Research on Credit Strategy Based on XGBoost Algorithm and Optimization Problem

Yunke Cheng
School of Electronic Information of Wuhan University, Wuhan, China
1244660970@eis.whu.edu.cn

Abstract. Small and medium-sized enterprises are the important foundation of national economic and social development, and play an important role in expanding employment, increasing income, improving people's livelihood and national taxation. In this paper, the credit decision-making of small and medium-sized enterprises is deeply studied, and reasonable credit risk quantification results and targeted credit strategies are given. To solve the first problem, this paper establishes a multi-index scoring model based on XGBoost algorithm. According to the extracted features, the credit risk of 123 enterprises is quantitatively analyzed, and the enterprises are scored according to the quantitative results. Prioritize them by score and enterprise credit rating, take interest rate and loan amount as decision variables, and bank profit as objective function, establish a multivariate and single-objective credit strategy optimization model, and get the optimal credit strategy for each enterprise when the annual total credit is fixed. In view of the second question, this paper takes the data in Annex 1 as the training set based on the first question, and extracts the features of the relevant data in Annex 2. Using the enterprise multi-index classification scoring model based on XGBoost algorithm, the credit rating of 302 enterprises is predicted and their credit risks are quantitatively analyzed, and the enterprises are scored according to the quantitative results. Prioritize them by score and enterprise reputation grade, establish the same optimization model as the first one, and formulate specific credit rules to make the results meet the limit conditions of total credit, thus determining the interest rate and loan amount of each enterprise, and obtaining specific credit strategies. In view of the third question, this paper takes the epidemic situation in SARS-CoV-2 as an example, and comprehensively considers the impact of this sudden factor on different industries and different types of enterprises on the basis of the second question. Based on the impact of corporate profits, the data are updated, the indicators are extracted again, and the XGBoost enterprise multi-index classification scoring model is used to predict the credit rating of 302 enterprises again and quantitatively analyze the credit risk, and the enterprises are scored according to the quantitative results. Once again, they are prioritized according to the score and the reputation level of the enterprise. Use the optimization model in question 2 to adjust the corresponding credit strategy. The model established in this paper considers more indicators, and the quantitative and prediction results are reasonable. The algorithm and model can comprehensively consider the impact of unexpected factors and adjust the credit strategy in time. In addition to being...
applied to credit decision-making of small and medium-sized enterprises, it can also be extended to other types of financial fields, providing ideas for solving related problems.

**Keywords:** Credit decision; XGBoost algorithm; Multi-index classification scoring model; optimization problem.

1. Restatement of problems

In practice, because small, medium and micro enterprises are relatively small in scale and lack of mortgage assets, banks usually provide loans to enterprises with strong strength and stable supply and demand relationship according to credit policies, transaction notes information of enterprises and the influence of upstream and downstream enterprises, and can give preferential interest rates to enterprises with high reputation and low credit risk. First, the bank evaluates the credit risk of small and medium-sized enterprises according to their strength and reputation, and then determines whether to lend or not and the credit strategies such as loan amount, interest rate and term according to factors such as credit risk.

A bank's loan quota for enterprises that are determined to lend is 100,000 to 1 million yuan; The annual interest rate is 4% ~ 15%; The loan term is 1 year. Annexes 1-3 respectively give the relevant data of 123 enterprises with credit records, 302 enterprises without credit records and the statistical data of the relationship between loan interest rate and customer churn rate in 2019. According to the actual data and information in the annex, the credit strategy for small and medium-sized enterprises is studied by establishing a mathematical model to solve the following problems:

**Question 1:** The credit risk of 123 enterprises in Annex 1 is quantitatively analyzed, and the credit strategy of the bank to these enterprises is given when the annual total credit is fixed.

**Question 2:** On the basis of Question 1, the credit risk of 302 enterprises in Annex 2 is quantitatively analyzed, and the credit strategy of the bank to these enterprises is given when the annual total credit is 100 million yuan.

**Question 3:** The production, operation and economic benefits of enterprises may be affected by some unexpected factors, and the unexpected factors often have different impacts on different industries and different types of enterprises. Considering the credit risk of each enterprise in Annex 2 and the influence of possible unexpected factors (such as SARS-CoV-2 outbreak) on each enterprise, the credit adjustment strategy of the bank is given when the annual total credit is 100 million yuan.

