预测商品册需求

执行概要

关键决策

公司今年是否向邮寄名单中新增的250名客户寄送产品目录册?

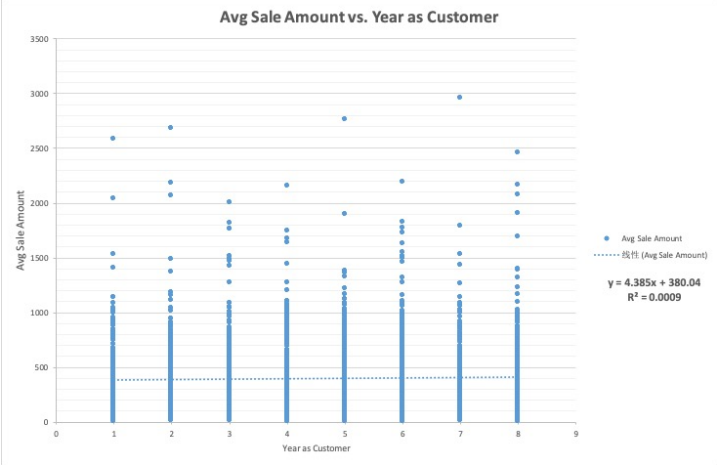
需要获取的数据

数据项	数据名称	数据来源	解释
1	Customer Segment	p1-customers.xlsx	在建模过程中建立虚拟变量
2	Avg Num Products Purchased	p1-customers.xlsx	在建模过程中建立预测变量
3	Avg Sale Amounts	p1-customers.xlsx	在建模过程中建立目标变量
4	Customer Segment	p1-mailinglist.xlsx	在分析过程中用作预测变量
5	Avg Num Products Purchased	p1-mailinglist.xlsx	在分析过程中用作预测变量
6	Avg Sale Amounts	p1-mailinglist.xlsx	在分析过程中用作目标变量
7	Score_Yes	p1-mailinglist.xlsx	顾客会对生产目录有所反应且进行购买的概率
8	平均毛利率50%	业务统计	通过产品目录册出售的所有产品的平均毛利率（价格减去成本）
9	成本6.5$	业务统计	印刷和寄送每本产品目录册成本

分析，建模和验证

1. 通过绘制每个数值型预测变量和目标变量之间的散点图来直观地查看是否有线性关系。
   - Avg Num Products Purchased

   随着平均购买产品数量的增加，平均销售额也以近似线性的趋势增长。



- Year as Customer

随着Year as Customer的增加，不会看到平均销售额增加（或减少）的趋势。可以得出结论，Year as Customer不是初步探索多元线性回归模型的一个很好的变量。

2. 对所有涉及到的单变量进行分析和回归:

数据项	数据分析对象	数据分析过程	数据分析结论
1	Avg Num Products Purchased	SUMMARY OUTPUT 回归统计 Multiple R 0.855754217 R Square 0.73231528 Adjusted R S 0.732202476 标准误差 176.0070633 观测值 2375 方差分析 df SS MS F ignificance F 回归分析 1 201109435.1 201109435.1 6491.906448 0 残差 2373 73511948.03 30978.48632 总计 2374 274621383.1 Coefficients 标准误差 t Stat P-value Lower 95% Upper 95% 下限 95.0% 上限 95.0% Intercept 44.01516317 5.704322669 7.71610684 1.75315E-14 32.82919075 55.20113558 32.82919075 55.20113558 Avg Num Pro 106.2801833 1.319064914 80.57236777 0 103.6935443 108.8668224 103.6935443 108.8668224	R的平方=0.7323>0.7, P<0.05,具有统计显著性，该数据可以参与多变量的线性回归分析
2	Customer Segment	SUMMARY OUTPUT 回归统计 Multiple R 0.9148156 R Square 0.83688758 Adjusted R 0.83618768 标准误差 137.512507 观测值 2366 方差分析 df SS MS F ignificance F 回归分析 5 229065911 45813182.3 3028.419779 0 残差 2361 44645777 18909.6895 总计 2366 273711688 Coefficients 标准误差 t Stat P-value Lower 95% Upper 95% 下限 95.0% 上限 95.0% Intercept 585.799494 14.839192 39.4765089 3.6317E-262 556.700295 614.898693 556.700295 614.898693 Credit Card -281.51137 11.9423762 -23.572476 1.6461E-110 -304.93 -258.09274 -304.93 -258.09274 Loyalty Clu 0 65535 #NUM! 0 0 0 0 Loyalty Clu -431.44536 12.7031437 -33.963668 #NUM! -456.35583 -406.53488 -456.35583 -406.53488 Store Maili -527.65274 13.8732322 -38.033872 2.5671E-247 -554.85773 -500.44776 -554.85773 -500.44776 Avg Num Pro 66.9080778 1.51779131 44.0825281 0 63.9317357 69.8844199 63.9317357 69.8844199	R的平方=0.8369>0.7, P<0.05,具有统计显著性，该数据可以参与多变量的线性回归分析
3	Responded to Last Catalog	SUMMARY OUTPUT 回归统计 Multiple R 0.198931667 R Square 0.039573808 Adjusted R 0.039167536 标准误差 333.46859 观测值 2366 方差分析 df SS MS F ignificance F 回归分析 1 10831813.83 10831813.83 97.40725849 1.52919E-22 残差 2364 262879874.5 111201.3005 总计 2365 273711688.3 Coefficients 标准误差 t Stat P-value Lower 95% Upper 95% 下限 95.0% 上限 95.0% Intercept 417.7003098 7.117667086 58.68500236 0 403.7427925 431.6578271 403.7427925 431.6578271 If Responde -261.301947 26.47567162 -9.86951156 1.52919E-22 -313.219892 -209.384003 -313.219892 -209.384003	R的平方=0.03960<0.3, 不存在线性关系，该数据不可以参与多变量的线性回归分析
4	Year as Customer	SUMMARY OUTPUT 回归统计 Multiple R 0.029781864 R Square 0.000886959 Adjusted R S 0.000465926 标准误差 340.0365645 观测值 2375 方差分析 df SS MS F ignificance F 回归分析 1 243578.0156 243578.0156 2.106623132 0.146794828 残差 2373 274377805.1 115624.8652 总计 2374 274621383.1 Coefficients 标准误差 t Stat P-value Lower 95% Upper 95% 下限 95.0% 上限 95.0% Intercept 380.0388359 15.28292813 24.86688628 1.6908E-121 350.0695612 410.0081105 350.0695612 410.0081105 # Years as C 4.384997179 3.021175081 1.451421073 0.146794828 -1.53941893 10.30941329 -1.53941893 10.30941329	R的平方=0.0009，几乎接近0， 不存在线性关系，该数据不可以参与多变量的线性回归分析

3. 调整的 R 平方为0.8366>（.70）,是强相关模型。对于所选择的预测变量，多元线性回归模型产生的 p 值都低于 0.05 ，所以具有统计学意义。以下是多元线性回归结果：

SUMMARY OUTPUT								
回归统计								
Multiple R	0.9148102							
R Square	0.8368777							
Adjusted R Square	0.8366024							
标准误差	137.48321							
观测值	2375							
方差分析								
	df	SS	MS	F	gnificance F			
回归分析	4	2.3E+08	5.7E+07	3039.744236	0			
残差	2370	44796869	18901.6					
总计	2374	2.75E+08						
	Coefficient	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	303.46	10.57571	28.6944	1.1227E-155	282.7249	324.20208	282.7249	324.20208
Loyalty Club and	281.84	11.90986	23.6643	2.5804E-111	258.4839	305.19358	258.4839	305.19358
Loyalty Club Only	-149.36	8.972755	-16.645	6.34584E-59	-166.951	-131.7605	-166.951	-131.7605
Store Mailing Lis	-245.42	9.767776	-25.125	1.0503E-123	-264.572	-226.2635	-264.572	-226.2635
Avg Num Products	66.98	1.51504	44.2075	0	64.00526	69.947147	64.00526	69.947147

- 根据提供的数据，最佳线性回归方程：

  Y = 303.46 + 66.98* (Avg. Num Products Purchased)+ 281.84 (If Type:Loyalty Club and Credit Card) – 149.36 (If Type: Loyalty Club Only) – 245.42 (If Type: Store Mailing List) + 0 (If Type: Credit Card Only)

建议

公司应该向这250个客户寄送产品目录册。

预测利润

1. 使用线性回归模型来确定每个新客户的平均销售额(Avg Sale Amounts)
   - 平均销售额=303.46 + 66.98* (Avg. Num Products Purchased)+ 281.84 (If Type:Loyalty Club and Credit Card) – 149.36 (If Type: Loyalty Club Only) – 245.42 (If Type: Store Mailing List) + 0 (If Type: Credit Card Only)
2. 接下来，计算每位客户的预测销售额
   - 预测销售额=平均销售额 X 购买商品的概率（Score_Yes）X 毛利率 — 宣传册成本
3. 合计这250名新客户的预测销售额，得出发送新的产品目录的预测利润
   - 预测利润=21987.96 $ >10000 $

因此，公司应该向这250名客户寄送产品目录册。