# Problem 2: CART-RF-ANN

## An Insurance firm providing tour insurance is facing higher claim frequency. The management decides to collect data from the past few years. You are assigned the task to make a model which predicts the claim status and provide recommendations to management. Use CART, RF & ANN and compare the model’s performances in train and test sets.

### 2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it.

There are 3000 Observations and 9 Attributes. Claimed is the target variable while all others are the predictors. Out of the 9 columns, 5 are object type, 2 are int and 2 are float type.

Object - Type, Claimed, Channel, Product Name, Destination.

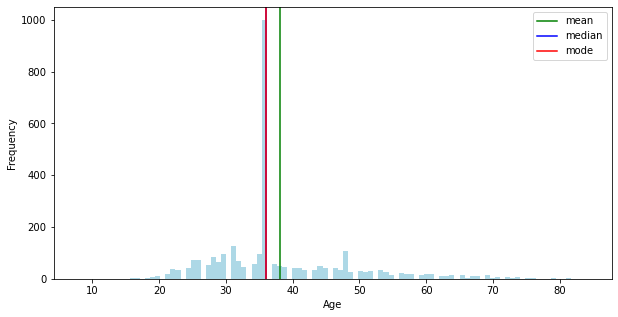
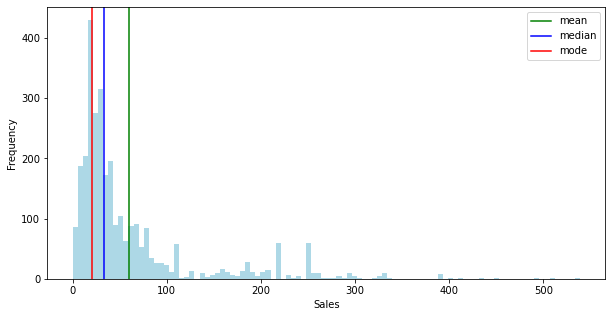
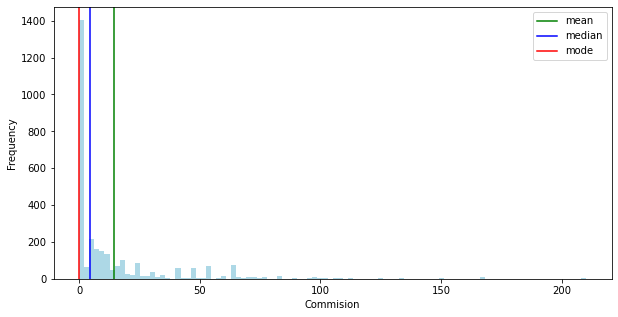
Int - Age and Duration

Float - Commision and Sales

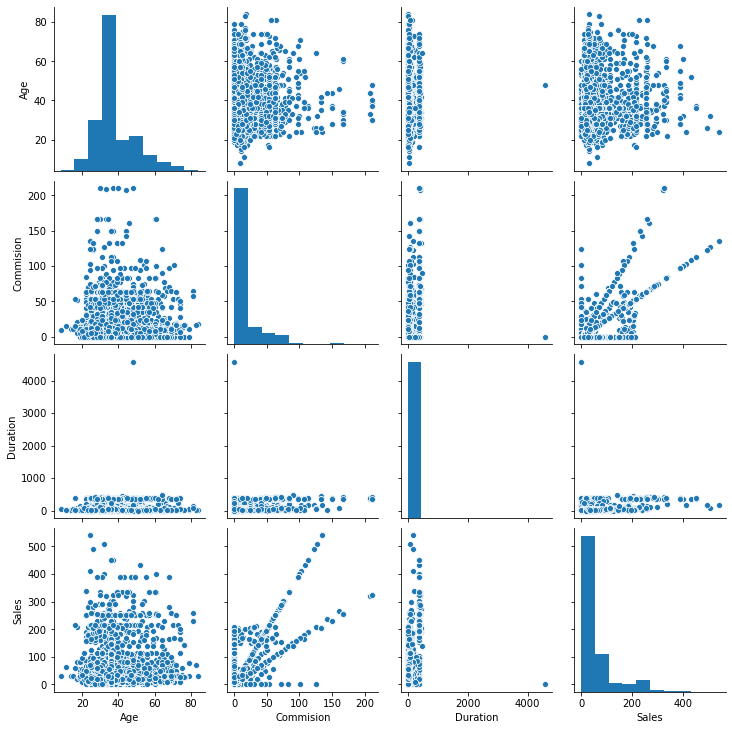
Since label encoding is already present in the data, some of the nominal variables are displayed as int.

