Sri Lanka Institute of Information Technology

SLIIT

COMPUTING | BUSINESS | ENGINEERING

5.CHARLES BOOK CLUB

FDM (Mini group project)

2018.10.07

Group members

- ✓ K.t.p.m. kariyawasam IT16063310
- ✓ S.y.senanayake IT16125308
- ✓ S. I Krusanth IT16122338
- ✓ A. A. Arshad IT16032316
- ✓ A. Arraamuthan IT16086326
- ✓ Gowshalini Rajalingam IT16113800
- ✓ L N kodithuwakku IT16038660

Table of Contents

Introduction

Abstract

This study is intended to aid the CBC Book club which was primarily focused on delivering the tailored book offerings for its customer base via targeted mailing. For this purpose, extracting targeted customer list has been an issue and through this technique, the targeted campaign would be more adequate to be followed. The main goal of this study is to generate a suitable customer list which would positively impact the selling of specialty books.

Data Exploration

The dataset consists of 4000 samples with each record consisting of 24 features. Each feature is described below with their corresponding definition.

Seq# -Sequence number in the training data

ID# -Customer Identification number in test database

Gender -O=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 'The Art History of Florence" was bought, = 0 if not

Related purchase - Number of related books purchased

Mcode- Range corresponds to total money spent; 0-25=1, 26-50=2, 51-100=3, 101-200=4, 201+=5

Rcode- Range corresponds to recent purchase ;0-2=1, 3-6=2, 7-12=3, 13+=4

Fcode- Range corresponds to total number of purchases ;1=1, 2=2, 3+=3

Yes_Florence-bought_art_history_of_florence=1, else 0

No_Florence-bought_art_history_of_florence=0, else 1

Data Preprocessing

All the attributes were found to be numerical

```
In [15]: df.dtypes
Out[15]:
Seq#                int64
ID#                 int64
Gender              int64
M                   int64
R                   int64
F                   int64
FirstPurch          int64
ChildBks            int64
YouthBks            int64
CookBks             int64
DoItYBks            int64
RefBks              int64
ArtBks              int64
GeogBks             int64
ItalCook            int64
ItalAtlas           int64
ItalArt             int64
Florence            int64
Related Purchase    int64
Mcode               int64
Rcode               int64
Fcode               int64
Yes_Florence        int64
No_Florence         int64
```

Figure 1. Data types of attributes

In the lines of pre-processing, it is found that the following fields are highly co relevant than the other attributes which are too low to be noticed.

- Gender
- M
- R
- F
- FirstPurch
- Related purchase
- Florence

Index	ted Purchase	Florence
Seq#	0102938	-0.018865
ID#	0103943	-0.018829
Gender	0133754	-0.059338
M	22495	0.0345612
R	340395	-0.0596788
F	43442	0.0796338
FirstPurch	1485	0.0352965
Related Purchase		0.120133
Florence	20133	1
Mcode	19909	0.0166292
Rcode	36179	-0.0634062
Fcode	36921	0.0529668

Figure 2. Correlation of independent attributes between dependent (Florence) attribute

