Ca' Foscari University of Venice

Master's Degree Program

in Data Analytics for Business and Society

Final Thesis

AI FOR FINANCIAL PLANNING

A Business Case

Supervisor
Prof. Giuliano Bonollo

Graduand
Niusha Khodadad Jafari
Mat. 893610

Academic Year
2023 – 2024

Abstract

In today's business landscape, the increase of data presents both challenges and opportunities for financial planning within companies. Access to extensive datasets offers the potential for informed decision-making, predictive insights, and enhanced efficiency. Leveraging data effectively is crucial for businesses to adapt to market changes, innovate, and manage risks. In essence, in this data-rich environment, utilizing data for financial planning is essential for a company's success and resilience in a dynamic marketplace.

This thesis serves as a comprehensive exploration of the intersection between business intelligence (BI) and artificial intelligence (AI), illuminating their integration to enhance financial planning processes within organizations. Employing a specific company's financial data as a case study, this research aims to provide a tangible demonstration of how the synergy between BI and AI can revolutionize financial planning methodologies. By delving into the intricacies of this integration, the goal is to develop a cutting-edge and sophisticated approach to crafting financial plans that not only meet immediate objectives but also pave the way for long-term strategic success within the organization.

Utilizing a set of tools such as Microsoft Power BI, Python, and R, coupled with advanced statistical models and artificial intelligence techniques, this research endeavors to evolve financial planning methodologies within companies. Through the integration of predictive analytics, the thesis seeks to enhance the accuracy and efficacy of financial forecasts, enabling the generation of actionable insights and timely alerts for potential risks and opportunities.

In summary, this thesis delves into the synergy between traditional financial planning practices and cutting-edge AI technologies, with the ultimate objective of empowering businesses to navigate complex economic landscapes with confidence and foresight.

Table of Contents

Section 1

Introduction

This thesis investigates the integration of AI with BI tools and its impact on data analysis and decision-making processes within organizations.

The structure of the thesis consists of six chapters, each dedicated to different aspects of the research journey. Chapter 1 provides the background, research objectives, significance, scope, limitations, and challenges. Moving forward, Chapter 2 delves into the transformation of BI to AI and the utilization of artificial intelligence in business intelligence practices. In Chapter 3, we explore the datasets and technical tools used, while Chapter 4 outlines the case study approach and methodology. Subsequently, Chapter 5 presents findings and discusses implications, and Chapter 6 concludes with key insights and recommendations for future work.

1.1 Background

In today's data-driven world, the ability to analyze and interpret vast amounts of data is crucial for making informed business decisions. Business Intelligence (BI) and data analytics play a pivotal role in this process, enabling organizations to gain insights, identify trends, and make strategic decisions. The integration of Artificial Intelligence (AI) with BI has further revolutionized the field, offering advanced analytical capabilities and enhancing the decision-making process.

As organizations increasingly rely on data to drive their strategies, the need for sophisticated tools that can handle complex data sets and provide actionable insights has become more apparent. Traditional BI platforms, while effective in processing and visualizing data, often lack the advanced analytical capabilities required to fully leverage the potential of big data. This gap has been bridged by the integration of AI technologies, which offer powerful data processing, predictive analytics, and machine learning capabilities.

This thesis explores the integration of BI and AI tools, specifically focusing on the use of Power BI, R, and Python, to enhance data analysis and decision-making in the context of financial planning for a distillation company. By leveraging these tools, the study aims to demonstrate how advanced analytics can transform managerial dashboards and provide valuable insights for business decision-makers.

1.2. Research Objective

The aim of this study is to illustrate the profound impact of integrating AI with BI in practice, particularly on analysis and managerial dashboards, ultimately leading to enhanced decision-making processes. Through this exploration, I aim to shed light on how the synergy between AI and BI tools transforms traditional business intelligence practices.

By leveraging AI alongside BI tools, this study seeks to showcase how it enriches the depth and breadth of analysis conducted within organizations. Furthermore, I aim to examine how this integration affects traditional business intelligence practices, emphasizing the evolution from static reporting to dynamic, predictive analytics.

Finally, this thesis aims to respond to the following question: How does the integration of AI with BI tools influence analysis and managerial dashboards, and what is its impact on financial decision-making processes within organizations?

1.3 Significance of the Study

The significance of this study lies in its potential to bridge the gap between traditional BI platforms and advanced AI-driven analytics. By integrating tools such as Power BI, R, and Python, this research aims to provide a comprehensive framework for enhancing data analysis and decision-making capabilities in the business environment.

The insights gained from this study can help business leaders make more informed decisions, optimize financial planning processes, and improve overall organizational performance. Additionally, this research contributes to the growing body of knowledge on the practical applications of AI and BI in the field of business analytics, offering valuable guidance for future studies and implementations.

1.4 Scope of the Study

This study focuses on the integration of BI and AI tools within the context of financial data analysis for a distillation company. The primary tools examined in this research are Power BI, R, and Python. The dataset used for analysis includes financial transactions, customer information, and other relevant data over a four-year period (2020-2023).

The study will not only explore the technical aspects of integrating these tools but also assess their impact on decision-making and managerial effectiveness. While the primary focus is on financial data, the methodologies and findings can be applicable to other industries and types of data, providing a broader relevance.

1.5 Limitations and Challenges

Despite the significant advantages of integrating AI within BI tools, conducting this thesis posed several notable limitations and challenges:

- Complexity of Integration: Integrating multiple tools (Power BI, R, and Python) required substantial technical expertise and increased the overall complexity of the analysis process.
- Performance Bottlenecks: Handling large datasets and complex computations often resulted in performance issues, requiring optimization. Power BI, already known for its slower processing, experienced further delays due to this integration.
- Maintenance and Updates: Sometimes updates and maintenance were necessary to ensure compatibility between the integrated tools and to meet the requirements of all three platforms in terms of data types and coding standards.
- Time Constraints: The integration process and subsequent analysis were time-consuming, adding to the challenges of ensuring seamless collaboration between platforms. Setting up prerequisites for an optimal environment within Power BI for the effective use of R and Python presented challenges.
- Limitations: Running R and Python within Power BI had its limitations, such as the inability to fit models and display results directly, which affected the analysis process.

Despite these challenges, the study aimed to provide valuable insights into the integration of BI and AI tools for enhanced data analysis and decision-making.

Section 2

From BI to AI

In today's dynamic business realm, the trajectory from Business Intelligence to Artificial Intelligence is a significant journey filled with innovation, transformation, and smart strategies. This chapter explains the world of BI and its variants and uncovers the connection between BI and AI, guiding us through this transformation, which is vital for any company to maintain its competitiveness and remain in the ever-changing business world.

2.1. Data Management

Data serves as the cornerstone upon which organizations derive insights, make decisions, and foster innovation. Initially raw, data encompasses individual facts or statistics in various/s forms, including numbers, text, images, or sounds, gathered through observations, measurements, or research. To unlock its potential, data undergoes a transformation along the data value chain:

i. From Data to Knowledge

Data undergoes processing, organization, or structuring to acquire meaning and context, transitioning into information. For example, numerical data may reveal revenue or costs, textual data could relate to addresses, images may represent chest X-rays, and voice recordings might capture interviews. As this information is utilized to better understand or accomplish tasks, it turns into knowledge. For instance, revenue data aids in investigating company performance, geographical clustering of addresses facilitates location-based insights, chest X-rays contribute to cancer diagnosis, and voice recordings help measure anxiety levels. Data from external sources, whether structured (e.g., databases, JSON files, XML) or unstructured (e.g., images, sounds, PDFs), undergoes storage and archiving, processing, and utilization for reporting, visualization, and other purposes.

ii. Data Storage and Processing

Data storage and archiving rely on systems designed to maintain accessibility for both internal and external resources. Among these systems, databases stand out as structured collections of data, optimized for transactional processing. Used extensively for operational purposes such as managing day-to-day transactions and supporting applications like CRM, ERP, and OLTP systems, databases come in various types including SQL (e.g., Oracle, SQL Server, Snowflake), NoSQL (e.g., Blockchain), and multidimensional formats. Data processing involves manipulation and transformation of raw data into useful information through techniques such as ETL (Extract, Transform and Load) or ELT (Extract, Load and Transform) processes, facilitating integration of data from multiple sources into a consistent format.

iii. Leveraging Data Warehouses and Data Marts

Utilizing Data Warehouses and Data Marts enhances data management efficiency.

Data Warehouses act as centralized repositories for extensive historical data, tailored for analytical queries, reporting, and deep data analysis. These repositories consolidate information, employing methodologies like data cleansing, transformation, and aggregation to ensure data integrity and consistency.

Data Marts, as specialized subsets of Data Warehouses, target specific functional areas or departments within an organization, providing refined data selections meticulously tailored to meet the unique analytical requirements of particular user groups.

iv. Data Utilization

Data is utilized for various purposes, including:

- Directional reporting (e.g., dashboards/synthetic dashboards, Key Performance Indicators (KPIs))
- Operational reporting (e.g., detailed tables of master data and transactions, operational measures, and alarm thresholds)
- Data Mining (e.g., identification of significant clusters and patterns, identification of hidden relationships between entities)
- Simulations (e.g., definition of different scenarios, parameterization of processes)
- Predictive Analysis (e.g., machine learning techniques for projections and diagnoses)
- Flows to external entities (e.g., regulatory flows such as AML and CESOP, reconciliation flows for invoices, commissions, and other transactions)

2.2. Business Intelligence (BI)

Business intelligence is a technology-driven process containing of a set of software and methods intended to collect, examine, and understand data related to business activities.

