Rossmann 销售额预测开题报告

李榕 优达学城

2019 年 9 月 9 日

提案

本文对 Rossmann 零售店，1115 个分店，自 13 年 2 月到 15 年 7 月之间的销售额进行分析，综合店铺各类因素，提出用于预测 15 年 8 月到 15 年 9 月各零售点销售额的解决方案。

背景描述

Rossmann（劳诗曼）成立于 1972 年，是德国最大的日化用品超市，截至 2012 年，Rossmann 有近 2600 家连锁店，其中 1600 多家位于德国境内。商家通过一系列折扣活动来刺激销售额，包括定期和不定期的折扣活动，而各个分店，由于各自的所属品类，地理位置，周边竞争者的地理位置，节假日等情况对于销售额产生了对应的影响。准确的预测各个门店的销售额，并分析影响销售额的因素，对于门店物品数量准备，改变打折策略，最终提升销售额起着重要的作用。

问题陈述

本课课程需要解决依据 Rossmann 各个商店 13 年 2 月到 15 年 7 月之间的销售额数据，结合商店类型、商店的竞争者位置、是否是节假日、当天是星期几等综合因素，来对 Rossmann15 年 8 月到 15 年 9 月之间的销售额进行预测。据此，问题可以被拆解为如下几点：

1、店铺的销售额与店铺的品类是不是有关系，有怎样的影响？

2、店铺周围是否有竞争者，对销售额是否有影响，有怎样的影响？

3、店铺竞争者的位置远近，对销售额是否有影响，有怎样的影响？

4、当天的日期是正常工作日还是星期日，对销售额是否有影响，有怎样影响

5、当天是否是公共假期，对销售额是否有影响，有怎样影响

6、当天日期是否是学校假日，对销售额是否有影响，有怎样影响

7、当天是是否发生促销，对销售额是否有影响

8、当天是否发生了季节性促销，对销售额的影响怎样

9、明确了对销售额的影响因素之后，进行特征工程的构建，对空值进行删除补充

10、考虑采用回归模型来进行不同店铺销量的预测。

11、评估预测的效果如何

数据集和输入

分析数据集中的各类变量，情况如下：

1、每个商店的一些信息：

store 商店 ID

storetype 商店类型

Assortment 商店分类

CompetitionDistance 竞争者的距离

CompetitionOpenSinceMonth 竞争者开启的月份

CompetitionOpenSinceYear 竞争者开启的年份

Promo2 是否有广告投放

Promo2SinceWeek 广告从那一周开始投放

Promo2SinceYear 广告从哪一年开始投放

PromoInterval 广告投放的周期情况

2、训练集及测试集数据

store 商店 ID

dayofweek 周几

date 日期

sales 销量

customer 顾客量

open 店铺当天是否开张

promo 是否有促销

StateHoliday 国家假日情况

SchoolHoliday 学校假日情况

依照题目内容，本提案中，对如下内容 ：

商店类型

商店分类

竞争者距离

竞争者月份&年份

广告投放时间

广告投放周期

星期

店铺开张情况

假日情况

如上变量作为输入，将 sales 销量作为输出进行模型的构建。

将训练样本：train.csv 文件 分为训练集与测试集进行模型的构建和效果的评估

进而对 test.csv 文件内数据进行预测，得到本次项目的结果

解决方案

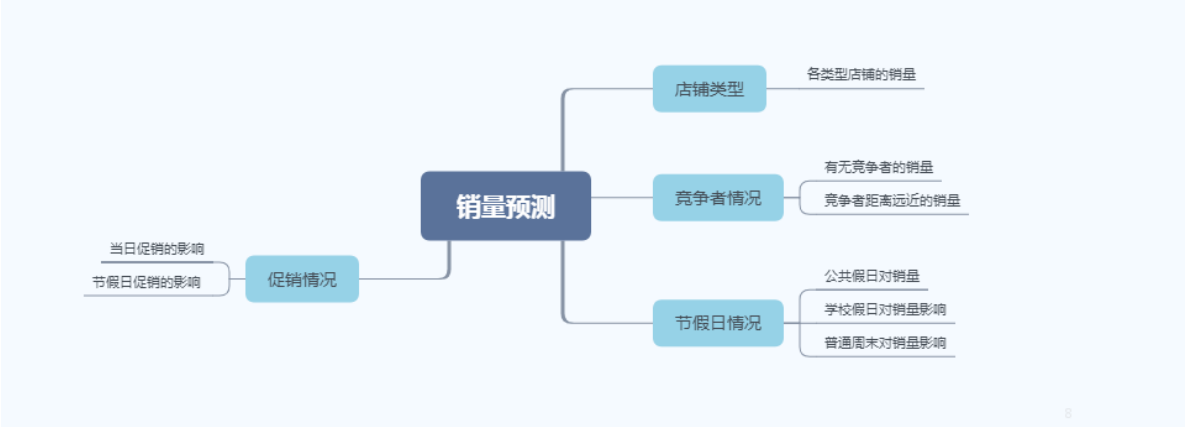
本问题属于回归问题，回归问题的目标是为了使得预测值与真实值更加接近。

为了解决回归问题计划采用的方式：

1、数据的分析与探索：

确定目标变量：Sales(销量)

