NATIONAL RESEARCH UNIVERSITY

HIGHER SCHOOL OF ECONOMICS

Graduate School of Business

DECISION ANALYSIS

GROUP PROJECT

"Analysis of the metal recovery of gold-bearing ores and prediction of the gold recovery rate"

Students:

Panasenkov Vladimir

Elena Tkachenko

Piskaeva Zlata

Yuanyuan Yu

Moscow 2022

Table of contents.

Abstract

With the increasing demand for gold, the massive development and utilization of mineral resources has made the mineral resources increasingly depleted. The resource recovery of gold-bearing ores can alleviate the shortage of gold resources, reduce waste and improve resource utilization. And many of the metal resources contained in gold-bearing ores also have great economic value. Therefore, it is necessary to analyze the metal recovery of gold-bearing ores and predict the gold recovery rate in order to provide guidance for production optimization.

In this paper, a prototype machine learning model is used to predict the gold recovery rate of gold-bearing ores. Firstly, a selection of data from the training dataset is used for preliminary data analysis. In the first step, the correctness of the enrichment efficiency calculation is checked by determining the MAE value corresponding to the measurement error of the enrichment efficiency. In the second step, the concentration of Au, Ag and Pb at various stages of purification process were analyzed to provide a preliminary analysis of the extraction of metals from gold-bearing ores. In the third step, the size distribution of the raw material granules in the training sample was compared with the test sample and most of the particles were found to be in the range of 40 - 100 μm in size. In the fourth step, the total concentration of all metals at different stages was analyzed.

Next, a model was built for prediction. Some of the data from the training samples were selected as influencing factors and resulting indicators and the defined function sMAPE was used as a metric. Linear regression, decision tree regression and random forest regression were used for prediction and cross-validation was used for model correction. The random forest regression model was found to be better and predicted a smape of 10.57% for the test sample, while dummy regression was used to obtain a smape of 9.44% for the test sample.

Finally, based on the above data analysis and model predictions, we were able to obtain the concentration, particle size and gold recovery rate of the recovered metal in the gold-bearing ore at each stage, analyze the reasons for the differences in the concentration of the recovered metal at each stage and draw relevant conclusions based on the predicted gold recovery rate.

Keywords: gold recovery; predictive analysis; machine learning

Chapter 1. Introduction

1.1 Background of the research

Gold has always been one of the most popular metals in the world. The crazy gold rush makes it sought after by various gold mining companies, mineral experts, and amateur gold prospectors. As shallow, high-grade gold deposits continue to be mined and utilised, the easy-to-mine resources are becoming depleted, so companies have turned their attention to recovering gold from gold-bearing ores rather than mining them further.

And with the opening up of the gold market, prices are becoming more and more transparent and profit margins are getting lower and lower for gold refining companies, which requires us to refine companies from management and technology to be effective. Refining methods are applied in different enterprises, the results are not the same, the same refining methods equipment and equipment level of different, the results are not the same, the recovery rate has high and low. Compared to the cost, the recovery rate of the high and low directly determines the effectiveness of the enterprise. Therefore, it is imperative to improve the gold recovery rate.

1.2 Significance and purpose of the research

1.2.1 Significance of the research

The gold recovery rate is closely related to the efficiency of gold refining enterprises and has become an important indicator for evaluating the productivity of enterprises, which is important for assessing the utilization of production resources and thus optimizing the production process.

Based on a statistical perspective, this paper uses four algorithms: linear regression, decision tree regression, random forest regression and dummy regression to construct a predictive model based on machine learning. It provides a basis for further optimisation of production planning and strategy development and assessment of production efficiency and effectiveness, and is of great practical importance.

1.2.2 Purpose of the research

In this paper, we propose to use four algorithms: linear regression, decision tree regression, random forest regression and dummy regression to construct a prediction model to predict the recovery rate of the most important product, gold, by selecting relevant index data, followed by some data processing based on machine learning, to describe and analyze the concentration and particle size of gold, silver and lead recovered from gold-bearing ores and to assess the current production resource utilization and thus optimize production efficiency. to assess the current production resource utilization and thus optimize production efficiency.

