Technocolabs
BUILD WITH AI

Blueprint of The project

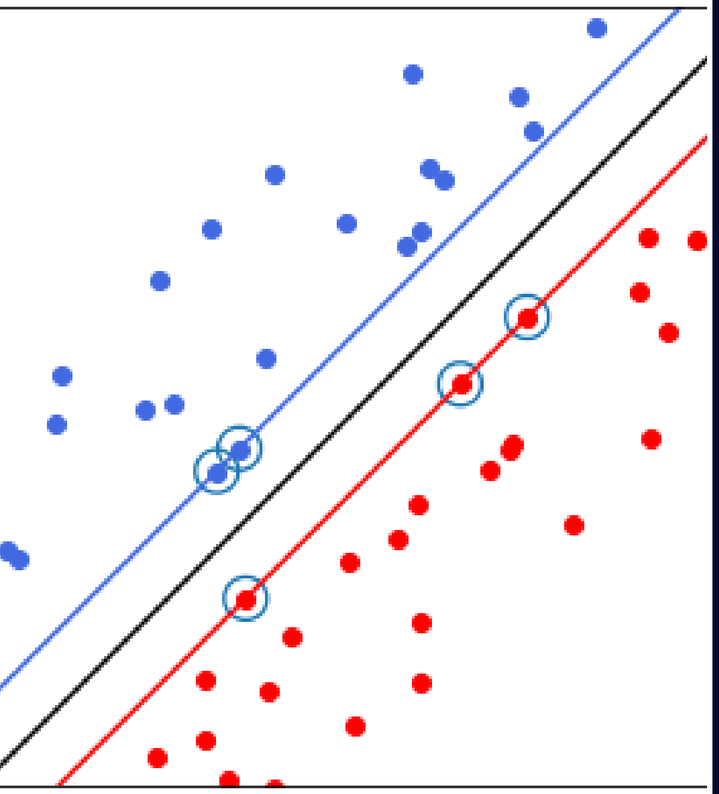
Predicting market volatility using macro headlines

DAHMANI Salah Eddine

INTRODCUTION

On June 26, 2015, months of debt negotiation between the Greek government, headed by Prime Minister Alexis Tsipras, and its creditors, including the IMF and fellow Eurozone countries, broke off abruptly. Tsipras announced a snap referendum regarding the terms of the pending bailout. By the following morning, S&P500 was down significantly. As political and economic uncertainty grew in Europe, investors across the world were moving their funds away from risky assets that could be negatively affected by a Greek default, and towards safer assets. Days later, when the Greek situation had 'resolved', the process reversed as equities rallied.

These market movements are not uncommon; just weeks later, crisis erupted in the Chinese equity markets, and asset values were responding to the uncertainty this provoked. In this paper, we investigate how macroeconomic sentiment immediately affects the volatility in liquid markets, by measuring market volatility through the VIX (volatility index of the S&P 500). Using data pulled from Twitter, we are researching how breaking economic news affects the markets, and subsequently how to predict which news stories can increase volatility. We will consider a tweet 'significant', in that the news presented in the tweet contributes to volatility in the market, if within 30 minutes of the tweet being tweeted, the volatility of the asset increases by one-fifth of a standard deviation.



By employing AI techniques for classification and prediction.

Using a training set, I will identify key words and their associated probability of increasing volatility in the markets using Naive Bayes, SVM and logistic regression. The input of the algorithm will be the word count of each bucket of tweets in a dictionary. Then I use the three prediction methods to output a predicted increase in the VIX, which will be a binary variable. In order to create a viable trading strategy, I aim to predict moves with above 51% accuracy.

Business Problem !

Ask common men, they will say that – stock's price go down more than it goes up.

Why we think like this? Because we don't know how to predict if a stock will go up or down.

