DA6813
Mason Schweder (cko441)
Roddy Caldwell (tcg315)
Joshua Gutierrez (eqt138)
Nancy Lopez Ramirez (hpn459)

Bookbinders Case Study

I. Executive Summary

We used a dataset with 1,600 observations to fit three models attempting to predict whether a customer would purchase a particular book. The three models examined were a multiple linear regression, a logistic regression and a support vector machine. The logistic regression model resulted in an accuracy rate of 86.96% and proved to be the most profitable of the three. We proposed using the logistic regression model to identify the predictors with the most influence on a customer's decision and to assist in designing targeted marketing campaigns.

II. The Problem

For this case study, we'll be using a data set provided by the Bookbinders Book Club (BBBC), which was established in 1986 in order to sell specialty books using direct marketing. The BBBC wants to know if they can improve the effectiveness of their marketing campaigns using predictive modeling; specifically ordinary linear regression, a logit model, and support vector machines. To explore this possibility, we are using data regarding a recent mailing campaign which included a specialized brochure for the book *The Art History of Florence*. We will use a subset of the database that includes 1600 observations: 400 customers who purchased the book and 1200 who did not. Our goal is to determine if any or all of the methods of interest will be effective in optimizing the BBBC's future mailing campaigns.

This report will discuss the steps we took in order to accomplish our goal. First, we'll discuss some literature that we reviewed while pursuing these modeling techniques. Next, we'll discuss the data set, and how we were able to clean and manipulate the data in order to serve our needs. We'll then discuss the explorations that we performed, followed by the results of those explorations. We'll end with our conclusions, recommendations, and answers to some relevant questions regarding the BBBC's goals.

III. Review of Related Literature

One important concept we note in our analysis is the tradeoff between accuracy and profit maximization. We find later in our model performance output that while SVM and logistic perform similarly in regard to accuracy, they do differ in terms of their misclassification rate, which in turn affects the profitability of the marketing campaign based on that model. We found past research to support this scenario in Stripling's examination of customer churn [1]. Stripling actually used a separate measure, EPMC, expected maximum profit measure for customer

churn, to account for the discrepancy between accuracy and profit. He remarks “that model selection based on the AUC leads to suboptimal profit”. This is why when we evaluate model performance, we do not only look for accuracy, but also cost-benefit performance.

In our initial explorations, we were getting suspicious results from running the logistic regression and SVM models. With further research, we were able to find an article titled “When and Why to Standardize a Variable” by Deepanshu Bhalla on a site called “Listen Data.” This article goes into detail about the different statistical techniques that go into standardizing variables, as well as guidance about when to do so. Using this article, we were able to identify the lack of standardization of our variables as our problem. Bhalla specifically references logistic regression models and support vector machines as instances in which standardizing variables is important, and reassures us that it is not necessary for linear regression models.

When performing the logistic regression, we used an article from the website “STHDA,” which stands for “Statistical Tools for High-throughput Data Analysis.” The article we used cites a chapter from *Machine Learning Essentials: Practical Guide in R* by Alboukadel Kassambara. This article helped us to better understand the assumptions that a logistic regression model operates under, and how to test them using the libraries “tidyverse” and “broom.”

Regarding cost-benefit analysis, we were able to refer to Kanisetty’s article where he performs a similar return on investment calculation for a marketing campaign [2]. Although he coded in python, the general strategy still applies and informed our analysis to answer some questions posed by the bookbinder customer in the case study.

IV. Data

The data we were given to work with was already clean, organized, and weighted for us. We were given a testing data set of 1600 observations to use in building the models, and a testing data set of 2300 observations to use in determining the accuracy and effectiveness. There were no missing values.

The only transformation we had to make was to standardize the continuous variables. We had several numeric variables that were using different scales, to include the number of books purchased, total money spent on books, and months since first and last purchases. Standardizing these variables allowed us to use them in our models and assess them accurately.

V. Methodology and Results

We fit a multiple linear regression model of “Choice” on all predictors, except for “Observation” because it is an index variable. Predictors were then selected using backward selection. “First_purchase” was removed first, which was unsurprising given that it is highly correlated to “Last_Purchase” with a correlation coefficient of 0.814 in the training set. “P_Youth” was removed next as it had a p-value above 0.05. The final iteration of the linear model ended up including “Gender”, “Amount_purchased”, “Frequency”, “Last_purchase”, “P_Child”, “P_Cook”, “P_DIY”, “P_Art” (see Appendix for coefficient estimates). This model has a small p-value of 2.2e-16 which would indicate it is a valid model, but it has an R-square value of 0.22. In other words, it only explains 22% of the variance observed.

