毕业项目开题报告

王珏

November 22st, 2019

项目背景

Rossman 是欧洲的一家连锁药店，其在欧洲 7 个国家/地区拥有 3,000 多家店铺。因经营需要，Rossman 的经理需要提前 6 周来预期店铺的销售额。不同的店铺面临的经营情况大相径庭，他们的销售受到许多因素的影响，包括促销，竞争，学校和州假期，季节性和地区性。

该项目是一个典型的监督学习问题，关于监督学习，我已经学习了许多模型，包括感知机，决策树，随机森林，朴素贝叶斯等。由于我个人的工作行业关系，时常会遇到这类监督学习的问题，因此，希望利用这个机会来提高自己的分析能力，因此我选择了 Rossman 销售额预测。

问题描述

监督学习可以分为回归问题和分类问题，我们需要预测 Rossman 商店未来的销售额，因此这是一个回归问题。从参赛者经验上看，只使用其给出的特征难以得到一个较好的效果，创造衍生特征成为了必选项。同时，该数据集较大，其训练集包含了上百万条数据，因此在模型训练时，还需要兼顾运算速度。最后，训练集中的特殊数据需要进行事先处理，比如当店铺关门时，其销售额为 0，这种情况需要和开门时销售额为 0 区分开来。此外，还有部分数据缺失的问题。

数据集和输入

在这个项目中，Kaggle 提供的数据集中包含了训练集（train）和测试集（test），另外还包含了一个店铺的信息表（store）和名为 sample_submission 的文档。

训练集包含了 1115 家店铺在一段时间内营业数据，具体见下图：

1. Id:测试集内(商店、日期)的组合。
2. Store:表示每个商店的唯一Id。
3. Sales:任意一天的销售额，也是我们要预测的字段。
4. Open:是否开门，0=关门，1=开门。
5. StateHoliday:国家假日，一般假日国家假期都会关门，所有学校在公共假日都会关门，a=公共假日，b=东部假日，c=圣诞节，0=不是假日。
6. SchoolHoliday:校园假日，指当日学校是否关闭。
7. StoreType:商店类型，有四种，abcd。
8. Assortment:分类级别，a=基础，b=额外，c=扩展。
9. CompetitionDistance:竞争对手距离。
10. CompetitionOpenSince[Month/Year]:给出最近竞争对手的开张时间。
11. Promo:表示商店当天是否进行促销
12. Promo2:表示商店是否进行持续的促销活动，0=没有参数，1=参与。
13. Promo2Since[Year/Week]:商店开始持续促销的年/星期。
14. PromoInterval:持续促销活动开始的间隔，"Feb,May,Aug,Nov"表示给定商店某一年的2589月开始持续促销活动。

我们需要将每个店铺每日的营业情况，和店铺自身的信息相关联，对特征进行分析建模，并对测试集中的数据进行预测，并将结果按照 sample_submission 文件的格式进行提交。

通过的训练集进行观察，我们发现存在缺失数据以及无用数据。

```
Data columns (total 10 columns):
Store                        1115 non-null int64
StoreType                    1115 non-null object
Assortment                   1115 non-null object
CompetitionDistance          1112 non-null float64
CompetitionOpenSinceMonth    761 non-null float64
CompetitionOpenSinceYear     761 non-null float64
Promo2                       1115 non-null int64
Promo2SinceWeek              571 non-null float64
Promo2SinceYear              571 non-null float64
PromoInterval                571 non-null object
dtypes: float64(5), int64(2), object(3)
memory usage: 87.2+ KB
```

上图可以看出，CompetitionOpenSinceMonth，CompetitionOpenSinceYear，Promo2SinceWeek，Promo2SinceYear，PromoInterval 数据存在缺失值。

同时，也观察到存在部分店铺在某些日期不开门，导致当日销售额为 0。

解决方案

针对缺失值，我考虑用过去一段时间的平均值进行代替，对于不开门时产生的数据，我会关门前后若干天的平均值代替。如果最终效果不好，也会采用别的方法，甚至删除数据。

对于衍生特征，我会参考其他人的意见以及自身的想法，结合 2-3 个特征进行处理。

由于数据较多，因此在选择模型时，会更关注其处理速度。选择运算速度较快的模型进行最终的预测。

基准模型

根据该项目获奖者的介绍，本项目我将使用 XGBoost 作为基准模型，XGBoost 是经过优化的分布式梯度提升库，旨在高效，灵活且可移植。它在 Gradient Boosting 框架下实现了机器学习算法。其特点是速度快，模型表现较好。

于此同时，我会适当采用其他模型进行对比，在速度和准确性上观察 XGBoost 模型的优势有多大。

评估指标

本次项目的评估指标由 Kaggle 给出，为均方根百分比误差(RMSPE)，其计算公式如下图：

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

RMSPE 与 均方误差 RMSE 相比，区别在于在计算误差时要除以真实值 yi，我认为这样的好处当 yi 的取值范围很大时，比如最大值与最小值相差几十万倍，那么当高值预测出现误差时，产生的误差就会非常大，从而影响到了对其他值预测误差的评估。 通过将误差除以 yi，相当于在归一化这些误差，这样能够平等对待较大值与较小值。

项目设计

在这个项目中，我们需要达到 Kaggle 前 10%的目标，为了实现这个目标，我们需要进行如下工作：

1、 数据观察：删掉对分析没有用处的特征

2、 数据预处理：补充缺失数据，删除不合理的数据。

3、 特征分析：对特征的分布情况，与销售额的相关性进行分析，根据其他参赛者的经验，我将选择一部分特征进行加工，产生若干个衍生变量用于分析。

4、 数据划分：默认是采用随机划分的方式来进行学习，但是也有建议是截取其中连续的一段数据作为验证集，原因是需要让模型学习到数据的连续性。我计划先采取随

机划分数据的方式，如果准确率无法提升的话，再采取按日期划分。

5、 确定评估指标：确定一个合适的评估指标，并将其用于评价模型训练结果的好坏。

6、 模型选择：对备选模型进行训练，并选择一个合适的模型进行调参，以提高预测的准确率。

7、 模型调参：首先，基于 XGBoost 模型的参数，我认为首先应该保持一个大的 eta（学习率）的条件下，对其他参数进行选择，比如较为重要的 max_depth 等。通过模型效果的相对好坏，来决定使用的参数，最后，通过减小 eta ，增加 num_round 参数来让模型的结果更进一步。由于在数据观察的部分，我们观察到数据与日期具有强相关性，背后的商业逻辑则是零售业是一个有着明显周期性的行业。因此我认为如果将预测的数据也画进整体的数据图中的话，我仍然能够观察到上述规律，一次侧面判断我预测的准确性。

最终，将预测结果提交给 Kaggle，并根据得分进行进一步调参，直至结果达到前 10%。

资源引用：

https://www.kaggle.com/c/rossmann-store-sales/data

https://www.kaggle.com/c/rossmann-store-sales/discussion/18024#latest-652735

https://dnc1994.com/2016/04/rank-10-percent-in-first-kaggle-competition/