Working paper on the improvement of corporate probability of default modeling exercises

Vidal Marvin C. Gabriel

De La Salle University, Philippines

Security Bank Corporation

INTRODUCTION

In 2018, the International Accounting Standard Board (IASB) replaced IAS-39, Financial Instruments: Recognition and Measurement with IFRS-9, Financial Instruments. This is due to the criticisms faced by IAS-39, to be backward-looking which caused for the delay of credit loss recognition during the financial crisis in 2008 (Barth and Landsman (2010)). The new IFRS-9 focuses on forward-looking assessment of expected credit loss (ECL). Since then, banks under the supervision of a central bank calculates ECL and assigns provisions on its suite of portfolios,i.e., corporate or wholesale, retail, auto loan, housing loan, to name a few. The accuracy of ECL is very critical on every bank as over estimation of ECL leads to lack of possible revenue while underestimation puts the bank at risk of liquidity shortage during bank runoff.

ECL can be calculated using the following equation:

$$ECL = PD * EAD * LGD \tag{1}$$

where *PD* is the probability of default, *EAD* is the exposure at default and *LGD* is the loss given default. Of the three components, *PD* is greatly affected by economic activity. This is due to the response of the corporate borrowers against a current or impending economic shock. According to Gavalas and Syriopoulos (2014), there is a direct impact on PD with the shifting of economic activity.

Of all the portfolios, wholesale, or wholesale banking segment (WBS) henceforth generates the highest possible income as well as the highest possible loss due to the amount being leased. Thus, for this study, the focus will be on the improvement of WBS portfolio.

The Philippine banking sector especially Security Bank Corporation (SBC), models WBS-PD through the following steps:

1. Scorecard PD – assigns initial PD on the accounts.
    - The WBS portfolio is divided into two categories, currently into big and small corporates using gross revenue as a filter [1].
    - Scorecard PD is represented by a logistic model [1]. At present, there is only one model for both big and small corporate. The models only account for financial performance, e.g. days-pass-due, gross balance, months on books.
2. The result of the scorecard PD model will be used to identify the risk rating of each account. Risk ratings are defined by each bank, usually it ranges from 1 to 10 where 1 is the least risky. There are also management overrides wherein the risk managers override the scorecard PD definition of risk ratings and rely on their expert judgments to assign risk ratings.
3. From the risk ratings, the accounts are divided into three categories:
    - Stage 1 – low risk ratings, maturing in the next 12 months (12-month PD). The result of the scorecard PD model is adjusted using the second model (discussed below) and is projected for 12 months.
    - Stage 2 – uses lifetime PD for the next 12 months. The result of the scorecard PD model is adjusted using the result of the second model (discussed below) and is projected for 15 years.
    - Stage 3 – Non-performing loans. These accounts are considered loss in the portfolio.
4. Macroeconomic model – uses macroeconomic covariates as the independent variables and the bad rate of the portfolio defined as,

$$\text{Bad rate} = \frac{\text{No. bad accts in the next 12 months}}{\text{Total no. of accts}} \tag{2}$$

[1]This may differ from bank-to-bank depending on their modeling techniques and definitions or risk.

as the target variable. The wide collection of macroeconomic variables is trimmed down to a parsimonious model which is usually a simple linear combination[1].

5. The initial PDs for Stage 1 and Stage 2 accounts are adjusted using an algorithm wherein the average of the WBS PD should be close to the result of the second model. Currently, SBC uses goal-seek to adjust the scorecard PD. Additionally, ARIMA (auto-ARIMA)[1] is being used to forecast 12-month ahead and 15-month ahead for the independent variables to come up with a forward-looking PD.

1 PROBLEM WITH THE CURRENT PROCESS

The use of logistic regression is quite restrictive because of symmetric assumption. In other words, the absolute magnitude of the positive and negative impacts of the change in covariates on PD is the same. This may lead to biased estimates on scorecard PDs as the response may be asymmetric, i.e., PDs might go up faster during stress conditions, and goes down slower during growth conditions.

Another problem is the use of aggregated bad rate as the dependent variable for the macroeconomic model and the neglect of industry specific effects on the model. Bruneau et al. (2024) have found that disaggregated PDs using industry specific models improve forecasts. This is obvious as one macroeconomic model will not account for the behaviors of the different industries, e.g. mining, education, construction, retail. Further, an exploration of a state-of-the-art forecasting model may be more beneficial than just using GLM. Guth (2022) made a study on the comparison of different modeling techniques for forecasting credit loss.

