**Approach & Result**

**Document**

**1. Opening of new stores or relocating stores:** Estimate sales that would be generated by a new location given the characteristics of the new store and location.

**2. Identify high performance stores:** Identify stores that are exceeding expectations so that their success formula can be applied to other store **3. Identify low performance stores:** Identify stores that aren't performing as well as expected and take

appropriate decisions including closing them down

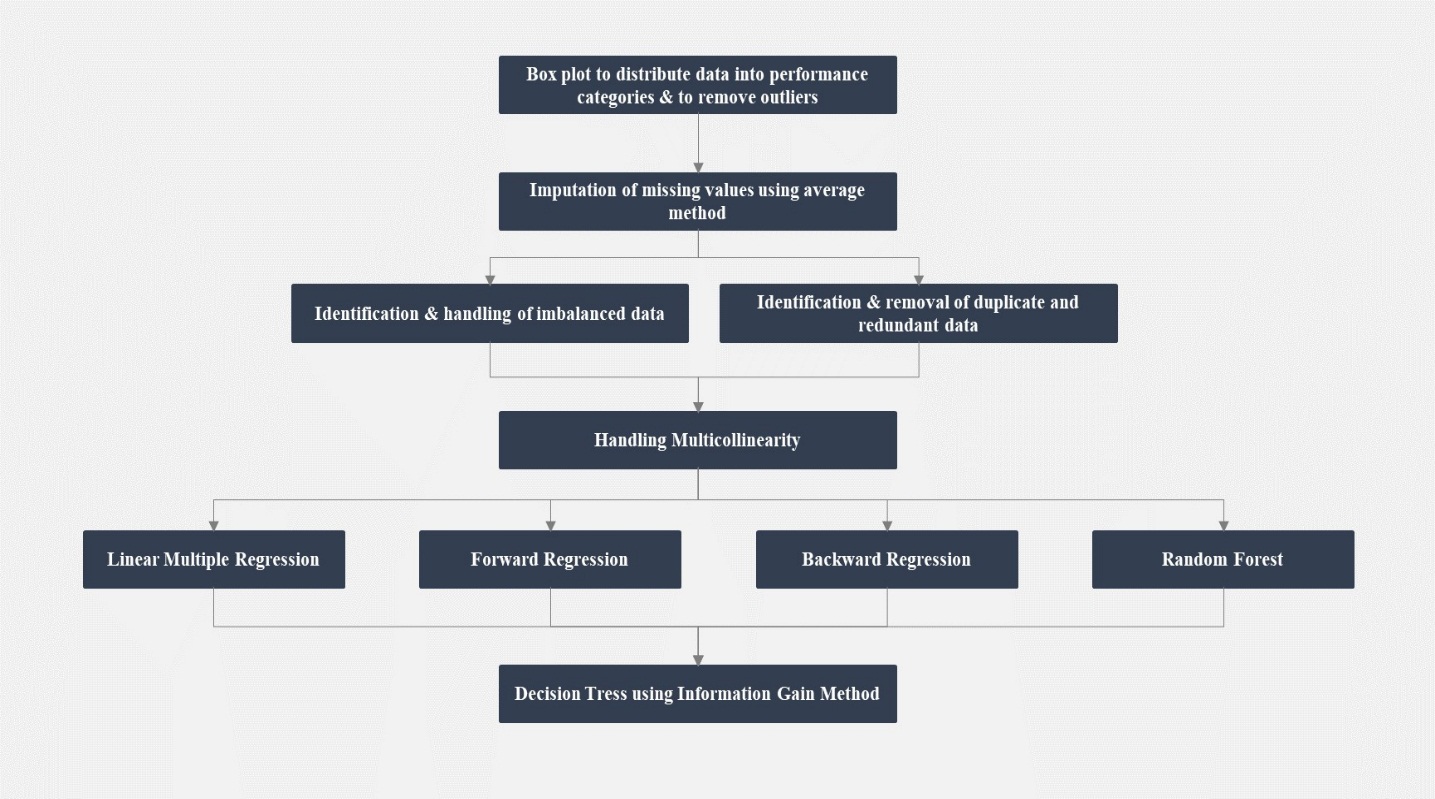
Below flowchart summarizes the approach used for solving the above problems.

Various statistical techniques were used to first clean the data for outliers, imbalanced data, duplicate data, redundant data and to prepare it for regression modelling by eliminating multicollinearity

Box plot was used to distribute data by revenue into three performance categories, namely Excellent, Good & Bad.

