Classification of Credit Risk Data by Using Deep Learning Algorithms

OZER OZDEMIR
Department of Statistics
Eskisehir Technical University
Eskisehir
TURKEY

NESLIHAN TURKOZU
Department of Statistics
Eskisehir Technical University
Eskisehir
TURKEY

Abstract: - It is very important to be able to give the right amount of credit to the right customer, at the right time, in order to effectively manage the increasing credit usage and demands of customers. As a result, it is aimed to increase efficiency in repayment of loans. Along with increasing efficiency, it is aimed to minimize risk and increase profitability. In line with these purposes, Artificial Neural Networks Method, one of the deep learning methods, which is one of the sub-branches of machine learning, was applied in this study in order to determine whether the customer will repay the loan and compared with algorithms such as Logistic Regression and Random Forest.

Key-Words: - Deep learning, Artificial neural networks, Credit risk, Classification algorithms

1 Introduction

In the economy, the financial situations and decisions of companies create some risks. These financial risks; loans, interest rates, exchange rates, cash management and commodity prices. Credit risk is the possible change in net profit and fair value of equity due to non-payment or late payment. It refers to the probability of non-repayment of loans given by banks. Banks deduct the provisions of the non-repayable loans over their incomes. This situation has a reducing effect on the profits of banks [1], [2].

The idea of a machine that can think and resemble a human intellectually was first put forward by Alan Turing [3]. Although a machine that can pass the Turing Test in his article has not been produced yet, computers have become high-performance in many jobs that require human skills [4]. The main reason behind this success is machine learning algorithms, which eliminate the necessity of establishing a mathematical model and train itself according to the data available without pre-programming [5].

The common point of machine learning studies is that the algorithms used are algorithms that are measured with a certain performance system for certain tasks and that gain experience with the performance grade they get while performing these tasks and improve in performing the tasks [6]. Autonomous helicopter driving with machine learning [7], cancer diagnosis [8], language-to-language translation [9], drug design [10] have been carried out in many different areas.

Deep learning involves the use of ANN algorithms and other ANN-based algorithms that mimic the way the brain works, based on the idea that machines can think like humans. It is one of the sub-branches of machine learning. While deep learning algorithms work better on big data systems with complex relationships, they are more expensive in terms of time and money than machine learning algorithms.

The first studies on ANN were made in the 1940s [11]. Laying the foundations of Hebb Theory and the effect of the intensity of connections on the learning processes of neurons were also examined in this period [12].

Deep learning applications are actively used in the finance sector as well as in many other sectors. There are many methods used by banks in the sector to predict the risks of loans they give to their customers.

[13] used Bagging and Boosting (AdaBoost) algorithms, which are collective learning models,

and RO, ANN, DSA methods to classify for risk analysis in their studies. The accuracy rates were 71.06% in AdaBoost model, 69.59% in ANN model, 69.01% in DSA model, 58.50% in RO model and 55.98% in Bagging model.

[14] developed a credit scoring model to evaluate the individual loan application of a customer coming to a bank. The average correct classification rate of the ANN model according to the credit worthiness of customers was 65.3%, and the average correct classification rate of the C5.0 KA model was 61.5%.

[15] tested four different versions of LR, RO, gradient boosting models and stochastic gradient descent DSA in credit risk modeling. As a result of the study, it has been observed that tree-based models are more stable than models based on multilayer ANN based on the AUC (area under the ROC curve) criterion. Gradient boosting and RO algorithms were followed by ANN in the success order.

[16] in his study, for the first time, compared the GKO algorithm with genetic algorithm, ant colony optimization, particle swarm optimization, evolution strategy and population-based incremental learning methods to show that it can be an improvement in training multi-layered sensors. It has been shown that the results obtained with the GKO algorithm are very competitive and achieve high values in zooming.

[17] used Bayesian networks management to identify and compare the potential risks of suppliers by passing supplier parameters through the Bayesian network model.

