White paper

Model Risk Management in Banking

Gary Class Industry Strategist, Financial Services at Teradata

teradata.

Version 1.0

Introduction

2 Introduction

3 The Teradata Signal Framework

4 Introduction to model risk management

6. Legal framework

7 Banking regulations

9. Monitoring feature quality

11 Measuring model performance

14 Model operations

15 Artificial intelligence models require new techniques

16 Model risk management banking practice

18 Teradata Trusted AI

19 How Teradata can help

Models play a crucial role in today's banking industry, shaping decisions that range from credit approvals to business lending. However, when models fail, they can expose institutions to significant financial and operational risks. One notable example is Zillow's 2020 venture into the residential housing market. The company applied its Automated Valuation Model to predict home prices but failed to account for market volatility, resulting in over $420 million in losses within just a few months.

This underscores the critical need for robust model risk management. In banking, model risk management ensures that models are rigorously tested, validated, and continuously monitored to avoid costly errors. Thanks to today's ever-increasing regulatory scrutiny, it's no longer enough to simply develop models; banks must implement frameworks that govern their safe and effective use. This paper explores the key components of model risk management in banking and how institutions can mitigate the risks posed by their models with cloud data and analytics solutions from Teradata.

The Teradata Signal Framework

While developing and deploying a model is the focus of attention, no value is created until the model is deployed into production to support a business application. Getting a model from the "path to production" in technology to the "point of decision" within a customer-facing business application is extremely challenging. No economic value is generated by expenditures on the data and analytics infrastructure to develop a model until the model is utilized to inform business decisions that ultimately impact customers.

The goal of the Teradata ecosystem is to provide an integrated framework where data serves as the foundational layer for analytics, encompassing the collection, organization, and storage of raw materials, refined by Teradata's extensive experience in optimal multi-thread processing and in-database analytics. Banks need to expand their focus beyond the data processing pipeline and model operations platform to visibility into how the signal generated from models is successfully consumed at the customer-facing point of decision in servicing workflows and enterprise applications.

Signal Framework

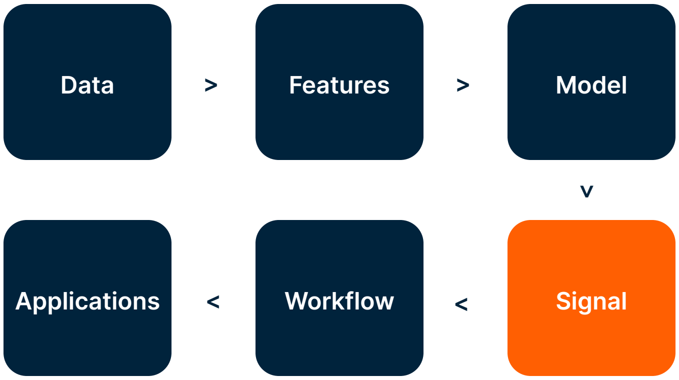


Exhibit A: The Teradata Signal Framework

Features are derived from the raw data and are crucial to enhance model performance and robustness. The model encodes intelligence by crawling multi-dimensional data to predict an outcome or decide an action via the distillation and propagation of the salient signal. The signal is the actionable insight derived from the model and represents the distilled intelligence that the organization can exploit to make informed decisions. In practical terms, the signal is what the users interact with as it's the final product of the data refining and model learning process. Intelligence refers to the application of signal in the workflow; it's the process of interpreting, optimizing, and applying the insights derived from the data to facilitate decision-making by the organization.

Signal Framework	
Data	Raw operational and administrative data: structured, semi-structured, or unstructured
Features	Transformation of data into standardized, harmonized, and conformed attributes
Model	Quantitative method to process input data into estimates of outcomes
Signal	Actionable insight generated by a unique pattern in the data as identified by the model
Workflow	Standardized business process to achieve a business outcome by leveraging the signal
Applications	Configuration of software tools that enable delivery and monitoring of workflows

Exhibit B: The Teradata Signal Framework explained

Introduction to model risk management

Changes in the regulatory framework after the Great Recession and a focus on managing reputational risk requires banks to adopt a rigorous model risk management approach to the governance of models both during development and after deployment. The critical activity to mitigate risk is a comprehensive approach to model validation, which incorporates a focus on input data quality and model output stability. The Federal Reserve's "Supervision and Regulation 11-7 Guidance on Model Risk Management," which we will discuss at length in this whitepaper, provides a durable framework to address model risk as outlined in Exhibit C.

Model risk management	
Feature quality	Integrity and validity of the process to transform data into the highly curated, documented, and catalogued feature store used for model development
Conceptual soundness	Quality of the model design and construction, the methods used, and the variables selected for the model
Ongoing monitoring	Confirmation that the model is appropriately implemented, being used as designed, and is performing as intended
Outcomes analysis	Comparison of model outputs to corresponding actual outcomes

Exhibit C: Model risk management, supervision and regulation framework

The Teradata Signal Framework provides a robust template for achieving analytical maturity, with the focus on the application of models in downstream workflows to generate economic value at the customer-facing "point of decision." With the increased regulatory focus on bank's model development and deployment, Teradata has developed a comprehensive infrastructure to provide complete support for model risk management, as well.

