Loan Default Prediction using Machine Learning: A Review on the techniques

Mafas Raheem
School of Computing
Asia Pacific University of Technology & Innovation
Kuala Lumpur, Malayisa
raheem@apu.edu.my

Swee Hong, Wong
MSc in Data Science and Business Analytics
Asia Pacific University of Technology & Innovation
Kuala Lumpur, Malayisa
TP065619@mail.apu.edu.my

Abstract – This paper aims to discuss the demand of machine learning models in the banking industry and using it to predict the loan default. The prediction is not merely for the purpose of credit scoring, it can also widen its access into capital reserve, risk management, loss forecasting and marketing campaign. At the same time, machine learning techniques offer benefits in managing big data to develop a more sophisticated scoring model as compared to the traditional econometrics' approaches. The performance of the model also relies on data pre-processing steps and business/credit consideration put into during its development process and it must be governed by a set of regulated frameworks. Missing value imputation, outlier treatment and variable transformation are some common practices in preparing the dataset, combination of treatments with business/credit consideration and variable predictive power assessment are rarely seen and worth further exploration. Logistic Regression has been a popular choice in credit scoring given its ability to generate probability as well as its high interpretability to end-users. Supervised learning like Tree-based models, Support Vector Machine, Artificial Neural Networks are gaining its popularity in recent decade and the ensemble methods should not be neglected as well. The model's end to end lifecycle should not be driven merely by achieving high accuracy but also other aspects like justifiable, transparency and ethical standards.

Keywords—Machine Learning Model, Credit Scoring, Data Pre-processing, Business driven analysis

I. INTRODUCTION

While the society is moving rapidly towards the digital era, banks are no longer the only financial provider with digital banking licences awarded to non-bank companies who are also capable in offering the lending services via the digital platforms. To maintain market competitiveness and operation efficiency, banks constantly seek new methods to improve their services in order to enhance its customer experience. The machine learning is one of the fast-growing areas in the financial services in supporting human functions to make prediction, decision and recommendation. Its applications include credit scoring, loss forecasting, fraud detection, customer service (chatbot), product offering and others (OECD, 2021). With lower human intervention and involvement, the data driven strategies bring potential cost savings and higher margin to the bank.

Credit underwriting is one of the crucial functions in the banks and machine learning (ML) techniques offer a variety option of descriptive and predictive analysis to empower this function. Being enabled by the abundance of affordable and computing capacity, ML models can now ride on both traditional data and non-traditional big data to identify patterns, learn quickly, and make fast and yet accurate approving decision to filter out risky and unqualified applicants. It is a gamechanger in transforming the old business model and benefits both the banks and their customers. In return of this great efficiency, the banks' growth is accelerated, customer stickiness increased and asset quality improved. An accurate prediction of customer risk using ML is hence crucial not only to the revenue, as well as loan lost management and cost control.

The success of a highly predictive model not only depends on the model algorithms, its performance also heavily relies on the model development process starting from problem statement understanding, data exploration, data pre-processing, customer segmentation and sampling, model training and testing. As the model is a mathematical representation of the business problem, its development process should be guided by balance between art and science to deliver the solution to the business quandary. On the other hand, there are also a set of regulatory accepted principles for the use of ML to be complied in voiding the heightening machine misuse risk.

This work discussed on the various data mining techniques used in the model development, its role and limitation in the banking industry, as well as potential areas that can be explored to address some issues faced in the bank's daily environment.

