Stat 551 Final Project White Paper

Nikhil Karthik Pamidimukkala

April 29, 2019

1. Introduction

The data we have is a credit card transaction data for 9997 customers. After several stages of analysis of this, in this step the analysis is focused on good customers so that they can be given incentives to make the them use the card more to make money. These incentives can include credit line increases or annual fee waivers. However, defining a a good or bad customer should be done with great care as wrong analysis could cost millions.

2. Data Pre-processing

Certain criterias are adopted for filtering out bad customers after the data analysis in previous step which include.

- Customers with Days Delinquent >0 in their first transaction record in the data
- Customers with non-blank external status (Account froze, revoked, closed etc.)
- Customers who have an ending/opening balance which exceeds credit limit.

All the customers who meet any of the above criteria are removed from the data.

3. Defining Bad

Even after filtering out the bad customers from the data, a Binary response variable Bad(1/0) is created based on the following criteria:

- Customers with Final transaction record in data / final month having Days Deliquent greater or equal to 90 are categorized as Bad(Bad = 1)
- Customer with Final transaction record in data / final month having external status other than 'Closed' and 'Open' with this status being month 7 or greater are categorized as Bad.
- All the other customers are categorized as Good (Bad = 0).

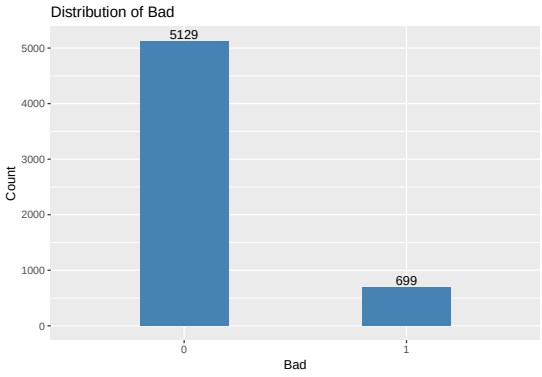

Figure 1: Distribution of Bad

4. Variable Creation

To find the predictive power in Variables, Supervised Discretization has been used which bins a continuous feature mapped to a target variable. The central idea is to find those cutpoints that maximize the difference between the groups. This is done using the smbinning function which conditional inference trees to determine cut points. After examining whether there were any significant with the variables, certain transformation of have variables have been done. To find directionality, bad rate vs binned variables was plotted.

- **Months on Book** : Months on books was found to have significant splits and it was binned to whether bad rates where changing with months on books. Figure 2 shows that Bad Rate decreases with increase in Months on Books.

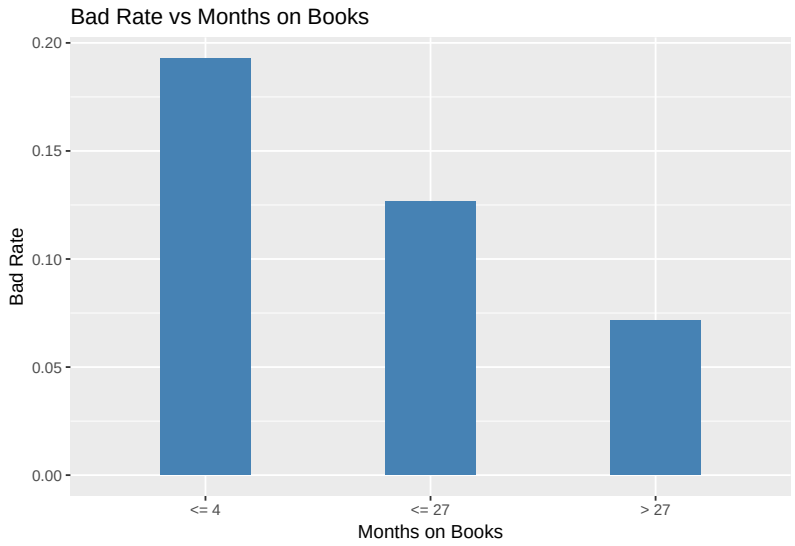

Figure 2: Bad Rate vs Months on Books

- **Due Proportion** : Due Proportion is ratio of total minimum payment due to credit limit. It is binned using smbinnig fuction to see whether bad rate chanes with Dueproportion. Figure 4, shows that Bad Rate increasing with increase in Due Proportion.