These problems contribute to the volatility of final PD estimates. Risk managers prefer a stable PD estimate as this may disrupt provisions and liquidity allocations on a quarterly basis. This may also cause the bank to lose profit and signals failure to manage credit risks in the portfolio.

2 RESEARCH QUESTION AND OBJECTIVE OF THIS STUDY

To address the current problems, this study tries to answer the following questions:

1. How can we improve the current forecasting process?
    - Is there a significant impact on the PD estimate if we consider asymmetry?
    - Is there a significant impact on the macroeconomic model if we consider industry-specific over aggregated PD?
    - Can non-parametric models improve the macroeconomic models?

From the questions above, the following objectives are formulated:

1. Subdivide the portfolio into industry instead of gross revenue.
2. Develop asymmetric models for each category.
3. Explore other possible algorithms to develop macroeconomic models for each category.

3 DATA SOURCES

Datasets can be gathered from the following:

1. Security Bank Corporation – for the corporate portfolio and its financial performance on the account level.
    - Historical optimist data
    - Historical bad rates of WBS portfolio.
    - Other NPL data
2. Bangko Sentral ng Pilipinas and Philippine Statistics Authority for the macroeconomic covariates.

4 LITERATURE REVIEW

4.1 Resampling method to model scorecard PD.

One of the challenges in developing scorecard PD model is the bias of non-default to default. This is due to the limited number of defaulting accounts as compared to the number of healthy accounts. Similar problem has been encountered by Zhu et al. (2019) in which they made used of resampling technique called synthetic oversampling technique (SMOTE). This technique uses the following procedure:

1. From each x_i in the minority sample set, a nearest neighbor is adopted to calculate the K nearest neighbor.
2. Then, for each sample x_i, several samples are randomly selected from its k-nearest neighbors.
3. Lastly, with the assumption that the selected neighbor is x_n. The randomly selected neighbor x_n produces new samples using the following equation,

$$x_{new} = x_i + rand(0,1) * |x - x_n| \tag{3}$$

Just like Zhu et al. (2019), it is acknowledged that for this study, similar problem will be encountered since the number of good accounts outnumbered bad accounts significantly. Thus, the use of SMOTE is beneficial for this study.

4.2 ROC-AUC to assess scorecard PD model.

The output of a scorecard PD model is between 0% and 100% and the actual values are in a binary mode (good = 0, bad = 1). Thus, the assessment of the model's performance is not as straight forward as a continuous data result. To solve this problem, this study can follow previous studies (e.g. Tang and Chi (2005); Zhu et al. (2019); Alonso and Carbó (2021)) wherein they used receiver operating characteristics (ROC) and area under the curve (AUC).

To perform an ROC-AUC analysis, confusion matrix will be established first. This matrix consists of actual and predicted classes. It is a two-dimensional matrix wherein the columns represent the actual classification while the rows represent the predicted classification. Table 1 shows a sample of confusion matrix.

	Actual bad (1)	Actual Good (0)
Predicted bad (1)	True positive	False positive
Predicted good (0)	False Negative	True Negative

Table 1: Confusion matrix.

On the one hand, true positive (TP) is defined as the correct positive predictions while false positive (FP) is the failed predictions of bad wherein the model predicted a default, but the actual value is good. On the other hand, true negative (TN) is defined as the correct negative predictions while false negative (FN) is the failed predictions of good wherein the model predicted good, but the actual value is bad.

The challenge here is the definition of bad and good. Specifically, at what threshold of PD is considered bad? ROC-AUC can be used to gain insight from a variety of threshold values. From the confusion matrix, the following equations can be established,

$$\text{True positive rate (TPR)} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{False positive rate (FPR)} = \frac{FP}{FP + TN} \tag{5}$$

For each threshold values, a corresponding TPR and FPR can be generated. Plotting FPR in the x-axis while TPR in the y-axis will give the ROC curve wherein the diagonal line in the center from origin to point (1,1) is the called the random line. Threshold values that fall below this line should be discarded immediately as they imply bad performance while threshold values close to point (1,0) can be considered a good performing threshold.

Additionally, threshold values can be studied based on how conservative or liberal they are in predicting bad. That is, if threshold is close to (1,0), it is expected to be conservative in the sense that, it focuses on predicting correct bad accounts while threshold away from (1,0) allows for certain errors on their predictions.

One might be tempted to aim for (1,0) threshold. But, for credit risk perspective, this is a bad choice since if we allow the model to be so conservative, the effect of this is high scorecard PD on the portfolio. Thus, it is best to have a trade-off of FPR to have a more robust distribution of scorecard PD.

