Prevendo Demanda de um Catálogo

Compreensão do Negócio e dos Dados

1. Que decisões precisam ser feitas??

A decisão que o gerente precisa tomar é se irá enviar os catálogos para os novos 250 clientes baseados no modelo preditivo que irei construir

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

Primeiramente será necessário criar um modelo preditivo em "*p1-customers"* usando o customer_segment como variável alvo e como variáveis preditoras o average_number_of_product_purchased, score_yes, margin and cost_of_catalog. Posteriormente aplicar esse modelo na lista dos 250 novos clientes.

Análise, modelagem e validação

1. Como e por que você selecionou as variáveis de previsão em seu modelo?

O estudo da regressão linear foi feito em todas as variáveis em relação a *Avg_Sale_Amount (variável alvo)*, como mostra a figura 1, apenas *Avg_Num_Products_Purchased e Customer_Segment* tem o p-valor menor que 0.05, o que implica em uma significância estatística boa. Já a figura 2 mostra o scatterplot entre as variáveis e a confirmação da linearidade.

Report

Report for Linear Model Regressão_linear_variables_

Basic Summary

Call:

lm(formula = Avg.Sale.Amount ~ Customer.Segment + Store.Number + Responded.to.Last.Catalog + Avg.Num.Products.Purchased + X..Years.as.Customer, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-665.19	-67.82	-2.17	70.42	975.25

Coeficientes:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	435.318	104.854	4.152	3e-05 ***
Customer.SegmentLoyalty Club Only	-150.224	8.971	-16.746	< 2.2e-16 ***
Customer.SegmentLoyalty Club and Credit Card	282.455	11.897	23.743	< 2.2e-16 ***
Customer.SegmentStore Mailing List	-243.279	9.816	-24.784	< 2.2e-16 ***
Store.Number	-1.146	0.994	-1.153	0.2489
Responded.to.Last.CatalogYes	-28.085	11.253	-2.496	0.01264 *
Avg.Num.Products.Purchased	66.787	1.515	44.082	< 2.2e-16 ***
X..Years.as.Customer	-2.326	1.222	-1.904	0.05707 .

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Erro padrão residual: 137.25 em 2367 graus de liberdade
R quadrada múltipla: 0.8376, R quadrada ajustada: 0.8372
F estatístico: 1745 em 7 e 2367 graus de liberdade (DF), valor p < 2.2e-16

Type II ANOVA Analysis

Response: Avg.Sale.Amount

	Sum Sq	DF	F value	Pr(>F)
Customer.Segment	28425698.84	3	503.02	< 2.2e-16 ***
Store.Number	25054.95	1	1.33	0.2489
Responded.to.Last.Catalog	117319.82	1	6.23	0.01264 *
Avg.Num.Products.Purchased	36603783.15	1	1943.23	< 2.2e-16 ***
X..Years.as.Customer	68263.47	1	3.62	0.05707 .
Residuals	44586209.93	2367		

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Figura 1 Report do modelo preditivo - variaveis