The goal of BI is to improve the speed and accuracy of information and give managers a clear understanding of where their organization stands compared to competitors.

In the rapidly evolving business landscape characterized by globalization, liberalization, consolidations, and corporate expansions, BI has become a strategic tool for organizations aiming for competitive advantages.

Using BI allows companies to improve the quality and timeliness of information, giving a clearer view of business operations at specific moments.

Through BI tools, companies can analyze changes in market share, shifts in customer behavior and spending habits, customer preferences, capabilities of the company, market conditions and other information of this kind based on their specific needs. These insights help analysts and managers figure out the best adjustments to respond to changing market trends and make more valuable decisions.

In laying the groundwork for a comprehensive Business Intelligence framework, effective data management plays a crucial role. As we have explored earlier, effective data management includes primary processes such as data transformation, storage, processing, and utilization. These processes provide the basis, which BI is structured upon: Data Capture/Acquisition, Data Storage, and Data Access & Analysis.

Within this framework, Data Capture/Acquisition involves the collection of data from diverse sources, from internal databases to external sources like customer, supplier or government, while Data Storage focuses on organizing and securing this data for future analysis. Subsequently, Data Access & Analysis enable users to extract valuable insights from stored data, empowering informed decision-making and strategic planning. the front-end of BI provides user-friendly tools for business users to access data. It serves various user groups through BI tools, which include a set of software with reporting, analysis, and visualization features. This element acts as the interface allowing direct, interactive, or batch access to data, presenting understandable and business-oriented information suitable for non-technical users.

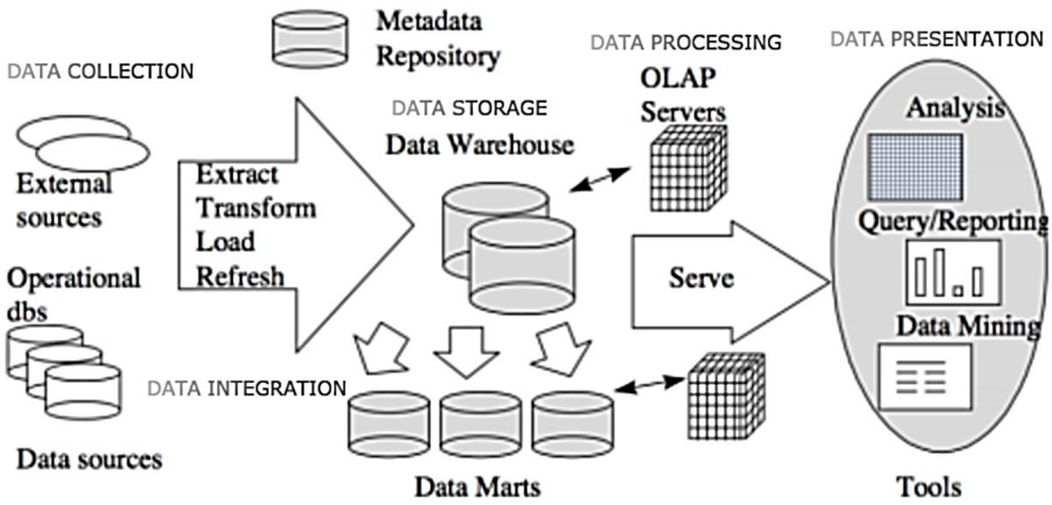

Figure 1: Framework of Business Intelligence[1]

Throughout time, BI embodies three main stages:

- Traditional BI: Focused on gathering data, business analysis, and presenting data visually. It uses various technological tools to create reports and predictions in order to improve decision-making efficiency. Tools like Data Warehouse (DW), Extract-Transform-Load (ETL), Online Analytical Processing (OLAP), Data Mining (DM), Text Mining, Web Mining, and Data Visualization are essential in this approach.
- Business Process Integration: Involves integrating business processes into BI, connecting business process management with strategy. Along with traditional BI tools, this approach includes Business Performance Management (BPM), Business Activity Monitoring (BAM), Service-Oriented Architecture (SOA), Automatic Decision Systems (ADS), and dashboards.
- Adaptive BI: Includes self-learning adaptive systems that suggest optimal actions and continually learn from past decisions by integrating Artificial Intelligence (AI) into BI systems.

[1] Source: - Rafi, A., &Quadri, S.M.K. (2012). Business Intelligence: An Integrated Approach. Retrieved from Business Intelligence Journal, January, Vol.5 No.1

2.3. Business Intelligence Variants

To achieve success in data-driven strategies, businesses should develop a comprehensive analytic framework consisting of three integrated functionality clusters: Descriptive Business Intelligence, Predictive Business Intelligence, and Prescriptive Business Intelligence.

i. Descriptive Business Intelligence

Descriptive Business Intelligence stands as the foundation of traditional BI methodologies, emphasizing the examination of historical data to reveal valuable insights into past events, performance trends, and essential metrics. Its main objective is to provide stakeholders with a comprehensive understanding of past business operations, employing techniques such as data aggregation, reporting, and data visualization to make information both understandable and valuable.

Descriptive BI encompasses several goals aimed at leveraging historical data for comprehensive insights. Firstly, it focuses on historical data summarization, seeking to extract complex datasets into meaningful overviews that enhance the understanding of past trends and patterns. Additionally, Descriptive BI aims to identify key metrics, shining a spotlight on crucial performance indicators across various organizational facets, including sales, inventory, workflow efficiency, and customer behavior. Lastly, the comprehensive goal is to facilitate informed decision-making, empowering decision-makers with valuable insights into historical patterns. This, in turn, enables them to make informed and strategic decisions regarding future actions and strategies.

Pros:

- Offers a user-friendly approach that does not require advanced statistical expertise, making it accessible to a broader audience within the organization.
- Simplifies complex information and provides users with comprehensible visuals.
- Effectively addresses essential business inquiries, allowing companies to compare their current state with both their past performance and that of competitors.

Cons:

- Centered on past data, it lacks the capability to explore the underlying causes of events, limiting its potential to provide in-depth insights.

- Lacking the predictive capabilities, it can only analyze historical data and is unable to foresee future outcomes or address real-time challenges.
- Incorrect use of metrics or compromised data quality might lead to incorrect or misleading conclusions, undermining the reliability of the outcomes.

ii. Predictive Business Intelligence

Predictive Business Intelligence represents an evolutionary leap in the field of data analytics, utilizing sophisticated statistical models, machine learning algorithms and data mining techniques to forecast future outcomes. It involves usage of historical and current data to predict trends and behaviors, providing organizations with a level of precision that spans from seconds to years into the future.

This advanced approach stands in clear contrast to descriptive BI, focusing on proactive decision-making by anticipating future scenarios based on historical data patterns. The creation of predictive models aims to identify correlations within selected datasets, serving as the foundation for forecasting future events.

The process of constructing predictive analytics frameworks involves a series of essential steps. First and foremost is the definition of the problem, a critical starting point that establishes the thesis and requirements for prediction. This could range from detecting fraud to optimizing inventory levels or forecasting potential flood levels. Subsequently, data is acquired from diverse sources and organized within data repositories, such as Big Query, laying the groundwork for the development of predictive models. To ensure the accuracy and reliability of these models, the raw data undergoes a pre-processing phase, involving the removal of inconsistencies, missing data points, or extreme outliers. Data scientists then utilize various tools and techniques, including machine learning, regression models, and decision trees, to create predictive models tailored to the specific problem and dataset. The final stages involve the validation of predictive model accuracy and the delivery of successful results to stakeholders, accessible through applications, websites, or data dashboards.

Pros:

- Allows organizations to make informed decisions ahead of time by predicting future trends.
- Allows manipulation of data in order to test various scenarios, assessing the influence changes might have on models and subsequently decisions.

- The ability to analyze millions of data points quickly, allows for faster decision-making and adapting strategies accordingly.

Cons:

- The accuracy of predictive analytics relies on the quality and availability of data, requiring constant inspection and adjustments.
- Computations regarding model creation and interpretation of results requires specific knowledge and expertise.

iii. Prescriptive Business Intelligence

Prescriptive Business Intelligence is the use of advanced processes and tools to analyze data and content, to identify the optimal course of action, offering recommendations for subsequent steps through a comprehensive consideration of all factors. Prescriptive BI not only predicts future outcomes based on historical data and current conditions but also provides insights into why it will happen, culminating in actionable recommendations.

Techniques employed in prescriptive analytics consist of the utilization of business rules, algorithms, machine learning, and computational modeling. These techniques draw insights from diverse datasets, allowing for a thorough analysis and recommendation of optimal strategies or decisions. Machine learning algorithms enhance the efficiency of prescriptive BI by rapidly analyzing large datasets and generating recommendations based on specific combinations of requirements.

Despite their capability to provide data-informed recommendations, it is crucial to acknowledge that algorithms cannot replace human discernment. Prescriptive BI serves as a tool to inform decisions and strategies, with human judgment providing valuable context and acting as guardrails to algorithmic outputs.

Pros:

- Empowers decision-makers with actionable recommendations; ensuring decisions are based on a thorough analysis of all relevant factors.
- Has the capability to navigate through complex and uncertain situations, providing clarity and direction despite rapidly changing circumstances. It can help prevent fraud, reduces risk, and increases efficiency.

Cons:

- The precision and reliability of recommendations relies directly on the quality of the input data. Consequently, the results are as effective as the data inputs.
- Works best when organizations ask the right questions and respond thoughtfully to the insights it offers. Simply having access to the technology or data analysis tools is not enough; it requires thoughtful interpretation and application of the insights to achieve successful outcomes.
- It may not offer entirely reliable solutions for long-term strategic planning, and continuous monitoring and adaptation is crucial.

