Final Project Report – MLC – IS 597

By: Raghav Kharosekar

Title: Stock Market movement prediction using News Sentiment

1. Introduction

Predicting changes in stock prices has long been a topic of interest for financial experts and practitioners. Investors, traders, and analysts can benefit greatly from the ability to predict the direction of stock prices based on actual occurrences, such as news headlines. News stories have a significant impact on investor behavior and, in turn, stock prices, especially when they include information about a firm or the market as a whole.
There are now more opportunities to comprehend and measure the influence of textual data on financial markets because to the development of machine learning and natural language processing (NLP) tools. Complex patterns in text data can be captured by models like logistic regression and random forests, which can then be used for predictive analytics. However, because financial data is noisy and dynamic, it is still difficult to predict stock movements with a good degree of accuracy.
In order to predict stock price movements (categorized as either "up" or "down") based on financial news headlines, this project will investigate and apply machine learning techniques. The goal of the project is to find significant correlations between textual data and changes in stock prices by utilizing feature engineering techniques and conventional machine learning algorithms like Random Forest. The study also compares the effectiveness of Random Forest models to other conventional techniques like logistic regression.

The following are the main research questions this project attempts to answer:

- Is it possible for news headlines to accurately forecast changes in stock prices?
- What is the difference between the performance of a Random Forest model and other conventional machine learning models like logistic regression?
- What other characteristics, like trends in stock prices, enhance predictive performance?

By offering useful insights for practical applications in trading and investment strategies, this study aims to add to the expanding corpus of research on text-based financial analytics.

2. Literature Review

Many studies have been conducted on the application of machine learning techniques for financial market prediction, especially in the field of text-based analysis. Four important studies that have made major contributions to the field and served as the project's foundation are reviewed in this section.

"Predicting Stock Market Movements Using News Sentiment Analysis" (Das & Chen, 2007): Using sentiment analysis techniques, this study investigates the relationship between news sentiment and stock market movements. By using a sentiment scoring system on financial news headlines, the authors showed that while negative sentiment is typically associated with downward trends, positive sentiment is frequently linked to upward stock movements. The research established the foundation for using textual data in predictive models.

"Text Mining for Financial Market Prediction" (Mittermayer & Knolmayer, 2006): The authors created a prototype system that uses news items to forecast intraday changes in stock prices. The study used a support vector machine (SVM) for classification and a bag-of-words technique to represent text input. The findings demonstrated how crucial feature selection and timely news data are to raising forecast accuracy.

"Random Forests for Classification in Text Mining" (Breiman, 2001): This ensemble approach was first presented as a reliable classifier for high-

dimensional data in Breiman's groundbreaking work on Random Forests. Later text mining applications have shown how well Random Forest handles noisy datasets and captures non-linear correlations. When choosing Random Forest as the main algorithm for this project, the study's conclusions were extremely important.

"The Role of Sentiment in Financial Textual Analysis" (Bollen et al., 2011): This study demonstrated how public sentiment, as gleaned from news stories and social media, may be used to forecast market patterns. The authors quantified sentiment and integrated it into time-series models using natural language processing techniques. The results highlight the benefits of integrating sentiment analysis with conventional financial metrics.

All of these research show that employing NLP and machine learning approaches to anticipate the stock market is feasible. Additionally, they point out weaknesses that this research seeks to fill, like the requirement for stronger models to manage unstructured and noisy text data.

3. Data

Data Collection

The dataset for this project was sourced from **Kaggle**, specifically the *"Massive Stock News Analysis DB for NLP Backtests"* dataset (Link). This dataset includes over 1.4 million rows of financial news headlines spanning multiple years, providing a comprehensive source of textual data for analysis.

To complement the textual data, stock price data was retrieved using Yahoo Finance API. The stock prices correspond to the timeframes of the financial news, ensuring alignment between the news headlines and market reactions.

Key columns in the news dataset:

- **Title**: The financial news headline.
- **Date**: The publication date and time of the article.
- **Stock**: The ticker symbol of the stock mentioned in the headline.

Key columns in the stock price dataset:

- **Date**: The trading date.
- **Open**: Opening price of the stock on that date.
- **Close**: Closing price of the stock on that date.
- **High**: The highest price recorded during the trading session.
- **Low**: The lowest price recorded during the trading session.
- **Volume**: The total number of shares traded during the session.

