Modeling Time-to-Trigger in Library Demand-Driven Acquisitions via Survival Analysis

Zhehan Jiang – University of Alabama
Sarah Rose Fitzgerald – University of Alabama
Kevin W. Walker – University of Alabama

Citation of published version:

Jiang, Z., Fitzgerald, S., Walker, K. (2019): Modeling Time-to-Trigger in Library Demand-Driven Acquisitions via Survival Analysis. *Library and Information Science Research*, 41(3). DOI: https://doi.org/10.1016/j.lisr.2019.100968

Library and Information Science Research xxx (xxxx) xxxx

Contents lists available at ScienceDirect

Library and Information Science Research

journal homepage: www.elsevier.com/locate/lisres

Modeling time-to-trigger in library demand-driven acquisitions via survival analysis

Zhehan Jiang*, Sarah Rose Fitzgerald, Kevin W. Walker

University Libraries, University of Alabama, Box 870266, Tuscaloosa, AL 35487, USA

A B S T R A C T

Conventional statistical methods (e.g. logistics regression, decision tree, etc.) have been used to analyze library demand-driven acquisitions (DDA) data. However, these methods are not well-suited to predict when acquisitions will be triggered or how long e-books will remain unused. Survival analysis, a statistical method commonly used in clinical research and medical trials, was employed to predict the time-to-trigger for DDA purchases within the context of a large research university library. By predicting which e-books will be triggered (i.e., purchased), as well as the time to trigger occurrence, the method tested in this study provides libraries a deeper understanding of factors influencing their DDA purchasing patterns. This understanding will help libraries optimize their DDA profile management and DDA budgets. This research provides a demonstration of how data science techniques can be of value for the library environment.

1. Introduction

Due to the rising costs of library resources and the stagnant budgets of academic libraries, cost effectiveness in collection development is vital. While some librarians believe it a less-than-ideal solution, many have suggested that libraries limit their collection spending to those items that are of immediate use to their patrons. One way to ensure usage follows from collection development spending is to allow patrons to purchase materials for the library. This is particularly easy with e-books, whose records can be added to a library's discovery layer before they are purchased for the library. The process of purchasing e-books based on patron usage is called Demand Driven Acquisitions (DDA) or sometimes Patron Driven Acquisitions. It is acknowledged that DDA is still viewed with resistance by some librarians, especially in libraries where funds are more available. This paper is not an argument for replacing all traditional collection development methods with DDA. It is an exploration of optimizing the utility of DDA programs.

In an e-book DDA program, a library selects a pool of titles based on criteria they communicate with a DDA vendor. These titles are then made visible in the library's online catalog or discovery layer as if they are titles owned by the library. However, these titles are not actually purchased by the library until sufficient usage by patrons occurs to trigger them. Triggers occur after a certain number of uses, such as when a patron downloads a book, prints from it, reads a certain number of pages, or accesses it for a certain amount of time, specified by or negotiated with the vendor. In order to increase offerings to patrons or limit spending, librarians can adjust the criteria that determine which titles appear in the pool according to various variables including publication year, publisher, subject, and price.

In an age where higher education is increasingly digital, libraries are growing their e-book collections. According to the annual Association of College and Research Libraries survey for fiscal year 2017, 36% of book titles at doctoral granting university libraries were e-books (Association of College and Research Libraries, 2019). DDA is one method libraries are using to develop these collections. In the last decade, many libraries have participated in e-book DDA programs and published about their various successes andDewland & See, 2014 challenges, some of which will be discussed further below (Breitbach & Lambert, 2011; ; Gilbertson, McKee, & Salisbury, 2014; Zhang & Downey, 2017). E-book DDA is now an important portion of academic library acquisitions. However, in order to maximize the utility of these programs and library budgets for such programs, more data is needed.

2. Problem statement

While DDA is a benefit to libraries in terms of providing content to their users based on demand for an item rather than librarian estimates of future demand, it can also be a challenge to libraries because of the unpredictability of the costs associated with allowing users to make library purchases. A machine learning approach to predicting when e-books will be triggered for purchase based on DDA data from a large research university is presented as one potential the method libraries can use to gain a deeper understanding of factors influencing their DDA purchasing patterns.

* Corresponding author.
E-mail address: zjiang17@ua.edu (Z. Jiang).

https://doi.org/10.1016/j.lisr.2019.100968
Received 28 January 2019; Received in revised form 11 June 2019; Accepted 2 July 2019

3. Literature review

In a higher education environment where costs are formidable for students and student success has been tied to the affordability of the texts they need for their education it is important that libraries support the real needs of their patrons, rather than building a collection of items librarians wish they would use. For a single university, collecting all academic texts is fiscally unrealistic and best left to libraries and consortia whose mission is preserving the scholarly record rather than university libraries with missions to serve the needs of their current and future patrons. Cost efficiency in collection development is about maximizing the utility of library budgets for patrons to support their research and learning. This study tests the viability of a method for predicting the needs of their patrons based on prior DDA purchasing history as a way of meeting their patrons' needs within a budget.

DDA has a variety of benefits for libraries. It leads to higher circulation than librarian selection or approval plans (Tyler, Falci, Melvin, Epp, & Kreps, 2013). Anderson et al. (2002) noted DDA's advantage to fill interdisciplinary gaps in collections, which do not necessary fall into any librarian liaison's subject responsibilities. This is especially important as funders call for increased interdisciplinarity from higher education. In addition, research has shown that DDA cost-per-use is lower than for librarian selected purchasing (Howland et al., 2014; Schroeder, 2012; Tyler, Xu, Melvin, Epp, & Kreps, 2011; Way & Garrison, 2011). The advantage in cost-per-use of DDA titles at the university of study has also been demonstrated (Walker & Arthur, 2018). However, even though there is evidence that DDA is more cost effective than librarian selection at this institution, the program still presents a challenge because the spending pressures on the library's budget. Therefore, this study examines the data to help with decisions about how to limit spending. Other libraries are likely to share this dilemma and benefit from the strategy outlined.

