机器学习导论大作业报告

基金间的相关性预测

李奉治 2016K8009929036

段江飞 2016K8009929011

姚依航 2016K8009929047

胡靖宇 2016K8009929050

陈子珩 2015K8009906006

1. **任务描述**

**赛题背景：**

基金间的相关性是基金的重要特征。根据金融学原理，一个基金组合的整体风险，不仅和其中各只基金的风险水平有关，还和这些基金之间的相关性有关。构造一个基金之间相关性小、或者说分散程度高的基金组合，能在保持一定收益水平的基础上，降低整体风险。因此，对基金之间的相关性进行预测，有助于我们构建一个好的基金组合。

**任务要求：**

根据给出的基金净值、基金业绩比较基准、对应指数行情、基金间相关性等数据，构建模型、算法进行训练。然后针对我们提供的测试样本，通过算法或模型预测出之后一段时间内基金间的相关性情况。

**数据描述：**

赛题数据包括200个基金的复权净值收益率，各只基金对应的业绩比较基准的收益率，基金间的相关性数据，同时期的重要市场指数收益率等。

收益率数据均为日度市场数据，直接获取于市场上公布的相关数据。基金间的相关性数据基于从对应日期开始向后61个交易日的市场数据和运营权重综合统计得出，按日度展示。

任务要求基于基金净值、基金业绩比较基准、对应指数行情，预测出测试数据期的下一个时间点、即2018-03-19对应的基金间的相关性。

训练数据包括四个csv文件

train\_fund\_return.csv(2015-09-29到2017-05-23):基金复权净值收益率

train\_fund\_benchmark\_return.csv(2015-09-29到2017-05-23)：基金业绩比较基准收益率

train\_index\_return.csv(2015-09-29到2017-05-23)：重要市场指数收益率

（Label）train\_correlation.csv(2015-09-30到2017-05-24)：基金间的相关性

测试数据：

test\_fund\_return.csv(2017-05-24到2018-03-16)

test\_fund\_benchmark\_return.csv(2017-05-24到2018-03-16)

test\_index\_return.csv(2017-05-24到2018-03-16)

test\_correlation.csv(2017-05-25到2017-12-14)

1. **实施方案**

**解决思路：**

采用xgboost和lightgbm的双层模型对于数据集进行特征提取进行预测。采用的原因是我们在调研的过程当中发现，在做这种预测性的工作的时候，lightgbm和xgboost有着特别优秀的表现。

