Final Project Report

Pham Hoang Nam

Part 1: Introduction and NASDAQ Market Implementation

- **Executive Summary**

This report details the development and implementation of a comprehensive stock market prediction and portfolio optimization system spanning both NASDAQ and Vietnamese markets. The project, undertaken over a period of 1 month, demonstrates the practical application of deep learning techniques in financial markets while highlighting the challenges and solutions encountered during implementation.

- **Initial Approach and Data Processing**

The project began with ambitious goals of creating a production-ready system worthy of inclusion in a professional portfolio. The first major challenge emerged immediately with data processing requirements that far exceeded typical academic implementations. Unlike previous experiences with single-file processing, this project demanded handling thousands of company files contained within compressed archives.

- **Data Loading Implementation**

The initial data processing challenge required developing a sophisticated file handling system. Using a combination of the zipfile library and pandas, I created a robust data loading pipeline capable of:

- Extracting multiple CSV files from compressed archives
- Handling various file encodings (utf-8 and latin-1)
- Implementing error handling for corrupted or incomplete files
- Processing files in batches to manage memory efficiently

Company Selection Methodology

The filtering criteria were carefully designed to ensure data quality and reliability:

- Companies must be listed in both NASDAQ and S&P 500 indices
- Minimum requirement of 120 historical data points
- Complete data availability across all required features
- Consistent trading history without significant gaps

This rigorous filtering process identified 115 qualified companies, providing a balanced dataset for model development and testing.

- **Market Sector Considerations in Model Application**

An essential aspect of my NASDAQ stock price prediction implementation concerns the appropriate application of trained models. My analysis demonstrates that prediction models perform most effectively when applied to companies within the same sector and industry. For instance, the model trained on Apple Inc. (AAPL) data is best suited for companies in the Technology sector, specifically within the Consumer Electronics industry.

This sector-specific approach is grounded in several key factors:

Industry-Specific Patterns

Companies within the same sector often exhibit similar trading patterns and respond comparably to market conditions. The Technology sector, for example, demonstrates distinct characteristics in terms of price movements, volatility, and trading volumes that differ significantly from other sectors such as Healthcare or Financial Services.

Common Market Drivers

Companies in the same industry typically face similar market drivers and challenges. For Consumer Electronics companies, these might include:

- Product development cycles
- Component supply chain dynamics
- Consumer spending patterns
- Technological advancement periods

Market Correlation

My analysis shows that companies within the same sector often demonstrate stronger price movement correlations compared to cross-sector relationships. This correlation strengthens the predictive power of our models when applied within their trained sector.

Implementation Guidelines

When deploying the prediction model:

1. Identify the sector and industry classification of the training data company
2. Apply the model primarily to companies within the same classification
3. Exercise caution when attempting cross-sector predictions
4. Consider developing sector-specific models for broader market coverage

This sector-specific approach significantly enhances the reliability and accuracy of our prediction system, providing more actionable insights for investment decision-making

Task 1.1: Multi-feature Extension

The initial implementation focused on expanding beyond single-feature prediction to incorporate a comprehensive set of market indicators:

Feature Selection and Justification:

1. Low Price: Captures support levels and downside risk
2. High Price: Indicates resistance levels and upside potential
3. Open Price: Represents market sentiment at trading start
4. Close Price: Reflects daily trading outcome
5. Adjusted Close: Accounts for corporate actions
6. Volume: Measures trading intensity and market interest

Model Architecture Development:

The deep learning architecture underwent several iterations before reaching its final form:

- Input Layer: Designed to handle multiple features simultaneously
- LSTM Layers: Two layers (32 and 64 units) for temporal pattern recognition
- Dense Layer: 100 units with ReLU activation for non-linear pattern detection
- Output Layer: Single unit for price prediction

Early Implementation Challenges:

A significant challenge emerged during initial testing when predicted values consistently fell below actual prices despite capturing correct patterns. This led to an intensive period of model architecture experimentation, with over 50 different configurations tested over several days. The breakthrough came after a mid-term break consultation with professors, revealing that the issue lay not in the model architecture but in the data preprocessing pipeline.

The key insight was the importance of proper denormalization after prediction. This experience highlighted that successful machine learning projects require meticulous attention to every step of the pipeline, not just model architecture. The corrected implementation achieved a test MSE of 0.0119, demonstrating strong predictive accuracy.

Data Normalization Process:

The final normalization implementation included:

- Feature-wise MinMax scaling
- Proper handling of time-series windows
- Careful denormalization of predictions
- Validation of scale consistency across features

Task 1.2: K-th Day Forecast Implementation

Building upon the success of the multi-feature model, we extended our prediction capabilities to forecast specific future days, with k=25 as our implementation target. This extension maintained our proven LSTM architecture while adapting the data preparation and training approach to accommodate longer-term predictions.

