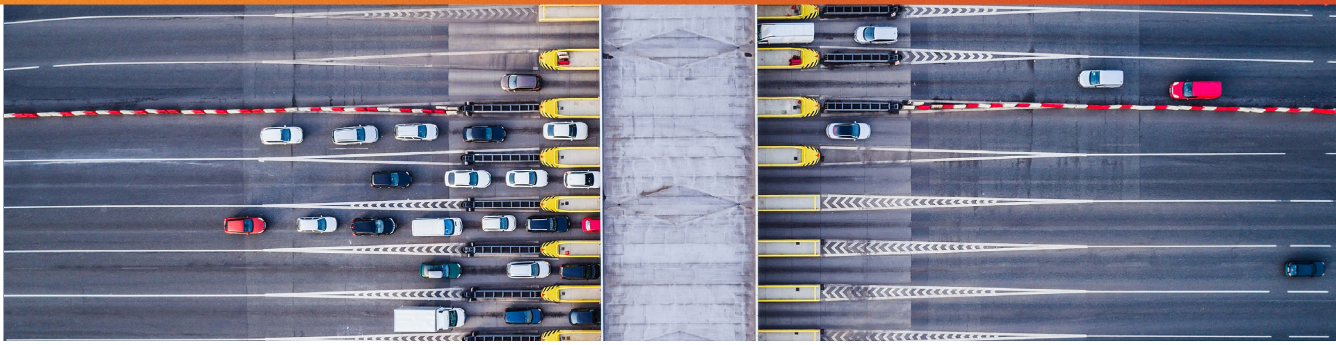
• • • POINT OF VIEW

protiviti®
Face the Future with Confidence



Validation of Machine Learning Models: Challenges and Alternatives

The potential of machine learning (ML) to deliver value to banks has created something of a gold rush in adopting this methodology for banking applications. ML can produce immense benefits when applied to complex nonlinear problems where there is a large amount of data, particularly unstructured data. Use cases for incorporating machine learning in banking include asset management, fraud detection, credit risk management and regulatory compliance, to name a few. More specifically, large banks are turning to ML models as an alternative to traditional models to gain faster, more accurate and insightful predictions and classifications in their risk management and financial management business decisions.

Because they are more complex and less transparent than traditional models, ML models pose a unique set of challenges to model risk management and model validation. While the complexity of ML systems brings an increased ability to derive actionable insights, it also introduces new dimensions of model risk. In our experience, regulators expect ML models to comply with the standards of SR 11-7 and OCC 2011-12,[1] the supervisory guidance for model risk management (MRM) that guides traditional model development and validation. These regulations also require that decision-makers understand a model's limitations and original intent and avoid using the model in ways that are inconsistent with that intent.

[1] SR 11-7 and OCC 2011-12, adopted by the FDIC as FIL 22-1017, were issued, respectively, by the Board of Governors of the Federal Reserve (FRB) and the Office of the Comptroller of the Currency (OCC).

SR 11-7 defines model risk as "the potential for adverse consequences from decisions based on incorrect or misused model outputs and reports." Model risk can occur when a model is built as it was intended but has fundamental errors that cause it to produce inaccurate outputs when viewed against the design objective and intended business use. It also can occur when a model is implemented or used incorrectly or inappropriately, or when its limitations or assumptions are not fully understood.

This paper discusses some of the challenges related to ML model validation and provides guidance for addressing issues that may impact banks.

> *"The increased complexity of machine learning models can create unique challenges for validation teams. Validators need to be prepared to use alternate methods or develop custom methods to meet regulatory requirements."*
>
> — Shaheen Dil, Senior Managing Director, Protiviti

Challenges in Validating ML Models

According to SR 11-7 and OCC 2011-12, model validators should assess models broadly from four perspectives: conceptual soundness, process verification, ongoing monitoring and outcomes analysis. The problem is that many model users and validators in the banking industry have not been trained in ML and may have a limited understanding of the concepts behind newer ML models. Even with a demonstrated interest in data science, many users do not have the proper statistical training and often resort to ML "plug and play" packages sold by third parties to develop ML models. This poses risk to the models' fitness for use, which is required by SR 11-7 and OCC 2011-12, and to the validation of the models as well, especially if the model developers and model users have a limited understanding of the algorithms powering the models and treat them like black boxes.