Most importantly, though, the multiple linear regression technique is not the best approach to a question dealing with a categorical response variable. This is evident in that the model only fulfilled 3 of the 4 OLS assumptions (see Diagnostics Plots in the Appendix). Its residuals plot shows a very clear pattern between the residuals and fitted values, obviously driven by the only two response values of 0 and 1. We cannot assume a linear relationship between the predictors and the outcome variables. The variances of the residual points are also not spread equally along the range of fitted values. The variance of the residuals actually decreases as the fitted values approach 0 and 1. Typically, heteroskedasticity could be addressed with a log or square root transformation of the response variable. But this would be a problem given that our response values are only 0 and 1. For example, Log of 1 is 0 and Log of 0 is undefined. The residuals are not normally distributed either, as they do not follow a straight line in the residuals Q-Q plot. Finally, there was some multicollinearity, as described above, specifically between “First_purchase” and “Last_purchase”. This was addressed by removing “First_purchase” from the model. Please refer to the Appendix for diagnostic plots.

The linear model was run on the test data set, despite it not fulfilling all 4 assumptions. When examining the predicted values, we see that the minimum value is negative -0.37 and the maximum is 0.94. Again, due to the fact that we are dealing with binary responses of 0 and 1, these predicted values are not particularly useful and are difficult to interpret. We went further, however, interpreting these predicted values as “crude probability estimates” [6] the way one would do with a logistic regression model and associated predicted values larger than 0.5 to be 1 and less than 0.5 to be 0. Doing so resulted in an accuracy rate of 89.87%.

Although the model itself is not particularly useful, its coefficient estimates did reinforce insights observed in the logistic regression model in terms of predictor significance and relationship to the response variable.

We fitted the logistic regression model using the scaled variables, and then used backward selection to remove the insignificant variables. Using this method, we only removed the variable “First_purchased”, since every other variable remained significant. Then, in checking the assumptions of the model, we noticed that “Last_purchase_scaled” had a problematic level of

collinearity, so we removed that variable. Removing “Last_purchase_scaled” caused “P_Youth_scaled” to become insignificant, so we continued the backward selection process and removed it as well. The remaining variables were all significant and had acceptably low levels of collinearity.

```
glm(formula = Choice ~ Gender + Amount_purchased_scaled + Frequency_scaled +
    P_Child_scaled + P_Cook_scaled + P_DIY_scaled + P_Art_scaled,
    family = binomial, data = bb.training)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.30792  -0.69156  -0.47311  -0.02466   2.84228

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -0.85087    0.10439  -8.151 3.62e-16 ***
Gender                  -0.81204    0.13457  -6.034 1.60e-09 ***
Amount_purchased_scaled  0.22738    0.07317   3.108  0.00189 **
Frequency_scaled        -0.69435    0.08137  -8.533  < 2e-16 ***
P_Child_scaled          -0.20799    0.07624  -2.728  0.00637 **
P_Cook_scaled           -0.30584    0.07567  -4.042 5.30e-05 ***
P_DIY_scaled            -0.19193    0.07316  -2.623  0.00870 **
P_Art_scaled             0.91400    0.07263  12.585  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1799.5  on 1599  degrees of freedom
Residual deviance: 1445.1  on 1592  degrees of freedom
AIC: 1461.1

Number of Fisher Scoring iterations: 5
```

This GLM model performed at an 86.96% accuracy on our test data set.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1907  189
         1  111   93
```

```
               Accuracy : 0.8696
                 95% CI : (0.8551, 0.8831)
    No Information Rate : 0.8774
    P-Value [Acc > NIR] : 0.8797

                  Kappa : 0.3119

 Mcnemar's Test P-Value : 8.765e-06

            Sensitivity : 0.9450
            Specificity : 0.3298
         Pos Pred Value : 0.9098
         Neg Pred Value : 0.4559
             Prevalence : 0.8774
         Detection Rate : 0.8291
   Detection Prevalence : 0.9113
      Balanced Accuracy : 0.6374

       'Positive' Class : 0
```

We can now interpret the variables and how they impact our response by analyzing the coefficients. The -0.812 for “Gender” shows us that females are more likely to buy *The Art History of Florence* than males. The positive 0.227 for “Amount_purchased_scaled” means that the more money a person spends on books in general, the more likely they are to buy this one. The -0.694 for “Frequency_scaled” tells us that a person is more likely to buy the book if they haven’t purchased many books recently. The negative coefficients on “P_Child_scaled”, “P_Cook_scaled”, and “P_DIY_scaled” tell us that if a person has bought one of these kinds of books, they are less likely to buy *The Art History of Florence*. However, the positive coefficient on “P_Art_scaled” means that a person is more likely to buy this book if they have bought an art book in the past.

In addition to the GLM we also ran two SVM models. Similar to the GLM we scaled all the numeric variables for our SVM analysis. First we ran a tuned svm function and had an accuracy of 90%. While initially 90% accuracy would be fantastic once taking a closer look at the confusion matrix we see our problem. Many true negatives are predicted and that contributes to nearly all of the accuracy in this model.

