**机器学习纳米学位**

##毕业项目 李榕 优达学城

2019年9月11日

**I. 问题的定义**

**项目概述**

本项目旨在为零售商Rossmann（劳诗曼）解决其销售额预测的问题，即利用过去已经发生的店铺的销量信息、店铺的类别、打折、 对手相关信息，对店铺近期未来的数据进行准确预测。

准确预测出销售额分析影响因素，对于零售商整体店铺布局，门店物品准备，改变打折策略以达到提升最终销售额的目的起着至关重要的作用。

为解决此问题，分析现有的一些数据，与每个店铺相关的信息包括：

|  |  |
| --- | --- |
| **列名** | **列含义** |
| store | 商店ID |
| storetype | 商店类型 |
| Assortment | 商店分类 |
| CompetitionDistance | 竞争者的距离 |
| CompetitionOpenSinceMonth | 竞争者开启的月份 |
| CompetitionOpenSinceYear | 竞争者开启的年份 |
| Promo2 | 是否有广告投放 |
| Promo2SinceWeek | 广告从那一周开始投放 |
| Promo2SinceYear | 广告从哪一年开始投放 |
| PromoInterval | 广告投放月份周期 |

每天生成的销量数据包括如下信息：

|  |  |
| --- | --- |
| **列名** | **列含义** |
| store | 商店ID |
| dayofweek | 周几 |
| date | 日期 |
| sales | 销量 |
| customer | 顾客量 |
| open | 店铺是否开门 |
| promo | 是否有促销 |
| StateHoliday | 日期状态 |
| SchoolHoliday | 是否是假日 |

本项目将根据如上已经发生的历史信息，来预测未来近期，每日每个店铺的销量（sales）的数据

* *需要解决的问题涉及哪个领域？做这个项目的出发点？有哪些相关的数据集或输入数据？*
* *问题的背景信息能够让完全没接触过这个问题的人充分了解这个问题吗？*

**问题陈述**

本课课程需要解决依据Rossmann各个商店13年2月到15年7月之间的销售额数据，结合商店类型、商店的竞争者位置、是否是节假日、当天是星期几等综合因素，来对Rossmann15年8月到15年9月之间的销售额进行预测。据此，问题可以被拆解为如下几点：

1、店铺的销售额与店铺的品类是不是有关系，有怎样的影响？

2、店铺周围是否有竞争者，对销售额是否有影响，有怎样的影响？

3、店铺竞争者的位置远近，对销售额是否有影响，有怎样的影响？

4、当天的日期是正常工作日还是星期日，对销售额是否有影响，有怎样影响

5、当天是否是公共假期，对销售额是否有影响，有怎样影响

6、当天日期是否是学校假日，对销售额是否有影响，有怎样影响

7、当天是是否发生促销，对销售额是否有影响

8、当天是否发生了季节性促销，对销售额的影响怎样

9、明确了对销售额的影响因素之后，进行特征工程的构建，对空值进行删除补充

10、考虑采用回归模型来进行不同店铺销量的预测。

11、评估预测的效果如何

期望得到结果：

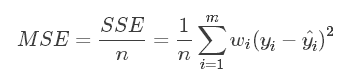
通过对train.csv数据进行训练，得到合适模型，模型好坏通过对预测值与真实值之间的差距来衡量；将模型应用于测试数据中（test.csv文件）得到对应商店每一日的销量数据

* *你是否清楚地定义了这个问题。站在读者的角度，他们能否明白你将要解决的问题是什么。*
* *你是否详尽地阐述了你将会如何解决这个问题？*
* *你期望什么样的结果，读者能明白你期望的这个结果吗？*