Table below shows a summary of the characteristics and differences between each cluster:

	Descriptive BI	Predictive BI	Prescriptive BI
Data Focus	Historical	Historical and Current	Historical and Current
Methodology	Traditional	Statistical and Machine Learning	Analytical and Machine Learning
Objective	Understanding Past Events	Forecasting Future Outcomes	Identifying Optimal Actions
Techniques	Data Aggregation, Reporting, Visualization	Predictive Modeling, Data Mining	Algorithms, Machine Learning
Decision Making	Based on Historical Patterns	Proactive Decision-making	Data-informed Recommendations
Pros	- User-friendly - Simplifies Information	Informed Decisions Ahead of Time	Empowers Decision-makers
Cons	- Lacks Predictive Capabilities -Unable to Explore Underlying Causes	- Accuracy relies on Data Quality - Requires Specific Knowledge	- Relies on Input Data Quality - Requires Thoughtful engagement with the data to prove useful

Table 1: Differences between Descriptive, Predictive and Prescriptive Business Intelligence

2.4. Integration of BI with AI

With the ever-growing volume of information and the emergence of big data, traditional BI faces challenges in handling day-to-day business operations. The demand for real-time processing of both structured and unstructured data is on the rise, making it increasingly demanding to implement effective decision-making methodologies and a more sophisticated intelligence system, commonly known as Artificial Intelligence (AI).

AI, along with its subsets, Machine Learning (ML) and Deep Learning (DL), enhances the resilience of day-to-day operations within organizations.

The capability of AI in collecting, processing, analyzing, and deriving insights from vast datasets, commonly referred to as Big Data, positions it as a crucial tool for organizations where the integration of BI with AI is extensively utilized for effective big data management and statistical analysis. This combination bridges the gap for business users, making analytics and big data insights accessible and understandable to a broader audience, ultimately revolutionizing decision-making processes, reshaping the landscape of data-driven decision-making in enterprises.

The unification of AI with BI involves combining the capabilities of AI technologies with traditional BI tools and processes to enhance data analysis, decision-making, and overall business insights. Below, I outline potential areas and approaches for integrating BI with AI to maximize the utility of these tools and enhance business insights:

- Data Collection and Preparation: BI typically involves collecting and preparing structured data from various sources. AI can contribute by automating the data collection process, identifying patterns in unstructured data, and assisting in data cleaning and transformation. AI algorithms can help organizations handle diverse data formats more effectively.
- Advanced Analytics: AI brings advanced analytical capabilities to BI by using machine learning algorithms for predictive and prescriptive analytics. These algorithms can analyze historical data, identify patterns, and make predictions about future trends. BI tools then leverage these AI-generated insights to provide a more comprehensive view of business performance.
- Natural Language Processing (NLP): NLP enables BI tools to understand and respond to human language. Users can interact with BI systems using natural language queries

or receive reports in a more human-readable format. This simplifies the user experience and makes BI accessible to a broader audience within the organization.

- Automated Reporting and Dashboards: AI can automate the generation of reports and dashboards by dynamically adapting to changing data patterns. This ensures that business leaders receive real-time, relevant insights without manual intervention. AI algorithms can identify key performance indicators (KPIs) and update visualizations accordingly.
- Anomaly Detection: AI-powered anomaly detection helps BI systems identify unusual patterns or outliers in data. This is valuable for detecting fraud, errors, or unexpected trends and allows organizations to respond quickly to unusual occurrences and take corrective actions.
- Personalized Insights: AI can enhance BI by providing personalized insights tailored to individual users. Machine learning algorithms can analyze user behavior, preferences, and historical interactions with the BI system to deliver customized reports and recommendations.
- Enhanced Analytics: The combination of BI and AI results in enhanced analytics, where AI algorithms assist users in exploring and interpreting data, revealing hidden insights that might be challenging for traditional BI methods alone.

In summary, the integration of BI with AI enhances data processing, analysis, and reporting capabilities, making business intelligence more powerful, adaptive, and user-friendly. This collaboration enables organizations to extract deeper insights from their data, leading to more informed decision-making processes.

To enhance comprehension regarding the application of various Business Intelligence types in real-world scenarios and their integration with artificial intelligence, let us examine a practical case involving a financial institution utilizing AI for fraud detection. The institution follows a systematic use case scenario that incorporates Descriptive, Predictive, and Prescriptive BI methodologies.

- Descriptive BI phase: Historical Fraud Analysis
    - ✓ Prepare Financial Dashboard

The institution initiates the process by developing a comprehensive financial dashboard that combine historical transaction data, including both flagged fraudulent transactions and legitimate ones.

The objective is to gain insights into historical fraud patterns, identifying common characteristics and categorizing past instances of fraud. Through Descriptive BI, the financial institution achieves an understanding of occurrence of fraud in the institution's history, laying the foundation for subsequent BI applications.

- Predictive BI phase: AI-Enhanced Fraud Prediction Models
    - ✓ Implement Predictive BI for Fraud Prediction

Moving forward, the institution utilizes Predictive BI tools equipped with machine learning algorithms to create advanced fraud prediction models.

Predictive BI enables real-time prediction of the likelihood of fraud by analyzing transaction patterns, customer behavior, and other pertinent features. This step allows for proactive identification and intervention in potential fraudulent activities.

The integration of AI with Predictive BI enhances the institution's ability to forecast and address potential fraud scenarios, contributing to a more proactive fraud detection system.

- Prescriptive BI phase: Automated Response and Optimization
    - ✓ Apply Prescriptive BI for Automated Response

In the final step, the institution employs Prescriptive BI together with AI to develop automated response systems triggered upon detecting potential fraud.

Thin step guides the system on optimal actions to take when fraud is identified, such as blocking transactions, alerting authorities, or implementing additional security measures. It ensures swift and effective responses, reducing the impact on the organization and its customers. Automated responses, strengthens the institution's ability to respond promptly and effectively to potential fraud, results in a comprehensive and advanced fraud detection system.

2.5. Endpoint

As a final point, the continuum of business intelligence (BI) encompasses three essential analytical approaches: descriptive, predictive, and prescriptive BI.

Each of the BI clusters, answers different types of questions and serves distinct purposes in extracting insights from data.

- Descriptive BI: It lays the groundwork by offering a retrospective view, summarizing historical data to understand past occurrences and performance trends, addressing the question, “What happened?”
- Predictive BI: Positioned to transform decision-making, predictive BI leverages statistical models and machine learning to forecast future trends and outcomes. It proactively answers the question, "What is likely to happen?" and provides a strategic edge in the competitive business landscape.
- Prescriptive BI: At the forefront of BI, prescriptive analytics not only predicts future outcomes but also prescribes optimal actions. It quantifies the effects of decisions for the best possible results, dealing with the question, "What should we do about it?"

As organizations evolve in the era of advanced analytics, the combination of descriptive, predictive, and prescriptive analytics provides a comprehensive approach to understanding and utilizing data. Descriptive analytics illuminates historical data, predictive analytics forecasts future trends, and prescriptive analytics recommends actionable strategies. This trio contributes to a holistic understanding of business performance, enabling organizations to make informed and strategic decisions. The transformative power of BI, closely linked with AI, equips organizations to navigate the complexities of the ever-evolving business landscape with confidence and foresight.

Machine Learning and Deep Learning, as part of AI technology, further amplify this transformative power. They automate complex decision-making tasks within BI systems, allowing machines to learn from historical cases and acquire problem-solving knowledge. Techniques like Artificial neural networks (ANN), case-based reasoning, genetic algorithms, and natural language processing (NLP) are seamlessly integrated into BI systems, harnessing the capabilities of AI. This strategic integration enhances the toolkit for data-driven decision-making, providing organizations with advanced capabilities to extract meaningful insights from their data. The synergy of BI with AI signifies a powerful paradigm shift in organizational decision-making, emphasizing adaptability and innovation in the face of evolving business challenges.

Section 3

Data and Technical Tools

As we delve deeper into the potentials of data analytics, the landscape becomes increasingly populated with diverse tools designed to empower Business Intelligence (BI). From traditional BI platforms to advanced AI-driven solutions, there are abundant options available, reflecting the growing importance of leveraging technology for informed decision-making.

This chapter is dedicated to exploring the forefront tools of innovation, considering the top tools and technologies available at the time, including Power BI, R, and Python. These platforms offer unique capabilities that contribute to the analysis and decision-making process.

In the following, I will provide an in-depth description of the datasets employed in this research, along with the carefully selected tool set chosen to carry out this study. Additionally, I will discuss the statistical models suitable for conducting this analysis based on similar research and academic articles, the available data and the research objectives. This exploration sets the stage for understanding how these tools, platforms, and models contribute to achieving the research objectives before delving into the details of the business case.

3.1. Data

To achieve the objectives of this study, a real-world dataset is provided. This dataset will be manipulated using a carefully selected set of tools to demonstrate how the integration of AI with BI tools enhances analysis and decision-making capabilities. In the following, I will proceed with discussing the characteristics and structure of the data.

The data used in this study is related to financial data from a distillation company that wishes to remain anonymous. It covers the inflows and outflows of the company over a four-year period from 2020 to 2023.

It's worth mentioning that the data has been anonymized, and all personal information such as names has been removed to ensure confidentiality and privacy.

The dataset consists of three separate Excel sheets:

i. Customers

This sheet contains information about the customers of the distillation company, including customer ID, industry sector, country, month, year, customer internal rating, and external rating for each specific month and year.

The internal rating is calculated based on historical transactions of the customer by the accounting department, ranging from one to seven. The external rating is purchased from a provider and ranges from one to seven as well.

