Forecast Rossmann Store Sales

项目背景：

Rossmann 在 7 个欧洲国家经营着 3000 多家药店。目前，Rossmann 门店经理的任务是提前 6 周预测他们的日常销售。商店的销售受到许多因素的影响，包括促销、竞争、学校和国家假日、季节性和地区。由于成千上万的个人经理根据自己的特殊情况预测销售情况，结果的准确性可能会非常不同。

在他们的第一个 Kaggle 竞赛中，Rossmann 挑战通过在德国各地的 1,115 家商店的每日销售额来预测未来 6 周的销售额。可靠的销售预测使商店经理能够制定有效的员工时间表，提高生产力和积极性。通过帮助 Rossmann 创建一个健壮的预测模型，您将帮助商店经理专注于他们最重要的事情:他们的客户和他们的团队

问题描述：

Rossmann 是欧洲的一家连锁药店。 在这个源自 Kaggle 比赛 Rossmann Store Sales 中，我们需要根据 Rossmann 药妆店的信息（比如促销，竞争对手，节假日）以及在过去的销售情况，来预测 Rossmann 未来的销售额。

输入数据：

输入的数据有两部分：

1. Rossmann 商店信息数据：
    1. Store（商店 ID）
    2. DayOfWeek（日期所在的周几）
    3. Date（销售时间）
    4. Sales（销售额）
    5. Customers（消费者数量）
    6. Open（商店当日是否开放）
    7. Promo（商店当日是否促销）
    8. StateHoliday（国家节假日，a、b、c 是国家节假日，0 表示非节假日）
    9. SchoolHoliday（国家公立学校节假日，1：节日，0 非节日）
2. 商店销售数据。
    1. Store（商店 ID）
    2. StoreType（商店类型）
    3. Assortment（商店级别）
    4. CompetitionDistance（竞争对手距离）
    5. CompetitionOpenSinceMonth（竞争对手开店月份，部分缺失）
    6. CompetitionOpenSinceYear（竞争对手开店年份，部分缺失）
    7. Promo2（是否持续促销）
    8. Promo2SinceWeek（持续促销的周）
    9. Promo2SinceYear（持续促销的年）
    10. PromoInterval（促销的月份）
3. 缺失数据

    如图 1 所示 CompetitionDistance 有三家数据缺失，Promo2 促销日有 571 天，竞争

对手开店时间大量缺失,缺失 206 条数据，如图 2 所示。

```
RangeIndex: 1115 entries, 0 to 1114
Data columns (total 10 columns):
Store                        1115 non-null int64
StoreType                    1115 non-null object
Assortment                   1115 non-null object
CompetitionDistance          1112 non-null float64
CompetitionOpenSinceMonth    761 non-null float64
CompetitionOpenSinceYear     761 non-null float64
Promo2                       1115 non-null int64
Promo2SinceWeek              571 non-null float64
Promo2SinceYear              571 non-null float64
PromoInterval                571 non-null object
dtypes: float64(5), int64(2), object(3)
memory usage: 87.2+ KB
```