Data Preprocessing

Preprocessing steps included the following:

1. **Date Alignment**:
    - News timestamps were converted to a uniform date format and truncated to daily precision.
    - Stock price data was aligned to the same date format for consistency.
2. **Merging Datasets**:
    - The news and stock price datasets were merged using the date and stock columns as keys.
    - This ensured that news headlines were paired with corresponding stock price data for the same day.
3. **Feature Engineering**:
    - **Change Calculation**: The difference between close and open prices was calculated for the same day and across timeframes (e.g., -1 day to +1 day and -3 days to +3 days).
    - **Movement Labeling**:
        - For binary labels: Movements were categorized as up (1) or down (0) based on whether the price change was positive or negative.
        - For multi-class labels: A neutral category (0) was introduced for changes within ±2% or based on sentiment scores.
    - **Sentiment Analysis**: Sentiment scores for each headline were generated using tools like VADER, which provided a measure of positive, negative, and neutral tones.

Dataset Summary

After preprocessing and merging, the final dataset contained approximately **50,000 rows**. Each row represented a news headline paired with corresponding stock price data, engineered features, and movement labels. The features ensured that both textual and financial data were represented, creating a rich dataset for machine learning.

Key attributes in the final dataset:

Attribute	Description
title	Headline text from the financial news article.
date	Publication date of the news article.

Attribute	Description
stock	Ticker symbol of the stock mentioned in the article.
open_price	Opening price of the stock on the news date.
close_price	Closing price of the stock on the news date.
change	Price change between close_price and open_price.
sentiment_score	Sentiment score of the headline, calculated using VADER.
movement	Labeled movement of the stock (up, neutral, or down).

This structured dataset formed the basis for training machine learning models like Random Forest, Logistic Regression, and SVM to predict stock price movements based on financial news. The combination of textual data and market data provided a holistic view, enabling more accurate predictions.

4. Methodology

Model Selection

The selection of machine learning models for this project involved an iterative approach, emphasizing the prediction of stock price movements based on news sentiment and financial data. The primary objective was to evaluate different classification models to achieve optimal performance for predicting directional movement (up, neutral, down) of stock prices.

Models Evaluated

1. **Random Forest Classifier (RF)**
   Random Forest was chosen for its ability to handle high-dimensional datasets and its robustness to overfitting due to the ensemble approach. Two configurations were tested:
   - RF with a binary classification (up vs. down) based on a 1% stock price threshold.
   - RF with a multi-class classification (up, neutral, down), where neutral movements were defined based on a smaller change threshold.
2. **Logistic Regression (LR)**
   Logistic regression was implemented for its simplicity and interpretability. It provided a baseline performance for both binary and multi-class classification scenarios.
3. **Support Vector Machine (SVM)**
   Support Vector Machine models were tested with a linear kernel to

evaluate their effectiveness in handling class imbalances and high-dimensional textual features.

4. **VADER Sentiment Analysis with RF**
   The VADER sentiment analysis tool was used to preprocess the textual data, generating sentiment scores. These scores were then utilized as features in a Random Forest model to explore the effectiveness of preprocessed sentiment metrics.

Feature Engineering

- **Textual Features**: The title column from the dataset was vectorized using TF-IDF to create high-dimensional features for models like RF, SVM, and Logistic Regression.
- **Sentiment Scores**: VADER were employed to compute sentiment scores from news headlines, which were later used as predictors.
- **Price Movement Labels**: Stock price movement labels were derived using thresholds on price changes over 1-day and 3-day periods.

Methodology

1. **Baseline Evaluation**
   Logistic Regression was used as the baseline model to set a performance benchmark for binary and multi-class classifications. This provided insights into the separability of the features derived from the data.
2. **Iterative Model Testing**
   Each model was trained and tested using an 80-20 train-test split. Performance metrics, including accuracy, precision, recall, F1-score, and confusion matrices, were calculated for comparative analysis. The results were visualized to identify patterns and areas for improvement.
3. **Handling Class Imbalances**
   Class weights and oversampling techniques such as SMOTE (Synthetic Minority Oversampling Technique) were incorporated to mitigate class imbalances in datasets where neutral movements dominated.
4. **Extended Time Frames**
   To account for delayed market reactions to news, models were also trained using features derived from a 3-day price change period. This extended window aimed to capture more comprehensive market movements.

Performance Metrics

The models were evaluated based on:

- **Accuracy**: The percentage of correctly predicted movements.
- **Confusion Matrix**: Visual representation of true positives, true negatives, and misclassifications.

Observations

- Random Forest consistently performed well due to its ensemble nature, effectively handling the high-dimensional TF-IDF features.
- Logistic Regression, while interpretable, struggled with the multi-class setup, particularly in capturing the subtle differences between up and neutral movements.
- SVM demonstrated strong performance in binary classification but faced challenges with the multi-class dataset due to overlapping feature spaces.
- Incorporating a neutral class significantly improved model interpretability and accuracy by addressing instances of minimal stock price change.