Despite the benefits of DDA programs, some weaknesses of these programs have been identified. Rather than allow these drawbacks to eliminate DDA as a tool for collection development, it is best to understand and give libraries a tool to address the shortcomings of DDA. Walters (2012) objects to DDA purchasing because it leaves the decisions of purchasing in the hands of students rather than prioritizing faculty as many library request programs do. While it is important for the network of academic libraries across the nation and the world to continue to offer access to esoteric materials needed by researchers on an irregular basis, an individual university library must offer ready access to the materials most often required by its patrons. This means purchasing items that reflect the needs of students, who make up the bulk of the patron base. Although Tyler, Hitt, Nterful, and Mettling (2019) suggest that books selected by librarians are more heavily cited than books selected through DDA in a majority of topics, their study did not account for the fact that librarians can preorder books, which means all DDA purchases come from a set of titles librarians did not choose to preorder. Regardless of the validity of the study, university libraries serve all readers, not just citers. Citation is a measure of utility to researchers everywhere, not a measure of utility to readers at the university purchasing the material. Many DDA programs require several uses of an item before a purchase is triggered. Because it reflects repeated patron needs, DDA is a way to collect materials which are clearly useful to patrons. It supports the information needs of learners who have fewer avenues for input into library collections compared with their professors. As opposed to print purchasing or interlibrary loan, e-book DDA offers instant access to titles which can be important to students who often have short course assignment deadlines compared to the longer deadlines of conference proposal and grant submissions which tend to shape the timing of faculty work.

Sens and Fonseca (2013) argue that it is the librarian's job to distinguish between titles of high or low value for library users, but research shows that users make quality choices (Tyler, Melvin, Epp, & Kreps, 2014). While Waller (2013) found that undergraduate collection building at one small liberal arts college led to poor circulation, that finding was based on print books purchased through ILL requests, not ebooks purchased through DDA integrated in a discovery layer. This makes a difference because print circulation is declining nationally, while ebook use is increasing. Now that the first stop for most users in locating information is more often a search engine than a library or its website, the primary role of librarians is to educate students on differentiating between resources based on their quality. An undergraduate education today should require of students that they select appropriate materials for their course projects. Asserting that students are not fit to judge what resources are right for their needs is in opposition to the assertion in American Library Association's (2015) Framework for Information Literacy that “Authority is constructed and contextual”. DDA gives students the opportunity to select resources which are appropriate to their personal contexts, which are not shared by collection development librarians. It is in line with the principles of student-centered learning, a growing trend among higher education institutions. To align themselves with the current state of higher education, librarians should be teaching students to make good resource choices, not making all the choices for them. That said, librarians continue to have important roles in selection, including evaluating collection strategies through processes like DDA.

Another objection Walters (2012) noted to DDA programs is that budgeting for them is difficult, because purchasing is in the hands of patrons. He also pointed out that many of the strategies employed by libraries to check their DDA spending are problematic. For instance, if purchasing is suspended after the library reaches its annual budget for DDA, that could lead to missed titles in the later parts of every fiscal year. The proposed survival analysis approach provides a way that librarians can get more information about the purchasing decisions their patrons are making in order to optimize DDA selections within a budget.

3.1. Data science and survival analysis

Recent advances in statistics and computer science, combined with an abundance of readily available data, have given rise to increased use of data science for decision making (Oliver, Kollen, Hickson, & Rios, 2019). Data science methods and products are transforming commerce, healthcare, and government and will continue to transform other sectors, including libraries (Oliver et al., 2019). The Federal Big Data Research and Development Strategic Plan, released under the Obama administration, explicitly identifies curators, librarians, and archivists as core specialists to help to meet a growing demand for analytical talent and capacity across all sectors of the national workforce (The Networking and Information Technology Research and Development Program, 2016). Meanwhile, libraries can adopt data science to improve their operations, which will help to improve services to patrons. For instance, Walker and Jiang (2019) used machine learning to handle DDA prediction, where Jiang and Fitzgerald (2019) deployed the change-point model to study institutional repositories.

Survival analysis is concerned with analyzing the amount of time that transpires between the start of an observation period and when some predetermined event occurs (Machin, Cheung, & Parmar, 2006, p. 22). Such analyses are frequently used within the context of clinical trials, in which death is often the predetermined event of concern. In those cases, the researcher wants to know how long it takes for a person under treatment to die or the probability that a patient survives in a given observation period. The researcher might also want to know what differences exist between groups in terms of survival status. Of course, there are cases where patients do not die during the course of the clinical trial. In these cases, data must be censored so that the missing value (e.g., time until death) is not a disqualifying factor for the case's inclusion in the. For example, if the trial followed patient progress for five years post-treatment, those cases where death does not occur may receive a time-to-event value of 60 months. Another example of the

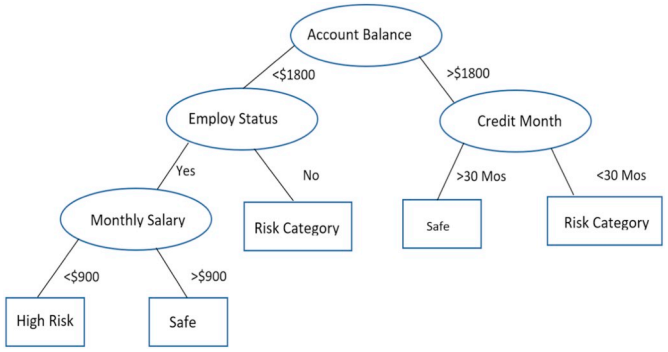

Fig. 1. A decision tree example used in many banks.

utility of survival analysis is churn studies conducted by companies to predict the adhesiveness of customers (Bradburn, Clark, Love, & Altman, 2003). In this research, e-books rather than patients or customers are the subjects under observation, while trigger rather than death is the event of interest. Detailed explanations of survival analysis and helpful examples of this method in action can be found in the work of Allison (2010) and Cox (2018).

The present survival analysis is based on random forests. The random forest (RF) family has proven to be a highly accurate and effective alternative to a classical decision tree model (Breiman, 2001). To understand the RF, one needs to understand the basic functions of its constituent unit—the decision tree (see Jiang, Walker, & Shi, 2019 for explanation). To illustrate this analytical device in action, Fig. 1 demonstrates how the evaluation of bankruptcy risk is carried out via decision tree. In this example, the set of predictors includes Account Balance, Employment Status, Credit Month, and Monthly Salary. Prediction outcomes (i.e., classes) are labeled as High Risk or Safe. Note that Fig. 1 shows only one tree. Theoretically, depending on the model parameters that a researcher employs (including the number of analyses carried out), such an analysis could encompass an infinite number of tree combinations—with each tree producing different results.

