Modern Banking: The Role of AI in Fraud Detection, Credit Assessment and Portfolio Optimization

Anders Bjerring

Supervisor: Douglas Eduardo Turatti

0. Abstract

This thesis investigates the comparative value of advanced Artificial Intelligence relative to traditional statistical and machine learning methods across three critical areas in banking fraud detection, credit risk assessment, and portfolio optimization. Using structured datasets that emulate real world banking environments, the analysis evaluates the predictive performance and practical applicability of ensemble algorithms, deep learning architectures, and generative AI tools.

While the findings are not broadly generalizable, they indicate that AI based models, particularly ensemble algorithms, tend to perform well in classification tasks, outperforming traditional logistic regression across several metrics. However, when applied to forecasting tasks such as return prediction in portfolio optimization, the results were more mixed. Although deep learning models achieved lower prediction errors in certain cases, these gains did not consistently lead to improved portfolio outcomes, underscoring the challenges of financial forecasting and the continuing relevance of market efficiency.

In addition to the empirical analysis, the thesis explores broader concerns related to explainability, organizational readiness, and regulatory compliance. SHAP was used to enhance model transparency, yet its reliability remains a challenge in high stakes financial contexts. Generative AI, though currently limited in predictive financial modeling, is emerging as a valuable tool for internal operations and productivity focused functions within banking.

These findings are situated within the broader landscape of AI adoption in western banking, where institutions are increasingly integrating AI into core operations. While fraud detection and credit scoring have seen meaningful improvements through AI, widespread implementation is often tempered by concerns over interpretability, model risk, and evolving regulatory expectations. As such, the thesis emphasizes that the effective deployment of AI requires not only technical advancement but also alignment with governance standards and banking sector specific conditions.

Contents

1. Introduction

Artificial Intelligence (AI) has become one of the most prominent and widely discussed technologies in recent years, often viewed as a driver of transformation across industries. Although AI itself is not a new concept, recent advances in computational power, model architectures, and access to large scale data have made real world deployment more practical and impactful. These developments have significantly expanded AI's capabilities, making it increasingly relevant for solving complex tasks, particularly in the banking sector, where decisions must be fast, accurate, and compliant with strict regulatory standards.

The banking sector represents a critical domain for AI implementation not only because of the complexity and volume of decisions involved, but also because banking touches nearly every individual and business in the economy to varying degrees. Whether through access to credit, financial planning, savings, or fraud protection, the way banks operate has widespread societal and economic impact (The European Central Bank, 2024). As such, improvements in banking efficiency, safety, and personalization through AI can have a large significance. In this context, AI is increasingly being explored for applications focused on predictive analytics and data driven decision making. These tasks align closely with AI's core strengths in pattern recognition, outcome forecasting, and processing large volumes of structured and unstructured data.

According to Citibank (2024), AI has the potential to increase global banking sector profits by up to 9% by 2028 equivalent to approximately $170 billion primarily through automation and productivity gains. Simultaneously, rapid progress in model development and the emergence of generative AI have further accelerated interest in AI's role within financial services. Generative tools are now being tested to support internal workflows, automate client communications, summarize complex documents, and introduce new capabilities for both operational and customer facing functions. However, despite growing interest and investment, widespread adoption of AI in banking remains limited. Many institutions are still in experimental or pilot phases and have yet to embed AI into their core decision making structures (Marous, 2024).

There are still several practical challenges that slow down progress. These include the fact that complex models can be difficult to interpret, may carry risks of bias, and are not always easy to align with regulatory requirements or internal governance structures. This is particularly relevant for advanced models like deep neural networks and ensemble methods, which may offer improved performance in some cases, but often provide limited transparency into how decisions are made.

These challenges underscore the importance of assessing AI not only on performance, but also on its transparency and alignment with regulatory expectations. This thesis explores the role of AI in modern banking through both empirical analysis and broader contextual reflection, focusing on three key areas fraud detection, credit risk assessment, and portfolio optimization. Using realistic datasets, it examines whether advanced

approaches provide meaningful advantages over traditional models. Beyond predictive accuracy, the study considers each methods interpretability, suitability for governance frameworks and practical applicability, to better understand the potential role and limitations of AI.

2. Research Area

This thesis investigates the implementation of AI within the banking sector, specifically exploring how and to what extent AI is being adopted in western banking, and whether advanced AI powered models offer improvements over traditional approaches. The analysis focuses on three key areas of banking, as they are central to the industry, rely heavily on data, and impact both internal processes and customer facing activities. They lie at the intersection of operational efficiency, financial performance, and risk control, while also influencing broader societal outcomes through access to financial services, trust in institutions, and economic stability. As such, advancements in these areas have the potential to shape not only how banks operate but also how they contribute to the functioning of the wider economy. By evaluating how well AI models perform in these domains, the thesis aims to clarify their practical value and limitations.. Through empirical analysis and detailed documentation, this thesis seeks to illustrate how AI is currently used in banking and how it may shape the sector in the future. To clearly address these considerations, the thesis examines the following problem statement:

How is AI being adopted within the banking sector, and to what extent do advanced models and methods provide practical improvements over traditional approaches in the areas of fraud detection, credit risk assessment, and portfolio optimization?

The problem statement is addressed through a combination of theoretical review and empirical analysis. Academic literature and industry reports provide context on current AI practices in banking, while empirical tests using real world and synthetic data evaluate the performance of advanced AI models relative to traditional approaches. Traditional models in this thesis refer to rule based systems and simpler ML algorithms such as logistic regression and decision trees, valued for their transparency and established use in finance. In contrast, advanced models include AI driven and more complex algorithms capable of capturing non linear patterns and adapting to dynamic data (Dumitrescu et al., 2022).

This dual approach aims to provide a well rounded understanding of AI's role in potentially enhancing decision making within the banking sector. Details on the methods and analytical framework are presented in the methodology section. The structure of the thesis is outlined on the following page.

2.1 Thesis Structure

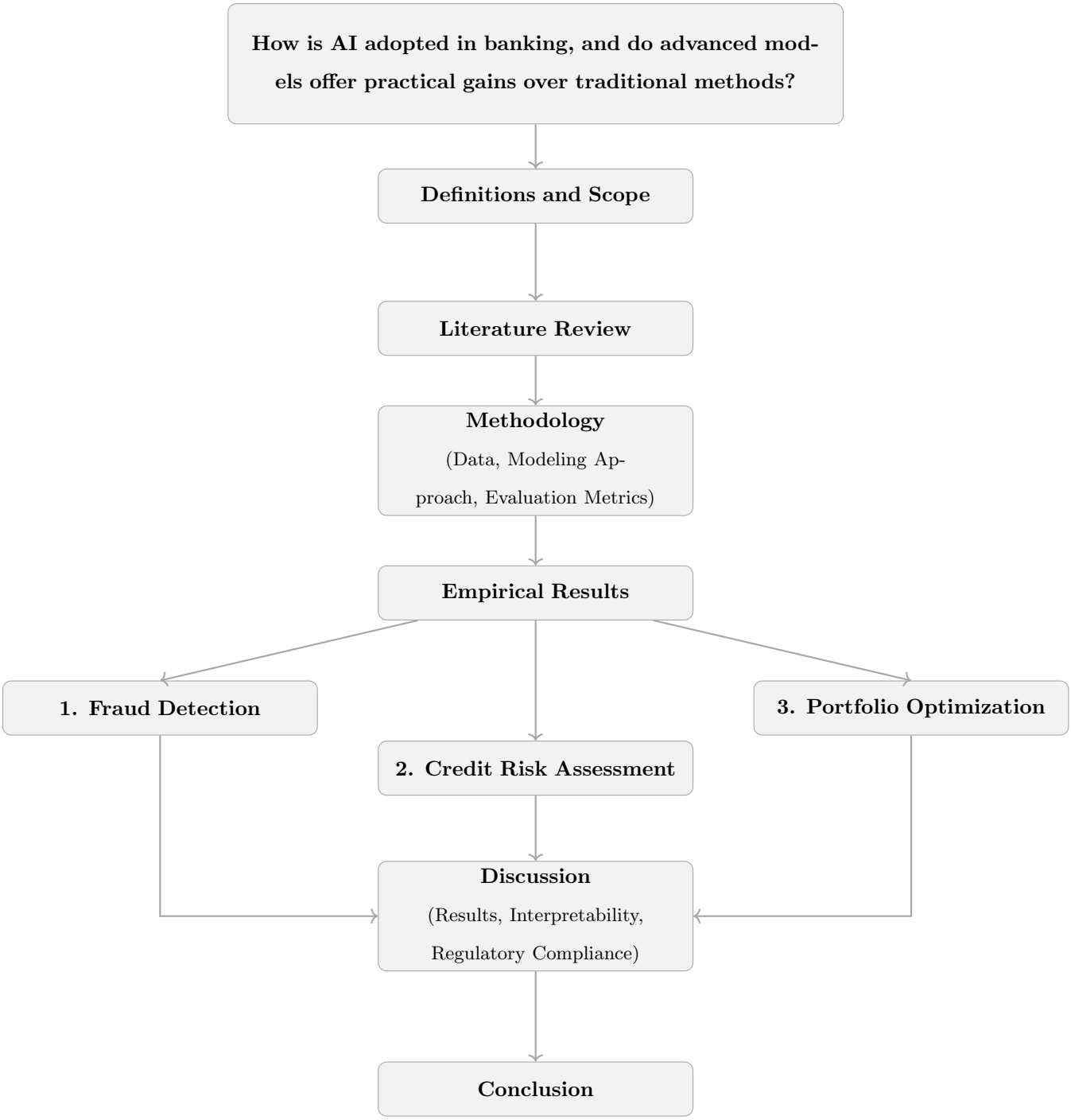

Figure 1: Thesis Structure Overview

2.2 Key Definitions and Scope of Thesis

To address the problem statement, it is necessary to clarify what AI and banking entail within the scope of this thesis. AI refers to a broad set of technologies whose impact depends on the context of application. Similarly, banking covers a wide range of functions, so this study focuses on areas where AI is currently being explored and shows particular promise in enhancing financial decision making.

2.2.1 The Banking Sector

Banking is foundational for the the global financial system, enabling the flow of capital, safeguarding deposits, providing credit, and facilitating investment across economies of all sizes. Although core banking functions such as lending, deposit taking, and payment processing are broadly similar worldwide, the organization, regulatory frameworks, and operational models of banking systems vary considerably across different countries and regions (Mishkin, 2019).

The thesis focuses on the western banking sector, where banks follow different rules depending on the country, along with some common international standards. In most western countries, banks are regulated by national authorities like central banks and other financial agencies. While the way regulation is organized varies, it usually includes rules to keep banks financially healthy, protect depositors, and monitor risks to the financial system. These differences are shaped by each countrys history, laws, and economy, resulting in a mix of systems that promote innovation and competition, but also create challenges for coordinating policies between countries (Mishkin, 2019).

While banking involves a broad range of services from customer interaction and payments processing to compliance and advisory functions certain areas are especially data intensive, analytically demanding, and closely tied to financial performance. Notably, fraud detection, credit assessment, and portfolio optimization represent critical domains that rely on large volumes of data, require continuous monitoring, and depend heavily on predictive modeling and realtime decision making (Mishkin, 2019).

Fraud prevention is a foundational responsibility within banking operations. As financial intermediaries, banks are frequent targets for a variety of fraudulent schemes, including identity theft, account takeovers, insider fraud, wire fraud, and money laundering. Preventing such activities is essential to maintaining the integrity of the banking system and upholding customer trust. To do this, banks implement controls, monitor for anomalies, and coordinate with regulatory authorities to prevent and respond to criminal activity. Fraud detection strategies rely heavily on realtime data monitoring, behavioral pattern recognition, and risk scoring systems (Mishkin, 2019).

Credit risk assessment also relies heavily on data analysis and systematic evaluation. Assessing a borrowers

ability to repay a loan is central to the stability and profitability of banks. Traditional assessments draw on historical credit behavior, income verification, financial ratios, and collateral analysis. At the same time, effective credit risk assessment should account for broader economic conditions, evolving borrower circumstances, and the composition of a banks loan portfolio. Accurate modeling plays a key role in pricing loans appropriately, meeting capital adequacy requirements, and managing exposure to default (Mishkin, 2019).

Portfolio optimization, particularly within banks investment and wealth management divisions, represents a third data driven area. Whether managing internal funds or acting on behalf of institutional and retail clients, banks construct and maintain portfolios that reflect risk tolerance, investment objectives, and market dynamics. This process involves the ongoing selection, weighting, and rebalancing of stocks to optimize returns while controlling risk. Most portfolio managers rely on quantitative models, financial forecasts, and risk metrics to guide decision making (Mishkin, 2019).

As stated, banking covers a wide array of functions, however these three areas reflect some of the sectors most data-intensive and analytically demanding tasks. Their reliance on ongoing monitoring, predictive assessment, and strategic decision making makes them especially relevant for evaluating the impact of AI. For this reason, these areas serve as the central focus of the empirical analysis in this thesis.

2.2.2 AI

AI refers to computational methods that enable machines to perform tasks typically linked to human intelligence, such as interpreting data, identifying patterns, making decisions under uncertainty, and generating content or responses. Unlike traditional software based on fixed rules, AI systems learn from data and generalize from past examples to function in complex or unfamiliar settings (Russell & Norvig, 2021). While AI spans a wide range of capabilities, this thesis focuses primarily on predictive AI models designed to classify outcomes or forecast events using structured data. These techniques are central to the empirical analysis of fraud detection, credit risk assessment, and portfolio optimization. The thesis also considers the growing role of generative AI in banking operations.

A central component of AI is machine learning (ML), which involves training algorithms to identify patterns in historical data and use those patterns to make decisions or predictions. ML models are the outcomes of this training process as they learn through exposure to data by adjusting internal parameters to minimize errors and improve accuracy over time. Broadly, ML encompasses two main types supervised learning and unsupervised learning, each with distinct purposes, methodologies, and areas of application (Russell & Norvig, 2021).

In supervised learning, the algorithm is trained on a dataset that contains both the input features and

the correct output for each data point. The goal is for the model to learn how to link inputs to the correct outputs, so it can make accurate predictions on new, unseen data. This learning process is guided by feedback where the model makes a prediction, compares it to the actual answer, and adjusts its internal settings to reduce the error. Over time, these adjustments help the model become more accurate. Supervised learning is widely used in situations where examples of correct answers are available (Russell & Norvig, 2021).

Unsupervised learning can be used with data that does not have labels or known outcomes. Instead of learning to predict a specific result, the model tries to uncover patterns, structures, or relationships within the data itself. The goal is often to find natural groupings, detect hidden patterns, or simplify complex information. Unlike supervised learning, there is no clear feedback or correct answer here the system organizes or represents the data based on what it finds. This approach is especially useful in exploratory data analysis, where the structure of the data may be too complex for manual interpretation. Common techniques include clustering, which groups similar data points together, and dimensionality reduction, which highlights the most important features in a dataset. Although supervised and unsupervised learning are distinct, they are not mutually exclusive and can enhance each other when combined. For example, unsupervised learning can be used as a preprocessing step to simplify or organize raw data before applying a supervised model. It can also help identify patterns or anomalies in the data that might influence how a supervised model is trained (Russell & Norvig, 2021).

Predictive AI is not a distinct technical category but rather a widely used application of primarily ML. It involves the use of algorithms to try an estimate future outcomes based on historical patterns. These systems generate forecasts or classifications that can help guide decisions, Predictive AI is especially valued in environments with large datasets and frequent decisions, where human judgment alone may be insufficient or inefficient. It relies on probabilistic models that continuously update as new data becomes available, making it particularly suitable for dynamic and data rich settings (Marous, 2024).

Deep Learning, which is a subfield of ML, focuses on models built with artificial neural networks. These networks consist of multiple interconnected layers that process data in stages, allowing the model to capture complex, nonlinear relationships. The depth of these models enables them to handle unstructured data such as text, images, or speech more effectively than traditional techniques. Traditional are based on predefined mathematical relationships and are generally more transparent but less suited for capturing more complex patterns. In contrast, deep learning models can learn these patterns automatically, though often in ways that are difficult to interpret. Models using deep learning is sometimes referred to as black boxes because it can be difficult to understand how the models makes its decision (Russell & Norvig, 2021).

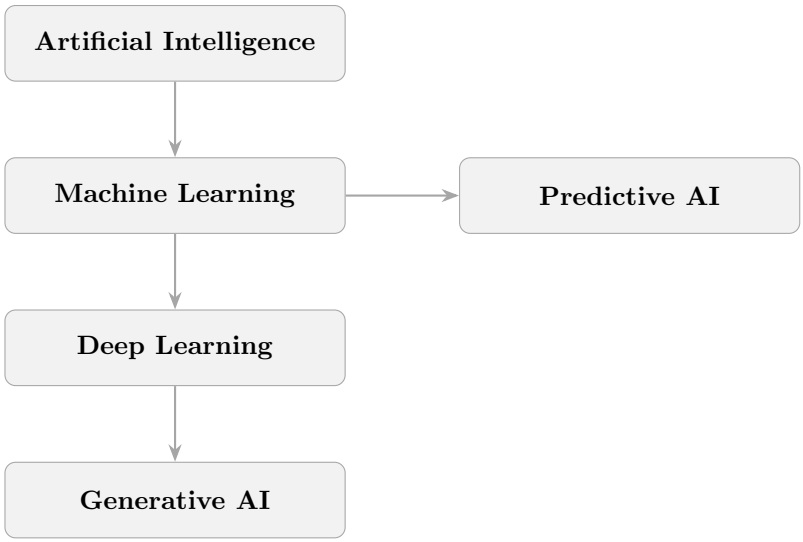

Figure 2: Layered structure of AI

Source: Figure adapted from Marous (2024)

Generative AI refers to a class of AI systems that are capable of producing new and meaningful content rather than only analyzing or classifying existing data. These models are typically built on deep learning architectures, such as transformer networks, and are trained on vast datasets to learn the underlying patterns, structures, and relationships within the data. Once trained, generative AI can produce outputs such as text, images, audio, or code that closely resemble the examples it has learned from (Feuerriegel et al., 2024).

One widely recognized application of generative AI is ChatGPT, developed by OpenAI. Introduced in late 2022, ChatGPT is based on large language models (LLM) and can generate humanlike responses to a wide range of prompts. Its capabilities include answering questions, composing text, summarizing documents, and assisting with code or calculations (OpenAI, 2022). Since its launch, ChatGPT has undergone continuous improvement, with new features such as multimodal input, memory for personalized interactions, and integration with external tools.

In all, AI is not a single technology but a layered system of statistical and computational tools that enable machines to learn, adapt, and act. From more interpretable models to complex, multilayered systems, AI methods vary in complexity and transparency.

2.2.3 Traditional Versus advanced AI Powered Methods

Traditional approaches in banking include both statistical systems and simpler ML models such as logistic regression and decision trees. Logistic regression, in particular, has been widely adopted for classification tasks like fraud detection and credit risk assessment due to its simplicity, transparency, and acceptance in regulatory settings. Many banks continue to rely on these models, which are often embedded in legacy

systems built around defined rules (Candemir, 2025).

In academic research, logistic regression has long been a standard method for binary classification tasks due to its interpretability, robustness, and ease of implementation. Given that real world banking models are often not disclosed and not publicly disclosed, logistic regression is commonly used as a stand in for traditional approaches in empirical studies (Anderson, 2007; Dumitrescu et al., 2022). Its transparency makes it a natural comparative baseline when evaluating more complex ML models. For example, studies such as Niu et al. (2019) and Zhang et al. (2024) use logistic regression to benchmark improvements in predictive performance.

Advanced ML algorithms, such as ensemble and boosting methods, have been developed to capture complex and nonlinear patterns in data patterns that traditional models may overlook. These techniques combine the outputs of several base models to produce more accurate and robust predictions. While they can enhance performance in areas like fraud detection and credit risk assessment, they also introduce greater model complexity and reduce transparency.

A similar contrast applies to portfolio optimization. Traditional approaches, based on Modern Portfolio Theory (MPT), focus on balancing expected return and risk using historical data and assumptions of stable market conditions. These methods often involve human judgment and relatively static strategies. In contrast, AI powered techniques use ML algorithms to process not only financial indicators but also alternative data sources, such as macroeconomic variables and online behavior. This enables more dynamic portfolio adjustments and a potentially greater capacity to model real world complexity (Mercanti, 2024). The models and algorithms used in this thesis will be examined in more detail in the methodology section.

2.2.4 AI in the Banking Sector

The purpose of this section is to explore how AI is being adopted within the banking sector, with a particular focus on western banks. While capturing the full complexity of AI implementation across all western banks is beyond the scope of this thesis, the sector shares sufficient structural and regulatory similarities to support a general overview. It is important to note that individual banks may be at different stages of AI adoption. However, the surveys and reports referenced here primarily focus on the large financial institutions and are considered broadly representative of current developments. This descriptive overview builds on the conceptual foundation introduced earlier and draws on recent empirical insights from the 2024 Digital Banking Report – State of AI in Banking by Jim Marous, supplemented by prior studies on AI adoption within the financial industry.

As early as 2017, the Financial Stability Board an international organization overseeing the global financial system had already identified several key areas within banking where AI could significantly enhance existing

practices and processes. These areas, illustrated in Figure 3, include customer interaction, operational optimization, trading and portfolio activities, and regulatory compliance. Notably, the core focus areas of this thesis credit scoring and risk management, fraud detection, and portfolio optimization were among those highlighted as having strong potential for further developments with AI (FSB, 2017).

