Incorporating BERT-based NLP and Transformer for An Ensemble Model and its Application to Personal Credit Prediction

Sophot Ky, Ju-Hong Lee, Kwangtek Na

Abstract

Tree-based algorithms have been the dominant methods used build a prediction model for tabular data. This also includes personal credit data. However, they are limited to compatibility with categorical and numerical data only, and also do not capture information of the relationship between other features. In this work, we proposed an ensemble model using the Transformer architecture that includes text features and harness the self-attention mechanism to tackle the feature relationships limitation. We describe a text formatter module, that converts the original tabular data into sentence data that is fed into FinBERT along with other text features. Furthermore, we employed FT-Transformer that train with the original tabular data. We evaluate this multi-modal approach with two popular tree-based algorithms known as, Random Forest and Extreme Gradient Boosting, XGBoost and TabTransformer. Our proposed method shows superior Default Recall, F1 score and AUC results across two public data sets. Our results are significant for financial institutions to reduce the risk of financial loss regarding defaulters.

Keywords : Credit Prediction | Transformer | BERT | Ensemble Modeling | Tabular Data

I. INTRODUCTION

Credit prediction, also known as credit scoring, is a process of assessing an individual' s creditworthiness or likelihood of defaulting on a loan. The word 'Default' in financial terminology refers to status when a borrower fails to make repayment on a loan back on scheduled as stated in the agreement. On the other hand, 'Non-default', means that a borrower repay back the borrowed money including interest on time. Whether in context of personal finance or corporate finance, losses regarding defaulters do more damage to the lenders than losses with respect to non-defaulters. Financial institutions such as banks hire loan specialists for this specific task of evaluating where the loan should be provided or not. However, humans are species with emotions and could make bias decisions. To aid the decision-making process, computational models that can learn from the existing data are equipped. Researchers have developed from statistical to machine learning and nowadays deep learning models for this specific problem [1-3].

The nature of credit prediction data is mostly stored in tabular format, which usually contains numerical, categorical, text or even image features and extremely imbalanced. Due to, the vast majority of good loans (i.e. non-defaulters) and few bad loans (i.e. defaulters). The model, in general, updates its parameters or weights from the batch of data that they see during training. If the training data is imbalanced, the model could not have much knowledge about the minority class data and weights its decision more on the majority class.

For many domains that data is stored in tabular format, tree-based ensemble algorithm such as Random Forest (RF) and Extreme Gradient Boosting (XGB) remain a popular choice due to theirs outstanding performances [4-6].

RF is a machine learning algorithm that join multiple decision trees together to output predictions. It works by creating an ensemble of individual decision trees which are trained in parallel on random subset of the training data. To reach the final prediction, each prediction of the decision trees are aggregated through majority voting [7, 8].

Manuscript : 2023.11.05
Revised : 2023.12.18
Confirmation of Publication: 2024.01.26
Corresponding Author : Ju-Hong Lee
E-mail : juhong@inha.ac.kr

XGB is also a machine learning algorithm that belongs the gradient boosting algorithm family. It has shown excellent performance in various competitions such as Kaggle and is used in various ways as a prediction model.

Ensemble modeling is a technique which multiple individual base models' predictions are combined to improve the overall prediction score and generalization. Individual base models could be different type of algorithms or trained on different data sets. Whenever individual base models are independent and of different modeling algorithms, using ensemble approach reduces the prediction error, and enhances robustness [9, 10].

1. Problem Statement

The main limitation of existing researches are that they only consider numerical and categorical features, and discard other features such as texts from the training and inferencing process. Second, state-of-the-art (SOTA) tree-based ensemble algorithms treat features in the tabular data as independent entity [6]. They do not inherently capture interactions or relationship between features within an observation. For instance, in RF or XGB models, each split at a node of the decision tree is based on a single feature's value and the process of partitioning continues this way. The interaction between features is indirectly captured as the model learns which features are more discriminative for making decisions at various splits. Thus, in this work, we would like to conduct a study to deal with the limitations mentioned above by using the Transformer architecture [11].

