



# Case Cientista de Dados



## Descrição:

Este desafio tem como objetivo resolver um problema de negócios atuando como um cientista e dados, e que deverá abranger os seguintes tópicos:

- Extração e ingestão da base de dados em alguma ferramenta (GUT, Colab, Databricks);
- Análise exploratória dos dados;
- Resolver **um** dos dois desafios de negócio propostos com uma solução de algoritmo **supervisionado**;
- Resolver **o** desafio de negócio proposto com abordagem **não-supervisionada**;
- Justificar as análises, premissas adotadas, métodos escolhidos e mostrar resultados – tanto de modelo quanto de negócios.



O projeto deve estar minimamente organizado e comentado, para que possa ser apresentado passo-a-passo de forma linear para entendimento da banca avaliadora.

Em caso de testes que falharam, avalie se é interessante deixar explícito (ex: testes de hiperparâmetros).



No dia da apresentação, é necessário compartilhar a tela e mostrar o código – podendo ser dentro da plataforma, ou uma extração **.html**, por exemplo. O intuito é passar por todas as etapas desenvolvidas, mostrando as técnicas aplicadas, explicando como funcionam, justificando o uso e citando outras técnicas que poderiam ser usadas (ou técnicas que não, sempre explicando o porquê).

A solução deve ser mostrada sob uma ótica de modelagem – mostrar as métricas de análise, explicando-as e justificando o uso – e uma de ótica de negócios, mostrando os ganhos de forma clara e coerente.

A metodologia de apresentação e a clara comunicação do trabalho realizado são de inteira responsabilidade do candidato.

O uso de algoritmos ou pacotes em particular é de livre escolha, onde a explicação dos motivadores será critério de avaliação – questões de negócio, de performance, de processamento computacional, de recursos e implantação em uma hipotética produtização, etc.

## **Critérios de Avaliação:**

O critério de avaliação recebido para cada tópico estará contido entre três definições:

- A) "Realizou essa etapa de forma adequada e demonstrou domínio do tema";
- B) "Realizou essa etapa de forma parcialmente adequada ou demonstrou domínio parcial do tema, com orientações sendo necessária para aprofundamento";
- C) "Não realizou a etapa de forma adequada ou não demonstrou domínio do tema".

Segue o conjunto de regras utilizado para definir a badge de cada participante:

**1) Badge de Expert:**

- No máximo **dois** "Não realizou a etapa de forma adequada ou não demonstrou domínio do tema"
- Pelo menos **treze** "Realizou essa etapa de forma adequada e demonstrou domínio do tema".



## 2) **Badge de Advanced:**

- No máximo **seis** "Não realizou a etapa de forma adequada ou não demonstrou domínio do tema".
- Pelo menos **cinco** "Realizou essa etapa de forma adequada e demonstrou domínio do tema".

## 3) **Badge de Driven:**

- No máximo **dezesseis** "Não realizou a etapa de forma adequada ou não demonstrou domínio do tema"
- No mínimo **cinco** "Realizou essa etapa de forma parcialmente adequada ou demonstrou domínio parcial do tema, com orientações sendo necessária para aprofundamento" ou critério superior.



#### 4) **Sem Badge - Lover**

- Pelo menos **dezesseis** "Não realizou a etapa de forma adequada ou não demonstrou domínio do tema"
- No máximo **cinco** "Realizou essa etapa de forma parcialmente adequada ou demonstrou domínio parcial do tema, com orientações sendo necessária para aprofundamento" **ou critério superior**.

