Knowledge-Based Systems 115 (2017) 133–151

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Applying computational intelligence methods for predicting the sales of newly published books in a real editorial business management environment

Pedro A. Castillo[a,*], Antonio M. Mora[a], Hossam Faris[b], J.J. Merelo[a], Pablo García-Sánchez[a], Antonio J. Fernández-Ares[a], Paloma De las Cuevas[a], María I. García-Arenas[a]

[a] *Department of Computer Architecture and Computer Technology, ETSIIT and CITIC, University of Granada, Granada, Spain*
[b] *Business Information Technology Department, King Abdullah II School for Information Technology, The University of Jordan, Amman, Jordan*

ARTICLE INFO

Article history:
Received 3 June 2016
Revised 12 September 2016
Accepted 16 October 2016
Available online 17 October 2016

Keywords:
Book sales forecasting
Decision-aid models
Feature selection
Regression

ABSTRACT

When a new book is launched the publisher faces the problem of how many books should be printed for delivery to bookstores; printing too many is the main issue, since it implies a loss of investment due to inventory excess, but printing too few will also have a negative economic impact. In this paper, we are tackling the problem of predicting total sales in order to print the right amount of books and doing so even before the book has reached the stores. A real dataset including the complete sales data for books published in Spain across several years has been used. We have conducted an analysis in three stages: an initial exploratory analysis, by means of data visualisation techniques; a feature selection process, using different techniques to find out what are the variables that have more impact on sales; and a regression or prediction stage, in which a set of machine learning methods has been applied to create forecasting models for book sales. The obtained models are able to predict sales from pre-publication data with remarkable accuracy, and can be visualised as simple decision trees. Thus, these can be used as decision-aid tools for publishers, which can provide a reliable guidance on the decision process of publishing a book. This is also shown in the paper by addressing four example cases of representative publishers, regarding their number of sales and the number of different books they sell.

1. Introduction

Publishing a book, in the same way as releasing any other cultural product with a physical substrate, implies several types of risks and costs due to the complexity, and derived expenses, of its processes of production, distribution, and storage. Most of these costs are associated with the number of books actually printed, the so called print run. But the problem is further complicated for completely new products, in the sense that they have been written by an unknown author or deal with a new topic with no track record. In order to make the print run as close as possible to actual sales, it is essential to have an estimation of those sales; that is the challenge we are taking in this paper.

In general, forecasting the sales of cultural products can only be based on the knowledge obtained from sales of other, similar, ones, but the definition of “similar” itself is fuzzy and, sometimes, subjective. Even if the author or the topic is known, making accurate estimations of future sales of the new product is quite difficult, and extremely important, so that production runs follow predicted sales. Indeed, several studies have demonstrated that improving sales predictions, that is reducing the error in estimated sales, results in an enhancement in the whole production process [1,2], so not only the costs associated with the book itself are affected.

Within the general area of cultural products, the problem of predicting sales is specially acute for book publishers who, among other providers of content in physical form, release new *products* more frequently than other industries, so that they face the problem of predicting sales quite often. The issue, in this case, is to print an adequate number of copies but not too many, as the unsold volumes will lead to losses and sunk inventory cost, whereas if the number of printed copies was not enough, a new print run can always be made, but this will result in temporary losses due

* Corresponding author. Fax: +34958248993.
E-mail addresses: pacv@ugr.es (P.A. Castillo), amorag@geneura.ugr.es (A.M. Mora), hossam.faris@ju.edu.jo (H. Faris), jmerelo@geneura.ugr.es (J.J. Merelo), pablogarcia@ugr.es (P. García-Sánchez), antares.es@gmail.com (A.J. Fernández-Ares), palomacd@ugr.es (P. De las Cuevas), mgarenas@ugr.es (M.I. García-Arenas).

http://dx.doi.org/10.1016/j.knosys.2016.10.019
0950-7051/© 2016 Elsevier B.V. All rights reserved.

to a lack of supply or in bad marketing for the customers, for instance. However, in practice predicting lower sales than the actual ones is not such a big problem because the new print run is ordered before the inventory is exhausted; nevertheless, this new print run presents the same problem as the initial one of printing only as many as forecasted, although its cost is not as high as the initial one.

Despite the fact that errors in one or the other direction do not have the same impact, it is extremely important to develop an accurate predictive method which could forecast the future sales a new book will achieve, in order to optimise production schedules, improve the publishing company profits and minimise losses [3]. This is the main objective of this paper.