2. Motivation

In defining a person's value or credit in our society, many factors are taken into account. Job title is one of those factors that can not be overlooked when identifying one's credit. Kaggle's lending club dataset contains 26 features, which one of those features is 'Job Title' where there are 173,105 unique entries in the data. Second, a borrower's information features should not be treated as independent entity. Annual income is heavily related to what kind of job or position a person have and one's career duration. These relationships are crucial in identifying borrowers' ability to pay back on the loan. Capturing these relationships is significant for making accurate predictions.

3. Related Work

Many successful stories of deep learning application in the real-world, especially in domain such as image, video, audio, and natural language processing, have led researchers and scientists to explore its potential in the tabular domain such as credit scoring, etc.

Yitan Zhu et al. [12], developed an image generator for tabular data algorithm (IGTD), that converts data from tabular to images by mapping features to pixel position in a way resembling features stay nearby one another in the image representations. They stated that a spatial relation between features are not well captured in most tabular data, and thus are not suitable for CNNs models. They evaluated IGTD on two data sets, and showed better performance comparing to models trained on the original tabular data.

Xin Huang et al. [13] from Amazon Web Service introduced TabTransformer in 2020 which utilized Transformer architecture on categorical feature. They showed their methods outperformed the existing SOTA deep learning models for tabular data on AUC score and resemble the tree-based ensemble models' performance. In addition, they found that highly correlated features result in embedding vectors that are close together in Euclidean distance. Later in 2021, Yury Gorishniy et al. [14], proposed FT-Transformer in which both categorical and numerical features are inputted to the Transformer.

Most studies focus on methods that apply on categorical and numerical features, and not taking into consideration of other features such as text feature. Nick Erickson et al. [15] also from AWS introduced a Transformer architecture that format categorical and numerical features into texts and then fed them into the Transformer along with other text features.

4. Outline

In section 2, we will discuss about the existing studies of tabular deep learning and how this research areas have developed. Then, in section 3, we describe the proposed method to deal with the problems in 1.1. Next, in section 4 go into the experiment settings and results. Finally, we conclude the paper in section 5.

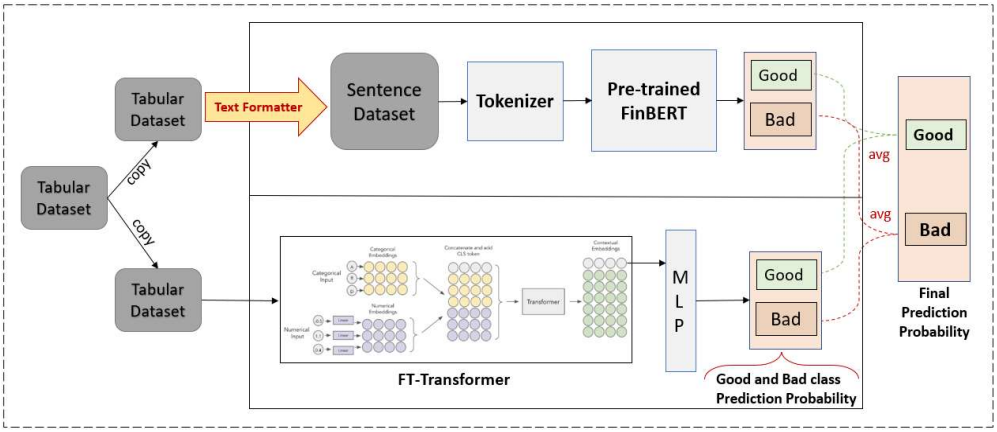

Fig. 1. Proposed Ensemble Framework

II. PROPOSED METHOD

In this section, we will discuss our proposed method which is an ensemble framework of two Transformer models to predict the likelihood of a loan being defaulted or non-defaulted in the personal credit prediction domain. The first base model is FinBERT to handle the text feature limitation [16]. The second base model is FT-Transformer which trains on categorical and numerical features. Its self-attention mechanism solve the feature relationship problem.

To give a quick overview of Fig. 1, an original tabular data set is duplicated to get two copies. The first copy is used as training data set for FinBERT and the other is to be used with FT-Transformer. Both FinBERT and FT-Transformer will be explored in more detail in the next section.

