Group 3 Project Proposal – Maksimas Ambrasas, Brandon Cullen, Abel Rajesh

Introduction

Our project focuses on a stock market dataset from Kaggle [1], which spans daily stock data from 1999 to 2021. It covers four US exchanges - NASDAQ, NYSE and S&P 500 and Forbes2000. This offers a broad perspective on market activity over two decades. There are 409 companies in the S&P500, 1564 companies in the NASDAQ, 1145 companies in the NYSE and 1076 in the Forbes2000. We will use Finnhub's stock APIs to fetch which sector each company is related to based on the ticker given in the dataset, e.g. DIS is Walt Disney and would be mapped to Media. Using this Kaggle dataset, we will form our own custom dataset that will be applicable for our research question. The dependent variable will be sector level growth; this is a feature defined as the percentage change in adjusted closing prices aggregated across all companies within a sector over a month and year timespans. The main inputs will be company-level features (Open, High, Low, Close, Adjusted Close, Volume), which will be aggregated into sector-level features such as average return, volatility, median growth and average volume.

Literature Review

There is a lot of information online related to using data to understand and predict industry growth. Recent research has explored using machine learning to analyse financial markets and identify growth trends. Many studies use company-level data, applying hybrid models that combine deep learning techniques such as LSTM and CNN with ensemble methods like XGBoost and Random Forest to predict stock price movements. These models capture temporal patterns in historical prices and the complex relationships among market indicators though most focus on short-term predictions rather than forecasting growth at the sector level [2].

Other studies examine high-growth companies using traditional machine learning methods including Logistic Regression, Random Forest and Gradient Boosting. They highlight the value of features such as market capitalisation, revenue, investment activity and macroeconomic indicators in identifying firms with strong growth potential. However, these approaches often concentrate on individual companies and do not account for broader industry trends or cross-sector interactions, limiting their usefulness for sector-wide forecasting [3].

In contrast, our project aggregates company-level data to sector-level features such as average return, volatility, median growth and average volume. We aim to predict sector growth rather than individual company performance, comparing the impact of monthly versus yearly averaged data. Furthermore, we will apply both time-series models (ARIMA or LSTM) and regression-based models to evaluate their effectiveness in forecasting which sectors are likely to experience the fastest growth over the coming month or year. This approach addresses the gap in existing research by focusing on sector-level forecasting.

Research Question

Our research question is: **"Predicting the fastest-growing sectors from aggregated company stock data: Comparison of time-series and regression based models"**. We would like to see the impact that monthly averaged and yearly averaged data has on our different models. Addressing this question will help identify sectors (such as technology or media) likely to experience the largest growth in the next month/year using historical data from NASDAQ, NYSE, S&P 500 and Forbes2000. By examining stock performance, trading volumes and potential market capitalisation, we aim to provide insights that are valuable for investors, portfolio managers and policymakers seeking to understand fast-growing sectors in the US economy.

References

[1] https://www.kaggle.com/datasets/paultimothymooney/stock-market-data/code

[2] https://www.researchgate.net/publication/359786453_Attention-based_CNN-LSTM_and_XGBoost_hybrid_model_for_stock_prediction

[3] https://www.sciencedirect.com/science/article/abs/pii/S0360835219304838