Signal Framework	Model risk management
Data	Feature quality
Features	Feature quality
Model	Conceptual soundness
Signal	Ongoing monitoring
Workflow	Ongoing monitoring
Applications	Outcomes analysis

Exhibit D: Teradata Signal Framework facilitates model risk management

Each of the critical elements of model risk management are well supported by Teradata. **ClearScape Analytics™** provides the in-database functions for the feature quality dimension of model risk management, encompassing the provenance and transformation of data and diagnostics of feature quality that requires constant oversight. ClearScape Analytics also includes the Teradata Feature Store, or the standardized transformation of enterprise data into variables, or features, suitable for use in a model. A feature store allows the separation of data engineering from data science and provides consistency between model training and scoring by standardizing, documenting, and cataloging model inputs. This is discussed at length in section V: Monitoring feature quality, below.

ClearScape Analytics supports all the model classification diagnostics related to model validation and enables an important capability for outcomes analysis. This is discussed at length in section VI: Measuring model performance. **Teradata ModelOps** also facilitates the automation of model development, supports "lifecycle management" of models in productivity, and facilitates the automation of model performance monitoring, which is important for ongoing monitoring. Developing the "model factory" is a necessary, but not sufficient condition for model risk management. This is discussed in section VII: Model operations.

ClearScape Analytics supports all the model classification diagnostics related to model validation and enables an important capability for outcomes analysis.

Signal Framework		Model risk management	
Data	Raw operational and administrative data: structured, semi-structured, or unstructured	Model risk management	Integrity and validity of the process to transform data into the highly curated, documented, and catalogued feature store used for model development
Features	Transformation of data into standardized, harmonized, and conformed attributes		
Model	Quantitative method to process input data into estimates of outcomes	Conceptual soundness	Quality of the model design and construction, the methods used, and the variables selected for the model
Signal	Actionable insight generated by a unique pattern in the data as identified by the model	Ongoing monitoring	Confirmation that the model is appropriately implemented, being used as designed, and is performing as intended
Workflow	Standardized business process to achieve a business outcome by leveraging the signal		
Applications	Configuration of software tools that enable delivery and monitoring of workflows	Outcomes analysis	Comparison of model outputs to corresponding actual outcomes

Exhibit E: Teradata Signal Framework facilitates model risk management

Legal framework

The most prevalent usage of models in banking is related to the provision of credit, while the earliest legal and regulatory activities related to models, as we know them today, were in the domain of consumer credit decisions. After World War II, applications for consumer loans were evaluated manually by credit underwriters using paper scorecards, which assigned points to information provided on the consumer's application, including a character assessment by the underwriter. Banks also leveraged credit bureaus that compiled records of consumer creditworthiness which banks contributed to and drew upon. With the goal of reducing costs and improving the validity of credit decisions, banks increasingly relied upon statistical algorithms to evaluate consumer loan applications, using both application and credit bureau data.

During the Civil Rights era of the 1960s, the US Congress passed the Equal Credit Opportunity Act incorporating the "fair lending doctrine," which asserts that no creditor should discriminate against a borrower in a credit transaction on a "prohibited basis." The interpretation of the law has evolved to include discrimination against members of the "protected class" based on race, color, ethnicity, gender, age, marital status, country of origin, disability or military status.

Fair lending violations generally fall into three categories: overt discrimination, disparate treatment, and disparate impact.

Fair lending violations generally fall into three categories:

- **Overt discrimination**, or when lenders openly discriminate against borrowers based on their belonging to a protected class.
- **Disparate treatment,** or when individuals are treated differently based on a prohibited basis; for example, borrowers may receive credit, but at terms that are disadvantageous.
- **Disparate impact,** which arises when a lender applies a neutral policy to all credit applicants, but that policy excludes or burdens certain protected classes. The burden of proof is on the lender to show that "there is no alternative policy or practice that could serve the same purpose without a discriminatory effect."[1]

The next evolution of regulatory policy regarding model risk management grew out of the banking capital adequacy crises of the Great Recession. Numerous bank failures and consumer skepticism regarding the soundness of many large banks compelled regulators to act. The Comprehensive Capital Analysis and Review or CCAR was launched in 2011 and placed emphasis on the capital planning process and the robustness of the process employed by participating bank holding companies (BHCs) in their internal risk assessment. Congress passed the Dodd-Frank Act in July 2010 requiring an annual regulatory "stress test" of banks' capital adequacy.[2] CCAR introduced the use of supervisory models based on detailed bank data, expanding capital adequacy calculations beyond balance sheet exposure to consider economic factors such as unemployment, interest rates, and home prices via a formal capital adequacy "stress test" exercise. Over the decade following the OCC's seminal "Model Validation Guidance" in 2000, supervisory attention was broadened from model validation to a more general concept of model risk management.[3] That evolution was reflected in a joint Supervisory Guidance issued by the OCC and the Board of Governors of the Federal Reserve System called "Supervision and Regulation 11-7 Guidance on Model Risk Management" issued April 4, 2011, now known simply as "SR 11-7."

[1] "Fair Lending in the Digital Age", Grant Thornton, June 9, 2023.

[2] Paul Glasserman and Gowtham Tangirala, "Are the Federal Reserve's Stress Test Results Predictable?", March 2015.

[3] Jeffrey A. Brown, Brad McGourty, Til Schuermann, "Model Risk and the Great Financial Crisis: The Rise of Modern Model Risk Management", 2015.

Banking regulations

Regulation SR 11-7 indicates that a "model refers to a quantitative method, system, or approach that applies statistical, economic, financial or mathematical techniques and assumptions to process input data into quantitative estimates." Obviously, this definition is quite broad and has had a profound impact on how banks approach which algorithms to include in model risk management, with a bias toward being overly inclusive.

