预测宣传册需求

第 1 步：理解业务和数据

关键决策：

1 需要作出什么样的决策？

1.1 本次项目的目的是预测给 250 个新客户发宣传册预期会带来的收益。

1.2 收益=(总预计收入*毛利率 50%) - 每本宣传册成本 6.5 美元*总人数，关键点是得到总预计收入

1.3 总预计收入需要得到这 250 个人每个人的预计收入；

1.4 需要找出每个人的预计收入由那些因素相关，建立回归模型。

2 作出这些决策需要获取哪些数据？

数据项	数据名称	数据来源	备注
1	Customer Segment	P1-Customers.xlsx	在建模过程中建立虚拟变量
2	Avg Sale Amount	P1-Customers.xlsx	在建模过程中建立目标变量
3	Avg Num Products Purchased	P1-Customers.xlsx	在建模过程中建立自变量
4	Customer Segment	P1-mailinglist.xlsx	应用于所建模型中的虚拟变量
5	Avg Num Products Purchased	P1-mailinglist.xlsx	应用于所建模型中的自变量
6	Score_Yes	P1-mailinglist.xlsx	计算每个人预计收入的概率
7	毛利率	项目详情	毛利率为 50%
8	每本宣传册的成本	项目详情	每本宣传册成本 6.5 美元/本
9	预计宣传册的总人数	P1-mailinglist.xlsx	共 250 人

