applied sciences

MDPI

Article

Application of Interpretable Machine Learning for Production Feasibility Prediction of Gold Mine Project

Kun Kang [1], Qishen Chen [1,2,*], Kun Wang [2], Yanfei Zhang [2], Dehui Zhang [2], Guodong Zheng [2], Jiayun Xing [2], Tao Long [2], Xin Ren [2], Chenghong Shang [1] and Bojing Cui [1]

[1] School of Earth Sciences and Resources, China University of Geosciences (Beijing), Beijing 100083, China; kangkun1998@163.com (K.K.); shchh515@163.com (C.S.); cuibojing00@163.com (B.C.)
[2] Institute of Mineral Resources, Chinese Academy of Geological Sciences, Beijing 100037, China; wkun1111@126.com (K.W.); zhyf1014@163.com (Y.Z.); zhdehui@cugb.edu.cn (D.Z.); zguodong@cags.ac.cn (G.Z.); xingjiayun123@163.com (J.X.); longtao@cags.ac.cn (T.L.); ren_xin@126.com (X.R.)
* Correspondence: chenqishen@126.com

Abstract: In the context of globalization in the mining industry, assessing the production feasibility of mining projects by smart technology is crucial for the improvement of mining development efficiency. However, evaluating the feasibility of such projects faces significant challenges due to incomplete data and complex variables. In recent years, the development of big data technology has offered new possibilities for rapidly evaluating mining projects. This study conducts an intelligent evaluation of gold mines based on global mineral resources data to estimate whether a gold mine project can be put into production. A technical workflow is constructed, including data filling, evaluation model construction, and production feasibility evaluation. Based on the workflow, the missing data is filled in by the Miceforest imputation algorithm first. The evaluation model is established based on the Random Forest model to quantitatively predict the feasibility of the mining project being put into production, and important features of the model are extracted using Shapley Additive explanation(SHAP). This workflow may enhance the efficiency and accuracy of quantitative production feasibility evaluation for mining projects, with an accuracy rate increased from 93.80% to 95.99%. Results suggest that the features of estimated mine life and gold ore grade have the most significant impact on production feasibility.

Keywords: gold mine project; production feasibility; miceforest imputation; random forest; SHAP value

check for updates

Citation: Kang, K.; Chen, Q.; Wang, K.; Zhang, Y.; Zhang, D.; Zheng, G.; Xing, J.; Long, T.; Ren, X.; Shang, C.; et al. Application of Interpretable Machine Learning for Production Feasibility Prediction of Gold Mine Project. *Appl. Sci.* **2023**, *13*, 8992. https://doi.org/10.3390/app13158992

Academic Editors: Yosoon Choi and Jieun Baek

Received: 28 June 2023
Revised: 28 July 2023
Accepted: 2 August 2023
Published: 5 August 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the accelerating internationalization process of mining enterprises in various countries, the investment demand for mining projects is also increasing. When mining enterprises make mining investments, they should comprehensively consider production conditions and evaluate the feasibility of the project entering the production phase. By adopting objective evaluation methods, a quantitative evaluation model is established to study and evaluate whether the project is worth putting into production so as to provide scientific decision-making basis. In recent years, with the development of big data technology, domestic and foreign scholars have successfully introduced big data, machine learning, and other information technologies into the fields of geological study, quantitative prediction and evaluation of mineral resources, and made a series of research achievements [1–10].

In recent decades, the utilization of data-driven modeling techniques has resulted in notable progress in the application of mineral resources [11,12]. Probabilistic methods like weights of evidence and logistic regression have become popular due to their clear expression of models and ease of interpretation [13,14]. Machine learning (ML) methods have emerged as promising tools for generating predictive models of mineral prediction, such as artificial neural networks, support vector machines, and Random Forest (RF) [14,15]. Research has demonstrated that machine learning techniques exceed traditional statistical

methods and exploratory empirical models in terms of performance, especially when dealing with complexly distributed input evidential features or nonlinear associations with mineralization [16,17]. Random Forest has been widely used in the prediction of mineral resources prospects and has achieved stable results. Daviran (2021) used the Random Forest algorithm to explore spatial relationships to gold mines [18]. Martins (2022) utilized the Random Forest algorithm to generate two maps of mineral potential on a district scale for deposits of Cu-Au [19]. Similarly, thirty-five evidence (predictor) layers, which indicate vectors to Au mineralization, were integrated using the Random Forest algorithm [20].

The precision of predictive mineral models that rely on data-driven approaches is largely dependent on the quality of the training datasets utilized [17,21]. The construction of predictive models utilizing machine learning is faced with significant challenges due to missing data in real-world datasets. In order to model based on machine learning, missing data must typically be interpolated using various strategies. The existing data interpolation strategies can be roughly divided into (1) single interpolation (mean, median), (2) non-mice interpolation (matrix decomposition, k-nearest neighbor), (3) multiple interpolation (by chain equation), (4) interpolation by means of integrated learning (Random Forest), and (5) deep learning (generation model, automatic encoder) [21]. Many traditional methods of Missing Value Interpolation (MVI) interpolate entire datasets without considering individual test data samples and without analyzing the types and patterns of missing data to determine an appropriate imputation model [22]. On the contrary, machine learning provides non-linear regression models that can be used as an alternative to linear regression equations in multiple imputation by chained equations (MICE) for more accurate missing value predictions. Multiple imputation by chained equations models are capable of capturing variability among multiple imputation models, while ensemble learning captures variability within a single model. Combining these two methods can lead to improved estimation of missing values [22]. It was demonstrated that chain equation Multiple Imputation is the more accurate model for data imputation because it conditionally models the missing variables of other variables [23]. Miceforest is a type of multiple imputation by chained equations algorithm that employs an ensemble of decision trees to improve missing data processing for machine learning and deep learning applications [24–26]. Miceforest utilizes predictive mean matching (PMM) to iteratively update missing values by comparing the estimate to other observed values in the dataset and selecting the value of the nearest neighbor to estimate lost data. This process enables multiple imputation by chained equations algorithms to improve performance by using decision tree-based regressors to impute missing values for categorical and binary variables in datasets [27,28]. Compared to other multiple interpolations and deep learning interpolation methods, Miceforest implements machine learning chain equation interpolation, which has fast and high memory utilization characteristics and can insert missing classification and regression data without too many settings; a user can also customize every aspect of the imputation process and output a series of available diagnostic maps, providing an effective way to evaluate the results [25,27]. Recent studies have demonstrated the effectiveness of multiple imputations by chained equations algorithms in imputing missing geological datasets with continuous variables. Additionally, these methods have been successfully utilized to impute missing values in the field of bioinformation and medical big data [29,30].

In addition, machine learning modeling is usually difficult to scientifically explain results due to the characteristics of the black box. Traditional feature importance metrics only indicate which features are important in predicting outcomes but do not provide a clear understanding of how these features affect the prediction results. Shapley Additive explanation (SHAP) value is a way to address the model interpretability, based on Shapley values, to determine the importance of that individual by calculating the contribution of the individual in the cooperation [31]. SHAP has been applied in multiple fields of research and has achieved good results [32]. Wang (2022) uses the SHAP value to select the first four features of the optimal model to study how each feature affects the wastewater parameters and finally improve the effluent quality control strategy [33]. L. Baptista (2022) used a

dataset composed of hundreds of jet engines to reveal the correlation of monotony, trend, and predictability among various features [34].

Previous big data research has mainly focused on mineral potential assessment, but there has been less research on the evaluation of mining projects themselves. At present, the feasibility study of mining projects is mainly conducted through manual due diligence, which requires a lot of manpower and material resources, and focuses on qualitative analysis, lacking quantitative comparison of similar projects using big data methods. Since data from mining projects have the characteristics of high dimension and concentrated distribution of missing fields, this paper constructs a technical workflow including data filling, evaluation model construction, and production feasibility evaluation. The dataset is established by the Miceforest imputation method. An evaluation model is based on a Random Forest algorithm to predict the possibility of the project entering the production stage and finally utilizes the SHAP value method to evaluate the interpretability of the factors affecting production. The workflow, from dataset construction to rapid project evaluation, will provide a reference for gold mine production feasibility assessment. It can have a certain significance for the mining development efficiency of gold mines.

