

Data Mining Project Report

3190105959 宁若汐


任务概要

Kaggle的销量预测比赛[Store Sales - Time Series Forecasting | Kaggle](#)。要求用2013-01-01到2017-08-15的9个商店每天33种商品的销售量数据，预测2017-08-16到2017-08-31的销售量。

贡献总结

1. 搭建Prophet模型，在Prophet模型上获得了0.50965的得分，排名240/657。


240



0.50965

4

8m



Your Best Entry!

Your most recent submission scored 0.50965, which is an improvement of your previous score of 0.60028. Great job!

Tweet this

2. 对多个数据表进行整合处理，在现有数据基础上提取出星期、客流量、人均销售额等多个特征，并对这些特征进行多种可视化分析，直观得出销售量的时间、空间变化规律。
3. 对比了SMA、linear regression和Prophet三种常见的时序预测模型在该任务上的表现。（表1）

表1. 几种方法在比赛中的得分（0至5分，其中0分代表表现最好，5分代表表现最差）

| | SMA (baseline) | Linear Regression | Prophet |
|--------|----------------|-------------------|---------|
| Result | 0.60028 | 2.76654 | 0.50965 |

技术方案

本节主要介绍Prophet的模型特点，将Prophet用于本预测任务的方法，以及简要介绍在其它两个模型上的尝试和结果。

Prophet模型特点

Facebook提出的Prophet模型设计主要基于时间序列分析领域的时间序列分解方法，可以将时间序列 y_t 分解成几个部分，分别是季节项 S_t ，趋势项 T_t ，剩余项 R_t 和节假日项 H_t 。即

$$y_t = S_t + T_t + R_t + H_t$$

Prophet就通过训练拟合上述各项，累加起来获得时间序列的预测值。

Prophet在预测未来值时，会在未来值之前的时间序列中选取一定数量的变点用于获取增长率的分布情况，通常变点的个数要求在两位数，因此数据集过小无法训练Prophet。

将Prophet用于销量预测

由于数据集分为多个商店的多个商品类型，各个商店的各种商品的销量间可能关系不大，因此将每所商店的每种商品分别作为一个单独的样本轨道进行预测（如图1），最后合并。共进行

