**MGS616 – Predictive Analytics**

**Project**

**Critical Analysis of Factors Affecting the Sales of Retail Stores in Seoul**

**Submitted By,**

**Apoorva John**

**Neeraja Narayan**

**Preethy Thomas**

**Yashaswini Huskur Lingarajaih**

**CONTENTS**

[**SEOUL RETAIL CASE PROBLEM STATEMENT** 3](#_Toc482319603)

[**THE DATA STORY** 5](#_Toc482319604)

[**STORE A** 5](#_Toc482319605)

# **THE MODELLING LIFE CYCLE**

## **DISCOVERY: SEOUL RETAIL CASE PROBLEM STATEMENT**

*Company Q has five franchise stores (A through E) located in an exclusive shopping district in Seoul, South Korea. Each store sells the same brands and items. As per guesstimates, roughly 95% of their sales come from Japanese tourists. Even though the stores bear the name of the same company, they are effectively competitors due to their close proximity to each other. The franchisee of Store B, Mr. Choe, is interested in identifying the factors that affect sales in his store and the extent to which these factors affect sales. He is also interested in identifying the similarities and dissimilarities of these impact factors across the five stores. Mr. Choe plans to use these identified factors for sales and operations planning in his store.*

*To help answer his questions, Mr. Choe has collected aggregated sales data on all four stores from September 1 2011 to March 16 2013. Here are the variables which are captured in the dataset:*

*Store ID*

*Store Name*

*Number of Customers*

*Number of Items Sold*

*Total Sales*

*Discount: This is the subset of Total Sales and reflects the proceeds from any items sold at a discount.*

*Average Sales per Customer*

*Average Sales per Item*

*Date*

*Day of week (weekends are thought to be better for business than weekdays)*

*Distances from Metro Stations X and Y (in meters and feet)*

*Distance from the nearest main thoroughfare (in meters and feet)*

*In addition, Mr. Choe was also able to collect data on the total number of Japanese tourists that visited Seoul. However, this data is available only from Feb 14 2012. Further, note that not all of these tourists might have visited Company Q’s retail stores. There might, however, be some correlation between the number of such tourists and some of the variables included in the dataset. Also, there may be differences across stores – for instance, the number of Japanese tourists may have a larger bearing on Store A’s revenues than on Store B’s.*

*Further, since most of the customers are Japanese, the currency exchange rates may affect their buying habits. The JPY (Japanese Yen)/KRW (Korean Won) ratio is provided for each day in the dataset.*

*Also, such tourists may be more likely to visit during Japanese national holidays. This effect may be more heightened if the holiday happens to fall on a Friday or a Monday, as it then implies an extended weekend vacation. The binary variable Holiday is set to 1 if the day is part of a holiday weekend such that a Japanese holiday falls on either a Friday or a Monday; otherwise, it is set to zero. For example, if there was a holiday on Friday March 2, 2012, then the binary variable for March 2, 3 and 4 is set to 1. Likewise, if the holiday falls on Monday March 5, the binary variable for March 3, 4 and 5 is set to 1.*

*Weather may also play a role in the shopping behavior. People may be more likely to shop and spend under certain weather conditions than others. For this purpose, the following data were collected: Actual High Temperature and Outlook (whether it was sunny, cloudy, rainy, snowy, etc).*

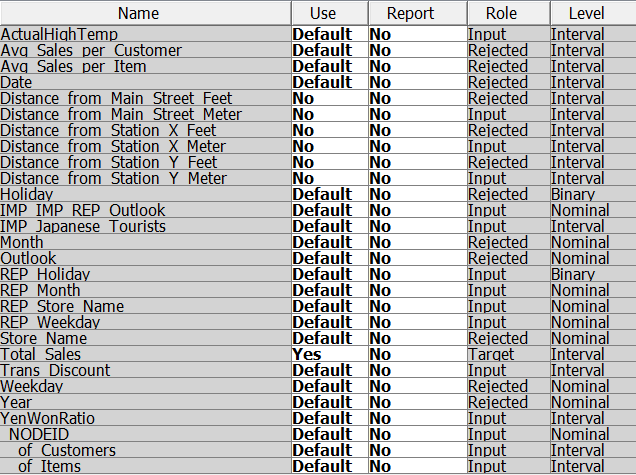
*Store A closed on March 17 2012 and reopened at the same location under new ownership on June 19 2012.*

*Store C closed on March 17 2012 and reopened at a different location under new ownership on September 2 2012.*

*Mr. Choe would like to know more about the factors that affect sales in each store and the importance of such factors. In writing your report to Mr. Choe, assume that he is a non-technical person and communicate your findings and analysis accordingly.*

## **DATA PREPROCESSING**

* Imported the file externally and checked the variables imported by checking the first 30 rows matching with the csv file.
* Had a discussion with the team and decided on variables to go ahead.



