Bookbinders Case Study
Brandi Rodriguez
VLJ466

I. Executive Summary

In this case study the Bookbinders Book Club (BBBC) considers the use of predictive modeling to improve the efficacy of its direct mailing program. They have developed a database containing all relevant information on 500,000 readers and would like to develop a response model that identifies the factors that influence their purchases. The case analysis uses a subset of the BBBC database to train linear regression, logistic regression, and support vector machine models. The linear regression model exhibited the poorest performance, proving to be unsuitable for making binary predictions. Logistic regression and SVM both performed well with accuracy rates of 89.57% and 90.97% respectively. The SVM was ultimately the most accurate model for predicting which customers will purchase *The Art History of Florence*, although the logistic regression provided insight on influential covariates**.** The findings from this case study can help BBBC with its targeting efforts for future mailing campaigns.

II. The Problem

Faced with intense competitive pressure from superstores and online superstores such as Amazon, book clubs are seeking alternative business models that offer greater flexibility and improve responsiveness to customer preferences. This case highlights how database marketing techniques and classification algorithms can help businesses work smarter to reach the right customer and understand their needs. The objective is to not only develop a highly accurate model, but to also identify the factors that most influenced customers to purchase the book. Although linear regression is not appropriate for this classification task, it will still be evaluated. The goal is to construct a classifier based on the training data that will correctly classify the observations using the available features. This report will include a review of related literature, discussion of the methodology, algorithms and data used, a comparison of the results, and will conclude with findings and recommendations for future modeling efforts to help BBBC to strategize and target the right customers to maximize profit.

III. Review of Related Literature

In *Support Vector Machines for Classification: a Statistical Portrait*, Yoonkyung Lee describes SVM as a 'hard' classification approach that departs from the more traditional 'soft' approach through the estimation of the underlying probabilities to predict class from methods, such as logistic regression (Yoonkyung Lee). The fact that there's no probabilistic interpretation makes SVMs difficult to interpret in comparison to other less complex classification models. In *An Application of Support Vector Machines to Customer Loyalty Classification of Korean Retail Company*, Nguyen explores the use of classification methods in R, including SVM, to classify loyal customers and determine which factors most have the greatest effect on customer loyalty for a Korean retail store. He highlights the risk of misclassifying customers as non-loyal when in fact they are loyal. This could be detrimental to company profits and makes customers feel they are not receiving proper treatment. Vice-versa, treating non-loyal customers as loyal creates an inefficient use of a company's time and resources, which can also dampen profits (Nguyen). In the study, the SVM performed the best with an accuracy rate of 95.6%, followed by a logistic regression (91.65%), discriminant analysis (90.33%), and random forest (89.45%).

IV. Methodology

For this case study, the results of a linear regression, logistic regression, and support vector machine model will be compared. The models were trained and executed after preprocessing and cleaning the dataset for missing values, high correlations, and influential observations. A detailed description of this process is included in the following section and the final code is included in Appendix

E. The analysis starts with linear regression which assumes the relationship between the dependent and independent variables is linear, homoscedasticity - the variance of residual is the same for any value of X, independence - observations are independent of each other, and normality - for any fixed value of X, Y is normally distributed. Linear regression is easy to interpret. However the target variable in this case is binary which violates the first assumption of linearity. It results in estimates outside the [0,1], making them difficult to interpret as probabilities. Linear regression is vulnerable to overfitting, not robust to outliers, and is limited by the assumption of a linear relationship between x and y, which is often not the case as is seen in this case study. To evaluate model performance, mean square error was calculated since a linear regression model cannot produce the confusion matrix since it's not a classification task.

Logistic regression assumes the outcome is a binary variable, that there's a linear relationship between the logit of the outcome and each predictor variable, there are no influential values (extreme values or outliers) in the continuous predictors, and there's no high intercorrelations (multicollinearity among the predictors). Although logistic regression models are easier to interpret, they can be rigid and sometimes cannot adequately model complex nonlinear relationships. It's vulnerable to overfitting and can easily be outperformed by more complex models, such as support vector machines. The logistic regression models were evaluated based on fit and performance metrics such as AIC, BIC, Accuracy, Sensitivity, Specificity and AUC.

SVM is a good alternative to logistic regression and works well in high dimensional spaces. It's very memory efficient, only relying on a subset of training points (the support vectors). SVM is ideal when you have a large number of datapoints because SVMs don't run out of memory, unlike other methods. It's also very versatile, since nonlinear kernels allow for higher flexibility for the decision boundary. SVM classifies observations by finding the optimal hyperplane that acts as a decision boundary to separate data into classes. Different kernels can be applied to find the best decision boundary. For this case, the default radial was used as well as a linear kernel. The radial returned the higher accuracy. SVM offers a more powerful solution to learn complex nonlinear functions (Bassey). SVM has parameters that can be tuned using tune.svm to perform a 10-fold cross validation. Gamma and cost can be used as arguments to tune the operation of the SVM where gamma is used by the kernel function and cost allows one to specify the cost of violating the margin (how heavily to weight misclassification). Kernels can also be specified and are used to transform the data to a higher dimension so it can be separated by hyperplanes. The SVM outperformed the logistic regression model, but it is difficult to interpret as there is no probabilistic interpretation. The best.parameters function was used to find the optimal parameters of gamma and cost, which were stored and then called when executing the SVM. Predictions were made using each model and tracked in the table seen in Appendix B: Results.

