Harsanyi-transformation Oriented Default Risk Prediction Based on FA-XGBoost in P2P Network Loan

Yicheng Gong¹, Lianying Zhou* and Yanna Zhang¹
¹College of Science, Wuhan University of Science and Technology, Wuhan, Hubei Province, China, 430065
* Lianying Zhou’s e-mail:13545256397@163.com

Abstract. In the traditional Harsanyi transformation, the virtual player "Nature" assigns the type of a real player according to the historical data with a certain probability distribution. However, sometimes it is difficult to obtain historical data. Therefore, this paper proposes a FA-XGBoost model to predict the probability distribution of a real player’s type based on a large number of relevant data. In order to eliminate the increase of the computational complexity caused by the large data dimension, factor analysis is firstly utilized to decrease the data dimension, and then the XGBoost algorithm is used to learn the probability distribution of real player’s type. “Nature” finally allocates the player's type according to the predicted probability distribution. This paper makes an empirical analysis based on the real loan data of Lending Club. The results show that the method can well guide the decision-making of P2P loan enterprises.

1. Introduction
Harsanyi transformation is an important method to analyze an incomplete game [1], and the key stage is the virtual player “Nature” selects the type for each real player. Scholars have used it to solve many problems. Xiong Fei et al. analyzed the incomplete information payment matrix based on the degree of individual confidence using the Harsanyi transformation to find its Bayesian Nash equilibrium [2]. Ma Lulu analyzed the problem of adverse selection before venture capital investment [3]. Zhang Pu et al. analyzed the distribution characteristics of stock volatility value and its influencing factors [4]. Li Jun and Li Wei obtained the Bayesian Nash equilibrium solution to determine the type of attacker [5]. Some scholars try to study how “Nature” assigns player’s types. Niu Xiaomeng determined the possibility of football player's kicking direction based on historical data [6]. Gong Yicheng used historical data to guide the randomly decision for group logistics companies [7]. However, Harsanyi transformation is still difficult to practice because of the difficulty in obtaining player's historical data. In 2013, Liu Tieyan and his team first proposed the concept of “game machine learning” [8]. His team eliminate the uncertainty of the game through online data and Markov chain [9-11].

To solve the problem of how "Nature" assigns the players’ type in Harsanyi transformation, this paper try to give a XGBoost based on factor analysis to predict the probability distribution.

2. Basic knowledge

2.1. Harsanyi transformation
An incomplete information static game G has N players can be shown as equation (1) [12].

\[ G = \{ A_1, ..., A_N; T_1, ..., T_n; p_1, ..., p_n; u_1, ..., u_n \} \] (1)
For the i-th player, the type $t_i$ belongs to the type space $T_i$. He will take the behavior $a_i$ from the behavior space $A_i$ with the probability of $p_i$, and gain the benefit $u_i$. The Harsanyi transformation transforms the incomplete information static game into a dynamic game with complete but imperfect information, and then it can be further analyzed by using perfect Bayesian equilibrium.

2.2. Algorithm Introduction

2.2.1. Factor analysis

Factor analysis is to analyze the factors of a number of comprehensive indicators and extract the common factors, the mathematical representation of the factor analysis is shown as Formula (2) [13].

$$
\begin{align*}
    x_1 &= \alpha_{11}f_1 + \alpha_{12}f_2 + \alpha_{13}f_3 + \ldots + \alpha_{1k}f_k + \beta_1 \\
    x_2 &= \alpha_{21}f_1 + \alpha_{22}f_2 + \alpha_{23}f_3 + \ldots + \alpha_{2k}f_k + \beta_2 \\
    x_3 &= \alpha_{31}f_1 + \alpha_{32}f_2 + \alpha_{33}f_3 + \ldots + \alpha_{3k}f_k + \beta_3 \\
    \vdots & \quad \vdots \\
    x_p &= \alpha_{p1}f_1 + \alpha_{p2}f_2 + \alpha_{p3}f_3 + \ldots + \alpha_{pk}f_k + \beta_p
\end{align*}
$$

In the model (2), $(x_1, x_2, x_3, ..., x_p)$ is the original observed variable. $(f_1, f_2, f_3, ..., f_p)$ is a common factor of $(x_1, x_2, x_3, ..., x_p)$. $\alpha_{ij}$ ($i = 1, 2, ..., p; j = 1, 2, ..., k$) is the load of the i-th original variable on the j-th factor. The larger the absolute value of $\alpha_{ij}$, the larger the load of the common factor $f_i$ for $x_i$. $(\beta_1, \beta_2, \beta_3, ..., \beta_p)$ is a special factor of $(x_1, x_2, x_3, ..., x_p)$, which is unobservable.