Data description in the annex:

1. Input invoice: the invoice issued by the seller when the enterprise purchases goods (purchases products).
2. Output invoice: the invoice issued for the buyer when the enterprise sells products.
3. Valid invoice: an invoice issued for normal trading activities.
4. Voided invoice: After invoicing the transaction, the transaction was cancelled for some reason, and the invoice was voided.
5. Negative invoice: a negative invoice that needs to be issued when the enterprise has entered the account and recorded tax after invoicing the transaction, and then the buyer returns and refunds for some reason.
6. Credit rating: If the bank evaluates manually according to the actual situation of the enterprise, the bank will not lend to the enterprise with a credit rating of D in principle.
7. Customer churn rate: the rate at which banks lose potential customers due to factors such as loan interest rate.
2. Problem analysis
To solve problem one, feature extraction is carried out according to the attachment data, and the enterprise multi-index scoring model based on XGBoost algorithm is established by using the extracted features. The credit risk of 123 enterprises with credit records is quantitatively analyzed, and each enterprise is scored. According to the quantitative analysis results, credit rating and Annex 3 data, the bank's credit strategy for these enterprises is given in detail.

To solve the second problem, based on the first problem, an enterprise multi-index classification scoring model based on XGBoost algorithm is established. Taking the data in Annex 1 as the training set of the model, the credit risk of 302 enterprises without credit records is quantitatively analyzed and the credit rating is predicted, and each enterprise is also scored. According to the scoring results, credit rating and the data in Annex 3, the bank's credit strategy for each enterprise when the annual total credit is 100 million yuan is given in detail.

In view of the third question, the industries are classified on the basis of the second question, and the credit risks of various enterprises and the different impacts of possible unexpected factors on different industries and different types of enterprises are comprehensively considered. Then, the credit risks of various enterprises are quantitatively analyzed and the credit rating is predicted, and each enterprise is scored. According to the scoring results, credit rating and the data in Annex 3, the bank's credit strategy for each enterprise when the annual total credit is 100 million yuan is given in detail.

3. Model hypothesis
In order to make the model simpler and more reliable, this paper makes the following assumptions:
(1) The invoice information in the annex completely records all the transactions of the enterprise in the corresponding time period;
(2) Questions 1 and 2 During the research period, all companies were operating normally, regardless of the impact of unexpected factors;
(3) The influence of output tax and input tax is not considered when analyzing the profitability of enterprises;
(4) The expected loan amount of all enterprises is greater than or equal to 1 million yuan;
(5) In principle, banks will not lend to enterprises with a credit rating of D.
(6) The attached data is authentic.

4. Symbol description

| Symbol | Description | Unit |
|--------|-------------|------|
| $E_i$  | Enterprise $i$ |      |
| $W_{ji}$ | Total input effective amount of Enterprise $i$ | Yuan |
| $W_{xi}$ | Total effective amount of output of Enterprise $i$ | Yuan |
| $T_i$  | Average number of transactions per month for Enterprise $i$ | Times/months |
| $M_i$  | Total number of trading months for Enterprise $i$ | Deal |
| $I$    | Annual interest rate of bank loans |      |
| $L$    | Customer churn rate |      |
| $E_i$  | The loan amount granted by the bank to Enterprise $i$ | Ten thousand yuan |
| $F$    | Scoring results of enterprise risk quantitative assessment | Points |
5. Model establishment and solution

5.1. Model establishment and solution of problem 1

5.1.1. Analysis of problems.

The first problem requires a quantitative analysis of the credit risk of 123 enterprises, and gives the credit strategy of the bank to each enterprise when the annual total credit is fixed. When banks formulate credit strategies, they need to consider factors such as enterprise strength, stability of supply and demand, enterprise profitability and enterprise reputation. The credit rating of the enterprise has been given in the title annex. And several other factors can be obtained from the quantitative indicators related to invoice data. Therefore, 15 indicators are introduced, including the total number of valid invoices, invalid invoices, negative invoices, effective amount of input/output, proportion of negative invoices, total number of trading months, average monthly transaction times of input/output and average monthly transaction amount of input/output. Based on XGBoost algorithm, the enterprise multi-index scoring model is established with the above 15 indicators, and the risk is quantified according to the enterprise score and loan amount. Based on the basic assumption of rational economic man, banks, as factor owners, should pursue profit maximization. Therefore, taking interest rate and loan amount as decision variables, and the expectation of bank profit as objective function, a nonlinear optimization model with two variables and one objective is established for each enterprise, and the credit strategy of the bank to each enterprise is obtained when the annual total credit is fixed.