V. Data

A subset of the BBBC database was used as the training dataset. It consisted of data for 400 customers who purchased *The Art History of Florence* after receiving a mailing containing a brochure advertisement for it, and 1200 customers who didn't. The testing dataset included 2,300 customers. The response variable for this analysis is Choice, representing whether a customer purchased the book or not, and the dependent variables include:

- Gender
- Amount purchased: total money spent on BBBC books
- Frequency: total # of purchases in the chosen period (used as a proxy for frequency)
- Last_purchase (recency of purchase): months since last purchase
- First_purchase: months since first 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: # of art books purchased.

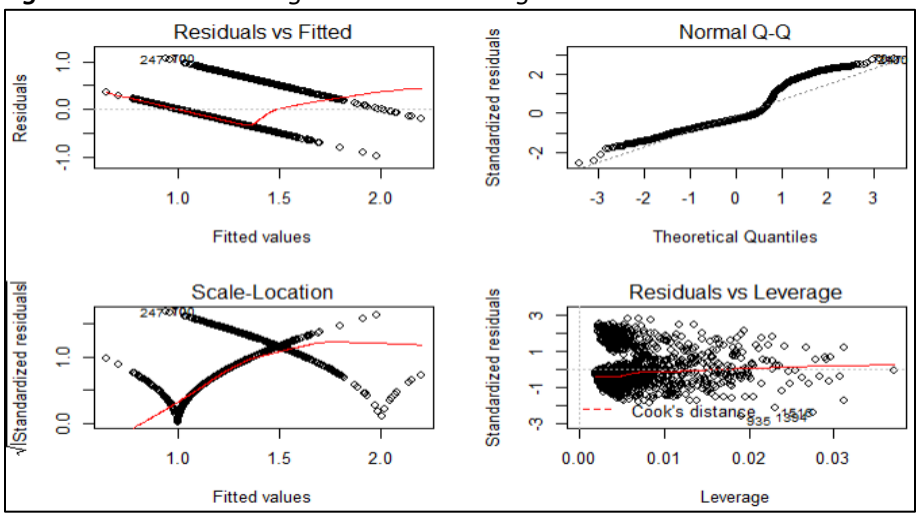
The pre-split training and testing datasets were imported and checked for duplicates and missing values (there were none). The categorical variables in both datasets, Choice and Gender, were converted to factor variables and several numeric variables were converted to integers because they are measures of discrete counts, and should not be treated as continuous numeric variables (Frequency, P_Child, P_Youth, P_Cook, P_DIY, P_Art). After cleaning the data, several plots were explored using the DataExplorer library to gain a better understanding of the distributions and relationships between the variables and the response (see plots in Appendix A). It revealed the majority of customers did not purchase the book and there were many more males sampled (gender = 1) than females. However, a larger proportion of females actually purchased the book. All numeric variables exhibited a right skewed distribution. Scatterplots are normally helpful for checking if there's a linear relationship between the independent variables and the response variable, as well as whether a data transformation may be needed to satisfy model assumptions of linearity, but they were unsurprisingly ineffective in this case since the response is a binary categorical variable. The variable 'Observation' is an indexing column and provides no real insight, so it was dropped from the dataset. Lastly, before getting started on model training, the data was checked for correlation. There were no exceptionally high correlations, but Last_purchase and First_purchase had the strongest correlation (.81). The variables weren't dropped but are later checked for muticollinearity by examining their VIFs.

V. Findings

This analysis compares the results of several linear and logistic regression models, as well as SVM. The results can be seen in Appendix B, as well as commentary and findings. To start, a full linear regression model was set as the baseline regression model. The First_purchase variable had a p-value greater than .05, indicating it is an insignificant predictor that is dropped in the second linear regression model. The first regression model had a mean square error (MSE) of .0926 and a low R^2 of .2401. Last_purchase, which had a high correlation to First_purchase, exhibited a high VIF of 18.11 and is subsequently dropped from the third linear regression model. The second linear regression model, which excluded First_purchase, was trained and resulted in the lowest MSE of all the linear regression models (.0924) and was ultimately selected as the final linear regression model. A third linear regression model was fit on all variables excluding Last_purchase instead of First_purchase, resulting in a lower R^2 and a .0005 increase to MSE. The last linear regression model is the same as the third model, but it excludes P_Youth as well because it was found to be an insignificant predictor by it's p-value greater than .05 in the prior linear regression model. This model actually resulted in the lowest R^2 and highest MSE.

Despite its selection as the final linear regression model, a look at its residuals plots prove that linear regression is still an unsuitable model for BBBC's classification task. The scatterplots previously plotted show the independent variables did not display linear relationships with the response variable. The first plot in the diagnostics plots is useful for checking the assumption of linearity and homoscedasticity. Instead of randomly scattered residuals with a straight and horizontal line centered around y = 0, which is characteristic of linearity, the residuals form a very distinctive pattern - two downward sloping lines and a bent line. To assess if the homoscedasticity assumption is met, the residuals should be equally spread around the y = 0 line, but they are not. The normality assumption can be evaluated by looking at the QQ plot. The normality assumption is violated, as the residuals do not follow closely along the 45-degree line. The third plot is useful for checking homoscedasticity. Ideally, the red line will be flat and horizontal with equally and randomly scattered data points, so clearly the homoscedasticity assumption is not satisfied. The fourth plot tells us there are a few influential points based on Cook's distance.

