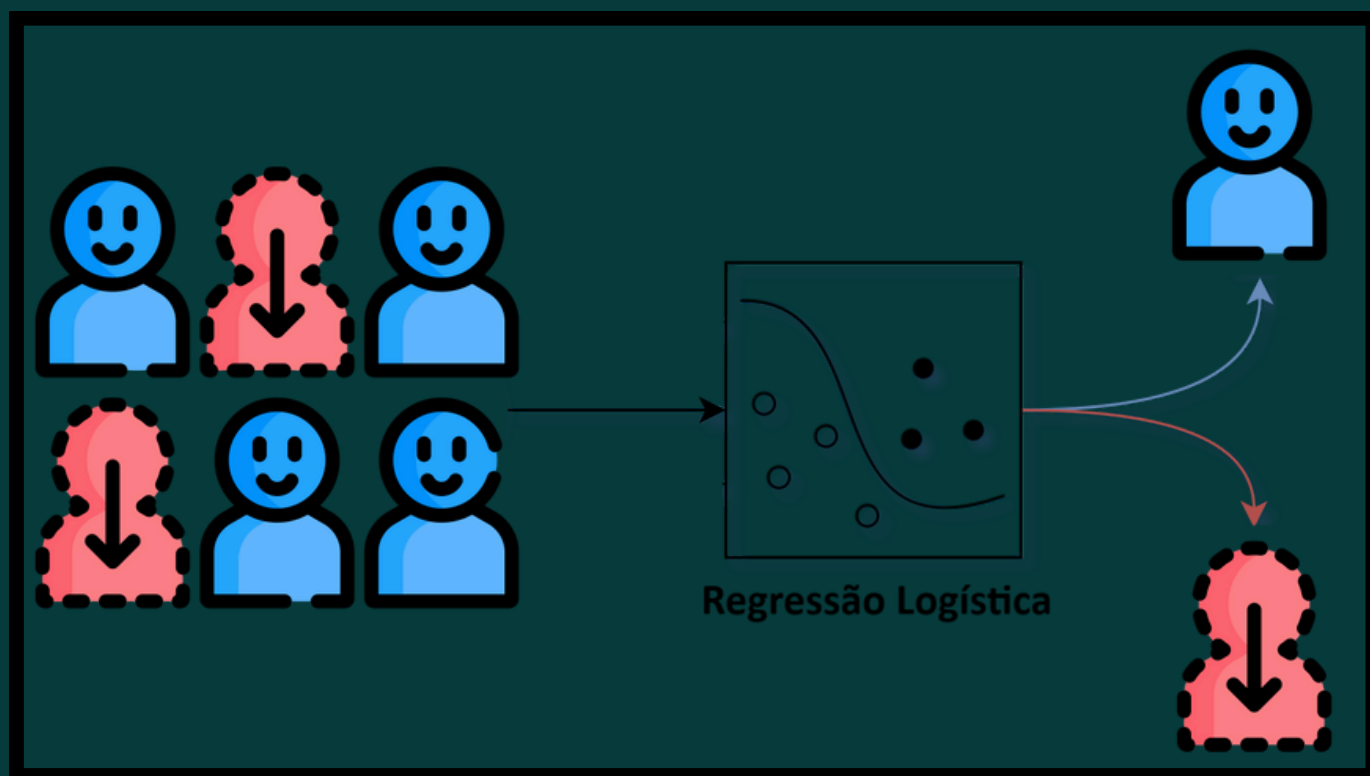




**WILLIAM  
IRINEU**

# PREVENDO CANCELAMENTO DE CLIENTES



# SUMÁRIO

---

**01**

**Contexto Inicial**

**02**

**Descrição dos Dados**

**03**

**Análise Descritiva**

**04**

**Criação do Modelo**

**05**

**Validação do Modelo**

**06**

**Interpretação do Modelo**

**07**

**Aplicação do modelo**

**08**

**Sugestões para Tomadas de  
Decisão da Empresa**

# 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 em 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 avanços, temos que perder clientes implica em perda de receita, custos adicionais com a reaquisição e um impacto negativo na reputação da marca.

A escolha do tema baseia-se em minha experiência de três anos em análise de dados, atuando em uma empresa de telecomunicações, onde analisorotatividade 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.

