Prodapt

Realizing the trustworthiness of AI systems

Implement AI reliability scorecard to accelerate trusted decision-making

Credit
Ishwarya MS | Kavya Rengaraj | Suneel Musunuru | Priyankaa A

Current state: The adoption of Artificial Intelligence (AI) is skyrocketing, but trust issues persist

- One-third of businesses leverage AI across a set of use cases to improve existing processes and open new channels of revenue
- However, lack of trust, transparency, and governance of AI systems are a major impediment to realize its true potential
- Rather than just augmenting human judgment, AI-based systems are now driving high-stake decisions
- Most organizations haven't taken key steps to ensure their AI is trustworthy and responsible, such as reducing bias and explaining AI-powered decisions

Source: Gartner

AI implementations today lack the following

- Mechanisms to arrive at **fair** and **interpretable predictions**
- Techniques to understand the working of complex and opaque ML models
- Methodology to secure model against **adversarial attacks** leading to the leakage of Personally Identifiable Information (PII)

Impact of current AI systems

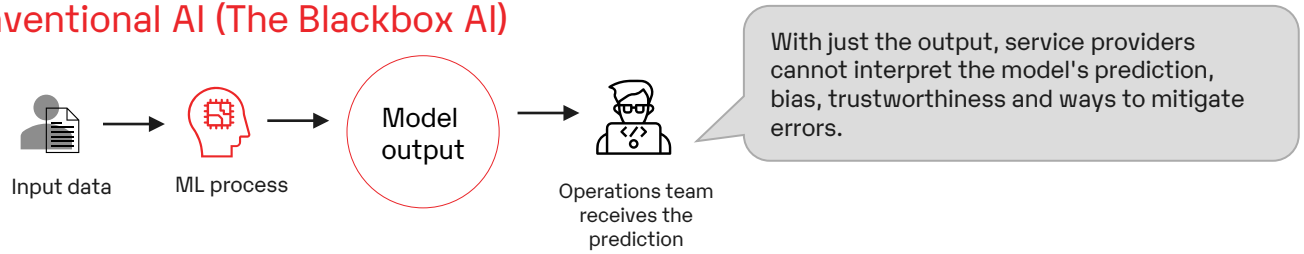
Growing bias in the systems	Legal and compliance fines and penalties	Increased costs due to multiple AI experiments

Responsible AI: The key to achieve a trustworthy AI system

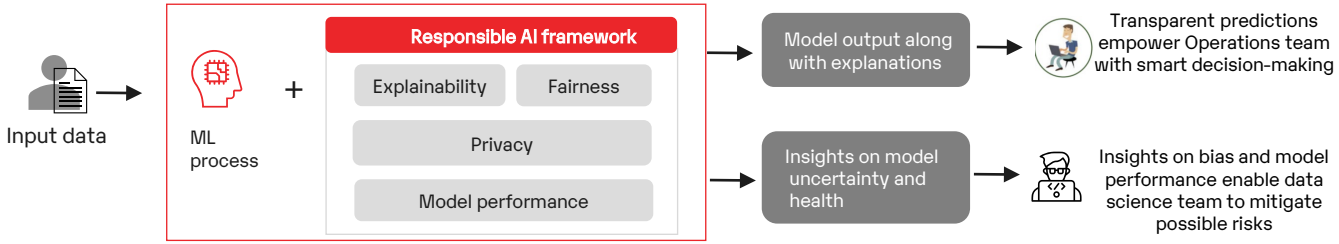
Despite the real value provided by AI systems, businesses struggle to address the risks arising from bias and privacy issues. **Responsible AI** assists service providers with recognizing, preparing, and mitigating the potential effects of AI. It also improves transparent communication, end-user trust, model auditability, and the productive use of AI.

According to Gartner "By 2023, all personnel hired for AI development and training will have to demonstrate an understanding of ethical considerations of AI to ensure **responsible development** of AI".

