Evaluation Report: Model_A vs Model_B

1. Summary of Model Performance

Model	Test Set Type	Accuracy	Macro F1	Notable Class Behavior
Model_A (trained on balanced data)	Imbalanced	0.43	0.42	High recall for Class 1 and 5; struggles on Class 2
Model_B (trained on imbalanced data)	Balanced	0.40	0.40	High precision for Class 1; best recall for Class 4

2. Evaluation Matrix Observations

Model_A on Imbalanced Test Set:

```
Evaluation of balanced Model on imbalanced Test Set:
              precision    recall  f1-score   support

           1       0.35      0.64      0.45       200
           2       0.31      0.32      0.31       300
           3       0.47      0.34      0.39       500
           4       0.50      0.40      0.44       600
           5       0.47      0.57      0.52       400

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

- **Precision:**
  - Highest for Class 4 (0.50) and Class 5 (0.47).
  - Low for Class 2 (0.31), indicating false positives.
- **Recall:**
  - Strongest for Class 1 (0.64) and Class 5 (0.57), indicating successful capture of true positives.
  - Very low for Class 2 (0.32), meaning many class 2 instances are missed.
- **F1-Score:**
  - Class 5 has the best balance between precision and recall (0.52).
  - Class 2 performs the worst (0.31), due to both low precision and recall.

Model_B on Balanced Test Set:

```
Evaluation of Imbalanced Model on Balanced Test Set:
              precision    recall  f1-score   support

           1       0.70      0.32      0.44       400
           2       0.40      0.26      0.32       400
           3       0.30      0.39      0.34       400
           4       0.33      0.57      0.42       400
           5       0.56      0.47      0.51       400
```

```
    accuracy                           0.40      2000
   macro avg       0.46      0.40      0.40      2000
weighted avg       0.46      0.40      0.40      2000
```

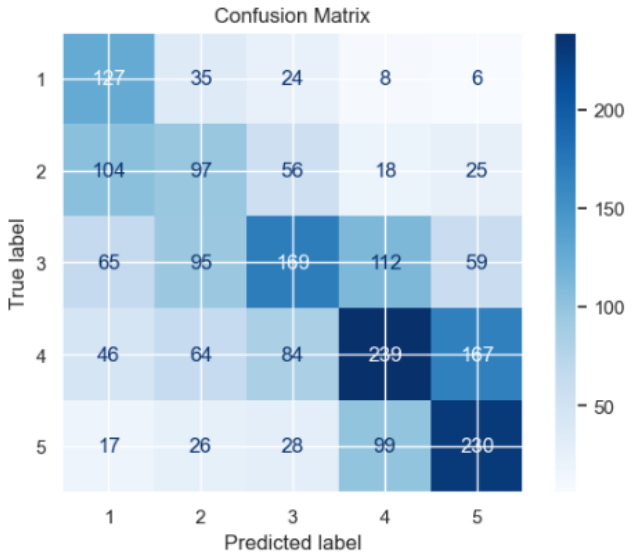
- **Precision:**
  - Highest for Class 1 (0.70), but with **very low recall (0.32)** many false negatives.
- **Recall:**
  - Best for Class 4 (0.57), meaning the model favors this class.
  - Poor recall for Class 2 (0.26) high false negative rate.
- **F1-Score:**
  - Again, Class 5 performs best (0.51), showing consistent model attention to this class.
  - Class 2 is again the weakest (0.32).

Insight: Although Model_B has high precision in some classes, the recall is generally poor it **misses many actual instances**, which is concerning in real-world use cases where missing a prediction is costlier than a false positive.

---

3 Observations from Confusion Matrices

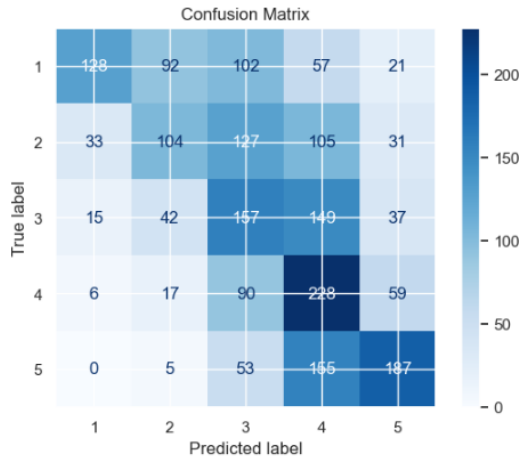
ModelA trained on imbalanced dataset



Model_A (on imbalanced test set):

- Confusion matrix shows **better recall** for **minority class 1** (127 out of 200).
- Model **spreads predictions fairly** across labels, a sign of **less bias**, due to balanced training.
- Predicts class 4 and 5 reasonably well too, despite imbalanced test data.

ModelB trained on Balanced Dataset



Model_B (on balanced test set):

- Model is **heavily biased** toward **majority class predictions** (from its imbalanced training).
- Struggles to generalize on balanced test data — confusion matrix shows misclassifications across all classes.
- Class 1 has high precision but **low recall** (many actual 1s misclassified).

4. Effect of Training Data Distribution on Generalization

- **Balanced training (Model_A)**:
  - Produces a model that generalizes more fairly across all classes.
  - Handles both imbalanced and balanced test sets better.
  - Improves recall for rare classes and avoids overfitting to dominant ones.
- **Imbalanced training (Model_B)**:
  - Leads to a biased model that struggles to correctly identify minority classes.
  - Performs poorly when tested on a balanced dataset due to **lack of exposure to rare classes**.

Aspect	**Balanced Training (Model_A)**	**Imbalanced Training (Model_B)**
Bias	Low -all classes treated fairly	High - favors majority classes
Generalization	Good across different distributions	Poor when test set is balanced
Recall on rare classes	Better	Very poor
Robustness	High	Low
Confusion spread	More evenly distributed	Concentrated around majority predictions

4. Recommendation:

Deploy Model_A (trained on balanced data)

Why?

- **More balanced performance** across all classes.
- **Higher macro F1-score**, which is crucial when each class is important.
- **Generalizes better** even when the test data is imbalanced — showing robustness.
- Reduces the risk of **neglecting minority classes**, which is critical in many applications (e.g., fraud detection, medical diagnosis, review moderation).