Connor Allison
Joel Fry
Nicholas Mata
Royalpriesthood Olola
DA-6813

Bookbinders Case Study

Executive Summary

The Bookbinders Book Club (BBBC) seeks to optimize its direct mail marketing strategy by implementing predictive modeling techniques. Our analysis compared three models: logistic regression, support vector machines (SVM), and linear regression. We found that linear regression was inappropriate due to the categorical nature of the response variable. Logistic regression yielded the highest sensitivity in identifying likely buyers, while the SVM model with a radial kernel achieved the highest accuracy. However, the sigmoid kernel SVM produced more profit than the radial kernel, though still less than logistic regression. The findings suggest that leveraging predictive analytics can significantly enhance BBBC's marketing efficiency, particularly in targeting high-response customers, potentially increasing profitability for future campaigns.

Problem

BBBC needs to refine its direct mail marketing strategy to improve customer response rates and maximize profitability. The challenge lies in identifying key customer characteristics that influence book purchases and developing a robust predictive model to guide future marketing efforts. Theoretically, the problem aligns with classification modeling in machine learning, where categorical response prediction is crucial for targeted marketing. The objective of this study is to evaluate logistic regression, support vector machines, and linear regression to determine the most effective model for customer targeting.

Literature Review

Predictive modeling in marketing has evolved significantly, particularly in customer segmentation and direct mail targeting (Cui & Curry, 2005). Logistic regression has been widely used due to its interpretability and effectiveness in binary classification tasks (Tranmer & Elliot, 2008). Support vector machines, particularly with non-linear kernels, have gained traction for their superior predictive power in complex datasets (Cui & Curry, 2005). Past research suggests that multicollinearity, data sparsity, and variable selection impact the effectiveness of predictive models. Furthermore, the rise of machine learning techniques has enabled businesses to move beyond traditional heuristics toward data-driven decision-making (Chan et al., 2022).

Methodology

To evaluate the effectiveness of predictive modeling for BBBC's direct mail strategy, we plan to test three different models: logistic regression, support vector machines (SVM), and linear regression. Each model will be assessed for its ability to predict customer purchases and its impact on company profits. Logistic Regression will be used to identify key predictors of customer purchases. We aim to find an optimal cutoff threshold that maximizes profitability by balancing true positives and false positives. Support Vector Machines with their best parameters will be used to determine if non-linear classification improves prediction accuracy and profitability. Although initially included, we anticipate that linear regression will be unsuitable due to the categorical nature of the response variable.

To compare the models, we will calculate their impact on BBBC's profits using the formula 9.55tp - 0.65fp, where each true positive generates $9.55 in revenue, while each false positive incurs a $0.65 cost. This formula is the end result of $0.65 for promotion mailing, $15 for ordering and sending a book to a customer, the 45% overhead per book, and $31.95 selling price of the book. We plan to use this metric to determine the best model for future campaigns and to ensure that our recommendations align with BBBC's financial objectives.

Data

As stated previously, the data used oversampled Choice being 1 relative to what BBBC observed in practice. We decided not to undersample the negatives, working with the full data

provided. We removed variables highly correlated with each other to prevent multicollinearity, and we made sure our models were free of multicollinearity using VIF. First_purchase and Last_purchase were dropped due to redundancy.

```{r}
bbbctrain ▷
  select(-First_purchase, -Last_purchase) ▷
  cor()
```

