NASDAQ 100 Trend Classification: Machine Learning's Approach to Identifying Major Turning Points

In this case study, we aimed at designing and implementing a classification model that uses broad economic and leading market indicators to forecast significant peaks and valleys in the NASDAQ 100 index from 2007 to 2023.

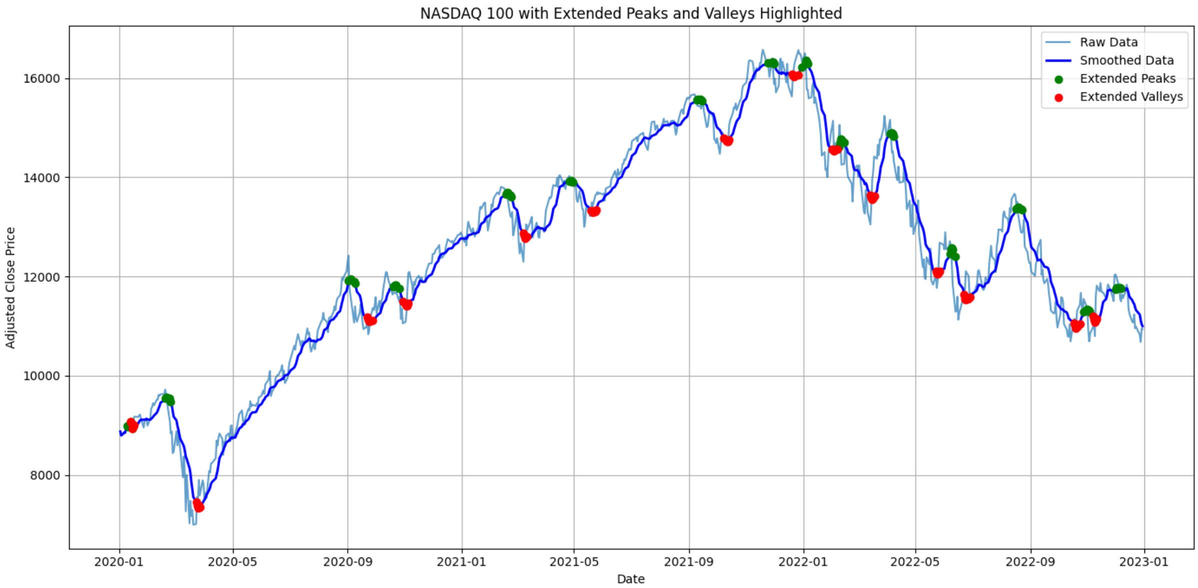
Data Sources:

Data for this study was obtained from Yahoo Finance, NASDAQ, and FRED. These sources provided various indices and leading economic indicators for the period spanning 2007-2023. The data wrangling phase ensured that these datasets were structured appropriately, making them ready for subsequent analysis.

Steps Followed:

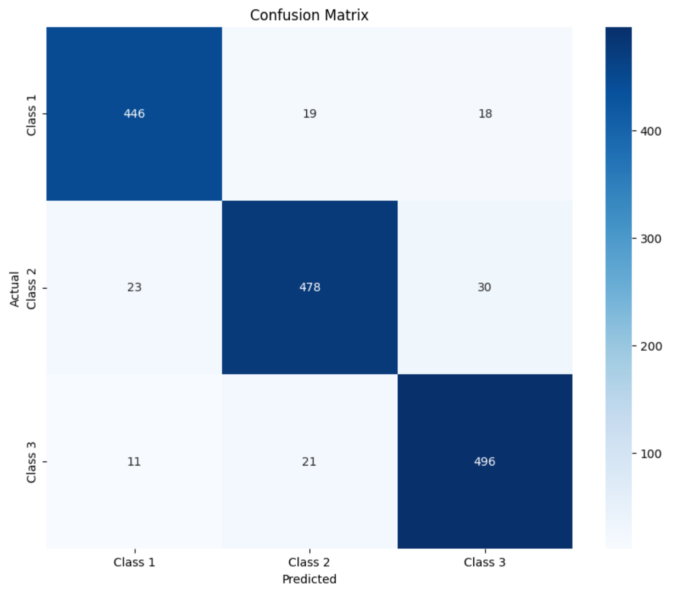
1. Data Smoothing: A moving average with a window size of 10 days was applied to the data to reduce short-term fluctuations and emphasize longer-term trends. SciPy's 'find_peaks' function was employed to identify significant peaks based on prominence. The data was inverted to determine valleys, and the 'find_peaks' function was used again. A prominence value of 100 was chosen after various experiments to capture the primary peaks and valleys.