$$54(\text{商店数}) \times 33(\text{商品数}) = 1782$$

次预测。

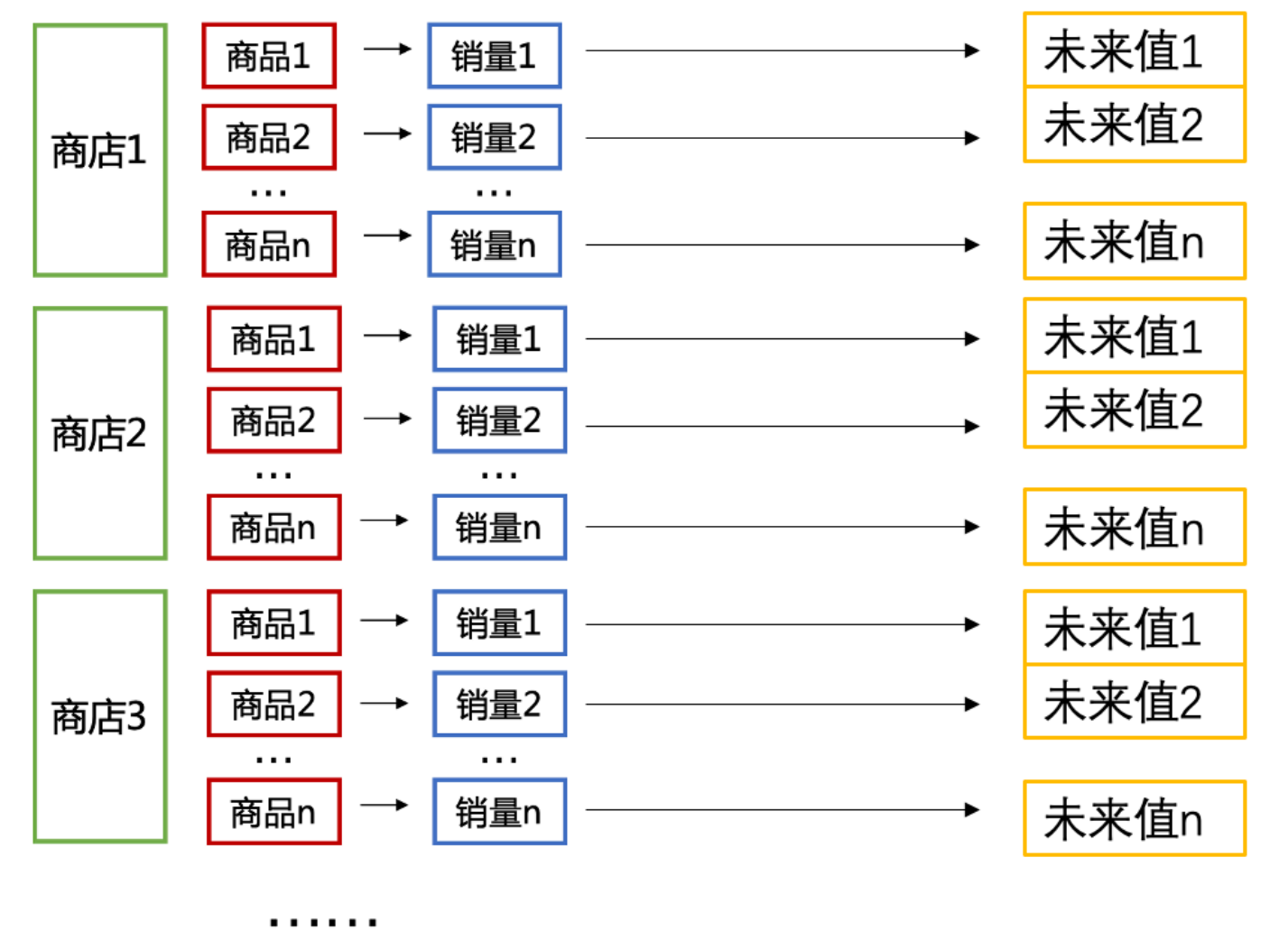


图1. 每个商店的每种商品样本轨道示意图

最后，将每个样本轨道的预测值根据日期、商店编号、商品种类分别填入对应的位置，整合获得完整的预测值。

模型对比与选择

实验初期还进行了与SMA（baseline）与Linear Regression两种方法的对比。其中SMA是简单的滑动平均方法，将前一段时间的平均值作为预测值输出；考虑到数据集特征较多，尝试使用Linear Regression处理多变量，但Linear Regression忽略了太多时序特征，表现并不好。最终选择了使用擅长处理时序特征、接受多变量输入的Prophet模型进行预测任务。

实验分析

数据预处理

本节介绍对数据集的分析，和特征提取工程的过程和结果。

该数据集由6个数据表构成，按照字段连接关系可以画出下方数据表关系图（图2）。由图可知训练集中特征不多，很多与销量有关的特征在其它数据表中，于是下面对数据表进行联合，将与训练集有关的数据合并到训练集中。

