目录

[一、解决方案概述 1](#_Toc477728826)

[1.1赛题介绍与题目分析 1](#_Toc477728827)

[1.2赛题总体思路 1](#_Toc477728828)

[二、数据清洗 1](#_Toc477728829)

[2.1、缺失值的处理 1](#_Toc477728830)

[2.2 异常值清除 2](#_Toc477728831)

[2.3 对天气的处理 3](#_Toc477728832)

[三、特征工程 3](#_Toc477728833)

[3.1 地理位置的处理 3](#_Toc477728834)

[3.2 商家位置特征向量化 4](#_Toc477728835)

[3.3 主要营业时间离散化 5](#_Toc477728836)

[四、特征选择 6](#_Toc477728837)

[五、分解时间序列 6](#_Toc477728838)

[5.1 去周期计算最优 6](#_Toc477728839)

[5.2基于历史记录寻找全局最优 8](#_Toc477728840)

[六、模型融合 8](#_Toc477728841)

# 一、解决方案概述

## 1.1赛题介绍与题目分析

口碑商家客流量预测从2000家口碑商店的历史人数流量记录（2015-06-25到2016-10-31）以及各个商店的信息、天气情况，在此基础上预测未来14天（2016-11-01到2016-11-14）各个商店人数流量。

问题转化为回归问题。评价指标是SMAPE( ),我们可以从商店历史人数流量记录、各个商店的信息、天气构建特征，考虑到评价指标的缘故，我们使用加权融合。

## 1.2赛题总体思路

本文首先从数据清洗开始，介绍的我们对缺失值的处理、对异常值的踢出方法以及对天气的中文相同含义的处理；其次是特征工程，包括对地理位置信息的特征构建、城市特征编码、商店分级类别含义描述编码、shop\_info表提取特征；我们使用xgboost、random forest训练，过程中我们对特征重要性排序，然后进行特征选择。由于时间序列的不稳定性，以及样本少等特点，所以需要考虑一些规则，例如，每个曜日在当前周所占的比例基本稳定等。最后一部分是模型的融合以及相关规则的融合，取得了不错的效果。

# 二、数据清洗

## 2.1、缺失值的处理

在数据收集过程中,由于各种原因可能造成数据集不完整，而该问题将严重影响数据挖掘的质量和结果的稳健性。因此，对缺失值进行补全就显得尤为重要。本次比赛是预测商家在未来14天内各自每天的客户流量，商家历史记录的完整程度会极大的影响模型的预测结果。显然，一个信息完整程度为100%的商家比完整程度为60%的商家更加容易准确的预测。从这一点出发，我们对缺失值进行了分析和处理：

