基于XGBoost的Rossmann商店销量预测分析

朱林

2018/4/5

# 一、问题定义

## 1.1、项目概述

在商业数据分析领域，对于分析的准确性要求越来越高，之前的主观推测方法，由于掺杂了很多人为因素，容易导致误判，所以很难满足要求。如今随着数据量增多、计算机算力提升迅速这两个因素的影响下，机器学习的方法有了天然的土壤，于是有很多机器学习的方法得以运用起来。本研究是基于欧洲一家连锁药店-Rossmann药妆商店的数据进行的分析，数据包括了：1115家商店的相关信息、一百万多条关于上述商店的日销量等信息。分析目的是，试图通过给定的销售量等数据，分析得出在接下来的一段时间内，上述商店的销售情况。

首先要对德国的商业环境有所了解，德国的法律规定，只有在少数周日商店才可以开门。[[1]](#footnote-1)可想而知，周日的销量可能不会和weekday的销量有太大差别。但是由于周日不开门，因此周六商店会开门很长时间，所以周六的销量可能会大大高于平日。基于上述信息，首先对给出的数据有有一定的了解。kaggle给出的数据主要包括三类：store-商店相关信息；train-商店的日销量等信息；test-测试数据。[[2]](#footnote-2)有哪些因素会影响一家商店的销量呢？日常逛街的时候，主要关注的信息有：星期几、商店是否有折扣，是否是假期、商店类型、到店人均消费。另外，还会关注周边的信息，比如商店和其他类似商店的远近比较。这些信息大多数都在给定的数据中可以找得到，另外一些需要经过数据处理才能得到。当然，还有一些周边信息，没有在数据中体现，比如当时的天气状况、气温等信息，这些信息也会影响人们去商店的意愿。

上述提到的机器学习方法中，有监督学习和非监督学习、半监督学习。另外现在图像识别问题中最常用的应该是深度学习方法。随着 Alpha Go 的巨大成功，宣告人类在围棋这个极其复杂的棋种，已经不可能超越机器了，机器也不再像以前一样需要人工输入大量棋局和指导，因为他们能够自己进行学习，强化学习让大家眼前一亮。该问题中，由于我们是基于给定的数据和标签结果，通过计算得出我们的判断，因此我们需要使用的是监督学习模型。

## 1.2、问题陈述

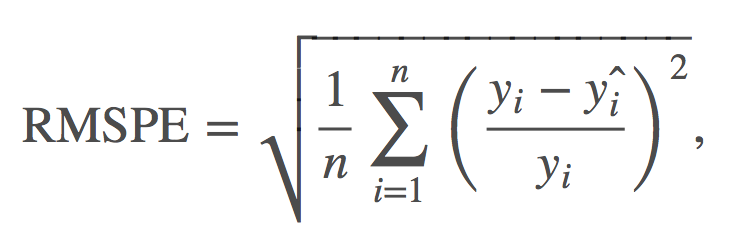
由于需要预估的销售量，是一个连续的数据，因此该问题属于一个回归问题，而且是有监督的回归问题。有监督学习的模型中，我选择 XGBoost 作为训练结果，因为他作为boost类别的方法，能够集合多个弱分类，综合判断得到结果，往往比单个模型得到的结果要好。

该问题的数据分为商店数据和训练数据，所以首先需要进行合并操作。其次，需要对数据进行一些预处理，去掉一些误差点等信息。最后再进行 XGBoost 模型的训练和调参。

由于该问题是基于 kaggle 的项目，因此项目最终结果需提交 kaggle 得到最终的结果。总共有3303支队伍参赛，评价标准是，起码需要达到总排行榜的前15%，力争进入前10%。由于有两个排行榜，Private榜 和 Public榜，分别表示随机取所有数据的 61% 和 39%。因此我选取更小随机性的 Private榜作为参照，进入前15% 和10% 的分数分别为：**0.11959** 和 **0.11773** 。

## 1.3、评价指标

最终需要输出一个csv文件，包含所有test数据的预测销量结果，将其提交到kaggle进行评价。该项目 kaggle 基于的评价指标 RMSPE，表示均方根误差的百分比值，能够将误差放到同一个量纲上进行比较。它的计算公式如下：



**y\_i** ： 销量误差的实际值

**yhat\_i** ： 销量误差的预测值

**n** ： 测试集的数目

当然，该判断中，排除了销量为0的数据不计算在内。

# 二、分析

## 2.1、数据的探索

数据集包括三个表格：

**store.csv-商店相关数据：**

**"Store"**：int，商店编号，是商店唯一编码

**"StoreType"**：object，商店类型，a,b,c,d

**"Assortment"**：object，产品级别，a=basic,b=extra,c=extended

**"CompetitionDistance"**：float，竞争对手距离，可以看成是，李竞争对手越近，顾客被分流的可能性就越大

**"CompetitionOpenSinceMonth"**：int，竞争对开门月份，和竞争对手开门年份合起来组成竞争对手开门时间，表示可能受到竞争对手影响的时间长度。

**"CompetitionOpenSinceYear"**：int，竞争对开门年份，同上。

**"Promo2"**：int，是否开展促销2。该促销是一个长周期的促销活动。0 代表商店未参与该活动，1 代表商店参与该活动。

**"Promo2SinceWeek"**：int，开展促销2的周。和促销2年份字段一起，表示促销2开展持续了多久。

**"Promo2SinceYear"**：int，开展促销2的年份，同上。

**"PromoInterval"**：object，促销间隔，表示促销2在一年的哪些月份开展。刚开始进行促销的时候，往往对于用户的吸引力最大。所以开始开展促销2活动的月份字段也是很重要的。

**train.csv-训练数据：**

**"Store"**：int，商店编号，是商店唯一编码

**"DayOfWeek"**：int，周几

**"Date"**：date，这条数据的日期

**"Sales"**：float，销量，该字段是我们需要预测的字段

**"Customers"**：int，客户数量，该字段在测试数据中不存在，所以不能依照该字段进行预测，但是为了充分榨取所有字段的剩余价值，将该字段与其他字段进行组合，构成 store 的特性，例如：客单价、每日客户数量等字段。

**"Open"**：int，该商店在当天是否开门，训练数据中有几条记录是没有开门与否的信息的，一般情况下是将其丢弃，但是我认为数据都有其存在的价值，这里不妨将其设置成open=1，即当天该商店是开门的状态。

**"Promo"**：int，是否进行促销活动，明显，由于是否进行促销对于当天的销量影响不小，因此这个字段也颇为重要。

**"StateHoliday"**：object，当天是否是国家法定假日，如果是假日，对于销量也会有很大影响，不同类型的国家假日对于销量的影响情况也不一样。a = public holiday, b = Easter holiday, c = Christmas, 0 = None。

