- **Title:** Stock Price Movement Prediction Using Machine Learning

- **Subtitle:** An End-to-End Data Science Project (EDA & Modeling)
- **Presented by:** Misheck Kibunja
- **Date:** 23rd February 2025

Project Overview & Business Problem

•**Overview:**

•This project develops a predictive model for stock price movements

•by leveraging historical market data and technical indicators.

•**Business Problem:**

•Investors struggle with predicting stock price changes due to market volatility and complex signals.

General Business Questions

- How have stock prices and trading volumes evolved over time?

- How do different technical indicators behave over time, and how do they correlate with stock price movements?

- Are there noticeable patterns in stock price movements before and after key technical signals?

- Which features (indicators, volume trends, price patterns) show the strongest relationship with stock price changes?

Business Questions Related to Modeling

- What is the best-performing machine learning model for stock price classification?
- How does model performance compare to a baseline (e.g., random guessing or simple moving average strategy)?
- What is the optimal time horizon (daily, weekly, monthly) for stock movement predictions?
- Can our model generalize across different stocks, or does it work best for specific sectors?
- How does our model’s performance compare to traditional technical analysis strategies used by traders?
- What is the financial impact of using our model for trading decisions?

Data & Dataset Description

- **Dataset Source:**
Historical stock data from Yahoo Finance.

- **Data Coverage:**
25 years of data for 20 diverse stocks across multiple sectors.

- **Key Features:**
Market data: Open, High, Low, Close, Adjusted Close, Volume.
Technical indicators: Moving Averages (EMA, MA), RSI, Bollinger Bands).

- **Target Variable:**
Binary indicator: 1 (upward movement) vs. 0 (downward movement).

Snapshot of the Dataset

	Date	Open	High	Low	Close	Volume	Ticker	Target	MA_50	MA_200	...	BB_Mid	BB_Upper	BB_Lower	OBV	ATR_14	RSI_Si
0	1990-10-15 00:00:00	0.202147	0.203920	0.188848	0.196828	201017600.0	AAPL	0	0.239980	0.267437	...	0.208088	0.234434	0.181742	2.743238e+09	0.012458	
1	1990-10-16 00:00:00	0.195055	0.195055	0.172003	0.177322	305233600.0	AAPL	1	0.237940	0.267015	...	0.205118	0.231263	0.178972	2.438005e+09	0.013341	
2	1990-10-17 00:00:00	0.179095	0.187962	0.177322	0.187962	309064000.0	AAPL	1	0.236113	0.266637	...	0.202990	0.227293	0.178687	2.747069e+09	0.013148	
3	1990-10-18 00:00:00	0.187962	0.203920	0.187962	0.202147	315000000.0	AAPL	1	0.234481	0.266325	...	0.201881	0.224015	0.179748	3.062069e+09	0.013349	

Data Preprocessing & Feature Engineering

- **Data Cleaning:**
  - Removed non-numeric columns (Date, Ticker) and handled missing values.

