Kaggle Corporación Favorita 商品销量预测比赛解决方案与引申学习

目录

一．赛题与数据介绍

1.1 赛题及背景

零售商有强烈的货物销量预测需求以合理安排进货量，补货频率，仓储空间，从而高效利用仓库，保证货物的新鲜程度，防止易腐烂食品的堆积和浪费。所以厄瓜多尔的 Corporación Favorita 全国连锁超市向 kaggle 提出了这一需求。

具体希望选手从给出的所有数据中提取信息，最终预测出他们 54 个连锁超市中将近 4000 种商品，在 2017 年 8 月 16 日-31 日中每天的销量具体数值。

这里非按“件”计数的商品为非整数销量，无需考虑计量单位。

1.2 数据

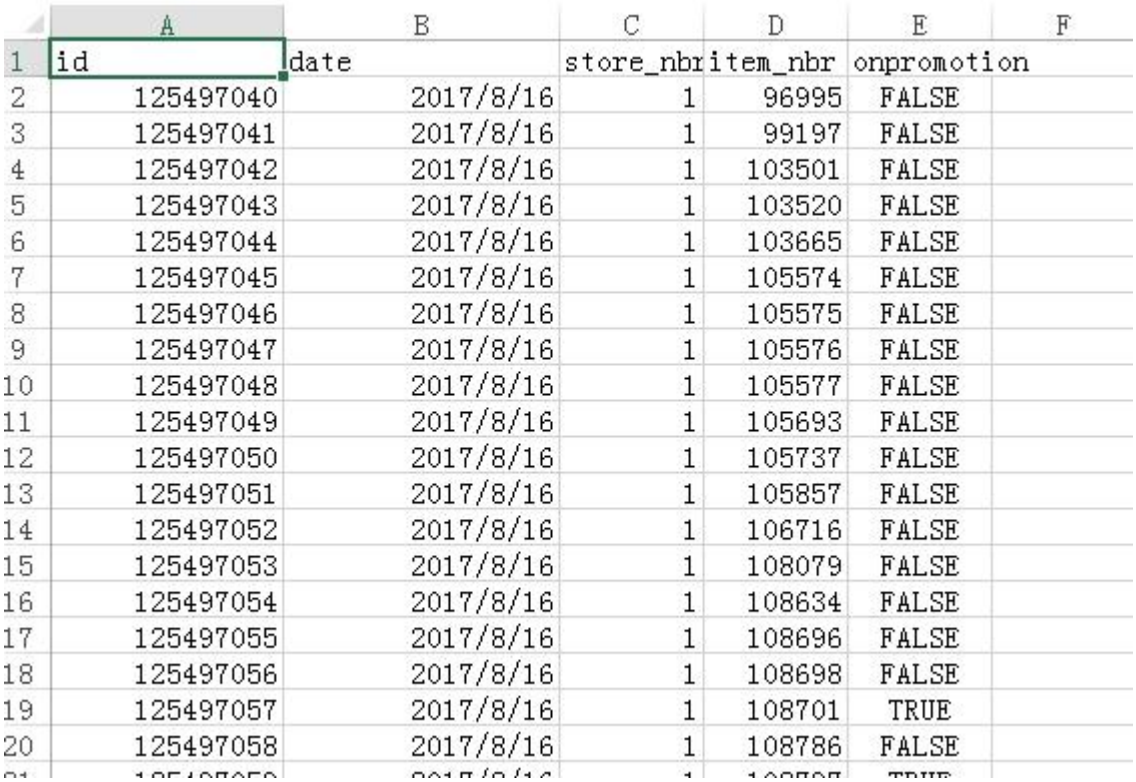
共有 7 张原始数据表，分别为：

（1） train.csv

	A	B	C	D	E	F	G	H
1	id	date	store_nbr	item_nbr	unit_sale	onpromotion		
2	0	2013/1/1	25	103665	7			
3	1	2013/1/1	25	105574	1			
4	2	2013/1/1	25	105575	2			
5	3	2013/1/1	25	108079	1			
6	4	2013/1/1	25	108701	1			
7	5	2013/1/1	25	108786	3			
8	6	2013/1/1	25	108797	1			
9	7	2013/1/1	25	108952	1			
10	8	2013/1/1	25	111397	13			
11	9	2013/1/1	25	114790	3			
12	10	2013/1/1	25	114800	1			
13	11	2013/1/1	25	115267	1			
14	12	2013/1/1	25	115611	1			
15	13	2013/1/1	25	115693	1			
16	14	2013/1/1	25	115720	5			
17	15	2013/1/1	25	115850	1			
18	16	2013/1/1	25	115891	6			
19	17	2013/1/1	25	115892	10			
20	18	2013/1/1	25	115894	5			
21	19	2013/1/1	25	119024	1			