Conventional AI (The Blackbox AI)



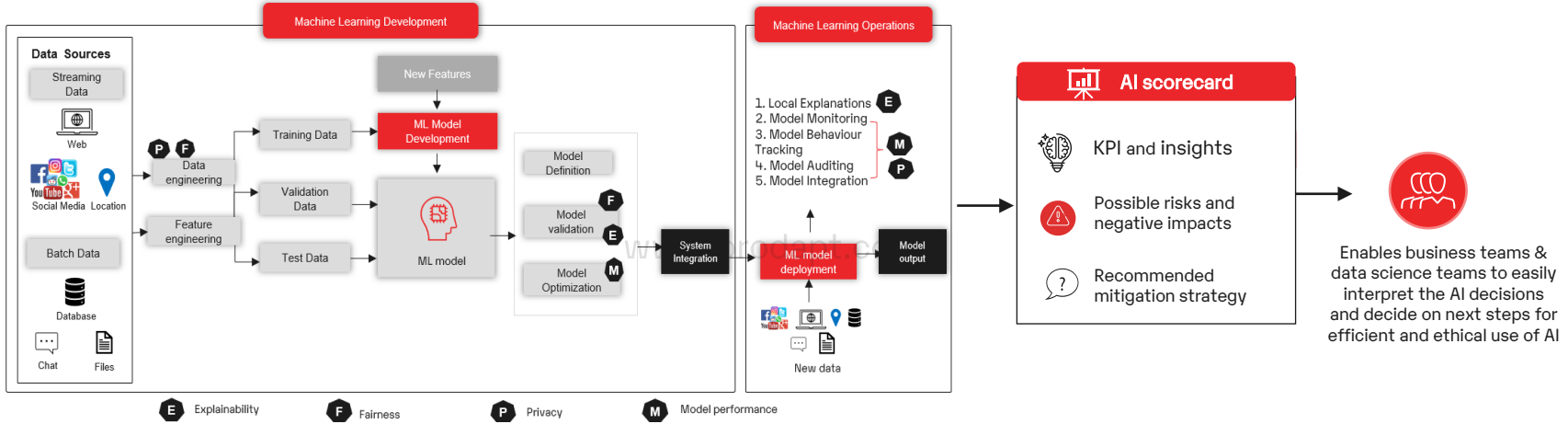
Responsible AI



AI reliability scorecard: An approach to implement Responsible AI

As AI progresses from proof-of-concept to powering business workflows, assessing AI pipelines is becoming increasingly crucial. Implementing AI scorecard assists in evaluating AI development and deployment pipelines along four major axes: **explainability**, **fairness**, **privacy**, and **model performance**.

Sample flow: Incorporating AI scorecard into AI pipeline



An AI scorecard flags out-of-the-bound KPIs when incorporated into the AI pipeline. Business and data science teams can use the AI scorecard to conduct in-process tuning of their algorithms, enabling creation of the right AI system and differentiated offerings.

Four key pillars for successfully implementing Responsible AI

1 Explainability

Gain a complete view of how AI systems make their decisions to improve transparency and trust in decision-making.

2 Fairness

Ensure that the AI system delivers unbiased predictions to all groups and individuals.

3 Privacy

Enable data protection in AI systems to prevent the inadvertent disclosure of sensitive information and system breaches.

