No：

**基于XGBosst的销售预测**

**2020年1月5日**

目 录

[1 项目概述 4](#_Toc29181818)

[1.1 背景信息 4](#_Toc29181819)

[1.2 问题陈述 4](#_Toc29181820)

[1.3 解决方案与评价指标 4](#_Toc29181821)

[2 分析 5](#_Toc29181822)

[2.1 数据探索 5](#_Toc29181823)

[2.2 可视化分析 8](#_Toc29181824)

[2.3 算法和技术 11](#_Toc29181825)

[2.4 基准模型 11](#_Toc29181826)

[3 方法 12](#_Toc29181827)

[3.1 数据预处理 12](#_Toc29181828)

[3.2 执行过程 12](#_Toc29181829)

[3.3 完善 13](#_Toc29181830)

[4 结果 13](#_Toc29181831)

[4.1 模型评价与验证 13](#_Toc29181832)

[5 项目结论 14](#_Toc29181833)

[5.1 思考 15](#_Toc29181834)

[5.2 待改进点 15](#_Toc29181835)

[参考文献 错误!未定义书签。](#_Toc29181836)

[致 谢 17](#_Toc29181837)

# 项目概述

## 背景信息

Rossmann在7个欧洲国家中有超过3000家药店。该项目的主要目的是利用这些药店过去几年的销售数据，去预测未来6周各药店的销量。

商店的销量受非常多的因素影响，包括促销信息、竞争对手信息、开学信息和国家节日（包括季节性的节日和当地的节日），以往的销量预测依靠各商店主管对于周期的判断，预测的准确度也因人而异。

更加精准的预测有助于商店制定更加合理的制定备货与促销策略，增大利润的同时，更好的迎合消费者的需求，提升购买体验。

## 问题陈述

本项目主要是利用1115家Rossmanns商店的历史销售数据，去预测未来6周每个商店的销量，历史销售数据属于时序数据，总体上看该问题属于典型的监督学习回归问题。

主要涉及的数据主要分为两个部分：

1. 历史销售数据：为每个商店按日期的时序型数据，包括1115家门店的编号、日期、销售额、访问客户数、营业状态、国家节日类型、校园日、促销状态信息；
2. 商店数据：包括商店编号、商店类型、竞争对手信息、促销月份、促销间隔信息。

