第三问

1、综述

针对该问，我们选择将五个分类预测模型分别建立，针对不同的指标，选择最优的分类预测模型，下面是每个指标的选择过程。

首先，对数据进行处理，由于部分方法对于数据是否标准化十分敏感，因此我们对所有数据都进行了标准化，但是，标准化有时反而不会改进模型的拟合效果，因此我们对未标准化、标准化的数据都进行了模型拟合，不过囿于篇幅所限，我们会直接给出较优结果。所以下文中我们在给出模型效果时，如果未加说明，则说明使用的是未标准化的数据；标准化的数据我们在介绍模型效果时即会特殊说明。

针对每个指标，我们都采用8种模型对数据进行拟合，之后通过比较模型的优劣来对每个指标选择适合的模型。对于每个拟合模型，我们将ADMET.xlsx中的1480个化合物当作训练集，494个化合物当作验证集，以此进行模型拟合以及效果的评价。同时，我们认为如果模型在训练后，在验证集上的准确率越高，说明该模型拟合效果越好；此外，我们还考虑了auc曲线，该曲线下方面积衡量了模型对真阳率与伪阳率的辨别效果，auc值越大说明辨别效果越好，所以下文中我们同时比较两个值来说明模型的优劣。

最终，想要预测5个ADMET性质值，对每个都建立8个模型拟合后，最终采用了1个SVM所建立模型，1个Adaboost算法所建立模型，1个随机森林所建立模型，2个LightGBM算法所建立模型。而由于随机森林在上文中已经介绍过，这里就不再赘述，主要说明剩下的7个算法。

2、方法介绍

1. 逻辑回归方法：

该方法主要用于二分类问题，其核心思想在于寻找决策函数，只要将变量值带入决策函数后，超过了决策边界，我们就认为其属于一个类别；反之，就认为它属于另一个类别。而逻辑回归与回归的区别就在于需要寻找阶跃函数，使得决策函数的连续输出值变为不连续的类别分类值，这样给定分子描述式，我们就可以给出该化合物相应的ADMET性质。

二、临近算法：

KNN是一个非常直观的分类算法，核心思想就是计算新样本点到过去样本的距离，然后找出K个距离最近的，我们认为新样本点与这K个点最相似，所以只要找出这K个点中哪个类别占比最大，我们就认为新样本点即为该类别。

三、决策树方法：

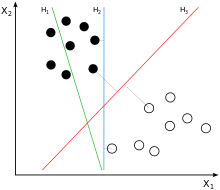
决策树也是分类方法中较为常见的一种，其核心思想就是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，并以此为根据进行分类。这种方法利用信息学理论中熵的概念，将对象属性与对象值之间构建一种映射关系。具体在本题中，可以将训练集中各个化合物的分子描述式看作对象属性，将已知的ADMET性质值看作对象值，那么就可以学习到一个分类器，通过该分类器就可以对新样本点进行预测。

四、支持向量机方法：

支持向量机是一种二分类模型，它将实例的特征向量映射为空间中的一些点，SVM 的目的就是想要画出一条线，以 “最好地” 区分这两类点，以至如果以后有了新的点，这条线也能做出很好的分类。SVM 适合中小型数据样本、非线性、高维的分类问题。

将实例的特征向量（以二维为例）映射为空间中的一些点，如下图的实心点和空心点，它们属于不同的两类。SVM 的目的就是想要画出一条线，以“最好地”区分这两类点，以至如果以后有了新的点，这条线也能做出很好的分类。

而在实际操作的时候，我们会发现给定超平面后，每一个类别，我们都能找到一个支持向量，支持向量上的点到超平面的距离都是一样的，所以我们只需考虑支持向量即可，这也是该方法名称的由来。



支持向量机方法有一些特点：

1、虽然可以找出无数条线将样本点进行区分，具体区别就是效果好不好，类似上图，蓝色和绿色线的分类效果就不是很好，而红线就是一个分类效果比较好的超平面。

2、支持向量机中，每做一个超平面，其两侧数据点都距离它有一个最小距离（垂直距离），这两个最小距离的和就是间隔。而我们的目的就是让间隔最大，这样我们分类的犯错概率才会更低。

3、训练好的SVM模型的算法复杂度是由支持向量的个数决定的，而不是由数据的维度决定的，所以SVM算法不太容易产生过拟合现象。

4、SVM训练出来的模型完全依赖于支持向量，所以即使训练集中所有非支持向量的点都被去除，重复训练过程，我们同样会得到完全一样的模型。

在我们实现的时候，我们直接调用了sklearn库来拟合SVM模型，但是其中还涉及核函数的选择问题，下面简单说明一下核函数。

至于为什么要使用核函数，因为在线性SVM中，转化为最优化问题时求解的公式计算都是以内积形式出现的，而内积的算法复杂度非常大，所以需要使用核函数来取代计算非线性映射函数的内积。

我们所考虑的，除了线性内核之外，就是高斯径向核函数以及S型核函数。

五、XGboost方法：

XGboost方法是一个基于决策树的分类算法，所以首先就要明确怎样的分类才算是合理的分类，什么才是关于分类的量化指标。所以在介绍该方法之前，首先我想说明关于量化分类效果的度量标准：信息增益

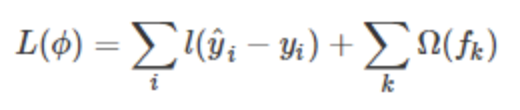
信息增益的核心思想就是以信息增益度量属性选择，选择分裂后信息增益最大的属性进行分裂，那什么是信息增益呢？为了准确定义信息增益，我们先定义信息论中广泛使用的一个度量标准，就是熵，它刻画了任意样本集的纯度。通过熵，我们就能评估一个决策树的分类效果好坏了。

而在单个决策树基础上，我们就可以定义集成学习了。通过构建多个决策树对数据集进行预测，然后用某种策略将多个决策树预测的结果集成起来，作为最终预测结果，就是集成学习。而根据集成学习的各个决策树之间有无依赖关系，分为Boosting和Bagging两大类，Boosting就是各分类器之间有依赖关系。XGboost就属于Boosting流派，而想要详细说明，还需要介绍一下GBDT，因为XGboost本质上还是GBDT，只是把速度效率发挥到了极致。