2. Methods and Materials

2.1. Methods

2.1.1. Miceforest Interpolation

Miceforest applies the chain equation to perform imputations iteratively and utilizes the Random Forest algorithm to produce imputations with the highest possible accuracy. The interpolation of chain equations "fills in" (interpolates) missing data in the data set through a series of iterative prediction models. In each iteration, each specified variable in the dataset was estimated using other variables in the dataset, running iterations repeatedly until the interpolation data converged with the original data and the interpolation ended [23,35]. We assume A = (A_1, A_2, . . ., A_p) to be a matrix with n rows and p columns, where each row represents an observation or instance, and each column represents a variable or feature. If a variable "As" has missing values at certain entries $i_{mis}^{(s)} \subseteq \{1, \ldots, n\}$. The dataset can be partitioned into four subsets:

(1) The observed values of variable As, represented as $y_{obs}^{(s)}$;
(2) The missing values of variable As, denoted by $y_{mis}^{(s)}$;
(3) The variables other than As with observations $i_{obs}^{(s)} = \{1, \ldots, n\} \backslash i_{mis}^{(s)}$ denoted by $a_{obs}^{(s)}$;
(4) The variables other than As with observations $i_{mis}^{(s)}$ denoted by $a_{mis}^{(s)}$.

The pseudo algorithm gives a representation of the Miceforest method. To start, estimate the missing values in A using mean imputation or other imputation techniques. Then, this means ordering the variables A_s, s = 1, . . ., p based on the number of missing values they have, with the variable having the least amount of missing values listed first and the one with the most missing values listed last. For each variable A_s, the missing values are imputed by first fitting an RF with response $y_{obs}^{(s)}$ and predictors $a_{obs}^{(s)}$; then, predicting the missing values $y_{mis}^{(s)}$ by applying the trained RF to $a_{mis}^{(s)}$. This suggests that the process of imputing missing values using RF is iterative and repeated until a certain stopping criterion is met.

The condition γ for stopping the process is achieved when the new imputed data matrix differs from the previous one for both continuous and categorical variables, and this difference is observed to increase for the first time. As soon as this happens, the imputation process is terminated. This ensures the accuracy and reliability of the imputed values and

avoids overfitting or underfitting the data. Here, the definition of the difference for the continuous variable set M is as follows

$$\Delta M = \frac{\sum_{j \in M} \left(A_{new}^{imp} - A_{old}^{imp} \right)^2}{\sum_{j \in M} \left(A_{old}^{imp} \right)^2} \tag{1}$$

and for the set of categorical variables P, its definition is as follows

$$\Delta P = \frac{\sum_{j \in P} \sum_{i=1}^{n} I_{A_{new}^{imp} \neq A_{old}^{imp}}}{\#NA} \tag{2}$$

the number of missing values in the categorical variables is denoted as $\#NA$.

The process of imputing Miceforest missing values is as follows (Figure 1):

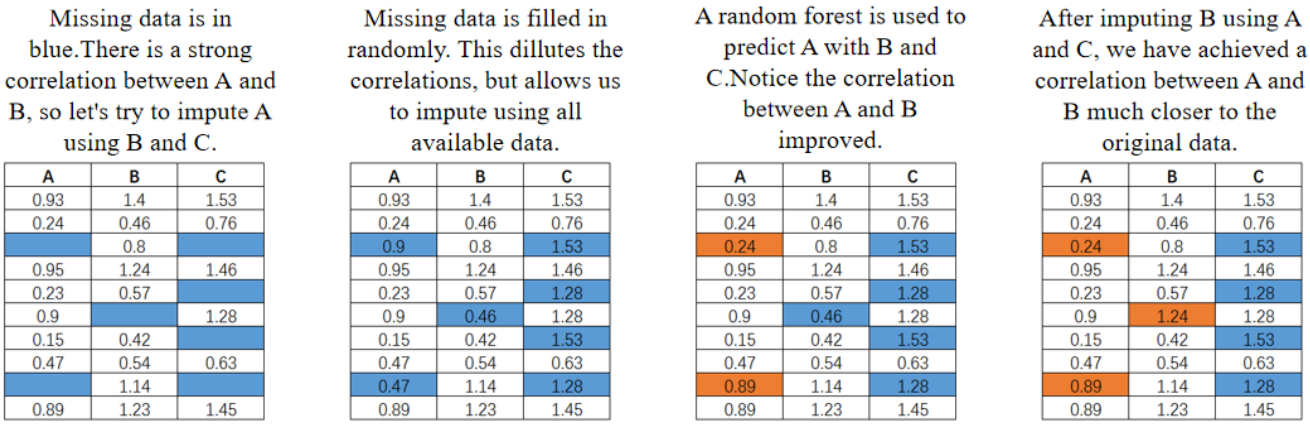


Figure 1. Schematic diagram of missing value filling process. Complete the filling process from left to right as shown in the picture. Use blue squares for missing values and orange squares for filling values.

For the miceforest process, this article used the Python package miceforest (version 5.6.2).

2.1.2. Random Forest Model

Random Forest (RF) is a classifier created by Breiman that combines multiple decision tree classifiers using the bagging method [36,37]. Random Forest combines multiple weaker decision trees, also known as weak classifiers, to improve the overall prediction accuracy. The approach takes inspiration from the Bagging method, and the so-called Bagging is:

(1) Randomly sample N training samples from the training set to create a new training set;
(2) Train M submodels on the new training set;
(3) For classification tasks, use a voting method to determine the final class prediction, which is based on the most frequent predictions amongst all submodels. For regression tasks, the predicted value is obtained through simple averaging of the submodels' predictions.

In general, integration algorithms can be broadly classified into three categories: Bagging, Boosting, and Stacking. Random Forest uses the bagging method, and the fundamental concept behind the Bagging method is to create several individual predictors, each of which makes independent predictions. The final prediction of the integration model is obtained by averaging or majority voting of the predictions made by these individual evaluators [38,39]. In this paper, the construction and evaluation of the model are realized by the use of the Random Forest algorithm. The specific process is shown in figure (Figure 2). For the Random Forest modeling process, this article used the Python package scikit-learn (version 1.0.2).

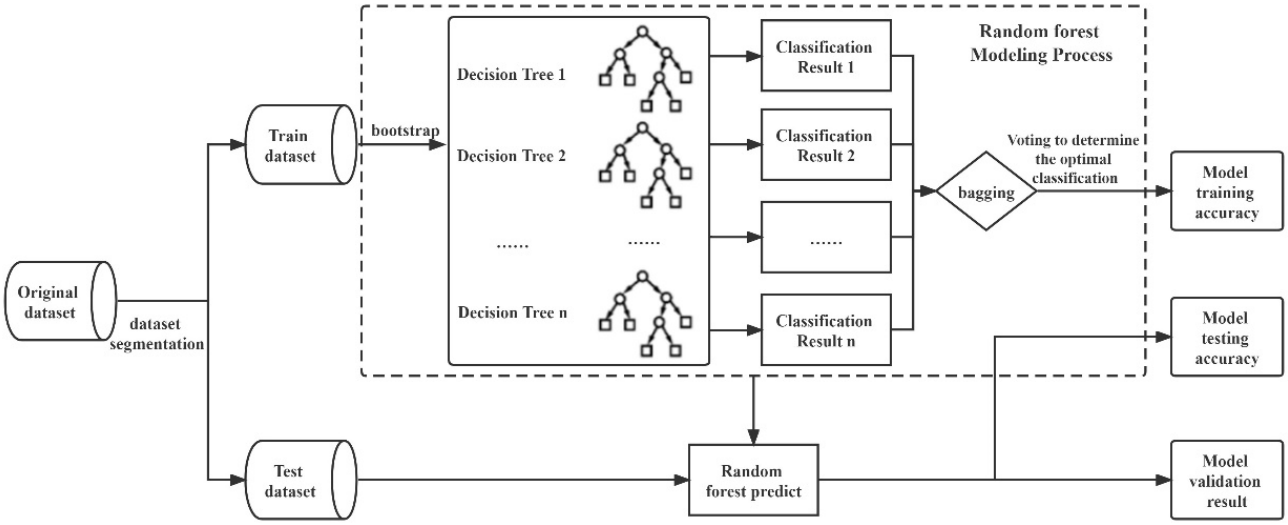


Figure 2. Flow chart of the Random Forest model construction.