II. LITERATURE REVIEW

A. Banking Supervision Agreement on Capital Requirement

Under the International Banking Supervision Agreement set by Basel Committee (Bank Negara Malaysia, 2019), it calls for measurable risk and quantified regulation guidelines to manage and find balance between business growth and credit risk. Under the guidelines, the financial institutions are to develop its internal rating-based system as part of the credit underwriting policy to risk rank customer in order to monitor capital requirement and most importantly to manage loan loss expectation. Capital requirement is an amount of cash that held in the safe and it has to be highly liquidity in order to sustain the sudden events like operating losses or honourable withdrawals. Loan loss occurs when money lent is not paid back. Too high of a capital reserved impacts bank's profit as its primary business is to utilize depositor's fund and lend to the borrower and earn from the interest charges. Hence, the Risk-Weighted Asset (RWA) is a measurement that permits the bank to evaluate risk emphatically in order to minimize

the amount of capital required and subsequently maximize the lending capacity.

One of the key components in RWA is Probability of default (PD), as the name suggests, PD measures customer risk in predicting the chance of it going into default in the up-coming 12 months and the term "default" occurs when customer no longer able to repay the loan instalment continuously up to 90 days (and above). It is also commonly used as a decision tool for loan application and a basis for interest rate assignment where good customer gets lower rate and will be compensated by higher risk customer who bears higher rate where banks can still make profit. Other than credit application scoring, PD is also used in the collection process for early delinquency detection, debt recovery and management (Wijewardhana et al. 2018).

The framework has promoted technique like Logistic Regression Algorithm to be an immensely effective tool for the development of PD given its ability to generate probability and high model transparency. Besides, classical linear and probit regressions are also commonly employed in credit models (Wong, 2021).

B. *Advantages and Downsides of Logistic Regression model*

Some advantages of Logistic Regression Model includes: no linear relationship is required between the dependant variable (in this case, the defaults) and its independent variables, the residuals do not need to fit into a normal distribution, homoscedasticity of independent variables are not required and no specific distributional form required from independent variable (Schreiber, 2018). The lack of assumption in the Logistic Regression Model makes it much easier for the financial institutions to adopt when it comes to variables creation and selection. This is because data extraction is strictly governed by Central Bank, Personal Data Protection Act (PDPA) as well as customer's consent leading to limited available sources of data that comes in both metrics and non-metrics format. It's output of probability offers finer selection criteria that can be used for pricing (Wong, 2021). Lastly, the model has high transparency with measurable variable weights reflected in the coefficient, making it easy to explain to end-users for user acceptance and production endorsement.

The downside of Logistic Regression Model is that it requires the absent if not little multicollinearity among the independent variables (Senaviratna, 2019) which could potentially leads to the following and subsequently impacts model performance:

- lacking variables diversity as only one variable will be selected into the final model as it has to meet the p-value significancy test
- prone to suffer in the event of macroeconomic change or customer base shift from a new marketing strategy

Besides, it has strict requirement on the appropriate outcome structure (both Binary logistic regression and Ordinal logistic regression).

C. *The demand of machine learning model*

The rise of digital economy and technologies have brough huge changes to traditional business ecosystem and triggered the banks to rethink about its business model and level of financial inclusion. One of the examples is eCommerce platforms that created an entirely new business ecosystem and purchasing behaviour given its low start-ups/financial cost, flexible shopping hours, wider range of goods with comparative and attractive prices that allows it to reach out to a broader customer base from both local and internationally to shop with ease at the tip of finger. The purchasing channel via websites and phone applications generated huge volume of transaction data that has not been seen before and are valuable in assessing business growth and strength and is still rarely used in credit decisioning model (Libor et al., 2021). Secondly, the ever-advancing big data analytic architecture like Hadoop has offered a cheaper, more flexible and faster way to manage, process and transform the high-volume data and enables the development and implementation of various machine learning algorithms.

The ongoing health crisis hit in 2020 has accelerated the digital economy to flourish, besides, the stir it has caused in the job market and spike in unemployment rate have led to increasing demand of short-term personal loan and working capital loan for finances management during the pandemic. On the other hand, the prolong lock down period raises both debt levels and impaired loan ratio, forces the bank to search for a quicker and stronger credit measurement to manage the business opportunities and credit risk associated with these events.

