Paper D02-2009

A Comparison of Decision Tree and Logistic Regression Model

Xianzhe Chen, North Dakota State University, Fargo, ND

▪ ABSTRACT

This paper applies a decision tree model and logistic regression models to a real transportation problem, compares results of these two methods and presents model building procedures as well. The data set is partitioned into train, validation and test data. Due to the skewness of some variables, the variable transformation technique has been conducted and a transformed logistic regression is built. The logistic regression models perform better than tree model, while the non-transformed logistic regression model and transformed regression model are indistinguishable. The non-transformed regression model is recommended for this transportation problem.

▪ INTRODUCTION

This paper illustrates how to develop decision trees and logistic regression model for a real transportation problem. There are eight kinds of commodities which need to be shipped from elevators located in North Dakota to six different locations in Minnesota by either rail or truck. The objective of this paper is to use SAS Enterprise Miner to model the shippers' transportation mode choice by decision tree model and logistic regression model. The commodities are shown in Table 1.

Table 1. Commodity Types

Commodity	Name
A	Wheat
C	Durum
E	Barley
G	Sunflower
H	Corn
J	Oats
K	Soybeans
L	Deb (Dry Edible Barley)

The target variable is binary and indicates the transportation mode choice, either rail or truck. In this paper, 0 stands for truck and 1 for rail. This data set contains about 5000 observations and a number of input variables. In order to eliminate irrelevant input variables and obtain a manageable size of input variables, an initial input variable selection has been conducted and the input variables have been reduced to six variables.

▪ PRELIMINARY INVESTIGATION

Before we build the statistical model, we first conduct some preliminary investigation in order to better understand the characteristics of the data. First of all, the metadata sample size is chosen as the default value of 2000 in order to use this information to assign measurement level and model role for each variable, which is shown in Figure 1. Then the target variable CHOICE has been identified as the target variable, and the other six predictor variables have been shown in Figure 2. From Figure 2, it shows that the target variable is a binary variable, i.e. truck or rail, the input variable COMMODITY is a categorical variable with 8 levels, and the other input variables are continuous.

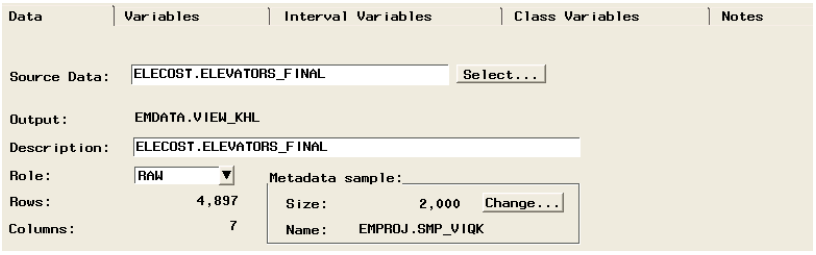
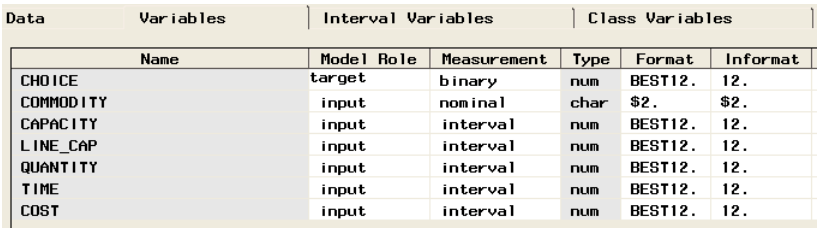


Figure 1. Metadata Sample Size



Name	Model Role	Measurement	Type	Format	Informat
CHOICE	target	binary	num	BEST12.	12.
COMMODITY	input	nominal	char	$2.	$2.
CAPACITY	input	interval	num	BEST12.	12.
LINE_CAP	input	interval	num	BEST12.	12.
QUANTITY	input	interval	num	BEST12.	12.
TIME	input	interval	num	BEST12.	12.
COST	input	interval	num	BEST12.	12.

