Title: Application of Artificial Intelligence in the area of Anti-Money Laundering

By:

Mark Pillon

Thesis

Submitted in partial fulfillment of the requirements

for the degree of MSc Data Science and Artificial Intelligence

OPIT - Open Institute of Technology

St. Julian's, Malta

Adviser:

Professor Lokesh Vij

Abstract

This thesis explores the application of automation and machine learning to improve Buffetti's Anti-Money Laundering (AML) processes, which are currently reliant on manual methods using Excel. The primary objective was to enhance the efficiency, accuracy, and scalability of these processes by implementing machine learning models, specifically the Isolation Forest algorithm, to detect suspicious financial transactions. The study addresses the limitations of manual AML operations, which struggle with large datasets and complex money laundering patterns.

The research methodology involved consolidating data from 13 transaction tables, developing Python scripts for automation, and integrating machine learning models. Tools such as Microsoft Power BI and Python were employed for data analysis and visualization, while the potential of Microsoft Fabric was considered for real-time scalability and data management.

Key findings revealed that machine learning significantly outperforms manual methods, particularly in detecting subtle anomalies indicative of money laundering. The Isolation Forest algorithm effectively identified suspicious transactions, highlighting its potential for improving AML compliance. However, challenges were identified, including the absence of regulatory feedback, Power BI's limitations with large datasets, and the lack of real-time implementation and model validation.

While machine learning offers significant improvements in AML accuracy, efficiency, and cost-effectiveness, further validation and real-world testing are required. Collaboration with regulators and stakeholders is also necessary to fully realize the potential of automated AML processes in large-scale, practical applications.

Acknowledgements

I would like to express my deepest gratitude to my advisor, **Lokesh Vij**, for his invaluable guidance, support, and encouragement throughout this research. His expertise and insights were instrumental in shaping the direction and outcomes of this thesis.

I am also sincerely thankful to the team at Buffetti—**Davide Cristiano**, **Matteo Pagnao**, **Giovanni Lanzilotti**, **Valeria Pizzutilo**, and **Vincenzo Ventrella**—for their collaboration and support. Their practical insights, access to essential data, and continuous assistance were crucial to the success of this project.

List of Figures/Tables:

Chapter 1: Introduction

1.1 The background of the problem

The primary issue under investigation is the inefficiency and error-proneness of the current manual Anti-Money Laundering (AML) processes at Buffetti Finance. While these manual processes are functional, they present six specific problems that significantly hinder the company's ability to meet the intricate and stringent regulatory demands imposed by both regional and supranational bodies, particularly those of the European Union. These problems include regulatory compliance risks, operational inefficiencies, high error rates, challenges in scalability and real-time monitoring, the loss of competitive advantage, and misalignment with the company's strategic goals.

Ensuring compliance with AML regulations is critical for any financial institution, and Buffetti Finance is no exception. Failure to comply can result in severe legal penalties, substantial financial losses, and irreparable damage to the company's reputation. For Buffetti Finance, adherence to these regulations is not merely a legal obligation; it is also fundamental to the company's operational integrity and the trust it has built with its customers.

The current manual processes at Buffetti Finance are notably inefficient, leading to extended processing times and potential delays in the identification and reporting of suspicious activities. By transitioning to automated AML processes, the company could significantly reduce processing times, enabling a quicker response to potential threats and thereby enhancing overall operational efficiency.

Manual processes are inherently prone to human error, which can result in missed detections of suspicious activities or false positives, both of which can be costly. The integration of

automation and machine learning technologies could dramatically improve the accuracy of anomaly detection, ensuring more reliable compliance with AML regulations and reducing the risk of costly mistakes.

As Buffetti Finance continues to expand its services, the volume of transactions increases, rendering manual monitoring increasingly impractical. Automated systems, however, are capable of handling larger datasets and providing real-time monitoring, which is essential for the timely detection and response to suspicious activities.

Moreover, by upgrading its AML processes, Buffetti Finance stands to gain a competitive advantage in the market. The ability to offer more robust and reliable services to clients will not only enhance customer satisfaction and trust but could also position Buffetti Finance as a leader in the fintech industry. This technological advancement can set a new benchmark for innovative compliance solutions in the sector.

Enhancing AML processes aligns closely with Buffetti Finance's broader mission of providing innovative and tailored fintech solutions. It supports the company’s vision of driving digital transformation and operational excellence, reinforcing its leadership in the Italian market and potentially beyond.

In conclusion, addressing the six specific problems—regulatory compliance risks, operational inefficiencies, potential error rates, scalability challenges, competitive disadvantages, and strategic misalignment—associated with the current manual AML processes is critical for Buffetti Finance. The exploration and implementation of automated, machine learning-enhanced AML solutions is not merely beneficial but essential for the company’s future success.

1.2 Problem Statement

1.2.1 Initial Question and Evolution of Research Focus

The research initially aimed to answer: "How can automation and machine learning technologies improve the efficiency and accuracy of Buffetti's Anti-Money Laundering (AML) processes compared to the current manual execution via Excel files?"

This question arose from the recognition that Buffetti's manual AML processes were time-consuming, error-prone, and costly, posing significant challenges in meeting the intricate regulatory demands.

1.2.2 Discovery and Emerging Problems

1.2.2.1 Accuracy and Reliability of AML Processes

Upon examining Buffetti's current manual processes, it became evident that accuracy in terms of scope was a major concern due to the over abundance in data and the analysis. This led to the following new questions:

- **What specific errors and inaccuracies are prevalent in the current manual AML processes?**
- **How do machine learning models compare in terms of accuracy and reliability for anomaly detection?**

1.2.2.2 Operational Cost Implications: The substantial allocation of human resources and the associated operational costs highlighted the need to explore the financial aspects of transitioning to automated processes:

- **How much time and human resources are currently allocated to manual AML processes?**
- **What are the projected cost savings and efficiency gains from automation?**

1.2.2.3 Timeliness and Regulatory Compliance: The time-intensive nature of manual processes raised concerns about meeting regulatory deadlines, prompting further inquiry into how automation could enhance compliance timeliness:

- **What are the current bottlenecks in the manual AML process that lead to delays?**
- **How can automation reduce these bottlenecks and enhance the speed of compliance activities?**

1.2.2.4 Effectiveness of Anomaly Detection: Initial analysis indicated that anomaly detection could be a feasible approach, but data complexity posed significant challenges:

- **What types of anomalies are most indicative of money laundering activities?**
- **How does the complexity of Buffetti's data impact the performance of anomaly detection algorithms?**

1.2.2.5 Data Handling and Integration Challenges: The complexity of Buffetti's data structure, consisting of 13 different transaction tables and limited data handling capabilities of PowerBI, necessitated new strategies for effective data management:

- **What are the challenges associated with Buffetti's 13 different transaction tables and the limitations of PowerBI?**
- **How can data be consolidated and pre-processed to facilitate effective analysis and machine learning implementation?**

1.2.2.6 Impact of Regulator Feedback: The lack of feedback from regulators became a critical issue, affecting the predictive capabilities of machine learning models:

- **How does the absence of regulator feedback affect the accuracy of AML models, particularly regarding false positives and false negatives?**
- **What strategies can be implemented to mitigate the effects of information asymmetry between regulators and Buffetti?**

1.2.2.7 Preliminary Insights from Initial Focus Stage: Focusing on the initial stage of the money laundering process (placement) provided some insights but also highlighted the need for a broader examination:

- **What patterns and trends have been identified in the placement stage of money laundering?**
- **How can these insights be expanded to cover the layering and integration stages?**

The initial question on improving AML processes through automation and machine learning evolved into a more comprehensive investigation, driven by the discovery of specific challenges and complexities within Buffetti's current system. This led to a series of new, more detailed research questions aimed at addressing the multifaceted nature of AML compliance, operational efficiency, data management, and the interplay with regulatory feedback. By systematically exploring these emerging questions, the research aims to develop a robust, automated AML solution that enhances Buffetti's compliance efforts and operational performance.

1.3 Research Objectives

The primary objective of this thesis is to develop and implement an automated Anti-Money Laundering (AML) process that enhances the efficiency and effectiveness of Buffetti Finance's compliance operations. To achieve this, the research is guided by the following specific objectives:

1.3.1 Develop an Anomaly Checking Model

Goal: The first objective is to create a machine learning model capable of identifying real-time anomalies in financial transactions.

Rationale: Detecting anomalies as they occur allows Buffetti Finance to address suspicious activities promptly. This capability reduces the risk of non-compliance and enhances the overall security of financial operations. Real-time anomaly detection is crucial for maintaining the integrity of AML processes and ensuring that potential threats are mitigated before they escalate.

1.3.2 Automate Delivery Operations

Goal: The second objective focuses on streamlining and automating compliance-related processes to improve operational efficiency and productivity.

Rationale: Automation will significantly reduce the time and human resources required to carry out AML processes. This, in turn, will minimize errors and ensure more consistent and reliable compliance with regulatory requirements. By automating these tasks, Buffetti Finance can achieve a higher level of accuracy and consistency in its AML operations, which is essential for maintaining regulatory compliance and operational integrity.

1.3.3 Evaluate Model Performance

Goal: The third objective is to assess the performance of the AML model using relevant metrics and benchmarks.

Rationale: Evaluating the model's performance is crucial for ensuring its effectiveness in

identifying anomalies. This evaluation will involve comparing the model's accuracy and reliability against existing manual processes. Such an assessment will provide data-driven insights that are necessary for making informed decisions about further improvements and refinements to the AML system.

1.3.4 Provide Insights and Recommendations

Goal: The fourth objective is to offer actionable insights and recommendations to stakeholders based on the results of the AML model.

Rationale: Analyzing the outcomes and performance of the AML model will generate valuable insights for stakeholders. These insights will inform strategic decisions, enhance compliance strategies, and guide future improvements in AML processes. Providing clear and data-backed recommendations will ensure that Buffetti Finance remains proactive in its compliance efforts.

1.4 Specific Goals to Achieve Through the Research

In pursuing these objectives, the research aims to achieve the following specific goals:

1.4.1 Improve Accuracy in AML Processes

The research aims to develop a robust anomaly detection system that minimizes both false positives and false negatives, thereby enhancing the precision of identifying suspicious activities. Accurate detection is vital for maintaining compliance and reducing the risk of overlooking potentially illegal transactions.

1.4.2 Enhance Operational Efficiency

By implementing automation tools, the research seeks to reduce the manual workload, lower

operational costs, and accelerate the processing of AML compliance tasks. This goal is critical for optimizing the use of resources and improving the overall efficiency of Buffetti Finance's compliance operations.

1.4.3 Ensure Timely Compliance

The research aims to develop systems and processes that ensure Buffetti Finance meets all regulatory deadlines promptly. Timely compliance is essential to avoid penalties and maintain the company’s reputation for regulatory adherence.

1.4.4 Optimize Data Handling

The research will focus on creating strategies to manage and analyze Buffetti Finance’s complex transaction data efficiently. This goal is important for overcoming current limitations and enhancing the comprehensiveness of AML efforts, ensuring that the company can effectively monitor and respond to suspicious activities.

1.4.5 Facilitate Better Stakeholder Communication

The research will provide clear and actionable reports and recommendations to stakeholders based on comprehensive data analysis and model results. This goal aims to foster informed decision-making, ensuring that stakeholders have the necessary insights to guide the company’s strategic direction in AML compliance.

By achieving these goals, the research intends to significantly elevate Buffetti Finance's AML compliance capabilities. This will not only ensure regulatory adherence but also optimize operational performance and reduce associated costs, positioning Buffetti Finance as a leader in compliance innovation within the fintech industry.

1.5 Thesis Statement

This research investigated how automation and machine learning technologies could improve Buffetti's Anti-Money Laundering (AML) processes compared to the current manual methods executed via Excel files. The study demonstrated that automation, specifically through the use of machine learning models like the Isolation Forest, significantly enhances both the efficiency and accuracy of AML operations. However, the research also revealed challenges, such as the absence of regulator feedback, which limits the ability to fully evaluate the model's performance using traditional statistical methods. Despite these challenges, the findings underscore the potential of automation to transform AML processes, particularly when supported by robust cloud infrastructure like Microsoft Fabric.

1.6 Scope and Limitations

The scope of this research is centered on enhancing Buffetti's Anti-Money Laundering (AML) processes by leveraging automation and machine learning technologies. The study primarily focuses on the placement stage of money laundering, offering valuable insights into anomaly detection and resource allocation. However, this limited scope excludes the more intricate stages of layering and integration, underscoring the necessity for a broader analysis to fully capture the complexities of money laundering activities.

One of the significant limitations encountered was the dependence on traditional rule-based AML systems, which are static, require frequent updates, and struggle with semi-structured and unstructured data. This reliance often results in high false positive rates and inefficiencies, highlighting the need for more dynamic and adaptable machine learning solutions. However,

implementing machine learning models like Isolation Forest posed its own challenges, particularly in terms of data complexity and the resource-intensive nature of training and retraining these models.

Another limitation was the absence of feedback from regulators, which impeded the ability to assess the accuracy of the machine learning models, particularly in distinguishing between false positives and false negatives. This information asymmetry between financial institutions and regulators contributes to a Nash equilibrium, leading to suboptimal AML outcomes and complicating efforts to improve model performance.

The complexity of Buffetti’s data, spread across 13 transaction tables and limited by Power BI's constraints, also posed significant challenges. The necessity of maintaining client confidentiality further complicated data processing, although successful data compression and the identification of essential data columns were key achievements that facilitated more efficient analysis.

While the research made strides in automating and enhancing Buffetti's AML processes, it was constrained by the limitations of traditional AML systems, data complexity, lack of regulatory feedback, and the challenges of implementing machine learning models. Future research and development should focus on overcoming these limitations by exploring more efficient data management techniques, improving stakeholder communication, and considering advanced platforms like Microsoft Fabric to optimize data processing and analysis workflows.

1.7 Outline of Thesis

Chapter 1: Introduction

The introduction sets the stage for the research by highlighting the inefficiencies and inaccuracies inherent in traditional Anti-Money Laundering (AML) processes. It details the research question, which focuses on how automation and machine learning technologies can enhance the efficiency and accuracy of AML operations compared to the current manual execution using Excel files. This chapter establishes the rationale for the study, outlining the critical problems faced by Buffetti, such as time-consuming and error-prone processes, and the necessity for an innovative approach to meet regulatory demands effectively.

Chapter 2: Literature Review

The literature review provides an overview of existing research and key findings relevant to the study. It begins by examining traditional rule-based AML systems, which rely on static, predefined rules that require frequent updates to address evolving money laundering techniques. These systems are often inadequate for handling semi-structured and unstructured data, leading to high false positive rates. The review emphasizes the need for more dynamic and adaptable AML solutions that can process diverse data types and reduce reliance on manual adjustments.

The review continues by exploring various machine learning (ML) approaches to AML. It highlights how ML models, such as Decision Trees, Random Forests, and Neural Networks, offer significant improvements in detecting suspicious activities by learning patterns from training data. Studies reviewed demonstrate that ML techniques, including Support Vector Machines (SVM) and Random Forests, outperform traditional methods in reducing false positives and enhancing detection accuracy. However, these models require substantial training time and effort, which can be a limiting factor.

Another key area of focus in the literature review is anomaly detection. Recent studies have introduced methods like the Isolation Forest algorithm, which enhances anomaly detection by

identifying outliers and potentially suspicious transactions. These methods offer a more comprehensive approach compared to traditional rule-based systems and are particularly valuable in handling complex datasets.

The review also addresses the importance of explainable AI in AML. Many ML models lack transparency, making it challenging for domain experts to interpret and trust the results. The review underscores the need for machine learning models that not only improve detection rates but also provide interpretable insights that support expert evaluations and decision-making processes.

Information asymmetry and the lack of feedback from regulators are identified as significant challenges in the literature. The absence of regulator feedback undermines the predictive capabilities of AML models and complicates the assessment of accuracy, particularly concerning false positives and false negatives. The review highlights the need for improved communication between financial institutions and regulators to enhance the effectiveness of AML models.

Finally, the literature review examines data complexity and handling challenges. Existing studies reveal that the complexity of data, including multiple transaction tables, and the limitations of tools like Power BI can hinder AML systems' performance. There is a clear gap in developing efficient data handling and preprocessing techniques that can manage complex datasets and integrate seamlessly with machine learning models.

Chapter 3: Methodology

Chapter 3 outlines the research design and methodology used to address the research question. The study adopts an observational and descriptive approach, focusing on analyzing existing transaction data without experimental manipulation. It details the processes involved in data

collection, including extraction, preprocessing, and handling, with particular attention to maintaining client confidentiality and using tools like Power BI and Python for data analysis.

The chapter describes the implementation of anomaly detection techniques, specifically the Isolation Forest algorithm, to identify suspicious transactions. It explains the process of setting thresholds, managing data complexity, and evaluating the model's effectiveness. Additionally, the chapter explores the integration of technologies such as Microsoft Fabric and Azure services to enhance data processing capabilities and scalability.