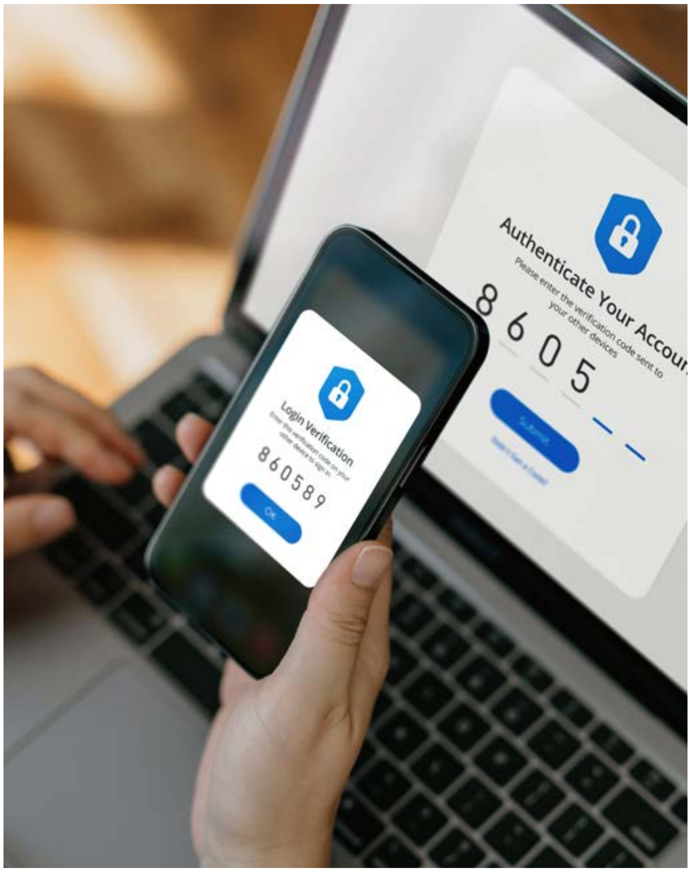
Moreover, SR 11-7 defines model risk as the "potential for adverse consequences from decisions based on incorrect or misused model outputs and reports." This definition of model risk emphasizes the application of the models in decision-making, where the potential for harm is realized.

Also telling is the emphasis on adversarial approaches to model risk mitigation in the regulation, characterized by the statement that an overarching principle is that managing model risk involves the "effective challenge of models" within the bank itself.

A very important policy delivered in SR 11-7 is that model risk governance should be "provided at the highest level by the board of directors and senior management when they establish an organization-wide approach to model risk management." By explicitly identifying model risk management as a board-level conversation, most banks organized this function under the chief auditor, who also reports directly to the board and not to the chief executive officer.

SR 11-7 places particular emphasis on model validation, or the "set of processes and activities intended to verify that models are performing as expected, in line with their design objectives and business uses." Simply put, it's critical to clearly define the business objectives of the model prior to any development efforts.

SR 11-7 indicates that "all model components—inputs, processing, outputs, and reports—should be subject to validation. This applies both to models developed in-house and those purchased or developed by vendors or consultants." This statement emphasizes that the bank leadership is always responsible for any model deployed, even if it was developed by a third party. This provision of SR 11-7 poses a major challenge for banks who now must convince third-party model vendors to provide sufficient details about the model to support model validation without straying into the arena of intellectual property rights.



Another notable dimension of model validation emphasized in SR 11-7 is that the "validation should be done by staff who aren't responsible for model development and use" on an annual cadence, at a minimum. This approach is in the spirit of separation of duties that pervades the approach to minimizing operational risk in banking by ensuring that oversight is truly independent of the line of business. Model validation is not a one-time process, and SR 11-7 emphasizes that validation should be conducted "on an ongoing basis after a model goes into use to track known model limitations and to identify new ones." This is consistent with the data quality doctrine of "validation in use" where problems in design or scope emerge only when the model is broadly deployed, and edge cases rise to the surface.

Within SR 11-7, model validation has three key pillars:

1. **Evaluation of conceptual soundness**, or assessing the quality of the model design and construction, as well as review of documentation and empirical evidence supporting the methods used and variables selected for the model.
2. **Ongoing monitoring**, or confirming that the model is appropriately implemented and is being used and performing as intended.
3. **Outcomes analysis** or comparing model outputs to corresponding actual outcomes.

Each of these model validation pillars provides unique challenges for banks to address in practice. The evaluation of conceptual soundness requires all parties to have a theoretical understanding of the benefits and limitations of the modeling technique deployed. This often leads banks into a natural conservatism where models with a foundation in classic statistics are more familiar while models based on neural architectures face are subject to greater scrutiny and can be actively resisted by compliance.

Ongoing monitoring requires a regular cadence of meetings involving the model developers (who often sit in a business unit) and model validators (who are generally aligned with a centralized audit function). These meetings are often plagued by a misalignment of goals where model developers have a bias toward action and validators embody a natural conservatism.

Outcomes analysis often involves the development of a comprehensive performance reporting infrastructure that's resource intensive and challenging to implement, as the data and features used to develop models can change abruptly as upstream business and data processing procedures change. While evaluating the performance of a model by analyzing model prediction error against historical data, or "back testing," is essential, it's often a poor indicator of model performance after deployment as both the input data and business conditions can change rapidly, as was demonstrated during the Pandemic.