2.1.3. SHAP Value Evaluation of Interpretability

Shapley Additive Explanation (SHAP) is an additive explanation model that draws inspiration from cooperative game theory and can provide detailed interpretations of the output of any machine learning model. In SHAP, all input features of a machine learning model are seen as "contributors" to the final prediction and are assessed based on their individual impact on the output [38,39]. The SHAP value is the value that is assigned to each feature in the instance, quantifying the contribution of the feature to the predicted outcome. In the case of the integrated tree model [40], used for classification tasks, the model produces a probability value as its output, which is to calculate the Shapley value of each feature. For Shapley value, its calculation method is as follows:

$$\varnothing_j = \sum_{S \subseteq (x_1,\ldots,x_M)\backslash x_j} \frac{|S|!(M-|S|-1)!}{M!}\left[f(S \cup x_j) - f(S)\right] \tag{3}$$

So as to measure the influence of the feature on the final output value formulation:

$$f(x) = g(z) = \varnothing_0 + \sum_{j=1}^{M} \varnothing_j z_j \,, \tag{4}$$

$$z \in \{0,1\}^M \tag{5}$$

In this Formulas (3)–(5), x_j represents the j-th feature of sample x, $\varnothing_j$ is the attribution value (Shapley value) of each feature, S is a subset of the feature, M denotes the number of input features, and z represents whether or not the corresponding feature exists (with 1 indicating its presence and 0 indicating its absence; such as text, after the word one-hot, all words will not appear in a sentence); $\varnothing_0$ is a constant.

For each project sample in the dataset, the SHAP model generates a result that measures the impact of the sample on the entire model. The calculation method used for SHAP values is comparable to that of linear models.

SHAP considers the i-th sample in a dataset, which we denote as x_i. Suppose that this sample has j features, with the value of the j-th feature being $x_{i,j}$. Additionally, assume that the machine learning model applied to this dataset generates a predicted value of y_i for the i-th sample. To calculate the SHAP value for this sample, SHAP starts by establishing a baseline score for the model, usually defined as the average value of the target variable across all samples and denoted as y_{base}. The SHAP value for the j-th feature

of the *i*-th sample can then be calculated using the following equation, which quantifies the contribution of that feature to the difference between the predicted value and the baseline:

$$y_i = y_{base} + f(x_{i,1}) + f(x_{i,2}) + \ldots\ldots + f(x_{i,j}) \quad (6)$$

The SHAP value for a given instance x_i and feature 1 is denoted as $f(x_{i,1})$. This value represents the extent to which the feature contributes to the final predicted value y_i for that instance. If $f(x_{i,1})$ is positive, it indicates that the feature has a positive effect on the prediction and improves its accuracy. Conversely, if $f(x_{i,1})$ is negative, the feature has a counterproductive effect on the prediction and reduces its accuracy.

The input to Tree SHAP is a tree-based machine learning model that has been trained using input data X. By utilizing this trained model along with the input data; Tree SHAP can generate an N × M matrix of SHAP values. Each value in this matrix represents the contribution of a specific feature to the prediction of a corresponding instance.

This section discusses how input instance vectors are related to their corresponding predictions in the context of black box machine learning algorithms used in prognostics modeling. Black box models do not rely on explicit causal relationships between inputs and predictions. However, the SHAP model can provide an output of decomposition factors or SHAP values that help explain the importance of individual feature values to the overall prediction (Figure 3). For SHAP value analysis, the Python package shap (version 0.41.0) is utilized in this paper.

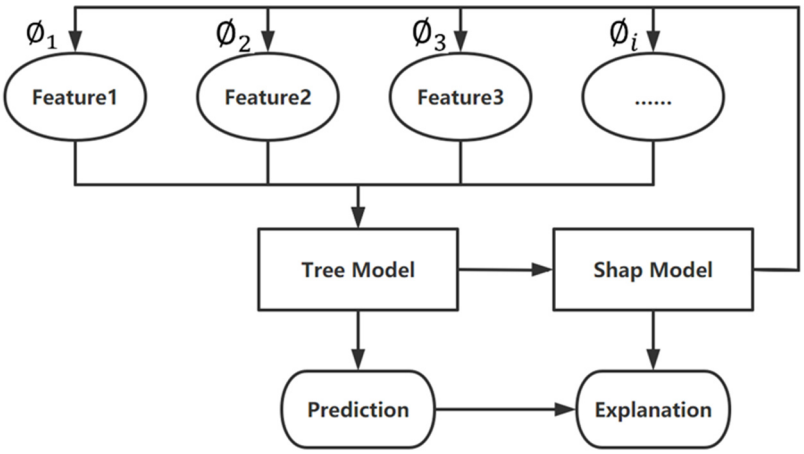


Figure 3. SHAP value model the decision flow chart.

2.2. Materials

The data used in this article is obtained from the metals and minerals sector of the global S&P Capital IQ platform. According to the official explanation provided by the platform, the development stage of projects is divided into Early Stage, Late Stage, and Mine Stage. The Early-Stage project is in the preliminary research situation, and some critical data are missing, so data from Late Stage and Mine Stage are chosen to be analyzed. The global gold mine project dataset is constructed as the source data, covering a total of 107 characteristics and a total of 4687 project samples in nine categories of information (Table 1): basic information, operator information, production information, owner information, production and reserve information, capital status, project transaction information, mine development information, and drilling information.

Table 1. Field details.

Category	Field Name	Category	Field Name
Basic information	Development Stage	Funding information	Funding Type
	Activity Status		Description
	Mine Type		Count of Capital Invested
	Count of Commodities		Capital Cost Type 1
	Country/Region		Capital Cost Type 2
	Primary Commodity		Capital Cost Type 3
	Study Year		Study Price per oz
	Country Risk Score Overall Current		LOM Cash Costs
	Country Risk Score Political Current		Cash Costs per oz
	Country Risk Score Economic Current		Mine Total Cost
	Country Risk Score Legal Current	Transaction Information	Deal Type
	Country Risk Score Tax Current		Deal Status
	Country Risk Score Operation Current		Deal Consideration
	Country Risk Score Security Current		Earn InYes/No
	Country Risk Outlook Overall		Joint VentureYes/No
	Country Risk Outlook Political		Deal Acquired (Announcement)
	Country Risk Outlook Economic		Total Deal Value (Announcement)
	Country Risk Outlook Legal		Total Deal Value (Completion)
	Country Risk Outlook Tax	Geological drilling information	Average Depth of Geologic Deposit Zone 1
	Country Risk Outlook Operation		Average Depth of Geologic Deposit Zone 2
	Country Risk Outlook Security		Average Depth of Geologic Deposit Zone 3
Operator information	Operator Market Capitalization		Average Depth of Geologic Deposit Zone 4
	Operator Total Enterprise Value		Significant Interval Yes/No
	Operator Total Debt		Ore Minerals Zone
	Operator Working Capital		Interval (meters) Drill Result 1
	Operator Total Capitalization		Interval (meters)Drill Result 2
Owner information	Count of Project Owners		Interval (meters) Drill Result 3
	Count of Project Royalty Holders		Depth (meters) Drill Result 1
	Historical Equity Ownership Percent		Depth (meters) Drill Result 2
	Historical Controlling Ownership Percent		Depth (meters) Drill Result 3
	Current Equity Ownership Percent		Exploration Purpose Drill Result 1
	Current Controlling Ownership Percent		Exploration Purpose Drill Result 2
	Owner Market Capitalization		Exploration Purpose Drill Result 3
	Owner Total Enterprise Value		Interval Value Drill Result 1
	Owner Total Debt Total Capitalization		Interval Value Drill Result 2
	Owner Working Capital		Interval Value Drill Result 3
	Total Debt		Grade x Interval Drill Result 1
	Current Liabilities		Grade x Interval Drill Result 2
	Reported EBITDA		Grade x Interval Drill Result 3
	EBITDA		Interval Grade Equivalent Drill Result 1
	Royalty Type		Interval Grade Equivalent Drill Result 2
Production information	Mill Capacity		Interval Grade Equivalent Drill Result 3
	Stripping Ratio		Max (Interval (feet))
	Waste to Ore Ratio	Resource endowments	Reserves (Ore Tonnage)
	Count of Mining Methods		Measured Indicated (Ore Tonnage Excl Reserves)
	Count of Processing Method		Inferred Resources (Ore Tonnage)
	Mining Method 1		Total Resources (Ore Tonnage Excl Reserves)
	Processing Method 1		Geologic Ore Body Type Zone
	Recovery Rate		In Situ Value (Measured Indicated Excl Reserves)
	Ore Processed Mass		Grade Reserves
	LOM Yearly Production		Contained Reserves
Mine development information	Estimated Mine Life		MillHead Grade
	Life of Mine Cash Flow (High Case)		In Situ Value (Reserves and Resources)
	Payback Period (High Case)		

Four of the 107 attribute fields in the Global Gold Project dataset are missing from the In Situ Value, MillHead Grade, Mine Total Cost, and Cash Costs. The following are the details of the missing fields (Table 2).