1.3 Research Methodology

1.3.1 Linear regression

Linear regression covers two main different types of linear regression models, namely the univariate linear as well as the multiple linear regression models. Generally speaking, when only one independent variable x is required to predict the work, the univariate linear regression model will be simpler in terms of model complexity due to the smaller number of independent variables x. However, it can also lead to a lack of fit of the data to the model due to the neglect of the use of the more highly correlated independent variables. The choice of the number of independent variables x to be used and the selection of the more highly correlated independent variables x for the construction of the linear model and the solution of the regression problem are therefore central to the linear regression algorithm. [1]The expression for the function of the univariate linear regression model is shown below.

$$\hat{y} = \beta_0 + \beta_1 x \tag{1.1}$$

where $\hat{y}$ is the predicted value of the model output and β_0, β_1 is the prediction coefficient, which can be calculated using the least squares method.

Multiple linear regression, on the other hand, uses multiple correlated independent variables for the regression problem and its model function expression is given by

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \tag{1.2}$$

Similar to the univariate linear regression model function, m is the total number of independent variables x used, and then $\beta_1, \cdots, \beta_m$ is the corresponding prediction coefficient for each independent variable.

1.3.2 Decision tree regression

Decision trees are a basic machine learning algorithm that can handle classification and regression problems. The implementation of the decision tree algorithm consists of three main parts: feature selection, decision tree generation and decision tree pruning. A decision tree can be graphically represented as a binary tree structure, where each internal node represents a judgement on an attribute, and each final leaf node represents the result of a classification regression. When dealing with regression problems, the CART (Classification And Regression Tree) algorithm is usually used. As the gold recovery prediction problem in this paper is a regression problem, we will focus on the implementation process of the CART algorithm.

Regression tree algorithm implementation process

Input: training data set $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$; assume X and Y are the input and output variables respectively.

Output: regression tree f(x).

---

[1] Linear regression theory

张涵夏.适用于线性回归和逻辑回归的场景分析[J].自动化与仪器仪表,2022(10):1-4+8.DOI:10.14016/j.cnki.1001-9227.2022.10.001.

https://kns.cnki.net/kcms/detail/detail.aspx?FileName=ZDYY202210001&DbName=DKFXTEMP

In the input space where the training dataset is located, recursively divide each region into two sub-regions and determine the output values on each sub-region to construct a binary decision tree by.

(1) Select the optimal cut variable j with cut point s and solve

$$\min_{j,s}[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2] \tag{1.3}$$

Iterate over variable j, scan cut point s for a fixed cut variable j, and select the pair (j,s) that minimizes the above equation .

(2) Using the selected pairs (j,s) delimit the region and determine the corresponding output values.

$$\widehat{c_m}=\frac{1}{N_m}\sum_{x_i\in R_m(j,s)}y_i\,,x\in R_m,m=1,2 \tag{1.4}$$

where , $R_1(j,s)=\{x|x^{(j)}\le s\},R_2(j,s)=\{x|x^{(j)}>s\}$

(3) Continue to call steps (1), (2) for both subregions until the stopping condition is satisfied.

(4) Divide the input space into M regions $R_1,R_2,\cdots,R_M$ Generate a decision tree.

$$f(x)=\sum_{m=1}^{M}\widehat{c_m}I(x\in R_m) \tag{1.5}$$

1.3.3 Random Forest Regression

Random Forest is an integrated algorithm that forms a strong learner by integrating multiple decision trees, each tree in Random Forest is independent of each other and unrelated. The basic principle is based on a statistical approach, where multiple samples are drawn from the original sample through a put-back sampling method, each sample is considered as a training set to build a decision tree, and then the prediction results of all the decision trees are averaged as the final prediction result. If m decision trees are built in a random forest, then the prediction model of the random forest for an unknown input (x can be a number or an n-dimensional vector) is :

$$f=\frac{1}{m}\sum_{i=1}^{m}f_i(x) \tag{1.6}$$

Usually, the number of decisions is more prone to overfitting, whereas random forests usually converge to a lower generalization error as the number of decision trees increases, avoiding the risk of overfitting.[2]

Random forest algorithm

Input: Training data set $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$; assume X and Y are the input and output variables respectively.