GBDT的原理很简单，就是把所有决策树的结果相加等于预测值，然后下一个决策树去拟合误差函数对预测值的残差，这样吧所有决策树的结论结合，就能得到最终结果。其中，GBDT用负梯度来近似残差，然后把均方差损失函数作为损失函数。

而到了XGboost算法，其核心思想在于不断添加决策树，不断的进行特征分裂来生长一棵树，每次添加一个树后，再拟合上次预测的残差。然后当我们训练完成后，只要给一个样本点，我们就可以根据这个样本的特征，在每棵树中找到一个对应的叶子节点，然后根据叶子节点的分数，得到该决策树对样本点的预测分数，最后将所有决策树的对应分数加起来就是该样本预测值。

显然，我们的目标是要让所有树的预测值尽量接近真实值，同时尽可能拥有较大的泛化能力，所以从数学角度看这是一个泛函最优化问题，那么就可以把目标函数简化如下：



这个目标函数就分为两部分：损失函数和正则化项，其中损失函数用来刻画预测分数与真实分数的差距，正则化定义复杂度。类似之前提到的GBDT的套路，XGboost也是需要将多棵树的得分累加得到最终的预测得分，那么每轮加入的那棵树就十分关键。这颗树就是要让我们的目标函数尽量降低，因此需要考虑损失函数，而XGboost的优势就是它利用泰勒展开，将展开的各项对应到目标函数中，这样，我们就把损失函数部分解决了。

关于正则化项，我们得说明我们是利用叶子节点集合以及叶子节点得分来表示的，每个样本都得落在一个叶子节点上。因此，XGboost对树的复杂度要求就包含了两部分：一个树里面的叶子节点个数和一个树上叶子节点得分的模平方，这样就把正则化部分解决了，至此XGboost算法介绍完成。

六、Adaboost算法：

Adaboost算法是boosting系列算法中最著名的算法之一，既可以用作分类，也可以用作回归。Boosting算法的工作机制是首先从训练集用初始权重训练出一个弱学习器1，根据弱学习的学习误差率表现来更新训练样本的权重，使得之前弱学习器1学习误差率高的训练样本点的权重变高，使得这些误差率高的点在后面的弱学习器2中得到更多的重视。然后基于调整权重后的训练集来训练弱学习器2.，如此重复进行，直到弱学习器数达到事先指定的数目T，最终将这T个弱学习器通过集合策略进行整合，得到最终的强学习器。其中Adaboost算法的主要特点在于其弱学习器权重系数的计算、更新样本权重和集合策略，尤其是Adaboost分类采用加权平均法的集合策略，使得其分类效果较好。

七、LightGBM算法：

LightGBM是个快速的，分布式的，高性能的基于决策树算法的梯度提升框架。可用于排序，分类，回归以及很多其他的机器学习任务中。虽然XGboost算法十分优异，但是在实际应用时速度太慢，而LightGBM在不降低准确率的前提下，速度提升了10倍左右，占用内存下降了3倍左右。因为它是基于决策树算法的，它采用最优的叶明智策略分裂叶子节点，然而其它的提升算法分裂树一般采用的是深度方向或者水平明智而不是叶明智的。因此，在LightGBM算法中，当增长到相同的叶子节点，叶明智算法比水平明智算法减少更多的损失。因此导致更高的精度，而其他的任何已存在的提升算法都不能够达。实际应用时主要体现在两点：一是不使用所有的样本点来计算梯度，而是对样本进行采样来计算梯度；二是不是使用所有的特征来进行扫描获得最佳的切分点，而是将某些特征进行捆绑在一起来降低特征的维度，使寻找最佳切分点的消耗减少。

3.具体应用

下面是对ADMET性质的五个指标进行分类预测：

1. Caco-2指标

首先观察数据，所有化合物中，有1215个化合物在该指标取值为0，759个化合物在该指标取值为1。同时，根据我们对该指标的生理意义理解，Caco-2度量了化合物被人体吸收的能力，特别的，该指标值为1时代表该化合物的小肠上皮细胞渗透性较好，那此时该化合物更容易被人体吸收，所以我们认为指标值为1时属于具有更好的ADMET性质。

1. 逻辑回归方法

该方法在验证集上的准确率为87.44939271255061%，然后auc曲线下方的面积为0.947381457891012，效果较为理想。

1. 标准化支持向量机方法

在使用该方法时，一个可以调整的即为所使用的内核，所以我们针对三种不同内核：linear、rbf、sigmoid分别计算了在验证集上的准确率，auc取值。

|  |  |  |
| --- | --- | --- |
| 内核 | 验证集准确率 | auc取值 |
| linear | 0.8825910931174089 | 0.9493984430290163 |
| rbf | 0.9149797570850202 | 0.9752653927813164 |
| sigmoid | 0.8299595141700404 | 0.896302193913659 |

显然，根据表格中数据，当使用rbf内核时，支持向量机模型在验证集上准确率和auc取值都较高，所以如果后续选择该模型，那么就应该使用rbf内核。

同时，在使用SVM时，参数gamma也是影响模型拟合效果的一个重要指标，所以我们考虑如何是gamma取到较优值，所以我们将gamma从10的-10次方到1，按照等比数列分50份，进行遍历，并通过svc.score函数来判断效果，最终gamma=0.007196856730011514时效果最好。

1. 标准化KNN方法

该方法在验证集上准确率为0.8724696356275303, auc值为0.9478060863411182，效果不如rbf内核的支持向量机方法。但是，在使用该方法时，仍然有参数可以调整的范围，KNN方法中，K就是一个可以调整的指标。同样，我们选择遍历的方法选择K最优的取值，但是由于K取值只能为整数，所以我们考虑范围就设置为为2到12之间的所有整数了。最终K=2时效果最好。

