Seleção Analista Data Science

# Introdução

Essa atividade é a segunda etapa do processo seletivo para a vaga de analista do setor de Data Science da Fortbrasil. O principal objetivo dessa etapa é conseguir entender o nível técnico de cada participante e principalmente as soluções de negócio que o participante trará nas questões abaixo.

# A Empresa

Prazer, somos a FortBrasil S/A, uma fintech genuinamente cearense que há mais de 15 anos atua no segmento financeiro, na concessão de crédito rápido e consciente para parceiros e para as classes emergentes. Atualmente, nossa maior operação está focada na região nordeste, mas estamos em expansão, a todo vapor, para todo o restante do Brasil. Nos últimos meses, já começamos operações em São Paulo, Brasília, Goiânia, dentre outros. Ao longo da nossa história, já conquistamos quase 2 milhões de clientes e mais de 300 clientes varejistas (privates). Em 2019, fechamos o ano com, aproximadamente, um bilhão em vendas. Uma grande conquista ocorrida no ano de 2019 foi o início da operação do cartão Fortbrasil Mastercard, onde consolidamos nossa expansão nacional.

# Pontos Importantes

O case será composto por 3 questões e deve ser enviado até o dia 01/04/2021 as 23:59 para os seguintes e-mails:

[gabriela.tabare@fortbrasil.com.br](mailto:gabriela.tabare@fortbrasil.com.br)

[elieyd.barreto@fortbrasil.com.br](mailto:elieyd.barreto@fortbrasil.com.br)

Para solucionar o case, o participante poderá utilizar **qualquer** **ferramenta gratuita** (Python, R, Power BI, Julia, Pentaho, mysql...) ou **ferramentas pagas que temos licenças** (Microsoft Office). Todas as respostas do case e sua respectiva apresentação deverá ser publicada no **GitHub ou Gitlab** e disponibilizado o link para que possa ser avaliado.

# Case

## Questão 1

O negócio da Fortbrasil consiste na oferta de crédito rápido e consciente para parceiros e clientes de classes emergentes, sendo nosso principal produto o cartão de crédito. Para a primeira questão, disponibilizamos algumas variáveis relacionadas à fatura dos clientes:

**ID\_CONTA:** Identificador da conta;

**DT\_VENCIMENTO:** Data vencimento da fatura;

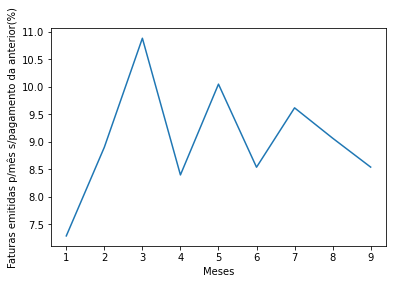
**DS\_ROLAGEM:** Indicador que define se o cliente pagou a fatura do Mês anterior ou não. Se o valor for **FX0** quer dizer que o cliente pagou a fatura do mês anterior e se **FX1** quer dizer que o cliente não pagou a fatura do mês anterior;

**VL\_FATURA:** Valor da Fatura atual.

Baseado nessas informações, responda:

* 1. **–** Qual o percentual de faturas emitidas por mês no qual os clientes não pagaram a fatura anterior?

O gráfico abaixo relaciona o percentual de faturas emitidas por mês no qual os clientes não pagaram a fatura anterior com os meses. Sendo os indicativos dos meses em números se referindo de forma ordenada aos meses do nosso calendário (1=JAN, 2=FEV,...).



**CÓDIGO GITHUB**

**1.2 –** Tendo como referência todos os clientes que tiveram fatura emitida no mês de setembro, gere uma base para esses clientes com os seguintes calculados:

• **Total de faturas emitidas** nos últimos 6 meses (sem contar com a fatura de setembro);

• **O valor médio de fatura** nos últimos 6 meses (sem contar com a fatura de setembro);

• **Quantidade de vezes que ele ficou sem pagar a fatura anterior** nos últimos 6 meses (sem contar com a fatura de setembro).

