**1.问题定义**

### 1.1项目概述

### 该项目通过对于德国Rossmann超过3000家药妆店之后六周的日销售数据进行预测。

### 1.2问题陈述

该问题是一个基于时间序列的预测问题。需要通过数据的清理，特征的筛选，使用机器学习模型针对三年中1000个药妆店的超过一百万条数据的日销售数据，对于未来6周的销售数据进行预测。

由于计算资源的有限性，无法在预测阶段使用Cross validation 和 GridSearch, 本问题的解决更加集中在数据的清理、筛选及特征的创造，而非模型参数。

### 1.3评价指标

### 通过已知的4万条用以测试的数据，以RMSPE（ROOT MEAN SQUARE PERCENTAGE ERROR）作为评价指标评定预测结果。

### 

### 本项目会先

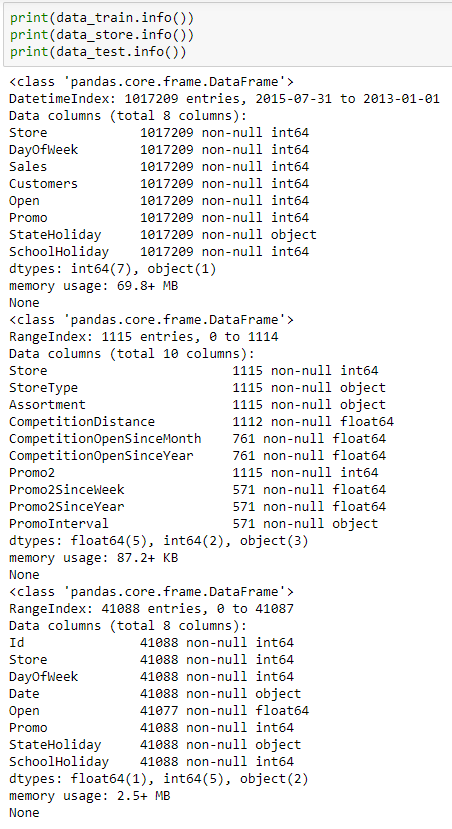
**2.分析**

(大概 2-4 页)

### 2.1数据的探索

在这一部分，你需要探索你将要使用的数据。数据可以是若干个数据集，或者输入数据/文件，甚至可以是一个设定环境。你需要详尽地描述数据的类型。如果可以的话，你需要展示数据的一些统计量和基本信息（例如输入的特征（features)，输入里与定义相关的特性，或者环境的描述）。你还要说明数据中的任何需要被关注的异常或有趣的性质（例如需要做变换的特征，离群值等等）。你需要考虑：

2.1.1数据类型，条目，及数据缺失如下：



2.1.2 StateHoliday, StoreType, Assortment, PromoInterval 为非数字，如果使用需要根据需要进行one hot encoding 或label\_encoder进行处理转化。

2.1.3Test-Open,Store-CompetitionDistance 有少量数据缺失。

CompetitionOpenSinceMonth, CompetitionOpenSinceYear,Promo2SinceWeek,Promo2SinceYear,PromoInterval 有大量数据缺失需要进行imputation。

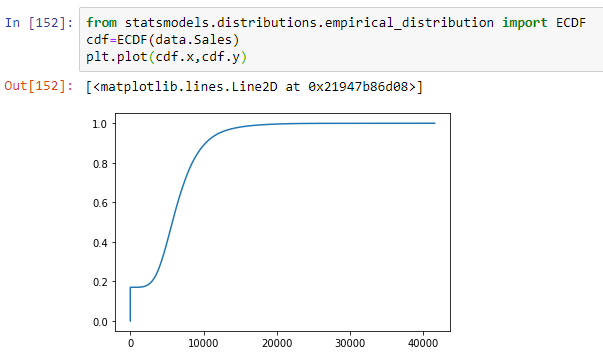
2.1.4训练集与测试集中的特征，测试集中多了ID, 而少了Customers。

### 2.2探索性可视化

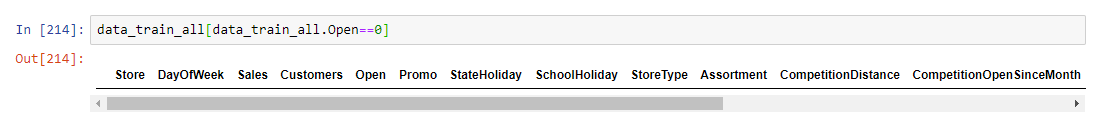
在这一部分，你需要对数据的特征或特性进行概括性或提取性的可视化。这个可视化的过程应该要适应你所使用的数据。就你为何使用这个形式的可视化，以及这个可视化过程为什么是有意义的，进行一定的讨论。你需要考虑的问题：

