Bookbinder Study Case

Alex Martinez, Josh Gardner, Cameron Playle, and Guillermo Gallardo

2024-09-25

Paper Starts Here

Executive Summary

Brief introduction of problem. Summarizes key findings. Summarizes insights behind key findings.

Our goal is to make a cost-effective decision for the book marketing campaign by either using a statistical model to target likely buyers or, if not worthwhile, sending the offer to everyone on our list to maximize profitability.

We used three models—GLM, SVM, and LDA. The model that outperformed the others is: ADD MODEL. What we found is ADD KEY FINDINGS

Our Problem

Clear description of the problem, from an application and theoretical point of view. Outlines the report.

For this study case our goal is to evaluate how effective three different models are and compare them with the option of creating this campaign without a model. We are trying to determine which model will provide the best balance between cost savings and profit. By analyzing the performance of each model and comparing it against the campaign, we will identify the most cost-effective approach that maximizes profit.

The three models we are comparing are the Linear Model (LM), Generalized Linear Model (GLM), and Support Vector Machine (SVM). Although we anticipate that the linear model may not perform well, we are still interested in understanding why it may not be the best fit for this case study. This exploration will help us gain valuable insights into the limitations of the linear model in this context and guide our decision-making process.

He said that we don't have to use the 50k population for our calculation to find which way would be best for the campaign. Email everyone in the 2300 or just a group of people based on our model.

Literature Review

Discusses and cites existing works in the theoretical and application realm.

The use of predictive modeling in marketing has been widely documented in both theoretical and applied settings. Predictive models, including linear regression, logistic regression, and support vector machines (SVM), have long been used to anticipate customer behavior based on historical data. In particular, database marketing, which focuses on analyzing customer data to personalize marketing efforts, has been a critical development in direct mail campaigns since the 1990s. According to Wilhelm (1994), database-driven approaches allow for the creation of tailored marketing strategies, improving response rates and customer engagement.

In the context of book clubs, predictive modeling can significantly enhance the efficiency of marketing campaigns. Doubleday's use of modeling techniques to analyze over 80 variables exemplifies how predictive analytics can identify the most influential factors in customer purchasing decisions. This approach aligns with the theoretical underpinnings of direct marketing and the application of consumer behavior models, which aim to increase the precision of marketing efforts (DM News, 1994).

Several studies have shown that logistic regression and SVM, in particular, offer high accuracy in binary classification problems such as purchase decisions, where the dependent variable is a choice between two outcomes. These methods, when combined with cost-benefit analyses, enable companies to target only those customers most likely to respond positively, thus maximizing profitability while minimizing wasted marketing spend. BBBC's interest in these models follows this established theoretical framework, as they seek to improve the efficacy of their direct mail program.

By utilizing predictive models, BBBC can determine which customers to target in their next campaign, reducing costs and increasing the likelihood of positive responses. Previous work in database marketing suggests that such targeted approaches can yield significantly better results than untargeted mailings, with predictive modeling emerging as a key tool for modern marketing efforts.

This approach is reinforced by the work of Levin and Zahavi (1994), who explored the application of machine learning techniques to marketing databases, confirming that models like SVM and logistic regression can offer actionable insights in direct mail marketing campaigns, improving response rates and customer satisfaction.

Methods

Discusses types of variables, sample size, and sampling techniques (if any). Discusses the model(s) and its assumptions and limitations.

We began our analysis by converting two variables, choice and gender, into factors. Additionally, we transformed *First_purchase* and *Last_purchase* into factor variables with four discrete buckets. However, we later chose not to include these variables in the final GLM model due to concerns about multicollinearity between them. The training dataset contained 1,600 observations, while the testing dataset had 2,300 observations.

Throughout the analysis, we utilized several models: logistic regression, Support Vector Machine (SVM), and Linear Discriminant Analysis (LDA). Although we initially considered using a linear regression model, it was not suitable for our problem because the response variable was binary.

For the logistic regression model, we made several assumptions, including that the dependent variable was binary, the observations were independent, and there was a linear relationship between the predictors and the log-odds of the outcome. We also assumed that there was no multicollinearity among the predictors and that there was no perfect separation in the data.

In the case of the SVM model, we assumed that the data was linearly separable and that the outcome was not imbalanced. To address any potential imbalance, we balanced the test data before building the SVM model.

