**Scientific Reports**
**Supplementary Information**

# Unraveling Green Innovation Drivers in High Carbon-Lock-in Process Industries: An Interpretable Machine Learning Approach with Temporal Validation

**Hao Huang 1,2, Yujun Hu 1*, Xinqiang Li 3, Yiwei Huang 1, Wenxuan Duan 1, Bowen Zhang 1, Anxue Li 1,2, Yang Li 1**

**1 School of Economics and Management, Shandong Institute of Petroleum and Chemical Technology, Dongying, China**
**2 Key Laboratory of Digital Transformation and Application of Energy and Chemical Enterprises in Dongying City, Dongying, China**
**3 PetroChina Xinjiang Oilfield Company Luliang Oilfield Operation Area, Karamay, China**

*** Corresponding author: Yujun Hu. E-mail: hu342758@163.com. ORCID: 0009-0005-3957-5375**

**Contents:**

Supplementary Table S1. Optimal hyperparameters of the XGBoost model.
Supplementary Table S2. List of abbreviations used in this study.
Supplementary Note: Detailed Results of the Diebold–Mariano Test.
Supplementary Note: Full Results of Random Perturbation and Hyperparameter Sensitivity Analysis.
Supplementary Note: Extended Variable Definitions and Variable Exclusion Rationale.
Supplementary Note: Detailed Stability Statistics of SHAP Feature Importance Rankings.

**Supplementary Table S1. Optimal hyperparameters of the XGBoost model.**

| Parameter | Description | Value |
|---|---|---|
| n_estimators | Number of gradient boosted trees | 300 |
| max_depth | Maximum tree depth for base learners | 4 |
| learning_rate | Boosting learning rate | 0.1 |
| reg_alpha | L1 regularization term on weights | 2 |
| reg_lambda | L2 regularization term on | 2 |

| | | |
|---|---|---|
| | weights | |
| min_child_weight | Minimum sum of instance weight (hessian) needed in a child | 3 |
| gamma | Minimum loss reduction required to make a further partition on a leaf node | 0.1 |
| subsample | Subsample ratio of the training instances | 0.8 |
| colsample_bytree | Subsample ratio of columns when constructing each tree | 0.7 |
| objective | Specify the learning task and the corresponding learning objective | reg:squarederror |
| eval_metric | Evaluation metric for validation data | rmse |
| random_state | Random number seed for result reproducibility | 42 |
| verbosity | Verbosity level of training log output | 0 |

**Supplementary Table S2. List of abbreviations used in this study.**

| **CART** | **Classification and Regression Trees** |
|---|---|
| CatBoost | Categorical Boosting |
| CEO | Chief Executive Officer |
| CNRDS | China Research Data Services Platform |
| CSMAR | China Stock Market & Accounting Research Database |
| CSR | Corporate Social Responsibility |
| DM | Diebold-Mariano |
| EU | European Union |
| GMM | Generalized Method of Moments |
| IRA | Inflation Reduction Act |
| ISO | International Organization for Standardization |
| LightGBM | Light Gradient Boosting Machine |
| MAE | Mean Absolute Error |
| NRBV | Natural Resource-Based View |
| RBV | Resource-Based View |
| RMSLE | Root Mean Squared Logarithmic Error |
| R&D | Research and Development |
| ROA | Return on Assets |
| SDG | Sustainable Development Goals |
| SHAP | SHapley Additive exPlanations |

| | |
|---|---|
| ST | Special Treatment |
| SVR | Support Vector Regression |
| VRIN | Valuable, Rare, Inimitable, Non-substitutable |
| XGBoost | eXtreme Gradient Boosting |

**Scientific Reports**

# Unraveling Green Innovation Drivers in High Carbon-Lock-in Process Industries: An Interpretable Machine Learning Approach with Temporal Validation

**Hao Huang 1,2, Yujun Hu 1*, Xinqiang Li 3, Yiwei Huang 1, Wenxuan Duan 1, Bowen Zhang 1, Anxue Li 1,2, Yang Li 1**

**1 School of Economics and Management, Shandong Institute of Petroleum and Chemical Technology, Dongying, China**

**2 Key Laboratory of Digital Transformation and Application of Energy and Chemical Enterprises in Dongying City, Dongying, China**

