STOCK MARKET PREDICTION USING ML

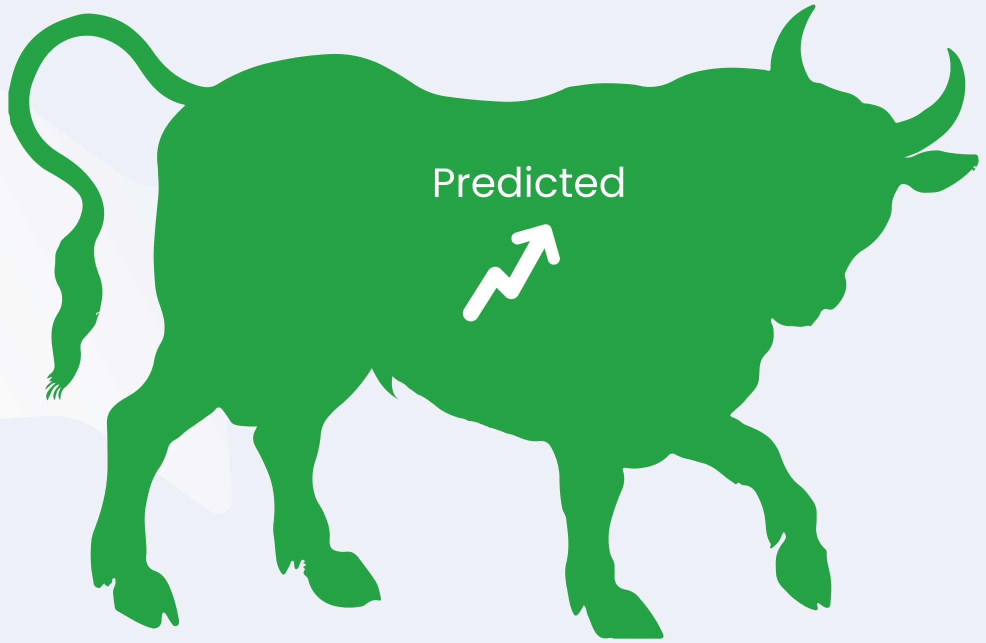
Predicting GOOGL Next-Day Price Movement

Yi Zhang, Songnan Zhao, Huizhen Zheng

INFO 6105 Final Project

2025.12.08

Problem & Objective



What

- Predict whether GOOGL stock will go UP or DOWN tomorrow
- Binary Classification: **Tomorrow_Up = 1 or 0**

The Challenge

- Stock market prediction is one of the most difficult problems in finance, which is **influenced by several factors**.
- Influenced by news, sentiment, global events, and market psychology.
- We aim to test whether ML + technical features can **perform above random guessing** and understand the ML pipeline.

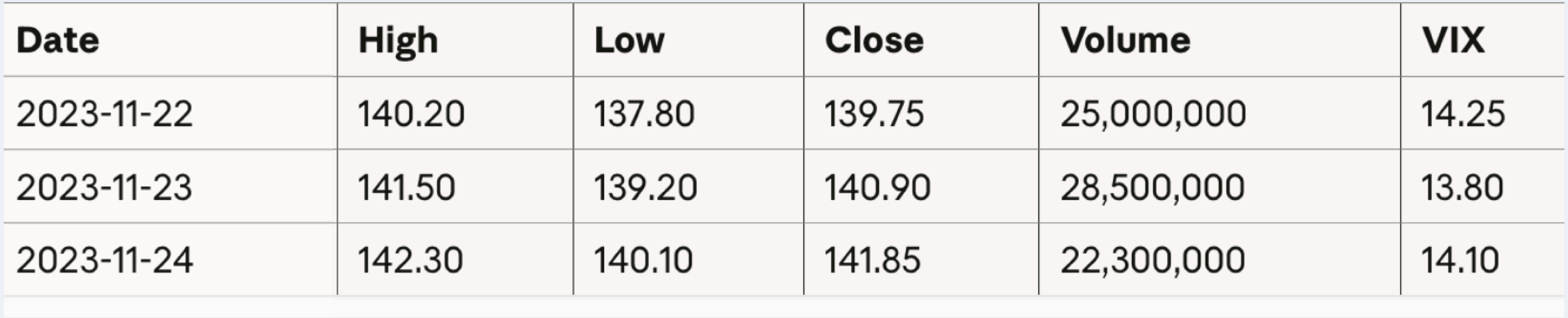
HOW

- Combine stock data + market sentiment data(**Yahoo Finance + FRED**)
- Compare **3 ML models**: Decision Tree → Random Forest → XGBoost
- Evaluate with **Accuracy, F1-Score, Precision**