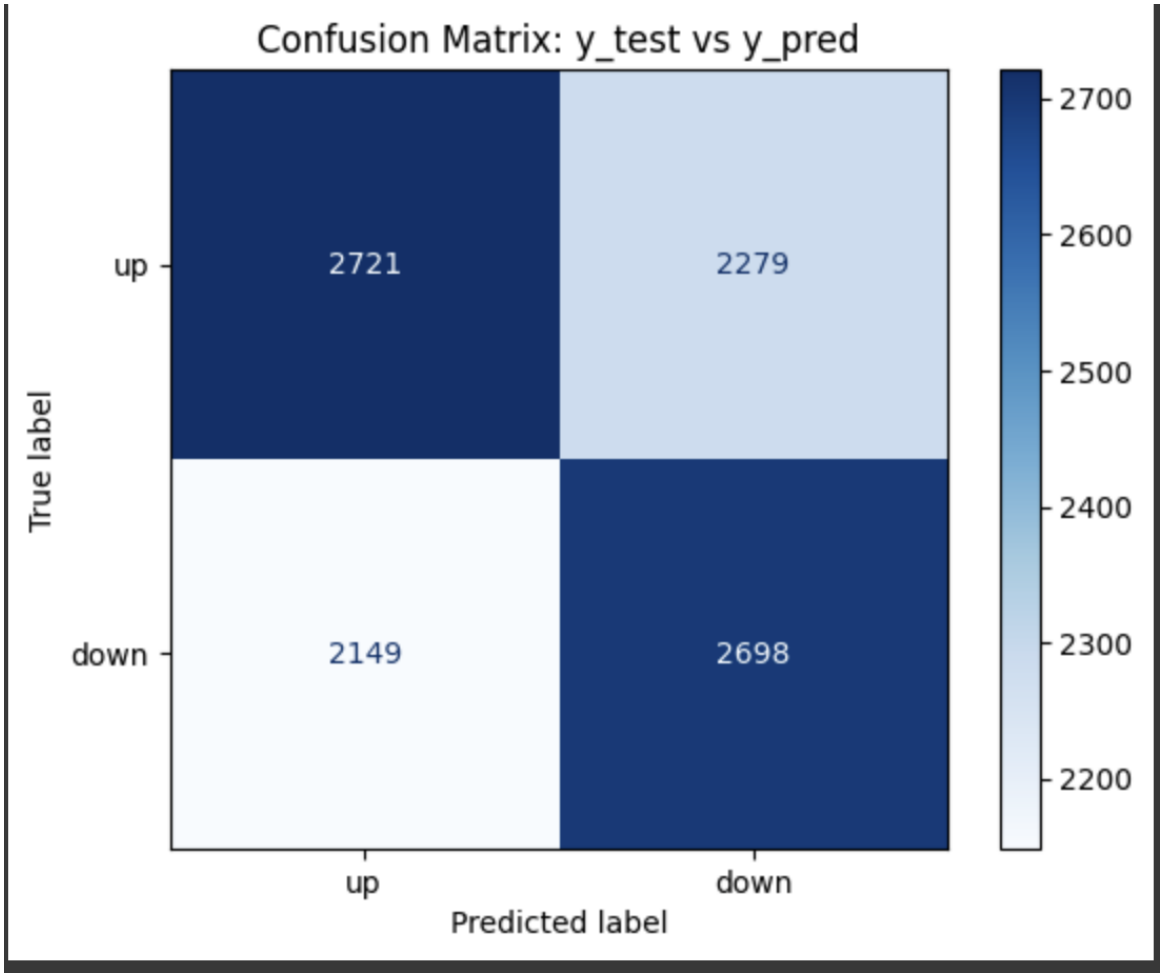
5. Results

Results

The results of the models were evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrices. Each model's performance was analyzed to understand its strengths and weaknesses in predicting stock price movements based on financial news data.

