**Customer Churn Prediction in Metro Cash and Carry, India**

Anurag Pandey

Harsha Vardhan Vemulapati

Rohit Sarkar

Rony Chacko

Shantanu Kulkarni

Contents

[Introduction 3](#_Toc468048029)

[Problem Statement 3](#_Toc468048030)

[Dataset Description 3](#_Toc468048031)

[Summary 5](#_Toc468048032)

[1. First Iteration 5](#_Toc468048033)

[2. Final Iteration 5](#_Toc468048034)

[Data Pre-processing 7](#_Toc468048035)

[1. Missing value treatment 7](#_Toc468048036)

[2. Outlier detection 8](#_Toc468048037)

[3. Data Partition 9](#_Toc468048038)

[4. A Priori statistics 9](#_Toc468048039)

[5. Logistic Regression (on raw data) 9](#_Toc468048040)

[6. Multicollinearity Check 10](#_Toc468048041)

[7. Variable selection Method 11](#_Toc468048042)

[8. Data Transformation 12](#_Toc468048043)

[a) Transform Variable node observation with optimal binning: 12](#_Toc468048044)

[b) Log transform 12](#_Toc468048045)

[c) Interactive binning 13](#_Toc468048046)

[d) Feature Creation 15](#_Toc468048047)

[e) Variable Standardization 15](#_Toc468048048)

[f) Interaction variable 16](#_Toc468048049)

[Data Modelling 17](#_Toc468048050)

[Model Comparison 17](#_Toc468048051)

[a) Decision Tree 19](#_Toc468048052)

[b) Regression(3) 20](#_Toc468048053)

[c) Neural network 21](#_Toc468048054)

[Conclusion 21](#_Toc468048055)

[Recommendations 22](#_Toc468048056)

Plan Of Action………………………………………………………………………………………………………………………………………..23

# Introduction

Retailers today are constantly finding innovative ways to draw insights from a growing repository of structured and unstructured data. Retailers that embrace a data-first strategy towards understanding their customers, matching them to products and parting them from their cash, secure a competitive advantage over their rivals in the marketplace. Major bricks n’ mortar chains have fought hard to keep up with, and in some ways better, the advances in technology driven by the online retail boom.

METRO Cash and Carry is represented in over 25 countries with over 750 business-to-business wholesale stores and a head count of over 110,000 employees. Its business model is defined by its customer base. Registered business customers visit a Cash & Carry outlet, select their own purchases and carry these back themselves instead of placing orders with multiple vendors.

METRO’s unique business model is defined by its registered wholesale customer base. Evaluating key transactional trends around this customer base is vital to METRO’s ongoing success in the retail industry.

# Problem Statement

METRO India wants to analyze key indicators of customer churn.

Specifically, we want to identify the key drivers of customer churn. This would involve answering questions around:

1. The impact of promotional vs non-promotional sales, discounts, advertisements on customer churn
2. Churn analysis across stock-up, immediate-needs, department-destination and seasonal products across a common basket size.
3. As a leading grocery retailer, METRO wants to identify the differentiating factors between food and non-food customer visits
4. METRO wants to analyze the visit patterns: repeat rate, trial rate, recent visit, tenure

METRO wants to understand the combined effect of these and other business variables on customer churn. Furthermore, METRO also needs recommendations on additional data to be collected to address these pertinent questions.

# Dataset Description

Our project uses first-hand data received from a METRO employee. Metro uses this sample live dataset to test new applicants who want to join the analytics wing of the organization. This is public and therefore available to all to use.

The dataset encapsulates a varied set of purchase behaviors of customers. This multivariate dataset has 10000 instances across 96 attributes, including 94 predictive attributes, 1 non-predictive and 1 goal (Target) variable. The one non-predictive attribute (Customer ID) of the dataset does not affect the outcomes of analysis and so will be disregarded for the analysis.

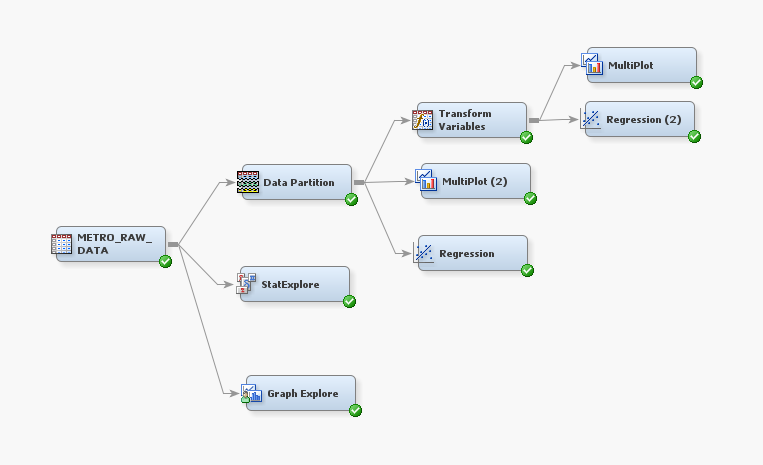
Variables are listed in the below attached excel sheet:



