**训练集特征**

**DayOfWeek：**星期几

**Store** - 商店id

**Id** - 测试集中id,表示（商店，日期）

**Sales -** 营业额

**Customers** - 顾客数量

**Open**- 商店是否开业

**Promo**- 商店是否短期促销

**StateHoliday** - 表示国家假日。正常情况下，除了少数例外，所有商店在国定假日都不营业。请注意，所有学校在公众假期和周末休课。a =公众假期，b =复活节假期，c =圣诞节，0 =没有。

**SchoolHoliday**- 指示(商店、日期)是否受公立学校关闭的影响

**商店特征**

**StoreType**- 商店模式:a, b, c, d

**Assortment**- 描述分类级别:a = basic, b = extra, c = extended

**CompetitionDistance** ：到最近竞争对手商店的距离

**CompetitionOpenSince[Month/Year]** - 最近的竞争对手开业的大概年份和月份

**Promo2**- 商店的连续促销:0 =商店不参与，1 =商店参与

**Promo2Since[Year/Week]** - 描述该店开始参与促销活动的年份和日历周

**PromoInterval** - 描述启动连续时间间隔Promo2，指定重新启动促销的月份。如，“Feb,May,Aug,Nov”是指该店每年2月、5月、8月、11月开始的每一轮。

测试集缺失数据分析：

缺失数据都来自于622店铺，从周1到周6而且没有假期，所以我们认为这个店铺的状态应该是正常营业的。

店铺缺失数据分析：

#查看是否Promo2系列的缺失是因为没有参加促销活动。

在实验中我们发现竞争对手信息的中值填充效果并不好，所以这里统一采用0填充。

#分析店铺销量随时间的变化

店铺的销售额是有周期性变化的，一年之中11，12月份销量要高于其他月份，可能有季节因素或者促销等原因。

#分析店铺6-9月份的销量变化

销售趋势与8，9月份类似，因为我们需要预测的6周在2015年8，9月份，因此我们可以把2015年6，7月份最近的6周数据作为hold-out数据集，用于模型的优化和验证。

特征处理：

将存在其他字符表示分类的特征转化为数字。

将时间特征进行拆分和转化，并加入'WeekOfYear'特征。

新增'CompetitionOpen'和'PromoOpen'特征,计算某天某店铺的竞争对手已营业时间和店铺已促销时间，用月为单位表示。

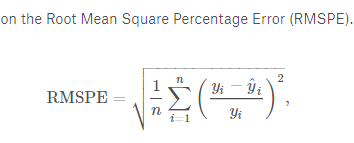
将'PromoInterval'特征转化为'IsPromoMonth'特征,表示某天某店铺是否处于促销月，1表示是，0表示否。

删掉训练和保留数据集中不需要的特征。

#分析训练数据集中特征相关性以及特征与'Sales'标签相关性

#拆分特征与标签，并将标签取对数处理，使其更加服从高斯分布

#定义评价函数rmspe

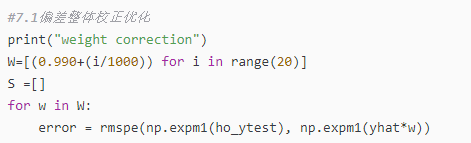




#分析偏差最大的10个预测结果

偏差调整:

当校正系数为0.995时，hold-out集的RMSPE得分最低：0.118889，相对于初始模型 0.125453得分有很大的提升



#7.2细致校正：以不同的店铺分组进行细致校正，每个店铺分别计算可以取得最佳RMSPE得分的校正系数.

细致校正后的hold-out集的得分为0.112010，相对于整体校正的0.118889的得分又有不小的提高。

然后我们用不同的seed训练10个模型,每个模型单独进行细致偏差校正后进行融合

#分析不同模型的相关性

模型融合可以采用简单平均或者加权重的方法进行融合。从上图来看，这10个模型相关性很高，差别不大，所以权重融合我们只考虑训练中单独模型在hold-out模型中的得分情况分配权重。

从模型特征重要性分析，比较重要的特征有四类包括1.周期性特征'Day'，'DayOfWeek'，'WeekOfYera'，'Month'等，可见店铺的销售额与时间是息息相关的，尤其是周期较短的时间特征；2.店铺差异'Store'和'StoreTyp'特征，不同店铺的销售额存在特异性；3.短期促销（Promo)情况:'PromoOpen'和'Promo'特征，促销时间的长短与营业额相关性比较大；4.竞争对手相关特征包括：'CompetitionOpen'，‘CompetitionDistance'，'CompetitionOpenSinceMoth'以及

'CompetitionOpenScinceyear'，竞争者的距离与营业年限对销售额有影响。

# - 作用不大的特征主要两类包括：1.假期特征：'SchoolHoliday'和'StateHoliday'，假期对销售额影响不大，有可能是假期店铺大多不营业，对模型预测没有太大帮助。2.持续促销(Promo2)相关的特征：'Promo2'，'Prom2SinceYear'以及'Prom2SinceWeek'等特征，有可能持续的促销活动对短期的销售额影响有限。