Archives available at journals.mriindia.com

International Journal of Recent Advances in Engineering and Technology

ISSN: 2347 - 2812

Volume 11 Issue 01, 2022

Explainable AI Models for Transparent Decision-Making in Business Analytics Dashboards

Sathish Kaniganahalli Ramareddy
Manager Technology, Publicis Sapient, USA
Email: reachsathishramareddy@gmail.com

Peer Review Information

Submission: 11 April 2022

Revision: 10 May 2022

Acceptance: 22 May 2022

Keywords

Explainable AI (XAI); Transparent Decision-Making; Business Analytics Dashboards; SHAP; LIME; Counterfactual Explanations; Model Interpretability; Responsible AI; Data-Driven Decision Support; Enterprise Intelligence; Trustworthy AI; AI Governance; Ethical Analytics

Abstract

Artificial intelligence-powered business dashboards are increasingly used to drive strategic and operational decisions across industries. However, traditional black-box machine learning models lack transparency, limiting trust, auditability, and widespread adoption in high-impact business settings. This study proposes an explainable AI-enabled dashboard architecture that integrates model interpretability techniques into enterprise analytics systems to support transparent and accountable decision-making. The methodology includes data preprocessing, model training, post-hoc explainability techniques such as SHAP and LIME, surrogate interpretable models, counterfactual analysis, and user-centric visual explanation components. The resulting system provides both predictive insights and intuitive explanations that highlight feature importance, model behavior, and decision rationale for individual instances and global patterns. Applications in finance, healthcare, marketing, supply chain management, and cybersecurity demonstrate the practical utility of explainable dashboards. Experimental evaluation and design considerations confirm that explainable interfaces enhance user trust, regulatory compliance, ethical governance, and adoption of AI-driven analytics in enterprise environments. This research contributes a scalable and interpretable approach to designing responsible business intelligence systems and establishes a foundation for future work in adaptive, real-time, and personalized AI explanations.

Introduction

In the era of data-driven enterprises, business analytics dashboards have evolved from static reporting interfaces into intelligent cognitive platforms capable of automated insights, predictive modeling, and real-time decision recommendations. Organizations across industries increasingly rely on machine learning (ML) and artificial intelligence (AI) to forecast business trends, identify risks, profile customers, and optimize operational workflows. Despite the remarkable predictive accuracy and automation capabilities of modern AI models, their widespread adoption in business analytics faces one critical barrier — the lack of transparency and interpretability [1]. Traditional black-box models such as deep neural networks, ensemble learning frameworks, and reinforcement learning agents often make highly accurate predictions, but they fail to provide human-interpretable justification for their outputs. This opacity can reduce managerial trust, limit regulatory compliance, and impede decision-makers from confidently adopting AI-generated insights in high-stakes business contexts.

Explainable AI (XAI) has emerged as a transformative paradigm to address this challenge by enabling AI systems to provide transparent, interpretable, and trustworthy predictions. XAI techniques generate human-understandable explanations such as feature contributions, rule-based rationales, counterfactual insights, and visual reasoning maps, allowing business users to understand *why* a model behaved in a specific way [2]. The integration of XAI into business analytics dashboards not only enhances user trust but also strengthens accountability, fairness, and ethical alignment, particularly in domains governed by strict regulations such as finance, healthcare, insurance, supply chain management, and banking. By making machine decisions auditable and comprehensible, XAI bridges the gap between technical complexity and managerial intuition, fostering responsible and confident AI adoption in executive decision workflows.

In business environments where decisions impact financial investments, customer relationships, regulatory reporting, and market strategies, the ability to explain AI-driven predictions becomes indispensable. Executives, analysts, and regulators increasingly demand interpretable insights rather than opaque outputs, especially when decisions involve risk scoring, anomaly detection, credit approvals, customer segmentation, and marketing automation. XAI empowers stakeholders to validate model fairness, detect data biases, and ensure that AI-powered dashboards align with organizational objectives and ethical standards. Moreover, transparent models enable democratic access to insights across non-technical business users, enhancing data literacy and decision reliability throughout the enterprise hierarchy [3].