2.2.2. Introduction to XGBoost Algorithm

XGBoost achieves the generation of weak learners by optimizing the loss function, and uses the first derivative and the second derivative of the loss function to improve the performance of the XGBoost algorithm through pre-sorting and weighted quantile [14].

Let the data set $D = \{(x_i, y_i)\}(|D| = n, x_i \in R^m, y_i \in R)$, CART space $F = \{f(x) = w_q(x)\}(q: R^m \rightarrow T, w \in R^T)$, $q$ is the leaf node coefficient of the tree structure, $T$ is the number of leaves, and each $f_k$ corresponds to an tree structure $q$ and leaf weight $w$. XGBoost algorithm is implemented by minimizing the regularization objective function, as shown in equation (3):

$$
\text{obj}(\theta) = \sum_i^n L(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)
$$

Among them

$$
\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|w\|^2
$$

In equations (3) and (4), $L$ is the loss function, $\Omega$ is the penalty term for the complexity of the model, $\gamma$ is the L1 regularization coefficient, and $\lambda$ is the L2 regularization coefficient.

2.3. Evaluation indicators

For the binary classification problem, four cases of correct prediction are recorded as Table 1, where TP-True Positives; FN- False Negatives; FP- False Positives; TN- True Negatives [15].

Table 1. Four cases of binary classification problem.

|                | Forecast is positive | Forecast is negative |
|----------------|----------------------|----------------------|
| Actually positive | TP                   | FN                   |
| Actually negative | FP                   | TN                   |

The accuracy (A), precision (P), recall (R) and F1 score are defined as shown in equation (5)-(8), a larger A means that the model is better, since P and R can hardly run the maximum value at the same time, a harmonic mean value F1 score is proposed.

$$
\begin{align*}
    A &= \frac{TP+TN}{TP+FP+FN+TN} \quad (5) \\
    P &= \frac{TP}{TP+FP} \quad (6) \\
    R &= \frac{TP}{TP+FN} \quad (7) \\
    F1 &= 2TP / (2TP+FP+FN) \quad (8)
\end{align*}
$$
3. FA-XGBoost driven Harsanyi transformation and its application

3.1. Harsanyi-transformation Oriented Default Risk Prediction Based on FA-XGBoost
In Harsanyi-transformation, the “Nature” is supposed to know the probability distribution of players’ type. However, a new player’s type is usually private, so it’s difficult to perform the Harsanyi transformation. This paper therefore proposes a FA-XGBoost model to learn the probability distribution \( p_t \). Suppose the new player has two types, \( t_1 \) and \( t_2 \), then “Nature” assigns the new player’s type \( t_1 \) according to the probability \( p_t \), and assigns the type \( t_2 \) according to the probability \( 1-p_t \).

Suppose the perfect Bayesian equilibrium corresponds to the probability distribution \( p_e \), comparing \( p_t \) and \( p_e \), players can know which strategy can lead to higher returns and how to make decisions.

3.2. A P2P Network Loan Default Risk Game Model Converted by Harsanyi
In the P2P network loan, a company has two strategies: “Approve” and “Refuse”. In order to obtain higher returns, applicants may hide their real information, so the game is an incomplete information game. This paper constructs a game of incomplete information under certain assumptions.

As shown in Figure 1, the top node is "Nature". It will distribute non-default and default loan applicants with the probability of \( p_g \) and \( 1-p_g \). The following ellipse represents the selection information set of P2P loan enterprises, including two nodes corresponding to natural selection, and the number of terminals represents the revenue of the P2P loan enterprises, when they reach the terminal along the corresponding path. Based on the expected return of the strategies, the perfect Bayesian equilibrium probability \( p_e \) be obtained, as shown in equation (9).

\[
p_e = \frac{(r+2)}{(3r+2)}
\]  

(9)

“Nature” will assign the loan applicant's type based on the probability distribution \( (p_g, 1-p_g) \). If \( p_g = p_e \), P2P loan enterprises chooses “Approve” or “Refuse” with the same benefits; if \( p_g > p_e \), “Approve” will have greater benefits; if \( p_g < p_e \), “Refuse” will have greater benefits.