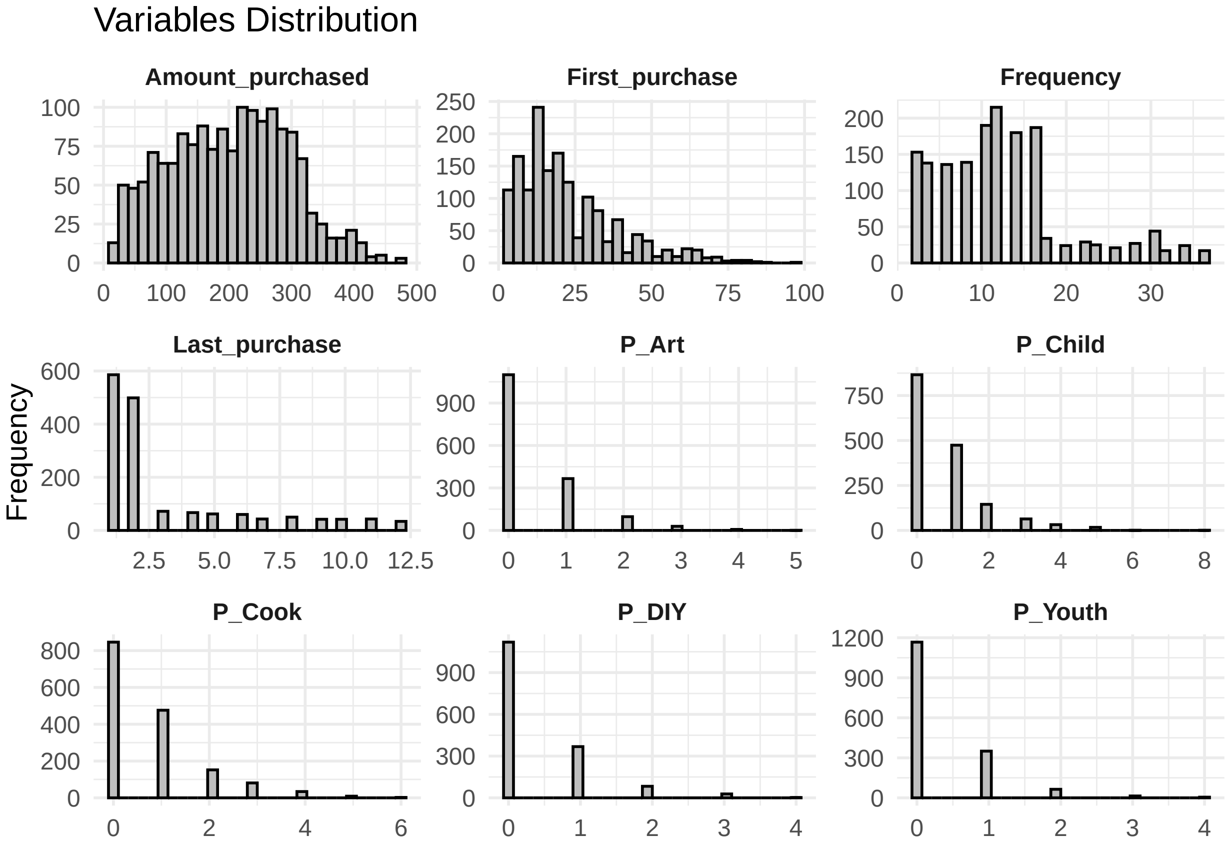
Finally, for the LDA model, we assumed that the data was normally distributed within each class, that there was homoscedasticity (i.e., equal variance-covariance structure across classes), that the classes were linearly separable, and that the observations were independent.

Data

Discusses how data was handled, i.e. cleaned and preprocessed. Discusses distributions, correlations, etc.

