

## Projeto 1: Prevendo Demanda de um Catálogo

### Passo 1: Compreensão do Negócio e dos Dados

#### Decisões Chaves:

*Responda estas perguntas*

**1. Que decisões precisam ser feitas??**

Precisamos identificar através da probabilidade, qual é a receita estimada para as receitas de vendas pelos catálogos.

Identificando que tipo de informações temos disponíveis em nosso banco de dados, e como podemos utilizá-la para gerar informações uteis na elaboração do nosso modelo preditivo, para que assim possamos começar a nossa análise de regressão linear para prever o comportamento das nossas vendas. E se o investimento nos catálogos irá trazer um retorno acima do esperado.

**2. Que dados são necessários para subsidiar essas decisões??**

No modelo de regressão iremos precisar das seguintes informações:  
Avg.Sale.Amount ~ Customer.Segment + Avg.Num.Products.Purchased

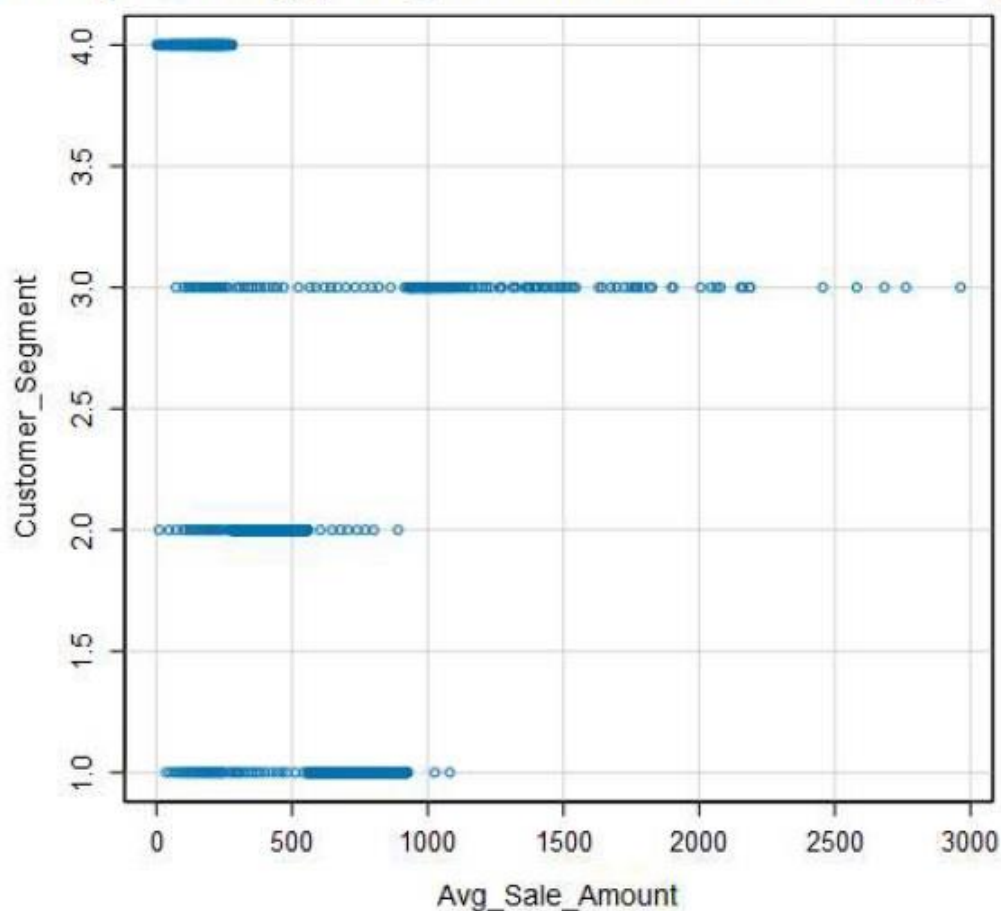
### Passo 2: Análise, modelagem e validação

