

Machine Learning Project Part2 – CART, Random Forest, Artificial Neural Network

By Ramprasad Mohan



Travel Insurance Claim Prediction

Table of Contents

1	Project Objective.....	3
2	Data dictionary.....	3
3	Step by step approach	3
4	Solution.....	4
4.1	Qn 1.....	4
4.2	Qn 2 a).....	9
4.3	Qn 2 b)	9
4.4	Qn 3.....	22
4.5	Qn 4.....	31
4.6	Qn 5.....	32
5	Conclusion.....	36
6	Appendix A – Source Code.....	36

1 Project Objective

The project objective is to solve the business problem of an Insurance firm providing tour insurance which is facing higher claim frequency. The management decides to collect data from the past few years. We are assigned the task to make a model which predicts the claim status and provide recommendations to management. Using CART, RF & ANN and comparing the models' performances in train and test sets we have to draw inferences.

Managerial Report is to be prepared.

2 Data Dictionary

Below are the ten variables or columns in the data set and its explanation:

- Age : Age of the insured customer
- Agency_Code: Code of tour firm
- Type: Type of tour insurance firms
- Claimed: Target or the dependent variable :Claim Status
- Channel: Distribution channel of tour insurance agencies
- Commission: The commission received for tour insurance
- Sales : Amount of sales of tour insurance policies
- Product : Name of the tour insurance products
- Duration : Duration of the tour
- Destination : Destination of the tour

3 Step by Step approach

We shall follow step by step approach to arrive to the conclusion as follows:

- Exploratory Data Analysis, descriptive statistics
- Splitting the data to training and testing data
- Build CART model using training set
- Predict using testing set and evaluate the model
- Build Random Forest model using training set
- Predict using testing set and evaluate the model
- Build ANN model using training set
- Predict using testing set and evaluate the model
- Compare the three models based on the accuracy to find which model is best suited.
- Conclusion Visuals and Recommendations

4 Solution

4.1 Qn 1 : Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it

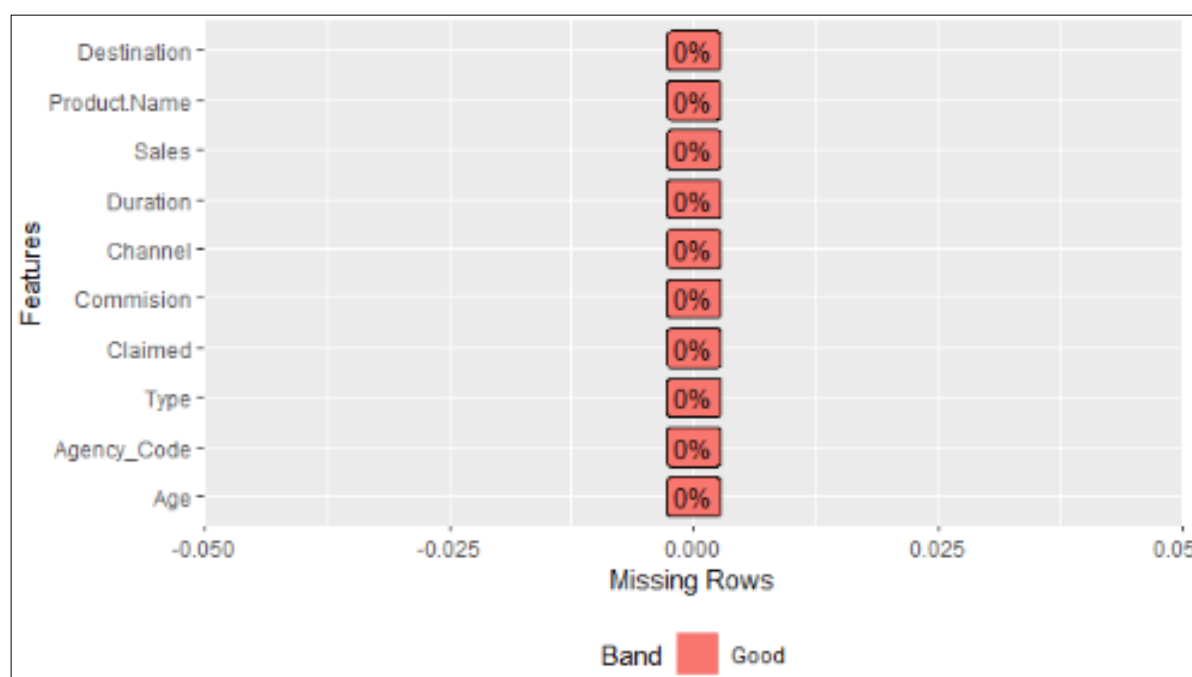
After importing the data, we are checking the top five rows of the insurance data. The insurance data has 3000 observations and 10 variables (The variable names are available in Data dictionary section)

Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product.Name	Destination
48	C2B	Airlines	No	0.70	Online	7	2.51	Customised Plan	ASIA
36	EPX	Travel Agency	No	0.00	Online	34	20.00	Customised Plan	ASIA
39	CWT	Travel Agency	No	5.94	Online	3	9.90	Customised Plan	Americas
36	EPX	Travel Agency	No	0.00	Online	4	26.00	Cancellation Plan	ASIA
33	JZI	Airlines	No	6.30	Online	53	18.00	Bronze Plan	ASIA
45	JZI	Airlines	Yes	15.75	Online	8	45.00	Bronze Plan	ASIA


Our target variable “Claimed” is of type character, for our machine learning models the target variable must be converted to either integer (1’s and 0’s) or factors. Hence after converting the target variable as factor below is the structure of the dataset.

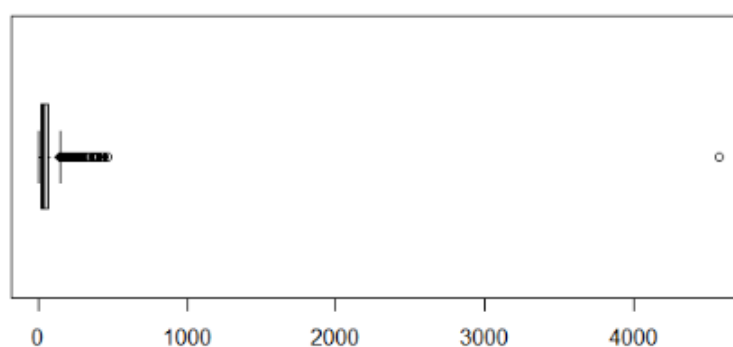
```
'data.frame': 3000 obs. of 10 variables:
 $ Age      : int  48 36 39 36 33 45 61 36 36 36 ...
 $ Agency_Code : chr  "C2B" "EPX" "CWT" "EPX" ...
 $ Type      : chr  "Airlines" "Travel Agency" "Travel Agency" "Travel Agency" ...
 $ Claimed   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 1 1 ...
 $ Commision : num  0.7 0 5.94 0 6.3 ...
 $ Channel   : chr  "Online" "Online" "Online" "Online" ...
 $ Duration  : int  7 34 3 4 53 8 30 16 19 42 ...
 $ Sales     : num  2.51 20 9.9 26 18 45 59.4 80 14 43 ...
 $ Product.Name: chr  "Customised Plan" "Customised Plan" "Customised Plan" "Cancellation Plan" ...
 $ Destination : chr  "ASIA" "ASIA" "Americas" "ASIA" ...
```

Next we are checking for any missing or null values in the dataset using the Data Explorer package, we could see that there are no null or missing values in the data and the dataset is fine for model building.

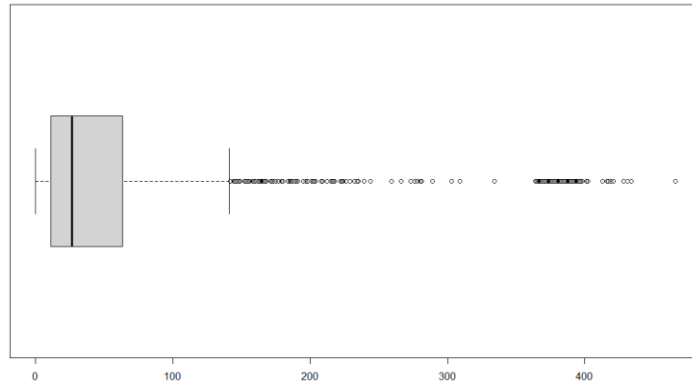


Next we are checking for outliers using the boxplot and summary tools function, we could see that the 'Duration' field has very extreme outlier record of value 4580 and an infeasible value of -1 (Duration of travel cannot be -1) which will affect our model accuracy or functioning. Below is the output of the field and box plot showing outlier.

