开题报告

张承博
2018 年 12 月 27 日

项目背景

Rossmann 公司在 7 个欧洲国家经营超过 3000 家连锁药店。目前，Rossmann 的商店经理需要提前预测未来 6 周的日销售情况。商店的销售情况受多个因素的影响，如促销、竞争、节假日、季节性和地方性。由于数千个商店经理基于其独特的环境预测销售情况，预测结果的准确性可能会有很大差异。可靠的销售预测能够使商店经理创建有效的员工时间表，从而提高其生产率和动力。

机器学习使用计算方法直接从数据中“学习”信息，对未见数据做出预测，而不依赖于预定方程模型。当可用于学习的样本数量增加时，这些算法可自适应提高性能。机器学习采用两种技术：监督式学习和无监督学习。监督式学习根据已知的输入和输出训练模型，让模型能够预测未来输出；无监督学习从输入数据中找出隐藏模式或内在结构。

如今，基于机器学习技术对销售预测的研究逐渐成熟，较传统数据分析方法其预测能力更加强大，同时具有更高可靠性。针对该领域涉及大数据和多变量的复杂任务或问题，可以考虑使用机器学习作为解决途径。

问题描述

Rossmann 希望借助稳健模型预测其在德国境内 1115 家商店未来 6 个月的日销售情况。
Rossmann 已提供 1115 家商店历史销售情况的数据集，该数据集包含特征（例如：顾客量、商店类型）和标签（即销售额）。因此，该问题是一个基于多特征的监督式回归预测问题，也是可被量化、可被测量且可被重现的。机器学习（监督式学习）中的回归方法可能是相关的潜在解决方案。

数据集和输入

提供的数据集文件包括：

- *train.csv* - 包含销售额（标签）的历史数据，用于训练
- *test.csv* - 不包含销售额（标签）的历史数据，用于测试
- *sample_submission.csv* - 正确格式的提交文件样本
- *store.csv* - 关于商店的额外信息

训练集包含从 2013-01-01 至 2015-07-31 共计 1017209 条记录，测试集包含 41088 条记录。

数据集特征字段包括：

- *Sales* - 给定日期当日销售额，即标签

- *Customers* - 给定日期当日顾客量
- *Open* - 表明商店是否开门营业：0 = 关门，1 = 开门
- *StateHoliday* - 法定假日类别：a = 公众假日，b = 复活节，c = 圣诞节，0 = 非假日
- *SchoolHoliday* - 表明是否受到学校关闭影响
- *StoreType* - 商店类型：a，b，c，d
- *Assortment* - 零售商品品类级别：a = 基础，b = 附加，c = 扩展
- *CompetitionDistance* - 最近竞争者距离
- *CompetitionOpenSince[Month/Year]* - 最近竞争者开始经营的大概时间（月/年）
- *Promo* - 表明给定日期当日是否有促销活动
- *Promo2* - 表明是否正在参与连续促销活动：0 = 未参与，1 = 正在参与
- *Promo2Since[Year/Week]* - 开始参与连续促销活动的时间（年/周）
- *PromoInterval* - 连续促销活动开始的月份

一些字段如样本编号 *Id* 和商店编号 *Store* 无益于预测模型构建而会影响模型泛化，因此它们需要被忽略。上述数据集包括：分类数据，如 *Open*、*StateHoliday*、*SchoolHoliday*、*StoreType*、*Assortment*、*Promo*、*Promo2*；数值数据，如 *Sales*、*Customers*、*CompetitionDistance*；时序数据，如 *CompetitionOpenSince*、*Promo2Since*、*PromoInterval*。分类数据代表定性的样本属性（例如，商店是否开门营业），是离散的；数值数据代表定量的样本属性（例如，当日客户量），是连续的；时序数据在后续过程中可能需要被处理，转换成分类数据。

经初步探索可知训练集在 *CompetitionDistance*, *CompetitionOpenSinceMonth*, *CompetitionOpenSinceYear*, *Promo2SinceWeek*, *Promo2SinceYear* 和 *PromoInterval* 字段存在缺失值；在 *Sales*、*Customers*、*CompetitionDistance* 字段存在异常值。

解决方案描述

本项目将采用 XGBoost 构建模型，XGBoost 是一种为执行速度和模型性能而设计的梯度提升方法（gradient boosting）的实现。它提供了并行树提升（也称为GBDT，GBM），可以快速准确地解决许多数据科学问题。考虑到 XGBoost 适合解决大型结构化数据集的回归预测问题，故引入 XGBoost 作为模型构建工具。对于数据输入方面，需清洗数据，插补缺失值，删除异常值，归一化数值特征，编码分类特征等。

基准模型

为对比本项目提出的解决方案和现有解决方案，可利用 Rossmann 发起的 Kaggle 相关竞赛的参与者的现有模型。

评估指标

根据 Kaggle 相关竞赛规则，提交的预测模型均采用 RMSPE（Root Mean Square Percentage Error）评估。RMSPE 计算方法如下：

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

其中 y_i 表示单个商店单日实际销售额，$\hat{y}_i$ 表示对应预测值。

项目设计

由于 Python 编程语言在数据科学和机器学习领域的成熟支持与广泛应用，本项目将基于 Python 研究开发。项目流程如下：

1. **探索数据**
   通过数据分析方法探索数据集，包括如下方面：
   a. **导入数据**
      利用 Pandas 库导入数据集文件至数据帧（DataFrame），联合多个数据源以令其合理化到同一数据集中。
   b. **检查数据**
      指定特征。检查统计量、缺失值和异常值。
   c. **可视化数据**
      利用 Seaborn 或 Matplotlib 库可视化数据并观察数据特点和趋势。
2. **预处理数据**
   清洗数据，处理缺失值和异常值。重新格式化某些特征数据，标准化或归一化数值特征，编码分类特征（如 *StateHoliday*）。
3. **训练和验证模型**
   利用 XGBoost 的 *XGBRegressor* 类在处理后数据集上训练模型。通过网格搜索（GridSearchCV）、交叉验证技术及 RMSPE 评估指标调整超参，选择最优化模型。输出并分析特征重要性（feature importance），进行特征选择。因数据集具有时序性，需通过时序性（而非随机）方式分割数据集。该阶段将应用 XGBoost 与 scikit-learn 机器学习库。
4. **测试和评估模型**
   使用最终模型生成测试集的预测结果并将其提交至 Kaggle，查看分数和排名。分析结果，进一步调整优化模型，提高性能。
5. **作出结论**
   回顾整个项目流程，对解决方案进行总结。

引用

1. *什么是机器学习,* https://ww2.mathworks.cn/discovery/machine-learning.html.
2. *Rossmann Store Sales*, www.kaggle.com/c/rossmann-store-sales.
3. *A Gentle Introduction to XGBoost for Applied Machine Learning*, https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning.
4. *Introduction to Boosted Trees*, https://xgboost.readthedocs.io/en/latest/tutorials/model.html.
5. *Predictions with XGboost and Linear Regression*, https://www.kaggle.com/mburakergenc/predictions-with-xgboost-and-linear-regression .
6. *Machine Learning Workflow*, https://cloud.google.com/ml-engine/docs/tensorflow/ml-solutions-overview.
7. *Time Series cross-validator*, https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html.