STAT 551 MARS Assignment

Yuchi Hu

April 21, 2019

1. Introduction

The data for this assignment comes from RetentionDataRaw.xlsx, which we have already used extensively in the Midterm. The data is composed of credit card monthly billing statements from February 2010 to November 2010 (shorter time frames for closed accounts). The task now is to analyze the data using multivariate adaptive regression splines (MARS) and logistic regression with binned variables. The response is the *Bad* variable, which we create and define, and the predictors are *Good Customer Score*, *Behavior Score*, and *Quarterly Fico Score*. Before we can do any modeling, we need to create the "model dataset" from the retention data.

2. Creating the Model Dataset

The retention data originally has 97,465 rows and 26 columns. After removing the rows with missing ID's, we are left with 91,502 rows representing 9,997 unique ID's. To create the model dataset, we need to remove customers that we deem to be too risky. We also need to define "good" and "bad" customers.

2.1 Remove Risky Customers

We found out from the Midterm that some customers are too risky to give more money to, so we will simply remove them from the data. The customer is deemed too risky if their *Row Num* = 1 (month 1) satisfies any of the following conditions:

- *Days Deliq* > 0 (more than 0 days delinquent)
- non-blank *External Status* (a blank *External Status* corresponds to an open account)
- *Opening Balance* $>$ *Credit Limit*
- *Ending Balance* $>$ *Credit Limit*

We find that 4,167 of the 9,997 unique ID's are too risky, so we are left with 5,830 unique ID's after removing the risky customers. Overall, we are left with 55,895 rows of data.

2.2 Define Good and Bad Customers

Next, for the remaining customers, we create the *Bad* variable (the response) to define whether they are good (*Bad*=0) or bad (*Bad*=1) customers. A customer is defined as bad if they satisfy any of the following conditions:

- *Days Deliq* ≥ 90 (90 or more days delinquent) in the final month
- *External Status* other than blank (open account) or "C" (closed account) in month 7 or later

We find that 610 unique ID's satisfy the first criterion and 669 unique ID's satisfy the second criterion. Customers who do not satisfy any of the two criteria are defined as good.

2.3 Final Step

After removing the risky customers and defining *Bad* for the remaining customers, the final step in the creation of the model dataset is to keep only *Row Num* = 1 (month 1) for each unique ID.

3. Exploratory Data Analysis on the Model Dataset

Now that we have created the model dataset, we can perform exploratory data analysis on it. The model dataset is composed of 5,826 rows (number of unique ID's) and 27 variables (26 original variables + *Bad*). The response is *Bad*, and the predictors are *Good Customer Score*, *Behavior Score*, and *Quarterly Fico Score*. For the MARS model, the predictors will be used as is, while for logistic regression, the predictors will be binned.

First, let's see how many good and bad customers are in the model dataset.

```
##    0    1
## 5073  753
```

We see that there are 5,073 good customers and 753 bad customers.

Figure 1: Histograms of the three scores grouped by Bad (Good=0, Bad=1).

Figure 1 shows the histograms of the three scores grouped by *Bad*. We see that good and bad customers have similar score distributions. We also see some 0's for *Good Customer Score* and *Quarterly FICO Score*; in the case of *Quarterly FICO Score*, the number of 0's is significant. We should also note that there are NA's for *Good Customer Score* as well.