7	Duration [integer]	Mean (sd) : 70 (134.1) min < med < max: -1 < 26.5 < 4580 IQR (CV) : 52 (1.9)		257 distinct values				3000 (100%)	0 (0%)
53	C2B	Airlines	Yes	64.8	Online	396	259.2	Silver Plan	ASIA
44	CWT	Travel Agency	Yes	208	Online	397	320	Gold Plan	Americas
35	C2B	Airlines	Yes	72.94	Online	397	291.75	Silver Plan	ASIA
55	C2B	Airlines	Yes	63.21	Online	398	252.85	Silver Plan	ASIA
59	C2B	Airlines	Yes	48.3	Online	398	193.2	Silver Plan	ASIA
26	C2B	Airlines	Yes	70.2	Online	398	280.8	Silver Plan	ASIA
27	C2B	Airlines	Yes	54	Online	401	216	Silver Plan	ASIA
28	C2B	Airlines	No	63.21	Online	401	252.85	Silver Plan	ASIA
31	CWT	Travel Agency	No	0	Offline	402	97	Customised	ASIA
31	C2B	Airlines	Yes	63.21	Online	413	252.85	Silver Plan	ASIA
27	C2B	Airlines	Yes	71.85	Online	416	287.4	Gold Plan	ASIA
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
34	CWT	Travel Agency	No	166.53	Online	421	256.2	Gold Plan	Americas
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
42	CWT	Travel Agency	No	132.99	Online	434	204.6	Gold Plan	ASIA
64	CWT	Travel Agency	No	90.09	Online	466	138.6	Silver Plan	ASIA
48	C2B	Airlines	No	0.09	Online	4580	0.32	Customised	ASIA




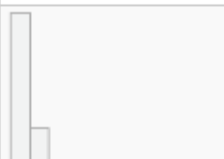
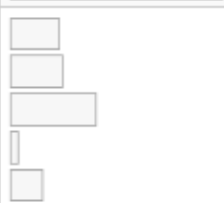
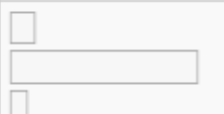
After removing the outlier value using R codes , the records are reduced to 2998 observations and below are the boxplot output after removing the extreme outliers. After removing there are still many outliers but if we remove them the number of records decreases and will affect the training and testing record numbers as well as the accuracy of the model building. Hence we are ignoring the below outlier values.



We are checking the summary if the dataset processed after removing outliers and below is the report generated.

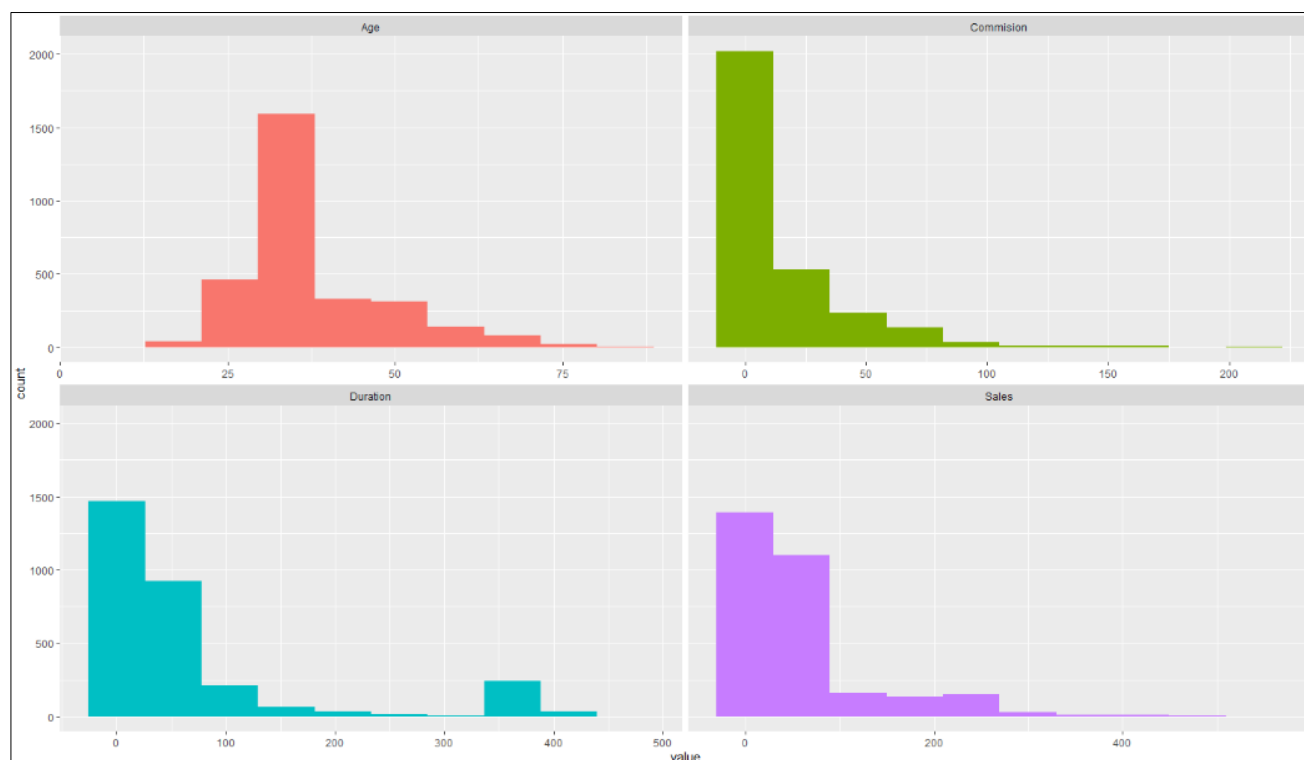
Dimensions: 2998 x 10
Duplicates: 139

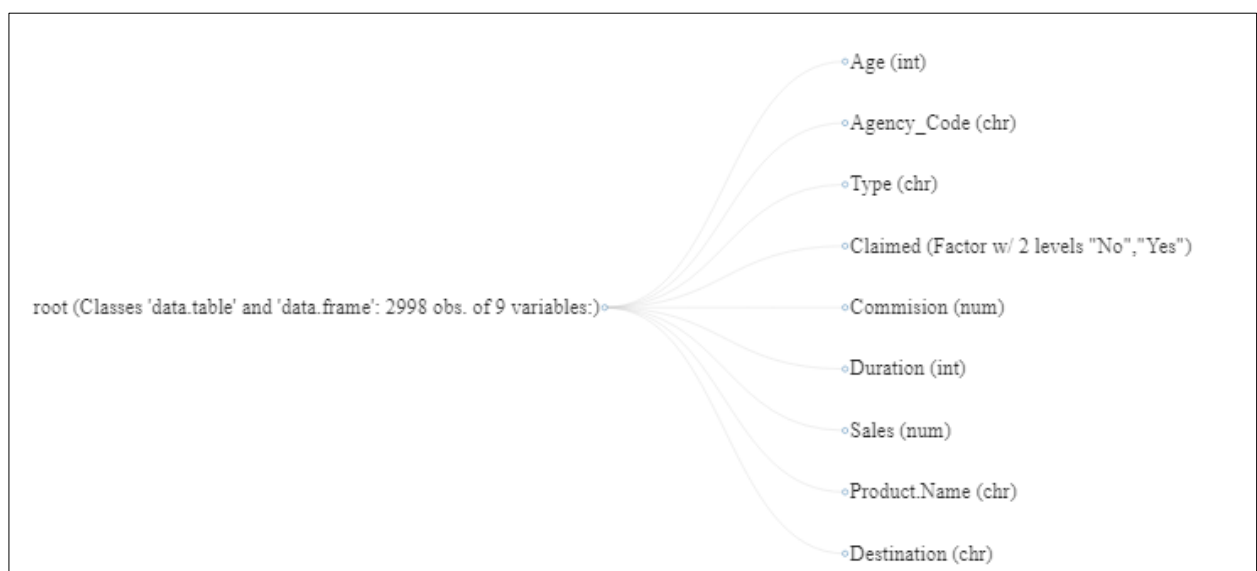
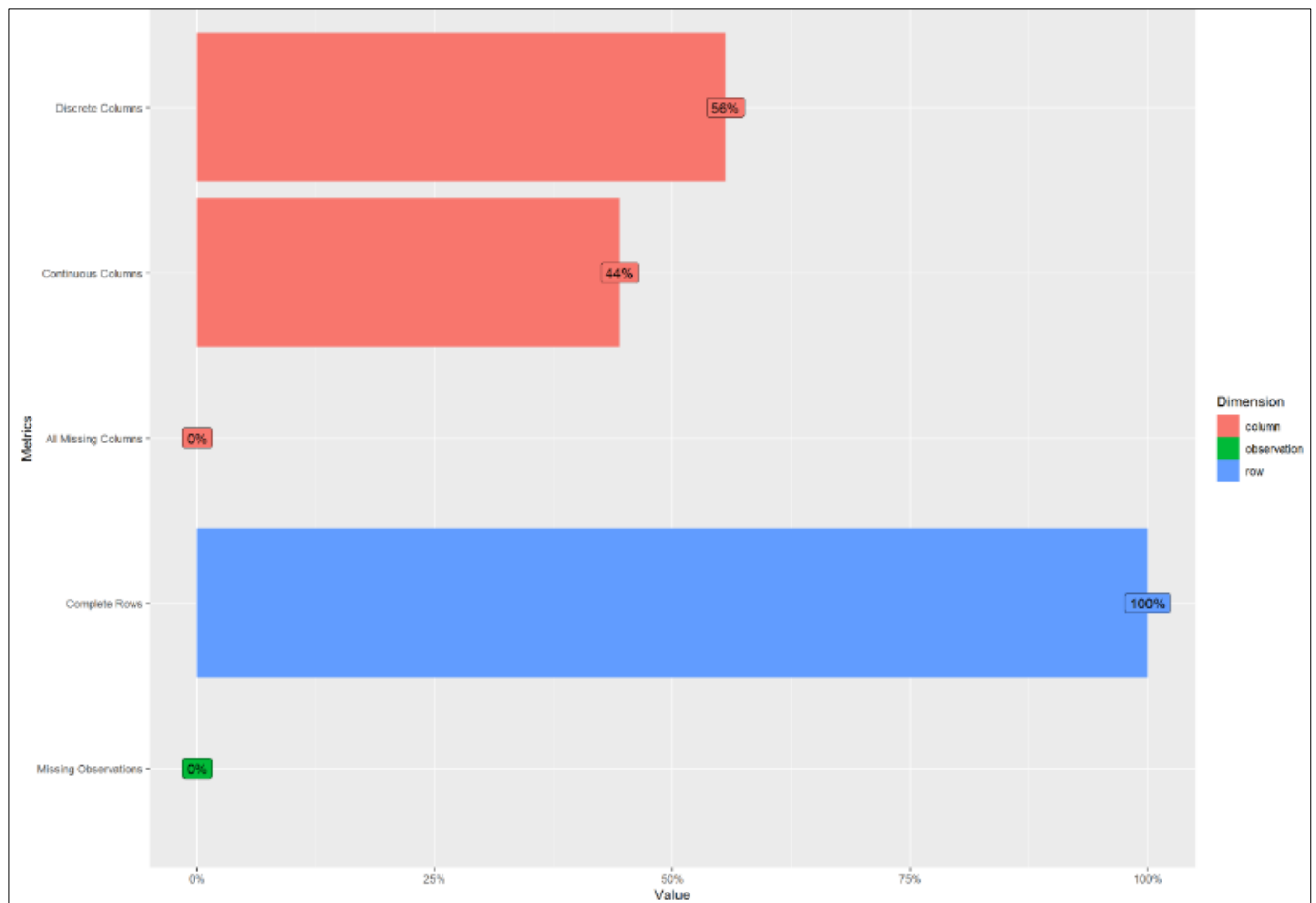
No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	Age [integer]	Mean (sd) : 38.1 (10.5) min < med < max: 8 < 36 < 84 IQR (CV) : 10 (0.3)	70 distinct values		2998 (100%)	0 (0%)
2	Agency_Code [character]	1. C2B 2. CWT 3. EPX 4. JZI	923 (30.8%) 472 (15.7%) 1365 (45.5%) 238 (7.9%)		2998 (100%)	0 (0%)
3	Type [character]	1. Airlines 2. Travel Agency	1161 (38.7%) 1837 (61.3%)		2998 (100%)	0 (0%)
4	Claimed [factor]	1. No 2. Yes	2074 (69.2%) 924 (30.8%)		2998 (100%)	0 (0%)
5	Commision [numeric]	Mean (sd) : 14.5 (25.5) min < med < max: 0 < 4.6 < 210.2 IQR (CV) : 17.2 (1.8)	323 distinct values		2998 (100%)	0 (0%)
6	Channel [character]	1. Offline 2. Online	46 (1.5%) 2952 (98.5%)		2998 (100%)	0 (0%)

