Quantifying Model Selection Risk

Joseph L Breeden

Deep Future Analytics LLC

1600 Lena St., Suite E3,

Santa Fe, NM 87505 USA

breeden@deepfutureanalytics.com

Nikolay Dobrinov

Deep Future Analytics LLC

1600 Lena St., Suite E3,

Santa Fe, NM 87505 USA

dobrinov@deepfutureanalytics.com

July 30, 2021

Abstract

Model selection is a critical step in model development and brings with it a significant level of model risk. This analysis considers a real-world example comparing the forecasts and uncertainties of three possibilities: 1. the model selected as best, 2. the best ensemble, and 3. the model not selected. A key part of this analysis is in propagating the prediction intervals for the individual forecasts using Monte Carlo simulation all the way through the lifetime probability of default calculations to obtain distributions for the lifetime forecasts. The results highlight two sources of risk: a systematic difference between the expectation value of the forecast and the median of the simulated distribution, and the 95% confidence interval from the distribution. The difference between the forecast and simulated median should be included as part of the loss reserve. The confidence interval from the distribution should contribute to the model risk component of the economic capital. This work also serves to illustrate how uncertainties can be computed for Current Expected Credit Loss (CECL) or IFRS 9 Stage 2 lifetime loss estimates.

Keywords: Model Risk Management, Ensemble Models, Prediction Intervals, Error Propagation

1 Introduction

Model risk management [7, 2, 10, 1, 12] has become one of the primary issues in model development and deployment. Identifying model risks is a broad and multifaceted problem. This is seen most recently in the model risks exposed

1

2

during the COVID-19 crises globally [5] and in the desired utilization of machine learning methods for credit risk [11, 6].

Among the many model risk components is model selection risk. Most lenders have short time histories relative to economic cycles. With only one or at most two economic cycles in the history of most lenders, many macroe conomic models can fit equally well in-sample but diverge significantly out of sample. This creates a model selection risk that is not as severe in credit scoring and many other lending applications.

The point-estimate of loss reserves varies with the methodology used to model losses. However, even within a specific modeling technique, the set of possible model specifications to choose from can be large. Different model spec ifications produce different paths for portfolio losses, while loss reserves are typ ically set based on a single selected model specification or a narrow ensemble of model specifications. The process of model specification selection introduces risk to the loss reserves, which maybe set too low or too high with respect to the required reserves. Such is the case when correlating changes in the macroe conomic environment to the performance of a portfolio of loans.

Changes in the macroeconomic environment are one of the main predictors of credit risk of interest to lenders and regulators. The macroeconomic factors that shape the behavior of borrowers are numerous. However, many lenders are forced to develop parsimonious models of default behavior due to sparse default event rates in their portfolios. Model selection criteria are then applied to select a narrower set of model specifications under which to derive the point-estimate of reserves. This narrower set of model specifications represents a narrower set of macroeconomic predictors employed to drive the loss forecast. This process limits the possible and plausible paths of default behavior and imposes a model selection risk on the point-estimate of the loss reserves.

Using real examples from the portfolio of a US lender with about $20 billion in assets, this article investigates the magnitude of model selection risk in credit risk modeling and its impact on loss reserves. The following sections explain a simple modeling approach used for this portfolio, the model selection process, how the models are used to create Current Expected Credit Loss (CECL) [3, 9], which is comparable to IFRS 9 Stage 2 lifetime loss estimates [8], and then explore various measures to quantify the model selection risk by looking at what could have come from the models not chosen.

2 Data

The data used for this study represents portfolios of Commercial and Industrial (C&I) Construction loans, Mortgage loans, and Non-Owner Occupied (NOO) loans. The portfolios are relatively small in terms of counts with the larger Commercial portfolios represented by on average 1500 healthy accounts and 3-6 default accounts per quarter, and the larger Consumer portfolios represented by on average 4000 healthy accounts and 18 default accounts per quarter. The historic data available for these portfolios begins 2006-Q4, and the models are

3

calibrated on data for the historic period of 2006-Q4 to 2019-Q4. Although data was available for 2020 during the COVID-19 pandemic, that data was excluded from the modeling due to the dramatic government interven tion. Consumers and businesses received direct assistance and loan forbearance. As of 2021, the ultimate consequences of the recession in terms of default are not yet known, so we cannot connect the pandemic causes to a default effect.

