EVALUATING RECOMMENDATION APPROACHES FOR EMPLOYEE BENEFIT PERSONALIZATION

VYTAUTAS MAGNUS UNIVERSITY
Faculty of Informatics

AUTHORS:
Eglė Rimkienė egle.rimkiene@vdu.lt
Marius Marmakas marius.marmakas@vdu.lt
Rimantė Povilavičienė rimante.povilaviciene@vdu.lt
Justinas Juozas Dainauskas justinas.dainauskas@vdu.lt
Tomas Krilavičius tomas.krilavicius@vdu.lt

1 INTRODUCTION

Employee benefit platforms offer budgets that employees can spend across many categories, but the large number of options can be overwhelming and lead to irrelevant suggestions. This reduces satisfaction, lowers provider visibility, and leaves budgets unused. Therefore, a key question arises: **which recommendation methods best balance efficiency, personalization, and fairness in this context?**

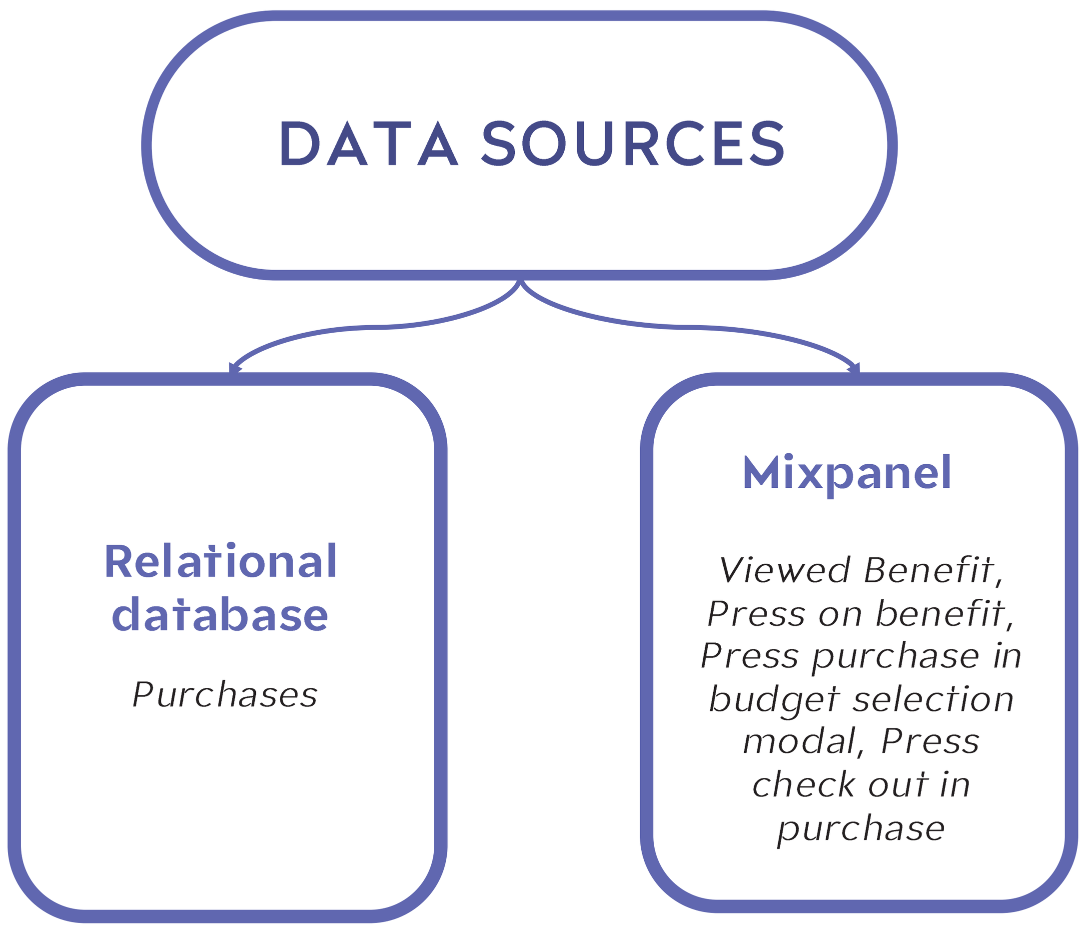
3 METHODS

Alternating Least Squares (ALS):
A matrix-factorization method

Two-Tower model:
A neural retrieval architecture

2 GOAL

To identify which recommendation method—**ALS** or **Two-Tower**—delivers the best balance of accuracy, personalization, and coverage for employee benefit platforms.