该表以流水的方式记录了从 2013.1.1-2017.8.15，每个超市每种商品的每日销量

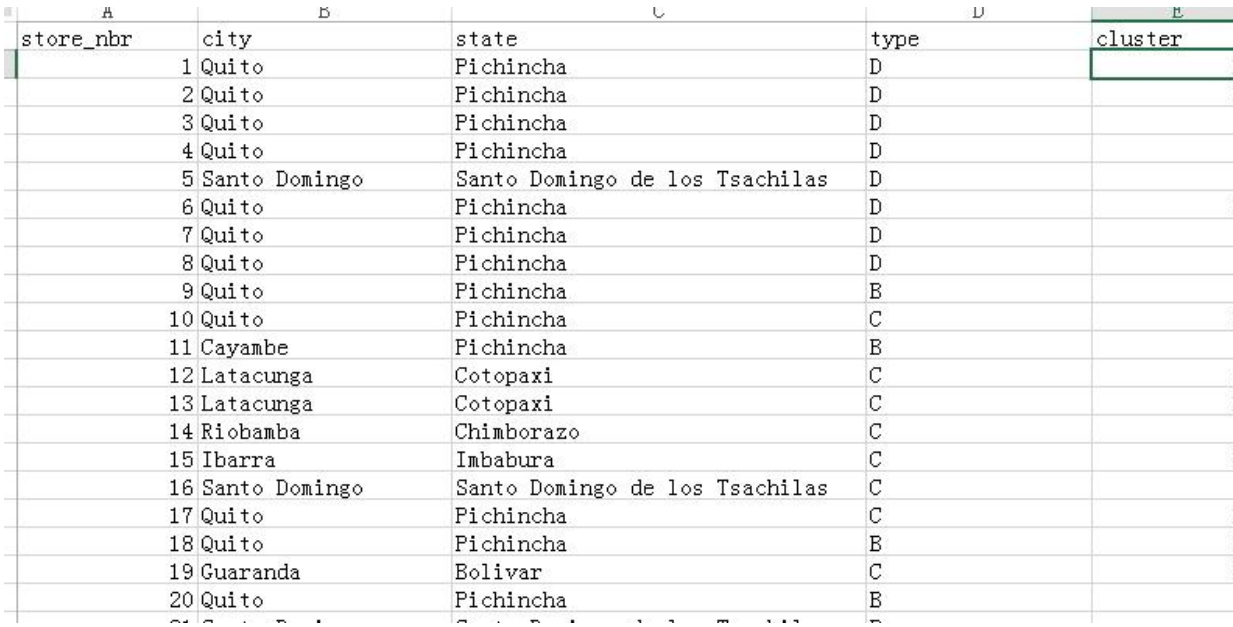
（2） test.csv



	A	B	C	D	E	F
1	id	date	store_nbr	item_nbr	onpromotion	
2	125497040	2017/8/16	1	96995	FALSE	
3	125497041	2017/8/16	1	99197	FALSE	
4	125497042	2017/8/16	1	103501	FALSE	
5	125497043	2017/8/16	1	103520	FALSE	
6	125497044	2017/8/16	1	103665	FALSE	
7	125497045	2017/8/16	1	105574	FALSE	
8	125497046	2017/8/16	1	105575	FALSE	
9	125497047	2017/8/16	1	105576	FALSE	
10	125497048	2017/8/16	1	105577	FALSE	
11	125497049	2017/8/16	1	105693	FALSE	
12	125497050	2017/8/16	1	105737	FALSE	
13	125497051	2017/8/16	1	105857	FALSE	
14	125497052	2017/8/16	1	106716	FALSE	
15	125497053	2017/8/16	1	108079	FALSE	
16	125497054	2017/8/16	1	108634	FALSE	
17	125497055	2017/8/16	1	108696	FALSE	
18	125497056	2017/8/16	1	108698	FALSE	
19	125497057	2017/8/16	1	108701	TRUE	
20	125497058	2017/8/16	1	108786	FALSE	

该表以相似的方式记录了“所有需要给出预测的【商品-商店】条目”和他们是否在2017.8.15-2017.8.31 进行促销活动的信息

（3） store.csv



	A	B	C	D	E
	store_nbr	city	state	type	cluster
	1	Quito	Pichincha	D	
	2	Quito	Pichincha	D	
	3	Quito	Pichincha	D	
	4	Quito	Pichincha	D	
	5	Santo Domingo	Santo Domingo de los Tsachilas	D	
	6	Quito	Pichincha	D	
	7	Quito	Pichincha	D	
	8	Quito	Pichincha	D	
	9	Quito	Pichincha	B	
	10	Quito	Pichincha	C	
	11	Cayambe	Pichincha	B	
	12	Latacunga	Cotopaxi	C	
	13	Latacunga	Cotopaxi	C	
	14	Riobamba	Chimborazo	C	
	15	Ibarra	Imbabura	C	
	16	Santo Domingo	Santo Domingo de los Tsachilas	C	
	17	Quito	Pichincha	C	
	18	Quito	Pichincha	B	
	19	Guaranda	Bolivar	C	
	20	Quito	Pichincha	B	

