DEPI R3 Group 79

AI & Data Science Track

رواد مصر الرقمية

# Stock Analysis

**Supervised by :**

Eng.Ahmed el-Safty

**Presented By:**

| | |
|---|---|
| Abdelrahman Sameeh | Malak Khaled |
| Habiba Mohamed | Mohamed Adham |
| Menna Fakharany | Sherouq Eldanaf |

# Team Roles

| | |
|---|---|
| 1.Data collection | Mohamed Adham |
| 2. Preprocessing | Abdelrahman Sameeh |
| 3. Data visualization | Shrouq eldanaf |
| 4.Feature engineering | Mohamed Adham |
| 5.Machine learning model training | Habiba , Abdelrahman |
| 6.Evaluation & Tuning | Malak Khaled |
| 7.Deployment & Testing | Menna Allah Mohamed |
| 8.Documentation | Habiba Mohamed |

# Project Objectives

Our project focuses on developing a Stock Analysis System that helps users **understand stock behavior ,market trends**, **evaluate stock performance**, and **make informed decisions** using data analytics and visualization techniques.

# Project Explanation and Key Features

**Project Overview:**

- Analyzes 5 years of S&P 500 stock data (2013–2018)
- Predicts next-day stock movement (up or down)
- Provides technical indicators and trend insight

# Data Preparation & Cleaning

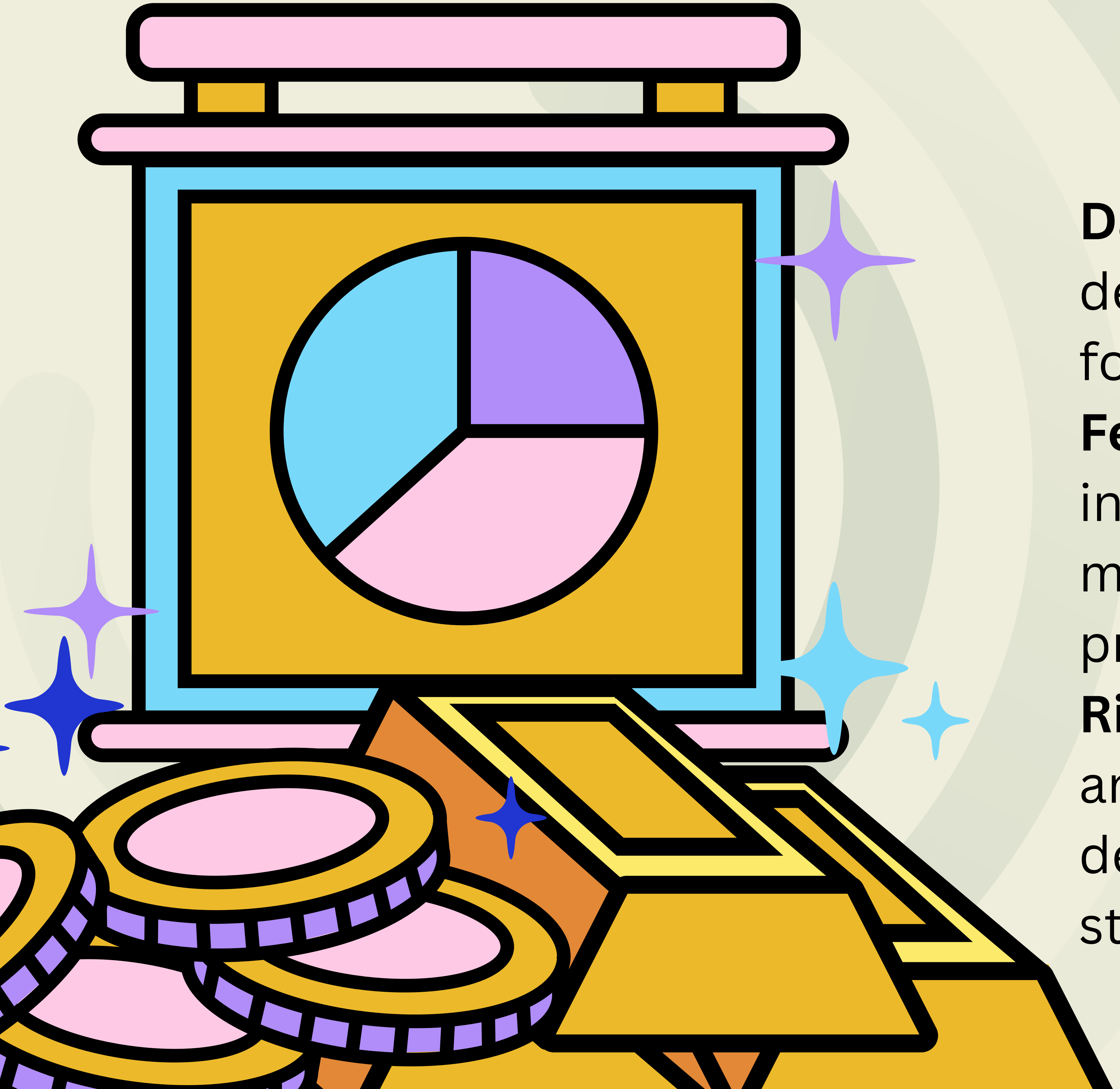
- **The First step is searching appropriate data about stocks :**

  The used dataset from Kaggle is : **S&P 500 stock data (2013–2018**)

- **The next step is checking data & cleaning it by:**

  - checking missing values
  - checking & modifying data types
  - checking duplicates
  - detecting outliers and remove them (but this decreased the efficiency of the model therefore, we reversed this step.)

