项目分析报告

预测邮寄产品目录带来的收入增长

1. 理解业务与数据

- 我们需要做出的决策：

  是否向这 250 名新客户发送产品目录？

- 决策所需信息：

  需要知道向这批客户发送产品目录带来的预期利润。

- 哪种分析模型能够获得所需信息：

  需要通过一个预测性模型来获得。

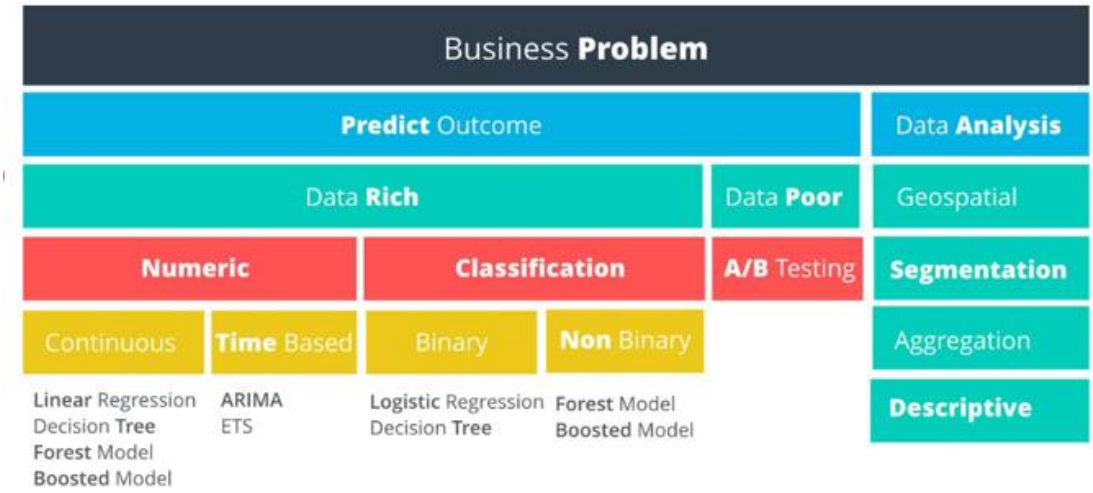
- 目前所拥有的数据：
  - Name：姓名，分类数据，数据来源：customers、mailinglist
  - Customer Segment：客户分类，分类数据，数据来源：customers、mailinglist
  - Customer ID：客户 ID，分类数据，数据来源：customers、mailinglist
  - Address：地址，分类数据，数据来源：customers、mailinglist
  - City：城市，分类数据，数据来源：customers、mailinglist
  - State：所在州，分类数据，数据来源：customers、mailinglist
  - ZIP：邮政编码，分类数据，数据来源：customers、mailinglist
  - Avg Sale Amount：平均销售额，数值型，数据来源：customers
  - Store Number：门店号码，分类数据，数据来源：customers、mailinglist
  - Responded to Last Catalog：对上一次目录的回应，分类数据，

数据来源：customers

- Avg Num Products Purchased：平均购买产品数，数值型，数据来源：customers、mailinglist
- # Years as Customer：作为顾客的年份，数值型，数据来源：customers、mailinglist
- Score_No：客户不购买的概率，数据来源：mailinglist
- Score_Yes：客户购买的概率，数据来源：mailinglist

- 预测模型的确立

  根据前述，已知这是一个预测性的问题，且已有丰富数据，而预期利润是一个连续型的数值型变量，按照以下方法图的指示，应当选用线性回归作为分析模型。



2. 分析、建模与验证

- 预测变量的选取：

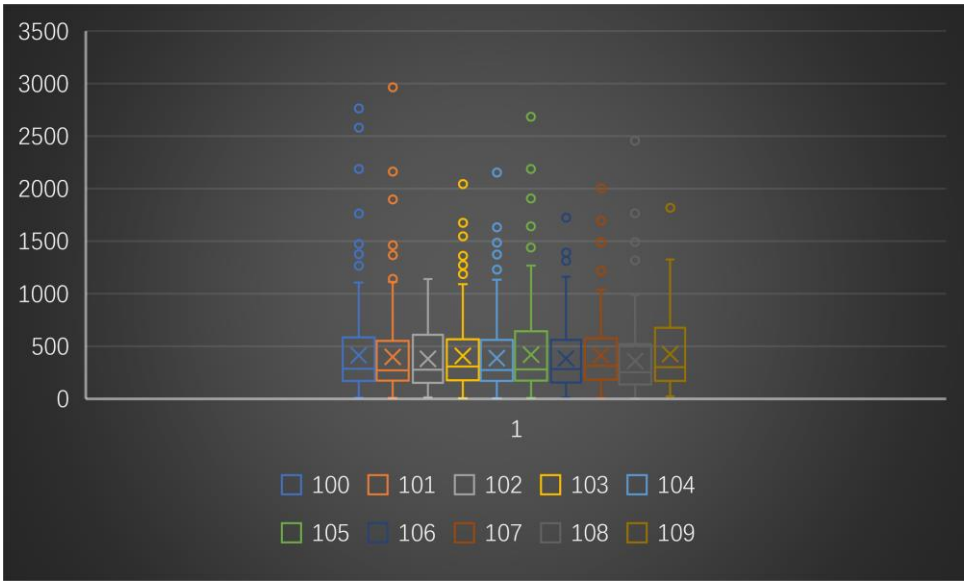
- 首先根据经验排除 Name、Customer ID、Address、State、ZIP 等变量
- Customer Segment：可能会对结果有影响，结合图表及单变量回归来看，可见箱线图显示出其规律，亦可通过单变量回归分析（调整 R^2>=0.5, P<0.05）