There were no values missing for any attributes

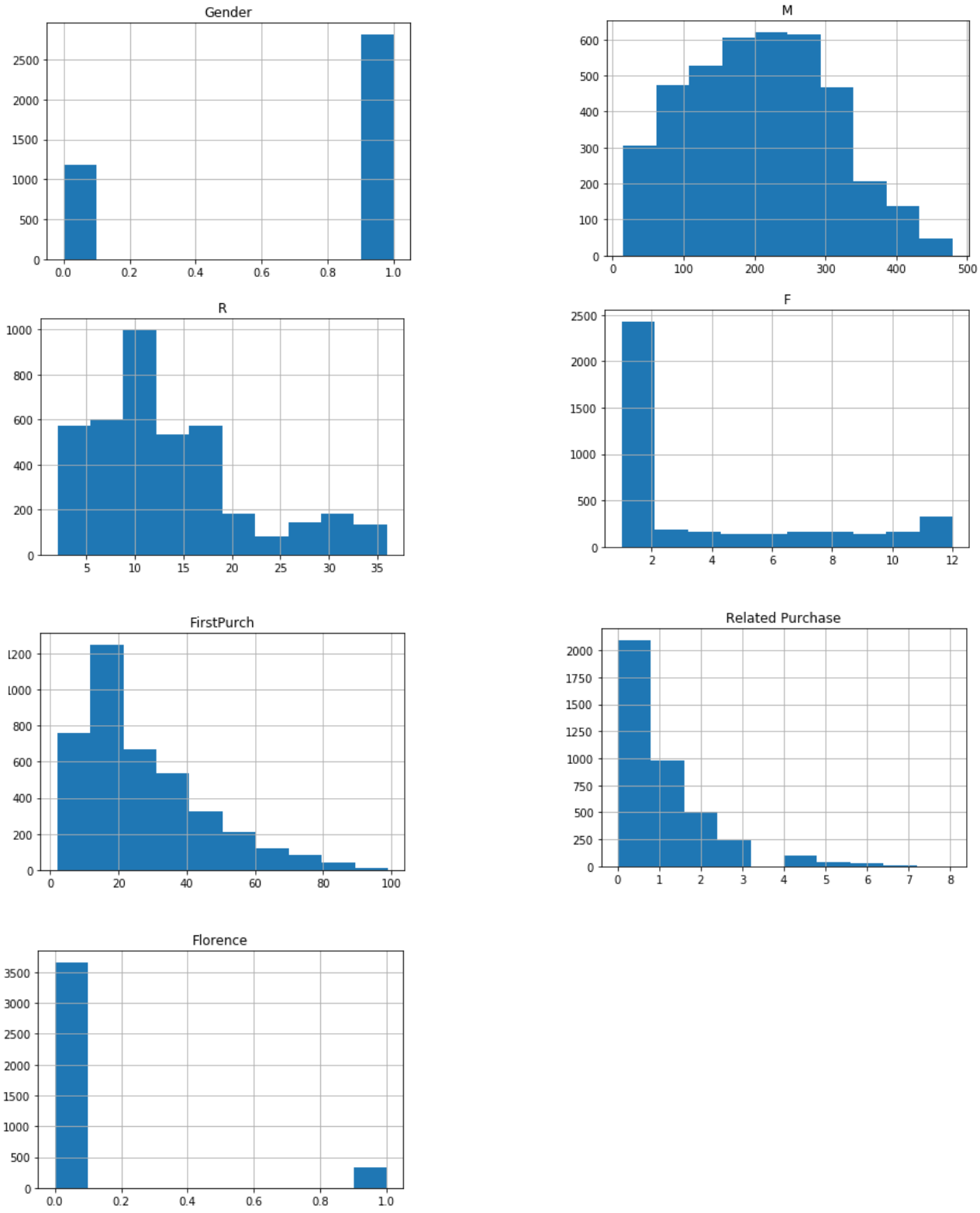


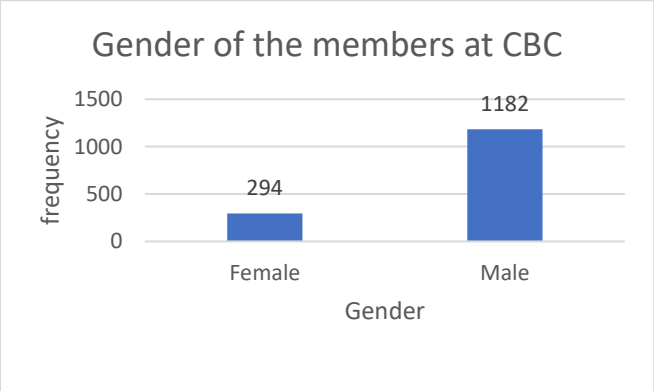
Figure 3. Frequency distribution of Attributes

1.Descriptive Analysis

The descriptive analysis has been done to understand the customer base. The analysis focused on 4 areas. Such as gender analysis, Analysing money spend on books. Popularity of books, most recent purchase data.

