CASE STUDY #2

BOOK BINDERS

DA 6813 DATA-ANALYTICS APPLICATIONS| 3-OCT-2021

KRISCHELLE JOYNER | KATHY KEEVAN |CHARLES REYES

Executive Summary

The linear regression, logistic regression and support vector machine models produced similar accuracy results. Our research shows that logistic regression was the best model for the Bookbinders dataset due to its repeated k-fold cross-validation that allowed a more reliable estimate of model performance. Frequency was the most significant variable that influenced subscribers to purchase the book. In our cost analysis, targeted mailing would be effective at reaching our audience by saving on marketing expenses and capitalizing on the opportunities to earn revenue.

Background

The book retailing industry had a rise in growth in the 1970's with the development of shopping malls. This created an influx of customers buying physical books in brick-and-mortar shops. By the 1990's, large superstores like Amazon started selling books en masse, putting pressure on mail order-firms such as the Bookbinders Book Club.

Historically mail-order firms, also called book clubs, provided readers with a negative option program where they would receive books on a monthly basis unless they opted out for a particular month. This program was a contractual agreement between the subscriber and book club. Subscribers could choose between several genres such as cooking, art, and children's books. By default, the club selection of the books for each month would be sent to the subscriber, but the subscriber could choose the books or how many to receive on a monthly basis.

In response to the competitive pressures of large superstores, book clubs are starting to look at alternative business models that will make them more responsive to their customer's preferences. Database marketing has become a competitive advantage for book clubs to understand their customer base better and to continue to get better through the use of data.

Since 1986, Bookbinders Book Club (BBBC) has been distributing books through direct marketing. BBBC built and maintained a comprehensive database about every one of its members at the start of its database marketing strategy. Currently, the company has a database of 500,000 subscribers and sends out a mailing every month.

Purpose of Study

A predictive analysis is being conducted by the BBBC to determine if the direct-mail model can be improved. A sample of 20,000 customers were selected from New York, Ohio, and Pennsylvania to receive a marketing brochure for *The Art of Florence* book to get an idea of the response rate. The results showed that 9.03% subscribers purchased the book.

In order to analyze these results further, we will use a subset of the database with these responses. In total, it includes data from 400 customers who bought the book and 1200 customers who did not, thus over-representing the response group.

The analysis will primarily focus on:

Which factors heavily influenced customers to buy the book?

Using Choice as our dependent variable (1 – purchased the book, 0 – did not purchase the book), we will explore several independent variables such as gender, purchase amount, frequency, first and last purchase, and several genres to see which variables worked best.

Does our model work well in a cost analysis scenario?

Using 50,000 customers from the Midwest region and some cost insights for the books, we will apply our best model through a cost analysis model and see if we can save money for our budgeting expenses.

Hypothetical research questions:

How often are customers purchasing books?

If we look at the number of months from the last purchase, we might be able to get an idea of our true month-to-month subscribers for marketing.

Does the type of a book play a role on whether it is purchased?

The type of books that book club members purchase might be more varied than what they normally read.

Review of Related Literature

The Bookbinders Book Club dataset is fabricated data for a fictitious company, created for educational purposes. It was originally published in 1995 as "A Case Study in Database Marketing" (Levin, Nissan and Zahav, Jacob, Tel Aviv University). It was later used as the basis of a case study under the name of "The Charles Book Club Case" (Bhandari, Vinni and Patel, Nitin).[1] Because these publications were for educational purposes, locating academic research (as opposed to student assignments posted online) proved challenging using searches for "Bookbinders Book Club", "BBB Club", or "Charles Book Club". The following are two studies that do examine different model types on the Bookbinders/Charles Book Club data.

evtree: Evolutionary Learning of Globally Optimal Classification and Regression Trees in R
Grubinger, et al.

