華東師範大學

EAST CHINA NORMAL UNIVERSITY

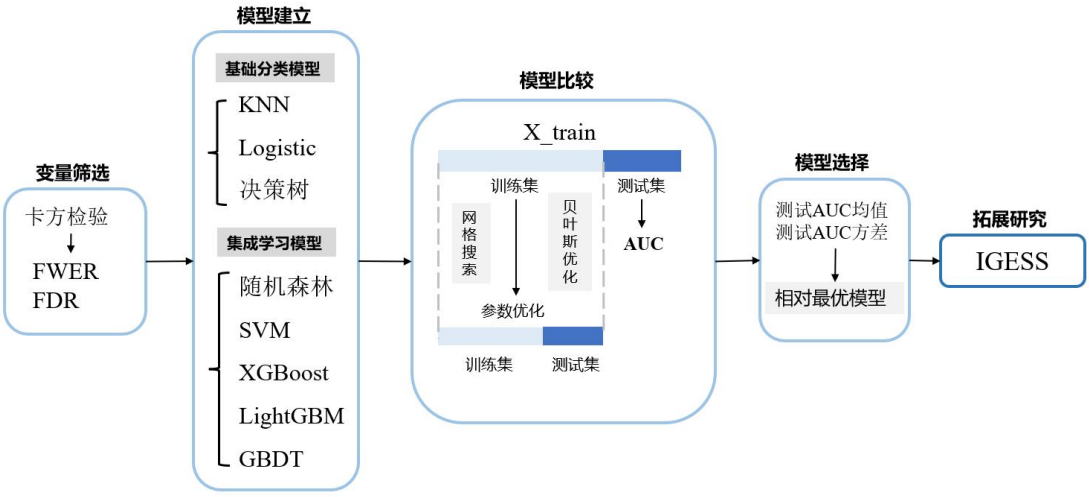
生物医疗中的统计与方法

小组作业 1——疾病预测报告

报告人：秦佳玥、王英璇、吴欣、陈明真

目 录

研究思路图



一、 筛选变量

由于数据集中有 248409 个 SNP，变量维数很大，而样本量只有 4081 个。这种情况不利于模型的建立，所以本小组成员首先考虑如何对变量进行有效筛选，即：既要保证数据维数的可操作性，又要保证筛选出的变量与疾病有较高的相关性。这一步操作可以减少模型的复杂度，避免过拟合，并且有利于模型解释。如果选择对的特征子集，可能会使得模型的准确度得到提升。

通过资料查阅，我们发现一些特征筛选方法如下图 1 所示。

本小组最先考虑的是嵌入法。没有考虑包裹法的原因是，虽然它可以将要使用的学习器的性能作为特征子集的评价标准，为其“量身定做”特征子集，但是其计算量十分庞大，考虑到数据集自身的样本量，我们没有尝试此方法。

我们尝试利用 LASSO 和 GBDT 进行特征筛选，并且 sklearn 中有提供 SelectFromModel 函数，可以直接调用模型挑选特征。虽然可以顺利运行，但是我们发现，利用 SelectFromModel 函数需要主观地确定需要筛选出多少变量。在缺乏经验以及理论支撑的情况下，我们最终没有选择此方法。

