John Brendan Owens

Data Mining

Final Project

6/26/17

Introduction

Data mining is the practice of automatically searching through large amounts of data to discover patterns and trends that would not be readily discoverable through standard statistical analysis. Data mining is increasingly important as the amount of structured and unstructured data rapidly grows. There are currently less than 20,000 exabytes of data, but by the year 2020 there will be over 50,000 exabytes of data. The amount of data will not be able to be analyzed through traditional methods, and we will need to rely on data mining to explore this data. Business Intelligence will rely on data mining to discover trends in these large amounts of data. One of the greatest abilities of data mining is to create predictive models.

Given a large dataset of Sam’s Club store visit information, I sought to create a predictive model to define aspects of a visit that would predict a higher than average visit purchase amount. I plan on creating a binary variable in Excel to flag whether the purchase amount for a visit is $50 higher than the average purchase amount. I then will use SAS Enterprise Miner to explore the relationship between variables. The data mining abilities of SAS Enterprise Miner allows for quicker automatic data exploration compared to traditional data analysis techniques. I will then use those variables to create models to predict whether a visit will result in a higher than average purchase total. I will conduct logistic regressions, decision trees, and neural network to determine the rules for predicting a greater than average visit purchase. The logistic regression offers an applicable model to predict if a visit will result in a larger than average purchase total. Logistic regression also gives the useful information of the weight of a variable in predicting the outcome. Decision tree is a beneficial model as it gives rules to define if a visit will be $50 over the average or below that amount. This model does not have the predictive ability of logistic regression, but it more clearly defines rules for variables that describe a positive outcome. The neural network does not develop rules for predicting the outcome, but it is the best predictive tool out of all the models.

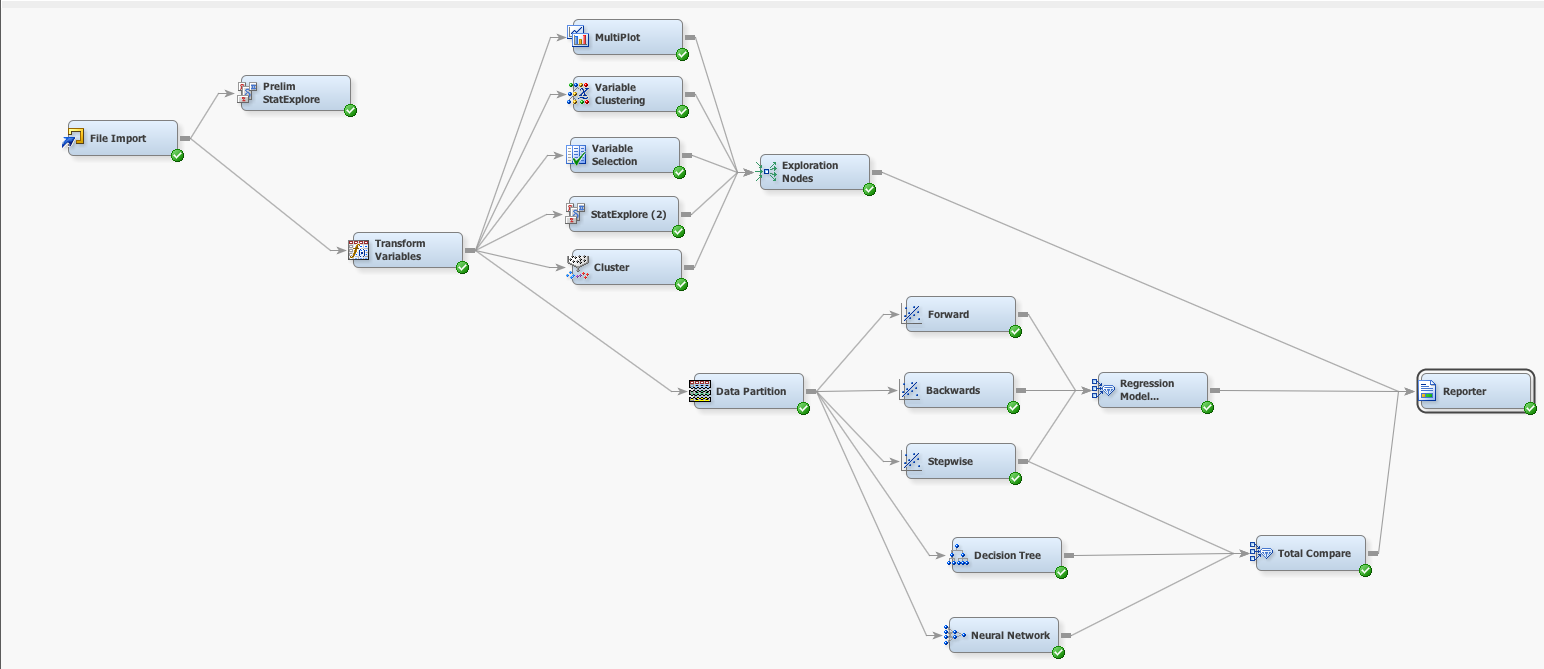
The question of what defines a visit that results in a higher than average purchase total, is important to increase the efficiency of Sam’s Club stores. From a business perspective, it is a better use to focus on driving these greater purchases because it allows for stores to be more efficient and thereby more profitable. I hope that my models give the ability to define the features of a greater than average purchase and to be able to accurately predict when those visits occur.