This research aims to develop a comprehensive framework that integrates explainable AI models into business analytics dashboards to enhance decision-making transparency and accountability. The work emphasizes the selection and evaluation of XAI techniques such as SHAP, LIME, integrated gradients, surrogate decision trees, and model-agnostic rule-based engines, and investigates their usability and effectiveness in real-world business dashboards. The study also explores visual interpretation components, such as feature contribution charts, heatmaps, force plots, and actionable explanation widgets, to ensure that interpretability is delivered in an intuitive and interactive manner suitable for executive users.

By addressing the intersection of AI interpretability and advanced business visualization platforms, this study contributes to building high-trust business intelligence ecosystems. It lays the foundation for transparent digital decision-making infrastructure, reducing risks associated with black-box AI deployments and empowering organizations to adopt intelligent analytics responsibly. Ultimately, the integration of XAI in business dashboards is not only a technological advancement but a strategic necessity for enterprises seeking sustainable competitive advantage, regulatory adherence, and ethical AI governance in the digital era.

Literature Review

The increasing adoption of artificial intelligence (AI) in enterprise decision systems has transformed the landscape of business analytics, enabling predictive insights, automation capabilities, and strategic intelligence across multiple industries. Business analytics dashboards have evolved from descriptive visualization platforms to advanced analytical engines powered by machine learning (ML) and deep learning models [4]. However, their reliance on complex black-box architectures poses challenges in interpretability and trust, especially when critical decisions are made based on AI-driven outcomes. To address these challenges, Explainable Artificial Intelligence (XAI) has emerged as a vital research domain that focuses on developing transparent and interpretable AI systems capable of justifying predictions and supporting responsible business decision-making [5].

Business Analytics and AI-Driven Decision Systems

AI-assisted dashboards have gained popularity for their ability to uncover hidden patterns, detect anomalies, and generate data-driven recommendations in real time. Platforms such as Power BI, Tableau, and Looker increasingly incorporate predictive capabilities to support strategic functions in finance, marketing, supply chain, and customer analytics [3]. Studies highlight that enterprises adopting AI-augmented dashboards observe enhanced operational efficiency and proactive decision-making capabilities. Yet, as these systems become more autonomous, business stakeholders demand greater transparency to validate model outputs and ensure alignment with organizational objectives and ethical standards [6].

Explainable AI: Concepts and Need

Explainable AI aims to enhance model interpretability while maintaining predictive accuracy. XAI techniques are broadly categorized into intrinsic models that are inherently interpretable (e.g., decision trees, linear regression, rule-based models) and post-hoc explanation tools such as LIME, SHAP, counterfactual reasoning, and gradient-based saliency maps. Literature emphasizes that transparency not only increases user trust but also assists enterprises in complying with regulatory frameworks like GDPR, which mandates algorithmic accountability for automated decisions [7]. Business contexts often involve sensitive decisions such as loan approvals, fraud detection, and churn prediction, where explainable reasoning is crucial for stakeholder acceptance and ethical assurance [8].

Interpretability Techniques in Business AI Models

Several studies evaluate interpretability techniques for enterprise AI adoption. SHAP and LIME have been widely used to provide feature-level contribution explanations in financial forecasting, credit scoring, and marketing analytics models [9]. Surrogate models such as interpretable decision trees and rule extraction frameworks are applied to explain deep learning-based predictions, making them suitable for managerial interpretation [10]. Hybrid frameworks combining white-box and black-box models are increasingly used to balance accuracy and interpretability in high-risk business environments [11]. Furthermore, user-centric visualization methods such as force plots, waterfall charts, and heatmaps enable intuitive understanding of model behavior, especially when integrated within BI dashboards [12].

Challenges and Research Gaps

Despite significant advancements, several challenges persist. High-dimensional business datasets, real-time inference needs, data privacy constraints, and model drift pose complexities for interpretable AI deployment [13]. Studies also emphasize the absence of standardized evaluation metrics for measuring explainability effectiveness, user trust, and cognitive load in decision environments [14]. Moreover, limited research exists on developing dashboard-level XAI interfaces tailored for executive users, particularly focusing on intuitive explanation design, cognitive ergonomics, and human-AI collaboration in enterprise decision workflows [15].

