Bookbinders Case Study

Eva Beyebach, Pablo Chacon, & Danya Saed

I. Executive Summary

This case study explores the strategic and operational challenges Bookbinders Book Club (BBBC) faces amid the shifting dynamics of book retail and distribution since its establishment in 1986. With the rise of superstores and online retailing altering consumer preferences and intensifying competitive pressures, BBBC has been prompted to innovate its customer engagement and sales strategies. Utilizing a robust database of 500,000 readers, BBBC is advancing into database marketing and predictive modeling to boost the efficiency of its direct marketing techniques.

This case study highlights Bookbinders Book Club's (BBBC) pivotal challenge in selecting the most suitable predictive modeling technique to forecast customer responses to marketing initiatives, crucial for crafting targeted and profitable strategies. Central to this endeavor is the thorough analysis of customer data—demographics, purchase history, and preferences—to pinpoint buying patterns and optimize marketing approaches for enhanced response rates and marketing ROI. BBBC's pursuit of the most effective and cost-efficient method underscores the broader imperative for adaptability and strategic innovation in the book industry's rapid evolution. Through sophisticated predictive modeling, BBBC aims to navigate this dynamic market landscape with greater operational flexibility and strategic acumen, leveraging data analytics to refine its marketing tactics and ensure sustained competitiveness.

II. The Problem

BBBC wants to determine which predictive model is best at improving the efficacy of direct mail programs. Their last model, which consisted of 20,000 customers across Northeastern USA that received a copy of the mailing list and a brochure for *The History of Florence*. Of the 20,000 clients that received the mailing list, 9.03% purchased the book. BBBC wants to maximize the number of clients that buy the book, and is exploring 3 models that may allow it to do so: linear regression, logistic regression, and SVM.

III. Review of Related Literature

In recent years, significant advancements have been made in utilizing data to decipher the factors influencing customer purchasing decisions. For instance, the 2021 study by Belcher et al, as presented on RPubs, employs logistic regression to investigate the determinants behind book purchases, analyzing the purchasing habits of 1,600 customers. This research extends beyond mere numerical analysis; it aims to unravel the narratives underlying these figures—specifically, the motivations leading individuals to prefer one book over another.

Furthermore, the research conducted by Thai et al. in 2023 adopts a distinct approach. Through the application of support vector machines (SVM) and linear regression, Thai et al. examine the potential to forecast book purchasing behavior. This investigation, by considering various perspectives and incorporating diverse data points, illuminates the nuanced methods through which purchasing behaviors can be predicted.

These investigations contribute significantly to the academic discourse, demonstrating the practical application of data to grasp the intricate decision-making processes of consumers. By leveraging methodologies such as logistic regression, SVM, and linear regression, scholars like Belcher et al. and Thai et al. are instrumental in guiding us through the complex terrain of consumer behavior analysis with increased accuracy and insight.

IV. Methodology

Our analysis explores logistic regression, support vector machines (SVM), and linear regression (lm) using R programming, with a focus on classification methods and their underlying assumptions. Logistic regression, a parametric method, assumes a linear relationship between the log odds of the response variable and the independent variables. It is suited for categorical (binary or ordinal) response variables and requires that independent variables do not exhibit multicollinearity. The glm() function in R facilitates model fitting and evaluation through maximum likelihood estimation, highlighting the importance of coefficients and model accuracy via odds ratios.

In contrast, SVMs employ a nonparametric approach, aiming to maximize the margin between classes using a hyperplane. This can be adjusted for non-linear boundaries through kernel functions. The e1071 package in R supports SVM, which does not assume data normality and handles both numerical and recoded categorical variables, focusing on maximizing margin and ensuring data points' independence.

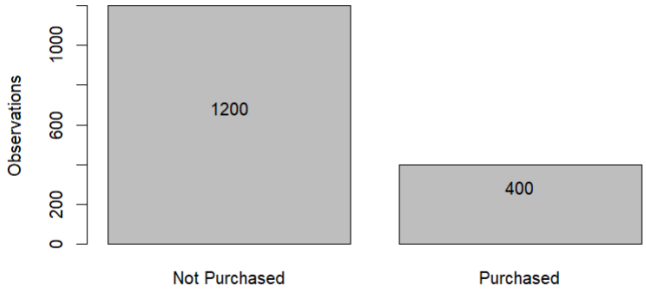
Linear regression, primarily used for modeling continuous outcomes, is not optimal for classification due to its inability to constrain predictions within a 0-1 range, potentially leading to inferior performance. Linear regression is implemented in R through the lm() function, which estimates the relationship between a dependent variable and one or more independent variables. The method provides diagnostic tools, such as the R-squared value and significance tests for coefficients, to assess model fit and the relevance of predictors, offering a clear perspective on variable interactions.

V. Data

The dataset provided has been pre-processed and analyzed for insights into the purchasing patterns of a book titled "The History of Florence". The dataset was already partitioned into training and testing sets, with the training set comprising 1600 observations across 12 variables, and the testing set containing 2300 observations. There

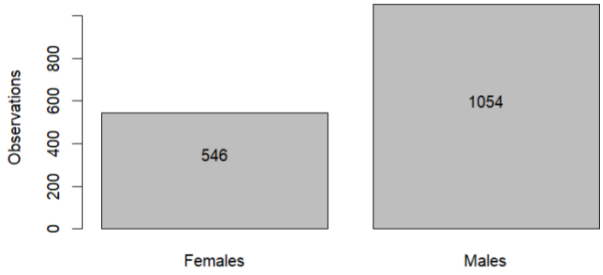
were no missing values or duplicates, and categorical variables like Choice and Gender were converted into factor variables for analysis.

An examination of the data revealed that it was not balanced, with three-quarters of the population not purchasing the book. Additionally, gender distribution was uneven, with two-thirds of the dataset being classified as one gender. No duplicate observations or missing values were found. After performing str(), we had to change Choice and Gender to factor, as both are categorical variables.



From the visualizations, it was noted that there was a much larger number of clients who did not purchase the book compared to those who did. Specifically, 75% of the dataset represented non-purchasers.

The distinction between male and female purchasers was marginal, with males slightly more inclined not to purchase the book.

