**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？**

**解答：**

（1）统计建模是通过假设一个合适的数据模型，然后根据数据估计模型参数。相反，机器学习方法不是从模型开始，而是使用一种算法来学习响应与预测变量之间的关系。机器学习方法是假设生成过程是复杂和未知的，试图通过观察输人和响应发现主导模式。

（2）传统统计方法选择了一个模型，没有检验任何其他假设。机器学习方法首先给出的是模型集合，包括多项式模型、指数模型等，目标是从不同模型选择最佳的一个。当然，最终的结果与回归算法结果可能是一致的。

（3）统计模型重点是描述数据与结果变量之间的关系，评估模型的合理性是通过置信区间、显著性检验和其他检验对回归参数进行分析，即统计推断。机器学习涉及训练集和测试集，模型优劣通过测试集评估。

（4）统计建模更多关注变量之间的关系和意义，是可解释的。然而，机器学习强调预测性能，而不在于模型是否具有可解释性。

**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doubs数据集，请根据问题填空。**

**解答：**

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mpg列。

library(vegan) # 加载函数包，用于生态指数计算

library(ade4) # 加载doubs数据所在包

data("doubs")

shannon\_index <- diversity(doubs$fish, index = "shannon") # 计算各样地鱼类Shannon多样性指数

doubs$mpg <- shannon\_index # 把Shannon指数添加为新列 mpg

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)

4）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)

5）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,

\_\_\_\_tuneGrid \_\_\_\_ = \_\_\_grid\_\_\_)

**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？**

**解答：**

递归特征消除(Becursive Feature Elimination，RFE)是指使用一个基模型来进行多轮训练,首先用所有特征训练模型得到每个特征权重,并剔除拥有最小权重特征,之后再基于其余特征训练,重复前面的过程,如此往复递归,直至剩余的特征数量达到所需的特征数量。

没有特征选择这个过程的原因：

（1）内在特征选择：树模型（比如随机森林）本身具有内置的特征选择机制。在构建每棵树的过程中，模型会选择最佳的特征来进行分裂，这意味着更重要的特征通常会出现在树的顶部，而不太重要的特征会逐渐被淘汰。这种内在的特征选择机制使得在使用树模型时很少需要手动进行特征选择。

（2）随机性：随机森林本身是通过引入随机性来提高模型的泛化能力，并减少过拟合。通过随机选择特征来构建每棵树，随机森林可以有效地处理高维数据集，并且不容易受到特征间的共线性影响。

（3）集成学习的优势：随机森林是一种集成学习方法，通过整合多棵决策树的结果来提高模型的性能。在整个模型中，每棵树都会选择自己认为最重要的特征来做决策，通过集成这些决策，模型可以更好地泛化到新数据集上，而不需要额外的特征选择步骤。

（4）模型调优：在使用caret包的train函数时，可以通过tuneGrid参数来指定模型调优的参数网格，其中包括了树模型的mtry参数（每次分裂时考虑的特征数量）。这个过程可以帮助找到最佳的模型参数，而不一定需要预先的特征选择。