Statistical Learning with Applications
CSC 542

Final Project
Statistical Learning Methods for Credit Score Predictions

Gabriel Huang
Nina Phan
College of Arts and Sciences, Miami Herbert Business School

Professor Vanessa Aguiar
University of Miami
May 2024

1. Introduction

Financial institutions have used credit scoring and decision models for many years to predict the risk associated with lending to individuals or other entities, but the introduction of machine learning has played an essential role in refining and improving the accuracy of these predictions for decision-making. Machine learning models' greater accuracy with their ability to analyze large volumes of different kinds of data creates the potential to increase access to credit for millions of people, especially underrepresented groups. However, as innovation in machine learning continues to grow, concerns about interpretability and fairness in machine learning models for credit scoring become more relevant for regulators today. Since credit scoring is a largely regulated industry of machine learning, models with less complexity and fewer input features are preferred in general, while remaining fair and unbiased to not discriminate against certain groups.

Achieving a robust, accurate, and unbiased credit scoring model depends on the type of model and deciding which input features are most relevant. Stakeholders of financial institutions today are concerned about whether certain models can be trusted for a sensitive application such as credit scoring, and with the rapid growth of new technologies and the creation of vast amounts of data on a per-second basis, the need for filtering out irrelevant features has increased in recent years. Machine learning models' ability to identify a wider range of relationships in training data increases concerns about the risk of replicating or even amplifying historical disparities in the credit context, hindering the initiative of expanding access to credit with fair lending. Features that may not be related to credit risk can lead to poor performance of models or a misleading interpretation of credit scoring.

The ability for stakeholders to a particular model to access information about its design, use, and performance is another crucial element towards trustworthiness with machine learning credit scoring models. Without sufficient transparency, firms nor regulators can evaluate whether a particular model is making credit decisions based on strong and fair relationships between an applicant's behavior and creditworthiness. The complexity of machine learning credit score models that makes it difficult to understand how a model assesses a particular individual's creditworthiness is what fuels the accuracy of these models, however. This tradeoff between strict regulation and the advancement of data analytics is investigated in our project, as we explore various feature selection models, and regression and classification statistical learning models for predicting the Probability of Default (PD) and credit score. The considerations that influence the choice of a model for credit scoring have led us to the following questions:

- What are the most important features of a credit scoring model for classification and regression?
- Do simpler models or more complex, less transparent models perform better for predicting PD and credit score?

- What combinations of feature selection and machine learning models are best for predicting PD and credit score?

This project provides a methodology for selecting key variables in establishing models for credit scoring and PD. We implement different classification algorithms for PD (logistic regression, Decision Trees, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Random Forest (RF)), regression algorithms for credit score prediction (linear regression, Decision Trees, Random Forest, and Boosting), and four feature selection methods (Forward Step Selection, Backward Step Selection, Best Subset Selection, and Lasso). We measured the performance and simplicity of these different methods using mean squared error, accuracy, and the number of selected features.

2. State-of-the-Art

With the increasing complexity of credit scoring, several approaches to designing robust and efficient credit scoring models have been proposed. Machine learning ensembles have gained prominence recently because of their ability to recognize and adapt to changes in data patterns and lending behavior. Tsholofelo Mokheleli and Tino Museba [4] applied the Adaptive Dynamic Heterogeneous Ensemble (ADHE) in which the models are updated regularly with changes in behavior from loan applicants. However, their study has limitations in being computationally heavy and using only one method of feature selection. Several other works have explored ensemble models to improve the adaptability and robustness of credit scoring models, but they overlook the issue of interpretability, which can raise privacy concerns and regulatory issues. Alagic et. al. [1] integrated mental health data into their machine-learning algorithms to enhance credit risk prediction. They found that XGBoost and Random Forest achieved the highest accuracy compared to simpler machine-learning models. However, they did not implement any feature selection techniques and their research has limitations in the ethical complications of including mental health data in the loan approval process.

Comparisons between conventional methods of machine learning and deep learning have been proposed in the field of credit score prediction, with several different datasets being utilized for our similar problem. Swati Tyagi [6] integrated explainable AI to improve the interpretability of non-linear “black box” models such as neural networks. However, neural networks did not perform better than other machine learning models such as Random Forest and LDA. Arram et. al [2] explored a new dataset from an American bank containing 34 features, with 24 being anonymized and the others being individual’s payroll information. They found that the multi-layer perceptron neural network performed best in terms of recall, their primary concern to protect banks from potential default customers. However, neural networks and the anonymized features do not provide much interpretability in credit scoring models. Sara Kornfield [3] conducted her research using 44 features of financial institutions rather than individuals to predict their probability of default. She found that logistic regression performed the best in terms

of accuracy for modeling the probability of default. The research, however, did not include any feature selection procedures for the conventional method, which could raise concerns about the interpretability of the model and the high correlation between features. Shrawan Trivedi [5] utilized a publicly available German credit scoring dataset, which included 20 features concerning individuals' financial and demographic characteristics, to predict the probability of default. He explored 3 feature selection techniques and found that Chi-Square feature selection and Random Forest produced the best false positive and false negative rates. The dataset, however, was not highly comprehensive, as most of the features concerned individuals' demographics and not financial information, and did not optimize how many important features should be included in their models.

Most recent research only focuses on the classification problem of predicting the probability of default and not improving the ability to predict credit scores. Our project will explore models of both classification for PD and regression for predicting credit score, while identifying key features using different feature selection techniques, to determine which model performs better in terms of accuracy and interpretability.