Table 2. Missing fields.

Vital Signs	Unit	Missing Percentage
In Situ Value (Reserves and Resources)	dollar	3.5%
MillHead Grade	g/tonne	64.2%
Mine Total Cost	dollar	71.0%
Cash Costs per oz	dollar	72.6%

3. Modeling and Results

3.1. Dataset Building

For the fields of the original data, in order to operate normally and perform better in the Random Forest model, it is necessary to formulate dataset feature engineering rules under the premise of meeting the requirements of rapid evaluation (Table 3). When processing single text type fields, such as "Royalty Type", this article performed dummy variable processing based on attribute classification. For multitext type fields such as "Deal Type", we count based on the number of text per value. As to discrete type (no size meaning) fields, such as "Owner Working Capital", we performed one-hot encoding processing on them. When dealing with discrete type (size meaning) fields such as "Country Risk Score Overall Current", it performed value mapping processing on them.

Table 3. Data processing rules.

Data Type	Method
Single text type	Dumb variable treatment is performed according to classification
Multitext type	Numeric mapping is performed after counting
Discrete type (no size meaning)	One-hot encoding
Discrete type (size meaning)	Value mapping
Continuous	No processing
Special fields	Attribute aggregation, addition, multiple overlay

The following describes how to extract other special field features.

(1) Geologic Ore Body Type Zone

For the field, Geologic Ore Body Type Zone is classified according to the gold ore genesis aggregation and dumb variable treatment (Table 4).

Table 4. Classification of gold mine types [41,42].

Geologic Ore	Type
Saddle Reefs	Orogenic
Mesothermal Lode Gold	Orogenic
Vein Hosted	Orogenic
Intrusive Related	Intrusive Related
Disseminated	Intrusive Related
Alkali Intrusion	Intrusive Related
Granite Related	Granite Related
Layered Mafic-Ultramafic Intrusion	Intrusive Related
Skarn (Metasomatic)	Skarn
Carbonate Replacement (incl Manto)	Skarn
IOCG Breccia Complex	IOCG
Iron Oxide Copper Gold (IOCG)	IOCG
Replacement	Replacement
Proterozoic Quartz Pebble Conglomerate	Placer
Paleoplacer (Buried)	Placer
Placer (Alluvial)	Placer
Placer (Beach)	Placer

Table 4. *Cont.*

Geologic Ore	Type
Jasperoid Hosted	Epithermal
Epithermal	Epithermal
Epithermal Low Sulphidation	Epithermal
Epithermal High Sulphidation	Epithermal
Hot Spring Au-Ag	Epithermal
Komatiitic Magmatic	Komatiitic Magmatic
Carlin Style Carbonate Replacement	Carlin
Flood Basalt (Dyke-Sill Complexes)	Flood Basalt
Laterite (Generic)	Laterite
Black Shale	Black Shale
Breccia Pipes	Porphyry
Collapse Breccia Pipes	Porphyry
Breccia Fill	Porphyry
Porphyry Deposit	Porphyry
Sedimentary Exhalative (SEDEX)	Sediment
Sediment Hosted (Reduced Facies)	Sediment
Supergene	Supergene
Volcanogenic Massive Sulfide (VMS)	VMS
Carb-Hosted (Mississippi Valley Type)	MVT
Banded Iron Formation (BIF)	BIF

(2) Max (Interval (feet))

Since the drilling footage data are not updated every year, if the data of a single year are taken, there will be many null values, so the five-year data of Interval (feet) from 2018 to 2022 is taken as the maximum value to generate a new field Max (Interval (feet)).

(3) Development Stage label division

According to the official interpretation of the Development Stage field (Table 5), gold mining projects are classified into "already put into production" status (Mine-Stage, Commissioning, Production, Operating, Satellite, Limited, Preproduction) and "not put into production" status (Late-Stage, Reserves Development, Advanced, Exploration, Prefeasibility/Scoping, Feasibility, Feasibility Started, Feasibility Complete, Construction Planncd, Construction Started, Expansion, Residual, Closed).

Therefore, 4689 projects were selected to constitute the original dataset of global gold mining projects, of which 1330 projects were put into production, and 3359 projects were not put into production. According to the average time limit of 7 years from the feasibility stage to the production stage of the gold mine, we take the projects that have been put into production as positive samples and the projects that have been researched but not put into production before 2015 and 2015 as negative samples. Then, randomly combine the positive sample project data and negative sample project data to form a training dataset and a verification dataset, a total of 911 projects.

3.2. Modeling Process

To begin with, extract a global gold mine projects dataset containing 107 characteristics from nine aspects, including production, operation, and trading of the project;

Secondly, the miceforest imputation method was utilized to simulate and fill in the missing data of global gold mines;

Thirdly, the Random Forest model was utilized to evaluate the feasibility of putting into production and verify the accuracy of the filled dataset;

Finally, use SHAP Value to extract the index features with high contribution to the model and analyze the extracted features.

The specific implementation process is shown in the following figure (Figure 4).

Table 5. S&P Capital IQ official explanation of the Development Stage.

Label	Development Stage	Description
Already put into production	Mine-Stage	Project that has made a decision to move forward with production.
	Commissioning	The mine is commissioned and production has started.
	Production	Commercial production has been achieved.
	Operating	The mine is fully operational.
	Satellite	A satellite of a central processing complex.
	Limited	Some ore and/or commodity is being produced. Closure.
	Preproduction	Ago-ahead decision has been made and the project is being readied for production.
Not put into production	Late-Stage	Project with a defined resource that has not yet reached a production decision.
	Reserves Development	An initial reserve/resource has been calculated.
	Advanced	Drilling is being completed to add additional reserves/resources.
	Exploration	Project is in the exploration phase.
	Prefeasibility/Scoping	Usually an in-house assessment that includes mining and proccssing methods, capital costs, NPV, IRR.
	Feasibility	Bankable feasibility study is underway to determine economic viability.
	Feasibility Started	Feasibility report has commenced.
	Feasibility Complete	Feasibility report is complete.
	Construction Planned	Construction is planned for the property
	Construction Started	Construction has begun at the property
	Expansion	Operator is engaged in an active capital expansion of the facilities.
	Residual	The operator has stopped mining ore and is leaching the residual ore.
	Closed	Operation has stopped, in many cases because the ore has been exhausted.

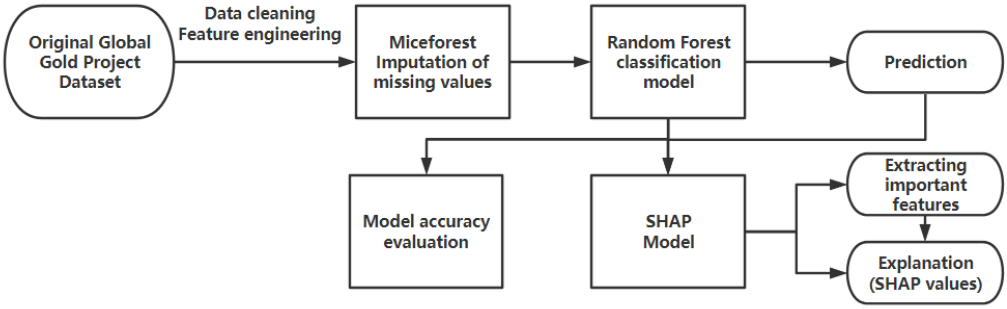


Figure 4. Technical route.

3.3. Results and Accuracy Verification

After data cleaning and feature engineering of the original data, the miceforest method was used to use 30 rounds of imputation iterations for the missing data. After the graphical diagnosis, it can be seen that the mean of the four variables with missing values gradually converges after multiple imputations (Figure 5). In the distribution line plot, it can be observed that the original data after imputation fit well with the distribution of the imputed data (Figure 6).