3 Model Development

The sparse data in these portfolios does not allow for the estimation of complex models. For that reason, the models developed for this study are macroeconomic time-series probability of default models [4]. Due to multiple quarters with zero defaults in the portfolio, a binomial fit with a logistic link function is used to fit the models. The macroeconomic indices are log or logit-transformed depending on their defined range. We do not allow nonstationary transformations for macroeconomic predictors that in the long term move in one direction, such as GDP, HPI, and disposable income. Such macroeconomic indicators are tested for predictive power only transforming to a rate-of-change. A moving average transform is only allowed for predictors that tend to have a cyclical and bounded time series, such as unemployment rate.

Also of interest is the lag in the impact of the macro-economic shocks on the portfolio as well as how prolonged the impact is. For that purpose, macroe conomic indices are also tested for predictive power with a lag of maximum of three quarters and the span over which a change or a moving average transform is calculated of maximum eight quarters. The latter transform is needed to cap ture the varying delay in responses to the macroeconomic shock by borrowers with different risk characteristics.

With this framework of creating logistic regression models using lagged transformations of macroeconomic factors, the lags and transformation win dows need to be optimized, as well as the selection of which macroeconomic factors are to be used. All possible lags and windows of all combinations of macroeconomic factors is far too large a search space, so this is done in stages. First, each macroeconomic factor’s lag and window are optimized in a univariate logistic regression. Then all combinations of optimized univariate models are tested to find the best multivariate model. This approach is not theoretically optimal, but it allows for the imposition of necessary constraints during the model selection process.

4 Model Selection

The candidate models, generally hundreds per product, were filtered according to the following criteria:

1. All coefficients must have Newey-West robust p-values *<*= 0*.*05.

4

2. All coefficients must have signs that align with business intuition. 3. The maximum variance inflation factor for any factor must be less than 2.

4. Out of sample tests under base and severe macroeconomic scenarios must show higher losses for the severe scenario.

Once all of these filters have been applied, the remaining models were ranked by AIC. For use by a business, a last step involves manual review by the man agement team to debate the features and relative merits of each model. An ensemble of two to four preferred models was chosen. For testing here, we con sidered the best model by AIC, the ensemble chose by the business, and the next 30 models not chosen.

5 Model Comparison

In a typical validation process, the models are tested against an out-of-time sample. However, the available data only spanned a period from 2008 through 2019. That period contains exactly one recession. Since the goal here is to model loan response to economic cycles, one must admit that this is an inherently ill specified problem. The statistically correct procedure would be to have at least several recessions in the training set and at least one recession (that is not an anomalous pandemic) in the test set. The practical reality is that no lenders possess such data, and yet models must be built to guide management on loss risks and for regulatory compliance.

The consequence of having a single recession for training is that the models may fit well in-sample but are poorly constrained to future events. In other words, many models may look equally good in-sample and yet give divergent answers out-of-sample. As will be seen below, this specific data set and these specific candidate models are not the worst example of the poorly fit nature of all stress testing models, but they are sufficient to explore model risk management regarding model selection.

Figures 1-3 show the historic data and time series forecasts for the best model according to AIC, the ensemble created via management selection, and the range of forecasts produced by the next 30 models that were not selected. All forecasts are shown as log-odds of default. Note that the difference between Best.AIC and Ensemble forecasts is between 0.3 and 0.4 in relative log-odds through the first several years under a baseline economic scenario. That difference translates to a roughly 60% spread between the forecasts in quarterly default rate. The Next.30 models span a range from -6.7 to -5.5 in log-odds, a roughly 75% spread between best and worst outcomes. Ironically, the spread between models is much more muted with the severe economic scenario.