Unfortunately, despite the adoption of SR-17, poor management practices will lead to problems. The most recent example is the failure of Silicon Valley Bank (SVB) in California. When interest rates spiked in early 2023, the value of SVB's investment securities declined rapidly, and to maintain solvency, the bank was forced to sell assets at a loss. When SVB announced a public offering to raise equity capital, the notification precipitated a spate of large dollar online withdrawals from Silicon Valley companies, a classic illustration of "risk contagion." "Clearly, the bank's risk modeling didn't anticipate the combination of interest rate and liquidity risk shocks it would face. While SVB maintained in regulatory findings that it conducted regular market risk analysis and interest rate risk hedging, it's apparent in hindsight that SVB's risk management practices were deficient. SVB was without a senior risk officer for about eight months in 2022 and none of members of the bank's risk committee on the board of directors had any background in risk management."[4]

In the United Kingdom, the Prudential Regulatory Authority (PRA) supervisory statement (issued May 2023 and coming into effect in May 2024) sets out the PRA's expectations for bank's model risk management. The UK approach to model governance seems based on and is largely consistent with Fed SR 11-7.

[4] Clifford Rossi, "Silicon Valley Bank: A Failure in Risk Management", March 2023.

Monitoring feature quality

While a lot of attention and excitement is focused on the modeling technique and related performance diagnostics, more attention should be paid to the health of the data, transformed into features, that is used to train and validate the model. Banks often struggle with managing the data "distillation process." Akin to petroleum refining, where crude petroleum is subject to a complex and time-consuming process to yield aviation fuel, banks' raw operational data is processed into the highly curated, documented, and catalogued "feature store" used for model development.

Model risk management	
Feature quality	Integrity and validity of the process to transform data into the highly curated, documented, and catalogued feature store used for model development
Conceptual soundness	Quality of the model design and construction, the methods used, and the variables selected for the model
Ongoing monitoring	Confirmation that the model is appropriately implemented, being used as designed, and is performing as intended
Outcomes analysis	Comparison of model outputs to corresponding actual outcomes

Model risk management encompasses the provenance and transformation of data and diagnostics of feature quality requiring constant oversight, including:

- **Missing values,** when a feature input is null or unavailable at the time of inference. Even when missing values are allowed in features, a model can see a lot more missing values than in the training set.
- **Range violation** is a feature input either out of expected bounds or is a known error. This happens when the model input exceeds the expected range of its values. It is quite common for categorical inputs to have typos and cardinality mismatches to cause this problem, i.e., free-form typing for categories and numerical fields like age.
- **Type mismatch happens** when the model expects data of a certain type (e.g., an integer) based on its training data, but is provided data of a different type (e.g., a string) at inference time. While it might seem surprising that type mismatches occur often, it's common for columns to get transposed or for the data schema to change unexpectedly.

To address this concern, there are several excellent data preparation utilities to handle outliers and missing values available in ClearScape Analytics. Monitoring data drift is an important consideration for evaluating the quality and integrity of the features used for model development and deployment, of which there are three dimensions:

- **Concept drift**, or a fundamental change in the underlying relationships between features and outcomes. For example, a housing price recession increases the riskiness of loan applications even though the applicant's income and credit scores haven't changed.
- **Feature drift,** or a change in the distribution of a model's inputs. For example, a sudden increase in loan applications from people who are much younger than expected.
- **Label drift**, or a change in a model's output distribution. For example, in "human in the loop" labelling where a new pool of human reviewers applies a different criterion to labeling outcomes.

There are two popular methodologies used in model performance monitoring to detect data drift:

- **Population Stability Index (PSI)** is a counting-based method of calculating drift. It divides the distribution into bins and counts the number of expected versus actual inputs in those bins. In that way, it estimates two distribution curves and how they differ. PSI is a number that ranges from zero to infinity and has a value of zero when the two distributions exactly match.
- **Jensen-Shannon Divergence (JSD)** is a popular drift metric for machine learning models and measures the statistical difference between two probability distributions, with a focus on measuring asymmetry.

Two popular methodologies are used in model performance monitoring to detect data drift: Population Stability Index (PSI) and Jensen-Shannon Divergence (JSD).

Measuring model performance

A critical component of model development prior to any deployment is to ensure the predictive validity of the model by instituting a set of comprehensive diagnostic tests. As this is critically important for model risk management, a review of these diagnostics is warranted.

Model risk management	
Feature quality	Integrity and validity of the process to transform data into the highly curated, documented, and catalogued feature store used for model development
Conceptual soundness	Quality of the model design and construction, the methods used, and the variables selected for the model
Ongoing monitoring	Confirmation that the model is appropriately implemented, being used as designed, and is performing as intended
Outcomes analysis	Comparison of model outputs to corresponding actual outcomes

Classification is a type of machine learning algorithm where the goal is to predict a categorical variable or class label based on a set of input features. The algorithm learns to classify new observations by training on a labelled dataset, where the class labels are already known. The most common type of classification is logistic regression where the algorithm models the probability of an event taking place. One crucial aspect of classification is selecting the appropriate features.

Too many features can lead to overfitting, where the model performs well on the training data but poorly on the test data. On the other hand, too few features can lead to underfitting, where the model fails to capture the underlying patterns in the data. In classification problems, a confusion matrix is used to visualize the performance of a classifier. The confusion matrix contains predicted labels represented across the columns with actual labels represented across the rows. Each cell in the confusion matrix corresponds to the count of occurrences of labels in the test data. The function works for multi-class scenarios, as well.