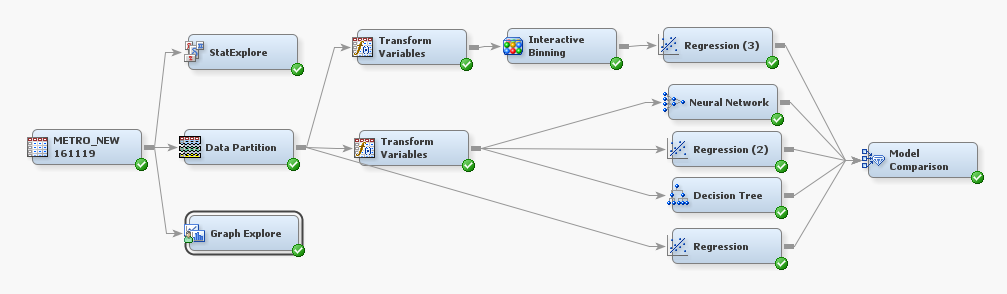
|  |  |
| --- | --- |
| How many observations in the dataset | 10000 |
| How many binary variables | 4 |
| How many nominal variables | 1 |
| How many interval variables | 91 |
| What is the outcome / target variable | customer churn or not |
| What is the level of the target variable (nominal, binary or interval) | binary |
| If binary or nominal: What percentage of the variables belong to each class. | 32.04% - churned customer 67.96% - consistent customer |
| If interval: What is the mean value of the target variable? | NA |
| Before doing any further processing, what is your prediction of the target variable? | Approximately 7 out of 10 customers will continue shopping in Metro Cash & Carry |

# Summary

## First Iteration



## Final Iteration



As shown above, the final model represents a significant upgrade from the first iteration previously outlined in the progress report. To identify the leading indicators of customer-churn the following steps were executed:

* Data pre- processing – Detection of missing values, outliers (Box plots), Multi plots
* Data Partition – Training sample – 60%, Validation sample – 20%, Test sample – 20%
* Logistic Regression for baseline accuracy – Analysis has been performed on the interval variables and accuracy and precision did not improve significantly than A priori statistics
* Variable selection – Multicollinearity check using Pearson correlation coefficient and backward selection method to eliminate non-significant variables by using chi-square statistics and significance value. And AIC criteria has been selected to maximize likelihood estimation
* Variable Binning – Optimal binning methodology has been used in binning methods which provides optimal numbers of bins considering Target variable
* Logistic Regression after optimal binning – Regression analysis has been performed on the final list of selected optimal binned variables and Accuracy improved on the 1st attempt by approximately 4% over and above regression results on raw data
* Log Transformation – were used to minimize skewness and to increase the level of significance of some variables
* Interactive Binning – After Fine classing of continuous variable, Coarse-classing has been performed by visualizing distribution of Event-count across segments and considering WOE and Gini Statistics to further improve binned groups significance and contribution
* Interaction variable – To account for the chemistry of explanatory variables, interaction variable was introduced to explain hidden information of response variable but this step did not improve the accuracy as desired
* Feature creation – Given that the data is one-dimensional in the direction of purchase pattern, in order to account for other dimension like price point-point preferences and other psychographic variables, some features were created. However, these variables improved accuracy only marginally
* Statistical Models – Logistic regression/ Neural Network / Decision tree, all have been performed on the processed data to find out key drivers of the churn analysis
* Model Comparison – **Regression 3** and **Decision** **Tree** were finally selected as the best models to consider drivers of churn and both the models highlighted the following key drivers of customer churn - **# of months visited by customers to Store , # of Nonfood visits by customer in year, Promotional sales $ value in Quarter 1 by customer , % change in trips to store by customer in Q2 over Q1 and breadth of purchase across categories in Q2 over Q1 , Bounce rate (churn after acquisition) increase in first 2 months and Visits in Q4(Festival time).**

# Data Pre-processing

## Missing value treatment

#### StatExplore observation:

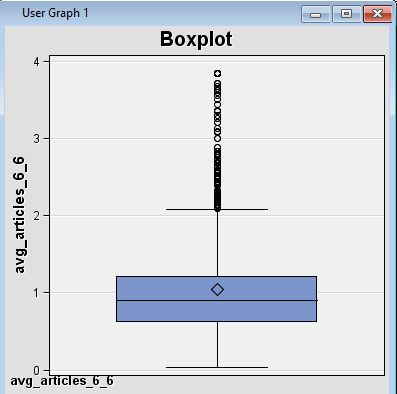
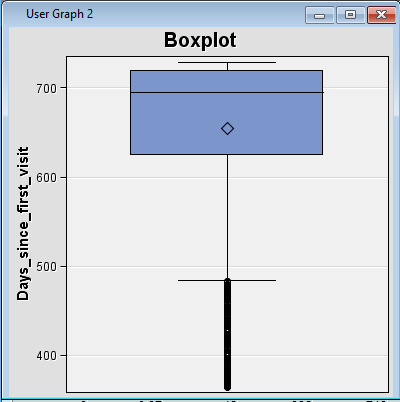


#### Inference:

The dataset has no missing values. So, labelling, imputing or dropping missing value records is not applicable to this dataset.

