# Problem: Insurance Firm Tour Insurance Case Study

## ****Problem Statement:****

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 models' performances in train and test sets.

## Data Description:

insurance\_part2\_data-1.csv is a dataset that contains the names of various colleges.

## Domain:

Insurance (Travel)

## Context:

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 models' performances in train and test sets.

## Attribute Information:

1. Target: Claim Status (Claimed)
2. Code of tour firm (Agency\_Code)
3. Type of tour insurance firms (Type)
4. Distribution channel of tour insurance agencies (Channel)
5. Name of the tour insurance products (Product)
6. Duration of the tour (Duration)
7. Destination of the tour (Destination)
8. Amount of sales of tour insurance policies (Sales)
9. The commission received for tour insurance firm (Commission)
10. Age of insured (Age)

## 2.1) Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis)

### Basic EDA summary:-

* Data contains 3000 rows and 10 columns.
* Data contains integer, float and object type columns.
* There is no null/missing data in any of these columns
* There are total of 139 duplicate data rows present in the dataset Some example of duplicate rows are as shown below :-

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Age** | **Agency\_**  **Code** | **Type** | **Claimed** | **Commision** | **Channel** | **Duration** | **Sales** | **Product Name** | **Destination** |
| **63** | 30 | C2B | Airlines | Yes | 15 | Online | 27 | 60 | Bronze Plan | ASIA |
| **329** | 36 | EPX | Travel Agency | No | 0 | Online | 5 | 20 | Customised Plan | ASIA |
| **407** | 36 | EPX | Travel Agency | No | 0 | Online | 11 | 19 | Cancellation Plan | ASIA |
| **411** | 35 | EPX | Travel Agency | No | 0 | Online | 2 | 20 | Customised Plan | ASIA |
| **422** | 36 | EPX | Travel Agency | No | 0 | Online | 5 | 20 | Customised Plan | ASIA |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **2940** | 36 | EPX | Travel Agency | No | 0 | Online | 8 | 10 | Cancellation Plan | ASIA |
| **2947** | 36 | EPX | Travel Agency | No | 0 | Online | 10 | 28 | Customised Plan | ASIA |
| **2952** | 36 | EPX | Travel Agency | No | 0 | Online | 2 | 10 | Cancellation Plan | ASIA |
| **2962** | 36 | EPX | Travel Agency | No | 0 | Online | 4 | 20 | Customised Plan | ASIA |
| **2984** | 36 | EPX | Travel Agency | No | 0 | Online | 1 | 20 | Customised Plan | ASIA |

With the data available it appears that same tour operator can give same tour with same plan to same age group people via same channel (say online) which will give same sale and commision to the insurance firm. More details like dates of travel and Insured person name etc. would be required here to find if these are genuinely duplicate records or not. So this doesn’t appears to be a duplicate data entry problem.

Hence at the moment, we are not removing the duplicates here.

### Univariate Analysis

### Summary

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **unique** | **top** | **freq** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **Age** | 3000 | NaN | NaN | NaN | 38.091 | 10.46352 | 8 | 32 | 36 | 42 | 84 |
| **Agency\_Code** | 3000 | 4 | EPX | 1365 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Type** | 3000 | 2 | Travel Agency | 1837 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Claimed** | 3000 | 2 | No | 2076 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Commision** | 3000 | NaN | NaN | NaN | 14.5292 | 25.48146 | 0 | 0 | 4.63 | 17.235 | 210.21 |
| **Channel** | 3000 | 2 | Online | 2954 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Duration** | 3000 | NaN | NaN | NaN | 70.00133 | 134.0533 | -1 | 11 | 26.5 | 63 | 4580 |
| **Sales** | 3000 | NaN | NaN | NaN | 60.24991 | 70.73395 | 0 | 20 | 33 | 69 | 539 |
| **Product Name** | 3000 | 5 | Customised Plan | 1136 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Destination** | 3000 | 3 | ASIA | 2465 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |

From the summary, we can see that :-

* Channel used in tour is primarily online (98%) as it contains 2954 rows out of 3000 rows.
* Around 60% tour insurance firm are 'Travel Agency' type.
* Around 37% of tours are customised plan only.
* Min value for duration is -1 which is strange as duration cannot be negative.
* Max duration value also seems to be very high (4580). This value is extremely high and its causing data to be skewed.
* Max age also appears to be high as compared to the other percentile values, so there appears to be outliers in age column too
* Max Commission , max sale and max duration for tours seemt be very high as compared to their mean. So there appears to be outliers in this data i.e. for few tours sale, duration and commission amounts are very high as compared to the other tours

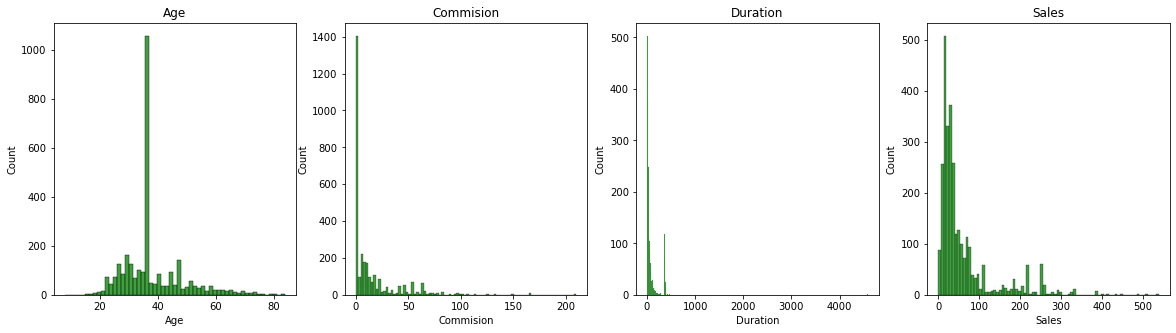
### Checking IQR, Coeffiecient of Variation, IQR, lower range and upper range of numerical cols with summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Age** | **Commision** | **Duration** | **Sales** |
| **count** | 3000 | 3000 | 3000 | 3000 |
| **mean** | 38.09 | 14.53 | 70 | 60.25 |
| **std** | 10.46 | 25.48 | 134.05 | 70.73 |
| **min** | 8 | 0 | -1 | 0 |
| **25%** | 32 | 0 | 11 | 20 |
| **50%** | 36 | 4.63 | 26.5 | 33 |
| **75%** | 42 | 17.24 | 63 | 69 |
| **max** | 84 | 210.21 | 4580 | 539 |
| **CV** | 0.27 | 1.75 | 1.91 | 1.17 |
| **Skew** | 1.15 | 3.15 | 13.78 | 2.38 |
| IQR | 10 | 17.23 | 52 | 49 |
| UR | 57 | 43.09 | 141 | 142.5 |
| LR | 17 | -25.85 | -67 | -53.5 |