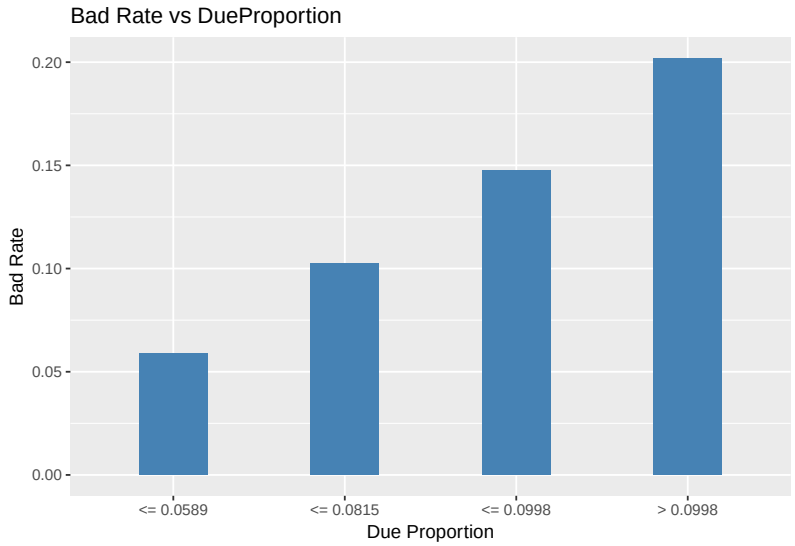

Figure 3: Bad Rate vs Due Proportion

- **Balance Proportion** : Balance Proportion variable is defined as follows
- If Opening and Ending Balance is negative , Balance proportion is zero.
- If Opening Balance is positive and Ending Balance Negative, Balance Proportion is zero.
- If Opening Balance is negative and Ending balance positive, Balance Proportion is Ending Balance/Credit Limit.
- If Opening and Ending balance is positive, Balance Proportion is mean of Opening and Ending Balance/ Credit Limit.

Figure 4 shows that Bad Rate increases as Balance Proportion increases.

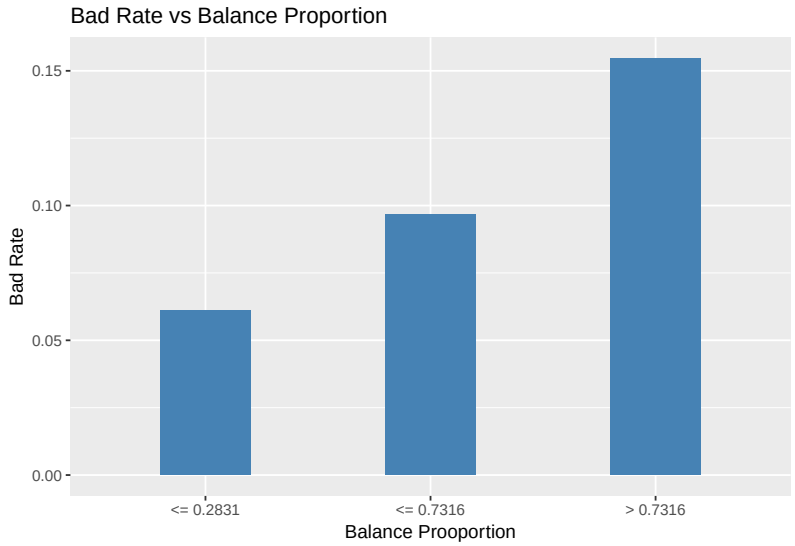

Figure 4: Bad Rate vs Balance Proportion

5. Model Building

The variables created monotonically change with Bad Rate. Finally we consider binned version of Months on Books, Binned version of Due Proportion and continuous Balance Proportion as predictors. MARS and Logistic Regression models are considered to model Bad(1/0). The data set is split into 60% training and 40% validation set.

5.1 MARS Model

Multivariate adaptive regression splines (MARS) provide a convenient approach to capture the nonlinearity aspect of polynomial regression by assessing cutpoints (knots) similar to step functions. The procedure assesses each data point for each predictor as a knot and creates a linear regression model with the candidate feature. The coefficients of the MARS model in Table 1 shows the coefficients of the MARS model. The values of these coefficients show what the Bad Rate vs Binned variable in figure 2,3,4 show.s i.e. as month in books increase the probability of bad decreases etc.

