Credit card default prediction by using Heterogeneous Ensemble

Wook Lee
School of Electrical Engineering
Korea University
Seoul, Korea
leewook94@korea.ac.kr

Sangmin Lee
School of Information Convergence
Kwangwoon University
Seoul, Korea
smlee5679@kw.ac.kr

Junhee Seok*
School of Electrical Engineering
Korea University
Seoul, Korea
jseok14@korea.ac.kr

***Abstract*—** Credit card companies calculate an accurate credit score by utilizing the personal information and credit data of new applicants. To analyze and predict credit ratings, there have been many studies using machine learning. However, previous research had limitations in improving prediction accuracy using single algorithms such as ensembles or deep learning and could not consider the problem of multiple histories of the same customer using different cards. This study proposes a hybrid algorithm that combines heterogeneous ensembles and TabNet, a deep learning algorithm specialized in tabular data, to address these issues. The study conducted comparative experiments with several state-of-the-art machine learning algorithms that have been used for credit card delinquency prediction.

Keywords—Gradient Boosting, TabNet, Machine Learning, Deep Learning, Ensemble Learning

I. Introduction

Credit card companies calculate a credit score by utilizing the personal information and individual credit data submitted by credit card applicants for new card issuance. The calculated credit score is used to determine the applicant's future creditworthiness, and credit card companies want to predict a member's delinquency possibility and take preemptive measures for debt management. However, the process from credit score calculation to delinquency prediction and preemptive measures is not automated, and the prediction accuracy is low, making it challenging to implement financial services. Recently, many problems have been solved by using machine learning and reinforcement learning in various fields. [1, 2, 3] Among them, products using deep learning have recently been talked about in the financial market.[4] This study aims to propose an artificial intelligence-based financial service system that accurately predicts whether users will default on their payments.

In this study, we propose a hybrid algorithm of heterogeneous ensemble and TabNet, which is a deep learning algorithm specialized in tabular data, to further improve the prediction performance of credit card delinquency level. We combined various learning models to improve the generalization performance through heterogeneous ensemble and extracted important variables and used TabNet as the final classifier. To evaluate the performance of the proposed method, we conducted comparative experiments with most recent learning algorithms commonly used for predicting credit card delinquency levels. The experimental results prove that the proposed algorithm shows robust prediction performance compared to other algorithms. Finally, to maximize the use of multi-account data for the same customer, we transformed the data into a mixture of continuous and categorical variables and applied the CatBoost model, which further improved the prediction performance.

II. Background

A. Ensemble Learning

Ensemble learning is an algorithm that combines multiple models to improve prediction performance. There three types of ensemble learning. First, stacking is to com.bine several models trained over the same dataset with different algorithms. The output of the models becomes the input of meta model, and the meta model is trained again. Second, bagging is to combine tens to hundreds of models using the same algorithm but trained over different data sets randomly resampled from the original data sets. Bagging is known to stabilize the base model. Thirdly, boosting combines multiple weak learners to create a single strong learner. In tabular datasets, ensemble models still show good performance.[5]

B. TabNet

TabNet is an artificial neural network algorithm that combines the advantages of gradient boosting. The feature transformer performs encoding, and the attentive transformer generates a mask based on the encoded results. The mask includes the activation values of the selected variables that will be used to make predictions. At each step, the encoded results obtained from the feature transformer go through a rectified linear unit (ReLU) function and are linearly combined, and then a fully-connected layer is applied to produce the final prediction value. This TabNet model structure improves prediction performance by utilizing sequential attention mechanism. At each step, feature selection is performed through the mask block to selectively combine the most important variables used at each time point, thereby enabling the model to learn a high-performance generalized model for structured data.[6]

III. Method

A. Data collection and preprocessing

In this experiment, we utilized the "Credit Card Approval Scoring Model" dataset [8]. The dataset classified customers' information and credit scores. We set credit score as the target variable (Y) and conducted experiments accordingly. The dataset had imbalanced data distribution according to the target variable, with 16,968 instances for Y = 2, 6,267 instances for Y = 1, and 3,222 instances for Y = 0. Moreover, the same customer's information had different target variables depending on the credit card, which resulted in the problem of varying credit scores for the same individual. To address this issue, we performed feature engineering by dividing time data, such as date of birth and start date of employment, into year, month, week, and day.

B. Model Architecture

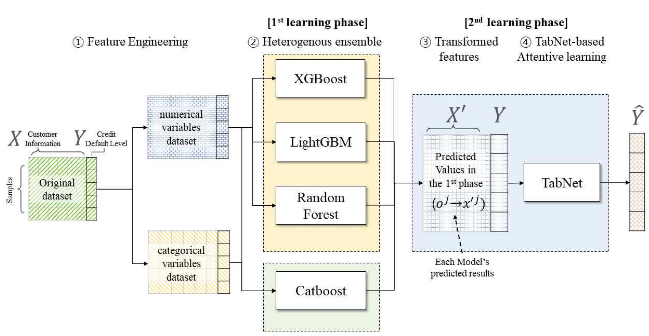

Fig. 1. proposed model architecture.

Fig. 1 is a diagram representing a stacking model that improves performance by using a meta-model to classify the predictions of a single model trained on input data. According to the 1st learning phase, numerical preprocessing was performed using LightGBM, XGBoost, and random forest for training. On the other hand, categorical preprocessing was performed using CatBoost's ordered Boosting technique, which showed better performance [6]. Finally, in the 2nd learning phase, the predictions were used as metadata and TabNet was trained on them. To address the data imbalance problem, the model was validated using stratified k-fold technique [7].

C. Catboost

Algorithm: CatBoost learning for categorical features

Description: Preprocessing categorical features and implementing of a permutation-based ordered boosting was conducted. Here, m denotes a CatBoost model.