Few things which can be noticed here is :-

* All columns are positively skewed (right skewed)
* Max values are very high as compared to the upper range (UR)

### Histogram



As seen in the histograms,

* Frequency of the records is too high for few values which is causing spikes in all the 4 histograms.
* For Age more than 1000 records (33%) are of age near about 35-40. It means 33% of the travellers are in age group 35-40.
* For commission around 1400 records (around 45%) have commission as 0. It means for almost 50% of tours no commission is earnred.
* Sale of tour insurance company is low for majority of the tours.
* In duration some values are extremely high which is causing x axis scale to be on higher side.

As seen in the summary and histograms, duration has negative value in it as a minimum value and maximum value of duration is extremely high.

### Analysing Duration

Checking for records less than or equals to 0

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Age** | **Agency\_Code** | **Type** | **Claimed** | **Commision** | **Channel** | **Duration** | **Sales** | **Product Name** | **Destination** |
| **1508** | 25 | JZI | Airlines | No | 6.3 | Online | -1 | 18 | Bronze Plan | ASIA |
| **1746** | 48 | C2B | Airlines | No | 0.14 | Online | 0 | 0.51 | Customised Plan | ASIA |
| **2628** | 37 | C2B | Airlines | No | 49.6 | Online | 0 | 124 | Bronze Plan | ASIA |

There seems to be some issues with duration here, as all other values like commission, sales, destination seems to contain data but duration of tour is 0 and negative for one row. These rows seem to be incorrect. We need to ask business for more clarification of these records. **As of now we are imputing duration with 0 for the negative duration row record.**

After imputation there is no record with negative duration of tour.

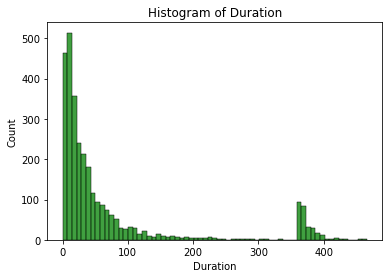
Now we are checking for duration of tours which are exceeding 400 (days).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Age** | **Agency\_Code** | **Type** | **Claimed** | **Commision** | **Channel** | **Duration** | **Sales** | **Product Name** | **Destination** |
| 27 | C2B | Airlines | Yes | 54 | Online | 401 | 216 | Silver Plan | ASIA |
| 31 | CWT | Travel Agency | No | 0 | Offline | 402 | 97 | Customised Plan | ASIA |
| 31 | C2B | Airlines | No | 46.96 | Online | 428 | 187.85 | Silver Plan | ASIA |
| 34 | C2B | Airlines | Yes | 68.08 | Online | 431 | 272.3 | Silver Plan | ASIA |
| 28 | C2B | Airlines | No | 63.21 | Online | 401 | 252.85 | Silver Plan | ASIA |
| 31 | C2B | Airlines | Yes | 63.21 | Online | 413 | 252.85 | Silver Plan | ASIA |
| 42 | CWT | Travel Agency | No | 132.99 | Online | 434 | 204.6 | Gold Plan | ASIA |
| 34 | CWT | Travel Agency | No | 166.53 | Online | 421 | 256.2 | Gold Plan | Americas |
| 30 | CWT | Travel Agency | Yes | 210.21 | Online | 417 | 323.4 | Gold Plan | Americas |
| 44 | C2B | Airlines | Yes | 63.21 | Online | 419 | 252.85 | Silver Plan | ASIA |
| 48 | C2B | Airlines | No | 0.09 | Online | 4580 | 0.32 | Customised Plan | ASIA |
| 64 | CWT | Travel Agency | No | 90.09 | Online | 466 | 138.6 | Silver Plan | ASIA |
| 27 | C2B | Airlines | Yes | 71.85 | Online | 416 | 287.4 | Gold Plan | ASIA |

As seen in the above data, out of 3000 row only 1 row contains 4580 as duration which seems to be very high. With such high value in duration, Sales and commission values are very less in contrast which again points that there is some problem in this figure

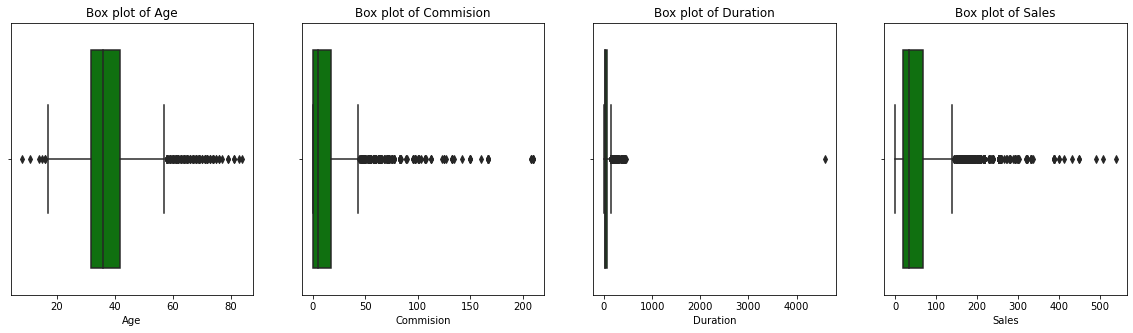
As of now we are imputing this value with the next highest value near about 466. It also appears that there could be 0 added accidently in this duration figure in this row. Hence we are imputing it with 458.

After imputing duration in one row from 4580 to 458, the duration histogram looks in much better shape.



