Sensitivity of Credit Scoring in case of Machine Learning Utilization

Zahra Moalleh
Matrikelnummer 5830879
Zahra.moalleh@student.uni-tuebingen.de

Abstract

Credit Scoring is a tough subject in *Banking & Financial Sector*. A variety of statistical approaches are applied to estimate the expected loss and credit risk through client credit evaluation. This procedure is sensitive since incorrect scoring can overstate the risk of default and exacerbate a financial institution's credit risk, which can lead to bankruptcy.

On the other hand, from the standpoint of data science, improving the capacity of analyzing and extracting useful information from large and complex datasets allows for the provision of insights that were before impossible or just marginally achievable.

The purpose of this work is to take a closer look at what should be addressed while adopting new high-power data analysis models and specifically machine learning. It aims to bring attention to the fact that some businesses are sensitive, and some uses of data analysis models make things more complicated for data scientists.

1 Introduction

Development of processing and data analysis provide new advancements in credit scoring like many other applications of data analysis. Nevertheless, it is still a debate that which methods are applicable in this problem. There is a huge literature regarding application of machine learning and specially classification models in credit scoring. Gunnarsson et al. (2021) investigated the applicability of deep learning in credit scoring and stated that: "Deep neural networks do not outperform their shallower counterparts and are considerably more computationally expensive to construct".
Most of the existing literature evaluate their classification model performance with widely known classification evaluation metrics and make their statements based on these standard performance indicators in machine learning context. Dastile et al. (2020) provides a useful figure of all evaluation metrics that were used in primary studies of their literature survey. They stated that the most utilized evaluation metrics are: The Percentage Correctly Classified (PCC), Operating Characteristics Curve (AUC), Type I Error and Type II Error. In this paper we want to shed light on the fact that what ever is the classification method to classify the loan demanders' eligibility, there are some classification evaluation metrics that become more sensitive when working with datasets like credit allocation. Credit allocation is a sensitive process of how a bank divides its financial resources of credit and poor credit allocation strategies could escalate the credit risk and even lead into bankruptcy of the financial Institution. The paper organised as follows. Section 2 provides a brief overview of Credit Scoring in Banking and Finance Sector. Section 3 includes .... and discussions regarding the are given in Sect. 4.

2 A brief overview of Credit Scoring in Banking and Finance Sector

The probability of loss arising from a customer's failure to repay the debt or loan obtained from a financial institution is referred as *Credit Risk*. While it's hard to predict who will fail on their

commitments, correctly analyzing and mitigating the credit risk can assist in reducing the magnitude of a loss. Default Risk, which occurs when borrowers are unable to fulfill contractual payments, is the form of credit risk we are focusing on in this study. "Credit scoring is a method that helps the bank to rationalize its process for credit granting decision" (Zaghdoudi, 2016). The general concept is to combine a series of financial indicators that can distinguish high and low performing consumers.

Here are some of these financial indicators related to what will be explored later in this study:

- *Risk Asset* is an asset owned by a bank or financial institution whose value may alter owing to various factors such as default risk.
- *Exposure at default (EAD)* is the entire value, in case of default, that a bank is exposed to.
- *Probability of Default (PD)* is the likelihood that a borrower would be unable to make anticipated repayments over a specific time horizon. It estimates the chance of a borrower failing to satisfy his or her financial commitments.
- *Loss Given Default (LGD)* is the amount of money, in case of default, lost by the financial institution.
- In drawing things to a close, the *Expected Loss* will be defined in the following way:

$$ExpectedLoss = EAD \times PD \times LGD \tag{1}$$

With the aforementioned definitions in mind, to examine closely at how the scoring process works, *Deutsche Bank* referred to Credit Scoring as a *Classification Problem*:

"The rating is a classification of the creditworthiness of borrowers or bond debtors carried out using a uniform procedure."

To this end a widely used dataset regarding *Credit Scoring Classification Problem* has been investigated in this paper which will be in detail presented in next section.

3 Data

The "Home Equity Dataset (HMEQ)[1] reports characteristics and delinquency information for 5,960 home equity loans. Table1 provides both dependent and independent variables description.

