Assignment 2: Logistic Regression Using SAS Enterprise Miner

Daanish Ahmed

DATA 640 9041

Fall 2017

dsahmed2334@yahoo.com

Professor Sounak Chakraborty

UMUC

October 16, 2017

Introduction

Logistic regression is a valuable predictive model for analyzing a categorical target. It involves using a logarithmic function with values between 0 and 1 to estimate the probability of the dependent variable having an expected class (Knode, 2016a). These models function the best on binary targets, and they often perform better when using variable selection methods such as forward selection, backward elimination, or stepwise selection. In this assignment, I will build three logistic regression models using SAS Enterprise Miner for a dataset on customer churn from a cellular phone company (see Figure 1 in Appendix). My goal is to predict whether a given customer will churn so that the company can identify potential churners and prevent them from leaving. I will create one model for forward selection, one for backward elimination, and one for stepwise selection. I will use appropriate variable transformations for each model to minimize their misclassification rates. And since we are most interested in cases where customers are likely to churn, I will use appropriate cutoff thresholds to balance the number of identified churners with the overall accuracy. I will then compare these models to determine which one will maximize customer retention and minimize financial losses for the company.

The dataset contains 15 columns and 2085 observations. The target variable is churn, which is a binary non-numeric (character) variable. Churn has two possible classes: no (meaning the customer did not churn) and yes (meaning the customer churned). Some of the inputs include account length, customer service calls, day calls, day minutes, evening calls, evening minutes, international plan, international calls, and international minutes (see Figure 2 in Appendix). The customer ID variable (labeled as “VAR1”) was set to rejected since it is not useful for the analysis. This leaves us with 14 variables remaining. Of the remaining variables, three are non-numeric: churn, international plan, and voice mail plan. All three of these non-numeric variables are binary.

The remaining inputs are all numeric interval variables. According to the descriptive statistics (see Figure 3 in Appendix), 0.047962% of values are missing for every single variable in this dataset. Since there are 2085 records, this means that there is one missing value in each variable (including the ID). And by examining the data in Figure 2, I found that the last row in the dataset is completely empty. We can therefore remove this row during data preparation. Regarding skewness, we see that the variables with the highest skewness are voice mail messages (1.31), international calls (1.19), and customer service calls (1.15). I will later examine the distributions of each interval input and apply appropriate transformations for the models I will build.

Data Preparation

I followed the SEMMA process during the implementation of my regression models (see Figure 1 in Appendix). I used nodes for sample (Data Partition, Filter), explore (StatExplore), modify (Replacement, Transform Variables), model (Regression), and assess (Cutoff, Model Comparison). My dataset was originally in .xlsx format, so I saved it as a .csv file in Excel and then imported it into SAS Enterprise Miner using the File Import node. Finally, I converted the file into a .sas7bdat file using the Save Data node. I began data exploration by using the StatExplore node to produce a bar plot of each input variable's worth (see Figure 4 in Appendix). According to the image, the four most significant variables are day minutes, customer service calls, international plan, and evening minutes. Likewise, the least important variables are account length, night calls, evening calls, voice mail plan, night minutes, international calls, and international minutes. During the model creation stage, I will experiment with removing unimportant variables from some of my models. This is to ensure that each of my models has the lowest possible misclassification rate.

By examining the prior probabilities for the output (churn), I found that this variable is skewed such that only 14.2% of customers will churn. To address this issue, I used decision weights to place a cost on incorrect predictions (see Figure 5 in Appendix). In this figure, decision 1 involves the company intervening to prevent a customer from churning, while decision 2 involves the company doing nothing. I set the cost of company intervention to -4.0, the cost of losing a customer to -10.0, and the profit from keeping a customer to 10.0. If a customer is expected to churn and the company appeals to him, then they will make 6.0. But if they appeal to a non-churner, they will waste 4.0. If they do nothing and a customer churns, they lose 10.0. And if they do nothing and the customer doesn't churn, they will gain a 10.0 profit. These weights will add insight to the costs of failing to make accurate predictions.

After this step, I proceeded to handle missing values, outliers, and incorrect entries. Since one of the rows contains every missing value in the dataset, we can remove that row using the Filter node. I set the property “keep missing values” to “no,” thus removing all observations with missing values. For both class and interval variables, I set the default filtering method to “none” to prevent the node from filtering records with outliers. Instead, I chose to replace outliers by using the Replacement node. By examining the descriptive statistics, I found that variables such as day minutes and evening minutes have values more than 3 standard deviations from the mean (see Figure 3 in Appendix). Thus, I set the interval variable default limits method to “standard deviations from the mean.” By using the default cutoff value of 3.0, this node will compute a replacement for any values more than 3 standard deviations from the mean. Finally, by using the replacement editor, we can verify that there are no incorrect entries in this dataset.

Next, I used the Data Partition node to split the dataset into training, validation, and test sets. I allocated 60% of data to the training set, 30% to the validation set, and 10% to the test set.

Finally, I transformed the variables to ensure they function effectively in each model. I used two types of transformation methods: “best” transform and maximum normal. According to SAS Enterprise Miner Reference Help, the “best” method involves applying several types of transformations to each input and selecting the transformation with the best Chi-squared test results. The Chi-squared test is useful for finding the significance of a relationship between two variables (Ray, 2016), which in this case refers to the relationship between the output (churn) and the inputs. The maximum normal method, as described by SAS Enterprise Miner Reference Help, also performs multiple transformations on each variable, but instead it uses the transformation that maximizes the normality of each variable’s distribution. This is useful for reducing the skewness of some inputs such as customer service calls and international calls—both of which have positive skews (see Figure 6 in Appendix). I experimented with matching different models with each of these transformation types and evaluated their misclassification rates. Through this process, I found that the "best" method works better with my forward selection model, while “max normal” is best for the backward and stepwise models. To eliminate non-numeric inputs, as well as handle variables like “voice mail messages” with spiked distributions (see Figure 6 in Appendix), I set the class output for every Transform Variables node to “dummy indicators.”

Model Development

Once data preparation was finished, I began to implement the logistic regression models. I created one model for forward selection, one for backward elimination, and one for stepwise selection. According to Knode (2016a), forward selection begins with an empty model and adds inputs one at a time, while backward elimination starts with all variables and removes them from the model. Stepwise selection begins similarly as forward selection, but allows for inputs to be added or removed (Knode, 2016a). I built these models using the Regression node by setting the regression type to "logistic regression" and choosing the selection models "forward," "backward," and "stepwise" for the appropriate models. For all three models, I set the selection criterion to "validation misclassification" to minimize the number of misclassified observations. Before running these nodes, I experimented with manually filtering out the least significant inputs from each model. During data exploration, I found that the five least important variables are account length, night calls, evening calls, voice mail plan, and night minutes (see Figure 4 in Appendix). These variables were excluded by editing the variables under the Regression node and setting their usage to "no." Through this process, I found that the forward regression model was the only model that saw improvement in its validation misclassification and average squared error. Therefore, this is the only model for which I will manually exclude insignificant inputs.