* As total sales were of Choe’s interest, made it has target variable
* Rejected the Average Sales per Customer and Average Sales per Item as Number of customer and Number of items column were upholding the same variables
* Rejected the Date and Year variable as we have months and days specified to see the sales effect
* Distance in meters were considered and rejected the feet variables
* Imputed the Japanese Tourist data with distribution to fill the missing data
* For the outlook column replaced the NA columns with the mode of the value of that month

## **MODEL PREPARATION AND BUILDING**

* Split the data into Store A, B, C, D and E using filter
* Applied the 70% to train the model and 30% to test the model
* Considered Distance as variable only for store C, as store C shifted the place. Rejected for Store A,B,D and E as the distance is unique
* For identifying identical variables, built the decision tree and selected the common variable among all stores
* After rejecting the common variable built 2 way and 4 way decision tree. Compared the model with mean square error. Picked the best model to see the dissimilar effect of captured attributes on each store, which helped us to give insight of each stores factors
* Built the Regression model of Stepwise, Forward and Backward. Selected the best fit model to tell the magnitude of significant attributes effecting the stores high and low sales
* Built the model package after building the required model.

# **THE DATA STORY**

## **STORE A**

Predominant factors that affect the sales of store A are

* Number of customers
* Number of Items
* Yen Won Ratio and
* Japanese Tourists

Maximum sales is achieved in the store for the following combination of factors:

* Number of Customers is greater than 191 and
* Number of Items is greater than 2221.

Least amount of sales occurs in the store when

* Number of customers is less than 191
* Number of Japanese tourists is less than 2559 and
* Yen Won ratio is less than 12.17.

Other than the obvious factors of number of items and number of customers improving sales, it can be observed that the sales of store shows a considerable increase when the Yen Won ratio is greater than 13.4 and the number of Japanese tourists is greater than 3047 at any point of time.

Studying the extent at which the predominant factors affect the sales of store A, it can be understood that:

* For every unit increase in Yen Won ratio, the sales of the store is expected to increase by 751260 Won (₩).
* For every single new customer visiting the store, the sales of the store is expected to increase by 45489 Won (₩).
* For every unit increase in number of sold items, the sales of the store is expected to increase by 3033 Won (₩).

Besides these predominant factors, the sales of the store shows a considerable appreciation during the months of July and August. Also, the sales shows a considerable improvement during the weekends. The extent at which these factors affect the sales are as follows:

* Every July, the sales of the store is expected to increase by 3280885 Won (₩).
* Every August, the sales of store is expected to increase by 4630474 Won (₩).
* Every Saturday, the sales of the store is expected to increase by 2060722 Won (₩).
* Every Sunday, the sales of the store is expected to increase by 1009379 Won (₩).

The factors which have a detrimental effect on the sales of store A are increase in actual temperature, and discount. Also, the sales of the store shows a decreasing trend during the months of February and March. The extent to which these factors affect the sales are as follows:

* For every unit increase in the actual temperature, the sales of the store is expected to decrease by 39307 Won (₩).
* Every February, the sales of the store is expected to decrease by 3374326 Won (₩).
* Every March, the sales of the store is expected to decrease by 2341273 Won (₩).
* For every unit increase in discount, the sales of the store is expected to decrease by 0.63 Won (₩).

## **STORE B**

Predominant factors that affect the sales of store B are

* Number of customers
* Number of Items

Maximum sales is achieved in the store for the following combination of factors:

* Number of Items is greater than 1340 and
* Number of Customers is greater than 170

Least amount of sales occurs in the store when

* Number of customers is less than 87 and
* Number of items is less than 459

Other than the obvious factors of number of items and number of customers improving sales, it can be observed that the sales of store shows a considerable increase when the Yen Won ratio is greater than 13.4, the day is a Friday, Saturday or Sunday and the number of Japanese tourists is greater than 3241 at any point of time.

Studying the extent at which the predominant factors affect the sales of store B, it can be understood that:

* For every single new customer visiting the store, the sales of the store is expected to increase by 32113 Won (₩).
* For every unit increase in number of sold items, the sales of the store is expected to increase by 5199 Won (₩).

Besides these predominant factors, the sales of the store shows a considerable appreciation during the months of May, July, August and October. Also, the sales shows a considerable improvement during Sundays. The extent at which these factors affect the sales are as follows:

* Every May, the sales of the store is expected to increase by 1365583 Won (₩).
* Every July, the sales of the store is expected to increase by 1050123 Won (₩).
* Every August, the sales of store is expected to increase by 1323485 Won (₩).
* Every October, the sales of store is expected to increase by 668937 Won (₩).
* Every Sunday, the sales of the store is expected to increase by 752656 Won (₩).

The factors which have a detrimental effect on the sales of store A is discount. Also, the sales of the store shows a decreasing trend during the months of February, March and September. The extent to which these factors affect the sales are as follows:

* Every February, the sales of the store is expected to decrease by 1191997 Won (₩).
* Every March, the sales of the store is expected to decrease by 1317982 Won (₩).
* Every September, the sales of the store is expected to decrease by 1233316 Won (₩).
* For every unit increase in discount, the sales of the store is expected to decrease by 0.9771 Won (₩).

