FERNANDO GABARDO

TELCO CUSTOMER CHURN

ABRIL

2023

TABELA DE CONTEÚDOS

Sumário

Índice figuras

Índice tabelas

01. DESCRIÇÃO DO CASO

Os integrantes do grupo, tinham em comum o desafio de gerar inteligência e insights sobre o *churn* dos clientes nas respectivas empresas em que trabalhavam. Inclusive, a primeira opção seria desenvolver um estudo com os dados reais de uma das empresas.

Entretanto, devido a questões de sigilo de dados empresariais, o professor Rafael deu a sugestão de trabalhar com um modelo de *churn* acadêmico disponível no Kaggle. Dessa maneira, seria possível ganhar experiência no trabalho com dados de *churn* e consequentemente levar esse conhecimento para suas respectivas empresas. O *dataset* escolhido foi fornecido pela IBM para fins acadêmicos

02. OBJETIVOS DO MODELO

Apesar da fonte de dados ser fictícia, é possível correlacionar com os objetivos de uma companhia de verdade. O objetivo “macro” do desenvolvimento do modelo é a redução do *churn*[1] da empresa. A redução do *churn* pode acontecer através dos seguintes planos de ação:

- **Identificar os fatores que contribuem com o *churn***

Ao listar quais são os principais fatores que contribuem com o *churn*, a companhia pode direcionar e priorizar seus recursos para resolver as causas raízes desses fatores. Dessa forma, é possível melhorar o resultado financeiro da empresa.

- **Disponibilizar um “*Score”* para as áreas de negócio**

Ao possuir a inteligência de quais clientes são os mais propensos ao *churn*, a área de negócio pode tomar ações preventivas nesse cliente e evitar que ele interrompa a parceria com a empresa.

[1] *Churn* é um indicador utilizado para medir o número de clientes que cancelam um serviço em determinado período.

03. DESCRIÇÃO DOS DADOS

O *Dataset* contém dados fictícios da empresa “Telco” que providencia serviços de telefone e internet para 7043 clientes e contém as seguintes informações:

- ✓ Quais saíram e ficaram na empresa (*churn*);
- ✓ Quais serviços os clientes adquiriram;
- ✓ Informações da conta;
- ✓ Informações demográficas.

Campos do Dataset

01. Customer ID – Campo com identificação única de cada cliente;
02. gender – Sexo do cliente;
03. SeniorCitizen – Se o cliente é idoso ou não;
04. Partner – Estado Civil;
05. Dependents – Se possui dependentes;
06. Tenure – Tempo como cliente (meses);
07. PhoneService – Se contratou Telefone;
08. MutipleLines – Se contratou +1 linha;
09. InternerService – Tipo de serviço de internet;
10. OnlineSecurity – Se contratou segurança online;
11. OnlineBackup – Se contratou backup online;
12. DeviceProtection - Se contratou proteção do dispositivo;
13. TechSupport – Se contratou suporte técnico;
14. StreamingTV – Se contratou “Streaming TV”;
15. StreamingMovies – Se contratou *streaming* de filmes;
16. Contract – Tipo do contrato;

17. PaperlessBilling – Se possui cobrança sem papel;
18. PaymentMethod – Método de pagamento;
19. MonthlyCharges – O quanto o cliente é cobrado mensalmente;
20. TotalCharges – O total cobrado ao cliente;
21. Churn – Se o cliente deixou a empresa.

04. RESULTADOS ENCONTRADOS PELA EDA

04.1 Distribuição das variáveis numéricas

Analisando as variáveis numéricas, é possível observar distribuições bem diferentes da distribuição normal.

Ao observar a faixa de cobrança, é possível notar que há uma maior frequência nas faixas iniciais, em seguida há uma queda e segue uma distribuição mais semelhante à normal.

No histograma de *tenure* (meses como cliente) verifica-se que há muitos clientes com pouco, ou muito tempo de casa.

