Forever Gone

Business Analytics using Data Mining

Group 5

Cynthia Chen
Edward Song
Huynh Nhat To
Olive Chang

1/12/2016

Part I Introduction

Problem Description

In the library, it has been a hard time for librarians to decide whether to purchase the replacement for a book that has been reported missing since some books may be found not long after the replacements have been bought. On the other hand, it would be irritating for us as students if books we're looking for have been report missing for a long time without replacements.

Business Goal

As a group of students who wish to make school better, our goal is to help the NTHU library identify which missing book is likely to be gone forever and which is very likely to be found, so they can decide whether to initiate finding a replacement. If we are successful, we would not only eliminate the waste on redundant spending, we would also contribute to enhancing the efficiency of NTHU library administration.

Data mining goal

We will attempt to identify which (kinds of) books once missing are most likely to remain missing. We will attempt to produce: a binary classification of books - forever missing or will be found; and a ranking by likelihood to remain missing. This is both a supervised and predictive task, as the records made available to us contain data showing whether the current status of the book (lost or found). Additionally, the solution we hope to produce can be applied both retrospectively and prospectively - the library can use our solution to classify books in its larger record depending on our current status now and in the future. The main outcome variables is "current item status"..

Part Ⅱ Data

Data collection from library

In November, we requested for data from the library and signed official agreements with the library administration pertaining to data handling. We negotiated with the librarian about the data we would receive - variables and records. But after we got the data, we realized there were several problems. Even though we had lots of columns, most of them were not related to our goal. Additionally, missing books are not as many as the librarian previously told us and most records in the dataset we received belonged to the same books.

Column	Description
id	Readers system number
reader idendity	Type of reader
department	Department of reader
item type	Type of item
transaction date	Date of transaction
transaction type	Type of transaction
bar code	Barcode No
current item status	The current processing status of the item
location	Library Branch Code
updating date of current item status	The item processing status update
date of last returning	Last return date
call number	Call Number
rule of call number type	Language of Item
history status of item	Problem with item in the past
updating date of histroy item status	History of the item processing status update
item title	Title
item author	Author
item ISBN_ISSN	ISBN
call number type	item type based on Chinese System or Library of Congress

Snapshot of raw data -
Both images are for the same records
e.g. row 1, image 1≡row 1, image 2

Each record is a library transaction: either borrowing a book or renewing a borrowed book.

	reader i	id	item typ	transaction da	transaction typ	bar code	current	location	updating d	date of last re
b62f7998ba9c50e3f	11	32	BK	2010/02/26	50	C363798		LB	2012/06/06	2015/11/09
b62f7998ba9c50e3f	11	32	BK	2009/05/08	62	C363798		LB	2012/06/06	2015/11/09
b62f7998ba9c50e3f	11	32	BK	2009/04/24	50	C363798		LB	2012/06/06	2015/11/09
b62f7998ba9c50e3f	11	32	BK	2009/06/04	62	C363798		LB	2012/06/06	2015/11/09
49dc0416eac4cd62	11	32	BK	2011/06/04	50	C501054		LB	2012/09/28	2015/10/04
97a5fe2620bc9b56d	11	3A	BK	2009/02/21	50	C429991		LB	2011/11/07	2015/09/23
97a5fe2620bc9b56d	11	3A	BK	2009/01/17	50	C429991		LB	2011/11/07	2015/09/23
359bfd6879cdfbb7fd	11	3A	BK	2013/04/11	50	C532411		LB	2012/04/10	2015/11/22
359bfd6879cdfbb7fd	11	3A	BK	2014/10/13	50	C532411		LB	2012/04/10	2015/11/22

Image 1

call number	rule of c	history s	updating da	Z13_TITLE	Z13_AUTHOR	Z13_ISBN_ISSN	call num
454 8473	8	B3	2012/03/27	材料熱力學	Gaskell,David R.	957-30084-5-9 (精裝	454
454 8473	8	B3	2012/03/27	材料熱力學	Gaskell,David R.	957-30084-5-9 (精裝	454
454 8473	8	B3	2012/03/27	材料熱力學	Gaskell,David R.	957-30084-5-9 (精裝	454
454 8473	8	B3	2012/03/27	材料熱力學	Gaskell,David R.	957-30084-5-9 (精裝	454
805.1895 873	8	B3	2012/09/26	新TOEIC 聽力題庫	Andrea, Allison	9789866186042 (平	805.1895
448.873 8437/2 v.1	8	B3	2011/10/05	微電子電路	Sedra,Adel S.	957-99921-8-5 (上冊	448.873
448.873 8437/2 v.1	8	B3	2011/10/05	微電子電路	Sedra,Adel S.	957-99921-8-5 (上冊	448.873
448.62 8954 2011 v.1	8	B3	2012/03/15	微電子電路 /	Sedra, Adel S	9789868085336 (上	448.62
448.62 8954 2011 v.1	8	B3	2012/03/15	微電子電路 /	Sedra, Adel S	9789868085336 (上	448.62

Image 2

Data preparation

Step 1. Compressing transactions into items.