As seen in Fig. 2, the RF is an ensemble of decision trees, each of which gives a biased classifier (as it only considers a random subset of the data). This ensemble process is like a team of experts, each with a little knowledge over the overall subject, but thorough knowledge in their area of expertise. Mogensen, Ishwaran, and Gerds (2012) compared the RF survival analysis with other strategies such as Cox regression or additive hazard regression in a large-scale simulation study, and found that the RF version, a non-parametric model with fewer

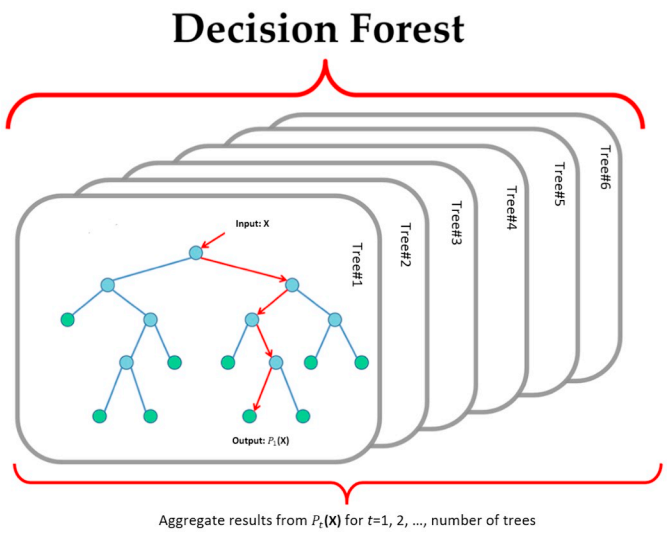

Fig. 2. A random forest is consisted of multiple decision trees.

required assumptions, provides more robust estimates than its competitors do, in many situations. To this end, this RF variant named random survival forest was adopted in this study (Ishwaran, Kogalur, Blackstone, & Lauer, 2008), to perform survival analysis; this RF variant is called RSF.

4. Methods

In a survival analysis, time to event (e.g., dropout and death) t is the parameter of interest and is assumed to follow a probability density distribution $p(t)$. This density is further defined by the survival function $S(t) = Pr(T > t) = \int_t^{\infty} P(T)dT$ for any $t > 0$. $S(t)$ represents the probability that the event occurs at time t. Given $p(t)$ and $S(t)$, one can model the event rate at time t by making $\lambda(t) = \frac{p(t)}{S(t)}$. Conventionally, $\lambda(t)$ is called the hazard function outlined as: $\lim_{\Delta t \to 0} (\Pr(t < T < t + \Delta t \mid T > t)) / \Delta t$. Via certain mathematical transformations, one will find that $S(t) = e^{-\int_0 \lambda(t)dt}$.

With a parametric framework, the likelihood function for right-censored survival data can be represented as $L(\theta|\{x_i, t_i, \delta_i\}_{i=1}^n) = \prod_{i=1}^n \lambda(t_i \mid x_i, \theta)^{\delta_i} S(t_i \mid x_i, \theta)$, where θ is the model parameters, δ_i is the event indicator (i.e., $\delta_i = 1$ if the event occurs and 0 otherwise), and n is the total number of subjects/individuals. Perhaps the most well-known modeling strategy is setting $\lambda(t|x, \theta) = \lambda_0(t) \exp\{x^T\theta\}$; this modeling variant is called the proportional hazard model (Cox, 1992). With the assumption about $\lambda(t|x, \theta)$, Cox (1992) proves that $\lambda_0(t)$ can be removed such that the likelihood function can be expressed as: $L_{partial}(\theta|\{x_i, t_i, \delta_i\}_{i=1}^n) = \prod_{i \in E} \frac{\exp\{x_i^T\theta\}}{\sum_{j:t_j \geq t_i} \exp\{x_j^T\theta\}}$, where E is the subset of the data with $\delta = 1$ and the subscript *partial* differentiates the likelihood from the previous one as $\lambda_0(t)$ is ignored. The proportional hazard model, despite the high popularity, is not always appropriate due to the strong assumption.

The previous two paragraphs can be boiled down to a few simple concepts. First, time-to-event is the central variable of interest within the context of survival analysis. Second, prior to any specific phenomenon via this method, it is expected that time-to-event for the individuals within the population under study will conform to a distribution pattern where the probability of the event for any particular member of that population will differ based on a one or more determinant factors. That is to say, not everyone will experience the event, and those who experience it will likely not experience it at the exact same time. Therefore, it can be understood the underlying phenomenon by determining how those differences arise, based on variations in the aforementioned determinant factors.

The typical DDA program presents parameters that are ripe for survival analysis. A number of e-books (N) are selected for inclusion within a DDA title pool, which was initiated at some point in time we can call Year 0. From Year 0 to the current year (Year 3), some titles will be triggered for purchase, additional titles will be added to the pool (and perhaps be triggered as well), and some titles will not have been triggered for purchase. In this scenario, the population is e-book titles rather than patients and the event is purchase rather than death.

The feasibility of using the survival modeling strategy was employed to predict which e-books within a DDA program's title pool were triggered for purchase. The vendor-provided data used in this analysis was generated by DDA purchasing and use activity within a large research library environment over a period of 24 months. A sample of 5617 e-book titles was analyzed. This includes 3417 triggered (i.e., purchased) and 2145 untriggered e-book titles. These titles represent works produced by 202 different publishers, across 14 publication years, and feature a total retail value of nearly $455 k. A total of 2145 of the e-books in the DDA pool were not accessed, but the remaining 62% were triggered for purchase.

For this study, the days to trigger variable represents the time

interval between when a title is added to the DDA pool and when it is triggered for purchase. For those titles that remained untriggered, the days to trigger is set to zero. In addition to days-to-trigger, publisher, publication year, retail price, and Library of Congress (LC) Classification subclass are included as variables of interest.

Unlike traditional parametric approaches (e.g., logistic and/or probit regression models), which are most often used to support statistical inference, the RF survival analysis used here is nonparametric and employed purely as a predictive method. That is to say, this study is not aimed at explaining why e-book titles are, or are not, triggered for purchase. Instead, this research aims to determine whether or not survival analysis can be used to make predictions about triggering phenomena within the context of active DDA program deployment.

