Project Report: Baseline Model for NIFTY 100 Directional Prediction

This report summarizes the first phase of the project, which involved building a simple machine learning baseline for predicting minute-by-minute directional movement of the NIFTY 100 index. The objective was to test the viability of basic technical indicators and linear models before proceeding to the complex crash prediction task. The results were not satisfactory, necessitating a major refinement in the project's second half.

1. Data Preparation and Feature Engineering

Data Used

The model was trained on minute-level historical data for the NIFTY 100 Index (NIFTY 100_minute.csv).

Target Variable Creation

The core classification target, , was created to predict the current minute's outcome:

- **if (Up or Flat minute).**
- **if (Down minute).**

Features Engineered

The final feature set () used for prediction consisted solely of two technical indicators derived from the price data:

1. **Relative Strength Index (RSI, 14-period): A momentum oscillator to measure the speed and change of price movements.**
2. **Exponential Moving Average (EMA, 21-period): A smoothed, trend-following average.**

Data Cleaning and Finalization

- **The date column was converted to datetime and the long-term price trend was plotted.**
- **Rows containing missing values (introduced by the rolling nature of RSI and EMA) were dropped.**
- **Original price columns (open, close, high, low) were deleted from the feature matrix to prevent direct price leakage into the model.**

2. Model Training and Evaluation

Methodology

1. **Data Split: The dataset was split into training (70%) and testing (30%) sets using a randomized split (train_test_split).**
2. **Model: A Logistic Regression model was chosen as a simple, linear baseline classifier.**
3. **Training and Prediction: The model was trained on the features () and used to generate predictions () on the test set ().**

Strategy Backtesting

The model's predictions were applied to the entire dataset to simulate a simple trading strategy:

1. **Strategy Logic: A hypothetical position was taken only when the model predicted an Up/Flat minute ().**
2. **Strategy Return: .**
3. **Performance: Cumulative returns () were tracked and compared against the raw market cumulative returns ().**

3. Conclusion and Future Work

Assessment of Results

The overall outcome of the baseline model was not satisfactory. While the final accuracy score was calculated, the strategy's cumulative returns likely do not show a compelling edge over a simple buy-and-hold (or random) strategy.

This poor performance is expected because the model was inherently flawed for the complexity of the task:

1. **Linear Limitation: A linear model (Logistic Regression) cannot capture the non-linear dynamics of high-frequency price movement.**
2. **Sequential Failure: The model ignores the sequence, treating each minute as independent, which is fundamentally incorrect for a time series problem.**
3. **Incorrect Goal: The model was not trained to predict the rare, high-impact crash event, but rather a near-random directional variable.**

Plan for the Next Project Half

The project will now pivot to the advanced, hybrid approach as originally planned, directly addressing the limitations of the baseline:

1. **Target Redefinition: The primary target will change from to the binary Crash Label (e.g., drop in the next 15 minutes).**
2. **Physics-Inspired Features: Integrate features derived from Econophysics, such as Rolling Volatility, Log Returns, and Hurst Exponent, to quantify market stress and instability.**
3. **Advanced Model: Implement and test Deep Learning models (LSTM/GRU) or Ensemble Trees (XGBoost), which are capable of handling sequential data, non-linearity, and, crucially, the extreme class imbalance of crash events.**
4. **Validation: Adopt Time Series Cross-Validation and focus evaluation metrics on Recall (to minimize missing a crash) and Score for the minority class.**