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>
<thead>
<tr>
<th>Commodity</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Wheat</td>
</tr>
<tr>
<td>C</td>
<td>Durum</td>
</tr>
<tr>
<td>E</td>
<td>Barley</td>
</tr>
<tr>
<td>G</td>
<td>Sunflower</td>
</tr>
<tr>
<td>H</td>
<td>Corn</td>
</tr>
<tr>
<td>J</td>
<td>Oats</td>
</tr>
<tr>
<td>K</td>
<td>Soybeans</td>
</tr>
<tr>
<td>L</td>
<td>Deb (Dry Edible Barley)</td>
</tr>
</tbody>
</table>

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.
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.
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.

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.
• 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.

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

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.

• 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](image)

**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](image)
**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](image)

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.

![Figure 15. Logistic Regression Results](image)

**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.
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.
**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.
**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.

**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.

![Figure 21. Comparison of Tree and Regression Model](image1)

![Figure 22. Prediction Results](image2)
• 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

• 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

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration.

Other brand and product names are trademarks of their respective companies.