中北大學

数据分析与统计建模

课 程 设 计

学　　院　理学院

专　　业　统计学

题　　目　宾堡集团的库存需求预测

学生姓名　孙玉林　学 号　1308064146

学生姓名　李姝　学 号　1308064120

学生姓名　张肖　学 号　1308064113

学生姓名　吴迪　学 号　1308064125

指导教师　杨明、宋娜

二〇一六年六月

摘要

随着科技的发展，食品行业也越来越重视产品的实际需求量，为了实现食品对客户的精准配送，减少退货和仓储带来的成本增加。宾堡集团给出了公司7周的详细销售数据，包括客户、销售途径、销售路线、销售站、销售地点、销售量、退回量、销售金额、产品种类等数据，让我们预测未来两周各产品的实际需求量，以实现大数据下的精准营销。

针对提供的大量销售明细数据，我们首先对数据进行探索性分析，发现数据的内在联系。对于7周的销售明细，分析销售量和退回率的分布。针对552个销售站，分别分析销售站的销量和退回率的分布，以及销售站和销售地点、周次的交互相关性，可发现大多数销售站的每天平均销量在(0，200)千的区间。分析33个销售地点的销售数据，可发现他们之间的销量差距和退回率差距很大。针对9条销售渠道，可发现渠道1为主要销售渠道约占比50%，并且退回率很低。该集团有超过2000个销售路线（仓库），但是超过2/3的销售路线每天的销售量不超过10000。该集团客户数量约为885416位，且约1/4的主要客户，消费了公司3/4的产品量。该集团生产过的产品有1719种，并且平均价格主要集中在（0,20）比索之间。

然后对需要预测的目标数据进行取自然对数变换，将数据分布转化为近似正态分布，然后将数据根据周次这个时间跨度变量，划分为生成数据集、训练数据集、检验数据集，用于机器学习模型的建立。

建立随机森林回归模型对目标变量进行预测。在200棵分类回归树下模型可得出，测试集中的平均偏差为0.5571、$R^2 = 0.3774$，主要预测变量为产品ID和销售路线，并且该模型的预测值偏于保守。

建立梯度提升机回归模型对目标变量进行预测。在25棵梯度提升回归树下模型可得到，测试集中的平均偏差为0.3902、$R^2 = 0.4476$，主要变量为销售路线。然后综合对比两个预测模型的优劣程度，可以发现梯度提升机回归模型预测效果更好。

关键字：大数据精准营销　　数据探索　　随机森林回归　　梯度提升机

目录

1、问题描述

宾堡集团的库存需求预测

最大限度地提高销量和最大限度地减少烘焙食品的退回

Bimbo 必须在沿其墨西哥 45000 条航线上超过 100 万家商店的货架上，对满足日常消费者对新鲜烘焙产品的需求量有着近似的估算。

目前，每天的库存计算是由直接交货的销售人员完成，他们必须在他们个人经验的基础上单方面对每家商店的供应量和需求量做出预测。对于只有一个星期保质期的面包，可接受的预测误差是很小的。

所以组织一场比赛，目的是保证供需平衡，即准备充足的食物且没有剩余。在这次比赛中，宾堡集团邀请 kagglers 开发一个模型，这个模型能够根据历史销售数据来准确预测库存需求。这样做将确保消费者可以购买到自己所需的商品，同时也减少了花在退款上的金额，这对于销售商和消费者都是一种最好的选择。

而且在这次比赛中，需要预测在规定的一周里，特定的商店中产品的供应量和市场需求量。提供的数据集主要包括墨西哥在 7 个星期内的销售交易记录。其中每周都有送货车运送产品给销售商，每一笔交易都包括售出和退回，退回的产品是未售出的和过期的产品。在某个星期的某个产品的需求被定义为本周的销售量减去下一周的退货量。

注意：

（1）：训练和测试数据集在时间的基础上被分割，以及公共和私人的排行榜(积分榜)数据分割。有可能是产品在训练集里，而不存在在测试集里。这是库存数据的预期行为，因为在任何时间都有新的产品被出售。

（2）：调整后的需求（demanda_uni_equil）总是> = 0 由于需求应该是 0 或者一个正数。（销售量－退回量）数据有时为负值的原因是，有时退回记录会延续几周。

数据集的描述：

（1）：train.csv－训练数据集（74180464 行*11 列）
（2）：test.csv－测试集
（3）：cliente_tabla.csv－客户名数据表
（4）：product_tabla.csv－产品名字数据表
（5）：town_state.csv－城镇和州名数据表

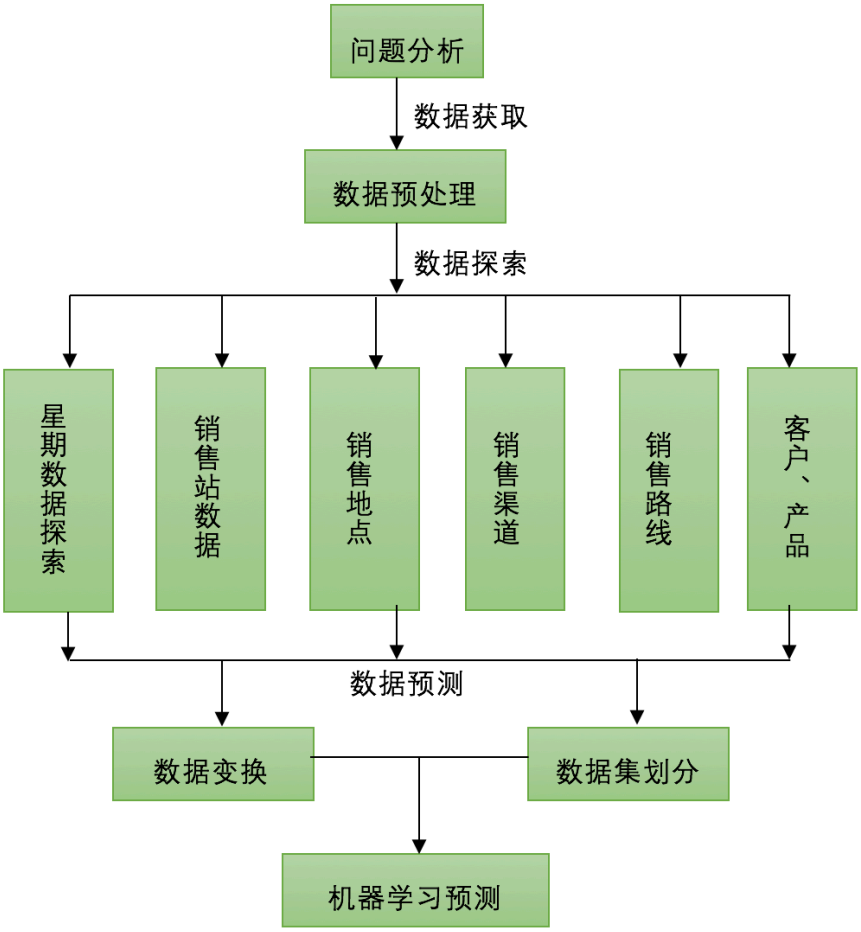
数据中变量名说明：

（1）：Semana－星期数
（2）：Agencia_ID－销售站 ID
（3）：Canal_ID－销售途径 ID
（4）：Ruta_SAK－销售路线 ID（几条路线＝几个销售仓库）
（5）：Cliente_ID－客户 ID
（6）：NombreCliente－客户名
（7）：Producto_ID－产品 ID
（8）：NombreProducto－产品名称

（9）：Venta_uni_hoy－本周销售量（整数）
（10）：Venta_hoy－本周销售量（单位：比索（墨西哥货币））
（11）：Dev_uni_proxima－下周退回量（整数）
（12）：Dev_proxima－下周退回量（单位：比索）
（13）：Demanda_uni_equil－调整后的需求量（整数）（需要预测的目标）

2、求解流程

针对大数据下的精准营销，我们按照下面的流程对问题进行求解：



3、相关变量的计算公式

（1）退回率的计算：

$$\text{本周退回率} = \frac{\text{下周退回量}}{\text{本周销量} + \text{下周退回量}}$$

（2）产品的平均单价：

$$产品的平均单价=\frac{产品销售额}{产品销量}$$

（3）实际需求量：

$$实际需求量=本周销售量-下周退回量$$

4、数据探索

4.1、按照星期划分的数据探索

我们首先将数据按照星期数划分，我们的训练数据集中一共有 7 周的数据。分析每个星期该集团的销售总量，和退回率的大小情况，得到的结果如下图：

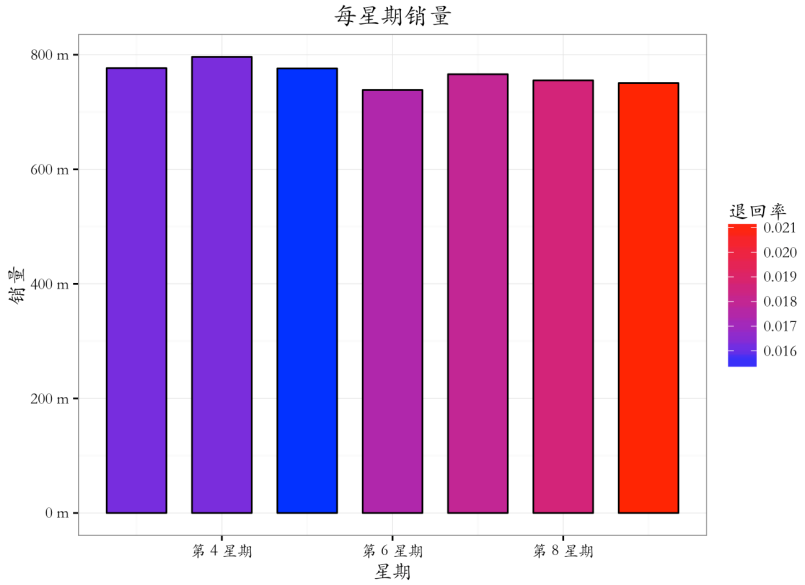


图 1：每星期的销售情况

从上图可以看出：每个星期的销售量，整体来看都相对较平均，相差不是很大。第五周的退货率最小，第 9 周的退货率最大。整体来看随着周数的增长，产品退货率越来越大，说明产品退货率有逐渐增加的趋势。