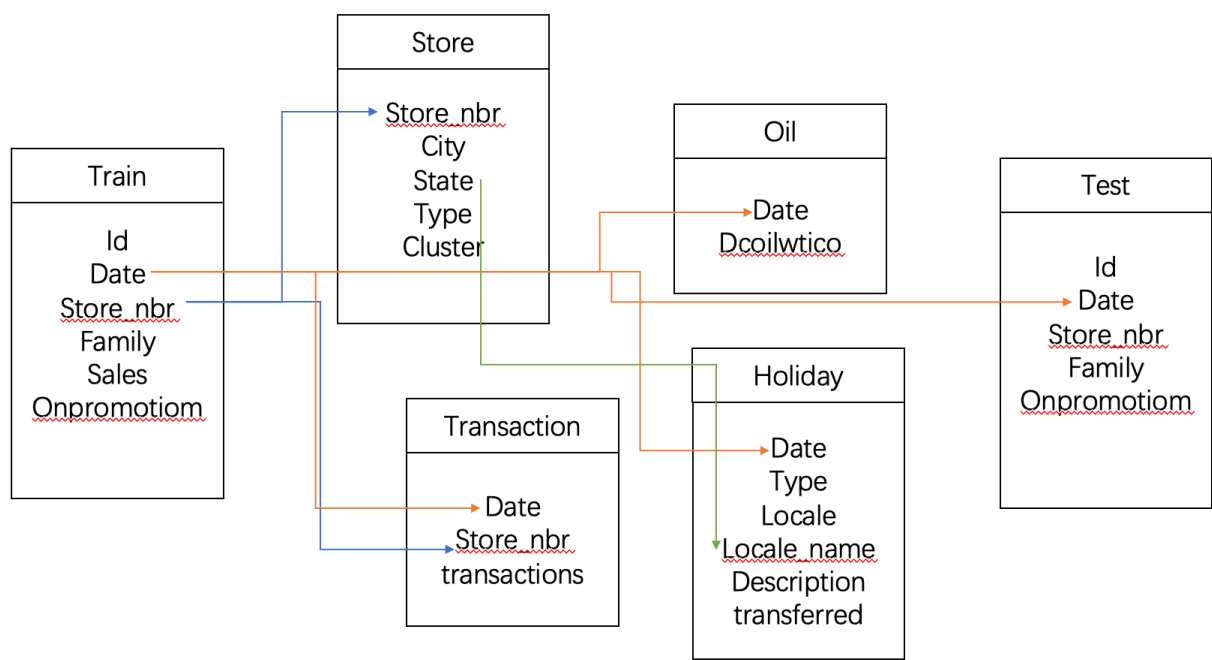


图2.训练集数据表关系图

考虑到训练集中没有的一些特征也可能对销售额产生影响，因此基于现有数据，提取了以下特征：

- 1. 商品平均单价：

$$SalePerTransaction = sales/transaction$$

将销售额（sales）与成交量（transaction）的商作为商品平均单价，每个商店对应一个该特征。

2. 某一周是一年的第几天：从date时间戳中直接读取获得，取值范围[1, 7]。

特征可视化

整合训练集后，获得一个3000888行，15列的数据表。对该数据表绘制协相关矩阵，获得各个变量之间的大致关系。（图3）

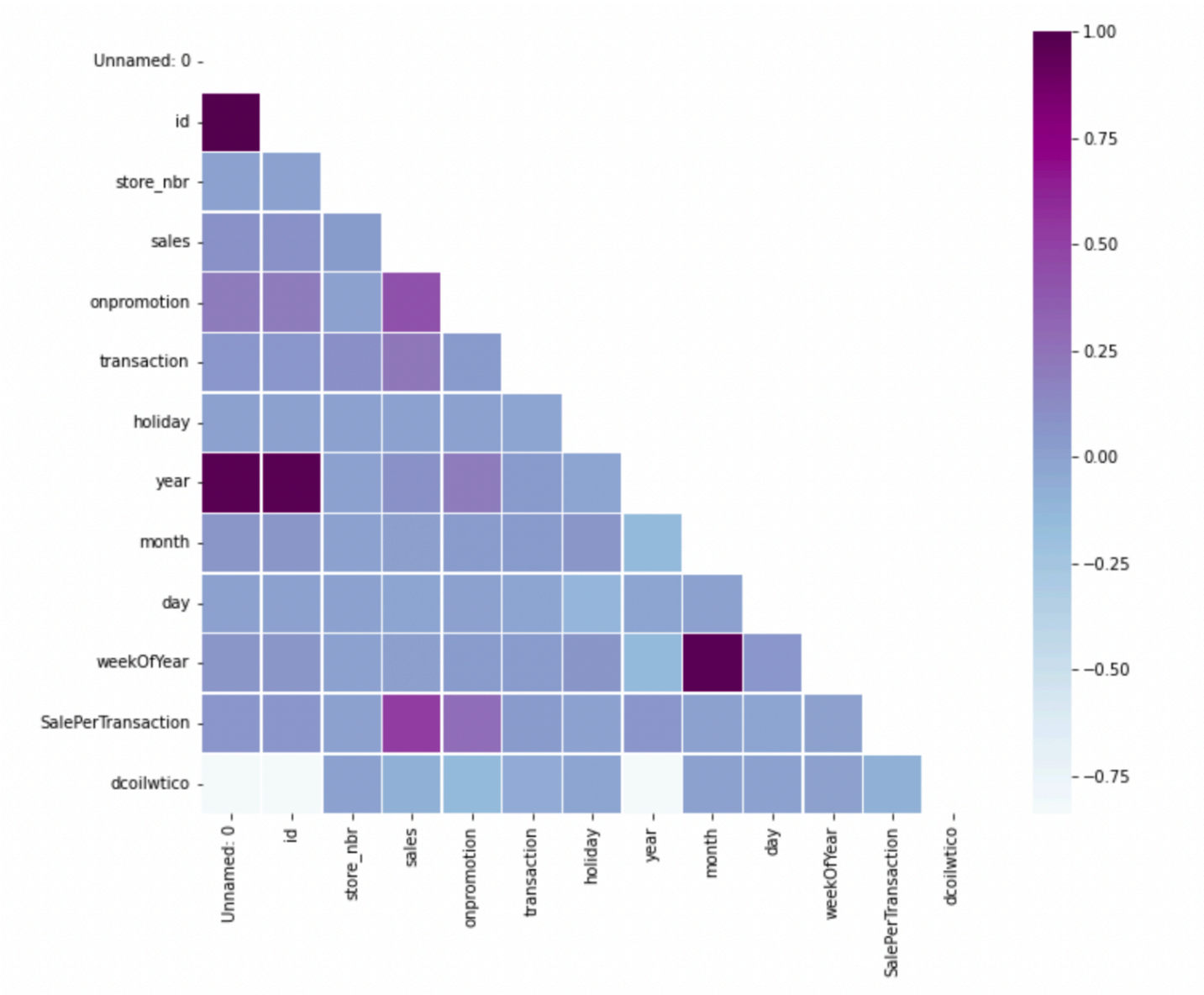


图3. 训练集各变量的协相关关系矩阵

从图中可以大致看出，销量（sales）与促销量（onpromotion）、成交量（transaction）有较强的相关关系，而与油价（dcoiltwico）呈偏蓝色的负相关。

