**Machine Learning Engineer Nanodegree**

## Rossman 销售预测开题报告

战柏瑞  
2018年

### 主要背景

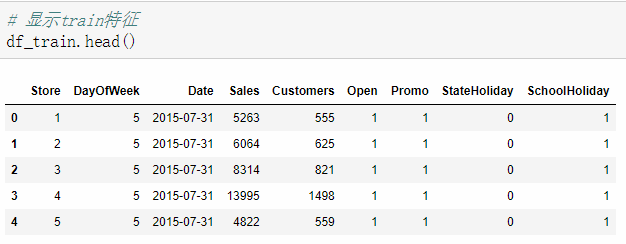
### 该项目是Kaggle上Rossmann公司举办的一个竞赛项目.Rossmann是德国第二大药品销售链,在欧洲7个国家拥有近3600家药店.公司由Dirk Rossmann建立于1972年[[1]](#endnote-1).在2015年Rossmann 的管理层被指派预测近6周的日销售额.销售额会被很多因素印象例如, 促销,竞争对手,学校以及州节假日,季节性,和地域性.[[2]](#endnote-2)

### 问题称述

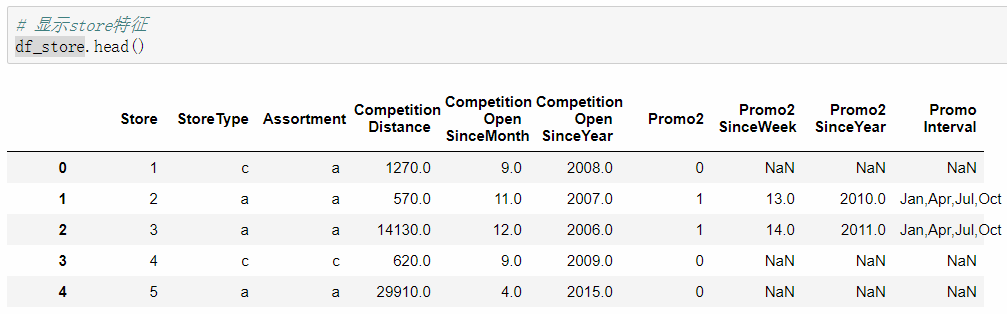
在分类中，我们想了解模型隔多久正确或不正确地识别新样本一次。而在回归中，我们可能更关注模型的预测值与真正值之间差多少。[[3]](#endnote-3) 通过对Rossmann数据分割,得到35天的销售数据,然后使用xgboost进行回归预测,因为此次项目更关注预测值和真实值之间的差,所以是一个回归问题.

### 数据集与输入

使用Kaggle website (<https://www.kaggle.com/c/rossmann-store-sales>)的数据,将Train.csv进行分割,从而获得最近的35天数据集作为测试集.



训练集具有9个特征,其中Sales作为结果应从特征中分离出来. Date特征中需要从字符串处理为年月日三个不同的特征.对于StateHoliday进行one-hot编码.最后一共有13个特征值.



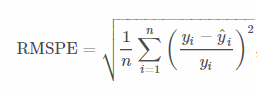
Store数据集是训练集的衍生部分,通过共有10个特征值,对其进行one-hot编码,最后有17个特征值.对于缺省值,competition distance可以用平均值meadian代替,别的缺省值由xgboost自行处理(默认处理0)[[4]](#endnote-4). 然后对store数据集和训练集以Store Id对应的方式进行合并填充

### 解决方案

此次项目通过XGboost来完成. XGBoost实现的是一种通用的Tree Boosting算法，此算法的一个代表为梯度提升决策树（Gradient Boosting Decision Tree, GBDT），又名MART(Multiple Additive Regression Tree)。GBDT的原理是，首先使用训练集和样本真值（即标准答案）训练一棵树，然后使用这棵树预测训练集，得到每个样本的预测值，由于预测值与真值存在偏差，所以二者相减可以得到“残差”。接下来训练第二棵树，此时不再使用真值，而是使用残差作为标准答案。两棵树训练完成后，可以再次得到每个样本的残差，然后进一步训练第三棵树，以此类推。树的总棵数可以人为指定，也可以监控某些指标（例如验证集上的误差）来停止训练。[[5]](#endnote-5)

### 评估标准

通过对测试集的 Root Mean Square Percentage Error (RMSPE)对已知模型进行评价.

[[6]](#endnote-6)

Yi为单个门店单天的销售 Yihat是对此进行的预测

### 基准模型

### 设定基准阈值为kaggle 排行榜前 10%（330/3303），也就是在 Private Leaderboard 上的分数要低于 0.11737。[[7]](#endnote-7)

### 项目设计

* 数据预处理

此阶段是对数据进行分割,以产生本地测试集以及训练集,对Store数据集进行合并,以及简单的时间转化成年月日,对其他特征进行one-hot编码.

* 模型搭建

此阶段对特征的重要性进行排序,去掉不必要的特征,还有重复特征.

* 模型训练

训练模型,对训练集进行分割,从而训练N次,每次对模型进行修正.

* 模型调参

通过参照XGboost API[[8]](#endnote-8)进行调参

例如,max\_depth尝试不同深度对模型的影响,从而选定合理的数值.

Learning\_rate,调整学习率,太大的学习率可能无法收敛.

n\_estimators:调整有几个boosted tree.

以及其他[[9]](#endnote-9)

* 模型评估

同评估标准

* 可视化
* 对必要以及可观的数据进行可视化,从而更好的了解数据

**参考文献**

1. <https://en.wikipedia.org/wiki/Rossmann_(company)> [↑](#endnote-ref-1)
2. <https://www.kaggle.com/c/rossmann-store-sales#description> [↑](#endnote-ref-2)
3. Udacity,机器学习(进阶)-课程12评估指标-分类指标与回归指标 [↑](#endnote-ref-3)
4. <http://xgboost.readthedocs.io/en/latest/faq.html?highlight=missing> [↑](#endnote-ref-4)
5. <http://www.a-site.cn/article/714295.html> [↑](#endnote-ref-5)
6. <https://www.kaggle.com/c/rossmann-store-sales#evaluation> [↑](#endnote-ref-6)
7. <https://www.kaggle.com/c/rossmann-store-sales/leaderboard> [↑](#endnote-ref-7)
8. <http://xgboost.readthedocs.io/en/latest/python/python_api.html> [↑](#endnote-ref-8)
9. [http://blog.csdn.net/sb19931201/article/details/52557382](%20http:/blog.csdn.net/sb19931201/article/details/52557382) [↑](#endnote-ref-9)