DDA Management With Predictive Modeling

John Vickery
North Carolina State University at Raleigh, john_vickery@ncsu.edu

Follow this and additional works at: <https://docs.lib.purdue.edu/charleston>

An indexed, print copy of the Proceedings is also available for purchase at:

<http://www.thepress.purdue.edu/series/charleston>.

You may also be interested in the new series, Charleston Insights in Library, Archival, and Information Sciences. Find out more at: <http://www.thepress.purdue.edu/series/charleston-insights-library-archival-and-information-sciences>.

John Vickery, "DDA Management With Predictive Modeling" (2016). *Proceedings of the Charleston Library Conference*.

<http://dx.doi.org/10.5703/1288284316457>

DDA Management With Predictive Modeling

John Vickery, Analytics Coordinator and Collections & Research Librarian for Social Sciences, North Carolina State University Libraries

Abstract

Demand-driven acquisitions (DDA) programs have become an integral part of academic libraries' collecting strategies. While DDA programs provide an effective way to build a just-in-time collection, it can be difficult to anticipate how many titles will be triggered for purchase and what the financial impact will be. This presentation will describe a project to build a predictive model to flag DDA titles that are likely to be triggered for purchase within the first year of being added to the catalog. By implementing a predictive model, collections and acquisitions departments can better plan the yearly DDA budget. In addition, titles with a high probability of being triggered for purchase can be purchased if they become ineligible for DDA. We will discuss how we combined text analytics and structured data as inputs to the model using a combination of Statistical Analysis System (SAS) and Python.

Introduction

This presentation describes a project to project to build a predictive model to flag DDA titles that are likely to be triggered for purchase within the first year of being added to the catalog. As demand-driven acquisitions (DDA) programs have become an integral part of academic libraries' collecting strategies it is important to investigate ways to anticipate how many titles will be triggered for purchase. In addition, as publishers consider reducing their participation in DDA programs, a predictive model can be used to simulate a demand-driven acquisition strategy by identifying titles that have a high probability of being used.

About the Data

The data used to train the model started with approximately 49,000 bibliographic records from the North Carolina State University (NCSU) libraries' Sirsi ILS. These records represented DDA titles added to the catalog between December 2011 and June 2015. Each record was identified as either currently untriggered or as having already been triggered for purchase. Purchase dates for each title were compared to the date the title was added to the catalog in order to calculate a binary variable indicating if the title was triggered for purchase within 12 months of being added to the catalog. This was represented as the variable *WithinYear* and was coded as 0 for nonpurchased and 1 for purchased. It was this variable that was used as the modeling target variable. Of the titles in the initial training data set, 11% had been purchased within 12 months of

being added to the catalog. The majority (95%) of the titles were published between 2011 and 2016. The two most common publishers represented in the data set were Routledge and Palgrave Macmillan.

In order to accommodate the relatively rare frequency of a title being purchased within 12 months, the initial data set was over sampled to a *WithinYear* = 1 frequency of 50%. The oversampling process resulted in a final training data set of 11,138 records. 100% of the *WithinYear* = 1 records ($n = 5569$) was combined with a random sample of 5,569 of the *WithinYear* = 0 records.

Data sources used to compile the training data set consisted of local bibliographic data from the libraries' Sirsi ILS and MARC records as well as data from external sources. The two sources for external data were Online Computer Library Center's (OCLC) Classify API and Proquest-Syndetic Solutions book summary data. The Classify web service available at http://classify.oclc.org/classify2/api_docs/index.htm was used to augment records where the library of congress classification was missing. The Syndetics book summary data was accessed via API. Note that a subscription to the Syndetics service is required to access the API. Book summary data and how it was used is discussed in more detail in following sections.

Software Used for the Project

Three software products were used to build the training data set and develop the predictive model. Base SAS was used for the majority of the data extraction and cleaning of Sirsi ILS data. Python

programs were used to retrieve data from both the OCLC Classify and Syndetic API's. Finally, SAS Enterprise Miner (with text miner) was used for the model training and validation. SAS Enterprise Miner was also used to score a hold-out data sample as a final test of the model.