1	Customer	Year	Month	External Rating	Internal Rating	Sector	Country
2	Customer1	2020	Agosto	5	4	Industry	Italy
3	Customer1	2020	Aprile	6	5	Industry	Italy
4	Customer1	2020	Dicembre	7	6	Industry	Italy
5	Customer1	2020	Febbraio	6	5	Industry	Italy
6	Customer1	2020	Gennaio	6	4	Industry	Italy
7	Customer1	2020	Giugno	7	5	Industry	Italy
8	Customer1	2020	Luglio	5	5	Industry	Italy
9	Customer1	2020	Maggio	5	5	Industry	Italy
10	Customer1	2020	Marzo	7	5	Industry	Italy
11	Customer1	2020	Novembre	5	5	Industry	Italy
12	Customer1	2020	Ottobre	5	4	Industry	Italy
13	Customer1	2020	Settembre	5	5	Industry	Italy
14	Customer10	2020	Agosto	5	5	Services	France
15	Customer10	2020	Aprile	5	5	Services	France

Figure 2: The First 15 Rows of Customer Data Illustrating the Data Structure and Column Names

ii. Inflows

Inflows data set represents transactions entering the company, which includes all payments made to the company by the customers purchasing company products. It includes information such as Transaction ID, customer ID, industry sector, due amount, whether the due amount has been collected (paid by the customer), the due date, and the payment date.

1	ID	Customer	Sector	Amount	Cashed	Due Date	Payment Date
2	INFLOW1419	Customer16	Commercial	4,045	YES	23/1/2020	9/10/2020
3	INFLOW1423	Customer9	Industry	4,441	YES	23/1/2020	13/10/2020
4	INFLOW1418	Customer16	Commercial	5,007	YES	23/1/2020	20/10/2020
5	INFLOW1433	Customer10	Services	2,136	YES	23/1/2020	22/10/2020
6	INFLOW1558	Customer2	Commercial	2,353	YES	23/1/2020	25/10/2020
7	INFLOW1545	Customer2	Commercial	2,707	YES	23/1/2020	25/10/2020
8	INFLOW1538	Customer15	Commercial	4,099	YES	23/1/2020	27/10/2020
9	INFLOW1542	Customer15	Commercial	4,531	YES	23/1/2020	27/10/2020
10	INFLOW1414	Customer11	Industry	4,854	YES	23/1/2020	31/10/2020
11	INFLOW1539	Customer18	Commercial	3,163	YES	23/1/2020	31/10/2020
12	INFLOW1543	Customer18	Commercial	3,664	YES	23/1/2020	31/10/2020
13	INFLOW1527	Customer1	Industry	5,334	YES	23/1/2020	1/11/2020
14	INFLOW1550	Customer16	Commercial	2,256	YES	23/1/2020	1/11/2020
15	INFLOW1555	Customer16	Commercial	2,766	YES	23/1/2020	1/11/2020

Figure 3: The First 15 Rows of Inflows Data Illustrating the Data Structure and Column Names

iii. Outflows

Outflows data set includes all the expenses paid by the company, such as buying raw materials, utilities, etc. It contains information such as Transaction ID, the subject of payment (suppliers ID, labor, utilities, leasing, mortgage), the type of transaction (acquisitions, labor, utilities, leasing, mortgage), due amount, due date, and the actual payment date.

1	ID	Subject	Type	Amount	Payment Dat	Due Date
2	OUTFLOW1	Supplier5	Acquisitions	7,578.64	18/2/2020	3/3/2020
3	OUTFLOW2	Supplier6	Acquisitions	7,020.77	20/2/2020	1/3/2020
4	OUTFLOW3	Leasing	Leasing	6,381.23	20/2/2020	2/3/2020
5	OUTFLOW4	Supplier2	Acquisitions	7,832.82	21/2/2020	5/3/2020
6	OUTFLOW5	Mortgage	Mortgage	7,126.76	21/2/2020	7/3/2020
7	OUTFLOW6	Supplier2	Acquisitions	7,565.29	22/2/2020	5/3/2020
8	OUTFLOW7	Supplier6	Acquisitions	7,127.96	23/2/2020	1/3/2020
9	OUTFLOW8	Supplier2	Acquisitions	7,088.65	23/2/2020	5/3/2020
10	OUTFLOW9	Supplier4	Acquisitions	7,931.84	24/2/2020	9/3/2020
11	OUTFLOW10	Supplier5	Acquisitions	8,283.69	25/2/2020	3/3/2020
12	OUTFLOW11	Supplier5	Acquisitions	8,381.25	25/2/2020	3/3/2020
13	OUTFLOW12	Supplier3	Acquisitions	7,721.93	26/2/2020	5/3/2020
14	OUTFLOW13	Supplier1	Acquisitions	6,807.72	26/2/2020	8/3/2020
15	OUTFLOW14	Supplier4	Acquisitions	6,660.41	26/2/2020	9/3/2020

Figure 4: The First 15 Rows of Outflows Data Illustrating the Data Structure and Column Names

3.2. Platforms

In order to utilize the data effectively and achieve the goals described earlier, a carefully selected set of tools has been chosen. Further, I will explore the features of each tool and discuss how they contribute to my analysis and research objectives.

i. Microsoft Power BI

Microsoft Power BI is a comprehensive business analytics tool developed by Microsoft, primarily focusing on empowering businesses. It enables users to visualize, analyze, and interpret data to derive meaningful insights for decision-making. With Power BI, users can create interactive reports and dashboards, offering visually appealing data representations.
A key feature is its ability to connect to various data sources, including databases, cloud services, Excel files, R, Python, and more, facilitating data consolidation for analysis. Power BI provides robust data preparation capabilities through its Power Query Editor, allowing users to shape and clean data before analysis.
Users can leverage advanced analytics features such as forecasting, clustering, and natural language processing for querying data. The platform facilitates collaboration and sharing, allowing secure publication of reports and dashboards to the Power BI service.
Seamless integration with Microsoft products like Excel, Azure, and Office 365 enhances productivity and extends its capabilities within the Microsoft ecosystem. Overall, Power BI is valued for its user-friendly interface, extensive visualization options, data connectivity, and integration capabilities, making it a popular choice for businesses seeking powerful data analytics solutions.
In addition to its advantages, Power BI has massive market reach and has been consistently selected by Gartner's Magic Quadrant as the market leader. This highlights its ability to deliver relevant, context-aware automated insights meeting the needs of business decision-makers and analytic content consumers.

ii. R Programming Language

R stands out as a programming language and environment renowned for its prowess in statistical computing and graphics. With a rich array of statistical techniques and graphical tools, R offers versatility for various data analysis tasks.

Statistical computing tasks, such as data analysis, modeling, and hypothesis testing, find extensive use in R. Its functionality spans from linear and nonlinear modeling to time-series analysis and clustering, catering to diverse analytical needs.

Notably, R is celebrated for its robust visualization capabilities, empowering users to craft a wide array of plots, charts, and graphs with ease. Packages like ggplot2 are favored for creating highly customizable and publication-quality graphics.

In addition to visualization, R provides powerful tools for data manipulation, allowing users to clean, reshape, and preprocess data efficiently. Popular packages like "dplyr" and "tidyr" streamline these tasks seamlessly.

Moreover, R's versatility extends to integration with other programming languages and tools, enabling seamless data analysis across various platforms and interfaces, including databases and web APIs.

Widely regarded as a cornerstone in data analysis, statistics, and research, R's flexibility and rich capabilities make it a preferred choice among data scientists and statisticians.

Furthermore, R's consistent presence in programming language rankings like the TIOBE Index underlines its popularity and widespread usage in statistical computing and data analysis tasks. It is often highlighted in surveys and reports focusing on data science and analytics tools, affirming its importance in the field.

iii. Python Programming Language

Python is a high-level, interpreted programming language celebrated for its simplicity and readability. Its design philosophy emphasizes clear and concise code, allowing developers to express complex concepts efficiently, often in fewer lines of code than languages like C++ or Java.

Python supports multiple programming paradigms, including procedural, object-oriented, and functional programming. Its extensive standard library offers a comprehensive suite of tools for a diverse range of tasks, from web development and data analysis to artificial intelligence and scientific computing. A notable strength of Python is its vibrant and active community, which has cultivated a vast ecosystem of third-party packages and libraries. Popular libraries include NumPy and Pandas for data manipulation, Matplotlib and Seaborn for data visualization, TensorFlow and PyTorch for machine learning, and Django and Flask for web development.

Python's versatility extends to its integration capabilities, allowing seamless interoperability with other languages and technologies. This adaptability, combined with its ease of learning and powerful features, makes Python a preferred choice for both novice and seasoned developers across various fields, including software development, data science, academia, and research. Python's prominence is reflected in its consistent ranking among the top programming languages in various indices. In the TIOBE Programming Community Index, Python frequently holds a leading position, often competing with stalwarts like Java and C. This index gauges the popularity of programming languages based on search engine queries, tutorials, and other online activities.

Additionally, the IEEE Spectrum ranking, which evaluates languages based on criteria such as job postings, search trends, and open-source contributions, regularly places Python at or near the top, indicating its widespread adoption and high demand in the job market.

Python is also highly regarded in specialized surveys, such as the Stack Overflow Developer Survey, where it is frequently listed as one of the most loved and most wanted languages. Its dominance in areas like data science, artificial intelligence, and web development further enhances its standing.

Overall, Python's flexibility, ease of use, and robust community support contribute to its high rankings in these various assessments, solidifying its status as a leading programming language in both general-purpose and specialized domains.