首先，对每一个商家统计其缺失值的个数，进一步得到各个商家的缺失比率。图1显示了含有缺失值的商家及其缺失的天数，其中横轴是商家的ID，纵轴是缺失天数。

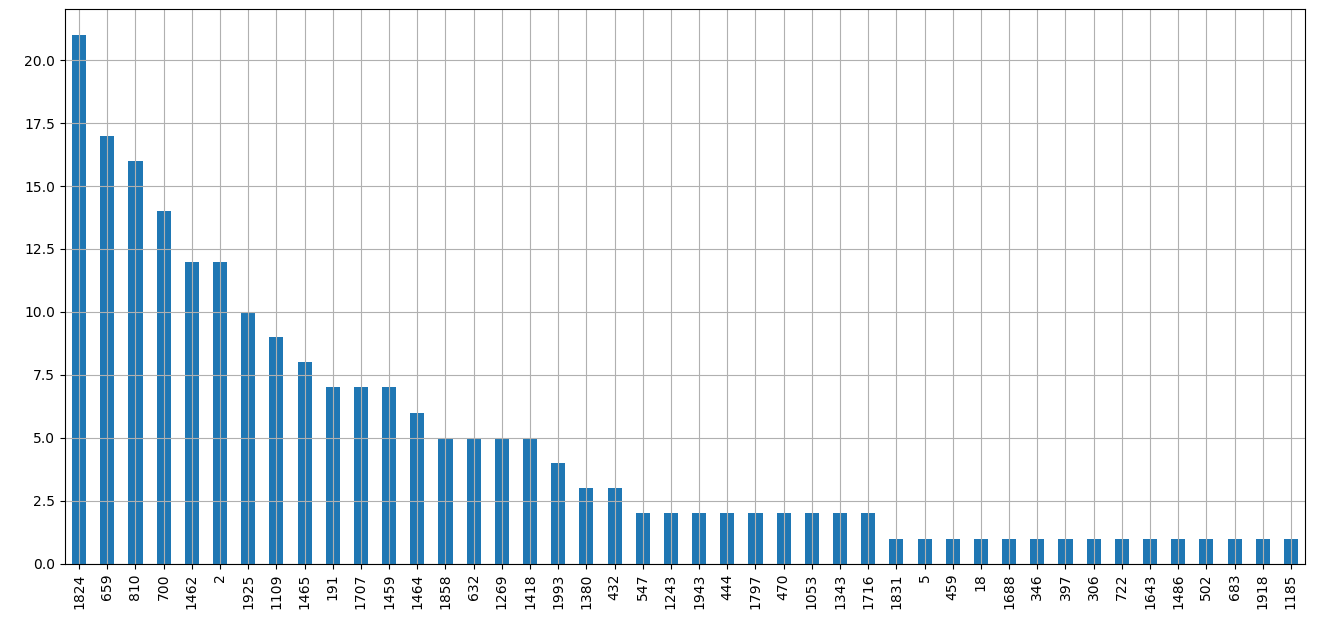


图1. 含有缺失值的商家与其缺失天数

由于十月的最后三个周的参考价值较高，所以我们从2016-10-09开始统计每个商家缺失的总天数(共计23天)。从上图可以看出，shopID\_1824的缺失天数高达21天，即缺失比率高达91%。对于这些缺失值，我们使用对应曜日的均值填充。例如：缺失值当天是周一，则将该店铺所有周一的数据求取均值并填入。

## 2.2 异常值清除

原始数据中每个商家虽然没缺失值，但有些数据是明显异常的。我们使用shopID1650举例，横坐标是日期（2016-09-01至 2016-10-31），纵坐标为每天的客流量，黑色虚线为每日造成消费人数流量，绿色虚线为每日中午消费人数流量，红色虚线为每日晚上消费人数流量。