**"SchoolHoliday"**：int，学校假日，表明当天该商店是否受到学校房价的影响，比如学校放寒假、暑假。依据经验，寒暑假的时候，小孩在家，大人应该会有更多时间陪小孩，所以逛街的可能性也会增加。

**test.csv-测试数据**

"Id","Store","DayOfWeek","Date","Open","Promo","StateHoliday","SchoolHoliday"

字段信息如上述描述所示。另外需要注意的是，最终进行测试的数据不会只有这几个字段，还应该包括之前提到的客单价等派生信息。

## 2.2、探索可视化

### 2.1.1、销量分布情况：

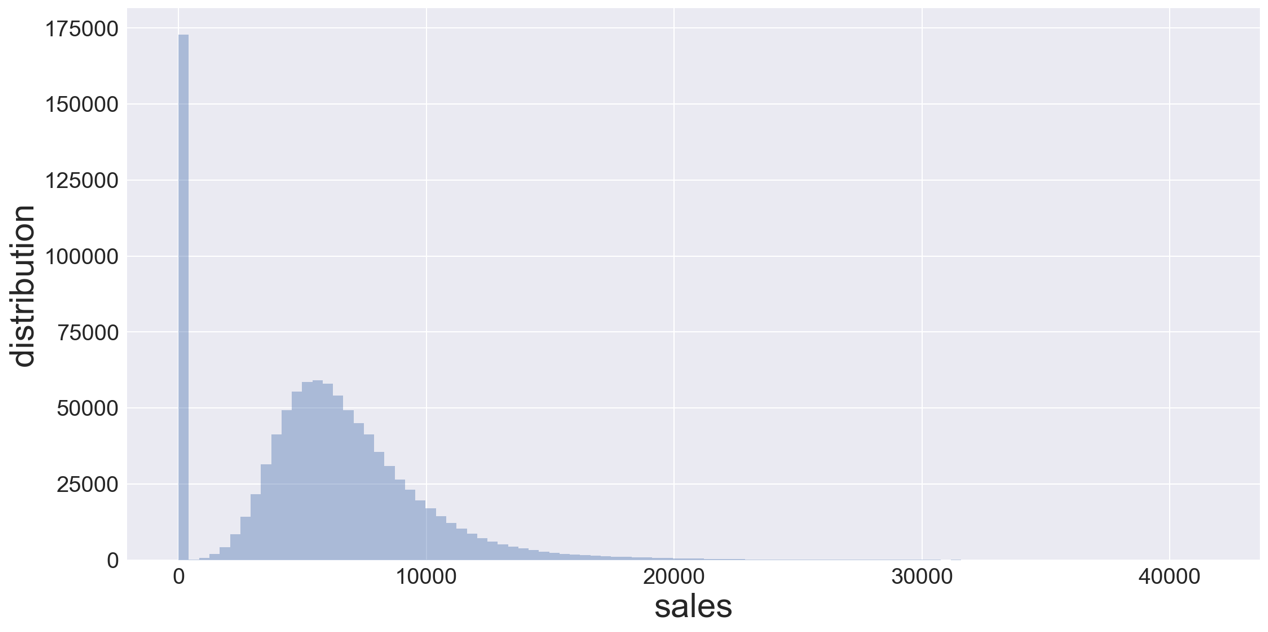


图1

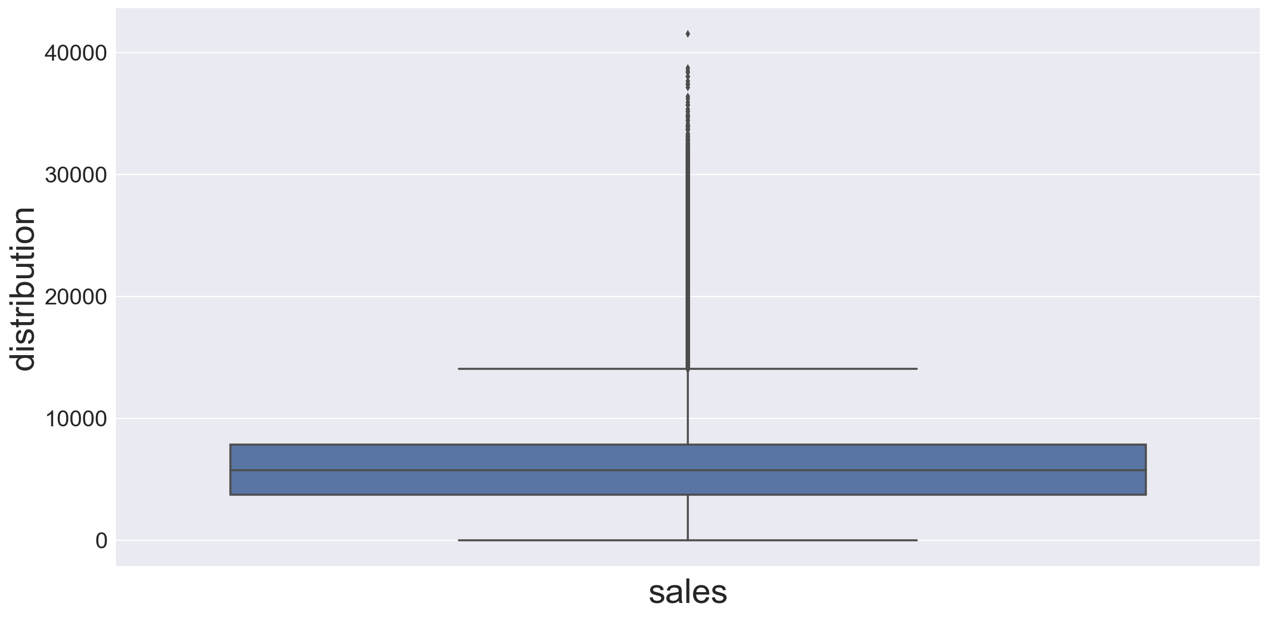


图2

**结论：**

图1，可以看出，除开高达150000+条销量为0的记录外，其余数据还是比较服从期望销量为5000左右的近似正态分布。