Discussion

To answer what features predicted a visit $50 greater than average, I first explored and cleaned data in Excel. I decided that VISIT\_NBR would be the identification variable for the data set, and by using this variable I could then create models to predict if a visit would be a positive outcome. I then chose TOTAL\_VISIT\_AMT as the variable to measure the total purchase amount. I chose this variable because it includes the total of all the items purchased in the visit without the sales tax. This allows stores in different states to be compared because it takes away the variation of sales tax between states.

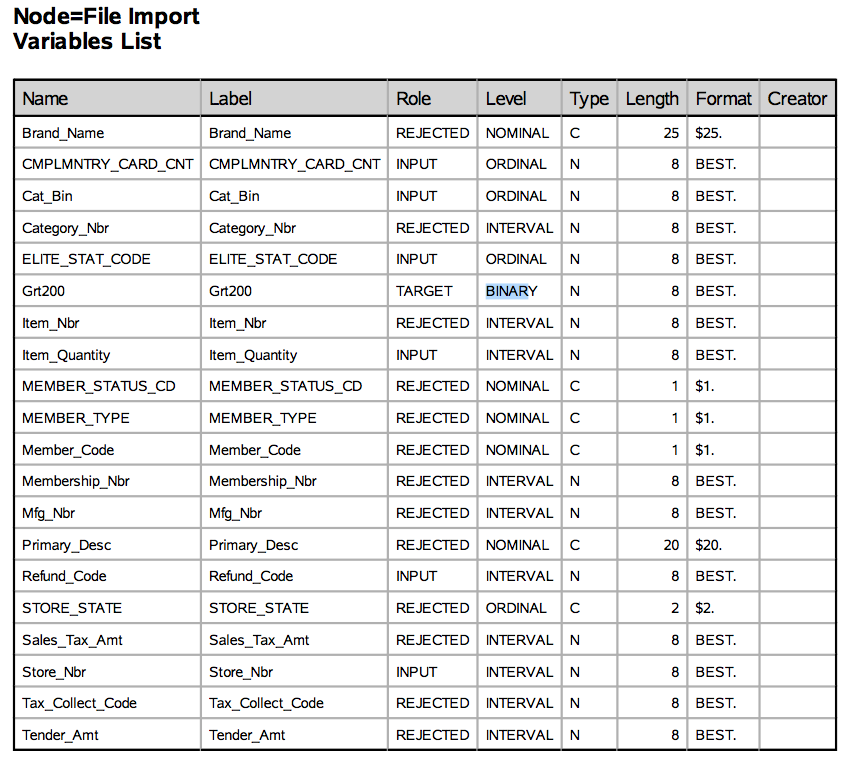
After determining TOTAL\_VISIT\_AMT would be the basis for the target variable, I found TOTAL\_VISIT\_AMT average to be $151, so I then created a binary variable to flag if TOTAL\_VISIT\_AMT was greater than $200. I created a new variable Grt200, and I coded it with a 1 if TOTAL\_VISIT\_AMT was greater than $200 and a 0 if it was less than. I define a positive outcome of the target variable, Grt200, as 1 or a purchase greater than $200. I define a negative outcome as a 0 in the target variable or a purchase below $200. While this is not the precise meaning of a positive or negative outcome, I use this terminology to help explain the models I create. I also created another variable called Cat\_Bin. I coded to create an ordinal variable for Category Number. This was the only data exploration/cleaning that I did in Excel, the rest was done in SAS Enterprise Miner.

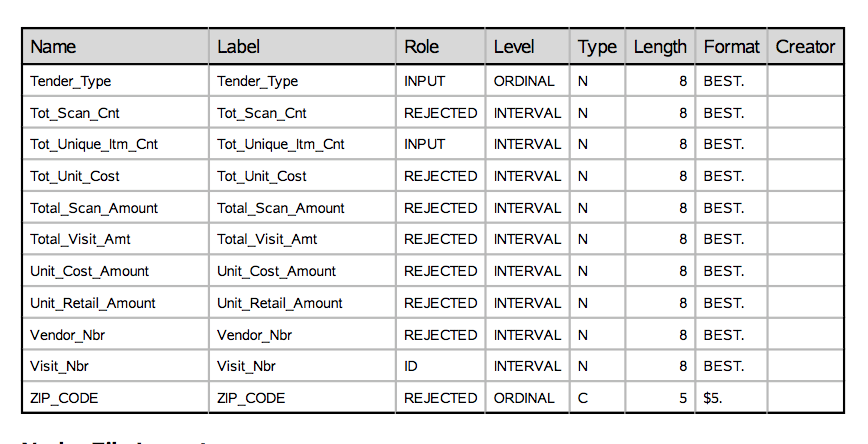
**Figure 1: Complete Diagram**



I first imported the data and decided which variables that I would keep and which would be rejected. I had already determined VISIT\_NBR would be the identification variable and Grt200 would be the binary target variable. I then rejected Item\_Nbr, all member information variables, Primary\_Descr, Business Credit Type Status, Tax Collect Code, and Vendor\_Nbr. I rejected these variables because of their form in this data set was very unorganized, did not have a proper relevant definition, were scrubbed from the data set, or were not relevant by definition to the target variable. I also opted to not use any of the purchase cost information variables such as Tot\_Unit\_Cost, Total\_Scan\_Amount, and Sales\_Tax\_Amt. I originally kept these variables included in preliminary modeling, but I realized, by definition these variables are too multi-collinear with the target variable, and they in essence share the meaning Total\_Visit\_Amt which is the basis for the target variable. I also rejected Unit\_Cost variables because they only have information for one item of the visit so are not predictive of the trends of the visit as a whole. This left me with 8 input variables.