4 Model performance

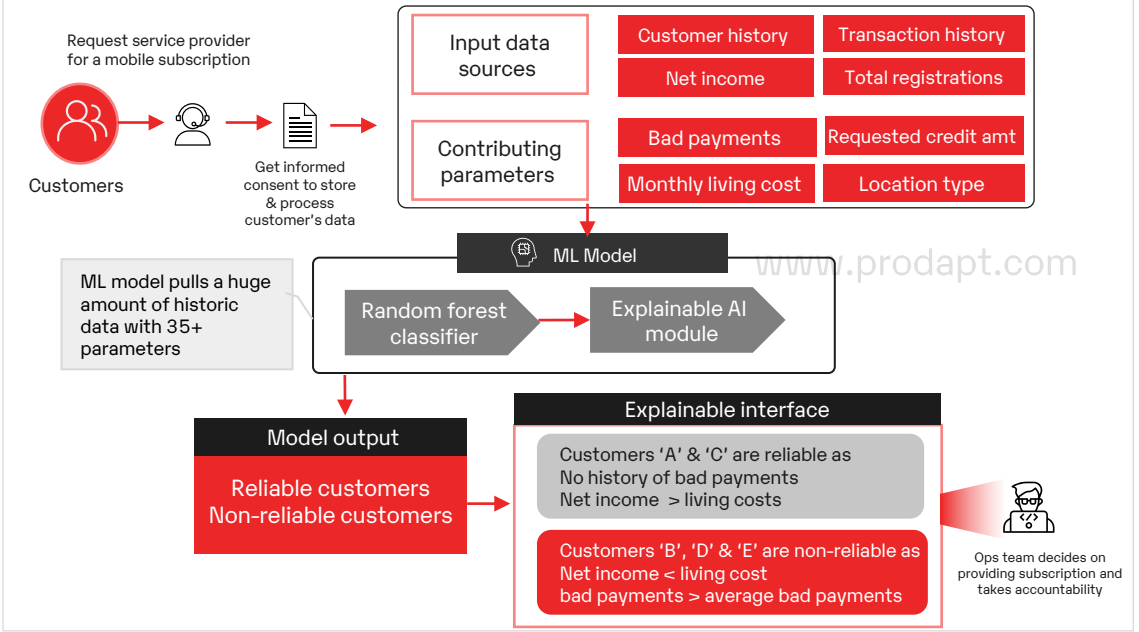
Perform continuous model evaluation to identify and mitigate unintended model behavior, drift in fairness, and explainability.

Assessment of the AI pipeline across these pillars mitigates the risk of serious harms, increasing cost savings. This insight deep dives into the 4 key pillars of Responsible AI and provides best practices for its effective implementation with **customer segmentation** as a sample use case.

Explainability: Gain a complete view of AI decision-making

With the increasing robustness of AI systems, interpreting the algorithm that derives results is becoming more challenging. Implementing explainable AI assists in describing a model, predicting its impact and gaining confidence while putting AI models into production.

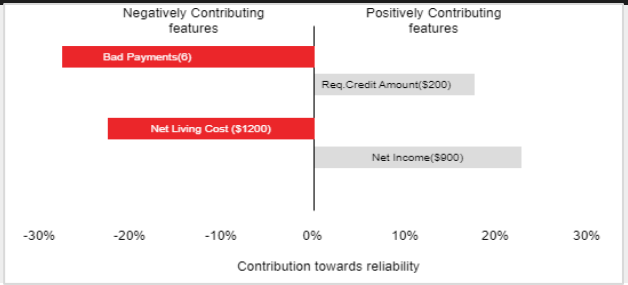
Sample use case: Implementing explainability in customer segmentation for trusted decision-making