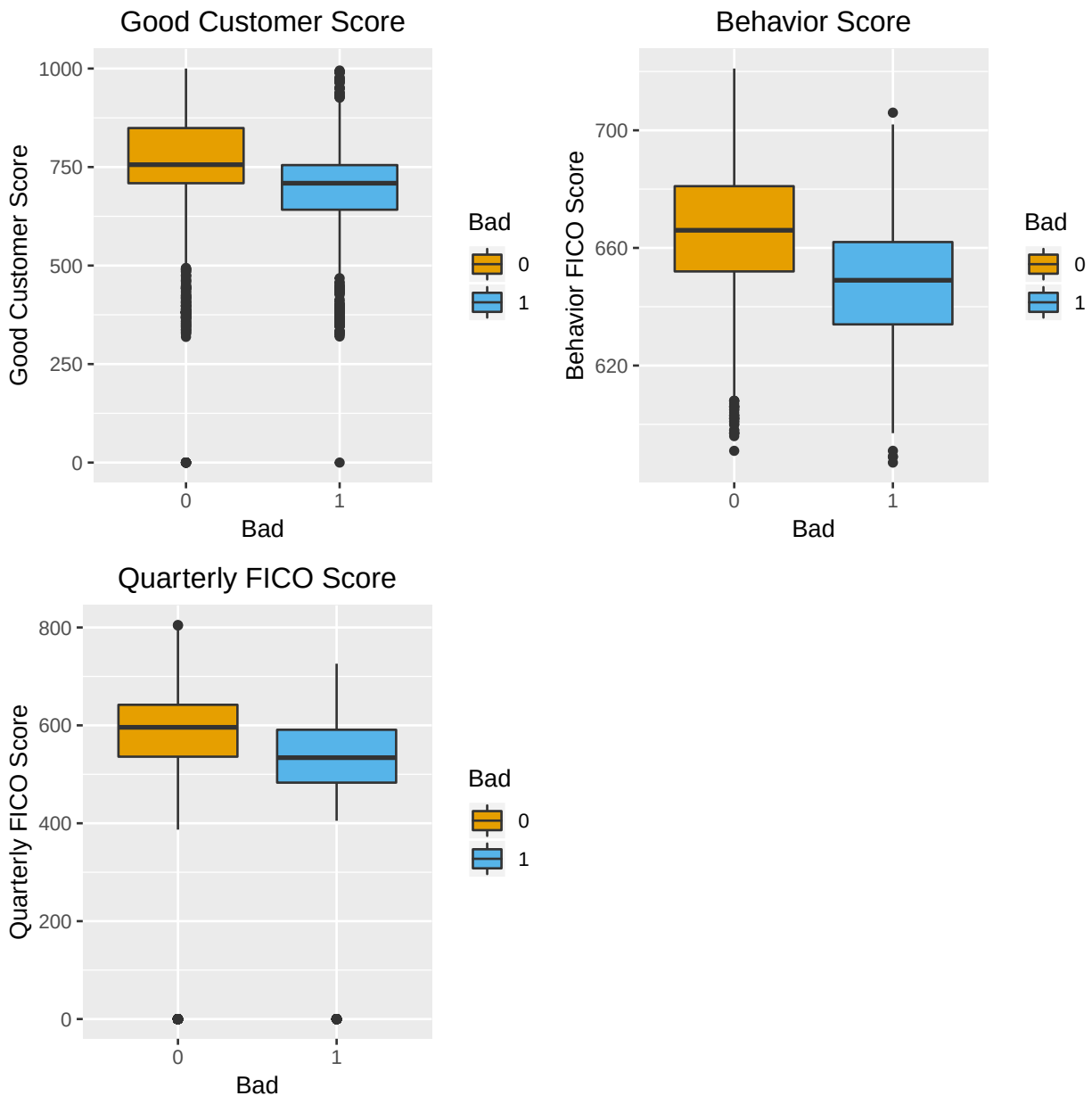

Figure 2: Boxplots of the three scores grouped by Bad (Good=0, Bad=1).

Figure 2 shows the boxplots of the three scores grouped by *Bad*. We see that good customers tend to have higher scores than bad customers.

As previously mentioned, *Good Customer Score* has 0's and NA's, and *Quarterly FICO Score* has 0's. In order to properly deal with these odd values, we should bin the scores; that is, convert the scores into categorical variables.

To bin the scores, we use **bin()** from the **OneR** package. We specify "cluster" (1D K-means clustering or Jenks natural breaks optimization) as the method of binning. This method reduces within-class variance and maximizes between-class variance. All of the 0's and NA's are put into their own bins, while the rest of the values are binned according to **bin()**.