# SOBRE OS DADOS E METODOS

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 análise descritiva dos dados pra obter outros resultados que auxilie a tomada de decisão.

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.**

O pré-processamento dos dados envolveu:

- Substituição de "No internet service" e "No phone service" por "No"
- 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.

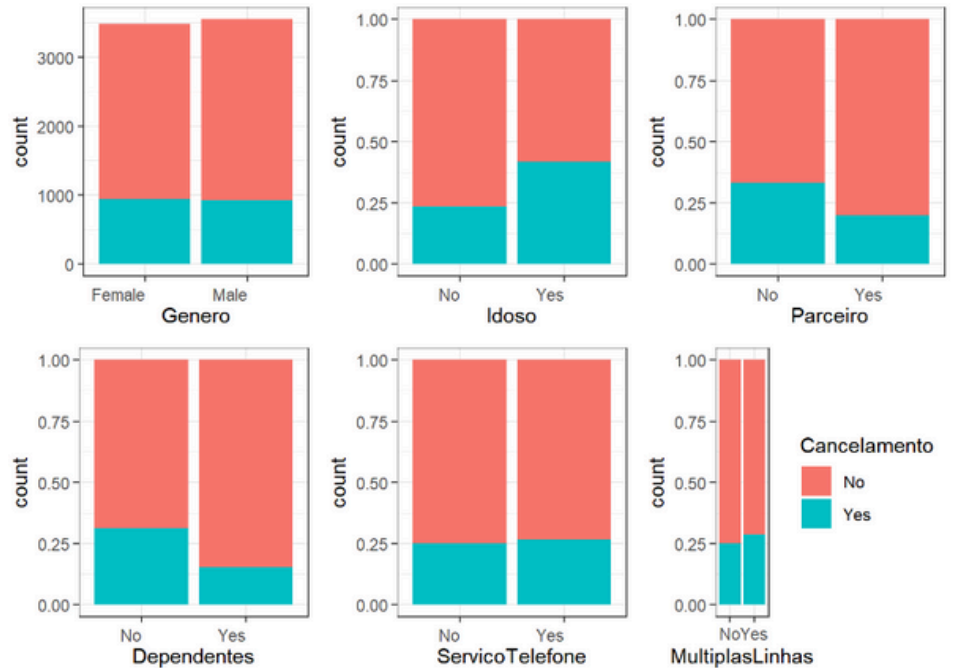
Descrição das Variáveis	Tipo	Categorias
ID do Cliente	Texto	Código Único
Gênero	Categórica	Female, Male
Idoso	Binária	No, Yes
Parceiro	Binária	Yes, No
Dependentes	Binária	No, Yes
Tempo de Contrato	Inteira	Intervalo de 1 a 72 (meses)
Serviço de Telefone	Binária	No, Yes
Múltiplas Linhas	Categórica	No, 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

# ANALISE DESCRITIVA

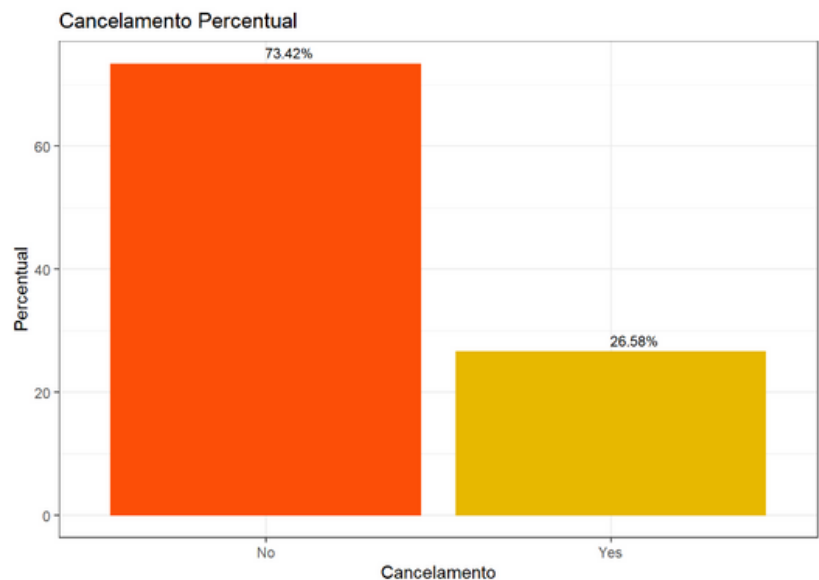
O pré-processamento dos dados envolveu:

- Substituição de "No internet service" e "No phone service" por "No"
- 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.