### Box Plots

As evident from box plots shown below, there are outliers present in all the numeric columns



### Checking value count of categorical variables

#### Checking value count of Type

Travel Agency 61.23

Airlines 38.77

Name: Type, dtype: float64

Around 60% of tours are of type Travel Agency and remaining 40% from Airlines

#### Checking value count of Agency\_Code

EPX 45.50

C2B 30.80

CWT 15.73

JZI 7.97

Name: Agency\_Code, dtype: float64

4 agencies are used in which EPX has the highest 45% of data

#### Checking value count of Channel

Online 98.47

Offline 1.53

Name: Channel, dtype: float64

Only 1% of tours come via channel "Offline" i.e. all the tours come via online channel

#### Checking value count of Product Name

Customised Plan 37.87

Cancellation Plan 22.60

Bronze Plan 21.67

Silver Plan 14.23

Gold Plan 3.63

Name: Product Name, dtype: float64

5 products are used here in which customised plan is the highest with 38% data

#### Checking value count of Destination

ASIA 82.17

Americas 10.67

EUROPE 7.17

Name: Destination, dtype: float64

For majority of the tours, Destination is Asia (82%) followed by America (11%) and Europe (7%)

#### Checking value count of Claimed

No 69.2

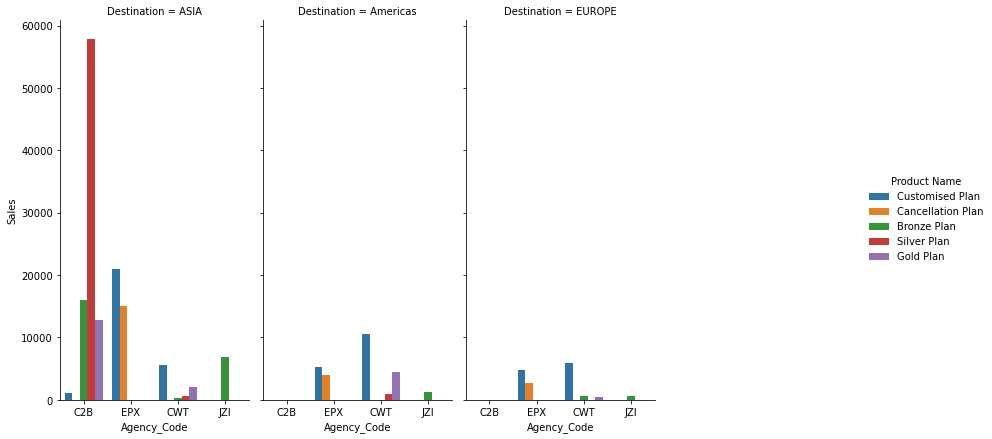
Yes 30.8

Name: Claimed, dtype: float64

Around 70% of tours have not claimed from insurance while 30% have claimed from the insurance. This target variable appears to be slighly imbalanced

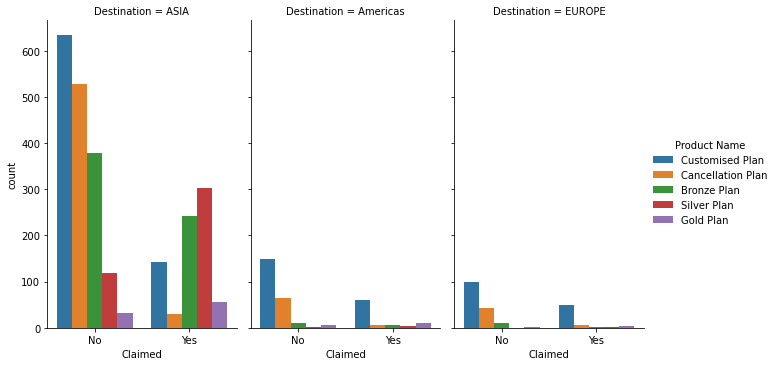
### Bivariate and Multivariate analysis

### Cat Plots

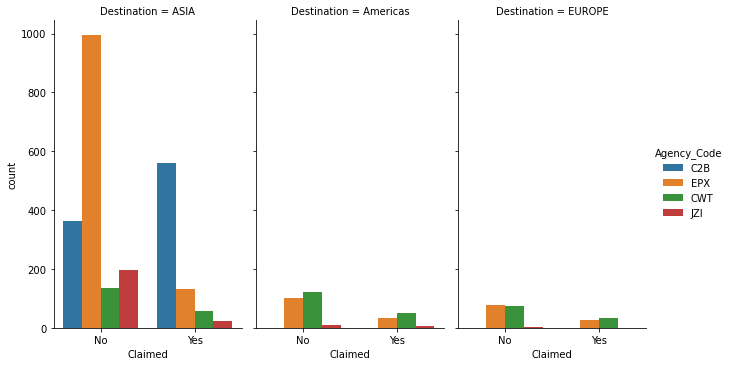


We can see observe from above graph that:-

* Agency Code 'C2B' is working for ASIA geography only. It has no sale for Ameria and Europe
* Agency code 'EPX','CWT', and 'JZI' are getting sales in all three geographies.
* Agency code 'JZI' sales figures are low compared to other agencies.



**Customised plan contains 38% of the total data. Still we can see that under Asia Geography, Bronze and Silver plan have taken more claim from insurance as compared to Customised Plan.**



**We can see that 'C2B' agency workining under ASIA demography is claiming more i.e around 580 claims and 380 no claims. It means 60% of the tours booked by this agency are getting claimed.**

We can create crosstab table for checking the exact stats agency wise

### Crosstab table between Agency\_Code and Claimed

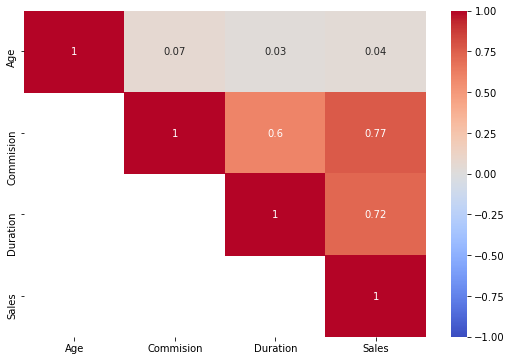
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Claimed** | **No** | **Yes** | **All** | **Claim\_percentage** |
| **Agency\_Code** |  |  |  |  |
| **C2B** | 364 | 560 | 924 | 60.61 |
| **CWT** | 331 | 141 | 472 | 29.87 |
| **EPX** | 1172 | 193 | 1365 | 14.14 |
| **JZI** | 209 | 30 | 239 | 12.55 |
| **All** | 2076 | 924 | 3000 | 30.8 |

From this crosstab table, its confirmed that 'C2B' agency has the highest claim percentage than the other and its causing total claims for the insurance firm to be on higher side

