Model Cross Testing

Overview:

Two Support Vector Machine (SVM) models were developed and evaluated:

- Model A: Trained on the balanced dataset (equal number of samples per rating).
- Model B: Trained on the imbalanced dataset (natural distribution of reviews).

Both models were tested on:

1. Their own test set (native evaluation).
2. The other dataset's test set (cross-test).

Performance Summary of Model A

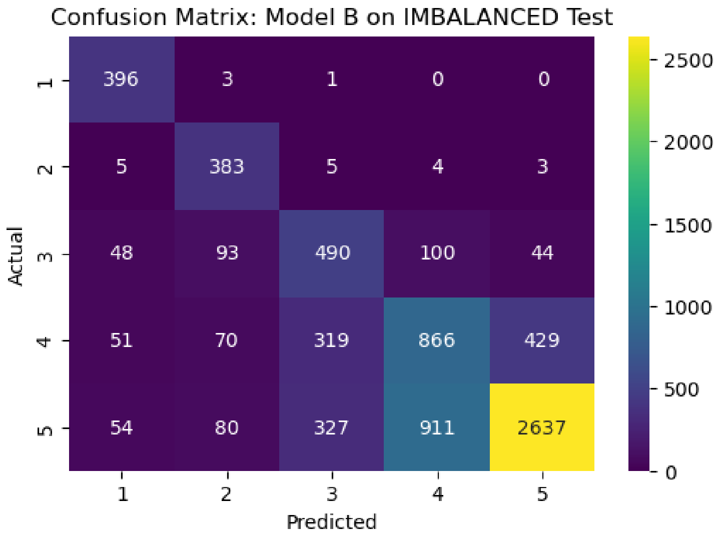
- Model A, on the balanced dataset:

Model achieved an accuracy of 62.75% on its native balanced test set, with a weighted precision of 0.6117, recall of 0.6275, and F1-score of 0.6167**.** Its macro-level metrics closely mirrored these results, all hovering around 0.61–0.62, indicating consistent performance across classes.

- Model A, on the imbalanced dataset:

When cross-tested on the imbalanced dataset, Model A's accuracy improved slightly to 65.2%, with a weighted precision of 0.6909, recall of 0.6520, and F1-score of 0.6594, while the macro F1 rose modestly to 0.6611. This suggests that the model adapted reasonably well to real-world class proportions, benefiting from the dominant presence of higher ratings.

```
=== Evaluation: Model B on IMBALANCED Test
Samples evaluated   : 7319
Accuracy            : 0.652002
Precision (weighted): 0.690913
Recall    (weighted): 0.652002
F1 Score  (weighted): 0.659355
Precision (macro)   : 0.612053
Recall    (macro)   : 0.747333
F1 Score  (macro)   : 0.661064
```



Performance Summary of Model B

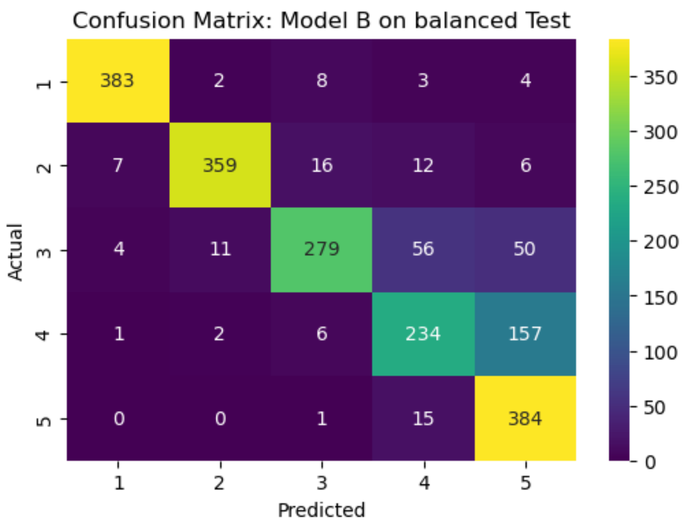
- Model B, on the imbalanced dataset:

This model demonstrated stronger and more stable performance overall. On its native imbalanced test set, it reached an accuracy of 67.76%, with a weighted precision of 0.6477, recall of 0.6776, and F1-score of 0.6526. Its macro precision, recall, and F1-scores stood at 0.6513, 0.6086, and 0.6193 respectively which is slightly lower, but indicative of steady behaviour across uneven class distributions.

- Model B, on the balanced dataset:

When evaluated on the balanced test set, Model B's performance surged substantially, achieving an impressive accuracy of **81.95%**, weighted precision of 0.8399, recall of 0.8195, and F1-score of 0.8189, with macro metrics matching these values exactly.

```
=== Evaluation: Model B on balanced Test
Samples evaluated   : 2000
Accuracy            : 0.819500
Precision (weighted): 0.839940
Recall    (weighted): 0.819500
F1 Score  (weighted): 0.818864
Precision (macro)   : 0.839940
Recall    (macro)   : 0.819500
F1 Score  (macro)   : 0.818864
```



Overall, the metrics clearly show that Model B not only performs better on its native data but also generalizes far more effectively to a different distribution. Its consistently high weighted and macro scores across both test sets indicate strong representational learning, robustness to data imbalance, and superior adaptability compared to Model A.

Observations

Effect of Training Data Distribution

- The Model A learned to treat all rating classes equally, which is good for fairness but makes it slightly less adaptive to real-world, imbalanced distributions.
- The Model B learned to mirror the natural class proportions, leading to stronger overall accuracy and generalization, especially when tested on both distributions.
- Interestingly, Model B maintained strong performance even on a balanced test set, indicating that exposure to dominant patterns in real data did not prevent it from correctly identifying minority classes.

Performance Behaviour Across Classes

- Both models struggled most with 3-star and 4-star ratings; these classes typically contain mixed sentiment and overlap semantically with adjacent ratings making them harder to separate.
- Model B consistently produced higher recall for high ratings (5-star) and better accuracy overall, suggesting stronger representational learning.

Conclusion

Based on the evaluation and cross-testing results, **Model B** (SVM trained on the imbalanced dataset) is recommended for deployment.

It achieved the highest overall accuracy (67.7%) on its native imbalanced test set and also demonstrated excellent generalization when tested on the balanced dataset, achieving 81.9% accuracy with consistently high precision and F1-scores across all classes.

Model A provided fairer performance across classes but lower overall accuracy, making it less suitable for real-world prediction scenarios.

Hence, Model B is selected for deployment, as it delivers superior accuracy, robustness, and generalization in both balanced and imbalanced contexts.