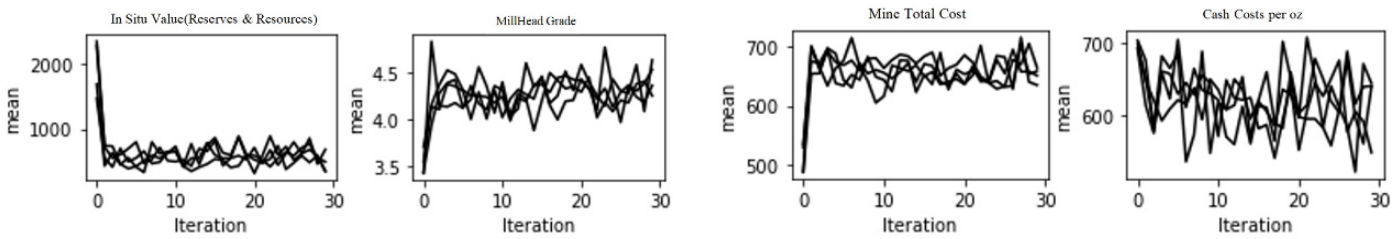


Figure 5. Convergence results during variable interpolation.

By randomly splitting the dataset into training and validation sets in a 7:3 ratio, 638 (70%) samples were randomly selected from the dataset of the gold mining project as the training set, and the remaining 273 (30%) samples were used as the validation set to

construct the Random Forest model. After modeling and ten-fold cross-validation of the model, the model training accuracy was 95.92%, and the model test accuracy was 93.80%. While the model training accuracy was 97.02%, and the model inspection accuracy was 95.99% after training modeling on the miceforest imputation dataset. Compared with the original data, the model accuracy of the filled dataset is improved by 2.19%. The true positive rate of the model has also increased from 96.07% to 97.80%. The ROC (Receiver Operating Characteristic Curve) is used to evaluate model accuracy. Obviously, we can see that the model using imputed data has higher accuracy (Figure 7) than the model using original missing value data (Figure 8).

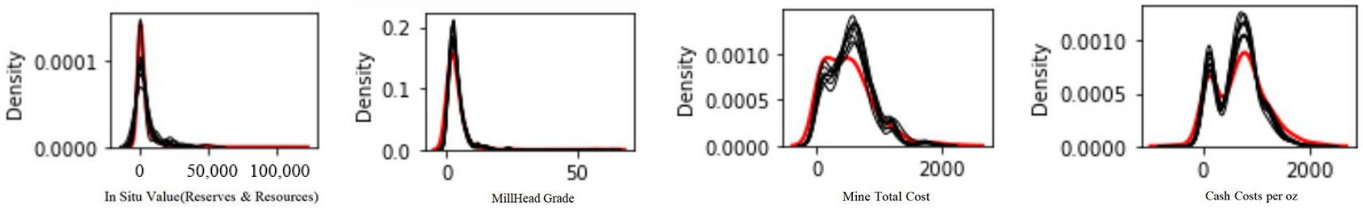


Figure 6. Data distribution and raw data distribution after missing values are filled, the red line represents the original data, while the black line depicts the imputed value for each dataset.

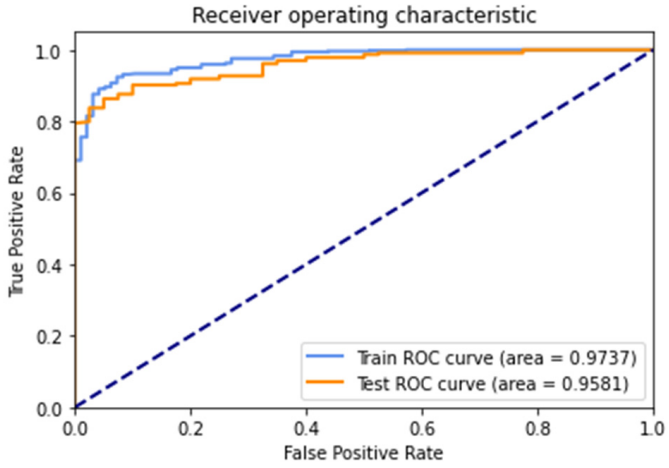


Figure 7. Random Forest modeling ROC without filling with missing values. This blue dotted line represents a random guess line, representing the performance level of the model in random prediction situations.

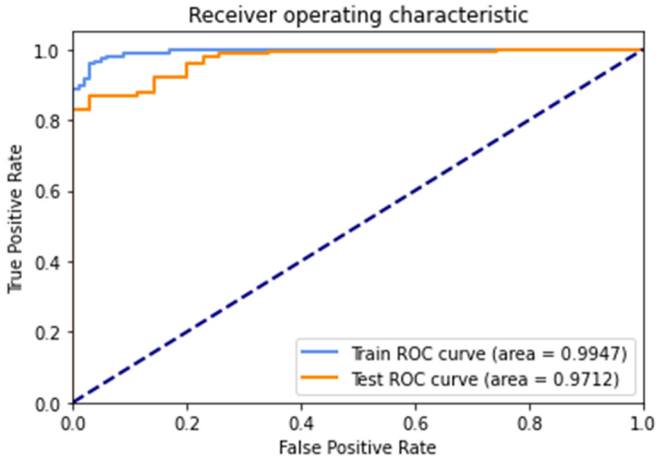


Figure 8. Missing values are filled in case of Random Forest modeling ROC. This blue dotted line represents a random guess line, representing the performance level of the model in random prediction situations.

The accuracy of the predictive model validation set is evaluated by constructing a confusion matrix visualization (Figure 9). The Random Forest algorithm model constructed by miceforest's imputed dataset correctly retrieved 220 of the 228 investable projects. The "recall rate" reached 96% (220/228), and among the 223 investable mining projects, three mining projects were misjudged, and the accuracy rate was 98.6% (220/223). That is to say, 1.4% of all investable mining projects will be misjudged as positive samples, and theoretically, according to this model, to study whether to put it into production, the failure rate is 1.4%.

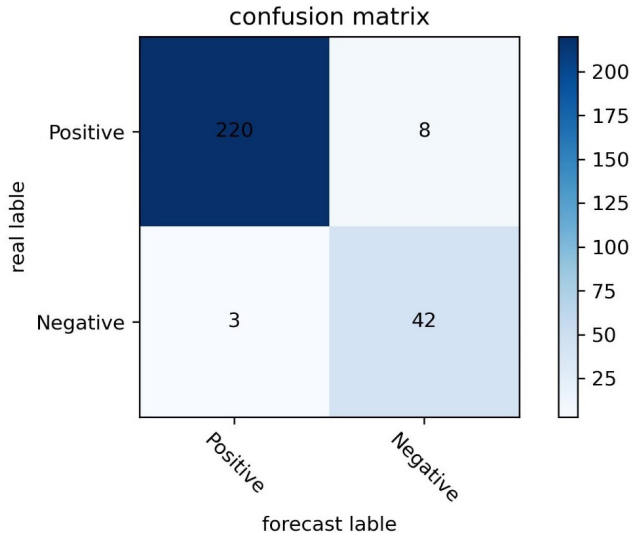


Figure 9. Confusion matrix. Positive samples represent the possibility of entering production; Negative samples represent that projects cannot be put into production.

4. Discussion

4.1. Feature Importance Analysis by SHAP Value Summary Plot

To investigate the positive and negative relationships between each feature's impact on the model's final output, the SHAP value method is used in this paper. As shown in the figure (Figure 10), each row in the figure corresponds to a feature, with the *x*-axis representing its SHAP value and each point signifying a sample. A redder color indicates a higher feature SHAP value, while bluer shades denote lower feature values. A wider range indicates that it contains more samples, a longer area range indicates that the feature has a greater impact on the model, the more left samples have a negative effect on the model, and the more right samples have a positive impact on the model. According to the average absolute value of SHAP, the ranking of features in importance from highest to lowest is descending. For example, we intuitively see that Estimated Mine Life is the most important feature affecting whether it can be put into production. The larger the value, the higher the probability of being put into production, and the amplitude of this feature has the greatest impact on the model, indicating that the life cycle of a gold mine is the most worthy factor when considering gold mine production feasibility; The smaller the LOM Cash Costs value, the greater the likelihood that it will be put into production, that is, the lower the cost of cash invested.