## **STORE C**

* Based on the analysis, it can be observed that the decision to change the location for Store C was not a strategically wise decision. It was observed that the average sales were comparatively higher by an average of 3 million Won in a span of 1.5 years before the store relocation.
* Also, it was observed that the average sales were maximum at the previous location in the month of March, October and November at an average of 11million Won (₩).
* After the store relocation, we observe a maximum average sale of only 8 billion Won (₩). This average sale is observed only when the average total discount the store offers on a day is greater than 8100 Won (₩).
* The average sales of Store C is the lowest at around 4 million Won (₩) when the average number of Japanese tourist in Seoul is less than 2102. We could conclude that most of the sales come from these tourists at Store C who are on the lookout for certain items on sale.
* Based on further analysis, it was observed that the following factors play a major role in predicting the sales for Store C and are the predominant factors.
  + **Month** – The calendar month in which the sales data are being recorded
  + **Discount** – The total discount amount given for each recorded day
  + **YenWonRatio** – Conversion rate
  + **Number of Customers** – The total number of customers visiting the store on a particular day.
  + **Number of Items** - The total number of items sold in the store on a particular day.
* The store sales is observed to decrease by 0.4679 Won (₩) for every unit increase in the discount value. The store manager would want to look at the products which are discounted.
* The sales at Store C is seen to have an increase in sales by 355,758 Won (₩) for every unit increase in the Yen/Won ratio.
* The store sales is observed to increase by 46419.3 Won (₩) for every increase in the number of customer and by 3,311.7 Won (₩) for every increase in the number of items sold.
* Also, it is observed that store sales in the month of February is expected to decrease by 546,064 Won whereas an increase of 768,023 Won (₩) is observed in the month of November.
* The most important observation here is that Store C is observed to be in a loss of around 5 million Won (₩) when other mentioned factors are not taken into consideration. It is extremely important for Store C to increase the number of customers visiting the store or the number of items being sold to expect a positive increase in the total sales.

As observed in the analysis, relocating Store C closer to Main Street and in the process closer to Store B has been disastrous. This could probably be attributed to the fact that two stores are so closely located and hence reducing the customer base for Store C. This relocation doesn’t seem to impact the sales of Store B though.

## **STORE D**

* It can be observed that the YenWonRatio is one of the most important predictors of the sales for   
  the Store D. When the Japanese Yen to Korean Won ratio is greater than or equal to 13.11, the sales seem to be doing approximately 1.67 times better in average sales compared to when the ratio is less than 13.11. This could be due to the large percentage of Japanese tourists shopping from the chain of retail stores and an increase in the Yen rate would definitely help boost sales.
* The other predictor variable that seems to impact sales is the different days of the week. In general, Store D seems to be generating more sales on weekends (Saturday and Sunday) than on weekdays. On weekends, store D generates an average sales of 4 million Won more than what is generated on weekdays.
* From the analysis, it can be observed that the maximum average sales is generated on weekends when the YenWonRatio is greater than or equal to 13.11 and the total discount given on items is more than or equal to 27900. The average sales generated on such days equals a value of more than 17 million.
* It can also be observed that on weekdays excluding Monday when the YenWonRatio is less than 13.11, store D seems to be performing considerably poorly with a comparatively low average sale of 6 million Won (₩).
* Based on further analysis, it was observed that the following variables help in predicting the sales of Store D:
* **Outlook –** The weather condition on the particular day when the sales data is being recorded
  + **Month** – The calendar month in which the sales data are being recorded
* **Weekday –** The day of the week when the sales data is being recorded
* **Discount** – The total discount amount given for each recorded day
* **Number of Customers** – The total number of customers visiting the store on a particular day.
* **Number of Items** - The total number of items sold in the store on a particular day.
* The store sales are observed to be impacted by the discounts given by the store. For every 1 unit increase in the discount value, the sales is seen to decrease by 0.74 Won (₩).
* We also observe that for every single new customer visiting the store, the store sales is seen to increase by 40181.7 Won (₩).
* The other analysis that can be reported is that for every new item that is sold by the store, the sales is seen to increase by 4327.4 Won (₩).
* Month and day of the week that a transaction takes place is also an important variable in predicting the sales. In July and August, the sales shows a positive trend, whereas in February, March and September the sales shows a negative trend. Similarly, weekends (Saturday and Sunday) is expected to boost sales whereas, Tuesday and Thursday the sales is observed to decrease. The extent to which the month and weekday variable impacts sales is summarized as below:
* Every February, sales decreases by 1292793 Won (₩)
* Every March, sales decreases by 1091557 Won (₩)
* Every September, sales decreases by 1217134 Won (₩)
* Every July, sales increases by 1507779 Won (₩)
* Every August, sales increases by 1352658 Won (₩)
* Every Saturday of the week, sales increases by 823725 Won (₩)
* Every Sunday of the week, sales increases by 514914 Won (₩)
* Every Tuesday of the week, sales decreases by 609178 Won (₩)
* Every Thursday of the week, sales decreases by 520121 Won (₩)

