Reading Notes — Stage 09 Feature Engineering

What Is Feature Engineering?

Feature engineering is the process of creating new variables ("features") from raw data to help models learn better patterns.
In many cases, **the quality of features matters more than the choice of model**.

Key Principles

- **EDA → Features**: Every feature idea should come from patterns noticed in EDA.
- **Hypothesis-Driven**: Each feature should embed a theory about the data.
- **Simple First**: Ratios, differences, and flags often outperform complex math.

Examples in Finance

- **Spend-to-Income Ratio**: How much a customer spends relative to their income.
- **Rolling Average of Transactions**: Captures short-term spending trends.
- **Interaction Features**: Income × Credit Score may signal creditworthiness.
- **Categorical Encodings**: Region encoded with one-hot or frequency values.

Categorical Encoding

- **One-Hot Encoding**: Expands each category into its own column.
- **Label Encoding**: Assigns integers to categories.
- **Frequency Encoding**: Uses the relative frequency of each category.
- **Target Encoding**: Replaces each category with the mean of the target variable for that category.
    - Example: If customers in "North" region default 15% of the time, then all North rows get 0.15.
    - Note: Powerful but prone to data leakage if applied incorrectly.

Interaction & Polynomial Features

- **Interaction Features**: Multiply or combine two variables to reveal relationships.
    - Example: Income × Credit Score may highlight higher-quality borrowers.
- **Polynomial Features**: Add squared or higher-order terms to capture curvature in data.

Pitfalls & Best Practices

- **Don't add features blindly**: Too many can cause overfitting or slow down training.
- **Check correlation/redundancy**: New features may simply duplicate existing information.
- **Document everything**: Future teammates (or you, in 2 weeks) should know *why* a feature exists.

Takeaway

Feature engineering is where creativity and domain knowledge meet data science.
Your job is to **translate raw patterns from EDA into meaningful, testable features** that improve your model's ability to learn.