These categories were used as target variables in a decision tree to identify decision rules on significant stores characteristics (identified using regression techniques), which can be used to identify a high performance or a low performance store

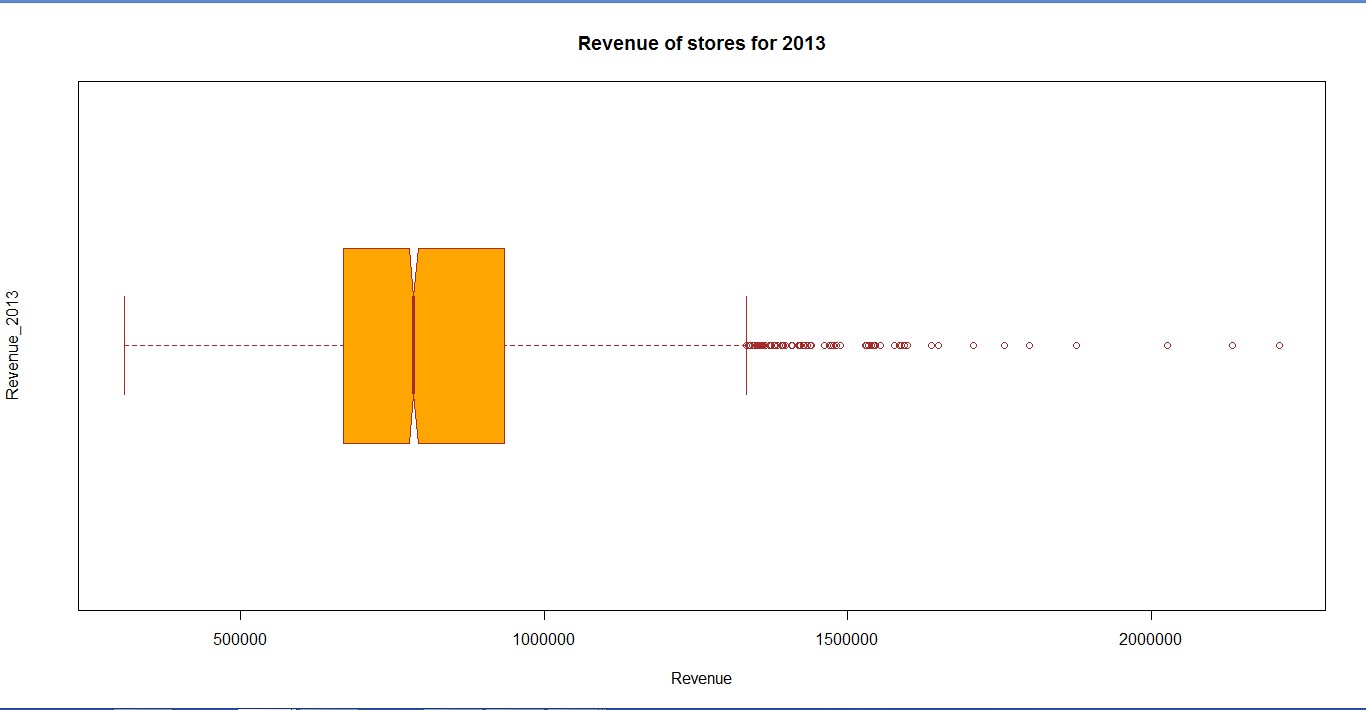
Regression model was also built to estimate sales for new stores



**II.1. Box Plot**

Below Box plot was created to distribute store data by revenue into three performance categories namely

Excellent, Good & Bad.



Below is the output of the above Box plot.

Results Value (Revenue) Min Value 309,408

First quartile 669,492.5

Second Quartile 785,471

Third Quartile 934,952.5

Fourth Quartile 1,332,707

Max Value 2,211,249

Data points outside the fourth quartile (revenue greater than 1,332,707) were treated as outliers and were deleted from the data set. After that, based on the identified quartiles, below conditions were used

to categorize store data into Excellent, Good & Bad categories as shown below.

Performance Categories Condition Count of Data Points

Excellent Revenue>= 934,952 640

Good 669,493< Revenue <=934952 1411

Bad Revenue<=669493 706

*Additionally, as it can be inferred from the above table, data is free from any imbalance anomalies.*

**II.2. Data Cleaning**

Below table summarizes the results of data cleaning activates performed.

**Activity Performed Row/Column Impacted Rationale**

Considered column **number of agreements**. Since number of agreements gave a better

Column deleted PERC\_CONVERTED\_TO\_AGREEMENT explanation.

**PERC\_CONVERTED\_TO\_AGREEMENT was**

deleted.

Duplicate columns (CYB07VBASE,

Column deleted CYB02V001

CENSUS\_DIVISION

CYB02V001). Only column CYB07VBASE was kept.

Considered only the **state** column.

Column deleted

CENSUS\_REGION U\_CITY

Column deleted PERC\_CYEA07V007 Values were almost negligible

Column deleted

SINGLE\_TENANT\_IND PAD\_IN\_SHOP\_CENTER\_IND COMP\_PRESENCE\_IND PAYLESS\_IND WALMART\_IND

TARGET\_IND AUTOZONE\_IND NUM\_PARKING\_SPACES

More than 80% of the rows had same values. Therefore, these column were deleted as these wouldn’t have made any variation in the model.

Rows deleted FRONTAGE\_ROAD Rows deleted where value was "**Unable to determine"** or "**Yes No"**

Values imputed

TOT\_ATTRITION\_2012

TOT\_ATTRITION\_2013

NUM\_ASSISTANT\_MANAGERS

NUM\_CUST\_ACC\_REPS NUM\_STORE\_MANAGERS NUM\_EMP\_PAY\_TYPE\_H AVG\_PAY\_RATE\_PAY\_TYPE\_S AVG\_PAY\_RATE\_PAY\_TYPE\_H

Missing values were imputed for these columns. Average value for the column was used for filling the missing data

**II.3. Handling Multicollinearity**

Multicollinearity is a problem because it can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable and difficult to interpret. To eliminate multicollinearity in the regression model, correlated continuous & categorical variables were removed as shown below.