1. 决策树方法

该模型在验证集上的准确率为0.8785425101214575，相应的auc取值为0.8677105449398443。

1. 随机森林方法

该模型在验证集上的准确率为0.8785425101214575，相应的auc取值为 0.970859872611465。

1. Adaboost算法

该模型在验证集上的准确率为0.9008097165991903，相应的auc取值为0.8970806794055203。

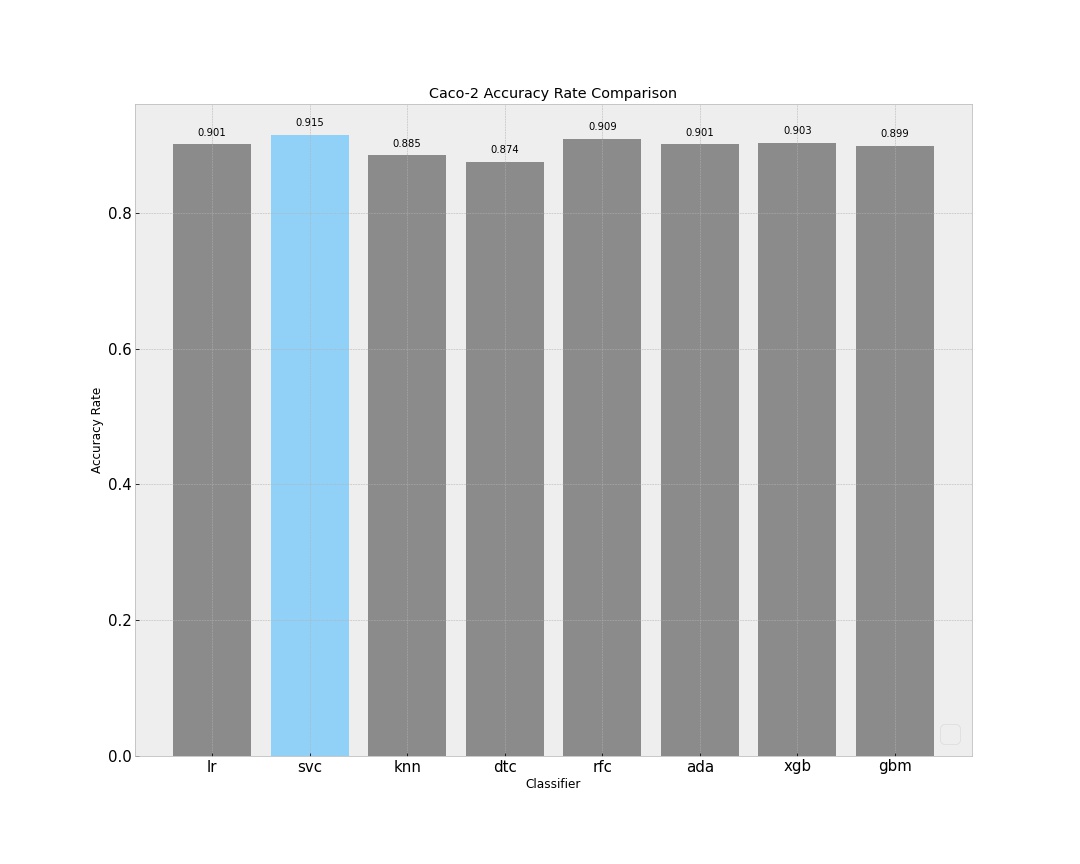
