Group 3: Clay Skiles, Seth Phillips, Yousaf Khaliq, John Allard

Project: Reproducibility

Paper Title: *"Forecasting directional movements of stock prices for intraday trading using LSTM and random forests" (arXiv:2004.10178v2)*

Paper Github*:* *https://github.com/pushpendughosh/Stock-market-forecasting/tree/master?tab=readme-ov-file*

Clay Skiles: Intraday-240,3-LSTM.py

Verifying Methodologies:

Verifying the methodologies of the research procedure provided by the authors proved to come to a standstill. The biggest challenge regarding this aspect served to be the lack of a complete dataset for stock prices from 1993-2018 from Bloomberg (as the authors state was their dataset of choice) to be used in the re-creation of the results without a massive paywall. For this particular example, a public Kaggle dataset with stock prices from the S&P500 from 1970-2018 was substituted to observe the logic of the methods, along with ChatGPT assisting in creation of the code.

In this verification procedure of the project, the Kaggle dataset was processed in the following ways:

- An S&P500 list was imported to filter only stock entries inside the Kaggle dataset which lie in the S&P500 index.
- Entries before 1990 were eliminated
- Open and closed stock prices data frames were created to mimic that of the provided dataframes on Kaggle with column labeling syntax

What is left are df_open and df_close datasets with 7220 daily entries for stock prices across 458 different stocks within the S&P500.

The first step is to structure the data such that there are no overlapping testing periods. The paper specifies that there should be 29 entire years-worth of data, and this data would be divided with a 4-year window and 1-year stride where study periods are divided into 3 years with trading periods divided into 1 year. This process was accomplished using ChatGPT, however missing days in the dataset meant that 25 study periods instead of 26 were created.

The next procedure as stated by the authors is to select features. The authors refer to the following process for selecting features:

“Given a prediction day $t := \tau$, we have the following inputs and prediction tasks:

Input: We have the historical opening prices, $op_t^{(s)}$, $t \in \{0, 1, \ldots, \tau-1, \tau\}$, (including the prediction day’s opening price $op_t^{(s)}$) as well as the historical adjusted closing prices $cp_t^{(s)}$, $t \in \{0, 1, \ldots, \tau-1\}$, (excluding the prediction day’s closing price, $cp_t^{(s)}$).

Task: Out of all n stocks, predict k stocks with the highest and k stocks with the lowest intraday return $ir_{\tau,0} := \frac{cp_t^{(s)}}{op_t^{(s)}} - 1$.”

Selecting k=10 allows the retrieval of the ten stocks with highest and lowest intraday returns as designated by this methodology. This would be the last step within the process which would be replicated without any discrepancies via the Kaggle dataset.

The step the Kaggle dataset runs into difficulties when it comes to scratch replication and reproducibility is in feature generation where "in a multi-feature setting rather than their single feature approach, we input the model with 240 timesteps and 3 features, and train it to predict the direction of the 241st intraday return." This step with producing the direction of the 241st intraday return proved impossible to replicate for the following reason as there was an issue with robust standardization not detecting 240 timestamps, which leads to input and target shapes which are empty, leading to no training samples available for modeling the LSTM. This could be due to many companies in the current S&P500 being created post-1990 where the feature generation is unable to detect 240 consistent timesteps across the board.

Therefore, the actual process given a dataset containing missing daily entries is un-reproducible. If the full Bloomberg dataset were included in the GitHub, perhaps this paper could be replicated from scratch given the amount of references it possesses in future works.

Verifying Intraday-3 LSTM Code:

The Intraday LSTM-3 code provided by the authors is capable of being reproducible with a dataset in a similar format as that of the Bloomberg dataset listed on GitHub. Considering the fact the full 29-year stock prices dataset is not available without a paywall for public use, a subset of 1990-1991 stock data was used in its place to ensure the algorithm published by the authors is accurate.

- The Close-1990, Open-1990, and SPXconst datasets were loaded, removing nulls, cleaned and parsed, also setting correct indices. (This is all dummy data, not real property and values formulated by Bloomberg themselves)
- The LSTM model, the methodology for creating training and test sets by year, as well as the Robust scaler method were defined as functions.
- The model was trained along with recording history for future time series predictions.
- Statistics were taken to compare with final paper (even though results won't match due to lack of access to 1990-2018 stock price dataframe)

In the end, a result for mean average returns prior to transaction charges came in at 1.2% for stock prices between 1990-1991. This result is sensical as the paper addresses said value between 1990-2018 is .64%, therefore the code provided by the authors can be trusted to reproduce a desired result if the dataframe is formatted similarly to the Bloomberg stock prices data.