### Correlation matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Age** | **Commision** | **Duration** | **Sales** |
| **Age** | 1 | 0.07 | 0.03 | 0.04 |
| **Commision** | 0.07 | 1 | 0.6 | 0.77 |
| **Duration** | 0.03 | 0.6 | 1 | 0.72 |
| **Sales** | 0.04 | 0.77 | 0.72 | 1 |

### Heat Map

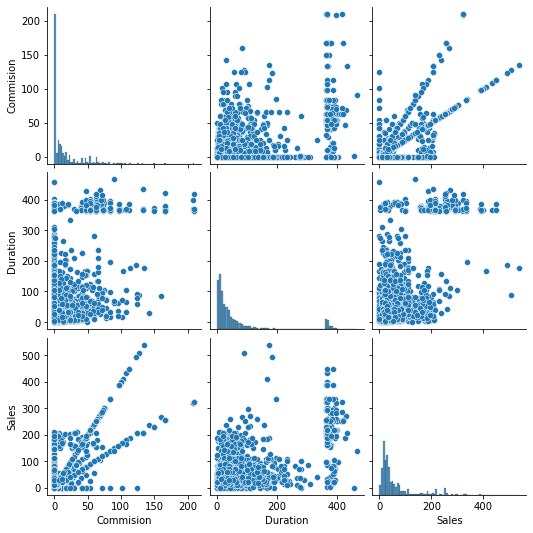
****

We can see in heatmap that

* Commission , Sales and Duration are positively correlated with each other.
* Age doesnt seems to be correlated with any other attributes.

### Pairplots

To check the correlation of 2 columns in more detail we can draw **pairplots/Scatter Diagram**



* As depicted in heatmap, we can see the positive correlation clearly between Sales and Duration.
* Between Duration and sales, there is slight correlation on positive side but there seems to be few tours in which sale is less but duration is high.

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

### Converting Object into int

Since all the three models required numeric data type as the input we need to convert/encode the categorical variables into the integers.

After encoding of object data type into integers, following coded values are obtained for various categorical data types.

|  |  |  |
| --- | --- | --- |
| **col\_name** | **actual\_value** | **coded\_value** |
| Agency\_Code | C2B | 0 |
| Agency\_Code | CWT | 1 |
| Agency\_Code | EPX | 2 |
| Agency\_Code | JZI | 3 |
| Channel | Offline | 0 |
| Channel | Online | 1 |
| Claimed | No | 0 |
| Claimed | Yes | 1 |
| Destination | ASIA | 0 |
| Destination | Americas | 1 |
| Destination | EUROPE | 2 |
| Product Name | Bronze Plan | 0 |
| Product Name | Cancellation Plan | 1 |
| Product Name | Customised Plan | 2 |
| Product Name | Gold Plan | 3 |
| Product Name | Silver Plan | 4 |
| Type | Airlines | 0 |
| Type | Travel Agency | 1 |

After encoding, sample of our input dataset is as shown below :-

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Age** | **Agency\_Code** | **Type** | **Claimed** | **Commision** | **Channel** | **Duration** | **Sales** | **Product Name** | **Destination** |
| 48 | 0 | 0 | 0 | 0.7 | 1 | 7 | 2.51 | 2 | 0 |
| 36 | 2 | 1 | 0 | 0 | 1 | 34 | 20 | 2 | 0 |
| 39 | 1 | 1 | 0 | 5.94 | 1 | 3 | 9.9 | 2 | 1 |
| 36 | 2 | 1 | 0 | 0 | 1 | 4 | 26 | 1 | 0 |
| 33 | 3 | 0 | 0 | 6.3 | 1 | 53 | 18 | 0 | 0 |

Now we can see that all values are numerical

Label Encoding has been done and all columns are converted to number

### Proportion of Categorical data type

Proportions of various categorical data type is as shown below :-

Proportion of Agency\_Code (in %)

2 46.0

0 31.0

1 16.0

3 8.0

Proportion of Type (in %)

1 61.0

0 39.0

Proportion of Claimed (in %)

0 69.0

1 31.0

Proportion of Channel (in %)

1 98.0

0 2.0

Proportion of Product Name (in %)

2 38.0

1 23.0

0 22.0

4 14.0

3 4.0

Proportion of Destination (in %)

0 82.0

1 11.0

2 7.0

### Proportion of 1s and 0s in Target Variable

No 69.0

Yes 31.0

This target variable is slightly imbalanced

### Building CART Model

### Extracting the target column into separate vectors for training set and test set

Target colum (Claimed) is extracted into a separate column before splitting the data set into training dataset and testing dataset.

### Splitting data into training and test set

After extracting the Target column, data set is split into 70:30 ratio i.e. 70% of input observations into train data set for building the model and 30% observations into test data set for testing and validating the model. This is done to keep the test data anonymous from the model and to check how our model is predicting the result of test data set.

### Checking the dimensions of the training and test data

After splitting, dimensions of training and test data sets are as shown below :-

X\_train (2100, 9) 🡪 input train

X\_test (900, 9) 🡪 input test

train\_labels (2100,)🡪 actual output train

test\_labels (900,) 🡪 actual output test

Total Observations (3000, 9)

We can see that data has been splitted into train(70%) and test(30%) successfully

### Building a Decision Tree Classifier

Initially we built the model using Gini index without giving any pruning parameters and created the dot file to visually see the decision tree.

Decision tree initially made is overgrown in the size and it is very tough to depict meaningful insight from such a long and complex decision tree.

Dot file is attached here for reference.