最后，基于分类数据，本小组采用卡方检验得到每个 SNP 的 p 值，再结合多重假设检验的控制方法，对变量进行了筛选，具体如下所示。

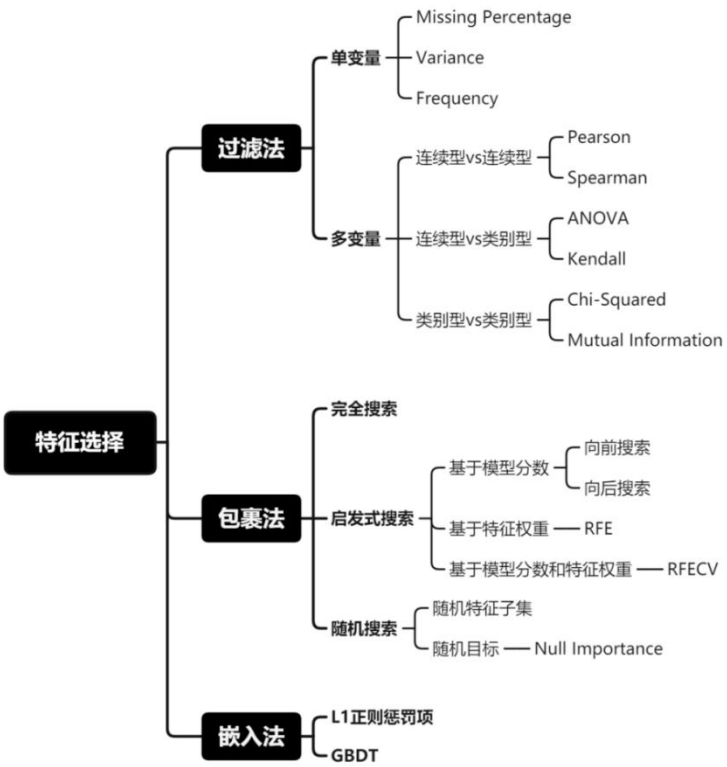


图 1：特征筛选方法

1.1 卡方检验

卡方检验可用于检验两个类别型变量之间的相关性。它建立的零假设是：两变量之间不相关，检验过程用卡方值衡量理论和实际的差异程度。卡方值的计算公式如下：

$$X^2 = \sum \frac{(observed - expected)^2}{expected}$$

卡方值越高，说明两变量之间具有相关性的可能性越大。

基于此理论，我们在 R 语言中将每个 SNP 和表现型 Y 组合成一个列联表，对其做卡方检验，可以得到相应的 p 值。

在卡方检验中存在 248409 个原假设，为了避免错误地拒绝过多的原假设，造成很多假阳性，通过以下两种方式控制第一类错误不要过度膨胀。

1.2 控制 FWER 的方法

FWER 为在 M 个假设检验中，发生至少一个第一类错误的概率。Bonferroni 校正为了将 FWER 控制在水平α，拒绝了所有 p-value ≤ α/M 的原假设。

基于此理论，我们筛选出 38 个 SNP（如图 2 所示）。

```
# 利用Bonfronni法进行筛选
sum(chisq_res<0.05/248409)
x <- X_train[,which(chisq_res<0.05/248409)]
dat <- as.data.frame(cbind(x,y_train))
```

```
[1] 38
```

图 2：基于 Bonferroni 方法筛选变量

Bonferroni 校正方法较为保守，真实的 FWER 通常低于名义 FWER，这往往会导致拒绝更多的原假设，造成功效过低。由图 3 可以进一步验证，在 m=248409 时，功效不理想。

