Reorienting Collection Analysis

Cost-Effective Item-Level Analysis and Machine Learning in Public Libraries

Ross Hanney

ABSTRACT

In public libraries, especially those in rural settings, it is important that every dime of library funding is leveraged effectively into serving the community. As part of a year-long project beginning in January 2023, we are evaluating item-level cost-effectiveness for each circulating item housed at the public library in Lakeville, Indiana. Through the use of big(ish) data, some custom Python scripting, and machine learning algorithms we hope to answer: How much money is saved by library patrons through their use of the public library's physical collection? How much money is saved by the community through the operation of a public library based on the use of the circulating collection? And are there any non-obvious traits which make an item or title a more or less cost-effective circulating asset? In this column, I will describe the scripts, share initial findings, discuss challenges, and investigate next steps.

INTRODUCTION

There is so much to be said about the value public libraries bring to the communities they serve. This column is not going to investigate the worth of programs or experiences, the social interactions taking place, the knowledge or education gained by patrons, or the many other benefits that endure because of the existence of public libraries.

Instead, this column will discuss the dollars and cents value that public libraries bring to their community using circulating collections. By using data available through the library's Integrated Library System (ILS), I examine how cost-effective each item is throughout any given month, quarter, or year, and what, if anything, makes an item more or less cost effective.

About the Community

This project focused on a public library branch located in Lakeville, Indiana, one of the ten locations operated by the St. Joe County Public Library system. Lakeville is a rural farming town with a population of 878 people.

What Prompted the Work?

During the past few years, the St. Joe County Public Library had received some concern from the community regarding the amount of money being used toward library operations. These concerns were spurred by the renovation of the largest location in downtown South Bend, which took two years and cost an estimated $38 million to complete. With the added financial pressure on households brought on by the COVID-19 pandemic and the marked increase in inflation, some patrons were beginning to voice concerns about the financial benefit of public libraries.

About the Author

Ross Hanney (corresponding author: r.hanney@sjcpl.org) is Staff Development Coordinator, St. Joe County (Indiana) Public Library. © 2023.

https://doi.org/10.5860/ital.v42i4.16987

PART I: THE MONEY SAVED

With the goal of determining the financial benefits of the library's collection, I took the first step towards devising an effective method to best measure the value of any given item in the library's collection at the branch. The cost of the item and the number of times it was checked out was reviewed. While not a perfect system, this method provides the total amount of money saved by library patrons who check out the material rather than purchase it for themselves. The formula was the following:

$$\text{Money\_Saved} = \text{Times\_Circulated} * \text{Item\_Cost}$$

Table 1 shows how the amount of money saved by patrons, without accounting for operating expenses, was calculated. The INIT CHKOUT and SECD CHKOUT fields relate to the first and second time the data was harvested from the library's database. The difference of these two variables yields the number of times the item was checked out within a certain time period. In this instance, it was a month. Each of the items above are from a different collection, a non-fiction book, holiday DVD, picture book, and a beginner reader.

Table 1.

BARCODE	INIT CHKOUT	SECD CHKOUT	TOT CHKOUT	PRICE	SAVED
3 1986 03488 2657	50	51	1	$19.95	$19.95
3 1986 03656 8510	189	190	1	$30.00	$30.00
3 1986 03703 8992	23	23	0	$16.00	$0.00
3 1986 05242 4986	44	46	2	$16.99	$33.98

The second step was to design a way to discover how much money the circulating collection saved the community. Uncovering this information was much more difficult considering the costs associated with the operation of the library and that not all operating costs are directly focused on the collection. For the purposes of this project, the cost of one full-time Circulation Assistant per 10,000 items, building and utility costs, subscription service costs necessary for collection management, and the cost of new materials were used to calculate operating expenses. The following formula was used:

$$\text{Money\_Saved} = (\text{Times\_Circulated} * \text{Item\_Cost}) - (\text{Operating\_Expenses} / \text{Number\_of\_Items})$$

While this method is not a perfect way of looking at the value of each item based on the cost to the community, it provides a framework for consistently measuring this metric, which met the needs for this evaluation.

PART II: METHODOLOGY AND DATA ANALYSIS

To figure out answers to the two questions (How much money is saved by library patrons through their use of the public library's physical collection? How much money is saved by the community through the operation of a public library based on the use of the circulating collection?), I harvested and analyzed select data from the library's ILS to gain insight into what makes each item more or less cost effective. At the beginning of each month, I exported the following item-level data from the ILS:

- item barcode or specific identifier
- total number of checkouts throughout the life of the item
- initial cost of the item at time of purchase
- item's collection area (Children's Picture Books, Adult Mystery, etc.)
- call number of the item
- publication statement (MARC 260: Publisher, publication date, region of publication)
- subject added entries (MARC 650: Common search terms linking to a controlled vocabulary)
- index terms for genre/form (MARC 655: Words or phrases used to aid in the discovery of the item on the library's online public access catalog)

Table 2 shows example data harvested from the ILS.

Table 2. Sample data harvested from the library's ILS.

BARCODE	TOTAL CHECK OUTS	PRICE	LOC	CALL #	MARC 260	MARC 650	MARC 655
3 1986 01685 5804	50	$15.95	lkvch	j 394.268 G352s, Easy	New York : Holiday House, c1994.	Saint Patrick's Day Juvenile literature.	NULL
3 1986 03069 0336	56	$20.00	lkvav	DVD MUSICAL Whi	Hollywood, Calif. : Paramount, c2000.	Entertainers United States Drama.; World War, 1939-1945 Veterans United States Drama.; Man-woman relationships United States Drama.	Christmas plays. Video recordings for the hearing impaired. Musical films. Christmas films. Feature films.
3 1986 06285 2234	6	$31.00	lkvlp	Lg Print Fiction Gra	NULL	Amish Fiction.; Interfaith dating Fiction.; Teenage pregnancy Fiction.; Large type books.	Religious fiction. Love stories. Christian fiction.