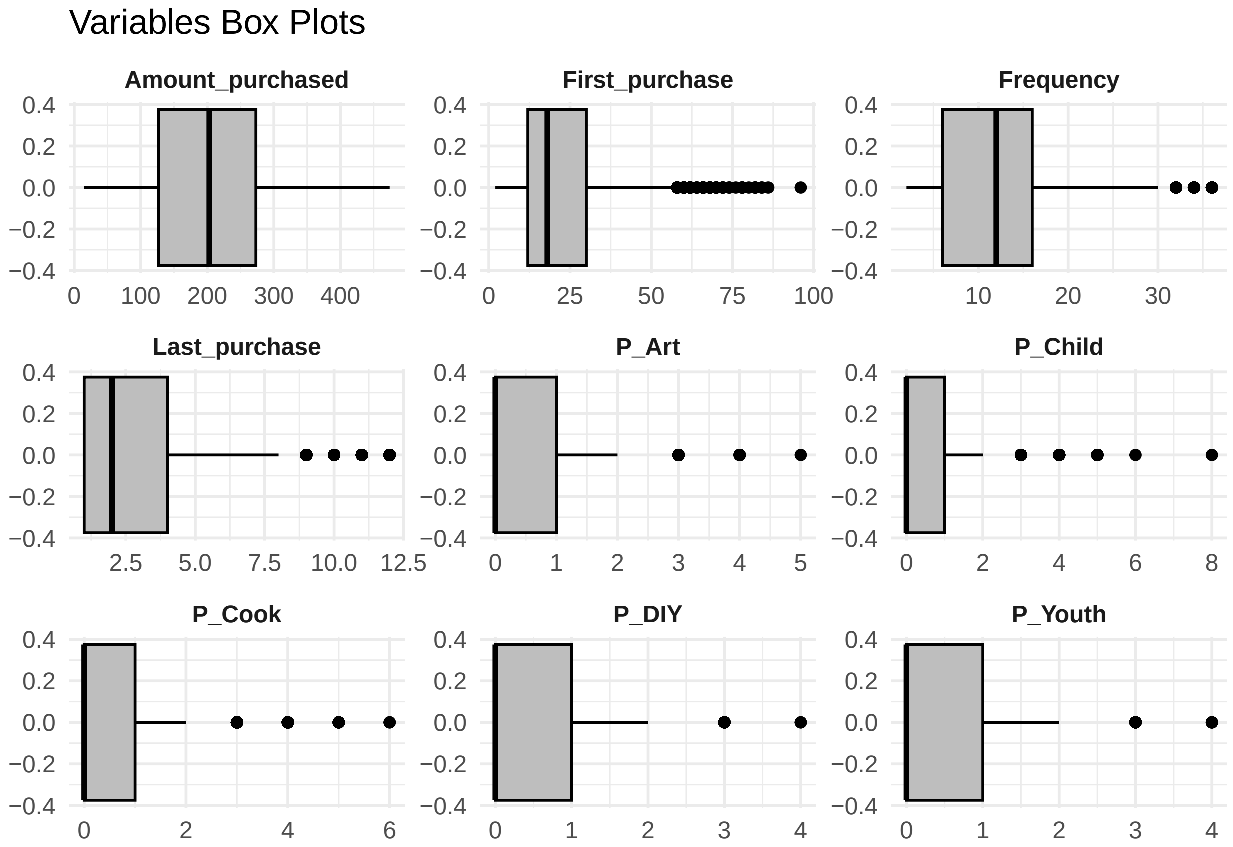
Distribution Plots

Below we can see the distribution of our variables. While many variables do not exhibit particularly informative distributions, two stand out for further analysis. Amount_purchased and First_purchase both show right-skewed distributions, indicating that most customers tend to make moderate purchases early on, with fewer customers making larger or later purchases.



Box Plots

The box plots reveal key patterns in the dataset. Amount_purchased shows a central range between 100 and 300 units, with a few high outliers, while First_purchase has most values below 50 but several outliers extending beyond that. Frequency is concentrated between 5 and 20 purchases, with a few customers making more frequent purchases.