In this paper, the authors test an R package called "evtree" against CART rpart and conditional inference trees. These trees use different algorithms for whether they are recursive (for example, forward stepwise search) and for tasks like segmenting data, pruning, number of branches, and missing value handling.[2] To facilitate model complexity, evtree splits to just two levels. To increase predictive accuracy in such a simple model, evtree uses evolutionary algorithms that mirror the Darwinian concepts of inheritance, mutation and natural selection. “rpart2” and “ctree2” are CART and conditional inference trees forced to two levels to be comparable to the two-level evtree model. Based on the misclassification rate and the evaluation function (the misclassification rate with a penalty for complexity), evtree performs better than rpart2 and ctree2, as well as ctree with unrestricted levels.[3]

	evtree	rpart	ctree	rpart2	ctree2
misclassification	0.243	0.238	0.248	0.262	0.255
evaluation function	660.680	655.851	694.191	701.510	692.680

The decision trees had an error rate of around 25%. While a decision tree is a popular choice for a classification model, we will not look at a decision tree model.

Logistic Regression - Modeling Dummy Dependent Variables
Darden School of Business, University of Virginia
Dr. Phillip E. Pfeifer

This case study from the Darden School of Business examines regressing Choice against just one variable in the dataset, “Recency” (“Last_purchase” in our dataset) in the Charles Book Club data (this is an important point, as we will discuss later). It looks at both linear and logistic regression of Choice~Recency.

When the dependent variable is a dummy variable, there are three reasons that a linear model is not appropriate:

- Non-normal residuals
- Heteroskedasticity (variability of the dispersion differs across elements of the vector)
- Model will fit only on a limited range of X values

Because y is either a zero or one, residuals cannot be normally distributed; they will cluster at the low and high ends of residual values with no values in between. Also, when Y is either zero or one, the ordinary regression line is affected because in theory it should eventually go below zero and above one. In the case of dummy variables, this can’t logically happen.

The case study presents the predictions of both a linear regression and a logistic regression. It is noted that the results are pretty similar between the two methods. The results of the logistic regression are slightly better, and it avoids illogical results like a negative probability that a customer would buy the book at some values of Recency.

Another point made by the case study author is the characteristics of the Recency variable. Dr. Pfeiffer noted that if you look at the distribution of the Charles Book Club data, it is clear that the mailings were sent every two months, and the value of Recency will always be an even number, because you see many observations at the exact data points, 2, 4, 6, etc.[4] This had implications for a regression against Recency. We should note here that our dataset does have odd numbers for Last_purchase, and it is unclear why the data used by Dr. Pfeifer would have different characteristics. It's likely that there is a difference between the Charles Book Club data and the BBB Club data that it is based on. Nevertheless, this is a reminder to check individual variables for non-normal distributions. We will look at this later in this case study.

Methodology and Assumptions

To find the best model, we will be training our models using 3 different techniques.

1. Linear Regression
2. Logit
3. Support Vector Machine (SVM)

Linear Regression

Although we will attempt to run a linear regression model based on BBBC's options for consideration, we do not believe this technique is appropriate for this particular case due to the following reason. Linear regression models are more suitable for y variables that have continuous values. In this case, we are looking to classify our predictions into a binary response (discrete values).

Logistic Regression

For the logit model, we will attempt the following techniques: no cross-validation, 5-fold cross-validation, and 5-fold repeated cross-validation to see which model will yield better results.

SVM

For the SVM model, we will attempt to apply tuned parameters to yield the best model possible. While SVMs might be good for accuracy, logit models are easy to interpret. We understand that some important factors to consider are accuracy and interpretability, so we will have to weigh both when selecting our final model.