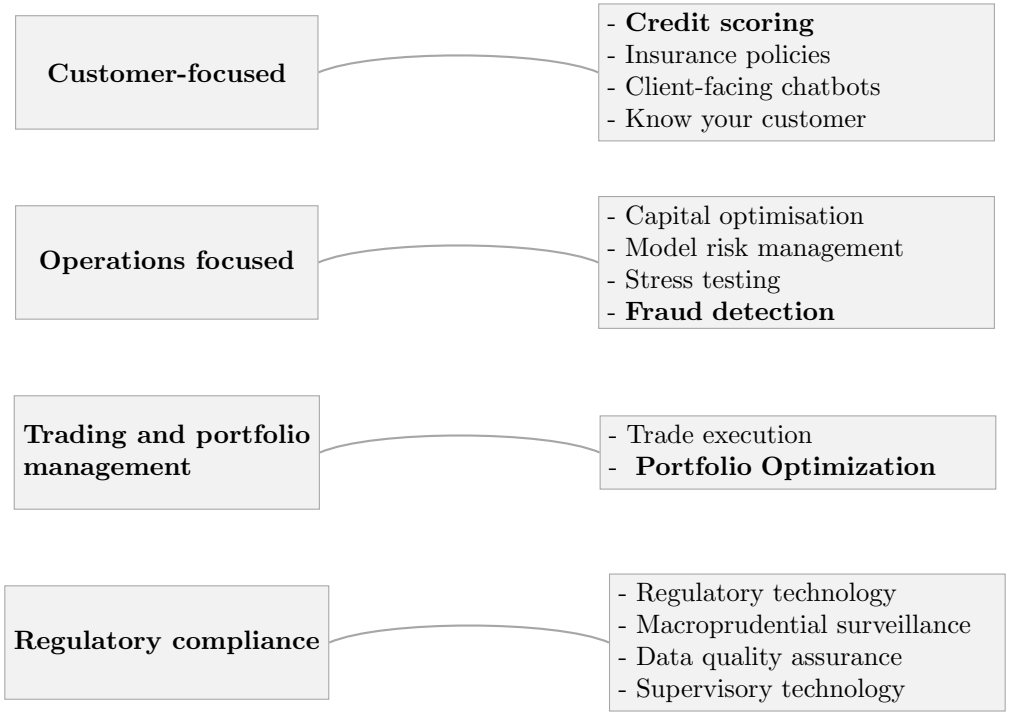

Figure 3: Key areas identified for AI implementation in banking

Source: Figure adapted from FSB (2017, as cited in Kaya, 2019).)

Recently with the rise of generative AI, a survey by The Harris Poll for the Digital Banking Report found that 57% of banking professionals reported using AI for general productivity tasks, including drafting internal communications and presentations. In marketing, 55% used generative AI to create content for campaigns and digital platforms, and 48% reported using AI chatbots and virtual assistants areas where interaction quality has improved significantly with the rise of improved generative models.

In more technical departments, generative AI is also being used. Half of surveyed professionals in IT and software development report using AI to assist with coding, while 49% of those in research and analysis roles use it to summarize complex market information. Another 40% apply AI for predictive modeling in risk related tasks. The poll suggest that AI is no longer limited to specific domains or early adopter teams and that is used across roles and functions, becoming an expected part of daily workflows in many banks.

As usage expands, so does its strategic importance. Reflecting this potential, the consulting firm McKinsey

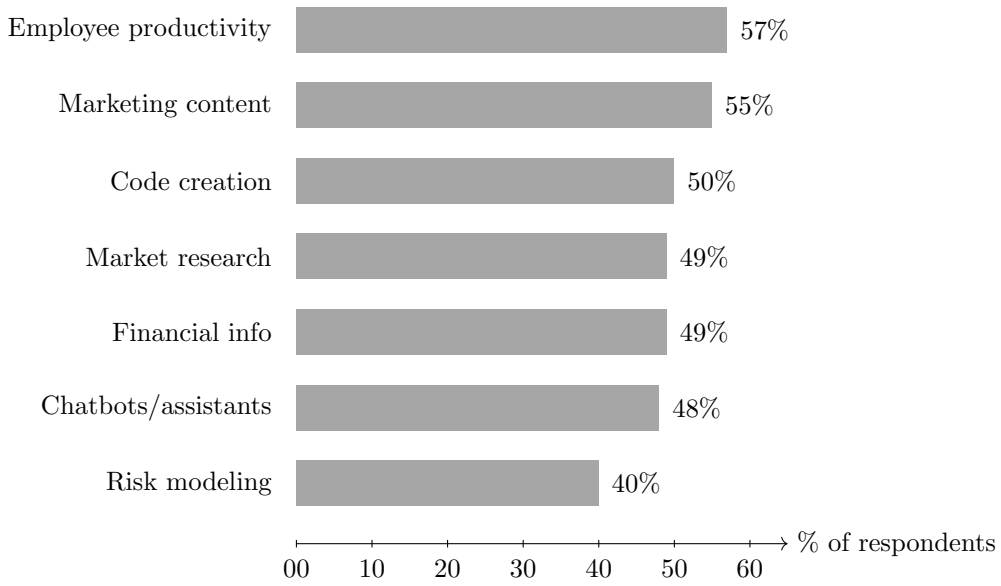

Figure 4: Use of Generative AI Within Banking

Source: Figure adapted from Marous (2024)

& Company emphasizes that generative AI could generate substantial value for the global banking sector (McKinsey, 2023). While much of this value is tied to productivity gains, the broader interest lies in its potential to drive innovation. Banks are increasingly exploring how generative AI can transform product development and enable new ways of interacting with customers.

Despite the growing hype, a more measured perspective is warranted. As McKinsey highlights in a separate report (McKinsey, 2024), most banks remain in the early stages of implementation. Many institutions are still in the process of developing the necessary infrastructure, governance frameworks, and internal capabilities to enable responsible deployment. While the potential seems large, realizing it will require deliberate action aligning innovation with technical maturity, regulatory requirements, and shifting customer expectations.

Looking again at the *State of AI in Banking report*, it becomes evident that broader organizational adoption of AI remains at an early stage across the financial sector. As shown in Figure 5, 68% of banking institutions classify their AI adoption as being in the beginning stage, while 21% report having not yet started. Only 11% describe their efforts as maturing, and none report having reached an advanced stage. This distribution reinforces McKinseys observation that, despite growing interest and experimentation, most banks are still in the foundational phases of building the infrastructure, governance, and internal capabilities necessary for full-scale deployment. It suggests that while momentum is building, widespread integration of AI into core banking operations is still a work in progress.

	Banks	FinTechs	Insurance
Not started	**21%**	**4%**	**6%**
Beginning	**68%**	**83%**	**88%**
Maturing	**11%**	**13%**	**6%**
Advanced	**0**	**0**	**0**

Figure 5: AI adoption stages across Banks, FinTechs, and Insurance sectors

Source: Figure adapted from (Marous, 2024)

FinTech companies, despite often being thought of as much more digitally advanced, show a similar pattern to banks when it comes to AI adoption. According to the data, 83% of FinTechs report that they are in the beginning phase of adoption, while 13% indicate a maturing level of integration. However, only 4% state that they have not yet started. For the insurance sector 88% of institutions are indicating that they are in the beginning phase. An additional 6% report having reached a maturing stage, and another 6% have not started. As in the banking sector, no institutions in either FinTech or insurance report having reached an advanced level of AI adoption.

In addition to being in the early stages of implementation, many organizations also report limited internal AI expertise. According to the State of AI in Banking report, 55% of banks assess their internal AI capabilities as either low or very low. This points to a gap between the growing interest in AI and the operational resources available to support its wider deployment. Taken together, the responses from banking, FinTech, and insurance institutions show a consistent trend across the financial industry that most organizations are still at the beginning of their AI implementations, with only a small share reporting more mature adoption. No sector reports any cases of fully advanced implementation. Importantly, AI adoption in finance is not only technically demanding but also financially expensive. Developing or purchasing AI solutions especially those tailored for complex financial tasks such as fraud detection or credit risk modeling requires substantial investment. In parallel, implementing the necessary regulatory and governance frameworks such as ensuring explainability, human oversight, auditability, and legal compliance adds significant cost and complexity (Marous, 2024)

However, looking into the potential of AI in a more matured stage, a survey conducted by OpenText for the *State of AI in Banking* report shows that the banking sector are beginning to form clearer expectations about where AI could deliver value as adoption reaches the advanced stage. The survey includes two parts

that shed light on how banks prioritize expected outcomes from AI and which specific use cases are believed to offer the greatest potential benefit. These insights offer an overview of the areas that institutions are likely to focus on as they move beyond early experimentation and toward broader integration.

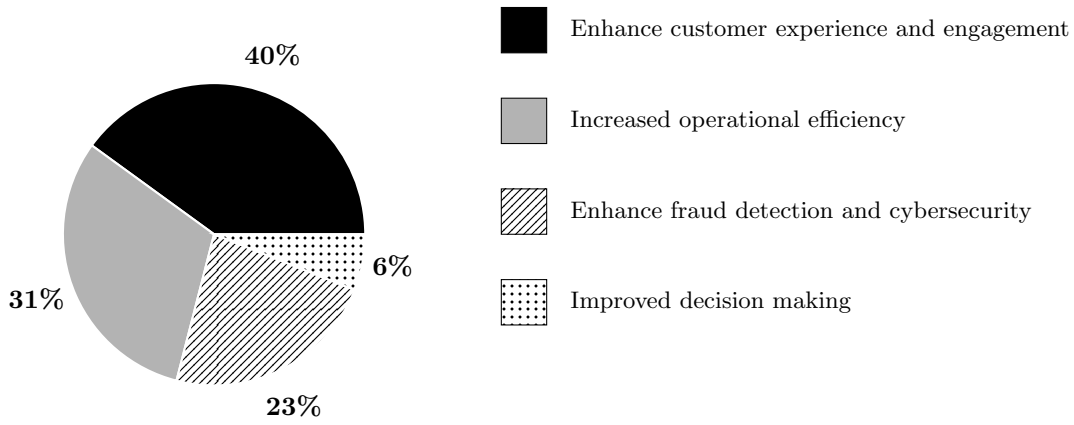

Figure 6: Primary benefit banks seek from AI enabled solutions

Source: Figure adapted from (Marous, 2024)

Figure 6 presents the results of a single selection question, where respondents from banks were asked to identify the primary benefit they expect from AI enabled solutions. The most frequently selected outcome was enhancing customer experience and engagement, chosen by 40% of respondents. This was followed by increased operational efficiency (31%), improvement in fraud detection and cybersecurity (23%), and improved decision making (6%).

In a separate, multi selection question from the same survey, banks were asked which AI use cases they believe will add significant value to banking over the next three to five years. The responses reveal a wide range of expected applications. Chatbots and virtual assistants were cited by 87% of respondents, followed by 83% who selected personalized marketing and product recommendations. Improving backoffice productivity was identified by 74%, and 65% pointed to fraud detection and prevention. Other selected areas included credit scoring and underwriting 62%, compliance and regulatory reporting 58%, customer retention and churn prediction 52%, and document processing and analysis 48%. The poll suggest that banks anticipate the future impact of AI across both customer facing services and operational infrastructure.

While interest in AI is growing, the report also shows that most banks are not developing these solutions that they anticipate to employ themselves. Instead, a majority prefer to work with third-party providers to access AI capabilities. According to a survey conducted by the *State of AI in Baking* report itself, 69% of banks indicated that their organization primarily acquires AI solutions from external vendors rather than

building them internally. This approach reflects a preference for leveraging existing technologies, platforms, and technical expertise particularly in a context where many institutions still report limited internal AI capabilities. For banks at the early stages of adoption, external partnerships may offer faster implementation, access to specialized tools, and reduced development costs compared to in-house development.

The reliance on third party providers also aligns with broader trends in digital transformation, where financial institutions increasingly collaborate with external technology firms to deliver innovation at scale. In the case of AI, this may include solutions for fraud detection, risk modeling, or customer service automation, delivered as modular tools or integrated systems. As adoption progresses, such partnerships are expected to remain a central feature of how banks engage with AI technologies (Marous, 2024).

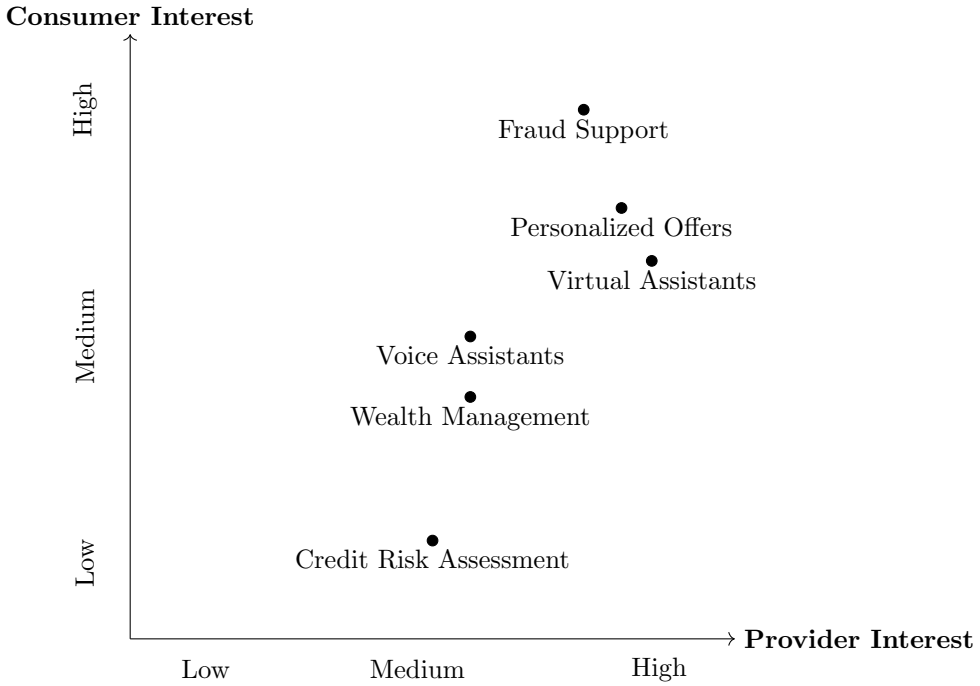

Figure 7: Interest in AI Capabilities by Banking and Consumers

Source: Figure adapted from (Marous, 2024)

A survey conducted by Insider Intelligence for the *State of AI in Banking* report gathered input from both banks and customers to assess expectations for various AI capabilities over the next three years. Figure 7 presents the results, comparing projected interest from providers, which are banks on the horizontal axis and from consumers on the vertical axis. Each point on the chart represents a specific AI use case, offering an overview of which capabilities are expected to receive the most attention from each group. Notably, credit risk assessment, fraud detection, and wealth management are among the highlighted use cases.

Fraud support is positioned in the top right corner of the chart, indicating strong interest from both banks

and consumers. This suggests it is a shared priority, closely aligned with institutional goals and public expectations. Wealth management appears near the center, reflecting a more moderate and balanced level of interest. While developments in portfolio-related services are gaining traction, they are not currently seen as urgent as other areas. Credit risk assessment is located in the lower middle, where institutional focus is moderate and consumer interest remains relatively low. Although levels of attention vary across these use cases, all three areas represent financially critical functions. This section has outlined the current state of AI in banking, with a focus on areas of application, employee use, organizational readiness, and future expectations. However, broader adoption is not only shaped by technological capability it is also influenced by concerns around algorithmic bias, regulatory compliance, and governance.

2.2.5 Regulatory Compliance and Governance

As use of AI becomes more common, it has attracted growing attention from regulators and raised serious ethical concerns particularly in the EU and the U.S. In both, rules and regulations are being updated to manage the risks of using AI in important areas like banking. While the EU has introduced a more unified and strict legal framework through the AI Act, the U.S. continues to rely on existing financial regulations and agency specific guidelines.

The EUs AI Act, adopted in 2024 and set to fully apply from 2026, is the most comprehensive legal framework for AI so far. It classifies AI systems by risk, and systems used in banking such as those for credit scoring and lending are considered high risk (KPMG, 2024). These must meet tough requirements for accuracy, transparency, robustness, and human oversight. The rules are supported by the General Data Protection Regulation (GDPR), which guarantees individuals the right to understand decisions made about them by automated systems.

Under the AI Act, banks and other financial institutions must monitor and document their AI systems closely. They must also test for bias, a situation where the model treats certain groups unfairly. This could be if an AI system is more likely to reject loan applications from women or ethnic minorities due to patterns in historical data. These systems must be regularly checked and corrected, and staff must be trained to understand how to manage and, if needed, override AI decisions (European Commission, 2024a). The European AI Office has been set up to help enforce these rules across member states (European Commission, 2024b).

In contrast, the U.S. takes a more flexible and decentralized approach. Rather than introducing a new large AI law, U.S. agencies apply existing regulations to AI use. The Consumer Financial Protection Bureau, for example, reminds lenders that they must still explain credit denials to customers, even if an AI system made the decision (CFPB, 2023). The Office of the Comptroller of the Currency and the Federal Reserve

also expect banks to apply existing model risk management rules to AI, which include requirements for explainability and testing (Mayer Brown, 2022). Banks are also expected to apply the same scrutiny to third party AI providers as they would to internal systems.

Although their regulatory approaches differ, both the EU and the U.S. share a focus on transparency, fairness, and human accountability. Regulators in both regions caution against the use of previously mentioned black box models, AI systems that generate decisions without offering clear explanations. In response, many financial institutions are investing in explainable AI, strengthening internal governance, and adopting ethical guidelines. These efforts are not only about compliance they also play a strategic role in preserving public trust and protecting the institutions reputation (European Central Bank, 2024).

Alongside explainability and oversight, another pressing ethical concern involves the types of data used to train and operate AI systems. A key issue is the growing use of alternative data which could be information from social media, mobile phone activity, or browsing history. While this can improve predictions, it also raises privacy and fairness concerns. If not properly reviewed, such data can reinforce discrimination or violate data protection laws. To avoid this, institutions are beginning to assess whether data sources are relevant, legal, and fair before using them in AI models (European Central Bank, 2024).

The EU tends to follow a cautious, top down approach, while the U.S. prefers a more flexible strategy with input from the private sector. For example, the U.S. Department of the Treasurys 2024 Request for Information acknowledged both the risks like bias and the benefits such as efficiency and innovation of AI in finance (U.S. Department of the Treasury, 2024). Meanwhile, the Department of State published a national AI strategy that promotes ethical and secure use across sectors (U.S. Department of State, 2024). If regulation, ethics, and governance are not taken seriously, the consequences can be significant and costly. Poorly governed AI systems can lead to discrimination, data misuse, legal penalties, or loss of customer trust. These risks are especially severe in the banking sector, where automated decisions directly impact peoples access to credit, savings, and financial stability. Using ethical principles such as fairness, transparency, and human oversight is increasingly been viewed not only as a regulatory necessity but also as an important enabler of long term, responsible innovation (Agu et al., 2024).

3. Literature Review

Building on the overview of current AI use in banking and the associated regulatory needs, this literature review shifts focus to research that specifically examines developments and testing of AI models and methods within the core financial domains of fraud detection, credit risk assessment, and portfolio optimization. While the previous section included broader uses such as customer service and communication tools, the following draws on studies that more precisely address the problem statement interest in the three key areas.

The review incorporates both recent empirical work and foundational studies, highlighting how AI has been evaluated for its ability to improve predictive performance, help with complex tasks, and support risk sensitive decision making. Common methodological approaches include ML, model benchmarking, and large scale data analysis. In addition to performance outcomes, the literature reflects ongoing concerns about data reliability, explainability, and the operational hurdles tied to the real world implementation. These insights help shape the direction of the thesis empirical analysis by pointing to areas where AI appears to hold high potential, while also revealing important practical constraints.

3.1 Fraud Detection

AI driven fraud detection has been and still is a critical area of research and application in the banking sector, as financial institutions seek more effective ways to combat increasingly sophisticated fraud schemes. Several studies have examined the impact of ML in enhancing fraud detection capabilities. Bao et al. (2020) provide an overview of AI-based techniques, highlighting the effectiveness of supervised learning models in recognizing known fraud patterns. However, they note that models may struggle with emerging fraud tactics, which often require unsupervised approaches such as clustering and anomaly detection to identify previously unseen behaviors.

Faisal et al. (2024) similarly emphasize the growing importance of realtime, AI powered fraud detection in modern banking environments. They argue that hybrid learning approaches combining supervised and unsupervised models tend to outperform single model methods by enabling systems to detect both established patterns and novel anomalies. This adaptability is particularly valuable in high-volume, fast-paced transaction settings. Bello et al. (2023) support this view, stressing the need to balance precision and recall in AI based fraud detection systems. Overly precise models risk missing subtle fraud signals, while high recall models may produce too many false alarms, straining investigative resources. Finding the right balance is seen as essential for both performance and operational credibility in the study.

The shift from exclusively rule based detection to AI driven methods has been widely explored. KPMG and Nets (2020) describe their Fraud Ensemble model, which integrates multiple ML algorithms to improve

detection accuracy in the context of payment card fraud. Their long term analysis indicates that such models can significantly reduce false positives while maintaining realtime responsiveness.

AI-based fraud detection systems are commonly evaluated using the aforementioned precision and recall and also F1-score metrics that capture both the accuracy and reliability of model performance. Johora et al. (2024) highlight that a diverse range of algorithms scored highly across all three, demonstrating advanced AI's potential to detect fraud effectively while keeping false positives manageable. However, they caution that setting thresholds too strictly can lead to increased false negatives, while overly sensitive models may block legitimate transactions both of which carry operational costs. Moreover, a lack of interpretability in many AI models remains a barrier, particularly in light of regulatory requirements for explainability.

Another significant challenge lies in data quality and class imbalance. Fraudulent transactions typically constitute only a small fraction of the overall transaction volume, resulting in datasets where non fraud cases vastly outnumber fraud cases. This imbalance makes it difficult for AI models to learn meaningful fraud patterns. Faisal et al. (2024) suggest data augmentation and synthetic case generation as ways to enhance model robustness by increasing exposure to diverse fraud scenarios. Javaid (2024) highlights that AI systems benefit from continuous learning, improving their ability to detect emerging money laundering techniques and adapt to evolving threats more effectively than static, rule based systems.

As AI powered fraud detection becomes more integrated into banking operations, the importance of robust data governance grows. Nathan et al. (2023) argue that AI introduces both opportunities and responsibilities, particularly around data integrity, quality, and regulatory compliance. They highlight the role of AI powered data governance in automating anomaly detection, realtime validation, and compliance monitoring tools that support both operational efficiency and alignment with data protection frameworks such as GDPR. They further emphasize the need for governance models that ensure transparency, fairness, and ethical standards, especially as AI systems increasingly generate high impact decisions at scale.

Taken together, existing research suggests that advanced AI powered models can significantly improve accuracy in fraud detection. These models have demonstrated advantages in identifying complex patterns, enabling real-time monitoring, and enhancing operational efficiency compared to traditional rule-based systems. As AI becomes more integrated into fraud detection processes, the focus of many studies has shifted from whether it performs better than traditional methods to how it should be implemented, particularly in relation to the role of human oversight, ongoing concerns about transparency, regulatory compliance, and the risks of false positives and negatives.

3.2 Credit Risk Assessment

Like fraud detection, AI in credit risk assessment has a history of academic and industry attention. AI and more precisely ML introduce more flexible, adaptive, and data rich approaches that promise improved predictive performance and operational efficiency in terms of determining credit scores and chances of default. The literature emphasizes that these technologies allow banks to move beyond linear assumptions and static scoring systems, toward advancded ML models that incorporate alternative data sources.