Data Source--Extract

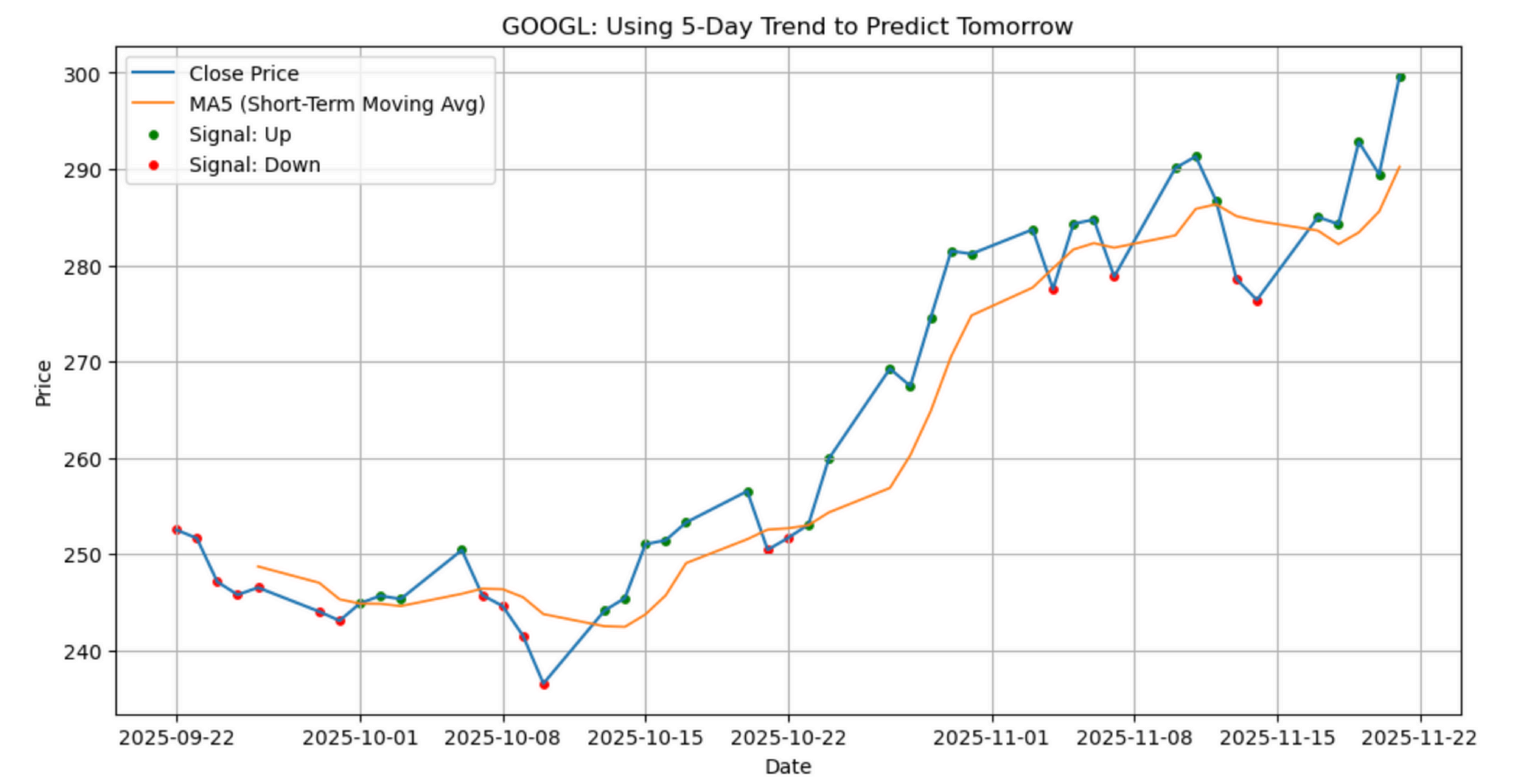
Original Source	Data Type	Features
Yahoo Finance API	Stock Data	High, Low, Close, Volume
FRED API	Economic Data	VIX (Chicago Board Options Exchange Volatility Index)