商店特征

分析商店数据表可知，商店类型共有ABCDE五种，分别属于13个地理位置上的集群（cluster）。
对每种类型的商店各取一个作为代表，查看逐年销售额的变化趋势。（如图4～图8）



图4. 类型A的46号商店逐年销售额



图5. 类型B的9号商店逐年销售额



图6. 类型C的13号商店逐年销售额



图7. 类型D的4号商店逐年销售额

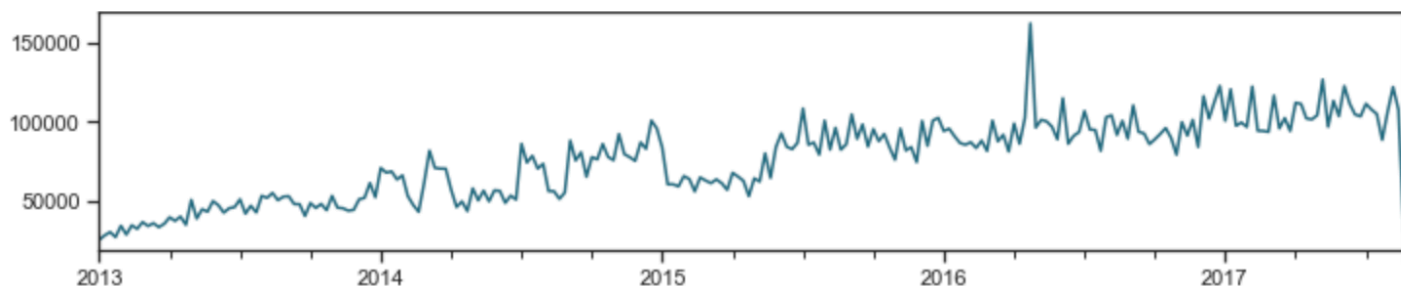
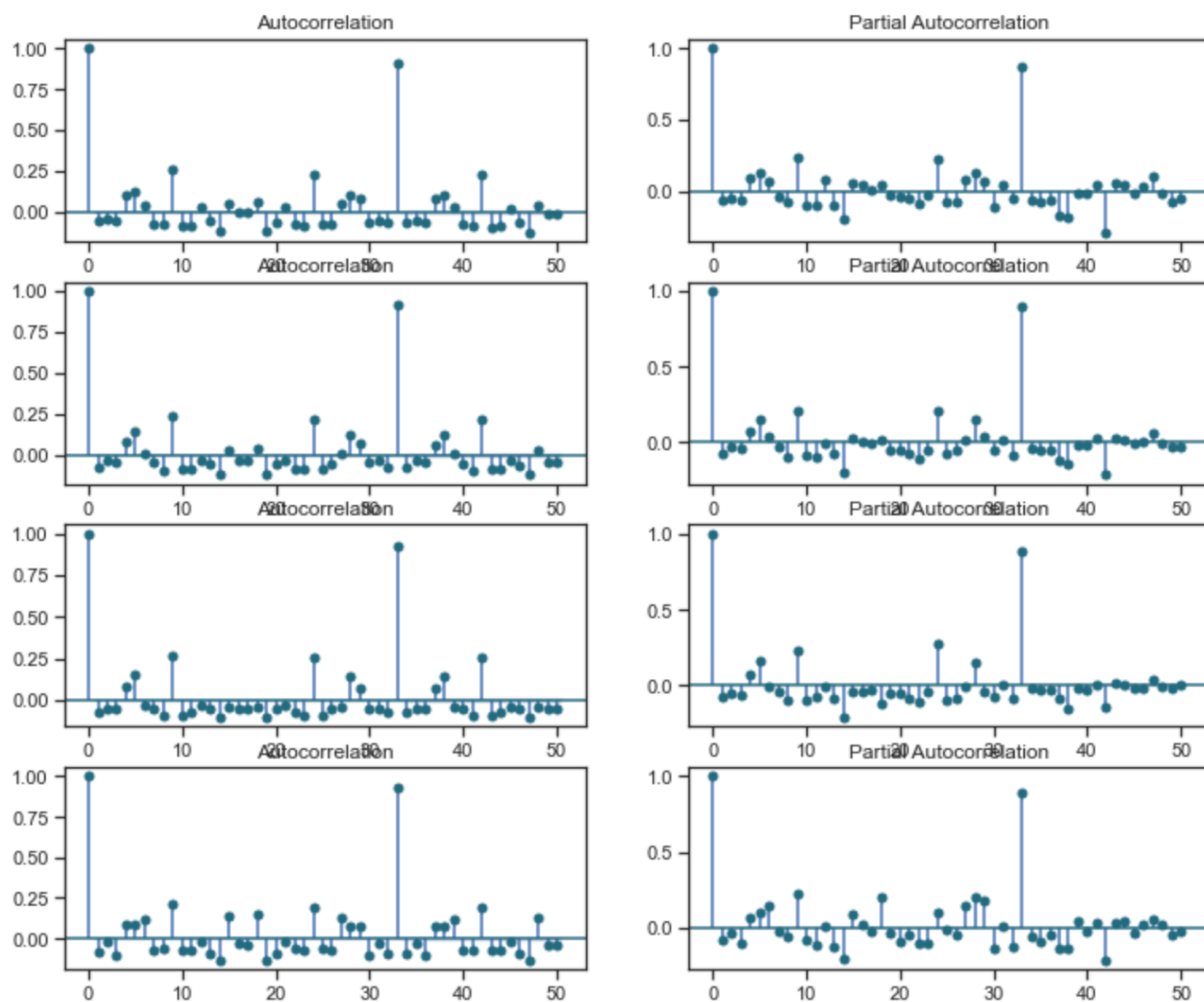