图2、可以看出，Sales的分布状况是一个长尾分布，超过40000之后的分布就很少了，可能存在离群值需要处理。

### 2.1.2、DayOfWeek分布状况：

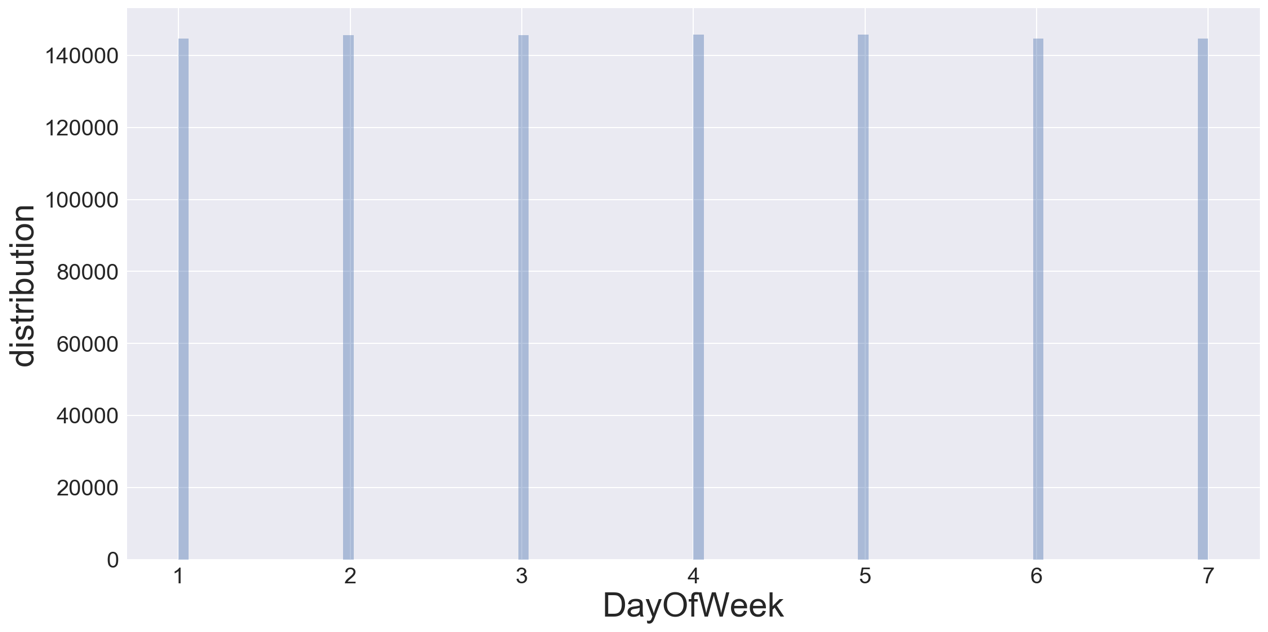


图3

注意，1-7分别代表的是周日开始到周六结束的分布情况。

**结论：**

1、一周内每天的数据都比较接近，不大可能出现因为采样的偏差导致最终对于某些天预测不准确的情况。

2、数据量都超过了140000万条，足够我们开展训练了。

### 2.1.3、Store分布状况：

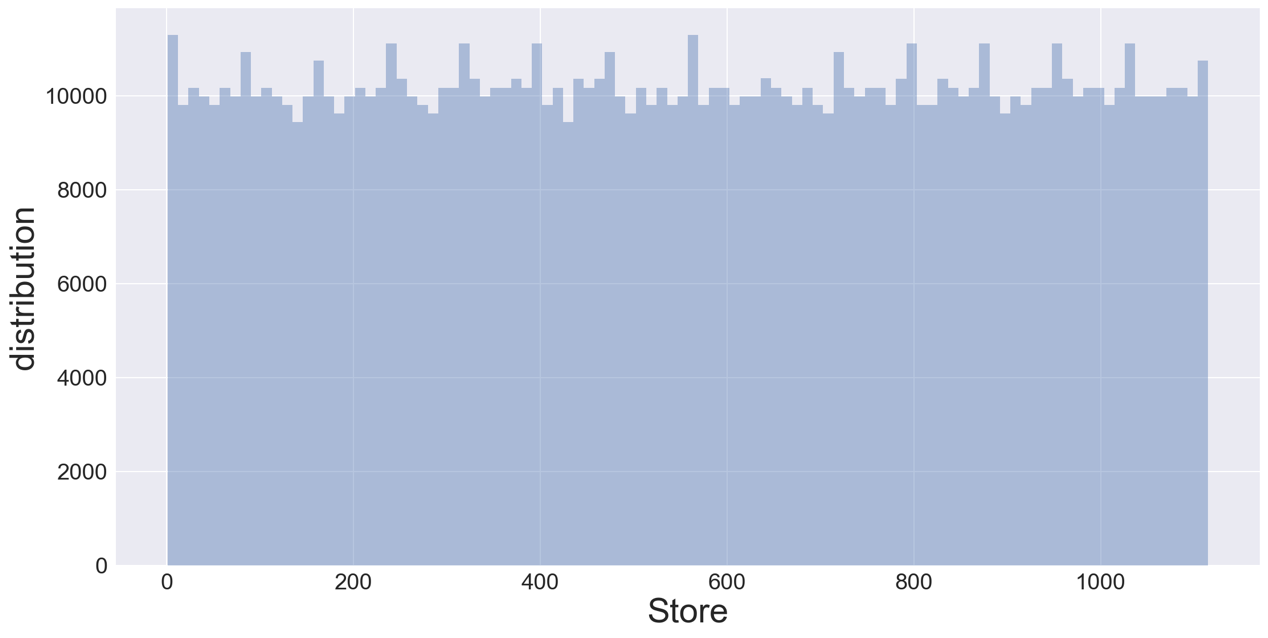


图4

**结论：**

对于1115家商店来说，采样分布均匀，同理也不会出现某些商店训练数据过少导致的不均匀的情况。

