MONITAUR

Machine Learning Assurance

Accelerating innovation in regulated industries with trust and transparency in Machine Learning

Table of Contents

Machine Learning Assurance

Machine Learning (ML) is a form of artificial intelligence whereby computer systems recognize patterns and make predictions or decisions without explicit programming. Innovative ML applications are at work in a variety of industries and use cases. For example, in health care, radiologists use models to examine X-rays and differentiate benign tumors from malignant or metastatic ones.[1,2,3] In finance, classification models determine if a customer is a suitable loan candidate or is at risk of payment default.[4]

ML model decision making is not always transparent or meaningful to humans. This is problematic as ML model decisions routinely affect high-stakes human activities. To advance model use, companies must establish relationships between technical deployments and the audit, compliance, risk management, quality assurance, and ethics functions. This includes providing applicable model controls. Without such controls, ML deployments will stall or deploy irresponsibly. In both cases, significant competitive and compliance risks emerge.

Enter: Machine Learning Assurance.

[1] Implementing Machine Learning in Radiology Practice and Research, Marc Kohli, Luciano M. Prevedello, Ross W. Filice, and J. Raymond Geis, American Journal of Roentgenology 2017 208:4, 754-760

[2] Hosny, A., Parmar, C., Quackenbush, J. et al. Artificial intelligence in radiology. Nat Rev Cancer 18, 500–510 (2018). https://doi.org/10.1038/s41568-018-0016-5

[3] McKinney, S.M., Sieniek, M., Godbole, V. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* 577, 89–94 (2020). https://doi.org/10.1038/s41586-019-1799-6

[4] https://medium.com/@DigiFi/evaluating-the-financial-impact-of-machine-learning-for-loan-underwriting-1015b3f229e5

Machine Learning Assurance (MLA) is the process of building trust in machine learning and its decisions.

This paper introduces:

- A description of the need for MLA
- Dynamics that are driving the adoption of MLA
- Core principles of MLA
- Guidance on processes and related controls for MLA
- An innovative solution to address MLA requirements
- Illustrative industry use cases

Using a robust MLA process and infrastructure, organizations can now confidently leverage the potential of ML within the framework of its proven management practices.

The need for MLA

Machine Learning Assurance has emerged as a sub-specialty within the larger field of Assurance because of the fundamental differences that machine learning applications and deployments impose upon their operators. Thanks to a growing number of high-profile public mishaps that have triggered legal and regulatory action, the problems of the “black box” of Machine Learning (ML) have entered not only the public consciousness but also boardrooms and courtrooms.

Such incidents highlight the need for a more proactive approach from both corporations and regulators with regard to machine learning models. Organizations have a responsibility to their customers and society at large to ensure that their models are safe, fair, and compliant in ways that are demonstrable to regulators. Internally, organizations also need to demonstrate that their machine learning is performant, delivering the desired results for the relevant business objectives.

Unfortunately, traditional approaches to assurance face challenges when mapped to initiatives dependent on machine learning models because of how these models work. Traditional statistical models utilize a small number of known inputs and extrapolate from narrow data sets to create larger inferences. On the other hand, machine learning models incorporate any number of inputs from massive data sets to identify patterns, predict relationships, and improve the model’s algorithm. The non-linear, unpredictable relationships between input and output make assurance more challenging. Simultaneously the ability for the models to evolve and scale creates enormous liabilities and risk for reputational harm.

Regulatory requirements are driving MLA adoption

Regulatory efforts are currently underway to contend with ML's expansion, including a legislative push for increased scrutiny and verification of model predictions.[5] As regulatory pressures mount, companies struggle to provide adequate model assurance.

As lawmakers prioritize consumer protection, the need for ML assurance and governance grows.

Consumer Data Privacy and Transparent Use

Agencies such as the Equal Employment Opportunity Commission (EEOC) and Federal Trade Commission (FTC) have regulations guarding against discrimination within a variety of industries. These, plus data-related regulations such as the European Union's Global Data Protection Regulation (GDPR), complicate fairness and bias practices for companies expanding ML use.

This confusion around how GDPR and AI/ML interact is a focal point of myriad articles and academic opinions. The ambiguity leaves companies scrambling for compliance answers. GDPR is intentionally vague; it gives regulators leeway to probe issues that they sense are unfair or anti-competitive.