A distribuição do total de cobranças é maior nos clientes com uma faixa menor de cobrança

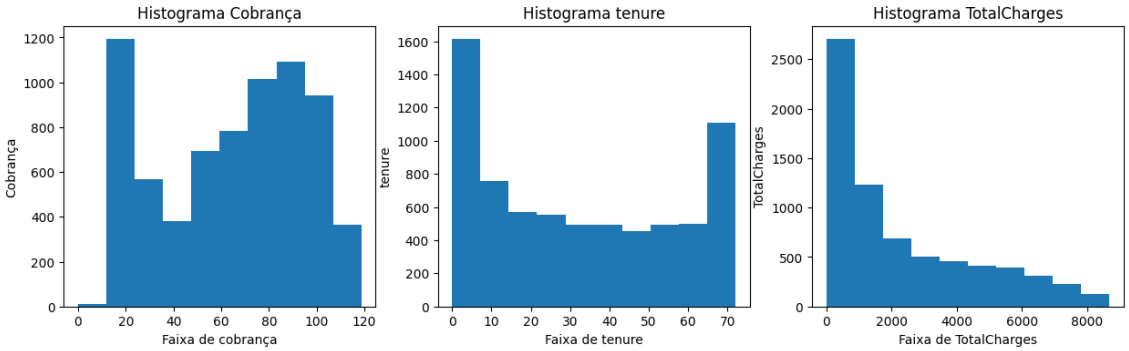

Figura 1 - Distribuição das variáveis numéricas

04.2 Distribuição as variáveis categóricas

Percebe-se que algumas variáveis não são uniformes na base. O *churn*, por exemplo está distribuído como: 73,5% (cliente ativo) 26,5% (cliente *churn*).

Verifica-se também que grande parte dos clientes (90,3%) não contrata o serviço de telefone.

