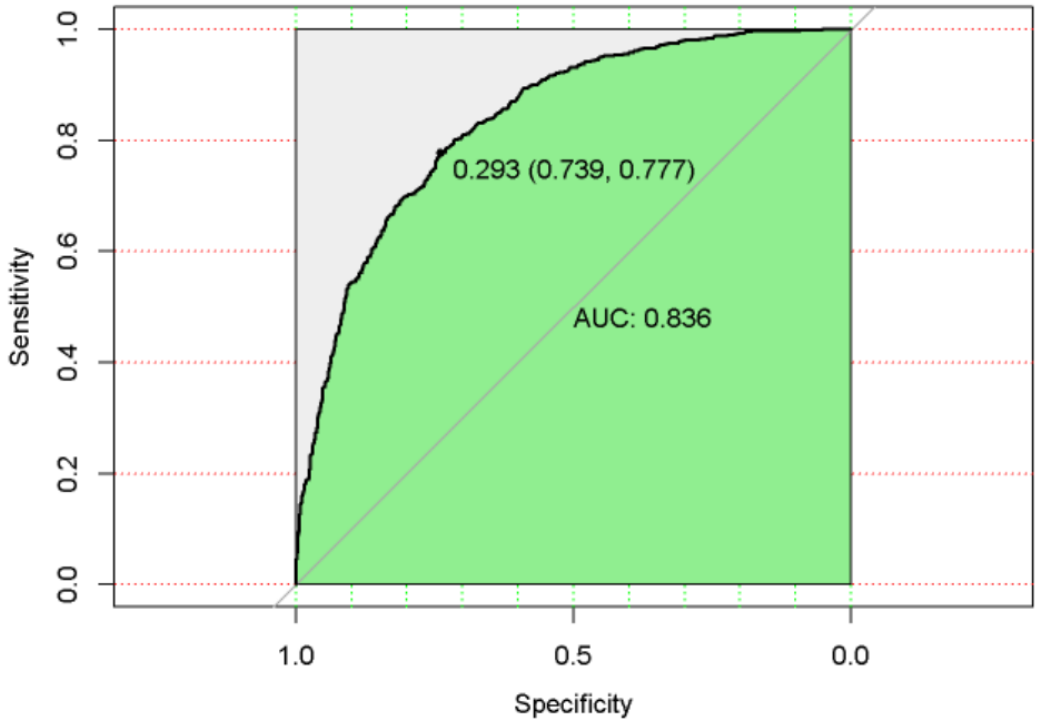
Métrica	Valor
Accuracy	79,91%
Sensitivity	55,08%
Specificity	88,90%

Figura 13 – Tabela Descrição das variaveis, tipo e categoria

Fonte: Elaborado pelo autor

Para a cutoff na Figura 13 utilizando um ponto de corte de 50% (ou 0.5), o que significa que, se a probabilidade de churn prevista para um cliente for superior a 50%, o modelo o classificará como "Sim"para cancelamento; caso contrário, será classificado como "Não". Considerando o ponto de corte 0.5 obtivemos informações referente desempenho do modelo, explicados em 2.0.6. Como podemos ver na Figura 13, quando usamos um corte de 0,50, obtemos uma boa acuracia e especificidade, mas a sensibilidade é muito menor.

Portanto, precisamos encontrar o corte de probabilidade ideal que dará a máxima precisão, sensibilidade e especificidade, para isso iremos utilizar a Curva Roc para encontrar o melhor ponto de corte, observe na Figura 14 o ponto ideal encontrado é 0,293 para o modelo final, onde as curvas de precisão, especificidade e sensibilidade se encontram e maximiza ambas.



Métrica	Valor
Accuracy	74,88%
Sensitivity	77,54%
Specificity	73,92%

Figura 14 – Tabela Descrição das variaveis, tipo e categoria

Fonte: Elaborado pelo autor

Interpretando as métricas de desempenho (2.0.6) encontradas com o ponto de corte 0,293 (Figura 14), temos que:

- Accuracy (Acurácia): 74,88O modelo faz previsões corretas em aproximadamente 75% dos casos.
- Sensitivity (Sensibilidade): 77,54% O modelo identifica corretamente 78% dos clientes que cancelaram. Essa métrica é importante para campanhas de retenção, pois indica que o modelo detecta a maioria dos cancelamentos reais.
- Specificity (Especificidade): 73,92% O modelo identifica corretamente 74% dos clientes que não cancelaram, embora ainda existam alguns falsos positivos (clientes que o modelo prevê como cancelamento, mas que realmente não cancelam).

