MACHINE LEARNING MODELS – VALIDATION AND THE STRESS TESTING OF CREDIT RISK

Michael Jacobs, Jr.
Senior Quantitative analytics & Modelling Expert
PNC Financial Services Group

Introduction and Motivation

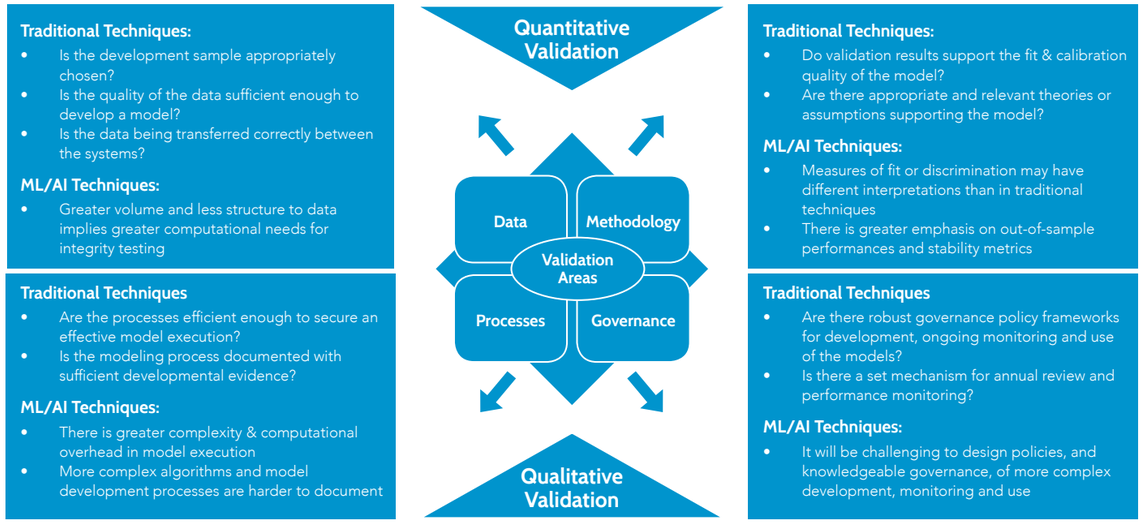
In the aftermath of the financial crisis, regulators have utilized stress testing as a means by which to evaluate the soundness of financial institutions' risk management procedures. The primary means of risk management, particularly in the field of credit risk, is through advanced mathematical, statistical and quantitative techniques and models, which leads to model risk. Model risk can be defined as the potential that a model does not sufficiently capture the risks it is used to assess, and the danger that it may underestimate potential risks in the future. Stress testing ("ST") has been used by supervisors to assess the reliability of credit risk models, as can be seen in the revised Basel framework and the Federal Reserve's Comprehensive Capital Analysis and Review ("CCAR") program.

ML / AI has, in fact, had several applications in finance well before the advent of this modern era. The high volume and accurate nature of the historical data, coupled with the quantitative nature of the finance fields, has made this industry a prime candidate for the application of these techniques. The proliferation of such applications has been driven by more powerful capabilities in computing power and more accessible ML / AI methodologies. The fields of financial (ie. credit, market, business and model) as well as that of non-financial (ie. operational, compliance, fraud and cyber) risk modeling is, and has been, a natural domain of application for ML / AI techniques. Indeed, many work-horse modeling techniques in risk modeling (eg. logistic regression, discriminant analysis, classification trees etc. can be viewed in fact as much more basic versions of the merging ML / AI modeling techniques of the recent period. That said, there are risk types for which ML /AI has a greater degree of applicability than others – for example, one would more likely find this application in data-rich environments such as retail credit risk scoring (eg. credit card, mortgages), as compared to relatively data-poor domains such as low default credit portfolios for highly rated counterparties (eg. sovereigns, financials, investment grade corporates). In the non-financial realm, we are seeing fruitful application in areas such as fraud analytics, where there is ample data to support ML / AL estimations.

