|  |
| --- |
| **机器学习** |
| 厄瓜多尔的大型杂货零售商销售预测 |
| 主 研 人：赵翰宇  参 研 人： |
| 审 核 人：  方 向：工业算法案例研究  版 本 号：A |
|  |

**版本更新**

|  |  |  |  |
| --- | --- | --- | --- |
| **日期** | **版本** | **更新描述** | **作者** |
| 2019/4/11 | A | 初稿 | 赵翰宇 |

**选题表格**

|  |  |  |  |
| --- | --- | --- | --- |
| 时间 | 竞赛名 | 竞赛背景描述（50字以内） | 类型（分类/回归） |
| 2019/4/3 | Corporación Favorita Grocery Sales Forecasting（kaggle） | 预测在厄瓜多尔不同的Favorita商店销售的数千种商品的单位销售量。 | 回归 |

目录

[1. 背景描述 4](#_Toc25575_WPSOffice_Level1)

[1.1 竞赛赛题描述 4](#_Toc5940_WPSOffice_Level2)

[1.2 评估指标描述 4](#_Toc12357_WPSOffice_Level2)

[2. 数据来源及描述性统计分析 5](#_Toc5940_WPSOffice_Level1)

[2.1 大赛数据来源 5](#_Toc22683_WPSOffice_Level2)

[2.2 数据的描述性统计 5](#_Toc16174_WPSOffice_Level2)

[2.2.1 数据基本情况描述： 5](#_Toc5940_WPSOffice_Level3)

[2.2.2 数据字段介绍： 6](#_Toc12357_WPSOffice_Level3)

[2.2.3 数据描述性统计 8](#_Toc22683_WPSOffice_Level3)

[3. 优秀算法思路 8](#_Toc12357_WPSOffice_Level1)

[3.1 方案一 8](#_Toc32293_WPSOffice_Level2)

[3.1.1 方案一特征提取 8](#_Toc16174_WPSOffice_Level3)

[3.1.2 方案一模型设计与建立 9](#_Toc32293_WPSOffice_Level3)

[3.1.3 方案一结果、排名等 9](#_Toc29922_WPSOffice_Level3)

[3.1.4 方案一算法流程图 9](#_Toc4787_WPSOffice_Level3)

[3.2 方案二 9](#_Toc29922_WPSOffice_Level2)

[3.2.1 方案二特征提取 9](#_Toc16174_WPSOffice_Level3)

[3.2.2 方案二模型设计与建立 10](#_Toc32293_WPSOffice_Level3)

[3.2.3 方案二结果、排名等 10](#_Toc29922_WPSOffice_Level3)

[3.2.4 方案二算法流程图 10](#_Toc4787_WPSOffice_Level3)

[4. 算法比较 10](#_Toc22683_WPSOffice_Level1)

[表4-1 算法比较 10](#_Toc7253_WPSOffice_Level2)

[5. 总结与展望 11](#_Toc16174_WPSOffice_Level1)

[5.1 总结 11](#_Toc23198_WPSOffice_Level2)

[5.2 建模思路 11](#_Toc16619_WPSOffice_Level2)

1. 背景描述

实体杂货店总是在购买和预测销售的困扰中。如果预测销售过多，食品杂货商就会被积压，易腐烂的商品会无法处理。如果没预测到位，流行的物品很快就卖光了，顾客只能干拿着钱发脾气。随着零售商增加具有独特需求，新产品，不断变换季节性口味和不可预测的产品营销的新地点，问题变得更加复杂。总部位于厄瓜多尔的大型杂货零售商CorporaciónFavorita，对此非常了解。他们经营着数百家超市，货架上有超过200,000种不同的产品。因此，零售商希望建立一个更准确地预测产品销售的模型。他们目前依靠主观预测方法，只需很少的数据来支持它们，而且很少有自动化来执行计划。他们很高兴看到机器学习如何通过在合适的时间拥有足够的正确产品来更好地确保他们取悦客户。

