REVIEWING THE RISKS OF AI TECHNICAL DEBT (TD) IN THE FINANCIAL SERVICES INDUSTRIES (FSIS)

Vinay Kumar Sankarapu

REVIEWING THE RISKS OF AI TECHNICAL DEBT (TD) IN THE FINANCIAL SERVICES INDUSTRIES (FSIS)

Vinay Kumar Sankarapu
vinay@aryaxai.com

September 3, 2024

ABSTRACT

AI is increasingly becoming an important catalyst for improving the efficiency and efficacy of the financial services industry. For this paper, we consider institutions that provide banking services, insurance services, payment services, and investment services as part of the financial services industry (FSI). The recent success of generative AI and predictive AI in the last few years generated enormous interest in deploying AI across FSI use cases. However, because it is highly regulated and lacks open-source datasets, there are not enough published resources on the production challenges, failures, and reasons behind them. Because of this, there is a growing technical debt regarding how AI is deployed in the FSIs. In addition to this, due to a lack of interdisciplinary skills in AI and the associated business risks, traditional risk managers and auditors struggle to create risk frameworks for AI deployments. In this paper, we will review the AI technical debt (TD) in FSIs and do an empirical study about the risks involved.

***K*eywords** Machine learning · Artificial Intelligence · AI Debt · Financial Services Industry · Banks · Insurance · AI Governance · ML Observability · AI Regulations · Explainable AI · AI Safety

1 Introduction

Financial Services as a term became mainstream after the reference in Gramm–Leach–Bliley Act of the late 1990s in the United States[1] allowed the creation of large financial conglomerates that offered various services under one roof. We use the financial services industry (FSIs) to represent financial institutions providing banking, insurance, payment, and investment services. Artificial Intelligence (AI) has seen unprecedented demand in FSIs in the last few years because of changing customer needs, improved preparedness, and evolving AI techniques, which have transformed traditional paper/human-based transactions into intelligent and data-based transactions. The benefits of AI teased FSIs for a very long time as other industries, like retail, Internet, etc., had been deploying AI at scale and were getting the benefits of AI [2]. In the last decade, FSIs have also undergone the digital transition [3], which involves migrating legacy to new-age software or cloud and modifying heavy paper-based processes into digital transactions. This made the industry capture a lot more data and a scalable software stack, which are the prerequisite requirements for deploying AI at scale[4].

The true success of AI deployments in other enterprise verticals is attributed to various factors, such as the availability of open-source/private data sets, rapid production feedback, fewer regulations, and the awareness of the use cases. Many of these factors can be attributed to the nature of business (retail, technology, supply chain, etc). However, highly regulated industries like FSIs are slow to deploy AI [2]. The reason could be regulatory and compliance challenges, closed data sets, lack of interdisciplinary skills required to design AI solutions, lack of maturity in AI architectures in FSIs etc. The current success of AI solutions in FSIs can be linked to the ability to learn from similar use cases in other verticals and deploy them in the FSIs (for example, recommendation systems, anomaly tracking, etc). However, other industries follow a very leisured risk framework as the business scale is prioritised over the risk/compliances. This inadvertently creates a Technical Debt in FSIs that could become a critical business continuity risk [5] resulting in loss of value or reputation or huge compliance troubles.

Ward Cunningham introduced Technical Debt (TD) 1992 as "shipping first-time code is like going into debt. A little debt speeds development so long as it is paid back promptly with a rewrite. . . The danger occurs when the debt is not repaid." [6]. This was later adapted to AI solutions in 2015 [7]. Sculley expanded the traditional software TD to Machine Learning and categorized various possible TDs in ML solutions. This became a reference for many future work [8] [9]. The proposed TD categorization is well-suited for traditional AI deployments. In this paper, we will review the AI Technical Debt in the Financial Services Industry (FSI) and empirically study the risks associated with these TD. Given that AI adoption is imperative in FSIs, creating a scalable AI solution to address technical debts can ensure reliable and sustainable adoption of AI in FSIs.

2 LITERATURE REVIEW

The adoption of Artificial Intelligence (AI) in the FSI is growing rapidly. By analyzing customer data, it can offer tailored product recommendations and improve customer satisfaction and loyalty in banks [10]. This also allows banks to understand customer preferences and behaviours mode deeply [11]. AI plays a crucial role in enhancing security measures within FSIs. ML is deployed to detect fraudulent activities by analysing transaction patterns and identifying anomalies [12]. The ability of these systems to analyze vast amounts of data in real-time and identify fraud significantly enhances the speed and accuracy of fraud prevention compared to traditional methods, thereby improving the overall efficiency of fraud management systems [11] [13]. Insurers have been exploring AI for various use cases like claims automation, pricing, underwriting, personalization, Data extraction, etc. In health insurance, ML was used to predict premiums based on features like age, gender, and lifestyle, achieving better accuracy and faster service delivery [14]. Jaskanwar compared the performance of various ML techniques like KNN, ANN, Random Forests, and SVMs in automating claims decisions for health insurance. In the property and casualty insurance sector, AI and ML prevent fraudulent claims, make swift business decisions, and enhance predictive analytics, improving customer satisfaction and reducing costs [15]. The implementation of AI-driven underwriting systems, such as the Hybrid Multiple Classifier System combining XGBoost, Random Forest, and SVM, has proven to be a cost-effective and time-saving solution, enhancing digital capabilities and customer focus[16]. AI's potential in fraud detection is immense, as it can analyze vast amounts of data to identify suspicious behaviour and develop predictive models tailored to detect fraudulent activities, thereby safeguarding the integrity of the insurance industry and protecting policyholders [17] [18]. Overall, the adoption of AI in FSIs is streamlining operations, reducing costs, and providing more personalized and efficient services to customers, ultimately transforming the industry landscape.

Ward Cunnigham first used technical debt (TD) to explain the long-term negative consequences of short-term compromises made in software development [6]. If not handled, these compromises can lead to significant damages and maintenance costs to the software in future. Effective Technical Debt Management (TDM) involves identifying, preventing, monitoring, prioritizing, and repaying TD to satisfy technical requirements and customer value, thereby preventing project failures and overruns [19]. Seaman et al. suggested a framework to manage TD with key components: TD identification, TD estimation, and decision-making. They provided options to categorize TD into - design/code debt, testing debt, documentation debt, defect debt and infrastructure debt. The resolutions may need humans or can be automated based on the TD type. However, it makes it easy to manage the business goals and track the pain points based on these categorizations. TD impact estimation was suggested to start with an estimated cost and arrive at the precise cost by refactoring occasionally. Estimating the TD impact cost will provide a metric to benchmark the urgency of TD resolution [19]. Despite its criticality, formal approaches to TD management are still not widely adopted, with less than 15% of practitioners using structured methods to manage TD [20]. Occasionally, accumulating TD can make maintenance nearly impossible or even the software redundant. So, addressing TD at the early stages is essential [6]. Therefore, understanding and managing TD is extremely important for software products.

