Casestudy 2 - BookBinders

Vishwa Koppisetti

January 30, 2019

1. Executive Summary

The Bookbinders Book Club (BBBC) is exploring whether to use predictive modeling approaches to improve the efficacy of its direct mail program. For a recent mailing they had a 9.03% response rate (1806 orders) for the purchase of a book, whose brochure they included with regular mailing of a selected 20,000 customers. BBBC then developed a database to calibrate a response model to identify the factors that influences these purchases

For further analysis to identify the following to evaluate its direct mail campaigns, we have a subset of the BBBC database available.

* The factors that most influenced customers to buy the book
* A highly accurate model among Ordinary Linear Regression, a logit model, and support vector machines

Summary of Data in Appendix A and B

Linear Regression: We run linear Regression without ‘observation’ variable and find all variables except First\_purchase to be significant. Among the significant ones, except Amount\_purchased, Last\_purchase and P\_Art all others seem to have a negative relationship with response.

This model helps in easy interpretation of how each variable effects the response however the R-squared and F-statistic are very low.

Since the assumption of linear regression is to have a continuous response variable with errors normally distributed, using a binary response violated this assumption and gives invalid inferences, standard errors and therefore confidence intervals.

For prediction part modeling binary outcome with linear regression we could get predicted values for some observations outside the range (0,1) (the range expected when considering probabilities for binary response), which is again invalid. We therefore need a better model for accuracy given the nature of the response variable, though the interpretation is better with this model in general.

Logistic Regression: We need to check two assumptions.

* No Multicollinearity: Last\_purchase has a VIF greater than 10 (17.7) and should be removed from the model since the information that this variable provides about the response is redundant in the presence of the other variables
* linear relationship between explanatory variable and log odds of response variable: We run a GAM function and conclude from the “Anova for nonparametric effects” table that none of the variables are significantly non-linear

The logistic regression model is fit with all variables except observation, Last\_purchase (to avoid multicollinearity).

All the predictors except P\_Youth have a significant relationship with probability of purchase.

Here the interpretation is not as straightforward and simple as in linear regression and is in terms of log odds of the response variable. However, we get a prediction accuracy of 89.57 which is 0.57% greater than when unaccounted for multicollinearity

When the model is used on the prediction dataset that has 2300 observation, we found that about 8% (182 customers) are likely to purchase the book

Support Vector Machine: the response variable was initially provided as numeric so has be converted into a factor variable. We run SVM with all except “observation” variable for 2 types of Kernels to check for any variation in accuracy.

Linear Kernel: The accuracy is 91 which is better that of logistic regression but not in terms of interpretation. There is a trade off between accuracy vs interpretation for logistic and SVM models; logistic having relatively lower accuracy but better interpretation and SVM providing better accuracy but hard to interpret. When the model is used on the prediction dataset that has 2300 observation, we predict that about 5% (117 customers) are likely to purchase the book

Radial Kernel: The accuracy is very slightly better than linear SVM so it would be better to go with the linear SVM since radial SVM is time-consuming only to provide almost the same accuracy. When the model is used on the prediction dataset that has 2300 observation, we predict that about 3% (72 customers) are likely to purchase the book

1. The Problem
2. Introduction/Background

Superstores especially online superstores such as Amazon are putting intense competitive pressure on book clubs, mail-order firms and retail outlets. In response to these pressures, book clubs are starting to look at alternative business models that will make them more responsive to their customer’s preferences

The Bookbinders Book Club (BBBC) is exploring whether to use predictive modeling approaches to improve the efficacy of its direct mail program. For a recent mailing, the company selected 20,000 customers in Pennsylvania, New York, and Ohio from its database and included with their regular mailing a specially produced brochure for the book *The Art History of Florence*. This resulted in a 9.03% response rate (1806 orders) for the purchase of the book. BBBC then developed a database to calibrate a response model to identify the factors that influences these purchases.

BBBC is evaluating three different modeling methods: ordinary linear regression, a logit model, and support vector machines. While they want to isolate the factors that most influenced customers to buy the book, they also want to develop a highly accurate model.

1. Purpose of study/importance of study/statement of problem

* Evaluate three different modeling methods: ordinary linear regression, a logit model, and support vector machines
* To determine the factors that most influenced customers to buy the book.
* To develop a highly accurate model

1. Questions to be answered/conceptual statement of hypotheses

* Which of the three; ordinary linear regression, a logit model, and support vector machines is best suited for this purpose of study
* Which factors have a significant role in purchase
* What are the advantages of one model over another; Which model provides better interpretation. Which model provides better prediction accuracy
* Are the assumptions of the model met
* Based on the model, which customers should Bookbinders target? How much more profit would you expect the company to generate using these models as compare to sending the mail offer to the entire list.