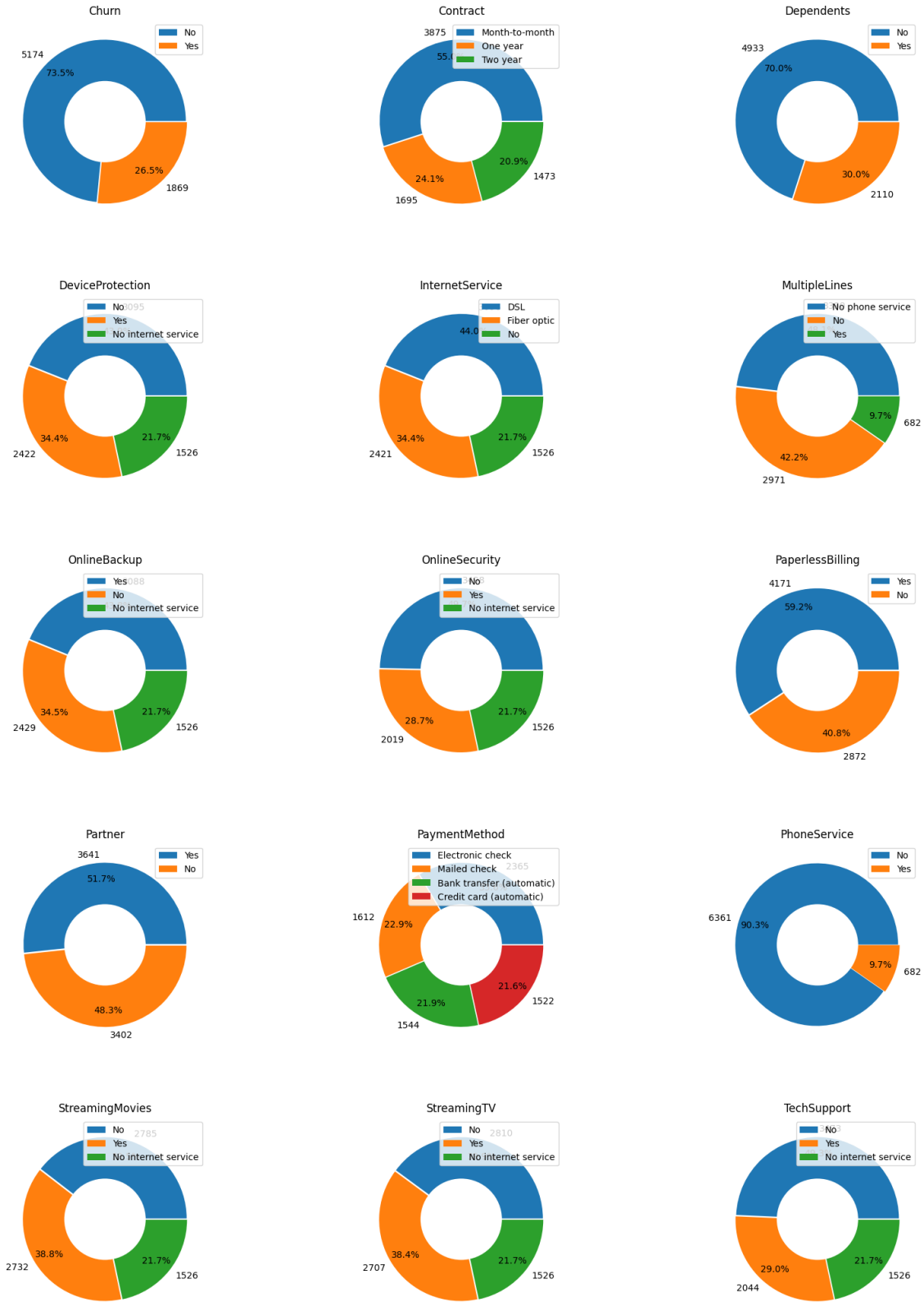

Figura 2 - Distribuição das variáveis categóricas

04.3 Correlação das variáveis categóricas com o *churn*

Ao analisar todas as variáveis categóricas e verificar a sua correlação com o *churn,* verifica-se que alguns fatores não fazem muito sentido estarem tão correlacionados assim. Por exemplo, o fator *PhoneService* a princípio não parece fazer sentido um cliente que também utiliza o serviço de telefone, ter uma propensão maior ao *churn*. No mercado, de maneira geral, quanto mais serviços um cliente utiliza, menor a chance de cancelar a parceria.

Tabela 1 - Correlação das variáveis categóricas com o *churn*

Fator	Principal Atributo	Correlação[2]
PhoneService	Yes	24,12
Contract	Month-to-month	23,49
Dependents	No	21,90
OnlineSecurity	No	20,74
TechSupport	No	20,53
PaperlessBilling	Yes	19,87
InternetService	Fiber optic	18,41
OnlineBackup	No	17,50
DeviceProtection	No	17,19
Partner	No	17,03
PaymentMethod	Electronic check	15,20
StreamingTV	No	13,37
gender	Female	13,33
StreamingMovies	No	13,31
MultipleLines	Yes	12,06