Coefficient	Value
Intercept	-2.5622088
monthscut<=27	-0.4549359
monthscut>27	-0.7245608
duepropcut<=0.0815	0.6105836
duepropcut<=0.0998	0.8165267
duepropcut>0.0998	1.0780799
h(BalProp-0.6124)	2.4546609

Table 1: Coefficients of MARS Model

5.1.1 ROC Curve for MARS

The ROC chart shows false positive rate (1-specificity) on X-axis against true positive rate (sensitivity) on Y-axis Ideally, the curve will climb quickly toward the top-left meaning the model correctly predicted the cases. The diagonal line is for a random model. ROC Curve can also help in selecting an optimal classification threshold which gives a balanced true positive and false postive rates.

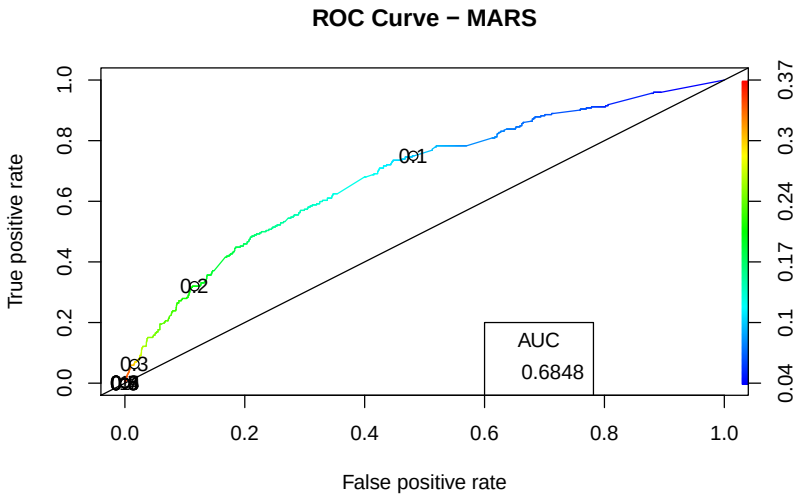

Figure 5: ROC Curve for MARS model

5.1.2 Confusion Matrix for MARS

The clasification threshold was set to 0.15. The Accuracy obtained by the MARS model is 71.16%. The specificity is 52.94 % and sensitivity is 73.69%. The positive class considered here is 0 i.e Good.

	Actual	
Prediction	0	1
0	1440	128
1	514	144

Table 2: Confusion Matrix for MARS Model

5.1.3 KS Curve

K-S is a measure of the degree of separation between the positive and negative distributions. The KS stat value is 0.27. Zero KS value indicates the model selects cases randomly from the population.

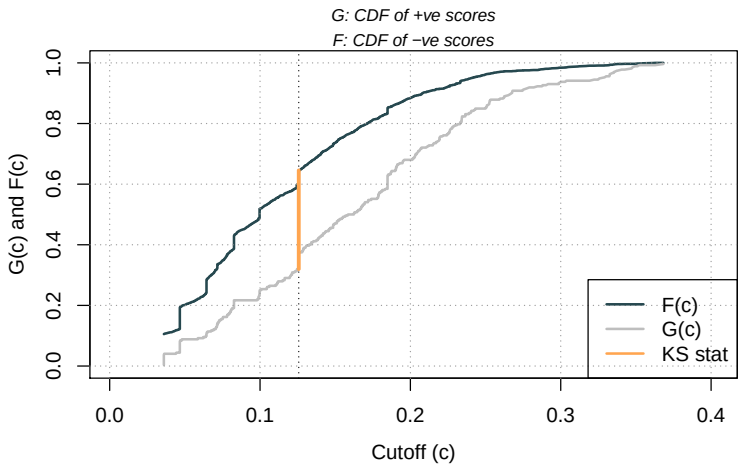

Figure 6: KS Curve for MARS model

5.1.4 Lift Chart for MARS

The lift chart shows how much more likely we are to select good customers than if we select a random sample of customers. The lift chart shows that considering only the first 10% of the customers will allow in selecting 2.4 times more good customers using the predictive model than done randomly.