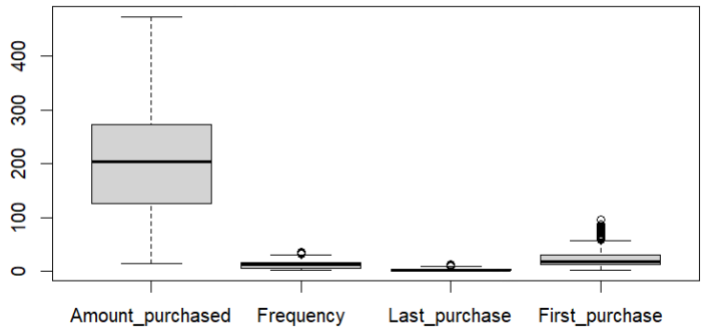


Upon further investigation, it was apparent that there were significantly more males than females in the dataset, which contributed to a larger proportion of males not purchasing the book.

```
[1] "The average purchase volume for purchasers of 'The History of Florence' is: 220.41"
[1] "The max purchase volume for purchasers of 'The History of Florence' is: 474"
[1] "The average purchase volume for purchasers of 'The History of Florence' is: 17"
```

```
[1] "The average purchase volume for non-purchasers of 'The History of Florence' is: 194.42"
[1] "The max purchase volume for non-purchasers of 'The History of Florence' is: 473"
[1] "The average purchase volume for non-purchasers of 'The History of Florence' is: 15"
```

The analysis of purchase volumes revealed the average, minimum, and maximum values for purchasers and non-purchasers of the book. For purchasers, the average purchase volume was reported at 220.41, the maximum at 474, and the minimum at 17. For non-purchasers, the average was 194.42, the maximum at 473, and the minimum at 15.



The analysis of the data revealed outliers in several variables related to purchases, such as 'Frequency', 'Last_purchase', 'First_purchase', and categories of book purchases including 'P_Child', 'P_Youth', 'P_Cook', 'P_DIY', and 'P_Art'. These outliers were identified through boxplot visualizations and were treated by standardizing extreme values to maintain the accuracy of the predictive model. For instance, values of 'Frequency' above 30 were capped at 30, while specific cutoff points were also applied to 'Last_purchase' and 'First_purchase', at 8 and 57 respectively. This approach was adopted across various variables to mitigate the skewing effects of outliers on the model, ensuring a more reliable and accurate predictive analysis.

```
[1] "The average purchase frequency for non-purchasers of 'The History of Florence' is: 1.77"
[1] "The average purchase volume for purchasers of 'The History of Florence' is: 2.62"
```

Finally, a new variable called 'Purchase_frequency' was proposed to provide insights into the frequency of book purchases by both purchasers and non-purchasers. This variable was calculated as the number of weeks between the first and last purchase divided by the frequency of purchases. The analysis showed that purchasers of "The History of Florence" bought books every 2.62 weeks on average, while non-purchasers did so every 1.77 weeks.

Overall, the data provided a detailed look at the purchasing patterns for "The History of Florence" and allowed for the identification of trends and potential areas for further investigation to understand the drivers behind the purchase decisions.

Logistic Regression:

In a comprehensive analysis utilizing logistic regression to predict book purchasing behavior, several important findings and adjustments were made to refine the predictive model. The initial model focused on a range of predictors, including `Gender`,

`Amount_purchased`, and `Frequency` of purchase, as well as specific types of book purchases like `P_Cook` for cookbooks and `P_Art` for art books. The outcome variable, `Choice`, indicated whether a client would purchase a book or not. Notably, `Gender1` displayed a negative association with book purchases, suggesting a gender-based preference in buying behavior, while `Amount_purchased` was positively correlated with the likelihood of a purchase. However, certain predictors like `Frequency` and `P_Cook` were linked to a decreased probability of purchasing, contrasting with `P_Art`, which significantly increased purchase likelihood.

Multicollinearity issues were identified through variance inflation factor (VIF) analysis, particularly between `Last_purchase` and `First_purchase`, leading to the removal of `Last_purchase` from the model to improve predictive accuracy. This adjustment not only addressed multicollinearity concerns but also led to a more interpretable and robust logistic regression model by enhancing model fit as evidenced by changes in the Akaike Information Criterion (AIC).

Further model refinement involved removing `First_purchase` due to its high VIF, resulting in an even better model fit. Despite all variables being significant predictors post-adjustment, `P_Youth` emerged as an insignificant factor, underscoring the nuanced relationship between client characteristics, purchasing behavior, and the likelihood of future book purchases.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1982  128
         1  114   76

               Accuracy : 0.8948
                 95% CI : (0.8815, 0.907)
    No Information Rate : 0.9113
    P-Value [Acc > NIR] : 0.9971

                  Kappa : 0.3283

 Mcnemar's Test P-Value : 0.4033

            Sensitivity : 0.9456
            Specificity : 0.3725
         Pos Pred Value : 0.9393
         Neg Pred Value : 0.4000
             Prevalence : 0.9113
         Detection Rate : 0.8617
   Detection Prevalence : 0.9174
      Balanced Accuracy : 0.6591

       'Positive' Class : 0
```

The model's performance, assessed through a confusion matrix, revealed a high accuracy rate of approximately 89.5%, with a notable sensitivity rate but a lower specificity, highlighting its strength in identifying non-purchasers over purchasers. This discrepancy pointed to the challenge of predicting positive cases in a dataset skewed towards non-purchasers.

Balanced Logistic Regression:

In our pursuit to refine the predictive accuracy of our model, we addressed a significant challenge posed by the class imbalance within our dataset, specifically, the disproportionate number of 'Non-Purchasers' compared to 'Purchasers'. To mitigate the

bias introduced by this imbalance, we utilized a balancing technique that equalized the representation of both purchasers and non-purchasers within the training and testing sets.

This approach was realized through a methodical resampling process, ensuring that both classes were equally represented. This balanced approach was hypothesized to improve model specificity without a substantial detriment to overall accuracy.

Upon application of the balanced logistic regression model, the results presented a distinct shift in performance metrics.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 168  70
##          1  36 134
##
##                Accuracy : 0.7402
##                  95% CI : (0.6948, 0.7821)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4804
##
##  Mcnemar's Test P-Value : 0.001349
##
##             Sensitivity : 0.8235
##             Specificity : 0.6569
##          Pos Pred Value : 0.7059
##          Neg Pred Value : 0.7882
##              Prevalence : 0.5000
##          Detection Rate : 0.4118
##    Detection Prevalence : 0.5833
##       Balanced Accuracy : 0.7402
##
##        'Positive' Class : 0
```

The balanced model achieved an accuracy of 74%, with sensitivity and specificity of 82.4% and 65.7% respectively. These figures contrast with the previous unbalanced model which exhibited an accuracy of 89.5%, and a sensitivity of 94.6%, but a specificity of only 37.3%. It's worth noting that the increase in specificity, from 37.3% to 65.7%, underscores the enhanced ability of the model to correctly identify true negatives, i.e., 'Non-Purchasers'.

