Stock Market Prediction

Deep Dive with Classic, Supervised Learning, and Deep Learning Approaches

CuriousMinds

- Sophia Yang (xya134)
- Weiwei Zhang (wza124)
- Yingzi Yuan (yya309)

Motivation

Stock price prediction is a well-known challenge due to its complexity and unpredictability.

Why is it an important project?

- Help investors make better decisions
- Reduce financial risks and increases profits
- AI models help uncover complex patterns in stock data and catch hidden trends

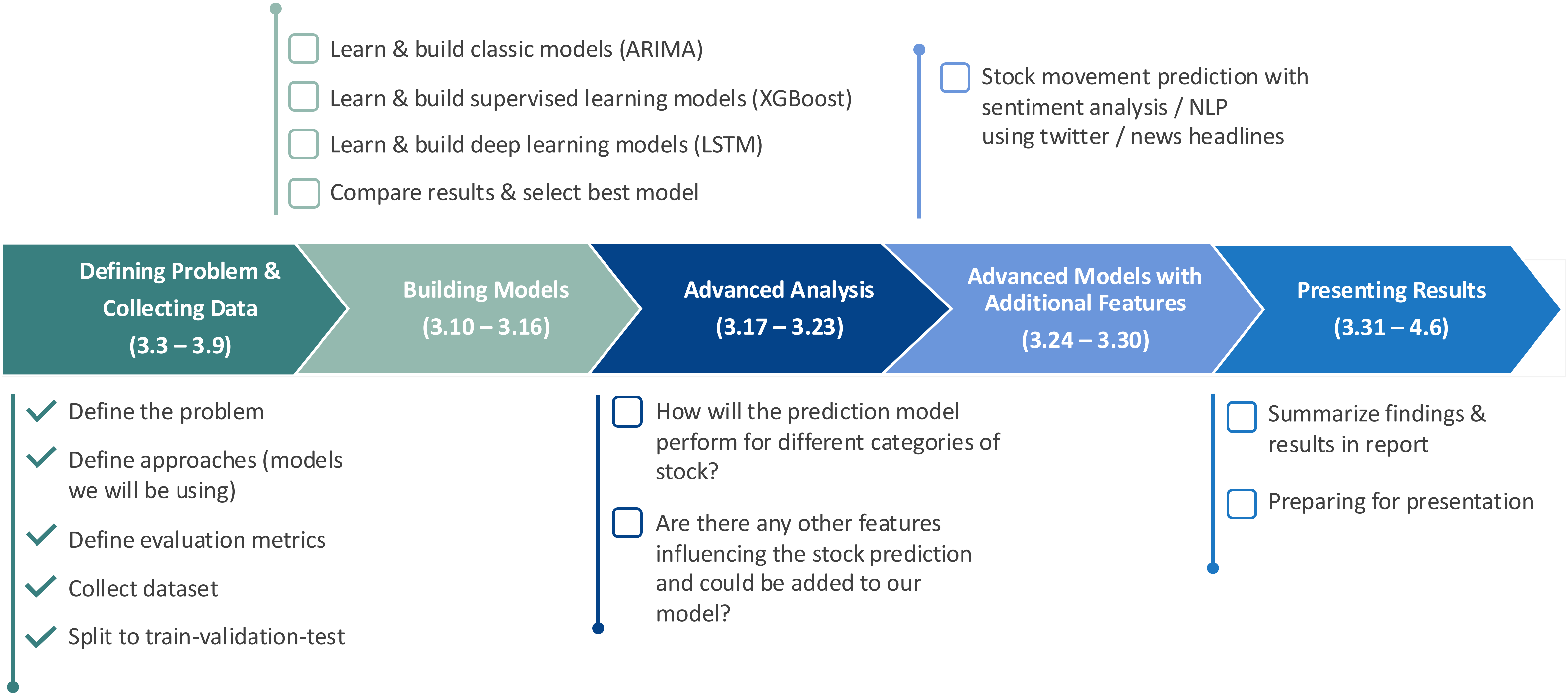
Objective

Building and comparing different prediction models (ARIMA, XGBoost and LSTM) to find the best way to forecast stock prices while exploring external factors to improve accuracy.