Since FinBERT is a model used in NLP, its nature input is sentence data. Therefore, the tabular data has to go through a formatting process that produce a sentence data set. FT-Transformer only works with numerical and categorical features. In this case, no conversion is needed. The features that are used in RF, XGB, and FT-Transformer are the same.

To effectively ensemble two of the base models, we tweaked prediction head to output soft prediction, i.e. probabilities of "Good" and "Bad", instead of hard prediction, i.e. either "Good" or "Bad". The probability predictions of each class is averaged to get the final prediction. In the final prediction layer, class with higher probability is selected. This approach of ensembling probabilities allow us to leverage each base model bias to achieve better generalization [17]. Also, we can make use of weighted average to give weights to the base models prediction. Example we want model_1 to make 70% and model_2 with 30% of the decision, we can give weights of 7:3 to model_1 and model_2 respectively when calculating ensemble.

1. FinBERT

FinBERT is a SOTA large language model that adjusts to the finance sector. It utilizes Google's BERT algorithm and training procedure (i.e. pretraining and finetuning) [18]. While BERT is pretrained on general text, such as BookCorpus and Wikipedia with 3.3 billion tokens in total. FinBERT is pretrained using 3 types of financial texts consisting of 4.9 billion tokens.

During pretraining, BERT follows two key training steps, masked language modeling and next sentence prediction, to gain deep contextualized understanding of a language. Masked language modeling involves masking of tokens in the input text and the model is trained to learn to predict these masked tokens, allowing BERT to infer the meaning and relationships between words, fostering deeper semantic understanding and capturing contextual nuances. Next sentence prediction trains the model to determine whether two input texts follow each other in the original text, enabling BERT to grasp the relationships between sentences, comprehend and generate coherent text.

The procedure of finetuning FinBERT from tabular data for text classification is as follows

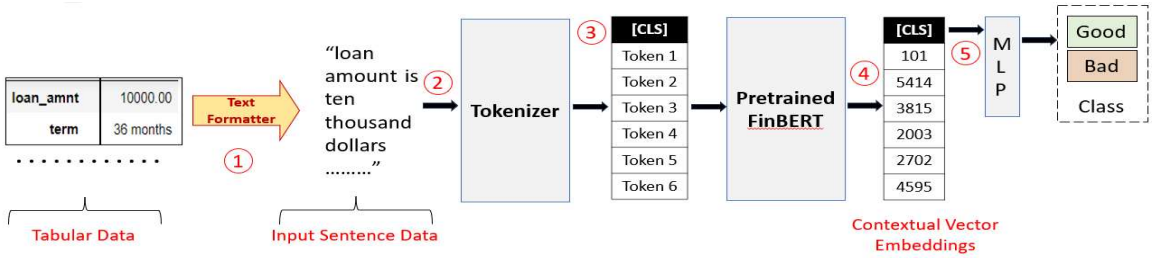

Fig. 2. Finetuning FinBERT for text classification procedure

[19, 20] (Fig. 2).

(1) The original tabular data is formatted to input sentence data via a 'Text Formatter' which is discussed in detail in section 3.3

(2) Every word in the sentence is tokenized by a Tokenizer

(3) A special classification token [CLS] is prepended at the beginning of the input sentence tokens

(4) The output vector of FinBERT are the contextual vector embeddings and the [CLS] token here contains all information of all the words in the input sentence.

(5) The [CLS] token is further used as input to downstream tasks, for example classification.

2. FT-Transformer

FT-Transformer is a modification of the Transformer architecture for tabular data. In a nutshell, all categorical and numerical features are transformed to embeddings and applies a stack of Transformer layers to the embeddings. Thus, every layer of the Transformer functions at the feature level of an individual object. Unlike NLP, positional embedding is not applied. Each training sample's features order are not considered through self-attention mechanism, every feature has knowledge about all other features. CLS token is also used for classification just like FinBERT.