第 2 步：分析、建模和验证

描述下你是如何设置线性回归模型的，使用了哪些变量，原因是什么，以及模型的结果。建议提供可视化图表（限 500 字以内）。

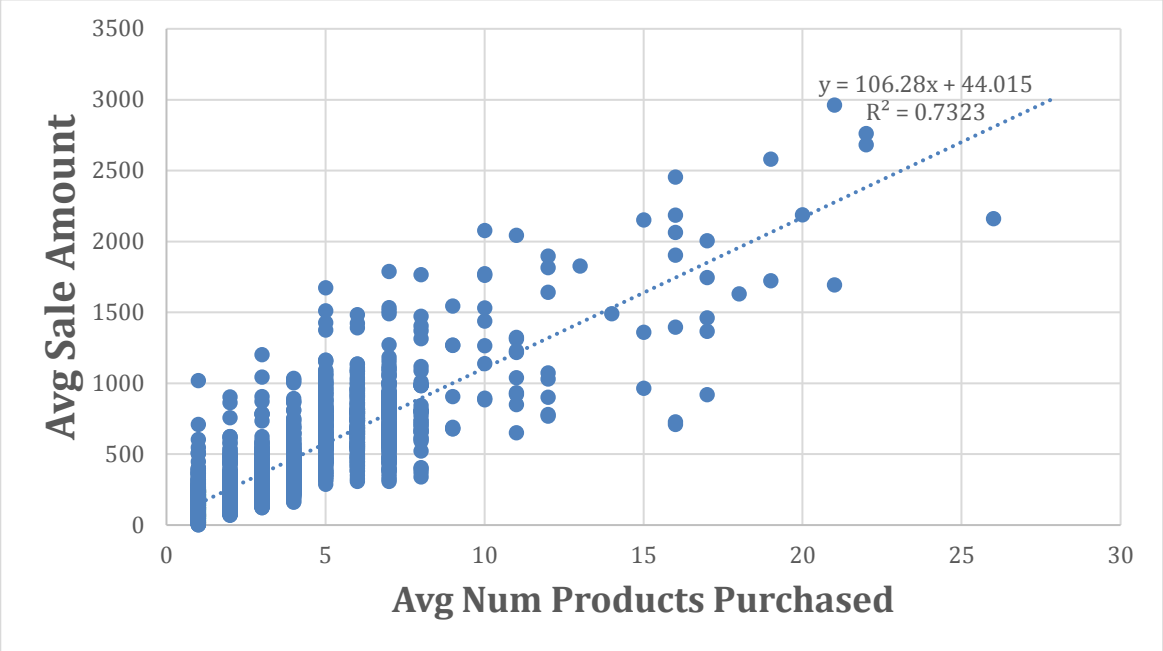
1 分析

1.1 观察确定要目标变量、自变量以及自变量的类型

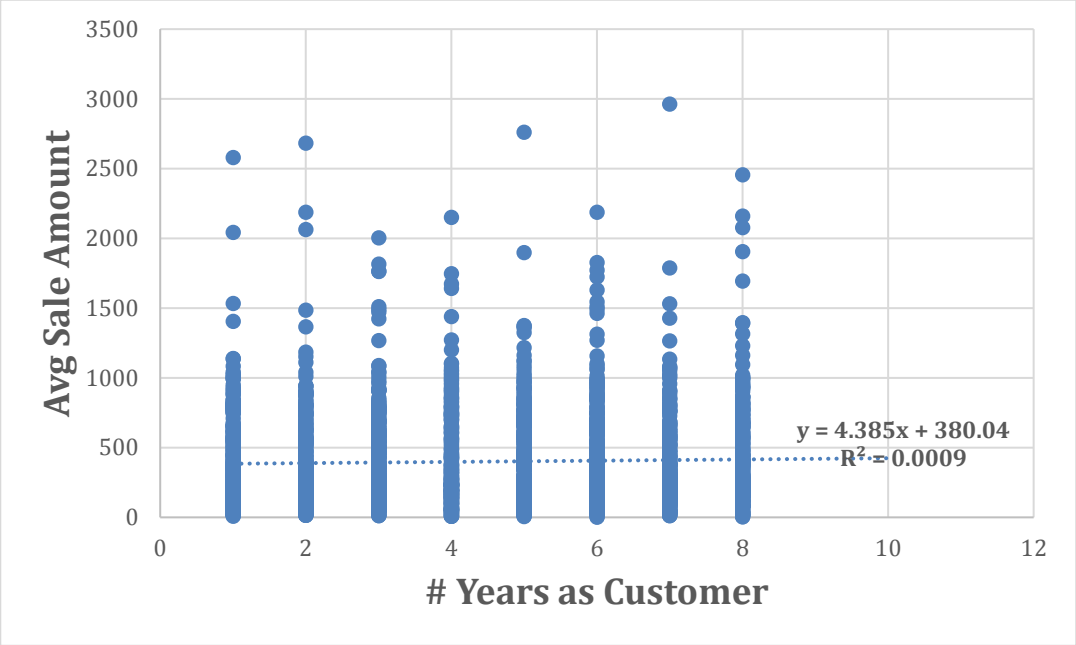
对比两张表提供的数据维度，目标变量 Avg Sale Amount 会与 Customer Segment、Store Number、Avg Num Products Purchased、# Years as Customer 有一定的关系，其中 Customer Segment、Store Number 是分类变量， Avg Num Products Purchased、# Years as Customer 是连续变量

1.2 确定 Avg Num Products Purchased、# Years as Customer 连续变量的相关性。

关于 Avg Sale Amount 分别与 Avg Num Products Purchased、# Years as Customer 得到的散点图如下，



Avg Sale Amount 与 Avg Num Products Purchased 的散点图如上，且 R 的平方大于 0.7，可以肯定 Avg Num Products Purchased 与 Avg Sale Amount 相关，是预测变量。



从 Aug Sale Amount 与# Years as Customer 的散点图以及 R 的平方可以看出，这两者没有太大关联，故排除# Years as Customer 为预测变量。

1.3 确定 Customer Segment 和 Store Number 这两个分类变量相关性。

1.3.1.1 Customer Segment，将 Credit Card Only 作为基础变量，Store Mailing List、Loyalty Club and Credit Card、Loyalty Club Only 为虚拟变量，用 if 函数进行数据处理转化为 0 或者 1；

Store Number 有 100 到 109 共 10 个不同店号，取 100 为基础变量，另外 9 个为虚拟变量用 if 函数进行数据处理转化为 0 或者 1；

1.3.1.2 对分类变量 Customer Segment 做回归分析，得到的 Adjusted R Square 为 0.7， 三个虚拟变量的 P 小于 0.05，说明该分类变量与目标变量 Avg Sale Amount 有相关性；