**整体流程：**

整体采用两层的结构来完成任务，首先第一层是xgboost和lgboost来提取特征，第二层采取lgb使用第一层的提取的特征作为输入，然后来得到预测的结果。

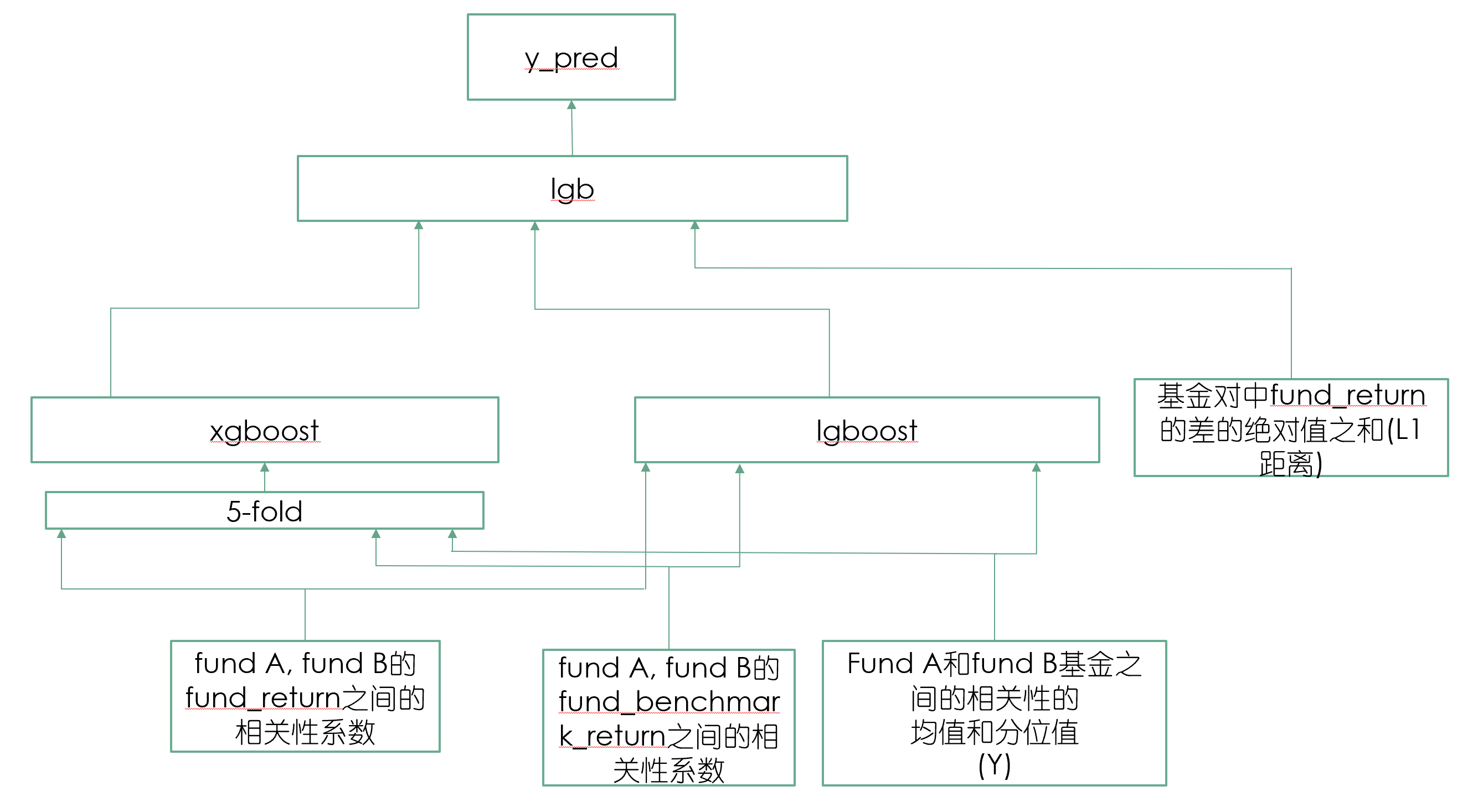
1. **具体实现**

**模型的设计和选择：**

调查各种算法，感觉值得一试的是随机森林和LGBT,在查关于LGBT资料的时候了解到了xgboost和lightgbm模型，最后决定尝试随机森林以及xgboost和lightgbm结合的模型，发现随机森林在自己的电脑上很难跑出结果，导致实验节奏会被拖慢，因此决定只用xgboost和lightgbm结合的模型

**具体模型：**

第一层的输入是基金对复权净值收益率之间的相关系数、基金业绩基准比较基准收益率之间的相关系数，和基金对间相关性的均值和分位值。对训练数据做5-fold后进行xgboost训练，并对所有的训练数据做lightgbm训练，两者的结果与基金对的基金复权净值收益率之差的绝对值之和作为第二层lightgbm的输入，得到y\_pred输出



**数据预处理方式：**

初步检查发现不存在数据为空的情况，因此不需要对数据进行清洗

但因为基金的相关性是时序相关的，要预测的是2018-03-19的基金相关性，为了预测的准确性，需要将测试集的数据也加入训练集里（在第二层中加入）

因此，第一层的训练集是train\_fund\_return.csv train\_fund\_benchmark\_return.csv train\_index\_return.csv train\_correlation.csv，第二层的训练集是第一层的训练集与train\_fund\_return.csv train\_fund\_benchmark\_return.csv train\_index\_return.csv train\_correlation.csv

**数据集划分方式：**

总体采用的数据集如下:

train\_fund\_return.csv train\_fund\_benchmark\_return.csv train\_index\_return.csv train\_correlation.csv train\_fund\_return.csv train\_fund\_benchmark\_return.csv train\_index\_return.csv train\_correlation.csv

其中

train\_fund\_return.csv train\_fund\_benchmark\_return.csv train\_index\_return.csv

train\_correlation.csv

作为第一层模型的输入，

train\_fund\_return.csv train\_fund\_benchmark\_return.csv train\_index\_return.csv

train\_correlation.csv

作为第二层模型的输入。

**对于correlation种类的选择：**

Correlation的种类有三种，分别是pearson, kendall和spearman，于是在计算基金对fundA-fundB的利益率之间的相关性时可以通过改变参数来计算不同的correlation，经过尝试，发现不同correlation之间得到的效果差距不大，而且由于kendall correlation和spearman correlation计算前要先排序，因此跑的速度会变得特别慢。并且，统计学中，只有pearson相关性计算的是两个连续型变量之间的相关系数，因此，最终决定采用pearson相关性

**对于读取数据的数据进行处理：**

读入数据时，发现都是<1的小数，因此将读入的数据都乘一个倍数，来提高精度，分别为10，100，1000，10000，100000，经过尝试发现，乘10000得出的预测准确率最高，因此采用了乘10000

**采用了5-fold：**

将训练数据分成5分，每次取4分进行训练，训练出5个模型，对验证级有5个预测结果，取均值送入xgboost

**lgb的调参：**

nrounds, feature\_fraction, learning\_rate, nround,加feature\_fraction。我们采取的方式是贪心式的调参，即在每一个超参数调到最佳后再进行下一个超参数的调参。