4.2、销售站的数据探索

销售该公司的产品的销售站共有 552 个，我们首先对销量前 100 的销售站的销量占比和退货率可视化分析，结果如下图：

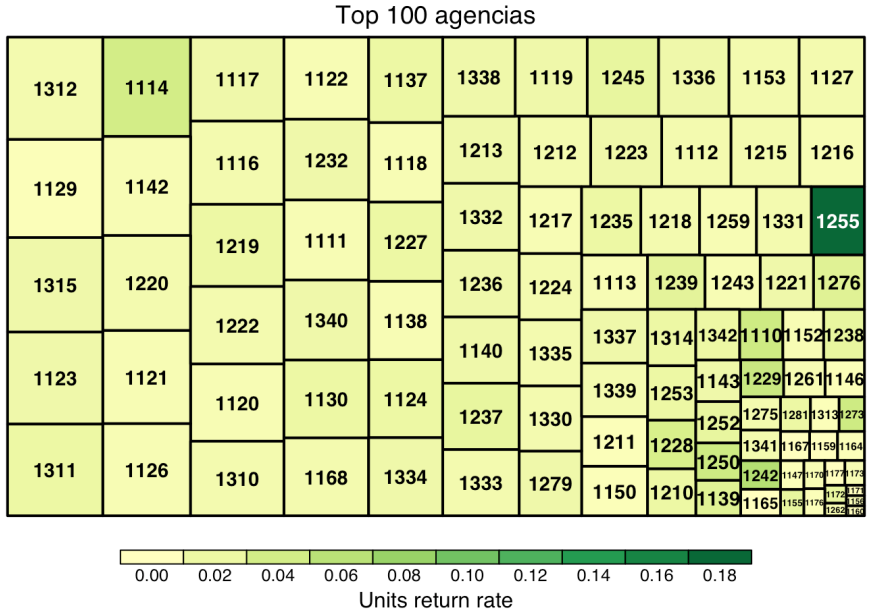


图 2:前 100 销售站销量树图

上面的树图中，其图形面积代表销量，不同颜色确定其退回率。由图可知：ID 为 1312、1129 的销售站销量最大，且退货率比较低；ID 为 1250、1139 的销售站则销量较小，退货率稍高；ID 为 1255 的销售站退回率最高，近达 18%。整体上，销量越小的销售站退货率就越大。

然后分析销售站的每天平均销量的分布情况，可以得到如下图示：

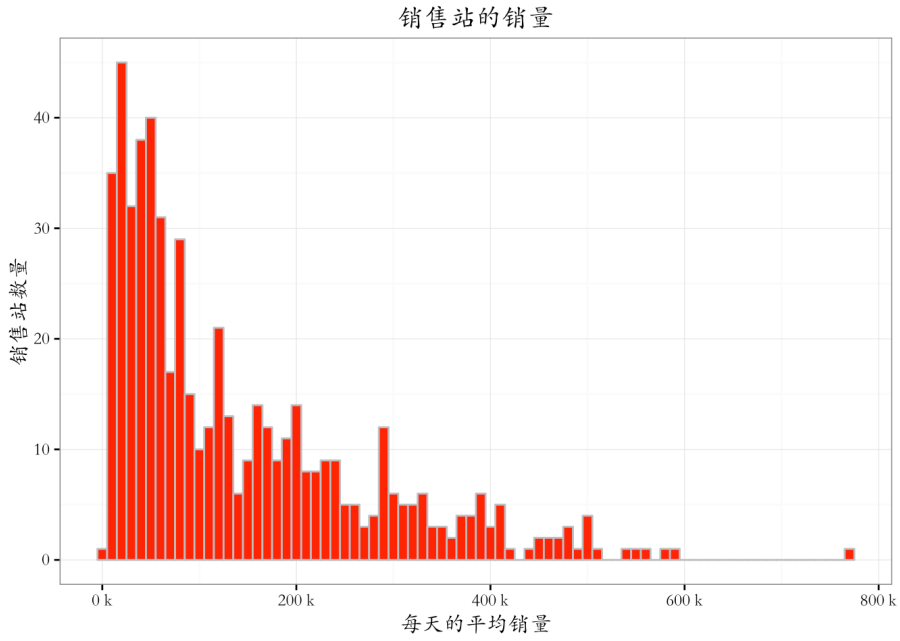


图 3:销售站销量分布图

由上图可知：每个销售站每天的平均销量分布相对较集中，大多数都分布在 0~200k（千）的区间，说明多数的销售站销量还是一般，只有少数的销量较高，每天平均销量达到 400k（千）的销售站数量更少。有极个别的销售站每天的平均销量接近 800k（千）。

最后我们拿出销量前 30 的销售站，查看他们的每星期销售量和退货率的情况，可以得到下图：

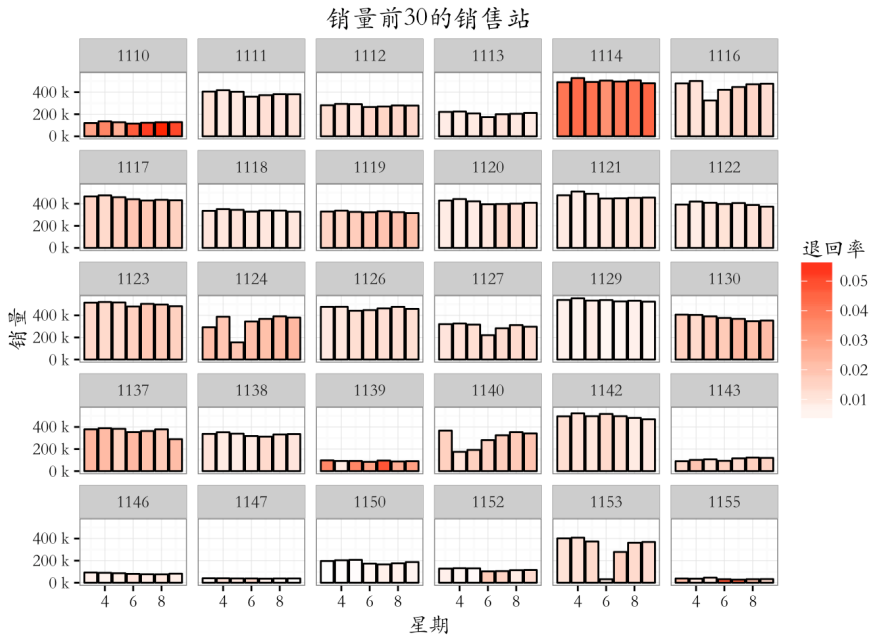


图 4:销售站的每星期销量和退回率图

从上图可以看出：在销量为前 30 的销售站中，有 21 个销售站每星期的销售量都在 200k 之上，有 8 个销售站每周的销售量都超过 400k；大部分销售站每个星期的销量均衡，只有 1124, 1140, 1153 会有大的波动；大部分的销售站退货率都在 3%以下，只有 1110，1114, 1139 销售站的退货率将近 5%，1139 销售站在七周之内的退货率有较大的波动。1114 的销量与退货率在销量前 30 的销售站中都是最高的，可能该店的宣传力度够好，然而产品质量及售后不到位。

4.3、销售地点（州）的数据探索

该公司一共在 33 个州销售自己的产品，接下来让我们看一看，每个州的销售情况，得到如下树图：

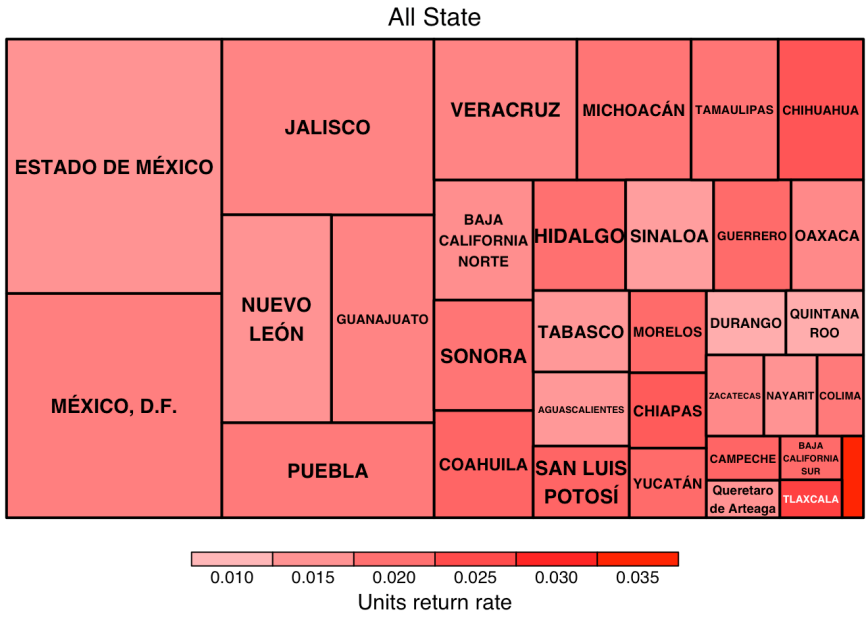


图 5:每个州的销售量和退回率树图

上图反映了所有州的销售量及其退货率，从图中可以看出，销售量前三的州分别为 ESTADD DE MEXICO、MEXICO.D.F 与 JALISCO，所有州的退货率基本都维持在[0.015,0.025]之内。ESTADD DE MEXICO 的销量最高，退货率相对较低，其他州都可以将其销售方法进行借鉴与参考。而销售量最小的州退货率也最高，约可达 3.5%。同样可以看出销量和退货率成反比的趋势。

