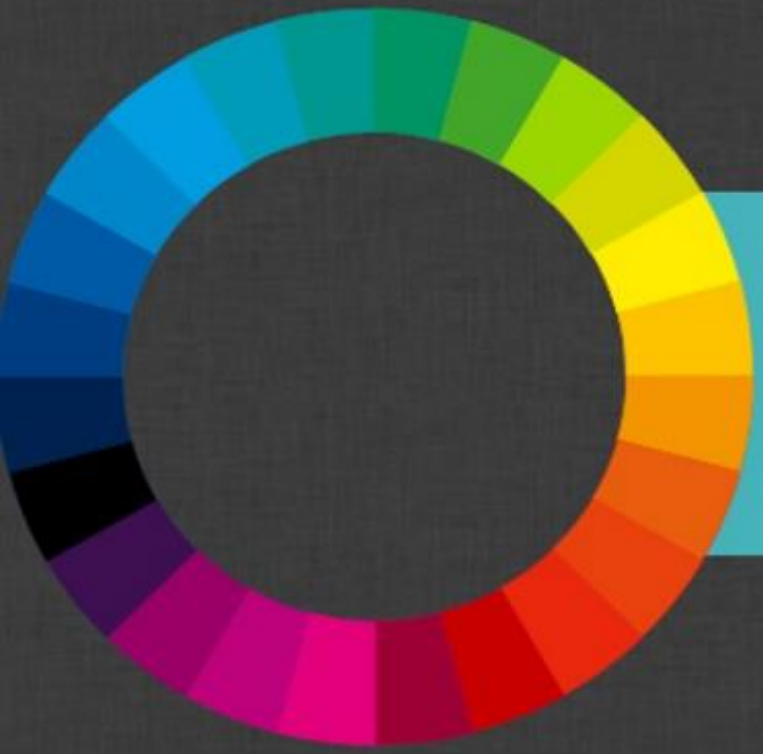


竞赛答辩

房租预测

CONTENTS

成员介绍

- 14号
- 15号

作品概述

关键技术阐述

- 数据清洗
- 异常值处理
- 缺失值填充

数据处理

特征工程

- 常规统计特征
- 基础特征
- 业务相关特征

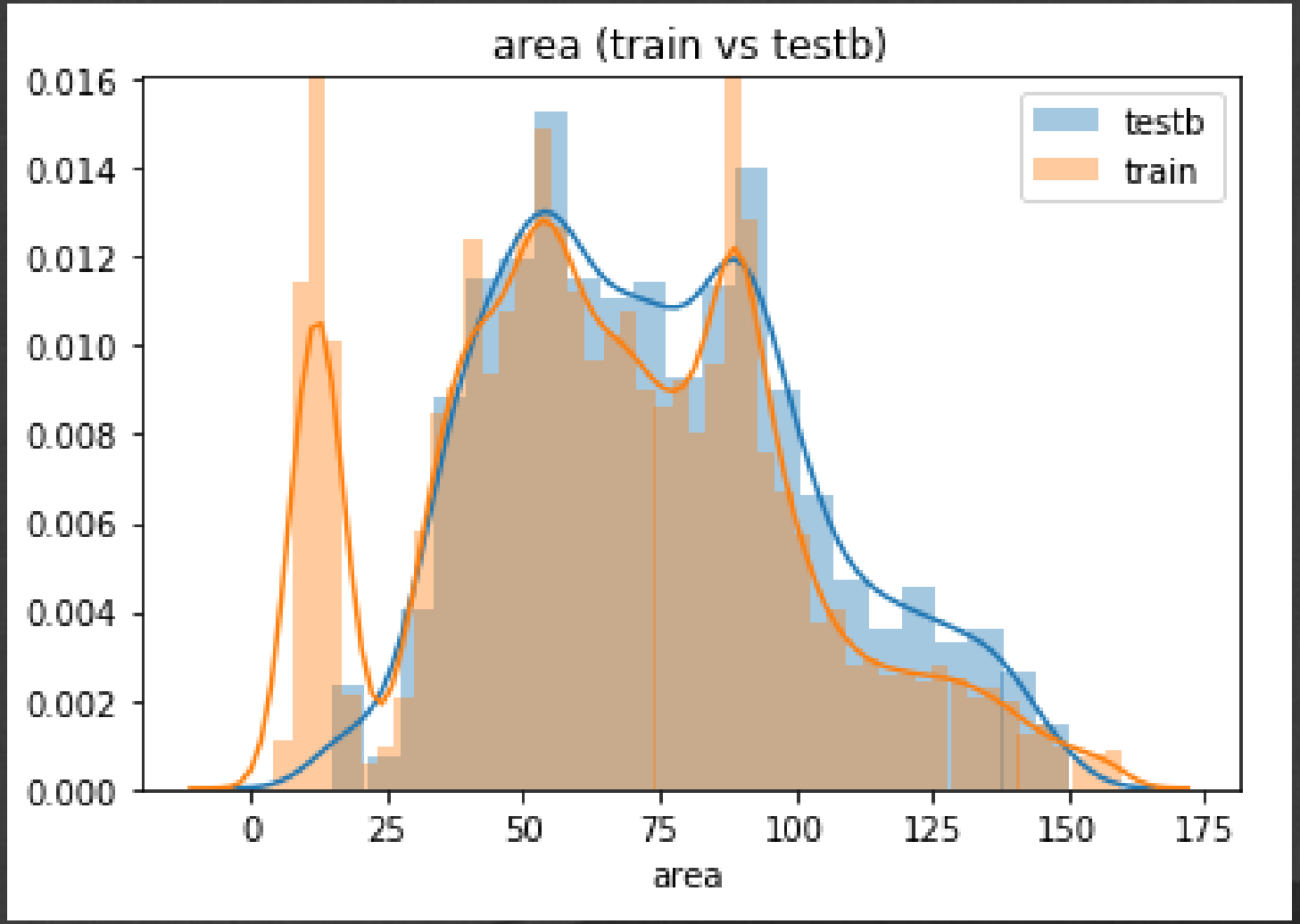
特征工程

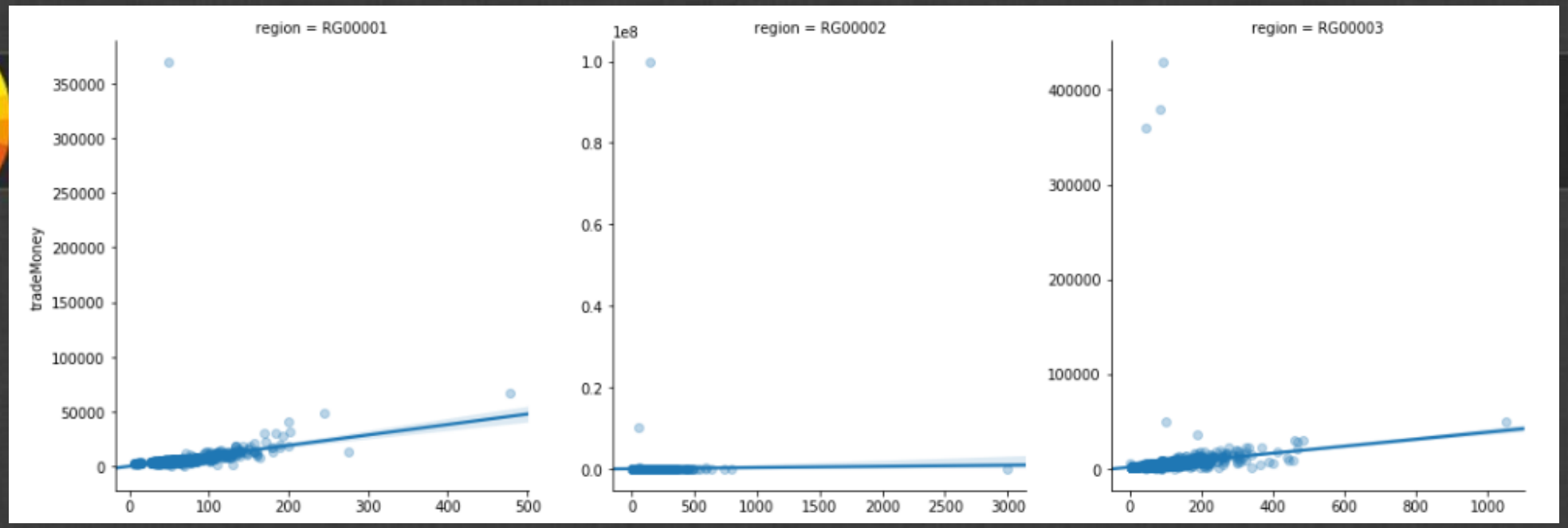
模型筛选

模型融合

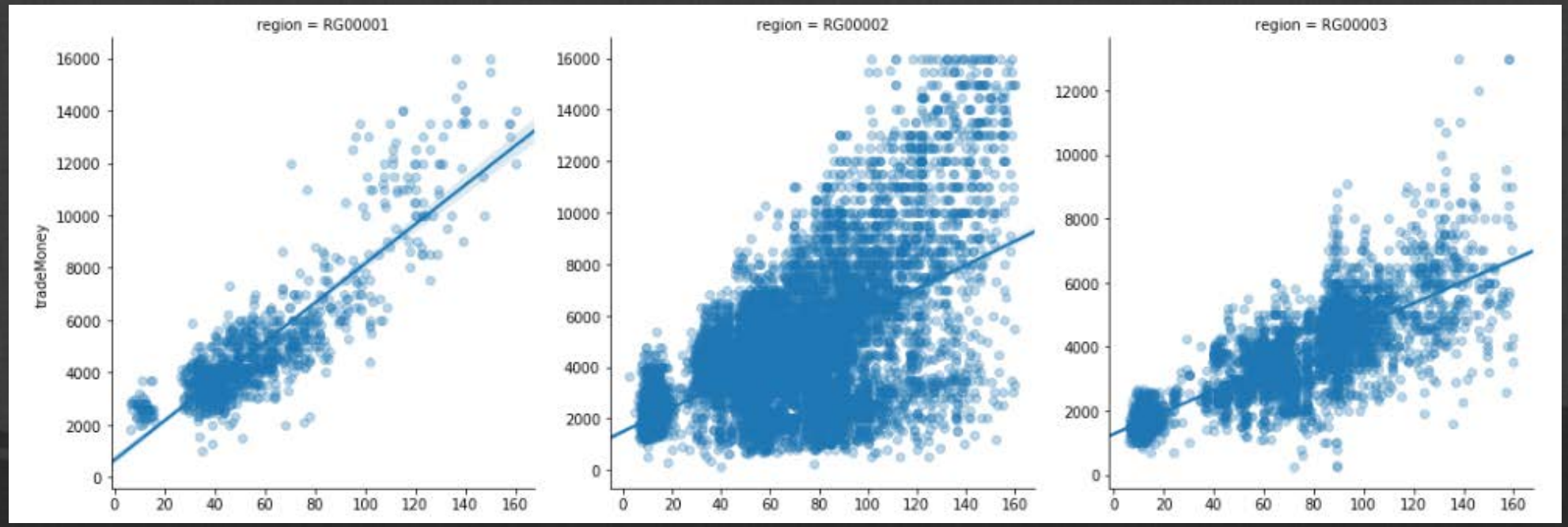
房屋面积

数据清洗前
训练集房屋面积
1-15056m²
测试集房屋面积
15-150m²





数据清洗，针对每个region的数据，按照area和tradeMoney两个维度进行异常值处理



数据清洗

数据清洗：

缺失值填充

- RentType对该问题来说，是一个非常重要的特征，然而，该字段有74%+的数据是缺失，因此，对其进行有效的填充是十分必要的。本方案对其制定了一套规则进行缺失值的填充。
- 该数据其他字段（如：pv，uv，buildYear等）也有缺失，这里我们采用众数填充，均值填充，最值填充等方式