该表记录了每个商店的城市，州，类型，其中后两个类型属于匿名信息

（4） transaction.csv

ate	store_nbr	transactions
2013/1/1	25	770
2013/1/2	1	2111
2013/1/2	2	2358
2013/1/2	3	3487
2013/1/2	4	1922
2013/1/2	5	1903
2013/1/2	6	2143
2013/1/2	7	1874
2013/1/2	8	3250
2013/1/2	9	2940
2013/1/2	10	1293
2013/1/2	11	3547
2013/1/2	12	1362
2013/1/2	13	1102
2013/1/2	14	2002
2013/1/2	15	1622
2013/1/2	16	1167
2013/1/2	17	1580
2013/1/2	18	1635
2013/1/2	19	1369

该表记录了每个超市的总日销

（5） holidays_events.csv

	A	B	C	D	E	F
	date	type	locale	locale_name	description	transf
	2012/3/2	Holiday	Local	Manta	Fundacion de Manta	FALSE
	2012/4/1	Holiday	Regional	Cotopaxi	Provincializacion de Cotopaxi	FALSE
	2012/4/12	Holiday	Local	Cuenca	Fundacion de Cuenca	FALSE
	2012/4/14	Holiday	Local	Libertad	Cantonizacion de Libertad	FALSE
	2012/4/21	Holiday	Local	Riobamba	Cantonizacion de Riobamba	FALSE
	2012/5/12	Holiday	Local	Puyo	Cantonizacion del Puyo	FALSE
	2012/6/23	Holiday	Local	Guaranda	Cantonizacion de Guaranda	FALSE
	2012/6/25	Holiday	Regional	Imbabura	Provincializacion de Imbabura	FALSE
	2012/6/25	Holiday	Local	Latacunga	Cantonizacion de Latacunga	FALSE
	2012/6/25	Holiday	Local	Machala	Fundacion de Machala	FALSE
	2012/7/3	Holiday	Local	Santo Domingo	Fundacion de Santo Domingo	FALSE
	2012/7/3	Holiday	Local	El Carmen	Cantonizacion de El Carmen	FALSE
	2012/7/23	Holiday	Local	Cayambe	Cantonizacion de Cayambe	FALSE
	2012/8/5	Holiday	Local	Esmeraldas	Fundacion de Esmeraldas	FALSE
	2012/8/10	Holiday	National	Ecuador	Primer Grito de Independencia	FALSE
	2012/8/15	Holiday	Local	Riobamba	Fundacion de Riobamba	FALSE
	2012/8/24	Holiday	Local	Ambato	Fundacion de Ambato	FALSE
	2012/9/28	Holiday	Local	Ibarra	Fundacion de Ibarra	FALSE
	2012/10/7	Holiday	Local	Quevedo	Cantonizacion de Quevedo	FALSE
	2012/10/9	Holiday	National	Ecuador	Independencia de Guayaquil	TRUE
	2012/10/12	Transfer	National	Ecuador	Traslado Independencia de Guayaquil	FALSE
	2012/11/2	Holiday	National	Ecuador	Dia de Difuntos	FALSE
	2012/11/3	Holiday	National	Ecuador	Independencia de Cuenca	FALSE

该表记录了厄瓜多尔的重要地方节日和全球性节日

（6） items.csv

item_nbr	family	class	perishable
96995	GROCERY I	1093	0
99197	GROCERY I	1067	0
103501	CLEANING	3008	0
103520	GROCERY I	1028	0
103665	BREAD/BAKERY	2712	1
105574	GROCERY I	1045	0
105575	GROCERY I	1045	0
105576	GROCERY I	1045	0
105577	GROCERY I	1045	0
105693	GROCERY I	1034	0
105737	GROCERY I	1044	0
105857	GROCERY I	1092	0
106716	GROCERY I	1032	0
108079	GROCERY I	1030	0
108634	GROCERY I	1075	0
108696	DELI	2636	1

该表记录了商品的所属小类和大类以及是否容易腐烂

（7） oil.csv

	dcoilwtico
2013/1/1	
2013/1/2	93.14
2013/1/3	92.97
2013/1/4	93.12
2013/1/7	93.2
2013/1/8	93.21
2013/1/9	93.08
2013/1/10	93.81
2013/1/11	93.6
2013/1/14	94.27
2013/1/15	93.26
2013/1/16	94.28
2013/1/17	95.49
2013/1/18	95.61
2013/1/21	
2013/1/22	96.09
2013/1/23	95.06
2013/1/24	95.35
2013/1/25	95.15

