Composition analysis and identification of ancient glass objects based on LightGBM

Quanming Chen [1, *], Guoxing Zhu [2], Quanfu Zhang [2]

[1] Department of Big Data Academy, Yunnan Agricultural University, Kunming, China

[2] School of Electrical and Mechanical Engineering, Yunnan Agricultural University, Kunming, China

* Corresponding Author Email: 2500929969@qq.com

Abstract. In this paper, for a batch of existing data related to ancient glass objects in China, by analyzing the relationship between the basic information of cultural relics and the information of chemical composition detected by cultural relics, a prediction model that can predict the chemical composition data before weathering by the chemical composition data after weathering is established, a classification model based on the existing data is built to simulate the original classification law, and the model is used to identify the type of unclassified cultural relics, and then by The importance of the feature selects barium oxide as a criterion for subcategory classification. The proportion of chemical composition of the classified glass artifacts was determined by feature engineering to determine whether the 0 in the detection data was filled in artificially, and it was used as a new feature to reduce the impact of extreme data 0 on the model. The data were then used for model training to obtain the LightGBM classification model. The mean value of the chemical components with the highest feature importance in the model was selected as the criterion for subclass classification, and the values in each class were divided into 2 subclasses. The chemical components with the top 3 feature importance were subjected to sensitivity analysis, and sensitivity judgments were made by the indicators of the model. After first performing feature engineering on the data information of the unknown category of artifacts, the type identification output results were performed using the classification model built in Problem 2. The top 3 chemical components of importance in the model were selected to let their values fluctuate up and down by 5%, and the model indicator curves were plotted for sensitivity analysis.

Keywords: Antique Glassware, Chemical composition, LightGBM Classification Model, Sensitivity Analysis.

1. Introduction

The main chemical composition of glass is silicon dioxide (SiO_2), in order to reduce the melting temperature of pure quartz sand is often added to a variety of accelerants, ancient glass due to the addition of different accelerants resulting in its main chemical composition is different [1-3]. Archaeologists have classified ancient glass into two types, high potassium glass and lead-barium glass, through chemical composition analysis [4-6]. At the same time, because ancient glass is easily affected by the burial environment and weathering, the internal elements in the weathering will exchange with the environmental elements, which will lead to changes in the chemical composition ratio and lead to the judgment of the category [7-10]. This paper intends to solve the following problems.

(1) Analyze the statistical laws of classified high potassium glass and lead-barium glass according to the test data; for each category take the appropriate chemical composition of the category for sub classification, get the detailed classification method and classification results, and analyze the reasonableness and sensitivity of the classification results.

(2) Analysis of the chemical composition of unknown categories of glass artifacts to determine the type of glass artifacts, as well as the sensitivity of the results of the analysis.

2. Building a Light GBM classification model

2.1. Feature engineering and model building

In the process of weathering, the main chemical composition of glass artifacts will be lost in large quantities, but this paper concludes that the main chemical composition is still the main factor in determining the type of artifacts. Secondly, we analyze the influence of different basic information on whether weathering, and conclude that ornamentation, color has no significant effect on whether weathering, but different types of glass have a certain influence on the judgment of whether weathering, according to this paper, the known glass artifacts will be used as a feature data for training the glass artifact type discrimination model.

a. Whether the proportion of chemical element content is manually added to the data: In the assumption of the model, the proportion of chemical element content of the vacancy is set to 0 in this paper, but in fact, some of the data in the original data would have been 0. In order to reduce the impact of extreme values such as 0 on the model, we choose to add the indicator of whether the element is detected by feature engineering.

b. The mean value of the elements of the repeatedly sampled artifacts: Considering the problem that different locations of the same artifact are sampled multiple times and the chemical composition ratio obtained from different locations on the same artifact is inconsistent, this paper chooses to use the mean value of the chemical element content of the artifacts sampled multiple times as the chemical element content. In the problem of feature engineering mentioned in the first has been extracted by way of number to determine whether the artifacts are repeatedly sampled, only the same number of different sampling points to obtain the proportion of each chemical composition of the average can be derived. As shown in Table 1. Results of mean processing of delineated components at some sampling points.

c. Whether the artifacts are weathered or not: In order to better make the discrete data better for model training choose a unique thermal encoding of the raw data to make the calculation between features more reasonable. As shown in Table. 2. Test of Light GBM prediction results.