Ref	Problem Addressed	Method / Focus	Key Findings / Contribution	Tools / XAI Techniques	Research Gap / Limitation	Section Theme
[1]	Lack of intelligent decision capability in dashboards	Analysis of AI-enabled dashboards	AI transforms dashboards into predictive systems	ML, BI platforms	Limited interpretability in enterprise use	AI in Business Analytics
[2]	Black-box AI limits adoption	Overview of XAI importance	Transparency improves trust and adoption	XAI theory	Practical enterprise frameworks lacking	XAI Need
[3]	Passive dashboards insufficient	BI platforms with AI integration	Power BI/Tableau increasingly include ML	BI + ML stack	XAI integration still emerging	AI-Driven Dashboards
[4]	Limited proactive decisions	Enterprise case studies	AI dashboards enhance operational efficiency	ML-based analytics	Explainability layer not addressed	Business Impact
[5]	Ethical & trust concerns in automation	Governance and transparency requirement	Transparency essential for trustworthy AI	N/A	No standardized trust frameworks	Ethical AI & Trust
[6]	Lack of clarity in model reasoning	Survey of XAI models	Categories: intrinsic vs post-hoc	LIME, SHAP, Rule-Based Models	Challenge scaling to enterprise data	Explainable AI Concepts
[7]	Regulatory compliance requirements	Policy & compliance study	GDPR mandates	Auditable AI frameworks	No enterprise operational model given	XAI & Compliance

			algorithm explanation			
[8]	Sensitive business decisions risk bias	Sector-specific cases	XAI essential in credit, fraud, churn systems	Credit scoring + XAI	Limited UI-driven explanations	High-Risk Decisions
[9]	Understanding ML decisions	Evaluation of SHAP/LIME	SHAP/LIME effective in finance	SHAP, LIME	Computational overhead in real-time dashboards	SHAP/LIME Usage
[10]	Explaining complex models	Surrogate explainability	Rule extraction improves transparency	Surrogate Trees	Surrogates may oversimplify	Surrogate Models
[11]	Trade-off between accuracy & interpretability	Hybrid AI modeling	Combining white-box + black-box models beneficial	Hybrid frameworks	UI and business usability not addressed	Hybrid XAI
[12]	Need for intuitive insights	Visual explanation tools	Force plots, heatmaps aid understanding	SHAP plots, heatmaps	Lacks user-experience validation	XAI Visualization
[13]	High data dimensionality & drift	Analysis of enterprise AI constraints	Data drift + privacy challenges	Drift monitoring	Real-time explainability difficult	XAI Challenges
[14]	Lack of explainability metrics	Evaluation need analysis	Gap in trust & cognitive load measurement	Trust metrics	No universal explainability KPIs	XAI Evaluation Gaps
[15]	Poor usability of enterprise XAI tools	Executive-focused XAI gap	Need for business-friendly dashboards	XAI UX principles	Minimal UX frameworks exist	Research Gap — Dashboard UI

The literature indicates strong momentum toward explainable AI in enterprise analytics. However, opportunities remain to design comprehensive frameworks that integrate predictive AI with interactive XAI visual modules, ensuring transparency, usability, regulatory compliance, and trust in business dashboards. This research aims to bridge these gaps by investigating practical XAI methods, designing executive-friendly explanation interfaces, and evaluating interpretability metrics to support transparent business decision-making.

System Architecture & Conceptual Framework

The proposed system introduces a comprehensive architecture designed to seamlessly integrate Explainable Artificial Intelligence (XAI) models into business analytics dashboards to enable transparent and trustworthy decision-making. The architecture combines advanced machine learning pipelines, explainability engines, and interactive visualization modules within a unified business intelligence framework. The overarching goal is to transform traditional dashboards into cognitive, interpretable, and decision-support systems that empower users with both predictive insights and understandable justifications behind AI recommendations. This section elaborates on the conceptual structure, key system components, and operational flow, illustrating how data moves from enterprise systems to interpretable and actionable insight delivery interfaces.

At the core of the architecture lies the **Data Processing Layer**, which acquires operational, transactional, and customer-related data from enterprise sources such as ERP, CRM, cloud warehouses, and IoT data lakes. The raw data undergoes a systematic transformation pipeline including cleansing, feature engineering, dimensionality reduction, and normalization. Robust data governance mechanisms, data access authorization, and role-based usage protocols ensure accuracy, consistency, and regulatory compliance during preprocessing. Once curated, the refined dataset is routed to the **AI Modeling Layer**, where multiple ML/DL models are trained and deployed based on business problems — such as classification for churn prediction, regression for revenue forecasting, or clustering for customer segmentation. The system supports model orchestration using widely adopted machine learning frameworks including Scikit-Learn, TensorFlow, PyTorch, and cloud-native services.

