🟡

Desafio: Preveja os usuários com alta chance de deixar seu Streaming

Escola	Dados
Habilidades	Classificação
Nº	7

🚀 Desafio

📎 **Arquivos do Desafio:**

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/9476d09e-5efc-4dee-8dd0-cbc871115ece/streaming_data.csv

- **ChatGPT:** Pode ser útil para iniciar sua pesquisa!
- **Python Graph Gallery**: repositório com o passo a passo de como gerar gráficos utilizando as principais bibliotecas de Python
- **SciKit Learn:** documentação com os principais modelos utilizados para classificação

Preveja os usuários com alta chance de deixar seu Streaming

> Utilize um modelo de classificação para mapear qual o perfil de usuários tem mais chance de deixar sua plataforma de streaming. Compreender quem é o perfil que está aumentando o churn do seu negócio é essencial para tomar ações que reduzam essas perdas, seja alterando critérios na venda ou modificando o produto.

Contexto

Você trabalha em uma plataforma de streaming e a diretoria está preocupada com o alto índice de usuários cancelando as suas assinaturas. Eles acreditam que é possível prever se um usuário tem mais chance de deixar a plataforma antes que isso aconteça, e com base nessa informação tomar ações para reduzir o churn.

Seu objetivo é criar um modelo de classificação capaz de prever se um usuário tem mais chance de cancelar a sua assinatura na plataforma ou não. Para isso, a empresa forneceu uma base de dados em csv contendo dados sobre as contas dos clientes.

Sobre os dados

Uma adaptação do problema de ecommerce, disponível no Kaggle. Acesse os dados aqui:

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/75a740fb-4146-455a-8d13-6a24ba56d2c8/streaming_data.csv

Os dados fornecidos possuem informações sobre as contas dos clientes na plataforma de streaming, divididos entre contas Basic, Standard e Premium, onde cada uma oferece uma gama maior de serviços que a anterior.

Coluna	Descrição	Tipo
client_id	Código de identificação do cliente	Int
age	Idade do cliente	Int

Coluna	Descrição	Tipo
gender	Gênero do cliente	String
region	Região de origem do cliente	String
subscription_days	Dias de assinatura ativa do cliente	Int
subscription_type	Tipo de conta	String
num_contents	Quantidade de conteúdos assistidos	Int
avg_rating	Avaliação média dos conteúdos da plataforma	Int
num_active_profiles	Número de perfis ativos na plataforma	Int
num_streaming_services	Quantidade de serviços de streaming que o cliente possui	Int
devices_connected	Quantidade de dispositivos conectados à conta	Int
churned	Se o cliente cancelou a conta ou não	Int

Como começar?

Desenvolva um modelo de classificação que seja capaz de prever se o cliente irá cancelar o serviço ou não, levando em consideração o seu perfil no streaming.

Teste com mais de um tipo de modelo para encontrar o que possuir a melhor performance em comparação com um baseline. Utilize gráficos e visualizações para auxiliar e enriquecer a sua análise.

Não se esqueça de documentar cada etapa, justificando as escolhas realizadas. É essencial informar os insights obtidos e como o serviço de streaming pode se beneficiar do uso do seu modelo para resolver o problema de negócio. Boa sorte!

🎯 Etapas de Desenvolvimento

Etapa 01) Análise exploratória dos dados (Data Understanding)

a. Carregue a base de dados;

b. Realize uma descrição estatística dos dados;

c. Verifique os tipos de dados

d. Verifique a quantidade de valores faltantes

> 💡 **Dica:** Utilize as funções .info(), .isna().sum(), faça algumas plotagens para entender a distribuição dos dados.

Etapa 02) Tratamento dos Dados (Data Preparation)

1. Substituir valores “NaN” por 0 Colunas → Time_on_platform, Num_streaming_services, Churned, Avg_rating, Devices_connected
2. Dropar linhas nulas nas colunas Gender, Subscription_type e Age
3. Transformando valores churned 0 e 1 por No e Yes
4. Transformando valores floats em valores inteiros

> 💡 **Dica:** Utilize as funções fillna(), dropna, replace, astype(int)

Etapa 03) Modelagem dos Dados - Regressão Logística

a. Definir variáveis X e y para o modelo
b. Realizar o .fit do modelo
c. Separar em train e test
d. Realizar a modelagem
e. Plotar matrix confusão
f. Printar métricas

> 💡 **Dica:** Utilize as funções LabelEncoder, .fit, .transform, get_dummies, MinMaxScaler, train_test_split, predict, assign, ConfusionMatrixDisplay

Etapa 04) Modelagem dos Dados - Tunning

a. Definir variáveis X e y para o modelo

b. Realizar o .fit do modelo
c. Separar em train e test
d. Realizar a modelagem
e. Plotar matrix confusão
f. Printar métricas

> 💡 **Dica:** Utilize as funções LabelEncoder, .fit, .transform, get_dummies, MinMaxScaler, train_test_split, predict, assign, ConfusionMatrixDisplay

Etapa 05) Modelagem dos Dados - Random Forest

a. Realizar a montagem do grid search
b. Realizar o .fit do modelo
c. Realizar o Tunning
d. Realizar a modelagem
e. Plotar matrix confusão
f. Printar métricas

> 💡 **Dica:** Utilize as funções grid_search.best_estimator_.get_params(), fit, assign, ConfusionMatrixDisplay

📝 Critérios de Avaliação

Os critérios de avaliação mostram como você será avaliado em relação ao seu desafio.

Critérios	**Atendeu às Especificações**	Pontos
Data Understanding	Para esta etapa os alunos precisam trabalhar a base de dados para entender com estão as distribuição dos dados, para isso eles precisam fazer as etapas: Describe, Info, isna().sum()	20

Critérios	Atendeu às Especificações	Pontos
Data Preparation	Para esta etapa os alunos precisam preparar a base de dados para realizar posteriormente a etapa de modelagem, para isso eles precisam utilizar minimamente as seguintes funções: fillna(), dropna, replace, astype(int), ou seja, substituição de valores, exclusão de valores, troca de valores, mudança de tipo de dado.	20
Modeling - Regressão Logística	Para esta etapa os alunos precisam modelar a base de dados para. Eles precisam utilizar minimamente as seguintes funções: LabelEncoder, .fit, .transform, get_dummies, MinMaxScaler, train_test_split, predict, assign, ConfusionMatrixDisplay	20
Modeling - Radom Forest	Para esta etapa os alunos precisam modelar a base de dados para. Eles precisam utilizar minimamente as seguintes funções: LabelEncoder, .fit, .transform, get_dummies, MinMaxScaler, train_test_split, predict, assign, ConfusionMatrixDisplay	20
Modeling - Tunning	Para essa etapa os alunos precisam tunar os modelos e utilizar minimamente as seguintes funçoes para atingir uma acurácia maior: grid_search.best_estimator_.get_params(), fit, assign, ConfusionMatrixDisplay	20

📅 Entrega

📎 **Como entregar:** Você deverá submeter o link compartilhável do colab!

💡 **Dica:** pense que essa documentação está sendo apresentada para o cliente final, seu modelo não pode ser uma “caixa preta”.