Table 1. Results of mean processing of delineated components at some sampling points.

Sampling Points	Before processing	After mean processing
03 Part 1	87.05	74.38
03 Part 2	61.71	74.38
06 Part 1	67.65	63.73
06 Part 2	59.81	63.73

The Light GBM model was used to analyze the classification pattern and to obtain a classification model that can be used to discriminate the type of glass artifacts of unknown types.

Table 2. Test of Light GBM prediction results.

Artifact Number	Type	Prediction Type
01	High Potassium	High Potassium
02	Lead barium	Lead barium
03	High Potassium	High Potassium
04	High Potassium	High Potassium
05	High Potassium	High Potassium
06	High Potassium	High Potassium
07	High Potassium	High Potassium
08	Lead barium	Lead barium
09	High Potassium	High Potassium
10	High Potassium	High Potassium

Subcategorization refers to the identification of new discriminative indicators to further classify the elements in the original categories. In this paper, we calculate the importance of each feature in the model after constructing the discriminant model, and select the mean value of the most important feature (lead oxide) in each category as the discriminant criterion for the subcategory classification.

As shown in Table. 3. Results of sub classification of high potassium glass.

The mean value of lead oxide content in the category of high potassium glass was calculated to be 0.27, and the mean value of lead oxide content in the category of lead-barium glass was calculated to be 33.35. The subcategories of high potassium glass and lead-barium glass were further classified according to these criteria. As shown in Table. 4. Lead barium glass sub-classification results.

In each category, further sub classification is made according to the discrimination criteria: data greater than the average value of lead oxide in this category are classified as category 1, while data less than and equal to the average value of lead oxide in this category are classified as category 2. As shown in Figure 1. Feature importance.

