Modeling Stock Market Returns using Deep Learning

Devansh Gupta* (2019160)
Divyansh Rastogi* (2019464)
Rupanshu Yadav* (2019475)

All authors denoted by * contributed equally, the ordering of naming is alphabetical.

Problem Statement

More than five trillion dollars worth of stocks are traded every single day. Investors can maximize their profits by investing rationally based on technical analysis. The time series prediction problem is complex due to the widely accepted semi-strong efficient market hypothesis and their volatile & non-stationary structure. Although these financial markets are shown to be predictable to a certain extent.

Stock market analysis methods fall into two categories:

1. Technical analysis involves the prediction of stock prices using historical data, primarily price and volume.
2. Fundamental analysis makes predictions based on the intrinsic value of the investment by studying macroeconomic features, industry features, and company's features, including financial reports, competitors, & viability.

Statistical models such as ARIMA, STAR are widely used for time series prediction purely based on the historical price of a stock. Machine learning has been intensively researched for its use in financial market forecasting where Decision trees, Xgboost, & Naïve Bayes all have shown to be beneficial in stock market price prediction. Deep learning methods & frameworks have also proven to be helpful in time series analysis with its ability to model complex non-linear relationships.

Related Work

- Volatility & non-stationary nature of stock markets
- Partial predictability of financial markets
- Predictive relationships of numerous financial & economic variables
- Statistical & Econometric ML methods:
    - ARIMA
    - SVM
    - Boosted Regression Trees
    - Logistic Regression
- Feature Selection & Extraction / Data Transformation Methods:
    - Genetic Algorithms + ANN / SVM
    - Autoencoders + DNN / SVM
    - PCA + ANN / DNN
- Deep CNN’s
- Deep CNN’s + LSTM

Dataset

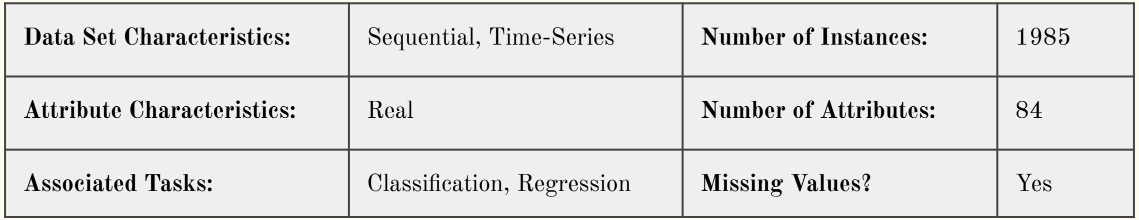
Dataset Description

The dataset contains several daily features of S&P 500, NASDAQ Composite, Dow Jones Industrial Average, RUSSELL 2000, and NYSE Composite from 2010 to 2017.

The dataset covers features from various categories of technical indicators, futures contracts, price of commodities, important indices of markets around the world, price of major companies in the U.S. market, and treasury bill rates.

Relevant papers:

- ❖ *CNNpred: CNN-based stock market prediction using a diverse set of variables.*
- ❖ *U-CNNpred: A Universal CNN-based Predictor for Stock Markets*



Data Set Characteristics:	Sequential, Time-Series	**Number of Instances:**	1985
Attribute Characteristics:	Real	**Number of Attributes:**	84
Associated Tasks:	Classification, Regression	**Missing Values?**	Yes

https://archive.ics.uci.edu/ml/datasets/CNNpred%3A+CNN-based+stock+market+prediction+using+a+diverse+set+of+variables

Baselines

ARIMA

1. ARIMA is widely used for prediction in time series data modeling.
2. ARMA requires the time series to be stationary; which means a constant mean, constant variance and non-seasonal.
3. Auto-Regressive(AR) and Moving Average(MA) models are parameterized by auto-correlation coefficient (p) and Partial autocorrelation coefficient (q).
4. A search is performed given the maximum values of p and q to find the best model.
5. An RMSE of 423.24 was observed. However, if the first testing point is included in the training data to predict the next point an RMSE of 49.72 was observed.

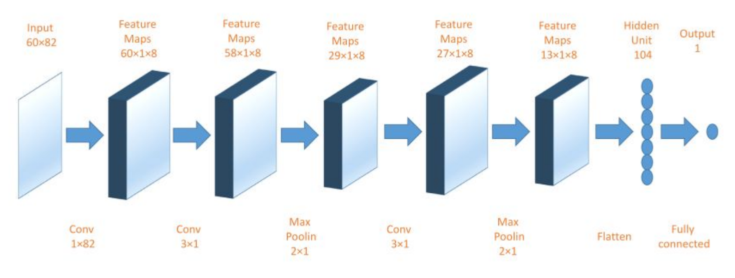
PCA + ANN

1. Forecasting daily return direction of the SPDR S&P 500 ETF index.
2. Study's model is extrapolated to the CNNPred dataset's S&P market based on resemblance of features.
3. Traditional PCA outperforms all non linear data transformation techniques on real world data.
4. PCA-represented dataset with 82 principal components is used.
5. Data Preprocessing:
   a. Outlier detection based on interquartile ranges
   b. Data standardization
   c. Train/Validation/Test ratio: 70/15/15.
6. PCA-represented dataset is classified using a ANN network comprising of 4 layers with RELU activation & Sigmoid activation for the last layer.
7. Binary cross entropy loss is used as loss criterion.
8. ADAM Optimizer with LR=10^{-4} is used over maximum of 100 epochs.
9. Early stopping is implemented with use of validation set.