## 解决方案与评价指标

本项目参考了Kaggle上若干经典解决方案后，拟采用XGboost模型进行回归预测，评价指标选用Kaggle针对该项目的指定指标RMSPE

RMSPE属于MSE的衍生指标，MSE全称为均方根误差，属于回归类分析的常见评价指标，其公式为（y\_true表示真实值，y\_pred表示预测值）：

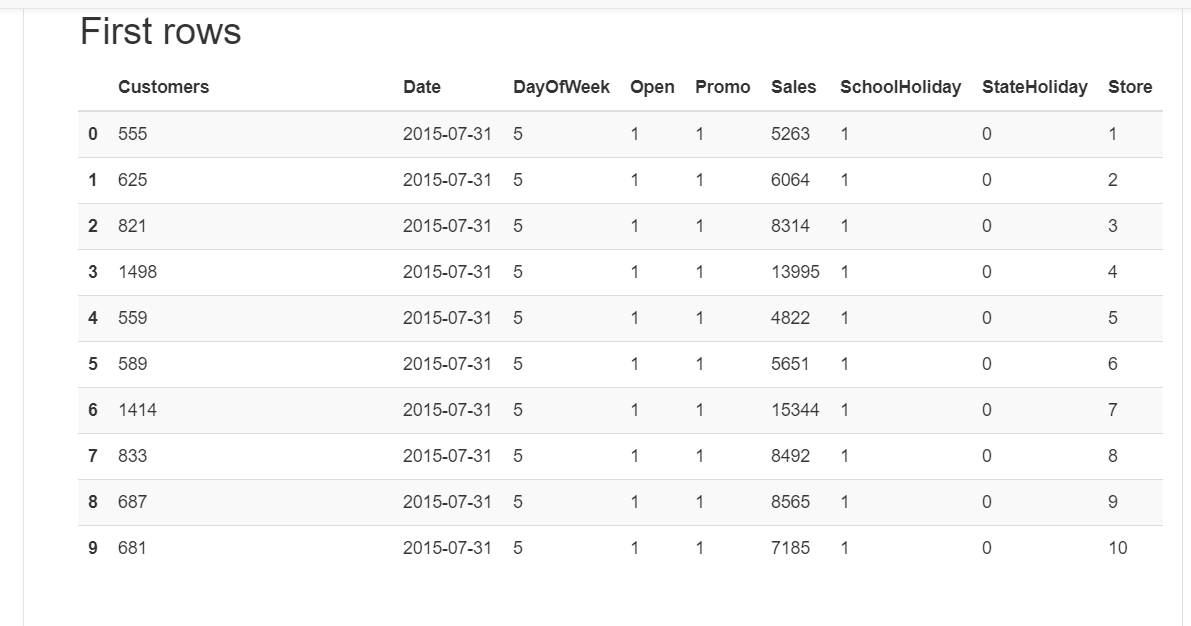
RMSPE的公式为：

# 分析

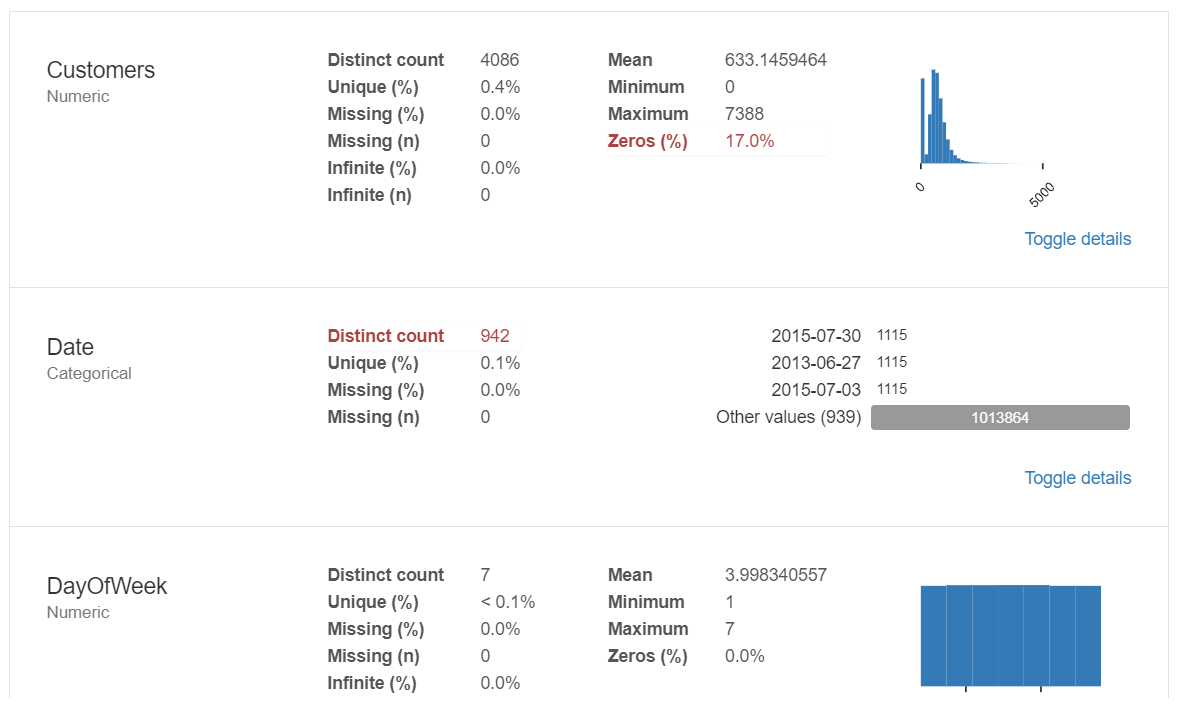
## 2.1 数据探索

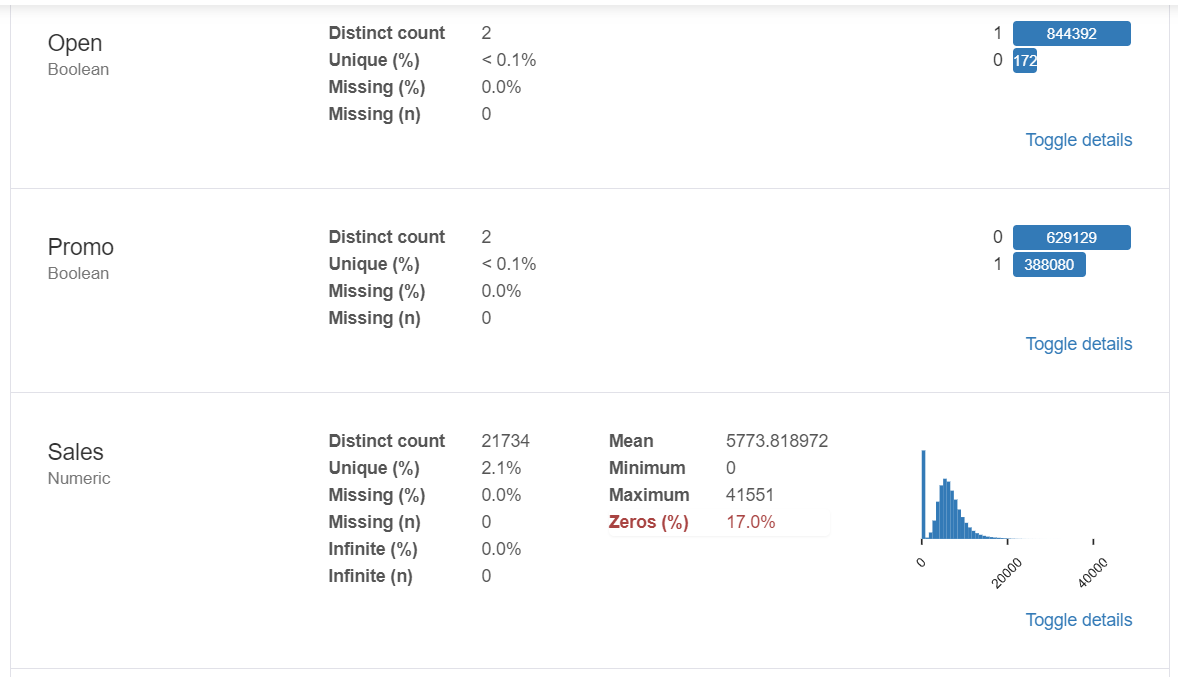
利用Pandas\_Profiling对销售数据集与商店数据集进行数据探索分析。

