Predicting Stock Volatility (COMP3125 Individual Project)

Gurucharan Naikar
School of Computing and Data Science,
Wentworth Institute of Technology

***Abstract*— This project uses data science to predict stock market volatility using machine learning. We use daily stock data from Yahoo Finance for S&P 500 companies from 2020 to 2025. By creating features like moving averages and daily returns, we characterize how the market behaves. We then use a K-Nearest Neighbors (KNN) model to classify stocks as either "High Volatility" or "Low Volatility" for the upcoming month. Our process includes cleaning the data, engineering features, and testing how well the model works using tools like confusion matrices.

Keywords*— *Stock Volatility, K-Nearest Neighbors, Machine Learning, Financial Analysis, Classification

I. Introduction (*Heading 1*)

The stock market is defined by volatility—the degree of variation in a trading price series over time. For investors and financial analysts, the ability to anticipate whether a stock will be "highly volatile" or "stable" in the near future is critical for risk management and portfolio optimization. High volatility can represent both a significant risk of loss and a significant opportunity for profit, depending on the investment strategy employed. Because financial markets are complex and noisy, predicting these shifts requires robust data analysis rather than intuition alone.

This project aims to build a data-driven classification model to predict stock volatility by analyzing historical patterns. The primary objective is to determine if past trading metrics can accurately forecast future risk. To achieve this, we will first perform a descriptive analysis to understand how daily price changes and trading volume differ between historically stable "blue-chip" stocks and high-growth tech stocks. We will then explore if there is a visible correlation between a stock's average daily trading volume and its subsequent volatility classification. Finally, the core predictive goal is to train a K-Nearest Neighbors (KNN) model to classify a stock as "High Volatility" or "Low Volatility" for the upcoming month based on data from the previous month.

II. Datasets

A. Source of dataset (Heading 2)

The data for this project was obtained from **Yahoo Finance**, a widely recognized and credible provider of financial market data. We accessed this data programmatically using the yfinance Python library, which scrapes official historical market data. The dataset was generated dynamically by pulling daily trading records for selected companies listed on the S&P 500 index. This method ensures the data is up-to-date and reflects actual market conditions.

Example: XXXX

B. Character of the datasets

The dataset covers a five-year period from 2020 to 2025 to capture different market conditions, including the post-pandemic recovery and recent economic shifts. The data consists of time-series records where each row represents a single trading day for a specific stock. The raw dataset includes the following features:

Feature Name	Unit	Description
Date	DateTime	The specific trading day.
Open	USD ($)	The price of the stock when the market opened.
High	USD ($)	The highest price reached during the day.
Low	USD ($)	The lowest price reached during the day.
Close	USD ($)	The final price when the market closed.
Volume	Count	The total number of shares traded that day.

To make this data ready for analysis, significant data cleaning and feature engineering were required. We removed rows with missing data, which usually correspond to market holidays. Furthermore, because raw stock prices vary wildly, for example, one stock costs $10 and another $1000, raw prices are not good for machine learning. We engineered new features to solve this, including a 7-Day Moving Average of Volume and Daily Return Percentages. Finally, we have created a new categorical target variable called Volatility Class. This was generated by calculating the standard deviation of returns for each month; months in the top 25th percentile of variance were labeled as High Volatility, and the rest as Low Volatility.

III. Methodology

This project treats stock prediction as a Supervised Classification problem. The goal is not to predict the exact price (regression), but to classify the *state* of the stock (High Risk vs. Low Risk). We chose this approach because financial regimes often cluster together; stocks behaving similarly to "high volatility" neighbors in terms of volume and price swings are likely to be volatile themselves.

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

A. Data Processing and Tools: The entire analysis pipeline is built using Python. We utilize Pandas and NumPy for data manipulation, specifically for resampling daily data into monthly volatility metrics. Seaborn and Matplotlib are used for exploratory data analysis to visualize the correlations between trading volume and price variance.

B. The Model (K-Nearest Neighbors): We selected the K-Nearest Neighbors (KNN) classifier for this task. KNN is a non-parametric method that makes predictions based on the majority class of the 'k' nearest data points in the feature space. We utilize the `Scikit-Learn` library to implement this model. Before training, we apply `StandardScaler` to normalize our input features (Volume and Daily Returns). This is a critical step because KNN calculates distances between points; if one feature has huge numbers (like Volume) and another has small numbers (like Returns), the model will be biased without scaling.

C. Model Evaluation: Since identifying high-risk stocks is critical for financial safety, simple accuracy is not enough. We will evaluate the model using a Confusion Matrix to visualize exactly how many high-volatility periods we correctly predicted versus how many we missed (False Negatives). We will also generate a ROC Curve and calculate the AUC (Area Under the Curve) to measure the model's ability to distinguish between the two classes at different threshold settings. Finally, we will use a chronological split for testing, ensuring we do not train the model on future data to predict the past.

[7].

IV. Results

In this section, present your findings using an appropriate method, such as equations, numerical summaries, or visualizations like charts and graphs. Clearly explain all results and provide guidance on how to interpret them. If any unexpected results arise, discuss possible reasons or contributing factors. To improve clarity and organization, consider using subsections (e.g., A, B) to separate different aspects of your results.

Example: After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. Result A

Example: XXX

1) For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

2) For papers with less than six authors: To change the default, adjust the template as follows.

a) Selection: Highlight all author and affiliation lines.

b) Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

c) Deletion: Delete the author and affiliation lines for the extra authors.

B. Results B

Example: Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

C. Results C

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

TABLE I. TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a] Sample of a Table footnote. (*Table footnote*)

Fig. 1. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

V. Discussion

Every method/project has its shortage or weakness. Please discuss the unsatisfied results in your project. And discuss the feasible suggestions of future work to revise/improve your result.

Example: xxx

VI. Conclusion

In this part, you should summarize your project. What important results did you find for your topic and what's the effect of this result on the real-world?

Example: xxx

Acknowledgment *(Heading 5)*

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

Use the IEEE format for the citation. The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..." Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord "Format" pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.