[2] Correlação feita através da função pd.crosstab

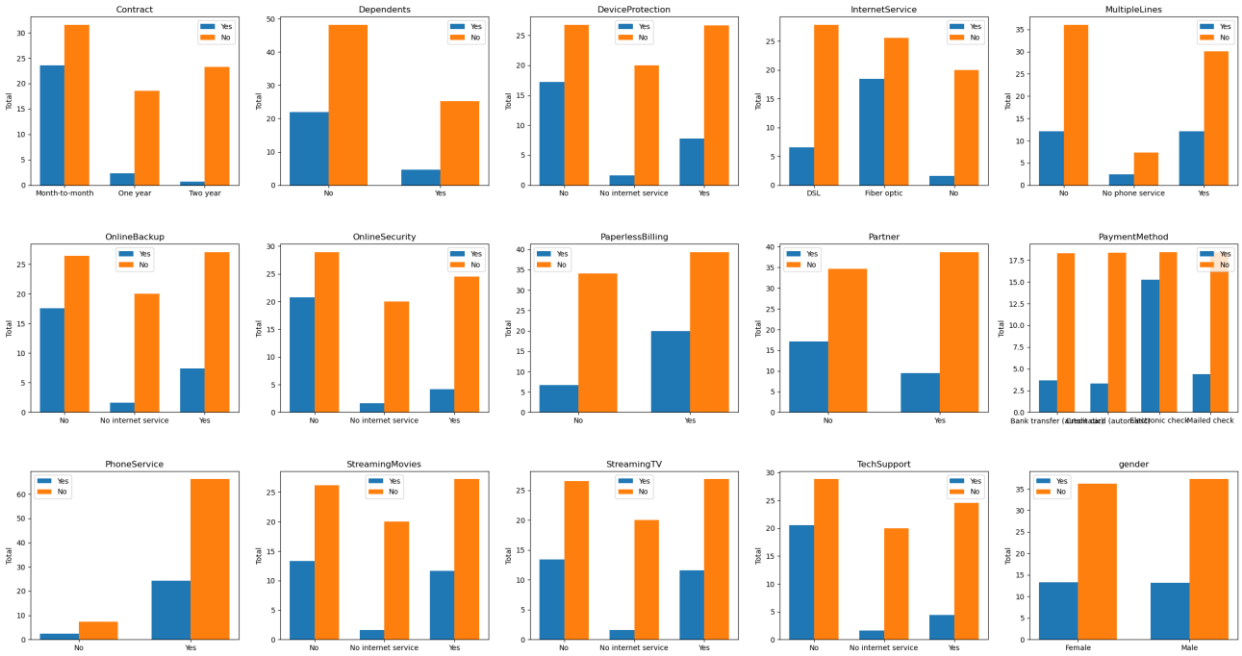

Figura 3 - Correlação das variáveis categóricas com o *churn*

Pensando nisso, utiliza-se o método estatístico do qui-quadrado[3] para analisar efetivamente as correlações das variáveis categóricas com o *churn*. Com isso, observa-se que o fator com maior correlação é o *Contract,* o que faz sentido, levando em consideração que se trata da duração do contrato de prestação de serviços (mensal, anual ou de 2 anos).

Tabela 2 - Método do qui-quadrado

fator	**p_value**	**result**
Contract	7,3262E-257	Dependent (reject H0)
OnlineSecurity	1,4007E-184	Dependent (reject H0)
TechSupport	7,4078E-180	Dependent (reject H0)
InternetService	5,8312E-159	Dependent (reject H0)
PaymentMethod	1,4263E-139	Dependent (reject H0)
OnlineBackup	7,7761E-131	Dependent (reject H0)
DeviceProtection	1,9594E-121	Dependent (reject H0)

[3] Alpha – 0,001

StreamingMovies	5,35356E-82	Dependent (reject H0)
StreamingTV	1,32464E-81	Dependent (reject H0)
PaperlessBilling	8,2362E-58	Dependent (reject H0)
Dependents	2,01966E-42	Dependent (reject H0)
Partner	3,9738E-36	Dependent (reject H0)
MultipleLines	0,003567927	Independent (H0 holds true)
PhoneService	0,349923989	Independent (H0 holds true)
gender	0,490488471	Independent (H0 holds true)

Através do qui-quadrado, a maioria das variáveis possui certa relação com o churn, exceto: *MultipleLines*, *PhoneService* e *gender*.

Os 5 fatores com maior correlação são: *Contract*, *OnlineSecurity*, *TechSupport*, *InternetService* e *PaymentMethod*. Além dos contratos de prestação de serviços de maior duração contribuírem para evitar o *churn*, os clientes que contratam segurança online e suporte técnico, também tendem a cancelar menos.

