Machine Learning Engineer Nanodegree

预测 Rossmann 未来销量项目

姚连英

2018 年 12 月 25 日星期二

项目背景

Rossmann 公司在 7 个欧洲国家拥有超过 3000 家药店。Rossmann 的经理的任务是提前六周预测药店的日常销售金额。药店销售额会受到包括促销、竞赛、学校和州的节假日，季节以及所在地址等等因素的影响。Rossmann 的经理目前已经根据实际情况，用个人的方式进行了种种预测，由于预测方式的不同，实际准确性上有较大的区别。

为此，Rossmann 公司举行了预测日常销售金额的竞赛项目。作为竞赛参与者，需要对在德国的 1115 家药店预测 6 周日常销售额。准确率高且可靠的销售预测模型可以为 Rossmann 公司各个药店的经理创建创建更加有效的员工日程安排，从而提高员工的劳动力与工作积极性。同时，为 Rossmann 建立预测模型，也可以帮助各位药店经理将重点和焦点放在如何提高客户满意度以及优化团队建设这些更加重要的方面上。

为解决此问题，需要先对历史销售数据进行探索性数据分析(EDA)，处理异常值，发现潜在结构；提取合适的特征，建立预测模型；对模型进行训练，获得模型较优参数，提高预测准确性。

问题陈述

为 Rossmann 公司的 1115 家德国药店建立预测未来 6 周的销售的模型。该任务是一个回归预测类问题，即根据现有的历史数据：各药店历史日期中的销售、药店规模、促销方案、假期安排等等数据，进行特征分析以及预测目标分析。从历史数据合理划分训练集和验证集，使用试 GBDT 类模型，例如 xgboost、lightgbm 等模型预测未来 6 周的销售情况。

数据集和输入

- 文档 train.csv - 包含销售的历史数据

  test.csv - 无销售的历史数据，用于输入模型后得出预测值，与实际销售比较计算准确率

  sample_submission.csv - 正确提交的数据格式

  store.csv - 关于商店的补充信息（包含商店规模、促销方案与时间等信息）

- 文档字段说明

  Id - 每个日期的每个药店的识别序号 Store - 每个药店的唯一序号

  Sales - 每个已知日期的药店的周转(turnover) (目标预测值)

  Customers - 已知日期的客户量

  Open - 药店是否营业的标示符: 0 = 不营业, 1 = 营业

  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. 所有的学校在公共假日和周末都是 close 的。 a =

public holiday, b =Easter holiday, c = Christmas, 0 = None

SchoolHoliday - 表明（商店，日期）是否受到公立学校关闭的影响

StoreType - 区分 4 种不同的商店规模: a, b, c, d

Assortment - 描述了一个分类级别: a = basic, b = extra, c = extended

CompetitionDistance - 距离最近的竞争对手商店的距离

CompetitionOpenSince[Month/Year] - 距离最近的竞争对手开业的时间（月/年）

Promo - 表示当天该药店是否在进行促销活动

Promo2 - 表示一些药店持续的促销活动: 0 = 药店未参与, 1 = 药店正在参与

Promo2Since[Year/Week] - 描述商店开始参与持续促销活动的年份和日历周

PromoInterval - 描述持续促销活动，启动的连续间隔，以促销月份命名 E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store

解决方案

1.探索数据。发现并处理异常以及空缺等数值；

2.分析数据相关性。根据基础数据，分析每个药店的销量与日期、节假日、促销等客观因素之间的相互关系，根据相互关系产生必要的新特征；

3.用历史数据训练 XGBoost 模型，确认最优模型参数； 4.用模型预测未来 6 周的销量。

基准模型

准备使用回归模型：XGBoost 模型

参数设定：

params = {"objective": "reg:linear",

"booster" : "gbtree",

"eta": 0.3,

"max_depth": 10,

"subsample": 0.9,

"colsample_bytree": 0.7,

"silent": 1,

"seed": 1301

}

用 XGBoost 模型递归地构建二叉决策树的过程，基于训练数据集生成决策树，使用最小二乘偏差（LSD）或最小绝对偏差（LAD）来获取最优分裂属性；用验证数据集对已生成的树进行剪枝并选择最优子树，这时损失函数最小作为剪枝的标准。

