Bookbinder Study Case

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

2024-10-6

Executive Summary

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.

By leveraging our GLM model with a specificity rate of 91%, we can accurately identify which customers should receive marketing flyers. This targeted approach is projected to yield a 200% increase in profit from art book sales compared to sending flyers to the entire customer base. The profit increase is largely attributed to the significant reduction in mailing costs by focusing on a smaller group of customers who are more likely to make a purchase.

Our Problem

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.

Literature Review

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

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.

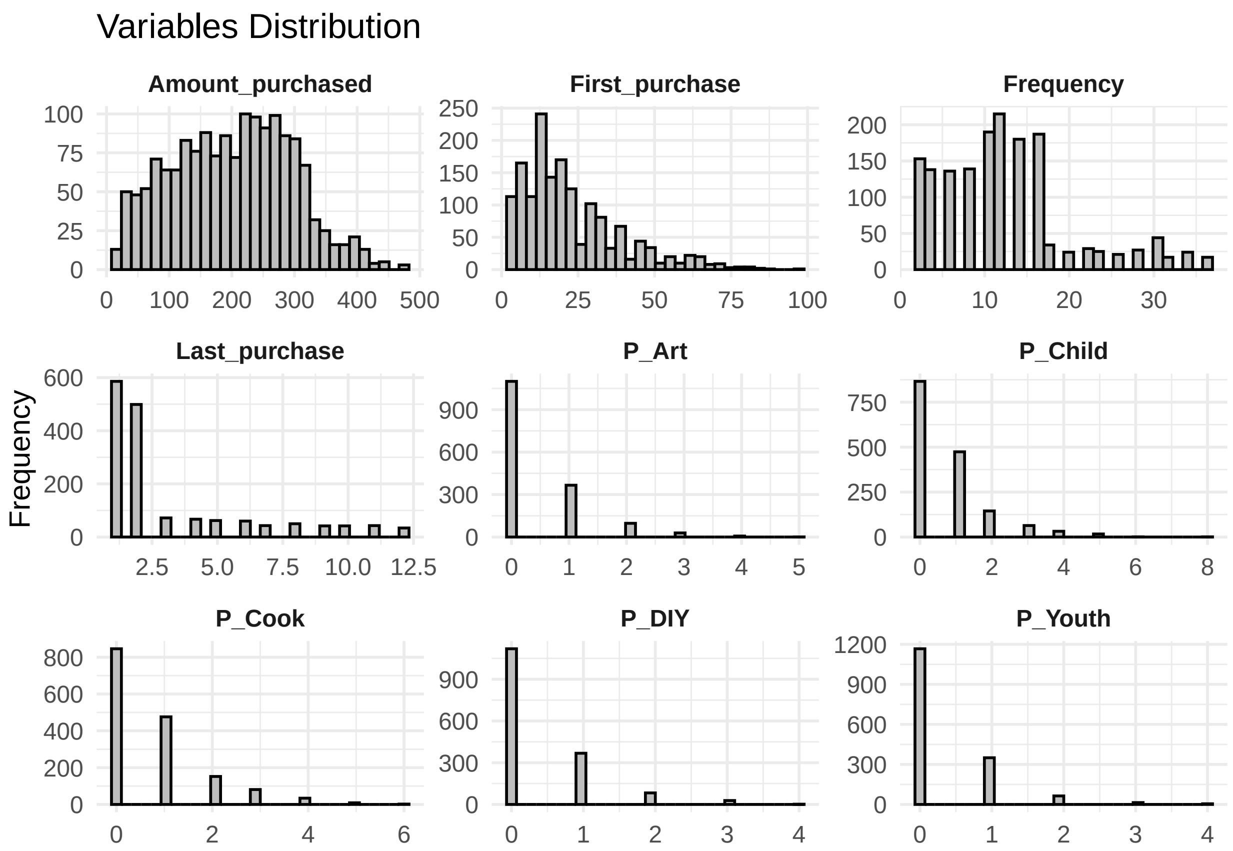
In the case of the SVM model, we assumed that the data was linearly separable and that the outcome was not imbalanced. To address any potential imbalance, we balanced the test data before building the SVM model.

