Machine Learning Capstone Proposal

Domain Background

Making stock market predictions based on World Daily News. Simply taking daily news and correlating it with the Dow Jones Industrial Average does not have enough accuracy to make a buy/sell decision. I will see if GICS sectors can better **classify** whether a stock sector is good to buy, sell or stand still. Stocks is about picking the right stocks. I would rather buy 5 stocks that do extremely well then predict 500 stocks that do marginally well.

In this case regression is not good because regression looks at how close we were to predicting end price. If a stock price goes above and beyond the predicted price is better than being close to predicted price. Minute 5:00 of the video below is a good explanation

Tastytrade. (2016, August 8). Predicting Stock Price with Machine Algorithms (Part 1). 8/8/2016. Retrieved 7/29/2019 from https://www.youtube.com/watch?v=brtRJxebL58&t=4s

Problem Statement

Based on the World Daily News, can you predict with confidence to buy or sell a stock with 80% accuracy and precision when you focus on certain GICS sectors?

Datasets and Inputs

Plan is to combine the News articles from Aaron7sun's Daily News for Stock Market Prediction dataset and Dominik Gawlik's dataset from New York Stock Exchange. Only gather data from the same date ranges (i.e. 1/4/2010-7/1/2016).

Stock Labeling:
A stock sector is considered up if it surpasses 1 standard deviation (1 sigma) away. A big move is if it passes 2 standard deviations away (2 sigma). 1 sigma = stock up, 2 sigma = stock extremely up, -1 sigma = stock down, -2 sigma = stock extremely down.

Final Dataset: cache/data.csv

- Dimensions: 17985 x 5.
- Columns: Date, News, GICS Sector, Label 1 Sigma, Label 2 Sigma.
- Dates: (1/5/2010-7/1/2016). Data taken daily.
- This is a multiclass classification with -1, 0, 1.
- Data is time based and imbalanced to have a value of 0. For best results, data will be split time based and then an algorithm that

Baseline Dataset: Combined_News_DJIA.csv

- Source: https://www.kaggle.com/lseiyjg/use-news-to-predict-stock-markets
- Dimensions: 1989 x 27.
- Columns: Date, Label (DJIA went up) and 25 columns for the top 25 news of the day.
- Usage: Used to plot the baseline using logistic regression

News Dataset: RedditNews.csv

- Source: https://www.kaggle.com/lseiyjg/use-news-to-predict-stock-markets
- Dimensions: 73608 x 2.
- Columns: Date, Label (DJIA went up) and 25 columns for the top 25 news of the day.
- Description: The top 25 world news from Reddit
- Usage: Group all the news articles by date then see if there is correlation between the news sentiment and the daily stocks

News Dataset: prices-split-adjusted.csv

- Source: https://www.kaggle.com/dgawlik/nyse
- Dimensions: 851264 x 7.
- Columns: Date, Symbol, Open, Close, Low, High and Volume.
- Description: NYSE stocks on a daily basis for the S&P 500. The opening and closing bell.
- Usage: Used to aggregate stock market changes by GICS sectors. Calculate positive change (1) if stock went up for 1 sigma. Calculate negative change (-1) if stock went down by 1 sigma. Label zero (0) if stock had no discernible change.

News Dataset: securities.csv

- Source: https://www.kaggle.com/dgawlik/nyse
- Dimensions: 505 x 8.
- Columns: Ticker Symbol, Open, Close, Low, High and Volume.
- Description: Description of each ticker symbol categorized by GICS Sector and sub industry
- Plan: Join the stock description with the daily S&P 500 stocks then join with the Reddit News

Solution statement

Instead of using the Dow Jones (30 stocks) use the NYSE dataset (500 stocks). Group the stocks by GICS sector and define a positive/negative change if it passes a one or two sigma threshold. Do a separate sentiment analysis for stocks going up/down by GICS sector. Utilize the bag of words and term frequency to find the featured n-grams that best correlate with the stocks.

Benchmark model

Use Aaron7sun's dataset and labeling as the basis for the benchmark. In Kaggle, users were only able to average 50-60% accuracy. For comparison sake we will also look at the precision as we want to limit false positives. Take the combined news dataset and do a standard sentiment analysis using a logistic regression model.

Sun, J. (2016, August). Daily News for Stock Market Prediction, Version 1. Retrieved 7/29/2019 from https://www.kaggle.com/aaron7sun/stocknews.

Evaluation metrics

In the purchasing and selling of stocks you are charged on a per transaction fee by your broker. It is better to purchase less stocks with better quality then many stocks in small quantity and hurt yourself on transaction fees. As a result, it is better to be selective and purchase fewer stocks with the least amount of risk then purchasing more stocks and risk losing money.

Thus, lowering false positives and having higher accuracy is key. While false negatives hurt, it is safer to focus on true positives. This is the reason accuracy and specificity will be used as evaluation metrics for this research.

Due to the imbalanced classes and multi-class issues, we will also calculate Cohen's kappa.

https://thedatascientist.com/performance-measures-cohens-kappa-statistic/

Project Design

Workflow.

Data Gathering: Group the S&P 500 stocks by sector and then average the stocks. Take the stock sectors and classify the sectors by -1, 0, 1 on whether to buy, sell or hold on the stock. This will normalize the data. The sector will be classified based on two magnitudes; 1 or 2 sigma (standard deviation). 1 sigma will classify 35% positive/negative results and 2 sigma will classify 5% positive/negative results. The stock dataset will then be combined with the news dataset by date. For further tuning, remove outlier stocks that do not correlate to other stocks in same GICS sector.

Data Analysis: Due to the time series data and the class imbalance, it will be difficult to get good results. We will use TimeSeriesSplit to break the dataset into n splits. To remove class imbalance we will try to use undersampling (ClusterCentroids) techniques as well as algorithm hyperparameters such as log loss.

Sentiment: data will be run through bag of words and term frequency analysis to come up with features that best correlate to the change in stock price

Data Modeling: When it comes to modeling we will first start by doing LogisticRegression on all sectors and then by each GICS sector. Afterwards, we will test various multi-class classification models such as XGBoost, LightGBM, RNN and a deep learning model like Keras.

https://towardsdatascience.com/multi-class-text-classification-with-lstm-1590bee1bd17

Afterwards, the. Finally, utilize a model such as logistic regression to determine whether the given stock sector can hit 80% accuracy

References

Gawlik, D. (2017, February). New York Stock Exchange. Retrieved 7/29/2019 from https://www.kaggle.com/dgawlik/nyse

Kampakis, S., Dr. (2016, May 8). Performance Measures: Cohen's Kappa Statistic. Retrieved 7/29/2019 from https://thedatascientist.com/performance-measures-cohens-kappa-statistic/

Li, Susan. (2017, April). Multi-Class Text Classification with LSTM. Retrieved 7/29/2019 from https://towardsdatascience.com/multi-class-text-classification-with-lstm-1590bee1bd17

Sun, J. (2016, August). Daily News for Stock Market Prediction, Version 1. Retrieved 7/29/2019 from https://www.kaggle.com/aaron7sun/stocknews.

Tastytrade. (2016, August 8). Predicting Stock Price with Machine Algorithms (Part 1). 8/8/2016. Retrieved 7/29/2019 from https://www.youtube.com/watch?v=brtRJxebL58&t=4s