7	Duration [integer]	Mean (sd) : 68.5 (105.8) min < med < max: 0 < 26.5 < 466 IQR (CV) : 52 (1.5)	255 distinct values		2998 (100%)	0 (0%)
8	Sales [numeric]	Mean (sd) : 60.3 (70.7) min < med < max: 0 < 33 < 539 IQR (CV) : 49 (1.2)	379 distinct values		2998 (100%)	0 (0%)
9	Product.Name [character]	1. Bronze Plan 2. Cancellation Plan 3. Customised Plan 4. Gold Plan 5. Silver Plan	649 (21.6%) 678 (22.6%) 1135 (37.9%) 109 (3.6%) 427 (14.2%)		2998 (100%)	0 (0%)
10	Destination [character]	1. Americas 2. ASIA 3. EUROPE	320 (10.7%) 2463 (82.2%) 215 (7.2%)		2998 (100%)	0 (0%)

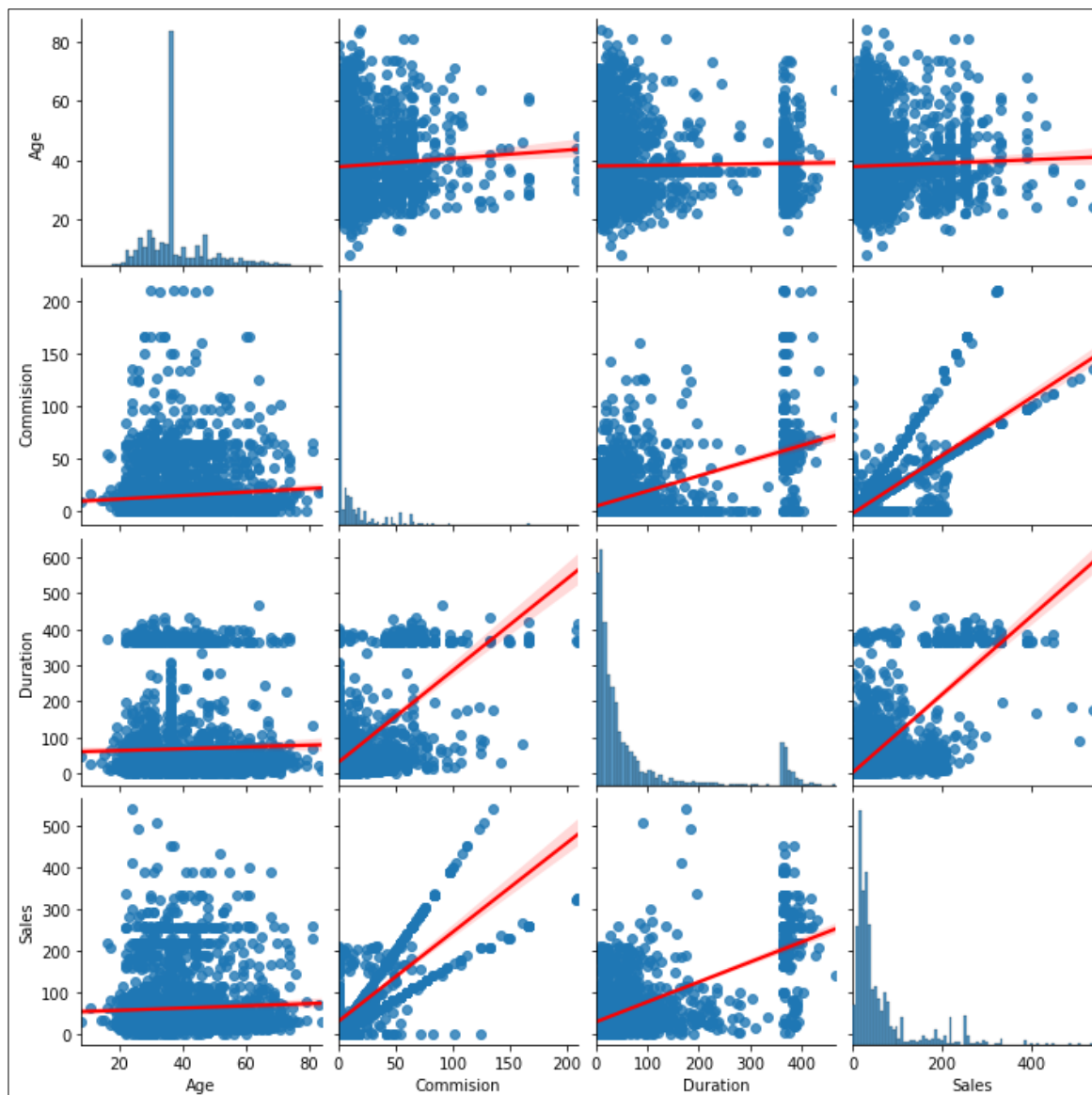
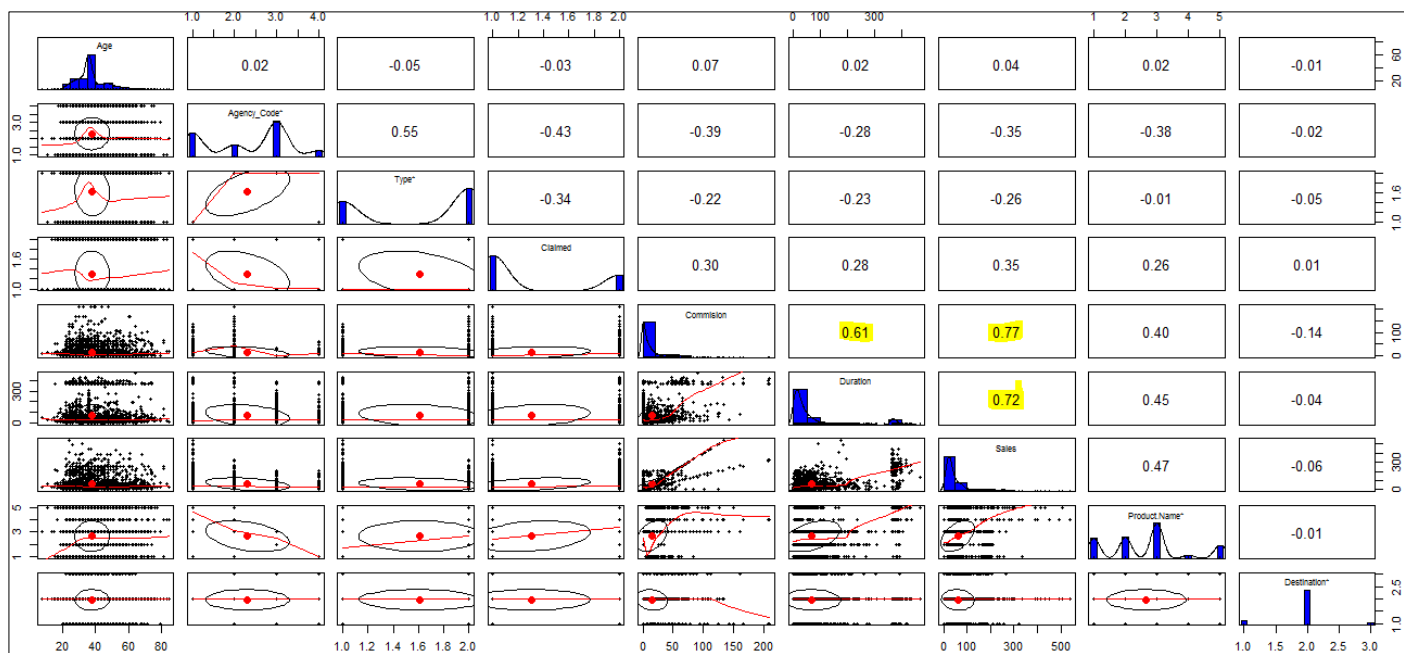
We could see that the 'Channel' field or column is highly insignificant as there are only 46 rows (1.5% of total records) having offline channel value hence removing this column will not have significance or affect our model building activity and model output. Hence we would be removing this column.