* 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？

2.2.1销售数据并无缺失的情况下，存在接近于20%的0数据

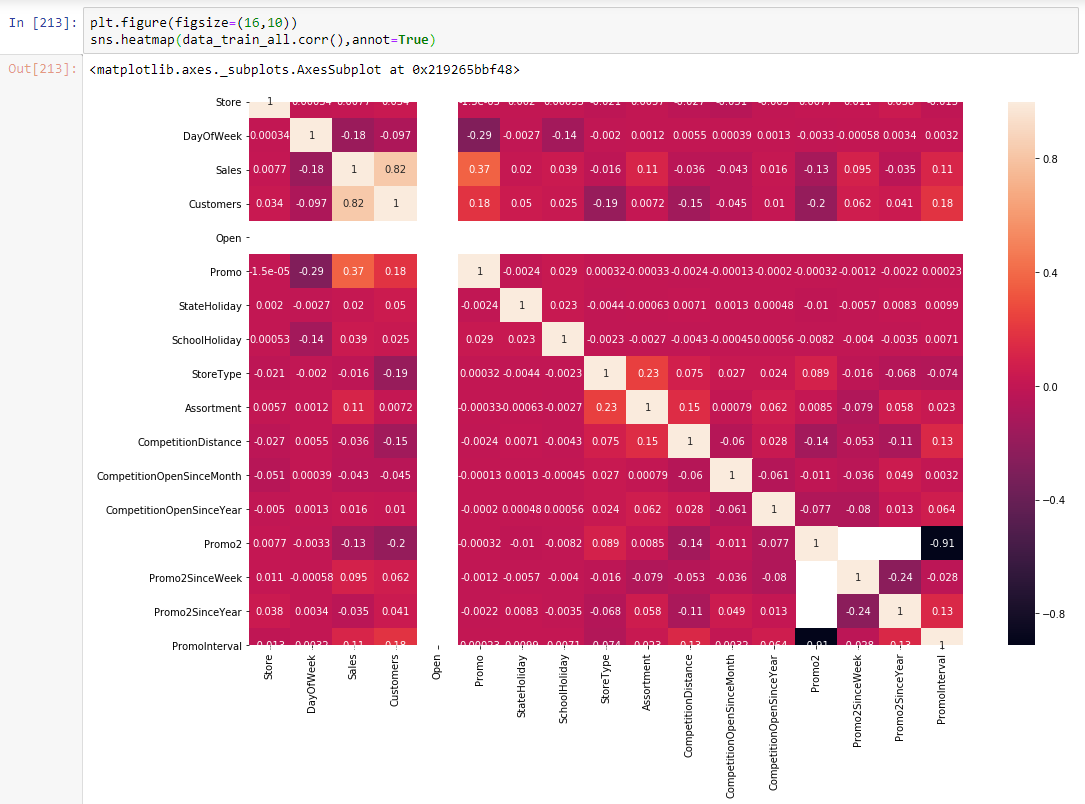


在清楚销售额为零的数据条目后，从相关性图标发现是由于店铺关门造成。通过显示删除销售额为零的数据条目后，剩下的数据条目Open 全部为非零，验证该发现。如下：



因此，Open特征在清楚销售为0的数据之后也可以删除。

销售额与客户数量，是否营业，促销，种类（Customers,Promo,SchoolHoliday,Assortment,CompetitionOpenSinceYear,Promo2SinceWeek,PromoInterval有正相关性）



### 算法和技术

在这一部分，你需要讨论你解决问题时用到的算法和技术。你需要根据问题的特性和所属领域来论述使用这些方法的合理性。你需要考虑：

* 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？

XGBoost 作为一个通用的ensemble learning的机器学习算法，适用于各种分类及预测场景。他通过将多个弱算法结合，调整若算法的权重，减少了bias,从而优化了学习正确率。他复杂度的可调节性，算法运行的并行及速度，相比于random tree， 及adaboost同样具有优势。

Eta eta [default=0.3, alias: learning\_rate]通过将其逐渐调小，及调节每一轮特征权重的节奏更精细和保守提高准确率

max\_depth [default=6] 通过逐渐增大而增加模型的复杂度以平衡和调节准确率和过拟合之间的关系。。

subsample [default=1] 选取随机数据的比例来避免过拟合。

colsample\_bytree 随机选取特征的比例来创建书。

可以通过feature importance， 对于结果进行再优化。

### 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）？

因为数据的庞大及计算资源有限，无法使用GridSearch进行参数优化， 因此将使用xgboost默认参数为基准模型，在此技术上针对某几个参数进行少量的手动调优。

