

Stock Price Prediction Using Lorentzian Distance Classifier

This Project aims to predict stock price movement using a Lorentzian Distance Classifier combined with sentiment analysis.

The dataset for stock price prediction includes essential indicators: stock prices, volume, sentiment scores, Exponential Moving Average (EMA), Relative Strength Index (RSI), momentum indicators, rolling sentiment, and Bollinger Bands. Stock prices reveal trends and volatility, while volume indicates market liquidity and investor interest. Sentiment scores from financial news reflect market sentiment, and technical indicators like EMA and RSI identify trends and potential overbought or oversold conditions.

By integrating these features, the model enhances predictive accuracy for identifying which of the five selected stocks yields the highest Return on Investment (ROI), allowing for the selection of a single stock for investment. Data was primarily sourced from Yahoo Finance, with limited success in gathering sentiment from Reddit and Twitter.

The model employs a Lorentzian Distance Classifier to predict binary stock price movements. Key features utilized in the model include Volume, Lag1, Exponential Moving Average (EMA), Relative Strength Index (RSI), Momentum, Sentiment_Value, and Rolling Sentiment. The target variable for the model is the price change, allowing for the identification of upward or downward movements in stock prices.

Data preprocessing involves transforming the data structure and scaling the variables using MinMaxScaler. The DataFrame is sliced to retain the most important features, and, given the small dataset, it is split into a 60:40 ratio for training and testing. For model tuning, GridSearchCV is employed to determine the best parameters for the Lorentzian Distance Classifier, utilizing KNN with a Minkowski metric ($p=1$). Evaluation metrics include Accuracy, Precision, Recall, F1 Score, and Return on Investment (ROI) for the investment simulation.

The evaluation metrics for each stock are as follows:

- **ONCO**: Accuracy 60%, Precision 50%, Recall 100%, F1 Score 67%, Simulation ROI 118.57%
- **CNEY**: Accuracy 75%, Precision 0%, Recall 0%, F1 Score 0%, Simulation ROI 132.01%
- **TNXP**: Accuracy 0%, Precision 0%, Recall 0%, F1 Score 0%, Simulation ROI 40.52%
- **APLD**: Accuracy 25%, Precision 33%, Recall 50%, F1 Score 40%, Simulation ROI 7.3%
- **KTTA**: Accuracy 60%, Precision 0%, Recall 0%, F1 Score 0%, Simulation ROI 0%

Interestingly, despite the low evaluation metrics for several stocks, the graphical representation of actual prices versus predicted up/down movements indicates the model's ability to accurately predict stock price trends. This suggests that while the model may struggle with precision in classification, it effectively captures overall market direction, leading to profitable simulation returns.

Sentiment analysis significantly enhances the model's accuracy, while other predictor variables also contribute meaningfully to the outcomes. Further exploration of additional features is necessary to identify which variables yield better predictions and evaluations, ultimately grasping the actual market direction.

For future improvements, rather than using forward fill to handle missing values, setting them to neutral sentiment may provide a more accurate reflection of market conditions. Forward fill was previously used due to the lingering effects of news on market behavior, even when subsequent days lacked significant updates. Additionally, testing various machine learning models and expanding the dataset and timeframe will help capture a broader perspective of the stock market, facilitating more robust predictions.