**II.3.1. Chi-square test for categorical variables**

Chi-square was used to identify correlated categorical variables. The test was used on the variables:

**FRONTAGE\_ROAD”, “STRIP\_SHOP\_CENTER\_IND”.**

**Null hypotheses**: FRONTAGE\_ROAD, STRIP\_SHOP\_CENTER\_IND are independent

Below is the result for chi square test.

Results Values

Chi-square test statistic (X2) 9.4608

Degrees of freedom (df) 1

P-value 0.002099

Since p-Value is less than the significance level of 0.05, null hypothesis was rejected and it was concluded that the two variables are in fact dependent. Therefore, the variable “**STRIP\_SHOP\_CENTER\_IND**” was deleted.

**II.3.2. Correlation for continuous variables**

Before applying correlation, continuous variables were normalized by using **Z-Score** methodology. This was to ensure that the variables are at the same scale to facilitate to accurate application of correlation.

Variables which were highly correlated that is with correlation coefficient greater than or equal to 0.9 were deleted. Below are the results after running correlation test between all the continuous variables in the data.

Highlighted cells in the below correlation matrix shows highly correlated variable pairs. For instance, **NAT\_CURR\_BURGLARY** is correlated to **NAT\_PAST\_BURGLARY**. Therefore, one of the correlated variable was deleted for each pair. Deleted variables were: **NAT\_PAST\_BURGLARY, NAT\_PAST\_MOT\_VEH\_THEFT**

**& NAT\_PAST\_ROBBERY.**

**Correlation Matrix** NAT\_CURR\_ROBBER Y

NAT\_CURR\_BURGLA RY

NAT\_CURR\_MOT\_VE H\_THEFT

NAT\_CURR\_BURGLARY 0.490253909 1 0.522482183

NAT\_PAST\_BURGLARY 0.462288233 0.952152028 0.464386853

NAT\_CURR\_MOT\_VEH\_THEFT 0.805064151 0.522482183 1

NAT\_PAST\_MOT\_VEH\_THEFT 0.77499171 0.520196538 0.92290718

NAT\_CURR\_ROBBERY 1 0.490253909 0.805064151

NAT\_PAST\_ROBBERY 0.971109376 0.483519924 0.740677959

Similarly, variable **“PERC\_CYB11V006”** was deleted for the below correlation matrix.

**Correlation Matrix** PERC\_CYB11V006 PERC\_CYB11V007

PERC\_CYB11V006 1 0.947791664

PERC\_CYB11V007 0.947791664 1

Only variable **CYA01V001** was kept and all the other variables were deleted for the below correlation matrix.

**Correlation Matrix** CYA01V001 CYA12V003

CYA12V001 0.9687948 0.8345908

CYA12V002 0.96853677 0.99108425

CYA12V003 0.9359856 1

CYA12V007 0.9338927 0.8742704

CYA12V008 0.9382616 0.8852796

CYB07VBASE 0.98560684 0.92023831

Total\_White\_Population 0.825679237 0.910799698

**II.4. Regression**

After performing all data cleaning activities & removing correlated variables, **Linear multiple regression**

model was built.

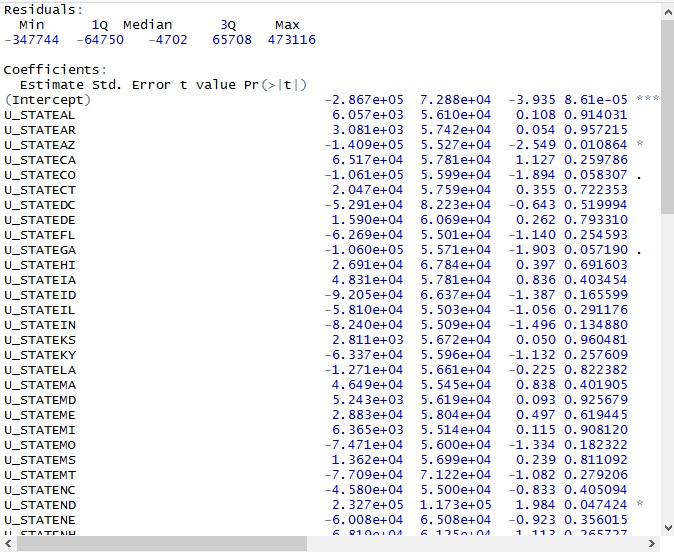
**Dependent Variable:** “revenue\_2013”