Figure 2. Identify Target Variable

Some statistical measures of the variables can be found in Figure 3 and Figure 4. From Figure 3 and 4, it is clear that the variables Capacity and Quantity are highly skewed, which indicates that a small percentage of observations may have a great impact. So we could perform a variable transformation in order to yield a better fitting model and compare the results of using transformed model with non-transformed model.

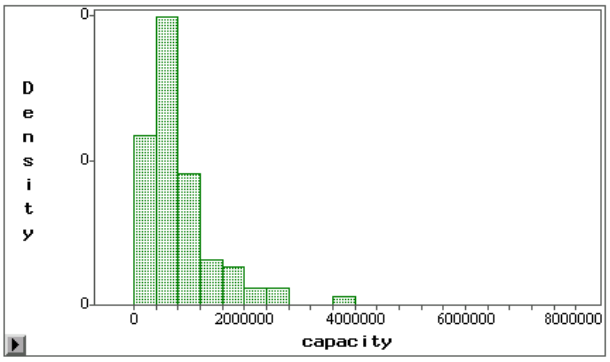


Figure 3. Distribution of Capacity

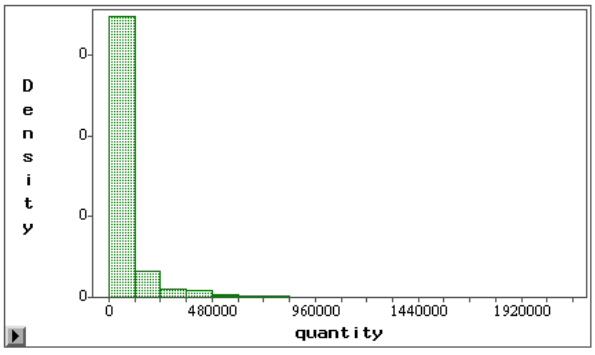


Figure 4. Distribution of Quantity

Data | Variables | Interval Variables | Class Variables | Notes

Name	Min	Max	Mean	Std Dev.	Missing %	Skewness	Kurtosis
CAPACITY	10000	8E6	826309	712164	0%	3.3621	22.457
LINE_CAP	0	99	40.619	32.884	0%	0.5519	-0.824
QUANTITY	25	1.9E6	82219	171959	0%	4.8955	33.433
TIME	2.1318	37.872	12.666	10.337	0%	1.1712	-0.144
COST	15.932	204.58	75.616	50.298	0%	1.083	-0.326

Figure 5. Interval Variables

Data | Variables | Interval Variables | Class Variables | Notes

Name	Values	Missing %	Order	Depends On
CHOICE	2	0%	Ascending	
COMMODITY	8	0%	Ascending	

Figure 6. Categorical Variables

From Figure 5 and 6, it is noted that there is no missing data for our data set. If in case there is missing data, then we need first impute the data set before we build the regression model, however, this step is not necessary for building decision trees.

We also need to specify the percentage of the data to allocate to train, validation, and test data. For our problem, the default value has been used, i.e. 40%, 30% and 30% for train, validation, and test, respectively.

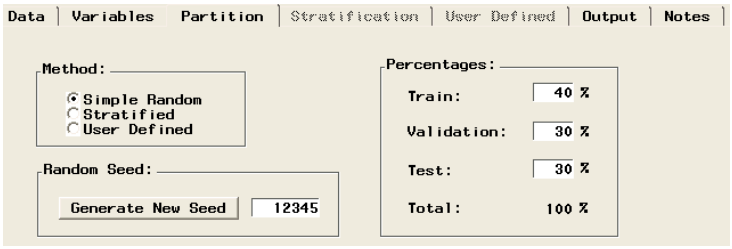


Figure 7. Percentage of Train, Validation and test

Next, we apply Insight node to inspect the characteristic of the variables. From Figure 8, we could see that commodity A and E are the most frequently transported products, which reflects the fact that wheat and barley are the major agricultural products of North Dakota.

