MACHINE LEARNING ALGORITUMS TO PREDICT SALES

A Capstone Project submitted in partial fulfilment of the requirements for the award of the degree of

MASTER OF SCIENCE IN BUSINESS

By

Manpreet Singh

Register No: V00953330

Under the guidance of

Jason Merrick, Ph. D.

VCU School of Business

Virginia Commonwealth University

November 2022

DECLARATION

I, Manpreet Singh, do hereby declare that the project entitled '**MACHINE LEARNING ALGORITUMS TO PREDICT SALES**' has been undertaken by me for the award of the degree of Master of Science in Business. I have completed this study under the guidance of Jason Merrick, School of Business and Management, Virginia Commonwealth University, USA.

I declare that this project has not been submitted for the award of any degree, diploma, associateship or fellowship or any other title in this University or any other University. I also declare that the project is an original work and not an adoption of any other project/work.

Place: Richmond, Virginia

Date:

(Name & Signature of the Candidate)

Manpreet Singh

Register No: V00953330

EXECUTIVE SUMMARY

The Book Industry-Approximately 50,000 new titles, including new editions, are published each year in the United States, giving rise to a $25 billion industry in 2001. In terms of percentage of sales, this industry may be segmented as follows:

- 16% Textbooks
- 16% Trade books sold in bookstores
- 21% Technical, scientific, and professional books 10% Book clubs and other mail-order books
- 17% Mass-market paperbound books
- 20% All other books

Book retailing in the United States in the 1970s was characterized by the growth of bookstore chains located in shopping malls. The 1980s saw increased purchases in bookstores stimulated through the widespread practice of discounting. By the 1990s, the superstore concept of book retailing gained acceptance and contributed to double-digit growth of the book industry.

Conveniently situated near large shopping centers, superstores maintain large inventories of 30,000–80,000 titles and employ well-informed sales personnel. Book retailing changed fundamentally with the arrival of Amazon, which started out as an online bookseller and, as of 2015, was the world's largest online retailer of any kind. Amazon's margins were small and the convenience factor high, putting intense competitive pressure on all other book retailers. Borders, one of the two major superstore chains, discontinued operations in 2011.

Subscription-based book clubs offer an alternative model that has persisted, though it too has suffered from the dominance of Amazon.

Historically, book clubs offered their readers different types of membership programs. Two common membership programs are the continuity and negative option programs, which are both extended contractual relationships between the club and its members. Under a continuity program, a reader signs up by accepting an offer of several books for just a few dollars (plus shipping and handling) and an agreement to receive a shipment of one or two books each month thereafter at more-standard pricing. The continuity program is most common in the children's book market, where parents are willing to delegate the rights to the book club to make a selection, and much of the club's prestige depends on the quality of its selections. In a negative option program, readers get to select how many and which additional books they would like to receive. However, the club's selection of the month is delivered to them automatically unless they specifically mark "no" on their order form by a deadline date. Negative option programs sometimes result in customer dissatisfaction and always give rise to significant mailing and processing costs.

In an attempt to combat these trends, some book clubs have begun to offer books on a positive option basis, but only to specific segments of their customer base that are likely to be receptive to specific offers. Rather than expanding the volume and coverage of mailings, some book clubs are beginning to use database-marketing techniques to target customers more accurately. Information contained in their databases is used to identify who is most likely to be interested in a specific offer.

This information enables clubs to design special programs carefully tailored to meet their customer segments' varying needs.

About the Problem- Database Marketing at Charles

The Club: The Charles Book Club (CBC) was established in December 1986 on the premise that a book club could differentiate itself through a deep understanding of its customer base and by delivering uniquely tailored offerings. CBC focused on selling specialty books by direct marketing through a variety of channels, including media advertising (TV, magazines, newspapers) and mailing. CBC is strictly a distributor and does not publish any of the books that it sells. In line with its commitment to understanding its customer base, CBC built and maintained a detailed database about its club members. Upon enrollment, readers were required to fill out an insert and mail it to CBC. Through this process, CBC created an active database of 500,000 readers; most were acquired through advertising in specialty magazines.

The Problem: CBC sent mailings to its club members each month containing the latest offerings. On the surface, CBC appeared very successful: mailing volume was increasing, book selection was diversifying and growing, and their customer database was increasing. However, their bottom-line profits were falling. The decreasing profits led CBC to revisit their original plan of using database marketing to improve mailing yields and to stay profitable.

A Possible Solution: CBC embraced the idea of deriving intelligence from their data to allow them to know their customers better and enable multiple targeted campaigns where each target audience would receive appropriate mailings. CBC's management decided to focus its efforts on the most profitable customers and prospects, and to design targeted marketing strategies to best reach them. The two processes they had in place were:

1. Customer acquisition:

- New members would be acquired by advertising in specialty magazines, newspapers, and on TV.
- Direct mailing and telemarketing would contact existing club members.
- Every new book would be offered to club members before general advertising.

2. Data collection:

- All customer responses would be recorded and maintained in the database.
- Any information not being collected that is critical would be requested from the customer.

For each new title, they decided to use a two-step approach:

1. Conduct a market test involving a random sample of 4000 customers from the database to enable analysis of customer responses. The analysis would create and calibrate response models for the current book offering.
2. Based on the response models, compute a score for each customer in the database. Use this score and a cutoff value to extract a target customer list for direct-mail promotion.

Targeting promotions were considered to be of prime importance. Other opportunities to create successful marketing campaigns based on customer behavior data (returns, inactivity, complaints, compliments, etc.) would be addressed by CBC at a later stage.

Data Mining Techniques

Various data mining techniques can be used to mine the data collected from the market test. No one technique is universally better than another. The particular context and the particular characteristics of the data are the major factors in determining which techniques perform better in an application. For this assignment, we focus on four fundamental techniques:

- K Nearest Neighbors
- Classification Trees
- Random Forest
- Boosted Trees
- Logistic Regression

In the direct marketing business, the most commonly used variables are the RFM variables:

- R = recency, time since last purchase
- F = frequency, number of previous purchases from the company over a period
- M = monetary, amount of money spent on the company's products over a period

The assumption is that the more recent the last purchase, the more products bought from the company in the past, and the more money spent in the past buying the company's products, the more likely the customer is to purchase the product offered.

For *The Art History of Florence*, CBC wants to use the following all the explanatory variables (including R, F, and M) in the data to predict whether a customer would order the book.

The Art History of Florence

Submitted By

Manpreet Singh

Contents

Overview

In this section, the primary goal is to do Data partitioning and exploration, build a classifier to understand the model's prediction capability, such as k-Nearest Neighbors , Classification Tree, Random Forest, Boosted Trees, Logistic Regression and Final Recommendation.

Objective of the Study

In this study, we like to understand the model performance between the classifier and conclude the model performance. In this assignment, we will build a classifier that predicts whether the person has purchased the Florence or not.

About Data

The Art History of Florence is ready for release. CBC sent a test mailing to a random sample of 4000 customers from its customer base. The customer responses have been collated with past purchase data. Each row (or case) in the spreadsheet (other than the header) corresponds to one market test customer. Each column is a variable, with the header row giving the name of the variable. The variable names and descriptions are given below

Variable Name Description

Seq# Sequence number in the partition ID# Identification number

in the full (unpartitioned) market test dataset Gender

0=Male,1=Female

M Monetary—Total money spent on

books R Recency—Months since last

purchase

F Frequency—Total number of

purchases FirstPurch Months since first

purchase

ChildBks Number of purchases from the category child books

YouthBks Number of purchases from the category youth books

CookBks Number of purchases from the category cookbooks

DoItYBks Number of purchases from the category do-it-yourself

books RefBks Number of purchases from the category reference

books (atlases, encyclopedias, dictionaries)

ArtBks Number of purchases from the category art books

GeoBks Number of purchases from the category geography

books

ItalCook Number of purchases of book title Secrets of Italian

Cooking ItalAtlas Number of purchases of book title Historical

Atlas of Italy ItalArt Number of purchases of book title Italian

Art

Florence = 1 if The Art History of Florence was bought; = 0 if not

Data Sampling

In this section, we will divide the data into two parts, namely training and testing. Wherein training details used for building the model and testing data is used for validation.

Data Analysis

Now let's try to understand the distribution between customer responses. Each row in the spreadsheet corresponds to one market test customer, and the variable Florence = 0 if The Art History of Florence was not bought Florence = 1 if The Art History of Florence was bought that means responded to the mailing list. The response rate for a given group of customers means the proportion of customers that responded by buying The Art History of Florence.

Art History of Florence

1.Data partitioning and exploration

First of all, I read the data into csv file and then check whether there is any null value in the data. The output for null values in each column is given below:

```
  Gender         M         R         F FirstPurch   ChildBks   YouthBks
       0         0         0         0          0          0          0
CookBks   DoItYBks   RefBks   ArtBks   GeogBks   ItalCook  ItalAtlas
       0         0         0         0          0          0          0
ItalArt   Florence
       0         0
```

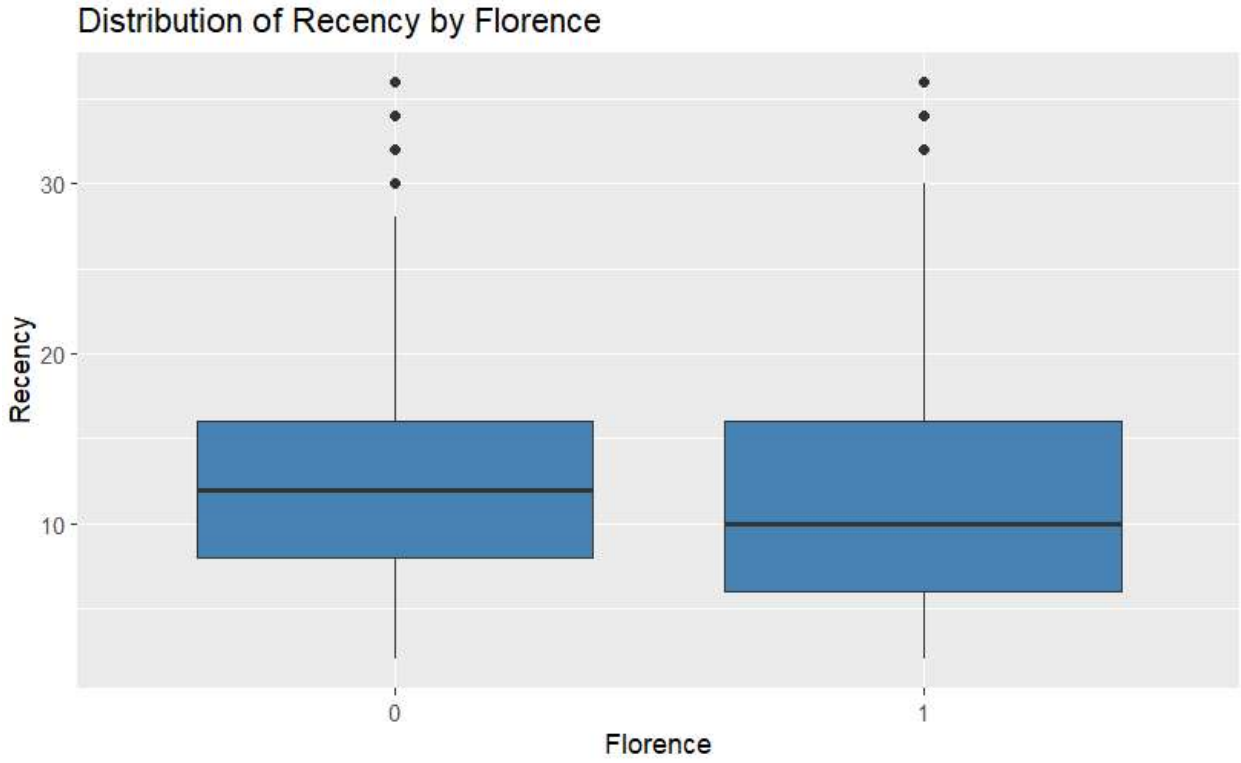
From above output, there is no null value in any of the column which means our dataset is already cleaned.

Response rate for the training data

The response for training data is 0.08708333.