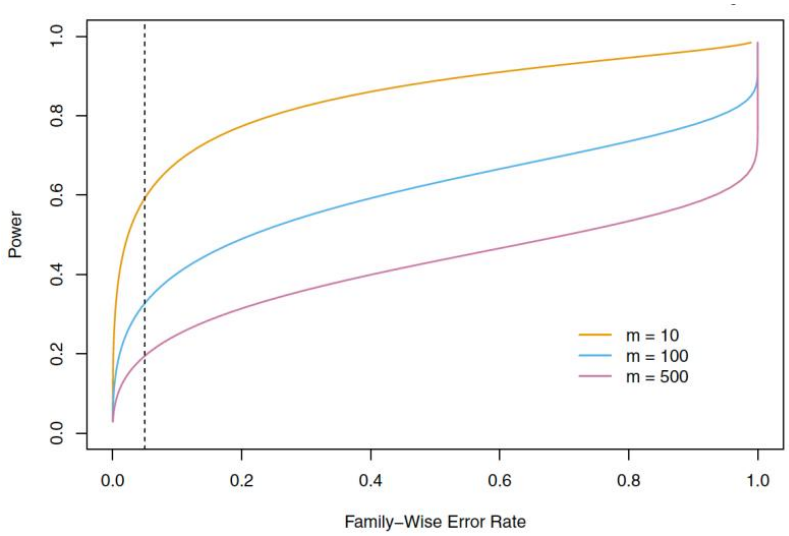


图 3：在一个 90%的原假设为真的模拟中，对于不同的原假设个数 m，power 随 FWER 的变化情况。虚线为 FWER=0. 05

1. 3 控制 FDR 的方法

FDR 是指在多重假设检验中第一类错误所占比例的期望。BH 方法将每个假设检验的 p-value 进行排序，找到满足条件的最小 k 值，拒绝前 k 个最小 p-value 对应的原假设，以此使得将 FDR 控制在水平α。

相较于 FWER，FDR 允许稍多一些的第一类错误来换取更大的功效。FDR 往往小于等于 FWER，所以拒绝的原假设更少。

基于此理论，我们选出了 65 个 SNP（如图 4 所示）。

```
p3<-p.adjust(unlist(p_value), method = "BH")
sum(p3<0.05)
X_BH <- X_train[,which(p3<0.05)]
```

```
[1] 65
```

图 4：基于 BH 方法筛选变量

经过进一步讨论，本小组更倾向于运用通过控制 FDR 方法筛选出的 SNP。具体原因是：我们认为当对生物医疗方面的数据进行变量筛选时，可以适当的允许一些冗杂变量的进入，以保证更多真正会影响疾病性状的 SNP 进入模型。

二、 模型选择的初步探索

在这一部分，本小组进行模型选择的初步探索，即建立不同的模型（建立模型时只应用网格搜索对参数进行初步粗略调整），并对模型进行比较。

具体步骤为：

1. 将数据集划分为训练集和测试集，在训练集上分别建立 p 种模型

2. 由于 p 种模型需要进行参数优化，所以利用交叉验证法和网格搜索寻找最优参数，操作如下所示：

（1）将训练集进一步分为五折，在其中四折上分别利用三个模型的网格点对应的参数建立模型，在剩余一折上算出目标值 AUC

（2）重复（1）操作五次，可以分别得到在对应网格点参数下的 AUC 均值

（3）对所有网格点进行上述操作后，可以选择出 p 个模型下的局部最优参数值

3. 利用局部最优参数值在训练集上分别建立 p 个模型，在测试集上计算出测试 AUC 值

4. 重复以上步骤 50 次，对 p 个模型的 AUC 求均值，即可比较出三个模型预测性能的优劣。

对于上述过程中网格点的选择，本小组也进行了一定的思考。网格搜索是对所有的网格点进行遍历，并计算出在对应参数下模型的测试 AUC，所以网格搜索可以保证在寻找范围内的局部最优性。但是考虑到遍历时计算量的庞大，寻找范围的设定一般不能过大。而寻找范围的设定往往是基于经验，参数优化不能一蹴而就，需要根据前面的搜索结果去调整寻找范围。所以本小组在多次尝试后选定了较为合适的寻找范围，保证了通过网格搜索出的参数下模型的 AUC 值不会明显低于最优水平。在后续分析中，我们会进一步对参数选择这一步骤进行优化。

2.1 基本分类模型

为达到根据 SNP 值预测表现型 Y 的目的，我们利用上述进行变量筛选后的数据建立分类模型。我们首先考虑建立一些常见的基本分类模型，如：KNN、Logistic 回归和决策树等，这些模型较为简单常见且易于解释，是非黑箱模型。

模型的简要原理及优点如下所示。

2.1.1 KNN

KNN（K-Nearest Neighbors）是一种经典的有监督机器学习方法，它通过测量不同数据点之间的距离，将新数据点分类到最靠近的 K 个已知类别中的一类。它不需要对模型进行假设，也不需要对模型进行参数估计，是一种非参数方法。由于 KNN 算法思想相对简单，易于实现，并且相对而言较为普适，我们选择该方法作为基本方法进行尝试。