4. Empirical analysis based on the Lending Club dataset

4.1. Data overview and pre-processing
This article uses the Lending Club’s transaction record in 2018, containing 495242 pieces of data, each containing 144 variables. Since this paper studies whether the loan applicant can finally repay on time, the loan status is selected as the tag value, and the remaining 143 variables are attribute values.

In order to facilitate the subsequent analysis and training of the model, we have done the following four steps to pre-process the original data:

- Screening target status. The loan status has two final states charge off and fully paid. So we delete the other six intermediate states, totaling 429762, the remaining data is 65480.
- Dealing with missing values. 44 variables are deleted, due to more than 40% missing values.
- Delete intermediate variables and irrelevant variables. 25 variables are deleted, and 74 variables remain.
- Quantitative qualitative data. As shown in Appendix A.
After data preprocessing, each sample has 74 variables, which will be used to proceed the research.

4.2. Application of FA-XGBoost Driven Harsanyi Transformation in P2P Network Loan

4.2.1. Factor Analysis Dimensionality Reduction

Based on the data preprocessing in 4.1, factor analysis was used to reduce the dimensionality of the data by SPSS software. After dimensionality reduction, the total variance was shown in Table 2.

| Ingredient | Initial eigenvalue | Extracting the sum of squared loads |
|------------|--------------------|-------------------------------------|
|            | Total Var %        | Acc %                               | Total Var % | Acc %   |
| 1          | 11.288             | 15.255                              | 11.288      | 15.255  |
| 2          | 6.298              | 8.511                               | 6.298       | 8.511   |
| ...        | ...                | ...                                 | ...         | ...     |
| 31         | 0.793              | 1.058                               | 0.793       | 1.058   |
| 32         | 0.752              | 1.003                               | 0.752       | 1.003   |
| ...        | ...                | ...                                 | ...         | ...     |
| 74         | -5.022E-16         | -6.787E-16                          | 100.000     |         |

According to the cumulative percentage of the fourth column in Table 2, the cumulative contribution rate of the first 31 principal components reaches 85.205%, so the 74 original variables are reduced to 31 new features. We named these 31 new features as \( F_1, F_2 \ldots F_{31} \).

4.2.2. FA-XGBoost Driven Harsanyi Transformation

In order to test the effect of building the model, the dimensionally reduced data needs to be divided into a training set and a test set, wherein the training set is used to learn the model, and the test set is used to test the model effect. Simple random sampling is used here, 7:3 is used for training and testing.

Through the training set learning, the XGBoost model gives the absolute value of the regression to each loan applicant of the test set, that is, the probability distribution \( p_t \) of the new player type, and then obtains the probability \( p_g \) of non-default. Based on the actual interest rate of the data, we use the average value of 13.42% as the loan interest rate \( r \). Therefore, the equilibrium probability \( p_e \) can be obtained according to the equation (9), as shown in the equation (10).

\[
p_e = \frac{r+2}{3r+2} = \frac{0.1342+2}{3 \times 0.1342+2} = 0.8883
\]  

So the game matrix based on the FA-XGBoost driven Harsanyi transformation is shown in Table 3.

| ID  | \( p_t \) | \( p_g \) | \( p_e \) | Forecasting strategy | Correct strategy |
|-----|----------|----------|----------|----------------------|-----------------|
| 1   | 0.000192 | 0.999808 | 0.8883   | Approve              | Approve         |
| 2   | 0.000187 | 0.999813 | 0.8883   | Approve              | Approve         |
| ... | ...      | ...      | ...      | ...                  | ...             |
| 19644| 0.000009 | 0.999991 | 0.8883   | Approve              | Approve         |

Comparing the third and fourth columns of Table 3, we can get the prediction strategy of the fifth column, comparing the fifth and sixth columns, we can get the confusion matrix as shown in Table 4.

| Forecasting strategy | Approve | Refuse | Total |
|----------------------|---------|--------|-------|
| Correct strategy     | 16368   | 4      | 16375 |
| Total                | 16372   | 3265   | 3269  |

