Eraste boko

Written Report: **Personal Credit Default Discrimination Model Based on Super Learner Ensemble**

Introduction

Historically, financial institutions have had difficulty deciding who should get loans because bank customers may be unable to repay them. Credit risk is the possibility that a customer cannot pay his loan over a specific period. A default discrimination model is a predictive model used by financial institutions for assessing their clients' credit risk. This report delves into the paper "Personal Credit Default Discrimination Model Based on Super Learner Ensemble" by Gang Li et al., presenting a new default discrimination model with enhanced accuracy and robustness. This report will be divided into five sections. The first section discusses the challenges and limitations of the previous credit risk assessment models and the need for the new model discussed in the research paper. The second section explains how the super learner heterogeneous ensemble model works in detail. The third section evaluates the performances of the super learner heterogeneous ensemble model, confirming its superiority over other default discrimination models using multiple metrics and real-world datasets. The fourth section examines the social impacts of implementing this new model with its implications for borrowers, lenders, and credit risk analysts/underwriters. Finally, the fifth section will review how this new system should be integrated ethically into society and how the balance between human judgment and AI prediction can enhance decision-making about potential borrowers'.

Problem context

The specific question AI is answering in the context of this research article is the improvement of personal credit default discrimination models using the super learner heterogeneous ensemble approach. Before the introduction of AI, traditional statistics methods such as the Z-score model, the probit analysis method, and the logistic analysis model were used to develop default discrimination models. These earlier methods had limitations. They may require strict assumptions and help with increasingly large index systems, limiting their practical application. With the rapid development of AI, the introduction of machine learning models like artificial neural networks (ANNs), support vector machines (SVMs), and decision tree (DT) have allowed techniques that can handle high dimensional data without strict assumption. These single classifiers performed better and solved nonlinear problems inherent in the credit dataset. (Shi et al. 2) Nevertheless, these classifiers have limitations in dealing with different credit risk assessment problems. These limitations led to the development of ensemble models, which combine multiple single classifiers to enhance their performance further. (Ghodselahi et al. 2) While ensemble models like Bagging, Boosting, and Stacking showed promise with better performance than single classifiers, the research performed was more focused on homogeneous ensembles, and heterogeneous ensemble models like stacking received less attention.

Another problem that arises is the indiscriminate inclusion of based classifiers. Previously, ensemble model research on homogeneous ensembles tended to integrate all candidate base classifiers without selecting only the ones with the best performances. This led to the inclusion of classifiers with poor predictive performance, which consequently reduced the performance of the final ensembled model. In response to that specific problem, a selective ensemble model approach was created. A selective ensemble model chooses base classifiers based on their performance and assigns different weights to optimize overall effectiveness. Researchers tested this method, which performed well but had yet to be applied to the default discrimination model. The selective ensemble acknowledges that no single model is universally best given the nature of the problem, data structures, and evaluation metrics. The super learner heterogeneous ensemble model addresses these issues using the selective ensemble model approach to surpass the limitations of traditional statistical methods and previous AI models used in personal credit default discrimination development. The super learner heterogeneous ensemble model not only improves prediction accuracy but also offers a customized solution to the specific data structure of each credit dataset. This advancement in AI for credit risk assessment is essential because it provides an enhanced credit risk assessment that improves the accuracy of credit risk prediction and significantly reduces losses due to bad debts while increasing profitability and better financial institutions' decision-making.

Approach to AI

The super learner, a heterogeneous ensemble model implemented in this paper, aims to solve the problem of personal credit default discrimination by predicting the probability of default based on the borrower features. This paper's super learner model is a heterogeneous ensemble model. It combines multiple single classifiers with ensemble models to increase the final model's overall predicting accuracy. It trains many different base classifiers using ten-fold cross-validation to minimize the weighted total loss of all base classifiers (Gang Li et al. 3). The super learner heterogeneous ensemble model is a heterogeneous ensemble model that employs a stacking algorithm. A heterogeneous ensemble model is an ensemble model that employs several base classifiers with various types of algorithms to ensure diversity in the final ensemble model developed. According to Sean a. Gilpin and Daniel m. Dunlavy in their paper "Heterogeneous Ensemble Classification," they said: "Using different types of base classifiers leads to diversity…. Different base classifier types can have different internal representations and may be biased in different ways. This leads to classifiers that will disagree with each other to some extent over a set of data instances covering a wide range of the feature space. This disagreement between the base classifiers is essential for the success of an ensemble classifier and is what we refer to as diversity" (Gilpin. et al.,93).