1. Outline of remainder of report (brief)

* Procedure followed to model the response variable on train set using for the different models.
* Assessing the performance of the model using test set
* Predicting the % of customers that are likely to purchase using prediction set

1. Review of Related Literature
   1. Acquaint reader with existing methodologies used in this area.

Book clubs are also beginning to use database marketing techniques to work smarter rather than expand the coverage of their mailings

Doubleday uses modeling techniques to look at more than 80 variables, including geography and the type of books customer’s purchase, and selects three to five variables that are most influential predictors. (1: DM News, May 23, 1994. n.d.)

The Bookbinders Book Club (BBBC) developed a database to calibrate a response model to identify the factors that influences these purchases.

BBBC is evaluating three different modeling methods: ordinary linear regression, a logit model, and support vector machines. While they want to isolate the factors that most influenced customers to buy the book, they also want to develop a highly accurate model.

(This case and data were developed by Professors Nissan Levin and Jacob Zahavi at Tel Aviv University. The materials have been adapted and further discussed in the book Marketing Engineering. n.d.)

1. Methodology
   1. Identification, classification and operationalization of variables.
   * 1600 observations of train set, 2300 observations of test set and prediction set
   * 12 variables with a binary dependent variable and 10 numeric independent variables. Classification and operationalization in Appendix A
   * All except response variable in the prediction set
   1. Statements of hypotheses being tested and/or models being developed.
   * To find a highly accurate model among Ordinary Linear Regression, a logit model, and support vector machines
   * To isolate the factors that most influence book purchase
   * To find a model with tradeoff between interpretability and accuracy
   1. Sampling techniques, if full data is not being used.

There is a separate dataset each for training the model, testing the model and for making predictions. Appendix B

* 1. Data collection process, including data sources, data size, etc. Primary/secondary?

(This case and data were developed by Professors Nissan Levin and Jacob Zahavi at Tel Aviv University. The materials have been adapted and further discussed in the book Marketing Engineering. n.d.)

For this case analysis, we will use a subset of the database available to BBBC as secondary data. It consists of data for 400 customers who purchased the book and 1200 customers who did not, thereby over-representing the response group. The dependent variable for the analysis is Choice – purchase or no purchase of the book. BBBC also selected several independent variables that it thought might explain the observed choice behavior.

* 1. Modeling analysis/techniques used

Different modelling techniques such as Ordinary Linear Regression, a logit model, and support vector machines are used to determine which model provides good accuracy along with interpretability

* 1. Methodological assumptions and limitations.

Since the assumption of linear regression is to have a continuous response variable with errors normally distributed, using a binary response violated this assumption and gives invalid inferences, standard errors and therefore confidence intervals.

For prediction part modeling binary outcome with linear regression we could get predicted values for some observations outside the range (0,1) (the range expected when considering probabilities for binary response), which is again invalid. We therefore need a better model for accuracy given the nature of the response variable, though the interpretation is better with this model.

Logistic Regression: We need to check two assumptions.

* + - No Multicollinearity: Last\_purchase has a VIF greater than 10 (17.7) and should be removed from the model since the information that this variable provides about the response is redundant in the presence of the other variables
    - linear relationship between explanatory variable and log odds of response variable: We run a GAM function and conclude from the “Anova for nonparametric effects” table that none of the variables are significantly non-linear

Limitations: Here the interpretation is not as straightforward and simple as in linear regression and is in terms of log odds of the response variable. However, we get a satisfactory prediction accuracy.

For SVM we choose radial over linear kernel if the predictive performance is observed to be much higher, over expense of time-consumption limitation of RBF

1. Data
   1. Data cleaning Appendix B

There are no missing values in any of the datasets. There are no invalid entrees for any of the features to be removed as outliers.

Data preprocessing Appendix B

The gender variable is converted to a factor of 2 levels: Male (1), Female (0). Response variable is made factor later during SVM modelling. To avoid multicollinearity for logistic regression Last\_purchase was removed but it led to only a 0.5% increase in accuracy

* 1. Data Limitations

There were no missing values in the dataset, but the number of observations is less, and data is unbalanced in terms of response class. Last\_purchase was highly correlated with other predictors so was removed since it provides redundant information about the response in the presence of the other variables. This led to fewer predictors than already available

1. Findings (Results)
   1. Results presented in tables or charts when appropriate

Qualitative analysis of relationship between response and explanatory variables through boxplots in Appendix C shows that while Frequency, Amount\_purchased, Last\_purchase, First\_purchase seem to have an effect on the purchase of the book, other variables do not seem to effect as much.

The frequency of gender variable from the summary table, Appendix B shows that female customers half as much as male customers.

The outliers technically detected for predictors contain information important to be retained for modelling hence not removed

There exists multicollinearity, Appendix E, with Last\_purchase being highly correlated with predictor variables and hence removed while modelling logistic regression

No significant nonlinear effects as observed in the ‘Anova for non-parametric table’ output from GAM, Appendix E

* 1. Results reported with respect to hypotheses/models.