Below is the histograms of the continuous variables, we could see that the three columns are highly left skewed. The age and commission columns has extreme frequency count of values on the distribution.





Below is the correlation plot for each combination and the density plot for the continuous variables done using `pairs.panels()` function. We could see that the duration, commission and sales have high correlation coefficient values compared to others.



4.2 Qn 2 a) : Data Split: Split the data into test and train

Next step is to split the data to training and testing data sets using the rsample package. By default we will split the proportions as 70% training dataset and 30% testing data set. Below are the first five rows of the randomly sampled training and test sets.

Top five rows of training set:

	Age	Agency_Code	Type	Claimed	Commision	Duration	Sales	Product.Name	Destination
1	48	C2B	Airlines	No	0.70	7	2.51	Customised Plan	ASIA
2	36	EPX	Travel Agency	No	0.00	34	20.00	Customised Plan	ASIA
3	39	CWT	Travel Agency	No	5.94	3	9.90	Customised Plan	Americas
4	36	EPX	Travel Agency	No	0.00	4	26.00	Cancellation Plan	ASIA
5	33	JZI	Airlines	No	6.30	53	18.00	Bronze Plan	ASIA
6	45	JZI	Airlines	Yes	15.75	8	45.00	Bronze Plan	ASIA

Top five rows of testing set :

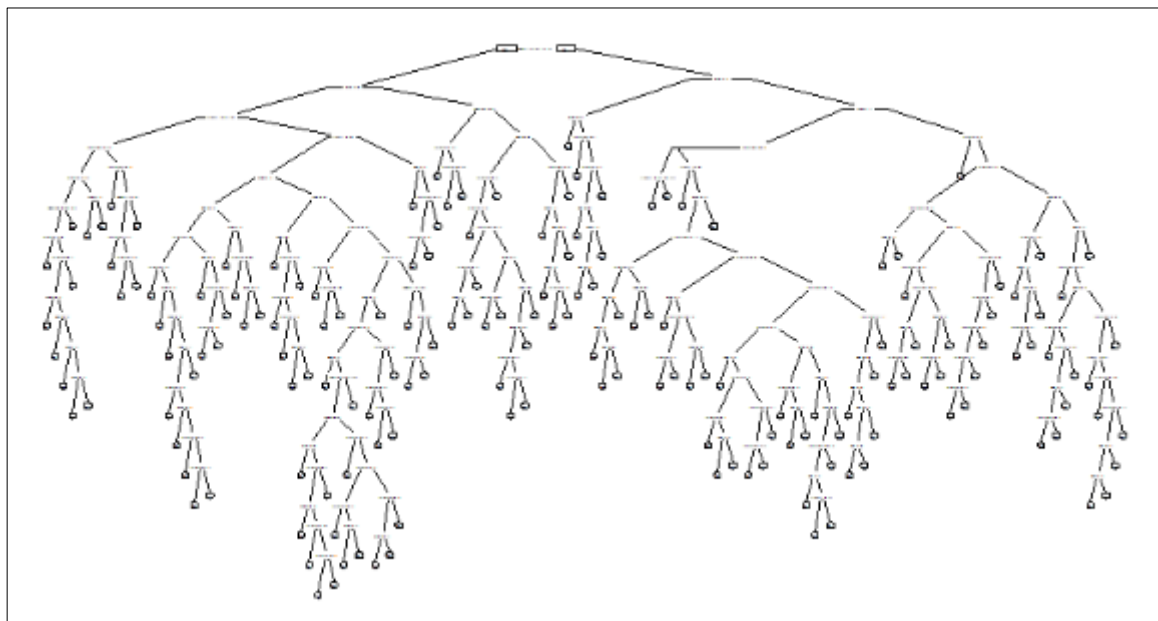
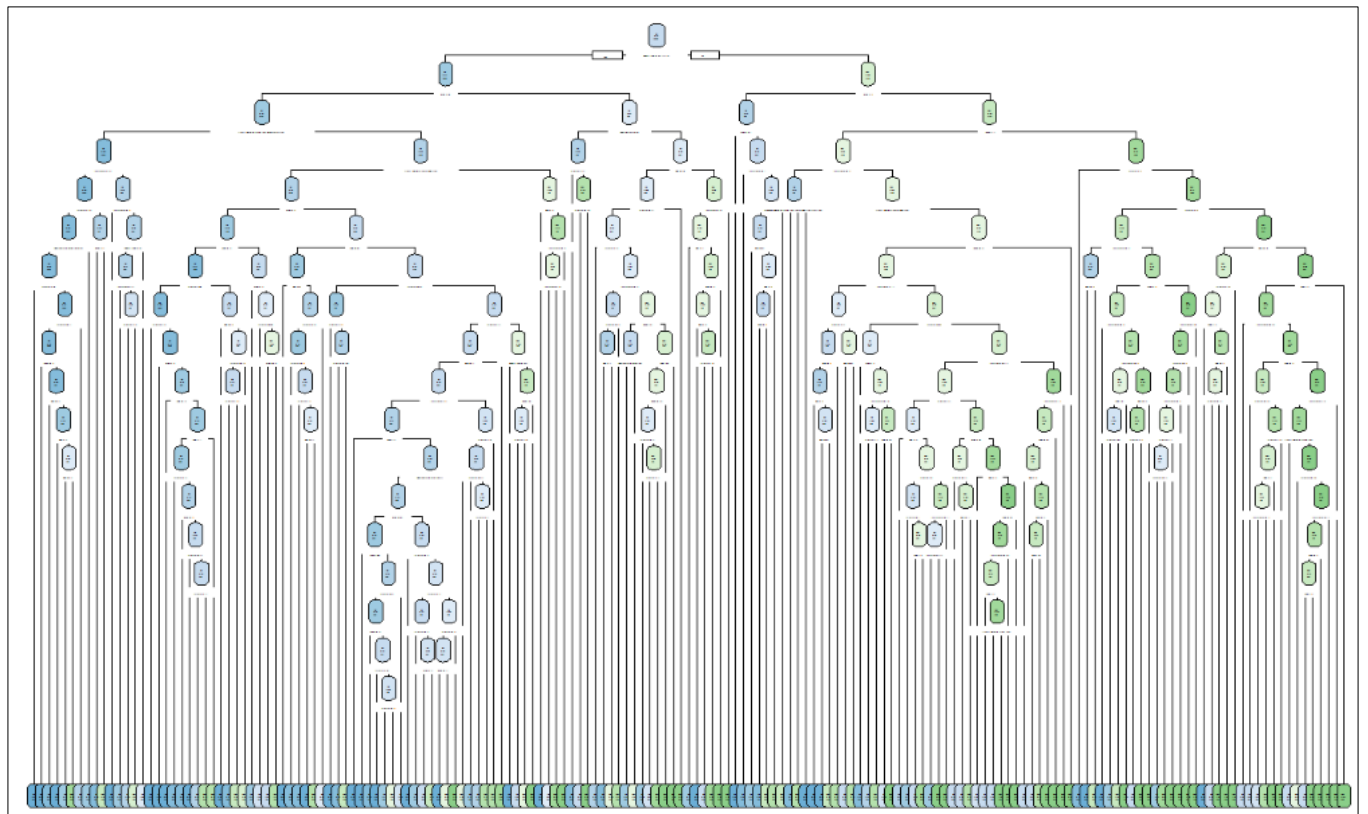
	Age	Agency_Code	Type	Claimed	Commision	Duration	Sales	Product.Name	Destination
7	61	CWT	Travel Agency	No	35.64	30	59.40	Customised Plan	Americas
9	36	EPX	Travel Agency	No	0.00	19	14.00	Cancellation Plan	ASIA
11	37	C2B	Airlines	Yes	46.96	368	187.85	Silver Plan	ASIA
13	36	EPX	Travel Agency	No	0.00	23	110.00	Customised Plan	EUROPE
15	31	CWT	Travel Agency	No	23.76	21	39.60	Customised Plan	ASIA
16	39	C2B	Airlines	Yes	54.00	366	216.00	Silver Plan	ASIA

4.3 Qn 2 b) : Build classification model CART, Random Forest and Artificial Neural Network

CART MODEL:

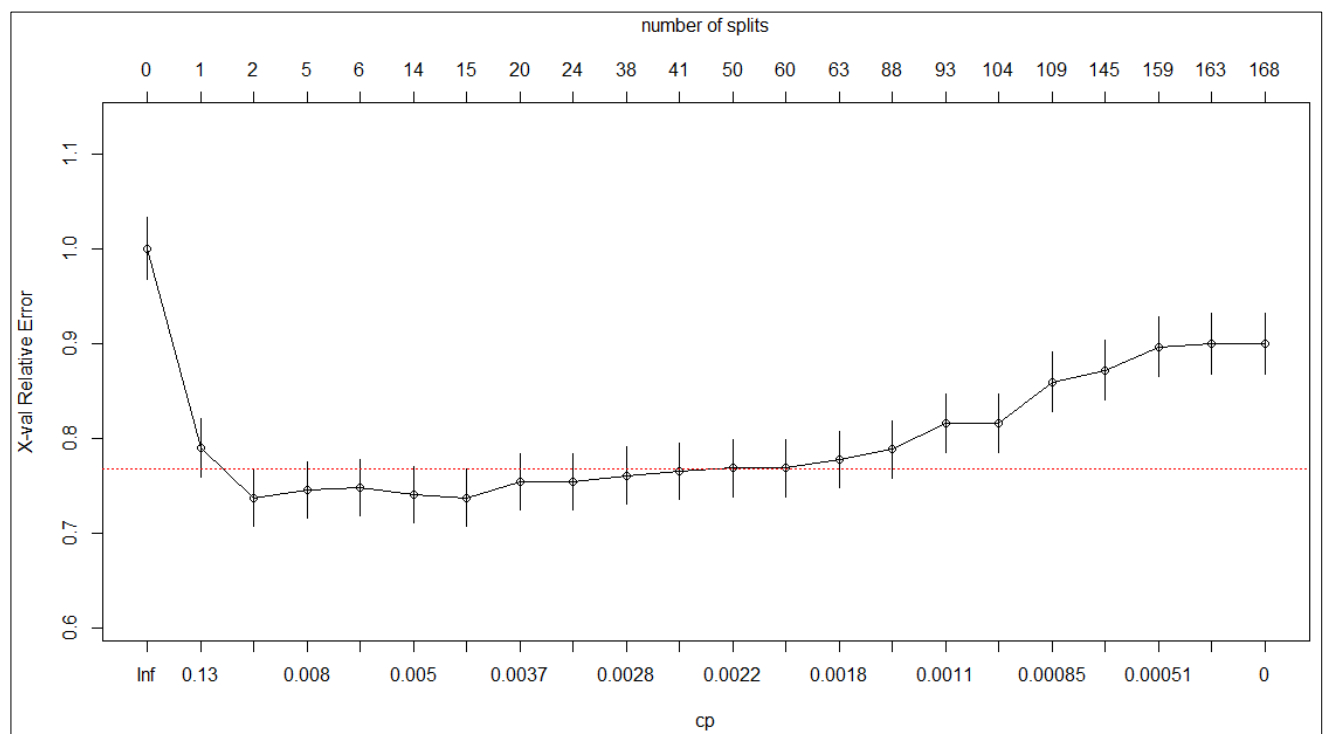
A Classification and Regression Tree (CART), is a predictive model, which explains how an outcome variable's values can be predicted based on other values. A CART output is a decision tree where each fork is a split in a predictor variable and each end node contains a prediction for the outcome variable

First we are building the cart model with the training dataset using the rpart package. We are providing the control parameters minsplit value as 9 which is equal and nearer to the number of variables. Next we are providing the minbucket value as round of minsplit value divided by three. We provide the cross validation value as 5. Once running the CART model below is the output of the random forest. We could see that the tree is large with many unwanted branches which shows that the post pruning technique must be applied. Also the first split happens with Agency code since the Gini gain for this variable is the highest.

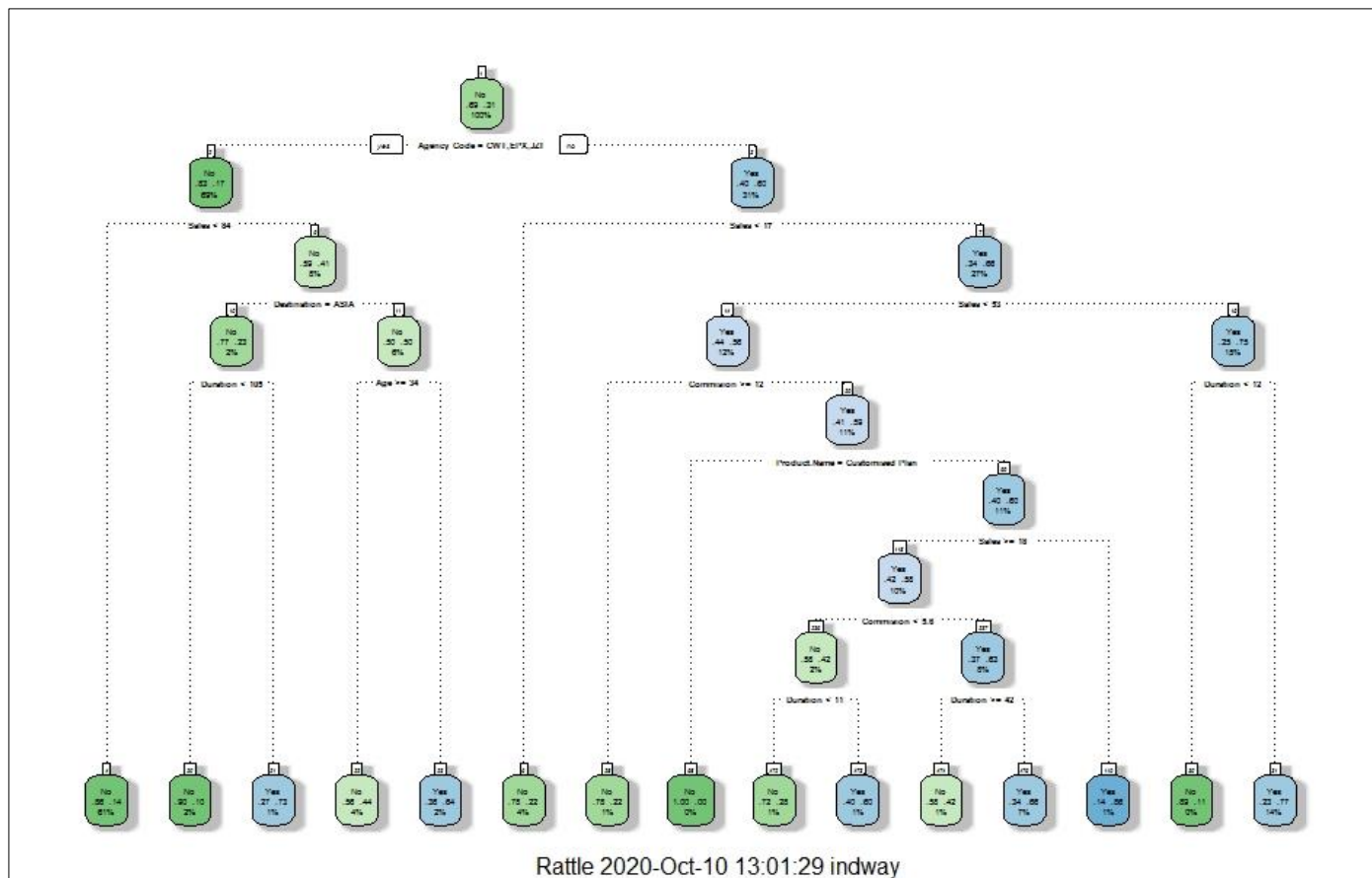


Next we print and plot the complexity parameter which helps us to decide the post pruning parameter.

