# **Q1：根据构建机器学习时间序列预测模型的相关代码，简述时间序列预测建模基本原理以及基本过程。**

1. 时间序列预测建模的基本原理：通过分析历史时间序列数据，寻找序列中的时间依赖性（即某点观测值与过去值之间的相关关系），分解出序列中的趋势、季节性和周期性行为等模式，构建模型，利用上述时间依赖性和模式对该时间序列的未来值进行预测。
2. 时间序列预测建模的基本过程：

|  |  |
| --- | --- |
| **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。 | |
| 1. library(tidyverse) 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) | **代码解读** |
| **数据预处理：**  读入待分析时间序列数据后，筛选感兴趣的VOLPla站点的CHE鱼类，保留DATE和BIOMASS字段，用ts()函数创建时间序列对象；  对该时间序列对象进行对数转化，减少数据偏斜，然后进行一阶差分，保证数据的平稳性。 |
| **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。 | |
| 1. lag\_order <- 2 2. horizon <- 2 3. ts\_mbd <- embed(ts.trd, lag.order + 1) 4. Y\_train <- ts\_mbd[, 1] 5. X\_train <- ts\_mbd[, -1] 6. y\_test <- window(ts, start = 2019, end = 2020) 7. x\_test <- ts\_mbd[nrow(ts\_mbd), c(1:lag\_order)] | **代码解读** |
| **训练集与测试集的划分和准备：**  lag\_order <- 2 规定构建滞后特征时使用的历史值数量，horizon <- 2 规定要预测未来的2个值；  用embed()对1994-2018年的BIOMASS时间序列构建矩阵，第1列作为target，第2、3列作为fearture；  用2019和2020年的BIOMASS时间序列作为测试集。 |
| 1. pred\_rf <- numeric(horizon) 2. for (i in 1:horizon){set.seed(1) 3. fit\_rf <- randomForest(X\_train, y\_train) 4. pred\_rf[i] <- predict(fit\_rf, X\_test) 5. y\_train <- y\_train[-1] 6. X\_train <- X\_train[-nrow(X\_train), ] } 7. pred\_rf | **代码解读** |
| **随机森林算法训练：**  使用randomForest包中的同名函数和准备好的训练集，进行随机森林算法训练，并用测试集的feature即x\_test产生2019和2020年的初步预测值，存放于pred\_rf。 |
| **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。 | |
| 1. exp\_term <- exp(cumsum(pred\_rf)) 2. last\_obs <- as.vector(tail(ts\_org, 1)) 3. backtrans\_fc <- last\_obs \* exp\_term 4. y\_pred <- ts(backtrans\_fc, start = 2019, frequency = 1) | **代码解读** |
| **预测结果逆向转换：**  因为在建模时，时间序列经过了对数化和一阶差分操作，所以要将pred\_rf恢复到原始尺度，并与2018年BIOMASS相乘，相对变化量转回绝对变化量。 |
| 1. library(fpp2) 2. forecast::accuracy(as.numeric(y\_pred), as.numeric(y\_test)) 3. ts\_fc <- cbind(ts,pred = c(rep(NA, length(ts\_org)), y\_pred)) 4. plot\_fc <- ts\_fc |> autoplot() + theme\_minimal() 5. plot\_fc | **代码解读** |
| **模型表现评估：**  使用forecast::accuracy（）计算预测值-经验值的RMSE等，作为模型表现的评估指标，再使用fpp2包可视化地比较经验数据和模型预测结果。  必要时，还可以根据当前评估结果，对模型进行进一步优化。 |

# **Q2：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？**

从时间戳构造建模特征的依据是时间戳通常包含一定的周期性、趋势性等多尺度的特征，也能指示潜在的外部因素影响。将这些特征加入模型有助于提高序列模型的预测精度。

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。