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

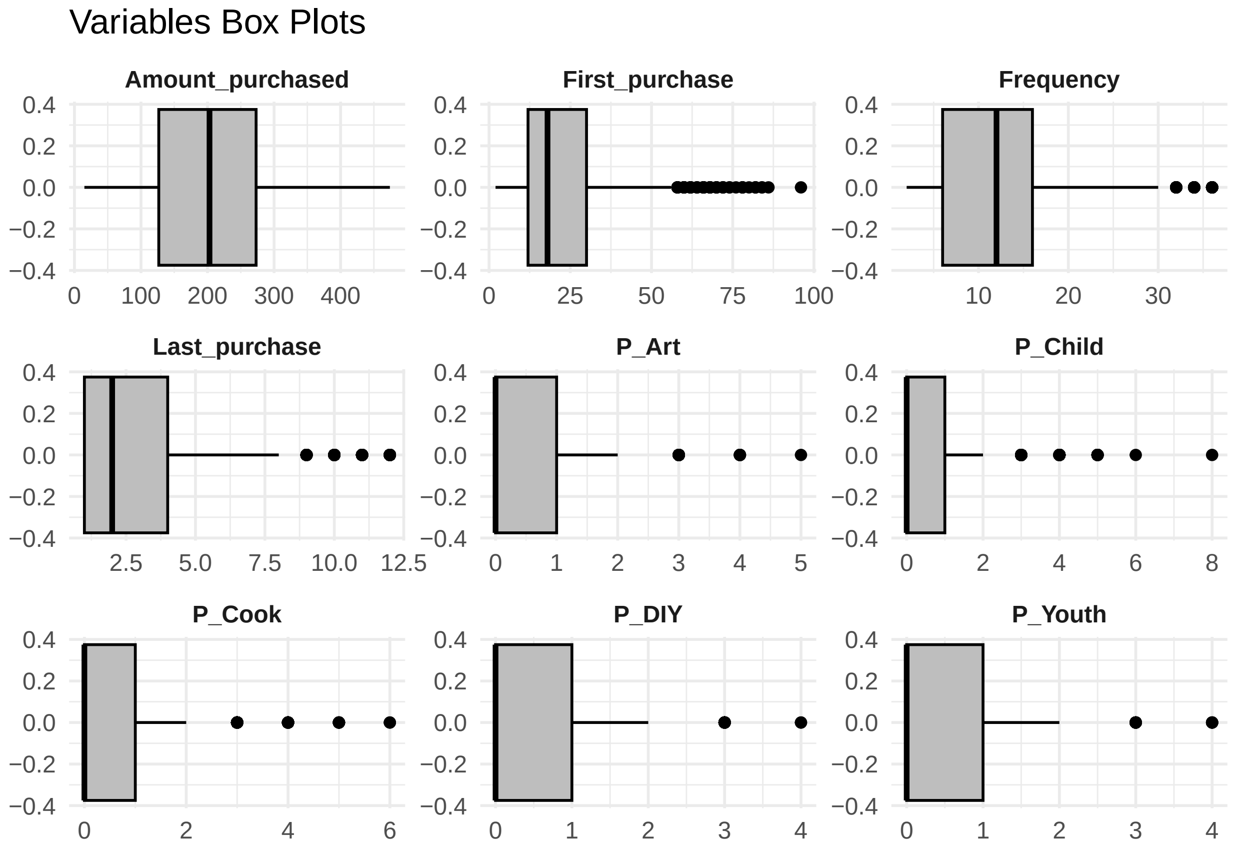
Distribution Plots

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.