## Outlier detection

#### Graph Explore node observation:

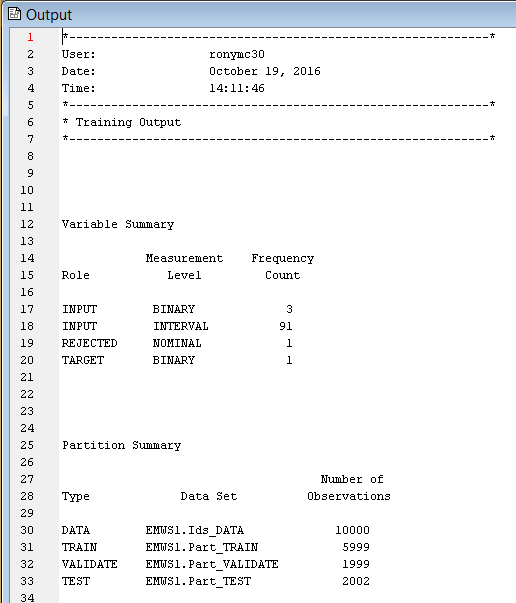


#### Inference:

The boxplots revealed outliers in most of the continuous variables. These were subsequently treated by capping the variables. Capping on the upper limit was done by replacing the variables with the 99 percentile value and on the lower limit with the 1 percentile value of the distribution.

## Data Partition

#### Data Partition node observation:



#### Inference:

We now proceed to partition our data into 3 partitions – TRAIN, VALIDATION & TEST. We reviewed parameters for both 50:25:25 and 60:20:20 partitions model. Finally, we selected the 60:20:20 mix as we observed a slightly better sampling mix.

## A Priori statistics

Examinations of the raw data reveals that currently 67.8% of customers will continue to shop with METRO while the remaining 32.2% will eventually churn. Therefore, the goal of our analysis will be to increase the accuracy to a value greater than 67.8%.

## Logistic Regression (on raw data)

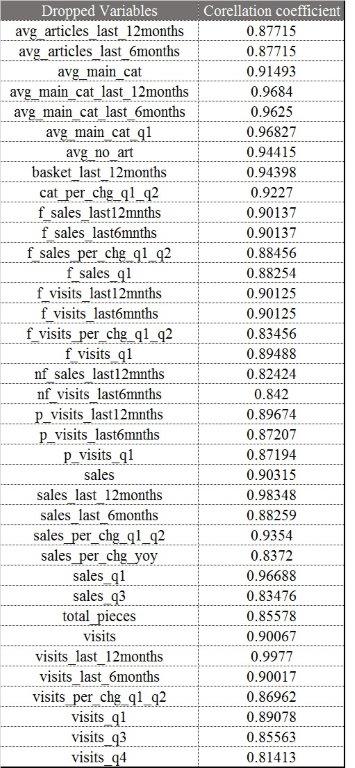
We perform logistic regression on the raw data (without removing any variables, without doing a multicollinearity check or feature selection) to study the present accuracy and precision of the data. This helps us to visualize the current accuracy and precision of the data and serve as a baseline to compare our improved model against. Precision and Accuracy on the raw dataset were computed and it has shown negligible change over A Priori Statistics. Our objective now is to increase the precision and accuracy of the model further.

We now proceed with data preprocessing by performing multicollinearity check and variable selection on the current data.

## Multicollinearity Check

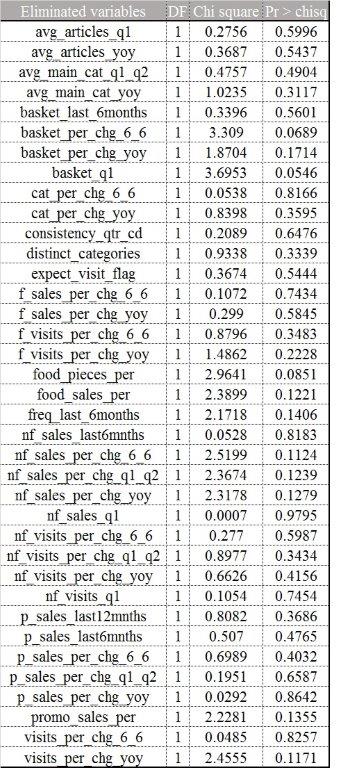
Highly correlated variables have a skewing effect on prediction output. So, we proceed to perform a multicollinearity check and carefully analyze which correlated variables can be removed.