1. XGboost算法

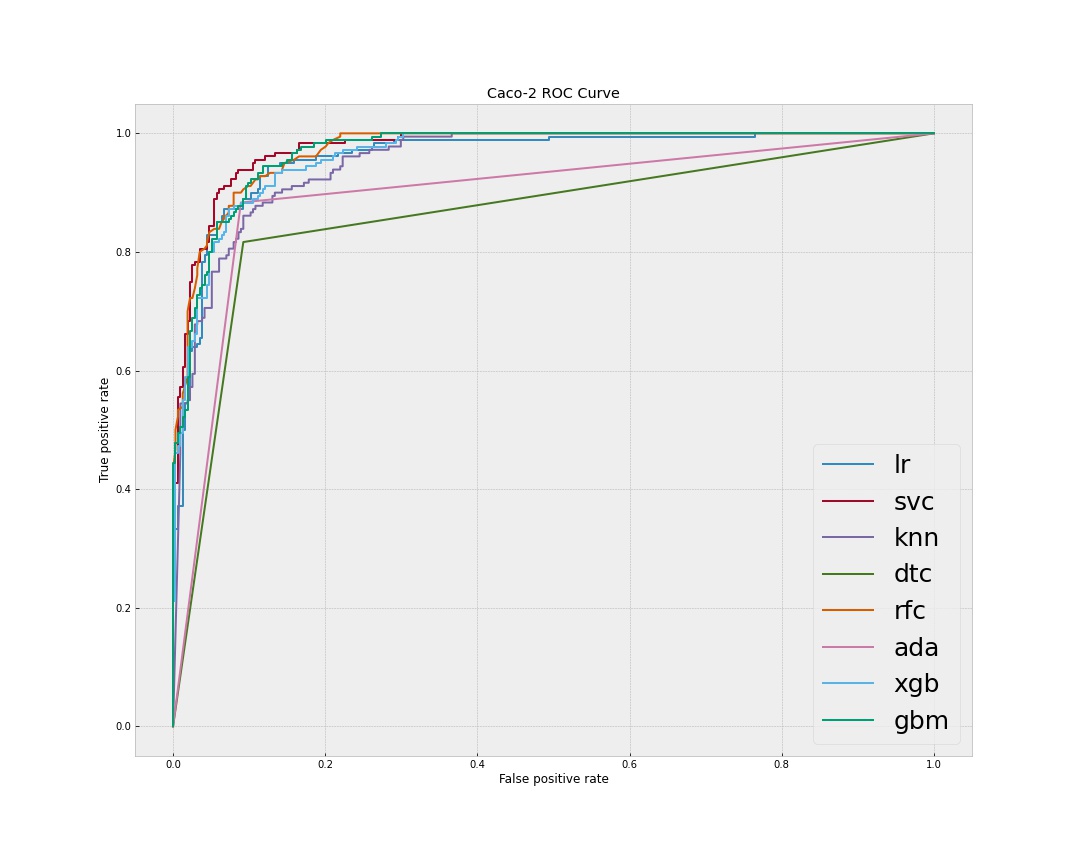
该模型相应的auc取值为0.96130。

8）lightGBM算法

该模型相应的auc取值为0.96518。

最终，针对该指标，我们选取的模型是支持向量机方法所建立模型，对应最高的验证集准确率为0.9149797570850202。



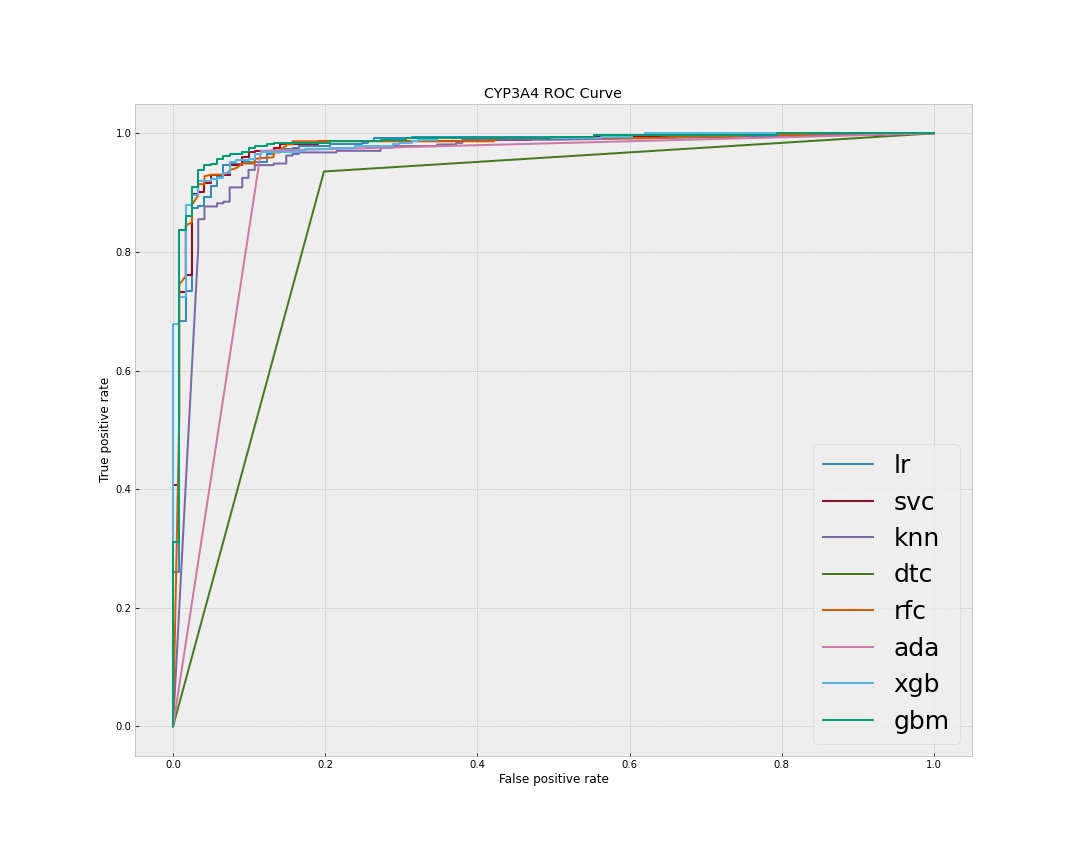
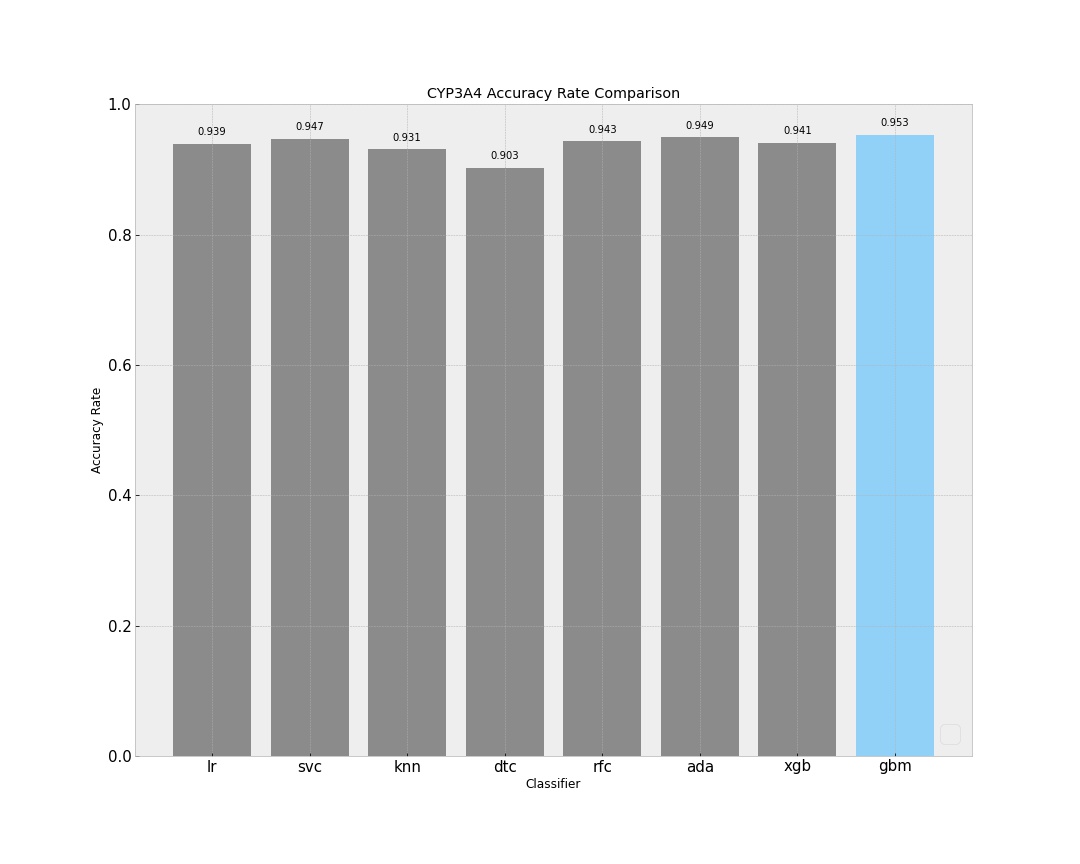


