Machine Learning-Based Termination Prediction Modeling

Yangha Chun

School of Artificial Intelligence, University Yongin, Yongin-si, Gyeonggi-do, Republic of Korea, yangha00@yongin.ac.kr

This study applied machine learning techniques to a prediction model for contract termination for small and medium-sized enterprises to analyze the factors for contract termination in small and medium-sized enterprises and sought ways to help improve the soundness of loans to small and medium-sized enterprises. The analysis of the causes of termination of existing small and medium-sized business loan agreements was largely based on fragmentary interpretations based on past cases using statistical techniques.

In this study, we use collected small and medium-sized business loan execution data to classify the characteristics of companies with a high likelihood of contract termination among fund applicants, that is, company patterns, and present a technique to predict contract cancellation.

Keywords: Machine Learning, Prediction Techniques, Pattern Classification, Correlation Analysis.

1. Introduction

Around the world, the importance of small and medium-sized businesses to the economy is being recognized, and various measures to support their growth are being discussed. This is because there is a more forward-looking expectation that small and medium-sized enterprises will not simply play a secondary role in the economic structure, but will provide quality jobs and establish themselves as the main axis of growth. Amid these expectations, the policy of supporting small and medium-sized enterprises has been steadily continued for the balanced development of the national economy and the establishment of a healthy business ecosystem, and in 1978, the government established the 'Small and Medium Business Startup and Promotion Fund (hereinafter referred to as the Mid-sized Enterprise Fund)' in accordance with the 'Small and Medium Business Promotion Act'. ' has been established to carry out projects such as policy funding support for small and medium-sized enterprises that are socially underprivileged. Machine learning networks are utilized in plants, animals, and fish sectors for disease detection (Cho et al., 2024; AlZubi, 2023; Wasik and Pattinson, 2024). Moreover, they also play a pivotal role in the manufacturing industry (Porwal, 2024).

In this study, Random Forest, k-NN (K-Nearest Neighbors), Gradient Boosting, Support Vector Machine, and Deep Neural Network, which have been widely used recently, were used to design prediction models. Neural Networks' machine learning techniques were used,

and the program was written in Python.

2. LITERATURE REVIEW

2.1. Policy Finance

Policy financial institutions were established with publicness as a priority in areas where market failures are expected, and because they are public institutions established to compensate for market failures in the financial sector and promote financial publicness, they are sensitive to market trends by their nature. In addition, policy financial institutions have the function of providing low-interest preferential financing for specific projects by directly or indirectly providing financial support to the government to supplement the market and promote publicness.

2.2. Machine Learning-Based Anomaly Detection and Pattern Classification Techniques

Anomaly detection refers to finding objects or data that show a different pattern than expected in the data being analyzed. Expressions such as anomaly, exception, outlier, discordant observation, aberration, peculiarity, and contaminant are used for such entities. Anomaly detection is widely used across society, including fraud, intrusion, safety-critical systems, and military surveillance. Clustering (Jain & Dubes, 1988; Tan et al., 2005) is used for exploratory data analysis and data visualization by forming similar entities into clusters. Clustering is originally an unsupervised technique, but semi-supervised clustering (Basu et al., 2004) has also been recently studied.

The Isolation Forest algorithm was first proposed by Fei Tony Liu, Kai Ming Ting, and Zhi-hua Zhou in 2008. It is one of the unsupervised anomaly detection methods and is mainly used to find outliers in existing data.

It is implemented based on a tree, divides random data, isolates all observations, and has the advantage of being able to operate efficiently even on data with many variables.

Abnormal data has a depth close to the top node, normal data has a depth close to the bottom node of the tree, and normal data has a depth close to the bottom node of the tree. The distance to the end node where a specific entity is isolated is defined as the outlier score, and the shorter the average distance, the higher the outlier score. If normal data is to be isolated, it has a depth close to the terminal node of the tree. Conversely, if abnormal data is to be inserted, it has a depth close to the root node. Figure.1 shows the Isolation Forest operation process.