Keeping the Pearson correlation threshold at 80%, we shortlist those variables that have a correlation coefficient above this figure for elimination. Finally, after excluding the **business-critical variables from the elimination shortlist**, the following 37 variables were eliminated:



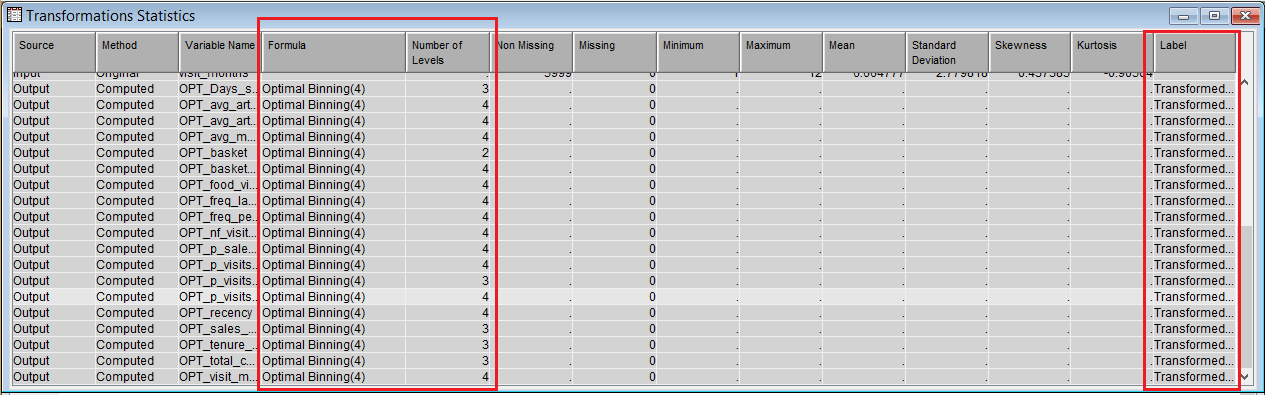
## Variable selection Method

To further narrow down the variables of interest, we proceed with variable selection. We use **backward logistic regression with Akaike Information criteria to maximize likelihood and significance value** **threshold** to identify the number of variables which affect the target variable significantly. Subsequently, we observed 37 variables are insignificant in predicting the churn rate. These variables are:



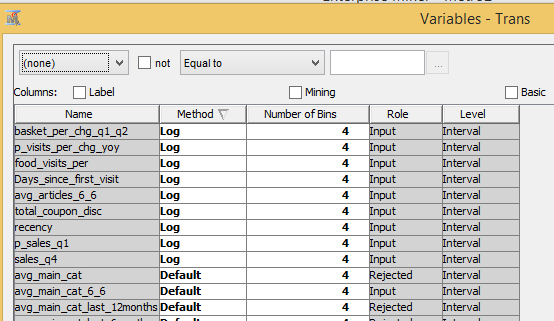
## Data Transformation

### Transform Variable node observation with optimal binning:



In the data transformation step, we use the **Transform Variable node (Optimal Binning method)** to arrive at the optimum number of bins required for our significant continuous variables. This enables us to run the data modelling steps on our model and obtain improved accuracy and precision. In the transform variable node, we have chosen number of bins to be 4 for the selected variables.