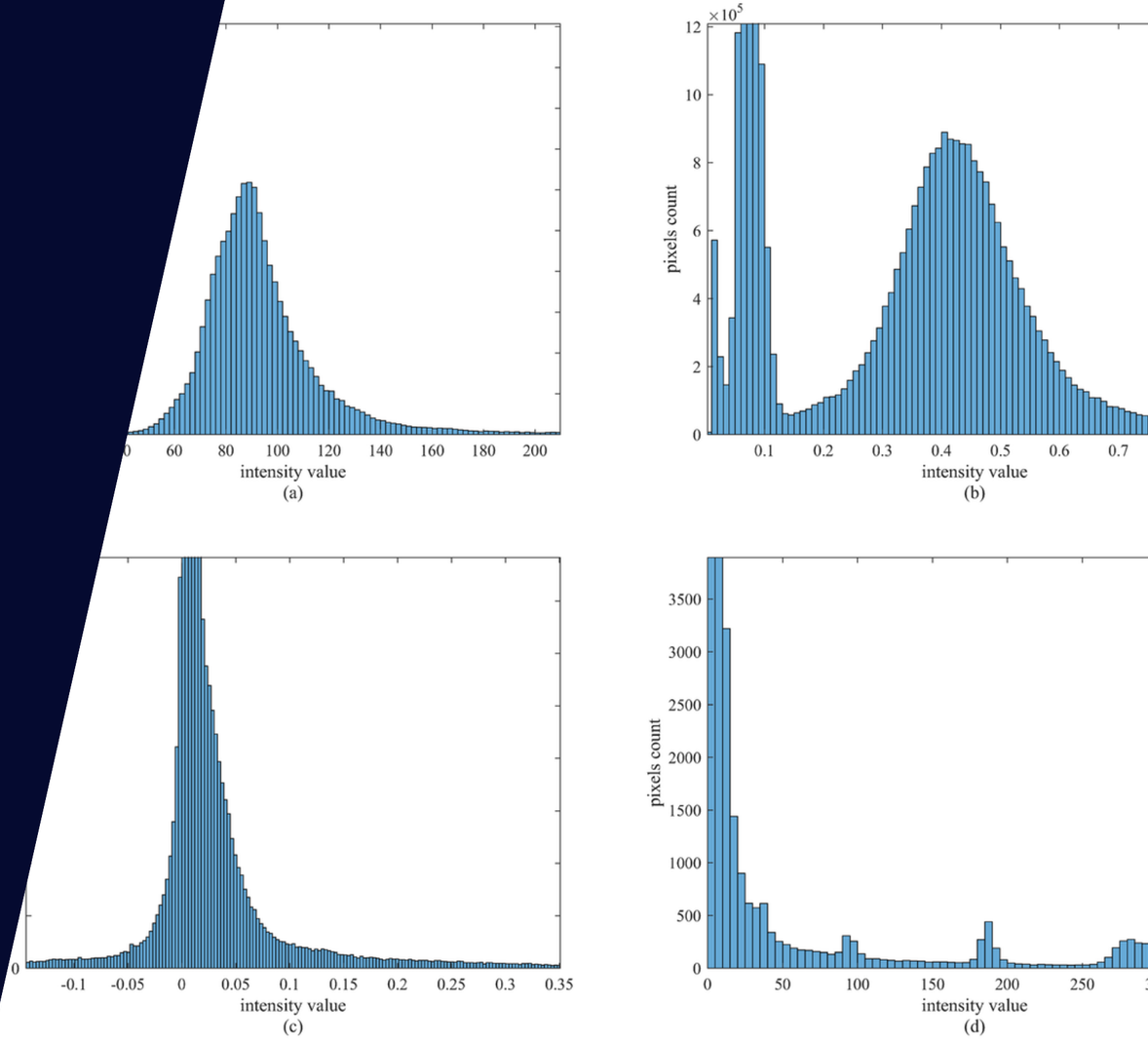
BUSINESS PROBLEM

Predicting market price movements is not easy because the market has lot of mouvements, reactions and consequeces and it varies by secons or minutes, we can do it by conducting technical analysis using historical and time-series data and many studies fall into developing techniques to analyze markets informations from texts of websites or social media posts.

Some methods include tokenizing each article, and matching them with stop word lists, using a subset of article terms as features, and using tagging of named entities to group nouns into predefined categories.

The latter two techniques are more frequently used in Question Answering (QA) systems, where themes have been predetermined. However, to analyze the sentiments within markets without knowing the objectives, the lexicon-based approach is a more

Robust technique, although it is subject to the effect of words around a single word, and trends can sometimes be hard to detected by using single-word weighting.
We'll use many approaches and models to predict the market the prices and their ups and downs.



Business constraints

This is not only our problem, even experts of stock market face a similar dilemma.
In short term (span of 2-3 months), stock price movement is mostly speculative. If there are more buyers, price goes up. If there are more sellers, price falls.

What triggers buying or selling? Quarterly or annual reports publication by the company. If results are positive, stock's price will go up. If results are negative, it might trigger a fall.

But in real world, factors effecting share price is more complex. It not only depends on the fundamentals of the company it represents, but also on hosts of other factors. It is a complex puzzle, and for common men like us, it is a hard nut to crack.

So with an AI model we'll use all the features that has a real influence on the market prices.

Data Overview

We have pulled macroeconomic news from Twitter; news sources will tweet a headline and the link to the accompanying news article, which allows for us to access the key topics frequently and in a rather condensed format.

For the data we have data of many periods, of last month, last 6 months, last year… we'll use first the data of one month after that we'll analyze the performance of our models if we've overfitting or under fitting and then we decide if we add more data or not, without forgetting that we'll use **70%** for training and **30%** for testing.

Our data has 6 features, the date, the close/last, the volume, Open, High and Low.

TYPE OF MACHINE LEARNING PROJECT

We have a classification **(supervised learning**) problem because we are not focusing on single-name stocks but rather market volatility so we'll use three models of classification to determine which one will fit better our data.

One is Support Vector Regression, which performs linear regression in the high dimension feature space to predict stock prices. Logistic Regression and Neural Networks are also plausible in analyzing stocks trends. However, all these methods are sensitive to bias and noise. In addition, the majority of past research focus on price prediction of specific stocks, and very few of them focus on the volatility of the overall markets. We have taken inspiration from various article and the techniques used, such as SVM and logistic regression.

Performance metrics

We want to compare the three classification models to determine which one best fits our data. Therefore, we use hold-out cross validation to devise 70% of our data for training, and the remaining 30% for testing, and compare the estimated generalization error/accuracy of each model.

The primary metrics for each model that we are interested in are accuracy, precision, and recall. Accuracy is the proportion of correctly classified data points to total number of data points. Precision or positive predictive value is the number of true positives over the total number of positives predicted, or the probability that a positively predicted data point is actually positive. Recall is the proportion of positive data points that are correctly classified.

Thank you!

Feel free to reach out to me
if you have any questions.