接下来，我们可以查看各个州的每星期的销量和退货率的情况，的到如下图示：

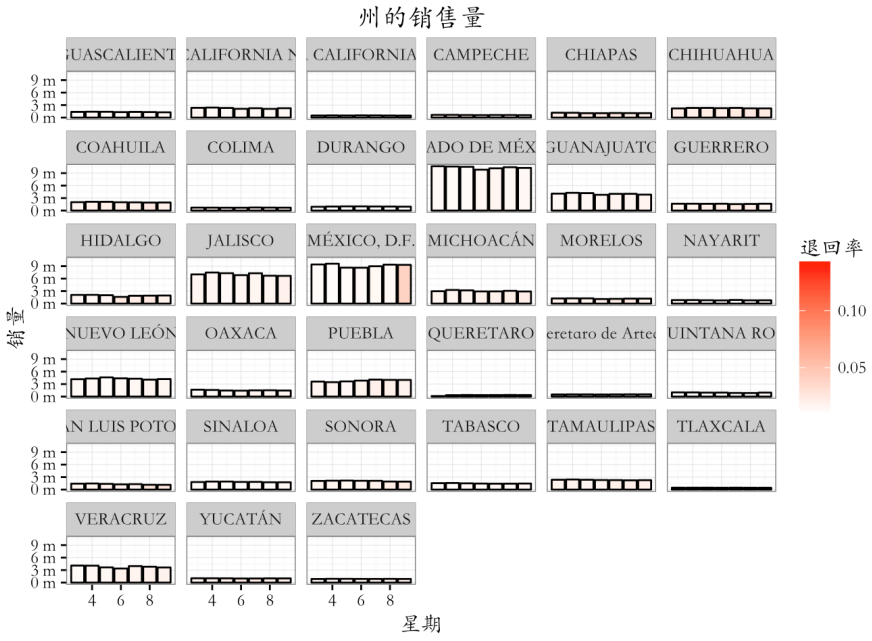


图 6:各州的每周销量和退货率

上图按照州的分布进行每星期的销售量、退货率进行统计，只有几个州的销量明显高于替他州，例如，ADO DE MEX 和 MEXICO.D.F.,大体销量销量都还可

以。在七周内每个州的销量分布及退货率基本都维持在同一水平上，且退货率都比较低，不到 0.05，只有 MEXICO，在第九周的退货率明显偏高。

4.4、销售渠道的数据探索

该公司一共有 9 条销售渠道，我们首先产看每个销售渠道的销量和退货率，可得到如下树图：

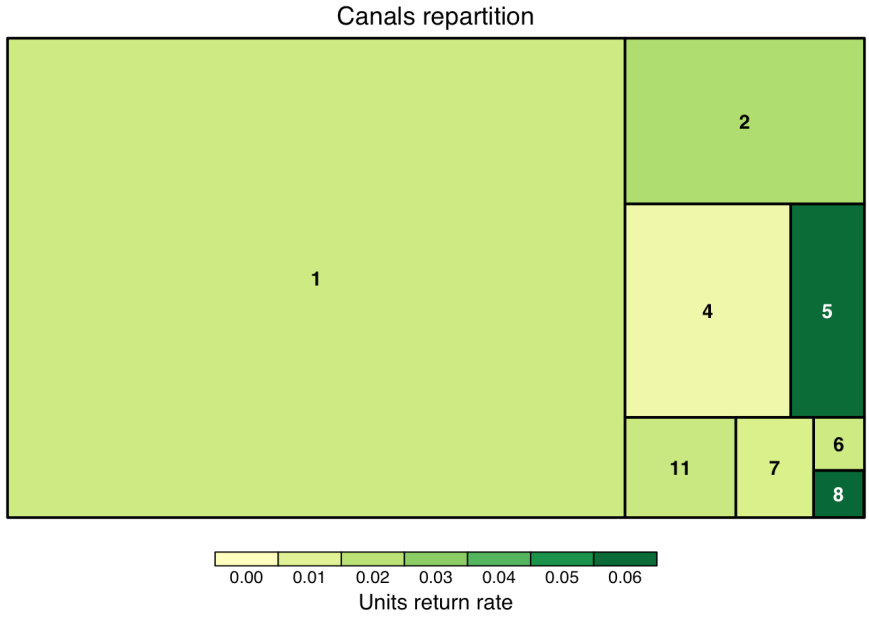


图 7:销售渠道的销量和退货率

上图显示出了 8 条销售渠道（之所以最后一条销售渠道不显示，是因为该渠道的销售量级和其他渠道相比量级太小），其中销售渠道 1 占销售量的绝大部分，退货率约为 2%，销售渠道 4 的退货率最低，几乎为 0，而销售渠道 5 和 8 的退货率最高，约为 6%。由此可知，大部分顾客都更愿意选择销售渠道 1，并且从退货率来看，销售渠道 4 的产品销售方式更好。

接下来我们查看 9 条销售渠道的各个星期的销售情况和退货情况，如下图：

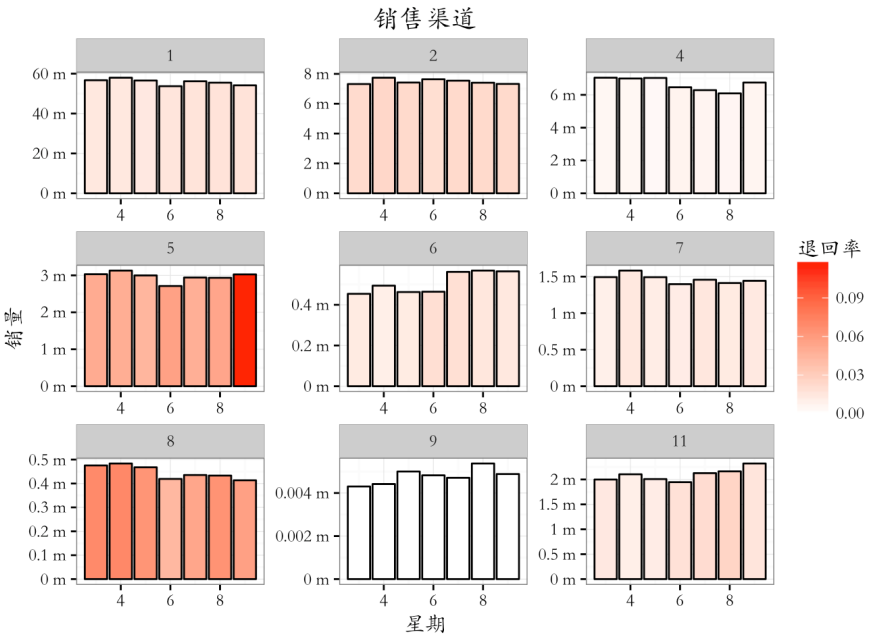


图 8:销售渠道的销量和退货率

上图呈现了不同的销售渠道每个星期所对应的销量，可见每种销售渠道每个星期的销量都维持在一个相对稳定的水平上；然而不同渠道的销量差距较大，渠道 1 销量最高，基本在 5 千万以上，渠道 9 销量最低，基本在 4 万左右。但是销售渠道 5、8 退回率都较高，特别是销售渠道 5 在第九周的退货率最高约 10%。

