预测目录需求

阮祥炬 2019/6

Step1：项目概述

有一家生产和销售高端家居用品的公司。去年该公司发出了第一份印刷的产品目录，并准备在未来几个月发出今年的产品目录。该公司的邮件列表中有250个新客户，他们希望将目录发送到这些目标客户的邮箱中。您被要求预测这250位新客户的预期利润。除非预期的利润贡献超过10,000美元，否则管理层不希望将目录发送给这些新客户。而且还有以下的信息：

1. 印刷和分发的成本为每个目录6.5美元。
2. 通过目录销售的所有产品的平均毛利率为50%。

把以前客户的购买历史作为分析的依据，分析出影响利润的关键因素，然后构建一个利润模型用来预测可以从新客户处获得的利润，最后使用模型预测可以从用户处获得的利润。

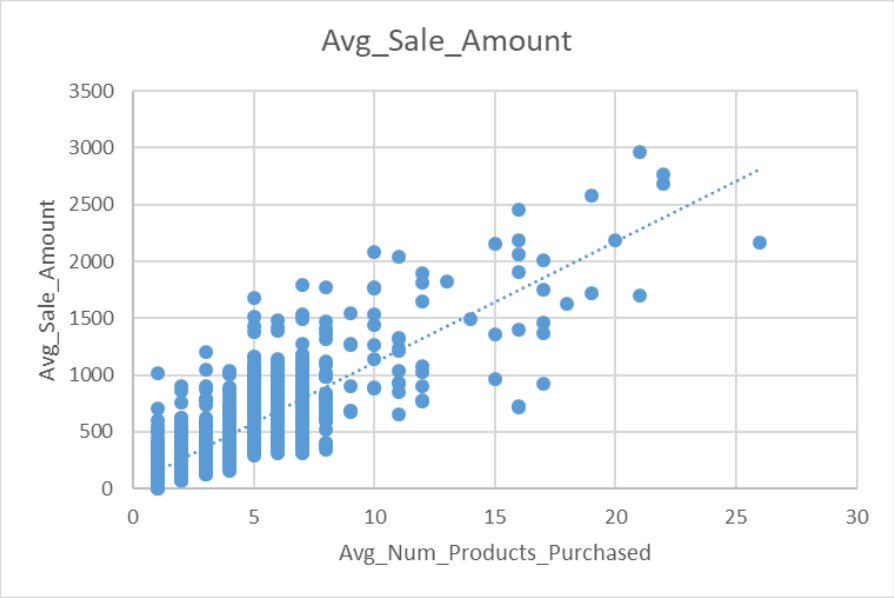
Step2：分析，建模，可视化

从现有的客户信息表中可以得到以下字段信息：

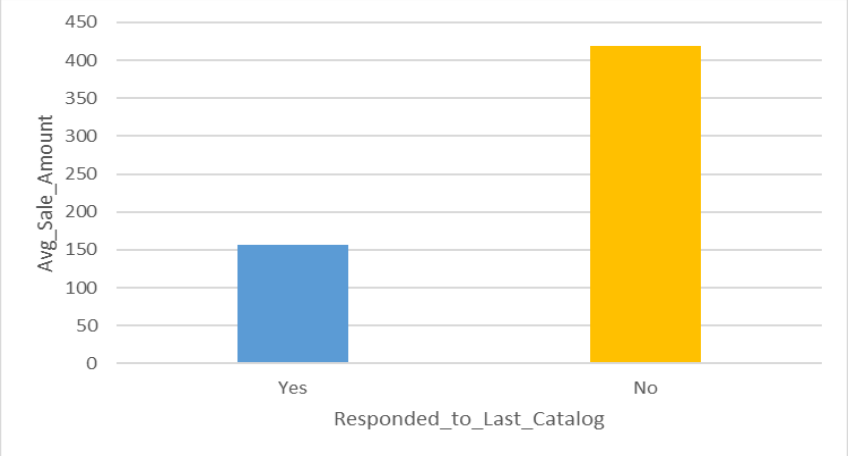
- Name & Customer ID
- Customer Segment
- Location（Address,City,State,and Zip Code）
- Store Number
- Responded to Last Catalog
- Average Number of Products
- Years as Customer
- Average Sale Amount(This will be our target variable)