Chapter 4: Results and Discussion

In Chapter 4, the research findings are analyzed and discussed. The effectiveness of the developed anomaly detection model is evaluated, highlighting improvements in accuracy and efficiency compared to manual methods. The chapter assesses the impact of automation on reducing manual labor, lowering operational costs, and addressing bottlenecks in traditional AML processes.

Challenges encountered during the research are discussed, including limitations in data handling and the need for better regulatory feedback. The chapter proposes solutions to these challenges, such as improving data preprocessing techniques and enhancing communication with regulators. It emphasizes the benefits of the proposed machine learning and automation solutions in streamlining AML processes and improving compliance.

Chapter 5: Recommendations

The concluding chapter summarizes the main findings of the research, reinforcing the advantages of machine learning and automation in enhancing AML processes. It provides practical recommendations for implementing ML models and automation in AML systems, with a focus

on utilizing Microsoft Fabric for improved data management. The chapter also outlines future research directions, suggesting areas for further exploration to address remaining gaps, such as reducing ML model training times and enhancing model explainability.

Chapter 6: Conclusion

This chapter revisits the key research questions and objectives that guided the study, providing a comprehensive summary of the main findings and conclusions drawn from the analysis. The chapter emphasizes the significant contributions that the research has made to the understanding of detecting potential money laundering activities through advanced data analytics. Additionally, it highlights the broader implications of the work for the field of Anti-Money Laundering (AML), demonstrating how the study's insights can inform and improve current practices and strategies.

Chapter 2: Literature Review

2.1 Literature Review

Overview of the literature review completed, key findings, and how they relate to the thesis.

Author(s)	Year	Title	Key Highlights
Nazanin et al.	2022	A Survey of Machine Learning Based Anti-Money Laundering Solutions	Adaptability of ML in ever changing Money Laundering methods
Salvatore et al	2024	Anomaly Detection in Cross-Country Money Transfer Temporal Networks	Examining network relationships
Ahmad et al	2022	Anti-Money Laundering Alert Optimization Using Machine Learning with Graphs	machine learning (ML) model to improve anti-money laundering

2.1.1 A Survey of Machine Learning Based Anti-Money Laundering Solutions, October 2022

As per Nazanin et al., the exponential growth of global financial transactions has outpaced traditional financial infrastructures, making it difficult to accurately detect money laundering, which involves disguising the origins of illegally obtained funds. Machine learning (ML) offers

promising solutions for identifying suspicious transactions. This survey examines various ML-based anti-money laundering (AML) solutions.

Traditional rule-based AML systems, while structured and hierarchical, are inadequate due to their reliance on predefined rules that require constant updates to counteract evolving laundering strategies. These systems also struggle with semi-structured and unstructured data. In contrast, ML can recognize and apply patterns learned during training, handling new scenarios without needing explicit programming for each possibility.

AML processes target three phases of money laundering: Placement, Layering, and Integration, each requiring robust techniques to manage large volumes of complex data. The main challenges in AML include minimizing false negatives and false positives. Effective data preparation, which involves profiling and cleansing data to ensure quality, is crucial. ML models, such as Decision Trees, Random Forests, and Neural Networks, play a key role in detecting suspicious activities, although they require significant training time and effort.

Studies have shown that ML techniques like Support Vector Machines (SVM) and Random Forests outperform traditional methods, reducing false positives and improving detection accuracy. Hybrid models that combine clustering, neural networks, and heuristic approaches offer enhanced AML capabilities. However, most commercial AML products still rely heavily on rule-based methods, with only a few incorporating advanced AI to improve performance.

In conclusion, while ML-based AML solutions show immense potential, they require comprehensive retraining with new data, and the learning process remains time-consuming. Future research should focus on reducing training times and improving the practical application of ML models in AML systems.

2.1.2 Anomaly Detection in Cross-Country Money Transfer Temporal Networks, February 20, 2024

This study done by Salvatore et al., explores anomaly detection in cross-country money transfers using temporal network analysis. By examining the evolving structures and relationships within these networks, the method identifies abrupt shifts in node roles, triggering further expert investigation. Analyzing a dataset of 80 million wire transfers, the approach automates Anti-Financial Crime (AFC) and Anti-Money Laundering (AML) processes, offering a comprehensive top-down view that addresses limitations of current paradigms.

Traditional anti-fraud systems, which are rule-based and rely on human reasoning, often fall short due to their static nature and lack of interpretability. This study proposes an anomaly detection system that ranks observations based on deviations from the norm, prioritizing outliers to enhance the likelihood of identifying true anomalies. The approach assumes domain experts will report and justify suspicious findings to authorities, emphasizing explainable AI to support expert evaluations.

The methodology frames anomaly detection as a ranked information retrieval problem, focusing on producing interpretable results that help experts filter noise and uncover insights. Centrality-based node rankings are used to detect anomalies, with the system shown to effectively identify relevant nodes without prior domain knowledge.

In conclusion, the proposed method filters out normal behaviors, leaving a focused set of data points for expert analysis. It does not predetermine the number of anomalies or specific

behaviors to look for, making it adaptable and efficient. The approach enhances anomaly detection by providing clear, actionable insights to support anti-money laundering efforts.

2.1.3 Anti-Money Laundering Alert Optimization Using Machine Learning with Graphs, June 17, 2022

This study conducted by Ahmad et al., proposes a machine learning (ML) model to improve anti-money laundering (AML) systems by significantly reducing false positives while maintaining high true positive detection rates. Given the impracticality of manually reviewing all transactions, banks heavily rely on AML systems, which often produce high false positive rates (95-98%). The proposed ML model aims to mitigate this issue by complementing existing rule-based systems.

The methodology involves a triage classifier that identifies suspicious transactions and reduces the number of false positives generated by rule-based systems. Profile features are utilized to characterize transaction histories for each account, aggregated over different time windows. Graph neighborhood features represent accounts as nodes and transactions as edges, with a sliding window limiting the number of events analyzed. Additionally, degree features measure node connections, while guilty walker features identify suspicious nodes based on their connections.

Results from testing the model on a real-world banking dataset show significant improvements in reducing false positives. Incorporating various features, such as neighborhood and degree features, enhances the model's performance, with the best results achieved using a one-day

sliding window. These findings underscore the potential of ML approaches to enhance AML systems' efficiency and accuracy while maintaining compliance and explainability.

2.2 Gaps or Limitations in the Current Literature

Dependence on Traditional Rule-Based Systems

Traditional Anti-Money Laundering (AML) systems largely rely on rule-based methods that are static and require frequent updates to keep up with evolving money laundering techniques. These systems struggle with semi-structured and unstructured data, often leading to high false positive rates.

There is a clear need for more dynamic and adaptable solutions that can efficiently process diverse data types and minimize false positives without necessitating constant manual adjustments.

Complexity and Training Time of ML Models

While machine learning models like Decision Trees, Random Forests, and Neural Networks show promise in AML, they demand significant training time and effort. These models require comprehensive retraining with new data, which can be both time-consuming and resource-intensive.

Further research is needed to reduce the training times of machine learning models and enhance their practical applicability in real-world AML systems.

Limited Integration of Explainable AI

The importance of explainable AI is well-documented, particularly in supporting expert evaluations and justifying suspicious findings to authorities. However, many machine learning models lack transparency, making it challenging for domain experts to interpret and trust the results.

There is a pressing need for the development of machine learning models that not only enhance detection rates but also provide interpretable and actionable insights for AML professionals.

Information Asymmetry and Lack of Regulator Feedback

The lack of feedback from regulators poses a significant challenge, undermining the predictive capabilities of machine learning models and complicating the assessment of accuracy regarding false positives and false negatives.

Strategies are required to address information asymmetry and improve communication between financial institutions and regulators to enhance the effectiveness of AML models.

Data Complexity and Handling Challenges

Existing studies indicate that the complexity of data—such as the presence of multiple transaction tables—and the limitations of data handling tools like Power BI can hinder the performance and scalability of AML systems.

There is a need for research focused on developing efficient data handling and preprocessing techniques that can manage complex datasets and integrate seamlessly with machine learning models.

2.3 How This Research Will Address These Gaps

2.3.1 Developing a Dynamic Anomaly Checking Model

- **Objective:** Create a machine learning model to identify real-time anomalies in financial transactions.
- **Addressing the Gap:** By focusing on developing a dynamic anomaly detection model, this research aims to provide a more adaptable and responsive AML solution that can handle various data types and reduce the dependency on static rule-based systems.

2.3.2 Streamlining and Automating AML Processes

- **Objective:** Automate delivery operations to improve efficiency and productivity.
- **Addressing the Gap:** Automation will reduce manual workload, lower operational costs, and enhance the speed of AML processes, addressing the inefficiencies of traditional systems.

2.3.3 Enhancing Model Performance and Practical Application

- **Objective:** Evaluate the performance of the AML model using relevant metrics and benchmarks.
- **Addressing the Gap:** This research will focus on improving the practical application of ML models by reducing training times and ensuring they can be effectively integrated into real-world AML systems.

2.3.4 Incorporating Explainable AI

- **Objective:** Provide insights and recommendations to stakeholders based on the results of the AML model.

- **Addressing the Gap:** By emphasizing explainable AI, the research will ensure that the ML models provide interpretable results, supporting domain experts in their evaluations and decision-making processes.

2.3.5 Mitigating Information Asymmetry

- **Objective:** Develop strategies to enhance communication and feedback loops between Buffetti and regulatory bodies.
- **Addressing the Gap:** The research will explore ways to improve the interaction between financial institutions and regulators, aiming to reduce information asymmetry and improve the accuracy of AML models.

2.3.6 Optimizing Data Handling and Integration

- **Objective:** Develop efficient data handling techniques to manage complex datasets.
- **Addressing the Gap:** By creating strategies for better data preprocessing and integration, this research will tackle the challenges posed by data complexity and limitations in current data handling tools.

This research aims to bridge the gaps identified in the current literature by developing a dynamic, efficient, and explainable AML solution that leverages machine learning for real-time anomaly detection. By addressing the limitations of traditional rule-based systems, reducing training times, incorporating explainable AI, improving communication with regulators, and optimizing data handling, the research will enhance Buffetti's AML compliance capabilities and operational efficiency.

Chapter 3: Methodology

3.1 Research Design

This study adopts an observational research design, focusing on descriptive analytics to explore the intricate world of financial transactions. The primary goal is to observe and analyze existing transaction data to uncover potential money laundering activities without manipulating the data or altering any experimental conditions. Given the complex nature of financial transactions, the study is inherently exploratory, seeking to identify patterns and anomalies that could inform more robust Anti-Money Laundering (AML) strategies.

3.2 Data Collection Methods

The foundation of this research lies in transactional data sourced directly from Buffetti with a vast amount of information spread across 13 distinct transaction tables. These tables encapsulate various facets of financial transactions. The data was initially extracted in CSV format, a process that required meticulous pre-processing to ensure consistency and completeness. To focus the analysis, key columns essential for AML compliance checks were identified, streamlining the dataset to ten crucial fields. The attributes were clearly defined and extracted from the raw data before it was utilized in Power BI for further analysis. The required attributes are:

3.2.1 Location of Transaction

The location of the transaction provides context regarding where the transaction was conducted. This is typically indicated by the AgencyId, which may be referred to as 'ExternalAgencyId' in

the 'Sintaplus transactions' table. Accurate identification of this column is crucial for understanding the geographic and institutional context of the transaction. This information must be consolidated to ensure that all transactions are correctly mapped to their respective locations before entering Power BI.

3.2.2 Transaction Parties

This attribute identifies the entities involved in the transaction, such as the sender and receiver. Columns such as CustomerId or Id should be examined to ascertain the parties involved. Proper identification of these parties is essential for tracing the flow of funds and understanding the relationships between different entities. Accurate consolidation of this data helps in creating a comprehensive view of transaction parties for AML analysis in Power BI.

3.2.3 Custody of Owners

This attribute specifies the custodians of the funds, including bank accounts associated with customers and businesses. Identifying columns that detail these custodians is important for assessing the control and ownership of the funds. This information must be consolidated to determine whether funds are being moved between accounts in a manner consistent with legitimate activities before the data is used in Power BI. Theoretically there could be two different custody accounts in each transaction

3.2.4 Purpose of Transfer

Understanding the purpose of each transaction provides insights into the reason behind the fund transfer. Columns such as 'TypeOfBullettinValue' should be analyzed to categorize transaction purposes. Knowledge of specific terms like Mav, Freccia, and Rav is necessary for correctly

interpreting the purpose and context of each transaction. Accurate consolidation of this data ensures that the purpose of each transfer is clearly defined in Power BI.

3.2.5 Unique Transaction Key

A unique transaction key is vital for tracking and differentiating individual transactions. This is typically represented by a unique transaction number, such as 'RequestTransactionID'. Ensuring each transaction has a distinct key helps in accurately monitoring and auditing transaction activities. This key must be consolidated to provide a reliable reference for each transaction in Power BI.

3.2.6 Transaction Amount

The amount of the transaction is a straightforward but crucial piece of information. Accurate recording of the transaction amount is essential for assessing the scale of financial activities and detecting anomalies or suspicious patterns. This attribute needs to be consolidated to ensure accurate representation of transaction amounts in Power BI.

3.2.7 Transaction Date

The date of the transaction is important for temporal analysis of financial activities. Accurate recording of the transaction date helps in identifying trends, patterns, and any time-based anomalies in the transaction data. This attribute should be consolidated to allow for effective time-based analysis in Power BI.

3.2.8 Payment Method

Identifying the method of payment used in the transaction (e.g., bank transfer, online payment) is essential for understanding the mechanism of fund transfers. It is important to determine whether there is a consistent identifier for payment methods across different tables and to record this information accurately. Proper consolidation of payment method data ensures that all relevant payment details are available for analysis in Power BI.

By identifying and consolidating these attributes before utilizing the data in Power BI, the integrity and comprehensiveness of the AML analysis are ensured. This preparatory step is crucial for accurate and effective visualization and analysis of financial transactions, ultimately enhancing the ability to detect and prevent illicit financial activities.

Maintaining client confidentiality was paramount throughout this process. To safeguard sensitive information, the data was securely uploaded to Power BI, where all subsequent analyses were conducted within a controlled environment. Every step in the data handling process was carefully designed to prevent any potential breaches, reflecting the ethical commitment underlying the research.

3.3 Tools and Technologies Used

Power BI served as the primary tool for data visualization and preliminary data exploration. However, given the tool's limitations in handling large datasets, strategic data management techniques were employed, including the use of compressed data files and incremental data loading. This ensured that the analyses remained efficient and effective.

To further enhance the data processing capabilities, Python was brought into play. Python scripts were instrumental in data consolidation and anomaly detection, with the Isolation Forest algorithm—a machine learning model—deployed to identify outliers within the transaction data. As the research progressed, stakeholder meetings and additional research highlighted the potential of Microsoft Fabric as a future development platform. Microsoft Fabric was considered for its integration capabilities, cost-effectiveness, and support for complex data processing, which could further enhance the scalability of the AML system. Additionally, various Azure services, such as Azure Data Factory and Azure Synapse Analytics, were explored for building automated data pipelines and improving the system's scalability.

3.4 Data Analysis Techniques

Anomaly detection was a critical component of the data analysis phase. The Isolation Forest method, an unsupervised learning algorithm, was chosen for its efficiency in handling large datasets with complex structures. This algorithm excels at identifying outliers by isolating observations that deviate significantly from the majority, making it well-suited for detecting irregularities in financial transactions.

To strike a balance between detecting significant anomalies and minimizing the risk of false positives, a 5% threshold was established. This threshold allowed for the classification of transactions as anomalies while maintaining a rigorous approach to data analysis. Given the complexity of the data and the limitations of Power BI, additional steps were necessary to consolidate and pre-process the data before the anomaly detection could be effectively applied.

The results of the anomaly detection were then presented through a Power BI dashboard. This dashboard was designed to be accessible and user-friendly, providing stakeholders with a visual summary of the detected anomalies. It enabled them to quickly identify and investigate suspicious transactions, making the insights gained from the analysis actionable and relevant.

3.5 Justification for the Chosen Methodology

The observational research design was chosen due to the inherent nature of AML efforts, which require detecting and analyzing existing patterns in financial data rather than manipulating variables in an experimental setting. This design aligns perfectly with the study's objectives, allowing for a comprehensive examination of transaction records to identify potential money laundering activities.

To effectively manage and analyze extensive financial data for Anti-Money Laundering (AML) purposes, it was crucial to address the challenges posed by the original dataset's large size and redundancy. The dataset, initially 1 megabyte in size, contained repeated data across multiple tables, complicating processing and analysis. By identifying and focusing on key attributes—such as transaction location, parties involved, custody, purpose, unique identifiers, amount, date, and payment method—the dataset was consolidated and reduced to 5,000 kilobytes (5 megabytes). This streamlined data eliminated redundancy and improved processing efficiency and accuracy in Power BI, enhancing the effectiveness of AML analysis.