5.1.2. Definition and analysis of enterprise rating index.

(1) Enterprise scale index

According to the data in Annex I, the transaction information of input and output bills respectively reflects the trading ability of the company, and reflects the influence of the company on upstream and downstream enterprises to a certain extent. The effective input amount $W_{ji}$ of company $i$ is defined to satisfy:

$$W_{ji} = \sum_{n=1}^{N} w_{jn}$$

where $w_{ji}$ is the total value of price and tax of the $n$th input invoice of the $i$th company after excluding invalid invoices and negative invoices, and $N$ is the total number of the above invoices of the $i$th company. Because there is a difference between the average monthly transaction times $T_{ji}$ of different companies and the total number of transaction months $M_{ji}$, the average monthly effective transaction amount of the input is obtained by taking the month as the calculation unit:

$$\bar{W}_{ji} = \frac{W_{ji}}{M_{ji}}$$

In the same way, the effective output amount $W_{xi}$ of company $i$ satisfies:

$$W_{xi} = \sum_{n=1}^{N} w_{xn}$$

In which $w_{xn}$ is the total value of tax and price of the $n$th output invoice of the $i$th company excluding invalid invoices and negative invoices, and $N$ is the total number of the above invoices of the $i$th company. The average monthly effective transaction amount of the output is obtained by taking the month as the calculation unit:

$$\bar{W}_{xi} = \frac{W_{xi}}{M_{xi}}$$

According to the transaction amount of input and output, it can be used as an index to measure the scale and strength of an enterprise.
(2) Stability index of supply-demand relationship

Voided invoices and negative invoices reflect the return and exchange of company customers and the return and exchange of goods purchased by the company itself. In this paper, the number and proportion of invoices are chosen as indicators to measure the stability of supply-demand relationship of enterprises. According to statistics, the total number of all kinds of invoices can be obtained, so the proportion of the total number of invoices with negative input $M_{j, negative}$ meets the following requirements:

$$R_{j, negative} = \frac{M_{j, negative}}{n_{total}}$$

Average monthly transaction times of input:

$$T_{ji} = \frac{n_{total}}{M_i}$$

In the same way, the proportion of the total number of negative invoices of company $i$ $M_{x, negative}$ is as follows:

$$R_{x, negative} = \frac{M_{x, negative}}{n_{total}}$$

Average monthly transactions of output:

$$T_{xi} = \frac{n_{total}}{M_i}$$

The abnormal situation of buying and selling returns of companies with stable supply-demand relationship is relatively small. Therefore, the frequency and proportion of abnormal bills can be used as an index to measure the stability of supply-demand relationship of enterprises.

(3) Profitability indicators

The profitability index can be obtained by making a difference between the average monthly effective transaction amount of output and the average monthly effective transaction amount of input in the enterprise scale index:

$$Y = \bar{W}_{xi} - \bar{W}_{ji}$$

5.1.3. Enterprise multi-index scoring model based on XGBoost algorithm. According to the indicators obtained from data processing, it is necessary to quantitatively analyze the credit risks of 123 enterprises. In this paper, whether an enterprise defaults or not is taken as a dependent variable, and a predictive analysis model is established based on XGBoost to predict the possibility of enterprise defaults, which is taken as the quantitative analysis result of enterprise credit risks. XGBoost algorithm principle is as follows:

It is assumed that the decision tree used is a binary tree, so as to continuously split the features. For example, the current tree node is split based on the $j$th eigenvalue, and it is assumed that the sample whose eigenvalue is less than $s$ is divided into the left subtree, and the sample whose eigenvalue is greater than $s$ is divided into the right subtree, that is:

$$R_1(j, s) = \{x \mid x^{(j)} \leq s\}$$

$$R_2(j, s) = \{x \mid x^{(j)} > s\}$$

XGBoost predictive analysis model constantly adds trees and splits features to grow a tree. Every time you add a tree, you will learn a new function and fit the residual error of the last prediction. When $K$ trees are obtained after training, the sample features will fall on a leaf node corresponding to each tree, and each leaf node will correspond to a score. Finally, the predicted value of the sample can be obtained by summing the scores corresponding to each tree. Defined as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \ f_k \in \mathcal{F}$$

Where $K$ is the number of trees and $f(x)$ is a function in the function space?

$$\mathcal{F} = \{f(x) = \omega_{q(x)}\}, \ (q: R^m \rightarrow T, \ \omega \in R^T)$$
\( q(x) \) means splitting sample \( x \) to a certain leaf node, and \( \omega \) is the score corresponding to the leaf node. XGBoost uses the following objective functions:

\[
\text{Obj}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k} \Omega(f_k)
\]

\[
\Omega(f) = \gamma T + \frac{1}{2} \lambda \| \omega \|^2
\]

Using square error as error function, there are:

\[
l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2
\]

After that, supervised learning is carried out for the proposed model, and the specific process is as follows:

\[
\hat{y}_i^{(0)} = 0
\]

\[
\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i)
\]

\[
\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i)
\]

\[
\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\]

Starting from the constant term prediction, a new function is added in each training cycle, and the related prediction is carried out in each round to obtain the predicted value. Insert \( f_t \) into the objective function to optimize the objective, so as to minimize the objective function:

\[
\text{Obj}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i)
\]

The formula is concretized, and the residual error in the formula is the difference between the predicted value and the true value in each round, and is substituted into the loss function to obtain:

\[
\text{Obj}^{(t)} = \sum_{i=1}^{n} (y_i - (\hat{y}_i^{(t-1)} + f_t(x_i)))^2 + \Omega(f_t) + \text{const}
\]

\[
= \sum_{i=1}^{n} [2(\hat{y}_i^{(t-1)} - y_i) f_t(x_i) + f_t(x_i)^2] + \Omega(f_t) + \text{const}
\]

To facilitate the calculation, Taylor expansion is used to approximate the objective function:

\[
\text{Obj}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i))^2 + \Omega(f_t) + \text{const}
\]

Taylor expansion is used to replace the original objective function approximately, and Taylor expansion satisfies:

\[
f(x + \Delta x) = f(x) + f'(x) \Delta x + \frac{1}{2} f''(x) \Delta x^2
\]

Definition:

\[
g_i = \partial \hat{y}^{(t-1)} l(y_i, \hat{y}^{(t-1)}), h_i = \partial^2 \hat{y}^{(t-1)} l(y_i, \hat{y}^{(t-1)})
\]

The objective function satisfies:

\[
\text{Obj}^{(t)} \approx \sum_{i=1}^{n} \left[ l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + \text{const}
\]

When the constant term is not considered, the expression in brackets in the above formula is composed of the sum of the error function and a first derivative and a second derivative. According to the \( t - 1 \) tree obtained before, only the first derivative and the second derivative of the first \( t - 1 \) functions need to be solved, and the model of the \( t \)-th tree can be obtained.

Next, the objective function is modified, and \( I \) is defined as the sample set of each leaf node, \( i \) is the \( i \)-th sample, and \( j \) represents the \( j \)-th leaf node:

\[ I_j = \{ i \mid q(x_i) = j \} \]

The quadratic function \( T \) is defined so that the number of \( T \) is the same as the number of trees obtained before, thus obtaining the modified objective function:
According to the obtained quadratic function, define:

\[ G_j = \sum_{i \in I_j} g_i, \quad H_j = \sum_{i \in I_j} h_i \]

Taking \( \omega \) as an unknown variable, assuming that the structure of the tree is a binary tree, a new objective function is obtained:

\[
\text{Obj}_{t}^{(t)} = \sum_{j=1}^{T} \left[ \left( \sum_{i \in I_j} g_i \right) \omega_j + \frac{1}{2} \left( \sum_{i \in I_j} h_i + \lambda \right) \omega_j^2 \right] + \gamma T
\]

