1. 直接将文件夹中所有文件放置在一起，然后运行Final\_B.py即可

2. 运行环境：

操作系统：win7

Python版本：3.6.2

所需包： sklearn\_pandas、sklearn、lightgbm、xgboost、keras(tensorflow)

3. 输入文件说明：

|  |  |  |
| --- | --- | --- |
| d\_train\_withA\_20180129.csv | 训练集数据 | 把原始的d\_train\_20180102.csv删去id 4228一行（血糖38.43），并将A榜结果合并所得。 |
| fea\_level1.csv | 额外制造的特征表1 | 来源于原特征四则及幂运算并由线性模型（贝叶斯岭回归）评估coef\_后得到的特征，第一列为特征名，第二列为\_coef总和 |
| fea\_poly\_LR.csv | 额外制造的特征表2 | 来源于原特征多项式运算并由线性模型（贝叶斯岭回归）评估coef\_后得到的特征，第一列为特征名，第二列为coef\_总和 |
| d\_test\_B\_20180128.csv | B榜测试集 | 未改动 |

4. 输出文件说明：运行Final\_B.py后会生成名为sub\*\*\*\*\*\_\*\*\*.csv的带时间戳的文件，即为最后提交的结果文件

5. 算法说明：

整体算法框架结合了c271828的10xCV取平均和吴飞的多模型加权求和方法。

添加了由线性模型评估过的特征、删去了部分原始特征（详见脚本中的RM\_LIST变量）、添加了一个血液检测指标的PCA、minmaxscaler处理了“甘油三酯”特征。

删去了异常血糖值id 4228，并将A榜血糖值加入到了训练集中。

最终使用了6641个样本和114个特征，使用lgb、xgb、NN、线性模型（弹性网回归）进行加权平均。

注：

c271828的技术圈代码：

https://tianchi.aliyun.com/forum/new\_articleDetail.html?spm=5176.8366600.0.0.57eaf00d3A8a7p&raceId=231638&postsId=3761

吴飞的技术圈代码：

<https://tianchi.aliyun.com/forum/new_articleDetail.html?spm=5176.8366600.0.0.46c5aaf6hqt9G7&raceId=231638&postsId=3873>

感谢两位大佬和豆腐大佬在技术圈的分享。