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.

Figure 1: Left-to-right diagram of the analysis and recommendation pipeline.

First, the team corrected the data extraction logic of a few fields that were troublesome in Autumn 2024. This allowed the team to create a more accurate dataset of both the library's collections and its checkout information via a pre-processing pipeline.

Next, the team utilized pre-existing subject fields used widely across libraries to classify 100% of the records to support high-level analysis. Alongside other features calculated and extracted from the datasets, the team improved their weighted scoring algorithm that ranks volumes by importance for on-site storage. This improvement arose from the addition of new features such as predicting checkouts of a book in the future.

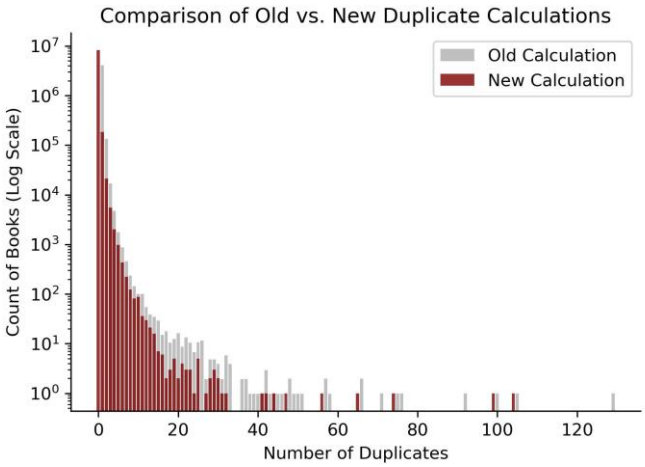

Figure 2: Distribution of the number of books that have a certain number of duplicate copies using both the Autumn quarter logic (grey) and improved Winter quarter logic (maroon). The shift shows a more accurate distribution of book duplicates.

With the library, the team optimized the scoring algorithm and extraction pipeline by examining existing and new features and established a replicable pipeline that provides the foundation for a reusable tool to assist with the library's storage issues in the future.