Since this decision tree was overgrown, different pruning parameters were added in the SKlearn GridSearchCV function to find the model with best results,

After running the grid search multiple times we have finalized upon the following parameters as the best parameters

Showing best parameters for the grid search

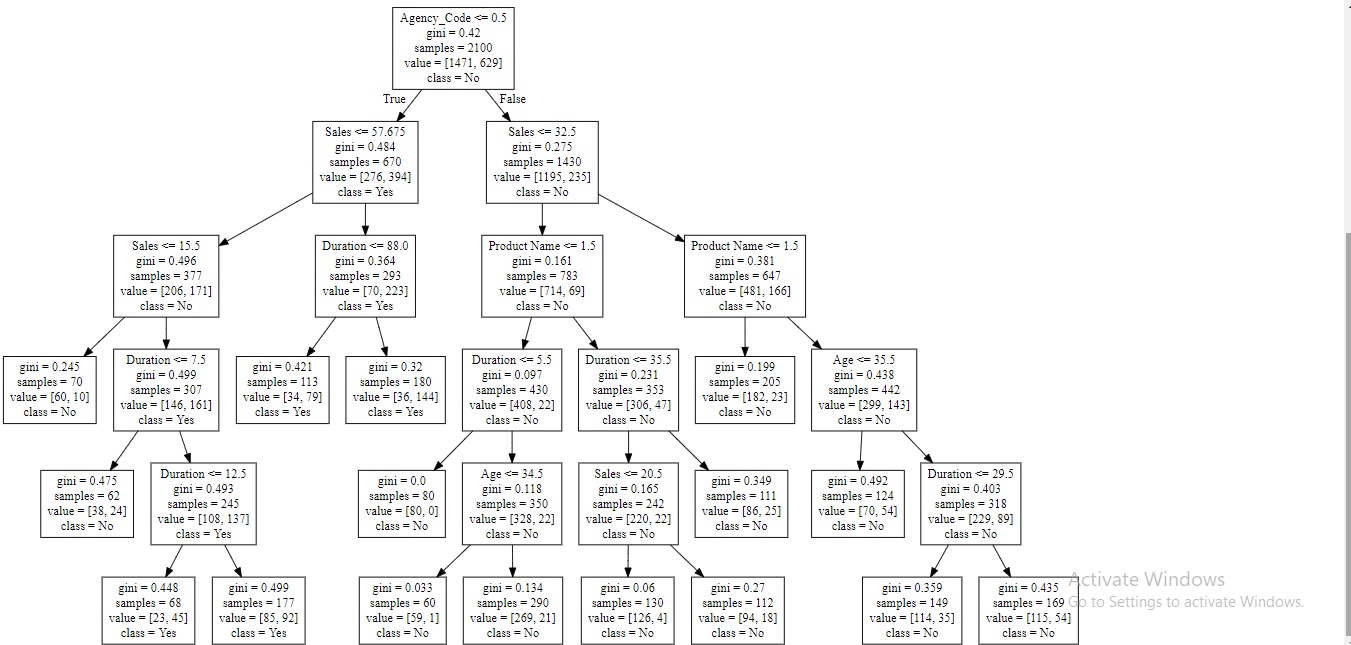
{'max\_depth': 5, 'min\_samples\_leaf': 5, 'min\_samples\_split': 15}

### Pruned Decision Tree dot file

Dot file is attached here for reference.

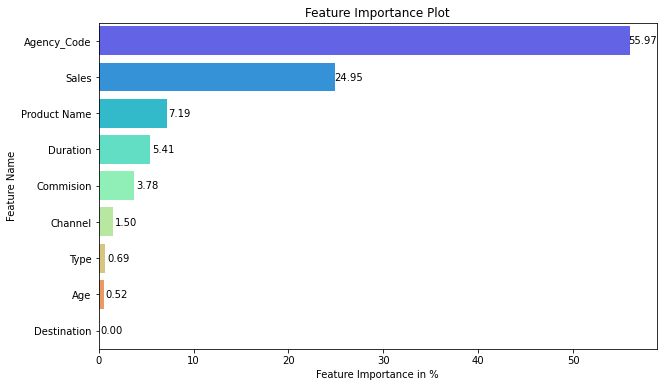


### Pruned Decision Tree visualize



### Variable Importance

After running the CART model, we can see in the graph features that are important in predicting the target variable value.



**As per this model, attributes like Type, Age and Destination have almost no importance in predicting the insurance claims. On the other hand ,most important feature is Agency\_Code followed by the Sales.**

### Model Evaluation

Now we are evaluating different ways to decide if the model is stable or not. Higher the score better will be the model.

### Model Score

This score gives the accuracy of the model.

#### Model score for the Training Data

0.8028571428571428

#### Model score for the Testing Data

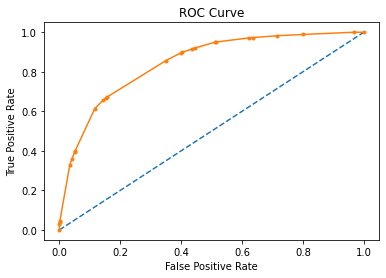
0.7733333333333333

### Measuring AUC-ROC Curve

In this evaluation technique we check the AUC (area under the curve) and ROC curve (receiver operating characteristics curve which is a graph between True Positive rate and False Positive rate)

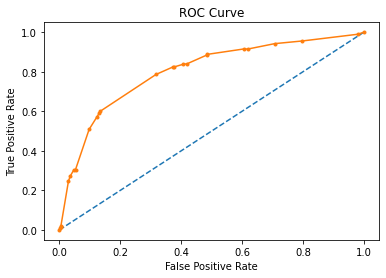
#### AUC and ROC for the training data

AUC for training data: 0.846

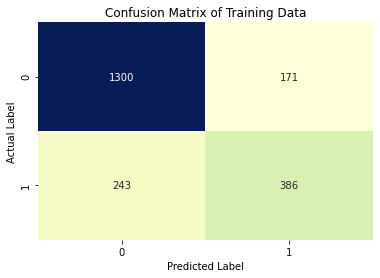


#### AUC and ROC for the testing data

AUC for testing data: 0.802



### Confusion Matrix for the Training data



### Classification Report for Training data

precision recall f1-score support

0 0.84 0.88 0.86 1471

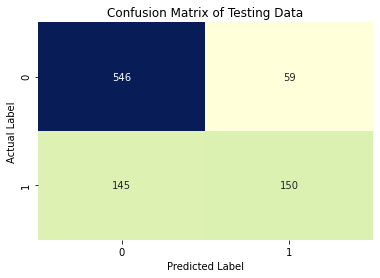
1 0.69 0.61 0.65 629

accuracy 0.80 2100

macro avg 0.77 0.75 0.76 2100

weighted avg 0.80 0.80 0.80 2100

### Confusion Matrix for the Testing data



### Classification Report for Testing data

precision recall f1-score support

0 0.79 0.90 0.84 605

1 0.72 0.51 0.60 295

accuracy 0.77 900

macro avg 0.75 0.71 0.72 900

weighted avg 0.77 0.77 0.76 900

### CART Conclusion

**Train Data:**

AUC: 84%

Accuracy: 80%

Precision: 69%

f1-Score: 65%

Recall: 61%

**Test Data:**

AUC: 80%

Accuracy: 77%

Precision: 72%

f1-Score: 60%

Recall: 51%

Except Recall, other measures for test and train are inline with each other. **Model is slighly overfitted as difference in recall in test and train is near about 10%**

With recall rate of 51%, model is only able to predict 51% of total tours which were actually claimed as claimed.