The first test considers how much the prediction interval for the models captures uncertainty of the forecasts as compared to the variation from model selection. Prediction interval is a combination of the confidence interval of the forecast with a specific input vector plus the error due to model fit. Although

5

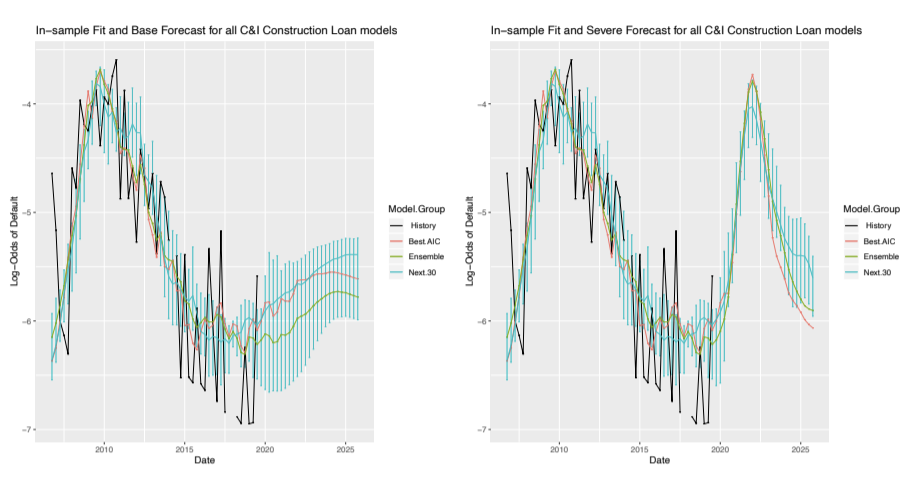


Figure 1: Plots for C&I Construction of the historic charge-off rate, best model, ensemble of best models, and range of the next 30 models with forecasts using a baseline or severe economic scenario.

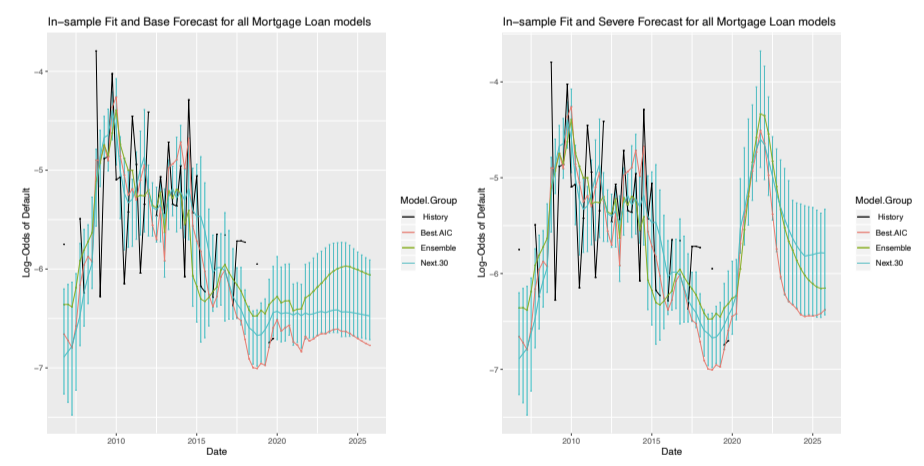
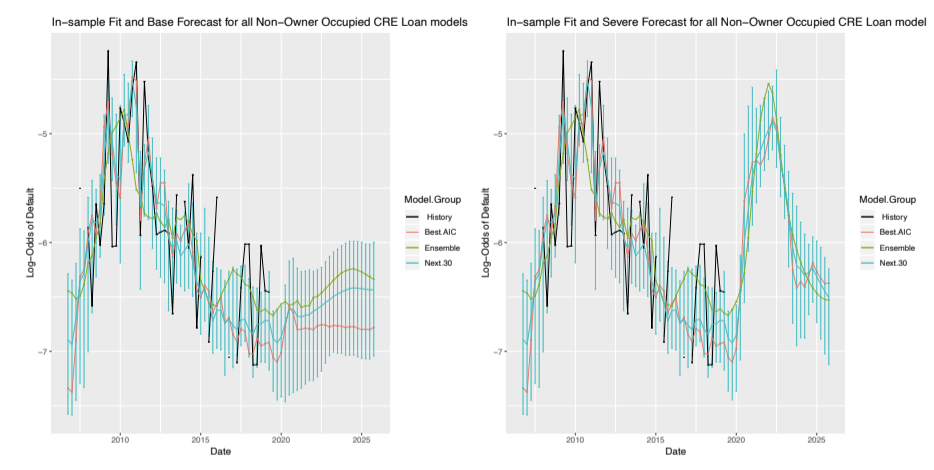


Figure 2: Plots for Mortgage of the historic charge-off rate, best model, ensemble of best models, and range of the next 30 models with forecasts using a baseline or severe economic scenario.