According to Table 4, there were 16375 applicants who did not default, and the companies chose to approve 16368 applications under the guidance of model. On the other hand, 3269 applications defaulted, and companies used the model chose to refuse 3265 applications, so the learning result is accurate.
Through equation (4)-(7), four evaluation indicators of the FA-XGBoost driven Harsanyi transformation P2P loan applicant default model can be calculated, as shown in equation (11)-(14).

\[
A = \frac{(16368+3265)}{19644} = 0.9994
\]

\[
P = \frac{16368}{16375} = 0.9996
\]

\[
R = \frac{16368}{16372} = 0.9998
\]

\[
F1 = \frac{(2 \times 16368)}{(2 \times 16368+7+4)} = 0.9997
\]

Equation (11)-(14) show that the accuracy is 0.9994 and the F1 score is 0.9997, therefore, the FA-XGBoost driven Harsanyi transformation P2P loan applicant default model established in this paper can help loan enterprises to make a good decision on whether to approve user loans or not.

5. Conclusion
This paper proposes a XGBoost based on factor analysis to solve the problem encountered in the practical application of Harsanyi transformation to help "Nature" assign the type of a player. Firstly, the factor analysis is used to reduce the large amount of relevant data, and then the XGBoost algorithm is used to perform regression prediction on the dimensionality reduction data. Finally, “Nature” allocates the player's type according to the predicted probability distribution. The test results show that the accuracy of the forecast reaches 0.9994. Therefore, this method can effectively help the loan enterprise to make decisions and avoid the default risk of users.

Part of Appendix A (To get full Appendix A, contact the corresponding author)

| Variable   | Replacement value |
|------------|-------------------|
| grade      | A-1,B-2,C-3,D-4,E-5,F-6,G-7 |
| ...        | ...               |
| loan_statu | Fully Paid-0, Charged Off-1 |

References
[1] Harsanyi, J.C. (2004) Games with incomplete information played by "Bayesian" players. I. The basic model. INFORMS.
[2] Xiong F., Liu Y., Si X.M., Cheng H. (2011) Group Decision-Making Simulation with Incomplete Information. System Engineering Theory and Practice, 31(1), 151-157.
[3] Ma L.L. (2012) Game analysis between venture capitalists and entrepreneurs in venture capital. Logistics Engineering and Management (3), 172-175.
[4] Zhang P., Chen L., Cao Q.L. (2018) Research on the Value of Stock Volatility Based on Heterogeneous Beliefs from the Perspective of Information. Management science, 31(2), 147-160.
[5] Li J., Li T. (2019) Information Physical Security Analysis of Smart Grid Based on Bayesian Sequential Game Model. Journal of Automation, 45(01), 100-111.
[6] Niu X.M. (2017) Data driven Hessani Conversion and Its Simulative Application in Penalty Game. (Doctoral dissertation).
[7] Gong Y.C., Zhang Y.N., Zhao J., Yu L., Pan T. (2017) Incomplete Information Game and Simulation of Logistics Supervision.In:the 2017 2nd International Conference on Applied Mathematics, Simulation and Modeling(AMSM2017). Bangkok.pp.165-178.
[8] He D., Chen W., Wang L., Liu T.Y. (2014) A game-theoretic machine learning approach for revenue maximization in sponsored search.
[9] Xu H., Gao B., Yang D., Liu T.Y. (2013) Predicting advertiser bidding behaviors in sponsored search by rationality modeling. Proceedings of the 22nd international conference on World Wide Web. ACM.
[10] Li H., Tian F., Chen W., Qin T., Liu T.Y. (2014). Generalization analysis for game-theoretic machine learning.
[11] Liu T.Y., Chen W., Qin T..(2015) Mechanism learning with mechanism induced data. Twenty-ninth Aaai Conference on Artificial Intelligence. AAAI Press,CA.
[12] Xie S.Y. (2017) Economic Game Theory. Fudan University Press, Shanghai.
[13] Xie H., Xue J., Zhang J. (2012) Research on the Relationship between Broadband Development and Economic Growth in China Based on Factor Analysis. In: Seminar on Broadband China Strategy and Innovation. Zhangjiajie, pp. 165-178.
[14] Chen T., Guestrin C. (2016) XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM.
[15] Li H. (2012) Statistical Learning Method. Tsinghua University Press, Beijing.