前海征信“好信杯”大数据算法大赛

团队:龙樱--PHM

目录

1 成员介绍

团队成员介绍

队长：张杰
简介：南京大学计算机系研二学生
邮箱：zhangj@lamda.nju.edu.cn
主页：http://lamda.nju.edu.cn/zhangj/

团队成员介绍

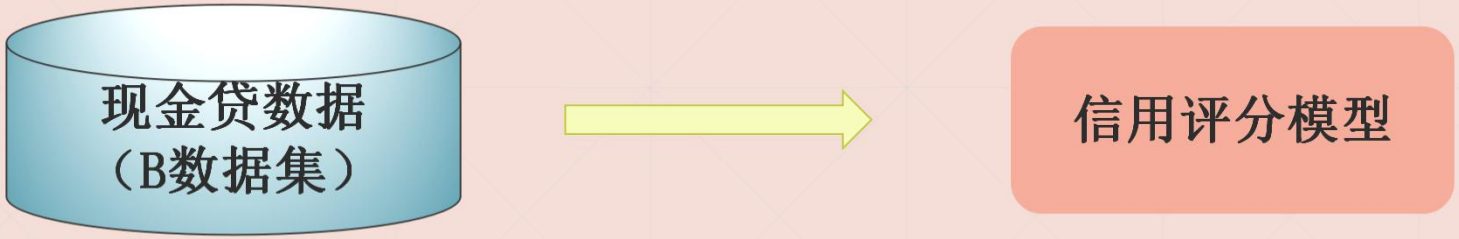
队员：李达
简介：北京航空航天大学可靠性与系统工程学院
系研二学生，技术指导-殷磊
邮箱：Lida_dreamer@outlook.com