King (2003) notes the "area under the curve" (AUC) provides a means of measuring predictive capacity for prediction models. This study adopts AUC as a measure of model value relative to the binary assignment problem presented by DDA e-book trigger status. Ideally, a perfect classification is present when AUC = 1. In practice, many "best-performing behavior checklists and inventories that are currently available yield AUC values between 0.8 to 0.9 under clinically realistic conditions and with valid reference standard diagnoses" (Youngstrom, 2013, p.1).

R software was used to analyze the data (R Core Team, 2018). Specifically, the package Survival (Lumley & Therneau, 2004) was used to execute the algorithmic core of the survival analysis, while the package gbm (Ridgeway, 2015) supports RF modeling functions within the Survival package framework. Additionally, the package pROC (Robin et al., 2011) was used to calculate the AUC for each model. Given the computational burden of these methods (Wu & Nagahashi, 2014), a workstation computing facility with four Intel® Core™ or Intel® Xeon® processors and 16 GB RAM was used. Table 1 shows three sample records of the dataset.

5. Findings

Fig. 3 shows the histogram of the days to trigger for all titles analyzed. The turquoise color represents the number of days it took to trigger items, while the black color represents the number of days items that have not yet been triggered have been in the pool. The majority of the triggered items were accessed within 50 days from the date they were included into the DDA pool, while the untriggered group spans the time with a longer tail. Similarly, the distribution of the retail price can be found in Table 2. As shown, the price range for triggered titles is wider than seen with untriggered titles. In addition, at all key distributional levels (i.e., mean, median, and some quantiles), the triggered group consistently yields equal or higher values. The top three publishers for the untriggered group were Taylor & Francis (631), McFarland & Company (162), and ABC-CLIO (88). The top three publishers for the triggered ones were Taylor & Francis (1141), Elsevier (127), and Oxford University Press USA (125). Note that Taylor & Francis appear in each group because it has a substantial collection volume and therefore they are not mutually exclusive.

In terms of Library of Congress (LC) class, this dataset does not include any category with one record only and therefore avoids the situation where the LC class provides meaningless information. The top three LC classes for the untriggered group were PN-Literature (General) (166), HV-Social pathology, Social and public welfare, Criminology (121), and LB-Theory and practice of education (84), while the highest statistics for the triggered group were HV (251), LB (163), and HQ-The family, marriage, women (151). Fig. 4 shows the density plot for the publication year: most of the triggered items were published more recently, where the majority of the untriggered ones were published between 2010 and 2015. From these descriptive statistics, it is evident that the selected predictors vary substantially in accordance with the trigger status.

The first model provided is a full model where all variables were loaded into the analysis. Using the default algorithmic configurations and convergence assessment named the out-of-bag method, which was suggested by Ridgeway (2015), the model was shown to converge at the 1689th iteration, as shown in the left panel of Fig. 5; that said, the results at the 1689th iteration are parameter estimates. The right panel of Fig. 5 shows the relative influence of the predictors; it shows that the publisher is more deterministic than any other predictor is, while the retail price is the least important in the model. Note that these influences are in a relative scale; unlike traditional models, the results yielded by the RF survival analysis do not provide further interpretability. Therefore, the model cannot be used to make an inference for other libraries. The AUC for the current model is 0.81. According to Youngstrom (2013), given the AUC is larger than 0.8, the current model possesses good prediction capacity.

This research models and test a prediction machine, thus for any arbitrary e-book record, the machine will predict (1) the time (in days) this record would take to be triggered and (2) the trigger status for a given time interval. For instance, if {Price = 73, Publisher = ABC-CLIO, LC Class = DT, Pub Year = 2017} is fed into the model, the predicted days for this item to be triggered is 87 days; similarly, if one is interested in the trigger status within 30 days since the item was included in the DDA pool, the predicted outcome is 0, meaning the title remains untriggered at 30 days.

To align with the relative influence result in Fig. 5, parsimonious models were constructed for fit comparison. Table 3 shows the relative predictive capacity of these models. As shown in Table 3, predictive capacity of the model increases as the model becomes more complex, though this will not always be the case, due to certain random processes inherent to the algorithm. Model 1 uses the most influential predictor within these data (i.e., publisher), resulting in a model with acceptable predictive power (AUC = 0.69) despite the model's incredible simplicity (i.e., use of only one predictor). Nevertheless, Model 2 sees a substantial boost in AUC with the addition of the publication year variable. Finally, the incremental gains witnessed with the added complexity of Model 3 and Model 4 are relatively small, though Model 4 (the model using all 6 predictors) produces an AUC higher than 0.8. These findings are consistent with the measures of relative predictor influence shown in Fig. 5. That is to say, one would expect that less influential predictors contribute less power to the predictive capacity of any model.

6. Discussion

The present survival analysis application demonstrates an effective deployment of machine learning within the library environment. The researchers successfully deployed survival analysis within a random forests learning framework to predict purchasing for a large research library's DDA program. Multiple models were tested within this

Table 1
Samples of the selected DDA pool.

ID	Title	Days to trigger	Triggered	Price	Publisher	Pub year	LC_class
16,303	Race and Work	73	Yes	$120	Taylor & Francis	2016	HD
16,404	Childhood Studies	80	Yes	$55	MIT Press	2012	HM
16,333	Search Society	0	No	$85	CRC Press	2008	PR