Notably, GDPR has provisions on profiling, which is the automated processing of personal data to make evaluations. This is a growing practice in a range of

[5] Burt, Andrew. "How Will the GDPR Impact Machine Learning?" O'Reilly Media. O'Reilly Media, Inc, May 16, 2018. https://www.oreilly.com/radar/how-will-the-gdpr-impact-machine-learning/.

industries, from healthcare to financial services. While automated individual decision-making leads to quicker and more consistent decisions, GDPR requires companies to provide meaningful information about the decision logic to consumers. Companies subject to GDPR not only need to prove assurance and governance, but they need to display well controlled environments and provide documentation to satisfy regulatory questions[6].

Healthcare Sector Regulation

In addition, the U.S. Food and Drug Administration (FDA) recently proposed a framework for AI and ML systems in the medical field called Software as a Medical Device (SaMD). It states that technologies using AI/ML "require a premarket submission to the FDA when the AI/ML software modification significantly affects device performance, or safety and effectiveness; the modification is to the device's intended use; or the modification introduced a major change to the SaMD algorithm..."[7] Essentially, businesses need a documented process for algorithmic changes (ACP).

The FDA also requires that real-world monitoring and a culture of quality control are in place. They propose the concept of "Good Machine Learning Practices" – and hold those applying for FDA approval accountable to have adequate Machine Learning Assurance and Governance practices.

[6] "ICO Consultation on the Draft AI Auditing Framework Guidance for Organisations." ICO. Information Commissioner's Office. Accessed March 5, 2020. https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-consultation-on-the-draft-ai-auditing-framework-guidance-for-organisations/.

[7] Center for Devices and Radiological Health. "Artificial Intelligence and Machine Learning in Software." U.S. Food and Drug Administration. FDA. Accessed February 27, 2020. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device.

Financial Services Sector Regulation

For financial service providers, the Federal Reserve and Office of the Comptroller of the Currency issued regulatory guidance SR 11-7. This requires financial models to meet a high degree of model governance and assurance. Although SR 11-7 is nearly a decade old, financial institutions still wrestle with it.[8] The increased compliance costs and model failures accompanying these monitoring, governance, and change management mandates prove particularly difficult.

Further, regulators demand that assurance levels be commensurate with model complexity. The existing statistical models that comply with SR 11-7 rely on highly manual processing and lack the complexity necessary for satisfactory ML model assurance.

The following table shows the current regulation in terms of how it's being enforced at organizations, the challenges with that enforcement, and how Monitaur can help.

[8] "SR 11-7: Guidance on Model Risk Management." The Fed - Supervisory Letter SR 11-7 on guidance on Model Risk Management -- April 4, 2011. The Federal Reserve, April 4, 2011. https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm.

Principles of Machine Learning Assurance

In order to provide assurance for ML models, companies must establish trust and confidence based on the following principles:

1. Operational System Context
2. Verifiability
3. Objective Third Parties

1. Operational System Context

To fully understand a process, stakeholders need answers to key questions:

- What are the key objectives and business goals?
- What risks or adverse events are possible?
- Where does the data come from?
- How is it transformed?
- Which model was used?
- How was testing conducted?
- How was the model deployed?
- What is the monitoring process?

Documentation from each step (along with accompanying flowcharts) in conjunction with Logging, Verifiability, and Reperformance establishes the process as thoughtful and controlled.

2. Verifiability

Verifiability is the ability to access, examine, and audit decisions within a time and condition specific context.

In financial auditing, there is an oft-repeated phrase: if it isn't documented, it didn't exist. The lack of consistent, comprehensive, readily accessible, and easy-to-understand model inputs and decision logs plagues verifiability for ML implementations. Without detailed and reliable **logging**, Machine Learning Assurance is unachievable.

For example, to verify whether an applicant was correctly assigned a $10,000 credit limit increase, the transaction behind the assignment – even if from months ago – needs to be accurately captured and available for a business review. This allows a "doublecheck" of the decision that resulted in a credit limit bump. This work is particularly essential in regulated industries. Model auditors can verify delivery of the same decision, using the same inputs, assuring the model's decision-making.

Reperforming a given transaction or decision is another critical part of any audit.

For example, in a financial audit, it is common for auditors to recalculate cash inflows and outflows. That being said, replicating a model's decision is complicated because of how the input data is transformed before being fed into the model. Managing past model versions, accommodating environmental shifts and navigating data privacy also complicate reperformance. Yet, model trust requires the capability to rerun, or reperform, a decision. This is a behavioral expectation of risk managers, auditors, and regulators.

