esprit

Se former autrement

HONORIS UNITED UNIVERSITIES

2023 / 2024

Software Engineering program – Data Science

SECOND-YEAR ENGINEERING CAPSTONE PROJECT – Intelligent Trading Agent

Authors :

- Chayma RHAIEM
- Zeineb BENFREDJ
- Aziz BEN ROMDHAN
- Yassine TRAIDI
- Ghofrane BEN RHAIEM
- Mohamed Jasser CHTOUROU

VALUE

This page intentionally left blank.

Acknowledgments

We would like to extend our deepest gratitude to **VALUE Digital Services** for entrusting us with the opportunity to contribute to their groundbreaking project. Their unwavering support, visionary guidance, and collaborative spirit have been instrumental in shaping our journey and fostering an environment of innovation and excellence.

We are immensely thankful to **Ms Sarra ZOUARI** and **Mr. Abderrahmen AROUS** for their invaluable mentorship and encouragement throughout this endeavor. Their expertise and commitment to excellence have been a constant source of inspiration.

We would also like to congratulate every member of the project team for their tireless dedication, passion, and camaraderie. Their collective efforts, creativity, and resilience have fueled our progress and transformed obstacles into opportunities.

Contents

List of Figures

List of Tables

List of Acronyms

Natural Language Processing
Natural Language Processing (NLP)

Large Language Model
Large Language Model (LLM)

Sustainable Development Goals
Sustainable Development Goals (SDG)

Key Performance Indicator
Key Performance Indicator (KPI)

Relative Strength Index
Relative Strength Index (RSI)

Named Entity Recognition
Named Entity Recognition (NER)

Exploratory Data Analysis
Exploratory Data Analysis (EDA)

Dividend Growth Rate
Dividend Growth Rate (DGR)

InterQuantile Range
InterQuantile Range (IQR)

open-ended investment company Société d'Investissement à Capital Variable
open-ended investment company Société d'Investissement à Capital Variable (SICAV)

Dividend Liquidity
Dividend Liquidity (DL)

1

Introduction

In the contemporary world, we see an increasing growth and development of the business landscape. New markets are constantly being integrated and many companies are becoming more and more competitive. This creates the need for financial and technological guidance and assistance. Many clients are in need of help when it comes to tasks like formulating and executing their business strategies, particularly in the areas of digital transformation and data management. Especially in the Tunisian market, as businesses increasingly embrace digitization for competitiveness, understanding the opportunities, challenges and the local business environment becomes necessary for stakeholders. In this report, we delve into the impact of digital services in financial and business consulting in Tunisia, as we work on this Intelligent Trading Agents project, supervised by the company Value. We explore the significance of helping clients define their digital and data strategies, develop business plans, and ensure the successful execution of these plans to achieve their business objectives.

2

Business Understanding

This chapter lays the groundwork for our project in intelligent trading portfolios within the Tunisian financial landscape. We delve into the project's context, challenges, and proposed solutions. Additionally, we will conduct a comparative study, focus on the state of the art in quantitative finance, and explore existing solutions in the market.

2.1 Project Scope

This project stands as the capstone integration project for students in the second year of the Software Engineering program at ESPRIT, specializing in the Data Science option. It represents a culmination of theoretical knowledge and practical skills acquired throughout the academic year. By delving into the intricate realm of quantitative finance and trading strategies,

2.1.1 Hosting Organisation

Within the dynamic landscape of digital innovation stands Value Digital Services, a Tunisian "Digital Native" company and consultancy established in 2019. Our collaboration with Value for this project is rooted in a shared commitment to creating enduring value for both the economy and society. As more than a service provider, Value serves as a nurturing ground for skills and expertise, fostering an environment where imagination, ethics, engagement, and excellence are paramount.

Figure 2.1: *Value Digital Services Logo [1]*

With a profound organizational culture, Value Digital Services operates through four specialized departments:

Digital Factory:

Crafting digital solutions from needs assessment to product maintenance.

Data Factory:

Specializing in data and big data projects, providing comprehensive data strategies.

AI Services:

Guiding clients in analytical use cases through artificial intelligence models, including predictive and analytical solutions.

Strategic Advisory:

Offering strategic guidance to clients in areas like marketing and finance.

Our project unfolds within the AI for Capital Markets team, situated in the AI Services department at Value Digital Services. This strategic positioning aligns with the expertise required for the development of intelligent trading portfolios within the Tunisian financial domain.

2.2 Project Context

Tunisia's financial sector is undergoing a rapid digital transformation, marked by increased data-driven decision-making and the need for sophisticated trading solutions. In this context, it is crucial to navigate the intricate demands of the Tunisian financial market. With a specific focus on the development of intelligent trading portfolios in order to provide tailored solutions that optimize capital allocation, manage risks effectively, and elevate overall portfolio performance.

This initiative delves into the intricacies of quantitative finance, trading, and portfolio optimization, addressing the dynamic risks associated with trading financial assets. Unlike traditional banking functions, our project adapts principles to the fluid environment of financial markets, where challenges involve predicting and managing risks.

2.2.1 Tunisian Market Analysis

The Tunisian stock market presents unique characteristics that our project considers:

- Growth and Digitalization: The banking sector in Tunisia has experienced significant growth and digitalization, paving the way for technological advancements in financial services. Our project aligns with this trend by introducing AI-powered trading solutions to further modernize the market.
- Limited Liquidity: Compared to more developed markets, the Tunisian market suffers from limited liquidity. Our intelligent agents aim to address this challenge by facilitating more efficient trades and potentially attracting new investors seeking liquidity.

- Investor Confidence and Security: Building investor confidence and ensuring security are crucial aspects of the Tunisian market. Our project adheres to rigorous security standards and transparent algorithms to foster trust among investors utilizing our intelligent agents.

2.2.2 Problem Statement

Despite its pivotal role in shaping financial landscapes, quantitative finance encounters critical hurdles in the context of the Tunisian stock market:

Deficiency in Decision Support Tools:

Tunisian investors currently lack access to sophisticated, data-driven decision-support tools. The absence of comprehensive market and company analysis tools forces investors to rely on empirical decisions, hindering their ability to make informed choices.

Inefficiencies in Identifying Opportunities:

The absence of automated financial monitoring tools results in delayed recognition of high-potential assets within the Tunisian market. This inefficiency hampers investors' ability to make timely and strategic investment decisions, impacting overall portfolio performance.

Inadequate Risk Management in a Volatile Market:

The Tunisian financial market is characterized by notable volatility. Addressing this requires the implementation of advanced risk assessment and diversification strategies. Currently, the lack of such strategies poses challenges for effective financial risk management.

Addressing these challenges becomes imperative for equipping investors with robust decision-making tools, enhancing market efficiency, and successfully navigating the complexities of the dynamic Tunisian financial landscape.

2.2.3 Objectives

Main Objective:

The primary goal of the project is to create portfolios that exhibit maximum returns, minimal volatility, and a defined level of diversification. The developed trading strategies should outperform a designated benchmark. The major challenge lies in risk minimization, detecting complex patterns within assets while adhering to portfolio diversification. This necessitates the use of theoretical and data-driven mathematical models.

Proposed Solutions

Value recognizes the importance of AI and Data Science in their activities, and that is why they have provided a large amount of data for this project. The data is mainly divided into financial data and market data, and another set of data that we will scrape from finance and trading articles.

This project aims to incorporate the data generously provided to leverage supervised and unsupervised learning to extract and train models that can help clients find the best financial strategies to manage their financial operations and have predictions on when and how to engage in investments in the Tunisian market.

Key Project Components

Data Collection Module:

Develop an automated module to collect traditional financial data and news from diverse sources, storing them in an appropriate database.

Financial Indicators Calculation Module:

Create a module that calculates performance and risk indicators for Tunisian market data and stocks of listed Tunisian companies.

Company Clustering Module:

Implement a module that utilizes machine learning algorithms to detect internal groupings of listed Tunisian companies based on their revenues and volatility.

Trading Strategies Module:

Develop a module that allows users to choose specific trading strategies tailored to their investment needs.

Visualization Module:

Design a module to visually present the results from the preceding modules to users, aiding them in decision-making.

Specific Portfolio Requirements:

Maximum Returns or Predefined Returns Minimum Volatility or Predefined Volatility Portfolio Diversification exceeding a Given Benchmark Consideration of Overall Market Risk Web-based Dynamic Dashboard: Implement a dynamic website with user-friendly dashboards, to enhance the user experience and facilitate efficient decision-making.

Additionally There is another aspect of this project that we will apply to serve our third main goal. Essentially, we will use the data that we will scrape from various articles

and websites to create an informative system on the different elements of the Tunisian market through NLP and Generative AI. This is a very broad project that encompasses many components of intelligent trading, and it will be beneficial in multiple aspects of the world of financial trading.

2.3 Methodology

2.3.1 CRISP-DM

CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining, is an industry-proven way to guide your data mining efforts.

As a methodology, it includes descriptions of the typical phases of a project, the tasks involved with each phase, and an explanation of the relationships between these tasks.

As a process model, CRISP-DM provides an overview of the data mining life cycle.

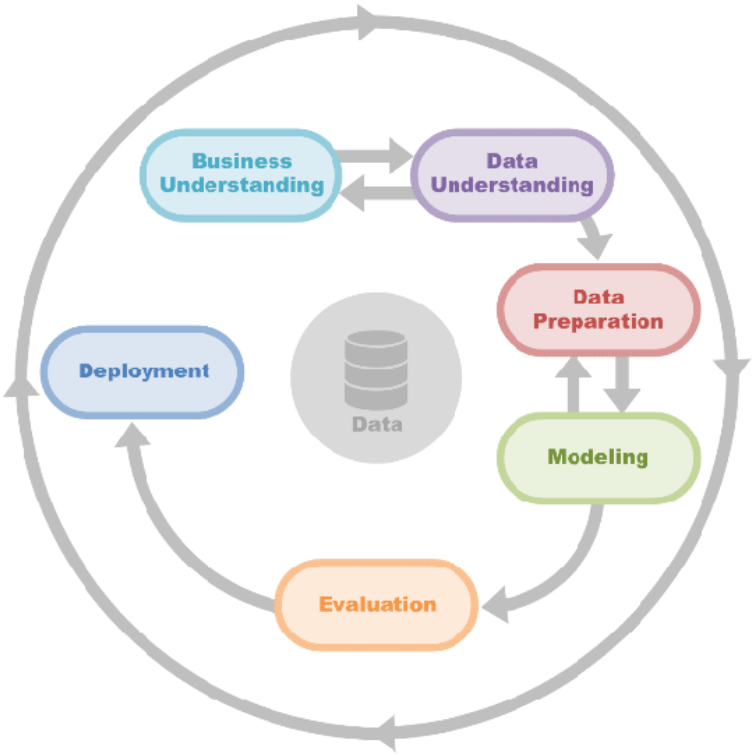


Figure 2.2: *« CRISP-DM framework » Lifecycle phases [2]*

The life cycle model consists of six phases with arrows indicating the most important and frequent dependencies between phases. The sequence of the phases is not strict. In fact, most projects move back and forth between phases as necessary.

Advantages:

- CRISP-DM Provides a systematic approach to data analysis, ensuring that data is cleansed and transformed effectively.

Disadvantages:

- CRISP-DM requires a significant amount of time and resources to implement, as it involves multiple stages and iterations.
- CRISP-DM may not be suitable for all types of data analysis projects, as it is primarily designed for structured data and may not be as effective for unstructured or complex data.

2.3.2 IBM Master Plan

The IBM Master Plan was a methodology developed by IBM in the 1950s and 1960s to systematically plan and manage large-scale business and technology projects.

Some key aspects of the IBM Master Plan methodology include:

- **Top-down, comprehensive planning:** The plan encompassed all aspects of the project from goals to tasks to ensure everything was considered and aligned.
- **Iterative approach:** The plan was not static but revisited regularly as the project progressed to allow for adjustments as needed.
- **Milestones and checkpoints:** Interim deadlines (milestones) were set to track progress and allow for course corrections if delays occurred.
- Breaking work into packages: The overall project was broken down into discrete work packages that could be estimated, scheduled and tracked independently.
- **Scheduling and resources:** Work packages were scheduled with estimated start/end dates and resource requirements to plan allocation of people, equipment etc.
- **Risk management:** Potential risks and issues were identified upfront along with contingency plans to mitigate risks to schedule and budget.
- **Documentation:** Planning documents including the master schedule, budgets and status reports were documented and updated regularly.

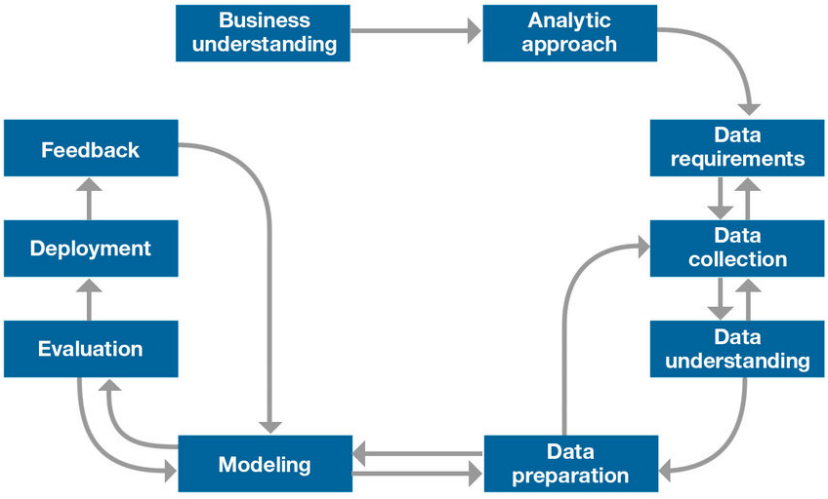


Figure 2.3: *« IBM Data Science Master Plan » Methodology Lifecycle [3]*

The IBM Master Plan Methodology contains 10 crucial steps:

business understanding, analytic approach, data requirements, data collection, data understanding, modeling, data preparation, deployment, evaluation, and monitoring.

Advantages

IBM Master Plan provides new practices extending the CRISP-DM framework.

Disadvantages

IBM Data Science Master has inherited certain drawbacks from the CRISP-DM methodology, especially the negligence of project management activities.

2.3.3 TDSP (Team Data Science Process)

The Team Data Science Process is a framework developed by Microsoft to guide data science projects. It consists of several stages, each addressing different aspects of the data science lifecycle.

TDSP's project lifecycle is like CRISP-DM and includes five iterative stages:

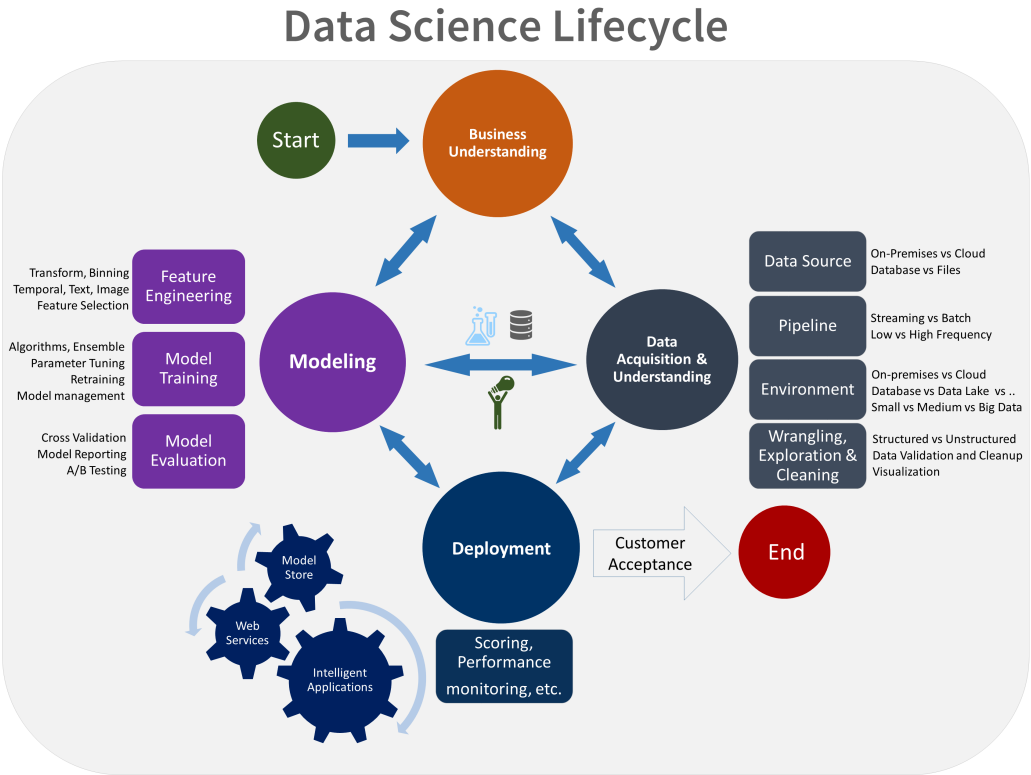


Figure 2.4: « *Microsoft TDSP» Methodology lifecycle [3]*

1. Business Understanding: define objectives and identify data sources
2. Data Acquisition and Understanding: ingest data and determine if it can answer the presenting question (effectively combines Data Understanding and Data Cleaning from CRISP-DM)

3. Modeling: feature engineering and model training (combines Modeling and Evaluation)
4. Deployment: deploy into a production environment
5. Customer Acceptance: customer validation if the system meets business needs (a phase not explicitly covered by CRISP-DM)

Advantages

- Agile: Emphasizes the need for incremental deliverables.
- Familiar: The product backlog, features, user stories, bugs, Git versioning, and sprint planning are familiar to those used to common software practices.
- Data Science Native: TDSP acknowledges that data science and software engineering are different, and is built for data science teams working on production-bound projects.
- Flexible: TDSP can be implemented as it is defined or in conjunction with other approaches such as CRISP-DM.

Disadvantages

- Fixed Sprints: TDSP leverages fixed-length planning sprints which many data scientists struggle with.
- Some Inconsistencies: Not all of Microsoft's documentation is consistent.

2.3.4 Comparison

The IBM Master Plan and CRISP-DM served as inspiration for the TDSP methodology, which was modified especially for use in the context of algorithmic trading and quantitative finance.

In contrast to CRISP-DM:

The real-time trading time constraints are given more weight by TDSP.

It maximizes data gathering, processing, and analysis to satisfy market performance standards.

It encourages backtesting, ongoing assessment, and quick feedback all the way through the model's life cycle.

In contrast to the IBM Master Plan:

Unlike the IBM Master Plan, which emphasizes the overall framework, the TDSP focuses on the details of trading.

It organizes the procedure around ongoing requirements for trading strategy modeling, implementation, and upkeep.

The stages are modified to fit the specific life cycle of financial market portfolios.

Finally, it appears that TDPS is the best strategy for our project.

2.3.5 Focus on TDSP

TDSP is suitable in the framework of quantitative finance and algorithmic trading. Considering particular limitations related to financial information and processes. Performance optimization to satisfy the stock market in Tunisia's requirements. Scalable methodology that ensures the necessary flexibility and ongoing development.

In the following sections, we will be detailing each stage of the lifecycle in TDSP:

Business Understanding

This stage involves understanding the problem at hand and defining the objectives of the data science project. It's crucial to comprehend the business context, stakeholder goals, and constraints.

Key Activities

- Identify the business problem or opportunity.
- Define project objectives and success criteria.
- Understand the business context and the impact of the project on stakeholders.

Data Acquisition and Understanding

In this stage, the focus is on collecting relevant data and gaining a deep understanding of its characteristics. This includes exploring data quality, structure, and potential challenges.

Key Activities

- Identify and gather relevant data sources.
- Explore and clean the data.
- Assess data quality and address missing or inconsistent data.
- Perform exploratory data analysis to understand patterns and relationships.

Modeling

The modeling stage involves developing and testing different models based on the project's goals. It includes selecting the most appropriate model and fine-tuning it for optimal performance.

Key Activities

- Select modeling techniques suitable for the problem.
- Create and train models using the chosen techniques.
- Evaluate and compare model performance.
- Fine-tune the selected model for better results.

Deployment

Once a satisfactory model is identified, the deployment stage involves implementing the model into a production environment, ensuring that it can be utilized for making predictions or generating insights.

Key Activities

- Develop the necessary infrastructure for deploying the model.
- Integrate the model into existing business processes.
- Monitor the model's performance in a real-world setting.
- Provide documentation and training for end-users and stakeholders.

2.4 Business Objectives & Data Science Objectives

Business Objective (BO1): Improving decision-making accuracy and efficiency in financial trading by enhancing the quality and reliability of data

DSO1: Prepare and process raw data for analysis, ensuring accuracy and consistency.

Business Objective (BO2): Enabling traders and investors to make informed decisions by predicting price movements and assessing risk in financial instruments

DSO1: Developing predictive models for forecasting price movements and evaluating risk in financial instruments to optimize trading strategies and investment decisions.

Business Objective (BO3): Harness NLP and LM for Deeper Insight into Tunisian Financial Markets

DSO1: Employ advanced NLP techniques, including sentiment analysis, topic modeling, and entity recognition, alongside sophisticated LM-based approaches, to extract nuanced insights from scraped news articles and financial reports pertaining to the Tunisian financial markets.

Business Objective (BO4): Analyze Assets' Risks and Returns

DSO1: Using feature engineering techniques and dividend profiles to obtain key metrics such as returns, volatility, and other risk measures for individual financial assets, providing a comprehensive understanding of their performance characteristics.

Business Objective (BO5): Uncovering actionable insights for improved trading outcomes and market understanding

DSO1: Apply clustering and segmentation techniques to identify distinct clusters within the market based on their similarities and patterns.

Business Objective (BO6): Present results synthetically and intuitively

DSO1: Develop user interfaces for easy visualization and exploration of analyses.

2.5 Metrics

Relative Metrics

2.5.1 Volatility

Definition: Volatility refers to the degree of variation of a trading price series over time. It is a statistical measure of the dispersion of returns for a given security or market index.

Role in Risk Indication

High volatility is often associated with higher risk. Traders and investors use volatility metrics, such as standard deviation or the VIX (Volatility Index), to assess the level of uncertainty or potential market fluctuations.

2.5.2 Skewness

Skewness measures the asymmetry of the probability distribution of a real-valued random variable. In finance, it often refers to the distribution of returns.

Role in Risk Indication

Positive skewness indicates a distribution with a longer right tail, suggesting potential for extreme positive returns. Negative skewness indicates a longer left tail, suggesting potential for extreme negative returns. Skewness can provide insights into the likelihood of extreme events.

2.5.3 Kurtosis

Definition: Kurtosis measures the tailedness or sharpness of the probability distribution of a real-valued random variable. In finance, it is used to understand the shape and thickness of the tails of a return distribution.

Role in Risk Indication

High kurtosis suggests fatter tails and a higher likelihood of extreme events, whether positive or negative. A leptokurtic distribution (high kurtosis) implies a higher probability of extreme outcomes, which can be relevant for risk assessment.

2.5.4 Absolute Metrics

Return on Investment (ROI)

The Return on Investment (ROI) is a crucial metric for assessing the profitability of the project. It measures the gain or loss generated relative to the amount invested. In the context of this project, the ROI can be calculated by comparing the benefits derived from the intelligent trading system implementation with the costs incurred during the development and deployment phases.

ROI Calculation Formula:

$$ROI = \frac{NetProfit}{TotalInvestment} \times 100$$

- **Net Profit:** The gains achieved from improved trading strategies, optimized portfolios, and enhanced decision-making tools.
- **Total Investment:** The overall cost involved in the development, implementation, and maintenance of the intelligent trading system.

SMART Metrics

To ensure the success and effectiveness of the project, it is essential to establish SMART metrics. SMART is an acronym for Specific, Measurable, Achievable, Relevant, and Time-Bound. These metrics provide clear criteria for evaluating project progress and success.

SMART Metrics for Project Objectives:

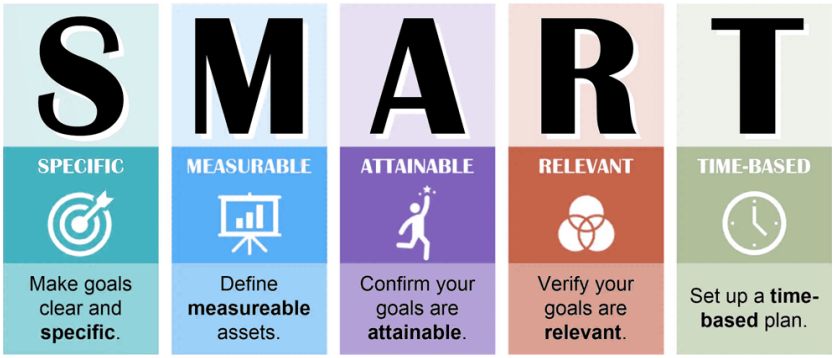


Figure 2.5: *SMART Metrics[4]*

1. **Specific:**
   - Develop and implement a user-friendly web-based platform for intelligent trading.
   - Enhance decision-making tools for investors in the Tunisian financial market.
2. **Measurable:**

- Achieve a minimum of 15% improvement in portfolio returns.
- Reduce the time required for financial data analysis by at least 30%.

3. **Achievable:**
    - Ensure the system is capable of handling a diverse range of financial instruments.
    - Implement trading strategies that outperform a predefined benchmark consistently.
4. **Relevant:**
    - Align project outcomes with the specific needs and challenges of the Tunisian financial market.
    - Address the lack of sophisticated data-driven tools and inefficiencies in identifying trading opportunities.
5. **Time-Bound:**
    - Complete the development and testing phases within the specified timeline.
    - Regularly assess and report progress against predefined milestones.