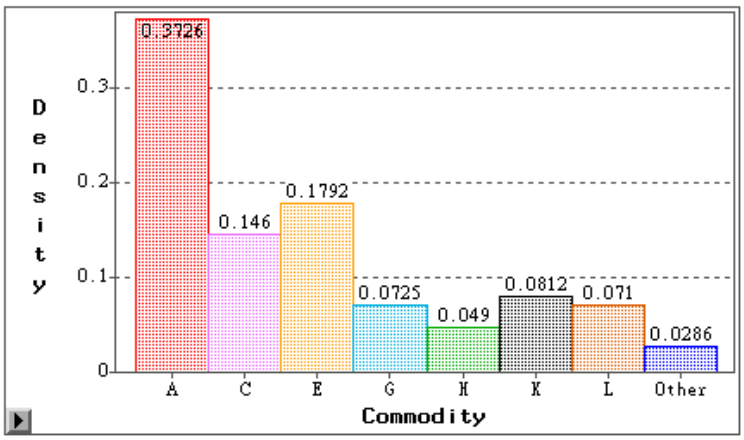


Figure 8. Commodity Density

▪ LOGISTIC REGRESSION MODEL

First of all, a logistic regression model without applying transformation is built, and the link function is chosen as logit (Agresti, 2002) as shown in Figure 9. The logistic model can be built by using SAS procedure Logistic, Genmod, while we build the model in SAS Enterprise Miner since it's quite convenient to compare the results with Decision Tree in Enterprise Miner.

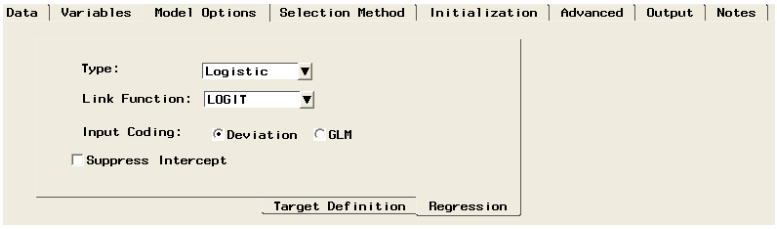


Figure 9. Logistic Regression Model

The variables selection method is chosen as stepwise and the entry and stay significance level is set at 0.05.

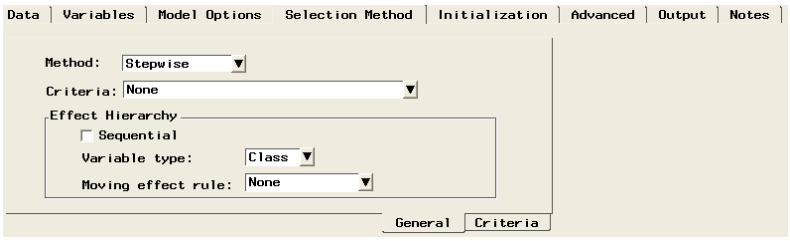


Figure 10. Stepwise Selection Method

After running the logistic model, we inspect the T-scores in Figure 11. The T-scores are ranked in decreasing order of their absolute values, which indicates that the higher the absolute value is, the more important the variable is. From Figure 11, it shows that cost, time and quantity are the most important model input variables.

