eISSN: 2976-1190

Journal of Engineering and Sciences

Vol. 2 Issue 1 November 30, 2023

Journal homepage: ***Journal of Engineering and Sciences (nepjol.info)***

Publisher: Research Management Cell, Pashchimanchal Campus, Lamachour, Pokhara

Multi-Class Credit Risk Analysis Using Deep Learning

Sagun Babu Paudel[1], Bidur Devkota[1*], Suresh Timilsina[2]

[1] *Faculty of Science and Technology, Gandaki College of Engineering and Science, Pokhara University, Nepal*

[2]*Department of Computer Engineering, IOE, Pashchimanchal Campus, Tribhuvan University, Nepal*

(Manuscript Received:13/08/2023; Revised:22/11/2023; Accepted: 24/11/2023)

Abstract

Credit risk prediction, reliability, monitoring and effective loan processing are the keys to proper bank decision-making. So, understanding the credit customer during the initial loan processing phase would help the bank prevent future losses. In this regard, this study aims to develop a credit risk evaluation model using deep learning algorithms. The model utilizes a credit risk analysis dataset published in Kaggle. The objective is to build deep learning models for predicting credit risk using real banking datasets published on Kaggle. Firstly, data preprocessing and feature engineering are done. Suitable features such as irrelevant and null valued features are identified and removed with techniques like the Karl Pearson correlation, information values, and weight of evidence. Next, data normalization is performed and target features are separated into three classes: high risk, medium risk and low risk. SMOTE-ENN (Synthetic Minority Oversampling Technique with Edited Nearest Neighbor) was applied to balance the dataset. State-of-the-art deep learning algorithms such as GRU (Gated Recurrent Units) Model and Bidirectional Long Short-Term Memory (Bi-LSTM) are implemented to train and learn from the pre-processed data. GRU and Bi-LSTM models performed well, with F1 scores of 0.92 and 0.93, respectively. The result of this investigation illustrates that deep learning models seem promising for evaluating and predicting multi-class problems.

Keywords: Bi-LSTM; Credit risk; Financial institutions; GRU; Loan default prediction; Risk mitigation; SMOTE-ENN

1. Introduction

1.1 Background

Credit risk analysis is the technique of determining the probability that a borrower would default on a loan. This process helps assess a borrower's trustworthiness, which is very important for lenders to make informed lending decisions and minimize the risk of losses. For proper credit risk analysis, lenders consider many factors (such as borrower's credit history, capital, capacity to repay, etc.). Various approaches like scoring models and financial analysis are in use by lenders for the purpose. Basically, when lenders calculate credit risk, they are trying to predict the chances of getting back both the interest and the main amount while releasing loans to customers. Borrowers with low credit risk can be charged lower interest rates. To avoid the maximum risk, the lender checks the borrower can pay the loan on time[2].

Deep learning models have shown superior predictive performance in various domains, which can be crucial in identifying potential credit defaults. Most of the literature has focused on credit risk analysis as a case of a binary classification problem and categorized the borrowers into two types, i.e., high risk or low risk [2,3]. Deep Learning models can be customized and tuned for multi-class credit risk analysis tasks[1].

In this research, a closer examination is conducted on deep learning methods for analyzing multi-class credit risk problems. Specifically, factors such as the loan amount, loan term, interest rate, installment amount, annual income, purpose of the loan, and total principal and interest payments are scrutinized. These key features hold a central position in the analysis of credit risk. This study uses these features to explore multi-class credit risk analysis using deep learning.

This study investigates and classifies the customers into 3 categories, i.e., high risk, low risk and medium risk. Exploring through the literature, it is known that most of the past works in the credit risk analysis domain deal with binary class credit risk analysis. Not much work has been accomplished for multi-class problems using deep learning. Hence, this study contributes by exploring and illustrating the use of deep learning models in multi-class evaluation

[*]Corresponding author. Tel.: +977- 9856066658,
E-mail address: im.bidur@gmail.com

problems.

1.2 Literature Survey

Over the past years, various studies have been accomplished to investigate credit risk evaluation problems. Zhang et al. [3] explored multi-class credit risk assessment problems with stacking integration. The study outlined how to tackle risk reduction by enhancing the process of selecting relevant features and incorporating a stacking approach with five distinct learners: Logistic Regression, Random Forest, GBDT, XGboost, and Light GBM. Promising results were obtained with F1 score of 0.8731.

