Predicting Volatility in Equity Markets Using Macroeconomic News

Anirudh Saxena

INTRODUCTION

There is a lot of research on predicting market price movements by conducting technical analysis using historical and time-series data . A large portion of the studies fall into developing techniques to analyze markets trends from texts such as nancial market news or social media posts .

In this project work, we investigate how macroeconomic sentiment immediately impacts the volatility in liquid markets, by measuring market volatility through the VIX (volatility index of the S&P 500).

By employing three machine learning techniques for classification, we aim to predict the volatility-inducing power of a single tweet and/or news story.

Business problem

Introduction

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 o abruptly.

Context

The market movements are not very uncommon.

Two examples include Chinese equity market and european market where asset values were responding to the uncertainty.

Problem statement

Predict the market movements using machine learning models.

Predicting volatility movements, with an accuracy between 57-67%.

DATASET AND FEATURES

1. Twitter data

The Twitter sentiment data are available in separate CSV files for each company. As an example, for the Microsoft (MSFT), the data are contained in the file twitter_data_MSFT.csv. The columns of the CSV files are:

Date
Number of negative tweets
Number of neutral tweets
Number of positive tweets
Total number of tweets

2. Financial data

The financial data are available in separate CSV files for each company. As an example, for the Microsoft (MSFT), the data are contained in the file financial_data_MSFT.csv.
Also, the data for the DJIA index are contained in the CSV file financial_data_DJIA.csv.
The columns of the CSV files are:
Date
High
Close
Open
Low

Figure shows a plot of the VIX since 2008 (daily data). While we used intraday data going back 2 months,
this illustrates how events, such as the Greek financial crisis, the Lehman Brothers default, the devaluation of the yuan, and other macro events affect the volatility.

Methods

For predicting movements in VIX, we use three different supervised learning methods:

1. Naive Bayes

2. Support Vector Machines

3. Logistic Regression with PCA

PERFORMANCE METRICS AND RESULTS

Accuracy and precision of model.

This is a regression type problem.

Compare the three supervised learning models to determine which one best ts our data.

Therefore,we use hold-out cross validation with 70% of our data for training, and the remaining 30% for testing, and compare the estimated generalization error/accuracy of each model.

Do not use k-fold cross validation, as we need our testing data to be the most recent 30% of the tweets.

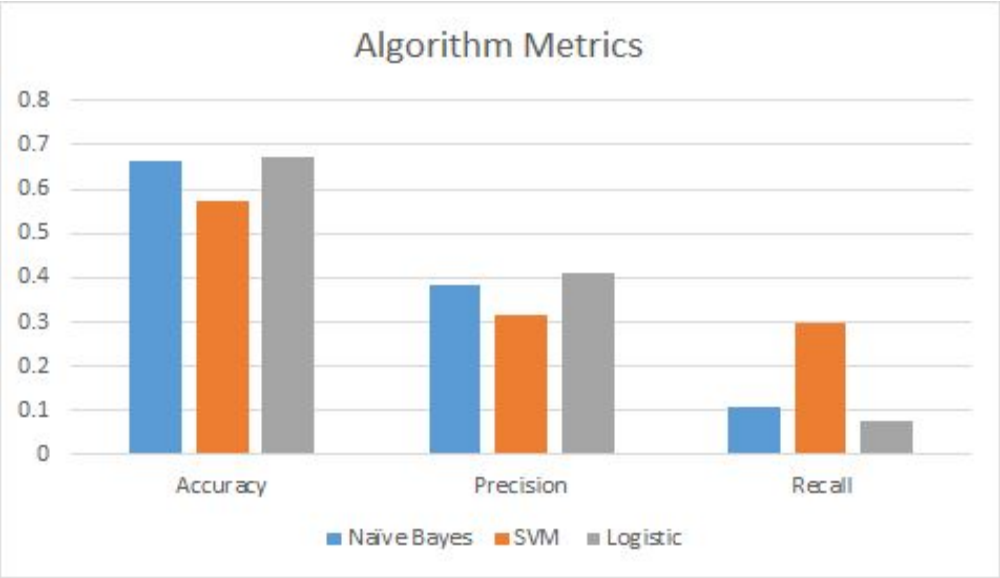
PERFORMANCE METRICS AND RESULTS

Confusion Matrices								
	True Value			True Value			True Value	
Naïve Bayes	Negative	Positive	SVM	Negative	Positive	Logistic	Negative	Positive
Predict Neg.	186	84	Predict Neg.	142	66	Predict Neg.	192	87
Predict Pos.	16	10	Predict Pos.	60	28	Predict Pos.	10	7

We see from figure that both Naive Bayes and Logistic Regression do a good job predicting negative data points, but do not predict very accurately the positive data points.

The SVM algorithm predicts more positive data points than Naive Bayes and Logistic Regression, but also has a lot more false positives, lowering its precision.

PERFORMANCE METRICS AND RESULTS



Logistic Regression has the best accuracy and precision at 67% and 41% respectively, with Naive Bayes trailing with 66% and 38% for accuracy and precision.

The most important metric for us is precision,
since we can see it as the probability that the position we enter into will be portable or not. Therefore,
since Logistic Regression has the highest accuracy and precision, we come to the conclusion that it is the best
model for our problem.

CONCLUSION

Using three supervised learning techniques, we have developed a methodology for predicting volatility movements, with an accuracy between 57-67%.

The Logistic Regression model outperformed both Naive Bayes and SVM due to less assumptions being made, and a lower chance that the model was overt.