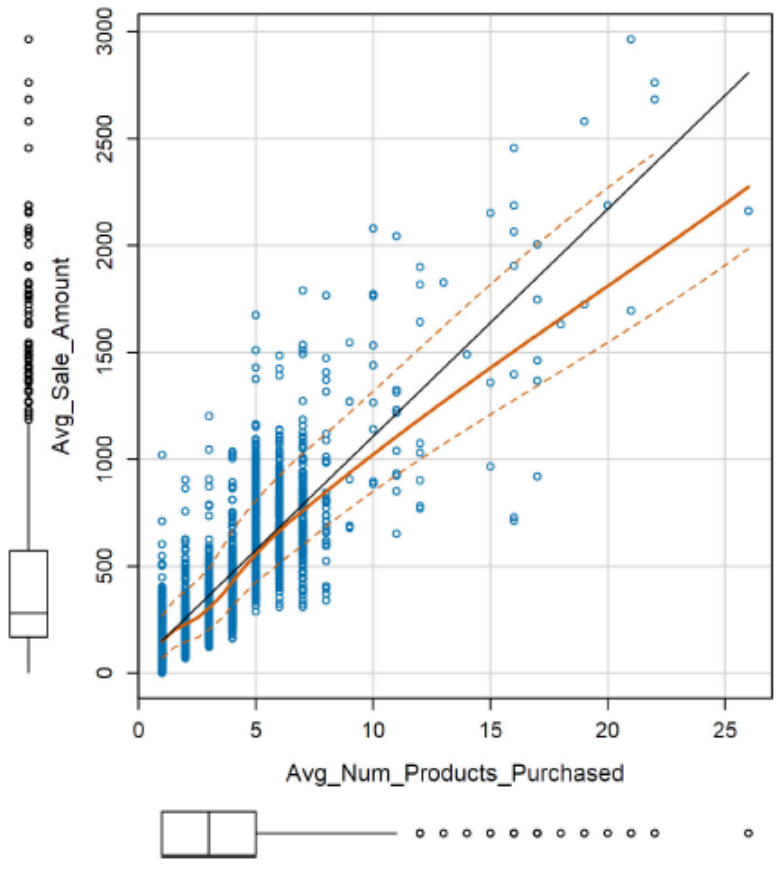

Figura 2 - Gráfico de disperssão entre Quantidade média de vendas e Média de produtos comprados.

2. Explique por que você acredita que seu modelo linear é um bom modelo.

- O valor-p “também chamado de nível descritivo ou probabilidade de significância”[1], é a probabilidade de se alcançar uma estatística de teste igual ou mais extrema que aquela avisada em uma amostra com a hipótese nula, ou seja, é a probabilidade de que o coeficiente seja zero. Para confiar na estimativa do coeficiente o valor-p deve estar abaixo 0.05, que no modelo os valores-p foram da ordem de 2.2^{-16} que indica uma relação entre os coeficientes e a variável alvo, sendo assim considera estatisticamente significativa.
- A significância estatística, "A significância estatística é um resultado que não é susceptível de ocorrer aleatoriamente, mas é provável que seja atribuível a uma causa específica"[2]. Na figura 2 mostra que todos os coeficientes têm três asteriscos (*) o que mostra uma boa significância estatística
- R^2 varia de 0 a 1 e representa a quantidade de variação na variável alvo explicada pela variação nas variáveis preditoras, ou seja, indica o quão bem os dados se encaixam na linha de tendências. Quanto maior o R^2, maior o poder explicativo do modelo. Ainda na Figura 2 o R^2 assume o valor de: 0.8369 o que mostra um poder explicativo aceitável, já que falta 0.1631 para atingir 1 na escala.

```
lm(formula = Avg.Sale.Amount ~ Customer.Segment + Avg.Num.Products.Purchased, data = the.data)

Residuals:
     Min       1Q   Median       3Q      Max 
-663.800  -67.300   -1.900   70.700  971.700 

Coeficientes:
                                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                               303.460     10.576   28.69  < 2e-16 ***
Customer.SegmentLoyalty Club Only        -149.360      8.973  -16.65  < 2e-16 ***
Customer.SegmentLoyalty Club and Credit Card 281.840  11.910   23.66  < 2e-16 ***
Customer.SegmentStore Mailing List       -245.420      9.768  -25.13  < 2e-16 ***
Avg.Num.Products.Purchased                 66.980      1.515   44.21  < 2e-16 ***
---
Significance codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Erro padrão residual: 137.48 em 2370 graus de liberdade
R quadrada múltipla: 0.8369,	R quadrada ajustada: 0.8366 
F estatístico: 3040 em 4 e 2370 graus de liberdade (DF), valor p < 2.2e-16
```

Figura 3 - Relatório Estatístico do Modelo Preditivo.