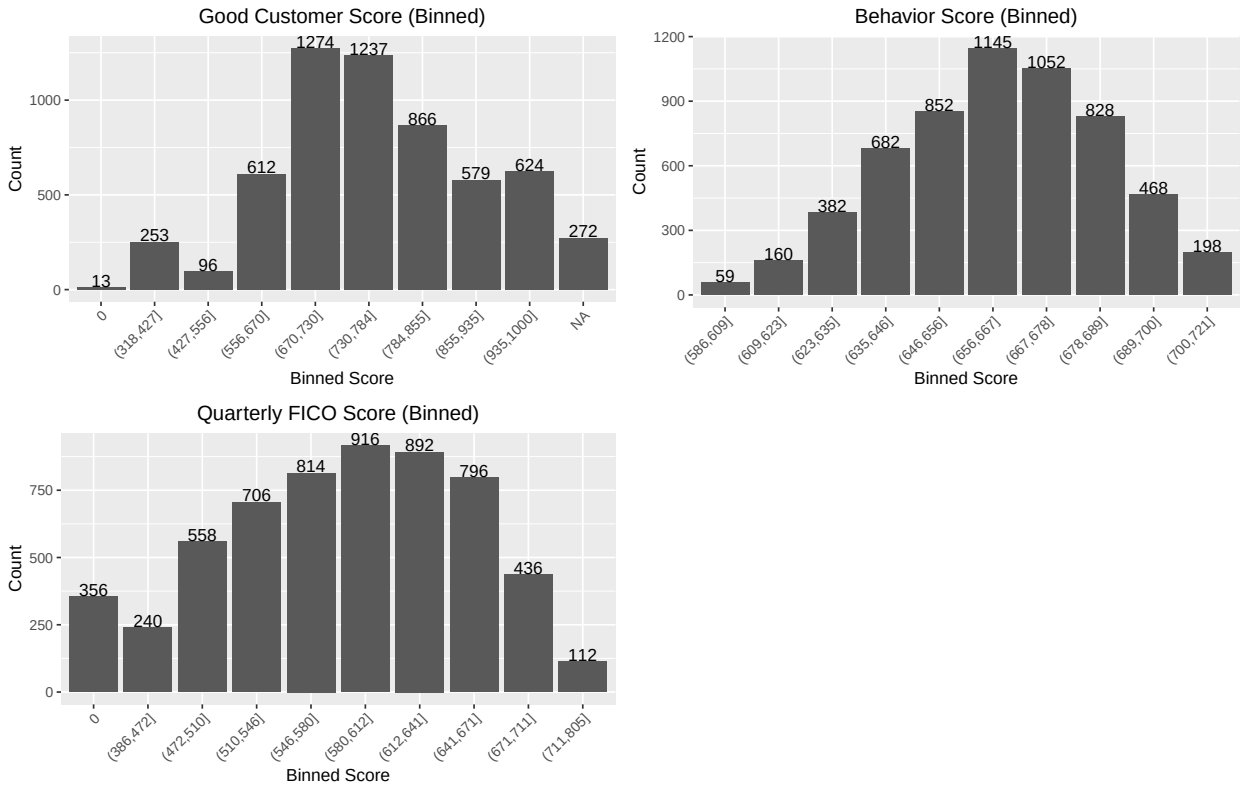

Figure 3: Bar charts of the binned scores.