评估指标

采用 RMSPE 评估，数学表达式为：

$$\text{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2},$$

注释：

y_i ：在某日期某个 store 的实际销售；

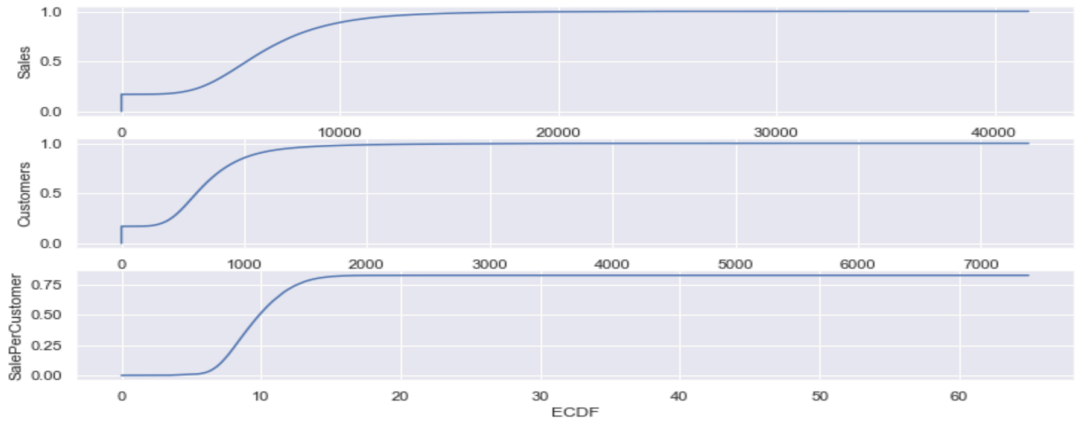
yhat_i ：经模型预测在某日期某个 store 的实际销售。

任何销量为 0 的预测在评价中都会被忽略。

项目设计

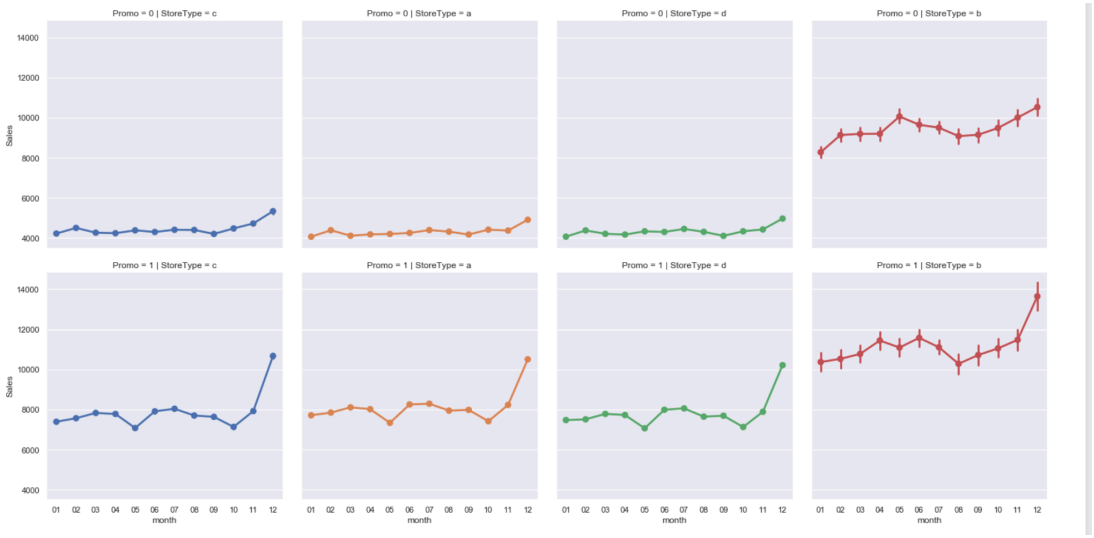
- 数据可视化分析

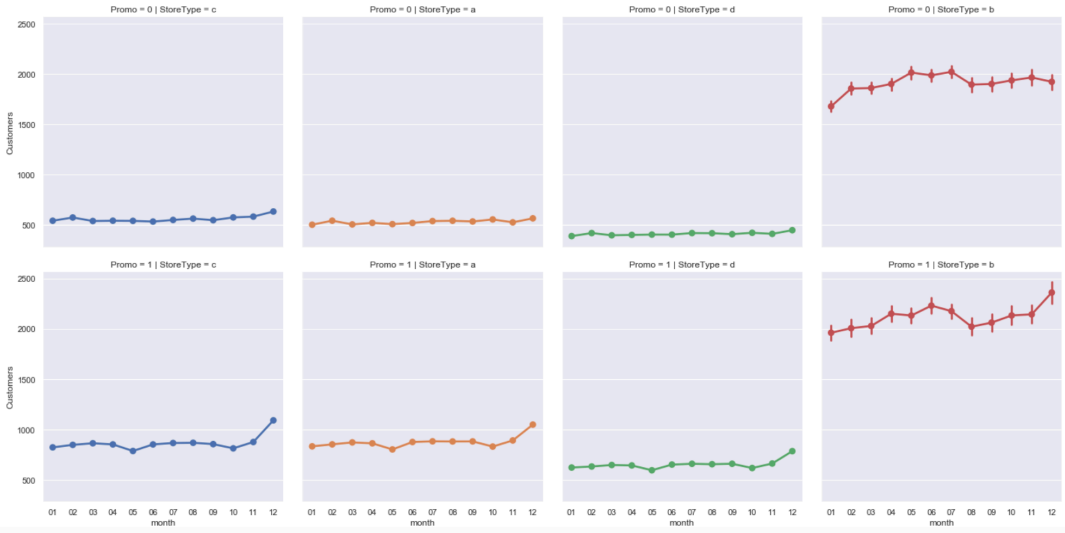
1. 关键数据的 ECDF 分析（Sales/Customers/SalePerCustomer）

