Stock Predictor

|  |  |  |
| --- | --- | --- |
| Daniel Hodges  Knoxville, USA  Dhodge12@vols.utk.edu | Robert King  Knoxville, USA  rking69@vols.utk.edu | Kush Patel  Knoxville, USA  pkx959@vols.utk.edu |
|  |  |  |

*Abstract*—This document is a model and instructions for LATEX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. \*CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.

*Index Terms*—component, formatting, style, styling, insert

# I. INTRODUCTION

# In the world of stock trading, accurately predicting stock price movements is a critical component of successful investment strategies. Traditional stock price forecasting is challenging due to the complex and fluctuating nature of financial markets, where prices are influenced by many different factors. For investors and analysts, understanding the potential trajectory of stock prices can provide a significant advantage, helping them make informed decisions and mitigate financial risks. This project addresses the need for an effective and accessible tool to forecast stock trends by developing a Stock Next Day Closing Price Predictor. By harnessing historical stock data and employing a machine learning model, specifically a Random Forest Regressor, this project aims to provide accurate next-day stock price predictions. The goal is to give traders and analysts a reliable prediction model that not only forecasts the closing prices but also classifies expected stock trends, ultimately aiding in better decision making and strategy development in the financial market.

# II. Data Exploration

## A. Dataset Summary

The dataset used in this project is made up of 2,408 daily entries of Berkshire Hathaway stock data, containing many important financial attributes: Date, Open, High, Close, Adjusted Close, and Volume. Each of these attributes is important in trend analysis and model training. The Date captures the trading date, which allows us to predict stock behavior over time. Open represents the initial stock price at market open, while High and Low indicate the maximum value and minimum price movements throughout any given day. The Close value marks the price of the stock at market close for the day. Adjusted Close accounts for dividends and stock splits, offering a more accurate reflection of value. Volume tracks the number of shares traded, which often has to do with stock volatility.

## B. Data Cleaning and Preprocessing

In order to have an effective training process, there were several steps taken to preprocess the data to be used. First, the Date field was converted from a string to a datetime format, helping in chronological sorting and easy trend extraction. Although our dataset was complete and was not missing any values, we have implemented missing data handling strategies such as row removal or interpolation to deal with any future data inconsistencies. To create a target variable for the next-day price prediction, the Close column was shifted by one day to produce a new column called Next Day Close. This shift allows our model to use today’s prices to predict tomorrow’s closing prices. We remove and filter out any null values tht result from this shift to make sure our data is complete to properly train the model.

## C. Exploratory Data Analysis

Exploratory Data Analysis was conducted in order to uncover trends, correlations, and significant patterns. Visualizing the Close and Adjusted Close prices over time reveals stock price trends and potential seasonal cycles in the market. These patterns can provide insight into stock behavior in response to annual events or shifts in the economy. Volume analysis highlights days with higher trading activity, which could be seen to correspond to important news or an increase in investor interest. Examining the correlations between Open, High, Low, and Close helps to assess the relationships within the price movements throughout the day, which could help in the model’s ability to capture daytime volatility.

# III. Baseline Solution

1. Introduction

For our baseline solution, we implemented a Random Forest Regressor to predict the next day’s closing price of the Berkshire Hathaway stock. Random Forest is an ensemble learning method that makes multiple decision trees during training and average their predictions to provide a more efficient and accurate forecast. This approach is effective for regression tasks like predicting stocks.

1. Existing Solutions

Predicting stock prices has traditionally been

approached using different methods, ranging from simple statistical techniques like Moving Averages (MA) and Exponential Moving Averages (EMA) to more complex machine learning models such as Long Short-Term Memory (LSTM) networks and Support Vector Machines (SVM). While statistical models focus on historical price trends and smoothing fluctuations over time, “moving averages smooth the data, but they may noy capture more complex patterns or sudden changes in a time series” (). Machine learning models, such as LSTM, are made to capture time dependencies but “are prone to overfitting and require large datasets to generalize well” (). Given the structured, tabular nature of stock data and the need for a model that balances accuracy and interpretability. Random Forest Regression was selected as the baseline model due to its robustness, ability to handle non-linear data, and resistant to overfitting.

1. Baseline Selection

The Random Forest Regressor (RFR) was chosen as the

baseline model for multiple reasons. First, Random Forest is an effective method that fits several decision trees and average their predictions, making it more robust and accurate than a single decision tree. Random forest is particular good at handing non-linear data between variables without requiring extensive tuning of the hyperparameters. This makes the RFR an ideal candidate for capturing the relationships between stock price variables such as “highs” and “lows”. Unlike some models like SVM or Neural Networks, which may require extensive tuning and is prone to overfitting with small datasets. Random Forest offers a strong balance between performance and simplicity.