		Model Predicted Value	
		0	1
Actual	0	True Negative (TN)	False Positive (FP)
Value	1	False Negative (FN)	True Positive (TP)

Exhibit F: Confusion matrix

True positive rate (TPR), also called sensitivity or recall, is the fraction of positives that the model classified correctly, out of all positives and is calculated as TP/(TP+FN).

		Model Predicted Value	
		0	1
Actual	0	True Negative (TN)	False Positive (FP)
Value	1	False Negative (FN)	True Positive (TP)

Exhibit G: Recall

The false positive rate (FPR) shows how often the model classifies something as positive when it's actually negative and is calculated as FP/(FP+TN). Precision measures how many inputs the model classified as positive that were, in fact, positive and is calculated as TP/(TP+FP).

		Model Predicted Value	
		0	1
Actual	0	True Negative (TN)	False Positive (FP)
Value	1	False Negative (FN)	True Positive (TP)

Exhibit H: Precision

Precision is often used in conjunction with recall. If you increase recall, you're likely to decrease precision because you'll make the model less "choosy" and, in turn, this will increase the false positive rate. Accuracy captures the fraction of predictions that were correct and is calculated as (TP+TN)/(TP+FP+TN+FN).

		Model Predicted Value	
		0	1
Actual	0	True Negative (TN)	False Positive (FP)
Value	1	False Negative (FN)	True Positive (TP)

Exhibit I: Accuracy

F1 score is the harmonic mean of precision and recall and is commonly used as an overall metric for model quality, since it captures both the desire to have good coverage of the target outcome while also minimizing prediction errors.

Fortunately, the "Classification Evaluator" function available in ClearScape Analytics supports all the model classification diagnostics discussed above and enables an important capability for model validation. Moreover, the proximity of the ClearScape diagnostics to the data is much more efficient than exporting the data to another application, such as SAS, for these diagnostics.

A receiver operator curve (ROC) shows how much a model can distinguish between classes. It is a graph that shows the performance of a classification model at various classification thresholds, ranging from 0 to 1. The ROC curve shows the tradeoff between sensitivity (or TPR) and specificity (1 – FPR).

Typically, a lower decision threshold identifies more positive cases, because you set a lower bar to classify an observation as positive. However, as you classify more observations as positive due to lenient threshold, you might misclassify more negative cases as positive as well. A better classifier makes fewer tradeoffs to catch more of both classes correctly. A ROC plot illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The starting value for the decision threshold is often the incidence in the population of the outcome that the model is trying to predict.

AUC stands for "area under the ROC curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve. AUC provides an aggregate measure of performance across classification thresholds. An AUC of 1 indicates a perfect classifier, an AUC of 0 indicates a classifier that always predicts the opposite of the actual class, while an AUC of 0.5 indicates a classifier that performs as good as random guessing.

Critically, the "Receiver Operating Characteristic" function is supported in ClearScape Analytics, along with many related diagnostics. The use of ClearScape in-database analytics radically accelerates the model development and validation process by "bringing analytics to the data."Regression is a type of machine learning algorithm that aims to establish a relationship between dependent and independent variables. The most common type of regression is linear regression, where the relationship between variables is assumed to be linear. In the regression process, a model is trained on a dataset consisting of input variables and corresponding output variables. The model tries to find the best fit line or curve that passes through the data points minimizing the difference between the actual and predicted values. Use most-frequently-used metrics to evaluate the performance of a regression model such as:

- **Mean squared error (MSE),** which is a measure of the average amount that the model deviates from the observed data; the ideal but unrealistic value is 0.
- **R-squared**, or the coefficient of determination, which measures how well a regression fits the real data; an R-squared equal to 1 indicates a perfect fit.

These metrics measure the accuracy of the model's predictions by comparing the predicted values with the actual values. Selecting the appropriate features or independent variables is one crucial aspect of regression. Too many features can lead to overfitting, where the model performs well on the training data but poorly on the test data. Alternatively, too few features can lead to underfitting, where the model fails to capture the underlying patterns in the data.The "TD Regression Evaluator" function, available in ClearScape Analytics, computes metrics to evaluate and compare multiple models and summarizes how close predictions are to their expected values. It takes the actual and predicted values of the dependent variables to calculate specified metrics, and the analyst can choose which metrics they want to calculate from a list of supported metrics.[5]

Model risk management	
Feature quality	Integrity and validity of the process to transform data into the highly curated, documented, and catalogued feature store used for model development
Conceptual soundness	Quality of the model design and construction, the methods used, and the variables selected for the model
Ongoing monitoring	Confirmation that the model is appropriately implemented, being used as designed, and is performing as intended
Outcomes analysis	Comparison of model outputs to corresponding actual outcomes

[5] This section is based on Fiddler.ai, "Model Performance Best Practices", 2022.

Model operations

Banks struggle to cross the wide chasm from developing a model to putting the model into production as there are numerous subprocesses that need to work in sequence, a process known as model operations, or “model ops” for short. Challenges with model deployment can be successfully addressed by the utilization of model ops on ClearScape Analytics, including the feature dtore, or the standardized transformation of enterprise data into variables, or features, suitable for use in a model.

A feature store allows the separation of data engineering from data science and provides consistency between model training and scoring by standardizing, documenting, and cataloging model inputs. Model ops also facilitates the automation of model development, supports “lifecycle management” of models in productivity, and facilitates the automation of model performance monitoring. Developing the “model factory” is a necessary, but not sufficient condition for model risk management.