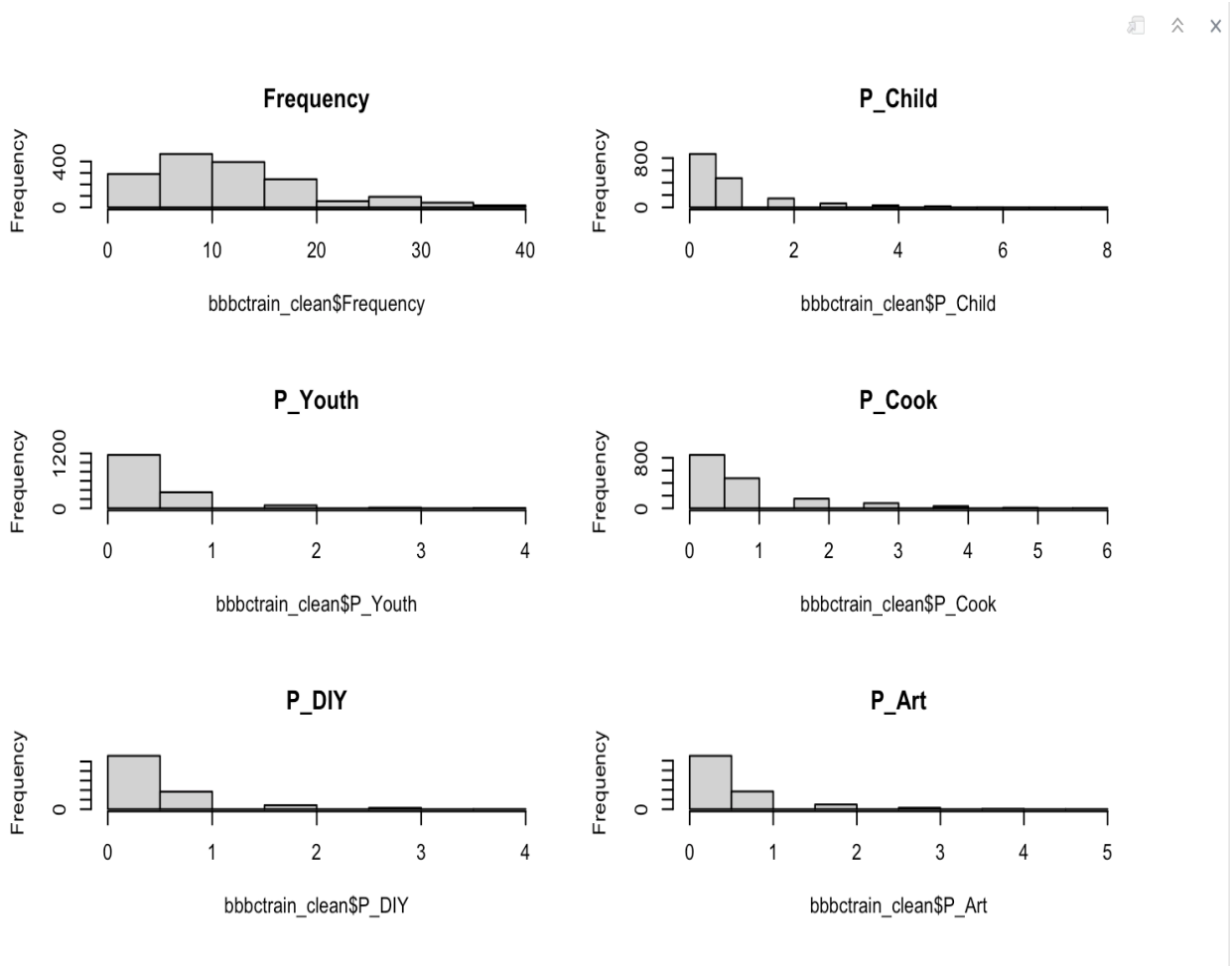
```
                     Choice       Gender Amount_purchased     Frequency       P_Child
Choice            1.000000000 -0.141558415       0.11815256 -0.2260181193  0.008523377
Gender           -0.141558415  1.000000000      -0.03060700  0.0321704951 -0.041475936
Amount_purchased  0.118152563 -0.030607000       1.00000000  0.0136664846  0.299313719
Frequency        -0.226018119  0.032170495       0.01366648  1.0000000000 -0.043327944
P_Child           0.008523377 -0.041475936       0.29931372 -0.0433279437  1.000000000
P_Youth           0.027608101 -0.014130306       0.18755727 -0.0095854745  0.174826719
P_Cook           -0.040256351 -0.026673876       0.30425340  0.0004968833  0.294706519
P_DIY            -0.005309265 -0.025946174       0.22331539 -0.0089634125  0.253837077
P_Art             0.357688817 -0.003500037       0.27248948 -0.0613754066  0.224512850
                     P_Youth        P_Cook        P_DIY        P_Art
Choice            0.027608101 -0.0402563507 -0.005309265  0.357688817
Gender           -0.014130306 -0.0266738763 -0.025946174 -0.003500037
Amount_purchased  0.187557270  0.3042533969  0.223315392  0.272489483
Frequency        -0.009585474  0.0004968833 -0.008963412 -0.061375407
P_Child           0.174826719  0.2947065185  0.253837077  0.224512850
P_Youth           1.000000000  0.1816566401  0.188683456  0.141751220
P_Cook            0.181656640  1.0000000000  0.271725126  0.191680761
P_DIY             0.188683456  0.2717251256  1.000000000  0.207791065
P_Art             0.141751220  0.1916807611  0.207791065  1.000000000
```