3. Objective Third Parties

No matter how well processes are designed and executed, errors may still be introduced by machine learning.

Beyond regulatory requirements, the prevalence of the consulting and auditing sectors attest to the value of an independent, third-party perspective. Maximizing assurance in high-risk, high-reward processes requires objective scrutiny, whether from internal or external auditors, so that practitioners aren’t evaluating their own work with rose-colored lenses. The benefits of objectivity also extend to assurance that legal and ethical standards are being met.

High-risk ML models should only be deployed with objective, third-party assurance. The repercussions of deploying an under-scrutinized system can be fatal, particularly in emerging domains like fully autonomous vehicles, cancer screening, and other applications with life and death consequences.

Putting the Principles into Practice

Naturally, the power of these principles relies on satisfying them in parallel. Objectivity is meaningless without understanding the operational context and the capability to verify the decisions of an ML application. Similarly, adhering to these principles of MLA only deliver value within a strong framework for ML deployment that enable internal and external auditors to practice them.

A Best Practices Methodology

Repurposing or expanding on proven methodologies supports the rigorous application of assurance practices for ML. The Cross Industry Standard Process for Data Mining (**CRISP-DM**) is an important framework in this respect.

Created in the 1990s by a European Union project on data analytics, the framework has become a de facto industry standard for how ML is conducted by practitioners, even if they have not explicitly or intentionally followed the framework. CRISP-DM was specifically tailored to enhance machine learning assurance in 2018.[9]

Its highest-level steps are:

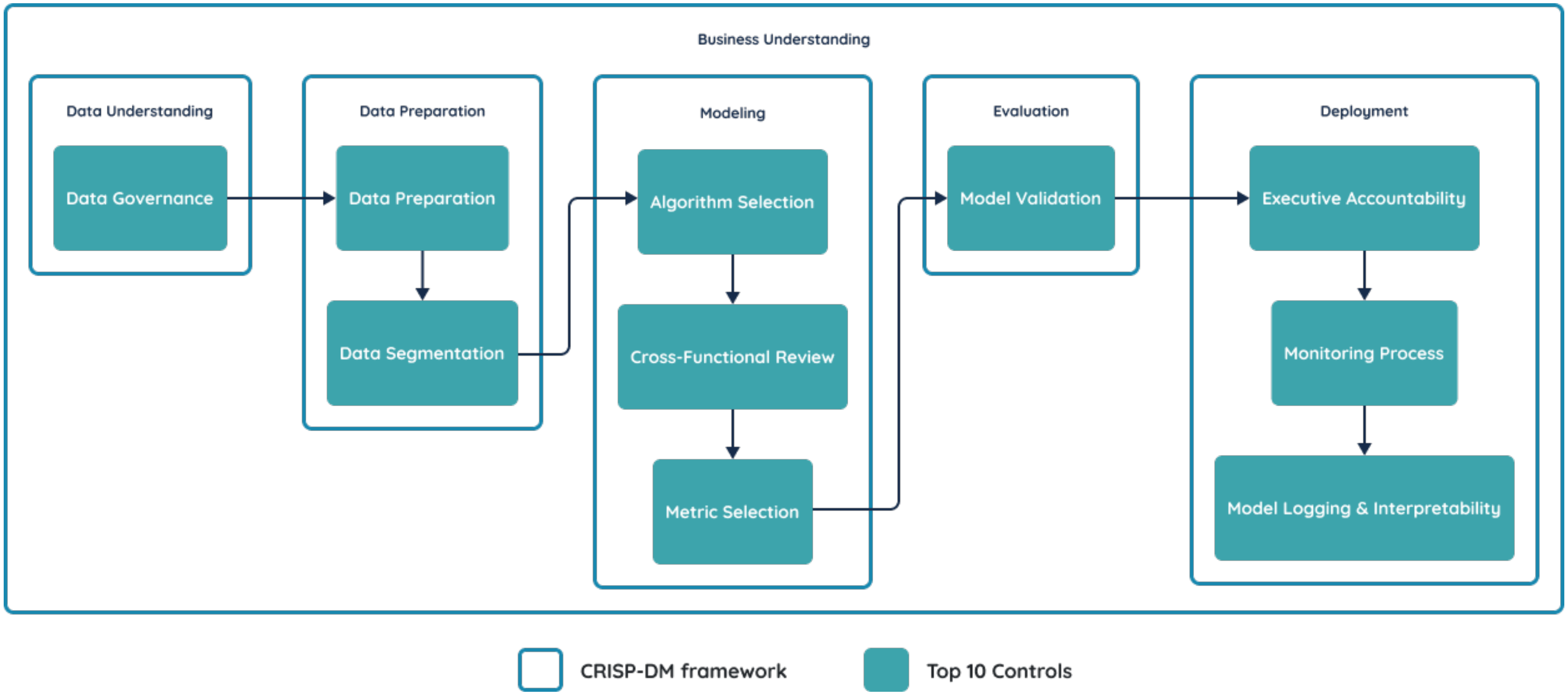
1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