图8. 类型E的28号商店逐年销售额

可以看出五种类型商店基本都服从2013年至2014年缓慢增粘，2014年末至2015年初急剧增长，2015年上下下跌，而后在高位波动的规律。

不同类型商店间，类型E的商店似乎在时间自变量上波动较小。

对上述五种类型代表商店进行自相关函数（ACF）和部分自相关函数（PACF）可视化，详细检查销售量与时间间隔的相关关系。（图9）



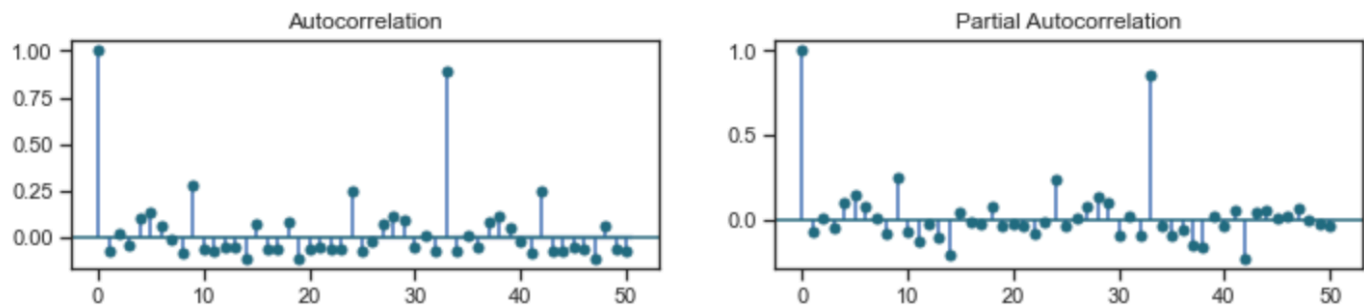


图9. 五类型商店的ACF和RACF图。其中左列分别为五种商店的ACF图，右列为RACF图。

该图中的横坐标代表不同天数间隔下，销量与时间间隔的相关性，纵坐标代表相关系数，即图中的直线越高，代表销量在对应天数下呈周期性。我们可以大致看出销量在每周（7天）、每月（30天）呈周期性，其中月的周期性强于周的周期性。

时序特征

绘制每日销量的变化趋势。（图10）

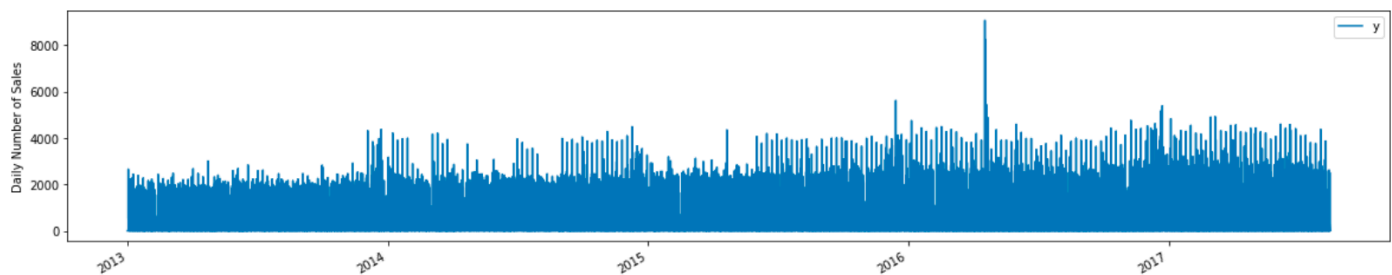


图10. 日销量逐年变化趋势

该趋势图显示销量在2013年至2014年平稳逐渐增加，2014年有小幅度增长，2015年有一段低谷，随后一直在高位波动，其中2016年有一个看起来较异常的高峰。这与之前根据商店销量作出的推测相符。