One should be careful though of how to approach citing this code in their own work: since the code is optimized for Bloomberg stock price datasets, there exists a concern when it comes to the redistribution of the dataset. The original authors only published dummy data on their Github to avoid infringement of illegal property reproduction as Bloomberg asks for a large monetary compensation for their complete dataset used in the code. If one uses this code in their own work, a similar approach to only publishing dummy data may for proof for reproducibility is wise to avoid misconceptions of who truly owns the data.

Stock prices, in general, are information in the public domain, but historical stock data is not public property.

Seth Phillips: Intraday-240,3-RF.py

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.

John Allard: NextDay-240,1-RF.py

First Attempt and Initial Observations

Upon first running the code, several key issues became immediately apparent. First, the code successfully generated results for the year 1993 but failed to print any predictions for the remaining data. This suggests a fundamental issue that could reside in multiple areas (how the training and test datasets are constructed, where the code looks for data, how the metrics are calculated). Additionally, the dataset included with the project files was not the one originally used in the published study. The original research relied on a proprietary Bloomberg dataset costing $500, however, the project provided a dataset containing constituent stock data that looks ambiguous and uninterpretable when performing EDA. This immediately created doubt in the accuracy of any results I could produce, and in whether the same features and financial indicators were used to produce each respective dataset. These initial observations highlight the broader challenges of reproducibility in data science and serve as a case study in the ethical implications of inaccessible datasets, which will be explored further in this paper.

Project Overview

This project aims to assess the reproducibility of a stock prediction model originally developed using a random forest algorithm. The original study analyzed stock performance using historical data spanning from 1990 to 2018, with predictions beginning in 1993 and extending to 2019. The model utilized the previous three years of stock data to generate forecasts for the subsequent year. The goal of this replication study was to recreate these results using the dataset provided in the project files. However, several challenges emerged, ranging from discrepancies in the dataset to difficulties in adjusting the test-year processing code.

Challenges in Data Availability and Structure

A major barrier to replication was the mismatch between the dataset provided and the dataset apparently used in the original study. The original implementation expected multiple CSV files named in the format "data/Close-{testyear}.csv" to represent stock price data for individual years. However, the project files only included a single dataset, `data/SPXconst.csv`. Lacking the original dataset, the code has to be modified to iterate of the provided dataset accurately, and my results are most likely guaranteed not to match those of the original study.

Furthermore, the dataset did not contain traditional stock price data. Instead, it contained stock constituent information, formatted as strings representing stock tickers (e.g., "S814", "S482", "S1221"). This means that instead of capturing numerical stock performance indicators like closing prices or returns, the dataset tracks which stocks were present in an index at any given time. The original study, by contrast, appears to have used a proprietary Bloomberg dataset, which costs approximately $500 to access. This restriction makes an exact replication of the original study impossible without incurring significant financial costs.

Ethical and Reproducibility Concerns

The reliance on proprietary data raises critical ethical and reproducibility concerns. Scientific integrity depends on the ability of researchers to validate and verify results using publicly accessible resources. The unavailability of the Bloomberg dataset means that any reproduction effort requires modifying the methodology, introducing additional sources of uncertainty and deviation from the original results. This issue highlights a broader ethical concern in financial research: studies that use paywalled datasets inherently limit the transparency and reliability of their findings.

Issues in Test-Year Data Processing

When adjusting the test-year code block to accommodate the available dataset, another significant challenge arose: the inability to generate stock predictions despite having data stored in both the training (`df_train`) and test (`df_test`) datasets. The primary function responsible for creating training and testing datasets, `create_stock_data()`, was modified to work with the single available CSV file. However, when integrated into the main test-year processing loop, no stock predictions were produced for any year.
Upon further examination, the issue likely stems from the data concatenation section of the code. While all expected columns for each year were present in the training dataset, none appeared in the test dataset, resulting in an inability to generate predictions. Despite debugging efforts, the exact root cause of this failure could not be conclusively determined. This highlights another aspect of the reproducibility challenge: even when code is available, modifications necessitated by data inconsistencies can lead to unexpected errors that prevent successful execution.