Data Preparation Modifications

The k-th day prediction required careful modification of our data preparation pipeline while maintaining the core model architecture. The key changes focused on label selection and window alignment. Instead of targeting the next day's price, we adjusted our labels to target the k-th day ahead. This modification allowed us to maintain our successful architectural design while extending our prediction horizon.

Our data preprocessing retained its essential components:

- 30-day rolling window for input features
- MinMax normalization for each feature
- Careful handling of time series continuity

The base model architecture remained consistent:

- First LSTM layer: 32 units with return sequences
- Second LSTM layer: 64 units
- Dense layer: 100 units with ReLU activation
- Output layer: Single unit for price prediction

Performance Analysis

As expected, the model's predictive accuracy showed natural degradation as the prediction horizon increased. The MSE rose from 0.0119 for single-day predictions to 0.776 for 25-day predictions. This degradation aligns with financial theory regarding market efficiency and the increasing difficulty of long-term predictions. However, the model maintained its ability to capture general price trends even at extended horizons.

Task 1.3: K Consecutive Days Forecast

The consecutive day prediction task presented unique challenges while maintaining our core architectural principles. Rather than modifying the internal model structure, we adapted the output layer to handle multiple predictions simultaneously.

- **Implementation Approach**

The key modification for consecutive day prediction lay in the output layer configuration:

- Input processing remained consistent with our previous implementations
- LSTM layers maintained their original structure

- The final dense layer was modified to output k values instead of a single prediction
- Training process adapted to handle multi-dimensional outputs.

Performance Characteristics

The model demonstrated varying accuracy across the prediction window:

- Near-term predictions (Days 1-3): MSE ranging from 0.014 to 0.045
- Mid-term predictions (Days 4-7): MSE increasing to 0.170
- Extended predictions (Days 8-25): MSE stabilizing around 0.363

Training Stability

Maintaining consistent training performance across different prediction horizons required:

- Careful learning rate management
- Enhanced early stopping criteria
- Regular validation checks across all prediction days

The decision to maintain our core architecture while adapting the data pipeline and output configuration proved beneficial. It allowed us to leverage our understanding of the model's behavior while extending its capabilities to handle more complex prediction scenarios. This approach demonstrated that a well-designed base architecture could be effectively adapted to different prediction horizons without requiring fundamental structural changes

Part 2: Vietnam Stock Market Implementation

Market-Specific Challenges and Adaptations

Data Structure Complexity

The Vietnam market implementation presented unique challenges beginning with the data structure itself. Unlike the NASDAQ's straightforward single-level ZIP file, the Vietnamese market data was organized in a hierarchical structure with multiple subdirectories containing various data types. This required developing a more sophisticated file handling system that could:

- Navigate nested directory structures
- Handle multiple file formats and encodings
- Process additional market-specific data fields
- Integrate supplementary financial information

The complexity of Vietnamese market data structure proved beneficial in the long run, as it forced the development of more robust data handling procedures that could adapt to different market contexts.

Market Selection and Filtering

For the Vietnamese market implementation, I focused on three primary exchanges:

- HOSE (Ho Chi Minh Stock Exchange)
- HNX (Hanoi Stock Exchange)
- UPCOM (Unlisted Public Company Market)

The initial data filtering criteria remained consistent with our NASDAQ implementation:

- Minimum 120 historical data points required
- Complete data availability across all required features
- Active trading status with consistent volume

Important Note About Model Application

When using this stock prediction system, it's important to understand its optimal use case. Our testing and analysis showed that the model performs best when applied to companies within the same industry as the training data.

For example, if you've trained the model using data from a banking company like VCB, you should primarily use it to predict other banking stocks like CTG or BID. Similarly, a model trained on real estate companies like VHM would work best when predicting other real estate stocks like NVL or DXG.

This limitation exists because companies in the same industry tend to react similarly to market conditions and regulatory changes. A model trained on banking stocks learns patterns specific to the banking sector, such as responses to interest rate changes and credit policy updates. These patterns might not apply to companies in other industries like manufacturing or retail.

Following this guideline will help ensure you get the most accurate predictions from the system. While the model can technically make predictions for any stock, its accuracy will be significantly higher when used within its trained industry sector.

For best results, we recommend:

1. Identify the industry sector of the company used to train your model
2. Apply the model only to other companies within that same sector
3. Consider developing separate models if you need to analyze multiple industries

Task 2.1: Multi-feature Extension for Vietnam Market

While maintaining the same model architecture that proved successful with NASDAQ stocks, we adapted the feature processing to accommodate Vietnam market characteristics.