From a predictive standpoint, Misheva et al. (2021) show that algorithms such as Random Forest and XGBoost significantly outperform traditional credit risk assessment methods. Their comparative analysis on large scale credit data reveals that more advanced ML based models deliver higher classification accuracy and better identification of high risk borrowers. However, the study also stress that performance alone is insufficient. To meet regulatory requirements and maintain institutional trust, their study incorporates the tools SHAP and LIME that assist in explanation of how advancded AI models make decisions. SHAP shows how much each feature contributes to a prediction, while LIME builds simple explanations for individual cases. These are tools, acording to the study, that potentially enable banks and compliance teams to understand and audit model outputs.

Zhang et al. (2024) further validate AI's potential in credit risk assessment, but using deep learning models. Their findings show strong results across the metrics precision and recall. These models are capable of capturing the non-linear relationships and subtle patterns in borrower data that are often overlooked by traditional approaches. Moreover, they can evolve over time, adapting to new risk indicators and reducing reliance on outdated assumptions. Its argured, that the adaptability is especially valuable in dynamic economic environments, where risk profiles can change rapidly due to macroeconomic shocks or borrower behavior shifts.

Likewise, Brown (2024) and Addo et al. (2018) both emphasize the value of AI based credit risk models in processing large and complex datasets, including behavioral and transactional information. While Addo et al. earlier work reflects the early application of ML in credit analysis, Browns more recent findings align with a matured perspective on how use of AI improves both accuracy and access to credit. Together, their research suggests that AI models can identify credible borrowers who might be overlooked by conventional scoring systems, thereby supporting more inclusive lending practices. As with fraud detection, credit datasets are often highly unbalanced, with relatively few default cases. This makes the use of AI and its ability to detect patterns in rare events especially valuable for minimizing risk and expanding financial access without compromising model reliability.

Looking further at alterative data sources, Mhlanga (2021) looks into how AI can leverage non traditional

data such as social media behavior, mobile usage, and geospatial data to build richer credit profiles. In markets where formal credit histories are lacking, such approaches enable institutions to assess risk with greater nuance, ultimately extending credit to underbanked or previously excluded groups. The paper suggests that AI not only improves model performance but also plays a role in promoting financial inclusion.

According to McKinsey (2024), generative AI is starting to help in credit risk assessment, particularly through administrative and decision support tasks. Most credit risk executives expect to implement at least one use case within the next year. While generative AI has not yet transformed the field, its noted that it is being piloted for tasks such as automating document review, generating credit memos, and facilitating communication around risk exposures. These applications may not directly improve model accuracy, but they enhance efficiency, consistency, and auditability within the risk function.

However, as in the case of fraud detection, key challenges are also investigated in studies. One of the most prominent is again model explainability. Misheva et al. (2021) emphasized that the complexity of AI models must be balanced with the regulatory requirement for transparency. In the EU, GDPR grants individuals the right to an explanation for decisions made solely by automated systems. As such, its argued that the use of interpretability tools is not only recommended but necessary for compliance.

Data quality and fairness also pose barriers to effective adoption. McKinsey (2024) reports that many financial institutions face issues such as incomplete borrower data, biased training sets, and underrepresentation of minority groups. Edunjobi & Odejide (2024) further emphasize that ethical principles such as fairness, accountability, and transparency must be integrated into model development. Failing to address these concerns can not only reduce predictive performance but also lead to discriminatory outcomes that damage institutional credibility and violate regulatory standards.

Organizational readiness is another recurring theme in the examined literature. Edunjobi & Odejide (2024) observe that many banks still operate on legacy systems that lack the flexibility and integration capabilities needed to support modern AI tools. Brown (2024) similarly notes that the success of AI initiatives depends not only on algorithmic sophistication but also on how well institutions align technology with operational infrastructure and governance processes.

Overall, the literature indicates that AI and more advanced AI hold significant longterm potential for improving credit risk assessment. These technologies can enhance risk estimation, support financial inclusion, and improve operational efficiency. However, unlocking this value depends on effectively addressing challenges related to transparency, data quality, ethical design, and institutional readiness.

3.3 Portfolio Optimization

Portfolio optimization is traditionally based on MPT introduced by Markowitz in 1952 (Markowitz, 1991). This approach focuses on building portfolios that balance risk and return by spreading investments across different assets. This model relies on historical estimates of expected returns, variances, and covariances and assumes rational behavior, normally distributed returns, and stable correlations among assets. While MPT remains foundational, its limitations particularly its sensitivity to input assumptions and inability to respond dynamically to market changes have prompted further research into more adaptive and datadriven approaches.

Complementing MPT, Famas (1970) Efficient Market Hypothesis (EMH) has framed much of the skepticism toward model based prediction. EMH posits that asset prices already reflect all publicly available information, rendering consistent outperformance nearly impossible. This hypothesis implies that even advanced quantitative or AI driven models should not be able to systematically identify mispricings or profitable anomalies. Nonetheless, the pursuit of such inefficiencies persists.

Lo's (2004) Adaptive Market Hypothesis (AMH) is often cited as a more flexible alternative to the EMH. AMH proposes that markets evolve over time as investors adapt to changing conditions, allowing for short-term inefficiencies that can potentially be exploited. From this perspective, the use of AI tools aligns with market dynamics, as these technologies are designed to adapt to new signals and shifting environments. This framework helps explain why AI is increasingly studied and applied in areas such as asset management and portfolio optimization.

In response to perceived limitations of traditional models, Rasekhschaffe & Jones (2019) provide an overview of how ML is being used in asset management. They highlight that these methods can work with large, complex datasets, capture non linear patterns, and avoid relying on strict assumptions. Their review shows that models like decision trees, random forests, and neural networks perform well at identifying predictive signals especially when combined in ensemble approaches. However, they also point out that performance often drops when tested on new data, and that a lack of interpretability remains a major challenge for wider use in institutions.

Gu, Kelly, & Xiu (2020) offer empirical support for the idea that ML can enhance stock return prediction. Their study applies a range of ML models to predict the cross section of stock returns using a large dataset of firm characteristics. The results show that ML methods outperform linear benchmarks, especially when allowing for interaction effects and non linearities. Notably, combining multiple models further improves performance, suggesting that ensemble learning may mitigate overfitting and improve robustness. However, the authors caution that model performance remains highly sensitive to the design of the learning task, and

emphasize the importance of careful feature selection, regularization, and validation techniques.

The limits of AI prediction are further examined in Mokhtari et al. (2021), who explore the use of ML to predict stock market behavior based on both technical indicators and sentiment data. Their study includes various classification and regression models applied to historical stock data and social media content. While some models show moderate predictive power in specific contexts, the overall results are mixed. The authors conclude that the inherent volatility and efficiency of financial markets restrict the practical value of predictive models and highlight the importance of hybrid approaches that combine quantitative modeling with domain expertise.

Beyond numerical models, recent research has also explored the application of LLMs in portfolio related tasks. Romanko et al. (2023) investigate the ability of ChatGPT to select stocks from the S&P 500 based on textual input and assess whether these selections can be effectively combined with traditional mean variance optimization. Their findings suggest that LLM assisted portfolios can outperform random benchmarks in out of sample testing, especially when the model is used as a pre-screening tool for investment idea generation. The study highlights the potential of LLMs to contribute qualitative insights in settings where structured data is limited.

Similarly, Ko & Lee (2024) look at how ChatGPT can be used to suggest stocks and find that combining its recommendations with traditional portfolio optimization can potentially lead to lower risk and better risk adjusted returns. However, they also point out important limitations of LLMs, such as imprecise calculations, inconsistent answers depending on how questions are asked, and difficulty understanding how the model reaches its conclusions. These issues reflect common concerns about the lack of transparency in AI generated outputs, especially in areas like finance and banking where clear reasoning is important.

Overall, the literature presents a cautiously optimistic perspective on the role of AI in portfolio optimization. While ML and LLMs have expanded the range of tools available to asset managers, their effectiveness is still shaped by factors such as data quality, model complexity, and the inherent uncertainty of financial markets. Most studies suggest that AI is unlikely to replace traditional financial approaches but can serve as a valuable complement especially when used in hybrid systems that balance analytical power with interpretability and human judgment. These dynamics will be examined and discussed, with a focus on the comparative performance and transparency of AI-driven portfolio strategies.

4. Methodology

4.1 Research Design

Elaborating on the previously shown *Figure 1: Thesis Structure Overview*, this thesis adopts a primarily quantitative research approach based on secondary data, complemented by a review of relevant academic literature and industry reports. By integrating empirical analysis with insights from current financial practice, the study aims to evaluate the technical performance of advanced AI models as well as their broader role and implications in the banking sector.

The empirical analysis draws on publicly available datasets selected for their relevance to real world banking applications. Although the data was not collected firsthand, it reflects the types of information commonly used by financial institutions, including transaction records, credit attributes, and asset price data. Using preexisting data supports transparency and reproducibility, while also allowing for a simplified yet realistic view of how AI can be applied in banking. This simplification is necessary due to limited access to actual financial data and the significant computational demands associated with training advanced models on large scale datasets challenges that fall outside the scope of this thesis.

Instead of relying on a single experiment or dataset, this thesis evaluates multiple algorithms across various datasets tied to different banking applications. This broader approach helps avoid drawing overly general conclusions from limited or narrow cases. While testing models in several contexts strengthens the reliability of the findings, it still does not justify firm claims about the overall superiority of advanced AI methods. Rather, the results contribute to a growing understanding of where these techniques may add value and under what conditions their use is most promising.

Whereas much of the existing literature concentrates on individual applications, this thesis takes a wider view by linking multiple financial tasks. The purpose is to explore how AI could be integrated responsibly into banking practices particularly in areas where it offers clear, data driven advantages. The intention is not to offer final answers, but to support continued discussion within both academic and industry settings by connecting model performance to practical demands and constraints in banking.

4.2 Technical Setup

The empirical analysis for the thesis was carried out using Python, a widely used programming language in data science and finance. Python is especially useful because it supports fast and flexible analysis of datasets, allows for the development of predictive models, and integrates well with tools for data visualization and reporting. It provides access to a wide range of libraries, which are collections of prewritten code that simplify common tasks such as preparing data, building models, and evaluating results. These libraries

help streamline the research process by reducing the need to write code from scratch. All analyses were performed on a local desktop computer. While this setup does not offer the computational power required for training very large scale or highly complex AI models, it was sufficient for the scope and objectives of this thesis. Working locally provided full control over the computing environment and allowed for efficient testing, adjustment, and replication of experiments using moderately sized datasets and models.

The complete code and datasets have been uploaded to GitHub[1], an online platform for storing and sharing code. This is done to ensure transparency in the research process, allowing the methods and analysis steps to be openly documented.

4.2.1 Use of Generative AI in the thesis

While this thesis explores the role of AI in banking, it also benefits from the use of AI tools. More specifically, the generative AI models ChatGPT 4o an Grok 3 are used, alongside applications like Grammarly and Writefull, to improve the clarity, flow, and readability of the text. These tools provided support with grammar correction and sentence structure. In addition, AI assisted coding tools such as GitHub Copilot supported the technical work by helping to correct syntax errors and resolve bugs in the code used for data processing and model testing. Importantly, all core content, ideas, analysis and conclusions were developed independently, based on personal academic interest and informed by previous empirical studies and literature on AI.

4.3 Data Sources

As stated, this thesis relies on publicly available datasets related to fraud detection, credit risk, and portfolio optimization. These publicly accessible datasets enable experimentation with ML in environments that to some degree resemble real financial settings, though limitations remain in terms of level of detail, how realistic the data is, and whether it includes all the information used in actual bank operations. Compared to fraud and credit assessment related data, which are often anonymized or synthetic due to their sensitive nature, portfolio optimization data is more openly available. The datasets selected for this study were chosen based on availability, relevance to each topic, and their ability to support meaningful analysis within the scope of the thesis and the problem statement.

Since the datasets were relatively clean and well structured, only limited exploratory data analysis was required to prepare them for modeling. This contrasts with real world financial data, which is often messy, incomplete, or inconsistent, and typically requires extensive preprocessing steps such as handling missing values and transforming variables before meaningful analysis can begin.

[1] `https://github.com/BjerringNK`

Data for Fraud Detection

Two datasets from Kaggle, a platform for data science competitions and dataset sharing, are used to examine fraud detection in the banking sector. These datasets provide testing for both traditional and advanced ML models while reflecting different types of financial fraud scenarios.

The first dataset[2] (Fraud Dataset A) was developed through a research collaboration between Worldline and the Machine Learning Group of the Université Libre de Bruxelles. It contains 284.807 real credit card transactions made by European cardholders in 2013. To protect sensitive information, the original features were transformed using Principal Component Analysis (PCA), a technique commonly used to reduce dimensionality while preserving the most important patterns in the data. This transformation also serves to anonymize the original variables. As a result, most input features are numerical and expressed as principal components (e.g., `V1, V2, ..., V28`), making the dataset well suited for ML without requiring categorical encoding. While this makes the dataset highly usable and well studied in academic literature, it also means that the specific contribution of each original variable to a fraudulent or non fraudulent outcome cannot be directly interpreted. A full list of the variables is available in Appendix A.

The second dataset[3] (Fraud Dataset B) consists of 6.362.620 synthetic financial transactions generated using the PaySim simulator. The simulator is based on real-world aggregated mobile money transaction data and systematically introduces fraudulent activity to support the testing and benchmarking of fraud detection models (Lopez-Rojas, 2018). Unlike Fraud Dataset A, this dataset does not use dimensionality reduction techniques such as PCA. As a result, the original variables remain intact, offering greater transparency and interpretability. Key variables include:

- `step`: A unit of time in hours, representing the chronological order of transactions across a simulated 30-day period.
- `type`: Type of transaction (e.g., Payment, Transfer, Cash Out)
- `amount`: The monetary value of each transaction, essential for detecting anomalies.
- `oldbalanceOrg`: Account balance of the sender before the transaction.
- `newbalanceOrig`: Sender's balance after the transaction.
- `isFraud`: Binary indicator of whether the transaction was fraudulent.

The dataset structure allows for testing of classification models in a highly imbalanced setting, mirroring the distribution of fraud in actual financial systems. A full list of variables is available in Appendix B.

[2] `https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud/data`
[3] `https://www.kaggle.com/datasets/ealaxi/paysim1/data`

Data for Credit Risk Assessment

As with fraud detection, two distinct datasets from Kaggle are used in this thesis. Incorporating multiple datasets helps ensure that conclusions about the effectiveness of advanced versus conventional models are not limited to a single data structure or scenario.

The third dataset[4] (Credit Risk Dataset A) consists of 32.581 synthetic loan applications designed to reflect the type of structured data typically used by financial institutions in credit risk assessments. The target variable, `loan_status`, indicates whether a loan was fully repaid or defaulted, forming the basis for classification modeling. The dataset contains a mixture of demographic and financial variables relevant to lending decisions. Among the most important variables are:

- `person_income`: Annual income of the applicant, a key indicator of financial capacity and creditworthiness.
- `loan_amnt`: The amount requested, which combined with income helps assess affordability and default risk.
- `loan_int_rate`: The interest rate on the loan, reflecting risk pricing by the lender.
- `loan_percent_income`: A derived metric showing the loan as a proportion of income, similar to a debt-to-income ratio.
- `cb_person_default_on_file`: Indicates past defaults
- `cb_person_cred_hist_length`: The number of years of credit history, relevant for distinguishing between experienced and new borrowers.

This dataset gives a foundation for testing and comparing various classification algorithms in credit risk assessment. A complete list of variables is available in Appendix C.

The fourth dataset[5] (Credit Risk Dataset B) is also a synthetic dataset consisting of 8.250 customer informations. It includes encoded demographic attributes and payment-related behavior, with the target variable classifying each customer as either High Risk or Not Risky. A full list of variables is provided in Appendix D, but among the most important variables are:

- `OVD_t1`, `OVD_t2`, `OVD_t3` – Indicators of the number of overdue payments by type, capturing different severities or durations of delinquency.
- `OVD_sum`: The total number of overdue days across all types, offering a cumulative measure of repayment problems.

[4] `https://www.kaggle.com/datasets/laotse/credit-risk-dataset/data1`
[5] `https://www.kaggle.com/datasets/praveengovi/credit-risk-classification-dataset/data`

- `pay_normal`: The number of times the customer made a normal (on-time) payment, reflecting overall payment behavior.
- `prod_limit`: The credit limit of the financial product
- `new_balance`: The most recent balance on the customer's credit product
- `label`: The target variable indicating high or low credit risk.

Both the fraud detection and credit risk datasets used suffer from significant class imbalance, where the minority class represents only a small portion of the data. Without addressing this imbalance, models tend to favor the majority class, resulting in poor detection of the minority class.

To address the issue of class imbalance in the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) is used. SMOTE helps improve the performance of ML models by generating new synthetic examples of the minority class. It does this by selecting existing samples from the minority class and creating new ones that are similar, based on nearby examples in the data. This approach adds variety without simply duplicating the data, helping the model potentially better learn the characteristics of rare cases, such as fraud or loan defaults. It allows the model to pay more attention to patterns that would otherwise be overlooked due to the small number of positive examples. Because of its simplicity and proven results, it is widely used in both academic research and practical applications (Chawla et al., 2002).

Data for Portfolio Optimization

In contrast to the structured, readymade datasets used in the fraud detection and credit risk classification components of this thesis, the portfolio optimization section is based on datasets that were constructed manually using publicly available data retrieved through APIs. This approach allowed for more flexibility and control in tailoring the datasets to the modeling needs of each portfolio experiment. Specifically, three separate datasets were assembled to support three distinct tests, each involving different levels of complexity. The data collection and preparation process is described here in general terms, while the exact time periods and variables used in each test are presented in the empirical results section.

In the first test, historical price data was collected for four U.S. stocks and one index using the yfinance library, which accesses Yahoo Finances financial data archives. The adjusted close price was used as the basis for all return calculations, as it incorporates dividends and stock splits and is considered the standard for accurate performance measurement. Weekly returns were computed as the percentage change in adjusted close prices from one week to the next. No alternative data was included in this setup.

For the second test, the dataset was expanded to include 20 U.S. selected randomly from a shortlist of large-medium and small cap. Alongside adjusted close prices, two categories of alternative data were incorporated

macroeconomic indicators and Google Trends search interest data. Macroeconomic variables, such as interest rates, inflation measures, and employment figures, were collected from official sources via the Federal Reserve Economic Data (FRED) API. Google Trends data was retrieved using the pytrends Python library, which provides weekly search interest scores. For each stock, data was collected for both the company name and its corresponding ticker symbol. These features were merged and aligned by date to match the weekly frequency of the financial data. All alternative variables were standardized prior to modeling to ensure consistency in scale across features.

In the third and final test, the dataset was further expanded to include adjusted close prices for all companies listed in the S&P 500 index as of 2025. The same preprocessing steps were applied, including weekly alignment and handling of missing values. As in the second test, the 3-month U.S. Treasury bill rate was included as a proxy for the risk-free rate.

Across all three tests, the datasets were prepared to ensure consistency in both time and format. Returns were calculated using log returns based on adjusted close prices. In cases where data was missing, values were either forward filled using the most recent available data or removed entirely, depending on the extent and significance of the missing information. The exact time periods used in each test are detailed in the empirical results section.

4.4 Algorithms and Methods Used

This section presents the algorithms and methods employed in the thesis to examine the key areas of banking. The objective is not to provide a detailed mathematical derivation of advanced algorithms such as the ensemble ones, but rather to offer a clear understanding of what these models are, how they are applied, and what they can contribute in practical settings. The section aims to highlight the strengths and limitations of each approach, as well as the types of insights they can generate. Particular emphasis is placed on model explainability, as the ability to understand and communicate model predictions is critical.

Fraud Detection and Credit Assessment

Fraud detection and credit risk assessment are both framed as classification problems, meaning the goal is to predict whether a case belongs to one of two categories such as fraud or not fraud, default or no default. Because of this similarity, a common modeling approach was used for both tasks. As explained in Section 2.2.3 *(Traditional Versus AI-Powered Methods)*, this thesis applies a combination of traditional and more advanced ML models to evaluate performance in these areas. The selected supervised learning models Logistic Regression, Decision Tree, Random Forest, and XGBoost represent a range of complexity, from more simple and interpretable approaches to more advanced ensemble based algorithms. To try an enhance performance, two unsupervised learning techniques were also applied K-Means clustering and Isolation Forest.

Logistic Regression

Logistic regression is a foundational statistical model commonly applied in the banking sector for binary classification problems such as fraud detection and credit risk assessment. It is particularly valued in finance for its balance of predictive power, interpretability, and its compatibility with regulatory requirements. According to James et al. (2013), logistic regression models the probability that a binary outcome Y equals 1 (e.g., fraud or default) given a set of explanatory variables $X = (x_1, x_2, ..., x_n)$. The probability function $p(X)$ is expressed using the logistic function:

$$p(X) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n}} \tag{1}$$

This formulation ensures that predicted probabilities lie strictly between 0 and 1, making it suitable for classification tasks where the response is either 0 or 1.

The logistic function can also be expressed in terms of the *odds*, which represent the ratio between the probability of an event occurring and not occurring:

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n} \tag{2}$$

Taking the natural logarithm of both sides gives the log-odds or *logit* form of the model:

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \tag{3}$$

This expression reveals that while the relationship between the predictors and the probability $p(X)$ is non-linear, the log-odds of the outcome is a linear function of the predictors.

- The intercept β_0 represents the log-odds of the outcome when all input features are equal to zero. It serves as a baseline from which the influence of other variables is added.
- Each coefficient β reflects the change in the log-odds of the outcome for a one-unit increase in x, holding all other variables constant.
- The exponential of the coefficient, e^{β}, represents the odds ratio, indicating how the odds of the outcome change with a one-unit increase in the corresponding variable.
- Since the relationship between X and $p(X)$ is non-linear, the change in predicted probability depends on the current value of X.

The coefficients $\beta_0, \beta_1, \ldots, \beta_n$ are estimated from data using maximum likelihood estimation. This approach identifies the values of the parameters that make the observed outcomes most probable under the logistic model (James et al., 2013). The likelihood function for n independent observations is:

$$\ell(\beta_0, \ldots, \beta_n) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i)) \tag{4}$$

Because of its interpretable structure, logistic regression remains a standard baseline in supervised classification tasks in finance. It allows analysts and decision makers to not only predict outcomes but also understand the contribution of individual input features in a transparent and statistically grounded way.