**3 PetroChina Xinjiang Oilfield Company Luliang Oilfield Operation Area, Karamay, China**

*** Corresponding author: Yujun Hu. E-mail: hu342758@163.com. ORCID: 0009-0005-3957-5375**

## Supplementary Note

### Detailed Results of the Diebold-Mariano Test

To further confirm the statistical significance of inter-group performance differences among the tree-based ensemble models, the Diebold-Mariano (DM) test based on the squared error loss function is employed. The null hypothesis is that the two competing models have equal out-of-sample prediction accuracy. The complete results are reported below.

| Comparison | DM Statistic | p-value | Significance |
|---|---|---|---|
| XGBoost vs CatBoost | -1.1347 | 0.2565 | Not significant |
| XGBoost vs | -1.3426 | 0.1794 | Not significant |

RandomForest

---

The DM test results show that the differences in prediction errors between XGBoost and the other two first-tier models fail to pass the 5% significance level, indicating no statistically significant difference in prediction accuracy among the three tree-based ensemble models. XGBoost is therefore retained as the core model on the grounds of prediction performance and attribution interpretability.

**Scientific Reports**

# Unraveling Green Innovation Drivers in High Carbon-Lock-in Process Industries: An Interpretable Machine Learning Approach with Temporal Validation

**Hao Huang 1,2, Yujun Hu 1*, Xinqiang Li 3, Yiwei Huang 1, Wenxuan Duan 1, Bowen Zhang 1, Anxue Li 1,2, Yang Li 1**

**1 School of Economics and Management, Shandong Institute of Petroleum and Chemical Technology, Dongying, China**

**2 Key Laboratory of Digital Transformation and Application of Energy and Chemical Enterprises in Dongying City, Dongying, China**

**3 PetroChina Xinjiang Oilfield Company Luliang Oilfield Operation Area, Karamay, China**

*** Corresponding author: Yujun Hu. E-mail: hu342758@163.com. ORCID: 0009-0005-3957-5375**

## Supplementary Note

### Full Results of Random Perturbation and Hyperparameter Sensitivity Analysis

#### 1. Random perturbation robustness (20 random seeds)

To eliminate accidental deviation caused by random initialization, model training and prediction are repeated with 20 distinct random seeds. The standard deviation of out-of-sample $R^2$ across the 20 experiments is only 0.0078, indicating that prediction performance is insensitive to random initialization.

#### 2. Univariate hyperparameter perturbation

Eight core hyperparameters (max_depth, learning_rate, reg_alpha, reg_lambda, min_child_weight, gamma, subsample, colsample_bytree) are adjusted one by one to

observe the fluctuation range of prediction performance. As shown below, under individual perturbation of all parameters, the range of out-of-sample $R^2$ falls between 0.0076 and 0.0241, all falling into the low sensitivity category.

| Perturbation Item | Range of $R^2$ | Sensitivity |
|---|---|---|
| max_depth | 0.0177 | Low sensitivity |
| learning_rate | 0.0241 | Low sensitivity |
| reg_alpha | 0.0211 | Low sensitivity |
| reg_lambda | 0.0096 | Low sensitivity |
| min_child_weight | 0.0133 | Low sensitivity |
| gamma | 0.0205 | Low sensitivity |
| subsample | 0.0076 | Low sensitivity |
| colsample_bytree | 0.0127 | Low sensitivity |

Note: An out-of-sample $R^2$ range below 0.05 under univariate perturbation is adopted as the criterion for low sensitivity.

**3. Joint perturbation (100 random parameter combinations)**

To simulate the extreme scenario of large-scale variation in the parameter space, 100 parameter combinations are randomly generated for the eight core hyperparameters. The standard deviation of out-of-sample $R^2$ under multi-parameter joint perturbation is only 0.0092. Even when multiple parameters are adjusted simultaneously, the model performance remains stable, further confirming that hyperparameter selection does not introduce systematic bias into the conclusions.

