Udacity MLND 毕业项目开题报告

学员：周慧明(james-zhou@foxmail.com)

项目：Rossmann 销售预测

一、项目背景

Rossmann 销售预测项目的目标是基于三年的销售数据来构建可以预测销售量的模型。销售额的提前精准预测对门店的精细化运营具有非常大的意义，包括提前配备合适的人力资源和物资数量，降低运营成本，提升营业额和用户体验。因为我从事新零售相关的工作，所以决定选择此课题作为毕业项目。

对于该问题有多种思考角度，一种方法是从时间序列（Forecasting Time-series data）角度思考，这是药店的销售额具有季节性。一般的问题可以使用 ARIMA(Auto-Regressive Integrated Moving Averages)，以及 facebook 的 Prophet [1] 进行建模，而更复杂的问题可以用循环神经网络如 LSTM[2]进行预测。

第二种是从机器学习的角度思考，分析数据的自变量（x：时间点/门店类型/促销等）和因变量（y：销售额），对数据做特征工程，选择相应的机器学习模型，得到最终的销售量预测模型。

时间序列方法和机器学习方法都能对流量预测的问题进行预测，在 kaggle 比赛‘’中，有人对这些方法进行了对比[3]。而在本次 Rossmann 销售预测项目中，有利用时间序列进行预测的模型[4],也有利用机器学习模型进行预测的模型[5]。

二、问题描述

Rossmann 拥有遍布在欧洲七个国家的 3000 个药店，该项目的目标是基于德国 1115 家药店三年的销售数据来构建可以预测销售量的模型，方便 Rossmann 进行提前六个星期的预测。

单个药店的销售额会被很多因素影响，包括是否有促销，周围的竞争者信息，学校和国家的假期，季节和位置。基于本项目中自变量（特征）很多，而这些特征对销售额有较为重要的影响，所以在本论文中使用机器学习方法建立销售额预测的模型。

三、输入数据

本次项目数据来源于 kaggle,总共有三组数据，训练集、测试集以及门店数据。

训练集主要记录了德国 1115 家药店在 2013/1/1~2015/2/20 的数据，测试集则记录了这些门店 2015/8/1~2015/9/17 的数据，其特征如下：
Store/DayOfWeek/Date /Open/Promo/StateHoliday/SchoolHoliday

门店数据则记录了 1115 家药店的门店特征数据，其特征如下：Store/StoreType/Assortment/CompetitionDistance/CompetitionOpenSinceMonth/CompetitionOpenSinceYear/Promo2SinceWeek/Promo2SinceYear/PromoInterval.

使用训练集和门店数据的时候要走以下数据分析流程：

1. 数据探索，对数据进行基础统计运算，包括计算最大小值，中位数等等；
2. 预处理，包括归一化、剔除异常值（包括很多门店没有营业的状况），独热编码等；
3. 对训练集数据进行切分，按照时间顺序，前 80%的数据用于训练，后 20%的数据用于验证。

四、解决办法

本项目拟采用机器学习的集成学习（**Gradient boosting**）方法来建立模型， 其通过加入新的弱学习器，来努力纠正前面所有弱学习器的残差，最终这样多个学习器相加在一起用来进行最终预测，准确率就会比单独的一个要高。

作为集成学习的一种方法，本项目采用 XGBoost[6],其特点就是**计算速度快，模型表现好，在各种数据挖掘大赛中有不错的表现，**多家公司已经在生产环境中使用。

除了选择 XGBoost 模型外，还需要做特征工程，即对训练集数据的特征进行筛选和工程，如时间点/门店类型和编号/促销/竞争情况/顾客数量等，其标签是销售额。

五、评估指标

此项目的评估标准跟 kaggle 比赛的评估标准一致，即使用预测准确度用均方根误差(RMSPE)来衡量，它是该模型在测试集里面的预测值与真值偏差的平方和观测次数 n 比值的平方根，具体如下：

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

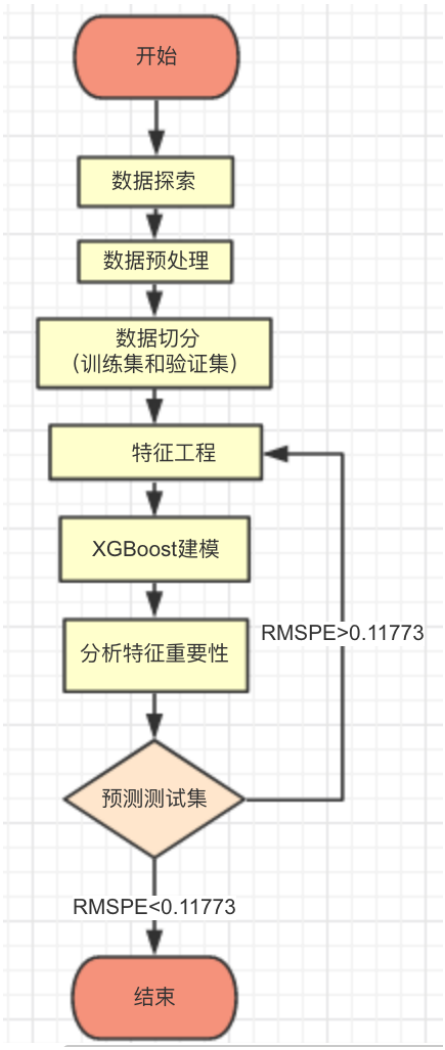
其中 n 代表预测的样本数量，y_i 代表某门店的一天的销售额，而 yhat_i 则代表模型的预测值。

学习得到 xgboost 模型之后用测试集的数据进行测试其 RMSPE 误差，对模型预测的准确度进行量化和评估。

六、基准模型

基准的模型是该门店销售额的中位数，用于跟 xgboost 模型进行对比。而 xgboost 模型的预测目标是其 RMSPE 误差进入 kaggle 比赛排行榜的 10%（总共 3303 个参赛者），即 RMSPE<0.11773。

- 七、设计大纲



- 八、参考文献

[1] prophet: https://facebook.github.io/prophet/
[2] Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras: https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/

[3] Predictive Analysis - Web Traffic Time Series Forecasting | Kaggle: https://www.kaggle.com/zoupet/predictive-analysis-with-different-approaches
[4] Time Series Analysis and Forecasting with Prophet: https://www.kaggle.com/elenapetrova/time-series-analysis-and-forecasts-with-prophet/notebook

[5] Model documentation 1st place: https://www.kaggle.com/c/rossmann-store-sales/discussion/18024

[6] XGBoost: https://github.com/dmlc/xgboost