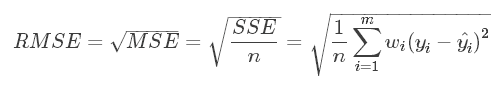
**评价指标**

本文处理的问题为回归问题，回归问题通过分析预测值与真实值之间的差距，来评估模型的好坏，常用的评价标准：

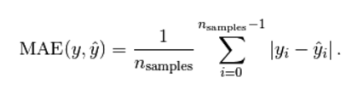
1）mse：均方误差——越接近于0越好



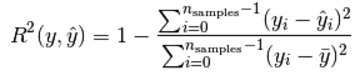
2）rmse：均方根误差——越接近于0越好



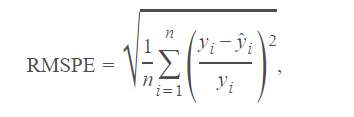
3）mae：平均绝对误差——越接近于0越好



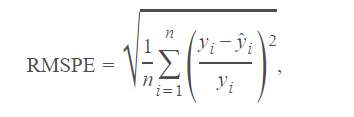
4）R2：决定系数——越接近1越好



**5）RMSPE——均方根百分比误差，为项目要求使用的评估方法。**



了解常用方法之后，本方案选择：RMSPE——均方根误差百分比。



yi 代表某店铺某天真实销量

i 代表对应店铺对应日期的预测销量

n代表总的记录数

指标整体衡量了预测值与真实值偏差的百分比均值

本项目需要得到的最终评估标准是以RMSPE来衡量，因而指标评估方案选择为RMSPE。

为啥选择RMSPE，而不是其他的

* *你是否清晰地定义了你所使用的指标和计算方法？*
* *你是否论述了这些指标和计算方法的合理性？*

**II. 分析**

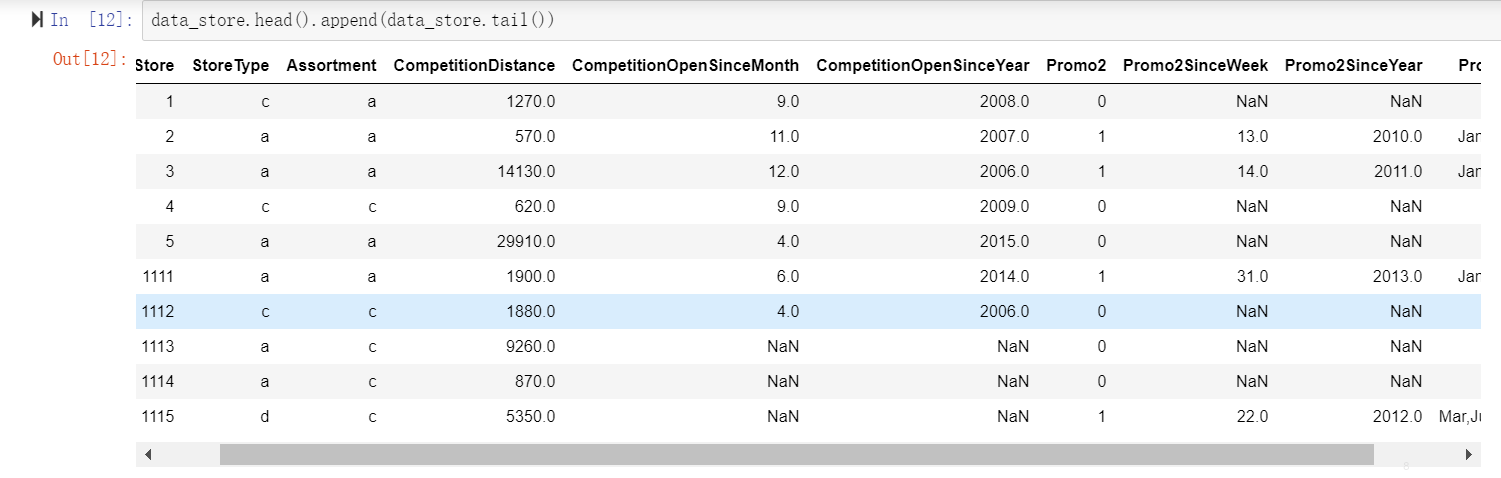
**数据的探索**

1 项目给定表内容分析

本项目共包含三个表文件：

1. store.csv 文件内容为每个店铺的说明情况：

数据情况如下：



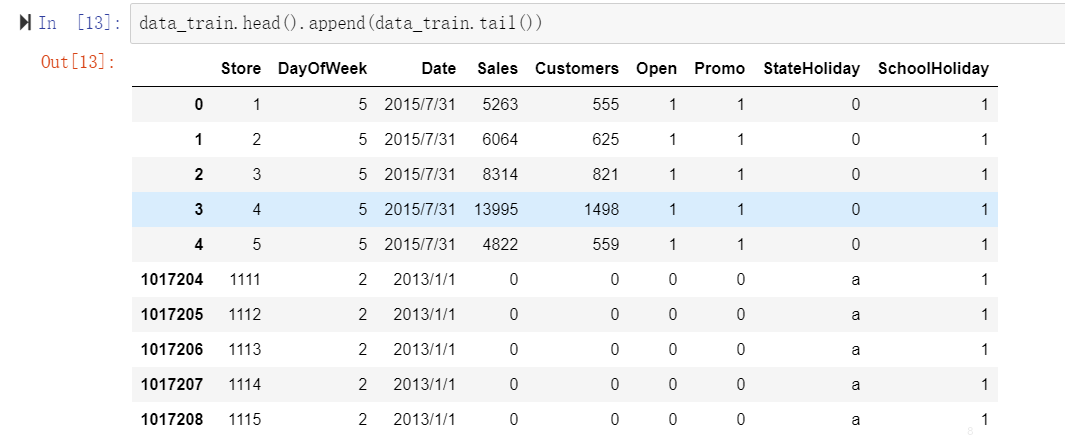
行情况：共1115行，代表1115个店铺，主键为Store

列情况：共10列，分析代表含义如下：

|  |  |
| --- | --- |
| **列名** | **列含义** |
| store | 商店ID |
| storetype | 商店类型 |
| Assortment | 商店分类 |
| CompetitionDistance | 竞争者的距离 |
| CompetitionOpenSinceMonth | 竞争者开启的月份 |
| CompetitionOpenSinceYear | 竞争者开启的年份 |
| Promo2 | 是否有广告投放 |
| Promo2SinceWeek | 广告从那一周开始投放 |
| Promo2SinceYear | 广告从哪一年开始投放 |
| PromoInterval | 广告投放月份周期 |

1. train.csv 文件内容为用于模型训练的数据

数据情况如下：

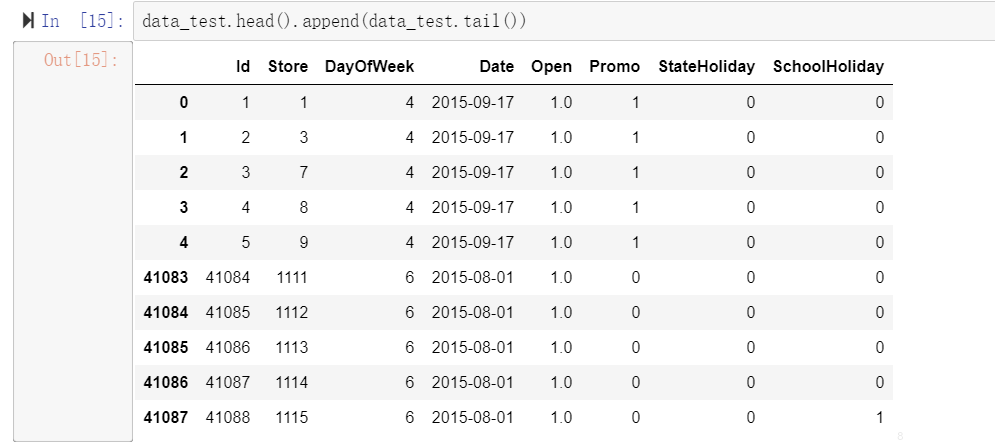


行情况：共计1017209行，代表从13年1月1日到15年7月31日的1115家店铺，每日的数据情况

列情况：共计列，分析各列代表含义如下：

|  |  |
| --- | --- |
| **列名** | **列含义** |
| store | 商店ID |
| dayofweek | 周几 |
| date | 日期 |
| sales | 销量 |
| customer | 顾客量 |
| open | 店铺是否开门 |
| promo | 是否有促销 |
| StateHoliday | 日期状态 |
| SchoolHoliday | 是否是假日 |

1. test.csv 文件内容为用于预测的结果数据



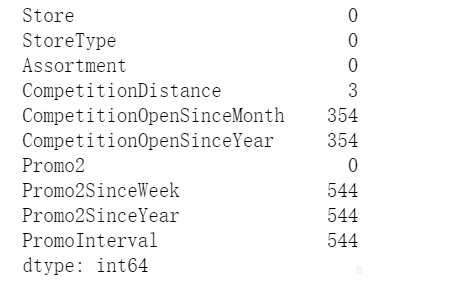
行情况：共计41088行，代表从15年8月1日到15年9月17日的店铺ID、促销信息、节假日信息等情况

列情况：

|  |  |
| --- | --- |
| **列名** | **列含义** |
| store | 商店ID |
| dayofweek | 周几 |
| date | 日期 |
| open | 店铺是否开门 |
| promo | 是否有促销 |
| StateHoliday | 日期状态 |
| SchoolHoliday | 是否是假日 |

2 表内容缺省值分析

1）store.csv缺省值分析



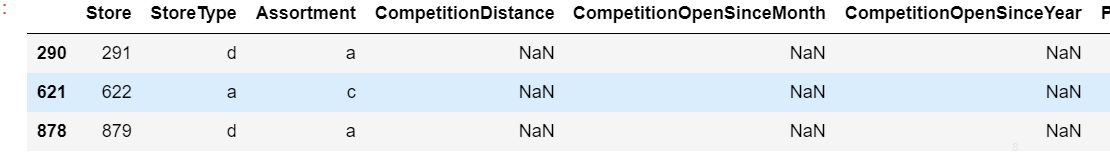
与竞争者相关的字段：

CompetitionDistance

CompetitionOpenSinceMonth

CompetitionOpenSinceYear

字段有空值，进一步分析

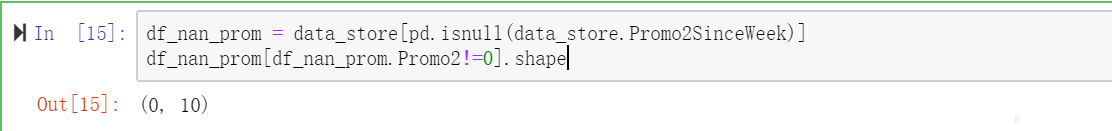


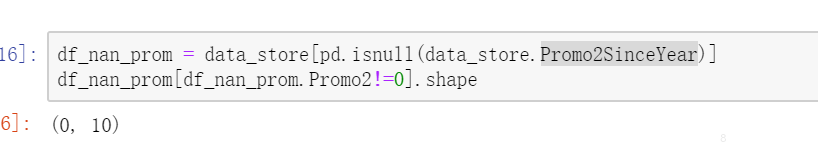
CompetitionDistance字段为空时，对应的CompetitionOpenSinceMonth，CompetitionOpenSinceYear字段同时为空，认定为店铺可能不存在竞争者，这里考虑把对应的值写为0