### Log transform

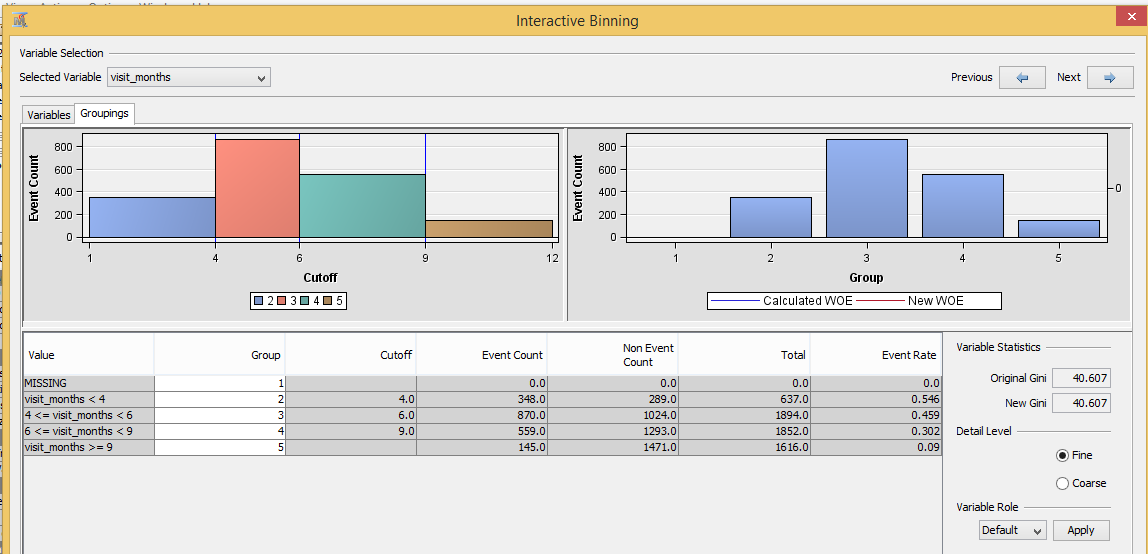


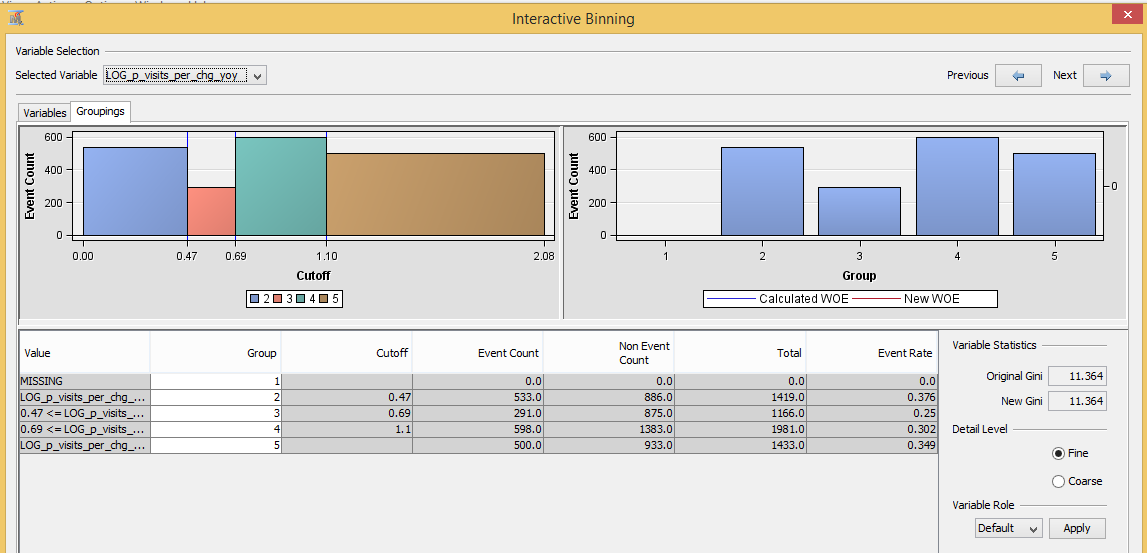
The purpose of Log transformation is to get the response variables into a nearly symmetric distribution. Some of the variables show higher spread of the dimension. A logarithmic transformation simplifies the complexity of variable spread.

### Interactive binning



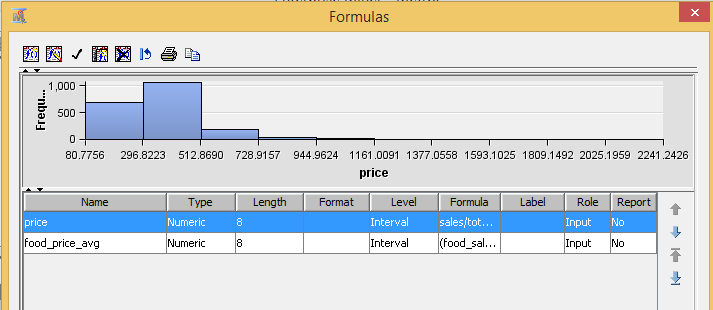
First, we created initial bins by choosing default values of interactive binning. Then, in some cases, we separated classes in better groups if Weight of Evidence (WOE) of dependent variable improves in explanatory variables. In other cases, we joined neighboring groups that exhibited similar homogeneous distribution of WOE of dependent variable





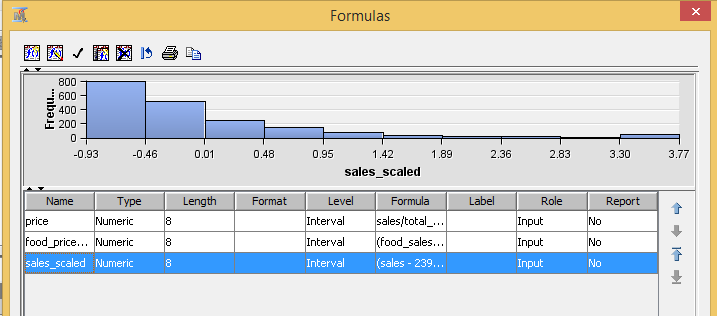
### Feature Creation

One of the limitation of this data is that it has just one dimension - aggregate transactional values at customer level. Other dimensions like psychographic variables (price point preferences, store trip pattern, tender type preference etc.) and demographic variables (Age, Income, Ethnicity, Gender, Household size etc.) are missing. Therefore, we derived features that might lead us to infer information related to purchase preference. **Average price point preference overall** and **Average price point of customer at food department level** are a couple of examples of derived features. However, other features, like the variability of price point preferences over time, which are very important causes of customer churn in India, cannot be determined with the given data.



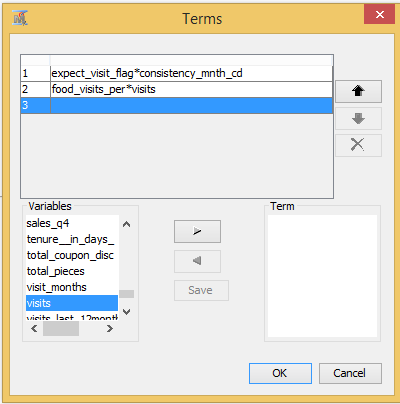
### Variable Standardization

As Metro Cash and Carry is a wholesale reatiler, it enjoys diverse array of customers. Some customers are the individuals who are coming to shop for monthly grocery while others are small scale corner shop owners making high volume store stocking purchase. Therefore, the absolute values of some variables are highly skewed. Neither logarthmic transformation nor binning of variables helped achieve variable significance. So, attempted rescaling the variables by standardizing those variables.