1. Implementation

The implementation of the Random Forest model has

several steps. First, the historical stock data for Berkshire, consisting of open, close, adjacent close, and low prices, was preprocessed. We then created a target variable, the Next Day’s Closing Price, by shifting the “Close” column down by one day. The dataset was then split into a testing (8%) and training (92%) sets, ensuring that the model could be tested on unknown data.

The RFR was configured with 100 trees (n\_estimators = 100), with default hyperparameters for maximum depth and minimum sample splits. The random forest classifier is a combination of tree classifiers such that each tree depends on the value of the random vector sampled independently and with the same distribution for all trees in the forest. The model was on trained on the training set, using features like open, close, adjacent close, high, low to predict the next day’s closing price. The random state was set to 24 to ensure reproducibility.

1. Baseline Performance

The model was tested on its ability to predict the next

Day’s closing price and classify stock trends as either “Uptrend” or “Downtrend”. The predicted closing prices were evaluated against the actual next-day prices. We also implemented a prediction difference to see how the model’s output aligned with the real data. Generally, the prediction differences were modest, suggesting that the model is effectively capturing the stock price movements. However, for some days there was more of a deviation between the predicted and actual closing prices. This behavior is common among models that rely on long term data, because the model doesn’t account for sudden external situations, such as market news or economic changes.

In addition to predicting closing prices, the model’s trend classification performance analyzed by comparing the “Predicted Trend” with the “Actual Trend” for each day. While the model aligned with the actual trend direction during stable periods, sometimes there was a misclassification trends in volatile conditions, predicting a downtrend trend when the actual trend was uptrend, and vice versa. For example, on 2015-07-31, the model predicted a Uptrend movement, however, it was a downtrend movement. These misclassifications indicate that while the baseline model provides valuable insight into stock price movement trends, it may benefit from additional features, such as technical indicators or sentiment data, to improve its performance during unpredictable market conditions. Overall, the model’s ability to track trends during stable periods demonstrates its potential, yet enhancements could make it more robust in dynamic market environments.

## A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and

## B. Units

* Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter”, not “webers/m2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.
* Use a zero before decimal points: “0.25”, not “.25”. Use

“cm3”, not “cc”.)

## C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

4. Baseline Model(1)

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

## D. LATEX-Specific Advice

Please use “soft” (e.g., \eqref{Eq}) cross references instead of “hard” references (e.g., (1)). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don’t use the {eqnarray} equation environment. Use {align} or {IEEEeqnarray} instead. The {eqnarray} environment leaves unsightly spaces around relation symbols.

Please note that the {subequations} environment in LATEX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you’ve discovered a new method of counting.

BIBTEX does not work by magic. It doesn’t get the bibliographic data from thin air but from .bib files. If you use BIBTEX to produce a bibliography you must send the .bib files.

LATEX can’t read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LATEX does not have precognitive abilities. If you put a \label command before the command that updates the counter it’s supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a \label command should not go before the caption of a figure or a table.

Do not use \nonumber inside the {array} environment. It will not stop equation numbers inside {array} (there won’t be any anyway) and it might stop a wanted equation number in the surrounding equation.

## E. Some Common Mistakes

* The word “data” is plural, not singular.
* The subscript for the permeability of vacuum *µ*0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
* In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
* A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
* Do not use the word “essentially” to mean “approximately” or “effectively”.
* In your paper title, if the words “that uses” can accurately replace the word “using”, capitalize the “u”; if not, keep using lower-cased.
* Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.
* Do not confuse “imply” and “infer”.
* The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
* There is no period after the “et” in the Latin abbreviation

“et al.”.

* The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

An excellent style manual for science writers is [7].

## F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

## G. Identify the Headings

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

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

## H. Figures and Tables

*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 copya |  |  |

aSample of a Table footnote.

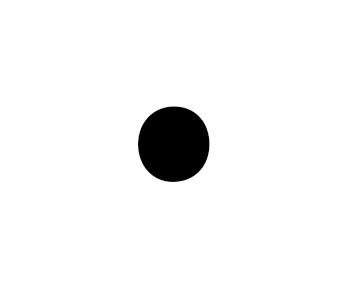


Fig. 1. Example of a figure caption.

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”.

### ACKNOWLEDGMENT

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

Please 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 *...*”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

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.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

### REFERENCES

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.
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.

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 the template text from your paper may result in your paper not being published.