机器学习工程师纳米学位

开题报告

优达学城

2018-12-26

预测 Rossmann 未来的销售额

1. 选题背景

随着经济全球化的发展，企业面临着更加复杂和残酷的市场竞争。能够快速准确的预测出来销售额从而合理的安排生产和库存，用低成本的产品快速满足客户要求成为企业关心的重点。传统的销售预测方法分为定性和定量两类，定性方法主要有市场调研、购买者期望分析、专家小组法等，定量方法主要有平均数趋势预测、因果预测分析、时间序列分析法等统计方法。随着大数据和人工智能技术的兴起，机器学习模型给销售额的预测带来了新的思路。

2. 问题描述

问题源自 Kaggle 竞赛，为欧洲的一家连锁药店 Rossmann 预测未来的销售情况。Rossmann 在欧洲的 7 个国家拥有 3000 多家连锁药店。需要帮助他们的管理者，对位于德国的 1115 家药店提前 6 周预测日销售额。可靠的预测值能够帮助他们制定有效的员工时间表，从而提高生产效率和积极性。

3. 数据说明

Kaggle 提供的数据集有三个，包含 1115 家店铺的基本信息 store 表（店铺类型、品类、竞争对手的距离及开业时间、是否连续促销及促销时间）、训练数据集 train 表（店铺编号、日期、星期数、当日销售额、客户数、开业状态、假期状态等），测试数据集 test 表（店铺编号、星期数、日期、开业状态、促销状态、假期状态等）。需要我们根据训练数据集和店铺的基本信息情况，预测出测试数据集中店铺在给出的日期和促销状态下的销售额。

4. 解决方案

首先观察数据的原始特征，根据数据特征做数据清洗、融合等工作，然后进行数据探索，通过可视化的工具查看数据了解数据特征，根据数据特征进行必要

的数据预处理。

将处理好的数据分割成训练集和验证集，并根据预测目标为模型选择合适的评价指标，参照题目可采用“均方根百分比误差（rmspe）”这个指标来衡量模型优劣。

根据训练数据的特征、维度、预测目标等选择合适的模型范围进行模型测试，可以考虑构造模型测试流水线进行模型选择。部分带有特征排序或选择的模型可以我们特色优化提供思路，帮助我们调整数据特征。

根据选择模型的实际情况，结合 rmspe 得分，进行模型调优，并将结果上传至 kaggle 提交页面，检测模型结果，直至达到预期要求。

5. 基准模型

本问题的最终目标是预测未来销售额，属于回归问题，解决回归问题可以考虑逻辑回归、SVR 模型，如果模型效果不理想可以考虑采用集成学习的模型来实现预测目标。

6. 评估指标

应题目要求采用 rmspe 指标来评价模型的预测效果，公式如下：

$$\text{rmspe} = \sqrt{\frac{(\frac{y_{pre}}{y}-1)^2}{n}}$$ 其中，n 为样本数量。

7. 方案设计

第一步：识别问题，剔除无效数据，根据数据的基本情况进行清洗和填充。

第二步：对训练集数据进行特征和标签的分离，并利用统计方法和线箱图、散点图等可视化方法进行数据探索，观察数据分布和统计特征。

第三步：根据数据特征进行数据预处理，对高偏度的特征进行转换、根据特征的数据范围进行数据缩放、并将类别特征进行 one-hot 编码转换。

第四步：将数据进行训练集和验证集的划分，构造模型训练的流水线，并定义模型平均指标。

第五步：利用模型训练流水线，采用逻辑回归、SVR、XGboost 等模型训练数据，并用验证集评估效果，根据 feature_importance 或 feature_selection 等属性，辅助进行特征的选择和优化。

第六步：结合第五步的结果，对选定模型和特征进行调参优化，直至 rmspe 满足 kaggle 排序要求为止。