Plot Response rate by Recency variable

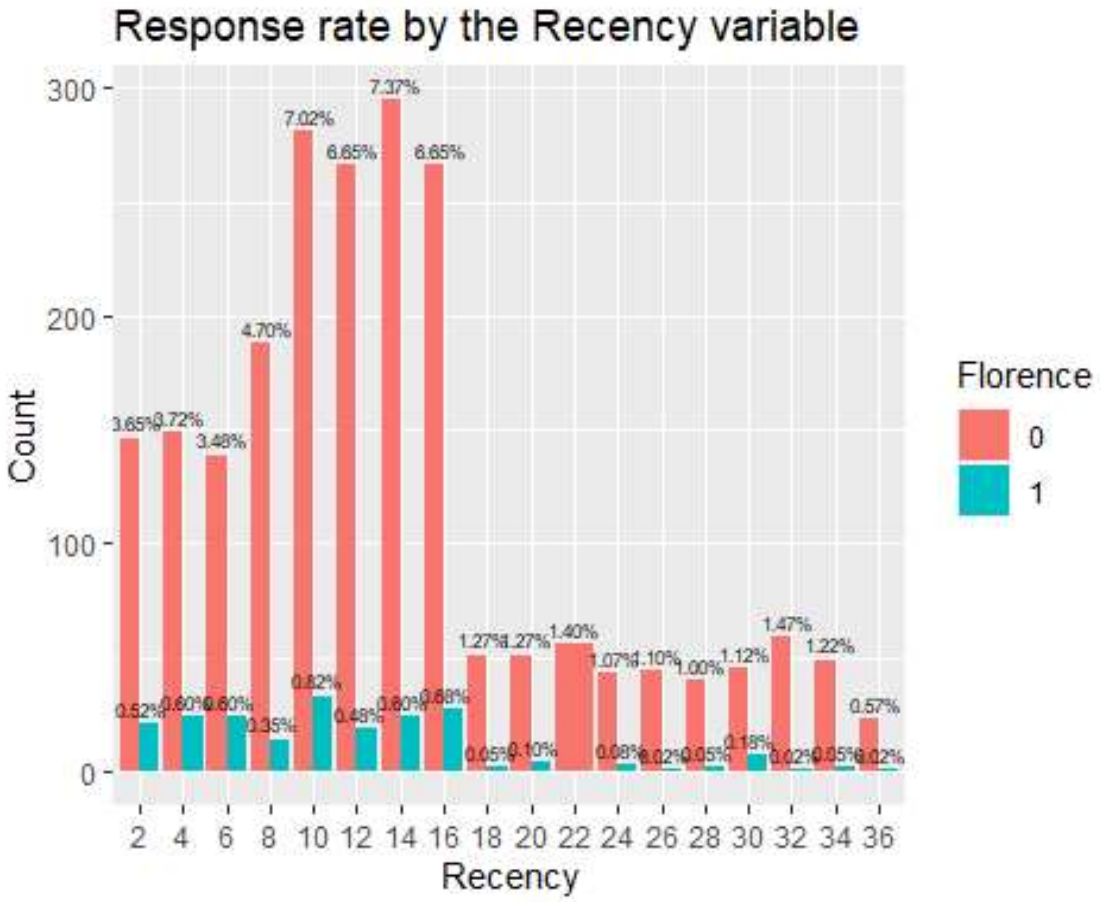
The plot for response rate by Recency variable is shown below:



Above plot shows that there is a difference in the mean for respondents and non-respondents.

For non-respondents the average value of recency is greater as compare to the respondents. I

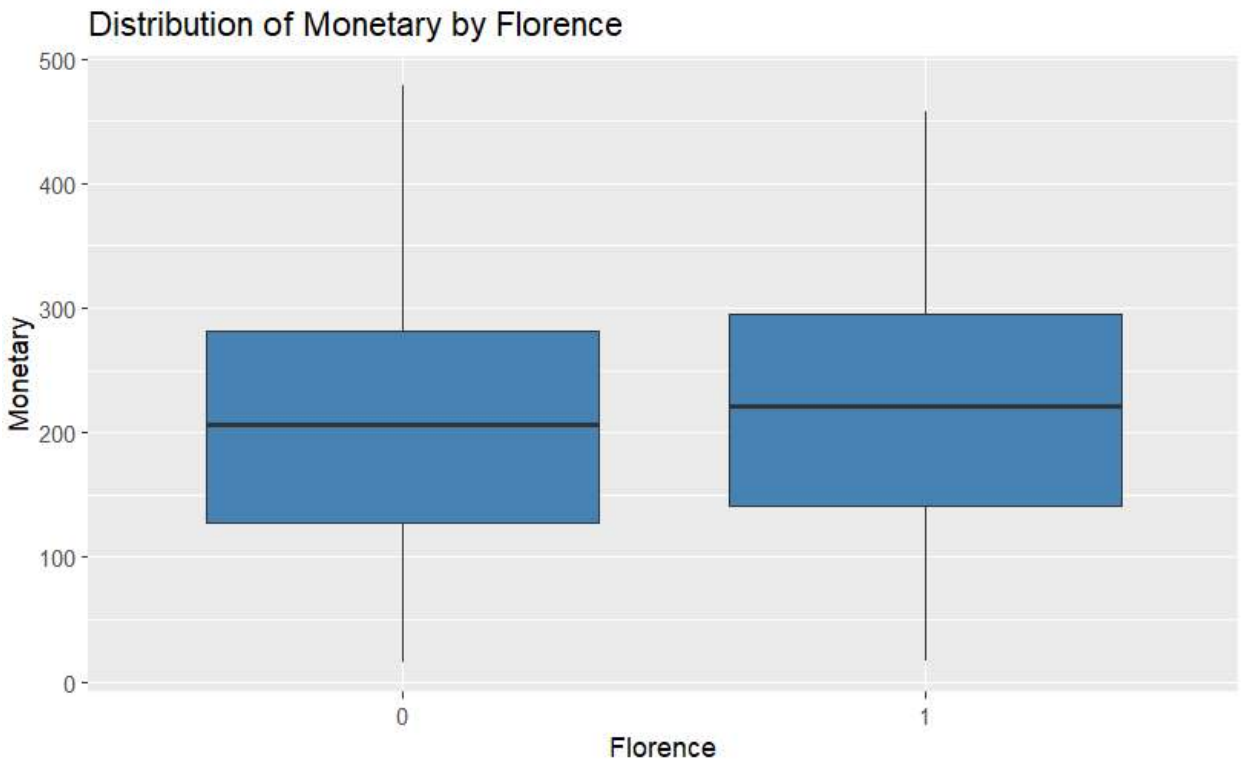
can conclude that customers with Florence campaign value of 1 are slightly more frequent as compare to the customers with Florence campaign value of 0.



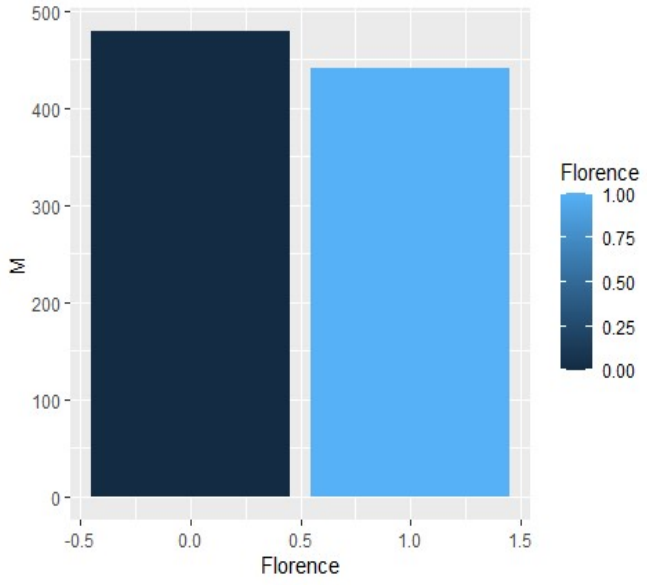
It is a necessary to understand the response rate by the recency visualization, we can see a higher chance if the recency score is less than 18.

Plot the response rate by the Monetary variable

The plot of response rate by the Monetary variable is shown below:

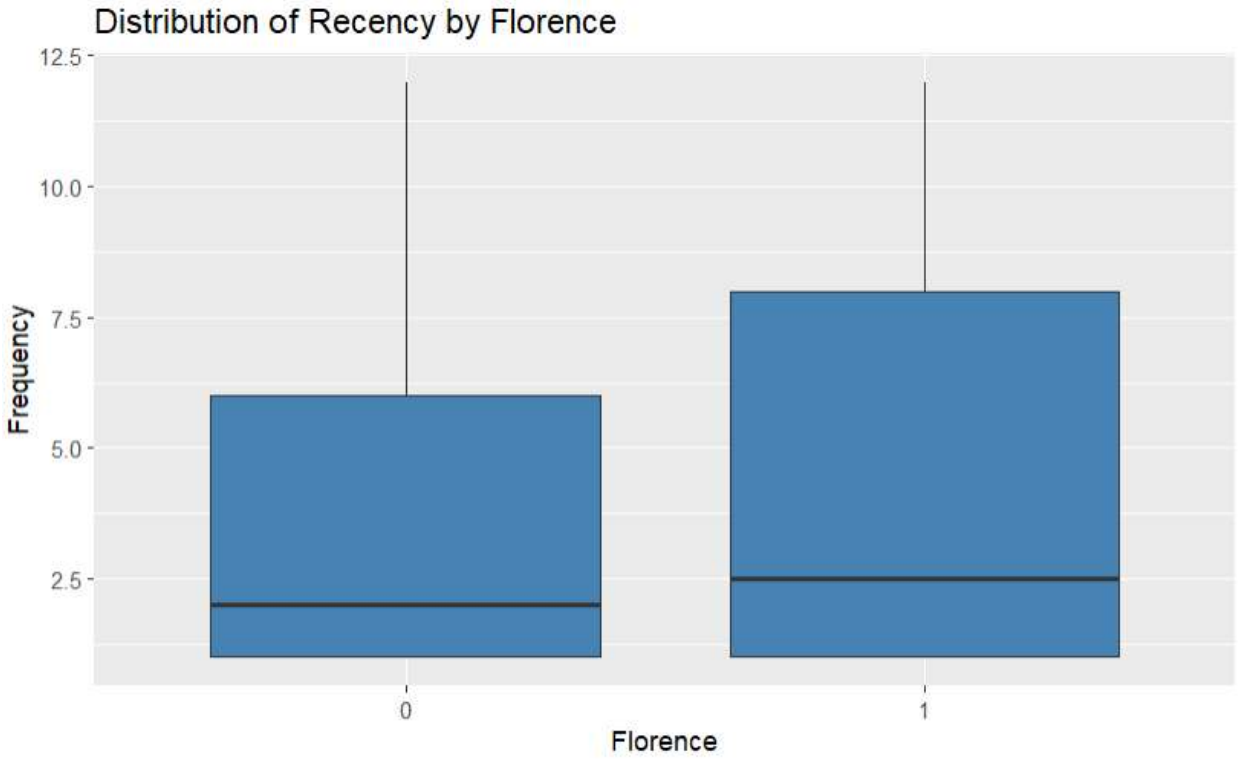


From above plot, it can be seen that customers who responded to Florence campaign have higher value of Monetary as compare to the customers who didn't respond to Florence campaign. I conclude that people who responded to the Florence campaign spent more money on the books as compare to people who didn't respond to the campaign.

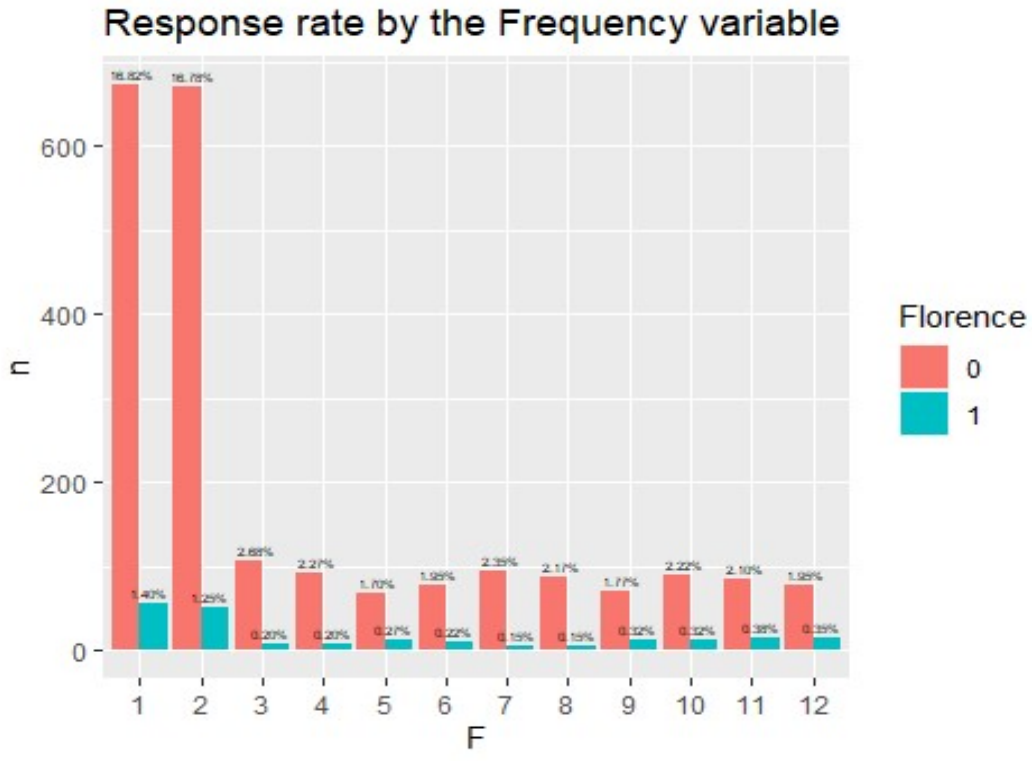


Plot the response rate by the Frequency variable

The plot of the response rate by the Frequency variable is shown below:



Above plot shows that people who responded to the Florence campaign made more average total number of purchases as compare to the people who didn't respond to the Florence campaign.