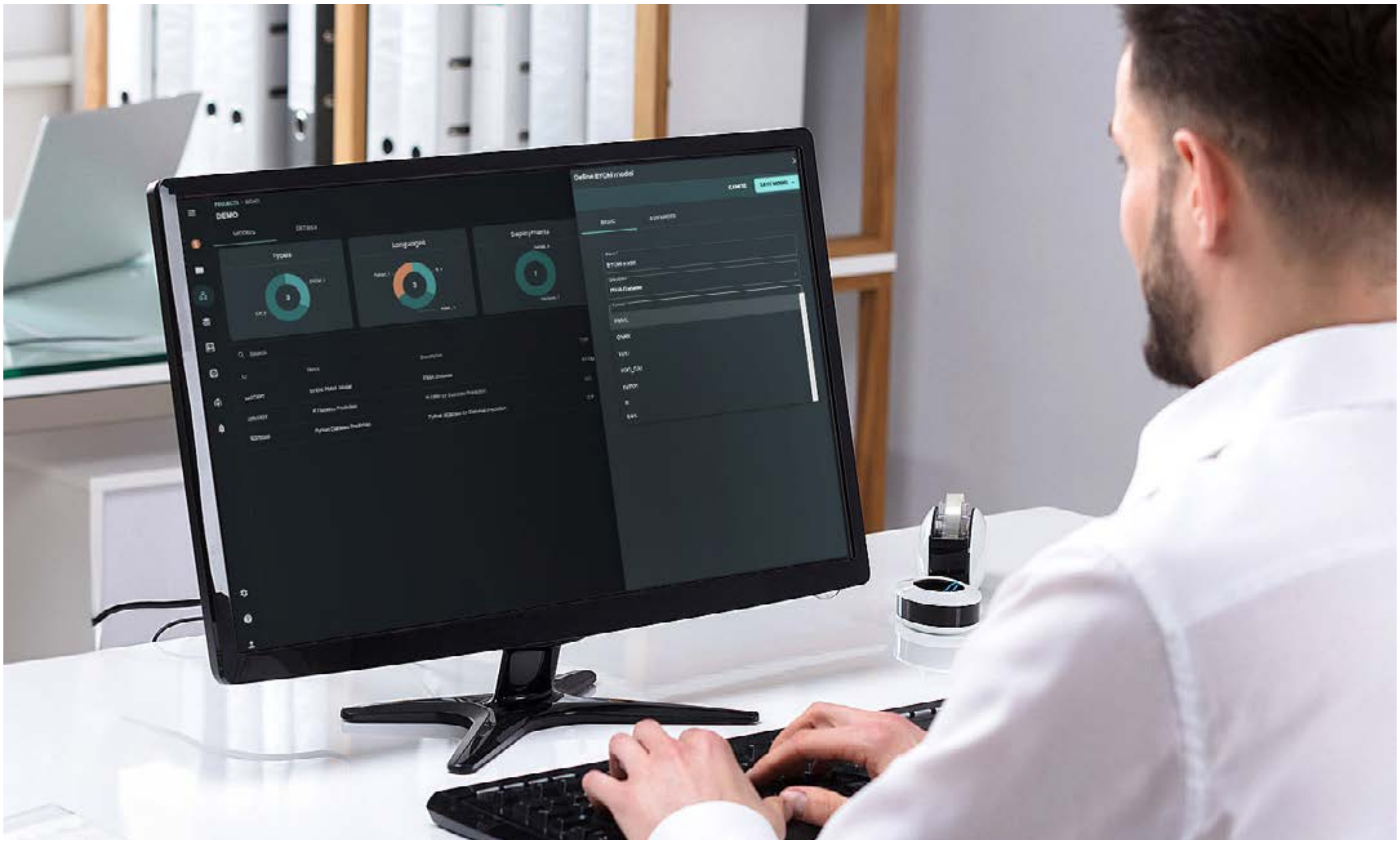


Exhibit J: Teradata ModelOps dashboard

Artificial intelligence models require new techniques

Model risk management for models based on neural architectures is very challenging, driven by both the scale and complexity of these models and the related challenges of transparency and explainability. FinRegLab is an independent, nonprofit organization based in Washington DC that conducts research to "drive the financial sector toward a responsible and inclusive marketplace." FinRegLab is focused on researching the application of artificial intelligence (AI) in financial services, recognizing that while these "techniques may enhance the accuracy and speed of models used to identify potential customers and assess their risks, they also carry significant risks of enhancing bias, eroding data privacy, and obscuring oversight of model's behavior."[6]

FinRegLab sponsored a paper[7] from the Stanford Graduate School of Business that "evaluates model diagnostic tools offered by seven technology companies, as well as several open-source tools applied to prevailing credit underwriting model techniques, including logistic regression, XG Boost (a decision tree model), and a neural network model." The paper is an outstanding introduction to the practical realities of model diagnostics in banking.

The open-source model diagnostic tools evaluated in the paper include:

- **Local Interpretable Model Agnostic Explanations (LIME),** which uses local linear surrogate models around a particular data point to approximate the model's output. The resulting local surrogate models are used to both explain the model's behavior around individual data points and to quantify the feature importance for the overall model.
- **Shapley Additive Explanations (SHAP),** which was developed to assess the unique contribution of players in a cooperative game and is a natural way to compute which features in a model contribute to a predicted outcome. Like LIME, SHAP explains how a model behaves locally. SHAP measures feature importance by conditionally averaging over features from a data point and quantifying how much the removed features impact the model output.
- **Permutation Importance (PERMUTATION),** which measures how important a feature is to a model by calculating how the feature impacts the model's accuracy. Permutation Importance values are calculated by randomly shuffling, or permuting, the values of the feature in the test dataset so that every data point has a new value for the feature and that value comes from a different data point.