Figure 1. *Final Linear Regression Model Diagnostics Plots*



Logistic regression models produced more sound results than linear regression. A full logistic regression model exhibiting an 89% accuracy rate was trained and set as the baseline. All predictors were significant except for First_purchase, similar to the full regression model. A total of four logistic regression models were trained (results in Appendix B), but the final logistic regression model contained all variables except Last_purchase, which exhibited a high VIF. It tied with the fourth logistic regression model (which excludes P-Youth as well) for the highest accuracy rate of 89.57%. However, logit4 had a slightly lower AUC. A review of the assumptions for logistic regression was conducted, followed by an interpretation of the odds ratios. The plots in Appendix C would typically be used to visually inspect if there is a linear relationship between the continuous predictor variables and the outcome. However, in this case most of the numeric variables are not continuous. Instead, they are discrete integer variables measuring counts (i.e. Frequency, P_Art, P_Child, P_Cook, etc.). Amount_purchased shows a roughly linear association with the Choice outcome in logit scale, aside from the points farthest to the left in the plot. The model was also checked for influential observations based on Cook's distance. The absolute standardized residuals were all below 3, indicating there are no outliers. If there were, they could be removed, the data could be transformed to a log scale, or a nonparametric method could be used instead.

The odds ratios of the final logistic regression model provided valuable insight on the most influential covariates. The coefficients of logistic models are not as intuitive to interpret, so it's common to use odds ratios for interpretation instead. Table 1 in Appendix D contains interpretations of the odds ratios for each of the significant variables, while Figure 1 (in Appendix D) graphically displays the magnitude of its impact on the odds of purchase. BBBC would have better odds of purchase if they were to target their female customers and those who have previously purchased art books, which makes sense since the response is whether a customer purchased another art book, *The Art of Florence*. The number of cookbooks, DIY, children and youth books negatively impact a customer's choice to purchase the book.

The support vector machine algorithm produced the best accuracy rate of 90.96% when applied to the testing dataset using all of the predictors provided. The optimal parameters were .05 for gamma and .7 for cost. Using the default radial kernel, it produced 785 support vectors and had 372 observations classified on one side of the hyperplane and 413 on the other. The fourth SVM used a linear kernel, but was a lower performing model.

V. Conclusion

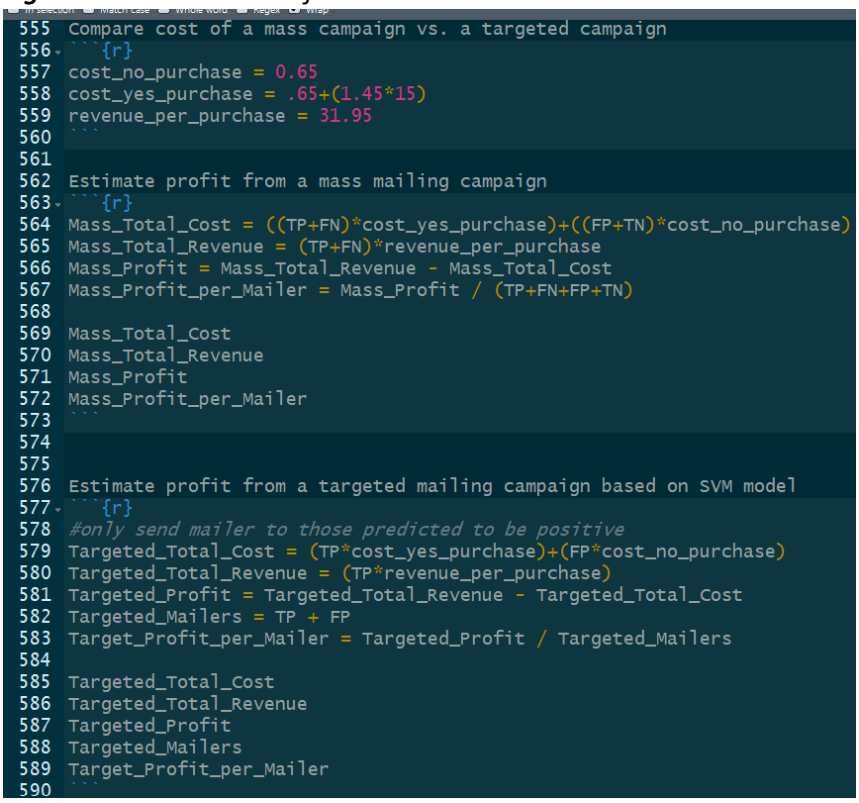
The SVM model can be used for future mailings to help generate profit more efficiently. For instance, assuming a scenario where a mailer costs $0.65 per addressee and the cost per book sold is

$22.40 ($15 per book + 45% overhead), a mass mailing to all 2,300 customers in the test dataset would generate more profit but much less efficiently than a targeted campaign would (returning only $0.25 in profit per mailer). Alternatively, a targeted campaign would generate less gross profit, but a much larger profit per mailer $4.20. A comparison of the results and calculations are presented in Table 1 and Figure 2 below.