# Benefits of EDA in Stock Analysis

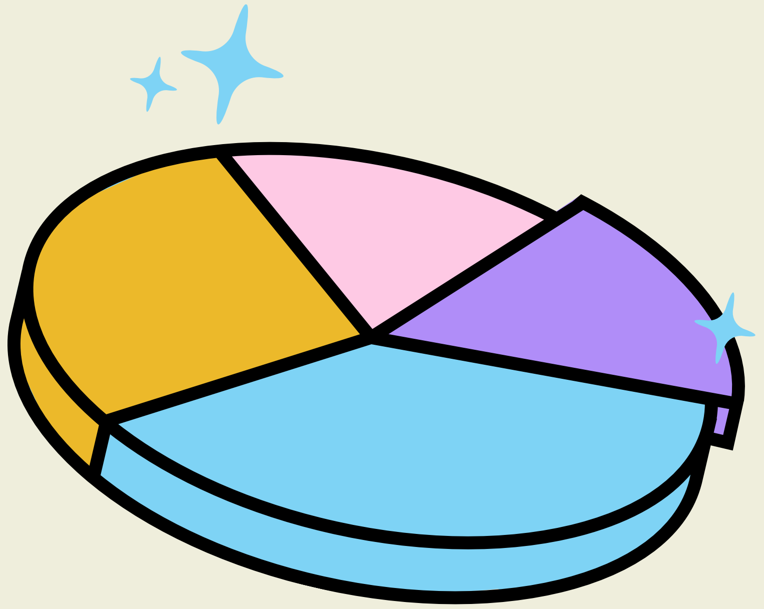


**Data Integrity:** Ensures accurate, clean data by detecting and correcting missing values, errors, and format issues early in the pipeline.

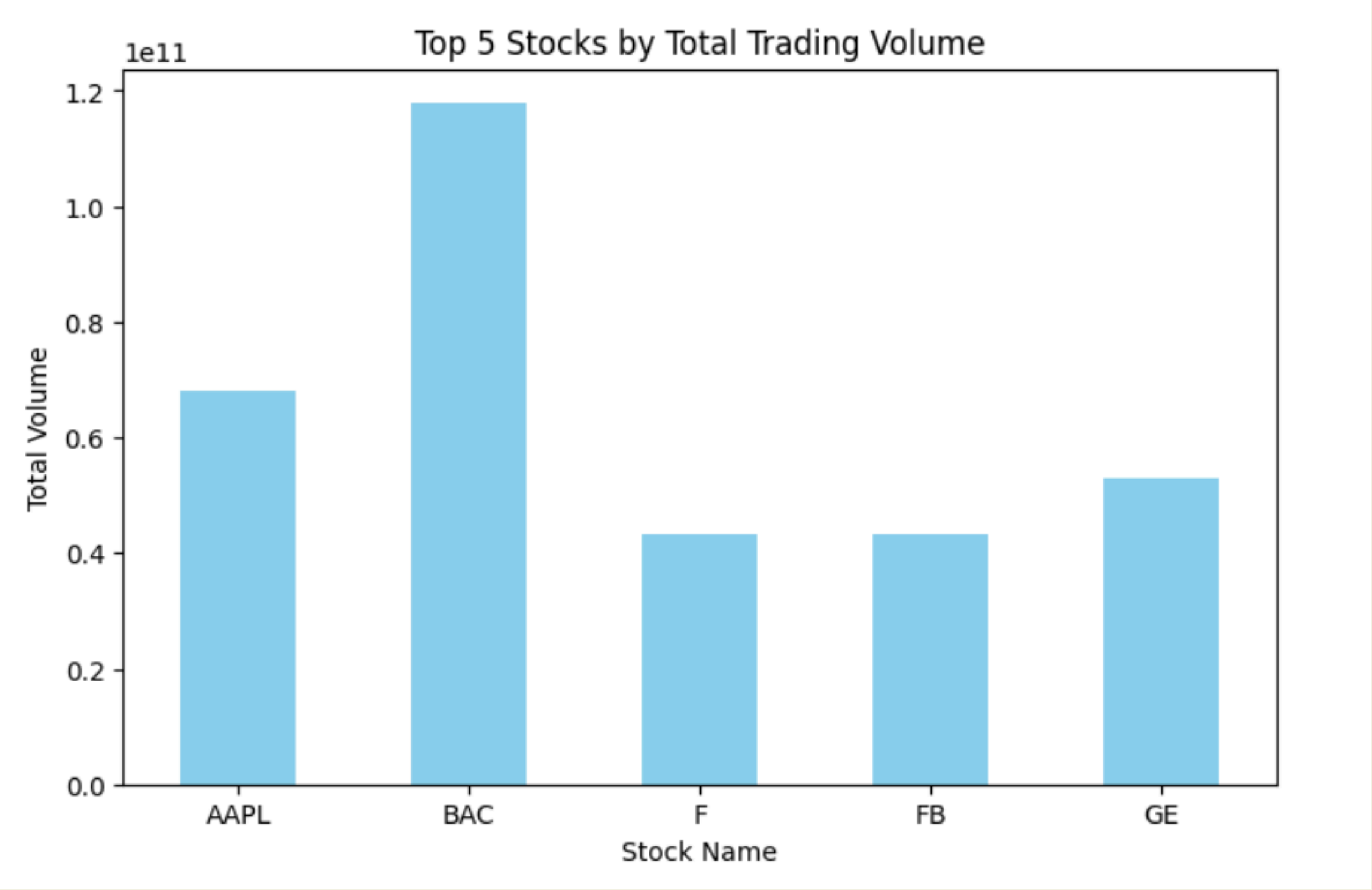
**Feature Engineering Insight:** Supports the creation of informative features like daily price ranges, volatility measures, and momentum indicators to improve predictive model performance.

