**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：

数据准备：收集和整理时间序列数据。在本代码中，数据来自1994-2020年的CHE鱼类生物量记录。

数据预处理：对数据进行清洗和转换，以适应模型的需求。如处理缺失值、异常值等。

特征工程：从原始数据中提取有用的特征，以提高模型的预测能力。在本代码中，除了生物量外，还从时间戳中提取了滞后特征。

模型选择：选择合适的时间序列预测模型。在本代码中，使用了随机森林模型。

模型训练：使用训练数据集来训练模型。在本代码中，通过随机森林模型对训练数据进行拟合。

模型验证：使用验证数据集来评估模型的性能。在本代码中，通过预测验证数据并计算预测准确性来评估模型。

模型优化：根据验证结果对模型进行调整和优化，以提高预测准确性。

预测：使用优化后的模型对未来的数据点进行预测。

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。

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。