```
      Confusion Matrix and Statistics

                Reference
      Prediction    0    1
               0 2033  167
```

```
         1   63   37

               Accuracy : 0.9
                 95% CI : (0.887, 0.912)
    No Information Rate : 0.9113
    P-Value [Acc > NIR] : 0.9724

                  Kappa : 0.1965

 Mcnemar's Test P-Value : 1.109e-11

            Sensitivity : 0.9699
            Specificity : 0.1814
         Pos Pred Value : 0.9241
         Neg Pred Value : 0.3700
             Prevalence : 0.9113
         Detection Rate : 0.8839
   Detection Prevalence : 0.9565
      Balanced Accuracy : 0.5757

       'Positive' Class : 0
```

The second SVM we ran was a non-tuned version using a linear kernel. The accuracy on this model was virtually the same (89.96%) but posed a similar problem as before. An overwhelming majority of the correct predictions are true negatives.

```
                svm.pred
Reference       0          1
           0    2016       80
           1    151        53

Accuracy 0.8995652
```

Below are the accuracy, precision and recall rates of the three models we examined.

MODEL	**ACCURACY**	**PRECISION**	**RECALL**
Linear	89.87%	.275	.397
Logistic	86.96%	.456	.330
SVM Tuned	90.00%	.370	.181

SVM Linear	89.96%	.260	.398

To see why we care if our model has an overwhelming number of true negatives and a very small amount of true positives, we must look at this from a contextual standpoint. Our objective is to maximize profit for a book company trying to optimize mailings promoting a specific book. While true negatives help save money by reducing mail, true positives are where the money will be made. Taking a look at our confusion matrices we see true positives account for the smallest portion of the confusion matrix. To help get a monetary grasp on which model would be best we calculated the cost and profit for the company if they mailed brochures to every client in their new database, as well as for each model we ran. The BBBC's mail campaign for the midwest includes data on 50,000 customers. Each brochure cost $0.65 to mail. The book costs $15 for the company to purchase and ship to the customer, and a 45% overhead for each book is allocated. The book will be sold for $31.95. Based on previous mail campaigns the company expects 9.03% of customers who receive a brochure to purchase the book. Using no model at all and mailing a brochure to every customer in the new database, the company should expect to spend $130,701.30 and have an income of $144,254.30. This results in a profit of $13,553.00. For the tuned SVM model we rather than sending the brochure to every customer the model will suggest who to send brochures to. For this specific SVM we approximate 4.3% of customers would be likely to purchase the book. Of those customers sent brochures, the model predicts 37% will purchase the book. Total cost for this process is $18,907.61 with an income of $25,698.91 giving a profit of $6,791.30. For the linear kernel SVM we approximate it will find 8.9% of customers worthy of sending a brochure and of those customers 26% will purchase the book. This gives a cost of $27,942.39, an income of $36,811.96 and a profit of $8,869.57. Both SVM models do lower the cost of the mailing campaign, however it also significantly reduces the profit. Based on our GLM results we expect to send a brochure to 8.9% of customers and 45% of those customers will purchase the book. This process would cost $46,855.43, have an income of $64,594.57, giving the company a profit of $17,739.13. The GLM simultaneously reduces the cost and increases the profit. For this reason we highly recommend the company invest in using GLM's for this campaign and future campaigns.

Model	**Predicted Cost**	**Predicted Income**	**Predicted Profit**
None	$130,701.30	$144,254.30	$13,553.00
Logistic	$46,855.43	$64,594.57	$17,739.13
SVM Tuned	$18,907.61	$25,698.91	$6,791.30

SVM Linear	$27,942.39	$36,811.96	$8,869.57

Advantages and Disadvantages of 3 models

Linear Regression:
We would not recommend using linear regression for modeling these campaigns. Given that our response variable is a two-level factor (categorical), linear regression would not be appropriate.

Logistic regression:
Logistic regression ended up being our top candidate for our model selection. It not only produced a high accuracy (~90%), but performed best in our cost-benefit analysis by having a lower misclassification rate. In terms of disadvantages of logistic models, they are more strict on following the parametric assumptions, which we reviewed in our logistic model analysis. Because our model did not meet all 4 assumptions, it caused us to have less confidence in the model output. An SVM in comparison is well suited in instances where we cannot satisfy the parametric assumptions.