We received 33,070 records, which each record is a library transaction. We compressed these transactions into individual items using the barcode as our identifier. But on the other hand, we created a column to count the number of times each bar code was present in a transaction. This was done to act as a proxy for book popularity. In the end, we had only 787 records. Of these records, only 38 books were classed as missing.

Step 2. Data Transformation

The variables in green in the adjacent figure are variables we retained as they are related to each item. We transformed the outcome "current item status" to binary with "missing" as success, and "found" as failure; the "location"; "item type"; and "rule of call number type"(a proxy for language) to dummies.

Part III Data mining methods

XLminer

We applied different methods to find a model which has the lowest error rate.

1. Logistic regression

First, we partition the data to training set and validation set with oversampling. Then we set "% Success in Training data" as 40 and "% Validation data taken away as test data" as 0.

Random Seed	12345
# training rows	47
# validation rows	393
# test rows	0
Selected output variable	B3_1
% Success in Training data	40
% Validation data taken away as test data	0
% Success in original data set	4.828462516

Image 3

The outcome seems bad. Our goal is to predict the transaction that the items may come back. As we described previously, the success means "missing", %Error of classification 1 is 57.89 which is very high.

Error Report

Class	# Cases	# Errors	% Error
1	19	11	57.89473684
0	374	108	28.87700535
Overall	393	119	30.27989822

Image 4

2. Classification trees

Before we use classification trees, we partition data with oversampling, and set "% Success in Training data" as 20 and "% Validation data taken away as test data" as 50.

# training rows	95
# validation rows	186
# test rows	207
Selected output variable	B3_1
% Success in Training data	20
% Validation data taken away as test data	50
% Success in original data set	4.828462516

Image 5

We use random tree to find the potential classifier of the data and keep other setting the same as logistic regression.

The outcome is extremely useless since the error of our success class is 100% which means none of the forever missing book would be predicted using this model.

Error Report			
Class	# Cases	# Errors	% Error
1	9	9	100
0	177	1	0.564972
Overall	186	10	5.376344

Image 6

3. Naive Bayes

We keep the same default and other settings to try Naive Bayes. The outcome is also worse than logistic regression.

The Model created by XLMiner is not useful since its predictability whether is due to software problem or data problem so we try to us can provide better result.

Error Report			
Class	# Cases	# Errors	% Error
1	9	7	77.77778
0	177	20	11.29944
Overall	186	27	14.51613

Image 7

RapidMiner Process

We split the data into training and test set, each containing 50% of our success class. Next we oversampled for the success class in our training set creating a 50-50 split. We next applied RapidMiner's bootstrap operator - sampling with replacement - to increase the size of the training set. Then we passed this training set into an ensemble of four data mining methods which decide the class through voting. The methods are Naive Bayes, Logistic Regression, Random Forest (30 trees), Linear Regression. Finally, we applied this ensemble to our test data set to generate predictions. Image 9 contains the confusion matrix for the test set.