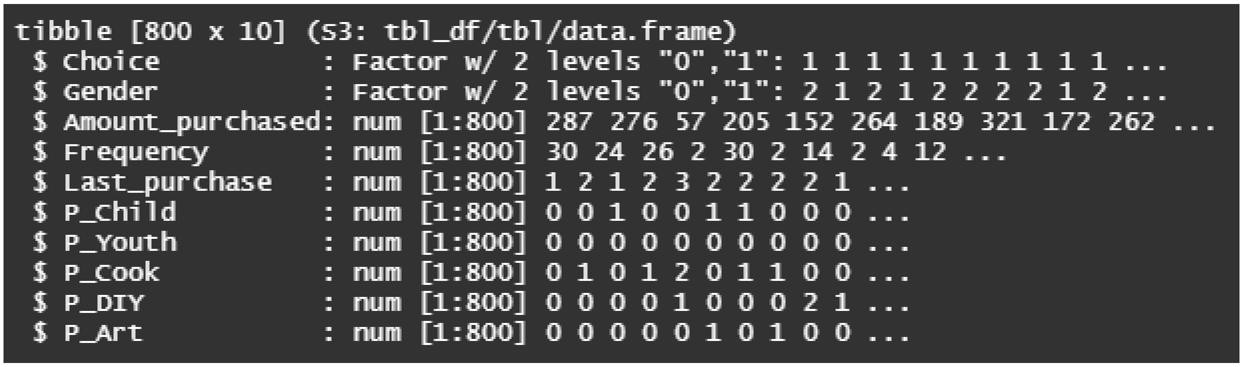
Data

The BBBC dataset is a subset of a targeted mailing to 20,000 customers in Pennsylvania, New York and Ohio for the purpose of developing a model to predict customer purchasing decisions for their broader customer base. Our initial dataset is a subset of the campaign data and consists of 1600 observations with 12 variables describing the data. This will be used as our training set and was provided by Dr. Arka Roy. 400 of the observations are customers who purchased the book while the remaining 1200 observations are customers who did purchase the book. Our testing dataset has 2300 observations and 12 variables also provided by Dr. Arka Roy. See appendix for variable descriptions.

We performed the following tasks to pre-process and clean the data, and adjust for its limitations.

1. We removed the “Observation” variable from both the training and test data sets since it is row numbers, not actual data.
2. We noticed a correlation between the “First_purchase” and “Last_purchase” variables, so we removed the “First_purchase” variable from the train and test datasets as it seems the last purchase will be more important.
3. We did not find any missing data to remove.
4. All of the variables have a number data type. We changed the “Choice” and “Gender” variables to factors as they are binary on the training and test datasets.
5. Since the data is unbalanced, we removed 800 non-book purchasers from the 1200 non-book observations from the “Choice” variable. This created a balanced dataset with 400 non-purchasers and 400 purchasers of “The Art of Florence” book in our training dataset.

The structure of the final data shows the Choice and Gender variables as factors, all others as numeric, and 800 observations for the 400 purchasers and non-purchasers.



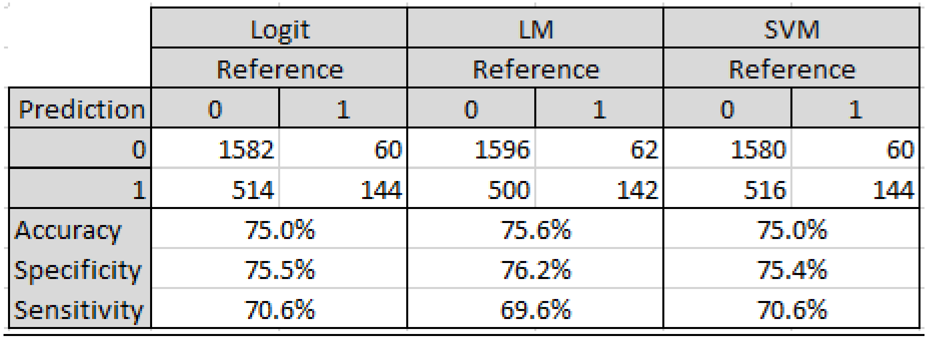
```
tibble [800 x 10] (S3: tbl_df/tbl/data.frame)
 $ Choice         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ Gender         : Factor w/ 2 levels "0","1": 2 1 2 1 2 2 2 2 1 2 ...
 $ Amount_purchased: num [1:800] 287 276 57 205 152 264 189 321 172 262 ...
 $ Frequency      : num [1:800] 30 24 26 2 30 2 14 2 4 12 ...
 $ Last_purchase  : num [1:800] 1 2 1 2 3 2 2 2 2 1 ...
 $ P_Child        : num [1:800] 0 0 1 0 0 1 1 0 0 0 ...
 $ P_Youth        : num [1:800] 0 0 0 0 0 0 0 0 0 0 ...
 $ P_Cook         : num [1:800] 0 1 0 1 2 0 1 1 0 0 ...
 $ P_DIY          : num [1:800] 0 0 0 0 1 0 0 0 2 1 ...
 $ P_Art          : num [1:800] 0 0 0 0 0 1 0 1 0 0 ...
```

Another limitation of the data is that most of the variables are right-skewed, with the variables for type of book especially so. This makes sense logically; while there may be book lovers who are interested in a subject and purchase many of those books, most customers will buy just a few books in a certain genre, and many will never buy a book in that genre. Thus, for the different book genres most total customers purchase 0-2 books. Another striking characteristic of the data is that almost all of the Last_purchase observations show zero months since the last purchase. This is good, because it shows that a large majority of our subscribers are purchasing items month-to-month.

