**机器学习（进阶）纳米学位**

**毕业开题报告**

**项目题目：Rossmann Sales预测**

**学员： 宣 扬**

**Email: *xuanyang84@gmail.com***

**2018.2.5**

# 项目背景

本毕业项目是来自Kaggle的竞赛题（2015年9月30日~2015年12月15日），项目目标是建立稳定的预测模型，为Rossmann公司在德国的1,115家商店未来6周的销售额进行预测。Rossmann公司是德国第二大的药品连锁商店，在全欧洲拥有超过3,600家商店【1】。

在应用机器学习预测模型之前，对销售额的预测工作是由上千个独立的经理人完成的，但由于销售额受促销、竞争者、假期、季节等因素影响，以及商店经理自身的环境和经验不同，导致预测的准确度变化很大，因此利用机器学习模型来进行该预测工作，可以在很大程度上提高预测的准确度和稳定性。

目前Kaggle上已有3,303只队伍参加了挑战，第一名是Gert，他所采用的模型和技术使其预测得分高达0.10021。

选择这个题目的主要原因是因为在我的实际工作中，有很多数据和问题需要通过回归方法来预测，例如对某组试验数据进行建模，从而预测未试验工况下的设备性能，为设备设计提供参考。通过这个项目的学习和训练，希望自己可以对机器学习中的回归方法有更深入的了解，对数据的处理、模型的调参和模型性能评估有更好的认识，让我更好的处理日常工作中的问题。

# 问题陈述

从描述来看，本项目是一个典型的数据回归问题。利用已积累的销售数据、商店特征、销售日特征等信息，通过对选择的回归算法进行训练，建立起商店销售额与训练数据中各项特征之间的回归预测模型，从而可以根据未来6周内的各种情况来预测销售额。

将利用验证数据来修正模型的各项超参数，并利用已有测试数据来测试模型的准确性。可以通过对比使用相同数据特征（测试数据集）情况下模型预测值与测试数据真实值之间的差距来评估模型性能，差距越小，模型的预测能力越好。

预测的内容是商店的销售额，用具体的数字表示，因此该问题是容易被测量和观察的（例如绘制点图、曲线等方法）。

# 数据集及输入

## 概述

本项目所使用的数据全部来自与Kaggle网站【2】。数据集共包括4个文件：

1. train.csv 用于训练的历史数据，包括销售额
2. test.csv 用于测试的历史数据，不包括销售额
3. sample\_submission.csv 具有正确格式的提交文件
4. store.csv 关于商店的其他补充信息

## 训练数据train.csv

利用jupyter notebook和pandas对训练数据进行初步探索，共有1,017,209条数据，每条数据包括9个数据域(Data Field)，如表X所示。

|  |  |  |
| --- | --- | --- |
| 数据域 | 意义 | 数据类型 |
| Store | 每个商店独有的ID | Int64，Category |
| DayOfWeek | 每周的第几天 | Int64，Category |
| Date | 记录的日期 | object，TimeStamp |
| Sales | 对应日期的商店销售额 | Int64，Real Number |
| Customers | 当天到店的消费者数 | Int64，Real Number |
| Open | 当天商店是否开张 | Int64, Category |
| Promo | 当天商店是否有优惠 | Int64, Category |
| StateHoliday | 当天是否是国家节日 | object，Category |
| SchoolHoliday | 当天是否是学校假期 | Int64, Category |

进一步探索发现，训练数据记录了1115个商店的数据，记录日期从2013年1月1日到2015年7月31日，共942天。但并不是所有的商店都有完整记录，其中934家店有完整的942天记录，1家店有941条记录，其余180家店有758条记录。

<https://en.wikipedia.org/wiki/Rossmann_(company)>

https://www.kaggle.com/c/rossmann-store-sales