# Fique atento às dicas!

Projeto	Análise Exploratória	<p>Analizou hipóteses voltadas ao problema e para a modelagem?</p> <p>Compreendeu as implicações das análises?</p> <p>Utilizou o resultado das análises para orientar decisões?</p>
	Dataprep + Feature Engineering + Feature Selection	<p>Realizou Dataprep adequado (ex: criação de público-alvo, target, análise de volumetrias, missings, outliers, etc)?</p> <p>Realizou Feature Engineering adequado (ex: construção de novas features, transformações de features, etc)?</p> <p>Realizou Feature Selection adequado (ex: selecionou variáveis significativas para o problema/modelo construído, etc)?</p>
	Modelagem	<p>Compreende a estrutura do modelo utilizado (ex: funcionamento do algoritmo, hiperparametros, vantagens e fraquezas etc.)?</p> <p>A modelagem é adequada ao problema + pipeline construído e entende os motivos disso?</p> <p>Compreende os impactos ao utilizar outras técnicas de modelagem no problema + pipeline construído?</p>
	Validação	<p>Realizou uma validação adequada para os resultados observados (ex: leakage, análise de overfit e underfit, tamanho da amostra, etc)?</p> <p>Compreende outras formas de validação que poderiam ser utilizadas (cross-validation, train test split, out of sample, out of time, etc)?</p> <p>Consegue entender como alterações no problema impactam a validação escolhida?</p>
	Definição e avaliação dos resultados	<p>Escolheu métricas de avaliação adequadas ao problema (ex: MAE, RMSE, accuracy, recall, auc, silhueta, etc)?</p> <p>Entende como outras métricas impactariam os resultados do case?</p> <p>Entende como mudanças na estrutura do problema impactariam diferentes métricas escolhidas?</p>
	Qualidade de código	<p>A solução apresentada está organizada e roda (ex: nomes intuitivos, comentários, ordenamento intuitivo etc.)?</p> <p>O código se encontra padronizado (ex: classes, funções e variáveis com nomes padronizados, clean code, pep8, etc)?</p> <p>A solução foi realizada com noções de implementação (ex: modularização, tempo de processamento, versionamento etc.)?</p>

**Fique atento às dicas!**



Apresentação	Soft Skills	Foi sucinto e conseguiu responder dentro do tempo indicado? Foi claro nas explicações? A apresentação foi bem estruturada com começo, meio e fim?
--------------	-------------	---

**Observação 1:** Conhecimentos sobre o tema de **viés e variância** são considerados como um critério transversal por todo processo de avaliação;

**Observação 2:** As referências indicadas nos exemplos são apenas sugestões sobre o direcionamento geral das perguntas realizadas durante a banca.

## Entregáveis:

A seguir são listados os entregáveis:

1. Código do notebook desenvolvido para a solução – export em .html ou outro formato legível – **OBRIGATÓRIO**;
2. Apresentação *Power Point* com a descrição do problema, os principais insights, os resultados de modelo e de negócios – **OPCIONAL**.

Lembrando que mesmo em caso de ter um *Power Point* montado para apresentação na banca, o candidato deverá ter o código **(em HTML ou com a ferramenta de escolha aberta)** disponível e aberto para mostrar detalhes do código quando questionado.

A organização do código auxilia para guiar a apresentação além de ser critério de avaliação.



## **Apresentação do Case:**

Durante a apresentação é esperado que o candidato demonstre como resolveu os temas de negócios abordando o problema como um cientista de dados, e consiga fazer conexões do case com seus outros conhecimentos não necessariamente aplicados.

O candidato terá 1hr30m (uma hora e trinta minutos) para apresentar o case e responder as perguntas dos avaliadores.

Boa sorte e esperamos que sua solução demonstre um domínio completo dos tópicos de Ciência de Dados mencionados!

## Case Cientista de Dados

Os dados abaixo representam o histórico de dois anos (2015-2017) de uma empresa que oferece serviço de streaming de música baseado em assinatura.

Quando os usuários se inscrevem no serviço, eles podem optar por renovar o serviço manualmente ou renovar automaticamente. Os usuários podem cancelar ativamente sua associação a qualquer momento.