Conceptual Soundness

Assessment of the conceptual soundness of models involves assessing the quality of the model design and construction, reviewing the model documentation, assessing empirical evidence, and confirming that the variable selection process used in the model is conceptually sound. Demonstrating the conceptual soundness of the models will be difficult if the math behind the ML theory used to design them is not well understood by the model developers, users and validators. The following factors should be considered during the assessment of conceptual soundness.

Data Integrity/Representativeness

SR 11-7 and OCC 2011-12 require that the data used for model development be representative of the bank's portfolio and market/business conditions. However, because ML uses large volumes of structured and unstructured data, the dimensionality of the ML modeling features is much broader and deeper, making it challenging to ensure data integrity and representativeness.

Bias

There is a tendency in ML to ignore bias in data because of the large sample sizes. Many times, the data used in models is generated by human decision-makers, so any inherent bias in human decision-making is carried over to the development data. Additionally, the data-generating process itself can be biased. For example, collection bias can occur if the data collection is conducted with too many exclusion criteria or the data is collected only for specific situations or scenarios. If loan officers have historically made biased decisions in rejecting individuals belonging to a certain race, gender or age group, for example, the development data will reflect these biases. If a machine-learning model is developed on this data to predict the risk of loan defaults, it will most likely discriminate against extending credit to individuals belonging to these groups.

Bias in ML models can trigger costly errors. One way to identify data bias is by benchmarking with other models or the opinion of subject-matter experts. Appropriate data de-biasing techniques should be used to remove bias from development data. In addition to traditional methods such as downscaling and quantile mapping, methods such as randomization and sample weighting should also be incorporated to correct data bias. The statistical soundness of selecting unbiased development and holdout data should be given extra emphasis for ML models.

Explainability Challenges

Machine learning models (especially neural network-based models) are difficult to explain and are often viewed as black boxes. Assessment of the variable selection process and explainability of driving factors become difficult due to the complexity and architecture of neural networks. Even if ML models perform better than traditional models, the lack of explainability may cause ML models to be restricted in use by model validation and MRM teams.

While ML can be a valuable tool in credit risk management, regulations in the credit area require that negative credit decisions be explained. Many other banking applications require understanding the driving factors behind models as well. In both cases, model validators need to set standards for requirements of explainability for ML models so that ML models with inadequate explainability are identified and remediated before they can be used in making credit decisions. While many organizations and vendors have been claiming proprietary methods for explaining ML models, most of these are still not verified or validated.

Parameter and Method Selection

ML model development techniques normally involve scaling, normalization, parameter optimization, randomization and activation functions. ML algorithms are fairly sensitive to the selection of these parameters/methods. The way normalization, parameter optimization and feature selection are conducted when developing ML models can impact test error estimation. Validators must therefore evaluate whether the choice of these items is conceptually sound.

Model Documentation

SR 11-7 and OCC 2011-12 require that model documentation be comprehensive and detailed enough so that a knowledgeable third party can recreate the model without having access to the model development code. The complexity of ML models and the model development process is likely to make documentation of ML models much more challenging than traditional model documentation. It is recommended that banks standardize their model development and validation procedures for ML models and provide a model documentation template that is consistent with regulatory expectations.

Process Verification

The process verification component of SR 11-7 and OCC 2011-12 requires that effective controls are in place to ensure proper model implementation. Before a model is implemented or changed in production, the bank must follow its model validation and approval processes. Given the computational complexity of ML models and the different deployment platforms, process verification can be challenging. SR 11-7 and OCC 2011-12 require that all models be approved by the validation team before use, and all significant model changes, as well as their impact on model output, must be assessed and validated before models are used. Therefore, validators need to be trained in implementation testing across the machine learning deployment platforms that the bank uses.

Some ML models are designed to redevelop automatically on a dynamic basis, which makes validation and approval of the model changes challenging. What constitutes a significant change to the model and how does one validate that change in cases when ML models are being dynamically redeveloped? Either the ML infrastructure has to have the capability to save model changes and related input/output data separately until a validation can be done, or the validation team has to determine how to validate the process of automatic redevelopment of ML models and evaluate whether this process poses any model risks.

