1.定义

1.1 项目概述

项目是针对 rosman 的历年销售额，进行分析建模以便可以预测未来的销售额。

Rossmann 是欧洲的一家连锁药店。 在这个源自 Kaggle 比赛 Rossmann Store Sales 中，我们需要根据 Rossmann 药妆店的信息（比如促销，竞争对手，节假日）以及在过去的销售情况，来预测 Rossmann 未来的销售额。

解决该问题涉及回归算法领域，数据集使用的 rosman 提供的销售数据以及门店信息数据。

1.2 问题陈述

需要对项目的输入数据进行特征处理，获取相关的数据来训练回归模型。通过最后的模型训练，可以有效的对未来的数据进行预测。

1.3 评价指标

使用 xgboost 集成算法以及 rmspe 评价指标。

$$\mathrm{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2}$$

2.分析

数据的探索

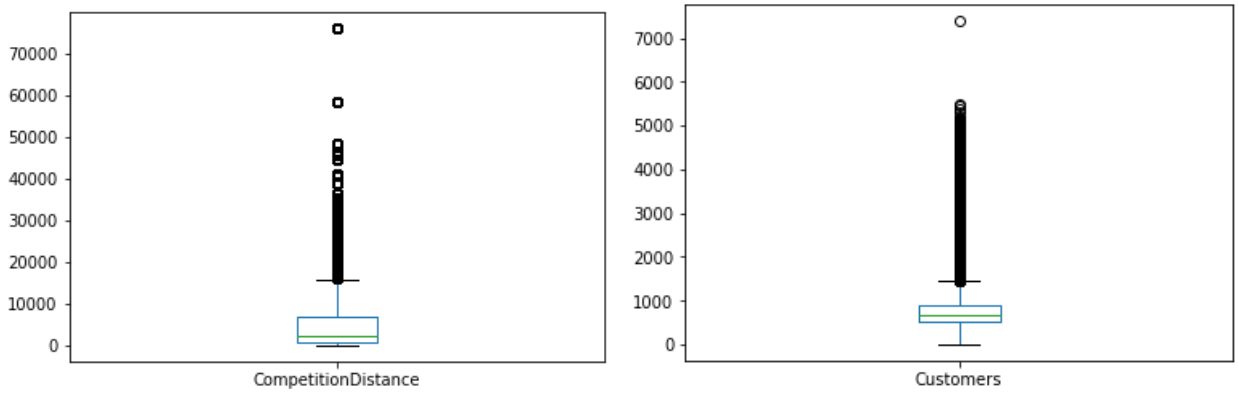
该训练数据集包括两类，历史销售数据以及每个商店的个体信息，我们可以通过结合这两个数据集来训练预测。

对两类数据合并后有如下的训练特征：

```
['Store', 'DayOfWeek', 'Date', 'Sales', 'Customers', 'Open', 'Promo',
       'StateHoliday', 'SchoolHoliday', 'StoreType', 'Assortment',
       'CompetitionDistance', 'CompetitionOpenSinceMonth',
       'CompetitionOpenSinceYear', 'Promo2', 'Promo2SinceWeek',
       'Promo2SinceYear', 'PromoInterval']
```

