天池精准医疗大赛

1. 特征工程
2. 由于乙肝特征指标与血糖并没有直接关系，而是当患上乙肝后，比较严重的时候会影响到血糖值，这里乙肝体检的缺失值超过70%，所以直接删除与乙肝相关的5个特征。
3. 将性别特征进行ONE-HOT编码，删除id，体检日期特征。
4. 根据论文和医学院同学提供的建议，添加7个具有医学统计意义的特征，分别为：AST/ALT, GFR, 中性粒细胞绝对值, 淋巴细胞绝对值,单核细胞绝对值，嗜酸细胞绝对值，嗜碱细胞绝对值。其计算公式见参考文献及代码。
5. 由于训练样本中存在血糖值为38的值，这里作为异常值处理，将该条样本删除，剩余5641条样本。
6. 对所有特征两两相乘、相加、相除等运算，生成上千个新特征，然后利用XGBoost特征重要度排序，最后选出最重要的10个特征，将这10个特征添加到原始特征中。
7. 经上面的特征处理，最后得到52个特征，利用XGBoost训练，进行特征重要度排序，按照特征重要度，从一个特征开始依次增加特征，最后得到52中情况的MSE，发现随着特征数目的增加交叉验证的误差先降低，然后再降低，在特征数目为32时，MSE达到最小值。因此我们最终选择32个特征。
8. 模型
9. 利用XGBoost单模型对以上得到的数据进行交叉验证，选出最优参数。
10. 利用lightgbm模型训练。
11. 最后分别得到两个模型预测的结果，提交的结果是xgboost模型得到的结果，成绩为0.6416，142名。然而发现lightgbm的结果成绩为0.6259，可以到达63名左右。