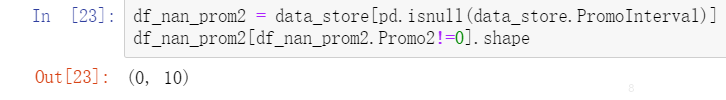
进一步观察CompetitionOpenSinceMonth，CompetitionOpenSinceYear，两者为空的列是一致的，认定为竞争者开启的年份和月份是未知的，考虑将对应的两列写为0

与促销相关字段

Promo2SinceWeek，Promo2SinceYear，Promo2Interval字段，均是当无促销时出现该情况，因而认为当日无促销

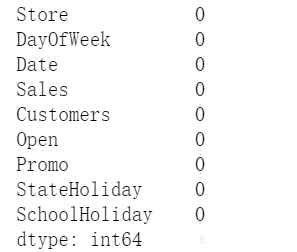






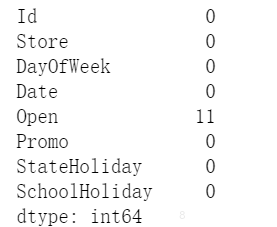
处理时，将对应字段填为0

2）train.csv缺省值分析

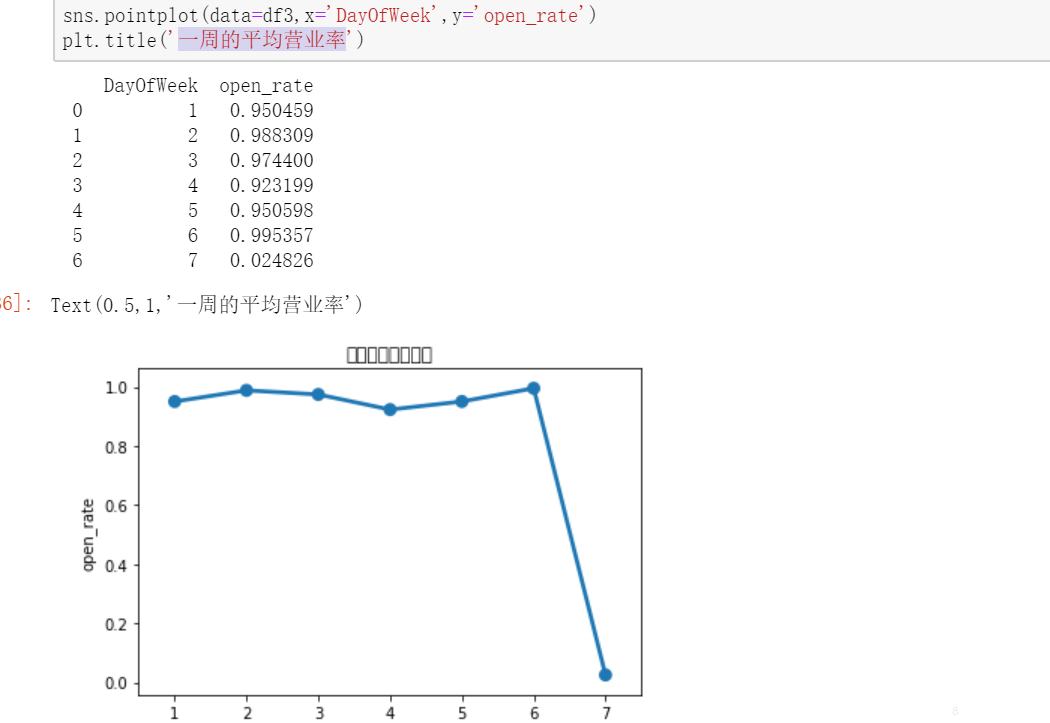


本文件中无缺省值

3）test.csv缺省值分析

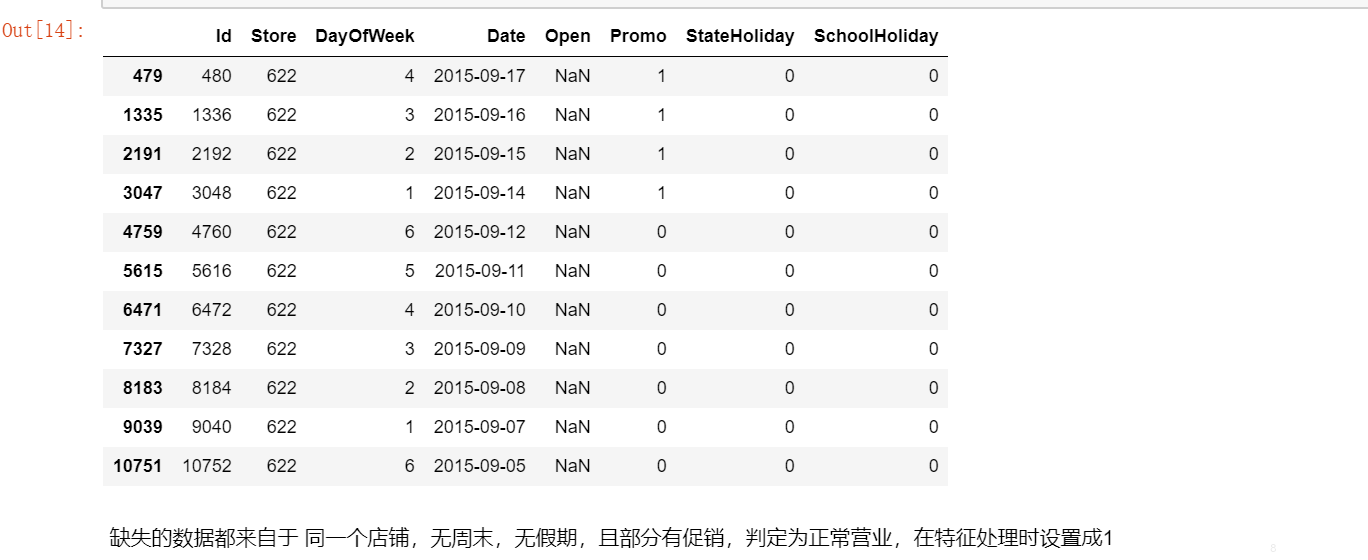


缺省值为open字段，代表商店是否开着。通过常识分析，商店是否开，可能与周几，节假日关系较大，进一步对train.csv文件中的DayofWeek字段的Open分布进行分析，得到下图：