**1.1 竞赛赛题描述**

本次竞赛是建立一个更准确地预测产品销售的模型，将预测在厄瓜多尔不同的Favorita商店销售的数千种商品的单位销售量。训练数据包括日期，商店和商品信息，是否正在推广该商品以及单位销售。其他文件包含可能对构建模型有用的补充信息。

**1.2 评估指标描述**

本次竞赛的评估指标为归一化加权均方根平方对数误差（NWRMSLE），计算如下：

当在大范围的数量级上预测值时，该度量是合适的。当预测值和真实数都很大时，它避免惩罚预测中的大的差异：当真值为50时预测5比在真值545时预测500惩罚的更多。

1. 数据来源及描述性统计分析

**2.1 大赛数据来源**

CorporaciónFavorita杂货店数据

这里是数据下载的超链接：https://www.kaggle.com/c/7391/download-all

**2.2 数据的描述性统计**

**2.2.1 数据基本情况描述：**

train.csv文件为训练数据，大小为125497041行6列，其中包括：

该目标unit\_sales通过date，store\_nbr以及item\_nbr和一个独特的id标记行。

目标unit\_sales可以是整数（例如，一袋薯片）或浮子（例如，1.5千克的干酪）。

负值unit\_sales表示该特定项目的回报。

该onpromotion列说明是否item\_nbr在促销指定date和store\_nbr。

onpromotion此文件中大约有16％的值是NaN。

注意：训练数据不包括unit\_sales商店/日期组合为零的项目的行。没有关于该商品在该日期是否有库存的信息，团队将需要决定处理该情况的最佳方式。此外，在训练数据中看到的少量项目在测试数据中看不到。

test.csv为测试数据，与date，store\_nbr，item\_nbr是组合要被预测，与沿onpromotion信息。

注意：测试数据包含少量未包含在训练数据中的项目。部分练习是根据类似产品预测新商品的销售情况。

sample\_submission.csv为正确格式的样本提交文件。

stores.csv为存储元数据，其中包括city，state，type，和cluster。

cluster 是一组类似的商店。

items.csv为项目的元数据，其中包括family，class，和perishable。

注：标记为perishable具有分数权重的项目1.25; 否则，重量是1.0。

transactions.csv为销售交易的每一个计数date，store\_nbr组合。仅包含在培训数据时间范围内。

oil.csv为每日油价。包括列车和测试数据时间范围内的值。（厄瓜多尔是一个依赖石油的国家，它的经济健康极易受到油价冲击的影响。

holidays\_events.csv为假期和活动，包含元数据

注意：请特别注意该transferred列。假期transferred正式落在该日历日，但由政府转移到另一个日期。一个transferred日子更像是一个正常的一天，而不是度假。要查找实际庆祝的那一天，请查找相应的行所在的type位置Transfer。例如，假期Independencia de Guayaquil从2012-10-09转移到2012-10-12，这意味着它在2012-10-12庆祝。类型Bridge的天数是添加到假日的额外天数（例如，延长长周末的休息时间）。这些通常是由Work Day通常不会安排工作（例如，星期六）的那种类型组成的，这种类型旨在偿还桥梁。 例如，假期通常在圣诞节周围（使圣诞节前夕成为假日）发生在假日日常假日。

补充说明：

公共部门的工资在每个星期的第15天和最后一天每两周支付一次。超市的销售可能受此影响。

2016年4月16日，厄瓜多尔发生7.8级地震。人们在救援工作中集会，捐赠水和其他首要需求的产品，这些产品在地震发生后数周内对超市的销售产生了极大的影响。

**2.2.2 数据字段介绍：**

表2-1 train.csv数据表字段介绍

|  |  |  |  |
| --- | --- | --- | --- |
|  | **变量含义** | **变量类型** | **缺失率** |
| **Date** | 测试集日期 | 日期 | 0% |
| **id** | Id | 离散 | 0% |
| **store\_nbr** | 商店编号 | 离散 | 0% |
| **item\_nbr** | 物品编号 | 离散 | 0% |
| **unit\_sales** | 单位销售量 | 离散 | 0% |
| **onpromotion** | 是否促销 | 布尔 | 16% |