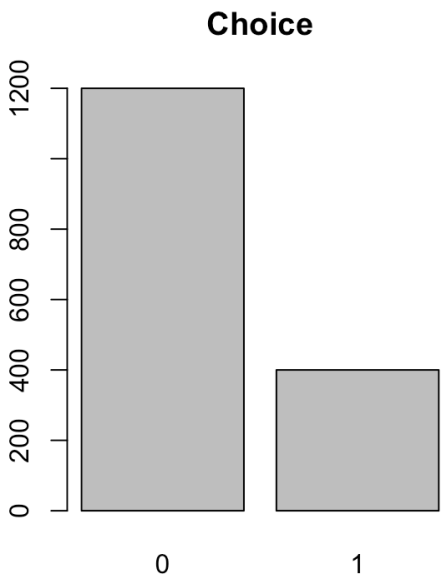
Factor variables such as Choice and Gender were encoded appropriately for classification models, effectively rendering a linear model useless because of the categorical response variable. Missing values were checked, and the dataset was found to be complete.

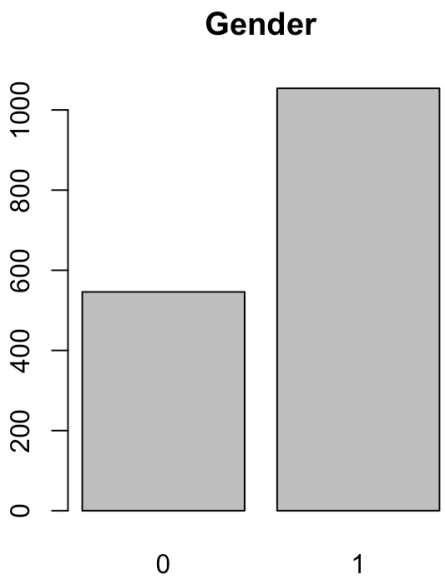
```{r}
str(bbbctrain_clean)
```

```
tibble [1,600 × 9] (S3: tbl_df/tbl/data.frame)
 $ Choice          : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
 $ Gender          : Factor w/ 2 levels "0","1": 2 2 2 2 1 2 2 1 2 2 ...
 $ Amount_purchased: num [1:1600] 113 418 336 180 320 268 198 280 393 138 ...
 $ Frequency       : num [1:1600] 8 6 18 16 2 4 2 6 12 10 ...
 $ P_Child         : num [1:1600] 0 0 2 2 0 0 2 0 3 2 ...
 $ P_Youth         : num [1:1600] 1 2 0 0 0 0 3 2 0 3 ...
 $ P_Cook          : num [1:1600] 0 3 1 0 0 0 2 0 3 0 ...
 $ P_DIY           : num [1:1600] 0 2 1 1 1 0 1 0 0 0 ...
 $ P_Art           : num [1:1600] 0 3 2 1 2 0 2 0 2 1 ...
```

All variables except Choice and Gender were found to be right skewed, and there were almost twice as many males as females.