2. Extending Peaks and Valleys: To provide better context, padding was added around each main peak and valley, marking 3 points on either side of each detected peak and valley. A 'Class' column was introduced in the dataframe for classification. Three labels were assigned: Extended Peaks as Class 1 and Extended Valleys as Class 3.



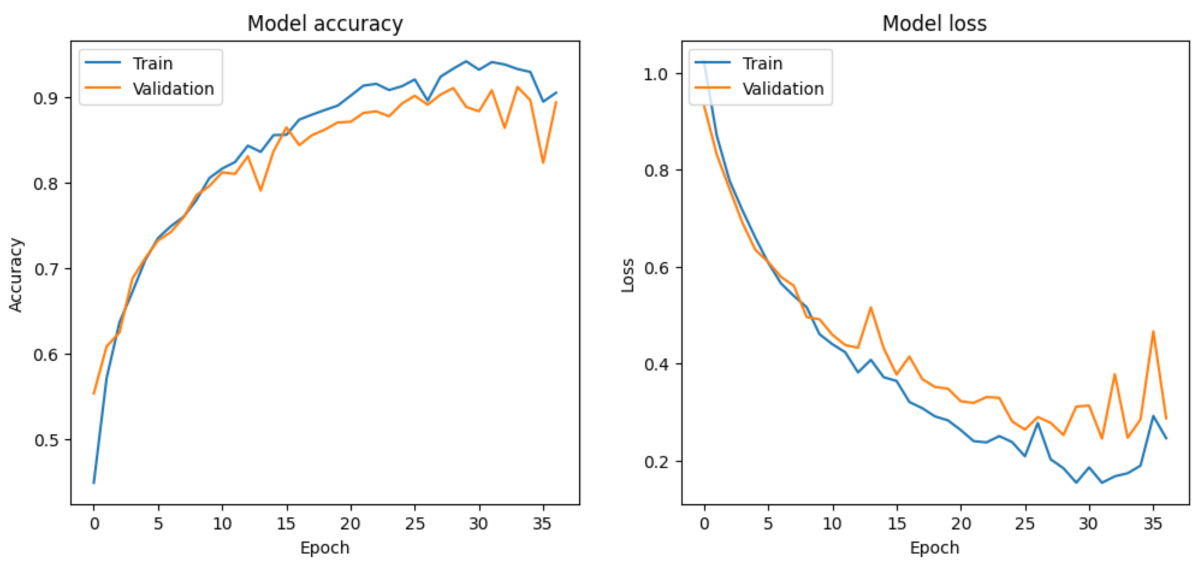
Model Performance:

The LSTM (Long Short-Term Memory) model was chosen and achieved an accuracy of approximately 92% on the test dataset. While the model exhibited high accuracy on the test set, it mainly predicted class 2 and under-predicted classes 1 and 3. Key steps in data preparation, such as data normalization, sequence creation suitable for LSTM, and addressing class imbalance using SMOTE, were instrumental in achieving optimal model performance.



The model architecture includes three LSTM layers and a final dense layer for multi-class classification.

By incorporating early stopping and learning rate adjustments, the model training was both efficient and robust, avoiding potential overfitting and ensuring convergence.



The orange and blue lines being close together for a significant number of epochs is a good sign, as it indicates the model is generalizing well to unseen data.

Conclusions:

The study successfully designed and implemented an LSTM model to predict significant peaks and valleys in the NASDAQ 100 index from 2007 to 2023. By utilizing broad economic and market indicators as features, the model achieved an accuracy of approximately 92% on the test dataset. Data preparation, especially steps like data normalization and addressing class imbalance using SMOTE, played a pivotal role in enhancing the model's performance.

Ideas for Further Research:

- Hyperparameter Optimization: Employ techniques such as grid search or Bayesian optimization to fine-tune model hyperparameters for enhanced performance.

- Deep Learning Variations: Explore other deep learning architectures like GRU (Gated Recurrent Units) or Transformer-based models for time series predictions.

- Feature Importance & Engineering: A deeper analysis of feature importance might reveal latent patterns in the data. Feature engineering can then be applied to derive new features or optimize existing ones for better model performance.

- Alternative Resampling Techniques: Besides SMOTE, consider techniques like ADASYN or under-sampling strategies to deal with class imbalance. The efficacy of these methods can be compared for optimal results.

Recommendations:

1. Operational Decisions: It is recommended to utilize the LSTM model for short-term stock market predictions, particularly in timeframes aligned with quarterly financial rhythms. Given the model's current accuracy, its predictions can greatly influence trading strategies and investment decisions.

2. Refinement Opportunities: Investing resources in refining the model is suggested, particularly by analysing instances of frequent misclassifications, especially between Classes 2 and 3. Such investigations may uncover subtle patterns in the data, paving the way for more sophisticated strategies.