Output: prediction results

---

[2] Decision Trees and Random Forests regression theory

李浪.基于机器学习的高层建筑风压预测[D].广州大学,2021.DOI:10.27040/d.cnki.ggzdu.2021.001068.

https://kns.cnki.net/kcms/detail/detail.aspx?FileName=1021634114.nh&DbName=CMFD2022

(1) From the training dataset, use Bootstrap to find the samples D_j[3]

(2) Use D_j as training data to train a decision tree. In the process of generating the decision tree, for each leaf node corresponding to a sample number greater than n_{min} the number of samples do the following.

a. From all d variables available for selection, a random variable d_1 is selected

b. Select the variable d_1 that produces the optimal division from the variables

c. Divide the node into two sub-nodes according to the optimal variable chosen

d. Repeat the above process until all leaf nodes correspond to a sample book smaller than n_{min}

(3) Combine the number of m decision trees:$T_1, T_2, \cdots, T_m$:

$$f(x) = \frac{1}{m} \sum_{j=1}^{m} T_j(x) \tag{1.7}$$

1.3.4 Dummy regression

In social science research, there are many categorical variables such as region, period, company, ethnicity, gender, literacy, occupation, etc. This information can be used in regression analysis to explain changes in the dependent variable, but the categorical variables must first be transformed into dummy variables before they can be introduced into the regression equation for the resulting regression results to have a clear meaningful interpretation.

Let x be a nominal variable with k classifications, with different coded values representing the type to which the case belongs in the data processing. Because there is no quantitative relationship at all between the categories of categorical variables, it is not possible to analyse the average change in y when x changes by one unit, as is the case with spacing variables. Therefore, the effect of each category on y must be analysed in terms of classes. Imagine that k dummy variables with values of 0 and 1 are used to represent the properties of each category, and that when a case belongs to the category represented by a dummy variable, the dummy variable is assigned a value of 1, otherwise it is 0, thus representing the properties of the category of the case. For variables coded with only two values, 0 and 1, the mean is the proportion of cases in the category coded as 1 to the total sample, so that it can be regressed. [4]

However, this coding results in k dummy variables that are linearly correlated and do not qualify for regression. To overcome this problem, one dummy variable is generally dropped. No information is actually lost by doing so, as the kth class attribute represented by the canceled dummy variable can be fully represented by the values taken by the other k-1 dummy variables, i.e. when all k-1 dummy variables

---

[3] Bootstrap article. ODE course

URL: https://habr.com/ru/company/ods/blog/324402/

[4] Dummy variable regression theory

马秋芳,孙根年,谢雪梅.基于虚拟变量回归的旅游花费模型构建[J].统计与决策,2008(22):62-64.

https://kns.cnki.net/kcms/detail/detail.aspx?FileName=TJJC200822023&DbName=CJFQ2008

take the value 0, then the case is the kth class. This category, which is not explicitly represented by dummy variables, is the reference group and is of particular interest when analysing regression results.

One way of formulating the common-slope model is

$$Y_i = \alpha + \beta X_i + \gamma D_i + \varepsilon_i \quad (1.8)$$

where D, called a dummy-variable regressor or an indicator variable, is coded 1 for men and 0 for women:

$$D_i = \begin{cases} 0 & \text{for one outcome} \\ 1 & \text{for another outcome} \end{cases}$$

- Thus, for one outcome the model becomes

$$Y_i = \alpha + \beta X_i + \gamma(0) + \varepsilon_i = \alpha + \beta X_i + \varepsilon_i \quad (1.9)$$

- and for another outcome

$$Y_i = \alpha + \beta X_i + \gamma(1) + \varepsilon_i = (\alpha + \gamma) + \beta X_i + \varepsilon_i \quad (1.10)$$

1.4 Data interpretation

We recover the metal from gold-bearing ores in various stages, as shown in the image below.

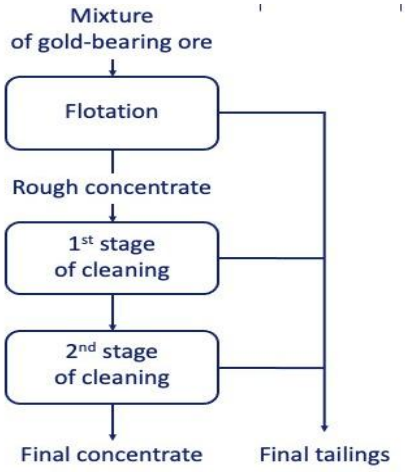


Image 1. Different stages of gold bearing ore processing

1step: Flotation

A mixture of gold-bearing ore is fed into the flotation plant. After enrichment, a rough concentrate and "dump tails" are obtained, that is, product residues with a low concentration of valuable metals.[5]

The stability of this process is affected by the unstable and suboptimal physico-chemical state of the flotation pulp (mixture of solid particles and liquid)

2step: Cleaning

The rough concentrate undergoes two purification. At the output, the final concentrate and new dump tails are obtained.