**Risk Assessment:** Identifies extreme price movements and periods of high volatility, aiding in the development of robust trading and risk management strategies.

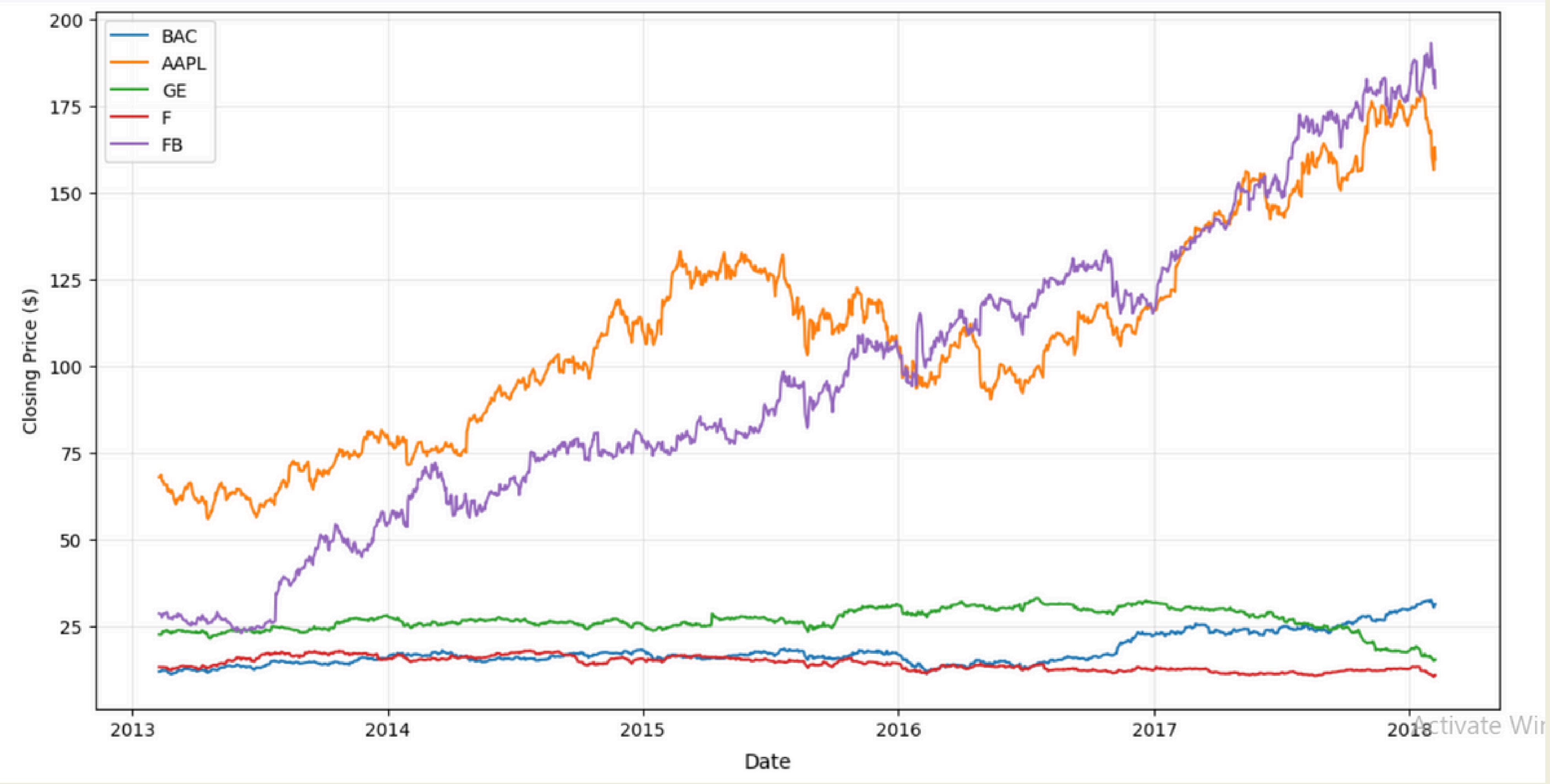
# Visualization:



Quickly highlights trends, volatility, and potential anomalies in the stock data before modeling.



