SCHOOL OF INFORMATICS AND COMPUTING

INDIANA UNIVERSITY
Bloomington

CSCI-B 565 Fall 2021 DATA MINING

FINAL PROJECT REPORT

AN EFFICIENT LEARNING MODEL FOR SHORT-TERM STOCKS PREDICTION

BY:
ABHIJEET VICHARYA (avicharya@iu.edu)
SYLVIA BODDU (sboddu@iu.edu)

Title: AN EFFICIENT LEARNING MODEL FOR SHORT-TERM STOCKS PREDICTION

Author:

Sylvia Boddu(sboddu@iu.edu)
Abhijeet Vichare(avichar@iu.edu)

Abstract:

For decades, the topic of stock prediction has been trending, and stock markets are always uncertain. People are very much interested in investing in stocks but knowing the uncertainty of the stock markets, most of them don't. Upon years there have been a lot of advancements in technology, and today we live in the era of big data, machine learning, and neural networks where there are advanced prediction models that learn and predict almost close to the real-world results. For our data mining final project, we thought of taking up the problem of stock market prediction due to its challenging nature. We have gone through multiple papers related to various stock market prediction methods and models. We got inspired by one research paper, where they used feature engineering to enhance the overall prediction accuracy of the results. After performing feature engineering and feature scaling, we train an LSTM model to make future price prediction trends.

Introduction

In this project, our goal is to build a model that would give us a short-term prediction of the stocks. We have designed our data mining model like any-other traditional Datamining project with significant steps like Preprocessing, Modeling, Visualization. Still, we have also added some advanced methods in the preprocessing to go along with our model efficiently and give us better results. To start with these special steps, we are doing Feature Engineering, followed by Recursive Feature Elimination, before performing PCA. As a result of the preprocessing, we get the top trending features ready to be fed into the model. We choose LSTM (Long Short-Term Memory) over other deep learning models because of its backpropagation techniques. As a variant neural network of RNN, even with one LSTM layer, the NN structure is still a deep neural network since it can process sequential data and memorize its hidden states through time. An LSTM layer comprises one or more LSTM units, and an LSTM unit consists of cells and gates to perform classification and prediction based on time series data.

Methods

- **Feature Engineering**

The input data from the dataset contains the general data of the stock, namely Open, High, Low, Close prices, and the volume of trade (OHLCV) of the stock for the given day. In-depth research has been done in the field by other researchers on how to extract information such as the volatility, health of the stocks, and the stocks' trend. We decided to leverage the research o and create new features from the existing data as inputs for the LSTM model.

The following are the technical indicators we created using the OHLCV data:

- Price Change: The difference in price of the current day and the previous day.
    i. $PC = (Close\ price_{now} - Close\ Price_{prev})$
- Price Change Percentage: The relative change in price from the previous day to the current day.
    i. $PC(\%) = \frac{(Close\ price_{now} - Close\ Price_{prev}) * 100}{Close\ Price_{prev}}$

- Simple moving average (SMA): The simple moving average of the previous n days. For the current project, we set the n for SMA as 10 days.
    i. $SMA = \frac{C_t + C_{t-1} + C_{t-2} + \cdots C_{t-n}}{n}$
- Moving Average Convergence Divergence (MACD): MACD is the trend following momentum indicator between two moving averages at different time periods for a stock.
- MACD Signal: The MACD signal is the 9-day exponential moving average (EMA) of the MACD.
- MACD Histogram (MACD HIST): MACD HIST is the distance between the MACD line and the MACD signal line.
- Commodity Channel Index (CCI): The CCI is a momentum-based oscillator that exhibits if the current equity is overbought or oversold.
- Momentum indicator (MTM): MTM is a momentum indicator based on the difference in closing price and the traded volume of the stock.
- Rate of Change (ROC): ROC is the ratio of current price difference and the previous price by the previous price.
- Relative Strength Index (RSI): The RSI is a momentum-based technical indicator that allows us to understand if equity is overbought or oversold.
- Stochastics (SLOWK and SLOWD): Stochastics are defined as the position of the current closing price with the previous high and low of n days.
- Chaikin Oscillator (ADOSC): The Chaikin Oscillator examines the strength of price moves and underlying selling and buying pressure to gauge the demand for security and possible turning points in the price.
- The Aroon (AR) indicator developed tells us whether an instrument is trending and how strong the trend is.
    i. Formula : AroonUp = ((Number of periods - Number of periods since highest high)/Number of periods) *100
    ii. AroonDown = ((Number of periods - Number of periods since lowest low) /Number of periods) *100
- Volatility Ratio (VR): The volatility ratio is a technical measure used to identify price patterns and breakouts. In technical analysis, it uses true range to understand how a security's price is moving on the current day compared to its past volatility.
- Bias: Deviation rate (BIAS) is an indicator that reflects the degree of deviation between the stock price and its moving average in a certain period.
    i. BIAS = [(Closing price of the day — N-day average price) / N-day average price] * 100%

- **Feature Selection**

After performing the feature engineering steps, we perform the feature selection process for each stock. We have 21 features from the original and the engineering features. Multiple features have information partially represented by the other feature. Also, all the features may not be important for prediction.

To solve the problems mentioned above we perform feature selection on the 21 features created. We selected the Recursive Feature Elimination (RFE) method in the current project because of its selection of features based on empirical evidence. As we have 5880 stocks in the dataset, we want the feature selection process to be as robust as possible. RFE selects the best 11 features from the 21 original features.

- **Dimensionality Reduction**

After running the feature selection process for a particular stock, we perform dimensionality reduction on the trimmed feature dataset. As mentioned before, the features may have multicollinearity, creating unwanted results in the model. Also, with increasing features, the model complexity increases, and we want the model to be fast for the practical purposes of running the model on ~6000 stocks.