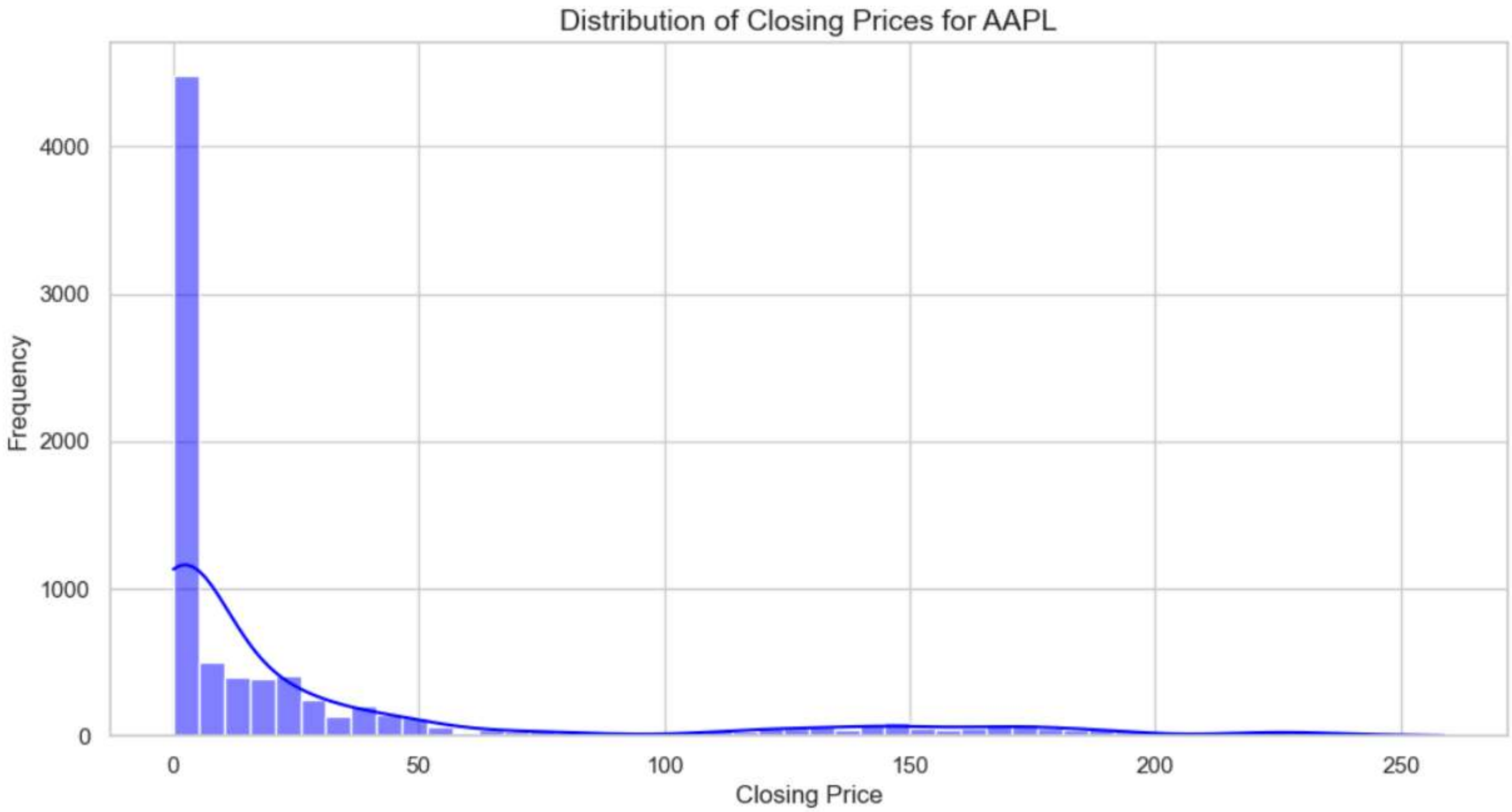
- **Feature Engineering:**
  - Calculated technical indicators such as EMA, RSI, and Bollinger Bands.
  - Created additional features like shifted closing prices (Close_Day_1, Close_Day_2, Close_Day_3).

- **Feature Selection:**
  - Applied Recursive Feature Elimination (RFE) to select the most predictive features.