Anomaly detection was selected as the primary analytical method due to the unique challenges posed by the research context. Traditional predictive models were deemed less effective, particularly without regulatory feedback and the information asymmetry between regulators and

financial institutions. The Isolation Forest algorithm was particularly well-suited to this task, as it does not rely on labeled data and is highly effective in identifying outliers within complex datasets.

The choice of tools and technologies further supports the study's objectives. Power BI was selected for its robust data visualization capabilities, essential for presenting findings to stakeholders in a clear and actionable manner. Python, with its flexibility and power in data processing and machine learning, enabled the implementation of sophisticated algorithms like Isolation Forest. The consideration of Microsoft Fabric and Azure services reflects a forward-looking approach, aiming to enhance the scalability and efficiency of the AML system.

Finally, the methodology was carefully designed with a strong emphasis on ethical considerations, particularly maintaining data confidentiality and fairness. These measures ensure that the research complies with ethical standards, producing results that are both reliable and unbiased.

This methodology outlines a thoughtfully considered approach to detecting money laundering activities through observational research. The use of advanced data analysis techniques, supported by powerful tools and technologies, ensures that the research is not only rigorous but also highly relevant to the evolving needs of AML efforts.

Chapter 4: Results/Findings

4.1 Overview of the Analyzed Data

The research focused on six months of transaction data, which was condensed into a manageable size for analysis. The data comprised 13 different transaction tables, which were consolidated into a single dataset containing the 10 essential columns identified as crucial for Anti-Money Laundering (AML) compliance checks.

- **Number of Transactions Analyzed:** 116,082
- **Number of Agencies Involved:** 797
- **Number of Customers Involved:** 57376
- **Number of Anomalies Detected:** 5815 (5% threshold)

4.2 Anomaly Detection Results

The Isolation Forest algorithm was used to identify potential anomalies within the dataset. The following results were obtained:

The anomaly detection analysis revealed significant insights into the nature and scope of potentially suspicious transactions within the dataset. Out of a total of 116,082 transactions analyzed, 5,818 were identified as anomalies, representing 5% of the total transactions. This percentage indicates a non-negligible portion of the dataset that may warrant further scrutiny, highlighting the effectiveness of the anomaly detection methods applied.

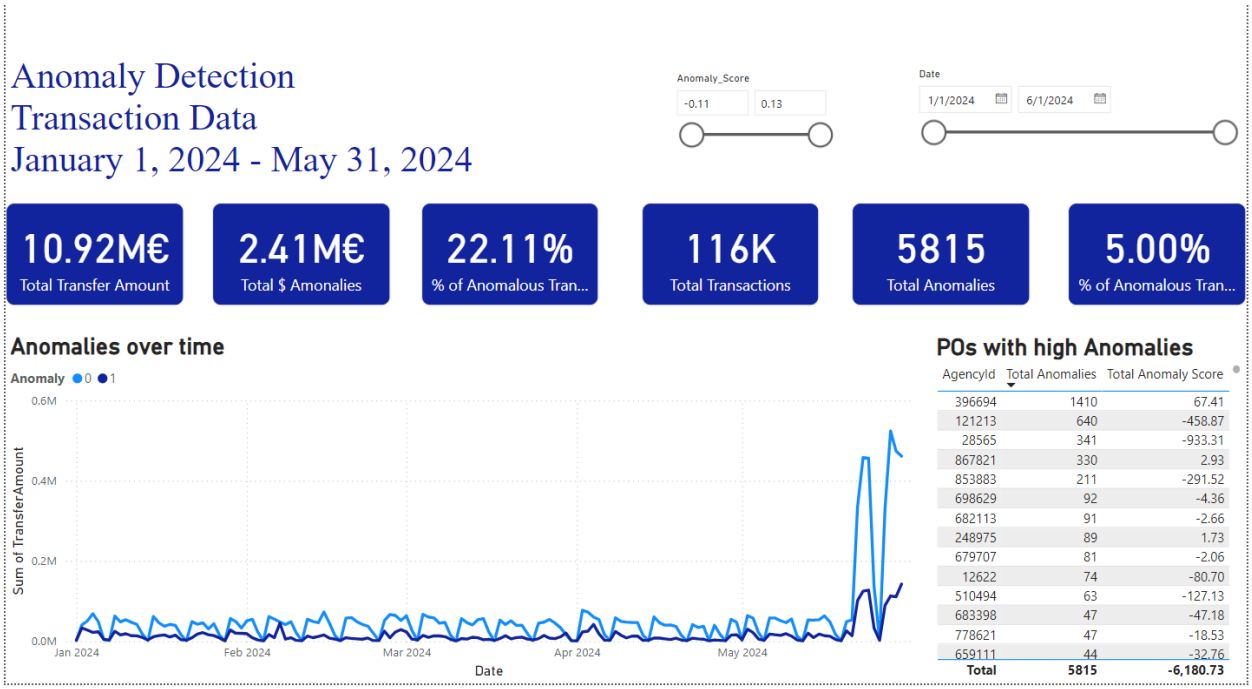
The financial impact of these anomalies is considerable. The total amount transferred across all transactions in the dataset was €10.92 million. However, the anomalies account for €2.41 million of this total, meaning that approximately 22% of the total transaction value is potentially suspicious. This disproportionate representation of anomalies in terms of monetary value suggests that anomalous transactions are not only frequent but also significant in size.

The fact that 5% of transactions are flagged as anomalies is consistent with typical thresholds used in anomaly detection models, such as the Isolation Forest method employed in this analysis. This method is particularly effective in identifying outliers in large datasets by isolating data points that differ significantly from the norm. The identification of 5,818 anomalies reflects the model's sensitivity to variations in transaction patterns, such as unusual transaction amounts, frequencies, or relationships between account holders.

Moreover, the €2.41 million associated with these anomalies underscores the potential risk posed by these transactions. Given that a quarter of the total transferred amount is linked to anomalies, this finding emphasizes the importance of investigating these transactions further to determine whether they represent legitimate activities or possible money laundering attempts. The high monetary value of these anomalous transactions could indicate sophisticated strategies aimed at obscuring the true nature of the funds being moved.

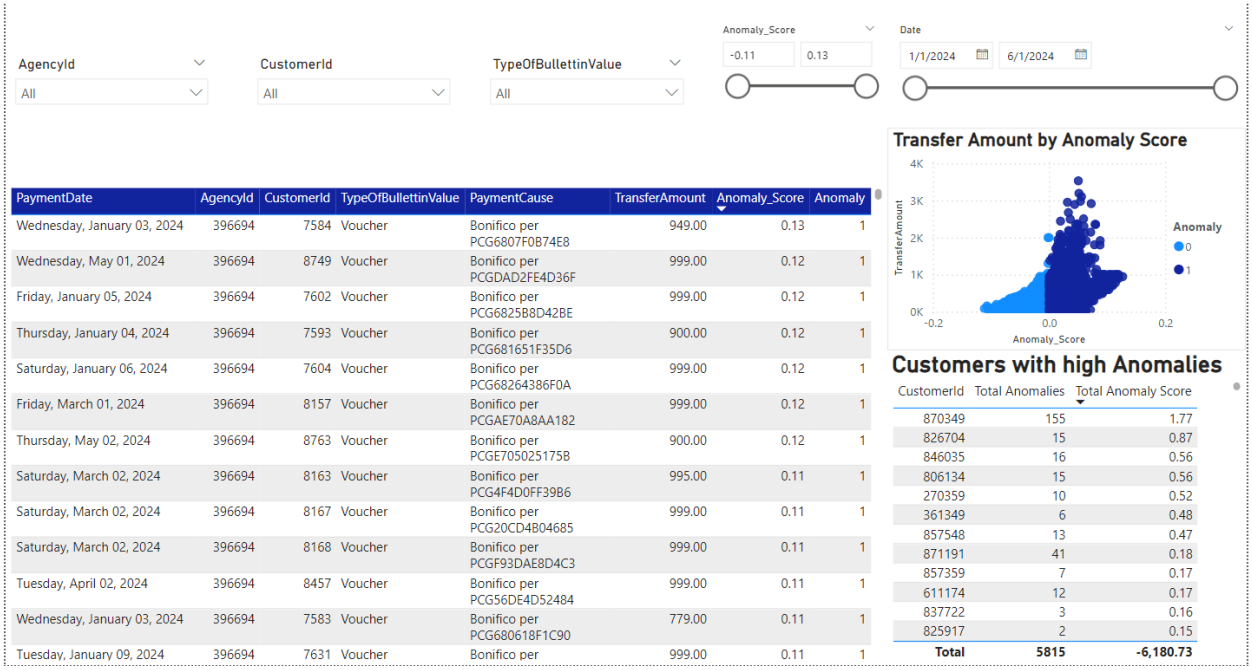
The anomaly detection results reveal a significant subset of transactions that are not only frequent but also financially substantial. The identification of these anomalies is a crucial step in the broader anti-money laundering (AML) efforts, as it helps to focus investigative resources on transactions that are more likely to be involved in illicit activities. Further analysis and validation of these anomalies are necessary to confirm their suspicious nature and to refine the detection models for even greater accuracy in future analyses.

Figure 4.1 Power BI showing summary tab ‘AML - Pillon v2.0’



The anomaly detection analysis uncovered several patterns that raise concerns about potential money laundering activities. When the transactions were sorted by anomaly score, a noteworthy trend emerged: numerous transactions were just slightly under €1,000. This pattern suggests that both clients and agencies might be deliberately circumventing hard rule restrictions, which often flag transactions above €1,000 for additional scrutiny. By keeping transactions just below this threshold, it appears that certain entities are attempting to avoid detection by automated compliance systems, which could indicate a sophisticated effort to launder money while evading regulatory oversight.

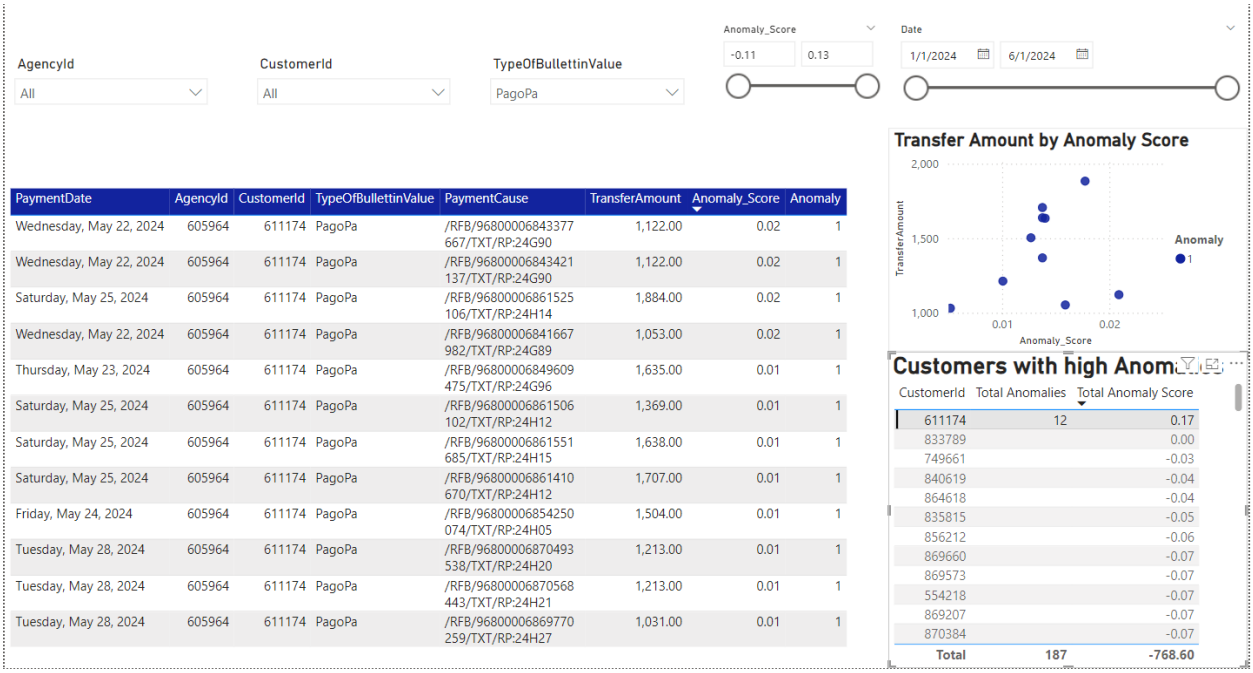
Figure 4.2 Power BI showing details tab ‘AML - Pillon v2.0’



Further analysis using the detailed tab focused on PagoPa transactions, revealing another suspicious pattern. One particular client exhibited 12 anomalies, all occurring at the same agency. This client's transaction history is especially concerning: between May 22, 2024, and May 28, 2024, nearly €15,000 was transferred, with no other transactions recorded for this client throughout the entire six months. The clustering of these anomalies within a short time frame, combined with the absence of other transactions, suggests a deliberate attempt to concentrate suspicious activity within a brief period, possibly to minimize the likelihood of detection. This behavior is atypical and warrants further investigation to determine whether these transactions are part of a larger money laundering scheme.

Figure 4.3 Power BI showing details tab with PagoPa selected under TypeOfBulletinValue selected and CustomerId 611174 selected under Customers with high Anomalies 'AML - Pillon

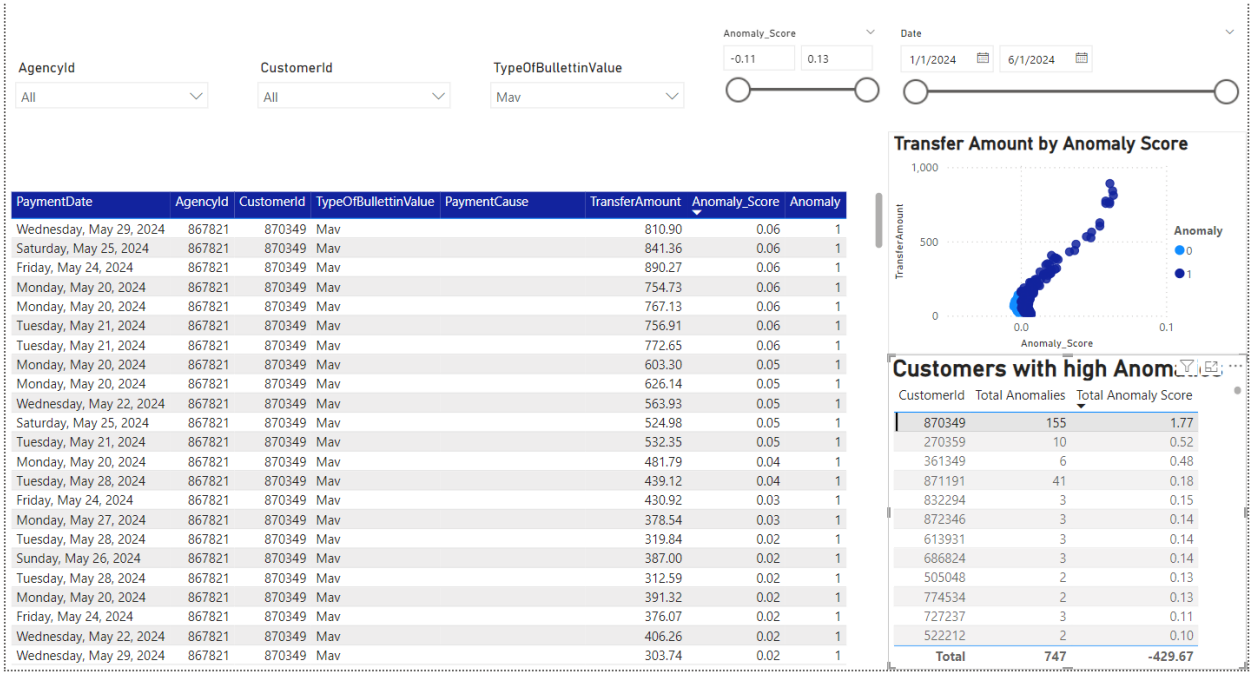
v2.0’



PaymentDate	AgencyId	CustomerId	TypeOfBullettinValue	PaymentCause	TransferAmount	Anomaly_Score	Anomaly
Wednesday, May 22, 2024	605964	611174	PagoPa	/RFB/968000068433777667/TXT/RP:24G90	1,122.00	0.02	1
Wednesday, May 22, 2024	605964	611174	PagoPa	/RFB/968000068434213137/TXT/RP:24G90	1,122.00	0.02	1
Saturday, May 25, 2024	605964	611174	PagoPa	/RFB/968000068615251106/TXT/RP:24H14	1,884.00	0.02	1
Wednesday, May 22, 2024	605964	611174	PagoPa	/RFB/968000068416671982/TXT/RP:24G89	1,053.00	0.02	1
Thursday, May 23, 2024	605964	611174	PagoPa	/RFB/968000068496094475/TXT/RP:24G96	1,635.00	0.01	1
Saturday, May 25, 2024	605964	611174	PagoPa	/RFB/968000068615061102/TXT/RP:24H12	1,369.00	0.01	1
Saturday, May 25, 2024	605964	611174	PagoPa	/RFB/968000068615511685/TXT/RP:24H15	1,638.00	0.01	1
Saturday, May 25, 2024	605964	611174	PagoPa	/RFB/968000068614106670/TXT/RP:24H12	1,707.00	0.01	1
Friday, May 24, 2024	605964	611174	PagoPa	/RFB/968000068542500074/TXT/RP:24H05	1,504.00	0.01	1
Tuesday, May 28, 2024	605964	611174	PagoPa	/RFB/968000068704935538/TXT/RP:24H20	1,213.00	0.01	1
Tuesday, May 28, 2024	605964	611174	PagoPa	/RFB/968000068705684443/TXT/RP:24H21	1,213.00	0.01	1
Tuesday, May 28, 2024	605964	611174	PagoPa	/RFB/968000068697702259/TXT/RP:24H27	1,031.00	0.01	1

Similarly, the analysis of Mav transactions revealed another red flag. A different client was associated with 155 anomalies within the six months, with several transactions processed daily, many of which were well above €500. The frequency and size of these transactions are unusual, particularly for Mav payments, which are typically smaller and less frequent. The consistent high value of these transactions, coupled with the large number of anomalies, suggests that this client could be engaged in layering, a common money laundering technique where illicit funds are moved through multiple transactions to obscure their origin. The daily processing of high-value transactions is a pattern that typically aligns with attempts to launder significant sums of money without drawing attention.