[18] tried to divide stocks into two classes as good and bad with a collective learning method consisting of LR, RO, DSA and combining RO and DSA in their study. After making feature selection with GA, the accuracy rate of the collective method was the highest with 92.7%. Other accuracy rates were 91.8%, 90.8% and 86.9% for RO, LR and DSA, respectively.

2 Methods

In this section, Artificial Neural Network and Backpropagation methods, which will be used in the analysis phase, will be expressed.

2.1 Artificial Neural Network

ANN examines the training data shown to it, establishes a relationship between these data and makes some generalizations. When faced with new unlabeled test data that he has never seen before, he uses generalizations based on the training data and is able to make inferences about the test data and decide [19].

When ANN is compared with traditional algorithms, while rules are set by using input-output information during learning in ANN, outputs in traditional algorithms; It is obtained by applying the entries to the set rules. While information and algorithms are precise in traditional algorithms, experiences are used in ANNs. However, ANN is slower and hardware dependent [20].

In the general working structure of ANN, the inputs come to the cell with a weight. Although these weights show the effect and importance of the input on the cell, a zero weight does not mean that that input is unimportant and can mean a lot for that network.

With the addition (joining) function, the net input to the cell is calculated. The 'Weighted Sum' is often used as the aggregate function. In the Weighted Sum, each input to the cell is multiplied by its own weight. Then the net input obtained is subjected to the determined activation function. According to the value of the activation function, the output is determined. When choosing the activation function, attention is paid to whether it is easy to calculate its derivative. Since the derivative of the activation function is also calculated in the feedback ANN, care is taken to choose an activation function that will not slow down the processes and whose derivative can be easily calculated. The most commonly used activation function is the 'sigmoid' function [21].

2.2 Backpropagation

Feedback ANN is the use of the GD algorithm in machine learning in ANN. The work of the model by moving from the input layer to the output layer creates the 'feed-forward ANN'. In 'feedback ANN', the output of a feed-forward network is fed back to the network as input to feed the model and performs back propagation. In this way, the weight coefficients are updated again, the error values and the cost function are reduced. The model is optimized.

The following mathematical approach was obtained when a notation suitable for the current machine learning literature [22] was used.

L : Total number of layers in the network
s_l: Number of units in layer l (excluding bias unit)
K : Number of output unit/class

Neural networks can have many output neurons. $h_\theta(x)_k$,, k. is stated as the resulting hypothesis in

the output. The regularization term for ANN is given in the used cost function (1).

$$J(\theta) = -\frac{1}{m}\sum_{t=1}^{m}\sum_{k=1}^{K}[y_k^{(t)}\log(h_\theta(x^{(t)}))_k + \left(1-y_k^{(t)}\right)\log(1-h_\theta(x^{(t)})_k)] + \frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(\theta_{j,i}^{(l)})^2 \quad (1)$$

Nested sums are used to calculate multiple output neurons. The nested sums before the parentheses in the first part of the equation provide as many loops as the number of output nodes.

3 Problem Solution

Home Loan Group is a bank that provides home loans to people. This bank seeks to broaden financial inclusion for people seeking loans by providing a positive and safe borrowing experience. Home Loan Group uses a variety of alternative data, including telecommunications and transaction information, to estimate its customers' repayment capabilities.

The problem is a classification problem. The ultimate purpose of the estimates is to answer the question of whether borrowers will be able to repay their debt.

Training data set is shown in Figure 1.

```
# Training data
app_train = pd.read_csv('../input/application_train.csv')
print('Training data shape: ', app_train.shape)
app_train.head()
```

```
Training data shape:  (307511, 122)
```

Figure 1: Training Data Set

Test data set is shown in Figure 2.

```
# Testing data features
app_test = pd.read_csv('../input/application_test.csv')
print('Testing data shape: ', app_test.shape)
app_test.head()
```

```
Testing data shape:  (48744, 121)
```

Figure 2: Test Data Set

The data set to be used in the training includes 307511 records (rows) and 122 attributes (columns). In addition, the test set to be estimated includes 48744 rows and 121 attribute columns.