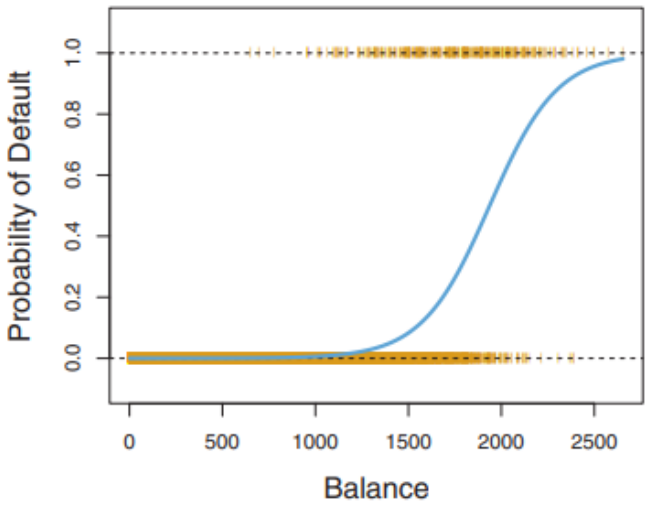

Figure 8: Logistic regression curve for binary classification

Source: Adapted from James et al. (2013)

Figure 8 shows an example of logistic regression where the probability of default is modeled using the feature *Balance*. As the balance increases, the predicted probability of default also rises, forming a smooth S-shaped curve typical of the logistic function.

Decision Tree

A decision tree is a supervised ML algorithm and can be used for both classification and regression tasks. It works by splitting the dataset into smaller groups based on feature values, using a tree like structure made up of decision nodes and branches. Each internal node checks a condition on a specific feature such as whether a value is above or below a threshold, and each branch represents the outcome of that test. This process continues until the data can no longer be split, and a final decision is made at a leaf node which could be classifying a customer as high risk or low risk (James et al., 2013).

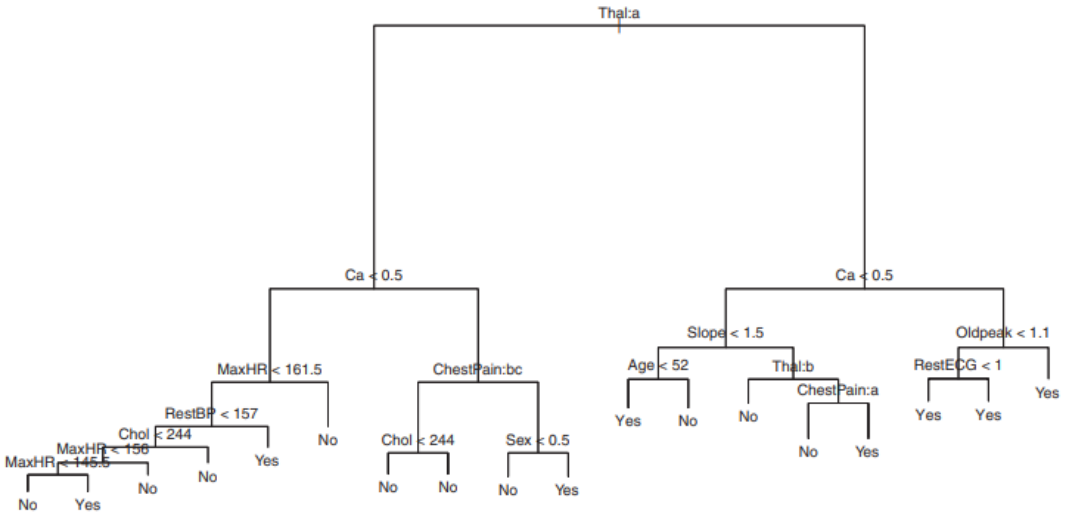

Figure 9: Decision Tree

Source: James et al. (2013)

Decision trees are well known for their interpretability and ease of use. The step by step logic visible in Figure 9 makes it possible to follow how input features lead to a final decision. Decision trees can work with both numerical and categorical variables and require relatively little data preparation. Decision trees can also capture non linear patterns in the data. However, there are some limitations. If the tree becomes too deep or complex, it may fit the training data too closely and perform poorly on new data. Decision trees can also be sensitive to small changes in the dataset, which may result in different tree structures. Although certain adjustments can help address these issues, decision trees do not always perform as well as more complex algorithms. Still, the simplicity and ability to explain predictions make it valuable (James et al., 2013).

Bagging and Random Forests

Bagging is an ensemble method designed to reduce the variance of high-variance models such as decision trees. As outlined by James et al. (2013), the core principle involves generating multiple versions of the training dataset by sampling with replacement—producing what are known as bootstrap samples. Each bootstrap sample is then used to train an individual model, typically a decision tree.

Let B represent the total number of decision trees to be constructed. For each $b = 1, ..., B$, the following steps are carried out:

- A bootstrap sample of the same size as the original dataset is generated by randomly sampling observations with replacement.
- A model $\hat{f}^{(b)}(x)$ is trained on this bootstrap sample.

- All B model predictions are aggregated to form the final output.

The overall bagged prediction function is:

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{(b)}(x) \tag{5}$$

where:

- $\hat{f}^{(b)}(x)$ is the prediction from the b^{th} tree trained on a bootstrap sample,
- B is the of such trees,
- The final prediction is obtained by averaging (for regression) and majority voting (for classification).

Bagging is particularly effective for models like decision trees, which tend to overfit. However, it can still produce correlated trees when one or more strong predictors dominate the splits, limiting the benefit of averaging (James et al., 2013).

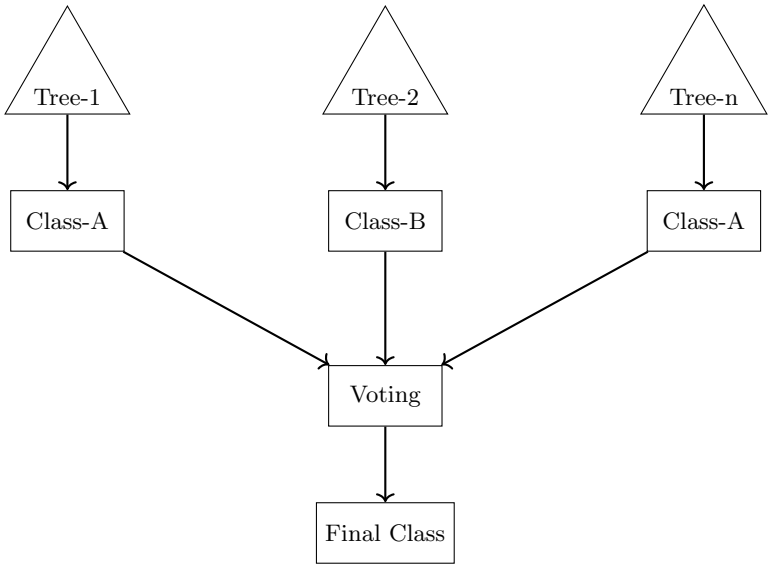

Figure 10: Random Forest classification via majority voting

Source: Own illustration, adapted from James et al. (2013)

Random Forests build upon bagging by injecting an additional layer of randomness. In addition to training each tree on a bootstrap sample, random forests also randomize the feature selection at each split. Specifically:

- Here p is the total number of predictors.
- At each split in the tree, only a random subset of m predictors ($m < p$) is considered.

This random feature selection ensures that not all trees use the same strong predictors at every split, thereby reducing correlation among trees and improving model variance reduction. When $m = p$, the model reverts to standard bagging. A key trade-off of random forests is reduced interpretability. Since predictions are made by aggregating many trees, each following different splitting paths, understanding individual predictions can become complex (James et al., 2013).

Boosting and XGBoost

Boosting is an ensemble method that improves predictive performance by combining a sequence of shallow decision trees, into a single strong learner. Unlike bagging, which builds trees independently on separate bootstrap samples, boosting fits each tree sequentially, where each new tree attempts to correct the errors made by the previous ensemble. As explained by James et al. (2013), boosting does not rely on bootstrap sampling. Instead, all trees are fit to different versions of the same training data, with each iteration focusing on the errors made by the previous model. The model is built in an additive manner, updating the function at each step:

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^{(b)}(x) \tag{6}$$

- $\hat{f}^{(b)}(x)$ is the output of the b^{th} weak learner, typically a shallow tree.
- λ is the learning rate, controlling how much influence each new tree has on the final prediction.
- B is the number of boosting iterations, where each iteration adds a tree to correct the errors from the previous model.
- Each new model is trained on the residuals, allowing the ensemble to improve incrementally over time.

To further enhance the base boosting method, **XGBoost** introduces several practical improvements that make it highly effective, especially for structured data problems. It includes mechanisms that penalize model complexity, helping to prevent overfitting when too many trees or overly deep trees are used. Additionally, it increases the diversity among trees by randomly selecting a subset of input features at each step, which reduces correlation and improves generalization. XGBoost removes unnecessary tree branches and uses parallel computing to speed up training. This makes it especially good for large datasets and strong performance in real world tasks (Chen & Guestrin, 2016).

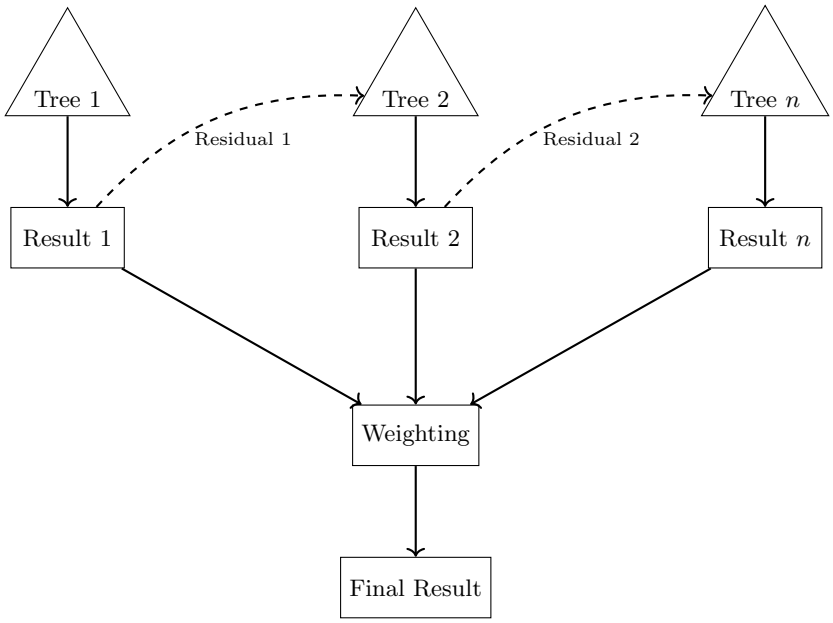

Figure 11: Boosting architecture: sequential tree fitting

Source: Own figure based on James et al. (2013)

Just like other ensemble methods such as Random Forest, XGBoost also suffers from reduced interpretability. Its predictions result from the combined effect of many trees added over time, which makes it harder to clearly understand the influence of individual features on specific predictions.

Unsupervised Learning: K-Means Clustering and Isolation Forest

K-Means Clustering is a commonly used unsupervised learning algorithm that groups data into clusters based on similarity. Each data point is assigned to the nearest cluster center, called a centroid, which is updated until the centroids stop changing significantly. To find a good number of clusters, the elbow method is often used. It involves plotting the models error for different numbers of clusters and looking for the point where the improvement slows down. This point, known as the elbow helps choose a number of clusters that balances accuracy and simplicity (Géron, 2022).

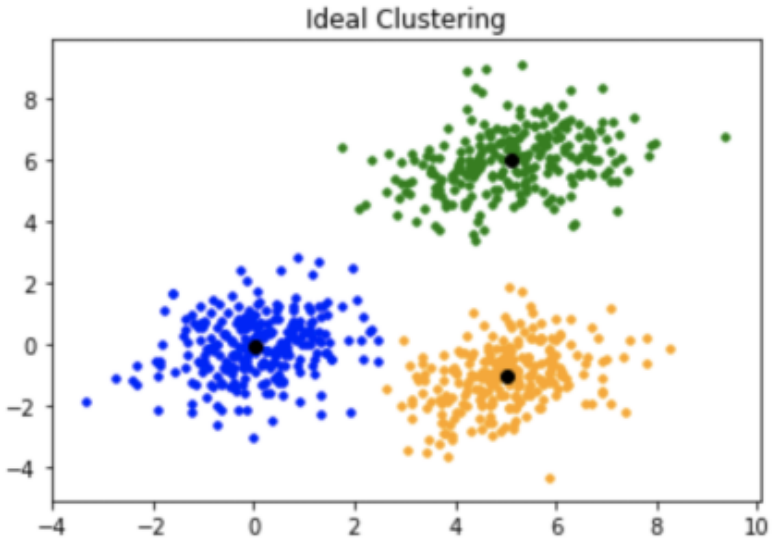

Figure 12: Ideal clustering structure using K-Means

Source: GeeksforGeeks.com (n.d.)

An ideal outcome for K-Means is when clusters are clearly separated, similarly sized, and tightly grouped. Figure 12 illustrates such a case, where the structure of the data aligns well with the assumptions of the algorithm. In these situations, K-Means is both efficient and effective. In practice, however, real world financial data often shows more complexity. Clusters may overlap, differ in size or shape, or include noise and outliers. This can make K-Means less reliable, as it may form groupings that do not reflect meaningful patterns. Still, K-Means can be a valuable tool especially when there is reason to believe that the data contains natural groupings. In areas such as credit analysis and fraud detection, clustering can in theory help identify different customer segments or detect unusual behavior that may warrant further investigation (Géron, 2022).

In contrast to K-Means, which aims to identify the general groupings or patterns across datasets, **Isolation Forest** is designed specifically for anomaly detection. Isolation Forest works by randomly selecting a feature and a split value to isolate individual observations. The number of random splits required to separate a data point from the rest of the dataset becomes a measure of how unusual that point is. Observations that can be isolated with fewer splits are more likely to be anomalies, as they deviate more strongly from the overall data structure.

Isolation Forest is particularly useful in high dimensional settings and performs well even when the underlying data distribution is complex or unknown. Unlike clustering methods like K-Means, it does not assume any specific shape or size of clusters, which makes it robust in detecting unusual or rare patterns that may otherwise go unnoticed (Géron, 2022). This makes it especially applicable in the context of this study, where

anomalies are rare but highly significant. When used together, K-Means and Isolation Forest can potentially provide complementary insights. While K-Means helps uncover broader structures and potential segments in the data, Isolation Forest focuses on identifying observations that fall far outside those patterns.

Portfolio Optimization

Unlike fraud detection and credit risk assessment, which in this thesis is framed as classification problems, portfolio optimization focuses on forecasting and decision making based on continuous financial variables. The following sections introduce several modeling approaches used to predict asset returns, ranging from classical statistical methods and regression versions of models previously used for classification, including the Random Forest Regressor and XGBoost Regressor as well as deep learning techniques. These models are evaluated to understand how different forecasting strategies influence portfolio optimization outcomes.

Time Series Forecasting with ARIMA, LSTM and GRU

The AutoRegressive Integrated Moving Average (ARIMA) model is a widely used statistical method for time series forecasting. It is designed to capture linear relationships in time series data by using its own past values and past forecast errors. ARIMA is especially effective when working with stationary time series data, where the statistical properties such as the mean and variance remain constant over time (Siami-Namini et al., 2018).

The ARIMA model consists of three main components:

- **AR (AutoRegressive)**: A component that uses a linear combination of past observations to predict the current value. The number of lagged terms is denoted by p.
- **I (Integrated)**: This refers to differencing the series d times to make it stationary. Stationarity is a necessary condition for applying ARIMA effectively.
- **MA (Moving Average)**: A component that incorporates past forecast errors to correct current predictions. The number of lagged error terms is denoted by q.

The general form of an ARIMA(p, d, q) model, after differencing d times to achieve stationarity, is expressed as:

$$x_t = c + \sum_{i=1}^{p} \phi_i x_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \tag{7}$$

- x_t: Differenced time series value at time t,
- c: Constant term (intercept),
- ϕ_i: Autoregressive coefficients,

- θ_j: Moving average coefficients,
- ε_t: Random white noise error at time t.

For ARIMA to generate valid forecasts, several assumptions must hold:

- **Linearity**: The model assumes that the time series can be represented as a linear function of past values and errors.
- **Stationarity**: The time series must be stationary.
- **White Noise Errors**: Residuals (ε_t) must be uncorrelated, have constant variance, and have a mean of zero.
- **Normality**: For constructing prediction intervals, the residuals are often assumed to follow a normal distribution, i.e., $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$.

To ensure valid ARIMA forecasts, stationarity was addressed by transforming price data into returns, and confirmed using the Augmented Dickey-Fuller test. Residual diagnostics included ACF and PACF plots, the Ljung-Box test for independence, and checks for constant variance. Models were fitted on training data and evaluated on separate test data.

Long Short-Term Memory (LSTM) networks are a type of deep learning model specifically designed to handle sequential data, such as time series. LSTM is part of the Recurrent Neural Network (RNN) family, which is good for analyzing data where the order of information matters. These models are built to retain information from previous time steps and use it when making predictions, allowing them to capture patterns that unfold over time. This makes LSTM particularly useful for financial forecasting tasks, where past trends and dependencies can perhaps influence future outcomes (Siami-Namini et al., 2018). Traditional RNNs have trouble remembering information over long sequences, which makes it hard for them to learn patterns that depend on what happened much earlier in the data. To solve this problem, LSTM networks were developed. LSTM improves the basic RNN by adding a special memory system called the cell state, along with gates that control what information should be kept, updated, or forgotten. These gates called the input gate, forget gate, and output gate help the model remember important information for longer and ignore whats not useful (Siami-Namini et al., 2018).

LSTMs ability to manage memory over long sequences makes it well suited for capturing complex, nonlinear patterns in time series data. Because of this, LSTM has become a popular choice for forecasting tasks, especially for the financial markets where data is often volatile, dynamic, and nonstationary. A related model, the **Gated Recurrent Unit (GRU)**, is a much newer and simplified version of LSTM. GRU

combines the key functions of LSTMs memory system into just two gates a reset gate and an update gate. It does not use separate memory cells, which makes the architecture simpler and the model smaller. As a result, GRUs often train faster and require fewer computational resources, making GRU a possible choice when data is limited and efficiency is important (Chung et al., 2014)

Mean-Variance Analysis and Optimization

Mean-variance analysis, as presented in Section 3.3, forms the foundational framework for analyzing the tradeoffs between risk and return in portfolio construction. It evaluates how different combinations of assets perform along the risk return spectrum, aiming to identify portfolios that either maximize expected return for a given level of risk or minimize risk for a desired level of return (Elton et al., 2014):

1. **Expected Portfolio Return:**

$$\mathbb{E}(R_p) = \sum_{i=1}^{n} w_i \mathbb{E}(R_i) \tag{8}$$

   - $\mathbb{E}(R_p)$: The expected return of the portfolio, calculated as the weighted average of the expected returns of individual assets.
   - w_i: The proportion of the total portfolio invested in asset i.
   - $\mathbb{E}(R_i)$: The expected return of asset i.

2. **Portfolio Variance:**

$$\sigma_p^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \sigma_{ij} \tag{9}$$

   - σ_p^2: The variance of the portfolio's return, representing total risk.
   - σ_{ij}: The covariance between the returns on assets i and j.
   - w_i, w_j: Portfolio weights on assets i and j, respectively.

Building on mean-variance analysis, mean-variance optimization provides a structured approach to selecting portfolio weights that maximize the risk adjusted return (Elton et al., 2014). This is commonly achieved by maximizing the Sharpe Ratio:

$$\text{Sharpe Ratio} = \frac{\mathbb{E}(R_p) - R_f}{\sigma_p} \tag{10}$$

- $\mathbb{E}(R_p)$: The expected return on the portfolio.
- R_f: The risk-free rate of return.
- σ_p: The standard deviation of the portfolio's return, representing total risk.

In the study, the optimization is subject to a full investment constraint, requiring that the sum of all portfolio weights equals one, and a non negativity constraint, which prevents shortselling by ensuring all weights are zero or positive. The resulting optimal portfolio lies on the efficient frontier, which represents the set of portfolios offering the best possible trade off between expected return and risk (Elton et al., 2014).

Mean-Variance Optimization using AI

Mean-variance optimization using historical returns is the traditional statistical approach to portfolio construction, offering a structured way to balance risk and return based on historical estimates. However, AI-based algorithms and methods such as Random Forest Regressor and XGBoost Regressor can be used to potentially enhance the prediction of future asset returns. These models cam be trained on a broad set of input features, which can include macroeconomic indicators, trend signals, technical variables, and other relevant market data. By learning complex, nonlinear relationships in the data, these models try to generate more accurate and timely return forecasts. The predicted returns are then incorporated into the mean-variance optimization process, effectively replacing static historical averages with dynamic estimates derived from alternative data sources and model based forecasting. This hybrid approach preserves the original structure of MPT while providing a way to test whether using AI can potentially lead to better performing portfolios in practice.

Generative AI

As outlined in Section 2.2.4, generative AI is already widely used in the banking sector to improve productivity in tasks such as summarising financial reports, extracting key insights, and drafting communication. In these cases, LLMs typically function as supportive tools, operating under human supervision or within tightly defined workflows. In this experiment, however, the model takes on a more autonomous role. Although the task is framed by a structured prompt, ChatGPT is responsible for the entire portfolio construction process from analysing anonymised price data to deciding how capital should be allocated across assets. There is no intermediate guidance, code assistance, or finetuning. This setup allows for an evaluation of how well a general purpose LLM can perform when asked to make investment decisions independently, within a clearly defined but selfcontained problem. The structure of the experiment is inspired by Perlin et al. (2025), who used Googles Gemini model to construct portfolios using anonymised financial statements and market data. Their results showed no consistent Gemini outperformance relative to a naive equal-weight portfolio or the S&P 500 index, and their conclusion explicitly called for comparative evaluations using other models:

> *"Future research could evaluate Geminis performance in comparison to other large language models, such as ChatGPT, to assess its relative effectiveness"*

This study extends the experimental framework introduced by Perlin et al. (2025) by evaluating how a different LLM (ChatGPTs o3) performs under similar anonymised and controlled conditions. Building on

their methodology, the simulation design preserves the key elements of anonymisation, return based decision making, and benchmark comparison. The data consist of weekly adjusted prices for 503 anonymised S&P 500 companies from 2015 to 2024, with all firm names and calendar dates removed. The stocks prices are scaled by a random factor drawn from the interval 0.01-0.99, maintaining the return dynamics while concealing identity. The weekly U.S. T-bill rate is used to represent the risk-free return throughout the simulations.
A practical modification is introduced in how the model is accessed. ChatGPT is used interactively, without API automation. This setup provides a freely accessible and low barrier alternative, allowing for the examination of LLM based portfolio construction without advanced infrastructure and coding. All decisions regarding stock selection and capital allocation are made by ChatGPT based solely on a single prompt. No further guidance, follow up questions, or adjustments are provided. To ensure full transparency and reproducibility, the exact prompt given to ChatGPT is presented below. The prompt was made with assistance from Grok 3:

```
Role
You are a financial analyst who specialises in empirical asset-pricing research and systematic portfolio
design.


Objective
Create and evaluate ChatGPTs Strategy|a rules-based allocation of USD 10 000 across five randomly-selected
equities and a risk-free asset|using nothing but historical price information.  Compare its performance
with:
Naive Strategy fixed equal weights in the same five stocks plus the risk-free asset (one-sixth of capital
in each).
S&P 500 index passive benchmark.


Data supplied
Weekly adjusted prices for 503 anonymised stocks (\Stock 1" ...  \Stock 503").
Weekly risk-free returns (Tbill_Weekly_Return).
Weekly S&P 500 index levels (benchmark only).
Stock prices have been rescaled by a random linear factor in [0.01, 0.99] to hide identities while preserving
return properties.  No company identifiers or dates are provided.


Task 1 Strategy Definitions
Strategy:  ChatGPTs Strategy
Devise a logical, price-based weighting scheme (e.g., momentum or mean-reversion) that distributes the
USD 10 000 across the five stocks and the risk-free asset.
Constraints:

  • Long-only, no leverage.
  • Weights ≥ 0 and sum to 1.0.
  • Base every decision solely on anonymised past prices.
  • Briefly justify the rule in your report.
```

Strategy: Naive Strategy
Equal weights in all six investable assets: 1/6 of capital in each of the five stocks and 1/6 in the risk-free asset.
Same long-only, full-investment constraints.