Challenges

- Noisy and Unpredictable Stock Data
- Model Complexity, Overfitting and Feature Challenges
- Tight Schedule with Heavy Workload and Extra Analysis

Project Timeline



Defining Problem & Collecting Data (3.3 – 3.9)

- ✓ Define the problem
- ✓ Define approaches (models we will be using)
- ✓ Define evaluation metrics
- ✓ Collect dataset
- ✓ Split to train-validation-test

Building Models (3.10 – 3.16)

- □ Learn & build classic models (ARIMA)
- □ Learn & build supervised learning models (XGBoost)
- □ Learn & build deep learning models (LSTM)
- □ Compare results & select best model

Advanced Analysis (3.17 – 3.23)

- □ How will the prediction model perform for different categories of stock?
- □ Are there any other features influencing the stock prediction and could be added to our model?

Advanced Models with Additional Features (3.24 – 3.30)

- □ Stock movement prediction with sentiment analysis / NLP using twitter / news headlines

Presenting Results (3.31 – 4.6)

- □ Summarize findings & results in report
- □ Preparing for presentation

Current Progress

- **Defined the problem** ✓

  Predict future stock prices based on historical data.

- **Decided on approaches and selected models** ✓

  ARIMA, XGBoost, LSTM.

- **Defined evaluation metrics** ✓

  RMSE (Root Mean Square Error) or MAPE (Mean Absolute Percentage Error)

- **Collected stock data using Yahoo Finance API (slide 6)** ✓

  Downloaded and preprocessed data from 2015 to 2025.

- **Split the dataset & Normalize features (slide 7)** ✓

  Training, validation, and testing sets based on time sequence.

Evidence of Data Collection

Source: Yahoo Finance

Range: 10 years (2015 – 2025)

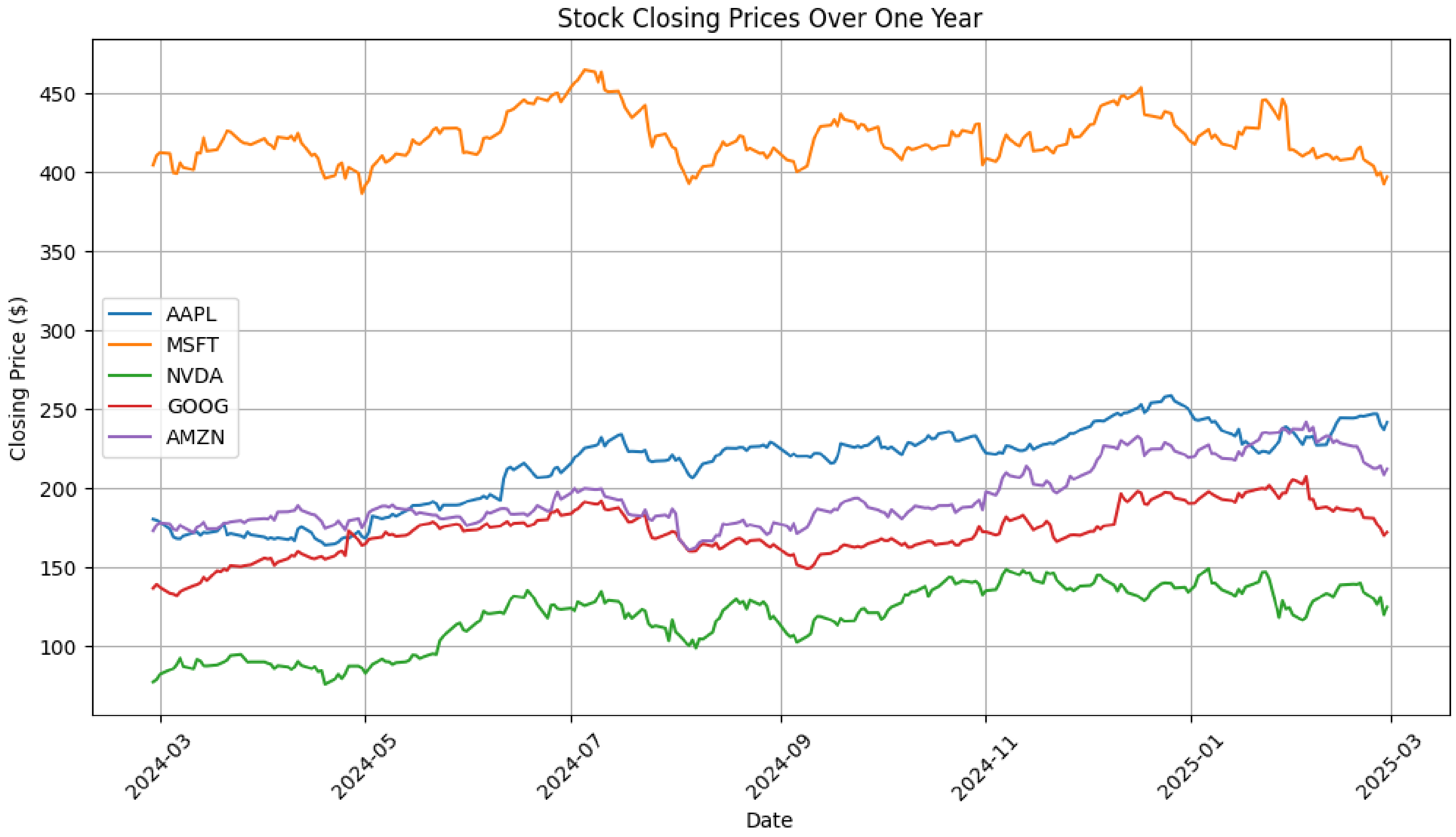
Features:

- close, high, low, open, volume

Datasets:

- Single stock for simplicity (Apple)
- Multi-stocks from different sectors

Sector	Stock Ticker
Technology	AAPL, MSFT, NVDA, GOOG, META
Finance	JPM, GS, BAC, WFC, MS
Energy	XOM, CVX, BP, COP, SLB
Healthcare	JNJ, PFE, MRK, UNH, ABBV
Consumer Goods	TSLA, AMZN, WMT, MCD, KO



Evidence of Dataset Splitting & Feature Normalization

Dataset Splitting

Training & Validation	Testing
The oldest 80% of data	The latest 20% of data

Training set will use **5-folds cross validation** for hyper-parameter tuning

Feature Normalization

1

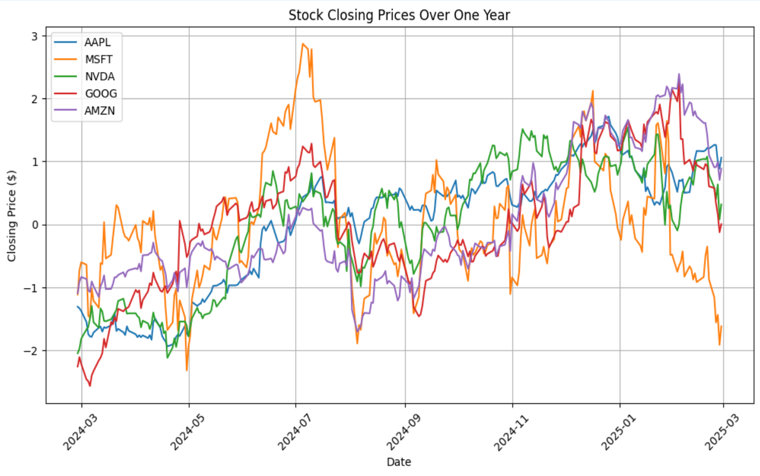
Min/Max Normalization

✗

Stock is ever changing, doesn't make much sense using min and max

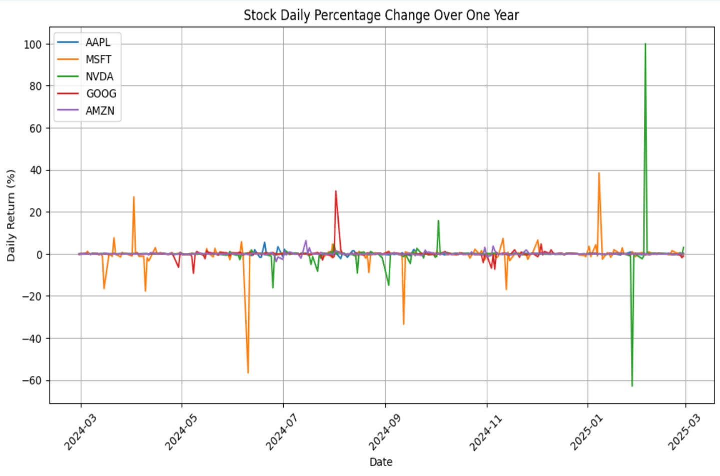
2

Standard Normalization



3

Use Percentage Change instead of absolute value


Stock Daily Percentage Change Over One Year
AAPL
MSFT
NVDA
GOOG
AMZN
Daily Return (%)
Date