最后我们分析，各个销售渠道和销售站之间的关系，结果如下图：

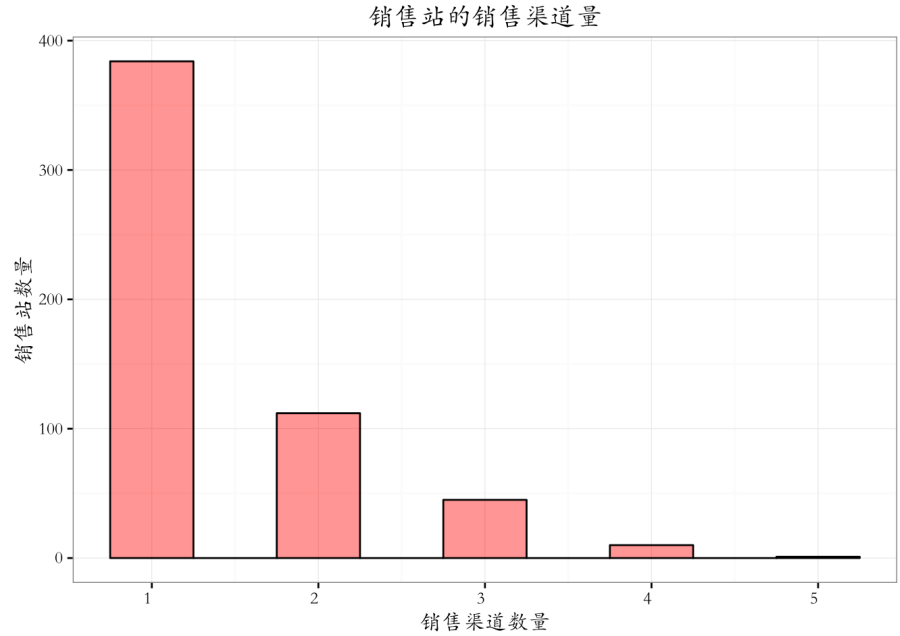


图 9:销售站对应的销售渠道

从上图可知：将近 400 个销售站只有 1 条销售渠道，只有不到 50 个销售站有超过三条的销售渠道。

4.5、销售路线的数据探索

该公司有大量的销售路线（销售路线对应着仓库），我们首先产看销售路线的每天的销量情况，如下图所示：

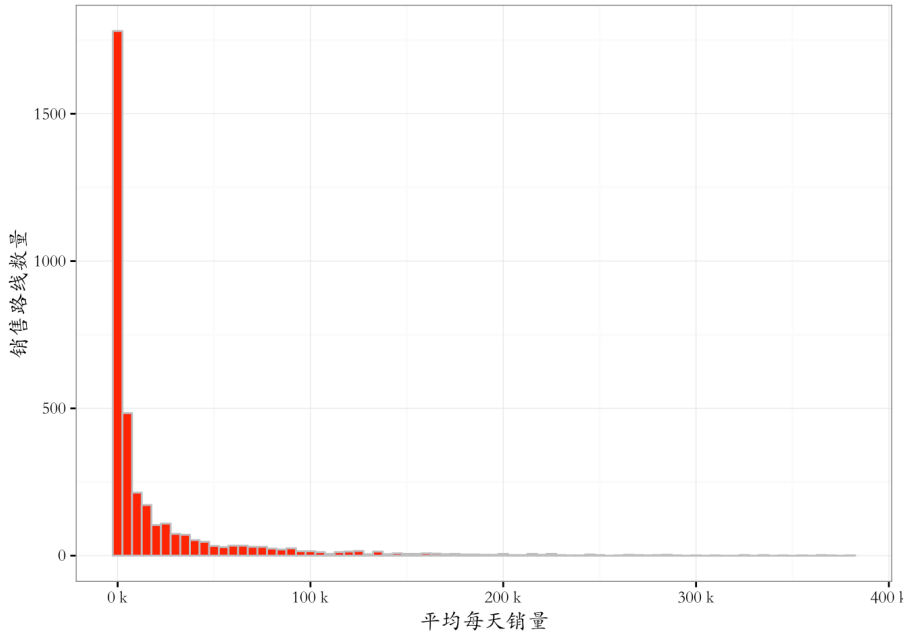


图 10:销售路线和每天销量

从平均每天销量和销售路线数量的条形图可以看出，大部分销售路线的平均销售量并不多，超过 2/3 的销售路线每天的销售量不超过 10000，且随着销量的增多销售路线的数量有递减非常迅速。只有极少数的销售路线销量超过 200k（千）。

接下来我们查看前 50 条销售路线和前 50 个销售站队销量和退货率的交互关系，如下图所示：

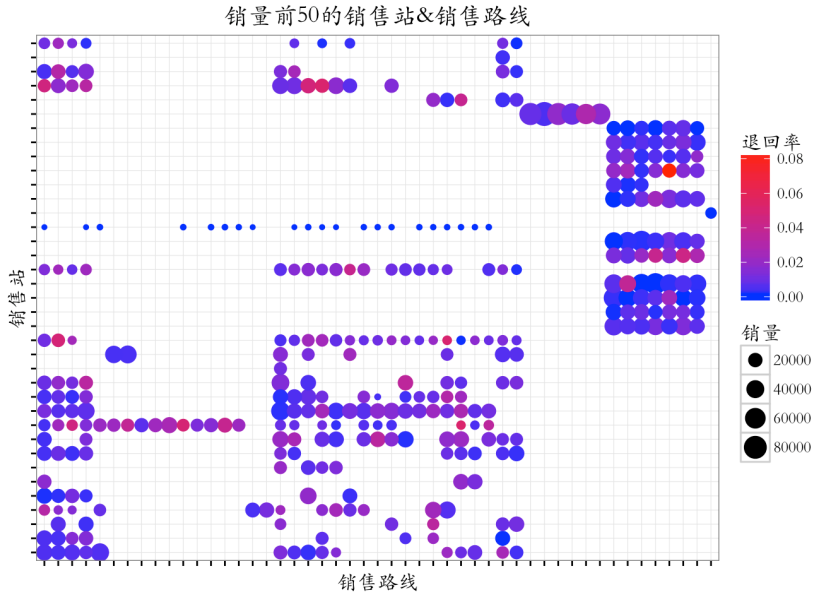


图 11：前 50 条销售路线和前 50 个销售站

从上图中可以看出：在图的右上侧点的 分布较为密集，点较大，颜色偏蓝，说明销售量大相应的退回百分比较小。

4.6、客户的数据探索

该公司的客户数量约为 885416 位，我们首先分析消费量前 100 的客户的消费情况，得到树图如下：

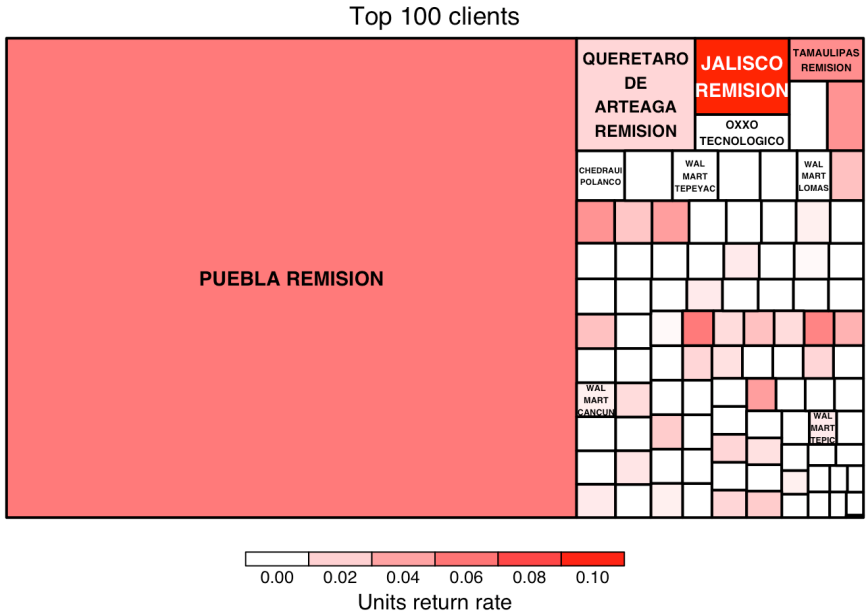


图 12:消费量前 100 的客户

上面树形图中面积最大的是 Puebla Remision，可见他是一个大客户，但是退货率也偏高。但其他客户与 Puebla Remision 花费相差较大，说明该店客户源较单一，应当多挖掘一些新的客户。并且可以发现客户的消费量和退货率成反比，会存在消费量较小的客户流失的概率较大的问题。

接下来我们分析，所有客户对该公司产品消费的累积贡献情况，如下图：

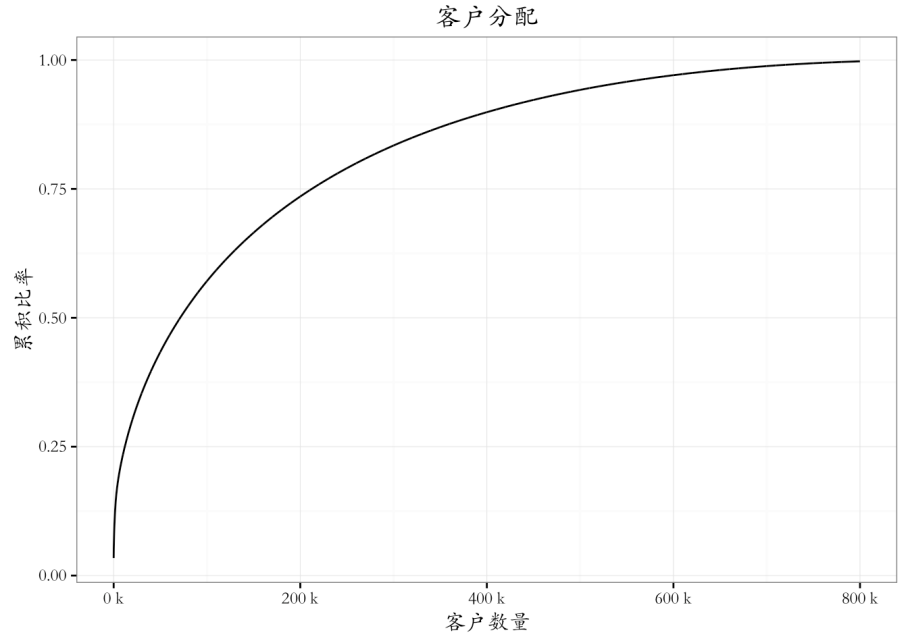


图 13:客户的累积消费

从上图可以看出：曲线随着客户数量的增多逐渐趋于平缓，销量的累积百分比逐渐趋于 1，可以看出前 20 万主要客户贡献了约 75%的销售量。可见前面约 1/4 的主要客户，消费了公司 3/4 的产品量。