### 2.1.4、holiday对于销量的影响

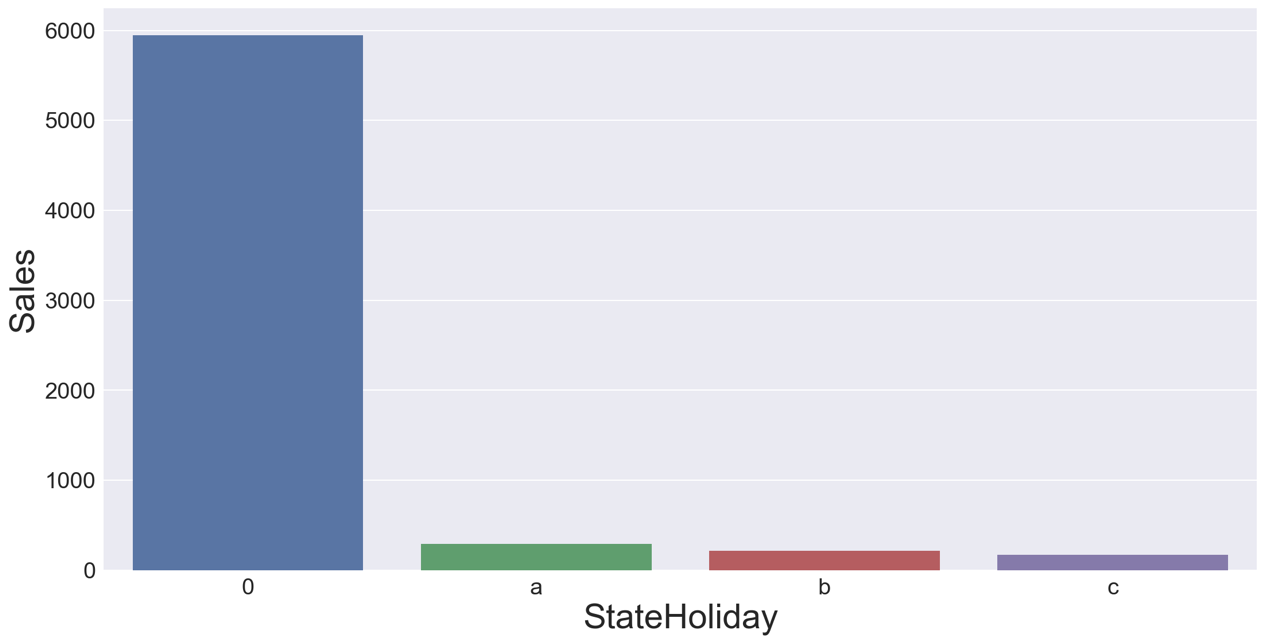


图5

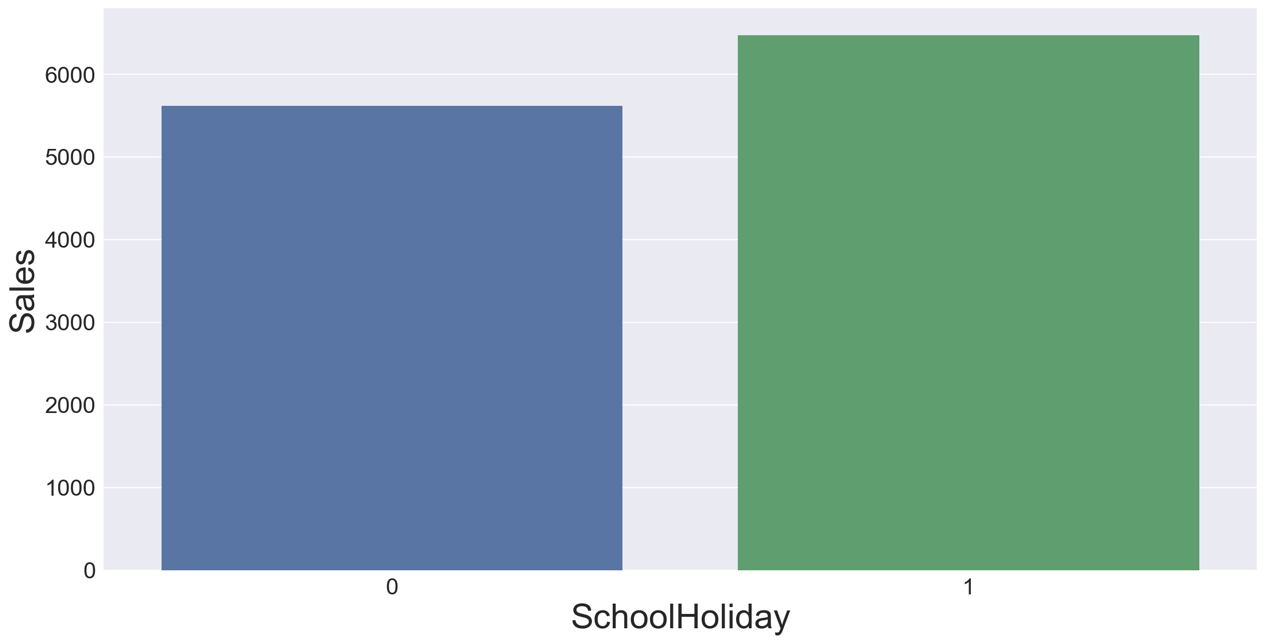


图6

**结论**

1、对于法定假日，无论是什么类型的假日，销量都会明显低于非节假日。

2、对于学校假日和非学校假日，前者的销量要比后者高很多，这也印证了之前的判断，即学校放假=》学生需要人陪=》家长带去逛街=》销量增加。

### 2.1.5、Customer与Sales之间的关系：

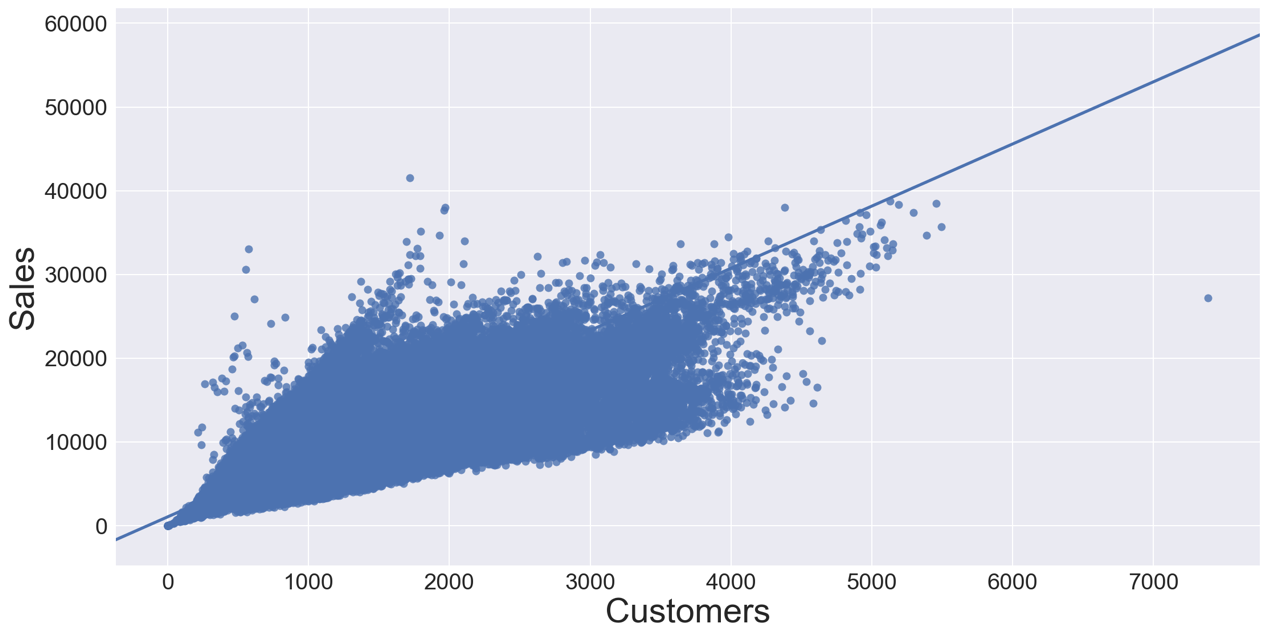


图7

**结论**

客户量与销量之间呈现一定的线性相关性，即客户量越大，销量也越大。说明，虽然客户量这个参数在测试数据中不存在（因为你无法预测未来的客户量，否则这就相当于利用了未来函数了），我们还是要重视这个参数，可以将其和其他参数综合之后，作为store的参数进行处理。

