NOVA IMS Information Management School

# MDSAA

Master Degree Program in

**Data Science and Advanced Analytics**

## Big Data Analytics

### Forecasting S&P 500 Price Direction Using Apache Spark


Diogo, Carvalho, number: 20240694

Rúben, Marques, number: 20240352

Tomás, Carvalho, number: 20240938

Group 54

**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**

Universidade Nova de Lisboa


June, 2025

# INDEX

# 1. INTRODUCTION & BACKGROUND

## 1.1. Problem Statement

Financial markets are complex, dynamic systems influenced by a variety of factors including investor behavior, macroeconomic indicators, and global events. Accurately predicting short-term price movements in these markets remains a significant challenge for both researchers and practitioners. This project focuses on forecasting the daily direction of the S&P 500 index by leveraging historical price data and macroeconomic indicators, specifically U.S. inflation rates. Using Big Data technologies, we aim to build a scalable machine learning pipeline capable of handling and modeling large volumes of financial time series data.

## 1.2. Research Motivation

The motivation behind this study is twofold. First, financial forecasting is a critical application of data science that can have significant economic impact—from guiding investment decisions to informing policy analysis. Second, the project serves as a practical implementation of Big Data Analytics tools and methodologies. By combining lagged market features with inflation data, we explore whether short-term price direction can be effectively predicted using scalable algorithms in Apache Spark. This aligns with our academic goal of demonstrating the power and limitations of distributed data processing in a real-world context.

# 2. DATA COLLECTION & PREPROCESSING

## 2.1. Data Sources

This project combines historical commodity market data with U.S. inflation rates. The market data was sourced from Investing.com, covering daily values like price, open, high, low, and forward prices. To provide economic context, monthly inflation data was taken from US Inflation Calculator, which reports figures based on the Consumer Price Index from the Bureau of Labor Statistics.

## 2.2. Data Characteristics

The dataset contains several hundred daily entries and covers approximately 25 years. Each observation corresponds to a specific trading date and contains numerical variables such as price movements, market indicators, and forward-looking estimates. The inflation data, provided monthly,

was integrated by aligning it to daily entries using forward-filling, ensuring every record had an associated inflation rate.

### 2.3. Data Cleaning and Preprocessing Steps

The cleaning and preprocessing phase involved some steps. First, the data was checked for missing values. The Volume column was dropped due to having 100% missing values. There were also some Inflation missing values, which were forward-filled, making sure the dates stayed in order.

To enrich the dataset for predictive modeling, some lagged features were created. These included one-day lagged values of the price, open, high, low, percentage change, forward open, forward price, and inflation. This allowed the model to incorporate temporal dependencies and trends from prior trading days. Additionally, a binary label was created to capture the direction of the daily price change, where an increase was labeled as 1, and a decrease was labeled as 0.

Finally, the dataset was cleaned by removing the row with missing values introduced by the lag operations. The relevant features were then assembled into a single vector column suitable for use in PySpark's ML pipeline. The dataset was split into training and testing in an 80/20 ratio.

## 3. METHODOLOGY & TOOLS

### 3.1. Machine Learning Techniques

We used supervised classification models Logistic Regression and Random Forest with PySpark MLlib to predict whether the daily price would increase or decrease, based on lagged market and economic features. The target variable was binary, indicating a positive or negative daily price change. Random Forest was further optimized with cross-validation and hyperparameter tuning. Model performance was assessed using ROC AUC, which measures a model's ability to distinguish between the two classes: higher values indicate better classification performance.

| Models | ROC_AUC |
|---|---|
| Logistic Regression | 0.5140 |
| Random Forest Untuned | 0.5206 |
| Random Forest Tuned | 0,5206 |

Table 1 – Model Results

### 3.2. Querying and Storage

All data manipulation and preprocessing steps were conducted using PySpark, allowing scalable transformations and handling of large datasets. The data was stored and processed in Spark DataFrames, and queries were written using PySpark's SQL-style API. This enabled efficient handling of missing values, lag feature creation, and aggregation operations throughout the pipeline.

### 3.3. VISUALIZATION

We used simple bar plots and grouped statistics to visualize the performance of the models, especially to assess prediction confidence. For example, predicted probabilities were grouped into buckets (e.g., 0.4–0.5, 0.5–0.6), and counts of true labels in each bucket were displayed to analyze how confident the model was across probability ranges. These insights helped confirm that the model lacked strong confidence in its predictions, with most scores clustered around 0.5.

### 3.4. STREAMING

We simulated a simple streaming setup by processing the data one day at a time, mimicking real-time ingestion. While not used in the final model, this helped demonstrate how the pipeline could support streaming scenarios.

## 4. CONCLUSION

This project demonstrated how Big Data Analytics, using Apache Spark MLlib, can be applied to stock market prediction with real-world, large-scale financial data. We designed a scalable pipeline to process historical prices from both the S&P 500 ETF and its corresponding futures contracts, along with inflation rates. By engineering lag-based features, we aimed to capture short-term dependencies, since futures prices reflect market expectations for the next day and can provide valuable predictive signals. Three models, the Logistic Regression, Random Forest, and a tuned Random Forest, were trained and evaluated. However, even after hyperparameter tuning, model performance (measured by ROC AUC) was only slightly better than random guessing, confirming that short-term market direction is inherently difficult to predict from price data alone.

Despite these modest results, the project highlights the important role of Big Data tools in efficiently handling and modeling large volumes of financial time series. A key lesson is that good predictive results are hard to achieve with historical prices and basic economic data, due to the high unpredictability of financial markets. The pipeline built here can be easily extended to test more advanced features such as macroeconomic indicators, sentiment analysis, or technical signals, and can be adapted for near real-time prediction using streaming data. Overall, this work demonstrates the value and potential of Big Data Analytics for financial modeling, while also underlining the practical challenges involved in forecasting short-term price movements.

## 5. CHANGES AFTER THE DEFENSE

Following the project defense, we implemented changes based on the feedback that our approach needed more financial and external data to enhance prediction accuracy. To address this, we integrated two key data sources into our dataset: inflation rates and S&P 500 futures prices. The addition of inflation data provided a broader economic context, while the inclusion of futures prices reflecting market expectations, this offered a forward-looking perspective that could improve the predictive power of our models.

We also engineered lag features from these new variables, capturing how past values influenced the target variable. These new features were incorporated into the model training pipeline to predict whether the market would go up or down the next day. This adjustment made our analysis more comprehensive and better aligned with real-world financial forecasting practices, directly responding to the recommendations from the defense.