Sheikh et al. [2] analyzed loan approval problems using machine learning algorithms like logistic regression. The model achieved an accuracy of 81.1%. Youlve et al. [5] demonstrate the application of principal component analysis to streamline dimensionality and extract the most pertinent indicators for credit decision systems. The proposed model achieved good performance with an accuracy of 97.6%.

Sarini et al. [10] accomplished a study titled "Easy ensemble with random forest to handle imbalanced data in classification." The results illustrated that Easy Ensemble and Random Forest can effectively handle data-imbalanced problems. The model achieved promising performance growth and a recall value of up to 0.82 while evaluating against different datasets.

Zhu et al. [1] provide some theoretical framework for multi-class credit risk analysis problems using ensemble machine learning; however, their framework is not supported empirically.

Clements et al. [12] presented a method for credit risk monitoring using deep recurrent and causal convolution-based neural networks. It is based on a credit card transaction sampling-based method that leverages lengthy historical financial data sequences. The outcomes showed promising results regarding considerable cost reductions and early credit risk detection.

Much of the literature encountered deals with credit risk analysis problems for binary and multiple classes using classic machine learning algorithms [2,10]. Few researchers have accomplished works using deep learning models for binary classification problems [13]. Another study has proposed frameworks for multi-class problems with deep learning methods. However, no experimental evaluation was made. The author suggested that bagging learners may be promising for multiclass problems [3]. The methodology proposed in our study aims to fill the gap in multi-class credit risk evaluation using state-of-the-art deep learning models.

1.3 Contribution

The proposed work empirically illustrates the use of deep learning models for multi-class credit risk analysis problems. Listed below are the contributions accomplished by this study.

- Use of multi-class (3-class) target classification on GRU and Bi-LSTM deep learning models.
- Comparing GRU and Bi-LSTM deep learning model using data balancing technique.

2. Materials and Method

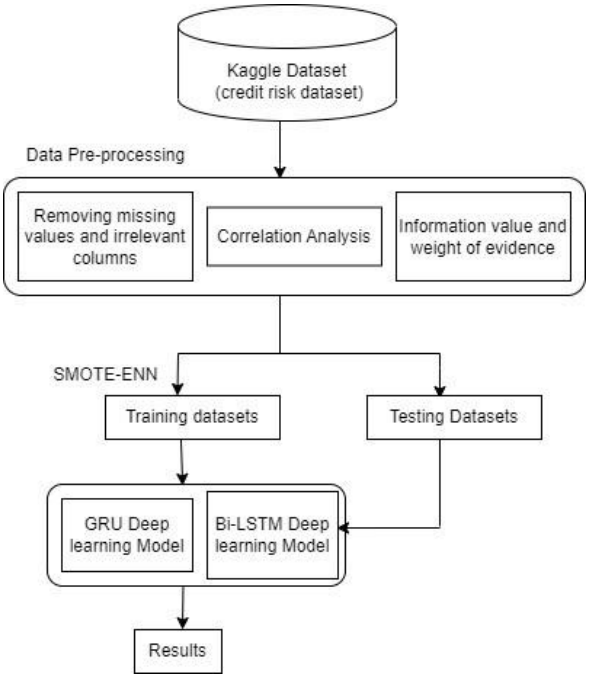

Figure 1: Overall Method

Figure 1 illustrates the proposed methodology for the credit risk analysis. Firstly, columns with missing values are removed to prevent potential inaccuracies during processing. Secondly, irrelevant columns are excluded as they do not contribute to model building. Thirdly, the Pearson coefficient correlation identifies and eliminates multicollinearity among features. Subsequently, the Weight of Evidence and Information Value is used in feature engineering to eliminate features lacking sufficient predictive information for credit risk assessment. Upon completing the data-preprocessing phase, the dataset is split into training and testing sets using ratios of 80/20, 70/30, and 50/50. Stratified sampling is used in training testing split, representing each class variable in credit risk analysis. Normalization is applied to scale the data for more accurate analysis.

Furthermore, deep learning algorithms like Gated Recurrent Units (GRU) and Bi-directional Long Short-Term Memory (Bi-LSTM) are employed and evaluated both with and without Synthetic Minority Oversampling Technique – Edited Nearest Neighbor (SMOTE-ENN) due to the imbalanced dataset. This comprehensive approach ensures a robust and thorough analysis of credit

risk assessment.

2.1 Dataset

A publicly accessible dataset from Kaggle [10] evaluates the proposed model. The dataset contains a total of 8,87,379 records and 74 features such as id, loan_amnt, int_rate, annual_inc, dti, total_payment, installment, loan_status, term, etc. loan_status is the target category to be classified in this study. Among the 8,87,379 records, 6,01,7799 records are of the current loan customers, which is not the focus of this study. The remaining records are categorized into three different loan status types. Charged-off and default customers are considered "high risk" type, having 36404 records. Late paying customers are taken as the "medium risk" type, having 11462 records. Fully paid customers are treated as the "low risk" type, having 1,53,937 records.