While accuracy and sensitivity experienced a decline of 15.5% and 12.2%, respectively, the significant increase in specificity suggests that the balanced model provides a more realistic assessment of purchasing behavior, as it is not overly influenced by the majority class. This improvement is critical, especially when the cost of false positives is high, or when an equitable distribution of prediction accuracy across classes is desired.

In summary, despite a reduction in overall accuracy, the balanced logistic regression model presented a more equitable distribution of predictive performance across both classes. This adjustment signifies a substantial advancement in our analytical capabilities, allowing for more nuanced and equitable decision-making processes. The

balanced model's enhancement in specificity is particularly advantageous in scenarios where the identification of true negatives is as consequential as the identification of true positives.

The second method of statistical modeling will be a linear regression model. Linear regression models aim to predict the value of a continuous dependent variable based on its predictors. Since the `Choice` variable in the data set is binary, this model will undoubtedly run into violations of the model's main assumptions, such as linearity in the dependent variable. However, the linear model should be performed to gauge its effectiveness.

Linear Regression:

In our exploratory data analysis, the scatter plot examining the relationship between the predictor variable 'Amount_purchased' and the binary outcome variable 'Choice' demonstrated an absence of a distinct linear pattern. This initial observation raised concerns about the suitability of a linear regression approach, as it contravenes the fundamental assumption of linearity required for such models.

Despite the preliminary indications of non-linearity, we proceeded to fit a linear regression model to the data. The results, encapsulated in the model summary, reveal several insights. The model includes an intercept and a suite of predictor variables, notably 'Gender1', 'Amount_purchased', and a range of product categories such as 'P_Child', 'P_Cook', 'P_DIY', and 'P_Art'. Within these predictors, variables like 'Amount_purchased' and 'P_Art' emerged as statistically significant with positive coefficients, indicating a positive influence on the likelihood of choosing a study. Conversely, 'Frequency', 'P_Child', 'P_Cook', and 'P_DIY' are inversely related to the choice, as evidenced by their negative coefficients and significant p-values. Notably, the variable 'Gender1' displays a negative relationship with 'Choice', suggesting a lower propensity for the category represented by 'Gender1' to make the targeted choice.

```
##
## Call:
## lm(formula = as.numeric(Choice) ~ . - Observation - Last_purchase -
##     First_purchase, data = BBBC_Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7782 -0.2509 -0.1212  0.1730  1.0847
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.3847880  0.0310625  44.581  < 2e-16 ***
## Gender1          -0.1226545  0.0204768  -5.990 2.59e-09 ***
## Amount_purchased  0.0003319  0.0001120   2.965 0.003077 **
## Frequency        -0.0119805  0.0013038  -9.189  < 2e-16 ***
## P_Child          -0.0322402  0.0139704  -2.308 0.021141 *
## P_Youth          -0.0055608  0.0177301  -0.314 0.753840
## P_Cook           -0.0497884  0.0137135  -3.631 0.000292 ***
## P_DIY            -0.0455746  0.0169938  -2.682 0.007398 **
## P_Art             0.2421983  0.0161149  15.029  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3876 on 1591 degrees of freedom
## Multiple R-squared:  0.2033, Adjusted R-squared:  0.1992
## F-statistic: 50.73 on 8 and 1591 DF,  p-value: < 2.2e-16
```

The model's explanatory power, as measured by the R-squared value, stands at 0.2033, explaining approximately 20.33% of the variance in the 'Choice' variable. The adjusted R-squared, which provides a more accurate measure by adjusting for the number of predictors, is slightly lower at 0.1992. Despite this modest explanatory power, the overall model significance is affirmed by a highly significant F-statistic.

One particular predictor, 'P_Youth', was noted for its lack of significance, indicating that the purchase of youth books does not significantly affect the decision associated with 'The History of Florence'. This could suggest that the influence of purchasing youth books on the choice in question is minimal or overshadowed by other factors in the model.

While the linear regression model has yielded some valuable insights, the binary nature of the dependent variable 'Choice' typically necessitates the application of a logistic regression model. This is substantiated by the comments within the screenshots that suggest the consideration or application of logistic regression, aligning with statistical best practices for binary outcome data.

Even though the Mean Squared Error of 0.15 is very low, the fact that the predictor variable is binary means that linear regression is inherently an inappropriate way to predict Choice, For that reason, linear regression will not be an apt method to derive results and predictions.

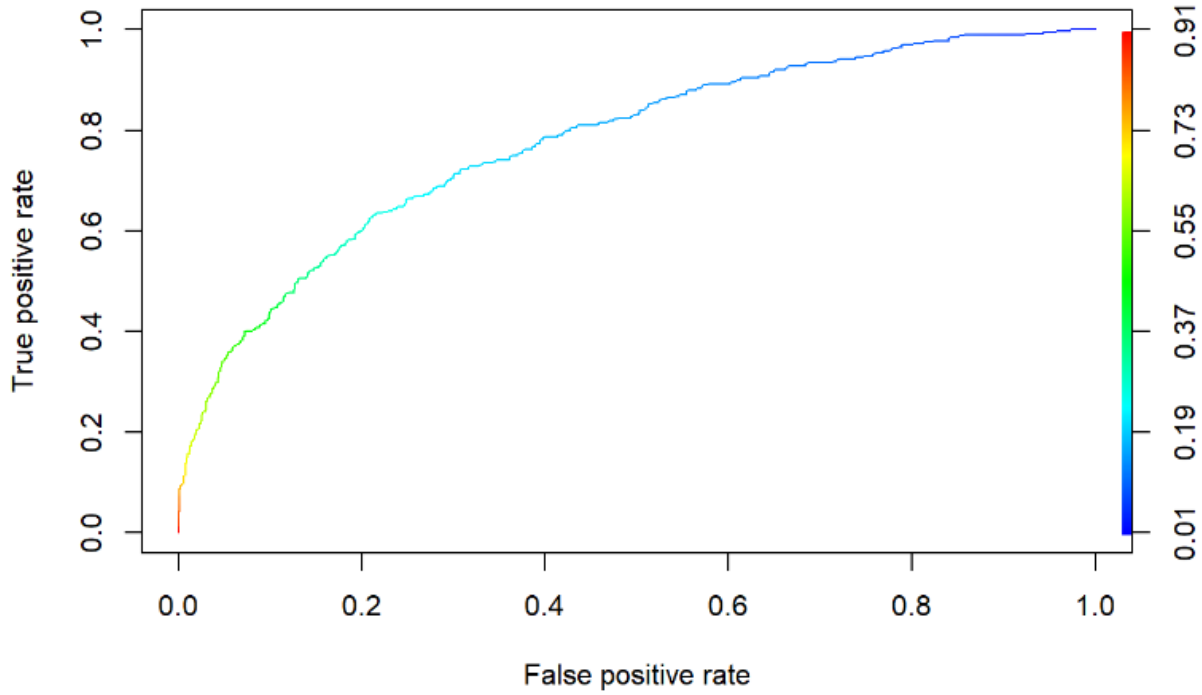
In conclusion, while the linear regression model provides a foundational understanding of the factors influencing customer choice, the analysis underscores the importance of selecting a model congruent with the data structure. Future analyses would benefit from employing logistic regression to better capture the dynamics of binary outcomes.

