Logistic Regression for Imbalanced Dataset

Why Logistic Regression?

Logistic Regression is a well-established linear model widely used in text classification tasks due to its simplicity, efficiency, and solid theoretical grounding. It is especially effective when paired with TF-IDF vectorized data, as it can handle high-dimensional, sparse feature spaces often found in natural language processing (NLP) tasks. It also offers probabilistic outputs and interpretable coefficients.

We selected Logistic Regression as a baseline algorithm to classify product review ratings (1 to 5 stars) based on review text. It is scalable and offers good performance with appropriate preprocessing and hyperparameter tuning.

Model Training Logic

The pipeline followed was:

1. **Text Preprocessing**:
    - Lowercasing
    - Punctuation removal
    - Stopword removal (NLTK/spaCy)
    - Lemmatization
2. **Vectorization**: TF-IDF (`max_features=5000`)
3. **Model**: `LogisticRegression` trained using `GridSearchCV` for hyperparameter optimization.
4. **Data Split**: Stratified 80/20 train-test split to preserve rating distribution.
5. **Scoring Metric**: `f1_macro` — chosen to give equal weight to all classes, especially for imbalanced datasets.

Hyperparameter Tuning Details

```
param_grid = {
    'C': [0.01, 0.1, 1, 10],                # Regularization strength (inverse)
    'penalty': ['l2'],                      # Regularization type
    'solver': ['liblinear', 'lbfgs', 'saga'],  # Optimization algorithm
    'max_iter': [100, 200, 500]              # Max training iterations
}
```

- **C** controls regularization: lower = stronger regularization.

- **penalty** 'l2' helps prevent overfitting in high-dimensional text data.
- **solver** options tested for stability and speed.
- **GridSearchCV** used 3-fold cross-validation to select the best combination.

Evaluation Result

After hyperparameter tuning and training, the best model achieved:

Metric	Value
Accuracy	43%
Macro F1	42%
Support	2000 test samples

Detailed classification report:

```
precision    recall  f1-score   support

           1       0.58      0.35      0.44       200
           2       0.40      0.25      0.31       300
           3       0.36      0.42      0.39       500
           4       0.43      0.54      0.48       600
           5       0.51      0.45      0.47       400

    accuracy                           0.43      2000
   macro avg       0.45      0.40      0.42      2000
weighted avg       0.44      0.43      0.42      2000
```

- **Class 4 & 5** performed better, likely due to more representation in training data.
- **Class 1 & 2** showed lower recall, suggesting the model struggles with minority classes in imbalanced settings

Interpretation

- The model demonstrates **moderate performance**, with higher confidence in predicting frequent ratings.
- It **fails to generalize** well on underrepresented ratings (like 1 and 2 stars), a common issue in imbalanced datasets.
- **Macro F1-score** of 0.42 highlights the need for more advanced balancing or feature techniques for improvement.

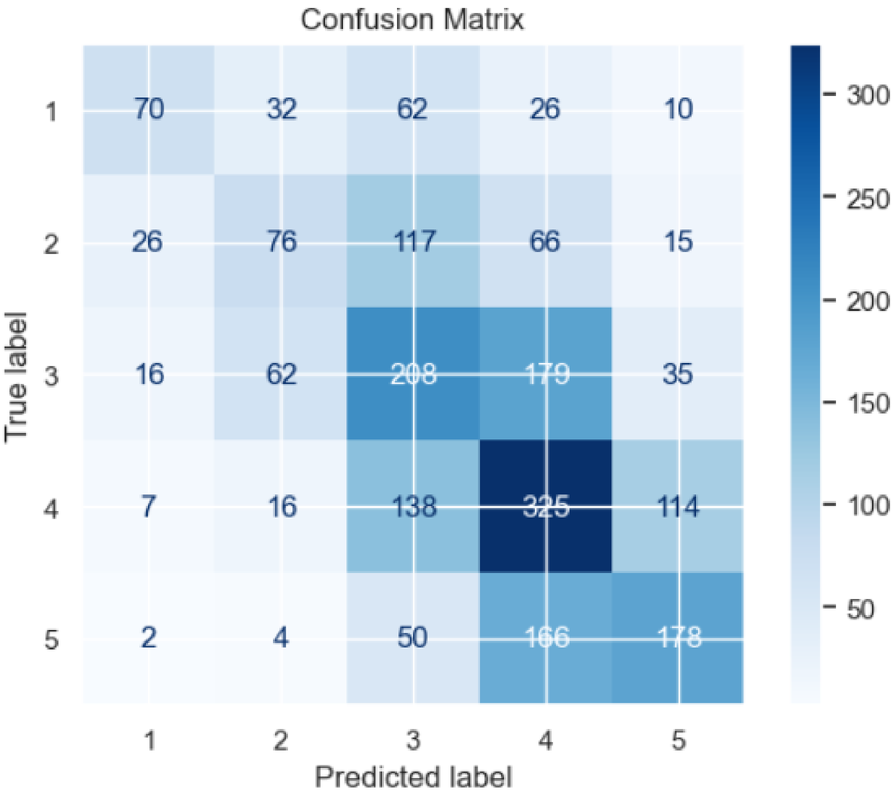
When to Use Logistic Regression

- High-dimensional, sparse feature space (like TF-IDF)
- Need for interpretable models
- Binary or multiclass classification with linearly separable data
- As a **baseline model** for text classification tasks

Limitations

- Assumes **linear decision boundaries**
- Sensitive to **correlated features**
- Struggles with **imbalanced classes** without tuning or sampling
- Less powerful than tree-based models or deep learning in non-linear settings

Confusion matrix



	1	2	3	4	5
1	70	32	62	26	10
2	26	76	117	66	15
3	16	62	208	179	35
4	7	16	138	325	114
5	2	4	50	166	178