出行产品未来14个月销量预测
---果粒橙团队解决方案说明

- 团队介绍
- 数据预处理
- 特征工程
- 模型构建
- 运行说明

果粒橙团队

- 团队成员
  - 沈伟臣 浙江大学 计算机科学与技术 学生
  - 戚家恒 浙江大学 计算机科学与技术 学生
  - 盛竹青 杭州安恒信息技术有限公司 设计师
- 最终成绩　A 156.059071 / B 149.694496

数据处理

- 缺失值处理
  - district_id2，5个缺失，根据district_id3进行填充
  - district_id4，若对应district_id3下所有district_id4均为-1，赋予一个新的district_id4
  - 效果不好最后没有采用
- price处理
  - product_quantity表中price属性有25条记录<-1，10957条记录为0，均应视作缺失。
- 预测后处理 对预测后的结果进行修正
  - 部分预测结果为负数，考虑用该商品前23个月的有效最小值替代，
    若前23个月均缺失，则置0
  - 有的商品从startdate开始有销售数据，而有的商品从cooperatedate开始有销售数据，
    预测时将在这两个时间点之前的结果置0

特征工程

- product_quantity orderattribute1
  - 每种出行产品对应唯一的orderattribute1，直接将该属性作为产品信息的一部分
- voter属性进行6级分箱，增强泛化性能
  - 划分[0,100,500,1000,2500,10000,inf]
- 根据历史订单统计产品的平均销售价格
  - sum(每单销售量*每单平均价格)/总销售量
- 添加自定义评分特征
  - 投票人数较多且用户评级高取2
  - 投票人数较多且用户评级低取0
  - 投票人数缺失取-1，其余为1
- 统计每个月节假日的天数
- 添加月份的one-hot结果和年份
- startdate, upgradedate, cooperatedate转为从该日期到当前时间的月数

特征工程

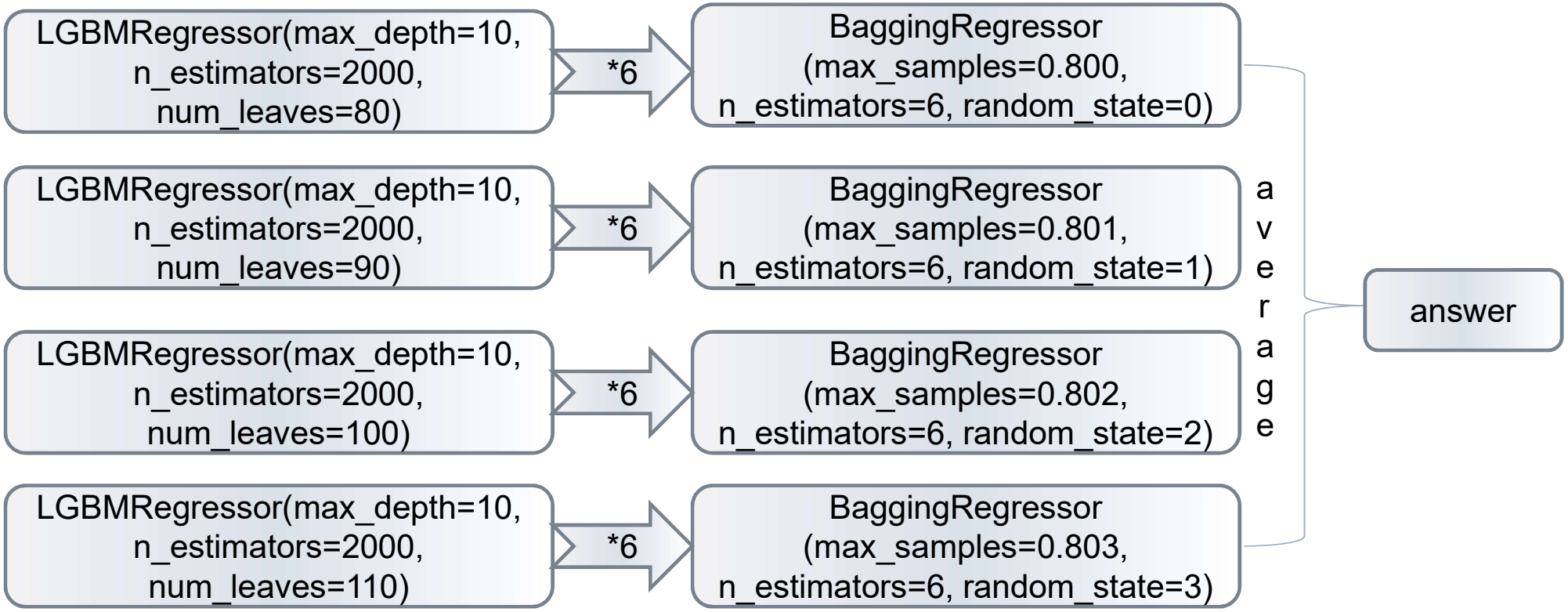
- product_info表原始特征
  - 'product_id', 'district_id1', 'district_id2', 'district_id3', 'district_id4', 'lat', 'lon', 'railway', 'airport', 'citycenter', 'railway2', 'airport2', 'citycenter2', 'eval', 'eval2', 'eval3', 'eval4', 'maxstock'
- product_info表处理过的特征
  - 'voters', 'startdate', 'upgradedate', 'cooperatedate'
- product_quantity表提取特征
  - 'orderattribute1', 'eval0', 'price'
- 其他日期相关特征
  - 'year', 'holiday', 'month_1', 'month_2', 'month_3', 'month_4', 'month_5', 'month_6', 'month_7', 'month_8', 'month_9', 'month_10', 'month_11', 'month_12', 'month79'
- 特征维度总共40维

模型构建

- 根据我们团队所使用的特征，采用线性模型并不能很好的对数据进行拟合，采用树模型作为基础模型
- 初期采用随机森林
  - 对若干决策树进行模型融合，提升泛化能力
  - 训练速度快，利于快速进行特征迭代
- 中期采用Bagging+GBDT
  - 泛化误差 = 偏差+方差+随机噪声
  - 使用GBDT对数据进行更精确的拟合，降低偏差
  - 采用Bagging来降低方差，进一步提高模型泛化能力
- 后期采用Bagging+LightGBM
  - 采用了对连续特征进行分箱的思想，并不精确划分特征，这从一定程度上带来了正则化的效果，同时支持离散特征的输入，是一种较好的GBM
  - 支持控制叶子结点数和最大深度等多种防过拟合的参数
  - 支持并行训练，训练速度显著快于传统GBDT

模型融合

最终使用了4组Bagging模型，每组使用6个LightGBM模型，共24个模型。4组Bagging的结果平均为最终结果。



运行说明

- 运行环境
  - Windows10
  - Python 3.5.2
  - lightgbm 0.1
  - scikit-learn 0.18.1
  - Pandas 0.19.2
  - Numpy 1.12.0
- 文件说明 /src
  - ctripfunc.py 特征处理函数
  - solution.py 主函数
  - prediction_lilei_20170320.txt 官方的提交样例，用于生成最终结果的格式
- 运行说明
  - 将原始数据文件放置在src目录下，直接运行solution.py即可，生成的l_bg46_lgb100_-1first.txt为最终提交文件