Date: 16 September, 2021

BLUEPRINT OF THE PROJECT

PREDICTING MARKET VOLATILITY USING MACRO HEADLINES

By

Sowmya Prakash

PROJECT LIFECYLE:

PROBLEM STATEMENT:

Predicting market volatility using macro headlines

INTRODUCTION:

Large market movements as a consequence of political and economic headlines are not uncommon, liquid markets are most susceptible to swings when news breaks.

Using the VIX as a proxy for market volatility, we investigate how macroeconomic news headlines affect changes in the VIX. We predict equity market vol using tweets from major news sources, hedge funds and investment banks, and notable economists.

Using a training set, we will identify key words and their associated probability of increasing volatility in the markets using Naïve Bayes, SVM and logistic regression.

DATA OVERVIEW:

Twitter provides a plethora of market data. In this project, we will use over 200,000 tweets from various accounts to predict upward movements in the VIX.

180,000 tweets from 70 accounts, including:

- Financial Newspapers
- Breaking News Sources
- Hedge Funds and Investment Banks
- Notable Economists and Analysts

Both Twitter data and market data is pulled over a maximum of 6 months. The Twitter API allows collection of 3200 most recent tweets per account.

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

DATA PREPROCESSING & EXPLORATORY DATA ANALYSIS:

Data Preprocessing -

- Data Cleaning -> remove blanks, null value and duplicates if any, handling outliers
- Parse the date into datetime data type, identify general statistics
- Standardizing the data

EDA -

- Visualizing various plots -> to analyze the distribution pattern of various attributes
- Understand the data pattern with respect to target variable (Close)
- Feature Selection -> remove unnecessary columns, identifying Correlation using Heat Map/Matrix

MODEL IMPLEMENTATION

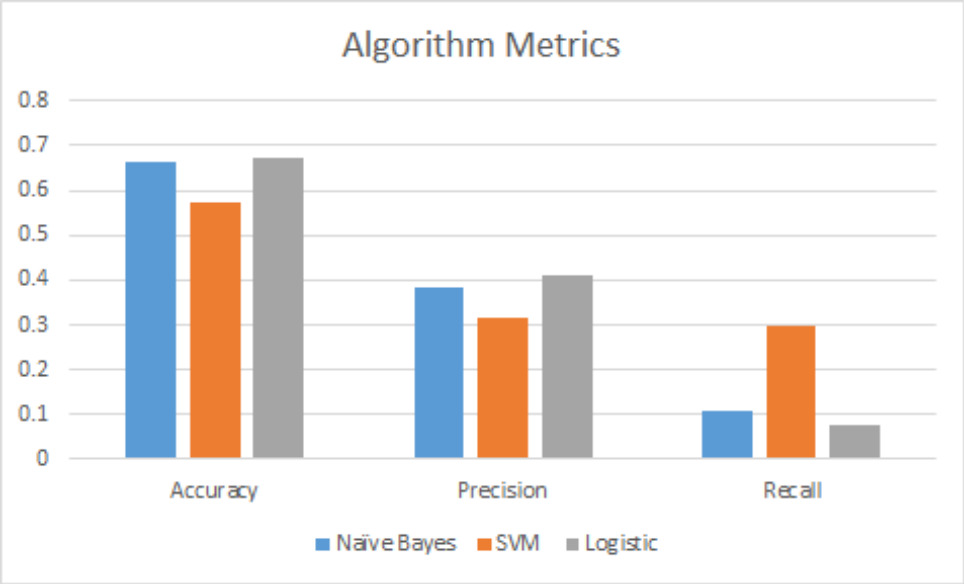
For predicting movements in VIX, we use three different supervised learning methods:

1. Naive Bayes
2. Support Vector Machines
3. Logistic Regression with PCA

RESULTS BASED ON RESEARCH PAPER:

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.



- We see that Logistic Regression has the best accuracy and precision at 67% and 41% respectively
- 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.

TYPE OF MACHINE LEARNING PROJECT:

The following problem statement is a Regression type problem.

Logistic regression is the best fit model. It is a model that measures the relationship between a categorical binary dependent variable, which we have taken to be whether or not the VIX has increased by a certain amount, and the independent variables (the features of our dictionary, from the tweets).
It estimates the probabilities of the categorical variable using logistic function.

PERFORMANCE METRICS:

Accuracy and Precision of model