Table 1. *Comparison of Profitability*

	Mass Campaign	Targeted Campaign
Number of Mailers	2300	80
Total Cost	$ 5,932.00	$ 878.50
Total Revenue	$ 6,517.80	$ 1,214.10
Profit	$ 585.80	$ 335.60
Profit per Mailer	**$ 0.25**	**$ 4.20**

Figure 2*. Calculations Performed in R*



```
555 Compare cost of a mass campaign vs. a targeted campaign
556 ```{r}
557 cost_no_purchase = 0.65
558 cost_yes_purchase = .65+(1.45*15)
559 revenue_per_purchase = 31.95
560 ```
561
562 Estimate profit from a mass mailing campaign
563 ```{r}
564 Mass_Total_Cost = ((TP+FN)*cost_yes_purchase)+((FP+TN)*cost_no_purchase)
565 Mass_Total_Revenue = (TP+FN)*revenue_per_purchase
566 Mass_Profit = Mass_Total_Revenue - Mass_Total_Cost
567 Mass_Profit_per_Mailer = Mass_Profit / (TP+FN+FP+TN)
568
569 Mass_Total_Cost
570 Mass_Total_Revenue
571 Mass_Profit
572 Mass_Profit_per_Mailer
573 ```
574
575
576 Estimate profit from a targeted mailing campaign based on SVM model
577 ```{r}
578 #only send mailer to those predicted to be positive
579 Targeted_Total_Cost = (TP*cost_yes_purchase)+(FP*cost_no_purchase)
580 Targeted_Total_Revenue = (TP*revenue_per_purchase)
581 Targeted_Profit = Targeted_Total_Revenue - Targeted_Total_Cost
582 Targeted_Mailers = TP + FP
583 Target_Profit_per_Mailer = Targeted_Profit / Targeted_Mailers
584
585 Targeted_Total_Cost
586 Targeted_Total_Revenue
587 Targeted_Profit
588 Targeted_Mailers
589 Target_Profit_per_Mailer
590 ```
```

The analysis for this case study indicates P_Art and gender were the most influential variables in predicting whether a purchase of *The Art History of Florence* was made by a customer. P_Art, First_purchase, and Amount_purchase positively impact the odds of a purchase being made, while being male, and increased frequency or purchases of cookbooks, DIY, children and youth books decreases the odds. It has also been applied to estimate the profitability of a targeted campaign. BBBC can automate the SVM and apply it to their direct marketing campaigns to improve its efficacy. Further tuning to the model could be explored, as well as including attributes from other forms of marketing such as email and phone calls, who have purchased a book from BBBC in the past.

SOURCES

Bassey, Patricia – *Logistic Regression Vs. Support Vector Machines (SVM)*, Axum Labs, 2019, https://medium.com/axum-labs/logistic-regression-vs-support-vector-machines-svm-c335610a3d16

Lee, Yoonkyung – *Support Vector Machines for Classification: a Statistical Portrait*. https://www.asc.ohio-state.edu/lee.2272/mss/svm.mimb.rev3.pdf

Nguyen, Phu – *An Application of Support Vector Machines to Customer Loyalty Classification of Korean Retail Company*, Journal of Information Systems, 2017, https://www.researchgate.net/publication/322266252_An_Application_of_Support_Vector_Machines_to_Customer_Loyalty_Classification_of_Korean_Retailing_Company_Using_R_Language_1