3.5 Interpretação dos coeficientes

Por fim, obtivemos do modelo os valores do coeficientes, observe os valores na Figura 15.

Coeficientes	Valores_Coeficientes
Intercept	-0,46557547
Tempo	-0,03670718
IdosoYes	0,40057739
ServicoTelefoneYes	-0,39326101
MultiplasLinhasYes	0,24896363
ServicoInternetFiberOptic	0,94164496
ServicoInternetNo	-0,74263758
BackupOnlineYes	-0,11692944
ProtecaoDispositivoYes	-0,01721924
StreamingTVYes	0,24968637
StreamingFilmesYes	0,21495696
ContratoOneYear	-0,73288236
ContratoTwoYear	-1,68905812
FaturamentoEletronicoYes	0,38170335
MetodoPagamentoCreditCardAutomatic	-0,03091178
MetodoPagamentoElectronicCheck	0,33622045
MetodoPagamentoMailedCheck	0,01954290

Figura 15 – Tabela Descrição das variaveis, tipo e categoria

Fonte: Elaborado pelo autor

Considerendo o que foi apresentado em 2.0.7, temos que realizar a interpretação das *odds ratio*, observe na Figura 16 que temos a variavel com a resposta a qual ele possui peso, a *odds ratio* e a interpretação, que é calculado com base na odd ratio, a formula de como obter a *odds ratio* e interpretação também estão presentes. Irei apenas interpretar alguns valores, o mesmo padrão segue para as outras variaveis (Figura 16):

- ServicoInternet (Fiber optic) (2,564): Clientes com serviço de internet via fibra óptica têm uma chance de cancelamento aproximadamente 156,42% maior em comparação aos clientes que utilizam a categoria de referência (DSL), mantendo todas as outras variáveis constantes.

- Tempo (0,964): Para cada unidade de aumento em Tempo (em meses), a chance de cancelamento diminui aproximadamente 3,60%, mantendo todas as outras variáveis constantes.
- Contrato (Two year) (0,185): Clientes com contrato de dois anos têm uma chance de cancelamento aproximadamente 81,53% menor em relação aos clientes com contrato mensal, mantendo todas as outras variáveis constantes.

Variável	Odds Ratio	Interpretação
(Intercept)	0,628	-37,22%
ServicoInternet (Fiber optic)	2,564	156,42%
Idoso (Yes)	1,493	49,27%
FaturamentoEletronico (Yes)	1,465	46,48%
MetodoPagamento (Electronic check)	1,400	39,96%
StreamingTV (Yes)	1,284	28,36%
MultiplasLinhas (Yes)	1,283	28,27%
StreamingFilmes (Yes)	1,240	23,98%
MetodoPagamento (Mailed check)	1,020	1,97%
ProtecaoDispositivo (Yes)	0,983	-1,71%
MetodoPagamento (Credit card (automatic))	0,970	-3,04%
Tempo	0,964	-3,60%
BackupOnline (Yes)	0,890	-11,04%
ServicoTelefone (Yes)	0,675	-32,51%
Contrato (One year)	0,481	-51,95%
ServicoInternet (No)	0,476	-52,41%
Contrato (Two year)	0,185	-81,53%

Figura 16 – Tabela Descrição das variaveis, tipo e categoria

Fonte: Elaborado pelo autor

Temos os dados de IC para *odds ratio*, permitindo observar as variaveis de risco(Vermelho), protetoras(Azul) e não significativa (cinza) indicadas pelos pontos no grafico Figura 17, além disso o grafico contem o valor do *odds ratio* e o valor da Interpretação.