---

[5] Gold mining process

URL: https://quote.rbc.ru/news/article/5f508fb39a794759bf6d23c3?ysclid=lamfiemvco212835478

Chapter 2 Analysis of recovered metals in gold ore

2.1 Calculation and analysis of concentration efficiency

First, we have to select the relevant indicators. The indicator we use to calculate the enrichment recovery is as follows:

•C- the proportion of gold in concentrate after flotation/purification

•F-the proportion of gold in the raw material/concentrate before flotation/purification

•T- the proportion of gold in tailings after flotation/purification

Next, according to the above indicators to calculate the enrichment recovery formula.

$$\text{Recovery} = \frac{C \times (F - T)}{F \times (C - T)} \times 100\% \quad (2.1)$$

Finally we calculate the MAE value.

```
MAE value: 9.73512347450521e-15
```

Image 2. MAE value

The MAE value is very small and corresponds to the measurement error of enrichment efficiency. We conclude that the calculation of the efficiency of enrichment of raw materials after flotation is correct.

2.2 Analysis of Au, Ag and Pb concentrate dynamics per step of production

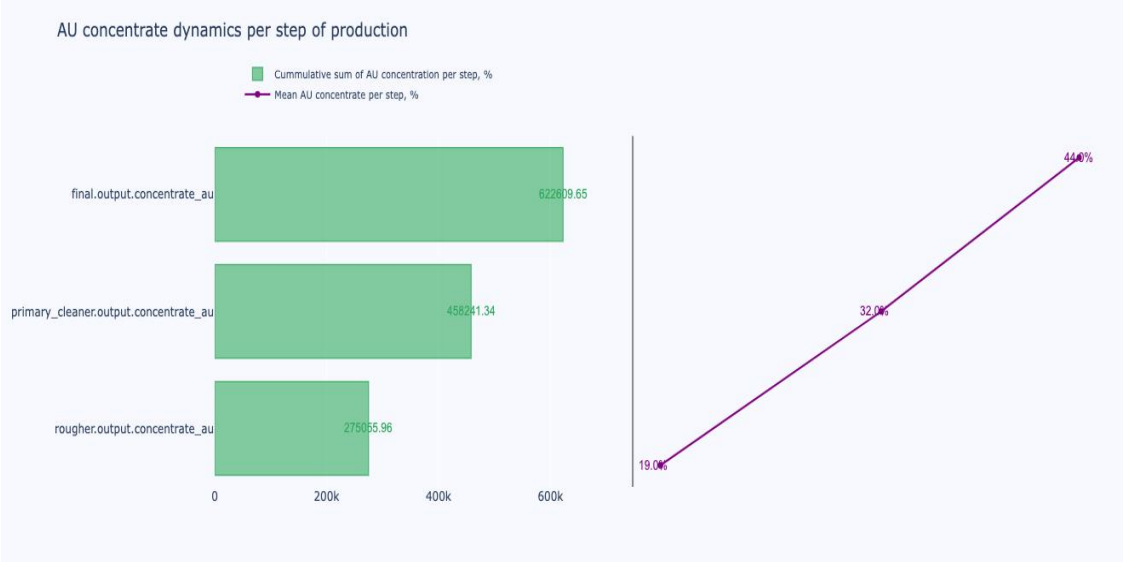
In this section we want to check the concentration of metals at various stages of purification.

According to the purification process consists of two main steps, the first cleaning and the second cleaning. For this purpose we have selected three indicators to analyze the metal concentration.

- Rougher.output.concentrate - represents the concentration after flotation and not after cleaning.
- Primary_cleaner.output.concentrate - represents the concentration after the primary cleaning.
- Final.output.concentrate - represents the concentration after both cleaning processes.

Then the gold-bearing ores contain three main metals: Au, Ag and Pb. We therefore focus on these three metals with respect to these indicators in this section.