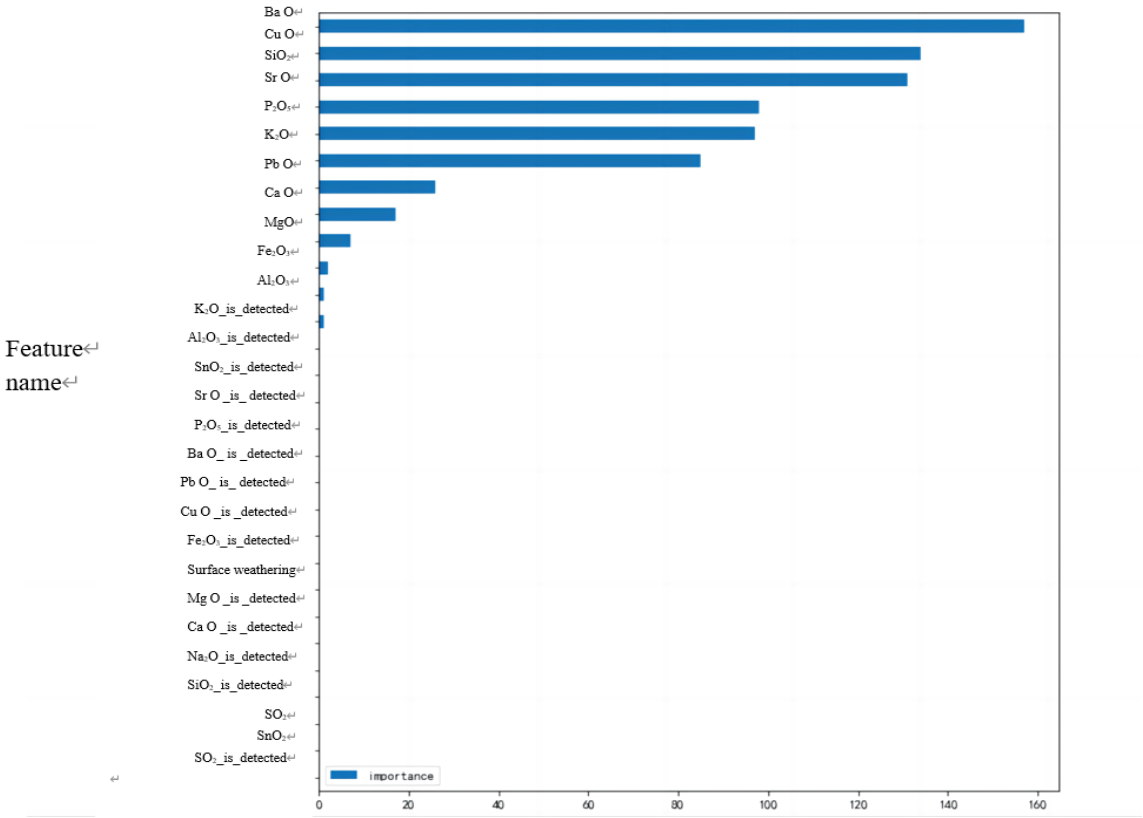

Figure 1. Feature importance.

Table 3. Results of sub classification of high potassium glass.

1 class number	Numerical value	2 class number	Numerical value
03 part 1	2 83	1	0
06 part 1	1.38	03 part 1	0
06 part 2	0.97	4	0
21	1.97	5	0
		7	0
		9	0
		10	0
		12	0
		13	0.11
		14	0
		16	0
		18	0

Table 4. Lead barium glass sub-classification results.

1 Class	Numerical value	2 Class	Numerical value
8	31.23	19	30.565
8 Severe weathering point	30.62	25 Unweathered spots	25.39
11	14.61	28	9.3
19	23 55	29	16 98
20	11.86	30 part 1	29.14
23 Unweathered spots	26.23	30 part 2	31.9
24	32 25	31	29 725
25 For weathering point	35.45	32	29.725
26	10.83	33	17.14
26 Severe weathering point	10.88	34	12.31
36	10.96	35	16.55
42	14.2	37	19.76
44	15.45	38	zh16.16
50	17.3	39	22.05
56		40	32.92

2.2. Reasonableness and sensitivity analysis

Rational analysis:

The analysis of the importance of the features shows that lead oxide is the most important feature, and based on the information in the question that lead oxide is the main component of lead-barium glass, it can be inferred that the results obtained from the model are reasonable and do not violate the established rules of the question.

Sensitivity analysis: In constructing this classification model, there are a large number of features, including the proportion of each chemical element and whether it is detected or not. In this paper, the top three features of barium oxide, copper oxide, and silica were selected for sensitivity analysis based on the importance of these features.

In this paper, the change curves of precision Precison_score, recall_score, and roc_auc_score of the model are plotted by setting their values to fluctuate up and down by ±5%, and their sensitivity is analyzed by observation.

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{TP}{TP+PN} \tag{2}$$

$$AUC = \frac{\sum_{positive} rank_i - \frac{M(1+M)}{2}}{TP+FP} \tag{3}$$

Similarly, the three feature controls of the top3 importance in the model are changed numerically by ±5% each time, and the change curves of the precision Precison_score of the model, recall recall_score of the model, and roc_auc_score of the model are plotted, and their sensitivity is analyzed by observation. As shown in Figure 2(BaO sensitivity test), Figure 3(CuO sensitivity test) and Figure 4(SiO_2 sensitivity test).