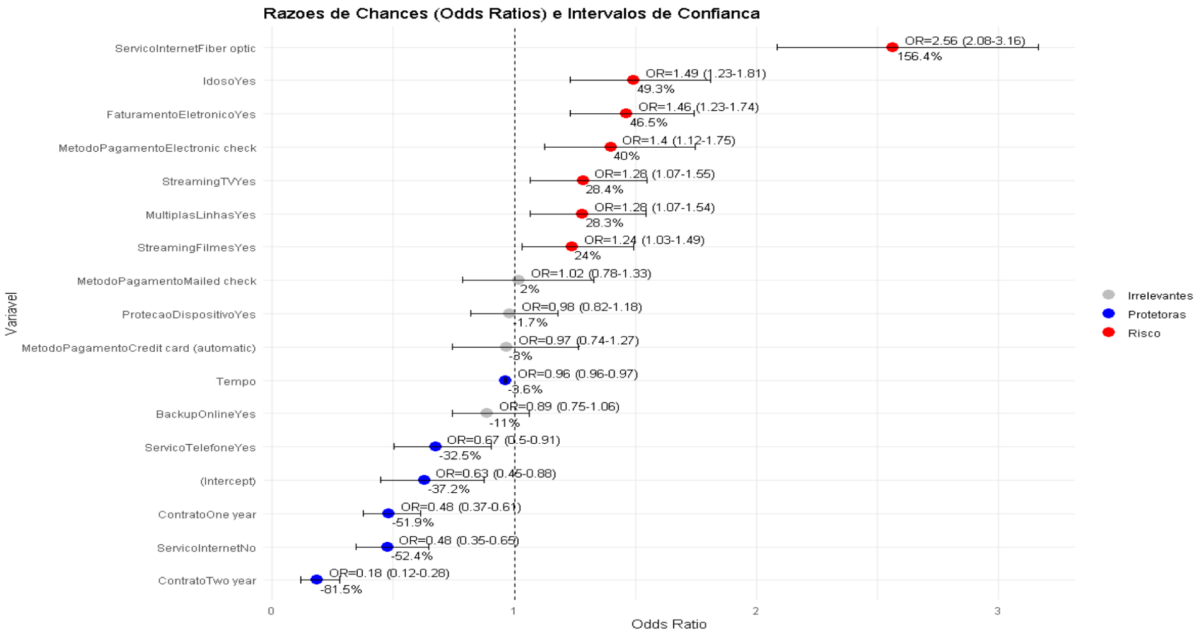


Figura 17 – Tabela Descrição das variaveis, tipo e categoria

Fonte: Elaborado pelo autor

3.6 Tomadas de descisões

Como conclusão, apresentamos as possíveis ações estratégicas recomendadas para a empresa com base nos insights obtidos através da análise e interpretação dos resultados do modelo. Esses insights são fundamentais para direcionar esforços e recursos de maneira eficiente, visando à retenção de clientes e à redução da taxa de churn. Cada um dos pontos a seguir destaca oportunidades para melhorar a satisfação dos clientes, aumentar o comprometimento com contratos de longo prazo e oferecer opções de faturamento que atendam melhor às preferências dos consumidores.

Os principais pontos de ação são:

- **Foco em Clientes com Fibra Óptica**
  - Clientes com fibra óptica têm 156,42% mais chance de cancelar em comparação com os clientes com DSL.
  - **Ação**: Realizar pesquisas de satisfação para entender as razões de insatisfação e oferecer pacotes promocionais de longo prazo.
- **Estimular Clientes a Migrar para Contratos de Longo Prazo**
  - **Intervenção**: Clientes com contrato de um ano têm uma chance de cancelamento 52% menor, e com contrato de dois anos, 81,53% menor.
- **Oferecer Opções Alternativas de Faturamento**
  - **Intervenção**: Clientes que optam pelo faturamento eletrônico apresentam 40% mais chance de cancelar.
  - **Ação**: Oferecer notificações ou lembretes antes de emitir a cobrança eletrônica e incentivar métodos alternativos.

Referências

CHAR, B. **Telco Customer Churn Dataset**. 2018. Acesso em: 14 nov. 2024. Disponível em: <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>. Citado na página 26.

CHURN-IBM. **IBM Telcom Churn**. <https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113>. Acesso em: 13 de setembro de 2024. Citado 2 vezes nas páginas 8 e 26.

IRINEU, W. **Prevendo o Churn de Clientes em Telecom com Regressão Logística no R**. 2023. Acesso em: 14 nov. 2024. Disponível em: <https://rpubs.com/WilliamIrineu/1244905----https://github.com/WilliamIrineu/RegressaoLogistica>. Citado na página 26.

MATTISON, R. **Telecom Churn Management: The Golden Opportunity**. São Paulo: XiT Press Oakwood Hills, Illinois, 2005. Citado na página 3.

ANEXO A – Conjunto de dados e Código

Qualquer duvida referente ao relatorio e codigo:

Para obter os codigo acesse o (IRINEU, 2023), aqui possui o link do Rpubs e também do github (aquivo RMD).

Para o conjunto de dados acesse (CHAR, 2018) e (CHURN-IBM,).