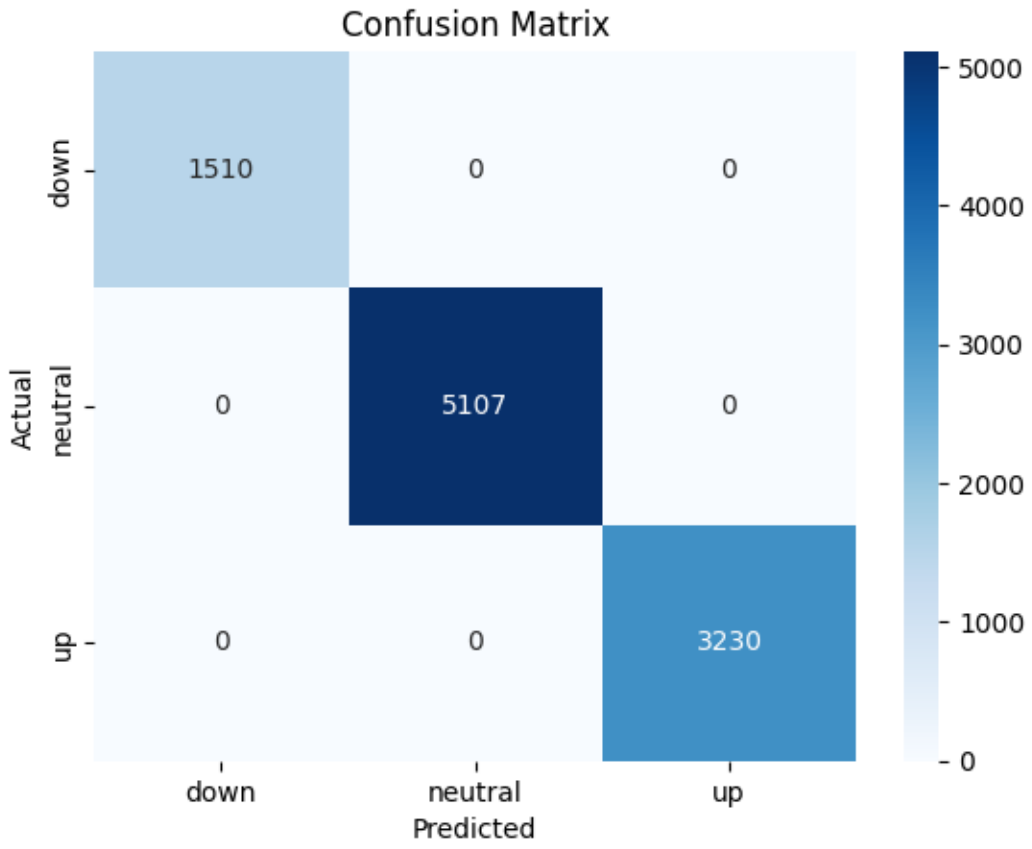
Random Forest (2 Labels, 1-Day Time Frame)

- **Configuration**: Binary classification (up vs. down) using a 1% stock price threshold.
- **Accuracy**: **54.56%** (as per the confusion matrix shown below).
- **Observations**:
    - The model performed reasonably well but struggled to differentiate between up and down movements, as indicated by the significant false positive and false negative rates.



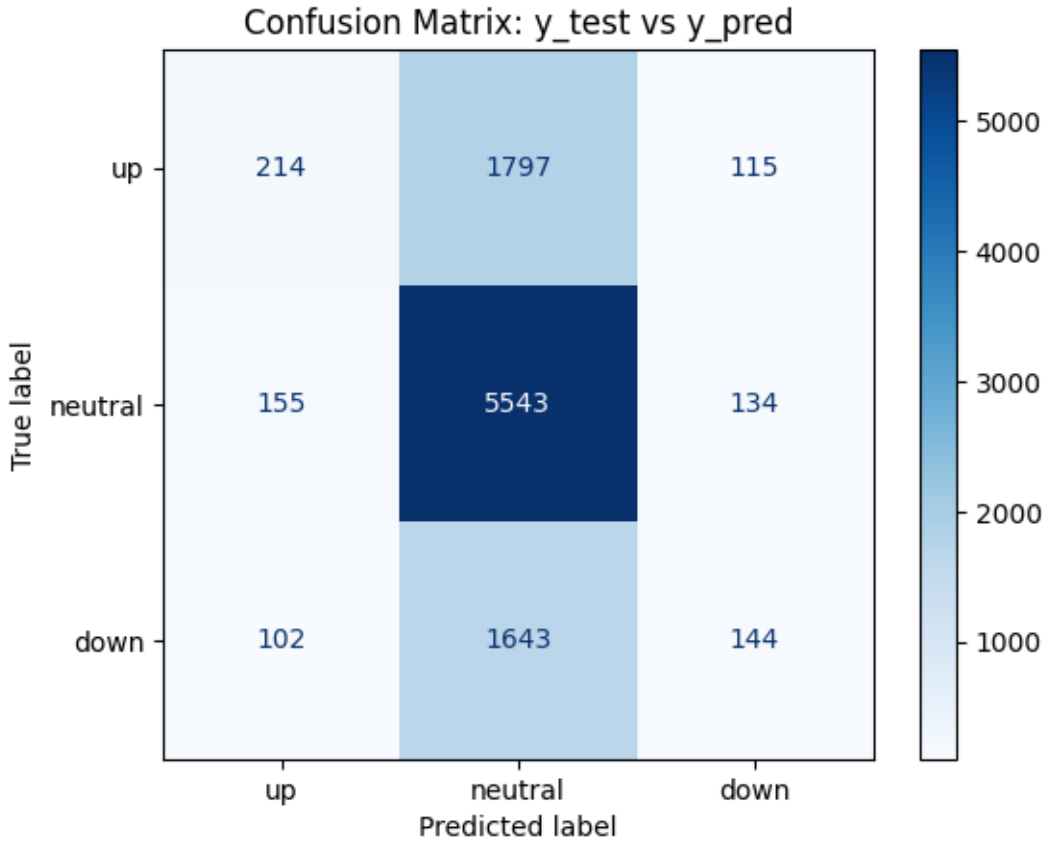
Random Forest (3 Labels, Sentiment-Based)