2 问题背景

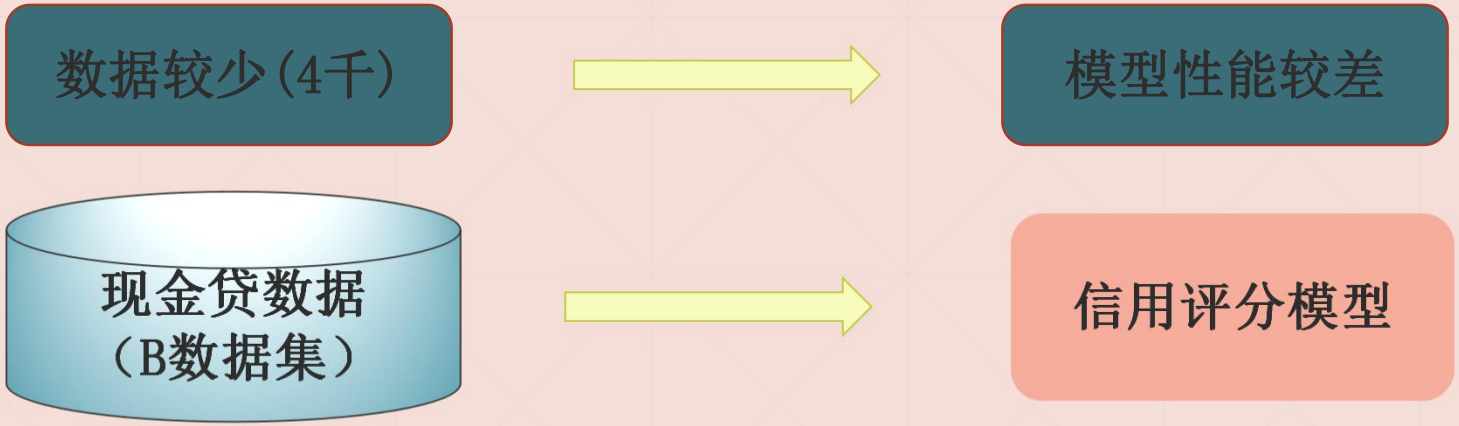
问题背景

现金贷的信用评分模型

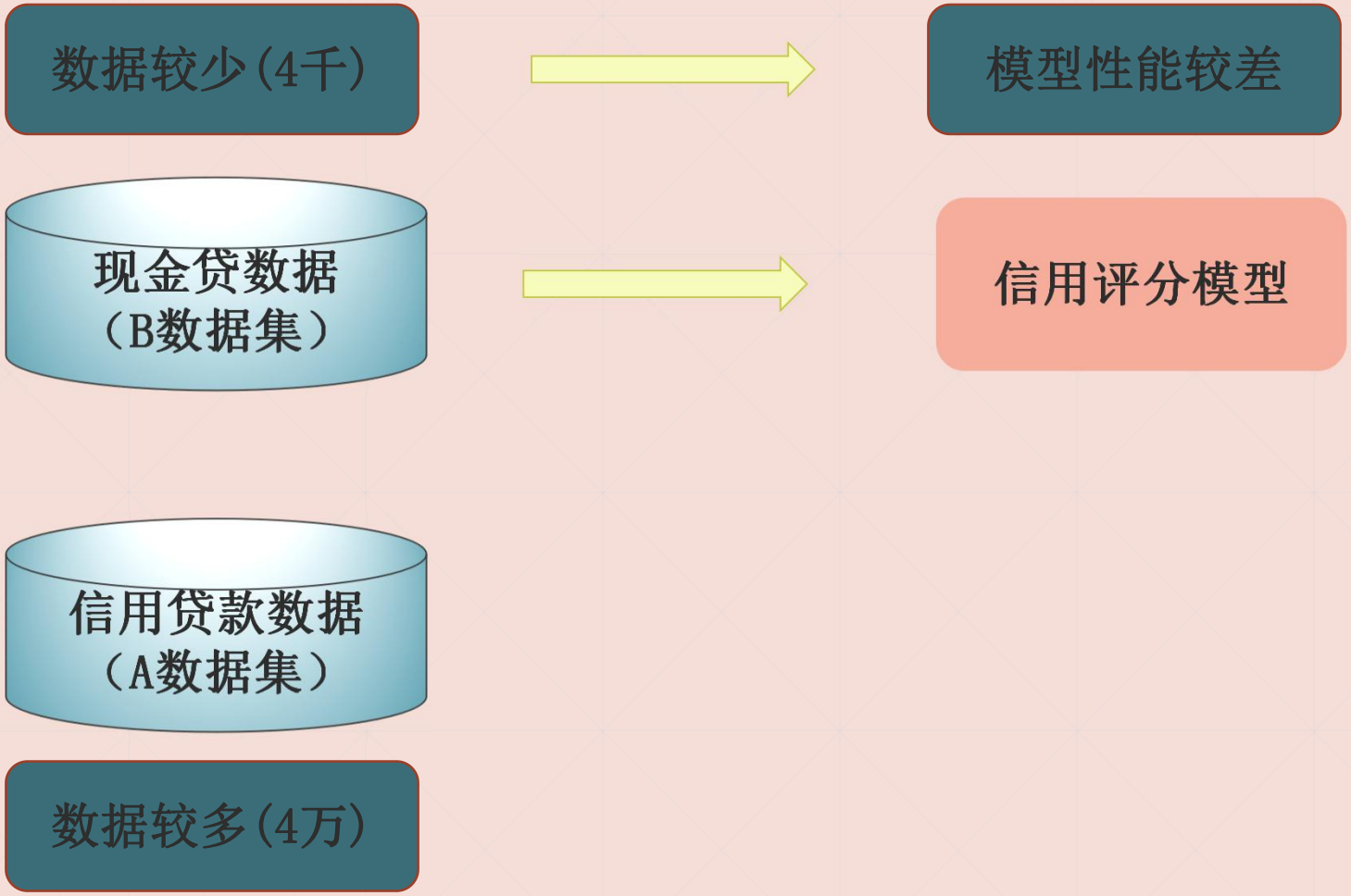
问题背景



问题背景



问题背景



问题背景


数据较少(4千)
模型性能较差
现金贷数据
（B数据集）
探索二者关系，将A中信息迁移到B中
信用贷款数据
（A数据集）
更好的信用评分模型
数据较多(4万)