Technical debt (TD) in Artificial Intelligence (AI) solutions is more complicated and multifaceted. TD in AI systems can arise from sub-optimal implementations, postponed tasks, and immature artefacts, which incur future costs in refactoring, rework, and maintenance [9] [6]. Algorithm debt refers to the sub-optimal implementation of algorithm logic, which can affect the performance and reliability of AI systems significantly [9]. Data testing debt refers to inadequate testing of input data, and reproducibility debt refers to a lack of consistency in reproducing the experimental results due to randomized algorithms and non-determinism in parallel learning [7]. Hidden debts are particularly concerning as they permute the risk of making future improvements more costly and complex[7]. Self-admitted technical debts (SATD) in AI are related to data dependency, code dependency, and awareness debts. Awareness debts, for instance, involve doubts about algorithmic procedures, design decisions, and proper API usage, which can lead to erroneous behaviour and increased maintenance efforts [8]. Tracking and visibility of TD for all stakeholders in the process is the key to managing TD properly [21]. Prioritising TD from a business perspective is important to business decision-making. [22]. Bogner et al proposed a new categorization of technical debts for general business along with anti-patterns [23].

AI is increasingly adopted in the financial services industry (FSI)s. Technical debt (TD) is a common issue in software development that can cause severe maintenance and usability challenges. Similarly, managing AI technical debt is crucial for the success of AI deployment. In highly regulated industries like FSIs, the damage is not just a failure of the AI; it can lead to severe damage to business and even business continuity risks.

3 Approach:

Categorizing and prioritising TD can ease the decision-making of the businesses and better risk management[22]. While there have been efforts to categorize TD in traditional environments, we aim to review TD for the Financial Services Industry (FSIs) and explore its risks. Unlike other verticals, FSIs have stringent regulatory compliance and a very high cost of failure. Making AI solutions acceptable in FSIs is very complex and needs multiple validations before the solution can be fully deployed in production environments in FSIs. Hence, documenting the TD with risks provides an acceptable framework for all stakeholders to validate and benchmark.

We will follow the following steps to arrive at this:

- Step 1: Propose the TD categorization
- Step 2: Empirical review of anti-patterns in each TD category
- Step 3: Reviewing the risks of the TD

4 Data inconsistencies between training and production pipelines:

4.1 Description:

When data pipelines and data formats shared between training and production environments differ, ML models suffer to maintain performance between training and production.

4.2 Risks:

This debt can manifest in various forms, including inefficiencies, maintenance challenges, and scalability issues, which can significantly impact the performance and reliability of machine learning systems. Many FSIs use legacy systems that are not entirely usable for AI integration. These are typically the workflow systems used for Core Banking, Loan Origination, Policy Admin, etc. To bypass this, FSIs use parallel systems like data lakes to provide training data instead of extracting it from the actual system. However, FSIs usually store limited data or processed data in these parallel systems, whereas the actual system captures different data. For example, the transaction details would typically override the old information with updated information in the data lake for use cases like the Claims process, where the claim management would've captured the transaction details each time of entry. These pipeline inconsistencies can cause the models to degrade more than usual in production. A similar debt is observed when FSIs migrate from one workflow system to another, where the previous data pipelines may not be well migrated to new systems, making the old models unusable with the new systems.

Given the complexity and ad-hoc nature of data processing workflows, it leads to a lack of standardisation and documentation that complicates maintenance and scalability. For instance, Xin et al. highlighted the challenges in managing data pipelines due to their intricate dependencies and the need for frequent updates to accommodate new data sources or changes in data schema [24]. In FSIs, ML models are built using diverse data sources, each with its own format and quality issues. Derakhshan et al. discussed how data heterogeneity creates complexities in the data processing and cleaning process, which becomes bottlenecks in the pipelines [25]. These require manual intervention and are prone to errors contributing to the debt. Not monitoring and debugging these pipelines risks the longevity of the AI solutions. And eventually, these become unscalable. Wu et al. highlighted that many existing pipelines are not designed with scalability in mind, leading to performance bottlenecks and increased latencies in data processing [26]. Addressing this debt requires standardized pipeline architectures, automated monitoring/debugging and design for scalability.

5 Excluding features without business oversight

5.1 Description:

Feature omission debt occurs when ML practitioners remove critical data features of the use case, often missing crucial information when providing predictions.

5.2 Risks:

Feature selection and omission for model building is often preferred through statistical observations rather than the opinions of business experts. This, in turn, created conflicts between the business experts and ML practitioners when selecting the final features. Simple observations like - highly missing values may exclude some of the essential features that can impact the outcome of the models. For example, in loan underwriting, some features are more than 90% nulls, like "Public Bankrupcies", which are mainly missing but are often the most critical features for an underwriter to make the final decision.

Excluding such features can lead to a technical debt where long-term costs outweigh the short-term gains of a simpler model. These debts can result in reduced model performance, increased risk of bias, and the need for frequent updates to the model to maintain accuracy and compliance[27][28]. Such exclusions can sometimes cause less accurate predictions. For instance, excluding features that capture the nuanced customer behaviour might simplify the model but can be less accurate in predicting credit risk [29]. This can get further compounded due to the dynamic nature of FSIs. The relevancy of features may increase, necessitating the inclusion of these omitted features in future models. This requires FSIs to continuously evaluate and update their models, which can be resource-intensive and complex [30][31]. Additionally, the legacy systems and data silos can limit the availability and quality of data for model training. This can lead to excluding important features, further exacerbating feature omission debt [32]

6 Over-engineering the features

6.1 Description:

When the raw features are highly engineered to create synthetic features that overly simplify or complicate the features used in the modelling.

6.2 Risks:

Raw features are processed through feature engineering techniques to create synthetic variables and are used in model training and inferencing. If such feature engineering techniques overcomplicate or capture limited information from raw features, ML models trained on these features may not effectively learn from the data and may miss out on critical information. One common area where this is more frequent is in processing "categorical" features or high cardinal features. ML practitioners, particularly in FSIs, ignore the categorical features and focus on numerical features, as many classic ML models can not take advantage of these categorical features, such as Xgboost, linear regression models, etc. This creates both feature omission debt and feature engineering debt.

ML practitioners use techniques like label encoding or one-hot encoding in high "cardinal" features. This creates very sparse data for very high cardinal features. To address this, ML practitioners will only focus on the top frequent values in a feature, encode only these values and encode the rest of the values into one single category, thereby reducing the dimensionality. This, in turn, limits the information captured in these synthetic features, resulting in poor model learning. Effective feature selection can mitigate feature engineering debt by ensuring that the most informative features are used and also reduce the model complexity and overfitting [33] [34]. Overcomplicating the feature engineering makes it challenging to explain the model functioning as traditional ML explainability uses feature importance as an explanation. It becomes hard to retrace the feature importance of complicated synthetic features and link them back to raw features. This further creates challenges for regulatory compliance and risk management. Integrating domain knowledge and expert input during feature engineering can enhance the relevance and quality of features selected and reduce the likelihood of debt.

7 Not testing enough!

7.1 Description:

It refers to testing models on limited or less data that doesn't reflect the full complexity of the production environment.

7.2 Risks:

Model testing in financial services is a critical component that ensures the reliability, accuracy, and robustness of financial models. Model robustness is a regulatory requirement for credit scoring models and other models [35][36][37]. By subjecting models to various stress tests, FSIs can demonstrate the robustness of their models to stakeholders and regulators

Finding sufficient training data is critical for building AI solutions. However, selecting the right test data is extremely important in FSIs. Testing data without a reasonable sample distribution of multiple scenarios would bias the models towards the provided samples during testing. This can lead to critical performance issues for the business. Stress testing the models will allow the users to find the gaps and design the right risk policies. Stress testing is also to be performed before each future model update. However, given the dynamic nature of the business and the models, the same testing strategy may not be ideal as the corner cases would have changed post-retraining or model updates. Driss El Maanaoui et al. stressed the need for continuous testing and validation to ensure the models remain relevant and accurate over time [38]. This adaptability is vital for maintaining the relevance and accuracy of financial models over time. Breeden discussed the role of model testing in improving the interpretability of complex financial models[39].