Monitaur advanced this framework by identifying the Top 10 Machine Learning Controls.[10] These controls are more implementation-specific than the CRISP-DM

[9] Andrew Clark, "The Machine Learning Audit—CRISP-DM Framework." ISACA Journal, vol. 1 (2018). https://www.isaca.org/resources/isaca-journal/issues/2018/volume-1/the-machine-learning-auditcrisp-dm-framework.

[10] Clark, Andrew. "Top 10 AI/ML Controls." Monitaur AI, March 18, 2020. https://monitaur.ai/blog/top-controls/.

framework; notably, they can be used in conjunction with the underlying principles of CRISP-DM, as shown in the below diagram.



1. Business Understanding

Before a ML algorithm is created, a Business Understanding of the proposed problem is necessary. In this stage, the project team evaluates key business questions to establish a common understanding of the model's opportunities and challenges. As the CRISP-DM framework is iterative, the business understanding section should be revisited often during a large project.

2. Data Understanding

Every data set has unique characteristics and must be treated specially. Without understanding the nature and idiosyncrasy of the data, an accurate model cannot be constructed. However, there is more to this step than meets the eye. Most data, besides categorical variables, have an inherent scale to them, such as Celsius or Fahrenheit.[11] Different stores of data have different considerations, such as the given schema of a relational database. Without this understanding, strong algorithmic models and subsequent audits cannot be built or accomplished.

Additionally, understanding the usage rights of data is key, as is identifying the GDPR ramifications or license limitations of the data. A thorough documentation of potential risks is also important. If regulations change, companies need to understand their data exposure.

Auditors need vigilance to understand the data variables and ensure that they do not conflict or introduce biases. Correlation and covariance matrices can be

[11] Andrew Clark, "The Machine Learning Audit—CRISP-DM Framework." ISACA Journal, vol. 1 (2018),https://www.isaca.org/resources/isaca-journal/issues/2018/volume-1/the-machine-learning-auditcrisp-dm-framework

examined to understand how the variables correlate and vary in response to one another. Enacting data governance controls provides the foundational level of technical assurance required to take the next step.

3. Data Preparation

Sufficient time should be allocated to data preparation. For relational data, there may not be much “wrangling” required to get the data into an amendable structure. However, with unstructured text, such as log files and web-scraped data, the preprocessing stage may be time consuming. Segregating data into train/test and validation sets prior to preprocessing is a key step. Data segregation ensures that the model is not overfitted to training data or that data transformation leakage occurs.

Standardized, reproducible, and statistically valid data preparation techniques should be applied to ensure human bias is not introduced. Here, use of race, gender, and other sensitive data variables, along with their proxies such as zip code or job title, should be considered to avoid biases or privacy issues.

4. Modeling

While modeling is the most prominent step in the process, in most ML projects, it is one of the shorter steps—at least in the initial implementation. Practitioners should recognize early on that a more complex model is not necessarily best for a given situation. Simpler models should be tested before shifting to more complex techniques since they are easier to maintain and deploy.

Teams must document what algorithm was used and what steps were taken to avoid overfitting or bias. Noting retraining times with model versioning capture is also crucial. Performance may differ after model updates, so these must stay subject to appropriate change management and governance controls.

5. Evaluation

Evaluation is where the accuracy and precision of the model is determined. Here, one must identify the appropriate testing to ensure against adverse outcomes, including unfairness or bias, and gauge the model's sensitivity to various inputs and outputs.

Thorough evaluation should occur prior to, as well as throughout, the model's deployment. The team should create a validation dataset across the range of all possible inputs paired with the predictions made by subject matter experts.