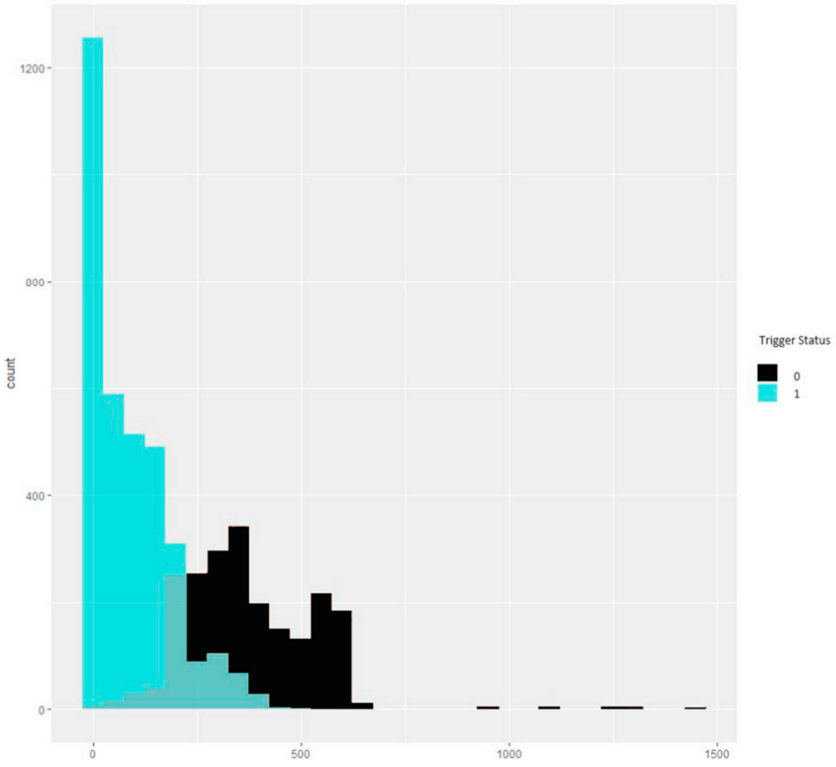

Fig. 3. Histogram of the days to trigger.

Table 2
The summary statistics for the retail price in US dollars.

Trigger	Retail price					
	Min	1st quantile	Median	Mean	3rd quantile	Max
Yes	4	60	112	102	145	494
No	4	15	33	46	45	261

framework, the most robust of which showed a predictive capacity of greater than 80% accuracy.

There are several possibilities of how forecasting DDA triggers might help to improve DDA program outcomes. For example, one might load into the model a list of titles that are not currently found in the pool. Then, based on each title's predicted time-to-trigger, a library could choose whether or not they want to clutter their pool with titles that would trigger beyond a pre-set trigger date of their choosing. Another approach would involve extrapolating the projected annual DDA program costs, based on predicted triggers (e.g., that will occur in less than 365 days). Armed with those data, one might either expand or trim their title pool to help prevent overspending. One might extrapolate detailed fund expenditures, based on LC sub-classifications—resulting in either a reassessment of fund allocations or DDA selection profile parameters.

6.1. Local insights & implications

This study revealed several interesting patterns within the local DDA data. For example, many general literature titles remain untriggered for long periods of time as compared to titles within other subject areas. One possible explanation is that patrons are less interested in e-book versions of literature titles. This is in line with the Ithaka S + R survey findings which show humanities faculty are more interested in print than electronic monographs and more interested in print monographs than other disciplines (Wolff-Eisenberg, Rod, & Schonfeld, 2016). Supporting this, librarians at the institution under study, through feedback received via their outreach work, have learned that English faculty have a strong preference for titles in tangible print format. Based on this pattern, it would seem advisable to continue the purchase of literature titles in tangible print format. Another potential explanation for untriggered literature titles is that the humanities have a much greater emphasis on monograph publishing than other disciplines, and, therefore, these titles comprise a larger proportion of the local DDA pool than offerings for other disciplines.

While it is possible that users at other libraries also prefer literature titles in tangible print, the usage and spending distributions witnessed in this study may not be generalizable to the broader community of research libraries. Each university has its own specialties, in terms of the degrees it offers, which shape the needs of its users. Additionally, the collection makeup of each library differs in ways that will shape the gaps that can be filled through a DDA program. There are also variations in how libraries subscribe to large e-book packages from vendors such as Wiley or Springer, which will impact what titles are included in their DDA pools (to avoid overlap). The authors recommend that each library investigate these trends among their own patrons and collections.

Another interesting local pattern discovered through this analysis is that titles published after DDA program implementation are more likely to be triggered within a short time than titles published before the DDA program was launched. This is likely due to the prioritization of more recently published titles within search results delivered by the library's discovery layer. If so, it would mean that titles published before DDA program launch are rarely appearing in the first page or two of search results. The lack of control libraries have over the search algorithms employed by vendor-provided discovery services may prove problematic for libraries seeking to fill historical gaps in their collections through DDA programs.

In cases where a proprietary discovery service provides the main point of contact between library patrons and collections, analyses such as those shared here are likely to prove as telling of the discovery layer

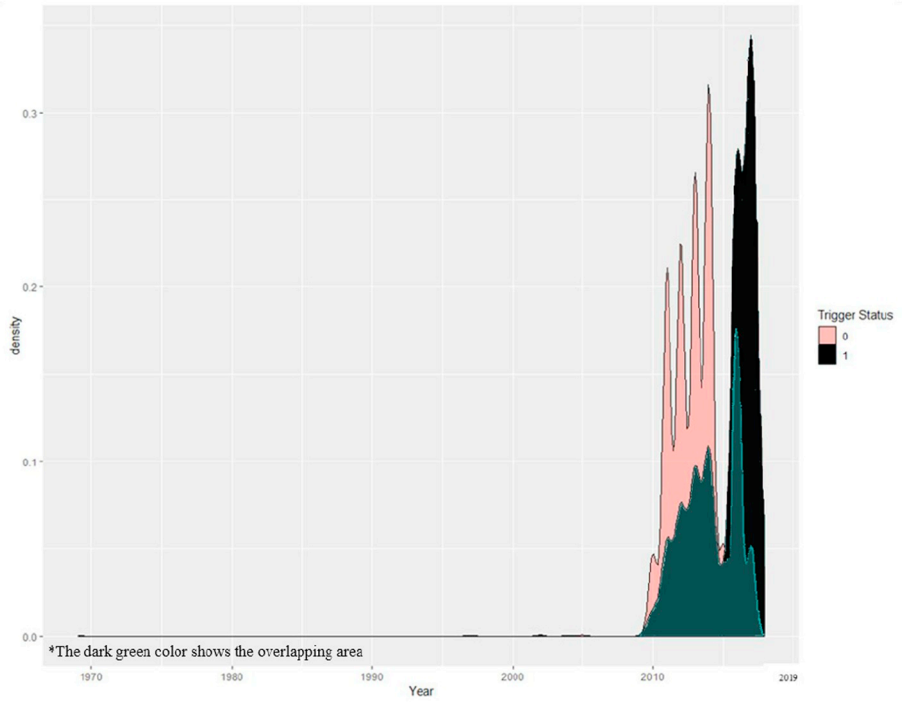

Fig. 4. Density plot of the publication year.