对于之后的指标，模型构建方法类似，所以我们直接展示结果

1. CYP3A4指标

CYP3A4：‘1’代表该化合物能够被CYP3A4代谢，‘0’代表该化合物不能被CYP3A4代谢，而指标值为1就说明该化合物进入人体后会被代谢，而被代谢一般来说即说明其结构发生变化，进而其药理特性一般也会发生变化，最终导致药物失效。所以该指标值为1，我们认为是该ADMET性质不好，指标值为0时，我们认为其有较好的ADMET性质。

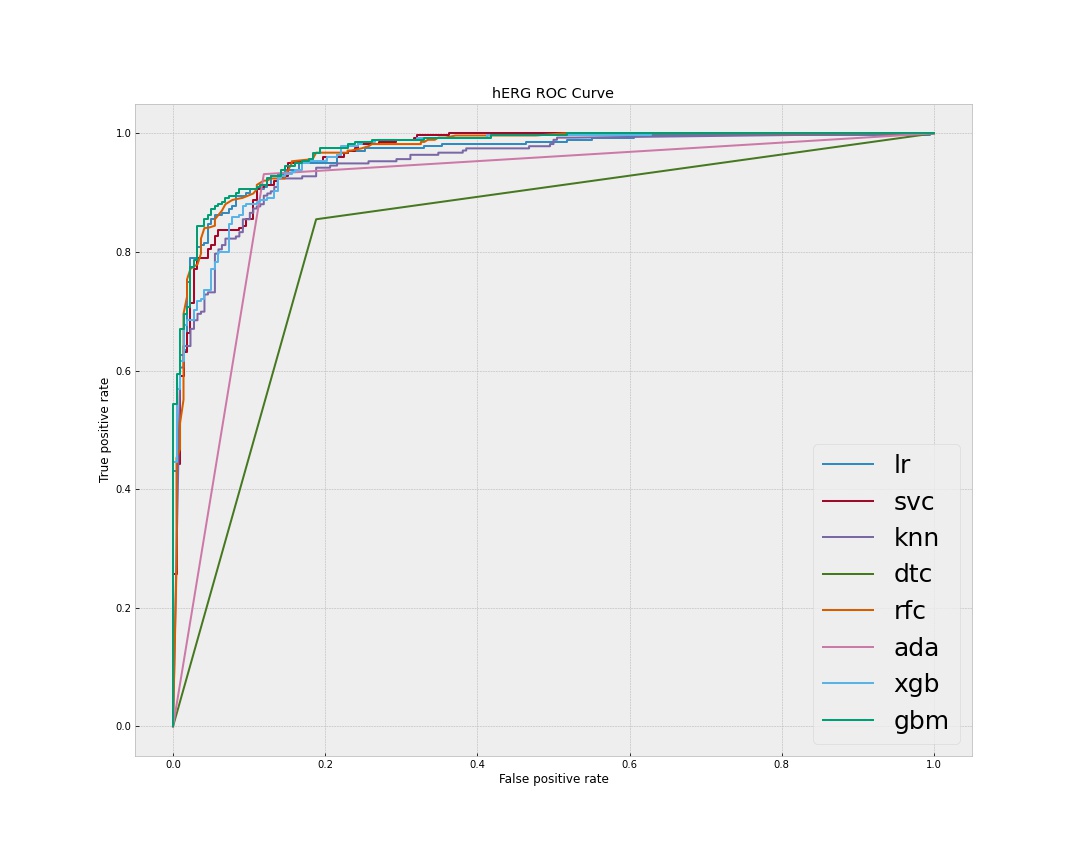
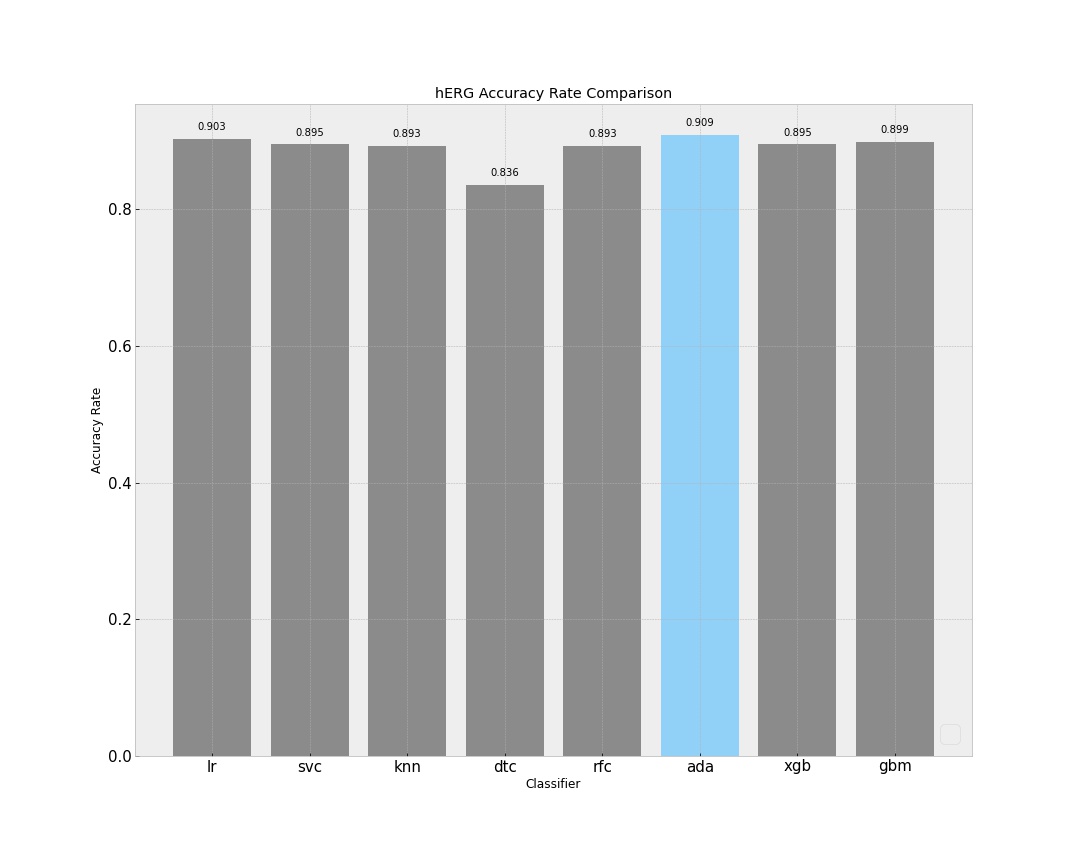
最终确定模型为LightGBM算法所建立模型，验证集准确率为0.9534412955465587。



