**Homework-09**

问题：

在doubs河流鱼群研究中，请筛选VOLPla站点有关CHE鱼类生物量和密度的记录，构建表3。其中，第1列为序号，第2列为时间戳（stamp），第3列为生物量（Biomass）。请根据下面问题作答。

|  |  |  |
| --- | --- | --- |
| no | Date | Biomass |
| 1854 | 1994-06-21 | 0.9500713 |
| 1855 | 1995-06-13 | 0.9519654 |
| 1856 | 1996-06-18 | 8.3632565 |
| ︙ | ︙ | ︙ |

* library(tidyverse) 表3 doubs河1994-2020年CHE鱼类个体生物量
* library(randomForest)
* data <- read.table("fishBiomassData.txt ")
* mydata <- data |>
* subset(STATION=="VOLPla" & SP == "CHE") |>
* select(DATE, BIOMASS)
* ts <- ts(mydata$BIOMASS, start=1994, frequency=1)
* ts\_org <- window(ts, end = 2018)
* ts\_trf <- ts.org |> log() |> diff(1)
* lag\_order <- 2
* horizon <- 2
* ts\_mbd <- embed(ts.trd, lag.order + 1)
* Y\_train <- ts\_mbd[, 1]
* X\_train <- ts\_mbd[, -1]
* y\_test <- window(ts, start = 2019, end = 2020)
* x\_test <- ts\_mbd[nrow(ts\_mbd), c(1:lag\_order)]
* pred\_rf <- numeric(horizon)
* for (i in 1:horizon){set.seed(1)
* fit\_rf <- randomForest(X\_train, y\_train)
* pred\_rf[i] <- predict(fit\_rf, X\_test)
* y\_train <- y\_train[-1]
* X\_train <- X\_train[-nrow(X\_train), ] }
* pred\_rf
* exp\_term <- exp(cumsum(pred\_rf))
* last\_obs <- as.vector(tail(ts\_org, 1))
* backtrans\_fc <- last\_obs \* exp\_term
* y\_pred <- ts(backtrans\_fc, start = 2019, frequency = 1)
* library(fpp2)
* forecast::accuracy(as.numeric(y\_pred), as.numeric(y\_test))
* ts\_fc <- cbind(ts,pred = c(rep(NA, length(ts\_org)), y\_pred))
* plot\_fc <- ts\_fc |> autoplot() + theme\_minimal()
* plot\_fc

1) 这是构建机器学习时间序列预测模型，根据代码，请简述时间序列预测建模基本原理以及基本过程。

答：基本原理：时间预测实际上就是利用回归模型，基于历史数据来预测未来（时间序列预测模型是一种用于预测未来数据点的统计模型，基于过去的观测数据。其基本原理是假设未来的数据点与过去的数据点之间存在某种关系，即时间上的相关性）