Input: Dataset $\mathbb{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
Output: Number of Iteration *T*, Model *M*
Number of Iteration *T*

σ ← random permutation of [1, n]

$m_i \leftarrow 0$ for $i = 1 \ldots n$
for *t* ← 1 to $\mathbb{D}$ do
  for *i* ← 1 to n do
    $r_i \leftarrow y_i - M_{\sigma(i)-1}(x_i)$
    end for
  for i ← 1 to n do
    $\Delta M \leftarrow$ *Update* $CatBoost\ ((x_i, r_i): \sigma_i \leq i)$
    $m_i \leftarrow m_i + \Delta m$
  end for
return *T, M*

It is essential to convert categorical variables into numerical ones for learning. Target encoding, which represents the mean value of Y for each level of the category, is one of the ways to preprocess categorical variables. CatBoost utilizes hold-out techniques for this purpose. The following is the equation for hold-out target encoding.

$$x_k^i = \frac{\sum_{j=1}^{p-1}[x_{\sigma_{j,k}} = x_{\sigma_{p,k}}] \cdot y_j + \alpha P}{\sum_{j=1}^{p-1}[x_{\sigma_{j,k}} = x_{\sigma_{p,k}}] \cdot y_j + \alpha} \qquad (1)$$

Here $x_{\sigma_{j,k}}$ represents the target statistics for each observation, obtained by summing up the target values of each observation and dividing by the total number of observations. However, when computing the first target statistics, there may be no previous observations to use for transformation. To address this issue, a prior value *P* and weight α are multiplied to obtain the target encoding. CatBoost uses a technique called ordered boosting, which splits the target encoding into these two parts and trains the model using a random permutation of the data.[6]

CatBoost uses ordered boosting technique to prevent target leakage and preprocess categorical data for training. The following pseudocode shows the process, where σ produces target encoding values using random permutations. The residual is trained using target encoding, and the value of is updated times. [6]

IV. Experiment

A. Setting

In this experiment, we conducted a comparative study of representative algorithms widely used in multivariate structured data using customer datasets of each credit card. By comparing the model prediction performance, we established a model based on the best performing training model and conducted an analysis of predictions.

B. Evaluation

To compare the prediction performance of the learning models, we used evaluation metrics such as log-loss, accuracy, and f1-score for multiclass classification models.

$$\log loss = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \quad (2)$$

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (3)$$

$$f1 = 2\, x \, \frac{precision\ x\ recall}{precision + recall} \quad (4)$$

C. Results

TABLE 1. presents the experimental results on the structured dataset compiled based on customer information. The performance improvement is demonstrated through the representative evaluation metrics of multi-class classification, including logloss, accuracy, f1-score, precision, and recall. In terms of accuracy, CatBoost outperforms heterogenous ensemble. On the other hand, heterogenous ensemble shows better performance in logloss and f1-score compared to other models. Regarding logloss, CatBoost performs better than other boosting models, likely due to the effective processing of categorical data using the ordered boosting method during training. Heterogenous ensemble achieves the best performance with a logloss of 0.6707, as it incorporates all the prediction values of these boosting models during training. Furthermore, CatBoost also shows higher f1-score than other boosting models, and heterogenous ensemble also demonstrates the highest f1-score of 0.5825.

TABLE I. PERFORMANCE COMPARISON BASED ON EACH MODEL

Algorithm	Metric		
	Logloss	***Accuracy***	***F1-score***
LightGBM	0.6871	0.7307	0.5625
XGBoost	0.6903	0.7296	0.5551
Random Forest	0.6914	0.7291	0.5401
CatBoost	0.6723	**0.7369**	0.5784
Heterogeous Ensemble - TabNet	**0.6707**	0.7362	**0.5825**

V. Conclusion

This paper proposes a method to improve the credit scoring performance using credit card customer data, and to handle duplicated independent variables. Boosting models, which show outstanding performance on structured data, were used, and a heterogeneous ensemble was constructed using TabNet, a high-performance deep learning structured data network. In addition, a preprocessing method was used to consider multiple account data for the same customer in the credit card data, using CatBoost's ordered boosting technique. As a result, the proposed method showed higher prediction performance compared to other models. Furthermore, the heterogeneous ensemble using TabNet demonstrated even more robust performance. Based on this model, it is possible to apply machine learning models to judge the delinquency information of multiple credit cards for an individual, which could potentially be used in practical services.

Acknowledgment

This work was supported by a grant from the National Research Foundation of Korea (NRF-2022R1A2C2004003)

References

[1] Kim I, Lee M, Seok J,(2023) “ICEGAN: Inverse Covariance Estimating Generative Adversarial Network”, Machine Learning-Science and Technology.
[2] Seo J, Kim I, Seok J,(2023) “Grid-wise simulation acceleration of the electromagnetic fields of 2D optical devices using super-resolution”, Scientific Reports.
[3] Hwang H-S, Lee M, Seok J,(2022) “Deep reinforcement learning with a critic-value-based branch tree for the inverse design of two-dimensional optical devices”, Applied Soft Computing.

[4] Yun H, Lee M, Kang YS, Seok J,(2020) “Portfolio Management via Two-stage Deep Learning with A Joint Cost”, Expert Systems with Applications.
[5] Arik, S. O., & Pfister, T. (2019). Tabnet: Attentive interpretable tabular learning. arXiv preprint, arXiv:1908.07442.
[6] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, Andrey Gulin (2019) CatBoost: unbiased boosting with categorical features
[7] R. Bey, R. Goussault, M. Benchoufi, R. Porcher(2020) Stratified cross-validation for unbiased and privacy-preserving federated learning
[8]dataset
https://mp.weixin.qq.com/s/upjzuPg5AMIDsGxlpqnoCg