1. hERG指标

hERG：‘1’代表该化合物具有心脏毒性，‘0’代表该化合物不具有心脏毒性，显然，具有心脏毒性的化合物是应该减少使用的，所以该指标为1时，我们认为其ADMET性质不好；指标值为0时我们认为其拥有更好的ADMET性质。

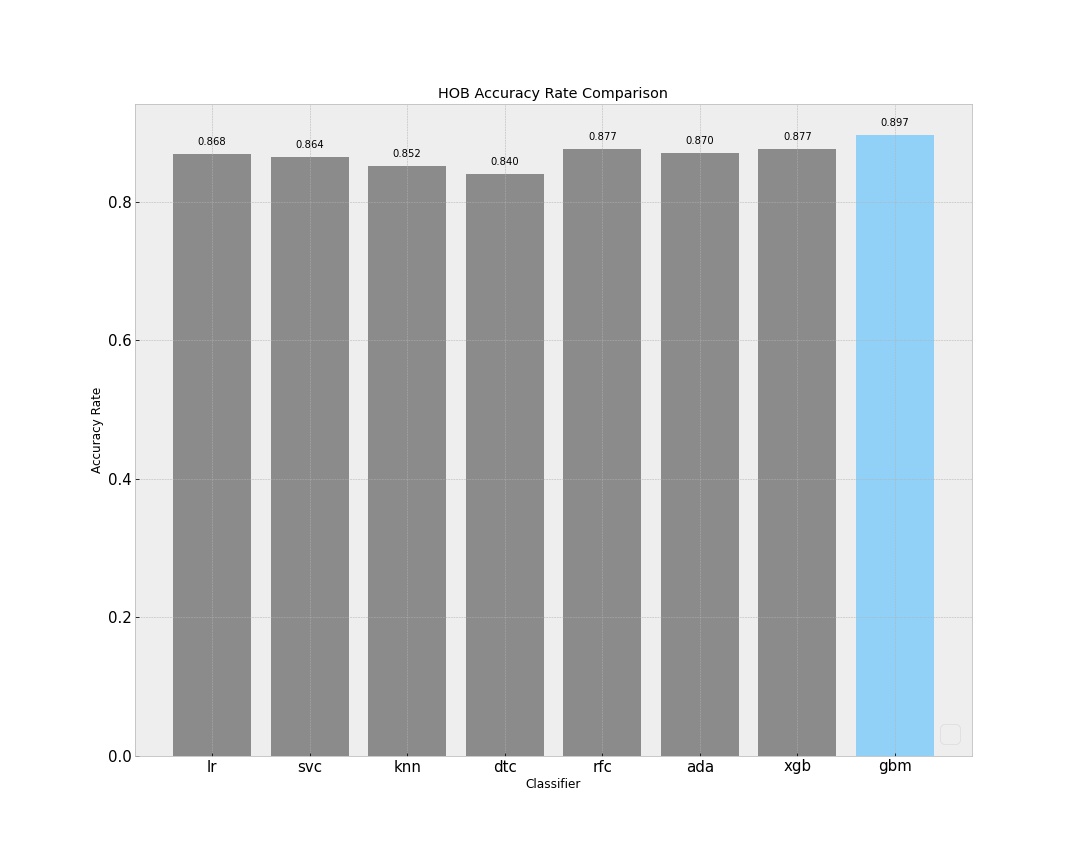
最终确定模型为Adaboost方法所建立模型，验证集准确率为0.9089068825910931。