下面我们分析客户和购买渠道的相关情况，可以得到如下数据表：

表 1:销售渠道和客户量

销售渠道数	1	2	3	4
客户数量	874022	6516	65	1

可以看出大多数客户只是通过一个购买渠道买到产品。有极少数的客户购买的产品是多个渠道送达的。

下面我们分析客户和销售路线的相关关系，得到结果如下图所示：

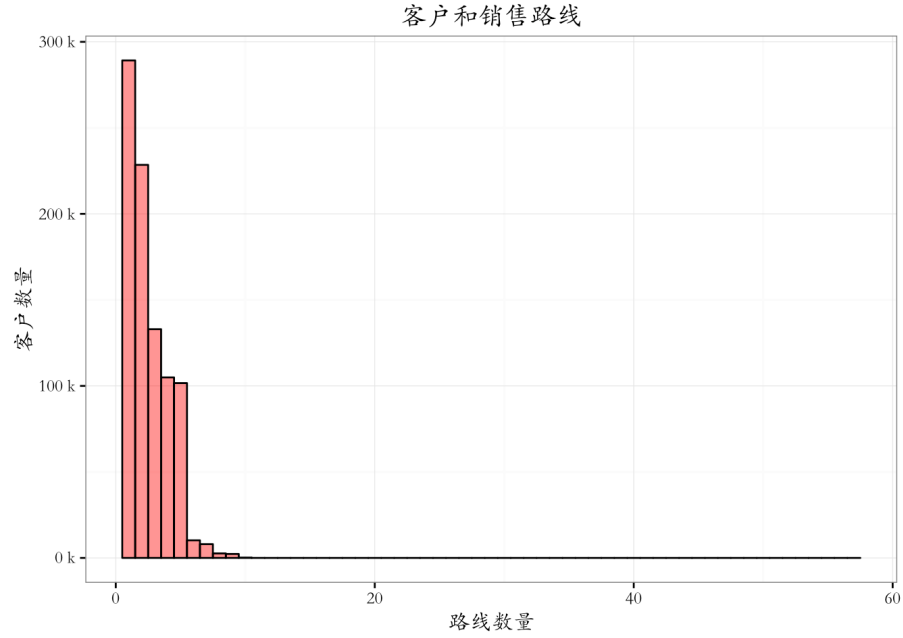


图 14:客户和销售路线

从上图可知：大部分客户购买的产品是通过不超过 5 条销售路线（仓库）来销售，只有极少数的客购买的产品来自超过 10 个仓库。

4.7、销售产品的数据探索

该集团销售过的产品有 1719 种，首先查看销售量排名前 100 的产品的销量和退货率树形图：

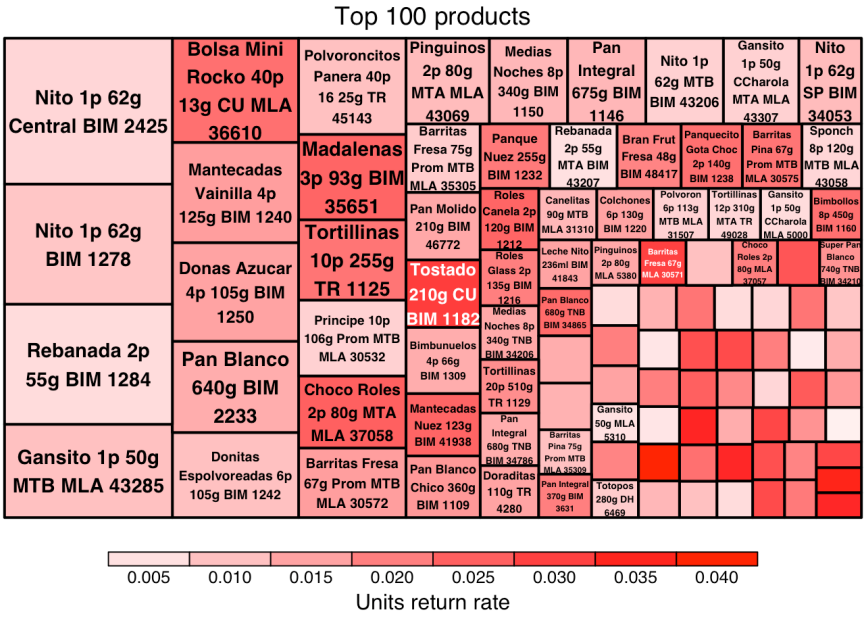


图 15:销量前 100 产品树图

从上图可以看出：销量较多的产品前三名是 Nito 1p 62g Central BIM 2425，Rebanada 2p 55g BIM 1284，Nito 1p 62g BIM 1278，且退回率都基本在 1.5%左右。产品 Bolsa Minl Rocko 40p 13g CU MLA 36610 的销量高，然而退回率也较高，约 3%，说明该产品应做相应的改进，减少退货率 ，有更好的提升空间。并且可以发现，退货率和销量并没有明显的相关性。

接下来我们分析产品的平均单价的相关分布：

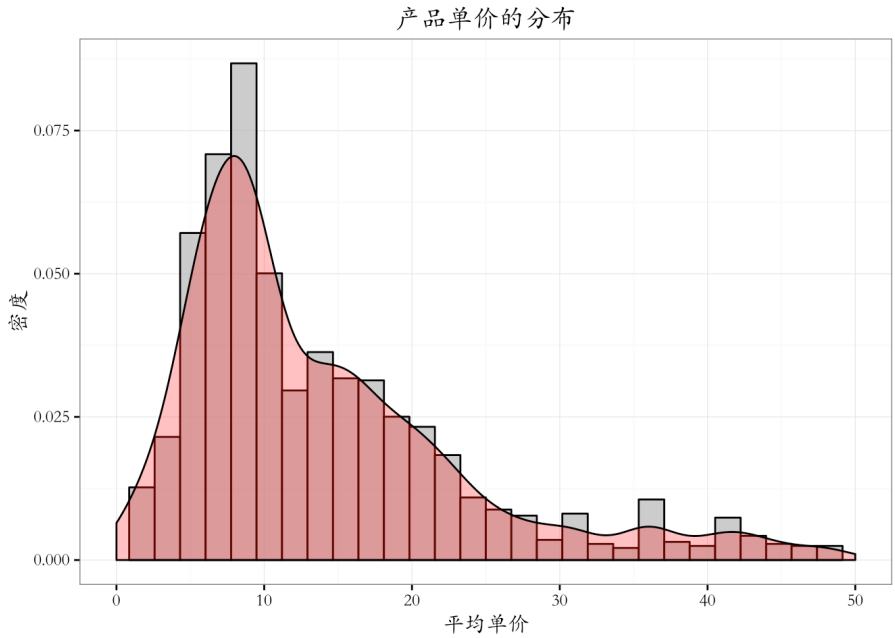


图 16：产品单价分布

从上图可以看出：每种产品的平均单价主要集中在（0,20）之内，其中单价为 8 的产品品种最多。该公司也销售单价大于 30 的较贵产品，但均不是主打产品。

接下来我们分析产品的种类与销售站的关系，得到的结果如下：

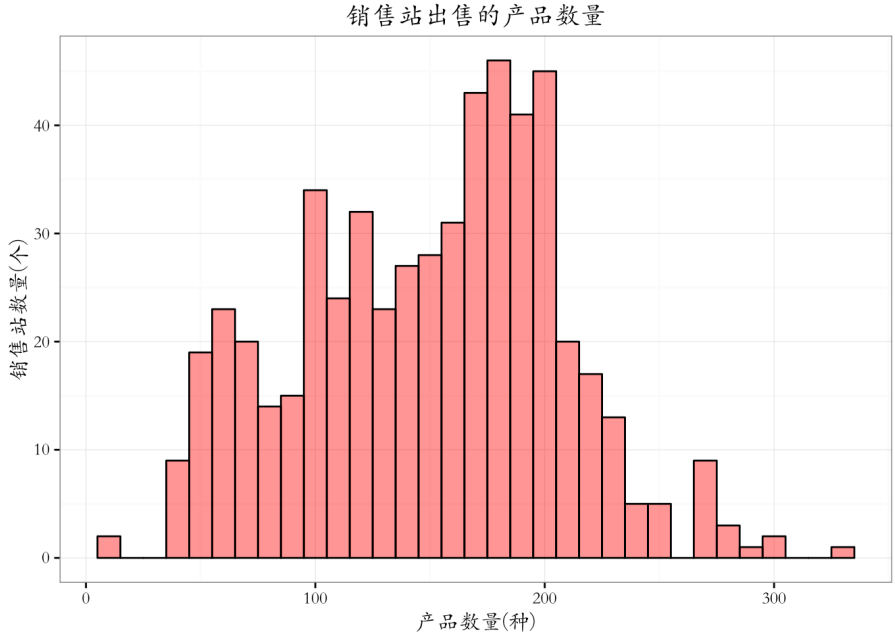


图 17：销售站出售的产品数

从上图可以看出，大部分的销售站点会销售 100～200 种的该公司产品；同时也有几个销售站，只销售该公司的几种产品；虽然该公司销售 1000 多种产品，但是没有哪个销售站会出售该公司的所有产品。