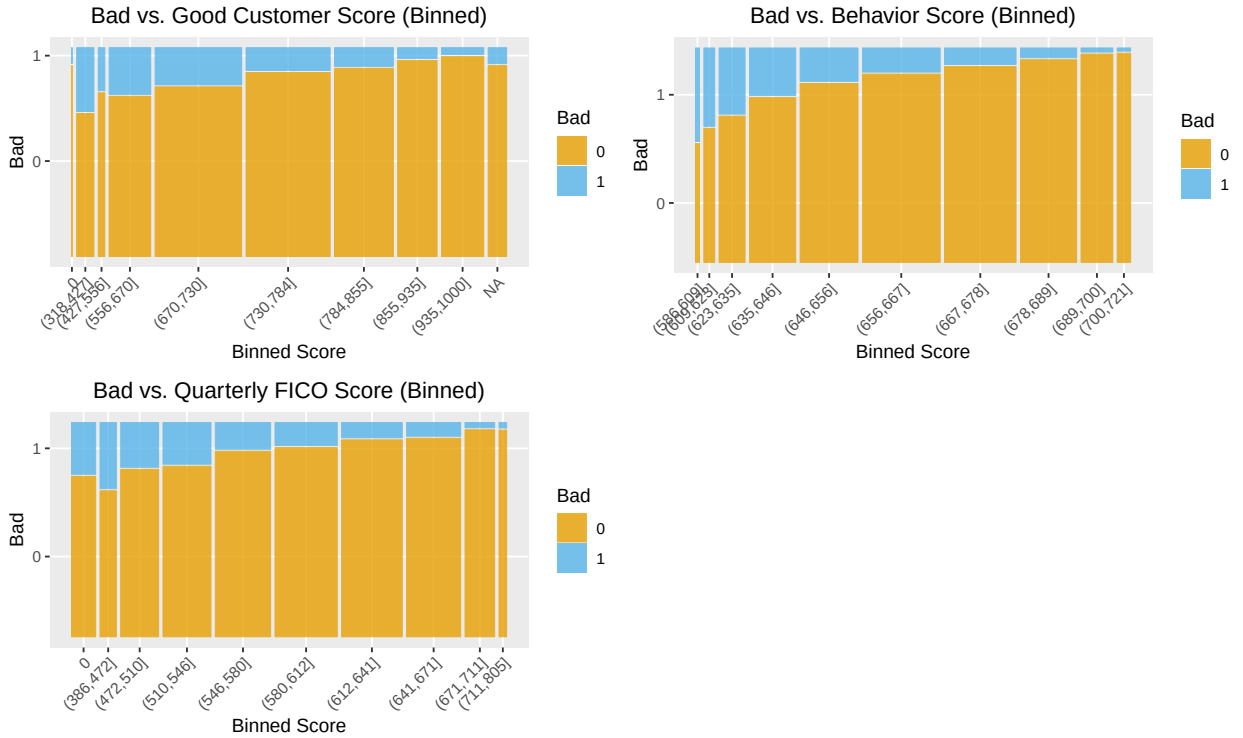

Figure 4: Mosaic plots of Bad vs. binned scores.

Figure 3 shows the bar charts of the three scores after binning. All three bar charts appear approximately normal. **Figure 4** shows the mosaic plots of *Bad* vs. the binned scores. We see that the relative proportion of bad customers (light blue) mostly decreases as scores increase.

4. The Models

Before building the models, we use **createDataPartition()** from the **caret** package to split the model dataset into training (70%; 4,080 observations) and validation sets (30%; 1,746 observations). This function partitions the data while maintaining the class ratios (stratified sampling).

4.1 MARS

We build a MARS model with *Bad* as the response and the three scores (unbinned) as the predictors.

