# 机器学习纳米学位

毕业项目 张立芹 优达学城

2018 年 06 月 04 日

## 项目背景：项目涉及的相关研究领域

- Rossmann 是欧洲的一家连锁药店，根据 Rossmann 药妆店的信息，比如促销、竞争对手、节假日、季节、气候、地域等因素，来预测 Rossmann 未来的销售额。本项目要求预测 Rossmann 的每家商户未来 6 周的每日销售额，我将提供创建一个预测模型。

## 问题描述：解决办法所针对的具体问题

通过过去的销售情况，通过分析后的数据进行创建模型，用于预测六周后的日销售情况。具体步骤和可能遇到的问题如下：

- 对数据进行清理与整理，进行缺失值与异常值处理。
- 对数据单变量分析、变量相关系分析。
- 使用 XGBoost 来建模
- 获取最佳模型。
- 预测并评价结。

## 输入数据：问题中涉及的数据或输入是什么

- *输入数据集*
  train.csv 表示过去的销售数据为训练集。
- test.csv 表示测试数据。
- sample_submission.csv 预测数据格式样本
- store.csv 表示门店信息

## 解决办法：针对给定问题的解决方案

- 查看缺失值，如果缺失值占比超过 90%，将移除该特征，其他的使用均值或中位数进行补充。根据单个特征分布情况，对异常值进行移除。
- 对特征值的相关度分析获取对模型有效变量。
- 使用 Train 和 Test，进行 XGBoost 模型训练与验证
- 通过不断的尝试和搭配 XGBoost 算法的高级用法来寻找到最适合的参数。
- 使用可视化结果展现

# 基准模型：用来与你的解决方案进行比较的一些简单的、过去的模型或者结果

- 对销售额预测的错误率在 10%， 使用 GB 模型来对比 XGBoost 模型的表现。

# 评估指标：衡量你解决方案的标准

- 使用 RMSPE 来做为验证函数，该值越低代表差异性越小。它是指模型的预测值和实际观察值之间的差异的一种衡量方式.

# 设计大纲：你的解决方案如何实现，如何获取结果

- 将 store 与 train 和 test 进行合并。
- 对数据做清洗和整理，对缺失占比为 90%的特征剔除，其他的使用均值或者中位数进行补充。
- 对异常值进行移除处理。
- 对数据变量做可视化分析。
- 进行模型训练。
- 配置 xgboost 参数并开始第一次的训练。
- 寻找最佳模型
- 记录训练的评估结果并保存模型。
- 用模型对 test 数据做 predict 并将结果按照 submission 的方式保存。