## **STORE E**

Predominant factors that affect the store E sales are:

* Discount offered
* Japanese tourist
* Month
* Holiday
* Yen Won ratio
* Customers number
* Item number

Store E has **more sales** effect due to discount offered irrespective of any Japanese tourist visiting the store to justify the statement, Store E has 2 million Won (₩) average sales for the discount offered above 5650 won (₩)

Store E has **less sales** when discount offered is less and Japanese tourists visiting Seoul number goes down in the month of September, December, January and February. Where discount offered is less than 2487 Won (₩) with average sales of 1 million Won (₩).

* For the store E a sales of 2.3 million Won (₩) ( can be expected irrespective of any captured parameters effect. So the budget can be planned accordingly ahead of time
* Significant month of sales concern are January, February. For January the sales are going down by 300 k Won (₩) and for February the sales are going down by 450 k Won (₩).

Store has to take action of attracting the tourists by changing the selling patterns, like in Sales in June is expected to be 291 k Won (₩) which is quite good compared to other months

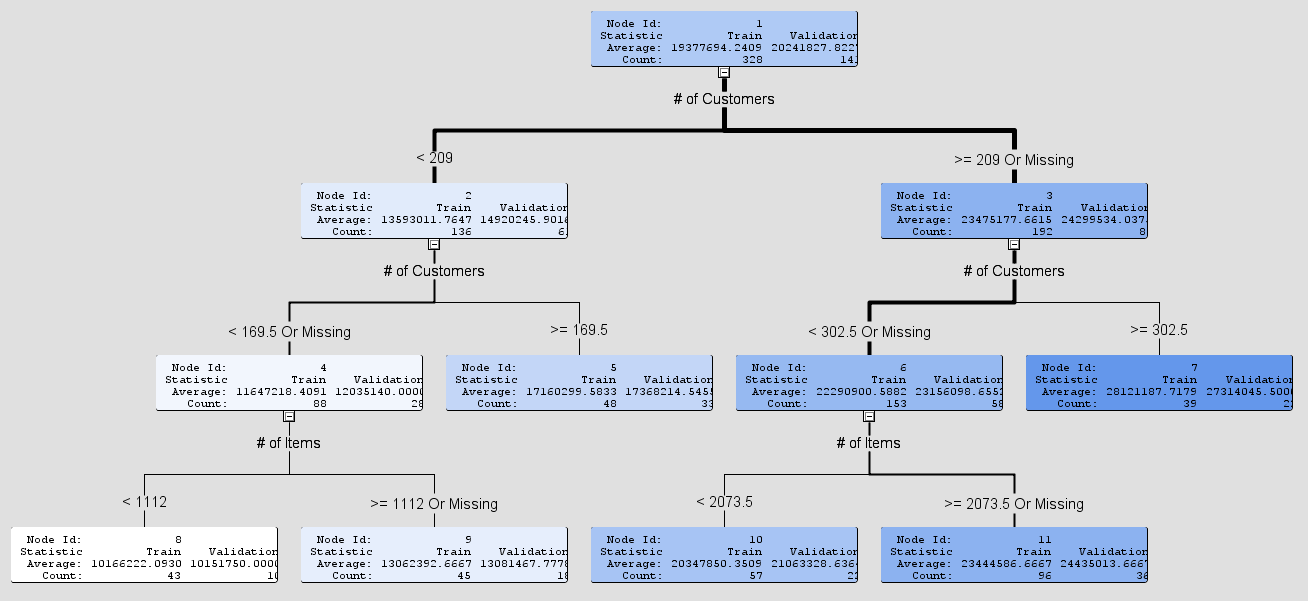
* Sales in July is expected to be 205 k Won (₩), August is expected to be 192 k Won (₩), November is expected to be 225 k. These months the store can keep the decision of selling items as it is
* For every increase in Yen Won ratio, the sales goes down by 175 k Won (₩). Need to collect the data to analyze the negative effect of sales. May be the products purchased are imported goods to attract the tourists, which is effecting the sales as Won value decreases in the international market
* On Holidays coming after and before week ends, the sales are quite impressive. Do open the store on these holidays, keep extra workers to fulfill the supply demand ratio. For every increase in customer coming to Shop E and purchasing will increase the sales by 17 k Won(₩). So if the store E really attracts the customers it can have a good sales.
* For every items increase in store, a profit of 2523 Won (₩) can be expected, so need the data of items selling in giving the forecast using time series.