Lastly, to compare competing models, AUC can be utilized. This is defined as the area under the curve of the series of thresholds of the model. A model with higher AUC is preferred to lower AUC.

4.3 Using asymmetric approach over symmetric to model macroeconomic PD.

Bruneau et al. (2024) presented a new credit risk model for corporate loan portfolio centered in Canadian banking sector. Their model maps relevant macroeconomic stress factors into aggregate and industry- specific probability of defaults (PD). Also, the model they developed is an improvement of the previous model by MacDonald and Traclet (2018) called Framework for Risk Identification and Assessment (FRIDA).

The incorporated improvements are:

1. Novel filter for extracting PD
2. Modelling of aggregated PD before going into industry decomposition anchored around the aggregate probability.
3. Expansion of the model's set of potential macro financial stress factors (eg. Risk factors associated with the foreign exposure of Canadian banks.
4. Shows how quantile regression is best at capturing non-linearities which might exist in adverse macro financial risk scenarios.

The last three items mentioned are beneficial to this study. This is because, most of the time the process to model PD is the use of linear symmetric models. As an effect, the magnitude of increase is equal to the magnitude of decrease in PD. In other words, the positive and negative impacts of the explanatory variables are assumed to be equal. A non-linear model accounts for asymmetric and non-linear relationships of the explanatory variables on PD. Further, the portfolio is modelled in aggregate and thus failed to incorporate the behavior of each account with respect to their industry. For number 1, this may not be beneficial in this study as the banks in the Philippines derive PDs based on Basel II and IFRS9 which is being regulated annually. Thus, this study will stick on how PDs are currently extracted.

In their exploration of explanatory variables, they have used 25 potential explanatory variables divided into subcategories. They also acknowledge the following problems arising from these variables:

1. Limited observations
2. The number of potential variables is large.
3. Possibility of lag orders.
4. Risk of over-fitting in the in-sample which may cause poor out-sample performance.

To overcome the above problems, they have used elastic net regularization as a model selection. In this study however, we can opt to use other model selection techniques instead of sticking with elastic net regularization.

To select the linear model, they have chosen one standard deviation MSE with up to one lag of macro financial explanatory variables. They have observed that PDs are best explained by:

1. $AR-1$ process
2. Contemporaneous unemployment rate, the real business credit, the policy rate differential between Canada and the United States, and the Canadian term premium.
3. First lagged values of the real business credit, the CAD/USD exchange rate, the five-year GoC bond yield, and the US financial stress index.

Apart from the specifications of the model, they also made overrides, i.e., (1) they did not penalize GDP and US industrial production, (2) they have also included their lagged values, (3) they overrule the elastic net by including a second lag of PD to account for serial correlation in the residuals. They have settled with an $AR-2$ process with an addition of additional predictors which takes the form:

$$PD_t^{\lambda} = \alpha + \sum_{i=1}^{2} \beta_i PD_{t-i}^{\lambda} + \gamma X_t + \epsilon_t \tag{6}$$

They used rolling 12-quarter-ahead root-mean squared error (RMSE) over the period of $t = 10$ through the end of the sample which showed favorable results.

For the non-linear model, they have extended Equation 1 to the following quantile regression (QR):

$$Q(\beta_q) = \sum_{t:PD_t^{\lambda} \geq \widehat{PD_t^{\lambda}}} q \left| PD_t^{\lambda} - PD_t^{\lambda} \right| + \sum_{t:PD_t^{\lambda} < \widehat{PD_t^{\lambda}}} (1-q) \left| PD_t^{\lambda} - PD_t^{\lambda} \right| \tag{7}$$

where $\widehat{PD_t^{\lambda}} = \alpha_q + \sum_{i=1}^{2} \beta_i PD_{t-i}^{\lambda} + \gamma_q X_t + \epsilon_t$ is the fitted value for quantile q, which takes values from 0.1 to 0.9. They have estimated Equation 7 using the majorize-minimize method of Hunter and Lange (2000).

Using QR, they have found the following:

1. AR process is more persistent in the upper tail than in the lower tail. This implies that the increase in PDs are faster which is persistent in times of crises as compared to it's decrease during economic growth.
2. Unemployment rate and US industrial production have an impact that is two to six times greater on PDs in the upper quantiles than in the lower quantiles.
3. Effects of real GDP and real business credit seem muted in the upper tails.
4. Thus, during stress testing wherein PDs are subjected to shocks which pushes PDs to upper quantiles, QR is preferred over a linear model.