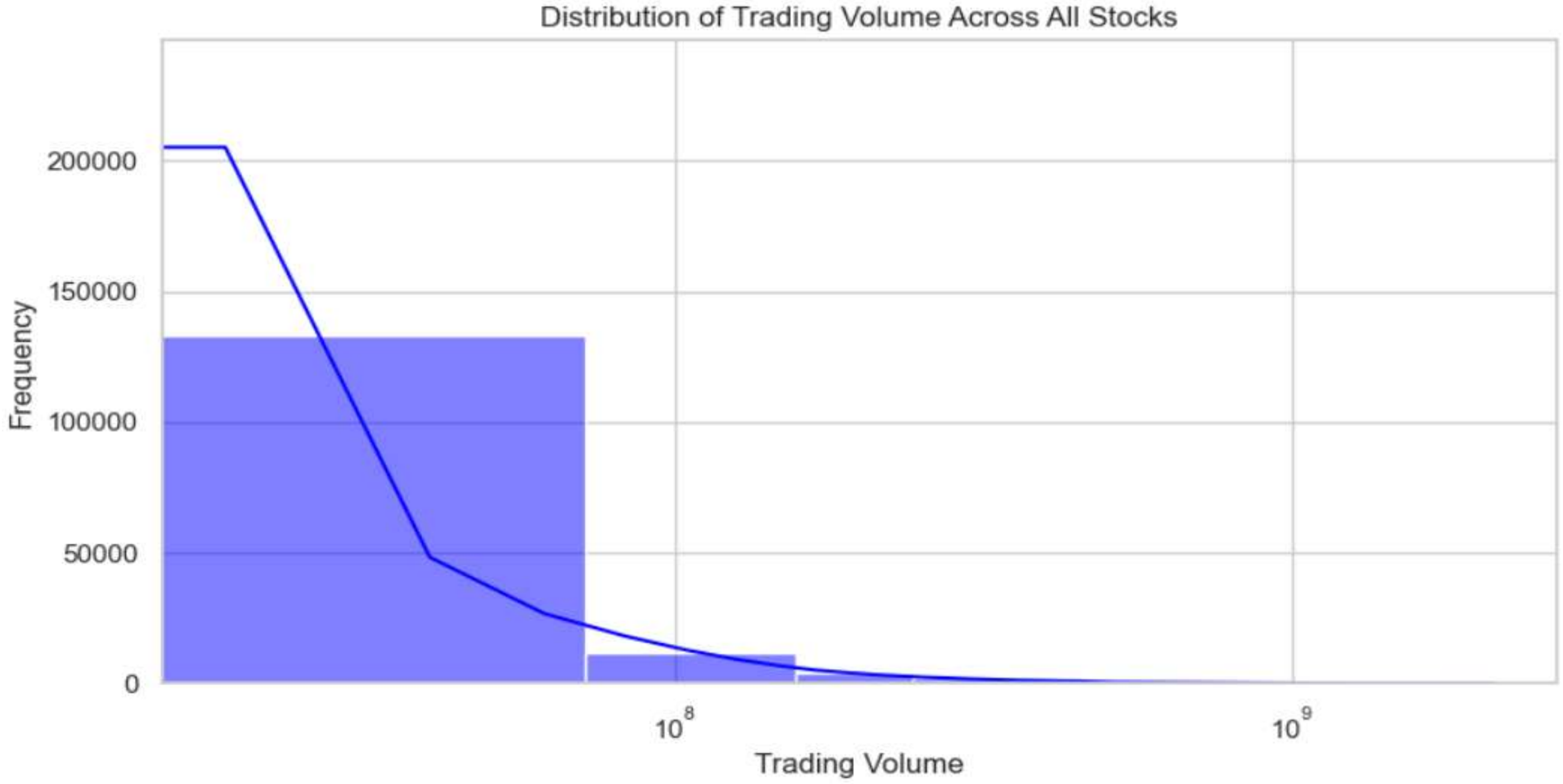
- **Train-Test Split:**
  - Data was split into training (80%) and testing (20%) sets with stratification.

Exploratory Data Analysis (EDA) Univariate Analysis

- Distribution of closing Prices



Univariate Analysis
Distribution of Trading Volume (Across All Stocks)



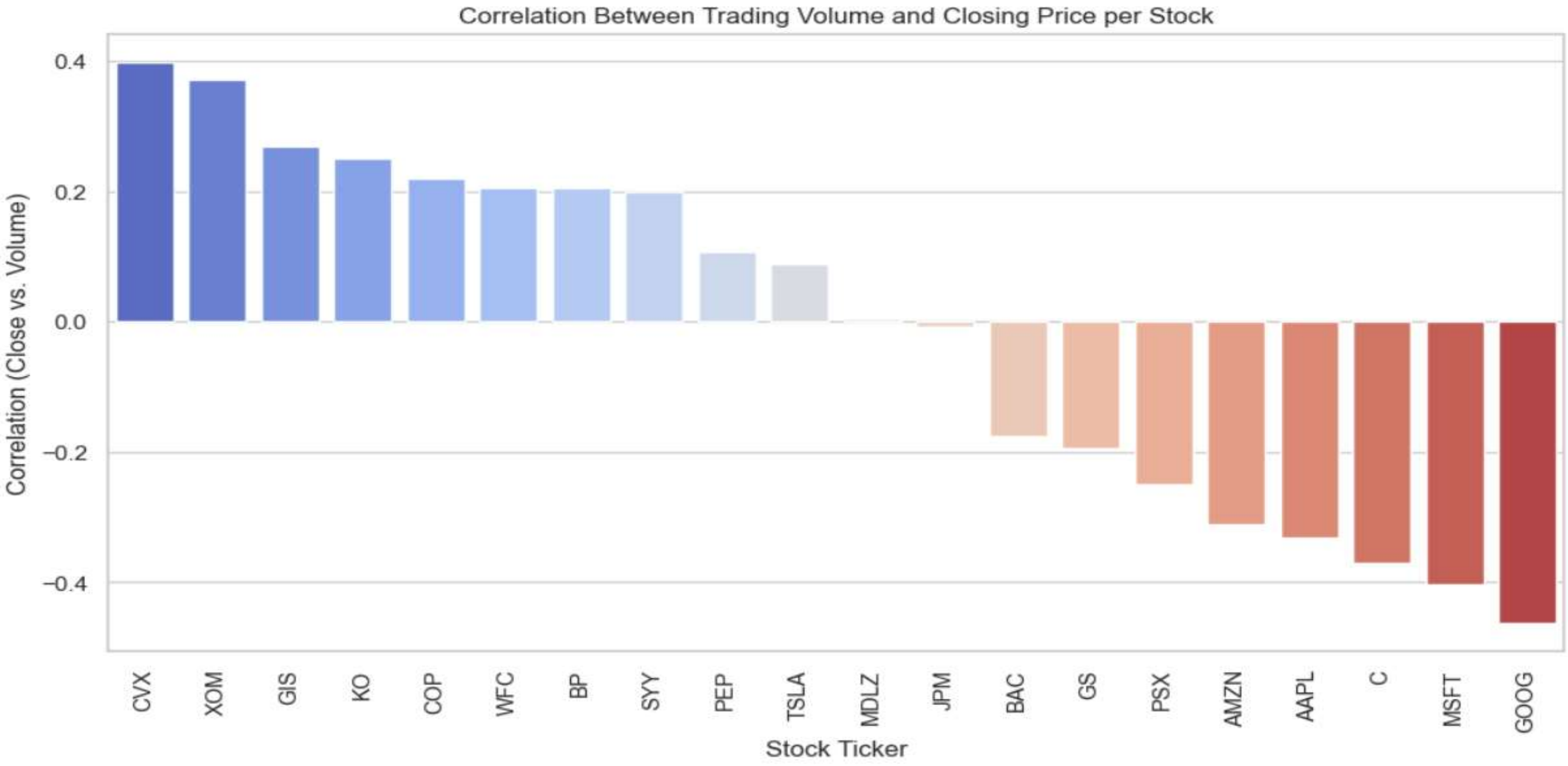
Analysis of the Distribution of Trading Volume