APPENDIX A: EXPLORATORY DATA ANALYSIS

Figure 1. *Bookbinders Categorical Predictors*

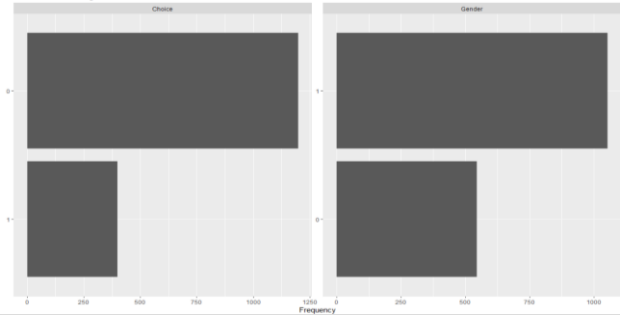


Figure 2. *Gender by Choice*

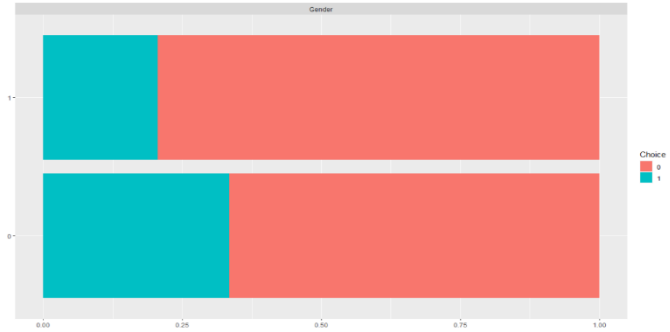


Figure 3. *Bookbinders Numeric Predictors*

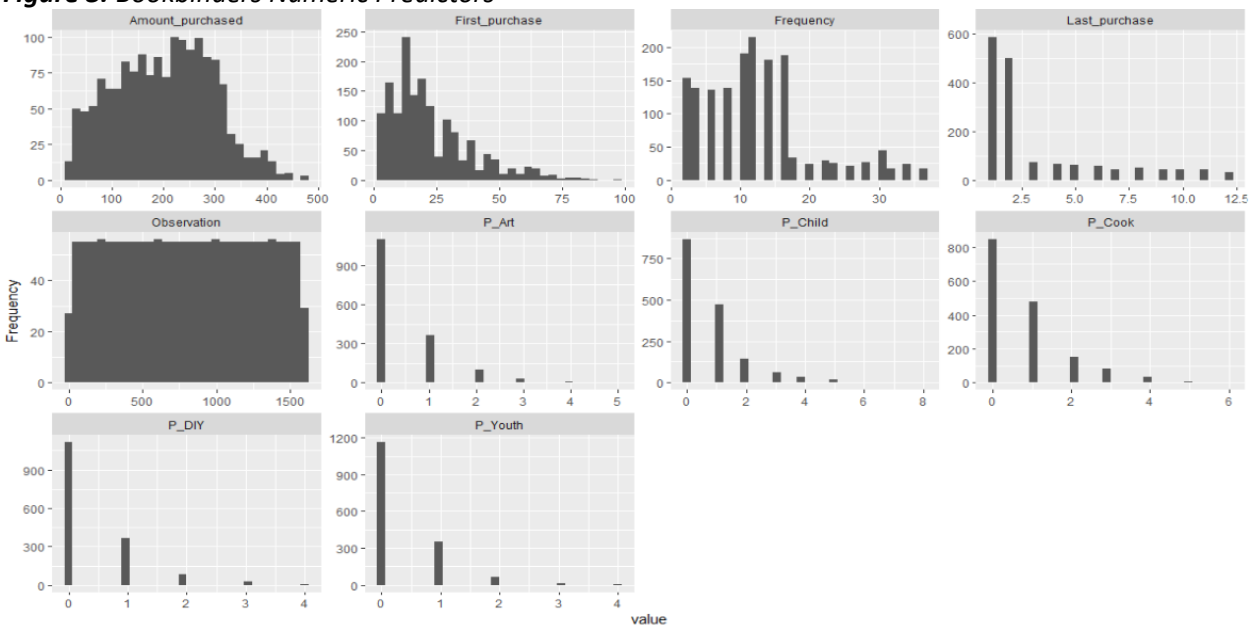


Figure 4. *Correlation Plot*

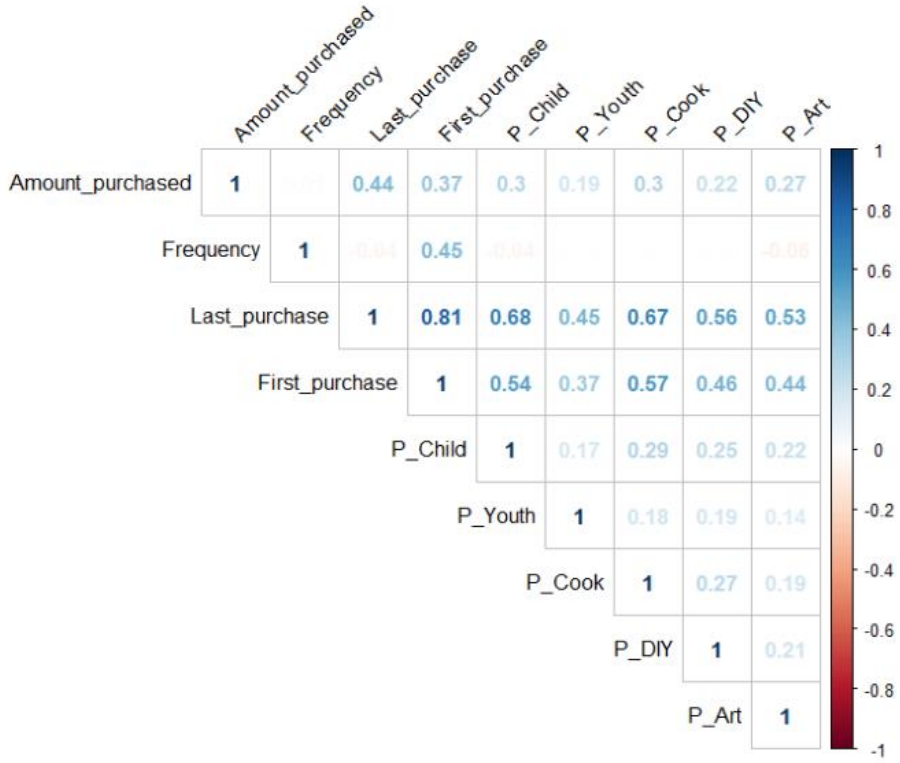
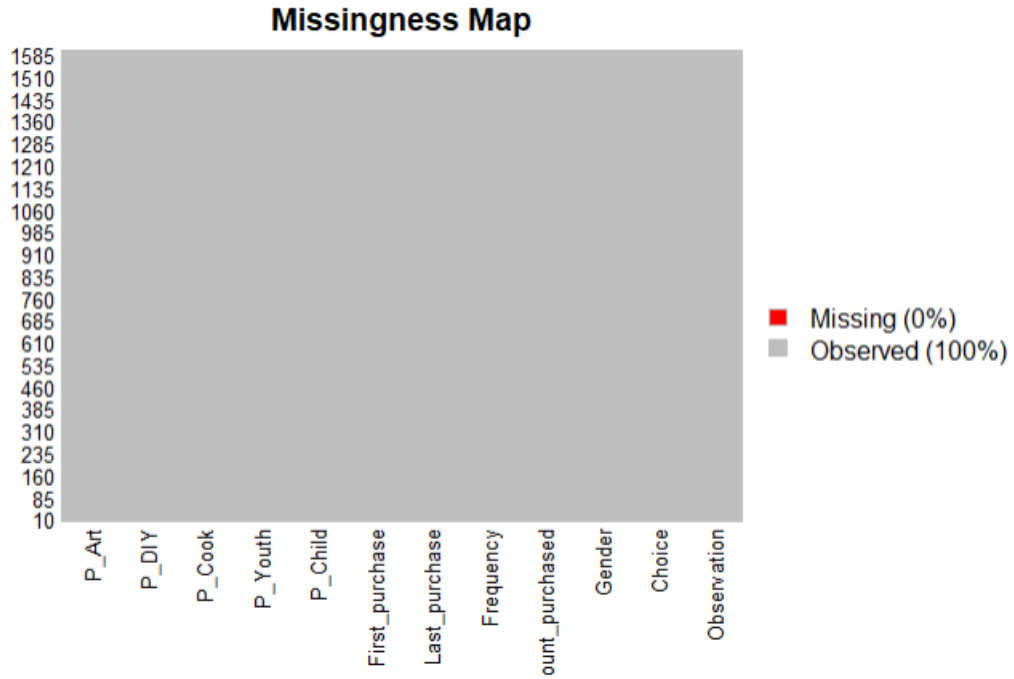