Gender Analysis

As the gender is considered as a nominal data, and there is no order to the segments, pie chart and bar chart can be used to display different percentages representing male and female customers at CBC. Females make up 70.45% of the customers whereas males make up only 29.55% of the total customers. The total number of female customers is more than the total number of male customers.



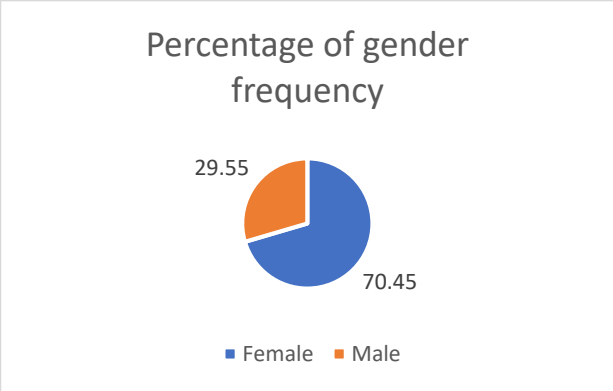


Figure 4. Bar char and pie chart of gender analysis

Monetary VS Gender

A typical customer spends $76.35 on Average to by books. Female customers are spending more money to buy books at CBC.

Sum of M	
Gender	Total
0	248033
1	584333
Grand Total	832366
Avg	208.0915

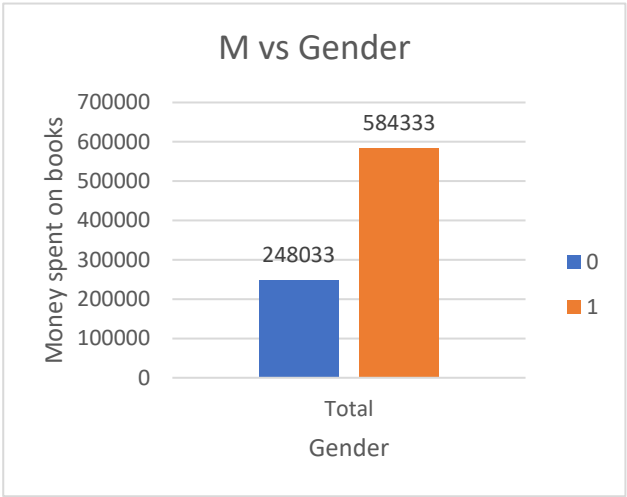
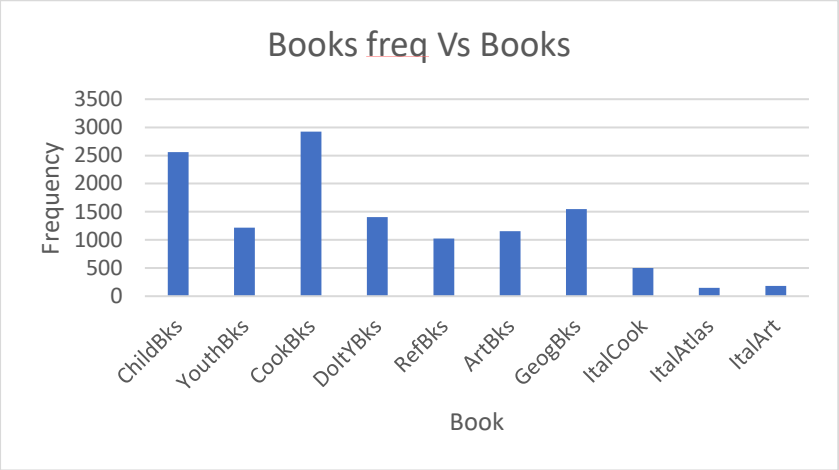


Figure 5. Table and Bar chart of money spent on books analysis

Popularity of books

Child and Cook books are most popular books among other books. The popularity is approximately similar for Youth, DoItY, Ref, Art and Geo Books.The popularity of ItalCook, ItalAtlas, ItalArt are less.