3. Text Formatter

'Text Formatter' converts data from tabular format to sentence format. Every tabular data has column (feature or attribute) headers, that is the top row of the table and acts as a title for the type of information of each column. Usually, meanings or descriptions about the column headers are found with the data set. In Fig. 3, original column headers, "loan_amnt" and "int_rate" are formatted to "loan amount" and "interest rate".

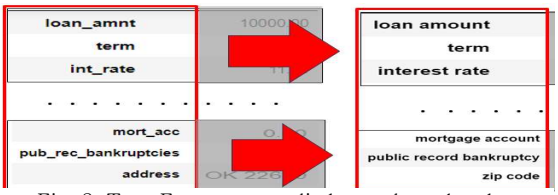

Fig. 3. Text Formatter applied on column header

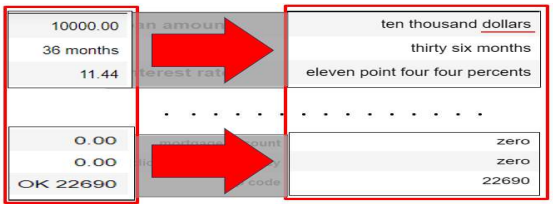

Fig. 4. Text Formatter applied on column value

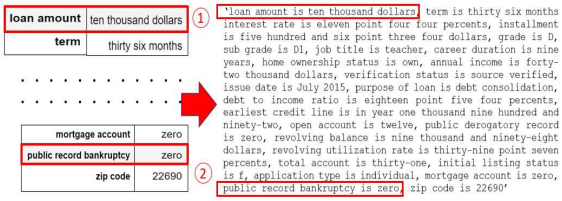

Fig. 5. Text Formatter – tabular data to text sentence

Under every column headers, there is a list of values called "column values or cell". They represent data points that are used during training and evaluating. Numerical and categorical values are formatted to text and additionally include words corresponds to what type of information it is. As shown in Fig. 4, loan amount with value of "10000.0" is formatted to "ten thousand dollars", the word "dollar" is added since we know that loan amount is corresponding to money. After the column headers and values are formatted, we concatenate every column and value into one long sentence shown in Fig. 5.

III. EXPERIMENT

This section discusses about data sets, training settings, evaluation metric and the results conducted from the experiment.

Table 1. Meta-data

	Lending Club	UCI Taiwan
# instances	396,030	30,000
# num. features	18	14
# cat. features	7	9
# text features	1	0
# classes	2	2
Class Ratio	80 : 20	78 : 22

Table 2. Hyper parameter settings

	FinBERT	FT-Transformer
# epoch	3	10
Batch size	8	64
Learning rate	1E-05	4E-03
Optimizer	AdamW	AdamW
Dropout	10%	10%
Activation	Tanh	GeGLU

1. Experiment Configuration

In this work, we used two public data sets, Kaggle's Lending Club data, a US peer-to-peer lending data, and a popular data set used in various studies, UCI Taiwan data set. (Table 1)

To observe our proposed method's performance, we compare with the two famous tree-based ensemble algorithms, RF and XGB, and Amazon's TabTransformer. The data set is split with 7:3 ratio which 70% is used for training and 30% for testing. The model's hyper parameter is set following Table 2. Due to, the nature of imbalance of the data, we additionally include two more experiments that employs data resampling techniques, on top of original (imbalanced) training, known as Random Under Sampling (RUS) and Random Over Sampling (ROS) [21]. Due to text features in the data set, algorithms such as smote or adasyn is not applicable.

Confusion matrix is used to calculate positive (default) recall, negative (non-default) recall, F1 score and Area Under the ROC Curve (AUC) to evaluate the performance of the models.

(1) Positive recall measures the models' ability to recall bad loan (defaulters) of all the bad loans in the data.

(2) Negative recall measures the ability to recall good loan (non-defaulters) of all the good loans in the data.

(3) ROC curve illustrates the balance between true positive rate (TPR) and false positive rate (FPR) across various decision thresholds. These rates are plotted as the threshold for classifying positive and negative outcomes varies, with the diagonal line representing random guessing. AUC measures the surface area underneath the ROC curve. AUC measures how well models distinguish between positive and negative classes, providing a single floating point value that summarizes the overall classification performance.