How does the super Learner model work?

Creating a super learner can be divided into three steps.

- Building the candidate base Classifier Library

-Calculating the Total Classification Loss of the Candidate Base Classifier

-Solving the Optimal Weight of Each Base Classifier and Building the Super Learner Heterogeneous Ensemble Model

Building the candidate base Classifier Library

This initial step involves building an algorithm library using the stacking algorithm. This library contains a wide variety of base classifiers. The candidate base classifier library in this paper uses first single classifiers like logistic regression (LR), lasso regression (Lasso), K-nearest neighbor (KNN), support vector machine (SVM), neural network (NN), decision tree (DT), then adds ensembled classifier with excellent performance like random forest (RF), GBDT, and XGBoost algorithms to further enhance the prediction performance and robustness of the model. These algorithms were chosen because logistic regression is one of the single classifiers used in credit scoring with good prediction. The ensemble methods added are known to perform better than traditional models (Gang Li et al. 5).

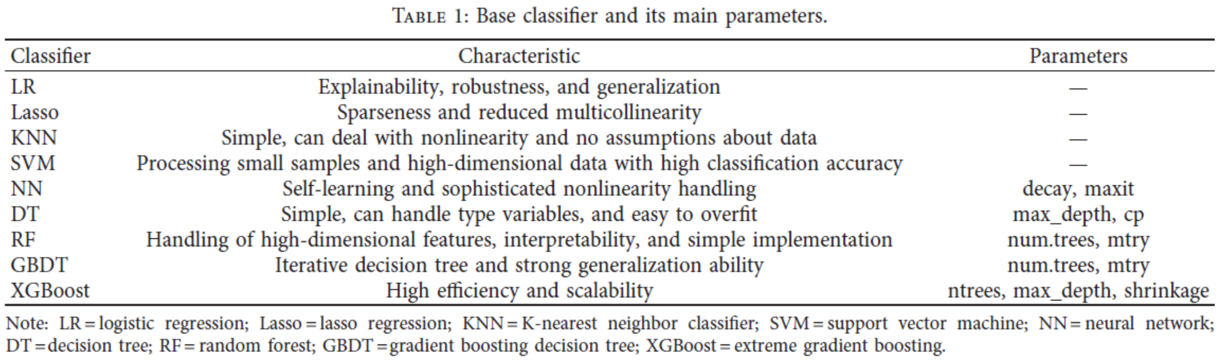

TABLE 1: Base classifier and its main parameters.

Classifier	Characteristic	Parameters
LR	Explainability, robustness, and generalization	—
Lasso	Sparseness and reduced multicollinearity	—
KNN	Simple, can deal with nonlinearity and no assumptions about data	—
SVM	Processing small samples and high-dimensional data with high classification accuracy	—
NN	Self-learning and sophisticated nonlinearity handling	decay, maxit
DT	Simple, can handle type variables, and easy to overfit	max_depth, cp
RF	Handling of high-dimensional features, interpretability, and simple implementation	num.trees, mtry
GBDT	Iterative decision tree and strong generalization ability	num.trees, mtry
XGBoost	High efficiency and scalability	ntrees, max_depth, shrinkage

Note: LR = logistic regression; Lasso = lasso regression; KNN = K-nearest neighbor classifier; SVM = support vector machine; NN = neural network; DT = decision tree; RF = random forest; GBDT = gradient boosting decision tree; XGBoost = extreme gradient boosting.

Table 1: Base classifier and its main parameters (Gang Li et al. 6).

Table 1 shows the different algorithms in the candidate library. These algorithms have different functions, contributing to the base classifier's diversity.