Figure 4.4 Power BI showing details tab with Mav selected under TypeOfBulletinValue selected and CustomerId 870349 selected under Customers with high Anomalies ‘AML - Pillon v2.0’



PaymentDate	AgencyId	CustomerId	TypeOfBullettinValue	PaymentCause	TransferAmount	Anomaly_Score	Anomaly
Wednesday, May 29, 2024	867821	870349	Mav		810.90	0.06	1
Saturday, May 25, 2024	867821	870349	Mav		841.36	0.06	1
Friday, May 24, 2024	867821	870349	Mav		890.27	0.06	1
Monday, May 20, 2024	867821	870349	Mav		754.73	0.06	1
Monday, May 20, 2024	867821	870349	Mav		767.13	0.06	1
Tuesday, May 21, 2024	867821	870349	Mav		756.91	0.06	1
Tuesday, May 21, 2024	867821	870349	Mav		772.65	0.06	1
Monday, May 20, 2024	867821	870349	Mav		603.30	0.05	1
Monday, May 20, 2024	867821	870349	Mav		626.14	0.05	1
Wednesday, May 22, 2024	867821	870349	Mav		563.93	0.05	1
Saturday, May 25, 2024	867821	870349	Mav		524.98	0.05	1
Tuesday, May 21, 2024	867821	870349	Mav		532.35	0.05	1
Monday, May 20, 2024	867821	870349	Mav		481.79	0.04	1
Tuesday, May 28, 2024	867821	870349	Mav		439.12	0.04	1
Friday, May 24, 2024	867821	870349	Mav		430.92	0.03	1
Monday, May 27, 2024	867821	870349	Mav		378.54	0.03	1
Tuesday, May 28, 2024	867821	870349	Mav		319.84	0.02	1
Sunday, May 26, 2024	867821	870349	Mav		387.00	0.02	1
Tuesday, May 28, 2024	867821	870349	Mav		312.59	0.02	1
Monday, May 20, 2024	867821	870349	Mav		391.32	0.02	1
Friday, May 24, 2024	867821	870349	Mav		376.07	0.02	1
Wednesday, May 22, 2024	867821	870349	Mav		406.26	0.02	1
Wednesday, May 29, 2024	867821	870349	Mav		303.74	0.02	1

These specific examples underscore the importance of not just relying on aggregate anomaly scores but also diving deeper into the transactional details to identify patterns that might indicate deliberate attempts to circumvent regulations. The detection of multiple anomalies concentrated around specific clients and agencies highlights the need for targeted investigations to determine whether these transactions are legitimate or part of illicit activities. The findings suggest that while the anomaly detection model is effective at flagging suspicious transactions, further manual review and investigative follow-up are critical for confirming the nature of these activities and ensuring compliance with anti-money laundering regulations.

Moreover, the display tool utilized in this analysis can serve as a powerful investigative resource. If an investigation is needed on a specific agency, the tool's functionality allows users to quickly isolate and examine all transactions associated with that agency. By selecting the agency from the dropdown menu, users can view transactions specifically ranked by anomaly score, enabling a focused investigation into potentially suspicious activities. This same application can be

extended to individual clients, allowing investigators to target specific accounts and explore their transactions in detail. This feature significantly enhances the tool's utility, making it not only a powerful detection mechanism but also a crucial component in the investigative process, providing AML professionals with the ability to efficiently prioritize and examine high-risk entities.

Chapter 5: Discussion

5.1 Interpretation of the significance in the findings

The findings of this analysis reveal both the challenges and potential solutions within the Anti-Money Laundering (AML) framework, emphasizing the need for a comprehensive, multi-faceted approach to effectively detect and prevent illicit financial activities.

5.1.1 Limited Scope in AML Processes

The focus on the placement stage of money laundering provides valuable insights but leaves out the more complex stages of layering and integration. This limited scope underscores the necessity for a broader analysis that encompasses all stages to capture the full extent of money laundering activities.

5.1.2 Regulatory Feedback Deficiency

The lack of feedback from regulators significantly hampers the predictive accuracy of machine learning models, making it challenging to distinguish between false positives and false negatives. This gap has led to a reliance on anomaly detection methods, but it also highlights the critical need for improved communication and collaboration with regulators to enhance model performance.

5.1.3 Information Asymmetry and Nash Equilibrium

The information asymmetry between financial institutions and regulators creates a Nash equilibrium, resulting in suboptimal AML outcomes. This equilibrium complicates efforts to

detect and prevent money laundering, suggesting a need for more balanced information sharing to improve the effectiveness of AML strategies.

5.1.4 Importance of Investigating Relationships and Transactions

Focusing on relationships between account owners and identifying excessive transactions within these relationships offers a strategic approach to uncovering layering strategies in money laundering. This method could reveal hidden patterns and connections, providing deeper insights into illicit financial activities.

5.1.5 Data Complexity and System Constraints

The complex data, spread across 13 transaction tables and limited to six months at a time in PowerBI, restricts the ability to perform comprehensive long-term analyses. Addressing these constraints is crucial for conducting more effective and wide-ranging AML investigations.

5.1.6 Challenges in Anomaly Detection

Implementing anomaly detection methods like Isolation Forest is straightforward, but the complexity of the data and system limitations require additional steps in data handling. This process is resource-intensive, underscoring the need for more efficient tools and techniques.

5.1.7 Data Confidentiality and Processing

The necessity of maintaining client confidentiality led to a complex data processing workflow, which, while necessary, added layers of complexity. However, successful data compression from 1 GB to 5 KB demonstrates the potential for more extensive and efficient data analysis.

5.1.8 Identification of Essential Data Columns

Identifying ten essential columns for compliance checks was a key achievement, ensuring that all critical aspects of transactions were included in the analysis. This step was vital for maintaining data integrity and improving the effectiveness of AML efforts.

5.1.9 Data Consolidation for Anomaly Detection

The development of a Python script to consolidate data ensured consistency and integrity, providing a solid foundation for anomaly detection. This streamlined the process, making the data ready for further analysis and potential automation in future workflows.

5.1.10 Effectiveness of the Isolation Forest Model

The Isolation Forest model effectively identified anomalies within the dataset, although future efforts should include better stakeholder communication about the model's mathematical operations to enhance understanding and trust.

5.1.11 Consideration of Azure Fabric

The exploration of Azure Fabric, following discussions with industry experts, revealed its potential advantages in terms of simplicity and cost-effectiveness. This consideration aligns with the project’s goals to optimize development processes and reduce costs while enhancing functionality.

5.1.12 Exploration of Microsoft Fabric for Data Management

Microsoft Fabric offers promising solutions to several key challenges in data management and analysis. By unifying organizational data into a single repository through OneLake, it simplifies data management and enhances analytics capabilities. Power BI's performance is notably

improved with Direct Lake, reducing the time needed to generate insights without the need for data imports. Fabric's Data Factory automates data processes, reducing manual effort and improving the efficiency of data pipelines. The platform’s support for incremental updates and seamless data integration further enhances its suitability for complex, dynamic datasets, making it a strong candidate for future projects.

5.2 Analysis of Results in Relation to the Research Question and Objectives

Primary Research Question

The research sought to determine how automation and machine learning could enhance Buffetti's AML processes, specifically compared to the manual execution via Excel files. The findings confirm that automation, particularly through machine learning models like the Isolation Forest, significantly improves both the efficiency and accuracy of Buffetti's AML operations. The deployment of machine learning algorithms allowed for the processing and analysis of large datasets, which would be unfeasible with manual methods due to time and resource constraints.

5.2.1 New Perspectives through Anomaly Detection

The study did not have access to specific data on the accuracy or time impact of Buffetti's current manual AML processes. As such, the existing manual processes were assumed to be limited in their ability to detect all potential issues due to their reliance on traditional methods, like Excel and hard rules, which may not capture the full complexity of money laundering activities. The introduction of anomaly detection through machine learning offers a fresh perspective,

potentially uncovering issues previously unnoticed. This approach not only enhances the detection of suspicious transactions but also serves as a valuable investigative tool, helping to identify patterns and irregularities that could indicate more sophisticated money laundering schemes. By providing a deeper level of insight, anomaly detection is expected to shed light on previously unknown issues, thereby improving the overall effectiveness of AML efforts.

5.2.2 Machine Learning Accuracy

Machine learning models, specifically the Isolation Forest algorithm, were more reliable and accurate in detecting anomalies compared to manual methods. The algorithm effectively identified suspicious transactions, representing a significant improvement over manual scrutiny, which often missed subtle patterns indicative of money laundering.

5.2.3 Resource Allocation to Manual Processes

The manual AML process is time-consuming and requires significant human resources. By contrast, automation can reduce the need for extensive manual labor, allowing staff to focus on higher-level compliance activities rather than data entry and manual analysis.

5.2.4 Cost Savings and Efficiency Gains

Automation projected significant cost savings and efficiency gains by reducing the time required to complete AML tasks and minimizing errors that could lead to regulatory fines. These gains not only streamline operations but also enhance the ability to meet regulatory requirements more effectively. Implementing automation effectively addressed these bottlenecks, speeding up the compliance process and reducing delays. The use of Power BI and Python allowed for more

efficient data processing and anomaly detection, thus enhancing the speed and accuracy of compliance activities.

5.2.5 Bottlenecks in Manual AML Processes

The manual AML processes were identified as having several bottlenecks, primarily due to the limitations of Excel in handling large datasets. These bottlenecks contributed to delays in compliance activities, further increasing the risk of non-compliance.

5.2.6 Types of Anomalies Indicative of Money Laundering

The study identified that anomalies, such as transactions slightly below regulatory thresholds or high-frequency, high-value transactions, are strong indicators of possible money laundering activities. The Machine learning model was particularly adept at recognizing these patterns.

5.2.7 Impact of Data Complexity on Anomaly Detection

The complexity of Buffetti's data, spread across 13 different transaction tables, posed challenges to anomaly detection. However, machine learning algorithms proved capable of handling this complexity, although additional data handling and preprocessing steps were necessary to ensure accuracy. It was later discovered and shared with the stakeholders that by utilizing Microsoft Fabric a unified scaleable pipeline can be developed for real time detection.

5.2.8 Challenges of Data Complexity and PowerBI Limitations

The data complexity and limitations of PowerBI, particularly in handling large datasets, necessitated careful data management. The study had overcome these challenges by condensing and preprocessing data, enabling effective analysis within PowerBI's constraints.

5.2.9 Data Consolidation and Pre-Processing

Data consolidation and preprocessing were crucial steps that facilitated effective machine learning implementation. The identification of essential columns and the development of Python scripts for data consolidation ensured that the analysis was both comprehensive and accurate.

5.2.10 Regulator Feedback Absence

The absence of regulator feedback was found to inhibit the predictive accuracy of AML models. This lack of communication led to a reliance on anomaly detection methods, which, while effective, could benefit from more precise input to reduce false positives and negatives.

5.2.11 Mitigating Information Asymmetry

The research highlighted the need for strategies to mitigate the information asymmetry between regulators and Buffetti. Suggestions include improving communication channels and sharing insights from anomaly detection with regulators to enhance the overall effectiveness of AML efforts.

5.2.12 Patterns in the Placement Stage

Significant patterns were identified in the placement stage of money laundering, particularly in terms of transaction amounts and frequencies. These insights could be expanded to cover the layering and integration stages, providing a more holistic view of money laundering activities.

5.3 Research Objectives

5.3.1 Develop an Anomaly Checking Model

The study successfully developed and implemented an anomaly detection model using the Isolation Forest algorithm, which effectively identified suspicious transactions.

5.3.2 Automate Delivery Operations

Although not developed but thoroughly researched, automation can be achieved through the integration of Microsoft Fabric and building pipelines that contain the machine learning algorithms and Python scripts, while streamlining the AML processes on Power BI and reducing the need for manual intervention.

5.3.3 Evaluate Model Performance

Despite taking into consideration and exploring the model's performance there was lack of feedback from executives and regulators that made it difficult to rigorously evaluate its effectiveness using statistical methods such as a confusion matrix. Without the ability to validate outcomes, the model's success is primarily demonstrated by its ability to process large volumes of data and highlight potential anomalies for further investigation. It is suggested that the model could add value, with enhanced data processing, identification of suspicious activity,

investigation support, adaptability to evolving threats, data driven decision making, and potential for automation. The definitive conclusions on its accuracy and overall performance remain inconclusive.

5.3.4 Provide Insights and Recommendations

The research provided actionable insights and recommendations for improving Buffetti's AML processes, emphasizing the importance of further automation, better data management, and enhanced communication with regulators.

The most exciting recommendation would be Microsoft Fabric. Microsoft Fabric can significantly enhance the project in several ways, particularly by addressing data management, scalability, and the integration of machine learning into Buffetti's AML processes. Here's how Microsoft Fabric can help:

5.3.4.1 Unified Data Management with OneLake

- **Centralized Data Repository:** Microsoft Fabric's OneLake provides a unified data storage solution, allowing Buffetti to consolidate data from multiple sources, including the 13 different transaction tables. This centralization simplifies data management, ensuring consistency and reducing the complexity involved in handling disparate datasets.
- **Scalability and Flexibility:** OneLake combines the scalability of data lakes with the structured performance of data warehouses, offering a robust storage solution. This enables Buffetti to manage large volumes of transaction data efficiently, making it easier to scale the AML system as the volume of data grows.

5.3.4.2 Enhanced Data Processing and Transformation

- **Automated Data Pipelines:** Fabric's Data Factory can automate data ingestion, transformation, and loading processes, which is crucial given the complexity and size of Buffetti's datasets. This automation reduces manual effort, minimizes errors, and ensures that data is processed and ready for analysis in a timely manner.
- **Advanced Data Transformation:** Data Flows Gen2, built on Power Query Online, offers sophisticated data transformation capabilities. This tool allows Buffetti to pre-process data more effectively, handling complex tasks such as cleaning, merging, and filtering data from various sources, which is essential before applying machine learning models.

5.3.4.3 Improved Performance with Direct Lake

- **Efficient Data Access:** Direct Lake in Microsoft Fabric enhances the performance of Power BI by enabling real-time access to data without needing to import it into the data model. This feature can significantly reduce the time required to generate insights and streamline the anomaly detection process, making it easier to work with large datasets.
- **Real-Time Analysis:** By bypassing the import process, Direct Lake allows for real-time updates and analysis, ensuring that the most current data is always used in AML efforts. This is particularly beneficial for identifying and responding to suspicious transactions as they occur.

5.3.4.4 Support for Machine Learning Integration

- **Integrated Machine Learning:** Microsoft Fabric supports the integration of machine learning models directly within the data environment. This means that Buffetti can

deploy and run anomaly detection models like Isolation Forest seamlessly within Fabric, reducing the need for external tools and simplifying the workflow.

- **Automated Model Deployment:** With Azure Machine Learning integration, models can be trained, deployed, and managed directly within the Fabric environment. This integration facilitates the continuous improvement of AML models as new data becomes available, ensuring that the system remains adaptive to evolving money laundering techniques.

