**1. Raw Trading Data**

**Essential Data**

* **OHLCV Data (Open, High, Low, Close, Volume):**
  + Across multiple timeframes (e.g., 1m, 5m, 1h, 1D).
* **Order Book Data:**
  + Bid-ask spread, order book depth, and volume.
* **Trade Flow Data:**
  + Timestamps, trade size, direction (buy/sell), and price.
* **Futures Data:**
  + Open interest, funding rates, liquidation data.

**Contextual Market Data**

* **Macroeconomic Data:**
  + Interest rates, inflation, and economic reports.
* **News Sentiment:**
  + Real-time sentiment analysis of financial news and social media.
* **Alternative Data:**
  + Google Trends, cryptocurrency wallet activity, on-chain analytics.

**2. Indicators**

**Momentum Indicators**

* Relative Strength Index (RSI)
* Moving Average Convergence Divergence (MACD)
* Stochastic Oscillator
* Commodity Channel Index (CCI)

**Trend Indicators**

* Moving Averages (Simple, Exponential, Weighted)
* Average Directional Index (ADX)
* Parabolic SAR

**Volatility Indicators**

* Average True Range (ATR)
* Bollinger Bands
* Donchian Channels
* Chaikin Volatility

**Volume-Based Indicators**

* On-Balance Volume (OBV)
* Volume-Weighted Average Price (VWAP)
* Chaikin Money Flow (CMF)

**3. Patterns**

**Candlestick Patterns**

* Single Candles:
  + Hammer, Shooting Star, Doji.
* Multi-Candles:
  + Bullish/Bearish Engulfing, Morning Star, Evening Star, Three Black Crows.

**Chart Patterns**

* Head and Shoulders (and Inverse)
* Double Top/Bottom
* Triangles (Ascending, Descending, Symmetrical)
* Flags and Pennants
* Cup and Handle

**4. Trend Analysis**

**Price Trends**

* Direction (uptrend, downtrend, sideways)
* Higher highs/lows or lower highs/lows.

**Support and Resistance Levels**

* Identify key price levels where reversals or consolidations occur.

**Multi-Timeframe Analysis**

* Detect higher-timeframe trends while trading on a lower timeframe.

**5. Engineered Features**

Feature engineering transforms raw data and indicators into ML-ready inputs:

**Statistical Features**

* Rolling statistics (mean, standard deviation, skewness, kurtosis) over various time windows.
* Price momentum and percentage changes.

**Relative Features**

* Price relative to moving averages (e.g., Close/50-day SMA).
* Volume relative to average volume.

**Lagged Features**

* Historical values of indicators to capture temporal dependencies.

**Derived Features**

* Delta between close prices of consecutive timeframes.
* Ratio of short-term to long-term moving averages (e.g., SMA 10/SMA 50).

**6. Labels**

The labels represent the outcome you aim to predict:

**Binary Classification**

* **Trade Direction**: Will the price go up or down? (1 for up, 0 for down)
* **Breakouts**: Will the price break above/below a resistance/support level?

**Multi-Class Classification**

* **Trend Prediction**: Sideways (0), Bullish (1), Bearish (2).
* **Pattern Detection**: Identify specific patterns or market structures.

**Regression**

* **Price Prediction**: Predict the exact future price.
* **Return Prediction**: Predict the percentage return over a future window.

**Other Labels**

* **Risk Metrics**: Predict maximum drawdown or volatility.
* **Profitability**: Whether entering a trade at a given time would be profitable.

**7. Data Enrichment for ML**

To improve the quality of your dataset:

* **Normalize Data**: Rescale features for compatibility (e.g., z-scores or min-max scaling).
* **Dimensionality Reduction**: Use techniques like PCA to reduce feature redundancy.
* **Label Lagging**: Ensure labels are correctly aligned with the features (e.g., future returns aligned to past features).
* **Handle Missing Data**: Impute or drop rows with missing values.

**8. Validation and Testing Considerations**

* **Train/Test Split by Time**: Ensure no data leakage by splitting chronologically.
* **Backtesting**: Validate on historical data while simulating live trading conditions.
* **Feature Importance Analysis**: Use SHAP values or feature importance from models like LightGBM to refine your dataset.

**9. Behavioral Indicators**

**Market Sentiment**

* **Social Sentiment Scores**: Aggregate sentiment from platforms like Twitter, Reddit, or crypto forums.
* **Fear and Greed Index**: Measures market emotion.
* **Google Trends Data**: Popularity of search terms related to specific assets.