4.2. Feature Correlation Analysis by Partial Dependence Plot

A Partial Dependence Plot (PDP or PD plot) illustrates the effect of one or two features on the predicted outcome of a machine learning model. It can demonstrate how individual features impact model predictions. The *x*-axis on the plot represents the actual value in the dataset, while the *y*-axis shows the SHAP value of the feature. This value illustrates the extent to which the feature value affects the predicted model output. The plot's color corresponds to another feature that could interact with the feature being visualized. As shown in the figure, taking Estimated Mine Life as an example, when Estimated Mine Life

is low, when it is 0–5 years, the SHAP value will experience a notable increase, and the probability that the project can be put into production will be greatly increased (Figure 11a). Similarly, when the Stripping Ratio is lower than 3%, the SHAP value will experience a notable increase, and the probability that the project can be put into production is greatly increased (Figure 11c).

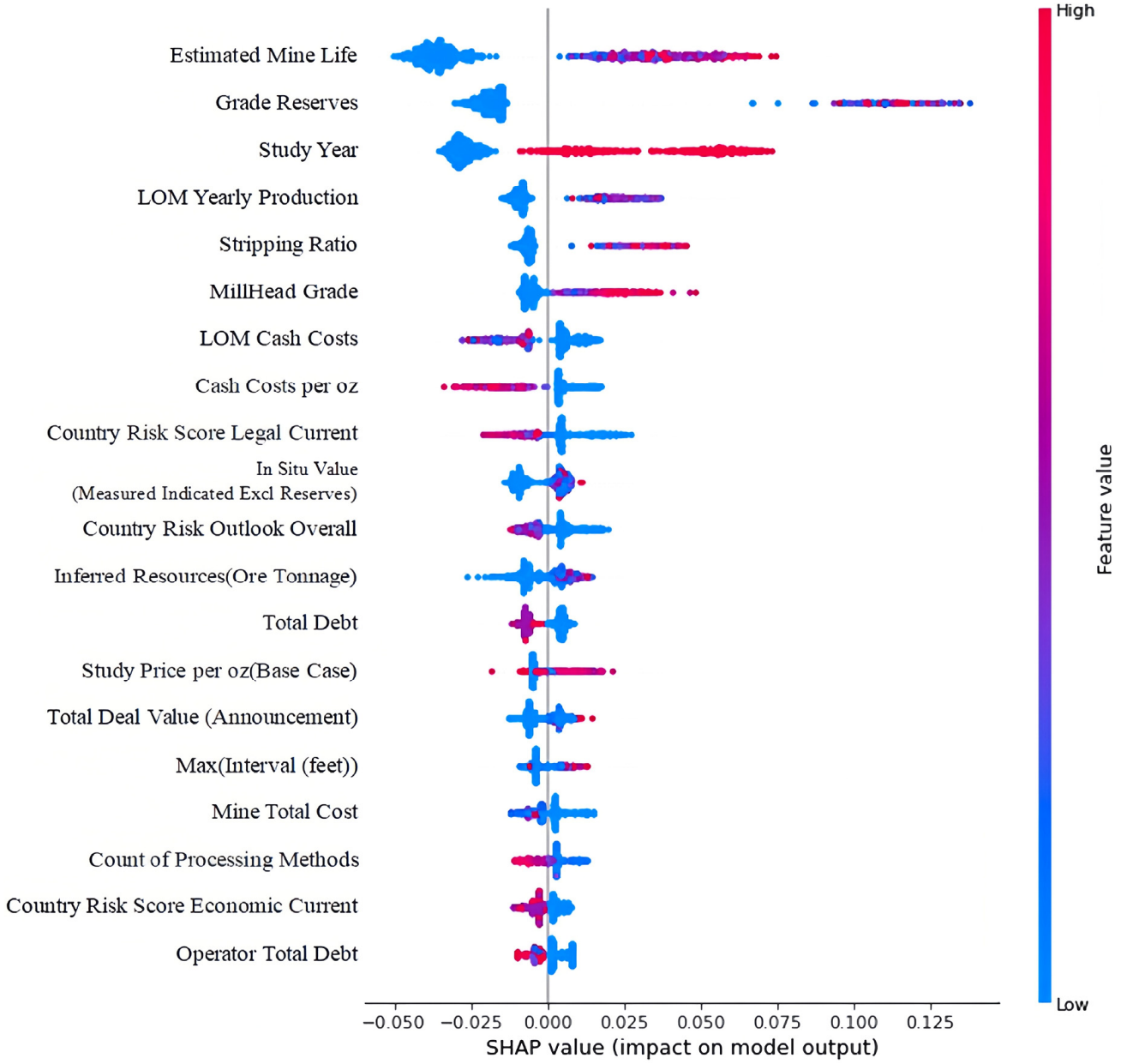


Figure 10. Ordering of all features in the model(SHAP value summary plot).

With the increase in Estimated Mine Life, the probability of the model being judged to be ready for production increased rapidly, and the positive impact of Estimated Mine Life on the model remained at a high level until the fifth year (Figure 11b). In the figure's top section (SHAP value > 0.03), the proportion of blue dots is higher among all points. This suggests that for a sample with low Cash Costs per oz, Estimated Mine Life has a greater impact on predicting whether a project will be able to go into production. In Figure 11d, most of the points in the figure are concentrated in the Stripping Ratio between 2% and 5%, and in the upper half of the graph (SHAP value > 0.04), the proportion of red dots is higher among all points, indicating that for the sample with a higher MillHead Grade, the

Stripping Ratio has a greater impact on predicting whether the project will be able to go into production.

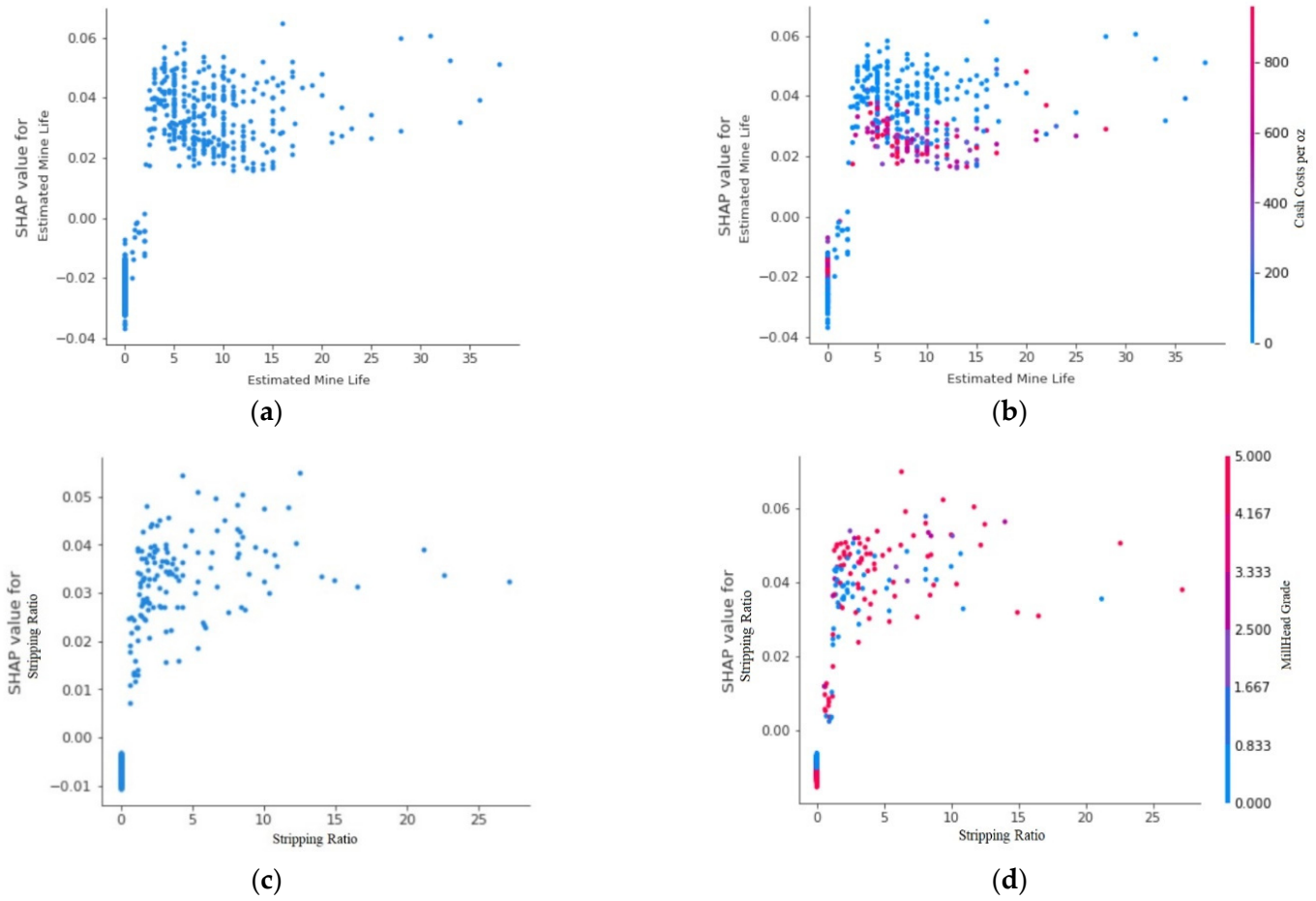


Figure 11. Distribution and correlation of important features (SHAP value dependence plot). (**a**) the distribution of Estimated Mine Life; (**b**) the correlation between Estimated Mine Life and Cash Costs per oz; (**c**) the distribution of Stripping Ratio; (**d**) the correlation between Stripping Ratio and MillHead Grade.