Monitoring and Reporting:

Regular monitoring and reporting on these metrics will be essential to track progress, make informed adjustments, and ensure that the project aligns with the defined objectives and expectations.

2.6 State-of-the-Art

In the rapidly evolving landscape of quantitative finance and trading, several key domains shape the current state-of-the-art. This section explores the prominent areas that influence the methodology and strategies in our project.

2.6.1 Quantitative Trading

According to Investopedia, Quantitative trading consists of trading strategies based on quantitative analysis, which rely on mathematical computations and number crunching to identify trading opportunities. Price and volume are two of the more common data inputs used in quantitative analysis as the main inputs to mathematical models.

Quantitative trading has been a lasting research area at the intersection of finance and computer science for many decades. In general, QT research can be divided into two directions. In the finance community, designing theories and models to understand and explain the financial market is the main focus. The famous capital asset pricing model (CAPM), Markowitz portfolio theory and Fama & French factor model are a few representative examples. On the other hand, computer scientists apply data-driven ML techniques to analyze financial data.

2.6.2 Mapping of Financial Players

Financial institutions play a crucial role in the landscape of the financial markets. Understanding the diverse players is essential for tailoring intelligent trading solutions. Here's a mapping of key financial players:

Banking-related:

Wholesale Banking: Involves financial services provided to large and mid-sized companies, including loans, credit, and other financial products.

Investment Banking: Specializes in facilitating capital raising, mergers and acquisitions (MA), equity and debt issuance, and securities trading.

Corporate Banking: Addresses the financial needs of corporations, offering services such as corporate lending, foreign exchange, and treasury operations.

Retail Banking: Caters to individual customers, providing services like payments, savings, mortgages, and loans.

Insurance-related:

Wholesale Insurance: Deals with large-scale insurance coverage for corporations and other entities.

Retail Insurance: Offers insurance products directly to individual customers, covering life, non-life, and commercial needs.

Asset Management:

Private Equity: Involves investments in private companies, often with the goal of acquiring or financing these companies.

Institutional Asset Management (AM): Manages investment portfolios on behalf of institutions such as pension funds and endowments.

Mutual Funds: Pools funds from many investors to invest in a diversified portfolio of stocks, bonds, or other securities.

Private Banking: Provides personalized financial and banking services to high-net-worth individuals.

Life and Pension Plans: Includes individual and group life insurance, unit-linked life products, and various pension plans.

Understanding the roles and functions of these financial players is vital for tailoring intelligent trading solutions that cater to the diverse needs of the market.

2.6.3 Machine Learning and Data Science in Finance

Machine learning and data science have revolutionized decision-making in finance, introducing cutting-edge techniques for analysis and prediction:

Supervised Learning

In financial markets, supervised learning is instrumental for predictive modeling. Models trained on historical data facilitate forecasting future trends, including movements in stock prices.

Unsupervised Learning

Through unsupervised learning, particularly clustering algorithms, finance professionals gain insights into market segmentation and hidden structures within datasets, enabling more informed decision-making.

Deep Learning

deep learning becomes an appealing approach owing to not only its stellar performance but also to the attractive property of learning meaningful representations from scratch

Natural Language Processing (NLP)

NLP techniques contribute significantly to market intelligence. Parsing through financial articles and news, NLP aids in sentiment analysis and the extraction of valuable information

Applications of machine learning in Trading and Finance

Applications in Trading: Machine learning finds diverse applications in finance, from algorithmic trading strategies to risk assessment models and the development of predictive analytics for financial markets.

2.6.4 Risk Management in Quantitative Trading

Effective risk management is paramount in quantitative trading to safeguard against potential losses. Key considerations in this domain encompass:

- Statistical Risk Models: Employing statistical models to evaluate and quantify risks associated with financial instruments and prevailing market conditions.

- Diversification Strategies: Prudent portfolio diversification serves as a risk mitigation strategy. Allocating assets across different classes reduces vulnerability to adverse market movements.
- Volatility Assessment: In dynamic markets like Tunisia, assessing and managing volatility is critical. Robust models are implemented to navigate and capitalize on changing market conditions.

2.7 Key Performance Indicators (KPIs)

- Return on investment (ROI) - The net gain or loss generated by the trading strategies, expressed as a percentage of the initial investment. This measures how profitable the strategies are.
- Annualized return - The average annual return generated by the trading strategies over a period of years. This gives a sense of the long-term growth potential.
- Sharpe ratio - A risk-adjusted return metric that measures the excess return per unit of volatility. A higher Sharpe ratio indicates better risk-adjusted returns.
- Sortino ratio - Similar to the Sharpe ratio but factoring only downside deviation below a target return. This gives a clearer picture of downside risk.
- Maximum drawdown - The peak-to-trough decline during a specific period. Too large a drawdown can indicate high risk.
- Hit rate - The percentage of trades that are profitable. Too low a hit rate may mean the strategies aren't robust enough.
- Exposure - How much of the portfolio is allocated on average. Higher exposure means higher risk but also potential upside.
- Diversification - How dispersed the holdings are across different assets. Higher diversification reduces risk.
- Outperformance vs benchmark - By how much the strategies outperform the target benchmark index on average. This is a key measure of success.
- Volatility - How much the returns fluctuate over time. The strategies should aim to generate returns with lower volatility than the benchmark.
- Processing time - How long it takes to execute the strategies after new data comes in. Faster processing is important for timely trades.
- Portfolio return: Measures the financial performance of the portfolio in terms of return relative to the initial target set.
- Portfolio beta: Measures the volatility of the portfolio relative to that of the market. It should be as close as possible to 1.
- Correlation with the benchmark: It should be as low as possible to ensure diversification.
- Diversification level: Measured by the number of securities in the portfolio. The level should not fall below the threshold set.
- Forecast accuracy: Evaluate the accuracy of prediction algorithms on new data.

2.8 Existing Solutions & Competitors

Understanding the competitive landscape is crucial for any company's success in the market. In this analysis, we delve into three key players in the field of financial technology and trading solutions: Talys, Amef Consulting, and Linedata. By comparing their offerings, target audience, technology, and market presence, we aim to gain valuable insights into their strengths, weaknesses, and their potential impact on Value's positioning and future strategies. This analysis will inform decision-making related to both internal resource allocation and external marketing efforts, ensuring Value remains competitive and adaptable in the evolving financial technology landscape

Table 2.1: *Relevant Competitors comparison*

Company	**Value**	**Talys**	**Amef Consulting**	**Linedata**
Target Audience	SMEs, Financial Institutions	Investment Banks, Asset Managers	Asset Managers, Hedge Funds	Large Financial Institutions
Market Share	N/A (Private Company)	Regional (France, Benelux)	Regional (France, North Africa)	Global
Pricing	Project-based, subscription options	Custom quotes	Custom quotes	Tiered pricing based on features and assets
Key Features	Algorithmic Trading, Quantitative Finance, AI-powered Strategies	Trading Platform, Portfolio Management, Risk Management	Consulting Services, Algorithmic Trading	Custody, Investment Management, Data Management
Technology	Proprietary AI Platform, Cloud-based	Proprietary Platform, Cloud-based	Open-source tools, Proprietary models	Proprietary Platform, Cloud-based
Differentiators	Focus on Tunisian market, AI expertise	Long industry experience, Global reach	Consulting-driven approach, Open-source integration	Global reach, Comprehensive product suite
Strengths	Agile team, Deep AI expertise	Established reputation, Strong client relationships	Expertise across asset classes, Customizable solutions	Scalability, Wide range of offerings
Weaknesses	Limited track record, Smaller company	Regional focus, Less emphasis on AI	Consulting-heavy model, Less focus on technology	Complex pricing structure, Can be expensive
Opportunities	Growing demand for AI-powered trading, Expansion into new markets	Increasing competition in algorithmic trading, Expansion into new asset classes	Growing demand for consulting services, Integration with new technologies	Continued expansion, Acquisition of new clients

2.9 SDGs

The development and implementation of intelligent trading agents in the Tunisian stock market present a unique opportunity to contribute to positive change and align with

several Sustainable Development Goals (SDGs) outlined by the United Nations. These goals represent a global call to action for tackling critical challenges related to poverty, inequality, climate change, and other essential aspects of sustainable development. Our project aims to harness the power of artificial intelligence (AI) to enhance efficiency, liquidity, and accessibility within the Tunisian market. As a result, it has the potential to contribute to the following SDGs:

2.9.1 SDG 8 - Decent Work and Economic Growth

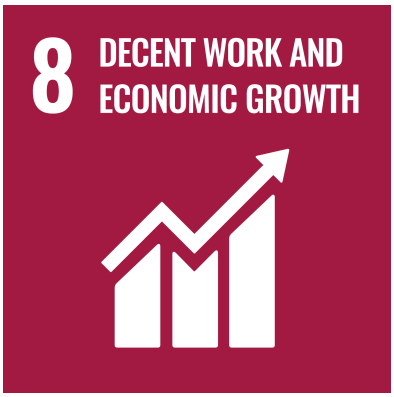


Figure 2.6: *Sustainable Development Goal 8 [5]*

The project aims to develop intelligent trading strategies and platforms which can help boost economic growth by facilitating investment and capital markets activity. More efficient trading can improve market liquidity.

2.9.2 SDG 9 - Industry, Innovation and Infrastructure

Figure 2.7: *Sustainable Development Goal 9 [5]*

Building advanced quantitative trading systems leveraging machine learning/AI techniques demonstrates progress towards developing sustainable infrastructure to support economic activity.

2.9.3 SDG 17 - Partnerships for the Goals

The project involves collaborating with financial institutions, data vendors, consultants etc. to build the necessary technical and data partnerships for success.

Figure 2.8: *Sustainable Development Goal 17 [5]*

2.10 Conclusion

In this chapter, we thoroughly presented the domain and the different components of the trading business with an emphasis on the Tunisian markets. We presented the company "Value" and illustrated its place in the market. In addition, we announced the problem statement and the idea of the project presented as a solution. We then explained our choice to work with the TDSP methodology through a comparative exploration of other Data Science project methodologies. We ended the chapter with a description of the metrics we used and a study of the KPI of this enterprise. This chapter represents the first step of the TDSP methodology which is business understanding. In the next chapter, we will start with the second phase which is Data Acquisition and Understanding.

3

Data Understanding & acquisition

This chapter provides an in-depth exploration of the data acquisition process, covering the diverse sources of data, methods of acquisition, and preparatory steps undertaken to ensure data quality and relevance for the future steps especially forecasting stocks and portfolio diversification.

3.1 Time Window Selection for Analysis

The time window selected for our analysis spans from 2019 to 2024. This period was chosen for several reasons:

1. **Significance of 2019:** The year 2019 marked the onset of the COVID-19 pandemic, which brought about profound changes in the stock market landscape globally. The pandemic disrupted economies, industries, and financial markets, making it a pivotal year for analysis.
2. **Capturing Dynamic Shifts:** By focusing on the years following 2019, we aim to capture the dynamic shifts and emerging trends that unfolded in the aftermath of the pandemic. This critical period provides insights into how markets responded to unprecedented challenges and adapted to new realities.
3. **Informed Decision-Making:** Analyzing data within this time window ensures that our analysis reflects relevant events and developments, enabling informed decision-making and strategic planning. By considering the post-2019 period, we can better understand the impact of recent events on market dynamics and investor behavior.

However, it's important to note that our analysis extends beyond the selected time window. We have also incorporated data from years prior to 2019, as provided to us, to gain a comprehensive understanding of historical trends and patterns in the market.

3.2 Data Sources and Acquisition

This section delves into the various sources from which data was acquired and the methodologies employed for data retrieval.

3.2.1 Market Data Acquisition

Market Data Provider Collaboration

VALUE digital services, a reputable data services provider, has granted us access to comprehensive datasets relevant to the Tunisian stock market. This collaboration ensured access to high-quality, curated data essential for conducting meaningful analysis and deriving actionable insights.

Batch-wise Data Acquisition

We Received market data in multiple batches, each containing distinct datasets capturing different aspects of the stock market. This structured approach to data acquisition facilitated systematic analysis and enabled targeted exploration of specific market dynamics and trends.

Historical Stock Prices Dataset

The historical stock prices dataset obtained from Value encompasses a wealth of information crucial for analyzing stock market trends and performance over time. This dataset is organized into multiple batches, each containing comprehensive data on various aspects of stock trading. Here, we outline the key columns present in these datasets:

Batch 1: Companies in the Tunisian Stock Trading Market (2008-2022)

- **Company Name (Valeurs):** Represents the official name of the company listed in the Tunisian stock market.
- **Nominal:** Denotes the nominal value of the stock.
- **Montant (Year):** Indicates the trading volume or amount for a specific year.
- **Date (Year):** Specifies the corresponding date for the trading volume.
- **Variation (Year):** Represents the variation in the stock's performance over the specified year.

This batch provides insights into the trading activity and performance of companies listed in the Tunisian stock market over the years, allowing for detailed trend analysis and comparative studies.

Historical Cotation Datasets (2008-2022)

- **Date:** Represents the date of the trading session.
- **Seance:** Indicates the session during which the trades occurred.
- **Groupe:** Categorizes the stock into relevant groups based on characteristics like industry.
- **Code:** Denotes the unique ticker symbol assigned to the stock.
- **Valeur:** Specifies the official name of the stock or company.
- **Open:** Represents the opening price of the stock.
- **Close:** Indicates the closing price of the stock.

- **Plus_Bas:** Denotes the lowest price of the stock during the trading session.
- **Plus_Haut:** Represents the highest price of the stock during the trading session.
- **Quantite_Negociee:** Specifies the total number of shares traded.
- **Nb_Transaction:** Indicates the count of individual trades.
- **Capitaux:** Denotes the total value of shares traded.
- **Ind_Res:** Represents a specific market indicator or result calculated from the trading data.

This dataset provides a granular view of the daily trading activity of stocks listed in the Tunisian stock market, facilitating detailed analysis of price movements, trading volumes, and market trends.

Historical Index Datasets (2008-2022)

- **SEANCE**: Represents the trading session or date.
- **CODE_INDICE**: Identifier for the specific index being tracked.
- **LIB_INDICE**: Name or label of the index.
- **INDICE_JOUR**: Index value at the end of the trading day.
- **INDICE_VEILLE**: Index value from the previous trading day.
- **VARIATION_VEILLE**: Change or variation in the index value from the previous trading day.
- **INDICE_PLUS_HAUT**: Highest value the index reached during the trading session.
- **INDICE_PLUS_BAS**: Lowest value the index reached during the trading session.
- **INDICE_OUV**: Opening value of the index at the beginning of the trading session.

Batch 2: Market Companies Datasets

In the context of financial markets, "security" refers to a financial instrument that can be bought or sold in the market, and the dataset captures data related to the pricing and trading activity of these securities.

Description: The dataset contains information regarding various market securities, including banks, assurances, leasing companies, and others. Here are the key columns present in this dataset:

- **Date:** Represents the date of the market data.
- **Price:** Indicates the closing price of the security on a given date.
- **Open:** Represents the opening price of the security on a given date.
- **High:** Denotes the highest traded price of the security during the trading day.
- **Low:** Signifies the lowest traded price of the security during the trading day.
- **Vol:** Represents the trading volume, indicating the total number of shares traded on a given date.
- **Change:** Indicates the percentage change in the closing price compared to the previous trading day.

- **Name:** Specifies the name of the company or security associated with the market data.

This data provides valuable insights into the pricing and trading activity of various securities in the financial market, facilitating analysis and decision-making processes for investors and analysts.

Batch 3: Macroeconomic Data

In addition to the market-specific data, our analysis incorporates macroeconomic indicators from key trading partners, including the United States, Europe, and Japan. These indicators provide valuable insights into global economic trends and their potential impact on the local market.

Data Sources

The analysis of macroeconomic data from key trading partners, including the United States, Europe, and Japan, provides valuable insights into the economic dynamics influencing Tunisia's trade and financial landscape.

The macroeconomic data is sourced primarily from the World Data Bank, a comprehensive repository of global economic and social development data. The datasets cover a wide range of indicators, including but not limited to GDP growth, inflation rates, unemployment rates, trade balances, and industrial production indices.

Metadata Tables

To ensure proper understanding and interpretation of the macroeconomic datasets, metadata tables containing the data dictionary are provided. These tables offer explanations of the column aliases used in the actual datasets, elucidating the meaning and significance of each indicator. Key components of the metadata tables include:

- **Indicator Name:** Provides a descriptive title for the economic indicator.
- **Alias:** Represents the abbreviated name or code used to reference the indicator within the dataset.
- **Description:** Offers a detailed explanation of the indicator, including its calculation methodology, units of measurement, and relevance to economic analysis.

By consulting the metadata tables, analysts can accurately interpret the macroeconomic data and incorporate it effectively into their market analysis and forecasting models.

Market Indexes Dataset

The market indexes dataset contains essential information regarding key market indicators and overall market performance. Here are the columns present in this dataset:

- **Date:** Represents the date of the market data.
- **High:** Denotes the highest value reached by the market index.
- **Low:** Signifies the lowest value reached by the market index.
- **Open:** Represents the opening value of the market index.
- **Close:** Indicates the closing value of the market index.
- **Name:** Specifies the name of the market index or associated security.

This dataset offers insights into broader market trends and dynamics, allowing for the analysis of overall market performance and benchmarking against specific market indices.

By leveraging these diverse datasets, we gain comprehensive insights into various aspects of the Tunisian stock market, enabling robust analysis and informed decision-making in our data science endeavors.

3.3 Scraping and Database Integration

3.3.1 News articles Scraping

For Economics news, we used Selenium, a powerful web scraping tool, to gather data from various financial market sources, specifically targeting news articles and company information from the Tunisian stock exchange (BVMT). Selenium allowed us to dynamically navigate through web pages, extract relevant information, and store it for further analysis.

Figure 3.1: *Selenium Logo [6]*

News and articles sources

Tunisian Stock Market (BVMT) Website

Extracting articles and news pertinent to stock trading activities and economic developments.

Agence Tunis Afrique News

Capturing comprehensive coverage of regional economic news and market analyses.

Tunisienumerique Economy Section

Scraping articles focusing on economic trends and business developments.

With Selenium, we automated the process of collecting news articles from the BVMT website, extracting data such as publication date, title, and content. This facilitated the

creation of a comprehensive dataset containing valuable insights into market trends and events.

Using SpaCy's pre-trained models, we successfully extracted company names mentioned in the scraped data. Additionally, we implemented a rule-based approach to capture additional company names that might have been missed by the NER model. This rule-based extraction method involved defining patterns and rules to identify company names based on specific criteria, such as capitalization patterns, presence of certain keywords, or syntactic structures.

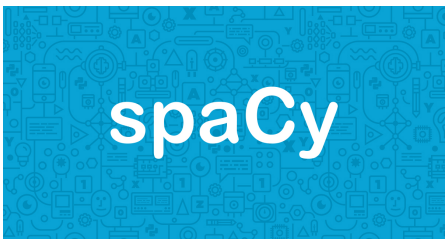


Figure 3.2: *SpaCy Logo [7]*

The combination of web scraping and NER provided us with a comprehensive dataset of company names, allowing for further analysis and insights into the financial market landscape. By leveraging these techniques, we augmented our dataset with valuable information, enabling more robust analysis and informed decision-making in our data science endeavors.

3.3.2 Web Scraping of Financial Data

We scraped financial data from the official *Financial Market Council Website* cmf.tn to obtain the financial statements of companies listed in the market watch. Our objective was to extract Income Statement tables, Cash Flow statements, and Balance Sheets for the years 2019 to present.

Figure 3.3: *Financial Market Coucil CMF Logo [8]*

We employed multiple techniques for this task:

- **OCR Techniques:** We utilized Optical Character Recognition (OCR) techniques, including pretrained models like PubLayNet and Camelot, along with tools like Pytesseract, to extract tabular data from scanned documents and images.
- **Parsing PDFs:** We parsed PDF documents using tools such as Tabula and PDFPlumber to extract structured data from financial reports available on cmf.tn.

- **Table Querying:** Additionally, we used Excel and querying techniques to extract specific tables and data points from the parsed financial statements, ensuring accuracy and completeness in our dataset.

By leveraging these techniques, we obtained comprehensive financial data for analysis, enabling us to gain insights into the financial performance and health of companies listed in the Tunisian stock market.

Exploratory Data Analysis (EDA)

In this section, we conduct exploratory data analysis (EDA) to gain insights into the time series nature of the acquired datasets. We analyze the temporal patterns, trends, and seasonality present in the data to better understand the dynamics of the market.

3.4 Selection of Temporal Data

We have selected a diverse range of years to analyze, encompassing both typical market conditions and periods marked by significant events. The chosen years include:

2013: Representative of a standard trading year, unaffected by major economic crises or extraordinary events. 2019: Another "normal" trading year, serving as a baseline for comparison with more recent data. 2020: A year of exceptional significance due to the global COVID-19 pandemic, providing insights into the market's response to unprecedented challenges.

3.5 Market Values - Dividends

Temporal Distribution of Key Variables

We begin by examining the temporal distribution of key variables, focusing on the first batch of datasets containing companies, their respective nominal values, trading amounts, and variations over time.

Example: Stock amount and Variation in 2020

We visualize the temporal distribution of trading amounts (Montant) and variations for the year 2020. This analysis offered insights into how market activity and volatility evolved during the challenging circumstances posed by the COVID-19 pandemic.

Comparative Analysis: 2019 vs. 2013 Next, we compare the data from 2019 with that of 2013 to discern any notable differences or similarities in market behavior between these two "normal" trading years. This analysis aids in identifying trends over time and assessing the market's performance across different economic conditions.

the figures above have shown:

1. Peaks: There are notable peaks on the graph that could correspond to periods of high financial activity or specific transactions that resulted in increased monetary amounts.

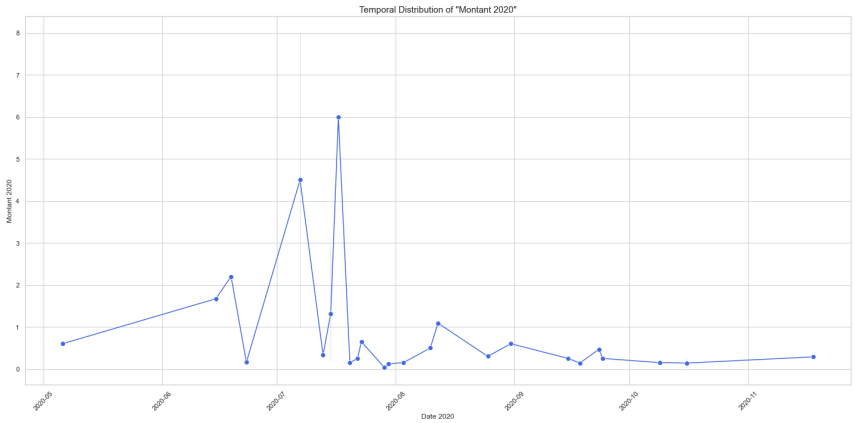
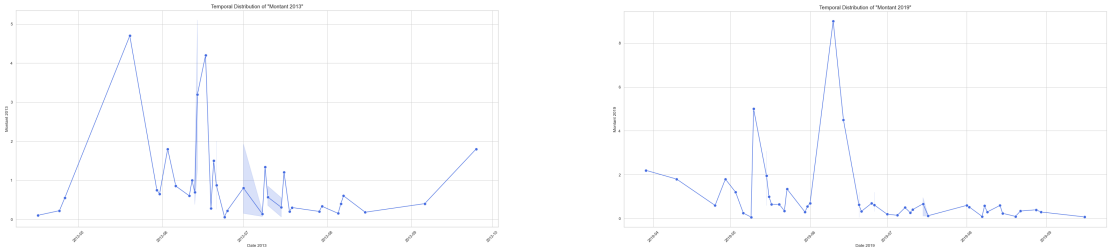


Figure 3.4: *temporal distribution of trading amounts and variations for the year 2020*



(a) *Temporal distribution for 2013* **(b)** *Temporal distribution for 2019*

Figure 3.5: *Temporal distribution of trading amounts and variations for the years 2013 and 2019*

2. Troughs: The graph also shows periods where the amount values drop, indicating lower financial activity or lesser monetary amounts being recorded for the companies.
3. End-of-Year Increase: There's an upward trend toward the end of the year, suggesting an increase in monetary values as the year progresses. This could be due to a variety of factors such as seasonal business patterns, end-of-year financial settlements, or other economic factors. However during the Covid-19 crisis we can see that it's not the case.

The figure depicting "Average Yearly Returns Over Time" illustrates the fluctuating nature of annual returns from 2009 to 2022. the trend generally shows many fluctuations over the years, a notable deviation is observed in 2020, marked by a significant fluctuation. This deviation coincides with the onset of the COVID-19 pandemic and the resulting market volatility and economic uncertainty. Such fluctuations underscore the dynamic nature of financial markets and the impact of external events on investment performance. Understanding these trends can guide investors in adapting their strategies to mitigate risks and capitalize on opportunities, particularly during periods of heightened market volatility.