SVM:

This report outlines the findings from the application of a Support Vector Machine (SVM) model to a binary classification task aimed at distinguishing between purchasers and non-purchasers of "The History of Florence". SVMs are supervised learning models renowned for their efficacy in binary classifications. The SVM implemented in this study used a linear kernel, as indicated by the kernel = "linear" parameter, a common choice for problems with linear separability.

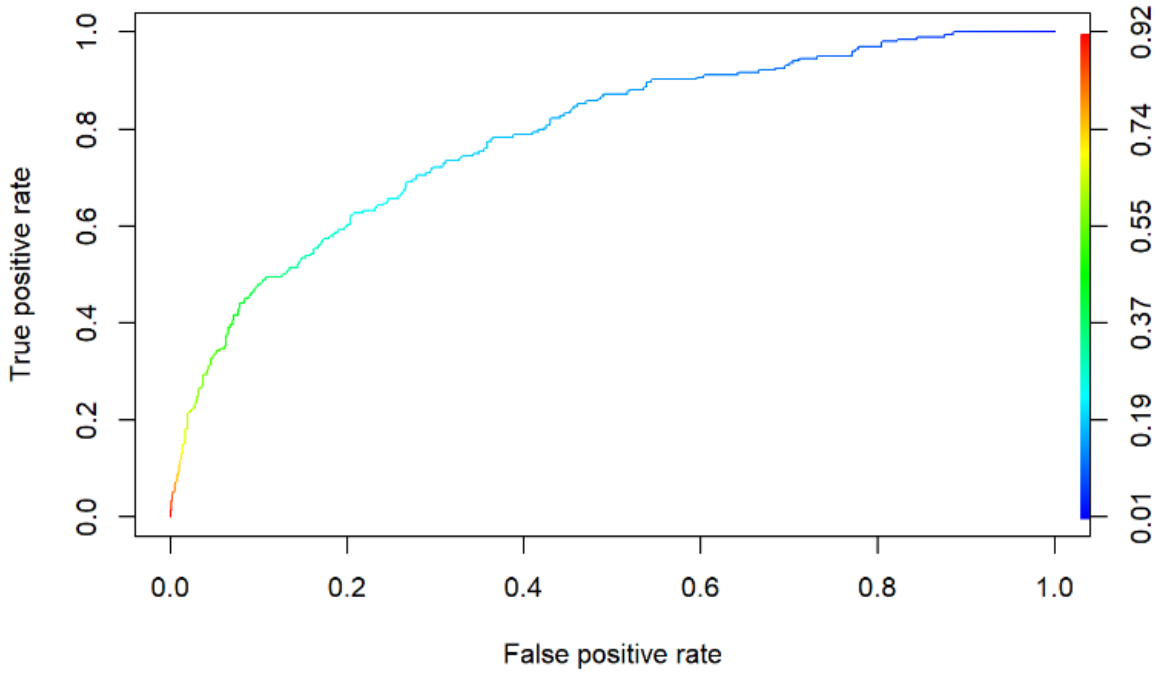
The model was trained on the BBBC_Train dataset, carefully excluding predictors such as 'Observation', 'Last_purchase', and 'First_purchase' to prevent issues of multicollinearity that could adversely affect the model's performance. The SVM setup involved C-classification with a cost parameter set to 1, and it utilized a total of 774 support vectors. These support vectors are pivotal as they define the model's separating hyperplane.

The model's performance on the training dataset was quantified using a confusion matrix, which revealed an accuracy rate of approximately 79.8%. Notably, the model demonstrated a high sensitivity of 94.92%, indicative of its strong ability to correctly identify actual non-purchasers. However, the specificity was relatively low at 34.5%, signaling a challenge in accurately classifying purchasers.



The Area Under the Curve (AUC) from the Receiver Operating Characteristic (ROC) curve stood at around 0.7758, reflecting a respectable discriminative ability of the model.

Further testing on the BBBC_Test dataset yielded an accuracy of 89.52%, which surpasses the training accuracy. Consistency was observed in sensitivity, with a marginally higher 94.99%, and specificity remained low at 33.33%.



Notably, the AUC on the testing set was 0.7852, slightly higher than the training set's AUC, which is indicative of the model's robustness and its capacity to generalize well to unseen data.

The SVM model exhibits a strong capacity to classify 'Purchasers' and 'Non-Purchasers' with high accuracy and consistency between training and testing datasets. Despite this, the low specificity is a concern that suggests the model may be biased towards predicting non-purchasers. It is recommended that further model tuning or alternative methods be explored to address the apparent class imbalance and improve the specificity without compromising the overall accuracy and sensitivity.

