18XDAJ DEEP LEARNING ASSIGNMENT PRESENTATION

Long Term Stock Prediction using Financial Statements

K V Pragathi and Shreya Ganeshe

19PD17 and 19PD12
Department of Applied Mathematics and Computational Sciences, PSG College of Technology

March 20, 2023

Table of Contents

INTRODUCTION

- Problems based on stock market prediction are on rise
- Many problems focus on short term predictions
- Focuses on building an end-to-end **LSTM** model for long term stock prediction based on historical financial statements

Dataset and Features

- Dataset - https://github.com/sugia/tradeX/tree/main/data. The dataset contains about 40000 datapoints.
- Balance sheet features - Includes features such as total current liabilities, long-term debt, other liabilities, deferred liability charges, misc. stocks, minority interest, total liabilities, common stocks, capital surplus, retained earnings, treasury stock, other equity, total equity, total liabilities and equity.
- Income statement features - total revenue, cost of revenue, gross profit, research and development, sales general and admin., non-recurring items, other operating items, operating income, add'l income/expense items, earnings before interest and tax, interest expense, earnings before tax, income tax, minority interest, equity earnings/loss unconsolidated subsidiary, net income-cont. operations, net income, net income applicable to common shareholders.
- Cash flow statement featues - net income, depreciation, net income adjustments, accounts receivable, changes in inventories, other operating activities, liabilities, net cash flow-operating, capital expenditures, investments, other investing activities, net cash flows-investing, sale and purchase of stock, net borrowings, other financing activities, net cash flows-financing, effect of exchange rate, net cash flow

Methods

Steps applied in data preprocessing stage:

- To construct the input of a LSTM model, 3 continuous data points are connected to be 1 training item, with 30 features from balance sheet, 18 features from income statement, 18 features from cash flow statement, for a total of 66 features in each data point.
- Training labels fall into five categories based on annual stock price percentage change one year after the last financial statement release date for an item. Those five categories are:
  - Annual percentage change smaller than -50%
  - Annual percentage change greater or equals to -50% and smaller than 0%
  - Annual percentage change greater or equals to 0% and smaller than 50%
  - Annual percentage change greater or equals to 50% and smaller than 100%
  - Annual percentage change greater or equals to 100%
- Two data augmentation techniques are applied:
  - One is to generate more training data by adding a small percentage change (from -5% to 5%) to each field in one item.
  - The other method is to ease out fields from one of the three financial statements (balance sheet, income statement, cash flow statement) to generate new items.
- In the training set, there are 4,135 items in the first category, 16,111 items in the second category, 16,312 items in the third category, 1,068 items in the fourth category, 374 items in the fifth category. To deal with the imbalance training data in each category, the final training set is constructed by random sampling 7,600 items from each category.

Model Construction

- The first layer is a LSTM layer with 256 nodes, with tanh as the activation function
- The second layer is a Dropout layer, with 0.1 as dropout rate
- The third layer is a Fully Connected layer, with 256 nodes, with tanh as the activation function
- The fourth layer is another Dropout layer, with 0.1 as dropout rate
- The last layer is a Softmax layer with 5 nodes