**Positioning Indicators**

* **COT Reports (Commitment of Traders)**: For futures markets, indicate commercial and non-commercial positions.
* **Funding Rates**: Show whether perpetual swaps are favoring long or short traders.

**Liquidity Analysis**

* **Slippage**: Measure of price movement when large orders are placed.
* **Market Impact Costs**: How order size affects execution price.

**10. Advanced Market Metrics**

**Order Book Dynamics**

* **Order Book Imbalances**: Difference between buy and sell volumes at various price levels.
* **Queue Dynamics**: Rate of order cancellation and replacement.
* **Order Book Spread Ratio**: Spread relative to mid-price.

**Trade Imbalances**

* **Buy vs. Sell Volume**: Net buying or selling pressure.
* **Volume Delta**: Difference between market buy and sell volumes.

**Volatility Skew**

* Distribution of implied volatilities for options at different strike prices (for options data).

**11. Derivatives Data**

* **Options Data**:
  + Put/Call ratios.
  + Implied volatility and volatility surfaces.
* **Futures Data**:
  + Basis (difference between spot and futures price).
  + Contango/Backwardation dynamics.

**12. Macro Correlations**

* Correlations with:
  + Broader market indices (S&P 500, NASDAQ).
  + Commodities like gold or oil.
  + Interest rates and bond yields (e.g., 10-year Treasury rates).

**13. Additional Chart Patterns**

* **Volume Patterns**:
  + Clustering of high-volume trades at specific price levels.
* **Fractal Patterns**:
  + Use fractal mathematics to detect repeating structures across timeframes.

**14. Time-Based Features**

* **Market Hours**:
  + Session information (Asian, European, US).
* **Day of Week**:
  + Trading tendencies based on weekdays or weekends.
* **Seasonality**:
  + Trends during specific times of the year (e.g., holiday effects).

**15. Cross-Asset Relationships**

* **Correlation Features**:
  + Price correlations between different trading pairs or assets.
* **Lead-Lag Relationships**:
  + Whether one asset leads another in price movements.
* **Sector Analysis**:
  + Group assets by sector (e.g., DeFi, gaming tokens) for relative performance.

**16. Risk and Reward Metrics**

* **Risk Metrics**:
  + Sharpe ratio, Sortino ratio, max drawdown.
* **Reward Metrics**:
  + Expected return over a fixed horizon.

**17. Alternative Features**

**Sentiment from Blockchain Activity**

* **On-Chain Metrics** (for crypto):
  + Active wallet addresses.
  + Token transfers and staking data.
  + Exchange inflows/outflows.

**Network Effects**

* **Transaction Growth**:
  + Rapid increases in on-chain transactions.
* **Hash Rate**:
  + For proof-of-work blockchains, indicates miner activity and network security.

**18. Market Regime Features**

* **Volatility Regimes**:
  + Classify periods as low-volatility or high-volatility.
* **Trend Persistence**:
  + Detect regime shifts using Hidden Markov Models or statistical methods.

**19. Technical Patterns via Advanced Techniques**

* **Fourier Transform Analysis**:
  + Extract cyclical patterns in price data.
* **Wavelet Analysis**:
  + Identify localized trends and cycles.

**20. Labels for Multi-Horizon Models**

* **Short-Term Movements**:
  + Predict returns or trend direction over minutes/hours.
* **Medium-Term Trends**:
  + Capture swing trading opportunities (e.g., daily trends).
* **Long-Term Predictions**:
  + Identify macro-level shifts over weeks/months.

**21. Data Augmentation for Robustness**

* **Synthetic Data Generation**:
  + Create new data using generative models (e.g., GANs) to simulate market behavior.
* **Noise Injection**:
  + Add random noise to train models that are resilient to market noise.

**22. Feature Interaction Engineering**

* **Pairwise Combinations**:
  + Create features by combining existing ones (e.g., RSI \* Volume).
* **Cross-Timeframe Interactions**:
  + Relate higher timeframe trends to lower timeframe behavior.

**23. Anomaly and Event Detection**

* **Unusual Price Movements**:
  + Detect outliers in price or volume.
* **Significant News Events**:
  + Tag data with events that significantly impact markets.

**End Goal**

The final dataset should be:

* **Rich in Information**: Include raw data, indicators, patterns, and trends.
* **Aligned with the Prediction Task**: Ensure features correlate well with the labels.
* **Ready for ML Models**: Clean, normalized, and preprocessed for algorithms.

Would you like help in structuring or implementing this dataset?