预测结果分析

将结果输入Prophet后，将得到的结果可视化。

首先对整个训练集和测试集绘制时间序列变化趋势。黑色的点代表训练集中的观察值，深蓝色的线代表平均值，浅蓝色的线代表值可能的上下界。2017年末段没有黑色点的蓝色线段区域，代表Prophet的预测值。（图11）

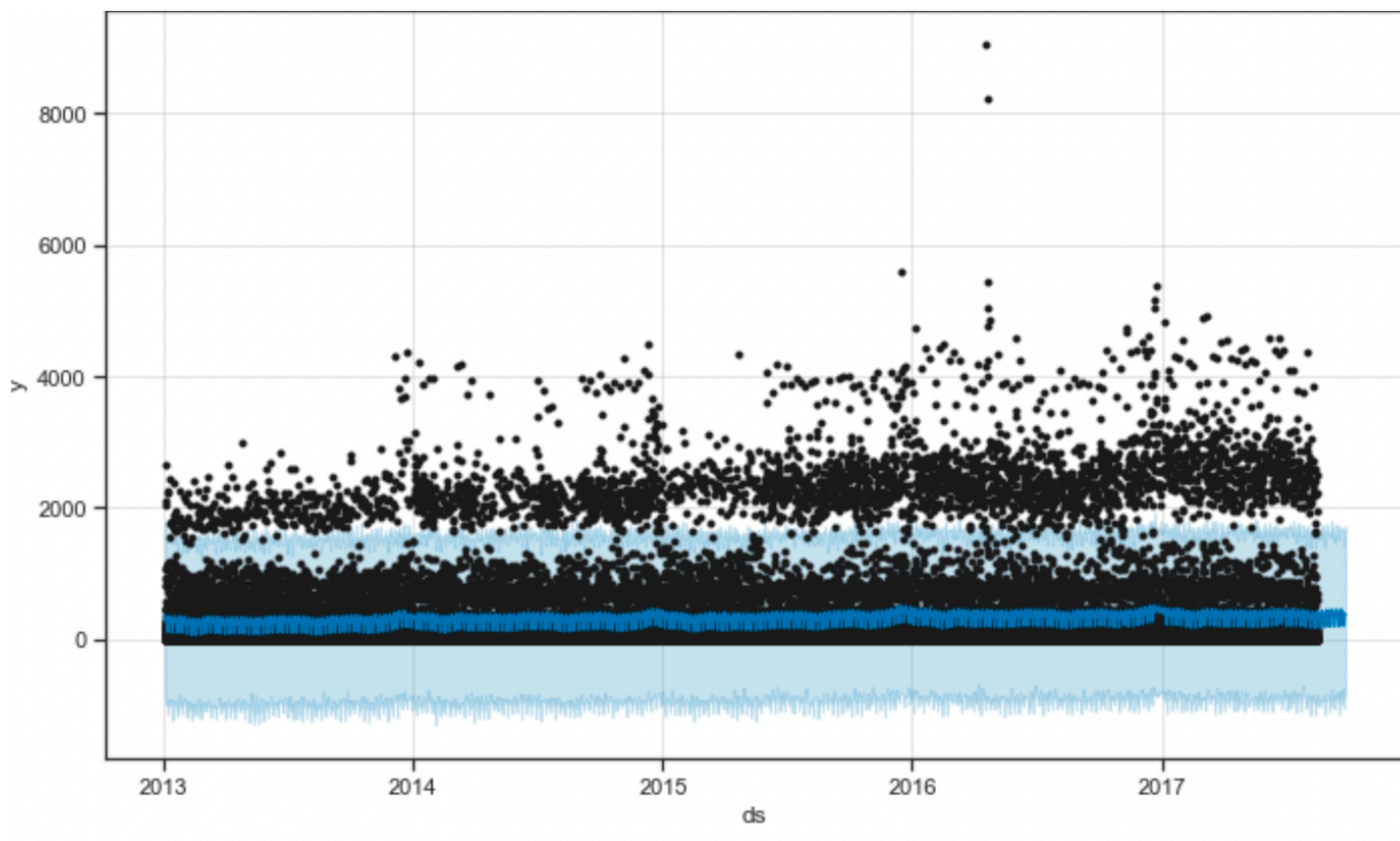


图11. 观测值与预测值的时间序列

接下来对日、星期、年、节假日和平均趋势对销售额的贡献拆解，分别绘制趋势图。（图12~图15）

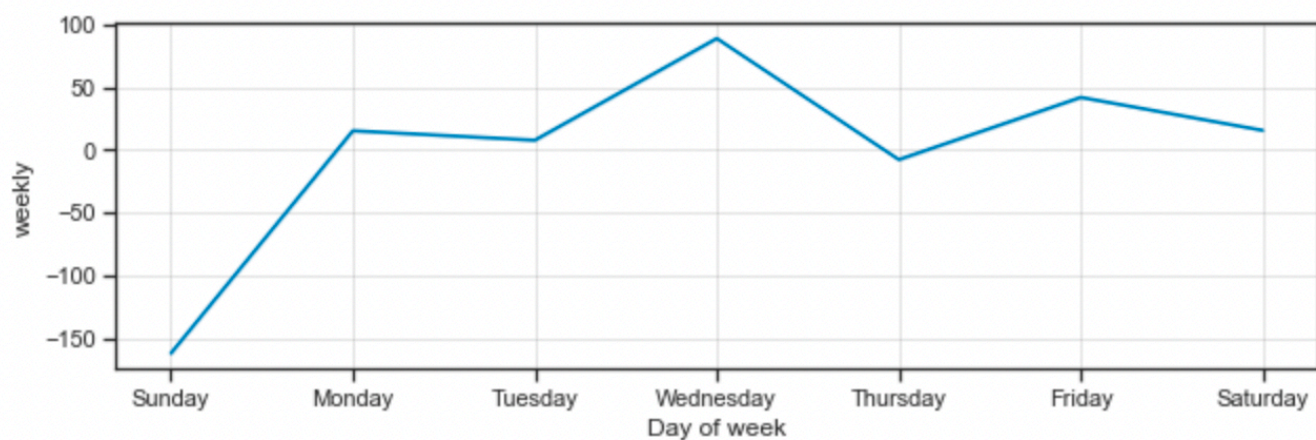


图12. 一周内不同日期对销售额的贡献

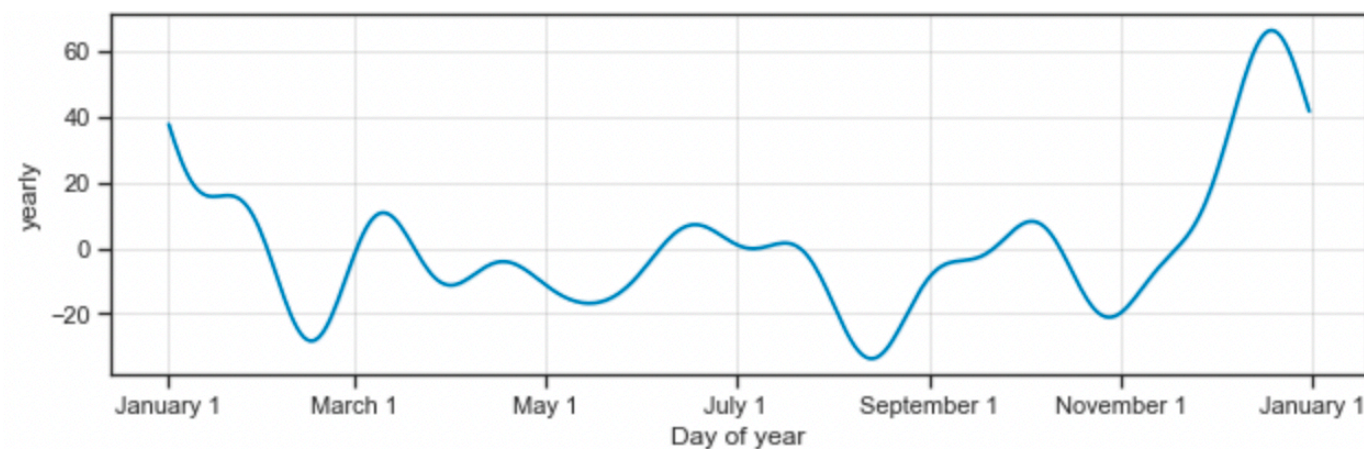


图13. 一年内不同月份对销售额的贡献

