An Ensemble Model for Credit Default Discrimination: Incorporating BERT-based NLP and Transformer

Sophot Ky[1], Ju-Hong Lee[2]
[1]Dept. of Electrical and Computer Engineering, Inha University
[2]Dept. of Computer Engineering, Inha University

sophotk@inha.edu, juhong@inha.ac.kr

ABSTRACT

Credit scoring is a technique used by financial institutions to assess the creditworthiness of potential borrowers. This involves evaluating a borrower's credit history to predict the likelihood of defaulting on a loan. This paper presents an ensemble of two Transformer based models within a framework for discriminating the default risk of loan applications in the field of credit scoring. The first model is FinBERT, a pretrained NLP model to analyze sentiment of financial text. The second model is FT-Transformer, a simple adaptation of the Transformer architecture for the tabular domain. Both models are trained on the same underlying data set, with the only difference being the representation of the data. This multi-modal approach allows us to leverage the unique capabilities of each model and potentially uncover insights that may not be apparent when using a single model alone. We compare our model with two famous ensemble-based models, Random Forest and Extreme Gradient Boosting.

1. INTRODUCTION

Default risk is the most crucial risk financial institutions are concerned about. In the conventional approach, loan applications are manually evaluated by domain experts to assess the likelihood of default (i.e., a borrower will not pay back). However, this process is labor-intensive and subject to human bias, leading to potential inconsistencies in decision making. As a result, a growing number of enterprises are dedicating their resources to the development of data-driven discrimination models aimed at assessing the creditworthiness of loan applications based on their information. Credit default discrimination is essentially a binary classification task, wherein a predictive model is utilized to determine whether a loan application is likely to result in default or non-default status, reflecting the borrower's ability to repay the loan in a timely manner.

Transformers have emerged as state-of-the-art deep learning model in the fields of natural language processing. Having achieved remarkable performance in the domain of machine translation and language modeling. Inspired by the success of this model, we have employed this concept into the finance domain, credit scoring application.

In this paper, we propose an ensemble model within a framework formed from two Transformer models by averaging their output likelihoods to get the final prediction whether a loan application is risky or safe. We utilize two distinct models, FinBERT [1] and FT-Transformer [2], which employ contrasting methodologies for processing input data. By harnessing this multi-modal approach, we can take advantage of the strengths of each model and learn insightful information from the data. FinBERT, a pretrained NLP model to analyze sentiment of financial text, takes a transformed textual sentence in natural language format as input, whereas FT-Transformer takes in the original tabular data and transforms numerical and categorical features into embeddings and applies a stack of Transformer layers to these embeddings.

Furthermore, we compare our ensemble model with two widely used ensemble models, Random Forest and Extreme Gradient Boosting (XGBoost).

The remainder of the paper is organized as follows. In section 2, we briefly review the related work, and the proposed model is described in section 3. Finally, section 4 and 5 describe the experiment results and conclusion of this paper.

2. RELATED WORKS

In [3], the author introduced an end-to-end feature embedded Transformer (FE-Transformer) into the field of credit scoring, where Transformer is applied on the input which is the user online behavior data. They conducted a detailed comparative analysis and demonstrated superiority of FE-Transformer against Logistic Regression and XGBoost.

In another study [4], a Network Intrusion Detection System (NIDS) based on Bidirectional Encoder Representations from Transformer (BERT) framework. The non-linguistic network traffic flows were organized into language-like structures, treating each flow as a word. Each flow consists of 6 features, *“Src pt”*, *“Dst pt”*, *“Proto+Flags”*, *“# Packets”*, *“# Bytes”*, *“Duration”*.

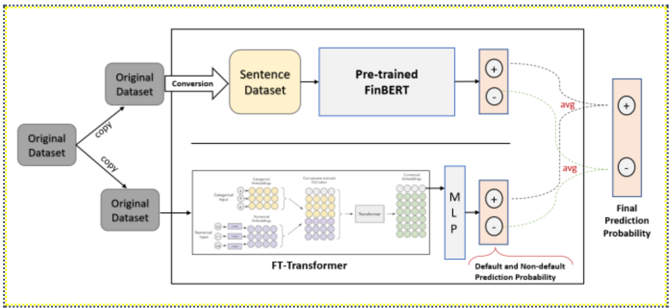

Figure 1. Proposed framework

3. PROPOSAL

Our proposed framework incorporates an ensemble model, as depicted in Figure 1.

As our two models utilize distinct data representation as inputs, we initially created two separate copies of the original tabular data set. For FinBERT fine-tuning, we transformed one of the copied datasets, which contains *M training examples*, into *M meaningful sentences* by treating each feature as a phrase and each row as a sentence. Further information regarding the data conversion can be found in the **Appendix** section. For FT-Transformer training, we followed the settings in [2].