### Interaction variable

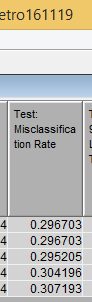
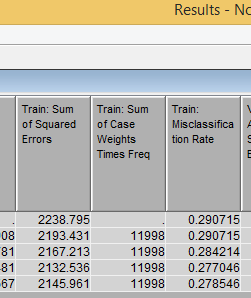
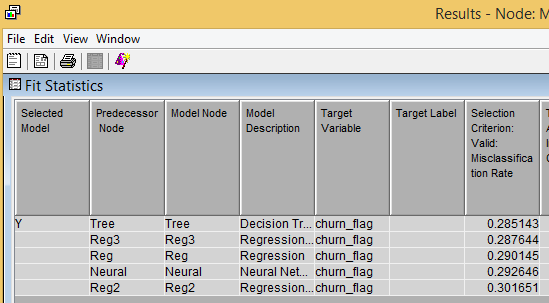
Despite these efforts, the accuracy of Model did not help the churn model. We hypothesized that relationships among the variables could describe the churning behavior of customer. So we created interaction variables that could help the interaction between variable. For example, Expect\_visit\_flag and consistency\_month\_cd of a customer were considered together and an interaction flag variable was created for the same.



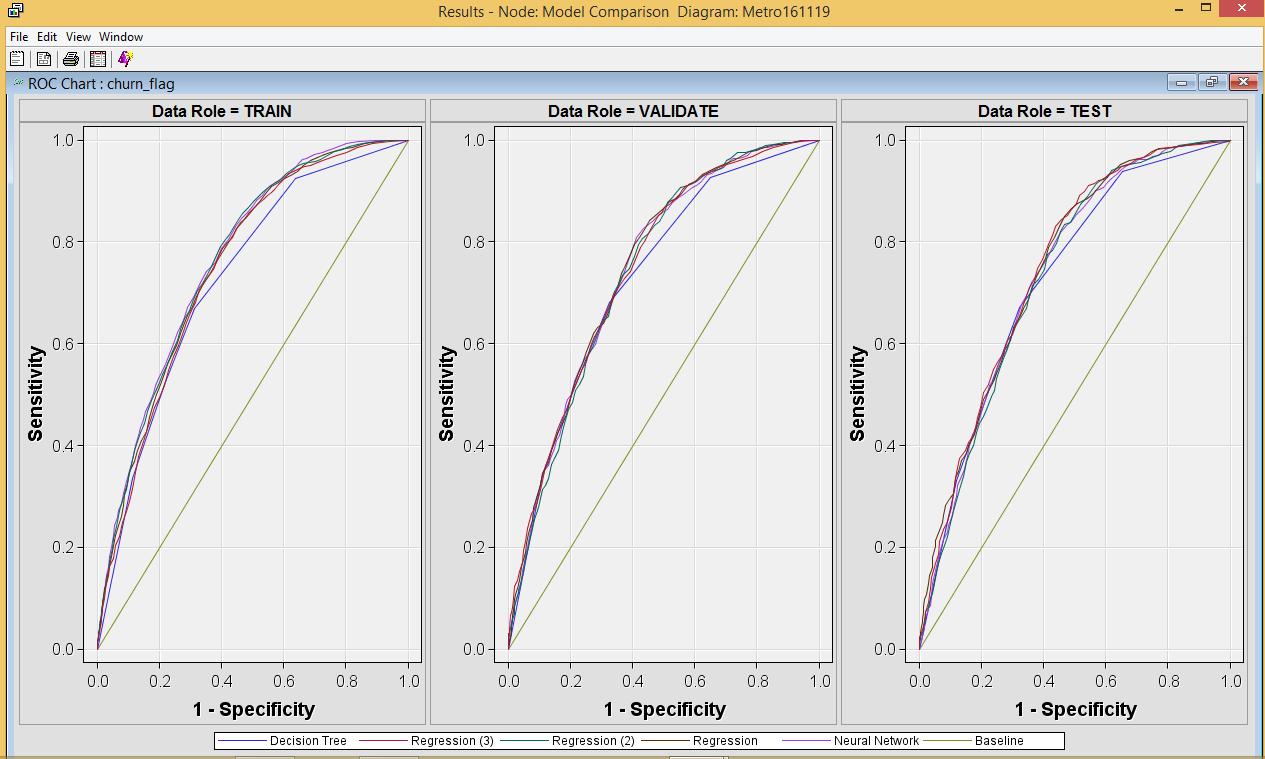
# Data Modelling

## Model Comparison

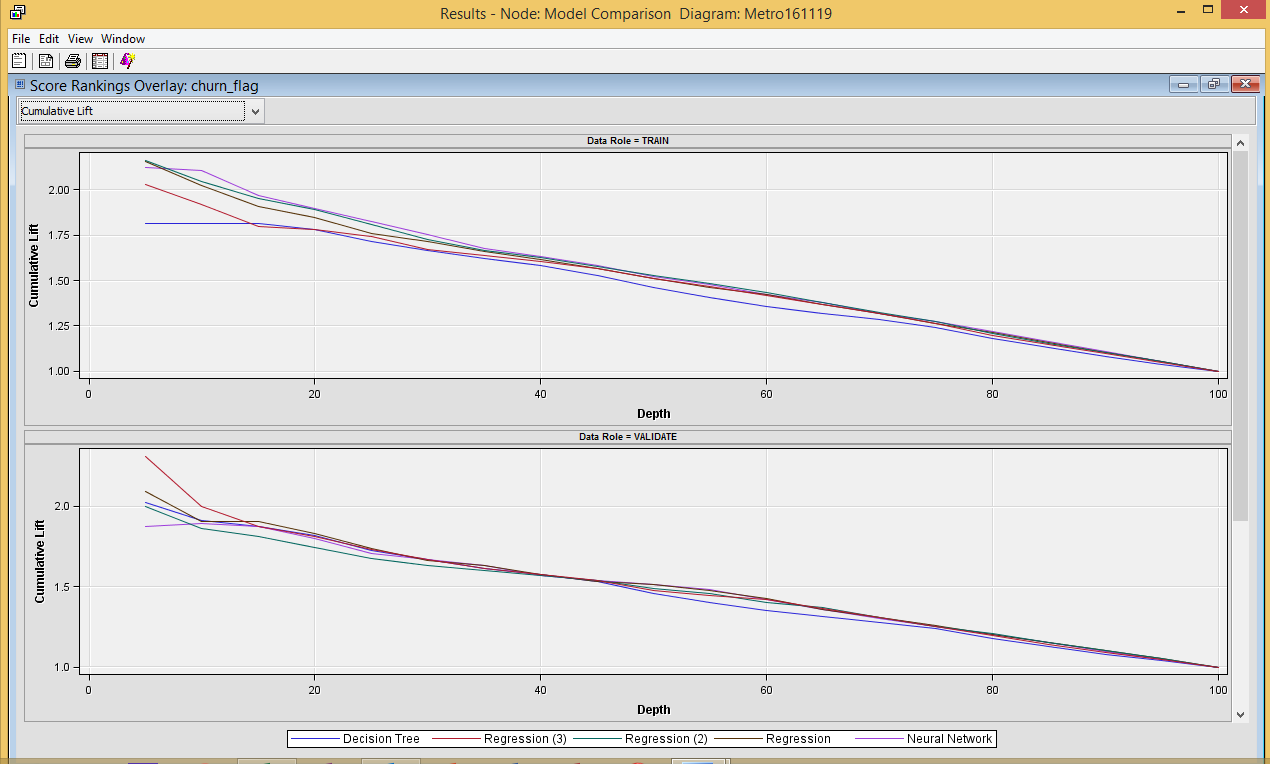