该表记录了历史原油价

二．我的解决方案概述

2.1 时间序列问题的整体解决方案

时间序列的预测问题最常用的两种方法是 ARIMA 时间序列模型和机器学习模型。

1.ARIMA 模型：

1. 在建模之前需要进行随机性-白噪声检验，判断序列是否为无信息序列
2. 判定为非白噪声序列后，进行平稳性检验，平稳与非平稳序列分别选择模型

3. 平稳序列均值方差是常数，非平稳序列有周期趋势的序列，差分后为平稳序列

可以看出，ARIMA 模型适用于具有较为规律周期性变化的时序预测问题，而在此问题当中，商品的销量不具备此种规律性，许多信息会被作为噪声剔除，因此不适于求解

2.机器学习模型：

机器学习模型对时间序列问题的求解逻辑是：

将 与目标预测区间性质相似的历史时间窗口 提取出来作为学习样本，通过构造关于历史销量，样本性质等方面的特征来描述样本。

实践中的预测表现比 ARIMA 好，具体的机器学习算法包括但不限于 LR,RF,GDBT,SVM

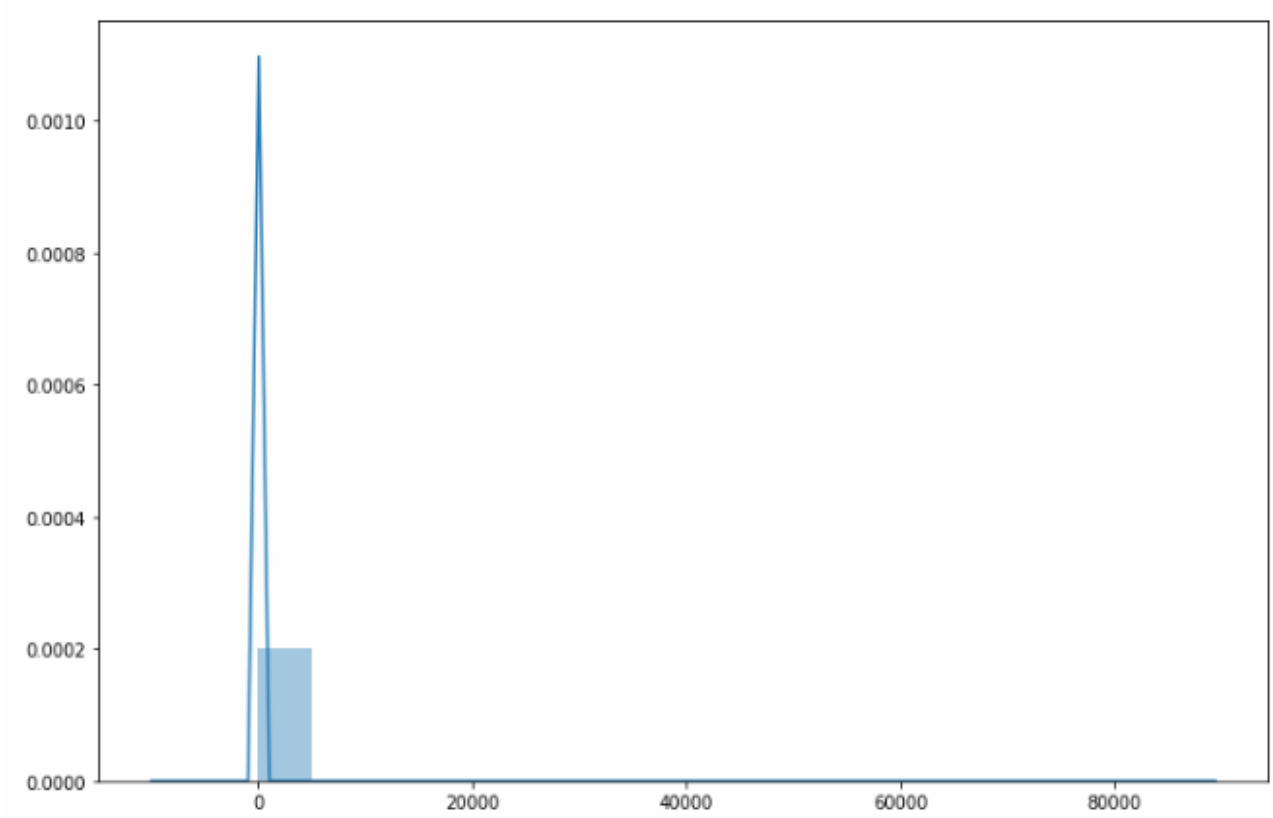
2.2 我的解决方案

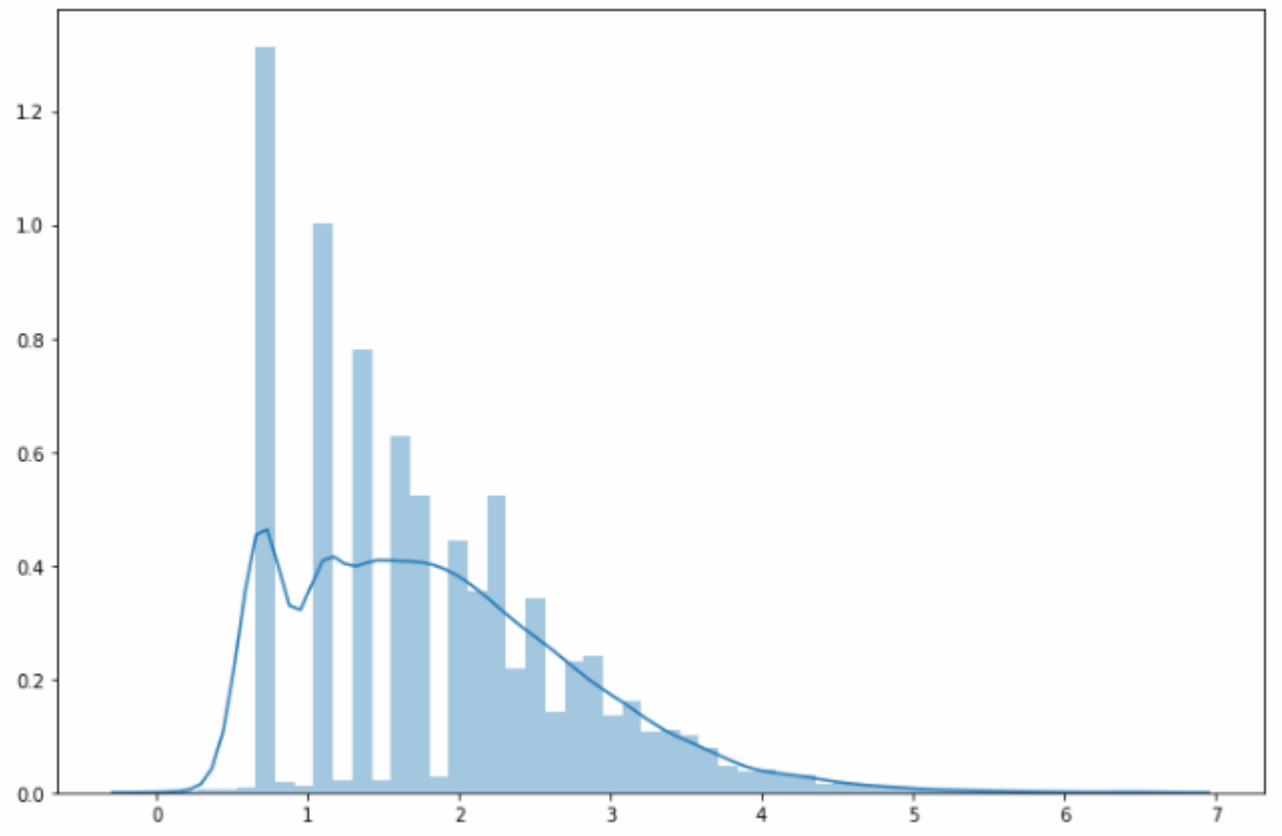