Of the 122 columns in the training set, 65 of them can take decimal values and 41 of them consist of numerical variables that can take integer values. The remaining 16 columns are categorical variables of character data type. Figure 3 shows how many categories each of these categorical data has.

```
# Number of unique classes in each object column
app_train.select_dtypes('object').apply(pd.Series.nunique, axis = 0)
```

```
NAME_CONTRACT_TYPE              2
CODE_GENDER                     3
FLAG_OWN_CAR                    2
FLAG_OWN_REALTY                 2
NAME_TYPE_SUITE                 7
NAME_INCOME_TYPE                8
NAME_EDUCATION_TYPE             5
NAME_FAMILY_STATUS              6
NAME_HOUSING_TYPE               6
OCCUPATION_TYPE                18
WEEKDAY_APPR_PROCESS_START      7
ORGANIZATION_TYPE              58
FONDKAPREMONT_MODE              4
HOUSETYPE_MODE                  3
WALLSMATERIAL_MODE              7
EMERGENCYSTATE_MODE             2
dtype: int64
```

Figure 3: Unique Number of Categories of Categorical Variables

After the data set was ready for preprocessing, the categorical variables with 2 categories were converted with Label Encoding. Categorical variables with more than two categories, on the other hand, were turned into a column with One Hot Encoding.

In algorithms using error terms such as logistic regression, n-1 category columns in each column are made into columns, whereas in random forest algorithms, n categories in each column are made into columns.

Label Encoding and One Hot Encoding are preprocessing functions in the Scikit-Learn library, which is a machine learning library.

Some missing values in the data set are filled with the median values of the columns they belong to. In addition, with min-max scaling, the values in the data set were kept in a constant range of 0-1.

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$

Polynomial features can be added when the relationship between variables is not linear. For example, a new feature can be added to the model by including the product of two variables in the model. In this way, multiple features are combined. For this purpose, a significant number of new features were created for the features in the data set. 35 terms were produced by including the 3rd degree powers of the selected 4 features and the terms of interaction with each other. With the generation of polynomial variables, it is possible to see which functions of the variables are related to the target variable and their effects on the model.

Machine learning application was made using Python programming language and Scikit-Learn and Keras libraries.

Random Forest and Logistic Regression algorithms were used to compare with the Artificial Neural Networks model. Although these algorithms are frequently used algorithms in the machine learning literature, they are preferred in practice because they are in the Scikit-Learn library and can be easily used with methods already available in the Scikit-Learn library, such as hyperparameter optimization and k-fold-cross validation.

The Random Forest algorithm, which consists of decision tree-based algorithms, evaluates the predictions made by using random parts of the data set of a determined number of decision trees by voting method and gives output accordingly [23]. Since it uses more than one model, the Random Forest algorithm is also included in the class of collective learning algorithms. The reasons for using the Random Forest algorithm is that this algorithm, which learns the random parts of the data set, is resistant to overfitting and can give good results.

Logistic Regression is widely used in classification problems. There are only two possible outcomes as the outcome is measured with a binary variable. The reasons for using the Logistic Regression algorithm is that the dependent variable is a powerful algorithm for modeling binary data as a categorical variable.

AUC, also known as AUC/ROC (Area Under Curve / Receiver Operating Characteristic) metric, was preferred as the evaluation criterion. It is designed to find the area under the curve formed by the dots with the True Positive Ratio (y-axis) and False Positive Ratio (x-axis) in Figure 4.