- Para o sexo masculino e feminino é quase igual
- Cancelamento é maior para os clientes que são idosos
- Clientes com parceiros e dependentes a taxa é menor
- Cliente com Serviço e múltiplas linhas são proporcionais.

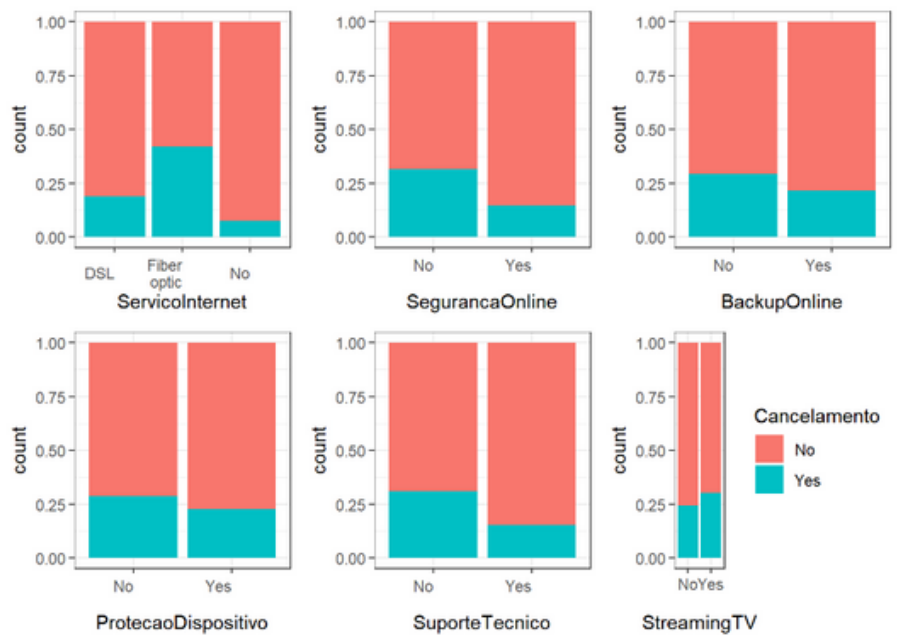


- Temos do Churn que cerca de 26% dos clientes deixaram de ser cliente no ultimo mês.