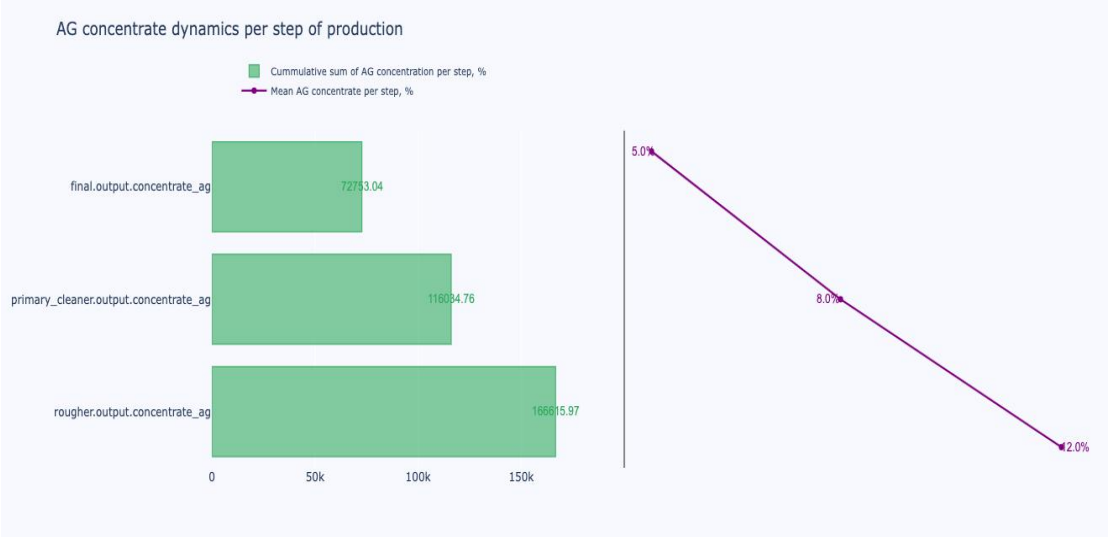
Firstly, we analyze the concentration of gold concentrate at each step of the purification process.



Graph 1. Au concentrate dynamics per step of production

The dashboard shows that as the raw material is refined, the concentration of gold increases.

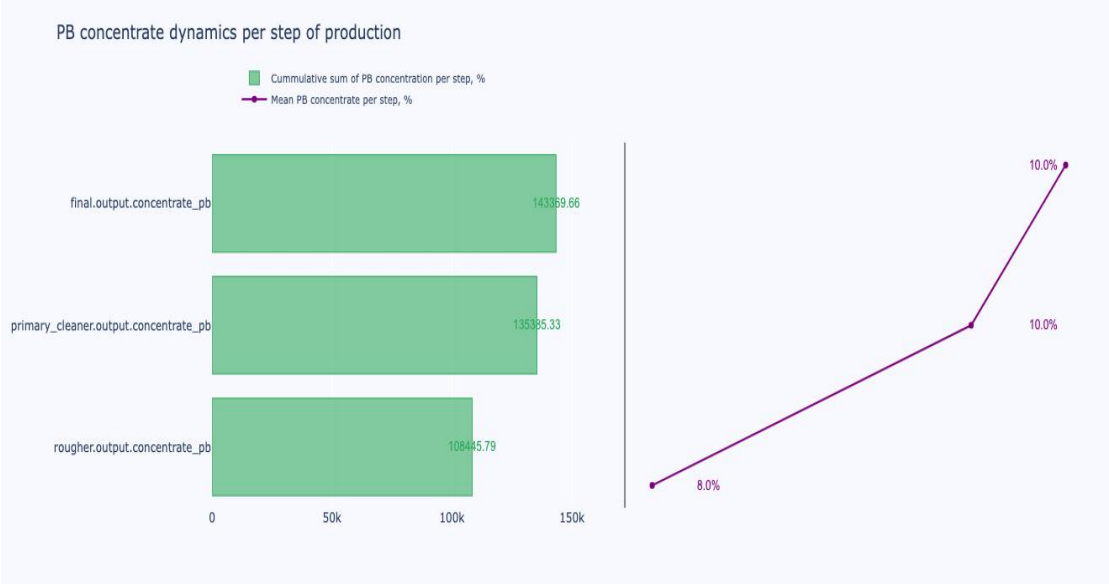
Secondly, we analyze the concentration of silver concentrate at each step of the purification process.



Graph 2. Ag concentrate dynamics per step of production

The concentration of silver decreases as the raw material is purified from impurities.

Thirdly, we analyze the concentration of lead concentrate at each step of the purification process.



