**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？

|  |  |  |  |
| --- | --- | --- | --- |
|  | | 机器学习建模 | 传统统计模型 |
| 建模目标 | 核心目标 | 执行建模任务时，使**预测准确率最大化** | 推断并**解读变量之间的关系**，以及对模型**机制的解释** |
| 适用场景 | 适用于**高维度、非线性、数据量大**的情况 | 适用于**低维度、线性关系或简单非线性关系、数据样本量小**的情况 |
| 建模算法 | 算法设计思想 | 以梯度下降等方法**优化损失函数**为思想核心，**依赖计算资源**；  侧重**模型泛化**，需要通过增加系数等方式**防止过拟合** | 以**概率分布假设及其验证**为思想核心，**依赖数学推导**；  侧重**参数显著性**，采用**变量筛选**等思想简化模型 |
| 模型复杂度 | 以**多参数复杂模型**居多，对非线性关系的建模能力强，但**可解释性弱** | 以**少参数的简单建模**方法为主，**可解释性强**，能够较清晰地解读各变量之间的关系 |
| 模型评估 | 评估指标 | 对模型的评估基于其**预测性能**，使用准确率、AUC、RMSE等指标 | 基于**检验统计量**：p值、置信区间、R2、AIC/BIC等 |
| 验证方法 | 使用**交叉验证**、独立的**测试集**评估模型的**泛化能力** | 使用**理论假设检验**进行验证 |

**问题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？

**递归消除选择**是使用递归特征消除（Recursive Feature Elimination, RFE）方法，通过递归迭代训练模型，以模型输出为标准进行特征重要度排序，逐步剔除不重要特征，以逐步实现特征子集的收缩、优化，最终达到指定的特征数量。

随机森林等树模型在决策树的生长过程中会自动计算变量重要性，且每次分割都只选择一个最佳特征，而随机森林方法本身通过特征随机子采样（mtry） 降低了对单个冗余特征的依赖，对冗余特征不敏感，基于上述原因，caret包中的随机森林等树模型没有显式的特征选择过程。

但caret包提供了rfe()函数，若遇到极端高维数据情况、算力限制、问题存在较高的解释性要求时，可以通过设置rfe()函数实现递归特征消除过程。