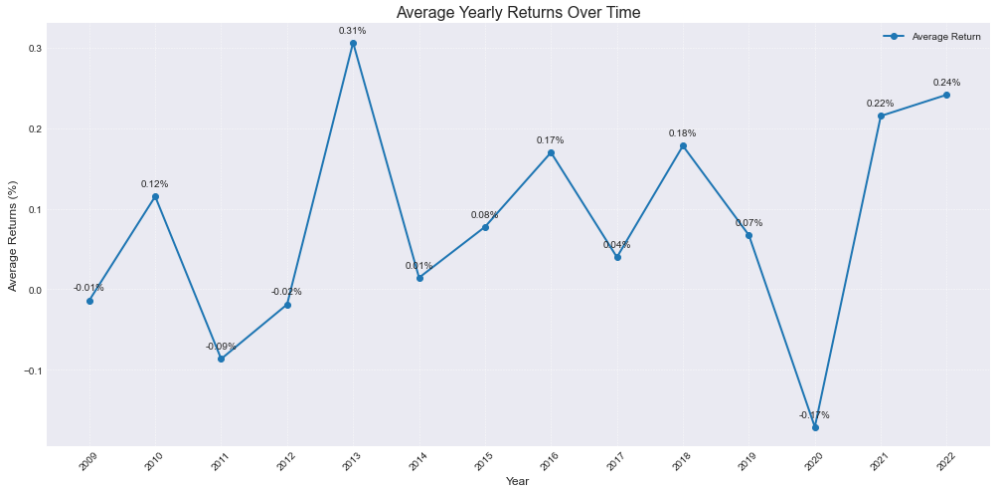
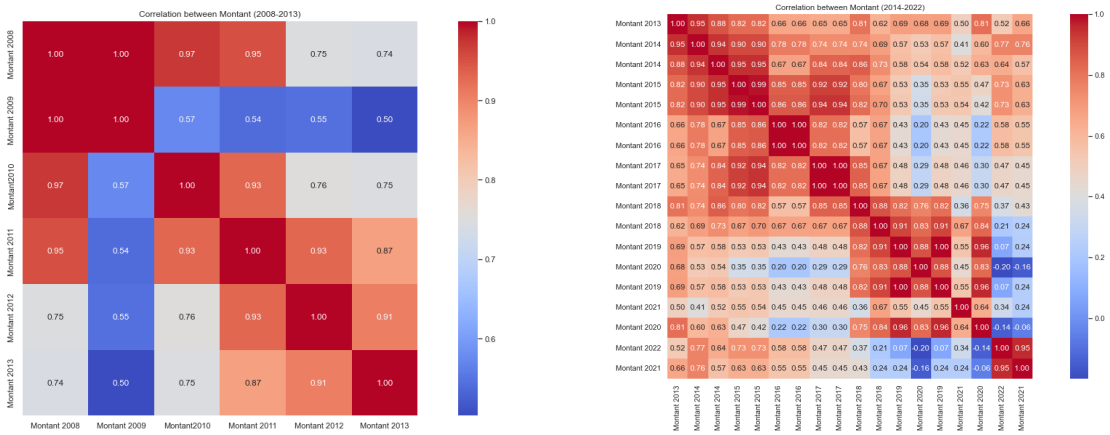


Figure 3.6: *Average Yearly Returns Over Time (2009-2022)*



(a) *Correlation Matrix of Trading Amounts (2008-2013)*

(b) *Correlation Matrix of Trading Amounts (2014-2022)*

Figure 3.7: *Correlation matrix illustrating the relationship between trading amounts from 2008 to 2013 and 2014 to 2022, providing insights into more recent market dynamics*

Correlation Analysis

The correlation analysis was conducted over two distinct time intervals: from 2008 to 2013 and from 2014 to 2022. The reason for this division is rooted in the composition of the companies within the dataset.

Strong Positive Correlations: Several pairs, like Montant 2014 with Montant 2015 (1.00), show close alignment, indicating shared performance factors.

Strong Negative Correlations: Others, such as Montant 2015 with Montant 2020 (-0.80), move inversely, suggesting divergent performances.

3.5.1 Feature Engineering

Dividend_Consistency:

The number of years a company has paid out dividends. This histogram shows how

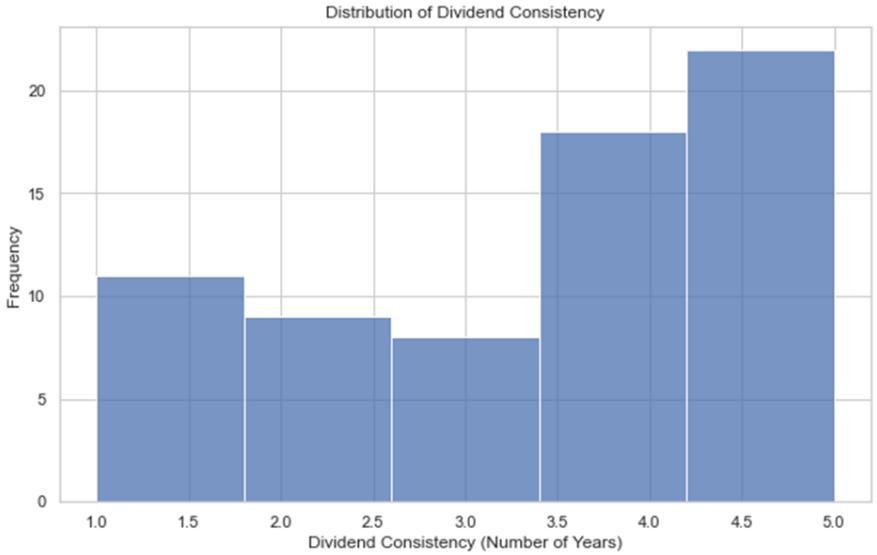


Figure 3.8: *Distribution of Dividend Consistency*

many companies have paid dividends consistently over the years. A higher number indicates a company has been consistent in paying dividends across the observed period, which could be seen as a positive indicator of financial stability.

Dividend_Variability: The standard deviation of the dividend amounts over the available years, measuring the variability in dividend payments. The distribution

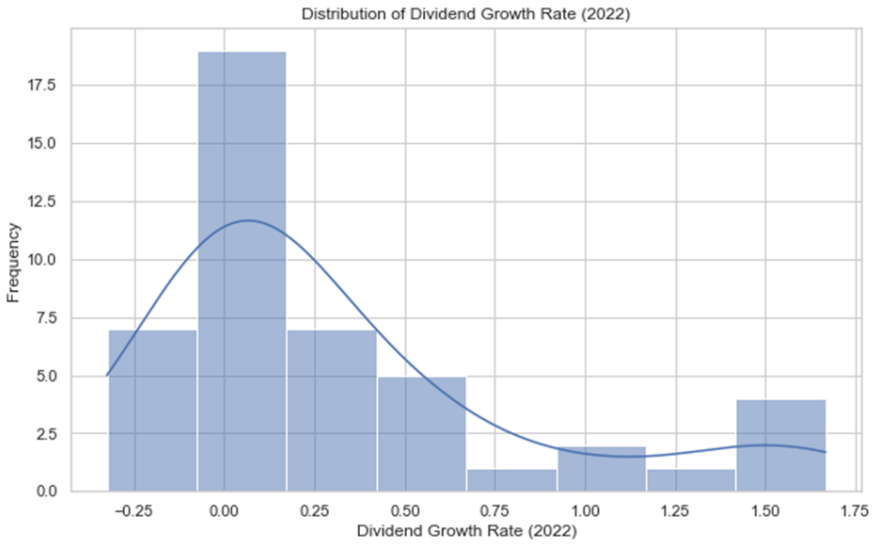


Figure 3.9: *Distribution of Dividend Variability*

of dividend variability, measured by the standard deviation of dividend amounts, illustrates how stable the dividend payments have been. Lower variability suggests more stable dividend payments, which might be preferable for risk-averse investors.

DGR_Year: The Dividend Growth Rate for each year, calculated as the year-over-year change in dividend amounts.

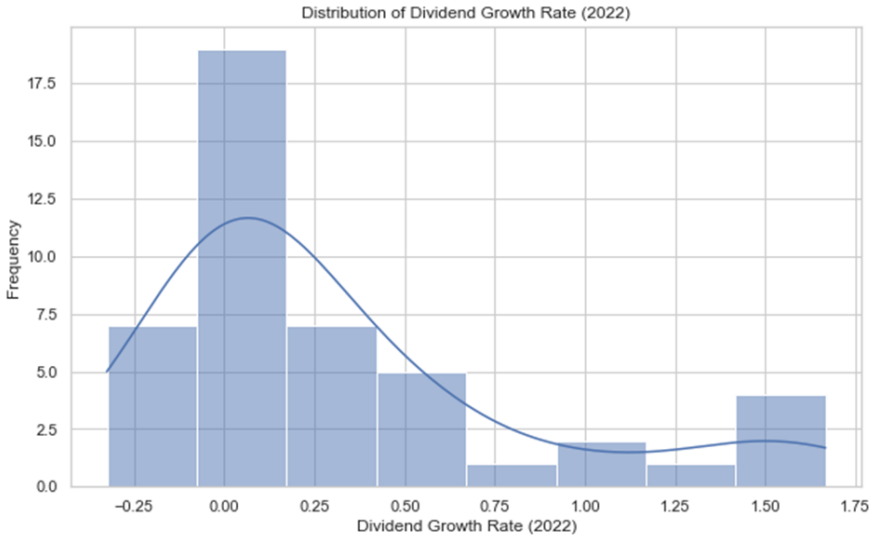


Figure 3.10: *Distribution of Dividend Growth Rate (DGR) for 2019 and 2020*

The histogram with the Kernel Density Estimate (KDE) overlay shows the distribution of Dividend Growth Rates (DGR) for 2019 and 2020

3.6 Historical Indexes Data

Our historical index data refers to records of past values of various stock market indexes over time. It provides a historical record of how market indexes have performed over different time periods.

3.6.1 Data Cleaning

1. **Header Removal:** The header row of the dataset is removed as it typically contains non-data information.
2. **Index Reset:** The index of the dataset is reset to ensure it starts from 0 and is sequentially ordered.
3. **Date Formatting:** The date column is converted to a standardized datetime format for uniformity and ease of analysis.
4. **Date Indexing:** The date column is set as the index of the dataset for chronological organization and efficient time-based querying.
5. **Invalid Date Removal:** Rows with invalid or unparseable date entries are eliminated to maintain data integrity.
6. **Numeric Conversion:** Numeric columns are converted to numerical data types to enable mathematical operations and analysis, ensuring consistency in data representation.

7. **Mean Imputation for Missing Values:** Missing values, if any, are addressed through mean imputation. This approach ensures data completeness while minimizing bias, thereby preserving the integrity of the dataset.

3.6.2 Feature Engineering

- **Percentage Change:**
  - Represents the percentage change in the index value from the previous day's close.
  - Indicates daily growth or decline in the index, useful for assessing short-term market trends and volatility.
- **Volatility:**
  - Measures the range of price fluctuations in the market.
  - Higher volatility suggests greater uncertainty and risk, while lower volatility implies stability.
  - Helps investors assess risk levels and adjust trading strategies accordingly.
- **Gap Up or Gap Down:**
  - Binary features indicating a gap up (positive change) or gap down (negative change) at the opening of the trading session compared to the previous day's close.
  - Gaps reflect significant market sentiment and can signal potential trading opportunities.
- **Moving Averages:**
  - Smooth out price data by calculating the average of past values over a specified time period.
  - Identify trends and potential reversal points in the market.
  - Shorter moving averages respond quickly to price changes, while longer moving averages provide a broader perspective on trends.
- **Relative Strength Index (RSI):**
  - Momentum oscillator measuring the speed and change of price movements.
  - Oscillates between 0 and 100, identifying overbought or oversold conditions.
  - RSI values above 70 indicate overbought conditions, while values below 30 suggest oversold conditions.
- **Bollinger Bands:**
  - Consist of a simple moving average (middle band) and upper and lower bands representing standard deviations from the moving average.
  - Identify overbought or oversold conditions and potential price reversal points.
  - Touching the upper band may indicate overbought conditions, while touching the lower band may suggest oversold conditions.

These features collectively provide insights into market behavior, trends, volatility, and trading opportunities. They are utilized by traders, analysts, and investors for decision-making, risk management, and optimizing trading strategies. Additionally, they serve as inputs to machine learning models for predicting future market movements.

Feature Insights

In this section, we conduct a comprehensive analysis of market trends and feature insights across different sectors and datasets, focusing specifically on the COVID-19 period of 2019. While exploring historical indexes data, we aim to showcase two key examples: the opening index and the occurrence of gaps over time. Through the examination of these specific metrics, we endeavor to uncover underlying patterns and trends that may inform trading strategies and investment decisions. By carefully analyzing these examples, we seek to provide actionable insights into the market dynamics during this critical period.

Opening Index

This observation underscores the widespread impact of the pandemic on global financial markets.

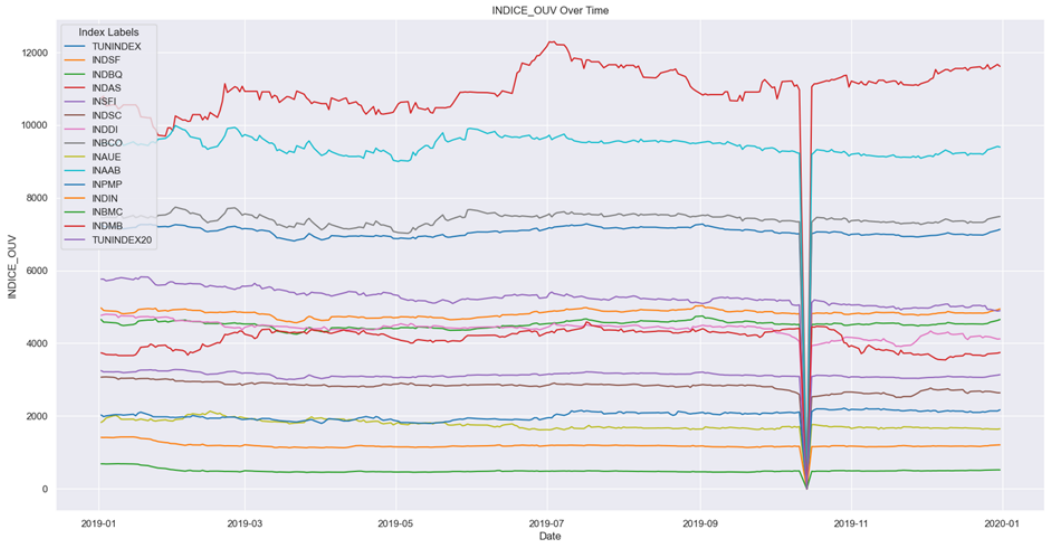


Figure 3.11: *Opening Index for 2019*

- **Magnitude of Decrease:**
  This plot exhibits a sharp downward trend in opening index values across different indexes during the COVID-19 period. This substantial decrease indicates a broad-based decline in market sentiment and investor confidence
- **Market Uncertainty:**
  The pronounced decline in opening index values suggests heightened uncertainty and volatility in financial markets. Investors may have reacted to the uncertainty

surrounding the pandemic by selling off assets, leading to a widespread decline in market indices.

- **Investor Response**:
  The steep decline in opening index values during the COVID-19 period may have prompted investors to reassess their investment strategies and risk tolerance. It underscores the importance of risk management and the need for diversified portfolios to mitigate the impact of such unprecedented events.

Gap Analysis

An analysis of gap occurrences over time during the COVID-19 period reveals a stark change in market dynamics:

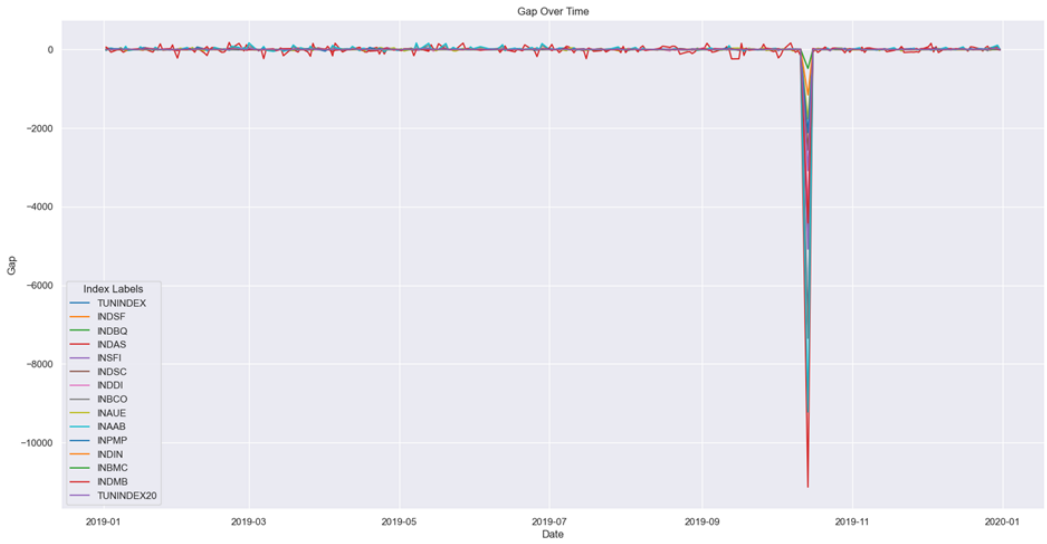
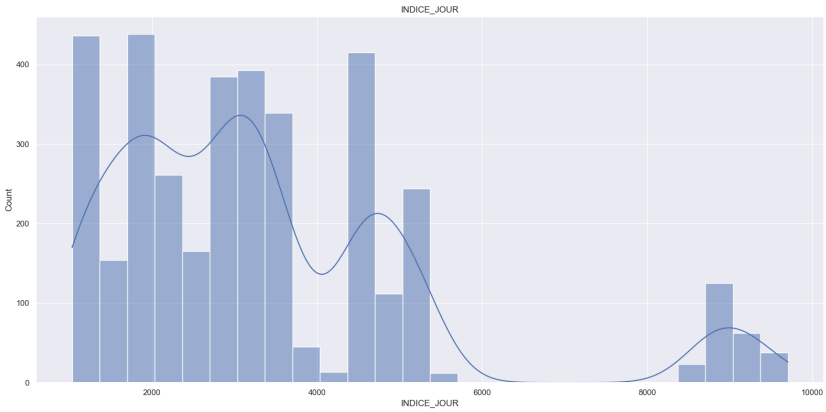


Figure 3.12: *Gap Analysis for the year 2019*

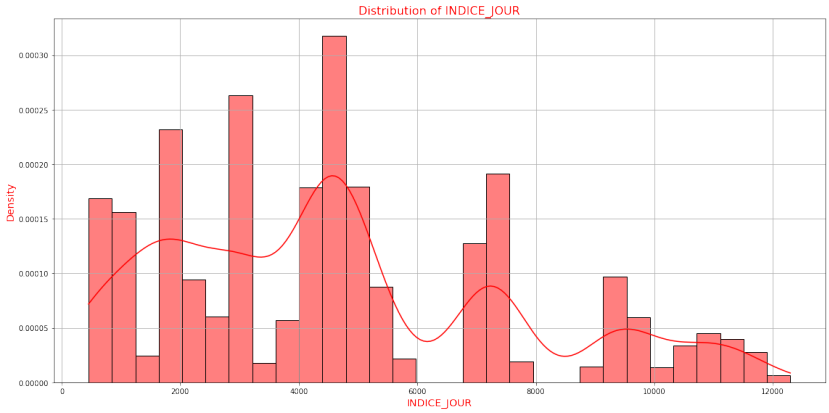
- **Stable Period Pre-COVID:**
  Before the onset of the pandemic, the plot shows relatively stable gap occurrences, with fluctuations around zero. This period likely reflects typical market behavior with occasional minor gaps due to routine market events or news releases.
- **Abrupt Change at COVID Onset:**
  As the COVID-19 pandemic emerged and spread globally, the plot demonstrates a sudden and significant increase in gap occurrences. This abrupt change indicates a rapid shift in market sentiment and heightened volatility.
- **Sharp Decline in Gap Values:**
  The gap plot shows a remarkable transition from predominantly small positive or negative gaps to a drastic decrease, with gap values plummeting to less than -10,000. This sharp decline suggests unprecedented market turmoil and panic selling as investors reacted to the uncertainties surrounding the pandemic.
- **Market Distress and Uncertainty:**
  The steep drop in gap values reflects the distress and uncertainty prevalent in financial markets during the COVID-19 crisis. Investors likely faced heightened fear and uncertainty, leading to extreme market reactions and large price gaps.
- **Impact on Investor Confidence:**
  The drastic change in gap occurrences and values serves as a barometer of investor confidence and risk appetite. The significant negative gaps indicate a widespread lack of confidence and a rush to sell assets, exacerbating market declines.
- **Long-Term Implications:**
  The sustained period of elevated gap occurrences and large negative gap values

during the COVID-19 period may have long-term implications for market participants. Investors may become more risk-averse, and policymakers may implement measures to restore market stability and investor confidence.

Market Indexes trend analysis



(a) *index values (day) for 2013*



(b) *index values (day) for 2019*

Figure 3.13: *comparison of histograms of index values for the years 2013 and 2019*

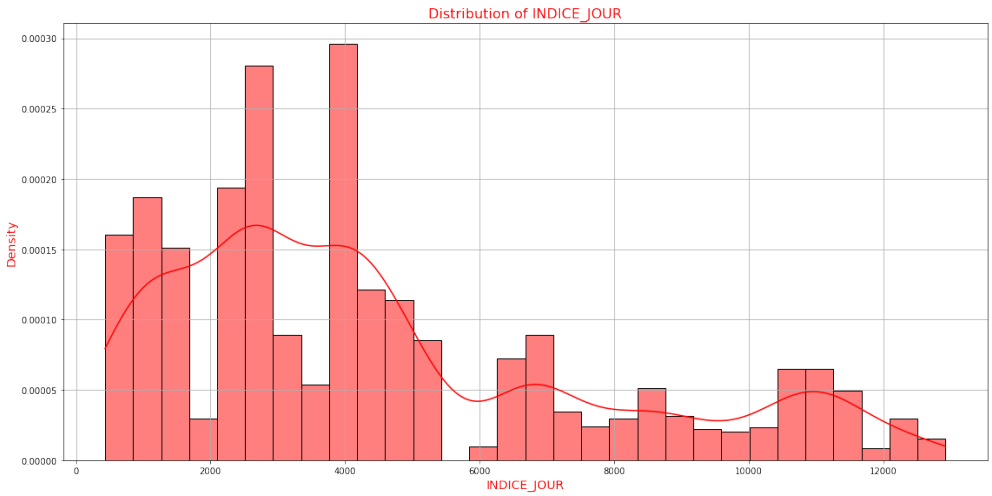


Figure 3.14: *Index Value (day) for 2020*

The comparison between the Indexes distribution in 2013 and 2019 shows a Positive Shift to the right, it suggests overall positive performance or growth in the market. This

could be due to factors such as economic expansion, increased investor confidence, or favorable policy changes that carried on to 2020 despite the Covid-19 crisis, this could be due to Sectoral Variations or resilience and confidence in the market's ability to challenges.

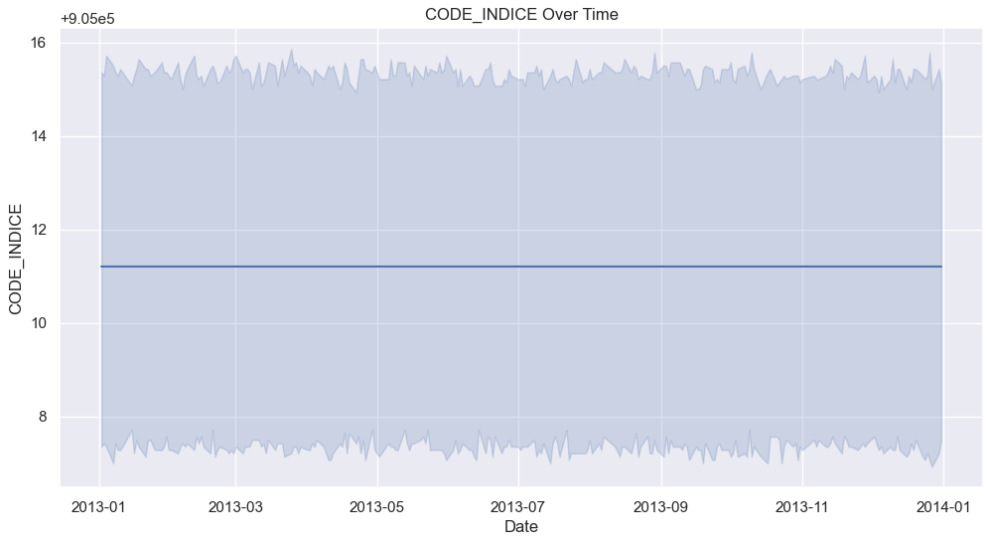
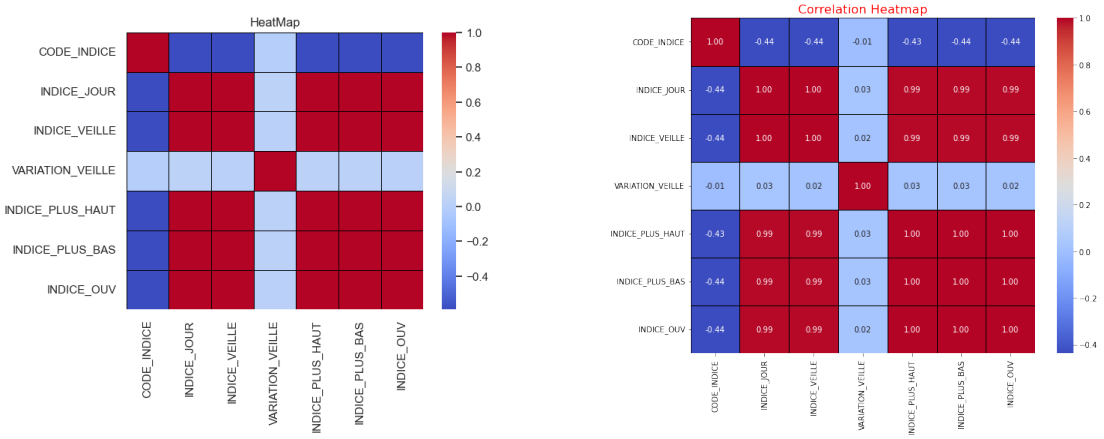


Figure 3.15: *Indexes consistency during 2013*

Temporal Stability: The index code values exhibit a consistent, straight-line trend throughout the year 2013, suggesting a period of stability and limited fluctuations in the tracked index during this timeframe.