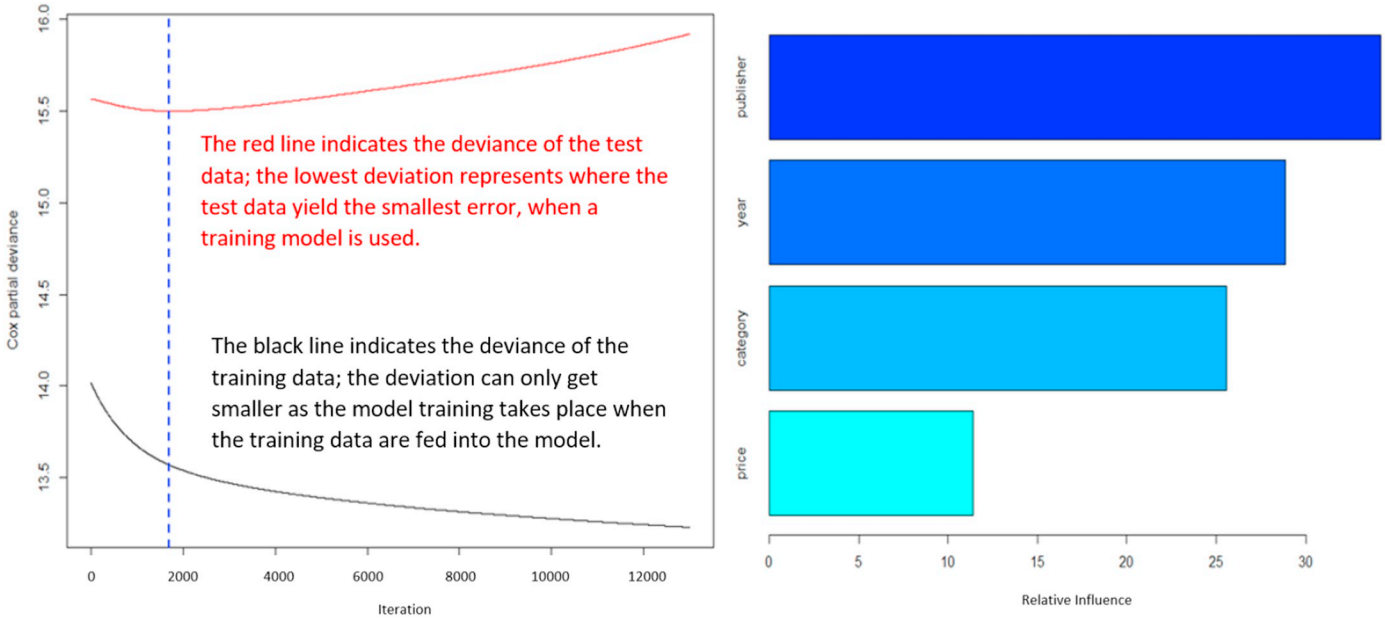

Fig. 5. Preliminary results of the survival analysis.

Table 3
Model comparison results.

Model ID	Predictors	Best iteration	AUC
1	Publisher	156	0.69
2	Publisher + Pub year	685	0.77
3	Publisher + Pub year + LC class	1233	0.79
4	Publisher + Pub year + LC class + price	1689	0.81

as patron preference. The algorithm determining the order of search results delivered to a library user will have a profound effect on what DDA titles that user is likely to select. Librarians should be cautious about trusting DDA mediated by a proprietary discovery layer to lead to purchases most relevant for patrons. Additional research into how discovery services affect DDA program outcomes is a worthwhile endeavor.

Vendor algorithms may be of particular concern when the same vendor is providing both DDA and discovery services. It is important for librarians to understand and guard against scenarios where DDA providers are prioritizing their own product within discovery layer search results. Sens and Fonseca (2013) warned of the danger in trusting vendors to populate discovery layer search results with their wares. Similarly, the ranking of unowned titles higher than those already owned is another risk worth guarding against, since patrons may end up purchasing redundant titles on the same topic. Each of these scenarios would likely result in DDA expenditures of greater benefit to vendors

than patrons. A discovery algorithm based on number of citations a book has received might be useful to ensure the discovery of influential titles on a topic. It is acknowledged that this method perpetuates a bias towards the frequently cited works, but we argue that patrons ought to be allowed to choose which sorting priority they prefer, or use both most recent and most cited.

In this study, triggered titles were on average 12% more expensive than titles that remained untriggered. Though additional analysis would be required for confirmation, this likely indicates that vendor pricing is related to a title's expected popularity. Limiting titles in the DDA pool to those published more than a year prior to load is one potential solution that could help to control program costs. However, this approach would delay the availability of the most current titles. For this reason, libraries might want to utilize this approach in a targeted fashion that places focus on selected disciplines where potential drawbacks are less impactful and/or title costs tend to be more significant.

It was unsurprising to see that Taylor & Francis was the publisher with the largest number of untriggered e-books in the pool (22% of total), since they are also the publisher with, by far, the largest number of e-books in the pool (25% of total). However, it is troubling to also discover that DDA program-level spending rate (40% of DDA expenditures) on Taylor & Francis titles is disproportionate in relation to their overall use (25% of total use). While it may be unreasonable to believe spending and use rates should align perfectly, a 15% difference between spending and use rates seems concerning. This pattern seems to contradict the notion that a bigger publisher can provide content at a lower price based on economies of scale. In contrast, Perseus Books, one of the top-three publishers within the DDA pool, in terms of titles purchased, shows a use rate that is more than 2% higher than their spending rate (3.24% vs. 1.03%).

This analysis shows that titles produced by university and society publishers experience two to four-times as much use-per-title than seen with larger corporate publishers. As an example, titles produced by the University of Tennessee Press feature an average use-per-title of 28, while titles from Taylor & Francis experience a use-per-title rate of only 8. A potential solution might be to add only selected titles from large corporate publishers to the DDA pool.

6.2. Limitations

The viability of survival analysis for DDA prediction is evident from this study, but a few limitations to the generalizability of this study's local findings should be noted. First, the dataset used here has been constructed with DDA program data for a single large research library. Moreover, titles for the program under study are identified for inclusion in the local DDA title pool based on a librarian-directed selection profile tailored to the needs of the local institution. These titles, and various related administrative mechanisms of the DDA program, are provided through a single, large library vendor.

In addition to the single data source issue, the e-book purchasing data being analyzed includes only single-trigger titles. There are various other DDA trigger models that exist and while data produced within these models can be simplified to mimic a single-trigger scenario, such an approach would lead to a loss of information caused by an artificial compression of variability within those data. For example, scenarios exist where a title might experience only two short-term loans (STL), which would not trigger a purchase under a three STL model. So to model that title as though it is identical to a title which does attain all three loans would skew findings in unpredictable ways.