2. Experiment Results

In term of risk management, default recall and AUC score are considered more important, because financial losses associated with defaulters are worse comparing to the loss caused by non-defaulters.

(1) Lending Club Results

From Fig. 6, our proposed method achieves the best default recall, F1 score, and AUC in all 3 settings, original, random undersampling and random oversampling. Meanwhile, RF seems to capture non-default recall better than XGB. Notice that in original imbalanced training result, the non-default recall is over 90% while default recall is lower than 45%. This is because the models have seen data from one class much more often than the other. Therefore, model bias is spot on.

(2) UCI Taiwan Results

Similar to Lending Club result, our proposed method also achieves the best default recall, F1 score and AUC in all 3 settings with UCI Taiwan data. (Fig. 7)

IV. CONCLUSION

In this paper, we introduced what credit prediction is, existing popular models that have been used, and mention two limitation of those models. We discussed the importance of text feature, job title, which is crucial in credit prediction. Also, introduced a text formatter that converts tabular data into text data. We touched on, in detailed, how to finetune FinBERT to fit our classification task. Furthermore, we introduced an ensemble framework of two Transformer models that incorporate tabular text feature into the multi-modal framework and displayed superior

	Original				Random Under Sampling				Random Over Sampling			
	OURS	RF	XGB	TabTransformer	**OURS**	RF	XGB	TabTransformer	**OURS**	RF	XGB	TabTransformer
Non-Default Recall	0.98	**0.995**	0.989	0.977	0.773	**0.805**	0.799	0.771	0.772	**0.982**	0.814	0.802
Default Recall	**0.514**	0.455	0.479	0.450	**0.841**	0.779	0.805	0.782	**0.837**	0.497	0.795	0.791
F1	**0.626**	0.618	0.618	0.610	**0.632**	0.624	0.626	0.622	**0.636**	0.626	0.634	0.628
AUC	**0.909**	0.889	0.907	0.877	**0.908**	0.891	0.905	0.887	**0.906**	0.89	0.906	0.901

Fig. 6. Lending Club results

	Original				Random Under Sampling				Random Over Sampling			
	OURS	RF	XGB	TabTransformer	**OURS**	RF	XGB	TabTransformer	**OURS**	RF	XGB	TabTransformer
Non-Default Recall	0.914	**0.939**	**0.939**	0.907	0.751	**0.779**	0.73	0.744	0.819	**0.91**	0.816	0.746
Default Recall	**0.448**	0.369	0.361	0.360	**0.664**	0.627	0.647	0.629	**0.601**	0.43	0.563	0.658
F1	**0.509**	0.464	0.455	0.476	**0.521**	0.499	0.498	0.511	**0.532**	0.490	0.505	0.512
AUC	**0.766**	0.76	0.759	0.749	**0.768**	0.764	0.747	0.749	**0.768**	0.758	0.757	0.757

Fig. 7. UCI Taiwan results

results comparing two SOTA models, Random Forest, and XGB, and TabTransformer which is the current dominant model using Transformer architecture across two publicly available data sets.

To the best of our knowledge, we are the first to tackle the multi-modal data in the domain of credit scoring system. Therefore, there are still rooms for improvement, such as using different text formatting algorithms, or techniques to speed up the training time.

REFERENCES

[1] Wang, Chongren, and Zhuoyi Xiao. "A Deep Learning Approach for Credit Scoring Using Feature Embedded Transformer," *Applied Sciences*, vol. 12, no. 21, 2022.

[2] Markov, Anton, Zinaida Seleznyova, and Victor Lapshin. "Credit scoring methods: Latest trends and points to consider," *The Journal of Finance and Data Science*, vol.8, pp. 180-201, 2022.

[3] Hayashi, Yoichi. "Emerging trends in deep learning for credit scoring: A review," *Electronics*, vol. 11, no. 19, 2022.

[4] Crook, Jonathan N., David B. Edelman, and Lyn C. Thomas. "Recent developments in consumer credit risk assessment," *European Journal of Operational Research*, vol. 185, no. 3, pp. 1447-1465, 2007.