Linear Regression: We run linear Regression without ‘observation’ variable and find all variables except First\_purchase to be significant. Among the significant ones, except Amount\_purchased, Last\_purchase and P\_Art all others seem to have a negative relationship with response. Appendix D

The interpretation is easy but doesn’t make much sense since the response is whether book is purchased or not

Interpretation: Gender has a negative relationship with response, its coefficient of regression is -0.13 which means that female customers purchase more than male customers by a factor of 0.13

Amount\_purhased has a positive relationship with response, its coefficient of regression is very small, 0.0003, which means for 1 unit increase in Total amount spent on BBBC books, the purchase increases by a factor of 0.0003, other predictors held constant. We have similar interpretation for P\_Art which has a positive relationship with response

Frequency has a negative relationship with response, its coefficient of regression is -------0.009, which means for 1 unit increase in Frequency, the purchase decreases by a factor of 0.009, other predictors held constant.

Last\_purchase has a positive relationship with response, its coefficient of regression is 0.1, which means for every 1 month increase since last purchase, this book purchase increases by a factor of 0.1, other predictors held constant.

P\_Child has a negative relationship with response, its coefficient of regression is -0.13, which means for 1 additional children’s book purchased, this book purchase decreases by a factor of 0.13, other predictors held constant. We have similar interpretations for P\_Youth, P\_Cook, P\_DIY which also share a negative relationship with being purchased

The R-squared and F-statistic are very low; The model could explain only 24% of the variation in response.

Since the assumption of linear regression is to have a continuous response variable with errors normally distributed, using a binary response violated this assumption and gives invalid inferences, standard errors and therefore confidence intervals.

For prediction part modeling binary outcome with linear regression we could get predicted values for some observations outside the range (0,1) (the range expected when considering probabilities for binary response), which is again invalid. We therefore need a better model for accuracy given the nature of the response variable, though the interpretation is better with this model in general.

Logistic Regression: We need to check two assumptions. Appendix E

* No Multicollinearity: Last\_purchase has a VIF greater than 10 (17.7) and should be removed from the model since the information that this variable provides about the response is redundant in the presence of the other variables
* linear relationship between explanatory variable and log odds of response variable: We run a GAM function and conclude from the “Anova for nonparametric effects” table that none of the variables are significantly non-linear

The logistic regression model is fit with all variables except observation, Last\_purchase (to avoid multicollinearity). Appendix F

All the predictors except P\_Youth have a significant relationship with probability of purchase. This means number of youth books purchased doesn’t affect the purchase of this book

Here the interpretation is not as straightforward and simple as in linear regression and is in terms of log odds of the response variable. However, we get a prediction accuracy of 89.57 which is 0.57% greater than when unaccounted for multicollinearity

Interpretation: Gender: The coefficient is -0.83, Appendix F which corresponds to the log of odds ratio between the male group and female group. The odds ratio equals 0.44 which means the odds for males are about 56% lower than the odds for females in purchasing the book.

Amount\_purchased: The coefficient is 0.002 which corresponds to odd ratio of 1.002 and shows that it has a positive relationship with response. For every additional dollar spent out of the total money to purchase BBBC books, we expect to see 0.2% increase in odds of purchasing the book, The Art History of Florence, other predictors held constant

P\_Child: The coefficient is -0.34 which corresponds to odd ratio of 0.71 and shows that it has a negative relationship with response. For every 1 children’s book purchased, we expect to see 29% decrease in odds of purchasing the book, other predictors held constant. We have similar interpretations for P\_Cook, P\_DIY which also share a negative relationship with probability of being purchased

Frequency: The coefficient is -0.12 which corresponds to odd ratio of 0.89 and shows that it has a negative relationship with response. For every 1 unit increase in Frequency, we expect to see 11% decrease in odds of purchasing the book, other predictors held constant

P\_Art: The coefficient is 1.077 which corresponds to odd ratio of 2.94 and shows that it has a positive relationship with response. For every 1 art book purchased, we expect to see 194% increase in odds of purchasing the book, The Art History of Florence, other predictors held constant. We have similar interpretations for First\_purchase which also shares a positive relationship with probability of being purchased

When the model is used on the prediction dataset that has 2300 observation, we predict that about 8% (182 customers) are likely to purchase the book

Support Vector Machine: the response variable was initially provided as numeric so has be converted into a factor variable. We run SVM with all except “observation” variable for 2 types of Kernels to check for any variation in accuracy. Appendix G

* Linear Kernel: The accuracy is 90.57% which is better that of logistic regression but not in terms of interpretation. There is a tradeoff between accuracy vs interpretation for logistic and SVM models; logistic having relatively lower accuracy but better interpretation and SVM providing slightly better accuracy but hard to interpret.