They then proceeded to develop industry specific PD models. However, they limit the possible explanatory variables on the chosen variables from the aggregated PD model with an addition to a few variables that may have an industry specific impact. Also, they only utilized a linear combination with an addition of the lag of the aggregate PD from the QR regression. Equation 8 shows the form of the industry specific model.

$$PD_t^{\lambda,ind} = \alpha + \sum_{i=1}^{2} \beta_i^{ind} PD_{t-i}^{\lambda,ind} + \sum_{i=1}^{2} \beta_i PD_{t-i}^{\lambda} + \gamma X_t + \epsilon_t \tag{8}$$

Using Equation 8, they have found the following:

1. Persistence of industry specific PDs is heterogeneous, with the majority being lower than the persistence of the aggregate PD process.
2. For most industries, an improvement in economic conditions reduces PDs, an increase in borrowing costs increases PDs, and a credit tightening through a decrease in real business credit raises PDs.
3. Some of those patterns differ, especially for the mining, quarrying, and oil industries.
4. Higher domestic borrowing costs are associated with lower PDs.

They have compared the industry specific model with the previous aggregate model (Equation 6) using the same method. They have found the following:

1. Industry specific ARX model improves upon $AR\left(2\right) + PD^{\lambda}$ model for all industries.
2. $AR\left(2\right) + PD^{\lambda}$ model improves upon the simple $AR\left(2\right)$ models.
3. For financial institutions industry, difference is less significant.

They also noted that, there are no constraints in the estimates which will ensure industry specific PD projections are consistent with aggregate PD projection for the same stress-test scenario. Thus, to address this, they made use of Equation 9 below.

$$PD_{T+h}^{\lambda,wa} = \sum_{ind} PD_{T+h}^{\lambda,ind} \frac{TL_T^{ind}}{TL_T} \tag{9}$$

where h is the quarter ahead in the projection and the weights are the share of loans of each industry over total loans for the most recently observed quarter, which are then assumed to be invariant for the period of the projection.

They then compare the aggregate PDs with the weighted average of projected industry specific PDs.

$$PDR^{\lambda}_{T+h} = \frac{PD^{\lambda,wa}_{T+h}}{PD^{\lambda}_{T+h}} \tag{10}$$

For each period, if the weighted average of industry specific PDs is below or above the aggregate PDs, then they respectively scale up or down the industry specific PDs by PDR^{λ}_{T+h}. This adjustment ensures that the aggregation of industry specific PDs matches the aggregate PD.

4.4 Modeling PD using naïve approach, from parametric to non-parametric.

The use of asymmetric approach can be beneficial in improving the bank's PD estimates as presented above. For this subsection, it will discuss the recent exploration of different parametric and non-parametric approaches to model PD. There have been numerous studies that cover the forecasting of PD (e.g. Aver et al. (2008); Bofondi and Ropele (2011); Gambera et al. (2000)). All of which uses a linear combination modeling technique. In recent years however, there have been attempts to use machine learning algorithms to better forecast PD (e.g. Guth (2022); Barbaglia et al. (2023); Bellotti et al. (2021); Alonso and Carbó (2021)). They have found that machine learning algorithms can be good candidates for PD projections.

Guth (2022), studied a naïve approach in modeling PD for forecasting and stress testing exercises. Their dataset consists of PD from 2002Q2 to 2019Q2 while the macroeconomic variables are from different categories, i.e., economic growth, employment, price indices, exchange rates, interest rates. Further, they explored a total of 43 models within 9 overarching categories.

For their initial analysis, they performed a logit transformation on PD to make sure it is between 0% and 100%. As for the covariates, they analyze the trend and cyclical components and performed a series of unit root tests, i.e., Dickey-Fuller (ADF), Elhott, Rottenberg & Stock (ERS), Philipps-Perron (PP), and Kwiatkowski-Phillips-Schmidt-Shin (KPSS). They found that without any treatments, the variables suffer from unit root and thus proceeded with first differencing.