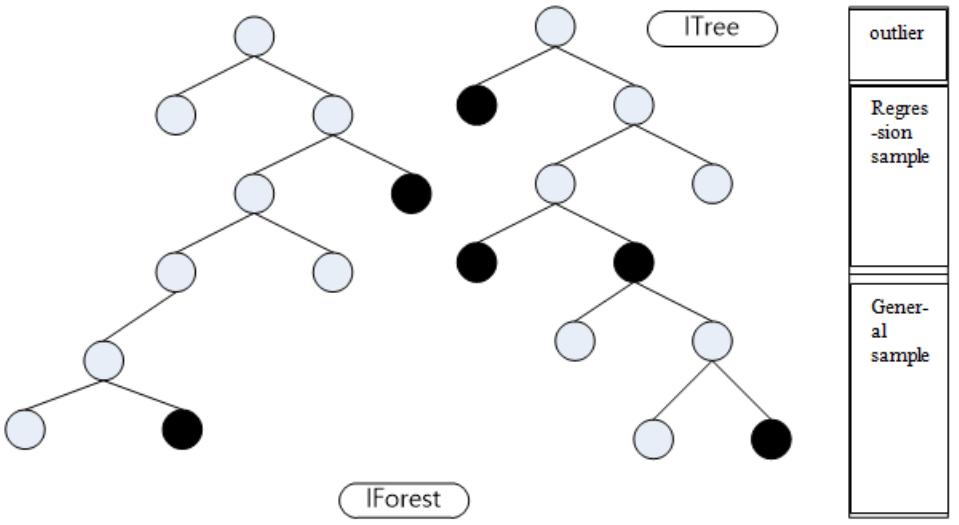

Fig. 1: Isolation Forest operation process

This means that almost all data in a normal distribution falls within a certain distance from the mean of the data set, and the value that measures how close the data is to the mean is called the standard deviation. It only applies to symmetrical, bell-shaped normal distribution curves, and the three cases below are called the 3 Sigma Rule or Empirical Rule.

1. If you use the standard normal distribution, approximately 68% of the values are within 1 standard deviation of the mean.
2. Approximately 95% of the values are within two standard deviations from the mean.
3. Almost all values (approximately 99.7%) are within 3 standard deviations of the mean.

The S-H-ESD algorithm is a method proposed by Twitter, a SNS service, and is a time series outlier detection method performed by combining existing statistical methods. Since the existing anomaly detection method uses standard deviation and average, outliers are already included in the calculation process, so it is not easy to detect outliers, and if the average and standard deviation change in time series and trends, outliers may be missed.

However, the MAD (Median, Absolute Deviation) calculation index and statistical technique used in the S-H-ESD algorithm is suitable for detecting outliers, so ESD uses the STL technique to remove time series and trend effects, leaving only random elements. Currently, the performance evaluation of the R-based S-H-ESD package released by Twitter has been conducted several times, and cases where outliers are detected well include invisible regression values, increase in noise, sudden rise and fall, etc., and outliers are not detected well. As shown in Figure 2, cases where this is not possible include flat signals, gradually increasing signals, and outliers in the direction of gradually increasing signal sounds.