2.1.2 Logistic Regression

Logistic 算法是一种常用的二元分类算法，它通过建立一个逻辑函数来预测一个事件的发生概率。其原理是将输入特征通过线性加权求和，再将结果经逻辑

函数转换为 0~1 之间的概率值，达到分类的目的。由于 Logistic 算法计算速度快、可解释性强，且可以很好地处理二元分类问题，因此它适用于我们的研究场景。

2.1.3 决策树

决策树（Decision Tree）以树结构（包括二叉树和多叉树）形式来进行预测和分析，通过将数据集递归地分解为子集，直到子集中的数据属于同一类别为止，构建出决策树模型。决策树算法在众多机器学习的算法中相对易于理解和解释，对于大型数据集具有良好的可扩展性，同时也是其他一些集成学习的基学习器，因此我们选择了决策树作为基本模型的一种。

本小组首先建立以上三种模型并进行比较。三种模型的预测效果如图 5 所示。其中，KNN 算法得到的 AUC 显著低于另外两种模型。我们发现：KNN 算法产生了维度灾难，当 p 很大时，与观测值最接近的 K 个观测可能在 p 维空间中距观测值很远，导致预测非常差，从而产生一个很差的 K 近邻拟合，虽然方差相对较小，但是偏差较大。决策树的预测效果不是十分理想，其方差较大，稳定性较差。并且本小组猜测样本标签值的不均衡可能会导致决策树的预测效果较差。福布斯杂志在讨论逻辑回归的优点时，甚至有着“技术上来说，最佳模型的 AUC 面积低于 0.8 时，逻辑回归常明显优于树模型”的说法。从目前的结果来看，Logistic 回归的效果相对较为理想。

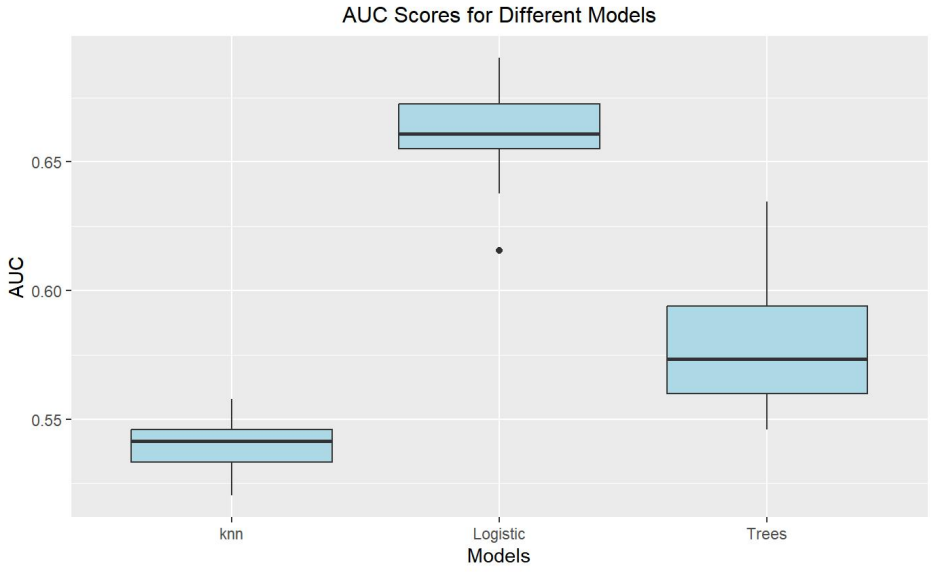


图 5：三种基本分类模型的预测性能对比

2.2 基于集成学习的分类模型

由于决策树的预测效果没有达到理想状态，但其 AUC 均值显示出弱学习器的特征，据此本小组尝试了以下一些集成学习方法进行建模，希望通过小幅度地增大方差来换取较大幅度的偏差减少。