04.4 Correlação das variáveis numéricas com o *churn*

Quanto maior o período de parceria do cliente com a empresa, menor a sua correlação com o *churn*. Para as cobranças mensais, não há muita diferenciação. Nas cobranças totais, assim como no tempo de casa, quanto maior o valor investido ao longo da “vida” como cliente, menor a correlação com o *churn*.

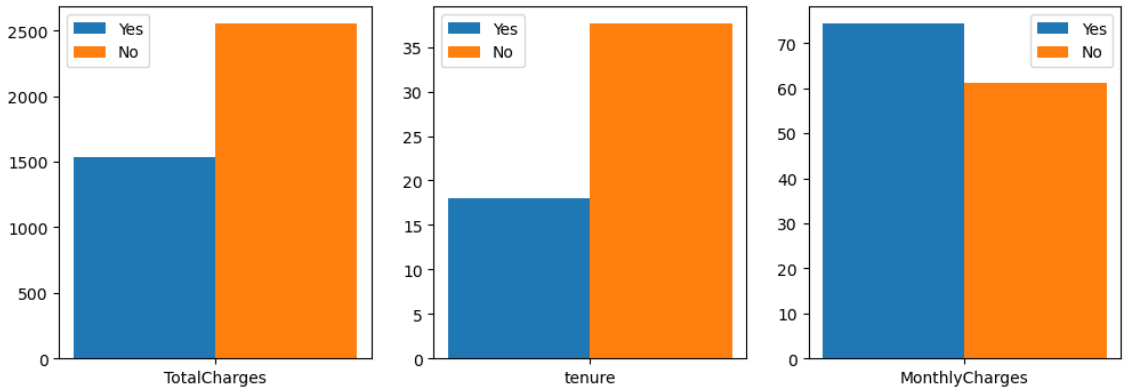

Figura 4 - Correlação das variáveis numéricas com o churn

04.5 Correlação multivariada com o *churn*

Ao correlacionar as variáveis numéricas com o *churn*, verifica-se alguns pontos interessantes. Observa-se de maneira mais clara a correlação entre o tempo em que o cliente é cliente com o *churn*, ou seja, quanto mais tempo, menor chance de o cliente encerrar a parceria.

Outro ponto interessante está na correlação entre *tenure* e *MonthlyCharges* (tabela abaixo), observa-se que o *churn* está mais concentrado nos clientes com alta cobrança mensal e pouco tempo como cliente. Comportamento semelhante está na correlação entre *MonthlyCharges* e *TotalCharges* na qual verifica-se que o *churn* está mais concentrado nas altas cobranças mensais.