### 2.1.6、商店类型与产品类型

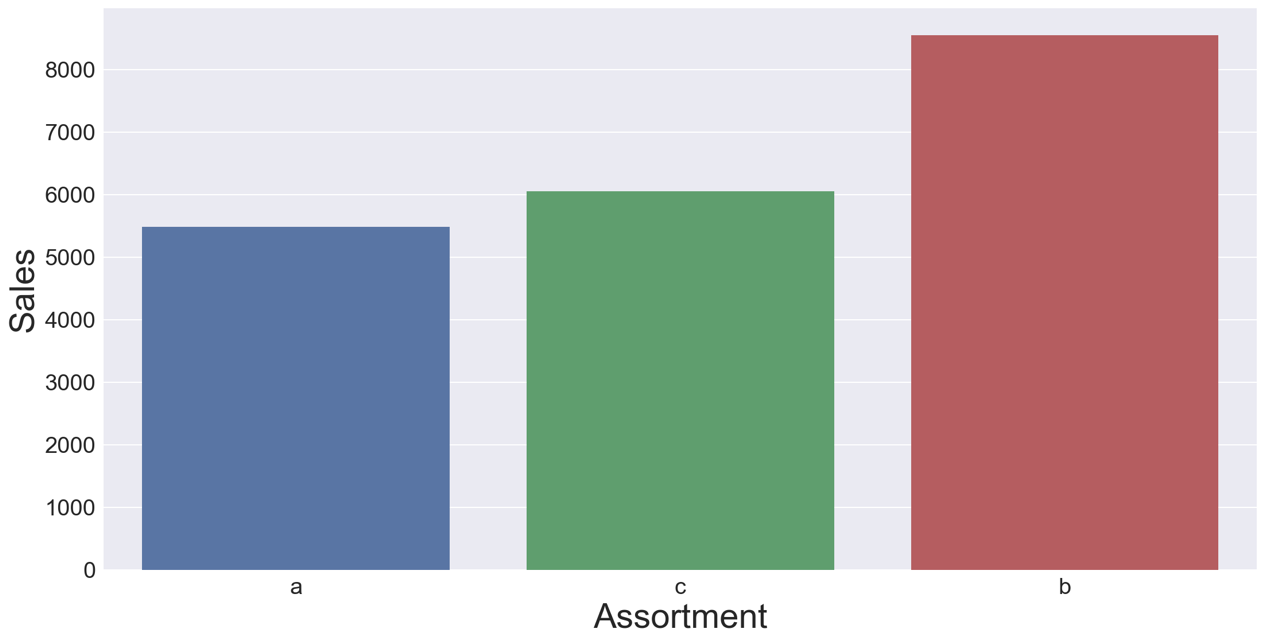


图8

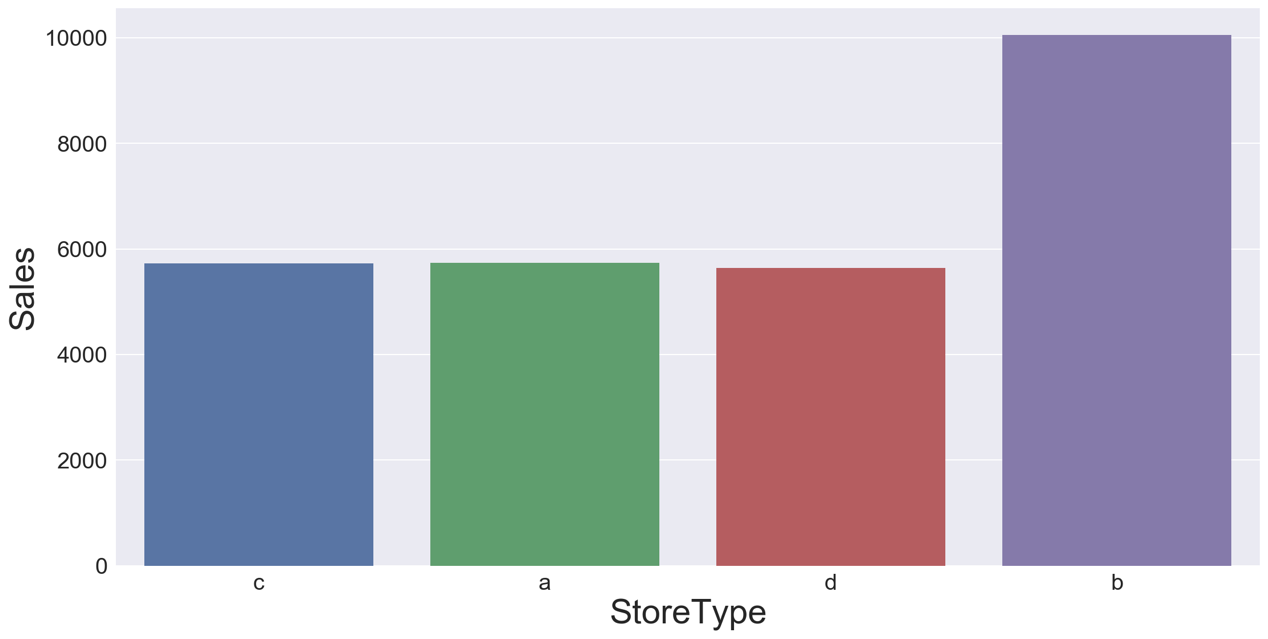


图9

**结论**

Assortment：a类型和c类型产品，销售量明显低于b类型的产品。

StoreType：a类型、c类型和d类型的商店，其销售数量明显低于b类型的商店。

说明以上两个参数，对于预测结果帮助是很大的。

### 2.1.7、promo促销与promo2促销对于销量的最终影响

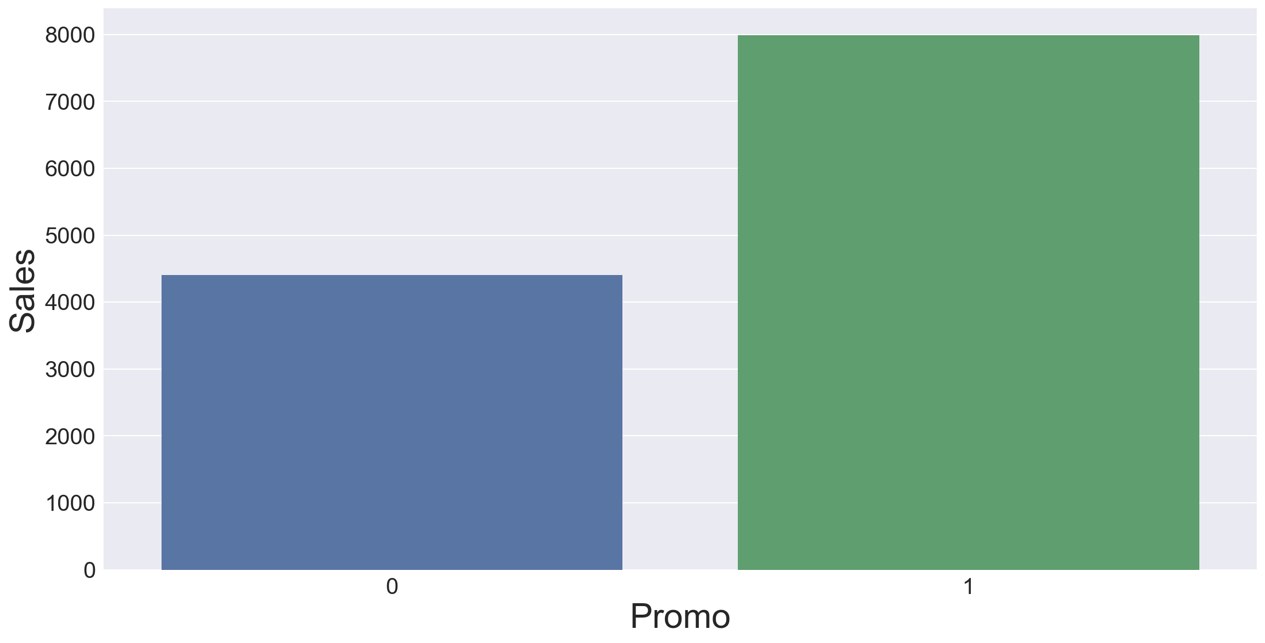


图10

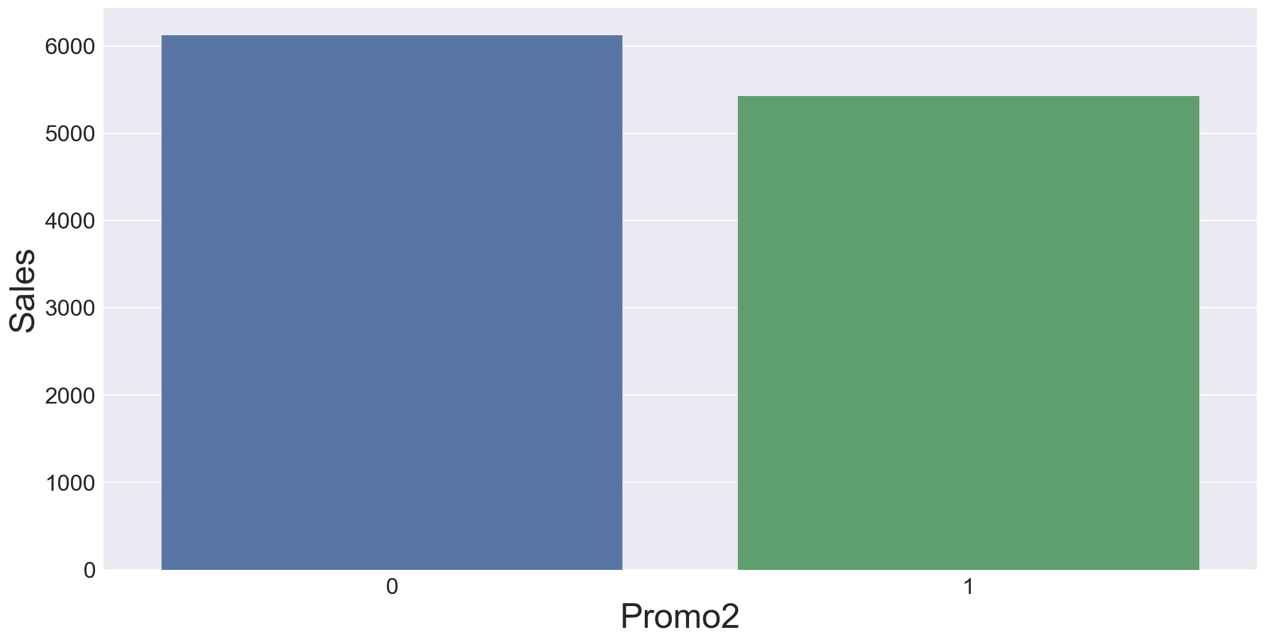


图11

**结论**

promo与promo2这两个参数，对于销售量sales，呈现两种截然相反的关系：

进行promo促销的商店当天销量要比没有进行促销的销量多了近一倍；但是进行promo2促销的商店，当天销量反而略低于没有进行促销的情况下的销量。

我觉得可能的原因是，进行短期促销，有利于用户从残存的记忆中快速对比出来，究竟价格和平时差了多少；但是长期促销，可能会让用户觉得这是一种很难销售出去的滞销品，进而不再进行购买。

## 2.3、算法和技术

此次分析需要用到的数据集由store、train和test三个部分组成，如果想要将数据划分好，首先需要做的就是如何将store很好的与train、test俩个数据集融合；另外，数据集合中“customer”属性在测试集中并不存在，所以需要额外的加工处理才行；数据存在一些缺失的情况，需要考虑到的对缺失数据的处理过程；数据集在训练集上也要切分为测试集与验证集，才能很好的确认训练过程的有效性。