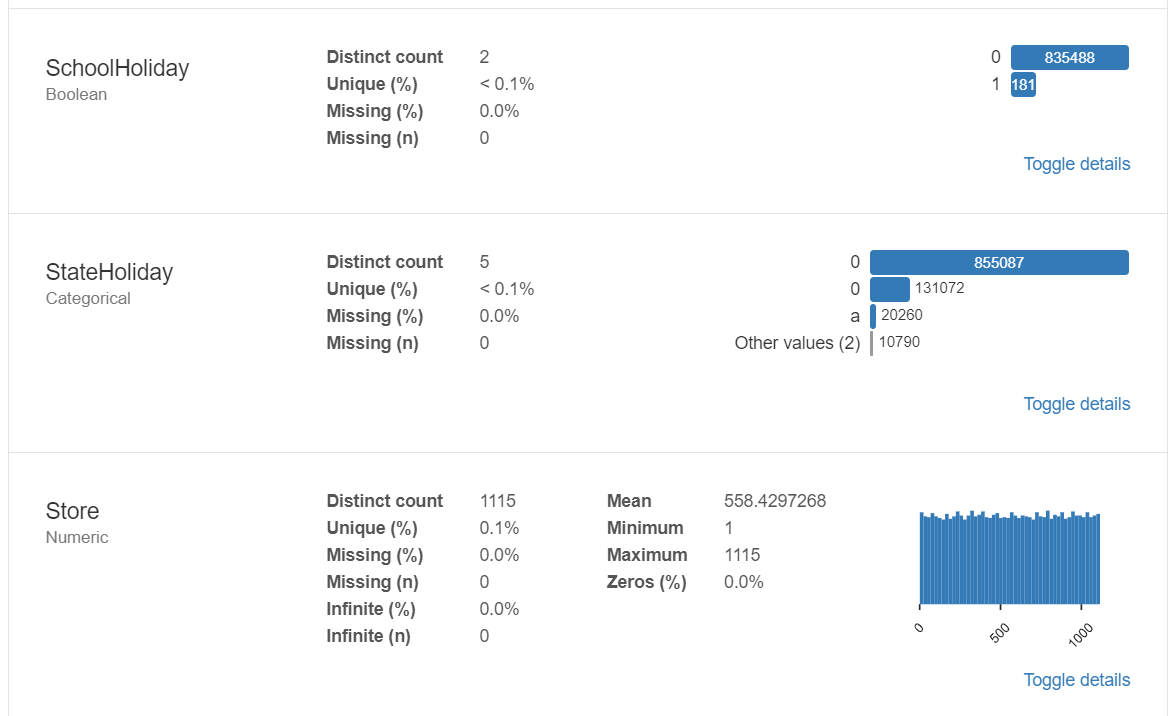
本项目的销售数据共计1,017,209行，共计9个特征，数据集前5行如下图所示：