即 星期为周日时，仅2%的店铺是营业状态。

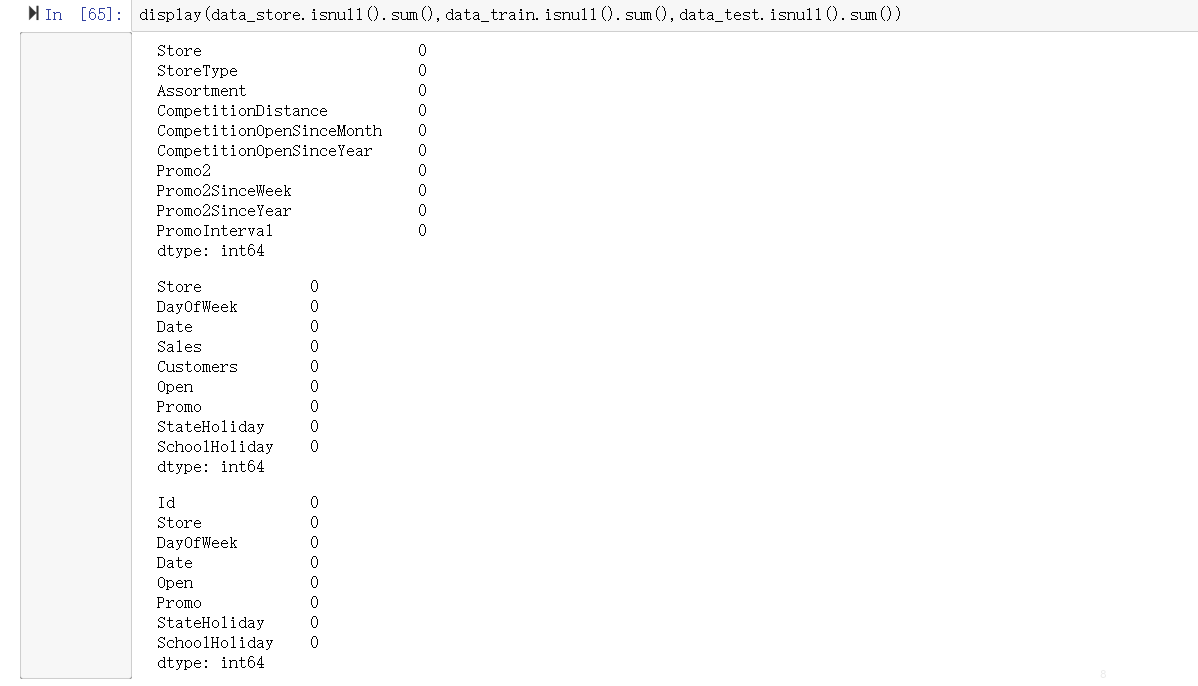
观察test.csv 中open字段缺失所对应的Dayofweek列



未出现周末的情况，因而直接将test.csv 文件中的Open为空的字段写为1

综上，对几个文件的缺省值进行填充，并判定，填充之后的数值是否仍有空值





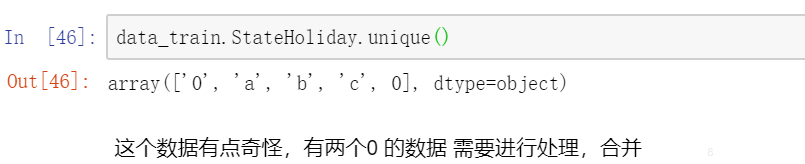
3 表内容异常值分析

1）store.csv

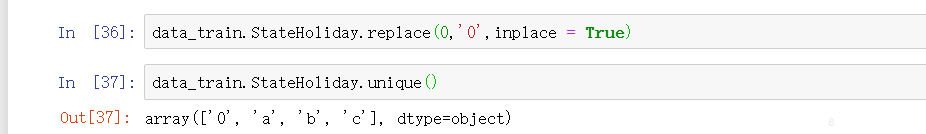
文件中，查看无明显异常值的情况，但是，StoreType, Assortment, PromoInterval为非数值表示，需要在特征处理时，进行数值转化

2）train.csv

明显异常集中在 StateHoliday字段，出现了一个数值0，和一个字符0，需要进行统一的数值化处理



这里将StateHoliday字段中 数值0统一转换为字符’0’



3）test.csv

未发现明显的数值异常情况，

4 数据集整体概览

如前所述，train.csv集合中，共计1017209行，代表从13年1月1日到15年7月31日的1115家店铺，每日的数据情况。

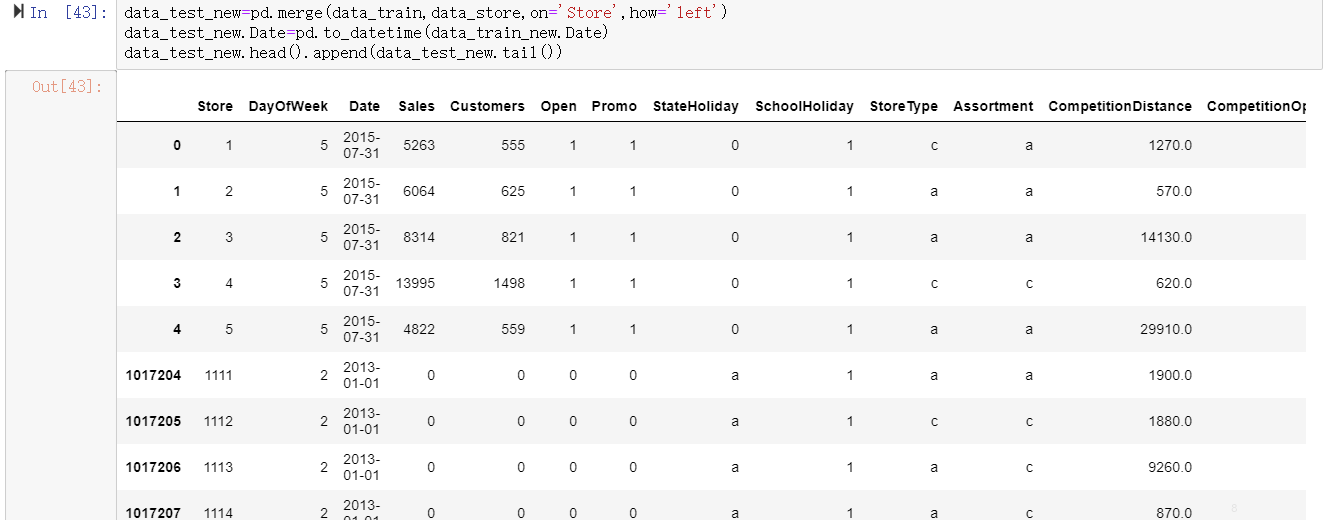
Test.csv共计41088行，代表从15年8月1日到15年9月17日的店铺ID、促销信息、节假日信息等情况

* *如果你使用了数据集，你要详尽地讨论了你所使用数据集的某些特征，并且为阅读者呈现一个直观的样本*
* *如果你使用了数据集，你要计算并描述了它们的统计量，并对其中与你问题相关的地方进行讨论*
* *如果你****没有****使用数据集，你需要对你所使用的输入空间（input space)或输入数据进行讨论？*
* *数据集或输入中存在的异常，缺陷或其他特性是否得到了处理？(例如分类变数，缺失数据，离群值等）*