Recommendations

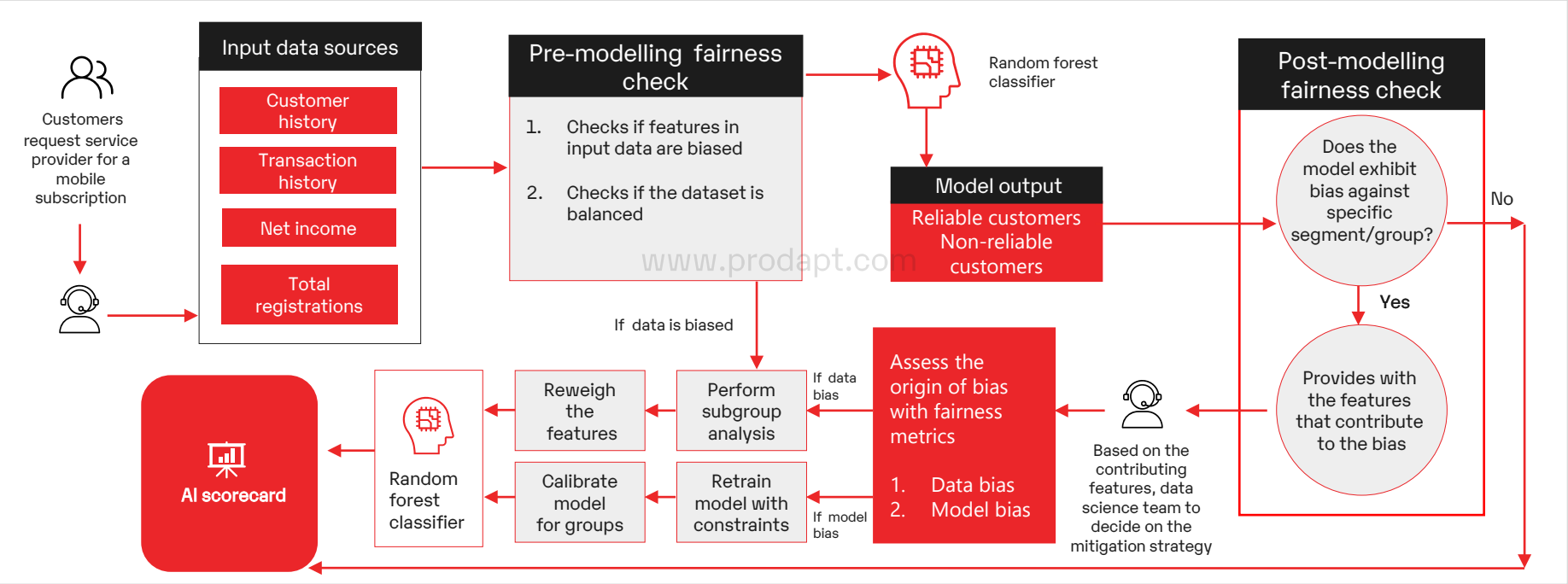
- Implement mechanisms like Local Interpretable Model-Agnostic Explanations **(LIME)** and SHapley Additive explanations **(SHAP)** to provide detailed explanations of predictions
- Develop **variable importance** and **partial dependence** reports to analyze the significance of each variable in the model predictions. This helps to understand the model behavior and improves transparency
- Implement **counterfactual analysis** to analyze model behavior in the absence of specific variables

Sample report: Local explanations showing why customer 'B' is not reliable



Fairness: Avoid biased predictions towards a sub-population

Inclination or prejudice against a certain group is unfair. AI systems can be affected by bias at any stage of prediction leading to reputational damage and revenue loss. Implementing fairness analysis prevents systematic advantages only to privileged groups and individuals.



Fairness: Avoid biased predictions towards a sub-population

Recommendations

- Implement root cause analysis to identify the features that contribute to the disparity. This assists in understanding how the bias crept into the ML model and determining a rapid mitigation strategy
- Ensure prediction (P) is statistically **independent** and **separated** from the sensitive feature (S) like 'Gender' for a given target class (T)
- Evaluate metrics such as **Demographic parity** and **Equalized odds** to measure independence and separation of the prediction
- Ensure the ratio of these metrics against the target class is '1' to avoid bias due to sensitive features
- Implement **Synthetic Minority Oversampling Technique (SMOTE)** to balance the data before model development
- Leverage model monitoring tools such as **MLFlow, and Amazon SageMaker Model Monitor** to determine if the model's fairness change over time
- Use tools like **Fairlearn** to generate reports on fairness issues using metrics across sensitive features and cohorts

Privacy: Enable data protection in AI systems

The evolution of AI systems magnifies the ability to use personal information, increasing the risk of privacy breaches and potential misuse of personal data. Principles of Responsible AI such as explainability, fairness, robustness and security of data processing are related to specific individual rights and provisions of corresponding privacy laws.

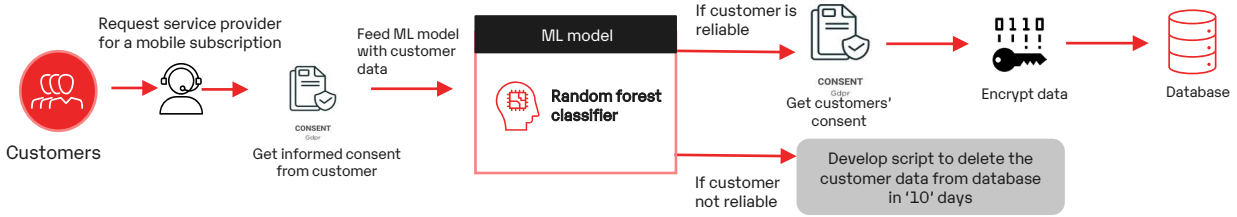
Gartner reported that"40% of organizations had an AI privacy breach and that, of those breaches, one in four was malicious."