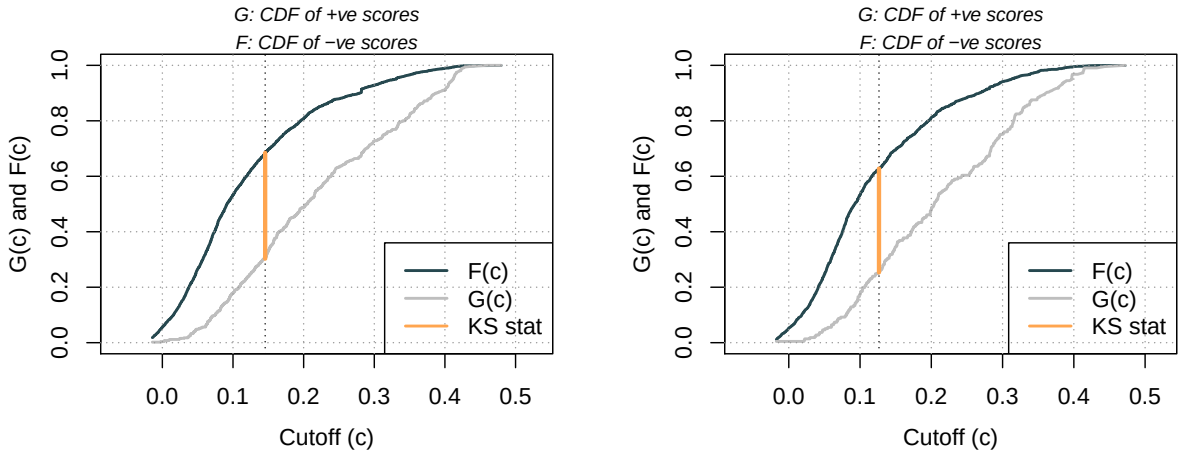

Figure 5: KS plots of training (left) and validation sets (right) for MARS.

Figure 5 shows the KS plots of the training and validation sets for MARS.

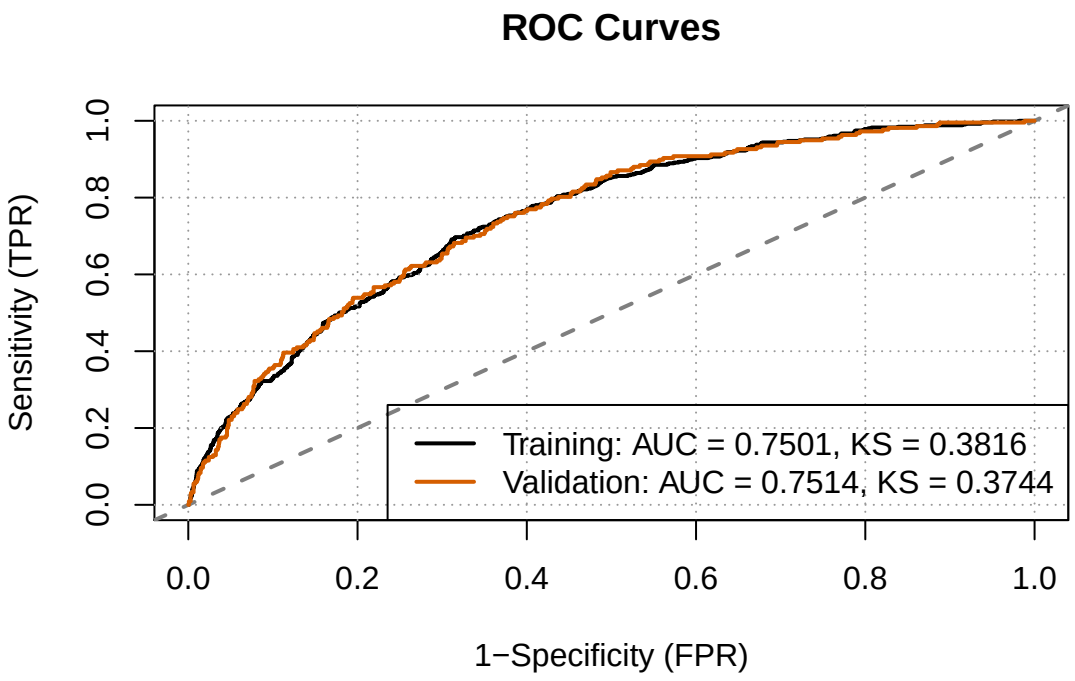

Figure 6: ROC curves of training and validation sets for MARS.

Figure 6 shows the ROC curves of the training and validation sets for MARS. The training and validation AUC's are 0.7501 and 0.7514, respectively. The training and validation KS statistics are 0.3816 and 0.3744, respectively.