Correlation Analysis

In the year 2020, which was marked by the COVID-19 crisis, we observe a noticeable increase in strong positive correlations among the market data indexes, as indicated by the prevalence of darker red squares in the heatmap. This suggests that during the crisis year, the market data indexes tended to move more in tandem compared to other years. This could be due to the global nature of the pandemic, which affected all sectors of the economy simultaneously, leading to more synchronized movements in the market data indexes



(a) *correlation between market data indexes for 2013*

(b) *correlation between market data indexes for 2019*

Figure 3.16: *correlation between different market data indexes for the years 2019 and 2013*

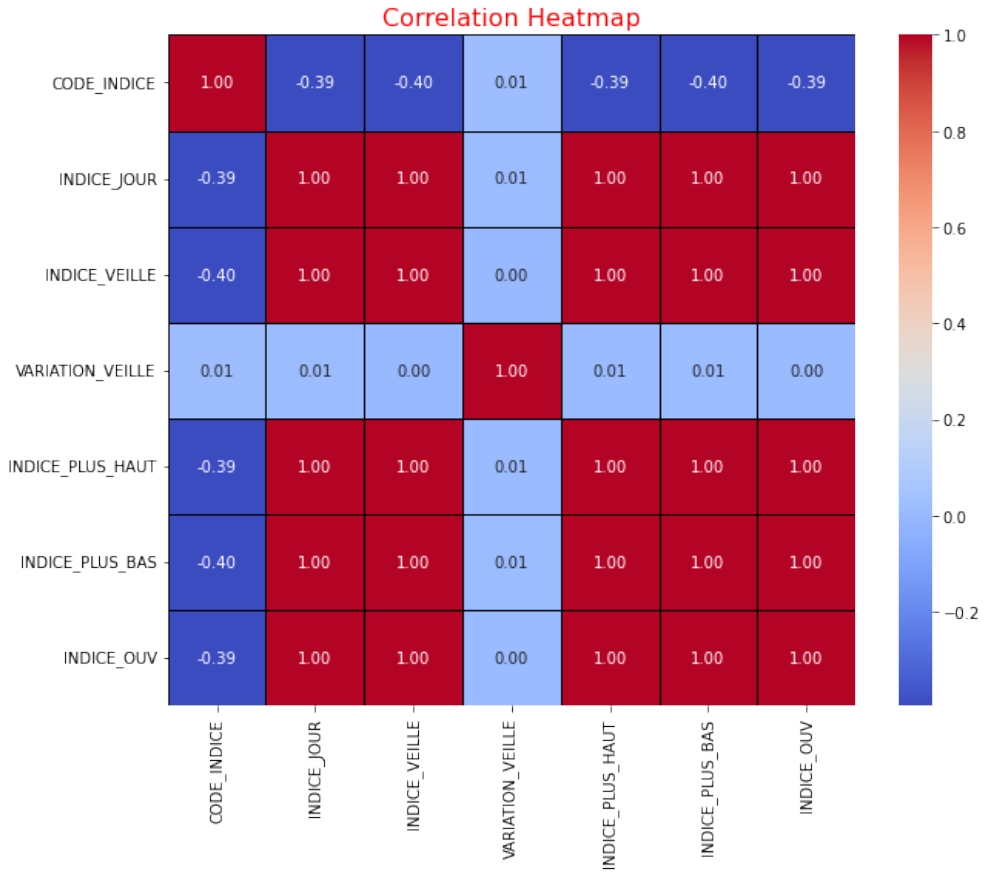


Figure 3.17: *correlation between market data indexes for 2020*

3.6.3 Market Companies

We conducted an analysis of the composition of our dataset, focusing on the distribution of companies across various sectors such as banks, leasing banks, and insurance companies. This analysis enables us to understand the relative prevalence of companies from key sectors and lays the foundation for more in-depth sector-specific investigations.

Key Points

1. Clean Data: The dataset is meticulously curated to ensure data integrity, with no missing values detected. This meticulous attention to data quality forms the foundation for robust analysis and informed decision-making.
2. Categorical Features: Crucial categorical features including Date, Vol. (Volume), Change (Change %), and Name (Company/Security Name) provide essential context for understanding market trends and dynamics.
3. Numerical Features: The dataset includes critical numerical metrics such as Price (Closing Price), Open (Opening Price), High (Highest Price), Low (Lowest Price), and Vol. (Volume), offering quantitative insights into market fluctuations and performance.

Outliers Analysis

In order to understand the Volatility of the market, we've used the range of values in the boxplot A larger range suggests higher volatility, which means higher risk but also potentially higher returns. We've also used outliers because they could indicate extraordinary events or errors in the data like the COVID-19 crisis, they could provide valuable insights into how such events might impact the market in the future.

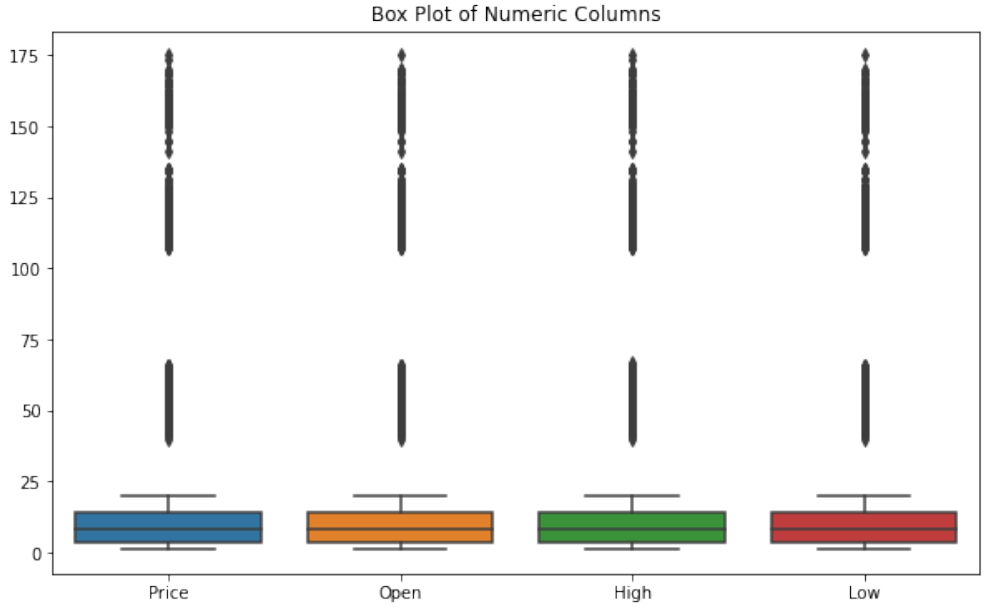


Figure 3.18: *Boxplot for Market companies data*

- Price: The median price is around 25, with a small interquartile range (IQR), indicating less variability in the price data. There are several outliers, suggesting some prices are significantly different from the rest.
- Open: The median open value is slightly higher than that of the price, with a similar small IQR. This suggests that the opening prices for the stocks are relatively stable, with a few exceptions.
- High: The median is similar to the 'Open' value, but there are more outliers above the upper whisker. This indicates that there are some instances where the highest price for the day was significantly higher than the typical range of high prices.
- Low: The median value is almost identical to 'Price' and 'Open', but there are outliers below the lower whisker. This suggests that there are some instances where the lowest price for the day was significantly lower than the typical range of low prices.

Data Cleaning

1. **Converting Date to DateTime:**
   The Date feature is converted into DateTime format to standardize date representation, enabling seamless temporal analysis and visualization.
2. **Converting Volume and Change % to Numeric:**
   Volume and Change % features are converted into numeric format to ensure consistency and compatibility for numerical computations and analysis.
3. **Dropping Unnecessary Columns:**
   Redundant columns such as Open, High, Low are dropped from the dataset to streamline the data structure and focus on essential information relevant to subsequent analysis.

4. **Mean Imputation for Missing Values:**
   Missing values, if any, are addressed through mean imputation. This approach ensures data completeness while minimizing bias, thereby preserving the integrity of the dataset.

Feature Engineering

- **Temporal Components (Day, Month, Year, Weekday):**
  Extracted from the Date feature, these temporal components provide granular insights into time-based patterns and trends. For instance, analyzing intraday patterns, seasonal trends, or weekday-specific anomalies becomes feasible with these extracted features.
- **IsWeekend:**
  This binary indicator distinguishes between weekend and weekday trading behavior, facilitating deeper analysis into how market dynamics vary across different days of the week.
- **DailyPriceChange:**
  Calculated as the percentage change in price from the previous day, this metric offers a comprehensive view of daily price movements and volatility, crucial for risk assessment and trading strategies.
- Moving Averages **(7-Day and 30-Day)**:
  These moving averages smooth out short-term fluctuations in price, providing a clearer picture of underlying trends and market sentiment over different time horizons.
- **Volume Metrics (VolumeChange, VolumePercentageChange):**
  These metrics capture changes in trading volume compared to the previous day, offering insights into shifts in market activity and investor interest.

Additional Features

- **DailyPriceAverage**:
  Represents the average price calculated from the opening, high, low, and closing prices, providing a smoothed measure of daily price levels.
- **HighLowRangePercentage**:
  Indicates the percentage difference between the high and low prices relative to the opening price, offering insights into intraday price volatility.
- RSI **(Relative Strength Index)**:
  A momentum oscillator that measures the speed and change of price movements, indicating overbought or oversold conditions in the market.
- **Bollinger Bands:**
  Boundary indicators based on volatility and standard deviation, used to identify potential price extremes and trading opportunities.
- **Volatility:**

Represents the standard deviation of daily price changes over a specified window, providing a measure of market uncertainty and risk.

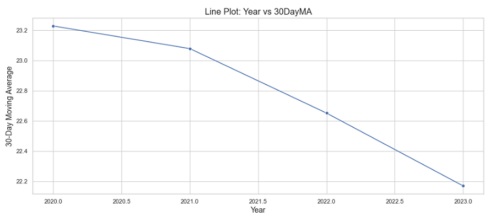
- **PriceUpDown:**
  A binary target variable indicating whether the price went up or down the next day, serving as a basis for classification or predictive modeling tasks.

These engineered features enhance the dataset's richness and analytical depth, assisting stakeholders to derive actionable insights and make informed decisions in the dynamic financial markets.

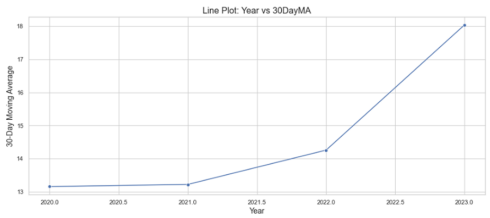
3.6.4 Analysis of Market Trends and Feature Insights

Trend Analysis: Insurance and Others vs. Banking

Insurance and Others Trend: The 30-day moving average consistently declines until 2023, suggesting a period of subdued performance or declining market sentiment within this sector.



(a) *30-day moving average for Insurance and other sectors*



(b) *30-day moving average for the Banking sector*

Figure 3.19: *trend analysis: 30-day moving average comparisons*

Banking Trend: Conversely, the 30-day moving average for the Banking sector exhibits a consistent upward trajectory from the 2021-2022 interval through 2023, indicating sustained growth and positive market sentiment.

Feature Insights: Volatility

Although all categories have an important rise in volatility after 2022, there are some differences across our datasets.

Volatility progressively increases over the years for the Insurance and Others dataset.

In contrast, the Banks dataset experiences a drastic rise in volatility in 2022 after a period of stability.

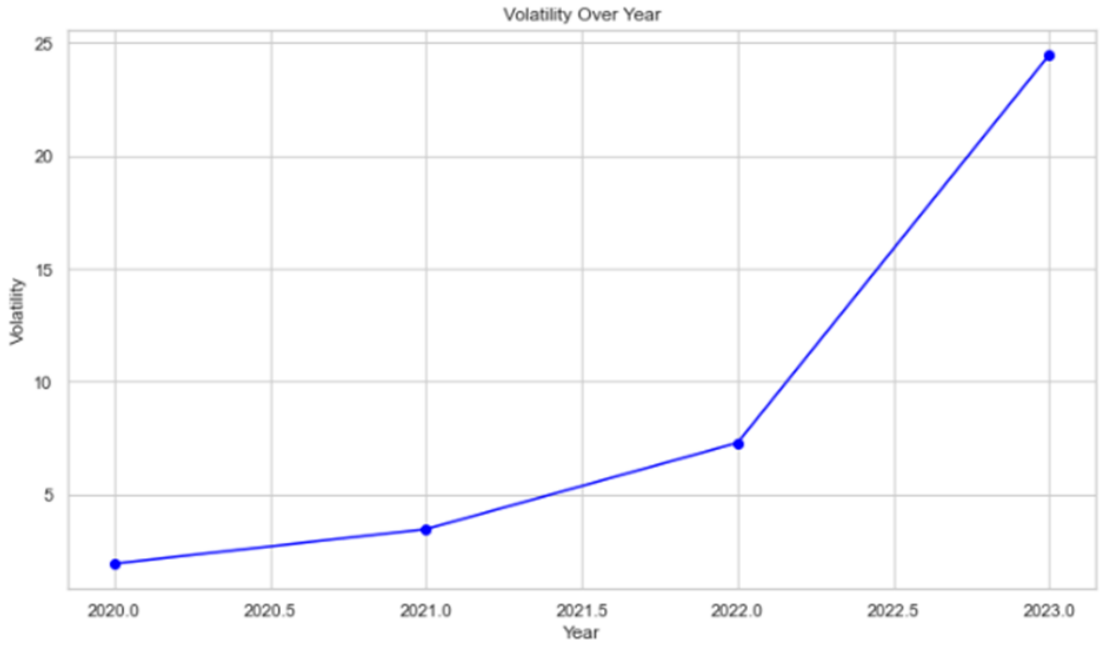


Figure 3.20: *Volatility Over Year - Insurance and Other sectors*

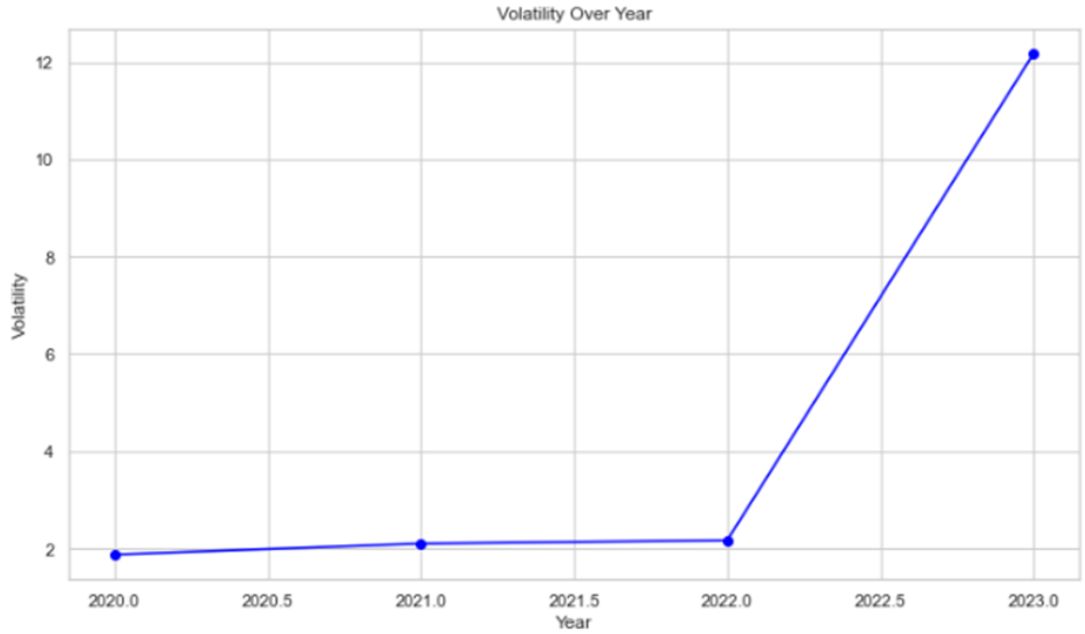


Figure 3.21: *Volatility Over Year - Banks*

Companies Price Movement Analysis

Across all datasets, prices generally exhibit a consistent downward trend over all weekdays. This observation suggests a prevailing bearish sentiment or downward pressure on prices irrespective of the day of the week.

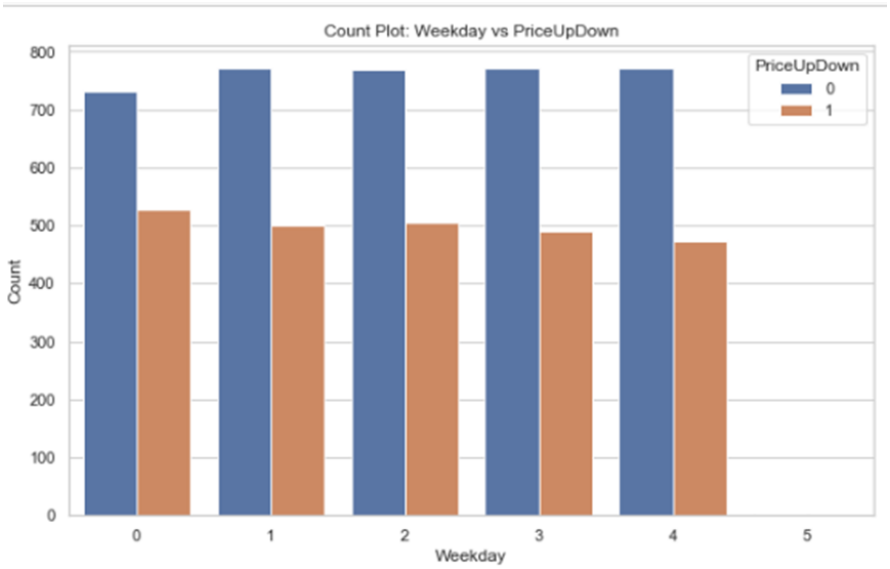


Figure 3.22: *Companies Price Movement Analysis Bar chart*

3.7 Cotation History

For the Cotation history we've concluded that 2017 could be a better year to make observations, we've concluded that over the observed period, the traded volume ('QUANTITE_NEGOCIEE') exhibited consistent flows with notable peaks in April 2017, November 2017, and a substantial surge in January 2018. These peaks suggest heightened market activity, potentially influenced by significant events or changes in investor sentiment.

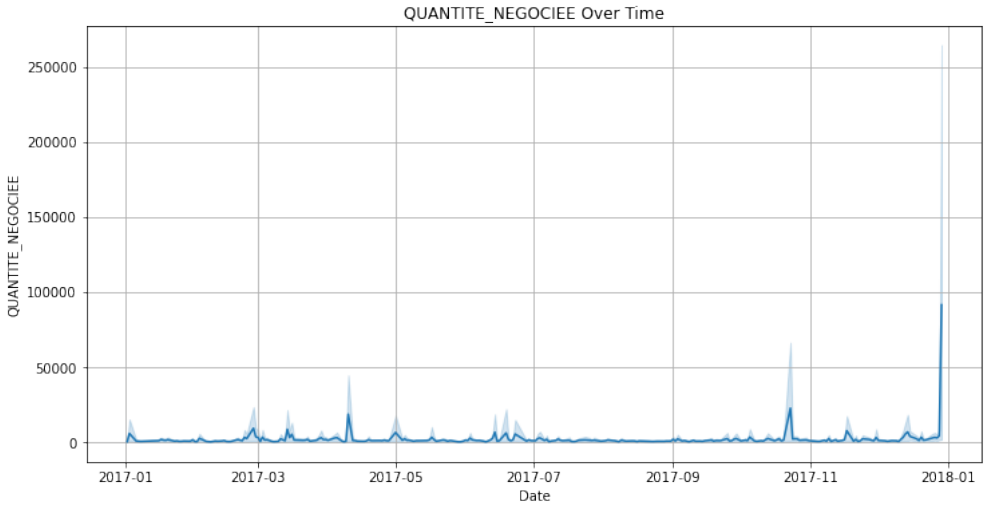


Figure 3.23: *trend of traded volume throughout the year 2017*

The number of transactions ('NB_TRAN') analysis highlighted peaks in January 2017 and sustained high activity in early 2017 to April 2017, indicating dynamic trading frequency.

In 2019, market activity witnessed notable fluctuations in both traded volume and

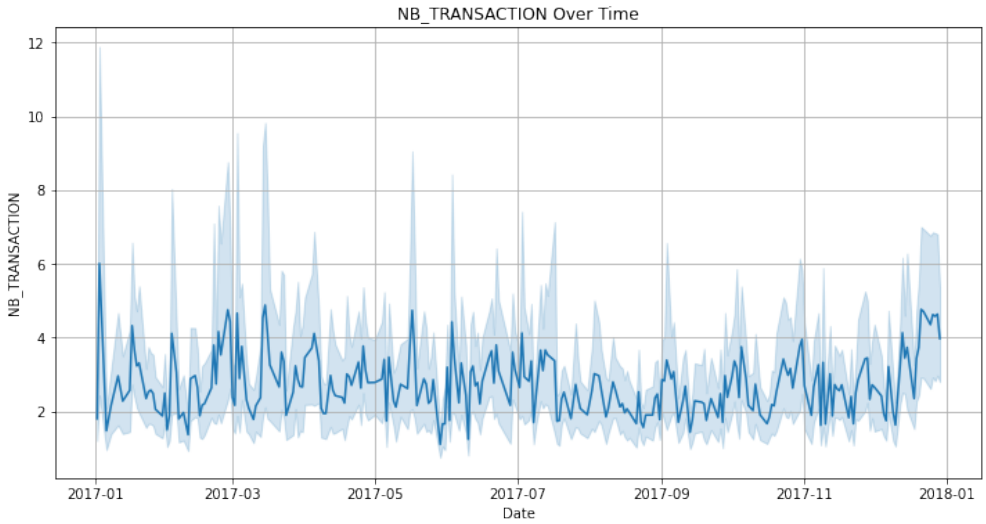
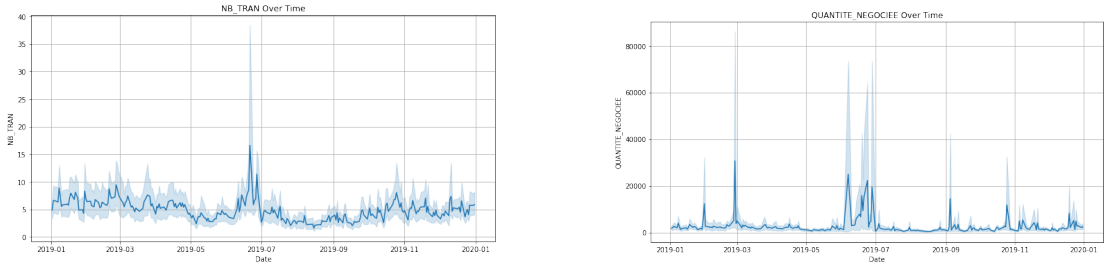


Figure 3.24: *trend of the number of transactions over the year 2017*

the number of transactions. A substantial surge in traded volume reached its pinnacle in March 2019, indicating heightened trading activity during that month. Following this peak, there were three subsequent lower peaks observed between June and July 2019, suggesting a sustained yet comparatively reduced level of trading intensity. Concurrently, the number of transactions exhibited stability initially, with minor variations. However, a significant spike in transaction numbers was observed in July 2019, indicating a noteworthy increase in trading activity during that specific period. Subsequent to this peak, the number of transactions returned to a stable phase until January 2020. The exponential growth in prices continued from January 2019 to January 2020, suggesting sustained positive market sentiment.



(a) *trend of the number of transactions over the year 2019*

(b) *trend of traded volume throughout the year 2019*

Figure 3.25: *Number of transactions and traded volume for 2019*

The year 2020 exhibited stability in the highest price, with notable short-term fluctuations. Trading volume had an extreme peak between January 2020 and February 2020, possibly related to a significant event, followed by a downfall and low volume until January 2021. The number of transactions displayed instability and responsiveness to market conditions.