3. Material and Methods

Our project utilizes a comprehensive data set of individuals' financial status and behaviors to predict credit scores and the risk of default to explore this opportunity of enhancing the performance and robustness of credit risk assessment models. The dataset contains 1000 observations and 84 features, but customer identification data was excluded. There were no missing values in the dataset. One of the problems with our data was an extreme data imbalance, as there were much more features that did not result in a credit default than observations that did. Exploring new ways to adjust for this data imbalance would be a good starting place for future research.

For our regression tasks aimed at predicting credit scores, we employed a myriad of algorithms—Multi-Regression, Bagging, Random Forest, and Boosting—chosen for their ability to manage complex datasets and accurately predict continuous outcomes. Similarly, our classification approaches included Logistic Regression, Decision Trees, Random Forest, and K-Nearest Neighbors (KNN), each selected for their effectiveness in discrete outcome prediction.

To refine our models, we applied various feature selection techniques such as Forward Stepwise Selection, Backward Stepwise Selection, and Lasso to optimize the input features and improve model performance. We assessed the effectiveness of our models using a split of 80% training data and 20% testing data. The evaluation metrics employed were mean squared error (MSE) for regression models, and accuracy, recall, precision, and F1 score for classification models, ensuring a comprehensive understanding of each model's predictive power and accuracy.

4. Results

The classification models did not demonstrate outstanding performance, although they were all very consistent. Logistic Regression achieved a 71% accuracy and a confusion matrix that struggled with Type II errors, as there were 45 observations with predicted values of 0 and actual values of 1. K-Nearest Neighbors reported slightly better accuracy at approximately 72%. Both Decision Trees and Random Forest showed a balanced performance in predicting classes with an accuracy of 68%, although they displayed a symmetrical balance of Type I and Type II errors. However, the best out of any of the models was the one that employed Lasso regression. This model's lasso regression only picked out nineteen of the features to use in the model, including interesting features such as `CAT_GAMBLING`, `CAT_MORTGAGE`, and `DEBT`.

Classification Method	Logistic Regression	K-Nearest Neighbors	Decision Tree	Random Forest	Lasso	Forward Stepwise	Backwards Stepwise
Accuracy	0.71	0.72	0.74	0.74	0.76	0.74	0.7
Precision	0.88	0.89	0.90	0.92	1	0.92	0.88
Recall	0.75	0.76	0.77	0.76	0.75	0.77	0.75
F1 Score	0.81	0.82	0.83	0.83	0.86	0.84	0.81

For regression, Linear Regression performed the best out of any of the models, with an MSE of 641.96 and an R-squared value of 0.844, indicative of a substantial fit to the data. Bagging produced an MSE of 816.21, which although was worse than Linear Regression, was better than the other tree base methods. Random Forest and Boosting techniques incurred higher MSEs of 926.69 and 934.77, respectively. These findings suggest that while capable of managing complex datasets, the advanced tree models may not capture the nuances of credit scoring as effectively as simpler models.

Regression Method	Linear Regression	Bagging	Forward Stepwise	Backwards Stepwise	Lasso	Random Forest	Boosting
MSE	0.155	0.210	0.157	0.164	0.155	0.266	0.201

While none of the models are particularly strong, predicting credit scores and defaults is inherently challenging due to the complex and often unpredictable nature of human behavior. Individuals' financial decisions can be influenced by a bunch of unforeseen personal and economic factors, making it difficult to capture and predict through models. The anarchic tendencies in human financial behavior mean that even the most sophisticated models can fail to forecast accurately, as they cannot account for all the variables at play. By examining the models

inside of some of the previous research done, it is apparent that our models perform just as well, if not better, than most of theirs, including those that used deep learning.

Our analysis indicates that for predicting credit scores, simpler models like Linear Regression may provide a better balance between accuracy and interpretability. Meanwhile, the choice among classification models may hinge on specific needs for sensitivity versus specificity. Future endeavors will focus on refining feature selection techniques and optimizing model parameters to enhance both classification and regression outcomes in the context of credit scoring.

References

[1] Alagic, Adnan, Natasa Zivic, Esad Kadusic, Dzenan Hamzic, Narcisa Hadzajlic, Mejra Dizdarevic, and Elmedin Selmanovic. 2024. "Machine Learning for an Enhanced Credit Risk Analysis: A Comparative Study of Loan Approval Prediction Models Integrating Mental Health Data" Machine Learning and Knowledge Extraction 6, no. 1: 53-77. https://doi.org/10.3390/make6010004

[2] Arram, Anas, et al. “Credit Card Score Prediction Using Machine Learning Models.” *Credit Card Score Prediction Using Machine Learning Models: A New Dataset*, arxiv.org/pdf/2310.02956.pdf. Accessed 7 May 2024.

[3] Kornfield, Sarah. *Predicting Default Probability in Credit Risk Using Machine ...*, www.diva-portal.org/smash/get/diva2:1437874/FULLTEXT01.pdf. Accessed 7 May 2024.

[4] Mokheleli, Tsholofelo & Museba, Tino. (2023). Machine Learning Approach for Credit Score Predictions. Journal of Information Systems and Informatics. 5. 497-517. 10.51519/journalisi.v5i2.487.

[5] Shrawan Kumar Trivedi, A study on credit scoring modeling with different feature selection and machine learning approaches, Technology in Society, Volume 63, 2020, 101413, ISSN 0160-791X, https://doi.org/10.1016/j.techsoc.2020.101413.

[6] Tyagi, Swati. *Analyzing Machine Learning Models for Credit Scoring ...*, www.aijbm.com/wp-content/uploads/2022/01/B510519.pdf. Accessed 7 May 2024.