Table 1: Gains Table (Validation)

Bucket	Obs	CObs	Depth	Resp	CResp	RespRate	CRespRate	CCapRate	Lift	CLift
1	166	166	0.1	61	61	0.37	0.37	0.28	2.81	2.81
2	166	332	0.2	39	100	0.23	0.30	0.46	1.80	2.30
3	166	498	0.3	30	130	0.18	0.26	0.60	1.38	2.00
4	166	664	0.4	25	155	0.15	0.23	0.71	1.15	1.79
5	166	830	0.5	21	176	0.13	0.21	0.81	0.97	1.62
6	166	996	0.6	18	194	0.11	0.19	0.89	0.83	1.49
7	166	1162	0.7	7	201	0.04	0.17	0.93	0.32	1.32
8	166	1328	0.8	8	209	0.05	0.16	0.96	0.37	1.20
9	166	1494	0.9	6	215	0.04	0.14	0.99	0.28	1.10
10	166	1660	1.0	2	217	0.01	0.13	1.00	0.09	1.00

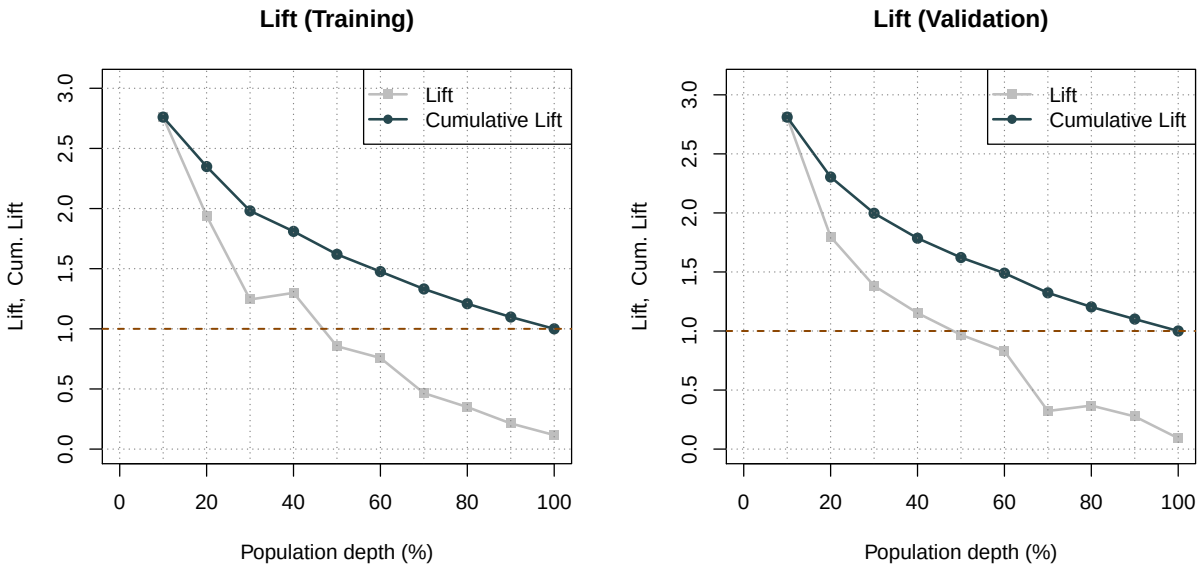

Figure 7: Lift and cumulative lift of training (left) and validation sets (right) for MARS.

Table 1 shows the gains table of the validation set, and **Figure 7** shows the lift and cumulative lift of the training and validation sets for MARS.

4.2 Logistic Regression

We build a logistic regression model with *Bad* as the response and the three scores (binned) as the predictors.