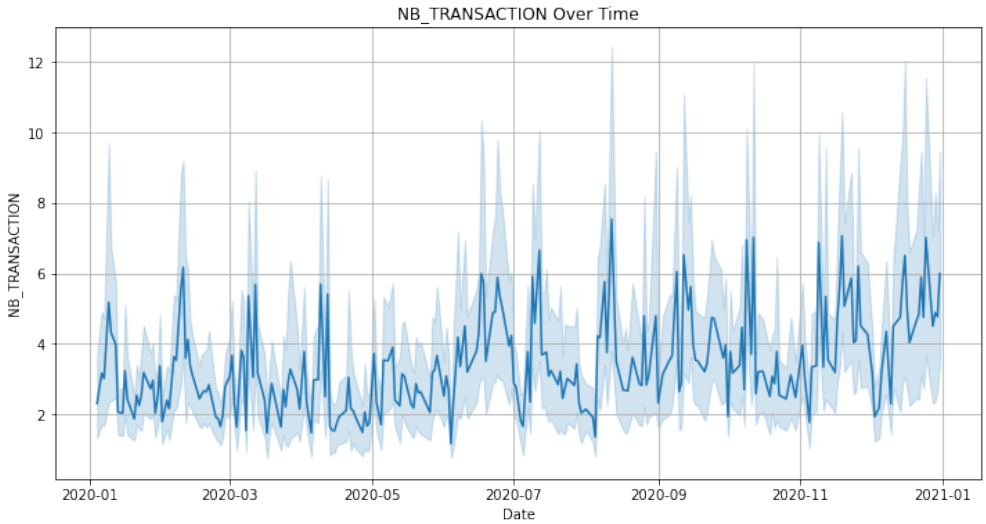


Figure 3.26: *trend of the number of transactions over the year 2020*

In 2022, a substantial surge in trading volume occurred between June and July, followed by sustained interest in September. The number of transactions rapidly increased in September 2022 and returned to a more normal range by November 2022. The analysis for 2023 reveals heightened market activity, with rapid variations in the number of transactions and distinct patterns in traded volume. A significant peak in January to February 2023 indicates a surge in trading, followed by a lower peak from May to June 2023 and additional peaks in April 2023 and July 2023. These patterns offer valuable insights into market dynamics and provide a foundation for the development of intelligent trading agents, allowing them to adapt to changing conditions and capitalize on periods of notable trading activity.

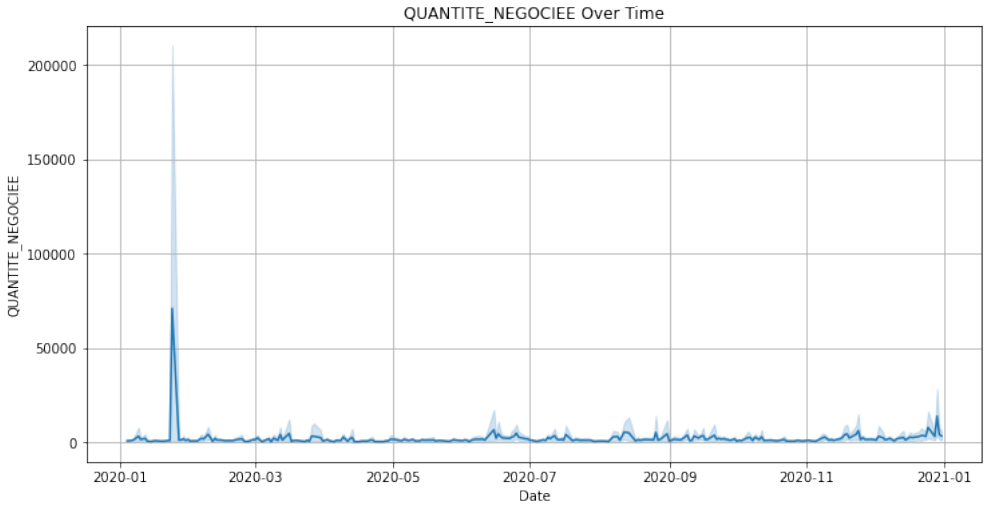


Figure 3.27: *trend of traded volume throughout the year 2020*

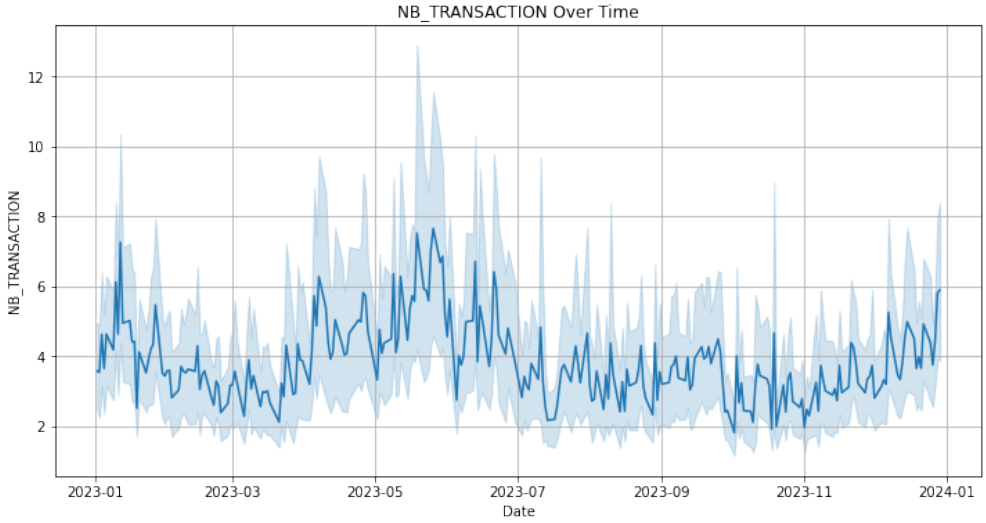


Figure 3.28: *trend of transactions throughout the year 2023*

3.8 Macroeconomic Data

Data Cleaning and Preprocessing

Prior to analysis, the macroeconomic datasets undergo rigorous cleaning and preprocessing to ensure data integrity and consistency:

- Handling missing values: Missing data points were identified in the US data but not in the European one. They were addressed through techniques such as imputation or deletion.
- Standardizing units: indicators are reported in different units or scales, normalization techniques were applied to standardize the data for comparative analysis.
- Removing duplicates: Duplicate entries or redundant observations were removed to streamline the dataset and prevent duplication of information.

Furthermore, the datasets may be transformed or aggregated to derive additional insights or construct composite indicators that better capture underlying economic trends.

3.8.1 Analysis and Insights

In the subsequent sections, we present a detailed analysis of the macroeconomic data, exploring trends, correlations, and causal relationships between different economic indicators. By examining the dynamics of GDP growth, inflation, unemployment, trade balances, and other key variables, we aim to provide valuable insights into the broader economic landscape and its implications for local market dynamics and investment opportunities.

Analysis of US Macroeconomic Data

The analysis of US macroeconomic data reveals several notable correlations and insights that provide valuable context for understanding the broader economic landscape. These insights shed light on the interplay between various economic indicators and their potential implications for market dynamics and investment strategies.

- **S&P 500 Index and S&P P/E Ratio:** The moderate positive correlation between the S&P 500 index and the price-to-earnings (P/E) ratio suggests that changes in stock prices are accompanied by shifts in market valuations relative to earnings. This correlation underscores the importance of considering both market performance and valuation metrics when assessing investment opportunities.
- **S&P 500 Index and Volatility Index:** The weak positive correlation between the S&P 500 index and market volatility implies that fluctuations in market performance have limited predictive power for volatility levels. This finding highlights the need for investors to incorporate additional risk management strategies beyond relying solely on market trends.
- **Retail Sales and Personal Consumption Expenditures:** The very high positive correlation between retail sales and personal consumption expenditures underscores the close relationship between consumer spending and overall economic activity. This correlation suggests that changes in consumer behavior can have significant implications for economic growth and retail sector performance.
- **Unemployment Rate and Total Nonfarm Payroll:** The very weak negative correlation between the unemployment rate and total nonfarm payroll suggests that changes in employment levels have minimal direct impact on the unemployment rate. This finding emphasizes the complex dynamics at play in labor markets and the need for nuanced analysis when interpreting employment data.
- **Consumer Price Index (CPI) and Producer Price Index (PPI) for Metals:** The high positive correlation between the CPI and PPI for metals highlights the interconnectedness of consumer prices and producer costs, particularly in industries

reliant on metal inputs. This correlation underscores the potential for input cost fluctuations to influence consumer inflation trends.

- **Government Employment and Federal Funds Rate:** The moderate negative correlation between government employment and the Federal Funds Rate suggests a potential relationship between fiscal policy and monetary policy decisions. This correlation implies that changes in government employment levels may influence central bank decisions regarding interest rates.
- **Trade Weighted Exchange Rate and CAD to USD Exchange Rate:** The moderate positive correlation between the trade-weighted exchange rate and the CAD to USD exchange rate indicates a relationship between broader currency movements and specific exchange rate dynamics. This correlation underscores the importance of considering global currency trends when analyzing exchange rate fluctuations.

Overall, the analysis of US macroeconomic data provides valuable insights into the complex interdependencies within the economy and their implications for financial markets. By understanding these relationships, investors can make more informed decisions and better navigate the dynamic economic landscape.

Analysis of European Macroeconomic Data

The analysis of European macroeconomic data provides valuable insights into the structure and dynamics of economic indicators across different countries and sectors. By examining the distribution, correlations, and trends within the dataset, we can gain a deeper understanding of the European economic landscape and its implications for investment decisions.

- **Data Description:** The dataset comprises 85 features, including categorical and numerical variables representing various economic indicators and attributes. The presence of unnamed features with all NaN values suggests that certain columns are not informative and can be dropped from further analysis. Key features such as SERIES, CNTRY, TRN, AGG, UNIT, REF, CODE, SUB-CHAPTER, and TITLE provide essential context for interpreting the data and understanding its structure.
- **Data Visualization:** The distribution of binary indicators such as TRN and AGG reveals insights into transaction frequencies and aggregation patterns within the dataset. For instance, the presence of more than 25,000 transactions of a certain type suggests common transactional activities, while instances where aggregated values approach 1.0 indicate high exposure to specific asset classes or investment strategies. Additionally, the correlation matrix provides a comprehensive overview of the relationships between different financial variables across European markets from 1960 to 2025.
- **Correlation Analysis:** The correlation matrix highlights the associations between various financial variables, such as transportation stocks (TRN), aggregate stocks (AGG), units, and reference indices (REF). Moderate positive correlations between

TRN and REF suggest some degree of association between transportation stocks and reference indices, while weak negative correlations with UNIT indicate inverse relationships with units. Similarly, moderate positive correlations between AGG and TRN imply stronger associations between aggregate stocks and transportation stocks.

- **Yearly Trends:** The correlation trends across different years indicate a gradual increase in correlations between financial variables from 1960 to 2025. This trend reflects a growing integration or interdependence among European financial markets over the decades, highlighting the evolving nature of economic relationships within the region.

Overall, the analysis of European macroeconomic data provides valuable insights for investors and analysts seeking to understand the complexities of the European economic landscape. By leveraging these insights, stakeholders can make more informed decisions and navigate the dynamic European markets effectively.

Conclusion

In conclusion, our data understanding and acquisition efforts, including web scraping and market, financial and macroeconomics data analysis, provided us with invaluable insights into financial market trends and dynamics. This chapter lays a solid foundation for subsequent analyses and modeling in our exploration of the financial markets.

4

Modelling

In this phase of the Team Data Science Process (TDSP), we focus on developing and implementing models to achieve the defined business objectives (BOs) and corresponding data sub-objectives (DSOs). Leveraging advanced techniques in machine learning, natural language processing (NLP), and predictive analytics, we aim to provide valuable insights and solutions to support decision-making in financial trading and investment.

4.0.1 Accomplished Business Objectives

During the modeling phase, we successfully addressed several key business objectives, as outlined below:

Business Objective (BO1)

Through rigorous data preprocessing and cleansing techniques, we ensured that the data used for modeling is reliable and suitable for analysis.

Business Objective (BO2)

Using machine learning algorithms and historical market data, we developed robust predictive models capable of forecasting price movements and evaluating risk factors associated with various financial instruments.

Business Objective (BO3)

By leveraging state-of-the-art NLP techniques and language models, we gained valuable insights from textual data sources, such as news articles and financial reports, enabling a deeper understanding of the Tunisian financial markets.

Business Objective (BO4)

Through advanced feature engineering and analysis of dividend profiles, we gained insights into the risks and returns associated with various financial assets, facilitating informed investment decisions.

Business Objective (BO5)

By employing clustering and segmentation techniques, we identified meaningful patterns within the market data, enabling us to uncover actionable insights for enhancing trading outcomes and market understanding.

4.1 Modeling: Risk Prediction in Financial Trading

Risk assessment and prediction through dimensionality reduction and clustering offers valuable insights for our companies across the different categories of our data: insurance leasing, banking, and investment funds (SICAV).

4.1.1 Price Movement Prediction

The primary objective of this analysis is to develop a reliable model capable of predicting whether the price of a financial instrument will rise or fall on a given date. Such a predictive tool is invaluable in financial trading as it empowers traders and investors to make informed decisions regarding buy or sell actions, thereby potentially capitalizing on market movements and maximizing returns while minimizing risks.

4.1.2 Model Evaluation Metrics

1. Accuracy: Across all models, accuracy ranged between 0.60 and 0.61, indicating a moderate level of predictive capability.
2. Precision, Recall, and F1-Score: Notably, all models demonstrated higher precision and recall for class 0 (indicating price increase) compared to class 1 (indicating price decrease), suggesting a potential class imbalance.
3. Cumulative Profit/Loss: Remarkably, all models yielded the same cumulative profit/loss of 158.495, indicating a consistent but modest financial outcome.
4. Sharpe Ratio: The Random Forest classifier exhibited a superior Sharpe ratio of 0.964 compared to other models, indicating a more favorable risk-adjusted return.
5. Brier Score: Across all models, Brier scores hovered around 0.24, suggesting fair but not optimal calibration.

4.1.3 Integration of Optuna and SMOTE

To enhance the predictive performance of the classifiers, we employed Optuna for hyperparameter optimization and SMOTE for addressing class imbalance. Optuna's optimization framework enabled us to iteratively search for the optimal combination of hyperparameters, such as boosting type, learning rate, number of leaves, maximum depth, and number of estimators, to maximize classification accuracy. Additionally, SMOTE was applied to the training data to synthetically generate new instances of the minority class, ensuring a more balanced distribution and improving the models' ability to generalize.

4.1.4 Model Selection

After comprehensive evaluation, the Random Forest classifier emerged as the most promising model for predicting price movement on the inserted date. It achieved comparable performance in terms of accuracy, precision, recall, and F1-score, while also

outperforming other models in terms of risk-adjusted return, as evidenced by its higher Sharpe ratio (0.964).

4.1.5 Conclusion

In conclusion, the integration of Optuna for hyperparameter optimization and SMOTE for addressing class imbalance significantly enhanced the performance of the Random Forest classifier in predicting price movement. Such predictive capability is invaluable in financial trading, enabling traders to make informed decisions, capitalize on market movements, and optimize trading strategies for favorable risk-adjusted returns.

4.2 Risk Assessment and Prediction for Financial Trading

In financial trading, accurately assessing and predicting the risk associated with different stocks is paramount for making informed investment decisions. Here, we outline the essential steps taken to assess and predict risk using a combination of technical indicators and machine learning algorithms.

4.2.1 Data Preprocessing and Risk Scoring

- **Technical Indicators**: We consider Relative Strength Index (RSI), Volatility, and Volume Change as key indicators of a stock's risk profile.
- **Risk Scoring Function**: Each indicator is assigned a score based on predefined thresholds, reflecting their impact on risk.
- **Overall Risk Score**: The individual scores are aggregated to derive an overall risk score for each stock.

4.2.2 Training the Model

- **Feature Selection**: RSI, Volatility, and Volume Change are selected as features for training the model.
- **Target Variable**: Risk levels (Low/High) based on the derived risk score are used as the target variable.
- **Train-Test Split**: The dataset is divided into training (80%) and testing (20%) sets to evaluate model performance.

4.2.3 Model Training and Evaluation

- **Algorithm Selection**: We employ an AdaBoostClassifier, a popular ensemble learning algorithm known for its ability to improve classification accuracy.
- **Model Training**: The AdaBoostClassifier is trained on the training dataset.
- **Model Evaluation**: The model's accuracy, precision, recall, and F1-score are evaluated using the test dataset.

4.2.4 Performance Metrics

- **Risk Profiling Accuracy (RPA)**: Measures the overall correctness of risk profiling for each stock.
- **High-Risk Precision (HRP)**: Evaluates the precision of identifying high-risk stocks.
- **Low-Risk Recall (LRR)**: Measures the recall of identifying low-risk stocks.
- **Risk Classification Fidelity (RCF)**: Evaluates the fidelity of risk classification across different risk levels.

4.2.5 Model Deployment and Prediction

- **Model Saving**: The trained **AdaBoostClassifier** model is saved for future use.
- **Model Loading**: The saved model can be loaded to make real-time predictions.
- **User Input**: Users can input RSI, Volatility, and Volume Change for a single stock.
- **Prediction**: The loaded model predicts the risk level (Low/High) for the given stock.

4.2.6 Results and Conclusion

- The model demonstrates high accuracy, precision, recall, and overall fidelity in risk classification, ensuring robust risk assessment for financial trading.
- With this approach, investors can make informed decisions by considering the predicted risk levels of individual stocks, thus optimizing their investment strategies and minimizing potential losses.

4.2.7 Future Considerations

Continuous monitoring and refinement of the model based on evolving market conditions and additional relevant features can further enhance its predictive power and applicability in real-world trading scenarios.

4.3 Risk Assessment through Dimensionality Reduction and Clustering

We employed various clustering algorithms on financial trading data to identify distinct market segments based on risk, return, and volatility characteristics. The dataset consisted of daily stock price data, which was preprocessed to extract relevant features and then subjected to dimensionality reduction using UMAP (Uniform Manifold Approximation and Projection).

4.3.1 Schema

Features Used:

- Date: The date of each trading day.

- Price: The closing price of the stock on each trading day.
- Vol.: The volume of stocks traded on each trading day.
- Change %: The percentage change in the stock price from the previous trading day.
- Day: The day component extracted from the date.
- Month: The month component extracted from the date.
- Year: The year component extracted from the date.
- Price Skewness: The skewness of the stock price distribution over a rolling window of 30 days.
- Price Kurtosis: The kurtosis of the stock price distribution over a rolling window of 30 days.
- Risk: The standard deviation of the stock price over a rolling window of 30 days, representing the riskiness of the stock.
- Return: The simple daily returns of the stock price.
- Median: The median of the stock price over a rolling window of 30 days.
- Q1: The first quartile (25th percentile) of the stock price over a rolling window of 30 days.
- Q3: The third quartile (75th percentile) of the stock price over a rolling window of 30 days.
- Average Q1: The average of the first quartile (Q1) of the stock price over a rolling window of 30 days.
- Average Q3: The average of the third quartile (Q3) of the stock price over a rolling window of 30 days.

4.3.2 Clustering Algorithms

Three clustering algorithms were applied:

1. **K-means Clustering**: Utilized to partition the data into four clusters based on similarities in risk, return, and volatility. The resulting clusters were interpreted as representing different risk-return profiles in the financial market.
2. **Hierarchical Agglomerative Clustering (HAC)**: Employed to group data points into clusters hierarchically, capturing underlying structures in the data. Similar to K-means, the clusters were interpreted based on risk, return, and volatility characteristics.
3. **Density-Based Spatial Clustering of Applications with Noise (DBSCAN)**: Used to identify clusters of varying densities in the data, effectively capturing outliers and noise. Clusters were again interpreted based on risk-return-volatility profiles.

4.3.3 Results and Interpretations

Upon performing the clustering analysis, distinct clusters emerged, each representing a unique combination of risk, return, and volatility:

- **K-means Clustering**: Four clusters were identified, characterized by different risk-return profiles. These clusters ranged from low-risk, low-return segments to very high-risk, very high-return segments, providing insights into the diverse market dynamics. The evaluation metrics used for K-means included:
  - **Cluster Cohesion Metric (CCM)**: The average distance between each data point and its corresponding centroid within a cluster. For K-means, the CCM was found to be 4.22, indicating relatively tight clusters.
  - **Inter-Cluster Distance Ratio (ICDR)**: The ratio of the total distance between cluster centroids to the average distance between data points and their respective centroids. In the case of K-means, the ICDR was calculated to be approximately 2.0, suggesting moderate separation between clusters.
  - **Noise Ratio (NR)**: The proportion of data points classified as noise by the algorithm. In K-means, the NR was found to be 0.0, indicating a clean separation of clusters without any outliers.

Figure 4.1: *U-Map Embeddings with K-means Clusters*

- **Hierarchical Agglomerative Clustering (HAC)**: The hierarchical clustering approach revealed similar clusters to K-means, albeit with slight variations in grouping. The evaluation metrics mirrored those of K-means, providing insights into cluster cohesion, inter-cluster distances, and noise levels.
  - **Cluster Cohesion Metric (CCM)**: For HAC, the CCM was calculated to be 3.54, indicating a slightly lower average distance between data points and their respective centroids compared to K-means.
  - **Inter-Cluster Distance Ratio (ICDR)**: The ICDR for HAC was 2.50, suggesting a similar level of separation between clusters as observed in K-means.
  - **Noise Ratio (NR)**: Like K-means, HAC also exhibited a NR of 0.0, indicating a clean separation of clusters without any outliers.

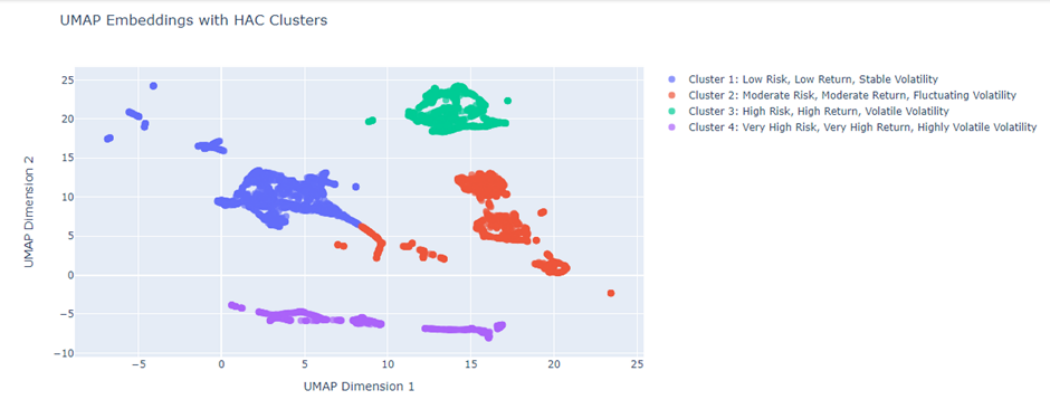


Figure 4.2: *U-Map Embeddings with HAC Clusters*

4.3.4 Results and Interpretations (Continued)

- **DBSCAN**: While DBSCAN identified clusters similar to K-means and HAC, its focus on density-based clustering provided additional insights into outlier detection and noise estimation, which could be crucial in risk management strategies. Metrics such as cluster cohesion, inter-cluster distance ratio, and noise ratio were used to evaluate the performance of DBSCAN in capturing meaningful clusters while identifying outliers and noise points.
  - **Cluster Cohesion Metric (CCM)**: DBSCAN yielded a CCM of 1.73, indicating a relatively higher average distance between data points and their respective centroids compared to K-means and HAC.
  - **Inter-Cluster Distance Ratio (ICDR)**: The ICDR for DBSCAN was calculated to be 17.00, indicating a significant distance between cluster centroids, which might suggest less distinct separation between clusters compared to K-means and HAC.
  - **Noise Ratio (NR)**: Similar to K-means and HAC, DBSCAN also exhibited a NR of 0.0, indicating a clean separation of clusters without any outliers.

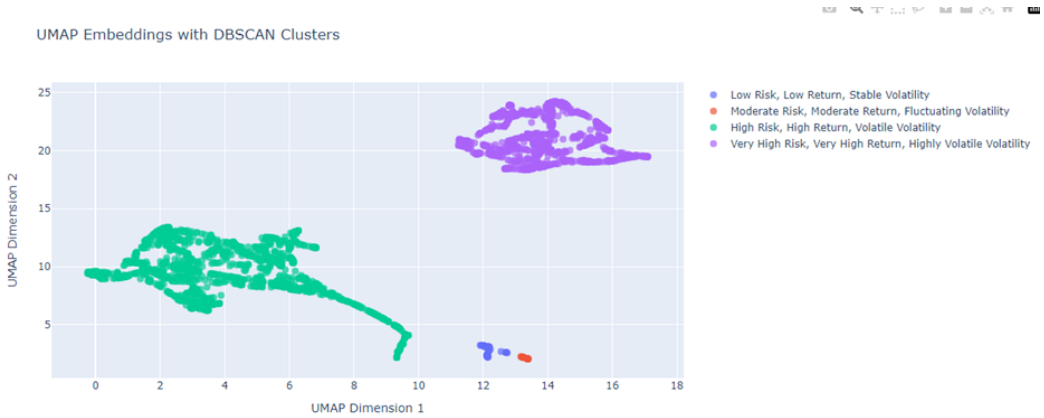


Figure 4.3: *U-Map Embeddings with DBScan Clusters*

Objective

The purpose of this modeling endeavor in a financial trading context is multifold:

1. **Risk Assessment**: By categorizing market segments based on risk, return, and volatility, traders and investors can better assess and manage their risk exposure within their portfolios.
2. **Market Sentiment Analysis**: Clustering analysis provides insights into market sentiment and investor behavior, aiding in decision-making processes such as market entry, exit, and asset allocation.
3. **Algorithmic Trading Strategies**: The identified clusters can serve as inputs for developing algorithmic trading strategies, automating trading decisions based on predefined risk-return parameters.