8 The Explainability Debt: Risks of Opaque or Misunderstood ML Models

8.1 Description:

Being unable to explain ML predictions or using a wrong explainability model creates explainability debt that has serious downstream risks.

8.2 Risks:

ML explainability refers to explaining the decision-making process of an ML model or understanding the model's functioning. It should help us understand why and how the system made a particular decision. Kuiper et al [40] interviewed 3 applications (consumer credit, credit risk, and anti-money laundering) in FSIs to understand the importance of explainability in AI solutions. There was a clear need for explainability in AI solutions used in FSIs. However, the urgency of the models in production made ML practitioners productionize the models without proper explainability around these decisions. The World Economic Forum in 2019 [41] notes that the opacity of AI poses serious risks for FSIs and can lead to loss of control because of a lack of transparency. This can lead to serious damage to consumer trust and confidence.

Local and model-agnostic techniques like SHAP and LIME are widely used for the explainability of individual predictions of black box models [42]. The LIME technique tries to understand the association between a specific example's features and the model's prediction by training an increasingly explainable model like the linear model with illustrations resulting from small changes to the original input (Dhinakaran [43]). At the end of the training, explanations can be found from the characteristics by which the linear model acquired co-efficient, irrespective of specific thresholds, after factoring in some normalization. The SHAP technique aims to calculate the contribution of every feature to the prediction to ascertain the impact of each input. SHAP explainability principles, entrenched in cooperative coalition game theory, generate Shapley values.

One of the most common challenges faced while using these post-hoc techniques is the lack of "stability". According to Zafar & Khan [44], lack of stability refers to explanation level uncertainty by which locally interpretable models encompass an uncertainty level linked to it due to black-box model simplification. In addition, the generated explanations can differ with varying selected features and feature weights due to the random nature of perturbation in LIME. In critical application areas such as financial services, generating inconsistent explanations could prove troublesome [44]. When features are correlated, the SHAP value method encounters the inclusion of unrealistic data instances. Elizebath et al [45] discussed mathematical issues like conditional versus interventional distributions and additive constraints raised using Shapely values in detail. Leon et al. observed that modified Back Propagation methods like Deep Taylor Decomposition, Layer-wise Relevance Propagation (LRP), Excitation BP, PatternAttribution, DeepLIFT, Deconv, RectGrad, and Guided BP are independent of the parameters of later layers questioning the faithfulness of the explanations generated using these methods [46].

Sometimes, these explainability techniques can fool the users too. Slack et al. [42] demonstrated that LIME and SHAP are unreliable as they proposed a novel scaffolding attack that hides the biases of a given classifier by allowing an adversarial entity to craft an arbitrary desired explanation. In a different study, Slack et al. [47] also demonstrated how Counterfactual explanation methods (CEM) could be manipulated. The adversarial models appear fair when evaluated using standard counterfactual explanations but provide much lower-cost recourse for specific subgroups when slightly perturbed. Kumar et al [48] presented Shapley residuals as a warning mechanism to alert practitioners to model complexities and essential interactions that may be missed when relying solely on Shapley values for interpretability.

This questions the faithfulness and stability of explainability methods. While most regulations mandate using explainable AI methods to decode "black box" models, the standards are not defined in terms of the accuracy of these explanations. "Explainability" itself is not a technical term, which means the outcome is open for interpretation [49]. Krishna et al. studied the disagreement problem in explainability from the responses of 25 data scientists from the technology and FSI sectors. It was observed that 84% of respondents admitted that they faced explainability disagreements regularly and many practitioners admitted to using ad hoc heuristics or being uncertain about resolving explanation disagreements in practice.

Given regulatory and compliance requirements, this poses a greater risk since any wrong explanation could lead to more confusion and setbacks for users and the community. Generating consistent and reliable true-to-model explanations can make debugging easier and deliver confidence to the user.

9 Drift Monitoring Debt

9.1 Description:

When models and data are poorly monitored during production, this leads to substantial model performance degradation due to drifts.

9.2 Risks:

There are multiple reasons why an AI system fails in production. The common issues that cause ML models to degrade in production are data drifts and model/concept drifts.

Data drift occurs when the data used in training viz-a-viz data used in production changes over time. This can occur for various reasons, such as evolving business/consumer behaviour, regulatory changes, market volatility, etc. For instance, Christensen et al. highlight the impact of high-frequency trading and market microstructure noise, which can introduce drift in financial data streams, affecting the accuracy of volatility estimations and risk assessments [50]. Pugliese et al. discuss the challenges posed by data drift in credit scoring models, where shifts in consumer behaviour or economic conditions can lead to model degradation over time [51]. Liu et al. focus on the role of anomaly detection in identifying data drift, particularly in fraud detection systems within financial services [52]. Anomalies can indicate shifts in transaction patterns, which may signal fraudulent activities or changes in consumer behaviour, necessitating timely intervention.

Model drift, also known as model decay or concept drift, refers to "the decay of a model's predictive power as a result of alterations in the environment" [53]. Model/Concept drift occurs when the relationship between input and target variables changes. The model provides inaccurate predictions since the definition of what it attempts to predict changes. The change can be gradual, sudden, or recurring. Li and Zhao observed that concept drift can lead to model degradation, necessitating frequent updates to maintain performance [54]. Similarly, Cavalcante et al. discussed how financial time series models are particularly susceptible to concept drift, resulting in significant prediction errors if not properly addressed [55]. Moreover, concept drift complicates risk management strategies. Neri points out that financial institutions rely heavily on historical data to assess risk, but concept drift can render these assessments obsolete, leading to potential financial losses [56]. Marques emphasizes the importance of adaptive algorithms in financial forecasting, noting that traditional models often fail to account for non-stationary data, leading to suboptimal predictions [57]. Lin et al. also highlighted the difficulty of distinguishing between noise and genuine concept drift, which can lead to unnecessary model updates and increased operational costs [58].

In conclusion, drift monitoring debt significantly impacts financial services by affecting the accuracy of models and complicating risk management. Significant delay in drift monitoring can make the model predictions redundant and may result in the complete failure of the AI solutions. Addressing these impacts requires adaptive modelling techniques and robust drift monitoring mechanisms. However, implementing these solutions must balance the need for accuracy with the associated computational and operational costs. As financial markets continue to evolve, ongoing research and innovation in handling concept drift will be crucial for maintaining the efficacy of financial models and strategies.

10 Bias/Fairness Debt.

10.1 Description:

Deploying or using biased ML models without addressing bias in the models.

10.2 Risks:

Biased or unfair models can lead to unfair treatment of individuals, perpetuate existing inequalities, and result in financial losses or reputational damage for companies. These models can lead to discriminatory practices and risk accumulation for one class of customers. Bias/unfairness can be induced into the model when trained on historical data reflecting societal biases[59]. This can be prominently seen in use cases like lending, where certain demographic groups are systematically disadvantaged in terms of loan approvals and interest rates[60] or in Insurance, where biased models may inaccurately assess the risk associated with certain individuals or groups, leading to unfair premium pricing [61]. Such behaviours can exacerbate existing economic disparities and limit marginalized communities' access to financial resources.

