1 问题描述

这个比赛的任务是预测妊娠期妇女是否患有妊娠期糖尿病，label 只有一列，0 表示未患病，1 表示患病。一共有 1000 条训练样本，85 维特征。线上测试样本为 200 条，采用 F1 值来评价结果好坏。由于样本较少所以很容易出现过拟合问题。在 85 维特征中，有 30 个是身体指标特征，诸如年龄、身高、体重、BMI、胆固醇指标等等，其他 55 个是基因特征，基因特征有 3 中取值 0,1,2 代表生物学中的 AA、Aa、aa。下面介绍赛题思路。

2 特征工程

1 连续特征类

查看数据的分布，采用不同的填充办法，比如平均值、中值、众数等

以平均值为标准值，添加和平均值的差值，以及差值的绝对值

对连续特征做归一化处理，由于后面需要对特征之间做运算，所以需要把 0 替换成极小值

归一化处理后做加减乘除和反除，以得到组合特征。

2 离散特征类

离散特征采用 one-ht 编码

编码后的特征做与、或、异或、同或处理

3 特征筛选

线性回归筛选特征+非线性 XGBoost 筛选

3 模型

线性筛选的特征用线性模型，非线性筛选的特征用非线性模型

4 融合

这里采用简单的加权融合的办法，最终结果线上 F1 值 0.6429，复赛排名 57。总体上来说思路比较简单，传统，和排名靠前的选手有很大的差距。下面根据答辩的情况，对每位选手的方案作出总结

5 对答辩选手的总结

在数据填充中，选手不是采用简单的平均值和中位数的填充办法，而是 Nuclear 和范数填充的办法。对于特征处理，选手先去掉了相关度非常小的几个特征以剔除噪声。然后分析所有特征对于 label 的单调性，实际上也类似于相关度，筛选出强特征和弱特征，在对强特征之间做组合，又一次通过单调性筛选出强特征。一般迭代 4,5 次可以得到不错的结果。这个的筛选指标不仅仅可以使用单调性、相关度等，还可以使用 IV 值，REFCV 的办法来筛

选。这里面 ACEID 也是一个非常不错的特征，但是缺失值很多，容易发生过拟合，选手们的办法是不填充缺失值或者干脆这个特征不用。

在模型选择上大家都大同小异，LR、SVM、RF、GBDT、XGB、LGB 都试一遍，看看效果，有一个选手用了 Catboost 的模型(不太了解)。在调参上面，有选手采用了遗传算法来调试出最佳的参数，思路不错。

后处理，，对于不确定的结果，正例和负例相接近的(患病概率为 0.49~0.50 之间的也预测为患病)，可以在一定程度上提高成绩。