2.2 Techniques and Algorithms

2.2.1 Feature Selection

Features with a high proportion of missing data could significantly impact the reliability of the results, as attempting to process them might introduce noise or distortion into the dataset and not contribute to the model-building process. Such features include 'id', member 'id', 'url', 'title', 'desc', 'policy code', and 'emp_title', which hold no predictive value and must be removed to streamline the dataset and enhance its usability. Similarly, Pearson Correlation is applied for the highly correlated features where a correlation of more than 0.8 was removed from the dataset [3].

Further, two important concepts are used to assess the significance of features, i.e., The weight of evidence and information value [6,7].

WOE = ln (% of non-events/ % of events) …. (i)

IV = ∑ (% of non-events/ % of events) * WOE...(ii)

where % of non-events is the percentage of observat***ions*** that do not belong to the event class and % of events ***is*** the percentage of observations that belong to the event class.

2.2.2 Gated Recurrent Unit (GRU)

The GRU model is selected as a potential algorithm for developing the model. It handles the long-range dependencies as vanishing gradient issues better than traditional Recurrent Neural Networks. The GRU model allows us to adapt quickly to changing trends or shifts in credit risk patterns. However, the effectiveness of a GRU model to ensure accurate credit risk assessment depends on the quality and quantity of data and careful model tuning [8].

2.2.3 Bidirectional Long Short-Term Memory (Bi-LSTM)

The Bi-LSTM model is an advanced version of LSTM model [8]. Bi-LSTM model is used for the model development because Bi-LSTMs process data in both forward and backward directions simultaneously, allowing them to consider past and future information for each time step. This enables a more comprehensive understanding of a borrower's financial behavior and credit history. Bi-LSTMs excel at capturing complex and non-linear relationships in the data, which is valuable for identifying subtle credit risk factors and early warning signals of potential defaults. However, the effectiveness of a Bi-LSTM model to ensure accurate credit risk assessment depends on careful data preprocessing, model tuning, and validation to ensure accurate and reliable credit risk assessments [8].

2.2.4 Synthetic Minority Oversampling Technique combined with Edited Nearest Neighbors (SMOTE-ENN)

SMOTE-ENN is a two-step technique used to handle imbalanced datasets in machine learning. SMOTE-ENN effectively addresses class imbalance by oversampling the minority class and removing noisy and irrelevant samples using ENN [11]. This creates a balanced dataset, which is crucial for training machine learning models like Bi-LSTM and GRU, as it prevents the models from being biased towards the majority class leading to more accurate and reliable credit risk assessments.

3. Results and Discussion

In this study, the raw dataset has a total of 74 features. Irrelevant features (e.g. 'id', 'url', 'title', 'desc', 'policy code', and 'emp_title') and features with more than 97% missing values have been removed. A total of 26 features were dropped. Now, the Pearson Correlation was applied to remove the highly correlated features, leading to 8 features removal. Furthermore, adhering to the Information Value and Weight of Evidence theory, 20 more features were filtered out and finally, 16 features were selected for further processing. One hot encoding was applied for the categorical features before applying Deep Learning algorithms.

The preprocessed dataset was split into training and testing sets and then GRU and Bi-LSTM were applied for analysis. The GRU model has 32-dimensional vectors for each input time step at 1st layer. The model includes the Sigmoid Activation function compiled by Adam Optimizer. The model is trained for 50 epochs, has a learning rate of 0.001 and a batch size of 32. The Bi-LSTM

model has 32-dimensional vectors for each input time step at 1st layer. Adam Optimizer compiles the model. 2nd layer is added to the model, which has 50 units, and a drop-out layer with a rate of 0.2 is added. The model is trained for 50 epochs, has a learning rate of 0.001, and a batch size of 32.
The weighted F1 score is used for performance evaluation as it considers precision and recall and provides a single value for overall performance.