In addition to creating operating risks, this can also create huge compliance risks. EU AI act in Europe is one of the first regulations that is specifically defined for AI usage in Low, Medium and High risk use cases. The Act introduces the concept of algorithmic audits to regularly assess AI systems for bias. These audits must be conducted to ensure compliance with the established standards and to identify any discriminatory outcomes that may arise from the use of the AI systems [62]. The Algorithmic Accountability Act in the U.S., mandates that companies conduct impact assessments of automated decision systems, including potential biases[63]. The Federal Trade Commission (FTC) has issued guidelines emphasizing the importance of transparency, accountability, and fairness in algorithmic decision-making. These guidelines encourage companies to conduct regular audits of their algorithms to identify and mitigate bias [64]. In India, RBI released a new circular on "Model Risk Management" of credit risk models probing on using unbiased models [36]. While the regulations enforce the usage of the process, the significant challenges remain in the lack of standardized metrics for measuring algorithmic bias. The National Institute of Standards and Technology (NIST) in the US has been working on a framework to improve the trustworthiness of AI systems, which includes addressing bias. This framework encourages organizations to adopt practices that ensure AI systems are fair, accountable, and transparent [65] . However, regulating the bias for these models is a major challenge. As discussed in "Explainability debt" it is possible to do scaffolding attacks on explainability methods to fool the auditors checking the bias.

11 Auditability debt.

11.1 Description:

Deploying or using AI solutions without capturing the right artefacts and following the right audit framework.

11.2 Risks:

As ML models are being used for increasingly complex and sensitive use cases in FSIs, the ability to audit is critical for managing risks and delivering regulatory compliance. The complexity of the ML models makes auditing challenging because of the vast amounts of data and model changes over time. As a result, organizations may struggle to identify and rectify issues, leading to an accumulation of audit debt that can undermine the reliability and trustworthiness of their systems [66]. One of the primary concerns with audit debt is the lack of transparency and accountability in algorithmic systems. Algorithms often operate as "black boxes," making it difficult to understand their decision-making processes. This opacity can lead to unchecked biases and errors, accumulating over time and resulting in significant audit debt. External auditors of AI systems are used to deliver external AI governance. But the effectiveness of the auditing depends heavily on accessibility to the model and model artificats[67]. Recent audits are limited only to query access to the models and observe the output. This highly limits the auditor's ability to perform stronger attacks, more thoroughly interpret models, and conduct fine-tuning. In addition to this, outside-the-box access to training and deployment information like methodology, code, documentation, data, deployment details, findings from internal evaluations, etc, allows auditors to scrutinize the development process and design more targeted evaluations.

Several strategies have been proposed to address audit debt in algorithms. One approach is the development of standardized auditing frameworks that provide guidelines for evaluating algorithmic systems. These frameworks can help ensure consistency and thoroughness in audits. Deborah et al. [68], suggested a comprehensive end-to-end audit framework that produces a set of documents which, when combined, will create an audit report that aligns with organizational values and principles accessing the fit of decisions made throughout the process. The framework follows different stages: scoping, mapping, artefact collection, testing, and reflection (SMACTR). Akula and Garibay ([69]) proposed a 7-phase audit depending on the access. The phases are Process access, Model access, Input access, Output access, Parameter control, Learning objective and white-box models. While artefacts for algorithm auditing are clear, the execution varies from case to case.

12 Models are prone to attacks

12.1 Description:

Deploying/using ML models with vulnerabilities and being open to attacks.

12.2 Risks:

ML models are prone to attacks that can compromise the integrity, confidentiality, availability and functioning of the AI solution. One prominent category of attacks is adversarial attacks, where malicious inputs are crafted to deceive ML models into making incorrect predictions. Xiangyu Qi et al. discuss the vulnerabilities of AI models to adversarial inputs, which can lead to incorrect predictions and decisions [70]. These adversarial attacks can be classified into white-box and black-box attacks. In White-box attacks, the attacker will have full access to the model and in black-box attacks, the access is limited to only query-access [71]. These attacks are highly significant in image classification where a slight perturbation to the input image can change the final prediction[72] [73]. Model poisoning attack refers to manipulating the training data to embed malicious behaviour into the model, leading to backdoor access to models to trigger specific output with specific inputs[74]. These can become really intrusive in federated learning environments, making it challenging to ensure data integrity [75]. Membership inferencing attacks aim to determine whether a specific data point was used in the model's training data, thereby compromising the data privacy [76]. Model extraction attack refers to training a surrogate model using the input and the output of the model to approximate the model functionality. This can further allow other attacks like adversarial or poisoning attacks using the extracted model [77].

Due to the sensitive nature of the business, protecting the ML models in production is crucial for FSIs. Training the models with adversarial examples can help them recognize and mitigate the malicious inputs designed to deceive them[78]. Using secure multi-party computation (SMPC) and homomorphic encryption allows the processing of encrypted data without revealing sensitive information. This can protect the data confidentiality and still enable ML models for decision-making [79]. Detecting anomalies in ML pipelines helps to identify attacks and provides an additional layer of defence against known and novel attacks [80]. Regular model audits and updates are essential to maintain the security and integrity of the AI solution. Balancing security with model performance and computational efficiency is a persistent challenge, as securing ML models may require more resources and impact the latencies of model inferencing. Deploying models without addressing model safety can carry a substantial debt for the FSIs.

13 Using shadowy pre-trained models

13.1 Description:

Using pre-trained models without sufficient documentation on the training data and the sources.

13.2 Risks:

While this is prominent in generative AI models where the model outputs can create copyright and trademark issues, it is also applicable in use cases where FSIs use pre-trained models, like in Credit rating where alternate data sets are used for credit scoring models. Using models that were trained on unauthorized data violates the data privacy laws of the land, such as the General Data Protection Regulation (GDPR) in Europe, the California Consumer Privacy Act (CCPA) in the United States, or the DPDP Act in India. These regulations mandate strict guidelines on data usage, and non-compliance can result in hefty fines and legal actions against financial institutions. Unauthorized usage of such data can lead to a breach of contractual obligations [81]. Such models also lack the accountability needed for FSIs.

14 Delayed, unchecked or no feedback to ML models

14.1 Description:

When ML models are used without any feedback in production or when the feedback is delayed or not validated correctly.

14.2 Risks:

ML models can iterate using feedback and adapt in a production environment. However, when such feedback is not provided or delayed, these models degrade over time and lack relevance in production. The nature of the business could also cause delays in feedback. In loan underwriting, while the decision can be made in real-time, the quality of the underwritten customer is only known after many months. Insurance companies rely heavily on ML for underwriting, claims processing, and customer service. In Health Insurance, claims can arise after years of underwriting the application. In life insurance, a death claim can be registered after decades. In such cases, it is hard to validate the efficacy of the models, and such debt can lead to the accumulation of tremendous risk over time. This can impact the decision-making and risk assessment of FSIs. One of the challenges of managing feedback debt is the complexity involved in integrating with diverse data sources. FSIs deal with vast amounts of structured and unstructured data, which could be time-consuming and complex to analyze and process the feedback in real-time. This can exacerbate feedback debt as the resources may not be readily available.

At the same time, sharing unchecked feedback can cause the models to learn inconsequential patterns and it would be very hard to track, rectify and fix such erroneous learning. Hence, the feedback loop in ML is critical, and any errors would propagate through the systems, leading to compounding errors over time[82]. Wyllie et al. highlighted that adversarial attacks can also manipulate the feedback data, intentionally skewing model output to favour specific outcomes[83]. The impact of incorrect feedback is also seen in model bias and fairness. Gupta et al. discussed how feedback can reinforce existing biases within the model [84]. There is also a strong need to create organisational changes on communication between the data scientists, engineers and domain experts to correctly interpret feedback data and make informed decisions on model updates [85].