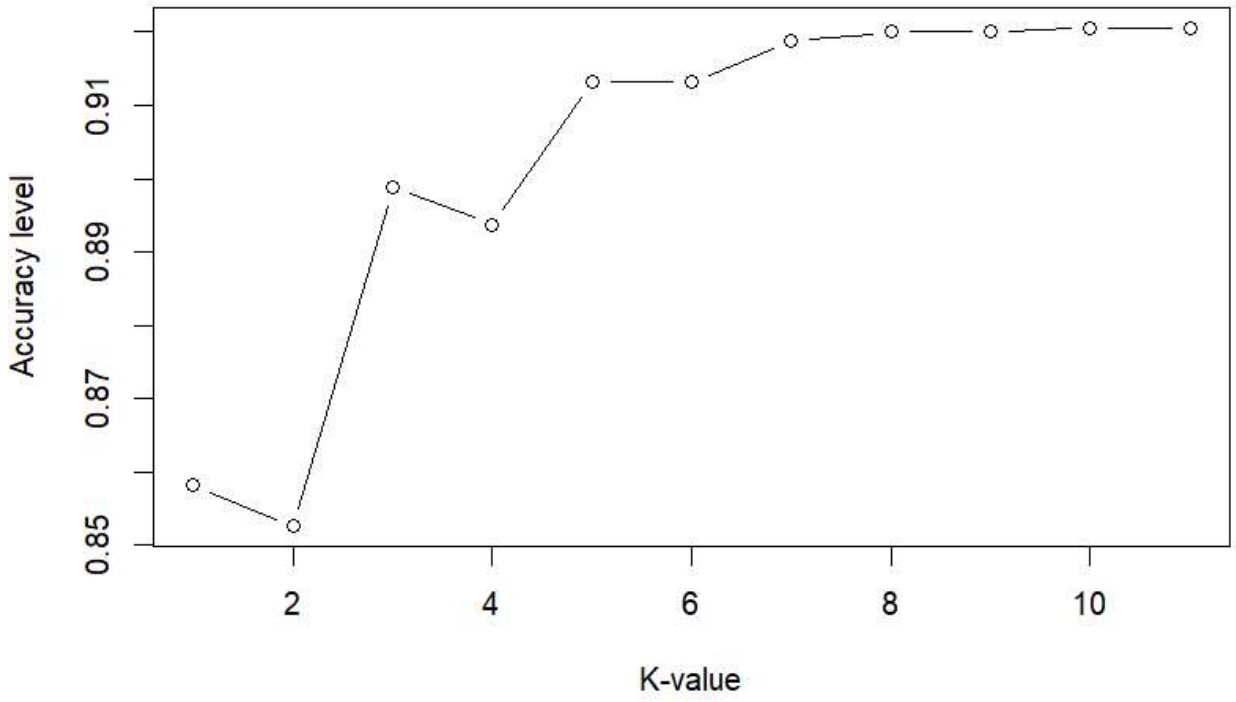


As per the visualization we can see that most of the people who are not frequent buyer has bought and we can also see a growing trend on people who are visiting over 9 times.

2. K-Nearest Neighbors

Find best value of K

In order to find best value of k, I implemented the knn model with values of k = 1, 2, ….. , 11 and then plot the accuracy of the model against each value of k. The plot is shown below:



From above plot the the value of k = 10 that balances between overfitting and ignoring the predictor information as this value provides best accuracy for test data.

Confusion matrix for best value of k

The confusion matrix for the best value of k is shown below:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1471    0
         1  127    2

               Accuracy : 0.9206
                 95% CI : (0.9063, 0.9334)
    No Information Rate : 0.9988
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0281

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

            Sensitivity : 0.9205
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.0155
             Prevalence : 0.9988
         Detection Rate : 0.9194
   Detection Prevalence : 0.9194
      Balanced Accuracy : 0.9603

       'Positive' Class : 0
```

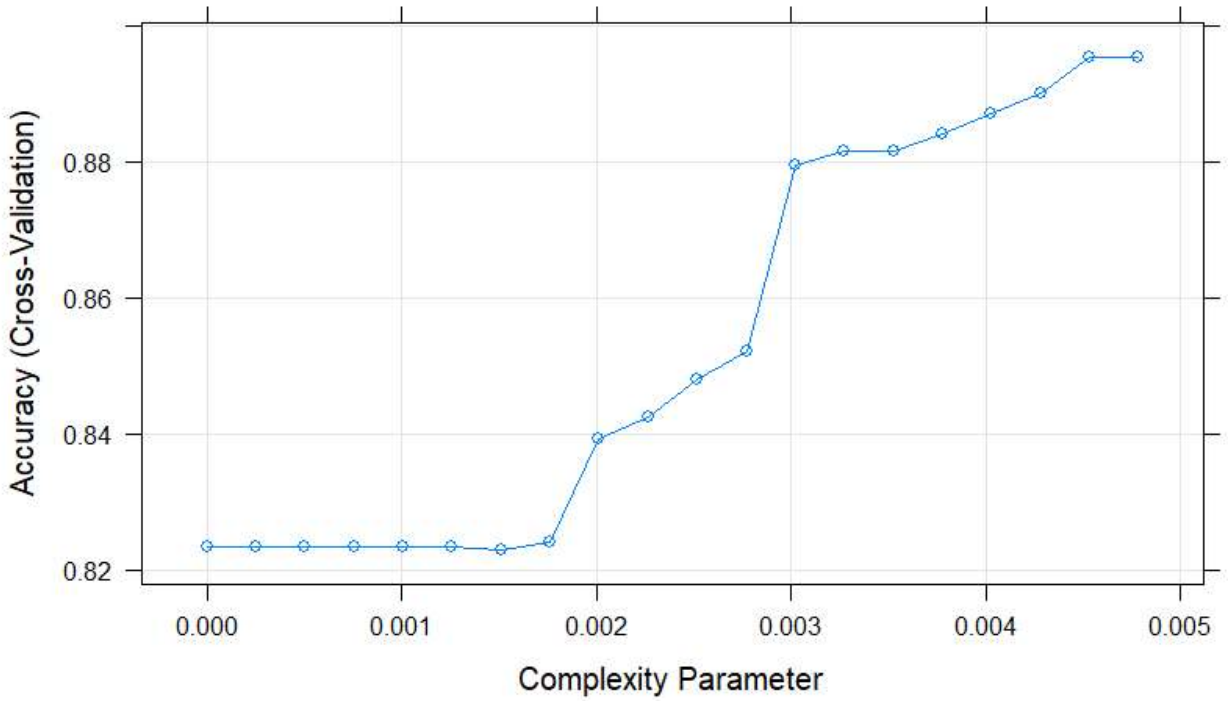
Above plot shows that the accuracy of the model is 0.92. The sensitivity value is 0.91995, the specificity value is 1 and the positive predicted value is 1. Since, the accuracy of model on test data is pretty high, so this model can be used practically for making predictions for this data set in future.

3. Classification Tree

For classification tree, I used 5-fold cross validation and fully implemented the tree. The output of tree is shown below:

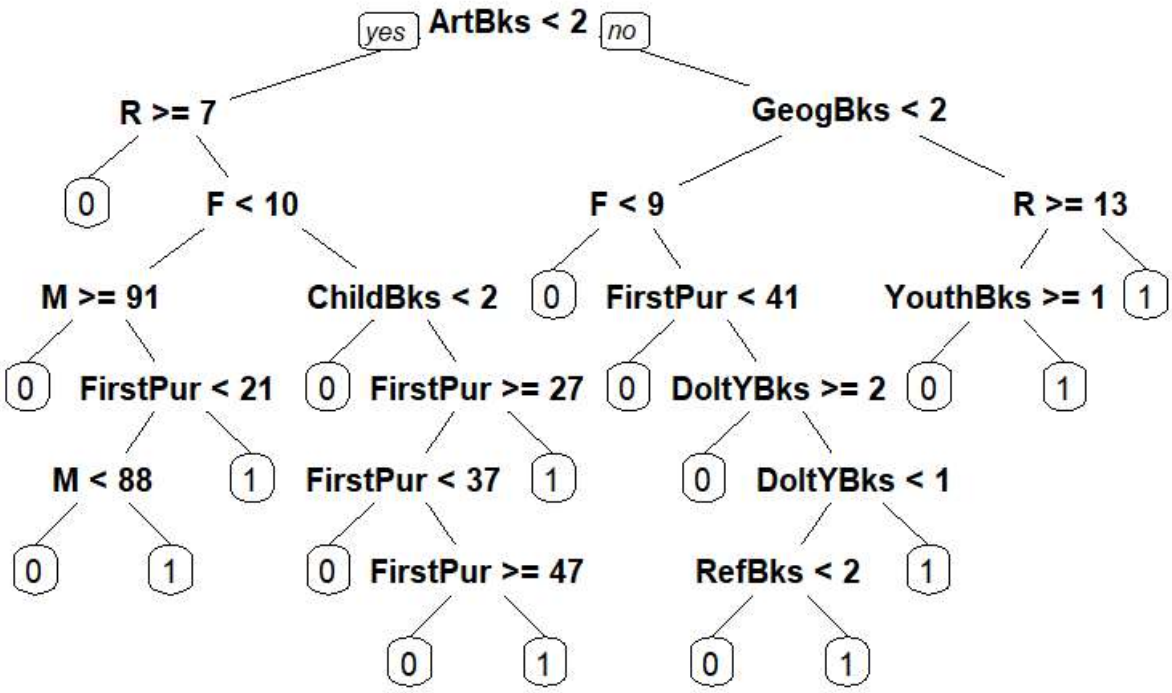
```
CART

2400 samples
  15 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 1920, 1920, 1920, 1920, 1920
Resampling results across tuning parameters:

  cp            Accuracy   Kappa
  0.0000000000  0.8233333  0.03181147
  0.0002518257  0.8233333  0.03181147
  0.0005036515  0.8233333  0.03181147
  0.0007554772  0.8233333  0.03181147
  0.0010073029  0.8233333  0.03181147
  0.0012591287  0.8233333  0.03181147
  0.0015109544  0.8229167  0.02749092
  0.0017627802  0.8241667  0.02878952
  0.0020146059  0.8391667  0.02244802
  0.0022664316  0.8425000  0.02616602
  0.0025182574  0.8479167  0.02049318
  0.0027700831  0.8520833  0.02596363
  0.0030219088  0.8795833  0.03116147
  0.0032737346  0.8816667  0.03404065
  0.0035255603  0.8816667  0.02869328
  0.0037773860  0.8841667  0.02737533
  0.0040292118  0.8870833  0.02805044
  0.0042810375  0.8900000  0.01991139
  0.0045328633  0.8954167  0.01584051
  0.0047846890  0.8954167  0.01584051