After building these models, I examined the graphs of the regression coefficients, which can be used to build the regression equation. The coefficients are also useful for determining how strongly an input is related to the output and whether the relationship is positive or negative (Frost, 2013). By examining the graph for my forward selection model (see Figure 7 in Appendix), we see that the most significant inputs are day minutes (three bins with values -4.69, -3.29, and -1.44), customer service calls (-2.23), and international plan (1.12). For the backward elimination model (see Figure 12 in Appendix), the important variables are customer service calls (2.45), international calls (-2.29), and international plan (0.97). Finally, the stepwise selection results (see Figure 16 in Appendix) indicate that the top inputs are customer service calls (2.37), international calls (-2.16), and international plan (0.95). The backward and stepwise models rely on many of the same variables with similar coefficients, yet the forward regression model does not. This is likely due to the backward and stepwise models using the same transformation method of "max normal," while the forward model uses the "best" transformation method. Using different transformations can cause the variables to be completely different, and thus the selection methods would have chosen different variables for each of these transformations.

Finally, I will change the cutoff thresholds to maximize each model's predictive performance. Cutoff thresholds are used to decide whether a prediction qualifies as a "1" or a "0," and they can be adjusted to maximize either the overall accuracy or the sensitivity—the number of true positives (Knode, 2016b). Since this dataset's output is heavily skewed, having a higher cutoff threshold will increase the accuracy but lower the sensitivity. Since we are interested in identifying possible churners, we want our models to have high sensitivity while still having acceptable accuracy. By looking at the classification rates for the forward selection model (see Figure 11 in Appendix), we see that 0.1 is a good cutoff threshold because it satisfies the mentioned conditions. I will use 0.1 as the cutoff value for all three models to allow for easier comparison between the models. Finally, I will connect an additional Cutoff node to the stepwise model using a threshold of 0.25 to study the impact of maximizing accuracy over sensitivity.

Results

I will now describe the results of my analysis to see which model is the best for predicting churn. First, I will examine the iteration plots for my three models. These plots show the misclassification rate for the training and validation sets during each step of the model building process. The plot for my forward regression model reveals that the model required five steps to train, and that the misclassification rate is higher for the validation set (see Figure 9 in Appendix). By moving the cursor over the graph in SAS Enterprise Miner, I found that the training and validation misclassification rates are 9.76% and 11.86% respectively (see Figure 21 in Appendix). This difference is not large enough to suggest that overfitting took place, especially since the test set's misclassification of 10.95% is lower than that of the validation set. The backward regression model also required five steps to train, but this time the training misclassification rate is higher than the validation rate (see Figure 14 in Appendix). Figure 21 shows that the misclassification for training, validation, and test data are 12.96%, 12.02%, and 14.76%. The stepwise model also has a higher misclassification for training data, but it is also the most complex model since it required six steps to train (see Figure 18 in Appendix). According to Figure 21, the misclassification for training, validation, and test data are 13.36%, 11.86%, and 13.33%. Since the test set is by far the smallest of the three samples with only 10% of the data, I will focus mostly on the validation results. Based on this, the forward and stepwise models appear to be the most accurate models since they have the lowest validation misclassification rates. But we must first examine the classification tables to see which model is better for predicting churn itself.

Classification tables provide us with the number of predicted "yes" and "no" instances compared to the actual number of "yes" and "no" cases for both the training and validation sets. This helps us to find the number of true positives, as well as the number of false positives and negatives. By examining the classification table for the forward selection model (see Figure 8 in Appendix), we see that the model correctly identified 37 out of 88 customers who will churn. This means that only 42% of churning customers were identified. The 51 churners who were not identified can cause notable losses for the company. We also see that 23 non-churners were identified as churners, which may lead to unnecessary costs since these customers do not need to be appealed to. By comparing the validation results to the training results, it is evident that the model is less accurate on validation data. But as mentioned earlier, the difference in accuracy does not appear to be high enough to suggest overfitting took place. Overall, we see that the model has an accuracy of 87.98%—but this is mainly due to correct identification of non-churners.

By looking at the classification table for the backward regression model (see Figure 13 in Appendix), we find that the model identified 24 out of 88 true positive cases in the validation set. It has a sensitivity of 27%, meaning that it is much weaker at identifying true positives compared to my first model. This model identified 11 non-churners as churners, which is lower than the 23 false positives of my previous model. However, this model failed to recognize a higher number of churners than the previous model (64 compared to 51). Failure to identify churners is much costlier than failure to identify non-churners (see Figure 5 in Appendix), and thus this model will lead to a higher cost for the company. The classification table for the stepwise model (see Figure 17 in Appendix) has similar results to that of the backward model. This model identified 22 true positives, has a sensitivity of 25%, and is altogether weaker than the forward model. However, the backward and stepwise models performed better on the validation data than on the training data, unlike the forward model. These two models also have accuracies similar to that of the first model, but this is once again due to the high number of true negatives identified.

To increase the number of true positives identified, I evaluated each model using a cutoff threshold of 0.1. By examining the cutoff statistics for the forward selection model (see Figure 10 in Appendix), we see that the model's sensitivity has been significantly improved. The model has successfully identified 163 churners in the training set, 74 in the validation set, and 27 in the test set. This leads to a true positive rate of 92.1% for training data, 84.1% for validation data, and 90% for test data. The only downside to this approach is that the overall classification accuracy decreases. However, the forward model's accuracy rates for the training, validation, and test sets are 83.0%, 80.8%, and 83.3%. These rates are only slightly lower than the model's original accuracy rates. When viewing the cutoff statistics for the backward model, we find the true positive rate to be 81.9% for training data, 78.4% for validation data, and 83.3% for test data (see Figure 15 in Appendix). Though the model now has a higher sensitivity, it is still weaker than the forward regression model. Furthermore, the overall classification rate for validation data is 63.8%, which is significantly worse than the model's original accuracy rate.

The results of the stepwise model are similar to those of the backward model. Its true positive rates for training, validation, and test data are 80.2%, 77.3%, and 86.7% (see Figure 19 in Appendix). Likewise, its overall accuracy is 62.5% on the validation data. By comparing my three models, it is apparent that the forward regression model is by far the strongest model with regards to both accuracy and for identifying churners. Finally, I evaluated the stepwise model using a cutoff threshold of 0.25 to determine if it is better to have a higher overall accuracy. This model now has a classification rate of 82.1% for the validation set (see Figure 20 in Appendix), which is significantly higher than its accuracy when using a cutoff of 0.1. However, its true positive rates decrease to 49.7%, 50%, and 53.3% for the training, validation, and test sets. Therefore, I would recommend using a cutoff threshold of 0.1 to maximize the number of identified churners and minimize financial losses.

Conclusion

In my analysis, I developed three logistic regression models to estimate the likelihood of a customer churning from the company. I found that the results differed the most when I used a different transformation type, not when I used a different variable selection method. For instance, the backward and stepwise methods were both impacted most by the variables "customer service calls," "international calls," and "international plan." In both cases, the variables had similar coefficients and only "international calls" had a negative relationship with churn. However, the forward selection model relied most heavily on "day minutes," "customer service calls," and "international plan"—where only "international plan" had a positive relationship with the output. When analyzing the accuracy and sensitivity, I found that the backward and stepwise models had similar results to each other but differed drastically from those of the forward model. Overall, I found that the forward regression model had the highest number of true positives, and its accuracy was the highest of my three models when setting the cutoff threshold to 0.1. When considering the costs from false predictions, the forward model has the lowest number of false negatives—thus producing the lowest expenses for the company. Therefore, this model is easily the most effective model for maximizing customer retention and minimizing costs.