15 Reproducibility Debt

15.1 Description:

Failure to replicate the results of ML models consistently.

15.2 Risks:

FSIs strive to maintain consistency and stability across the process. Risk management is built on delivering stability to the organization. Using an ML model without fully documented artefacts can result in huge reproducibility debt. One of the issues was a lack of standardization, documentation, poor version controls, and poor audit or governance frameworks. This is exacerbated by using complex models and proprietary datasets, which are not always available during validation [86] [87]. ML models typically involve multiple configurations that need thorough documentation of data reprocessing steps, hyperparameter settings and training procedures. The common reasons for the lack of documentation are time constraints and pressure to deploy models in production [88] [89]. Furthermore, FSIs operate a complex array of systems that often involve complex interactions between systems, creating a unique flow for each integration. Without standardization, documentation and continuous capturing of artefacts, it only adds more complexity and increases reproducibility debt. While this looks like an operations issue, failing to recreate the scenarios during an audit can lead to non-compliance or even higher penalties for violation of regulations.

16 Compliance & Governance Debt

16.1 Description:

Compliance & governance debt refers to using AI solutions without a regulatory or governance framework because of a lack of clarity in regulations or internal governance.

16.2 Risks:

AI is fairly new, and the techniques are evolving rapidly. It is very hard for regulators to keep up with the speed of AI innovations. These gaps arise from the rapid advancement of AI technologies, which often outpace the development of corresponding regulatory measures. Even today, there is no standardization of benchmarking or definitions. Also, there is no standardization between regulating bodies within the same geography or between geographies. For example, in the US, anti-model laundering (AML) and fraud prevention models are heavily regulated, whereas in Europe, this is not the case [90]. This disparity creates challenges for FSIs to create complete foolproof AI solutions. Moreover, the regulatory system fails to capture the dynamic nature of AI systems, which poses challenges for traditional regulatory approaches that are static and prescriptive. As these AI systems become autonomous, there is a need for adaptive regulatory mechanisms that can respond to changes in AI capabilities [91]. There is also a gap in the liability and accountability of AI decisions. There is no clarity on the liability when AI systems cause harm or make erroneous decisions [92]. So, FSIs need to deploy ML models that are compliant with internal AI governance standards that could align with possible regulations in the future and realign the internal AI governance with changes in regulations from time to time.

17 Stakeholder Debt: The Hidden Cost of Limited Participation in AI Development

17.1 Description:

When AI solutions are built and deployed with participation from only one/limited stakeholder group, they result in high dependency on that group.

17.2 Risks:

In FSIs, multiple process owners exist for a use case, such as data, IT, operations, risk, revenue, etc. For any given use case, all critical information about the ML model must be provided to all stakeholders to ensure transparency and validation from all users in approach and usage. However, because ML is highly technical, only Data Scientists or ML practitioners will often own the entire process and the other stakeholders are often limited to just the final usage. This creates a gap in understanding the AI solutions with the other stakeholders. A few issues, like Feature Selection debt, could be solved by involving all the stakeholders at each stage. Similar to generating explainability (local or global), the user group can validate the sufficiency of the explainability. This issue is also caused by the inability to provide the information in a language other stakeholders can understand, making it intrinsically dependent on ML practitioners.

Zhang et al. explored the role of participatory design in ML, suggesting the implementation of iterative feedback with stakeholders and the need to address potential issues early in the development process [93]. Yakota et al. suggest that stakeholder input is vital for identifying and prioritizing system features that align with the user's needs [94]. This reduces the misalignment of system functionality. Communication and transparency between stakeholders regarding system capabilities and limitations are also required so that there are fewer misunderstandings and misalignments in the deliverables. But at the same time, balancing the stakeholder's interests and noting the conflicting priorities can lead to compromises that may not satisfy all staekholders. Addressing them through structured negotiations and consensus-building is essential[95]. Enabling such a flow of information and knowledge about the AI solution is not scalable without proper tools. Castro-Herrera et al. stressed the need for tools and frameworks that facilitate stakeholder engagements through the ML lifecycle [96]. These tools can streamline communication and ensure that stakeholders' inputs are integrated systematically.

18 Model Risk Managing(MRM) Debt.

18.1 Description:

Model risk management debt occurs when FSIs use AI solutions with poorly tested model risk frameworks in production.

18.2 Risks:

Effective model risk management(MRM) should safeguard organizational objectives, shield/protect business assets, and warrant financial sustainability. This involves setting up feedback loops and performance metrics that allow for the ongoing assessment of model usage risk and the timely identification of any emerging risks. Various regulatory bodies like the Federal Reserve Supervisory in 07-2011 [37], Bank of England in 06-2022 [35], and Reserve Bank of India in 08-2024 [36] released model risk management guidelines. However, these are more applicable to traditional models.

Traditional MRM involves a more structured process of evaluating model assumptions and testing its performance and robustness. For instance, the definition of a model given by the Federal Reserve Supervisory states: "model refers to a quantitative method, system, or approach that applies statistical, economic, financial, or mathematical theories, techniques, and assumptions to process input data into quantitative estimates. A model consists of three components: an information input component, which delivers assumptions and data to the model; a processing component, which transforms inputs into estimates; and a reporting component, which translates the estimates into useful business information. . . . The definition of the model also covers quantitative approaches whose inputs are partially or wholly qualitative or based on expert judgment, provided that the output is quantitative in nature." However, in machine learning, the models learn from the data, and this data serves as the source of the functions that emerge from the ML models. Similarly in RBI's guidelines suggest that the models should be explainable, but as discussed in the explainability debt, there is no standard for "explainability" [49].

In contrast to traditional MRM, MRM in ML is complex due to the nature of ML models, which are often non-linear and involve high-dimensional data. The opacity of ML models necessitates new techniques for risk management. This paper highlights multiple risk sources when deploying AI in FSIs, such as biased models, poor drift monitoring setup, poor testing, lack of or inadequate explainability, data processing debt, model safety, etc. The model risk management (MRM) framework should address these concerns when deploying ML models in FSIs. Moreover, ML models are highly sensitive to the quality and quantity of data, making data management a crucial aspect of MRM in ML. Issues such as data bias, data drift, and overfitting are prevalent, requiring continuous monitoring and updating of models to ensure their reliability and accuracy in ML models. In contrast, traditional models are largely static and require significantly less frequent changes. Traditional models are subject to well-established regulatory requirements, whereas ML models still navigate a developing regulatory environment. This lack of clear guidelines can challenge organisations when implementing effective MRM practices for ML models, thereby creating an MRM debt.

To track the framework's effectiveness, it is also essential to identify the right metric for benchmarking each risk criterion. For example - how to quantify bias? If there is a greater risk of gender bias in the models, fine-tuning the metrics to reflect the sensitivity of the risk is needed. Unlike the traditional risk management model followed in FSI, MRM on AI solutions is more interdisciplinary and complex. It sometimes has technical dependencies like interpretability and transparency of AI models, which are vital components for designing risk management.