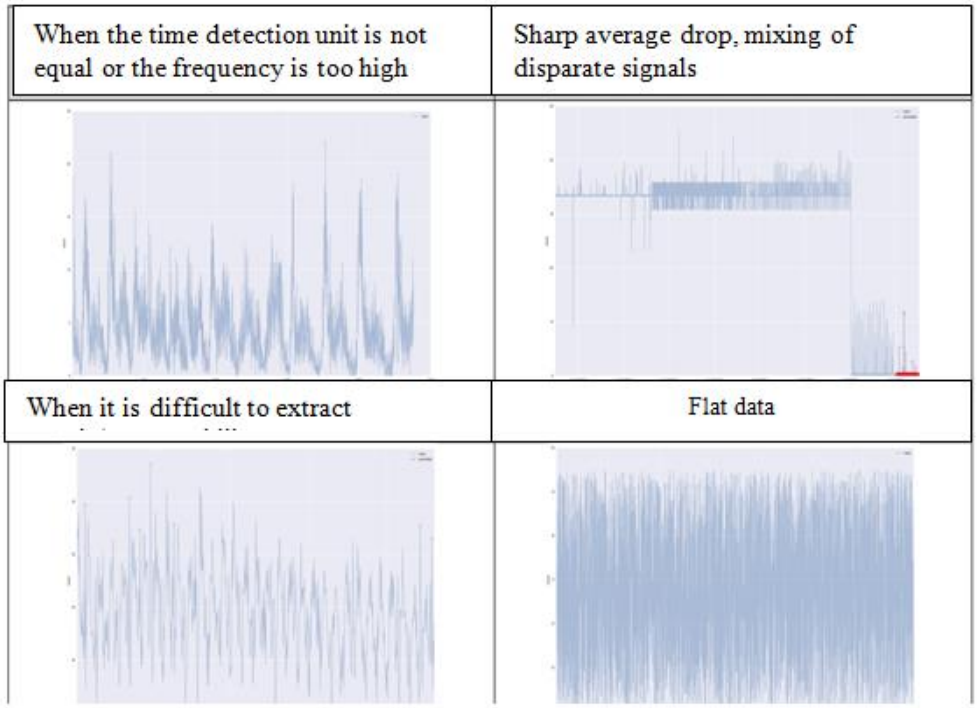

Fig. 2: An example where the S-H-ESD algorithm does not distinguish outliers well

3. METHODOLOGY

As shown in Figure 3, this study aims to design a contract termination prediction model using machine learning techniques targeting the Small and Medium Venture Business Corporation's five-year policy fund data and compare the predicted and actual values.

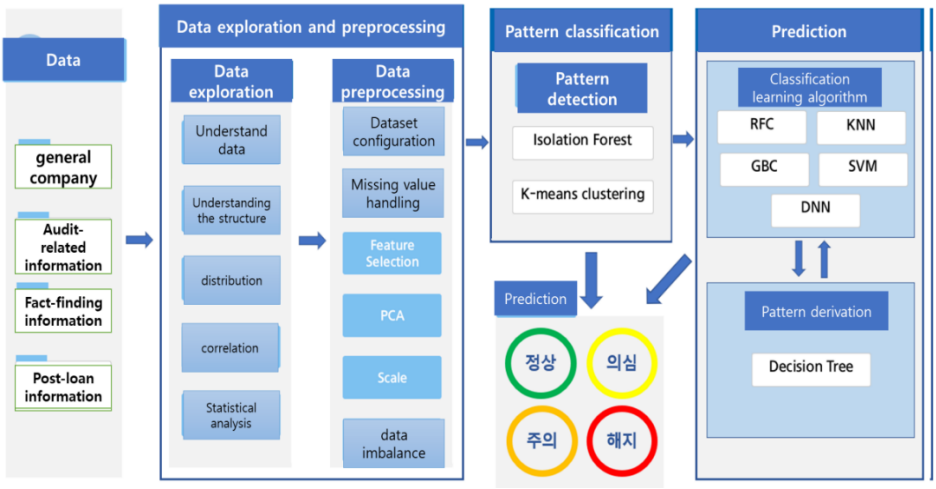

Fig. 3: Machine learning-based contract cancellation prediction model

Since the goal of the research is to classify contract termination patterns based on machine learning and predict the occurrence of contract termination in advance, machine learning techniques are used to classify the characteristics of companies with a high likelihood of contract termination among companies applying for policy funds, that is, company patterns, and We will present a technique for predicting contract termination.

3.1. Data Preprocessing

In relation to machine learning analysis, data preprocessing is an essential process and has a direct impact on model performance along with analysis results (Hyunhwa Song, 2019). Data preprocessing includes data integration, data reduction, data conversion, and data purification. Data conversion is a technology that converts data into a form that is easy for data analysis, such as conversion of data type, and can use methods such as normalization, summary, and hierarchy creation.