Recommendations

1. Facilitate privacy during data engineering

- Ensure **informed consent** from the customer to store and process the personal data, as per **GDPR** guidelines. This assists in consent- based access of customer's Personally Identifiable Information (PII) for reliability check

- **Encrypt** all sensitive data post getting informed consent from the customers. This ensures confidentiality during storage and access of customers' personal data

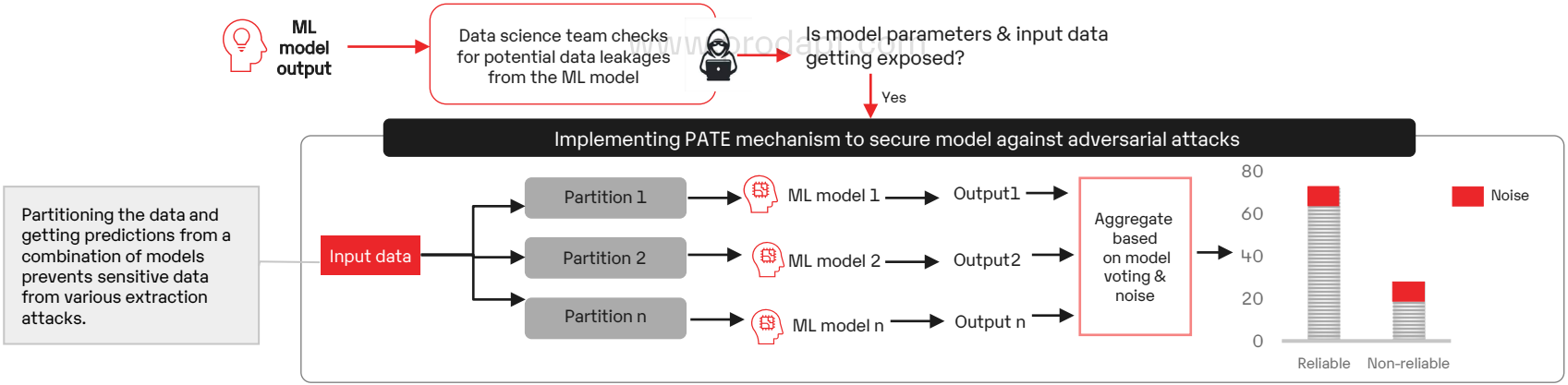


Privacy: Enable data protection in AI systems

Training data and model predictions may contain sensitive information. Hence it is vital to defend the models against malicious attacks and ensure customer data privacy.

