预测 Rossmann 未来的销售额

开题报告

Denny

2019/04/11

项目背景

随着数据信息化的迅猛发展，数据量呈爆炸性增长。如何合理有效地利用企业掌握的数据为企业决策服务成为各行业关注的焦点[1]。市场预测即依据过去的与现在的市场信息，利用现有的经验、知识以及方法对未来市场发展趋势进行推测与估计。市场的预测是企业计划的重要制订标准之一，可帮助企业降低风险，为企业提供科学的经营决策。早在 70 年代，计算机已经开始被应用于销售额的预测，计算机信息处理与传统的方法相结合，使用定量方法，以趋势及其变化为重点，依靠历史数据，建立线性或非线性函数完成预测。欧洲 Rossmann 连锁药店的数据作为本项目的数据，通过对原始数据的清洗和可视化分析，挖掘隐藏于数据背后的特征，进行特征提取。Rossmann 是欧洲的一家连锁药店，在 7 个欧洲国家拥有 3,000 家药店。商店销售受到诸多因素的影响，包括促销，竞争，学校和国家假日，季节性和地点。 成千上万的个人经理根据其独特的情况预测销售量，结果的准确性可能会有很大的变化。

问题描述

Rossmann 是欧洲的一家连锁药店。 在这个源自 Kaggle 比赛 Rossmann Store Sales 中，需要根据 Rossmann 药妆店的信息（比如促销，竞争对手，节假日）以及在过去的销售情况，来预测 Rossmann 未来的销售额。

与波士顿房价预测、寻找捐助者、科赛中提供营销方案这些项目类似，在给定的训练集中都给出了在测试集所需要预测的部分，属于有监督的回归问题

数据与输入

原始数据从 Kaggle 的 Rossmann Store Sales 比赛中获得 https://www.kaggle.com/c/4594/download-all，包含了 1115 家 Rossmann Store 的历史销售记录，共四个文件，分别是：