Data integration refers to a technology that integrates similar data and linked data to facilitate data analysis, and data reduction is used to reduce data unnecessary for analysis work to increase the efficiency of analysis without compromising uniqueness, such as dimensionality reduction and data compression. , Principal Component Analysis, etc. Data cleaning refers to processing noise data such as missing values, outliers, errors, or variance.

3.2 Data Exploration

To understand the collected data, we conduct exploratory data analysis, check statistics, missing values, outliers, etc., and check the overall distribution of the data. Through this, correlations between variables are obtained and visualized to identify variable tendencies. Data values were adjusted by applying a scaler to normalize the data and correct for differences in values between data. By applying principal component analysis, we extracted the main components of the analysis data, maximized the variance, made the data into a more easy-to-analyze form, compared the results, and then selected and applied them to obtain the results shown in Figure 4.

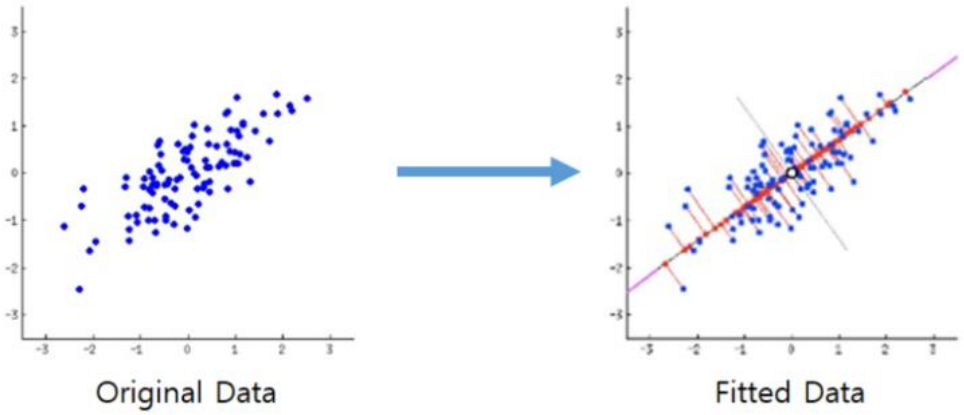

Fig. 4: Example of PCA application

3.3 Predictive Analytics

To construct a data set for predicting contract termination, the Small and Medium Venture Business Corporation's 2019 policy fund company's corporate status, financial and non-financial evaluation, application, rental, and recommended company data were used.

The data used for learning was 70% of the total analysis target data, and for pilot analysis, random forest (RFC), k-nearest neighbor (k-NN), and support vector machine (SVM) machine learning techniques were implemented and applied. The performance of each prediction model was compared using four classification values: total accuracy, precision, recall, and F1 score. The performance of each prediction model was compared using four classification values: total accuracy, precision, recall, and F1 score. The comparison value is expressed as a value between 0 and 1. The closer to 1, the better the performance can be judged as shown in Table 1. As can be seen, the performance of the support vector machine (SVM) model was found to be better than random forest (RFC) and k-nearest neighbors (k-NN).

Table.1: Performance comparison by prediction model

Classification	RFC	k-NN	SVM
accuracy	0.78	0.79	0.81
precision	0.77	0.78	0.80
recall rate	0.78	0.79	0.81
F1 score	0.76	0.78	0.80

4. RESULT AND DISCUSSION

The error matrix and classification report were used to evaluate the classification performance of the prediction model designed in the study. The error matrix is used to visualize the classification results by drawing the actual labels and classification results into a matrix, and the classification report is an indicator that evaluates the performance of the classification model by calculating accuracy, precision, recall, and F1 score. In this study, recall rate in addition to accuracy was used as an evaluation index to examine the performance of classification of contract cancellation data, which is a small number of data.

$$\text{Precision} = \frac{TP(TruePositive)}{TP(TruePositive) + FP(FalsePositive)}$$

As a result of comparing the performance of machine learning techniques compared with data within the classified deferral group, an error matrix was derived as shown in Figure 5.