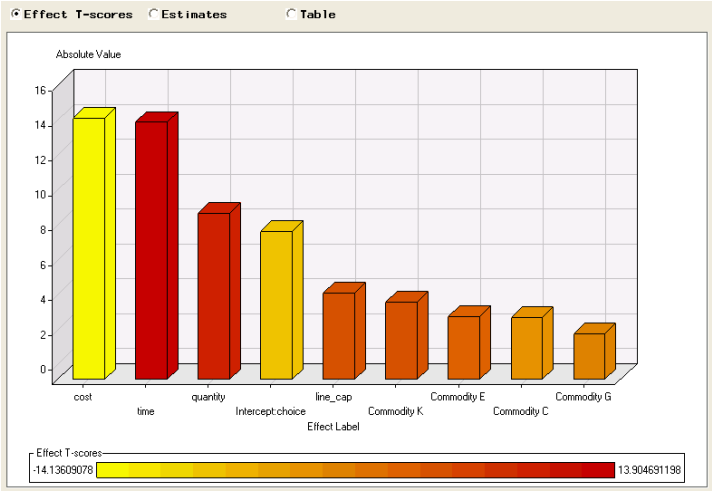


Figure 11. T-scores Ranking

▪ LIFT CHART

This chart divides observations into deciles according to their predicted response probability. For example, in

cumulative %response chart, the response is transportation mode choice of train (Choice = 1). For each observation, the logistic regression model predicts the probability that the mode choice is train. And all these observations are sorted by the predicted probability from highest to lowest. Then these observations are bagged into ordered bins, each including 10% of the whole data.

If the model is useful, then the top proportion of the observations will be relatively high while the predicted probability of response is high. For example, in the Figure 12, in the top 10%, almost 100% observations select train as the transportation mode. And in the top 50%, approximately 90% observations choose train. The base line is an estimate of the percentage of train expected if taking a random sample.

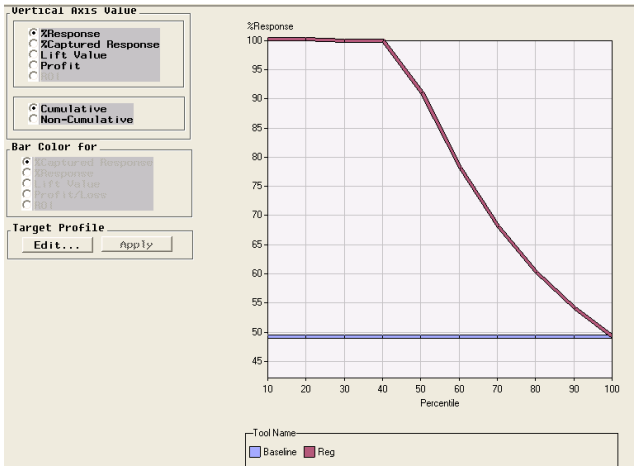


Figure 12. Cumulative %response Chart

▪ LIFT VALUE

The lift value chart reveals the same information from a different angle. The overall population response rate is close to 50% as shown in Figure 12. The lift can be calculated by dividing the response rate in a certain bin by the overall response rate. For example in Figure 13, in the top 10%, the response percentage is almost 100%, which is divided by 50%, and we can obtain the lift value above 2. This value indicates that the response rate in the top 10% is over two times as high as the response rate in the overall population.

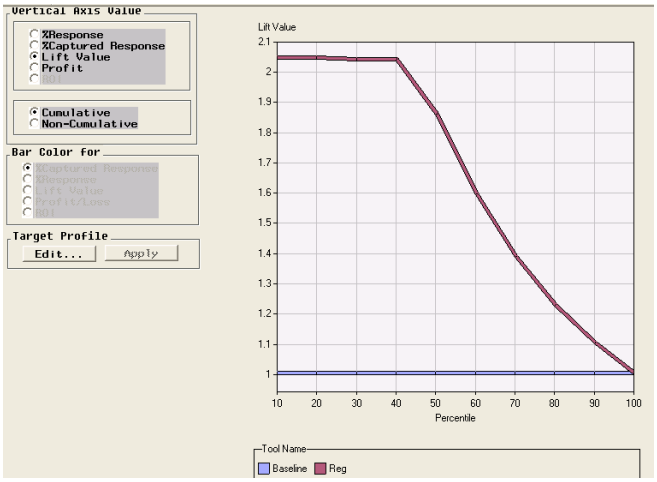


Figure 13. Lift Value Chart

▪ CAPTURED RESPONSE

Consider taking a random sample of 10% of the observations, it is expected that 10% of train is selected, similarly for 20%. From the %captured response chart in Figure 14, it shows that in the top 10% response, over 20% of those whose transportation mode choice is train, which equals a lift value over 20%/10%=2.