One of the shortcomings of my analysis is the question of whether my forward regression model is comparable with my other two models. This model not only used a completely different type of transformation from the other models, but it was also the only model in which I manually excluded unimportant input variables prior to model development. Due to these differences, the question remains whether forward selection is truly better than backward or stepwise selection. One suggestion for future improvement would be to implement backward and stepwise models using the "best" transformation type while following similar steps to those I used to build the forward regression model. If one of these new models is more effective at predicting customer churn, then it suggests that using the "best" transformation instead of "max normal" was the primary reason why my forward regression model was so effective. I would recommend further analysis because even though my forward selection model is the most accurate of my models, there is still room for improvement even when using a cutoff threshold of 0.1.

References

Frost, J. (2013, December 12). Regression Analysis Tutorial and Examples. Retrieved October 4, 2017, from http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-tutorial-and-examples

Knode, S. (2016a, August 19). *Regression Models*. Lecture presented at UMUC. Retrieved October 3, 2017.

Knode, S. (2016b, October 11). *Adjusting for Skewed Target Distribution*. Lecture presented at UMUC. Retrieved October 14, 2017.

Ray, S. (2016, January 10). A Comprehensive Guide to Data Exploration. Retrieved September 26, 2017, from https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/

Welcome to Logistic Regression Analysis. (n.d.). Retrieved October 6, 2017, from http://logisticregressionanalysis.com/

Appendix

Relevant SAS Enterprise Miner Output Images

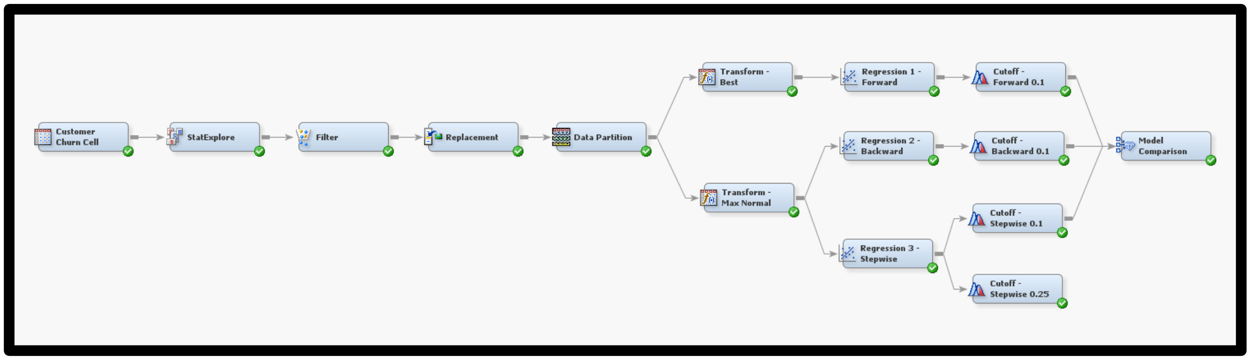

Figure 1. Process Flow Diagram for Customer Churn Logistic Regression Models.

Obs #	Churn	ï»¿Customer	Account Length	International Plan	Voice Mail Plan	Voice Mail Messages	Day Minutes	Day Calls	Evening Minutes	Evening Calls	Night Minutes	Night Calls	International Minutes	International Calls	Customer Service Calls
2028	Yes	2038	98	No	No	0	0	0	159.6	130	167.1	88	6.8	1	4
2029	No	2039	54	No	Yes	39	143.9	73	210.3	117	129.2	117	12.5	8	2
2030	No	2040	135	No	No	0	155.2	100	135.9	84	184.6	82	3.8	9	3
2031	No	2041	56	No	Yes	24	121.7	87	184	76	266.6	98	12.7	3	1
2032	No	2042	177	No	No	0	189.5	99	176.3	117	225.9	112	14.2	2	1
2033	Yes	2043	141	No	Yes	32	322.4	92	283.2	107	209.5	111	6.7	3	1
2034	No	2044	153	No	No	0	195.4	107	154.6	96	142.8	97	11.6	6	1
2035	No	2045	120	No	No	0	203.3	108	259.9	66	115.9	103	7.8	2	3
2036	No	2046	65	No	No	0	136.1	112	272.9	96	220.2	104	4.4	2	1
2037	No	2047	118	No	No	0	136.1	120	204.2	103	228.2	90	11.3	4	1
2038	Yes	2048	74	No	No	0	174.1	96	251.1	94	257.6	123	8.3	5	2
2039	No	2049	72	No	No	0	175.5	103	132.3	120	242.9	96	11.8	3	1
2040	No	2050	41	No	No	0	207.2	138	214.1	83	193	105	11.9	4	1
2041	No	2051	217	No	No	0	176.4	115	158.8	128	306.6	107	9.3	3	4
2042	No	2052	130	No	No	0	162.8	113	290.3	111	114.9	140	7.2	3	1
2043	No	2053	22	No	No	0	110.3	107	166.5	93	202.3	96	9.5	5	0
2044	No	2054	105	No	No	0	193.7	108	183.2	124	293.7	72	10.8	5	1
2045	No	2055	48	No	No	0	198.2	73	202.8	115	146.4	73	5.1	5	1
2046	No	2056	45	No	No	0	211.3	87	165.7	97	265.9	72	13.3	6	1
2047	No	2057	135	Yes	Yes	24	127.7	54	215	105	234.3	84	5.8	4	2
2048	No	2058	150	No	No	0	126	99	238.5	73	285.1	100	10.2	6	3
2049	No	2059	84	No	Yes	33	159.1	106	149.8	101	213.4	108	13	18	1
2050	No	2060	197	No	No	0	233.9	96	218.9	111	182.9	109	9.5	3	0
2051	No	2061	40	No	No	0	109.4	107	244.7	102	276.9	123	7.1	7	0
2052	No	2062	30	No	No	0	247.4	107	175.9	76	287.4	90	11.3	2	0
2053	Yes	2063	122	Yes	No	0	140	101	196.4	77	120.1	133	9.7	4	4
2054	No	2064	108	No	No	0	73.8	105	143.4	114	170.2	98	10.9	3	2
2055	Yes	2065	71	No	No	0	238	82	278.5	94	193.1	134	11.8	10	0
2056	No	2066	128	No	No	0	158.8	75	264.8	91	270	77	7.6	7	1
2057	No	2067	84	No	Yes	42	214.3	112	188.2	107	333.5	117	11.3	10	0
2058	No	2068	74	No	No	0	85.7	83	247.7	67	142.4	85	10.1	5	2
2059	No	2069	99	No	No	0	62.9	81	231	64	168.9	121	8.5	5	1
2060	No	2070	56	No	No	0	197	110	222.8	102	225.3	91	10.6	6	2
2061	No	2071	104	No	No	0	264	108	132.2	75	177.7	91	10.6	8	3
2062	No	2072	146	No	No	0	206.3	151	148.6	89	167.2	91	6.1	3	1
2063	No	2073	28	No	No	0	187.8	94	248.6	86	208.8	124	10.6	5	0
2064	No	2074	87	No	Yes	22	263.8	65	103.4	115	208.1	109	8.5	3	3
2065	No	2075	97	No	Yes	25	141	101	212	85	175.2	138	4.9	2	3
2066	No	2076	57	No	No	0	192.8	68	158	86	235.5	105	12.7	6	1
2067	No	2077	45	No	No	0	207.6	71	152.7	94	217.8	125	12.4	13	1
2068	No	2078	53	No	No	0	184.8	98	216.4	125	141.1	116	18.4	3	2
2069	No	2079	117	No	No	0	184.8	83	248.6	101	133.1	113	9.6	8	1
2070	No	2080	103	No	Yes	31	107.7	124	188.9	104	196.2	98	8.9	3	0
2071	No	2081	155	No	No	0	163	93	203.9	102	159	109	15.1	4	2
2072	No	2082	139	No	No	0	139	110	132.9	93	272	120	12.1	1	0
2073	No	2083	139	No	Yes	20	214.6	101	235.1	132	162.8	132	14.8	12	0
2074	No	2084	75	No	Yes	28	200.6	96	164.1	111	169.6	153	2.5	5	1
2075	No	2085	119	No	No	0	176.8	90	224.7	81	204.6	77	7.5	15	1
2076	No	2086	195	No	No	0	63.2	108	220.2	88	184	99	5.1	3	0
2077	No	2087	138	No	Yes	29	190.1	87	223.2	123	256.2	130	14.2	6	0
2078	No	2088	51	No	No	0	180.5	88	134.7	102	170.7	97	10	3	2
2079	Yes	2089	107	No	No	0	294.9	71	192.8	78	148.1	87	13.2	5	1
2080	No	2090	19	No	No	0	186.1	98	254.3	57	214	127	14.6	7	2
2081	No	2091	141	No	Yes	36	187.5	99	241.4	116	229.5	105	10.5	5	3
2082	No	2092	84	No	Yes	35	207.5	138	201	116	164.5	107	7.5	16	4
2083	No	2093	99	No	Yes	28	200.7	88	264.2	116	172.7	102	9.1	5	1
2084	Yes	2094	100	No	No	0	278	76	176.7	74	219.5	126	8.3	4	0
2085		.	.			.	.	.	.	.	.	.	.	.	.