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.004784689.
```

Above output shows that, I used 20 different values for complexity parameter and the model has highest accuracy for cp = 0.004784689. The plot of tree model using different values of complexity parameter and accuracy is shown below:



Above plot shows that the accuracy of model is highest for cp value of approximately 0.0047.

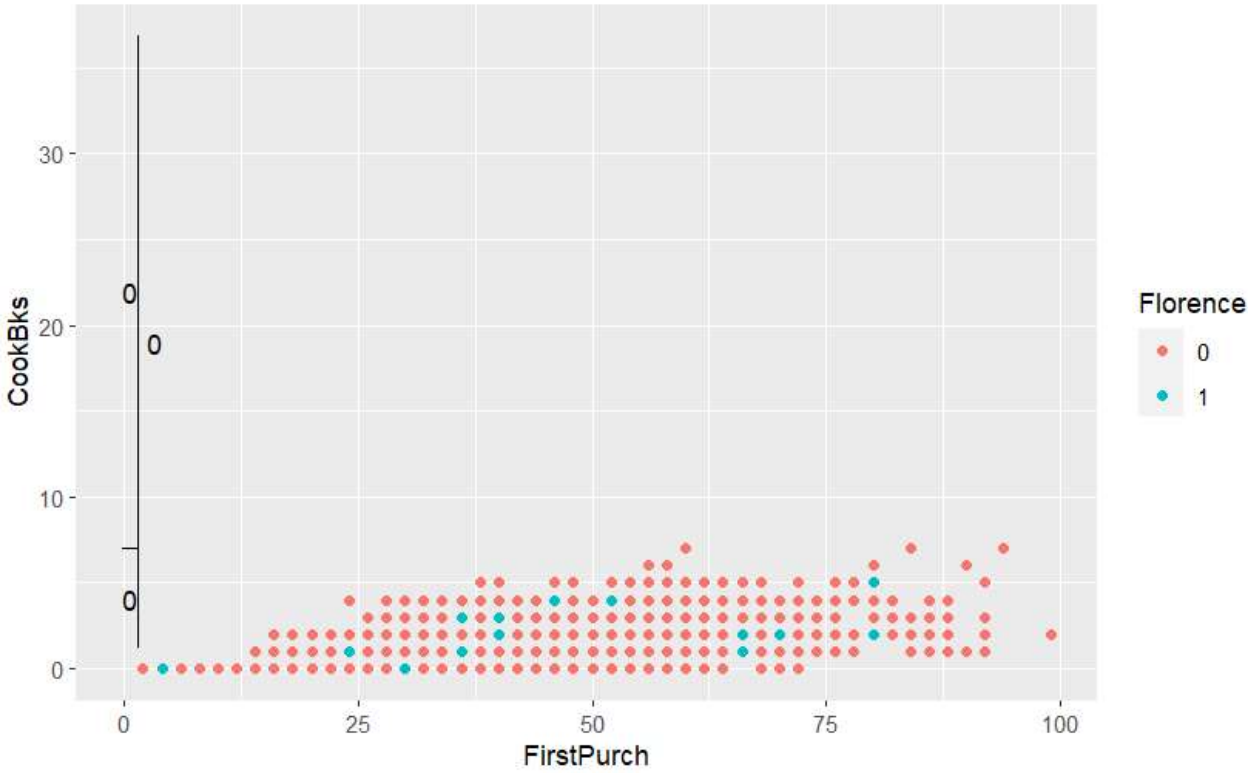
The plot for best pruned decision tree is shown below:



Above plot shows different rules used in making predictions in decision tree. From above plot, if the ArtBks value is less than 2 and R is greater or equal to 7 then value of Florence will be 0. If ArtBks value is greater than 2 and F value is less than 8 then the value of Florence will be 0. If ArtBks is less than 2 and R is greater than 7 and GoegBks is greater than 2 and if R is less than 13

Then if YouthBks is greater or equal to 1 then Floerence will be 0 and if YouthBks is less than 1 then FLoerence wil be 1. If ArtBks is less than 2 and R is less than 7 and F is greater than 10 and if ChildBks is less than 2 then Florence will be 0. Similarly the other rules can be interpreted also.

The scatter plot for decision boundary of decision tree is shown below:



From above plot, it can be seen that the model doesn't separate well the classes 0 and 1. The model predicted all the classes as 0 and didn't predict even a single point as 1. So, I concluded that this splitting doesn't seems reasonable with respect to the meaning of predictors.

Confusion matrix for optimal tree

The confusion matrix for optimal tree is shown below:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1448   23
         1  121    8

               Accuracy : 0.91
                 95% CI : (0.8949, 0.9236)
    No Information Rate : 0.9806
    P-Value [Acc > NIR] : 1

                  Kappa : 0.071

 Mcnemar's Test P-Value : 6.302e-16

            Sensitivity : 0.92288
            Specificity : 0.25806
         Pos Pred Value : 0.98436
         Neg Pred Value : 0.06202
             Prevalence : 0.98062
         Detection Rate : 0.90500
   Detection Prevalence : 0.91938
      Balanced Accuracy : 0.59047

       'Positive' Class : 0
```

From above confusion matrix, the sensitivity of the model is 0.92288, the specificity value is 0.25806, and positive predicted value is 0.98436 and the accuracy of the model on test data is 0.91. Since the decision boundary of the model is not good as it predicts all values 0, so this model is practically not good enough to predict the Florence value in future,

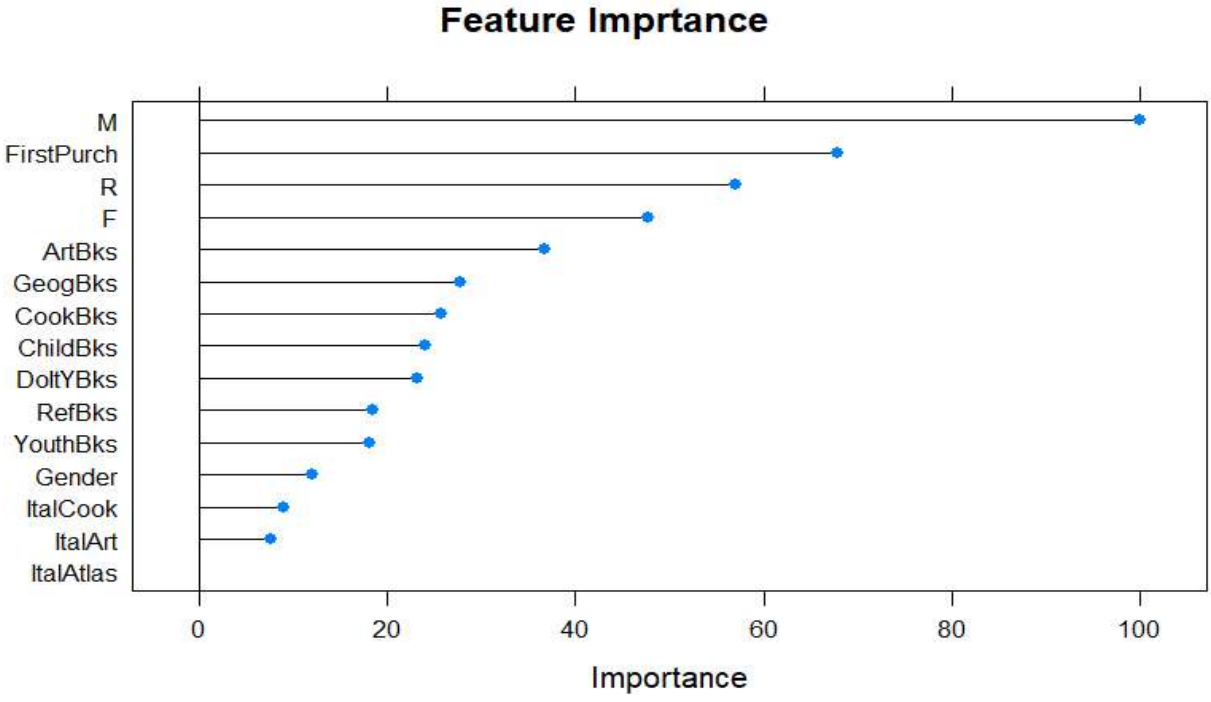
4. Random Forest

For Random Forest model, I trained the model using different values of hyper parameter mtry. The output for this model is shown below:

```
Random Forest

2400 samples
  15 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold, repeated 1 times)
Summary of sample sizes: 1921, 1920, 1919, 1920, 1920
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.9129174  0.00000000
   3    0.9129174  0.00000000
   4    0.9120814  0.02122243
   6    0.9100016  0.03785583
   7    0.9075007  0.03262923
   9    0.9058349  0.05008476
  10    0.9062490  0.04450544
  12    0.9054182  0.04267603
  13    0.9025007  0.02999189
  15    0.9033358  0.03813479

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 2.
```

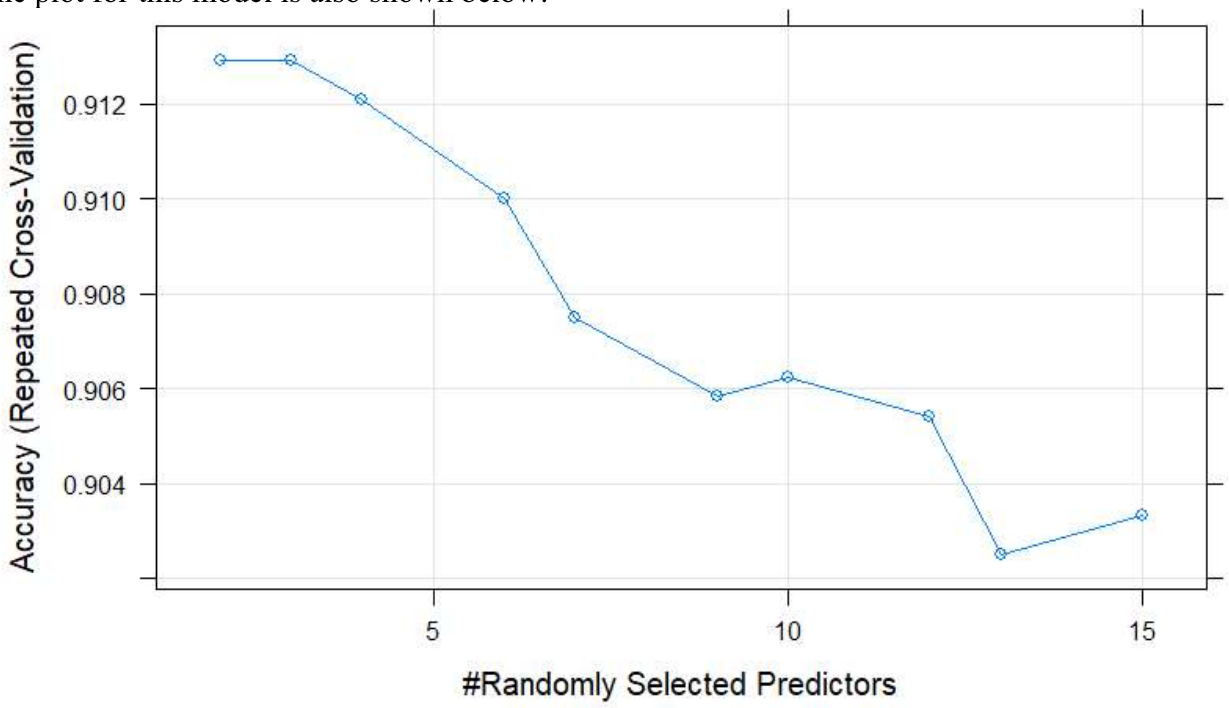
From above output the best value of mtry = 2.

The feature importance plot for random forest is shown below:



From above plot, it can be seen that for random forest model the most important variable is M, followed by FirstPurch, R and F. The plot also shows that the least important variables are ItalAtlas, ItalArt and ItalCook.