Calculating the Total Classification Loss of the Candidate Base Classifier

Once the candidate library is created, the ten-fold cross-validation method divides the original dataset into ten equal-sized subsets. It uses nine divided subsets as the training set, while the remaining subset is used as a test set to evaluate the classifier performance. This method ensures that every subset acts as a test set. The cross-validation method trains each base classifier in the candidate library and evaluates their prediction performance using the test set. The prediction result is stored, and the same procedure repeats itself ten times for each base classifier until each subset becomes a test set that could be used to evaluate the model prediction. After obtaining ten sets of predictions, the total classification loss of the base classifier is calculated by using the deviation of the ten sets of predictions of the base classifier against the ten-test set previously generated. The lower the total classification loss, the higher the discrimination performance (Gang Li et al. 4).

Solving the Optimal Weight of Each Base Classifier and Building The Super Learner Heterogeneous Ensemble Model.

Using a selective ensemble model approach to select the best-performing base classifiers or to give different base classifiers different weights for the ensemble model, the goal in this step is to minimize the total weighted loss of all base classifiers to obtain an optimal weighting model. The best-performing base classifiers are selected so that classifiers with a smaller total loss get larger weights. The weighting model is established with the constraint that the weight of each base classifier is nonnegative, and the sum of the weights is 1. "It may even happen that some classifiers are not considered during the ensemble process to make the prediction performance of the ensembled model better; that is, the weight value of the classifier is 0" (Gang Li et al. 7). Finally, the selected base classifiers are fitted on the complete dataset. Their optimal weights are used to construct the Super Learner heterogeneous ensemble model. This final model obtained includes the strength of the best optimal performance of the base classifiers and the capacity to dynamically adjust the weight of each base classifier to adapt itself to any dataset structure.

Data and features used in the Super Learner Heterogeneous Ensemble Model

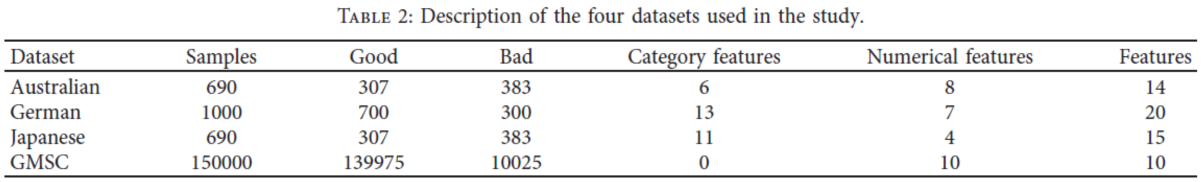

TABLE 2: Description of the four datasets used in the study.

Dataset	Samples	Good	Bad	Category features	Numerical features	Features
Australian	690	307	383	6	8	14
German	1000	700	300	13	7	20
Japanese	690	307	383	11	4	15
GMSC	150000	139975	10025	0	10	10

Table 2: Description of the four datasets used in the study. (Gang Li et al. 8).

The super learner heterogeneous ensemble model uses four real credit datasets: the Australian, Japanese, and German datasets from the UCI and a large real credit dataset of Give Me Some

Credit (GMSC) from the Kaggle platform. These datasets contain both categorical and numerical features.

The good column in Table 2 shows the number of people in the dataset that were able to pay their loans on time(non-default). The bad column in Table 2 shows the number of people in the dataset who were not able to pay their loans on time(default)

The Australian dataset contains 690 samples, with 307 non-default samples and 383 default samples. Each observation datum in the Australian set contains six categorical features, eight numerical features, and a result variable (accept or reject). The German dataset contains 1000 samples, with 700 samples, which are non-default and 300 default samples. Each observation datum in the German datasets contains 13 categorical features, seven numerical features, and a result variable (accept or reject). The Japanese dataset contains 690 samples, with 307 samples that are non-default and 383 that are default samples. Each observation datum in the Japanese set contains 11 categorical features, four numerical features, and a result variable (accept or reject). The GMSC dataset contains 15000 samples with 139975 samples, which are non-default and 10025 are default samples. Each observation datum in the GMSC set contains 10 numerical features and a result variable (accept or reject). The Japanese and the GMSC datasets contain missing values that must be addressed before the model construction.

Data Preprocessing

This paper uses a multistep data preprocessing to replace the missing data in the credit data set. The multistep includes three steps: missing value filling, qualitative index virtual coding, and data standardization.