### 3.方法

### 3.1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？

3.1.1对于非数字数据进行处理

### 

### 3.1.2将为零的销售数据进行删除

### 

### 3.1.3根据之前的分析，删除不相关特征

### 

### 3.1.4填充缺失数据

### CompetitionDistance 缺失数极少，不影响判断结果，用中位数填充。

### CompetitionOpenSinceMonth, CompetitionOpenSinceYear单纯数据缺失，用零代表没有进行填充。

### Promo2SinceWeek,Promo2SinceYear, PromoInterval的缺失通过相关性heatmap 看出他们完全由完全Promo2的开展与否决定。用零代表没有进行填充。

### 

### 3.1.5准备测试数据使测试集与训练集的特征相同。

### 

### 3.1.6为训练模型准备训练集

### 

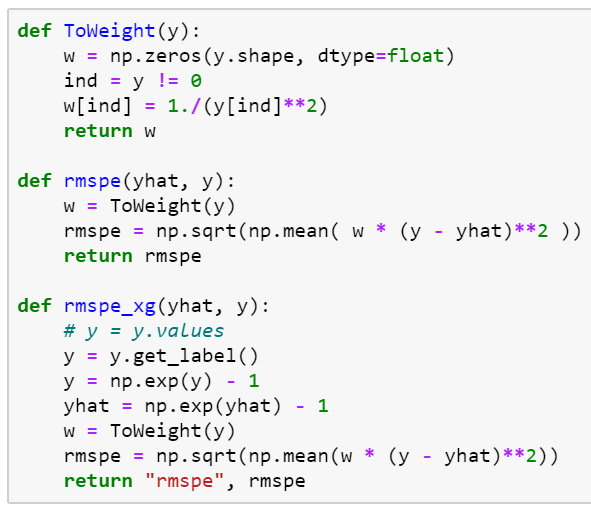
### 3.2执行过程

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）？