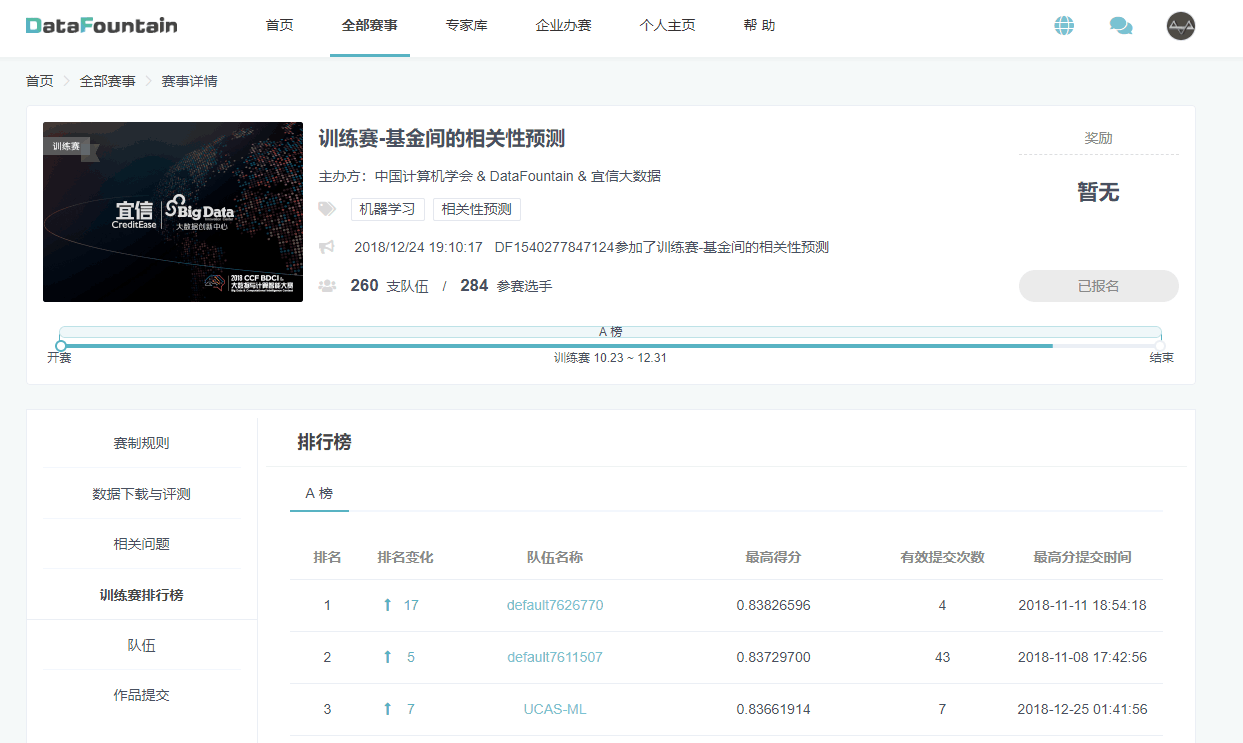
**stacking：**

因为距离预测时间最近的数据最重要，因此将最近的时间迭取多次来得到数据，对t1,t2的时间间隔和stack的次数进行调整，选择结果最优时的值(15,20)

1. **实验结果及分析**

我们经过多次调参数、修改特征后，最后在网站上的排名上升到了top3.

最终我们的得分为0.83661914

****

1. **总结展望**

本次实验我们采用的是先调研后代码的形式，没有一开始就进行代码的撰写，而是参考了许多其他的在AI的比赛方面很有经验的同学们的经验，参考了他们的模型，融入到了我们的模型之中，使得我们的模型效果提升了许多，并且使得我们的最终的结果比较满意。但是我们不光光参考了别人的模型，我们自己的模型也是有自己的创新，我们采用了5-fold的方法来增强我们学习的效果和最终的结果。但是我们的调参也是有着可以改进的地方，比如我们的调参采用的是贪心式的，得到的是局部最优，我们逐个调整每一个超参数，得到最优值，再调整下一个超参数。如果时间充足的话，其实采取整体遍历式的调参才是更好的。

接下来是我们整体的pipeline：

1.分析问题：

2.数据预处理：初步根据dataset提出特征，作为model的input

3.特征工程：

特征提取（主要从fund\_return和fund\_benchmark\_return中提取特征）

特征组合：3个从fund\_return中的features和fund\_benchmark\_return中提取的一个特征，以及correlation中提取的一个feature

4.构建模型：xgboost+lightgbm

5.参数调试

6.评估模型：

上交数据集后的在线测试结果，本地下载数据集的测试结果，在网站上的rank（top3）

**六、 任务分工**

**段江飞**：特征工程，添加 k-fold，前期调研xgboost，编写lgb和xgb模型代码

**李奉治**：调整xgb和lgb的参数，配置服务器，选题调研和组织讨论

**姚依航**：特征调整，编写lgb和xgb模型代码，ppt制作，文档撰写

**胡靖宇**：ppt制作，文档撰写

**陈子珩**：ppt讲解，文档撰写

**共同完成**：学习xgboost和lightgbm包的使用，随机森林和RNN等算法的实现