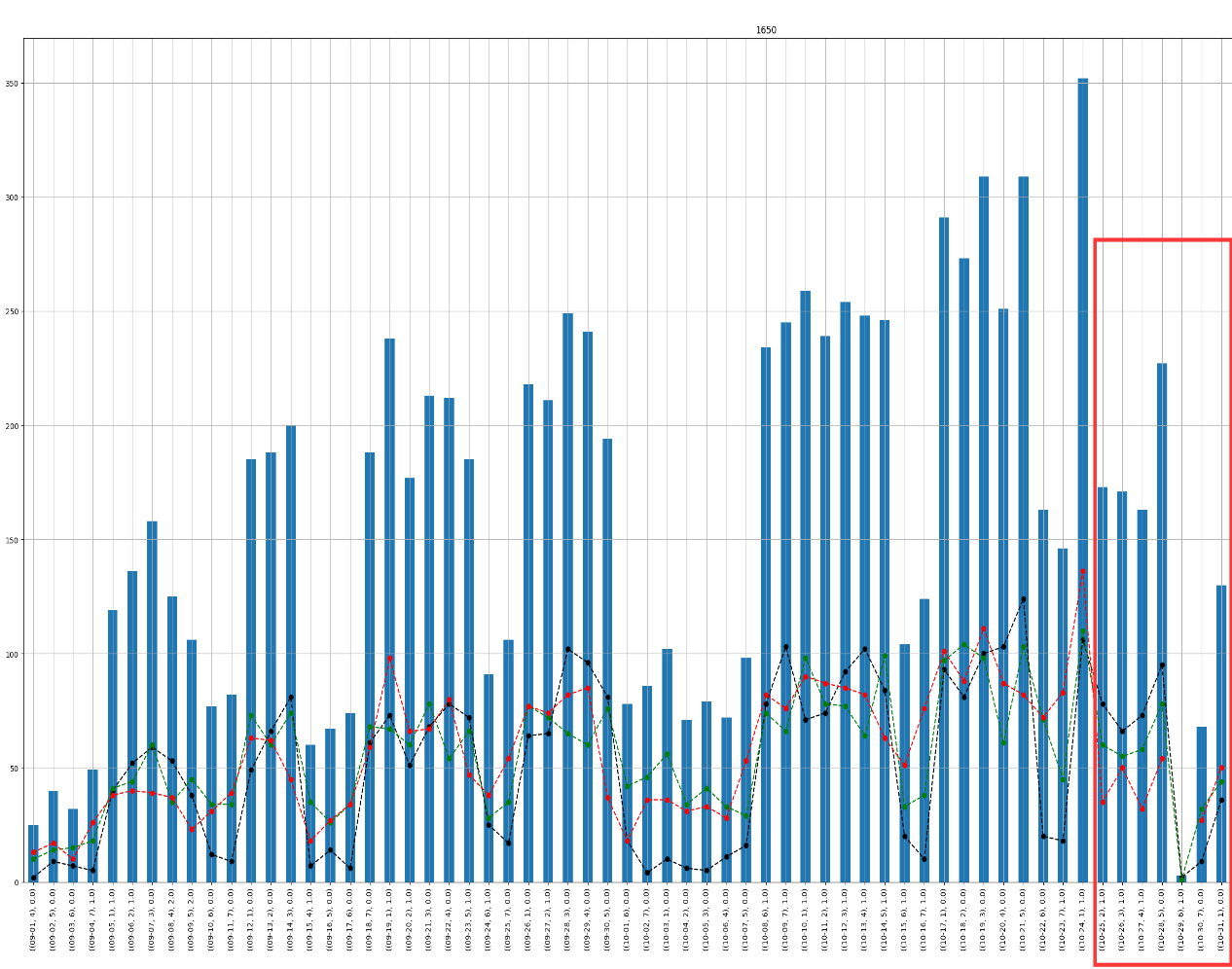


图2. shopID1650的异常数据

从图2可以看出，虽然该商家的客流量没有缺失，但红色方框中最后几天我们可以明显看出异常，故我们直接将这周的数据剔除。

## 2.3 对天气的处理

论坛中提供的Weather表中天气字段，字段都为中文字段，包含了很多含义相同而描述不同的字段。例如：对于“阴转小雨”和“小雨转阴”，由于我们不知道具体下雨的时间以及时长，所以我们认为这是同一种天气。

# 三、特征工程

## 3.1 地理位置的处理

对地理位置信息（类别型变量）最简单的处理方式是独热编码（one-hot），但如果直接使用会得到很高维的稀疏特征，影响模型学习，故我们在独热编码的基础上，做了特征选择。下面介绍具体的方法。

赛题提供中国各省的商家，我们统计了每个商家属于哪个省份。然后对省份分别计数。图3可视化了商家的分布热力图，颜色越深代表数量越多，其中数量最多的几个省份为北京、上海、浙江、广东、江苏、湖北。为了减少特征的维数，我们重点考虑这六个省份，其余的省份则用“其他”字段表示。