# **PRIME FINDINGS**

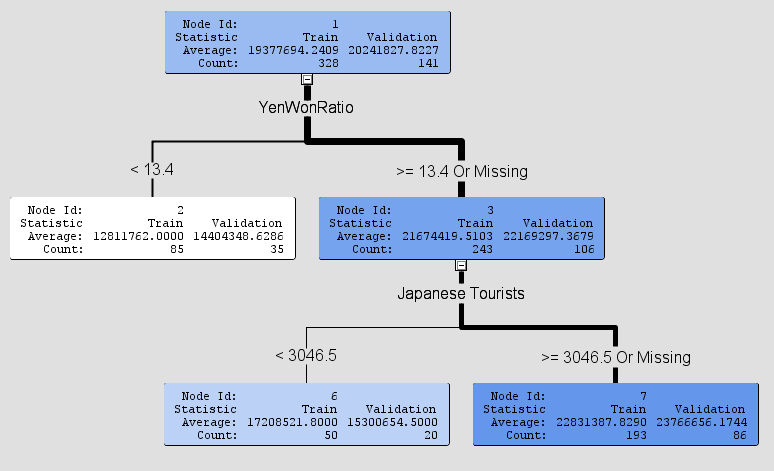
* The most important factors that contribute towards an improvement in Sales across the stores are:
* Number of Customers and
* Number of Items
* It can be seen that irrespective of other factors, the sales of store A and store A increase in the months of July and August and on Sundays.
* One of the factors that is found to increase the sales in both Store B and Store C is Yen Won ratio.
* Store B and Store D shows similar trends in increase and decrease of sales across months ie., Sales tend to decrease in both the stores in the months of February, March, and September and increase in the months of July and August. Sales also tend to increase during all Sundays.
* The sales tend to increase for both Store E and Store B during the months of July and August. Also, it could be observed that one of the factors that helps in improving the sales in both these stores is Yen Won ratio.
* It could be noted that the sales amount tend to decrease across the stores with an increase in the value of discount.

# **APPENDIX**

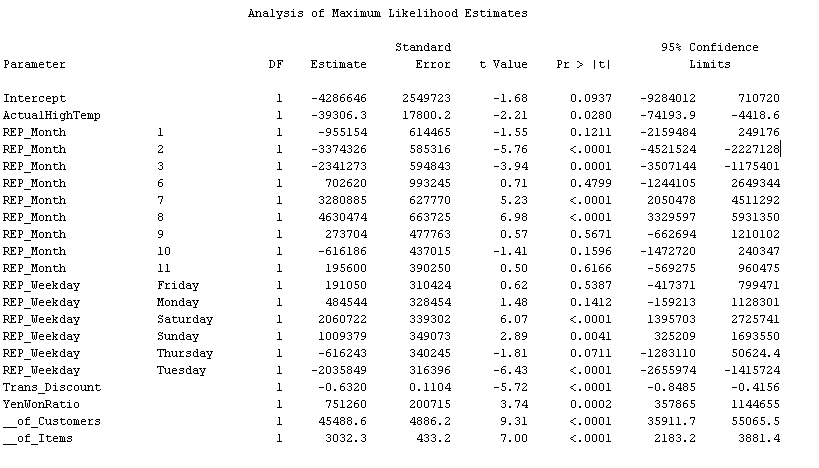
## **Store A – Overall Decision Tree**



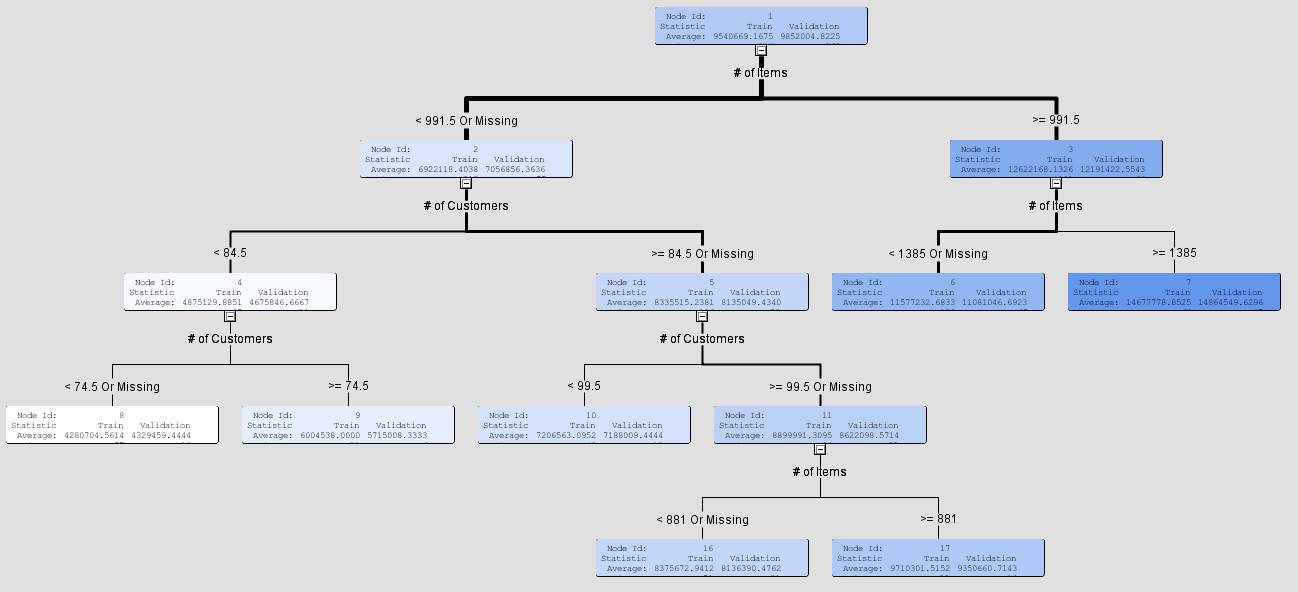
## **Store A – Decision Tree Without Number of Items and Number of Customers**