表2-2 **store.csv**数据表字段介绍

|  |  |  |  |
| --- | --- | --- | --- |
|  | **变量含义** | **变量类型** | **缺失率** |
| **store\_nbr** | 商店编号 | 离散 | 0% |
| **city** | 商店所在城市 | 字符串 | 0% |
| **state** | 商店所在州 | 字符串 | 0% |
| **type** | 商店类型 | {A,B,C,D} | 0% |
| **cluster** | 一组类似的商店 | 离散 | 0% |

表2-3 **item.csv**数据表字段介绍

|  |  |  |  |
| --- | --- | --- | --- |
|  | **变量含义** | **变量类型** | **缺失率** |
| **item\_nbr** | 商品编号 | 离散 | 0% |
| **family** | 家庭 | 字符串 | 0% |
| **class** | 类别 | 离散 | 0% |
| **perishable** | 权重判断 | {0,1} | 0% |

表2-4 **transactions.csv中的**数据表字段介绍

|  |  |  |  |
| --- | --- | --- | --- |
|  | **变量含义** | **变量类型** | **缺失率** |
| **date** | 日期 | 日期 | 0% |
| **store\_nbr** | 商店编号 | 离散 | 0% |
| **transactions** | 销售交易计数 | 离散 | 0% |

表2-5 **oil.csv中的**数据表字段介绍

|  |  |  |  |
| --- | --- | --- | --- |
|  | **变量含义** | **变量类型** | **缺失率** |
| **date** | 日期 | 日期 | 0% |
| **dcoilwtico** | 每日油价 | 离散 | 3.5% |

表2-6 **holidays\_events.csv**数据表字段介绍

|  |  |  |  |
| --- | --- | --- | --- |
|  | **变量含义** | **变量类型** | **缺失率** |
| **date** | 日期 | 日期 | 0% |
| **type** | 活动或假期种类 | 字符串 | 0% |
| **locale** | 活动的性质 | 字符串 | 0% |
| **locale\_name** | 活动名称 | 字符串 | 0% |
| **description** | 描述 | 字符串 | 0% |
| **transferred** | 是否转移节假日期 | 布尔 | 0% |

**2.2.3 数据描述性统计**

该竞赛数据维度较高，通过题注可知，时间序列对该数据集影响很大，因此要注重考虑时间维度。数据中字符串格式较多，要先进行转化。

3. 优秀算法思路

**3.1 方案一**

**3.1.1 方案一特征工程**

此方案的模型仅使用2017年数据来提取特征和构建样本，首先进行了数据预处理，训练数据：20170531-20170719或20170614-20170719，用不同的数据集训练不同的模型。验证集为：20170726 - 20170810。将train和test中的每列定义类型，单独存储时间序列，test按store\_nbr,item\_nbr 和date 来设置索引。item文件也按item\_nbr 设置索引。然后将onpromotion中的缺失值用False填充，用0来填补缺失或否定的促销和目标值。

方案选取的基本特征为

1，类别特征：商店，商品，家庭，类，集群......

2，dayofweek（仅适用于模特3）

选取的统计特征为：

使用一些方法来统计不同时间窗口中不同键的一些目标

一、时间窗口

1，最近的日期：[1,3,5,7,14,30,60,140]

2，等时窗口：[1] \* 16，[7] \* 20 ......

二，key：存储x项，项，存储x类

三、target：promotion，unit\_sales，0

四、方法

1，mean，median，max，min，std,

2，自上次出现以来的几天

3，相邻时间窗口之间的平均值差异（仅适用于等时间窗口）

不选取的无用的功能

1，假期

2，其他如：cluster x item，store x family ...