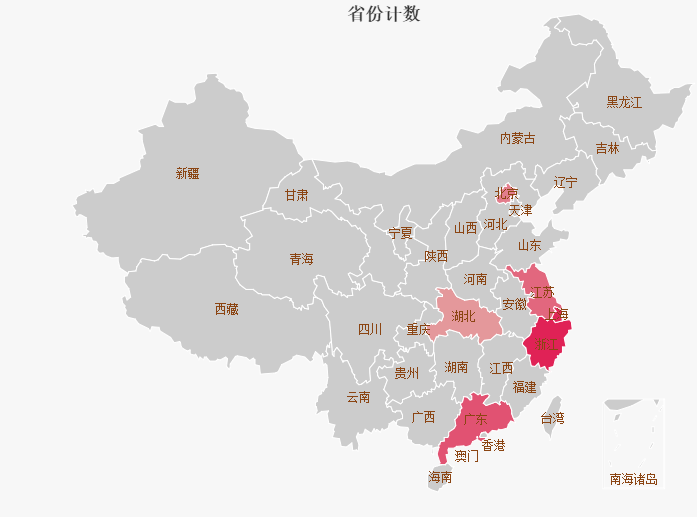


图3. 商家分布热力图

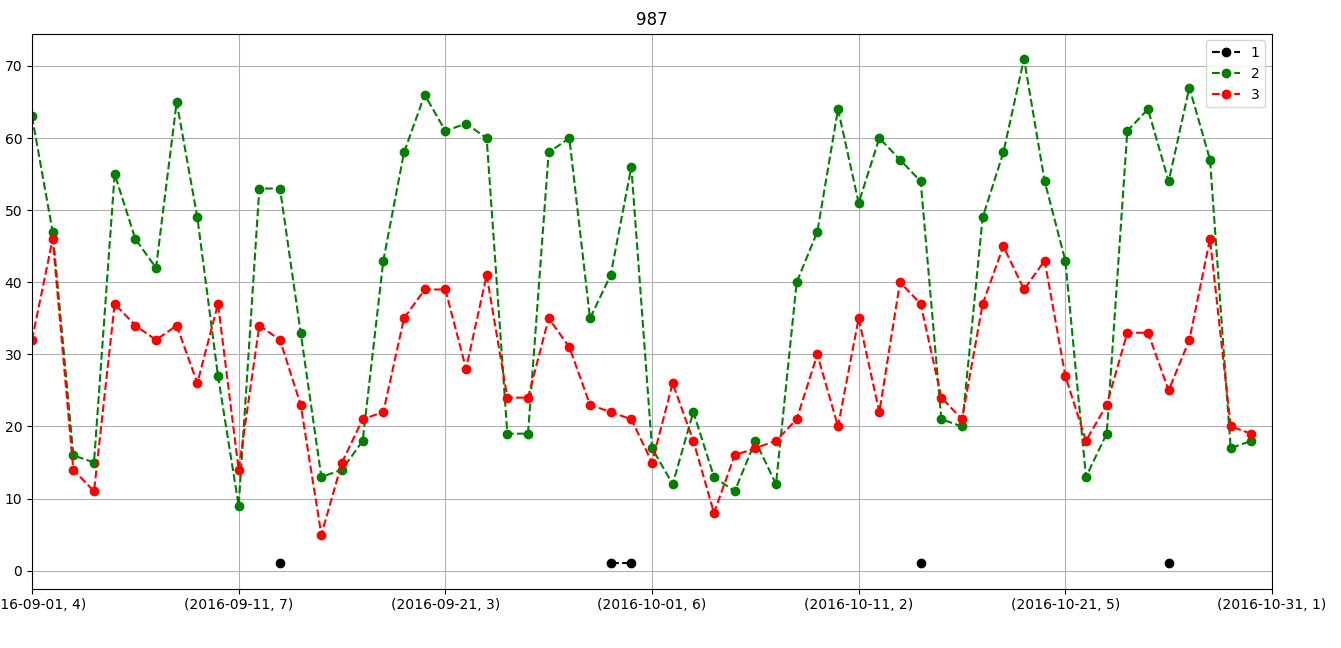
