TECHNISCHE UNIVERSITÄT DARMSTADT

Thesis (B.Sc. / M.Sc.)
Knock, Knock. Who's There? Dataset Expansion and Model Generalization for Parcel Delivery Optimization

Department 18
Electrical Engineering and Information Technology
Self-Organizing Systems Lab

Prof. Dr. Heinz Koeppl
Head of lab

Background: Building on our successful development of a machine learning model for predicting reliable delivery time frames in parcel services, this thesis focuses on the critical next phase: expanding our dataset and validating model generalization across diverse user populations. Our existing model, developed in collaboration with the startup Green Convenience, has shown promising results on an initial student dataset. However, to ensure robustness and real-world applicability, we need to evaluate performance across broader demographic groups and larger datasets. This thesis will explore innovative approaches to user acquisition and conduct comprehensive generalization studies.

Philipp Froehlich
Project supervisor

S3|06 206
Merckstrasse 25
64283 Darmstadt

philipp.froehlich@tu-darmstadt.de
https://www.bcs.tu-darmstadt.de

June 16, 2025

Objective: This thesis addresses the crucial challenge of model validation and dataset expansion in machine learning applications. With our working model as a foundation, the project will focus on the following key areas:

- **User Acquisition Strategy Development:** Research and implement innovative methods to identify and recruit new test users from diverse demographic backgrounds. This includes exploring cost-effective recruitment channels, incentive structures, and partnership opportunities. Budget considerations and cost-benefit analysis will be conducted in collaboration with Green Convenience.

- **Dataset Expansion and Curation:** Design and execute a systematic approach to collect a significantly larger dataset that represents a broader cross-section of the target population. This involves establishing data collection protocols, ensuring data quality and consistency, and addressing potential biases in the expanded dataset.

- **Generalization Studies and Cross-Dataset Validation:** Conduct comprehensive experiments to evaluate model performance across different datasets. This includes comparing results between the existing student dataset and the newly acquired larger dataset, analyzing performance variations across demographic groups, and identifying factors that influence model

generalization.

- **Statistical Analysis and Reporting:** Apply rigorous statistical methods to assess model robustness, identify potential overfitting or underfitting issues, and quantify confidence intervals for model predictions. Develop comprehensive reporting frameworks to communicate findings effectively to both technical and business stakeholders.

This thesis provides an excellent opportunity to gain hands-on experience in the practical challenges of deploying machine learning models in real-world scenarios, with particular emphasis on data acquisition strategies, model validation, and business considerations in the logistics sector.

Prerequisites:

- Background in data science and modeling (e.g., students of electrical engineering, information theory, computer science, medical engineering, physics, mathematics).
- Familiarity with Python and statistical analysis tools.
- Interest in experimental design and user research methodologies.

For further information, please contact Philipp Froehlich.