Closing Prices Over Time — Top 5 Companies



# Feature Engineering

**It was a helpful step to add some columns to increase accuaracy when dealing with stocks as :**

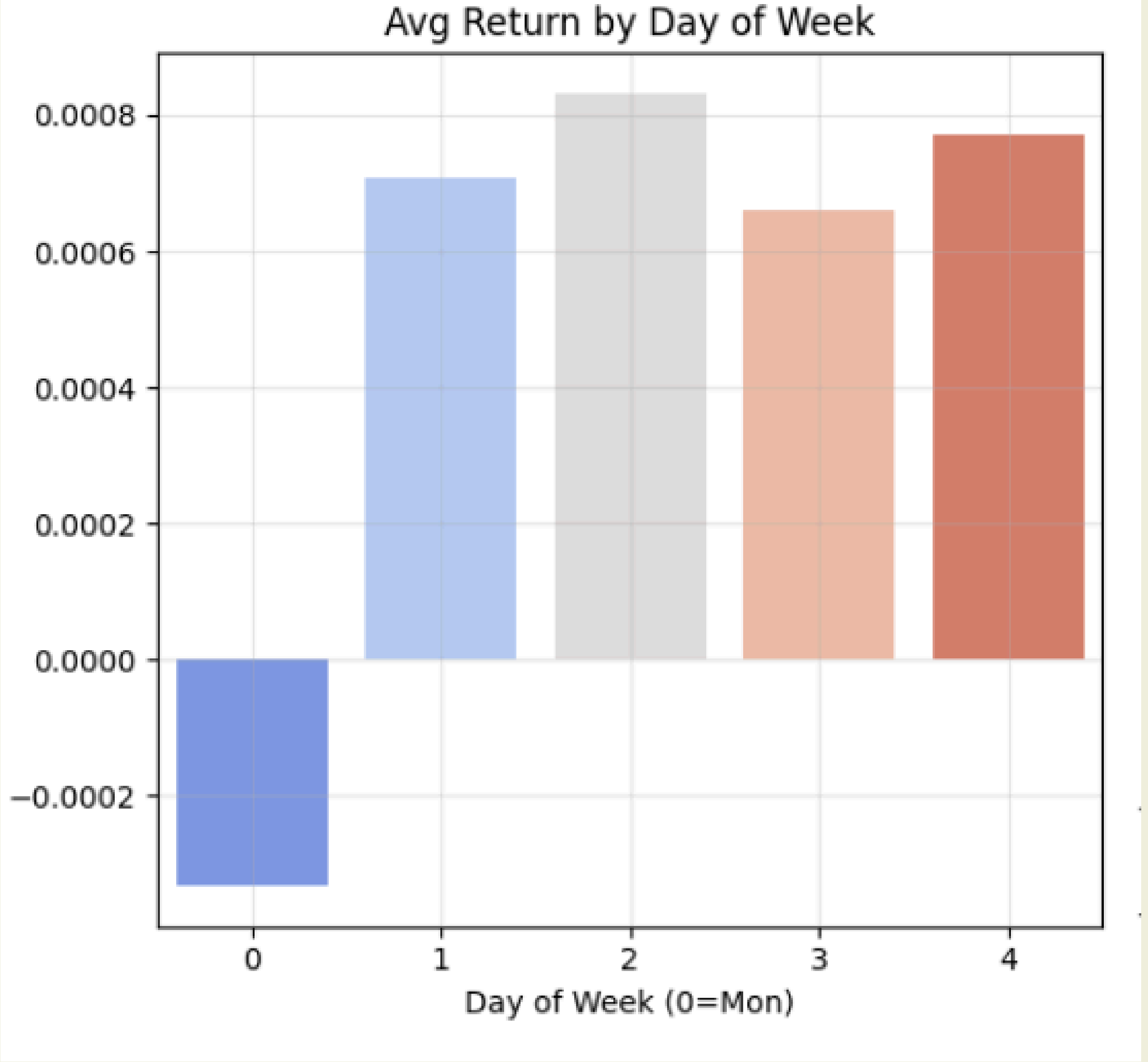
- **SMA column :** Average Price over last N days , it helps us to see general changes (trends) of the stock.
- **RSI column :** A number between 0-100 that indicates how fast and how much a stock has been going up vs down over a period.
- **Volatility columns :** Measures how much a stock's price fluctuates over a recent period (20 or 50 days).

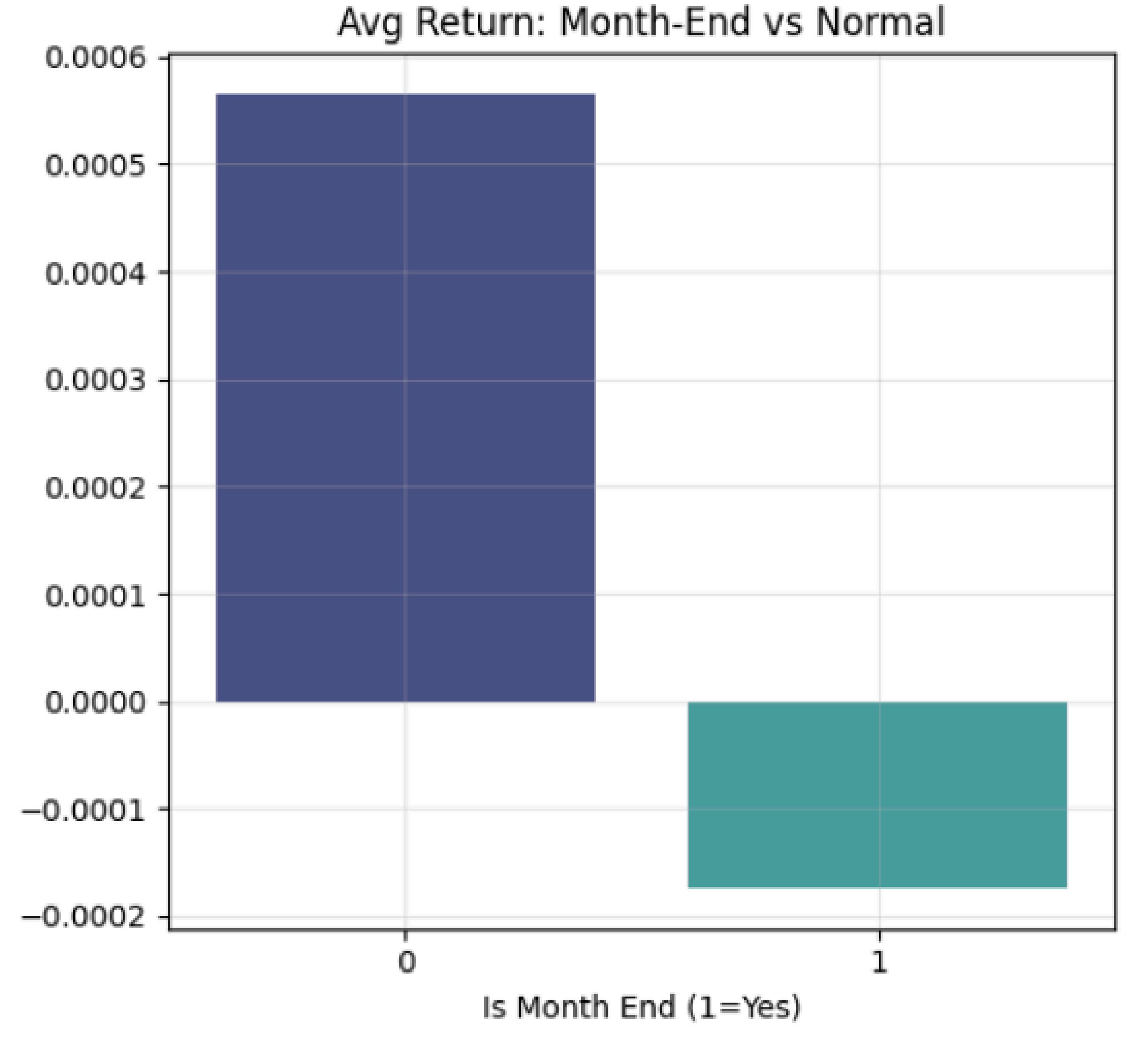
A higher value means the stock is more “unstable” or risky, while a lower value means it moves more steadily.