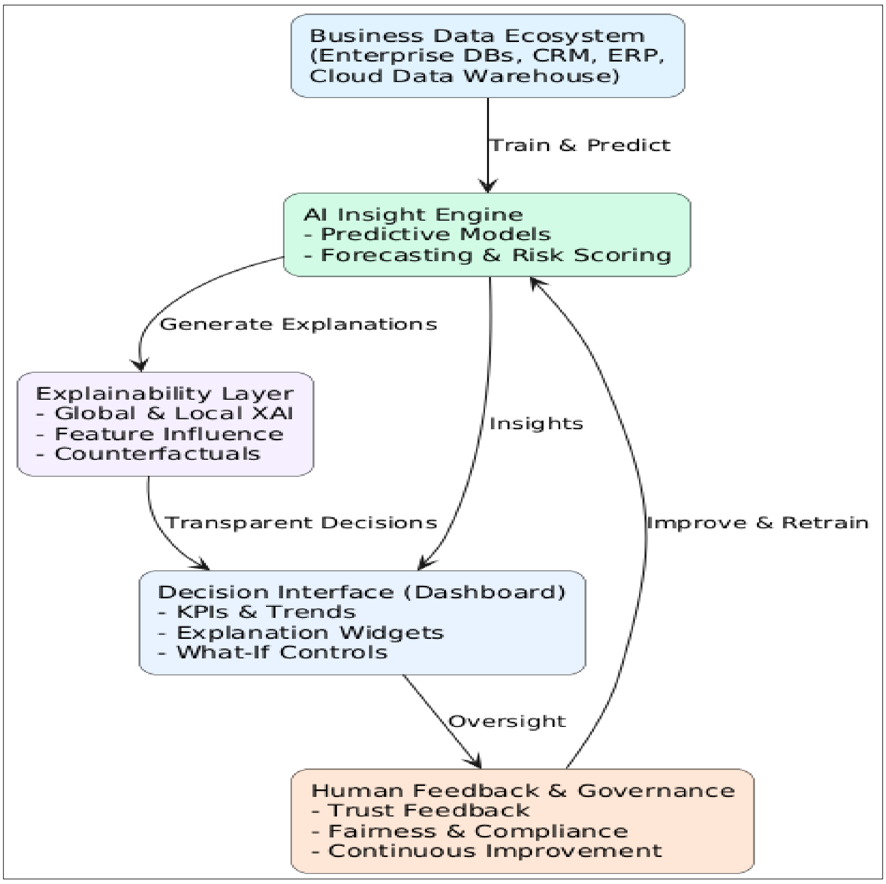

Figure 1. Proposed System Architecture

The **Explainability Engine** represents the core innovation in this architecture. It hosts model-agnostic and model-specific interpretability algorithms such as SHAP, LIME, Decision Rules, Integrated Gradients, Counterfactual Explainers, and Surrogate Decision Trees. These explanation tools function parallel to the prediction engine, deriving local and global interpretability metrics without disturbing real-time inference pipelines. Local interpretability modules generate case-specific explanations (e.g., why a particular customer may churn), whereas global interpretability models provide generalized feature-importance reasoning for executive-level understanding. Explanation outputs are formatted into intuitive insights such as contribution bar charts, SHAP force plots, waterfall charts, rule narratives, and counterfactual sliders.

The **Business Analytics Dashboard Layer** serves as the presentation and interaction tier, enabling business users to seamlessly consume predictions alongside explanations. Integration with modern BI platforms such as Power BI, Tableau, Qlik Sense, and Apache Superset ensures flexible deployment in enterprise environments. The dashboard provides configurable transparency widgets, enabling decision-makers to drill into model logic, assess influencing variables, and validate recommendations through visual storytelling. Security controls ensure role-based explainability, restricting sensitive model variables from unauthorized access while promoting transparency for decision-makers. This layer incorporates real-time streaming support for dynamic business operations such as fraud detection, revenue anomalies, sales forecasting, and supply chain risk alerts.