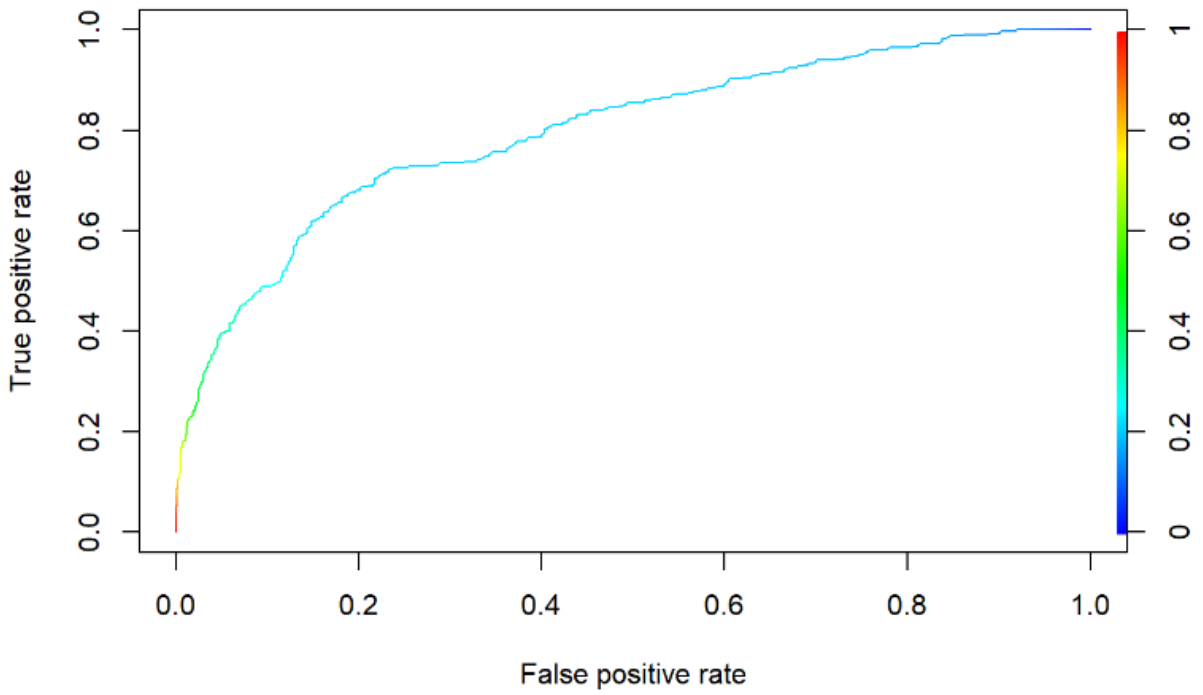
SVM Polynomial:

Our analysis proceeded with the application of a Support Vector Machine (SVM) employing a polynomial kernel, which is particularly suited for datasets where the relationship between the classes is not linear. The SVM was configured for C-classification with a polynomial kernel of degree 3 and a cost parameter set to 1. This model was trained on the BBBC_Train dataset, avoiding predictors that could introduce multicollinearity, such as 'Observation', 'Last_purchase', and 'First_purchase'.

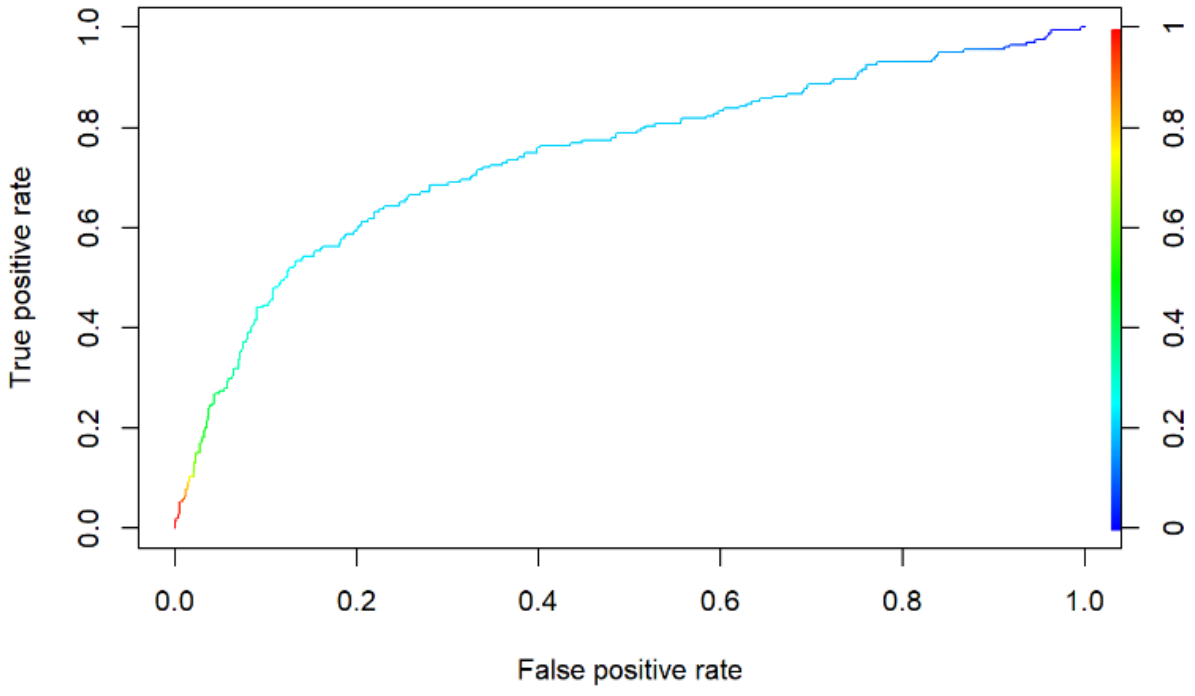
```
## Call:
## svm(formula = as.factor(Choice) ~ . - Observation - Last_purchase -
##     First_purchase, data = BBBC_Train, kernel = "polynomial")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  polynomial
##        cost:  1
##      degree:  3
##      coef.0:  0
##
## Number of Support Vectors:  745
##
##  ( 353 392 )
##
##
## Number of Classes:  2
##
## Levels:
##  0 1
```

The resulting SVM model utilized 745 support vectors to create the decision boundary between the classes, with a distribution of 353 for 'Non-Purchasers' and 392 for 'Purchasers'. This distribution of support vectors indicates the complexity and the space partitioning performed by the model to distinguish between the two classes.

Evaluation of the model's performance on the training data through a confusion matrix revealed an accuracy of 79.75%, with a sensitivity of 97.58% and a specificity of 26.25%. The high sensitivity rate underscores the model's adeptness at correctly identifying 'Non-Purchasers'. However, the specificity suggests that the model is less effective at correctly predicting 'Purchasers', a potential area for improvement.



Furthermore, the Area Under the Curve (AUC) for the training data was calculated to be approximately 0.7985, reflecting the model's substantial discriminative ability despite the imbalance in specificity and sensitivity.



When the model was applied to the test data (BBBC_Test), it achieved an accuracy of 89.78% and an AUC of 74.4%, which is slightly lower than the AUC from

the training set. This decrease in AUC on the testing data compared to the training set may indicate a loss in generalization when the model is exposed to new data.