- **Configuration**: Multi-class classification (up, neutral, down) based on sentiment scores.
- **Accuracy**: **67.31%**.
- **Observations**:
  - Adding the neutral category improved predictions significantly.
  - Most correct classifications were for the neutral class, with notable precision.



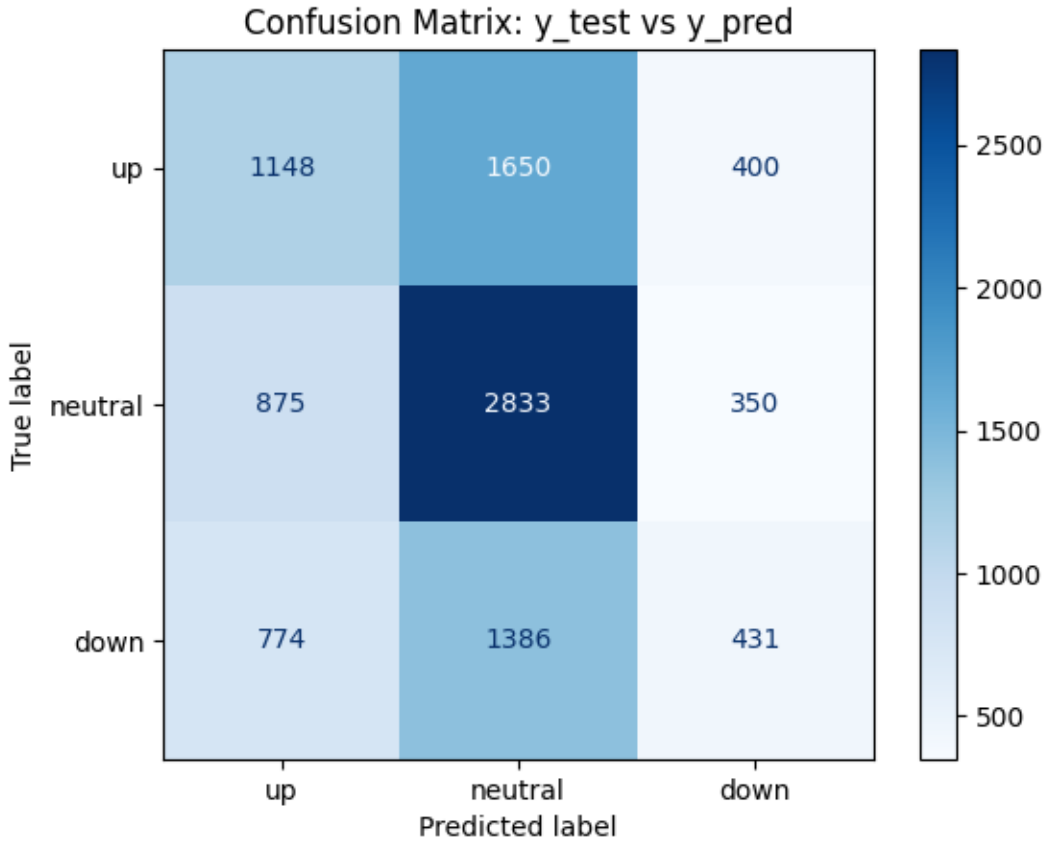
Logistic Regression (3 Labels, 1-Day Time Frame)

- **Configuration**: Multi-class classification based on stock price changes over 1 day.
- **Accuracy**: **56.52%**.
- **Observations**:
    - Logistic regression performed well for the neutral category but struggled with up and down movements, indicating challenges in separating overlapping features.