Missing Value Filling

The first step used in this paper to preprocess the credit data involves filling in the missing data in the dataset. The method used to fill in the missing values depends on the data type in the original dataset. A new category is created for categorical features to replace the missing values. For numerical features, the mean values of the features are used to replace the missing values. In the Japanese and the GMSC datasets, missing values are filled, and a complete dataset is obtained now.

Qualitative Index Virtual Coding

The second step involves using a virtual or hot encoding, as seen in class. This process quantifies the categorical variable into a binary variable (0, 1) according to the feature category, for a feature with a category k, k-1, dummy variables are created to prevent multicollinearity. Multicollinearity occurs when independent variables in a regression model are extremely correlated. This means that a linear relationship exists between them. (Alin 2) This second step is crucial because it ensures that the encoded variable remains independent, thus eliminating the risk of multicollinearity. The Japanese, Australian, and German datasets containing classification features are processed with the dummy variable created from the virtual encoding.

Data standardization

Models like support vector machines are based on distance metrics sensitive to the difference in the order of magnitude between data. To eliminate the numerical differences between features, the data is standardized. The Z-score standardization method is used in this paper to standardize the data. The Z-score standardization normalizes the features based on their mean and standard deviation. After this step, a complete dataset is obtained.

After the multistep preprocessing, the dataset is divided into training and test sets in an 8:2 ratio. 80% of the data is used for training the model, and 20% is used for testing the model's effectiveness. To further improve the model performance by optimizing the model parameters through the enumeration process.

Success Criteria

Four Indicators were used to measure the classification performance of the super learner created.

Model Accuracy

The model accuracy represents the model's overall accuracy in the context of credit risk prediction. A high accuracy means that the model can effectively identify people who are both non-defaulters (True positive) and defaulters (True Negative)

Area Under the curve (AUC)

The AUC is a probability metric that assesses the model's ability to differentiate between classes in the context of credit risk assessment (defaulter and non-defaulter). This value is derived from the Receiver operating characteristic (ROC) curve, which plots the true positive rate against the false positive rate under various threshold levels. A high AUC value indicates a better model performance in avoiding false classification while accurately identifying true cases.

Type I error (False positive Rate)

A type I error is a metric representing the number of times a non-defaulter is incorrectly identified as a defaulter. A lower type I error is good as it means that fewer borrowers will be wrongly classified as defaulters.

Type II Errors

A type II error is a metric representing the number of times a non-defaulter is incorrectly identified as a non-defaulter. A lower type II error is crucial for financial institutions because it represents the likelihood of giving loans to customers who will likely default.

The super learner created in this paper adapts its choice of base classifiers and their weight coefficient according to the specific dataset. This allows the model to optimize its performance based on the characteristics of each dataset. The Australian dataset, for instance, only used three based classifiers: LR, KNN, and GBDT, with GBDT having the highest weight coefficient, indicating its crucial role in building a super learner on the Australian dataset. Other classifiers with a weight of 0 were not considered in the ensemble process. Similar to the Australian dataset, for the German, Japanese, and GMSC datasets, the super learner chooses different combinations of classifiers, selecting only those that perform the best. The super learner has demonstrated strong performance across different datasets. It was only sometimes the top performer, but it maintained either the best or the second-best performance in most cases. Also, the super learner was compared with ten other models using the four-evaluation metric discussed earlier. The evaluation showed that while no classifier is the best for any dataset, the super learner's ability to adapt to different data structures and optimize its classifier selection makes it one of the best models for credit risk assessment. Among the ensemble models with the best performance, RF and XGBoost have high performance. Having said that, their discrimination accuracy is slightly inferior to the super learner. This shows that ensembled models are better than single classifiers in most cases.