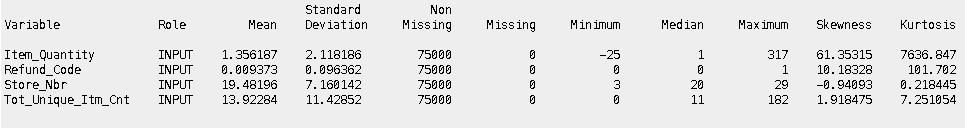
**Figure 2: Imported Data, Rejected and Input Variables**





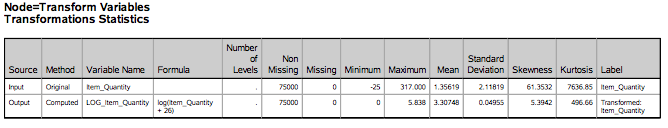
The 8 input variables I then sought to explore and use to create a predictive model were Complmntry\_Card\_Count, Cat\_Bin, Elite\_Stat\_Code, Item\_Quantity, Refund\_Code, Store\_Nbr, Tender\_Type, and Tot\_Uniq\_Itm\_Cnt. I found each of these variables important and showing different aspect of the visit. Complmntry\_Card\_Count and Elite\_Stat\_Code offered information of the type of member making the visit, in terms of their engagement with Sam’s Club. Item\_Quantity and Tot\_Uniq\_Itm\_Cnt offered information on how many items and unique items were bought in a visit. The Refund\_Code is a binary variable that flags if any of the items were returned, Tender\_Type is an ordinal variable showing how the purchase was made, and Store\_Nbr was used to see if there were difference between the stores.

**Figure 3: Preliminary StatExplore Results**



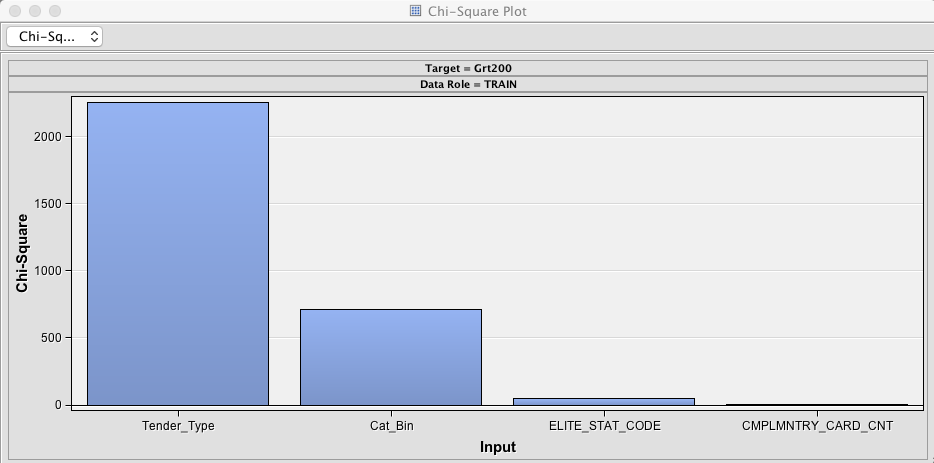
I did not assume normality for the distribution of any of these variables. I used a preliminary StatExplore node to explore the descriptive statistics of their distributions. As seen in Figure 3, I found Item\_Quantity did not have acceptable skewness and kurtosis levels. I corrected for this non-normal distribution by transforming the variable to be logarithmic Item\_Quantity. This transformation brought the kurtosis and skewness to more normal levels as seen in Figure 4. Kurtosis for the transformed variable was still too high, but it saw a significant reduction, making it acceptable for analysis.