Table 1: HMEQ variables description

Variable	Dependency	Type	Description
BAD	DV[2]	Binary	*1* = customer defaulted on the loan or is seriously delinquent *0* = Otherwise
REASON	IV[3]	Binary	*DebtCon* = debt consolidation *HomeImp* = home improvement
JOB	IV	Nominal	Occupational category
LOAN	IV	Interval	Requested loan amount
MORTDUE	IV	Interval	Amount due on existing mortgage
VALUE	IV	Interval	Value of current property
YOJ	IV	Interval	Years at present job
DEROG	IV	Interval	Number of major derogatory reports
DELINQ	IV	Interval	Number of delinquent credit lines
CLAGE	IV	Interval	Age of oldest credit line in months
NINQ	IV	Interval	Number of recent credit inquiries
CLNO	IV	Interval	Number of credit lines
DEBTINC	IV	Interval	Debt-to-income ratio

[1] Available online: https://www.kaggle.com/ajay1735/hmeq-data

The dataset contains twelve features to classify the dependent variable which is the probability of default into binary classes.

3.1 Feature Engineering

Some preprocessing has been done after acquiring insights into the dataset by glancing at descriptive statistics and data information.

Regarding *Missing data*, apart for the "BAD" and "LOAN" features, all others contain missing values. Each features's missing values do not surpass 10% of the total raws, except for "DEBTINC" which is amount to 21%. It is not a decent idea to drop the missing data, specifically in such case that roughly 56% of the total dataset will remain. Meanwhile, completing the missing data is outside the scope of this study and the incomplete data has been removed.

Concerning the *outliers*, a different outlook has been acquired on "DEBTINC" while looking to all features. Figure 1 depicts the fact that above a certain level of Dept-to-income ratio (about 45.56) people are more likely to default (class 1). The Y axis is reduced from 120 to maximum of 60 for better visualisation of the mentioned threshold, although there were still some data points between these levels regarding class 1.

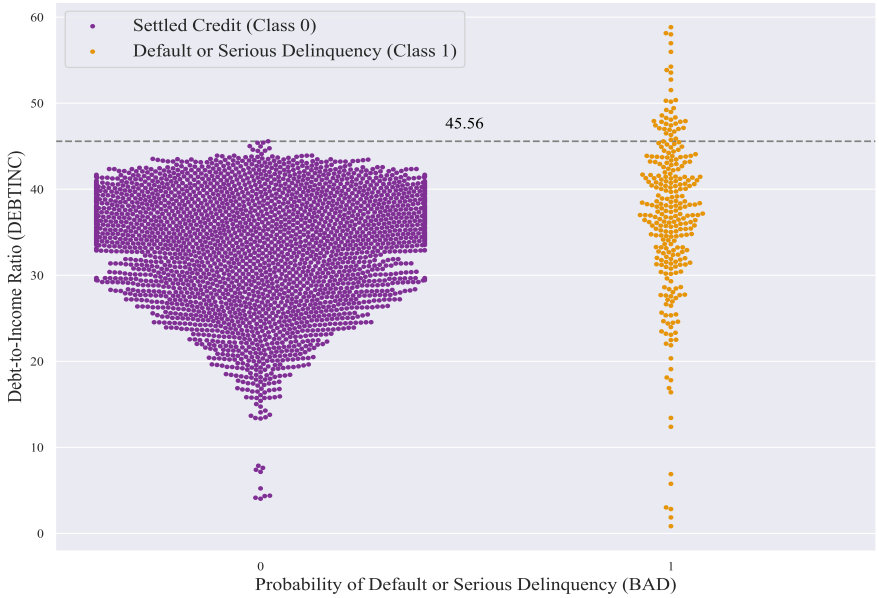

Figure 1: Debt-to-Income ratio threshold depending on settled credit class

Following step is to perform feature engineering to create processable feature set. To fix the qualitative "REASON" and "JOB" variables, converting to dummy variables and frequency encoding have been applied, respectively.

Last of all, the labels are unequally distributed containing 80% of "class 0" and 20% of "class 1". This issue give rise to an *Imbalanced Classification Problem*.