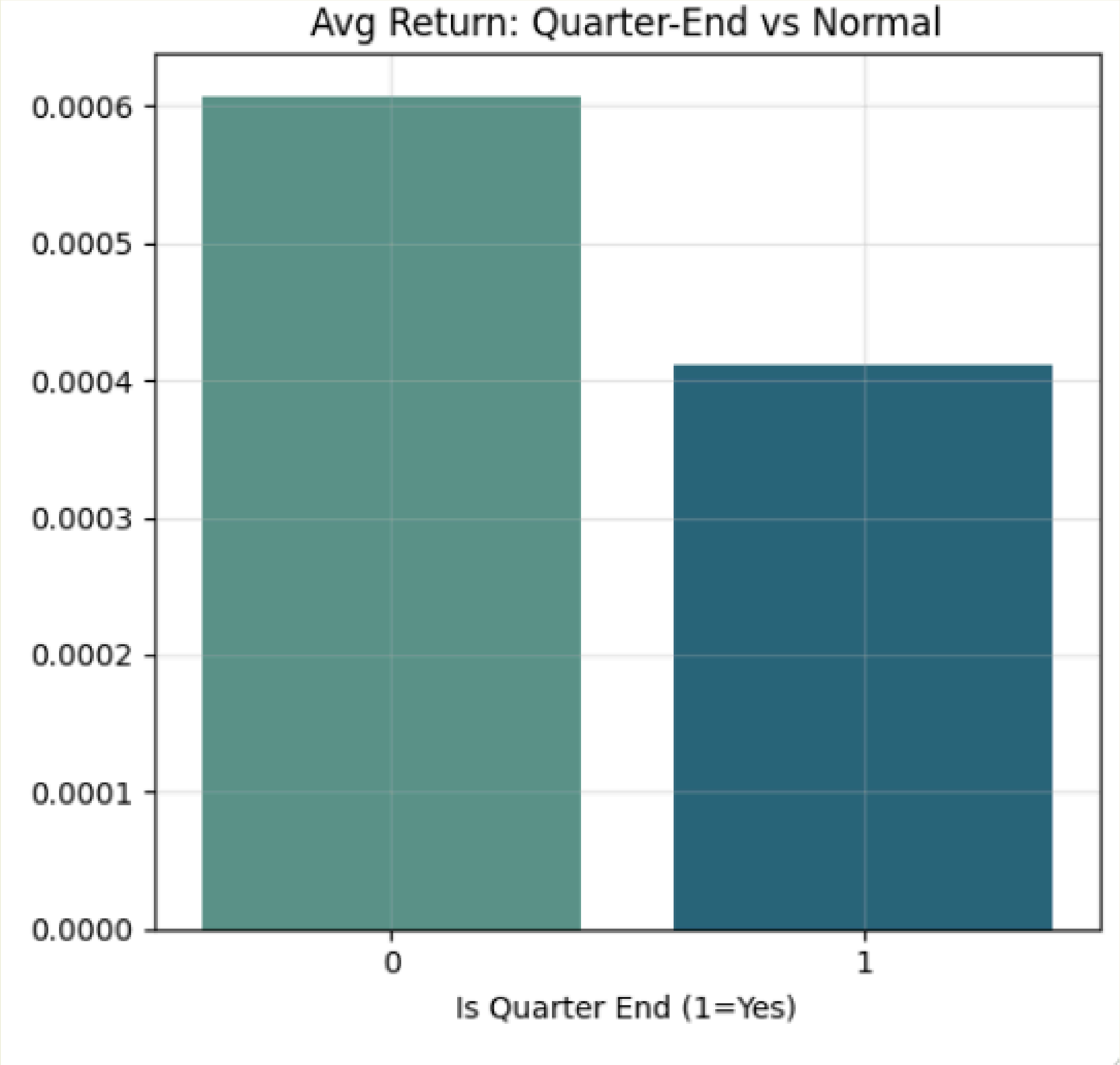
## Time Features as :

- **day_of_week :** Stock prices often follow weekly patterns, so this helps the model capture day-specific trends.
- **is_month_end :** Market behavior can change at the end of the month due to regular reporting and trading activity, making this a useful indicator.
- **is_quarter_end :** A quarter is defined as a group of three months.