**Scientific Reports**

# Unraveling Green Innovation Drivers in High Carbon-Lock-in Process Industries: An Interpretable Machine Learning Approach with Temporal Validation

**Hao Huang 1,2, Yujun Hu 1*, Xinqiang Li 3, Yiwei Huang 1, Wenxuan Duan 1, Bowen Zhang 1, Anxue Li 1,2, Yang Li 1**

**1 School of Economics and Management, Shandong Institute of Petroleum and Chemical Technology, Dongying, China**

**2 Key Laboratory of Digital Transformation and Application of Energy and Chemical Enterprises in Dongying City, Dongying, China**

**3 PetroChina Xinjiang Oilfield Company Luliang Oilfield Operation Area, Karamay, China**

*** Corresponding author: Yujun Hu. E-mail: hu342758@163.com. ORCID: 0009-0005-3957-5375**

## Supplementary Note

### Extended Variable Definitions and Variable Exclusion Rationale

#### 1. Extended variable definitions

The complete variable system of this study covers five dimensions with 20 indicators. Detailed definitions and data sources are presented in Table 1 of the main text. The extended computational formulas of the composite indexes are summarized below:

| Variable | Code | Extended definition / computation |
|---|---|---|
| Command-and-Control Environmental Regulation | ER | Composite index of urban pollutant emissions (industrial $SO_2$, industrial |

| | | |
|---|---|---|
| | | wastewater, industrial dust emissions), normalized and aggregated to reflect regional regulatory pressure |
| Market-Based Environmental Regulation | Sumprice | Pollutant discharge fee before 2018; environmental protection tax after 2018, as the market-based price signal of environmental regulation |
| Digitalization Level | DT | Proportion of digital technology-related items in intangible assets, as disclosed in the notes to year-end financial statements |
| Market Competition Intensity | HHI | Sum of squared market shares of firms in the industry (Herfindahl-Hirschman Index) |
| Government Subsidy | Amount | Natural logarithm of total government subsidies received by the firm in the current period, as disclosed in financial statements |

**2. Variable exclusion rationale**

Variables including CEO duality, executives political background and executives academic background show zero within-sample variation in the petrochemical listed firm sample, and thus are excluded from the formal modeling to avoid invalid estimation interference.

**3. Sample screening procedure**

First, samples of petrochemical listed companies are selected according to the industry classification standards of the China Securities Regulatory Commission. Second, firm-

year observations with excessive missing values in core variables are dropped to maintain the reliability of model estimation. Third, firms with abnormal financial status (ST, *ST and delisting risk) are excluded to avoid distorted estimation. After the above screening, a final unbalanced panel sample of 380 firm-year observations is obtained.

**Scientific Reports**

# Unraveling Green Innovation Drivers in High Carbon-Lock-in Process Industries: An Interpretable Machine Learning Approach with Temporal Validation

**Hao Huang 1,2, Yujun Hu 1*, Xinqiang Li 3, Yiwei Huang 1, Wenxuan Duan 1, Bowen Zhang 1, Anxue Li 1,2, Yang Li 1**

**1 School of Economics and Management, Shandong Institute of Petroleum and Chemical Technology, Dongying, China**

**2 Key Laboratory of Digital Transformation and Application of Energy and Chemical Enterprises in Dongying City, Dongying, China**

**3 PetroChina Xinjiang Oilfield Company Luliang Oilfield Operation Area, Karamay, China**

*** Corresponding author: Yujun Hu. E-mail: hu342758@163.com. ORCID: 0009-0005-3957-5375**

## Supplementary Note

### Detailed Stability Statistics of SHAP Feature Importance Rankings

The Kendall rank correlation coefficient (Kendall's tau) is employed to measure the consistency of the importance ranking of the top 10 core features under different hyperparameter perturbations.

#### 1. Univariate perturbation stability

Under individual perturbation of four core hyperparameters (max_depth, learning_rate, reg_alpha, reg_lambda), the Kendall's tau coefficients of the top 10 feature importance rankings are reported below, all above 0.92, indicating a very high level of ranking consistency.

| Perturbed Hyperparameter | Kendall's tau of Top-10 Ranking |
|---|---|

| | |
|---|---|
| max_depth | 0.9556 |
| learning_rate | 0.9222 |
| reg_alpha | 0.9444 |
| reg_lambda | 0.9556 |

**2. Joint perturbation stability**

Under the extreme scenario of multi-parameter combined perturbation, the average Kendall's tau coefficient of the top 10 feature importance rankings is 0.8904, and the minimum value still reaches 0.8222, maintaining strong overall ranking consistency.

**3. Conclusion**

Even when multiple sets of hyperparameters are adjusted simultaneously, the importance ranking of core features does not change substantially, and the threshold-type nonlinear patterns of firm size and R&D investment remain essentially unchanged. The SHAP attribution conclusions are therefore free from parameter-dependent contingency, and the identification results of driving factors have sufficient replicability.