- 1. Como e por que você selecionou [as variáveis de previsão \(veja texto suplementar\)](#) em seu modelo? Você deve explicar como as variáveis de previsão contínuas que você escolheu têm uma relação linear com a variável-alvo. Consulte esta [lição](#) para ajudar você a explorar seus dados e usar gráficos de dispersão para procurar relações lineares. Você deve incluir gráficos de dispersão em sua resposta.**

Utilizei as Vendas Médias como a variável a ser prevista. O tipo de segmento e a média de produtos vendidos como variáveis explicativas. As demais variáveis foram descartadas, pois sua utilização além de não ter um grande peso no modelo, seus p-valores não eram significativos a 95% de confiança estatística, e pouca correlação com a variável que estamos tentando prever.

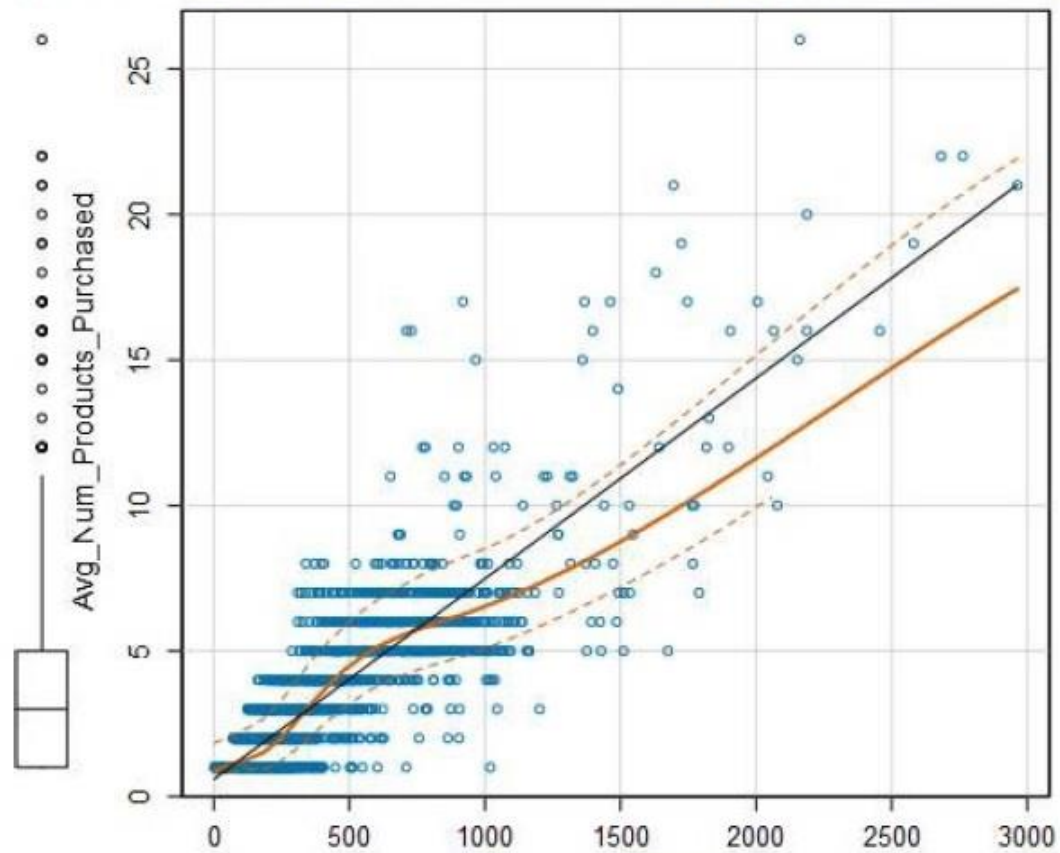
No gráfico abaixo, vemos que a variável referente ao segmento `Loyalty Club Only`, é a variável que tem o melhor desempenho, desde as médias de vendas mais baixas, até as mais elevadas, ou seja, é um indicativo claro que a loja deve investir principalmente nesse segmento, pois além de abranger todos os públicos de renda, é a que mais traz resultados para a empresa. Enquanto que a variável `Store Mailing List`, é a variável que teve a menor média de vendas, e também a menor quantidade de vendas.