Because we created a subset of the train data to balance observations for purchasers and non-purchasers, a large number of non-purchasers were removed from the model. While the distributions are still skewed, the skewness was improved. You can find graphs before and after balancing the data in the Appendix.

Findings

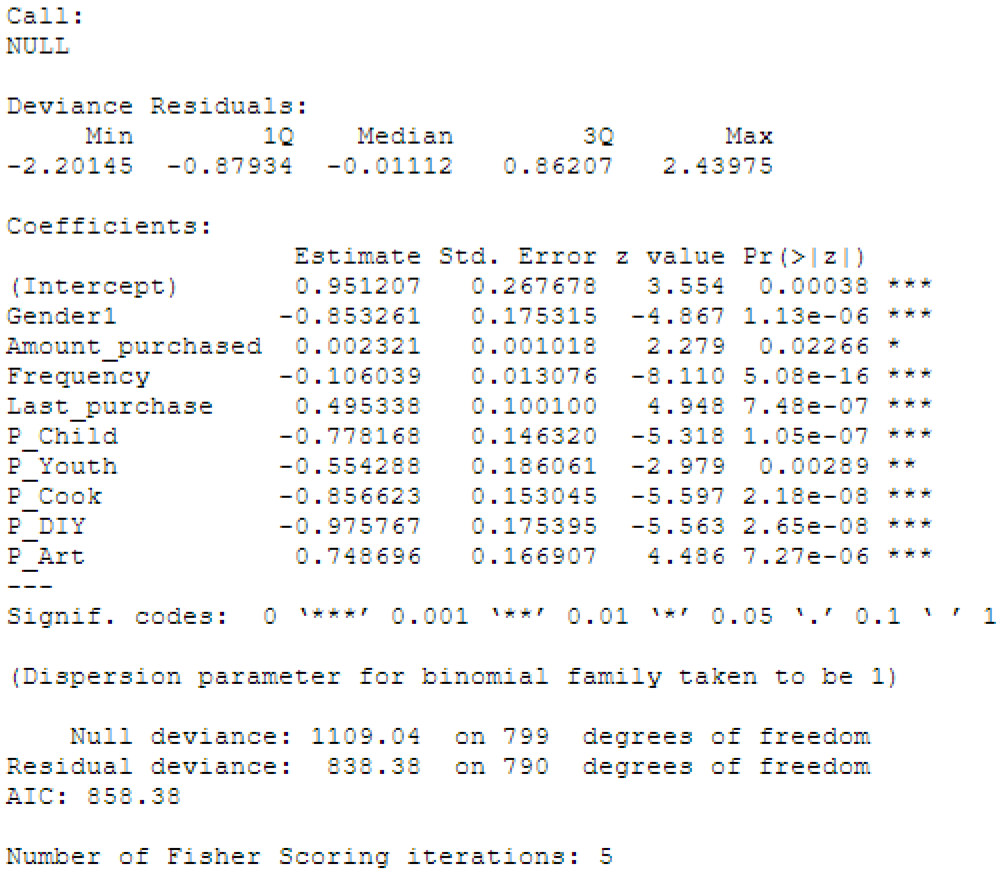
The linear model (LM), logistic model (logit), and support machine vector (SVM) models all performed about the same. A summary of the results:



	Logit		LM		SVM	
	Reference		Reference		Reference	
Prediction	0	1	0	1	0	1
0	1582	60	1596	62	1580	60
1	514	144	500	142	516	144
Accuracy	75.0%		75.6%		75.0%	
Specificity	75.5%		76.2%		75.4%	
Sensitivity	70.6%		69.6%		70.6%	

The differences in accuracy, specificity and sensitivity are fractions of a percent. It would be tempting in this case to use the linear model, as it is simpler and easier to interpret. As we discussed above in the literature review, it would be inappropriate to use a linear model on this data.