3.3. Integration of Platforms: The reasoning behind it

As discussed earlier, the integration of BI and AI leads to a deeper understanding of data, driving better strategic decisions and improving managerial dashboards' effectiveness.
A report by Gartner highlights the crucial role of intelligence in digital businesses, making analytics and BI top innovation investments for IT and business leaders. According to the report, there will be a more than 25% increase in the use of AI to drive personalized adaptive user interfaces in applications by 2026.

In finance, AI significantly enhances financial forecasting, market analysis, and trend identification, providing valuable insights for investment opportunities. AI aims to create machines capable of human-like thought processes, revolutionizing the financial services industry with applications like algorithmic trading and regulatory compliance. Moreover, AI and FinTech together make financial services more accessible and transparent, benefiting traditionally excluded populations.

Integrating R and Python within Power BI for my thesis provides a compelling advantage by combining advanced statistical and machine learning capabilities with robust data visualization and business intelligence features. R excels in statistical computing and graphics, ideal for tasks like time-series analysis and forecasting, while Python offers strong capabilities in machine learning, AI, and data manipulation.

Using R and Python within Power BI enables seamless integration of these advanced techniques into interactive dashboards and reports, facilitating better understanding and decision-making. This integration allows for real-time data processing, updates, and automation of analytical processes, reducing manual intervention and ensuring consistency. It leverages Power BI's strengths in data visualization while capitalizing on the analytical power of R and Python.

Consequently, at the heart of the analysis lies Power BI, serving as the cornerstone of this research. Power BI will form the foundation upon which this analysis is built, acting as the outline to integrate other tools seamlessly. The focus remains on a financial planning report, demonstrating how AI tools can enhance decision-making and transform managerial dashboards' effectiveness. Thus, this research frames within a business context, specifically financial dashboards, utilizing advanced analytical tools within the same domain.

3.4. Model Selection

This segment serves as an introduction to explore the statistical models suitable for conducting this thesis based on the available data. Regardless of their perceived utility, all models will be tested in the subsequent chapter, and the most suitable models for this case study will then be selected.

Drawing insights from similar research and academic articles and considering the characteristics of the dataset provided, let us explore some models that could be useful for carrying out the analysis:

In order to find a suitable model, certain points need to be taken into consideration. Machine Learning consists of two main branches: Supervised and Unsupervised.

Supervised learning involves training the model on a labeled dataset, where each data point has input-output pairs. This means that for each input data (features), there is a corresponding output label. The model learns to map input data to the correct output during training.

On the other hand, unsupervised learning works with an unlabeled dataset, where input data is provided without corresponding output labels. In unsupervised learning, the model attempts to find patterns, structures, or relationships in the data without explicit guidance.

Since our data is labeled, this chapter will focus on finding suitable models within supervised machine learning only.

Supervised machine learning is divided into two main categories: classification and regression. Classification aims to predict categorical outcomes, where input data is assigned to predefined classes or labels. For example, it determines whether an email is spam or not spam, identifies fraudulent transactions, or classifies images into various categories.

Regression, on the other hand, predicts continuous numerical or binary values. It is applied in scenarios such as forecasting stock prices, estimating house values, or predicting sales revenue based on input features.

Understanding these distinctions is crucial for selecting the appropriate modeling approach and techniques tailored to the specific characteristics of the data and the goals of the analysis. By recognizing whether the task involves predicting categories or numerical values, we can choose the most suitable models and methods for our analysis.

- Logistic Regression

Logistic regression is a powerful statistical model commonly used for binary classification problems. It is applied when the outcome or dependent variable is categorical with only two possible outcomes.

In logistic regression, the model predicts the probability of a binary response based on one or more predictor variables. The dependent variable represents the binary outcome, while the independent variables, also known as predictors or features, can be either continuous or categorical.

This model is widely used in various fields such as medicine, finance, and marketing. It's employed for tasks like predicting whether a customer will churn or not, identifying fraudulent transactions, or determining whether an email is spam or not spam.

- Decision Tree

A decision tree is a model that predicts the value of a target variable by learning simple decision rules. It has a hierarchical, tree-like structure, consisting of a root, nodes, branches, internal nodes, and leaf nodes.

Starting from the root, the decision tree evaluates features and selects splits based on maximum gain, creating partitions. This process repeats recursively until a final leaf node is reached.

In essence, a decision tree is a tool used in machine learning to aid in decision-making or predictions. It resembles a tree, where each question or decision leads to different branches. For example, to determine whether to wear a coat, the tree might first ask, "Is it cold outside?" If yes, the next branch might ask, "Is it raining?" Each answer leads to a final decision.

Decision trees can handle both categorical and continuous variables. Categorical variables are split based on their categories, while continuous variables are split based on threshold values.

More specifically, the tree divides data into groups based on answers to questions about the data. This continues until it reaches a final decision or prediction at the end of each branch. Decision trees are intuitive and easy to visualize but may create overly complex rules that generalize poorly to new data. Techniques like pruning (trimming back the tree) and ensemble methods (e.g., Random Forests) can enhance their accuracy and generalization ability.

- Random Forest

Random Forest is an enhanced version of the standard tree-growing algorithm with several key features:

- Bagging: It improves the accuracy of single decision trees by using bagging, which involves building multiple trees on different subsets of the training data and averaging their predictions.
- Fully Grown Trees: Trees in Random Forest are grown until they reach pure leaves for maximum accuracy.
- Random Input Selection: During node splitting, Random Forest randomly selects a subset of input features. This randomness ensures that each tree uses different features, reducing bias and introducing variance.

Each node is built on a small random subset of the input features, ensuring diversity among trees. This helps prevent overfitting and improves generalization.

Similar to Decision trees, Random Forest can handle both categorical and continuous variables. Categorical variables are split based on their categories, while continuous variables are split based on threshold values.

Overall, Random Forest combines the predictions of multiple trees to create a more robust and accurate model for various machine learning tasks.

- Autoregressive Integrated Moving Average (ARIMA)

ARIMA models are commonly used in time series analysis for forecasting. They predict future values of the time series based on past observations and help understand and capture the underlying patterns and dynamics in the time series data.

In these models, the dependent variable typically represents time series data, such as cash flow, expenses, or ratings, while the independent variable corresponds to time steps, which are usually equally spaced (meaning annual, monthly, daily and etc.)

ARIMA models aim to capture autocorrelation in the data by incorporating three key components:

- Autoregressive (AR) Component: It involves linear regression of the time series against its lagged values to capture serial correlation.
- Integrated (I) Component: This component involves differencing the time series data to make it stationary, removing trend and seasonality.
- Moving Average (MA) Component: It is a weighted average of error terms from previous time steps to capture random shocks or noise.

ARIMA models can handle both univariate and multivariate time series data.

Univariate time series involves only one variable observed sequentially over time.

Multivariate time series involves multiple variables observed sequentially over time, where each variable may depend on its own past values and the past values of other variables.

ARIMA models are widely used in various fields such as finance, economics, and weather forecasting for their ability to capture complex temporal patterns and make accurate predictions.

- Autoregressive Moving Average (ARMA)

ARMA model is suitable for analyzing and forecasting stationary time series data, meaning the statistical properties such as mean and variance remain constant over time.

It combines autoregression (AR) and moving average (MA) components to capture the temporal dependencies in the data and model the relationship between an observation and a linear combination of lagged observations and lagged forecast errors.

- Evaluation Methods

Below is listed some evaluation methods provide insights into different aspects of model performance for regression, classification and time series models, aiding in model selection and assessment.

- **Accuracy:** Measures the proportion of correctly classified instances among the total instances in classification tasks. (Classification)
- **Precision:** Proportion of true positive predictions among all positive predictions. Important for imbalanced classes in classification tasks. (Classification)
- **Recall (Sensitivity):** Proportion of true positive predictions among all actual positives. Crucial when false negatives are costly in classification tasks. (Classification)
- **F1 Score:** Harmonic mean of precision and recall, balancing both metrics in classification tasks. (Classification)
- **ROC Curve and AUC (Area Under the Curve):** ROC curve plots true positive rate against false positive rate. AUC represents the model's ability to distinguish between classes in classification tasks. (Classification)
- **Confusion Matrix:** Shows true positives, false positives, true negatives, and false negatives in classification tasks. (Classification)
- **Mean Absolute Error (MAE):** Average absolute difference between predicted and actual values in regression tasks. (Regression, Time Series)

- **Mean Squared Error (MSE):** Average of squared differences between predicted and actual values in regression tasks. (Regression, Time Series)
- **Root Mean Squared Error (RMSE):** Square root of MSE, providing interpretable measure of prediction error in regression tasks. (Regression, Time Series)
- **R-squared (R2):** Measures proportion of variance in the dependent variable explained by the independent variables in regression tasks. (Regression)

Section 4

Case Study and Methodologies

In this chapter, I will describe the processes involved in manipulating the data, the various models used, and the underlying logic guiding these choices. This section might initially appear to overcomplicate the approach by incorporating multiple tools and methodologies. However, it's essential to provide context for why these tools were chosen and how they contribute to achieving the project's goals.

The primary tool for this project is Power BI, chosen for its capability to present data interactively and intuitively. Unlike static presentations or raw HTML outputs, Power BI allows stakeholders, who may lack coding or statistical expertise, to engage with the data dynamically. This interactivity is crucial for decision-making, enabling managers to explore various scenarios effortlessly through selections within the dashboard.

Integration of R and Python within Power BI enhances the depth and utility of the report. These tools are employed to perform sophisticated data manipulations and analyses that are beyond the native capabilities of Power BI. By embedding these scripts into the Power BI environment, we can leverage advanced statistical methods and machine learning models, making the dashboard not only informative but also predictive.