The histogram of **trading volume** across all stocks is **right-skewed**

This suggests that **liquidity is concentrated in a handful of stocks.**

The generally long right tail indicates that some stocks occasionally experience **large volume spikes**, possibly due to earnings reports, news events, or major institutional trades.

Bivariate Analysis
Correlation Between Trading Volume and Stock Price



Multivariate Analysis

Modeling Approach – Overview of Techniques

•**Models Explored:**

1.Logistic Regression (baseline)

2.Decision Tree (basic unpruned and pruned versions)

3.Random Forest

•**Key Steps:**

•Data preprocessing and feature selection via RFE.

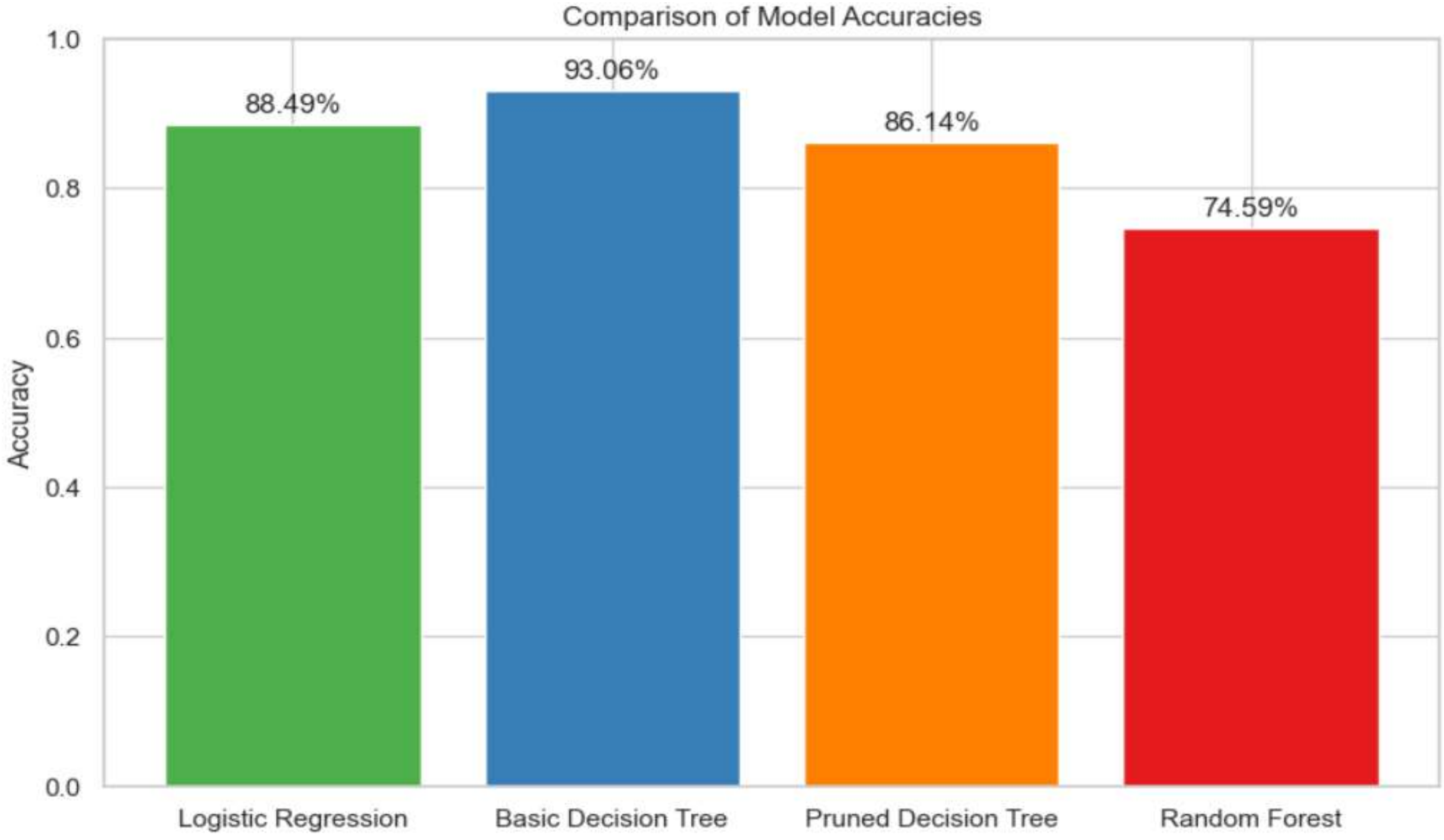
•Hyperparameter tuning using RandomizedSearchCV.