ChildBks	YouthBks	CookBks	DoItYBks	RefBks	ArtBks	GeogBks	ItalCook	ItalAtlas	ItalArt
2559	1219	2925	1403	1025	1156	1550	501	150	183

Figure 6. Table and Bar chart of popularity of books

Most recent purchase data

Most customers at CBC purchased their latest books in more than 13 months since the previous purchases. Customer frequency is increasing with recency.

Rcode	Frequency
1	294
2	558
3	1322
4	1826

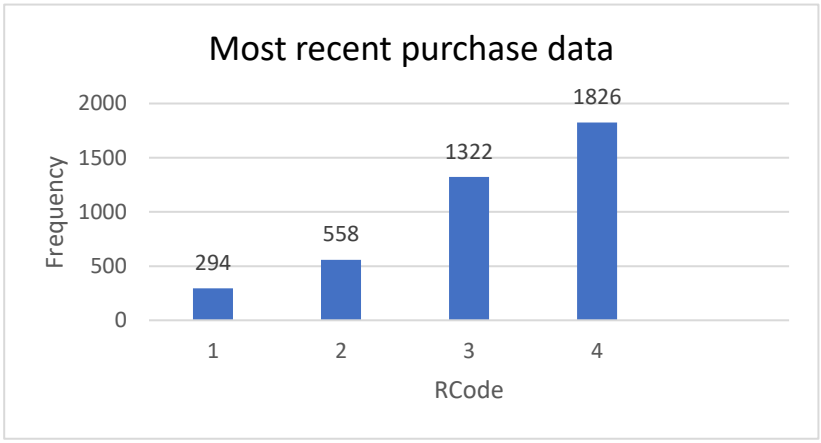


Figure 7. Table and Bar chart of most recent purchased data

2.Predictive Analysis

Segmentation of Dataset

Data set is divided into 3 parts. such as training set, Validation set, test set.

Training set is 45% of data set.

Validation set is 35% of data set.

Test set is 20% of data set.

Data Mining Techniques Used

We used classification technique as our data mining technique. Because in the given case study we were asked to find the target customers for direct mail promotion. Therefore, we used binary classification technique to find whether the customer is a target customer for direct mail promotion or not.

Fitted models

We fitted all the classification models and checked the accuracy score and the log loss for each model. Finally, we picked a model as best model which is fitted using GradientBoostingClassifier Algorithm.

Accuracy score= 90.357%

Log loss= 0.3758 (if log loss is 0 prediction is perfect)

Results

As we were asked to compute a score for each customer and use this score and a cutoff value to extract a target customer list for direct mail promotion we used probability scoring technique to score each customer.

The optimal cut off value is found where the true positive rate is high and false positive rate is low. The cut off value is 0.300772.

	fpr	tpr	1-fpr	tf	thresholds
2	0.467777	0.568047	0.532223	0.035825	0.300772

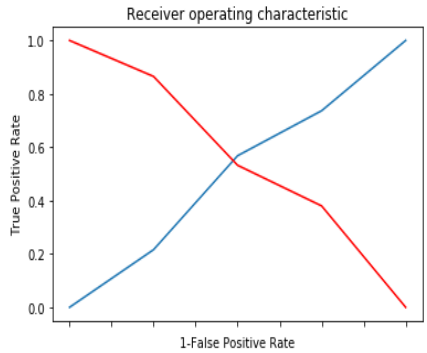


Figure 8. Test results

Visualization on Dash Board

We have created the dashboard which can predict

1.For a given book category of a new title List of target users

2.For a given customer the predicted results whether the customer will by the offered book or not

