**Method**

**数据预处理**

我们采用随机森林分类器来识别分类任务相关的预测因子。RF 是一种集成学习方法，通过构建多个决策树并对其预测求平均值，显著降低了过拟合的风险。它还能更有效地处理特征之间的复杂交互。19 个候选变量的重要性排列显示在表1 。缺失值通过预测均值匹配进行估算。

**机器学习**

对于机器学习，发现队列按照8：2被分为训练集和测试集。根据RF前12个变量构建预测模型，我们采用了5种机器学习算法来构建模型，包括逻辑回归、随机森林、SVM、GBM和决策树。对于每种方法我们均使用15倍分层交叉验证来训练模型，以确保对性能进行稳健的评估。每种机器学习算法的性能通过灵敏度、特异度、阳性预测值、阴性预测值、召回率、F1 分数和 AUC 来衡量。在测试集中对已建立的模型进行评估，并在验证队列中进行验证。Python 是用于数据分析的主要软件工具。

**模型解释**

我们使用 SHAP 分析来解释模型预测。计算每个预测变量的 SHAP 值以确定其特征重要性，即方向和范围对预测的影响。正 SHAP 值表示相应特征导致 AKI 风险较高，而负 SHAP 值表示相应特征导致 AKI 风险较低。SHAP 值的大小表示每个特征对预测性能的贡献程度。我们使用蜂群图来说明重要特征的排名，使用瀑布图来说明每个特征如何对单个预测做出贡献，并使用依赖关系图来说明特征和结果之间的关系。

**Results**

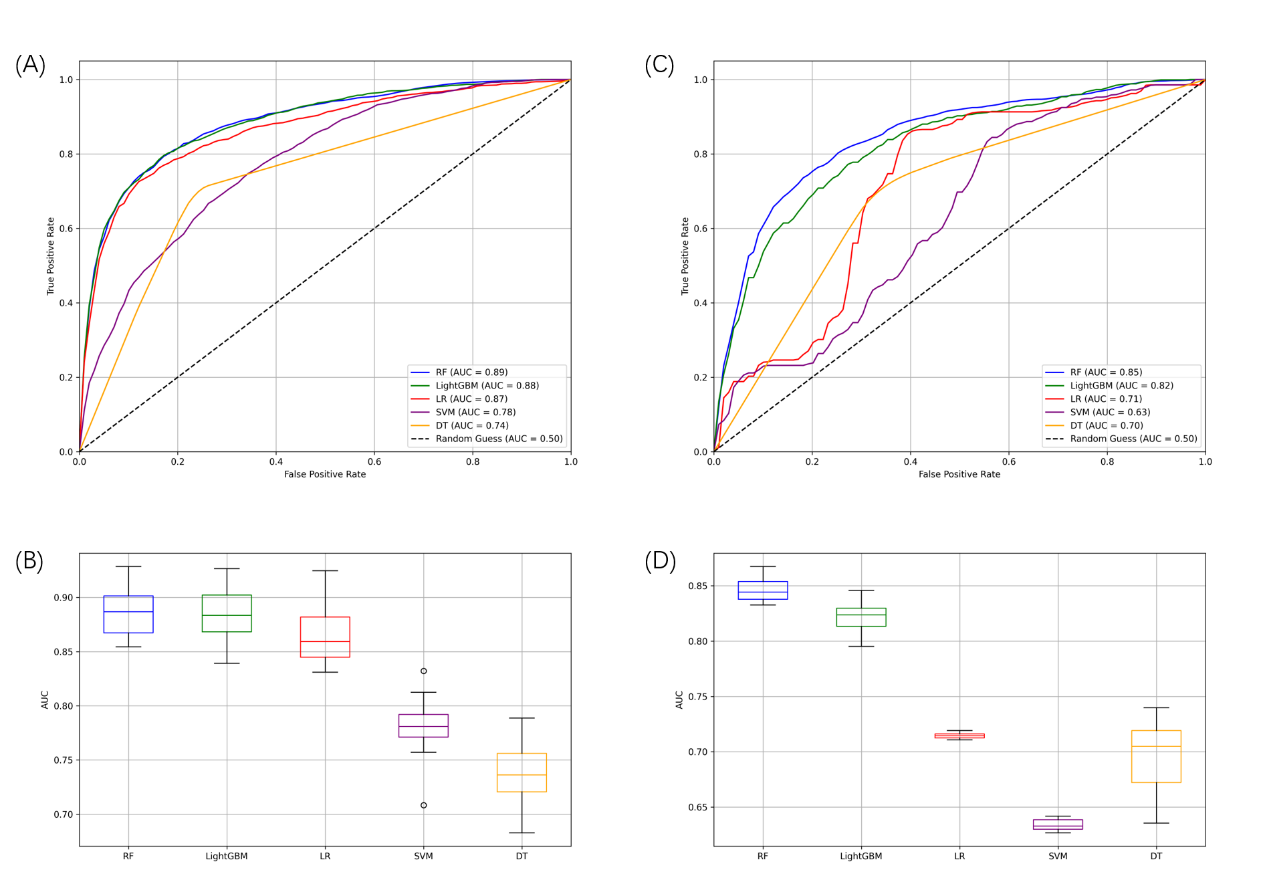
1、内部队列验证结果：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Table The efficacy of five machine learning models in predicting diabetic nephropathy in the internal validation cohort. | | | | | | | |
| Machine learner | Sensitivity | Specificity | Accuracy | Precision | AUC | Recall | F1 Score |
| RF | 0.72 | 0.88 | 0.81 | 0.81 | 0.89 | 0.72 | 0.76 |
| LightGBM | 0.73 | 0.87 | 0.81 | 0.80 | 0.88 | 0.73 | 0.76 |
| LR | 0.68 | 0.88 | 0.80 | 0.80 | 0.87 | 0.68 | 0.74 |
| SVM | 0.55 | 0.82 | 0.71 | 0.68 | 0.78 | 0.55 | 0.60 |
| DT | 0.68 | 0.76 | 0.73 | 0.66 | 0.74 | 0.68 | 0.67 |
| RF: random forest; LightGBM: light gradient boosting machin; LR: logistic regression; SVM: support vector machine; DT: decision tree. | | | | | | | |