Table 1: Weighted F1 score for 3 class classification

Split	**GRU Model**		**Bi-LSTM model**	
	Without Smoteen	With Smoteen	Without Smoteen	With Smoteen
80/20	0.91	**0.92**	0.92	**0.93**
70/30	0.91	0.90	0.92	**0.93**
50/50	**0.92**	0.90	**0.93**	**0.93**

Table 1 shows the performance metrics for GRU and Bi-LSTM models. Bi-LSTM exhibits the best performance against balanced data with the F1 score of 0.93 for all splits. The F1 score of the Bi-LSTM model increases when using the SMOTE-ENN because the combination of oversampling and noise reduction can enhance the ability of the model to generalize well to new and unseen data. Since the Bi-LSTM model with a 50/50 data split ratio outperformed other models, further analysis of the results in the upcoming discussions is based on it.

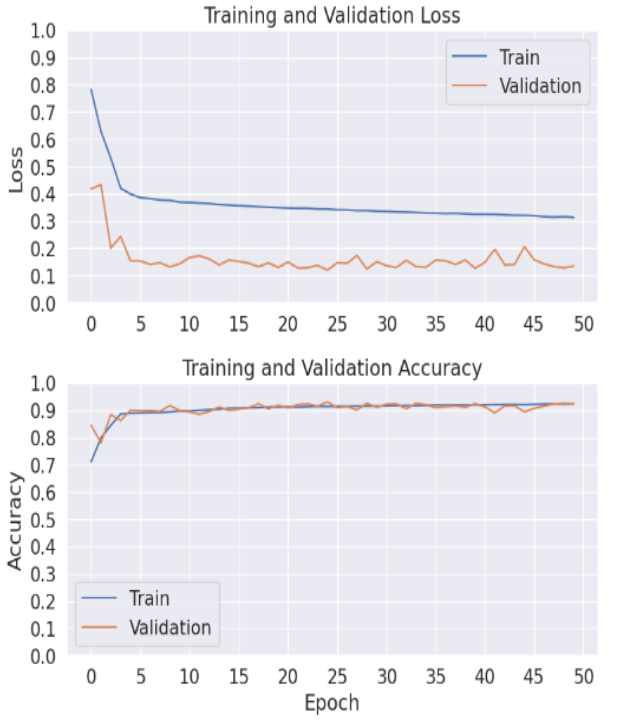

Figure 2: Training and Validation loss and accuracy of Bi-LSTM model for 50/50 split

Figure 2 contains the training and validation loss and accuracy of the Bi-LSTM model of the 50/50 split. A training loss of 0.3 indicates that the model's prediction is relatively close to the actual value in the training data. A validation loss of 0.1346 suggests that the model could capture meaningful patterns and relationships from the training data and apply them to new data. The decreasing trend indicates that the performance of the model increases. The model accurately predicted the class label for 92.34% of training instances. The model's prediction on validation data was quite accurate, with 92.35% of the instances predicted accurately.
Sheikh et al. [2] use the Kaggle datasets, where the predictive journey begins with data cleaning, preprocessing and handling missing values. The model gives the best accuracy rate of up to 81%. The accuracy of the model can be increased by using deep learning models. Likewise, in this study, the model gives the best F1 score of 0.93 for the separate Kaggle dataset compared to the above.

Actual ↓	**Confusion Matrix**		
Low Risk	76833	85	51
Medium Risk	145	3176	2410
High Risk	217	4359	13627
Predicted →	**Low Risk**	**Medium Risk**	**High Risk**

Figure 3: Confusion Matrix of the Bi-LSTM model for 50/50 split

Figure 3 represents the confusion matrix of the best model for a 50/50 train test split. Figure 3 shows that 93% of the data are accurately classified, and 7% of the data are misclassified in the above model. 7267 data were misclassified.

Table 2: Sample misclassified data

Loan amount	Interest rate	Annual income	Total payment	Predicted class	Actual Class
11000	19.99	45000	1717.74	Medium Risk	Fully paid
2400	13.99	17000	652.43	Low Risk	High Risk

Table 2 shows some samples of the misclassified data. Misclassification within a confusion matrix highlights

the instances where the model's prediction does not align with the actual outcomes. The first row shows that the loan amount of 11000 was a Grade A loan and fully paid, but the model misclassified it. The misclassification might have occurred because this record has a comparatively higher interest rate and, in the dataset, most of the records with higher interest rates are one of the default categories. This might have caused the model to classify it as a medium-risk category instead of a low-risk category. The second row shows that the loan amount of 2400 was a Grade C loan and high risk, but the model misclassified it. The misclassification might have occurred because this record has a comparatively lower interest rate, and in the dataset, most of the records with lower interest rates are low-risk. This might have caused the model to classify it as low-risk rather than high-risk.