Figure 2. Customer Churn at a Cellular Phone Company Dataset.

Name	Role	Level	Type	Number of Levels	Percent Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
Account_Length	Input	Interval	Numeric	.	0.047962	1	232	100.8714	40.90292	0.068997	-0.13497
Churn	Target	Binary	Character	2	0.047962	.	.	.	.	.	.
Customer_Service_Calls	Input	Interval	Numeric	.	0.047962	0	9	1.571017	1.327448	1.152134	2.047751
Day_Calls	Input	Interval	Numeric	.	0.047962	0	163	100.2361	20.38864	-0.17372	0.418226
Day_Minutes	Input	Interval	Numeric	.	0.047962	0	346.8	179.9886	54.56822	-0.03005	0.095213
Evening_Calls	Input	Interval	Numeric	.	0.047962	0	168	99.8095	20.03431	-0.11128	0.330131
Evening_Minutes	Input	Interval	Numeric	.	0.047962	0	351.6	201.8364	50.12722	-0.05516	-0.02615
International_Calls	Input	Interval	Numeric	.	0.047962	0	18	4.524952	2.480428	1.197296	2.220202
International_Minutes	Input	Interval	Numeric	.	0.047962	0	20	10.21171	2.797119	-0.25924	0.659338
International_Plan	Input	Binary	Character	2	0.047962	.	.	.	.	.	.
Night_Calls	Input	Interval	Numeric	.	0.047962	33	175	100.2634	19.70518	0.052944	0.004829
Night_Minutes	Input	Interval	Numeric	.	0.047962	43.7	395	201.6883	50.08343	0.016151	0.052053
VAR1	Rejected	Interval	Numeric	.	0.047962	11	2094	1052.5	601.7433	0	-1.2
Voice_Mail_Messages	Input	Interval	Numeric	.	0.047962	0	51	7.797025	13.48652	1.3136	0.074732
Voice_Mail_Plan	Input	Binary	Character	2	0.047962	.	.	.	.	.	.

Figure 3. Descriptive Statistics of Customer Churn Dataset.

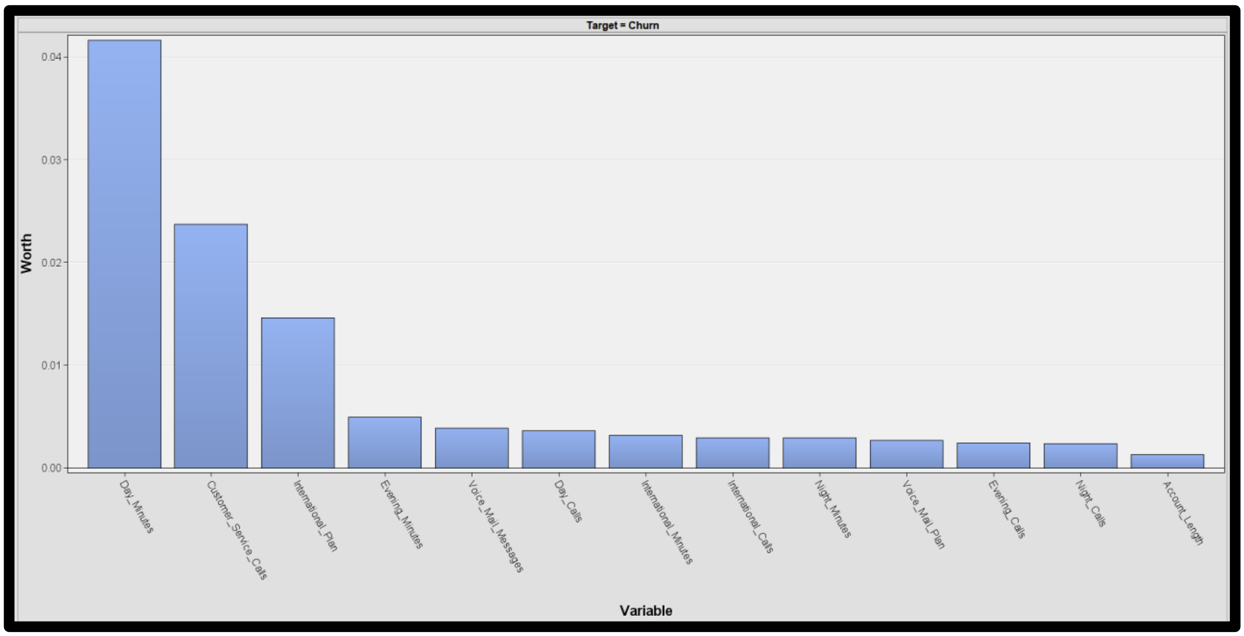

Figure 4. StatExplore Plot of Input Variable Worth.