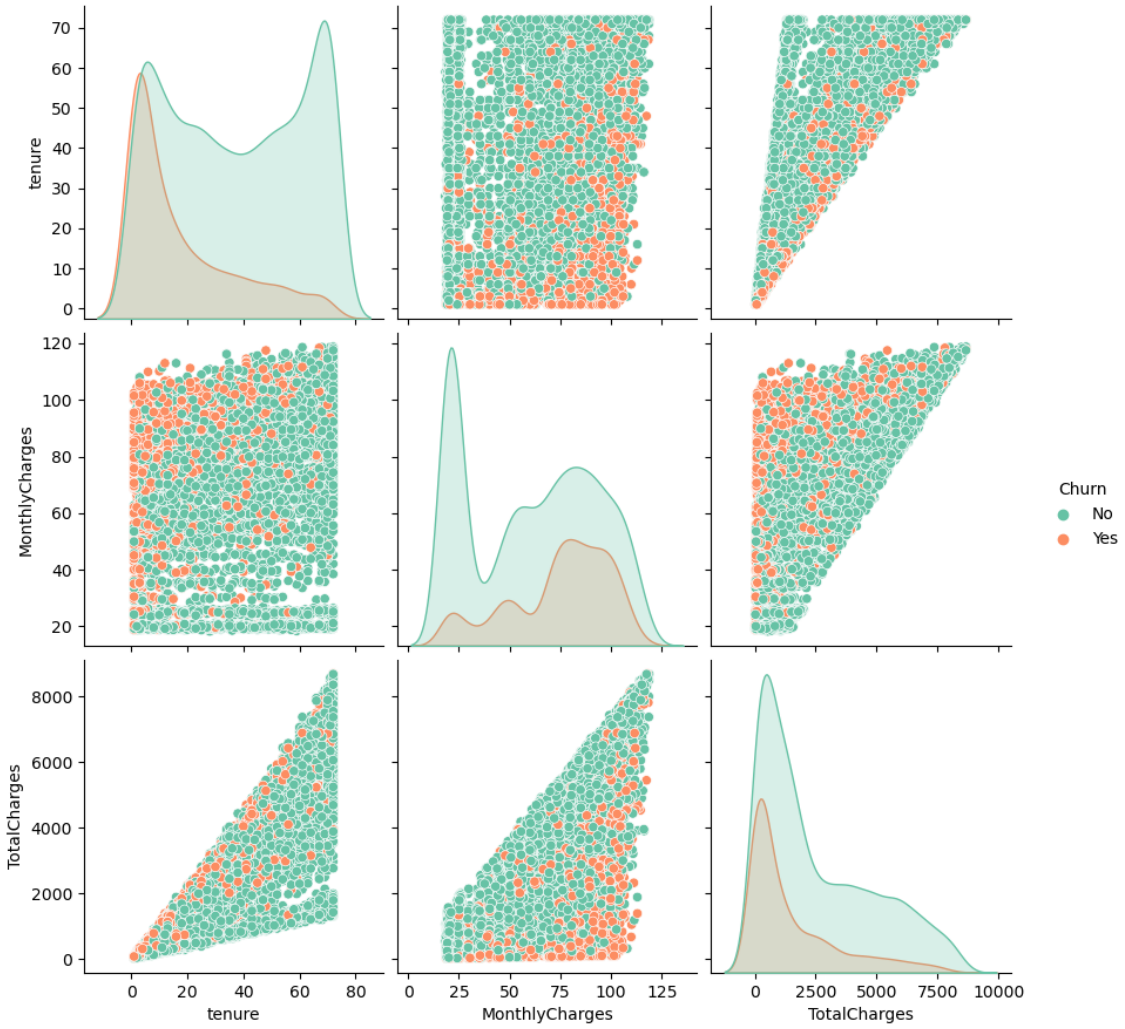

Figura 5 - Correlação multivariada com o *churn*

05. DATA WRANGLING

Foram identificadas 11 linhas com valores nulos as quais foram removidas da análise. Além disso, o campo *TotalCharges* veio originalmente como *string* e teve seu *dtype* modificado para *float*.

Não se identificou necessidade de gerar novas *features* e foram mantidos os campos originais.

06. SELEÇÃO DO MODELO

Considerando os principais objetivos do projeto sendo:

a. Identificar os fatores que contribuem com o *churn*
b. Disponibilizar um "*Score"* (alerta 0 e 1) para as áreas de negócio

O primeiro, de certa forma já foi atingido através da análise exploratória de dados. Para criar o "*Score*" deve-se criar um modelo de classificação com o objetivo de as áreas de negócio tomarem ações proativas para evitar que o cliente cancele o serviço.

Para os testes do modelo, foi utilizado o pacote sklearn. 70% da base foi separada para o conjunto de treino e 30% para o conjunto de teste.

06.1 Seleção do Algoritmo

Levando em consideração que a parcela dos clientes que deram *churn* é minoria nesse conjunto de dados, os indicadores mais importantes para o sucesso no negócio desse modelo seriam: *Precision* e *Recall*.

