Projeto 1: Prevendo Demanda de um Catálogo

Compreensão do Negócio e dos Dados

O projeto visa a subsidiar a decisão de enviar catálogos impressos com produtos da empresa - acessórios para casa e decoração – para 250 novos clientes.

A gerência deseja enviar o catálogo do ano somente se o lucro esperado for superior a US$ 10.000,00; portanto, faz-se necessário estimar o lucro resultante do envio desses catálogos para os novos clientes.

A margem bruta média de todos os produtos vendidos por meio do catálogo é 50% e o custo de impressão e distribuição desse item é U$ 6,50 por unidade. Dessa maneira, é possível estimar o lucro a partir das vendas, por meio da multiplicação desse valor por 0,5 e, a seguir, a dedução do custo total de US$ 1.625,00.

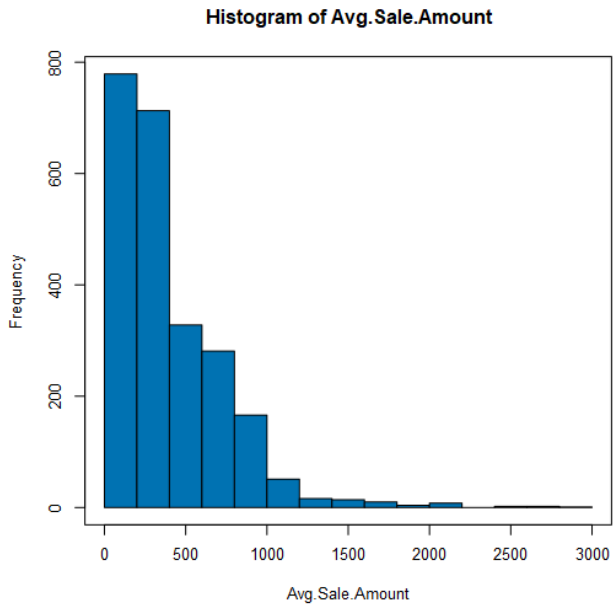
Os dados necessários para a predição almejada são, portanto, o valor de venda de períodos anteriores e outros que estejam disponíveis e possam agregar informação ao modelo.

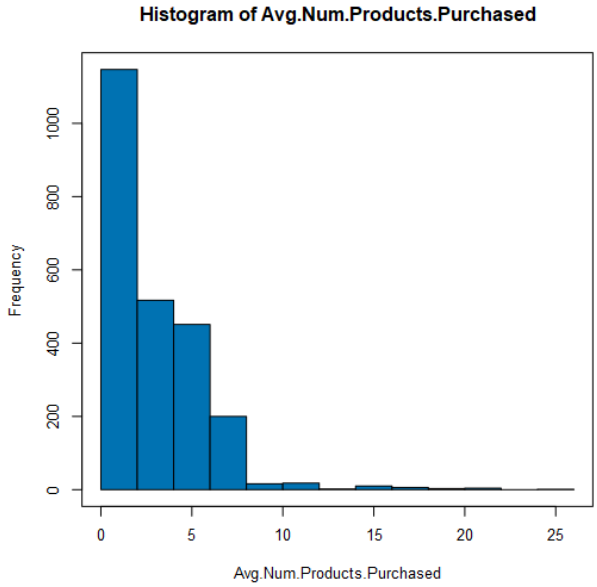
Análise, modelagem e validação

No caso em análise, como o número a ser obtido – lucro esperado – é uma estimativa, a análise a ser feita é a preditiva. Ademais, a média de venda por consumidor (Avg Sale Amount) – variável alvo por meio da qual é calculado o lucro – está disponível; portanto, o modelo é rico em dados. Por outro lado, o resultado buscado – lucro esperado e, por decorrência, as vendas – é um número; em consequência, e considerando que a predição não está baseada em tempo, deve ser utilizado o modelo de regressão linear múltipla.

Os pressupostos desse modelo são linearidade, normalidade, homocedasticidade, inexistência de autocorrelação entre os erros do modelo e ausência de multicolinearidade entre as variáveis preditivas.

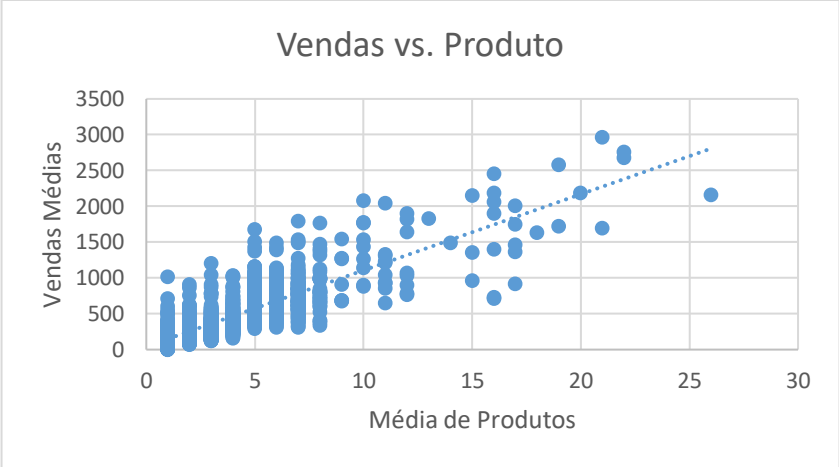
Quanto ao quesito normalidade dos dados, os histogramas que se seguem revelam uma distribuição não normal das variáveis média de vendas e número médio dos produtos comprados, o que pode sugerir alguma inexatidão nos resultados previstos pelo modelo de regressão linear, vez que esse método estatístico assume que os dados seguem uma distribuição normal. Assim sendo, é recomendável a confirmação da adequação do modelo mediante uma análise dos resíduos.



