Predicting Market Volatility Using Macro Headlines

Introduction

- Data from financial markets offer challenging problems and have attracted much interest from quantitative researchers and traders. With easy availability of electronic data, there is significant recent growth in their analysis, and such analysis translating into automated trading systems.
- The analysis shows that the recent news of macroeconomic variable (such as Tweets & News Headlines from different data sources) can be used to improve the prediction of Stock Market Volatility through the VIX (Volatility Index).
- Large market movements as a consequence of political and economic headlines are hardly uncommon now-a-days. Liquid markets are most susceptible to swing when such news breaks.
- Using macro economics news headlines and tweets from major news sources, we will try to predict market volatility using VIX as a proxy for market Volatility.

Dataset Analysis

- Our Data source is macroeconomic News, from Twitter, which consists 180,000 tweets from 70 accounts including Financial Newspapers, Breaking news sources, Hedge Funds and Investment Banks, Notable Economists & Analysts.
- Both Twitter Data and market data is pulled over a period of 6 months.
- Using subset of tweets, dictionary of 10,000 word have been created, which includes token results from '#' feature from twitter.
- Tweets were grouped together in 30-minutes increments.

Figure 1 shows a plot of the VIX since 2008 (daily data). While we used intra-day 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 effect the volatility.

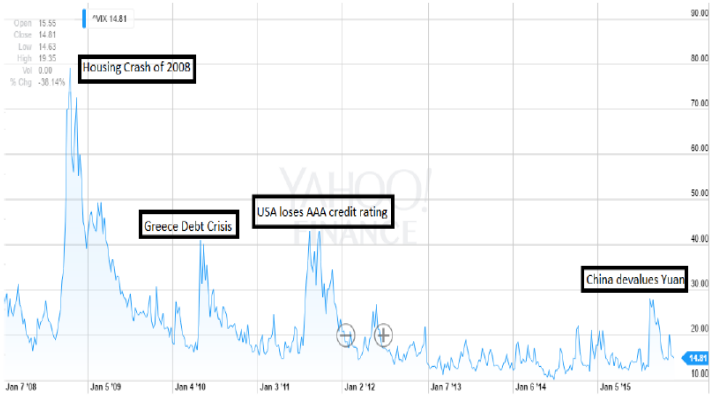

FIGURE 1.

Machine Learning Models for Predicting Market Volatility

We are using three different Supervised Learning Models for our prediction:

- **Naive Bayes,**
- **Support Vector Machine (SVM),**
- **Logistic Regression & PCA.**

- **Naive Bayes:** We can use Multiclass Naïve Bias classifier algorithm for twitter and news dataset analysis. The key assumption of this model is the conditional independence of the features.

- **Support Vector Machine (SVM):** In order to perform Linear Regression in high dimension feature to maximize the functional margin. Obtain a desirable linear hyper-plane to maximize the linear separation between classes.

- **Logistic Regression & PCA:** Logistic Regression is a model that measures the relationship between categorical binary dependent variable. It estimates the probabilities of the categorical variable using the Logistic function i.e.: Sigmoid Function.

$$P(y=1;x,\theta)=h_x(x)=\frac{1}{1+\exp^{-\theta^T x}}$$ Sigmoid Function

Principal Component Analysis is a process to reduce the dimension of the dataset (twitter & news dataset) by using Orthogonal transformation.

Comparison & Analysis of Different Machine Learning Algorithm Results

We want to compare three classification models which one best fits to our dataset. We are taking 70% of the dataset for model training (& cross validation) and remaining 30% dataset has been kept for model testing.

All three models, SVM, Naive Bayes & Logistic Regression, performed almost similar, accuracy is around 60-64%.

Method	Naive Bayes	SVM	Logistic
Accuracy	0.624	0.5586	0.6461

Figure 2: Accuracy from Three Methods

Here, we can see Logistic Regression & PCA provides maximum Accuracy among all three model.

For Classification models, the metrics for evaluation the models are:

Precision, Recall, Accuracy.

$$Acc = (TP + TN)/M, \quad Prec = TP/(TP + FP), \quad Recall = TP(TP + FN),$$

where TP stands for true positive, TN true negative, FP false positive, FN false negative, and M number of data points.

In order to reduce the dimension of covariates form our Dictionary which contains 10,000 tokens, we performed PCA algorithm on twitter dataset.

Confusion Matrix for all three Models:

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

Conclusion

Logistic Regression outperformed over other two algorithms:

- Naive Bias Assumption is for conditional independence of the features which is not applicable here for twitter and news dataset analysis.
- Also, Logistic Regression has a lower chance to overfit due to the less assumption being made. Here, PCA has been performed before applying Logistic Regression in order to lower the number of dependent features which actually helped to lessen the chance of Overfitting.

Future Scope Analysis

- With more concentrated twitter data & taking only macroeconomic headlines from news data, would definitely increase the accuracy and precision of the models.
- By modifying and shortening the dictionary size and including only market related words and tokens, we can lower the chance of Overfitting the model, resulting better accuracy and precision.
- Taking data form other data sources (such as Bloomberg) also will enrich our dataset collection, which results more clean data, using which we can obtain more accurate market volatility prediction.
- Eliminating data which are not related to S & P500 movements, we can foresee to get more accurate outcome.