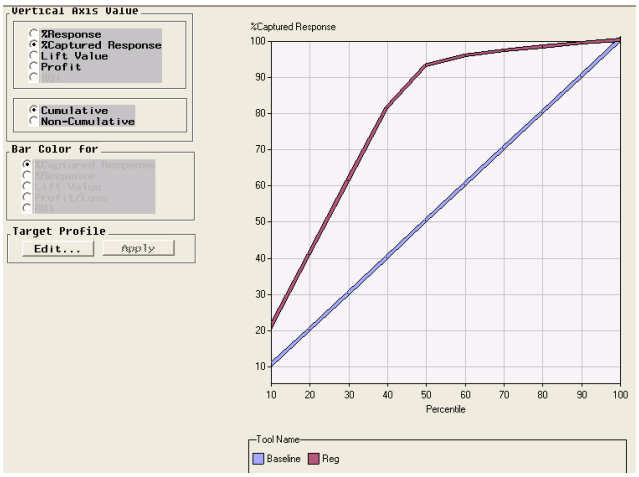


Figure 14. %Captured Response Chart

Generally, lift value decreases as the selected proportion of the data increases. The model which has higher lift value is usually preferred when comparing different models for the same proportion of data. And the parameter estimates are shown in Figure 15. It is noted that the variable Capacity is insignificant in this logistic regression model.

Effect T-scores Estimates Table

Effect Name	Effect Label	Parameter Estimate	Effect T-scores
CommodityA	Commodity A	0.0224627817	0.1120293299
CommodityC	Commodity C	-0.834256553	-2.724012562
CommodityE	Commodity E	0.6445751493	2.753786168
CommodityG	Commodity G	-0.615320568	-1.772974752
CommodityH	Commodity H	0.6343540131	1.7648018395
CommodityJ	Commodity J	-0.071381733	-0.124036735
CommodityK	Commodity K	1.0041148545	3.6139259518
Intercept	Intercept:choice=1	-2.187698919	-7.631143097
capacity	capacity	.	.
cost	cost	-0.416432715	-14.13609078
line_cap	line_cap	0.0136374312	4.1304982894
quantity	quantity	0.000015324	8.6640757991
time	time	3.3325536482	13.904691198

Figure 15. Logistic Regression Results

▪ VARIABLE TRANSFORMATION

From Figure 3 and 4, it is clear that the variables Capacity and Quantity are highly skewed. Hence we transform them by using logarithm. And the transformed figures are shown in Figure 16 and 17. From the transformation results in Figure 18, it is noted that the transformed variable Capacity is still insignificant, and parameter estimate of Quantity is changed significantly. And from T-scores ranking, it reveals that cost, time and quantity are still the most important model input variables.