The plot for this model is also shown below:



From above plot, the model has highest accuracy for mtry = 2. I made predictions on test data using best model and the confusion matrix is shown below:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1471    0
         1  129    0

               Accuracy : 0.9194
                 95% CI : (0.9049, 0.9322)
    No Information Rate : 1
    P-Value [Acc > NIR] : 1

                  Kappa : 0

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

            Sensitivity : 0.9194
            Specificity :     NA
         Pos Pred Value :     NA
         Neg Pred Value :     NA
             Prevalence : 1.0000
         Detection Rate : 0.9194
   Detection Prevalence : 0.9194
      Balanced Accuracy :     NA

       'Positive' Class : 0
```

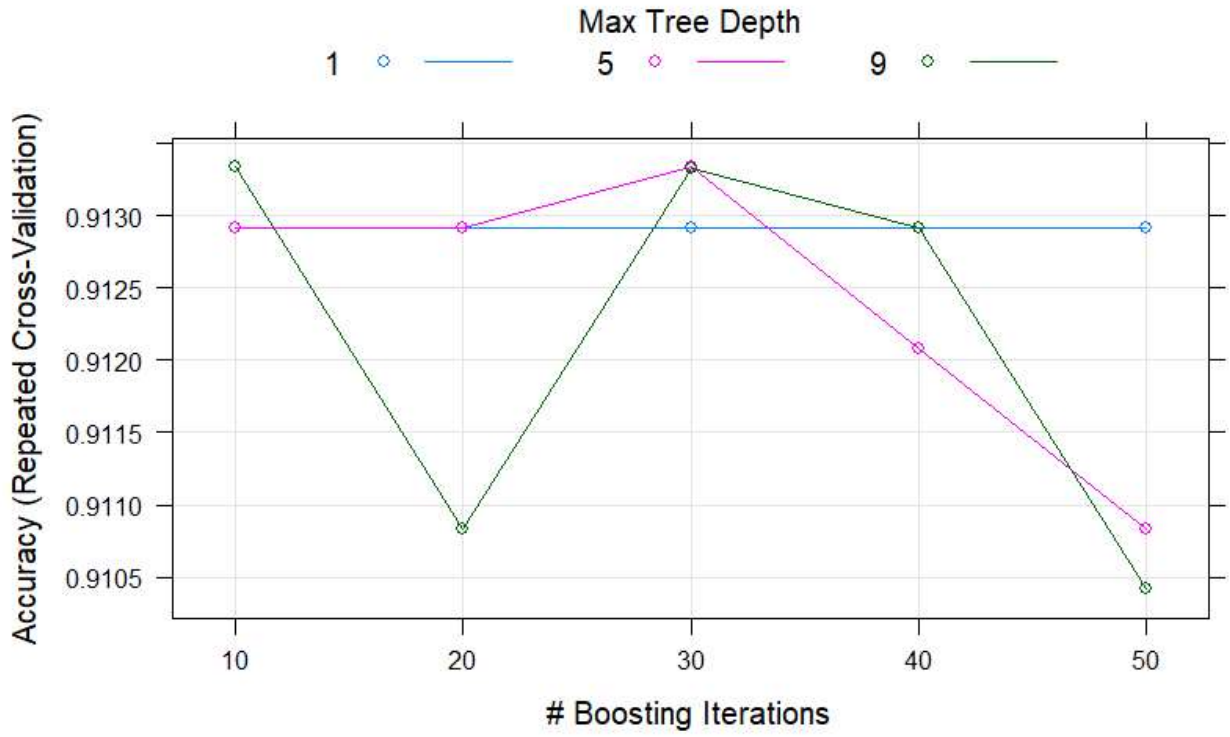
From the output, the Sensitivity value is 0.9194, the Specificity value is 0, the positive predicted value is 0 and the accuracy of the model on test data is 0.9194. Since this model has good accuracy on the test data so this model can be used for making predictions in future and this model is practically a good model.

5. Boosted Trees

For Boosted trees, I used gradient boosting model and tuned it using different hyper parameters. The output for this model is shown below:

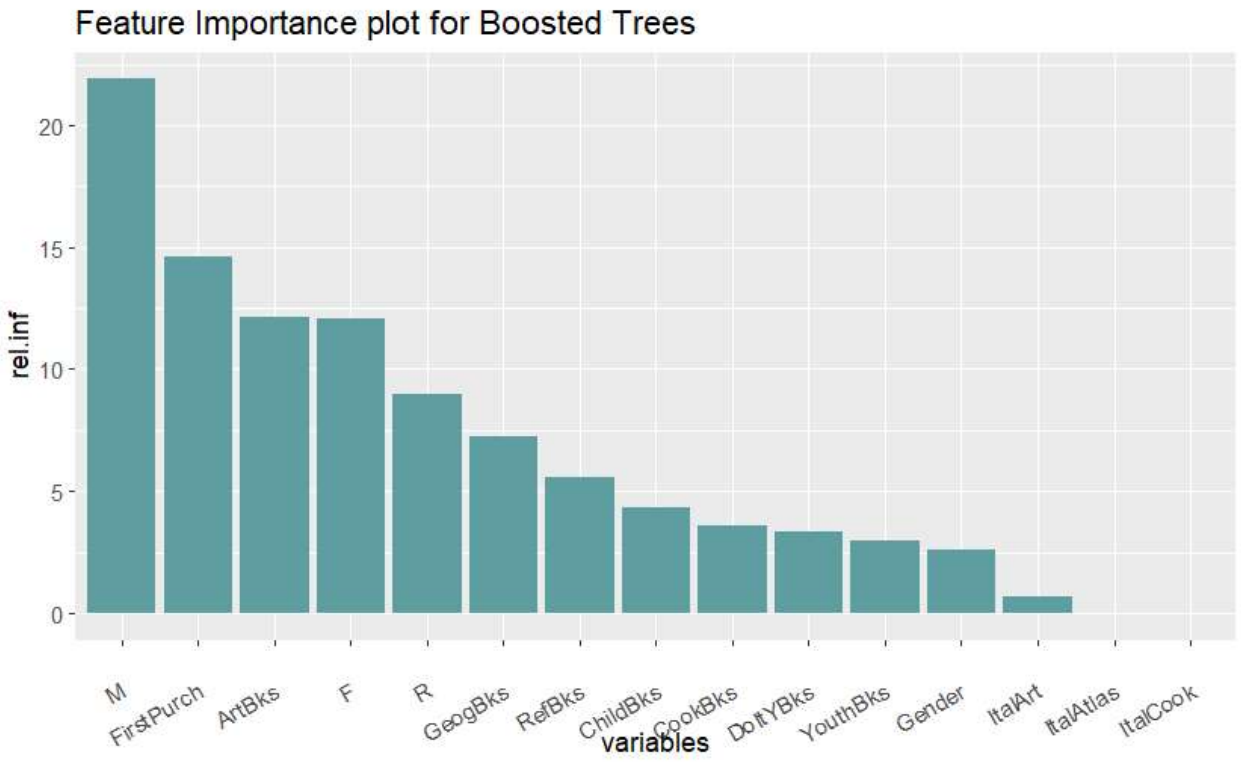
```
Stochastic Gradient Boosting

2400 samples
  15 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold, repeated 1 times)
Summary of sample sizes: 1921, 1920, 1919, 1920, 1920
Resampling results across tuning parameters:

  interaction.depth  n.trees  Accuracy   Kappa
  1                  10       0.9129174  0.00000000
  1                  20       0.9129174  0.00000000
  1                  30       0.9129174  0.00000000
  1                  40       0.9129174  0.00000000
  1                  50       0.9129174  0.00000000
  5                  10       0.9129165  0.01526467
  5                  20       0.9129165  0.03042388
  5                  30       0.9133332  0.04593221
  5                  40       0.9120832  0.06335136
  5                  50       0.9108340  0.05983695
  9                  10       0.9133332  0.01622759
  9                  20       0.9108314  0.01749897
  9                  30       0.9133323  0.06598279
  9                  40       0.9129165  0.07230075
  9                  50       0.9104165  0.05867776

Tuning parameter 'shrinkage' was held constant at a value of 0.1
Tuning parameter 'n.minobsinnode' was
 held constant at a value of 3
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were n.trees = 10, interaction.depth = 9,
shrinkage = 0.1 and
 n.minobsinnode = 3.
```

The plot for this model is shown below:



From above plot, it can be seen that the highest accuracy is for max_tree_depth 10, and number of iterations is 9.

The feature importance plot for this model is shown below:



Above plot shows that, according to boosted trees model, the most important variable is FirstPurch, followed by Monetary and Recency. The plot also shows the least important variables are RefBks and ItalAtlas.
I made predictions on test data using best model and the confusion matrix is shown below:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1470    1
         1  127    2

               Accuracy : 0.92
                 95% CI : (0.9056, 0.9328)
    No Information Rate : 0.9981
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0267

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

            Sensitivity : 0.9205
            Specificity : 0.6667
         Pos Pred Value : 0.9993
         Neg Pred Value : 0.0155
             Prevalence : 0.9981
         Detection Rate : 0.9187
   Detection Prevalence : 0.9194
      Balanced Accuracy : 0.7936

       'Positive' Class : 0
```

From the output of confusion matrix for boosted trees, the Sensitivity of model on test data is 0.9205, the Specificity value is 0.6668, the positive predicted value is 0.9993 and the overall accuracy of the model is 0.92. Since the accuracy of the model is pretty good, so this model can be used practically to make predictions for Florence in future.

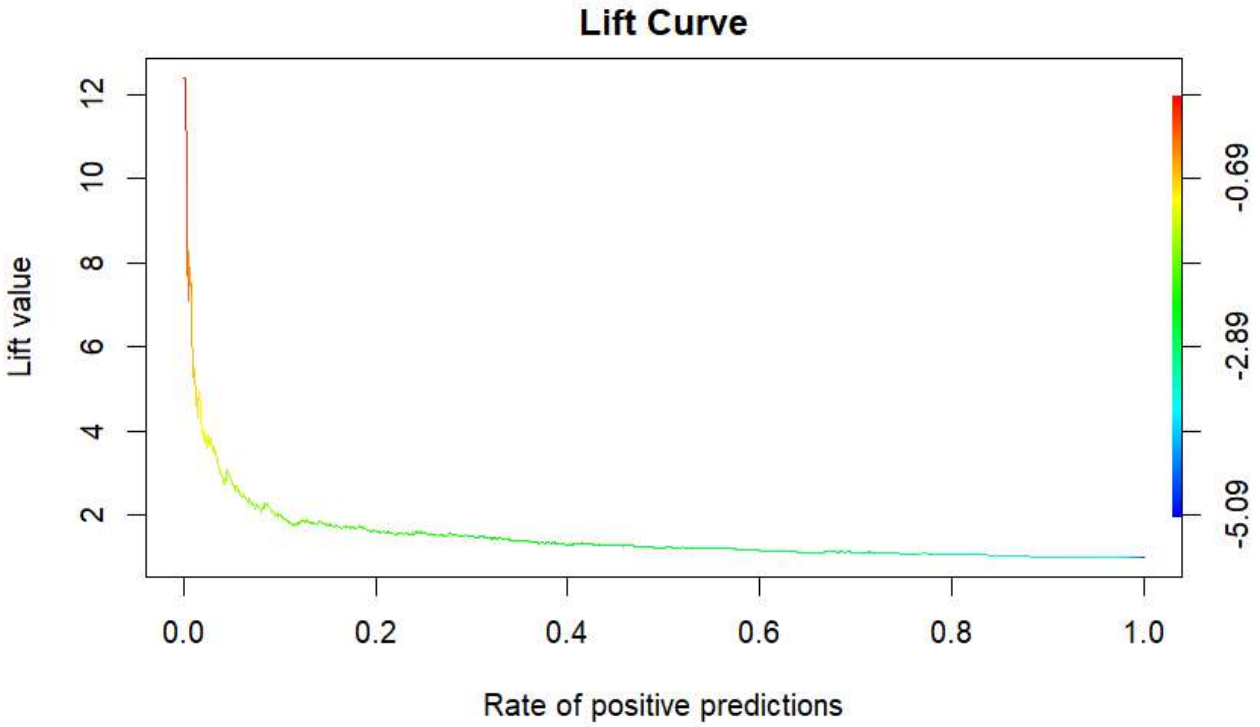
6. Logistic Regression

I implemented logistic regression on training data and made predictions on test data. The summary for logistic regression is shown below:

```
Call:
glm(formula = Florence ~ ., family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0966  -0.4385  -0.3669  -0.2963   2.8092

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.3143644  0.2584677  -8.954  < 2e-16 ***
Gender      -0.3389381  0.1563721  -2.168 0.030196 *
M            0.0001257  0.0008788   0.143 0.886245
R           -0.0299126  0.0155749  -1.921 0.054787 .
F            0.2048026  0.0649261   3.154 0.001608 **
FirstPurch  -0.0005054  0.0110170  -0.046 0.963407
ChildBks    -0.1773853  0.1057749  -1.677 0.093541 .
YouthBks    -0.2621029  0.1517956  -1.727 0.084225 .
CookBks     -0.3732984  0.1098566  -3.398 0.000679 ***
DoItYBks    -0.1841962  0.1329718  -1.385 0.165983
RefBks      -0.2272505  0.1542431  -1.473 0.140663
ArtBks       0.4823741  0.1012087   4.766 1.88e-06 ***
GeogBks      0.2408928  0.0976610   2.467 0.013639 *
ItalCook    -0.0614844  0.1981087  -0.310 0.756290
ItalAtlas   -0.8621905  0.5085769  -1.695 0.090018 .
ItalArt      0.2700183  0.2793200   0.967 0.333695
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1419.5  on 2399  degrees of freedom
Residual deviance: 1326.3  on 2384  degrees of freedom
AIC: 1358.3

Number of Fisher Scoring iterations: 5
```

From above output, I observed that the variables monetary, FirstPurch, YouthBks, RefBks, ItalCook, ItalAtlas and IralArt don’t have significant relation with dependent variable Florence based on 0.05 confidence level.

The lift chart for logistic regression is shown below:



Above plot represents the Lift Curve for logistic regression model. From above plot, it is pretty evident that with increase in rate of positive predictors the lift value decreased and reached to the minimum value when the rate of positive predictions is 1.