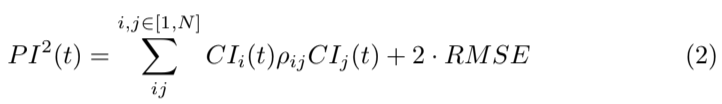
6

Figure 3: Plots for Non-owner Occupied CRE of the historic charge-off rate, best model, ensemble of best models, and range of the next 30 models with forecasts using a baseline or severe economic scenario. 

straight-forward in linear regression models, error is not well defined for logistic regression models. As a rough approximation, the root-mean-square-error of a given model was computed in log-odds of default. The two standard deviation confidence interval *CI*(*t*) was combined with the RMSE to obtain the prediction interval *P I*(*t*).

*P I*2(*t*) = *CI*2(*t*) + (2 *· RMSE*)2 (1)

For the ensemble model, the RMSE was computed after averaging the fore casts of the component models. The confidence interval of the ensemble was computed by combining confidence intervals of the component models with con sideration for the correlation of the residuals between models. With *ρ* as the correlation matrix between the model residuals  *i*(*t*) for model *i*, the prediction interval is



where N is the number of models. Note that *ρ* is computed on the model resid uals, because the goal is to measure the correlation between the uncertainty distributions. Each time *t* can be seen as a sample from those distributions. We do not measure *ρ* on the forecast time series directly, because that would simply measure the correlation between the mean forecasts, essentially the macroeco nomic correlation.

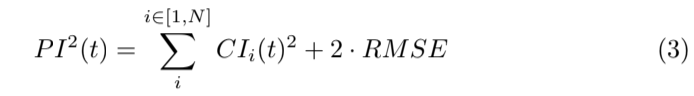
In the examples here, two models were selected for each ensemble, so *ρ* = 0*.*92*,* 0*.*74*,* 0*.*70 for C&I Construction, Mortgage, and Non-Owner Occupied CRE, respectively.

The next 30 models were never used as an ensemble forecast, because some

7

were clearly less desirable. Rather, we want to look again at the implied range of possible outcomes. Figures 1-3 showed the range of forecasts, but we also want to look at the range including their confidence intervals. That was also computed according to Equation 4 and is shown in Figures 4-6 as Next.30.pi.

Applying Equation 4 to a large number of models creates a prediction interval that grows at least as *√N*. This is easily seen if the models were uncorrelated. In that case, Equation 4 simplifies to



If all the models had identical confidence intervals,

*P I*2(*t*) = *N · CI*(*t*)2 + 2 *· RMSE* (4)

Therefore, a more stable representation of the uncertainty from the models not selected might be to use *CI*(*t*) = 2*·σ*(*t*), where *σ*(*t*) is the standard deviation of the forecasts of all models computed at each time step *t*. That result is shown in Figures 4-6 as Next.30.sd.