模型的简要原理及优点如下所示。

2.2.1 随机森林

随机森林（Random Forest、RF）一种集成学习算法，基于决策树构建，通过随机选择特征子集和样本子集进行训练，最后通过投票或平均来进行预测。该模型在高维数据和大规模数据集上表现良好、具有较好的稳定性。因此，本小组尝试用随机森林进行建模。

2.2.2 SVM

支持向量机（Support Vector Machine，SVM）是一种监督学习算法，用于二分类和多分类问题。它的目标是找到一个最优的超平面，将不同类别的样本分开，并最大化样本与超平面之间的间隔。该模型在高维空间中表现良好，在处理大规模数据集时效率较高，且可以通过选择不同的内核函数来适应不同类型的数据，因此本小组尝试用 SVM 进行建模。

2.2.3 XGBoost

XGBoost（eXtreme radient Boosting）是经过优化的分布式梯度提升库，在并行计算效率和预测泛化能力上都表现非常优秀。由于我们的数据量大，且注重预测性能，因此本小组尝试用 XGBoost 方法。

2.2.4 LightGBM

LightGBM（Light Gradient Boosting Model）是一种基于梯度提升决策树的框架，通过使用基于直方图的决策树学习算法来提高效率。该模型在大规模数据集上训练速度快、内存占用低、具有较好的准确性和泛化能力，因此，本小组尝试

用 LightGBM 进行建模。

2.2.5 GBDT

梯度提升决策树（Gradient Boosting Decision Trees，GBDT）是一种集成学习算法，通过串行训练多个决策树来提高预测性能。并且，每个决策树都试图纠正前一个树的预测误差，使得模型在训练过程中逐步改进。由于该模型可以处理类别型变量和非线性关系，因此，本小组尝试用 GBDT 进行建模。

在这一步，本小组通过设置随机数种子来保证在每一次循环内，建立这 5 种集成学习模型时所用的测试集与前述建立基本分类所用的测试集是相同的，以便于可以合理地对进行 8 种方法的预测效果进行比较。（在代码部分，为了保证简洁性，我们在一个 for 循环下面展示了 8 种方法的模型初步建立。）

本小组分别建立以上 5 种集成学习模型并进行比较，具体步骤如前述所示，几种模型的预测效果如图 6 所示。本小组发现：利用集成学习后模型的预测性能有了显著提升。五个集成学习模型的 AUC 均高于 0.6，且 AUC 中位数都高于 0.65。由于此处只进行了粗略调参，所以我们暂时无法断定 GBDT、LightGBM、RF 和 SVM 这四种方法中哪个最优，需要进一步进行分析。

