# 机器学习工程师纳米学位

## 毕业项目

优达学城

2019-1-11

## 预测Rossmann未来的销售额

### Ⅰ问题的定义

#### 项目概述

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。

本次项目问题源自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，基于历史数据对位于德国的1115家药店预测未来6周的销售额。项目主要涉及三个数据集，包含店铺基本信息的store.csv，共1115个店铺开店情况、竞争对手情况、促销情况的数据；train.csv，包含1017209从2013年1月年至2015年7月1115个店铺每天的销售额、用户数等数据；test.csv，包含41088条从2015年8月1日至2015年9月17日间每天的假期状态、每个店铺的促销状态等数据。我们需要借助store表和train表的数据构建预测模型，再利用test表和store表结合的数据，预测test表中列出的店铺在当日的假期及促销条件下会产生的销售额。

#### 问题陈述

本项目是一个回归预测问题，目标是根据给出的数据信息，构建一个合适的预测模型，为店铺预测出具体某天的销售额。

为实现这一目标，首先，我们将通过数据探视了解数据的基本信息、分布情况，因在训练集中的数据有1017209条，是1115个店铺按时间序列记录的销售数据，在数据探视时，可能需要采用多维度统计分析及数据可视化的方法，全方位了解数据。在充分了解数据后，对于缺失数据、异常数据进行清洗规整，此外，为了扩充特征范围，可以考虑通过一定方法对原始数据加工产生新的特征。数据规整完成后，根据数值范围，对数据进行归一化或ont-hot转换。然后，构建数据模型训练流水线及评估指标，帮助我们在逻辑回归、SVR、XGboost等预测模型中选出最合适的模型，并对合适的预测模型进行特征和参数的优化，使其达到最优。最后，将利用最优模型对测试集进行预测，实现预测目标。

#### 评价指标

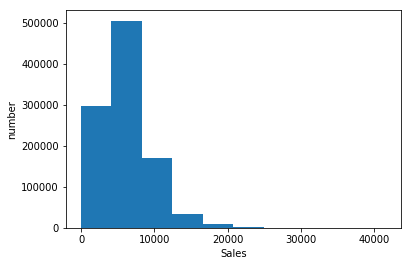
针对回归问题的评价指标通常有平均绝对误差（MAE）、平均平方误差（MSE）、均方根误差（RMSE）

### 数据说明

Kaggle提供的数据集有三个，分别是包含1115家店铺的基本信息store表（含店铺类型、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等字段）；包含1017209条数据的训练数据集train表（含店铺编号、日期、星期数、当日销售额、客户数、开业状态、假期状态等字段），样本取值从2013-01-01到2015-07-31；包含41088条日数据的测试数据集test表（含店铺编号、星期数、日期、开业状态、促销状态、假期状态等字段），预测区间从2015-08-01到2015-09-17。store表中为店铺的基本信息数据，计算时需要将store表融合到train和test中。

三张表的数据多为连续型，部分时间、类型、状态等字段是离散型数据，处理时可以进行one-hot转换。store表中，竞争对手距离为nan的数据项，其竞争对手开业时间均为nan，可用0填充月份用1900填充年份，竞争对手距离不为nan的数据项，其竞争对手开业时间为nan的，用频数最多的开业月份和年份值填充，另外，竞争对手距离数据值分布范围较广，要对数据值做对数转换。店铺连续促销开始年的nan值用1900填充，促销开始周nan值用0值填充，其他少量离散数据的缺失可以新增一个类型项做填充（如，0、1，缺失值可以用2填充），其他连续型数据的少量缺失值采用均值填充，日期数据计划作为连续型数据来处理。

根据题目要求，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。对训练数据集中的销售额字段进行基本分析发现，整体销售额最大值为41551，最小值为0，平均值为5773.8，75%的数据都小于7856，属于偏态分布。因销售额的统计是按每天每家店的维度统计的，在进行数据分析时还应考虑按店或按天的维度分布统计后的情况。



### 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。

### 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。

### 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：

其中，n为样本数量。

### 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，并结合原始数据情况加工新的特征。

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：将数据按照随机选择的方式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。

### 参考文献

[1] Hongyu Xiong (hxiong2), Xi Wu (wuxi), Jingying Yue (jingying), Drugs store sales forecast using Machine Learning.

[2] 特征选择<https://www.cnblogs.com/stevenlk/p/6543628.html>.

[3] 监督学习项目2：为CharityML寻找捐献者

<https://classroom.udacity.com/nanodegrees/nd009-cn-advanced/parts/e340f873-83cc-44aa-8530-bb78bf834995/modules/d1dd6eea-c2a7-421d-b1ce-7702c44aab31/lessons/9eccaf6a-09e5-4fdb-a32f-17d942669bc3/concepts/8465cfb8-6864-4ad0-b945-626c4297d0ed>

<https://blog.csdn.net/pipisorry/article/details/52574156>

<https://blog.csdn.net/tox33/article/details/81141485>