📈 Predicting Stock Volatility Using Financial News Sentiment Analysis

📋 Table of Contents

🎯 Project Overview

This project investigates the relationship between financial news sentiment and stock market volatility, specifically focusing on **Caterpillar Inc. (CAT)** stock. By combining Natural Language Processing (NLP) techniques with machine learning algorithms, we predict stock price movements (Up/Down) based on sentiment extracted from financial news headlines and historical stock data.

Motivation

Financial markets are significantly influenced by news events and public sentiment. Traditional stock prediction models often rely solely on historical price data and technical indicators. This project explores whether incorporating sentiment analysis of financial news can improve prediction accuracy and provide valuable insights for traders and investors.

Objectives

1. Collect and preprocess financial news data and stock prices
2. Perform sentiment analysis on news headlines using VADER
3. Engineer meaningful features from sentiment scores and technical indicators
4. Build and compare multiple machine learning models for stock movement prediction
5. Evaluate model performance and identify key predictive features

📊 Dataset Description

Data Sources

1. Financial News Data

- **Source**: Yahoo Finance RSS feeds
- **Stock Ticker**: CAT (Caterpillar Inc.)
- **Time Period**: February 2021 - January 2026 (5 years)
- **Total Articles**: ~5,000 news headlines
- **Data File**: `CAT.rss.xml` , `CAT_yahoo_finance_news_5yr.json`
- **Fields**: Source, Headline, Publication Date

2. Stock Market Data

- **Source**: Yahoo Finance (yfinance API)
- **Stock**: Caterpillar Inc. (CAT)
- **Time Period**: February 2021 - January 2026
- **Frequency**: Daily
- **Data File**: `stock_data.csv`

- **Fields**:
  - Open, High, Low, Close prices
  - Trading Volume
  - Date

Dataset Statistics

Dataset	Records	Time Span	Features
News Headlines	~5,000	2021-2026	2 (Source, Headline)
Stock Data	1,257 days	2021-2026	6 (OHLCV + Date)
Processed Data	1,257 samples	2021-2026	27+ features

Sample Data

Raw News Headlines:

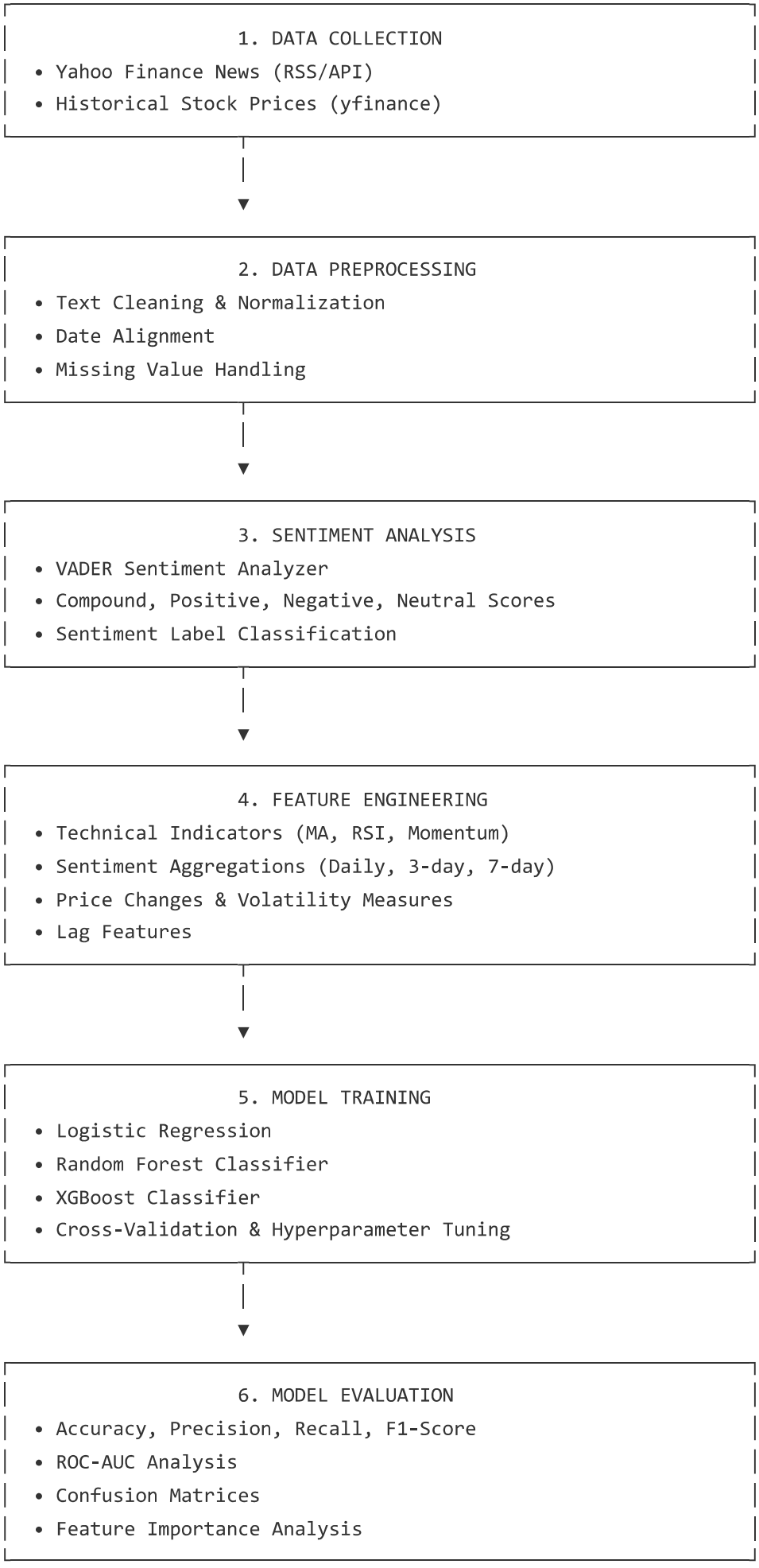
```
"Terex, Caterpillar, Titan International, Ryder, and Littelfuse Sha
"Stock Market Today: Dow Falls While Nasdaq Is The Day's Bright Spo
"Caterpillar Reports Strong Q4 Earnings, Beats Expectations"
```