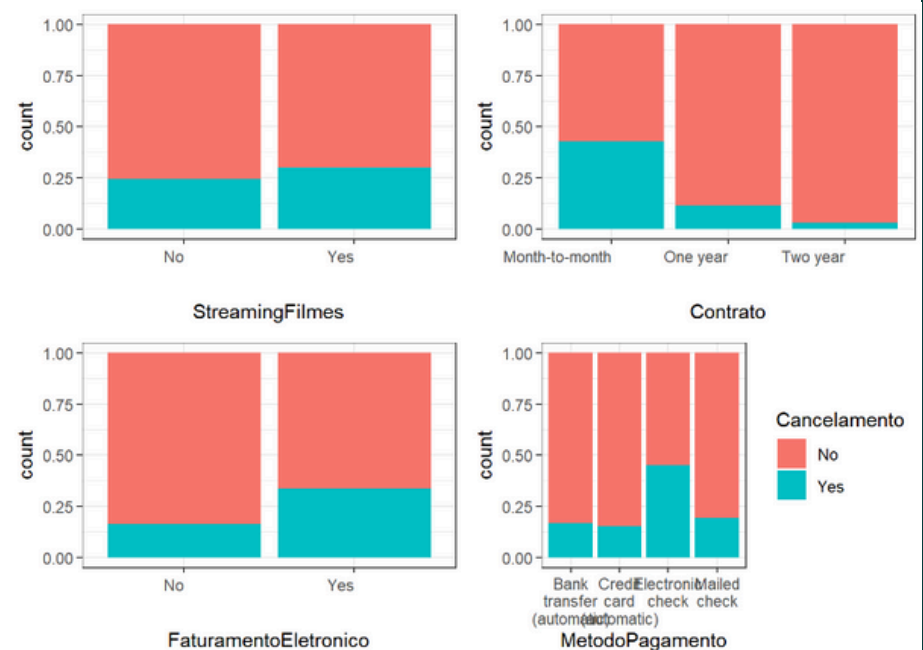


# ANALISE DESCRITIVA

- Para o tipo de serviço em que o cliente usa Fibra Ótica a taxa é maior
- Cliente que mais saíram no ultimo mes sao aqueles que nao possuem Segurança Online, Backup Online, Proteção de Dispositivo e Suporte Técnico



- 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



# 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 e utilizando a função de ligação logit, utilizamos a função logística (logit) para transformar o preditor linear em uma probabilidade.

Para refinar o modelo, utilizamos o AIC da 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.

## Treino e Teste


```
library(caTools)

set.seed(123)
indices <- sample.split(telco$Cancelamento, SplitRatio = 0.7)
train <- telco[indices, ]
validation <- telco[!indices, ]
```



## Modelo Geral

```
#Build the first model using all variables
model_1 = glm(Cancelamento ~ ., data = train, family = binomial(link = "logit"))
summary(model_1)
```



## Usando Step AIC

Com o step AIC iremos escolher as melhores variáveis para o melhor modelo

#Modelo 2

```
model_2 <- stepAIC(model_1, direction="both")
```

# MEDINDO O PROGRESSO

Por fim chegamos ao modelo 2 (Figura Abaixo ) que mantém apenas as variáveis mais significativas, esse modelo final identifica as variáveis mais impactantes no churn, oferecendo insights claros sobre os fatores que levam clientes a cancelar o serviço.

```
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)
```

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



# APERFEIÇOANDO O MODELO

Para melhorar e validar o modelo, foi realizada uma 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 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 satisfaz a não multicolinearidade. Observe a estrutura:

## MULTICOLINEARIDADE

```
library(car)
vif(model_2)
```

##	GVIF
## Tempo	39.314335
## CobrançaMensal	83.252283
## Idoso	1.099051
## ServicoTelefone	5.916309
## MultiplasLinhas	2.165843
## ServicoInternet	44.492207
## BackupOnline	1.885703
## ProtecaoDispositivo	1.974924
## StreamingTV	4.342736
## StreamingFilmes	4.336961
## Contrato	1.714739
## FaturamentoEletronico	1.126871
## MetodoPagamento	1.399229
## Categoria_Tempo	41.559587