In Figure 1, we depict the model validation function as the nexus of four core components (data, methodology, processes and governance), and two dimensions in the spectrum from quantitative to qualitative validation methodology. The graphic highlights examples of some differences in the context of ML/AI modeling methodology. It is clear that many of the validation elements that have been the practice for traditional models will carry over to the ML / AI context, and the differences will be in emphasis or extensions of existing techniques. >>

> *"In the aftermath of the financial crisis, regulators have utilized stress testing as a means by which to evaluate the soundness of financial institutions' risk management procedures. The primary means of risk management, particularly in the field of credit risk, is through advanced mathematical, statistical and quantitative techniques and models, which leads to model risk."*

Figure 1 – The Model Validation Function and Challenges in ML/AI Modeling Methodologies



An Empirical Study of AI / ML and Stress testing

As part of the Federal Reserve's CCAR stress testing exercise, U.S. domiciled top-tier bank companies BHCs are required to estimate potential losses under stressed operating conditions. The adverse scenario is described by quarterly trajectories for key macroeconomic variables ("MVs") over the next nine quarters, or for thirteen months to estimate loss allowances. In addition, the Federal Reserve generates its own supervisory stress scenarios, so that firms are expected to apply both BHC and supervisory stress scenarios to all exposures. Jacobs (2018) considers a diverse set of macroeconomic drivers representing varied dimensions of the economic environment, and a sufficient number of drivers balancing the consideration of avoiding over-fitting by industry standards (ie. at least 2-3, and no more than 5-7, independent variables) are considered.

The model selection process imposed the following criteria in selecting input and output variables across both multiple time series vector autoregressive ("VAR") and the ML / AI multivariate adaptive regression spline ("MARS") models:

- Transformations of chosen variables should indicate stationarity;
- Signs of coefficient estimates are economically intuitive;
- Probability values of coefficient estimates indicate statistical significance at conventional confidence levels;
- Residual diagnostics indicate white noise behaviour;
- Model performance metrics (goodness of fit, risk ranking and cumulative error measures) are within industry-accepted thresholds of acceptability;
- Scenarios rank order intuitively (ie. severely adverse scenario stress losses exceeding scenario base expected losses).

Similarly, we identify the following loss segments (with loss measured by Gross Charge-offs – "GCOs") according to the same criteria, in conjunction with the requirement that they cover the most prevalent portfolio types in typical traditional banking institutions:

- Commercial and Industrial ("C&I");
- Commercial Real Estate ("CRE");
- Consumer Credit ("CONS").

In the case of C&I, the best model for the quarterly change in charge-off rates was found, according to the model selection process, to contain the transformations of the following macroeconomic variables:

- Real GDP: lagged 4 quarters
- Corporate Bond Spread to Treasuries: 2 quarter change lagged 4 quarters

In the case of CRE, the best corresponding model is found to be:

- BBB Corporate Yield: 4 quarter change lagged 2 quarters
- Unemployment Rate: lagged 1 quarter

In the case of CONS, the best corresponding model is found to be:

- BBB Corporate Yield: 3 quarter change lagged 4 quarters
- Unemployment Rate: lagged 1 quarter

In Table 1, we present a comparison of model performance metrics (Generalized Cross Validation – GCV, Squared Correlation - SC, Root Mean Squared Error – RMSE, Cumulative Percentage Error – CPE and Aikaike Information Criterion - AIC) across key time periods (development sample: 3Q99-4Q14, full sample: 3Q99-3Q16, downturn period: 1Q08-1Q10, out-of-sample: 4Q14-3Q16) for the MARS and VAR estimations. The author makes the following conclusions in comparing the model estimation results: >>

Table 1: Loss Estimation Results for C&I, CRE and CONS Segments– VAR and MARS Model Performance Metrics Comparison (Historical Y9 Credit Loss Rates and Federal Reserve Macroeconomic Variables 4Q99–3Q14)