•Evaluation using accuracy, confusion matrices, and classification reports.

•**Class Imbalance:**

•Some models incorporated class balancing techniques.

Modeling Results – Logistic Regression vs. Decision Tree vs. Random Forest



Modeling Results

Comparison with Previous Models:

Model	Test Accuracy	Precision (0/1)	Recall (0/1)	F1-Score (0/1)	Observations
Logistic Regression	88.49%	0.88 / 0.89	0.88 / 0.89	0.88 / 0.89	Best overall performance, highly balanced model
Basic Decision Tree (Unpruned)	93.06%	0.93 / 0.93	0.93 / 0.93	0.93 / 0.93	Likely overfitting, extremely high accuracy
Pruned Decision Tree (No Class Balancing)	86.14%	0.85 / 0.87	0.87 / 0.85	0.86 / 0.86	Balanced performance, reduced overfitting
Further Pruned Decision Tree (Class Balancing)	52.38%	0.54 / 0.52	0.30 / 0.75	0.38 / 0.61	Overcorrection of class imbalance led to poor accuracy
Random Forest	74.59%	0.80 / 0.71	0.65 / 0.84	0.72 / 0.77	Improved over pruned trees, but still less effective than logist regression

Brief Interpretation

- **Logistic Regression:**
- Accuracy ~88.49%
- Balanced metrics across classes.

- **Basic Decision Tree (Unpruned):**
- Accuracy ~93.06% (but overfitting issues).

- **Pruned Decision Tree (Without Class Balancing):**
- Accuracy ~86.14%

- **Pruned Decision Tree with Class Balancing:**
- Accuracy ~52.38% (underperforming due to imbalance overcorrection).

- **Random Forest:**
- Accuracy ~74.59%
- Improved over some pruned trees but not as effective as logistic regression.

Final Model Selection & Comparison

•**Best Model So Far:**

•The basic logistic regression model, with an accuracy of ~88.49%, emerged as the best overall due to its strong generalization, balanced performance, and interpretability.

•**Model Comparison:**

•Although the unpruned decision tree showed higher accuracy, its overfitting undermines reliability.
•Pruned decision trees and random forest did not achieve competitive accuracy when adjusted for overfitting or class imbalance.

•**Conclusion:**

•Logistic regression offers the best trade-off between accuracy, robustness, and interpretability.

Future Work & Improvements

•**Advanced Models:**

•Experiment with ensemble methods like Gradient Boosting, XGBoost, or even deep learning approaches.

•**Feature Engineering:**

•Incorporate additional data such as news sentiment, macroeconomic indicators, and earnings reports.

•**Time Horizon Optimization:**

•Evaluate performance for different prediction intervals (daily, weekly, monthly).

•**Backtesting:**

•Quantify the financial impact through simulated trading strategies and risk analysis.

Conclusion

- This project demonstrates a complete end-to-end process for predicting stock price movements using machine learning.

- Our best model (logistic regression) outperforms traditional methods and provides a strong basis for further enhancements.