The confusion matrix for logistic regression is shown below:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1471    0
         1  127    2

               Accuracy : 0.9206
                 95% CI : (0.9063, 0.9334)
    No Information Rate : 0.9988
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0281

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

            Sensitivity : 0.9205
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.0155
             Prevalence : 0.9988
         Detection Rate : 0.9194
   Detection Prevalence : 0.9194
      Balanced Accuracy : 0.9603

       'Positive' Class : 0
```

From the output of logistic regression, the sensitivity value is 0.9205, the specificity value is 1, the positive prediction value is 1 and the accuracy of the model 0.9206. Since the accuracy

of the model is pretty good, the model is practically good and can be used for making predictions.

7. Final Recommendation

The table for Sensitivity, Specificity, positive predicted value and the overall accuracy for each model is given below in table:

Model	Sensitivity	Specificity	Positive Predicted Value	Accuracy
KNN	0.9205	1.00	1.0000	0.9206
Classification Tree	0.9229	0.26	0.9843	0.9100
Random Forest	0.9194	0.00	0.0000	0.9194
Boosted Trees	0.9205	0.67	0.9993	0.9200
Logistic Regression	0.9205	1.00	1.000	0.9206

From above table, I observed that from all models, Logistic regression and KNN model has highest values for Sensitivity, Specificity and positive predicted value and accuracy which means Logistic Regression model fits and generalize well to the data and performs better as compare to other models.

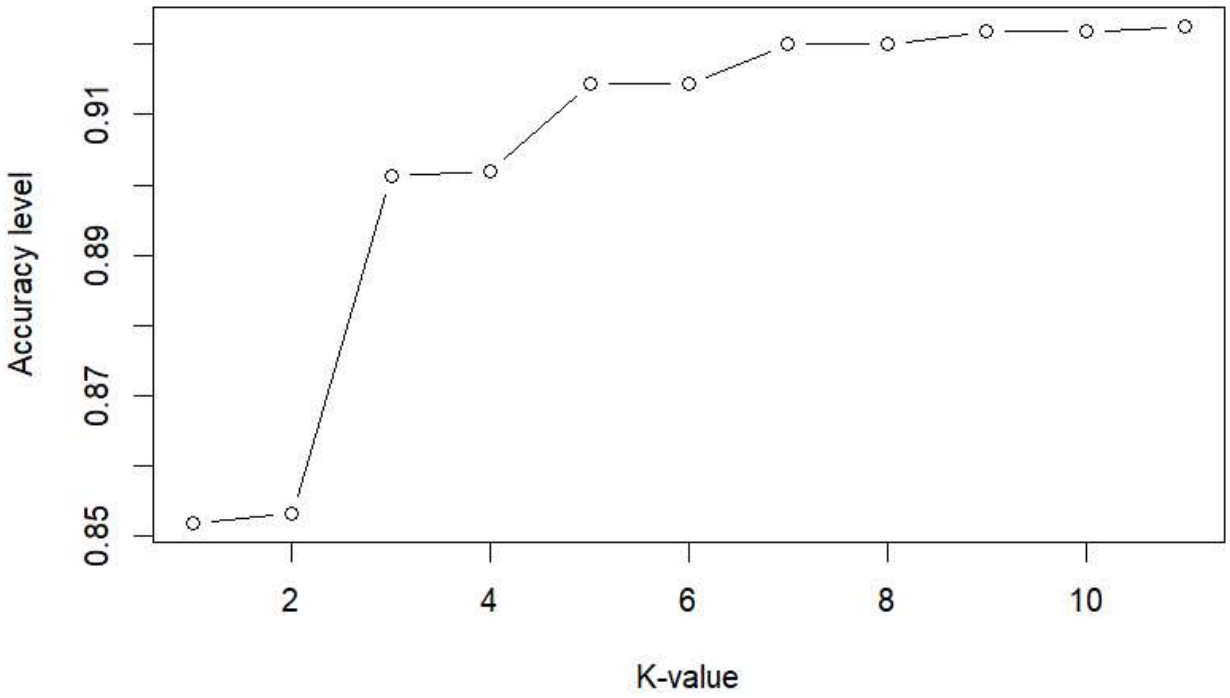
Exploratory & Supplementary Work

The Models here on are just for exploring more possibilities with the given data .The seed value in the below analysis is 101.

1. K-Nearest Neighbors

Find best value of K

In order to find best value of k, I implemented the knn model with values of k = 1, 2, ….. , 11 and then plot the accuracy of the model against each value of k. The plot is shown below:



From above plot the the value of k = 10 that balances between overfitting and ignoring the predictor information as this value provides best accuracy for test data.

Confusion matrix for best value of k

The confusion matrix for the best value of k is shown below:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1475    0
         1  124    1

               Accuracy : 0.9225
                 95% CI : (0.9083, 0.9351)
    No Information Rate : 0.9994
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0147

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

            Sensitivity : 0.9225
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.0080
             Prevalence : 0.9994
         Detection Rate : 0.9219
   Detection Prevalence : 0.9219
      Balanced Accuracy : 0.9612

       'Positive' Class : 0
```

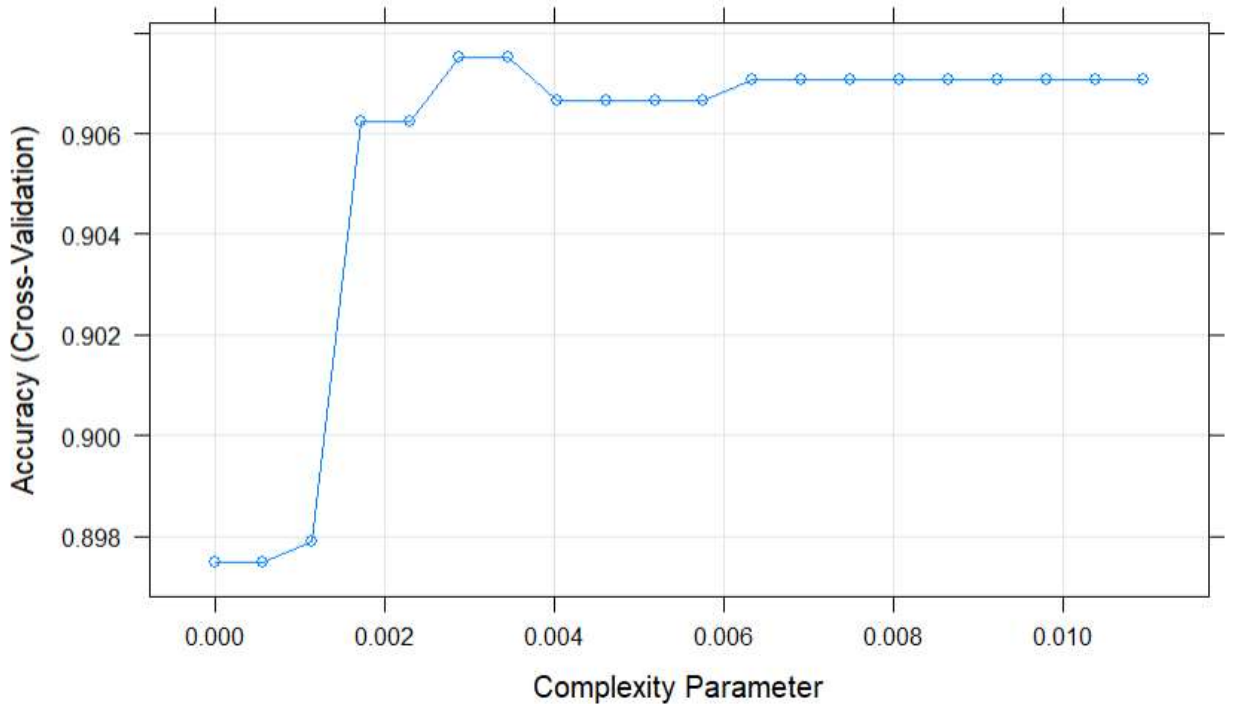
Above plot shows that the accuracy of the model is 0.9225. The sensitivity value is 0.9225, the specificity value is 1 and the positive predicted value is 1. Since, the accuracy of model on test data is pretty high, so this model ican be used practically for making predictions for this data set in future.

3. Classification Tree

For classification tree, I used 5-fold cross validation and fully implemented the tree. The output of tree is shown below:

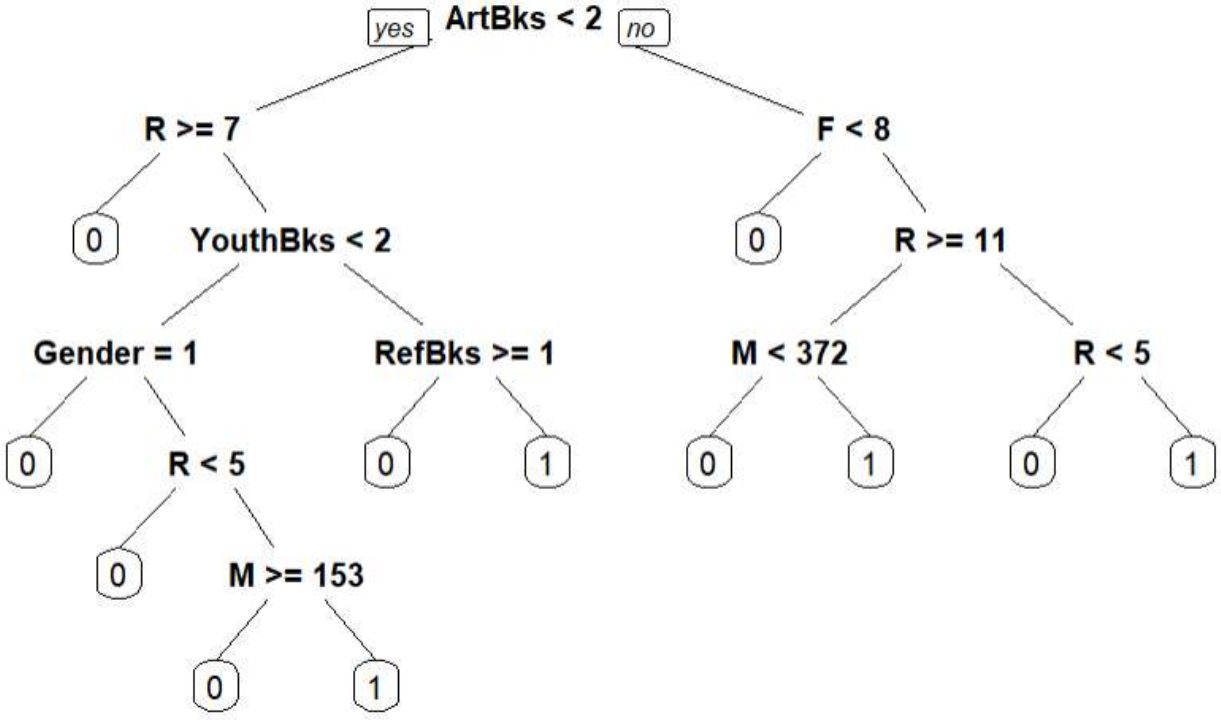
```
CART

2400 samples
  15 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 1919, 1920, 1920, 1921, 1920
Resampling results across tuning parameters:

  cp            Accuracy   Kappa
  0.0000000000  0.8974990  0.04673516
  0.0005765588  0.8974990  0.04673516
  0.0011531175  0.8979156  0.04756316
  0.0017296763  0.9062507  0.06474291
  0.0023062351  0.9062507  0.06474291
  0.0028827938  0.9075007  0.06803509
  0.0034593526  0.9075007  0.06803509
  0.0040359114  0.9066665  0.05228330
  0.0046124701  0.9066665  0.05228330
  0.0051890289  0.9066665  0.05228330
  0.0057655877  0.9066665  0.05228330
  0.0063421464  0.9070832  0.04048078
  0.0069187052  0.9070832  0.04048078
  0.0074952640  0.9070832  0.02591124
  0.0080718227  0.9070832  0.02591124
  0.0086483815  0.9070832  0.02591124
  0.0092249403  0.9070832  0.02591124
  0.0098014991  0.9070832  0.02591124
  0.0103780578  0.9070832  0.02591124
  0.0109546166  0.9070832  0.02591124

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.003459353.
```

Above output shows that, I used 20 different values for complexity parameter and the model has highest accuracy for cp = 0.003459353. The plot of tree model using different values of complexity parameter and accuracy is shown below:

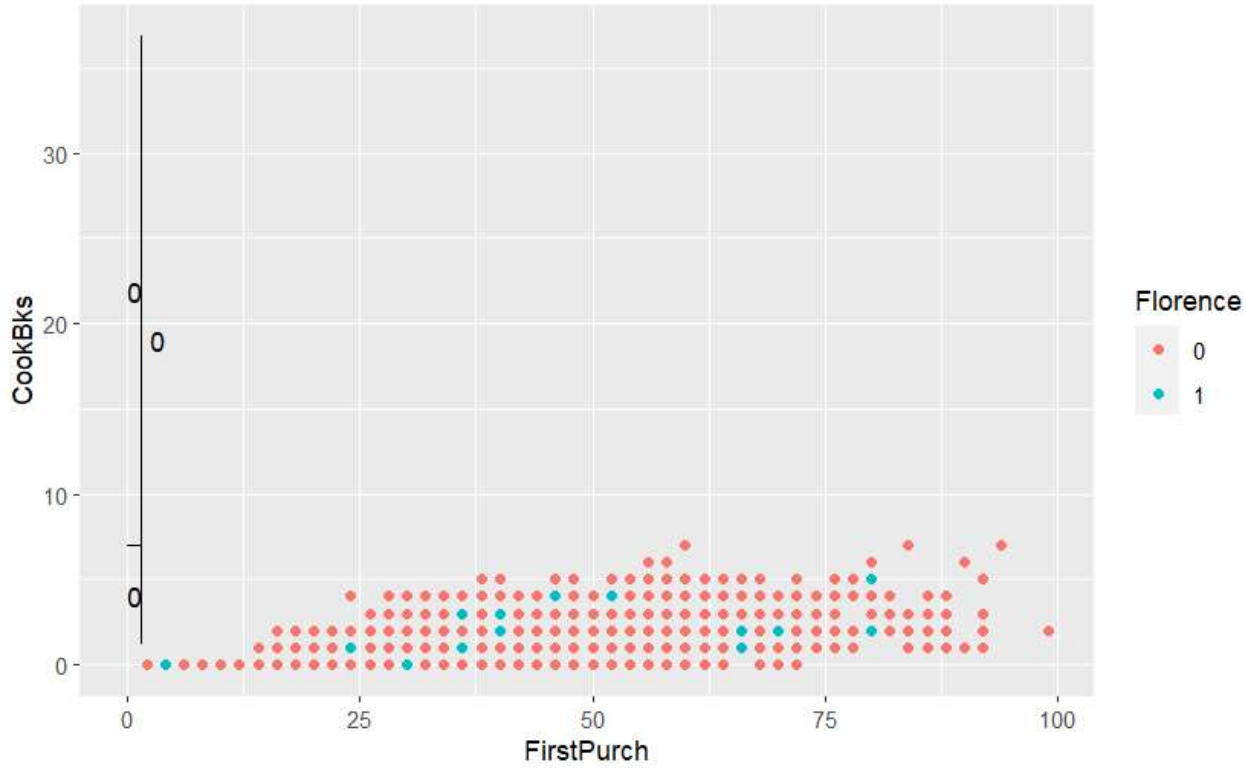


Above plot shows that the accuracy of model is highest for cp value of approximately 0.011.

The plot for best pruned decision tree is shown below:



Above plot shows different rules used in making predictions in decision tree. From above plot, if the ArtBks value is less than 2 and R is greater or equal to 7 then value of Florence will be 0. If ArtBks value is greater than 2 and F value is less than 8 then the value of Florence will be 0. If ArtBks is less than 2 and R is less than 7 and YouthBks is greater than 2 and if RefBks is less than 1 then Florence value will be 0 and if RefBks is greater or equal to 1 then the Florence value will be 0. Similarly the other rules can be interpreted also.
The scatter plot for decision boundary of decision tree is shown below:



From above plot, it can be seen that the model doesn't separate well the classes 0 and 1. The model predicted all the classes as 0 and didn't predict even a single point as 1. So, I concluded that this splitting doesn't seems reasonable with respect to the meaning of predictors.

Confusion matrix for optimal tree

The confusion matrix for optimal tree is shown below:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1448   27
         1  122    3

               Accuracy : 0.9069
                 95% CI : (0.8916, 0.9207)
    No Information Rate : 0.9812
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0087

 Mcnemar's Test P-Value : 1.352e-14

            Sensitivity : 0.9223
            Specificity : 0.1000
         Pos Pred Value : 0.9817
         Neg Pred Value : 0.0240
             Prevalence : 0.9812
         Detection Rate : 0.9050
   Detection Prevalence : 0.9219
      Balanced Accuracy : 0.5111

       'Positive' Class : 0
```

From above confusion matrix, the sensitivity of the model is 0.9223, the specificity value is 0.1, and positive predicted value is 0.9817 and the accuracy of the model on test data is 0.9069. Since the decision boundary of the model is not good as it predict all values 0, so this model is practically not good enough to predict the Florence value in future.

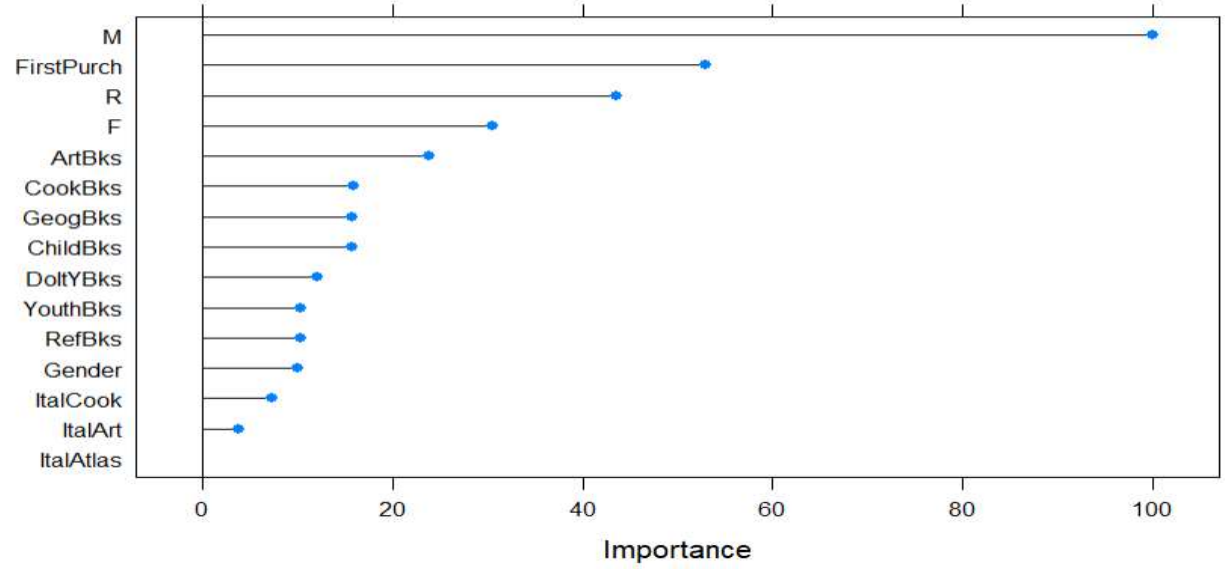
4. Random Forest

For Random Forest model, I trained the model using different values of hyper parameter mtry. The output for this model is shown below:

```
note: only 14 unique complexity parameters in default grid. Truncating the grid to
14 .

Random Forest

2400 samples
  15 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 2160, 2159, 2160, 2159, 2160, 2160, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.9109746  0.004505536
   3    0.9113918  0.019943085
   4    0.9119457  0.037829629
   5    0.9123629  0.054858582
   6    0.9118079  0.056098693
   7    0.9119474  0.067569461
   8    0.9112529  0.081404900
   9    0.9112530  0.091757389
  10    0.9109746  0.097646917
  11    0.9102801  0.103577964
  12    0.9109723  0.109675138
  13    0.9100024  0.108817932
  14    0.9101424  0.112994683
  15    0.9087506  0.106049127

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 5.
```

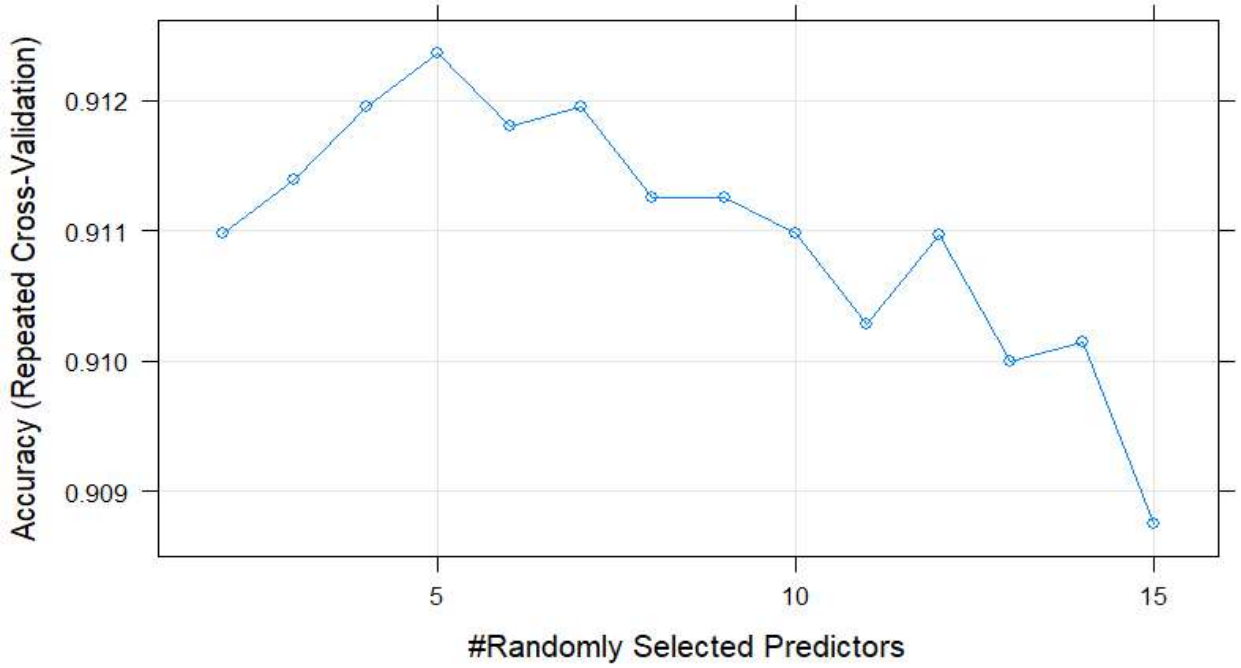
From above output the best value of mtry = 5.

The feature importance plot for random forest is shown below:



From above plot, it can be seen that for random forest model the most important variable is M, followed by FirstPurch, R and F. The plot also shows that the least important variables are ItalAtlas, ItalArt and ItalCook.

The plot for this model is also shown below:



From above plot, the model has highest accuracy for mtry = 4. I made predictions on test data using best model and the confusion matrix is shown below:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1467    8
         1  124    1

               Accuracy : 0.9175
                 95% CI : (0.9029, 0.9305)
    No Information Rate : 0.9944
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0045

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

            Sensitivity : 0.9221
            Specificity : 0.1111
         Pos Pred Value : 0.9946
         Neg Pred Value : 0.0080
             Prevalence : 0.9944
         Detection Rate : 0.9169
   Detection Prevalence : 0.9219
      Balanced Accuracy : 0.5166

       'Positive' Class : 0
```

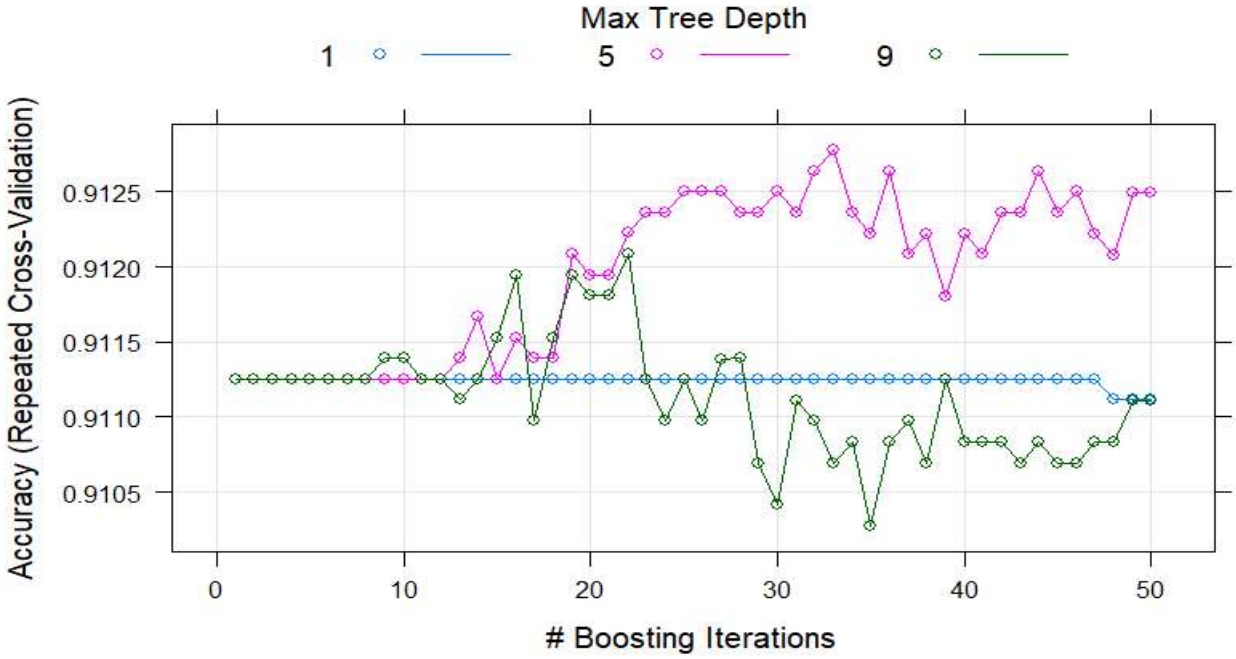
From the output, the Sensitivity value is 0.9221, the Specificity value is 0.1111, the positive predicted value is 0.9946 and the accuracy of the model on test data is 0.9175. Since this model has good accuracy on the test data so this model can be used for making predictions in future and this model is practically a good model.