In addition, back in the old days, loan approval went through very subjective underwriting process, causing inconsistent decision made by different standards and sometimes the process is prolonged when volume of loan application is high and the hands are tight. The credit model would allow decision to be made on a standardized basis with higher speed (Hoang et al., 2022).

There are some common machine learning methods that are widely recognized, the unsupervised learning is a technique used to study the underlying structure of the dataset, find similarities and hidden patterns without data label without dependant variable. In contrast, supervised learning is trained on labelled data using past behaviour in order to predict future event on a previously unobserved dataset. They differ from the traditional econometrics that study and explain the relationship between variables as shown in Table 1.

Machine learning models are supervised learning, its fundamental idea in a nutshell is using a set of data to train the model on characteristics and later the real data from the out-of-sample will be fed into the model to generate the output of the prediction. While the use of machine learning in the banking industry is still at its infancy stage, but the speed of its application is growing rapidly in the recent years (Majid, 2019).

TABLE I. APPROACHES AND PURPOSES

Approach	Data	Method	Usage/Purpose
Traditional Econometrics	Labelled dataset (with dependent variable)	Linear Regression (OLS)	Relationship Explanation
Supervised Learning	Labelled dataset (with dependent variable)	Supervised Learning	Predictions
Unsupervised Learning	Labelled dataset (without dependent variable)	Unsupervised Learning	Data Structure Inference

The application of of machine learning has been reflected in publications over the years where:

- in 2018, the number publications have increased 3x as compared the average of last 8 years
- in 2019, it has further increased 5x
- in 2020, the increase was nearly 7x
- in 2021, there were almost 11x publications more than before on machine learning (Hoang et al., 2022)

In many researches, machine learning algorithms have been proven to be a great tool in predicting default. It has the ability to deal with high-dimensional data from both credit or non-credit histories to predict the probabilistic nature of patterns and take a constructive decision.

D. Data Preparation and Pre-Processing

In most researches on the development of ML models for default predictions, it started with data cleaning process. Mode values assignment is common in missing values treatment for categorical variable, it replaces missing value with the highest frequency of each label in the variable. Mean values assignment is suitable for numerical variable, it can be taken at sample level or by label's level (Orji et al., 2022), the second method establishes a probability relationship with the dependent variable but may cause overfitting issue. There are other methods that are hardly mentioned and used:

Global constant where new value like "Unknown", "N/A" are imputed to replace the missing values where other imputed value may not make logical sense to the observations, in example, number of children is replaced with average value but its marital status = single.

Hot deck imputation an imputation technique that fill up missing value in non-respondent variable (recipient) when similar characters/trends are observed in other respondent variables (donor), however, it could lead to a discounted precision for variable with high missing values rate.

A combination of multiple imputation approaches or deterministic method is also worth further exploration to make the best fitting value.

In the extremely large dataset with number of variables, irrelevant variables are removed based on business' recommendation (Ereiz, 2019). An alternative way that could be considered is elimination by variable predictive power. The variable predictive power is also a measurement to evaluate the missing values imputation methods.

Besides missing values, outliers' detection via interquartile range and boxplot are some of the guidance commonly used. Sqrt transformation and log Transformation can be applied by taking the square root or log values of the numerical attribute to smooth the outliers and normalise skewed distribution. Percentile capping is another easy method to implement by capping the outliers at 1% and 99% percentile boundaries. Binning method is suitable for categorical variable (Cheng et al., 2021).

E. Dependent Variable (DV) and Imbalanced Class

Dependent variable (DV) is the purpose of the research, and it decides the type of data mining techniques to be used, for example, discriminant analysis is suitable when DV has more categories, linear regression can be used when DV is continuous. In classification problem, the key to the prediction performance is the count and accuracy of the dependant variable. A higher count of DV in the observations helps to improve prediction power (Daniel et al., 2022). In real life, the loan default data is known with its imbalanced classification problem for a simple reason that the bank has many credits and risk control in place starting from the point of loan onboarding up to collection activities to prevent loses. Imbalance class refers to dataset with skewed class proportion, larger distribution forms the majority class, and the rest is minority class. Minority class is often the more interest and importance for the prediction (Aida, 2015). It poses a challenge for ML as most classification algorithms were designed with equal class size assumption and the problem has high impact on the model accuracy. Ironically, imbalanced classification problem has not gotten more attention than it should.