5.3.4.5 Streamlined Collaboration and Accessibility:

- **Cross-Departmental Collaboration:** Fabric's platform enables easier collaboration across different departments within Buffetti, ensuring that AML efforts are coordinated and that insights are shared effectively. This enhances the collective ability to detect and respond to suspicious activities.
- **User Empowerment:** Fabric's tools are designed to be accessible to users at various skill levels, empowering more team members to engage with the data, run analyses, and contribute to AML processes without requiring deep technical expertise.

5.3.4.6 Cost-Effectiveness and Future-Proofing:

- **Cost Efficiency:** Fabric's pay-as-you-go model and the consolidation of multiple data and analytics tools into a single platform can reduce operational costs. This is particularly important for Buffetti, as it aims to enhance its AML processes without incurring excessive expenses.
- **Future-Proof Technology:** By adopting Microsoft Fabric, Buffetti positions itself to take advantage of ongoing advancements in data analytics and machine learning. The

platform's integration capabilities and regular updates ensure that the AML system remains cutting-edge and adaptable to future needs.

Microsoft Fabric can play a critical role in enhancing Buffetti's AML processes by providing a scalable, unified, and efficient data management and analytics platform. It supports the entire workflow—from data ingestion and transformation to machine learning model deployment and real-time analysis—making it a powerful tool for improving the efficiency and effectiveness of AML operations.

Also consider that Microsoft Fabric's capabilities extend beyond AML, offering significant benefits for other departments within Buffetti as well. Its centralized data management with OneLake can streamline data access and integration across various functions, such as finance, operations, and customer service. This unification enhances cross-departmental collaboration and data consistency, facilitating more informed decision-making and cohesive strategies.

For finance, Fabric’s automated data pipelines and advanced transformation tools can simplify budgeting, forecasting, and reporting, reducing manual effort and improving accuracy. In operations, the platform can optimize logistics and supply chain management by providing real-time insights and facilitating efficient data processing. Customer service can leverage Fabric to gain a comprehensive view of customer interactions and feedback, enhancing support and engagement strategies.

5.4 The implications of your findings for the field

5.4.1 Inadequacy of Rule-Based AML Systems

The findings confirm that traditional rule-based Anti-Money Laundering (AML) systems are insufficient for addressing modern money laundering techniques. These systems often struggle with handling complex, unstructured data and generate high false positive rates. This underscores the need for more dynamic and adaptive solutions, such as machine learning models, which can better identify anomalies and adapt to evolving patterns in financial transactions.

5.4.2 Importance of Explainable AI

The findings highlight the necessity of integrating explainable AI models in AML systems, especially when stakeholders may lack technical expertise. While machine learning models can enhance detection accuracy, their complexity often makes them opaque to users. Thus, developing models that offer clear, interpretable outputs is crucial for gaining stakeholder trust and ensuring actionable insights in regulatory contexts.

5.4.3 Role of Cloud Infrastructure

The research demonstrates that cloud infrastructure, such as Microsoft Fabric, is essential for managing and analyzing large volumes of data efficiently. Cloud solutions facilitate centralized data management, scalable storage, and advanced analytics capabilities, which are vital for processing complex AML data and supporting sophisticated machine learning techniques.

5.4.4 Optimizing Regulatory Strategies

The findings raise important questions about optimizing regulatory strategies in the face of information asymmetry. The observed Nash equilibrium, where both regulators and financial institutions face suboptimal conditions due to asymmetric information, suggests a need for

improved feedback mechanisms and collaborative strategies to enhance AML effectiveness and regulatory compliance.

5.4.5 Impact of Data Complexity

Finally, the research reveals that the complexity of a firm's data significantly influences project progress and effectiveness. Just as cloud infrastructure plays a critical role, the management of intricate data structures and integration challenges are pivotal in advancing AML systems. Effective data consolidation and preprocessing are crucial for leveraging advanced analytics and achieving meaningful insights.

5.5 Research Limitations

While this study demonstrates the potential of automation and machine learning in enhancing Buffetti's Anti-Money Laundering (AML) processes, several limitations should be acknowledged to contextualize the findings and their applicability. The research on automating Buffetti's AML processes using machine learning revealed several key limitations. The lack of data on the accuracy of existing manual processes, absence of stakeholder feedback, and reliance on anomaly detection without complementary methods hindered comprehensive model validation. Limitations in Power BI's handling of large datasets and the complexity of Buffetti's data further affected the analysis. Additionally, the study did not explore real-time implementation or the full money laundering lifecycle, while initial cost and resource requirements for automation were not accounted for, limiting the generalizability of findings.

5.5.1 Absence of Ground Truth for Manual Processes

A significant limitation of the research is the lack of access to specific data on the accuracy and time impact of Buffetti's current manual AML processes. Due to the unavailability of precise metrics, the study assumes that manual processes, relying on Excel and hard rules, are less effective at detecting sophisticated money laundering activities. While automation, particularly through machine learning models, has shown improvements in efficiency and accuracy, the absence of a direct, quantified comparison between manual and automated processes limits the ability to definitively assess the scale of these improvements.

5.5.2 Lack of Stakeholder Feedback

The absence of feedback from key stakeholders, including regulators and executives, further constrains the research. Without their input, the performance of the machine learning model, particularly its ability to reduce false positives and negatives, could not be rigorously validated. Although the model effectively identified anomalies in large datasets, the lack of real-world feedback limits the conclusions that can be drawn about its practical implementation in regulatory compliance scenarios.

5.5.3 Over-Reliance on Anomaly Detection

The research focuses primarily on anomaly detection techniques, specifically through the Isolation Forest algorithm, to identify suspicious transactions. While effective, this approach may not fully capture the complexity of money laundering schemes, which can involve patterns that are not easily detected by anomaly detection alone. Additionally, the model's reliance on anomaly detection, without complementary methods or regulatory feedback, introduces the risk

of both false positives and false negatives, which could affect the overall effectiveness of the system.

5.5.4 Limitations of Power BI and Microsoft Fabric

Although the study highlights the advantages of using Microsoft Fabric and Power BI for data management and machine learning integration, these tools have limitations. Power BI, in particular, struggles with large datasets, which necessitated data preprocessing and consolidation. This process, while necessary, may have introduced biases or resulted in the loss of potentially important information, impacting the model's performance. Furthermore, the practical implementation of Microsoft Fabric is still theoretical within the scope of this research, as it was not deployed in a live environment during this study.

5.5.5 Data Complexity and Preprocessing

The complexity of Buffetti's transactional data, spread across 13 different tables, presented challenges in terms of data handling and anomaly detection. While machine learning algorithms were able to manage this complexity after significant preprocessing, the necessity of consolidating and transforming the data could have influenced the results. Preprocessing may have introduced biases, simplified certain anomalies, or led to the exclusion of subtle patterns that could affect the model's overall accuracy.

5.5.6 Incomplete Model Validation

Due to the lack of statistical validation techniques such as confusion matrices, the evaluation of the machine learning model's performance is primarily based on its ability to handle large datasets and highlight potential anomalies. The absence of a more comprehensive validation

process means that the findings regarding the model's accuracy and effectiveness in detecting suspicious transactions remain inconclusive. Further testing, particularly with real-world data and stakeholder feedback, is necessary to fully assess the model's capabilities.

5.5.7 Generalizability of Findings

The findings of this research are specific to Buffetti's AML processes and the structure of its data. Other organizations, with different data structures, transaction volumes, or regulatory requirements, may face different challenges when implementing similar machine learning models. As such, the generalizability of the results to other contexts is limited.

5.5.8 Limited Focus on the Money Laundering Lifecycle

While the study successfully identified patterns in the placement stage of money laundering, the model's effectiveness in detecting activities during the layering and integration stages remains unexplored. Money laundering schemes often involve multiple stages, and this research focused primarily on one stage, limiting the understanding of the model's broader applicability across the entire lifecycle of money laundering.

5.5.9 Lack of Real-Time Implementation

The machine learning models developed in this research were not tested in a live, real-time environment. Without real-time deployment, it is challenging to evaluate the system's responsiveness to evolving money laundering schemes or its effectiveness in dynamic regulatory compliance processes. Further research is required to test the models under real-world conditions to fully assess their practical utility.

5.5.10 Initial Cost and Resource Constraints

While the study suggests significant cost savings and efficiency gains from automation, it does not account for the initial investment required to implement machine learning models, Microsoft Fabric, and Power BI at scale. Additionally, ongoing maintenance and updates of the system may require specialized expertise, which could introduce further resource constraints. These factors may affect the feasibility of widespread adoption within the organization or similar firms.

5.6 Future Research Direction:

Optimizing Regulatory Strategies for the Commons: Nash Equilibrium in Asymmetric Information Games for Anti-Money Laundering

5.6.1 Abstract

This research proposal aims to address the challenge of asymmetric information in anti-money laundering (AML) scenarios through the application of game theory, specifically by solving for a Nash equilibrium. The discrepancy in information between financial institutions and regulators often leads to a suboptimal equilibrium with significant social costs. This study will develop a game-theoretic model to identify and analyze the Nash equilibrium in AML contexts characterized by asymmetric information. The objectives are to quantify the social costs associated with the current equilibrium and propose strategies to mitigate these costs through improved regulatory feedback mechanisms.

Drawing on domain knowledge in data science, finance, and artificial intelligence, this research will employ analytical techniques to solve for the Nash equilibrium and assess its implications.

The significance of this study lies in its potential to enhance the effectiveness of AML efforts, inform more efficient regulatory practices, and provide broader insights into managing asymmetric information in other industries.

The preliminary work has highlighted the need for dynamic regulatory feedback to shift the equilibrium towards more socially optimal outcomes. The proposed timeline includes comprehensive literature review, model development, data collection and analysis, and the identification of practical interventions. This research aspires to contribute significantly to both the fields of game theory and AML, ultimately advancing the understanding and management of asymmetric information in critical economic systems.

5.6.2 Introduction

The global financial system faces significant challenges from money laundering activities, which threaten economic stability and security. Anti-money laundering (AML) efforts are crucial in combating these illicit activities. However, one of the persistent challenges in AML is the problem of asymmetric information, where financial institutions and regulators operate with different levels of information. This discrepancy creates a scenario similar to a Nash equilibrium, where both parties adopt strategies that lead to suboptimal outcomes with high social costs.

This research aims to explore the Nash equilibrium in AML scenarios characterized by asymmetric information and to develop strategies to mitigate the associated social costs. By addressing this equilibrium, the research seeks to improve the effectiveness of AML efforts and enhance the overall value of common resources.

5.6.3 Background Information

Background and context of anti-money laundering (AML):

In Italy, anti-money laundering (AML) regulations constitute a vital framework aimed at preventing the financial system from being exploited for illicit activities, including money laundering and terrorist financing. These regulations are anchored by Legislative Decree No. 231/2007, which aligns with European Union directives and international standards established by organizations such as the Financial Action Task Force (FATF). This legislative foundation mandates stringent measures for financial institutions, including banks, insurance companies, and other entities, to implement robust controls and procedures to detect and prevent money laundering activities.

The oversight and enforcement of AML compliance in Italy are primarily managed by the Financial Intelligence Unit (FIU), known locally as the Unità di Informazione Finanziaria (UIF), operating within the Bank of Italy. The UIF plays a pivotal role in receiving, analyzing, and disseminating reports of suspicious transactions submitted by reporting entities. These entities are obligated to conduct thorough customer due diligence (CDD) procedures, verifying the identity of clients, assessing associated risks, and monitoring transactions for suspicious activities. Enhanced due diligence is required for clients deemed high-risk.

Central to Italy's AML regime are reporting requirements mandating financial institutions to file suspicious transaction reports (STRs) to the UIF upon encountering transactions or activities indicative of potential money laundering or terrorist financing. These reports serve as crucial triggers for investigations conducted by law enforcement authorities to combat financial crime effectively.

Non-compliance with AML regulations in Italy carries significant penalties, including substantial fines and administrative sanctions. The UIF conducts regular inspections and audits to ensure that financial institutions maintain effective AML controls and adhere to regulatory requirements. Italy also actively engages in international cooperation efforts, including mutual legal assistance treaties (MLATs) and collaborations with international organizations like the FATF, to exchange financial intelligence and enhance global AML standards and practices.

Challenges faced by Italy in combating money laundering include the increasing complexity of financial transactions facilitated by digital technologies. To address these challenges, Italy continues to innovate by adopting advanced analytics, artificial intelligence, and blockchain technology to enhance transaction monitoring capabilities and strengthen its AML framework.

In summary, Italy's AML framework underscores its commitment to safeguarding the integrity of its financial system and protecting against financial crime. The regulatory regime emphasizes compliance, transparency, and international collaboration to mitigate risks and uphold the integrity of financial transactions within Italy and across international borders. Continued adaptation to emerging threats and technological advancements remains crucial for the effectiveness of Italy's efforts in combating money laundering and ensuring financial stability and security.

5.6.4 Brief overview of Nash equilibrium and its relevance to AML

Nash equilibrium, a foundational concept in game theory, describes a scenario where each participant in a game makes decisions based on the actions of others, with no player able to improve their outcome by unilaterally changing their strategy. In the realm of anti-money

laundering (AML) and financial regulation, Nash equilibrium can be applied to understand the strategic interactions between regulators and financial institutions.

One critical aspect is the lack of timely and meaningful feedback from regulators to financial institutions regarding suspicious transactions and compliance efforts. Financial institutions rely on regulatory feedback to refine their AML strategies, improve detection capabilities, and enhance compliance frameworks. Without this feedback loop, financial institutions operate under uncertainty, potentially leading to suboptimal AML practices and increased compliance costs.

Financial institutions incur significant operational costs in maintaining robust AML programs, including investments in technology, staff training, and compliance resources. The absence of timely feedback disrupts the equilibrium where regulators and financial institutions could otherwise cooperate more effectively to combat money laundering and financial crime. Regulators play a pivotal role in providing guidance, clarifying regulatory expectations, and sharing intelligence that informs financial institutions' AML strategies.

In Nash equilibrium terms, the lack of feedback creates a situation where financial institutions make decisions based on incomplete information about regulatory expectations and enforcement outcomes. This informational asymmetry affects strategic behaviors and compliance efforts, potentially resulting in inefficiencies and higher costs for financial institutions aiming to meet regulatory requirements.

Moreover, the dynamic nature of AML regulations and enforcement actions introduces further complexities. Changes in regulatory policies, new technological advancements, and evolving criminal tactics continuously reshape the AML landscape. Nash equilibrium analysis helps illuminate how these factors influence the strategic interactions between regulators and financial

institutions, highlighting the importance of effective communication and feedback mechanisms in achieving mutual compliance objectives.

To summarize, Nash equilibrium provides a theoretical lens to explore the strategic dynamics between regulators and financial institutions in the context of AML. It underscores the significance of timely feedback from regulators in fostering cooperation, improving compliance outcomes, and reducing operational costs for financial institutions engaged in the fight against money laundering and financial crime. Enhancing feedback mechanisms can potentially lead to more effective AML practices and a stronger overall regulatory framework in safeguarding the integrity of the financial system.

5.6.5 Research Problem

In the realm of anti-money laundering (AML), financial institutions and regulatory bodies engage in a continuous struggle to detect and prevent illicit financial activities. A significant challenge in this context is the problem of asymmetric information, where financial institutions have access to detailed transaction data, while regulators often rely on reports and disclosures that may lack granularity and immediacy. This disparity in information availability and quality creates a situation analogous to a game-theoretic scenario, where each party's actions are interdependent and strategic.

The core of this research problem lies in the formation of a Nash equilibrium under conditions of asymmetric information. Financial institutions may adopt strategies that minimize their compliance costs while meeting regulatory requirements, whereas regulators aim to maximize detection and deterrence of money laundering with limited information. This strategic interplay

often results in an equilibrium where neither party has the incentive to change their strategy, despite the presence of high social costs, such as the proliferation of undetected money laundering activities and the inefficient allocation of resources towards compliance and enforcement.

Current AML practices do not adequately address the strategic nature of interactions between regulators and financial institutions, nor do they sufficiently quantify the social costs associated with the resulting Nash equilibrium. This gap in understanding and methodology prevents the development of more effective and cooperative approaches to AML.

The research problem, therefore, is to:

1. Develop a game-theoretic model that accurately represents the strategic interactions between financial institutions and regulators in an AML context with asymmetric information.
2. Solve for the Nash equilibrium in this model to understand the strategic behaviors and outcomes of both parties.
3. Quantify the social costs associated with the equilibrium, highlighting the inefficiencies and potential harms to society.
4. Propose and evaluate strategies for reducing these social costs, potentially through improved feedback mechanisms and regulatory interventions that promote more optimal behavior and outcomes.

By addressing this problem, the research aims to provide a deeper understanding of the dynamics at play in AML efforts and to suggest practical solutions that enhance the effectiveness of these efforts while reducing the overall social cost.

5.6.6 Objectives

The primary objectives of this research are to:

1. To develop a game-theoretic model that identifies and addresses the Nash equilibrium in AML scenarios with asymmetric information.
2. To quantify the social costs associated with the current equilibrium in AML practices.
3. To propose strategies for reducing these social costs through improved regulatory feedback mechanisms and other interventions.