In summary, the clustering analysis conducted on financial trading data offers valuable insights into market segmentation, risk assessment, and portfolio management, contributing to informed decision-making and strategy formulation in the dynamic world of financial markets.

4.4 MARKET INDEXES DATA SEGMENTATION

4.4.1 Feature Selection

we employed variance thresholding as a method for feature selection to identify the most significant features for our clustering model. Variance thresholding is a straightforward approach that removes features with low variance, under the assumption that features with low variance do not contribute significantly to the model's predictive power.

```
Selected features after variance thresholding: Index(['INDICE_JOUR', 'Percentage_Change', 'Volatility', 'Gap', 'Gap_Up',
       'Gap_Down', 'MA_5', 'MA_10', 'Upper_BB', 'Lower_BB'],
      dtype='object')
```

Figure 4.4: *variance thresholding feature selection*

Upon applying variance thresholding to our dataset, we observed that all features were retained. This outcome indicates that each feature in our dataset possesses a sufficient level of variance, suggesting that they all contribute valuable information. This is a crucial finding as it confirms the relevance and necessity of each feature in our analysis, ensuring that no critical data is overlooked in our clustering and segmentation processes.

U-MAP Segmentation

The U-MAP Technique is very relevant to our project, in the figure below, we can observe the clusters obtained form the market indexes data we've obtained.

Figure 4.5: *Market Indexes clusters with U-MAP*

There are two main clusters visible in the image, differentiated by colors yellow and purple.

Cluster Locations

The purple cluster is primarily located in the central to right portion of the plot, with UMAP1 values ranging roughly from 8.5 to 10. The yellow cluster is more spread towards the left, with UMAP1 values from about 7.5 to 8.5.

Cluster Composition

- Yellow Cluster: Includes indices like "AGROALIMENT BOISSO", "ASSURANCES" and "BIENS CONSOMMATION".
- Purple Cluster: Comprises indices such as "TUNINDEX", "SERVICES FINANCIER", "STES FINANCIERES", "BANQUES", "MATERIAUX DE BASE", "SERVICES AUX CONSO", "BATIM" and "PROD MENAGER SOIN".

Results and Interpretat

The separation along the UMAP dimensions indicates that these clusters have distinct features that are being captured and visualized through the UMAP reduction technique.

This silhouette score provides a quantitative basis for evaluating the effectiveness of the clustering and supports the visual insights gained from the UMAP plot. It suggests that the clustering approach is reasonably effective, although there may be room for further refinement or exploration of alternative clustering parameters to enhance the separation and definition of the clusters.

Market indexes segmentation using PCA

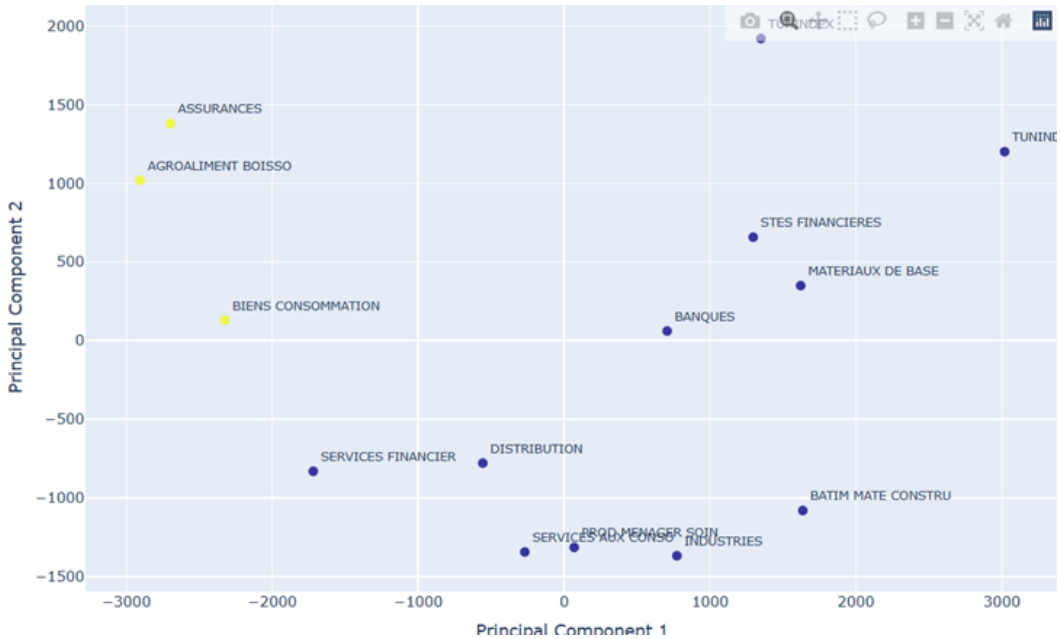


Figure 4.6: *Market indexes segmentation using PCA*

Sectors like "TUNINDEX" and "STES FINANCIERES" are positioned far to the right, suggesting they might have unique characteristics or are influenced heavily by factors captured predominantly in PC1. "SERVICES FINANCIER" and "DISTRIBUTION" are near the center, indicating a more balanced influence from the underlying metrics. Vertical distribution (along PC2) might suggest sensitivity to different types of financial metrics not captured by PC1. This PCA plot is useful for visualizing the overall structure of the data and understanding the relationships and differences among various sectors in the market. It helps in identifying which sectors are outliers and which ones share common characteristics, potentially guiding investment decisions or further financial analysis.

Conclusion

To conclude, both UMAP and PCA provided relatively good results in visualizing and segmenting the market index data. However, in most of our datasets, UMAP proved to be a bit more effective. UMAP tends to preserve more of the local structure of the data, making it particularly useful for capturing the nuances of complex datasets, which might explain its slightly superior performance in our analysis.

4.5 Dividend Profiles

In this section, we delve into the segmentation of companies based on their dividend policies. Dividends play a crucial role in investor decision-making and are indicative of a company's financial health and shareholder value proposition. By segmenting companies based on their dividend characteristics, we aim to uncover distinct groups with varying dividend behaviors.

4.5.1 Dividend Segmentation

K-means Clustering

The K-Means clustering algorithm identified 2 main clusters in the data based on the 'Average DL (Dividend Liquidity)' feature.

- 61 companies were assigned to cluster 0 and 11 companies to cluster 1.
- Cluster 0 has a mean 'Average DL' of 0.186966 which is lower than cluster 1's mean of 0.722078.
- Cluster 0 generally has lower 'Average DL' values based on the statistics. The minimum, 25%, 50% and 75% values are all lower than cluster 1.
- Companies assigned to cluster 0 have an 'Average DL' ranging from 0.02 to 0.434286 (lower end of the scale).
- Companies in cluster 1 have an 'Average DL' ranging from 0.463571 to 1.396429 (higher end of the scale).
- The new 'Needs_Improvement' column identifies companies in cluster 0 (with lower Average DL) as needing potential improvement, while cluster 1 companies currently do not need improvement.

4.5.2 Regression Analysis

Dividend Prediction

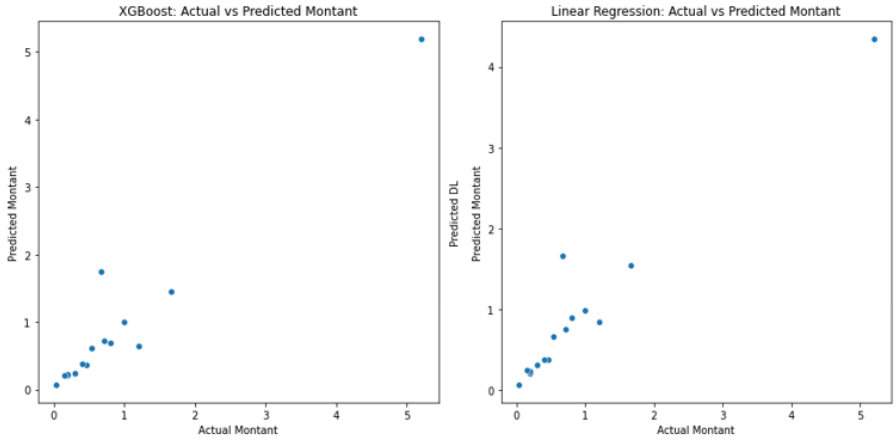


Figure 4.7: *Predicted Values vs Actual Values of Dividends*

The XGBoost model showed tighter clustering of data points around the regression line compared to linear regression, indicating superior ability to learn complex patterns and make more precise dividend predictions. XGBoost outperformed linear regression, likely due to its capability to model nonlinear relationships effectively.

Dividend Liquidity Prediction

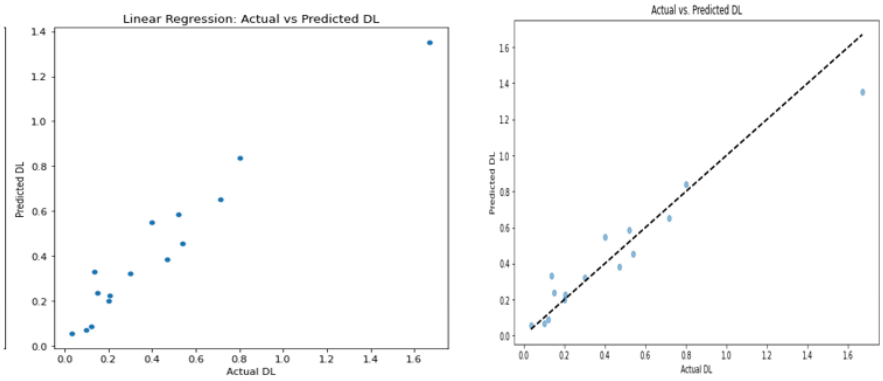


Figure 4.8: *Predicted Values vs Actual Values of Dividends Liquidity*

Random forest demonstrated superior performance in predicting dividend liquidity compared to linear regression, as evidenced by tighter point clustering around the trendline. This suggests random forest's ability to capture complex relationships in the data, leading to more accurate predictions.

Model Evaluation

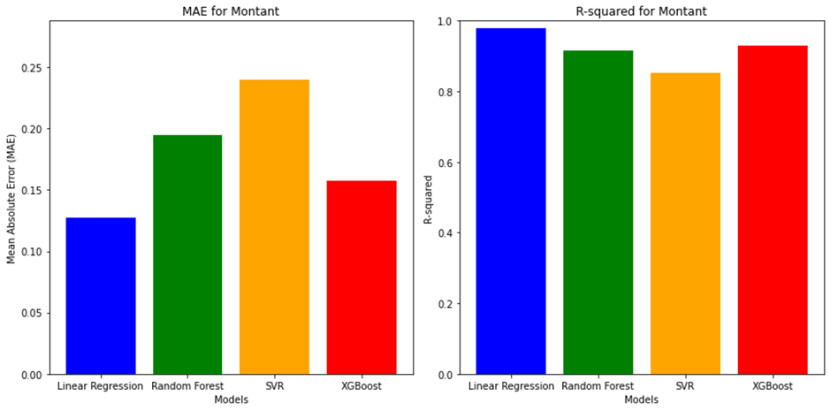


Figure 4.9: *Models Comparison for Dividends prediction*

XGBoost exhibited the lowest prediction errors and highest R-squared value (0.9309099) among models, indicating superior accuracy and explanation of variability in dividend predictions. Linear regression, while achieving high R-squared, showed limited predictive ability compared to XGBoost.

Dividend Liquidity Prediction:

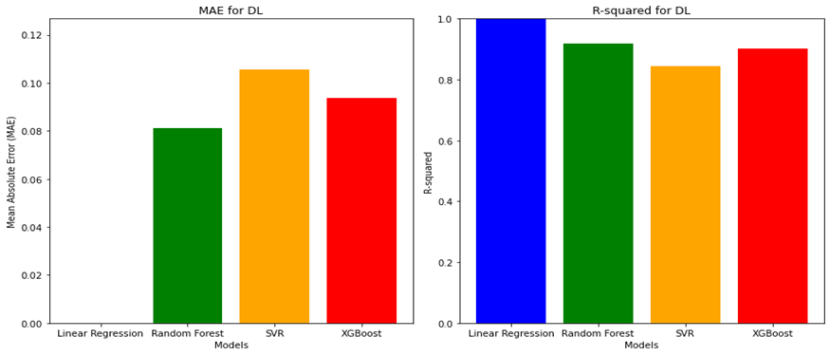


Figure 4.10: *Models Comparison for Dividends Liquidity prediction*

XGBoost attained the lowest prediction errors and highest R-squared (0.903053), outperforming other models in predicting dividend liquidity. Random forest also showed competitive performance, highlighting its effectiveness in capturing DL patterns.

In conclusion, XGBoost emerged as the top-performing algorithm in both dividend and dividend liquidity prediction tasks, showcasing its superior predictive capability and ability to model complex relationships in the data.

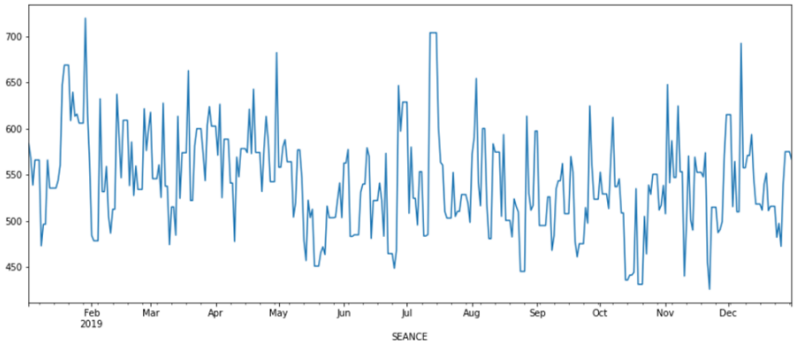


Figure 4.11: *SEANCE column index*

4.5.3 Time Series Forecasting

A fundamental step in time series analysis is indexing dates, which we achieve by setting the SEANCE column as the dataframe's index.

Given the heterogeneous observation points within a day, resampling the time series data is imperative for consistency and meaningful analysis.

Time series analysis employs statistical methods to discern trends, patterns, and predict future values, often depicted using line charts for visual clarity.

Decomposing time series data elucidates underlying trends and variations, offering insights crucial for effective forecasting. Model evaluation entails rigorous examination

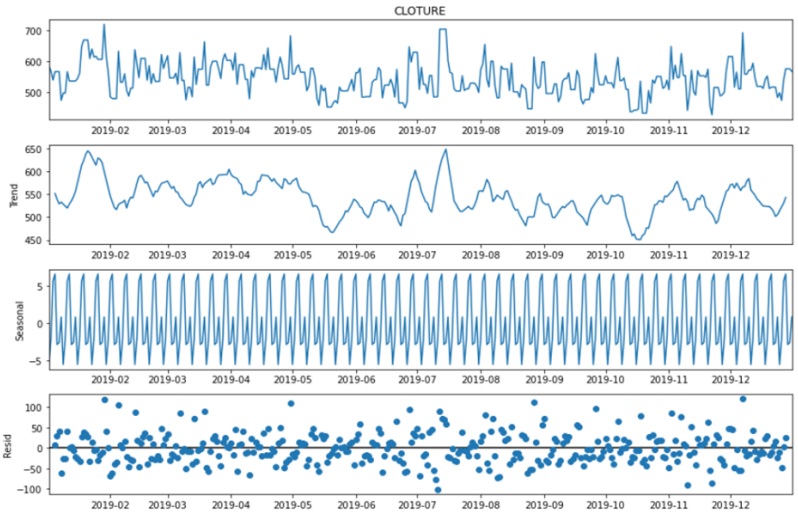


Figure 4.12: *Time series decomposition*

of standardized residual plots, histograms with density estimation, normal Q-Q plots, and correlograms, ensuring robustness and reliability of the forecasting model.

The final visualization juxtaposes observed data with forecasted values, providing a comprehensive view of the model's predictive performance, with observed data represented in blue and forecasted values along with confidence intervals depicted in orange with grey shading.

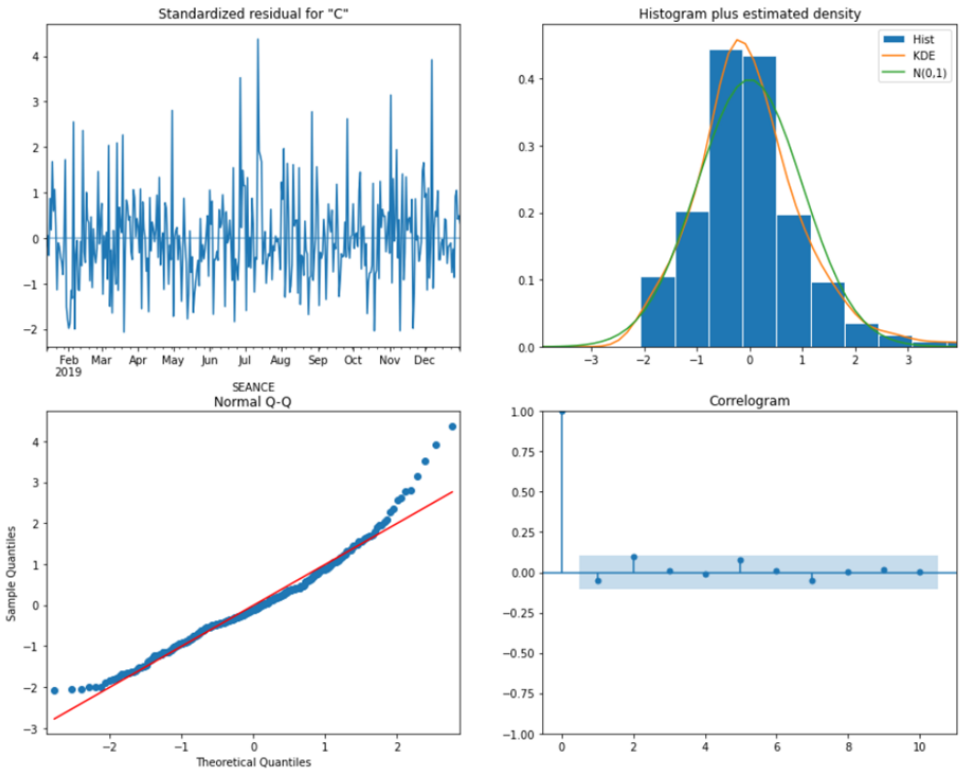


Figure 4.13: *Time series analysis*

4.5.4 Forecasting Models Evaluation

The analysis compares three forecasting models: Prophet, LSTM, and XGBoost, focusing on their performance in predicting the closing price of the TUNALIM index.

Prophet

Prophet forecasts the closing price from 2013 to 2025, showing a declining trend from 2023 onwards. The widening shaded area indicates increasing uncertainty in the forecast. The model suggests a bearish outlook for the index, with prices expected to decrease over the next few years.

Figure 4.14: *Closing price using Prophet*

LSTM

The LSTM model displays the actual and predicted closing prices over time. While the red line closely follows the blue line, indicating overall accuracy, there are deviations, particularly during peaks and sharp movements. Overall, the LSTM model performs reasonably well in capturing trends and patterns in the data.

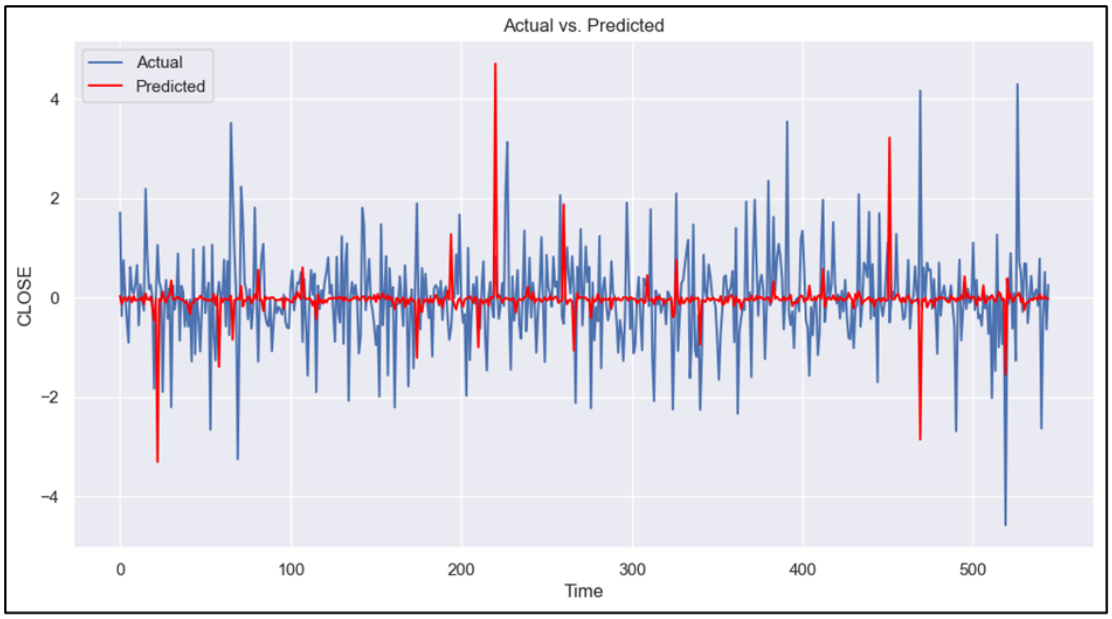


Figure 4.15: *Closing price forecasting using LSTM*

XGBoost

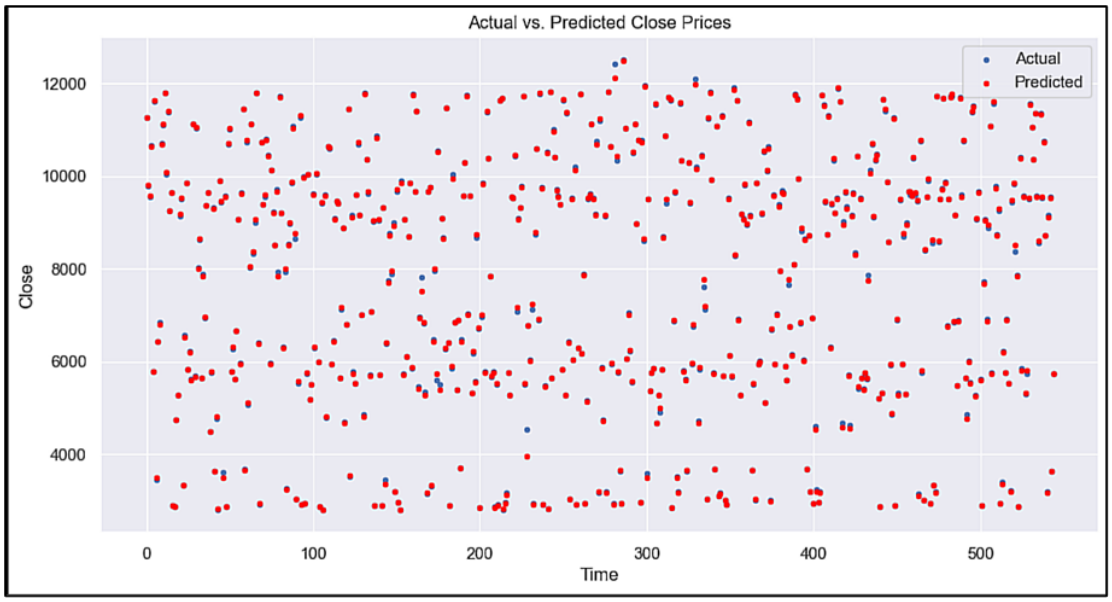


Figure 4.16: *Actual vs Predicted Values using Xgboost*

The XGBoost model scatter plot compares actual and predicted closing prices over time. While the predicted prices generally follow the actual prices, there are notable deviations. Despite its high performance metrics, such as MSE and R2, overfitting is observed, raising concerns about its reliability.

Model Evaluation Metrics

TUNALIM — **EVALUATION METRICS**

Model/Metrics	Prophet	LSTM	Xgboost
Mean Squared Error (MSE)	2288128.17	0.989647	1963.57
Mean Absolute Error (MAE)	204.77	0.66	23.64
Accuracy (within 5.0% deviation)	82.83%	0.18%	99.82%
R-squared (R2)	0.98	-0.166 ⚠	0.999 ⚠

Figure 4.17: *Models Comparison for TUNALIM Index*

A comparison table of evaluation metrics reveals XGBoost's exceptional performance in metrics like MSE, MAE, accuracy, and R2. However, overfitting is evident. While Prophet may not match XGBoost's raw metrics, it demonstrates strong results with

an R2 of 0.98 and reasonable accuracy, making it a prudent choice for forecasting the closing price.

Modeling for Sentiment Analysis in Financial News

In this section, we delve into the modeling approach employed for sentiment analysis in financial news articles, specifically focusing on content related to the Tunisian stock market. The methodology integrates a pre-trained NLP model and Named Entity Recognition (NER) techniques to enhance sentiment analysis accuracy and granularity.

Pre-trained NLP Model: FinBERT

We leveraged the FinBERT model, a state-of-the-art pre-trained NLP model fine-tuned specifically for financial sentiment analysis tasks. FinBERT is trained on financial text corpora and possesses a deep understanding of financial language nuances, making it well-suited for our task of analyzing sentiments in Tunisian stock market news articles.