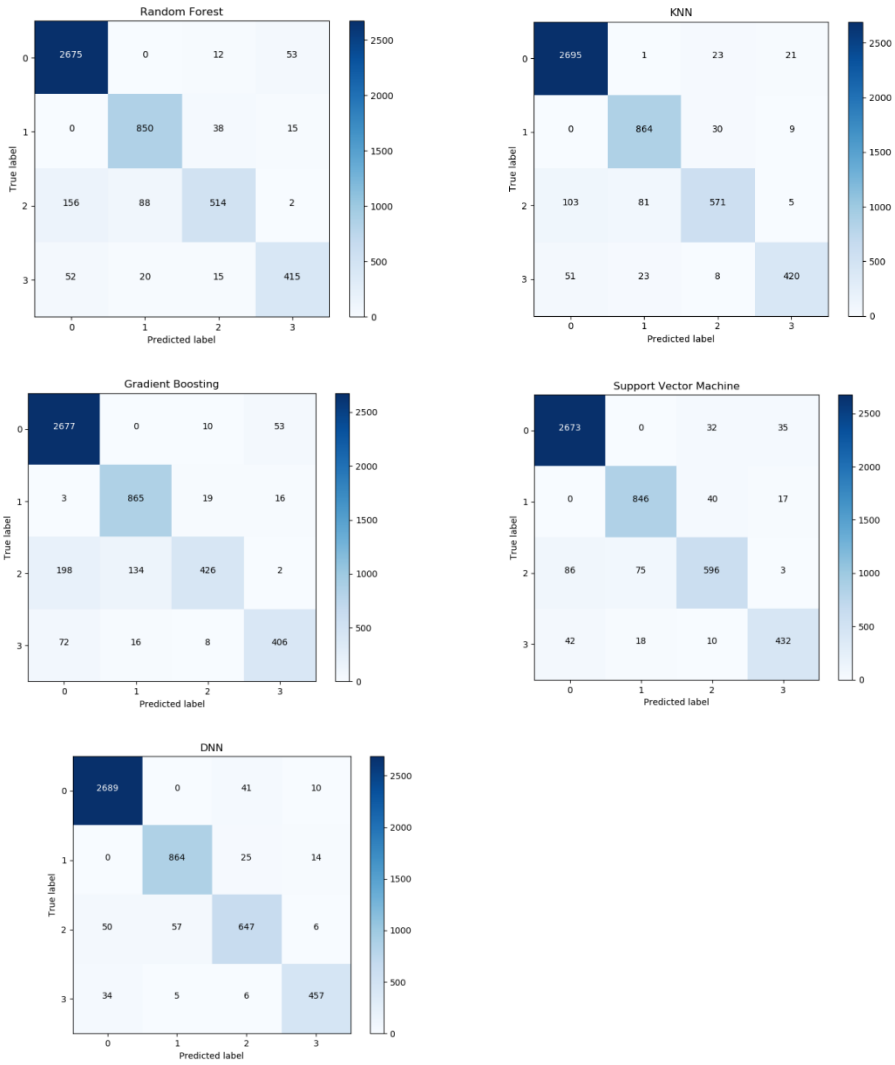

Fig. 5: Error matrix for each deferred group contract termination prediction model

All of the machine learning techniques that compared classification performance showed an accuracy of over 89%, and looking at the results in Table 2, the classification performance of the deep neural network showed the highest accuracy of 95%.

Table.2: Comparison of deferred group classification performance

classification	accuracy	precision	recall rate	F1-score
Random Forest	0.91	0.91	0.91	0.90
k-NN	0.93	0.93	0.93	0.93
Gradient Boosting	0.89	0.89	0.89	0.88

SVM	0.93	0.93	0.93	0.93
DNN	0.95	0.95	0.95	0.95

5. CONCLUSION AND FUTURE WORK

In this paper, we proposed a model that detects and classifies abnormal variables that define contract termination and predicts contract termination using application, recommendation, and rental data for small and medium-sized business policy funds over the past five years. When approaching the contract cancellation rate over the past five years using statistical techniques, the contract cancellation rate during the two-year loan grace period is 1.4%, while the contract cancellation rate during the five-year loan repayment period is a whopping 15%.