单变量回归分析	
Adjusted R Square	**0.70199017**
Loyalty Club and Credit Card: P-value	**1.2112E-121**
Loyalty Club Only：P-value	**3.5029E-124**
Store Mailing List：P-value	**0**

- City：可能对结果有影响，但从单变量回归来看，该变量对目标变量的解释度很弱

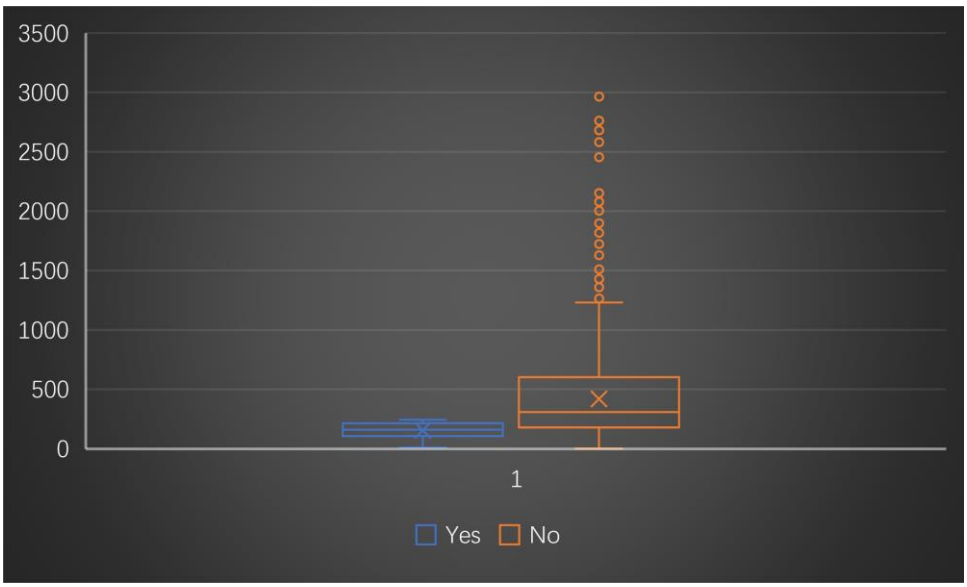
回归统计	
Adjusted R Square	-0.00242812
Denver:P-value	0.653878407
Aurora:P-value	0.670083437
Arvada:P-value	0.777595069
Lakewood:P-value	0.437528387
Broomfield:P-value	0.982162584
Westminster:P-value	0.976787778
Centennial:P-value	0.609797788
Littleton:P-value	0.687332806

- Store Number：可能会有影响，但图表中未见明显规律，单变量回归检测亦不支持其与目标变量间存在线性关系



回归统计	
Adjusted R Square	-0.000639123
100：P-value	0.649510485
101：P-value	0.371574747
102：P-value	0.30169929
103：P-value	0.550678485
104：P-value	0.22058413
105：P-value	0.808175198
106：P-value	0.172444842
107：P-value	0.703721449
108：P-value	0.053817547

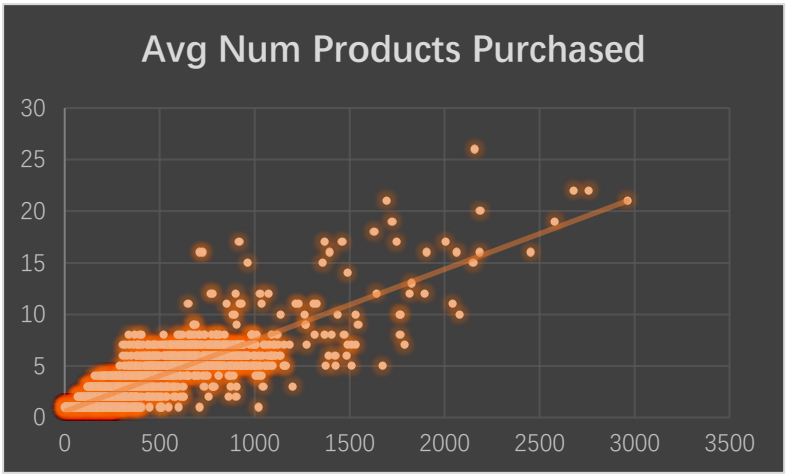
- Responded to Last Catalog：按照经验判断，该变量应对结果有影响，同样以下箱线图也显示两者有区别，但单变量回归的分析统计量显示两者之间缺乏线性关系。