This research found that the Performance of the Bi-LSTM model is better than that of the GRU model. Applying SMOTE-ENN to the Bi-LSTM model resulted in performance improvement with an increase in F1 score from 0.91 to 0.92, showcasing the ability of the model to leverage the enhanced dataset for improved predictions. Conversely, the F1 score of the GRU slightly decreases from 0.91 to 0.90 with SMOTE-ENN. It is due to the introduction of noise through oversampling and the existing effectiveness of the model. These findings highlight the significance of model architecture and data characteristics in class imbalance handling. The F1 score in the Bi-LSTM model increases because it has two LSTM layers, which can capture both forward and backward temporal dependencies in the sequence data [8]. This model helps segment customers into different risk categories so that the bank can measure the associated risk.

4. Conclusions and Future Works

4.1 Conclusions

The proposed Credit Risk Analysis Model using deep learning algorithms (GRU and Bi-LSTM) yielded promising results for 3 class classification scenarios. The model implemented with Bi-LSTM outperformed GRU and obtained the best performance with an F1 score of 0.93 while using a balanced dataset. Thus, the research and study show that deep learning techniques can be used for analyzing credit risk. The bi-LSTM model gives a better F1 score than the GRU model in the case of the deep learning models. Further, using the data balancing techniques.

The study's limitations are employing deep learning to solve the credit risk analysis changes over time and from place to place, so human expertise is also needed for changing economic conditions and shifting borrower behaviors. Training deep learning models can be computationally intensive and time-consuming.

4.2 Future Works

Generative AI can be a valuable tool in credit risk analysis, but it should be used with human expertise. ADASYN (Adaptive Synthetic Sampling) is an advanced version of SMOTE that aims to oversample minority data by considering data density can be used. Misclassification errors on the confusion matrix can be lowered.

Acknowledgment

This work is supported by the Faculty of Science and Technology, Gandaki College of Engineering and Science, Pokhara, Nepal.

References

[1] Zhu, F., Chen, X., and Li, G. Multi-classification assessment of personal credit risk based on stacking integration, *Procedia Computer Science*, 214 (2022) 605-612.

[2] Sheikh, M. A., Goel, A. K. and Kumar, T. An Approach for Prediction of Loan Approval using Machine Learning Algorithm, *International Conference on Electronics and Sustainable Communication Systems (ICESC)*, (2020) 490-494.

[3] Zhang, T., and Li, J. Credit risk control algorithm based on stacking ensemble learning, *IEEE International Conference on Power Electronics, Computer Applications (ICPECA)*, (IEEE 2021) 668-670.

[4] Gogtay, N. J., and Thatte, U. M. Principles of correlation analysis, *Journal of the Association of Physicians of India*, 65(3) (2017) 78-81.

[5] Youlve, C., Kaiyun, B., and Jiangtian, C. Credit decision system based on combination weight and eXtreme Gradient Boosting algorithm, *Journal of Physics: Conference Series*, 1955 (1) (2021) 012081.

[6] Jie, S., Li, J., and Fujita, H. Multi-class imbalanced enterprise credit evaluation based on asymmetric bagging combined with light gradient boosting machine, *Applied Soft Computing*, 130 (2022) 109637.

[7] Guoping, Z. Metric divergence measures and information value in credit scoring. *Journal of Mathematics*, (2013) 1-10.

[8] Mateus, B., Mendes, M., Farinha, J. T., Assis, R., and Cardoso, A. M. Comparing LSTM and GRU models to predict the condition of a pulp paper press. *Energies*, 14 (21) (2021) 6958.

[9] Sarini, A. and Prasetyo, G. V. Easy ensemble with random forest to handle imbalanced data in classification. *Journal of Fundamental Mathematics and Applications*, 3(1) (2020) 39-46.

[10] R.G. Kaggle Credit Risk dataset (2021) see

[https://www.kaggle.com/dsv/2327131].

[11] Mbunge, E., Sibiya, M. N., Takavarasha, S., Millham, R. C., Chemhaka, G., Muchemwa B., and Dzinamarira T. Implementation of ensemble machine learning classifiers to predict diarrhea with SMOTEENN, SMOTE, and SMOTETomek class imbalance approaches, *2023. Conference on Information Communications Technology and Society*, Durban, South Africa, (2023) 1-6.

[12] Clements, J. M., Xu, D., Yousefi, N., Efimov, D. Sequential Deep Learning for Credit Risk Monitoring with Tabular Financial Data, *arXiv preprint arXiv:2012.15330* (2020).

[13] Wang, H., Kou., and Peng, Y. Multi-class misclassification cost matrix for credit ratings in peer-to-peer lending. *Journal of the Operational Research Society*, 72(4) (2021) 923-934.