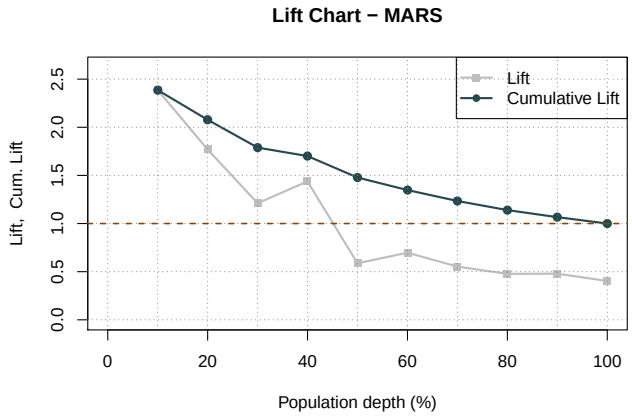

Figure 7: Lift Chart - MARS

5.1.5 Gains Table for MARS

Depth of File	N	Cume N	Mean Resp	Cume Mean Resp	Cume Pct of Total Resp	Lift Index	Cume Lift	Mean Model Score
10	222	222	0.29	0.29	23.9%	240	240	0.27
20	223	445	0.22	0.25	41.5%	176	208	0.20
30	222	667	0.14	0.22	53.3%	118	178	0.17
43	299	966	0.13	0.19	68.0%	109	157	0.13
50	147	1113	0.11	0.18	73.9%	89	148	0.12
64	304	1417	0.06	0.16	80.9%	51	127	0.09
70	141	1558	0.11	0.15	86.4%	87	123	0.08
80	222	1780	0.06	0.14	91.2%	48	114	0.07
90	223	2003	0.06	0.13	96.0%	48	107	0.05
100	223	2226	0.05	0.12	100.0%	40	100	0.04

Table 3: Gains Table for MARS Model

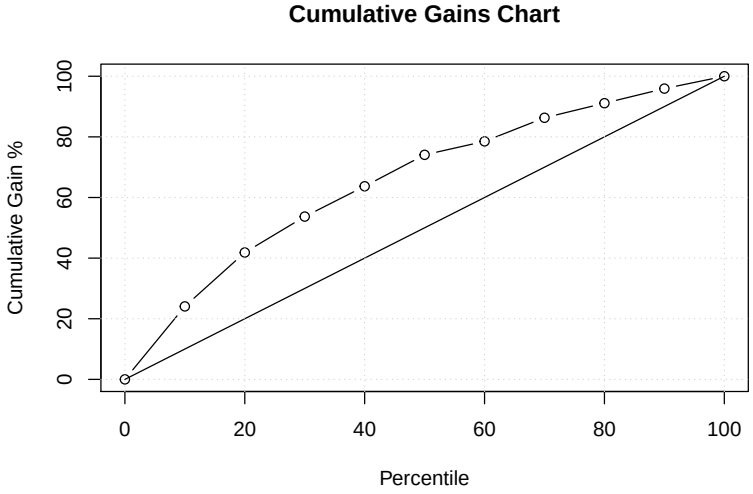

Figure 8: Gains Chart - MARS

5.2 Logistic Regression

A logistic regression model is fit on the training data to model the response variable Bad(1/0). All the model coefficients are significant and indicate the same outcomes as the MARS model. i.e. Probability of Bad decreases with increase in months on books, increases with increase in DueProportion and increases with as Balance Proportion increases. The coefficients are reported in Table 4.

Coefficient	Value
Intercept	-2.8749
monthscut<=27	-0.4239
monthscut>27	-0.7067
duepropcut<=0.0815	0.5435
duepropcut<=0.0998	0.7413
duepropcut>0.0998	0.9808
BalProp	1.0622

Table 4: Coefficients of Logistic Regression Model

5.2.1 ROC Curve for logistic regression

The ROC Curve of the logistic regression is shown in Figure 8. Looking at the ROC Curve, the threshold value for classificcation between 0.1 and 0.2 is appropriate to maintain the balance between True Postive Rate and False Positive Rate.

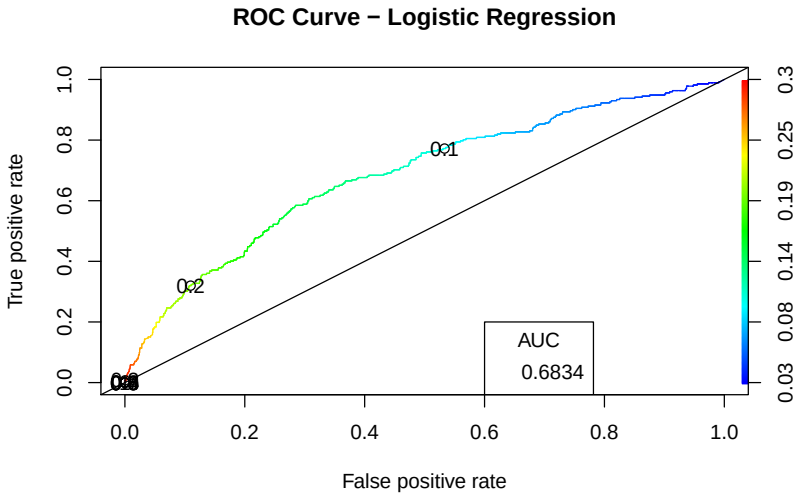

Figure 9: ROC Curve for Logistic Regression model

5.2.2 Confusion Matrix for Logistic Regression

The clasification threshold was set to 0.15. The Accuracy obtained by the logistic regression model is 71.43%. The specificity is 54.41 % and sensitivity is 73.80 %. The positive class considered here is 0 i.e Good.