The results of the selected logistic model show that all variables are significant. Some are negatively correlated with purchasing *Art of Florence:* males are less likely to purchase than females; customers who previously purchased a children's book, youth book, cookbook, or do-it-yourself book were less likely to purchase the *Art of Florence* book. Not surprisingly, customers who had previously purchased an art book were much more likely to purchase *Art of Florence.* The time of last purchase and the total amount purchased were also positively correlated with buying *Art of Florence.*



```
Call:
NULL

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.20145  -0.87934  -0.01112   0.86207   2.43975

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.951207   0.267678   3.554  0.00038 ***
Gender1          -0.853261   0.175315  -4.867 1.13e-06 ***
Amount_purchased  0.002321   0.001018   2.279  0.02266 *
Frequency        -0.106039   0.013076  -8.110 5.08e-16 ***
Last_purchase     0.495338   0.100100   4.948 7.48e-07 ***
P_Child          -0.778168   0.146320  -5.318 1.05e-07 ***
P_Youth          -0.554288   0.186061  -2.979  0.00289 **
P_Cook           -0.856623   0.153045  -5.597 2.18e-08 ***
P_DIY            -0.975767   0.175395  -5.563 2.65e-08 ***
P_Art             0.748696   0.166907   4.486 7.27e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1109.04  on 799  degrees of freedom
Residual deviance:  838.38  on 790  degrees of freedom
AIC: 858.38

Number of Fisher Scoring iterations: 5
```

Accuracy

We found all the variables to be statistically significant with an alpha of .05 in the linear and logistic regression models. This means that all of our independent variables influenced our subscribers to purchase *The Art of Florence*. These variables include: gender, the purchase amount, frequency, number of months since last purchase, children, art, youth, cooking, and DIY books. Although the linear regression accuracy is highest, we used the logit model repeated cross-validation as our final model. We went into detail about why linear regression isn't the best model in the conclusion section.

Specificity and Sensitivity

Although the Specificity and Sensitivity percentages are fairly close between all three models, we will go into detail on the Logit model results. In our logistic regression model, the sensitivity is 70.6% and the specificity is 75.5%. This model will correctly identify 70.6% of the subscribers who will purchase a book, but it will also fail to identify 29.4%. The model will correctly identify 75.5% who will not purchase a book, but it will also identify 24.5% of people as purchasing the book when they will not. These are fair numbers when we compare them with the other models.

Conclusions and Recommendations

Advantages and Disadvantages

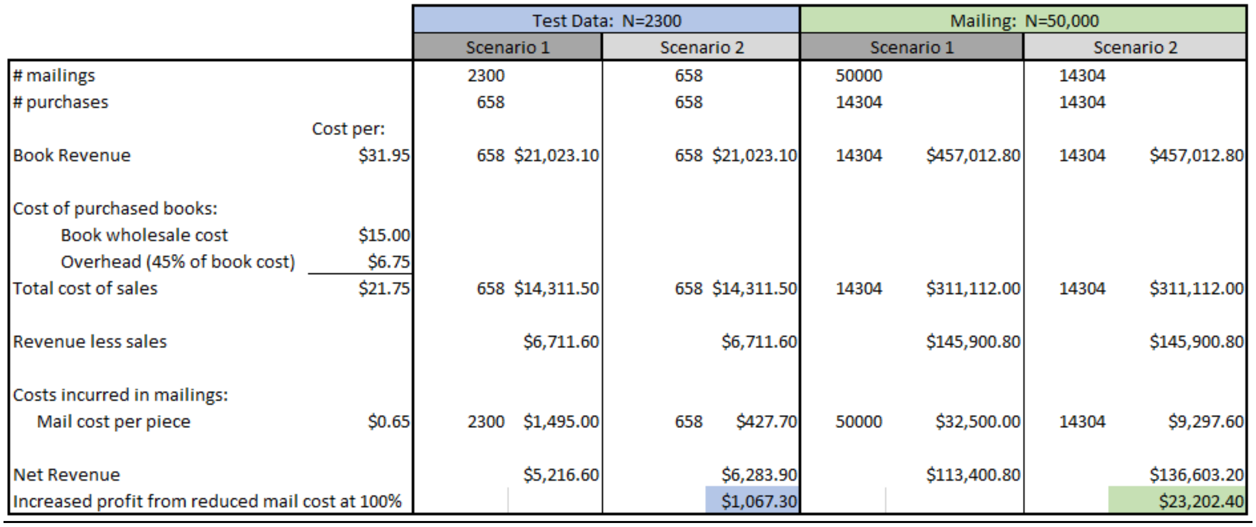
In our analysis, we were able to convert our data to fit the requirements of linear regression using continuous values and convert our predicted responses to binary responses. However, finding the proper threshold to convert our predicted values into binary responses was not always clear. In our case, we used a threshold of 1.5, but when introducing new data to the model, this threshold can vary and will not be reliable for accurate predictions.