Figure 5. *Plot Missing Values*



APPENDIX B: MODEL RESULTS

NAME	MODEL	SIGNIFICANT PREDICTORS	METRICS	FINDINGS	FINAL MODEL?
linear1	linear = lm(Choice~., data=train_linear)	All except First_purchase	R2 = .2401 Adj R2 = .2353 MSE = .0926	The full regression model has a low MSE but only 24%of the variance in Choice is predictable by the model. Last_purchase has an 81% correlation to First_purchase and has a high VIF of 18.77, so fit a model without it.	**N**
linear2	linear2 = lm(Choice~. - First_purchase, data=train_linear)	All	R2 = .2395, Adj R2 = .2352 **MSE = .0924**	Dropping First_purchase (because it was an insignificant predictor in linear1) led to a .002 improvement to MSE. Once again, Last_purchase has a high VIF (13.920175). Because linear2 had the lowest MSE, it was selected as the final linear regression model.	**Y**
linear3	linear3 = lm(Choice~. - Last_purchase, data=train_linear)	All except P_Youth	R2 = .2156 Adj R2 = .2111, MSE = .0929	Dropping Last_purchase (because of its high VIF) negatively impacted results, with a lower R2 and .0005 increase to MSE. First_purchase had a high VIF (7.18).	**N**
linear4	linear4 = lm(Choice~. - Last_purchase - P_Youth, data=train_linear)	All	R2 = .2148 Adj R2 = .2108 MSE = .0930	Dropping Last_Purchase and the next least significant predictor, P_Youth, resulted in the highest MSE (.0930).	**N**
logit1	logit1 = glm(Choice ~., data = train, family = "binomial")	All except First_purchase	AIC = 1414.159 BIC = 1473.315 Accuracy = .89 Sensitivity = .9389 Specificity = .3873 AUC = .8	Baseline logistic model	**N**
logit2	logit2 = glm(Choice ~.- First_purchase, data = train, family = "binomial")	All	AIC = 1413.496 BIC = 1467.273 Accuracy = .8913 Sensitivity = .9413 Specificity = .3775 AUC = .801	Fitting the logistic regression model after dropping the insignificant predictor, First_purchase, led to some improvement. It resulted in a lower AIC, BIC, and a higher accuracy and sensitivity rate, although specificity is slightly lower. There is a small positive impact to AUC which increased by .001.	**N**
logit3	logit3 = glm(Choice ~.- Last_purchase, data = train, family = "binomial")	All except P_Youth	AIC = 1456.978 BIC = 1510.756 **Accuracy = .8957** Sensitivity = .9480 Specificity = .3578 **AUC = .796**	Fitting the logistic regression model after dropping Last_purchase because it had a high VIF in logit1, resulted in a higher AIC and BIC but had the highest accuracy so far out of the logistic models. Sensivity is also the highest out of the models thus far, but specificity and AUC have decreased and are the lowest out of the logistic models so far.	**Y**
logit4	logit4 = glm(Choice ~.- Last_purchase - P_Youth, data = train, family = "binomial")	All	AIC = 1457.149 BIC = 1505.549 Accuracy = .8957 Sensitivity = .9485 Specificity = .3529 AUC = .795	Removing the insignificant predictor, P_Youth from logit3 did not have much of a payoff. Accuracy remained the same (.8975) but AUC decreased by .001. Since logit3 has the higher accuracy rate and AUC, it will be selected as the final logistic regression model.	**N**