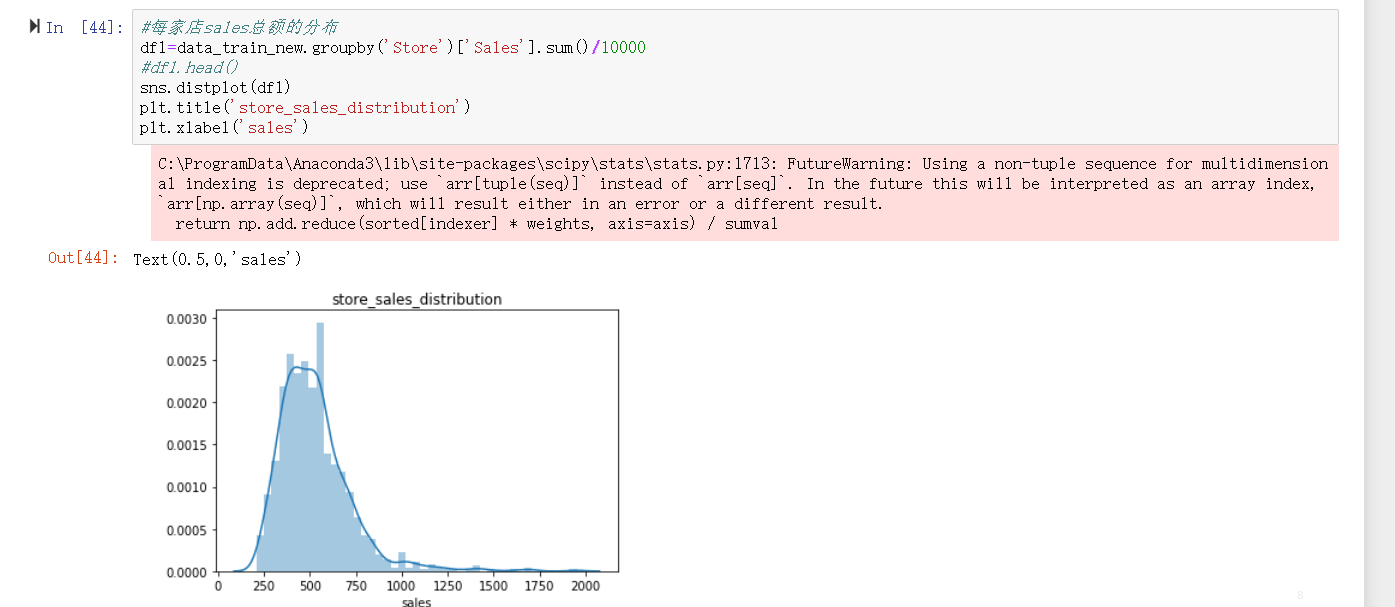
**探索性可视化**

分析的目的是需要将各类的变量对销量的数据产生的影响进行分析，因而分别将train.csv ,test.csv 与store.csv 进行连接，得到新的数据集方便进行下一步分析。



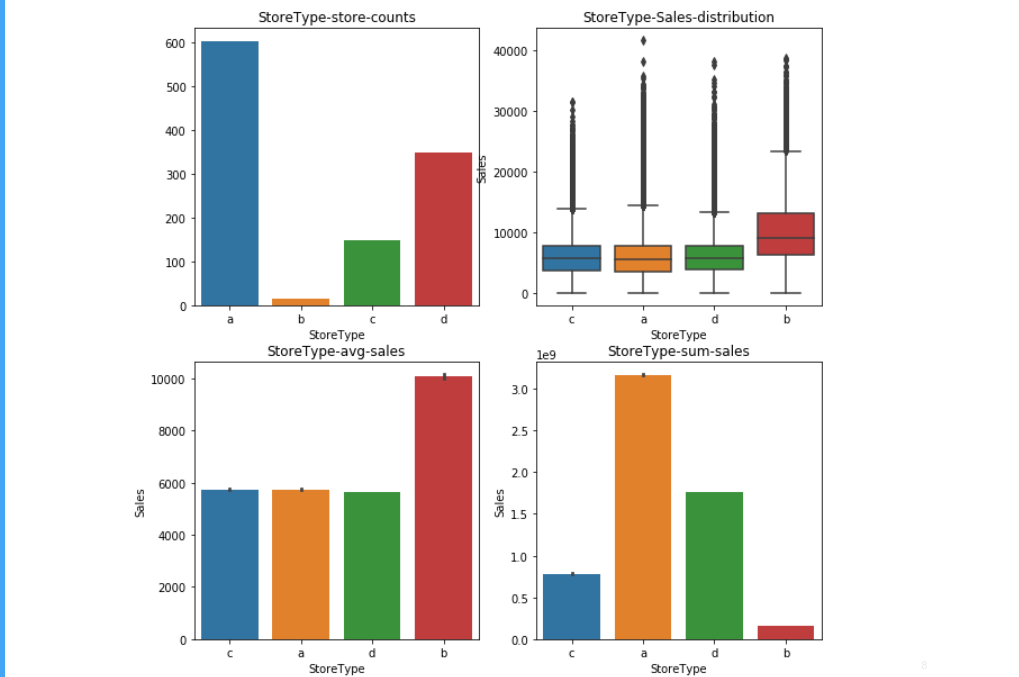


1. 店铺整体销量的频率分布情况：

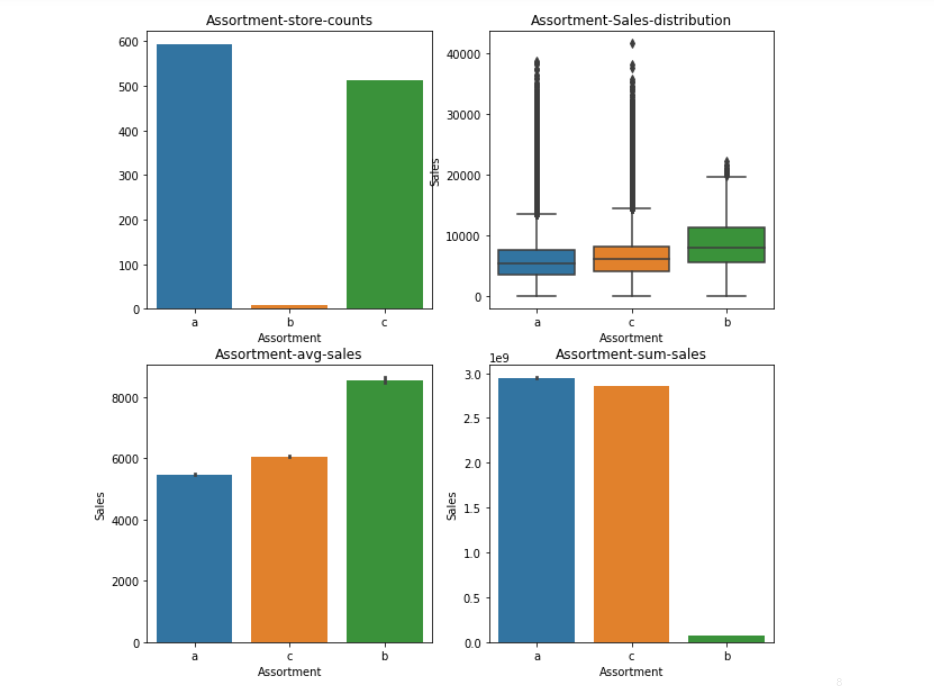


绝大部分的店铺销量销量低于1000W，销量在1000W以上的店铺很少。

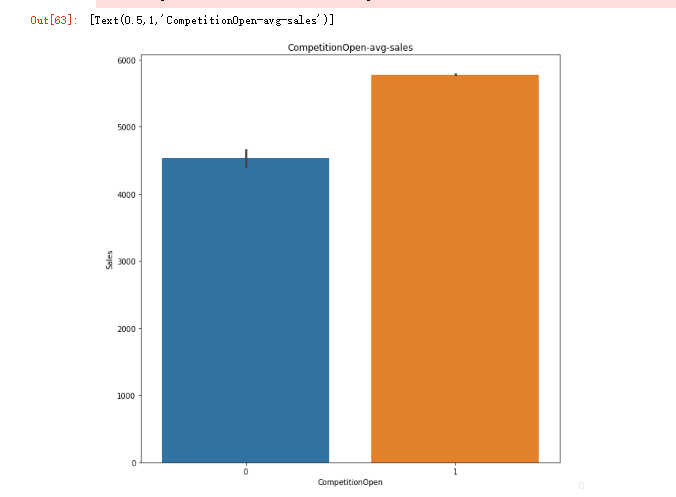
1. 店铺的不同类型以及对应类型的销量情况



1. 从店铺数量的分布来看，a店铺的数量最多，其次为d类，b类店铺的数量最少
2. 从销量分布来看，a,c,d店铺的销量分布情况相对接近，b类店铺的销量整体分布偏高
3. 从日均销量对比来看，b类店铺的日均销量最高，a,b,c 相对接近
4. 从总销量来对比，与店铺的数量分布类似，a店铺带来的销量最多，其次为d，b销量最少
5. 不同分类级别（Assortment）情况