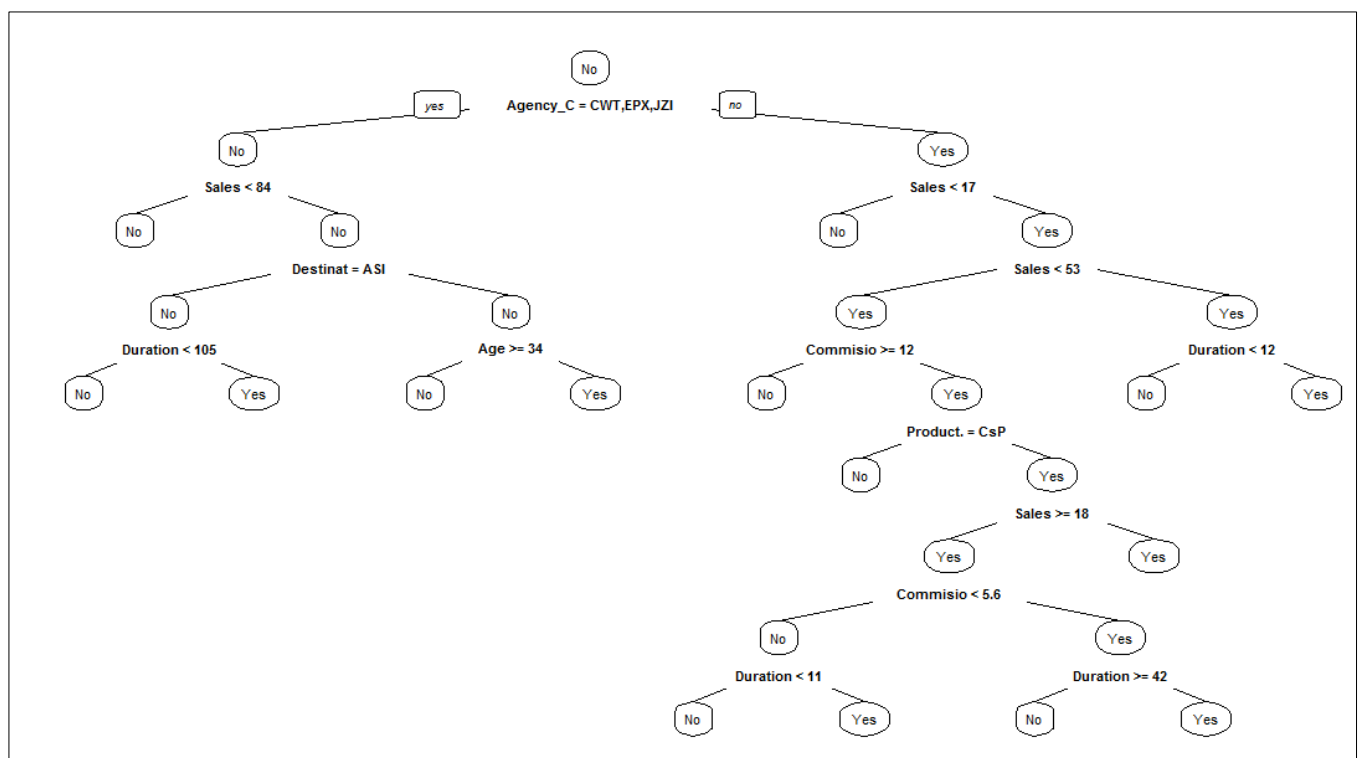
	CP	nsplit	rel error	xerror	xstd
1	0.21020093	0	1.00000	1.00000	0.032698
2	0.07727975	1	0.78980	0.78980	0.030390
3	0.01030397	2	0.71252	0.72179	0.029452
4	0.00618238	5	0.68161	0.73107	0.029586
5	0.00540958	6	0.67543	0.72643	0.029519
6	0.00463679	14	0.63060	0.72798	0.029541
7	0.00386399	15	0.62597	0.72643	0.029519
8	0.00360639	20	0.60433	0.73725	0.029674
9	0.00309119	24	0.58887	0.73879	0.029696
10	0.00257599	38	0.54405	0.73879	0.029696
11	0.00231839	41	0.53632	0.74189	0.029739
12	0.00216383	50	0.51468	0.74189	0.029739
13	0.00206079	60	0.48686	0.74343	0.029761
14	0.00154560	63	0.48068	0.81762	0.030744
15	0.00123648	88	0.43740	0.81607	0.030725
16	0.00103040	93	0.43122	0.83771	0.030991
17	0.00092736	104	0.41886	0.83771	0.030991
18	0.00077280	109	0.41422	0.86708	0.031336

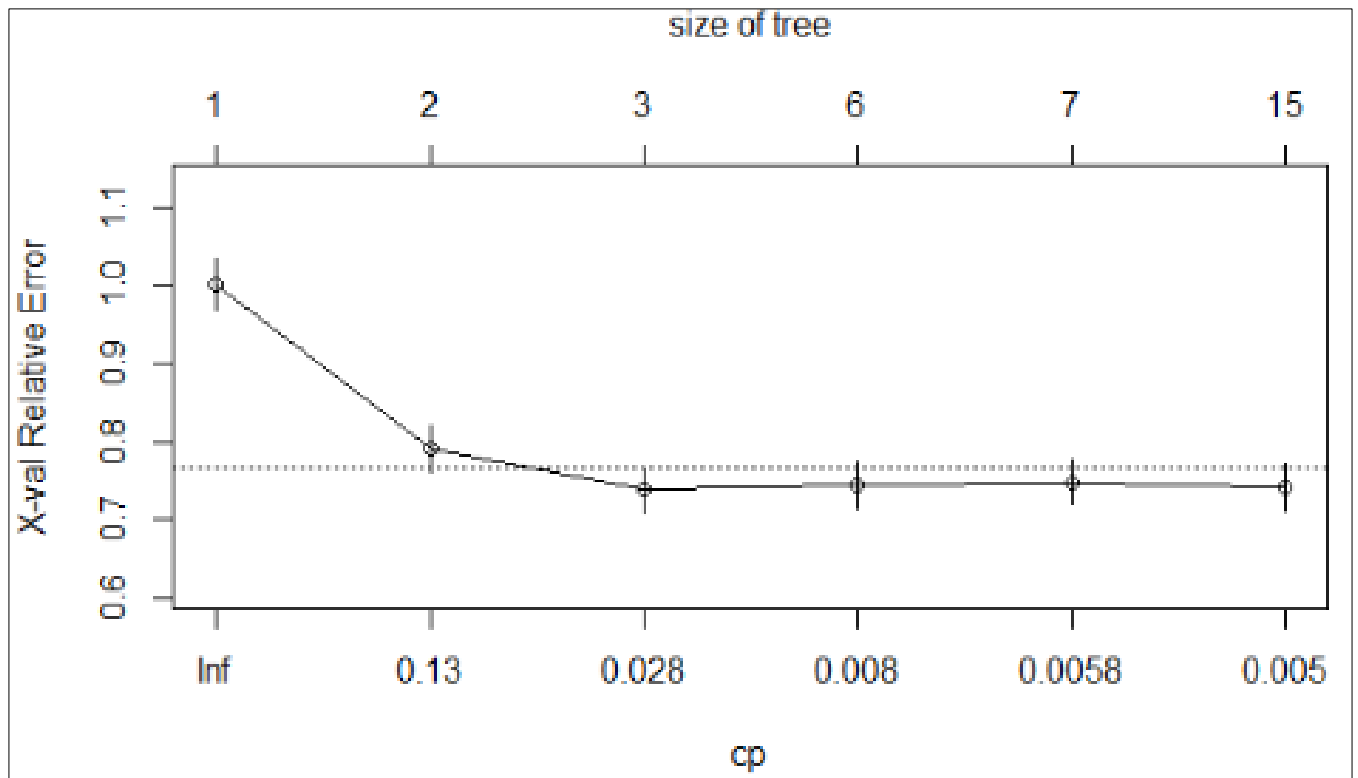


When checking the cp values we check the cross validation error (xerror), we could see that the value decreases until a certain point and increases again. The error value of 0.73725 is the least error variable beyond which the error values increases again and the same is visualized in the above plot. Hence we are taking the cp value as 0.00463679 for pruning our CART tree. After using the `prune()` function below is the final output of the pruned tree.

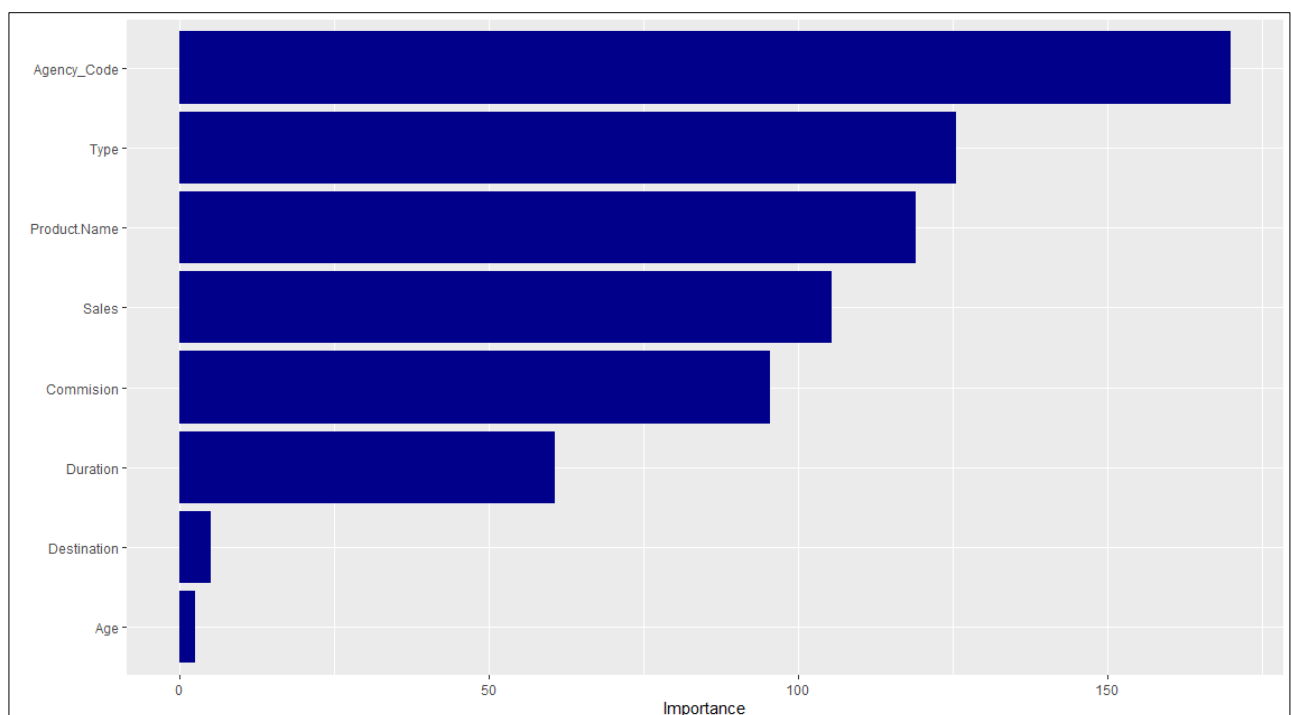


Below displays the first split which happens with the Agency code:





Next we are finding the variable importance for this pruned model using the VIP package and below is the plot generated. As per cart model we can see that the Agency Code, Product name, Sales and commission has the highest importance when it comes to independent variable that provides best predictors for our dependent variable which is claims.