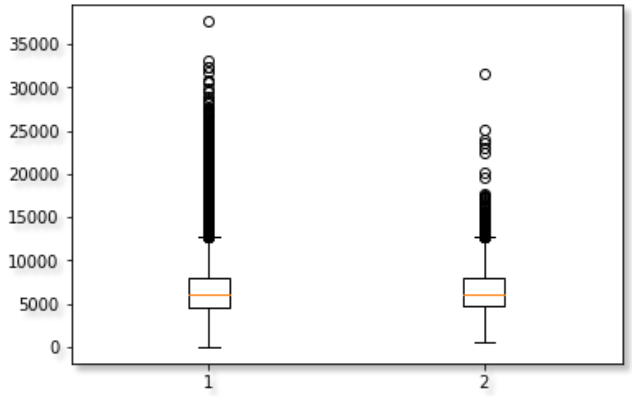
分析异常值：

通过对 customer，CompetitionDistance 存在异常值

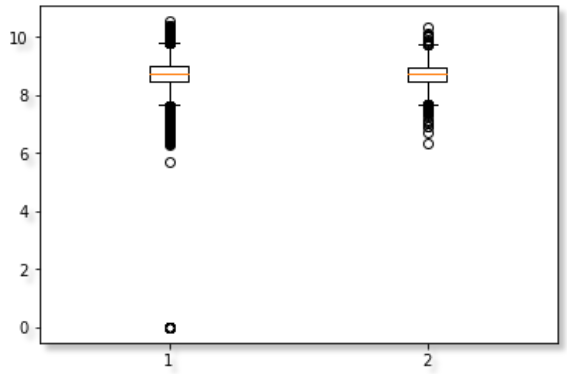


我们可以对异常值进行删除，但是考虑到我们这次用的算法是 xgboost，所以不需要对异常值做太多的处理

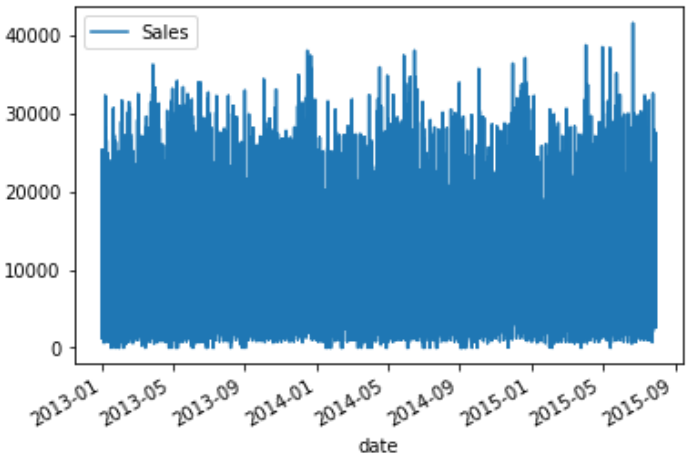
目标数据是 Sales 特征，对这个特征的盒线图显示，数据存在比较大的异常偏差。



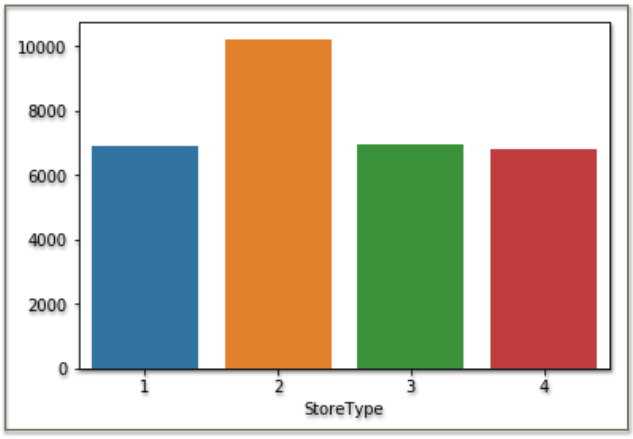
所以对标签数据进行平滑处理，使其服从正态分布。使用 log 平滑处理。



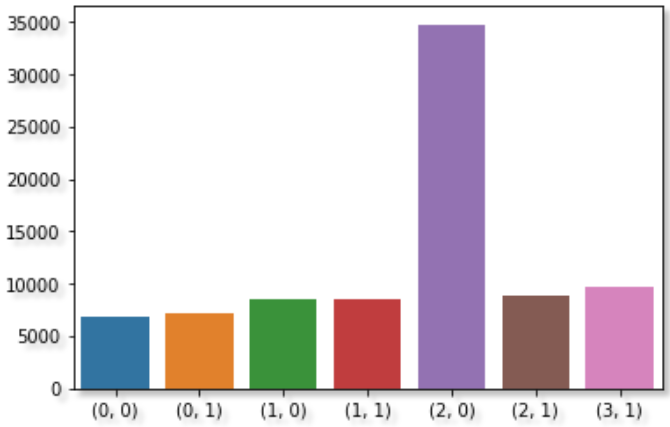
Sales 也是这次项目的预测标签，查看下从 13 年到 15 年的 sales 数据趋势



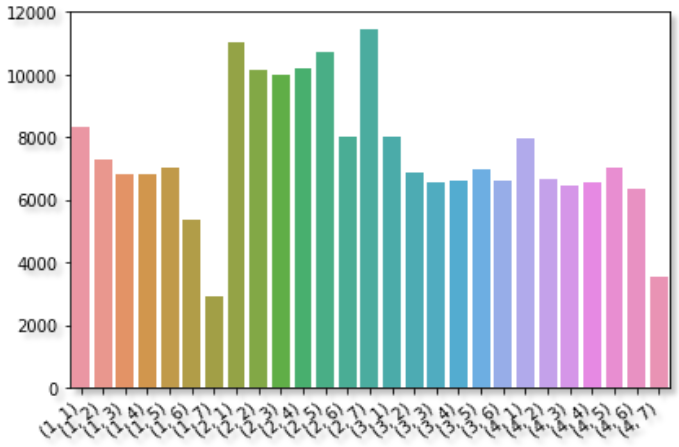
StoreType 对于销售额的影响很大，从柱形图就可以看出不同 type 的平均销售额是有差别的

