1.) A cloth manufacturing company is interested to know about the segment or attributes causes high sale.

Approach - A Random Forest can be built with target variable Sales (we will first convert it in categorical variable) & all other variable will be independent in the analysis.

Solution:

**Business Problem:** To prepare a model for sale.

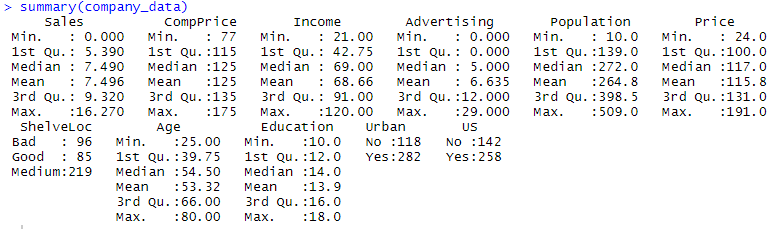
**Datasets:**

Dependent Variable: Sales

Independent Variable: CompPrice, Income, Advertising, Population, Price, ShelveLoc,

Age, Education, Urban, US

**EDA:**



For converting target variable sales into categorical variables need to create bins.

So three bins namely ‘Low’, ‘Average’, ’High’ is created to divided the data of sales variable.

A new variable in the name of salesstatus is added to the datasets which consist record of Low Average and High sales.

**Data Partitioning:**

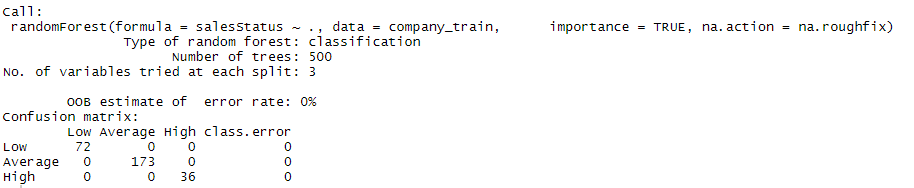
The data is divided into test and train by 7:3 ratio.

**Building Model:**

A model is built using random forest algorithm.

company\_rf <- randomForest(salesStatus ~ ., data=company\_train, na.action=na.roughfix, importance=TRUE)

Model summary



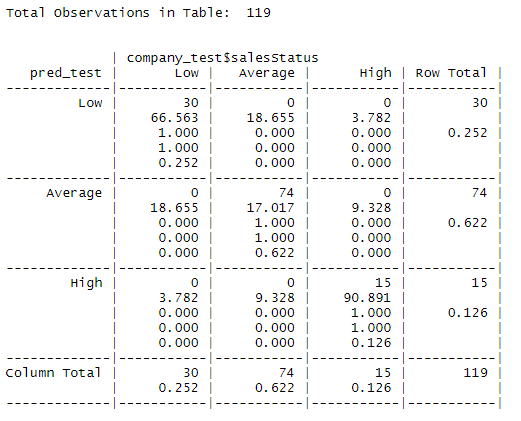
No. of trees = 500 which means 500 decision tree is used to get the average result

Training accuracy is 100 %, as error rate is 0.

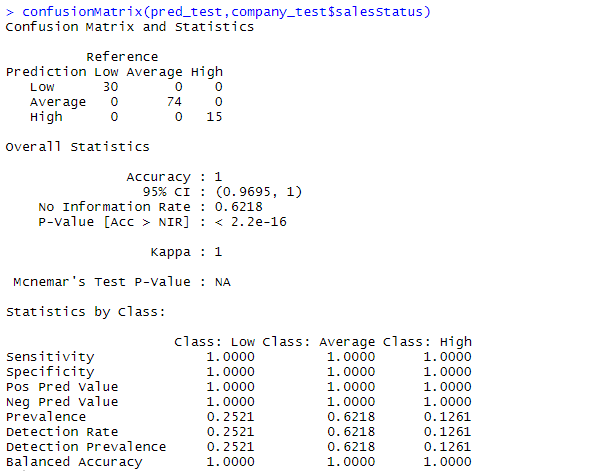
**Evaluation:**

The performance of model is evaluated on test datasets.