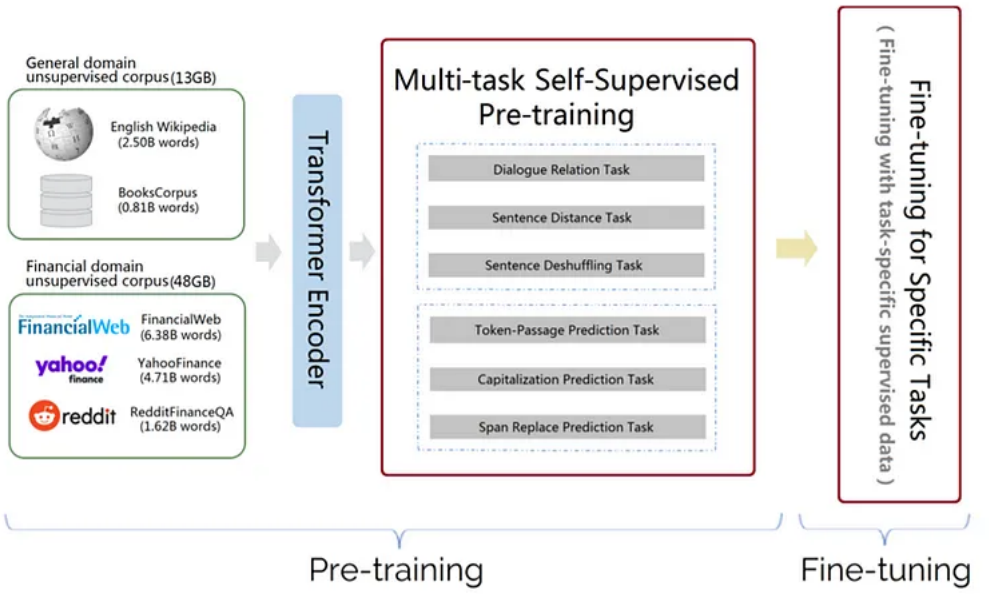


Figure 4.18: *FinBERT Pretraining Architecture [9]*

Named Entity Recognition (NER)

Named Entity Recognition (NER) is employed to identify and extract entities of interest from the input text. We used the spaCy library, which offers robust NER capabilities, to perform entity extraction. Entities detected by the NER system are cross-referenced with a custom entities dictionary tailored for the Tunisian stock market domain.

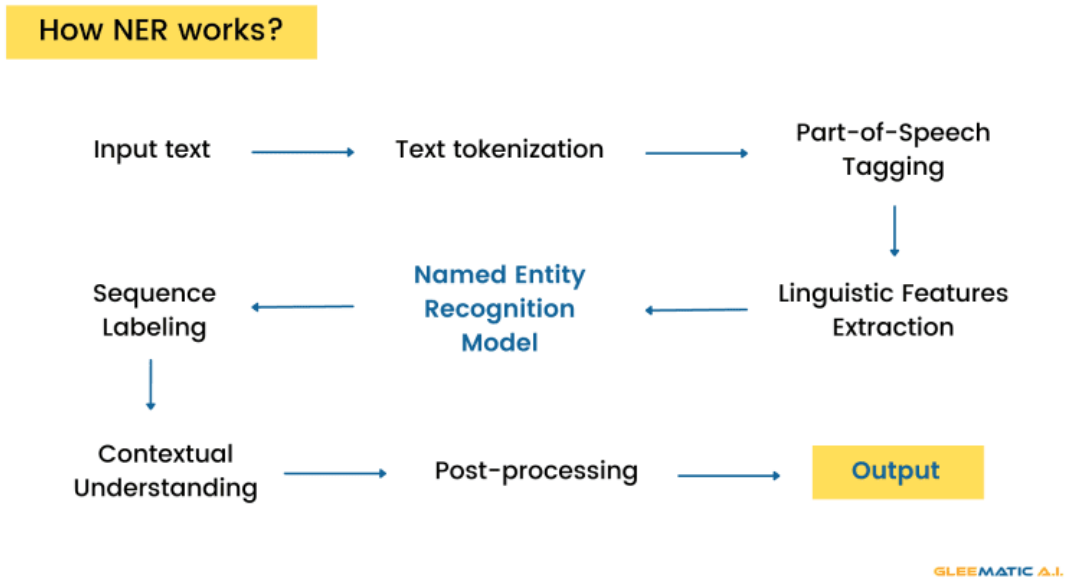


Figure 4.19: *Named Entity Recognition Architecture [10]*

Methodology

1. **Named Entity Recognition (NER)**: Entities relevant to the Tunisian stock market, such as company names, stock symbols, and financial terms, are extracted using spaCy's NER functionality. Additionally, a custom entities dictionary is utilized to enrich the entity extraction process with domain-specific knowledge.
2. **Sentiment Analysis**: For each detected entity, sentiment analysis is conducted. The context surrounding the entity is considered, and sentiment scores are calculated using a combination of the FinBERT model's output and predefined sentiment weights from the custom dictionary.

Example Usage and Evaluation

We can Consider an excerpt from a Tunisian financial news article:

Using the implemented methodology, sentiment analysis is performed on the provided text. Entities are extracted, sentiments are analyzed, and sentiment scores are computed for each entity based on the context and the model's output.

The performance of the model is evaluated based on accuracy in sentiment analysis and the relevance of extracted entities.

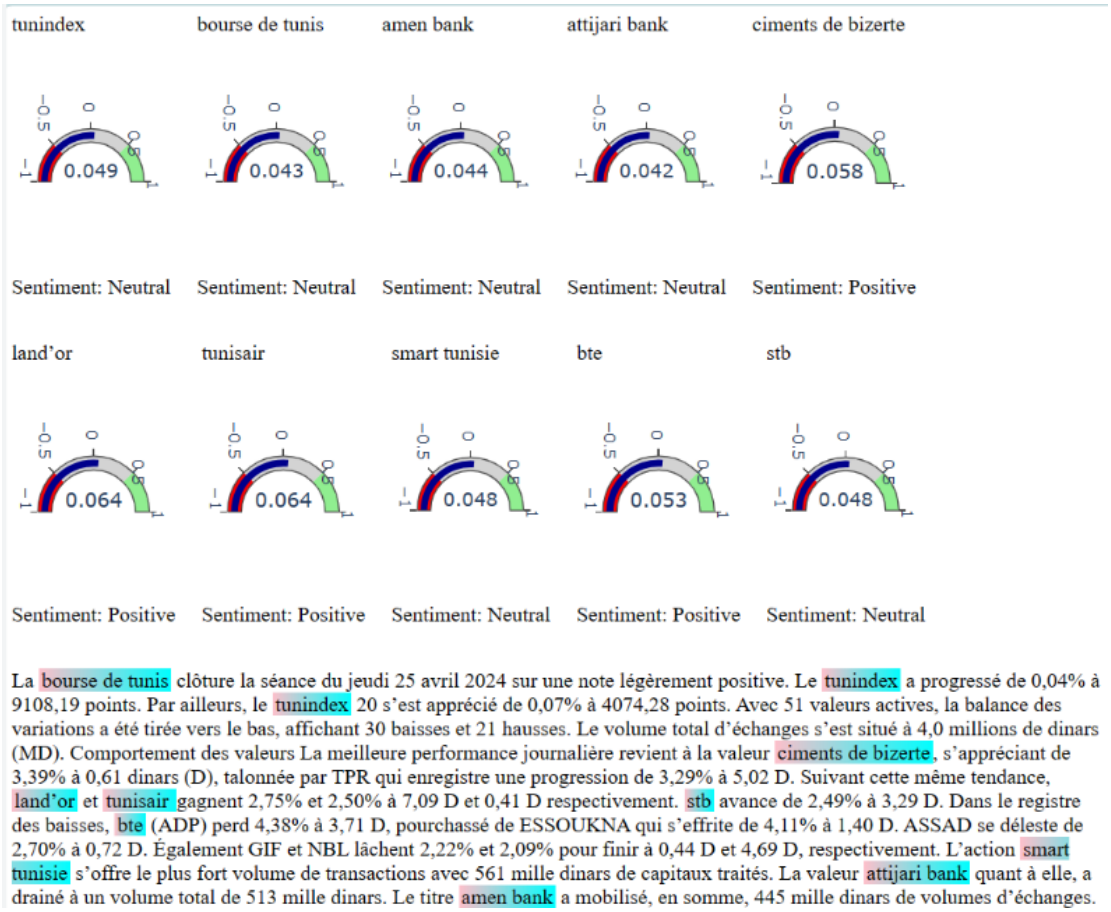


Figure 4.20: *FinBert Sentiment Analysis Usage Example*

The sentiment analysis results revealed interesting patterns. For the articles scraped from BVMT, the majority exhibited a neutral sentiment, with a smaller proportion showing a negative sentiment. This could be indicative of the objective and factual nature of the content on the official stock market website.

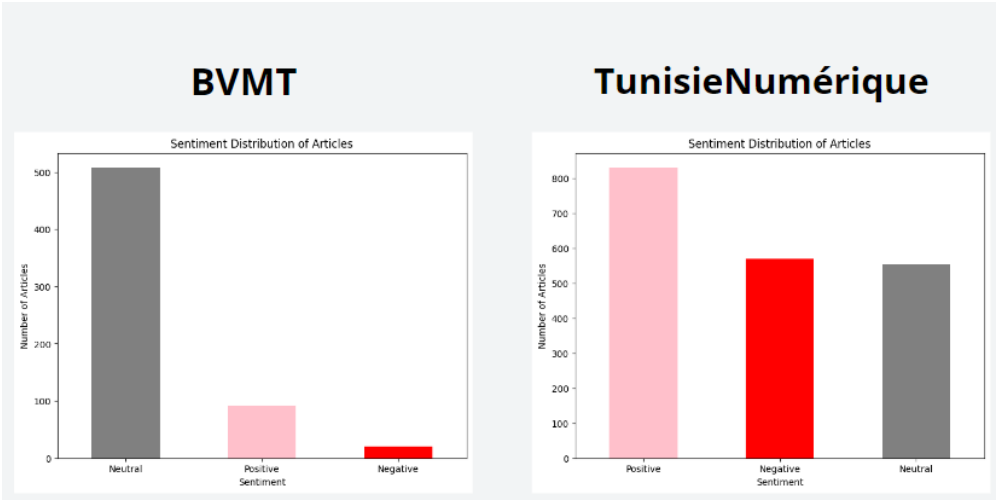


Figure 4.21: *Comparative Sentiment Analysis of different Stock Market News Sources*

On the other hand, the sentiment distribution for the articles from TunisieNumérique was quite different. A significant portion of the articles displayed a negative sentiment, followed by a smaller proportion of neutral articles, and a very minimal representation of positive sentiment. This suggests that the media website might have a tendency to report more on negative events or issues related to the stock market.

These findings highlight the importance of considering the source of information when performing sentiment analysis, as different sources can portray varying sentiment distributions. This could potentially impact decision-making processes based on this analysis.

The modeling approach detailed in this section provides a robust framework for sentiment analysis in Tunisian stock market news articles. By integrating a specialized pre-trained NLP model and leveraging advanced NER techniques, we achieve enhanced accuracy and granularity in sentiment assessment.

Conclusion

This chapter provided a deep dive into risk assessment and prediction in financial trading, covering techniques such as technical indicator-based scoring and machine learning model training. We also explored dimensionality reduction and clustering for market segmentation. These methods offer crucial insights for traders and investors, empowering them to navigate market dynamics effectively and make informed decisions. As technology evolves, these strategies remain fundamental in achieving success in financial trading.

5

Deployment

In this final chapter, we focus on the deployment phase of our project, where we translate the insights gained from our data analysis and modeling efforts into actionable strategies and tools that traders and investors can leverage in the real world. Our deployment strategy is aligned with the overarching business objectives and data science objectives outlined throughout this project.

5.1 Project Scope

The project scope defines the boundaries and objectives of the deployment phase within the broader context of the TDSP methodology.

5.1.1 Actors Identification

The application targets various individuals involved in the trading domain and financial markets. These include:

Financial Consultants

Financial consultants utilize the application to analyze market trends, assess risk, and provide informed recommendations to clients regarding investment strategies and portfolio management.

Traders

Traders rely on the application to access real-time market data, analyze financial indicators, and develop trading strategies to execute buy and sell orders efficiently.

Investors

Investors use the application to evaluate the performance of their investment portfolios, assess risk exposure, and make data-driven decisions to maximize returns.

Market Analysts

Market analysts leverage the application to conduct in-depth market research, identify emerging trends, and generate insights to support investment decisions and market predictions.

Portfolio Managers

Portfolio managers utilize the application to monitor the performance of investment portfolios, rebalance asset allocations, and optimize portfolio strategies based on market conditions and risk profiles.

Risk Managers

Risk managers rely on the application to assess and mitigate risks associated with investment portfolios, monitor compliance with risk management policies, and ensure regulatory compliance.

Business Stakeholders

Business stakeholders, including executives, managers, and decision-makers, utilize the application to gain strategic insights into market trends, monitor key performance indicators, and make informed decisions to drive business growth and profitability.

5.1.2 Functional and non-functional specifications

Functional Needs Identification

Data Collection Module

This module addresses **Business Objective (BO1)** by automatically collecting financial data and relevant news from various sources, thereby improving the quality and reliability of data for more accurate and efficient decision-making. The corresponding **Data Science Objective (DSO)** involves preparing and processing this raw data, ensuring its accuracy and consistency for further analysis.

Financial Indicators Calculation Module

This module fulfills **BO4** by calculating performance and risk indicators for Tunisian market data and listed Tunisian company stocks. The associated **DSO** involves developing predictive models for forecasting price movements and evaluating risk in financial instruments to optimize trading strategies and investment decisions.

Company Clustering Module

This module addresses **BO5** by using clustering techniques to detect internal groupings of Tunisian listed companies, enabling the identification of similarities and patterns within

the market. The corresponding **DSO** involves applying clustering and segmentation techniques to identify distinct clusters in the market, thereby uncovering insights to improve trading outcomes and market understanding.

Trading Strategy Module

This module fulfills **BO2** by allowing users to select trading strategies based on their needs, relying on forecasts and risk assessments provided by other modules. The associated **DSO** involves developing predictive models for forecasting price movements and evaluating risk in financial instruments to optimize trading strategies and investment decisions.

Visualization Module

This module addresses **BO6** by presenting the results of other modules to the user in a synthetic and intuitive manner, facilitating informed decision-making. The corresponding **DSO** involves developing user-friendly interfaces for visualization and exploration of analyses.

Specific Needs for the Final Portfolio

For the final portfolio, specific functional needs include:

- Maximum or predefined return, linked to **BO2**.
- Minimum or predefined volatility, linked to **BO4**.
- Portfolio diversification, linked to **BO5**.
- Outperforming a given benchmark, linked to **BO2**.
- Considering the overall market risk, linked to **BO4**.

Non-functional Needs Identification

The non-functional specifications highlight some characteristics that the system could have such as: flexibility, as it can be modified according to the customer's preferences, simplicity, as its simple design makes it accessible to all, and his strict security system

- (A) Scalability : The architecture of the application allows the evolution and maintenance at the level of its different modules in a flexible way thanks to the implemented distributed massive data architecture.
- (B) Ergonomics and friendliness : The application offers a user-friendly and easy-to-use interface which requires no prerequisites so that it can be used by most Consultants. This is perhaps one of the most important things to have in our project. A qualified Consultant would be able analyze and understand portfolios from simply looking to the graphs and indicators.
- (C) Configurability : The application offers many configuration parameters to adjust the graphs and views according to what the user wants to see, the time interval, the phenomena he wants to follow...

- (D) Security : We can see that the information that the application offers is important and very critical. So, the security is a very important factor. We don't want this information to be accessible to everyone. For this reason, this module has his own sign-in to access this information. Moreover, the dependencies we use in this project are all internal in order to avoid any external hacker attack using an external dependency.

5.2 Development Tools

Various development tools and technologies are employed to streamline the development process and ensure the deployment of a robust and scalable IntelligentTrading agent. This section provides an overview of the key tools utilized in the project.

5.2.1 Streamlit

Figure 5.1: *Streamlit Logo*

Streamlit chosen for its fast and rewarding development of interactive web applications with Python.

5.2.2 ATLAS MongoDB

ATLAS MongoDB serves as the database for storing and managing structured and unstructured data. It provides a flexible and scalable solution for handling large volumes of data in real-time.

Figure 5.2: *ATLAS MongoDB [11]*

5.2.3 Langchain

Langchain is employed for natural language processing tasks, such as sentiment analysis and entity recognition. It offers advanced capabilities for processing and analyzing text data.

Figure 5.3: *Langchain Logo [12]*

5.2.4 Hugging Face Transformers

The **Hugging Face Transformers** library is used for implementing advanced natural language processing models. It provides a wide range of pre-trained models and tools for fine-tuning models on specific tasks.

Figure 5.4: *Hugging Face Logo [13]*

5.2.5 Llama and Ollama

Figure 5.5: *Meta's Llama 2 LLM [14]*

We channeled the capabilities of **Llama** and **Ollama** for conversational AI, text and document analysis in our application, enhancing the application's functionality and modernity. Specifically, the integration of **Llama 2**, a cutting-edge language model,

played a pivotal role in empowering our application with advanced natural language understanding capabilities.

By leveraging Llama 2, we were able to achieve state-of-the-art performance in tasks such as text generation, question answering, and context-based responses.

The sophisticated architecture of Llama 2 enabled our application to comprehend and generate human-like responses, thereby enriching the user experience and ensuring a more engaging interaction. Moreover, the versatility of Llama 2 allowed us to seamlessly integrate it into our application architecture, facilitating efficient deployment and scalability.

5.2.6 Pandas, NumPy, Matplotlib, Seaborn

Pandas, **NumPy**, **Matplotlib**, and **Seaborn** are essential libraries for data manipulation, analysis, and visualization in Python. **Pandas** provides data structures and functions for working with structured data, while **NumPy** offers support for numerical operations and array manipulation. **Matplotlib** and **Seaborn** are powerful visualization libraries used for creating various types of plots and charts to visualize data effectively.

Figure 5.6: *Primary Data Analysis Toolkit [15]*

5.2.7 Plotly

In addition to the core data manipulation and visualization libraries mentionned previously, we have integrated Plotly into our data science ecosystem. Plotly's interactive and versatile plotting functionalities allowed us to create dynamic and engaging visualizations, enabling deeper exploration and understanding of our data inside of our website's dashboards.

5.2.8 Scheduler

A **scheduler** tool is employed to automate data collection, model training, and other recurring tasks. It ensures the efficient execution of scheduled tasks and helps in

Figure 5.7: *Plotly's logo [16]*

managing workflow processes.

5.2.9 VSCode Studio

Visual Studio Code is used as the integrated development environment (IDE) for coding and debugging application components. It provides a lightweight and customizable environment for software development.

Figure 5.8: *VSCode Logo [17]*

5.3 Solution Design and Strategy

In our design strategy, we will present the different architectures and technologies we used, as well as a high-level overview of the project.

5.3.1 Market Sentiment module

The Market Sentiment Analysis Module is deployed within the application's cloud-based infrastructure, ensuring scalability, reliability, and accessibility. The deployment architecture consists of:

- **Cloud Infrastructure**: The module is hosted on a cloud platform, leveraging its computational resources and scalability features to handle varying workloads and user demands effectively.

- **Streamlit Dashboards**: Interactive dashboards developed using Streamlit provide users with intuitive access to sentiment analysis results, visualizations, and article exploration features.
- **Real-time insights** The module periodically scrapes articles from online sources, such as Tunisie Numerique and BVMT (Bourse des Valeurs Mobilières de Tunis), using Selenium WebDriver. Scraped articles are stored in MongoDB collections for further analysis.

Deployment Process

The deployment process involves the following steps:

1. **Configuration**: The Market Sentiment Analysis Module is configured to connect to the MongoDB databases where articles are stored. Streamlit dashboards are set up to visualize sentiment analysis results and facilitate user interaction.
2. **Integration**: The module is seamlessly integrated into the larger trading application, ensuring cohesive functionality and a unified user experience.
3. **Testing**: Rigorous testing is conducted to validate the module's performance, including article scraping, sentiment analysis accuracy, and dashboard responsiveness.
4. **Deployment**: The fully tested module is deployed to the cloud infrastructure, making it accessible to users via web browsers or dedicated client applications.

User Interaction

Users interact with the Market Sentiment Analysis Module through the Streamlit dashboards, where they can find:

Word Cloud Visualization

The Market Sentiment Analysis Module includes a word cloud visualization that provides a graphical representation of the most frequently occurring words in the latest articles,from which you could additionally filter through the displayed words.

- The process involved:

Text Preprocessing:

Text underwent preprocessing, including tokenization, removal of common stopwords, and lemmatization to ensure meaningful representation.

TF-IDF Calculation:

- TF-IDF scores were computed to measure the importance of words in the corpus of articles.
- Selection of Top Words: Words with the highest TF-IDF scores were selected, prioritizing relevance and distinctiveness.
- Insights: The word cloud highlights key themes and topics in the articles, aiding in intuitive understanding.

Figure 5.9: *Word Cloud Visualization and Filtering*

- Larger font sizes indicate higher TF-IDF scores, emphasizing significant terms.
- Users can interact with the cloud to explore specific topics and associated sentiments.

Real-time Article Display with Sentiments

Users can view real-time articles scraped from online sources along with their associated sentiments. The sentiment analysis is performed on-the-fly, providing insights into the prevailing market sentiment for each article.

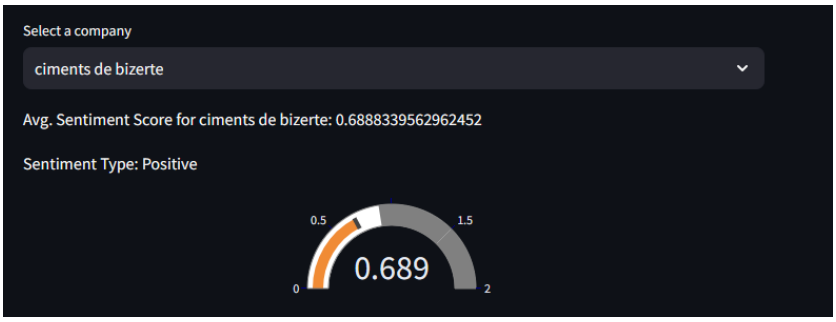


Figure 5.10: *Recent Sentiment on selected Company - Ciments de Bizerte*

Top and Bottom Performing Entities

The module identifies and highlights the top-performing and bottom-performing financial entities from the most recent articles based on sentiment analysis scores. This feature allows users to quickly assess which entities are influencing market sentiment positively and negatively.

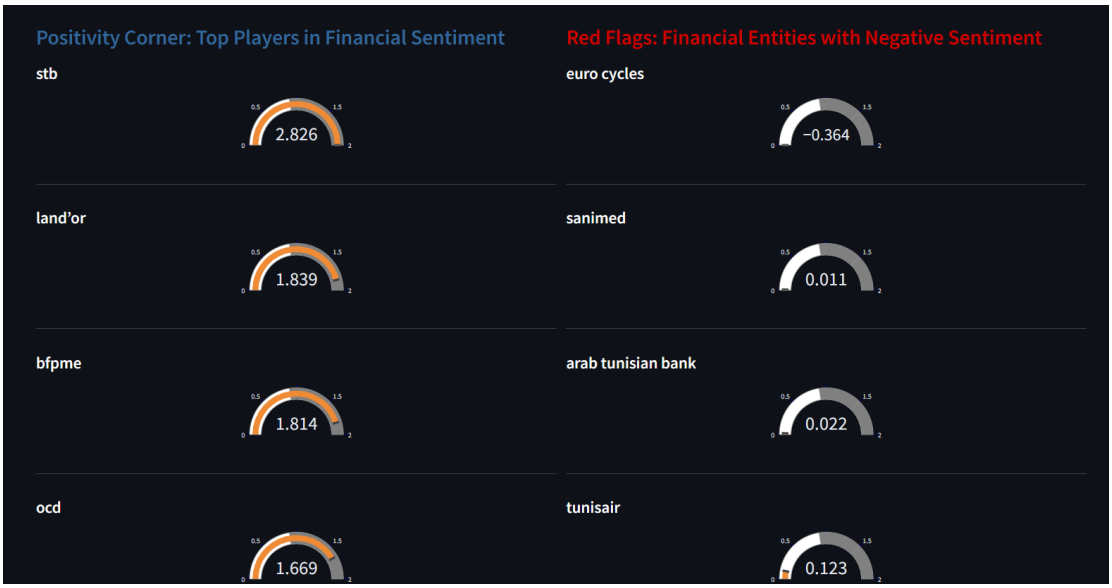


Figure 5.11: *Sentiment Scores For Best and Worst Companies from Recent News*

5.3.2 PDF GPT Integration for Financial Statements and Bulletins

The `PdfGpt` module utilizes advanced natural language processing techniques to analyze the textual content of uploaded Financial Staement and Bulletins. It employs Llama 2 Large language model (LLM) , along with a retrieval-based question-answering (QA)

model, to extract relevant information from the PDF files and generate accurate responses to user queries.

Functionality

The main functionalities of the `PdfGpt` class include:

- Uploading a PDF file.
- Processing the uploaded PDF file to extract text.
- Splitting the text into smaller chunks for analysis.
- Creating embeddings for the text chunks using Hugging Face embeddings.
- Generating responses to user questions based on the content of the PDF using the retrieval QA model.