接下来分析销量前 50 的产品和前 50 的销售路线之间的交互关系，如下图：

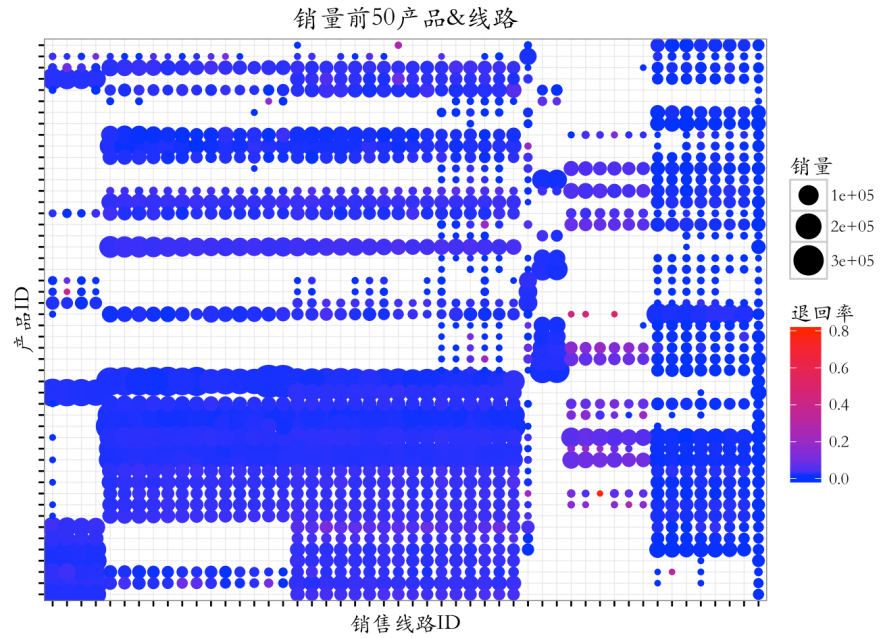


图 18：产品与路线的关系

从上图可以看出：从图中可以看出，在图的左下侧点的分布较为密集，点较大，颜色偏蓝，说明销售量大相应的退回百分比较小。而且也显示出了产品的销量越小，退回率越大。

最后分析产品量和客户量之间的关系，得到如下图所示：

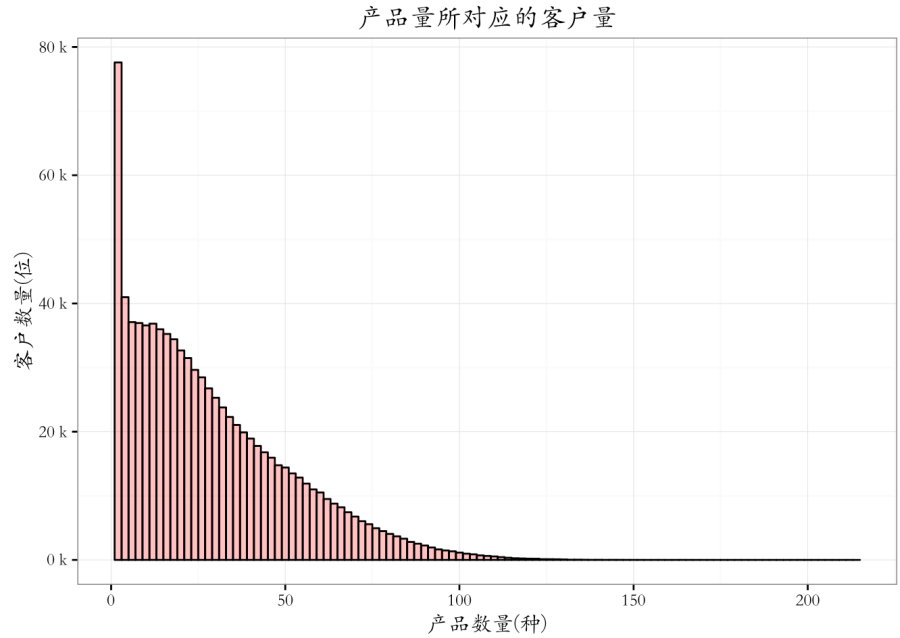


图 19：产品量所对应的客户量

从上图可以看出：大部分的客户消费过该公司的产品数量小于 50，而且只消费过 1～2 种产品大约有 8 万客户，同时也有很少一部分客户消费过该公司超过 150 种的产品。

5、数据预测准备

5.1、数据变换

使用机器学习对需求量进行预测

在原始数据中需求量的数据分布较分散，它的五数具体数值如表：

表 2:原始需求量五数

Min.	1st Qu	Median	Mean	3st Qu	Max.
0	14	40	99.93	104	15120

从上表可以看出，数据的分布时有偏的，数据的分布图形如下:

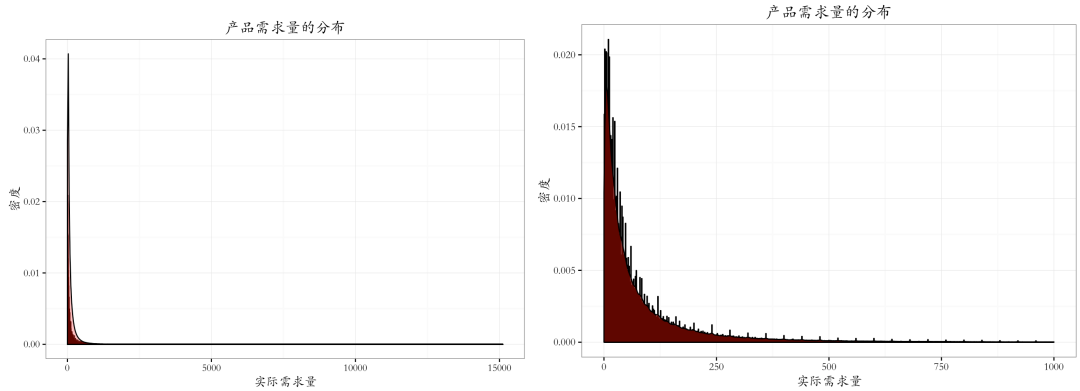


图 20：原始需求量数据分布

上图均为产品调整后需求量的分布，可以看出，需求量主要集中在 0～250 之间，只有极少数的需求量会在 1000 以上。为了使用机器学习方法进行预测时准确度的提升，需要对数据进行取对数变换，变换公式如下：

$$\tilde{y} = \ln(y + 1)$$

上式中：y:原始的预测变量

$\tilde{y}$:变换后的预测变量

上面的公式中之所以进行加 1 运算，是因为原始数据中存在 0，而对 0 进行自然对数运算会的到负无穷，只有加 1 才能得到将原始数据转化为近似正态分布。

预测变量变换后的数据的五数如下：

表 3:自然对数变换后五数

Min.	1st Qu	Median	Mean	3st Qu	Max.
0	2.89	3.892	3.824	4.828	9.624

数据的分布图如下图：

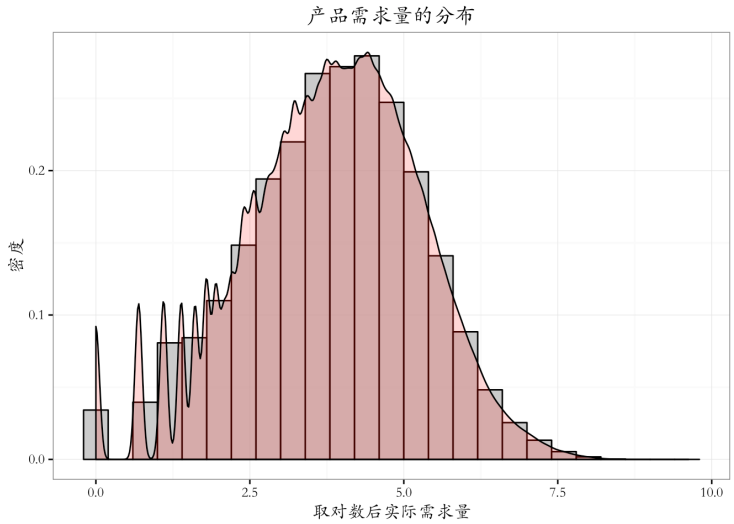


图 21：变换后需求量数据分布

从上图可以看出，变换后的数据近似服从正态分布，达到了我们训练模型，进行预测的要求。

在使用模型得到预测的数据后，并不是最终的需求量，需要对预测数据进行指数变换得到最终的预测数据，变换的公式如下：

$$pre = e^{\tilde{y}} - 1$$

上式中：pre:预测数据集的预测变量

$\tilde{y}$:预测模型预测得到的变量

5.2、训练数据集的划分

因为每时每刻都会有新的产品诞生，所以说下周的数据中会出现这周数据中没有的新产品销售信息。所以说我们需要将数据集按照星期划分为三个部分，分别为生成数据集、训练数据集、验证数据集，因为星期数量有限，所以可以存在某些数据集之间的数据存在相同周次的数据，以提高模型的预测能力。然后为了提高对新产品的预测能力，将新产品需求量数据使用所有类别产品需求量均值的中位数来代替。

数据集的划分步骤如下图：

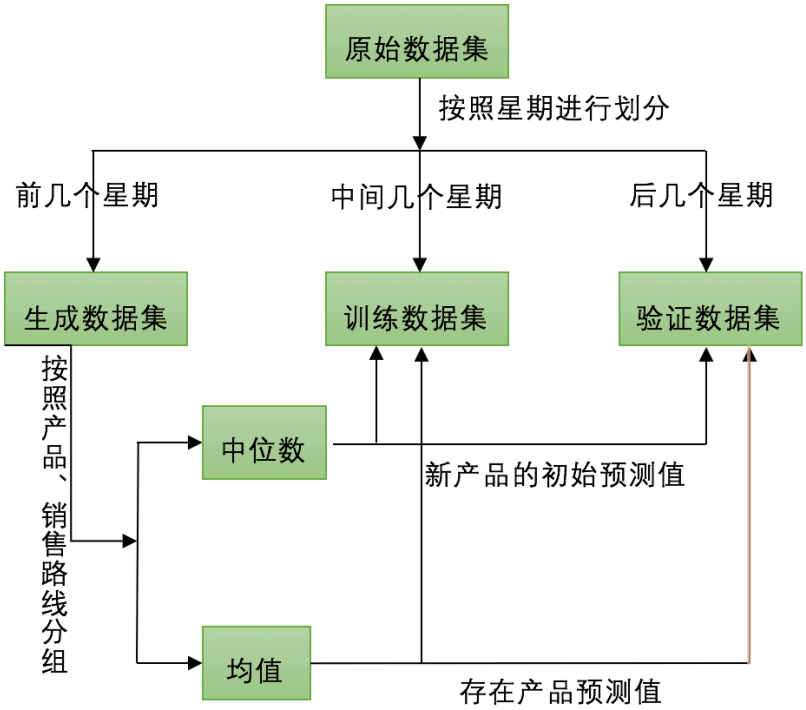


图 22：数据集划分流程图

通过上面的步骤，将原始的训练数据集，划分为方便进行机器学习预测需求量的数据。

6、随机森林回归模型

6.1、随机森林算法综述

随机森林算法(random forest)是由 Breiman 和 Cutler 在 2001 年提出的一种基于分类树的算法。它通过对大量分类树的汇总提高模型的预测精度，是取代神经网络等传统机器学习方法的新的模型。随机森林算法计算速度很快，在处理大数据时表现优异。随机森林步需要顾虑一般回归分析面临的多元共线性问题，不用做变量的选择，对离群值也不敏感。

随机森林可以解释若干字变量$(x_1, x_2, \cdots, x_k)$对因变量y的作用，在因变量为连续变量时，则进行随机森林回归。随机森林会随机地在原数据中重新选择 n 个观测值，其中有的观测值被选择多次，有的没有被选到，这是 Bootstrap 重新抽样的方法。同时，随机森林随机地从 k 个自变量选择部分变量进行回归树节点的确定。这样，每次构建的回归树都可能不一样。一般情况下，随机森林随机地生成几百个至几千个回归树，然后把每个回归树预测值进行加权平均，得到最终的回归结果。（Breiman，2001）

