DATA MINING AND ANALYSIS FINAL REPORT

Prediction of daily stock movements on the US market

Project ID : 5DMACP11

Under the guidance of

Dr. P G Sunitha Hiremath

TEAM MEMBERS:

1) Shivaraj Chattannavar	01FE17BCS191
2) Niketan Doddamani	01FE17BCS120
3) Gauravi Naik	01FE17BCS112
4) Namrata Nyamagoudar	01FE17BCS248

ABSTRACT

Stock price modelling and prediction have been challenging objectives for researchers and speculators because of noisy and non-stationary characteristics of samples. However, stock return do exhibit some predictability. With the development in deep learning, the task of feature learning can be carried out more effectively by efficient designing of the network. It has been observed that the use of neural networks for training the model is comparatively better than other state of art methods.
Keywords :
Stock price prediction, Stock return, neural networks

INTRODUCTION

The Stock of a corporation is all of the shares into which ownership of the corporation is divided.

Stock Movement shows the movement (in and out) of the stocks due to sale to customer transfer, sales return or stocking in inventories.
The goal of this challenge is to predict the sign of the returns at the end of about 700 days for about 700 stocks. A return, also known as a financial return, in its simplest terms, is the money made or lost on an investment over some period of time.

A return can be expressed nominally as the change in dollar value of an investment over time. A return can also be expressed as a percentage derived from the ratio of profit to investment. Returns can also be presented as net results (after fees, taxes, and inflation) or gross returns that do not account for anything but the price change.

The returns are thus such that a value of 0 means that the selling price of the stock is the same as that of purchasing price by the shareholder and positive (>0) means the shareholder earned a profit that is the selling price was greater than the purchasing price and inversely negative return(<0) means that shareholder incurred a loss. The end of the day return is the price change between the beginning bidding price of the stock and the final bidding price of the stock.

RELATED WORK

The paper on Deep Learning-Based Feature Engineering for Stock Price Movement Prediction gives us an insight into how deep neural networks are trained to predict the future stock movement. The authors of the paper have designed an end-to-end model named multi-filters neural network (MFNN) specifically for feature extraction on financial time series samples and price movement prediction task. Both convolutional and recurrent neurons are integrated to build the multi-filters structure, so that the information

from different feature spaces and market views can be obtained. The authors apply the MFNN for extreme market prediction and signal-based trading simulation tasks on Chinese stock market index CSI 300. Experimental results show that this network outperforms traditional machine learning models,statistical models, and single-structure(convolutional, recurrent, and LSTM) networks in terms of the accuracy, profitability, and stability. In our case as we are to predict whether the end of the day return shows a positive sign or negative sign, we design a simple neural network which learns on the input training data and predicts more accurately the positive sign or the negative sign as compared to the benchmark classifier, Light GBM classifier as provided as an example by the challenge providers.

PROBLEM STATEMENT

Prediction of daily stock movements on the US market

Duration : From Jan. 1, 2019 to Jan. 1, 2020

The challenge is to predict the sign of the returns means the price change over some time interval at the end of about 700 days for about 700 stocks. The returns are residual returns because part of the overall market movement was removed from raw stock returns.

UNDERSTANDING OF DATA

Input Train Data :
Number of tuples - 745327
Number of attributes – 74

Input Test Data :
Number of tuples - 319769
Number of attributes - 74

Description of data

Input Data:

- ID : Unique row identifier that matches input and targets.
- Equity code : Unique stock identifier.
- Date : Unique date identifier.
- Time intervals in steps of 5 minutes, in the form of HH:MM:00

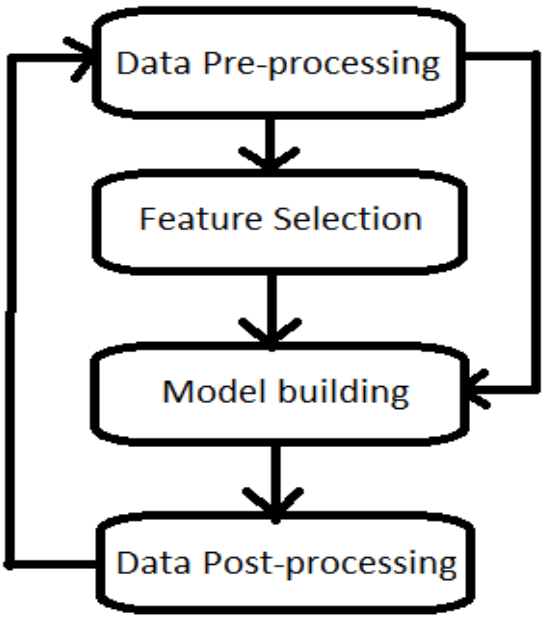
Output Data:

- ID : Corresponds to ID pair for the given target is given.

- End_of_the _day_return : Represents stock return from 3.30 pm to 4 pm

METHODOLOGY

We used Neural Network for building the model for prediction. We had a dataset of input_train , input_test and output data. The datasets had a huge number of tuples which contained null values. First we needed to tackle those null values and then build a model. We followed the KDD process which is iterative process.



In Data preprocessing, we first analysed the data sets. The input data contains the attributes such as ID, date, equity code, time intervals of 5 minutes starting from 9:30:00 to 15:20:00. As the date attribute was anonymized, we eliminated that attribute. Likely to the date attribute, equity code was also anonymized but we did not drop this attribute because there were 680 unique codes which later helped us in grouping the ID's.

Next we found some of the attribute values were missing values in the input dataset which would affect the end result. Handling with these missing values itself was a big task.Firstly, we filled those missing values with NaN. Later those NaN values is filled with nearly appropriate values that is, the values were filled with linear interpolation method.

After the data is preprocessed, we tried to build the models with Polynomial regression, Random forest, Light GBM and Neural Networks.

Among all these models, Neural Networking model gave more accurate results. The output data contains the end of the day return values which are compared to 0.5 with their respective ID's.

The end of the day return values are compared to 0.5, if the value is greater than 0.5, it is said to be a 'positive return' and if lesser, it is said to be a 'negative return'.

Our proposed neural network contains 9 layers with 10 epochs and linear activation function. The loss is calculated through mse (mean squared error) and 'adam' as optimizer.

The predicted output is compared with the actual output (From Custom metrics), then the accuracy for our model is given.

RESULTS

1)Polynomial Regression : The model failed due to a memory error and it was working fine till degree 5.

Accuracy : 0.5083

2) Light GBM : The model behaviour was very good compared to polynomial regression model but was overfitting

Accuracy : 0.5208

3) Neural Network : The model was convincing enough compared to the rest of the models.

Accuracy : 0.5238

REFERENCES

1)https://challengedata.ens.fr/

2)https://ideas.repec.org/a/bla/manch2/v63y1995i0p85-102.html

3)https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/neuralnet_model.htm

4)Long, Wen & Lu, Zhichen & Cui, Lingxiao. (2018). Deep learning-based feature engineering for stock price movement

prediction. Knowledge-Based Systems. 164. 10.1016/j.knosys.2018.10.034.