19 CONCLUSION

AI Technical Debt (TD) is gaining momentum, as it documents various gaps in the ML system design that could cause the system failure. This paper focused primarily on reviewing TDs in the Financial Services Industry (FSI). FSIs are highly regulated, and any failure can cause severe financial damage, create enormous regulatory challenges and damage the customers' trust. In FSIs, AI is undoubtedly proving its value due to applications ranging from fighting financial crime, automating multiple critical decisions (underwriting, claims, product construct, etc.) and assisting with innovative digital experiences for clients. As AI gets mainstream and starts being deployed into multiple operations, any failure is not just an operational risk but a business continuity risk. Hence, financial institutions should formulate appropriate steps to guarantee the stability of these systems and manage the risks of using AI systems.

In this paper, we have highlighted the risks of Technical Debt (TD) in AI solutions arising from compromises or poor design of the AI systems in FSIs. The reasons for these TDs are multifaceted and complex. However, certain debts are highly critical for FSIs, and it is important for FSIs to design AI systems addressing these TDs. TDs, like explainability debt, are pretty complex because of non-standardization in the definition and the risk of using wrong explainability models. Model risk management debt is incredibly different from traditional ML and is highly complex and evolving. Other TDs, like Feature Engineering debt, Feature exclusion debt, Feedback debt, etc, carry heavy risks but are addressable. Compliance and regulatory debt are underdeveloped, but regulators try to introduce more explicit regulations over time. Hence, FSIs should focus on integrating an internal governance framework to align and meet the governance expectations. Finally, TDs like model attacks are new and need more research to explore the gaps in ML and provide insights on addressing this TD.

20 Future Work

While this paper focused on FSIs, many of the components apply to any other highly regulated industry, such as Health care, Manufacturing, Government, etc. We hope this paper will encourage additional development in Technical Debt (TD) areas in highly regulated industries, where mission-critical use cases are scrutinized more and involve very high risk. There is no one-size-fits-all when it comes to deploying AI. Hence, it is essential to identify the risks and pitfall areas so that the industry can work to create solutions, tools, and products that effectively address these challenges.

References

[1] Cara S. Lown, Carol L. Osler, Philip E. Strahan, and Amir Sufi. The changing landscape of the financial services industry: what lies ahead? *Economic Policy Review*, pages 39–54, October 2000.

[2] Carsten Maple, Lukasz Szpruch, Gregory Epiphaniou, Kalina Staykova, Simran Singh, William Penwarden, Yisi Wen, Zijian Wang, Jagdish Hariharan, and Pavle Avramovic. The ai revolution: opportunities and challenges for the finance sector. *arXiv preprint arXiv:2308.16538*, 2023.

[3] Christian Eckert and Katrin Osterrieder. How digitalization affects insurance companies: overview and use cases of digital technologies. *Zeitschrift für die gesamte Versicherungswissenschaft*, 109(5):333–360, 2020.

[4] Deborah Leff and Kenneth T. K. Lim. *The key to leveraging AI at scale*, pages 171–175. Springer Nature Switzerland, Cham, 2023.

[5] Narayan Ramasubbu and Chris F. Kemerer. Technical debt and the reliability of enterprise software systems: A competing risks analysis. *Management Science*, 62(5):1487–1510, 2016.

[6] Ye Yang, Dinesh Verma, and Philip S Anton. Technical debt in the engineering of complex systems. *Systems Engineering*, 26(5):590–603, 2023.

[7] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. *Advances in neural information processing systems*, 28, 2015.

[8] David OBrien, Sumon Biswas, Sayem Imtiaz, Rabe Abdalkareem, Emad Shihab, and Hridesh Rajan. 23 shades of self-admitted technical debt: an empirical study on machine learning software. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2022, page 734–746, New York, NY, USA, 2022. Association for Computing Machinery.

[9] Emmanuel Iko-Ojo Simon, Melina Vidoni, and Fatemeh H. Fard. Algorithm debt: Challenges and future paths. In *2023 IEEE/ACM 2nd International Conference on AI Engineering – Software Engineering for AI (CAIN)*, pages 90–91, 2023.

[10] Galib Reza. An analysis of ai tools adoption in finance industry. *Poonam Shodh Rachna*, 2024.

[11] Dr B Vinothkumar. Banking innovations through artificial intelligence. In *Futuristic Trends in Artificial Intelligence Volume 3 Book 2*, pages 69–79. Iterative International Publishers, Selfypage Developers Pvt Ltd, May 2024.

[12] Lawrence Damilare Oyeniyi, Chinonye Esther Ugochukwu, and Noluthando Zamanjomane Mhlongo. Implementing ai in banking customer service: A review of current trends and future applications. *International Journal of Science and Research Archive*, 2024.

[13] Geetha Manoharan, G Nithya, K Rajchandar, Abdul Razak, Swati Gupta, Subhashini Durai, and Sunitha Prurushottam Ashtikar. AI in finance and banking. In *Advances in E-Business Research*, pages 1–28. IGI Global, May 2024.

[14] Keshav Kaushik, Akashdeep Bhardwaj, Ashutosh Dhar Dwivedi, and Rajani Singh. Machine learning-based regression framework to predict health insurance premiums. *International Journal of Environmental Research and Public Health*, 19(13):7898–7898, 2022.

[15] Muralikrishna Dabbugudi. Artificial intelligence on property and casualty insurance. *European Journal of Electrical Engineering and Computer Science*, 6(6):26–30, 2022.

[16] Chun Lei He, Dave Keirstead, and Ching Y Suen. A hybrid multiple classifier system applied in life insurance underwriting. In *Pattern Recognition and Artificial Intelligence*, Lecture notes in computer science, pages 171–176. Springer International Publishing, Cham, 2020.

[17] Dharmendra Singh, Rasha Ahmed Al Mamari, Amal Khalil Al-Zadjali, and Omar Abdulaziz Al Ansari. Fraud in insurance and the application of artificial intelligence (AI) in preventing fraud. In *Transforming the Financial Landscape With ICTs*, pages 134–164. IGI Global, May 2024.

[18] Venkata Ramana Saddi, Swetha Boddu, Bhagawan Gnanapa, Nasmin Jiwani, and T. Kiruthiga. Leveraging big data and ai for predictive analysis in insurance fraud detection, 2024.

[19] Carolyn Seaman and Yuepu Guo. Measuring and monitoring technical debt. In *Advances in Computers*, volume 82, pages 25–46. Elsevier, 2011.

[20] Danyllo Albuquerque, Everton Tavares Guimaraes, Graziela Simone Tonin, Mirko Barbosa Perkusich, Hyggo Almeida, and Angelo Perkusich. Perceptions of technical debt and its management activities - a survey of software practitioners. In *Proceedings of the XXXVI Brazilian Symposium on Software Engineering*, SBES '22, page 220–229, New York, NY, USA, 2022. Association for Computing Machinery.

[21] Nanette Brown, Yuanfang Cai, Yuepu Guo, Rick Kazman, Miryung Kim, Philippe Kruchten, Erin Lim, Alan MacCormack, Robert Nord, Ipek Ozkaya, et al. Managing technical debt in software-reliant systems. In *Proceedings of the FSE/SDP workshop on Future of software engineering research*, pages 47–52, 2010.

[22] Rodrigo Rebouças de Almeida, Rafael do Nascimento Ribeiro, Christoph Treude, and Uirá Kulesza. Business-driven technical debt prioritization: An industrial case study. In *2021 IEEE/ACM International Conference on Technical Debt (TechDebt)*, pages 74–83, 2021.

[23] Justus Bogner, Roberto Verdecchia, and Ilias Gerostathopoulos. Characterizing technical debt and antipatterns in ai-based systems: A systematic mapping study. In *2021 IEEE/ACM International Conference on Technical Debt (TechDebt)*. IEEE, May 2021.