Every company prepares its quarterly results and publishes them publicly so, that people can analyze the company's performance. These quarterly results affect the stock prices heavily which is why we have added this feature because this can be a helpful feature for the learning model.


Avg Return by Day of Week
0.0008
0.0006
0.0004
0.0002
0.0000
−0.0002
0
1
2
3
4
Day of Week (0=Mon)


Avg Return: Month-End vs Normal
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
−0.0001
−0.0002
0
1
Is Month End (1=Yes)


Avg Return: Quarter-End vs Normal
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0
1
Is Quarter End (1=Yes)

# ML Modeling Overview

**Objective:**

- Build a classification model to predict **next-day stock direction (Up vs. Down).**

Models we used and why we used them.

- **LightGBM :** Fastest + strongest accuracy on large datasets.
  Handles nonlinear relationships and class imbalance very well.
- **XGBoost :** has proven performance on financial and stock-related prediction tasks.
- **Random Forest :** Simple to tune and provides a strong baseline.

**Expected Value:**

- Produce a predictive signal that performs better than random guessing.
- Enhance trading decisions by integrating ML-driven insights.

# Models Comparison

- **Data Split:** Used TimeSeriesSplit (no shuffling) to strictly avoid lookahead bias.
- **Purpose:** Identify the best-performing algorithm for financial time-series.

| Model | Accuracy | Weighted F1 |
|---|---|---|
| LightGBM | 0.5351 | 0.5354 |
| XGBoost | 0.5349 | 0.5349 |
| Random Forest | 0.5279 | 0.5298 |

**LightGBM** was selected as the **best model** because it had the highest F1-score, which is most important for balanced performance in class prediction.

**Accuracy** differences are small, but **F1** gives a clearer measure for the imbalanced financial dataset.

This table highlights **why we chose LightGBM for hyperparameter tuning.**

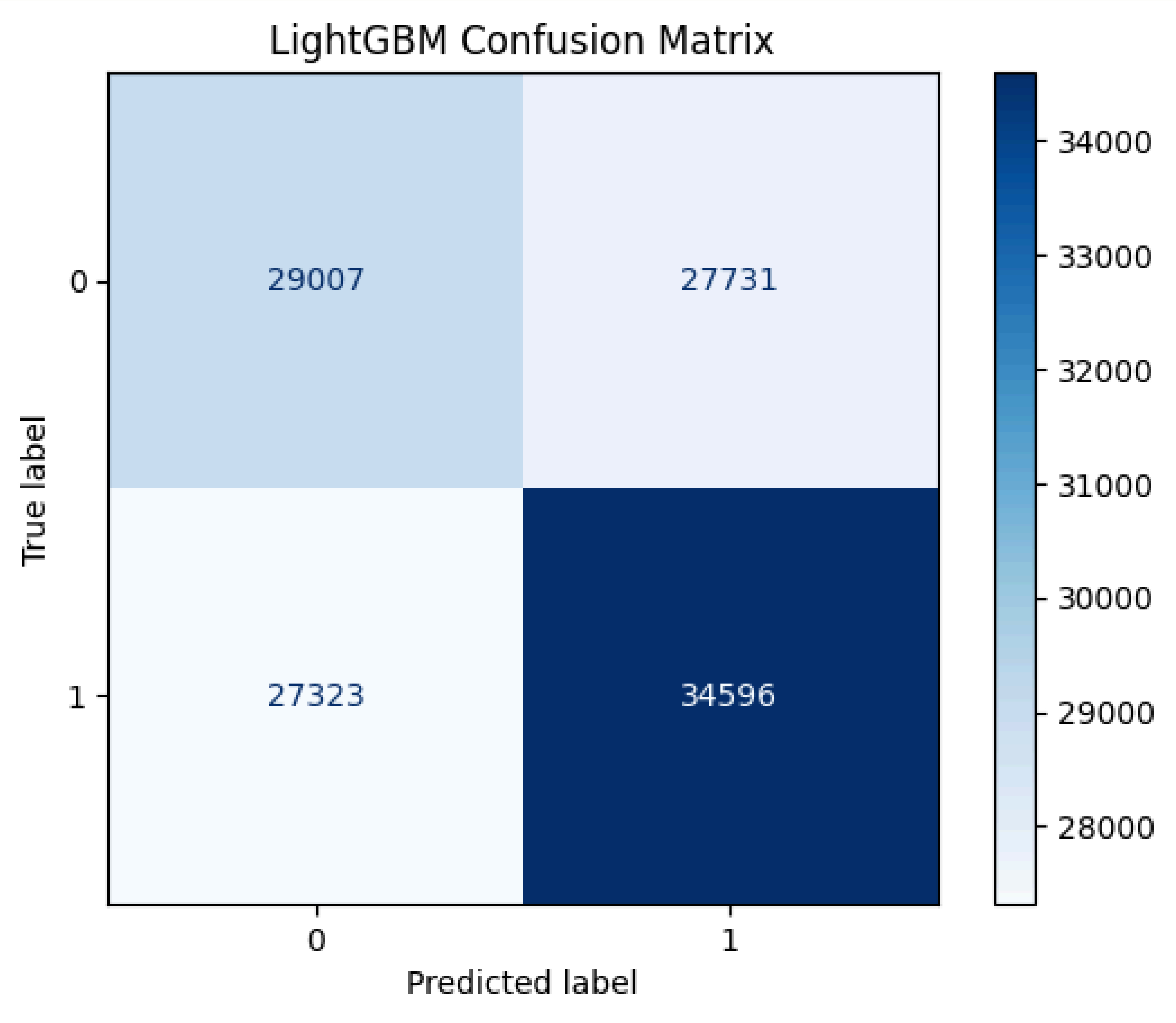
# LightGBM: Final Results&Insights

| Metric | Result | Insight |
|---|---|---|
| Accuracy | 0.536 | Overall correct predictions |
| F1-Score | 0.5398 | Best balanced score after optimization |
| Recall | 0.56 | Detects Up moves for long positions |