Social Impact

While there is no section on the social impacts in the paper, we can infer that with the accuracy and robustness of the model, an increase in prediction accuracy directly translates to massive profitability for the financial institution because it will reduce losses due to bad debts and increase profitability. Except for financial institutions, the other stakeholders that this new model will affect are borrowers and credit analysts/underwriters. The model affects borrowers by determining their default risk, influencing loan approval decisions and interest rates. This model can directly affect credit analysts or underwriters in assessing a borrower's creditworthiness. Regarding ethical considerations or concerns, the model might be biased or inaccurate, lacking representativeness, and not transparent or accountable (Credit Scoring Approaches Guidelines-Final-Web 41). While the research in the paper focused on improving the model accuracy, no part of the research looked into how to avoid or mitigate bias in the dataset set used or the model created. This is a critical ethical concern as it can unfairly affect loan approvals for certain groups of borrowers. If the dataset does not represent the broader population, the super learner model might not perform equitably across different demographic groups. There is a tradeoff between accuracy and Interpretability. The super learner is designed as a black-box predictive model whose complexity makes it difficult for stakeholders to understand how decisions are made. This problem is critical, especially for a system like credit risk assessment, because it can impact many people's lives. Rachel Thomas, director of the University of San Francisco's Center for Applied Data Ethics, says, "When AI makes decisions that impact people's lives, then not having an explanation is incredibly frustrating," she says. "But an explanation alone is not sufficient. There needs to be some sort of system for recourse, such as the ability to appeal decisions" (4).

Guidance for Coexistence

Although there is no section on how this new system should be integrated ethically into society and how the balance between human judgment and AI prediction can enhance decision-making, several guidelines can be implemented. The system created with the super learner heterogeneous ensemble model needs to be regularly tested for biases, and the developers need to mitigate any biases they find as soon as possible to have a fair system. Additionally, financial institutions need to pair up the model prediction with a credit risk assessment expert's judgment before deciding about a potential borrower. Also, adequate training needs to be given to the people interacting with the system to understand the model's capability, limitations, and risks. Furthermore, there should be a periodical feedback process where the system prediction and human judgment are reviewed to identify areas of improvement and how to enhance the decisions made by both the system and the human.

Conclusion

Gang Li et al.'s research on the super learner heterogeneous ensemble model represents a significant advancement in credit risk assessment. This model introduces an effective approach to predicting personal credit default by overcoming the limitations of traditional statistical methods and previous AI models. The application of this model holds substantial implications for financial institutions, borrowers, and credit risk analysts/underwriters. For financial institutions, increased predictive accuracy can reduce loss from bad debts and improve profitability. On the other hand, borrower benefits from fairer and more accurate assessments of their creditworthiness. Credit risk analyst/underwriters may see their job change and include a new tool or methodologies. Despite these advantages, the model's complexity and potential bias raise ethical concerns. It amplifies the need for continuous evaluation and mitigation of biases and the importance of balancing AI predictions with human judgment. This approach enhances decision-making and ensures that the model remains fair and representative of diverse demographic groups. As financial institutions increasingly rely on AI for decision-making, the super learner heterogeneous ensemble showcases the potential of AI to transform the financial industry.

Work Cited

Alin, A. (2010), Multicollinearity. WIREs Comp Stat, 2: 370-374. https://doi.org/10.1002/wics.84

Clark, Lindsay. “Grilling the Answers: How Businesses Need to Show How AI Decides.” *ComputerWeekly.com*, 9 Jan. 2020, www.computerweekly.com/feature/Grilling-the-answers-how-businesses-need-to-show-how-AI-decides.

Ghodselahi, Ahmad, and Ashkan Amirmadhi. “Application of Artificial Intelligence Techniques for Credit Risk Evaluation.” *International Journal of Modeling and Optimization*, Jan. 2011, pp. 243–49, doi:10.7763/ijmo.2011.v1.43.

Gilpin, Sean A., and Daniel M. Dunlavy. "Heterogeneous ensemble classification." *CSRI SUMMER PROCEEDINGS 2008* 90 (2008).daniel-dunlavy-2009-SAND2009-0203P.pdf (sandia.gov)

Li, Gang, et al. “Personal Credit Default Discrimination Model Based on Super Learner Ensemble.” *Mathematical Problems in Engineering*, vol. 2021, Mar. 2021, pp. 1–16, doi:10.1155/2021/5586120.

Shi, Si, et al. “Machine Learning-driven Credit Risk: A Systemic Review.” *Neural Computing and Applications*, vol. 34, no. 17, July 2022, pp. 14327–39, doi:10.1007/s00521-022-07472-2.

CREDIT SCORING APPROACHES GUIDELINES-FINAL-WEB. thedocs.worldbank.org/en/doc/935891585869698451-0130022020/CREDIT-SCORING-APPROACHES-GUIDELINES-FINAL-WEB.