The practical part of this thesis begins with the development of a Power BI financial planning dashboard. Importing the dataset into Power BI marks the first step. Data formatting is crucial as Power BI sometimes fails to detect the correct data types automatically. Therefore, upon uploading the dataset, all data formats were precisely adjusted based on the content of each column to ensure accuracy and consistency.

One of the fundamental aspects of creating a Power BI dashboard is establishing the data model. In Power BI, the data model organizes and structures data by defining relationships between different tables. This enables users to perform complex analyses and visualizations without the need for manual data merging. The data model supports relationships, data types, and hierarchies, providing a foundation for advanced data manipulations using calculated columns, measures, and DAX expressions.

For this specific project, the data model follows a snowflake schema. The 'inflows' table acts as the fact table, connected to the 'customers' table by customer ID and to 'outflows' by payment date. Since the data is monthly, ensuring a complete calendar for the study period is essential. A separate date table was created, containing unique date values from the year 2020 and connected to the payment date of the 'inflows' table. This ensures comprehensive time series visualizations and captures all transaction dates accurately.

The next step involved creating the 'Number of Delayed Days (NDD)' column, calculating the difference between the due date and the payment date. Null values related to unsettled transactions were assigned a placeholder value of 999 for consistency.

Similar procedures were applied to the 'customers' and 'outflows' tables, ensuring consistent data formatting and completeness.

After data cleaning, simple calculations were performed using DAX measures to facilitate visualization. Financial calculations such as summing the inflows and outflows were conducted. Additionally, the 'inflows' were categorized into actual and expected invoice groups based on settlement status. Another grouping has been made for detailed analyzing the data based on location (Italy or abroad).

Finally, different sheets were created within the dashboard, each dedicated to analyzing various aspects of the data including an overview, inflows, outflows, and customer analysis. Multiple related visuals were employed to illustrate trends and proportions effectively.

Up to this point, the development follows a traditional BI application approach, laying the groundwork for more advanced analyses and predictive modeling in subsequent stages.

Before starting the advanced analysis, some assumptions were made that form the ground work of the analysis.

One of the primary assumptions made is that I assumed that there might be a relation between the customer being delayed, its internal and external ranking and the due amount as well as the industry and country of origin.

Another assumption is that there should be some patterns of the type seasonality or trend within the data since it is a time related financial data.

These assumptions are the basis of model selection and tool selection to carry out this analysis. Consequently, it is essential to acknowledge these assumptions and recognize the limitations of the chosen approach.

As highlighted throughout previous chapters, the ultimate goal is to integrate R and Python within Power BI to create an advanced, interactive dashboard that facilitates decision-making for managers and stakeholders. By leveraging these tools, we can unlock the full potential of the data for predictive analytics.

To achieve this, a set of packages have been utilized including NumPy, Pandas, Matplotlib, and Seaborn, each offering unique capabilities.

NumPy is fundamental for numerical computing in Python, providing powerful tools for working with arrays and matrices. It offers efficient numerical operations and mathematical functions, making it essential for data manipulation and numerical computations.

Pandas is a versatile library for data manipulation and analysis. It provides data structures like DataFrame and Series, allowing for easy handling of structured data. Pandas offers functionalities for data cleaning, filtering, grouping, and merging, making it indispensable for data preprocessing and analysis tasks.

Matplotlib is a comprehensive library for creating static, interactive, and publication-quality visualizations in Python. It offers a wide range of plot types and customization options, making it suitable for generating various charts, plots, and graphs to visualize data effectively.

Seaborn is built on top of Matplotlib and specializes in statistical data visualization. It provides high-level interfaces for creating informative and attractive statistical graphics. Seaborn simplifies the process of creating complex visualizations such as categorical plots, distribution plots, and regression plots, enhancing the presentation of data analysis results.

Then, the 'inflows' and 'customers' data tables were merged. Although Python and R alone are flexible with data types, their integration within Power BI requires strict adherence to data formatting rules. This ensures seamless functionality within the Power BI environment.

The handling of null values and outliers was consistent with the earlier steps, but with additional columns introduced for further analysis. One such column is 'delay,' indicating whether a customer was late in paying their debt. This was derived by comparing the due date with the payment date, with a delay indicated by a value of 1. To balance the dataset, delays of up to 15 days were categorized as non-delayed (0). Categorical columns were also converted to numeric formats to facilitate analysis and modeling.

Python was employed to predict customer payment behavior, specifically the likelihood of a customer paying their due amount on time. This required testing and implementing several predictive models. However, due to certain limitations and strategic decisions, the analysis was conducted directly in the Python application rather than within Power BI.

There are two main reasons for this approach:

1. **Power BI's Limitations**: Power BI supports the integration of Python scripts for data manipulation and visualization. However, it has specific requirements for how these scripts are utilized. Any data manipulation or modeling performed using Python within Power BI must result in a visual output. This means that while it is possible to fit models using Python, the results need to be visualized within Power BI, such as through ROC curves or residual plots. Additionally, when working with Excel data in Power BI, Python scripts can be used to manipulate data, but the outputs must be incorporated into a visual format within the Power BI report. This limitation makes it impractical to use Power BI solely for backend model fitting without visualization.
2. **Managerial Focus**: The primary goal is to enhance the managerial dashboard, not to display the step-by-step analysis within Power BI. Displaying the intricate modeling process would confuse managers and stakeholders, who need clear, actionable insights rather than technical details. Additionally, troubleshooting within Power BI is more challenging than in Python. The main objective is to predict whether a customer will be delayed in their payment, making the binary 'delay' column the dependent variable.

The next step is to identify the best features for the models.

- **Logistic Regression** Logistic regression is a foundational model for binary dependent variables. It models the relationship between predictor variables and the binary outcome, which in our case is the 'delay' in payments. The independent variables included industry sector, due amount, settlement status, external rating, internal rating, and country.

The dataset was split into a training set (70%) and a testing set (30%) to evaluate accuracy and precision. After training the model, ROC curves and confusion matrices were plotted to assess model performance. The ROC curve helps visualize the trade-off between sensitivity and specificity, while confusion matrices provide insight into the model's classification performance.

- **Decision Tree** A decision tree is suitable for capturing non-linear relationships between independent and dependent variables. Similar to logistic regression, 'delay' was the dependent variable, and the same set of independent variables was used. The dataset was split into training and testing sets, and ROC curves and confusion matrices were used for evaluation.

  Decision trees provide a clear understanding of how decisions are made by partitioning the data based on feature values. They are interpretable and can handle both numerical and categorical data. Evaluation metrics such as accuracy, precision, recall, and F1-score were computed to assess model performance.

- **Random Forest** Random Forest is an ensemble learning method that builds on decision trees to create a more robust predictive model. It reduces overfitting, handles complex interactions, and provides feature importance.

  The feature selection and modeling procedure were consistent with logistic regression and decision tree models. Random Forest works well with both categorical and continuous data, making it suitable for our dataset. Evaluation metrics such as accuracy, precision, recall, and feature importance were analyzed to assess model performance.

Given the time-related nature of the dataset, leveraging R and its powerful forecasting tools is essential for accurate future projections. For this purpose, I utilized the fpp3 library in R, which offers a comprehensive and user-friendly toolkit designed specifically for time series analysis and forecasting. The fpp3 library seamlessly integrates with the tidyverse, allowing for efficient data manipulation and visualization while supporting a variety of forecasting models, including exponential smoothing, ARIMA, state space models, and machine learning methods. Its key features include handling complex seasonal patterns, automatic model selection, and robust forecast accuracy evaluation. The library's flexibility, ease of use, and incorporation of state-of-the-art methods make it an invaluable resource for producing reliable forecasts across diverse scenarios.

As with Python, not all analytical steps need to be displayed within the Power BI app, so the detailed analysis was conducted in R. To manipulate and forecast data using fpp3, it is essential to structure the data as a tsibble. A tsibble is a powerful data structure for managing time series data in R, combining the capabilities of traditional time series objects with the flexibility and ease of use of tibbles. It supports both regular and irregular time series, integrates seamlessly with the tidyverse, and provides robust handling of multiple time series and missing data. This makes it an essential tool for time series analysis within the fpp3 forecasting framework.

After creating a tsibble and the necessary time steps, several plots are used to observe potential seasonality or trends in the data. One such plot is the Autocorrelation Function (ACF), which helps identify repeating patterns and detect seasonality by displaying significant correlations at specific lags. High spikes in the ACF plot indicate significant correlations, and repeating high spikes at fixed-length lags suggest seasonality.

Another valuable tool is the lag plot, which helps identify relationships between lagged values of the series. A lag plot can reveal non-randomness in the data, indicating that past values influence future values. A noticeable pattern at fixed-length lags in the lag plot suggests seasonality.

Once seasonality is observed, moving averages are used to estimate the trend at time ttt. Moving averages of the order of the periodicity of the seasonality help smooth out the seasonal component, allowing for the identification of the trend-cycle component.

Exponential smoothing is then employed to produce forecasts based on weighted averages of past observations, with more recent observations given heavier weights. This method is adaptable for data with or without trend and/or seasonal patterns. Simple exponential smoothing, for instance, balances giving more weight to recent observations without entirely discarding past data. This approach produces forecasts that are more relevant and reflective of recent trends while acknowledging the influence of past observations. Although this method takes into account all past data, it adjusts the weights so that earlier observations are progressively less influential, making the forecast more responsive to recent changes.

Once the analysis is complete, I will incorporate various visuals in Power BI to display the predictions. These visuals will be designed to be as interactive as possible, enabling the audience to inspect different case scenarios and gain deeper insights. By doing so, managers and stakeholders will be able to explore the forecast data dynamically, enhancing their ability to make informed decisions based on the predictive models.