In conclusion, the polynomial kernel SVM has demonstrated a high level of accuracy in classifying 'Purchasers' and 'Non-Purchasers'. However, the lower specificity and the drop in AUC when applied to the test data suggest that there is room for further optimization, possibly by adjusting the kernel's parameters or exploring alternative models to achieve a more balanced classification performance.

SVM sigmoid:

The classification challenge of distinguishing between purchasers ('1') and non-purchasers ('0') of a product was addressed using an SVM model equipped with a sigmoid kernel. The choice of a sigmoid kernel is particularly interesting as it introduces the capability to model complex, non-linear decision boundaries, akin to logistic regression.

Upon training, the SVM model with a sigmoid kernel was configured for C-classification with a cost parameter of 1. The model utilized a total of 745 support vectors to construct the decision boundary, indicative of the model's complexity and its strategy for class partitioning. Notably, the support vectors were almost evenly distributed between the two classes, with 353 for 'Non-Purchasers' and 392 for 'Purchasers'.

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction   0   1
##          0 989 257
##          1 211 143
##
##               Accuracy : 0.7075
##                 95% CI : (0.6845, 0.7297)
##    No Information Rate : 0.75
##    P-Value [Acc > NIR] : 0.99995
##
##                  Kappa : 0.1889
##
##  Mcnemar's Test P-Value : 0.03751
##
##             Sensitivity : 0.8242
##             Specificity : 0.3575
##          Pos Pred Value : 0.7937
##          Neg Pred Value : 0.4040
##              Prevalence : 0.7500
##          Detection Rate : 0.6181
##    Detection Prevalence : 0.7788
##       Balanced Accuracy : 0.5908
##
##        'Positive' Class : 0
```

The performance of the sigmoid kernel SVM on the training data was assessed through a confusion matrix. The model achieved an accuracy of 70.75%, with a sensitivity of 82.42% and a specificity of 35.75%. While the sensitivity is commendable, indicating a strong ability to detect 'Non-Purchasers', the specificity highlights room for improvement in correctly classifying 'Purchasers'.

```
## [1] 0.6067573
```

The Area Under the Curve (AUC) for the training data stood at approximately 0.6067, suggesting a moderate discriminative ability of the model.

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction    0    1
##          0 1698  116
##          1  398   88
##
##                Accuracy : 0.7765
##                  95% CI : (0.7589, 0.7934)
##     No Information Rate : 0.9113
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1487
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.8101
##             Specificity : 0.4314
##          Pos Pred Value : 0.9361
##          Neg Pred Value : 0.1811
##              Prevalence : 0.9113
##          Detection Rate : 0.7383
##    Detection Prevalence : 0.7887
##       Balanced Accuracy : 0.6207
##
##        'Positive' Class : 0
```

When applied to the test dataset, the sigmoid kernel SVM model maintained a high accuracy of 77.65% and improved specificity to 43.14%, while retaining a high sensitivity of 81.01%.

```
## [1] 0.6531734
```

The AUC on the testing dataset was 0.6532, a marginal improvement from the training AUC, but still reflecting a moderate level of discrimination between the two classes.

In summary, the application of an SVM with a sigmoid kernel has demonstrated a capacity for high sensitivity in the binary classification task. The model's accuracy and AUC on the test data suggest a fair generalization ability. However, the specificity levels indicate that the model is less adept at identifying true 'Purchasers', which could be a focal point for further model tuning or exploration of alternative approaches to enhance the balance between sensitivity and specificity.

SVM RBF:

VI. Findings

The three models that were run produced different results. As stated earlier, linear regression is not an appropriate prediction method due to a lack of linearity in the DV and due to it being a binary variable. The other two classification methods performed well, and had high accuracy and sensitivity, with specificity around 30%. Due to the large amount of "Purchasers" (1s), the data is heavily skewed to "Purchasers" (0) and therefore results in much higher sensitivity than specificity. The results of the Logistic Regression & SVM model are outlined below:

Logistic Regression
Accuracy: 89.5%
Sensitivity: 94.5%
Specificity: 37.3%

Balanced Logistic Regression
Accuracy: 74%
Sensitivity: 82.4%
Specificity: 65.7%

SVM Model Linear Unbalanced
Accuracy: 89.5%
Sensitivity: 95%
Specificity: 34.5%
AUC: 78.5%

SVM Model Balanced
Accuracy: 73.5%
Sensitivity: 82.8%
Specificity: 64.2%
AUC: 81.6%

Overall, all 3 models are very accurate, with accuracy never dropping below 70%. When taking a detailed view, there are vast differences between the 3 models, notably in specificity. A main concern with the Unbalanced Logistic Regression & Unbalanced SVM Models is that there are much higher numbers of "Non-Purchasers" than "Purchasers" in the training and testing datasets. This is exactly why accuracy and sensitivity are so high since the model can predict "Non-Purchasers" with near-perfect accuracy since there are so many of them! However, "Purchasers" are just as important, if not more important for the predictions. In the end, the "Purchasers" are driving the

revenue of "The History of Florence", and they're the target audience for these predictive models. In that regard, the Balanced Logistic Regression model beats the two other models, with a much higher specificity of 65.7%. It's possible that collecting more data on "Purchasers" could increase this figure above 70%, but currently, it is still much better than the 30%-35% range obtained by the Unbalanced SVM & Unbalanced Logistic Regression. Comparing Balanced Logistic Regression against the Balanced SVM, Logistic Regression has the higher specificity and is therefore the better model. When comparing both Balanced SVM models, the Logistic Regression is slightly better, edging out the SVM model in every category except Sensitivity, but only by 0.8%. If the goal is to find the best and most balanced model, the Balanced Logistic Regression provides the best results.