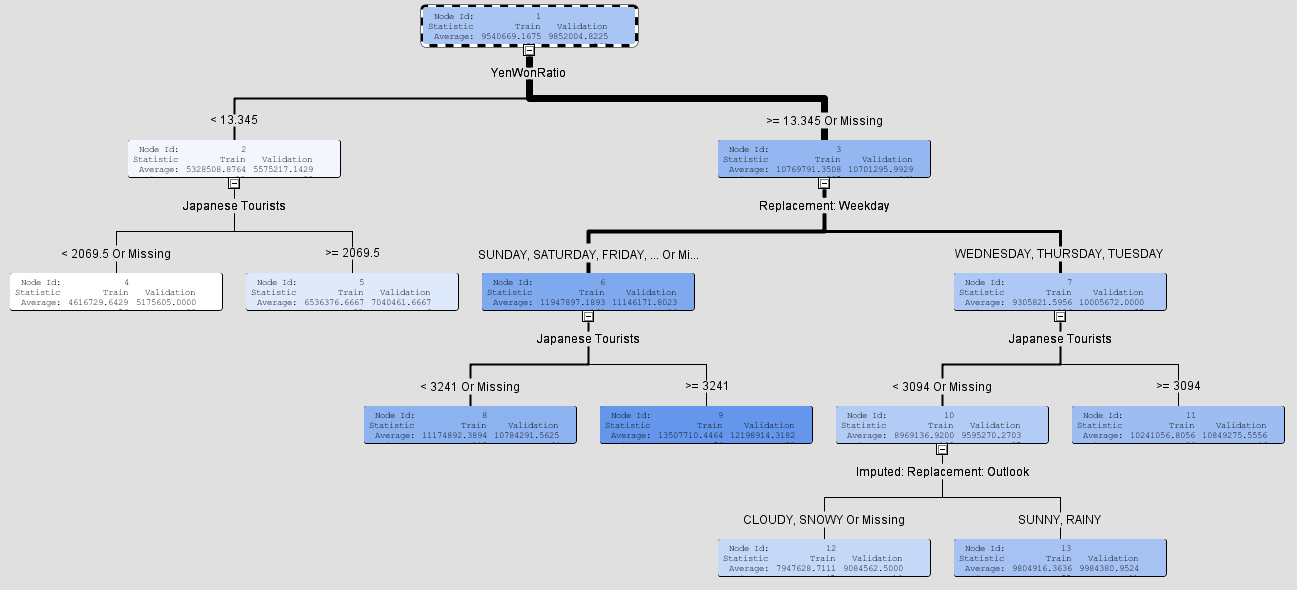
## **Store A - Stepwise Regression**



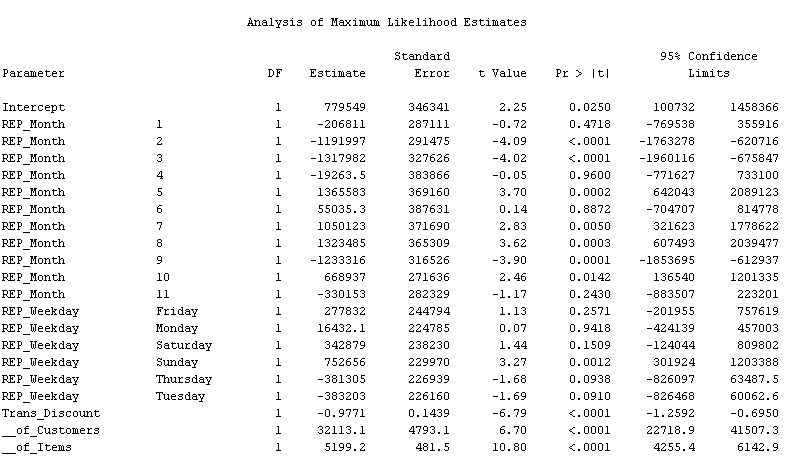
## **Store B – Overall Decision Tree**



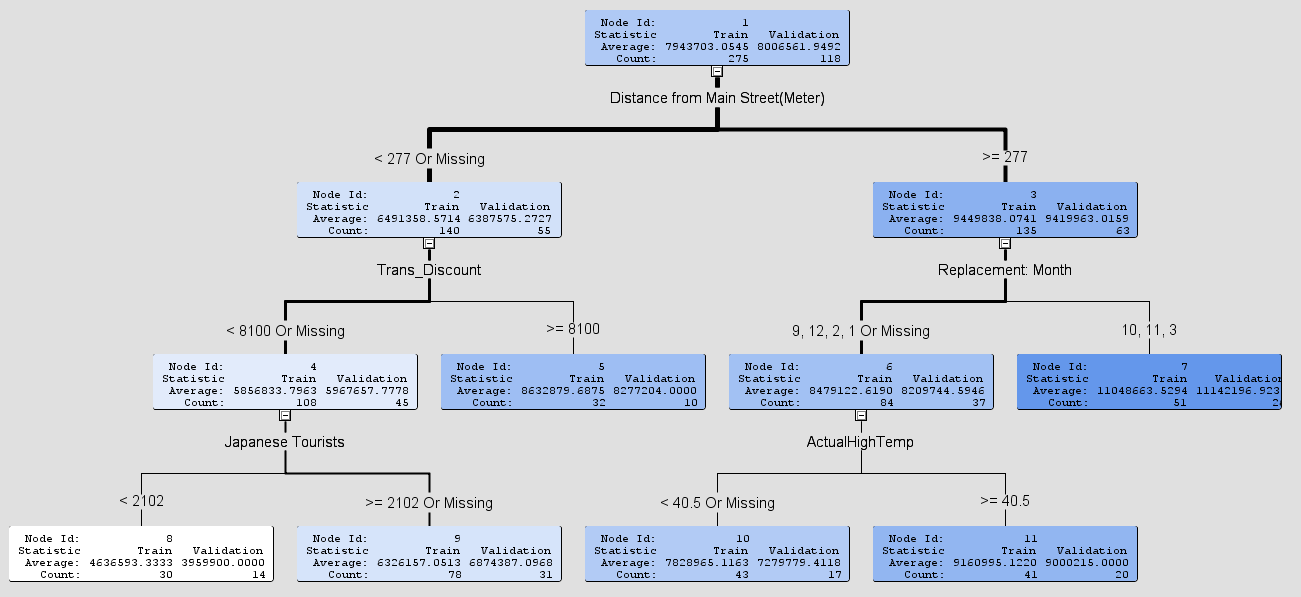
## **Store B – Decision Tree Without Number of Items and Number of Customers**



