THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

University of Chicago Library

The University of Chicago Library is facing a storage crisis, adding 80,000 volumes annually to an overflowing collection. As such, the library engaged the Data Science Clinic to develop a systematic process to identify materials to relocate to off-site storage.

This quarter, the team scaled the entire analysis pipeline to run remotely, allowing for more reproducible execution and faster processing. The team also implemented checkpoints that preserve intermediate results, reducing rerun time due to errors. To support collaboration, the team improved the documentation for each stage of the pipeline. These changes will allow future teams to rapidly get up to speed on the project.

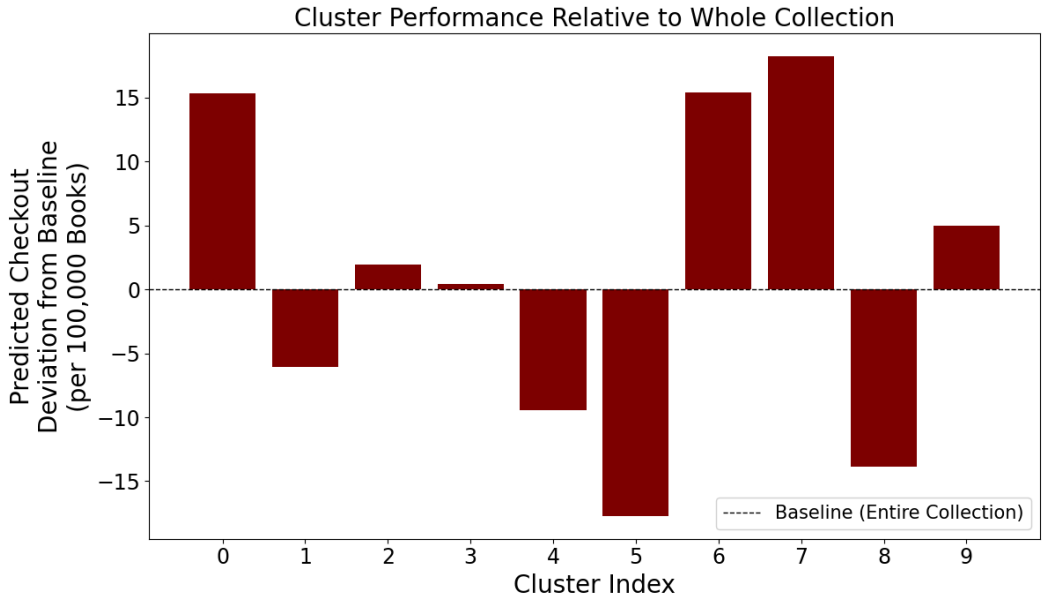

Figure 1: Cluster Performance scores clusters on their projected checkout rate relative to the entire collection's projected checkout rate. Projections are generated from historical usage and library attendance data. For each cluster, a separate algorithm assigns a 3-word label describing the topics captured by the cluster.

Key features used to assess a book's importance for on-site storage were also recalculated. The clustering and feature generation algorithms were optimized, enabling the team to iterate quickly and catch bugs early.

Each book is scored using a weighted sum of features, with the lowest-scoring items recommended for relocation. This quarter, the team set up a framework to assess the sensitivity of this scoring system to changes in feature weights. The team also consulted with Library staff about the data quality and perceived importance of various features. For example, Library staff recommended that the cluster performance feature, shown in Fig. 1, should be weighed more heavily. These insights will guide further refinement of the model and ensure alignment with the Library's institutional priorities.