Segue exemplo:

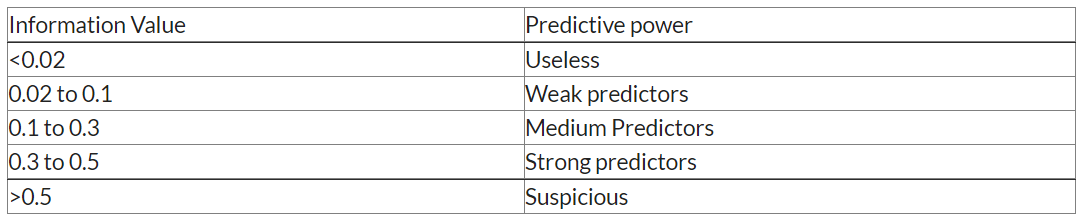


**CÓDIGO GITHUB**

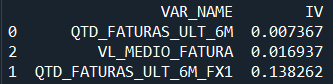
**1.3 –** Utilizando como referência a base calculada na questão anterior, identifique qual das 3 variáveis calculadas tem o maior potencial de preditivo em relação a variável DS\_ROLAGEM do mês de setembro.

**CÓDIGO GITHUB**

A variável com maior potencial preditivo será a que tiver o seu *Information Value* mais de acordo com a seguinte tabela.

****

Após a execução do código temos como IV para cada uma das variáveis os seguintes valores da Figura abaixo.



QTD\_FATURAS\_ULT\_6M\_FX1 é uma variável que possui um potencial preditivo médio

VL\_MEDIO\_FATURA é uma variável que possui um potencial preditivo inútil

QTD\_FATURAS\_ULT\_6M é uma variável que possui um potencial preditivo inútil.

Dessa forma, QTD\_FATURAS\_ULT\_6M\_FX1 é a variável que apresenta o melhor potencial preditivo.

## Questão 2

O ciclo de crédito do cartão Fortbrasil consiste em 3 etapas: Aquisição, Manutenção e Cobrança. O cliente entra em cobrança com 5 dias de atraso, e quando atinge a marca de 65 dias tem o cartão cancelado e então para realizar o pagamento da dívida precisa realizar um acordo. A FortBrasil irá fazer uma nova campanha para aumentar a adesão aos acordos e gostaria de traçar uma estratégia de cobrança utilizando as informações históricas e técnicas de modelagem preditivas. Segue abaixo as definições das variáveis:

**ID\_CONTA:** Identificador da conta;

**DT\_ACORDO:** Data do acordo;

**NU\_DIAS\_ATRASO:** Número de dias após o vencimento que esse cliente está sem pagar;

**VALOR\_CRELIQ:** Valor da dívida do cliente quando ele ultrapassa os 65 dias de atraso;

**DIVIDA\_ATUAL:** Dívida atualizada com juros;

**RESPOSTA:** Se for 1, significa que o cliente aderiu aquele acordo. Ou seja, efetuou o pagamento da entrada. Se for 0, caso contrário.

**QTD\_PARCELAMENTO\_XM**: Quantidade de parcelamentos de fatura aderidos X meses antes da fatura que ele deixou de pagar.

**LIMITE:** Limite total do cliente;

**QTD\_EXTRATOS:** Quantidade de Faturas emitidas;

**QTD\_FX0:** Quantidade de faturas emitidas em dia;

**QTD\_FX1:** Quantidade de faturas emitidas no qual o cliente não pagou a fatura do mês anterior;

**QTD\_FX2:** Quantidade de faturas emitidas no qual o cliente não pagou as duas ultimas faturas anteriores;

**QTD\_ACIONAMENTO:** Quantidade de vezes que ligamos para os clientes para cobrar a sua dívida;

**QTD\_CP:** Quantidade de vezes que a ligação foi atendida, mas não houve contato com o cliente;

**QTD\_CPC:** Quantidade de vezes que o próprio cliente atendeu a ligação de cobrança.

**DEFINIÇÃO**: Acordo é uma proposta de negociação da dívida por parte da Fortbrasil para clientes que estão com mais de 180 dias de atraso.

Baseado nessas informações, responda:

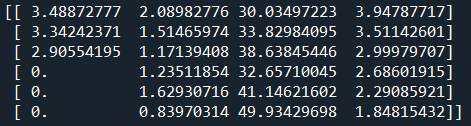
**2.1** – Qual o percentual de adesão mensal por faixa de atraso (Histórico)?

**CÓDIGO GITHUB**

Foi necessário estabelecer algumas faixas de atraso para a realização da questão. Dessa forma, optou-se por usar a variável NU\_DIAS\_ATRASO para o estabelecimento dessas faixas. Sendo assim, temos as seguintes faixa de atraso:

* Faixa label 0: Faixa de atraso de 181 até 240 dias;
* Faixa label 1: Faixa de atraso de 241 dias até 300 dias;
* Faixa label 2: Faixa de atraso de 301 dias até 360 dias;
* Faixa label 3: Faixa de atraso de 361 dias até 420 dias;
* Faixa label 4: Faixa de atraso de 421 dias até 480 dias;
* Faixa label 5: Faixa de atraso de 481 dias até 540 dias.