Implications for Financial Research and Best Practices

This project underscores several key lessons regarding reproducibility in financial research:

1. **The necessity of publicly available datasets:** Studies that rely on proprietary data limit the ability of independent researchers to verify their findings. Providing at least a representative sample of the data used would improve reproducibility.
2. **Detailed documentation of preprocessing steps:** The provided code did not include clear documentation on how the dataset was originally structured or processed. More transparency in data handling would facilitate replication efforts.

3. **Robustness to data format changes:** A reproducible study should be flexible enough to accommodate minor deviations in dataset format without requiring extensive modifications to the codebase.

In conclusion, the inability to access the original dataset, coupled with issues in test-year data processing, rendered a full replication of the stock prediction study infeasible. This case serves as a cautionary example of the challenges in financial research reproducibility and highlights the need for more open-access data and clearer methodological documentation.

Yousaf Khaliq: NextDay-240,1-LSTM.py

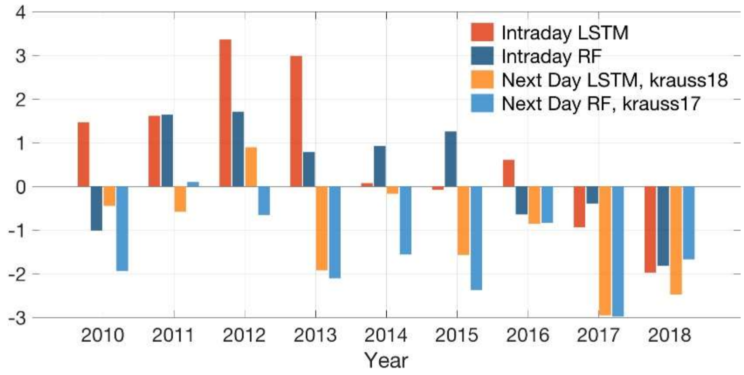
Reproduction process: I skimmed through the paper to get an idea of it's goals, problems it's trying to solve, methodology, and the kinds of results it conveyed. The paper had 6 different scripts and I chose the NextDay-240,1-LSTM.py script. I downloaded the script from github and immediately tried to run it.

After clearing up all deprecated code (and removing the GPU based LSTM the code used and replacing it with a CPU based, normal LSTM) I ran the script with the data provided in the github. That data proved completely insufficient, in fact, it was dummy data. The paper obtained data from Bloomberg which had a significant paywall. So I then obtained all the necessary data from the Yahoo Finance package.

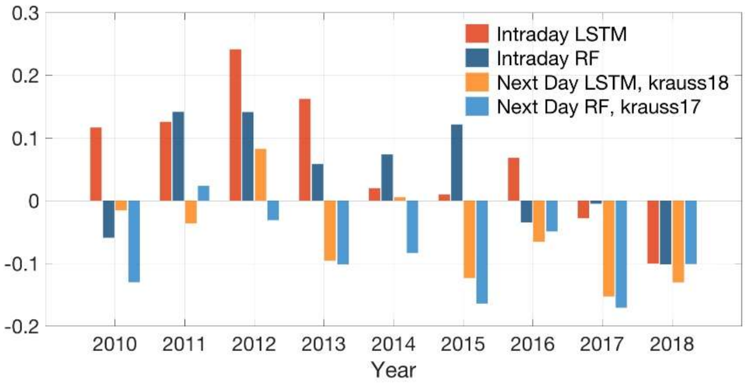
My results: The paper analyses all 500 companies stock price action over the course of 28 years. We simply did not have the time to process that much data so I pulled only 5 companies and used that data. I then attempted to recreate the visualizations and results that the paper had, as seen below:

<u>ORIGINAL RESULTS:</u>

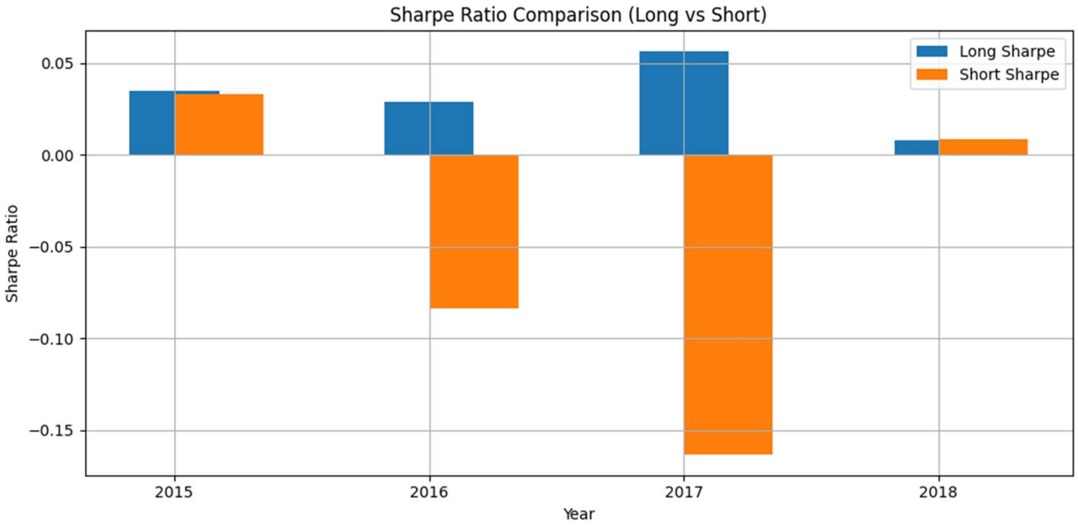
Average Sharpe Ratio

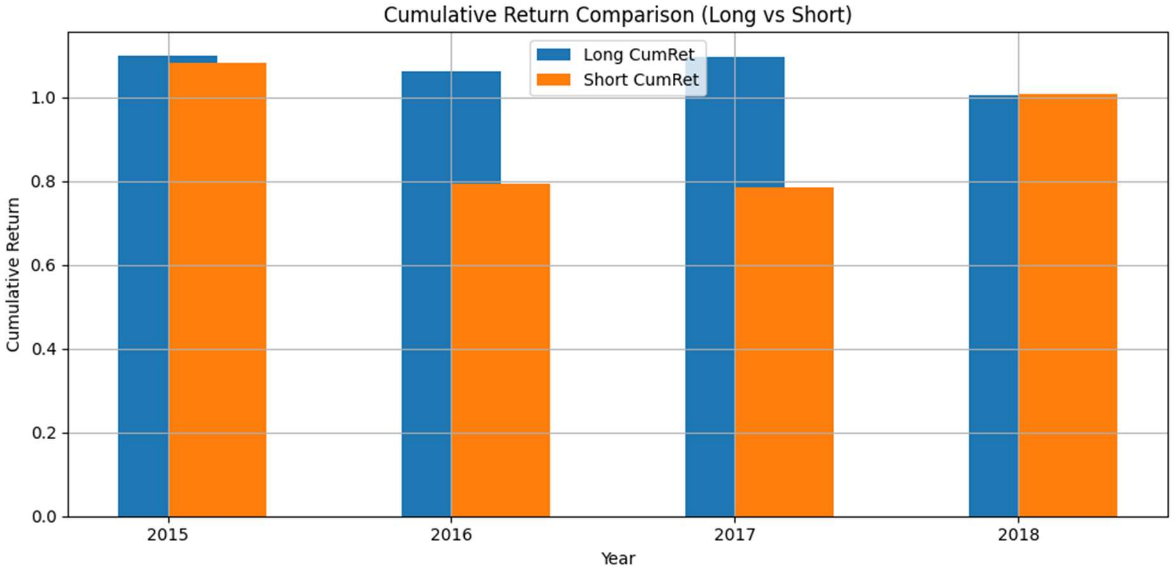


Cumulative Daily Returns



<u>MY RESULTS:</u>





Comparison:

I was only able to run the code for the last 4 years in the data set. The only results that correlate with the original paper's are the short position sharpe ratio, both showing negative for the NextDay LSTM. All other results are basically night and day.

Ethical Analysis:

This paper claims high accuracy and predictive power of stock price movement direction. This could very easily influence any investor, most likely retail investors, to make risky financial decisions.

Amongst the scientific community, a paper like this could slow down the progress of other research teams in that they would waste time trying to reproduce the results and possibly reach a dead end.

Without reproducibility, a paper loses credibility and raises concerns about data leakage or even fabrication.