When the model is used on the prediction dataset that has 2300 observation, we predict that about 5% (117 customers) are likely to purchase the book

* Radial Kernel: The accuracy is very slightly better than linear SVM, 90.96% so it would be better to go with the linear SVM since radial SVM is time-consuming only to provide almost the same accuracy. When the model is used on the prediction dataset that has 2300 observation, we predict that about 3% (72 customers) are likely to purchase the book

C. Factual information kept separate from interpretation, inference and evaluation.

For a recent mailing, the company selected 20,000 customers in Pennsylvania, New York, and Ohio from its database and included with their regular mailing a specially produced brochure for the book The Art History of Florence. This resulted in a 9.03% response rate (1806 orders) for the purchase of the book. BBBC then developed a database to calibrate a response model to identify the factors that influences these purchases.

1. Conclusions and Recommendations

We can conclude that though interpretation is easiest among the three for linear regression, it is not suitable in terms of nature of response variable and the prediction results, resulting outside range (0,1). SVM is known for providing better accuracy but is hard to interpret and since there was not much improvement in accuracy on switching from logistic to SVM, I recommend logistic regression to be used which provides a better balance between linear regression and SVM in terms of interpretability and accuracy respectively. However, if accuracy is of top priority then linear SVM should be adopted over Radial(nonlinear) SVM since both have the same accuracy for this dataset and linear runs faster.

As per the model recommended (logistic) and per its prediction results of 8% of customers likely to make a purchase, in that 8%, the company should target female customers (gender being positive and male as having negative relationship) who have made their first purchase quite a while ago (First\_purchase being significant and having positive relationship) who buy art books(P\_Art being significant and having positive relationship) and are among the group that spends greater amounts in total while buying BBBC books(Amount\_purchased being significant and having positive relationship).

For the recommended model only 8% (182 customers) of 2300 customers were predicted to buy the book. Since mailing cost is $0.65/addressee if we send the mail offer to the entire list of 2300 it would be $1495 but if we send it to only 182 members who are likely to purchase, we spend only $118.3 saving about $1377.

Furthermore, if 182 people purchase the book for selling price $31.95 totaling up to $5814.9 out of which the company profits $10.2/addressee [31.95- (15+15\*0.45)], which is about $1856.4.

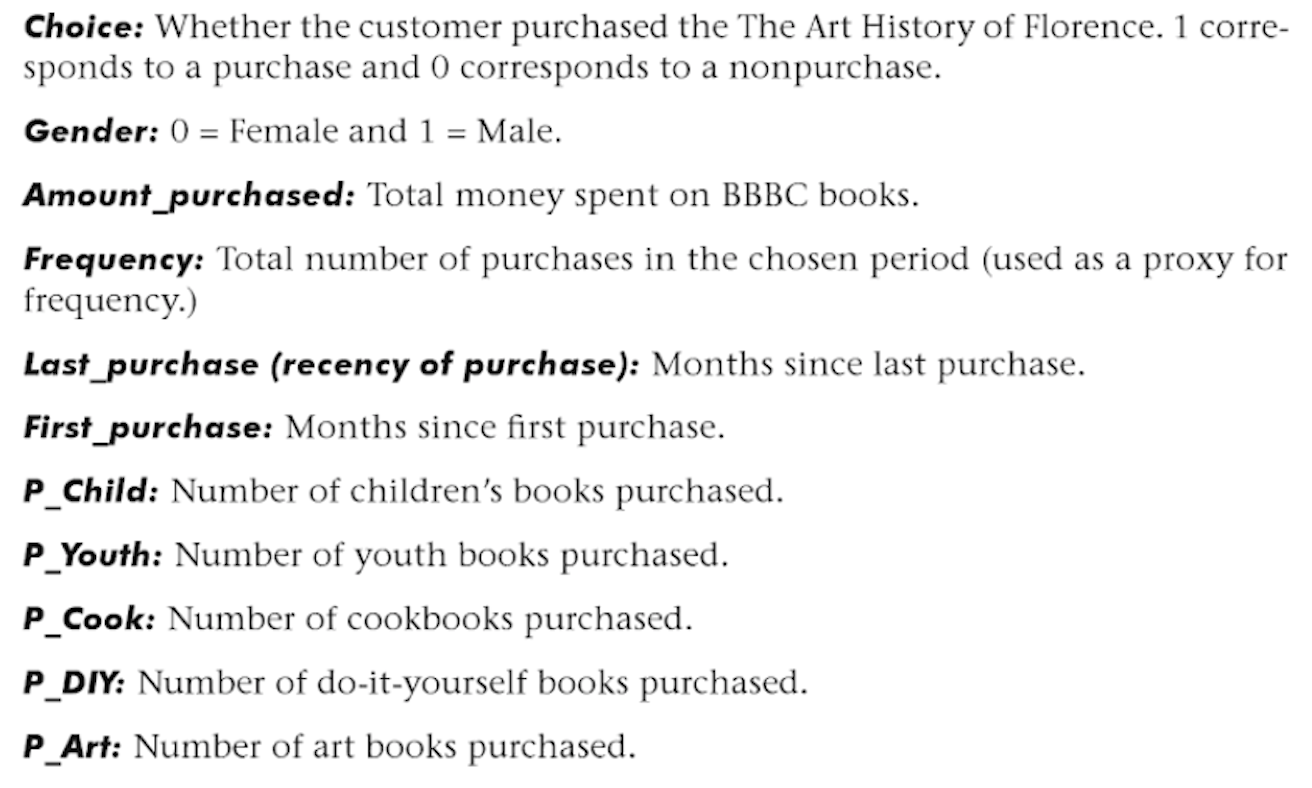
So, the company spends $118 and profits $1856.4 when sending to only 182 customers. $1738 profit