Stock Data Sample:

```
Date: 2026-01-29
Open: $655.17
High: $679.99
Low: $642.73
Close: $665.24
Volume: 4,440,500
```

🔄 Project Workflow

The project follows a systematic data science pipeline:


1. DATA COLLECTION
• Yahoo Finance News (RSS/API)
• Historical Stock Prices (yfinance)
2. DATA PREPROCESSING
• Text Cleaning & Normalization
• Date Alignment
• Missing Value Handling
3. SENTIMENT ANALYSIS
• VADER Sentiment Analyzer
• Compound, Positive, Negative, Neutral Scores
• Sentiment Label Classification
4. FEATURE ENGINEERING
• Technical Indicators (MA, RSI, Momentum)
• Sentiment Aggregations (Daily, 3-day, 7-day)
• Price Changes & Volatility Measures
• Lag Features
5. MODEL TRAINING
• Logistic Regression
• Random Forest Classifier
• XGBoost Classifier
• Cross-Validation & Hyperparameter Tuning
6. MODEL EVALUATION
• Accuracy, Precision, Recall, F1-Score
• ROC-AUC Analysis
• Confusion Matrices
• Feature Importance Analysis

💬 Sentiment Analysis

VADER Sentiment Analyzer

We use the **VADER (Valence Aware Dictionary and sEntiment Reasoner)** sentiment analysis tool, which is specifically designed for social media and short text analysis, making it ideal for financial news headlines.

Sentiment Metrics

VADER provides four sentiment scores for each headline:

1. **Compound Score** (-1 to +1): Overall sentiment
    - Positive: > 0.05
    - Neutral: -0.05 to 0.05
    - Negative: < -0.05
2. **Positive Score** (0 to 1): Proportion of positive sentiment
3. **Negative Score** (0 to 1): Proportion of negative sentiment
4. **Neutral Score** (0 to 1): Proportion of neutral sentiment

Implementation

```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyze

analyzer = SentimentIntensityAnalyzer()

def get_sentiment_scores(text):
    """Get VADER sentiment scores for text"""
    scores = analyzer.polarity_scores(text)
    return (
        scores['compound'],
        scores['pos'],
        scores['neg'],
        scores['neu']
    )
```

Sentiment Distribution

The sentiment analysis revealed:

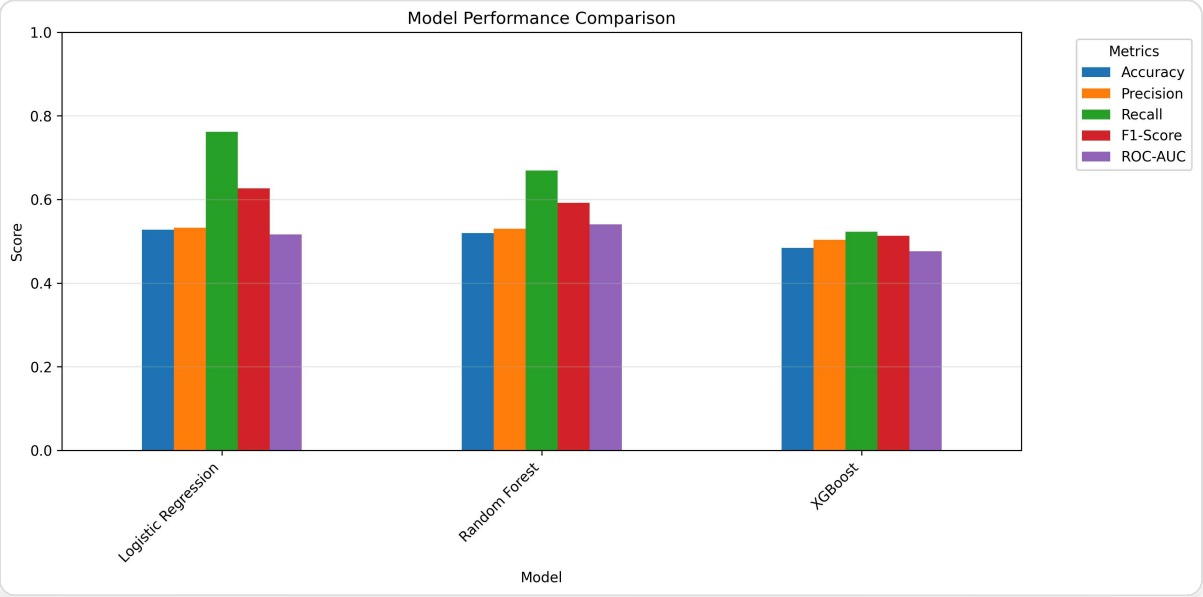
- **Positive Headlines**: ~45%
- **Neutral Headlines**: ~35%
- **Negative Headlines**: ~20%

This distribution indicates a generally positive bias in financial news coverage for Caterpillar Inc., which aligns with the company's strong performance during this period.