Through the use of interactive visuals in Power BI, the complex results of the R-based forecasting and Python based prediction models will be presented in a user-friendly format. This approach ensures that the technical depth of the analysis is leveraged to provide clear, intuitive insights, thereby facilitating better decision-making for all stakeholders involved.

Section 5

Results and Discussions

In this chapter, we delve into the outcomes obtained from applying the various models and methodologies discussed in the previous chapter to our financial dataset. The primary focus is on analyzing the predictive performance of the models and extracting actionable insights to aid decision-making. By exploring the results in detail, I aim to evaluate the effectiveness of the predictive analytics approach employed in this study and discuss its implications for financial planning and management.

The predictive models, including logistic regression, decision trees, random forest, and time series forecasting using R, were applied to address different aspects of the financial data. Now, we examine how well these models performed in predicting customer payment behavior, identifying important factors influencing delays, and forecasting future trends. Additionally, we discuss any challenges encountered during the analysis and potential areas for improvement.

Finally, the effect of these analyses on the overall informativeness and usability of the Power BI dashboard, highlighting how the insights gained contribute to better decision-making for financial planning and management stakeholders will be discussed.

To begin with, the Power BI dashboard starts with a traditional business intelligence section consisting of four sheets: Overview, Inflows, Outflows, and Customers.

Overview: The Overview sheet provides a snapshot of the data, including total inflows, total outflows, net flow, and their trends over time. Users can quickly grasp the financial performance at a glance and identify any notable trends or fluctuations.

Inflows: The Inflows sheet delves deeper into the details of the inflows, showcasing actual inflows, expected inflows, inflows from Italy and abroad, and the proportion of inflows based on sector and country. Additionally, visualizations include the monthly average delay, providing insights into payment delays over time.

Outflows: Moving to the Outflows sheet, users can explore detailed information related to outflow transactions. Visualizations include outflow trends, proportion of outflows by category, outflows by category for each quarter, and the top 5 suppliers by outflows. This helps in understanding expenditure patterns and vendor relationships.

Customers: The Customers sheet addresses details related to customers, offering valuable insights into customer behavior and relationships. A map visualization displays the amount of inflows by country, with bubble size representing the inflow amount. Additionally, users can view the top 5 customers by inflow and the median of internal and external ratings, providing an overview of customer importance and satisfaction levels.

This traditional BI approach lays the foundation for deeper analysis and predictive modeling discussed in subsequent sections, providing stakeholders with structured reports and visualizations to support decision-making.

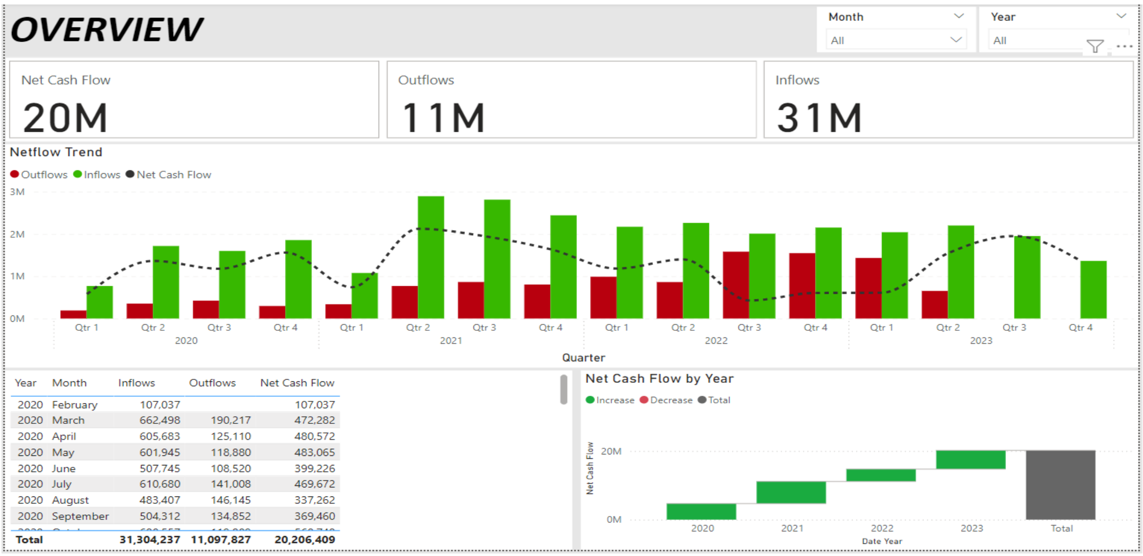

Figure 5: An example of traditional Power BI dashboard

After conducting time series analysis in R, including plotting the autocorrelation function, lag plot, and seasonal plot, it became evident that the data does not exhibit any significant seasonality or trend of fixed-length. In other words, the data appears to resemble white noise. Further investigation revealed that only a segment of the data had been downloaded randomly from the company's database. As a result, precise prediction based on this data would be challenging, and the range of prediction would be as wide as if a random number had been selected.

Below is the prediction of the net flow for the next six months, which shows a wide range of possible outcomes. The prediction intervals at the 80% and 95% confidence levels are also displayed.

The 80% prediction interval provides a range within which we expect the net flow to fall 80% of the time, giving a narrower but still reliable range. Conversely, the 95% prediction interval is wider, providing a range within which we expect the net flow to fall 95% of the time. This wider range accounts for more variability and uncertainty in the forecast, offering a higher degree of confidence that the actual net flow will fall within this interval.

The significant width of these intervals highlights the high degree of uncertainty in the predictions, which is likely due to the data quality issues and lack of discernible patterns or trends in the dataset. This uncertainty reinforces the need for more accurate and comprehensive data to improve the reliability of future forecasts.

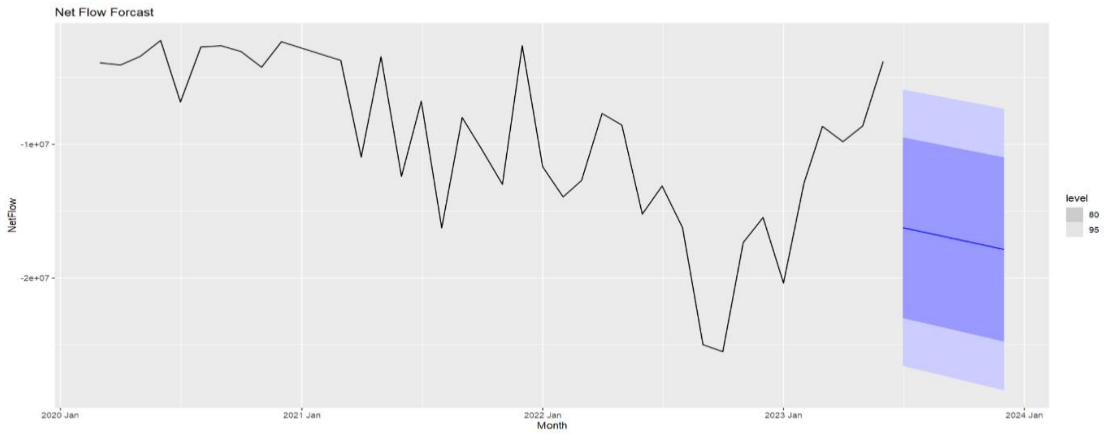

Figure 6: Linear Plot of Net Flow prediction for a period of 6 months

However, data integrity issues also affected the results of correlation between features and model fitting.

Below is the correlation matrix heatmap illustrating the relationships between the dependent variable (delay) and the independent variables (industry sector, due amount, settlement status, external rating, internal rating, and country). The heatmap reveals approximately no significant correlation between the dependent variable and these independent variables, indicating that the assumed predictors may not be strongly related to predicting customer payment delays. This lack of correlation contributes to the poor performance of the predictive models, highlighting the need for further data exploration and the identification of more relevant features to improve model accuracy and reliability.

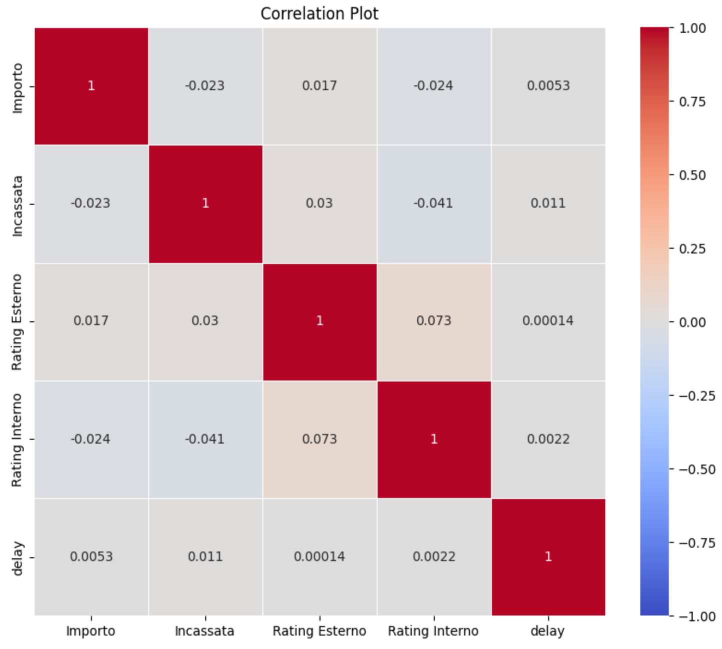

Figure 7: Correlation Matrix heatmap showing the relationship between 'delay' dependent variable and assumed related features

The accuracy of all three models—logistic regression, decision tree, and random forest—is around 50%, which is not an acceptable range for prediction. This suggests that the probability of a customer being late with payments is approximately 50%, similar to random chance.