问题背景（赛题）

数据较少(4千)

现金贷数据
（B数据集）

信用贷款数据
（A数据集）

数据较多(4万)

模型性能较差

探索二者关系，将A中信息迁移到B中

更好的信用评分模型

3 算法流程

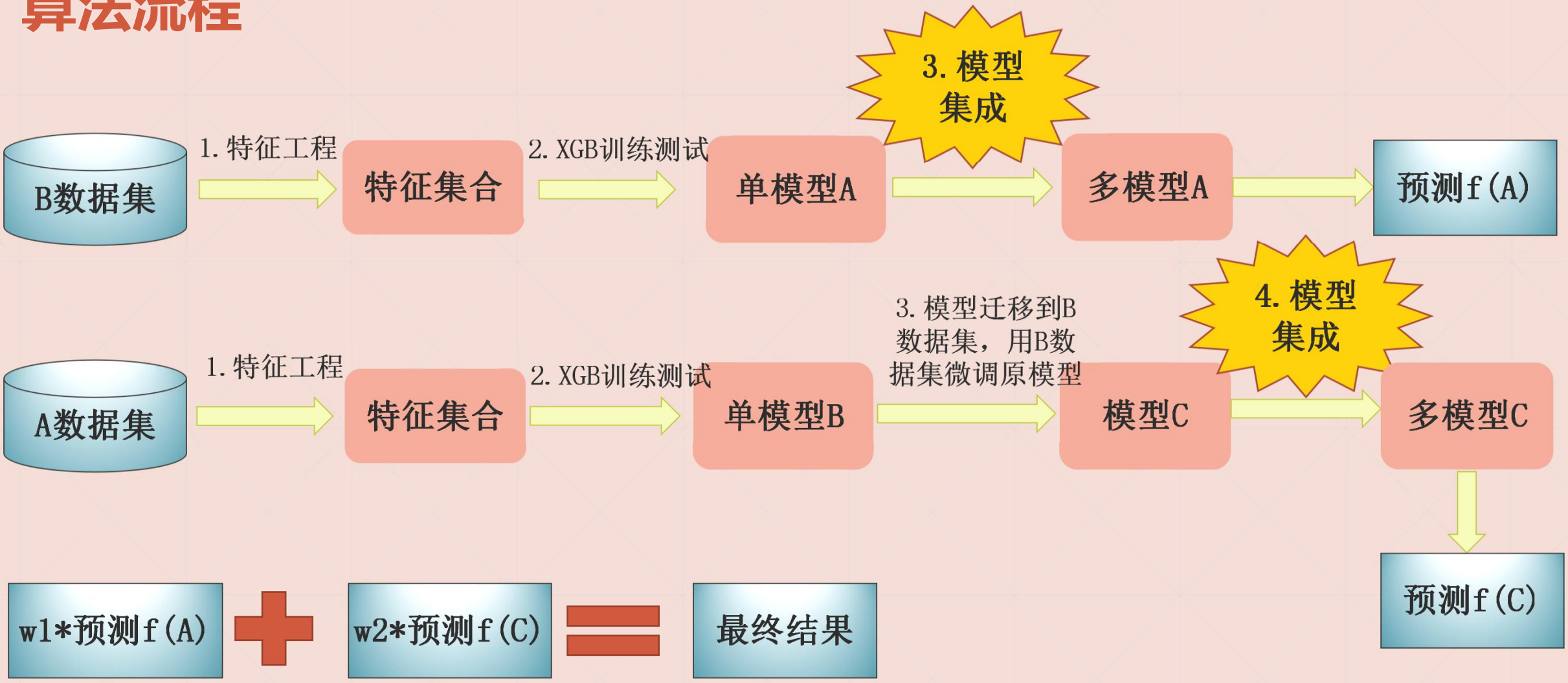
算法思路

① 对4千条业务B数据做简单特征工程，采用XGBoost加bagging的方式获得模型A

② 对4万条业务A数据采用相同的方式做特征工程，采用XGBoost建立模型B（以A数据为训练数据，B数据为验证数据调参），将模型B作为我们的迁移模型，再在模型B的基础上利用4千条业务B数据进行微调获得模型C，使模型C在能更好的拟合4千条业务B中的数据。