对不同的历史时间窗口构造均值，最大最小，偏度等统计量特征，历史窗每周特定日期的平均销量特征，关于产品和商店的类别型特征，使用 GBDT 模型，单独预测目标区间的每日销量并合并。

在这里我并没有使用节假日，原油价格等数据，同时我选择的特征提取时间窗口全部在 2017 年之后，因为我认为半年的整体经济周期比较平稳，在半年之内寻找的学习样本是那个“和目标预测区间性质更为相似”的合理样本提取集。

三．数据探索

3.1 商品销量整体分布





将所有不同类型的商品的日销量铺平发现整体分布十分极端，绝大部分分布在 0-10 之间，而有极其长尾的高销量商品（上图），经过 log 变化之后，将乘法量级的差异转换为加法量级的差异（下图），将预测目标从预测销量转换为预测销量的对数可以使得模型在学习的过程中更容易照顾到长尾数据，若直接使用销量作为预测目标，长尾数据会极其容易被对待为离群异常数据，在优化过程中丧失其信息

3.2 目标预测商品样本

经过仔细查证 test.csv，我发现 test.csv 中需要被我们预测的商品是**某 3901 种商品和 54 个超市的笛卡尔积，**而通过对 train 中商品在不同时间窗口的存在与否我发现，这 3901 种商品和这 54 个超市组成的 21 万条待预测目标，它们有部分从未出现在历史上，也就是说现实中并没有这种搭配，或者属于上架的新商品。21 万种组合构成如下：

