Mine Engineering 矿山工程, 2025, 13(4), 801-809
Published Online July 2025 in Hans. https://www.hanspub.org/journal/me
https://doi.org/10.12677/me.2025.134091

Hans汉斯

基于机器学习的致密气气井动态储量预测模型研究

王琳曼，王　璐，赵安琪

重庆科技大学石油与天然气工程学院，重庆

收稿日期：2025年6月6日；录用日期：2025年7月10日；发布日期：2025年7月24日

摘　要

精确预测气井的动态储量是致密气高效开发的关键基础。本文针对致密气藏气井动态储量预测问题，基于X致密气藏某区块400口致密气井的实际数据，采用机器学习方法开展致密气井动态储量预测。首先，通过对原始数据进行预处理，其次，利用皮尔逊相关系数法分析，筛选出影响动态储量的主控因素。接着，运用机器学习的方法建立了动态储量的预测模型，最后，使用粒子群优化算法对超参数进行优化。结果表明利用随机森林算法预测结果良好，能够很好地预测气井的动态储量。该技术为提升气井动态储量的预测能力提供了新途径。

关键词

致密气，动态储量预测，机器学习

Study on Dynamic Reserve Prediction Model for Tight Gas Wells Based on Machine Learning

Linman Wang, Lu Wang, Anqi Zhao

School of Petroleum Engineering, Chongqing University of Science and Technology, Chongqing

Received: Jun. 6th, 2025; accepted: Jul. 10th, 2025; published: Jul. 24th, 2025

Abstract

Accurate prediction of gas wells' dynamic reserves is a critical foundation for the efficient development of tight gas reservoirs. Aiming at the technical problems in predicting dynamic reserves of tight gas wells, this study uses machine learning methods to predict dynamic reserves of tight gas

文章引用: 王琳曼, 王璐, 赵安琪. 基于机器学习的致密气气井动态储量预测模型研究[J]. 矿山工程, 2025, 13(4): 801-809. DOI: 10.12677/me.2025.134091

wells based on the actual data of 400 tight gas wells in a block of the X Tight Gas Reservoir. First, the original data is preprocessed. Second, the Pearson correlation coefficient method is used for analysis to screen out the main controlling factors affecting dynamic reserves. Then, machine learning methods are applied to establish a prediction model for dynamic reserves. Finally, the particle swarm optimization algorithm is used to optimize hyperparameters. The results show that the random forest algorithm provides good prediction results and can effectively predict the dynamic reserves of gas wells. This study provides a new technical means for predicting the dynamic reserves of gas wells.

Keywords

Tight Gas, Dynamic Reserve Prediction, Machine Learning

1. 前言

致密气作为非常规天然气资源的重要组成部分，在我国能源安全和低碳转型中占据着关键地位。低渗透致密气田的资源储量超过国内非常规天然气总量的 50%，其高效开发对于缓解常规气田产量下降的压力、优化能源供给结构具有深远意义。然而，致密气具有孔隙度、渗透率和含气饱和度较低，以及储层物性差等特点，动态储量预测面临较大挑战。由此可见，动态储量评估结果的精确性对于气田开发方案的设计不可或缺。故而，探索一种适用于致密气井动态储量的预测方法，对推动致密气藏的经济高效开发有着显著的现实意义。

随着数据分析、人工智能等技术在油气智能勘探开发及生产领域的兴起，油气工业领域迎来了新的发展机遇[1]。近年来，许多研究者通过应用人工智能技术，在油气领域取得了显著进展，通过收集油气田的多种特征数据并构建机器学习模型，朱庆忠等人基于随机森林算法建立煤层气直井产气量模型[2]；柳洁等人基于复合机器算法建立致密气井产能预测模型[3]；唐钦锡等人基于 XGBoost 的压裂水平井建立产能预测[4]。学者们能够对油气井的产量和地层特性进行预测、识别与分析，这为该领域的进一步发展提供了动力[5]。

为提升致密气藏动态储量预测的准确性，本文通过文献调研，采用机器学习方法对地质参数、工程参数等影响气井动态储量的关键因素进行了深入分析，构建了高精度的动态储量预测模型，该技术为提升气井动态储量的预测能力提供了新途径。

2. 数据整理与预处理

2.1. 数据整理

针对 X 某区块的 400 口实际气井数据，影响气井动态储量的因素繁多，主要分为地质因素和工程因素两大类。本文选取的地质因素包括渗透率、有效厚度、含水饱和度、地层压力等；工程因素包括施工排量、前置液量、陶粒用量、砂比、含砂浓度等。数据的整理与预处理为后续主控因素识别及动态储量预测模型的建立提供了基础。