Finally, the architecture includes a **Monitoring and Feedback Loop**, responsible for performance tracking, drift detection, explanation quality evaluation, and user trust feedback collection. Continuous monitoring ensures that models remain fair, accurate, and interpretable over time. User feedback is systematically captured through interaction logs and rating prompts to quantify explanation usefulness and cognitive load. This feedback becomes input for iterative model refinement, enabling a cycle of learning and improvement. Overall, the conceptual framework emphasizes transparency, accountability, and business relevance, ensuring that AI-driven dashboards not only produce reliable insights but also support ethical governance and informed decision-making across enterprise environments.

Methodology

The proposed methodology adopts a structured, multi-stage pipeline that ensures seamless integration of machine learning models and explainability modules into a business analytics dashboard. Each stage is designed to progress logically from raw data acquisition through interpretability deployment and continuous organizational feedback, ensuring that transparency and user trust are embedded throughout the analytical lifecycle.

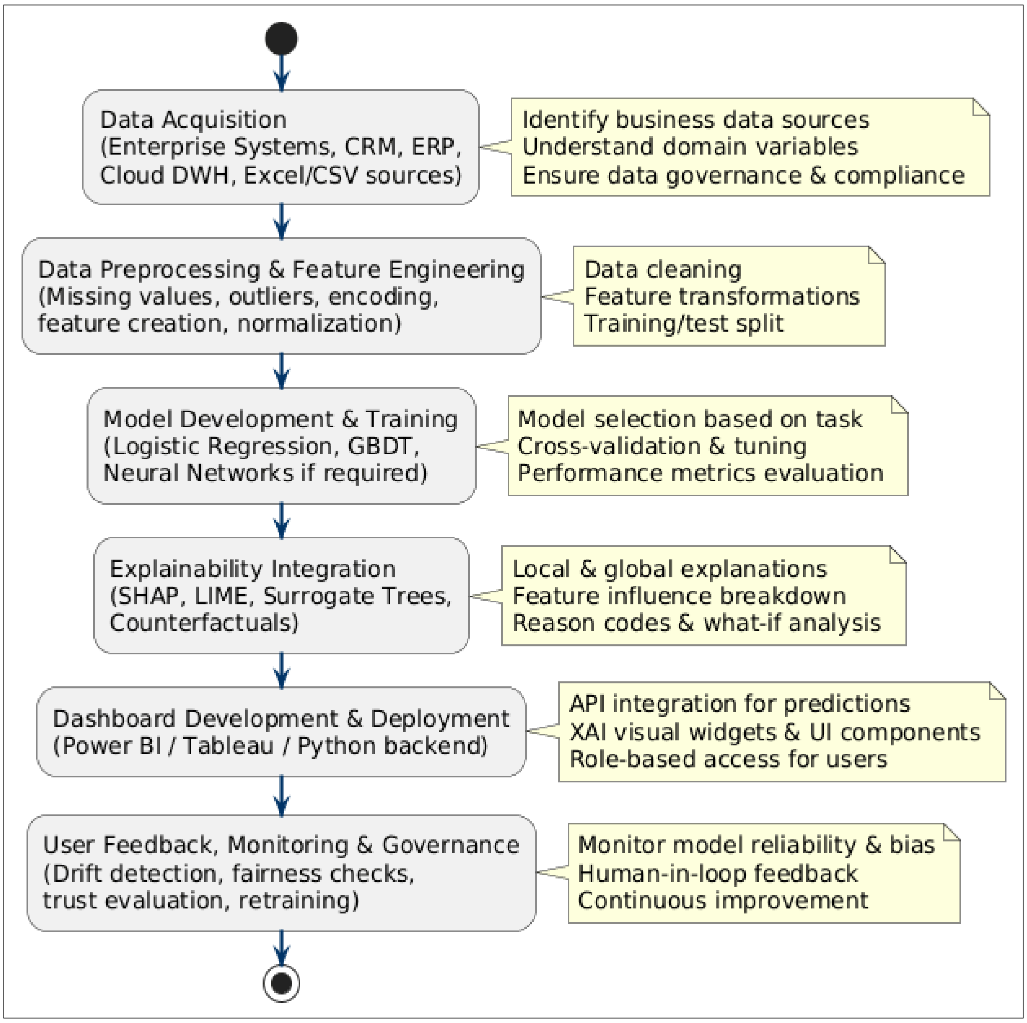
Data Acquisition
(Enterprise Systems, CRM, ERP,
Cloud DWH, Excel/CSV sources)
Identify business data sources
Understand domain variables
Ensure data governance & compliance
Data Preprocessing & Feature Engineering
(Missing values, outliers, encoding,
feature creation, normalization)
Data cleaning
Feature transformations
Training/test split
Model Development & Training
(Logistic Regression, GBDT,
Neural Networks if required)
Model selection based on task
Cross-validation & tuning
Performance metrics evaluation
Explainability Integration
(SHAP, LIME, Surrogate Trees,
Counterfactuals)
Local & global explanations
Feature influence breakdown
Reason codes & what-if analysis
Dashboard Development & Deployment
(Power BI / Tableau / Python backend)
API integration for predictions
XAI visual widgets & UI components
Role-based access for users
User Feedback, Monitoring & Governance
(Drift detection, fairness checks,
trust evaluation, retraining)
Monitor model reliability & bias
Human-in-loop feedback
Continuous improvement