**Figure 4: Transformation Statistics**

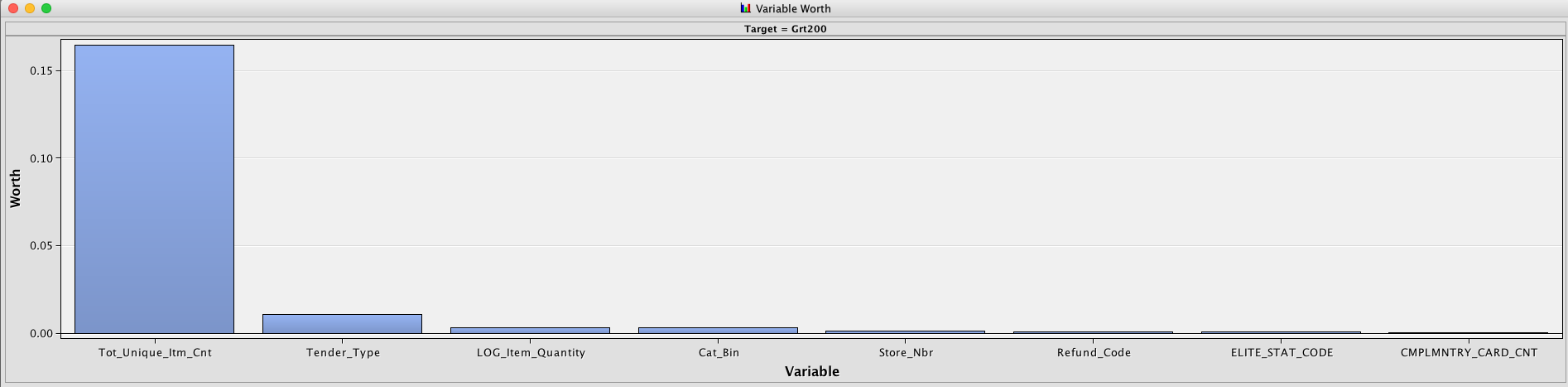


After I checked for normality and transformed the non-normal variable, I then sought to explore the variables and find their connections and trends with each other and the target variable. I did a cluster and a variable cluster analysis, but did not gain any beneficial insight to the patterns based on this node. The most important information on which variables would have high predictive powers came from the StatExplore node. As noted in Figure 5, I found Tender\_Type and Cat\_Bin had high correlations with the target variable and would likely lead to greater predictive ability in my models. I also found, in Figure 6, that Tot\_Unique\_Itm\_Cnt had a high Variable Worth and would also contribute to higher predictability.

**Figure 5: Chi Square of Variables from StatExplore Node**

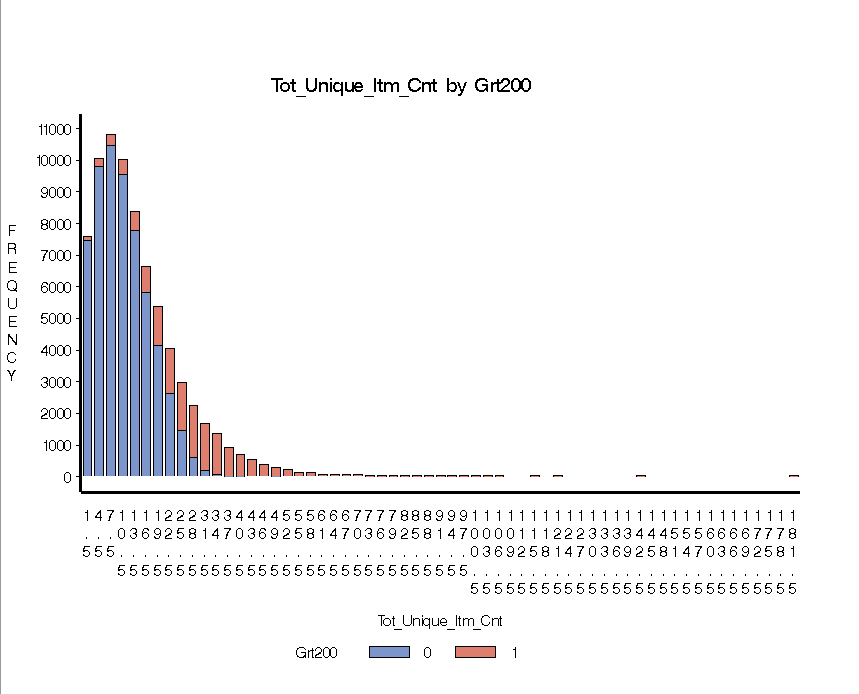


**Figure 6: Variable Worth from StatExplore Node**

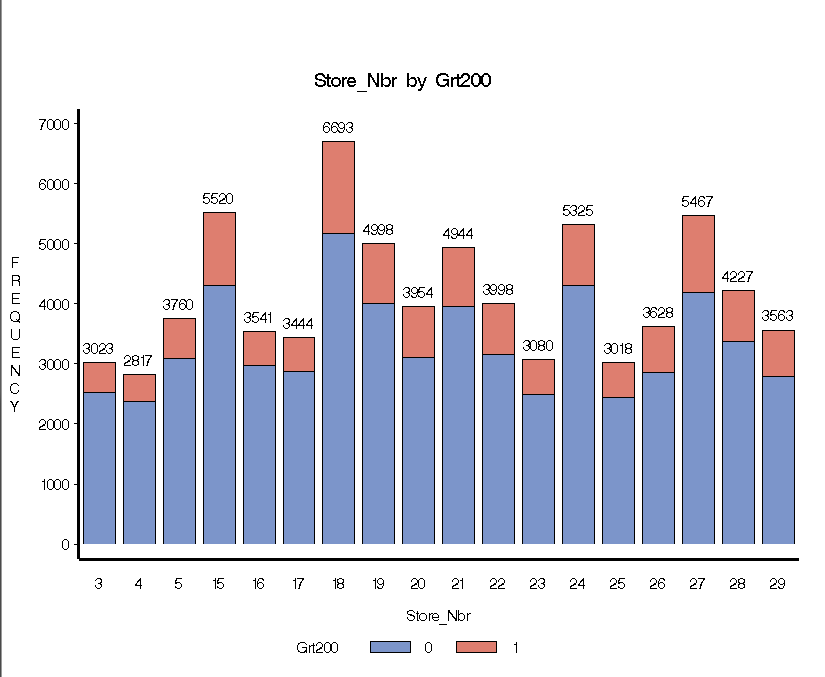


I also confirmed these results with the Multiplot. In Figure 7, it is shown that higher Tot\_Unique\_Itm\_Cnt is highly related to the total purchase amount being greater than $200. Another interesting graph from Multiplot, shown in Figure 8, illustrates that each store had similar rates of positive and negative outcomes for the target variable. This indicates that the stores do not differ on the rate of visits that spend $50 greater than the average amount, and different stores are not an indicator of the target variable.