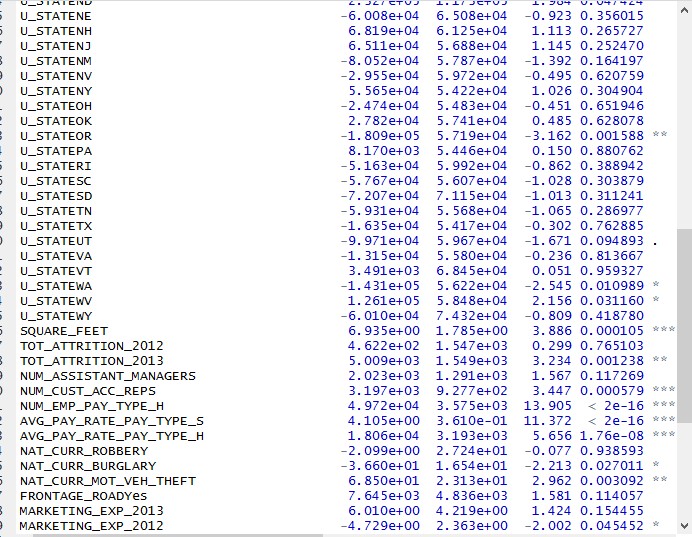
**Independent Variables:**

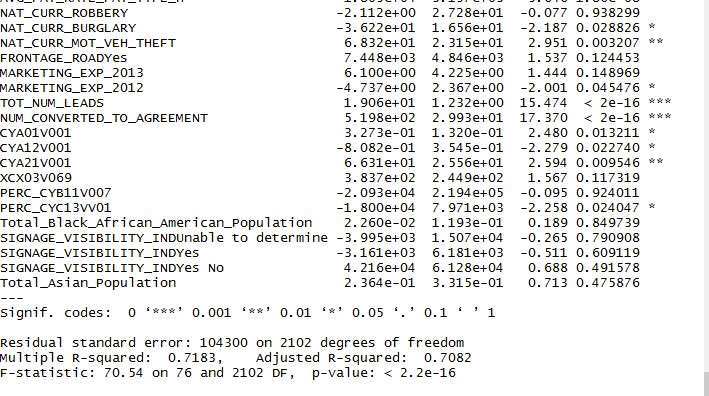
“U\_STATE”, “SQUARE\_FEET”, “TOT\_ATTRITION\_2012”, “TOT\_ATTRITION\_2013”, “NUM\_ASSISTANT\_MANAGERS”, “NUM\_CUST\_ACC\_REPS”, “NUM\_STORE\_MANAGERS”, “NUM\_EMP\_PAY\_TYPE\_H”, “AVG\_PAY\_RATE\_PAY\_TYPE\_S”, “AVG\_PAY\_RATE\_PAY\_TYPE\_H”, “NAT\_CURR\_ROBBERY”, “NAT\_CURR\_BURGLARY”, “NAT\_CURR\_MOT\_VEH\_THEFT “, “FRONTAGE\_ROAD”, “MARKETING\_EXP\_2013”, “MARKETING\_EXP\_2012”, “TOT\_NUM\_LEADS”, “NUM\_CONVERTED\_TO\_AGREEMENT”, “CYA01V001”, “CYA12V001”, “CYA21V001”, “XCX03V069” “PERC\_CYB11V007”, “PERC\_CYC13VV01”, “Total\_Black\_African\_American\_Population”, “Total\_Asian\_Population**”**

Below are the summary screen shots (portioned into three for sake of clarity) of the results of the regression model