According to the objective function, the optimal solution of \( \omega \) is solved, and its derivative can be obtained as follows:

\[
\omega_j^* = -\frac{G_j}{H_j + \lambda}, \quad \text{Obj}_{t}^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \lambda T
\]

Based on the above principles, all the 15 indexes related to enterprise scale, stability of supply and demand relationship and profitability in 5.1.2 are used to quantitatively analyze the 123 enterprises in Annex I, and a comprehensive index with a size between 0 and 1 is output. If the obtained index is greater than 0.5, it is judged that the enterprise does not default; Otherwise, it is judged as default.

104 data were randomly selected as training sets in Annex 1. The XGBoost algorithm is adopted, the depth parameter of the tree is set to 5, the number of trees is set to 300, the training set is put into the classifier, the corresponding labels are searched for classification learning, and the remaining 19 data are taken as test sets, and the accuracy of the model is evaluated, and the accuracy is 89.47%, so the model is reasonable and feasible.

The index is multiplied by 100 and converted into a percentage system as the result of credit risk quantification. The higher the score, the lower the risk. See appendix for the quantitative score of each company.

5.1.4. Fitting the functional relationship between loan annual interest rate and customer churn rate.

By analyzing the data in Annex III, it is found that the bank customer churn rate \( L \) and the loan annual interest rate \( I \) approximately satisfy the logarithmic function relationship. In Excel, the relationship between the annual interest rate \( I \) of bank loans and the customer churn rate \( L \) under three credit ratings A, B and C is fitted, and the analytical formula and fitting function image are shown as follows:
Figure 1. Fitting curve of relationship between loan annual interest rate and customer churn rate when rating A.

\[ y = 0.669\ln(x) + 2.2386 \]
\[ R^2 = 0.9823 \]

Figure 2. Fitting curve of relationship between loan annual interest rate and customer churn rate when rating B.

\[ y = 0.6506\ln(x) + 2.1576 \]
\[ R^2 = 0.9887 \]
That is, there is an approximate relationship between $L$ and $I$:

$$L = \begin{cases} 
0.669 \ln I + 2.2386, & \text{If the rating is A.} \\
0.6506 \ln I + 2.1576, & \text{If the rating is B.} \\
0.6586 \ln I + 2.168, & \text{If the rating is C.}
\end{cases}$$

The goodness-of-fit $R^2$ of the three formulas are calculated to be 0.9823, 0.9887 and 0.9919, which are very close to 1, indicating that the function fitting effect is good, the regression model is significant, and there is a strong correlation between the annual loan interest rate and the customer churn rate.

5.1.5. Establishment of repayment probability function. Assume that $E$ is the loan amount and $F$ is the scoring result of enterprise risk quantitative assessment. In order to calculate the expectation of the bank’s profit after lending to this enterprise for one year, it is necessary to introduce the enterprise repayment probability function $P(E, F)$. The function should at least satisfy the following conditions: the range of $P(E, F)$ is [0,1]; The function approximately satisfies: $P(E, 100) = 1$, $P(E, 0) = 0$, $P(10,100) = 1$ at the boundary; The function decreases monotonously with $E$ and increases monotonously with $F$.

Due to the complexity of economic behavior, it is difficult to get an accurate expression of repayment probability. However, a function satisfying the above conditions can be constructed, which is positively related to the repayment probability and can be used to approximate the repayment probability. Therefore, the following functions can be defined:

$$P(E, F) = \frac{1}{1 + e^{\frac{E}{10} + \frac{F}{100} - \frac{1}{100}}}$$

$E$ is divided by 10, and $\frac{1}{F}$ by $\frac{1}{100}$ is to normalize the data and eliminate the influence of magnitude and dimension.

