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??

- Temos que decidir se os catálogos serão enviados, e para isso os lucros esperados devem ser superior a US$10.000. A empresa tem na sua lista de email, 250 novos clientes para quem eles querem enviar o catálogo. A nossa solução tem que prover uma analise preditiva calculando a probabilidade de lucro para ajudar na tomada de decisão.

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

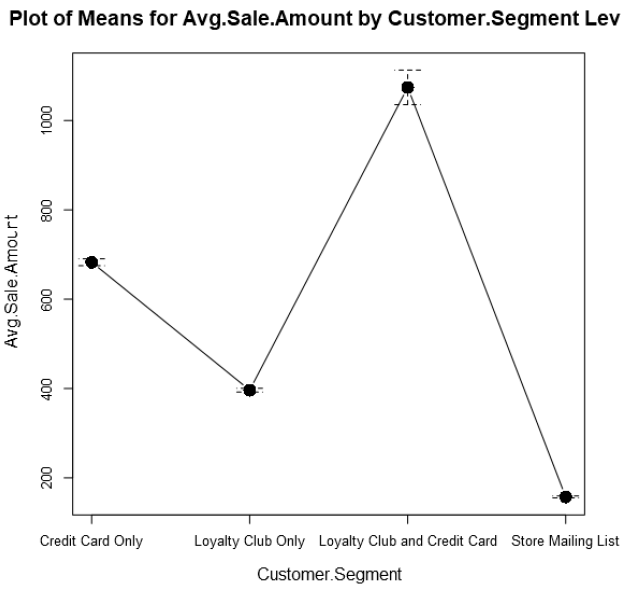
- O arquivo **p1-customers.xlsx** contem as informações dos 2300 clientes, o nosso modelo preditivo será construído com este conjunto, o arquivo **p1-mailinglist.xlsx** contem os dados para os quais vamos a prever as vendas, ele contem os campos Score_No: A probabilidade de que o cliente **NÃO VAI** responder ao catálogo e não vai fazer uma compra, e Score_Yes: A probabilidade de que o cliente **VAI** responder ao catálogo e fazer uma compra. O custo de impressão e distribuição é de US$6,50 por catálogo. A margem bruta média (preço - custo) de todos os produtos vendidos através do catálogo é 50%

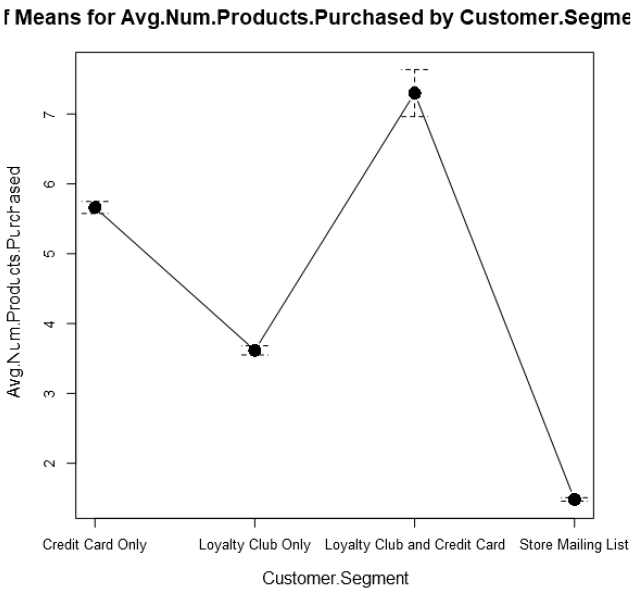
Passo 2: Análise, modelagem e validação

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.

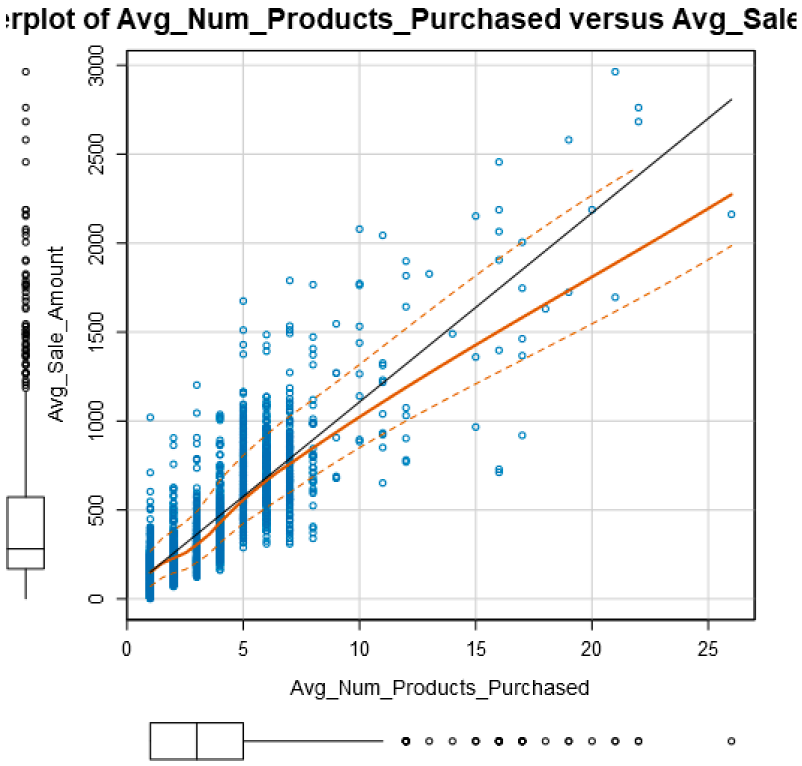
Record #	FieldName	Avg Sale Amount	Avg Num Products Purchased	# Years as Customer
1	Avg Sale Amount	1	0.855754	0.029782
2	Avg Num Products Purchased	0.855754	1	0.043346
3	# Years as Customer	0.029782	0.043346	1