The Stanford paper sponsored by FinRegLab found that "there are diagnostic tools which can help lenders address transparency challenges associated with machine learning underwriting models. However, there was no single tool that performed best across all regulatory requirements."

[6] FinRegLab website, 2023.
[7] Laura Blattner, Jann Spiess, "Machine Learning Explainability and Fairness: Insights from Consumer Lending", FinRegLab, 2022.

Model risk management banking practice

A thoughtful point of view on the benefits and realities of model risk management in banking is articulated by Agus Sudjianto, the former chief model risk officer at Wells Fargo and, prior to that, Lloyds Bank and Bank of America. Dr. Sudjianto has numerous observations regarding best practices in model risk management at banks, which he has discussed at numerous conferences and podcasts.

In his opinion, "the obsession with model performance, as measured by the error of prediction, is misguided" and the evaluation mindset should focus on the impact of model failure because "when models are wrong, they create wrongs." Model risk management is fundamentally about model design, model testing, model validation, and model use.

Objective and performance

Given George Box's aphorism that "all models are wrong," a critical management activity is to identify the model's objective and the threshold tolerance for error. For Dr. Sudjianto, the goal should be to start with the threshold acceptance criteria as the guiding principle in model design, validation, and testing. No model is likely to perform well under extreme conditions that the model was not designed to tolerate. Drawing on his experience in mechanical engineering, Dr. Sudjianto describes how even the best-designed truck engine will not perform in extreme conditions like sub-zero air temperatures in the Arctic or a dust storm in the desert.

In his view, model validation and testing are primarily about what he calls "model hacking," or expending effort to identifying the conditions where the model will fail, which can often be unforeseen. It is important not to be consumed with traditional statistical methodologies, or what he described as formally testing a hypothesis against the assumption of a "magical functional form that generates the data." Testing is essential to managing operational risk, as "nobody deploys software without testing!"

Dr. Sudjianto emphasizes the durability and continued relevance of the key fundamental tenants of model risk management that are identified in the Fed's SR 11-7: evaluation of the model's conceptual soundness, analysis of the actual outcomes versus what the model predicted, and ongoing monitoring of the model suitability and performance relative to expectations.

Model governance

To routinize these fundamental tenants of model risk management, Dr. Sudjianto developed a wide-ranging operational control process to ensure that risk is minimized by comprehensive governance of model development and deployment.

The first step in the control process is to create a bank-wide inventory of all models, using the broadest interpretation of "model" from SR 11-7. Next, each model is rated on its inherent level of risk, both the risk of material negative outcomes should the model fail and the inherent complexity of the model itself. The highest risk-rated models are the first to be subjected to model validation.

Extensive documentation for each model is required as a prerequisite of model validation, which identifies the source data, its transformation into features, a discussion of potential alternative model specification, and the theoretical justification for the model form finally selected. The nominated model is then back-tested on historical data and extensive model performance diagnostics are required, such as the aforementioned confusion matrix, AUC, as well as some type of sensitivity analysis.

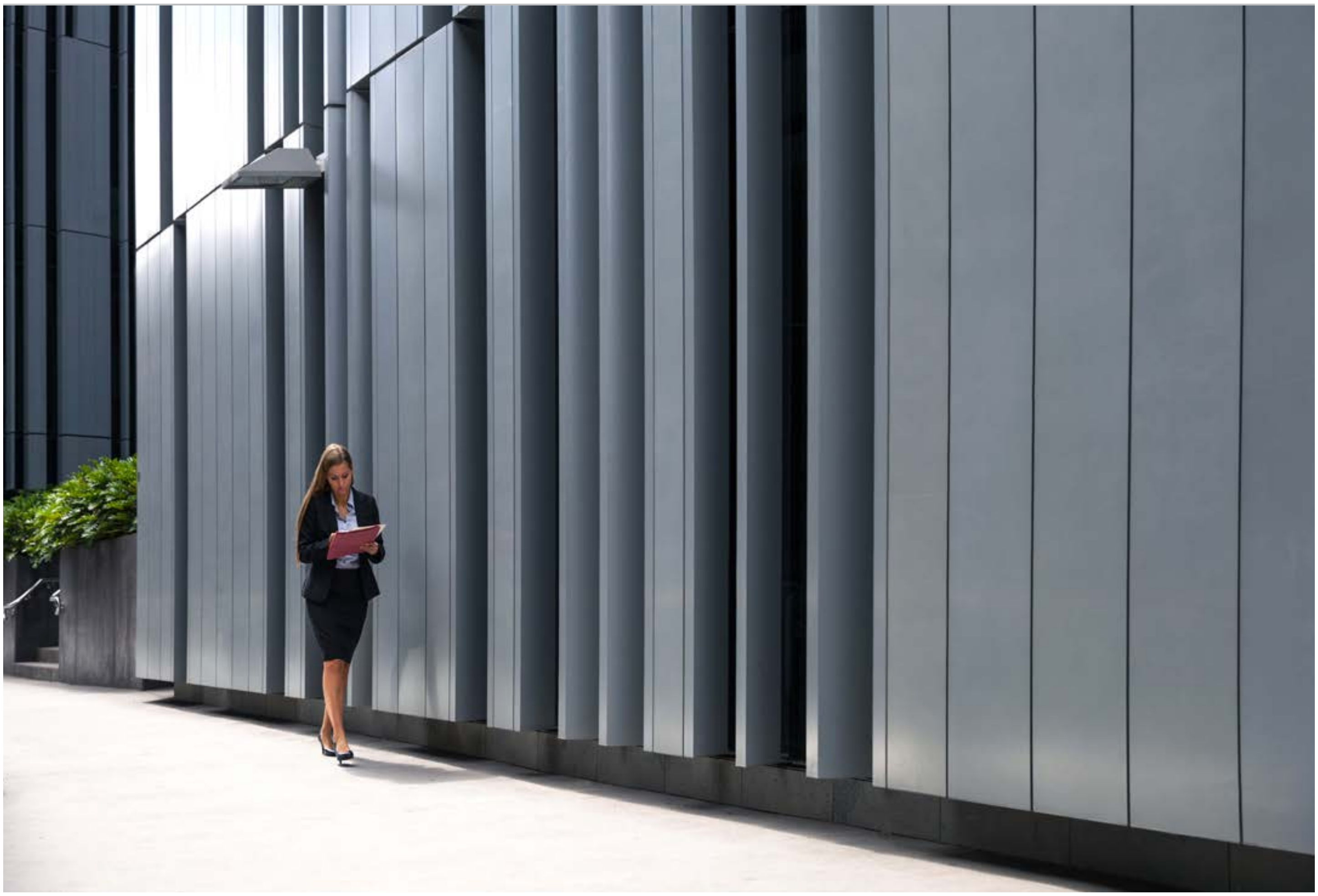
This packet of model documentation and model performance results during development is submitted to a separate and autonomous model validation team, who then challenges the model by probing for problems using the fundamental tenants of model risk management: conceptual soundness, prediction accuracy and robustness, and evaluation over multiple scenarios.