2.2. 数据的预处理

在气田实际生产中，由于人为操作、设备故障等原因，数据常出现缺失或噪声问题。为确保后续模

型构建的准确性，本研究结合石油工程专业知识和数据特性进行数据预处理，主要步骤包括重复值、缺失值、异常值处理和数据标准化。异常值采用箱型图法进行识别和处理，缺失值通过 K 最邻近填补法(KNN)进行填补，最后采用 Min-Max 标准化方法对数据进行归一化处理。

(1) 箱型法处理异常值

异常值(又称离群点)是指远离绝大多数样本点的数据点，这类数据在数据集中通常呈现出统计意义上的极端性。在数据探索阶段，需对这类异常值进行识别并实施合理处理。本文采用箱型法处理异常值，该方法基于数据的分位数特性识别异常点。如图 1 所示，下四分位数(Q1)为数据的 25%分位点对应值，中位数(Q2)为 50%分位点对应值，上四分位数(Q3)为 75%分位点对应值。四分位差(IQR)计算公式为 IQR = Q3 − Q1，上须和下须分别通过以下公式确定：上须 = Q3 + 1.5IQR，下须 = Q1 − 1.5IQR。当变量数据值大于上须或小于下须时，通常可将这类数据点视为异常点。

本文基于 Python 的 matplotlib 模块，通过 boxplot 函数对各特征数据列进行箱线图绘制，实现异常值的可视化检测。针对检测出的异常数据点(即超出箱线图须线范围的极端值)，采用删除处理以净化数据集。

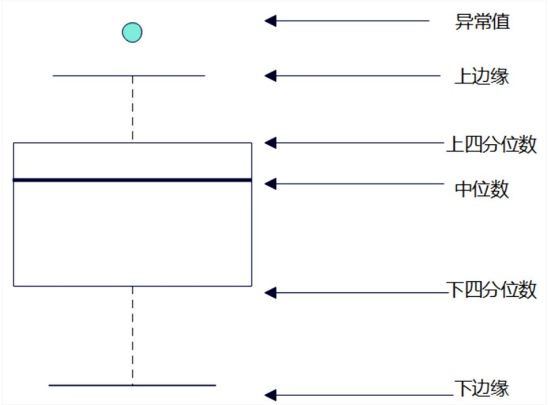

Figure 1. Example of a box plot
图 1. 箱型图示例图

(2) K 近邻法(KNN)处理缺失值

K 近邻法(KNN)是一种通过计算缺失值的 K 个最相似观测值的加权平均来填补缺失值的技术。其核心思想是基于欧几里得距离度量样本间的相似性。具体步骤为：首先，针对包含缺失值的样本，使用欧几里得距离计算其与数据集中其他完整样本之间的距离，筛选出距离最近的 K 个邻居。然后，提取这 K 个邻居中对应特征的有效数值，计算各邻居与缺失值样本的距离倒数，并归一化为权重。最后，通过权重对有效数值进行加权平均，得到缺失值的填补结果。

(3) 数据的标准化

数据归一化(如 Min-Max 标准化)是数据预处理的核心手段。在实际应用中，地质参数(如渗透率、有效厚度)、工程参数(如施工排量、陶粒用量)与气井动态储量的量纲体系与取值范围差异显著。此类差异可能导致机器学习模型训练时出现梯度失衡，例如量纲较大的特征主导损失函数优化方向，造成模型收敛效率下降。数据标准化方法通过对原始数据进行线性变换，消除不同特征间的量纲干扰，使各维度数据处于相同尺度，从而为后续机器学习模型的高效训练与精准预测奠定基础。本文采用的是 Min-Max 标准化，公式如下：

$$y_i = \frac{x_i - x_{\max}}{x_{\max} - x_{\min}} \tag{2.1}$$

式中，x_i 表示页岩气井开采中的原始数据，x_{max} 和 x_{min} 分别为原始数据中的最大值、最小值，y_i 为归一化处理之后的无量纲数据。

本文对原始 400 组数据依次进行异常值处理、缺失值填补和数据归一化后，最终得到 380 组可用样本，以此构建 x 区块致密气井动态储量数据集。

2.3. 数据集划分

本文将数据集划分为训练集和测试集两部分，其中训练集与测试集的样本划分比例为 8:2。训练集用于拟合模型参数，使模型学习数据特征与动态储量之间的映射关系；测试集用于评估模型对未知数据的泛化能力，避免过拟合(模型过度学习训练数据中的噪声和局部模式，导致在真实场景中预测失效)。具体划分见表 1。