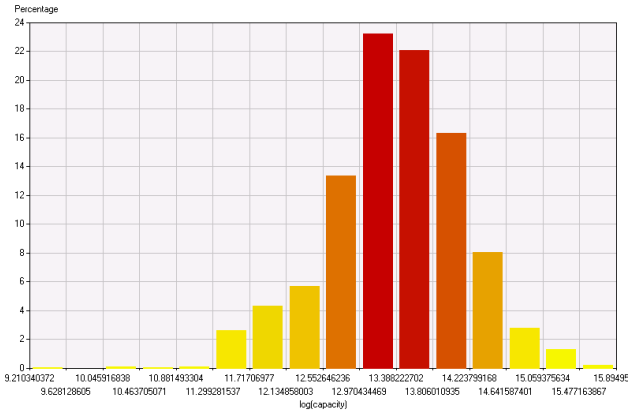


Figure 16. Transformed Variable Capacity

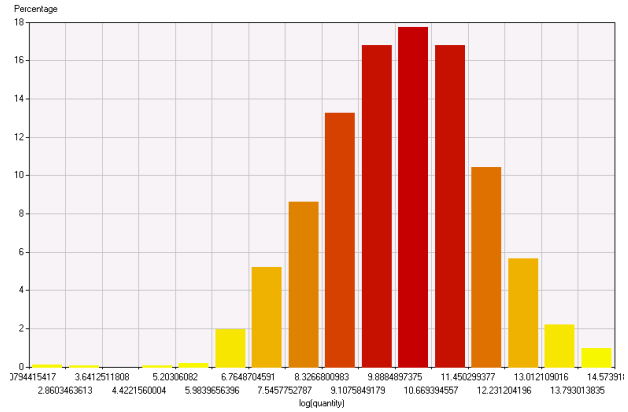


Figure 17. Transformed Variable Quantity

Effect T-scores Estimates Table

Effect Name	Effect Label	Parameter Estimate	Effect T-scores
CAPA_JP0	log(capacity)	.	.
CommodityA	Commodity A	-0.015012456	-0.07562276
CommodityC	Commodity C	-0.775591661	-2.673036034
CommodityE	Commodity E	0.8553680784	3.7465785063
CommodityG	Commodity G	-0.830872086	-2.395049025
CommodityH	Commodity H	0.5269240138	1.4484629571
CommodityJ	Commodity J	0.0149883261	0.0262468024
CommodityK	Commodity K	0.925163847	3.2941184411
Intercept	Intercept:choice=1	-9.098573303	-11.13792145
QUAN_BSF	log(quantity)	0.7583600996	10.068544175
cost	cost	-0.372805189	-14.54227368
line_cap	line_cap	0.0161780509	4.9714309278
time	time	2.9718186361	14.295541126

Figure 18. Transformation Results

▪ LIFT COMPARISON

From Figure 19, the transformed logistic regression is slightly better than the non-transformed logistic regression model from percentage top 40% to 60%. While from 60% to 80%, the non-transformed regression model is slightly better than the transformed logistic regression. After 80%, they are indistinguishable. Hence, for this transportation problem, the variable transformation technique does not improve the model significantly.

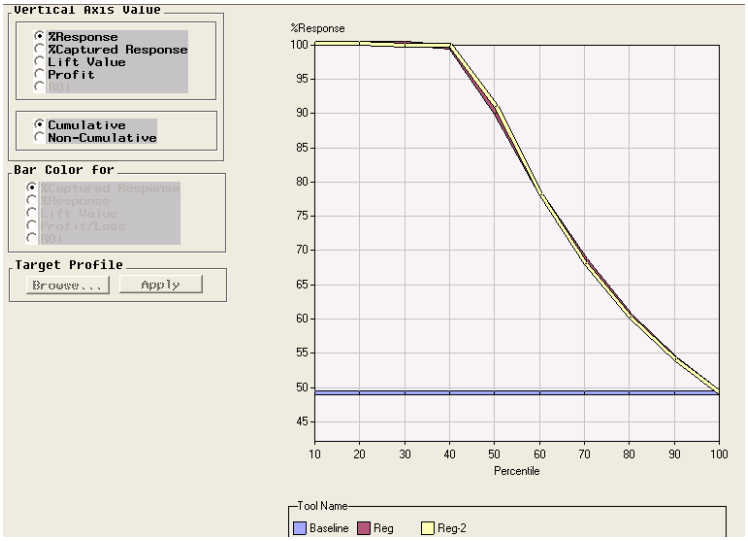


Figure 19. Comparison of Non-transformed and Transformed Models

▪ DECISION TREES

The decision trees is a quite popular data mining technique (Hastie, Tibshirani and Friedman, 2009), which is ease of use, robustness with missing data and ease of interpretability. Generally, decision trees are flexible, while regression models are relatively inflexible, for example, you have to add additional terms, i.e. interaction terms, polynomial terms. And decision trees can deal with missing values without imputation, while regression model usually has to impute missing values before building the model, although there is not missing data in our data set. And decision trees are nonparametric and highly robust, while regression models are parametric and sensitive to influencing points. We use the Tree node in Enterprise Miner to analyze the transportation data. The Gini reduction method is chosen as the splitting criterion