	Actual	
Prediction	0	1
0	1440	128
1	514	144

Table 5: Confusion Matrix of Logistic Regression Model

5.2.3 KS Curve for Logistic Regression

K-S is a measure of the degree of separation between the positive and negative distributions. The KS stat value is the maximum distance between 0.27 between two the CDF and for logistic regression it is 0.29.

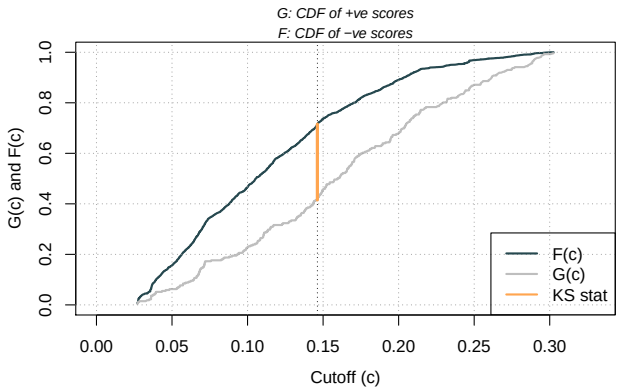

Figure 10: ROC Curve for Logistic Regression model

5.2.4 Lift Chart for Logistic Regression

The lift chart shows how much more likely we are to select good customers than if we select a random sample of customers. The lift chart shows that considering only the first 10% of the customers will allow in selecting 2.5 times more good customers using the predictive model than done randomly.

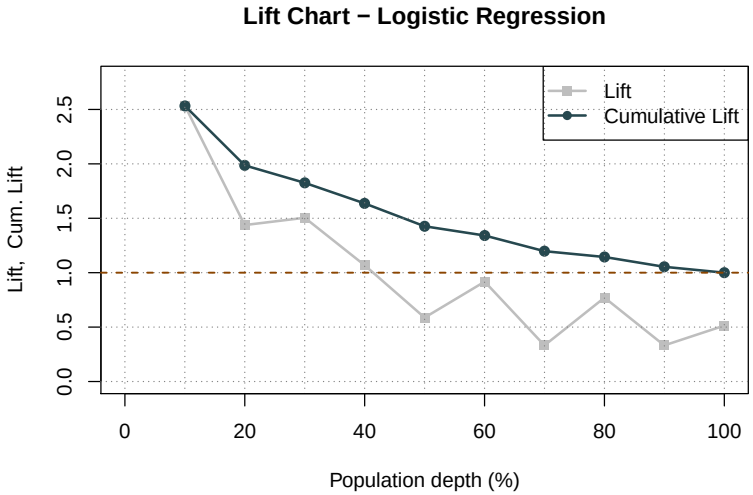

Figure 11: Lift Chart for Logistic Regression model

5.2.5 Gains Table for Logistic Regression

Depth of File	N	Cume N	Mean Resp	Cume Mean Resp	Cume Pct of Total Resp	Lift Index	Cume Lift	Mean Model Score
10	222	222	0.31	0.31	25.4%	254	254	0.25
20	223	445	0.17	0.24	39.7%	143	199	0.20
30	222	667	0.18	0.22	54.8%	151	183	0.16
40	223	890	0.13	0.20	65.4%	106	164	0.14
50	223	1113	0.07	0.17	71.3%	59	143	0.12
60	222	1335	0.11	0.16	80.5%	92	134	0.10
70	223	1558	0.04	0.15	83.8%	33	120	0.08
80	222	1780	0.09	0.14	91.2%	74	114	0.07
90	223	2003	0.04	0.13	94.9%	37	105	0.05
100	223	2226	0.06	0.12	100.0%	51	100	0.03

