Finity Banking Series

October 2020

Using Ensemble Models to Mitigate Concept Drift and Data Variance in Credit Modelling

Authors: Dr. Paul Beinat and Sen Nagarajan

Introduction

Measurement of credit risk by banks relies heavily on predictive models. Credit risk models aim to exploit patterns or relationships observed in historic data to predict future default outcomes accurately. This is challenging when relationships between default and the drivers of default change usually in response to business and economic environments evolving. Banks using credit models observed this during the Global Financial Crisis (GFC) particularly in regions such as North America and Europe which suffered an economic decline. Banks are now facing a similar situation where economies have entered a recession caused by a pandemic. This recession will be different to the previous economic downturns. Over this period, the predictive accuracy of current models is likely to degrade.

Credit models are especially important over a downturn to help a bank continue to lend while managing its risk. Accurate credit models ensure that the bank is scoring new loans accurately, as well as, setting adequate loan provisions and capital. Credit models can also inform credit management and collections operations, through providing insights into prioritising and selecting actions for troubled or defaulted loans.

Classes of models used in credit modelling handle changes in the underlying data differently. Data can change either as the sample changes e.g. simply from collecting new historic data over time, or as the environment changes. These changes are referred to as data variance and concept drift respectively. Banks have historically relied on generalised linear models (GLMs); in particular, logistic regression models to predict default behaviour. GLMs are relatively simple, tractable, and easy to interpret. Theoretical and experimental studies show that data variance drives GLM model variance, i.e. GLMs are sensitive to the sample used to train the model. Ensemble models however have properties that handle data variance and concept drift, better.

Ensemble models however are considered opaque and don't have simple model parameters that can be interpreted like a GLM. The choice between ensemble models and GLMs is a trade-off between predictive accuracy and interpretability. Modern techniques can be applied to understand what drives predictions from ensemble methods and these should be used as part of model validation and review. While model choice and model fitting is important, of course modelling teams still need to understand the data thoroughly, as well as the context in which the model is applied. There may still be applications where GLMs are the best choice for the given purpose perhaps due to factors such as simplicity and interpretability. However, where there is more freedom, alternative approaches should be explored, particularly where predictive accuracy is important.

In this article, Dr. Paul Beinat and Sen Nagarajan discuss how ensemble model classes are better at handling data variance and concept drift. Dr. Beinat worked with a large analytics vendor in the US exploring the impact of concept drift in a large data set of mortgage loans over the GFC. As part of this work he studied the performance of different models and has validated these over the GFC. He has also explored the impact of data variance in an experimental study on a large auto insurance data set for a US insurer. Sen Nagarajan has worked with Australian banks to develop credit risk models for internal capital models, credit risk stress testing and economic capital models where selecting models that work across a range of economic conditions is critical.

Concept Drift and Data Variance

Concept drift and data variance degrade the ability of a model to predict accurately. Data variance describes the variation observed in samples of data. Models trained on different data samples from the same population will have different parameters and produce different predictions for the same set of predictor variables. The prediction error of a model can be decomposed into the square of bias, model variance and an irreducible error component. Bias is a measure of how well the model fits features of the data, while variance is how model calibration changes in response to training data variance. At the extremes, a model that simply uses the mean of the data for all predictions exhibits low variance, the mean will almost be the same for different samples from the data population, but the consequence is high bias – all estimates are biased towards the mean. A model that is essentially a memory of the data exhibits low bias, all the estimates match the data, but high variance since it will create a very different memory model on a new data sample. Neither would be acceptable models in the real world. Ideally the chosen model will minimise both bias and variance, however there is always some trade-off. Where a model class has high variance, then the model parameters and predictions will change as the data sample changes. GLMs reduce prediction error by having low bias, however they consequently exhibit high model variance, driven by the training data used.

The model variance exhibited by the statistical models, such as GLMs, have other implications. There are no reliable statistics that indicate when the optimum model has been reached. These statistics are also subject to the same variance as their underlying models.

Concept drift is when the underlying relationships between predictor variables and the outcome that the model is trying to predict changes unexpectedly. This is observed in many situations as the environment changes over time. Two different forms of drift occur. One is where the joint probability of the predictors and variables changes, as a simple, example perhaps we have not experienced interest rates so low before. Another is where the relationships between the predictors and dependent variable change. A key objective of model monitoring is to detect such concept drifts. Degrading model performance then triggers a review of the model and potentially a recalibration of the model to correct for concept drift. Better models should be less susceptible to concept drift, that is, the predictive capability is less eroded by concept drift. This does not remove the need for monitoring and periodic retraining of models for concept drift. There is a potential for financial loss where concept drift results in poor predictive accuracy e.g. misclassification of ‘bad’ loans as ‘good’.