Below is the ROC curve of the logistic regression model fitted on the data, with an accuracy of 0.57. The area under the curve (AUC) shows that the model performs slightly better than random guessing.

The accuracy of 0.57 means that the model correctly predicts whether a customer will be late with payments approximately 57% of the time. However, since the AUC is close to 0.5, it indicates that the model's discriminatory power is weak, and it performs only slightly better than random chance.

This suggests that while the model makes predictions slightly better than chance, it is not reliable enough for practical use without further improvement or additional data.

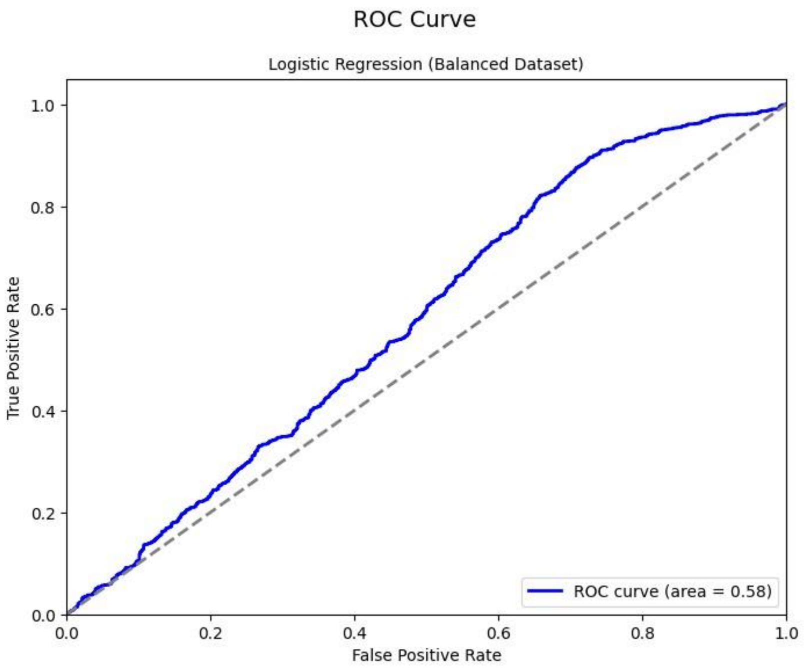

Figure 8: Roc Curve Plot as an evaluation of the Logistic Regression Model fitted in Python weak functionality of the model

Additionally, cross-validation was performed to obtain a better estimate of model performance. The cross-validated scores indicate significant variability in the model's performance, with scores ranging from [0.52, 0.43, 0.68, 0.58 and 0.46], suggesting either model instability or data quality issues.

Given the consistent variability in cross-validation scores across all three models, it appears that there is a data quality issue. This could also imply that the features used to predict the dependent variable (payment delay) are irrelevant, as indicated by the correlation plot earlier.

Despite the limitations encountered in the analysis, let's explore the potential impact on the dashboard if accurate forecasting of future inflows and outflows were possible. While the current models may not yield precise predictions due to data quality issues, considering the hypothetical scenario of accurate forecasting highlights the significance of advanced analysis techniques.

Regardless of the accuracy of the models and predictions with the data available, the method to conduct advanced analysis and integrating R and Python with Power BI is correct. Despite the challenges faced, leveraging R and Python within Power BI for enhanced data analysis and visualization remains valid. This integration opens up possibilities for more in-depth insights and better decision-making.

With a complete and integrated dataset, the results could be more insightful and actionable. The potential benefits of using R and Python within Power BI are evident, despite the current limitations.

This discussion emphasizes the importance of advanced analytics and the integration of R and Python with Power BI in financial planning and management. Despite the challenges posed by data quality issues, the approach presented in this study lays the groundwork for future research and improvements.

Section 6

Conclusions and Recommendations

In this section, we summarize the insights gained from the analysis, discuss the limitations encountered, and provide recommendations for future work. The integration of R and Python within Power BI was explored to enhance the capabilities of a managerial dashboard, illustrating how advanced analytical tools can be seamlessly combined with powerful data visualization techniques. This study aims to shed light on the practical implications of such integration, providing a comprehensive understanding of both the benefits and challenges involved.

During the analysis, it became evident that conducting advanced analysis simultaneously in R or Python, while using Power BI for visualization, is essential. While Power BI excels in visualizing data, it's not the optimal tool for advanced analysis, necessitating data manipulation in R or Python first. Despite this redundancy, integrating R and Python with Power BI demonstrates the potential benefits of enhancing a managerial dashboard's depth and insight.

One significant limitation of this study was the dataset's quality, lacking integration and scraped randomly from the company database. This led to absent seasonality and suboptimal model performance. However, the study illustrates how integrating R and Python within Power BI can provide a more comprehensive analysis.

If a complete dataset had been used, the results would likely have been more accurate and insightful. Nevertheless, this study serves its purpose by showcasing the process and potential benefits of incorporating advanced analytical tools into Power BI, regardless of model accuracy and prediction, which is not the primary goal of this thesis.

Advantages:

- Enhanced analytical capabilities: R and Python offer advanced statistical methods and machine learning models that can significantly improve the depth of analysis.
- Flexibility in data manipulation: Both R and Python provide extensive libraries for data manipulation, allowing for more complex data handling tasks.
- Seamless integration: Power BI allows for the embedding of R and Python scripts, making it possible to combine advanced analysis with intuitive data visualization.

Disadvantages:

- Data handling redundancy: In some cases, it may be necessary to manipulate data twice, once in R or Python and again in Power BI.
- Complexity: Integrating these tools can add complexity to the analysis process, requiring additional expertise and resources.
- Performance limitations: Power BI may have limitations in handling very large datasets or highly complex models, which can impact the performance of the integrated solution.

The results drawn upon similar analysis and integration demonstrate the transformative impact of AI in financial planning, revolutionizing the industry's practices. Key milestones include the adoption of various AI technologies, significantly improving analytical capabilities and strategic decision-making.

Future work should focus on improving data quality, exploring more sophisticated modeling techniques, and refining the dashboard for accurate and valuable insights in financial planning and management. As I continue to refine my approach, integrating advanced analytics with Power BI will play a crucial role in unlocking the full potential of financial data for organizations.

In conclusion, despite challenges, integrating R and Python within Power BI offers significant potential for enhancing data analysis and visualization in managerial dashboards. Future work should focus on ensuring data integrity and completeness to fully realize these tools' potential. This study marks the beginning of a research journey aimed at exploring and optimizing the integration of these tools for improved decision-making processes.

References

[1] L. Zavarella, Extending Power BI with Python and R, Birmingham, UK: Packt Publishing Ltd., November 2021.

[2] S. Weidman, Deep Learning from Scratch: Building with Python from First Principles, O'Reilly.

[3] V. Vashisht, N. Jakhmola, P. Manjarwar and N. Nikhil, "An Effective Approach for Integrating Microsoft Power BI Application with Python for Predictive Analytics," in *Micro-Electronics and Telecommunication Engineering; Proceedings of 4th ICMETE*, Springer, 2020, p. 627.

[4] J. VanderPlas, Python Data Science Handbook, O'Reilly, 2021.

[5] K. Schlegel, "Magic Quadrant for Analytics and Business Intelligence Platforms," *Gartner,* April 2023.

[6] B. Moghaddam and M. Zohuri, "From Business Intelligence to Artificial Intelligence," *Journal of Material Sciences & Manufacturing Research,* 2020.

[7] B. Marr, "15 Amazing Real-World Applications Of AI Everyone Should Know About," *Forbes,* May, 2023.

[8] R. A. Khan, "Business Intelligence: an Integrated Approach," *Business Intelligence Journal,* January, 2012.

[9] S. K, A. K. Nakkella and N. S. Ommi, "Role and Applications of Artificial Intelligence in Business Analytics: A Critical Evaluation," in *Contemporary Research in Environmental Science, Management, IT, Pharmaceutical and Social Sciences*, Michigan, USA, Selfypage Developers Pvt Ltd., December 2023.

[10] R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice, Australia.

[11] F. Fawcett and T. Provost, Data Science for Business, O’Reilly Media, 2013.

[12] E. Cebotarean, "Business Intelligence," *Journal of Knowledge Management, Economics and Information Technology,* 2011.

[13] L. CAO, "AI in Finance: Challenges, Techniques and Opportunities," *University of Technology Sydney,* vol. Vol. 1, no. No. 1, p. 40, June 2021.

[14] S. S. Cao, W. Jiang, L. (. Lei and Q. (. Zhou, "Applied AI for finance and accounting: Alternative data and opportunities," *Pacific-Basin Finance Journal,* April 2024.

[15] H. Canitz, "Descriptive, Predictive and Prescriptive Analytics Explained," *LinkedIn,* Augest, 2023.

[16] "Prescriptive Analytics: From Insight to Action - Bridging the Gap with Business Intelligence," *Virtux BI Solutions,* October, 2023.

[17] "Five Major Trends Shaping the Evolution of Analytics and Business Intelligence," *Gartner,* October 2019.

[18] "Exploring the Evolution of Business Intelligence: From Descriptive to Predictive Analytics," *Virtux BI Solutions,* July, 2023.

[19] "5 Top Use Cases for AI in Corporate Finance," *Gartner,* October 2022.

[20] L. D. Oyeniyi, C. E. Ugochukwu and N. Z. Mhlongo, "Transforming Financial Planning with AI-Driven Analysis: A Review and Application," *Finance & Accounting Research Journal,* vol. 6, no. 4, p. 22, 2024.