Results

Summarizing Model Performance: Based on the costs and selling prices given by BBBC, we estimated profits with the formula 9.55tp - 0.65fp, where every true positive yields $9.55 and every false positive costs the company $0.65. The logistic regression model with an optimized cutoff threshold of 0.204318 yielded the highest profit of $1074.56, representing an over 80% increase from the $585.80 profit when mailing to the entire list.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1369   42
         1  727  162

               Accuracy : 0.6657
                 95% CI : (0.646, 0.6849)
    No Information Rate : 0.9113
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1778
```

```
Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.79412
            Specificity : 0.65315
         Pos Pred Value : 0.18223
         Neg Pred Value : 0.97023
             Prevalence : 0.08870
         Detection Rate : 0.07043
   Detection Prevalence : 0.38652
      Balanced Accuracy : 0.72363

       'Positive' Class : 1
```

The cutoff threshold was manually tweaked to find the best ratio of true positives to true negatives to generate maximum profit, but ultimately the computed optimal cutoff was the best.

threshold	tp	fp	profit
0.204	162	729	1073.25
0	204	2096	585.8
0.5	73	110	625.65
0.4	96	209	780.95
0.3	120	385	895.75
0.2	162	749	1060.25
0.1	190	1417	893.45
0.25	138	530	973.4
0.21	156	696	1037.4
0.205	160	726	1056.1
0.203	162	734	1070
0.2043	162	728	1073.9
0.204318	162	727	1074.55
radial svm	37	40	327.35
sigmoid svm	76	144	632.2

The radial kernel SVM model had the highest accuracy but produced a lower profit of $327.35, making it less effective than the mass mailing method.

```
Call:
svm(formula = form_svm, data = bbbctrain_clean, gamma = tuned$best.parameters$gamma,
    cost = tuned$best.parameters$cost)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  0.9
```

```
    true
pred    0    1
   0 2056  167
   1   40   37
```

The sigmoid kernel SVM yielded $632.20, which was an improvement in profitability over the radial kernel but still significantly lower than the logistic regression model.

```
Call:
svm(formula = form_svm, data = bbbctrain_clean, kernel = "sigmoid", gamma =
tuned$best.parameters$gamma,
    cost = tuned$best.parameters$cost)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  sigmoid
       cost:  0.9
     coef.0:  0
```

```
    true
pred    0    1
   0 1952  128
   1  144   76
```

In summary, logistic regression had high interpretability and strong sensitivity for identifying likely buyers. It was also the most profitable model. Radial SVM had the highest accuracy but lowest profitability, making it a suboptimal choice for business decision-making. Sigmoid SVM had moderate sensitivity and accuracy, producing higher profit than the radial kernel SVM but significantly less than logistic regression. Linear Regression was inapplicable due to the categorical response variable.

The significance of the predictor variable P_Art suggests that BBBC should target individuals who purchase more books. Conversely, the Frequency variable is negatively correlated with purchases, indicating that individuals who have made multiple recent purchases are less likely to respond to additional offers. In summary, targeting those who purchase more

books over longer periods of time will maximize profits.

```
Call:
glm(formula = Choice ~ Gender + Amount_purchased + Frequency +
    P_Child + P_Cook + P_DIY + P_Art, family = binomial, data = bbbctrain_clean)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.2894506  0.2026211  -1.429  0.15314
Gender1         -0.8120440  0.1345723  -6.034  1.6e-09 ***
Amount_purchased  0.0023859  0.0007678   3.108  0.00189 **
Frequency       -0.0885491  0.0103772  -8.533  < 2e-16 ***
P_Child         -0.1964495  0.0720116  -2.728  0.00637 **
P_Cook          -0.2940503  0.0727520  -4.042  5.3e-05 ***
P_DIY           -0.2823065  0.1076089  -2.623  0.00870 **
P_Art            1.2441330  0.0988603  12.585  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusion

BBBC's planned mail campaign in the Midwest presents an opportunity to leverage the logistic model for customer selection. Given the model's ability to increase profit, BBBC should target individuals with a high likelihood of purchase based on significant predictor variables, particularly P_Art, while avoiding high-frequency purchasers. Applying the logistic model in this campaign can significantly increase profitability compared to mass mailing.