分析各类特征与目标变量之前的关系，对数据能够有明确的认识，分析思路如下：



2、特征工程的构建：

通过相关的分析过程之后，需要对模型的输入进行特征的构建：

1）去除掉无用特征：商店 ID，顾客量（结果）等

2）非数值的特征数值化

3）通过当前特征衍生出更多有价值特征

最终作为输入放到模型中来

3、评估指标

本文处理的问题为回归问题，回归问题通过分析预测值与真实值之间的差距，来评估模型的好坏，常用的评价标准：

1）mse：均方误差——越接近于 0 越好

$$MSE = \frac{SSE}{n} = \frac{1}{n}\sum_{i=1}^{m} w_i(y_i - \hat{y}_i)^2$$

2）rmse：均方根误差——越接近于 0 越好

$$RMSE = \sqrt{MSE} = \sqrt{\frac{SSE}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{m} w_i (y_i - \hat{y}_i)^2}$$

3）mae：平均绝对误差——越接近于 0 越好

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i| .$$

4）R2：决定系数——越接近 1 越好

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{\text{samples}}-1} (y_i - \bar{y})^2}$$

5）RMSPE——均方根百分比误差，为项目要求使用的评估方法。

$$\mathrm{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{y_i} \right)^2},$$

4、模型选择

本问题是一个回归问题，因而考虑采用回归模型来求解：

1）**线性回归，**通过线性回归构造出来的函数一般称之为了线性回归模型。

$$h_\theta(x) = \theta_0 + \theta_1 x$$

2）**回归树**，通过构建决策树来获取到区间均价，分裂点方法与分裂问题不同，分类问题采用的是，信息增益或信息增益率 基尼系数等方法；回归树采用最小化误差平方的方式，

从特征变量中，找到一个变量 j，找到一个取值，这个取值能够使得按照取值划分之后的两个集合，对于目标变量的误差平方最小；接着依次遍历所有的变量，得到特征空间的划分。然后依次进行划分，最终得到子向量空间，向量空间的均值就是预测值。

3）**随机森林**

1. 假如有 N 个样本，则有放回的随机选择 N 个样本(每次随机选择一个样本，然后返回继续选择)。这选择好了的 N 个样本用来训练一个决策树，作为决策树根节点处的样本。

2. 当每个样本有 M 个属性时，在决策树的每个节点需要分裂时，随机从这 M 个属性中选取出 m 个属性，满足条件 m << M。然后从这 m 个属性中采用某种策略（比如说信息增益）来选择 1 个属性作为该节点的分裂属性。

3. 决策树形成过程中每个节点都要按照步骤 2 来分裂（很容易理解，如果下一次该节点选出来的那一个属性是刚刚其父节点分裂时用过的属性，则该节点已经达到了叶子节点，无须继续分裂了）。一直到不能够再分裂为止。注意整个决策树形成过程中没有进行剪枝。

4. 按照步骤 1~3 建立大量的决策树，这样就构成了随机森林了。

4）**xgboost**

xgboost 由多个相关联的树联合。

xgboost 不断对残差进行预测，xgboost 每个决策树是逐一被添加进入。下一棵决策树加入的输入，会取决于前一棵决策树产生的残差。

单个树生成方式：

4.1 不断便利特征和特征的取值，然后进行分类，计算 loss function 最小值，然后再选择一个特征分裂，又得到一个损失函数的最小值，找到分裂效果最好的（即分裂前后损失函数变化最大）的特征进行分裂

4.2 持续 4.1 的过程，不断进行分裂

4.3 如下条件停止分裂：

a）分裂的增益小于某个阈值

b）分裂达到最大深度时，停止，最大深度理论上应当是是一个超参数

c) 样本权重之和，小于设定阈值时，停止分裂

4.4 计算前一棵树的残差数据，然后将 目标函数定义为 残差，继续输入到模型中，进行重复的模型构建，直到触发 ：增益小于阈值，树的数量小于一定量，样本的权重之和小于设定的阈值，停止整个过程。

5、基准模型

基准模型考虑使用线性回归模型来进行预测

项目设计

本提案用于解决销量预测问题，设计总体思路如下：

1、明确要解决的问题——通过已经确定的商店类别、竞争者、节假日等因素，对未来销量进行预测

2、数据分析——对相关的数据变量与销量之前的关系进行分析，看销量的多少与某个数据变量之前的关系如何

3、对数据进行特征工程的构建，去除分析中无用的变量，增加有效的特征变量，用于对模型进行输入

4、构建模型的评估标准，采用通用的对回归模型的评估标准，应项目的需求，采用rmspe

5、选择模型，进行预测

6、为预测的数据中得到结果