ORIGINAL DATASET

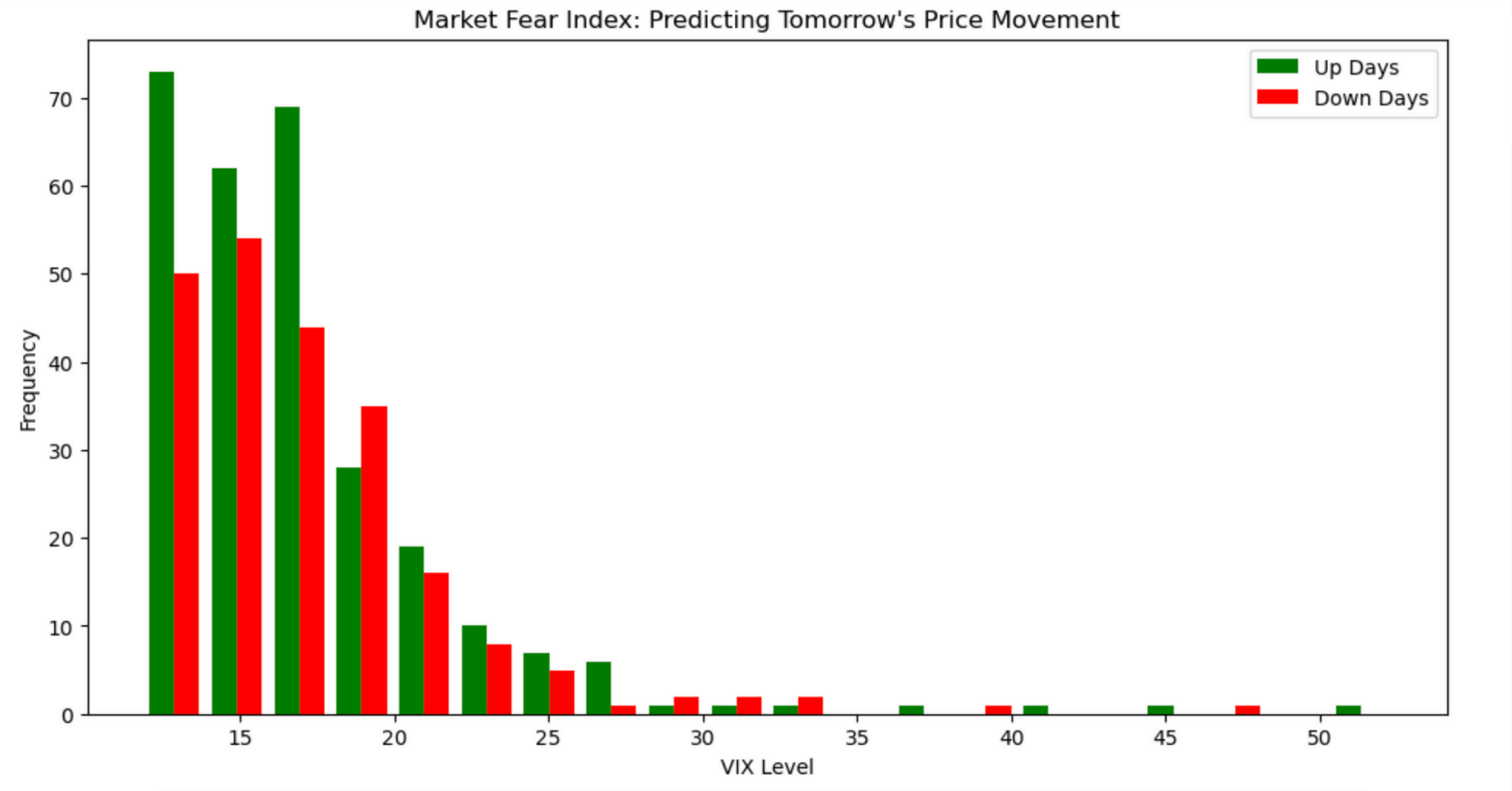


Date	High	Low	Close	Volume	VIX
2023-11-22	140.20	137.80	139.75	25,000,000	14.25
2023-11-23	141.50	139.20	140.90	28,500,000	13.80
2023-11-24	142.30	140.10	141.85	22,300,000	14.10

Visualization



- The Short_Trend_Signal feature (Close > MA5) helps predict next-day price movement
- Green points (price > MA5) often precede continued upward movement



- Shows the distribution of VIX levels for up vs down days
- Helps identify patterns: lower VIX → higher up probability
- Validates that our engineered features have predictive value for the ML model

Feature Engineering

Features	Data Meaning	Why It Matters
Returns	Daily price change %	Trends tend to persist (momentum effect), Stocks that have risen recently are likely to continue rising in the short term.
Volume_Change	Volume change %	Market activity signal
