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。

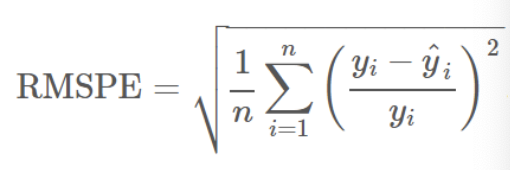
Rossmann是欧洲的一家连锁药店。 在这个源自Kaggle比赛[Rossmann Store Sales](https://www.kaggle.com/c/rossmann-store-sales)中，我们需要根据Rossmann药妆店的信息（比如促销，竞争对手，节假日）以及在过去的销售情况，来预测Rossmann未来的销售额。

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

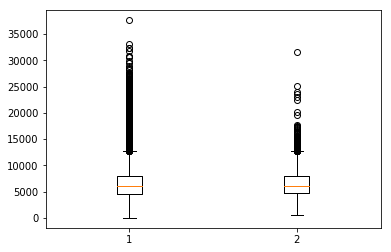
需要对项目的输入数据进行特征处理，获取相关的数据来训练回归模型。通过最后的模型训练，可以有效的对未来的数据进行预测。

1.3 评价指标

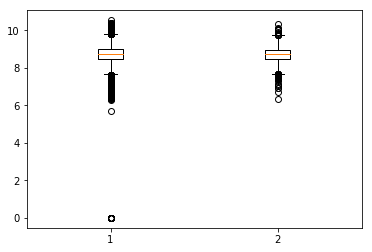
使用xgboost集成算法以及rmspe评价指标。

2.分析

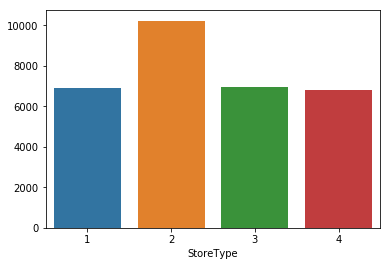
数据的探索

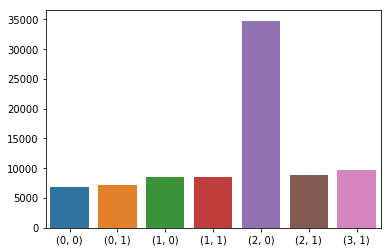
目标数据是Sales特征，对这个特征的盒线图显示，数据存在比较大的异常偏差。

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。

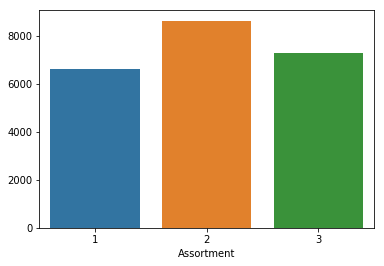


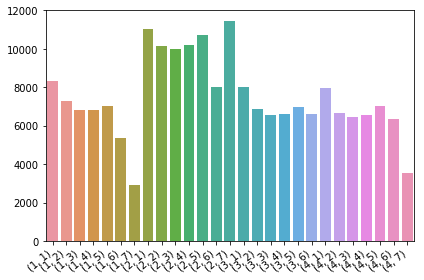
StoreType对于销售额的影响很大，从柱形图就可以看出不同type的平均销售额是有差别的



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