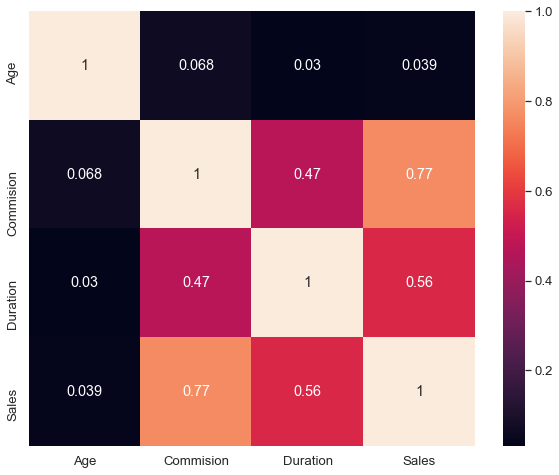
It appears there are also no missing values.

From the descriptive analysis, we can say that all the continuous variables are almost positively skewed.



There are positive correlations between variables. Overall the magnitude of correlations between the variables are very less.

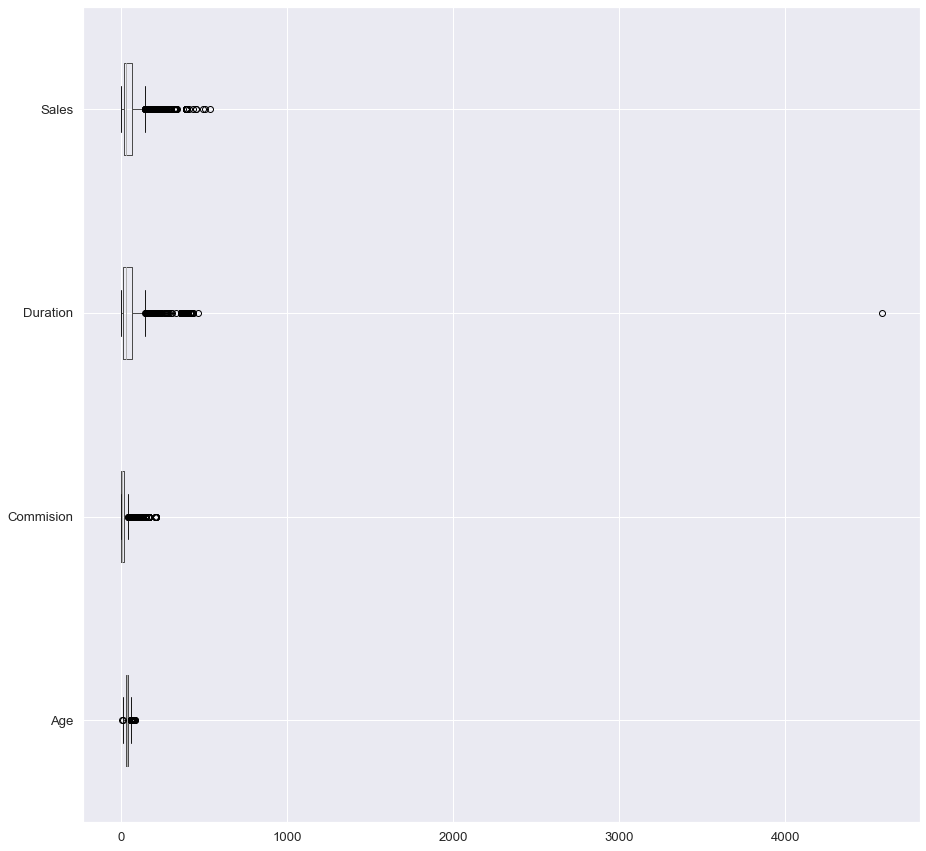




There are 139 duplicates found. Duplicates are removed by using the duplicated() function.

There are outliers in all the variables. Random Forest and ANN can handle the outliers.