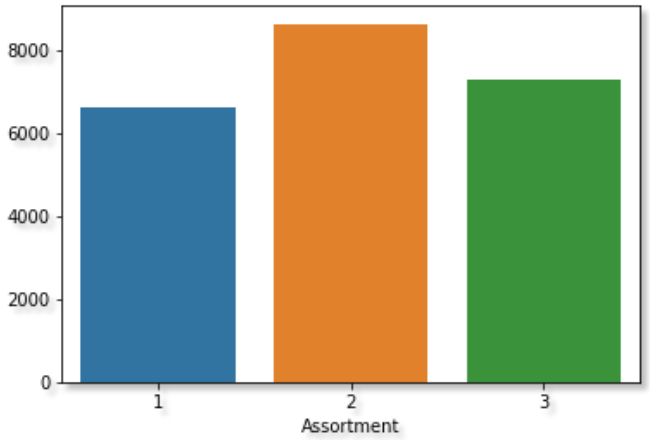


发现在不同的节假日的情况下，学校的开关也对销售额有很大的影响



从直觉上来说 Assortment 也会影响销售额





DayOfWeek 对于销售额也影响

通过绘制图获取的 feature 特征包括

```
['DayOfWeek', 'Sales', 'Promo', 'StateHoliday', 'SchoolHoliday',
       'StoreType', 'Assortment', 'CompetitionDistance',
       'CompetitionOpenSinceMonth',
'CompetitionOpenSinceYear', 'Promo2',
       'Promo2SinceWeek', 'Promo2SinceYear']
```

在数据集中有很多标签数据，统一进行 label encode

```
mappings = {'0':0, 'a':1, 'b':2, 'c':3, 'd':4}
data.StoreType.replace(mappings, inplace=True)
data.Assortment.replace(mappings, inplace=True)
```

```
data.StateHoliday.replace(mappings, inplace=True)
```
使其适合模型的训练运算

因为使用的模型是 xgboost,所以从原理上来说就不需要进行一些类似离群数据，缺失数据的预处理。

在数据集的选择上，82 原则将训练数据分为 80%训练，20%验证数据集。

算法和技术

主要运用的是 xgboost 算法。因为这个问题是回归问题，xgboost 内部使用的 CART tree，这种结构可以处理分类回归问题。而且 xgboost 的特点是它能够自动利用 CPU 的多线程进行并行，同时在算法上加以改进提高了精度。也是一种集成方法。集成方法的有点就是采用很多弱学习器最后合并成强学习器，效果会比较好。

Xgboost 在学习的过程中， 或不断的通过上一次学习的模型的残差进一步学习，最终将损失缩小。

基准模型
RMSPE 的得分在 0.12 左右

方法

数据预处理

数据中的标签数据都要进行 label encode.
目标数据存在很大的离群值，所以需要进行平滑处理。
在分析数据的过程中，发现时间序列对于数据学习很有效果，所以对 date 进行特征提取，获取年月日的数据