O percentual de adesão mensal para cada uma das faixas de atraso obtidos está representado na seguinte matriz

****

Em que as linhas indicam a faixa de atraso (1ª Linha = 1º faixa de atraso, 2ª linha = 2ª faixa de atraso), enquanto as colunas indicam os meses de novembro/2018, março/2019, abril/2019 e junho/2019. Ou seja, teríamos que para o mês de novembro de 2018 uma adesão mensal de 3,48% em relação à faixa de atraso de 181 até 240 dias.

**2.2** – Qual modelo você utilizaria para traçar uma estratégia objetivando o aumento da adesão dos acordos? (Descreva a técnica utilizada)

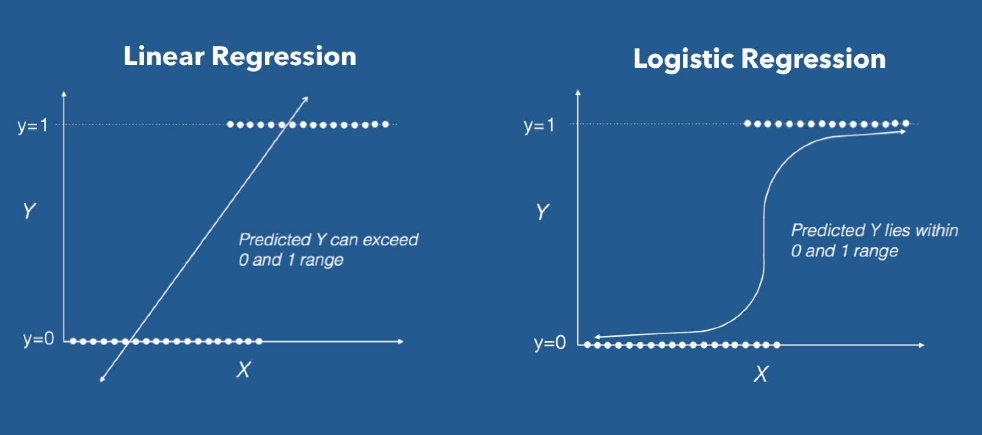
O modelo utilizado será o modelo de regressão logística.

O objetivo da estratégia é aumentar a adesão de acordos. É sabido que a variável RESPOSTA é a variável responsável por indicar se um cliente aderiu ao acordo (“1”) ou se não aderiu ao acordo (“0”). Por a variável RESPOSTA ser uma variável categórica (valores sendo 1 ou 0), o modelo de regressão logística é adequado a ser utilizado para traçar a estratégia alvo, pois com esse modelo é possível obter valores preditivos no que se refere à probabilidade de um cliente aderir ou não ao acordo.

É válido comparar o modelo da regressão linear com o modelo da regressão logística para fazer alguns apontamentos:

* Enquanto a regressão logística se caracteriza por ter uma curva em forma de uma reta, a regressão logística tem uma curva em um formato de “S”. A reta fornecida pela regressão logística não é útil para modelos com saídas discretas (categóricas, conforme mencionado mais acima), pois a saída desses modelos são apenas 0 ou 1 e a reta fornecida pela regressão linear pode exceder esses valores.
* Já a curva fornecida pela regressão logística é chamada de sigmoide e apresenta esse formato parecido com um “S”. Essa curva se adequa bem aos dados discretos, não os excedendo. Na Figura 1 é mostrada uma curva característica dos dois tipos de regressão para elucidar mais claramente o porquê de utilizar o modelo de regressão logística

Figura 1 - Regressão Linear (Linear Regression) x Regressão Logística (Logistic Regression)



Após essa breve explicação sobre a regressão logística e alguns motivos para ela ser utilizada, pode-se entrar em mais detalhes em relação ao modelo. A curva de saída tem como equação: .

Em que:

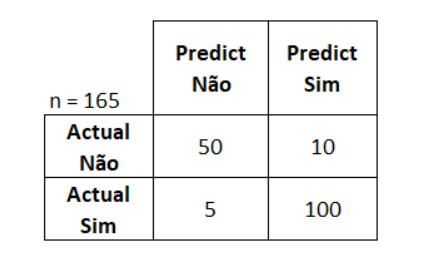
* ;
* é o número de variáveis independentes.

