问题：

1. 从建模目标、建模算法和评估方面，阐述机器学习建模与传统统计模型有哪些不同？

答：

①建模目标上的不同，对于机器学习模型的目标是通过大数据量的复杂特征来预测未知数据的最优结果（黑盒优化），对于传统统计模型的目标是使得小样本满足统计检验假设来验证变量间因果关系（白盒解释）。

②建模算法上的不同，对于机器学习多采用增加模型的容量逼近数据分布，而统计模型则通过数学约束保证可解性。比如机器学习通常采用神经网络、深度学习、迭代优化等复杂度高的大模型进行数据预测；而传统统计模型使用线性模型和数学的统计检验等进行数据结果的解释。

③评估模型性能上，机器学习模型主要通过测试集的F1-score或者AUC-ROC的值来评估性能的优劣。机器学习通过K折交叉验证来防止过拟合。传统统计模型主要通过P值置信区间或者模型似然值来评估统计结果的可靠性。传统统计模型可以通过残差分析和假设检验对拟合结果进行可靠度的评估。

2. 生态学上常用树模型，包括随机森林、提升回归树。建模步骤包括：数据预处理、拆分数据集、选择特征、算法和训练模型、模型评估等。caret为各种机器学习算法提供了统一模板，加载mtcars数据集，请根据问题填空。

1）对于doubs中的鱼群数据，按照样地，计算各样地鱼类Shannon多样性指数，并新增##列。

答案：

###按照样地计算doubs中鱼群数据的鱼类Shannon多样性指数

library(ade4)

data(doubs)

library(vegan)

#检查otu表格的列和

otu <- doubs$fish

colSums(otu)

#抽平

otu\_flat <-as.data.frame(t(rrarefy(t(otu), min(colSums(otu)))))

#计算香农指数

#以e作为底数表示方法

otu\_flat$mpg <- diversity(otu\_flat, index = 'shannon', base = exp(1))

2）利用train()，训练随机森林（randomForest）模型

model\_rf <- train(mpg ~ ., data = training\_data, method = " rf \_")

3）通过trainControl()，向train()添加重采样10-fold cross-validation，以优化参数

fitControl <- trainControl(method = " repeatedcv ", number = 10, repeats = 5)

model\_rf <- train(mpg ~ ., data = training\_data, method = " rf ", trControl =fitControl)

3）在train()中，增加中心化和标准化等数据预处理，提高模型精度

model\_rf <- train(mpg ~ ., data = training\_data, method = " rf ",

preProcess = c('scale', 'center'),

trControl =fitControl)

4）rf有mtry和tree两个参数，可以通过expand.grid()设置调优，并在train()添加

grid <- expand.grid(.mtry=c(1:10))

model\_rf <- train(mpg ~ ., data = training\_data, method = " rf ",

preProcess = c('scale', 'center'),

trControl =fitControl,

\_\_ tuneLength \_\_\_ = \_5\_\_)

3. 什么是递归消除选择？在caret包中，为何选择随机森林等树模型时，没有特征选择这个过程？

答：

（1）递归特征消除算法，是一种机器学习算法。其主要算法原理是基于模型的特征选择方法，通过反复训练模型和剔除不重要特征的方式来选择最优的特征子集。大概的步骤是：①用训练数据训练一个初始模型，并计算每个特征的重要得分；②根据特征重要性得分进行排序，从中选择得分最低的若干个特征剔除；③在剩余的特征上重新训练模型并计算特征的重要性得分；④特征数达到预设目标或者全部剔除则停止算法，否则回到第二步；⑤选择剩余的特征作为最终的特征子集。

（2）因为随机森林树模型具有动态特征选择的原理，在构建树的过程即实现了特征的随机选择。而且冗余的特征对于树模型的影响不大。