📊 Results and Performance

Model Comparison

We evaluated three machine learning models for predicting stock price movements:



Model Performance Comparison Across All Metrics

Performance Summary

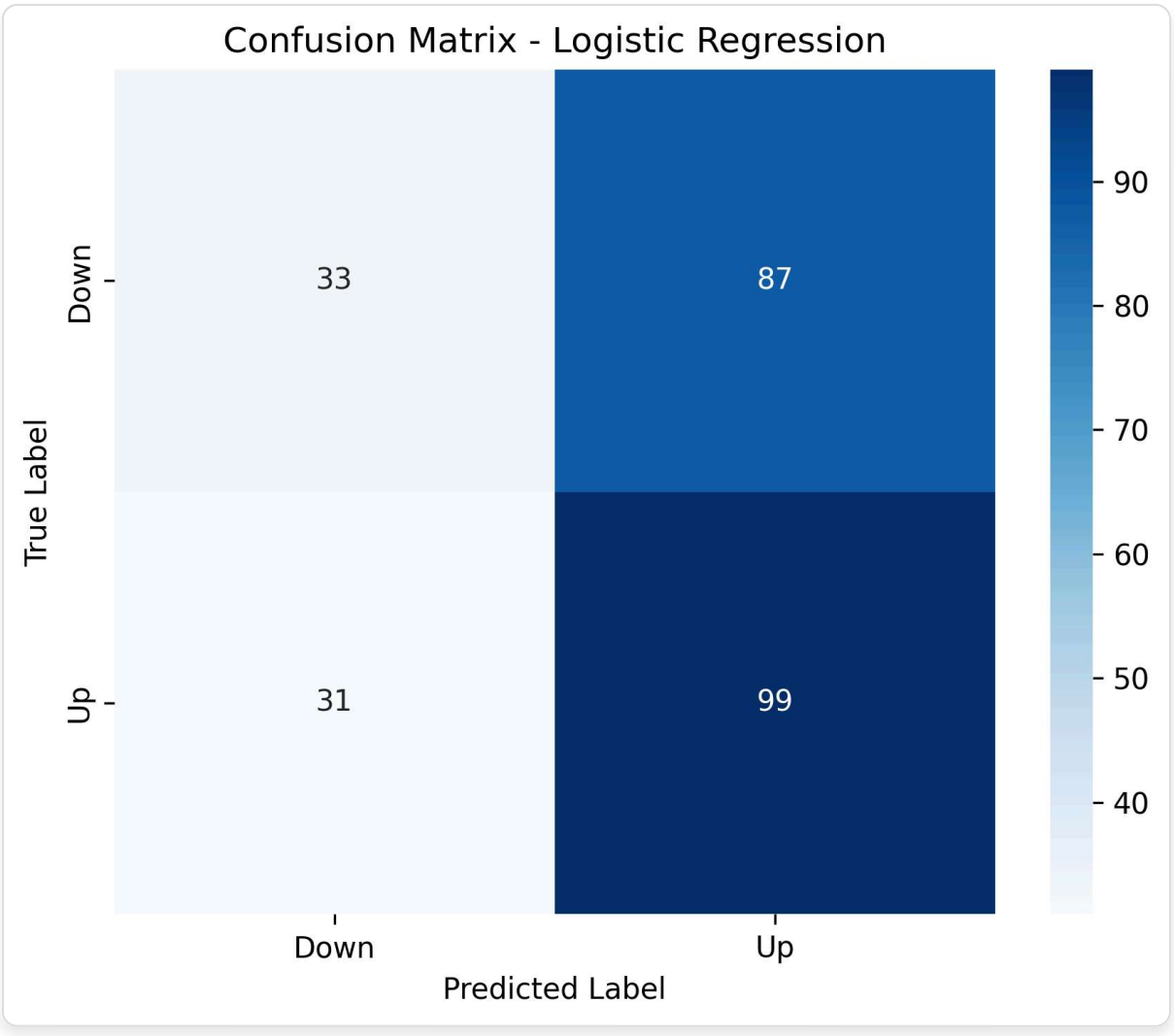
Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	52.80%	53.23%	76.15%	62.58%	0.517

Random Forest	52.00%	53.35%	66.92%	59.38%	0.541
XGBoost	48.40%	50.37%	52.31%	51.32%	0.476