Precision is 72% of test data which means, out of total tours predicted by model as claimed , 72% were actually claimed.

F1-score is the harmonic mean of precision and recall, it takes into the effect of both the scores and this value is low if any of these 2 values are low.

**Agency Code, Sales, Product Name, Duration and Commission (in same order of preference) are the most important variables in determining if a tour will be claimed from insurance company**

Since we are building a model to predict if a tour will be claimed for insurance or not, for practical purposes, we will be more interested in correctly classifying 1 (taking insurance claim) than 0(not taking insurance claim).

If a tour not claiming by the insurer is incorrectly predicted to be claimed by the model, then the impact on cost for the insurance company would be bare minimum. But if a tour claimed by the insurer is incorrectly predicted to be not claimed by the model, then the cost impact would be very high for the insurance company. Hence recall rate (actual data point identified as True by model) is very important in this scenario.

F1 score which is dependent on recall and precision is also an important factor in scenario.

#### As Recall rate of test dataset is very poor around 50% thus this doesnt looks good enough for classification

### Ensemble Random Forest Classifier

This model also requires only numeric data type. Since we have already extracted the target column into separate column, encoded the categorical variables and splitted the data in test and train data set in the CART model (mentioned above in detail), we will directly use these variables as a input to this Random Forest classifier model.

### Building a Random Forest Classifier

Initially after building the model, grid search cross validation was run on the model multiple times to find out the best parameters for accurate results.

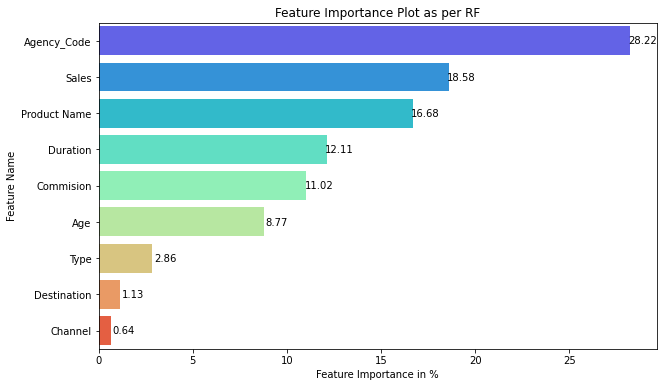
We have finalized upon the following parameters as the best parameters

Showing best parameters for the grid search

{'max\_depth': 15, 'max\_features': 5, 'min\_samples\_leaf': 2, 'min\_samples\_split': 40, 'n\_estimators': 500, 'oob\_score': True}

### Variable Importance

After running the CART model, we can see in the graph features that are important in predicting the target variable value.

****

**As per this model, attributes like Type, Destination and Channel have almost no importance in predicting the insurance claims. On the other hand ,most important feature is Agency\_Code followed by the Sales.**

### Model Evaluation

Now we are evaluating different ways to decide if the model is stable or not. Higher the score better will be the model.

### Model Score

This score gives the accuracy of the model.

#### Model score for the Training Data

0.8319047619047619

#### Model score for the Testing Data

0.7744444444444445

#### Out of Bag Score for Model

0.7871428571428571

#### Error rate

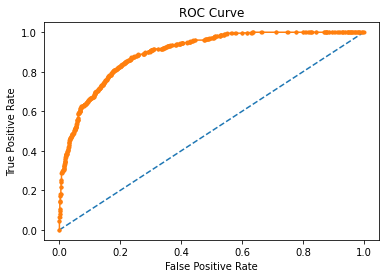
21.285714285714285%

### Measuring AUC-ROC Curve

In this evaluation technique we check the AUC (area under the curve) and ROC curve (receiver operating characteristics curve which is a graph between True Positive rate and False Positive rate)

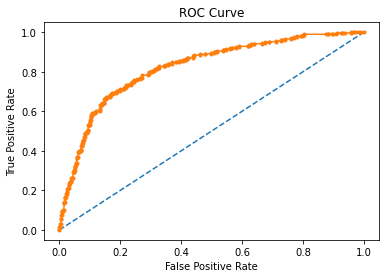
#### AUC and ROC for the training data

AUC for training data: 0.899

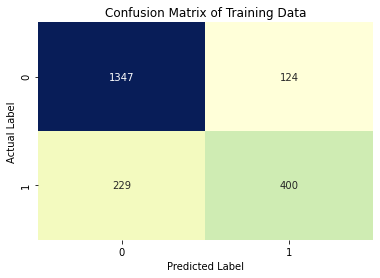


#### AUC and ROC for the testing data

AUC for testing data: 0.823



### Confusion Matrix for the Training data



### Classification Report for Training data

precision recall f1-score support

0 0.85 0.92 0.88 1471

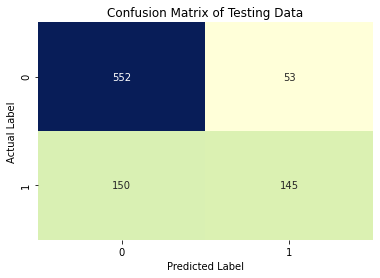
1 0.76 0.64 0.69 629

accuracy 0.83 2100

macro avg 0.81 0.78 0.79 2100

weighted avg 0.83 0.83 0.83 2100

### Confusion Matrix for the Testing data



### Classification Report for Testing data

precision recall f1-score support

0 0.79 0.91 0.84 605

1 0.73 0.49 0.59 295

accuracy 0.77 900

macro avg 0.76 0.70 0.72 900

weighted avg 0.77 0.77 0.76 900

### Random Forest Conclusion

**Train Data:**

AUC: 90%

Accuracy: 83%

Precision: 76%

f1-Score: 69%

Recall: 64%

**Test Data:**

AUC: 82%

Accuracy: 77%

Precision: 73%

f1-Score: 59%

Recall: 49%

**Its a case for overfitting where train dataset is giving better result as compared to test.**

With recall rate of 49%, model is only able to predict 49% of total tours which were actually claimed as claimed.