回归统计	
R Square	0.039743702
Responded to Last Catalog: P-value	1.0296E-22

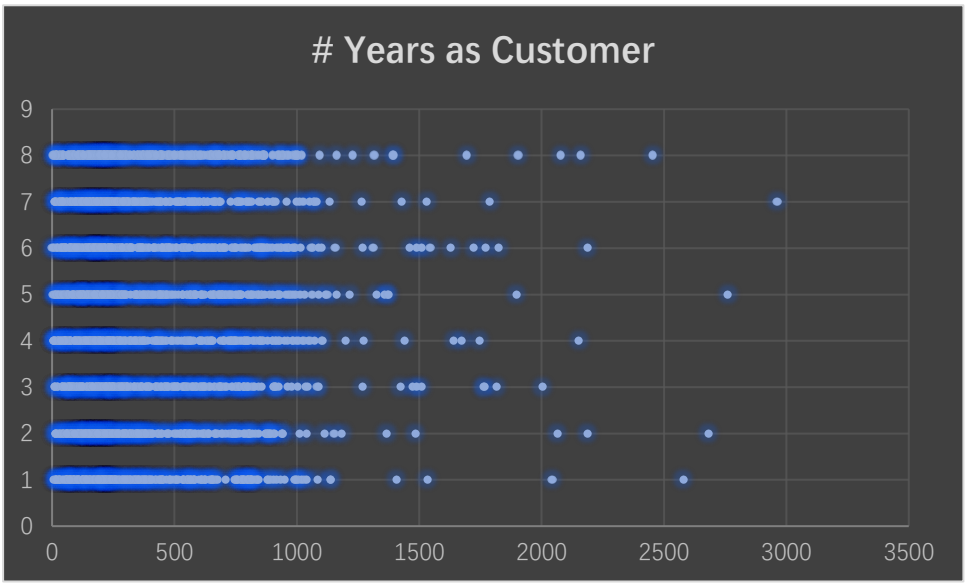
- Avg Num Products Purchased：从以下的散点图可以看到明显

的相关性，同时回归分析也显示该变量与目标变量有线性关系（相关系数>0.8，R^2>0.7，P<0.01）



回归统计	
相关系数	**0.855754217**
R Square	**0.73231528**
Avg Num Products Purchased：P-value	**0**

- # Years as Customer：主观考虑这个变量应会对结果有影响，但散点图未能体现出明显相关线性规律，同时相关系数、R^2 以及系数估计的 P 值都不符合要求。



回归统计	
相关系数	**0.02978186**
R Square	**0.00088696**
# Years as Customer：P-value	**0.14679483**

综上所述，应选择以下变量作为预测变量：

- Customer Segment
- Avg Num Products Purchased

- 多元线性回归模型的建立与验证
  - 多元线性回归模型方程：

    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)

  - 多元线性回归模型的检验：

回归统计	
Adjusted R Square	**0.836602397**
P-value	
Intercept	**1.1227E-155**
Avg Num Products Purchased	**0**
Loyalty Club and Credit Card	**2.5804E-111**
Loyalty Club Only	**6.34584E-59**
Store Mailing List	**1.0503E-123**

整个模型的调整 R^2>0.8，各项系数估计值的 P 值都<0.05，因此可以认为该多元线性回归模型是适用的。

3. 演示/可视化:

- 给出的建议：

  建议向这 250 名客户发送产品目录。

- 预计发送产品目录带来的预期利润达到 21987.43 美元

- 得出建议的过程：

  运用经验常识、可视化图表以及单变量回归分析等方法下筛选出合适的预测变量后，基于这些预测变量构建出如下多元线性回归模型：

  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）

  其调整 R^2 及各个系数估计值的 P 值都满足要求，因此该模型可以用于估计预期收入额。

将该模型运用到 250 名新客户的资料中，计算每一名客户对应的预计销售额，预期销售额（预计销售额乘以购买概率）、预期利润。将计算所得的 250 名客户的预期利润加总，可得到向这些客户发送产品目录带来的预期利润总共 21987.43 美元。

按照要求，如果这些新客户带来的预期利润超过一万美元，那么管理层就会向他们寄送产品目录册。因此我们建议管理层向他们寄送产品目录。