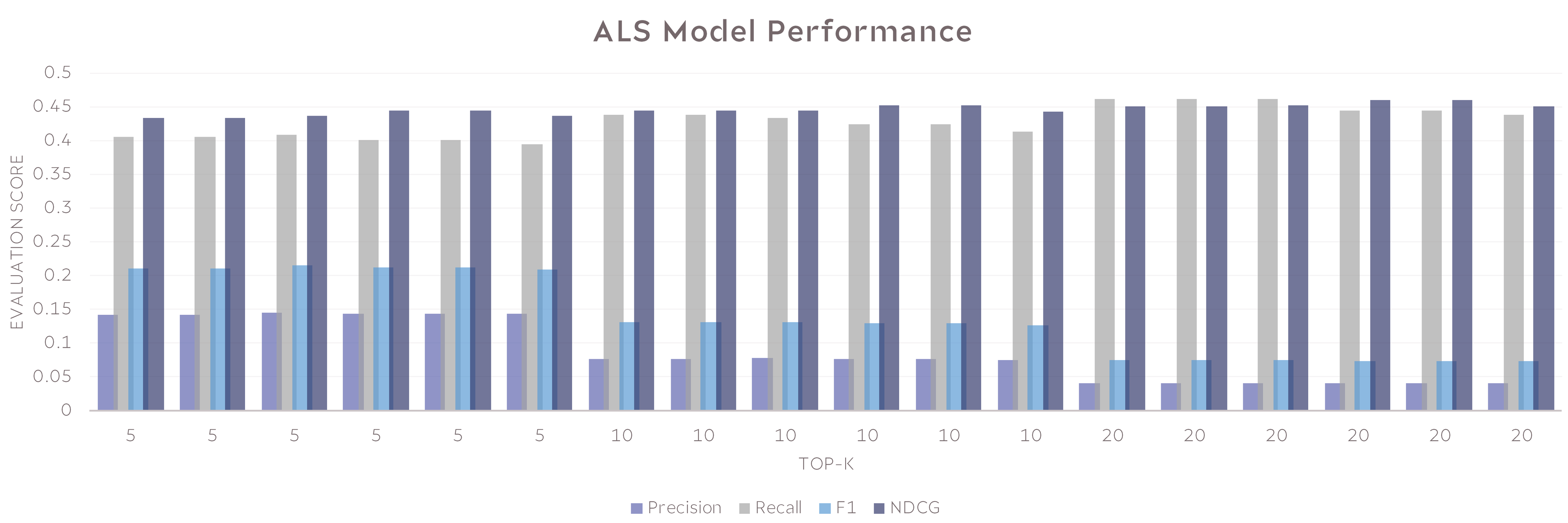


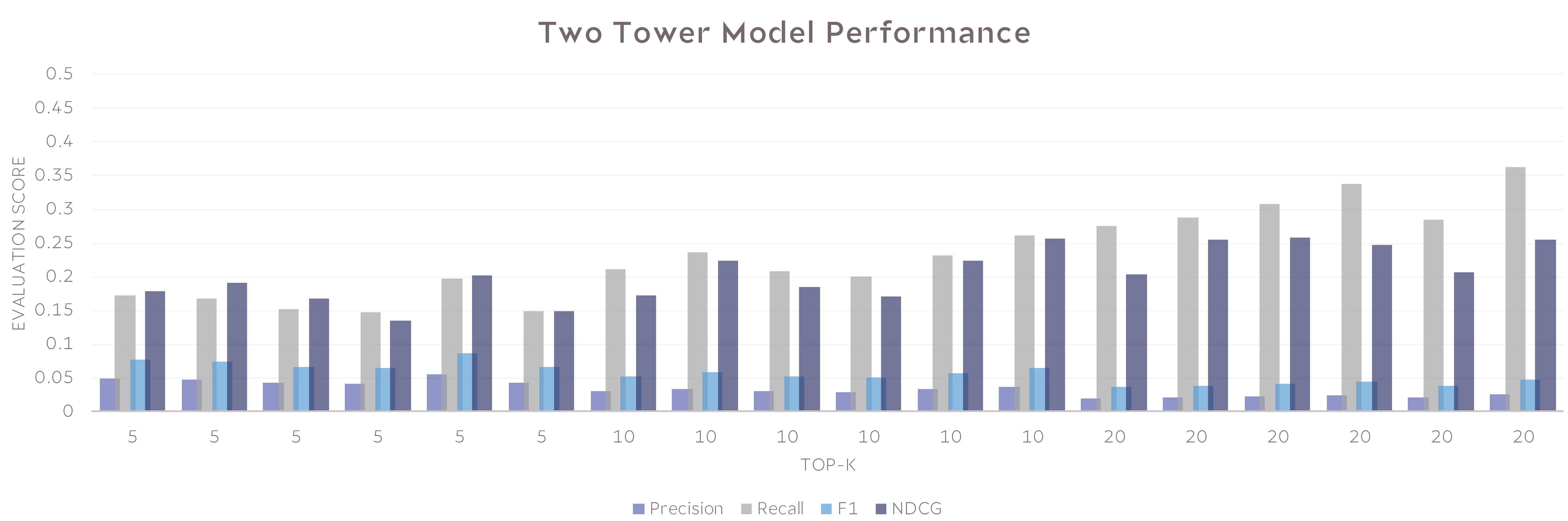
4 EXPERIMENTS

No	Purchase history (DB)	Mixpanel events
1	2022-10-01 → 2024-09-30 (2 yrs)	2024-04-01 → 2024-09-30 (6 mo)
2	2022-10-01 → 2024-09-30 (2 yrs)	2024-07-01 → 2024-09-30 (3 mo)
3	2022-10-01 → 2024-09-30 (2 yrs)	2024-09-01 → 2024-09-30 (1 mo)
4	2023-10-01 → 2024-09-30 (1 yr)	2024-04-01 → 2024-09-30 (6 mo)
5	2023-10-01 → 2024-09-30 (1 yr)	2024-07-01 → 2024-09-30 (3 mo)
6	2023-10-01 → 2024-09-30 (1 yr)	2024-09-01 → 2024-09-30 (1 mo)

Experiments with varying training windows were performed to assess temporal stability and the impact of data sparsity. The data from **2024-10-01 to 2024-11-19** was held out for testing to objectively evaluate how well the model generalizes to unseen, recent user behavior.

5 RESULTS





6 CONCLUSIONS

- ALS achieves higher Precision, Recall, F1, and NDCG across all Top-K settings.
- ALS results are more stable across different data windows and sparsity conditions.
- ALS is simpler to train and tune, as it uses matrix factorization rather than a complex neural architecture.
- Two-Tower provides broader coverage, but its accuracy and ranking quality remain lower in this context.
- Two-Tower performance improves with larger K, though overall scores remain below ALS.

Overall: ALS best satisfies the practical requirements of this system, providing accurate, consistent, and deployment-ready employee-benefit recommendations. Two-Tower remains a promising direction for improving cold-start coverage in future iterations.

7 FUTURE PLANS

- Hybrid models (ALS + embeddings)
- Cold-start improvements
- Fairness analysis
- Online A/B testing
- Deployment optimizations

benme

CARD
CENTRE FOR APPLIED RESEARCH AND DEVELOPMENT