Model Inputs

Model inputs were purposefully restricted to data that would be apparent to a user while viewing a title in NCSU Libraries' catalog. Figure 1 highlights data used as potential model inputs. Specifically, the following data items were selected as potential inputs: Publisher, publication year, Library of Congress subclass, and Syndetic Solutions book summary. The Syndetic book summary data can be considered unstructured, text data while the other inputs are structured. In order to be used as model input, the unstructured book summary data must be converted to a structured format. The following section describes the process of deriving structured topics from the book summary.

Topic Modeling Process

In order to use the Syndetic book summary data as input to the predictive model, the unstructured text must be converted into structured numeric data. The method used for this project is referred to as *topic modeling*. Topic modeling is a process of automatically associating a document with collections or terms characterizing a theme or idea. In this case, each summary represents a "document." The topic modeling algorithm assigns a score to each document and term. If the association is strong enough, the document is assigned to a particular topic. Each document can be assigned to one or more topics or even no topics. It is these numeric scores representing topics or themes that are then used as inputs to the predictive model. As mentioned in a previous section, the text miner module of SAS Enterprise Miner software was used to model topics from the Syndetic book summary data. Chakraborty, et al. (2013) provide an excellent overview and case study of the topic modeling process. Blei (2012) also provides an overview of topic modeling.

Model Inputs:

- Publisher
- Publication year
- LC class
- Summary

Figure 1. Potential model inputs.

Extracting topics from unstructured text is a multistep process. Figure 2 illustrates the main steps involved. Summary data for each DDA title was gathered using a Python program to fetch the book summary from the Syndetic Web service. Note that a subscription to their service would be required. In the next step, the software (SAS Enterprise Miner with Text Miner) parses the collection of book summaries into a term-document frequency matrix where each element of the matrix equals the number of times that term appears in a document. During this step, low-value words such as *the*, *and*, *a*, *etc.* are dropped from the analysis. The following step involves refining the term list by interactively dropping or combining terms. For example, in the case of this project, the term *book* could be dropped

as each summary is referring to a book, and the term, therefore, does not add value to the analysis. In the next step, the SAS software algorithm reduces the term-document matrix dimensions using the singular value decomposition method and extracts topics from the document collection. While topics can have many terms associated with them, the top five terms for each topic are displayed as output. Table 1 lists examples of topics as represented by their top five terms. The final step is to interpret the generated topics and determine if they make sense. If the topics do not make sense, it may be necessary to refine the list of terms that are combined or dropped. The numeric score for each topic can then be used as inputs to a predictive model.

The Topic Modeling Process

Figure 2. Topic modeling process.

Table 1. Example topics represented by their top five terms (+ symbols indicate that the term has been stemmed or synonymized).

Top five terms
engineering,+engineer,+engineer,+design,+application
clinical,+treatment,medical,+therapy,+professional
mathematics,+application,mathematical,+problem,+solution
+disease,molecular,+biology,+cell,+protein
software,+design,+basic,+guide,+learn
+woman,+man,male,+identity,sexual
+introduction,accessible,concise,comprehensive,+philosophy
+material,+engineer,+polymer,+technology,+material science
food,+food,nutrition,+product,food
+law,legal,+court,+right,legal
data,big data,+process,+application,+technique

Figure 3. A portion of the final decision tree.

Figure 4. Caption: An 8% misclassification rate.

Results

A step-wise logistic regression model (Cramer & Howitt, 2004) and a decision tree model (Salkind, 2010) were trained on the data. The models' performance was then compared based on misclassification rate. Based on misclassification

rate, a decision tree model was selected as the best performing. The final model relied solely on inputs derived from the topic modeling process described previously. These topics are shown in Table 1. Figure 3 shows a graphical representation of a portion of the final decision tree model.

References

- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84.
<https://doi.org/10.1145/2133806.2133826>
- Chakraborty, G., Pagolu, M., & Garla, S. (2013). *Text mining and analysis: Practical methods, examples, and case studies using SAS*. Cary, NC: SAS Institute.
- Cramer, D., & Howitt, D. (2004). *The SAGE dictionary of statistics*. Thousand Oaks, CA: SAGE Publications Ltd.
<https://doi.org/10.4135/9780857020123>
- Salkind, N. J. (2010). *Encyclopedia of research design*. Thousand Oaks, CA: SAGE Publications Ltd.
<https://doi.org/10.4135/9781412961288>