Precision is 73% of test data which means, out of total tours predicted by model as claimed , 73% were actually claimed.

F1-score is the harmonic mean of precision and recall, it takes into the effect of both the scores and this value is low if any of these 2 value is low.

Agency Code, Sales, Product Name, Duration and Commission (in same order of preference) are the most important variables in determining if a tour will be claimed from insurance company

Since we are building a model to predict if a tour will be claimed for insurance or not, for practical purposes, we will be more interested in correctly classifying 1 (taking insurance claim) than 0(not taking insurance claim).

If a tour not claiming by the insurer is incorrectly predicted to be claimed by the model, then the impact on cost for the insurance company would be bare minimum. But if a tour claimed by the insurer is incorrectly predicted to be not claimed by the model, then the cost impact would be very high for the insurance company. Hence recall rate (actual data point identified as True by model) is very important in this scenario.

F1 score which is dependent on recall and precision is also an important factor in scenario.

#### As Recall rate of test dataset is very poor around 50% thus this doesnt looks good enough for classification

### Artificial Neural Network Classifier

Like Random Forest and CART, this model also requires only numeric data type. Since we have already extracted the target column into separate column, encoded the categorical variables and splitted the data in test and train data set in the CART model (mentioned above in detail), we will directly use these variables as a input to this Random Forest classifier model.

### Scaling the test and train data for Neural Networks

However for ANN classifier, we need scaled input data. Thus scaling is required for ANN model. We have used Sklearn StandardScaler function to scale the test and train data. Train data is fit and transformed, however test data is only transformed (not fitted).

Scaled output for trained dataset

array([[-0.19192502, 0.72815922, 0.80520286, ..., -0.5730663 ,

0.24642411, -0.43926017],

[-0.19192502, 0.72815922, 0.80520286, ..., -0.26910565,

0.24642411, 1.27851702],

[-0.97188154, -1.28518425, -1.24192306, ..., 1.74601534,

1.83381865, -0.43926017],

...,

[-0.19192502, 0.72815922, 0.80520286, ..., 0.02103862,

0.24642411, -0.43926017],

[ 0.58803151, 1.73483096, -1.24192306, ..., -0.60069909,

-1.34097044, -0.43926017],

[-0.19192502, -1.28518425, -1.24192306, ..., -0.53852532,

1.83381865, -0.43926017]])

Scaled output for test dataset

array([[-1.55684893, -0.27851251, 0.80520286, ..., 0.18683534,

-1.34097044, 2.99629421],

[ 1.66047173, -1.28518425, -1.24192306, ..., -0.48325974,

-1.34097044, -0.43926017],

[-0.87438698, -1.28518425, -1.24192306, ..., -0.62833187,

-1.34097044, -0.43926017],

...,

[-0.19192502, -1.28518425, -1.24192306, ..., -0.47635155,

-1.34097044, -0.43926017],

[ 1.07550434, 1.73483096, -1.24192306, ..., -0.43490237,

-1.34097044, -0.43926017],

[-0.28941958, 1.73483096, -1.24192306, ..., -0.49016794,

-1.34097044, -0.43926017]])

### Building a ANN Classifier

Initially after building the model, grid search cross validation was run on the model multiple times to find out the best parameters for accurate results.

We have finalized upon the following parameters as the best parameters

Showing best parameters for the grid search

{'activation': 'relu', 'hidden\_layer\_sizes': 350, 'max\_iter': 10000, 'solver': 'adam', 'tol': 0.0001, 'verbose': True}

### Model Evaluation

Now we are evaluating different ways to decide if the model is stable or not. Higher the score better will be the model.

### Model Score

This score gives the accuracy of the model.

#### Model score for the Training Data

0.8166666666666667

#### Model score for the Testing Data

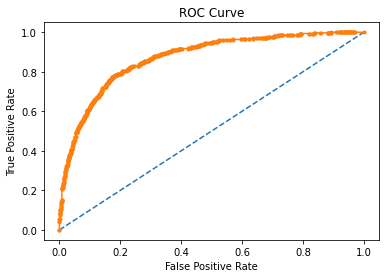
0.7722222222222223

### Measuring AUC-ROC Curve

In this evaluation technique we check the AUC (area under the curve) and ROC curve (receiver operating characteristics curve which is a graph between True Positive rate and False Positive rate)

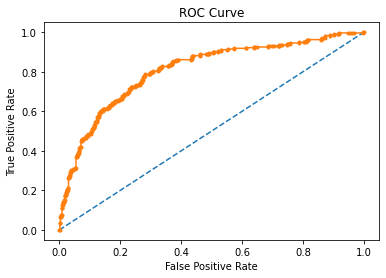
#### AUC and ROC for the training data

AUC for training data: 0.871

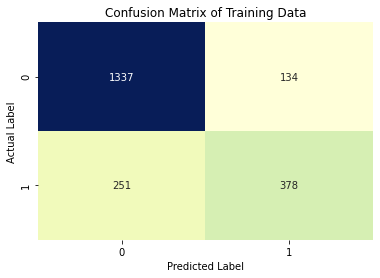


#### AUC and ROC for the testing data

AUC for testing data: 0.809



### Confusion Matrix for the Training data



### Classification Report for Training data

precision recall f1-score support

0 0.84 0.91 0.87 1471

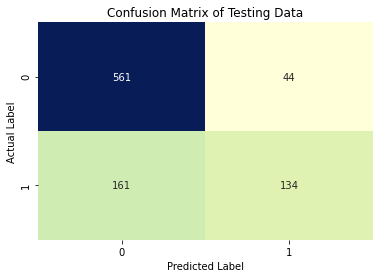
1 0.74 0.60 0.66 629

accuracy 0.82 2100

macro avg 0.79 0.75 0.77 2100

weighted avg 0.81 0.82 0.81 2100

### Confusion Matrix for the Testing data



### Classification Report for Testing data

precision recall f1-score support

0 0.78 0.93 0.85 605

1 0.75 0.45 0.57 295

accuracy 0.77 900

macro avg 0.76 0.69 0.71 900

weighted avg 0.77 0.77 0.75 900

### Artificial Neuro Network Conclusion

**Train Data:**

AUC: 87%

Accuracy: 82%

Precision: 74%

f1-Score: 66%

Recall: 60%

**Test Data:**

AUC: 81%

Accuracy: 77%

Precision: 75%

f1-Score: 57%

Recall: 45%

**Its a case for overfitting where train dataset is giving better result as compared to test.**

With recall rate of 45%, model is only able to predict 45% of total tours which were actually claimed as claimed.