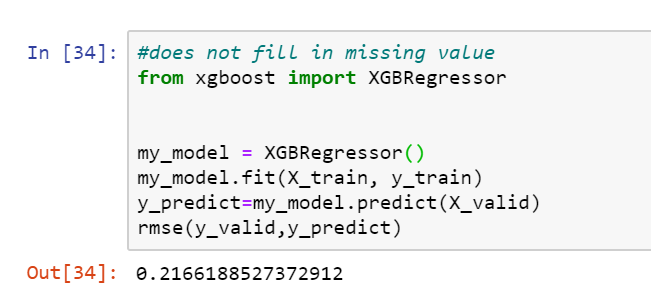
3.2.1

对于评价指标的实现

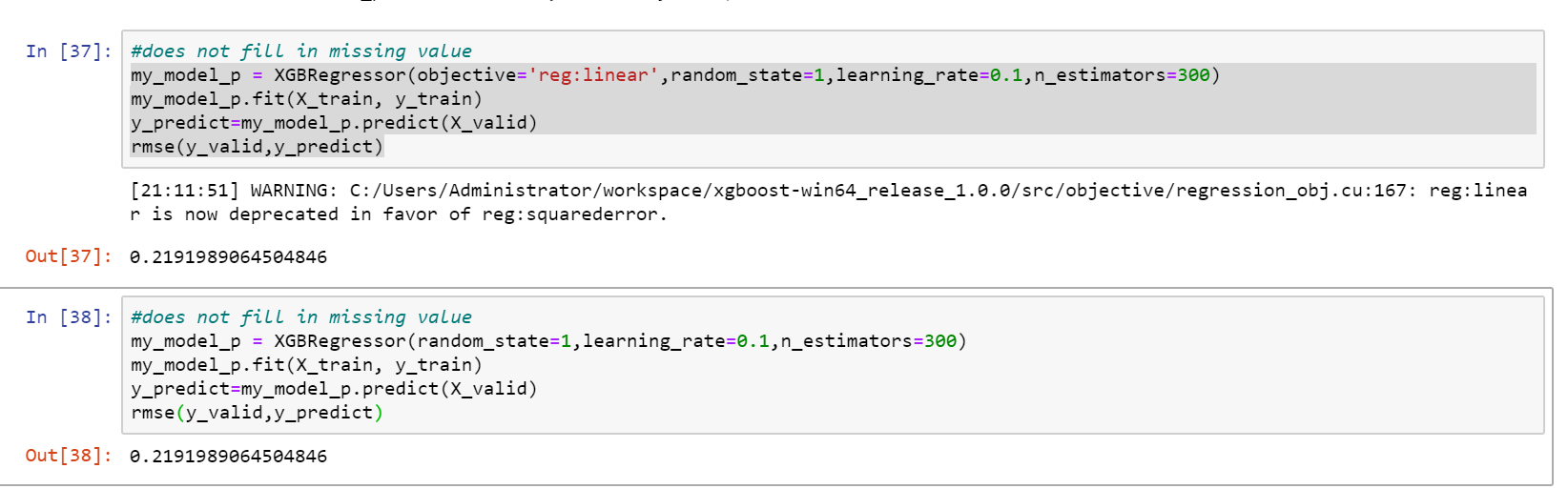








尝试使用由缺失数据的训练集训练模型，尝试XGBboost默认自动处理缺失数据的效果，同时超市比较linear和squarederror对结果的影响



### 3.3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？

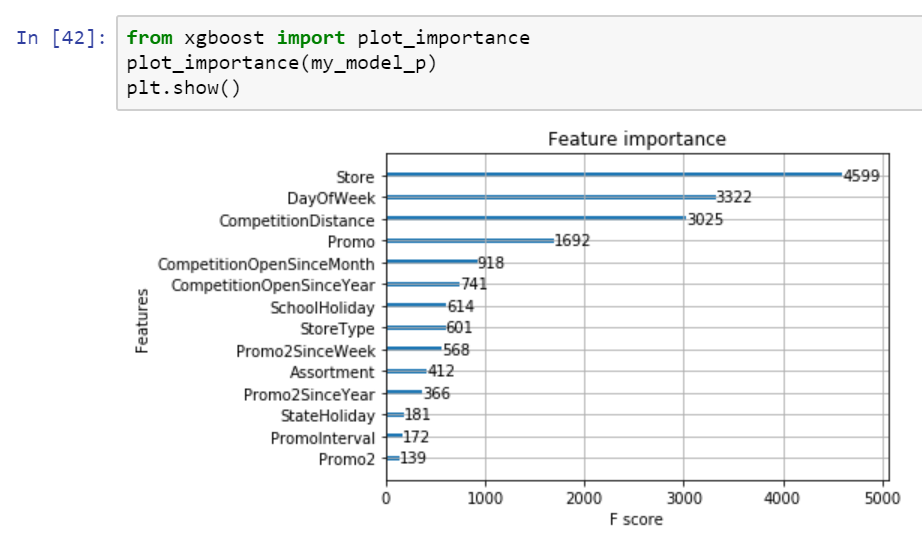
3.3.1新特征的生成及运用，

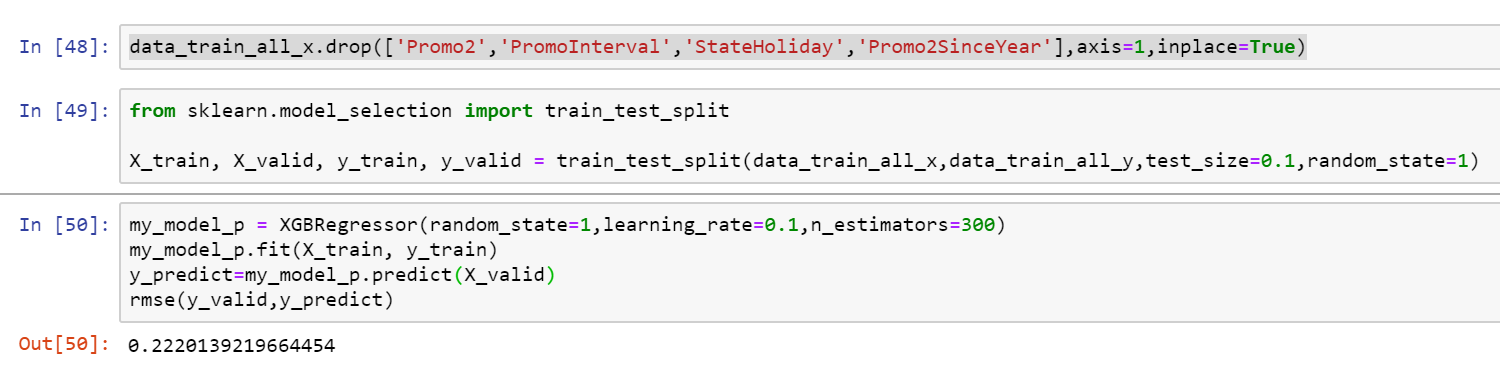
3.3.2销售数据中边缘数据的剔除，尝试通过Tukey异常值定义法，将异常阶定义为1.5倍的四分位距（interquartile range，IQR）。

* 将指定特征的 25th 分位点的值分配给 Q1 。使用 np.percentile 来完成这个功能。
* 将指定特征的 75th 分位点的值分配给 Q3 。同样的，使用 np.percentile 来完成这个功能。
* 将指定特征的异常阶的计算结果赋值给 step。
* 选择性地通过将索引添加到 outliers 列表中，以移除异常值。