Enter weight values for the decisions.

Level	DECISION1	DECISION2
YES	6.0	-10.0
NO	-4.0	10.0

Figure 5. Decision Weights for Customer Churn Predictions.

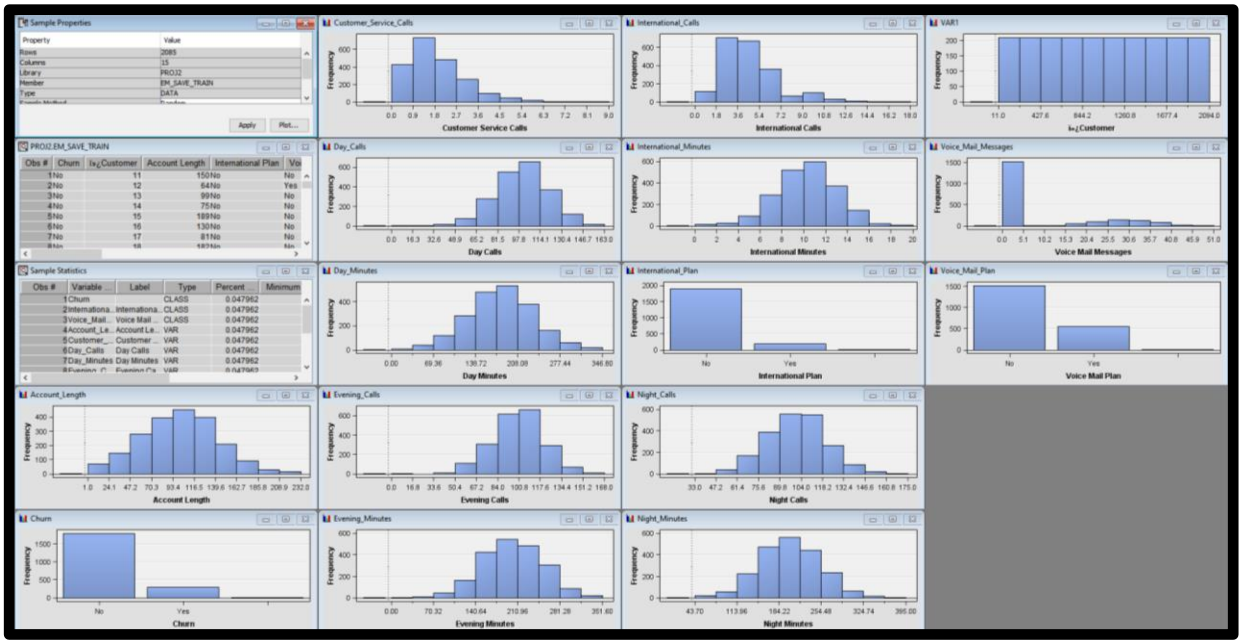

Figure 6. Initial Distributions of Variables in Customer Churn Dataset.

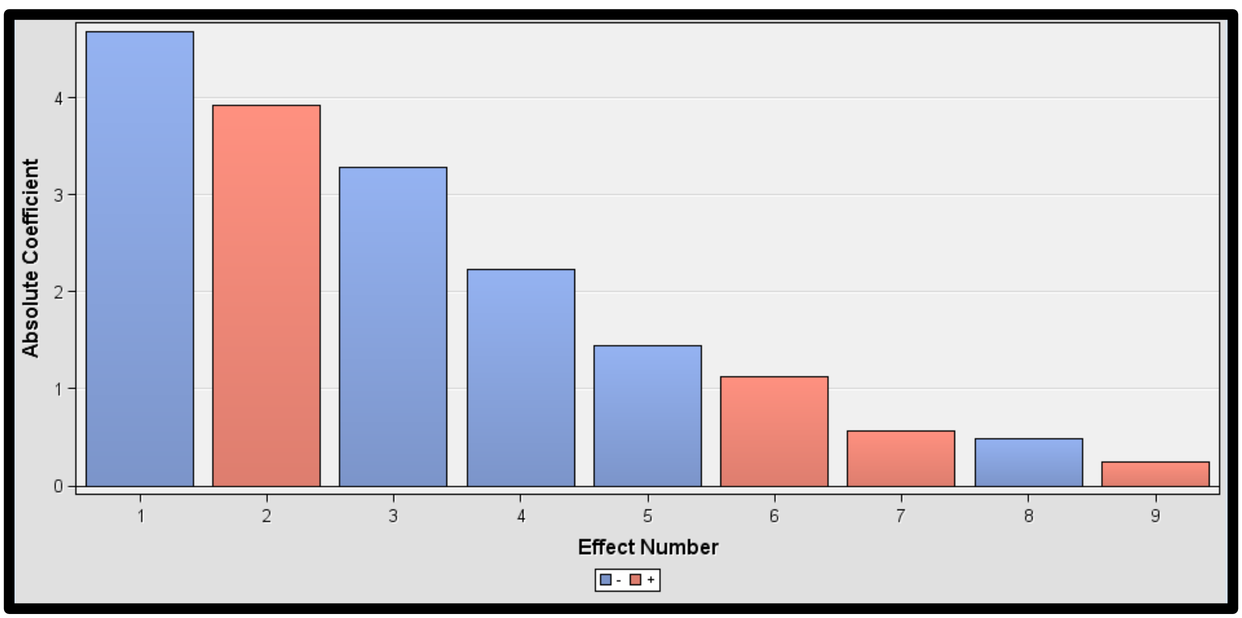

Figure 7. Graph of Regression Coefficients for Forward Selection Model.

Classification Table

Data Role=TRAIN Target Variable=REP_Churn Target Label=Replacement: Churn

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
NO	NO	92.4175	96.5517	1036	82.88
YES	NO	7.5825	48.0226	85	6.80
NO	YES	28.6822	3.4483	37	2.96
YES	YES	71.3178	51.9774	92	7.36

Data Role=VALIDATE Target Variable=REP_Churn Target Label=Replacement: Churn

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
NO	NO	90.9574	95.7090	513	82.2115
YES	NO	9.0426	57.9545	51	8.1731
NO	YES	38.3333	4.2910	23	3.6859
YES	YES	61.6667	42.0455	37	5.9295

Figure 8. Classification Table for Forward Selection Model.