2. Ensure privacy in ML models

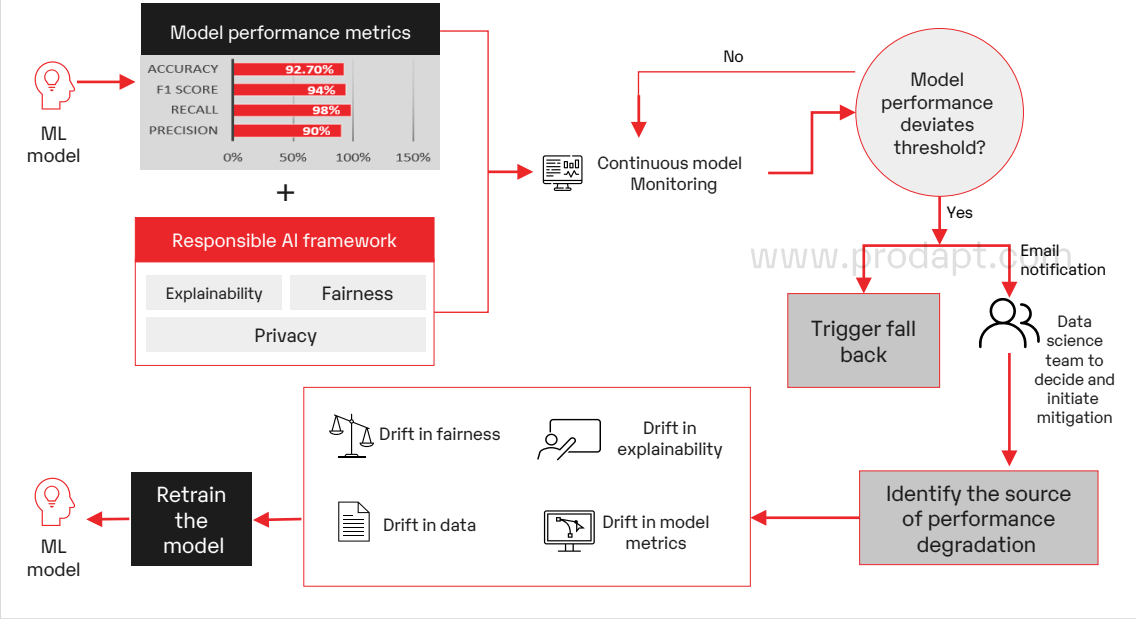
- Perform defensive measures to protect the training data and ML model from data extractions
- Implement the **Private Aggregation of Teacher Ensembles** (PATE) mechanism to secure the model from adversarial attacks and ensure privacy, especially when the model possesses intellectual properties (e.g., business trends or patterns)
- Leverage **differential privacy** which adds noise, safeguarding the real data from attackers



Model performance: Perform continuous model evaluation to identify and mitigate unintended model behavior

As the AI/ML models often interact with various real-world events, model predictions and accuracy can degrade over time. Investigating model behaviors using insights based on fairness, explainability, and model quality is essential to scaling AI. Continuous model evaluation empowers businesses to compare model predictions, quantify model risk and optimize performance.

Sample flow of continuous model evaluation to mitigate possible drifts



Recommendations

- Leverage tools such as **MLFlow, Amazon SageMaker Model Monitor,** and **Vertex AI Model Monitoring** to monitor the models for data and model quality, bias, and explainability
- Create baselines to analyze input features and bias and track the drift. Also, set alerts to notify the data science team whenever features exceed the threshold
- Implement methodology like Kullback-Leiber Divergence and Population Stability Index to identify drift by comparing the difference between two subsets of data

AI reliability scorecard to accelerate smart decision-making

Provides a unified view of model health with scores across four pillars for the operations and data science team to mitigate potential risks and arrive at smart decisions

AI RELIABILITY SCORECARD

MODEL CARD

Model ID : dt_101

Family : Sklearn

Type : Classification

Name : Decision Tree Classifier

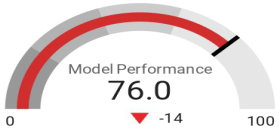
Input : 3333 rows and 19 columns

Output : Labels

Accuracy : 92.7

MODEL PERFORMANCE

Is your AI pipeline accountable, technically robust and safe?



Result : Good

EXPLAINABILITY

Is your AI pipeline transparent?

Result : Excellent

Value drop and increase is spotted based on previous iteration

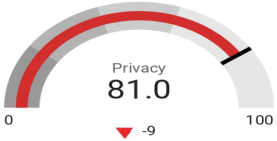
FAIRNESS

Is your AI pipeline diverse, non-discriminative and fair?

Result : Good

PRIVACY

Does your AI pipeline adhere to data governance principles?



Result : Excellent

PERFORMANCE RANGE :
Poor : 0-40 **Average :** 40-60 **Good :** 60-80 **Excellent :** 80-100

The identified risks and call to actions will be notified to the data science team for risk mitigation

FINDINGS

DRIFT IDENTIFIED Info

Mitigation Strategy : Retrain Model

Fix | Trigger | SendMail

BIAS DETECTED Info

Business benefits achieved by a leading service provider after implementing Responsible AI

Implementing the four pillars as discussed in this insight, resulted in the following benefits.

Reduction in the impact of model bias

30% reduction in OpEx

4X faster and trustworthy decision-making

Lower risk and cost of model governance

Thank you

insights@prodapt.com