Next with the pruned CART tree model we will predict the Claimed value for the testing set and then evaluate the model's accuracy by building the confusion matrix and the ROC curve.

The next question covers the model accuracy.

RANDOM FOREST:

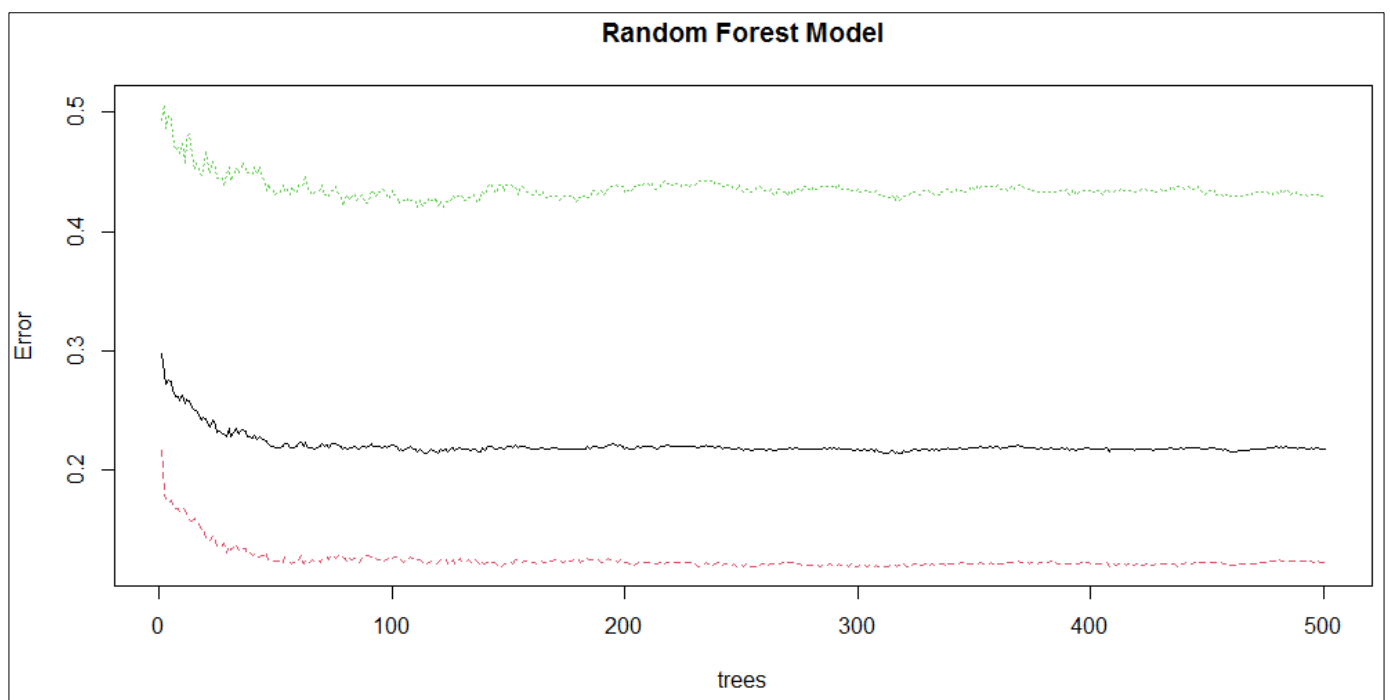
Random Forest is a learning method that operates by constructing multiple decision trees. The final decision is made based on the majority of the trees and is chosen by the random forest. It is an ensemble learning method for classification, regression that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees

Based on the same trained and test data set we now perform the random forest model. Below are the proportion scores generated for the training and test dataset. We could see that the train and test data has equal proportions of Yes and No claims which shows both are equally split.

No	Yes
0.691758	0.308242

No	Yes
0.6918799	0.3081201

Using the RandomForest package we now build the model. We are providing the number of trees parameter as 501 trees, mtry or the number of variables to perform bootstrapping as 4 and the nodesize (the number of variables to be left in the terminal nodes) as 10. With all these parameters we run the model first and below is the output plot generated where the green line shows the error rate of predicted “Yes” , black line shows the Out of Bag error rate and the Red shows the “No” predicted error rate. We could see that the “Yes” error rate smoothens at a value of 400 trees along with OOB error rate which shows that the tuning parameter for Random forest tuning can be chosen to be 401 trees.



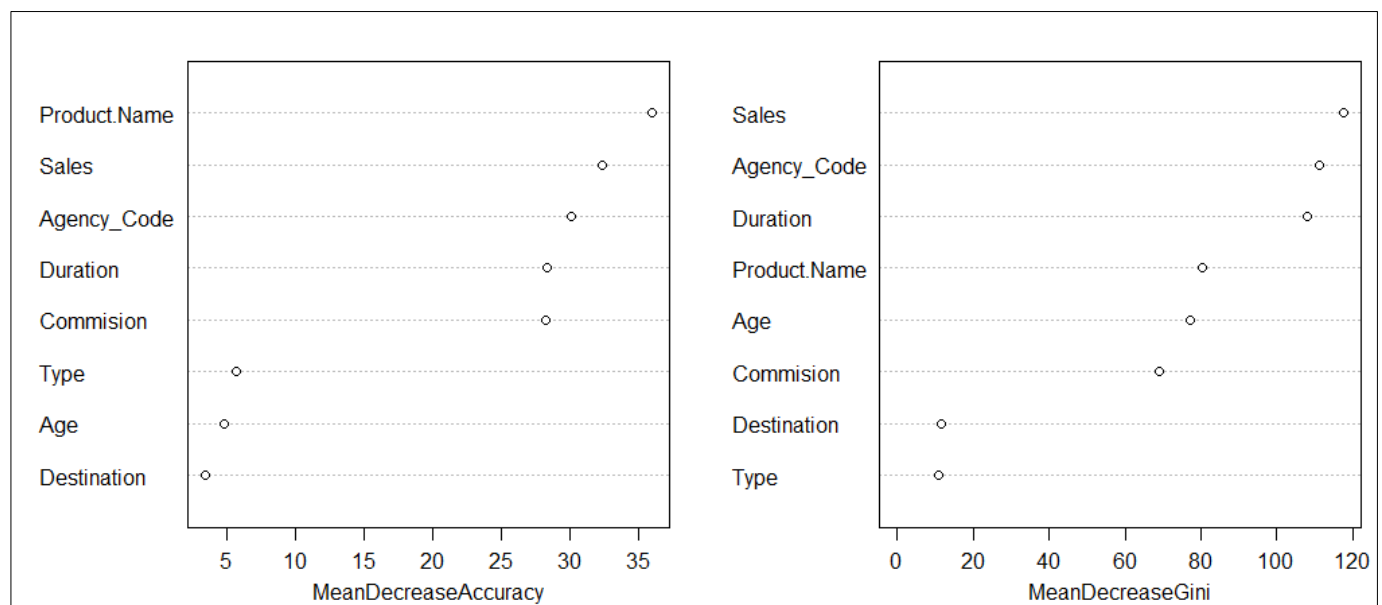
Below is the summary of the random forest model

	Length	Class	Mode
call	7	-none-	call

type	1	-none-	character
predicted	2099	factor	numeric
err.rate	1503	-none-	numeric
confusion	6	-none-	numeric
votes	4198	matrix	numeric
oob.times	2099	-none-	numeric
classes	2	-none-	character
importance	32	-none-	numeric
importanceSD	24	-none-	numeric
localImportance	0	-none-	NULL
proximity	0	-none-	NULL
ntree	1	-none-	numeric
mtry	1	-none-	numeric
forest	14	-none-	list
y	2099	factor	numeric
test	0	-none-	NULL
inbag	0	-none-	NULL
terms	3	terms	call

In addition we are checking the importance values of the model and plotting the same