其中各特征类型如下：





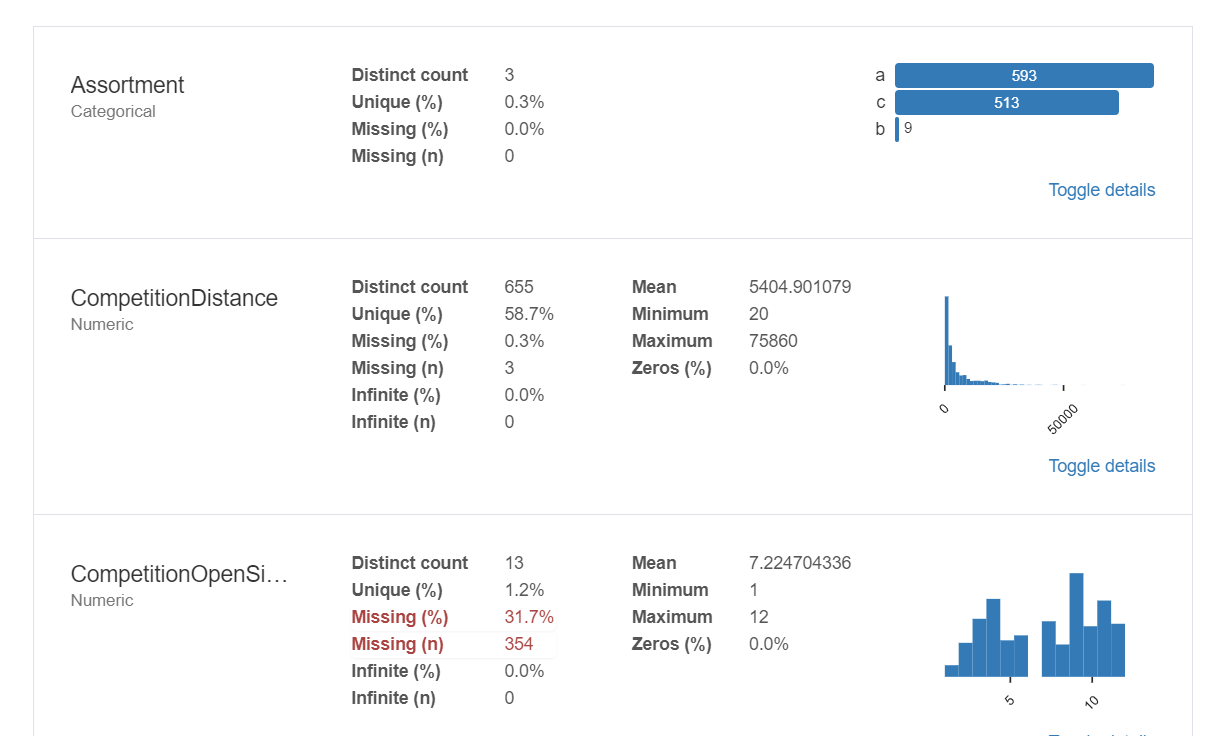


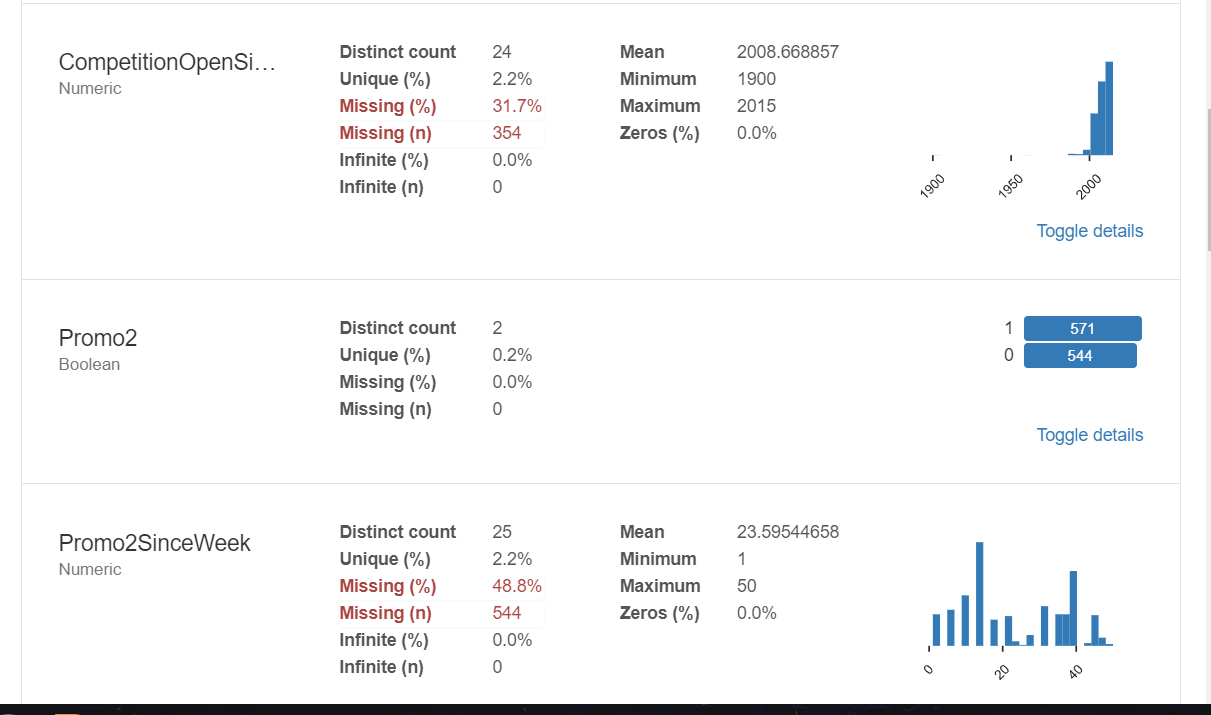
综上可以看出，销售数据集中共有4个数值型特征、2个类别型特征和3个布伦型特征。缺失值方面，仅有Customers与Sales两个特征有0值存在，且个数相等，进一步分析发现缺失原因为该商店当日为非营业状态，后期在数据预处理阶段进行0值补缺。