数据处理

类别特征处理

较小特征：one-hot

较大：Word2Vec

数值特征处理

将种数较少的转为类别特征，

利用这些特征进行聚类

长尾数值归一化

特征工程

房屋特征

- 房屋面积
- 房型
- 出租方式
- 楼层
- 总楼层数
- 朝向
- 装修

位置信息

- 小区名（W2V）
- 板块名
- 区域名
- 建筑年代

配套设施

学校、交通
医疗、购物
生活设施

市场信息

- 当月二手房、新房交易面积、均价、套数
- 当月土地供应幅数、面积、成交幅数、成交总价
- 现有办公人数
- 常驻人口
- 当月流入人口
- 线上浏览次数、总人数
- 线下看房次数

特征工程

交叉特征

房屋面积与金钱类特征的交叉，
区域、板块与金钱类特征的交叉，
配套设施的组合特征等。

组合特征

对特征取均值、最大值、最小值、标准差、计数、one-hot等统计量；
按照一个或多个特征进行分组，并求取各组的统计特征

业务特征

带看次数与浏览次数比值
成交面积与供应面积比值等

模型融合

模型筛选

- LightGBM
- XGBoost
- GBDT

模型融合

- Stacking
- 简单加权融合

单模结果

- 最终单个模型

```
K-Fold score:0.917358
```

```
from sklearn.metrics import r2_score
def online_score(pred):
    print("预测结果最大值：{},预测结果最小值：{}".format(pred.max(),pred.min()))
    # a榜测分
    conmbine1 = pd.read_csv("./newe/sub_b_919.csv",engine = "python",header=None)
    score1 = r2_score(pred, conmbine1)
    print("对比919分数:{}".format(score1))
```

```
online_score(predictions)
```

```
预测结果最大值：15437.373721186812,预测结果最小值：1221.6220418448754
对比919分数:0.9910714692899146
```

尝试

- 尝试

  Catboost 和xgboost

  Stacking融合都没有单模好 哈哈哈

体会

- 数据清洗很重要

  提分关键

- 特征筛选

  费时，最后的时候能提升一点

- 模型调参

  网格搜索

THANK YOU