Predicting Volatility in Equity Markets Using Macroeconomic News

Introduction

In this project 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). Large market movements as a consequence of political and economic headlines are hardly uncommon; liquid markets are most susceptible to swings when news breaks. We predict equity market vol using tweets from major news sources, hedge funds and investment banks, and notable economists

Dataset and Features

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. In addition, hedge funds, banks, and analysts will frequently tweet either articles or short views on the market. We have selected 70 accounts that we deemed relevant, and tweeted with a 2 significant frequency. The Twitter API allows for the most recent 3200 tweets per account to be exported from the website, which provides at least a month of tweets (and hence, news) per account. Because we are investigating the immediate market reaction, and therefore using intraday data, this is more than sufficient. This amounted to more than 200,000 tweets. 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.

Methods

1. *Naive Bayes:* The first classification model we consider is the Multi-factor Naive Bayes. The key assumption of this model is the conditional independence of the features.
2. *Support Vector Machines (SVM):* Support vector machine method, It performs linear regression in the high dimension feature to maximize the functional margin.
3. *PCA and Logistic Regression*: We performed logistic regression. Logistic regression makes the least amount of assumptions on the dataset, and since we cannot be sure that the assumption of conditional independence holds true in the Naive Bayes algorithm, logistic regression makes sense.

Results

After the comparison of three classification models we want to find the model which one best fits our data. The primary metrics for each model that we are interested in are accuracy, precision, and recall. 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.We used following formulas:

$$Acc = (T\,P + T\,N)/M,\ P\,rec = T\,P/(T\,P + F\,P),\ Recall = T\,P(T\,P + F\,N),$$

where T P stands for true positive, T N true negative, F P false positive, F N false negative, and M number of data points. We first look at the confusion matrix to get an idea of how each model is performing.

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

FIGURE 2.

We see from **figure 2** 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. These observations are made more precise in figure 3, which compares each of the models main metrics to each other.

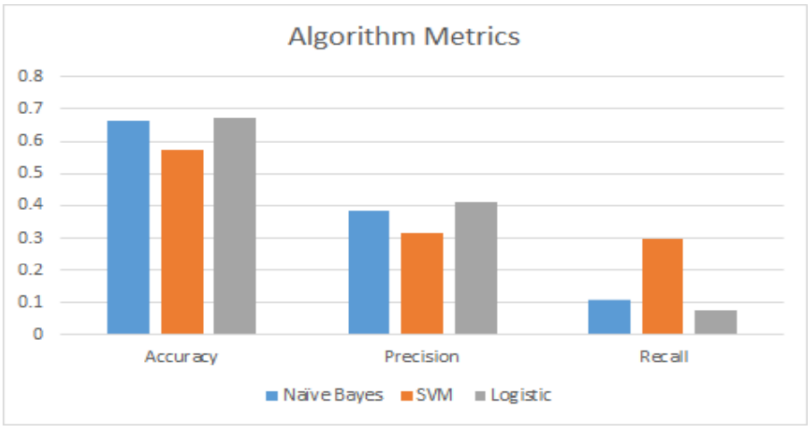

FIGURE 3.

From **figure 3**, we see that 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. Although SVM does much better than both Naive Bayes and Logistic Regression in recall, we still consider SVM to be the worst performing model for our purposes. This is because our trading strategy will only enter into a position in the market if we predict a positive result from the data. Thus, the most important metric for us is precision, since we

can see it as the probability that the position we enter into will be profitable 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 overfit.