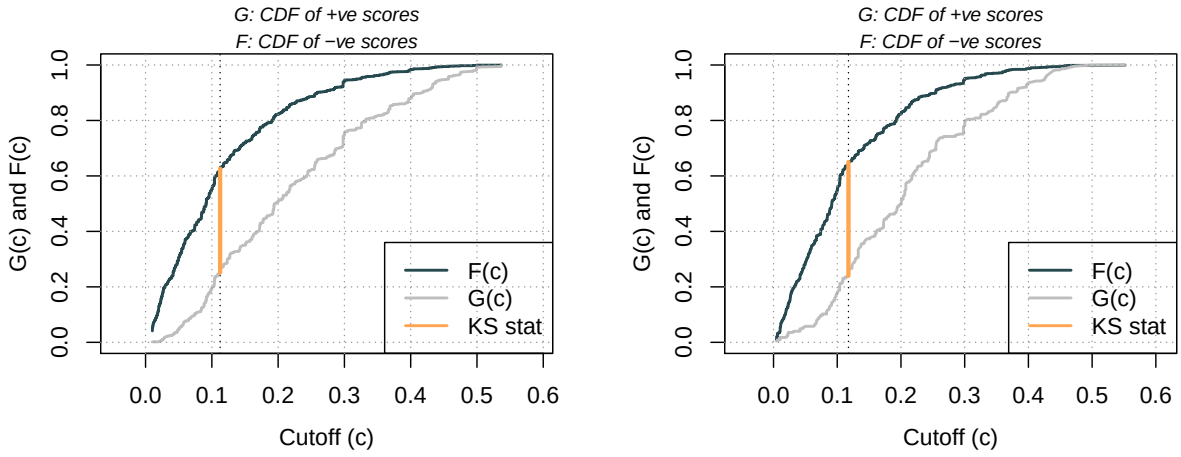

Figure 8: KS plots of training (left) and validation sets (right) for logistic regression.

Figure 8 shows the KS plots of the training and validation sets for logistic regression.

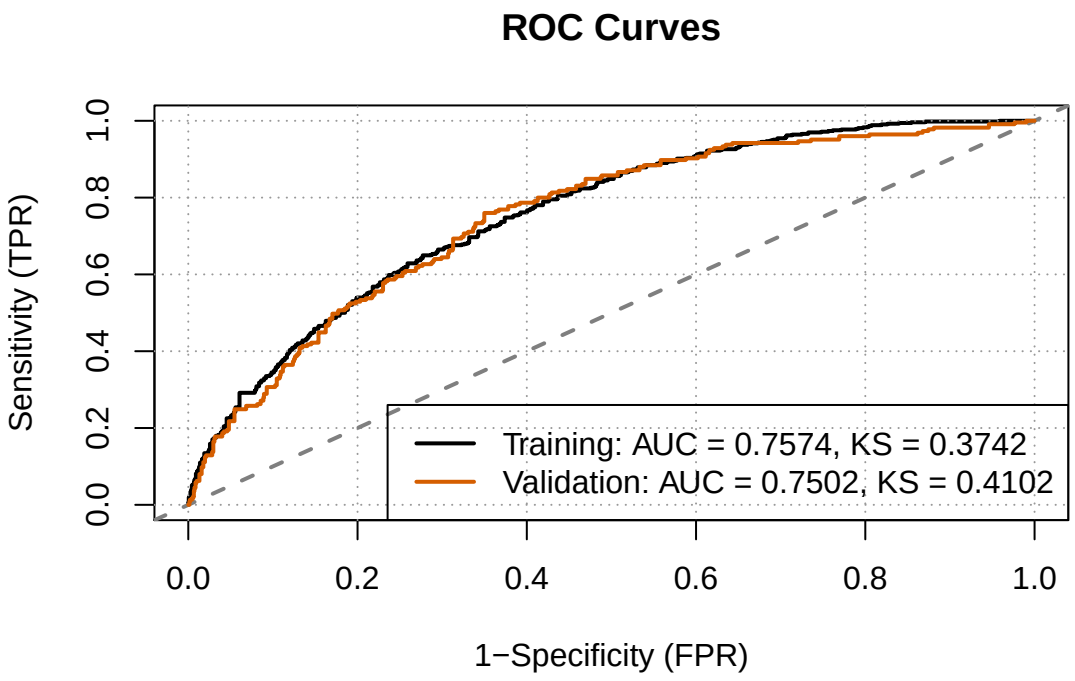

Figure 9: ROC curves of training and validation sets for logistic regression.

Figure 9 shows the ROC curves of the training and validation sets for logistic regression. The training and validation AUC's are 0.7574 and 0.7502, respectively. The training and validation KS statistics are 0.3742 and 0.4102, respectively.

