毕业项目开题报告

王珏

November 19st, 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

我们的问题是预测 Rossman 商店未来的销售额，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的日期、营业额、客户人数、是否在促销期、是否在节假日外的特征。测试集具备相同的特征，但是需要我们预测其营业额数据。

店铺的信息包括店铺的促销期，竞争对手的情况，以及店铺的类型和级别数据。

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。

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：

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

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、 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。

5、 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。

6、 模型调参：我会采用网格搜索作为调参的主要方式。

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/