All their 43 models follow the following structure,

$$y = f(X_t, X_{t-1}, \dots, X_{t-p}) + \epsilon \tag{11}$$

where $y = (y_1, \dots,\ y_T)\prime$, the dependent variable is a function of contemporaneous and lagged predictions of the independent variable $X = (x_1, \dots,\ x_T)\prime$. In their paper, they have used up to P=4 lags. Also, the 43 models are categorized in the following:

1. Generalized linear model (GLM) – contrary to the previous study by Bruneau et al. (2024), AR was not used in their candidate models.
2. Model averaging
3. Exponential smoothing
4. Generalized additive models
5. Multivariate adaptive regression splines
6. Support vector machines
7. Gaussian process
8. Tree-based models
9. Neural networks

The paper then proceeded with the search for hyperparameters for their 43 models. The author used grid search technique. However, the issue with this technique is the possibility of large number of combinations that must be evaluated. To solve this problem, the author combined grid search with expert judgement. Cross-validation strategy was introduced to generalize the parameters. The paper applied a rolling-origin evaluation, starting with an initial training set of $T_{t,1} = 1, \dots,\ 52$ and a fixed holdout set of $T_h = 12$ which represents 12 quarters to be forecasted. The author also recognized the data set is small and some of the models (particularly non-parametric) may suffer from over-fitting. To assess for the performance of each model, mean absolute error (MAE) was chosen as a metric.

After performing necessary treatments on each model, the paper found instabilities on the model, it is being dropped due to distorting outputs.

Therefore, 21 models have been dropped leaving the collection of models to 22, from eight categories. The paper has introduced an over-fitting threshold which made the remaining threshold improve and stabilize forecast. It has also been noted that models struggle at first until a certain length of the training set is reached. This confirms the expectation that there is a large variation in results across the models. Thus, a proper comparison is needed to gain more insight into the driving factors. *[One issue I come up here is the identification of over-fitting threshold. This was not extensively discussed in the paper.]*

From the comparison, the paper noted that the overall best performance stems from **tree-based models**. Specifically, **classical random forests** and **Bayesian additive regression trees (BART)**. The paper further analyzed the results of the model using regression for multiple comparison with the Best (RMCB) which revealed that the winning model is BART. **However, the paper noted that there seven mode models that are not statistically different from BART, i.e., spike-and-slab prior random forest, independent component regression, lasso, principal component regression, neural network, support/relevance vector machine.**

For the interest of this study, the eight stated models from Guth (2022) will be used as candidate models for the macroeconomic model (second model).

5 POSSIBLE CHALLENGES

1. Extensive WBS-ECL exercises only started in 2018 and this is done quarterly. Thus, the number of data points might be a challenge and we might face the problem of over-fitting.

2. Further sub-categorizing the dataset into industry specific might worsen number 1 and the distribution might focus on one to two industries only. Thus, if such cases occur during the study, sub-categorizing using other modes (e.g., gross revenue, gross balance, months on books) might be more beneficial.

Reference

Alonso, A. and Carbó, J. M. (2021). Understanding the performance of machine learning models to predict credit default: a novel approach for supervisory evaluation.

Aver, B. et al. (2008). An empirical analysis of credit risk factors of the slovenian banking system. *Managing Global Transitions*, 6(3):317–334.

Barbaglia, L., Manzan, S., and Tosetti, E. (2023). Forecasting loan default in europe with machine learning. *Journal of Financial Econometrics*, 21(2):569–596.

Barth, M. E. and Landsman, W. R. (2010). How did financial reporting contribute to the financial crisis? *European accounting review*, 19(3):399–423.

Bellotti, A., Brigo, D., Gambetti, P., and Vrins, F. (2021). Forecasting recovery rates on non-performing loans with machine learning. *International Journal of Forecasting*, 37(1):428–444.

Bofondi, M. and Ropele, T. (2011). Macroeconomic determinants of bad loans: evidence from italian banks. *Bank of Italy Occasional Paper*, (89).

Bruneau, G., Ferreiro, J. O., Plummer, A., Tremblay, M.-C., and Witts, A. (2024). Gabriel bruneau-latest. *Methods*.

Gambera, M. et al. (2000). *Simple forecasts of bank loan quality in the business cycle*, volume 230. Federal Reserve Bank of Chicago Chicago, IL.

Gavalas, D. and Syriopoulos, T. (2014). Bank credit risk management and rating migration analysis on the business cycle. *International Journal of Financial Studies*, 2(1):122–143.

Guth, M. (2022). Predicting default probabilities for stress tests: A comparison of models. *arXiv preprint arXiv:2202.03110*.

Hunter, D. R. and Lange, K. (2000). Quantile regression via an mm algorithm. *Journal of Computational and Graphical Statistics*, 9(1):60–77.

MacDonald, C. and Traclet, V. (2018). The framework for risk identification and assessment. Technical report, Bank of Canada.

Tang, T.-C. and Chi, L.-C. (2005). Predicting multilateral trade credit risks: comparisons of logit and fuzzy logic models using roc curve analysis. *Expert Systems with Applications*, 28(3):547–556.

Zhu, L., Qiu, D., Ergu, D., Ying, C., and Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science*, 162:503–513.