Cross tabulation



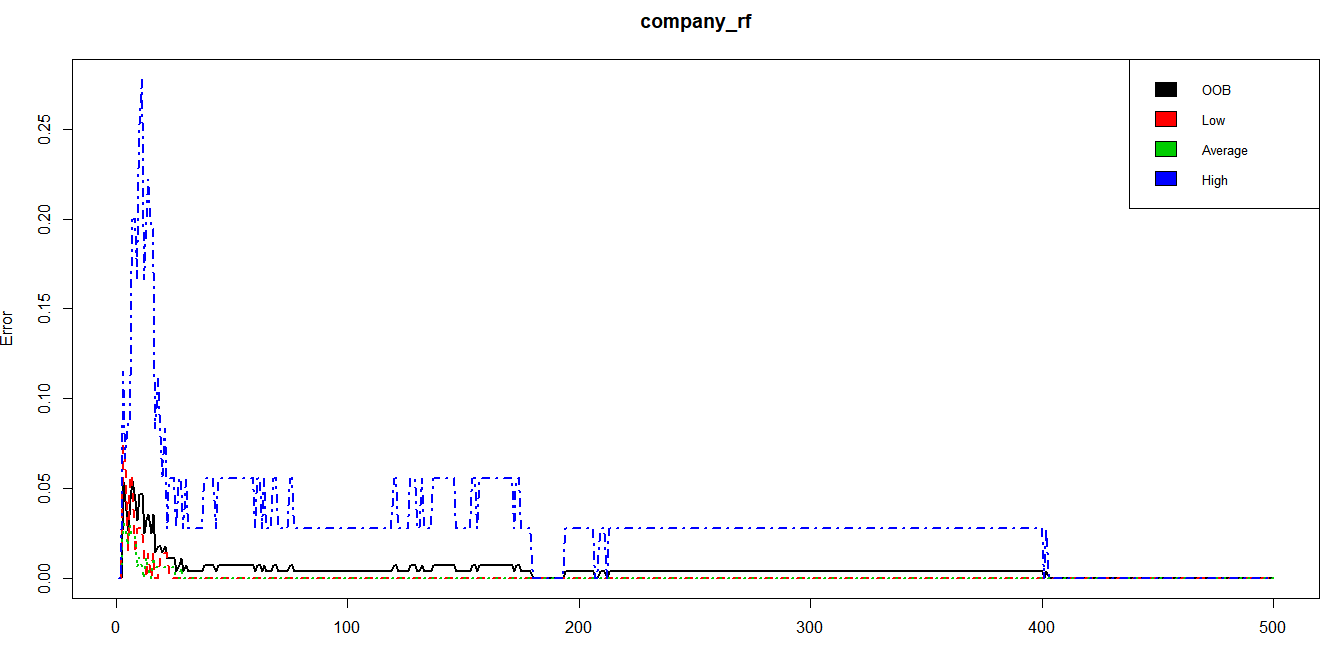
From the above cross table of predicted salesstatus and actual test data, there is no misclassifcation

Confusion matrix 

Accuracy = 100 %

So, the model seems to be very good model with 0 % error.

**Visualization:**



2.) Use Random Forest to prepare a model on fraud data treating those who have taxable

income <= 30000 as "Risky" and others are "Good"

Solution:

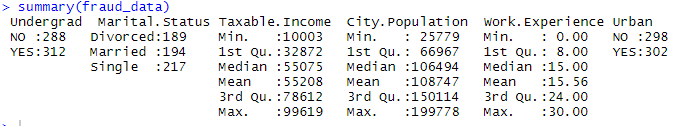
**Business Problem:** To prepare a model to predict risky or good for fraud data.

**Datasets:**

Dependent Variable: Taxable.Income

Independent Variable: Undergrad, Marital.Status, City.Population, Work.Experience, Urban

**EDA:**



Adding one more variable as status includes record as “Risky” and “Good” by dividing the datasets into >30000 as “Good” and others as “Risky”.

Data Partitioning

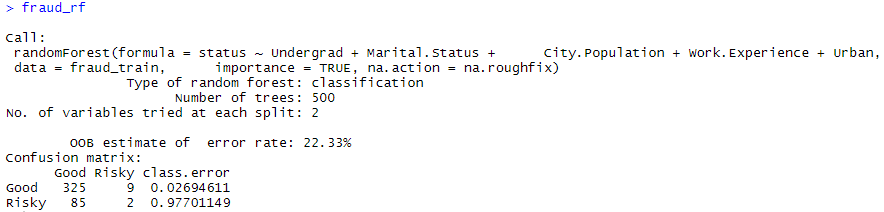
Dividing the datasets into train and test datasets

**Model Building:**

A model is build on train datasets

fraud\_rf <- randomForest(status~Undergrad+Marital.Status+City.Population+Work.Experience+Urban,data=fraud\_train, na.action=na.roughfix,importance=TRUE)

Model Summary



Number of trees created is 500

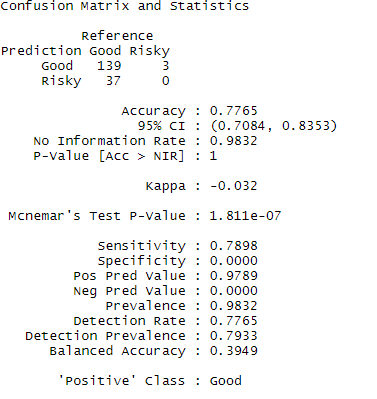
The Out-of-bag OOB error estimate rate is 22.33%

Model accuracy for train data = 92.63%

**Evaluation:**

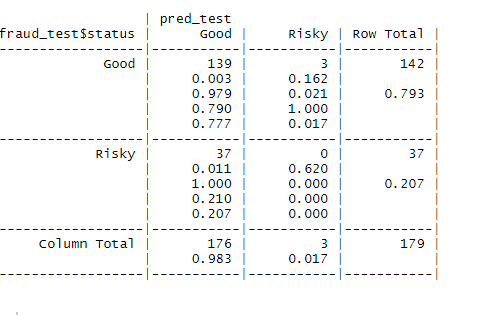
Evaluation is checked on test datasets.

Confusion Matrix



If we look at the Confusion Matrix, we can see that classification error is quite low. And accuracy is 77.65 %. This shows that our RF model is performing average in classifying the test data set.

Cross tabulation



Total correctly classified is 139 and 40 is misclassified.

**Visualization:** 