Ongoing Monitoring

SR 11-7 and OCC 2011-12 require that model performance, identified model risks and limitations be monitored on a frequency commensurate with the frequency of model use. As mentioned above, some ML systems use automated processes to redevelop the models without any human intervention or ongoing monitoring. This aspect of ML systems needs to be monitored and validated, especially when the model changes significantly from a prior version. This could be more frequent than in traditional models. Advanced algorithms can correct for statistical errors but they cannot distinguish errors with high business costs from those with low business costs, so it is important to understand and validate the business logic behind the automated re-estimation of ML models. The capability to perform ongoing monitoring of ML models should be developed, and validation teams must put a high emphasis on this process.

Outcomes Analysis

Outcomes analysis helps evaluate model performance and tests for model accuracy and stability. ML models (especially neural networks) are prone to overfitting and underfitting problems. Specifically, simpler models lead to underfitting, or high bias (see Figure 1), where more complex models lead to overfitting, or high variance (see Figure 2). The methodology used by ML model development techniques to address the bias-variance tradeoffs should be carefully examined by model validators.

• • • Figure 1: Simple models can underfit and lead to high bias

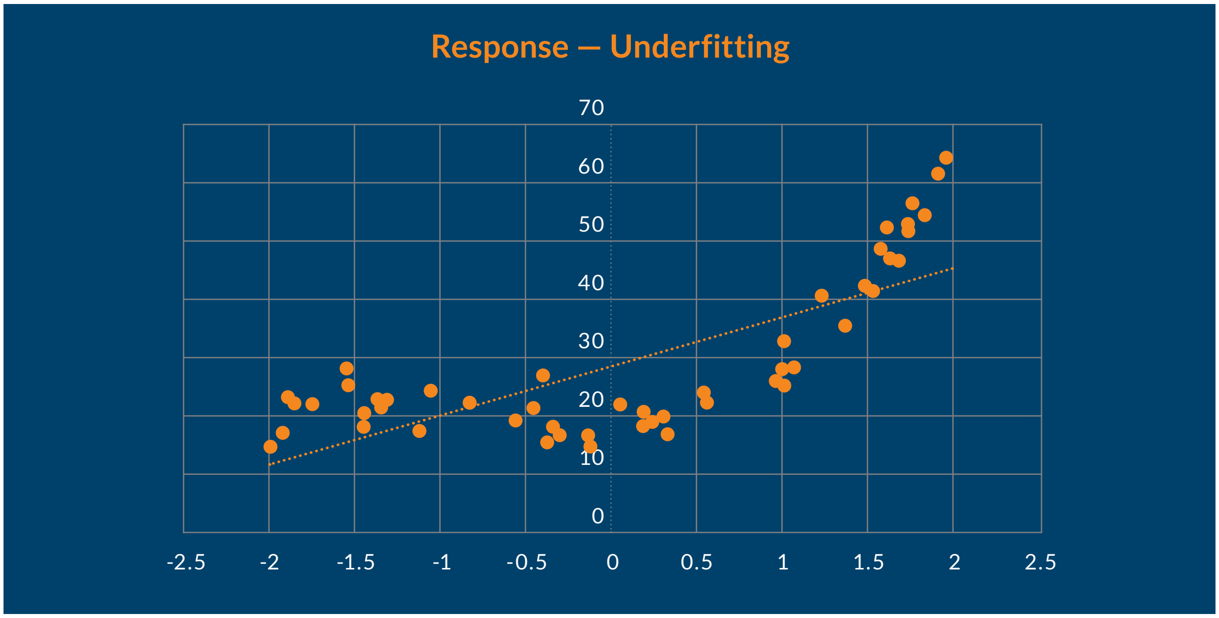
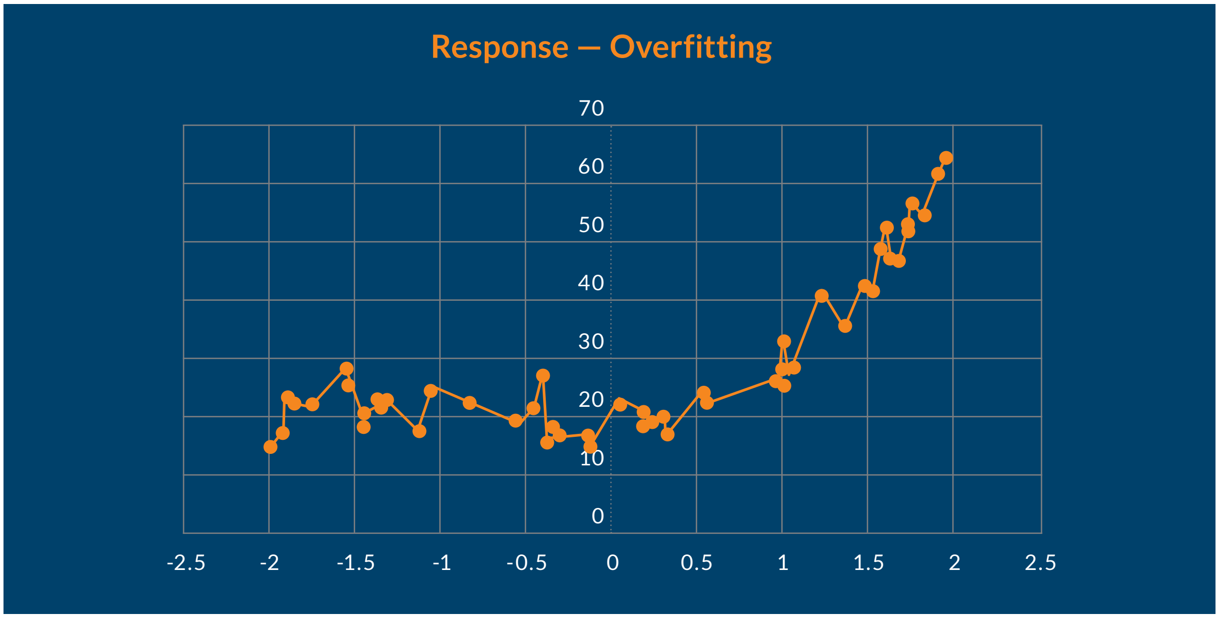