图 1

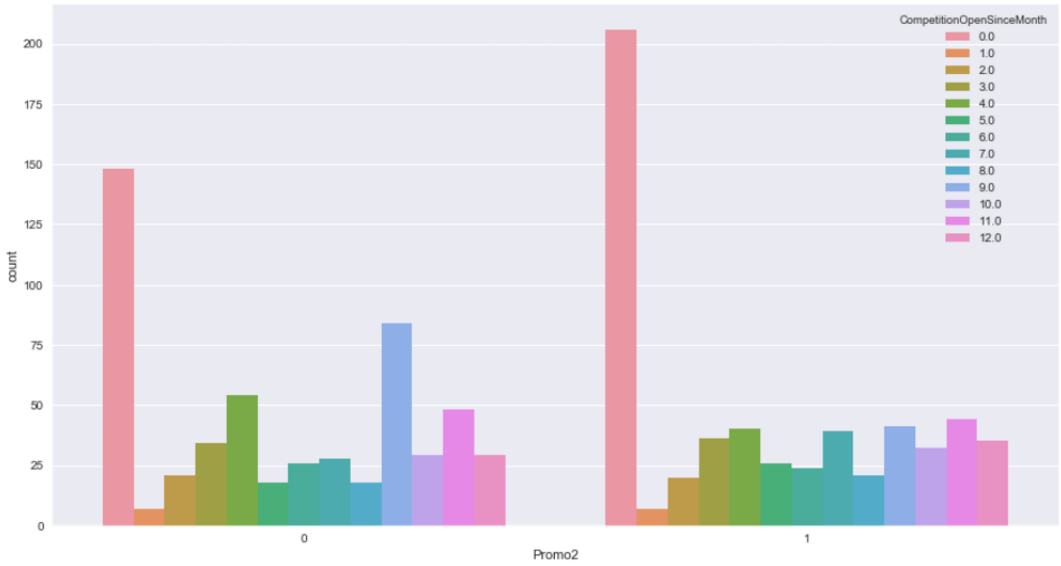

图 2

解决办法：

根据数据集通过数据集中数据的时间序列特性，编写每条数据的生成时间框函数，从而能够获取每个门店前 3、5、7 日各自的销售数据来对数据进行特征工程，最后建立模型的 X 数据集合 Y 数据集。根据问题要求选择选择回归模型，为了提高准确率选择集成模型。

基准模型：

根据项目背景和问题，为未来 6 周各个门店的销售额进行预测，本项目开始参考 Omar El Gabry 的特征分析方法：

1. 分析不同类型商店的平均销售。
2. 分析一周七天的平均销售额。
3. 分析连续几个月的销售额。
4. 进行促销对比分析。
5. 进行持续促销对比分析。

6. 是否公共节假日对比分析。
7. 通过箱型图查看销售额，用户数分布。
8. 对商店类型进行对比分析。
9. 对商店等级进行对比分析。
10. 查看竞争对手距离分布。
11. 查看用户数分布。
12. 进行商店之间相关性分析。

通过分析发现一周的不同时间、节假日、促销、商店类型等对销售额有影响，通过 one-hot-encoding 编码实现特征工程，在根据 XGboost 算法分类器把成百上千个分类准确率较低的树模型组合起来，成为一个准确率很高的模型；这个模型会不断地迭代，每次迭代就生成一颗新的回归树，在对每个商店单独建模预测数据，最后通过搜索优化提高模型的准确率。