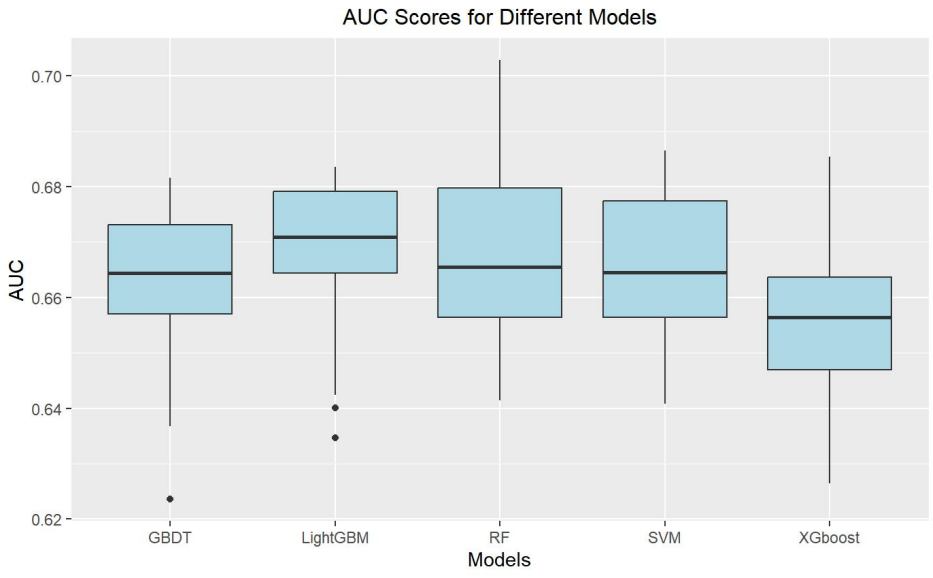


图 6：5 种集成学习方法的预测性能比较

综上所述，基于对 8 种分类模型的比较（如图 7 所示），集成学习模型的预测效果更为突出。

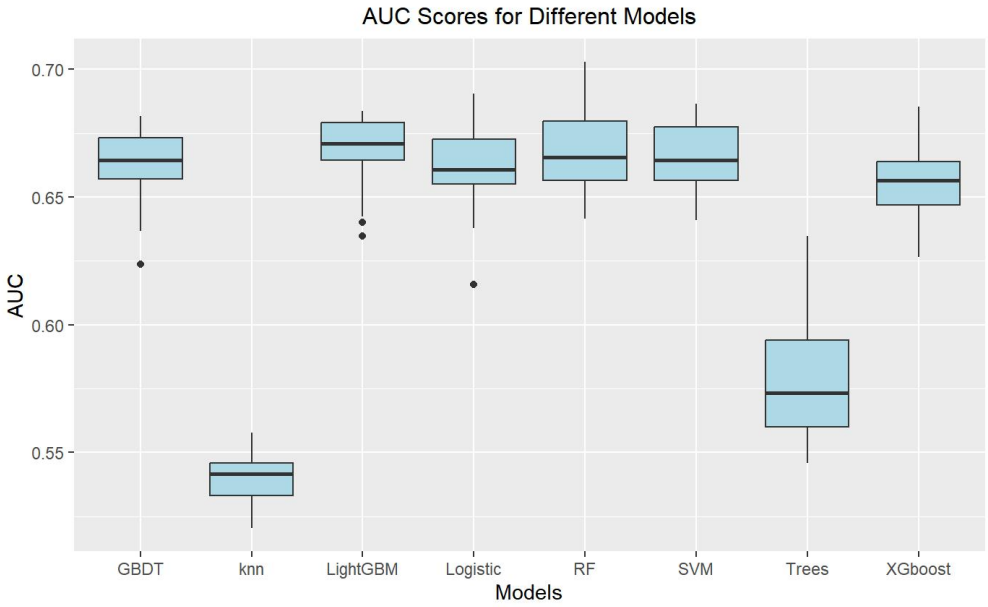


图 7：7 种集成学习方法的预测性能比较

三、 模型的进一步比较

在这一部分，本小组运用贝叶斯优化对 GBDT、LightGBM、RF 和 SVM 这四种方法进行更大范围的调参后，再次进行比较。

在数据量和参数范围都较大的情况下，利用网格搜索进行参数优化的速度过于缓慢，而贝叶斯优化采用高斯过程，会考虑到之前的参数信息来不断地更新先验信息，利用已有的先验信息去找到使目标函数达到全局最大的参数。该方法迭代次数少，计算速度快，可以节约计算开销。

本小组主要利用 python 中 scikit-optimize 包的 BayesSearchCV 函数进行参数优化，该函数基于交叉验证实现的贝叶斯超参数寻优，可以设置超参组的个数，运算量可控，有利于控制计算机运算开销和运算时间。优化结果如图 8 所示，可以发现 GBDT 模型的测试 AUC 中位数高且方差最小，其运行一次的时间为 107.3734s，表现较为良好。与之相比，LightGBM 模型的测试 AUC 中位数略低于 GBDT 模型，方差也略大于 GBDT 模型的方差，约为 0.0075。但是 LightGBM 模