We use python flask to create the dashboard

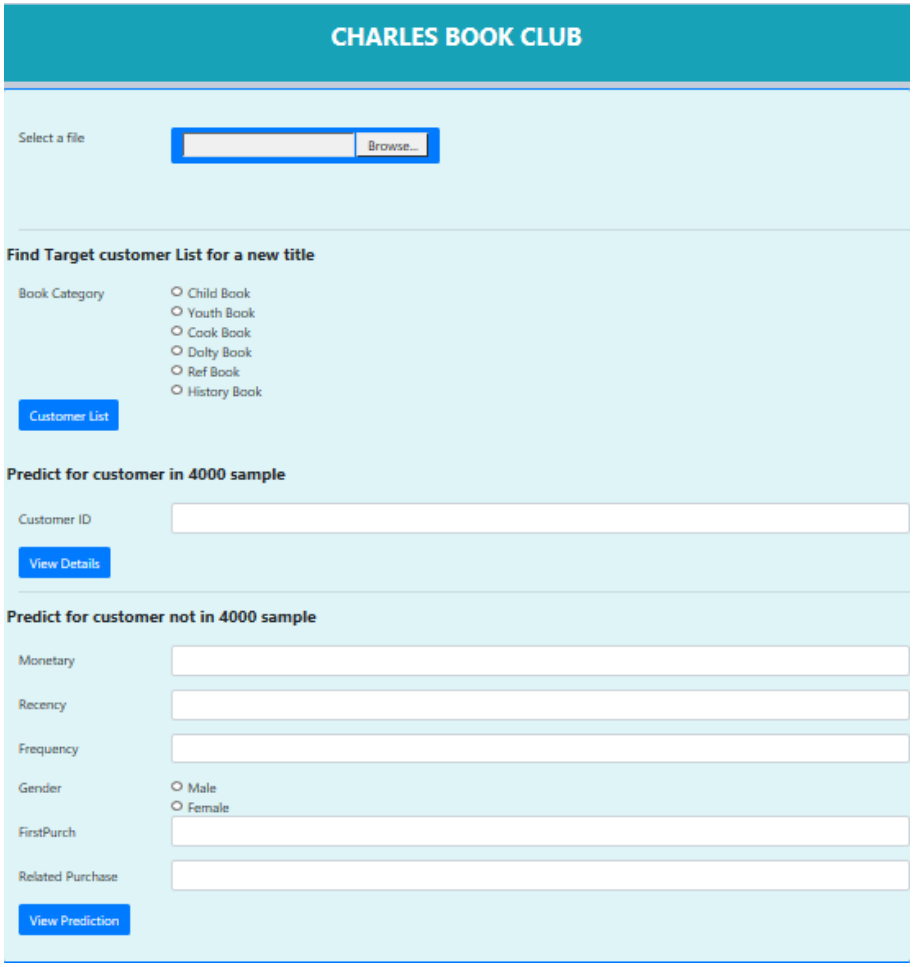


Figure 9. index page

TARGET CUSTOMER LIST

	Seq#	ID#	Gender	M	R	F	FirstPurch	Related Purchase	Yes_Florence	No_Florence	ArtBks	ChildBks	CookBks	DoItYBks	Fcode	Florence	GeogBks	ItalArt	ItalAtlas	ItalCook	Mcod
19	20	145	1	393	12	11	50	5	1	0	2.0	3.0	2.0	0.0	3.0	1.0	3.0	0.0	0.0	0.0	5.0
39	40	330	0	211	6	11	44	6	0	1	2.0	3.0	2.0	0.0	3.0	0.0	1.0	1.0	0.0	2.0	5.0
56	57	440	1	458	10	12	44	6	1	0	2.0	1.0	3.0	1.0	3.0	1.0	0.0	1.0	1.0	2.0	5.0
86	87	708	0	277	30	11	78	5	1	0	2.0	2.0	1.0	1.0	3.0	1.0	3.0	0.0	0.0	0.0	5.0
89	90	723	1	215	2	8	36	5	1	0	3.0	0.0	0.0	0.0	3.0	1.0	2.0	0.0	0.0	0.0	5.0
101	102	874	1	274	14	11	68	6	0	1	1.0	4.0	2.0	1.0	3.0	0.0	2.0	1.0	0.0	2.0	5.0
112	113	966	1	266	2	6	30	2	0	1	1.0	1.0	0.0	0.0	3.0	0.0	1.0	0.0	0.0	0.0	5.0
168	169	1394	1	275	8	10	62	6	0	1	2.0	1.0	1.0	2.0	3.0	0.0	0.0	2.0	1.0	1.0	5.0
193	194	1671	0	250	12	12	36	6	1	0	3.0	1.0	3.0	2.0	3.0	1.0	0.0	1.0	1.0	1.0	5.0
243	244	2124	1	441	10	12	72	5	1	0	2.0	2.0	0.0	2.0	3.0	1.0	3.0	0.0	0.0	0.0	5.0
287	288	2527	0	415	16	12	68	4	1	0	1.0	5.0	0.0	2.0	3.0	1.0	3.0	0.0	0.0	0.0	5.0
314	315	2812	0	268	12	7	46	3	1	0	2.0	0.0	2.0	1.0	3.0	1.0	0.0	0.0	0.0	1.0	5.0
411	412	3684	1	144	6	8	38	5	0	1	1.0	0.0	2.0	1.0	3.0	0.0	1.0	1.0	1.0	1.0	4.0
563	564	5112	1	338	26	5	40	5	0	1	1.0	0.0	2.0	0.0	3.0	0.0	1.0	1.0	1.0	1.0	5.0
576	577	5228	1	406	14	10	72	3	0	1	1.0	1.0	3.0	1.0	3.0	0.0	2.0	0.0	0.0	0.0	5.0
597	598	5417	0	288	6	11	70	4	1	0	2.0	3.0	2.0	1.0	3.0	1.0	0.0	0.0	0.0	2.0	5.0