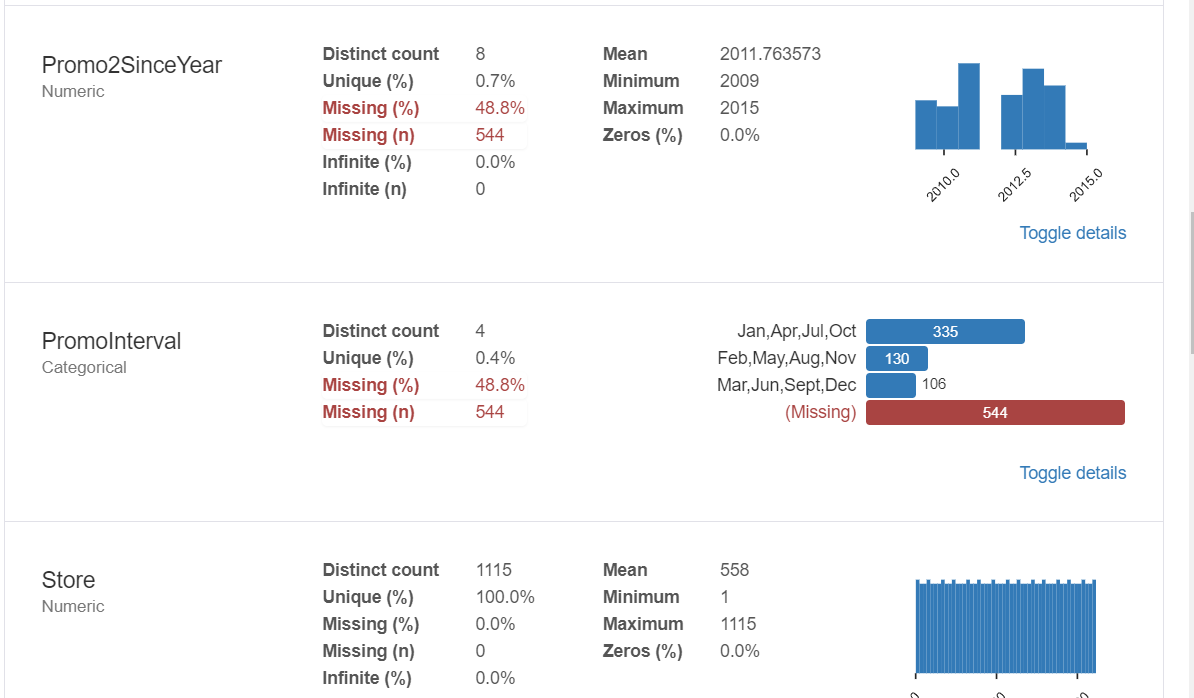
本项目的商店数据共计1,115行，共计10个特征，数据集前5行如下图所示：



其中各特征类型如下：



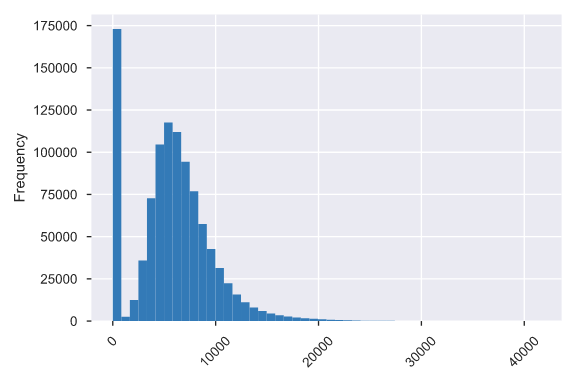




综上可以看出，商店数据集中共有6个数值型特征、3个类别型特征和1个布伦型特征。缺失值方面，CompetitionOpenSinceMontn、CompetitionOpenSinceYeaR、Promo2SinceWeek、Promo2SinceYear、PromoInterval含有一定比例的缺失值，主要原因为部分商店周边暂未有竞争对手或者未参与过促销活动，后期将对这些特征进行映射补零操作

## 2.2 可视化分析

**整体销量额偏态分布：**



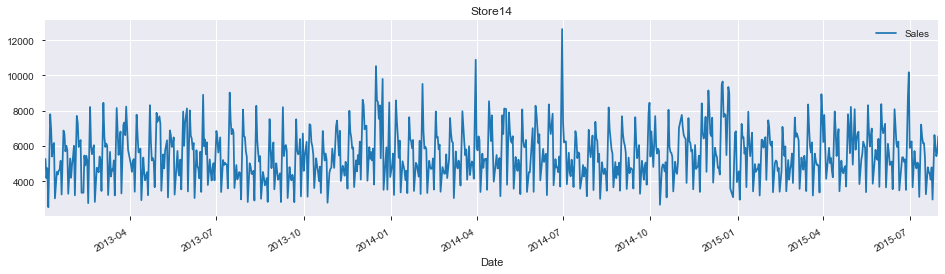
查看所有销售额的分布，会发现销售额成偏态分布，且有大量的0值存在，0值记录应整体剔除。同时剔除后，应使用log函数将销售额平滑成正态分布。