Table 1. Dataset partitioning
表 1. 数据集划分

	样本数占比	样本数量
训练集	80%	304
测试集	20%	76

3. 主控因素分析

开展影响致密气井动态储量因素的重要性评估，能够更精准地判别各因素对动态储量的贡献差异，进而从繁杂的影响因素中提炼出起主导作用的关键因素。本文首先采用皮尔逊(Pearson)相关系数法分析相关因素的线性相关性，接着采用随机森林算法在训练过程中进行特征选择，将各个参数对模型训练的重要性进行排序。

(1) 皮尔逊相关系数法

皮尔森相关系数分析法是由皮尔逊提出的并广泛地应用数据分析之中的统计指标[3]。相关系数范围[−1, 1]，正数表正相关(变量同向变化)，负数表负相关(变量反向变化)，绝对值越接近 1，相关性越强；越接近 0，相关性越弱，0 表示无线性相关。如图 2 所示，可以得出，相关性的大小关系，有效厚度 > 渗透率 > 地层压力 > 含水饱和度 > 砂比 > 陶粒用量 > 施工排量 > 前置液量 > 含砂浓度。

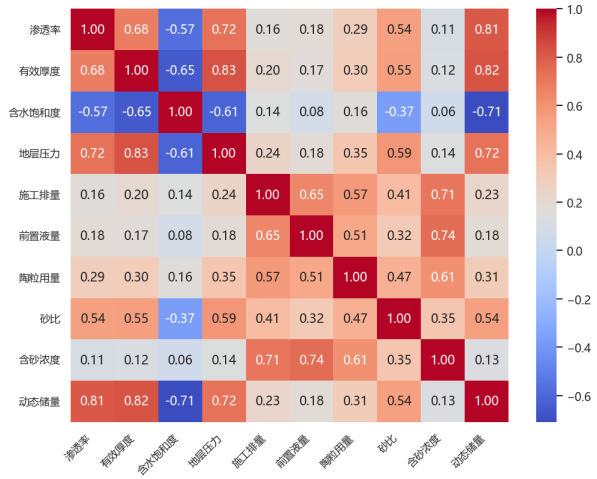

Figure 2. Factor correlation coefficient diagram
图 2. 因素相关系数图

(2) 随机森林算法特征选择

本研究运用随机森林算法对各影响因素与动态储量的关联程度进行量化评估与主控因素解析。随机森林可以通过计算特征在决策树中的节点分裂增益来量化特征的重要性。具体来说，每个特征在所有决策树中的平均不纯度减少量(如基尼不纯度或均方误差)可以作为该特征的重要性指标。在本文中，通过随机森林模型的特征重要性分析，可以明确各主控因素对动态储量的具体贡献程度。

如图 3 所示，渗透率、有效厚度、含水饱和度、地层压力的累计贡献率达到 70%，是影响动态储量的核心地质参数；而工程参数中，施工排量和前置液量的累计贡献率仅为 16%，对动态储量的影响显著低于地质参数。