**Figure 7: Tot\_Unique\_Itm\_Cnt from Multiplot Node**

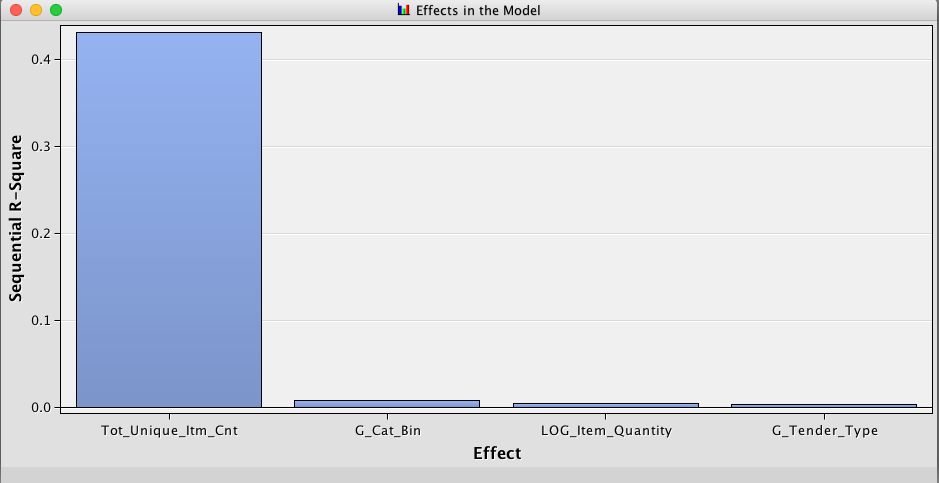


**Figure 8: Store\_NBR from Multiplot Node**



These effects and correlations were also confirmed in the Variable Selection node. Tot\_Unique\_Itm\_Cnt again had a very high worth in potential models and a high correlation with the target variable as seen in Figure 9.

**Figure 9: Variable Worth from Variable Selection Node**

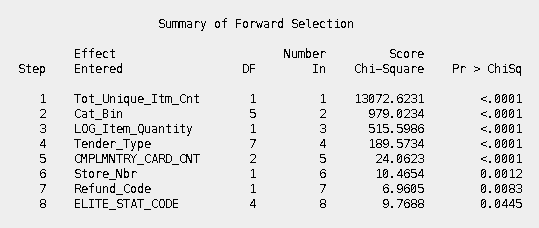


From these explorative nodes, I discovered the important information of how correlated Tot\_Unique\_Itm\_Cnt is to the target variable. I also learned there were not differences between the stores in the rate of positive target variable outcomes, which meant that I did not have to account for differences between the stores in the modeling. After I conducted these explorative test, I partitioned the data to 40% test data and 60% validation data for modeling. I then created three different logistic regression with each having a different selection method of Backwards, Forwards, or Stepwise. I also created a decision tree and a neural network.

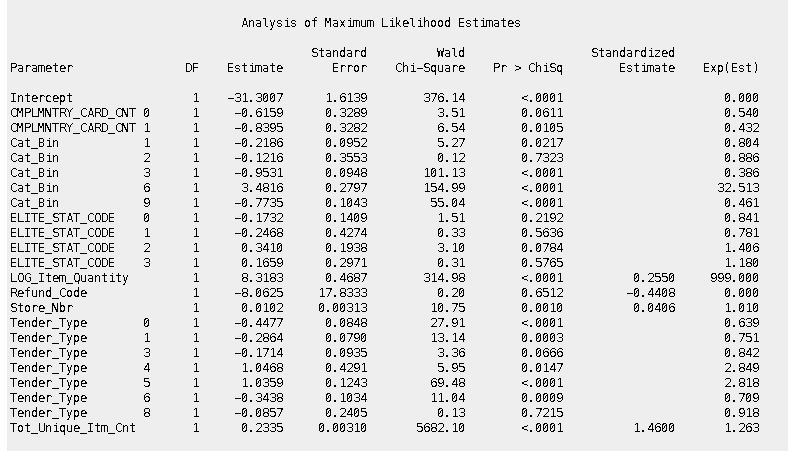
Discussion of Models

The first model produced was a Forward Logistic Regression. This model use every variable, with Tot\_Unique\_Itm\_Cnt, Item\_Quantity, Cat\_Bin, and Tender\_Type having the strongest predictive effects, based on Chi-Square scores from Figure 10. Looking further into Analysis of Maximum Likelihood Estimates in Figure 11, the item quantity variables again showed high worth and positive correlation to the target variable, but other than those variables, Cat\_Bin 3 and 6 had the highest effect. Cat\_Bin 3 had a high effect with a negative correlation. Cat\_Bin 3 relates to Personal Care products therefore when people are buying things for personal care it is likely that they will not spend greater than $200. While Cat\_Bin 6 had a high effect with a positive correlation and relates to Auto supplies, purchase of Cat\_Bin 6 products related to positive outcomes of the target variable. This model had a miscalculation rate of 0.09 for both Training and Validation data, Figure 12. The most important information from this model is the positive correlations and predictive abilities of the item quantity variables, and that Cat\_Bin 6 had a large positive prediction effect while Cat\_Bin 3 had a large negative prediction effect. However, the effects from Cat-Bin 3 and 6 could be due to Personal Care items not being particularly expensive and Auto items being very expensive.