5.6.7 Literature Review

Remarks on literature review

The study of Nash equilibrium in games with asymmetric information has been well-documented in economic and game theory literature, with seminal works by Akerlof (1970) on market mechanisms under quality uncertainty, and by Spence (1973) and Stiglitz (1975) on signaling and screening in labor markets. Nash's (1950) foundational concept of equilibrium points in n-person games provides a critical underpinning for understanding strategic interactions in scenarios where players have incomplete information about each other.

However, the application of these principles to Anti-Money Laundering (AML) scenarios remains underexplored. Traditional AML methods, such as those highlighted by Baltoiu et al. (2019) and Bakhshinejad et al. (2022), primarily focus on the detection and reporting of

suspicious activities using machine learning and statistical analysis. While these approaches are effective in identifying anomalies, they often overlook the strategic behaviors and interactions between financial institutions and regulators.

Recent studies, such as those by Jayantilal et al. (2017) and Xi (2015), have started to explore the application of game theory to AML policies and network payment systems, respectively. These works demonstrate the potential of game-theoretic models to enhance our understanding of AML dynamics. Similarly, Dupuis (2021) discusses the cat-and-mouse game between money launderers and regulators in a Central Bank Digital Currency (CBDC) context, highlighting the strategic depth of AML challenges.

Moreover, research by Araujo (2010) and Luo & Gu (2018) extends the application of game theory to evolutionary and e-commerce environments, respectively. These studies suggest that game theory can provide valuable insights into the optimization of AML strategies, especially when considering the evolving tactics of money launderers.

Theoretical contributions by Hardin (1968) on the Tragedy of the Commons and by Kreps & Scheinkman (1983) on precommitment and competition further inform the understanding of strategic interactions in regulated environments. These concepts are particularly relevant to AML, where the common good—i.e., a stable and transparent financial system—can be undermined by individual incentives to engage in or overlook illicit activities.

Furthermore, empirical research on anomaly detection and machine learning, such as the works by Savage et al. (2016) and Zhao et al. (2024), provides a foundation for integrating advanced analytical techniques with game-theoretic models. These studies underscore the importance of

developing robust methods for detecting suspicious activities while accounting for the strategic adaptations of both financial institutions and criminals.

This research aims to fill the gap in the literature by applying game-theoretic principles specifically to the AML context. By modeling the interactions between financial institutions and regulators as a game with asymmetric information, this study seeks to provide a novel approach to understanding and addressing the challenges posed by such information asymmetries. The goal is to develop a framework that not only enhances the detection of money laundering activities but also incentivizes cooperation and compliance among all stakeholders, thereby reducing the overall social cost and improving the effectiveness of AML efforts.

5.6.8 Methodology

This research will utilize a game-theoretic approach to model the interactions between financial institutions and regulators in AML scenarios. The model will incorporate elements of asymmetric information and strategic decision-making. Data collection will involve obtaining financial transaction data and regulatory reports to inform the model. Analytical techniques will be employed to solve for the Nash equilibrium and to quantify the associated social costs. The research will also explore potential interventions and feedback mechanisms that could shift the equilibrium towards more optimal outcomes.

5.6.9 Expected Results

Hypotheses

In the context of Anti-Money Laundering (AML), the lack of feedback from regulators to financial institutions creates a Nash equilibrium characterized by suboptimal compliance efforts and increased social costs. By incorporating a structured feedback mechanism from regulators, financial institutions can adjust their AML strategies more effectively, leading to a reduction in money laundering activities and a decrease in overall social costs. This study hypothesizes that:

Primary Hypothesis: Implementing a structured feedback mechanism from regulators to financial institutions will disrupt the current Nash equilibrium, resulting in enhanced compliance efforts and reduced money laundering activities.

Secondary Hypothesis: Financial institutions that receive regular, detailed feedback from regulators will exhibit a significant decrease in the incidence of false positives and false negatives in their AML detection systems, leading to more efficient allocation of resources.

Tertiary Hypothesis: The introduction of feedback mechanisms will lower the social cost associated with AML by fostering greater transparency and cooperation between financial institutions and regulators, ultimately benefiting the common good of a stable and transparent financial system.

Initial Condition

- The current state of AML efforts involves asymmetric information between financial institutions and regulators.

- This asymmetry leads to a Nash equilibrium where both parties settle into strategies that are individually rational but collectively suboptimal.

Main Assertion

- The lack of feedback from regulators is a key factor contributing to this equilibrium.
- The equilibrium results in high social costs, including inefficient resource allocation and undetected money laundering activities.

Proposed Intervention

- Introducing enhanced feedback mechanisms from regulators to financial institutions.
- Such mechanisms would provide more timely, detailed, and actionable information.

Expected Outcome

- Improved feedback will encourage financial institutions to adopt better compliance strategies.
- This will lead to a more optimal equilibrium with lower social costs.
- Enhanced detection and prevention of money laundering activities.

5.6.10 Expected Findings

The findings of this research are expected to make significant contributions to both game theory and AML practices. By providing a deeper understanding of the Nash equilibrium in AML scenarios, the research will inform more effective regulatory strategies and policies. Additionally, the insights gained from this study could be applied to other industries facing

similar asymmetric information challenges, thereby enhancing the overall efficiency and equity of various economic systems.

5.6.11 Timeline

Year 1

Months 1-3: Literature Review and Framework Development

- Conduct a comprehensive review of existing literature on Nash equilibrium, asymmetric information, and anti-money laundering (AML).
- Identify gaps in the current research and refine the research problem.
- Develop a conceptual framework for the game-theoretic model.

Months 4-6: Model Development

- Formulate the game-theoretic model representing the strategic interactions between financial institutions and regulators.
- Define key parameters and variables, considering both theoretical and practical aspects of AML scenarios.

Months 7-9: Data Collection and Preliminary Analysis

- Collect relevant data from financial institutions and regulatory bodies.
- Conduct preliminary analysis to validate the assumptions of the model.
- Identify any data gaps and plan for additional data collection if necessary.

Months 10-12: Initial Model Testing

- Implement the initial version of the game-theoretic model.
- Conduct initial tests to solve for the Nash equilibrium under current AML conditions.
- Analyze preliminary results and refine the model as needed.

Year 2

Months 13-15: Detailed Analysis and Model Refinement

- Perform a detailed analysis of the Nash equilibrium and the associated social costs.
- Refine the model based on initial findings and incorporate any additional variables or constraints.

Months 16-18: Quantification of Social Costs

- Develop methods to quantify the social costs associated with the current Nash equilibrium.
- Use statistical and analytical techniques to measure these costs.

Months 19-21: Proposal of Interventions

- Identify and propose regulatory feedback mechanisms and other interventions to shift the Nash equilibrium towards more optimal outcomes.
- Develop a theoretical framework for these interventions.

Months 22-24: Simulation and Testing of Interventions

- Implement the proposed interventions within the game-theoretic model.

- Conduct simulations to test the impact of these interventions on the Nash equilibrium and social costs.
- Analyze the results and refine the interventions as needed.

Year 3

Months 25-27: Validation and Sensitivity Analysis

- Validate the refined model and interventions using additional data and real-world case studies.
- Conduct sensitivity analysis to assess the robustness of the model under various scenarios and assumptions.

Months 28-30: Final Model and Policy Recommendations

- Finalize the game-theoretic model and the proposed regulatory feedback mechanisms.
- Develop comprehensive policy recommendations based on the findings of the research.

Months 31-33: Thesis Writing

- Begin drafting the thesis, incorporating all research findings, analysis, and policy recommendations.
- Ensure the thesis is well-structured and thoroughly documented.

Months 34-36: Review and Defense Preparation

- Review and revise the thesis based on feedback from advisors and peers.
- Prepare for the thesis defense, including presentation and discussion of key findings and contributions.

Month 36: Thesis Submission and Defense

- Submit the final thesis to the academic committee.
- Defend the thesis in front of the academic panel.

5.6.12 Proposed Reference Abstracts and Literature

Akerlof, G. (1970). 'The market for lemons: Quality uncertainty and the market mechanism'. *Quarterly Journal of Economics*, 84(3), 488–500.

Araújo, R. (2010). *An evolutionary game theory approach to combat money laundering*. Journal of Money Laundering Control, January 2010.

Baltoiu, A., Patrascu, A., & Irofti, P. (2019). *Community-Level Anomaly Detection for Anti-Money Laundering*. University of Bucharest, Romania.

Bakhshinejad, N., Soltani, R., Nguyen, U. T., & Messina, P. (2022). *A Survey of Machine Learning Based Anti-Money Laundering Solutions*. Department of Electrical Engineering and Computer Science, York University.

Bertrand, J. L. F. (1883). 'Theorie Mathématique de la Richesse Sociale Par Léon Walras: Recherches Sur Les Principes Mathématiques de la Theorie Des Richesse Par Augustin Cournot'. *Journal des Savants*, 67, 499–508.

Campbell, D. (2018). Sections 3.1–3.4, 3.6–3.8 and 3.10 in *Incentives: Motivation and the Economics of Information* (3rd ed.). Cambridge, UK: Cambridge University Press.

Cournot, A. (1838). *Recherches sur les Principes Mathématiques de la Theorie des Richesses*. Paris, France: Hachette.

Dupuis, D. (2021). *Money Laundering in a CBDC World: A Game of Cats and Mice*. Social Science Research, February 25, 2021.

Eddin, A. N., Bono, J., Aparício, D., Polido, D., Ascensao, J. T., Bizarro, P., & Ribeiro, P. (2022). *Anti-Money Laundering Alert Optimization Using Machine Learning with Graphs*. University of Porto, Portugal.

Gibbons, R. (1992). *A Primer in Game Theory*. Harlow, UK: FT Prentice Hall.

Hardin, G. (1968). 'The Tragedy of the Commons'. *Science*, 162, 1243–1248.

Jayantilal, S., Jorge, S., & Ferreira, A. (2017). *Portuguese Anti-money Laundering Policy: A Game Theory Approach*. European Journal on Criminal Policy and Research, July 7, 2017.

Klemperer, P. (1999). 'Auction theory: A guide to the literature'. *Journal of Economic Surveys*, 13(3), 227–286.

Kreps, D., & Scheinkman, J. (1983). 'Quantity precommitment and Bertrand competition yield Cournot outcomes'. *Bell Journal of Economics*, 14(2), 326–337.

Luo, S., & Gu, B. (2018). *Research on Optimal Anti-money Laundering Model Based on E-Commerce Environment*. DEStech Transactions on Engineering and Technology Research.

Nash, J. F. (1950). 'Equilibrium points in n-person games'. *Proceedings of the National Academy of Sciences*, 36, 48–49.

Osborne, M., & Rubinstein, A. (1994). *A Course in Game Theory*. Oxford, UK: Oxford University Press.

Savage, D., Wang, Q., Chou, P., Zhang, X., & Yu, X. (2016). *Detection of Money Laundering Groups Using Supervised Learning in Networks*. University of Melbourne, Australia.

Spence, M. (1973). 'Job market signaling'. *Quarterly Journal of Economics*, 87(3), 355–374.

Stiglitz, J. (1975). 'The theory of 'screening', education, and the distribution of income'. *American Economic Review,* 65(3), 283–300.

Vilella, S., Capozzi Lupi, A. T. E., Fornasiero, M., Moncalvo, D., Ricci, V., Ronchiadin, S., & Ruffo, G. (2024). *Anomaly Detection in Cross-Country Money Transfer Temporal Networks*. DISIT, Universita degli Studi del Piemonte Orientale; Dipartimento di Informatica, Universita degli Studi di Torino; Anti Financial Crime Digital Hub Corso Inghilterra.

Varian, H. (1992). Chapter 17 'Exchanges'. *Microeconomic Analysis*.

Wen-jun, T. (2010). *The Theoretical Basis of Anti-money Laundering Supervision in Financial Industry*. Journal of Hunan Financial and Economic College.

Xi, Z. (2015). *Research on anti-money laundering game theory model and strategies of network payment industry players*. Systems Engineering - Theory and Practice.

Ya-mei, G. (2007). *A Game Analysis of Anti-Money Laundering Auditing*. Economic Survey.

Zhao, H., Zi, C., Liu, Y., Zhang, C., Zhou, Y., & Li, J. (2024). *Weakly Supervised Anomaly Detection via Knowledge-Data Alignment*. Hong Kong University of Science and Technology (Guangzhou); CreateLink Technology China.

Chapter 6: Conclusion

6.1 Restating Research Questions and Objectives

The journey of this research began with the question: **"How can automation and machine learning technologies improve the efficiency and accuracy of Buffetti's Anti-Money Laundering (AML) processes compared to the current manual execution via Excel files?"** This question was rooted in the recognition that Buffetti's existing manual AML processes were not only time-consuming and error-prone but also costly, posing significant challenges in meeting stringent regulatory requirements.

As the research unfolded, the initial focus expanded to address emerging complexities within Buffetti's AML operations. These included concerns about the accuracy and reliability of the current manual processes, the high operational costs associated with these processes, and the time delays that jeopardized regulatory compliance. The study also explored the effectiveness of anomaly detection in identifying potential money laundering activities, the challenges posed by Buffetti's complex data structure, and the impact of limited feedback from regulators on the performance of machine learning models.

In response to these challenges, the research sought to achieve several specific objectives:

6.1.1 Develop an Anomaly Checking Model

The primary objective was to create a machine learning model capable of real-time anomaly detection in financial transactions, thereby enhancing Buffetti's ability to address suspicious activities promptly and reduce the risk of non-compliance.

6.1.2 Automate Delivery Operations

The research aimed to streamline and automate AML-related processes to improve operational efficiency, reduce errors, and ensure consistent compliance with regulatory standards.

6.1.3 Evaluate Model Performance

A crucial objective was to assess the performance of the AML model by comparing its accuracy and reliability against existing manual processes, providing data-driven insights for further refinement.

6.1.4 Provide Insights and Recommendations

Finally, the research sought to offer actionable insights and recommendations to stakeholders, guiding strategic decisions and future improvements in Buffetti's AML processes.

Through these objectives, the research ultimately aimed to enhance the accuracy, efficiency, and timeliness of Buffetti Finance's AML operations, optimize data handling strategies, and facilitate better communication with stakeholders. Achieving these goals not only bolstered Buffetti's compliance capabilities but also positioned the company as an innovator in the fintech industry, setting a new standard for AML practices.

6.2 Main Findings and Conclusions

This research provided a thorough examination of Buffetti Finance's Anti-Money Laundering (AML) processes through the application of advanced machine learning techniques and data analytics. The analysis of six months of transaction data, encompassing over 116,000

transactions across 797 agencies and involving 57,376 customers, yielded several important findings.

6.2.1 Anomaly Detection and Financial Impact

The Isolation Forest algorithm identified 5,818 anomalies, representing 5% of the total transactions. These anomalies were significant not only in number but also in financial terms, accounting for €2.41 million, or approximately 22% of the total transaction value of €10.92 million. This indicates that a substantial portion of the financial transactions analyzed may warrant further investigation, underscoring the effectiveness of the anomaly detection model in highlighting potentially suspicious activities.

6.2.2 Patterns of Potential Money Laundering Activities

Several concerning patterns emerged from the anomaly detection results, particularly regarding the transactions that fell just below €1,000. This trend suggests deliberate attempts by clients and agencies to evade regulatory thresholds and avoid detection. Furthermore, the analysis of specific cases, such as a client transferring nearly €15,000 within a single week and another associated with 155 anomalies involving high-value transactions, pointed to possible money laundering strategies like structuring and layering.

6.2.3 Importance of Detailed Transactional Analysis

The research demonstrated that while aggregate anomaly scores are useful for identifying broad trends, a deeper dive into individual transactions is crucial for uncovering more subtle patterns of suspicious activity. For instance, the concentration of anomalies within specific clients and

agencies highlighted the need for targeted investigations, which are essential for confirming the nature of these transactions and ensuring regulatory compliance.

6.2.4 Utility of the Display Tool in Investigations

The custom display tool developed during this research proved to be an invaluable asset for AML professionals. It enables quick isolation and examination of transactions associated with specific agencies or clients, prioritized by anomaly score. This functionality not only enhances the detection process but also significantly improves the efficiency of subsequent investigative efforts.

6.2.5 Conclusions

The findings from this research illustrate the potential of machine learning, particularly anomaly detection algorithms, to significantly enhance the efficiency and accuracy of AML processes at Buffetti Finance. The identification of high-value, suspicious transactions indicates that automation and advanced data analytics can play a critical role in detecting and preventing money laundering activities. However, the study also reveals the necessity of combining automated detection with manual review to ensure the reliability and thoroughness of AML efforts. This integrated approach is essential for maintaining regulatory compliance and safeguarding the financial integrity of institutions.