664	665	5902	1	351	10	9	62	5	1	0	2.0	2.0	2.0	1.0	3.0	1.0	0.0	2.0	0.0	1.0	5.0
669	670	5912	1	399	14	8	48	8	1	0	2.0	0.0	2.0	0.0	3.0	1.0	3.0	1.0	1.0	1.0	5.0

Figure 10. Target customer list page

Probability score is:0.04695594803072611
Customer Will not Buy the Book

Key	Value
ID	25
Gender	1
total_expenditure	297
months_since_last_purchase	14
number_of_purchases	2
months_since_first_purchase	22
ChildrensBooks_purchased	0
YouthBooks_purchased	1
CookBooks_purchased	1
DoityourselfBooks_purchased	0
Dict_Encycl_Atlases_purchased	0
ArtBooks_purchased	0
GeographyBooks_purchased	0
Secrets_Italian_Cooking	0

Figure 11. Predict for specific customer who is in the 4000 sample page

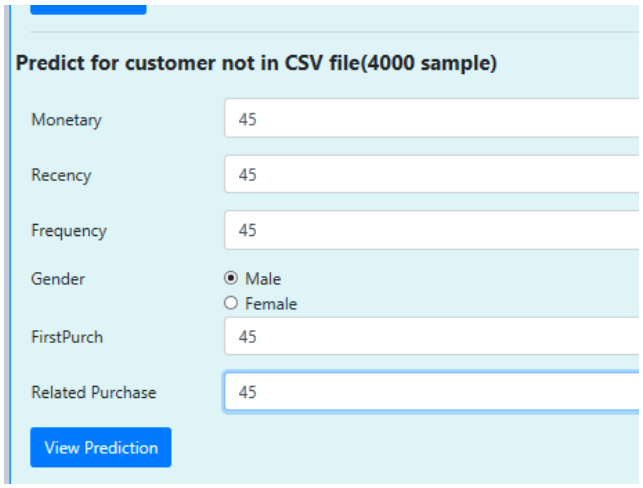


Figure 12. Predict for specific customer who is not the 4000 sample page

Probability score is:0.9972687843810307
Customer Will Buy the Book

Figure 13 The result of figure 8

Conclusion

Gradient Boosting Classifier Algorithm was found to be the best model to fit and a probabilistic model approach was used to derive scores and an appropriate cutoff. If the predicted probability for given data (values of independent variable) is greater than the cutoff threshold value,0.300772, then the customer is eligible for direct mail promotions for specialty book title. That is, they have higher tendency to buy that given book title under given promotional mail campaign. Thus, CBC book club could achieve their goal realized in an effective way provided that they follow this approach.