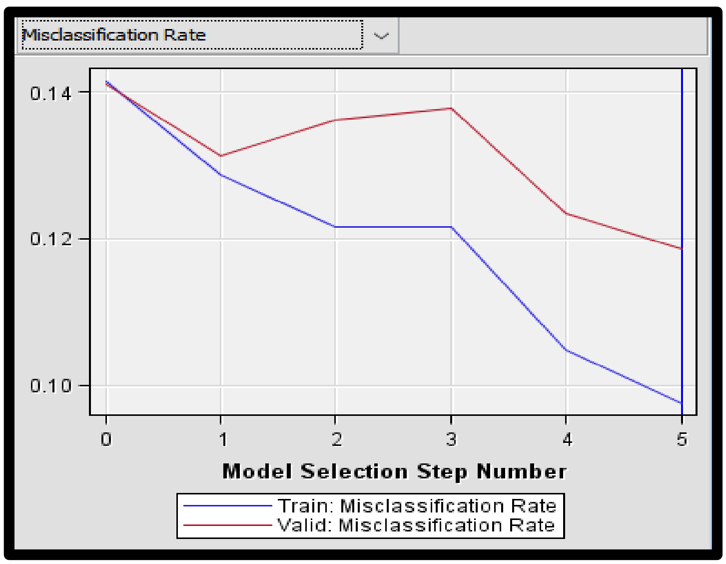

Figure 9. Iteration Plot for Forward Selection Model.

Cutoff	Counts of True Positives	Counts of False Positives	Counts of True Negatives	Counts of False Negatives	Counts of Predicted Positives	Counts of Predicted Negatives	Counts of False Positives and Negatives	Counts of True Positive and Negatives	Overall Classification Rate	Change Count True Positives	Change Count False Positives	True Positive Rate	True Negative Rate	False Positive Rate	Misscl. cost prior 0.1416 equal cost structure	Misscl. cost prior 0.1 equal cost structure	Misscl. cost prior 0.2 equal cost structure	Misscl. cost prior 0.3 equal cost structure	Misscl. cost prior 0.4 equal cost structure	Misscl. cost prior 0.5 equal cost structure	Event Precision Rate	Non Event Precision Rate	Overall Precision Rate	Data Role
0.11	27	32	148	3	59	151	35	175	83.33333	0	0	90	82.22222	17.77778	0.166764	0.17	0.162222	0.154444	0.146667	0.138889	45.76271	98.01325	71.88798	TEST
0.1	163	199	874	14	362	888	213	1037	82.96	0	0	92.0904	81.45387	18.54613	0.1704	0.174825	0.164188	0.153552	0.142915	0.132279	45.02762	98.42342	71.72552	TRAIN
0.1	74	106	430	14	180	444	120	504	80.76923	0	0	84.09091	80.22388	19.77612	0.192285	0.193894	0.190027	0.18616	0.182293	0.178426	41.11111	96.84685	68.97898	VALIDATE
0.1	27	32	148	3	59	151	35	175	83.33333	0	0	90	82.22222	17.77778	0.166764	0.17	0.162222	0.154444	0.146667	0.138889	45.76271	98.01325	71.88798	TEST
0.09	163	199	874	14	362	888	213	1037	82.96	0	0	92.0904	81.45387	18.54613	0.1704	0.174825	0.164188	0.153552	0.142915	0.132279	45.02762	98.42342	71.72552	TRAIN

Figure 10. Cutoff Statistics for Forward Selection Model with 0.1 Cutoff Threshold.

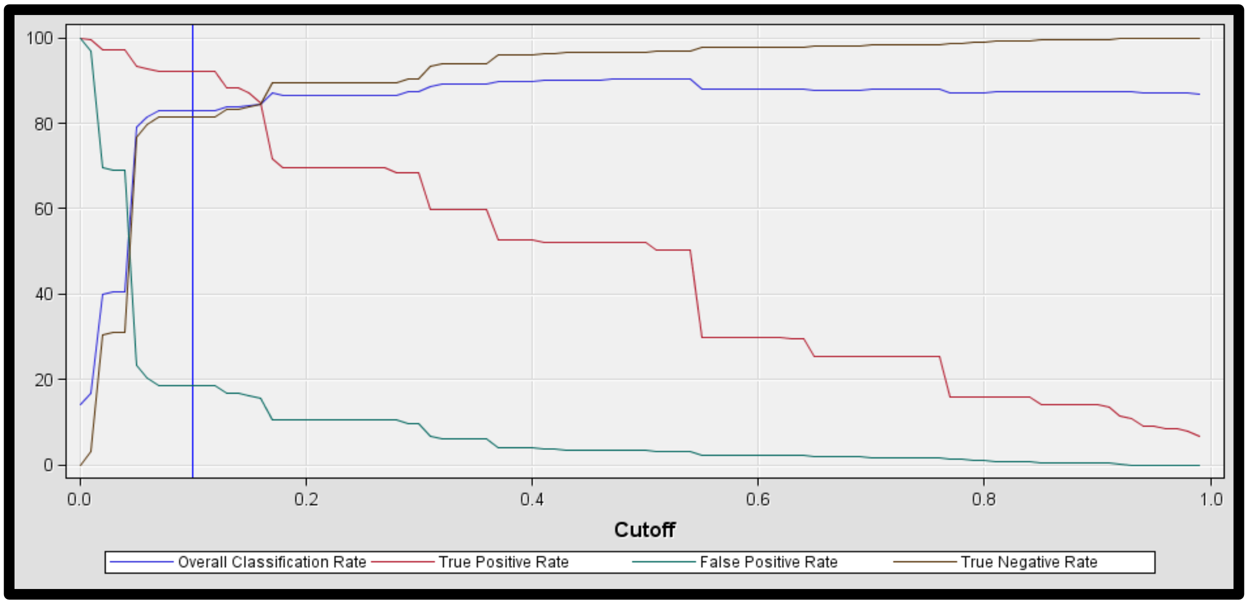

Figure 11. Classification Rates for Forward Selection Model with 0.1 Cutoff Threshold.

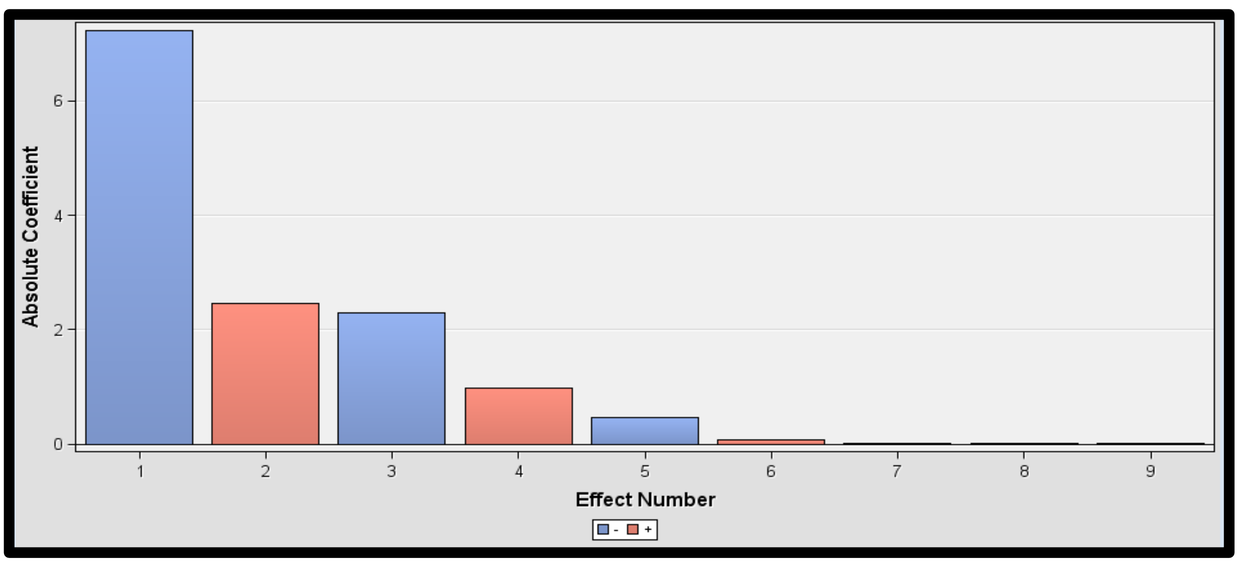

Figure 12. Graph of Regression Coefficients for Backward Elimination Model.