After running the tree model, we obtain the misclassification rate in Figure 20. It shows that how large a tree is needed. If the misclassification rate between training and validation data is close across all sub-trees, then choose the least number of leaves. From the results, the sub-trees which have 10 to 20 leaves have the smallest value of misclassification rates for validation. Therefore, the sub-tree with 10 leaves has been chosen. Since the leaves are large, we do not show the tree plot here. For this transportation problem, the number of leaves is so many that it is not easy to interpret.

Misclassification Rate

Leaves	Training	Validation
1	0.4967	0.4894
2	0.2864	0.2934
3	0.1975	0.2001
4	0.1975	0.2001
5	0.1276	0.1266
6	0.1276	0.1266
7	0.0893	0.0830
8	0.0893	0.0830
9	0.0852	0.0810
10	0.0817	0.0790
11	0.0817	0.0790
12	0.0817	0.0790
13	0.0817	0.0790
14	0.0817	0.0790
15	0.0817	0.0790
16	0.0817	0.0790
17	0.0817	0.0790
18	0.0817	0.0790
19	0.0817	0.0790
20	0.0817	0.0790
21	0.0750	0.0817
22	0.0750	0.0817

Figure 20. Misclassification Rate

▪ COMPARISON OF TREE AND REGRESSION MODEL

After running assessment node, we could obtain the comparison result of tree model and regression model, which shows that the transformed regression model and non-transformed regression model are both better than tree model from 10% to 80%, after 80%, they are all indistinguishable. And the performance of transformed regression model and non-transformed regression model are very close. Hence, for this transportation problem, the non-transformed logistic regression model is preferred.

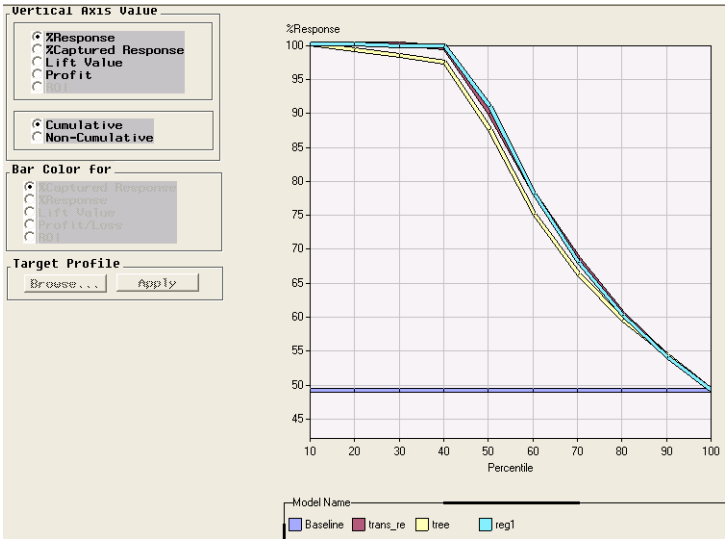


Figure 21. Comparison of Tree and Regression Model

▪ PREDICTION

We use the non-transformed logistic regression model to predict the test data set. By using score node in Enterprise Miner, we obtain the prediction results in Figure 22. For example, for observation 1, the probability of choosing choice 1 is 0.06 and the probability of choosing choice 0 is 0.94, which implies that there is 94% chance to choose truck (choice = 0). And the whole program flow chart is shown in Figure 23.