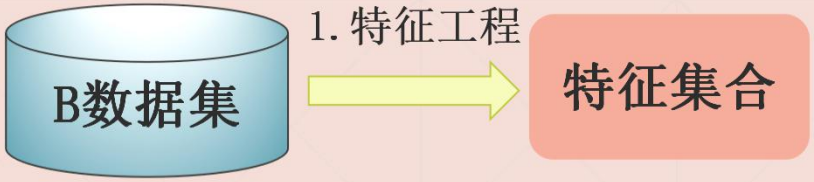
③ 利用模型A和模型C分别对测试集进行预测获得预测结果f(A)和f(C),然后对f(A)以及f(C)加权(f(A)附上较大权重，f(C)附上较小权重)获得我们最终的结果。

算法流程


B数据集
1. 特征工程
特征集合
2. XGB训练测试
单模型A
3. 模型集成
多模型A
预测f(A)
A数据集
1. 特征工程
特征集合
2. XGB训练测试
单模型B
3. 模型迁移到B数据集，用B数据集微调原模型
模型C
4. 模型集成
多模型C
预测f(C)
w1*预测f(A)
w2*预测f(C)
最终结果

算法流程1：获取多模型A

1.1 特征工程



- 统计每一个用户的缺失值的个数
- 统计每维特征中不同元素的个数，如果个数大于2，小于10, 就进行one-hot编码

1.2 单模型

- XGBoost作为基分类器,在4千条业务B数据上cv调参，选最好的参数进行训练测试。

参数名称	参数值	参数名称	参数值
booster	gbtree	objective	binary:logistic
eval_metric	auc	lambda	3
num_boost_round	130	alpha	5
subsample	1	eta	0.05
colsample_bytree	0.9	max_depth	4
base_score	0.25		

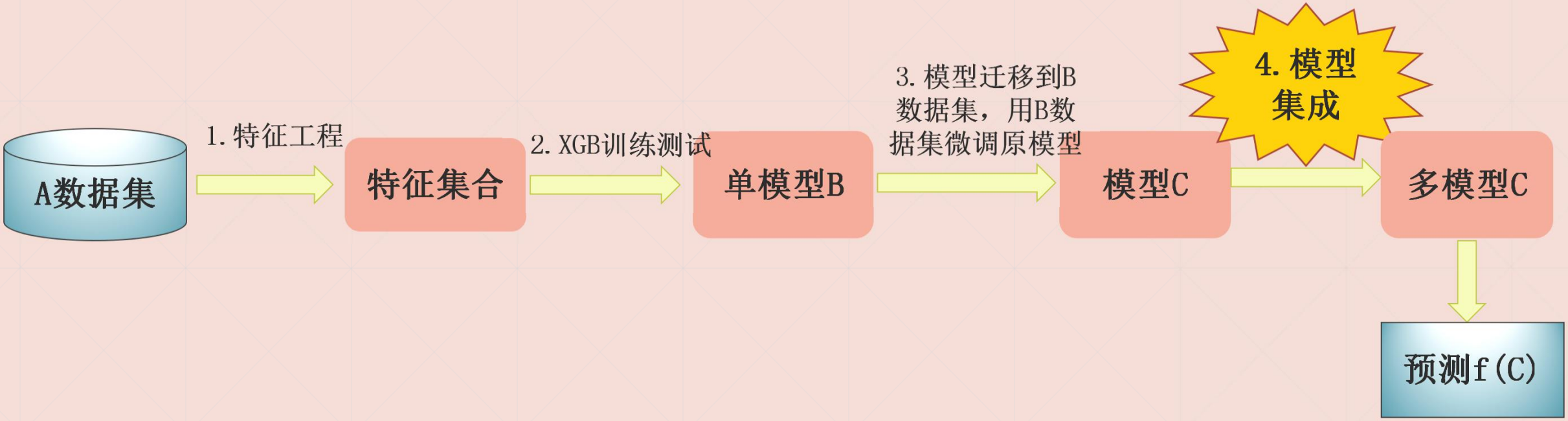
Table 1：XGB线上效果最好的参数，与线下cv结果最好的参数在迭代次数上略有不同