**Figure 10: Summary of Forward Selection**



**Figure 11: Analysis of Maximum Likelihood Estimates**

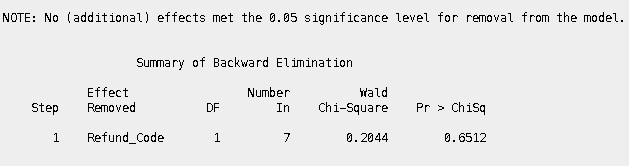


**Figure 12: Forward Model Miscalculation Rate**

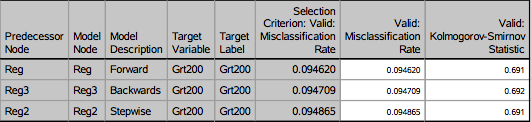


I then conducted another logistic regression, but I used a backwards selection method. The Backwards Model only difference from the Forward Model was the removal of Refund Code as a variable, Figure 13. This removal did not significantly decrease the misclassification rate compared to the Forward Model, Figure 14. The Backwards Model also showed the same effects as the Forward Model.

**Figure 13: Removal of Refund Code from Backward Model**

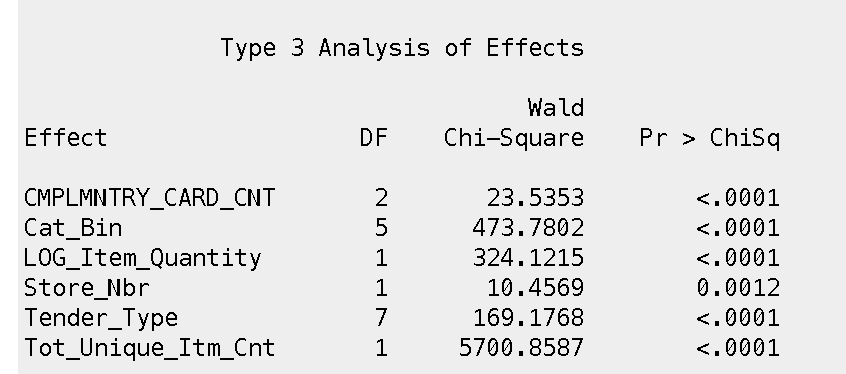


**Figure 14: Regression Models Comparison, Misclassification Rates**



I then created another logistic regression using stepwise selection. The Stepwise Model was the same as the Backwards Model, expect the Stepwise Model did not have Elite\_Stat\_Code, Figure 15. This did not change the miscalculation rate significantly, Figure 14. I prefer the Stepwise model out of the 3 logistic regression models because it has the same miscalculation rate as the other two while removing two variables with very lower effects. However, I am concerned with the future efficacy of these models. I believe they might be over-fitted to this dataset and not be applicable to new data. While each model has the same miscalculation rate for the train and validation data, I believe the concerning factor is how low the miscalculation rate is. These models could be very useful for predictive purposes, but they are not particularly useful for developing rules that define a visit of greater than $200. I then created a decision tree to develop rules that indicate a positive outcome.

**Figure 15: Stepwise Logistic Regression Model**

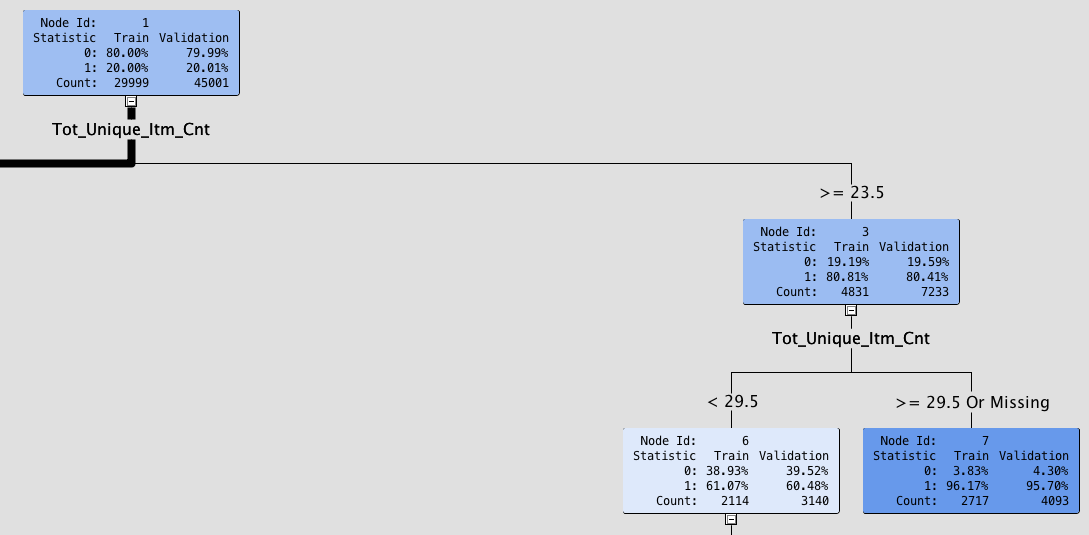


From the Decision Tree Model, I was able to create several rules to determine if the target variable would be a positive outcome for being greater than $200 for the purchase or less than. The first rule, from Figure 16, being,

If Tot\_Unique\_Itm\_Cnt>=29.5

Then there is a 95% chance the total visit amount is greater than $200.

