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 |
| ︙ | ︙ | ︙ |

1. library(tidyverse) 表3 doubs河1994-2020年CHE鱼类个体生物量
2. library(randomForest)
3. data <- read.table("fishBiomassData.txt ")
4. mydata <- data |>
5. subset(STATION=="VOLPla" & SP == "CHE") |>
6. select(DATE, BIOMASS)
7. ts <- ts(mydata$BIOMASS, start=1994, frequency=1)
8. ts\_org <- window(ts, end = 2018)
9. ts\_trf <- ts.org |> log() |> diff(1)
10. lag\_order <- 2
11. horizon <- 2
12. ts\_mbd <- embed(ts.trd, lag.order + 1)
13. Y\_train <- ts\_mbd[, 1]
14. X\_train <- ts\_mbd[, -1]
15. y\_test <- window(ts, start = 2019, end = 2020)
16. x\_test <- ts\_mbd[nrow(ts\_mbd), c(1:lag\_order)]
17. pred\_rf <- numeric(horizon)
18. for (i in 1:horizon){set.seed(1)
19. fit\_rf <- randomForest(X\_train, y\_train)
20. pred\_rf[i] <- predict(fit\_rf, X\_test)
21. y\_train <- y\_train[-1]
22. X\_train <- X\_train[-nrow(X\_train), ] }
23. pred\_rf
24. exp\_term <- exp(cumsum(pred\_rf))
25. last\_obs <- as.vector(tail(ts\_org, 1))
26. backtrans\_fc <- last\_obs \* exp\_term
27. y\_pred <- ts(backtrans\_fc, start = 2019, frequency = 1)
28. library(fpp2)
29. forecast::accuracy(as.numeric(y\_pred), as.numeric(y\_test))
30. ts\_fc <- cbind(ts,pred = c(rep(NA, length(ts\_org)), y\_pred))
31. plot\_fc <- ts\_fc |> autoplot() + theme\_minimal()
32. 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：

1. 时间序列预测建模的**基本原理**是利用过去已有数据的趋势、季节性和自相关性来建立数学模型预测未来值，也就是说利用了过去值与未来值的相关性。比如在机器学习中，时间序列的滞后变量被用作特征（X），已有值作为目标（Y），通过监督学习建模。

②参考以上代码和数据，时间序列预测模型的建模的**基本过程**是

1）**数据预处理**，包括选择特定站点VOLPla和鱼类CHE；随后提取DATE时间列和BIOMASS生物量列构建ts时间序列；接着划分训练集ts\_org（1994-2018）和测试集y\_test（2019-2020）。

2）**进行时间序列变换**（平稳化），使用对数差分变换来消除指数增长趋势和方差变化（ log() |> diff(1) ）。

3）**构建滞后特征**，使用embed()生成滞后矩阵ts\_mbd，每行包含Y\_train和X\_train值。

4）**进行模型训练和预测**，用randomForest() 拟合Y=f(Xlagged)，然后迭代预测。

5）将得到的结果**进行反变换和评估预测可靠性**，将对数差分预测转换回原始尺度，再使用forecast::accuracy()计算测试集（2019-2020）的预测误差（可计算RMSE、MAE）

6）**可视化结果**，用autoplot()绘制原始序列与预测值，进行效果的直观对比。

（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？**

答：可以从时间戳提取并构造建模特征，是利用时间本身的趋势性（直接提取年份、月份或者时间序号）、周期性（根据季节或者月份）和关键时间点信息（根据特定的时间点—月初、节假日、禁渔期等）。

其依据主要是以下三点：

1）**日期时间特征**：时间序列中的数值可能与特定日期/季节存在相关关系，例如温度-降水，工作日-车流量等，此时可将日期/季节作为特征纳入模型。

2）**滞后特征**：通过时间戳可以捕捉一定时段前后两点的相关关系，例如“往年今日”与当前时点的相关性。

3）**窗口特征**：时间序列所对应的系统可能存在一定时段内或长期的趋势，可以通过滑动窗口捕捉局部趋势和长期趋势，或采取窗口内的加和、最大/最小值等作为建模特征。