Classification Table

Data Role=TRAIN Target Variable=REP_Churn Target Label=Replacement: Churn

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
NO	NO	88.1811	98.0429	1052	84.16
YES	NO	11.8189	79.6610	141	11.28
NO	YES	36.8421	1.9571	21	1.68
YES	YES	63.1579	20.3390	36	2.88

Data Role=VALIDATE Target Variable=REP_Churn Target Label=Replacement: Churn

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
NO	NO	89.1341	97.9478	525	84.1346
YES	NO	10.8659	72.7273	64	10.2564
NO	YES	31.4286	2.0522	11	1.7628
YES	YES	68.5714	27.2727	24	3.8462

Figure 13. Classification Table for Backward Elimination Model.

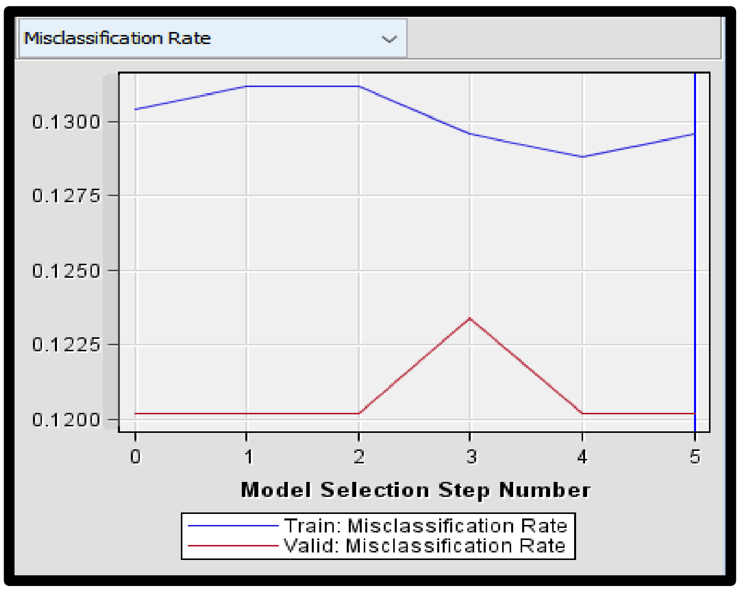

Figure 14. Iteration Plot for Backward Elimination Model.

Cutoff	Counts of True Positives	Counts of False Positives	Counts of True Negatives	Counts of False Negatives	Counts of Predicted Positives	Counts of Predicted Negatives	Counts of False Positives and Negatives	Counts of True Positive and Negatives	Overall Classification Rate	Change Count True Positives	Change Count False Positives
0.11	24	65	115	6	89	121	71	139	66.19048	1	6
0.1	145	412	661	32	557	693	444	806	64.48	7	37
0.1	69	207	329	19	276	348	226	398	63.78205	0	21
0.1	25	70	110	5	95	115	75	135	64.28571	1	5
0.09	148	442	631	29	590	660	471	779	62.32	3	30

Cutoff	True Positive Rate	True Negative Rate	False Positive Rate	Misscl. cost prior 0.1416 equal cost structure	Misscl. cost prior 0.1 equal cost structure	Misscl. cost prior 0.2 equal cost structure	Misscl. cost prior 0.3 equal cost structure	Misscl. cost prior 0.4 equal cost structure	Misscl. cost prior 0.5 equal cost structure	Event Precision Rate	Non Event Precision Rate	Overall Precision Rate	Data Role
0.11	80	63.88889	36.11111	0.338298	0.345	0.328889	0.312778	0.296667	0.280556	26.96629	95.04132	61.00381	TEST
0.1	81.9209	61.60298	38.39702	0.3552	0.363652	0.343334	0.323016	0.302698	0.282381	26.03232	95.3824	60.70736	TRAIN
0.1	78.40909	61.3806	38.6194	0.362082	0.369166	0.352137	0.335109	0.31808	0.301052	25	94.54023	59.77011	VALIDATE
0.1	83.33333	61.11111	38.88889	0.357422	0.366667	0.344444	0.322222	0.3	0.277778	26.31579	95.65217	60.98398	TEST
0.09	83.61582	58.80708	41.19292	0.3768	0.38712	0.362312	0.337503	0.312694	0.287885	25.08475	95.60606	60.3454	TRAIN

Figure 15. Cutoff Statistics for Backward Elimination Model with 0.1 Cutoff Threshold.

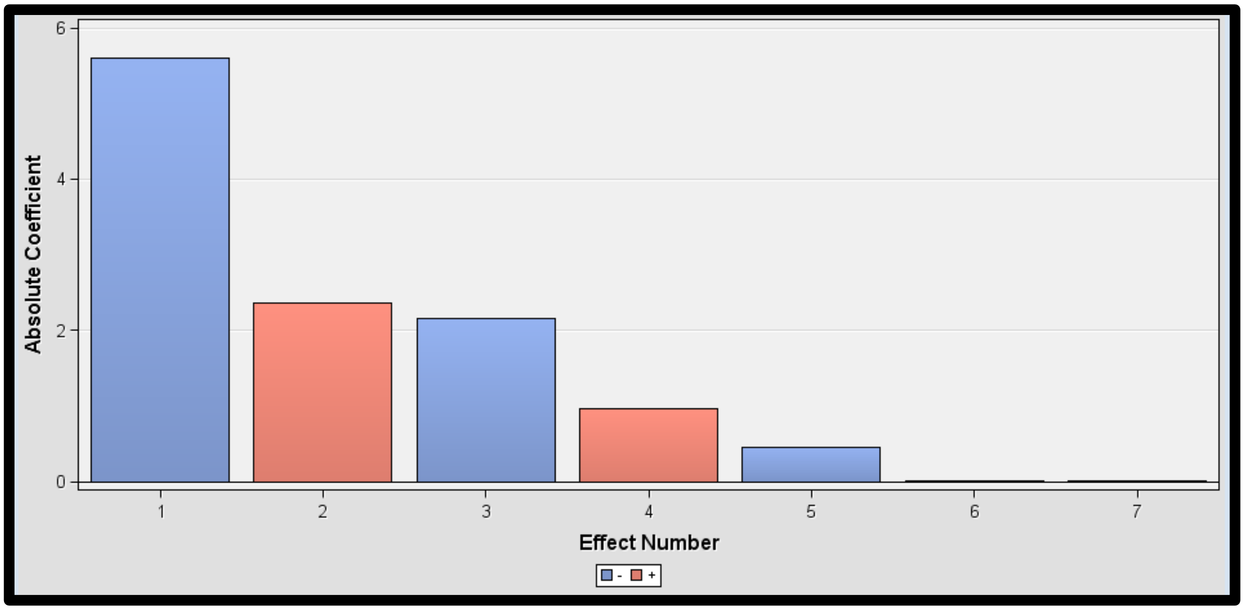

Figure 16. Graph of Regression Coefficients for Stepwise Selection Model.