Task 2 Simulations
Evaluate both strategies across four investment horizons:
Horizon Weeks: 4 (1 month), 24 (6 months), 52 (12 months), 156 (36 months)
For each horizon, run 1 500 simulations following the steps below:

- Random start date: Draw a start week uniformly such that sufficient forward data exist for the horizon.
- Random stock set: Select five distinct stocks satisfying, over the 5-year look-back period ending at the start week:
  - complete weekly price history (no missing values),
  - no weekly price below USD 1.
- Form allocations: Determine weights for ChatGPT's Strategy using only the anonymised price histories. Apply the Naive 1/6-each rule.
- Hold period: Compute weekly portfolio returns over the horizon.

Task 3 Performance Metrics
For every simulation calculate:

- Weekly portfolio returns for ChatGPT, Naive, and the S&P 500.
- Beta (OLS slope versus S&P 500).
- Annualised total return.
- Annualised Sharpe ratio.

Task 4 Reporting Requirements
Produce two tables:

- Table 1 { Summary Statistics (median across the 1 500 runs per horizon):
  Horizon, Simulations, Beta (median), Annualised Return (median), Annualised Sharpe (median)
- Table 2 { Comparative Performance (percentage of simulations where ChatGPT's Strategy outperforms):
  Horizon, Simulations, % Returns > S&P 500, % Sharpe > S&P 500, % Returns > Naive, % Sharpe > Naive

Constraints & Implementation Notes

- Allocation decisions must rely exclusively on the anonymised price histories; no fundamentals or external data.
- Handle any missing values by forward-filling or discarding incomplete series.
- No short selling or leverage; weights are non-negative and sum precisely to USD 10 000.
- Document and explain the logic behind ChatGPTs Strategy clearly in your output.

ChatGPT model o3 was selected for this experiment due to its state of the art reasoning capabilities and its position at the forefront of analytical AI models. Released in December 2024, o3 is specifically designed for complex, logic driven tasks that require deep pattern recognition, structured decision making, and mathematical precision. With support for an extended context window and strong performance across technical benchmarks (OpenAI, 2024)

Hyperparameter Tuning

In this thesis, hyperparameter tuning was applied to each model to improve predictive performance beyond default configurations while ensuring a fair and consistent basis for comparison. Default settings, while easy to implement, are often not optimal for specific datasets and can lead to either underfitting or overfitting. As such, model performance can be significantly enhanced by tuning key hyperparameters that govern the structure and learning behavior of each algorithm. To balance the trade off between training time and identifying strong model settings, this thesis used RandomizedSearchCV instead of a full grid search. RandomizedSearchCV is a hyperparameter tuning tool available in the Python library scikit-learn. It works by randomly sampling a fixed number of combinations from a specified range of possible hyperparameter values, rather than exhaustively testing every single combination. This makes it much more efficient when working with complex models and large datasets, where grid search can quickly become too slow or computationally expensive. Despite its faster runtime, `RandomizedSearchCV` still provides good coverage of the parameter space and is widely used in practice for tuning models with many hyperparameters.

For each model, relevant hyperparameters were selected based on official documentation. While simpler algorithms like logistic regression has relatively few tuning options and was used with mostly standard settings, the ensemble models required more detailed tuning. The most influential hyperparameters for these models include:

Random Forest

- `n_estimators`: Number of decision trees in the forest.
- `max_depth`: Maximum depth of each tree.
- `min_samples_split`: Minimum number of samples required to split an internal node.
- `max_features`: Number of features considered when looking for the best split.

XGBoost

- `learning_rate`: Controls the contribution of each tree to the overall prediction.
- `n_estimators`: Number of boosting rounds/ trees added sequentially.
- `max_depth`: Maximum depth of each tree.

- `subsample`: Fraction of the training data used for each boosting iteration.
- `colsample_bytree`: Fraction of features randomly sampled for each tree.

Tuning was performed using 5-fold cross-validation on the training data to ensure that selected hyperparameter configurations generalized well to unseen data. Tuning ranges were chosen to be realistic and based on values commonly used in practice, ensuring that no model was given an unfair advantage. The goal was not to maximize individual model performance at all costs, but to evaluate models under consistent and comparable conditions. A fixed random seed of 10 was applied throughout the experiments to ensure reproducibility. Full details on the tuning ranges, selected values, and complete training code are available on GitHub.[6]

4.5 Evaluation Metrics

Fraud Detection and Credit Assessment

To assess the performance of the classification models used in fraud detection and credit risk assessment, this thesis applies evaluation metrics commonly used in existing and aforementioned literature. As highlighted in the literature review (e.g., Johora et al., 2024; Faisal et al., 2024; Bello et al., 2023), metrics such as precision, recall, F1-score, and AUC are standard tools for evaluating AI models in classification tasks, particularly in scenarios involving class imbalance. All reported performance results are based on a separate testing set that was held out from training to provide an unbiased estimate of each models generalization ability. A consistent 70/30 train-test split was applied across all experiments, and models were never evaluated on the same data used for training.

Precision measures the proportion of predicted positive cases that are actually positive. High precision indicates a low false positive rate.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{11}$$

Recall, also known as sensitivity or true positive rate, measures the proportion of actual positive cases that are correctly identified.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{12}$$

F1-score is the harmonic mean of precision and recall. It provides a more balanced measure when there is an uneven class distribution or when both false positives and false negatives are important.

[6] `https://github.com/BjerringNK`

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

In fraud detection, recall is often viewed as an important metric, as it reflects the models ability to detect a large share of fraudulent cases. Missing fraud can lead to compliance issues and reputational damage. Ensuring that as many suspicious transactions as possible are flagged can be seen as prioritized, even if this results in a higher number of false positives. Precision still matters, but incorrectly flagging legitimate activity typically leads to secondary review in current fraud detection method, rather than direct harm. In credit risk assessment, precision could be seen as more important, as it reflects how accurately the model identifies applicants who are likely to default. Approving high risk borrowers can lead to financial exposure, so minimizing false positives is desirable. Still, recall also plays a role in ensuring that truly risky applicants are not overlooked. This thesis reports precision, recall, and F1-score for all classification tasks to provide a balanced view of model performance.

The **Receiver Operating Characteristic (ROC)** curve is a commonly used method for assessing the performance of binary classification models. It plots the True Positive Rate against the False Positive Rate at various threshold levels, helping to visualize how changes in the threshold affect the balance between correctly identified positives and incorrectly identified negatives. The ROC curve highlights the trade off between sensitivity (True Positive Rate) and specificity (1 – False Positive Rate). A model that performs well will have a curve that bends sharply toward the top-left corner, indicating strong ability to correctly classify positive cases while minimizing false positives (Fawcett, 2006).

The ROC curve is particularly useful in settings with imbalanced datasets, such as fraud detection and credit risk assessment, where relying solely on single metrics can be misleading. By focusing how the model behaves across different decision boundaries, the ROC curve gives a more complete picture of classifier performance.

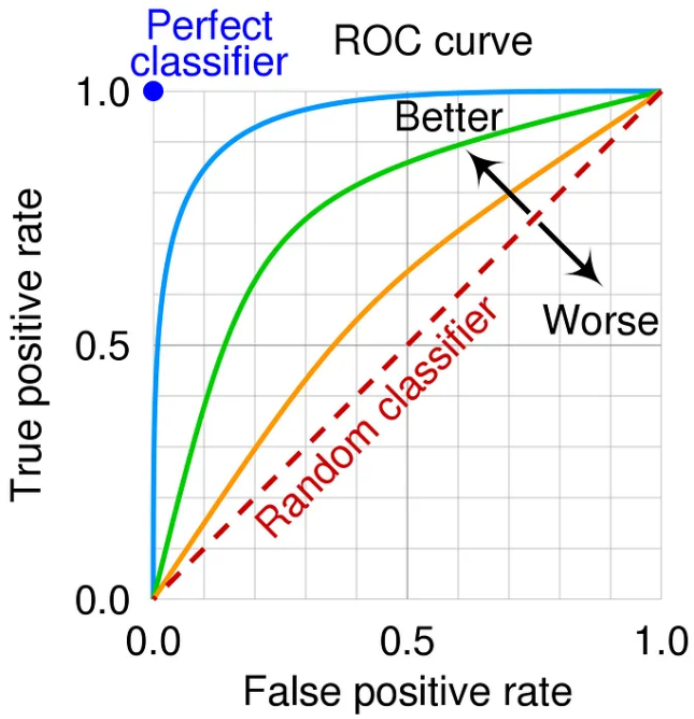

Figure 13: Example of ROC curve.

Source: (CMgleee, 2021)

To quantify the performance shown by the ROC curve, the **Area Under the Curve (AUC)** is calculated. The AUC represents the probability that the model will correctly rank a randomly chosen positive instance higher than a randomly chosen negative one. A value of 0.5 indicates that the model performs no better than random guessing, whereas a value of 1.0 reflects perfect classification performance (Fawcett, 2006).

The AUC is especially valuable for model comparison, as it summarizes the ROC curve into a single scalar value that is threshold independent. This makes it easier to assess overall model quality when the optimal threshold for classification is unknown or varies across applications.

SHAP for Interpretability

As stated, a well known drawback of introduced advanced ML models is their reduced interpretability. While these models can sometimes offer improved predictive accuracy compared to simpler alternatives, their internal decision making processes can difficult to communicate, making them less transparent and harder to justify for actual use in the banking sector.

To address this limitation, SHAP is employed in this study to enhance model interpretability. SHAP is a method that works with any ML algorithm and quantifies how much each feature contributes to a specific prediction, based on principles from cooperative game theory (Lundberg & Lee, 2017). By assigning each

feature a SHAP value, the method provides a consistent explanation of how input variables influence model outputs. These explanations can be aggregated across many predictions to assess overall feature importance or examined at the individual level to interpret specific outcomes.

Figure 14: Example of SHAP output

Source: (Lundberg, 2018)

SHAP offers two key strengths. First, it provides both local and global interpretability, allowing insight into individual predictions as well as the models overall behavior. Second, it is grounded in theoretical foundations, ensuring fairness and consistency in how contributions are assigned to features. But SHAP also comes with certain limitations. The computational cost can be high, especially for large datasets or complex models, as calculating exact Shapley values is resource intensive. In practice, SHAP relies on approximations that may sacrifice precision for efficiency. Additionally, while SHAP values can show which features are influential, they do not always reveal complex feature interactions or causal relationships.

Portfolio Optimization

Root Mean Squared Error (RMSE) is a commonly used evaluation metric for regression and time series forecasting tasks. It measures the average magnitude of prediction error, penalizing larger deviations more severely due to the squaring of residuals (Siami-Namini et al., 2018). RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{14}$$

- y_i denotes the actual observed value
- $\hat{y}_i$ is the predicted value
- n is the number of observations

A lower RMSE value can indicate better model performance, as it implies that predictions are, on average, closer to the true values. However, a notable limitation of RMSE is its sensitivity to outliers. Because the errors are squared before averaging, a few large errors can disproportionately influence the final value.

This can make RMSE misleading in datasets where extreme values occur frequently or are not necessarily important.

Evaluating Portfolios

When comparing traditional portfolio optimization methods with those incorporating AI, the evaluation is based on three key metrics: annualized return, annualized volatility, and the Sharpe ratio. The annualized return reflects the average yearly growth of the portfolio, offering a standardized measure of performance over time. Annualized volatility captures the typical yearly fluctuations in returns, indicating the level of risk or uncertainty associated with the portfolio. While return and volatility are informative on their own, the Sharpe ratio, as stated, provides a more comprehensive measure by expressing the excess return earned per unit of risk. All metrics are reported in annualized terms to ensure consistency and comparability across different investment horizons.

5. Empirical Results

This section presents the empirical results of the study, with a primary focus on evaluating the predictive accuracy of the developed models. The results are presented through figures and tables, illustrating how the models perform when applied to datasets that incorporate a combination of real world data and synthetically generated samples. These findings provide an evidence based foundation for assessing the second part of the problem statement: *To what extent do advanced models and methods provide practical improvements over traditional approaches in the areas of fraud detection, credit risk assessment, and portfolio optimization?* By highlighting where predictive performance is most evident, this section helps to identify the potential value of using more advanced AI methods in a financial context. However, the practical relevance, reliability, and limitations of implementing these approaches in real world banking environments are explored more thoroughly in the following discussion section.

5.1 Fraud Detection

Results from Fraud Dataset A

As previously described, Fraud Dataset A is based on real world transaction data that was transformed using PCA before being made publicly available. Due to the significant class imbalance where fraudulent cases represent only a small fraction of the dataset SMOTE was applied to the training data to balance the classes prior to model training. An example of hyperparameter tuning using RandomizedSearchCV applicable to all models used in both the fraud and credit risk experiments is provided in the Appendix E. The following presents the results based on this setup.

Fraud Dataset A	**Class 0 (Non-Fraud)**			**Class 1 (Fraud)**		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Logistic Regression	0.9998	0.9764	0.9880	0.0624	0.9054	0.1168
Decision Tree	0.9997	0.9980	0.9989	0.4232	0.8378	0.5624
Random Forest	0.9997	0.9998	0.9997	0.8601	0.8311	0.8454
XGBoost	0.9997	0.9997	0.9997	0.8435	0.8378	0.8407

Table 1: Classification Report for Fraud Dataset A

As seen in the table above, all models perform extremely well on the non fraud class, with precision and recall values close to 1. This is due to the strong class imbalance in the dataset, where non fraudulent transactions make up the vast majority of observations. This is why the differences between the models become much clearer and interesting when focusing on the fraud class. Logistic Regression achieves the highest recall with 90.54%, meaning it successfully identifies most fraud cases. However, its very low precision of 60.24%

indicates that the majority of its fraud predictions are incorrect, resulting in a large number of false positives which then results in a low F1-score of 11.68%.

The Decision Tree model improves this balance slightly, with a better precision of 42.32% and a higher F1-score of 56.24%, though its recall is slightly lower than that of Logistic Regression. The more advanced ensemble models, Random Forest and XGBoost, both show strong and balanced performance on the fraud class. Random Forest achieves the highest F1-score at 84.54%, with XGBoost performing similarly at 84.07%. Both models reach precision and recall values around 84%, suggesting that they are more reliable at identifying fraud without flagging too many false positives, compared to the models based on Logistic Regression and Decision Tree for this dataset.

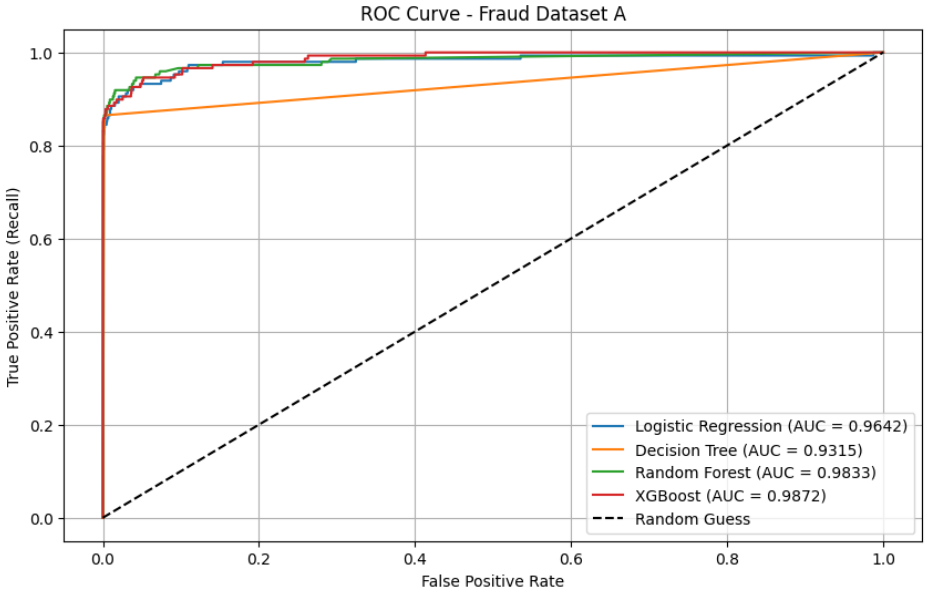

Figure 15: ROC Curve – Fraud Dataset A

Figure 15 presents the ROC curves for the evaluated models on Fraud Dataset A. All models achieve very high AUC values: XGBoost reaches 98.72%, Random Forest 98.33%, Logistic Regression 96.42%, and Decision Tree 93.15%. The ensemble methods XGBoost and Random Forest clearly achieve the highest scores, suggesting stronger ranking performance in distinguishing between fraudulent and non fraudulent transactions. These results are in line with previous tests using this dataset and are consistent with values reported in the literature as previously referenced. That said, the scores are optimistic compared to what would be expected under real world conditions. This will not be repeated for each result that follows, as the broader implications of this will be addressed in the discussion.

Fraud Dataset B

Fraud Dataset B is, as presented, based on a synthetic dataset constructed from one month of real financial transactions. The task remained a binary classification problem, with a significant imbalance between non fraudulent and fraudulent cases.

Fraud Dataset B	**Class 0 (Non-Fraud)**			**Class 1 (Fraud)**		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Logistic Regression	0.9999	0.9828	0.9913	0.0704	0.9070	0.1307
Decision Tree	0.9988	0.9897	0.9947	0.1066	0.8605	0.1897
Random Forest	0.9998	0.9989	0.9993	0.4658	0.7907	0.5862
XGBoost	0.9998	0.9989	0.9983	0.5294	0.8372	0.6486

Table 2: Classification Report for for Fraud Dataset B

Table 2 presents the evaluation metrics for each model. The results for the fraud class reveal some differences. The models using Logistic Regression The Decision Tree suffers from low F1-scores. The ensemble models once again perform best Random Forest reaches an F1-score of 58.62%, and XGBoost performs strongest overall with 64.86%. XGBoost also achieves the highest precision 52.94% and has a recall of 83.72% indicating what is a more effective trade off between correctly identifying fraud and limiting false positives.

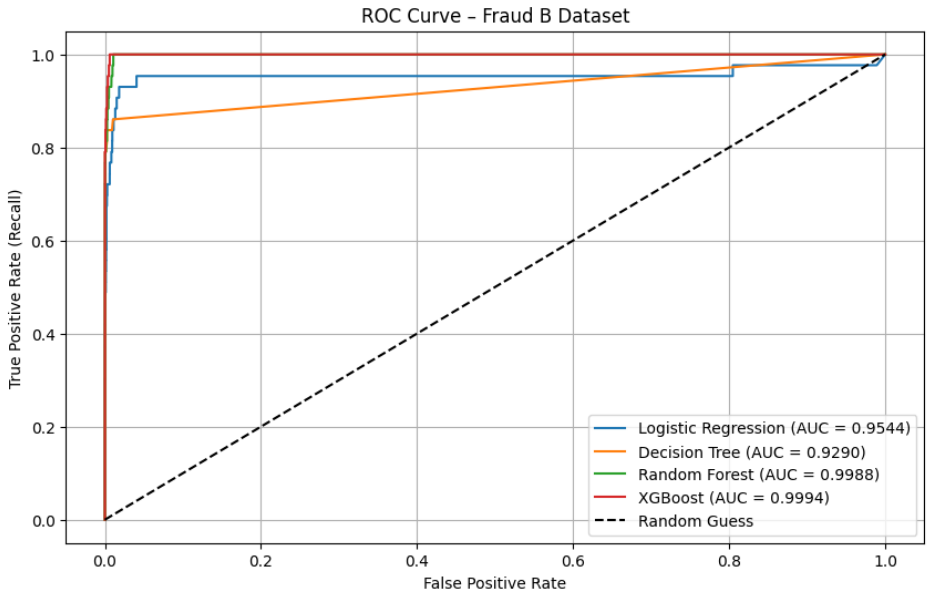

Figure 16: ROC Curve – Fraud Dataset B

Figure 16 shows the ROC curves for all models. XGBoost and Random Forest achieve the highest AUC scores, at 99.94% and 99.88% respectively, followed by Logistic Regression at 95.44% and Decision Tree

at 92.90%. The superior AUC scores obtained by the ensemble methods again suggest a strong ability to distinguish between fraudulent and non fraudulent transactions. The ensemble models appear to be particularly effective at ranking observations and assigning higher predicted probabilities to true fraud cases while assigning lower probabilities to legitimate transactions for this dataset.

5.2 Credit Risk Assessment

The evaluation of credit risk follows a structure similar to that used in the fraud detection section, as the task is likewise framed as a binary classification problem. The same evaluation metrics are applied to ensure comparability across models. Although credit risk assessment can be approached using multiclass formulations such as low, medium, high risk, the datasets used in this study are based on a binary distinction whether an applicant is considered high risk or not, or whether a loan is estimated to end in default.

This section distinguishes itself from the fraud detection section by expanding the empirical results to include unsupervised learning methods. Additionally, model interpretability is addressed using SHAP, which provides a view of feature contributions to individual predictions for the advanced ML algorithms used.

