# 机器学习工程师纳米学位

## 报告人/ 时间

焦龙

2019-06-09

## 开题报告

### 领域背景

Rossmann在欧洲7个国家经营着3,000多家药店，而预测销售额是每个商店经理的日常工作任务。由于每个商店经理的预测只能根据自身情况进行预测，导致准确性存在很大偏差，我们可以通过构建预测模型来帮助商店经理完成未来6周销售额的预测任务，以便商店经理提前规划销售策略及相关准备工作。

在机器学习领域中，对销售额进行预测是监督学习常见的回归问题，加上Rossmann有足够的数据，我们完全可以实现对这3,000多家药店进行精准销售额预测。

### 问题描述

对1115家店进行提前6周的销售额预测， 其中所需的数据已从kaggle下载，包括训练数据，测试数据，商店数据。

### 数据集和输入

由于数据的最小粒度为每个商店每天的销售情况，因此可以判定这些数据是专门针对预测每个店销售额进行数据收集整理的，所以使用这些数据可以进行销售额预测。

数据集中包含3个文件,分别为train.csv、test.csv、store.csv。其中train.csv是训练数据、test.csv是测试数据，这两个文件的字段相同。store.csv存储的是与商店有关的信息，主要是为了对训练数据进行扩充，train.csv与store.csv这两个文件通过Store字段进行数据合并。

train.csv有1017209条数据，所有数据都是非空数据，无需对其进行填充，销售数据的时间跨度是从2013-1-1日到2015-7-31日。store.csv有1115条数据，也就是说有1115家商店，其中"竞争对手"信息有空数据，需要采用平均值方式对其进行填充。而"促销信息"也存在空数据，由于有将近一半的店铺不参与促销，因此无需对促销信息中的空数据进行填充。

数据集中必要字段含义及特征如下：

Store- 每家商店的唯一ID

Sales - 任何一天的营业额（这是您预测的）

Customers- 特定日期的客户数量

Open - 商店是否营业的指标：0 =关闭，1 =打开

StateHoliday - 表示州假日。通常情况下，除了少数例外，所有商店都会在州假期关闭。请注意，所有学校在公众假期和周末都关闭。a =公共假日，b =复活节假期，c =圣诞节，0 =无

StoreType - 区分4种不同的商店模型：a，b，c，d

Assortment- 描述分类级别：a =基本，b =额外，c =扩展

CompetitionDistance - 距离最近的竞争对手商店的距离（以米为单位）

CompetitionOpenSince [Month / Year] - 给出最接近的竞争对手开放时间的大致年份和月份

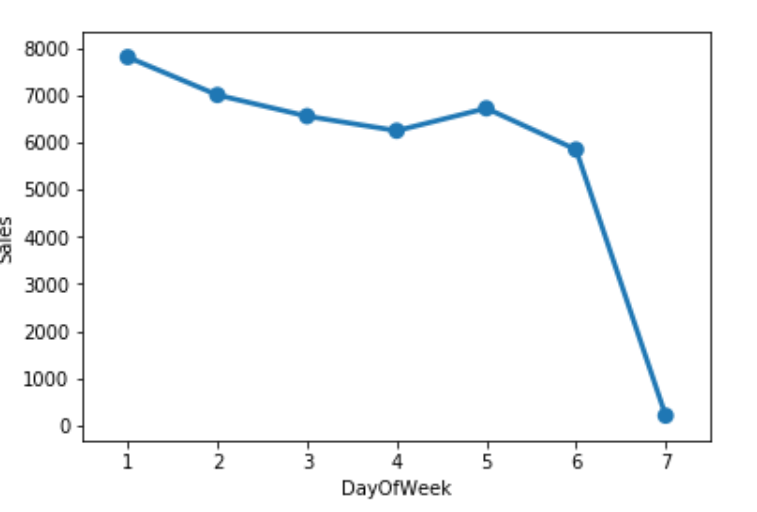
Promo- 指示商店当天是否正在运营促销

Promo2 - Promo2是一些商店的持续和连续促销：0 =商店不参与，1 =商店参与

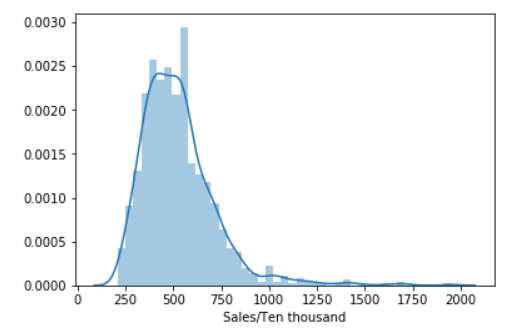
Promo2Since [年/周] - 描述商店开始参与Promo2的年份和日历周

PromoInterval - 描述Promo2启动的连续间隔，命名重新开始促销的月份。例如，“2月，5月，8月，11月”意味着每一轮在该商店的任何给定年份的2月，5月，8月和11月开始