High_Low_Ratio	Volatility	Measures how much the price jumped around during the day. Big jumps = nervous market
VIX	Market Fear	High VIX typically correlates with market declines.
VIX_Change_5d	5-day VIX trend	Sentiment momentum
Short_Trend_Signal	Price > 5-day MA	Price above 5-day MA suggests upward momentum; below suggests downward pressure

Dataset Overview

- Time Period: Nov 2023 - Nov 2025 (2 years)
- Total Samples: 497 trading days
- Train/Test Split: 80% / 20% (397 / 100 days)

DECISION TREE

How it works:Gini Impurity

Gini Impurity: Split data to make each group as "pure" as possible

$$Gini = 1 - (P\_up)^2 - (P\_down)^2$$

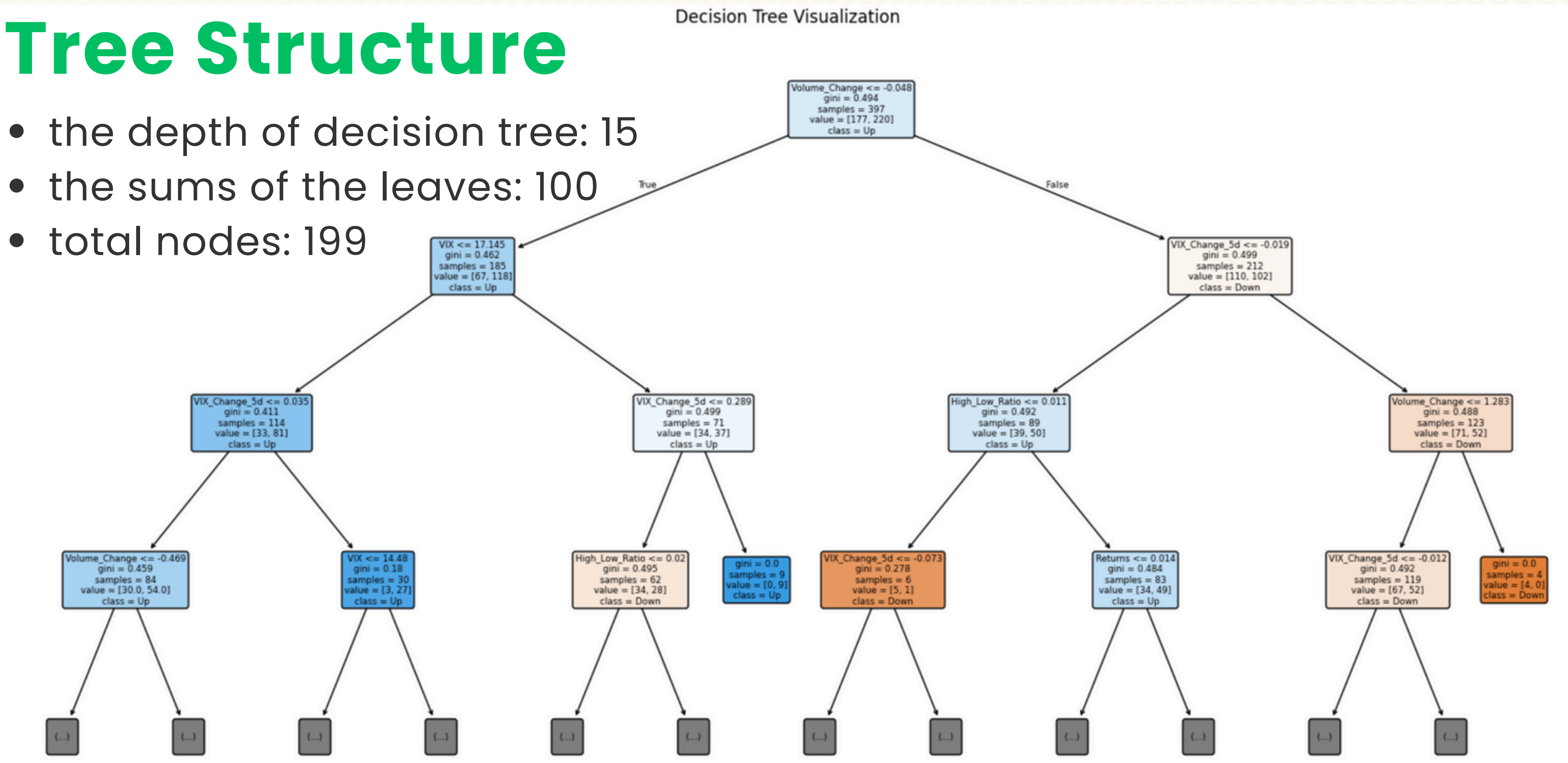
- ✓ Calculate Gini of current node
- ✓ Try all features * ALL thresholds
- ✓ Pick split with LOWEST Gini
- ✓ Repeat for each child node
- ✓ Stop when node is pure or too small