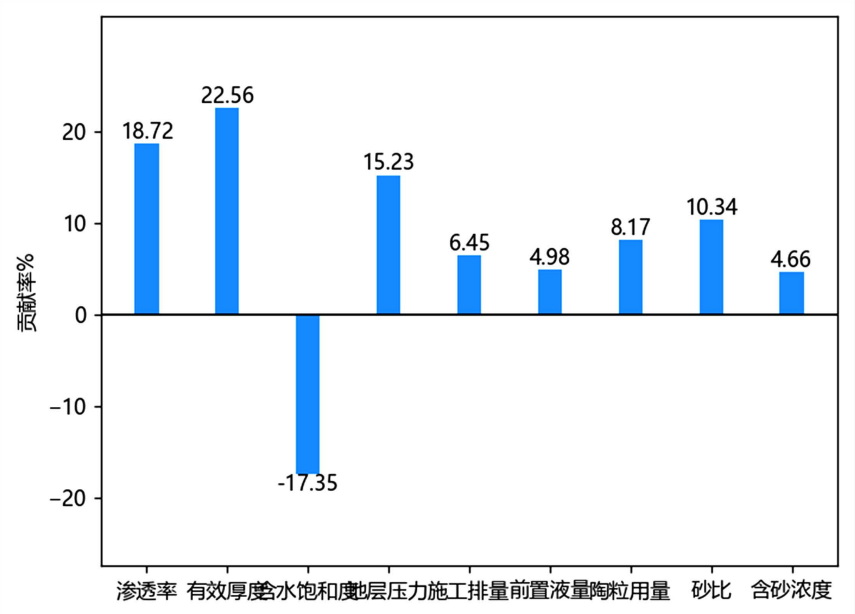

Figure 3. Statistical chart of parameter contribution rate
图 3. 参数贡献率统计图

结合皮尔逊相关系数法(强相关)与随机森林特征选择结果(高贡献率)分析，地质因素的影响程度远高于工程因素。因此，筛选出渗透率、有效厚度、含水饱和度、地层压力和砂比共 5 个参数作为影响气井动态储量的主控因素。主控因素的筛选是构建气井动态储量预测模型的关键前提。

渗透率：渗透率是衡量岩石中流体流动能力的重要参数。较高的渗透率意味着气体更容易在储层中流动，从而提高气井的动态储量。在模型中，渗透率对动态储量的影响可以通过其在决策树中的重要性来量化。

有效厚度：有效厚度反映了气藏中具有产气能力的岩层厚度。有效厚度越大，气井的动态储量通常越高。在随机森林模型中，有效厚度的高重要性表明其对动态储量的贡献显著。

含水饱和度：含水饱和度表示储层中水的含量。较高的含水饱和度会降低气体的有效孔隙体积，从而减少动态储量。模型中含水饱和度的负相关性反映了这一物理现象。

地层压力：地层压力是驱动气体流动的关键因素。较高的地层压力有助于气体的产出，从而增加动态储量。在模型中，地层压力的重要性表明其对动态储量的直接影响。

砂比：砂比反映了压裂过程中砂的用量。适当的砂比可以提高裂缝的导流能力，从而增加动态储量。在模型中，砂比的重要性表明其对动态储量的间接影响。

4. 气井动态储量预测模型

4.1. 机器学习算法

(1) 随机森林(RF)

随机森林是一种基于集成学习(Ensemble Learning)的机器学习算法，通过构建多棵决策树并结合它们的预测结果来提升模型的准确性、鲁棒性和泛化能力。其核心思想源于 bootstrap aggregating (装袋法)，并引入双重随机性：

样本随机：通过有放回抽样生成多组训练集(每组含 n 个样本)，约 36.8%未抽中样本作为袋外数据用于验证。特征随机：节点分裂时随机选 K 个特征(K << 总特征数)，降低树间相关性以避免过拟合。结果集成：回归任务取所有树预测值的平均值，公式为：

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \tag{4.1}$$

其中，T 为树的数量，$f_t(x)$ 为第 t 棵树对样本 x 的预测值。

(2) 支持向量机(SVM)

分类目标为在线性可分的特征空间中找到一个超平面，使不同类别样本的间隔(margin)最大化，从而降低分类误差、提升模型泛化能力。对于非线性数据，可通过核函数(如 RBF、多项式核)将样本映射到高维空间，再寻找最优超平面。其核心优势包括：适用于小样本、高维度数据(如基因表达、图像特征)；通过间隔最大化和正则化项控制过拟合，抗噪声能力强；核技巧可高效处理非线性分类问题，无需显式高维映射。

(3) XGBoost 算法

XGBoost 属于梯度提升树(Gradient Boosting Tree, GBT)的一种，核心思想是迭代训练多个弱学习器(决策树)，并将它们组合成一个强学习器；每棵树通过拟合前一轮预测的残差(利用梯度下降优化)来逐步降低模型整体误差。相较于传统 GBT，XGBoost 通过引入正则化、稀疏感知算法和并行处理等机制，进一步提升了模型的泛化能力和训练效率。

4.2. 模型超参数的设置

超参数优化是机器学习模型训练中至关重要的环节，指的是对模型训练前手动设定、无法通过数据直接学习的超参数进行调优的过程。其核心目标是通过系统性地搜索不同的超参数组合，找到使模型在测试集上表现最佳的参数配置，从而提升模型的泛化能力和预测精度。不同算法的超参数类型和作用各异，以随机森林、支持向量机为例，其部分超参数如表 2 所示。本文采用粒子群优化算法对超参数进行优化。