**Scatterplot of Avg\_Sale\_Amount versus Customer\_Segm**



Abaixo trazemos um segundo gráfico que foi utilizado para embasar a decisão de utilizar essas variáveis no modelo de regressão. É a relação entre AVG Num Product Purchased e AVG Sale Amount. Ou seja, é a média de produtos vendidos em e a Média de Vendas.

Scatterplot of Avg\_Sale\_Amount versus Avg\_Num\_Products\_P



Podemos facilmente identificar que há uma certa correlação positiva entre as variáveis. Porém existem muitos pontos que estão longe da reta de regressão linear, isso indica que a predição do modelo tende a não ser tão eficiente. Principalmente, quanto maior a média de compras menos efetivo se mostra a predição da média de vendas.

2. **Explique por que você acredita que seu modelo linear é um bom modelo. Você deve justificar o seu raciocínio usando os resultados estatísticos criados pelo seu modelo de regressão. Para cada variável selecionada, por favor justificar por que cada variável é uma boa opção para o seu modelo, usando os valores-p e valores R-quadrado produzidos pelo seu modelo.**

O arquivo customer apresenta os dados de 3000 clientes que incluem nome e informação de contato, customer segment (segmento de compra), customer ID, average sales (média de vendas), store number (número da loja), se eles responderam ao último catalogo, média de número de produtos comprados e número de anos deles como consumidor.

O meu modelo, é um bom modelo dentro daquilo que extrai de informações dos arquivos. Os P- Valores todos ficaram muito abaixo dos 0,05, indicando que todos tem significância estatística a níveis de 95% de confiança. O  $R^2$  0,8369 e o  $R^2$  ajustado em 0,8366. É um

bom modelo preditivo. Conforme esse modelo for sendo treinado, esse resultado pode até melhorar.

As variáveis utilizadas são uma boa opção para o modelo, pois elas tem relação com a quantidade de vendas médias. Outras variáveis poderiam ser usadas, como por exemplo, as cidades, porém essas foram descartadas pois não havia significância estatística que sustentasse a sua utilização, e por isso foram descartadas.

- 3. Qual é a melhor equação de regressão linear com base nos dados disponíveis? Cada coeficiente não deve ter mais de 2 dígitos após o decimal (ex: 1,28)**

**Importante: A equação de regressão deve estar na forma:**

```
Y = 303,46 - 149,36 * Customer.SegmentLoyalty Club Only + 281,84 *  
Customer.SegmentLoyalty Club and Credit Card + -245,42 * Customer.SegmentStore  
Mailing List + 0 * Store_mailing_list + 66,98 * Avg.Num.Products.Purchased
```

## Passo 3: Apresentação/Visualização

*Use os resultados do modelo para fornecer uma recomendação. (limite de 500 palavras)*

*No mínimo, responder à estas perguntas:*

- 1. Qual é a sua recomendação? A empresa deve enviar o catálogo para estes 250 clientes?**

Sim a empresa deve enviar o catálogo para os clientes, pois segundo a previsão do modelo o lucro está acima do lucro esperado pela gerencia que é de U\$10.000,00.

- 2. Como você chegou na sua recomendação? (Por favor, explique a sua lógica para os revisores poderem lhe dar feedback sobre o seu processo)**

Com base nos valores do modelo treino, utilizamos essas variáveis para testar e estimar a média das vendas para esses 250 clientes. Após isso, multipliquei o valor previsto pela probabilidade de o cliente comprar pelo catálogo. Essa nova variável é a Receita Esperada, subtraí, os 50% de margem, e também o custo de U\$6,50 \* 250 catálogos. O Resultado foi superior ao valor mínimo esperado de lucro que foi definido pela gerência.

- 3. Qual é o lucro esperado do novo catálogo (assumindo que o catálogo é enviado para estes 250 clientes)?**

O lucro esperado é de U\$21.987,43 dólares.