1. 其中 5 万种没有在 2017 年时段出现
2. 大约 1 万条仅在 4.1 日以前出现过之后就再也没有出现

以越是靠近目标区间的时间段的出现情况筛选，就有越多的组合被筛选出去。

从这个分析中我认为，商品的最早出现时间和最迟出现时间距离预测区间的日期数应该是一个较具有区分度的样本特征，我也在后面的特征工程中构造了这一特征

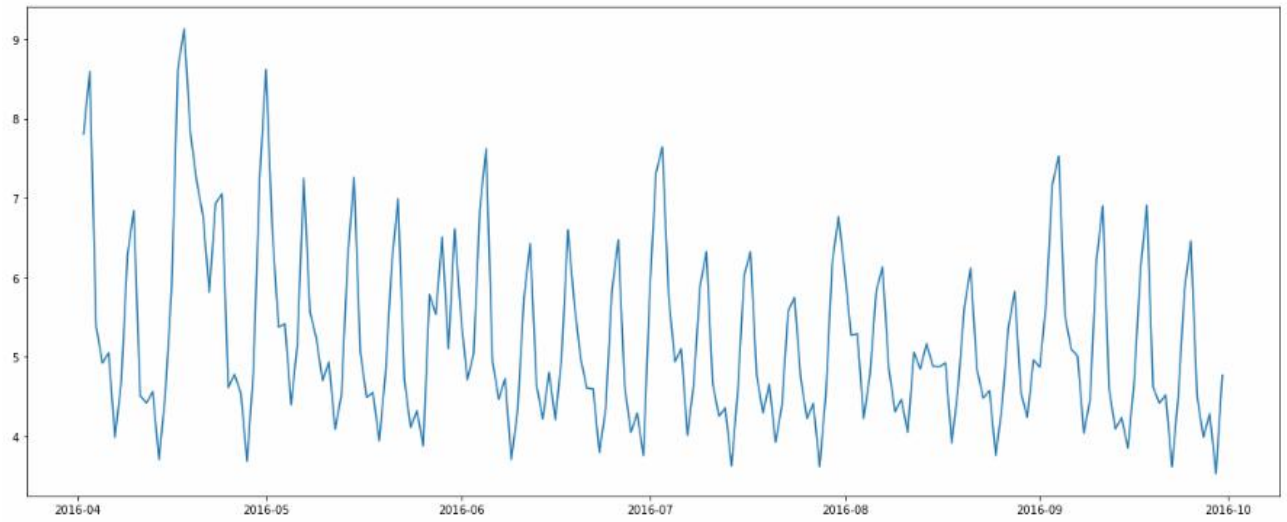
另一方面，未出现过的组合在 test 表中有促销信息存在，经过统计，5 万条 2017 年未出现过的组合，在 16 天的预测区间中，至少出现过一次促销记录的组合，有不到 1000 种，应作为上新商品对待，

而这部分的内容该如何预测我当时并没有处理，在比赛后了解到其他人的处理方式是这样的：

1. 寻找和此类条目具有相同商品大类和商品小类，且上新时有促销的商品集
2. 找到商品集中每个商品最初出现促销的日期，并选择其与待预测商品在目标区间开始打折的日期在 16 天内相对位置一样的一段 16 日历史价格序列
3. 对这些取平均，作为对于上新商品销量的预估