### 2.3.1、预处理

经过上面的分析可以得出，数据都是来自于kaggle的项目。对于customer属性，如果丢弃该属性，这个属性的相关信息就不能很好的利用了，如果要将其利用起来，需要将其作为商店的属性进行处理。每家商店由于所处的地理位置的不同、商店面积的大小，药品的种类不同，都会带来人流量、客单价的不同。因此，将customer作为商店属性的一部分，能够充分利用属性中包含的信息。

另外，“Date”属性是一个object类型，但其实质上是一个“YYYY-MM-DD”的日期类型，所以需要将其处理，拆分成“year”、“month”、“day”三个字段，进行处理。

### 2.3.2、数据集拆分

测试集已经被kaggle单独区分出来，现在需要处理的就是训练集。由于我们需要判别训练过程的好坏，需要将train数据集合再进行切分，需要将其分为训练集和验证集两种类型，为了让训练过程对训练集合的利用最大化，采取了1%的数据集作为验证集、99%的数据集作为训练集进行处理。这样能够最大化的利用训练数据，同时也能兼顾到对训练过程的监控。

### 2.3.3、模型算法

由于kaggle赛事结束，对第一名的采访资料[[3]](#footnote-3)和他对自己训练参数的总结[[4]](#footnote-4)，我决定采用xgboost算法进行训练。xgboost全称是“extreme gradient boosting”，他是基于GBDT的加强版本，我认为选择该算法的理由有：

1、xgboost在代价函里面加入了正则化项，正则项里包含了叶子节点的数目、每个叶子节点输出的score的L2平方和。能够有效控制模型复杂度，减少过拟合的发生。

