FinTech Data Curation: Critical Analysis of Minimal Predictive Features

When trying to predict the stock market, it's easy to think that adding more and more data will make the model smarter. But in reality, most of that extra information is just noise and doesn't help. I decided to keep only a small set of features that really matter. These features focus on two main things: how new information affects the market, and how trading activity reflects that information. By combining basic price and volume data with simple news measures, this set is enough to predict short-term price moves without being too complex.

Structured Features (Market Data)
I included the standard **OHLCV data** (Open, High, Low, Close, Volume). These numbers show the market's behavior in a simple way. The closing price is like the final decision of the day, highs and lows show the mood swings, and volume tells us how strong the buying or selling pressure is.

On top of that, I added a few more signals:

- **Daily Return**, which shows momentum from one day to the next.
- **5-Day Volatility**, which tells us if the market is about to get shaky.
- **Moving Averages (5-day and 10-day)**, which act like the market's memory—short-term vs. slightly longer-term trends.
- **Volume Z-Score**, which points out unusual spikes in trading, often linked to big institutional moves.

Unstructured Features (News Data)
I also added two news-based features. **News Count** simply shows how much attention a stock is getting—more news usually means more activity and volatility. **Sentiment Score** (using VADER) captures whether the news is positive or negative, which often pushes the market in the same direction.

Why This Small Set is Enough
I didn't add technical indicators like RSI or MACD because they are basically built from the same price and volume data I already have, so they don't add anything new. I also skipped macroeconomic data because it changes slowly and doesn't really help with predicting the next day's price of a single stock.This minimal set still covers the most important aspects:

1. **Momentum** (returns, averages)
2. **Risk** (volatility)
3. **Attention** (news count)
4. **Sentiment** (tone of news)
5. **Participation** (volume activity)

By keeping it lean, the model avoids overfitting and focuses only on what really drives prices in the short term.