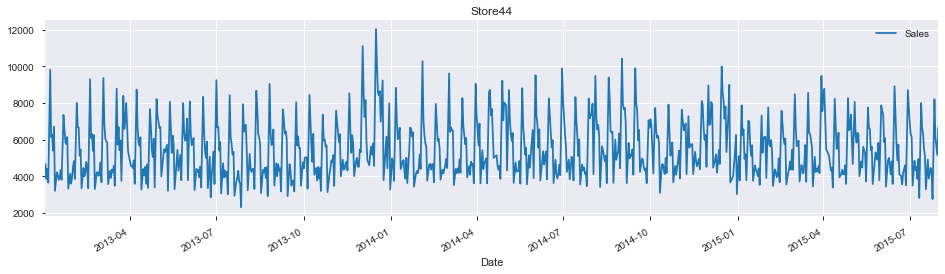
**销售类数据集的特征间关联性：**



对销售类数据集的绘制关联热图（如上图），颜色深浅表示的为各特征间Pearson关联系数的大小，红色表示正向关联，蓝色表示负向关联。由上图可见，除客户数量、营业状态外，销量还与促销状态成正相关。因此，在后期特征抽取时，应重点抽取促销类信息。

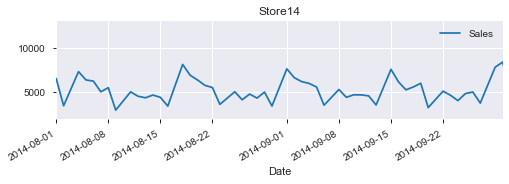
**商店销售的周期性：**





分别选取商店编号为14、44号商店绘制销量图，发现上述两家商店销售额呈现一定的周期性波动。





更进一步，查看选中14号商店的2014年6-7月销量和7-9月销量，发现销量波动呈现一定相似性。从波动规律可以看出，由于销量周期性的存在，短期销量预测具有可行性，同时在特征选取时，应保留时间特征，确保数据集中有特征可以体现时序性。

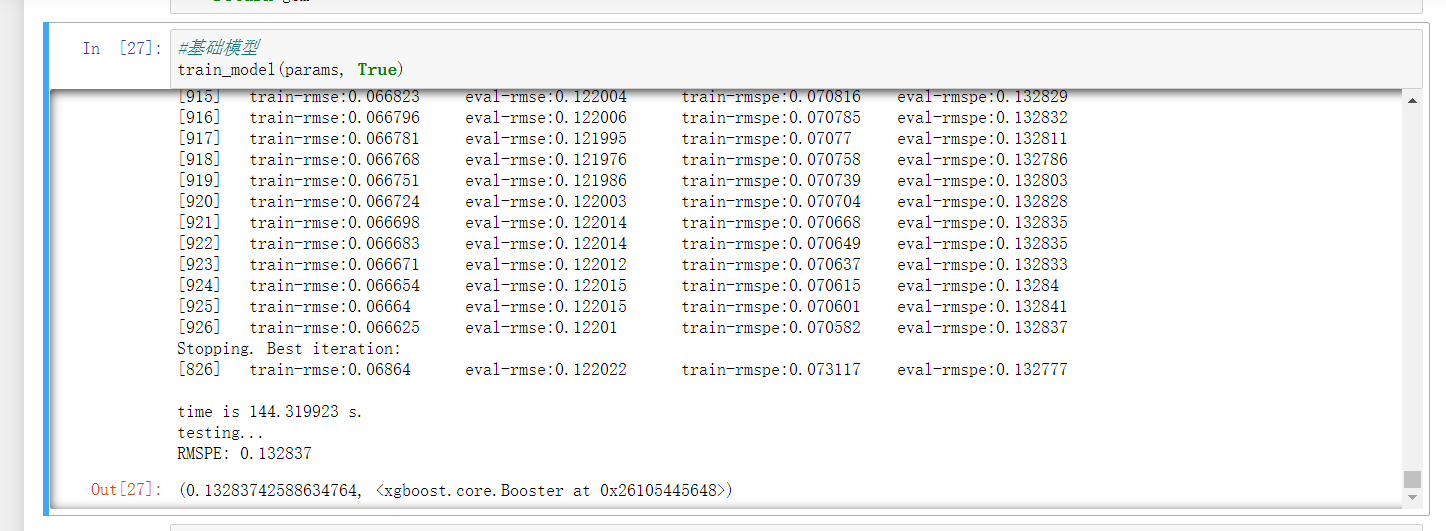
## 2.3 算法和技术

本项目将采用XGboost模型进行回归预测。XGboost本质属于Boosting 分类器模型，属于集成学习模型，它基本思想是把成百上千个分类准确率较低的树模型组合起来，成为一个准确率很高的模型。这个模型会不断地迭代，每次迭代就生成一颗新的树。一般来说，gradient boosting的实现是比较慢的，因为每次都要先构造出一个树并添加到整个模型序列中，而 XGBoost 的特点就是计算速度快，模型表现好[1]。