If it sends mail to all customers and only 182 purchase, though it profits $1856.4 from selling the books, if will spend $1495 to mail everyone this offset the profit and leaves only $361 profit

Appendix A

str(booktrain)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 1600 obs. of 12 variables:  
## $ Observation : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ Choice : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ Gender : num 1 1 1 1 0 1 1 0 1 1 ...  
## $ Amount\_purchased: num 113 418 336 180 320 268 198 280 393 138 ...  
## $ Frequency : num 8 6 18 16 2 4 2 6 12 10 ...  
## $ Last\_purchase : num 1 11 6 5 3 1 12 2 11 7 ...  
## $ First\_purchase : num 8 66 32 42 18 4 62 12 50 38 ...  
## $ P\_Child : num 0 0 2 2 0 0 2 0 3 2 ...  
## $ P\_Youth : num 1 2 0 0 0 0 3 2 0 3 ...  
## $ P\_Cook : num 0 3 1 0 0 0 2 0 3 0 ...  
## $ P\_DIY : num 0 2 1 1 1 0 1 0 0 0 ...  
## $ P\_Art : num 0 3 2 1 2 0 2 0 2 1 ...



Appendix B

library(readxl)  
booktrain = readxl::read\_excel(here::here("BBBC-Train.xlsx"))  
booktest = readxl::read\_excel(here::here("BBBC-Test.xlsx"))  
bookpredict = readxl::read\_excel(here::here("BBBC-Predict.xlsx"))

#Check for missing values  
anyNA(booktrain)

## [1] FALSE

anyNA(booktest)

## [1] FALSE

anyNA(bookpredict)

## [1] FALSE

#Convert gender to factor variable  
booktrain$Gender = as.factor(booktrain$Gender)  
booktest$Gender = as.factor(booktest$Gender)  
bookpredict$Gender = as.factor(bookpredict$Gender)

summary(booktrain)

## Observation Choice Gender Amount\_purchased Frequency   
## Min. : 1.0 Min. :0.00 0: 546 Min. : 15.0 Min. : 2.00   
## 1st Qu.: 400.8 1st Qu.:0.00 1:1054 1st Qu.:126.8 1st Qu.: 6.00   
## Median : 800.5 Median :0.00 Median :203.0 Median :12.00   
## Mean : 800.5 Mean :0.25 Mean :200.9 Mean :12.31   
## 3rd Qu.:1200.2 3rd Qu.:0.25 3rd Qu.:273.0 3rd Qu.:16.00   
## Max. :1600.0 Max. :1.00 Max. :474.0 Max. :36.00   
## Last\_purchase First\_purchase P\_Child P\_Youth   
## Min. : 1.000 Min. : 2.00 Min. :0.0000 Min. :0.0000   
## 1st Qu.: 1.000 1st Qu.:12.00 1st Qu.:0.0000 1st Qu.:0.0000   
## Median : 2.000 Median :18.00 Median :0.0000 Median :0.0000   
## Mean : 3.199 Mean :22.58 Mean :0.7394 Mean :0.3375   
## 3rd Qu.: 4.000 3rd Qu.:30.00 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :12.000 Max. :96.00 Max. :8.0000 Max. :4.0000   
## P\_Cook P\_DIY P\_Art   
## Min. :0.00 Min. :0.0000 Min. :0.000   
## 1st Qu.:0.00 1st Qu.:0.0000 1st Qu.:0.000   
## Median :0.00 Median :0.0000 Median :0.000   
## Mean :0.76 Mean :0.3912 Mean :0.425   
## 3rd Qu.:1.00 3rd Qu.:1.0000 3rd Qu.:1.000   
## Max. :6.00 Max. :4.0000 Max. :5.000

Appendix C

par(mfrow=c(3,3))   
boxplot(Frequency ~ Choice, data = booktrain, main = "Frequency vs Purchase")  
boxplot(Amount\_purchased ~ Choice, data = booktrain, main = "Amount\_purchased vs Purchase")  
boxplot(Last\_purchase ~ Choice, data = booktrain, main = "Last\_purchasevs Purchase")  
boxplot(First\_purchase ~ Choice, data = booktrain, main = "First\_purchase vs Purchase")  
boxplot(P\_Child ~ Choice, data = booktrain, main = "P\_Child vs Purchase")  
boxplot(P\_Youth ~ Choice, data = booktrain, main = "P\_Youth vs Purchase")  
boxplot(P\_Cook ~ Choice, data = booktrain, main = "P\_Cook vs Purchase")  
boxplot(P\_DIY ~ Choice, data = booktrain, main = "P\_DIY vs Purchase")  
boxplot(P\_Cook ~ Choice, data = booktrain, main = "P\_Cook vs Purchase")