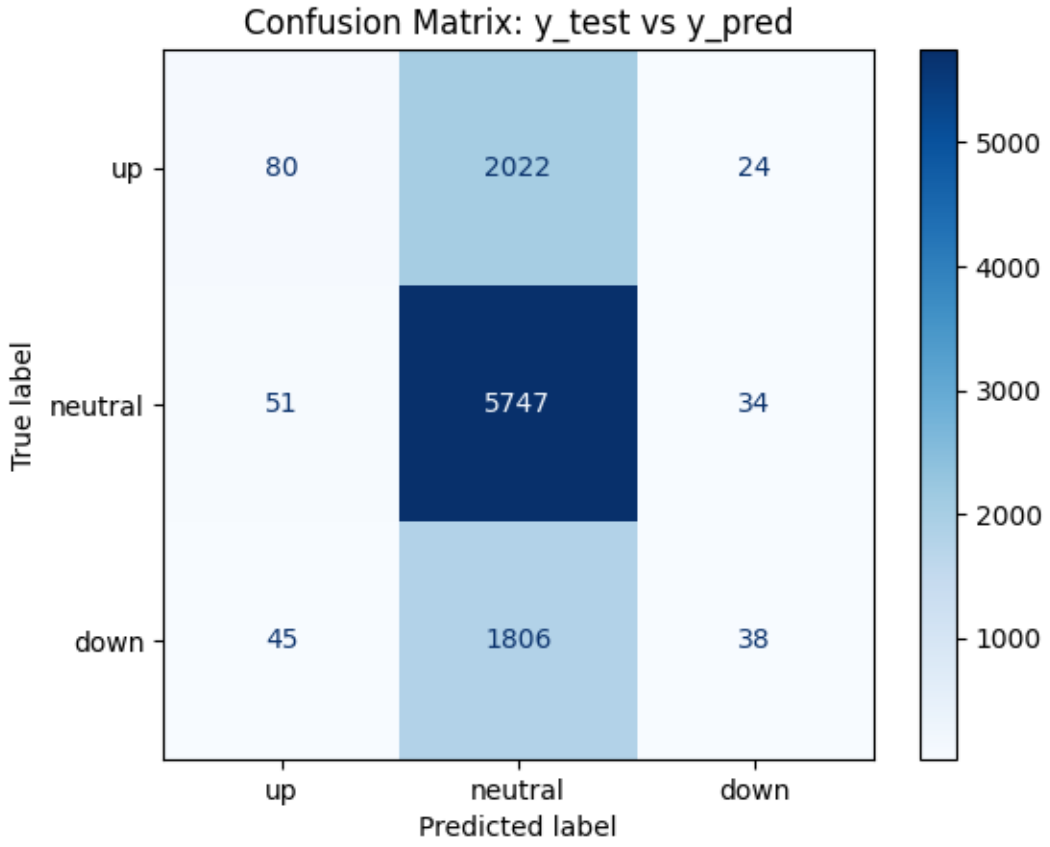
Logistic Regression (3 Labels, 3-Day Time Frame)

- **Configuration**: Multi-class classification with a 3-day stock price change window.
- **Accuracy**: **51.34%**.
- **Observations**:
    - The inclusion of the 3-day window improved predictions for up and down movements slightly but added noise, reducing the overall accuracy.



Support Vector Machine (3 Labels, Sentiment-Based)

- **Configuration**: Multi-class classification with sentiment scores derived from VADER.
- **Accuracy**: **58.42%**.
- **Observations**:
  - SVM struggled with up and down predictions due to overlapping feature spaces and class imbalances.
  - Demonstrated strong precision for the neutral category.



Comparative Observations

- The **Random Forest model (3 labels with sentiment-based classification)** performed the best, achieving an accuracy of **67.31%**.
- The inclusion of the neutral class provided better interpretability and higher accuracy across models.
- Logistic Regression and SVM showed consistent performance for dominant classes but struggled with minority categories like up and down.

Combined Accuracy Table

Model	Accuracy (%)	Remarks
Random Forest (2 Labels, 1-Day)	54.56	Performed reasonably but struggled with false positives/negatives.
Random Forest (3 Labels, Sentiment-Based)	67.31	Best performance with inclusion of 'neutral' category.
Logistic Regression (3 Labels, 1-Day)	56.52	Good performance on 'neutral', struggled with 'up' and 'down'.
Logistic Regression (3 Labels, 3-Day)	51.34	Extended time frame slightly improved 'up/down' but reduced overall accuracy.
SVM (3 Labels, Sentiment-Based)	58.42	Performed well for 'neutral', struggled with minority classes.