## Removendo

CobrançaMensal CobrançaTotal Categoria\_Tempo

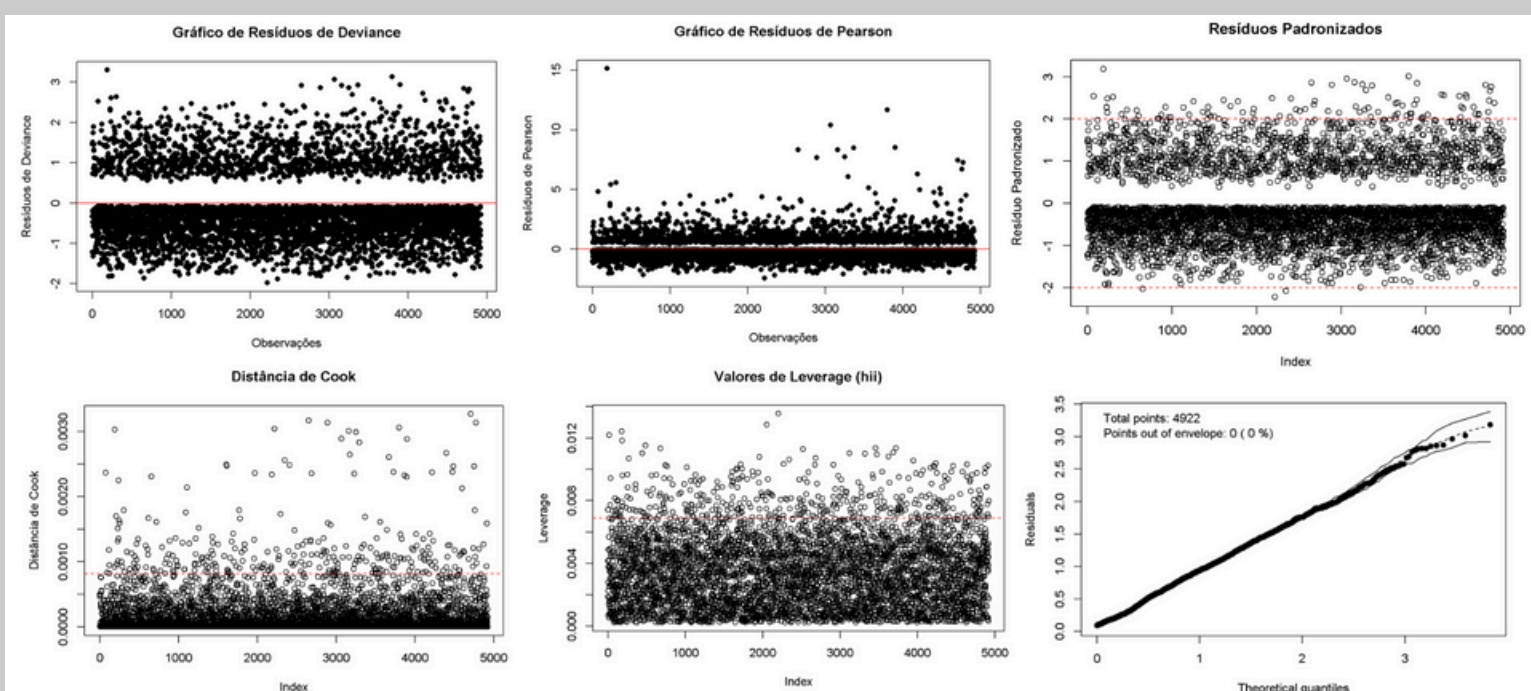
```
model_2=glm(formula = Cancelamento ~ Tempo + Idoso +
  ServicoTelefone + MultiplasLinhas + ServicoInternet + BackupOnline +
  ProtecaoDispositivo + StreamingTV + StreamingFilmes + Contrato +
  FaturamentoEletronico + MetodoPagamento,
  family = binomial(link = "logit"), data = train)
```

```
library(car)
vif(model_2)
```

##	GVIF	Df	GVIF^(1/(2*Df))
## Tempo	2.051899	1	1.432445
## Idoso	1.086695	1	1.042447
## ServicoTelefone	1.392930	1	1.180224
## MultiplasLinhas	1.441062	1	1.200442
## ServicoInternet	2.097860	2	1.203494
## BackupOnline	1.183903	1	1.088073
## ProtecaoDispositivo	1.253554	1	1.119622
## StreamingTV	1.449552	1	1.203973
## StreamingFilmes	1.441859	1	1.200775
## Contrato	1.490932	2	1.105006
## FaturamentoEletronico	1.118965	1	1.057812
## MetodoPagamento	1.354376	3	1.051857

# ANALISANDO OS RESÍDUOS

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.



O gráfico na direita inferior, avalia se os resíduos padronizados seguem a distribuição teórica esperada. Neste caso, todos os 4.922 pontos estão dentro do envelope, indicando uma excelente adequação do modelo aos dados. A ausência de pontos fora do envelope sugere que os resíduos seguem perfeitamente a distribuição proposta.

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 análise anterior estão atrapalhando ou não a modelagem, esse teste possui como hipóteses:

- Se o p-value maior que 0.05 o modelo se ajusta bem aos dados.
- Se o p-value menor que 0.05 o modelo não se ajusta bem aos dados.