# 

Appendix D

# Linear Regression

book\_lm = lm(Choice ~ .-Observation, data = booktrain)  
summary(book\_lm)

##   
## Call:  
## lm(formula = Choice ~ . - Observation, data = booktrain)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.9603 -0.2462 -0.1161 0.1622 1.0588   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.3642284 0.0307411 11.848 < 2e-16 \*\*\*  
## Gender1 -0.1309205 0.0200303 -6.536 8.48e-11 \*\*\*  
## Amount\_purchased 0.0002736 0.0001110 2.464 0.0138 \*   
## Frequency -0.0090868 0.0021791 -4.170 3.21e-05 \*\*\*  
## Last\_purchase 0.0970286 0.0135589 7.156 1.26e-12 \*\*\*  
## First\_purchase -0.0020024 0.0018160 -1.103 0.2704   
## P\_Child -0.1262584 0.0164011 -7.698 2.41e-14 \*\*\*  
## P\_Youth -0.0963563 0.0201097 -4.792 1.81e-06 \*\*\*  
## P\_Cook -0.1414907 0.0166064 -8.520 < 2e-16 \*\*\*  
## P\_DIY -0.1352313 0.0197873 -6.834 1.17e-11 \*\*\*  
## P\_Art 0.1178494 0.0194427 6.061 1.68e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3788 on 1589 degrees of freedom  
## Multiple R-squared: 0.2401, Adjusted R-squared: 0.2353   
## F-statistic: 50.2 on 10 and 1589 DF, p-value: < 2.2e-16

# Logistic Regression

book\_log = glm(formula = Choice ~ .-Observation, data = booktrain, family = binomial)  
  
  
summary(book\_log)

##   
## Call:  
## glm(formula = Choice ~ . - Observation, family = binomial, data = booktrain)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.38586 -0.66728 -0.43696 -0.02242 2.72238   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.3515281 0.2143839 -1.640 0.1011   
## Gender1 -0.8632319 0.1374499 -6.280 3.38e-10 \*\*\*  
## Amount\_purchased 0.0018641 0.0007918 2.354 0.0186 \*   
## Frequency -0.0755142 0.0165937 -4.551 5.35e-06 \*\*\*  
## Last\_purchase 0.6117713 0.0938127 6.521 6.97e-11 \*\*\*  
## First\_purchase -0.0147792 0.0128027 -1.154 0.2483   
## P\_Child -0.8112489 0.1167067 -6.951 3.62e-12 \*\*\*  
## P\_Youth -0.6370422 0.1433778 -4.443 8.87e-06 \*\*\*  
## P\_Cook -0.9230066 0.1194814 -7.725 1.12e-14 \*\*\*  
## P\_DIY -0.9058697 0.1437025 -6.304 2.90e-10 \*\*\*  
## P\_Art 0.6861124 0.1270176 5.402 6.60e-08 \*\*\*  
## ---  
## 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: 1392.2 on 1589 degrees of freedom  
## AIC: 1414.2  
##   
## Number of Fisher Scoring iterations: 5

Checking for the 2 model assumptions of multicollinearity and Linear relationship between explanatory variable and log odds of response

Appendix E

**Multicollinearity**

#install.packages("car")  
library(car)

## Warning: package 'car' was built under R version 3.5.2

## Loading required package: carData

mulcol = car::vif(book\_log);mulcol

## Gender Amount\_purchased Frequency Last\_purchase   
## 1.023359 1.232172 2.490447 17.706670   
## First\_purchase P\_Child P\_Youth P\_Cook   
## 9.247748 2.992269 1.761546 3.229097   
## P\_DIY P\_Art   
## 1.992698 1.938089

# Linear relationship between explanatory variable and log odds of response variable

library(gam)

## Loading required package: splines

## Loading required package: foreach

## Loaded gam 1.16

gam\_book = gam(Choice ~ Gender + s(Amount\_purchased) + s(Frequency) + s(Last\_purchase) + s(First\_purchase) + s(P\_Child) + s(P\_Youth) +s(P\_Cook) +s(P\_DIY) + s(P\_Art), data = booktrain, family=binomial,select = TRUE)  
summary(gam\_book)