**Summary (1/3):**





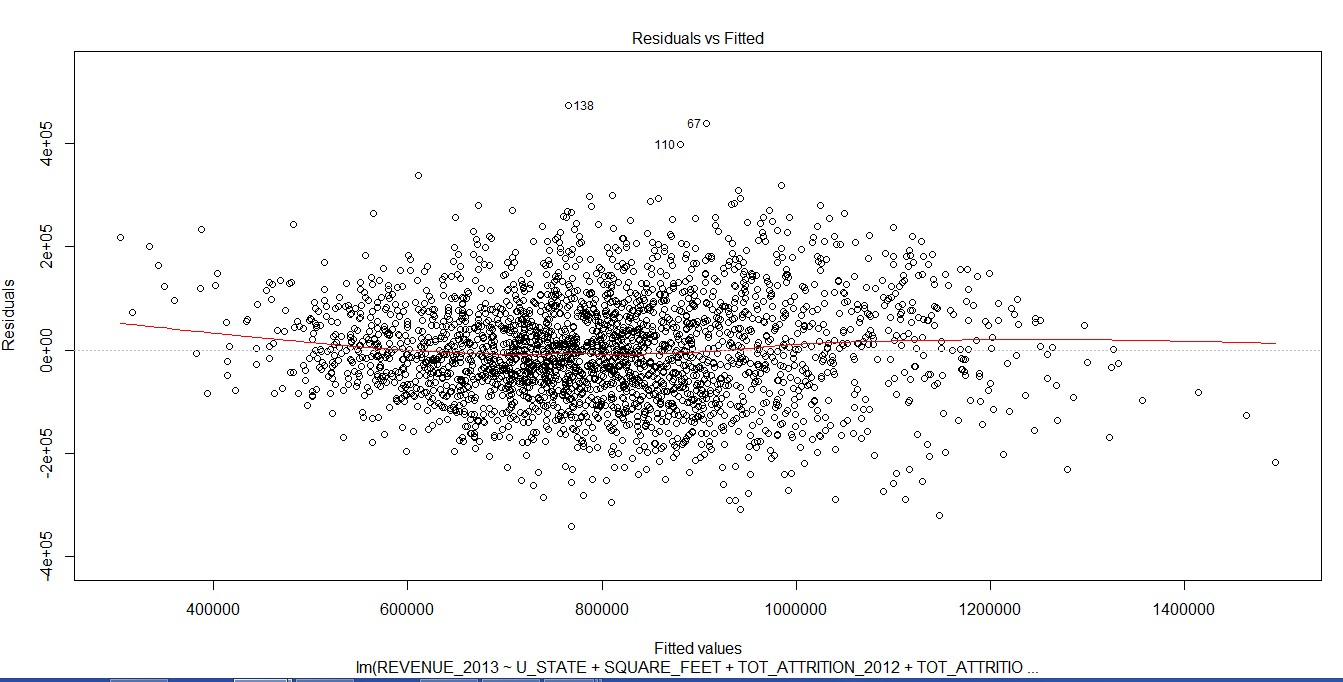


As evident from the above summary, **R square value is moderately high at 71.83%** and **adjusted R square at 70.82%** is close to R square. This implies that the model explains the variability of the response data to a good extent.

**II.4.1.** Testing Regression Model

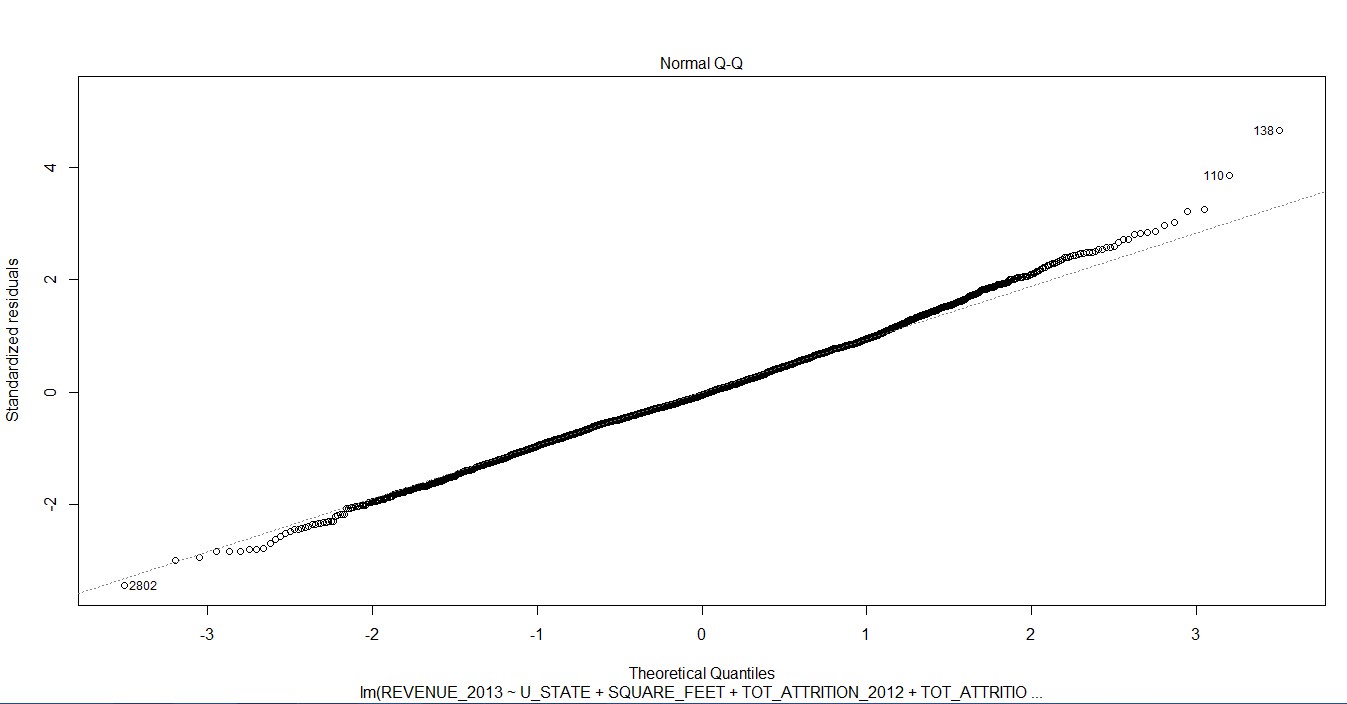
In addition to looking at R-square, other tests were also conducted to validate the model as below.