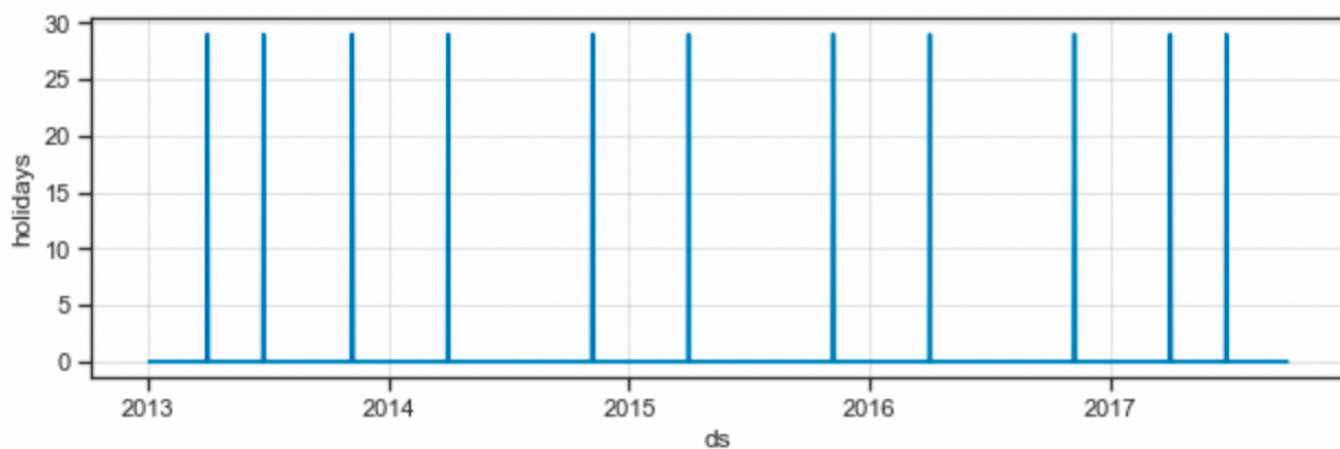


图14. 假期对销售额的贡献

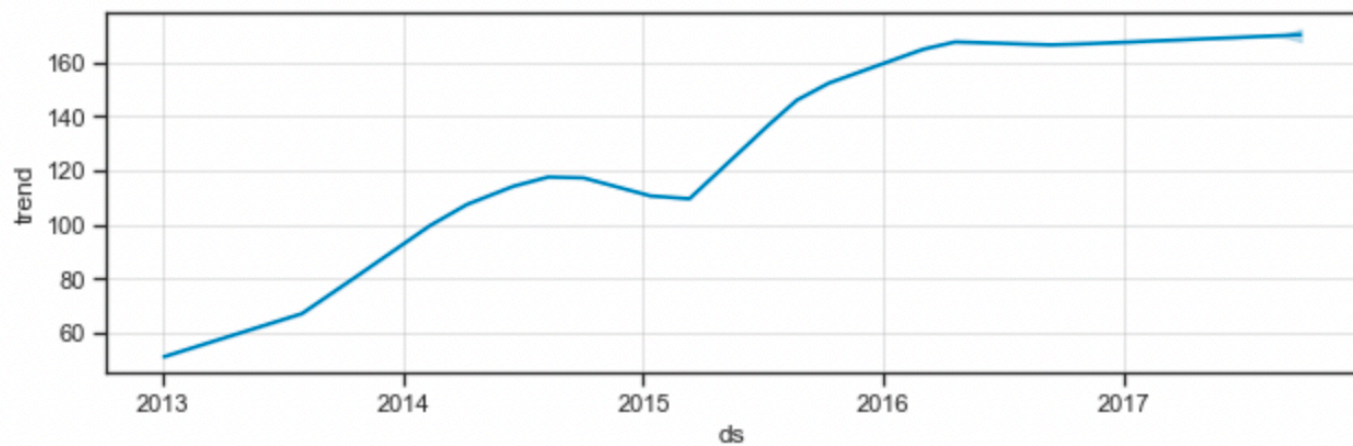


图15. 销售额增长趋势对销售额的贡献

从图中得出，一周中周三销量较高，周日销量为0，与该地商店周日闭店的背景信息相符；一年中，每年年末销量最高，与圣诞节期间销售量较高的背景信息相符，每年约有两次较影响销量的节日；销售量大致呈逐年增长趋势。

模型总结

经过实验对比得出，各个模型的优缺点大致如下。

SMA

优点：简单易做，输入特征极少。

缺点：只将前面一段时间窗口的平均值作为预测值，结果较不准确。

Linear Regression

优点：便于输入多个特征建模。

缺点：自然忽略掉了数据集的时序特征，而由于时序特征几乎是该销量数据集一个最大的影响因素，该缺点导致线性回归模型极不适用于时序预测任务。

Prophet

优点：对数据的时序特征具有较强的建模能力，同时支持较多特征变量输入。

缺点：对除时序之外的其它特征（如商店类型、距离，各个商品之间的影响关系）没有独特的处理方法，需要进行很多手动处理后输入，否则会忽略这些因素的一些因果联系。

附件

- time-series-analysis-and-forecasts-with-prophet.ipynb
项目的主要代码文件，包含数据预处理、数据可视化和搭建Prophet模型部分
- other_models
文件夹，包含前期搭建的其它模型代码
- submission
文件夹，包含三种模型分别生成的预测值表格
- presentation.pptx
展示用的ppt文件