2、外部队列验证结果：

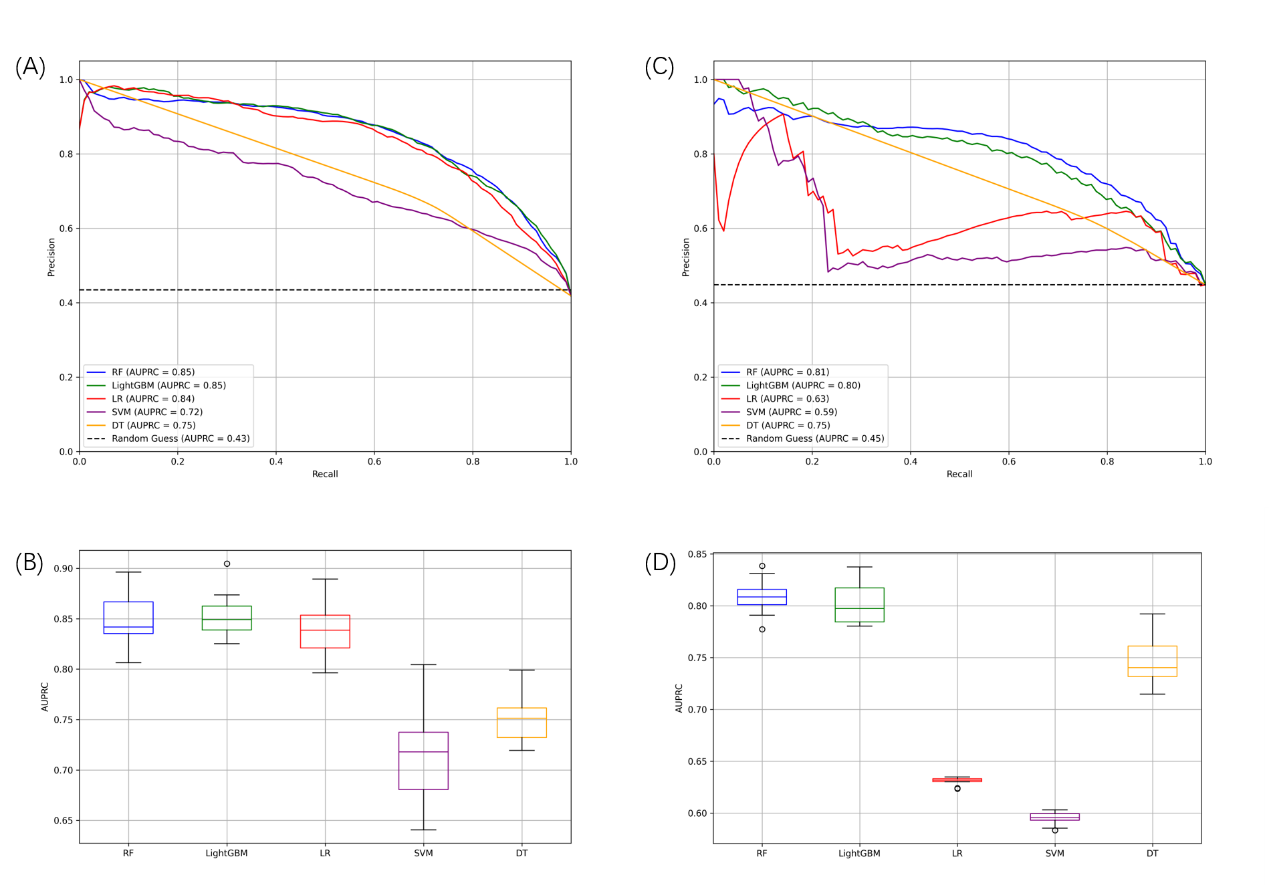
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Table The efficacy of five machine learning models in predicting diabetic nephropathy in the external validation cohort. | | | | | | | |
| Machine learner | Sensitivity | Specificity | Accuracy | Precision | AUC | Recall | F1 Score |
| RF | 0.81 | 0.73 | 0.77 | 0.71 | 0.85 | 0.81 | 0.76 |
| LightGBM | 0.81 | 0.71 | 0.75 | 0.69 | 0.82 | 0.81 | 0.75 |
| LR | 0.82 | 0.63 | 0.71 | 0.64 | 0.71 | 0.82 | 0.72 |
| SVM | 0.48 | 0.63 | 0.56 | 0.51 | 0.63 | 0.48 | 0.50 |
| DT | 0.73 | 0.65 | 0.69 | 0.63 | 0.70 | 0.73 | 0.68 |
| RF: random forest; LightGBM: light gradient boosting machin; LR: logistic regression; SVM: support vector machine; DT: decision tree. | | | | | | | |

3、ROC：



1. B为内部验证队列，C、D为外部验证队列

4、PR曲线



5、SHAP值