随机森林算法的一些特点：

（1）：随机森林是一个有效的预测工具。很多数据显示能够达到同提升算法和自适应装袋(Adaptive Bagging)算法一样好的效果，中间不需要反复改变训练集，对噪声的稳健性比提升算法好。

（2）：适合高纬输入变量的特征选择，不需要提前对变量进行删减和筛选 。

（3）：能够提高分类和回归问题的准确率，同时也能避免过拟合现象的出现。

（4）：当数据集中存在大量缺失值时，能够对缺失值进行有效的估计和处理。

（5）：能够在分类或回归过程中估计特征变量或者解释变量的重要性。

（6）：随着随机森林中树的增加，模型的泛化误差(generalization error)已经被证明趋于一个上界，这说明随机森林树对未知数据有较好的泛化能力。

6.2、随机森林算法求解流程

随机森林回归在该问题中的求解流程如下图：

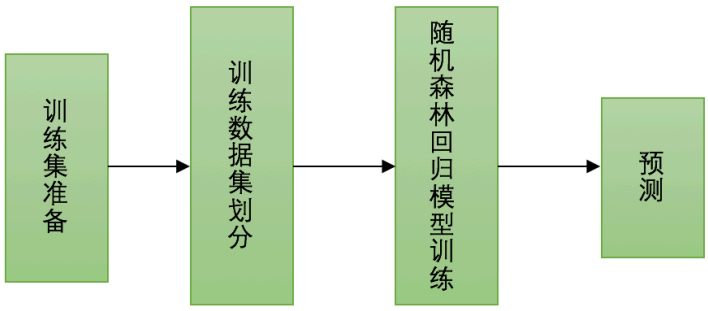


图 23：随机森林回归流程

我们借助 R 语言平台的 H2o 包（H2O 为初创公司 Oxdata 推出的开源机器学习项目）来进行随机森林回归模型的建立。使用 5、6、7、8 四周总共 4159 万多条数据作为训练集来训练模型，使用 8、9 两周总共 2081 万多条数据作为测试集，来评测模型的好坏。

6.3、随机森林回归求解结果

训练拥有 200 棵回归树的随机森林回归模型，得到的模型结果主要信息如下图：

```
## Model Details:
## ==============
##
## H2ORegressionModel: drf
## Model Key:  randomForest1
## Model Summary:
##   number_of_trees model_size_in_bytes min_depth max_depth mean_depth
## 1             200            60660176        20        20   20.00000
##   min_leaves max_leaves mean_leaves
## 1        368     104998 26256.45500
##
## H2ORegressionMetrics: drf
## ** Reported on training data. **
## Description: Metrics reported on Out-Of-Bag training samples
##
## MSE:  0.5565291
## R2 :  0.3818045
## Mean Residual Deviance :  0.5565291
##
##
## H2ORegressionMetrics: drf
## ** Reported on validation data. **
##
## MSE:  0.557135
## R2 :  0.3774114
## Mean Residual Deviance :  0.557135
```

图 24：随机森林回归结果

从上面的回归模型结果可以看出，200 棵树的在训练集中的到的平均偏差为 0.5565、$R^2 = 0.3818$，测试集中的平均偏差为 0.5571、$R^2 = 0.3774$。可以从偏差大小看出随机森林回归模型的拟合效果较好。

各个变量的重要程度如下表：

表 4:随机森林回归变量重要程度

随机森林回归变量最要程度			
variable	relative_importance	scaled_importance	percentage
Producto_ID	1136510336.000000	1.000000	0.385112
Canal_ID	985370368.000000	0.867014	0.333897
Ruta_SAK	581313536.000000	0.511490	0.196981
Agencia_ID	153383392.000000	0.134960	0.051975
Cliente_ID	94541512.000000	0.083186	0.032036

从上表可以看出占比最大的两个变量为产品的 ID、销售路线。重要性最小的变量为客户 ID。

使用随机森林模型对所给的约 700 万条测试集进行预测，得到的预测结果数据分布情况如下：

表 5：预测结果五数

Min.	1st Qu	Median	Mean	3st Qu	Max.
0	4	4	4.965	5	102

预测数据的分布直方图如下图：

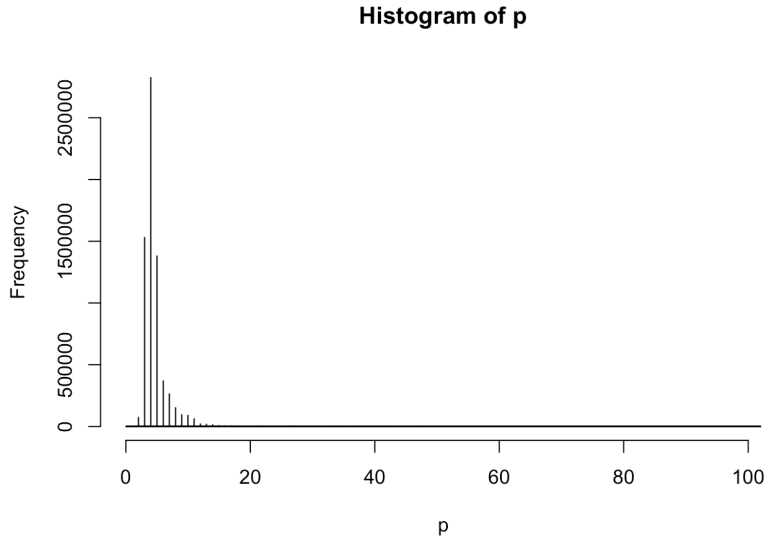


图 25：随机森林回归预测结果分布

从随机森林预测出的结果中可以发现，预测的结果虽然很符合原始数据的分布是相似的，但是并没有出现数值大于 102 的预测需求数值。可以看出随机森林回归预测的结果比较保守。

7、梯度提升机回归模型

7.1、Gradient Boosted Machines 综述

Boosting

Boosting 是一族可将弱学习器提升为强学习器的算法。这族算法的工作机制类似：先从初始训练集训练出一个基学习器，再根据基学习器的表现对训练样本分布进行调整，使得先前的基学习器做错的训练样本在后续受到更多的关注，然后基于调整后的样本分布训练下一个基学习器；如此重复进行，直至基学习器数目达到事先指定的值 T，最终将这 T 个基学习器进行加权结合。

Boosting 算法要求基学习器能对特定的数据分布进行学习，可通过“重赋权法”(re-weighting)实施，即在训练过程的每一轮学习中，根据样本的分布为美衣个训练样本重新赋一个权重。对无法接受带权样本的基学习算法，可以通过“重采样法”(re-sampling)来处理，即在每一轮学习中，根据样本分布对训练集重新进行采样，再用重采样的样本集对基学习器进行训练。

Gradient Boosting

Gradient Boosting 是一种 Boosting 的方法，它的主要思想是，每一次建立模型是在之前建立模型损失函数的剃度下降方向。损失模型描述的是模型的不靠谱程度，损失函数越大，则说明模型约容易出错。如果我们的模型能够让损失函数持续下降，则说明我们的模型在一直的改进，而最好的方法就是让损

失函数在其梯度的方向上下降。当该模型的预测数据为连续数据时，则可视为梯度提升机回归模型。

7.2、梯度提升机算法流程

使用梯度提升机回归模型(Gradient Boosted Machines)对该问题进行求解释，求解流程图如下：

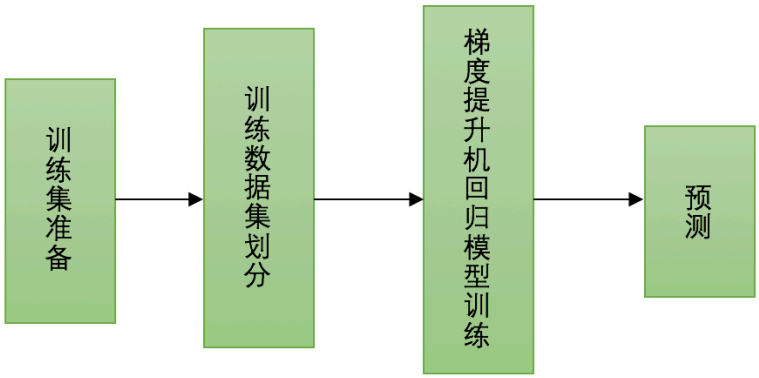


图 26：梯度提升机回归流程

我们借助 R 语言平台的 H2o 包（H2O 为初创公司 Oxdata 推出的开源机器学习项目）来进行梯度提升机回归模型的建立。使用 5、6、7、8 四周总共 4159 万多条数据作为训练集来训练模型，使用 8、9 两周总共 2081 万多条数据作为测试集，来评测模型的好坏。

7.3、梯度提升机回归求解结果

我们训练拥有 25 棵回归树的梯度提升机回归模型，可以得到的模型结果如下图所示：

```
## Model Details:
## ==============
##
## H2ORegressionModel: gbm
## Model Key:  gbm1
## Model Summary:
##   number_of_trees model_size_in_bytes min_depth max_depth mean_depth
## 1              25               10747         5         5    5.00000
##   min_leaves max_leaves mean_leaves
## 1         28         32    31.68000
##
## H2ORegressionMetrics: gbm
## ** Reported on training data. **
##
## MSE:  0.3820492
## R2 :  0.4587036
## Mean Residual Deviance :  0.3820492
##
##
## H2ORegressionMetrics: gbm
## ** Reported on validation data. **
##
## MSE:  0.3902167
## R2 :  0.4475754
## Mean Residual Deviance :  0.3902167
```