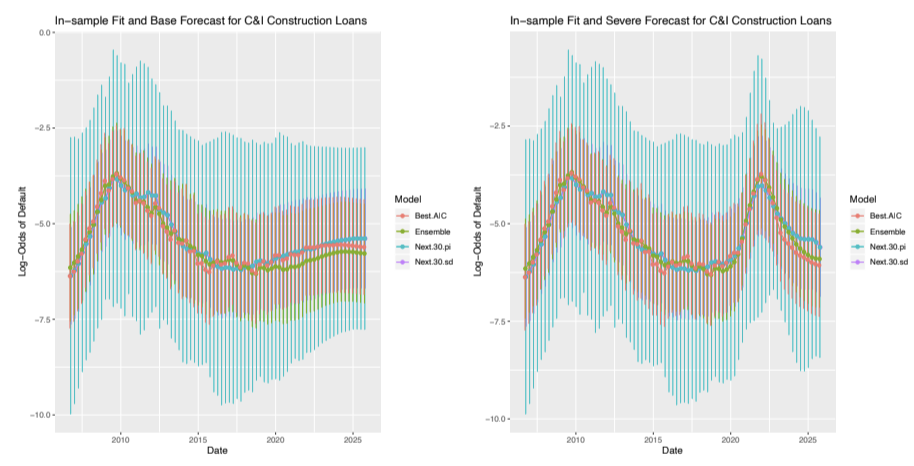
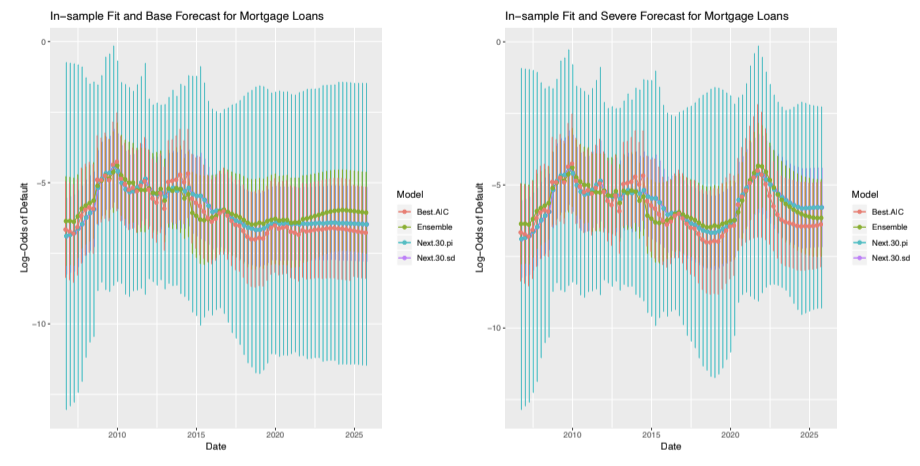


Figure 4: Plots for C&I Construction of the in-sample fits and forecasts of the best model and ensemble model showing two standard deviation prediction intervals. ”Next 30 pi” refers to combining the confidence intervals of the next 30 models. ”Next 30 sd” shows twice the standard deviation of the forecasts of those models.

At first the prediction interval results in Figures 4-6 seem remarkably sim ilar across the tests. This occurs because for small *N*, the prediction interval is dominated by RMSE. For large *N* in the Next 30, the standard deviation happens to give a similar answer. This result is shown for all loan types.

The final step in the process was to compute the lifetime loss forecast for CECL. A constant quarterly attrition rate was used for each product as the

8

Figure 5: Plots for Mortgage of the in-sample fits and forecasts of the best model and ensemble model showing two standard deviation prediction intervals. ”Next 30 pi” refers to combining the confidence intervals of the next 30 models. ”Next 30 sd” shows twice the standard deviation of the forecasts of those models.

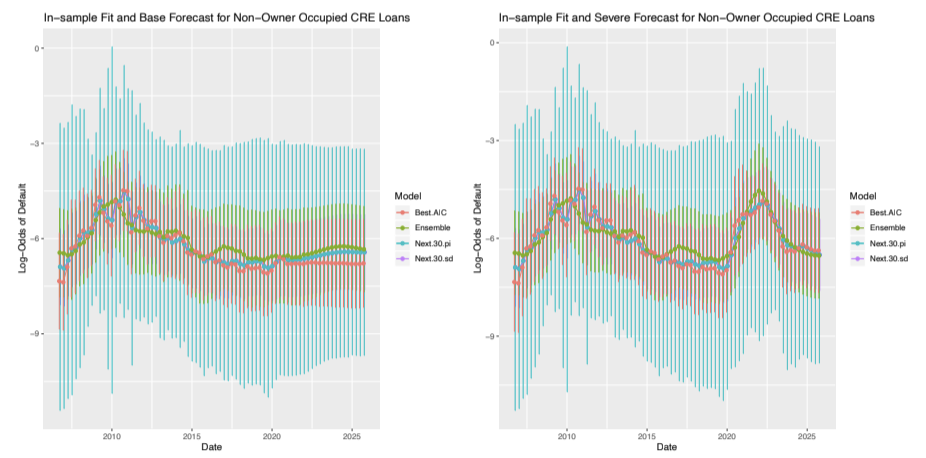
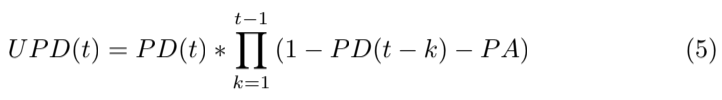


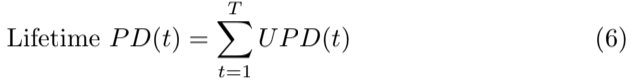
Figure 6: Plots for Non-owner Occupied CRE of the in-sample fits and fore casts of the best model and ensemble model showing two standard deviation prediction intervals. ”Next 30 pi” refers to combining the confidence intervals of the next 30 models. ”Next 30 sd” shows twice the standard deviation of the forecasts of those models.

