杂货店商品销量预测

# 赛题描述



Brick-and-mortar grocery stores are always in a delicate dance with purchasing and sales forecasting. Predict a little over, and grocers are stuck with overstocked, perishable goods. Guess a little under, and popular items quickly sell out, leaving money on the table and customers fuming.

购买和**销量预测**对于实体杂货店而言至关重要。如果预测过多，那么杂货商就需要烦恼那些积压的且易腐烂的商品。但是如果预测少了，受欢迎的商品很快就抢购一空，将会无货可卖而且还有失去顾客的危险。

The problem becomes more complex as retailers add new locations with unique needs, new products, ever transitioning seasonal tastes, and unpredictable product marketing. Corporación Favorita, a large Ecuadorian-based grocery retailer, knows this all too well. They operate hundreds of supermarkets, with over 200,000 different products on their shelves.

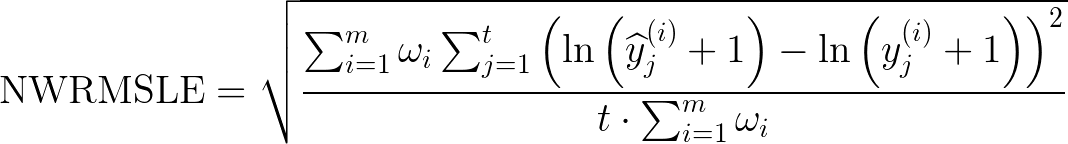
随着零售商添加更多售卖的地点，这些地点有独特的需求、新的产品、不断变换的季节性口味以及难以预测的产品市场，这个问题变得更加复杂。总部位于厄瓜多尔的大型杂货零售商Corporación Favorita对此非常了解。他们经营者数百家超市，货架上有超过200,000种不同的产品。

Corporación Favorita has challenged the Kaggle community to build a model that more accurately forecasts product sales. They currently rely on subjective forecasting methods with very little data to back them up and very little automation to execute plans. They’re excited to see how machine learning could better ensure they please customers by having just enough of the right products at the right time.

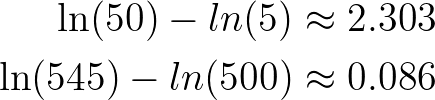
Corporación Favorita向Kaggle社区提出了一个难题，建立一个能够精确预测商品销量的模型。他们目前还依赖于只需要很少数据支撑的主观预测方法，且很少自动执行计划。他们非常想知道机器学习如何能够更好地保证他们能够取悦客户——在正确的时间提供足够多正确的商品。

# 评价指标

归一化加权均方根对数误差（Normalized Weighted Root Mean Squared Logarithmic Error）：



为什么要取对数呢？来看一个直观的例子：



# 数据详情

train.csv

|  |  |
| --- | --- |
| id |  |
| date |  |
| store\_nbr |  |
| item\_nbr |  |
| unit\_sales | int or float (may be negative) |
| onpromotion | 是否促销，有16%为NaN |

备注：

1. 训练数据中不包含unit\_sales为0的商品（对于某天的某个店铺）
2. 不包含库存信息
3. 训练数据中有些产品在测试数据中没有出现

test.csv

|  |  |
| --- | --- |
| id |  |
| date |  |
| store\_nbr |  |
| item\_nbr |  |
| onpromotion |  |

备注：

1. 测试数据中有部分的产品在训练数据中没有出现

stores.csv

|  |  |
| --- | --- |
| store\_nbr |  |
| city |  |
| state |  |
| type |  |
| cluster | 分组（类似的商店） |

items.csv

|  |  |
| --- | --- |
| item\_nbr |  |
| family |  |
| class |  |
| perishable | 是否容易腐烂 |

transactions.csv 每个店铺每天的交易量

|  |  |
| --- | --- |
| date |  |
| store\_nbr |  |
| transactions |  |

备注：只包含训练数据

oil.csv 每天的石油价格

|  |  |
| --- | --- |
| date |  |
| dcoilwtico |  |

holidays\_events.csv

|  |  |
| --- | --- |
| date |  |
| type |  |
| locale |  |
| locale\_name |  |
| description |  |
| transferred |  |

# 特征工程

促销天数特征（6个特征）

* 前i（14、60、140）天中每个店铺商品的促销天数
* 后i（3、7、14）天中每个店铺商品的促销天数

促销销量和正常销量特征（24个特征）

* 前i（3、7、14、30、60、140）天平均促销销量
* 前i（3、7、14、30、60、140）天促销销量和（幂次衰减）
* 前i（3、7、14、30、60、140）天平均正常销量
* 前i（3、7、14、30、60、140）天正常销量和（幂次衰减）

销量统计特征（42个特征）

* 前i（3、7、14、30、60、140）天每个店铺商品销量的平均一阶差分
* 前i（3、7、14、30、60、140）天每个店铺商品销量和（幂次衰减）
* 前i（3、7、14、30、60、140）天每个店铺商品销量的均值
* 前i（3、7、14、30、60、140）天每个店铺商品销量的中位数
* 前i（3、7、14、30、60、140）天每个店铺商品销量的最小值
* 前i（3、7、14、30、60、140）天每个店铺商品销量的最大值
* 前i（3、7、14、30、60、140）天每个店铺商品销量的标准差

销量统计特征2（42个特征）

* 一周前的前i（3、7、14、30、60、140）天每个店铺商品销量的平均一阶差分
* 一周前的前i（3、7、14、30、60、140）天每个店铺商品销量和（幂次衰减）
* 一周前的前i（3、7、14、30、60、140）天每个店铺商品销量的均值
* 一周前的前i（3、7、14、30、60、140）天每个店铺商品销量的中位数
* 一周前的前i（3、7、14、30、60、140）天每个店铺商品销量的最小值
* 一周前的前i（3、7、14、30、60、140）天每个店铺商品销量的最大值
* 一周前的前i（3、7、14、30、60、140）天每个店铺商品销量的标准差

有销量和有促销的天数特征（30个特征）

* 前i（7、14、30、60、140）天每个店铺商品有销量的天数
* 前i（7、14、30、60、140）天中每个店铺商品距离上一次有销量的天数
* 前i（7、14、30、60、140）天中每个店铺商品距离第一次有销量的天数
* 前i（7、14、30、60、140）天每个店铺商品有促销的天数
* 前i（7、14、30、60、140）天中每个店铺商品距离上一次有促销的天数
* 前i（7、14、30、60、140）天中每个店铺商品距离第一次有促销的天数

后15天的促销情况特征（3个特征）

* 后15天每个店铺商品有促销的天数
* 后15天中每个店铺商品距离最后一次有促销的天数
* 后15天中每个店铺商品距离最近一次有促销的天数

前15天的销量（15个特征）

* 前i（1到15）天每个店铺商品的销量

前4（20）周每个星期几的平均销量（14个特征）

* 前4周每个星期（一到日）的平均销量
* 前20周每个星期（一到日）的平均销量

前16天到后15天每天的促销情况（32个特征）

# df

# df

# df

# df