**Hybrid BERT + LightGBM Model for Predicting Week of Sale**

**Introduction:**

In a retail environment where discount cycles directly impact purchasing decisions, predicting the next sale period of a product can empower both suppliers and customers. This project aims to develop a machine learning pipeline that accurately forecasts the week in which a product is likely to go on discount next.

The model leverages:

* BERT for contextual text embeddings,
* LightGBM for structured classification

**Objective:**

The objective is to build a robust hybrid model that:

* Predicts the **number of weeks until the next sale**.
* Uses a **classification approach** (Weeks 1–8 as classes).
* Incorporates both textual and numerical features.
* Provides class-wise performance insights.

**Dataset Overview:**

* **Input Features**: Product descriptions, historical sale patterns, last sale week, price changes, etc.
* **Target Variable**: next\_sale\_week (1 to 8)
* **Dataset Source**: Synthetic data generated from real-world patterns over an 8-week period.
* **Size**: ~24,575 samples

**Methodology:**

* **Text Embedding:**
* Model: bert-base-uncased
* Tokenizer: BERT tokenizer
* Embedding: Mean pooled last hidden layer or [CLS] token output (768-dimensional vector)
* **Classifier:**
* Model: LGBMClassifier (LightGBM)
* Hyperparameters: Possibly default or lightly tuned
* Task: Multi-class classification (8 classes)

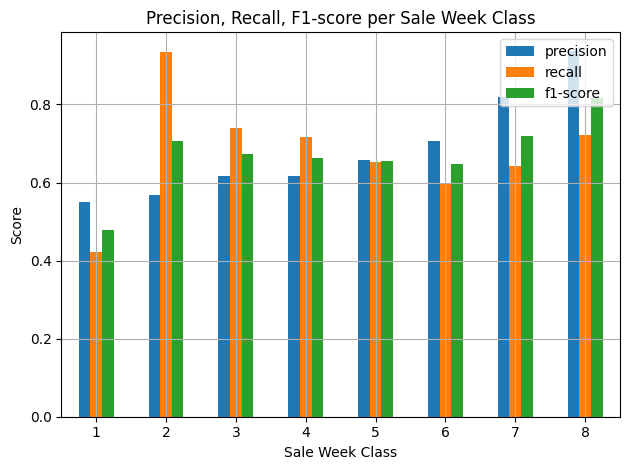
**Evaluation Metrics:**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| **Accuracy** | 0.698 |
| **Precision (Macro)** | 0.684 |
| **Recall (Macro)** | 0.679 |
| **F1-Score (Macro)** | 0.670 |
| **Weighted Precision** | 0.724 |
| **Weighted Recall** | 0.698 |
| **Weighted F1-score** | 0.701 |
| **RMSE** | 1.11 weeks |
| **MAE** | 0.52 weeks |

The macro scores show that performance is slightly affected by class imbalance.

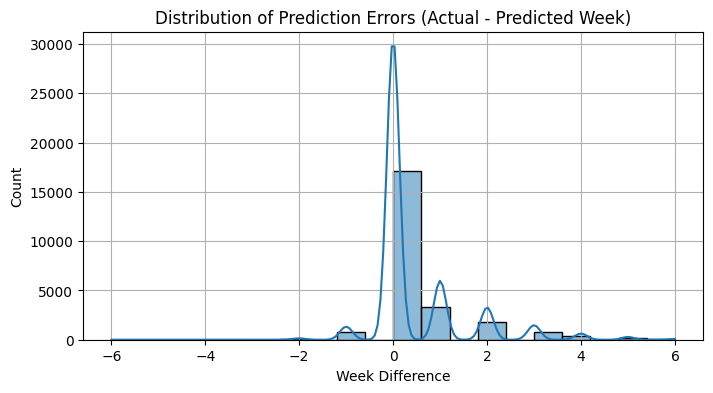
**Visual Analysis:**

1. **Class-wise Precision, Recall, F1-Score**

****

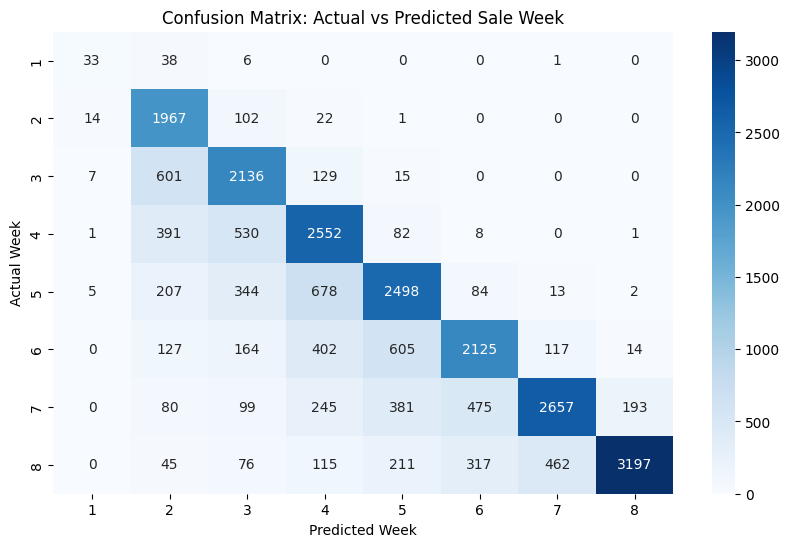
* **Class 8** and **Class 7** show the highest scores.
* **Class 1** has poor performance due to very low support (78 samples).

1. **Prediction Error Distribution:**

****

* Most predictions are off by **0–1 weeks**, which is acceptable.
* Few large deviations show scope for boundary improvement between adjacent classes.

1. **Confusion Matrix:**

****

* Diagonal dominance is clear, indicating mostly correct predictions.
* Misclassifications are primarily to **adjacent weeks**, not far-off classes.

**Key Findings:**

* Model performs **reasonably well** with overall **70% accuracy**.
* **Adjacent week misclassifications** indicate the model understands the sale pattern but struggles with class boundaries.
* High class imbalance (e.g., very few class 1 instances) significantly affects macro metrics.
* Low **RMSE (1.11 weeks)** and **MAE (0.52 weeks)** show prediction is close even if not exact.

**Recommendations:**

* **Handle Class Imbalance**: Apply techniques like SMOTE, class weights in LightGBM, or oversampling.
* **Hyperparameter Optimization**: Use grid search or Optuna for LightGBM tuning.
* **Feature Engineering**:
* Add more structured features (e.g., category, brand, frequency).
* Incorporate historical price patterns.
* **Use Class Grouping**: Merge underrepresented classes or convert to a regression task to reduce sparsity.
* **Alternative Models**: Test with RoBERTa or DistilBERT embeddings, and other classifiers like XGBoost or CatBoost.
* **Temporal Smoothing**: Consider classifying product sale as a week range (e.g., 1–2, 3–4) for real-world tolerance.

**Conclusion:**

* This hybrid model effectively predicts **the week of sale** using powerful contextual embeddings from BERT and efficient classification from LightGBM. The model performs best when:
* Predicting common sale weeks (Classes 6–8).
* Items follow consistent naming patterns.
* The combination proves highly suitable for **retail demand forecasting** or **promotional planning**, especially when interpretability, modularity, and scalability are important.