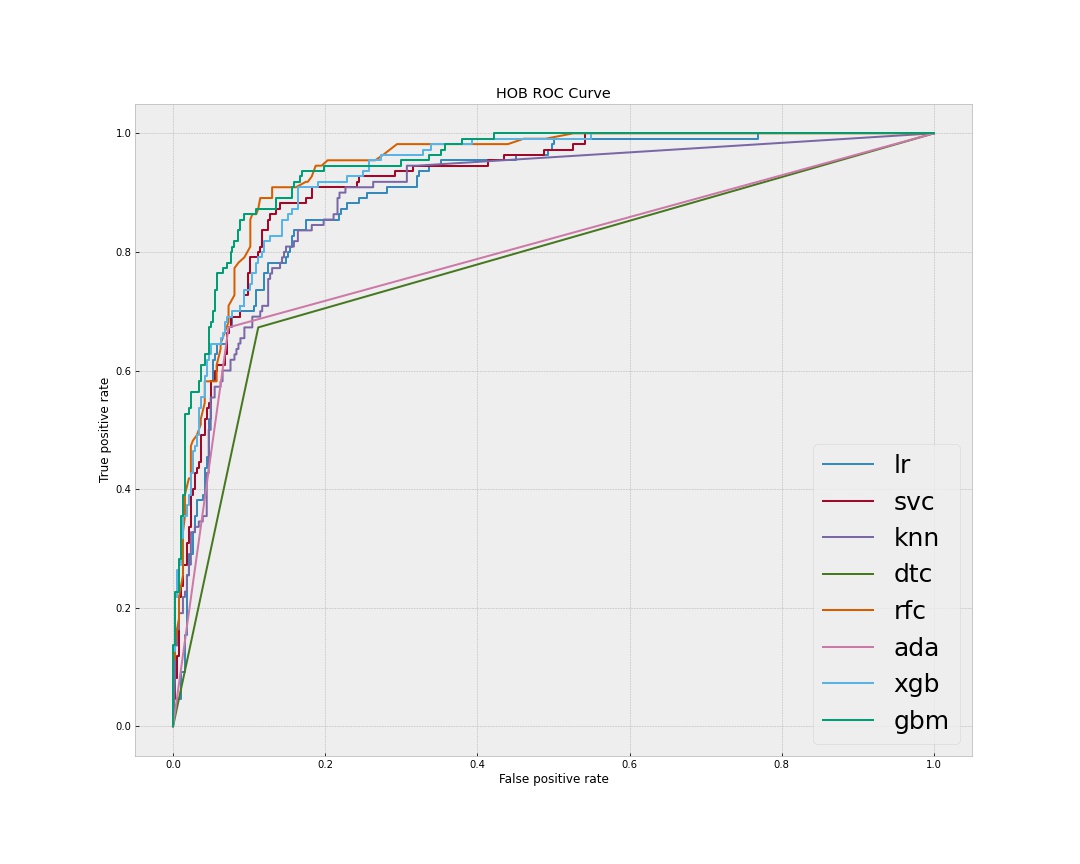


1. HOB指标

HOB：‘1’代表该化合物的口服生物利用度较好，‘0’代表该化合物的口服生物利用度较差，显然，口服生物利用度较好的化合物更容易被病人使用，所以，该指标值为1时，我们认为其ADMET性质更好。

最终确定模型为LightGBM所建立模型，对应的验证集准确率为0.8967611336032388。





1. MN指标

MN：‘1’代表该化合物具有遗传毒性，‘0’代表该化合物不具有遗传毒性。显然，具有遗传毒性的药物病人不希望去使用，我们认为其药理特性不好，所以指标值为1时，我们认为其ADMET性质不好；指标值为0时，我们认为其ADMET性质较好。

针对该指标，我们首先观察数据，一共1974个化合物，1514个化合物取1值，460个指标取0值，我们认为取0值的样本太少，存在样本不平衡问题，所以考虑增加样本。同时，由于样本本身数目就不多，所以使用过采样技术是较为合理的，在这里我们使用合成少数类过采样技术（SMOTE）。

SMOTE算法是基于随机过采样算法的一种改进方案，由于随机过采样采取简单复制样本的策略来增加少数类样本，这样容易产生模型过拟合的问题，即使得模型学习到的信息过于特别(Specific)而不够泛化(General)，SMOTE算法的基本思想是对少数类样本进行分析并根据少数类样本人工合成新样本添加到数据集中，具体实现