Given these insights, BBBC should invest in logistic regression for its direct mail campaigns, as it maximizes profitability and provides clear insights into customer purchasing behavior. To streamline future modeling efforts, BBBC should automate the selection of optimal cutoff points and model training. Implementing machine learning pipelines can enhance efficiency, and continuous model validation should be integrated to ensure the approach remains effective over time. Additionally, exploring additional methods, such as random forests, may provide further improvements in predictive performance.

Works Cited

Chan, J. Y. L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z. W., & Chen, Y. L.

(2022). Mitigating the multicollinearity problem and its machine learning approach: a review.

Mathematics, *10*(8), 1283.

Cui, D., & Curry, D. (2005). Prediction in marketing using the support vector machine. *Marketing Science*,

24(4), 595-615.

Tranmer, M., & Elliot, M. (2008). Binary logistic regression. *Cathie Marsh for census and survey research,*

paper, *20*, 90033-9.

Appendix

Read in train and test sets and drop `Observation`

```{r}
bbbctrain = read_excel('BBBC-Train.xlsx')
bbbctrain = bbbctrain[,-1]
```

```{r}
bbbctest = read_excel('BBBC-Test.xlsx')
bbbctest = bbbctest[,-1]
```

Checking for missing variables

```{r}
anyNA(bbbctest)
anyNA(bbbctrain)
```

```
[1] FALSE
[1] FALSE
```

Checking for correlated variables

```{r}
cor(bbbctrain)
```

Drop `First_purchase` because it was highly correlated with `Last_purchase` and check for correlation again

```{r}
bbbctrain ▷
  select(-First_purchase) ▷
  cor()
```

Drop `Last_purchase` because of its moderate-to-high correlation with four other variables `P_Child`, `P_Cook`, `P_DIY`, `P_Art`.

Checking for correlation among remaining variables.

```{r}
bbbctrain ▷
  select(-First_purchase, -Last_purchase) ▷
  cor()
```

Correlation looks good across the remaining variables: `Choice`, `Gender`, `Amount_purchased`, `Frequency`, `P_Child`, `P_Youth`, `P_Cook`, `P_DIY`, `P_Art`

Initializing new objects for clean data.

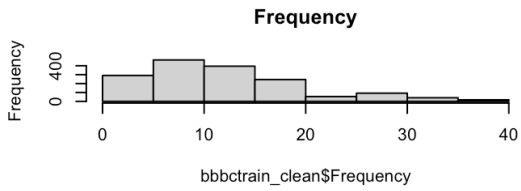
```{r}
bbbctrain_clean = bbbctrain |>
  select(-First_purchase, -Last_purchase)
```

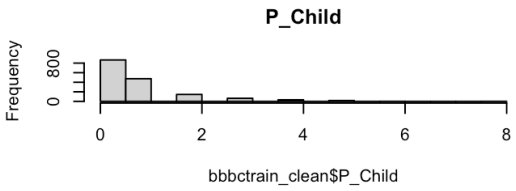
```{r}
bbbctest_clean = bbbctest |>
  select(-First_purchase, -Last_purchase)
```