Table 6: Gains Table of Logistic Regression Model

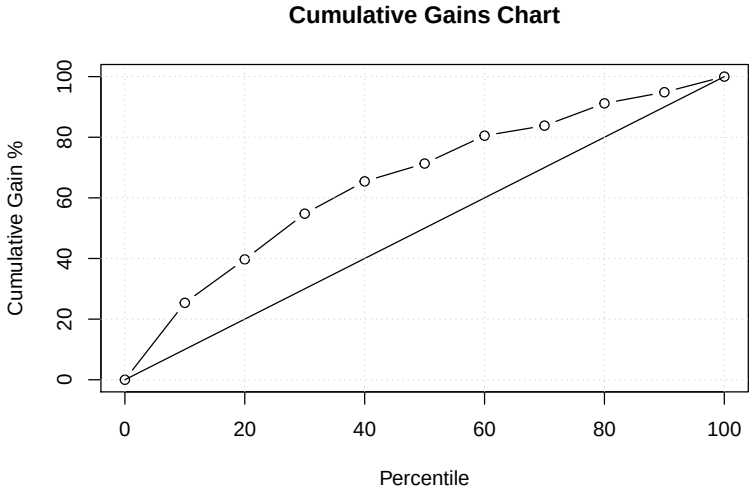

Figure 11: Lift Chart for Logistic Regression model

6. Gains Table for Good Customer Score

Good Customer Score is considered just another model whose probability score is converted into points. To compare this model with our's, we use a gains table. Comparing the Gains Charts of logistic regression in Figure 11 and Gains Chart of Good Customer Score in Figure 12, we can say the percentage of Good customers covered at the first few deciles is higher for the logistic regression model than the Good Customer Score Model. Therefore we can say our model performs better than the Good Customer Score Model.

Depth of File	N	Cume N	Mean Resp	Cume Mean Resp	Cume Pct of Total Resp	Lift Index	Cume Lift	Mean Model Score
10	223	223	0.07	0.07	5.5%	55	55	980.50
20	222	445	0.04	0.05	8.8%	33	44	908.10
30	224	669	0.06	0.06	13.6%	47	45	838.71

```
##   40   221   890   0.07   0.06   19.5%   59   49   792.02
##   50   229  1119   0.06   0.06   24.6%   50   49   760.07
##   60   223  1342   0.14   0.07   36.0%  114   60   738.95
##   70   225  1567   0.14   0.08   47.4%  113   67   720.00
##   80   215  1782   0.20   0.10   62.9%  160   79   698.02
##   90   221  2003   0.17   0.10   76.8%  141   85   655.09
##  100   223  2226   0.28   0.12  100.0%  231  100   473.12
```

Table 7: Gains Table of Good Customer Score

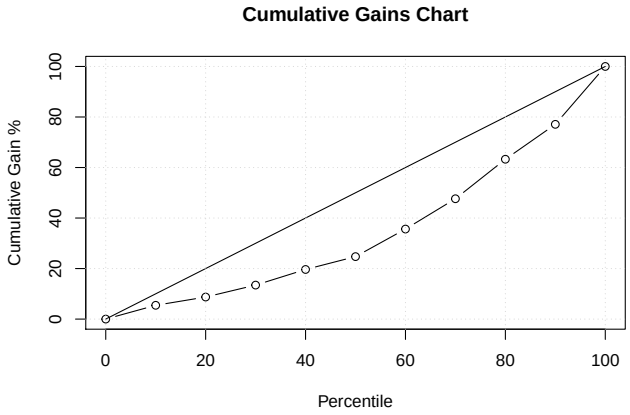

Figure 12: Gains Chart of Good Customer Score

7. Profitability in Gains Table

Determining Profit for particular customer is never a easy task. But we adopt the following definition to calculate profit.

- Profit (Bad=0): Sum up Net Payments for all rows of an account.
- Profit (Bad=1): Take the ending balance and make it a loss. Make it negative.

Based on the above definition we calculate the profitability in gains table which is shown in the following table. The Average Profit, Profit and Cumulative Profit at each decile of the customer population is shown. The final cumulative profit gained is 1119256.65.

Depth of File	N	Cume N	Average Profit	Profit	Cumulative Profit
10	222	222	261.0730	58219.27	58219.27
20	223	445	417.5650	93117.00	151336.27
30	222	667	389.2591	86415.52	237751.79
40	223	890	448.1318	99933.39	337685.18
50	223	1113	567.2192	125922.67	463607.85
60	222	1335	510.0153	113733.42	577341.27
70	223	1558	608.5361	135703.55	713044.82
80	222	1780	522.0956	115905.23	828950.05
90	223	2003	725.7239	161836.43	990786.48
100	223	2226	578.6945	128470.17	1119256.65

Table 8: Profitability in Gains Table