**Homoscedasticity Test**



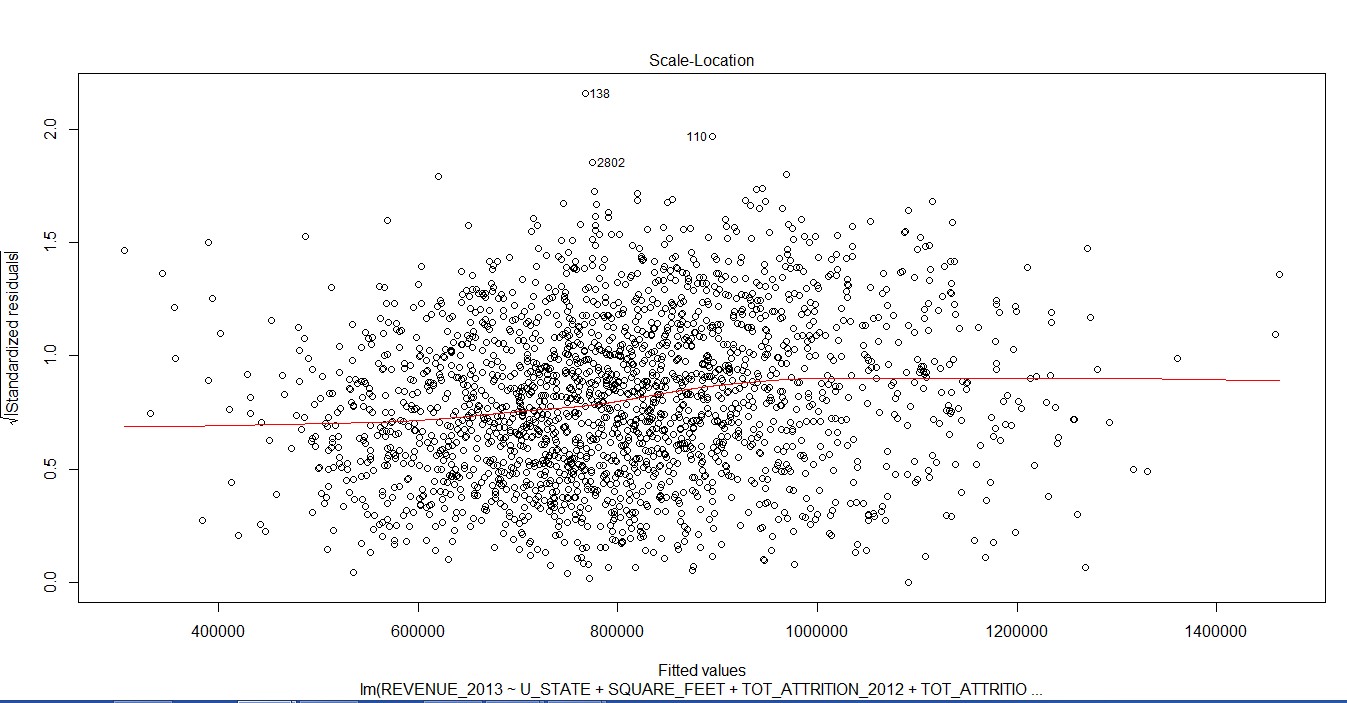
Homoscedasticity describes a situation in which the error term (that is, the “noise” or random disturbance in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables. From the above graph, it was inferred that there is Homoscedasticity in the model which means there are not outliers in the model & the true variance and covariance are not underestimated.

**QQ plot**



The QQ plot was a straight line which indicates the errors are normally distributed.

**Standardized Residual Plot**



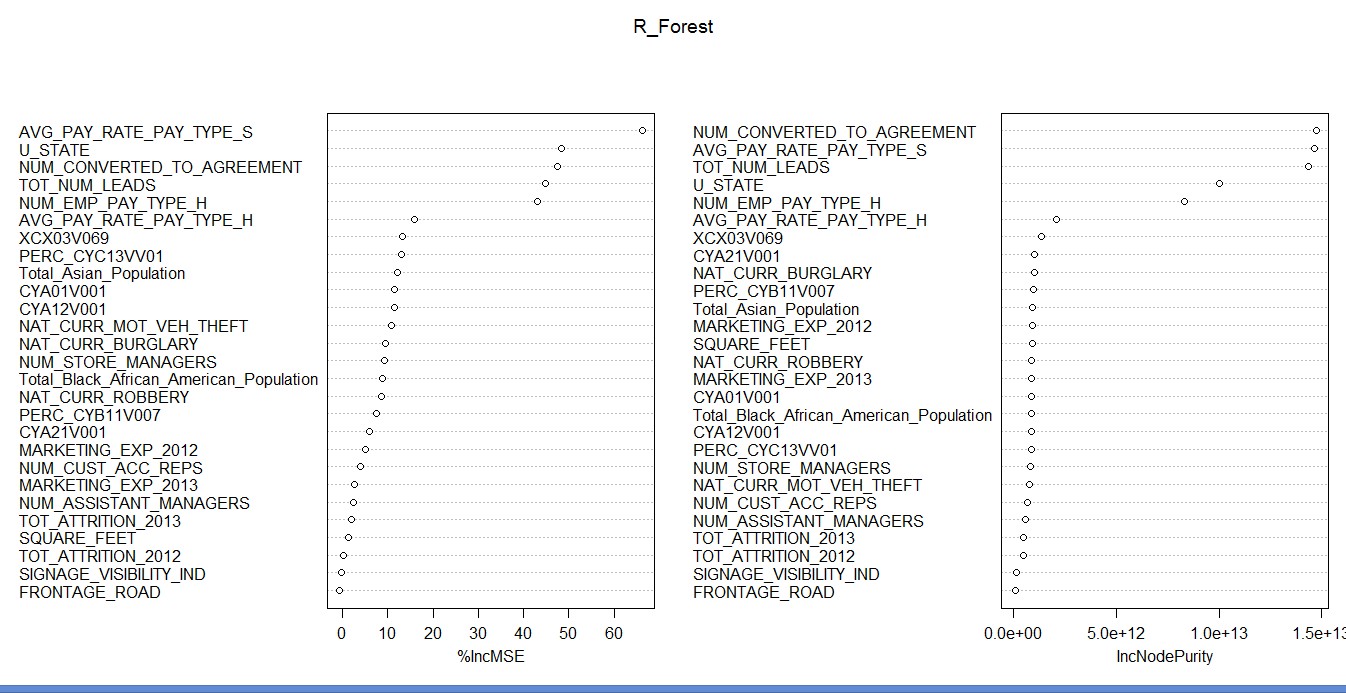
The Standardized Residual plot was homogeneously distributed, and no patterns were observed. This indicates that error terms