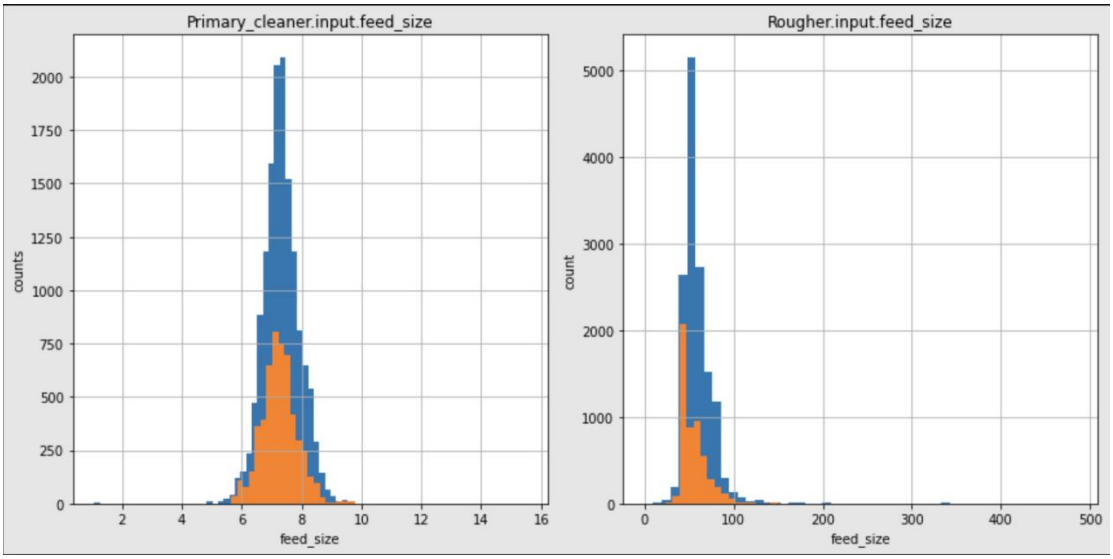
Graph 3. Pb concentrate dynamics per step of production

The concentration of the metal increases during purification at the stage of primary purification, but remains practically unchanged after the secondary purification of raw materials.

2.3 Granules size analysis of the raw material during the cleaning process

In this section we focus on comparing the particle size of raw gold-bearing ore without flotation to that of a rough concentrate that has been flotation washed and not primary cleaned.

In addition to this, we contain two samples, a training sample and a test sample, so we will also discuss the particle size between these two samples in this section. The results are shown in the graph below.



Graph 4. Comparison of the size distribution of raw material granules on the training and test samples

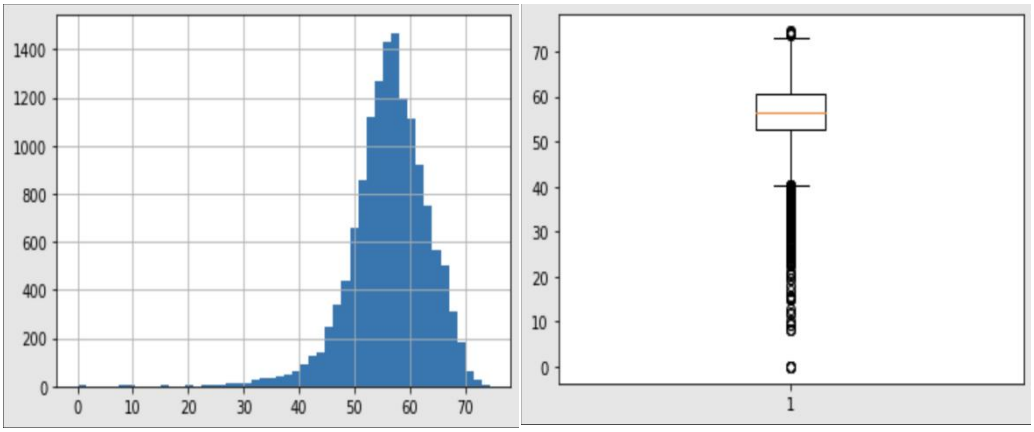
The distribution for the test sample is slightly different from that for the training sample - the first one is slightly shifted down. But in general, the sizes of most particles lie in the optimal range of 40–100 μm for both samples. So yes, the samples are eligible for evaluation.

2.4 Analysis of the concentration of all metals at each stage of the purification process

We analyze the total concentration of all substances at different stages. The indicators to be selected for our analysis of the purification process belong to three different stages.

The first being the total concentration of material without flotation (i.e. the total concentration of all material in the mixture of gold-bearing ore); the second being the total concentration of all material in the rough concentrate after flotation; and the third being the total concentration of all material in the final concentrate after cleaning.