Figure 2: Complex models can overfit and lead to high variance



Although standard out-of-sample backtesting works well for traditional models, it may not work well for ML models. Normally, k-fold cross-validation is a recommended technique for detecting and preventing overfitting in ML models. Validators should evaluate the use of normalization and feature selection within the context of k-fold cross-validation to ensure no information from the training data sample is leaked into testing data.

Sensitivity analysis of ML models may be hard to interpret, especially if there is a lack of explainability of neural network-based models. Since the inputs and outputs in neural network models are not linked as they are in statistical models using linear or logistic regression, performing sensitivity analysis can become computationally intensive, and the interpretation of sensitivity analysis results can be difficult to sort out. This challenge is the same as the one identified in the Explainability Challenges section above.

Vendor Models

SR 11-7 and OCC 2011-12 require that all vendor and third-party models be subjected to the same rigor as internally developed models. The same policy applies to third-party ML models. The proprietary nature of vendor models has prevented their comprehensive validation. Typical approaches to vendor-model validation have consisted of outcomes analysis, sensitivity analysis and benchmarking. In the ML space, vendors may have developed the models based on proprietary data and may be unwilling to share the development and holdout data required for backtesting and other validation testing. As discussed above in the Outcomes Analysis section, both outcomes analysis and sensitivity analysis can be challenging, especially for vendor models. Banks have to use alternative approaches such as proof-of-concept, periodic review of conceptual soundness, and more frequent ongoing monitoring to assess the applicability of the vendor model to the bank's needs.

Conclusion

While machine learning has the potential to enhance the quality of quantitative models in terms of accuracy, predictive power and actionable insights, the increased complexity of these models poses a unique set of challenges to model validators. Simply using traditional model validation methods may lead to rejecting good models and accepting bad ones. Model validators need to understand these challenges and develop customized methods for validating ML models so that these powerful tools can be deployed in the banking industry with greater confidence while minimizing model risk.

About Protiviti's Model Practice

Protiviti's Model Validation team consists of professionals specialized in machine learning, with many years of industry experience at leading banks and technology firms and a deep knowledge of all aspects of SR 11-7 and OCC 2011-12. Protiviti can assist with both the developing of ML model governance and validation frameworks and with ML model validations.

Contacts

Shaheen Dil	**Suresh Baral**	**Lucas Lau**
+1.212.603.8378	+1.212.471.9674	+1.212.603.8398
shaheen.dil@protiviti.com	suresh.baral@protiviti.com	lucas.lau@protiviti.com

protiviti®