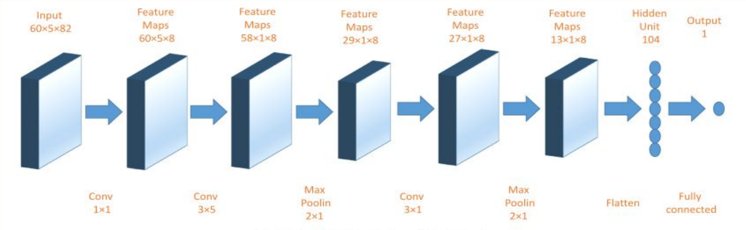
https://jfin-swufe.springeropen.com/articles/10.1186/s40854-019-0138-0

CNNPred

- CNNPred 2D
  - Using data from a single market to predict the trend of the future prices using 2D convolutional filters



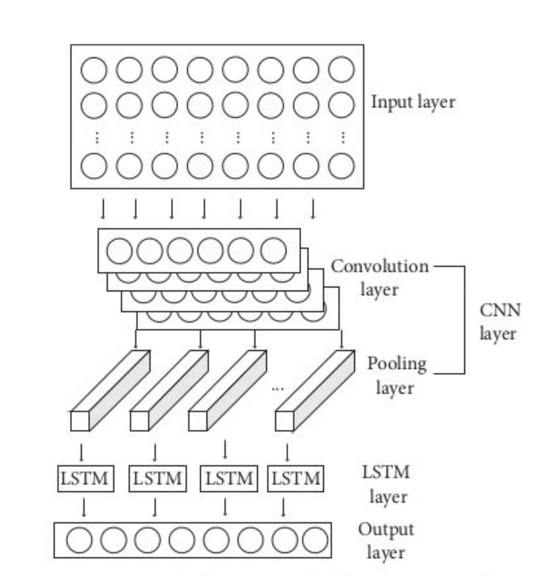
- CNNPred 3D
  - Using data from a multiple markets to predict the trend of the future prices using 3D convolutional filters to capture data across markets

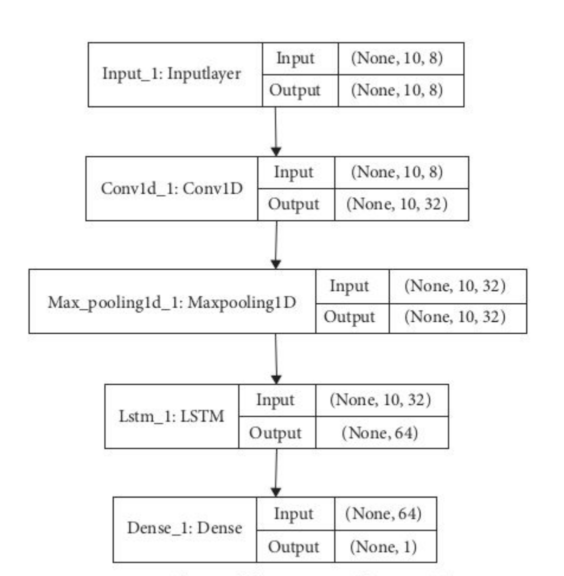


https://www.sciencedirect.com/science/article/abs/pii/S0957417419301915

CNN + LSTM

- Considers learning sequence aware features as the stock market
- Explored the price prediction perspective
    - Regression viewpoint
- Initially extracts features from input data using CNN layer and then uses an LSTM to model the time-series data





Layer		Shape
Input_1: Inputlayer	Input	(None, 10, 8)
	Output	(None, 10, 8)
Conv1d_1: Conv1D	Input	(None, 10, 8)
	Output	(None, 10, 32)
Max_pooling1d_1: Maxpooling1D	Input	(None, 10, 32)
	Output	(None, 10, 32)
Lstm_1: LSTM	Input	(None, 10, 32)
	Output	(None, 64)
Dense_1: Dense	Input	(None, 64)
	Output	(None, 1)

https://www.hindawi.com/journals/complexity/2020/6622927/

Future Work

Future Work

- Dynamics learning through Neural ODEs
  - Motivation: Stock Market Prices tend to follow the equations of Brownian Motion
- Generative Adversarial Networks for Time Series Data
- Graph Attention Networks or Hypergraph Convolutional Networks
  - Use company relations as an additional form of supervision to build graphs or hypergraphs