Hence, outliers are not treated for now. We will keep the data as it is.



### 2.2 Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network

After splitting data into test and train, dimensions of the training and test data are:

|  |
| --- |
| X\_train (2002, 8)  X\_test (859, 8)  train\_labels (2002,)  test\_labels (859,) |

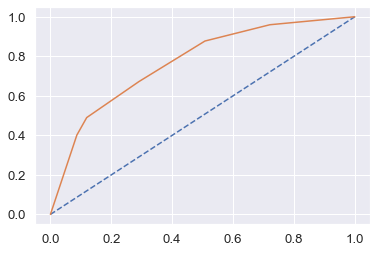
Product Name is the most important variable for predicting claimed status

### 2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model

#### CART model:

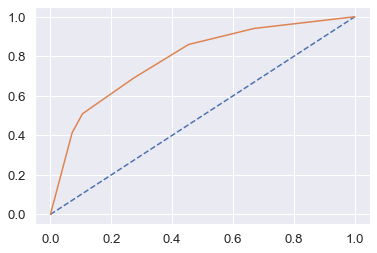
For the train Data, AUC - 77% , Accuracy - 76%, Precision - 66%, f1-Score - 56% :

AUC and ROC curve for the training data



For the test Data, AUC - 79%, Accuracy - 77% , Precision - 69%, f1-Score - 59%:

### AUC and ROC curve for the test data



Training and Test set results are almost similar, and with the overall measures high, the model is a good model.

Product Name is the most important variable for predicting claimed status

RF model:

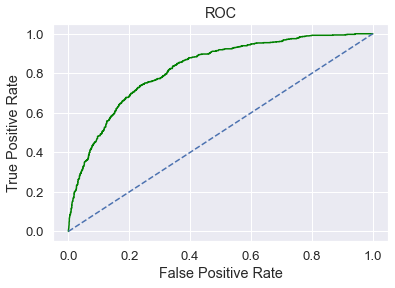
For the train data, AUC: 82%, Accuracy: 76%, Precision: 72%, f1-Score: 54%

For the test data, AUC: 81%, Accuracy: 76%, Precision: 71%, f1-Score: 53%

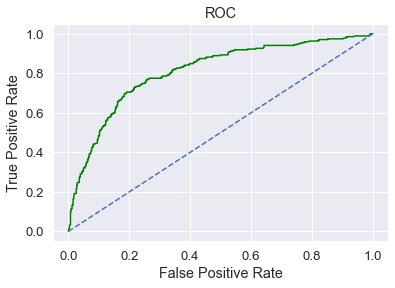
Training and Test set results are almost similar.

Product Name is again the most important variable for predicting claimed status.