这个方法通过简单的规则，假设新商品大概销量处于同类商品上新时的平均水平，学到了。

3.3 总体销量的时序形态

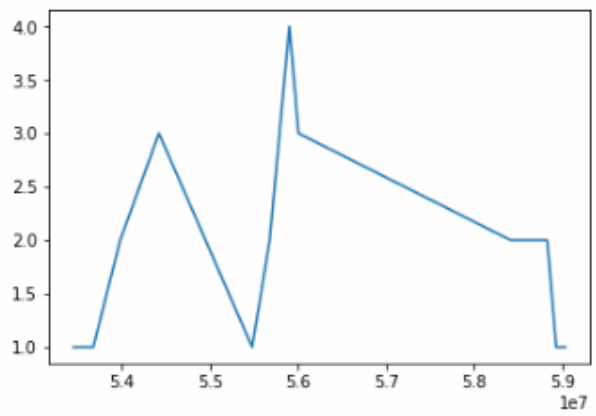


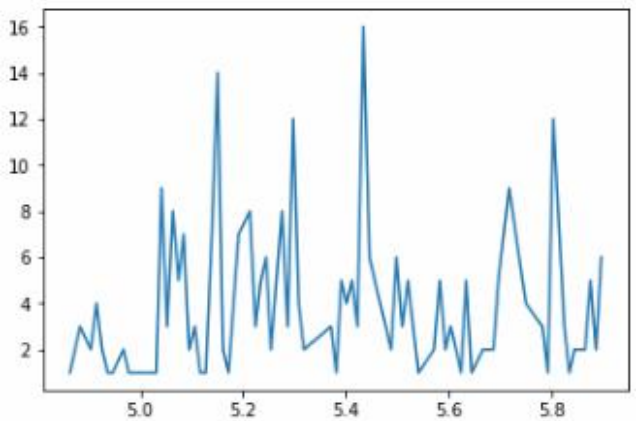
如图可见整体销量有月份的周期性变化：即月初一周的销量较高，附加信息：厄瓜多尔大部分企业的发薪日是在月初，同时在周内的起伏规律更加明显，一周销量的高低排序是 6>7>5>2/1/3>4.

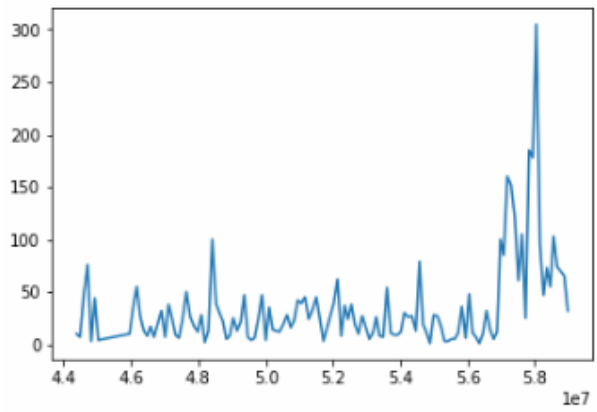
因而时间性质对样本也是具有很大区分度的，也应该加入特征的构造

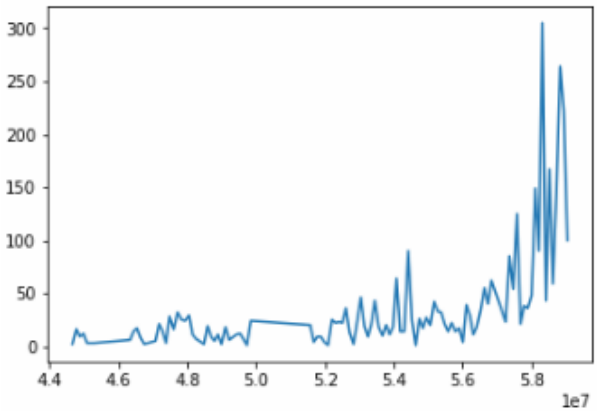
3.4 不同条目的销量变化模式

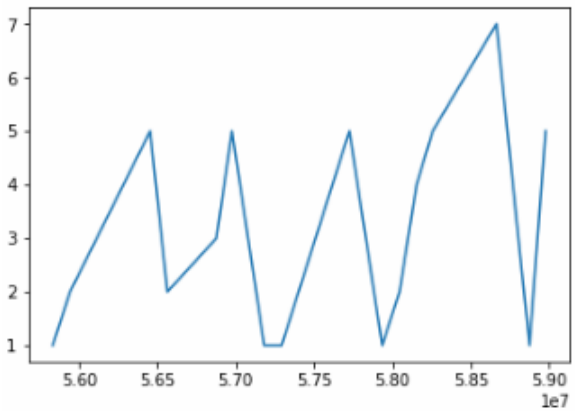
随机选择了一些【商品-商店】组合查看历史销量情况

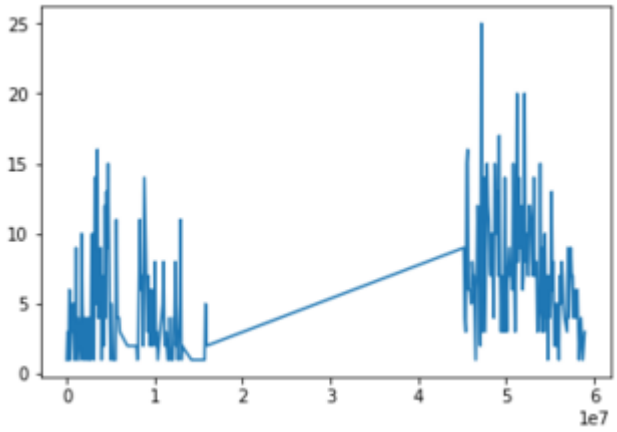












1. 部分商品的销量较为稳定，尤其是低销量商品，而部分商品的销量变化较为明显
2. 部分商品有较为规律的周期变化而部分商品没有
3. 部分商品会有高销售中夹杂 0 销量状态，猜测是断货，下架等原因
4. 不同商品的有销量记录存在的密度差异很大