**3.1.2 方案一模型设计与建立**

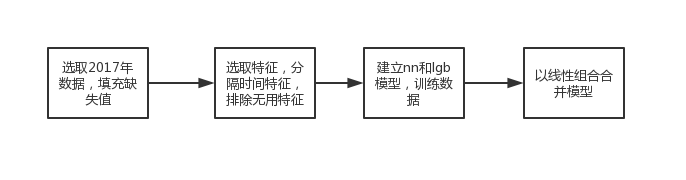
方案一建立了4个大类模型，第一个为对应16天的16个lgb模型。第二个为对应16天的16个nn神经网络模型。第三个为包含16天的数据训练的1个lgb模型，第四个为一个基于sjv的代码的神经网络模型。

## 最终在进行模型组合时，团队发现stacking表现得不是很好，于是最终采取了4类模型的线性组合，final submission = 0.42\*model1 + 0.28 \* model2 + 0.18 \* *model*3 + 0.12 \* model4。

**3.1.3 方案一结果、排名等**

此方案获得了第一名，公共排行榜上为0.50487，排名第14位，私人排行榜上为0.50918，排名第一位。

**3.1.4 方案一算法流程图**



**图3-1**

**3.2方案二**

**3.2.1 方案二特征工程**

方案二使用随机抽样128序列的小批量数据。然后随机选择解码/目标日期的开始。因此，可以说某种程度上模型将为每次训练迭代看到不同的数据。因为总数据集大约是170000（seq）x 365天。能够很好地处理过度拟合。

方案也注重unit\_sales和onpromotion特征，因为可用于更好地捕获季度和年度模式。对新旧商品不采用模型来预防过拟合处理，而是采用损失函数。

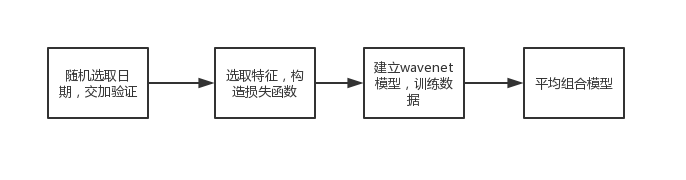
**3.2.2 方案二模型设计与建立**

方案选用wavenet模型，建立了5个模型，用指数加权平均来平均预测值。

**3.2.3 方案二结果、排名等**

此方案获得了第二名，公共排行榜上为0.50444，排名第12位，私人排行榜上为0.51296，排名第二位。

**3.2.4 方案二算法流程图、**



**图3-2**

4. 算法比较

**表4-1 算法比较**

|  | **评估指标** | **特征工程** | **基础算法** | **基本库** |
| --- | --- | --- | --- | --- |
| **算法1** | **NWRMSLE** | **选取时间节点，删除无用特征** | **Lgb和神经网络** | **Python的Sklearn** |
| **算法2** | **NWRMSLE** | **构造损失函数** | **Wavenet神经网络** | **Python的Sklearn** |

算法一、二都进行了常规的数据预处理操作，在特征建立上，算法一更为详细的研究了各特征的重要性，将时间节点着重提出，并删去了无用的特征。算法二同样也挖掘了相关性大的特征，用于区分时间序列对预测的影响。

在模型建立上，算法一仅选取了2017年的数据，尽管大幅缩减了数据量，但是对预测未免有一些局限的偏差。算法一分别建立每天的lgb和16天总的lgb模型进行预测，能更合理的从局部到总体的进行训练。再加上16个神经网络的模型，共同加权线性组合出预测结果，可以确保模型的训练较好。算法二随机选取时间节点和数据，用深度神经网络来进行训练，训练效果较好，但有些忽略时间序列的影响。

5. 总结与展望

**5.1 总结**

此次销售预测竞赛的关键在于大数据量的选取和时间维度的处理，设计的算法以神经网络等模型组合为主。对于特征的处理，采取取均值的效果较好。

**5.2 建模思路**

首先进行数据预处理，以False和0填充缺失值。其次，提取出特殊促销时期的数据，其余时间以日期作为区分，将每年每月的同一日的数据取均值进行数据聚合，提取出31个时间特征值。在算法上，用lgb算法，对时间特征值分别进行建模，预测效果会更佳。