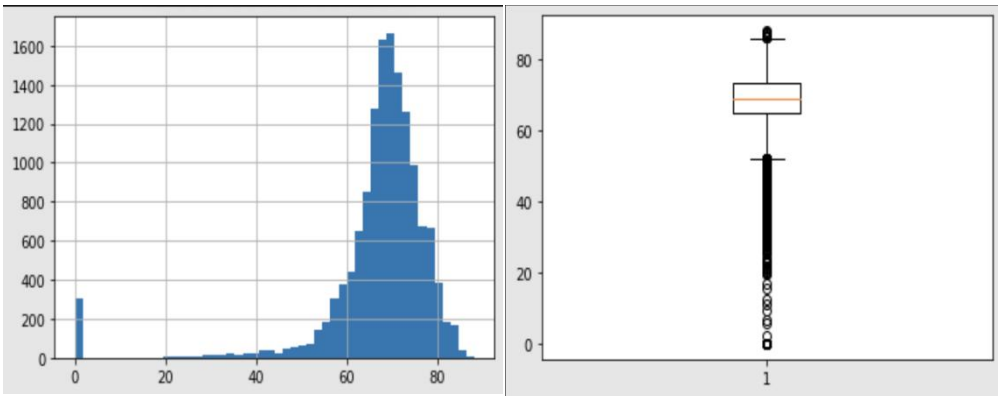
First we analyze the total concentration of all substances in the raw material, i.e. the mixture of gold-bearing ore.



Graph 5. Rougher input feed total

As can be seen from the graph, at the stage of raw materials, the distribution of the concentration of substances is in the range from 40% to 70%.

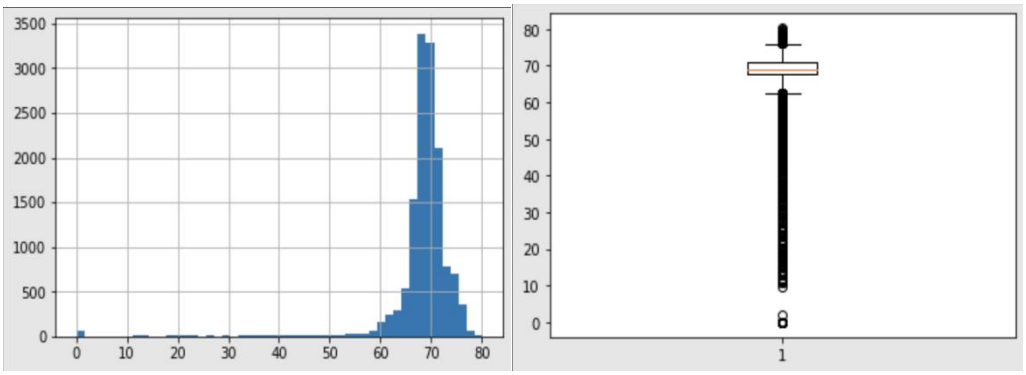
Next we analyze the total concentration of all substances in the crude concentrate.



Graph 6. Rougher output concentrate total

At the stage of crude concentrate, the total concentration of all substances is shifted to the right in the range from 50% to 85%.

Finally we analyze the total concentration of all substances in the final concentrate.



Graph 7. Final output concentrate total

At the stage of the final concentrate, the total concentration of all substances goes into the range from 65% to 75%.

There are zero anomalies in total concentrations - and it is better to "kill" them. The fact is that at least a gram of something is in the "total ore" - salt + gold + silver + lead - there must definitely be something .. And if we have zero, then there is a high probability of inaccurate measurements.

Chapter 3: Prediction of gold recovery rate

3.1 Model construction and selection of indicators

We use the Symmetric mean absolute percentage error (sMAPE)as a metric to evaluate the model. It is expressed not in absolute values, but in relative ones. Equally takes into account the scale of both the target attribute and the prediction. The sMAPE formula is as follows:

$$\text{sMAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i-\hat{y}_i|}{(|y_i|+|\hat{y}_i|)/2}\times 100\% \quad (3.1)$$

Where, y_i-the value of the target attribute for the object with the ordinal number i in the sample on which the quality is measured

- $\hat{y}_i$-the prediction value for the object with the sequence number i in the sample
- N - Number of objects in the selection

What's more, in order to calculate the total sMAPE value, we need to calculate the sMAPE value under the rougher concentrate stage and the sMAPE value of the final concentrate. The total sMAPE formula is as follows:

$$\text{total sMAPE} = 25\% \times \text{sMAPE(rougher)} + 75\% \times \text{sMAPE(final)} \quad (3.2)$$