Figure 2. Proposed Methodology

Data Acquisition and Understanding

This stage focuses on gathering relevant business information from diverse sources such as CRM systems, ERP platforms, customer transaction logs, cloud data warehouses, and digital customer touchpoints. Business domain experts collaborate with data engineers to contextualize the variables and define relationships between operational, financial, and customer-centric attributes. During this stage, regulatory considerations—including GDPR, HIPAA, and institutional data-governance policies—are mapped to the data flow to ensure ethical and compliant usage. The output of this stage is a comprehensive dataset library with documented semantics and business meaning, forming the foundation for reliable model development.

Data Preprocessing and Feature Engineering

Once the data sources are established, thorough preprocessing is conducted to enhance quality and reliability. This includes handling missing values using suitable imputation techniques, addressing outliers through statistical filtering, normalizing and scaling numerical fields, encoding categorical variables, and splitting data into training and testing segments. Simultaneously, feature engineering is performed with a strong focus on business interpretability, ensuring engineered indicators reflect real-world behavior rather than abstract transformations. The resulting dataset is structured, cleaned, and enriched with features that improve both predictive power and clarity in subsequent explanations.

Model Development and Training

Following preprocessing, predictive models are developed according to the business problem—classification for churn detection, regression for demand forecasting, or clustering for customer segmentation. Models range from interpretable baselines such as logistic regression and decision trees to more advanced algorithms including gradient boosting models and neural networks where necessary. The training phase employs cross-validation and hyperparameter tuning to balance performance and generalization capability. Evaluation metrics such as accuracy, recall, F1-score, ROC-AUC, or mean squared error are computed, ensuring that selected models demonstrate acceptable predictive integrity before proceeding to interpretability integration.

Explainable AI Module Integration

This stage represents the core innovation, where explainability techniques are layered on top of trained models. Model-agnostic and model-specific explainers—such as SHAP for feature-attribution analysis, LIME for localized perturbation-based interpretation, surrogate decision-tree methods for rule extraction, and counterfactual explainers for scenario exploration—are integrated into the inference pipeline. The explainability outputs are then formatted into intuitive narratives, ranked feature contributions, and visual interpretability components. The output of this phase is a transparent reasoning engine capable of revealing why a model generated a particular prediction, aligning computational logic with business decision expectations.

Dashboard Design and Deployment

With prediction and explanation components ready, the system is embedded into a business analytics dashboard such as Power BI, Tableau, Looker, or an internal enterprise BI interface. The backend exposes model predictions and explanation artifacts through secure APIs. The dashboard interface is engineered to present business KPIs alongside interpretability widgets, including feature importance graphs, explainer text panels, threshold indicators, and what-if scenario sliders. Real-time or scheduled inference is supported based on organizational needs. Access control layers ensure that different user roles—from executives to analysts and auditors—see explanations tailored to their decision context and authority level.