**Hyperparameter Tuning**

- **Method**: GridSearchCV with TimeSeriesSplit to respect time order.
- **Final Hyperparameters:**
  - n_estimators = 200
  - max_depth = 7
  - learning_rate = 0.05
  - class_weight = "balanced" (important for handling imbalanced returns)

# LightGBM: Final Results&Insights



**Insight:**

- The hyperparameter-tuned LightGBM model shows a balanced performance between predicting downward (0) and upward (1) stock movements.

# Strategic Recommendations

## Core Finding

- LightGBM delivers a consistent predictive edge (accuracy ~53.6%).
- Although modest, this is meaningful in a market where baseline accuracy ≈50%.

## Understanding the Challenge

- Stock direction prediction = **high noise, low signal.**
- Accuracy >60% is unrealistic in daily short-term forecasting.
- The achieved performance is strong for a directional signal model.

# Deployment

## Live Deployment

The app is hosted on **Streamlit Cloud**.

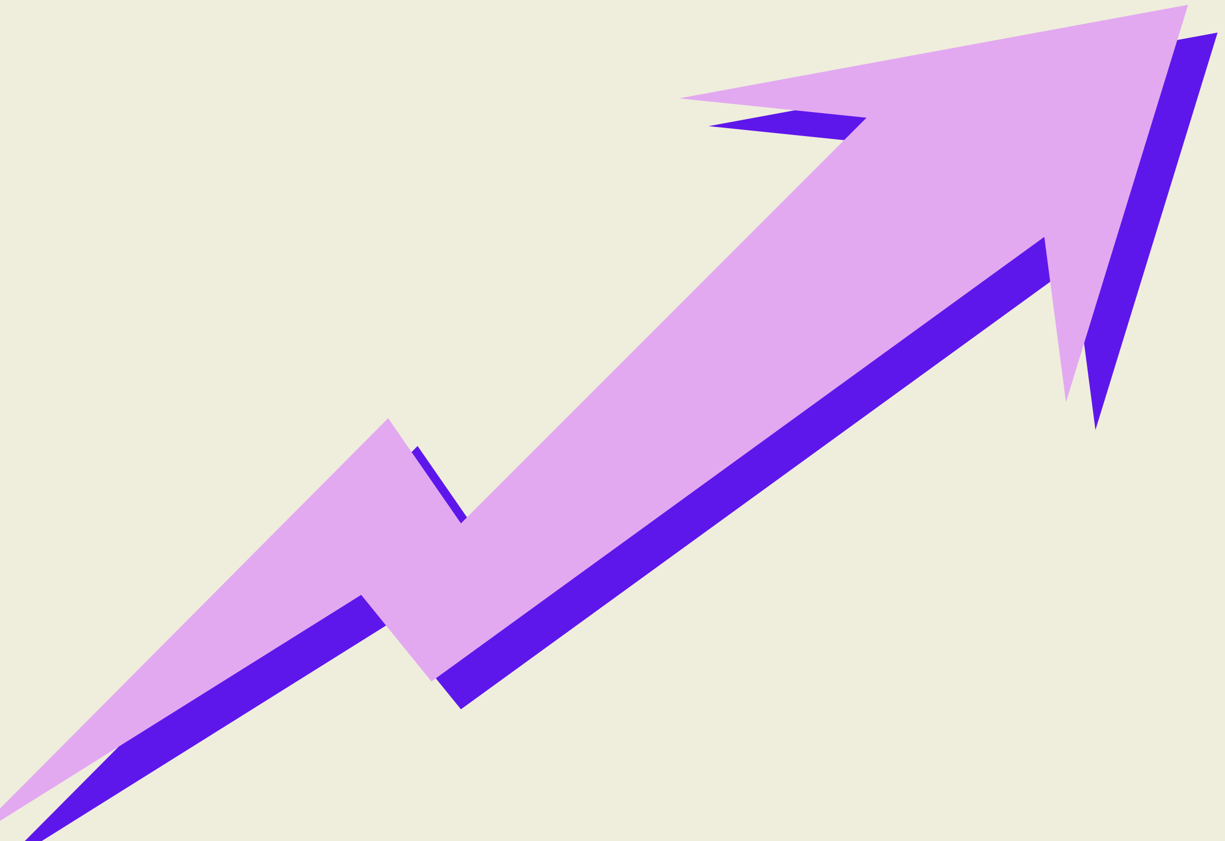
## Key Features

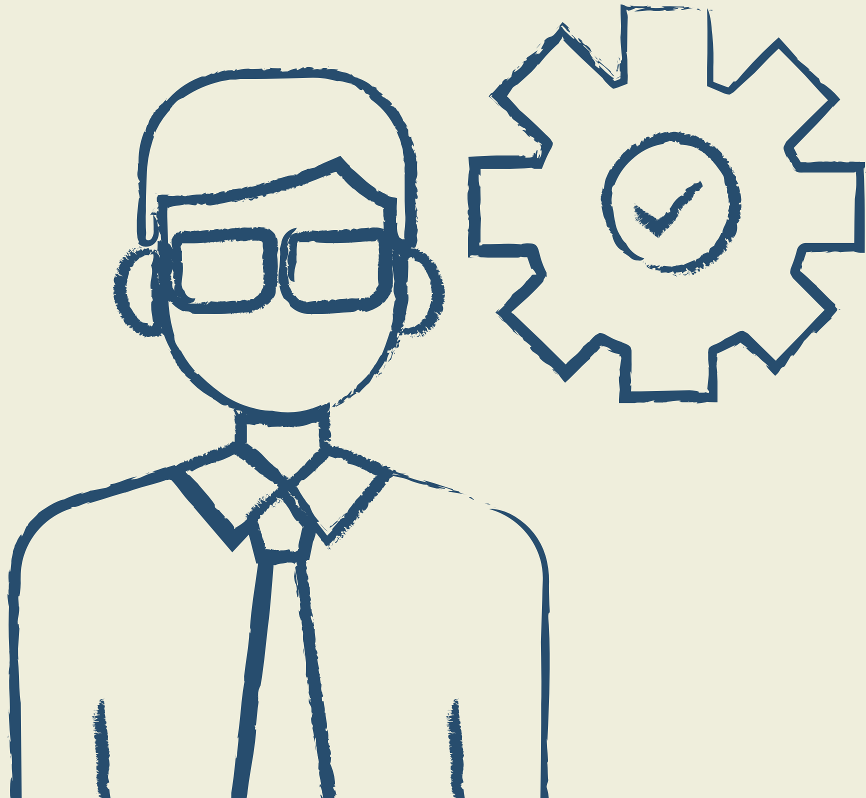
- **Interactive EDA:** Filterable stock list, time-series charts, histograms, and correlation heatmaps.
- **Single-Stock Predictions:** Use your trained model to predict the next movement (Up / Down).
- **Real-Time Stock Insights:** Fetch live prices via Yahoo Finance and get instant predictions.
- **Auto Live Predictions:** Monitor top-performing stocks with automatic predictions.

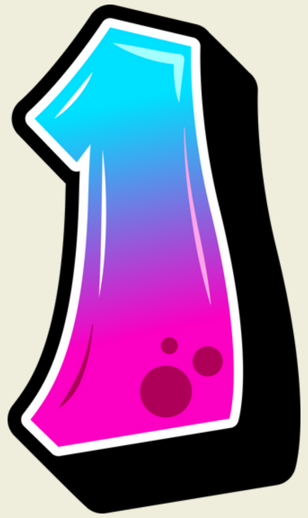
# Why Use This App?

- Make more informed investment decisions.
- Quickly visualize trends and anomalies in stock data.
- Run both historical and live predictions without coding.
- Ideal for both beginners and professional traders looking for quick insights.

# Use Cases For Each Prediction







## 1.Single-Stock Predictions

**Use Case:** A trader wants to check if a specific stock (e.g., AAPL) will go up or down tomorrow.

**How:** User selects a ticker → model predicts Up / Down → user can decide to buy, hold, or sell.

**Benefit:** Quick decision support for individual stocks without scanning the entire market.

# 2.Real-Time Stock Insights

**Use Case:** An investor wants live prices and trends for a chosen stock.

**How:** App fetches current stock price from **Yahoo Finance** → shows live chart → provides prediction based on model.

**Benefit:** Combines market data + ML insight for actionable decisions in real-time.