[2]Dependent Variable
[3]Independent Variable

4 Classification Results

Many recent studies like Dumitrescu et al. (2020) mentioned ensemble classification based on decision tree as the best performing model in credit scoring context. Therefore, a *Bootstrapping Random Forest Classifier* with 100 number of trees and 70-30% train-test dataset split using scikit-learn package is implemented. The result of RF classifier is compared with a dummy classifier as a baseline. Figure 2 & 3 provide heatmap regarding each classifier.

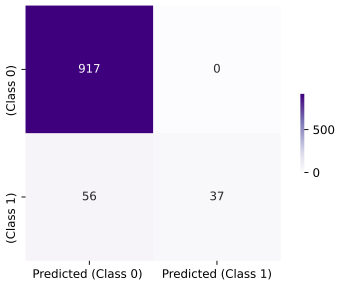

Figure 2: Random forest classifier heatmap

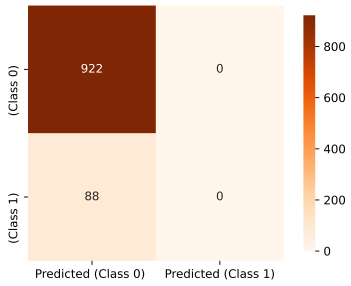

Figure 3: Dummy classifier heatmap

In case of interest, the model could be trained with more extensive real or even simulated data, the missing data could be fixed and some other hyper parameters could be tuned to gain a better prediction. However, as previously stated, the purpose of this research is not to train a more accurate model; rather, it is to give a sector-based application insight for data scientists who believe their job is done after developing a decent model as measured by standard evaluation criteria. All codes is available as online supplemental materials[4].

In the last section more detailed explanation on this subject is provided.

5 Results discussion from Credit Scoring Perspective

The goal of credit scoring is to categorize customers into different categories, each of which has its own PD and LGD as outlined in section 1. According to the above-mentioned binary classification model, there are two types of customers: the best and the worst. If probability of defaults (PD) equal to 1% and 99% assumed for each group. Considering the test dataset in the analysis 9.2% of the costumers defaulted in reality, but only 3.16% of them were predicted correctly by the model.

Considering Figure 2, among the real amount of 88 defaults, 32 were correctly predicted (True Positive) and 56 were wrongly classified as settled credit. Recalling Equation (1) the total expected loss could be written as follows:

$$\sum_{n=1}^{TP+FN} ExpectedLoss(i) = \sum_{n=1}^{TP+FN} EAD(i) \times PD(i) \times LGD(i) \tag{2}$$

which also could be written as :

$$\sum_{n=1}^{TP+FN} ExpectedLoss(i) = \sum_{n=1}^{TP} EAD(i) \times PD(i) \times LGD(i) + \sum_{n=1}^{FN} EAD(i) \times PD(i) \times LGD(i) \tag{3}$$

According to Equation (3), the second component, the summation of false negatives will result in an underestimate the estimated loss by wrongly lower estimated probability of default. To sum it up, a per unit rise in Type II error could also have inflated effect by impressive underestimation of PD. As previously stated, poor scoring model results in increase of expected losses and possibly bankruptcy.

[4]Available online: https://github.com/moalla299/CreditScoring_DataLiteracy

References

Dastile, X., Celik, T. & Potsane, M. (2020) *Statistical and machine learning models in credit scoring: A systematic literature survey. Applied Soft Computing Journal 91.*

Dumitrescu, E., Hué, S., Hurlin, Ch. & Tokpavi, S. (2020) *Machine Learning or Econometrics for Credit Scoring: Let's Get the Best of Both Worlds. HAL open science.*

Gunnarsson B. R., Broucke S., Baesens B., Óskarsdóttir M. & Lemahieu W. (2021) *Deep learning for credit scoring: Do or don't? European Journal of Operational Research.* **295**(1):292-305

Zaghdoudi Kh., Djebali N., Baesens B. & Mezni M. (2016) *Credit Scoring and Default Risk Prediction: A Comparative Study between Discriminant Analysis and Logistic Regression. International Journal of International Journal of Economics and Finance* **8**(4):39-53