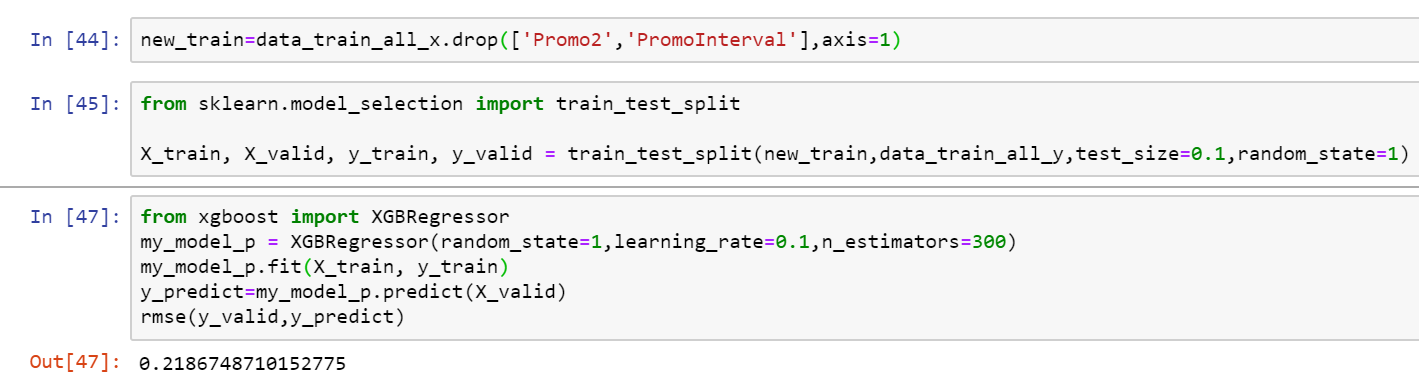
log\_data[~((log\_data[feature] >= Q1 - step) & (log\_data[feature] <= Q3 + step))])

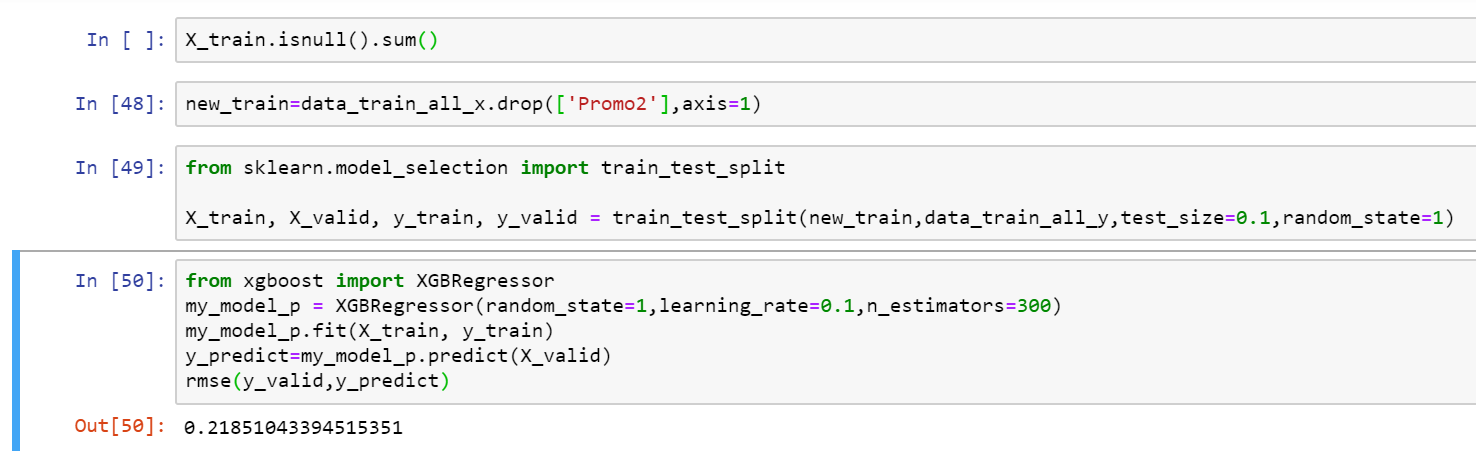
然而，使用此种方法没有异常数据。将系数改为1.4倍的四分位距后数据太多，影响了预测效果。因此，在最后使用了手动的方式进行了outlier的剔除。