Thus, all the above tests helped to validate the assumptions taken in the model.

Additionally, stepwise regression models (both backward and frontward) were also run and gave the same results.

**II.4.2. Random Forest**

Random forest was used along with the Linear multiple regression to select a list of common significant variables. This was to done to validate the results of the liner regression. Dependent & Independent variables similar to linear multiple regression model were used. Below is the variable importance plot of the random forest.



Using the results of both the **Linear Multiple Regression model** & the **Random forest**, common top significant variables were identified. Using these significant variables, regression equation was built.

**Regression Equation**

**Sales**= (-2.960e+05)

+ (2.942e+00\*AVG\_PAY\_RATE\_PAY\_TYPE\_S)

+ (5.653e+02\*NUM\_CONVERTED\_TO\_AGREEMENT)

+ (6.022e+04\*NUM\_EMP\_PAY\_TYPE\_H)

+ (1.534e+01\*TOT\_NUM\_LEADS)

+ (2.150e+04\*AVG\_PAY\_RATE\_PAY\_TYPE\_H)

+ (3.837e+02\*XCX03V06)

+ (-1.426e+05\*U\_STATEAZ)

+ (2.326e+05\*U\_STATEND)

+ (-1.813e+05\*U\_STATEOR)

+ (-1.441e+05\*U\_STATEWA)

+ (1.255e+05\*U\_STATEWV)

**II.5. Decision Tree**

Decision tree was built to come up with decision rules that can be used to evaluate the performance of a store. It was build using the **Information Gain Methodology.**

**Dependent Variables**

Performance categories based on revenue, as identified earlier using the box plot, were used as dependent variables.

**Independent Variables**

Below significant variables identified using the regression were used as independent variables in the decision tree.