4.3. 模型评价指标

回归模型常用的 3 种评估指标包括均方根误差(RMSE)、均方误差(MSE)和决定系数(R^2)。

均方根误差(RMSE)：是 MSE 的平方根，是衡量模型预测值和数据实际值误差大小的一种常用指标。

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.2}$$

均方误差(MSE)：均方误差的概念来源于最小二乘法。在最小二乘法中，目标是通过找到一条最佳拟合的直线或曲线，使预测值与实际值之间的平方误差总和最小[4]。公式为：

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{4.3}$$

决定系数(R^2)是衡量回归模型拟合优度的核心指标，表示因变量的变异中能被自变量解释的比例。其取值范围为[0, 1]。

$$R^2 = 1 - \frac{\sum_i \hat{y}^{(i)} - y^{(i)}}{\sum_i \overline{y}^{(i)} - y^{(i)}} \tag{4.4}$$

图 4~6 的模型测试集等值线图和表 3 的模型评估指标可知：随机森林模型的训练集和测试集上均方根误差(RMSE)均较低，表明其回归误差最小、性能最优。该模型在训练集和测试集上的 R^2 分别为 0.864 和 0.847，说明模型对训练数据的拟合能力与对未知数据的泛化能力均较强，可靠性较高。XGBoost 模型的各项指标次之，其训练集和测试集 R^2 分别为 0.858 和 0.824，回归性能略低于随机森林模型。

Table 2. Description of main built-in parameters of the algorithm
表 2. 算法主要内置参数说明

算法名称	参数范围	最佳参数
随机森林	n_estimators: (10, 100, 10)	35
	max_depth: (1, 30, 1)	18
	max_features: (0.1, 1)	0.487
	min_samples_split: (2, 10, 1)	5
支持向量机	C: (1, 250)	218.14
	Kernel：[‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’]	
	Gamma: (0, 10)	4.56
	Max_iter: (100, 500, 1)	380
	Espsilon: (0.01, 1)	0.084
XGboost 算法	n_estimators: (10, 300, 10)	240
	min_child_weight: (1, 10, 1)	3
	Subsample: (0.1, 1)	0.7
	max_depth: (2, 20, 1)	8

Table 3. Model evaluation metrics
表 3. 模型评估指标

算法	样本集	MSE	RMSE	R^2
XGboost	训练集	0.956	0.978	0.858
	测试集	1.256	0.952	0.824
随机森林	训练集	0.912	0.955	0.864
	测试集	0.945	0.935	0.847
支持向量机	训练集	1.422	1.192	0.788
	测试集	2.456	1.598	0.645