Table 2: Gains Table (Validation)

Bucket	Obs	CObs	Depth	Resp	CResp	RespRate	CRespRate	CCapRate	Lift	CLift
1	175	175	0.1	58	58	0.33	0.33	0.26	2.57	2.57
2	174	349	0.2	44	102	0.25	0.29	0.45	1.96	2.27
3	175	524	0.3	34	136	0.19	0.26	0.60	1.51	2.01
4	174	698	0.4	30	166	0.17	0.24	0.74	1.34	1.85
5	175	873	0.5	19	185	0.11	0.21	0.82	0.84	1.64
6	175	1048	0.6	14	199	0.08	0.19	0.88	0.62	1.47
7	174	1222	0.7	13	212	0.07	0.17	0.94	0.58	1.35
8	175	1397	0.8	4	216	0.02	0.15	0.96	0.18	1.20
9	174	1571	0.9	5	221	0.03	0.14	0.98	0.22	1.09
10	175	1746	1.0	4	225	0.02	0.13	1.00	0.18	1.00

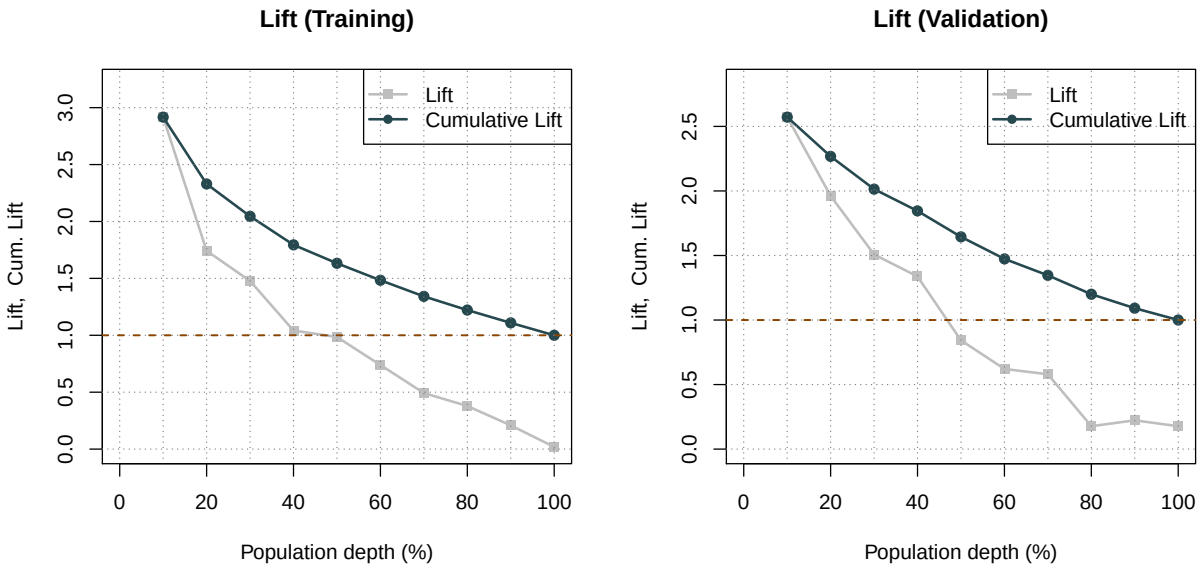

Figure 10: Lift and cumulative lift of training (left) and validation sets (right) for logistic regression.

Table 2 shows the gains table of the validation set, and **Figure 10** shows the lift and cumulative lift of the training and validation sets for logistic regression.

5. Conclusion

The MARS and logistic regression models performed about equally well. The validation AUC's for MARS and logistic regression are 0.7514 and 0.7502, respectively. The KS statistics for MARS and logistic regression are 0.3744 and 0.4102, respectively. There are advantages and disadvantages to both models. For MARS, we used the original scores as predictors. These scores are continuous, and *Good Customer Score* contains NA's, which cannot be handled by MARS. Thus, we had to remove the NA's, potentially costing us valuable information. Despite that, MARS can provide great predictive power by capturing non-linear relationships through piecewise linear models. For logistic regression, we binned the scores, converting them into categorical variables. In this way, we can include 0's, NA's, or any other odd values we encounter as distinct categories. On the other hand, binning can result in too many bins and loss of information.