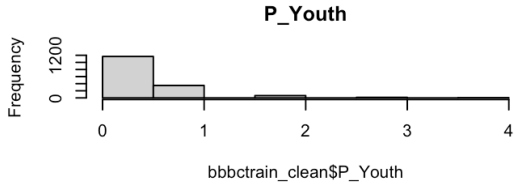
Exploring data and plots

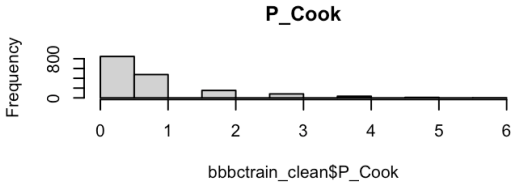
```{r}
str(bbbctrain_clean)
```

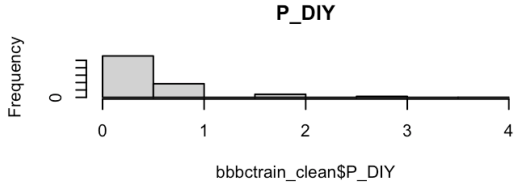
```{r}
par(mfrow = c(3 ,2))
hist(bbbctrain_clean$Frequency, main = "Frequency")
hist(bbbctrain_clean$P_Child, main = "P_Child")
hist(bbbctrain_clean$P_Youth, main = "P_Youth")
hist(bbbctrain_clean$P_Cook, main = "P_Cook")
hist(bbbctrain_clean$P_DIY, main = "P_DIY")
hist(bbbctrain_clean$P_Art, main = "P_Art")
```

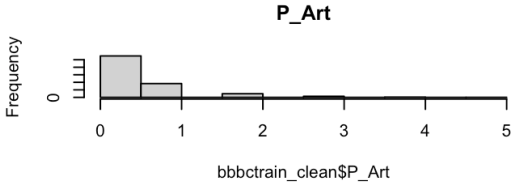




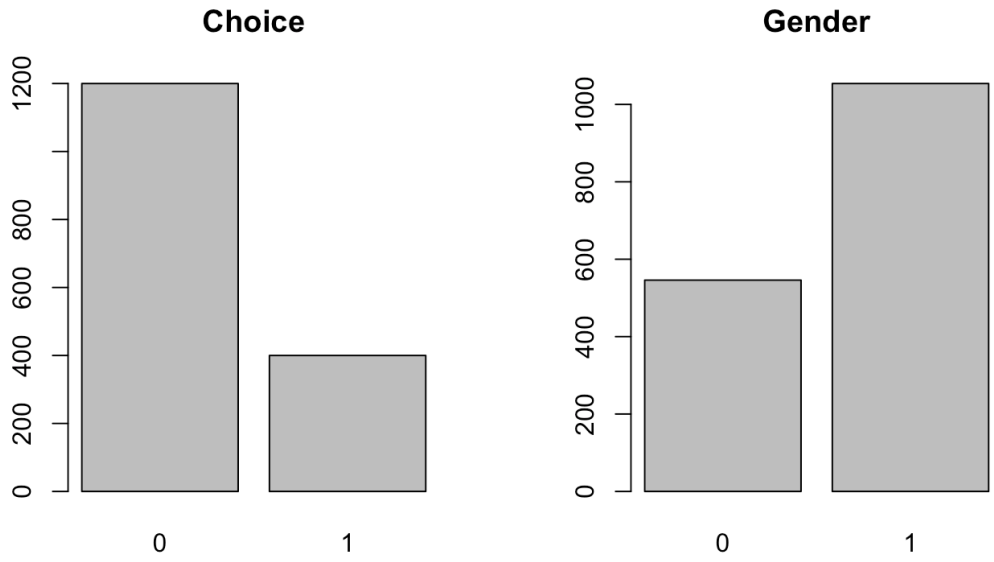








```r
par(mfrow = c(1,2))
plot(bbbctrain_clean$Choice, main = 'Choice')
plot(bbbctrain_clean$Gender, main = 'Gender')
```