SUMMARY OUTPUT								
回归统计								
Multiple R	0.838073244							
R Square	0.702366762							
Adjusted R Square	**0.70199017**							
标准误差	185.6701605							
观测值	2375							
方差分析								
	df	SS	MS	F	Significance F			
回归分析	3	1.93E+08	64294977.17	1865.060055	0			
残差	2371	81736452	34473.40851					
总计	2374	2.75E+08						
	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	682.6789474	8.353695	81.7217902	**0**	666.2976428	699.0603	666.2976	699.060252
X Variable 1	-525.3174221	10.04477	-52.29760376	**0**	-545.0148655	-505.62	-545.015	-505.6199787
X Variable 2	391.4805372	15.73157	24.88503082	**1.2112E-121**	360.6314839	422.3296	360.6315	422.3295904
X Variable 3	-286.346374	11.37206	-25.17981126	**3.5029E-124**	-308.6465897	-264.046	-308.647	-264.0461582

对分类变量 Store Number 做回归分析，得到 Adjusted R Square 为-0.0006， 三个虚拟变量的 P 远大于 0.05，故该变量与目标变量 Avg Sale Amount 无相关性；

回归统计								
Multiple R	0.05616378							
R Square	0.00315437							
Adjusted R Square	**-0.00063912**							
标准误差	340.2244786							
观测值	2375							
方差分析								
	df	SS	MS	F	Significance F			
回归分析	9	866257.512	96250.83467	0.831521322	0.586971828			
残差	2365	273755125.6	115752.6958					
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	412.5054908	18.84329109	21.89137178	**6.9575E-97**	375.5544082	449.4565734	375.5544082	449.4565734
X Variable 1	-15.04157775	27.82918332	-0.540496556	**0.58890553**	-69.61370358	39.53054807	-69.61370358	39.53054807
X Variable 2	-32.0774908	41.43508224	-0.774162595	**0.43891203**	-113.3303431	49.17536148	-113.3303431	49.17536148
X Variable 3	-6.012290798	29.48772702	-0.203891293	**0.83845599**	-63.83676702	51.81218542	-63.83676702	51.81218542
X Variable 4	-26.23263895	27.99613022	-0.937009463	**0.34884936**	-81.13214222	28.66686433	-81.13214222	28.66686433
X Variable 5	6.735361661	27.10325911	0.248507445	**0.80376339**	-46.41325033	59.88397365	-46.41325033	59.88397365
X Variable 6	-30.483512	27.64217331	-1.102789989	**0.27023058**	-84.68891726	23.72189327	-84.68891726	23.72189327
X Variable 7	1.497429556	29.44910336	0.050848053	**0.95945089**	-56.25130693	59.24616604	-56.25130693	59.24616604
X Variable 8	-53.17315746	30.10436733	-1.766293803	**0.07747556**	-112.2068453	5.860530403	-112.2068453	5.860530403
X Variable 9	14.65717192	32.24898158	0.454500304	**0.64951048**	-48.58203493	77.89637878	-48.58203493	77.89637878

2 建模

目标变量 Avg Sale Amount

连续变量 Avg Num Products Purchased

分类变量 Customer Segment（设三个虚拟变量 Store Mailing List、Loyalty Club and Credit Card、Loyalty Club Only；

用 excel 里数据分析里的回归工具得到以下信息，Adjusted R Square=0.834，P 值小于 0.05，说明该模型合理。

SUMMARY OUTPUT								
回归统计								
Multiple R	0.914810204							
R Square	0.836877709							
Adjusted R Square	**0.836602397**							
标准误差	137.4832081							
观测值	2375							
方差分析								
	df	SS	MS	F	Significance F			
回归分析	4	229824514	57456128.51	3039.744236	0			
残差	2370	44796869.07	18901.63252					
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	**303.4634713**	10.57571483	28.69436972	**1.1227E-155**	282.72486	324.2020827	282.72486	324.2020827
X Variable 1	**66.97620492**	1.515040358	44.20753848	**0**	64.00526313	69.94714671	64.00526313	69.94714671
X Variable 2	**-245.417744**	9.767775616	-25.12524388	**1.05E-123**	-264.572015	-226.263474	-264.572015	-226.263474
X Variable 3	**281.8387649**	11.90985741	23.66432739	**2.5804E-111**	258.4839461	305.1935838	258.4839461	305.1935838
X Variable 4	**-149.355722**	8.972754792	-16.64547014	**6.34584E-59**	-166.950984	-131.7604598	-166.950984	-131.7604598