What is Decision Tree

A tree-structured model that makes predictions by asking a series of Yes/No questions

Tree Structure

- the depth of decision tree: 15
- the sums of the leaves: 100
- total nodes: 199

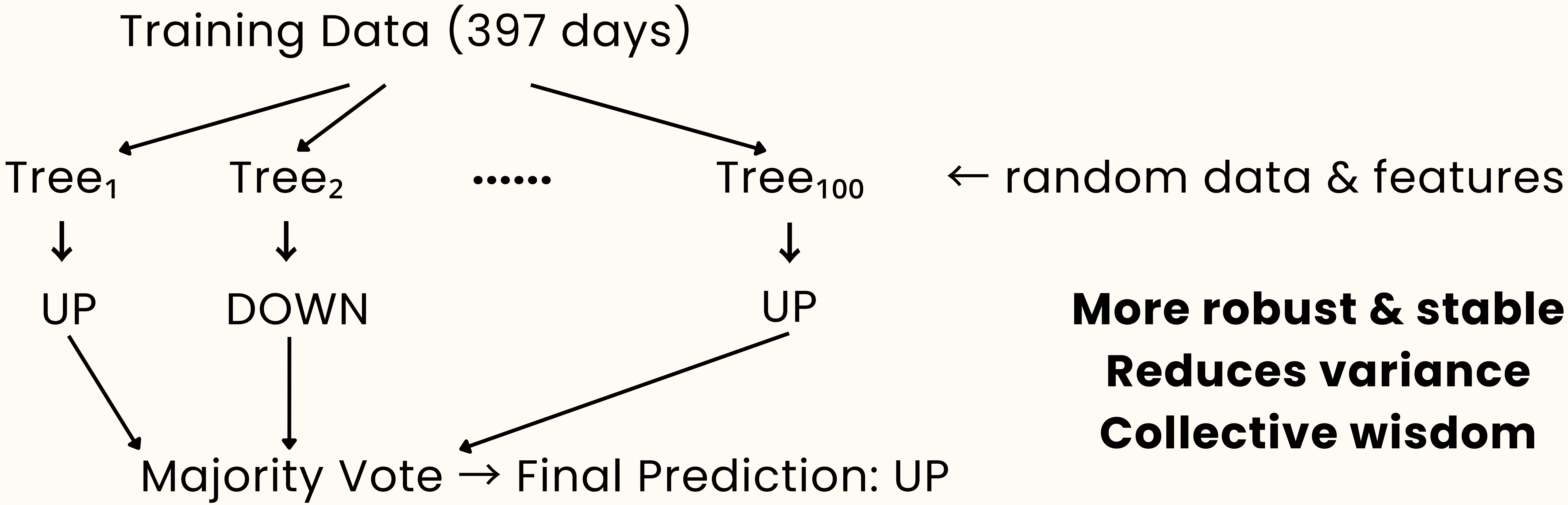


- Model may be **overfitting** — learning noise rather than patterns(accuracy is only 53%).
- failing to capture true market patterns.

RANDOM FOREST

Core Idea

Many different decision trees vote together → Majority wins



More robust & stable

Reduces variance

Collective wisdom

XGBOOST

What is XGBoost

An advanced ensemble method that builds trees sequentially, each tree learning from the mistakes of previous ones

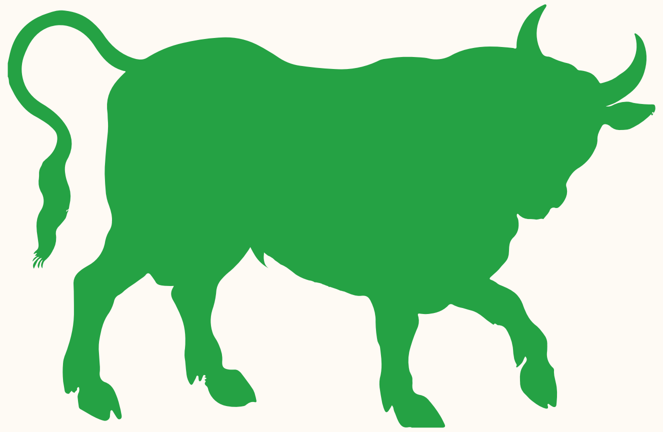
Hyperparameter Tuning Process

We tested 8 combinations (2×2×2) to find the sweet spot:

Parameter	Options Tested	Winner	Why It Won
n_estimators	[150, 300]	150	Enough trees without overdoing it
max_depth	[3, 5]	3	Less overfitting
learning_rate	[0.01, 0.1]	0.01	Small steps work better for noisy data

Key Insight:
Conservative settings (shallow trees + slow learning) work best for GOOGL's price patterns in our 2-year dataset

XGBOOST RESULTS



Confusion Matrix:

	Pred_Down	Pred_Up
Act_Down	16 (TN)	24 (FP)
Act_Up	17 (FN)	43 (TP)

How We Calculate Each Metric:

Accuracy = (True Negative + Ture Positive) / (Total)
= (16 + 43) / 100 = 0.59

Precision = True Positives / (True Positives + False Positives)
= 43 / (43 + 24) = 0.64

Recall = True Positives / (True Positives + False Negatives)
= 43 / (43 + 17) = 0.72

F1-Score = 2 × (Precision × Recall) / (Precision + Recall)
= 2 × (0.64 × 0.72) / (0.64 + 0.72) = 0.68

What This Means:

- Accuracy 0.59: Overall correctness
- Precision 0.64: When we predict UP, we're right 2 out of 3 times
- F1-Score 0.68: Good balance between precision and recall

