基于XGBoost的Rossmann商店销量预测分析

朱林

2018/4/5

# 一、问题定义

## 1.1、项目概述

在商业数据分析领域，对于分析的准确性要求越来越高，之前的主观推测方法，由于掺杂了很多人为因素，容易导致误判，所以很难满足要求。如今随着数据量增多、计算机算力提升迅速这两个因素的影响下，机器学习的方法有了天然的土壤，于是有很多机器学习的方法得以运用起来。本研究是基于欧洲一家连锁药店-Rossmann药妆商店的数据进行的分析，数据包括了：1115家商店的相关信息、一百万多条关于上述商店的日销量等信息。分析目的是，试图通过给定的销售量等数据，分析得出在接下来的一段时间内，上述商店的销售情况。

首先要对德国的商业环境有所了解，德国的法律规定，只有在少数周日商店才可以开门，可想而知，周日的销量可能不会和weekday的销量有太大差别。但是由于周日不开门，因此周六商店会开门很长时间，所以周六的销量可能会大大高于平日。基于上述信息，首先对给出的数据有有一定的了解。kaggle给出的数据主要包括三类：store-商店相关信息；train-商店的日销量等信息；test-测试数据。有哪些因素会影响一家商店的销量呢？日常逛街的时候，主要关注的信息有：星期几、商店是否有折扣，是否是假期、商店类型、到店人均消费。另外，还会关注周边的信息，比如商店和其他类似商店的远近比较。这些信息大多数都在给定的数据中可以找得到，另外一些需要经过数据处理才能得到。当然，还有一些周边信息，没有在数据中体现，比如当时的天气状况、气温等信息，这些信息也会影响人们去商店的意愿。

上述提到的机器学习方法中，有监督学习和非监督学习、半监督学习。另外现在图像识别问题中最常用的应该是深度学习方法。随着 Alpha Go 的巨大成功，宣告人类在围棋这个极其复杂的棋种，已经不可能超越机器了，机器也不再像以前一样需要人工输入大量棋局和指导，因为他们能够自己进行学习，强化学习让大家眼前一亮。该问题中，由于我们是基于给定的数据和标签结果，通过计算得出我们的判断，因此我们需要使用的是监督学习模型。

## 1.2、问题陈述

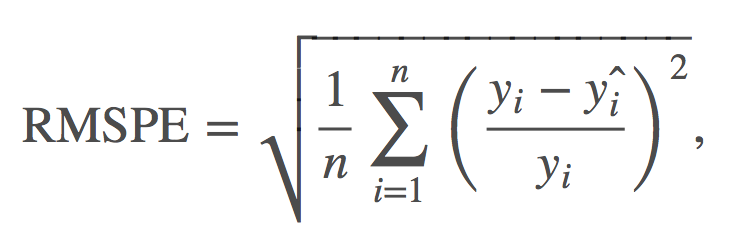
由于需要预估的销售量，是一个连续的数据，因此该问题属于一个回归问题，而且是有监督的回归问题。有监督学习的模型中，我选择 XGBoost 作为训练结果，因为他作为boost类别的方法，能够集合多个弱分类，综合判断得到结果，往往比单个模型得到的结果要好。

该问题的数据分为商店数据和训练数据，所以首先需要进行合并操作。其次，需要对数据进行一些预处理，去掉一些误差点等信息。最后再进行 XGBoost 模型的训练和调参。

由于该问题是基于 kaggle 的项目，因此项目最终结果需提交 kaggle 得到最终的结果。总共有3303支队伍参赛，评价标准是，起码需要达到总排行榜的前15%，力争进入前10%。由于有两个排行榜，Private榜 和 Public榜，分别表示随机取所有数据的 61% 和 39%。因此我选取更小随机性的 Private榜作为参照，进入前15%和10%的分数分别为：0.11959 和 0.11773。

## 1.3、评价指标

最终需要输出一个csv文件，包含所有test数据的预测销量结果，将其提交到kaggle进行评价。该项目 kaggle 基于的评价指标 RMSPE，表示均方根误差的百分比值，能够将误差放到同一个量纲上进行比较。它的计算公式如下：



y\_i ： 销量误差的实际值

yhat\_i ： 销量误差的预测值

n ：测试集的数目

当然，该判断中，排除了销量为0的数据不计算在内。

# 二、分析

## 2.1、数据的探索

## 2.2、探索可视化

## 2.3、算法和技术

## 2.4、基准模型

# 三、方法

## 3.1、数据预处理

## 3.2、执行过程

## 3.3、完善

# 四、结果

## 4.1、模型的评价与验证

## 4.2、合理性分析

# 五、项目结论

## 5.1、结果可视化

## 5.2、对项目的思考

## 5.3、需要做出的改进