为了得出线性模型中需要的预测因子，需要使用双变量分析法来研究数据集合中各变量和平均销售额（Average Sale Amount）之间的关系。如果一个变量与目标变量之间存在某种线性关系，我们就可以假设它可以作为线性分析的一个预测因子。

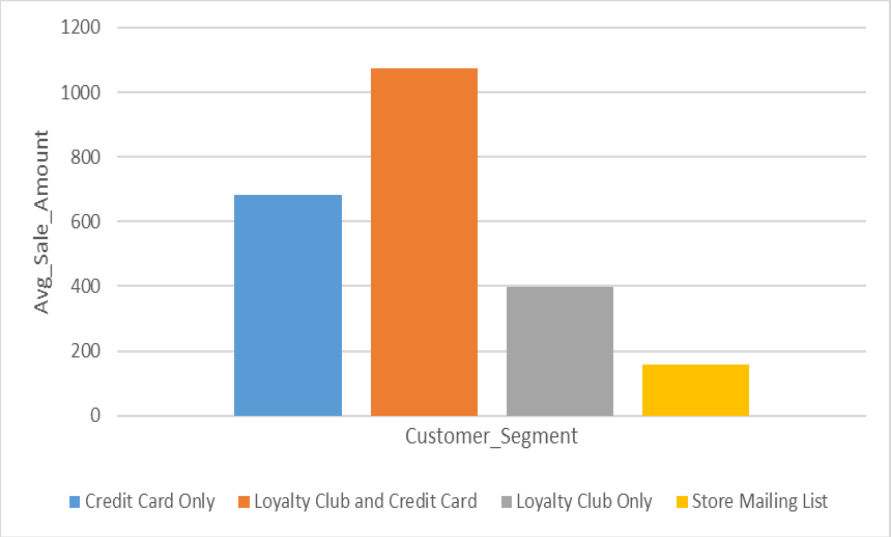
通过对客户购买信息数据中产品平均销售额（Avg_Sale_Amount）和产品平均购买量（Avg_Num_Products_Purchased）的双变量分析可得到如下的散点图，该图表明产品平均销售额和产品平均购买量之间存在着很强的线性关系。



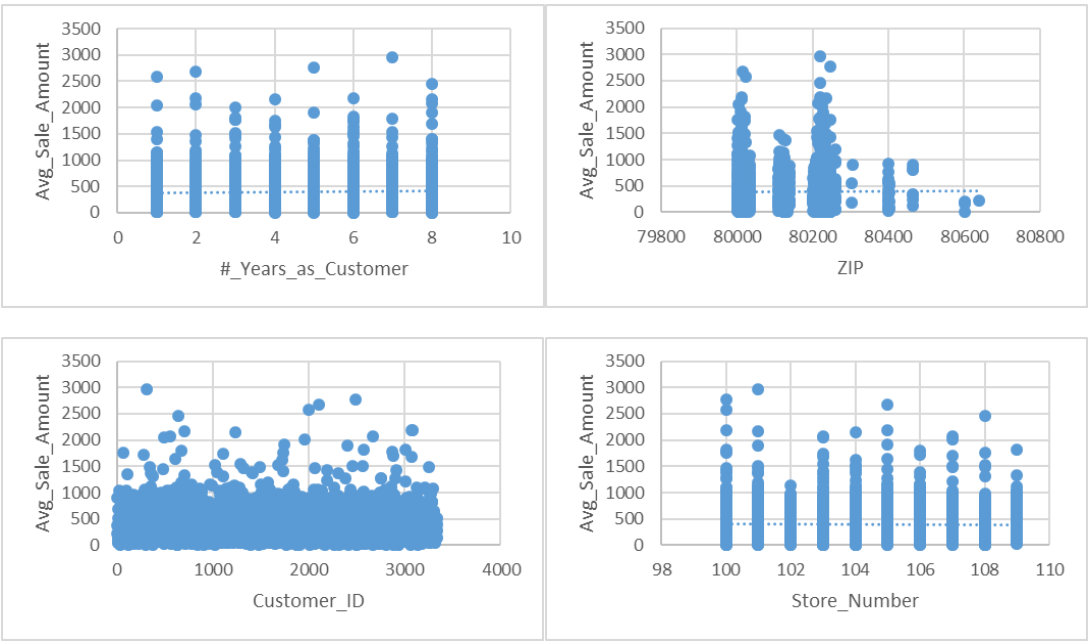
而通过对于是否回复最后一次目录（Responded_to_Last_Catalog）和产品平均销售额（Avg_Sale_Amount）之间的条形图分析可知，两者间也存在着较强的关系。