## TESTE DE HOSMER-LEMESHOW

```
#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
```

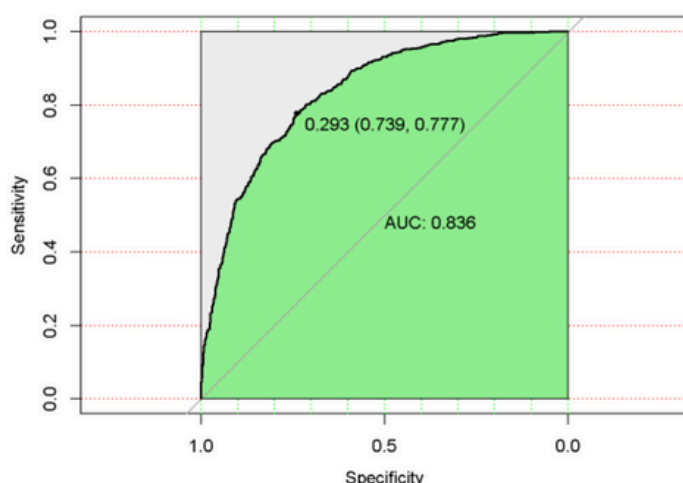
```
##           p-value =  0.66222
```

O p-value 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

Tendo encontrado o melhor modelo, é hora de avaliar se o modelo é bom ou não, com os dados de teste. a Curva ROC e a AUC (Área Sob a Curva), que medem a eficácia do modelo na distinção entre clientes que cancelam ou não.

Definimos ainda um ponto de corte (cutoff), que determina o limite de probabilidade para classificar um cliente como "Sim" (irá cancelar) ou "Não" (não irá cancelar), esse ponto é escolhido com a curva roc

## PONTO DE CORTE IDEAL E MÉTRICAS



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



Acerta 78%



Acerta 74%

As principais métricas de desempenho são acurácia, sensibilidade e especificidade, e que é interpretado da seguinte maneira:

- Acurácia (74,88%): O modelo prevê corretamente 75% dos casos.
- Sensibilidade (77,54%): O modelo identifica corretamente 78% dos clientes que cancelaram.
- Especificidade (73,92%): O modelo identifica corretamente 74% dos clientes que não cancelaram.

# 78%

**Antecipamos 78% dos cancelamentos**  
Essa inteligência permite direcionar os clientes em risco ao time de retenção, otimizando esforços e resultados.

# PRÓXIMOS PASSOS

Na figura abaixo, destacamos a interpretação das odds ratio, métricas fundamentais para entender a influência de cada variável sobre a probabilidade de cancelamento. A odds ratio quantifica o impacto de uma variável, indicando o quanto aumenta ou diminui a chance de cancelamento, seguindo a fórmula e critérios descritos na seção.

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%

Odds Ratio:  $\text{Exp}(\text{Coeficientes})$

Interpretação:  $(\text{Exp}(\text{Coeficientes}) - 1) * 100$



A interpretação das odds ratio revela insights valiosos sobre os fatores que influenciam o cancelamento de clientes. Por exemplo:

- Internet via Fibra Óptica: Clientes com este serviço têm uma chance 156% maior de cancelar em comparação aos que usam DSL, indicando a necessidade de investigar e melhorar a experiência para esse segmento.
- Tempo de Relacionamento: A cada mês adicional, a chance de cancelamento reduz em 3,6%, reforçando a importância de estratégias para aumentar a fidelidade ao longo do tempo.
- Contrato de Dois Anos: Clientes com esse tipo de contrato têm 81% menos chance de cancelar em relação aos contratos mensais, evidenciando o valor de incentivar contratos de maior duração.

Essas descobertas não apenas explicam o comportamento dos clientes, mas também fornecem um mapa claro para ações direcionadas que maximizem a retenção e otimizem nossos resultados.

# PRÓXIMOS PASSOS

---

O modelo apresentado classifica as variáveis em três categorias estratégicas: boas (protetoras), ruins (de risco) e irrelevantes, cada uma com implicações diretas para as ações empresariais.



## 01 — Boas (Protetoras)

Contratos de longo prazo, como dois anos (-81,5% de risco), e maior tempo de relacionamento (-3,6% por mês) reduzem o cancelamento, destacando a importância de incentivar fidelização.



## 02 — Ruins (De Risco)

Internet via fibra óptica (+156,4% de risco) e faturamento eletrônico (+46,5% de risco) exigem ações para melhorar a experiência e reduzir insatisfações.



## 03 — Irrelevantes

Variáveis como proteção de dispositivo e pagamento via cheque não têm impacto significativo.

Ao entender quais variáveis protegem, aumentam ou não influenciam o risco de cancelamento, a empresa pode priorizar esforços para fidelizar clientes por meio de contratos mais longos e intervenções em serviços de maior risco, como internet via fibra óptica. Esse modelo oferece uma base sólida para ações direcionadas, otimizando recursos e maximizando resultados.