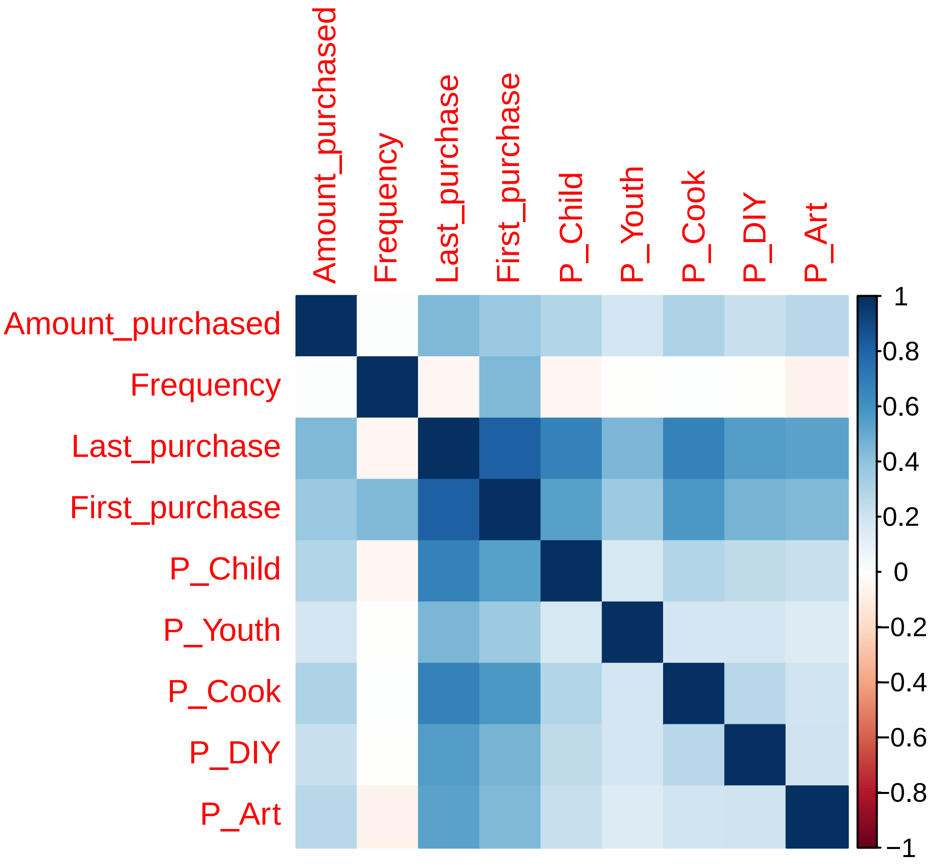
Box Plots

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.



Correlation

The graph below shows the variables with the highest correlations, with *First_purchased* and *Last_purchased* having the strongest correlation. Based on this observation, we created labels for these two variables to test their impact on the model's performance. However, our results indicated that including them did not improve the model, so we chose to exclude them.



Results

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.

We initially ran a GLM excluding the Observation variable. From the results, we identified that Last_Purchase had the highest VIF, prompting its removal. In the second iteration, First_purchase showed the highest VIF, so we excluded it as well. In the third iteration, P_Youth had a p-value greater than 0.05, leading us to remove it from the model. Our final model, shown below, includes all the remaining 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

GLM 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

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%

LDA

The LDA model summary presents the coefficients for each variable and the overall model accuracy. Key variables like Gender, Frequency, P_Child, P_Cook, and P_DIY have negative coefficients, indicating a negative relationship with the outcome, while Amount_purchased and P_Art show positive coefficients, with P_Art having the strongest positive impact (1.0525). The model achieved an overall accuracy of 0.7457, suggesting that it performs moderately well in predicting the target outcome based on these variables.

LDA Model Coefficients and Accuracy

Summary of the LDA Model and its Performance

Term/Metric	Value
Gender	−0.7452
Amount_purchased	0.0021
Frequency	−0.0761
P_Child	−0.1390
P_Youth	0.0112
P_Cook	−0.2121
P_DIY	−0.2781
P_Art	1.0525
Model Accuracy	0.7457

SVM

Our ouput shows our gamma and cost for our best model. Our gamma is 0.05 and Cost is 0.1.

```
##   gamma cost
## 4  0.04  0.1
```

Conclusion

If we send mailers to all 50,000 customers, we project a profit of approximately $16,970. This calculation is based on the expected 4,848 purchases (derived from the proportion of customers who bought a book in our test data) multiplied by a $10.20 profit per book, minus the $0.65 mailing cost per address. However, by using our model, we could reduce mailing costs by around $29,000 by targeting only the most likely buyers.

In comparing models, we found that both the SVM and LDA models achieved more balanced specificity and sensitivity, while the logistic regression model had a higher sensitivity compared to specificity. Ultimately, we chose the logistic regression model because it offered the highest sensitivity. This decision was based on our goal to identify customers likely to purchase a book, where false positives are acceptable, especially given our model's 91% specificity.

I recommend that the company consider building an in-house team to manage this process, as the model predicts a 200% increase in profit when applied. This increase comes primarily from the significant reduction in mailing costs while still targeting the right customers.

To streamline our process, we focused only on the most significant variables for the GLM model, excluding three variables that were less relevant. This refinement enhanced the efficiency of data collection and improved the overall effectiveness of the model. For instance, insights from our LDA model revealed that past purchases of art books are strong predictors of future purchases, enabling us to better target customers. By incorporating these insights, we can enhance the accuracy and efficiency of our marketing efforts, ultimately making the entire process more streamlined and cost-effective.