[24] Doris Xin, Hui Miao, Aditya Parameswaran, and Neoklis Polyzotis. Production machine learning pipelines: Empirical analysis and optimization opportunities, 2021.

[25] Behrouz Derakhshan, Alireza Rezaei Mahdiraji, Zoi Kaoudi, Tilmann Rabl, and Volker Markl. Materialization and reuse optimizations for production data science pipelines. In *Proceedings of the 2022 International Conference on Management of Data*, SIGMOD '22, page 1962–1976, New York, NY, USA, 2022. Association for Computing Machinery.

[26] Jiang Wu, Hongbo Wang, Chunhe Ni, Chenwei Zhang, and Wenran Lu. Data pipeline training: Integrating automl to optimize the data flow of machine learning models, 2024.

[27] Kartik Athreya and Hubert P. Janicki. Credit exclusion in quantitative models of bankruptcy: Does it matter?, 2006.

[28] Majid Bazarbash. Fintech in financial inclusion: Machine learning applications in assessing credit risk, 2019.

[29] Ayushi Pillay, Monika Arya, Sumit Kumar Sar, Bhupesh Kumar Dewangan, Tanupriya Choudhury, Ketan Kotecha, and Sanjana Dewangan. ML-based feature selection technique for imbalanced data streams. In *2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. IEEE, October 2023.

[30] Mark Drakeford and Darshan Sachdev. Financial exclusion and debt redemption. *Critical Social Policy*, 21(2):209–230, 2001.

[31] Barbu Bogdan Popescu and Lavinia Ştefania Ţoţan. Econometric modeling of banking exclusion, 2013.

[32] Khanh Van Nguyen, Md Rafiqul Islam, Huan Huo, Peter Tilocca, and Guandong Xu. Explainable exclusion in the life insurance using multi-label classifier. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2023.

[33] Shengbo Chang, Changqi Wang, and Cheng Wang. Automated feature engineering for fraud prediction in online credit loan services. In *2022 13th Asian Control Conference (ASCC)*, pages 738–743, 2022.

[34] Nilesh Kumar Sahu, Manorama Patnaik, and Itu Snigdh. Feature engineering for various data types in data science. In *Advances in Data Mining and Database Management*, pages 1–16. IGI Global, 2021.

[35] Bank of England. Ps6/23 – model risk management principles for banks, 17 May 202317 May 202317 May 2023.

[36] Reserve Bank of India. Regulatory principles for management of model risks in credit, 08 August 2024.

[37] Federal Reserve System. Sr 11-7: Guidance on model risk management, 04 April 2011.

[38] Driss El Maanaoui, Khalid Jeaab, Hajare Najmi, Youness Saoudi, and Moulay El Mehdi Falloul. Machine learning in finance case of credit scoring. In *Lecture Notes in Networks and Systems*, Lecture notes in networks and systems, pages 8–16. Springer Nature Switzerland, Cham, 2024.

[39] Joseph L Breeden. Validation of stress testing models. In *The Analytics of Risk Model Validation*, pages 13–25. Elsevier, 2008.

[40] Ouren Kuiper, Martin van den Berg, Joost van der Burgt, and Stefan Leijnen. Exploring explainable AI in the financial sector: Perspectives of banks and supervisory authorities. In *Communications in Computer and Information Science*, pages 105–119. Springer International Publishing, 2022.

[41] Galaski R. McWaters R., Blake M. Navigating uncharted waters: A roadmap to responsible innovation with ai in financial services, 23 October 2019.

[42] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods, 2020.

[43] Dhinakaran. What are the prevailing explainability methods?, 2021.

[44] Muhammad Rehman Zafar and Naimul Khan. Deterministic local interpretable model-agnostic explanations for stable explainability. *Machine Learning and Knowledge Extraction*, 3(3):525–541, 6 2021.

[45] Indra Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Eduardo Scheidegger, and Sorelle A. Friedler. Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, 2020.

[46] Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why modified BP attribution fails. *CoRR*, abs/1912.09818, 2019.

[47] Dylan Slack, Sophie Hilgard, Himabindu Lakkaraju, and Sameer Singh. Counterfactual explanations can be manipulated. *CoRR*, abs/2106.02666, 2021.

[48] Indra Elizabeth Kumar, Carlos Eduardo Scheidegger, Suresh Venkatasubramanian, and Sorelle A. Friedler. Shapley residuals: Quantifying the limits of the shapley value for explanations. In *Neural Information Processing Systems*, 2021.

[49] Leilani H. Gilpin, Andrew R. Paley, Mohammed A. Alam, Sarah Spurlock, and Kristian J. Hammond. "explanation" is not a technical term: The problem of ambiguity in xai, 2022.

[50] Kim Christensen, Roel Oomen, and Roberto Renò. The drift burst hypothesis. *Journal of Econometrics*, 227(2):461–497, 2022.

[51] Victor Ulisses Pugliese, Renato Duarte Costa, and Celso Massaki Hirata. Comparative evaluation of the supervised machine learning classification methods and the concept drift detection methods in the financial business problems. In Joaquim Filipe, Michał Śmiałek, Alexander Brodsky, and Slimane Hammoudi, editors, *Enterprise Information Systems*, pages 268–292, Cham, 2021. Springer International Publishing.

[52] Zongying Liu, Wenru Zhang, Hui Sun, Shaoxi Li, and Hangqi Li. Handling concept drift in financial time series data: Recurrent xavier on-line sequential extreme learning machine. In *Proceedings of the 2022 5th International Conference on Signal Processing and Machine Learning*, SPML '22, page 82–87, New York, NY, USA, 2022. Association for Computing Machinery.

[53] Kurtis Pykes. Model drift in machine learning, 2021.

[54] Haoli Li and Tao Zhao. A dynamic similarity weighted evolving fuzzy system for concept drift of data streams. *Information Sciences*, 659:120062, 2024.

[55] Rodolfo C. Cavalcante and Adriano L. I. Oliveira. An approach to handle concept drift in financial time series based on extreme learning machines and explicit drift detection. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2015.

[56] Filippo Neri. Domain specific concept drift detectors for predicting financial time series, 2021.

[57] Bruno Silva, Nuno Marques, and Gisele Panosso. Applying neural networks for concept drift detection in financial markets. In *CEUR Workshop Proceedings*, volume 960, pages 43–47. CEUR Workshop Proceedings, 2012. Workshop on Ubiquitous Data Mining, UDM 2012 - In Conjunction with the 20th European Conference on Artificial Intelligence, ECAI 2012 ; Conference date: 27-08-2012 Through 31-08-2012.

[58] Hong-Che Lin and Kuo-Wei Hsu. An empirical study of concept drift detection for the prediction of taiex futures. In *2013 IEEE 6th International Workshop on Computational Intelligence and Applications (IWCIA)*, pages 155–160, 2013.

[59] Dushyant Sengar. Implications of algorithmic bias in financial services. In *Revolutionizing the Global Stock Market*, pages 60–82. IGI Global, April 2024.

[60] Reginald E. Bryant, Celia Cintas, Isaac Wambugu, Andrew Kinai, Abdigani Diriye, and Komminist Weldemariam. Evaluation of bias in sensitive personal information used to train financial models, 2019.

[61] Wei Luo, Akib Mashrur, Antonio Robles-Kelly, and Gang Li. Bias-regularised neural-network metamodelling of insurance portfolio risk. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020.

[62] Eu ai act: first regulation on artificial intelligence, 08 June 2023.

