Final Project: Machine Learning Practice using FastBox High-Frequency Data

Project Requirements

1. Data

1. **Data Source:** Utilize FastBox high-frequency **futures data** with half-second snapshots.
2. **Time Span:** Choose an appropriate time span that balances data richness and computational efficiency:
   - **Suggested Range:** 5 days to 2 weeks (adjust based on hardware capacity).
3. **Notes:**
   - Exclude the `amount`, as it contains slippage errors and is unreliable.
   - Perform data cleaning, including handling missing values and outliers.

2. Prediction Target

1. **Target Definition:** Predict the return of a selected instrument over a fixed time horizon:
$$\text{Return} = \frac{P_{t+h} - P_t}{P_t}$$
where P_t is the price at time t, and P_{t+h} is the price at time $t + h$.
2. **Prediction Horizon:** Define one or multiple horizons (e.g., 1 second, 3 seconds, 5 seconds, 10 seconds).
3. **Prediction Object:** Select at least one instrument for analysis; multiple instruments are encouraged to enhance generalizability.

3. Feature Engineering

1. **Objective:** Extract predictive signals from snapshot data.
2. **Considerations:**
   - **Trading Delay:** Ensure features do not leak future information.
   - **Slippage:** Consider the impact of slippage on actual transaction prices.

3. **Feature Examples:**
    - **Price Features:** Moving averages, bid-ask spread.
    - **Volume Features:** Order book depth, rate of volume change.
    - **Dynamic Features:** Volatility, buy/sell pressure ratio.

4. Model

1. **Model Selection:** Use at least one model for prediction:
    - **Basic Models:** Linear regression, Ridge regression.
    - **Advanced Models:** Random forest, Transformer, etc.
2. **Training and Validation:** Use cross-validation or rolling window validation to ensure robustness.
3. **Evaluation Metrics:**
    - **Regression Metrics:** MSE, MAE, R^2.
    - **Trading Performance:** Cumulative return, Sharpe ratio.

5. Evaluation and Results Analysis

1. **Robustness Testing:** Assess model performance across different time horizons and instruments.
2. **Transaction Cost Analysis:** Simulate the impact of slippage and fees on returns.
3. **Visualization:** Include performance comparison charts for models, return curves, and prediction results.

6. Deliverables

1. **Report:**
    - Format: PDF, approximately 10 pages.
    - Content:
        - Project background and objectives.
        - Data description and preprocessing.
        - Feature construction methods.
        - Model training and validation.
        - Results and analysis.
        - Conclusions and recommendations.
2. **Slides:** A 5-minute presentation in PDF or PPTX format.
3. **Code:** Submit a Jupyter Notebook file with annotations and results to ensure reproducibility.