Credit Dataset A

Credit Dataset A addresses the binary classification problem of predicting whether a loan ends in default or not, based on a variety of borrower and loan related features. These include variables such as income level, employment length, loan amount, and loan intent. The aim was to evaluate the ability of the different models to correctly identify instances of default, which represent the minority class.

Credit Dataset A	**Class 0 (Non-Default)**			**Class 1 (Default)**		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Logistic Regression	0.8970	0.5877	0.7101	0.2460	0.6659	0.3593
Decision Tree	0.9630	0.9471	0.9549	0.7578	0.8197	0.7875
Random Forest	0.9379	0.9835	0.9602	0.8924	0.6779	0.7705
XGBoost	0.9795	0.9966	0.9880	0.9816	0.8966	0.9372

Table 3: Classification Report Credit Dataset A

As observed across earlier datasets, most models perform strongly on the majority class here, loans that did not default. However, Logistic Regression stands out with noticeably lower performance, achieving an F1-score of just 71.01% for the majority class, compared to over 95% for the other models. The differences become even clearer when evaluating the minority class. Logistic Regression shows the weakest performance, with a precision of 24.60% and recall of 66.59%, resulting in a low F1-score of 35.93%.

The Decision Tree model offers a stronger balance, improving both precision and recall to 75.78% and 81.97% respectively, and reaching an F1-score of 78.75%. Random Forest and XGBoost, demonstrate notably higher performance. Random Forest achieves a precision of 89.24% and recall of 67.79%, yielding an F1-score of 77.05%. XGBoost again shows the strongest classification performance on the default class, with a precision of 98.16%, recall of 89.66%, and an F1-score of 93.72%.

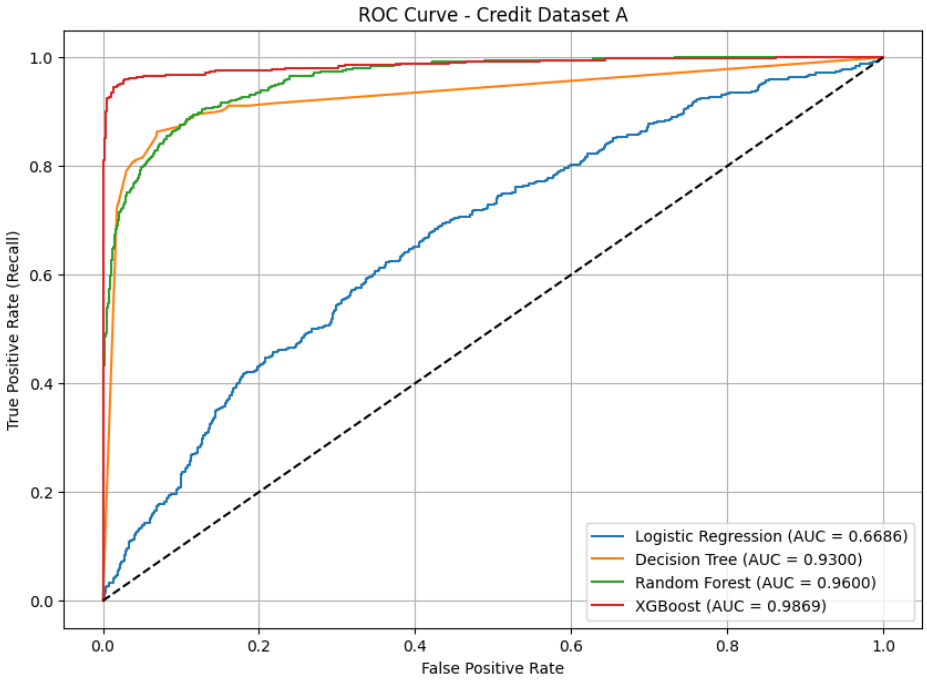

Figure 17: ROC Curve – Credit Dataset A

From the ROC Curve in Figure 17 it shows that the ensemble methods again show strong ranking performance, with XGBoost reaching the highest AUC at 98.69%, followed by Random Forest at 96.00%. The Decision Tree model also performs well with an AUC of 93.00%. Logistic Regression lags behind, with a lower AUC of 66.86%, indicating limited ability to distinguish between default and non default cases for the dataset.

Credit Dataset B

This dataset evaluates whether a customer is classified as high credit risk or not, based on a combination of behavioral and demographic information. The data includes variables describing payment history such as the number of overdue payments, the total number of overdue days, the number of ontime payment, credit limits and account balances. In addition to payment history, the dataset incorporates a set of encoded categorical variables representing customer demographic characteristics. The target variable is binary, indicating whether a customer is deemed high or normal risk in terms of giving out a loan, credit etc..

Credit Dataset B	**Class 0 (Normal Risk)**			**Class 1 (High Risk)**		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Logistic Regression	0.8667	0.9489	0.9060	0.7187	0.4723	0.5700
Decision Tree	0.9230	0.9869	0.9539	0.9369	0.7023	0.8028
Random Forest	0.9247	0.9929	0.9576	0.9648	0.7707	0.8165
XGBoost	0.9304	0.9912	0.9599	0.9585	0.7319	0.8300

Table 4: Classification Report Credit Dataset B

The performance patterns across models largely shows similarities to previous presented results. Logistic Regression achieves high precision on high-risk predictions 71.87% but fails to recall more than half of actual high-risk cases 47.23%, leading to the lowest F1-score of 57.00%. The Decision Tree model performs more consistently, increasing both precision and recall 93.69% and 70.23%, respectively, which results in a stronger F1-score of 80.28%.

On the last dataset the ensemble models continue to perform better than the deemed traditional models. Random Forest achieves strong scores across both classes, including a recall of 77.07% and an F1-score of 81.65% for high risk customers. XGBoost provides the most balanced outcome, reaching 95.85% precision, 73.19% recall, and an F1-score of 83.00%, suggesting the most stable prediction performance for the minority class.

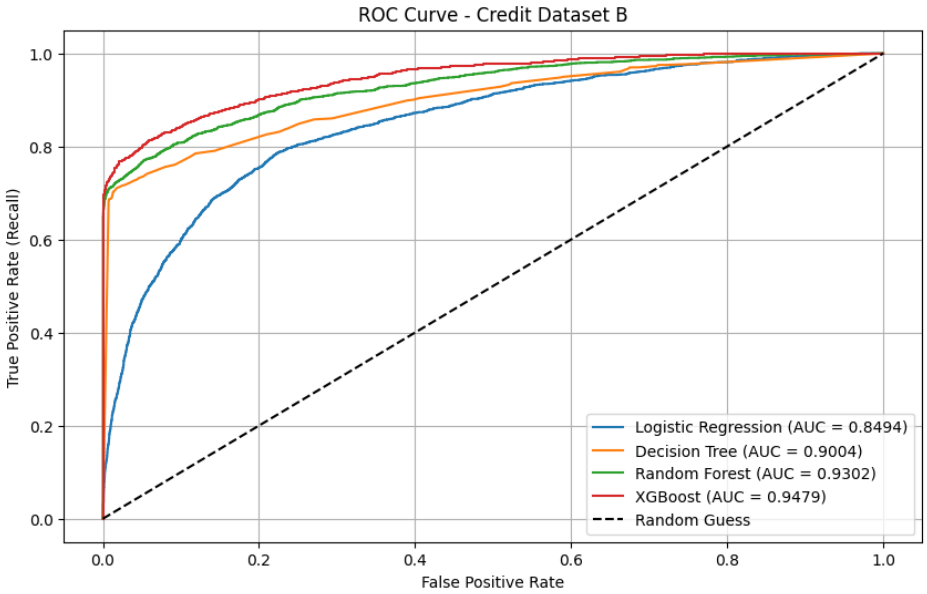

Figure 18: ROC Curve – Credit Dataset B

This pattern of stronger performance by the ensemble methods is also evident in the last plot of the ROC curve. XGBoost achieves the highest AUC at 94.79%, followed closely by Random Forest at 93.02%. Both

models outperform the simpler approaches, with Decision Tree reaching an AUC of 90.04% and Logistic Regression trailing behind at 84.94%. These results reinforce the trend observed across datasets, where ensemble models consistently show superior ability to distinguish between high risk and more normal customers.

5.2.1 Inclusion of Unsupervised ML

To further explore underlying risk structures in the data and support the classification task, Credit Dataset B was reevaluated using the two unsupervised learning techniques previously presented K-means clustering and Isolation Forest. These methods are applied to try an uncover latent borrower profiles and detect anomaly patterns that may not be captured by supervised models alone. Both K-means and Isolation Forest can be used as standalone exploratory tools to better understand the structure and distribution of risk and as sources of additional features that enrich the supervised ML models. This approach is supported by multiple studies, including Faisal et al. (2024), who, as previously noted, highlight the effectiveness of hybrid learning strategies in banking applications. They argue that combining supervised and unsupervised methods can improve model performance by enabling models to capture both established patterns and previously unseen anomalies.

K-Means Clustering

To determine a suitable number of clusters for the borrower population, the elbow method was applied. As illustrated in Figure 19, the inertia representing the cluster sum of squares decreases a little bit up to k=3 Based on this inflection point, three clusters were selected as a practical and interpretable solution for segmenting the data.

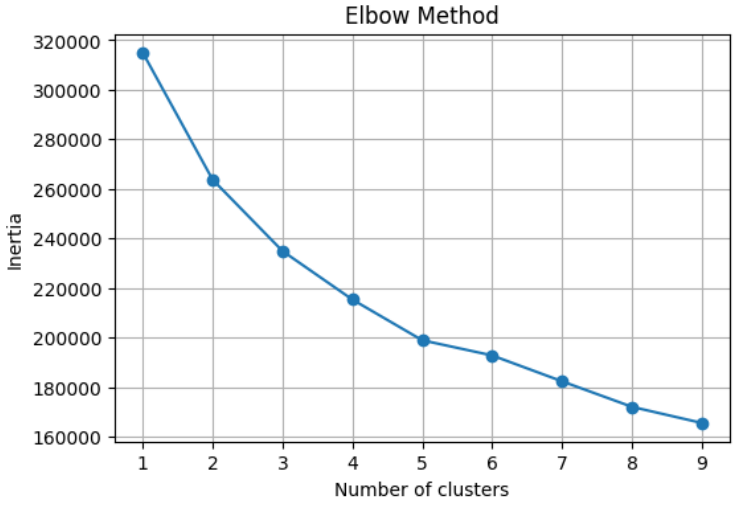

Figure 19: Elbow Method – Optimal Number of Clusters

To visualize the results, PCA was applied to reduce the dimensionality of the feature space while preserving as much variance as possible. This allows the clustered structure to be projected into two dimensions

which makes for easier interpretation. Figure 20 displays the resulting clusters using the PCA transformed components, where each point represents a borrower and each color indicates a cluster assignment. While PCA aids with visualization, it reduces the datas complexity and can obscure important variance. The plot shows distinct groupings that may reflect underlying borrower profiles.

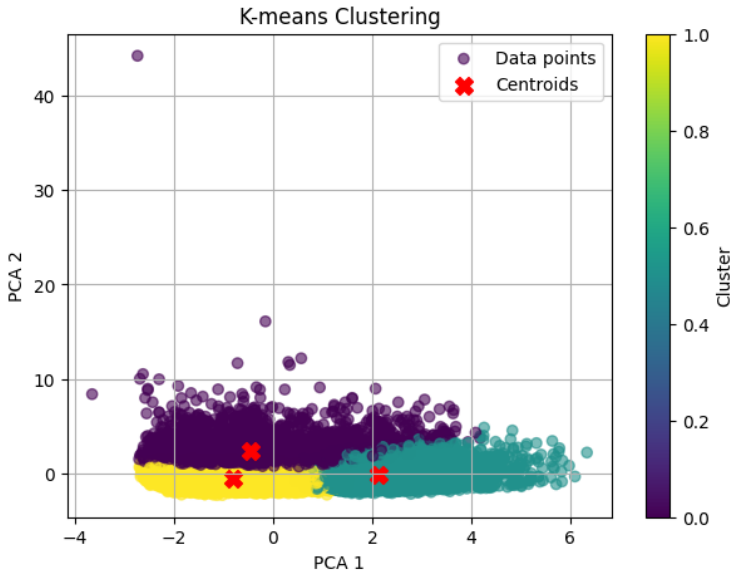

Figure 20: K-means Clustering of Borrowers

In addition to serving as a visual tool, the cluster assignments generated by K-means were also incorporated into the supervised models as a new categorical feature. This hybrid approach was done to introduce structural insights from unsupervised learning, which could help capture patterns not fully reflected in the original variables.

Isolation Forest

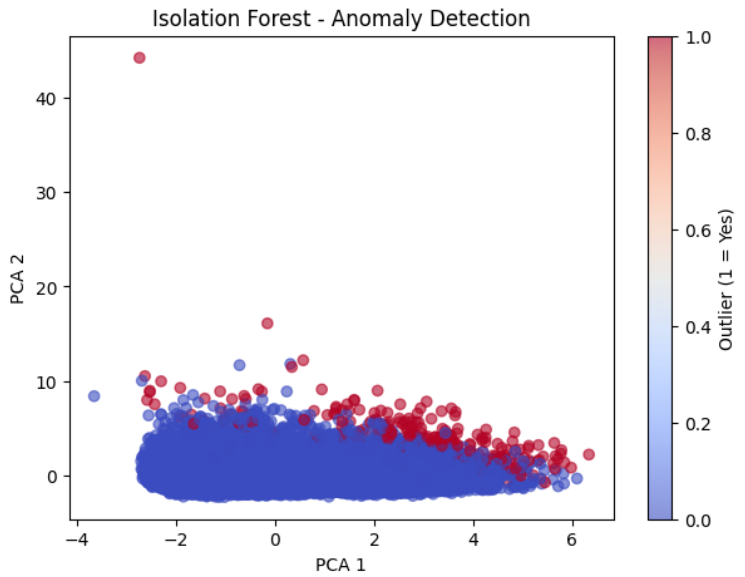

Figure 21: Detected anomalies using Isolation Forest with a 1% contamination threshold

Isolation Forest was applied to identify anomaly observations that deviate from the general structure of the credit data. A contamination rate of 0.01 was used to reflect the assumption that only a very small proportion of the observations of approximately 1% are actual anomalies. This choice was made to reflect the nature of credit datasets, where defaults and high risk borrower behavior are relatively rare. The visualization in Figure 21, based also on PCA for dimensionality reduction, highlights these detected outliers in red. While the plot helps reveal dispersion and isolation in a simplified 2D form, it is important to note that PCA may not fully capture the complexity of the original feature space. As with K-means, the outlier flags from Isolation Forest were also added as a new binary feature to the supervised models.

Evaluation Metrics Incorporating Unsupervised Machine Learning Features

Credit Dataset B with UML	**Class 0 (Non-Default)**			**Class 1 (Default)**		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Logistic Regression	0.8684	0.9483	0.9066	0.7198	0.4804	0.5762
Decision Tree	0.9333	0.9162	0.9547	0.7157	0.7630	0.7386
Random Forest	0.9245	0.9903	0.9563	0.9530	0.7707	0.8122
XGBoost	0.9316	0.9909	0.9603	0.9574	0.7367	0.8327

Table 5: Classification Report Credit Dataset B with UML

The inclusion of features derived from unsupervised learning techniques specifically, cluster labels does not result in a improvement in the evaluation metrics for the classification models. As shown in Table 5 and Figure 22, the precision, recall, F1-scores and AUC values remain broadly comparable to those presented before the integration of these additional features. In some cases, minor declines can be observed, while in others the values are maintained at similar levels.

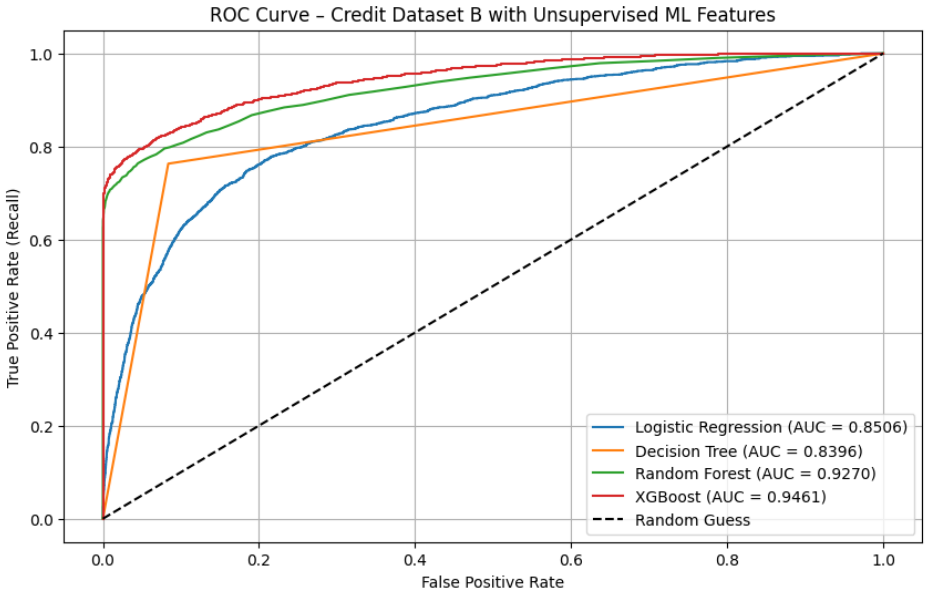

Figure 22: ROC Curve - Credit Dataset B with UML Features

These results could suggest that, given the specific dataset and experimental setup used in this analysis, the integration of unsupervised learning features did not substantially change the predictive performance of the models. However, this does not diminish the relevance of hybrid approaches, which may still offer valuable insights in different contexts or with further refinement. The outcome may reflect the characteristics of the data or the alignment between the unsupervised patterns and the supervised task. Notably, the inclusion of these features did not negatively impact model stability, and they can still contribute to a broader understanding of model behavior particularly when combined with the interpretability tool SHAP.

SHAP and the Interpretability of Ensemble Models

To conclude the evaluation of fraud detection and credit risk classification models, SHAP was used to interpret the XGBoost model trained on Credit Dataset B. As shown in Figure **??**, several features stand out in terms of their impact on the models predictions. `person_income`, `loan_grade`, and `loan_percent_income` are among the most influential variables, with high income generally pushing predictions toward a lower default risk. Similarly, high loan burden or lower loan grade appears to increase predicted risk which are consistent with general expectations.

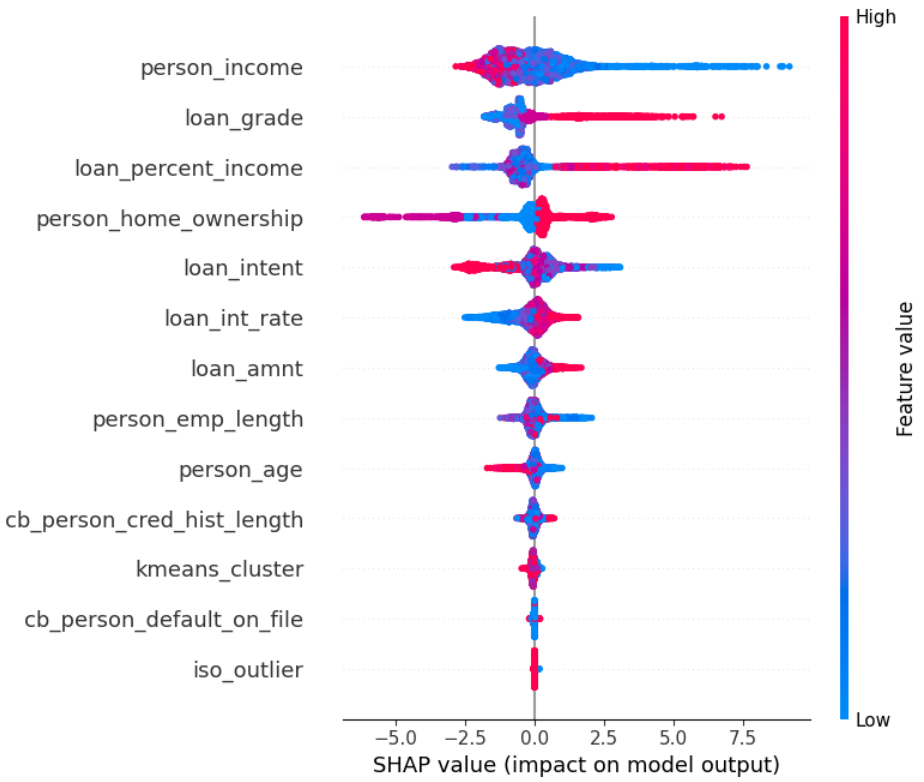

Figure 23: Plotted SHAP Values

The SHAP plot also allows for the evaluation of the two features derived from unsupervised learning kmeans_cluster and iso_outlier. Compared to the other inputs, their contribution to the model is modest. While both are visible in the plot, they have a small average impact on prediction outcomes. This suggests that while clustering and anomaly detection may add structure or flag unusual behavior in the dataset, their utility in improving the XGBoost models predictive performance appears limited in this specific setup which is also why the inclusion of the features did not affect the results.

5.3 Portfolio Optimization

5.3.1 Time Series Forecasting

The following section presents results from the first of three empirical tests conducted for portfolio optimzation. The purpose of this test is to compare the forecasting accuracy of traditional time series models and advanced deep learning methods when predicting financial returns a key input in portfolio construction. This design follows the structure introduced by Siami-Namini et al. (2018) in their comparison of ARIMA and LSTM models and is extended here by including GRU as a third model.

The analysis is based on weekly adjusted closing price data for the period from 2015-01-01 to 2025-01-01. Five assets were selected the Dow Jones Industrial Average (DJI) and four randomly chosen stocks from

the index Apple (AAPL), Johnson & Johnson (JNJ), JPMorgan Chase (JPM), and Boeing (BA). Forecasts were generated using both weekly and monthly return series derived from these prices.

In the original study by Siami-Namini et al. (2018), a fixed ARIMA (5,1,0) specification was applied uniformly across all time series. To maintain comparability with that setup, the same model structure was adopted in this thesis. While this allows for straightforward benchmarking across methods, it also represents a methodological shortcoming, as each time series may in fact require a different (p, d, q) configuration to achieve optimal forecasting accuracy.

To try and address this, the pmdarima Python library was used to test whether data driven model selection could improve accuracy. This library performs stepwise model selection by minimizing information criteria, and is commonly used to automatically identify the best ARIMA order for each time series. However, no notable change in RMSE was observed when switching from the fixed ARIMA(5,1,0) to the individually optimized models. For that reason, and to maintain comparability across series and methods, the fixed specification was retained throughout the main analysis. No hyperparameter tuning was applied to the LSTM and GRU models default configurations were used and are shown in Appendix F.

