机器学习纳米学位毕业项目

开题报告

项目背景

本次毕业项目的选题是 Rossmann 销售预测，来源于 Kaggle 比赛平台的项目 Rossmann Store Sales。Rossmann 是一家遍布于欧洲 7 个国家的药妆公司并拥有 3000 余家门店。目前 Rossmann 公司要求各门店经理预测未来六周的每天营业额，从而可以使门店经理们根据预测制定有效的员工排班，从而提高门店效率和激励员工。由于各家门店经理单独进行销售预测，方法不尽相同，预测的准确性也是相差很大。现在公司提供了过去 3 年的各家门店的历史每日营业额和影响门店销售的一些因子比如促销，竞争对手，学校和法定假日等数据，希望有一个可靠的预测模型来帮助所有门店经理更加合理的预测销售。这是一个典型的可以用机器学习的监督模型来解决的预测问题。由于包含了时间序列数据，在建模上会与课程中学到的监督学习模型有所不同，所以想尝试挑战一下如何解决时间序列数据的机器学习模型问题。

问题描述

本项目要预测的是位于德国的 1115 家门店的未来六周的每天每店的销售额。具体来说要预测时间段 2015/08/01~ 2015/09/17 内的每天每店的销售额。销售额的单位和提供的训练数据相同。预测数据文件参考 test.csv 文件，

数据集

数据集可以从 Kaggle 上获得。包含了 4 个文件：

train.csv - 包含了历史的销售 1115 家店的销售数据，这份数据作为建模的训练数据，同时留出最后 6 周的数据作为验证数据。训练数据集的记录数在 100 万条左右

test.csv – 包含了要测试的 1115 家店和日期的基本数据，没有销售额。这是项目的测试数据，测试通过在 Kaggle 上提交预测结果来评判。测试数据在 4 万条左右

sample_submission.csv - 这是在 Kaggle 提交测试数据的模板

store.csv- 1115 家门店的相关信息，比如竞争对手，促销，节假日等。这些都是影响门店销售的变量，是模型的输入。

下面是一些主要字段的描述：

Id – 序列号，这个字段没有用途

Store – 门店号，由于各家门店的销售额趋势不同，Store Id 也是模型的重要参数

Sales – 销售额，这是模型要预测的数据，也就是 Y

Customers – 客流量，这个数据应该和 Sales 高度相关，由于在测试数据中不存在，应该也是一个 Y，但是本项目不考虑预测此数据

Open - 0 表示不营业，1 表示营业

StateHoliday – a 表示公共假日，b 表示复活节，c 表示圣诞节，0 表示非节日

SchoolHoliday – 表示门店是否受学校放假的影响

StoreType – 门店的 4 种类型：a, b, c, d

Assortment - 表示门店的 3 个分类层次: a, b, c

CompetitionDistance - 最近竞争对手门店的距离

CompetitionOpenSince[Month/Year] – 竞争对手门店的开店日期

Promo – 表示当日门店是否有促销

Promo2 – 表示门店是否参与连续促销活动

Promo2Since[Year/Week] - 表示门店开始参与 Promo2 的日期

PromoInterval – 表示 Promo2 开始的月份

解决方案

据前面所述，本项目是根据过去的历史销售数据来预测未来的销售，这是一个基于大量数据的预测问题。由于销售数据是基于时间序列的数据，一般可以利用统计学上的时间序列模型来初做初步的预测。时间序列模型主要有 ETS 和 ARIMA 两种模型，ETS 模型是基于数据的趋势和周期性的变化规律，而 ARIMA 模型是基于描述数据的自相关性。要做进一步的预测，可以利用机器学习方法的监督学习方法来建模和预测。本项目的数据集中除了时间序列数据之外，还有一些其他特征，比如门店的促销，竞争对手，节假日等信息，这些都可以作为监督学习的输入特征，输出是每天的销售额。预测目标销售额是一个连续型的数字，就要用到机器学习里的回归模型。回归模型从简单的线性回归，岭回归，决策树回归，到复杂的集成方法比如随机森林，梯度提升回归算法（GBRT, XGBOOST）等，都可以用于给定数据的回归模型训练和预测。根据 Kaggle 上的历史表现，集成的回归算法的效果在监督学习的案例中普遍不错，我们也将以集成回归算法为主来解决本项目的销售额预测问题。

基准模型

在 Kaggle 上该项目有 3303 位参与者，其提交的模型预测结果在 Private LeaderBoard 上的 RMPSE 评分在 0.10021 到 1.0000 之间，中间值是 0.12967，考虑到预测中的随机因素，我们预测模型的结果以 0.12967 为基准。

评价指标

本项目的预测结果是销售额，销售额的准确率以 RMPSE（Rooted Mean Percentage Squared Error）来衡量。

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

其中 y_i 是某店在某日的实际销售额，$\hat{y}_i$ 是对应的预测值，RMPSE 的 0.0~1.0 之间，RMPSE 值越小，准确率越高。如果任一店的任一日实际销售额 y_i 为 0，则不计入公式中。

项目设计

项目的分析流程可以按照 CRISP-DM 的框架进行，CRISP-DM 是跨行业的数据挖掘标准流程。分为六个阶段：

1）业务理解 – 明确项目背景和需要解决的问题，下载数据集，明确项目的评价标准，将业务问题转化机器学习问题
2）数据理解 – 对数据进行探索分析，包括数据的统计分析，缺失值，相关性，分布形态和异常值等
3）数据预处理 – 包括缺失值的插补，异常点的处理，数据的转换，标签数据的编码，特征的创建，训练集和验证集的划分。
4）数据建模 – 回归模型的初步选择，建模和比较，特征的加工，考虑选择的模型包括随机森林，梯度提升回归树，XGBoost 算法
5）模型优化，验证与评价 – 对最终选定模型的参数优化，特征的重要性评估和选择，对验证集的数据进行验证与评分
6）模型测试 – 对测试集的数据进行预测，并提交预测结果到 Kaggle 上评分。

参考文献

1 Rossmann Store Sales, Kaggle link: https://www.kaggle.com/c/rossmann-store-sales/data

2 OTexts, ARIMA vs ETS, https://www.otexts.org/fpp/8/10

4 How to Rank 10% in Your First Kaggle Competition, https://dnc1994.com/2016/05/rank-10-percent-in-first-kaggle-competition-en/