		Model Performance Metrics	Development Sample	Full Sample	Downturn Period	Out-of-Time Sample
Multivariate Adaptive Regression	Commercial and Industrial	Generalized Cross Validation	3.47E-06	2.93E-06	1.87E-05	3.18E-05
		Squared Correlation	34.16%	39.28%	79.68%	27.37%
		Root Mean Squared Error	1.80E-03	1.58E-03	2.49E-03	1.41E-02
		Cumulative Percentage Error	-1.55E-07	-1.34E-08	-4.20E-02	-9.53E-02
		Aikaike Information Criterion	-5.07E+02	-6.29E+02	-8.60E+01	-9.41E+01
Vector Autoregression		Generalized Cross Validation	5.11E-06	3.12E-06	2.02E-05	4.07E-05
		Squared Correlation	24.05%	33.26%	12.68%	24.88%
		Root Mean Squared Error	2.20E-03	1.73E-03	3.67E-03	1.78E-02
		Cumulative Percentage Error	-2.12E+00	2.33E+00	6.58E-01	-1.84E+00
		Aikaike Information Criterion	-4.86E+02	-5.75E+02	-7.88E+01	-8.24E+01
Multivariate Adaptive Regression	Commercial Real Estate	Generalized Cross Validation	3.41E-06	3.14E-06	9.61E-04	7.23E-05
		Squared Correlation	63.43%	62.57%	4.19%	0.61%
		Root Mean Squared Error	1.77E-03	2.08E-03	1.20E-02	2.16E-03
		Cumulative Percentage Error	-1.07E-15	1.58E-15	-2.93E-01	-4.83E-01
		Aikaike Information Criterion	-5.30E+02	-5.52E+02	-6.74E+01	-8.31E+01
Vector Autoregression		Generalized Cross Validation	4.862E-06	3.76E-06	1.50E-03	5.14E-04
		Squared Correlation	42.73%	52.43%	3.68%	0.52%
		Root Mean Squared Error	3.03E-03	2.84E-03	1.62E-02	1.79E-02
		Cumulative Percentage Error	-1.03E+00	4.00E-01	-3.77E-01	-1.80E+01
		Aikaike Information Criterion	-4.54E+02	-4.75E+02	-6.37E+01	-4.55E+01
Multivariate Adaptive Regression	Consumer Credit	Generalized Cross Validation	9.83E-06	8.35E-06	5.07E-05	2.92E-05
		Squared Correlation	44.74%	46.23%	24.95%	2.66%
		Root Mean Squared Error	3.03E-03	2.46E-03	4.50E-03	2.70E-03
		Cumulative Percentage Error	1.16E-15	-9.71E-17	-2.59E-01	8.58E-01
		Aikaike Information Criterion	-4.57E+02	-5.38E+02	-8.22E+01	-7.71E+01
Vector Autoregression		Generalized Cross Validation	1.78E-05	1.68E-05	5.44E-05	5.11E-04
		Squared Correlation	39.23%	39.83%	20.00%	1.16%
		Root Mean Squared Error	3.79E-03	2.77E-03	5.55E-03	1.79E-02
		Cumulative Percentage Error	-9.89E-01	-5.26E-01	-4.73E-01	1.22E+02
		Aikaike Information Criterion	-2.33E+02	-2.61E+02	-6.94E+01	-4.55E+01

RISK INSIGHTS | EMERGING TECHNOLOGY

- We observe that, generally, across metrics and time periods, the MARS model outperforms the VAR model.
- There are some notable differences across segments – in particular, for the CRE and CONS portfolios, the out-of-sample performance of the VAR model is much worse than the MARS model.
- Furthermore, the MARS model is generally more accurate over the downturn period than the VAR model, showing more accuracy and less underprediction.
- Finally, according to the CPE measure of model accuracy – one preferred by regulators – the MARS model performs better by several orders of magnitude.

Across modeling segments, the author observes that the MARS model exhibits greater separation in cumulative loss between Base and either Adverse or Severe scenarios. Furthermore, while the Base scenarios are lower in the MARS than in the VAR model, in the former model the Adverse and Severe scenarios have much higher cumulative losses than in the latter model. Furthermore, the spread in cumulative losses between the Severe (Adverse) and Base scenarios is greater in the MARS model than in the VAR model.

In summary, this study has examined a critical input into the stress testing process, the macroeconomic scenarios provided by the prudential supervisors to institutions for exercises such as the Federal Reserve's CCAR program. The author analyzed a common approach of a VAR statistical model that exploits the dependency structure between both macroeconomic drivers, and has proposed a challenger model. Across modeling segments, he observes that the MARS model exhibits greater separation in cumulative loss between Base and either Adverse or Severe scenarios. ■

"Across modeling segments, the author observes that the MARS model exhibits greater separation in cumulative loss between Base and either Adverse or Severe scenarios. Furthermore, while the Base scenarios are lower in the MARS than in the VAR model, in the former model the Adverse and Severe scenarios have much higher cumulative losses than in the latter model."