型运行一次的时间为 63.07s。在综合考虑时间效率和预测效率之下，本小组成员认为应选择 LightGBM 作为最终的预测模型。

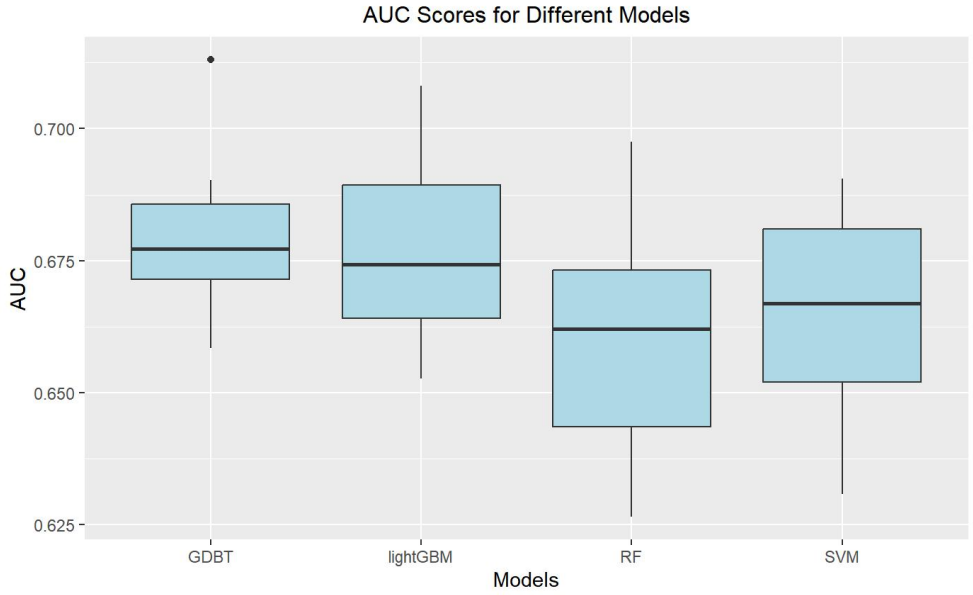


图 8：基于贝叶斯优化的 4 种集成学习方法的预测性能比较

最后，我们在全部数据集上分割训练集和测试集，在训练集上训练各种参数下的 LightGBM 模型，在测试集上计算相应的 AUC 值。循环 100 次后选择出最优参数值，并在全部数据集上进行建模，从而得出最终预测模型。带入数据集 X_test，得到最终预测集 y_pred。

四、 拓展思考

考虑到上述的筛选变量只利用了 CD.Rdata 数据而未考虑 CD_P.Rdata 数据，并且模型建立只是基于已有的适用于二分类变量预测的基础模型，本小组查阅了有关高维全基因组关联分析的文献。全基因组关联分析（GWAS）是在全基因组水平上分析高密度的 SNP 与性状相关性的分析，从而发现影响复杂性状的基因变异的一种统计方法。很多学者致力于改善筛选方法以及预测模型，目的是提高预测准确度和计算效率。为了解决如何最有效地利用现有的数据资源，戴明伟等 2017 年提出了一种统计方法，即 IGESS，它通过整合个体水平的基因型数据和汇总统计数据，提高识别风险变异的统计能力，并改善风险预测的准确性。于是，本小组通过调用戴明伟等开发的 IGESS 包，利用已有的 7 个 GWAS 的 summary

statistics（p 值）和 4081 个人的基因型数据进行变量筛选和模型预测。但是在数据分析过程中，由于电脑内存不足无法运行高维数据。因此，本小组思考或许可以通过以下方法对 248409 个 SNP 进行初步筛选，达到减少数据维数的目的：基于通过卡方检验计算得到的 p 值和 7 个汇总 p 值，通过控制 FDR 方法（fdr<0.1）筛选出 317 个 SNP。在这之后，本小组调用 IGESS 包，利用 4081 个人的 317 个 SNP 进行进一步的变量筛选和模型预测。在调用过程中，该统计方法运行效率明显高于上述的集成学习模型(其运行一次时间约为 0.236s)且 AUC 较高（平均 AUC 为 0.696）。IGESS 方法预测效果如图 9。