In Spain, the process of releasing a book goes like this: the publisher decides, based on past experience, how many books to print. These books are distributed to points of sale, but also given out to reviewers and literary magazines; the number of these will depend on how many were printed. Sometimes books go to the *novelty* table of the bookstore, that is, they are prominently displayed, shown in the window, or highlighted using props or other kind of advertisement. From these displays they then move to shelves or, in some cases and eventually, to storage. Books stay for sale in the store as long as the publisher or the store wants or until the bookseller decides to return them to the publisher. Books returned to the publisher are pulped, sold back or given to the author, or finally distributed in used-books channels such as book stalls or second-hand bookstores. If the print run has been fully distributed and there are more requests from bookstores, on the other hand, a new print run is made and distributed to these shops. Even if the process has changed slightly in the last few years since Amazon started to sell in Spain, in the sense that many books are sold through this new channel and do not undergo the cycle novelty table-shelf-storage-back to publisher, which actually happened after the data used for this paper was collected, it roughly stays the same. Other countries will have a similar model, although the scale will be different.

Our intention in this paper is to find out the main factors influencing sales in order to create a tool that the publisher can use to decide how many books should be printed, as well as how to leverage these printed copies to maximize sales using the decision variables under his control. That is why, using data obtained from a company that sells software for publishers, we analyse them and compute predictive models that can be mainly used as decision-aid tools for book publishers. With these models, publishers will be able to combine their expert knowledge about the market with the created forecasting models in order to get a reliable estimation of book sales, and, based on it, act consequently in order to maximize the books sold for a particular print run.

This will improve the current decision flow that the publishers follow, which consist in analysing (in a subjective way) the quality of the book and its features (author, genre, etc), and take a ‘fuzzy’ decision roughly between printing 300 copies (standard book) or 5000 (best seller).

This process might be tedious, especially when the amount of data is quite large. Moreover, different experts can reach different predictions from the same dataset [4].

With these issues in mind, in this paper, we firstly present an analysis on a real dataset, provided by the Spanish publishing company Trevenque Editorial S.L.[1]. This study has been conducted by means of data mining and visualisation techniques. Then, a feature selection process is performed using three different methods, in order to find out what are the relevant variables describing a book in the prediction, or estimation, procedure. This procedure is performed by applying a set of regression algorithms to the different datasets (those with a different number of variables), yielding a set of models which could be used as the desired tool.

Given this, the main contribution of this paper to the state of the art is the development of the aforementioned methodology to process book sales data, in which: firstly, the most relevant variables are identified, and then, these variables are considered to ‘refine’ the data in order to conduct accurate predictions on future sales. To our knowledge this is the first work in which regression methods have been applied to estimate sales for new launched books. Moreover, as stated before, we have used in the study a real dataset of book sales (related to the Spanish market), which is another point to remark.

The value of this methodology is contrasted considering four different publishing companies being representative as top sellers, mid-range sellers, the most varied, and one which sells a medium amount of different books. From this, we draw insight on what are the important factors when selling a book and some rule of thumbs extracted from the prediction methods.

The rest of this paper is structured as follows: next, Section 2 presents a comprehensive review of the approaches found in the bibliography related to sales prediction in similar scopes. Then, Section 3 introduces the problem of the estimation of print run for new books, along with a description of the considered dataset. Section 4 details the methodology considered in the study. These are reported in Section 5, which is also devoted to analyse the obtained results in feature selection and book sales forecasting using different datasets, with different amounts of features, for all the publishing companies. Obtained results for the four considered special cases are commented in Section 6. Finally, conclusions and future lines of work are presented in Section 7.

2. State of the art on sales forecasting

As far as we know, no previous attempt can be found in the literature in which prediction methods have been applied to estimate sales for new launched books in a whole country market. However, similar techniques to those proposed in this paper, e.g. regression models [5], neural networks [6] or fuzzy systems [7], have been applied to solve sales prediction problems in other industrial sectors.

Specifically, time-series prediction methods [5,8–12] is perhaps the most used technique to tackle sales forecasting problems, although the efficiency of these techniques strongly depends on the field of application and the correctness of the problem data. However, since they require a large amount of data for predicting sales, these methods are not the most suitable for this task [13]. Moreover, this kind of methods cannot be applied before the book is published, as is the objective of this paper. In this case, just a few variables are known in advance (pre-sales data) [13–15], and some of them are categorical, which makes them harder to process, even more if the number of categories is high. As stated in those works, given these issues, finding adequate forecasting techniques and selecting the best method to use is also a problem. An additional issue makes this a complex problem, since the predictions are influenced by external variables that must be taken into account, such as seasonality, promotions, or fashions that expert managers might subjectively apply [16,17]. For this reason, usually the methods used to predict book sales have been generally based on experts’ experience, who analyse data about sales and, taking into account their knowledge of the industry, their experience, and their perception about trends, could make decisions on the companies production, i.e. how many books should be printed when a new book is published.

[1] http://editorial.trevenque.es Trevenque offers management systems and services for bookshops, publishers, and distributors.