9	Int	Nom	Int	Int	Int	Int	Int	Int	Int	Int
1469	choice	Commodity	capacity	line_cap	quantity	time	cost	P_choice1	P_choice0
1	0	E	10000	0	513	9.2209	76.6581	6.1283620E-02	9.3871638E-01
2	0	A	30000	10	22179	2.9969	24.9149	1.1135521E-01	8.8864479E-01
3	0	A	30000	10	40190	4.2251	35.1252	1.2347041E-01	8.7652959E-01
4	0	L	50000	1	2000	6.1900	51.4606	2.3446296E-02	9.7655370E-01
5	0	L	55000	1	12000	3.4589	28.7558	3.8327250E-02	9.6167275E-01
6	0	L	80000	1	14259	3.7562	31.2268	3.8185136E-02	9.6181486E-01
7	0	A	82000	25	5296	5.2160	43.3632	8.1779829E-02	9.1822017E-01
8	0	A	84000	0	7467	22.0277	181.8386	1.2544023E-02	9.8745598E-01
9	0	G	95000	10	800	4.5780	38.0592	3.7432941E-02	9.6256706E-01
10	0	A	95000	10	7083	6.1655	51.2572	6.1889854E-02	9.3811015E-01
11	0	C	95000	10	2644	6.1655	51.2572	2.5499814E-02	9.7450019E-01
12	0	C	95000	10	36359	6.1655	51.2572	4.2023028E-02	9.5797697E-01
13	0	C	95000	10	2991	6.1655	51.2572	2.5632284E-02	9.7436772E-01
14	0	E	95000	10	6012	6.1655	51.2572	1.0785791E-01	8.9214209E-01
15	0	J	95000	10	4633	6.1655	51.2572	5.4686520E-02	9.4531348E-01
16	0	A	95000	10	1637	5.8928	48.9895	5.9153023E-02	9.4084698E-01
17	0	A	95000	10	2567	5.8928	48.9895	5.9951165E-02	9.4004884E-01
18	0	C	95000	1	10134	6.1655	51.2572	2.5302748E-02	9.7469725E-01
19	0	C	95000	1	11642	6.1655	51.2572	2.5878958E-02	9.7412104E-01
20	0	C	95000	1	1698	6.1655	51.2572	2.2302819E-02	9.7769718E-01
21	0	E	95000	1	2991	6.1655	51.2572	9.2638064E-02	9.0736194E-01

Figure 22. Prediction Results

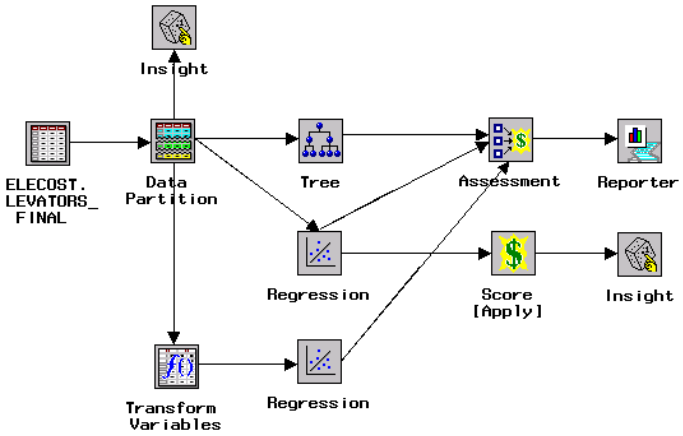


Figure 23. Flow Chart

▪ CONCLUSION

This paper illustrates how to develop decision tree and logistic regression model for a real transportation problem. The decision tree in this case contains quite a lot leaves and is not easy to interpret, while the non-transformed logistic regression model perform better than tree model. Also the non-transformed regression model and transformed regression model are indistinguishable. Hence, for this problem, we recommend to use the non-transformed logistic regression model.

▪ REFERENCES

- Alan Agresti, 2002. *Categorical Data Analysis*, Second Edition. John Wiley & Sons, Inc., Hoboken, New Jersey.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman, 2009. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Second Edition. Springer, New York.

▪ CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Xianzhe Chen

Enterprise: North Dakota State University

Address: 12th Ave North

City, State ZIP: Fargo, ND 58105

Work Phone: 701-231-5763

E-mail: Xianzhe.Chen@ndsu.edu