We have performed Minmax scaler on the data before performing dimensionality reduction on the dataset. After scaling the data, we perform a dynamic dimensionality reduction of the features. We select k columns that explain at least the c% variance of the dataset. After testing various c% values, we fixed the value of c as 96% variance. For example, if 2 out of 9 features can explain at least 96% variance of the dataset, we select 2 component outputs from the PCA.

After performing PCA, we input the preprocessed data to the LSTM model.

- **LSTM model:**

The processed data from the PCA is delivered into the LSTM Model, where we have ten units in the first layer. These units are also known as memory units. Here we perform the major LSTM propagations, calculate the final output probability, and send it in via a dense output layer; based on the final probability, we decide if the stock goes up or down.

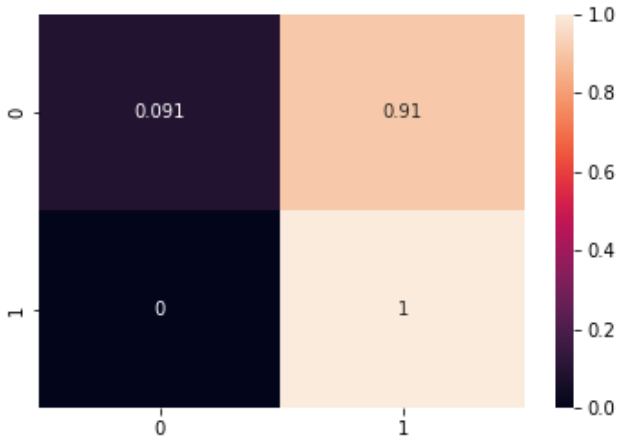
Results

Below are the results of the top 10 and the bottom 9 results. Where f1 represents true positive and true negative in a balance, TNR- true negative rate, FNR- false negative rate, FPR- false positive rate, TPR-true negative rate.

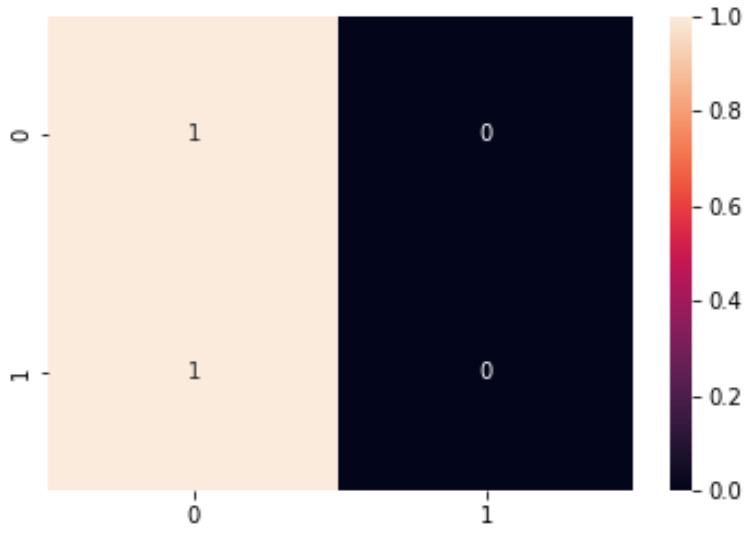
name	F1 score	TNR	FNR	FPR	TPR
DTYL	0.77	0.09	0.91	0	1
DEA	0.77	0.00	1.00	0	1
MPW	0.75	0.00	1.00	0	1
LIZI	0.75	0.33	0.67	0	1
FFC	0.73	0.00	1.00	0	1
GPN	0.73	0.00	1.00	0	1
PLAN	0.73	0.00	1.00	0	1
EGP	0.72	0.00	1.00	0	1
ISRG	0.72	0.00	1.00	0	1
REGN	0.72	0.00	1.00	0	1

name	F1 score	TNR	FNR	FPR	TPR
ATXI	0.00	1.00	0.00	1	0
AVAL	0.00	1.00	0.00	1	0
AVT	0.00	1.00	0.00	1	0
ATOS	0.00	1.00	0.00	1	0
AVXL	0.00	1.00	0.00	1	0
AWX	0.00	1.00	0.00	1	0
AY	0.00	1.00	0.00	1	0
BANR	0.00	1.00	0.00	1	0
AX	0.00	1.00	0.00	1	0

Below is the confusion matrix of the DTYL.



Below is the confusion matrix of AX.



Discussion

The best F1 score of the model is 0.77 for stock DTYL and DTA.

As an extension to this project the following can be worked on to produce other requirements:

- Increase the prediction time period
- Increase the lookback period on prediction
- Find a way to leverage trends of other stocks in the current prediction model.

The project made is very versatile can be used on any stock data to do decently accurate predictions.

Reference

Papers:

Short-term stock market price trend prediction using a comprehensive deep learning system | Journal of Big Data | Full Text

https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00333-6

Improving S&P stock prediction with time series stock similarity | Papers With Code

https://paperswithcode.com/paper/improving-sp-stock-prediction-with-time

HATS: A Hierarchical Graph Attention Network for Stock Movement Prediction

https://arxiv.org/pdf/1908.07999v3.pdf

Learning Multiple Stock Trading Patterns with Temporal Routing Adaptor and Optimal Transport

https://arxiv.org/pdf/2106.12950v2.pdfDeep

Attentive Learning for Stock Movement Prediction From Social Media Text and Company Correlations

https://aclanthology.org/2020.emnlp-main.676.pdf

Correlation between stocks:

http://snap.stanford.edu/class/cs224w-2015/projects_2015/Predicting_Stock_Movements_Using_Market_Correlation_Networks.pdf