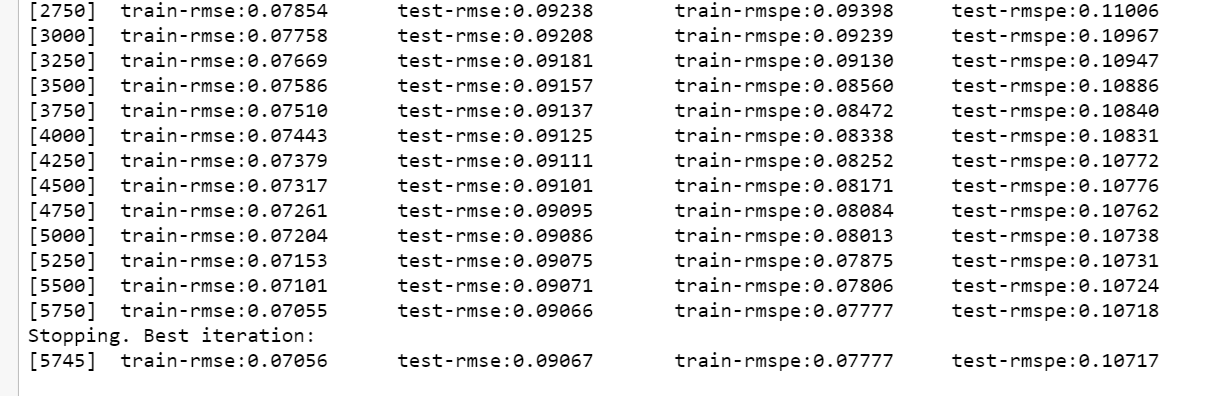


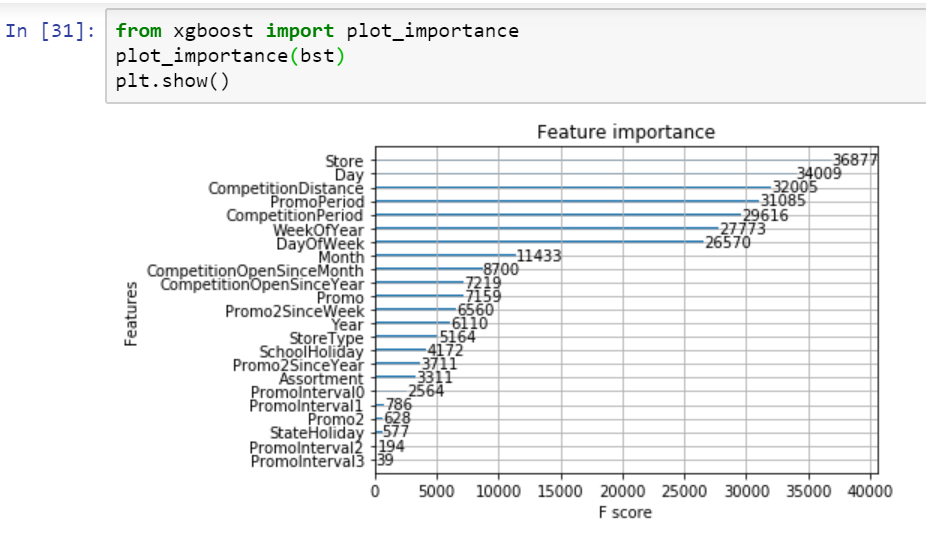
3.3.3 当删除最后四位特征是，预测结果增加，因此减少了特征删除的数量。

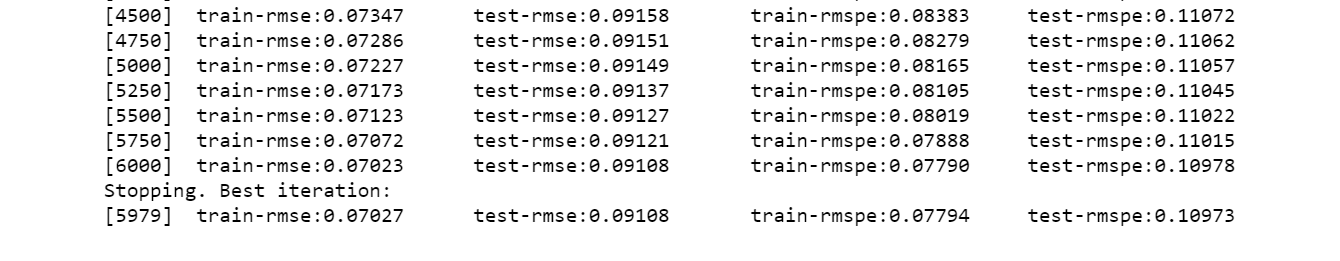
0.2199901377133706

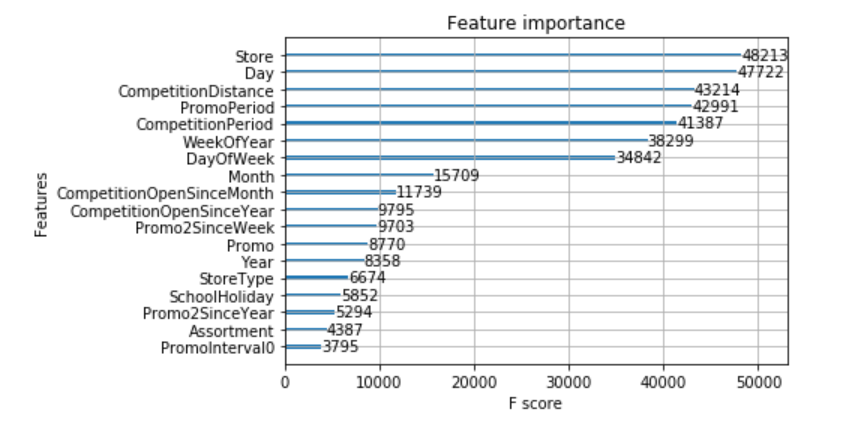








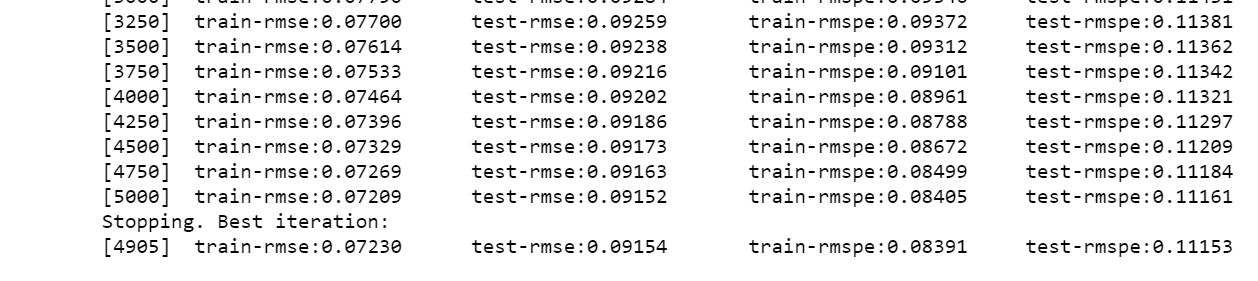