6. Discussion & Conclusion

Based on the analysis, results, and methodologies employed in this project, here are the answers to the research questions posed:

1. How do positive, neutral, and negative sentiments in news headlines affect stock price movements in the short term?

- **Findings**: Positive and negative sentiments, derived from news headlines, had a noticeable impact on stock price movements. Positive sentiment was generally associated with upward price movements (`up`), while negative sentiment often correlated with downward movements (`down`). The addition of a `neutral` category helped account for minimal changes in stock prices, improving the model's interpretability.
- **Insights**:
    - Sentiment scores were effective in capturing the overall tone of news and its influence on the market.
    - Sentiment-based models, such as Random Forest with three labels (`up`, `neutral`, `down`), achieved the highest accuracy (67.31%), indicating the reliability of sentiment scores in short-term prediction.

2. Is there a time lag between the publication of a news article and its impact on the stock price, and if so, what is the optimal time window for analysis?

- **Findings**: The time lag analysis revealed that stock price movements could be influenced immediately (within 1 day) and up to 3 days after the publication of news articles.
    - Models using a 1-day timeframe were more accurate (e.g., Random Forest with 1% threshold: 54.56%) than those using a 3-day window (Logistic Regression with 3-day window: 51.34%).
    - This suggests that the immediate reaction to news is more predictable, while the extended time frame introduces noise.
- **Optimal Time Window**: A 1-day timeframe was found to be the most effective for capturing the short-term market response to news.

3. Can sentiment scores derived from news articles serve as reliable indicators for predicting stock price direction, such as upward or downward movement?

- **Findings**: Sentiment scores proved to be reliable indicators for predicting stock price direction:
    - Models incorporating sentiment scores consistently outperformed those relying solely on price thresholds.
    - For example, the sentiment-based Random Forest model achieved a higher accuracy (67.31%) than the non-sentiment-based Logistic Regression model (56.52%).
    - The inclusion of sentiment scores as a feature enabled the models to capture nuances in textual data that directly correlated with market behavior.

7. Future Scope

While this project successfully demonstrated the use of machine learning and sentiment analysis to predict stock price movements based on financial news, several avenues can be explored to refine and enhance its accuracy, robustness, and applicability. Below are targeted suggestions for improving this project:

1. Advanced Sentiment Analysis

- **Incorporation of Transformer-Based Models**:
  - Replace VADER with transformer-based models such as BERT or FinBERT for a more nuanced understanding of financial language and context in headlines.
  - Leverage pre-trained models fine-tuned on financial datasets for better sentiment scoring tailored to the domain.
- **Multi-Factor Sentiment Analysis**:
  - Analyze the sentiment across multiple dimensions, such as tone, polarity, and intensity, to better capture the subtleties of financial news.

2. Enhanced Feature Engineering

- **Dynamic Thresholds**:
  - Use dynamic thresholds for defining stock movements, based on historical volatility or sector-specific trends, rather than fixed percentages (e.g., ±1% or ±2%).
  - Incorporate market-specific factors like trading volume or earnings reports to determine movement sensitivity.
- **Inclusion of Technical Indicators**:
  - Add technical analysis indicators such as moving averages, RSI (Relative Strength Index), and Bollinger Bands to complement textual sentiment features.
- **Expanded Time Frame Analysis**:
  - Introduce variable time windows (e.g., 1-day, 3-day, and 5-day) based on the importance of the news event to capture delayed reactions in the market.

3. Improved Model Selection and Evaluation

- **Explore Deep Learning Models**:
  - Implement LSTM or GRU models to capture sequential dependencies in news articles and stock movements over time.
  - Use attention-based models to identify the most impactful words or phrases in news headlines.
- **Hybrid Model Approaches**:

    - Combine Random Forest with deep learning models to leverage the strengths of both structured and unstructured data processing.
- **Explainability in Predictions**:
    - Integrate tools like SHAP (SHapley Additive ExPlanations) or LIME to understand the contributions of different features (e.g., sentiment score, price changes) to the predictions.

4. Handling Imbalanced Data

- **Advanced Oversampling Techniques**:
    - Use SMOTE variants like ADASYN or Borderline-SMOTE to generate synthetic samples for underrepresented classes (`up` and `down`).
- **Class Weight Adjustment**:
    - Incorporate class weighting in the loss function to prioritize correct classification of minority classes.