Classification Table

Data Role=TRAIN Target Variable=REP_Churn Target Label=Replacement: Churn

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
NO	NO	87.8763	97.9497	1051	84.08
YES	NO	12.1237	81.9209	145	11.60
NO	YES	40.7407	2.0503	22	1.76
YES	YES	59.2593	18.0791	32	2.56

Data Role=VALIDATE Target Variable=REP_Churn Target Label=Replacement: Churn

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
NO	NO	88.8889	98.5075	528	84.6154
YES	NO	11.1111	75.0000	66	10.5769
NO	YES	26.6667	1.4925	8	1.2821
YES	YES	73.3333	25.0000	22	3.5256

Figure 17. Classification Table for Stepwise Selection Model.

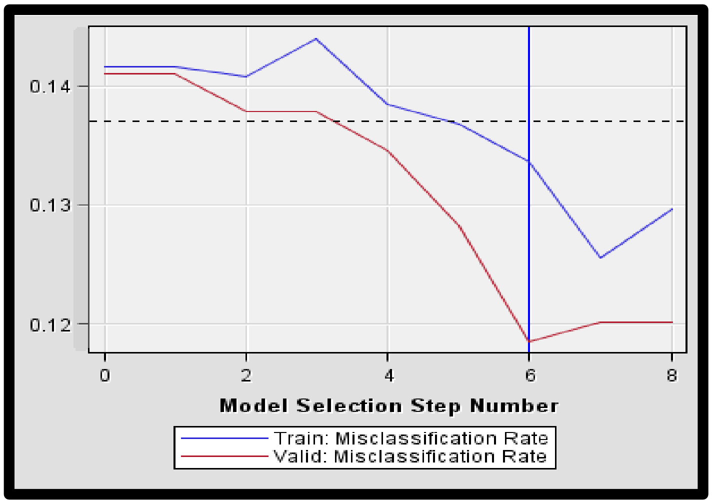

Figure 18. Iteration Plot for Stepwise Selection Model.

Cutoff	Counts of True Positives	Counts of False Positives	Counts of True Negatives	Counts of False Negatives	Counts of Predicted Positives	Counts of Predicted Negatives	Counts of False Positives and Negatives	Counts of True Positive and Negatives	Overall Classification Rate	Change Count True Positives	Change Count False Positives	True Positive Rate	True Negative Rate	False Positive Rate	Missclassification cost prior 0.1416 equal cost structure	Missclassification cost prior 0.1 equal cost structure	Missclassification cost prior 0.2 equal cost structure	Missclassification cost prior 0.3 equal cost structure	Missclassification cost prior 0.4 equal cost structure	Missclassification cost prior 0.5 equal cost structure	Event Precision Rate	Non Event Precision Rate	Overall Precision Rate	Data Role
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	TEST
0.1	142	414	659	35	556	694	449	801	64.08	2	38	80.22599	61.41659	38.58341	0.3592	0.367025	0.348215	0.329406	0.310597	0.291787	25.53957	94.95677	60.24817	TRAIN
0.1	68	214	322	20	282	342	234	390	62.5	1	16	77.27273	60.07463	39.92537	0.374901	0.382056	0.364858	0.347659	0.330461	0.313263	24.11348	94.15205	59.13276	VALIDATE
0.1	26	67	113	4	93	117	71	139	66.19048	1	5	86.66667	62.77778	37.22222	0.338396	0.348333	0.324444	0.300556	0.276667	0.252778	27.95699	96.5812	62.26909	TEST
0.09	149	449	624	28	598	652	477	773	61.84	7	35	84.18079	58.15471	41.84529	0.3816	0.392427	0.366401	0.340375	0.314349	0.288323	24.91639	95.70552	60.31095	TRAIN

Figure 19. Cutoff Statistics for Stepwise Selection Model with 0.1 Cutoff Threshold.

Cutoff	Counts of True Positives	Counts of False Positives	Counts of True Negatives	Counts of False Negatives	Counts of Predicted Positives	Counts of Predicted Negatives	Counts of False Positives and Negatives	Counts of True Positive and Negatives	Overall Classification Rate	Change Count True Positives	Change Count False Positives	True Positive Rate	True Negative Rate	False Positive Rate	Missclassification cost prior 0.1416 equal cost structure	Missclassification cost prior 0.1 equal cost structure	Missclassification cost prior 0.2 equal cost structure	Missclassification cost prior 0.3 equal cost structure	Missclassification cost prior 0.4 equal cost structure	Missclassification cost prior 0.5 equal cost structure	Event Precision Rate	Non Event Precision Rate	Overall Precision Rate	Data Role
0.26	16	26	154	14	42	168	40	170	80.95238	0	3	53.33333	85.55556	14.44444	0.190071	0.176667	0.208889	0.241111	0.273333	0.305556	38.09524	91.66667	64.88095	TEST
0.25	88	120	953	89	208	1042	209	1041	83.28	2	10	49.71751	88.8164	11.1836	0.1672	0.150935	0.190034	0.229133	0.268232	0.30733	42.30769	91.45873	66.88321	TRAIN
0.25	44	68	468	44	112	512	112	512	82.05128	1	9	50	87.31343	12.68657	0.179701	0.164179	0.201493	0.238806	0.276119	0.313433	39.28571	91.40625	65.34598	VALIDATE
0.25	16	28	152	14	44	166	42	168	80	0	2	53.33333	84.44444	15.55556	0.199609	0.186667	0.217778	0.248889	0.28	0.311111	36.36364	91.56627	63.96495	TEST
0.24	91	134	939	86	225	1025	220	1030	82.4	3	14	51.41243	87.51165	12.48835	0.176	0.160983	0.197082	0.233181	0.26928	0.30538	40.44444	91.60976	66.0271	TRAIN

Figure 20. Cutoff Statistics for Stepwise Selection Model with 0.25 Cutoff Threshold.

Selected Model	Predecessor Node	Model Node	Model Description ▲	Target Variable	Target Label	Train: Misclassification Rate	Selection Criterion: Valid: Misclassification Rate	Test: Misclassification Rate
Y	CUT	Reg	Regression 1 - Forward	REP_Churn	Replacement: Churn	0.0976	0.11859	0.109524
	CUT2	Reg2	Regression 2 - Backward	REP_Churn	Replacement: Churn	0.1296	0.120192	0.147619
	CUT3	Reg3	Regression 3 - Stepwise	REP_Churn	Replacement: Churn	0.1336	0.11859	0.133333

Figure 21. Model Comparison Based on Misclassification Rate.