基本过程：

* library(tidyverse) #加载tidyverse包
* library(randomForest)#加载随机森林包
* data <- read.table("fishBiomassData.txt ") #读取fishBiomassdata文件
* mydata <- data |>
* subset(STATION=="VOLPla" & SP == "CHE") |>#筛选出VOLPla站点和CHE鱼类的数据
* select(DATE, BIOMASS)#选择日期和生物量列
* ts <- ts(mydata$BIOMASS, start=1994, frequency=1)#创建时间序列对象，从1994年开始，频数为1，表明以年为单位
* ts\_org <- window(ts, end = 2018)#创建时间序列原始对象，结束于2018年
* ts\_trf <- ts.org |> log() |> diff(1)#对时间序列取对数并求一阶差分
* lag\_order <- 2#设置滞后阶数为2，即使用过去的第二个值进行预测
* horizon <- 2 #设置预测范围为2，即预测未来的第二个值
* ts\_mbd <- embed(ts.trf, lag.order + 1)#创建嵌入向量，在将转换后的数据作为机器学习的输入前，要把当前向量形式转化为矩阵，即嵌入k维序列，k=lag.order+1,这就是时间延迟嵌入
* Y\_train <- ts\_mbd[, 1] #提取训练集的因变量
* X\_train <- ts\_mbd[, -1] #提取训练集的自变量
* y\_test <- window(ts, start = 2019, end = 2020) #创建测试集，开始于2019，结束于2020
* x\_test <- ts\_mbd[nrow(ts\_mbd), c(1:lag\_order)]#创建测试集的自变量
* pred\_rf <- numeric(horizon)#初始化预测变量
* for (i in 1:horizon){set.seed(1) #设置随机种子
* fit\_rf <- randomForest(X\_train, y\_train) #使用随机森林模型进行预测
* pred\_rf[i] <- predict(fit\_rf, X\_test) #进行预测
* y\_train <- y\_train[-1] #更新训练集的因变量
* X\_train <- X\_train[-nrow(X\_train), ] }#更新训练集的自变量
* pred\_rf#输出结果
* exp\_term <- exp(cumsum(pred\_rf))#将预测变量转换为原始尺度
* last\_obs <- as.vector(tail(ts\_org, 1)) #获取时间序列最后一个观测值
* backtrans\_fc <- last\_obs \* exp\_term #进行反向转换
* y\_pred <- ts(backtrans\_fc, start = 2019, frequency = 1)#创建预测的时间序列对象
* library(fpp2)#加载fpp2包
* forecast::accuracy(as.numeric(y\_pred), as.numeric(y\_test))#计算预测准确性
* ts\_fc <- cbind(ts,pred = c(rep(NA, length(ts\_org)), y\_pred)) #合并原始时间序列和预测时间序列
* plot\_fc <- ts\_fc |> autoplot() + theme\_minimal() #绘制时间序列图
* plot\_fc3#输出结果

从上述代码可以看出基本过程：

**第一步，数据预处理和探索性分析：**

对于待分析历史时间序列数据，进行缺失值补齐和异常值检测；

检查时间序列的季节性、变化趋势等，对时间序列进行必要的相关性分析，比如自相关（ACF）和偏自相关（PACF），为建模寻找合适的特征，提高预测的准确性。

**第二步，构建滞后特征，模型训练及预测：**

选择合适的建模方法（传统统计模型、机器学习、深度学习等），此处选择随机森林；构建滞后特征，使用历史值预测未来值；然后在此基础上进行训练集和测试集的划分（必要时还要划分验证集，进行时间序列交叉验证），进行模型训练，得到初步预测结果。

**第三步，后处理，包括预测结果的逆向转换，模型评估和优化等：**

在模型训练时，使用的数据是经过平稳化操作的，因此预测值需要逆向转换才能还原为原始量纲；然后利用逆向转换的预测结果和经验值进行模型表现评估，必要时进行预测结果的校准，并进一步对时间序列预测模型进行优化

2）除了基于原始值（生物量）构造滞后特征外，还可从时间戳提取并构造建模特征，请简述从时间戳构造建模特征的依据是什么？

时间序列包含趋势，季节性以及周期性

时间趋势：时间序列数据通常具有趋势性，即数据点随时间的变化而变化。通过提取时间戳中的年份、月份、季节等信息，可以捕捉到这种趋势性。

周期性：许多时间序列数据具有周期性特征，例如季节性变化。通过提取时间戳中的周、月、季度等信息，可以捕捉到这种周期性。

根据书本内容还有以下依据：

（1）日期时间特征：即观察值与具体日期或季节有关，如日低温与雨季等有重要关系，可以选择把日期或季节作为特征

（2）滞后特征：如今年的11月份数据与以往年份的11月份数据更相关，即要关注历史上的今天

（3）窗口特征：建模精度不仅采用或选择的滑动窗口大小密切相关，而且与窗口内的均值和最大值等具体数据也有关

数据下载网址：

<https://figshare.com/articles/dataset/Data_for_Contemporary_loss_of_genetic_diversity_in_wild_fish_populations_reduces_biomass_stability_over_time_/13095380?file=39686458>

或者从GitHub下载[fishBiomassData.txt](https://github.com/flliu315/homework2025/blob/main/fishBiomassData.txt)：

<https://github.com/flliu315/homework2025>

要求：

一周内上传至各自GitHub托管的homework中。