特征工程便是要提取出这些能分辨样本之间差异性的指标，作为描述样本的特征。

四．特征工程

4.1 历史销量统计量特征

1, 3, 7, 14, 28, 60, 140 天内的平均值，最大值，最小值，方差，偏度，峰度，指数**加权平均值**

这里引申一下关于数据分布的处理，在另一个比赛中，有选手全部将特征分布偏度超过某阈值的特征用其 log 形式代替了，这是一个聪明的做法

这里的指数加权平均值的权重随着日期呈指数衰减

4.2 为以上 1, 3, 7, 14, 28, 60, 140 天内的平均值构造差分

为了捕获长期销量与短期销量的相对变化，构建这一特征

4.3 星期特征

周 X 的历史 4,10,20 周的平均销量

4.4 历史移动窗口平均

以一天为步长向前移动两周，14 次，分别计算 3,7,14,28,60,130 日内的平均销量

这个特征的构造是由于，预测目标是 16 天的单日销量，移动平均值可以捕获一系列起始点不一的均值窗口

4.5 促销特征

1, 3, 7, 14, 28, 60, 140 累积促销天数

预测窗口每日是否打折

4.6 类别特征

将商品的大，小类，商店的类别，商店的 id 特征进行 onehot

五．训练集验证集的选取

5.1 训练集的选取

数据分析显示有一个超市 4.20 开业，再基于越邻近的销量数据更能反映目标区间的销量这一性质，选择 2017.5.31 后的 16 天作为第一个训练集区间，并向后以一周为步长取 8 个训练窗口。

这里查找了 2017.8.16 的星期数并使训练集的星期与之相同

5.2 验证集的选取

将 2017.7.26 后的 16 天作为验证集

六．模型的设计

6.1 lightGBM

由于 lightGBM 的高效和在 kaggle 竞赛中的火爆 public kernel 的试验，受制于数据量的庞大，顾直接使用了 lightGBM 模型，而且'min_data_in_leaf'也设置为了 1500 这样一个相对更大的数值

分别使用了 1.3.5.7 四个随机种子，分别得到结果并做平均

参数设置：

'feature_fraction': 0.7，'subsample': 0.9 bagging

'min_child_weight': 150，'min_child_samples': 10，'min_data_in_leaf': 1500，'num_leaves': 33,控制过拟合

模型通过调参从 CV:0.516,LB:0.529 达到了 CV:0.512,LB:0.524 lightGBM 单模型的最好成绩排在了 410 左右，共 1600 多支队伍

其实应该绘图表示出 train 和 test 的指标变化来观察过拟合的，可是由于机器受限，训练时间过长，模型加速能力有限，参数空间的寻优并没有做出很多尝试

6.2 神经网络 NN 模型

是在发掘比赛后期 CNN,RNN 巨大的预测能力之后所做的尝试

首先使用 sklearn.neural_network.MLPRegressor 的全连接网络，8 层，每层 128 个神经元，使用 2017.5.1 起始的原始价格数据，模型的表现并不好，而且仅仅使用原始价格数据，没有使用促销和类别信息

之后使用了 keras 库，将类别特征作为 embedding 层输入，将促销与否，原始价格信息经过 concatenate，合并之前的 embedding 类别信息，构造了 6 层 256-128-64-32-16-1 的全连接网络，添加了 dropout 层 0.1 和 BatchNormalization。

之后学到 dropout 应该在网络的前半段使用，防止过拟合，或者衰减，而后半段设置的过大会影响模

型的表达能力

6.3 其他尝试

线性模型没有尝试，尝试了多种树模型：ExtraTrees，RandomForest，GradientBoosting，以及用 Bagging 封装的 DecisionTree

七. 引申学习

赛后认真学习了第五名的源码和第一名的源码，收获颇丰

第五名方案：共三个模型进行融合

【Kaggle】Favorita Grocery Sales Forecasting 5th 从源码到方案解读—lightGBM：

https://zhuanlan.zhihu.com/p/33657040

【Kaggle】Favorita Grocery Sales Forecasting 5th 从源码到方案解读—cnn：

https://zhuanlan.zhihu.com/p/33729906

【Kaggle】Favorita Grocery Sales Forecasting 5th 从源码到方案解读—seq2seq：

https://zhuanlan.zhihu.com/p/33712097

第一名方案：共两个模型进行融合

【Kaggle】Favorita Grocery Sales Forecasting 1st 从源码到方案解读--lightGBM

https://zhuanlan.zhihu.com/p/33671145

【Kaggle】Favorita Grocery Sales Forecasting 1st 从源码到方案解读--LSTM

https://zhuanlan.zhihu.com/p/33672879

感悟与教训：

在整个参赛过程中应该迅速建立起反馈，搭建代码架构 pipeline，快速迭代产生结果，并不停通过对错分样本的分析寻找新特征，而不是不断用分析的思路构造各种各样的特征