However, one advantage of the logistic model that we appreciate is greater interpretability. There is a direct relationship between a given parameter and the response variable, where you can say for example “with an 1 unit increase in predictor X, the odds of having a book purchase increases by a factor of 1.844”. Further, there is clarity in terms of which predictors have the highest impact on the response variable, and the relative strength of all predictors, provided by the parameter coefficients. That can also help us in the feature selection process by indicationing which predictors are significant. Ultimately, that will help us provide specific recommendations to the customer, about which customer attributes they should target in their marketing campaign.

Support Vector Machine:
Even though SVM produces a very high accuracy (~90%), it has one major drawback. Unlike Linear and Logistic (parametric) models, the SVM is very difficult to interpret. The SVM does not produce clean coefficient values that can explain the strength of individual predictors. The best we can examine the output is by looking at relative importance of the predictors. In this way, SVM is more of a ‘black box’ that is great for highly accurate models, but presents a challenge when it comes to offering concrete recommendations for the customer, for example, which customer attributes to focus on in this context.

Based on the advantages and disadvantages of these three models, we would recommend to focus on utilizing the logistic regression model, and continuing to refine only one model, as opposed to using all 3. Logistic offers the most in this scenario; not only a high accuracy

(rivalling SVM), but also high interpretability, which allows the bookbinders club to focus on specific customer attributes and use targeted marketing in the future. If they were to expand their database by collecting many more customer attributes, at the point, it may be worth it to start exploring SVM since it's a machine learning model better suited to large datasets.

VII. Conclusions and Recommendations

If we were to make a recommendation based on the results of our GLM model about who the BBBC should prioritize for their next direct marketing campaign, it would be women who have bought art books in the past and that spend a lot of money on books, but who have not made many recent book purchases.

However, in checking the assumptions of the logistic regression model, we noticed that it only satisfies three of the four assumptions this model performs under. The model does not technically satisfy the assumption that there is a linear relationship between the logit of the outcome and each of the predictor variables. This means that the model could possibly be misleading or inaccurate. The results of the model are easy to interpret, and the prediction accuracy was high when performed on the test data, but we would be obligated to mention that caveat alongside our recommendations to the BBBC.

Simplification and Automation

BBBC could leverage this logistic regression model to streamline marketing and sales processes. For instance, the customer profile that is interested in purchasing an art book may be significantly different than the customer who purchases a cooking book. If BBBC has a proper data collection and management framework in place, it could easily replicate this experiment and develop more models geared towards various genres available. While developing more models may seem to be a step away from simplification, we envision that it would actually boil down all the potential areas of focus down to, simply, genre. Having various models for each genre ready to help target the right customers each month would allow BBBC sales teams 1) the ability to diversify and potentially increase its monthly selections; 2) the flexibility to launch multiple, more effective marketing campaigns.

With automation and growth in mind, though not necessarily simplification, a recommendation for the long term would be to capture more granular data regarding the characteristics of the monthly selections (author, length, etc.) and of the customers themselves. This type of information may help inform targeted marketing content used to, not just target potential purchasers but, actually *influence* the consumers' purchase decisions.

Regardless of the number of models built, they need to be made available to sales and marketing teams in a form that allows for quick business decisioning and content creation.

APPENDIX

Logistic Regression

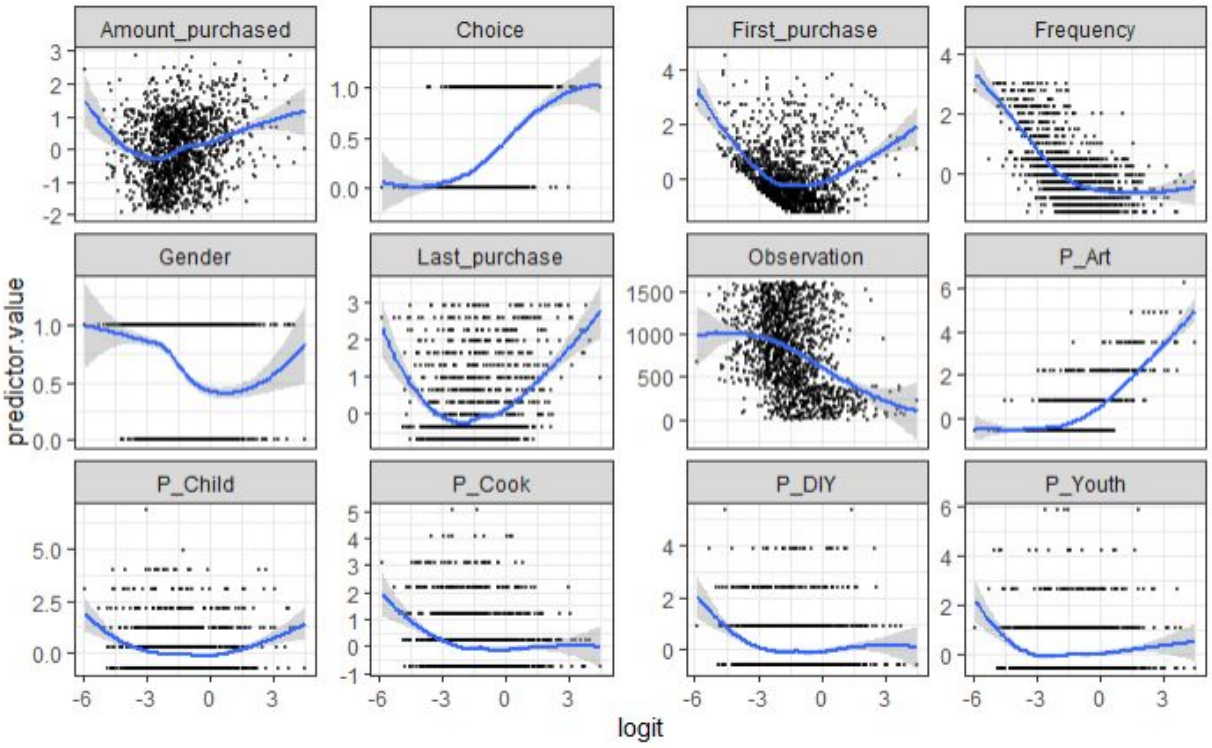
Code for the confusion matrix:

```
bb.testing$PredProb = predict.glm(glm3, newdata = bb.testing, type =
'response')
bb.testing$PredChoice = ifelse(bb.testing$PredProb >= 0.5, 1, 0)
caret::confusionMatrix(as.factor(bb.testing$Choice),
as.factor(bb.testing$PredChoice))
```

Checking the assumptions of the GLM model. The logistic regression method assumes that:

1. The outcome is a binary or dichotomous variable like yes vs no, positive vs negative, 1 vs 0. Our model satisfies this assumption.
2. There is a linear relationship between the logit of the outcome and each predictor variable.

```
ggplot(mydata, aes(logit, predictor.value))+
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_bw() +
  facet_wrap(~predictors, scales = "free_y")
```

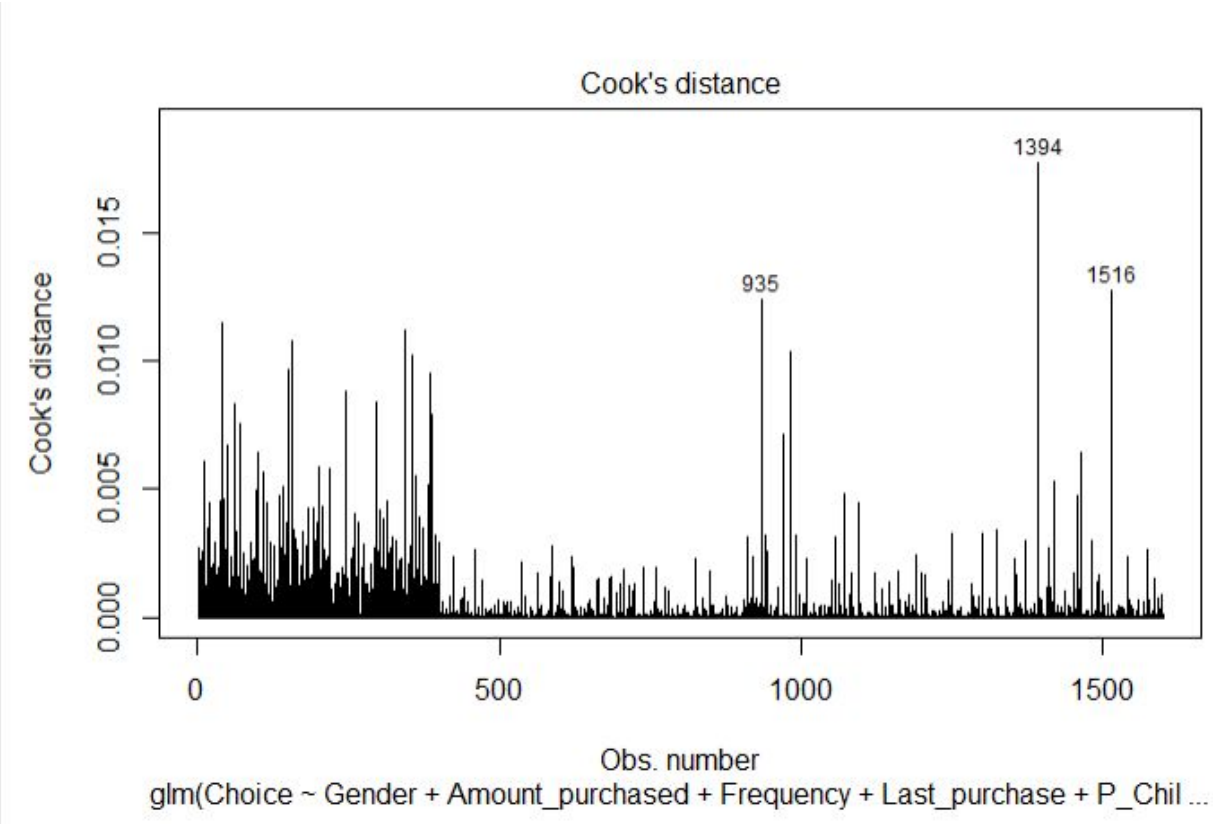


None of these seem to have a linear relationship with the logit of the outcome. This means that our model does not satisfy this assumption.

3. There are no influential values (extreme values or outliers) in the continuous predictors.

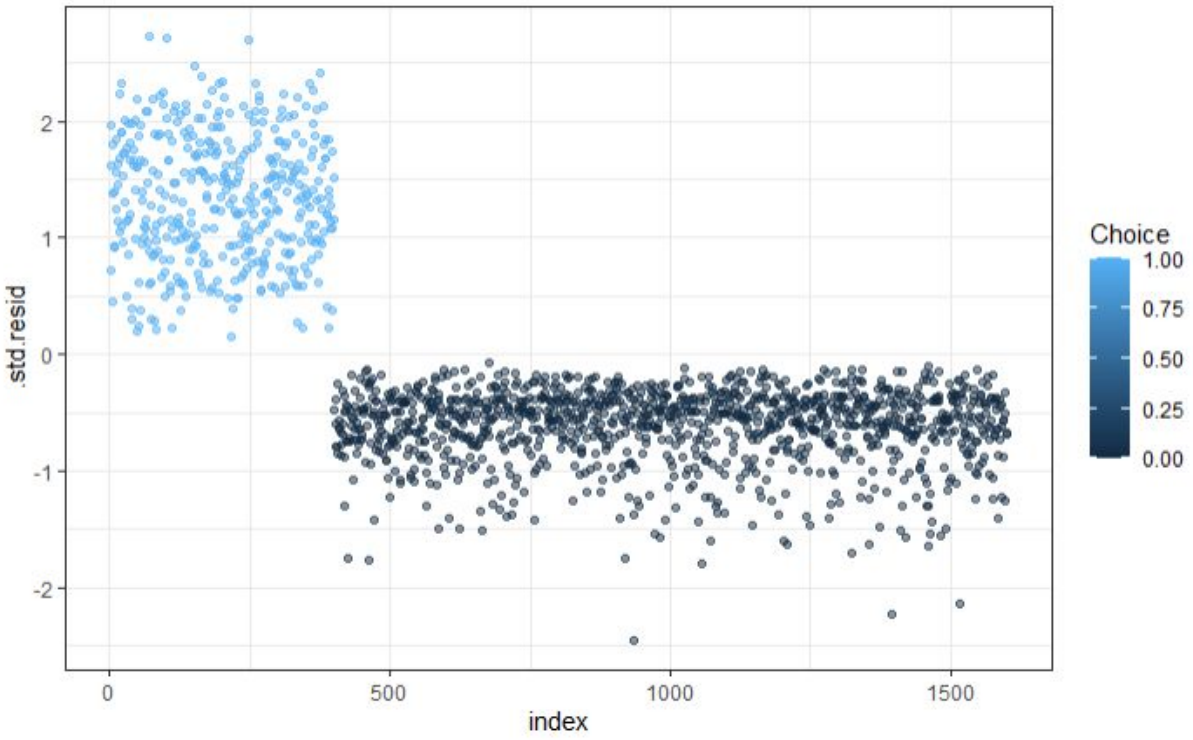
```
plot(glm3, which = 4)
```

DA6813
Mason Schweder (cko441)
Roddy Caldwell (tcg315)
Joshua Gutierrez (eqt138)
Nancy Lopez Ramirez (hpn459)