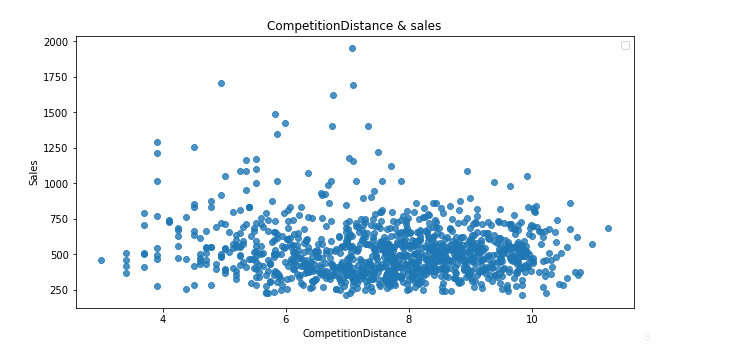


1. 店铺数量对比来看，a类店铺最多，c类次之，b类店铺最少，即基本店铺和扩展型店铺较多，额外店铺少
2. 从销量分布来看，b类店铺的销量分布整体比a，c更高
3. 从平均销量来看，b类店铺平均销量最高
4. 店铺的总销量与店铺的总量规律一致
5. 与竞争对手的关系
6. 是否有竞争对手



有竞争对手平均销量会稍高些

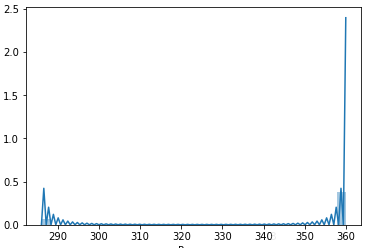
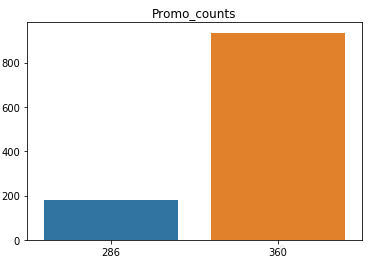
1. 竞争对手距离与平均销量的关系