F. Machine Learning model

The discriminant analysis was one of the basic methods employed in credit scoring since 1965 as shown in Fig. 1. The progression of the methodology has accelerated in the early 21st century. The growth of the algorithm addresses some of the limitations faced particularly from 2 dimensions, firstly the raw, larger volume, unstructured alternative data (including the non-credit/financial data) and secondly a more sophisticated techniques for descriptive and predictive analysis (Anil et al., 2021):

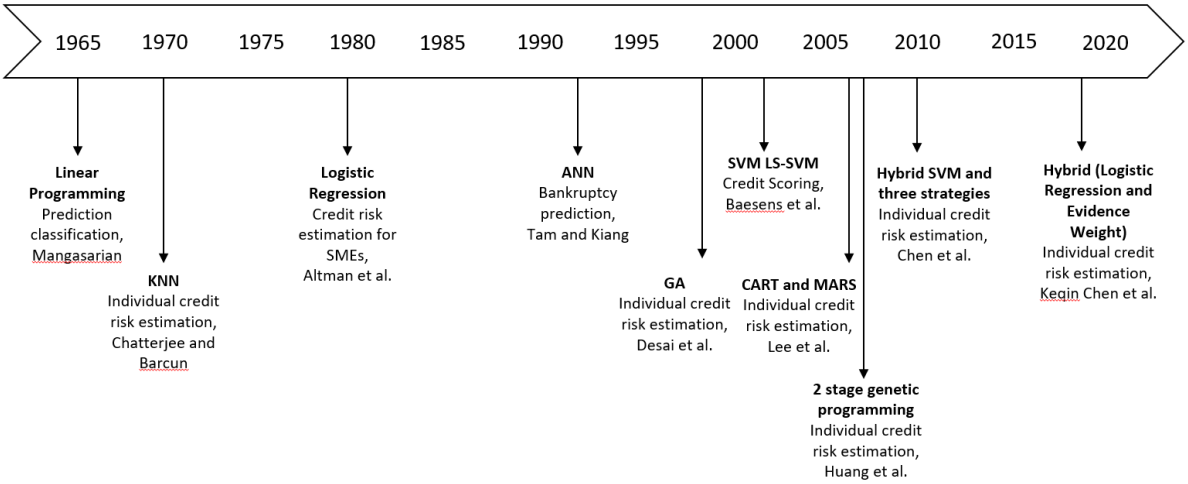

Fig. 1. Models/Analytic Techniques Introduced for Credit Risk

Recent machine Learning related literatures to banking were surveyed to compare against the statistics done in (Anis et al., 2021) work, results almost tally with popular machine learning models like Logistic Regression, Decision Trees, Random Forest, SVM, ANN that are repeatedly used as shown in Table II. Both Logistic Regression and Tree-Based Models (Decision Tree, Random Forest) were constantly found to outperform the rest most of the time:

TABLE II. FREQUENCY AND PERFORMANCE OF ML MODELS TESTED IN RECENT WORKS

Source	Logistic Regression	Random Forest	ANN	Decision Tree	KNN	Others
Zoran (2019)	94.70%	94.60%	82.10%			
Ugochukwu et, al. (2022)	80.00%	95.55%				
Jovanne et, al. (2022)	77.31%			78.38%	78.38%	NB – 76.54%
Bomvalue et, al. (2022)	81.50%	81.30%	76.50%	79.80%	75.40%	SVM – 80.8%
Swati (2022)	67.40%	72.50%	72.40%	66.34%		LDA – 67.4% QDA – 53.6% Stacking – 67.6%
Frequency Anil et, al. (2021)	4	3	3	2	NA	SVM – 3 Hybrid - 2