In alignment with the experimental findings in [5], which suggest the advantages of ensembling probabilities, we tailored this approach to suit our purpose, considering our use of only two base models, which averaging the hard predictions is not applicable. Accordingly, we modified each of our Transformer models to output the probabilities of obtaining 0 for non-default class and 1 for default class, instead of hard values of 0 or 1. Subsequently, we averaged the probabilities of 0 and 1 from both models and employed the resulting averaged probabilities as our final output, wherein the class with the higher probability serves as the hard prediction.

We used two datasets to conduct experiments for this study.

- UCI default of credit card: data of customers default payments in Taiwan.
- Kaggle Lending Club: a US peer-to-peer lending company.

The details of the datasets are shown in Table 1.

Table 1. Description of the two data sets.

	Taiwan	*Lending Club*
# Instances	30,000	396,030
# Numerical Features	14	14
# Categorical Features	9	7
# Classes	2	2
Class ratio	78 : 22	80 : 20

As the class distribution in both of our datasets is imbalanced, it can result in the model being biased towards predicting one class more frequently than the other. To address this issue, we employed a random under sampling (RUS) technique, where we subsampled the majority class training samples to align with the number of instances in the minority class. This approach helps to mitigate the potential bias caused by the class imbalance in our data.

The original data is split into train and test set with the ratio of 70 to 30 respectively.

We evaluate the effectiveness of our approach by comparing its performance against popular and widely used ensemble models in various machine learning tasks, Random Forest (RF) and XGBoost (XGB) classifiers. The performance of each classifier is assessed based on its ability to accurately recall default and non-default loan applications as well as its Area Under the Curve (AUC) score.

4. EXPERIMENT RESULT

Table 2 and 3 show the results of the three ensemble models. Based on the tables, our proposed ensemble model shows superior performance in terms of recalling default loan application, and AUC compared to Random Forest and XGBoost. While our model may have lower performance in recalling non-default applications, accurately identifying default loan applications is generally considered more important in terms of risk management. This allows lenders to identify high-risk loan applications, which helps mitigate potential financial losses associated with defaults.

Table 2. Evaluation result of models on Lending Club data set.

	RF	***XGB***	***OURS***	***Gain (%)*** ***RF | XGB***
Non-default recall	**0.805**	0.799	0.729	*-7.6 | -7*
Default Recall	0.779	0.805	**0.877**	***9.8 | 7.2***
AUC	0.891	0.905	**0.910**	-

Table 3. Evaluation result of models on Taiwan data set.

	RF	***XGB***	***OURS***	***Gain (%)*** ***RF | XGB***
Non-default recall	**0.779**	0.730	0.695	-8.4 | -3.5
Default Recall	0.629	0.647	**0.701**	**7.2 | 5.4**
AUC	0.764	0.747	**0.772**	-

5. CONCLUSION

In this paper, we introduce a framework that incorporates an ensemble of two Transformer based models for the purpose of discriminating the default risk of loan applications. The first transformer model is FinBERT, a pre-trained natural language processing model that analyzes sentiment in financial text. The second Transformer model is FT-Transformer, a modified version of the Transformer architecture designed for tabular data.

We concluded that Transformer models hold promise in the field of credit scoring when applied to tabular data set. However, further research could be conducted to explore the effective methods of converting tabular data into linguistic sentences in order to fully harness the capabilities of BERT.

REFERENCES

[1] Araci, Dogu. "Finbert: Financial sentiment analysis with pre-trained language models." *arXiv preprint arXiv:1908.10063* (2019)

[2] Gorishniy, Yury, et al. "Revisiting deep learning models for tabular data." *Advances in Neural Information Processing Systems* 34 (2021): 18932-18943.

[3] Wang, Chongren, and Zhuoyi Xiao. "A Deep Learning Approach for Credit Scoring Using Feature Embedded Transformer." *Applied Sciences* 12.21 (2022): 10995.

[4] Nguyen, Loc Gia, and Kohei Watabe. "Flow-based network intrusion detection based on BERT masked language model." *Proceedings of the 3rd International CoNEXT Student Workshop*. (2022): 7-8.

[5] Ju, Cheng, Aurélien Bibaut, and Mark van der Laan. "The relative performance of ensemble methods with deep convolutional neural networks for image classification." *Journal of Applied Statistics* 45.15 (2018): 2800-2818.

APPENDIX

In tabular data, column headers serve as a reference or guide to interpret the data in the table and provide context for understanding the values in the corresponding columns. They can be used to identify the type of data in a column, such as amounts, date, interest rates or other relevant information. They are usually placed at the top of the row of the table and are often formatted differently.
All we need to do is combine these column descriptions and values in the columns and provide a more meaningful sentence. But before that, we must transform the column headers and their values to meaningful text, as shown in Figure 2. This transformation includes removing symbols like "_" and correcting misspelled words.

Before

loan_amnt	term	int_rate	installment
21000.00	60 months	15.61	506.34

After

loan amount	term	interest rate	installment
twenty-one thousand dollars	sixty month	fifteen point size one percents	five hundred and six point three four dollars

Figure 2. An example of data conversion

Additionally, any numerical values in the table are transformed to textual numbers and necessary additional string (*"dollars"* and *"percents"*) are concatenated as needed.