竞争对手距离变大或者变小，对销量的影响并不大

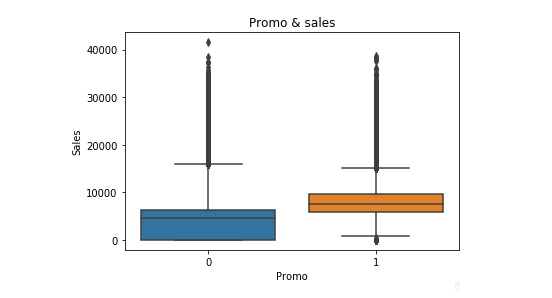
1. 与促销的关系

1）促销情况的频率分布

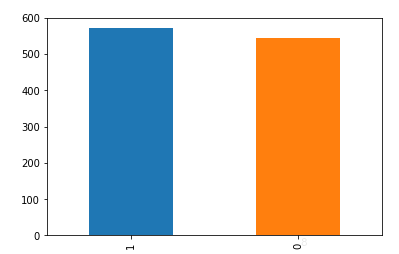
整体促销次数的频率分布图如上。

1. 促销对销量的影响



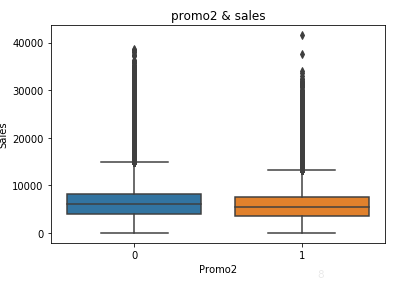
促销之后整体的销量要好于不促销的情况

1. 季节性促销的次数



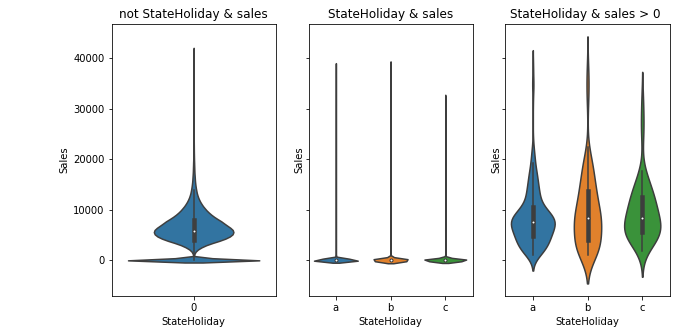
季节性促销与不促销次数基本接近

1. 季节性促销与销量的影响



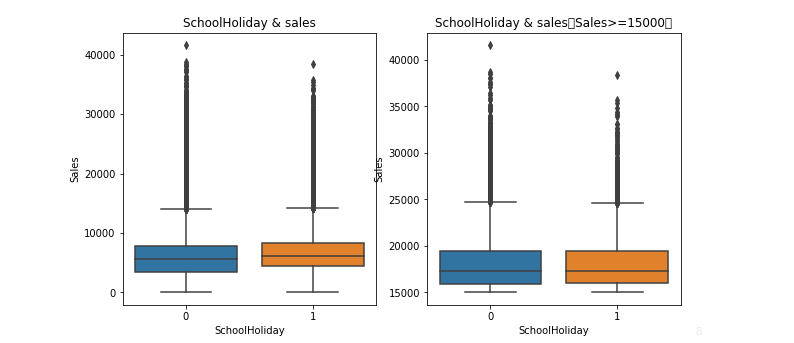
季节性促销与否与销量的影响关系不大

1. 探索法定假日、周末、学校假日对销量的影响
2. 法定假日的影响：



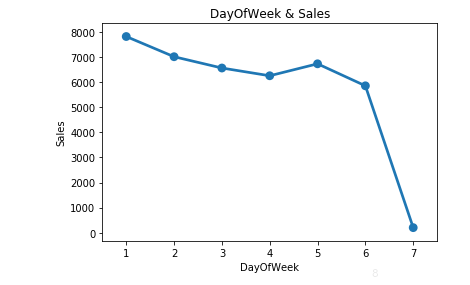
法定假日时，销量很低，说明大部分商店不营业

1. 学校假日对销量的影响

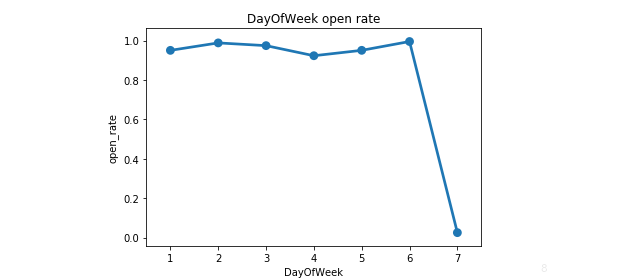


学校的假日基本，整体会比无假日稍高，但高销量数据中差异不明显，说明学校的假日影响到的群体仅有学生这一小部分群体，且学生的消费水平并不高。

1. 星期与销量的关系

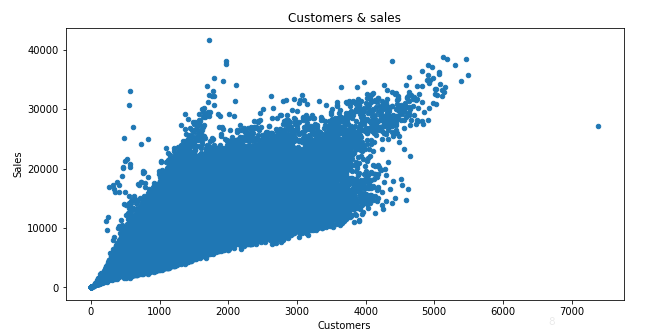


从周一到周日，日均销量逐渐下滑，周日比周一到周六明显降低。



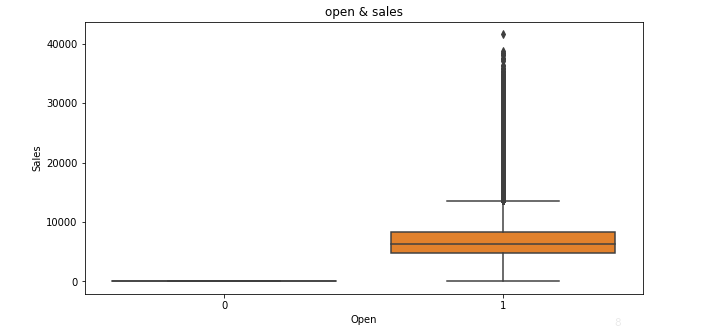
在看店铺的打开率，周日98%店铺不开门，无销量

1. 顾客量与销量的关系



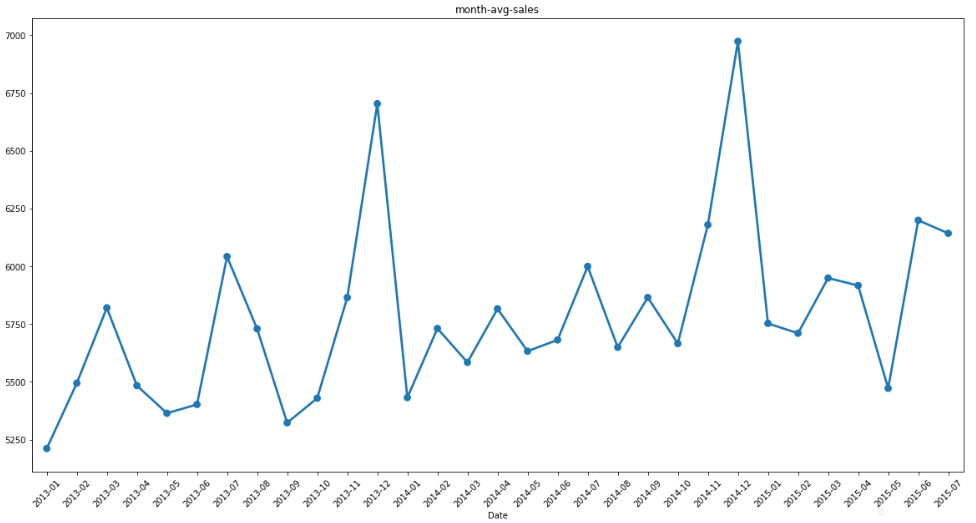
顾客量与销量呈现明显得相关性，不难理解，顾客量多自然买东西的人多，但由于在测试的数据中没有顾客量，因而该数据无法作为特征，输入模型训练

1. 店铺是否开启与销量关系



与预想的结果进行相互的验证，确实如果店铺不开那么就是没有销量，符合我们的尝试。

1. 探索销量随时间的变化规律
2. 每个月份的平均销量



每年的年底是高峰，之后回落

* *你是否对数据中与问题有关的特性进行了可视化？*
* *你对可视化结果进行详尽的分析和讨论了吗？*
* *绘图的坐标轴，标题，基准面是不是清晰定义了？*

**算法和技术**

本问题是一个回归问题，因而考虑采用回归模型来求解：

1）**线性回归，**通过线性回归构造出来的函数一般称之为了线性回归模型。

C:\Users\Administrator\AppData\Local\YNote\data\weixinobU7Vjomn6r2tiNT6zdH5ZvbZEhI\628d8d90f2584e349b8fc70dd3257a8a\p=49&f=jpeg.jpeg

2）**回归树**，通过构建决策树来获取到区间均价，分裂点方法与分裂问题不同，分类问题采用的是，信息增益或信息增益率 基尼系数等方法；回归树采用最小化误差平方的方式，从特征变量中，找到一个变量j，找到一个取值，这个取值能够使得按照取值划分之后的 两个集合，对于目标变量的误差平方最小；接着依次遍历所有的变量，得到特征空间的划分。然后依次进行划分，最终得到子向量空间，向量空间的均值就是预测值。

3）**随机森林**

　　1. 假如有N个样本，则有放回的随机选择N个样本(每次随机选择一个样本，然后返回继续选择)。这选择好了的N个样本用来训练一个决策树，作为决策树根节点处的样本。

　　2. 当每个样本有M个属性时，在决策树的每个节点需要分裂时，随机从这M个属性中选取出m个属性，满足条件m << M。然后从这m个属性中采用某种策略（比如说信息增益）来选择1个属性作为该节点的分裂属性。

　　3. 决策树形成过程中每个节点都要按照步骤2来分裂（很容易理解，如果下一次该节点选出来的那一个属性是刚刚其父节点分裂时用过的属性，则该节点已经达到了叶子节点，无须继续分裂了）。一直到不能够再分裂为止。注意整个决策树形成过程中没有进行剪枝。

4. 按照步骤1~3建立大量的决策树，这样就构成了随机森林了。

4）**xgboost**

xgboost由多个相关联的树联合。

xgboost不断对残差进行预测，xgboost每个决策树是逐一被添加进入。下一棵决策树加入的输入，会取决于前一棵决策树产生的残差。

单个树生成方式：

4.1 不断便利特征和特征的取值，然后进行分类，计算loss function最小值，然后再选择一个特征分裂，又得到一个损失函数的最小值，找到分裂效果最好的（即分裂前后损失函数变化最大）的特征进行分裂

4.2 持续4.1的过程，不断进行分裂

4.3 如下条件停止分裂：

a）分裂的增益小于某个阈值

b）分裂达到最大深度时，停止，最大深度理论上应当是是一个超参数

c) 样本权重之和，小于设定阈值时，停止分裂

4.4 计算前一棵树的残差数据，然后将 目标函数定义为 残差，继续输入到模型中，进行重复的模型构建，直到触发 ：增益小于阈值，树的数量小于一定量，样本的权重之和小于设定的阈值，停止整个过程。