# 3.Auto Live Predictions (Top Stocks)

**Use Case:** Portfolio managers or traders want to monitor multiple top-performing stocks automatically.
**How:** Predefined stock list is monitored → app fetches live prices → model predicts next-day **movement** → top gainers are highlighted.
**Benefit:** Saves time, highlights potential opportunities, and supports proactive trading across multiple stocks.

# Challenges Faced

1. **Time-series data handling** – avoiding data leakage between past and future.
2. **Financial data volatility** – stock prices are noisy and non-linear, making predictions difficult.
3. **Model selection** – choosing a model that generalizes well without overfitting.
4. **Huge dataset size** – GitHub couldn’t host the large dataset, so Streamlit Cloud couldn’t load it.
5. **Platform limitations** – free hosting with Streamlit & Ngrok has bandwidth and runtime limits.

# Solutions

1. **Chronological train/test split** – ensured test data is completely unseen during training.
2. **Ensemble tree-based models** – LightGBM, XGBoost, and Random Forest handled noise and captured non-linear patterns.
3. **Hyperparameter tuning with GridSearchCV** – optimized the best model (LightGBM) using time-aware validation folds.
4. **Google Drive** - Store the dataset on Google Drive and load it directly from there.
5. **Streamlit Cloud** - Deploy the project on Streamlit Cloud for stable, continuous hosting.

# Outcomes

**Outcomes**

1. **High-performing LightGBM model** – achieved strong F1-score and accuracy on unseen test data.
2. **Robust predictive system** – capable of estimating next-day stock movement direction (Up/Down).
3. **Ready for future integration** – can be expanded into a dashboard, trading alert system, or AI-based financial service.
4. **User-friendly dashboard** – interactive Streamlit interface for visualizing trends, indicators, and predictions.
5. **Real-time data integration** – successfully connected with live stock data through yfinance.

# Future Improvements

- Deploying on AWS EC2, Azure App Service, or Google Cloud Run.
- Combine historical stock data with global events and news using NLP and advanced models to improve the accuracy of future market movement predictions.
- Automated Alerts: Notify users of predicted trends or unusual market movements.
- Portfolio Optimization: Suggest buy/sell strategies based on model predictions.
- Interactive Visualizations: Real-time charts, heatmaps, and comparative dashboards.

# Thank You