6. Deployment

When a model is ready for deployment, consider if it accomplishes its business purpose. In this stage, one must ensure that deployment controls can provide timely diagnosis of model misbehavior. And model predictions must be logged and monitored with sufficient detail for local interpretability of outcomes. Following these procedures will add necessary visibility and audibility of the model.

During and after deployment, one should also be sensitive to "technical debt." This represents technical compromises or short-cuts introduced when a project progresses under time and resource pressures.

Periodic validation tests and tests for feature drift should occur to ensure that the model performs as expected. Unless a model is self-learning, it will not change when deployed. But data inputs may change, which can degrade model performance.

Across the full lifecycle of ML design, implementation, and deployment, leadership is responsible for model outcomes and erratic performance affecting customers. As with any business process, appropriate risk management procedures should be in place to ensure that business goals are met without negatively disrupting the operations of the corporate entity.

Monitaur: Delivering Continuous Assurance

Monitaur is a Machine Learning Assurance platform addressing the needs of companies using ML models to make decisions in regulated industries.

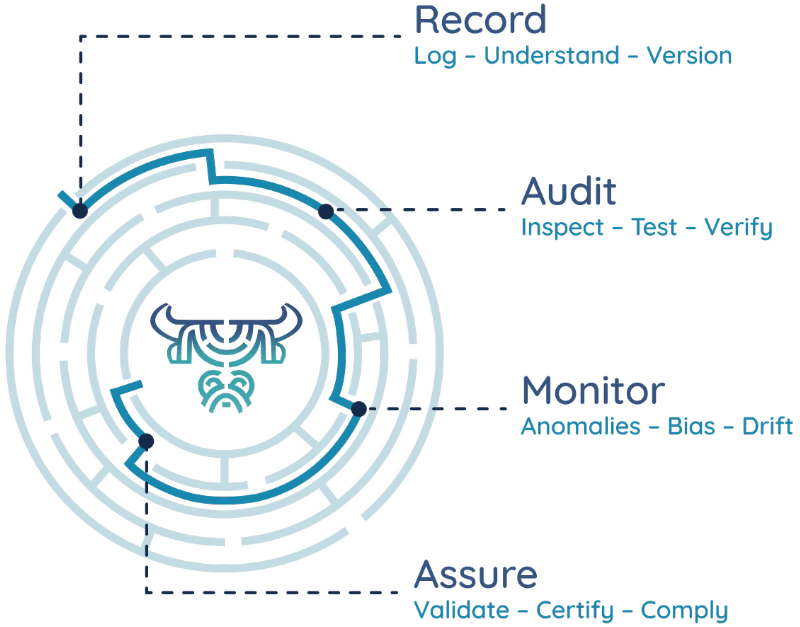
Monitaur delivers the transparency and confidence necessary to manage compliance and unlock innovation, providing an advanced solution for MLA based on flexible and widely supported practices.. Monitaur enables recording, auditing, and monitoring of ML models so that MLA can be achieved. Additionally, Monitaur's technology fully supports the CRISP-DM framework in the Evaluation and Deployment steps so that internal and external teams can determine the Operational Context, Verifiability, and Objectivity that provide thorough Machine Learning Assurance to business owners and regulators alike.

Monitaur is platform and environment agnostic, with planned support for all classical machine learning and deep learning implementations in Python, R, and Java. With full model versioning and transaction reproducibility, Monitaur allows counterfactuals for auditability of ML models[12]. Monitaur can be deployed on-premise, on the Monitaur Cloud, or on a customer's cloud.

[12] Note: A counterfactual is the reperformance of a specific transaction by running the inputs against the exact model version used in the original transaction.

Monitaur's Four Core Components

Five Stages of MLA	Monitaur Products
Operational System Context	Record Audit Monitor Assure
Verifiability	Record Audit
Objective Third Parties	Assure



Record
Log – Understand – Version

Audit
Inspect – Test – Verify

Monitor
Anomalies – Bias – Drift

Assure
Validate – Certify – Comply

Record Product

Log

The Monitaur client library records a ML model's meta information along with related inputs and decisions. Each transaction is recorded as it runs on production infrastructure.

In addition, Monitaur versions models based on changes to the production and trained model file hashes. This enables the solution to reperform and validate results from the past. When environmental changes occur, the platform automatically generates alerts.

Understand

As transactions are sent to a back-end API, Monitaur obtains the underlying influences for each of a model's decisions via the Anchors or GradCAM interpretability frameworks. This explains which inputs are most impactful to a specific decision. Monitaur also enables users to pass in their own interpretability, if preferred.