U\_STATE\_AZ

U\_STATE\_ND

U\_STATE\_OR

U\_STATE\_WA

U\_STATE\_WV

AVG\_PAY\_RATE\_PAY\_TYPE\_S

NUM\_CONVERTED\_TO\_AGREEMENT

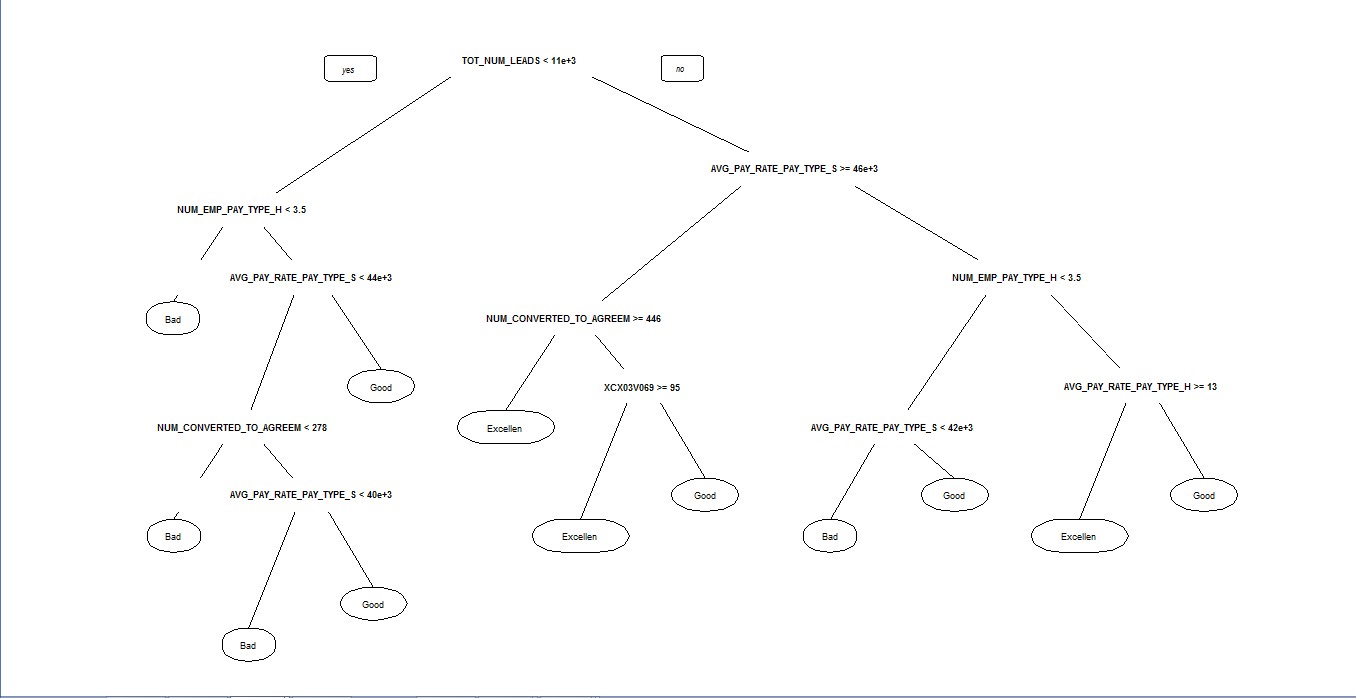
NUM\_EMP\_PAY\_TYPE\_H

TOT\_NUM\_LEADS+AVG\_PAY\_RATE\_PAY\_TYPE\_H

XCX03V06

Please note that the new variables were created for the states based on output of the regression equation. These are U\_STATE\_AZ, U\_STATE\_ND, U\_STATE\_OR, U\_STATE\_WA & U\_STATE\_WV. These variables could take tow values either 0 (which implies that store is not in specified state) or 1 (which implies that store is in specified state).

Below is the decision tree built using the above inputs.



We have assumed that the performance catego ry “ Excellent” equates to the high perfo rm ing sto res & the

cate go ry “ Bad” equate s to lo w perfo rm ing sto res. **Final summarized results of the decision tress are given in the next section.**

**III. Results**

Below are the results for the Problem 2.

1. Linear multiple regression model was used to come up with regression equation to estimate the sales at a new store given characteristics of the new store & location. Additionally, forward & backward regression was also done which gave exactly the same significant variables as linear multiple regression. Random forest was used along with the Linear multiple regression to select a list of common significant variables. This was to done to validate the results of the liner regression.

Below is the regression equation to estimate sales of a new store.

**Sales**= (-2.960e+05)

+ (2.942e+00\*AVG\_PAY\_RATE\_PAY\_TYPE\_S)

+ (5.653e+02\*NUM\_CONVERTED\_TO\_AGREEMENT)

+ (6.022e+04\*NUM\_EMP\_PAY\_TYPE\_H)

+ (1.534e+01\*TOT\_NUM\_LEADS)

+ (2.150e+04\*AVG\_PAY\_RATE\_PAY\_TYPE\_H)

+ (3.837e+02\*XCX03V06)

+ (-1.426e+05\*U\_STATEAZ)

+ (2.326e+05\*U\_STATEND)

+ (-1.813e+05\*U\_STATEOR)

+ (-1.441e+05\*U\_STATEWA)

+ (1.255e+05\*U\_STATEWV)

*As evident fr om the above equation, it’s not advisable to open a store in the states: AZ, OR & WA since these have negative impact on sales because of negative regression coefficients. On the other hand, states: ND & WV are favourable locations to open new stores.*

2. Decision tree was used to come up with the conditions that can be applied to the stores characteristics to evaluate if a store is high performing or low performing. Below table summarizes the various decision rules identified. There are three decision rules to identify a high performing store & four

decision rules to evaluate a low performing store.

Please note that for each decision rules, all the conditions on store characteristics should be satisfy.

**Performance** Decision Rules

Tot\_Num\_Leads>=11,000

Avg\_Pay\_Rate\_Pay\_Type\_S>=46,000

Num\_Converted\_To\_Agreem>=446

Tot\_Num\_Leads>=11,000

Avg\_Pay\_Rate\_Pay\_Type\_S>=46,000

High Performance Stores

Low Performance Stores

Num\_Converted\_To\_Agreem<446

XCX03V069>=95

Tot\_Num\_Leads>=11,000

Avg\_Pay\_Rate\_Pay\_Type\_S<46,000

Num\_Emp\_Pay\_Type\_H>=3.5

Avg\_Pay\_Rate\_Pay\_Type\_H>=13

Tot\_Num\_Leads<11,000

Num\_Emp\_Pay\_Type\_H<3.5

Tot\_Num\_Leads<11,000

Num\_Emp\_Pay\_Type\_H>=3.5

Avg\_Pay\_Rate\_Pay\_Type\_S<44000

Num\_Converted\_To\_Agreem<278

Tot\_Num\_Leads<11,000

Num\_Emp\_Pay\_Type\_H>=3.5

Avg\_Pay\_Rate\_Pay\_Type\_S<44000

Num\_Converted\_To\_Agreem>=278

Avg\_Pay\_Rate\_Pay\_Type\_S<40,000

Tot\_Num\_Leads>=11,000

Avg\_Pay\_Rate\_Pay\_Type\_S<46,000

Num\_Emp\_Pay\_Type\_H<3.5

Avg\_Pay\_Rate\_Pay\_Type\_S<42,00