本项目还将利用GPU加速XGboost的训练过程，完成参数的网格搜索，达到模型参数的最优化。

## 2.4 基准模型

本项目以XGboost初始参数的模型为基准模型[2]，测得基准RMSPE得分0.132837。同时，项目目标为结业最低要求，即测试集的最低RMSPE为0.11773



# 方法

## 3.1 数据预处理

Null值处理。商店数据的Null值主要集中于竞争信息相关与促销信息相关的特征，Null值的主要原因为该商店附近未有竞争对手或未进行促销，因此针对该部分Null的填充方法为使用0值填充。另外测试集中，也存在Null，主要为营业状态，对于该部分Null使用1值进行填充，默认商店处于营业状态。

日期格式转换。销售数据集中，存在日期格式的特征Date。由可视化分析中可知，日期特征是重要的时序型特征，也是表达周期性的关键特征，但由于日期型无法直接用于模型计算，故将日期拆分为年、月、日三个整数型特征。

销售数据集与商店数据集合并。由于销售数据集与商店数据集都是模型训练的关键数据，因此利用pandas的merge方法，将两个数据集合并为一个数据集用于模型训练。

拆分训练集与验证集。将合并后的数据集进一步拆分为训练集与验证集，由于本项目的目标是预测未来6周的销售额，故将合并后的数据集中，最后6周的1115家商店的数据拆分出来作为验证集，其余数据作为训练集。同时，由于当商店处于未营业状态时，销售额为0，这些数据在训练或者验证集中都对模型训练具有一定的干扰，故从上述数据集中剔除。

字符型（分类型）特征进行映射转换。针对StateHoliday、Assortment与StoreType三个字符型特征，由于数据探索是发现这三类特征的值属于离散型，且值不超过5种类型，故针对5种取值进行映射转换，分别映射为0-5的整数。

其他特征变换。针对特征竞争对手营业时间，将其变换为竞争对手至今营业天数，促销开始时间转换为促销存续时间，同时针对促销月份信息，生成新特征“是否处于促销月份”，最大程度保留促销信息[3]。

## 3.2 执行过程

1. 使用Google 云环境配置GPU服务器
2. 编写相关函数，包括网格搜索函数与RMPSE函数
3. 设置模型初始参数值
4. 设定网格搜索参数，循环调试得出最佳参数
5. 使用最佳参数训练模型
6. 将训练后的模型在验证集进行验证
7. 将模型预测测试集，得到待提交数据
8. 根据验证集的预测结果，计算整体偏差，并根据修正参数对待提交数据修正
9. 提交数据，得到结果分数

## 3.3 完善

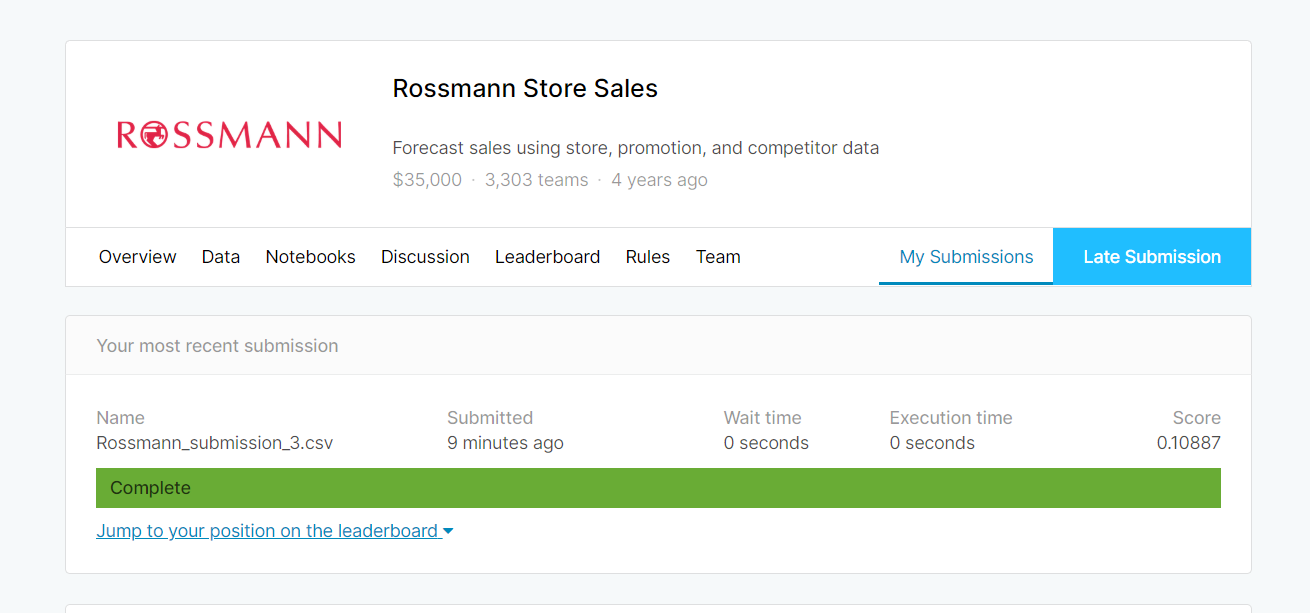
1. 初始参数得到的模型，其在验证集的RMPSE得分为0.13457
2. 后续分别对max\_depth、max\_depth、gamma、subsample与colsample\_bytree、reg\_alpha、reg\_lambda、learning\_rate参数进行了网格搜索，最终确定最佳参数[4]
3. 第一次网格搜索训练时，使用CPU进行训练速度较慢，后经查阅资料发现XGboost支持GPU，故配置GPU环境加速训练过程。整体网格搜索过程大约为2小时
4. 确定最佳参数后并训练模型，得到第一版提交结果，提交Kaggle后发现RMPSE得分为0.11948，与最低要求相差0.002分
5. 进一步查阅相关资料，应用验证集数据进行偏差修正，再次提交后达到分数要求