GLMs Short-comings

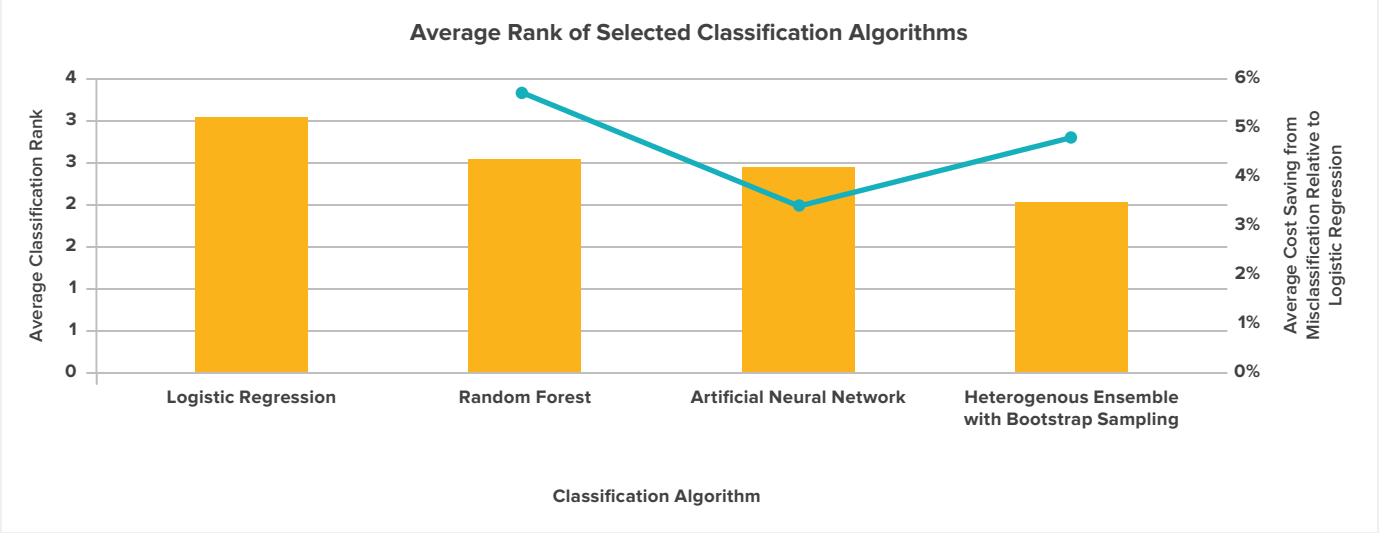
GLMs have been used in a number of applications in risk modelling both in insurance and banking. Credit risk models have applied logistic regression to predict loans going into default. They are easy to implement and interpret. This has been important as credit risk models are subject to governance and need to under go detailed validation and challenge. However GLMs do not handle data variance and concept drift as well as other model classes.

Paul has carried out research on a large car insurance claims dataset from a US insurer. In this test, the data is partitioned such that the characteristics of partitions are similar. Different models are trained on both partitions and the prediction compared using a test dataset. When the GLMs model parameters were selected using stepwise procedures using either AIC or BIC, these models would select a different numbers of parameters in a different order when compared between both datasets. In addition, the parameters of the GLM models were very different. When the models were applied for test dataset, predictions varied more compared to ensemble models which had better accuracy. This outcome is backed by research that proves GLMs are worse than GBMs and Random Forests in handling data variance.

Similarly GLMs appeared to perform worse when there is concept drift. In an experiment using a proprietary US home lending data set, logistic models trained on data prior to the GFC performed worse post GFC compared to ensemble model types. The ensemble models had a significantly better error rate compared to GLMs in the out-of-time validation. This has significant implications for model reliability.

Alternative Modelling Approaches

Ensemble models, for example, random forests and gradient boosting machines (GBMs), have variance reduction properties and are very flexible. Random forests use an average of multiple models and include an element of randomness that works to reduce data variance. Research into classification algorithms demonstrates that ensemble algorithms can achieve better predictive accuracy than traditional approaches such as logistic regression. One example[1] of a comprehensive study compared 41 models across 8 different datasets evaluated on a range of criteria using a test dataset. The average ranking across the different datasets and evaluation metrics is shown below. The authors also calculated the reduction in losses from misclassification from the different approaches. Both the random forest and a heterogeneous ensemble outperformed logistic regression in predictive accuracy.



The ability of ensemble models to handle data variance also extends to concept drift. Factors such as the number of observations, number of predictors, and mix of predictor variables also influences how well an ensemble model handles data variance relative to GLMs. As described above, Paul's experimental analysis on partitions of the insurance dataset, ensemble methods trained on different data samples produced more consistent predictions on the test dataset. In the credit modelling experiment, ensemble models produced better predictive outcomes over periods such as the GFC when conditions changed markedly, i.e. they handled concept drift better.

Some of the arguments against novel ensemble models are lack of transparency and interpretability. Credit models can be classified as medium or high risk in a bank's model risk framework depending on their purpose. Therefore credit risk models need to be validated and be subject to review and challenge. GLMs are tractable and it is easier to explain how the GLM makes its predictions. An examination of the model parameters provides insights into how each variable influences the prediction and the importance of the predictors. Model developers and validators are familiar with tests and methods for goodness of fit, statistical significance of predictors and analysis of residuals when using GLMs. Care should be taken with such statistical significance measures such as Z scores, P values and standard errors, as these are calculated only during the last iteration of the GLM, when it takes its smallest step to the solution.

While ensemble models are not tractable, significant insight into how variables influence predictions can be uncovered. For example, there are approaches to understand variable importance, variable interactions and how changes in predictor variables interact with the outcome variable for ensemble methods. By applying appropriate model diagnostic techniques newer model classes can be readily managed and governed within a bank's existing model risk framework.

Conclusion

Contemporary machine learning models, particularly ensemble model types, have variance reduction characteristics which mean these models are less sensitive to the data sample used to train the models. Ensemble models also appear to handle concept drift better, as observed using credit data over the GFC.

While the interpretation of ensemble models is different, significant insights can still be drawn on how these models use predictor variables in the outcome.

1 Lessmann, Stefan & Baesens, Bart & Seow, Hsin-Vonn & Thomas, Lyn C., 2015. "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," European Journal of Operational Research, Elsevier, vol. 247(1), pages 124-136.

Dr. Paul Beinat

paul.beinat@finity.com.au Mobile | +61 408 080 622 Sydney Office | +61 2 8252 3461

Sen Nagarajan

sen.nagarajan@finity.com.au Mobile | +61 434 195 740 Melbourne Office | +61 3 8080 0900