The function image of the repayment probability $P$ on the enterprise score $F$ when the loan amount $E$ is 100,000-yuan, 500,000 yuan and 1 million yuan respectively is as follows:

**Figure 3.** Fitting curve of relationship between loan annual interest rate and customer churn rate when rating C.
5.1.6. Establishment of optimization model. The loan interest rate for an enterprise is $I$, the loan amount is $E$, and the risk assessment score is $F$. According to the scoring situation of enterprise credit risk in 5.1.3 and the original credit rating, considering the annual interest rate of bank loans in Annex 3 and the loss rate of customers with different credit ratings, an optimization model is constructed. If the loan amount given to each enterprise is optimized as a decision variable, the problem will be too large to be solved effectively. Therefore, we can solve an optimization problem for each enterprise and get the optimal loan amount and interest rate for it. Taking the loan amount $E$ and interest rate $I$ as decision variables, the bank's profit expectation as objective function, and combining with the subject constraints, the following optimization model is established:

$$\text{max} \quad I \times E \times (1 - L) \times P(E, F)$$

$$s.t. \begin{cases} 
  \frac{1}{1 + e^{10 - \frac{100}{100 - F}}} < 1 \leq 100 \\
  4\% \leq I \leq 15\% \\
  L = \begin{cases} 
    0.669 \ln I + 2.2386, & \text{If the rating is A.} \\
    0.6506 \ln I + 2.1576, & \text{If the rating is B.} \\
    0.6586 \ln I + 2.168, & \text{If the rating is C.} 
  \end{cases}
\end{cases}$$

With $Y_1 = I \times (1 - L)$, the optimization problem is: It can be seen that the objective function is a multiplication of four factors, in which $I \times (1 - L)$ is only related to the decision variable $I$, and $E \times P(E, F)$ is only related to the decision variable $E$. Therefore, in order to maximize the objective function, $I \times (1 - L)$ and $E \times P(E, F)$ only need to maximize. At this time, the optimization problem can be transformed into two independent optimization problems for solving.

With $[y]_1 = I \times (1-l)$, the optimization problem is:

$$\text{max} \quad I \times (1 - L)$$

$$s.t. \begin{cases} 
  4\% \leq I \leq 15\% \\
  L = \begin{cases} 
    0.669 \ln I + 2.2386, & \text{If the rating is A.} \\
    0.6506 \ln I + 2.1576, & \text{If the rating is B.} \\
    0.6586 \ln I + 2.168, & \text{If the rating is C.} 
  \end{cases}
\end{cases}$$

Using MATLAB to solve the problem, the interest rate when obtaining the optimal solution is:

![Figure 4. Function image of repayment probability $P$](image)
That is to say, the enterprises with credit rating A uniformly adopt the interest rate of 5.7762\%, the enterprises with credit rating B uniformly adopt the interest rate of 6.2083\%, and the enterprises with credit rating C uniformly adopt the interest rate of 6.2446\%. This shows that banks give preferential interest rates to enterprises with high reputation, which is consistent with real life experience.

With $Y_2 = E \times P(E, F)$, the optimization problem is:

$$
\begin{align*}
\text{max} & \quad E \times P(E, F) \\
10 \leq E & \leq 100
\end{align*}
$$

$$
P(E, F) = \frac{1}{1 + e^{10 \times (E - 100)/(100 - F)}}
$$

Substitute the $f$ value obtained in 5.1.3, and solve it by MATLAB, and get the specific loan amount of each company, see Appendix. At this time, the total annual credit of the bank is 83.255 million yuan.

To sum up, when the fixed value of the annual total credit is equal to 83.255 million yuan, the bank obtains the maximum profit. At this time, the interest rates of different credit grades are 5.7762\% for A grade, 6.2083\% for B grade and 6.2446\% for C grade respectively. See the appendix for the specific loan amount of each company. When the fixed value of the annual total credit is more than (or less than) 83.255 million yuan, the total income is not the maximum.

5.2. Model establishment and solution of Problem 2

5.2.1. Analysis of Problems. Question 2: On the basis of Question 1, this paper makes a quantitative analysis of the credit risks of 302 enterprises in Annex 2, and gives the bank's credit strategy for these enterprises when the annual total credit is 100 million yuan. Take the data in Annex 1 as a training set for supervised learning, and use the learned model to score the enterprises in Annex 2 and grade A, B, C and D. Using the optimization model in the first question, and making specific credit rules, the results meet the limit conditions of total credit, determine the interest rate and loan amount of each enterprise, and get the specific credit strategy.