```
model.data <- augment(glm3) %>%
  mutate(index = 1:n())
model.data %>% top_n(3, .cooksd)
ggplot(model.data, aes(index, .std.resid)) +
  geom_point(aes(color = Choice), alpha = .5) +
  theme_bw()
```

DA6813
Mason Schweder (cko441)
Roddy Caldwell (tcg315)
Joshua Gutierrez (eqt138)
Nancy Lopez Ramirez (hpn459)



Using these tests, we see that there are no influential points, which means our model satisfies this assumption.

4) There are no high intercorrelations (i.e. multicollinearity) among the predictors.

```
car::vif(glm3)
```

Gender	Amount_purchased	Frequency	P_Child
1.020262	1.204294	1.015220	1.208836

P_Cook	P_DIY	P_Art
1.221915	1.169499	1.225022

These values are well below the threshold of acceptable collinearity, meaning that our model satisfies this assumption as well.

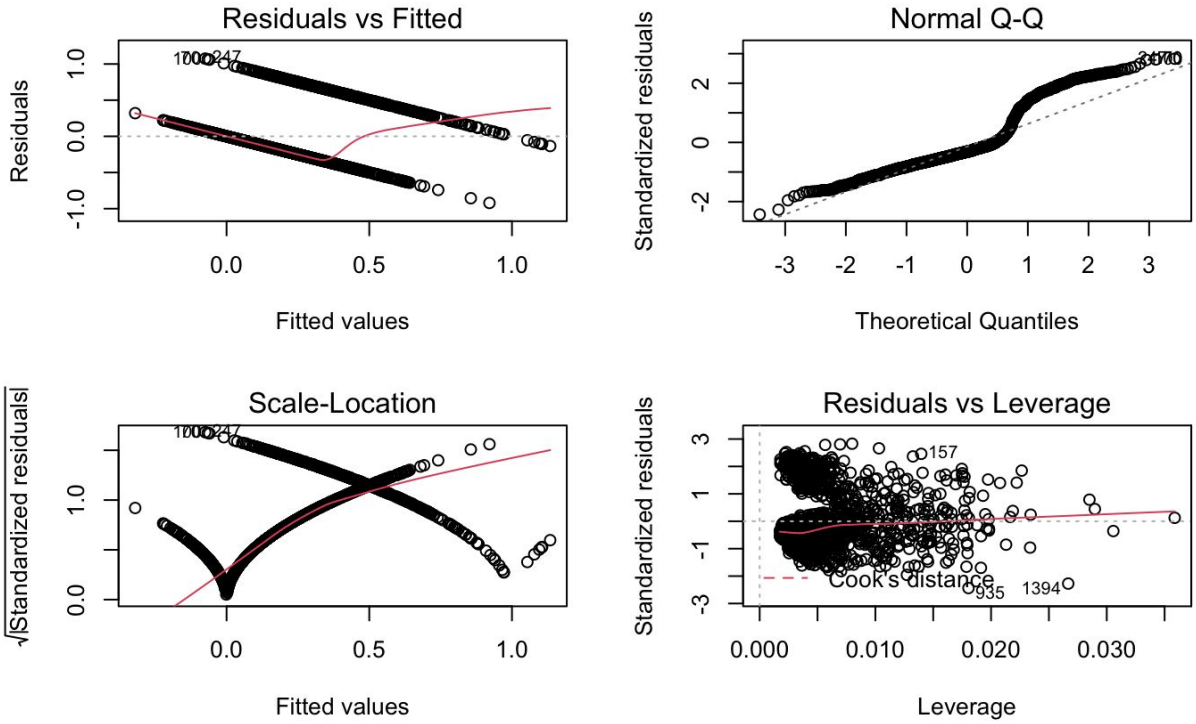
DA6813
Mason Schweder (cko441)
Roddy Caldwell (tcg315)
Joshua Gutierrez (eqt138)
Nancy Lopez Ramirez (hpn459)

Multiple Linear Regression

```
Call:
lm(formula = Choice ~ . - First_purchase - Observation - P_Youth,
    data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9211 -0.2495 -0.1187  0.1429  1.0750

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.3682284  0.0299539  12.293  < 2e-16 ***
Gender           -0.1293787  0.0201563  -6.419 1.81e-10 ***
Amount_purchased  0.0002736  0.0001118   2.447   0.0145 *
Frequency        -0.0111288  0.0012213  -9.112  < 2e-16 ***
Last_purchase     0.0541629  0.0091969   5.889 4.72e-09 ***
P_Child          -0.0909394  0.0146050  -6.227 6.09e-10 ***
P_Cook           -0.1068639  0.0148103  -7.215 8.28e-13 ***
P_DIY            -0.1039742  0.0187033  -5.559 3.17e-08 ***
P_Art             0.1528195  0.0177456   8.612  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3815 on 1591 degrees of freedom
Multiple R-squared:  0.2282,      Adjusted R-squared:  0.2244
F-statistic: 58.82 on 8 and 1591 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression - Diagnostic Plots

DA6813
Mason Schweder (cko441)
Roddy Caldwell (tcg315)
Joshua Gutierrez (eqt138)
Nancy Lopez Ramirez (hpn459)



Multiple Linear Regression - Confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2011   85
         1  148   56

               Accuracy : 0.8987
                 95% CI : (0.8856, 0.9107)
    No Information Rate : 0.9387
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2718

 Mcnemar's Test P-Value : 4.871e-05

            Sensitivity : 0.9314
            Specificity : 0.3972
         Pos Pred Value : 0.9594
         Neg Pred Value : 0.2745
```

```
            Prevalence : 0.9387
        Detection Rate : 0.8743
  Detection Prevalence : 0.9113
     Balanced Accuracy : 0.6643

      'Positive' Class : 0
```

Multiple Linear Regression - Code

```
#Loading the data
train = read_xlsx('BBBC-Train.xlsx')
test = read_xlsx('BBBC-Test.xlsx')
#Exploring the data
dim(train)
dim(test)
names(train)
summary(train)
summary(train$Choice)
hist(train$Choice)
sum(train$Choice == '1')
sum(train$Choice == '0')
### The train data set has n = 1600, where Choice=1 in 400 and Choice=0 in 1200.
### The response variables values are somewhat unbalanced.
hist(train$Amount_purchased)
hist(train$Frequency)
hist(train$Last_purchase)
linear.mod = lm(Choice~.-First_purchase -Observation - P_Youth, data=train)
summary(linear.mod)
par(mfrow = c(2,2))
plot(linear.mod)
cor(train)
pred.linear.mod = predict(linear.mod, newdata=test)
summary(pred.linear.mod)
pred.linearmodel = ifelse(pred.linear.mod>= 0.5, "1", "0")
caret::confusionMatrix(as.factor(test$Choice), as.factor(pred.linearmodel))
```

Read in the training and testing datasets from xlsx