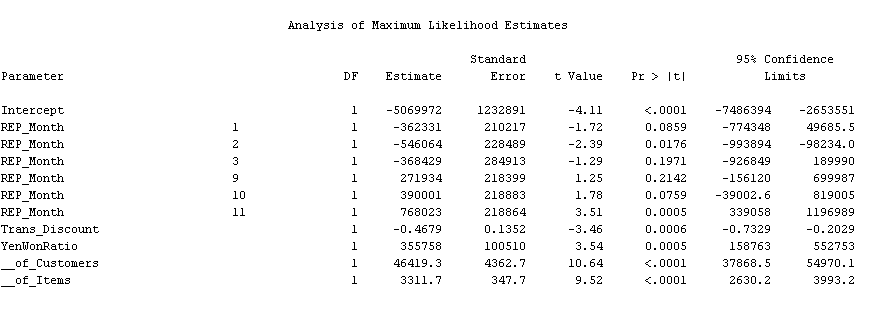
## **Store B - Stepwise Regression**



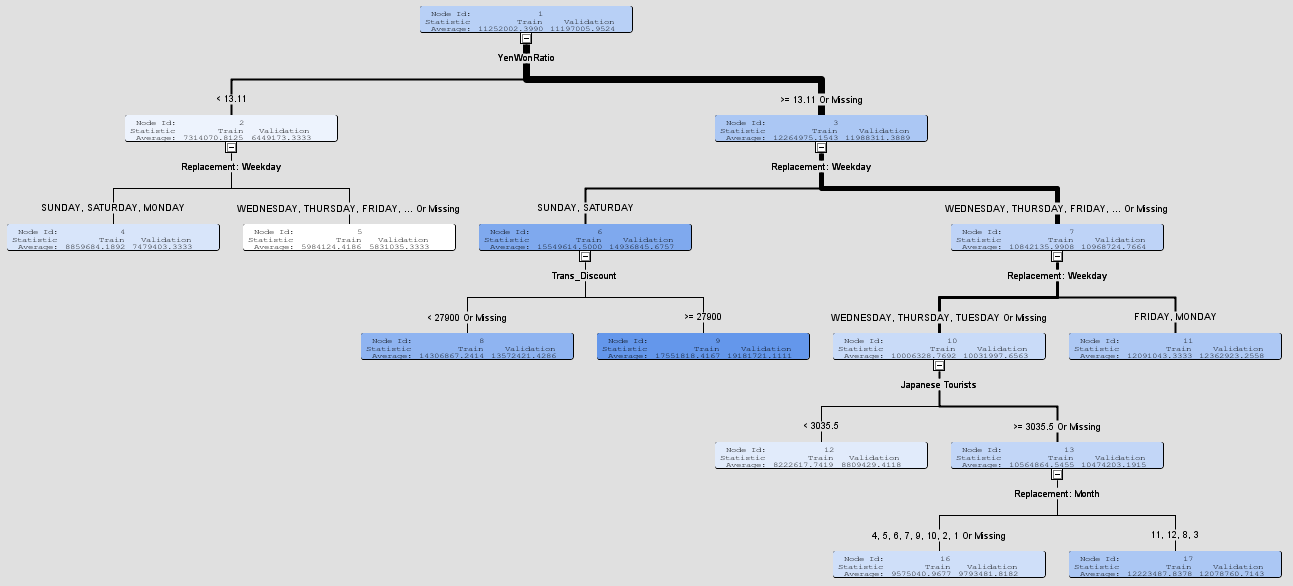
## **Store C – Decision Tree Without Number of Items and Number of Customers**



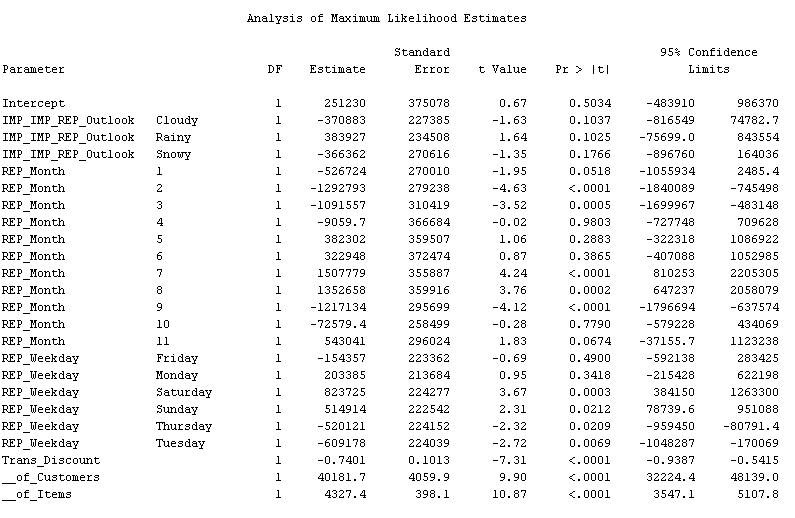