NAME	MODEL	SIGNIFICANT PREDICTORS	METRICS	FINDINGS	FINAL MODEL?
SVM1	form1 = Choice ~ . set.seed(2021) tuned = tune.svm(form1, data=train, kernel = "radial", gamma = seq(.01, .1, by = .01), cost = seq(.1, 1, by = .1)) svm1 = svm(form1, data=train, gamma = tuned$best.parameters$gamma, cost = tuned$best.parameters$cost)	N/A	**Accuracy = .9096** Sensitivity = .9800 Specificity = .1863	The optimal parameters were .05 for gamma and .7 for cost. Using the default kernel, there were 785 support vectors.	**Y**
SVM2	form2 = Choice ~ . - Last_purchase set.seed(2021) tuned2 = tune.svm(form2, data=train, kernel ="radial", gamma = seq(.01, .1, by = .01), cost = seq(.1, 1, by = .1)) svm2 = svm(form2, data=train, gamma = tuned2$best.parameters$gamma, cost = tuned2best.parameters$cost)	N/A	Accuracy = .9074 Sensitivity = .9819 Specificity = .1422	Dropping Last_purchase (because it exhibited a high VIF in prior models) negatively impacted model accuracy, which decreased by .0022. The optimal parameters were .01 for gamma and 1 for cost. There were 782 support vectors.	**N**
SVM3	form3 = Choice ~ . - First_purchase set.seed(2021) tuned3 = tune.svm(form3, data=train, kernel = "radial", gamma = seq(.01, .1, by = .01), cost = seq(.1, 1, by = .1)) svm3 = svm(form3, data=train, gamma = tuned3$best.parameters$gamma, cost = tuned3$best.parameters$cost)	N/A	Accuracy = .9065 Sensitivity = .9709 Specificity = .2451	First_purchase was dropped because of its high p-value in prior models and high VIF. The optimal parameters were .1 for gamma and .8 for cost. There were 804 support vectors.	**N**
SVM4	form4 = form1. set.seed(2021) tuned4 = tune.svm(form4, data=train, kernel = "linear", gamma = seq(.01, .1, by = .01), cost = seq(.1, 1, by = .1)) svm4 = svm(form4, data=train, gamma = tuned4$best.parameters$gamma, cost = tuned4$best.parameters$cost)	N/A	Accuracy = .8991 Sensitivity = .9594 Specificity = .2794	Using same formula as SVM1 but using a linear kernel instead. The optimal parameters were .05 for gamma and .7 for cost. There were 739 support vectors.	**N**