Através dessa função temos que para qualquer valor de x pertencente ao conjunto dos números reais irá me dar como saída um valor entre 0 e 1. Também é usado um ponto de *cutoff* para fazer com que os valores de saída fiquem sendo exatamente 0 ou 1. Geralmente esse *cutoff* é de 0,5. Dessa forma, o modelo tem como saída 1 todos os valores que estão acima de 0,5, enquanto os valores abaixo ou igual a 0,5 irá nos dar como predição uma saída de 0.

O modelo de regressão logística final pode ser simplificado até chegar nessa fórmula: . Que nos dá como saída uma probabilidade.

**2.3** – Quais indicadores e ferramentas você utilizaria para avaliar a performance/aderência desse modelo? (Descreva os indicadores utilizados)

Apesar de não ser usada (embora pudesse), irei descrever uma matriz de confusão por conta de sua nomenclatura que facilita o entendimento dos outros indicadores.



Pode-se observar que a matriz de confusão usada como exemplo tem como características:

* ser de ordem 2 (possui 2 linhas e 2 colunas);
* O valor *n = 165* se refere a quantidade de observações que temos, ou seja, o total de predições feitas pelo modelo responsável por gerar essa matriz de confusão é igual a 165;
* Possuímos, no caso, como variáveis preditas duas variáveis “Sim” e “Não”;
* Por exemplo, se o exemplo se trata de uma concessão de crédito a determinado cliente temos que: o modelo obteve “Sim” como saída para predição 110 vezes, enquanto obteve “Não” como saída 55 vezes. Das 110 predições com “Sim”, 100 delas estavam corretas (Actual Sim x Predict Sim) , enquanto 10 estavam erradas. Das 55 predições com “Não”, 50 delas estavam corretas, enquanto 5 estavam erradas.

Usando a nomenclatura oficial da matriz de confusão temos que:

* Verdadeiro Positivos (VP): Se referem às observações que o modelo conseguiu prever a saída “Sim” corretamente. No exemplo dado acima, podemos dizer que 100 clientes são VP;
* Verdadeiro Negativos (VN): Se referem às observações que o modelo conseguiu prever a saída “Não” corretamente. No exemplo dado acima, podemos dizer que 50 clientes são VN;
* Falso Positivos (FP): Se referem às observações que o modelo classificou como “Sim”, mas na verdade eram “Não”. No exemplo dado acima, podemos dizer que 10 clientes são FP;
* Falso Negativos (FN): Se referem às observações que o modelo classificou como “Não”, mas na verdade eram “Sim”. No exemplo dado acima, podemos dizer que 5 clientes são FP.

Existem diversos indicadores que podem ser utilizados para avaliar o desempenho do modelo, dentre eles pode-se destacar mais especificamente alguns:

* Acurácia

Indicador que nos dá a relação entre as saídas que foram preditas corretamente pelo modelo e o total de observação. Basicamente, é uma taxa obtida da divisão do número de acertos (VP + VN) pelo número total de observações (VP+VN+FN+FP). O seu uso é recomendado para problemas em que temos um equilíbrio entre as variáveis categóricas (classes). Já para modelos em que não existe esse equilíbrio entre as classes, a acurácia pode dar uma impressão falsa no que se refere ao seu desempenho (*overfitting*).

* Precisão

Número de exemplos classificados como pertencentes a uma classe, que realmente são daquela classe (VP), dividido pela soma entre este número, e o número de exemplos classificados nesta classe, mas que pertencem a outras (FP).

* Recall

Número de exemplos classificados como pertencentes a uma classe, que realmente são daquela classe, dividido pela quantidade total de exemplos que pertencem a esta classe, mesmo que sejam classificados em outra. Ou seja, VP dividido por (VP+FP).

* AUC - Area Under the ROC Curve

Nela, mede-se a área sob uma curva formada pelo gráfico entre a taxa de exemplos positivos, que realmente são positivos, e a taxa de falsos positivos. Acaba sendo uma métrica interessante para modelos com uma desproporcionalidade entre as classes.