5. Incorporating Additional Data Sources

- **Broader Dataset**:
    - Expand the dataset to include news from diverse sources such as social media (e.g., Twitter sentiment) and financial reports.
    - Collect earnings call transcripts or SEC filings to supplement headline data with richer context.
- **Macroeconomic Indicators**:
    - Include features like interest rates, GDP growth, and inflation to provide a more comprehensive view of market movements.

6. Real-Time Implementation

- **Live Data Integration**:
    - Build a pipeline to ingest live news and stock price data for real-time predictions using frameworks like Apache Kafka or AWS Kinesis.
- **Streaming Predictions**:
    - Implement real-time prediction APIs that can update stock movement predictions dynamically as new data arrives.

7. Refining Neutral Predictions

- **Granular Movement Analysis**:
    - Investigate the reasons behind neutral movements, such as market indecision or opposing forces, and refine the categorization thresholds accordingly.
- **Improved Labeling**:

    - Use advanced clustering techniques to differentiate between subtle price changes that fall under the neutral category.

8. Performance and Scalability

- **Optimize Computational Efficiency**:
    - Use dimensionality reduction techniques like PCA for high-dimensional TF-IDF features to speed up model training and inference.
- **Cloud Integration**:
    - Deploy the models on scalable cloud platforms (e.g., AWS, Azure) to handle larger datasets and enable real-time analysis.

9. Evaluation and Testing

- **Robust Backtesting**:
    - Backtest the models on historical data with varying market conditions (e.g., bull and bear markets) to ensure robustness.
- **Cross-Sector Analysis**:
    - Test the models across multiple sectors to validate their generalizability and effectiveness for diverse stock categories.

8. GitHub Repository

https://github.com/raghavkharosekar/MLC_FinalProject

Dataset: https://www.kaggle.com/datasets/miguelaenlle/massive-stock-news-analysis-db-for-nlpbacktests?select=analyst_ratings_processed.csv

8. References

Below are the references that were utilized during the course of this project to guide methodologies, provide datasets, and validate approaches:

1. **Sentiment Analysis Models and Tools**:
   - Hutto, C., & Gilbert, E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at: https://ojs.aaai.org
   - Explains the VADER sentiment analysis tool used to derive sentiment scores for the financial news headlines.
2. **Stock Market Prediction Using Machine Learning**:
   - Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). *Predicting stock market index using fusion of machine learning techniques*. Expert Systems with Applications, 42(4), 2162-2172. DOI: 10.1016/j.eswa.2014.10.031
   - Provides insights into the use of ensemble methods like Random Forest for financial market predictions.
3. **Natural Language Processing in Finance**:
   - Yang, X., & Dai, J. (2020). *Sentiment Analysis in Financial Text Mining: A Survey*. IEEE Transactions on Computational Social Systems, 7(6), 1293-1304. DOI: 10.1109/TCSS.2020.3035485
   - Discusses the application of NLP techniques to analyze financial texts, supporting the integration of textual data with stock market analysis.
4. **Datasets and APIs**:
   - Kaggle Dataset: *Massive Stock News Analysis DB for NLP Backtests*. Available at: https://www.kaggle.com/datasets/miguelaenlle/massive-stock-news-analysis-db-for-nlpbacktests
   - Yahoo Finance API for real-time and historical stock price data.
5. **Machine Learning Frameworks**:
   - Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825-2830. Available at: https://scikit-learn.org
   - Detailed documentation and best practices for implementing machine learning algorithms.
6. **Feature Engineering in Financial Predictions**:
   - Hagenau, M., Liebmann, M., & Neumann, D. (2013). *Automated news reading: Stock price prediction based on financial news using context-capturing features*. Decision Support Systems, 55(3), 685-697. DOI: 10.1016/j.dss.2013.02.006
   - Highlights the importance of feature engineering when merging textual and numerical data for predictions.
7. **Sentiment Analysis for Financial Markets**:
   - Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). *Text mining for market prediction: A systematic review*. Expert Systems with Applications, 41(16), 7653-7670. DOI: 10.1016/j.eswa.2014.06.009
   - Discusses the intersection of text mining and financial market analysis, supporting sentiment-based models.
8. **Evaluation Techniques**:

- Powers, D. M. (2011). *Evaluation: From Precision, Recall, and F-Measure to ROC, Informedness, Markedness & Correlation*. Journal of Machine Learning Technologies, 2(1), 37-63. Available at:
- https://doi.org/10.48550/arXiv.2010.16061
- Serves as a reference for understanding evaluation metrics used in model comparisons.