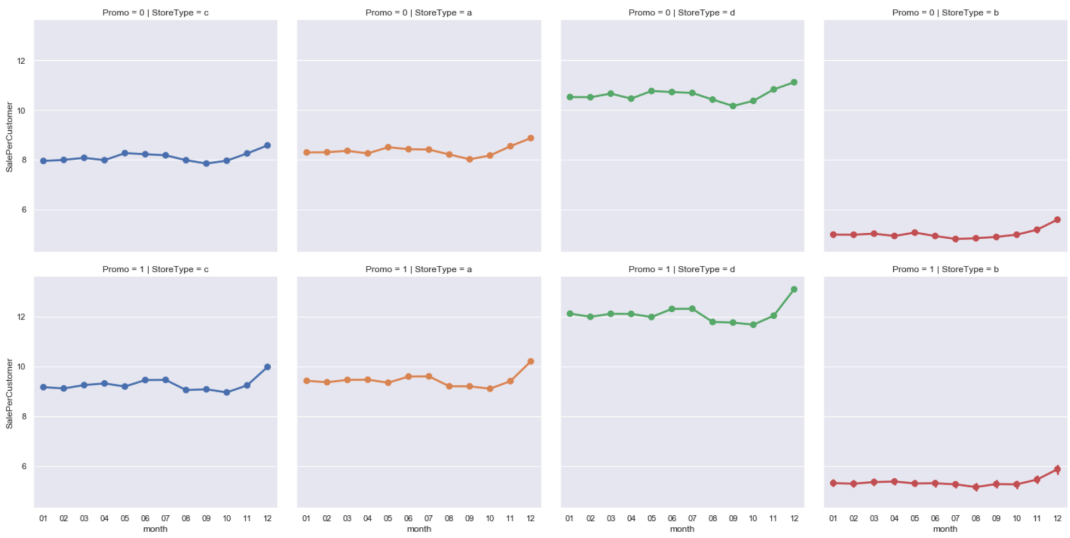


训练集（train.csv）当中有约 20%的数据没有 Sales 和 Customers 记录，主要原因是 store 未营业，极少数是由于极端情况导致在营业情况下无营业额。

2. 关键数据逐月分析


Promo = 0 | StoreType = c
Promo = 0 | StoreType = a
Promo = 0 | StoreType = d
Promo = 0 | StoreType = b
Promo = 1 | StoreType = c
Promo = 1 | StoreType = a
Promo = 1 | StoreType = d
Promo = 1 | StoreType = b
Sales
month


Promo = 0 | StoreType = c
Promo = 0 | StoreType = a
Promo = 0 | StoreType = d
Promo = 0 | StoreType = b
Promo = 1 | StoreType = c
Promo = 1 | StoreType = a
Promo = 1 | StoreType = d
Promo = 1 | StoreType = b
Customers
month



根据 Sales/Customers 的变化，可合理推测，store 的销售与 StoreType 有直接关系，且虽然 StoreType 为 b 的 store 的客户量和销售额高于其他类型的 store，但是人均销售却是最低的，推测 b store 主要销售小类多量商品；

4 类 store 的销量在月份上基本呈现相同的波动趋势，推测实际销量与月份有较大的联系。

- 数据特征处理

1. "Open"为空赋值为 1，默认所有无 Open 状态的 Store 均为营业状态；
2. train.csv 训练集当中只看 open 为 1 且 sales>0 的子集记录；
3. 合并 store.csv 和 train.csv&test.csv,在训练集和预测集上增加 store 状态明细，扩张特征；
4. 所有空值填 0；

5. 对分类向量'StoreType', 'Assortment', 'StateHoliday'中将 0abcd 转换成数值 01234；
6. 从日期中抽出 Year, Month, Day, DOW(Day of Week), WOY(Week of Year)；
7. 创建竞争对手开店月数，优惠月数，每条记录的所处月份是否是优惠月，作为新的特征向量。

- 建立模型

1. 选择模型 XGBoost 模型
2. 参数设定：

   params = {"objective": "reg:linear",

   "booster" : "gbtree",

   "eta": 0.3,

   "max_depth": 10,

   "subsample": 0.9,

   "colsample_bytree": 0.7,

   "silent": 1,

   "seed": 1301

   }
3. 拆分训练集，取其中 10%数据作为验证集，验证模型预测准确率情况
4. 用训练完成的模型预测测试集，获得预测值。

参考文献与链接：

1. 模型建立： https://www.jianshu.com/p/7467e616f227

2. 数据可视化与处理：https://www.kaggle.com/elenapetrova/time-series-analysis-and-forecasts-with-prophet