

机器学习工程师 纳米学位

开题报告

毕业项目：预测 Rossmann 未来的销售额

一、项目背景

Rossmann 是欧洲的一家连锁药店，目前一共拥有大约3,000家药店，它们分布在 7 个欧洲国家。现在的任务是通过历史数据，预测 6 周后的每日销售量[1]。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。每个人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。

可靠的销售预测对门店的精细化运营具有非常大的帮助。销售量的预测能够使商店的管理人员创建有效的员工时间表，提高生产力。此外，良好的销售预测模型，还可以让管理人员调整供应链，制定合理的促销策略与竞争策略。另外还可以帮助门店，提前准备合适的人力资源和物资数量，降低成本，提高营业额以及用户体验。

二、问题描述

本项目所需要解决的问题是：通过 Rossmann 给出的历史数据集，创建一个销售额预测模型，用来预测德国各地 1,115 家店铺的 6 周销量，来帮助 Rossmann 获得更大的收益。

该问题的解决，可以从机器学习和数据挖掘的角度思考，分析数据的自变量（即各种属性）和因变量（即销售额），进行训练，最后得到销售量预测模型。Rossmann 给出的数据集比较丰富，是一个明显的有监督的回归问题。

单个商店的销售会受到很多因素的影响，包括是否促销，附近有没

有竞争，是否是国家假日，季节性和地点等待因素。这些特征都有可能对销售量造成影响。

三、数据集与输入

本项目主要用到了 3 个数据集，并有 1 个数据集作为提交数据集的参考样本，它们来自 Kaggle。

- (1) **train.csv** --- 历史数据包括 sales 的数据
- (2) **test.csv** --- 历史数据不包括 sales 的数据
- (3) **store.csv** --- 提供各个店铺的具体信息的数据
- (4) **sample_submission.csv** --- 一个最后提交数据集的参考样本

其中特征包含：Store 的 ID、日期、当天顾客的数量、是否促销、是否是州放假、是否是学校放假，其中猜想节假日，促销和每周的某天是影响的最大因素。

对数据进行处理之前，需要经历两个阶段：

- (1) 数据探索，弄清楚每个属性的含义，并对数据进行基础统计；
- (2) 预处理，包括归一化、剔除异常值、特征的重编码等。

四、解决方案陈述

特征工程是不可或缺的，好的特征工程对模型的精确度至关重要，集成学习也是很重要的，为了得到更好的预测模型，可以运用好优秀的集成学习方法。我们在实际运用中需要在提高分数的同时又降低过拟合的风险，这个需要用到不少技巧，尤其是大多数单个模型很难实现 **test error** 的持续下降。

本项目采用的集成学习为 XGBoost，优点是高效、灵活，已经被广泛使用，并且采用 CV, K-Folders 等交叉验证的方法确保预测的准确性。

五、 评估指标与基准模型

用 pandas 来载入数据，并 matplotlib 和 seaborn 提供的绘图功能做一些简单的可视化来理解数据。其中分类问题的常用评估指标有 Accuracy, Precision, Recall, F score 等；回归问题有 MAE, MSE, RMSPE 等。由于该项目解决的是回归问题，所以采用与 Kaggle 一样的评估指标，RMSPE，即均方根误差。目标是，XGBoost 得到的模型的 RMSPE 小于 0.11773。计算公式如下：

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

六、 项目设计

第一步：业务理解与数据获取：该项目的任务是监督学习的回归，从Kaggle获取数据集。

第二步：数据探索：通过简单的描述统计与可视化来理解数据。

第三步：数据预处理：处理数据缺失、异常值等问题。

第四步：特征工程：将某些特征中的数值进行转换，更好地进行挖掘。

第五步：模型训练：将前面准备好的数据用来训练模型。

第六步：模型评估：评估模型的好坏，如果表现不错则走到下一步，否则从头开始分析问题，重新改进模型。

第七步：预测未知数据：用来预测未来的销售额。

附：参考文献

[1] Kaggle Rossmann Store Sales Overview:
<https://www.kaggle.com/c/rossmann-store-sales>