APPENDIX C: LOGISTIC REGRESSION PLOTS

Figure 1. *Linearity of Predictors to the Response Variable, 'Choice'*

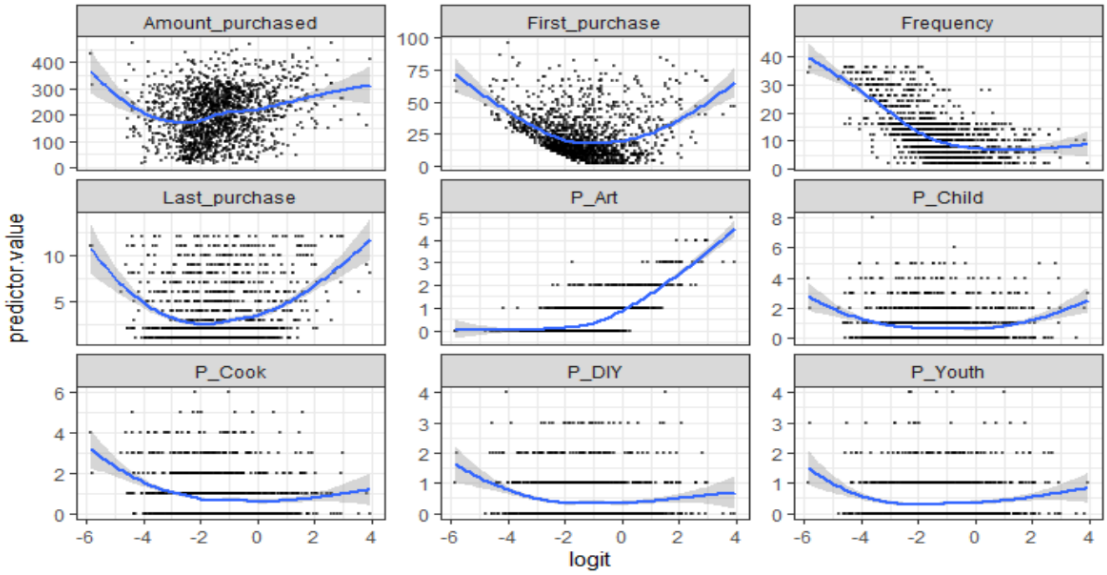


Figure 2. *Cook's Distance, Highlighting 2 most Influential Observations*

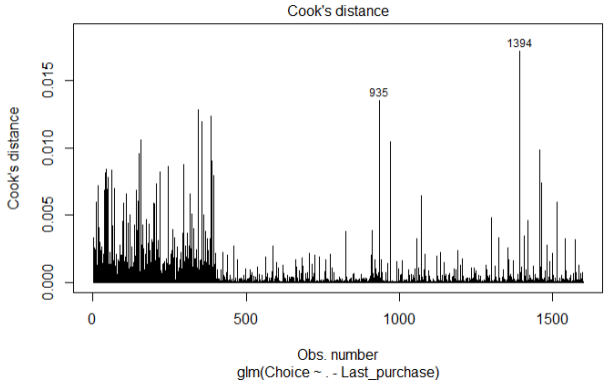


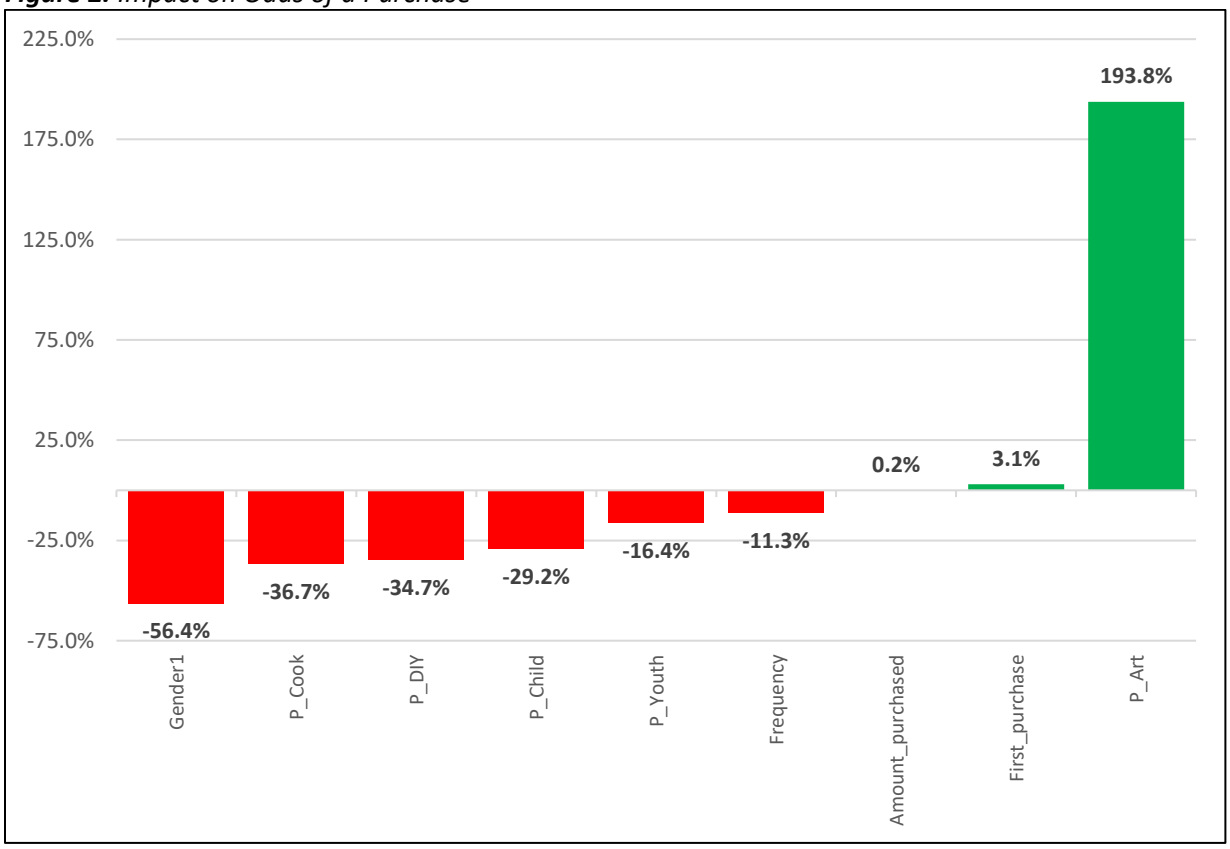
Figure 3*. Standardized Residuals Plot*

APPENDIX D: INTERPRETATION OF LOGISTIC REGRESSION ODDS RATIOS

Table 1. *Interpretation of Odds Ratios*

VARIABLE	ODDS RATIO	INTERPRETATION
Gender1	**0.436**	The odds of a purchase are 100(.436 - 1) = **56.4%** lower for males than females.
P_Cook	**0.633**	For each additional cook book purchased, the odds of purchase decrease by 100(.633-1) = **36.7%**.
P_DIY	**0.653**	For each additional DIY book purchased, the odds of purchase decrease by 100(.653-1) = **34.7%.**
P_Child	**0.708**	For each additional children's book purchased, the odds of purchase decrease by 100(.708-1) = **29.2%**.
P_Youth	**0.836**	For each additional youth book purchased, the odds of purchase decrease by 100(.836-1) = **16.4%**.
Frequency	**0.887**	Each additional purchase in the chosen period leads to a 100(.887-1) = **11.3%** decrease in the odds of purchasing.
Amount_purchased	**1.002**	Each increase in Amount_purchased leads to a 100(1.002-1) = **.2%** increase in the odds of a purchase.
First_purchase	**1.031**	For each additional month since the first purchase was made, the odds of purchasing increase by 100(1.031-1) = **3.1%**.
P_Art	**2.938**	For each additional art book purchased, the odds of purchase increase by 100(2.938-1) = **194%**.

Figure 2. *Impact on Odds of a Purchase*



APPENDIX E: R CODE