Além disso, a precisão alta é importante para evitar o retrabalho da área de negócio (abordar clientes falsos positivos) e o *Recall* para evitar ao máximo os falsos negativos. Assim o indicador ideal seria o F1 (média geométrica entre *Precision* e *Recall).*

Para uma análise inicial do algoritmo mais indicado para esse problema de negócio, foram desenvolvidos diversos modelos com alterações mínimas aos Hyperparâmetros:

Tabela 3 - Seleção do algoritmo

Modelo	Acurácia	Precisão	Recall	F1 Score
Gradient	0,80	0,65	0,53	0,58
Regressão Logística	0,80	0,64	0,53	0,58
LigthGBM	0,79	0,62	0,52	0,57

Adaboost	0,79	0,63	0,51	0,56
Xgboost	0,78	0,60	0,51	0,55
Randon Forrest	0,78	0,61	0,48	0,54
KNN	0,77	0,59	0,45	0,51
Decision Tree	0,73	0,49	0,51	0,50
SVM	0,78	0,65	0,35	0,46

Em uma primeira análise, o modelo de Gradient Boosting foi o que apresentou o melhor desempenho.

06.1.1 Grid Search

Para selecionar o melhor algoritmo, foi feito o *Grid Search* com algumas combinações de *hyperparâmetros*:

Tabela 4 - *Grid Search*

Modelo	Hyperparâmetros	F1
Regressão Logística	{'C': 10, 'dual': False, 'penalty': 'l2', 'random_state': 25}	0,603
Adaboost	{'learning_rate': 0.1, 'n_estimators': 200, 'random_state': 25}	0,599
Xgboost	{'booster': 'gblinear', 'eta': 1}	0,599
Gradient	{'learning_rate': 0.1, 'loss': 'exponential', 'n_estimators': 100}	0,597
LigthGBM	{'bagging_seed': 1, 'extra_trees': False, 'min_data_in_leaf': 100}	0,586
Randon Forrest	{'criterion': 'gini', 'min_samples_leaf': 3, 'n_estimators': 50, 'random_state': 25}	0,583
SVM	{'C': 1, 'degree': 3, 'kernel': 'linear'}	0,535
Decision Tree	{'criterion': 'log_loss', 'min_samples_leaf': 3, 'min_samples_split': 2, 'random_state': 100}	0,528
KNN	{'algorithm': 'auto', 'leaf_size': 1, 'n_neighbors': 5, 'weights': 'uniform'}	0,518

Observa-se que a seleção de *hyperparâmetros*, no geral, melhorou o desempenho de todos os modelos. Após o ajuste, o algoritmo com o melhor valor de F1 foi a Regressão Logística, o qual foi o algoritmo selecionado.

07. CONCLUSÕES

Através da análise exploratória de dados, observa-se alguns pontos-chave da jornada do cliente os quais contribuem significativamente com o *churn*. É possível aplicar diversos incentivos e ações com base nesses pontos para reduzir o *churn* da empresa de maneira geral.

O melhor resultado de F1 obtido com os dados disponíveis e os algoritmos testados foi de 60,3%. Dessa forma, já seria um ponto de partida para a atuação do time de negócio de forma preventiva, sem muito retrabalho.

Apesar disso, para uma melhor visão da aplicabilidade do Score dentro do negócio, seria necessário levantar quantos clientes efetivamente deixam de cancelar o serviço quando a empresa entra em contato proativo, em comparação com algum grupo de controle.

Com os testes, pode-se verificar o impacto no negócio e justificar uma nova etapa no projeto, visando melhorar ainda mais os resultados desse modelo.

Possíveis próximos passos:

- Testes no impacto do *churn*;
- Enriquecimento de dados;
- Maior profundidade na seleção dos *Hyperparâmetros*;
- Disponibilizar o Score como probabilidade de *churn*.

08. LINKS IMPORTANTES

- ✓ Kaggle - https://www.kaggle.com/datasets/blastchar/telco-customer-churn
- ✓ Repositório Git - https://github.com/fggabardo/Telco-Customer-Churn-Gabardo
- ✓ Método do qui-quadrado - http://www.leg.ufpr.br/lib/exe/fetch.php/disciplinas:ce001:teste_do_qui-quadrado.pdf