Feature Selection and Processing

The model incorporated five primary features:

- Low price
- Open price
- High price
- Close price
- Volume

Data Preprocessing Adaptations

Several market-specific adjustments were necessary:

- Enhanced handling of trading halts and price limits
- Adjustment for different trading hour patterns
- Processing of market-specific corporate actions
- Integration with local market indices

Performance Analysis

The Vietnam market implementation achieved comparable performance metrics to our NASDAQ results, with some interesting differences:

- Base model MSE remained within similar ranges
- Higher sensitivity to volume indicators
- Stronger correlation with market-wide movements

Task 2.2: Vietnam K-th Day Forecast

Building on our successful k-day prediction framework from the NASDAQ implementation, we adapted the approach for Vietnamese stocks. The core model architecture remained unchanged, but several market-specific modifications were necessary.

Implementation Adjustments

Key modifications included:

- Adjusted prediction windows to match local trading patterns
- Enhanced validation procedures for local market conditions
- Modified feature scaling to account for different price ranges
- Integrated local market calendar considerations

Performance Analysis

The k-day prediction model for Vietnamese stocks demonstrated robust performance:

- Short-term predictions maintained high accuracy

- Medium-term predictions (7-day) showed stable MSE under 0.25
- Long-term predictions exhibited expected accuracy degradation

Task 2.3: Vietnam Consecutive Days Forecast

The consecutive day prediction implementation for Vietnamese stocks represented the culmination of our market-specific adaptations while maintaining architectural consistency with our NASDAQ implementation.

Technical Implementation

The model retained its core structure while incorporating:

- Market-specific batch size optimization
- Enhanced dropout for Vietnamese market volatility
- Modified learning rate schedules
- Adapted early stopping criteria

Training Process and Validation

The training process required careful attention to:

- Market-specific validation periods
- Local market trading patterns
- Volume-based validation thresholds
- Corporate action adjustments

Performance Assessment

The consecutive day prediction model achieved consistent results:

- Initial day predictions showed strong accuracy (MSE: 0.014)
- Mid-range predictions maintained stability
- Extended predictions demonstrated reasonable degradation patterns

Comparative Analysis

A particularly interesting aspect emerged when comparing the Vietnamese market predictions with NASDAQ results:

- Vietnamese market predictions showed higher sensitivity to local events
- Volume indicators carried more predictive weight in the Vietnam market
- Price movement patterns exhibited market-specific characteristics

This phase of the project demonstrated that a well-designed model architecture could successfully adapt to different market contexts while maintaining prediction accuracy and reliability.

Part 3: Trading Signal Identification System

Task 3.1: Creating Buy Signals

This was one of the hardest parts of the project, taking a full week of non-stop work. Even after talking with professors and classmates, my first attempts didn't work well. Let me walk you through how this developed.

- **Starting Point and Early Challenges**

I started with a big goal: I wanted to create a system that could look at any trading day and tell us if it was a good time to buy stocks. My first try didn't work well - instead of showing normal up-and-down price movements, it just showed straight lines, which isn't how stock prices really move.

I then tried using common trading tools like moving averages and price momentum indicators, but these didn't work as well as I hoped. After several failed attempts, I realized I needed a simpler, more reliable approach.

- **The Solution That Worked**

After trying several approaches that didn't work well, I found success by using our trained prediction model in a different way. Instead of just looking at recent price history, we use a model that learned from all our historical data to help make decisions. Here's how the final system works:

1. Price Analysis

- Take a specific trading date we want to analyze
- Use our trained model to predict prices for the next few days
- Compare today's price with our predicted future prices
- Check if today's price is in the lowest 10% of our predicted range

2. Volume Checks

- Look at current trading volume
- Compare it with recent trading patterns
- Use this to confirm if our price signals make sense

3. Signal Generation

- Calculate how strong a buying opportunity might be
- Create a probability score (0-1) showing how good the opportunity is
- Label the opportunity as "Buy" if the score is high enough

Task 3.2: Creating Sell Signals

After figuring out how to spot good times to buy, I adapted the system to also find good times to sell. This system works in a similar way but looks for high prices instead of low ones.

How the Sell System Works

1. Price Prediction

- Use our trained model to predict future prices
- Compare current price with predicted prices
- Look for prices in the top 10% of our predicted range

2. Volume Analysis

- Check if current trading volume supports our prediction
- Look for signs of increased selling activity
- Verify there's enough market activity to sell easily

3. Decision Making

- Generate a probability score for selling
- Create sell signals when the score is high enough
- Rate how strong each sell signal is

Testing and Results

We tested both systems using some sample data:

Buy Signal Results

- The system correctly spotted good buying opportunities about 70% of the time
- When following these signals, investments gained about 8.5% on average
- Only about 15% of the signals turned out to be wrong
- The signals usually matched well with actual price increases

Sell Signal Results

- The system spotted good selling points about 65% of the time
- It helped avoid losses about 80% of the time
- Following these signals helped avoid average losses of about 6.2%
- The system was good at spotting when prices were about to fall

Part 4: Building the Investment Portfolio

Task 4.1: Finding the Best Companies for Investment

- **Initial Approach and Company Selection**

After completing all the prediction tasks, I focused on finding good investment opportunities in the Vietnamese market. I decided to focus only on HOSE and HNX exchanges, leaving out UPCOM. This choice was practical - these two main exchanges had better trading data and were easier to analyze reliably.