While not perfect, the model provides actionable trading signals

Good for Users: "Buy and hold" strategies, identifying entry points

Results Comparison

Key Insight: XGBoost outperforms in all metrics

3

Decision Tree

Accuracy:0.53

F1-Score: 0.58

Precision:0.63

2

Random Forest

Accuracy:0.53

F1-Score: 0.62

Precision:0.60

XG Boost

Accuracy:0.59

F1-Score: 0.68

Precision:0.64

Future Work & Improvements

Expand Data Sources

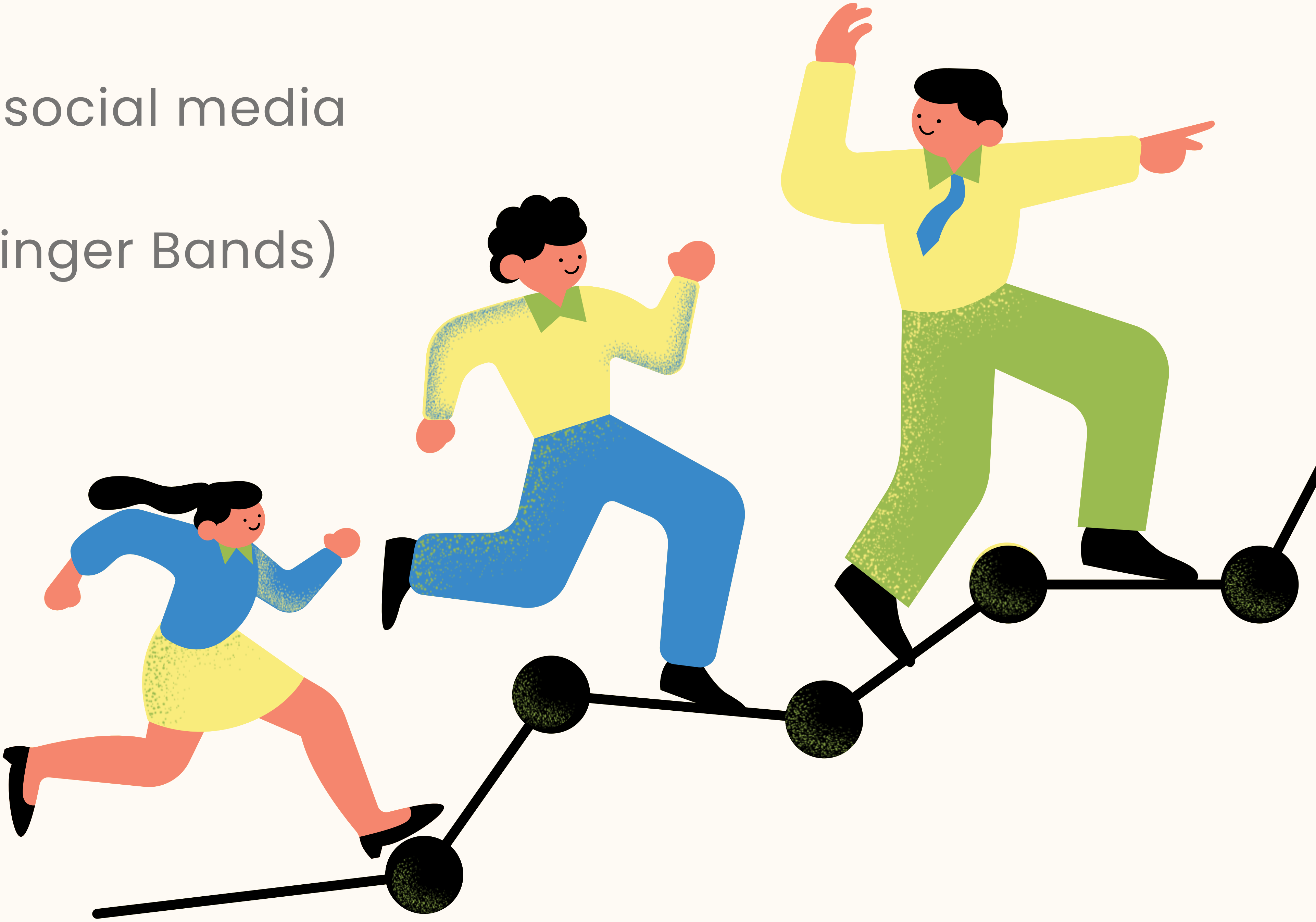
- Integrate news APIs (e.g.NewsAPI,GDELT),social media sentiment (Twitter/Reddit)
- Add technical indicators (RSI, MACD, Bollinger Bands)

Practical Implementation

- Build real-time prediction system
- Backtest on multiple stocks/sectors

Expected Impact

- Target: 65% accuracy with enhanced features



Thank You