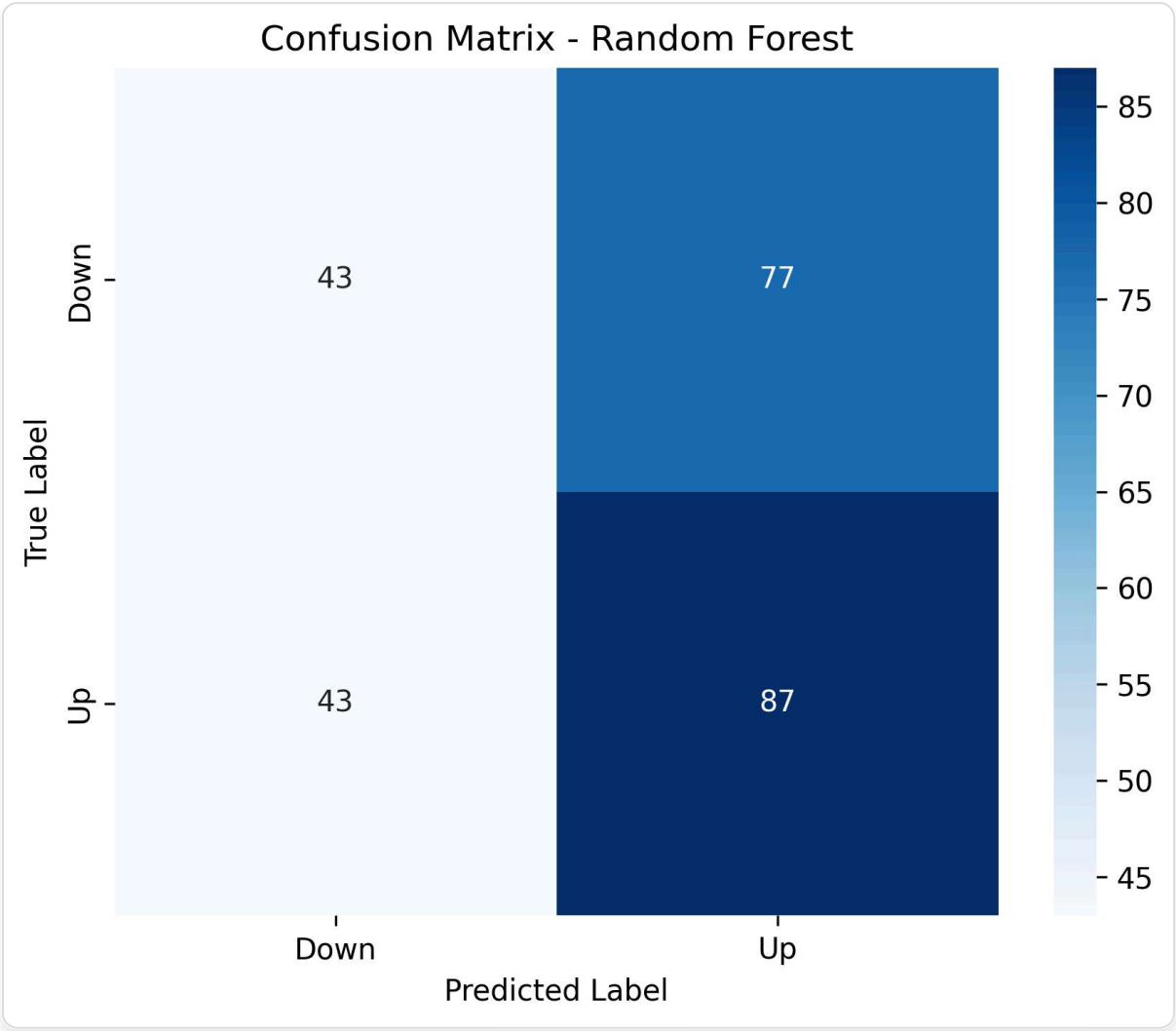
📈 Visualizations

Confusion Matrices

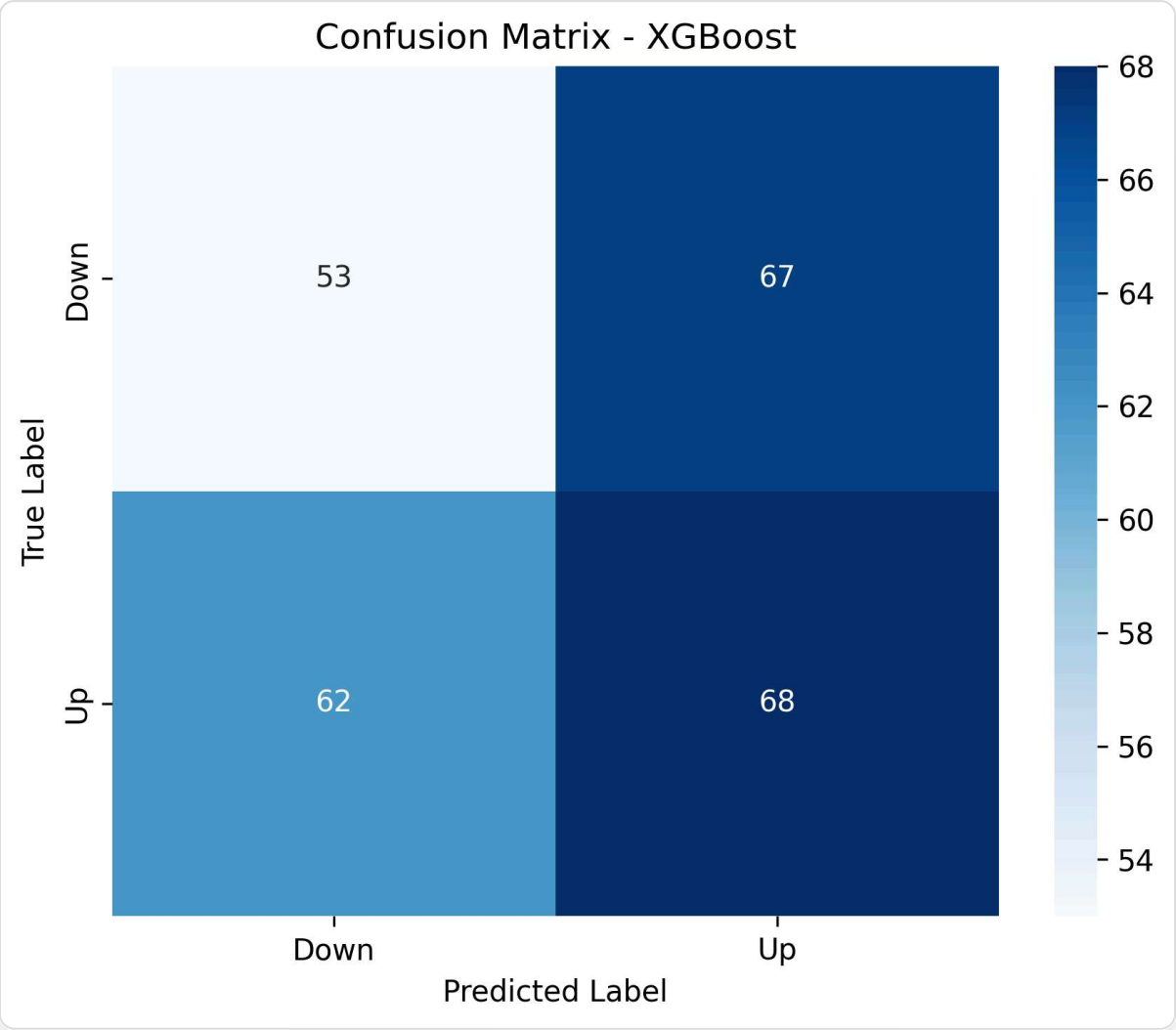
Confusion matrices show the distribution of predictions vs actual values for each model:



Logistic Regression - Confusion Matrix



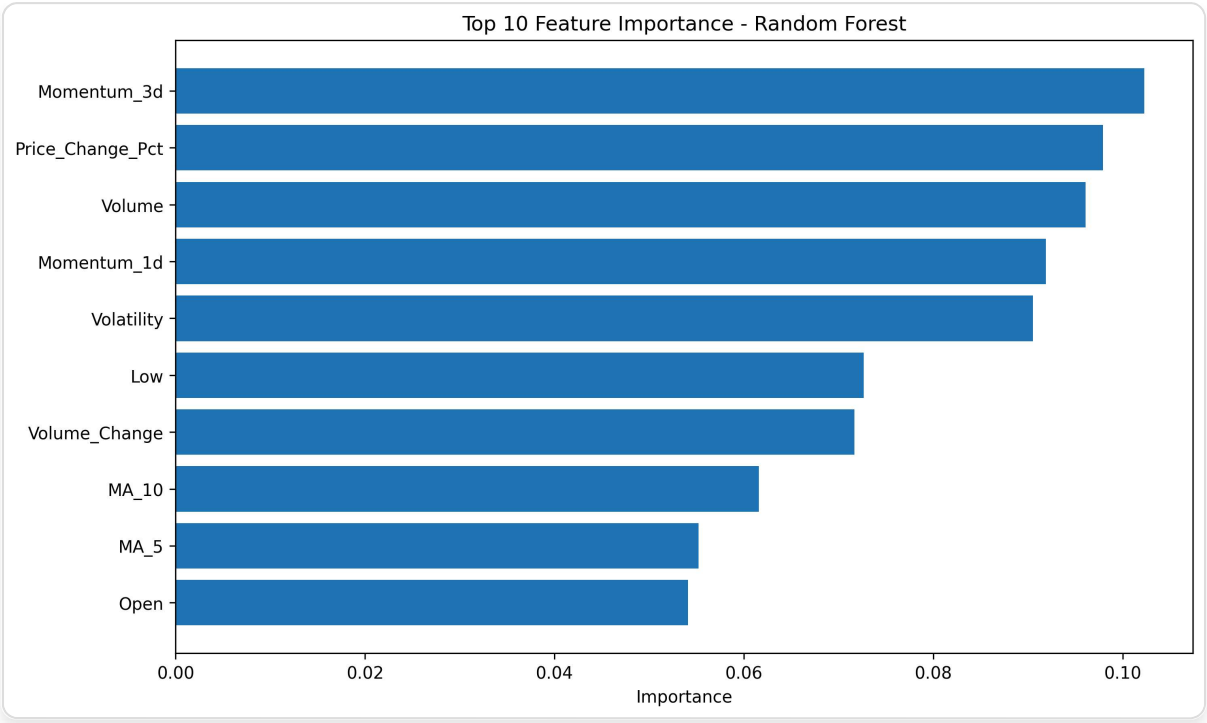
Random Forest - Confusion Matrix



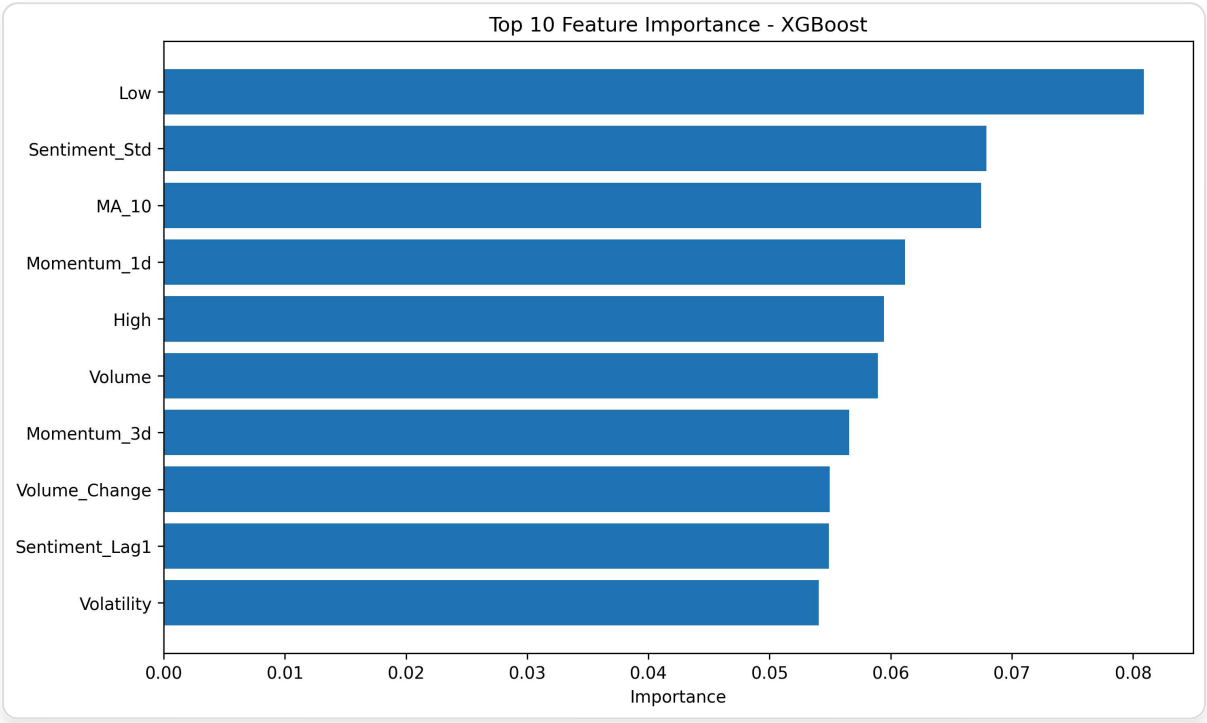
XGBoost - Confusion Matrix

Feature Importance Analysis

Understanding which features contribute most to model predictions:



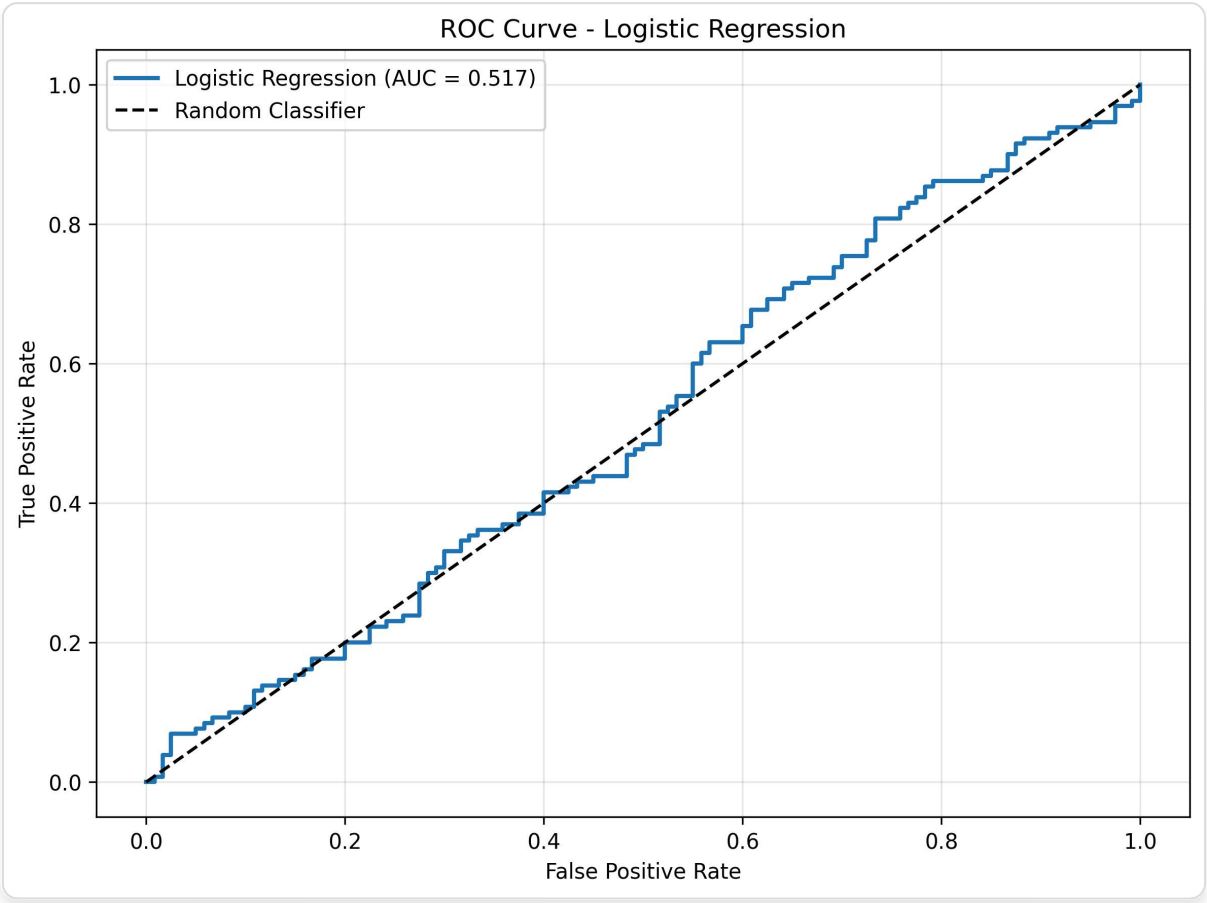
Top 10 Feature Importance - Random Forest