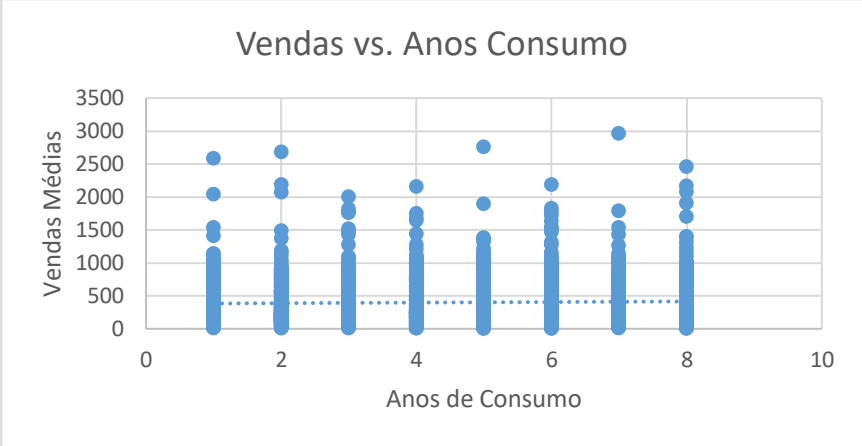


Para verificar se as variáveis numéricas devem ser incluídas no modelo, é necessário verificar se elas possuem uma relação linear com a variável alvo, que é a média de vendas.

O gráfico de dispersão da variável média do número de produtos comprados em relação à variável alvo indica uma forte relação linear. No caso, o aumento no número médio de produtos comprados é acompanhado por uma elevação nas vendas médias:



Já o gráfico de dispersão referente aos anos como consumidor evidencia a ausência de relação de linearidade (linha de tendência plana) com a variável resultado:



As demais variáveis – categóricas - podem ser analisadas por meio da ferramenta Alteryx, a qual, ao compor um modelo de regressão linear múltipla, apresenta o seguinte resultado:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	315.165	11.861	26.571	< 2.2e-16	***
Customer.SegmentLoyalty Club Only	-149.781	8.963	-16.711	< 2.2e-16	***
Customer.SegmentLoyalty Club and Credit Card	282.467	11.897	23.742	< 2.2e-16	***
Customer.SegmentStore Mailing List	-242.842	9.809	-24.756	< 2.2e-16	***
Responded.to.Last.CatalogYes	-27.982	11.254	-2.486	0.01297	*
Avg.Num.Products.Purchased	66.848	1.514	44.147	< 2.2e-16	***
X..Years.as.Customer	-2.313	1.222	-1.893	0.05845	.

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

Residual standard error: 137.26 on 2368 degrees of freedom
Multiple R-squared: 0.8376, Adjusted R-Squared: 0.8371
F-statistic: 2035 on 6 and 2368 degrees of freedom (DF), p-value < 2.2e-16

Esse levantamento evidencia que as variáveis '#Years.as.Customer' e 'Responded.to.Last.Catalogues' apresentam uma significância muito baixa - 0,05845 e 0,01297, respectivamente -, de forma que não devem ser incluídas no modelo.

Retirando essas variáveis, o Alteryx apresenta as seguintes informações:

Basic Summary

Call:

lm(formula = Avg.Sale.Amount ~ Customer.Segment + Avg.Num.Products.Purchased, data = 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 degrees of freedom (DF), p-value < 2.2e-16

Esse levantamento, que considera o segmento do consumidor como uma variável categórica preditiva, demonstra um R-quadrado ajustado de 0,8366, um pouco menor que o R-quadrado de 0,8369. De outro lado, um modelo em função apenas da variável numérica contínua referente ao número médio de produtos comprados (abaixo) apresenta um R-quadrado ajustado de 0,7322 e um R-quadrado de 0,7323. Tendo em vista que ambas estatísticas aumentam com a aludida variável categórica, o modelo deve incluí-la como uma variável preditiva.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	44.02	5.704	7.716	1.75e-14	***
Avg.Num.Products.Purchased	106.28	1.319	80.572	< 2.2e-16	***

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

Residual standard error: 176.01 on 2373 degrees of freedom
Multiple R-squared: 0.7323, Adjusted R-Squared: 0.7322
F-statistic: 6492 on 1 and 2373 degrees of freedom (DF), p-value < 2.2e-16

Por outro lado, o p-valor de 2,2e-16, bem próximo a zero, indica uma baixa probabilidade de que os resultados observados ocorram por acaso, e sugere que existe uma relação real entre as variáveis preditivas e a variável alvo.

