QUANTITATIVE FINANCE AND RESEARCH PROJECT REPORT

CAIC SUMMER OF TECH 2025 – WEEK 2

1)Introduction:

This project applies machine learning and statistical models to analyse and predict the behaviour of selected large-cap U.S. stocks. The focus is on understanding financial indicators, engineering relevant features, using models, and simulating a trading strategy.

Stocks Analysed:

- Apple (AAPL)
- Amazon (AMZN)
- Google (GOOGL)
- Microsoft (MSFT)
- Tesla (TSLA)

2) Dataset Overview:

The dataset has been taken from the website Kaggle, which was initially named “Price and Volume Data for All US Stocks”. It has been filtered to show the data only for the past 10 years. It has a multi-indexed structure with ticker as the outer index and date as the inner index.

3) Data Processing:

First, I checked for missing and duplicate values in the dataset. Then I removed the outliers and filtered the dataset to include only the last 10 years of available data. Then I sorted the data chronologically and standardised the formats to ensure data integrity.

QUANTITATIVE FINANCE AND RESEARCH PROJECT REPORT

4) Feature Engineering:

To improve the quality of the dataset and consequently improve the analysis, I incorporated the following features into the dataset:

- Moving Averages (7-day, 30-day)
- RSI-14, MACD
- Rolling Volatility (30d)
- High-Low Spread, Intraday Volatility%
- Lagged Daily Returns
- Volume-based metrics (change, ratio)

5) Exploratory Data Analysis:

I analysed the dataset using the features and found the following:

- TSLA showed the highest average return
- AMZN was the most volatile (Nov 2008)

I also addressed the skewness of the close prices via log transformation and included some correlation plots and volatility charts.

6) Modelling:

For modelling, I chose AMZN as my sample stock. I chronologically split it into 80% training and 20% testing datasets. I proceeded to try the following models:

- Linear Regression (lag features)
- ARIMA (time series model)
- Random Forest Regressor (technical indicators)

QUANTITATIVE FINANCE AND RESEARCH PROJECT REPORT

7) Evaluation Metrics:

I chose Mean Absolute Error and Direction Accuracy as my evaluation metrics. The results were as follows:

Model	MAE	Direction Accuracy
Linear Regression	0.0140	54.99%
ARIMA	138.71	46.76%
Random Forest	0.1631	60.10%

8) Back testing Strategy:

My back testing strategy was:

- Generate buy/sell signals: If the predicted price > current price, sell; otherwise, buy.
- Calculate the hypothetical profit/loss over the test period.

I used the Random Forest model, since it gave the best results. The results were as follows:

- Strategy Profit: $436.99
- Buy & Hold Profit: $339.73

9) Conclusion:

In this project, we explored and modelled historical stock data using a combination of technical indicators and machine learning techniques. The Random Forest Regressor delivered the best prediction accuracy and trading profitability performance among the models tested.