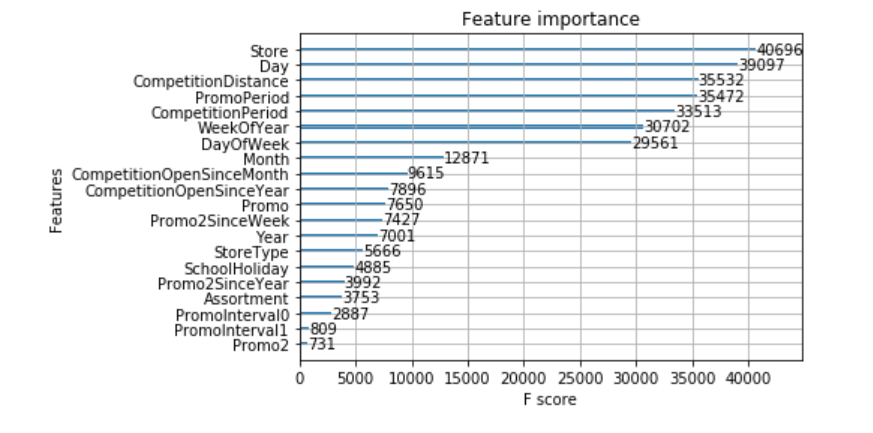


data\_train\_all.drop(['PromoInterval2','PromoInterval3'],axis=1,inplace=True)

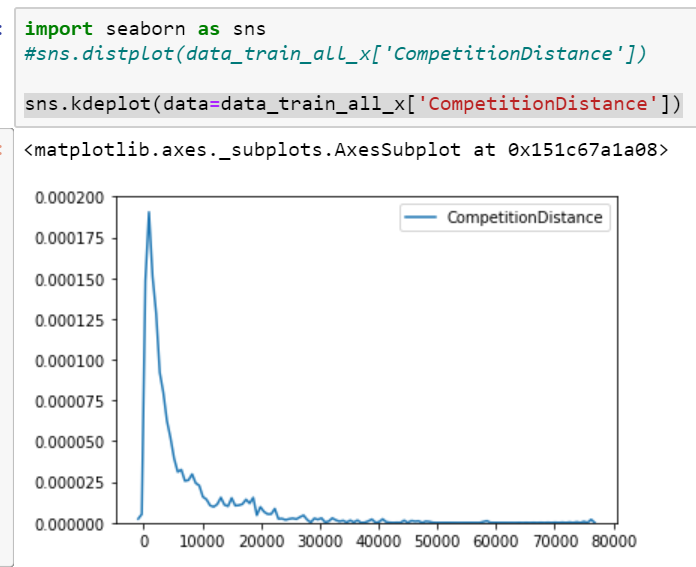
has the best score so far

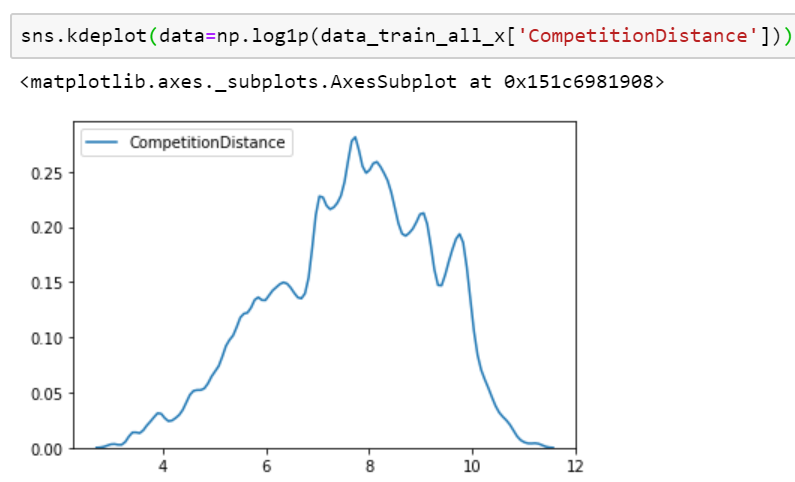
3.3.4在参数方面增加了**min\_child\_weight，并不断增大其数值，明显的改善了预测错误。**