Another limitation lies in the availability of predictors. Theoretically, the more information a model possesses, the higher accuracy the prediction can reach. If the DDA vendor had provided us with more variables, it might have been geared to produce more accurate predictions. Similar to what other library researchers have encountered, limited access to data places a ceiling effect on potential results. For example, Kohn (2018) used logistic regression to model e-book use with only three available variables, which resulted in low levels of model fit.

Although the local findings are not generalizable, this research has demonstrated the viability of survival analysis as a basis for predictive modeling of DDA purchasing. Machine learning involves development of real-time predictive models that change as the underlying data change. That is to say, a library defined by much different environmental factors than the library supplying this study's data, experiencing DDA title usage and purchasing patterns very different from those in this study, could deploy the exact same methods employed by this study to produce a model with predictive capacity equal to the model developed here. The underlying descriptive statistics may differ, causing unseen alterations to the underlying learning algorithm, but the manner in which those descriptive statistics are explored would be very similar.

While this research did find success in producing a highly accurate predictive model of DDA purchasing, there are potential improvements worth noting. For example, predictive modeling would be greatly improved if a wider array of data were included in such models. It would be helpful if vendors could more readily supply an expanded range of data, such as authors' publication records, the number of chapters, and the average length of each chapter. The inclusion of additional data from other external sources—a practice known as data merging or data joining—would also add value to these types of activities. Authors' publication sales records, product reviews on Amazon's website, and other data external to the vendor records could improve the capacity of predictive models. Finally, the incorporation of local user data, such as major area of study, ethnicity, age, and gender could provide important information to link with library content usage. These variables could help predict variation in demand for books related to those majors and identities. In addition, if books are listed in a syllabus, it is almost certain that they will be triggered.

7. Conclusion

A method of machine learning was used to predict DDA purchasing patterns. Survival analysis, within a random forests learning framework, was proven effective as a predictive modeling approach for use with DDA-generated data. In testing four separate models, it was found that, while the simplest of these models featured only fair predictive capacity, the most complex model featured what can be described as “good” predictive capacity (as evidenced by an AUC score of 0.81). Furthermore, the researchers surmise that with additional data added to the model, a higher AUC of over 0.90 might be achieved (i.e., what would be considered an “excellent” predictive model).

Purchasing patterns showing that more recently published materials are both purchased more readily, as well as cost more on average, raise concerns regarding both DDA vendor pricing models and the impact of proprietary discovery service algorithms on purchasing patterns. Moving forward, additional research that addresses these questions will provide great value to librarianship. As universities struggle with budgetary constraints, DDA will find continued importance as a tool that can bolster the overall return on investment for library dollars. At the same time, it will be of paramount importance for libraries to utilize innovative, forward-looking tools, like machine learning, to help better manage acquisition decisions. Research that develops and demonstrates effective, ongoing, and sustainable implementations of machine learning in support of better library outcomes, will help to sustain the continued impact of the academic library in the turbulent, fast-paced environment of technology-driven information consumption.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.lisr.2019.100968.

References

Allison, P. D. (2010). *Survival analysis using SAS: A practical guide.* Cary, NC: SAS Institute.

American Library Association (2015). Framework for information literacy for higher education. Retrieved from http://www.ala.org/acrl/standards/ilframework.

Anderson, K. J., Freeman, R. S., Hérubel, J. M., Mykytiu, L. J., Nixon, J. M., & Ward, S. M. (2002). Buy, don't borrow: Bibliographers' analysis of academic library collection development through interlibrary loan requests. *Collection Management, 27*(3/4), 1–11.

Association of College and Research Libraries (2019). ACRL trends and statistics survey. Retrieved from https://acrl.countingopinions.com/pireports/view_pireport.php?report_id=70402&fx.

Bradburn, M. J., Clark, T. G., Love, S. B., & Altman, D. G. (2003). Survival analysis part II: Multivariate data analysis–an introduction to concepts and methods. *British Journal of Cancer, 89*, 431–436. https://doi.org/10.1038/sj.bjc.6601119.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Breitbach, W., & Lambert, J. E. (2011). Patron-driven ebook acquisition. *Computers in Libraries, 31*(6), 16–20.

Cox, D. R. (1992). Regression models and life-tables. In S. Kotz, & N. Johnson (Eds.). *Breakthroughs in statistics (pp.* 527-541*).* New York, NY: Springer.

Cox, D. R. (2018). *Analysis of survival data.* New York, NY: Routledge.

Dewland, J. C., & See, A. (2014). Notes on operations: Patron driven acquisitions - determining the metrics for success. *Library Resources and Technical Service, 59*(1), 13–23.

Gilbertson, M., McKee, E. C., & Salisbury, L. (2014). Just in case or just in time? Outcomes of a 15-month patron-driven acquisition of e-books at the University of Arkansas libraries. *Library Collections, Acquisitions, and Technical Services, 38*, 10–20. https://doi.org/10.1080/14649055.2014.924072.

Howland, J. L., Schroeder, R., & Wright, T. (2014). Brigham young university's patron-driven acquisitions: Does it stand the test of time? In K. Bridges (Ed.). *Customer-based collection development: An overview (pp.* 115-126*).* London, England: Facet.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics, 2*, 841–860. https://doi.org/10.1214/08-AOAS169.

Jiang, Z., & Fitzgerald, S. R. (2019). Promoting institutional repositories via visualizations: A change-point study. *New Review of Academic Librarianship, 25*(1), 95–112. https://doi.org/10.1080/13614533.2018.1547775.

Jiang, Z., Walker, K., & Shi, D. (2019). Applying adaboost to improve diagnostic accuracy. *Methodology, 15*(2), 77–87. https://doi.org/10.1027/1614-2241/a000166.

King, S. L. (2003). Using ROC curves to compare neural networks and logistic regression for modeling individual noncatastrophic tree mortality. In J. W. Van Sambeek, J. O. Dawson, F. Ponder, E. F. Loewenstein, & J. S. Fralish (Eds.). *Proceedings of the 13th Central Hardwood forest Conference: Gen Tech Rep NC-234,* 349-358. St. Paul, MN: US Department of Agriculture, Forest Service, North Central Research Station.