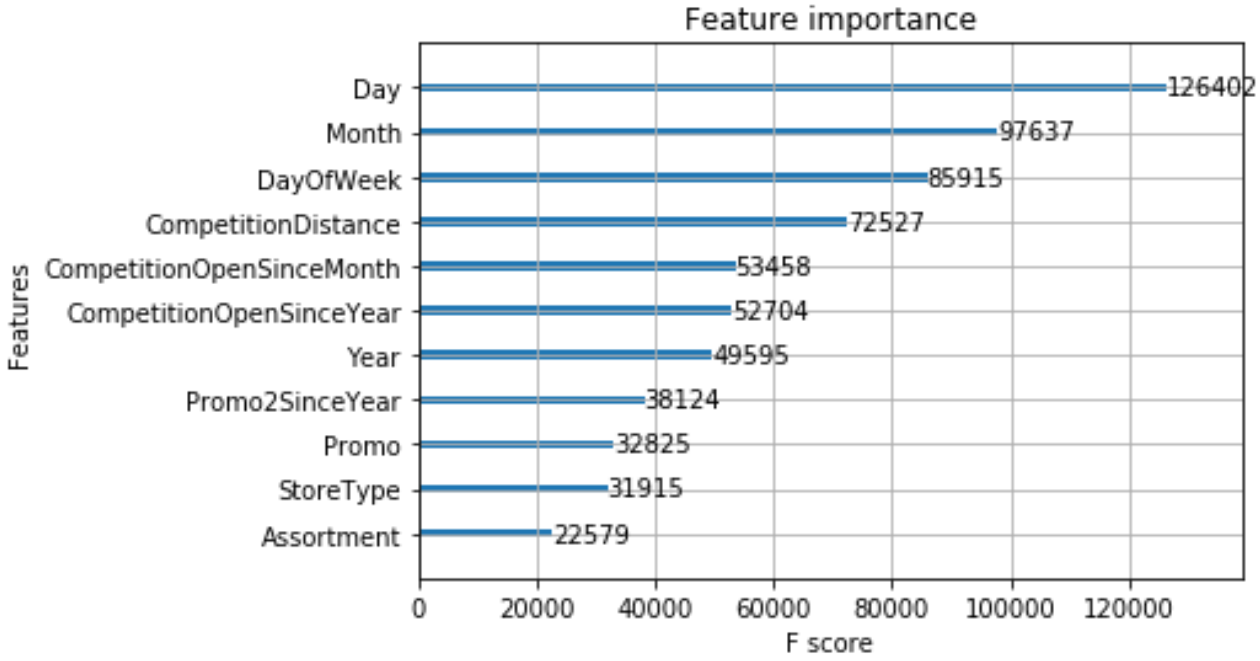
执行过程

一开始对数据分别进行预处理，特征提取和特征选择，然后进行 train valid 的划分，定义 xgboost 相关的参数，然后进行训练。
在训练期间，准备着手针对

"eta"，"max_depth"，进行调参，eta 设计到学习的速率问题，max_depth 可以增加树的深度从而提升拟合度。

完善

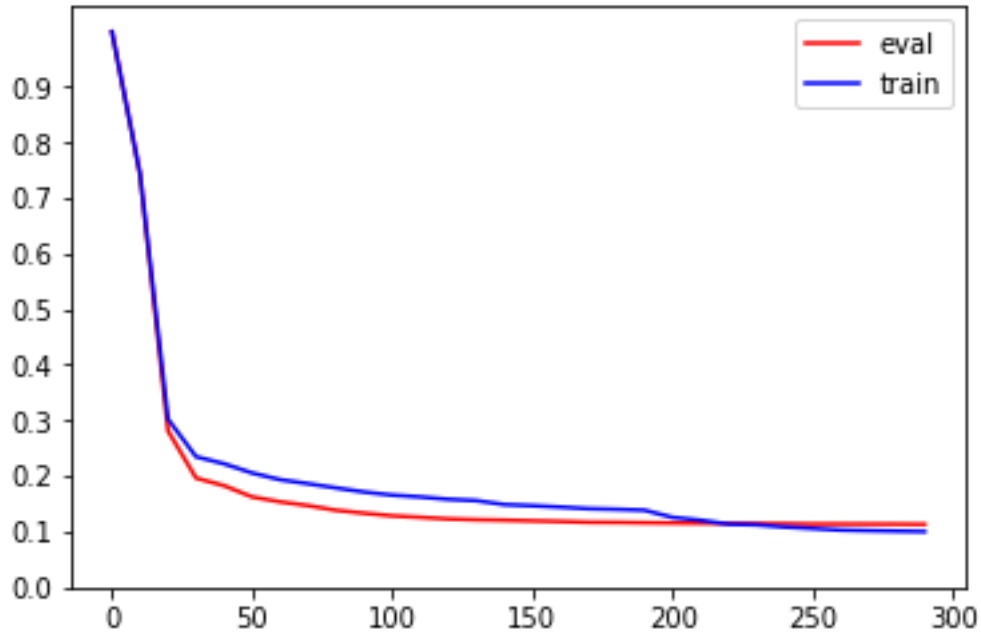
在训练过程中，结果都不是特别好都在 0.17-0.18 左右，这是后尝试加入时间信息，结果提升很多，从最后的 feature important 上也能看到，时间序列确实很重要。



模型的评价与验证

使用 CV 验证，调试参数，获取最终的模型，结果在 0.11605.

选取最终的模型根据 train 和 valid 数据的 RMSPE 的值结果选择。偏差和方差都比较合理。



总体来说，现在这个模型表现还是比基准模型好很多。

结论

Rosman 项目通过分析数据，可视化，对数据进一步获取了解，然后运用回归处理技术。有很多回归技术可以用，但是 xgboost 是目前最为流行的，速度快强大，容易使用上手。

目前这个模型的得分在 kaggle 上排名：1809.

效果不是特别理想，应该还有更有效的时间特征可以提取。

引用：

https://xgboost.readthedocs.io/en/latest/tutorials/index.html

https://zhuanlan.zhihu.com/p/54334329

https://www.kaggle.com/c/rossmann-store-sales/discussion/18024