通过做出的条形图发现产品平均销售额（Avg_Sale_Amount）和客户细分（Customers_Segment）也有较强的关系。使用信用卡和优惠卡购物的客户贡献了较高的产品平均销售额（Avg_Sale_Amount），而使用商店邮件列表购物（Store Mailing List）的客户贡献的销售额较低。



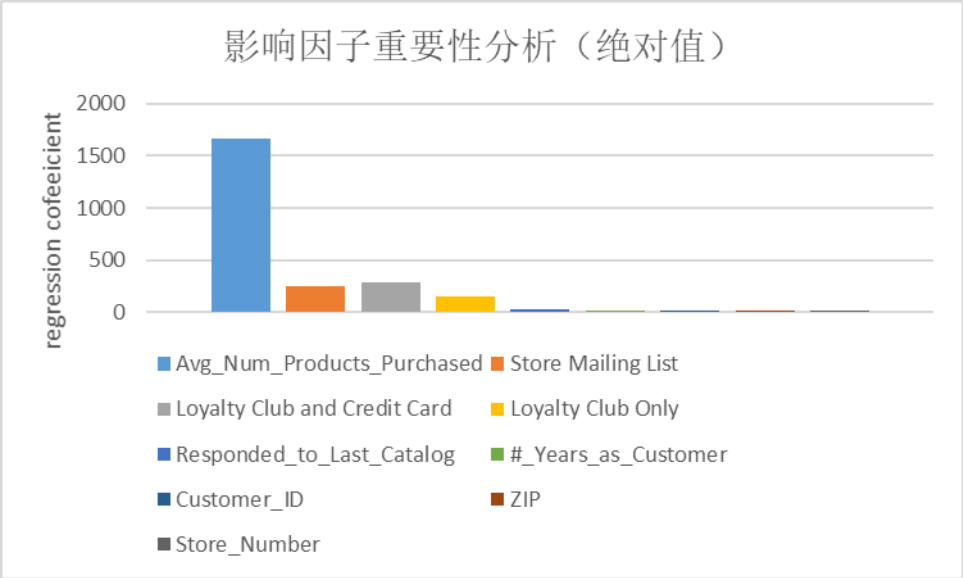
其余的指标中，包括客户年限（#Years_as_Customer），邮编（ZIP），店铺编号（store number），客户编号（customer ID number）各指标间的散点分布均匀，对于产品平均销售额（Avg_Sale_Amount）没有太大的影响，所以他们不会作为线性回归模型的变量。



由于变量（Responded_to_Last_Catalog）属于二进制变量，所以为了使它能够在线性回归模型中使用，需要对其进行处理，所以令 No = 0，Yes = 1。客户细分（Customer Segment）有多个非数字化选项，所以将其设置为哑变量。

Customer Segment	Store Mailing	Loyalty Club	Loyalty Clu	Credit Card
Store Mailing List	1	0	0	0
Loyalty Club and Credit Card	0	1	0	0
Loyalty Club Only	0	0	1	0
Credit Card	0	0	0	1