4.3. Project Validation by SHAP Value Force Plot

In order to verify the feasibility of the model, this study selects two projects (projects A and B) from the projects that have been researched but not put into production in the past five years for verification and verification. SHAP value force plot gives us interpretability for single model predictions that can be used for error analysis to find explanations for specific instance predictions. In the plot, the significance of features can be demonstrated by the length of the bar in the SHAP values. Longer bars mean higher SHAP values, and variables are more important. Among them, features that will increase the SHAP value (resulting in production) are displayed in red, and features that will decrease the SHAP value (resulting in non-production) are displayed in blue.

For instance, Figure 12a shows that due to the short Estimated Mine Life, Low In Situ Value (Measured Indicated Excl Reserves), and High Country Risk Score Legal Current, even if the Reserve (Ore Tonnage) is high, the model estimates that the project is more likely to be non-productive (probability f(x) = 0.47).In Figure 12b, the project is forecast to be production-ready. Although the project has two risk characteristics: low Stripping Ratio (Stripping Ratio = 1.75) and Low In Situ Value (Measured Indicated Excl Reserves), its Estimated Mine Life, Study Year, and MillHead Grade are all at a high level, which puts the project at a production-ready level.

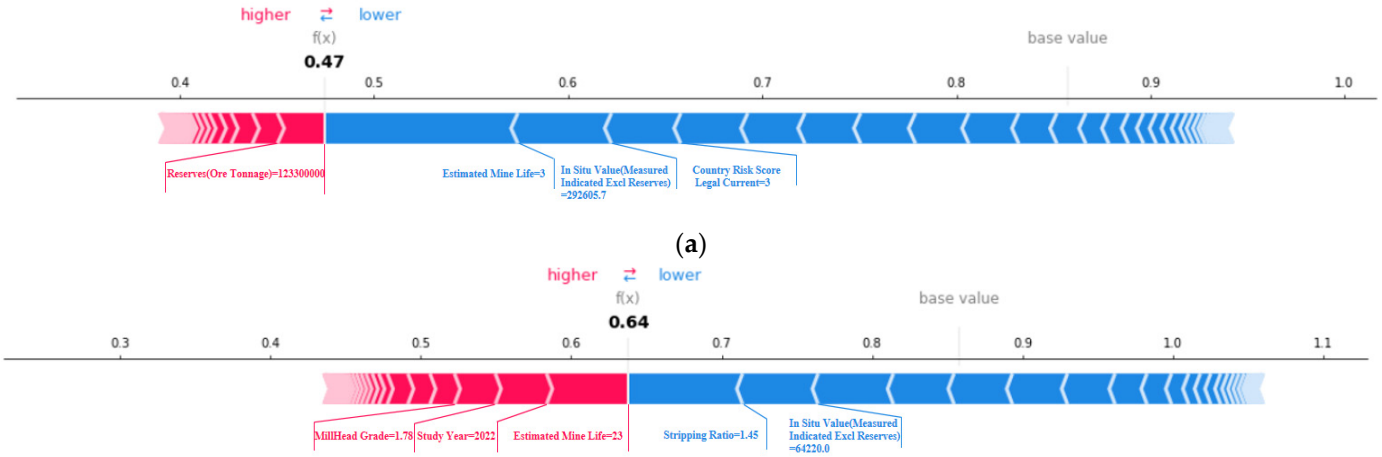


Figure 12. Verification project (SHAP value force plot). (**a**) analysis of the possibility of project A's production and its influencing factors; (**b**) analysis of the possibility of project B's production and its influencing factors.

5. Conclusions

In this paper, a technical workflow is constructed to predict the production feasibility of the mining project and whether it will be put into the production stage according to the characteristics of the high feature dimension of the global gold mine project data dataset and the concentration of missing fields in the project data. The global gold mine project dataset is taken as the research object, and the workflow is completed based on miceforest imputation and Random Forest algorithm to complete, including the construction of an index system, missing value imputation, model construction, and evaluation. Through the modeling and analysis of existing production projects in the world, the important influencing factors that determine whether gold mines are put into production are obtained, and the empirical analysis of actual projects is given. Combined with examples in the geological field, the algorithm process of imputation and interpretability evaluation is derived in detail, which provides theoretical support for geologists to deeply understand the mathematical principles of the miceforest imputation algorithm and SHAP value. Through the research of this paper, the conclusions are mainly as follows:

(1) The accuracy rate in the prediction of the validation dataset imputed by miceforest was obviously improved. After constructing a Random Forest model on the dataset imputed by miceforest, the accuracy rate in the prediction was improved from 93.80% to 95.99%.

(2) The Random Forest algorithm was used to construct the production feasibility model of the global gold mine project. It is verified by the accuracy, recall rate, and false positive rate that the model has high training accuracy and inspection accuracy, which proved the accuracy of the model construction, and there was no obvious overfitting scenario.

(3) The estimated production life and ore deposit grade can be obtained in the feature importance ranking results of the SHAP value algorithm model as the most important factors affecting the production of gold mining projects.

(4) This paper proposes a workflow for the evaluation of global gold mining projects' production feasibility, which achieves satisfactory application results and provides new ideas for mining project assessment research and mining development efficiency in the era of big data.

Author Contributions: Conceptualization, K.K., Q.C. and K.W.; funding acquisition, Q.C.; methodology, K.K., K.W. and Y.Z.; visualization, K.K., X.R., C.S. and B.C.; writing—original draft, K.K., D.Z., G.Z. and J.X.; writing—review and editing, K.K. and T.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Chinese Academy of Engineering strategic research and consulting project (Grant 52922023002); the National Natural Science Foundation of China (Grants 42271281; 92062111); the China Geological Survey Program (Grants DD20230040; DD20221694; DD20211405; DD20190674).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: 3rd Party Data. Restrictions apply to the availability of these data. Data was obtained from S&P Global and are available [from the URL: https://www.capitaliq.spglobal.cn, accessed on 1 March 2023)] with the permission of S&P Global.

