Debugging myself

Amy Ma's data-related life

BUSINESS ISSUES / **CLUSTER ANALYSIS** / **DATA MINING** / **GRADIENT BOOSTING** / **MACHINE LEARNING** / **PYTHON** / **SUPERVISED LEARNING** / **UNCATEGORIZED** / **UNSUPERVISED LEARNING**

Predicting Stock Returns with sentiment analysis and LSTM

- ASIDE
- NOVEMBER 27, 2016
- YUJINGMA45
- LEAVE A COMMENT

This project inspired by a recent acquisition activity is Bass Pro to acquire Cabela's. I would like to look at the revenues and the market share of Cabela's and one of its competitors, Dick's Sporting Goods, prior to acquisition and see if there are any features/signals that can be seen in the last few months prior to acquisition. To test this effect with anterior data, I used both stock market and social media datasets to predict stock returns for Cabela's and its main competitors. The test RMSE for my model is around 0.019.

Implementation

Assumptions

1. Stocks in the same industry are driven by the same signals and are correlated with each other.
2. The historical returns for (n-1) days can be used to predict the return of nth day. That is, a stock's returns over a long enough trading period contain information about the next day.
3. The official Twitter accounts are extremely active before big changes, and the messages should be more positive.
4. Stock momentum (i.e. no reversion back to zero) is correlated with retweets and favorite related tweets.

Dataset

Based on the assumptions, I used two main datasets:

- Daily stock price and corporate action for CAB, DKS, HIBB and S&P 500 Index from 2011-10-11 to 2016-10-07. The data was obtained from Yahoo! Finance and amigobulls, which includes the P/E ratio, P/S ratio and stock price for a given day.

- Historical twitter data sent by these companies from 2011-10-11 to 2016-10-07. The data was gathered using Twitter's API. It contains the timestamp, tweet messages, number of retweets and favorites. Since my analysis is on a daily basis, I aggregate the tweets by date. In addition, more than 160 million public tweets are used to do sentiment analysis. The data can be downloaded from this website (http://help.sentiment140.com/for-students/).

Approach

Daily stock returns are calculated by using adjusted close price and dividends (according to historical corporate action). Twitter data is aggregated to count daily retweets and favorites. The time series data are preprocessed using independent component analysis, window methods and lead-lag correlation, before being fed into the final three models. At the same time, the tweets are converted to numerical features as more positive (+1) or negative (-1) by using doc2vec. Then the mood data and other data are fed to the final model, which blends an AR model, ridge regression and LSTM neural network. Here is a flow diagram of the approach and the illustration of the cross-validation process:

1. Independent Component Analysis

   "The idea of ICA is to linearly map the observed multivariate financial time series into a space of statistically independent components(ICs)." (Back, A. D., & Weigend, A. S. ,1997)

   The estimated ICs have two main categories: Infrequent and smaller fluctuations (due to the major change of the whole stock market); Frequent with large shocks

   Here are the three signals which are founded by ICA:

2. Sentiment analysis

Results

1. Cabela's daily returns are significantly autocorrelated with the past 4 days; Dick's Sporting Goods' daily returns are significantly autocorrelated with the past 17 days, suggesting more stability. Indeed, market capitalization for Cabela is approx. 4.18Bn, and 6.35Bn for Dick's.
2. Abnormal number of retweets can capture the big change in the companies. Cabela's official Twitter account get much more retweets on before the acquisition.

3. There is a lead-lag effect between the current Dick's Sporting Goods' daily returns and the others' returns for past 2 and 29 days. However, there is no pattern for the current Cabela's and others' daily return, but it significantly related to its passed P/E ratio and P/S ratio.

Further Work

1. Adding new data

- Google trends
- Key statistics in Yahoo! Finance
- Seeking Alpha — news titles

2. Improving cross-validation process