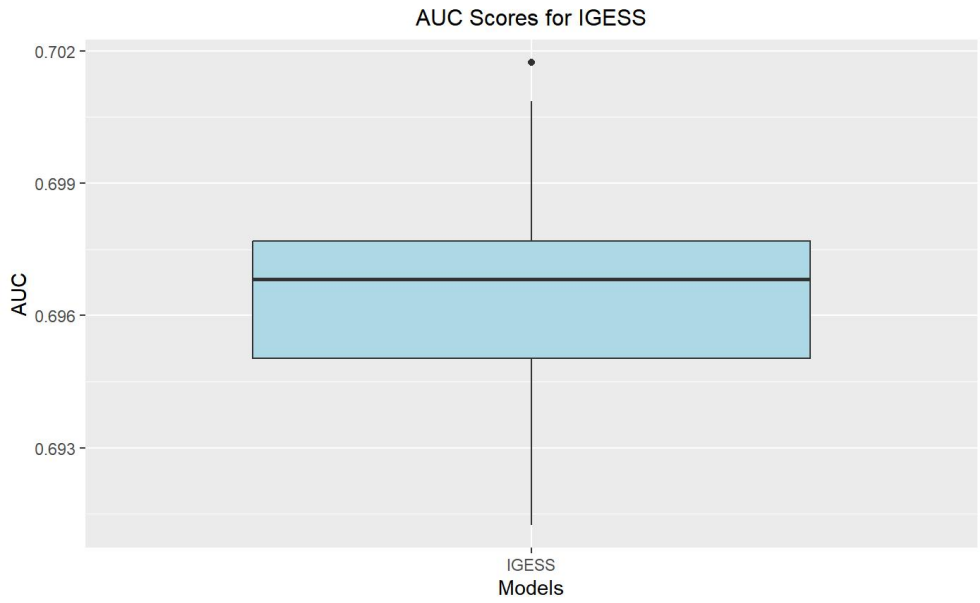


图 9：IGESS 预测性能

五、 评价与反思

通过这次大作业，本小组成员第一次面对大数据。在分析时，我们尝试了很多以前从未使用过的变量筛选的办法。在使用很多更高级的筛选方法时（如 lasso），本小组遇到了电脑内存不够的问题，这一问题并未得到解决。但同时也引发了我们的思考，究竟应该如何进行解决，或许可以在云平台上面代为运行程序，又或许可以将变量筛选方法混合使用，如：先用卡方检验进行初步筛选（将 FDR 控制在 0.1 水平），再对初步筛选后的变量进行 lasso 回归，达到压缩估计的目的。

利用集成学习建立模型时，本小组发现用网格搜索十分的耗时，所以我们尝试利用贝叶斯优化进行参数优化，由于是第一次使用，缺乏经验，导致得出的结

果无法保证最优性。随着学习的进一步深入，本小组会不断精进自己构建模型的能力，丰富自己的经验。

通过拓展思考，本小组了解了许多经常使用的模型以外的方法，我们也尝试将这些方法运用到此次作业中，有成功有失败，但是收获颇丰。

参考文献

[1] Dai Mingwei,Ming Jingsi,Cai Mingxuan,Liu Jin,Yang Can,Wan Xiang,Xu Zongben. IGESS: a statistical approach to integrating individual-level genotype data and summary statistics in genome-wide association studies.[J]. Bioinformatics (Oxford, England),2017,33(18).

[2] 杨文宇,吴成秀,肖英杰,严建兵.基于Adaptive Lasso的两阶段全基因组关联分析方法[J/OL].作物学报:1-11[2023-05-04].http://kns.cnki.net/kcms/detail/11.1809.S.20230302.1544.007.html