Below is some basic information on the highly used models except for Logistic Regression as it has been mentioned in the earlier sections:

Decision Tree

Decision Tree organizes a series of root in a flowchart structure to make classification choices, it starts at root node that represents the sample being analysed where information gain is calculated to further split into sub-nodes (with largest information gain), the process repeats recursively with other never-selected attributes and subsequently reaching the decision nodes. Decision Tree classifier can produce probability of each class by looking at the ratio. It is easy to visualize and interpret but prone to overfitting issue when the tree is deep, this problem can be overcome by limiting its maximum depth (Tyagi, 2022).

Random Forest

Bias increases when limiting the maximum depth of the Decision Tree, Random Forest is a good way to combat overfitting without sacrificing bias. Random Forest is an ensemble model as its name suggests, a collection of decision trees whose result will be aggregated into single final result; thus, the decision is much more robust. Unlike Decision Tree, Random Forest is inherently less interpretable (Orji et al., 2022).

Artificial Neural Networks (ANN)

ANN attempt to mimic human brain's neurons network and make decision in a humanlike manner, it is often described as a puff cake that are made up of 3 types of layers, naming the input layer, output layer and hidden layer(s) (Vladimir et al., 2021). However, ANN works well only with the numerical variables.

Support Vector Machine (SVM)

SVM is a method for classification that was first developed by Cortes and Vapnick (1995). It can categorize the data observations by separators called hyperplane in n-dimensional feature space: in the background, SVM tries to solve the convex optimization problem that maximizes the margin and where the constraints say that each category should be on the right side of the hyperplane. The biggest pros of SVM are that they are easy to understand, implement and interpret. Furthermore, they are effective when the training sample is smaller. The simplicity of SVM also

cause problems like in some applications, datapoints cannot be separated by hyperplane (Chitambira et.al, 2022).

The mentioned scoring models introduced can evaluate customer well, however, there are other important aspects to look into like decision stability, model inclusion for linear and non-linear dataset and a more effective variance avoidance method. These are some of the characters of Ensemble model, (Karalar et al., 2021) ran various combination of Ensemble models, compared them to its sub-models and proven its ability in boosting the model accuracy where the best combination is Extra Trees + Random Forest + Logistic Regression.

G. Model Selection

There are a few considerations to put in when it comes to final model selection. Model accuracy is of all the most important criteria as it evaluates how well the model can address the problem statement. However, it should not be the only standard, in most of the researches, a few morels were tested and there will be other candidate models with close accuracy. Hence, the ability of the models to generalize and not overfitting to training set can also be part of the consideration to avoid bias towards the research goal (Chitambira, 2022). Thirdly, the regulator has raised its emphasis on model transparency in the recent releases making it an upcoming topic and model interpretability should also be part of the selection process. Fourthly, in loss managing, sometimes it will also be helpful in look into the precision result (Type I Error) to detect false positive rather than only looking at the overall accuracy (Ereiz, 2019) (Lai-Yan, 2021).

H. Drawbacks in Machine Learning Models

Machine learning models' prediction power rises with its complexity, and is always known as a black box with low interpretability and hence may not be useful as a root cause analysis or for problems that seek deep understanding of variables that impact the prediction target. It normally works better with large dataset (both rows and tuples) and unfortunately it has always not been the case in real life dataset, especially for small competitors in the market or unpopular research questions (IMF, 2019). Besides, it also requires high computational costs in term of machine computing power, time etc.