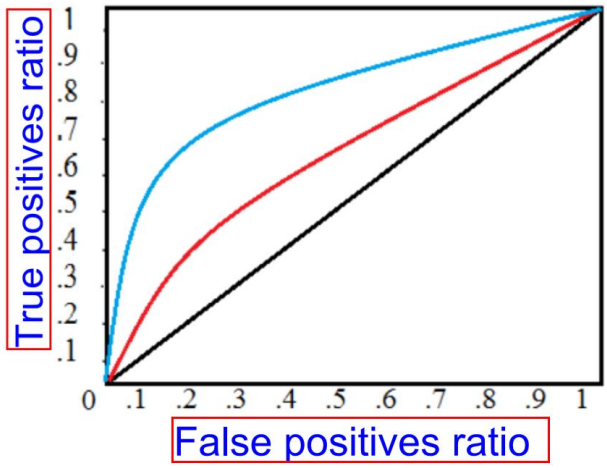

Figure 4: ROC/AUC Graph

The True Positive Ratio indicates the proportion of correctly predicted positive samples among all positive samples, while the False Positive Ratio indicates the proportion of incorrectly predicted positive samples among all negative samples.

Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a certain decision threshold. The area under the ROC curve (AUC) is a measure of how well the two groups can be distinguished.

The closer the ROC curve is to the upper left corner, that is, the larger the area under the curve; It approaches 1, and the overall accuracy of the test is increasing.

$$True\ positive\ ratio = \frac{True\ positives}{positives} \quad (3)$$

$$False\ positive\ ratio = \frac{False\ positives}{negatives} \quad (4)$$

4 Conclusion

Running the Logistic Regression model with the regularization parameter gives better results than the Classical Logistic Regression model. For this reason, C value, which is the regularization parameter in the Logistic Regression model, has been reduced by taking 0.0001. The regularization parameter controls the amount of overfitting. A low C value will reduce overfitting.

In the Random Forest model, predictions were made using 100 different decision trees.

The architecture of the Artificial Neural Network model is shown in Figure 5.

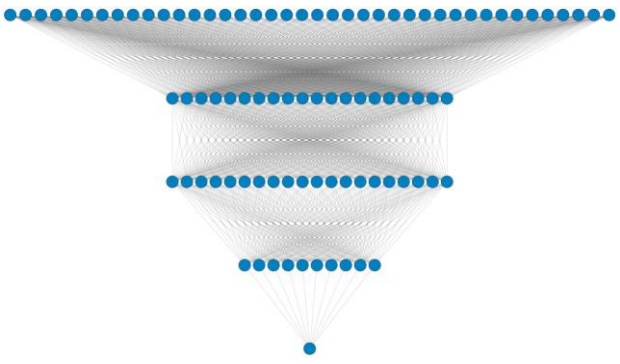

Figure 5: Architecture of ANN model

In the Artificial Neural Network model, there are 1 input layer, 1 output layer and 3 hidden layers

between these layers. There are 277 nodes in the input layer, 80 in the 1st hidden layer, 80 in the 2nd hidden layer, 40 in the 3rd hidden layer, and 1 node in the output layer.
ReLu (Rectified Linear Units) and Sigmoid functions are preferred as activation functions (Shown in Figure 6). Backpropagation was used as the optimizing method.

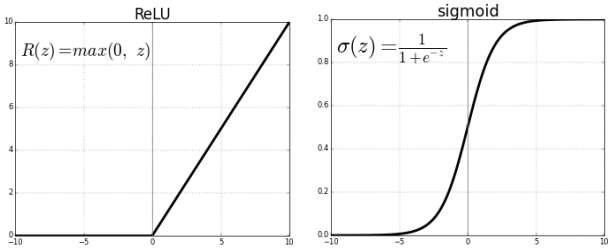

Figure 6: RELU and Sigmoid activation functions

When the parameters are optimized, the iteration number is 30, the loss function is Binary Cross Entropy, and the learning rate is 0.01.
The estimation successes according to the results of the three models applied are given in Table 1.

Table 1: The estimation successes according to the results of the three models

Algorithm	ROC/AUC Value
Logistic Regression	0.67505922
Random Forest	0.70322869
Artificial Neural Network	0.73725195

According to the results obtained, the Artificial Neural Networks model gave the highest result among the applied methods, with an accuracy rate of approximately 74%. This result is followed by the Random Forest model, which makes predictions with 70% accuracy. In the Logistic Regression model, estimation was made with an accuracy rate of 67%. Considering the results, it has been seen that machine learning algorithms can be used for credit and similar problems when sufficient data is collected.