```{r}
set.seed(12345)
library(caret)
library(readxl)
```

```{r}
```

```
bb.training <- read_xlsx("C:/Users/Mason/Documents/Analytics
Applications/Assignment2/BBBC-Train.xlsx")
bb.testing <- read_xlsx("C:/Users/Mason/Documents/Analytics
Applications/Assignment2/BBBC-Test.xlsx")
```

```{r}
# head(bb.training)
#Converting Choice to a factor variable (requirement for running caret logistic model)
bb.training$Choice = as.factor(bb.training$Choice)
bb.testing$Choice = as.factor(bb.testing$Choice)
```

```{r}
## Scale variables
bb.training$Amount_purchased = scale(bb.training$Amount_purchased)
bb.training$Frequency = scale(bb.training$Frequency)
bb.training$Last_purchase = scale(bb.training$Last_purchase)
bb.training$First_purchase = scale(bb.training$First_purchase)
bb.training$P_Child = scale(bb.training$P_Child)
bb.training$P_Youth = scale(bb.training$P_Youth)
bb.training$P_Cook = scale(bb.training$P_Cook)
bb.training$P_DIY = scale(bb.training$P_DIY)
bb.training$P_Art = scale(bb.training$P_Art)
bb.training$Gender = as.factor(bb.training$Gender)
bb.testing$Gender = as.factor(bb.testing$Gender)
bb.testing$Amount_purchased = scale(bb.testing$Amount_purchased)
bb.testing$Frequency = scale(bb.testing$Frequency)
bb.testing$Last_purchase = scale(bb.testing$Last_purchase)
bb.testing$First_purchase = scale(bb.testing$First_purchase)
bb.testing$P_Child = scale(bb.testing$P_Child)
bb.testing$P_Youth = scale(bb.testing$P_Youth)
bb.testing$P_Cook = scale(bb.testing$P_Cook)
bb.testing$P_DIY = scale(bb.testing$P_DIY)
bb.testing$P_Art = scale(bb.testing$P_Art)
```

```{r}
library(e1071)
tuned <-
tune.svm(Choice~Gender+Amount_purchased+Frequency+Last_purchase+First_purchase+P_
Child+P_Youth+P_Cook+P_DIY+P_Art, data=bb.training, gamma=seq(.01, 0.1, by=.01),
cost=seq(.1, 1, by=.1), scale=TRUE)
##run SVM with the tuned cost and gamma
```

```
resultsSVM <-
svm(formula=Choice~Gender+Amount_purchased+Frequency+Last_purchase+First_purchase
+P_Child+P_Youth+P_Cook+P_DIY+P_Art, data=bb.training,
gamma=tuned$best.parameters$gamma, cost=tuned$best.parameters$cost)
##make predictions
predSVM<- predict(resultsSVM, bb.testing, type="response")
##compute the confusion matrix
confusionMatrix(data=predSVM, bb.testing$Choice)
```

```{r}
library(rminer)
m =
fit(Choice~Gender+Amount_purchased+Frequency+Last_purchase+First_purchase+P_Child+P
_Youth+P_Cook+P_DIY+P_Art, data = as.data.frame(bb.training), model = "svm", kpar =
list(sigma = 0.10), C = 2)
Importance(m, data = as.data.frame(bb.training))
```

DA6813
Mason Schweder (cko441)
Roddy Caldwell (tcg315)
Joshua Gutierrez (eqt138)
Nancy Lopez Ramirez (hpn459)

REFERENCES

1. Stripling E, vanden Broucke S, Antonio K, Baesens B, Snoeck M. Profit maximizing logistic model for customer churn prediction using genetic algorithms. Swarm and Evolutionary Computation. 2018 Jun 1;40:116-30.
2. Kanisetty, Sai. Logistic Regression in Python to evaluate profitability of Sales-Marketing System. 2018 Jan 2. Available from: https://towardsdatascience.com/logistic-regression-in-python-to-evaluate-profitability-of-sales-marketing-system-fe2261964fa4
3. Kassambara, Alboukadel. Logistic Regression Assumptions and Diagnostics in R. Statistical tools for high-throughput data analysis. 2018 Nov 3. Available from: http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/
4. Bhalla, Deepanshu. WHEN AND WHY TO STANDARDIZE A VARIABLE. [cited 2020 Sep 14]. Available from: https://www.listendata.com/2017/04/how-to-standardize-variable-in-regression.html
5. Marketing Automation Simplified: The Small Guide to Big Ideas. Oracle Eloqua. [cited 2020 Sep 14]. Available from: https://static1.squarespace.com/static/51b949f4e4b0c43b09f8b97f/t/570301ceb6aa607cbb9a534f/1459814863739/O-Eloqua_Marketing_Automation_Simplified_eBook.pdf
6. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning with applications in R*. New York: Springer.