3.3.5 数据形态方面对于skew的数据分布，目的在于将特征数据的分布拉近与正太分布，例如：





Log\_ CompetitionDistance =np.log(data\_train\_all\_x['CompetitionDistance'])

data\_train\_all\_x['CompetitionDistance']= np.log(data\_train\_all\_x['CompetitionDistance'])

Log\_ CompetitionOpenSinceMonth =np.log(data\_train\_all\_x['CompetitionOpenSinceMonth']))

data\_train\_all\_x['CompetitionOpenSinceMonth']= np.log(data\_train\_all\_x['CompetitionOpenSinceMonth']))

Log\_ Promo2SinceWeek =np.log(data\_train\_all\_x['Promo2SinceWeek'])

data\_train\_all\_x['Promo2SinceWeek']= np.log(data\_train\_all\_x['Promo2SinceWeek'])

Log\_CompetitionPeriod =np.log(data\_train\_all\_x['CompetitionPeriod'])

data\_train\_all\_x['CompetitionPeriod']= np.log(data\_train\_all\_x['CompetitionPeriod'])

Log\_ PromoPeriod =np.log(data\_train\_all\_x['PromoPeriod'])

data\_train\_all\_x['PromoPeriod']=np.log(data\_train\_all\_x['PromoPeriod'])

**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？*

该模型在一方面通过max\_depth, min\_child\_weight的应用直接控制和增加了复杂度。另一方面通过subsample and colsample\_bytree增加了随机性，使其对于噪音更加的健全。

最优结果运行两次，结果。。。。

**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？*

对于有限特征中，新特征的生成，需要有更多的商业经验和项目经验，从而生成对现存特征有较低的相关性且数据pattern新颖的特征。

**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：

对于，Gamma，Alpha参数对于模型的影响和效果在计算资源允许的情况下有待于实验应用。

* *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？



Kaggle 0.12147

基于18

SageMaker1

Round1

从'bst:max\_depth':12,到'bst:max\_depth':14,



Kaggle0.18266 18 变差



Round2

Feature,competitionDistance with log1p





Roud3

从'bst:max\_depth':12,到'bst:max\_depth':13,





Round4

'bst:max\_depth':12, 'min\_child\_weight':8,





提高了，错误减小了

Round5

'min\_child\_weight':8,





错误增大.

最终锁定

'min\_child\_weight':7,

SageMaker2

Round 1

从min\_child\_weight':6到min\_child\_weight':8



Kaggle 0.20818 变差



从min\_child\_weight':6到min\_child\_weight':7

Round2

到min\_child\_weight':7



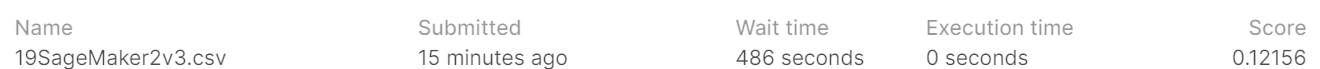


Kaggle 0.18599 改进了

Round 3

max\_depth':11 和SageMaker3 Round 2 比





无差别，放弃max\_depth的调整

SageMaker3

Gamma =0到Gamma =0.2



Kaggle 0.19376 变差



Round 2

Feature 完全没有log



Kaggle 0.12156



Round 3 18SageMaker3v3.csv

Old

data\_store.CompetitionDistance.fillna(data\_store.CompetitionDistance.median())

new

data\_store.CompetitionDistance.fillna(0)





Kaggle 0.12250 实际上测试集上数据变差了

Round 4

Add promomonth





Round 5

Add promointervel 0,1 back





从第二轮开始所有特征不使用log

1Why

W=[(0.990+(i/1000)) for i **in** range(20)]

2

error = rmspe(np.expm1(ho\_ytest), np.expm1(yhat\*w))

ho\_ytest

yhat

yhat = gbm.predict(xgb.DMatrix(ho\_xtest))

3

error = rmspe(np.expm1(ho\_ytest), np.expm1(yhat\*w))

train = train.sort\_values(['Date'],ascending = False)

ho\_test = train[:6\*7\*1115]

ho\_train = train[6\*7\*1115:]

*# split the last 6 weeks data as hold-out set (idea from Gert https://www.kaggle.com/c/rossmann-store-sales/discussion/18024)*

train =data\_train.sort\_values(['Date'],ascending = False)

ho\_test = train[:6\*7\*1115]

ho\_train = train[6\*7\*1115:]

ho\_test = ho\_test[ho\_test["Sales"] > 0]

ho\_train = ho\_train[ho\_train["Sales"] > 0]