Evaluating The Chances & Risks of Machine Learning in Predicting Mode Choice For The Mobility Coin System Based On Stated Preference Survey Data

Master‘s Thesis of Berhane Gebremariam Desta

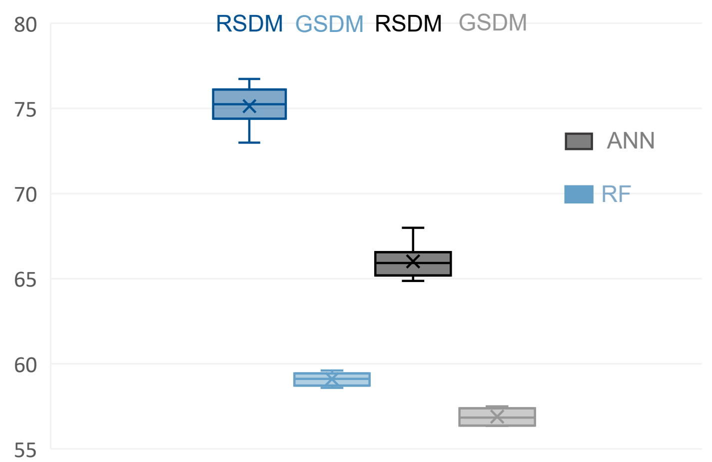
Mentoring:
Dipl.-Ing. Philipp Servatius
MSc. Isabella Waldorf

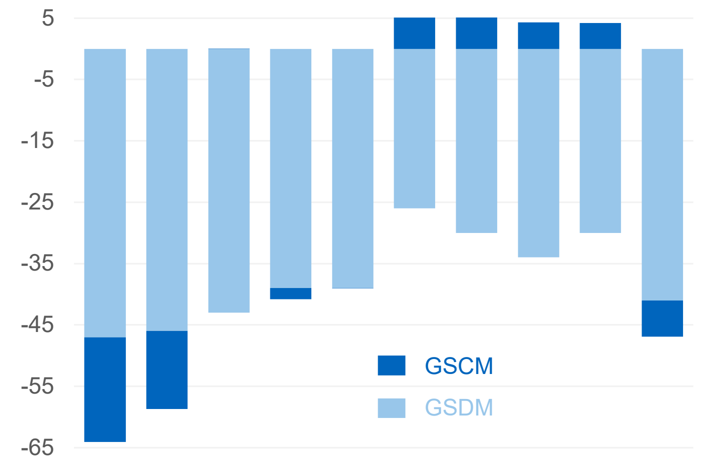
Background

The rapid growth of AI and ML has been driven by the availability of large datasets, fast & low-cost computing power, and open-source tools like PyTorch and TensorFlow, alongside advances in DL and NLP. Their ability to model complex, non-linear interactions has boosted their use in choice modeling. However, key gaps in the literature remain, including inconsistent findings on predictive performance, methodological flaws in model development, and limited interpretability of results.
This study addresses these issues using the Mobility Coin SP survey data, which includes responses from 1,349 participants answering 12 choice questions, six for the status quo and six for the Mobility Coin scenario.

Data leakage and overfitting

Using random sampling on a panel dataset led to significant data leakage. Models trained this way showed over a 15% drop in accuracy when evaluated on unseen individual-level test data.



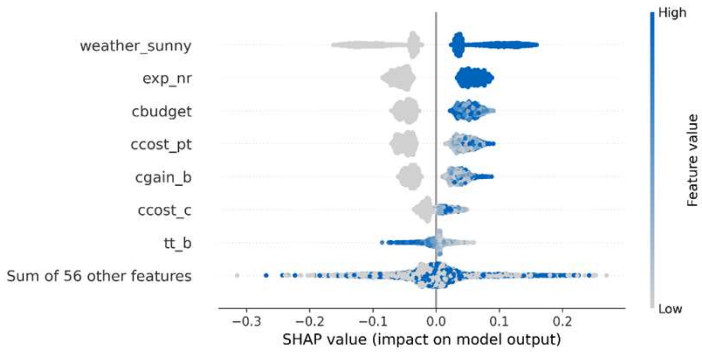


Aggregation Bias

Using discrete performance metrics like accuracy led to over a 130% increase in RMSE compared to continuous (probabilistic) metrics. This effect was especially pronounced in the underestimation of modes with low market share, such as MoCo Walk (as shown in the figure above).

Explainability

Sunny weather, Mobility Coin revenue, and travel time were identified as the most influential factors in choosing MoCo Bike, based on both RF and ANN models trained on grouped samples. In particular, higher Mobility Coin revenue increased the likelihood of users choosing biking, emphasizing the system's potential to promote sustainable mobility, while longer travel time had the opposite effect.



In-sample validation

In-sample validation showed a slight improvement in predictive performance compared to OOS validation; however, the difference was not statistically significant under ten-fold cross-validation for both RF & ANN models.

Conclusion

The choice of sampling method and performance metrics should be made carefully, with OOS validation recommended for large datasets. Model explainability helps modellers assess plausibility and gain insights into user behaviour.

R – random, S – sampled, D – discrete , M – model , G – grouped , C – continuous , OOS – out-of-sample