Isso posto, a melhor equação de regressão linear com base nos dados disponíveis é:

*Y = 303.46 - 149.36 * Customer_Segment_Loyalty Club Only + 281.84 * Customer_Segment_Loyalty Club and Credit Card - 245.42 * Customer_Segment_Store Mailling List + 0 * Customer_Segment_Cash Only*

Apresentação/Visualização

Recomenda-se que a empresa envie catálogos impressos com seus produtos - acessórios para casa e decoração – para os 250 novos clientes, haja vista que o lucro esperado com tal procedimento é superior a US$ 10.000,00.

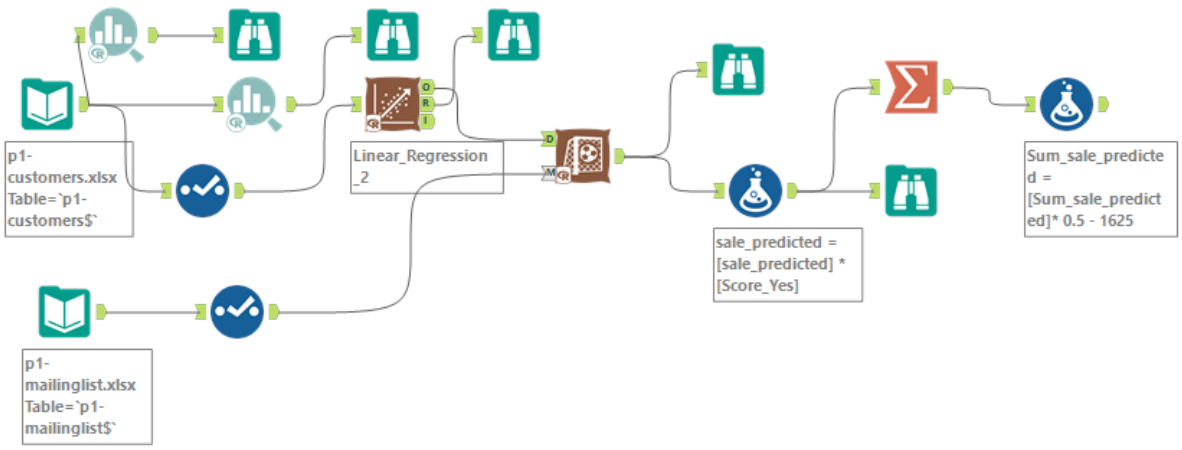
A análise dos dados deu-se mediante a construção de um modelo de regressão linear múltipla para a estimativa da variável alvo atinente às vendas, com a inclusão de variáveis preditivas relacionadas ao segmento do cliente e ao número médio de produtos comprados.

As demais informações disponíveis - '# Years.as.Customer' e 'Responded.to.Last.Catalogues' - apresentaram uma significância muito baixa - 0,05845 e 0,01297, respectivamente -, de forma que não foram incluídas no modelo.

O R-quadrado (coeficiente de determinação) varia entre 0 e 1, indicando, em percentagem, o quanto o modelo consegue explicar os valores observados. Quanto maior o R-quadrado, mais explicativo é o modelo. No caso em tela, o R-quadrado apurado indica que 83,66% da variável dependente é explicada pelas variáveis preditivas presentes no modelo. Ademais, o p-valor de 2,2e-16, bem próximo a zero, indica uma baixa probabilidade de que os resultados observados (a estimativa dos coeficientes) ocorram por acaso.

O lucro estimado resultante do envio dos catálogos para os novos clientes importa em US$ 21.987,44. Esse valor foi obtido a partir do somatório das estimativas de venda por consumidor. Sobre esse montante foi aplicado o percentual de 50% para a apuração da margem bruta média e, desse resultado, foi deduzido o custo de US$ 1.625,00, relativo aos custos de impressão e distribuição dos 250 catálogos (custo unitário de US$ 6,50).

O fluxo utilizado no aplicativo Alteryx consta a seguir:



Fontes:
http://apps.einstein.br/revista/arquivos/PDF/1173-ECv7n1_3-4.pdf
https://pt.wikipedia.org/wiki/R%C2%B2
http://www.portalaction.com.br/analise-de-regressao/analise-dos-residuos
https://help.alteryx.com/current/pt-br/Formula.htm
https://community.alteryx.com/t5/Alteryx-Designer-Discussions/Need-help-with-the-formula-tool-for-a-basic-division-problem/td-p/53941