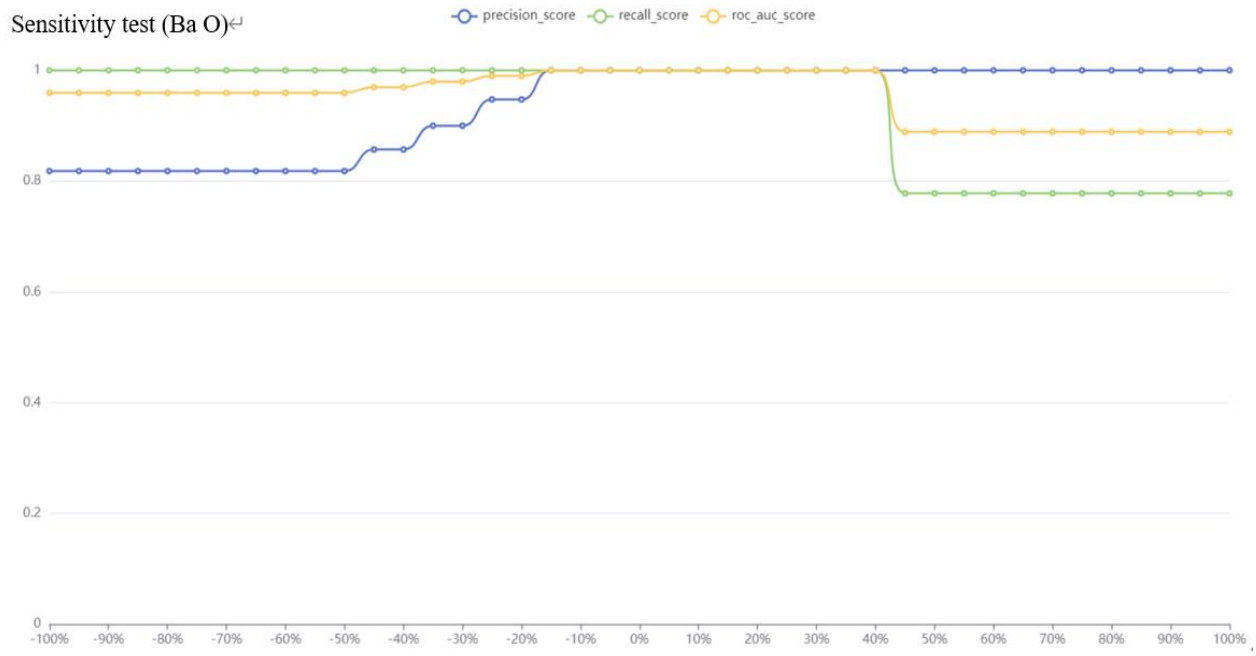

Figure 2. BaO sensitivity test.

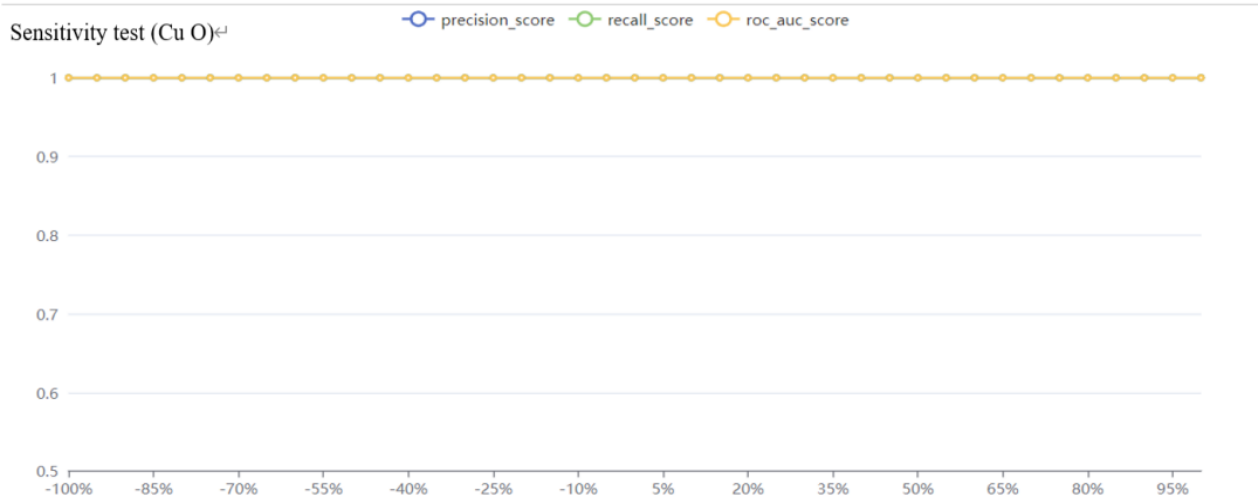

Figure 3. CuO sensitivity test.

