**How LightGBM is Used in the BERT + LightGBM Hybrid Model?**

1. **BERT Extracts Semantic Meaning**

* We take the product’s item\_name (e.g., *"Up & Go Breakfast Drink Vanilla 3x250ml"*)
* Use **DistilBERT** to convert that name into a **768-dimensional vector**
* This vector captures the **semantic context** (brand, size, type, flavor, etc.)

1. **PCA Reduces Dimensions**

* 768 is too high-dimensional for LightGBM
* So we apply PCA to reduce it to 100 dimensions
* This keeps the most informative signals while reducing noise and complexity

1. **Structured Features are added**

We combine the reduced BERT features with classic tabular data like:

* item\_price
* unit\_price
* DiscountRate
* PriceCapped
* Week\_num
* was\_on\_special\_last\_week

1. **LightGBM trains on combined data**

* We now have a single feature matrix:

X = [structured features + PCA-reduced BERT embeddings]

y = next\_on\_sale\_week (multi-class label: Week 1–8)

* LightGBM is trained as a multi-class classifier to predict the correct week

**Why Use LightGBM?**

|  |  |
| --- | --- |
| **Reason** | **Benefit** |
| Handles tabular + numerical data well | Perfect for structured retail features |
| Very fast to train | Ideal when BERT embeddings are precomputed |
| Easy to interpret | Feature importance, tree plots |
| Supports calibration | Used with CalibratedClassifierCV for better probabilities |

**Summary:**

**BERT** handles the **semantic understanding** of product names.  
**LightGBM** does the **actual week classification**, using both **semantic + structured signals**.

**Final Summary & Conclusion:**

**Key Takeaways:**

|  |  |
| --- | --- |
| **Milestone** | **Accomplishment** |
| Dataset Generation | Realistic 8-week synthetic dataset based on original data |
| Quality Assurance | Advanced logic, statistical, and visual checks |
| Model Development | High-performing hybrid of BERT and LightGBM |
| Results | Reliable predictions on discount week likelihood |

**Limitations**

* Synthetic logic doesn't reflect **real market demand or inventory factors**
* BERT was not **fine-tuned** — potential for future accuracy improvements
* Lack of real transactional feedback