图 27：梯度提升机回归结果

从上面的回归模型结果可以看出，25 棵回归树在训练集中得到的平均偏差为 0.382、$R^2 = 0.4587$，测试集中的平均偏差为 0.3902、$R^2 = 0.4476$。可以从偏差大小看出该模型的拟合效果较好。

各个变量的重要程度如下表：

表 6:梯度提升机变量重要程度

梯度提升机变量重要程度			
variable	relative_importance	scaled_importance	percentage
Ruta_SAK	1120714.750000	1.000000	0.595817
Producto_ID	281057.156250	0.250784	0.149421
Agencia_ID	211715.000000	0.188911	0.112556
Cliente_ID	201189.531250	0.179430	0.106907
Cacal_ID	66396.335938	0.059245	0.035299

从上表可以看出，梯度提升机回归模型中最重要的变量是销售路线（每一个销售路线代表一个仓库），而最不重要的变量为销售途径。这和随机森林回归模型的到的结果有很大的不同。

使用梯度提升机回归模型对所给的约 700 万条测试集进行预测，得到的预测结果数据分布情况如下：

表 7: 梯度提升机回归预测结果五数

Min.	1st Qu	Median	Mean	3st Qu	Max.
−0.885	2.172	2.172	5.343	5.229	3056.159

从上面的五数表中可以看出最小值小于 0，因为需求量不能小于 0，所以将小于 0 的数据用 0 来代替，并且对数据进行四舍五入的到整数值，计算出新的预测值五数表如下：

表 8：调整后梯度提升机回归预测结果五数

Min.	1st Qu	Median	Mean	3st Qu	Max.
0	0	3.057	5.356	3.057	3056

得到的预测需求量的直方图分布如下：

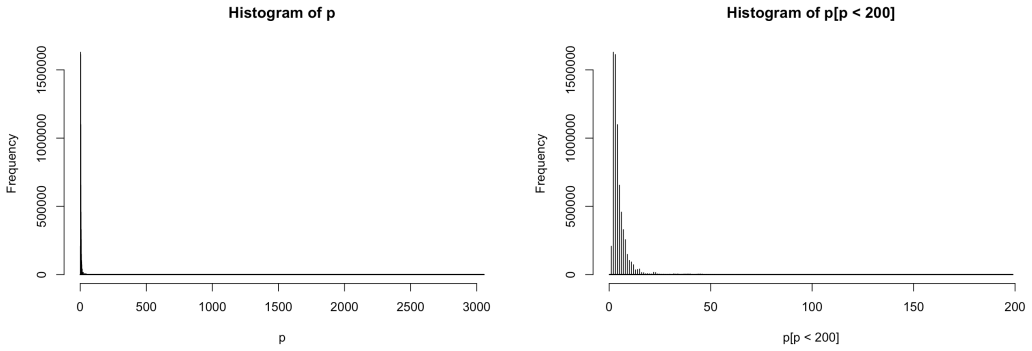


图 28:梯度提升机回归预测值分布

从上面的预测结果可以看出主要的预测需求量集中在 10 左右，但是也有需求量很大的预测值，如最大值为 3056，这样的预测数据和训练集的数据分布更加接近。

8、预测模型对比

我们使用了两种机器学习方法，对目标变量进行了预测，下面我们通过两个模型的多方面表现，对两模型的好坏进行对比，详情见下表：

表 9：模型参数对比表

相关参数	随机森林回归模型	梯度提升机回归模型
回归树数量	200	25
平均深度	10	5
测试集平均偏差	0.5571	0.3902
测试集R^2	0.3774	0.4475
最重要变量	产品 ID	销售路线 ID
最不重要变量	客户 ID	销售渠道 ID
模型训练时间	约 60 分钟	约 3 分钟
预测结果	更加保守	更加大胆

从上面的两种机器学习预测方法中，可以发现梯度提升机回归模型的的平均偏差更小，相关系数的平方R^2更大，所以得到的预测效果更好。而且梯度提升机回归模型只需要训练 25 棵回归树，花费 3 分钟，预测精度就超过随机森林回归 200 棵回归树达到的预测精度，大大减少了训练模型花费的时间，所以在此问题中，梯度提升机回归模型的效果更好。

9、参考文献

[1] 李欣海　随机森林模型在分类与回归分析中的应用[J]
[2] 吴喜之 复杂数据统计方法－基于 R 的应用[M] 中国人民大学出版社
[3] 周志华 机器学习[M] 清华大学出版社
[4] 王星 大数据分析：方法与应用[M] 清华大学出版社
[5] 黄文 王正林 数据挖掘：R 语言实战[M] 电子工业出版社
[6] kaggle 网站　https://www.kaggle.com/
[7] Jared P. Lander(美)著　蒋家坤 译　R 语言：实用数据分析与可视化技术[M]　机械工业出版社
[8] Greg Ridgeway　Generalized Boosted Models:A guide to the gbm package
[9] 马春鹏 模式识别与机器学习[M]
[10] Package“h2o” R 语言 H2o 包的帮助文档

10、附录

1:数据探索 R 语言程序

因为程序太长，所以这里只贴出部分程序，详细的 R 程序见相应的 PDF 文件

```
## 2016 年暑期课程设计
## 问题：Grupo Bimbo Inventory Demand
##  宾堡集团的库存需求
## 最大限度地提高销售和最大限度地减少烘焙食品的退回
## Daitu
## start:2016.06.21
## 参考借鉴 kaggle 上的公开程序
setwd("/Users/Daitu/数据分析/kaggle/Grupo Bimbo")
getwd()
## 加载包
library(data.table)
library(ggplot2)
library(dplyr)
library(treemap)
## 读取数据####
## 1:读取训练集
system.time({
  traindata <- fread("train.csv",sep=",",header = TRUE)
})
head(traindata)
## 2:读取客户名单数据
cliente_tabla <- fread("cliente_tabla.csv",sep=",",header = TRUE)
head(cliente_tabla)
## 3:读取产品名单数据
producto_tabla <- fread("producto_tabla.csv",sep=",",header = TRUE)
head(producto_tabla)
## 4:读取城镇和国家（州）数据
town_state <- fread("town_state.csv",sep=",",header = TRUE)
head(town_state)
## 数据的描述统计####
## 1:分析数据的周数：Semana
Semana <- data.frame(table(traindata$Semana))
colnames(Semana) <- c("Semana","Freq")
# 3        4        5        6        7        8        9
# 11165207 11009593 10615397 10191837 10382849 10406868 10408713
## 条形图
ggplot(data = Semana,aes(Semana,Freq)) +
```

```
  geom_bar(stat = "identity", width = 0.6,fill = "lightblue",colour =
"black") +
  theme_grey(base_family = "STKaiti") +
  scale_y_continuous() +
  labs(x="周数",y="数据条数",title = "每周数据的记录数")
Semana2 <- traindata %>%
  group_by(Semana) %>%
  summarise(Units = sum(Venta_uni_hoy),
            Return_Units = sum(Dev_uni_proxima),
            NetU = sum(Demanda_uni_equil)) %>%
  mutate(Retern_Rate = Return_Units / (Return_Units + Units))
Semana2
dim(Semana2)  # 几个星期的数据
## 可视化每周的情况
ggplot(data = Semana2,aes(Semana,NetU,fill = Retern_Rate)) +
  geom_bar(stat = "identity",color = "black",width = 0.7) +
  theme_bw(base_family = "STKaiti") +
  scale_y_continuous(labels=function(x)paste(x/100000, "m"))+
  scale_x_continuous(labels = function(x)paste("第",x,"星期")) +
  scale_fill_gradient(name="退回率", low="blue", high="red")+
  labs(x = "星期",y = "销量",title = "每星期销量")
```

2:随机森林预测的关键程序：

```
## 训练 randomForest; 使用参数以保持整体运行时间在 20 分钟内
## this model is fit on Semana 6 & 7 & 8, and evaluated on Semana 9.
rf<-h2o.randomForest(
  x=predictors,          ## 建立模型的预测变量
  y="target",   ## target: using the logged variable created earlier
  training_frame = newVal,   ## H2O frame holding the training data
  validation_frame = newFinal,  ## extra holdout piece for three
layer modeling
  model_id="randomForest1",      ## internal H2O name for model
  ntrees = 200,             ## 使用 200 棵树建立模型
  sample_rate = 0.8,        ## use 80% of the rows each scoring round
  weights_column = "mean_target"
)
## 查看模型
summary(rf)
```

3:梯度提升机回归 R 程序

```
## 训练 GBM; 使用参数以保持整体运行时间在 20 分钟内
## this model is fit on Semana 6 & 7 & 8, and evaluated on Semana 9.
g<-h2o.gbm(
  training_frame = newVal,       ## H2O frame holding the training
data
```

```
  validation_frame = newFinal,  ## extra holdout piece for three
layer modeling
  x=predictors,                 ## 建立模型的预测变量
  y="target",                   ## target: using the logged variable
created earlier
  model_id="gbm1",              ## internal H2O name for model
  distribution = "gaussian",        ## 目标数据的分布
  ntrees = 25,                  ## 使用 25 棵树建立模型
  learn_rate = 0.3,             ## lower learn_rate is better, but
use high rate to offset few trees
  score_tree_interval = 5,      ## score every 5 trees
  sample_rate = 0.6,            ## use 60% of the rows each scoring
round
  col_sample_rate = 0.8,        ## use 4/5 the columns to decide each
split decision
  offset_column = "mean_target"
)
## 查看模型
summary(g)
```

11、分工情况

相关工作	孙玉林 1308064146	李姝 1308064120	张肖 1308064113	吴迪 1308064125
确定问题	是	是	是	是
问题分析	是	是	是	是
数据探索	是	是	是	是
模型确定	是	是	是	是
程序实现	是			
论文摘要	是			
论文数据探索		是	是	是
论文机器学习	是			
论文参考文献		是		
论文附录			是	
论文排版				是
PPT 制作		是	是	是
PPT 修改	是	是	是	是