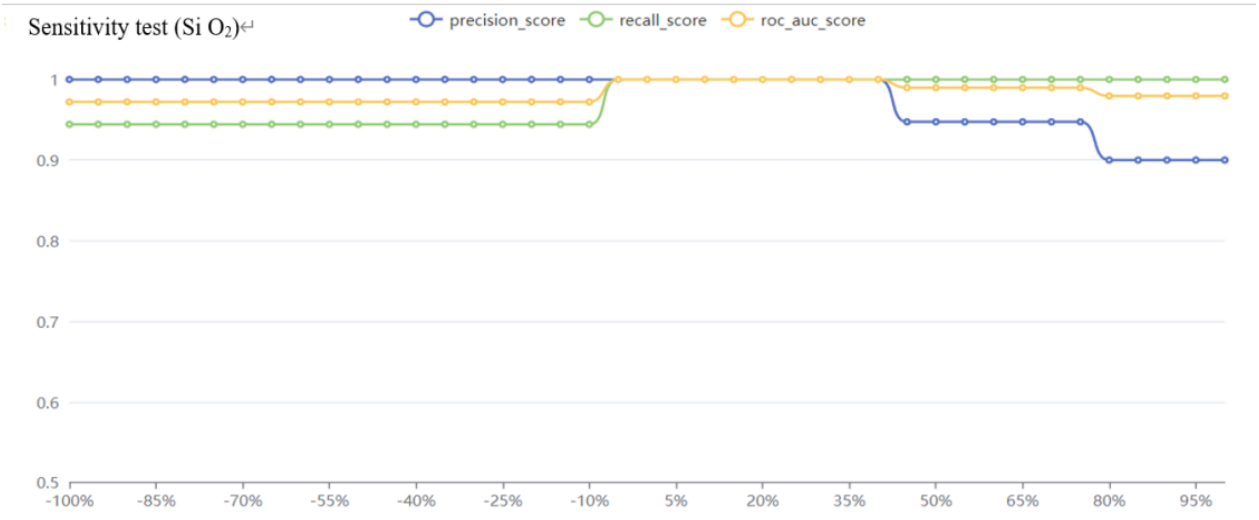

Figure 4. SiO_2 sensitivity test.

The above image curves reveal that:

When the value of barium oxide fluctuates between -10% and 30%, the model has almost no effect, but below -50%, the accuracy of the model is lower than the original one by nearly 20%, and above 40%, it is obvious that the recall_score and roc_auc_score of the model decrease rapidly by nearly 5% and 10% respectively.

When the value of copper oxide is lower than 0%, there is almost no effect on the model, and the system results are stable, but the accuracy of the model and roc_auc_score change when the value fluctuates between 0% and 5%.

The accuracy and roc_auc_score of the model change when the value of silica is lower than 30%, and the system results are stable.

From the above analysis, it is easy to see that the trained model has excellent stability, and the performance of the model will not be affected when the features fluctuate.

3. Classification using LightGBM classification model

3.1. Request unknown category of glass artifact identification type.

Using the model constructed above, it is possible to identify the type of glass artifacts from the available data. Since the model constructed in Problem 2 is to be used, feature engineering is also required first.

Feature engineering:

a. Fill vacant values: Fill the vacant data with 0.

b. Discriminate whether the element is detected or not: As in problem 2, the keyword is used to discriminate whether the element is detected or not.

The processed data is fed into the type discrimination model constructed above, and the output is the class of the unknown glass artifacts identified by the model. The identification results are shown in Table 5.

Table. 5. Results of data identification for unknown categories.