I. Regulatory control

To-date, in most countries, there is no clear guidance on the use of AI and ML in the financial industry. However According to (Monetary Authority of Singapore, 2022), there are some principles to promote fairness, ethics, accountability, and transparency in decision making data analytics. It aims to ensure:

- there is already an existing customer segmentation (for example: MASS, High Net Worth, Premier) in the bank, using unsupervised technique like clustering analysis may enlarge the scale, hence there must be justifiability where no group of customers are disadvantaged in the analytics.
- secondly, the banks are to align with the social ethical standards and adherence these values in the use of AI and ML.
- thirdly, it also stressed on the transparency of data used to both internal and external parties' awareness where clear explanations are to be made when the data subjects are in doubts.
- besides, the data and model are to be reviewed and validated regularly to avoid any potential bias so that it is in line with its intended objectives. FEAT also urged the banks to develop its internal framework and government team to ensure the appropriate approvals are obtained in all its AI/ML initiatives.

The proposals are meant to make sure while AI and ML are fulfilling its huge potential in providing some great solutions to the banks but at the same time, they should also be strengthening its social obligation.

III. CONCLUSION

A practical model in predicting defaults does not only rely on its accuracy, it asl seek risk and business insights throughout the model's lifecycle – from the beginning stage of problem understanding, data collection, data preparation, development, validation, model selection, approval, deployment and up to monitoring. The process cycle repeats when new data coming in and model tested with out of time sample, it is an ever-improving activities that progress along with the business environment. Besides, regulatory governance is an important role to guide the development process is not bias and trustworthy. As they say: Data is the new oil, refining it is what makes it valuable.

REFERENCES

Wong, L. Y. (2021). Predicting Subprime Customers' Probability of Default Using Transaction and Debt Data from NPLs.

Schreiber-Gregory, D., Bader, K. (2018). Logistic and Linear Regression Assumptions: Violation Recognition and Control

Orji, U. E., Ugwuishiwu, C. H., Nguemaleu, J. C., & Ugwuanyi, P. N. (2022, April). Machine Learning Models for Predicting Bank Loan Eligibility. In 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON) (pp. 1-5). IEEE.

Alejandrino, J. C., Jovito Jr, P., & Murcia, J. V. B. (2023). Supervised and unsupervised data mining approaches in loan default prediction. International Journal of Electrical and Computer Engineering (IJECE), 13(2), 1837-1847.

Ereiz, Z. (2019). Predicting Default Loans Using Machine Learning (OptiML). 1-4. 10.1109/TELFOR48224.2019.8971110.

Ozgur, O., Karagol, E. T., & Ozbugday, F. C. (2021). Machine learning approach to drivers of bank lending: evidence from an emerging economy. Financial Innovation, 7(1), 1-29.

Hoang, D., & Wiegratz, K. (2021). Machine Learning Methods in Finance: Recent Applications and Prospects. European Financial Management.

Chitambira, B. (2022). Credit Scoring using Machine Learning Approaches.

Tyagi, S. (2022). Analyzing Machine Learning Models for Credit Scoring with Explainable AI and

Optimizing Investment Decisions. arXiv preprint arXiv:2209.09362.

Bank Negara Malaysia (2019). Capital Adequacy Framework (Basel II – Risk-Weighted Assets)

Karalar, H., Kapucu, C., & Gürüler, H. (2021). Predicting students at risk of academic failure using ensemble model during pandemic in a distance learning system. International Journal of Educational Technology in Higher Education, 18(1), 1-18.

Aida, A, Shamsuddin, S., Ralescu, A. (2015). Classification with class imbalance problem: A review. 7. 176-204.

OECD (2021) Artificial Intelligence, Machine Learning and Big Data in Finance

Wijewardhana, Udani. (2018). A Mathematical Model for Predicting Debt Repayment: A Technical Note. Australasian Accounting, Business and Finance Journal. 12. 107-115. 10.14453/aabfj.v12i3.8.

Cheng, L. C., Wu, C. C., & Chen, C. Y. (2019). Behavior analysis of customer churn for a customer relationship system: an empirical case study. Journal of Global Information Management (JGIM), 27(1), 111-127.