预测 Rossmann 未来的销售额

一 领域背景

数据的兴起，互联网而技术与各个领域的快速融合，使得各个领域发展产生了巨大的变化，尤其是利用机器学习预测销售额的技术日趋完善，一方面互联网的快速发展使得各个领域的销售数据更加的系统化、完整化，另一方面大数据 AI 的兴起，使得这些数据变的更有经济价值，通过对历史销售数据的学习与反馈，从而预测将来的销售情况，方便市场更好的调动资源，使得人力及物资资源得到最优化的配置和最大化的利用。

二 问题陈述

本项目采用 Rossmann 的每天的销售情况的数据，通过对 Rossmann 历史销售数据的学习，预测未来的销售情况。从项目的定义来看，本项目为一个回归问题，属于监督学习，所以要采用监督学习的模型来对数据进行学习及预测。

Rossmann 是欧洲的一家连锁药店，在 7 个欧洲国家拥有 3,000 家药店。目前，Rossmann 店经理需要提前六周预测其日销量。 商店销售受到诸多因素的影响，包括促销，竞争，学校和国家假日，季节性和地点。 成千上万的个人经理根据其独特的情况预测销售量，结果的准确性可能会有很大的变化，可靠的销售额预测可以使得 Rossmann 店经理 制作员工时间表以提高生产力和动力。本项目的目标是帮助 Rossmann 店经理建立一个稳定的预测模型，预测未来的销售额，从而帮助 Rossmann 店经理分析出影响销售的关键要素。

三 数据集和数据输入

本项目的数据集是从 Kaggle 网站上下载下来的，从数据集来看，分成三个部分：store.csv（店铺信息表），train.csv (训练集) 和 test.csv（测试集）

训练集数据 train.csv 中包，

其中 store.csv（店铺信息表），该表中有 1115 条记录，记录的是 1115 个店的信息，其属性详细说明如下：

属性名	说明	备注
store	店 ID	
StoreType	店的类型 a,b,c,d	
Assortment	店的分类	a basic，b:extra,、c:extended
CompetitionDistance	与竞争者的距离	
CompetitionOpenSinceMonth	附近竞争者开始营业的月份	2：表示从 2 月份开始营业
CompetitionOpenSinceYear	附近竞争者开始营业的年份	2008：表示附近竞争者从 2008 年开始营业
Promo2	是否有连续的促销活动	0 没有，1 有
Promo2SinceWeek	店面开始促销活动的周	
Promo2SinceYear	店面开始促销的年	

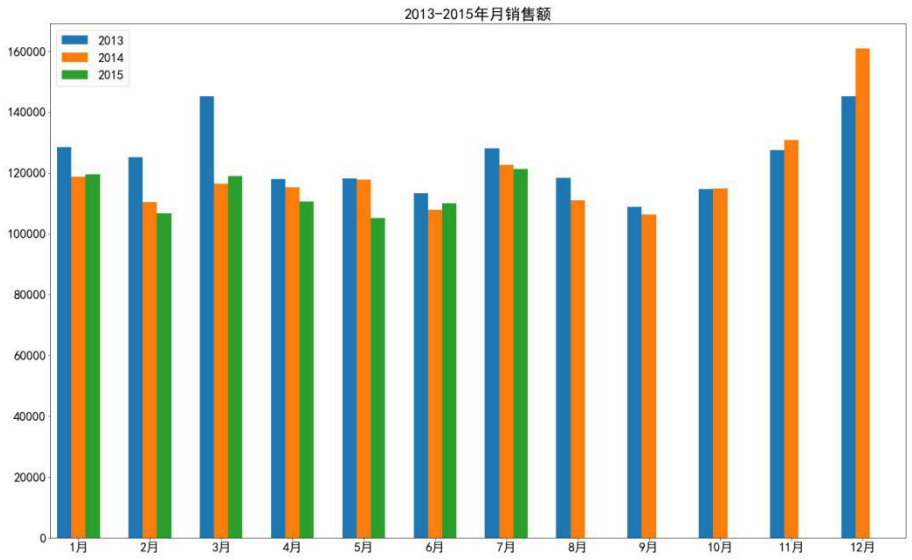
train.csv（训练集）表里有 101，7210 条数据，包含了从 2013 年开始到 2015 年 7 月的每天的销售额。

其属性详细说明如下

属性名	说明	备注
Store	店 ID	
DayOfWeek	星期几	1-7：星期一~星期日
Date	日期	
Sales	销售额	
Customers	顾客数量	
Open	是否营业	0 否 1 是
Promo	是否有促销活动	0 否 1 是
StateHoliday	国定假日	a = public holiday, b = Easter holiday, c = Christmas, 0 = None
SchoolHoliday	是否为学校假期	0 否 1 是

从上面两个表的数据属性来初步看，影响销售额的因素有很多，所处的日期月份，是否是节假日，学校是否放假，是否有促销活动，所在的日期月份，周围的竞争者等等 都构成影响销售额的因素，但具体的因素的重要性及影响性需要通过数据分析得出最后结论。

从训练集数据来看，筛选 open=1 并且 Sales>0 的数据，作为训练数据，总共 844392 条数据，为了看下 整体销售情况，对每个月的销售额进行了汇总比较，下面的柱形图呈现其对比情况：



从上图中可以看出，每年 5,6,8,9 这几个月的销售额偏低一些，具体因素 还需进一步分析。

测试集 test.csv 总共 41,089 条数据，主要预测 2015 年 8,9 月的数据，其中 open 属性有为空的状态，填充默认值为 1，即为营业状态，测试时候选取 open=1 的数据进行预测，open=0 的时候 Sales = 0

四 解决方案

根据上面对数据集的各个属性的分析， 该问题定位成回归的问题，可以按照监督式学习的模型选取几种模型进行学习并进行对比，选取最有模型，再进行预测。

五 基准模型

根据数据情况，想尝试几种模型，随机森林 bagging 和 boosting （xgboost） ，几种模型，从效率和效果上来做个对比，最后选择一种模型进行调优详细预测。

六 评估指标

采用模型的 rmspe 数值来作为评估的标准

七 项目设计

1 数据加载与清洗：对数据进行加载和清洗，查看数据的缺失情况，并处理缺失数据；

2 数据探索：查看数据的几个基本数据特征，了解数据的基本分布情况；

3 分析属性：根据表中的属性思考是否可以 进行进一步的转换变成更有明显意义的属性，比如 可以增加 根据 Promo2SinceWeek/Year 属性进行转变，进而构成是否在促销期间，表示该销售记录所在的日期是否正在进行促销活动；

4 数据训练：准备采用 xgboost 和随机森林两种模型进行训练并对比训练效果，数据集的划分采用随机划分的方式；

5 针对不同的训练模型，对比其 RMSPE 的值，并查看每种训练模型下各个特征重要性的分布情况。