Number	Predicted results
A1	High Potassium
A2	Lead Barium
A3	Lead Barium
A4	Lead Barium
A5	Lead Barium
A6	High Potassium
A7	High Potassium
A8	Lead Barium

3.2. Sensitivity Analysis

Similarly, the three features (BaO, CuO, and SiO_2) of the top 3 importance in the model were controlled to change numerically by ±5% each time, and the curves of the precision Precison_ score, the recall recall_score, and the roc_auc_score of the model were plotted to analyze their sensitivity by observation. As shown in Figure 5. BaO sensitivity test, Figure 6. CuO sensitivity test and Figure 7. SiO2 sensitivity test.

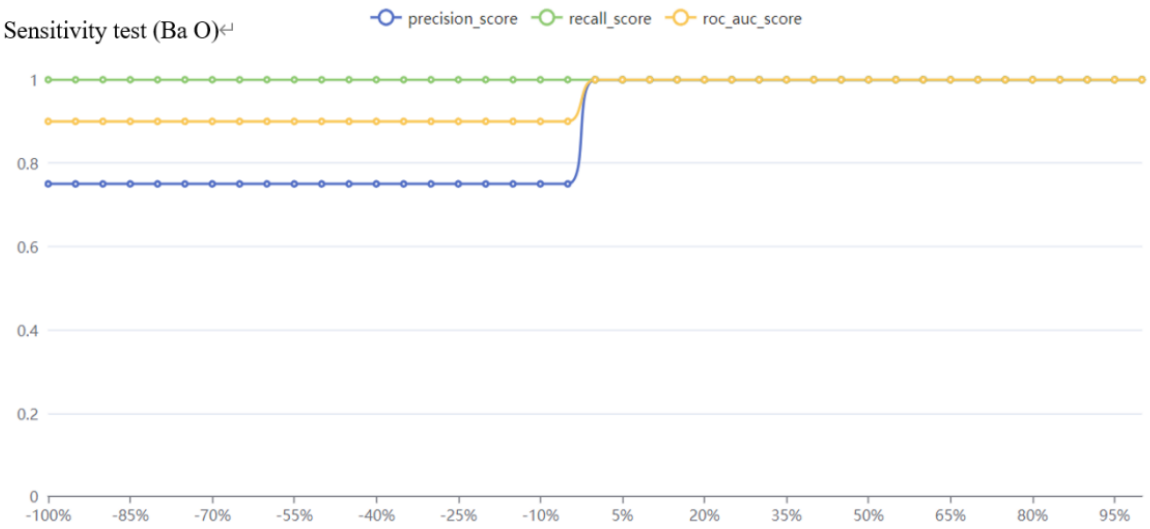

Figure 5. BaO sensitivity test.