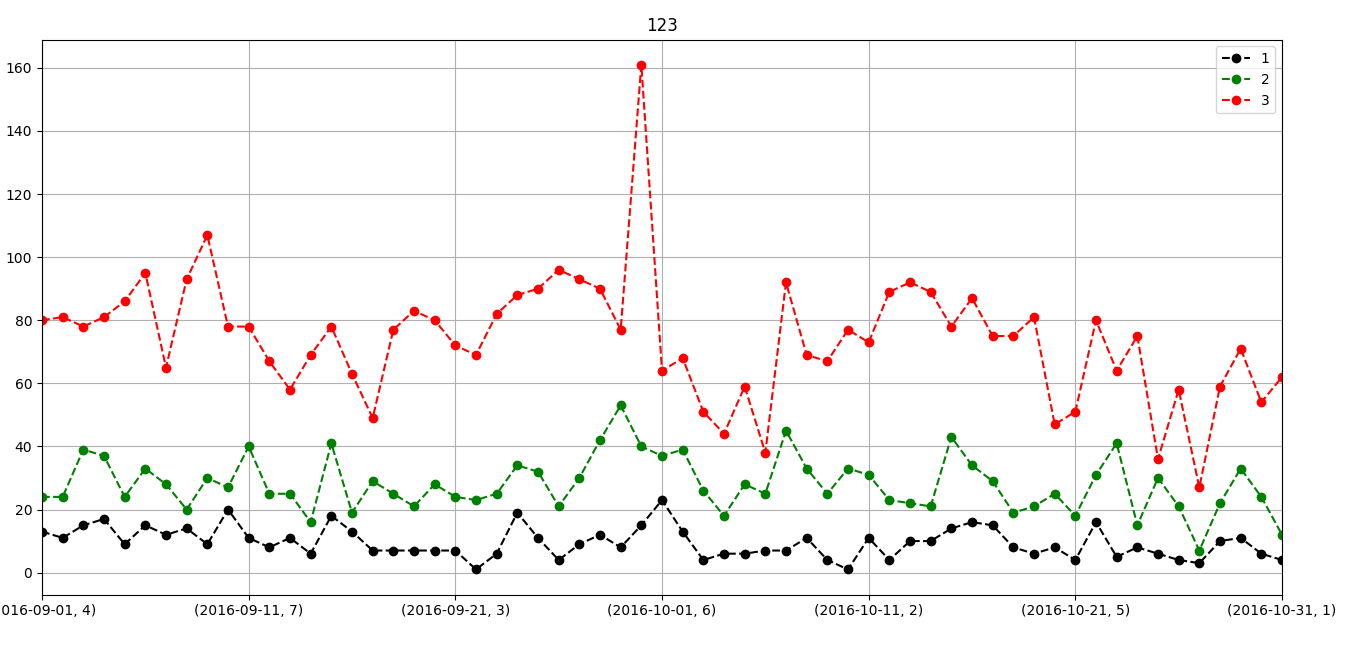
## 3.2 商家位置特征向量化