2、xgboost对于缺失值，首先会不考虑缺失数值，将其分别分到左子树和右子树，分别计算损失，然后再选取最优的那一个，如果训练时候没有缺失，预测时候出现了缺失，则默认分到右子树。[[5]](#footnote-5)

3、相比于GBDT以CART作为基分类器，xgboost还支持线性分类器，相当于利用了带L1和L2正则化项的logistic回归或线性回归。

4、xgboost首先会对特征进行排序并保存下来，然后再进行节点分裂时，计算节点特征增益就可以并行进行，节省时间。由于该数据集训练条数达到了100万，对于如此大数据量的训练，xgboost模型能够快速训练得出结果，节省时间。

基于以上讨论，数据经过预处理、数据集拆分，然后再利用xgboost算法进行训练，得到最优结果的可能性会大大增加。

## 2.4、基准模型

由于该分析是基于kaggle的项目进行的，对于结果的判断也需要上传到kaggle上才能得到最终得分[[6]](#footnote-6)。总共参赛队伍有3303支，我给自己定的目标是：

**优**-前10%；

**良**-前15%。

由于包含两个榜单-Priavte（基于61%的测试数据）和Public（基于31%的测试数据），因此我选取覆盖度更高的Private榜单作为模型的衡量标准。榜单进入前15% 和10% 的分数分别为：0.11959 和 0.11773 。这个结果是有参考意义的，因此我决定采用上述的基准来评价模型表现。

# 三、方法

## 3.1、数据预处理

3.1.1、异常值处理

首先我们需要对数据异常值进行处理，在预处理章节和可视化章节看到，数据集合存在着一些异常值，其可能的来源主要是机器原因或者人为原因导致的，它们对于预测过程带来很大影响，数据源头被污染，接下来无论怎么进行调参都无法达到满意的结果。