[1] Glossário Inglês-Português de Estatística, Sociedade Portuguesa de Estatística e Associação Brasileira de Estatística
[2] Investopedia

3. Qual é a melhor equação de regressão linear com base nos dados disponíveis?

1	*Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3*
2	$Y = 303.46 + b_1 * (\text{Loyalty Club Only}) + b_2 * (\text{Loyalty Club and Credit Card}) + b_3 * (\text{Store Mailing List}) + b_4 * (\text{Avg.Num.Products.Purchased})$
3	Y = 303.46 + -149.36*(Loyalty Club Only) + 281.84 *(Loyalty Club and Credit Card) + (-245.42)*(Store Mailing List) + 66.98*(Avg.Num.Products.Purchased)

Apresentação/Visualização

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

Baseado no modelo preditivo, a empresa deve sim enviar o catálogo para os novos 250 clientes.

2. Como você chegou na sua recomendação?

Cheguei a essa conclusão, porque o modelo mostrou-se estatisticamente aceitável em todos os requisitos (R2, valor-p e significância estatística), e o lucro previsto é positivo, o que mostra ser interessante para a empresa.

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

Lucro esperado = (soma da receita esperada x margem bruta) - (custo do catálogo x 250)

= (47,225.87 x 0.5) – (6.50 x 250)

= 23,612.44 – 1,625

= $21,987.44

Distribuição de Variáveis

Variáveis como endereço, nome, Estado, ID do cliente, número da loja e CEP não são variáveis preditoras importantes, pois são únicas para cada valor ou irrelevantes na previsão da venda usando o bom senso; Já Cidade, quem respondeu ao último catálogo e número de anos como cliente pode parecer um bom indicador, pois não é um ID exclusivo, e o modelo de regressão linear mostrou que eles são estatisticamente insignificantes.

Mais dados da categoria de itens comprados, duração da rotatividade de itens serão úteis para entender o comportamento de compra do cliente, onde podemos explorá-lo para segmentar nossos clientes e personalizar o catálogo.

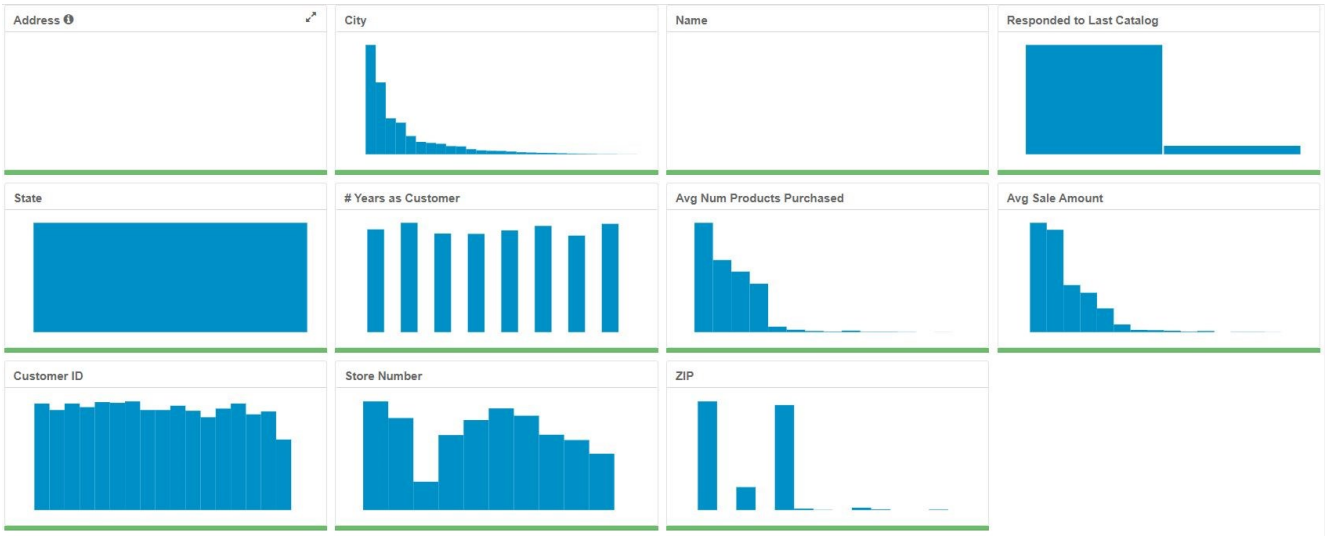

Figura 4 Distribuição de cada variável no dataset P1-customers

Alteryx Workflow