# E NA PRÁTICA É O QUE ?

Supondo que exista um cliente chamado João que possui as seguintes características: João é um cliente com 10 meses de contrato, idoso, que utiliza telefone com múltiplas linhas, internet fibra óptica e streaming de TV, mas não contratou serviços adicionais como backup online. Ele tem um contrato de 1 ano, prefere faturamento eletrônico e paga via cheque eletrônico.

Coeficientes	Valores_Coeficientes	Valores_Cliente	(Coeficiente x Cliente)
Intercept	-0,466	1	-0,4656
Tempo	-0,037	10	-0,3671
IdosoYes	0,401	1	0,4006
ServicoTelefoneYes	-0,393	1	-0,3933
MultiplasLinhasYes	0,249	1	0,2490
ServicoInternetFiberOptic	0,942	1	0,9416
ServicoInternetNo	-0,743	0	0,0000
BackupOnlineYes	-0,117	0	0,0000
ProtecaoDispositivoYes	-0,017	0	0,0000
StreamingTVYes	0,250	1	0,2497
StreamingFilmesYes	0,215	0	0,0000
ContratoOneYear	-0,733	1	-0,7329
ContratoTwoYear	-1,689	0	0,0000
FaturamentoEletronicoYes	0,382	1	0,3817
MetodoPagamentoCreditCardAutomatic	-0,031	0	0,0000
MetodoPagamentoElectronicCheck	0,336	1	0,3362
MetodoPagamentoMailedCheck	0,020	0	0,0000
		<b>Probabilidade</b>	<b>0,64566</b>



## VAI CANCELAR

João possui probabilidade de cancelamento de 65%, fale com ele para reter ele e mudar essa previsão

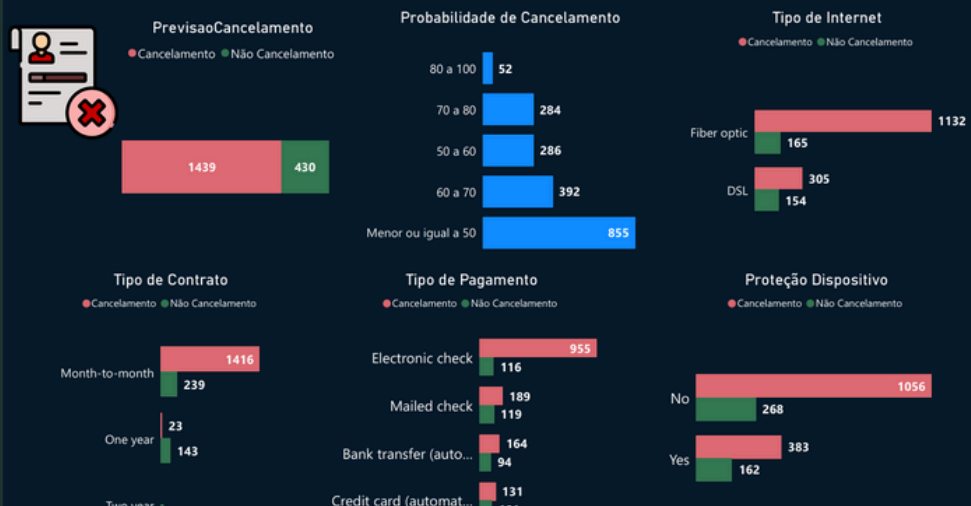
Com base no nosso modelo preditivo, observe o resultado da Figura 18, a probabilidade de João cancelar o contrato é de 65% o que o coloca em uma zona de risco. Usando o ponto de corte de 0,293, o modelo classifica João como um cliente propenso ao cancelamento.