1.3 模型集成预测

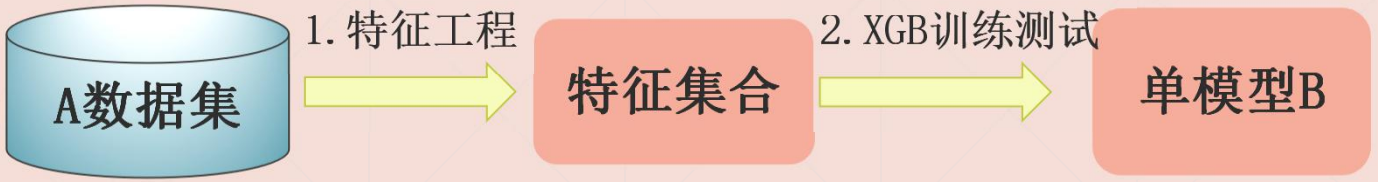
- 生成多个较大diversity的模型
  - XGB模型选用不同的参数(树的深度,列采样,正则化参数的设置等)
  - 单个模型选用不同的迭代次数

- 选用简单的集成方式(均值),获得预测结果f(A) → 线上0.6009

算法流程2: 获取多模型C

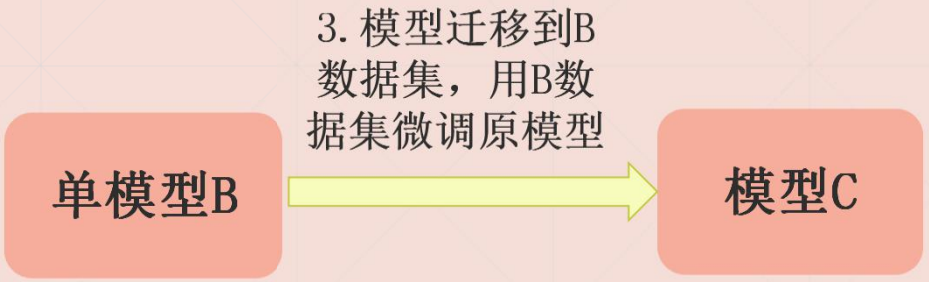


2.1 获取模型B



- 对4万条业务A数据做一样的特征工程
- 以4万条业务A数据为训练集，4千条业务B数据为验证集进行调参，将在验证集上获得最好效果的模型作为模型B

2.2 获取模型C



- 将模型B进行迁移，以B模型作为初始模型，利用4千条业务B数据在模型B的基础上进行cv参数微调获得模型C，使模型C在能更好的拟合4千条业务B中的数据

模型C = xgb.train(params,train_data,base_model）

- **paras: xgb的参数，包含树的深度，步长等**
- **train_data: 4千条业务B数据作为微调用的新数据**
- **base_model：模型B，表示以模型B为初始模型**