VII. Profitability Analysis

This report will also seek to answer the following question: "How much more profit could the company expect to make by using these models as opposed to sending the mail offer to the entire mailing list?" The question addresses the notion that it could be either more or less profitable to use the model to target predicted "Purchasers" of "The History of Florence" than to simply send the mailing list to everyone on the list and see what happens, even if the model would predict them to be "Non-Purchasers" of the book. Before answering this question, notable metrics should be recorded from the base-case scenario, which will be compared against the effectiveness of the models

```
mailing_price <- 0.65
book_cost <- 15
overhead <- .45 * book_cost
book_price <- 31.95
```

mailing_price is the mailing cost per addressee, so it costs $0.65 to send the mailing list to each addressee
book_cost is the cost of purchasing and mailing the book
overhead is applied to each book, and it is 45% of the cost of each book
book_price is the retail price of the book

To solve this question, potential revenue & costs will be calculated for the number of predicted purchasers in both the Logistic Regression & SVM models. The total profit of each model will be compared against the result of simply mailing the list out to everyone in the testing sample, assuming that 9.06% of customers buy the book:

```
customers_mailed <- 2300
orders_rec <- round(customers_mailed*0.0906,0)
yield <- 0.0906
mailing_costs <- customers_mailed*mailing_price
book_costs <- (orders_rec * book_cost) + (orders_rec * overhead)
revenue <- book_price * orders_rec
profit <- revenue - book_costs - mailing_costs
```

```
"Total revenue from the 1,806 book purchases was: $ 6645.6"
"The total costs to send the mailing list and purchase/send the book were: $ 6019"
"Total profit was: $ 626.6"
"The success of the campaign is 0.09"
```

Balanced Logistic Regression Profitability

Firstly, the Balanced Logistic Regression model will be analyzed. Below is the confusion matrix for the testing Logistic Regression model:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 168  70
         1  36 134

               Accuracy : 0.7402
                 95% CI : (0.6948, 0.7821)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4804

 Mcnemar's Test P-Value : 0.001349

            Sensitivity : 0.8235
            Specificity : 0.6569
         Pos Pred Value : 0.7059
         Neg Pred Value : 0.7882
             Prevalence : 0.5000
         Detection Rate : 0.4118
   Detection Prevalence : 0.5833
      Balanced Accuracy : 0.7402

       'Positive' Class : 0
```

Of the 408 people in the sample, 168 are true "Non-Purchasers". There is no need to spend money mailing these customers the mailing list since there is no chance they would purchase the

book. 70 customers in the model were classified as "false negatives", which means that the model predicts them to be "Non-Purchasers" when in reality they did purchase "The History of Florence". These clients should be sent the mailing list since there is a very real chance that they could purchase the book. Lastly, 134 clients were correctly predicted as "Purchasers". These clients must be on the mailing list.

Let's calculate the metrics for the Logistic Regression:

```
customers_mailed <- 134+70
orders_rec <- 134
yield <- orders_rec/customers_mailed
mailing_costs <- customers_mailed*mailing_price
book_costs <- (orders_rec * book_cost) + (orders_rec * overhead)
revenue <- book_price * orders_rec
profit <- revenue - book_costs - mailing_costs
print(paste("Total revenue from the 134 book purchases was: $",round(revenue,2)))
print(paste("The total costs to send the mailing list and purchase/send the book were: $", round(book_costs+mailing_costs,2)))
print(paste("Total profit was: $",round(profit,2)))
print(paste("The success of the campaign is", round(yield,2)))
```

```
"Total revenue from the 134 book purchases was: $ 4281.3"
"The total costs to send the mailing list and purchase/send the book were: $ 3047.1"
"Total profit was: $ 1234.2"
"The success of the campaign is 0.66"
```

Of the 204 customers that received the mailing list, 134 ended up purchasing "The History of Florence" based on the model. Let's compare key metrics against the base model:

Revenue: $6,645.6 base vs. $4,281.3 Log
Total Costs: $6,019 base vs. $3,047.1
Profit: $626.6 base vs. $1,234.2 Log
Success Rate: 9.06% base vs 66% Log

Overall, the Logistic Regression is more profitable than the base case model of sending everyone the mailing list. Total costs are 97.5% lower for the Logistic Regression model, and profit is 49.2% higher. Even though revenue is much higher in the base case, the benefit of this is erased by the high mailing and book costs. Also, the success rate of 66% for the Logistic Regression is 628.5% higher, which could provide even higher profits with a larger sample and/or in the population.

Balanced SVM Model Profitability

Now for the Balanced SVM Model. Once again, the confusion matrix must be analyzed for the linear kernel SVM, which provided the best results:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 169  73
         1  35 131

               Accuracy : 0.7353
                 95% CI : (0.6897, 0.7775)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4706

 Mcnemar's Test P-Value : 0.0003704

            Sensitivity : 0.8284
            Specificity : 0.6422
         Pos Pred Value : 0.6983
         Neg Pred Value : 0.7892
             Prevalence : 0.5000
         Detection Rate : 0.4142
   Detection Prevalence : 0.5931
      Balanced Accuracy : 0.7353

       'Positive' Class : 0
```

The SVM model correctly classifies 169 customers as non-purchasers, so they will not be sent to the mailing list. There are 73 false negatives, people whom the model said would NOT purchase "The History of Florence" but did. 131 people were correctly predicted as purchasers of the book. Overall, 204 people should be sent to the mailing list to optimize the results. Let's analyze costs and revenue for the SVM Model:

```
customers_mailed <- 204
orders_rec <- 131
yield <- orders_rec/customers_mailed
mailing_costs <- customers_mailed*mailing_price
book_costs <- (orders_rec * book_cost) + (orders_rec * overhead)
revenue <- book_price * orders_rec
profit <- revenue - book_costs - mailing_costs
```

```
"Total revenue from the 1,806 book purchases was: $ 4185.45"
"The total costs to send the mailing list and purchase/send the book were: $ 2981.85"
"Total profit was: $ 1203.6"
"The success of the campaign is 0.64"
```

This is how the results compare against the base case:

Revenue: $6,645.6 base vs. $4,185.5 SVM
Total Costs: $6,019 base vs. $2,981.9 SVM
Profit: $626.6 base vs. $1,203.6 SVM
Success Rate: 9.06% base vs 64% SVM

Overall, profit is 92% higher with the Balanced SVM than for the base case, a massive improvement. The success rate of 64% is very comparable to the 66% success rate of the Logistic Regression mode. Costs decreased by 37% against the base case, and revenue decreased by 37% against the base case, but the decline in revenue was offset by the decrease in costs. Overall, the SVM model is much better than the base case model.

Logistic & SVM Profitability Comparison:

Revenue
$4,185.5 SVM vs. $4,281.3 Log
Costs
$2,981.9 SVM vs. $3,047.1 Log
Profit
$1,203.6 SVM vs $1,234.2 Log
Success Rate
64% SVM vs 66% Log
Orders Placed
131 SVM vs 134 Log

Best Profitability Model

Overall, the Logistic Regression model is the best at delivering higher profit and a higher success rate. Even though costs increased for the Logistic Regression model, this is offset by higher profits and a higher success rate.