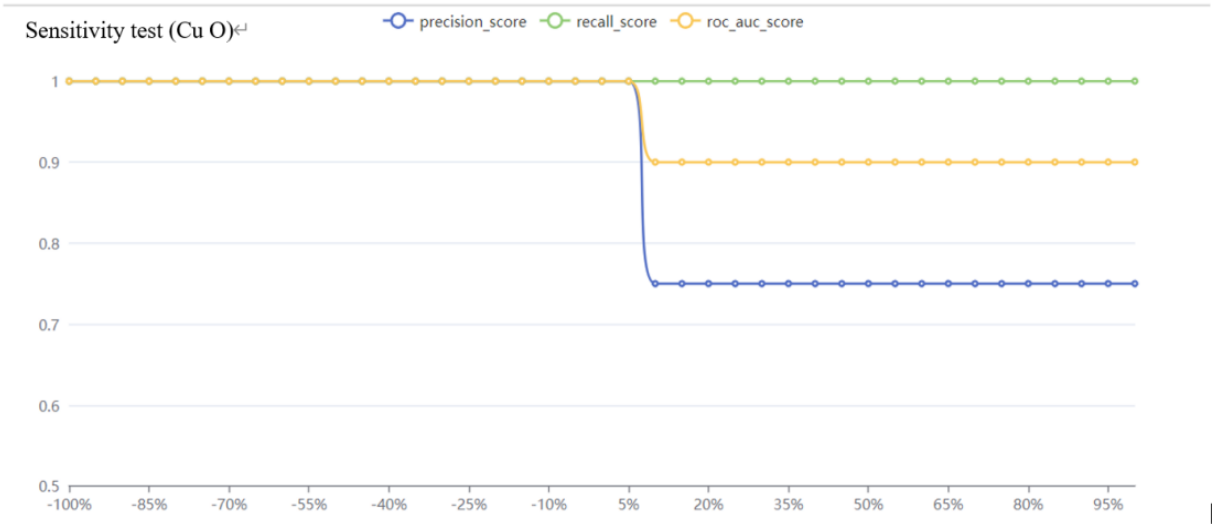

Figure 6. CuO sensitivity test.

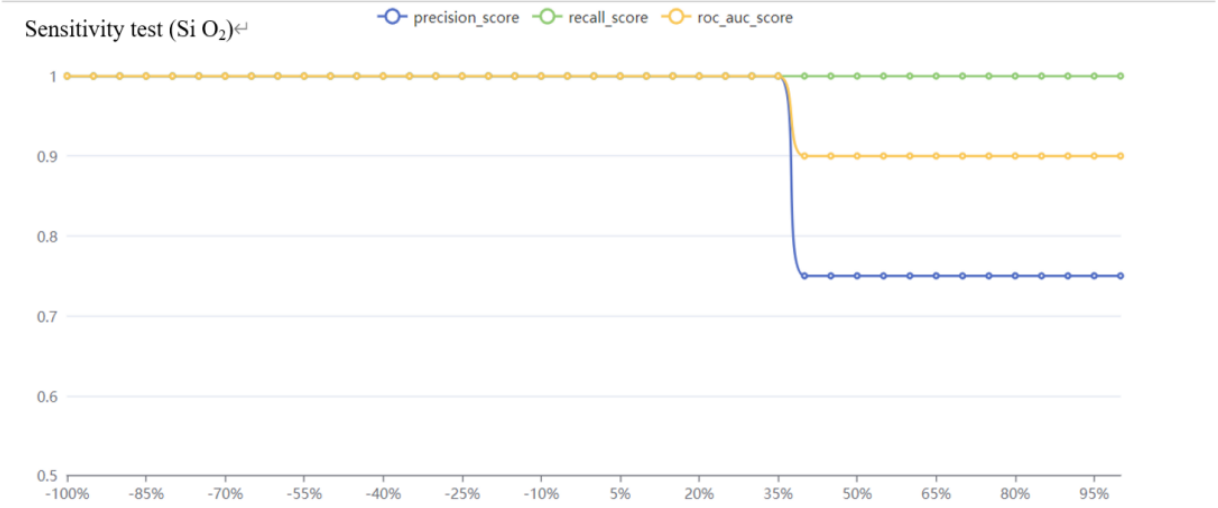

Figure 7. SiO_2 sensitivity test.

Analysis by image:

When the value of PbO fluctuates between - 10% and 40%, the model has almost no effect, but below - 50% the accuracy of the model is almost 20% lower than the original one.

The recall_ score and roc_ auc_ score of the model decreased rapidly when the value was higher than 40%, and decreased by 25% and 10%, respectively, compared with the original model.

When the number of copper oxide fluctuates arbitrarily, the three judgment indexes of the observation analysis model do not receive significant influence, and the system performs stably.

When the value of silica fluctuates between -5% and 40%, the model is almost unaffected, but below -5% the ecall_ score of the model is not affected.

The ecall_ score and roc_ auc_ scor of the model are lower than the original by nearly 20%, and the accuracy of the model decreases when the value is higher than 40%, which is lower than the original accuracy by nearly 5%, but it tends to stabilize again at 40%-70% until it decreases again when it exceeds 70%, which is lower than the original model data by nearly 10%.