2.3 模型集成

- 生成多个较大diversity的模型
  - XGB模型选用不同的参数(树的深度,列采样,正则化参数的设置等)
  - 单个模型选用不同的迭代次数

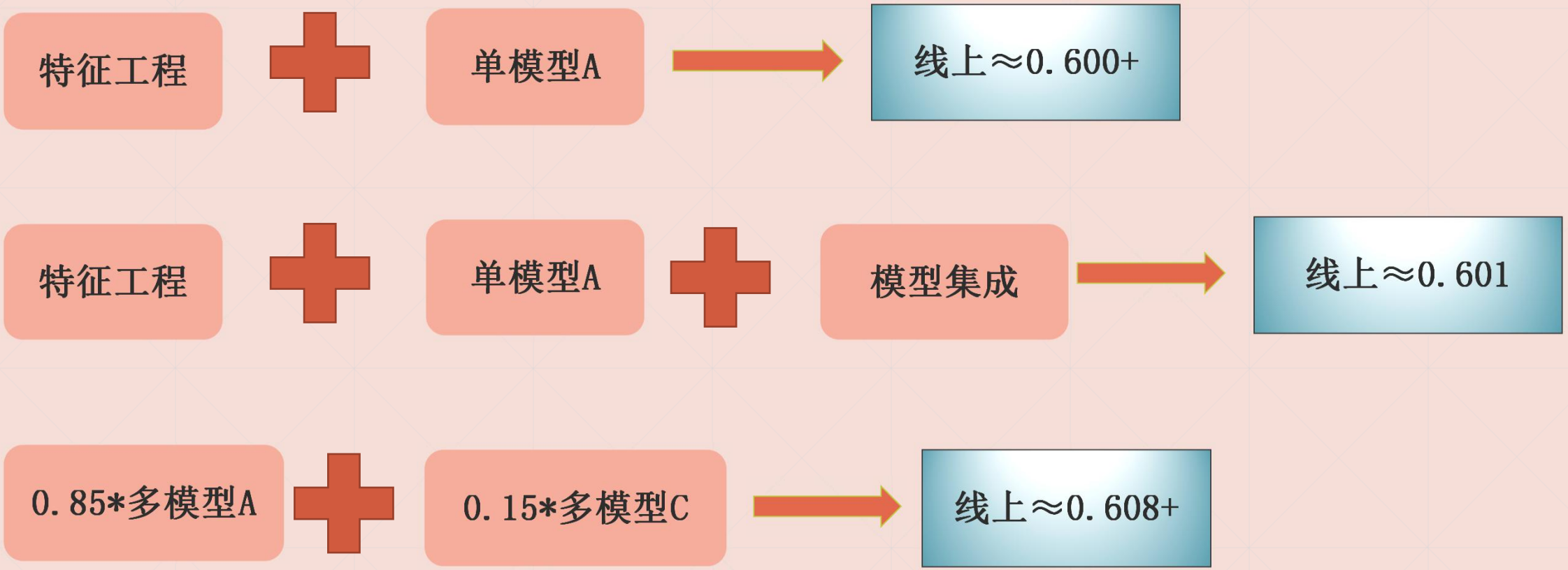
- 选用简单的集成的方式(均值),获得预测结果f(C)

算法流程3: 模型融合获得最终提交结果

- 多模型A是在目标数据上直接获得,和测试数据相关性大,对其赋予较大权重;
- 多模型C是根据业务A数据训练并在业务B数据上微调得到,能辅助业务B数据的判断和预测,赋予相对较小的权重。

4 实验结果

实验结果


特征工程
单模型A
线上≈0.600+
特征工程
单模型A
模型集成
线上≈0.601
0.85*多模型A
0.15*多模型C
线上≈0.608+

5 小结

小结

- 方法总结
  - 方法**以模型为主**(**目标域模型+迁移模型**),特征工程较少
  - 把**目标域训练得到的模型作为主模型1**,将在源域上训练得到的模型迁移到目标域并利用目标域数据进行fine-tune得到的模型作为**主模型2**，最后利用合理的加权方式得到最终结果,合理的考虑了迁移模型与目标域模型的关系。
  - **模型简单易懂**,通过目标域数据对源数据模型的fine-tune的来做迁移,算法部分由XGBoost和简单的bagging构成，构建两大主模型的关系的方式也很易懂。

小结

- 改进空间
  - 增加特征工程
  - 对最后一步的双模型融合的参数进行进一步的调优
    - 因为提交的次数原因，只进行了4次调优尝试
  - 集成时采用不同的模型,lightGBM,随机森林等

Thanks

请各位评委专家批评指正