##   
## Call: gam(formula = Choice ~ Gender + s(Amount\_purchased) + s(Frequency) +   
## s(Last\_purchase) + s(First\_purchase) + s(P\_Child) + s(P\_Youth) +   
## s(P\_Cook) + s(P\_DIY) + s(P\_Art), family = binomial, data = booktrain,   
## select = TRUE)  
## Deviance Residuals:  
## Min 1Q Median 3Q Max   
## -2.04928 -0.65347 -0.43696 -0.02308 2.58130   
##   
## (Dispersion Parameter for binomial family taken to be 1)  
##   
## Null Deviance: 1799.473 on 1599 degrees of freedom  
## Residual Deviance: 1368.961 on 1562 degrees of freedom  
## AIC: 1444.961   
##   
## Number of Local Scoring Iterations: 7   
##   
## Anova for Parametric Effects  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Gender 1 18.05 18.052 18.9679 1.416e-05 \*\*\*  
## s(Amount\_purchased) 1 13.84 13.835 14.5368 0.0001428 \*\*\*  
## s(Frequency) 1 49.73 49.730 52.2521 7.608e-13 \*\*\*  
## s(Last\_purchase) 1 11.92 11.922 12.5262 0.0004130 \*\*\*  
## s(First\_purchase) 1 0.01 0.015 0.0154 0.9013644   
## s(P\_Child) 1 17.28 17.277 18.1534 2.160e-05 \*\*\*  
## s(P\_Youth) 1 4.45 4.453 4.6783 0.0306965 \*   
## s(P\_Cook) 1 70.41 70.409 73.9801 < 2.2e-16 \*\*\*  
## s(P\_DIY) 1 70.68 70.682 74.2669 < 2.2e-16 \*\*\*  
## s(P\_Art) 1 29.39 29.392 30.8824 3.218e-08 \*\*\*  
## Residuals 1562 1486.60 0.952   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Anova for Nonparametric Effects  
## Npar Df Npar Chisq P(Chi)  
## (Intercept)   
## Gender   
## s(Amount\_purchased) 3 3.7321 0.2919  
## s(Frequency) 3 3.0778 0.3798  
## s(Last\_purchase) 3 1.0140 0.7979  
## s(First\_purchase) 3 1.7596 0.6238  
## s(P\_Child) 3 2.7637 0.4295  
## s(P\_Youth) 3 4.7069 0.1946  
## s(P\_Cook) 3 2.2032 0.5313  
## s(P\_DIY) 3 0.8153 0.8458  
## s(P\_Art) 3 4.6311 0.2009

Appendix F

#removing Last\_purchase  
book\_logm = glm(formula = Choice ~ .-Observation-Last\_purchase, data = booktrain, family = binomial)  
  
  
summary(book\_logm)

##   
## Call:  
## glm(formula = Choice ~ . - Observation - Last\_purchase, family = binomial,   
## data = booktrain)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.46171 -0.68074 -0.46620 -0.00855 2.80519   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.1489829 0.2095375 -0.711 0.477079   
## Gender1 -0.8302649 0.1350384 -6.148 7.83e-10 \*\*\*  
## Amount\_purchased 0.0022691 0.0007747 2.929 0.003399 \*\*   
## Frequency -0.1194992 0.0152620 -7.830 4.89e-15 \*\*\*  
## First\_purchase 0.0306235 0.0108454 2.824 0.004748 \*\*   
## P\_Child -0.3456948 0.0908420 -3.805 0.000142 \*\*\*  
## P\_Youth -0.1789417 0.1226235 -1.459 0.144489   
## P\_Cook -0.4578299 0.0950443 -4.817 1.46e-06 \*\*\*  
## P\_DIY -0.4265209 0.1209960 -3.525 0.000423 \*\*\*  
## P\_Art 1.0778036 0.1144995 9.413 < 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: 1437.0 on 1590 degrees of freedom  
## AIC: 1457  
##   
## Number of Fisher Scoring iterations: 5

mulcol1 = car::vif(book\_logm);mulcol1

# no variable with VIF > 10

## Gender Amount\_purchased Frequency First\_purchase   
## 1.021977 1.220305 2.173240 6.886806   
## P\_Child P\_Youth P\_Cook P\_DIY   
## 1.904631 1.320305 2.060140 1.462770   
## P\_Art   
## 1.603865

# accessing the performance of model with test set  
PredProbt = predict.glm(book\_logm, newdata = booktest, type="response")  
Predcht = ifelse(PredProbt >= 0.5,1,0)  
caret::confusionMatrix(as.factor(booktest$Choice),as.factor(Predcht))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 1987 109  
## 1 131 73  
##   
## Accuracy : 0.8957   
## 95% CI : (0.8824, 0.9079)  
## No Information Rate : 0.9209   
## P-Value [Acc > NIR] : 1.0000   
##   
## Kappa : 0.3215   
## Mcnemar's Test P-Value : 0.1752   
##   
## Sensitivity : 0.9381   
## Specificity : 0.4011   
## Pos Pred Value : 0.9480   
## Neg Pred Value : 0.3578   
## Prevalence : 0.9209   
## Detection Rate : 0.8639   
## Detection Prevalence : 0.9113   
## Balanced Accuracy : 0.6696   
##   
## 'Positive' Class : 0   
##