故该模型的回归方程式如下：

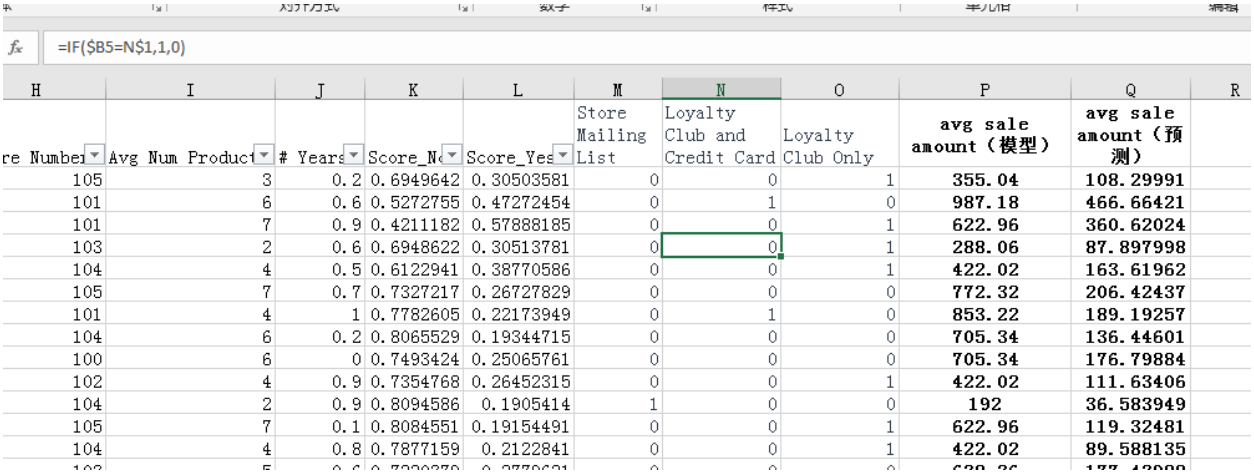
Avg Sale Amount= *303.46*+ 66.98* Avg Num Products Purchased)-245.42 (if type: Store Mailing List)+ 281.84(if type: Loyalty Club and Credit Card)-149.36 (if type: Loyalty Club Only)

第 3 步：演示/可视化:

根据你的模型结果给出建议。（限 500 字以内）

假设向这 250 个客户发送了宣传册，

对表”P1-mailinglist”里的 customer segment 的虚拟变量用 if 函数做处理，将回归模型方程式应用到其中，得到每个人预测的 avg sale amount，再乘以每个人会购买的概率，得到每个人的预计收入。处理结果截图如下：



H	I	J	K	L	M	N	O	P	Q	R
re Numbe	Avg Num Product	# Years	Score_N	Score_Yes	Store Mailing List	Loyalty Club and Credit Card	Loyalty Club Only	avg sale amount（模型）	avg sale amount（预测）	
105	3	0.2	0.6949642	0.30503581	0	0	1	355.04	108.29991	
101	6	0.6	0.5272755	0.47272454	0	1	0	987.18	466.66421	
101	7	0.9	0.4211182	0.57888185	0	0	1	622.96	360.62024	
103	2	0.6	0.6948622	0.30513781	0	0	1	288.06	87.897998	
104	4	0.5	0.6122941	0.38770586	0	0	1	422.02	163.61962	
105	7	0.7	0.7327217	0.26727829	0	0	0	772.32	206.42437	
101	4	1	0.7782605	0.22173949	0	1	0	853.22	189.19257	
104	6	0.2	0.8065529	0.19344715	0	0	0	705.34	136.44601	
100	6	0	0.7493424	0.25065761	0	0	0	705.34	176.79884	
102	4	0.9	0.7354768	0.26452315	0	0	1	422.02	111.63406	
104	2	0.9	0.8094586	0.1905414	1	0	0	192	36.583949	
105	7	0.1	0.8084551	0.19154491	0	0	1	622.96	119.32481	
104	4	0.8	0.7877159	0.2122841	0	0	1	422.02	89.588135	

用 sum 函数得到表中 Q 列即 mailinglist 里的

预计总收入为 47225.91 美元

总收益=47225.91*50%-250*6.5=21987.96 元

预期收益大于一万元，所以建议公司给这些客户发送宣传册。