Model governance for artificial intelligence

Dr. Sudjianto has articulated how the conceptual framework of SR 11-7 can be applied to AI models. For the "evaluation of conceptual soundness," it's important to understand the mechanism of how the algorithm transforms inputs into outputs. Special attention must be paid to the convergence algorithm and criteria, most notably the problem of "benign overfitting" that can result from the "gradient descent" convergence optimization technique. For conceptual soundness, it's also important to pay close attention to input control, the constraints on inputs and outputs, as well as model variable selection, most notably the plausibility of causality.

For "outcomes analysis," it's critical to identify the weakness in the model by interrogating the robustness against "noise" such as perturbation of the input variables and the certainty and reliability of the corresponding output. In this case, perturbation reveals the model's robustness.

Here again, Dr. Sudjianto emphasizes the importance of establishing model performance acceptance criteria beforehand. Finally, Generative AI models face a unique challenge for model risk management as banks need to be rigorously evaluated for adherence to the "fair lending doctrine" and to be vigilant to guard against "Large Language Model toxicity."



Teradata Trusted AI

For Teradata, Trusted AI involves the creation of a secure, ethical and governed framework that encompasses the data, features and models that lead to the generation of reliable signals. The concept of signal is critical; Teradata facilitates and democratizes the access by enterprise application workflows of AI-generated signals.

Teradata believes that banks must cultivate an environment where data, AI models, and people work together to create both value and accountability throughout the AI model lifecycle. Keeping people at the center of the deployment of AI technology is critical to ensure data security, environmental stability, and regulatory compliance. AI model deployments must promote explainability and transparency to prevent discrimination against members of a protected class.

The Federal Reserve's "Supervision and Regulation 11-7 Guidance on Model Risk Management" provides banks a durable framework to address model risk with a focus on evaluating the model's conceptual soundness, analysis of the actual outcomes versus what the model predicted, and ongoing monitoring of the model's suitability and performance relative to expectations.

The critical activity to mitigate model risk is a comprehensive approach to model validation, which incorporates a focus on input data quality and model output stability. Teradata supports effective and efficient model risk management with the data preparation utilities and model performance diagnostics available in ClearScape Analytics, attribute curation in the Teradata Feature Store, and model lifecycle management via Teradata ModelOps.

Teradata believes that banks must cultivate an environment where data, AI models, and people work together to create both value and accountability throughout the AI model lifecycle.

How Teradata can help

Teradata partners with businesses in financial services and many other industries to create impactful customer experiences through AI, machine learning, and advanced analytics. Using customer journey analytics, we help banks evaluate and improve their service delivery processes by expediting issue resolution and eliminating channel friction to radically increase customer satisfaction and engagement.

We empower banks' customer journey analytics through:

- **Complete data harmonization:** Integrating data and accelerate data preparation with the most resource-efficient cloud platform and advanced in-database analytics
- **Rapid AI innovation:** Using preferred model training tools and technologies via our open and connected ecosystem
- **The most cost-effective performance:** Operationalizing and scaling Trusted AI through robust governance, automated lifecycle management, and massively parallel processing

Teradata provides the flexible, proven solutions banks need to innovate faster, enrich customer experiences, and deliver value—all with the transparency and security of Trusted AI that banks need. To learn more about how Teradata can empower your customer experience with customer journey analytics, visit our **website** or **talk to an expert** today.

About Teradata

At Teradata, we believe that people thrive when empowered with trusted information. That's why we built the most complete cloud analytics and data platform for AI. By delivering harmonized data, Trusted AI, and faster innovation, we uplift and empower our customers—and our customers' customers—to make better, more confident decisions. The world's top companies across every major industry trust Teradata to improve business performance, enrich customer experiences, and fully integrate data across the enterprise. **Learn more at teradata.com**

17095 Via Del Campo, San Diego, CA 92127   Teradata.com

teradata.