In our study, a logistic regression model was the most efficient to train, was easy to interpret, and was also easily implemented. We ran a regular logistic regression model with no cross-validation, 5-fold cross-validation, and 5-fold repeated cross-validation. Our results were exactly the same for all three models. The cross validation model created a dataset and then evaluated a logistic regression model on it using cross-validation ten times. The average accuracy of the classification was reported. It was expected that repeated k-fold cross-validation would result in a more reliable estimate of model performance than a single k-fold cross-validation. This may mean less statistical noise, but that wasn't the case. It seems that all three models had the same Accuracy, Specificity, and Sensitivity because logistic regression has no hyperparameters to tune over; the estimates for the coefficients will always be given by maximum likelihood. Repetition of k-fold cross validation will not change the estimates of the parameters/coefficients.

The advantages of using support vector machines on data is that they work relatively well when there is a clear margin of separation between classes. They also work well on small datasets like our 800 observations. Comparing the logistic model to the support vector machine model, we found the logistic model to have a slightly higher accuracy. A disadvantage of using the support vector machine is that it doesn't work as well when there are a few features and a larger amount of observations. After cleaning our data, we had 10 features and 800 observations for the support vector machine model.

Cost Analysis

The BBB Club wishes to predict who is a likely purchaser of their book to project the savings of sending to a targeted group rather than the whole membership. To calculate the net increase in profit, the only differential is the cost to mail the advertisement to all members versus the cost to mail the advertisement to only those members who are likely to purchase it. The revenue, cost of books sold and overhead per book sold are the same in both scenarios. First we can calculate the difference in mailing expense, i.e. the change in expected profit, for the test group. Then we can extrapolate that calculation to the 50,000 mailings. Purchasers were 28.6% of our test data and non-purchasers were 71.4% of the test data; this translates to 14,304 purchasers and 35,696 nonpurchasers for the 50,000 group. If we were 100% confident of the predictions, we could take the difference between total mailings and mailings with a purchase, times $0.65/mailing. The number in green is the maximum predicted increased profit: $23,202. Because our predictions are not 100% accurate, with a 75% accuracy rate we can say the minimum predicted profit is 75% of that amount, and the actual incremental profit will be in a range between $17,402-$23,202.



		Test Data: N=2300				Mailing: N=50,000			
		Scenario 1		Scenario 2		Scenario 1		Scenario 2	
# mailings		2300		658		50000		14304	
# purchases		658		658		14304		14304	
	Cost per:								
Book Revenue	$31.95	658	$21,023.10	658	$21,023.10	14304	$457,012.80	14304	$457,012.80
Cost of purchased books:									
Book wholesale cost	$15.00								
Overhead (45% of book cost)	$6.75								
Total cost of sales	$21.75	658	$14,311.50	658	$14,311.50	14304	$311,112.00	14304	$311,112.00
Revenue less sales			$6,711.60		$6,711.60		$145,900.80		$145,900.80
Costs incurred in mailings:									
Mail cost per piece	$0.65	2300	$1,495.00	658	$427.70	50000	$32,500.00	14304	$9,297.60
Net Revenue			$5,216.60		$6,283.90		$113,400.80		$136,603.20
Increased profit from reduced mail cost at 100%					$1,067.30				$23,202.40

Final recommendations

Mailing only to likely book purchasers increased the net revenue from the mailing by 20%, from $113,400 to $136,603. This is a significant increase which clearly shows the value of targeted mailings. Developing the capability to model this in-house could increase overall profitability of the BBBC.