我们这里采用箱型图判别的方法进行处理，原理如图所示：

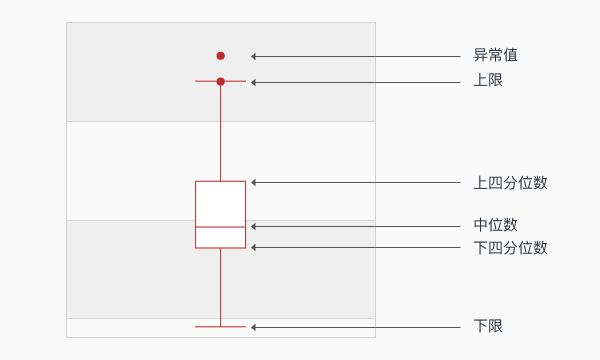


图12

上图可以表示数据的上限、下限、中位数，上四分位表示中位数和最大值之间的的中位数，下四分位表示中位数和最小值之间的中位数。一般用“Q1”代表上四分位，用“Q3”代表下四分位。用IQR表示Q1与Q3的差值，异常值被定义为大于（Q1+1.5\*IQR）和小于（Q3-1.5\*IQR）。

通过下图可以看到，在我对异常值进行去除之后，并没有影响到整体的分布，并且去除之后，长尾的那部分数据被去掉，整个数据看起来更加平滑，也更加贴近于正态分布。

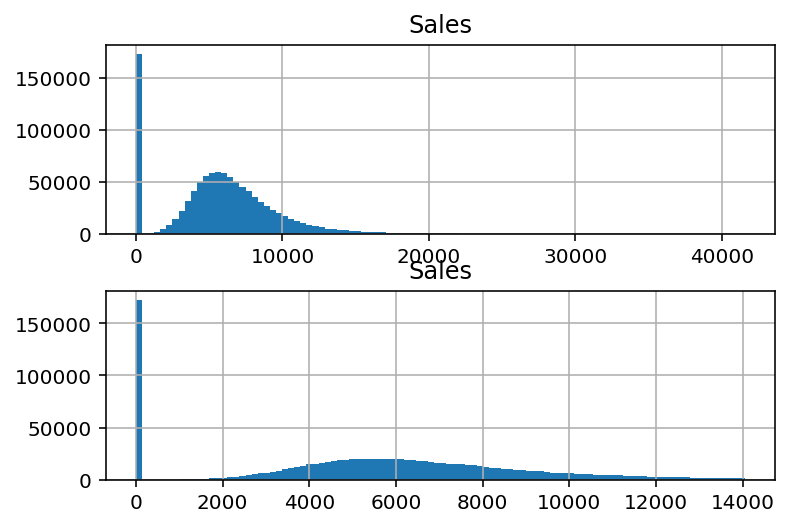


图13

3.1.2、OneHot编码

由于机器学习方法，对于数据类型的限制，需要将类比的数据进行Onehot编码才能输入到模型中去进行计算。需要将“StoreType”、“Assortment”、“Stateholiday”其中的“abcd”对应到多个字段中去。对于“ProniInterval”字段，需要将12个月分别对应到12个字段中去，得出算法可以处理的数据类型。

3.1.3、时间处理

由于销售数据是带时间序列的，其“Date”字段代表当天的时间，由于“Date”字段实际为“YYYY-MM-DD”格式，需要将其转化为“Date\_year”、“Date\_month”、“Date\_day”。观察Store为1的商店其走势情况可以看出，时间周期对其影响还是很大的，年末销量会有大曾，季度与季度之间的销量也有很大区别，因此需增加字段来表示当天在一年中处于什么阶段，于是新增两个字段“DateOfYear”、“WeekOfYear”。

3.1.4、特征组合

考虑到当前属性虽然能够进行预测，但是和排行榜的衡量标准来比较，发现差距还是很大，因此需要新增数据。

首先，是将Customer字段与Sales字段等相结合，处理成“客单价”、“客流量”等商店属性。

其次，根据对排行榜第一名的Gret的采访，他与属性的处理很有借鉴意义，于是将数据分为"有/无促销情况下的销量"、“2013/2014/2015不同年份的销量平均值”、“第一到四季度销量的平均值”等特征属性。

第三，由于对竞争对手开业时间分布在“CompetitionOpenSince[Year/Month]”两个字段，如果不进行处理就很难分析，因此将其处理成，截止这条记录发生时，竞争对手开业的总月数就可以方便分析了。同理也可以处理“Promo2Since[Year/Month]”字段。

第四、节假日前后，会进行一些打折活动，因此对于商家的销量有一定的影响。由于之前的分析看出，学校假对于销量影响并不是很大，而且学校假期普遍较长，因此这里只对“a类”法定假日进行分析，得出距离各个“a类”法定假日的距离。

第五、最近一周的Holiday数量也会对销售量有一定影响，因此算出最近一周时间内，“学校假”和“法定假”各自的数量。

3.1.5、缺失值处理

对于缺失值，处理方式有很多，直接删除是一种方法，均值插入也是一种方式。这里我的处理方式是：不作处理。因为XGBoost有一套对于缺失值的处理方式：

训练集中的缺失值处理方法：分别分配到左右子树，选取损失函数最优的解决方案；

测试集中的缺失值处理方法：默认分配到右子树。

因此，我们可以放XGBoost去处理缺失的数据。

## 3.2、执行过程

## 3.3、完善

# 四、结果

## 4.1、模型的评价与验证

## 4.2、合理性分析

# 五、项目结论

## 5.1、结果可视化

## 5.2、对项目的思考

## 5.3、需要做出的改进

1. #### 张慰. 德国周日与节日法律保护之评述. 德国研究. 2017

   [↑](#footnote-ref-1)
2. **https://www.kaggle.com/c/rossmann-store-sales** [↑](#footnote-ref-2)
3. http://blog.kaggle.com/2015/12/21/rossmann-store-sales-winners-interview-1st-place-gert/ [↑](#footnote-ref-3)
4. https://www.kaggle.com/c/rossmann-store-sales/discussion/18024 [↑](#footnote-ref-4)
5. https://blog.csdn.net/jasonzhangoo/article/details/73061060 [↑](#footnote-ref-5)
6. https://www.kaggle.com/c/rossmann-store-sales/leaderboard [↑](#footnote-ref-6)