将城市特征里的城市计数并取log，再等值离散到若干个区间。例如：我们将商家所在城市离散到一个6维向量，向量“100000”表示该商家被判定在第一个区间。我们分别尝试了离散到3-10个区间的情况，通过使用交叉验证的方法，我们发现在离散到6个区间具有更优的结果，如图4所示，横轴表示区间个数，纵轴表示Loss。

图4. 不同的区间对评分的影响

## 3.3 主要营业时间离散化

通过用户的支付时间可以反向估计商家的营业时间，而营业时间又可以根据消费习惯离散化为早上、中午、晚上。图5和图6分别描述了shop\_id为987和321的商家各时段的客流量情况，其中横轴为日期，纵轴为客流量，黑色虚线、绿色虚线、红色虚线为每日早上、中午、晚上消费人数流量。

图5.shop\_id为987的商家各时段客流量图6.shop\_id为123的商家各时段客流量

对比图5和图6可知：不同的商家的主要营业时间可能是不一样的。例如shop\_id为987的店铺主要营业时间是中午，而shop\_id为123的店铺主要营业时间是晚上。所以我们构建一个三维特征。“100”表示主要早上营业，“010”表示主要晚上营业，“001”表示主要晚上营业。

# 四、特征选择

在特征工程部分，我们构建了一系列位置信息相关的特征、主要营业时间特征。所有特征加起来近300维，而样本很少，很容易导致过拟合，需要做降维处理。然而根据以往经验，在数据挖掘类的比赛中，使用降维算法的效果往往不好，如PCA。

除了降维算法之外，我们也可以通过特征选择来降低特征维度。特征选择的方法很多：最大信息系数（MIC）、皮尔逊相关系数、L1正则化、基于模型的特征排序方法等。在本次比赛中，我们选用了比较简单且有效的是最后一种。首先使用训练集训练Xgboost或者Random Forest，在模型训练完成后输出特征重要性，我们只需保留topN的特征即可，从而达到特征选择的目的。

# 五、分解时间序列

## 5.1 去周期计算最优

从宏观的角度来看，口碑是从2015年06月开始上线运行的，所以用户总数是呈现一个总体上升的趋势。而这种趋势，正是解决本赛题的关键，我们姑且推测消费总人数会随着用户总数的增加而增加。图7为所有商家每天的客流量图，其中横轴是日期，纵轴是客流量。可以看到：消费总人数也是呈现总体上升的趋势，这与我们的推测是一致的。即我们可以认为在未来的一段时间内，消费的人数也会呈现出总体上升的趋势。

图7. 全部商家每天的客流量曲线

从微观的角度来看，对于单个商家来说，由于我们的目标是预测每天的客流量，如果其客流量有较强的波动，模型的表现往往不是很好。图8为shop\_id为7的商家的客流量曲线，其中横轴是日期，纵轴是客流量。由图8可以看出该商家的客流量近似以7天作为一个周期，通过多次可视化商家的客流量曲线，我们发现客流量常以7天为周期，因此我们将时间序列分解成每周消费人流量总数和每个曜日在当前周所占的比例，这样就达到了平滑曲线的目的。shop\_id为7的商家平滑后的客流量曲线如图9所示，所示曲线明显比图8所示的曲线平稳，这样更有利于我们预测。