9

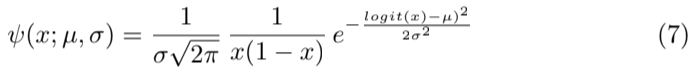
average rate in the history. Because the default probability modeled above was conditional on active accounts in the previous month, this competing risks approach leads to the following formula for quarterly unconditional probability of default.



The lifetime CECL loss rate comes from summing the *UP D*(*t*),



Model uncertainty is relatively simple to assess in the space of log-odds of default, because the model errors fit a normal distribution well. That means we have a logit-normal distribution when the forecasts and confidence intervals are converted to probabilities. If the mean and deviation are *µ* and *σ* in log-odds, then the density function for *P D*(*t*) is given by.



Having a closed form representation for the distribution *ψ*(*x*) for the values of *P D*(*t*) is nice, but not particularly helpful with Equations 5 and 6 because no closed form solution exists for multiplying logit-normal distributions.

To incorporate the prediction intervals in *UP D*(*t*) and Lifetime *P D*, a Monte Carlo simulation is the most obvious approach. At each time step *t*,

*logit*(*P D*(*t*)) = *N* (*µ* = *M*(*t*)*, σ* = *P I*(*t*)*/*2)

where *M*(*t*) is the forecast from one of our candidate models or ensembles, and *P I*(*t*) is the two standard deviation prediction interval.

For this test, 10,000 samples were taken at each time step *t* and fed through the equations to get a distribution for Lifetime PD. For the current study we focused only on default probabilities without simulating balance pay-down, so this is not the final answer for a CECL calculation. The results shown for Next 30 were created by simulating each model separately before computing the percentiles.

Table 5 show the 95% confidence range from the tabulated distributions for estimating Lifetime PD. The calculations were done for all three loan products. The lifetime *T* was 3, 5, and 6 years for C&I Construction, Mortgage, and Non-Owner Occupied CRE, respectively. For Next 30, the prediction interval used was the twice the standard deviation of the model forecasts plus twice the composite RMSE.

A review of Table 5 shows that the original forecast is systematically below the 50th percentile of the distribution generated via Monte Carlo. Forecasts C&I Construction, Mortgage, and Non-Owner Occupied CRE were low by 16.1%,

10

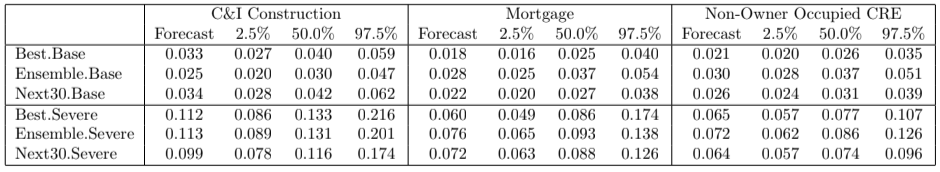


Table 1: Lifetime PD results for Base and Severe macroeconomic scenarios showing the 95% confidence intervals.

22.3%, and 16.3% respectively. For a single logit-normal distribution for the forecast of *P D*(*t*), the median of the distribution for *P D*(*t*) is just the inverse logit of *µ*, the mean in logit space. Therefore, a single forecast should not show a shift from the median with the transform.

However, when combining multiple forecasts of *P D*(*t*) to get to the lifetime estimate via Equations 5 and 6, at least in this example, the net forecast no longer matches the median of the distribution as obtained numerically via Monte Carlo. The final distribution is skewed toward higher loss rates, apparently as a natural consequence of the loan loss rates being close to zero. If we were modeling rates that were close to 1.0, the final distribution would likely be skewed toward lower values.