Accordingly, in this study, based on machine learning, we separately implemented a prediction model based on the loan grace period and a prediction model based on the loan repayment period in order to analyze in detail the characteristics of companies canceling contracts. When analyzing the results of this study based on pattern rules, the deferral group has variables such as total assets at the time of application for funds, the date of receipt, and the loan processing period, and the repayment group has variables such as the date of corporation establishment or business registration, the date of application, and the date of recommendation. was found to be a significant variable affecting contract termination, which can also be seen as a high correlation between the loan period and contract termination. In addition, it is believed that the variables derived from the pattern rules of this study can be defined as predictive indicators for contract termination and utilized in the organization's data analysis system.

ACKNOWLEDGEMENTS

This study was supported by Yong In University’s 2023 academic research grant

References

1. A Likas, N Vlassis, JJVerbeek. (2003) “The global K-Means Clustering Algorithm” Pattwm recognition, Volume 36, Issue 2 451-461
2. AlZubi, A.A. (2023). Artificial Intelligence and its Application in the Prediction and Diagnosis of Animal Diseases: A Review. Indian Journal of Animal Research. 57(10): 1265-1271. https://doi.org/10.18805/IJAR.BF-1684
3. Aly, M. Survey on multiclass classification methods. Neural Netw, 19, 1-9, 2015
4. Boodhun, N., & Jayabalan,(2018) M. Risk prediction in life insurance industry using supervised learning algorithms. Complex & Intelligent Systems, 4(2), 145-154
5. Breiman. L. (2001), “RANDOM FORESTS. Machine Learning”. 45(1), 5-32
6. Cho, O.H., Na, I.S. and Koh, J.G. (2024). Exploring Advanced Machine Learning Techniques for Swift Legume Disease Detection. Legume Research. https://doi.org/10.18805/LRF-789
7. CW Olanow, RL Watts, WC Koller (2011) “An algorithm (decision tree) for the Management of Parkinson’s disease”, Neurology AAN Enterprises
8. E. Osuna, R Freund, FGirosi (1997) “An improved training algorithm for support vector

machines" IEEE Neural Networks for signal Processing VII
9. Omar, S., Ngadi, A., & Jebur, H. H.(2013) Machine learning techniques for anomaly detection: an overview. International Journal of Computer Applications, 79(2)
10. Porwal, S., Majid, M., Desai, S. C. Vaishnav, J. & Alam, S. (2024). Recent advances, Challenges in Applying Artificial Intelligence and Deep Learning in the Manufacturing Industry. Pacific Business Review (International), 16(7), 143-152.
11. Rawte, V., & Anuradha, (2015) G. Fraud detection in health insurance using data mining techniques. In 2015 IEEE International Conference on Communication, Information & Computing Technology (ICCICT) 1-5
12. Wasik, S. and Pattinson, R. (2024). Artificial Intelligence Applications in Fish Classification and Taxonomy: Advancing Our Understanding of Aquatic Biodiversity. FishTaxa, 31: 11-21.
13. YK Kim, JH Han (1995) "Fuzzy k-NN algorithm using modified K-selection" IEEE International Conference On Fuzzy Sytem
14. Y Freund, L Mason (1999), "The alternating decision tree learning algorithm" icml academia. edu
15. Y. J. Lee, J. H. Nang(2016), "A Personal Video Event Classification Method based on Multi-Modalities by DNN-Learning," Journal of Korea Information Science Society, Vol. 43, No. 11, 1281-1297
16. ZY WU, Y He, Q Li(2018) "Comparing Deep Learning with Statistical Control Methods for Anomaly Detection" Volume 1, WDSA/CCWIJoint Conference