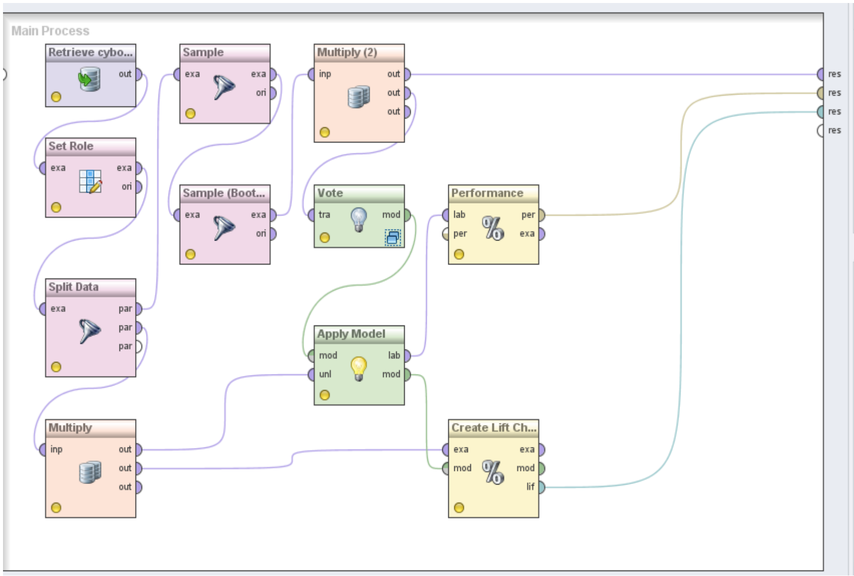


Image 8

classification_error: 46.31%

	true 0	true 1	class precision
pred. 0	198	5	97.54%
pred. 1	177	13	6.84%
class recall	52.80%	72.22%	

Image 9

Part IV Conclusion & Recommendations

Conclusion

1. Given the relatively few number of missing books, our model can not be guaranteed to reliably inform the library staff.
2. Compared to all books in library, the amount of missing books seems relatively very little. The information we want to have from missing books is insufficient. Which means, they don’t really either put effort on this problem or try solving it.

Recommendations

1. According to the record we got, we couldn’t build a useful model, the result appears randomly, so we suggest NTHU Library create more columns such as “the times looking for missing books” , “the location where they find missing books” ...etc. It would be more reasonable and useful to build a model if there are more detail records for missing books.

2. The definition of our data are not quite clear and some are complicated. For example, the librarian couldn’t know that a book is missing or not until readers inform of this. Also, we don’t know how many times librarians trying to find missing books before they define the book is missing. This makes the process of sorting data difficult. As a result, we suggest NTHU Library should make more effort on defining missing books, and collecting details as well.

Appendix

The Outcome of Logistic Regression

Logistic Regression

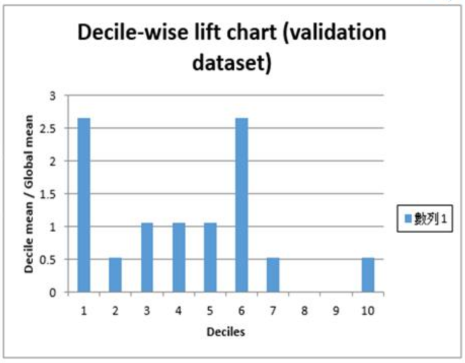
Method

Error Report

Error Report			
Class	# Cases	# Errors	% Error
1	19	11	57.89473684
0	374	108	28.87700535
Overall	393	119	30.27989822

Confusion Matrix

Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	8	1
0	108	26



The Outcome of Naïve Bayes

Naive Bayes

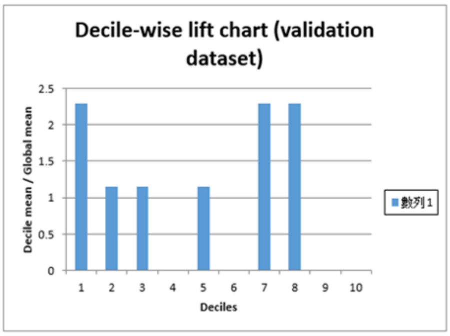
Method

Error Report

Error Report			
Class	# Cases	# Errors	% Error
1	9	7	77.77778
0	177	20	11.29944
Overall	186	27	14.51613

Confusion Matrix

Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	2	7
0	20	157



The Outcome of Classification Tree

Classification Tree

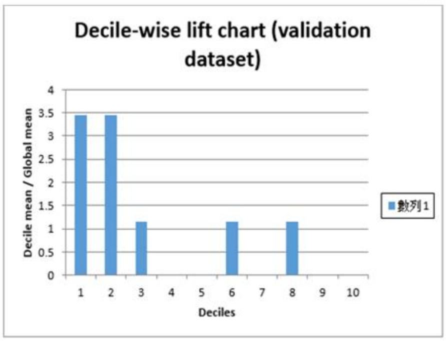
Method

Error Report

Error Report			
Class	# Cases	# Errors	% Error
1	9	9	100
0	177	1	0.564972
Overall	186	10	5.376344

Confusion Matrix

Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	0	9
0	1	176



The Outcome of Rapid Miner