# 结果

## 4.1 模型评价与验证

最终模型各参数经网格搜索后，设置较为合理，除gmma值外，所有模型初始参数均得到优化，最终模型的RMSPE得分明显高于基准模型，同时解决了项目设定的目标。

经Kaggle评分，模型结果直接得分为0.11948，整体数据修正后达到0.10887，达到项目结业最低要求，结论可靠。

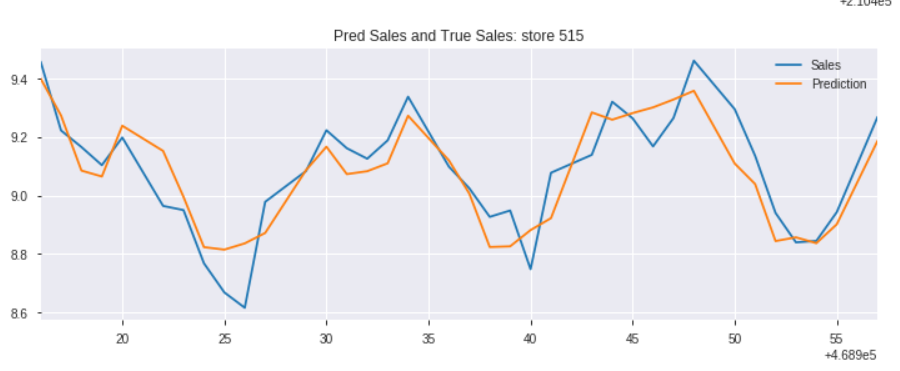


# 项目结论

## 结果可视化









上述四图为随机抽取验证集中四户商店，并对实际销售额与预测销售额进行对比，其中蓝线代表实际销售额，黄线代表预测销售额。由上图可见，模型在验证集中拟合能力较好，预测趋势与实际趋势同频，模型效果较好。

## 思考

本项目的第一个困难点在于如何在较短时间内训练出合理的模型。本次项目第一次在深度学习以外的场景使用GPU加速训练模型，加快了模型训练速度，因此尝试了网格搜索方法进行调参。若采用CPU进行训练，在较短时间内无法完成调参工作。

自主编写网格搜索函数。虽然XGboost也提供了Sci-kit learn的API，但在调试过程和GPU支持过程中还是遇到很多问题，故重新编写了简单调参函数与调参流水线。

该项目对时序性数据进行回归预测有很大的启发。时序的时间预测按照思维惯性一般会采用时序预测模型（例如ARIMA），本次使用XGBoost这种提升树模型对时序性数据进行预测，且效果良好，为后续该类场景应用打开了新的思路。

## 待改进点

引入外部特征数据。其他案例中，引用包括温度在内的其他有效信息，可进一步提升模型预测准确度。

抽取更多特征信息。参照比赛第一名的方案，作者抽取了包括客均单价、上季度销量等在内的统计特征，作为新特征加入训练集，得到了较好的结果。

# 参考文献

[1] Kaggle 神器 xgboost( <https://www.jianshu.com/p/7e0e2d66b3d4>)

[2] [4]XGboost数据比赛实战之调参篇(完整流程) (<https://blog.csdn.net/sinat_35512245/article/details/79700029>)

[3] XGBoost Feature Importance( <https://www.kaggle.com/cast42/xgboost-in-python-with-rmspe-v2/code>)

# 致 谢

本人在毕业设计以及学习课程的时间里，得到了Udacity的导师、群学友的热心帮助和支持，在此向他们表示诚挚的感谢。

台 亮

二O二O年一月