6.3 Contributions of the Research

This research makes several significant contributions to the field of Anti-Money Laundering (AML) and financial compliance, particularly in the integration of machine learning, cloud infrastructure, and regulatory strategy optimization.

6.3.1 Challenging the Adequacy of Traditional AML Systems

One of the key contributions of this research is the identification of the limitations inherent in traditional, rule-based AML systems. These systems, while effective in straightforward scenarios, fall short when confronted with modern, sophisticated money laundering techniques. By demonstrating the superiority of machine learning models in detecting complex and evolving patterns within large datasets, this research advocates for a shift away from rigid, rule-based systems towards more adaptive and intelligent solutions. This has significant implications for financial institutions, as it suggests a pathway to more accurate and efficient AML practices that can keep pace with the growing complexity of financial crimes.

6.3.2 Advocating for Explainable AI in AML

The research also underscores the importance of integrating explainable AI into AML processes. While machine learning models offer enhanced detection capabilities, their complexity can make them difficult to understand for non-technical stakeholders. By highlighting this issue, the research contributes to the growing discourse on the necessity of developing AI models that not only perform well but also provide interpretable and actionable insights. This is crucial for gaining the trust of stakeholders, ensuring regulatory compliance, and making informed decisions based on the model's outputs.

6.3.3 Demonstrating the Role of Cloud Infrastructure

Another major contribution is the demonstration of how cloud infrastructure, specifically Microsoft Fabric, can be leveraged to manage and analyze large volumes of complex financial data efficiently. The research illustrates that cloud-based solutions are not merely a convenience but a necessity in modern AML efforts. By facilitating centralized data management, scalable

storage, and advanced analytics, cloud platforms enable the processing and analysis of intricate datasets that are critical for effective anomaly detection and AML compliance. This finding supports the broader adoption of cloud technologies in financial compliance operations.

6.3.4 Insights into Optimizing Regulatory Strategies

The research also contributes to the optimization of regulatory strategies, particularly in the context of information asymmetry between financial institutions and regulators. The observed Nash equilibrium scenario, where both parties operate under suboptimal conditions due to a lack of information sharing, highlights the need for improved feedback mechanisms and collaborative strategies. This insight is crucial for both regulators and financial institutions as it points towards the potential for more effective AML enforcement through enhanced communication and cooperation.

6.3.5 Addressing the Impact of Data Complexity

Finally, the research brings to light the significant impact that data complexity has on the effectiveness of AML systems. By demonstrating the challenges and solutions related to managing complex data structures across multiple transaction tables, the research contributes to the understanding of how data management practices can either hinder or enhance the effectiveness of AML processes. This emphasizes the importance of robust data consolidation and preprocessing strategies as foundational elements of any successful AML system.

6.4 Implications for the Field

The findings from this research have far-reaching implications for the field of Anti-Money Laundering (AML) and financial compliance, underscoring the necessity for a multi-dimensional and technologically advanced approach to combating illicit financial activities.

6.4.1 Need for Comprehensive AML Strategies

The limited scope of traditional AML processes, which often focus primarily on the placement stage of money laundering, highlights a critical gap in addressing the more sophisticated stages of layering and integration. This research reveals the importance of developing comprehensive strategies that encompass all stages of money laundering to effectively capture the full spectrum of illicit activities. By broadening the scope of AML efforts, financial institutions can better anticipate and mitigate the complex techniques used by money launderers.

6.4.2 Enhancing Regulatory Collaboration

The findings emphasize the significant role that regulatory feedback plays in refining machine learning models used in AML. The current deficiency in feedback mechanisms not only hampers the predictive accuracy of these models but also underscores the broader issue of information asymmetry between regulators and financial institutions. To enhance the effectiveness of AML strategies, there is a clear need for improved communication and collaboration with regulators. This will enable the development of more accurate and reliable models, ultimately leading to better detection and prevention of money laundering activities.

6.4.3 Addressing Information Asymmetry

The research sheds light on the impact of information asymmetry, which creates a Nash equilibrium where both financial institutions and regulators are operating under suboptimal

conditions. This equilibrium complicates AML efforts, suggesting that a more balanced approach to information sharing is essential. By fostering greater transparency and cooperation between parties, the field can move towards more effective AML outcomes, reducing the risk of money laundering and enhancing compliance.

6.4.4 Advancing Data-Driven AML Techniques

The focus on investigating relationships and transaction patterns as part of anomaly detection offers a strategic advancement in identifying potential money laundering activities. By leveraging machine learning to uncover hidden patterns and connections within complex datasets, financial institutions can gain deeper insights into illicit financial activities. This data-driven approach not only enhances the detection of suspicious transactions but also provides a more nuanced understanding of how money laundering schemes operate.

6.4.5 Overcoming Data and System Constraints

The challenges associated with managing complex data structures and system limitations, such as those encountered in PowerBI, highlight the need for more robust data management solutions. The research demonstrates that while current tools are effective to a point, there is a pressing need for more advanced, scalable systems that can handle large datasets over extended periods. The exploration of cloud-based solutions like Microsoft Fabric offers promising avenues for overcoming these limitations, enabling more comprehensive and efficient AML investigations.

6.4.6 Efficiency in Anomaly Detection and Data Processing

The implementation of anomaly detection methods, particularly the Isolation Forest model, underscores the effectiveness of machine learning in identifying suspicious transactions.

However, the resource-intensive nature of data handling and processing highlights the importance of optimizing these workflows. The successful data compression and consolidation achieved in this research point to the potential for streamlining AML processes, making them more efficient and scalable for future applications.

6.4.7 Leveraging Cloud Infrastructure for AML

The consideration and exploration of platforms like Azure Fabric and Microsoft Fabric reflect the growing importance of cloud infrastructure in managing and analyzing AML data. These platforms offer significant advantages in terms of simplicity, cost-effectiveness, and functionality. By unifying data and automating processes, cloud-based solutions can enhance the efficiency and effectiveness of AML systems, making them better equipped to handle the complexities of modern financial transactions.

6.4.8 Pioneering Explainable AI in AML

The research also emphasizes the critical need for explainable AI in AML processes. As machine learning models become increasingly integral to detecting financial crimes, ensuring that these models provide interpretable and actionable insights is paramount. This not only builds trust among stakeholders but also ensures that the outputs of these models can be effectively used in regulatory and compliance contexts.

References

Bakhshinejad, Nazanin, et al. *A Survey of Machine Learning Based Anti-Money Laundering Solutions*. Department of Electrical Engineering and Computer Science, York University, October 2022.

Vilella, Salvatore, et al. *Anomaly Detection in Cross-Country Money Transfer Temporal Networks*. DISIT, Universita degli Studi del Piemonte Orientale, Dipartimento di Informatica Universita degli Studi di Torino, Anti Financial Crime Digital Hub Corso Inghilterra, 20 Feb. 2024.

Eddin, Ahmad Naser, et al. *Anti-Money Laundering Alert Optimization Using Machine Learning with Graphs*. University of Porto, 17 June 2022.

Yen Wee, Lim. "Unsupervised Outlier Detection with Isolation Forest." *Medium*, 17 Mar. 2022, https://medium.com/@limyenwee_19946/unsupervised-outlier-detection-with-isolation-forest-eab398c593b2.

Minder, Randy. *Microsoft Fabric: The Future of Data Analytics for Power BI, Data Factory, Data Engineering, Data Science and Data Warehousing*. Udemy, July 2024, www.udemy.com/course/microsoft-fabric-k/learn/lecture/38690850?start=0#content. Accessed 5 Sept. 2024.

Enterprise DNA. *Beginners Guide to Power BI*. Certificate ID dc0823b2-20b5-4093-9e23-d74db51ba126, issued 20 May 2024. Enterprise DNA, www.enterprisedna.co/course/beginners-guide-to-power-bi.

Enterprise DNA. *Ultimate Beginners Guide to DAX*. Certificate ID 547565e2-2c65-4c76-9d58-ffc907c0578c, issued 22 May 2024. Enterprise DNA, www.enterprisedna.co/course/ultimate-beginners-guide-to-dax.

Appendices

A1.1 Explainable AI - IForest model translated into Italian

Panoramica del Funzionamento dell'Isolation Forest

Tradizionalmente, altri metodi cercano di costruire un profilo dei dati normali per poi identificare i punti dati che non si conformano a tale profilo come anomalie.

La parte brillante dell'Isolation Forest è che può rilevare direttamente le anomalie usando l'isolamento (quanto un punto dato è distante rispetto al resto dei dati). Questo significa che l'algoritmo può operare con una complessità temporale lineare, come altri modelli basati sulla distanza, come i K-Nearest Neighbors.

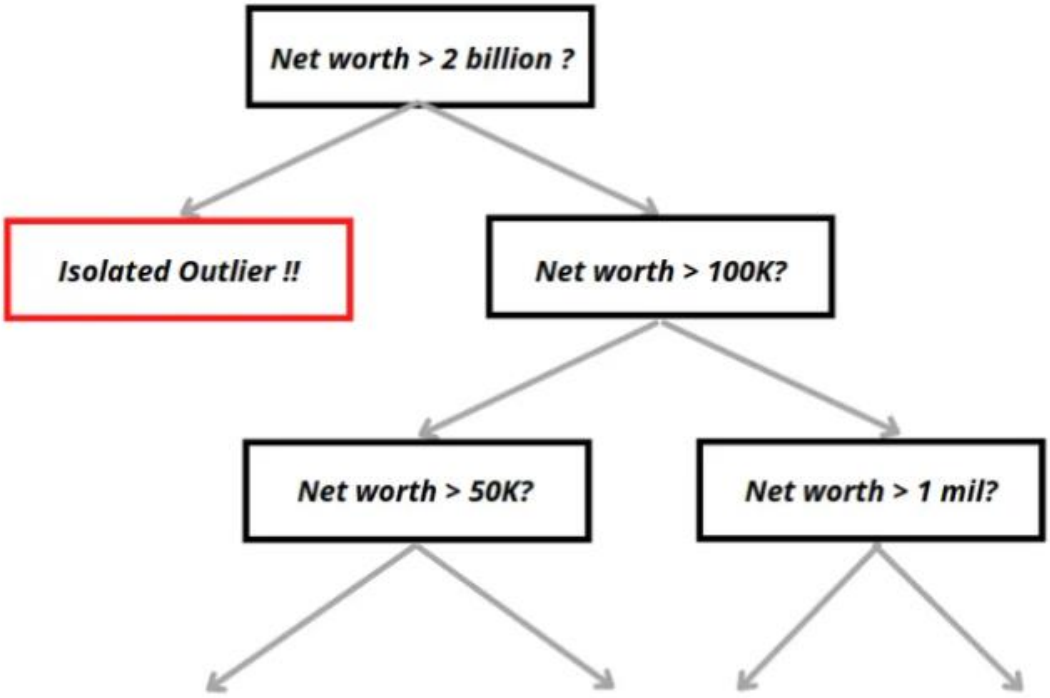
L'algoritmo funziona concentrandosi sugli attributi più evidenti di un outlier:

- Ci saranno solo pochi outlier
- Questi outlier saranno differenti

Isolation Forest lo fa introducendo (un insieme di) alberi binari che generano ricorsivamente partizioni selezionando casualmente una caratteristica e poi selezionando casualmente un valore di divisione per la caratteristica. Il processo di partizionamento continuerà fino a quando non separerà tutti i punti dati dal resto dei campioni.

Dato che viene selezionata solo una caratteristica da un'istanza per ciascun albero, si può dire che la profondità massima dell'albero decisionale è in realtà uno. Infatti, l'estimatore base di un'Isolation Forest è in realtà un albero decisionale estremamente casuale (ExtraTrees) su vari sottoinsiemi di dati.

Un esempio di un singolo albero in un'Isolation Forest può essere visto qui sotto.



Dato gli attributi di un outlier menzionati sopra, possiamo osservare che un outlier richiederà meno partizioni in media per essere isolato rispetto ai campioni normali. Ogni punto dati riceverà quindi un punteggio basato su quanto facilmente viene isolato dopo un numero X di iterazioni. I punti dati che hanno punteggi anomali saranno quindi contrassegnati come anomalie.

Dettagli

In termini più formali, dividiamo ricorsivamente ogni istanza di dati selezionando casualmente un attributo q e un valore di divisione p (all'interno del minimo e massimo dell'attributo q) fino a quando tutte le istanze non sono completamente isolate. L'Isolation Forest fornirà quindi una classifica che riflette il grado di anomalia di ciascuna istanza di dati in base alla lunghezza dei loro percorsi. La classifica o i punteggi sono chiamati punteggi di anomalia, calcolati come segue:

H(x) : il numero di passaggi fino a quando l'istanza di dati x è completamente isolata.

E[H(x)] : la media di H(x) da una raccolta di alberi di isolamento.

Le metriche hanno senso, ma un problema è che il numero massimo possibile di passaggi dell'iTree cresce in ordine di n mentre allo stesso tempo i passaggi medi crescono in ordine di log n. Questo porterà a problemi in cui i passaggi non possono essere confrontati direttamente. Pertanto, sarà necessario introdurre una costante di normalizzazione variabile in base a n.

c(n), la costante di normalizzazione della lunghezza del percorso con la seguente formula:

$$c(m) = \begin{cases} 2H(m-1) - \frac{2(m-1)}{n} & \text{for } m > 2 \\ 1 & \text{for } m = 2 \\ 0 & \text{otherwise} \end{cases}$$

dove H(i) è il numero armonico che può essere stimato come ln(i) + 0,5772156649 (costante di Eulero).

L'equazione completa del punteggio di anomalia:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}$$

Pertanto, se facciamo passare l'intero set di dati attraverso un'Isolation Forest, possiamo ottenere il suo punteggio di anomalia. Utilizzando il punteggio di anomalia s, possiamo dedurre se ci sono anomalie ogni volta che ci sono istanze con un punteggio di anomalia molto vicino a uno.

Reference:

Yen Wee, L. (2022, March 17). Unsupervised outlier detection with isolation forest. Medium.

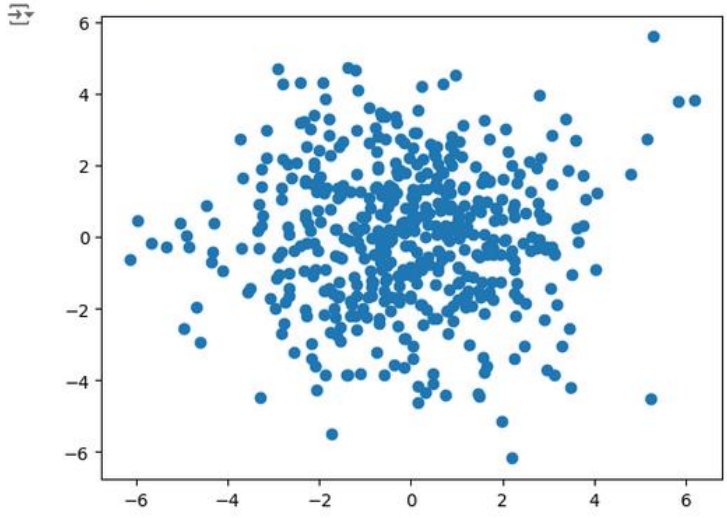
https://medium.com/@limyenwee_19946/unsupervised-outlier-detection-with-isolation-forest-eab398c593b2

iforest_example.ipynb_

Implementation of Isolation Forest

Per prima cosa, importiamo rapidamente alcuni moduli utili che utilizzeremo in seguito. Generiamo un dataset con punti dati casuali utilizzando la funzione make_blobs().

```
[ ] import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt
    from sklearn.datasets import make_blobs
    data, _ = make_blobs(n_samples=500, centers=1, cluster_std=2, center_box=(0, 0))
    plt.scatter(data[:, 0], data[:, 1])
    plt.show()
```



Possiamo facilmente individuare alcuni outlier poiché questo è solo un caso d'uso 2-D. È una buona scelta per dimostrare che l'algoritmo funziona. Si noti che l'algoritmo può essere utilizzato su un set di dati con più caratteristiche senza alcun problema.

Inizializziamo un oggetto isolation forest chiamando IsolationForest().

Gli iperparametri utilizzati qui sono per lo più predefiniti e raccomandati dall'articolo originale.

Il numero di alberi controlla la dimensione dell'ensemble. Abbiamo scoperto che le lunghezze dei percorsi di solito convergono bene prima di t = 100. A meno che non sia specificato diversamente, utilizzeremo t = 100 come valore predefinito nel nostro esperimento.

Empiricamente, abbiamo scoperto che impostare il campione del sottoinsieme a 256 fornisce generalmente dettagli sufficienti per eseguire il rilevamento delle anomalie su un'ampia gamma di dati.