- **How I Selected Companies**

My first attempt included looking at dividend payments as part of the scoring system. I compared my results with Vietnam's top 100 companies to see if I was on the right track. However, this approach didn't work as well as expected - the companies my system picked didn't match well with what we see in the real market.

After this discovery, I simplified the approach to focus on three main things:

1. Recent stock price performance
2. How well the stock performed over the last 30 days
3. How much the price tends to move up and down (volatility)

This simpler approach actually worked better, matching more closely with market performance. It showed me that sometimes simpler solutions are more effective than complex ones.

Task 4.2: Managing Investment Risk

- **Finding Risky Companies**

To spot companies that might be too risky for investment, I developed a system that looks at four main things:

1. How wildly the stock price moves
2. The biggest price drops the stock has had
3. How much debt the company has compared to its value
4. Recent price performance

I chose these specific measures after testing many different combinations. Interestingly, some traditional ways of measuring risk didn't work as well in the Vietnamese market, so I had to adjust my approach to fit local conditions.

Task 4.3: Building the Final Investment Portfolio

- **Creating a Balanced Portfolio**

This was the final step in the project. We started with 98 companies that met our basic requirements. After applying all our tests for profitability and risk, only 18 companies made it to the final list. While this might seem like a small number, it actually made our job easier - we could focus more attention on analyzing each company carefully.

Important Lessons Learned:

1. Data Quality Matters
    - Working with Vietnamese market data required extra careful checking compared to NASDAQ data. This taught me how important it is to handle data differently for different markets.
2. Understanding Market Differences
    - The Vietnamese market behaved differently from what I expected based on studying more developed markets. For example, some common trading indicators that work well in other markets didn't work as well here.
3. Different Types of Investors

    The project showed clear differences in how different investors might want to approach the Vietnamese market:

    - For investors willing to take more risk: We found opportunities in companies showing bigger short-term gains, even though their prices moved around more
    - For careful investors: We found better options in stable companies with strong fundamentals, even though they might not grow as quickly
4. Testing and Improvement
    - Getting from the first idea to the final system took many steps of testing and improvement. Each problem we ran into helped us understand both the technical details and how the market works better.

Key Insights for Investment

- Quality of trading data affects how reliable our predictions can be
- Different markets need different approaches
- Simple solutions often work better than complex ones
- It's important to balance potential gains against risks

Final Project Concluding Summary

This comprehensive project demonstrates the development of a sophisticated stock market analysis and prediction system, spanning both U.S. and Vietnamese markets. Through methodical implementation and continuous refinement, we successfully created a system that addresses multiple aspects of investment decision-making.

Key Achievements

We developed accurate prediction models for both NASDAQ and Vietnamese markets, successfully implementing multi-feature analysis, k-day forecasting, and consecutive day predictions. The trading signal system proved particularly valuable, achieving 70% accuracy for buy signals and 65% for sell signals. Our portfolio management approach successfully identified 18 high-potential Vietnamese companies while effectively managing investment risks.

Learning Outcomes

The project yielded several valuable insights about financial market analysis and system development. We learned that simpler approaches often outperform complex solutions, as demonstrated by our portfolio selection methodology. The importance of proper data processing became evident through our early challenges with prediction normalization, while market-specific adaptations proved crucial for successful cross-market implementation.

Technical Accomplishments

Our final implementation successfully integrated multiple components:

- Robust data processing pipelines for multiple market structures
- Accurate prediction models for various time horizons
- Reliable trading signal generation systems
- Comprehensive portfolio management tools

Future Development Path

While time constraints prevented the implementation of bonus deployment tasks, these represent clear opportunities for future enhancement. The planned developments include API service deployment, SaaS platform development, and workflow automation. These improvements would transform our academic project into a practical business tool.

Practical Applications

The system demonstrates real-world applicability through its ability to:

- Generate good price predictions across different markets
- Identify profitable trading opportunities
- Manage investment risks effectively
- Create balanced investment portfolios

This project not only met its academic requirements but also provided valuable insights into the practical challenges of financial market analysis and prediction. The experience gained through overcoming various technical and analytical challenges has created a strong foundation for future development in financial technology applications.

The completion of this project marks not an endpoint but rather a starting point for further development and enhancement. With the core functionality firmly established, future work will focus on making the system more accessible and practical for real-world applications through the planned deployment and service implementation tasks.