## **Store C – Stepwise Regression**



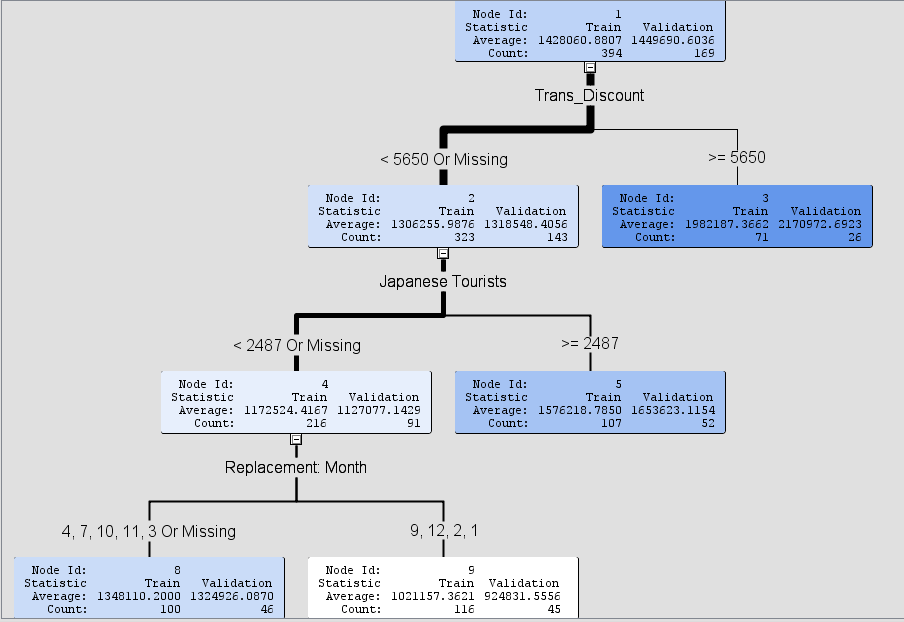
## **Store D – Decision Tree Without Number of Items and Number of Customers**



## **Store D – Stepwise Regression**



## **Store E – Decision Tree Without Number of Items and Number of Customers**



## **Store E – Stepwise Regression**