# predicting the % of customers that are likely to purchase using prediction set  
PredProbp = predict.glm(book\_logm, newdata = bookpredict, type = "response")  
Predchp = ifelse(PredProbp >= 0.5,1,0)  
p = table(Predchp)  
purchased = p[2]\*100/(p[1]+p[2]);purchased

## 1   
## 7.913043

Appendix G

# SVM

library(e1071)  
  
model\_svm = as.factor(Choice) ~ . -Observation  
book.tune=tune.svm(model\_svm,data=booktrain,kernel="linear",   
 cost=seq(0.1, 1, by = 0.1)) # checking best cost by 10 fold cv  
summary(book.tune$best.model)

##   
## Call:  
## best.svm(x = model\_svm, data = booktrain, cost = seq(0.1, 1,   
## by = 0.1), kernel = "linear")  
##   
##   
## Parameters:  
## SVM-Type: C-classification   
## SVM-Kernel: linear   
## cost: 1   
## gamma: 0.09090909   
##   
## Number of Support Vectors: 738  
##   
## ( 366 372 )  
##   
##   
## Number of Classes: 2   
##   
## Levels:   
## 0 1

book\_svm = svm(formula = model\_svm, data = booktrain,cost = book.tune$best.parameters$cost)  
summary(book\_svm)

##   
## Call:  
## svm(formula = model\_svm, data = booktrain, cost = book.tune$best.parameters$cost)  
##   
##   
## Parameters:  
## SVM-Type: C-classification   
## SVM-Kernel: radial   
## cost: 1   
## gamma: 0.09090909   
##   
## Number of Support Vectors: 810  
##   
## ( 366 444 )  
##   
##   
## Number of Classes: 2   
##   
## Levels:   
## 0 1

# accessing the performance of model with test set  
svm.test = predict(book\_svm, booktest)  
caret::confusionMatrix(as.factor(booktest$Choice),svm.test)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 2031 65  
## 1 152 52  
##   
## Accuracy : 0.9057   
## 95% CI : (0.893, 0.9173)  
## No Information Rate : 0.9491   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.2773   
## Mcnemar's Test P-Value : 5.281e-09   
##   
## Sensitivity : 0.9304   
## Specificity : 0.4444   
## Pos Pred Value : 0.9690   
## Neg Pred Value : 0.2549   
## Prevalence : 0.9491   
## Detection Rate : 0.8830   
## Detection Prevalence : 0.9113   
## Balanced Accuracy : 0.6874   
##   
## 'Positive' Class : 0   
##

# predicting the % of customers that are likely to purchase using prediction set  
svm.pred = predict(book\_svm, bookpredict)  
l = table(svm.pred);l

## svm.pred  
## 0 1   
## 2183 117

purchased\_svm = l[2]\*100/(l[1]+l[2]);purchased\_svm

## 1   
## 5.086957

## radial basis kernel

radial\_svm = tune.svm(model\_svm,data = booktrain, gamma = seq(.01,0.1, by = 0.01), cost = seq(0.1, 1, by = 0.1))  
radial\_book = svm(formula = model\_svm, data = booktrain, gamma = radial\_svm$best.parameters$gamma, cost=   
 radial\_svm$best.parameters$cost)  
  
summary(radial\_book)

##   
## Call:  
## svm(formula = model\_svm, data = booktrain, gamma = radial\_svm$best.parameters$gamma,   
## cost = radial\_svm$best.parameters$cost)  
##   
##   
## Parameters:  
## SVM-Type: C-classification   
## SVM-Kernel: radial   
## cost: 0.7   
## gamma: 0.03   
##   
## Number of Support Vectors: 777  
##   
## ( 379 398 )  
##   
##   
## Number of Classes: 2   
##   
## Levels:   
## 0 1

# accessing the performance of model with test set  
svm.radial.test = predict(radial\_book, booktest, type = "response")  
caret::confusionMatrix(as.factor(booktest$Choice),as.factor(svm.radial.test))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 2058 38  
## 1 170 34  
##   
## Accuracy : 0.9096   
## 95% CI : (0.8971, 0.921)  
## No Information Rate : 0.9687   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.2098   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.9237   
## Specificity : 0.4722   
## Pos Pred Value : 0.9819   
## Neg Pred Value : 0.1667   
## Prevalence : 0.9687   
## Detection Rate : 0.8948   
## Detection Prevalence : 0.9113   
## Balanced Accuracy : 0.6980   
##   
## 'Positive' Class : 0   
##

# predicting the % of customers that are likely to purchase using prediction set  
svm.radial.pred = predict(radial\_book, bookpredict, type = "response")  
r = table(svm.radial.pred);r

## svm.radial.pred  
## 0 1   
## 2228 72

purchased\_radial = r[2]\*100/(r[1]+r[2]);purchased\_radial

3.130434