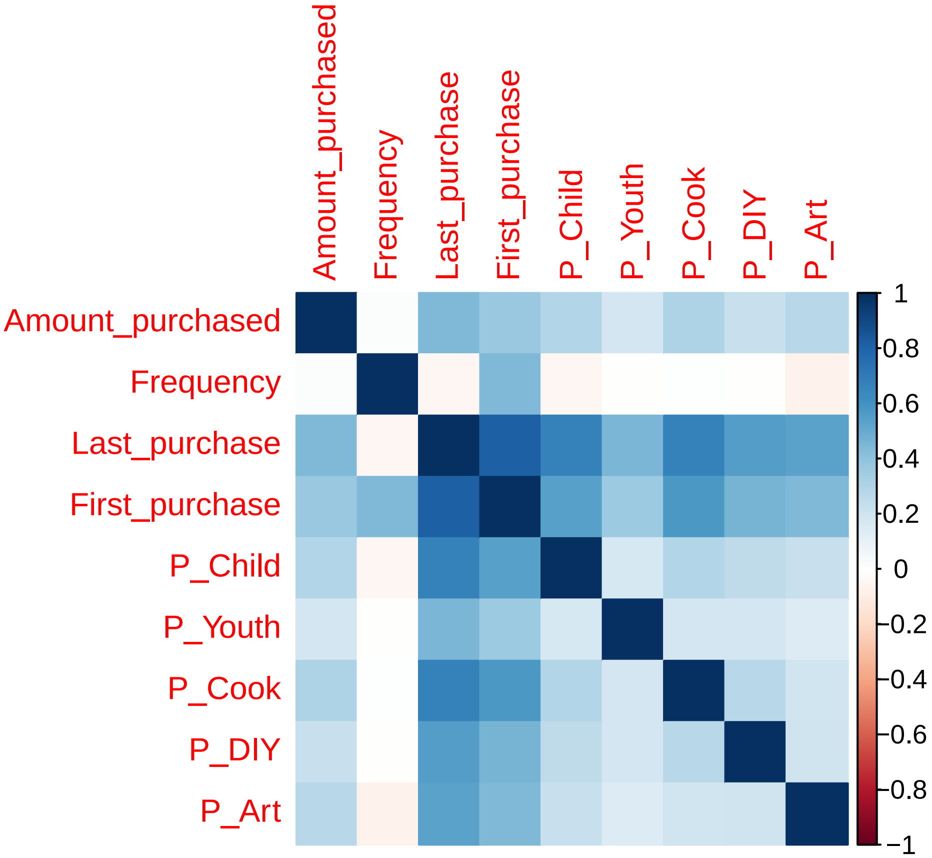


Clean up

We modified two variables into factors, and we created categories for ADD STUFF HERE?

Correlation

The graph below highlights the variables with the highest correlations. We observed that *first_purchased* and *last_purchased* exhibit the strongest correlation. Based on this, we decided to create labels for these two variables to test their potential impact on our model's performance.. **ADD STUFF ABOUT RESULTS. DID IT WORK OR DID IT MAKE IT WORST?**



Results

Presents and discusses the results from model(s). Discusses relationships between covariates and response, if possible, and provides deep insights behind relationships in the context of the application.

GLM Results

Here we will be discussing the outputs from our logistic model we applied to the data set to best predict if someone will purchase a book - Choice (1/0). We first begin with all the variables in the data set to see which independent variables are significant.

For our model, we first ran the GLM removing the observations variable and from our results ew noticed that the variable *Last_Purchase* had the largest VIF so we decided to remove that from the model. On our second run of the GLM, we found that *First_purchase* now had the largest VIF so we decided to remove from our second iteration of the model. On our third iteration, we found that P_Youth had a p-value greater than 0.05 so we removed it from our final model. Our final model (below), includes all the significant variables

GLM Final Model

Logistic Regression Results

Summary of Model Coefficients

Variable	Estimate	Standard Error	t-value	p-value
(Intercept)	0.3731	0.0303	12.3297	0.0000
Gender	−0.1264	0.0204	−6.2061	0.0000
Amount_purchased	0.0004	0.0001	3.2893	0.0010
Frequency	−0.0112	0.0012	−9.1034	0.0000
P_Child	−0.0277	0.0100	−2.7674	0.0057
P_Cook	−0.0429	0.0102	−4.2166	0.0000
P_DIY	−0.0386	0.0152	−2.5366	0.0113
P_Art	0.2183	0.0140	15.6104	0.0000

Table 2: Variance Inflation Factors

	Variable	VIF
Gender	Gender	1.00
Amount_purchased	Amount_purchased	1.22
Frequency	Frequency	1.01
P_Child	P_Child	1.21
P_Cook	P_Cook	1.21
P_DIY	P_DIY	1.15
P_Art	P_Art	1.14

Odds Ratio

Before we get into the confusion matrix. Lets explain the relationship of each independent variable to the dependent variable. We first compute the exponential of our coefficient ratio to get the odds ratio.

Table 3: Odds Ratios from the Logistic Model

	Variable	Odds Ratio
Gender	Gender (Male)	0.8812904
Amount_purchased	Amount Purchased	1.0003680
Frequency	Purchase Frequency	0.9888287
P_Child	Child Books Purchased	0.9727118
P_Cook	Cook Books Purchased	0.9579945
P_DIY	DIY Books Purchased	0.9621561
P_Art	Art Books Purchased	1.2439095

We observe the following key findings from the model:

- Gender (Male): Males decrease the odds of a client buying a book by a factor of 0.88.
- Amount of Books Purchased: A larger amount of books purchased increases the odds of a client buying a book by a factor of 1.
- Purchase Frequency: A higher frequency of books purchased decreases the odds of a client buying a book by a factor of 0.99.
- Child Books Purchased: A higher purchase of child books increases the odds of a client buying a book by a factor of 0.97.