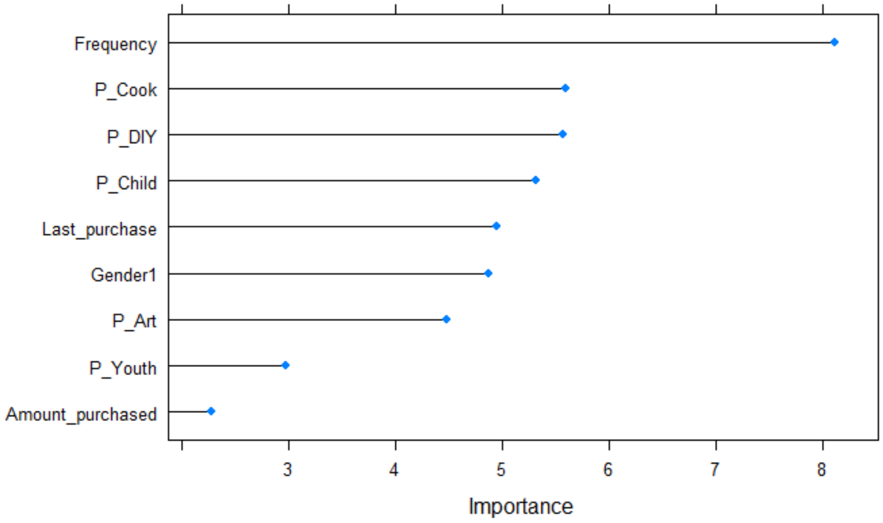
As the firm runs the model on a regular basis, they can refine their methodology. Perhaps not all of the variables will continue to be predictive, or will not be predictive for every type of book; for example, the number of customers who purchased a children's book may not be the best predictor of which ones will purchase an art book.

Appendix

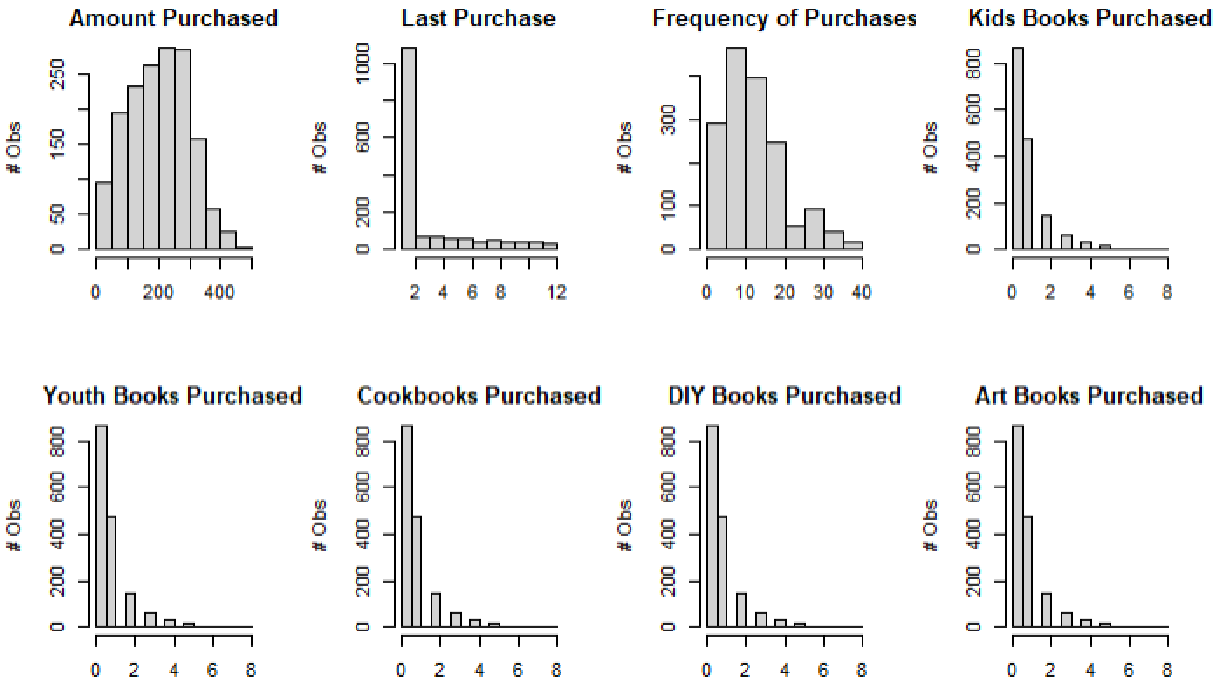
The variables describing the data include:

