# 机器学习纳米学位

毕业项目 薛威 优达学城

2018 年 01 月 08 日

## 项目背景：项目涉及的相关研究领域

- Rossmann 在欧洲经营 7 国经营着 3000 家药店，目前，Rossmann 商店的经理被要求预测他们未来六周的日销售情况，商店销售收很多因素的影响，包括促销，竞争，学校放假和法定节假日，节气性和区域性。由于数千经营者依据他们独特的情况预测销售情况，结果的准确性可能有很大不同。 在这个项目中，将挑战预测 6 周的 1115 家德国境内的 Rossmann 商店的每日销售额，可靠的销售预报可以让商店经营者增加工作效率和积极性创建更高效的工作人员安排。通过帮助 Rossmann 创建一个强壮的预测模型，你将帮助经营者停留在关注对他们来说什么是最重要的：他们的客户和他们的团队。
- 选择该项目的主要原因有两个，第一针对已有数据分析的预测是我今后发展的一个主要的方向和兴趣点，第二日后在 kaggle 进行数据比赛先以此作为一个试水项目

## 问题描述：解决办法所针对的具体问题

我要通过对旧的销售记录的学习建模，然后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：

- 对数据做清洗和整理，可能遇到的问题有异常值的处理，缺失值的处理。
- 对数据做单个变量分析，多个变量分析，取出影响结果的高相关度的变量，可能会遇到如何取出相关度高的特征值的问题。
- 使用 XGBoost 来建模，如何调整训练模型的数据用来适合 XGBoost 模型的输入要求。
- 如何找到最合适的超参。
- 预测并评价结果，根据 Kaggle 的描述使用 RMSPE 来做为评价标准，需要对结果做一个可视化，需要补充相关的知识。

## 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 - 包括销售额的历史数据训练用

  包含有

  "Store","DayOfWeek","Date","Sales","Customers","Open","Promo","StateHoliday","SchoolHoliday"字段。

  test.csv - 包括销售额的历史数据测试用 historical data excluding Sales

  包含有

  "Id","Store","DayOfWeek","Date","Open","Promo","StateHoliday","SchoolHoliday"字段。

  sample_submission.csv - 预测数据格式样本

  包含有"Id","Sales"字段。

  store.csv - 关于商店的附加信息

  包含有

  "Store","StoreType","Assortment","CompetitionDistance","CompetitionOpenSinceMonth",

  "CompetitionOpenSinceYear","Promo2","Promo2SinceWeek","Promo2SinceYear","PromoInterval"字段。

- *数据集特征如下*

  Id - 测试集中表示一条记录的编号。

  Store - 每个商店的唯一编号。

  Sales - 任意一个给定日期的销售营业额。

  Customers - 给定那一天的消费者数。

  Open - 商店是否开门标志，0 为关，1 为开。

  StateHoliday - 表明影响商店关门的节假日，正常来说所有商店，除了极少数，都会在节假日关门，a=所有的节假日，b=复活节，c=圣诞节，所有学校都会在公共假日和周末关门。

  SchoolHoliday - 表明商店的时间是否受到公共学校放假影响。

  StoreType - 四种不同的商店类型 a，b，c 和 d。

  Assortment - 描述种类的程度，a = basic, b = extra, c = extended。

  CompetitionDistance - 最近的竞争对手的商店的距离。

  CompetitionOpenSince[Month/Year] - 最近的竞争者商店大概开业的年和月时间。

  Promo - 表明商店该天是否在进行促销。

  Promo2 - 指的是持续和连续的促销活动。: 0 = 商店没有参加, 1 = 商店正在参加。

  Promo2Since[Year/Week] - 表示参加连续促销开始的年份和周。

  PromoInterval - 描述持续促销间隔开始，促销的月份代表新的一轮，月份意味着每一轮的开始在哪几个月。

# 解决办法：针对给定问题的解决方案

- 查看缺失值的特征类型和特点，补充 media 值或者设置上下限随机生成。根据各个特征的分布情况找到并 dropout 异常值。
- 对特征值的相关度做一个排序选取有正向作用的特征。
- 将数据切割为 Train 和 Test，并进行 matrix 来符合 XGBoost 的输入要求
- 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。
- 学习相关的可视化结果的知识来完成最后的结果的提交

# 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

- 对于销售额的预测人为假设错误率在 10%， 我用一些其他的模型比如说 GB 模型来对比 xgboost 的表现。

# 评估指标：衡量你解决方案的标准

- 我使用 Kaggle 提供建议的 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

- 先对数据做清洗和整理，处理缺失值，统一数据类型，拆分部分的特征，合并部分的特征，剔除部分特征。
- 将 store 与 train 和 test 数据合并起来，剔除和拆分部分特征。
- 对数据变量做可视化分析，去掉异常值。
- 将 Train 和 Store 合并处理后的数据做 train 和 test 拆分并 matrix 化准备开始训练。
- 配置 xgboost 参数并开始第一次的训练。
- 对第一次训练的结果做总结，再对特征做进一步的优化，对参数做进一步的优化
- 开始第二次训练，并通过配置 xgboost 参数的方式来寻找最佳超参
- 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。