Audit Product

Verify

The Monitaur GRC (Governance, Risk, and Compliance) web application enables a non-technical user to find individual ML transactions and verify their inputs and decisions for audit and compliance requests.

Inspect

With full model versioning and transaction reproducibility, Monitaur supports counterfactuals for model auditing. Counterfactuals offer the ability to reperform transactions with the same model version and inputs, or even with slight variations, for “what-if” analyses.

Monitor Product

Monitaur allows users to set alerts, rules, and application controls to proactively manage risks, bias, and compliance. It establishes metrics for model and feature drift to detect model degradation. As well, it introduces bias controls for controlled variables such as age, sex, or race.

Assurance Product

Monitaur provides automated audit reports exportable for a third-party audit or for tracking over time. This supports continuous and independent validation of ML compliance. Monitaur also facilitates model white paper creation and workflow process development around proven machine learning governance frameworks.

Monitaur Capabilities

What does Monitaur record?

- **Model inputs:** Every considered input for a specific transaction such as patient age or blood pressure in healthcare or credit score and income in financial services.
- **Model predictions:** “The patient doesn’t have cancer”; “the customer’s credit limit is $5,000.”
- **Model version:** “Model version 1.1 was used.”
- **Meta model information:** Examples: Developer name, business owner name, library or model algorithm used, and model whitepaper.

Benefits from Monitaur’s recordings:

- Verifiability of ML predictions and audit trail creation.
- Transaction and time specific model versioning and changes.
- Transaction interpretability.
- Audibility of models (including the ability to rerun transactions against the specific version of the model used for each transaction).

Conclusion

Assurance is the fundamental objective of compliance and risk management, and the emergence of wide-scale deployments of machine learning places new demands on those professionals who provide businesses with assurance. While assurance often focuses on regulatory and audit compliance, it really does much more than that: it provides the confidence and trust to business leaders to launch powerful new initiatives using ML; and it helps unleash the innovation of data science and engineering teams by freeing them to focus on solutioning instead.

Organizations need to establish assurance for ML with intentionality. With Monitaur's robust MLA platform, organizations can speed innovation responsibly and with confidence, ensuring that the company is safe, fair, and compliant in pursuing all of its goals.

Illustrative Examples

The following are illustrations of applications of Monitaur in particular industries.

HEALTHCARE CASE STUDY

Verifying Decisions for Hospital Readmissions

American hospitals commit significant time and resources towards reducing readmissions. Not only do reduced readmissions signify effective care, but minimizing them drives cost and profit improvements.

Consider an illustrative example of Monitaur platform use:

ACME Hospital was experiencing a high level of readmissions. They contracted data science consultancy Data Wranglers to build a ML model to reduce occurrences.

But teams from ACME and Data Wranglers ran into issues. ACME's compliance department struggled to verify individual patient decisions and raised concerns. Simply: they couldn't comply with ongoing requirements to inspect and audit readmissions data. Further, the compliance team could not ascertain, access, or understand the patient data used in decision making.

Using Monitaur, ACME and Data Wranglers resolved the compliance team's concerns by offering transaction level logging and ongoing model monitoring. Monitaur's transaction recording, model version change detection, prediction understanding, and reperformance capabilities restored ACME leadership's confidence in the model – and transformed patient care innovations throughout the hospital.

INSURANCE CASE STUDY

Explaining Customer Approvals for Life Insurance

Rapid, accurate, and simplified insurance underwriting is a competitive advantage for insurers. Intelligence here creates new product offerings, improves service levels, and reduces costs. However, applying ML technology is challenging in this highly regulated market.

Consider an illustrative example of Monitaur platform use:

Pinnacle Insurance asked its data science team to create a model to make immediate, independent decisions about consumers for a new life insurance product.

The algorithm's goal was to quickly and accurately determine if an individual qualified for an insurance policy. The data science team built a model displaying high accuracy around applicants' risk profiles. **However, the model never advanced beyond the test environment. State insurance regulators expressed concerns about using ML models because of their complexity. The inability to reperform transactions and verify inputs was also an issue.** The state regulator needed a reliable way to access and evaluate the model before it would approve the new offering.

After engaging Monitaur, Pinnacle successfully deployed the model. Transaction level recording, input, output and version verifiability, and reperformance capabilities were what the regulators needed to green-light the product.