Each of the datasets exported consisted of, on average, 12,000 rows, one row for each item in the Lakeville branch library's collection.

PART III: THE SCRIPTS

Working with such large datasets by hand would have been nearly impossible due to the time it would have taken and would have likely resulted in the presence of human errors. For these reasons, I wrote Python scripts to handle the data within these datasets. These scripts generally fall into two categories: Data Pre-Processing and Data Analysis. Scripts are available in Github at https://www.github.com/SardineDude1/CEILA.

The scripts contained within the Data Pre-Processing category prepare the data to be analyzed and include format manipulation, concatenation, and the generation of training data.

The scripts which fall under the Data Analysis category generate useful information in understanding the pre-processed data. One of these scripts consists of a multi-input neural network model that combines text and numeric inputs, concatenation layers, and a dense output layer for classification. In other words, it is a predictive model which uses harvested item-level data and predicts whether the item would be cost effective or not. The neural network uses the item's collection, the publication statement, the number of subject added entries, and the cost of the item to predict the value rating of the item on a scale of 1 to 3, with 1 being not cost effective and 3 being cost effective.

PART IV: INITIAL FINDINGS

Money Saved

On average, patrons of the public library in Lakeville, Indiana collectively saved approximately $18,500 each month by checking items out rather than purchasing them. After accounting for operating expenses, the community saved on average an estimated $8,500 each month.

In St. Joseph County, the library is funded mainly by property taxes. The average Lakeville household pays $3,471.40 per year in property taxes. However, only a small portion of those tax dollars ever reach the library. These funds amount to an estimated $16 per month per resident. After library operating expenses are considered, over half of the taxpayer’s obligation is recouped through their use of the circulating collection.

Machine Learning

Once complete, the neural network was able to provide accurate predictions about an item's cost-effectiveness, with an average accuracy rating of 92% . This rating was determined by separating the available data into training data (90% of the dataset) and testing data (10% of the dataset). Once the model was trained using the training data, it was asked to make predictions about the testing data it was not trained on. For example, if a dataset was 12,000 items, the neural network would be trained on 10,800 items. When asked to predict whether or not an item was cost-effective on the remaining 1,200, it got 1,104 predictions correct.

After an investigation into the internal architecture of the trained model, I discovered that the most important feature in predicting if an item is more or less cost effective is the publication statement. However, the number of subject fields was also highly weighted and was only slightly less important when forming predictions. The third most important feature was the genre or collection to which the item belongs. which was still weighted more than twice as important as the cost of the item.

These preliminary findings are based on initial results of the predictive model. More work is required to determine if this is a viable means of predicting an item's cost effectiveness using only these four variables.

PART V: CHALLENGES AND LIMITATIONS

Several challenges were encountered throughout the course of this project, including the availability of data, the amount of time needed to pre-process data and train an artificial neural network on non-specialized equipment, and my own knowledge of artificial intelligence and neural networks.

The infrastructure for data analysis at the item-level was not something our library had given a lot of thought to before this project. There were no formal guidelines or staff training on how to analyze this kind of circulation data or even where to find this data. However, beginning with the ILS user manual and ending with a series of long conversations with the database librarian, a temporary framework was established, and I was able to get the data needed.

Once the data was acquired, it needed to be processed. It took approximately 20 minutes of processing time to prepare a single dataset of about 12,000 lines. This does not include the time it takes to check for errors in preprocessing scripts, the training of the neural network, or backing up the data. This process could have been sped up with higher quality or specialized equipment, but there was none available, so I made do with the resources I had on hand.

The last barrier encountered during this project was the limited knowledge of neural networks and artificial intelligence model architecture. It turns out, being a public librarian does not prime someone for an easy foray into the world of AI development or even computer programming. However, with much time and energy, a functional neural network was born after hours and hours of testing and troubleshooting.

PART VI: CONCLUSION AND NEXT STEPS

I was pleasantly surprised to discover the initial amount of money saved by patrons who check items out at the library rather than purchasing the items for themselves. This metric could be used to aid in demonstrating the dollar and cents value library users get access to through their local library.

The findings surrounding the money saved by the community, albeit an estimated amount, was also encouraging. With this metric in mind, perhaps steps could be taken to increase the monetary value the library presents to the community through its circulating collection. While public libraries are so much more than the materials they make available to their patrons, continuing to provide materials that aid in the economic support of the community should be of some consideration.

Finally, after analyzing the results from the neural network, some non-obvious traits were discovered to have an impact on the cost-effectiveness of items. The publication statement being the highest weighted feature in forming a prediction was quite striking. This suggests that the MARC 260 field has the highest influence on an item's circulation, and by extension its cost-effectiveness, out of the four inputs investigated. It was equally remarkable that the cost of the item had the least amount of influence in making a prediction. This seems to indicate that the cost of a circulating object holds little influence on how well the item will circulate.

While a more comprehensive look at the relationships between item-level data and its effects on circulation is needed before implementing this neural network as a collection development tool, the information gained from looking at how much money patrons and the community save by using the library is certainly usable now. In fact, it just might make it into the next library newsletter.