



DAY 25 — LightGBM & CatBoost

| Next-generation Gradient Boosting

PART 1 — LightGBM (Light Gradient Boosting Machine)

1 Why LightGBM Exists

Problems with XGBoost:

- Slower on large datasets
- Memory heavy
- Level-wise tree growth causes inefficiency

 **LightGBM solves this**

2 What Is LightGBM? (CORE IDEA)

| LightGBM is a fast, memory-efficient gradient boosting framework designed for large datasets.

3 The BIG Innovation: Leaf-Wise Tree Growth (VERY IMPORTANT 🔥)

Traditional Trees (XGBoost):

- Grow **level by level**

LightGBM:

- Grows **leaf by leaf**
- Expands the leaf with **maximum loss reduction**

 **Result:**

- Faster convergence
- Better accuracy

- Higher overfitting risk (needs control)
-

4 Why LightGBM Is FAST

LightGBM uses:

- Histogram-based splitting
- Efficient memory usage
- Parallel learning

✅ Works extremely well on **millions of rows**

5 Important LightGBM Parameters (MUST KNOW)

num_leaves

Controls complexity (VERY IMPORTANT)

- More leaves → more complex
 - Must match `max_depth`
-

learning_rate

Same concept as before.

Lower = safer.

n_estimators

Number of boosting rounds.

max_depth

Limits tree depth to prevent overfitting.

min_data_in_leaf

Minimum samples per leaf.

Critical for stability.

6 When to Use LightGBM

- ✓ Very large datasets
 - ✓ Many numerical features
 - ✓ Need speed
 - ✗ Very small datasets
 - ✗ High risk of overfitting without tuning
-

🟡 PART 2 — CatBoost (Categorical Boosting)

7 Why CatBoost Exists

Problems with other models:

- One-hot encoding explodes features
- Target leakage risk
- Categorical data handling is painful

👉 **CatBoost solves categorical features elegantly**

8 What Is CatBoost? (CORE IDEA)

| CatBoost is a gradient boosting algorithm optimized for categorical features.

9 The MAGIC: Ordered Target Encoding (VERY IMPORTANT 🔥)

Instead of one-hot encoding:

- CatBoost encodes categories using target statistics
- Prevents data leakage
- Uses permutation-driven learning

📌 This is HUGE for real-world data.

10 Why CatBoost Is So Stable

- Symmetric trees
- Built-in regularization

- Handles missing values
- Minimal tuning required

 Works well **out-of-the-box**

1 1 Important CatBoost Parameters

depth

Tree depth (default = 6)

learning_rate

Controls step size.

iterations


Number of trees.

loss_function

- Classification: **Logloss**
 - Regression: **RMSE**
-

cat_features

Specify categorical column indices.

 This is the killer feature.

1 2 LightGBM vs CatBoost vs XGBoost (DECISION TABLE)

| Situation | Best Choice |
|---------------------------|--------------------|
| Huge numeric dataset | LightGBM |
| Many categorical features | CatBoost |
| General tabular ML | XGBoost |
| Minimal tuning needed | CatBoost |
| Kaggle competitions | XGBoost / LightGBM |

One Sentence to Remember

XGBoost is powerful, LightGBM is fast, CatBoost is smart with categories.