本方案选择采用xgboost方法作为算法。

1. 代码实现的xgb相关参数说明

<https://www.cnblogs.com/Allen-rg/p/10563362.html>

<https://www.cnblogs.com/baby-lily/p/10663328.html>

xgboost.train(

params,

dtrain,

num\_boost\_round=10,

evals=(),

obj=None,

feval=None,

maximize=False,

early\_stopping\_rounds=None,

evals\_result=None,

verbose\_eval=True,

learning\_rates=None,

xgb\_model=None)

params 这是一个字典，里面包含着训练中的参数关键字和对应的值

dtrain 训练的数据

num\_boost\_round 提升迭代次数，即生成多少基模型

evals 这是一个列表，用于对训练过程中进行评估列表中的元素。形式是evals = [(dtrain,’train’),(dval,’val’)]或者是evals = [(dtrain,’train’)],对于第一种情况，它使得我们可以在训练过程中观察验证集的效果。

**obj**：自定义目的函数

**feval**：自定义评估函数，在本文中评估函数是RMSPE，需要自定义一个RMSPE的函数传入此参数中

**maximize**：是否对评估函数进行最大化

early\_stopping\_rounds：早期停止次数 ，假设为100，验证集的误差迭代到一定程度在100次内不能再继续降低，就停止迭代。这要求evals 里至少有 一个元素，如果有多个，按最后一个去执行。返回的是最后的迭代次数（不是最好的）。如果early\_stopping\_rounds存在，则模型会生成三个属性，bst.best\_score，bst.best\_iteration和bst.best\_ntree\_limit

**evals\_result**：字典，存储在watchlist中的元素的评估结果。

**verbose\_eval** ：(可以输入布尔型或数值型)，也要求evals里至少有 一个元素。如果为True,则对evals中元素的评估结果会输出在结果中；如果输入数字，假设为5，则每隔5个迭代输出一次。

**learning\_rates**：每一次提升的学习率的列表，

**xgb\_model**：在训练之前用于加载的xgb model。

模型参数：

params = {"objective": "reg:linear",

"booster" : "gbtree",

"eta": 0.03,

"max\_depth": 10,

"subsample": 0.9,

"colsample\_bytree": 0.7,

"silent": 1,

"seed": 10

}

* *你所使用的算法，包括用到的变量/参数都清晰地说明了吗？*
* *你是否已经详尽地描述并讨论了使用这些技术的合理性？*
* *你是否清晰地描述了这些算法和技术具体会如何处理这些数据？*

**基准模型**

在这一部分，你需要提供一个可以用于衡量解决方案性能的基准结果/阈值。这个基准模型要能够和你的解决方案的性能进行比较。你也应该讨论你为什么使用这个基准模型。一些需要考虑的问题：

* *你是否提供了作为基准的结果或数值，它们能够衡量模型的性能吗？*
* *该基准是如何得到的（是靠数据还是假设）？*

**III. 方法**

*(大概 3-5 页）*

**数据预处理**

在这一部分， 你需要清晰记录你所有必要的数据预处理步骤。在前一个部分所描述的数据的异常或特性在这一部分需要被更正和处理。需要考虑的问题有：

* *如果你选择的算法需要进行特征选取或特征变换，你对此进行记录和描述了吗？*
* ***数据的探索****这一部分中提及的异常和特性是否被更正了，对此进行记录和描述了吗？*
* *如果你认为不需要进行预处理，你解释个中原因了吗？*

**执行过程**

在这一部分， 你需要描述你所建立的模型在给定数据上执行过程。模型的执行过程，以及过程中遇到的困难的描述应该清晰明了地记录和描述。需要考虑的问题：

* *你所用到的算法和技术执行的方式是否清晰记录了？*
* *在运用上面所提及的技术及指标的执行过程中是否遇到了困难，是否需要作出改动来得到想要的结果？*
* *是否有需要记录解释的代码片段(例如复杂的函数）？*

**完善**

在这一部分，你需要描述你对原有的算法和技术完善的过程。例如调整模型的参数以达到更好的结果的过程应该有所记录。你需要记录最初和最终的模型，以及过程中有代表性意义的结果。你需要考虑的问题：

* *初始结果是否清晰记录了？*
* *完善的过程是否清晰记录了，其中使用了什么技术？*
* *完善过程中的结果以及最终结果是否清晰记录了？*

**IV. 结果**

*（大概 2-3 页）*

**模型的评价与验证**

在这一部分，你需要对你得出的最终模型的各种技术质量进行详尽的评价。最终模型是怎么得出来的，为什么它会被选为最佳需要清晰地描述。你也需要对模型和结果可靠性作出验证分析，譬如对输入数据或环境的一些操控是否会对结果产生影响（敏感性分析sensitivity analysis）。一些需要考虑的问题：

* *最终的模型是否合理，跟期待的结果是否一致？最后的各种参数是否合理？*
* *模型是否对于这个问题是否足够稳健可靠？训练数据或输入的一些微小的改变是否会极大影响结果？（鲁棒性）*
* *这个模型得出的结果是否可信？*

**合理性分析**

在这个部分，你需要利用一些统计分析，把你的最终模型得到的结果与你的前面设定的基准模型进行对比。你也分析你的最终模型和结果是否确确实实解决了你在这个项目里设定的问题。你需要考虑：

* *最终结果对比你的基准模型表现得更好还是有所逊色？*
* *你是否详尽地分析和讨论了最终结果？*
* *最终结果是不是确确实实解决了问题？*

**V. 项目结论**

*（大概 1-2 页）*

**结果可视化**

在这一部分，你需要用可视化的方式展示项目中需要强调的重要技术特性。至于什么形式，你可以自由把握，但需要表达出一个关于这个项目重要的结论和特点，并对此作出讨论。一些需要考虑的：

* *你是否对一个与问题，数据集，输入数据，或结果相关的，重要的技术特性进行了可视化？*
* *可视化结果是否详尽的分析讨论了？*
* *绘图的坐标轴，标题，基准面是不是清晰定义了？*

**对项目的思考**

在这一部分，你需要从头到尾总结一下整个问题的解决方案，讨论其中你认为有趣或困难的地方。从整体来反思一下整个项目，确保自己对整个流程是明确掌握的。需要考虑：

* *你是否详尽总结了项目的整个流程？*
* *项目里有哪些比较有意思的地方？*
* *项目里有哪些比较困难的地方？*
* *最终模型和结果是否符合你对这个问题的期望？它可以在通用的场景下解决这些类型的问题吗？*

**需要作出的改进**

在这一部分，你需要讨论你可以怎么样去完善你执行流程中的某一方面。例如考虑一下你的操作的方法是否可以进一步推广，泛化，有没有需要作出变更的地方。你并不需要确实作出这些改进，不过你应能够讨论这些改进可能对结果的影响，并与现有结果进行比较。一些需要考虑的问题：

* *是否可以有算法和技术层面的进一步的完善？*
* *是否有一些你了解到，但是你还没能够实践的算法和技术？*
* *如果将你最终模型作为新的基准，你认为还能有更好的解决方案吗？*

\*\* 在提交之前， 问一下自己... \*\*

* 你所写的项目报告结构对比于这个模板而言足够清晰了没有？
* 每一个部分（尤其**分析**和**方法**）是否清晰，简洁，明了？有没有存在歧义的术语和用语需要进一步说明的？
* 你的目标读者是不是能够明白你的分析，方法和结果？
* 报告里面是否有语法错误或拼写错误？
* 报告里提到的一些外部资料及来源是不是都正确引述或引用了？
* 代码可读性是否良好？必要的注释是否加上了？
* 代码是否可以顺利运行并重现跟报告相似的结果？