[63] S.3572 - 117th congress (2021-2022): Algorithmic accountability act of 2022., 2022, February 3.

[64] Carmel Shachar and Sara Gerke. Prevention of Bias and Discrimination in Clinical Practice Algorithms. *JAMA*, 329(4):283–284, 01 2023.

[65] Reva Schwartz, Apostol Vassilev, Kristen K. Greene, Lori Perine, Andrew Burt, and Patrick Hall. Towards a standard for identifying and managing bias in artificial intelligence, 2022-03-15 04:03:00 2022.

[66] Lina Bouayad, Balaji Padmanabhan, and Kaushal Chari. Audit policies under the sentinel effect: Deterrence-driven algorithms. *Information Systems Research*, 30(2):466–485, 2019.

[67] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. Black-box access is insufficient for rigorous ai audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 2254–2272, New York, NY, USA, 2024. Association for Computing Machinery.

[68] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

[69] Ramya Akula and Ivan Garibay. Audit and assurance of AI algorithms: A framework to ensure ethical algorithmic practices in artificial intelligence. *International Conference on Human-Computer Interaction 2021*, 07 2021.

[70] Xiangyu Qi, Yangsibo Huang, Yi Zeng, Edoardo Debenedetti, Jonas Geiping, Luxi He, Kaixuan Huang, Udari Madhushani, Vikash Sehwag, Weijia Shi, Boyi Wei, Tinghao Xie, Danqi Chen, Pin-Yu Chen, Jeffrey Ding, Ruoxi Jia, Jiaqi Ma, Arvind Narayanan, Weijie J Su, Mengdi Wang, Chaowei Xiao, Bo Li, Dawn Song, Peter Henderson, and Prateek Mittal. Ai risk management should incorporate both safety and security, 2024.

[71] Annapurna Jonnalagadda, Debdeep Mohanty, Ashraf Zakee, and Firuz Kamalov. Modelling data poisoning attacks against convolutional neural networks. *Journal of Information & Knowledge Management*, 23(02):2450022, 2024.

[72] Mahmoud Ghorbel, Halima Bouzidi, Ioan Marius Bilasco, and Ihsen Alouani. Model for peanuts: Hijacking ml models without training access is possible, 2024.

[73] Anoop Singhal. Modeling and security analysis of attacks on machine learning systems. In *Proceedings of the 10th ACM International Workshop on Security and Privacy Analytics*, IWSPA '24, page 1–2, New York, NY, USA, 2024. Association for Computing Machinery.

[74] Junzhe Song and Dmitry Namiot. On real-time model inversion attacks detection. In *Distributed Computer and Communication Networks: Control, Computation, Communications: 26th International Conference, DCCN 2023, Moscow, Russia, September 25–29, 2023, Revised Selected Papers*, page 56–67, Berlin, Heidelberg, 2024. Springer-Verlag.

[75] M. Surekha, Anil Kumar Sagar, and Vineeta Khemchandani. A comprehensive analysis of poisoning attack and defence strategies in machine learning techniques, 2024.

[76] D. E. Namiot and T. M. Bidzhiev. Attacks on machine learning models based on the pytorch framework. *Automation and Remote Control*, 2024.

[77] Vimal Kumar, Juliette Mayo, and Khadija Bahiss. Admin: Attacks on dataset, model and input. a threat model for ai based software, 2024.

[78] Anubha Pandey, Himanshu Chaudhary, Alekhya Bhatraju, Deepak Bhatt, and Maneet Singh. Improving the robustness of financial models through identification of the minimal vulnerable feature set. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF '23, page 297–304, New York, NY, USA, 2023. Association for Computing Machinery.

[79] Robert Nikolai Reith, Thomas Schneider, and Oleksandr Tkachenko. Efficiently stealing your machine learning models. In *Proceedings of the 18th ACM Workshop on Privacy in the Electronic Society*, WPES'19, page 198–210, New York, NY, USA, 2019. Association for Computing Machinery.

[80] Baddiri Narsimha, Ch. V. Raghavendran, Pannangi Rajyalakshmi, G Kasi Reddy, M. Bhargavi, and Pallavi Naresh. Cyber defense in the age of artificial intelligence and machine learning for financial fraud detection application. *International Journal of Electrical and Electronics Research*, 2022.

[81] Ricardo Muñoz-Cancino, Cristián Bravo, Sebastián A. Ríos, and Manuel Graña. *Assessment of Creditworthiness Models Privacy-Preserving Training with Synthetic Data*, page 375–384. Springer International Publishing, 2022.

[82] Shreya Shankar, Labib Fawaz, Karl Gyllstrom, and Aditya Parameswaran. Moving fast with broken data. *arXiv.org*, abs/2303.06094, 2023.

[83] Sierra Wyllie, Ilia Shumailov, and Nicolas Papernot. Fairness feedback loops: Training on synthetic data amplifies bias, 2024.

[84] Abhishek Gupta and Erick Galinkin. Green lighting ml: Confidentiality, integrity, and availability of machine learning systems in deployment, 2020.

[85] Hemadri Jayalath and Lakshmish Ramaswamy. Enhancing performance of operationalized machine learning models by analyzing user feedback. In *Proceedings of the 2022 4th International Conference on Image, Video and Signal Processing*, IVSP '22, page 197–203, New York, NY, USA, 2022. Association for Computing Machinery.

[86] Hana Ahmed and Jay Lofstead. Managing randomness to enable reproducible machine learning. In *Proceedings of the 5th International Workshop on Practical Reproducible Evaluation of Computer Systems*, P-RECS '22, page 15–20, New York, NY, USA, 2022. Association for Computing Machinery.

[87] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, August 2023. Publisher: Elsevier.

[88] Riccardo Albertoni, Sara Colantonio, Piotr Skrzypczyński, and Jerzy Stefanowski. Reproducibility of machine learning: Terminology, recommendations and open issues, 2023.

[89] Zohaib Hassan, Christoph Treude, Michael Norrish, Graham R. Williams, and Alex Potanin. Characterising reproducibility debt in scientific software:a systematic literature review, 2024.

[90] Joseph L. Breeden Peter Quell, Anthony Graham Bellotti and Javier Calvo Martin. Machine learning and model risk management. *Tech Report 2021-01, 8 March 2021, Version 1.0, ©2021 MRMIA*, 8 March 2021,.

[91] Laura Lucaj, Patrick van der Smagt, and Djalel Benbouzid. Ai regulation is (not) all you need. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1267–1279, New York, NY, USA, 2023. Association for Computing Machinery.

[92] Colleen E. Batey. Regulating ai. *Information technology and law series*, pages 139–183, 2022.

[93] Angie Zhang, Olympia Walker, Kaci Nguyen, Jiajun Dai, Anqing Chen, and Min Kyung Lee. Deliberating with ai: Improving decision-making for the future through participatory ai design and stakeholder deliberation. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–32, April 2023.

[94] Takuya Yokota and Yuri Nakao. Toward a decision process of the best machine learning model for multi-stakeholders: a crowdsourcing survey method. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '22 Adjunct, page 245–254, New York, NY, USA, 2022. Association for Computing Machinery.

[95] Xiaoli Tang. Stakeholder-oriented decision support for auction-based federated learning. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8514–8515. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Doctoral Consortium.

[96] Carlos Castro-Herrera and Jane Cleland-Huang. A machine learning approach for identifying expert stakeholders. In *2009 Second International Workshop on Managing Requirements Knowledge*, pages 45–49, 2009.