At this stage, we applied multiple models to the preprocessed and transformed data. The model output parameters were compared and analyzed. Of all the models, the Decision Tree, Regression(3) and Neural Network exhibited the best metrics.



Misclassification Rate of the Training and Validation datasets



ROC Charts for the Training, Validation and Test datasets

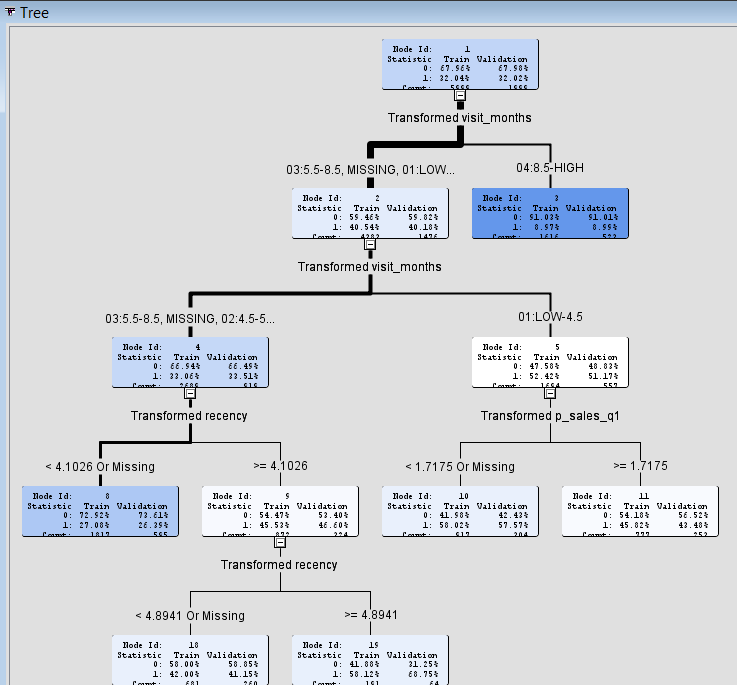


Cumulative lift curve of training and validation model

### Decision Tree

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Training Model Prediction** | |
| Churn | Retain |
| **Actual** | Churn | 643 | 1279 |
| Retain | 465 | 3612 |
|  | | **Validation Model Prediction** | |
| Churn | Retain |
| **Actual** | Churn | 219 | 421 |
| Retain | 149 | 1210 |