train.csv 训练集，包含 1.02m 条销售的历史数据，时间跨度为 2013/01/1—2015/7/31

test.csv 测试集包含 41.1k 条销售的历史数据，时间跨度为 2015/08/1—2015/9/17

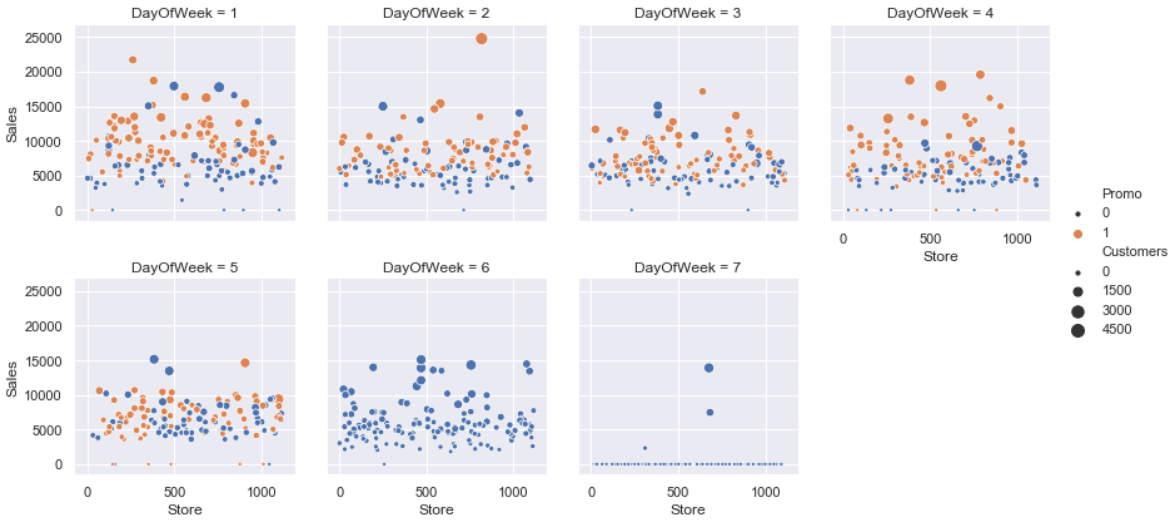
sample_submission.csv 正确的输出格式的样本

store.csv 关于商店的补充信息

数据域包括：

- Id - 表示测试集中的商店与日期
- Store - 每个商店有唯一的 ID
- Sales - 任何一天的营业额（需要预测的部分）
- Customers - 某一天的客户数量
- Open - 商店是否开放的指标: 0 = 关闭, 1 = 开放
- StateHoliday - 国家节假日指标。 一般除极个别商店外所有的商店都在国家节假日关闭. 所有学校在公众假期和周末都关闭. a =公共假期, b =复活节假期, c = 圣诞节, 0 = 无
- SchoolHoliday - 表明商店是否受到公立学校关闭的影响
- StoreType -区分 4 种不同的商店模型: a, b, c, d
- Assortment - 描述了一个分类级别: a = 基础, b = 额外, c = 扩展

- CompetitionDistance - 距离最近的竞争对手商店的距离
- CompetitionOpenSince[Month/Year] - 给出了最接近的竞争对手开业时间的大致年份和月份
- Promo - 表示商店当天是否正在营销促销
- Promo2 - Promo2 是一些商店持续不断的促销活动。: 0 =商店未参加, 1 = 商店正在参加
- Promo2Since[Year/Week] - 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。



Sales 预测数据可视化

根据“星期”、“是否进行促销”、“顾客量“三个属性对 Sales 进行了初步可视化，横轴为店铺，纵轴为销量。可以看出顾客量与销售额成正比，促销能够提高销售额，而周六没有店铺促销，周日几乎所有店铺休息等信息。

在训练集中并不存在缺省值，但在 StateHoliday 这一列存在非 string 类型的 0，需要进行类型统一处理。在测试集中 Open 列存在 11 个 NaN，后续需作填充或删除，StateHoliday 存在列与训练集同样的问题，作同样的处理即可。

使用 train.csv 对模型进行训练，然后使用 test.csv 测试模型并输出结果。但 train.csv，test.csv 都和 store.csv 分离，所以要得到好的结果需要 store.csv 分别与 train.csv，test.csv 结合

基准模型

考虑到这个项目与波士顿房价项目具有一定相识性，同样是回归问题，并且波士顿房价项目作为入门项目，其中所用到的模型简单易于理解，所以决定选用决策树作为基准模型。

在通过决策树得到相对好的成绩后考虑使用 Kaggle 中最热的 XGboost 模型进行预测

评估标准

根据 kaggle 官方的测评指标对于该任务这里只能使用 rmspe 进行测评，对于测试集只能获取 rmspe 的分数。

项目设计

第一步：对数据进行探索性的分析，为之后的处理和建模提供必要的结论，用 pandas 来载入数据，用 matplotlib 或 seaborn 可视化数据以便更好的理解数据。

第二步：数据清洗，对缺失数据、异常值等非正常数据进行处理，降低数据集噪音。

第三步：对必要的分类变量进行转换，如 string—>num，对数据进行归一化处理

第四步：数据特征分析，分布分析，分布分析揭示数据分布特征和分布类型；对比分析，绝对数比较，相对数比较；统计量分析，均值、中位数、众数等；相关性分析，散点图、散点图矩阵、计算相关系数；

第五步：划分模型，使用随机划分，之后开始训练基准模型，并获取预测结果

第六步：按照第六步划分模型，训练 xgboost 模型并获取结果

第七步：优化模型，使用特征工程，提取有用的信息，挖掘更深层次的特征，利用已有属性集构造出新的特征属性，加入现有属性集合中，提高结果的精度。

第八步：模型评价，除了测试集的 rmspe 值外，从特征重要性表等角度出发，进行 PCA 并且将其可视化。

[1] 赵啸彬，基于数据挖掘的零售业销售预测[D].上海：上海交通大学，2010