# PAPER ID 118 - REVIEWER 12345

## ENHANCING E-COMMERCE CONVERSION THROUGH PERSONALIZED RECOMMENDATION SYSTEMS

This paper compares three approaches to recommendation systems—collaborative filtering, content-based filtering, and a hybrid ensemble—using transaction data from a Czech online bookstore (52,347 transactions, Jan–Jun 2023). The authors report that the hybrid model achieved the best performance, with a 14.2% increase in conversion rate compared to collaborative filtering and a 9.3% uplift in average basket value compared to content-based filtering. Results are summarized in Table 3 (p. 8).

## Strengths

- Dataset quality: The dataset is substantial, covering more than 50,000 transactions, and includes both browsing history and purchase outcomes. This allows for a robust analysis.
- Evaluation metrics: The authors use precision, recall, F1-score, and AUC (Table 2, p. 6), which is appropriate for recommendation tasks.
- Applied relevance: The results are tied to managerial implications, e.g., prioritizing hybrid recommenders for premium customers (p. 12).
- Structure: The manuscript is well-structured, moving logically from motivation, literature, methodology, results, to discussion.

## Weaknesses / Limitations

- Insufficient statistical validation:
- While mean F1-scores are reported (e.g., hybrid = 0.82, collaborative = 0.78, content-based = 0.74), no statistical tests are performed to verify significance. A paired t-test or Wilcoxon signed-rank test across cross-validation folds would strengthen claims.
- No effect sizes (e.g., Cohen’s d) are reported, making it hard to gauge the magnitude of improvements.
- Cross-validation procedure unclear:
- The authors state they used “five-fold cross-validation” (p. 5),

but do not clarify whether folds were stratified by user or by transaction, which is crucial in recommendation settings to avoid data leakage.

- The hybrid model is described as a “weighted combination” of collaborative and content-based outputs (p. 5), but the weights are not specified. Were they optimized using grid search, logistic regression, or learned adaptively?
- No details on hyperparameter tuning are provided (e.g., neighborhood size in collaborative filtering, similarity thresholds in content-based).
- Missing robustness checks:
- The authors do not test the stability of their results under alternative evaluation metrics such as NDCG or MAP, which are standard in recommender research.
- No sensitivity analysis was conducted to examine how results change with different training/test splits.

## SUGGESTIONS FOR IMPROVEMNT

- Perform statistical testing on model performance differences (paired t-tests or non-parametric alternatives such as Wilcoxon signed-rank). Report p-values and effect sizes to validate the significance of improvements.
- Clarify the cross-validation setup: ensure folds are split by user, not by transaction, to avoid inflating performance.
- Provide a detailed description of the hybrid model, including weighting mechanisms, training procedures, and hyperparameter tuning strategy.
- Extend evaluation with additional metrics (e.g., NDCG, MAP, Hit Ratio at k) to align with recommender system standards.
- Add robustness checks (e.g., sensitivity to hyperparameters, alternative data splits).
- Expand the ethical discussion: explain anonymization of user data, fairness considerations, and compliance with GDPR.
- Update the literature review with more recent references, such as Zhang et al. (2022, RecSys) on neural collaborative filtering and Li & Chen (2023, Information Processing & Management) on explainability.

Decision:
**ACCEPT WITH MAJOR REVISIONS**