- Em este gráfico podemos apreciar que a variável Avg_Sale_Amount tem uma relação forte com Avg_Num_Products_Purchased, ja com # Years_as_Customer a relação não e boa.





- Podemos observar que a variável categórica Customer_Segment estabelece uma forte relação com Avg_Sale_Amount, os clientes do segmento "Loyalty Club and Credit Card" são os que mais produtos compram e tem maior media de compras por produto.



- Em este gráfico de dispersão entre a variável preditora numérica, no caso Avg_Num_products_Purchased, e a nossa variável alvo Avg_Sales, podemos demonstrar visualmente a correlação entre as duas variáveis, podemos dizer então que se o cliente compra mais produtos a probabilidade da sua media de compra aumenta linearmente.

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.

Report

Report for Linear Model Sales_Forecast

Basic Summary

Call:
lm(formula = Avg.Sale.Amount ~ Customer.Segment + Avg.Num.Products.Purchased, data = inputs$the.data)

Residuals:

Min	1Q	Median	3Q	Max
-663.8	-67.3	-1.9	70.7	971.7

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	303.46	10.576	28.69	< 2.2e-16	***
Customer.SegmentLoyalty Club Only	-149.36	8.973	-16.65	< 2.2e-16	***
Customer.SegmentLoyalty Club and Credit Card	281.84	11.910	23.66	< 2.2e-16	***
Customer.SegmentStore Mailing List	-245.42	9.768	-25.13	< 2.2e-16	***
Avg.Num.Products.Purchased	66.98	1.515	44.21	< 2.2e-16	***

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

Residual standard error: 137.48 on 2370 degrees of freedom
Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366
F-statistic: 3040 on 4 and 2370 DF, p-value: < 2.2e-16

Type II ANOVA Analysis

Response: Avg.Sale.Amount

	Sum Sq	DF	F value	Pr(>F)	
Customer.Segment	28715078.96	3	506.4	< 2.2e-16	***
Avg.Num.Products.Purchased	36939582.5	1	1954.31	< 2.2e-16	***
Residuals	44796869.07	2370			

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

- O p-valor é a probabilidade de que os resultados observados (a estimativa do coeficiente) ocorram por acaso, e que não existe uma relação real entre o preditor e a variável alvo. Em outras palavras, o p-valor é a probabilidade de que o coeficiente seja zero. Quanto menor o p-valor, maior a probabilidade de existir uma relação entre o preditor e a variável alvo. Se o p-valor é alto, não devemos confiar na estimativa do coeficiente. Quando uma variável preditora tem um p-valor abaixo de 0,05, a relação entre ele e a variável alvo é considerada como sendo estatisticamente significativa. No nosso modelo os valores-p são menores que 0.05, por tanto as variáveis são estatisticamente significativas.
- R-quadrado varia de 0 a 1 e representa a quantidade de variação na variável alvo explicada pela variação nas variáveis preditoras. Quanto maior o r-quadrado, maior o poder explicativo do modelo. No nosso modelo, o valor R-quadrado é de 0.8366, então os coeficientes são significantes e o r^2 é relativamente alto (> 0,7), podemos chegar a conclusão que nosso modelo é viável.

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)

Y = 303.46 + 66.98 * Avg_Num_Products_Purchased - 245.42 * (Se Type:Customer_Store_Mailing_List) + 281.84 * (Se Type: Customer_Loyalty_Club_and_Credit_Card) - 149.36 * (Se Type: Customer_Loyalty_Club_Only) + 0 * (Se Type: Credit_Card_Only)

Passo 3: Apresentação/Visualização

1.Qual é a sua recomendação? A empresa deve enviar o catálogo para estes 250 clientes?
Prevendo um lucro de aproximadamente $21987, a empresa deve enviar o catalogo para os 250 clientes.

2.Como você chegou na sua recomendação?
Prevendo as vendas com o nosso modelo linear calculamos a media das vendas e multiplicamos o valor total por a probabilidade de que sejam efetuadas (Score_Yes), depois multiplicamos o resultado por 0.5 e restamos os 6.50 do custo de impressão dos catálogos, assim calculamos o lucro.
Predicted_Sales = Avg_Sales * Score_Yes
Profit = (Predicted_Sales * 0.5) - 6.5

3.Qual é o lucro esperado do novo catálogo (assumindo que o catálogo é enviado para estes 250 clientes)?
Profit = $ 21987.44