# E NA PRÁTICA É O QUE ?

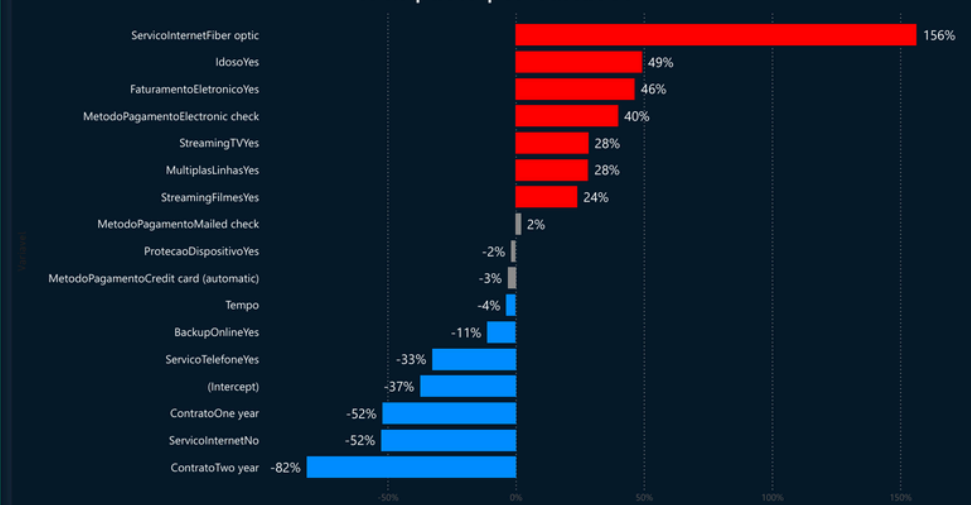
A fórmula matemática para a probabilidade de cancelamento é a seguinte:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}}$$

## Previsão e Probabilidade



## % Impacto por Variável



Link Power Bi:

<https://app.powerbi.com/reportEmbed?reportId=3837c029-a290-47a7-bb92-63490f957b26&autoAuth=true&ctid=fa7bebb9-e009-4b77-b8a7-947b125d4814>

```
1 Probabilidade =
2 VAR Logit =
3     -0.46558 * // Intercepto
4     -0.03671 * Telecom_Churn[Tempo] +
5     0.40058 * IF(Telecom_Churn[Idoso] = "Yes", 1, 0) +
6     -0.39326 * IF(Telecom_Churn[ServicoTelefone] = "Yes", 1, 0) +
7     0.24896 * IF(Telecom_Churn[MultiplasLinhas] = "Yes", 1, 0) +
8     0.94164 * IF(Telecom_Churn[ServicoInternet] = "Fiber optic", 1, 0) +
9     -0.74264 * IF(Telecom_Churn[ServicoInternet] = "No", 1, 0) +
10    -0.11693 * IF(Telecom_Churn[BackupOnline] = "Yes", 1, 0) +
11    -0.01722 * IF(Telecom_Churn[ProtecaoDispositivo] = "Yes", 1, 0) +
12    0.24969 * IF(Telecom_Churn[StreamingTV] = "Yes", 1, 0) +
13    0.21496 * IF(Telecom_Churn[StreamingFilmes] = "Yes", 1, 0) +
14    -0.73288 * IF(Telecom_Churn[Contrato] = "One year", 1, 0) +
15    -1.68906 * IF(Telecom_Churn[Contrato] = "Two year", 1, 0) +
16    0.38170 * IF(Telecom_Churn[FaturamentoEletronico] = "Yes", 1, 0) +
17    -0.03091 * IF(
18        Telecom_Churn[MetodoPagamento] = "Credit card (automatic)" ||
19        Telecom_Churn[MetodoPagamento] = "Electronic check" ||
20        Telecom_Churn[MetodoPagamento] = "Mailed check", 1, 0) +
21    0.33622 * IF(Telecom_Churn[MetodoPagamento] = "Electronic check", 1, 0) +
22    0.01954 * IF(Telecom_Churn[MetodoPagamento] = "Mailed check", 1, 0)
23
24 VAR ExpLogit = EXP(Logit)
25
26 RETURN
27     ExpLogit / (1 + ExpLogit)
28
```



# CONCLUSÃO

---

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



## 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.

# AGRADECIMENTOS

---



## Contato

**William Irineu -**  
**[kdowillian@gmail.com](mailto:kdowillian@gmail.com)**

Universidade Federal  
de Goiás - Disciplina  
de MLG 2024/2 -  
Curso de Estatística



## Fontes:

- <https://rpubs.com/WilliamIrineu>
- <https://github.com/WilliamIrineu/RegressaoLogistica>
- <https://medium.com/@kdowillian>