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：Rossmann 商店信息数据和商店销售数据。

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几日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：

通过使用 XGboost、random forest、bagging 这三个模型进行比较分析选取最优模型。

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

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

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。
2. 对数据集中的缺失值，错误值进行删除替换。
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、bagging 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。