**Figure 16: Decision Tree Rule 1**



This is my favorite of the rules because it takes the most important factor, based on all the explorative results, and uses it to make a simple rule on if there will be a positive outcome. The simplicity and accuracy of this rule makes it great for defining purchases that are greater than $200. According to this rule, the greatest factor in a purchase being more than the average is the number of unique items purchased. The second-best rule is one that determines when purchases are not greater than $200. It is stated, in Figure 17 as,

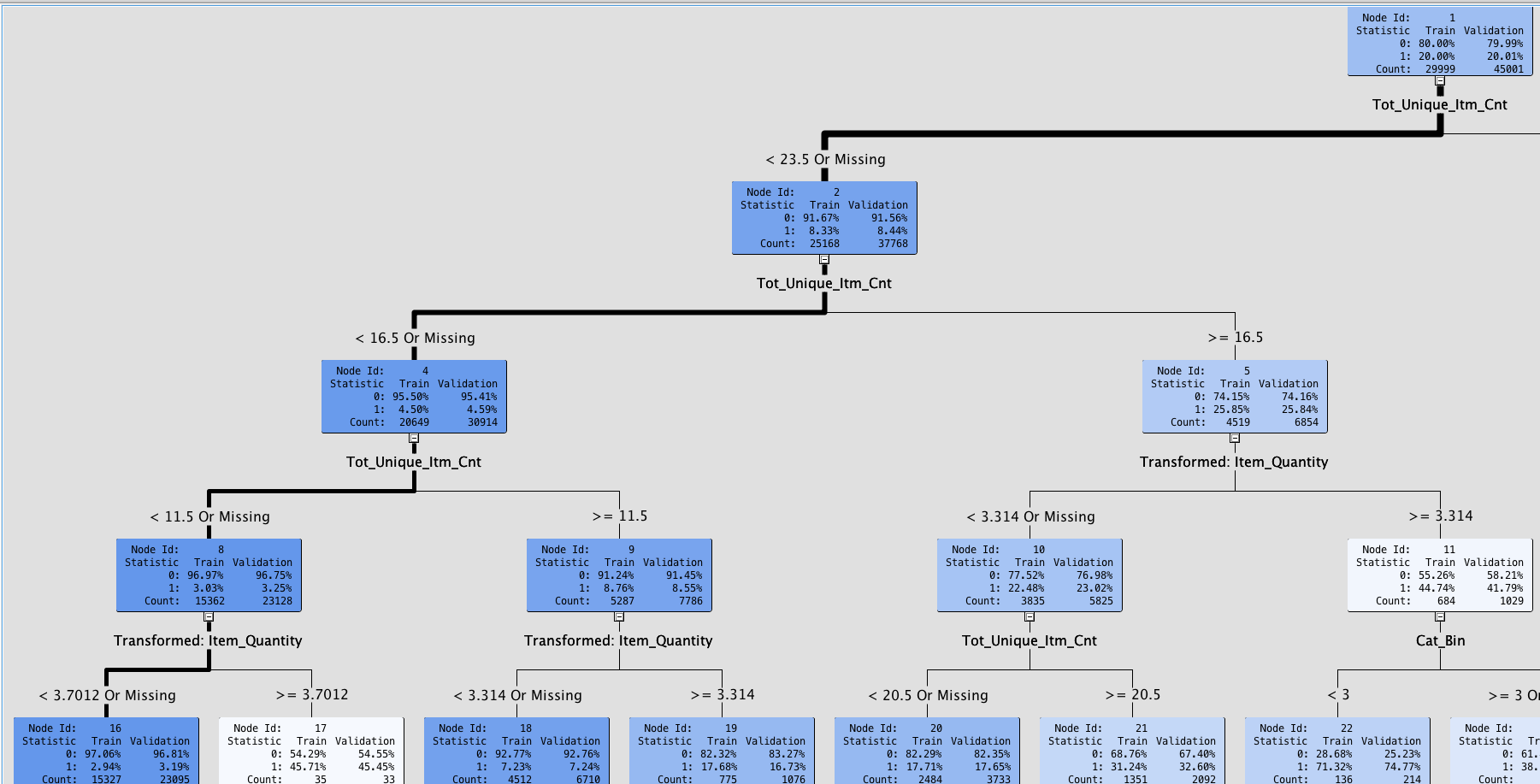
If Tot\_Unique\_Itm\_Cnt <11.5

And Transformed Item Quantity<3.7

Then there is a 97% chance the total visit amount is less than $200

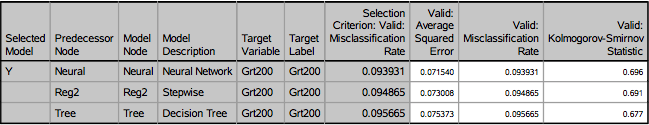
This is another great rule because it shows that the biggest determinant on the outcome of the target variable is based on the item quantity variables. While other rules were computed for the Decision Tree, they narrowed down to too small amount of data for accuracy and were not very applicable to the data set as a whole. The most efficient and applicable rules only had definitions based on the item quantity variables.

**Figure 17: Decision Tree Rule 2**



The Decision Tree model created clear rules for defining what was likely a greater than $200 purchase. I also created a Neural Network to see if I could improve upon the predictability of the Decision Tree and Logistic Regressions. I created a Neural Network and then used the Model Comparison node to compare it to my other models.