Each model was trained using a 70/30 split, with 70% of the data allocated for training and the remaining 30% reserved for out-of-sample evaluation. To ensure the validity of the results, the previously discussed model assumptions such as stationarity and residual independence were tested for before model estimation.

Ticker	Model	RMSE	Reduction vs ARIMA (%)
AAPL	ARIMA	0.040761	–
	LSTM	0.035129	13.82
	GRU	0.035043	14.03
JNJ	ARIMA	0.023333	–
	LSTM	0.022062	5.45
	GRU	0.022414	3.94
JPM	ARIMA	0.039542	–
	LSTM	0.034189	13.54
	GRU	0.035304	10.72
BA	ARIMA	0.054166	–
	LSTM	0.050034	7.63
	GRU	0.048046	11.30
DJI	ARIMA	0.023914	–
	LSTM	0.019862	16.94
	GRU	0.020478	14.37
Average	ARIMA	**0.036343**	–
	LSTM	**0.032255**	**11.88**
	GRU	**0.032657**	**10.87**

Table 6: Weekly RMSE and percentage

Table 6 presents the RMSE and relative improvements achieved by LSTM and GRU over ARIMA on weekly return data. Across all five assets, both LSTM and GRU outperformed ARIMA. On average, LSTM achieved an 11.88% reduction in RMSE, while GRU delivered a slightly lower average improvement of 10.87%. The

strongest relative performance occurred for the DJI and AAPL series, where RMSE reductions exceeded 14% in both neural models. These improvements suggest that deep learning models may better capture nonlinear patterns and dependencies that are not well handled by linear ARIMA processes.

Ticker	Model	RMSE	Reduction vs ARIMA (%)
AAPL	ARIMA	0.086123	–
	LSTM	0.068081	20.95
	GRU	0.064027	25.66
JNJ	ARIMA	0.053834	–
	LSTM	0.043397	19.39
	GRU	0.053668	0.31
JPM	ARIMA	0.082389	–
	LSTM	0.075977	7.78
	GRU	0.073738	10.50
BA	ARIMA	0.122724	–
	LSTM	0.125526	-2.28
	GRU	0.124648	-1.57
DJI	ARIMA	0.052229	–
	LSTM	0.043408	16.89
	GRU	0.043468	16.77
Average	ARIMA	**0.079860**	–
	LSTM	**0.071278**	**12.35**
	GRU	**0.071510**	**10.73**

Table 7: Monthly RMSE and percentage reduction

Table 7 summarizes the results for monthly returns. Compared to the weekly setup, the monthly data yielded slightly larger average reductions in RMSE 12.35% for LSTM and 10.73% for GRU. As in the weekly test, the best results were achieved for AAPL and DJI, with GRU reducing RMSE for AAPL by over 25%.

Not all series showed consistent improvements. For example, in the case of BA, both GRU and LSTM slightly underperformed compared to ARIMA. The performance of a model can depend on specific characteristics of the time series, such as noise levels, autocorrelation patterns and structural volatility.

While the goal of these results is not to establish definitive model superiority, the patterns observed provide some initial insights. Inspired by the comparative setup in Siami-Namini et al. (2018), where substantial performance gaps were showcased between ARIMA and deep learning models, this test yielded more moderate, but still noticeable differences. Even when using relatively simple architectures and default hyperparameters, both LSTM and GRU models generally produced more accurate forecasts than ARIMA on the specific time series datasets used in this study. In practice, performance can be very much influenced by factors such as feature engineering, big hyperparameter tuning, model architecture, and training strategies. There remains considerable potential for refinement in both ARIMA and more advanced methods to further improve forecasting outcomes, although this comes with an increased risk of overfitting.

5.3.2 Forecast-Driven Allocation

This section presents traditional and AI driven approaches to portfolio allocation, with a specific emphasis on the methods used to forecast asset returns. The benchmark model is classical mean-variance optimization, in which expected returns are estimated using historical averages and portfolio weights are chosen to maximize the Sharpe ratio. The risk-free rate is taken from the 3-Month U.S. Treasury Bill yield, converted to a weekly frequency to align with the return data.

As a first step in the evaluation, a traditional allocation strategy based on historical mean returns was implemented. Weekly return data from 2020 to the end of 2023 was used to estimate expected returns and covariances, which were then input into the classical mean-variance optimization framework. The shorter span was chosen to ensure compatibility with weekly Google Trends data, as retrieving trend data beyond five years at a weekly frequency is limited by platform constraints. Portfolio performance was evaluated over the full year of 2024, without any reallocation during the test period. This served as the performance baseline against which all forecasting driven strategies are compared.

Given the performance of both LSTM and GRU models in the previous section, GRU was selected for the forecasting based simulations due to its efficiency and ability to deliver results comparable to that of LSTM. In this design, the only change relative to the classical benchmark was the method used to estimate expected returns, where rather than relying on historical averages, weekly returns were forecasted using GRU. These forecasts were then integrated into the same mean-variance optimization framework.

A total of 20 stocks were randomly selected from a pool of large-, mid-, and small-cap companies the full list is provided in the Appendix G. To avoid basing results on a single combination of stocks, 100 simulations were run. In each simulation, a random subset of 6 to 10 stocks was available and used to construct a portfolio. The same random seed of 10 was used across all methods to ensure that each strategy was evaluated on identical stock selections. This was done to provides a more fair comparison and a more reliable picture of how each strategy performs across a range of scenarios and to not over interpret any single outcome.

Date	ADBE	ADBE_trend	FedFundsRate	CPI	Inflation	Unemployment
2020-05-04	0.06840	1.555741	-1.06715	-1.724338	-0.844265	4.08469
2020-05-11	-0.006014	1.458107	-1.06715	-1.724338	-0.419171	4.08469
2020-05-18	0.054640	1.262839	-1.06715	-1.724338	-0.419171	4.08469
2020-05-25	0.003478	0.579401	-1.06715	-1.724338	-0.419171	4.08469
2020-06-01	0.016296	0.286500	-1.06715	-1.724338	-0.419171	4.08469

Table 8: Sample of weekly data for ADBE, Google Trend and selected macroeconomic indicators

The analysis was extended by introducing the previously used ensemble based ML algorithms, XGBoost and Random Forest, to forecast future returns. Both models were trained using weekly returns alongside

additional input features derived from macroeconomic indicators and Google Trends data, as described earlier. These alternative data sources were included to examine whether they could enhance the accuracy of return predictions and ultimately improve portfolio performance when applied within the same Sharpe ratio maximization framework.

Method	Return	Volatility	Sharpe Ratio
Historical Returns	0.4279	0.3414	1.3252
GRU Forecasting	0.2691	0.2415	1.0280

Table 9: Median 2024 Portfolio Metrics Based Only on Returns

Table 9 presents the median portfolio performance metrics for 2024 using traditional historical mean returns and GRU based forecasts as inputs to the mean-variance optimization framework. While GRU offers a more dynamic, model driven approach to forecasting returns, the results indicate that it did not lead to better portfolio outcomes in this setting. The traditional strategy based on historical average returns yielded a higher Sharpe ratio, indicating superior risk adjusted performance within the context of this specific dataset and evaluation period.

The data including alterative data using XGBoost and Random Forest did not achieve stronger risk adjusted performance. As shown in Table 10, both methods resulted in Sharpe ratios below 0.83, which is lower than in earlier tests that used price data alone. There are a few likely reasons for this. The added variables may not have been strongly linked to short term price movements, or the timing between signals and returns may not have aligned well.

Method	Return	Volatility	Sharpe Ratio
XGBoost Forecasting	0.2148	0.2622	0.8243
Random Forest Forecasting	0.2132	0.2223	0.8182

Table 10: Median 2024 Portfolio Metrics from Forecasting-Based Strategies

If all strategies are compared to the actual S&P 500 index, none outperform it from a Sharpe maximization perspective. From Table 11, it can be seen, that S&P 500 delivered a Sharpe ratio of 1.68 in 2024, significantly exceeding the performance of all model based portfolios.

Method	Return	Volatility	Sharpe Ratio
Actual S&P 500	0.2577	0.1267	1.6781

Table 11: S&P 500 Performance Metrics in 2024

This outcome can also be attributed to the strong market performance in 2024, characterized by high returns and relatively low volatility, which contributed to the S&P 500s superior risk-adjusted results. While not

conclusive, the findings underscore the challenges of consistently outperforming the market using active forecasting strategies, even when supported by advanced ML models and enriched with alternative data sources. These results should not be seen as a dismissal of predictive modeling rather they illustrate how sensitive forecasting performance is to factors such as timing, data quality, and broader market conditions. Moreover, the findings align with the EMH, which suggests that financial markets tend to incorporate available information efficiently making systematic outperformance through active strategies difficult to achieve.

5.3.3 Portfolio Optimization with Generative AI

This section presents the final set of empirical results, focusing on the performance of portfolios constructed using generative AI under anonymized and unbiased conditions. As outlined in the methodology, the experiment extends the simulation design proposed by Perlin et al. (2025), who examined the effectiveness of Google Gemini model in selecting stocks from anonymized financial data. The technical setup in this study differs by using ChatGPT in an interactive setting, but the core structure remains aligned anonymized stock data, randomized evaluation periods, and benchmark comparisons against a naive portfolio and the S&P 500 index.

The results based on 1,500 independent simulations for each investment horizon. In every simulation, a random set of five anonymized stocks is selected from a dataset of 503 S&P 500 stocks, with historical weekly data spanning from 2015 to 2024. Identifiers and dates are excluded, and stock price series were rescaled by a random factor to preserve anonymity. The goal was to examine whether LLMs like ChatGPT o3 can formulate a viable investment rule based solely on anonymized past prices.

Horizon	**Beta**		**Ann. Return**		**Ann. Sharpe**	
	ChatGPT	Naive	ChatGPT	Naive	ChatGPT	Naive
1 Month	0.90	1.05	10.4%	15.4%	0.78	1.00
6 Months	0.76	0.83	11.3%	14.3%	0.65	0.79
12 Months	0.73	0.81	9.0%	13.7%	0.47	0.66
36 Months	0.81	0.82	8.1%	10.8%	0.40	0.54

Table 12: Performance of ChatGPTs Strategy across 1 500 simulations

Source: Table structure adapted from Perlin et al. (2025)

Tables 12 summarize the median outcomes from the simulations. These include beta, annualized return, and annualized Sharpe ratio across the four horizons 1 month, 6 months, 12 months, and 36 months. Table 13 shows the results for the S&P 500 index under the same conditions.

As observed in Table 12, ChatGPT portfolios display a gradual decline in beta, return, and Sharpe ratio as the horizon increases. At the shortest horizon, the AI based strategy yields a median return of 10.4% and a Sharpe ratio of 0.78. At the 36-month horizon, these figures decrease to 8.1% and 0.40, respectively. The naive strategy produces marginally higher median returns and Sharpe ratios at most horizons. The S&P 500 outperforms both strategies in terms of risk adjusted returns in the first three time periods.

Horizon	**Ann. Return**	**Ann. Sharpe**
1 Month	24.5%	1.89
6 Months	19.9%	1.25
12 Months	17.6%	0.98
36 Months	8.4%	0.40

Table 13: Median performance of the S&P 500 across the same 1 500 simulations

Source: Table structure adapted from Perlin et al. (2025)

Table 14 shows how frequently ChatGPTs portfolios outperformed the benchmarks in each simulation. These proportions refer to the number of times the AI-generated portfolios achieved higher returns or Sharpe ratios than the naive portfolio or the S&P 500.

Across all horizons, ChatGPT's strategy was outperformed by the S&P 500 in terms of total return in the majority of simulations. Return outperformance rates ranged from 48.6% at the 1-month horizon to 46.2% at 36 months indicating that ChatGPT lagged the index more often than not. A similar pattern is observed when comparing against the naive benchmark where while the AI portfolios came close to same performance at the 1-month horizon, it fell further behind at longer horizons, dropping to 35.3% at 36 months. In terms of Sharpe ratio, ChatGPT's underperformance was even more pronounced. It achieved a higher Sharpe than the S&P 500 in only 35.6–45.5% of simulations, depending on the horizon. Against the naive strategy, the trend is clearly declining from 44.5% at 1 month to just 31.9% at 36 months showing that the AI strategy struggled to maintain competitive risk- djusted returns over time.

Months	% Returns > S&P500	% Sharpe > S&P500	% Returns > Naive	% Sharpe > Naive
1	48.6%	42.8%	49.6%	44.5%
6	46.1%	39.1%	46.5%	41.3%
12	43.6%	35.6%	39.6%	34.7%
36	46.2%	45.5%	35.3%	31.9%

Table 14: Proportion of simulations in which ChatGPTs Strategy outperformed benchmarks

Source: Table structure adapted from Perlin et al. (2025)

The overall pattern closely mirrors the findings reported by Perlin et al. (2025). In this study, the portfolios constructed by ChatGPT tended to underperform both the S&P 500 index and the naive benchmark across all investment horizons. The AI generated strategy did not show a consistent edge in terms of Sharpe ratio, and its relative performance appeared to weaken over longer time frames. These outcomes suggest that anonymized setups may limit the models ability to detect patterns that consistently deliver superior returns. While ChatGPT can generate plausible investment strategies from historical prices, the lack of contextual and structural information may restrict its effectiveness. These findings point to the broader limitations of using LLMs in isolation for portfolio construction under strict anonymity constraints where potential for bias is very low.

6. Discussion

Traditional vs Advanced AI

The central part of the problem statement addressed in this thesis is whether advanced AI models offer meaningful improvements over traditional statistical methods in core banking tasks. This comparison is complex, as model performance is influenced by factors such as data structure, feature selection, and the specifics of each use case. Statistical significance testing was not applied, as the aim is not to generalize beyond the specific datasets and model configurations used. Instead, models are evaluated under similar conditions to allow for practical, scenario specific comparison. In this context, relying on significance tests could oversimplify the evaluation by reducing it to a single metric, failing to account for the variability and sensitivity inherent in real world financial data. Moreover, achieving statistical significance in a single dataset does not prove that advanced AI methods are superior overall, especially given the dynamic and institution specific nature of banking applications.

However, empirically, the results from both this thesis and the literature review suggest that AI powered methods may offer advantages over traditional approaches in certain banking tasks particularly classification problems such as fraud detection and credit risk assessment. In these areas, ensemble models like Random Forest and XGBoost consistently outperformed logistic regression in terms of both F1-score and AUC. These findings align with existing research reviewed in this thesis.

While logistic regression continues to serve as a benchmark in practice, favored for its interpretability, ease of implementation, and compliance with regulatory standards, its linear structure limits its ability to fully model the complexities of real world banking datasets (Dumitrescu et al., 2022). In this study, logistic regression proved competitive in some settings. Still, more flexible models, particularly XGBoost, demonstrated superior performance across the evaluated datasets, consistent with previous studies advocating for ensemble methods in improving classification accuracy.

In contrast, the portfolio optimization component yielded more varied and inconclusive results. Forecasting models based on deep learning architectures, such as LSTM and GRU, achieved moderately lower prediction errors than traditional ARIMA models, but these improvements did not consistently translate into better portfolio outcomes under Sharpe ratio maximization. Compared to classification tasks, return forecasting was more sensitive to input design, data frequency, and the size of the training window. As Gu et al. (2020) argue, ML can uncover return dynamics that traditional models may overlook. Still, this study finds that such predictive gains do not necessarily lead to enhanced asset allocation performance. In several simulations, portfolios based on AI driven forecasts failed to outperform naive or passive investment strategies. This limitation reflects both modeling challenges and the constraints of available data. It also resonates with

the EMH, which suggests that asset prices already incorporate all publicly available information, making consistent outperformance difficult to achieve (Fama, 1970). Moreover, as AI becomes more widely adopted in the financial sector, any edge it provides may diminish, underscoring that the value of AI lies not merely in its use, but in how it is integrated. Despite the rise of new technologies, the foundational principles of portfolio theory particularly diversification and the risk return trade off introduced by Markowitz remain central to asset management practice (Martin, 2021)

The limitations identified in this study likely stem from both the intrinsic difficulty of financial forecasting and the methodological choices made such as restricted feature sets, data constraints, and potential overfitting. Furthermore, the use of LLMs, while potentially helpful in supporting technical and fundamental analysis or automating aspects of the research process, did not result in performance improvements with mean-variance optimization when used on its own in this study.

While the observed gains were uneven, the findings support the view that AI holds potential to add value in banking applications. However, significant challenges remain particularly regarding how these models are interpreted, trusted, and integrated into real world decision making processes. Although model accuracy may be further improved through technical refinement, practical deployment involves more than predictive performance alone. Issues such as explainability, regulatory compliance, and user confidence play a critical role in determining the practical application utility of AI systems in the banking sector.

Interpretability and Transparency

Although the advanced AI models used in this thesis demonstrated strong predictive performance in classification tasks, their limited interpretability remains a major barrier to adoption in banking. Traditional models continue to be favored due to their transparency and compatibility with regulatory frameworks, as their outputs can be directly linked to input features. In contrast, the advanced models such as those using XGBoost and deep neural networks often operate as the previously stated black boxes, offering little insight into the decision-making process, an big issue of particular concern in domains like credit scoring and fraud detection. This lack of transparency is a key reason why, despite encouraging results, AI has not been broadly implemented in practice (Misheva et al., 2021).

To mitigate this, SHAP was applied to the AI models in this thesis to provide feature level explanations. The use of SHAP and its rankings aligned with what was expected, helping to identify the most influential variables in determining customer risk profiles. While this contributed to greater transparency, the use of SHAP does not fully resolve the underlying explainability problem. A study by Huang & Marques-Silva (2023) shows that SHAP can misrepresent feature importance, particularly in high dimensional or highly interactive datasets common in banking. It may assign high relevance to features that are irrelevant or

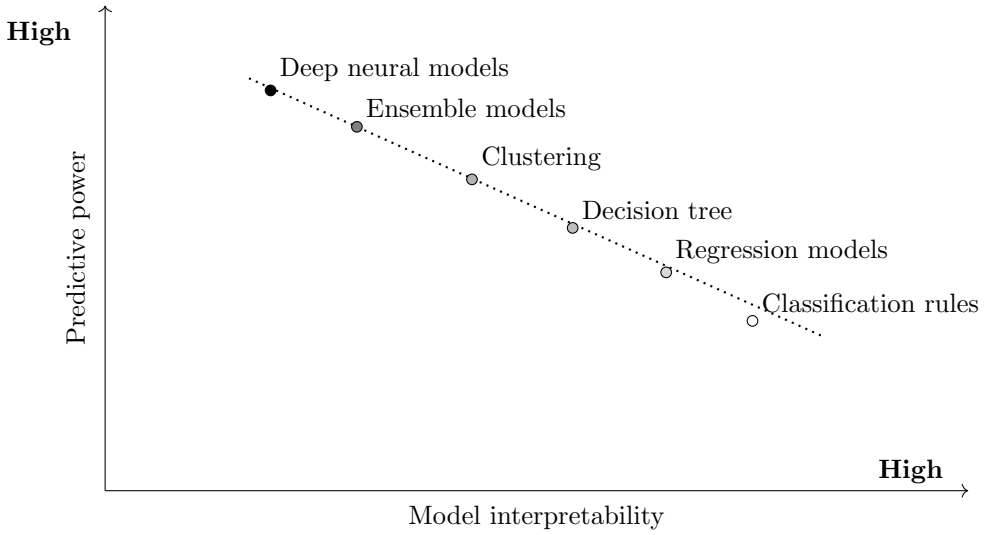

Figure 24: Trade-off between Predictive power vs. Interpretability

Source: Figure adapted from (Kumar et al., 2021)

miss key predictors, leading to explanations that appear plausible but are not theoretically grounded. This reflects a broader trade off between performance and interpretability. Even with tools like SHAP, complex models often fall short of the level of transparency required by financial institutions and regulators As a result, high performing AI models still face resistance not only because they are hard to interpret, but also because of practical challenges with the data they rely on (Huang & Marques-Silva, 2023).

Data Challenges

Many of the challenges faced in this thesis were not just about model choice, but also about the data used. In banking, the quality and structure of data are often just as important as the algorithms themselves. One key issue in fraud detection and credit risk assessment was the imbalance between classes. Fraud cases or defaults were rare compared to the total dataset, making it harder for models to detect them accurately. To address this, SMOTE was used to generate synthetic examples, which helped improve performance on the minority class. However, synthetic data can also lead to overfitting and does not fully replace the value of real world examples.

Another challenge was the lack of access to detailed and high frequency real world banking data. While the datasets used were appropriate for academic testing, being commonly used in the reviewed literature, actual banking data containing transaction histories, behavioral indicators etc. are not publicly available due to confidentiality and regulatory constraints. This limitation reduced the depth and simplified the the analysis somewhat. As a result, the models tested in this thesis may have underperformed relative to what could be achieved with institutional data. The findings therefore provide a useful indication of AI's potential but do

not fully reflect its capacity in real world banking environments.

In the portfolio optimization section, similar data limitations were present. While historical price data is widely available and straightforward to use, access to good and meaningful alternative data sources is far more limited. Such data can be commercially restricted, difficult to collect in a structured format, or require web scraping, which may raise legal and ethical concerns. As a result, the forecasting models in this thesis were trained on a relatively narrow set of inputs and a limited time window, which ultimately constrained their ability to reflect the full complexity of real financial markets. This also limited the opportunity to explore the full potential of more advanced models that are capable of leveraging very large and high dimensional datasets.

The data related constraints also affected the performance of unsupervised learning methods explored in this thesis. Without access to richer feature sets or behavioral patterns, the clustering and anomaly detection results were underwhelming. While unsupervised techniques are often proposed as a promising alternative when labels are scarce, their effectiveness relies heavily on high quality, multi-dimensional data. As highlighted by Bello et al. (2023), recent literature suggests that combining supervised and unsupervised methods in hybrid frameworks can yield better performance and robustness particularly in domains like the areas in banking, where both structured labels and hidden patterns can coexist. Such hybrid approaches could be a valuable approach for future research and practical deployment.

AI Applications in Practice

A key part of addressing the problem statement was not only to compare model performance, but also to examine how AI has been adopted within the banking sector. Understanding the current pace and scope of adoption helps contextualize the technical findings, showing not just what AI can do in theory, but how it is being used in practice. Many financial institutions remain in early experimental phases, exploring use cases without moving toward full scale implementation (Marous, 2024). This gap between technical potential and actual adoption comes not only from cost and infrastructure limitations but also from concerns around compliance, governance, and organizational readiness. Even when performance improvements are observed, they may not justify the added complexity and risk involved in deploying very advanced and non transparent models at scale. In this environment, simpler and more interpretable approaches often remain the preferred option particularly when they already meet regulatory and operational requirements and allow for meaningful human interaction and oversight in the decision making process (Marous, 2024).