经过这样的平滑，对于目标问题，我们只需要预测一个未来两周的基线，然后再把每个曜日的比率乘以基线，即可得到预测值。

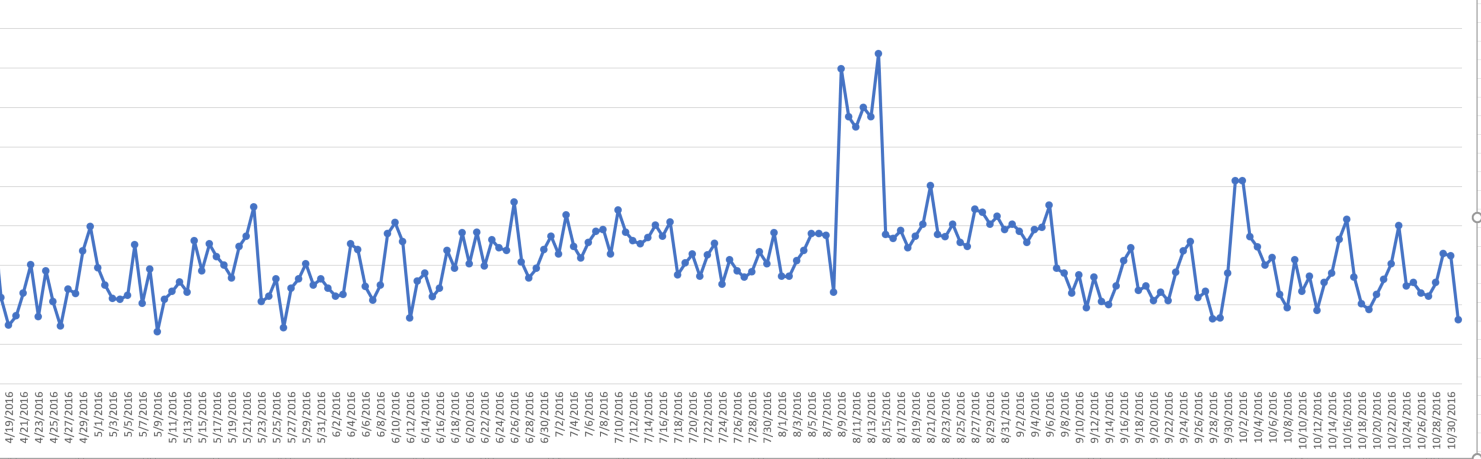


图8. shop\_id为7的商家客流量曲线

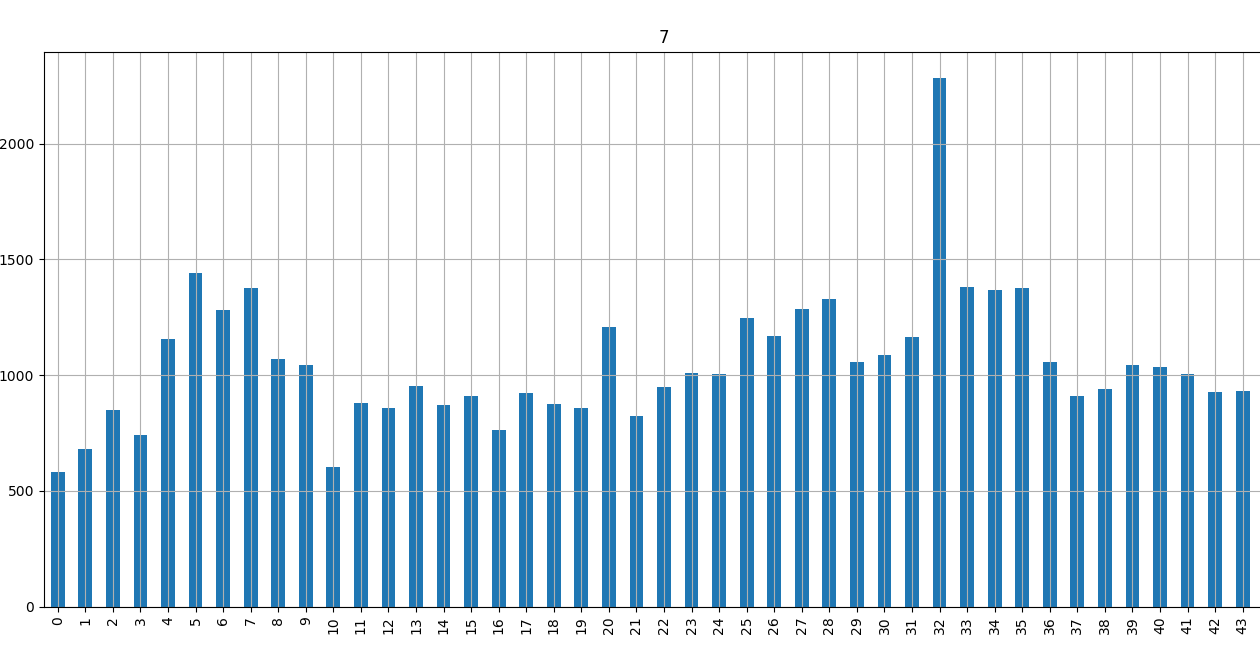


图9. shop\_id为7的商家平滑后的曲线

## 5.2基于历史记录寻找全局最优

赛题提供2015-06-25到2016-10-31的数据，我们可以对每个商家每个曜日寻找一个全局最优。以shop\_id为1的商家周一的数据为例，我们在2015-06-25到2016-10-31的数据中统计出每个周一的值，对于每一个周一的值，使用赛题给出的评价公式去验证其他的周一的值，进而得到全局Loss最小的周一的值。



图10. shop\_id为1的商家每周一的客流量

# 六、模型融合

受bagging思想的启发，我们对单模型xgb做了进一步改进，首先通过调参为xgb寻找一组不错的参数，我们在这些参数在一定的小范围内随机波动，同时对特征进行了随机抽样，训练了多个xgb子模型进行bagging。例如，我们xgb的subsample参数取值5，10，15，20，25这几个取值随机取值，然后subsample参数取值在0.7-0.8之间随机取值。这种方法在参数和特征上都引入了多样性，最后的bagging的效果还是有所提升的。