[5] Li, Yu. "Credit risk prediction based on machine learning methods," *2019 14th International Conference on Computer Science & Education (ICCSE). IEEE*, 2019.

[6] Li, Hua, et al. "XGBoost model and its application to personal credit evaluation," *IEEE Intelligent Systems*, vol. 35, no. 3, pp. 52-61, 2020.

[7] Misra, Siddharth, Hao Li, and J. He. "Noninvasive fracture characterization based on the classification of sonic wave travel times," *Machine learning for subsurface characterization*, pp. 243-287, 2020.

[8] L Breiman. "Random Forests," *Springer,* vol. 45, pp.5-32, 2001.

[9] Kotu, Vijay, and Bala Despande. "Chapter 2-data science process," Data science, 2nd edn, Morgan Kaufmann (2019):, pp.19-37.

[10] Deshpande, V.K., and V.Kotu. "Predictive Analytics and Data Mining," *Elsevier Inc*, 2015.

[11] Vaswani, Ashish, et al. "Attention is all you need," *Advances in Neural Information Processing Systems 30* , 2017.

[12] Zhu, Yitan, et al. "Converting tabular data into images for deep learning with convolutional neural networks," *Scientific reports*, no. 11325, 2021.

[13] Huang, Xin, et al. "Tabtransformer: Tabular data modeling using contextual embeddings," arXiv preprint arXiv:2012.06678 (2020).
[14] Gorishniy, Yury, et al. "Revisiting deep learning models for tabular data," Advances in Neural Information Processing Systems 34 (2021): 18932-18943
[15] Erickson, Nick, et al. "Multimodal automl for image, text and tabular data," *Proceeding of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4786-4787, 2022.
[16] Huang, Allen H., Hui Wang, and Yi Yang. "FinBERT: A large language model for extracting information from financial text," *Contemporary Accounting Research*, vol. 40, no. 2, 2023.
[17] Ju, Cheng, Aur□lien Bibaut, and Mark van der Laan. "The relative performance of ensemble methods with deep convolutional neural networks for image classification," *Journal of Applied Statistics*, vol. 45, no. 15 2018.
[18] Devlin, Jacon, et.al "BERT: Pretraining of Deep Bidirectional Transformers for Language Understandin," arXiv preprint arXiv:1810.04805 (2018).
[19] Sun, Chi, et al. "How to fine-tune bert for text classification?," Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18-20, 2019, Proceedings 18. *Springer International Publishing* , 2019.
[20] Howard, Jeremy, and Sebastian Ruder. "Universal language model fine-tuning for text classification," arXiv preprint arXiv:1801.06146 (2018).
[21] Kotsiantis, Sotiris, Dimitris Kanellopoulos, and Panayiotis Pintelas. "Handling imbalanced datasets:A review," *GEST international transactions on computer science and engineering*, vol. 30, no. 1 pp. 25-36, 2006.

Author

Sophot Ky is a Korean Air's merit-based Cambodian scholar who received his B.Sc. degree in computer science and engineering from Inha University, Incheon, South Korea, in 2022, where he pursues M.S. degree with Department of Electrical and Computer Engineering. He has high interest in ML/DL, the Transformer architecture, NLP, and Automatic Speech Recognition.

Ju-Hong Lee received the B.S and M.S. degrees in computer engineering from Seoul National University, South Korea, in 1983 and 1985, respectively, and the Ph.D. degree in computer science from KAIST, Daejeon, South Korea, in 2001. He is currently a Professor of computer science and enginering with Inha University, Korea. He is also the CEO of Qhedge Company Ltd. He has developed the credit for portfolio management, index tracking, and arbitrage trading. His research interests include machine learning, data mining, financial engineering.

Kwangtek Na received the B.Sc. degree in civil engineering and M.S. degree in computer science and engineering from Inha University, South Korea, in 2013 and 2017, respectively, where he is currently pursuing the Ph.D. degree with Department of Electrical and Computer Engineering. He is also reseaching machine learning with Hanwha Group. His research interests include statistical machine learning, reinforcement learning, and recommender systems.