- **Choice:** 0 – did not purchase book, 1 – purchased book
- **Gender:** 0 – Female, 1 – Male
- **Amount_purchased:** The total money spent on BBBC books
- **Frequency:** Number of purchases in a chose time frame
- **Last_purchase:** Months since last purchase (recency of purchase)
- **First_purchase:** Months since first purchase (recency of purchase
- **P_child:** Number of children's books purchased
- **P_Youth:** Number of youth books purchased
- **P_Cook:** Number of cookbooks purchased
- **P_DIY:** Number of do-it-yourself books purchased
- **P_Art:** Number of art books purchased

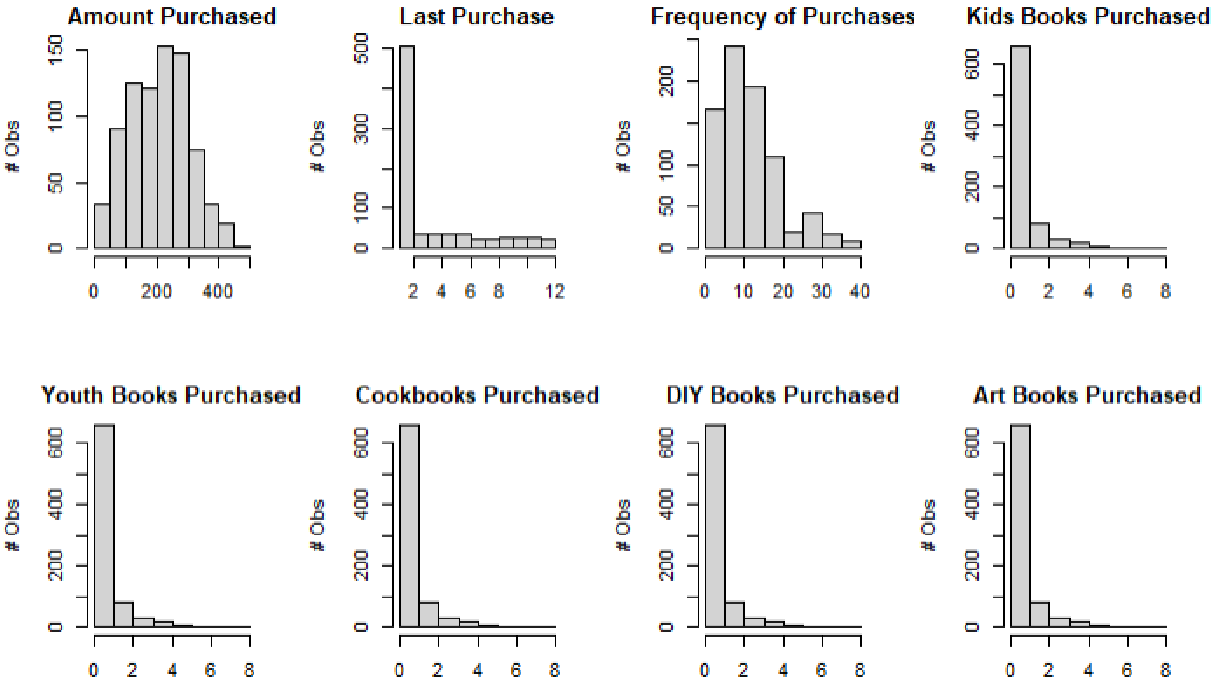
Chart listing variables in final logistic regression model in order of importance.



Histograms Before Data Balancing



Histograms After Data Balancing



Bibliography

CART data mining algorithm in plain English [WWW Document], 2018. . Hacker Bits. URL https://hackerbits.com/data/cart-data-mining-algorithm/ (accessed 9.30.21).

Grubinger, T., Zeileis, A., Pfeiffer, K.-P., 2014. evtree: Evolutionary Learning of Globally Optimal Classification and Regression Trees in R. J. Stat. Soft. 61, 1–29. https://doi.org/10.18637/jss.v061.i01

Pfeifer, P.E., 2006. Logistic Regression: Modeling Dummy Dependent Variables. Darden Business Publishing, University of Virginia.

Endnotes

[1] (Pfeifer, 2006)

[2] (“CART data mining algorithm in plain English,” 2018)

[3] (Grubinger et al., 2014)

[4] (Pfeifer, 2006)