Implementation in Streamlit App

The `PdfGpt` class is seamlessly integrated into our Streamlit application, providing a user-friendly interface for interacting with PDF files and obtaining insightful responses. The key components of the implementation include:

- File upload section: Users can upload a PDF file through the Streamlit interface.
- Text input section: Users can input questions related to the content of the PDF file.
- Response generation: The application displays the generated responses to user questions in real-time.

Example Interaction

Figure 5.12 illustrates an example interaction with our Streamlit application using the `PdfGpt` class. The user uploads a PDF file, asks a question, and receives a response based on the content of the PDF.

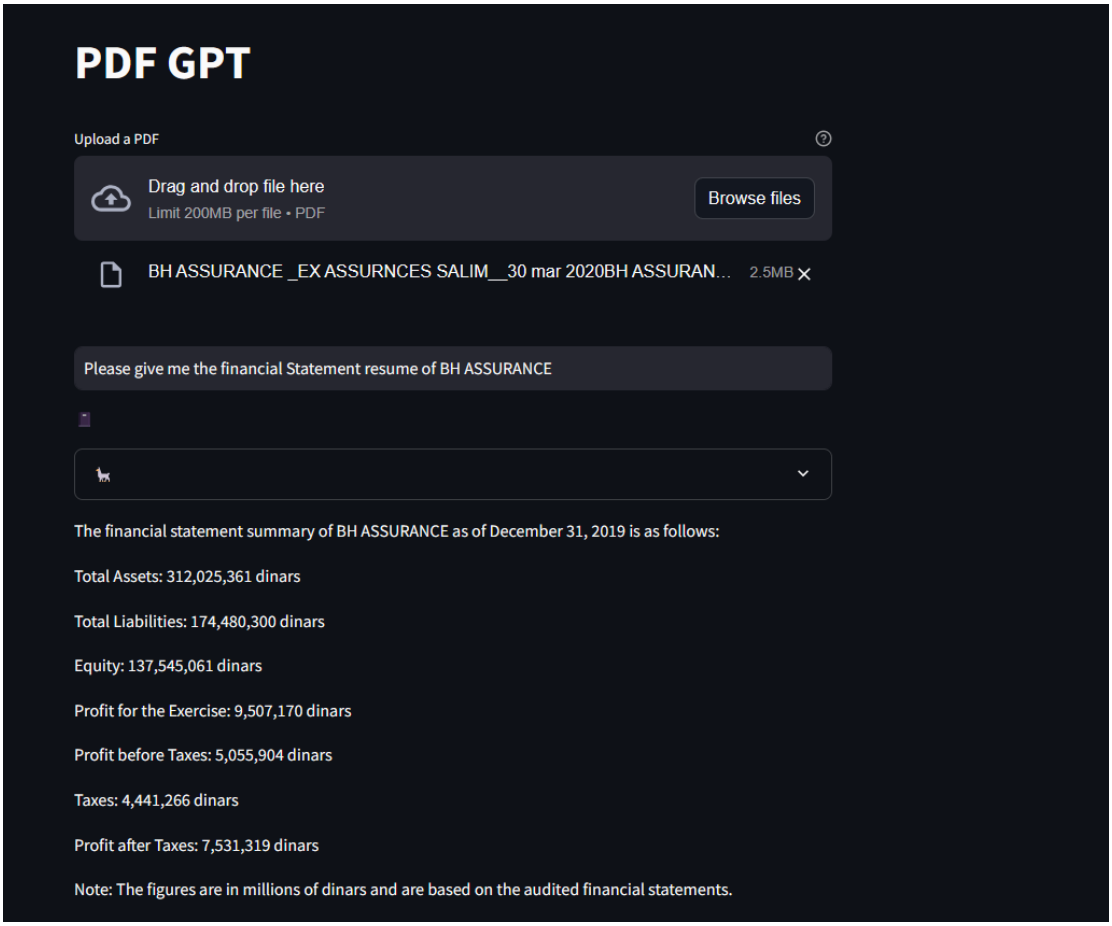


Figure 5.12: *interaction with PDF GPT over Financial Statement Document*

5.3.3 Integration of Prediction and Forecasting Models

Prediction and forecasting models developed during the modeling phase are integrated into Streamlit dashboards. This integration enables users, such as traders and investors, to access predictive insights directly within the application. Key aspects of this integration include:

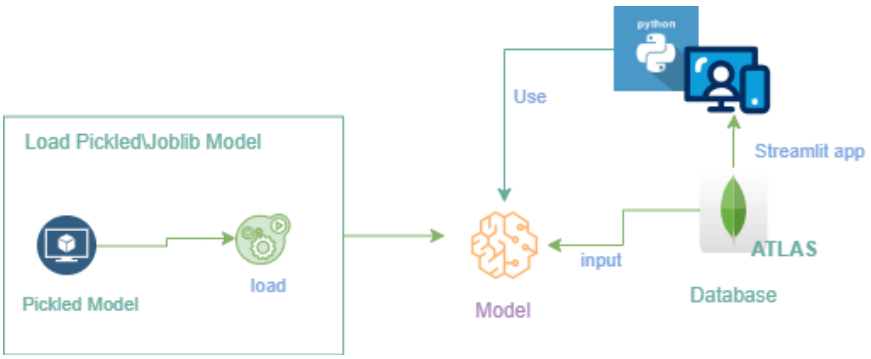


Figure 5.13: *Predictive Models Integration*

Real-time Data Integration

Prediction models are fed with real-time market data retrieved from ATLAS MongoDB, allowing for up-to-date analysis and forecasting.

Figure 5.14: *Real-Time Company Stock Prices Trends*

Model Output Visualization

The output of prediction and forecasting models, including dividends predictions and stock price predictions, is visualized within the Streamlit dashboards using interactive charts, graphs, and tables.

Dividends Predictions Interface

The Dividends Predictions Interface empowers users to forecast dividends for chosen companies, providing valuable insights into future income streams. Users can select a company of interest and visualize both historical dividend payments and predicted values over time.

Figure 5.15: *Companies Dividends Predictions Interface*

Through ***interactive plotting capabilities***, users can explore the impact of various factors on dividend payouts and make informed investment decisions based on projected earnings.

Stock Price Prediction Interface

The Stock Price Prediction Interface enables users to simulate stock price predictions based on specific financial measures such as Relative Strength Index (RSI), volume change, and volatility. Users can input custom parameters and visualize the predicted price movements over a chosen time horizon.

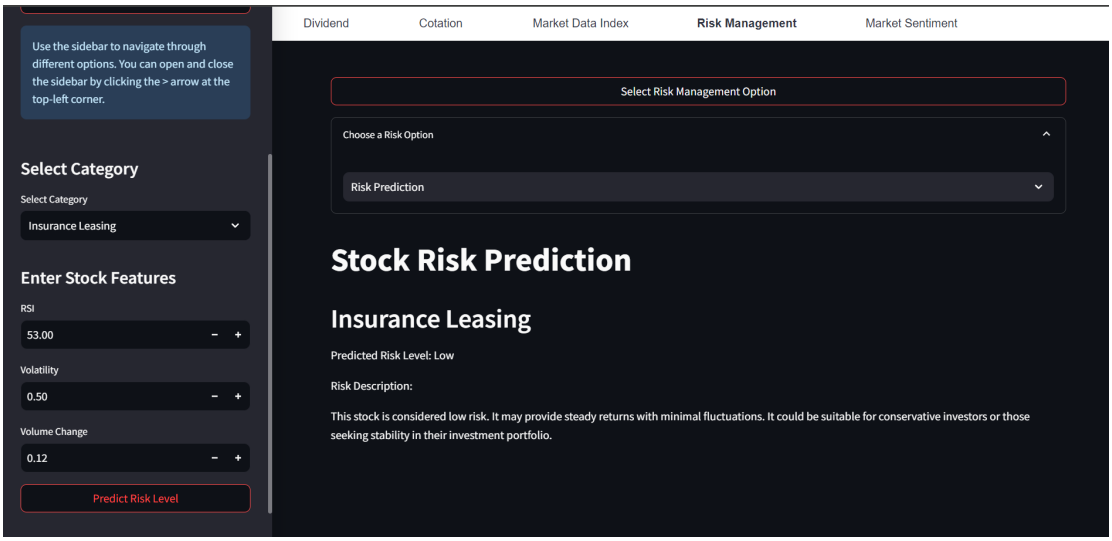


Figure 5.16: *Stock Price Prediction Interface with Risk Parameters*

By adjusting parameters and exploring different scenarios, users can gain deeper insights into potential market trends and outcomes, aiding in strategic decision-making and risk management.

5.4 Risk Assessment Models Deployment

In addition to the modules outlined earlier, our deployment strategy incorporates tailored models specifically designed for risk assessment in the Tunisian stock market across four key categories of data: "Insurance Leasing," "Bank," "Other," and "Bank Leasing Sicav." Leveraging the Streamlit framework, we've developed user-friendly web applications that allow stakeholders to interact seamlessly with the risk assessment models and gain actionable insights.

5.4.1 Model 1: Price Prediction per Day

We've implemented machine learning algorithms trained on historical market data to forecast daily stock price movements, with a particular focus on assessing the associated risk. By incorporating features such as Daily Price Change, Moving Averages, and Relative Strength Index (RSI), this model provides accurate predictions on whether stock prices are likely to rise or fall. This aids investors in evaluating potential risks associated with their investment decisions.

5.4.2 Model 2: Risk Assessment

Our risk assessment tool enables users to evaluate the risk level associated with individual stocks across various categories. By analyzing features such as Relative Strength Index (RSI), Volatility, and Volume Change, the model assesses the risk profile of stocks, providing insights into potential investment risks. It caters to investors seeking to manage risk exposure and make informed decisions based on their risk tolerance and investment goals.

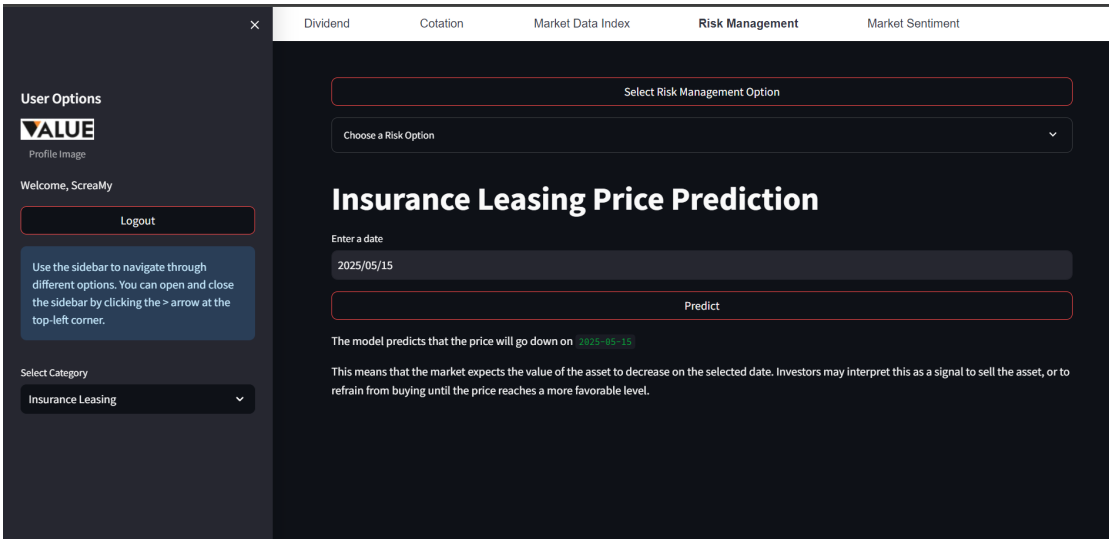


Figure 5.17: *Price Prediction with Risk assessment Interface*

5.4.3 Model 3: Market Segmentation

In the market segmentation explorer, clustering algorithms and dimensionality reduction techniques are employed to identify distinct market segments based on risk profiles. By clustering stocks based on risk, the tool helps users uncover underlying market trends and patterns related to risk. This visualization tool is instrumental for investors looking to manage risk exposure, identify diversification opportunities, and optimize their investment strategies accordingly.

5.4.4 Use Cases

- **Investment Decision Support:** Investors can utilize the risk assessment tools to make informed decisions on buying, selling, or holding stocks based on their risk tolerance.
- **Risk Management:** Financial analysts and risk managers can employ these tools to assess and manage the risk exposure of their investment portfolios.
- **Strategic Planning:** Risk analysts can use the risk assessment models to identify emerging risks, trends, and opportunities in the stock market, enabling them to develop strategic risk mitigation strategies.

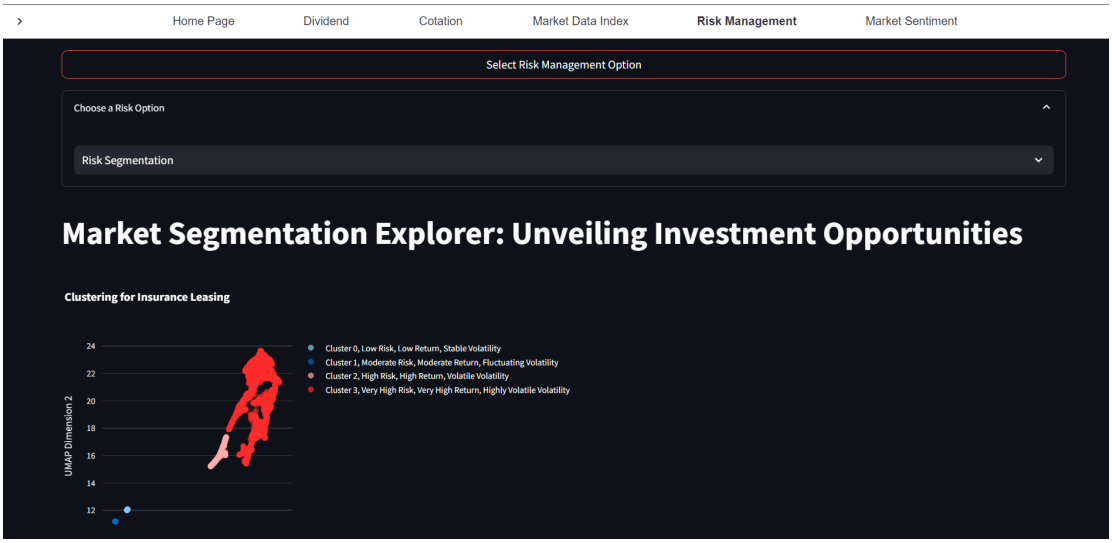


Figure 5.18: *Identifying Market Segments using Risk indicators*

We've also used other measures to mirror the segmentation techniques used by the Official Tunisian Stock Market Website such as Cotation as shown in the figure below, aiming to classify the companies into 3 Compartiments (A, B and S)

Figure 5.19: *Classification of Companies using Cotation Values and Capital*

Overall, the deployment of these risk assessment models represents a significant advancement in leveraging data science and machine learning techniques to empower stakeholders with actionable insights for risk management and decision-making in the dynamic landscape of financial markets.

5.5 Tunisian Market Indexes Forecasting Models Deployment

Our application includes intuitive interfaces that provide users with valuable insights into Tunisian index prices and associated risk indicators. These interfaces leverage advanced forecasting techniques and data visualization to enhance decision-making and risk management capabilities.

5.5.1 Index Price Forecasting Interface

The Index Price Forecasting Interface offers users a comprehensive view of predicted trends in Tunisian index prices. By analyzing historical data and incorporating machine learning algorithms, the interface generates forecasts for future price movements, enabling users to anticipate market trends and make informed investment decisions.

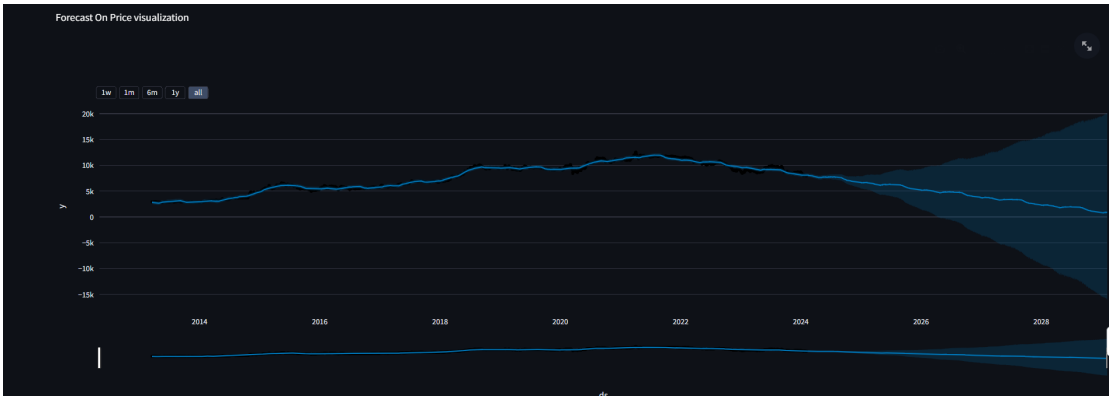


Figure 5.20: *Index Price Forecasting Interface displaying predicted trends in Tunisian index prices*

5.5.2 Risk Indicators for Tunsian Stock Market Indexes

Through interactive visualization tools and customizable filters, users can delve deeper into specific risk indicators such as volatility, beta, and correlation coefficients for each index. This granular level of analysis empowers users to identify trends, assess risk exposure, and formulate informed investment strategies tailored to their risk tolerance and objectives.

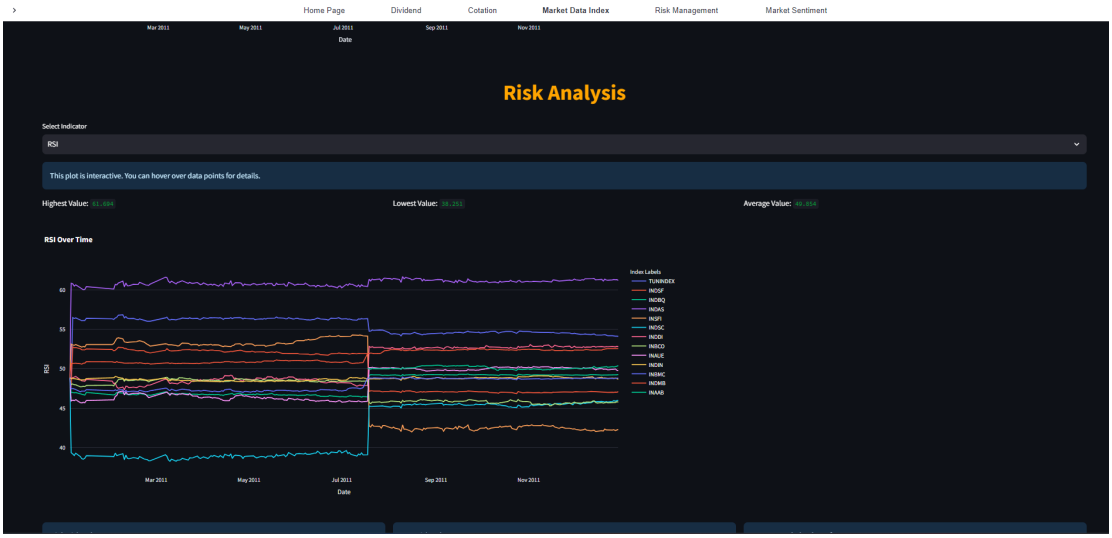
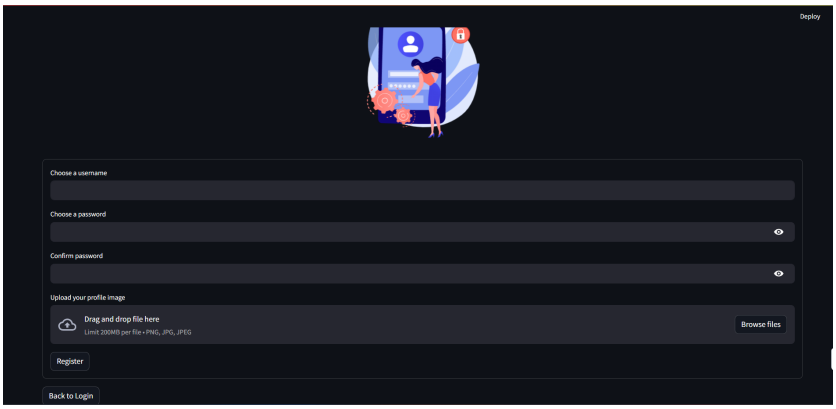


Figure 5.21: *Risk Indicators Interface showcasing risk metrics across multiple indexes*

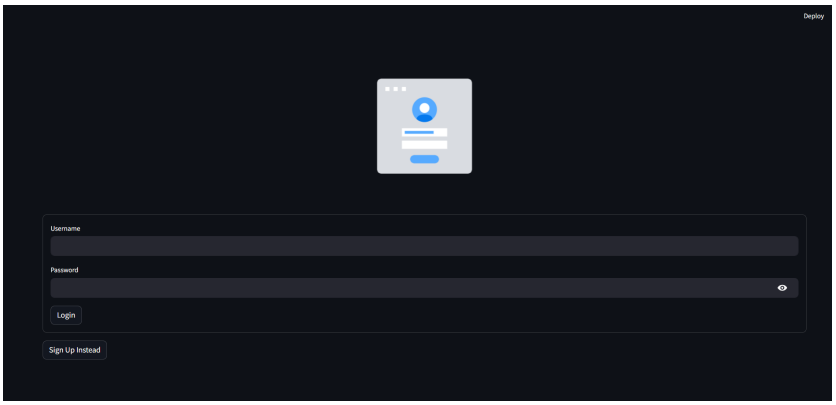
By centralizing risk information for all indexes within a single interface, our application facilitates comprehensive risk management and strategic decision-making in the Tunisian financial market. Users can leverage this interface to stay abreast of market dynamics, mitigate potential risks, and capitalize on emerging opportunities with confidence.

5.6 Authentication Interfaces

As part of ensuring secure access to our application, we've implemented authentication interfaces using Streamlit in Python. These interfaces include functionalities for user registration and login, each equipped with necessary fields such as username, password, and attributed role (admin, superuser).



(a) *Registration interface*



(b) *Login Interface*

Figure 5.22: *Authentication Interfaces*

Conclusion

In this chapter, we've detailed the deployment of our intelligent trading agent project in the Tunisian market. Through the implementation of various modules and risk assessment models, we've provided stakeholders with tools for informed decision-making and risk management. Leveraging Streamlit interfaces, we've ensured user-friendly access to these tools. Moreover, the integration of risk assessment models tailored specifically for the Tunisian stock market represents a significant advancement in leveraging data science and machine learning techniques to empower stakeholders with actionable insights for risk management and decision-making.

6

Conclusion and Perspectives

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

6.1 Conclusion

In this project, we have successfully developed and deployed an intelligent trading agent tailored for the Tunisian market. Through comprehensive data analysis, modeling, and deployment phases.

The integration of advanced forecasting models, risk assessment techniques, and user-friendly interfaces has significantly enhanced the usability and effectiveness of our application.

6.2 Perspectives

Looking ahead, there are several avenues for further improvement and expansion of our intelligent trading agent:

- **Enhanced Predictive Models:** Continual refinement and optimization of our predictive models can improve accuracy and reliability, providing users with more actionable insights.
- **Integration of External Data Sources:** Incorporating additional data sources such as social media sentiment, macroeconomic indicators, and geopolitical events can enrich our analysis and enhance the robustness of our trading strategies.

- **Machine Learning Interpretability:** Developing interpretable machine learning models can enhance transparency and trustworthiness, enabling users to better understand the underlying factors driving predictions and recommendations.
- **Expanded Market Coverage:** Expanding our coverage to include other markets beyond the Tunisian market can broaden our user base and provide opportunities for global investment strategies.
- **User Feedback and Iterative Development:** Continuously soliciting feedback from users and stakeholders and iteratively refining our application based on their input is essential for ensuring relevance and effectiveness in a dynamic market environment.

By embracing these perspectives and responsive to market dynamics and user needs, we can further solidify our position as a leading provider of intelligent trading solutions in the Tunisian financial market and beyond.

References

1. Value Official website: https://www.value.com.tn/
2. IBM Documentation: https://www.ibm.com/docs/fr/spss-modeler/18.5.0?topic=dm-crisp-help-overview
3. Microsoft TDSP Documentation: https://learn.microsoft.com/en-us/azure/architecture/data-science-process/overview
4. Sopcat SMART metrics guide: https://www.sopact.com/guides/smart-metrics#:~:text=SMART%20metrics%20refer%20to%20a,objectives%20are%20concrete%20and%20attainable.
5. United Nations Department of Economic and Social Affairs: https://sdgs.un.org/goals/goal8
6. Selenium official website: https://www.selenium.dev/
7. Spacy official website: https://spacy.io/
8. Financial Market Council CMF Website: https://www.cmf.tn/
9. FinBERT Architecture: https://www.ijcai.org/proceedings/2020/0622.pdf
10. NER IBM documentation: https://gleematic.com/what-is-named-entity-recognition-ner/
11. StreamLit official website: https://streamlit.io/
12. MongoDB Atlas official website: https://www.mongodb.com/cloud/atlas/
13. Langchain official website: https://www.langchain.com/
14. HuggingFace official website: https://huggingface.co/
15. Meta: https://ai.meta.com/blog/large-language-model-llama-meta-ai/
16. Data Analysis Toolkit: https://advancedanalyticssolutions.co.uk/data-analysis-toolkit/
17. Plotly website: https://plotly.com/
18. Visual Studio website: https://code.visualstudio.com/

This page intentionally left blank.