5. Boosted Trees

For Boosted trees, I used gradient boosting model and tuned it using different hyper parameters. The output for this model is shown below:

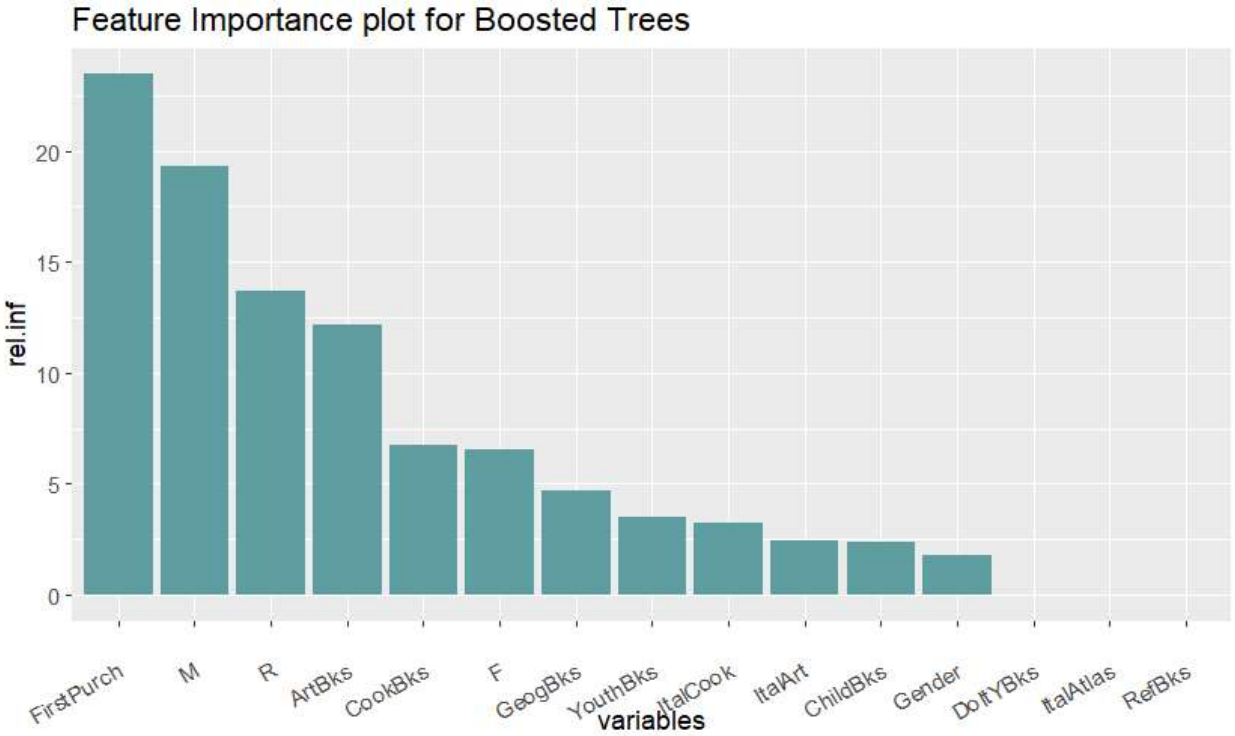
```
Stochastic Gradient Boosting

2400 samples
  15 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 2161, 2161, 2160, 2160, 2160, 2159, ...
Resampling results across tuning parameters:

  interaction.depth  n.trees  Accuracy   Kappa
  1                  10       0.9112529   0.0000000000
  1                  20       0.9112529   0.0000000000
  1                  30       0.9112529   0.0000000000
  1                  40       0.9111140  -0.0002672776
  1                  50       0.9109751  -0.0005351130
  5                  10       0.9104189   0.0056962298
  5                  20       0.9104189   0.0271972502
  5                  30       0.9106938   0.0523870550
  5                  40       0.9102789   0.0556678066
  5                  50       0.9098616   0.0530295431
  9                  10       0.9119485   0.0306988865
  9                  20       0.9115324   0.0466066309
  9                  30       0.9112534   0.0586582179
  9                  40       0.9102806   0.0609814875
  9                  50       0.9101412   0.0625891214

Tuning parameter 'shrinkage' was held constant at a value of 0.1

Tuning parameter 'n.minobsinnode' was held constant at a value of 3
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were n.trees = 10, interaction.depth =
 9, shrinkage = 0.1 and n.minobsinnode = 3.
```

The plot for this model is shown below:



From above plot, it can be seen that the highest accuracy is for max_tree_depth 5, and number of iterations is 32.

The feature importance plot for this model is shown below:



Above plot shows that, according to boosted trees model, the most important variable is FirstPurch, followed by Monetary and Recency. The plot also shows the least important variables are RefBks and ItalAtlas.

I made predictions on test data using best model and the confusion matrix is shown below:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1474    1
         1  125    0

               Accuracy : 0.9212
                 95% CI : (0.907, 0.934)
    No Information Rate : 0.9994
    P-Value [Acc > NIR] : 1

                  Kappa : -0.0012

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

            Sensitivity : 0.9218
            Specificity : 0.0000
         Pos Pred Value : 0.9993
         Neg Pred Value : 0.0000
             Prevalence : 0.9994
         Detection Rate : 0.9213
   Detection Prevalence : 0.9219
      Balanced Accuracy : 0.4609

       'Positive' Class : 0
```

From the output of confusion matrix for boosted trees, the Sensitivity of model on test data is 0.9218, the Specificity value is 0, the positive predicted value is 0.9993 and the overall accuracy of the model is 0.92. Since the accuracy of the model is pretty good, so this model can be used practically to make predictions for Florence in future.

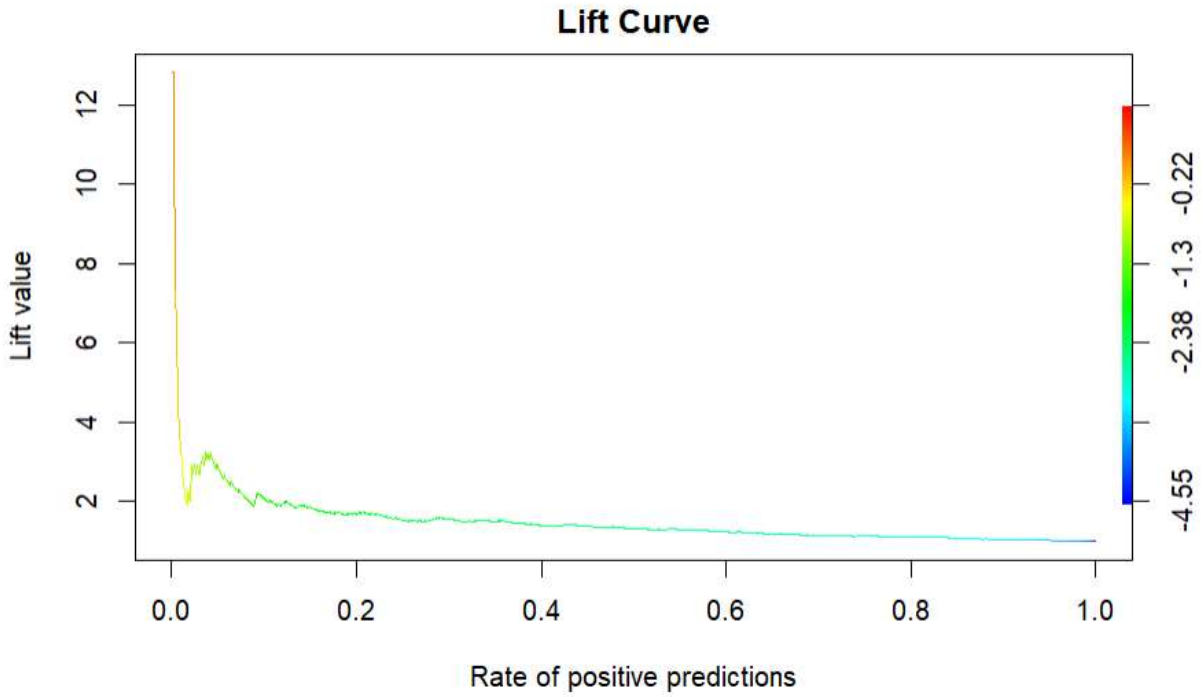
6. Logistic Regression

I implemented logistic regression on training data and made predictions on test data. The summary for logistic regression is shown below:

```
Call:
glm(formula = Florence ~ ., family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2248  -0.4417  -0.3641  -0.2947   2.8536

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.297e+00  2.545e-01  -9.025  < 2e-16 ***
Gender      -4.120e-01  1.529e-01  -2.694  0.00705 **
M            1.432e-05  8.792e-04   0.016  0.98700
R           -3.002e-02  1.551e-02  -1.935  0.05300 .
F            2.220e-01  6.887e-02   3.224  0.00127 **
FirstPurch   4.608e-03  1.121e-02   0.411  0.68110
ChildBks    -2.922e-01  1.109e-01  -2.635  0.00842 **
YouthBks    -1.638e-01  1.417e-01  -1.156  0.24762
CookBks     -4.914e-01  1.139e-01  -4.314 1.61e-05 ***
DoItYBks    -2.438e-01  1.373e-01  -1.775  0.07588 .
RefBks      -1.313e-01  1.479e-01  -0.888  0.37467
ArtBks       4.179e-01  9.913e-02   4.215 2.50e-05 ***
GeogBks      1.424e-01  1.025e-01   1.390  0.16449
ItalCook     1.390e-01  1.879e-01   0.740  0.45948
ItalAtlas   -2.974e-01  3.566e-01  -0.834  0.40432
ItalArt      7.149e-01  2.705e-01   2.643  0.00823 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1438.3  on 2399  degrees of freedom
Residual deviance: 1336.0  on 2384  degrees of freedom
AIC: 1368

Number of Fisher Scoring iterations: 5
```

From above output, I observed that the variables monetary, FirstPurch, YouthBks, RefBks, GeogBks, ItalCook and ItalAtlas don’t have significant relation with dependent variable Florence based on 0.1 confidence level.

The lift chart for logistic regression is shown below:



Above plot represents the Lift Curve for logistic regression model. From above plot, it is pretty evident that with increase in rate of positive predictors the lift value decreased and reached to the minimum value when the rate of positive predictions is 1.
The confusion matrix for logistic regression is shown below:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1475    0
         1  123    2

               Accuracy : 0.9231
                 95% CI : (0.909, 0.9357)
    No Information Rate : 0.9988
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0291

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

            Sensitivity : 0.9230
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.0160
             Prevalence : 0.9988
         Detection Rate : 0.9219
   Detection Prevalence : 0.9219
      Balanced Accuracy : 0.9615

       'Positive' Class : 0
```

From the output of logistic regression, the sensitivity value is 0.923, the specificity value is 1, the positive prediction value is 1 and the accuracy of the model 0.9231. Since the accuracy of the model is pretty good, the model is practically good and can be used for making predictions.