A notable recent development, however, is the rise of generative AI. As discussed earlier, generative tools such as ChatGPT and also Microsofts Copilot are being explored and actively used already for tasks like document summarization, internal communication support, and customer interaction. Yet their role remains

limited when it comes to working with sensitive internal banking data. Privacy concerns, data protection regulations, and the risk of unintended information disclosure significantly restrict how generative AI can be integrated into core systems and used by bank employees. Furthermore, generative models are rarely embedded into workflows where full auditability and reliability are required, limiting their application in high stakes decision making contexts as of now. Beyond regulatory and data constraints, broader adoption of AI in banking is also shaped by operational and institutional barriers. Many banks continue to rely on legacy IT infrastructure that is not designed to support ML models at scale. As looked into earlier, successful deployment often requires dedicated teams across data science, compliance, IT, and business operations resources that are not always available or aligned. Cultural resistance can also slow adoption, especially when models are difficult to interpret or challenge established practices. In some cases, even when technical capacity exists, institutional trust in AI systems remains low particularly for decisions that directly impact customers or financial risk exposure since the consequences of incorrect or non transparent model behavior can be severe (Marous, 2024).

Regulatory Considerations

As emphasized throughout, regulation continues to be one of the most decisive factors shaping how AI is used and can be used within banks. Even when the models themselves appear promising, their real world application depends on how well they align with legal, ethical, and supervisory frameworks. This means a models value depends not just on its technical performance, but also on how well it fits into the realities of the financial sector. Ethical concerns, such as fairness, accountability, and the potential for biased or discriminatory outcomes, are especially important in banking, where algorithmic decisions can directly influence an individuals financial opportunities, credit access, or fraud exposure. Addressing these concerns is not only a regulatory requirement it is also crucial for maintaining institutional credibility and public trust. As AI continues to evolve, its adoption in finance must be accompanied by careful consideration of how these systems affect both individuals and society more broadly (Max et al., 2021).

From a regulatory perspective in the western context, the EU has taken a structured, risk based approach through the AI Act, which sets strict rules for high risk applications like credit scoring and fraud detection (European Commission, 2024c). Key principles include transparency, human oversight, and high data quality. At the same time, the EU supports responsible innovation. Programs such as InvestAI aim to channel significant funding into the development of AI, while the European AI Office ensures consistent rule enforcement across Member States (European Commission, 2025). The European Central Bank (2024) has also recognized that, when implemented responsibly, AI has the potential to strengthen risk management and improve operational efficiency in the financial sector.

The U.S., by contrast, takes a more decentralized and sector driven path. Rather than imposing unified

legislation, agencies such as the Department of the Treasury engage stakeholders directly to explore both the risks and benefits of AI in finance. The U.S. Department of State (2024) has published a broader AI strategy emphasizing ethical and secure use, particularly in diplomatic and public-facing systems. While the he EU leans toward a precautionary, top down model, the U.S. emphasizes flexibility and private sector input. These regulatory differences shape how and how fast AI can be adopted widely in banking. While regulation may appear to slow implementation, it can also build trust and provide the structure necessary for responsible innovation. If aligned with institutional capabilities, regulation may not only constrain AI, but also help enable it.

7. Conclusion

This thesis examined whether advanced AI models offer meaningful improvements over traditional statistical and ML methods in key areas of banking. Through a combination of empirical analysis and contextual assessment, the study explored fraud detection, credit risk assessment, and portfolio optimization, as well as the broader landscape of AI adoption in western banking institutions. The findings suggest that while AI based approaches can surpass traditional techniques in specific tasks, their broader implementation remains constrained by institutional, regulatory, and interpretability related factors.

In the domains of fraud detection and credit scoring, ensemble methods such as Random Forest and XGBoost yielded higher F1-scores and AUC values than logistic regression, reflecting their ability to model complex, non-linear relationships within financial data. However, their increased complexity also raises concerns regarding transparency. To address this, SHAP values were used to provide local explanations of model behavior. While this improved transparency to some extent, existing concerns remain particularly in high stakes financial applications regarding the consistency and reliability of feature attributions. Unsupervised methods showed limited value in this setting, though recent literature points to the potential of hybrid models under more varied and data rich conditions.

Portfolio optimization led to more mixed results. Forecasting based strategies using deep learning models such as LSTM and GRU achieved lower prediction errors compared to traditional benchmarks, but these gains did not consistently translate into better portfolio performance when evaluated using Sharpe ratios. These results reflect the broader challenge of predicting financial markets, which are often considered efficient with respect to publicly available information thereby limiting the possibility of sustained outperformance through data driven prediction alone.

While the results indicate that advanced AI models may offer enhanced predictive capabilities, they also support the broader perspective in existing research that performance metrics alone do not suffice to establish practical utility. The integration of AI within banking operations remains influenced by various external factors, including regulatory preparedness, data quality, and institutional capacity. Despite heightened interest particularly in generative models, which currently represent one of the most widely adopted forms of AI in banking, actual implementation within core banking systems has progressed cautiously. As the field advances, financial institutions that combine technological innovation with robust governance and regulatory compliance seem more likely to derive lasting value. Drawing on both the empirical results and prior studies, this thesis suggests that advanced AI can provide practical improvements over traditional methods and when applied responsibly improve decision making in modern banking.

8. Appendix

A: Features Fraud Dataset A

	Column Names
0	Time
1	V1
2	V2
3	V3
4	V4
5	V5
6	V6
7	V7
8	V8
9	V9
10	V10
11	V11
12	V12
13	V13
14	V14
15	V15
16	V16
17	V17
18	V18
19	V19
20	V20
21	V21
22	V22
23	V23
24	V24
25	V25
26	V26
27	V27
28	V28
29	Amount
30	Class

B: Features Fraud Dataset B

	Column Names
0	step
1	type
2	amount
3	nameOrig
4	oldbalanceOrg
5	newbalanceOrig
6	nameDest
7	oldbalanceDest
8	newbalanceDest
9	isFraud
10	isFlaggedFraud

C: Features Credit Dataset A

	Column Names
0	person_age
1	person_income
2	person_home_ownership
3	person_emp_length
4	loan_intent
5	loan_grade
6	loan_amnt
7	loan_int_rate
8	loan_status
9	loan_percent_income
10	cb_person_default_on_file
11	cb_person_cred_hist_length

D: Features Credit Dataset B

	Column Names
0	label
1	id
2	fea_1
3	fea_2
4	fea_3
5	fea_4
6	fea_5
7	fea_6
8	fea_7
9	fea_8
10	fea_9
11	fea_10
12	fea_11
13	OVD_t1
14	OVD_t2
15	OVD_t3
16	OVD_sum
17	pay_normal
18	prod_code
19	prod_limit
20	update_date
21	new_balance
22	highest_balance
23	report_date

E: Example of hyperparameter tuning Fraud and Credit

```python
# Hyperparameter distributions
param_dist = {
    'n_estimators': [100, 150, 200],
    'max_depth': [None, 10, 20, 30],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'max_features': ['sqrt', 'log2']
}

# Initialize base model
rf = RandomForestClassifier(random_state=10)

# Setting up RandomizedSearchCV
random_search = RandomizedSearchCV(
    estimator=rf,
    param_distributions=param_dist,
    n_iter=10,
    scoring='roc_auc',
    cv=5,
    random_state=10,
    n_jobs=-1
)

# Fits the model on SMOTE-resampled and scaled training data
random_search.fit(X_train_smote, y_train_smote)

# Best model
best_rf = random_search.best_estimator_

# Predict on the scaled original test set
y_pred_rf = best_rf.predict(X_test_scaled)
y_proba_rf = best_rf.predict_proba(X_test_scaled)[:, 1]

# Evaluate
print("Best Parameters:", random_search.best_params_)
print("\nRandom Forest 🌲 Classification Report")
print(classification_report(y_test, y_pred_rf, digits=4))
print(f"AUC: {roc_auc_score(y_test, y_proba_rf):.4f}")
```

```
Best Parameters: {'n_estimators': 100, 'min_samples_split': 5, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': 20}
```

F: Default hyperparameters of LSTM and GRU

```python
# LSTM Model
model = Sequential()
model.add(LSTM(units=64, return_sequences=True, input_shape=(look_back, 1)))
model.add(Dropout(0.2))
model.add(LSTM(units=32))
model.add(Dropout(0.2))
model.add(Dense(1))

model.compile(loss='mean_squared_error', optimizer='adam')
```

```python
# GRU model
model = Sequential()
model.add(GRU(units=64, return_sequences=True, input_shape=(look_back, 1)))
model.add(Dropout(0.2))
model.add(GRU(units=32))
model.add(Dropout(0.2))
model.add(Dense(1))

model.compile(loss='mean_squared_error', optimizer='adam')
```

G: Stocks selected test 2 portfolio optimization

```python
# Parameters
tickers = [
    'AVGO', 'BRK-B', 'JNJ', 'PG', 'HD', 'MRK', 'KO', 'PEP', 'COST', 'ADBE',
    'INTC', 'WLK', 'SMCI', 'DECK', 'MANH', 'NDSN', 'POOL', 'HUBB', 'STLD', 'RPM'
]

start_date = '2020-05-01'
end_date = '2025-01-01'
train_end_date = '2023-12-31'
```

References

[1] Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit risk analysis using machine and deep learning models. *Risks, 6*(2), 38.

[2] Agu, E. E., Abhulimen, A. O., Obiki-Osafiele, A. N., Osundare, O. S., Adeniran, I. A., & Efunniyi, C. P. (2024). Discussing ethical considerations and solutions for ensuring fairness in AI-driven financial services. *International Journal of Frontline Research in Multidisciplinary Studies*, 3(2), 001-009.

[3] Anderson, R. (2007). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press.

[4] Atadoga, A., Obi, O. C., Onwusinkwue, S., Dawodu, S. O., Osasona, F., & Daraojimba, A. I. (2024). AI's evolving impact in US banking: An insightful review. *International Journal of Science and Research Archive, 11*(1), 904-922.

[5] Bao, Y., Hilary, G., & Ke, B. (2020). Artificial intelligence and fraud detection. In *Innovative Technology at the Interface of Finance and Operations: Volume I* (pp. 223-247).

[6] Bello, O. A., Ogundipe, A., Mohammed, D., Adebola, F., & Alonge, O. A. (2023). AI-Driven Approaches for Real-Time Fraud Detection in US Financial Transactions: Challenges and Opportunities. *European Journal of Computer Science and Information Technology, 11*(6), 84–102.

[7] Boobier, T. (2020). *AI and the Future of Banking*. John Wiley & Sons.

[8] Bostrom, N., & Yudkowsky, E. (2018). The ethics of artificial intelligence. In *Artificial intelligence safety and security* (pp. 57-69). Chapman and Hall/CRC.

[9] Brenner, L., & Meyll, T. (2020). Robo-advisors: A substitute for human financial advice? *Journal of Behavioral and Experimental Finance, 25*, 100275.

[10] Brown, M. (2024). Influence of artificial intelligence on credit risk assessment in banking sector. *International Journal of Modern Risk Management, 2*(1), 24–33.

[11] Brynjolfsson, E., & McAfee, A. (2017). The business of artificial intelligence. *Harvard Business Review, 7*(1), 1-2.

[12] Candemir, M. (2025). Traditional logistic regression vs. modern machine learning in credit scoring: A practical overview. *Towards AI*, 2025.

[13] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

[14] Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

[15] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

[16] Citibank. (2024). *AI in Finance: Citi publishes new report*. Retrieved from `https://www.citigroup.com/global/news/press-release/2024/citi-publishes-new-report-ai-in-finance`

[17] CMGlee, Wikimedia Commons contributors. (2021). *ROC curve*. Retrieved from `https://upload.wikimedia.org/wikipedia/commons/archive/1/13/20240927054927%21Roc_curve.svg`

[18] Cohen, G. (2022). Algorithmic trading and financial forecasting using advanced artificial intelligence methodologies. *Mathematics, 10*(18), 3302.

[19] Consumer Financial Protection Bureau (CFPB). (2023). *CFPB issues guidance on credit denials by lenders using artificial intelligence*. Retrieved from `https://www.consumerfinance.gov/about-us/newsroom/cfpb-issues-guidance-on-credit-denials-by-lenders-using-artificial-intelligence`

[20] Crouhy, M., Galai, D., & Mark, R. (2000). A comparative analysis of current credit risk models. *Journal of Banking & Finance, 24*(1–2), 59–117.

[21] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2017). *Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy*. IEEE Transactions on Neural Networks and Learning Systems.

[22] De Prado, M. L. (2018). *Advances in financial machine learning*. John Wiley & Sons

[23] Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). *Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects*. European Journal of Operational Research, 295(2), 631–647.

[24] Edunjobi, T. E., & Odejide, O. A. (2024). Theoretical frameworks in AI for credit risk assessment: Towards banking efficiency and accuracy. *International Journal of Scientific Research Updates, 7*(01), 092–102.

[25] Elton, E. J., Gruber, M. J., Brown, S. J., & Goetzmann, W. N. (2014). *Modern portfolio theory and investment analysis* (9th ed.). Wiley.

[26] European Commission. (2024a). *Proposal for a Regulation laying down harmonised rules on Artificial Intelligence (AI Act)*. Retrieved from `https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence`

[27] European Commission. (2024b). *European AI Office*. Retrieved from https://digital-strategy.ec.europa.eu/en/policies/ai-office

[28] European Commission. (2024c). *Europe's approach to artificial intelligence*. Retrieved from https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

[29] European Commission. (2025). EU launches InvestAI initiative to mobilise 200 billion for artificial intelligence. *European Commission Representation in Luxembourg*. Retrieved from https://luxembourg.representation.ec.europa.eu/actualites-et-evenements/actualites/eu-launches-investai-initiative-mobilise-eu200-billion-investment-artificial-intelligence-2025-02-11_en

[30] European Central Bank. (2024). *The impact of AI in financial services: benefits, risks and regulation*. *Financial Stability Review*. Retrieved from https://www.ecb.europa.eu/press/financial-stability-publications/fsr/special/html/ecb.fsrart202405_02~58c3ce5246.en.html

[31] Faisal, N. A., Nahar, J., Sultana, N., & Mintoo, A. A. (2024). Fraud Detection in Banking: Leveraging AI to Identify and Prevent Fraudulent Activities in Real-Time. *Journal of Machine Learning, Data Engineering and Data Science, 1*(01), 181–197.

[32] Fama, E. F. (1970). Efficient capital markets. *Journal of Finance, 25*(2), 383–417.

[33] Fan, M. H., Chen, M. Y., & Liao, E. C. (2021). A deep learning approach for financial market prediction: Utilization of Google trends and keywords. *Granular Computing, 6*(1), 207–216. https://doi.org/10.1007/s41066-020-00205-3

[34] Fawcett, T. (2006). *An introduction to ROC analysis*. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

[35] Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering, 66*(1), 111–126.

[36] GeeksforGeeks. (n.d.). *ML — K-Means Algorithm*. Retrieved from https://www.geeksforgeeks.org/ml-k-means-algorithm/

[37] Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media

[38] Ghodselahi, A., & Amirmadhi, A. (2011). Application of artificial intelligence techniques for credit risk evaluation. *International Journal of Modeling and Optimization, 1*(3), 243.

[39] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223–2273.

[40] Financial Stability Board. (2017). *Financial stability implications from fintech.* Retrieved from `https://www.fsb.org/2017/06/financial-stability-implications-from-fintech/`

[41] Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review, 61*(4), 5-14.

[42] Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: Deep portfolios. *Applied Stochastic Models in Business and Industry, 33*(1), 3-12.

[43] Huang, X., & Marques-Silva, J. (2023). The inadequacy of Shapley values for explainability. *arXiv preprint arXiv:2302.08160.*

[44] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R.* Springer.

[45] Javaid, H. A. (2024). Revolutionizing AML: How AI is Leading the Charge in Detection and Prevention. *Journal of Innovative Technologies, 7*, 1–9.

[46] Johora, F. T., Hasan, R., Farabi, S. F., Akter, J., & Al Mahmud, M. A. (2024). AI-Powered Fraud Detection in Banking: Safeguarding Financial Transactions. *The American Journal of Management and Economics Innovations, 6*(06), 8–22.

[47] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255-260.

[48] Kaya, O. (2019). *Artificial intelligence in banking: A lever for profitability with limited implementation to date.* Deutsche Bank Research, EU Monitor.

[49] Ko, H., & Lee, J. (2024). Can ChatGPT improve investment decisions? From a Portfolio Optimization perspective. *Finance Research Letters, 64*, 105433.

[50] Kumar, A., Dikshit, S., & Albuquerque, V. H. C. (2021). Explainable artificial intelligence for sarcasm detection in dialogues. *Wireless Communications and Mobile Computing, 2021*(1), 2939334.

[51] KPMG. (2024). *Decoding the EU Artificial Intelligence Act*. Retrieved from `https://kpmg.com/us/en/articles/2023/decoding-eu-ai-act.html`

[52] Lopez-Lira, F., & Tang, Y. (2023). Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. *Available at SSRN 4433342.*

[53] Lo, A. W. (2007). Efficient markets hypothesis. *Palgrave Macmillan.*

[54] Lopez-Rojas, E. (2018). *Synthetic Financial Datasets For Fraud Detection (PaySim)* [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/ealaxi/paysim1

[55] Lu, T., Zhang, Y., & Li, B. (2019). The value of alternative data in credit risk prediction: Evidence from a large field experiment.

[56] Lundberg, S. M., (2018). SHAP (SHapley Additive exPlanations) Documentation. Retrieved from *https://shap.readthedocs.io/en/latest/*

[57] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.

[58] Lynn, T., Mooney, J. G., Rosati, P., & Cummins, M. (2019). *Disrupting Finance: FinTech and Strategy in the 21st Century* (p. 175). Springer Nature.

[59] Mercanti, L. (2024). AI in Portfolio Optimization: An Overview. *Medium*, 2023. Retrieved from `https://leomercanti.medium.com/ai-in-portfolio-management-an-overview-9144d777f2d3`

[60] Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Vintage.

[61] Markowitz, H. M. (1991). Foundations of portfolio theory. *The Journal of Finance, 46*(2), 469–477.

[62] Marous, J. (2024). *State of AI in Banking*. Digital Banking Report, August 2024.

[63] Martin, R. A. (2021). PyPortfolioOpt: portfolio optimization in Python. *Journal of Open Source Software, 6*(61), 3066. https://doi.org/10.21105/joss.03066

[64] Max, R., Kriebitz, A., & Von Websky, C. (2021). Ethical considerations about the implications of artificial intelligence in finance. *Handbook on Ethics in Finance*, 577–592.

[65] Max, A., Stevens, C., & Lewis, R. (2021). Ethical concerns in AI-driven finance: Accountability and fairness in algorithmic decision-making. *Journal of Financial Ethics*, 18(2), 45–63.

[66] Mayer Brown. (2022). *Supervisory Expectations for Artificial Intelligence Outlined by US OCC*. Retrieved from `https://www.mayerbrown.com/en/insights/publications/2022/05/supervisory-expectations-for-artificial-intelligence-outlined-by-us-occ`

[67] McKinsey, Chui, M., Harryson, M., Manyika, J., Roberts, R., Chung, R., van Heteren, A., & Nel, P. (2023). *The economic potential of generative AI: The next productivity frontier*. McKinsey & Company.

[68] McKinsey, Kremer, A., Govindarajan, A., Singh, H., Kristensen, I., & Li, E. (2024). Embracing generative AI in credit risk. *McKinsey & Company, Risk & Resilience Practice*, July 2024.

[69] Mishkin, F. S. (2019). *The economics of money, banking, and financial markets* (12th ed.). Pearson.

[70] Mokhtari, S., Yen, K. K., & Liu, J. (2021). Effectiveness of artificial intelligence in stock market prediction based on machine learning. *arXiv preprint arXiv:2107.01031*.

[71] Mhlanga, D. (2021). Financial inclusion in emerging economies: The application of machine learning and artificial intelligence in credit risk assessment. *International Journal of Financial Studies, 9*(3), 39.

[72] Misheva, B. H., Osterrieder, J., Hirsa, A., Kulkarni, O., & Lin, S. F. (2021). Explainable AI in credit risk management. *arXiv preprint arXiv:2103.00949*.

[73] Nathan, C., Alalyani, A., Serbanoiu, A., & Khan, D. (2023). AI-Powered Data Governance: Ensuring Integrity in Banking's Technological Frontier. *Unpublished manuscript*

[74] Nets & KPMG. (2020). *Fighting Fraud with a Model of Models: Utilising Artificial Intelligence to Prevent Payment Card Fraud*. Nets Whitepaper.

[75] Niu, X., Wang, L., & Yang, X. (2019). *A Comparison Study of Credit Card Fraud Detection: Supervised versus Unsupervised*. arXiv preprint arXiv:1904.10604.

[76] OpenAI. (2024). Introducing o3 and o4-mini. *OpenAI Index*. Retrieved from `https://openai.com/index/introducing-o3-and-o4-mini/`

[77] OpenAI. (2022). *ChatGPT*. Retrieved from `https://openai.com/index/chatgpt/`

[78] Perlin, M. S., Foguesatto, C. R., Müller, F. M., & Righi, M. B. Can AI beat a naive portfolio? An experiment with anonymized data. *Manuscript in preparation or unpublished work*.

[79] Rasekhschaffe, K. C., & Jones, R. C. (2019). Machine learning for stock selection. *Financial Analysts Journal, 75*(3), 70–88.

[80] Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach*. Pearson.

[81] Romanko, O., Narayan, A., & Kwon, R. H. (2023). ChatGPT-based investment portfolio selection. *Operations Research Forum, 4*(4), 91.

[82] Siami-Namini, S., Tavakoli, N., & Siami Namin, A. (2018, December). A comparison of ARIMA and LSTM in forecasting time series. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1394–1401). IEEE.

[83] U.S. Department of the Treasury. (2024). *Artificial Intelligence in Financial Services: Request for Information*.

[84] U.S. Department of State. (2024). *Enterprise Artificial Intelligence Strategy FY 2024–2025*. Retrieved from `https://2021-2025.state.gov/artificial-intelligence/`

[85] Zhang, X., Xu, L., Li, N., & Zou, J. (2024). Research on credit risk assessment optimization based on machine learning. *Applied and Computational Engineering, 69*, 173

[86] Öztornaci, B. (2024). Simplified structure of XGBoost [Figure].