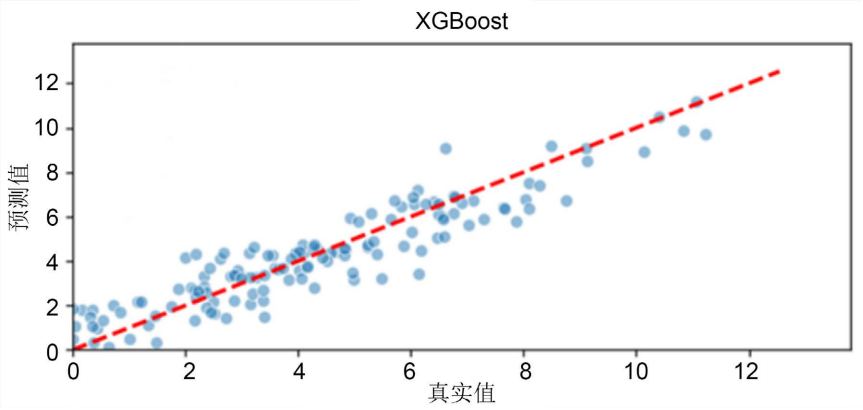

Figure 4. Contour plot of XGBoost model test set
图 4. XGBoost 模型测试集等值线图

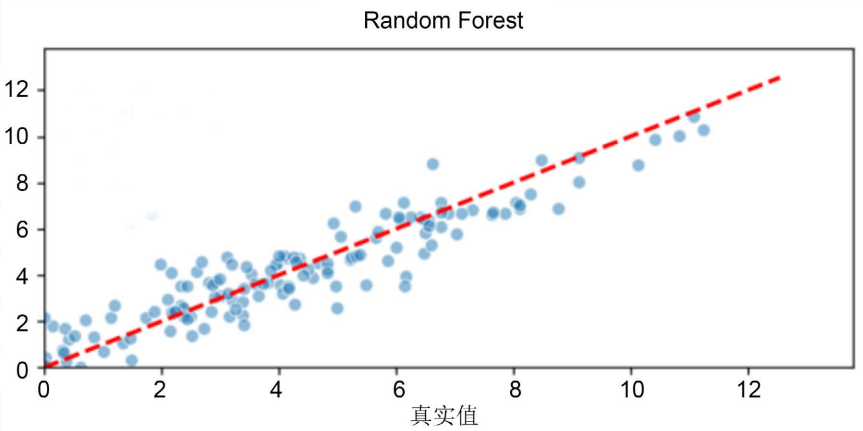

Figure 5. Contour plot of random forest model test set
图 5. 随机森林模型测试集等值线图

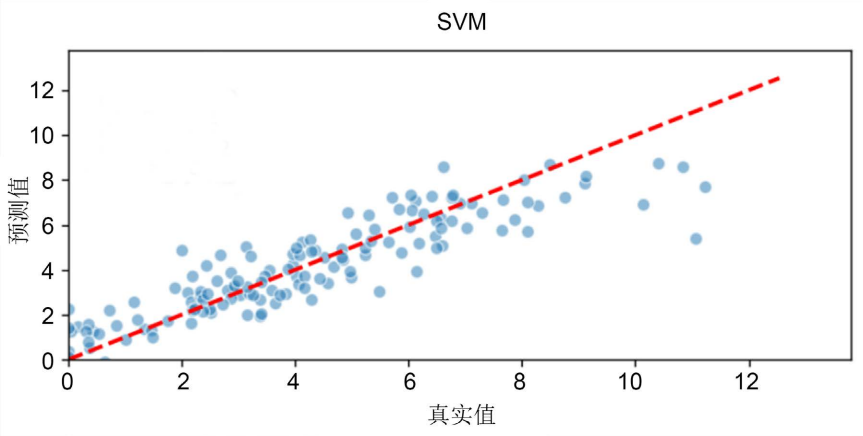

Figure 6. Contour plot of SVM model test set
图 6. SVM 模型测试集等值线图

5. 模型应用

为进一步验证致密气井动态储量预测模型的准确性，选取 X 区块未参与建模的气井参数，输入至随

机森林预测模型进行动态储量预测。通过对比预测值与真实值，对两者的拟合度进行评估：预测值与真实值的散点图显示，数据点集中分布在 y = x 直线(45˚对角线)附近，决定系数(R^2)为 0.89，表明模型解释了 89%的真实值变异，对未知数据的泛化能力较强。

6. 结论

(1) 对 400 组原始数据依次进行异常值处理(箱型图法)、缺失值填补(KNN 法)和数据归一化(Min-Max 标准化)，最终获得 380 组有效样本，消除了量纲差异与数据噪声，为模型构建提供了高质量数据集。

(2) 本研究通过随机森林模型对致密气井动态储量进行预测，不仅实现了高精度的预测结果，还通过特征重要性分析和决策树可视化，详细解释了各主控因素对动态储量的影响。研究表明，渗透率、有效厚度、含水饱和度和地层压力是影响动态储量的核心地质参数，而砂比是重要的工程参数。这些因素通过不同的物理机制影响气井的动态储量，模型的预测机理与地质和工程原理相一致。

(3) 本研究基于筛选出的 5 项主控因素，分别构建随机森林、支持向量机(SVM)和 XGBoost 三种机器学习模型进行动态储量预测。通过对比发现：随机森林模型泛化能力和预测精度优于其他模型；XGBoost 模型次之。综合评估表明，随机森林模型更适用于致密气井动态储量的高精度预测。

基金项目

重庆科技大学研究生科技创新项目“基于静态数据的致密气井产量预测模型研究”(YKJCX2420139)。

参考文献

[1] 林伯韬, 郭建成. 人工智能在石油工业中的应用现状探讨[J]. 石油科学通报, 2019, 4(4): 403-413.

[2] 朱庆忠, 胡秋嘉, 杜海为, 等. 基于随机森林算法的煤层气直井产气量模型[J]. 煤炭学报, 2020, 45(8): 2846-2855.

[3] 柳洁, 田冷, 刘士鑫, 等. 基于复合机器算法的致密气井产能预测模型——以鄂尔多斯盆地 SM 区块为例[J]. 大庆石油地质与开发, 2024, 43(5): 69-78.

[4] 唐钦锡, 王涛. 基于 XGBoost 的压裂水平井产能预测[J]. 中国石油和化工标准与质量, 2023, 43(24): 15-17.

[5] 郎岳, 张金川, 王焕第, 等. 页岩气地质评价智能化的应用与展望[J]. 大庆石油地质与开发, 2022, 41(1): 166-174.