Human Feedback, Monitoring, and Governance

The final stage implements continuous monitoring and governance mechanisms to ensure ongoing trustworthiness and reliability. User interactions with explanations are analyzed to capture interpretability usefulness, cognitive load, and trust-building feedback. Model performance drift and data drift are monitored, while fairness diagnostics ensure non-discriminatory predictions over time. Feedback from end users, auditors, and governance teams informs model retraining cycles, explanation refinement strategies, and dashboard usability improvements. This feedback loop establishes a self-correcting system where explainability and accuracy evolve with business dynamics and user expectations, reinforcing both operational trust and institutional compliance.

Applications

Explainable AI–enabled dashboards are transforming data-driven decision-making across industries by providing transparent, accountability-focused intelligence. Instead of merely generating predictions, these dashboards clarify *why* decisions are suggested, which improves managerial trust, regulatory compliance, and organizational adoption of AI systems. The applications span multiple strategic and operational domains.

1. Customer Churn Prediction & Retention

In customer-centric businesses such as telecom, banking, insurance, and retail, XAI dashboards help identify customers at risk of leaving and reveal the factors that influence churn, such as reduced engagement, high complaints, or price sensitivity. Decision-makers can assess customer-specific explanations and design targeted retention policies, discounts, or proactive engagement campaigns. Counterfactual dashboards also allow managers to explore what changes (e.g., improved service, reduced pricing) could prevent churn.

2. Credit Risk Scoring & Financial Decisioning

In credit and lending environments, explainability plays a central role in compliance and fairness. XAI dashboards enhance credit-scoring models by showing feature contributions like credit history, debt ratios, employment stability, or spending patterns. This transparency fulfills regulatory expectations (e.g., RBI, GDPR, FCA) and helps loan officers justify approvals or rejections, reducing bias and improving responsible lending practices.

3. Fraud Detection & Transaction Monitoring
Explainable AI enhances fraud-analytics dashboards by not only flagging suspicious transactions but also explaining the underlying indicators such as behavioral anomalies, unusual location patterns, or abnormal spending frequency. Financial institutions can validate alerts more efficiently, detect false positives, and accelerate investigation workflows while maintaining audit trails for regulatory checks.

4. Sales Forecasting & Revenue Analytics
Businesses rely on forecasting dashboards to predict sales performance and demand fluctuations. With XAI, leaders can understand which factors—seasonality, promotions, customer demographics, or product mix—drive sales outcomes. This leads to more informed budgeting, capacity planning, and targeted promotional strategies.

5. Supply Chain Optimization & Logistics Planning
Supply chain managers use XAI dashboards to predict demand variability, transport delays, supplier risks, and inventory imbalances. Explanations on lead time deviations, fuel costs, supplier reliability scores, and logistics bottlenecks enable transparent decisions, improving supply network resilience and cost efficiency.

6. HR Analytics & Talent Management
HR teams leverage XAI dashboards to predict employee turnover, hiring needs, or performance trends. Transparent explanations allow leaders to identify workload imbalance, training needs, skill gaps, and engagement issues without bias. This ensures ethical, evidence-driven talent strategies and supports fairness in workforce decisions.

7. Healthcare Diagnostics & Patient Risk Stratification
In healthcare delivery, AI-enabled dashboards assist clinicians by predicting patient outcomes, disease progression, or likelihood of readmission. Explainability ensures medical staff can understand risk contributors such as vital patterns, lab results, or treatment history, supporting clinical trust, ethical care, and regulatory standards in medical diagnosis.

8. Marketing Personalization & Campaign Optimization
Marketing teams use explainable dashboards to optimize customer segmentation, targeting, and personalization campaigns. Understanding the drivers behind customer conversions—such as engagement frequency, product interest, or discount sensitivity—enables ethical personalization and prevents intrusive or biased targeting.

9. Cybersecurity Threat Detection
Security operation centers (SOCs) apply XAI dashboards to identify intrusion patterns, abnormal network traffic, and insider threats. Transparent reasoning helps analysts validate alerts faster, differentiate legitimate anomalies from threats, and strengthen cyber-defense posture with evidence-based insights.

10. ESG Compliance & Financial Governance
Corporate sustainability reporting and ESG dashboards integrate XAI to track energy use, emissions, ethical sourcing, and governance metrics. Explainable models increase confidence in ESG scoring systems, ensuring audit-readiness and transparent compliance reporting for stakeholders and regulators.