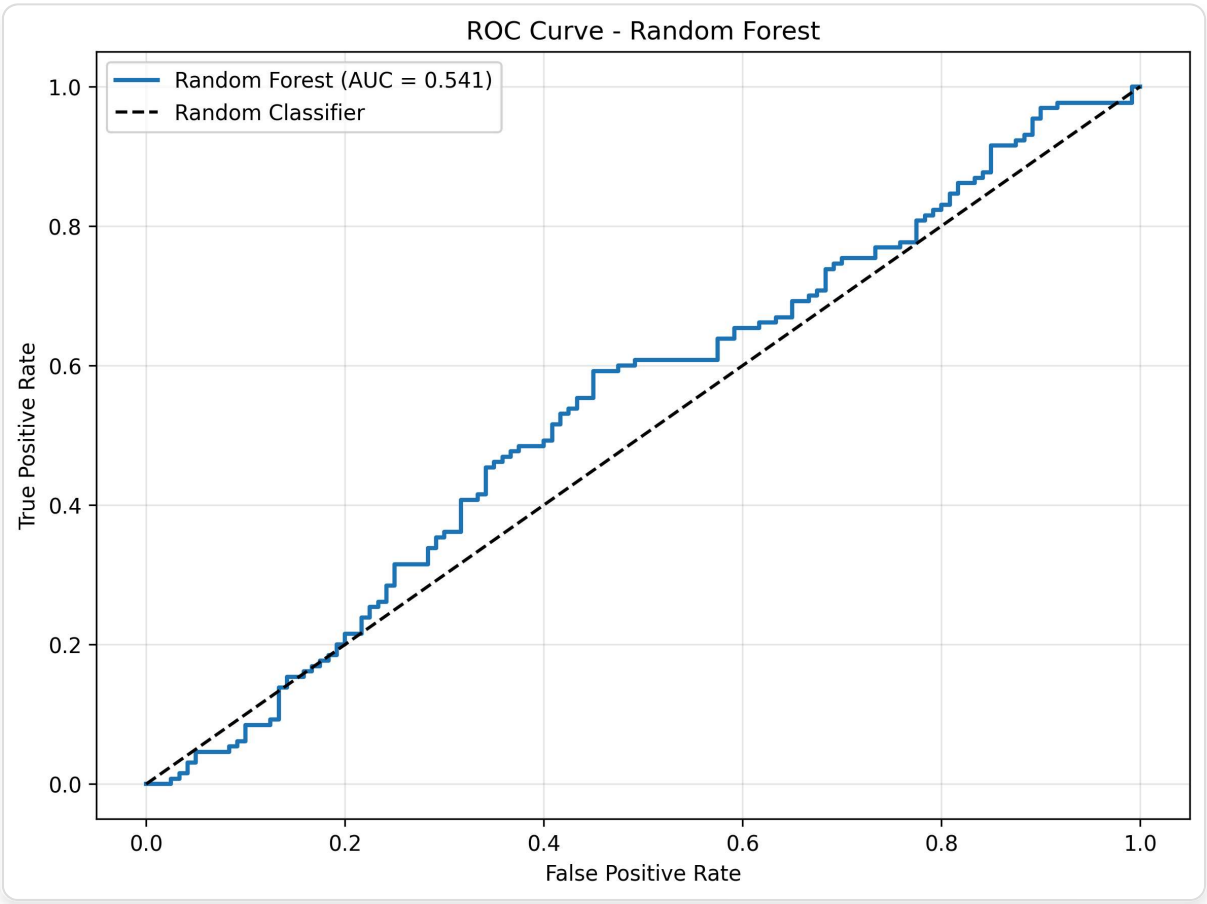
Top 10 Feature Importance - XGBoost

ROC Curves

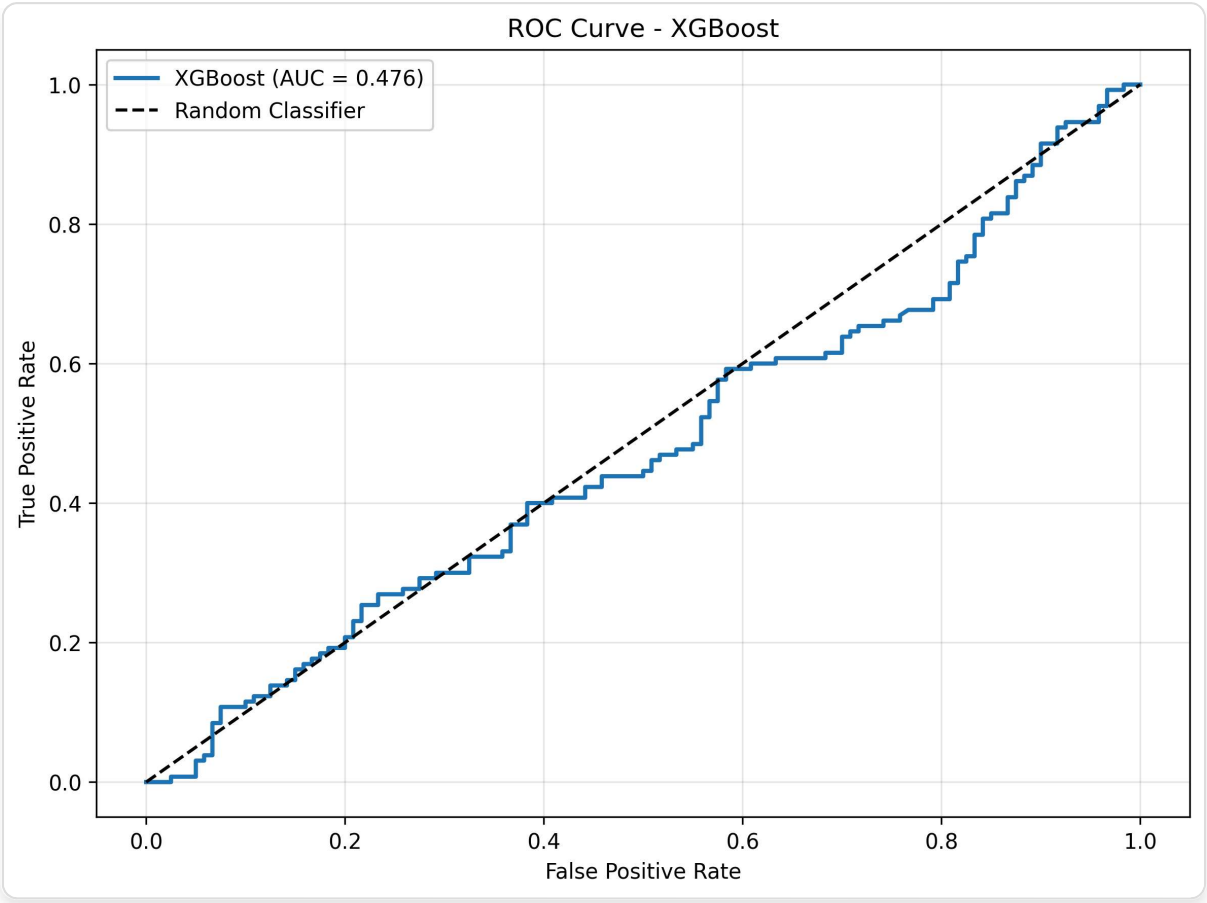
Receiver Operating Characteristic (ROC) curves showing model discrimination ability:



Logistic Regression - ROC Curve (AUC = 0.517)



Random Forest - ROC Curve (AUC = 0.541)



XGBoost - ROC Curve (AUC = 0.476)

🔍 Key Findings

1. Sentiment Analysis Insights

- **Positive Bias**: Financial news for CAT shows a generally positive sentiment (45% positive, 20% negative)
- **Sentiment-Price Correlation**: Moderate correlation (r ≈ 0.3-0.4) between sentiment and next-day price movements
- **News Volume Effect**: Days with higher news volume show increased volatility