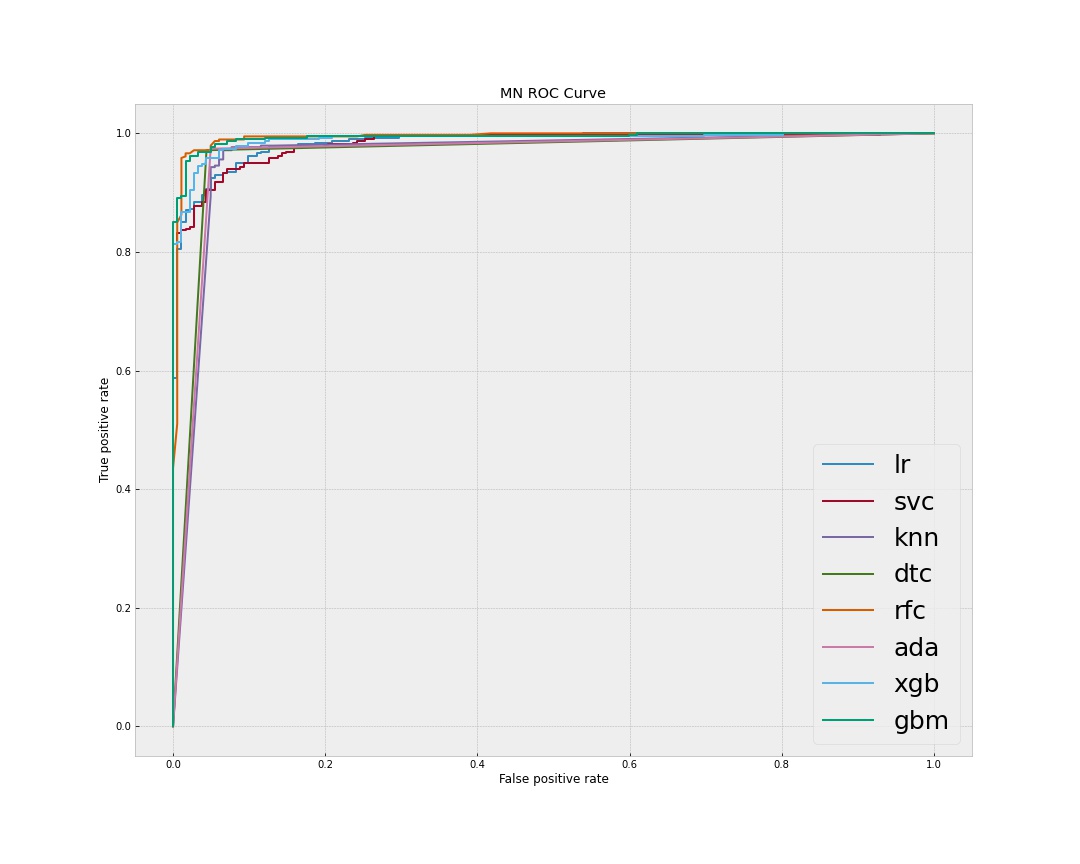
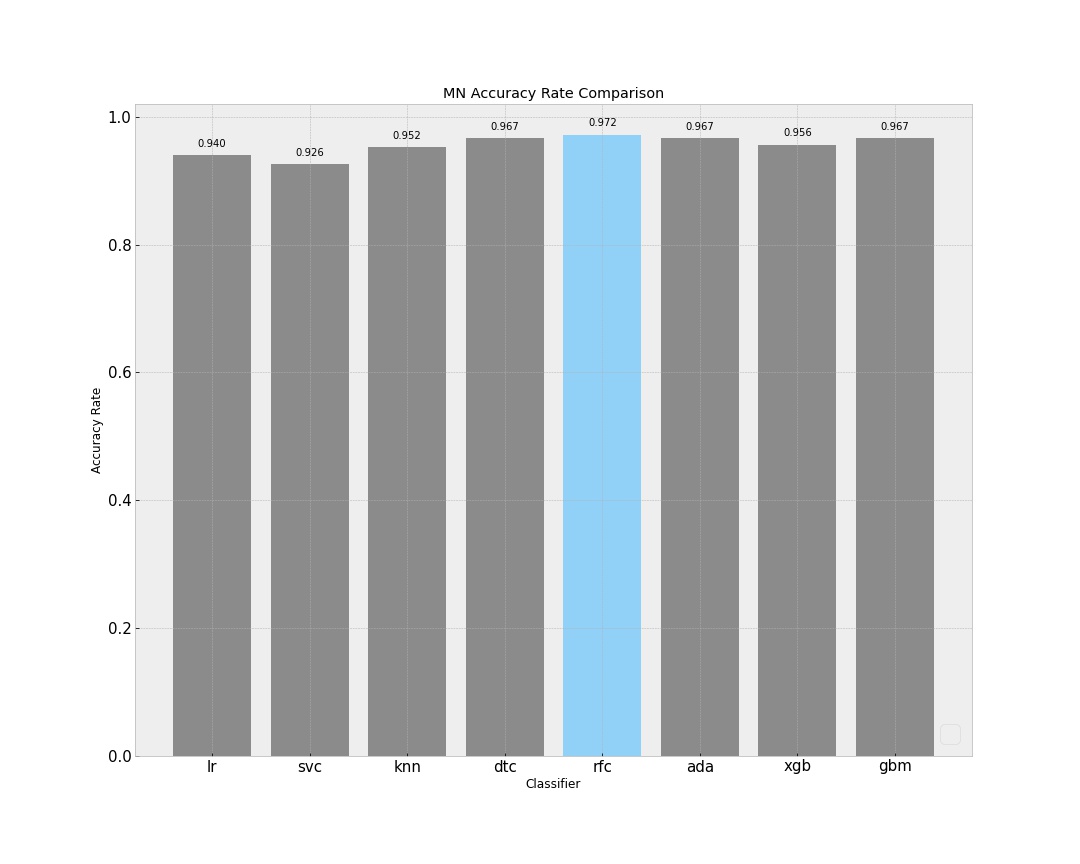
1)对于少数类中每一个样本x，以欧氏距离为标准计算它到少数类样本集中所有样本的距离，得到其k近邻。

(2)根据样本不平衡比例设置一个采样比例以确定采样倍率N，对于每一个少数类样本x，从其k近邻中随机选择若干个样本，假设选择的近邻为xn。

(3)对于每一个随机选出的近邻xn，分别与原样本按照如下的公式构建新的样本。如下：

IMG_256

对增加完样本后的数据，使用前文中提到的8个模型进行模拟，最终结果同样是采用随机森林所建立模型预测效果最好，验证集准确率为0.971830985915493。



以上就是我们对想要预测这五个ADMET性质所操作的具体流程，我们也是根据这五个拟合模型，对预测集—“ADMET.xlsx”的test表进行预测的，具体的结果已经补充在test表了。

第四问

针对该问，我们要同时满足两个条件：1、要保证优良的ADMET性质，对此，我们考虑用数学公式将ADMET性质的五个指标值进行限制。在之前的分析中，我们可以知道，CYP3A4、hERG和MN指标这三个指标为零时，说明相应的药理性质较差。我们将五个ADMET性质，相应的定义为自变量X1到X5，那么此时，至少三个性质较好就等价于X2+X3+X5<=X1+X4，具体实现时也是这样限制的。2、要尽可能搜索化合物的分子描述符的范围，使得相应的化合物生物活性尽可能高。

对此，我们所考虑的方法是在局部最优解的范围进行搜索，寻找相应的分子描述符范围，即：我们将这1974个化合物的生物活性从高到低进行排序，之后对每个化合物，验证其是否满足至少三个ADMET性质良好这个条件，不行的话就寻找下一个，直到寻找到符合条件的化合物为止。这样，我们找到了既满足至少三个ADMET性质良好同时生物活性也较高的化合物，我们将其设定为初始值。之后，进行小范围搜索。

考虑到第一二问我们已经将749个分子描述式通过变量选择减少到了20个，所以此时我们也仅考虑这20个选择后的分子描述式的应如何取值。下面介绍我们的搜索方法：

我们将这20个分子描述式的初始值都设为我们前面选择的化合物相应的分子描述式的值，然后我们将其中19个固定，对剩下的那一个进行搜索。我们以该分子描述式值的1/10作为步长分别向上向下进行搜索，每走一个步长，我们就利用第三问建立的5个模型，对新的5个ADMET性质值进行估计，然后判断其是否满足至少三个ADMET性质较好。如果条件满足，那么此时再利用第二问建立的生物活性回归模型，计算相应的生物活性，如果该生物活性不低于初始条件下的生物活性，我们就认为该分子描述式的取值是我们想要的，就纳入取值范围。类似的，每次固定19个分子描述式值，对剩下的那个进行搜索，我们就可以得出剩下的分子描述式的取值范围，最终就可以得到这20个分子描述式的取值范围。

