**Homework-05**

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？

（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，试图通过观察输入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）统计模型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biodiv列。

install.packages("vegan")

library("vegan")

library("dplyr")

doubs$biodiv <- diversity(doubs$fish, index = "shannon", MARGIN = 1)

print(doubs)

2）利用train()，训练随机森林（randomForest）模型#如果要分割训练集，需要使用createDatapartition()

model\_rf <- train(biodiv ~ ., 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 = " cv ", number = 10, repeats = 5)

model\_rf <- train(biodiv ~ ., 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(biodiv ~ ., 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(biodiv ~ ., 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？

递归特征消除（RFE）是一种通过递归方式逐步移除不重要特征以筛选出对模型性能贡献最大的特征子集的方法。它通常在模型训练之前或之后独立进行，适用于特征数量较多且可能存在冗余的情况。在caret包中，随机森林等树模型没有直接嵌入特征选择过程，主要是因为随机森林本身具有内置的特征重要性评估机制，能够有效处理大量特征且对特征数量的敏感性较低。此外，caret包的设计理念是将特征选择作为独立模块提供，用户可以根据需要灵活选择是否进行特征选择以及选择哪种方法。这种模块化设计不仅提高了灵活性，还使得特征选择方法（如RFE）可以独立于模型训练过程，适用于多种模型，进一步增强了caret包的通用性和适应性。

要求：

一周内上传代码至自己GitHub托管的homework中。