### AUC and ROC curve for the training data



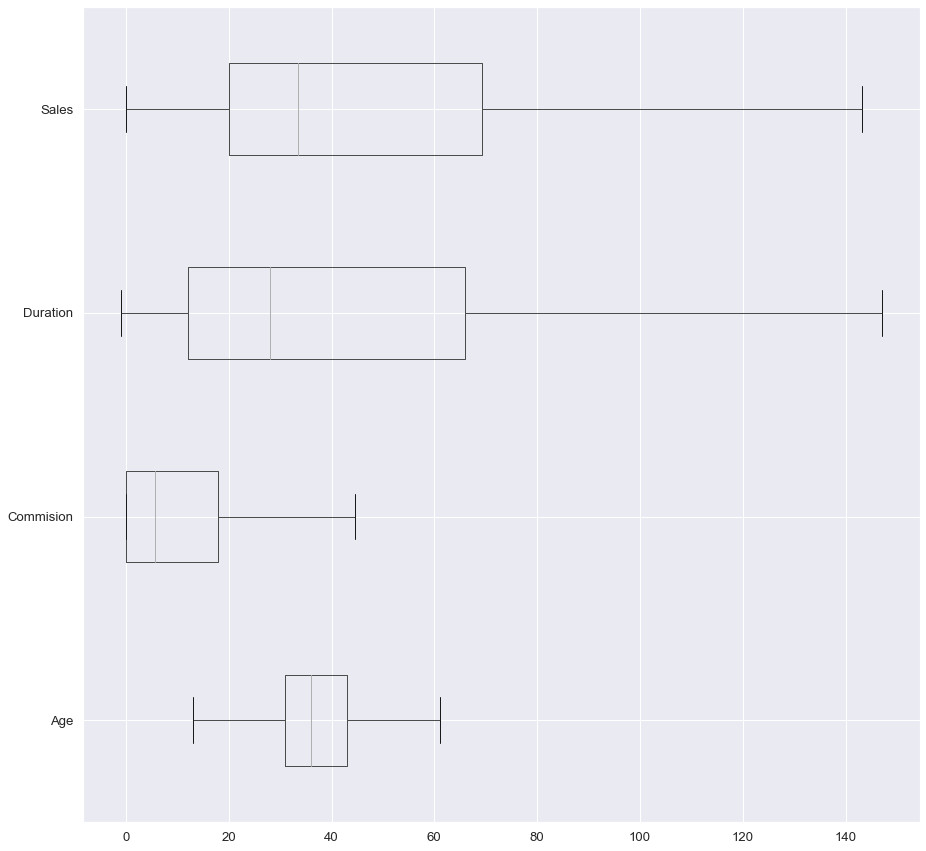
AUC and ROC curve for the test data



ANN Model:

For ANN model, treated the outliers and scaled the data using a standard scaler. Training data is fitted into the standard scaler and transformed. We scaled the test data. We should not fit for test data. Because we already fitted the train data and using the mean and std of the training data, we have to create or standardize the test data to ensure that the scaling is uniform between training and testing.

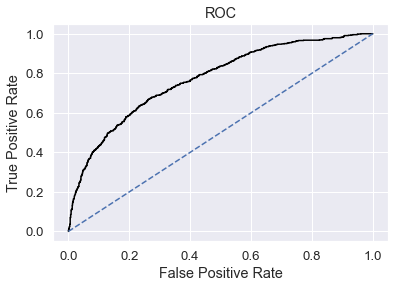
After treating the outliers for continuous variables,



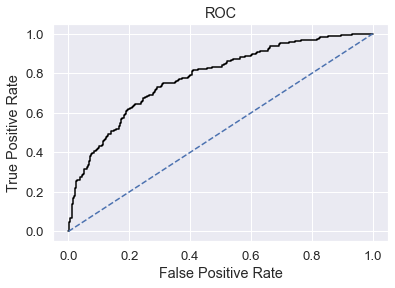
And after scaling and running the model, train data is AUC: 77%, Accuracy: 75%, Precision: 69%, f1-Score: 52% and test data is AUC: 78%, Accuracy: 76%, Precision: 71%, f1-Score: 50%.

Training and Test set results are almost similar.

### AUC and ROC curve for the training data

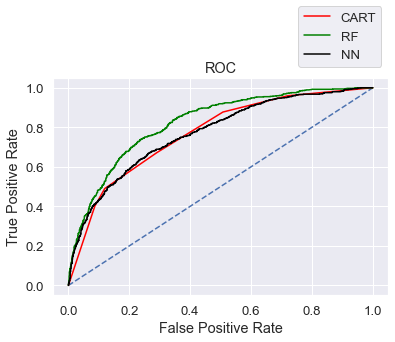


### AUC and ROC curve for the test data

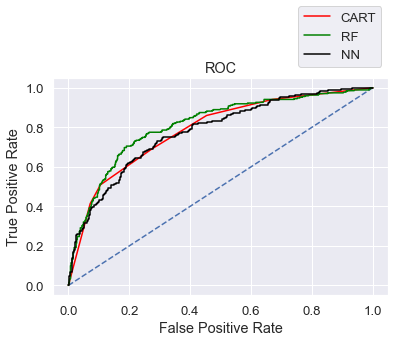


### 2.4 Final Model: Compare all the models and write an inference which model is best/optimized.

After comparing all the three models, ROC Curve for the 3 models on the Training data



ROC Curve for the 3 models on the Test data



Out of the 3 models, Random Forest has slightly better performance than the Cart and Neural network model.

### 2.5 Inference: Based on the whole Analysis, what are the business insights and recommendations

Overall all the 3 models are reasonably stable enough to be used for making any future predictions.

From Cart and Random Forest Model, Product Name is found to be the most useful feature amongst all other features for predicting claim status.