**Reproducibility Summary Report: Forecasting Directional Movements of Stock Prices Using LSTM and Random Forests**

**Project Overview** This document outlines the process and challenges of reproducing the results from the research paper titled *"Forecasting directional movements of stock prices for intraday trading using LSTM and random forests" (arXiv:2004.10178v2)*. The original study proposed a trading strategy leveraging Random Forest and CuDNNLSTM models to predict intraday directional stock movements using S&P 500 historical data.

**Steps Taken to Reproduce the Study**

1. **Literature Review & Objective Definition**
   * Understood the original study’s goals: to compare model performance on intraday return prediction.
   * Focused on reproducing the Random Forest component using public tools and data.
2. **Code Migration and Environment Setup**
   * Adapted original codebase (which relied on proprietary Bloomberg data and an SPX constituents file) to use public data from Yahoo Finance (yfinance).
   * Implemented Random Forest pipeline for training and prediction.
3. **Data Acquisition and Preparation**
   * Attempted to replace SPXconst.csv with approximated static ticker lists.
   * Used yfinance to download historical Open and Adjusted Close prices.
   * Created training and test datasets using rolling 3-year windows.
4. **Model Execution and Evaluation**
   * Trained Random Forest models for each year (2015–2019).
   * Simulated a simple long-short trading strategy to evaluate predictions.
   * Measured output: daily return averages.
5. **Code Adjustments for Compatibility**
   * Rewrote deprecated or missing functions.
   * Replaced unavailable Statistics class with placeholder metrics.
   * Handled missing and incomplete ticker data.
   * Improved error handling and modularization.

**Comparison with Original Results**

| **Metric** | **Original Study (RF)** | **Reproduced (RF, partial)** |
| --- | --- | --- |
| Daily Return (%) | ~0.54 | Not replicable (yet) |
| Sharpe Ratio | ~5.20 | Not replicable (yet) |
| Max Drawdown | ~19.7% | Not replicable |

* **Observation**: Key results were not replicable due to data constraints and model environment differences.

**Discrepancies and Missing Information**

* **SPXconst.csv**: Crucial to replicate accurate stock universe; not publicly available.
* **Bloomberg Data**: Proprietary intraday historical data unavailable; substituted with daily yfinance.
* **Statistics Module**: Custom class used to summarize strategy performance missing.
* **Exact LSTM Model**: CuDNNLSTM requires GPU and specific preprocessing that is not trivial to recreate.

**Sources of Variation**

* **Data Granularity**: Original used intraday data; reproduction used daily Open/Close.
* **Constituents Filter**: Study filtered stocks monthly; reproduction used static ticker sets.
* **Feature Engineering**: Manually engineered features differ from those in original paper.
* **Time Periods**: Full test periods (1993–2018) in original; reproduction focused on 2018–2019.

**Reproducibility Challenges**

* **Impact**: Core results could not be validated; model effectiveness unverified.
* **Resource Limitations**: No access to Bloomberg, CuDNNLSTM training environment, or custom code libraries.
* **Scientific Communication**: Original paper lacked full code and data-sharing; made assumptions hard to interpret.
* **Stakeholder Implications**: Investors and researchers relying on reproducible results cannot trust model validity without full transparency.

**Recommendations**

1. **Data Availability**: Authors should provide open alternatives or sample datasets.
2. **Code Transparency**: Include full training, preprocessing, and evaluation scripts.
3. **Documentation**: Describe every data input, label method, and hyperparameter clearly.
4. **Ethical Disclosure**: If models are tied to financial performance, clearly state risks and limitations.
5. **Stakeholder Communication**: Ensure findings are accessible to non-technical decision makers.

**Ethical Considerations**

* **Bias and Misuse**: Financial models without proper validation may mislead or disadvantage certain investors.
* **Transparency**: Results should be replicable to maintain scientific integrity.
* **Responsibility**: Researchers must ensure that their claims are verifiable and not solely dependent on proprietary infrastructure.

**Conclusion** While the full replication of the study’s results was not achievable using public resources, this exercise highlights the importance of open data, thorough documentation, and transparent methodology in financial machine learning research.

Future work could involve sourcing or approximating constituent datasets more accurately and integrating intraday data streams if available.