3.2 Empirical analysis of the predicted gold recovery rate

We use cross-validation to check the model and find the best hyperparameters of

the model for model correction to get the best smape value.

```
Model tested: LinearRegression()
Best Score: 180.98445150813168
Best Hyperparameters: {'copy_X': True, 'fit_intercept': False, 'n_jobs': None}
Model tested: RandomForestRegressor(random_state=0)
Best Score: 12.603258682763611
Best Hyperparameters: {'max_depth': 9, 'n_estimators': 1}
Model tested: DecisionTreeRegressor(random_state=0)
Best Score: 14.328919704764557
Best Hyperparameters: {'max_depth': 9, 'min_samples_split': 3}
```

Image 3. Model checking (cross-validation)

As can be seen from the image, the smallest smape value in the training sample is shown by the random forest regression model, so we used this model to test the smape value in the test sample and the results are shown below.

Image 4. Test sample smape value

The value of the resulting metric on the test sample turned out to be 10.57%, which is a fairly representative value for the model.

The metric is quite low. And that's good. We need to understand that our model predicts better than just the average.

For example, substitute the median values of the target train into the predictive metric formula or use Dummy Regression.

Image 5. Test sample smape value

Comparison with the constant model showed that the sMAPE metric on the constant model is slightly lower than on the predicted values of the random forest model, so it is rather difficult to talk about the adequacy of the predictions made by the model.

Chapter 4 Conclusion

This paper describes the concentration and particle size at different stages of the process of recovering metals from gold-bearing ores, and also proposes the use of machine learning algorithms to predict gold recovery and evaluate optimized models using a cross-validation approach, leading to the following conclusions.

1. During the cleaning process, the concentration of gold has been gradually increasing, the concentration of silver has been gradually decreasing, and the concentration of lead is increasing in the first cleaning process and remains the same in the second cleaning process.

2. In the cleaning process, the sizes of most particles lie in the optimal range of 40–100 μm for both train and test samples.

3. During the purification process, at the stage of raw materials, the distribution of the concentration of substances is in the range from 40% to 70%. At the stage of crude concentrate, the total concentration of all substances is shifted to the right in the range from 50% to 85%. At the stage of the final concentrate, the total concentration of all substances goes into the range from 65% to 75%.

The purification process allows the concentration of metals to be more concentrated in a certain range, allowing a better assessment of the quality of the gold-bearing ore and the level of purification. This shows a good level of purification.

4. Among the three prediction models, the random forest model had the lower smape value and the best fit for the prediction model evaluation, and the optimal hyperparameters of the random forest model were obtained using cross-validation to check that the max-depth was 9 and the n-estimators were 1.

5. Our comparison using the constant model revealed that the smape values were slightly lower than those of the random forest model, making it difficult to verify the adequacy of the predictions made by the model and requiring considerable work to be done to discuss them.

References

1. Zyfra company (data provider)
URL: https://www.zyfra.com/
2. Random forest article
URL: https://habr.com/ru/company/ruvds/blog/488342/
3. Bootstrap article. ODE course
URL: https://habr.com/ru/company/ods/blog/324402/
4. Classification with decision trees. ODE course.
URL: https://habr.com/ru/company/ods/blog/322534/
5. Gold mining process
URL:
https://quote.rbc.ru/news/article/5f508fb39a794759bf6d23c3?ysclid=lamfiemvco212835478
6. Linear regression theory
张涵夏.适用于线性回归和逻辑回归的场景分析[J].自动化与仪器仪表,2022(10):1-4+8.DOI:10.14016/j.cnki.1001-9227.2022.10.001.
URL:
https://kns.cnki.net/kcms/detail/detail.aspx?FileName=ZDYY202210001&DbName=DKFXTEMP
7. Decision Trees and Random Forests regression theory
李浪.基于机器学习的高层建筑风压预测[D].广州大学,2021.DOI:10.27040/d.cnki.ggzdu.2021.001068.
URL:
https://kns.cnki.net/kcms/detail/detail.aspx?FileName=1021634114.nh&DbName=C

MFD2022
8. Dummy variable regression theory
马秋芳,孙根年,谢雪梅.基于虚拟变量回归的旅游花费模型构建[J].统计与决策,2008(22):62-64.
URL:
https://kns.cnki.net/kcms/detail/detail.aspx?FileName=TJJC200822023&DbName=CJFQ2008