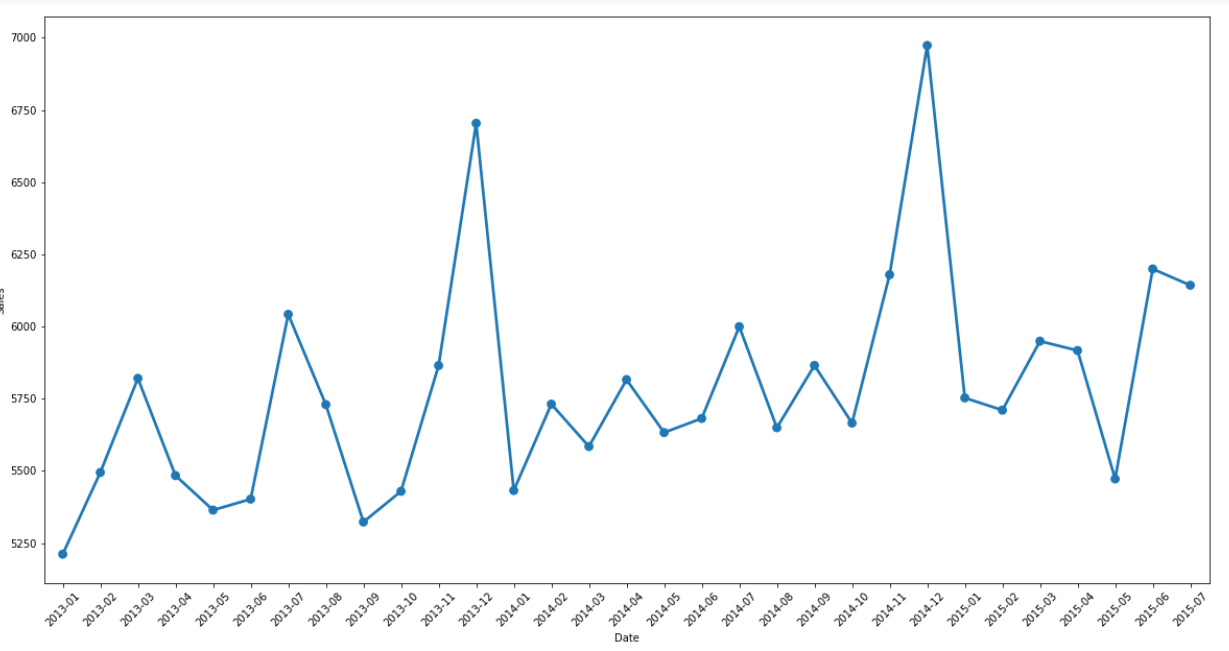
我们可以对和销售额相关的字段进行分析，观察那些征与标签存在必然联系，首先看一下DayofWeek与Sales的关系



通过观察，发现周日（7）的时候销售额几乎为0，说明在周日商店是不营业的。数据集中所有商店销售额的分布情况如下：



通过观察发现，销售额属于正偏态分布，大部分商店的销售额集中在500万附近。月份与销售额之间的关系如下：



通过观察，明显发现10月至12月是销售是销售的高峰期，是销售产品最好的时间段。

### 解决方案

对数据中的商店类别字段、时间等字段进行数据预处理，及数据归一化。通过处理后数据的维度很高，因此需要对其进行主成分分析。在训练模型上使用推荐的 XGBoost ，指标使用指定的RMSPE 评价指标进行模型评估及调参。

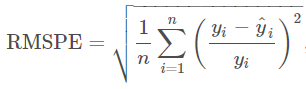
### 基准模型

train.csv数据中基准销售额预测：

平均值 ：5773.818972

RMSPE值：0.4553159240500434

### 评价指标



y\_i ：表示单日商店的销售额

yhat\_i：表示相应的预测。

n ：表示数据的数量，需要注意的是在该评分指标中，每天销售额为0的商店销售数据需要被忽略。

### 项目设计

0. 导入train.csv、store.csv数据。

1. 将train.csv、store.csv合并为一个大表。

2. 数据预处理。

3. 数据划分为训练数据和测试数据

4. 训练模型

5. 与基准模型对比评价训练模型指标。

6. 优化模型。

7. 结果上传kaggle 。

采用模型：

xgboost是一个优化的分布式梯度增强库，可在分布式系统中对海量的数据进行并行训练，因此采用xgboost进行模型训练，可以实现高效，灵活，便捷。

数据划分：

由于我们要预测未来6周的销售额，因此在分割数据的时候，需要按日期时间对数据进行排序，然后获取6周连续的数据作为测试数据。不可随机对数据进行划分。

模型优化：

1. 使用“网格搜索”对模型参数进行优化。
2. 对数据进行特征工程，如对StoreType进行One-Hot编码、归一化、挖掘新特征。
3. 可以采用XGBoost + LR 模型融合

### 参考文献

【 1 】 XGBoost 官方文档：https://xgboost.readthedocs.io/en/latest/

【 2 】 Kaggle 数据说明： https://www.kaggle.com/c/rossmann-store-sales/data

【 3 】 Pandas 官方文档：http://pandas.pydata.org/pandas-docs/stable/index.html

【 4 】 特征工程：<https://www.zhihu.com/question/29316149>

【 5 】 模型融合：https://zhuanlan.zhihu.com/p/42123341