Conclusion

This work presented a comprehensive framework for integrating explainable artificial intelligence into business analytics dashboards to support transparent, reliable, and trust-driven decision-making in modern enterprises. While conventional AI systems have enabled substantial improvements in predictive accuracy and automation, their opaque nature has limited trust and adoption among business leaders, regulators, and operational stakeholders. By embedding interpretability mechanisms such as SHAP, LIME, surrogate models, and counterfactual analysis directly into the analytical workflow, the proposed approach bridges the gap between machine predictions and human understanding. The resulting system not only delivers accurate forecasts and risk insights but also provides clear, intuitive reasoning behind each model output, ensuring transparency, accountability, and ethical alignment. The proposed methodology emphasizes end-to-end integration — from data preprocessing and model development to explainability layer deployment, dashboard design, user interaction, and governance. With interactive visual explanation components, role-based access, and continuous feedback loops, the system promotes informed executive decision-making while enabling analysts and auditors to validate fairness, detect bias, monitor drift, and maintain regulatory compliance. Applications across domains including finance, healthcare, retail, supply chain, HR analytics, and cybersecurity demonstrate the versatility and real-world impact of explainable dashboards in enhancing operational visibility and strategic intelligence. Overall, explainable AI represents a necessary evolution in enterprise analytics, combining the power of machine learning with human interpretability and institutional governance. This work contributes to the development of responsible AI ecosystems and lays the foundation for future research in real-time interpretability, automated fairness assurance, user trust quantification, and adaptive explanation personalization. The integration of XAI into business dashboards is not merely a technological enhancement, but a strategic imperative for organizations seeking to build trustworthy, accountable, and future-ready intelligence platforms.

References

José Manuel Alonso, Concetta Castiello, Luis Magdalena, and Carmelo Mencar, *Explainable Fuzzy Systems—Paving the Way from Interpretable Fuzzy Systems to Explainable AI Systems*. Cham, Switzerland: Springer, 2021.

Judea Pearl and Dana Mackenzie, *The Book of Why: The New Science of Cause and Effect*. London, U.K.: Penguin, 2018.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2018.

Lilian Edwards and Michael Veale, "Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for," *Duke Law & Technology Review*, vol. 16, p. 18, 2017.

David Gunning, Mark Stefik, James Choi, Tim Miller, Simone Stumpf, and Guang-Zhong Yang, "XAI: Explainable artificial intelligence," *Science Robotics*, vol. 4, no. 37, 2019.

Amal Adadi and Mohammed Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52,138–52,160, 2018.

Thomas Rieg, Johannes Frick, Heinrich Baumgartl, and Ralf Buettner, "White-box machine learning for cardiovascular disease electrocardiograms," *PLoS One*, vol. 15, no. 12, e0243615, 2020.

Carissa Véliz, Claudia Prunkl, Michael Phillips-Brown, and T. M. Lechterman, "We might be afraid of black-box algorithms," *Journal of Medical Ethics*, vol. 47, pp. 339–340, 2021.

Niloofar Mehrabi, Fred Morstatter, Ninareh Saxena, Kristina Lerman, and Aram Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, pp. 1–35, 2021.

Chris Finlay and Adam Michael Oberman, "Scaleable input gradient regularization for adversarial robustness," *Machine Learning with Applications*, vol. 3, 100017, 2021.

Aparna Das and Peter Rad, "Opportunities and challenges in Explainable Artificial Intelligence (XAI): A survey," *arXiv preprint arXiv:2006.11371*, 2020.

Zachary Chase Lipton, "The mythos of model interpretability," *Communications of the ACM*, 2018.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. ACM SIGKDD*, 2016, pp. 1135–1144.

Haofan Wang, Zifan Wang, Mingyang Du, Fan Yang, Zhe Zhang, Sirui Ding, Pierre Mardziel, and Xia Hu, "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *IEEE/CVF CVPR Workshops*, 2020.

Ramprasaath Ramesh Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *IEEE ICCV*, 2017, pp. 618–626.

Shane T. Mueller, Robert R. Hoffman, W. Lewis Clancey, Alexandra Emrey, and Gary Klein, "Explanation in human-AI systems: Meta-review," *arXiv:1902.01876*, 2019.