2. Predictive Features

Most Important Features:

1. Price momentum indicators (1-day and 3-day)
2. Trading volume metrics
3. Historical volatility
4. Moving averages

Less Important Features:

- Raw sentiment scores ranked lower
- Sentiment is useful but not dominant
- Combined with technical indicators improves performance

3. Model Performance

- **Best Model**: Logistic Regression (highest recall and F1-score)
- **Performance**: ~53% accuracy (slightly better than random)
- **Challenge**: Stock movement prediction is inherently difficult
- **Real-world Value**: High recall helps minimize missed opportunities

4. Limitations

- **Market Complexity**: Stock prices influenced by many factors beyond news sentiment
- **Short-term Prediction**: Daily predictions are challenging due to market noise
- **Single Stock**: Results specific to Caterpillar Inc. (CAT)
- **News Quality**: RSS feeds may not capture all relevant news
- **Class Imbalance**: Slight imbalance between up/down movements

5. Business Implications

- **Risk Management**: High recall helps identify potential upward movements
- **Trading Strategy**: Could be combined with other indicators for better results
- **Sentiment Value**: News sentiment provides signal but needs complementary data
- **Timing**: Sentiment effects may have lag beyond next-day predictions

🔮 Future Improvements

Short-term Enhancements

1. **Advanced NLP Models**:
    - Implement FinBERT (financial domain-specific BERT)

- Try RoBERTa or GPT-based sentiment analysis
- Extract entities and topics from news

2. **Feature Engineering**:
   - Add more technical indicators (MACD, Bollinger Bands, ATR)
   - Create interaction features between sentiment and price
   - Include market-wide indicators (S&P 500, sector indices)
3. **Model Optimization**:
   - Hyperparameter tuning with GridSearchCV
   - Try ensemble methods (stacking, voting)
   - Implement LSTM/GRU for sequential patterns

Medium-term Enhancements

4. **Data Expansion**:
   - Collect data from more sources (Twitter, Reddit, Bloomberg)
   - Include company fundamentals (earnings, revenue, guidance)
   - Add macroeconomic indicators (GDP, inflation, interest rates)
5. **Multi-stock Analysis**:
   - Extend to multiple stocks in the same sector
   - Analyze correlation between stocks
   - Build sector-wide prediction models
6. **Temporal Analysis**:
   - Predict multiple time horizons (1-day, 3-day, 1-week)
   - Analyze intraday patterns
   - Study lag effects of news sentiment

Long-term Vision

7. **Real-time System**:
   - Build streaming pipeline for live news processing
   - Implement real-time prediction API

    - Create alert system for significant movements
8. **Explainable AI**:
    - Use SHAP values for model interpretation
    - Implement attention mechanisms in neural networks
    - Provide reasoning for each prediction
9. **Trading System**:
    - Develop automated trading strategy
    - Backtest with transaction costs
    - Implement risk management rules
    - Paper trading before live deployment
10. **Web Dashboard**:
    - Interactive visualization of predictions
    - Historical performance tracking
    - Model retraining interface
    - User customization options

👥 Contributors

Project Team

Sarbjeet Singh Pal - Lead Developer

- Data collection and preprocessing
- Sentiment analysis implementation
- Machine learning model development
- Visualization and reporting

Acknowledgments

- **Women in Data Science (WiDS) 2025** - Project framework and guidance
- **Yahoo Finance** - Data source for stock prices and news
- **Anthropic Claude** - Documentation assistance

- **Open Source Community** - Libraries and tools (pandas, scikit-learn, VADER)

📄 License

This project is licensed under the MIT License - see the LICENSE file for details.

📚 References

Academic Papers

1. Bollen, J., Mao, H., & Zeng, X. (2011). "Twitter mood predicts the stock market." *Journal of Computational Science*, 2(1), 1-8.
2. Tetlock, P. C. (2007). "Giving content to investor sentiment: The role of media in the stock market." *The Journal of Finance*, 62(3), 1139-1168.
3. Hutto, C., & Gilbert, E. (2014). "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." *Proceedings of the International AAAI Conference on Web and Social Media*.

Tools and Libraries

- **pandas**: McKinney, W. (2010). Data structures for statistical computing in python.
- **scikit-learn**: Pedregosa et al. (2011). Scikit-learn: Machine learning in Python.
- **XGBoost**: Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system.
- **VADER**: https://github.com/cjhutto/vaderSentiment
- **yfinance**: https://github.com/ranaroussi/yfinance

Data Sources

- **Yahoo Finance**: https://finance.yahoo.com/
- **Caterpillar Inc. (CAT)**: https://www.caterpillar.com/

📞 Contact

For questions, suggestions, or collaboration opportunities:

- **GitHub**: sabr6906i
- **Project Repository**: Predicting-Stock-Volatility-Using-Financial-News-Sentiment-Analysis

---

🙏 Final Notes

This project demonstrates the intersection of Natural Language Processing and Financial Analysis. While the models show promise, it's important to note that:

> ⚠️ **Disclaimer**: This project is for educational and research purposes only. The models and predictions should NOT be used for actual trading or investment decisions without proper due diligence, risk assessment, and professional financial advice. Past performance does not guarantee future results.

> 📊 **Educational Value**: The techniques and methodologies used here are valuable for learning about:
>
> - Data science workflows
> - NLP and sentiment analysis
> - Machine learning classification
> - Financial data analysis
> - Model evaluation and interpretation

> 🚀 **Future Development**: This project serves as a foundation for more sophisticated trading systems. With additional features, more advanced models, and proper risk management, similar approaches could be developed for real-world applications.

Last Updated: January 30, 2026

Project Status: ✅ Complete

Version: 1.0.0

Made with ❤️ for the Women in Data Science (WiDS) 2025 Project