Kohn, K. (2018). Using logistic regression to examine multiple factors related to e-book use. *Library Resources & Technical Services, 62*, 54–65.

Lumley, T., & Therneau, T. (2004). The survival package. *R News, 4*(1), 26–28.

Machin, D., Cheung, Y. B., & Parmar, M. (2006). *Survival analysis: A practical approach.* Hoboken, NJ: John Wiley & Sons.

Mogensen, U. B., Ishwaran, H., & Gerds, T. A. (2012). Evaluating random forests for survival analysis using prediction error curves. *Journal of Statistical Software, 50*(11), 1–23.

Oliver, J. C., Kollen, C., Hickson, B., & Rios, F. (2019). Data science support at the academic library. *Journal of Library Administration, 59*, 241–257. https://doi.org/10.1080/01930826.2019.1583015.

R Core Team (2018). *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* Vienna, Austria, Available online at http://www.Rproject.org/.

Ridgeway, G. (2015). gbm: Generalized boosted regression models. [Computer software version 2.1.4]. Retrieved from https://CRAN.R-project.org/package= *gbm*.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). Proc: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*(1), 77.

Schroeder, R. (2012). When patrons call the shots: Patron-driven acquisition at Brigham Young University. *Collection Building, 31*(1), 11–14.

Sens, J. M., & Fonseca, A. J. (2013). A skeptic's view of patron-driven acquisitions: Is it time to ask the tough questions? *Technical Services Quarterly, 30*, 359–371.

The Networking and Information Technology Research and Development Program. *The federal big data research and development strategic plan.* Retrieved from https://www.nitrd.gov/PUBS/bigdatardstrategicplan.pdf.

Tyler, D. C., Falci, C., Melvin, J. C., Epp, M. L., & Kreps, A. M. (2013). Patron-driven acquisition and circulation at an academic library: Interaction effects and circulation performance of print books acquired via librarians' orders, approval plans, and patrons' interlibrary loan requests. *Collection Management, 38*(1), 3–32.

Tyler, D. C., Hitt, B. D., Nterful, F. A., & Mettling, M. R. (2019). The scholarly impact of books acquired via approval plan selection, librarian orders, and patron-driven acquisitions as measured by citation counts. *College & Research Libraries, 80*, 525.

Tyler, D. C., Melvin, J. C., Epp, M., & Kreps, A. M. (2014). Patron-driven acquisition and monopolistic use: Are patrons at academic libraries using library funds to effectively build private collections? *Library Philosophy and Practice, 1149*. Retrieved from https://digitalcommons.unl.edu/libphilprac/1149/?utm_source=digitalcommons.unl.edu%2Flibphilprac%2F1149&utm_medium=PDF&utm_campaign=PDFCoverPages.

Tyler, D. C., Xu, Y., Melvin, J. C., Epp, M., & Kreps, A. M. (2011). Just how right are customers? An analysis of the relative performance of patron-initiated interlibrary loan monograph purchases. In J. M. Nixon, R. S. Freeman, & S. M. Ward (Eds.). *Patron-driven acquisitions: Current successes and future directions*. Hoboken: Taylor and Francis.

Walker, K. W., & Arthur, M. A. (2018). Judging the need for and value of DDA in an academic research library setting. *The Journal of Academic Librarianship, 44*, 650–662. https://doi.org/10.1016/j.acalib.2018.07.011.

Walker, K. W., & Jiang, Z. (2019). Application of adaptive boosting (AdaBoost) in demand-driven acquisition (DDA) prediction: A machine-learning approach. *The Journal of Academic Librarianship, 45*, 203–212. https://doi.org/10.1016/j.acalib.2019.02.013.

Waller, J. H. (2013). Undergrads as selectors: Assessing patron-driven acquisition at a liberal arts college. *Journal of Interlibrary Loan, Document Delivery & Electronic Reserve, 23*, 127–148. https://doi.org/10.1080/1072303X.2013.851052.

Walters, W. H. (2012). Patron-driven acquisition and the educational mission of the academic library. *Library Resources and Technical Service, 56*, 199–213. https://doi.org/10.5860/lrts.56n3.199.

Way, D., & Garrison, J. A. (2011). Financial implications of demand-driven acquisitions: A case study of the value of short term loans. In D. Swords (Ed.). *Patron-driven acquisitions: History and best practices (pp.* 137-156*).* Berlin, Germany: Walter De Gruyter.

Wolff-Eisenberg, C., Rod, A. B., & Schonfeld, R. C. (2016). *Ithaka S+R US faculty survey 2015.* https://doi.org/10.18665/sr.277685.

Wu, S., & Nagahashi, H. (2014). Parameterized adaboost: Introducing a parameter to speed up the training of real adaboost. *IEEE Signal Processing Letters, 21*, 687–691.

Youngstrom, E. A. (2013). A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology, 39*, 204–221.

Zhang, Y., & Downey, K. (2017). Ebook ROI: A longitudinal study of patron-driven acquisition models. *Computers in Libraries, 37*(5), 4–8.

Zhehan Jiang is the Data Services Librarian and an Assistant Professor at the University of Alabama Libraries. He holds a PhD from the University Kansas in educational statistics and research methods, a Master's degree from the University of California, Los Angeles School of Education and Information Studies in social research methodology, and a Bachelor's degree in finance from the University of San Francisco. He has published in Behavior Research Methods, Structural Equation Modeling: A Multidisciplinary Journal, Applied Psychological Measurement, Educational and Psychological Measurement, Multivariate Behavioral Research, Frontiers in Psychology, Psychometrika, Methodology, and Methodological Innovations. His research focuses on quantitative research methods.

Sarah Rose Fitzgerald is the Assessment Librarian and an Assistant Professor at the University of Alabama Libraries. She holds a PhD in higher, adult, and lifelong education from Michigan State University, a Master's degree in library and information science from Wayne State University, and bachelor's degree in English from the University of Michigan. She has published in the Journal of Academic Librarianship and the Journal of Higher Education Policy and Management. Her research interests include faculty work, information seeking in higher education, and scholarly communication.

Kevin W. Walker is the head of Assessment & Government Information and an Associate Professor at the University of Alabama Libraries. He holds a PhD in political science from Auburn University. He has published in the Journal of Academic Librarianship, Methodology, and Methodological Innovations. His research interests include library assessment, information seeking, and analytics application.