References:

[1] Mandacı, P. E.. Türk Bankacılık Sektörünün Taşıdığı Riskler ve Finansal Krizi Aşmada Kullanılan Risk Ölçüm Teknikleri. *Sosyal Bilimler Enstitüsü Dergisi*, 67-84, 2003.

[2] Uğur, A., & Karaca, S. S.. Türkiye'deki Bankacılık Sektöründe Risk ve Karlılık Analizi. *Muhasebe Bilim Dünyası Dergisi*, 123-134, 2008.

[3] Turing, A. M.. Computing Machinery and Intelligence. *Computers & Thought* (s. 11-35). MIT Press, 1995.

[4] He, K., Zhang, X., Ren, S., & Sun, J.. *Deep Residual Learning for Image Recognition.* arXiv:1512.03385, 2015.

[5] Samuel, A. L.. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 210 – 229, 1959.

[6] Mitchell, T.. *Machine Learning.* McGraw Hill, 1997.

[7] Ng, A. Y., Coates, A., Diel, M., Ganapathi, V., Schulte, J., Tse, B., . . . Liang, E.. Autonomous inverted helicopter flight via reinforcement learning. *Experimental Robotics IX*, 363-372, 2006.

[8] Cruz, J. A., & Wishart, D. S.. Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics 2*, 2006.

[9] Bahdanau, D., Cho, K., & Bengio, Y.. *Neural machine translation by jointly learning to align and translate.* arXiv:1409.0473, 2014.

[10] Burbidge, R., Trotter, M., Buxton, B., & Holden, S.. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers & Chemistry*, 5-14, 2001.

[11] McCulloch, W. S., & Pitts, W. H.. A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, s. 115-133, 1943.

[12] Hebb, D. O.. *The Organization of Behavior: A Neuropsychological Theory.* New York: John Wiley & Sons, 1949.

[13] Hamori, S., Kawai, M., Kume, T., Murakami, Y., & Watanabe, C.. Ensemble Learning or Deep Learning? Application to Default Risk Analysis. *J. Risk Financial Manag*, 2018.

[14] Sönmez, F.. Kredi Skorunun Belirlenmesinde Yapay Sinir Ağları ve Karar Ağaçlarının Kullanımı: Bir Model Önerisi. *ABMYO Dergisi 40*, 1-22, 2015.

[15] Addo, P. M., Guegan, D., & Hassani, B.. Credit Risk Analysis Using Machine and Deep Learning Models. *Risks*, 2018.

[16] Mirjalili, S.. How effective is the Grey Wolf optimizer in training multi-layer perceptrons. *Appliled Intelligence*, 150-161, 2015.

[17] Lockamy, A., & McCormack, K.. Modeling Supplier Risks Using Bayesian Networks. *Industrial Management & Data Systems*, 313-333, 2012.

[18] Fu, X., Du, J., Guo, Y., Liu, M., Dong, T., & Duan, X.. A Machine Learning Framework for

Stock Selection. *arXiv:1806.01743v1 [q-fin.PM]*, 2018.
[19] Ergezer, H., Dikmen, M., & Özdemir, E.. Yapay Sinir Ağları ve Tanıma Sistemleri. *Pivolka*, 14-17, 2003.
[20] Pirim, H.. Yapay Zeka. *Journal of Yaşar University*, 81-93, 2006.
[21] Çayıroğlu, İ.. *İleri Algoritma Analizi-5 Yapay Sinir Ağları. http://www.ibrahimcayiroglu.com/Dokumanlar/IleriAlgoritmaAnalizi/IleriAlgoritmaAnalizi-5.Hafta-YapaySinirAglari.pdf* , 2015.
[22] Ng, A.. CS229 *Lecture notes. CS229: Machine Learning: http://cs229.stanford.edu/notes/cs229-notes-deep_learning.pdf*, 2018.
[23] Breiman, L.. *Random forests. Machine learning*, Springer, 5-32, 2001.