According to the above conclusions, the Light GBM model in Problem 2 has a stable index fluctuation of less than 30% in identifying unknown categories of artifacts, which can be judged as a stable model that does not affect the accuracy of the model due to rare numerical fluctuations. The model has a good generalization capability for the identification of glass artifacts in the previous categories.

4. Conclusions

This paper combines the advantages of Light GBM's leaf-wise leaf growth strategy with depth limitation to achieve a more accurate prediction of the chemical composition before glass weathering without overfitting. By combining the advantages of Light GBM with depth-limited leaf-wise leaf growth strategy, the prediction of the chemical composition content before glass weathering is not over-fitted and the prediction results are more correct. Given that extreme values such as manually filling in 0 may affect the accuracy of the model, new features are introduced to distinguish between manually added 0 values and detected 0 values, which greatly reduces the impact of the quarterly one on the model construction. The prediction accuracy of the LightGBM model was used to predict the chemical composition of unclassified glass artifacts, and the resulting glass type predictions were high and accurate in terms of model scores.

The LightGBM model quickly predicts the chemical composition of two types of glass before weathering and presents them as histograms, giving a better visualization of the chemical composition of two types of glass before weathering.

References

[1] Mocioiu Oana Cătălina, Mocioiu Ana-Maria, Neagu Simona, Enache Mădălin. Development of anticorrosive and antibacterial coatings for preservation of glass heritage objects [J]. Materials Today: Proceedings, 2021, 45(P5). Fangfang. Research on power load forecasting based on Improved BP neural network [D]. Harbin Institute of Technology, 2011.

[2] Melada Jacopo, Ludwig Nicola, Micheletti Francesca, Orsilli Jacopo, Gargano Marco, Grifoni Emanuela, Bonizzoni Letizia. Visualization of defects in glass through pulsed thermography [J]. Applied optics, 2020, 59(17). Ma Kunlong. Short term distributed load forecasting method based on big data [D]. Changsha: Hunan University, 2014.

[3] Lavinia de Ferri, Francesco Mezzadri, Roberto Falcone, Valeria Quagliani, Fabio Milazzo, Giulio Pojana. A non-destructive approach for the characterization of glass artifacts: The case of glass beads from the Iron Age Picene necropolises of Novilara and Crocefisso-Matelica (Italy) [J]. Journal of Archaeological Science: Reports, 2020, 29(C).

[4] Archaeology and Anthropology; Studies from A. Carter et al Further Understanding of Archaeology and Anthropology (Glass Artifacts at Angkor: Evidence for Exchange) [J]. Science Letter, 2019.

[5] Alison Carter, Laure Dussubieux, Martin Polkinghorne, Christophe Pottier. Glass artifacts at Angkor: evidence for exchange [J]. Archaeological and Anthropological Sciences, 2019, 11(3).

[6] Tomasz Purowski, Luiza Kępa, Barbara Wagner. Glass on the Amber Road: the chemical composition of glass beads from the Bronze Age in Poland [J]. Archaeological and Anthropological Sciences, 2018, 10(6).

[7] Ben Ford. The Glass and Ceramic Assemblage of the Mardi gras Shipwreck [J]. Historical Archaeology, 2017, 51(3).

[8] . Chemistry - Inorganic Chemistry; Studies from Max-Planck-Institute for Solid State Research in the Area of Inorganic Chemistry Reported [Glass-Induced Lead Corrosion of Heritage Objects: Structural Characterization of K (OH) center dot 2PbCO (3)] [J]. Chemicals & Chemistry, 2017.

[9] Bette Sebastian, Eggert Gerhard, Fischer Andrea, Dinnebier Robert E. Glass-Induced Lead Corrosion of Heritage Objects: Structural Characterization of K (OH) ·2PbCO3. [J]. Inorganic chemistry, 2017, 56(10).

[10] A. Blomme, P. Degryse, E. Dotsika, D. Ignatiadou, A. Longinelli, A. Silvestri. Provenance of polychrome and colourless 8th–4th century BC glass from Pieria, Greece: A chemical and isotopic approach [J]. Journal of Archaeological Science, 2017, 78.