— Fei Tony Liu, Kai Ming Ting (Autore dell'articolo originale, Isolation Forest)

N_estimators qui rappresenta il numero di alberi e max_sample rappresenta il campione del sottoinsieme utilizzato in ogni iterazione. Max_samples = 'auto' imposta la dimensione del sottoinsieme come min(256, num_samples). Il parametro contamination qui rappresenta la proporzione di outlier nel set di dati. Di default, la soglia del punteggio di anomalia seguirà quanto indicato nell'articolo originale. Tuttavia, possiamo fissare manualmente la proporzione di outlier nel set di dati se abbiamo qualche conoscenza preliminare. Lo impostiamo a 0.03 qui per scopi dimostrativi. Successivamente, adattiamo e prevediamo l'intero set di dati. Restituisce un array costituito da [-1 o 1] dove -1 rappresenta un'anomalia e 1

rappresenta un'istanza normale.

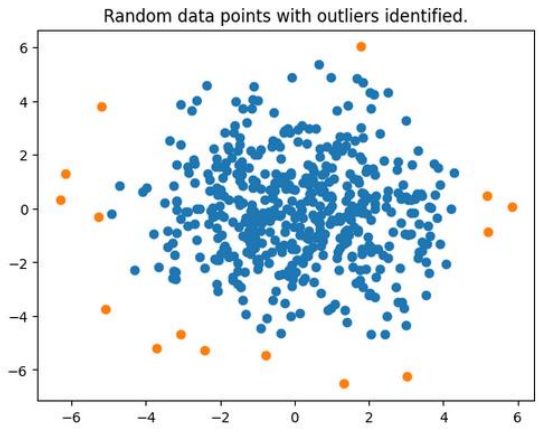
```
[2]  from sklearn.ensemble import IsolationForest
```

```
[3]  iforest = IsolationForest(n_estimators=100, contamination=0.03, max_samples='auto')
     prediction = iforest.fit_predict(data)
     print(prediction[:20])
     print("Number of outliers detected: {}".format(prediction[prediction < 0].sum()))
     print("Number of normal samples detected: {}".format(prediction[prediction > 0].sum()))
```

```
[ 1  1  1  1  1  1 -1  1  1  1  1  1  1  1  1  1  1  1  1  1]
Number of outliers detected: -15
Number of normal samples detected: 485
```

Dopodiché graficheremo gli outlier rilevati dall'Isolation Forest.

```
[4]  normal_data = data[np.where(prediction > 0)]
     outliers = data[np.where(prediction < 0)]
     plt.scatter(normal_data[:, 0], normal_data[:, 1])
     plt.scatter(outliers[:, 0], outliers[:, 1])
     plt.title("Random data points with outliers identified.")
     plt.show()
```



Possiamo vedere che funziona piuttosto bene e identifica i punti dati intorno ai margini.

Possiamo anche chiamare decision_function() per calcolare il punteggio di anomalia di ciascun punto dati. In questo modo possiamo capire quali punti dati sono più anomali.

```
[5] score = iforest.decision_function(data)
    data_scores = pd.DataFrame(list(zip(data[:, 0],data[:, 1],score)),columns = ['X','Y','Anomaly Score'])
    display(data_scores.head())
```

	X	Y	Anomaly Score
0	0.345348	-1.774484	0.187192
1	-2.257057	-0.067699	0.180781
2	3.914781	-1.303171	0.069867
3	-2.158931	-0.261864	0.187487
4	2.167824	0.185833	0.181555

Selezioniamo le prime 5 anomalie utilizzando i punteggi di anomalia e quindi plottiamole nuovamente

```
[6] top_5_outliers = data_scores.sort_values(by = ['Anomaly Score']).head()
    plt.scatter(data[:, 0], data[:, 1])
    plt.scatter(top_5_outliers['X'], top_5_outliers['Y'])
    plt.title("Random data points with only 5 outliers identified.")
    plt.show()
```

Conclusione

Isolation Forest è un modello di rilevamento degli outlier fondamentalmente diverso in grado di isolare le anomalie con grande velocità. Ha una complessità temporale lineare, il che lo rende uno dei migliori per gestire set di dati ad alto volume.

Si basa sul concetto che, poiché le anomalie sono "poche e diverse", sono più facili da isolare rispetto ai punti normali.

Reference:

Yen Wee, L. (2022, March 17). Unsupervised outlier detection with isolation forest. Medium. https://medium.com/@limyenwee_19946/unsupervised-outlier-detection-with-isolation-forest-eab398c593b2

A1.2 Python Data Code used to combine data tables and formulate anomalies

```
# Step 1: Mount Google Drive (if using Google Drive)
```

```
from google.colab import drive
drive.mount('/content/drive')
```

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).

```
# Step 2: Import Necessary Libraries
import pandas as pd
import glob
import os
from datetime import datetime
import logging
```

```
# Set up logging
logging.basicConfig(level=logging.INFO)
logger = logging.getLogger(__name__)
```

```
# Step 4: Read and Combine CSV Files
dfs = []
total_rows = 0
for file in files:
    try:
        df = pd.read_csv(file)
        total_rows += len(df)
        dfs.append(df)
        logger.info(f"Successfully read {file} with {len(df)} rows.")
    except Exception as e:
        logger.error(f"Error reading {file}: {e}")

combined_df = pd.concat(dfs, ignore_index=True)
logger.info(f"Combined DataFrame has {len(combined_df)} rows and
{combined_df.shape[1]} columns.")
```

```
# Step 5: Parse and Sort by Date
def parse_date(date_str):
    for fmt in ("%Y-%m-%d %H:%M:%S", "%Y-%m-%d", "%A, %B %d, %Y"):
        try:
            return datetime.strptime(date_str, fmt)
        except ValueError:
            continue
```

```
    logger.error(f"No valid date format found for {date_str}")
    return pd.NaT  # Return NaT for invalid dates


# Step 6: Apply the Custom Date Parsing Function
combined_df['PaymentDate'] = combined_df['PaymentDate'].apply(parse_date)
combined_df.dropna(subset=['PaymentDate'], inplace=True)  # Remove rows with
invalid dates
combined_df.sort_values(by='PaymentDate', inplace=True)


# Verification

# Check the structure of the combined DataFrame
logger.info(f"First few rows of the combined DataFrame:\n{combined_df.head()}")
logger.info(f"Last few rows of the combined DataFrame:\n{combined_df.tail()}")

# Check for missing values
missing_values = combined_df.isnull().sum()
logger.info(f"Missing values in each column:\n{missing_values}")

# Verify data types
data_types = combined_df.dtypes
logger.info(f"Data types of each column:\n{data_types}")

# Randomly sample a few rows to verify content
sample_rows = combined_df.sample(5)
logger.info(f"Sample rows from the combined DataFrame:\n{sample_rows}")

# Verify the total number of rows
if total_rows == len(combined_df):
    logger.info(f"Row count verification passed: {total_rows} rows in total.")
else:
    logger.error(f"Row count verification failed: {total_rows} rows expected, but
got {len(combined_df)} rows.")

# Output the first few rows of the combined DataFrame to visually inspect
combined_df.head()
```

	TransactionId	AgencyId	CustomerId	FeesAmount	TransferAmount	PaymentType	PaymentCause	TypeOfBullettinValue	PaymentDate
94837	5IRLI8	707104	499731.0	1.0	10.00	0	Iliad ricarica 10€	PhoneRecharge	2024-01-01
62288	360630017852892448000010000 1IT	28565	826674.0	1.3	530.00	0	Pagamento B123	B123	2024-01-01
97250	76FRIE	706998	499726.0	1.8	20.00	0	Vodafone Ricarica OL 20€	PhoneRecharge	2024-01-01
57504	360630017852852448000010000 1IT	28565	NaN	1.3	92.11	0	Pagamento B896 000065802027893727	B896	2024-01-01
105295	C9SV0G	687852	499750.0	2.5	50.00	0	Amazon Pin 50 Euro	PurchaseCode	2024-01-01

```
combined_df.shape
```
```
(116321, 9)
```

```
pip install pycaret --quiet
```

```
import pandas as pd
import datetime
from datetime import timedelta
import plotly.express as px
import numpy as np
```

```
combined_df.info()
```
```
<class 'pandas.core.frame.DataFrame'>
Index: 116321 entries, 94837 to 4268
Data columns (total 9 columns):
 #  Column           Non-Null Count   Dtype
--- ------          --------------  -----
 0  TransactionId    116087 non-null  object
 1  AgencyId         116321 non-null  int64
 2  CustomerId       89952 non-null   float64
 3  FeesAmount       116321 non-null  float64
 4  TransferAmount   116321 non-null  float64
 5  PaymentType      116321 non-null  int64
 6  PaymentCause     112854 non-null  object
 7  TypeOfBullettinValue  116321 non-null  object
 8  PaymentDate      116321 non-null  datetime64[ns]
dtypes: datetime64[ns](1), float64(3), int64(2), object(3)
memory usage: 8.9+ MB
```

```
# Count the number of unique items in each column of the DataFrame
unique_counts = combined_df.nunique()

# Print the unique counts
print(unique_counts)
```

```
TransactionId        116082
AgencyId             797
CustomerId           57376
FeesAmount           35
TransferAmount       16052
PaymentType          5
PaymentCause         60948
TypeOfBullettinValue 17
PaymentDate          151
dtype: int64
```

```
from pycaret.anomaly import *
s = setup(combined_df, ignore_features = [
    "TransactionId",
    "CustomerId",
    "PaymentCause"
], session_id = 123)
```

	Description	Value
0	Session id	123
1	Original data shape	(116321, 9)
2	Transformed data shape	(116321, 24)
3	Ignore features	3
4	Numeric features	4
5	Date features	1
6	Categorical features	1
7	Rows with missing values	25.7%
8	Preprocess	True
9	Imputation type	simple
10	Numeric imputation	mean
11	Categorical imputation	mode
12	Maximum one-hot encoding	-1
13	Encoding method	None
14	CPU Jobs	-1
15	Use GPU	False
16	Log Experiment	False
17	Experiment Name	anomaly-default-name
18	USI	b16e

```
model = create_model('iforest', fraction = 0.05)
```

```
model
```

```
IForest(behaviour='new', bootstrap=False, contamination=0.05,
  max_features=1.0, max_samples='auto', n_estimators=100, n_jobs=-1,
  random_state=123, verbose=0
```

```
results = assign_model(model)
```

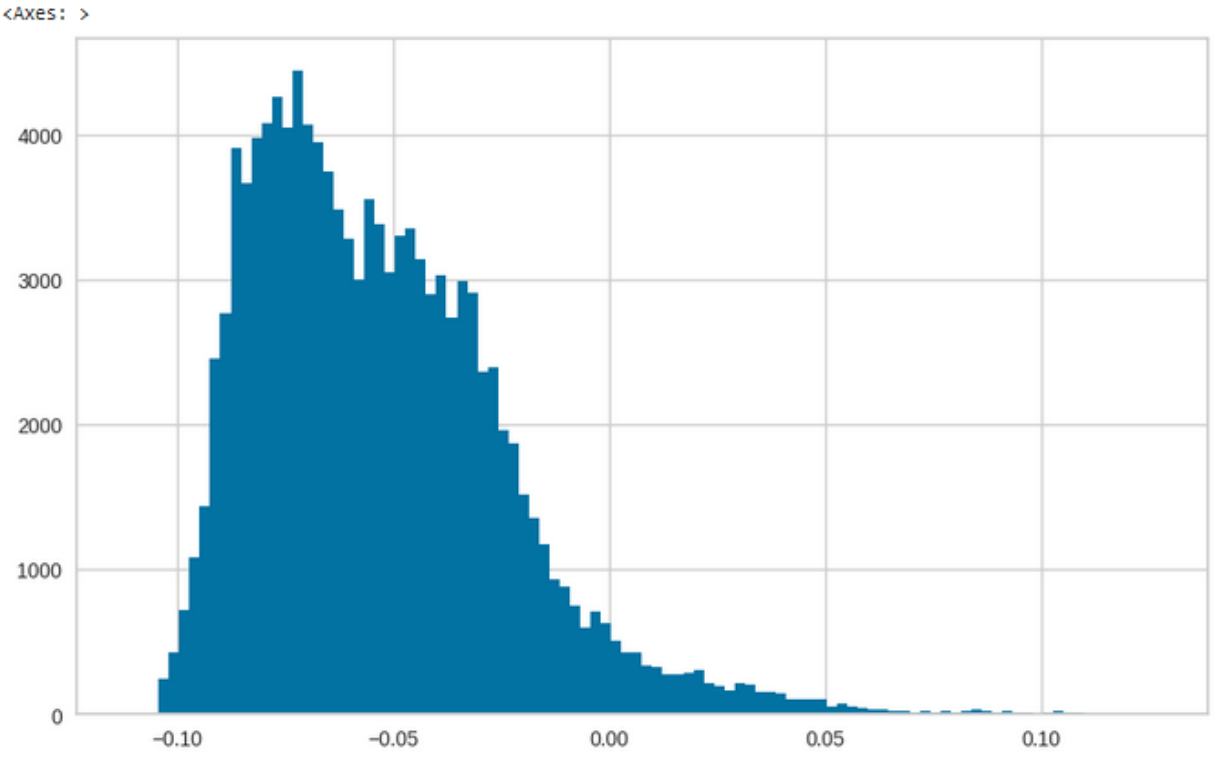
```
results.sort_values(by = 'Anomaly_Score', ascending=False)
```

	AgencyId	FeesAmount	TransferAmount	PaymentType	TypeOfBullettinValue	PaymentDate	Anomaly	Anomaly_Score
46644	396694	0.0	949.000000	5	Voucher	2024-01-03	1	0.126582
46499	396694	0.0	999.000000	5	Voucher	2024-05-01	1	0.119929
46672	396694	0.0	999.000000	5	Voucher	2024-01-05	1	0.119794
46907	396694	0.0	900.000000	5	Voucher	2024-01-04	1	0.119733
46479	396694	0.0	999.000000	5	Voucher	2024-01-06	1	0.118962
...	...	...	...	...	...	...	...	...
45593	4870	1.5	51.270000	1	Rav	2024-02-27	0	-0.104317
45582	810196	2.0	43.660000	0	Rav	2024-03-08	0	-0.104495
45585	119468	1.5	109.900002	1	Rav	2024-02-12	0	-0.105936
45592	778253	1.5	157.190002	1	Rav	2024-04-22	0	-0.107058
45588	119468	1.5	86.709999	0	Rav	2024-02-07	0	-0.111413

116321 rows × 8 columns

```
results['Anomaly_Score'].hist(bins=100, figsize =(10,6))
```

```
<Axes: >
```




```
results_subset = results[['Anomaly', 'Anomaly_Score']]
```

```
# Merge the selected columns into 'combined_df' based on the index
combined_results_df = combined_df.join(results_subset)
```

```
combined_results_df
```

	TransactionId	AgencyId	CustomerId	FeesAmount	TransferAmount	PaymentType	PaymentCause	TypeOfBullettinValue	PaymentDate	Anomaly	Anomaly_Score
94837	5IRLI8	707104	499731.0	1.00	10.00	0	Iliad ricarica 10€	PhoneRecharge	2024-01-01	0	-0.039624
62288	3606300178528924480000100001IT	28565	826674.0	1.30	530.00	0	Pagamento B123	B123	2024-01-01	1	0.046060
97250	76FRIE	706998	499726.0	1.80	20.00	0	Vodafone Ricarica OL 20€	PhoneRecharge	2024-01-01	0	-0.038835
57504	3606300178528524480000100001IT	28565	NaN	1.30	92.11	0	Pagamento B896 000065802027893727	B896	2024-01-01	0	-0.039869
105295	C9SV0G	687852	499750.0	2.50	50.00	0	Amazon Pin 50 Euro	PurchaseCode	2024-01-01	1	0.037335
...	...	...	...	...	...	...	...	...	...	...	...
2791	396800006885704189	512573	513299.0	0.89	16.00	4	/RFB/96800006885704189/TXT/Repertor	PagoPa	2024-05-30	0	-0.078016
2792	396800006885698735	512573	513299.0	0.89	16.00	4	/RFB/96800006885698735/TXT/Repertor	PagoPa	2024-05-30	0	-0.078016
2793	396800006885695503	512573	513299.0	0.89	16.00	4	/RFB/96800006885695503/TXT/Repertor	PagoPa	2024-05-30	0	-0.078016
2784	396800006885728857	512573	513299.0	0.89	16.00	4	/RFB/96800006885728857/TXT/Repertor	PagoPa	2024-05-30	0	-0.078016
4268	396800006883717330	776477	778891.0	2.75	208.00	4	/RFB/96800006883717330/TXT/RP:24H39	PagoPaPra	2024-05-30	0	-0.047877

116321 rows × 11 columns

```
# Step 6: Export Combined DataFrame to CSV
output_path = '/content/drive/My Drive/Buffetti/Payments/combined_data.csv'  # If using Google Drive
```

```
combined_results_df.to_csv(output_path, index=False)
print(f"Combined file saved to: {output_path}")
```

Combined file saved to: /content/drive/My Drive/Buffetti/Payments/combined_data.cs