- DIY Books Purchased: A higher purchase of DIY books increases the odds of a client buying a book by a factor of 0.96.
- Art Books Purchased: A higher purchase of art books increases the odds of a client buying a book by a factor of 1.24.

With this model, we then use it on our bbc_test sample to see how well it predicts and determine the most optimal cutoff to have the highest Sensitivity. Here are the following results:

GLM Confusion Matrix

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2005   91
##          1  140   64
##
##                Accuracy : 0.8996
##                  95% CI : (0.8866, 0.9116)
##     No Information Rate : 0.9326
##     P-Value [Acc > NIR] : 1.000000
##
##                   Kappa : 0.3032
##
##  Mcnemar’s Test P-Value : 0.001588
##
##             Sensitivity : 0.41290
##             Specificity : 0.93473
##          Pos Pred Value : 0.31373
##          Neg Pred Value : 0.95658
##              Prevalence : 0.06739
##          Detection Rate : 0.02783
##    Detection Prevalence : 0.08870
##       Balanced Accuracy : 0.67382
##
##        ’Positive’ Class : 1
##
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2095    1
##          1  199    5
##
##                Accuracy : 0.913
##                  95% CI : (0.9008, 0.9242)
##     No Information Rate : 0.9974
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0428
##
```

```
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.833333
##             Specificity : 0.913252
##          Pos Pred Value : 0.024510
##          Neg Pred Value : 0.999523
##              Prevalence : 0.002609
##          Detection Rate : 0.002174
##    Detection Prevalence : 0.088696
##       Balanced Accuracy : 0.873293
##
##        'Positive' Class : 1
##
```

```r
#| echo: false

# Predict the probability on bbc_test dataset
bbc_test$PredProb <- predict.glm(glm1, newdata = bbc_test, type = 'response')

# Define a function to extract metrics from the confusion matrix
get_metrics <- function(cutoff) {
  # Create prediction indicators based on cutoff
  bbc_test$Pred_Y <- ifelse(bbc_test$PredProb >= cutoff, 1, 0)

  # Calculate confusion matrix
  cm <- confusionMatrix(as.factor(bbc_test$Choice), as.factor(bbc_test$Pred_Y), positive = '1')

  # Extract the relevant metrics
  data.frame(
    Cutoff = cutoff,
    Accuracy = cm$overall['Accuracy'],
    Sensitivity = cm$byClass['Sensitivity'],
    Specificity = cm$byClass['Specificity']
  )
}

# Get metrics for cutoff values 0.5 and 0.8
metrics_0.5 <- get_metrics(0.5)
metrics_0.8 <- get_metrics(0.8)

# Combine the results into a single data frame
metrics_table <- rbind(metrics_0.5, metrics_0.8)
```

Table 4: Model Performance Metrics at Different Cutoffs

	Cutoff	Accuracy	Sensitivity	Specificity
Accuracy	0.5	0.8995652	0.4129032	0.9347319
Accuracy1	0.8	0.9130435	0.8333333	0.9132520

We see the best cutoff for the highest sensitivity is at 0.8. With this, the performance of our model on the bbc_test data set is 91% accurate overall, and our sensitivity is 83% with a specificity of 91%

SVM

LDA

Conclusion

Concludes with a summary of the aim and results. Discusses alternative methods that can be used.

Based on our analysis, the **ADD MODEL**

Add details on if it is better to use model or just send campaign to everyone in the list.

STUFF FROM THE PDF

Summarize the results of your analysis for the three models. The training, testing, and prediction data can be found on Blackboard.

Interpret the results of the models. In particular, for models the influential covariates and their coefficients, provide insights.

BBBC is considering a similar mail campaign in the Midwest where it has data for 50,000 customers. Such mailings typically promote several books. The allocated cost of the mailing is $0.65/addressee (including postage) for the art book, and the book costs $15 to purchase and mail. The company allocates overhead to each book at 45% of cost. The selling price of the book is $31.95. Based on the model, which customers should Bookbinders target? How much more profit would you expect the company to generate using these models as compare to sending the mail offer to the entire list.

Please also summarize the advantages and disadvantages of the three models, as you experienced in the modeling exercise. Should the company develop expertise in either (or all) of these methods to develop in-house capability to evaluate its direct mail campaigns.

How would you simplify and automate your recommended method(s) for future modeling efforts at the company?