Precision is 75% of test data which means, out of total tours predicted by model as claimed , 75% were actually claimed.

F1-score is the harmonic mean of precision and recall, it takes into the effect of both the scores and this value is low if any of these 2 value is low.

Since we are building a model to predict if a tour will be claimed for insurance or not, for practical purposes, we will be more interested in correctly classifying 1 (taking insurance claim) than 0(not taking insurance claim).

If a tour not claiming by the insurer is incorrectly predicted to be claimed by the model, then the impact on cost for the insurance company would be bare minimum. But if a tour claimed by the insurer is incorrectly predicted to be not claimed by the model, then the cost impact would be very high for the insurance company. Hence recall rate (actual data point identified as True by model) is very important in this scenario.

F1 score which is dependent on recall and precision is also an important factor in scenario.

#### As Recall rate of test dataset is very poor around 45% thus this doesn’t looks good enough for classification

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

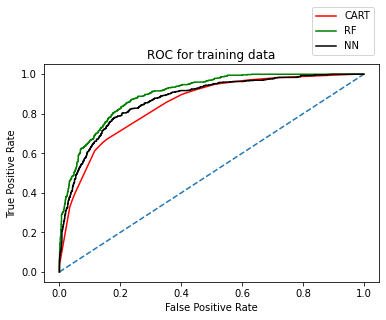
### Comparison of the performance metrics from the 3 models[¶](http://localhost:8889/notebooks/Scripts/Data%20Mining/Project/Raghav_Gupta_Data_Mining_Project_21-03-2021.ipynb#Comparison-of-the-performance-metrics-from-the-3-models)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | AUC | Recall | Precision | F1 Score |
| CART Train | 0.8 | 0.85 | 0.61 | 0.69 | 0.65 |
| CART Test | 0.77 | 0.8 | 0.51 | 0.72 | 0.6 |
| Random Forest Train | 0.83 | 0.9 | 0.64 | 0.76 | 0.69 |
| Random Forest Test | 0.77 | 0.82 | 0.49 | 0.73 | 0.59 |
| Neural Network Train | 0.82 | 0.87 | 0.6 | 0.74 | 0.66 |
| Neural Network Test | 0.77 | 0.81 | 0.45 | 0.75 | 0.57 |

We can see following points :-

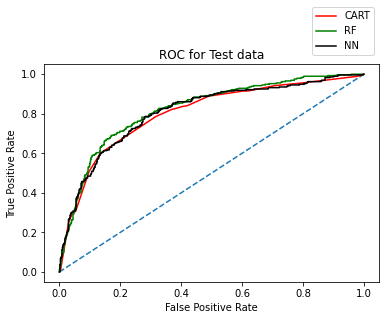
* AUC value is highest for model Random Forest for test and train data.
* Accuracy is also high for model Random Forest for test and train data.
* F1 score of Random forest is high for Train data set but for test data set CART has the highest.
* Recall for train data set is higher than 10% as compared to recall for test data set in all the three dataset. Hence it is a case of overfitting for all three models.

### ROC Curve for the 3 models on the Training data



We can see that area under the curve is more for Random Forest for training data

### ROC Curve for the 3 models on the Test data



We can see that area under the curve is slightly more for Random Forest for test data

### Conclusion

We are building a model to predict if a tour will be claimed for insurance or not, for practical purposes, we will be more interested in correctly classifying taking insurance claim than not taking insurance claim.

Hence recall rate (actual data point identified as True by model) is very important in this scenario. F1 score which is dependent on recall and precision is also an important factor in scenario.

**In all the three models, train data set is showing better result than the test data set. So it’s a case of over fitting in all three models. But still, the model is useful only in predicting class 0 (tour not getting claimed), and not class 1 (tour getting claimed). We need more sample data to improve the results as the dataset is imbalanced.**

AUC is more than 80% for all three models. Hence AUC is not a problem in any model. Accuracy is also showing similar result for all three models in test data set.

**So we are more interested in Recall and F1 score for comparing three models.**

**Out of three models, all three are showing similar results. However CART is slightly giving better F1 score and recall rate than other 2 models. Hence its performance is slightly better than the other 2 models.**

Overall in all the 3 models recall rate is coming as 45-50% which is very low. So all three models are providing 45-50% chance of correct predictions of tour getting actually claimed as claimed which doesn’t look good enough.

**From Cart and Random Forest Model, the variable Agency code is found to be the most useful feature amongst all other features for predicting whether tour will be claimed or not.**

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

We have tried to run the three different models (CART/ Random Forest/ Artificial Neuro Network) for predicting whether tour is getting claimed or not for an insurance firm.

Based on the reports and analysis done it was found that all three models were not good enough to predict tour getting claimed (class 1). The model is useful only in predicting tour not getting claimed (class 0) for which we are getting good results in all three models.

Training data set is giving better result as compared to test dataset. There is an imbalance in the Target dataset (30:70 ratio with claimed tour have 30% of observations).

So our recommendation to the business is as shown below:-

* In order to further improve the predictive model results for finding the tours which will be claimed in future more accurately, more data sample is required.
* Current model is useful to predict when tours are not getting claimed with more than 90% accuracy.
* During analysis, attribute Agency code is found to be the most useful feature amongst all other features for predicting whether tour will be claimed or not.
* It was found out during exploratory data analysis and in decision tree that around 60% of the tours booked by agency code ‘C2B’ are getting claimed. This percentage is quite high than normal (30%) claims.
* So business can check with the C2B agency and find out the reasons for getting higher claims. If these reasons are genuine then business can increase the premium amount of insurance for the tours booked by C2B agency to cover up their losses.
* There were few observations found in which duration of the tour was 0 or negative. Duration cannot be negative or zero for any tour. So business can check the source of data whether correct information is flowing from that or not.
* It was found out during exploratory data analysis that Customised plan contains the most observations (38% of the total observations). However we can see that under Asia geography, Bronze and Silver plan have taken more claim from insurance as compared to Customised Plan. So business can check why these plans are getting more claims in Asia geography.