评估指标：

提交的内容是根据均方根百分比误差(RMSPE)计算的。RMSPE 被计算为

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

其中 yi 表示单个商店在一天内的销售额，yhat_i 表示相应的预测。任何一天和商店 0 的销售都被忽视得分。

设计大纲：

1. 通过对 Rossmann 的销售数据通过探索分析，发现销售额与其它特征之间的关系。

1）一周分析，发现数据与周几有很大影响。

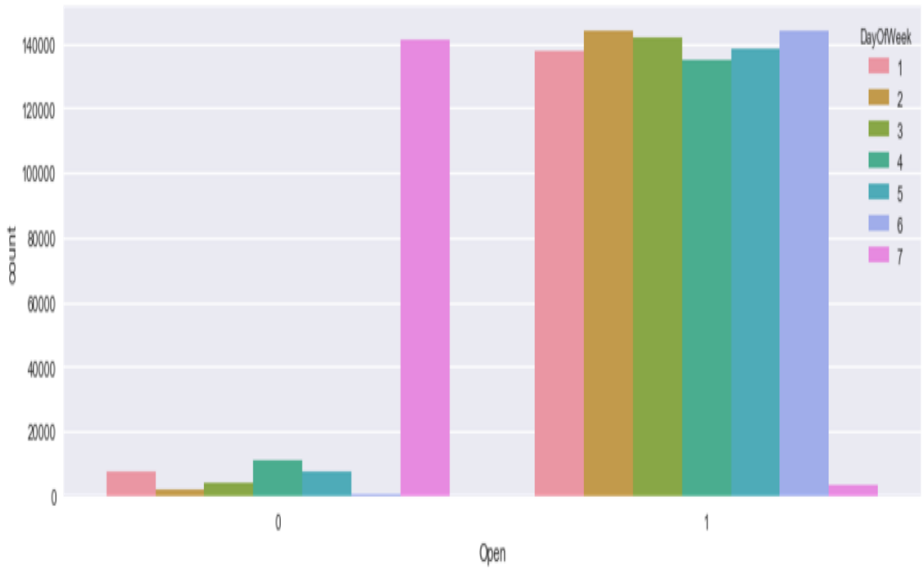

图 3

2）相关分析，发现销售额与客户量、商店是否营业、促销、一周的第几天高相关如图 4 所示。

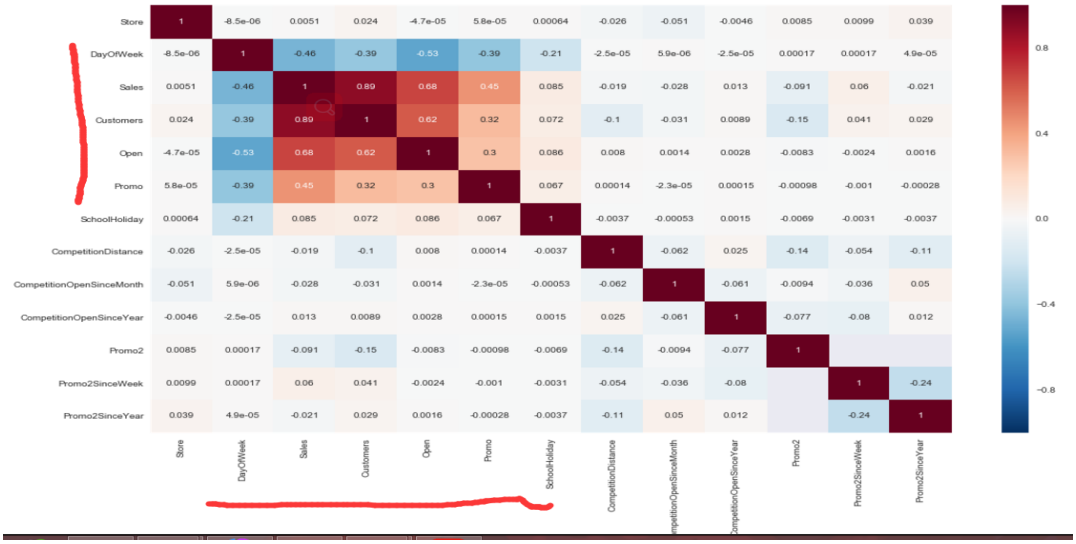

图 4

2. 对数据集中的缺失值，错误值进行删除替换，对其中销售额额为 0 的数据进行删除。
3. 根据销售额的特征值与其它特征值进行相关分析，特征比较进行特征筛选。
4. 建立模型数据的 x，y 值，通过 XGboost、random forest、回归树这三个模型进行销售额预测建模。
5. 通过对比分析选择准确率最高的模型。
6. 通过 sciki-learn 包中参数优化函数搜索最优参数。从而使模型准确率提高。
7. 最后对测试集数据进行预测。
8. 该项目业务流程图如图 5 所示

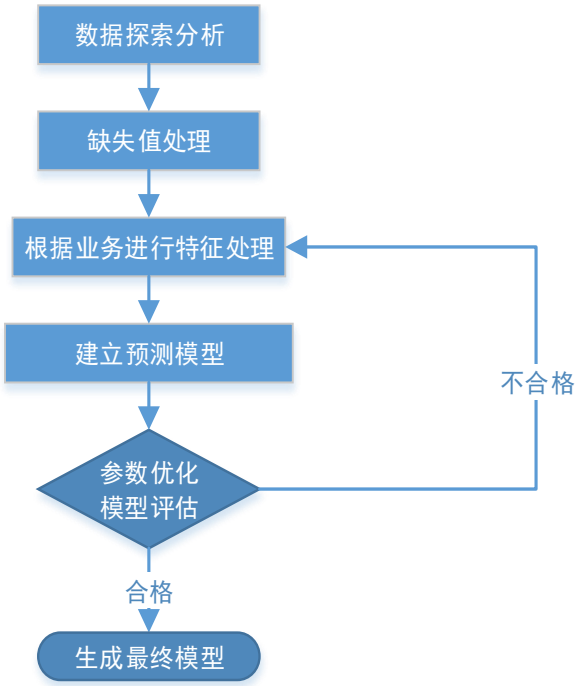

图 5