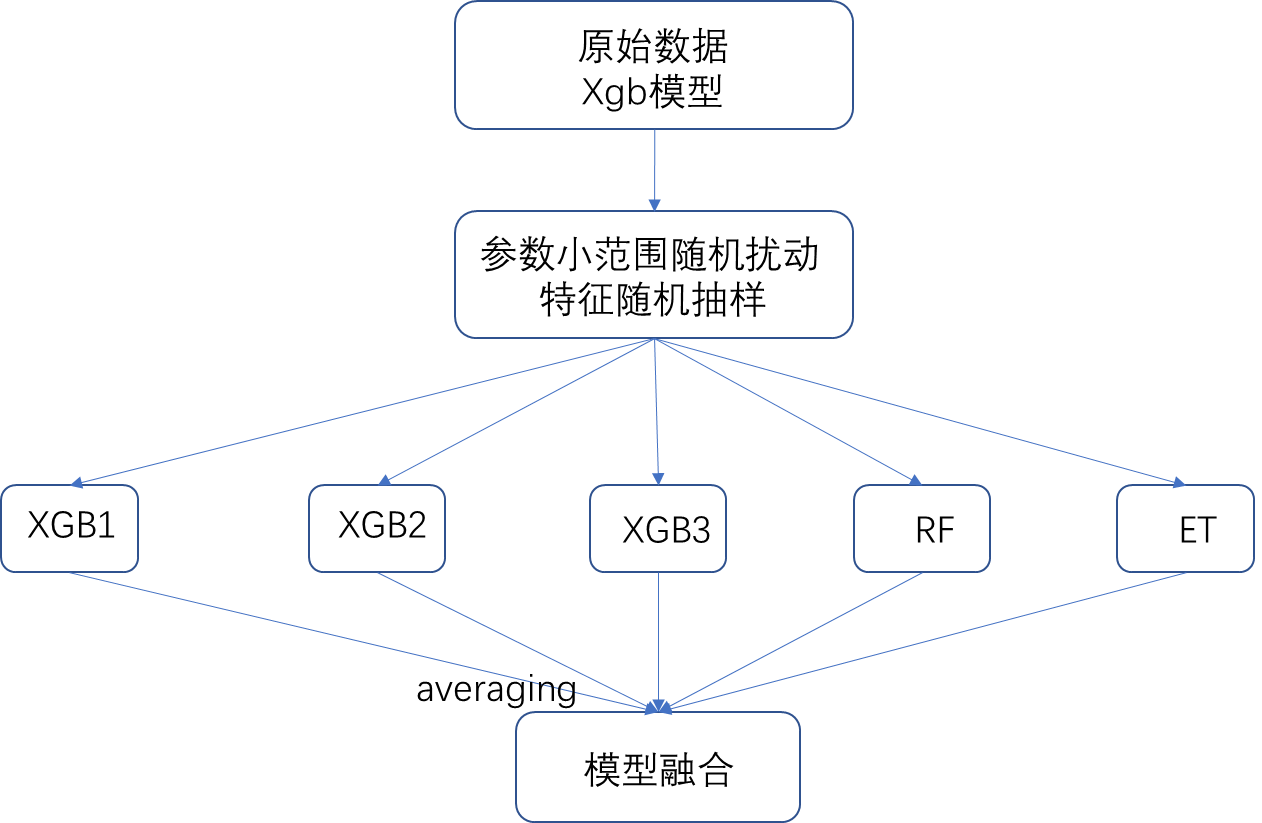


图11.模型融合

模型融合的关键是在于模型的差异性。模型的差异性主要体现在：不同模型、相同模型不同参数、不同训练数据集。因此在融合之前要参考一下各个模型的相关性，可以使用Pearson相关系数评价。