Roaldb86 / **Predicting-stock-market-direction**

⊙ Watch   0   ★ Star   0   ⑂ Fork   0

<> Code   ⊙ Issues 0   ⑂ Pull requests 0   ▥ Projects 0   ⎍⎍ Insights

Branch: **master** ▾   **Predicting-stock-market-direction** / **README.md**

Find file   Copy path

Roaldb86 Update README.md   a2a5cba 3 minutes ago

1 contributor

74 lines (47 sloc)   5.76 KB

Raw   Blame   History

Data scientist Nanodegree

Capstone Proposal

Roald Brønstad
November 14 2018

Proposal

I want to investigate the possibility of predicting tommorows stock market direction based on historical data. My intention is to create a dataset with historcial daily open, high, low, closing price and volume. I will also expand the dataset with technical indicators such as momentum, relative strength index, moving average etc. The difference between closing price today and closing price tommorow (prediction) will be either positiv or negativ, indicating direction

Domain Background

Predicting stock prices and market direction can surely be considered the holy grail within finance. Algorithmic trading is replacing humans and the finance sector is exploring how new technology can be utilized every day.

Traditionally the finance sector has relied on econometric tools such as linear regression to try and predict the market, today machine learning and deep learning networks are taking over the task.

Im educated within finance and have some investment background. I would love to use this Capstone project to try and combine data science with finance and also get some more experience analyzing time series which is often used in real-world data science projects. I will also use this project as an opportunity to explore artificial neural networks a bit more.

I found this academic research paper which is very relevant:
https://www.researchgate.net/publication/303380588_Predicting_the_Direction_of_Stock_Market_Index_Movement_Using_an_Optimized_Artificial_Neural_Network_Model

Problem Statement

I will experiment with ANN and some machine learning algorithms to do a binary classification for next days stock market direction. If tomorrows closing price is higher than todays it will be labeled as 1, if direction is negativ the label will be 0. I will look at the Oslo stock exchange benchmark index OSEBX historical daily data for training the models. I will split the data in to a training and test set and evaluate the model against the test set.

Datasets and Inputs

Dataset with historical open, high, low, closing price and volume for the OSEBX index can be found on this site: https://www.oslobors.no/ob_eng/markedsaktivitet/#/details/OSEBX.OSE/overview

The dataset is from 2001 and has about 4400 rows. I would prefer a more detailed dataset (hourly or minute data) but cant find any. I will also compute technical indicators such as momentum, relativ strength and moving average to supply as features.

As a sidenote, the dataset has been adjusted for dividends already

Solution Statement

A model that can classify next day movement as either 1 or 0 for up or down. I will experiment with different machine learning and neural networks to find an optimal model. The models will be evaluated by their ROC curves to find the best performing one. A good model is one that can predict movement significantly better than random guess. There will probably be a skewness in the data towards more observations in the positive direction as the stockmarket has been moving up since 2001.
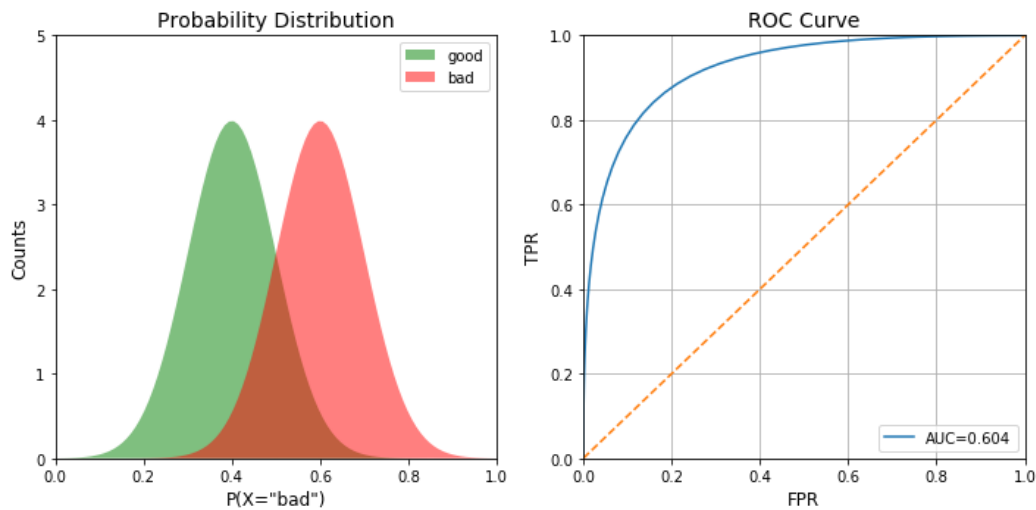
Benchmark Model

The traditional tool for predicting the stock market is regression. I would like to use linear regression as that is what we used when I studied finance and it would be interesting to compare machine learning and neural networks to that. However since this is a binary classification problem i will use logistic regression as a benchmark to the more advanced models this project will develop

The benchmark and project models can be compared by their ROC curves and AUC on the testdata.

Evaluation Metrics

I will use ROC curve and area under curve as an evaluation metric for this project.

A roc curve is derived from a confusion matrix of the test data (true/false positive and true/false negatives) A binary classifier will classify based on the distribution of probabilities. (see picture below) A treshold needs to be set unless the model can predict 100% a ROC curve shows true positive rate (true positive / all positive) againts a false positive rate (false positive / all negative) plotted for all discriminating tresholds in the distributions. The diagonal line shows pure guessing between two classes.

picture from:
https://towardsdatascience.com/receiver-operating-characteristic-curves-demystified-in-python-bd531a4364d0

Project Design

The first part of my project will be to engineer a dataset. Prices and volume are already given but I want to derive other features from them. RSI*, momentum and moving averages should be included but I will also explore further posibilities. I will check for outlier values and missing values in the dataset and normalize the data before training models on it.

The second stage is to split the data to a training set and test set. Since the data is a time series, I need to split accordingly. I will also have to derive the target value (1 or 0) from the closing prices.

Then i will make a simple logistic regression model as a benchmark and create a few machine learning models such as SVM and random forest to compare it with.

Finally i will build an artificial neural network with pytorch to see if a deep learning model can outperform the benchamrk and the machine learning models. I will experiment with hyperparameters and network architecture underway to optimize my models. The models will be evaluated by ROC-curves and AUC.

*definition of RSI:

RSI is a momentum indicator that measures the magnitude of recent price changes to analyze overbought or oversold conditions. It is primarily used to attempt to identify overbought or oversold conditions

Read more: Relative Strength Index (RSI) https://www.investopedia.com/terms/r/rsi.asp#ixzz5WqNgZOGW