Acknowledgments: We are grateful to the National Natural Science Foundation of China and the China Geological Survey for their financial support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zuo, R.; Xiong, Y.; Wang, J.; Carranza, E.J.M. Deep learning and its application in geochemical mapping. *Earth Sci. Rev.* **2019**, *192*, 1–14. [CrossRef]
2. Xiong, Y.; Zuo, R. Recognition of geochemical anomalies using a deep autoencoder network. *Comput. Geosci.* **2016**, *86*, 75–82. [CrossRef]
3. Zaki, M.M.; Chen, S.; Jicheng, Z.; Feng, F.; Qi, L.; Mahdy, M.A.; Jin, L. Optimized Weighted Ensemble Approach for Enhancing Gold Mineralization Prediction. *Appl. Sci.* **2023**, *13*, 7622. [CrossRef]
4. Qi, C. Big data management in the mining industry. *Int. J. Miner. Metall. Mater.* **2020**, *27*, 131–139. [CrossRef]
5. Li, W.L.; Gao, S.Y.; Han, C.H.; Wei, G.H.; Song, X.; Yang, J.K. A brief analysis on data mining for deep-sea mineral resources based on big data. *Procedia Comput. Sci.* **2019**, *154*, 699–705. [CrossRef]
6. Yu, P.; Chen, J.; Chai, F.; Zheng, X.; Yu, M.; Xu, B. Research on model-driven quantitative prediction and evaluation of mineral resources based on geological big data concept. *Geol. Bull. China* **2015**, *34*, 1333–1343.
7. Chen, Q.; Yu, W.; Zhang, Y.; Tan, H. Resources-Industry 'flying geese' evolving pattern. *Resour. Sci.* **2015**, *37*, 871–882. (In Chinese)
8. Chen, Q.; Yu, W.; Zhang, Y.; Tan, H. Mining development cycle theory and development trends in Chinese mining. *Resour. Sci.* **2015**, *37*, 891–899. (In Chinese)
9. Chen, Q.; Zhang, Y.; Xing, J.; Long, T.; Zheng, G.; Wang, K.; Cui, B.; Qin, S. Methods of Strategic Mineral Resources Determination in China and Abroad. *Acta Geosci. Sin.* **2021**, *42*, 137–144. (In Chinese)
10. Wang, K.; Chen, Q.; Zhang, Y.; Wang, F.; Xing, J.; Zheng, G.; Long, T.; Zhang, T.; Cui, B. A Discussion on a Comprehensive Evaluation Method for Overseas Copper Mine Investment Projects: A Case Study of Africa. *Acta Geosci. Sin.* **2021**, *42*, 229–235. (In Chinese)
11. Li, B.; Liu, B.; Guo, K.; Li, C.; Wang, B. Application of a maximum entropy model for mineral prospectivity maps. *Minerals* **2019**, *9*, 556. [CrossRef]
12. Li, X.; Yuan, F.; Zhang, M.; Jia, C.; Jowitt, S.M.; Ord, A.; Zheng, T.; Hu, X.; Li, Y. Three-dimensional mineral prospectivity modeling for targeting of concealed mineralization within the Zhonggu iron orefield, Ningwu Basin, China. *Ore Geol. Rev.* **2015**, *71*, 633–654. [CrossRef]
13. Porwal, A.; Carranza, E.J.M. *Introduction to the Special Issue: GIS-Based Mineral Potential Modelling and Geological Data Analyses for Mineral Exploration*; Elsevier: Amsterdam, The Netherlands, 2015; Volume 71, pp. 477–483.
14. Zuo, R. Machine learning of mineralization-related geochemical anomalies: A review of potential methods. *Nat. Resour. Res.* **2017**, *26*, 457–464. [CrossRef]
15. Wang, K.; Ai, Z.; Zhao, W.; Fu, Q.; Zhou, A. A Hybrid Model for Predicting Low Oxygen in the Return Air Corner of Shallow Coal Seams Using Random Forests and Genetic Algorithm. *Appl. Sci.* **2023**, *13*, 2538. [CrossRef]
16. Elahi, F.; Muhammad, K.; Din, S.U.; Khan, M.F.A.; Bashir, S.; Hanif, M. Lithological Mapping of Kohat Basin in Pakistan Using Multispectral Remote Sensing Data: A Comparison of Support Vector Machine (SVM) and Artificial Neural Network (ANN). *Appl. Sci.* **2022**, *12*, 12147. [CrossRef]
17. Xi, N.; Yang, Q.; Sun, Y.; Mei, G. Machine Learning Approaches for Slope Deformation Prediction Based on Monitored Time-Series Displacement Data: A Comparative Investigation. *Appl. Sci.* **2023**, *13*, 4677. [CrossRef]
18. Daviran, M.; Maghsoudi, A.; Ghezelbash, R.; Pradhan, B. A new strategy for spatial predictive mapping of mineral prospectivity: Automated hyperparameter tuning of Random Forest approach. *Comput. Geosci.* **2021**, *148*, 104688. [CrossRef]

19. Martins, T.F.; Seoane, J.C.S.; Tavares, F.M. Cu–Au exploration target generation in the eastern Carajás Mineral Province using Random Forest and multi-class index overlay mapping. *J. S. Am. Earth Sci.* **2022**, *116*, 103790. [CrossRef]
20. Harris, J.R.; Naghizadeh, M.; Behnia, P.; Mathieu, L. Data-driven gold potential maps for the Chibougamau area, Abitibi greenstone belt, Canada. *Ore Geol. Rev.* **2022**, *150*, 105176. [CrossRef]
21. Leke, C.; Marwala, T.; Paul, S. Proposition of a theoretical model for missing data imputation using deep learning and evolutionary algorithms. *arXiv* **2015**, arXiv:1512.01362.
22. Valdiviezo, H.C.; Van Aelst, S. Tree-based prediction on incomplete data using imputation or surrogate decisions. *Inf. Sci.* **2015**, *311*, 163–181. [CrossRef]
23. Van Buuren, S. *Flexible Imputation of Missing Data*; CRC Press: Boca Raton, FL, USA, 2018; pp. 87–126.
24. Xu, D.; Sheng, J.Q.; Hu, P.J.; Huang, T.; Hsu, C. A deep learning–based unsupervised method to impute missing values in patient records for improved management of cardiovascular patients. *IEEE J. Biomed. Health* **2020**, *25*, 2260–2272. [CrossRef] [PubMed]
25. Zhao, X.; Shen, W.; Wang, G. Early prediction of sepsis based on machine learning algorithm. *Comput. Intell. Neurosc.* **2021**, *2021*, 6522633. [CrossRef] [PubMed]
26. Stekhoven, D.J.; Bühlmann, P. MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics* **2012**, *28*, 112–118. [CrossRef]
27. Akande, O.; Li, F.; Reiter, J. An empirical comparison of multiple imputation methods for categorical data. *Am. Stat.* **2017**, *71*, 162–170. [CrossRef]
28. Li, L.; Prato, C.G.; Wang, Y. Ranking contributors to traffic crashes on mountainous freeways from an incomplete dataset: A sequential approach of multivariate imputation by chained equations and Random Forest classifier. *Accid. Anal. Prev.* **2020**, *146*, 105744. [CrossRef]
29. Slade, E.; Naylor, M.G. A fair comparison of tree-based and parametric methods in multiple imputation by chained equations. *Stat. Med.* **2020**, *39*, 1156–1166. [CrossRef]
30. Resche-Rigon, M.; White, I.R. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Stat. Methods Med. Res.* **2018**, *27*, 1634–1649. [CrossRef]
31. Lundberg, S.M.; Lee, S. A unified approach to interpreting model predictions. In Proceedings of the Advances in Neural information Processing Systems 2017, Long Beach, CA, USA, 25 November 2017.
32. Liu, Y.; Liu, Z.; Luo, X.; Zhao, H. Diagnosis of Parkinson's disease based on SHAP value feature selection. *Biocybern. Biomed. Eng.* **2022**, *42*, 856–869. [CrossRef]
33. Wang, D.; Thunéll, S.; Lindberg, U.; Jiang, L.; Trygg, J.; Tysklind, M. Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods. *J. Environ. Manag.* **2022**, *301*, 113941. [CrossRef]
34. Baptista, M.L.; Goebel, K.; Henriques, E.M.P. Relation between prognostics predictor evaluation metrics and local interpretability SHAP values. *Artif. Intell.* **2022**, *306*, 103667. [CrossRef]
35. Samad, M.D.; Yin, L. Non-linear regression models for imputing longitudinal missing data. In Proceedings of the 2019 IEEE International Conference on Healthcare Informatics, Xi'an, China, 1 June 2019.
36. Breiman, L. Using iterated bagging to debias regressions. *Mach. Learn.* **2001**, *45*, 261–277. [CrossRef]
37. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
38. Butnariu, D.; Kroupa, T. Shapley mappings and the cumulative value for n-person games with fuzzy coalitions. *Eur. J. Oper. Res.* **2008**, *186*, 288–299. [CrossRef]
39. Štrumbelj, E.; Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.* **2014**, *41*, 647–665. [CrossRef]
40. Lundberg, S.M.; Erion, G.G.; Lee, S. Consistent individualized feature attribution for tree ensembles. *arXiv* **2018**, arXiv:1802.03888.
41. Brown, P.E.; Hagemann, S.G. MacFlinCor and its application to fluids in Archean lode-gold deposits. *Geochim. Cosmochim. Acta* **1995**, *59*, 3943–3952. [CrossRef]
42. Groves, D.I.; Goldfarb, R.J.; Santosh, M. The conjunction of factors that lead to formation of giant gold provinces and deposits in non-arc settings. *Geosci. Front.* **2016**, *7*, 303–314. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.