5.2.2. Enterprise multi-index classification scoring model based on XGBoost algorithm. XGBoost algorithm is adopted, and the depth parameter of trees is set to 5, and the number of trees is set to 300. Since the reliability of the algorithm has been verified in question one, the data originally used as the test set in annex one is also put into the training set, the training set is put into the classifier, the corresponding labels are searched for classification learning, and the scores of 302 enterprises in annex 2 are output.

Then, using all 15 indexes related to enterprise scale, stability of supply and demand relationship and profitability in 5.1.2, and using the same algorithm, the credit rating A, B, C and D of the enterprise is predicted, and the distribution of A, B, C and D is shown in the annex.

5.2.3. The establishment of the optimization model of the second problem. According to the enterprise credit ratings A, B, C and D predicted above, using the same ideas as the first question, an optimization problem is constructed for each non-D enterprise:

$$
\max \quad I \times E \times (1 - L) \times P(E, F)
$$
\[
P(E, F) = \frac{1}{1 + e^{\frac{100}{100 - F}}} \quad 10 \leq E \leq 100
\]

\[
s.t. \begin{cases}
P(E, F) = \frac{1}{1 + e^{\frac{100}{100 - F}}} \quad 10 \leq E \leq 100 \\
0.669 \ln I + 2.2386, \quad \text{if the rating is A.} \\
0.6506 \ln I + 2.1576, \quad \text{if the rating is B.} \\
0.6586 \ln I + 2.168, \quad \text{if the rating is C.}
\end{cases}
\]

According to the calculation, the total loan amount is about 110 million yuan, which exceeds the limit of 100 million yuan. Therefore, it is necessary to stipulate some loan limits, that is, change the value range of \( E \) in the limit conditions of optimization problems.

According to Pareto's law in economics, in any specific group, the important factors usually account for only a few, while the unimportant factors account for the majority, so as long as the important factors can be controlled, the overall situation can be controlled. Usually, a small number of high-quality customers will contribute most of the profits. Therefore, when formulating rules, we should give maximum interest rate concessions and loan quota support to Class A customers, so as to ensure that such customers can be retained. For example, if the upper limit of loans for Class A enterprises is still 1 million, the loans for Class B enterprises shall not exceed 800,000, and the loans for Class C enterprises shall not exceed 600,000. That is, the restriction on \( E \) is changed to:

\[
10 \leq E \leq \begin{cases}
100, & \text{Level A} \\
80, & \text{Level B} \\
60, & \text{Level C}
\end{cases}
\]

By calculation, the total loan amount at this time is about 100.9 million yuan, which basically meets the requirements of the project. See appendix for the optimal loan amount and interest rate for each enterprise under this restriction.

5.3. Model establishment and solution of problem 3

5.3.1. Analysis of problems. In view of the third problem, according to the actual situation, the production and operation of enterprises will be affected by certain unexpected factors. First of all, based on natural language, enterprises are classified into different categories (for example, self-employed, engineering, food, transportation, science and technology, trade, medical care, service, cultural media), and the proportion of each type of enterprise is as follows:

![Figure 5. Pie chart of different types of enterprises](image-url)
Under the influence of the sudden factors of SARS-CoV-2 epidemic, the corresponding economic benefits of enterprises will be affected, so the index data should be reconstructed, and then the credit risk of enterprises should be quantitatively analyzed and the credit rating of enterprises should be re-evaluated based on the XGBoost prediction analysis model in question 2. According to the revised data, the enterprises are prioritized again.

According to the score and credit rating of the enterprise, the optimization model is adopted to adjust the credit adjustment strategy of the bank when the annual total credit is 100 million yuan.

5.3.2. XGBoost model solution. According to the National Bureau of Statistics data [1-8], the impact of the epidemic situation on different enterprises in the first quarter of 2020 is obtained. The transaction volume of self-employed households decreased by 9.4% year-on-year; The transaction volume of engineering enterprises decreased by 17.7% year-on-year; The transaction volume of food enterprises decreased by 15.1% year-on-year; The transaction volume of transportation enterprises decreased by 15.8% year-on-year; The transaction volume of technology enterprises decreased by 8.18% year-on-year; The transaction volume of service enterprises decreased by 1.8% year-on-year; The turnover of cultural media enterprises decreased by 14.3% year on year. According to the above data, adjust the output indicators in Annex 2 and re-extract the characteristic indicators. Using XGBoost enterprise multi-index classification scoring model and new characteristic index value in question 2, the quantitative analysis of credit risk and the prediction of credit rating of enterprises are carried out again.