The second observation comes from comparing the median values for Best, Ensemble, and Next 30. Here we note that these values do not show a systematic pattern. The range of 50th percentile values across the models, if computed as (max(*x*) *−* min(*x*))*/*mean(*x*), is 30%, 39%, and 36% for the Base scenarios and 13%, 8%, and 15% for the Severe scenarios.

Therefore, we see two clear sources of model selection risk. Using the fore casts without incorporating the prediction intervals through the calculations all the way to the final answer in this case created a systematic under-prediction. This might not always be the case for all problems, because we do not have a proven relationship between the distribution of the lifetime *P D* and the com ponent monthly values. The example here is evidence enough that it should be tested. In the case where the forecast does not match the median of the distribution, the institution’s loss reserves should be increased to cover this under-prediction relative to the asymmetric risk distribution.

The range of possible outcomes from the different models should be included as a model risk component in economic and regulatory capital. Loss reserves are set according to the institutions best expectation of future losses, so that should be set using the chosen model and any known bias adjustments. Uncertainty is always included in capital. At least in this example, propagating the prediction intervals through all of the calculations provides a final estimation uncertainty for the 95% confidence interval (between 2.5% and 97.5%) that is comparable to the uncertainty seen when comparing different models. This does not provide proof in all cases, but it suggests that careful error propagation calculations for a single model provide a reasonable estimate of the model selection risk, even without considering the range of outcomes that could have been observed by

11

studying all of the unselected models.

6 Conclusions

Model selection risk is a component of any model development project. Models that incorporate macroeconomic time series elements afford the opportunity to explicitly study this risk. The simplest finding of this work is that propagat ing the prediction intervals of point forecasts through the many steps typically needed to get to a final forecast does, in fact, capture much of the uncertainty that was revealed by looking at the forecasts of the many unselected models. This can be done even when multiple models are not explicitly available.

Logistic regression models that are the underpinning of so much of the mod eling in lending should have normally distributed prediction intervals in logit space. However, when they are converted back to probability space and com bined with many other model elements, the final forecast cannot be assumed to have a simple distribution. Therefore, Monte Carlo simulation as used here was found to be an effective method to propagate the errors and obtain results useful to assessing the model risk.

Once the distribution of possible answers is known, any adjustment to the expected value can be made. Equally important, this approach quantifies the model risk component needed when computing economic capital.

References

[1] Tiziano Bellini. *IFRS 9 and CECL Credit Risk Modelling and Validation: A Practical Guide with Examples Worked in R and SAS*. Academic Press, 2019.

[2] Dennis E. Bennett. Governance and organizational requirements for effec tive model risk management. *Journal of Risk Model Validation*, 11(4):97– 116, 2017.

[3] Financial Accounting Standards Board. Financial instruments - credit losses (topic 326). Technical Report 2016-13, June 2016.

[4] Joseph L. Breeden. *Living with CECL: Mortgage Modeling Alternatives*. Prescient Models LLC, 2018.

[5] Joseph L. Breeden, LiMing Brotcke, Bin Duan, Andy Johnson, Charles Maner, and Paul ONeal. Impacts of covid-19 on model risk management. *MRMIA Best Practices*, 1:85–125, 2021.

[6] European Commission. On artificial intelligence - a european approach to excellence and trust. Technical Report Brussels, 19.2.2020 COM(2020) 65 final, 2020.

12

[7] FRB. Supervisory guidance on model risk management. Technical report, Board of Governors of the Federal Reserve System, April 2011.

[8] IASB. IFRS 9 financial instruments. Technical report, IFRS Foundation, 2014.

[9] Joseph McPhail and Lihong McPhail. Forecasting lifetime credit losses: Modelling considerations for complying with the new FASB and IASB cur rent expected credit loss models. *Journal of Risk Management in Financial Institutions*, 7(4):375–388, 2014.

[10] Massimo Morini. *Understanding and Managing Model Risk: A practical guide for quants, traders and validators*. John Wiley & Sons, 2011.

[11] Florian Ostmann and Cosmina Dorobantu. AI in financial services. Tech nical report, 2021.

[12] Seppe vanden Broucke and Bart Baesens. Managing model risk: Lessons and experiences from industry and research on the challenges and dangers of analytical models. 2021.