Building the linear model with our cleaned training data.

```{r}
lm1 = lm(form1, data = bbbctrain_clean)
summary(lm1)
```

```
Call:
lm(formula = form1, data = bbbctrain_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9501 -0.2518 -0.1273  0.1509  1.1211

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.3731865  0.0302933  12.319  < 2e-16 ***
Gender           -0.1263728  0.0203683  -6.204 6.99e-10 ***
Amount_purchased  0.0003688  0.0001123   3.283  0.00105 **
Frequency        -0.0112345  0.0012344  -9.101  < 2e-16 ***
P_Child          -0.0275983  0.0100284  -2.752  0.00599 **
P_Youth          -0.0014841  0.0159946  -0.093  0.92609
P_Cook           -0.0428346  0.0102155  -4.193 2.90e-05 ***
P_DIY            -0.0384262  0.0153017  -2.511  0.01213 *
P_Art             0.2183323  0.0140081  15.586  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3856 on 1591 degrees of freedom
Multiple R-squared:  0.2114,    Adjusted R-squared:  0.2075
F-statistic: 53.32 on 8 and 1591 DF,  p-value: < 2.2e-16
```

We need 2 variables, including the response, to be factors, so a linear model would not be appropriate for this data set.

Getting data ready for a logistic regression model.

set factor variables for logistic

```{r}
bbbctrain_clean$Choice = as.factor(bbbctrain_clean$Choice)
bbbctest_clean$Choice = as.factor(bbbctest$Choice)

bbbctrain_clean$Gender = as.factor(bbbctrain_clean$Gender)
bbbctest_clean$Gender = as.factor(bbbctest$Gender)
```

```{r}
str(bbbctrain_clean)
```

```
tibble [1,600 × 9] (S3: tbl_df/tbl/data.frame)
 $ Choice          : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
 $ Gender          : Factor w/ 2 levels "0","1": 2 2 2 2 1 2 2 1 2 2 ...
 $ Amount_purchased: num [1:1600] 113 418 336 180 320 268 198 280 393 138 ...
 $ Frequency       : num [1:1600] 8 6 18 16 2 4 2 6 12 10 ...
 $ P_Child         : num [1:1600] 0 0 2 2 0 0 2 0 3 2 ...
 $ P_Youth         : num [1:1600] 1 2 0 0 0 0 3 2 0 3 ...
 $ P_Cook          : num [1:1600] 0 3 1 0 0 0 2 0 3 0 ...
 $ P_DIY           : num [1:1600] 0 2 1 1 1 0 1 0 0 0 ...
 $ P_Art           : num [1:1600] 0 3 2 1 2 0 2 0 2 1 ...
```

building logistic model (glm)

```{r}
glm1 = glm(form1, data = bbbctrain_clean, family = binomial)
summary(glm1)
```

Checking for multicolinearity using `vif()`

```{r}
vif(glm1)
```

```
      Gender Amount_purchased        Frequency          P_Child          P_Youth 
    1.020217         1.213528         1.015899         1.215500         1.081019 
      P_Cook            P_DIY            P_Art 
    1.228798         1.179821         1.229491 
```

No multicolinearity present, so we can look to improve the model. We will use stepwise selection to find best variables. After finding best variables, multicolinnearity will be checked again.

```{r}
glm1_step = step(glm1, direction = "both")
summary(glm1_step)
vif(glm1_step)
```

Finding the most optimal cutoff point.

```{r}
proc_log = roc(bbbctest_clean$Choice, bbbctest_clean$PredProb)
```

```{r}
plot(proc_log)
```

```{r}
coords(proc_log, "best", ret = "threshold")
```

Next we want to fit a support vector machine and to make predictions on our response variable like we did with the logistic regression.

radial kernel support vector machine

```{r}
form_svm = Choice ~ .
```

```{r}
tuned = tune.svm(form_svm, data = bbbctrain_clean, gamma = seq(.01, .1, by = .01), cost = seq(.1, 1, by = .1))
```

```{r}
mysvm = svm(formula = form_svm, data = bbbctrain_clean, gamma = tuned$best.parameters$gamma, cost = tuned$best.parameters$cost)
summary(mysvm)
```

Trying a sigmoid kernel svm

```{r}
mysvm = svm(formula = form_svm, data = bbbctrain_clean, kernel = 'sigmoid', gamma = tuned$best.parameters$gamma, cost = tuned$best.parameters$cost)
summary(mysvm)
```