5.3.3. Solution and analysis of optimization model. According to the score and reputation grade obtained after considering the impact of epidemic situation, the optimization model of problem 2 is adopted to solve:

\[
\begin{align*}
\text{max} & \quad I \times E \times (1 - L) \times P(E,F) \\
\text{s.t.} & \quad P(E,F) = \frac{1}{1 + e^{\frac{E - 100}{100-F}}} \\
& \quad 10 \leq E \leq \begin{cases} 
80, & \text{Level A} \\
60, & \text{Level B} \\
4\% \leq I \leq 15\% 
\end{cases} \\
& \quad L = \begin{cases} 
0.669\ln l + 2.2386, & \text{If the rating is A} \\
0.6506\ln l + 2.1576, & \text{If the rating is B} \\
0.6586\ln l + 2.168, & \text{If the rating is C} 
\end{cases}
\end{align*}
\]

See appendix for the adjusted credit strategy. The total loan amount is 58,844,200 yuan, which meets the limit that the total loan amount in the title is less than or equal to 100 million yuan.

Excel is used to draw the box diagram of the bank loan amount to each enterprise before and after the epidemic situation:
It can be seen from the calculation results and Figure 6 that under the influence of epidemic situation, banks need to cut the loan amount to enterprises in order to obtain the maximum benefit. From life experience, it can be known that sudden factors similar to epidemic will lead to a sharp decrease in the profits of most enterprises and increase the uncertainty of economic operation. At this time, banks tend to keep their own assets in order to avoid risks, which is consistent with the calculation results.

6. Model evaluation and promotion

6.1. Advantages of the model
1. XGBoost algorithm in machine learning theory is adopted in this paper, which can select features by itself and adapt to various loss functions, and does not need to normalize features, which is fast, efficient and fault-tolerant.
2. The enterprise multi-index scoring classification model is established, and the enterprise credit risk is analyzed accurately and quantitatively. After training, the accuracy of this model is approximately 90%, which has high credibility.
3. The indexes extracted in this paper can fully reflect the size of the company, the stability of the relationship between supply and demand, and the profitability of the enterprise, which is practical and scientific.
4. When this paper puts forward the credit strategy, it comprehensively considers the enterprise rating and the enterprise reputation rating, and considers the relationship between the loan amount and annual interest rate and the customer churn rate in the optimization objective function. The credit strategy given is very targeted.

6.2. Defects of the model
1. XGBoost will split the leaves in the same layer when building trees, which is not easy to over-fit. However, if there are too many leaf nodes, the splitting gain will be low, which will lead to extra overhead.
2. When XG Boost trains the model, the parameter setting is cumbersome, so it is necessary to constantly adjust the parameter optimization model.

6.3. Improvement and Popularization of Model
1. An efficient algorithm can be found to optimize the credit strategies of all enterprises.
2. According to the actual situation, when unexpected factors appear in the third question, in order to maintain healthy economic development and social stability, banks should not only maximize their own profits, but also consider providing certain credit assistance to enterprises with operational
difficulties to help them tide over the difficulties. At this time, it is necessary to adjust and optimize the existing credit strategies.

References
[1] Source: National Bureau of Statistic
[2] http://www.stats.gov.cn/tjsj/zxfb/202004/t20200419_1739678.html
[3] http://www.stats.gov.cn/tjsj/zxfb/202004/t20200417_1739328.html
[4] http://www.stats.gov.cn/tjsj/sjjd/202003/t20200327_1735115.html
[5] Explanation of XGBoost algorithm principle
[6] https://blog.csdn.net/u013714645/article/details/73609230
[7] http://blog.itpub.net/31542119/viewspace-2199549/
[8] https://www.cnblogs.com/wj-1314/p/9402324.html