VIII. Best Predictors

From the models and coefficients, we found out that *P_Art* (people buying Art books) had a positive impact on people buying "The History of Florence" book. From the logistic regression analysis (after removing variables with high multicollinearity), we found out that one increase in *Choice* would have a 1.343 increase in *P_Art*. *Gender1* was after *P_Art* the variable with higher coefficients (-0.789). *Gender1* being Male and being a negative coefficient, is interpreted as men being less likely to buy "The History of Florence" book.
Therefore, based on coefficients and predictors, BBBC should mostly target women who have bought Art books in the past, to make sure that they will get more profit from their books. Based

on our analysis they should also use a balanced logistic regression method to do so, to increase its profitability.

IX. Assumptions & Limitations

Three models were analyzed during this report: Linear Regression, Logistic Regression, and SVM. All 3 models are unique in their ways, and have assumptions and limitations that must be understood.

SVM models are extremely versatile classification models. SVM has various advantages:

* Effective on data sets with multiple features
* Effective in data sets where the number of features > the number of observations
* Can leverage different kernels for data that is not linear

However, they do have some limitations and disadvantages:

* The model is hard to interpret due to the inability to provide probabilities
* Works best on smaller data sets due to its high training time
* Can be computationally expensive to run for large and complex data sets

Logistic regression is a popular method for binary classification problems. It has several advantages:

* Interpretability: One of the main strengths of logistic regression is its interpretability. Coefficients can be directly related to the odds ratio for each of the predictors, making it easier to understand the impact of each variable.
* Probability Estimates: Unlike some other classification methods, logistic regression provides probabilities for the outcomes, allowing for a nuanced understanding of the predictions.
* Simplicity and Efficiency: Logistic regression is straightforward to implement and computationally not as demanding as more complex models, making it a good baseline model for binary classification problems.
* Flexibility with Feature Transformation: It can handle both linear and non-linear effects using transformations or interaction terms.

Despite its advantages, logistic regression comes with limitations and disadvantages:

* Assumption of Linearity: Logistic regression assumes a linear relationship between the log odds of the dependent variable and each predictor variable. This can be overly simplistic for some real-world scenarios where the relationship is not linear.
* Performance with Complex Relationships: It may not perform well when there are complex relationships between features that are difficult to capture with linear boundaries or when the data is highly dimensional without regularization.
* Vulnerability to Overfitting: In cases where the data set includes a large number of features, logistic regression can become prone to overfitting, although techniques like regularization can help mitigate this.
Sensitive to Imbalanced Data: Logistic regression can be sensitive to unbalanced data, leading to biased models that favor the majority class. Special techniques like oversampling, under-sampling, or penalization methods are often required to handle this issue.

Linear regression is primarily used for predicting the value of a continuous dependent variable based on one or more predictor variables. The method assumes a linear relationship between the independent and dependent variables, and it estimates the dependent variable as a weighted sum of the independent variables, plus an intercept.

When you have a binary dependent variable (e.g., 0 or 1, representing categories like "no" or "yes," "fail" or "pass"), using linear regression can lead to several issues:

Non-linearity: The assumption of linearity is violated when the outcome variable is binary. The relationship between the predictor(s) and the probability of the outcome being 1 (or 0) is not linear but S-shaped when plotted on a graph (logistic function). Therefore, the linear model does not fit the data well.

Prediction outside of bounds: Linear regression predictions can fall outside the range of 0 and 1, which doesn't make sense for binary outcomes because probabilities must be between 0 and 1.

Homoscedasticity violation: Linear regression assumes that the variance of the error terms is constant across all levels of the independent variables. However, for a binary dependent variable, the variance of the error terms is not constant and depends on the predicted probability, leading to heteroscedasticity.

Error distribution: Linear regression assumes that the residuals (errors) are normally distributed. However with a binary dependent variable, the error distribution can deviate significantly from normality, especially since the outcome can only take on two values.

Because of these reasons, logistic regression is typically used when the dependent variable is binary. Logistic regression models the probability that the dependent variable belongs to a particular category. Unlike linear regression, logistic regression does not assume a linear relationship between the independent and dependent variables. It uses the logistic function to ensure that the predictions fall between 0 and 1, making it suitable for binary (and categorical) outcomes.

X. Best Model for Case

Overall, both the Balanced/Unbalanced SVM & Logistic Regression models have comparatively similar results. All 4 are very accurate and efficient classification models that allow the company to determine who is most likely to buy and not buy "The History of Florence". However, there are important differences that cannot be overlooked. Particularly, the effect of balancing the data between "Purchasers" (1) & "Non-Purchasers" (0) provided much higher specificity for both SVM & Logistic Regression, which gave the models much more power in predicting the "Purchasers" of the book. After having analyzed both the models' results and profitability, the best model for this case is the **Balanced Logistic Regression model.** As previously mentioned, the Specificity of 65.7% is the highest of all 4 models run in this report. Since the company is seeking to maximize the number of "Purchasers", this model is the best at doing so. In terms of profitability, this model also generated the most amount of profit: $1,2343.2 out of 134 sales. As touched upon earlier, there are also a significant number of false negatives that could increase the total number of sales, and this should be explored further. For the sake of this report, Balanced Logistic Regression is the best model for predicting "Purchasers".

XI. Conclusion

After cleaning the data, doing visualizations, deleting outliers, and removing variables (multicollinearity), we performed SVM (linear, polynomial, Sigmoid, RBF), Logistic Regression, and Linear Regression. Since we saw that the data was unbalanced, we did a logistic and SVM model which balanced the data and performed the best. Since we were trying to predict people who purchase the book (Specificity), we decided to choose Logistic Regression (Balanced), as the best model, due to their high accuracy and specificity. It was also the model that performed the best and gave the company more profit. This model is great for predicting categorical variables and is very easy to interpret.
The best predictors were Female and P_Art, therefore we suggest the company mainly target women who have already bought Art books. BBBC should also use Balanced Logistic Regression for future Analysis.

To simplify and automate the recommended methods for future modeling efforts at the company we would suggest BBBC collect more data from their clients (more observations), and add more variables to the survey such as kids or marital status, to be able to have more predictors and therefore do a better classification analysis. We would suggest the company do a more comprehensive analysis of each predictor to be able to classify each customer more and look for patterns and trends among the data. Once they have enough data from the customers, they could do the same analysis with other category books to see how they can make more profit with other books and improve their customer retention.