Para esse cenário, temos os seguintes desafios:

A) Sabendo que existe a seguinte ação de retenção para clientes: Quando detectamos que um cliente não renova a assinatura, oferecemos 3 meses grátis. Porém, identificamos que essa ação é muito reativa e entendemos que uma abordagem proativa seria mais efetiva.

Sendo assim, é proposto que você crie um modelo classificador para prever clientes que serão churn 3 meses no futuro (ou seja, clientes que possuem assinatura ativa no período analisado e 3 meses depois desse período ele não é mais ativo, ou porque cancelou ou não renovou a assinatura) e indique os clientes que serão direcionados para a ação de forma proativa.

Assumindo que, usando a ação de forma proativa, 50% dos clientes que iriam cancelar (Verdadeiro Positivo) respondem de forma positiva e continuam ativos por mais um ano, qual sua avaliação sobre sua solução?

Mínimo esperado:

- Criação de target;
- Feature Engineering;
- Feature Selection;
- Predictive Modeling;
- Quantidade de clientes retidos e resultado financeiro da ação.



B) O comitê executivo precisa de visibilidade de rentabilidade das assinaturas dos clientes para antecipar tendências.

O custo é dado por:

$$C(u,t)=50+0.0051u+0.0001t$$

Em que  $u$  é a quantidade de músicas únicas que o cliente ouviu no mês de referência e  $t$  é o tempo total em segundos que o cliente ouviu no mesmo período;

Desenvolva um modelo para estimar a Margem Líquida (Preço – Custo) do produto e avalie sua performance em  $M+1$

A partir do métrica escolhida, o que você pode concluir sobre o resultados?

Mínimo esperado:

- Criação de target usando a função dada;
- Feature Engineering;
- Feature Selection;
- Predictive Modeling;
- Conclusão do resultado (métrica escolhida, intervalo de confiança, etc).

C) Considerando o problema escolhido anteriormente (Churn ou Rentabilidade), realize uma análise não-supervisionada dos clientes com objetivo de aprofundar a compreensão sobre características deles.

Algumas sugestões de possíveis direcionamentos para sua análise:

- a) Análise de clientes com diferentes perfis de uso da plataforma, com as variadas estimações de rentabilidade/churn;
- b) Análise de perfis de clientes com diferentes volatilidade/incerteza nas respostas de rentabilidade/churn;
- c) Análise de erros sistemáticos cometidos pelos modelos do case supervisionado;
- d) Análise da variação temporal no comportamento dos clientes da base.

Para ter acesso as bases e resolver o case, clique abaixo:

## Kaggle

<https://www.kaggle.com/datasets/gcenachi/case-data-master-2024>





## Dicionário de dados:

User log data has been aggregated and converted to parquet format

## Tables

transactions.csv

transactions of users up until 3/31/2017.

- msno: user id
- payment\_method\_id: payment method
- payment\_plan\_days: length of membership plan in days

- plan\_list\_price: in New Taiwan Dollar (NTD)
- actual\_amount\_paid: in New Taiwan Dollar (NTD)
- is\_auto\_renew
- transaction\_date: format %Y%m%d
- membership\_expire\_date: format %Y%m%d
- is\_cancel: whether or not the user canceled the membership in this transaction.

## user\_logs.csv

- daily user logs describing listening behaviors of a user. Data collected until 3/31/2017.
- msno: user id
- date: format %Y%m%d
- num\_25: # of songs played less than 25% of the song length
- num\_50: # of songs played between 25% to 50% of the song length
- num\_75: # of songs played between 50% to 75% of of the song length
- num\_985: # of songs played between 75% to 98.5% of the song length
- num\_100: # of songs played over 98.5% of the song length
- num\_unq: # of unique songs played
- total\_secs: total seconds played



## **members\_v3.csv**

User information. Note that not every user in the dataset is available.

- msno
- city
- bd: age. Note: this column has outlier values ranging from -7000 to 2015, please use your judgement.
- gender
- registered\_via: registration method
- registration\_init\_time: format %Y%m%d