**Figure 18: Model Comparison Miscalculation Rate Results**



**Figure 19: Model Comparison Lift Graph**

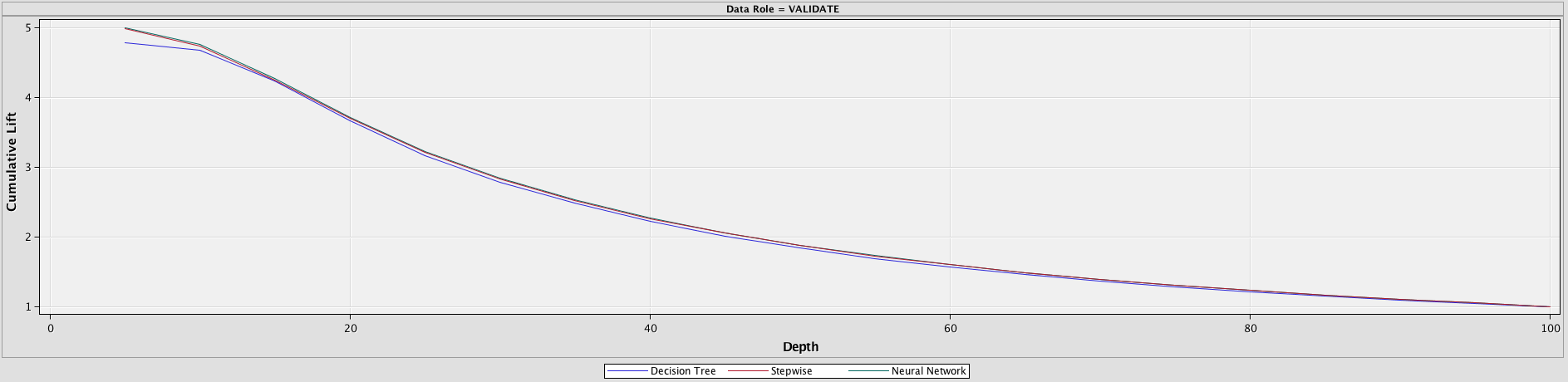


Figure 18 shows that the Neural Network was able to predict the outcome of the target variable with greater accuracy than the other two models. However, this improvement is minimal. As shown in Figure 19, the comparison of the Lift between each model is practically indistinguishable and each model offers very strong predictive ability.

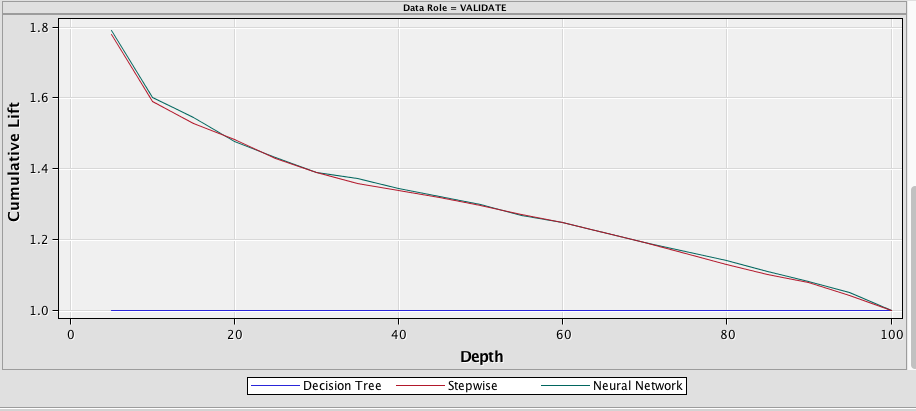
Each model offered different and important information. The logistic regression models showed how much predictive power the item quantity variables had as well as illustrating the predictive ability and differences between Cat\_Bin3 and Cat\_Bin 6 groups. The Decision Tree then showed that the most defining characteristics of a visit total greater than $200 is the number of unique items and overall number of items. The Neural Network then proved to be the most predictive of the models. I prefer the Decision Tree the best though. It created the simplest rules for predicting the outcome variable while maintain the high degree of accuracy as the other models. From a business perspective, the Decision Tree offered the most insightful discovery.

Conclusion and Recommendations

The major conclusion from this data exploration and modeling is that the defining predictor of Sam’s Club Visit that result in purchases greater than $200 is a greater number of unique items and overall items purchased. Sam’s Club should focus on driving customers to make more purchases in a visit if their goal is to raise the average amount per visit. This would increase the efficiency of the stores and raise profit margins.

I do believe these models are inherently flawed from their simplicity. It seems very obvious that purchasing more items results in greater amount for that purchase. The total amount for visit and items purchased in a visit are very similar variables in their nature. I did run the Diagram, removing the item quantity variables, and I found the models to still be very predictive.[[1]](#footnote-1) These models had misclassification rates of 20%, and they primarily used Tender\_Type and Cat\_Bin for their prediction. These would have been good models on their own, just not as predictive as the models with the item quantity variables.

**Figure 20: Model Comparison without Item Quantity Variables.**



I believe the best way to improve this dataset and these models is to add more information about the members. I believe there would be better models to predict if a visit resulted in purchases greater than $200, if information on how many times the member visited or how long they have been members. It would also be useful to have information on every item purchased per visit. This would give a greater insight into what items are predictive of purchases greater than $200 and give more insight to those trends.

This exercise in data mining was a good start in examining the trends of Sam’s Club consumers. I believe that the most important aspect of data mining is simply finding a place to begin to look for insights and then building upon those results.

1. I overlooked the collinearity of the item quantity variables and my target variable of purchases greater than $200. I should’ve spent more time on exploration of models without these variables, but I had overlooked it till I did not have time to. [↑](#footnote-ref-1)