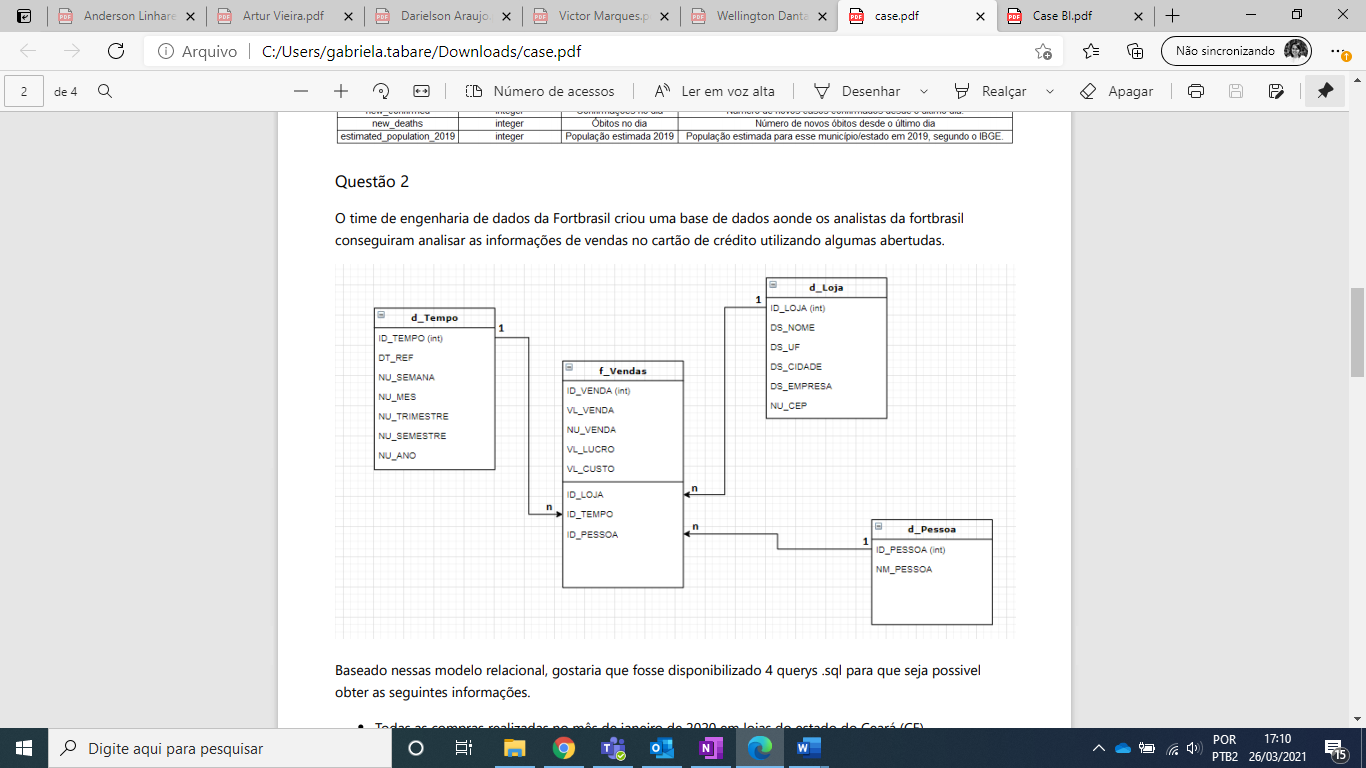
**2.4** – Apresente o modelo desenvolvido utilizando a técnica do item (2.2) e as técnicas de avaliação descritas no item (2.3).

**CÓDIGO GITHUB**

**2.5 –** Crie um relatório analítico no Power BI para acompanhar a adesão dos acordos e criar insights de como melhorar essa adesão

## Questão 3

O time de engenharia de dados da Fortbrasil criou uma base de dados onde os analistas conseguiram avaliar as informações de vendas no cartão de crédito utilizando algumas aberturas:



Baseado nesse modelo relacional, disponibilize 4 querys .sql para que seja possível obter as seguintes informações:

**3.1** - Todas as compras realizadas no mês de janeiro de 2020 em lojas do estado do Ceará (CE)

* ID da pessoa
* Nome da pessoa
* Data Referência da Venda
* Valor da Venda

**CÓDIGO GITHUB**

**3.2** - Quantidade de compras por cliente no mês de março de 2020

* ID da pessoa
* Quantidade de compras

**CÓDIGO GITHUB**

**3.3** – Todos os clientes que não fizeram compras no mês de março de 2020

**CÓDIGO GITHUB**

**3.4** – Data da última compra por cliente

**CÓDIGO GITHUB**