Confusion Matrix

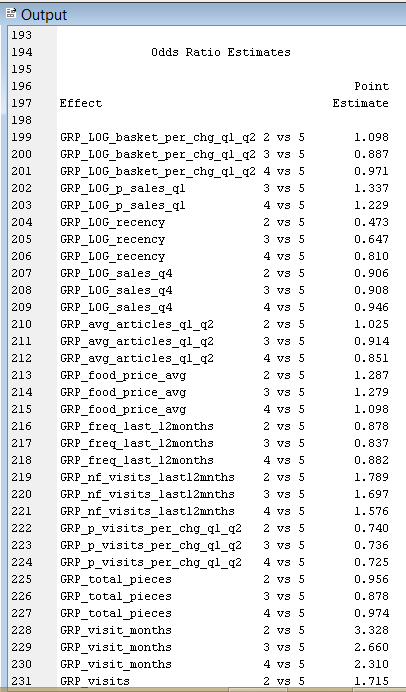


Decision Tree

### Regression(3)

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Training Model Prediction** | |
| Churn | Retain |
| **Actual** | Churn | 728 | 1194 |
| Retain | 550 | 3527 |
|  | | **Validation Model Prediction** | |
| Churn | Retain |
| **Actual** | Churn | 240 | 400 |
| Retain | 175 | 1184 |

Confusion Matrix



Odds Ratio Estimates

### Neural network

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Training Model Prediction** | |
| Churn | Retain |
| **Actual** | Churn | 769 | 1153 |
| Retain | 509 | 3568 |
|  | | **Validation Model Prediction** | |
| Churn | Retain |
| **Actual** | Churn | 240 | 400 |
| Retain | 185 | 1174 |

Confusion Matrix

# Conclusion

Going by the validation misclassification rate, the Decision Tree proved to be the best among all the models, followed closely behind by Regression(3). Going by the training misclassification rate, Neural Network model performed better than the rest.

As our main objective is to determine the key drivers that are responsible for customer churn, we selected the **Regression(3)** model as the best model for our project. Also, the variance in misclassification rate among the validation, training and test datasets was minimal for **Regression(3)**.



# Recommendations



* No. of months visited by a customer across a year if less than **9 months** then there are **130%** higher chances that customer can churn and if **visit\_month < 4** then chances of churn increases to **230%.**
* If **nonfood visits** are **less than 8** in a year, then there **58% more chance that the customer will churn.**
* If customer’s food department purchase is less than 55k in a year, then there is ~ **30% chances that a customer will leave.**
* An average metro **customer visits almost 33 times to store in a year**, If somebody visits **1/3rd of average visits then he is 72% more likely to chur**n and if someone visits **½ of average visits** in a year then he is **40% more likely to churn.**
* Normal customer buys almost **4000 Rs. promotional products in Q1**, If someone buys **less than 800 Rs promotional products in Q1. He is 33% more likely** to churn.
* If **percentage of increase in customer visit is more than 33% between Q1 and Q2** then **25% more likelihood of person staying** with Metro. If Breadth of articles (# of articles across categories) increase in q2 comparing to q1 then 15% more likelihood to that customer will stay with Metro.
* If somebody visits **in Q4 at least 1 time, he is 25% more likely to stay with Metro**.
* After joining Metro, in first 10 days of a customer shows **50% more probability that he will stay with Metro** but after **2 more months the chances of staying as a loyal member decreases by 30%**.

# Plan of Actions for Metro

As it is clear from above recommendations that Metro should focus about certain important factors like No of months visited by customer to store (Repeat rate approximations), first 2 months of acquired customer, Quarter 2 and Quarter 4 purchase pattern etc. Following are certain action items which can help Metro to retain customers:

* To increase number of months visited by a potential churn customer, Metro should provide promotional coupons to redeem within a certain no of weeks in coming future, It will help to increase customer’s repeat rate to store and chances of churn will decrease.
* Target Newly acquired customers by providing Non-food departments offer. As New Customers show a decrease in likelihood of staying with Metro in first 2 months and customer engagement across nonfood departments also helps customers to increase retention. Both combined will help customer to continue with Metro.
* Promotional offers should be planned in Q1 (Like 10% discount on purchase above 4000 Rs.). It will help Customers to buy more than 4000 Rs in Q1 which came as prominent indicator for customer retention
* Customers who visit frequently to store and make a purchase only in Food department, should be given coupons to redeem Non-Food department. It will help those customers to expose to Non-Food departments of Metro and Non-Food Visits showed an important indicator for Customer retention
* As increase in no of articles in Q2 helps in retention, more options to purchase across categories will help these customers to buy more or bundling of products with a promotional price point will help customers to buy more no of articles
* Quarter 4 promotional sales period can be better organized as visit during Q4 helps customers to be loyal to Metro