在构建线性模型的过程中，最重要的一步是选出线性回归方程最有效的影响因子，包括那些转换为二进制变量和哑变量。对所有影响因子使用最大最小归一化处理，然后考虑各影响因子对于客户平均销售额的影响。



对客户数据进行一系列的必要处理后，然后进行回归分析得到上图。分析上图可知产品平均购买量（Avg_Num_Products_Purchased）对产品平均销售额（Avg_Sale_Amount）的影响最大。客户细分（Customers_Segment）对于销售额也有较大的影响。这与先前的假设一致。而客户是否回复最后一次目录对于销售额的影响不大，与先前的假设不符，排除掉它的影响。其他的因素，客户年限，客户编号，客户邮编，商店编号等对模型的影响较低，可以忽略，这与先前

的假设一致。

综上，可以得到如下四个构建线性回归方程最重要的影响因子：

- Avg_Num_Products_Purchased
- Store Mailing List
- Loyalty Club and Credit Card
- Loyalty Club Only

根据这四个影响因子构建线性回归模型，并得到最终的线性回归方程：

Average Sale Amount =

303.46 + 66.98 * (Avg_Num_Products_Purchased)

- 245.42 * (if segment:Store Mailing List)

+ 281.84 * (if segment:Loyalty Club and Credit Card)

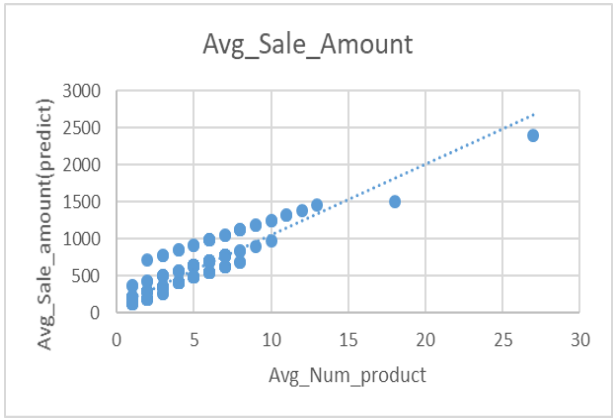
- 149.36 * (if segment:Loyalty Club Only)

+ 0 * (if segment:Create Card Only)

计算得出该模型的 R 平方约为 0.84，大于 0.8 的拟合度，说明该模型拟合优度较高。

Step3:可视化

上一步骤中建立的线性模型将用于预测 250 个邮件列表中的潜在客户，每个客户的平均销售额。基于客户细分和平均购买的产品数量进行预测。将数据带入线性回归模型，得到下面两幅图，分别是客户平均购买量和客户细分对客户平均销量的预测。



发送这 250 个目录的预期收入是预期销售的产品和客户响应并购买的概率的总和

Revenue = Sales(Avg_Sale_Amount(predict)) x P(Customer will purchase(Score_Yes))

此外，预测的利润还必须考虑到通过目录销售的所有利润的平均毛利率以及印刷每份目录的成本。

Profit = （Revenue x Gross Margin）- Cost of Catalog

将上面的两个公式代入新客户的信息表中，通过计算得到如下的表格（部分）

Score_Yes	Avg_Sale_Amount(predict)	Revense(predict)	profit(predict)
0.305035807	355.04	108.2999129	47.64995646
0.472724537	987.18	466.6642084	226.8321042
0.57888185	622.96	360.6202373	173.8101186
0.305137811	288.06	87.89799784	37.44899892
0.387705855	422.02	163.6196249	75.30981246

对新客户信息表中的信息进行聚合得到下面的一个汇总表。

Avg_Sales_Amount(predict)	138295.2
Revenue(predict)	47225.91
Profit(predict)	21987.96

从表中可知250位客户的预期总销售额为$138295.2.客户响应后购买的销售额为$47225.91.最终可以获得的利润为$21987.96.该利润大于预期的$10000,所以可以向这250名新客户发放邮件目录。