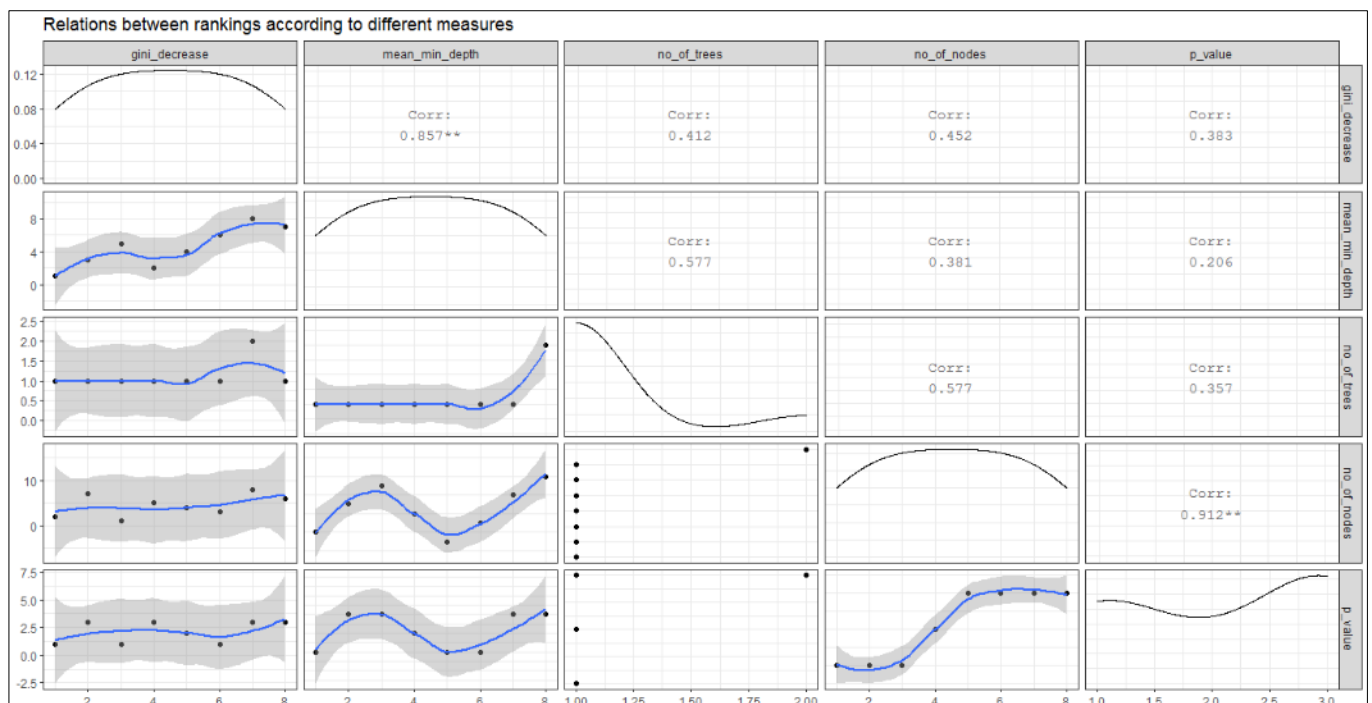
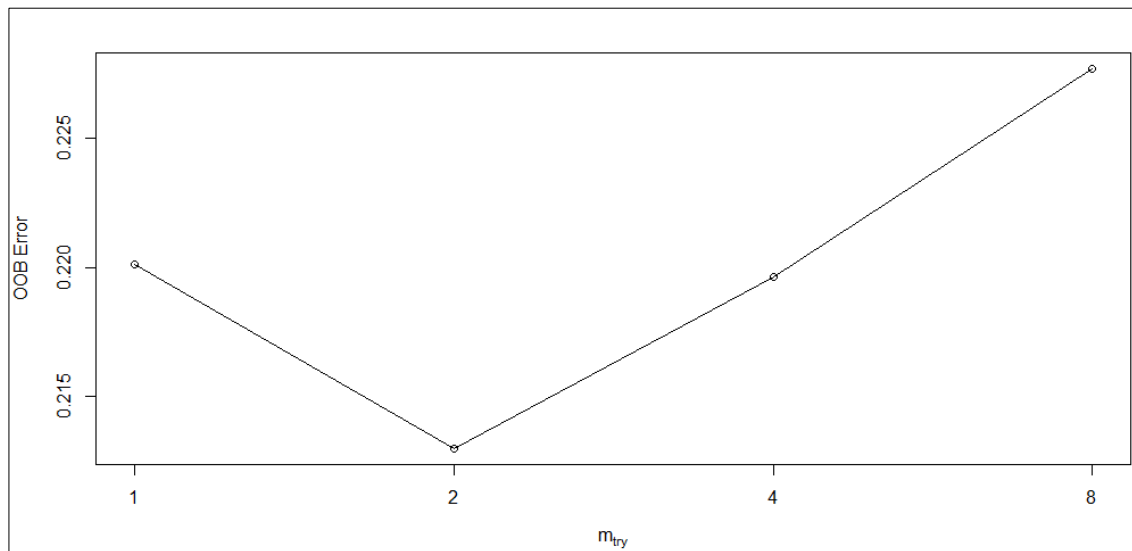
	No	Yes	MeanDecreaseAccuracy	MeanDecreaseGini
Age	9.12430252	-4.251875	4.840464	77.11574
Agency_Code	10.36174949	27.574275	30.107681	111.28534
Type	0.02157416	6.177304	5.684899	10.70500
Commision	-3.53340355	30.124477	28.312777	69.13925
Duration	-1.50816162	30.813025	28.360824	107.96570
Sales	-6.95632549	37.028926	32.427874	117.58865
Product.Name	24.00446929	16.427724	35.959257	80.36501
Destination	-4.87426502	12.744677	3.471388	11.66192

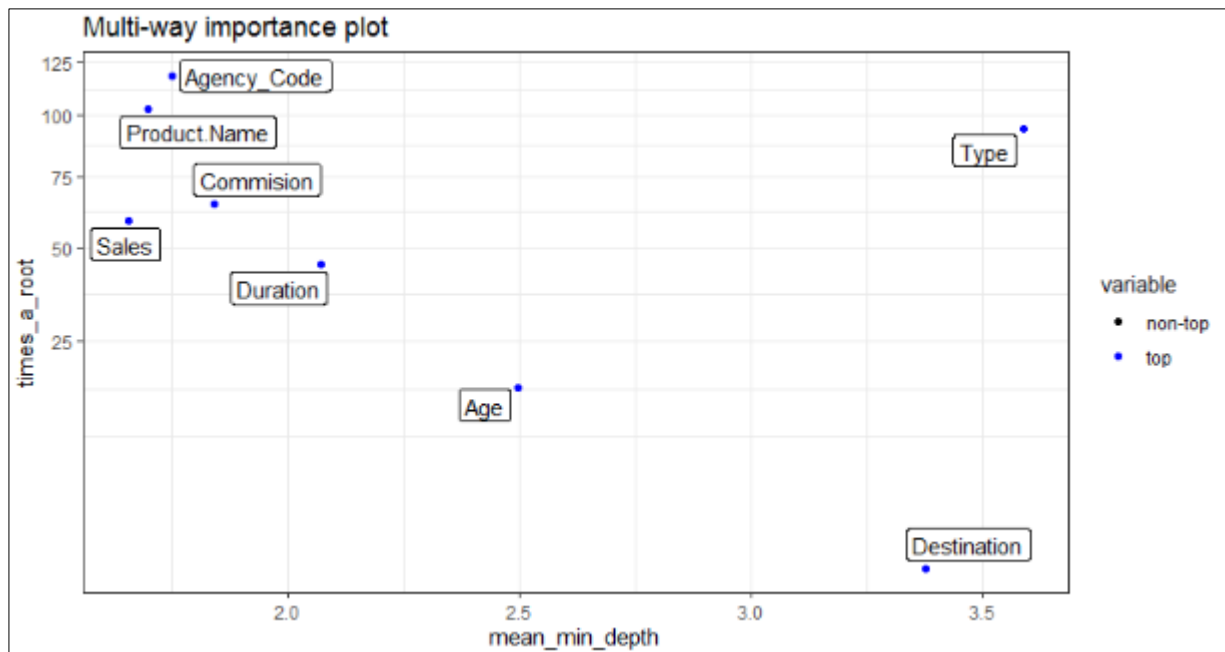


We could see that the Product name has the highest mean decrease in accuracy with a value of 35% followed by Sales , Agency code and duration. This shows that these three variables are important for prediction and without which there will be huge reduction in the accuracy. The product name, sales, agency code, duration contribute to around 90% accuracy which shows high weightage for these products.

Tuning the random forest:

Next with the required parameters we are tuning our build random forest model to accurately find the best random forest model. Using Tune RF() function we are now providing the ntreeTry tree size as 401, mtry start value as 4 and step factor as 0.5. Below is the consistent output from the tune function. We could see that the mtry value from 2 increase and flattens at a value of 8, this shows that the mtry value between 2 and 8 would be appropriate.





The final tuned RF model is now used for predicting claims in the test data set and the model is evaluated. Once predicted we build the confusion matrix, ROC curve , auc and other parameters to evaluate the model accuracy which is discussed in the next section.

Artificial Neural Network (ANN)

Neural Networks are a machine learning framework that attempts to mimic the learning pattern of natural biological neural networks. Biological neural networks have interconnected neurons with dendrites that receive inputs, then based on these inputs they produce an output signal through an axon to another neuron. We will try to mimic this process through the use of Artificial Neural Networks (ANN). The process of creating a neural network begins with the most basic form, a single perceptron.

Data pre - processing : Unlike CART and Random forest models, the ANN requires all the categorical values to be converted to numeric and all the target or dependent variable to be converted to numerical type. Also the neural network requires data to be normalized and scaled as ANN algorithm is based on weight calculations and any large values or data with different scales/ units will have large impact on the weights and the nodes. For all the categorical variable conversion we will use dummies library to convert each categorical values as separate columns with 1's and 0's . We then convert our target variable from string to numbers. We then scale the data using normalization and then split the data to training and testing set. Below is the top five rows of the final preprocessed data.

Age	Agency.CodeC2B	Agency.CodeCWT	Agency.CodeEPX	Agency.CodeJZI	TypeAirlines	TypeTravel Agency	Commision	Duration
0.94697619	1.4991178	-0.4321971	-0.9141149	-0.2936037	1.2576680	-1.2576680	-0.54287057	-0.5815341
-0.19995403	-0.6668365	-0.4321971	1.0935895	-0.2936037	-0.7948572	0.7948572	-0.57033431	-0.3263123
0.08677852	-0.6668365	2.3129872	-0.9141149	-0.2936037	-0.7948572	0.7948572	-0.33728481	-0.6193448
-0.19995403	-0.6668365	-0.4321971	1.0935895	-0.2936037	-0.7948572	0.7948572	-0.57033431	-0.6098921
-0.48668659	-0.6668365	-0.4321971	-0.9141149	3.4048158	1.2576680	-1.2576680	-0.32316059	-0.1467117
0.66024363	-0.6668365	-0.4321971	-0.9141149	3.4048158	1.2576680	-1.2576680	0.04759999	-0.5720815

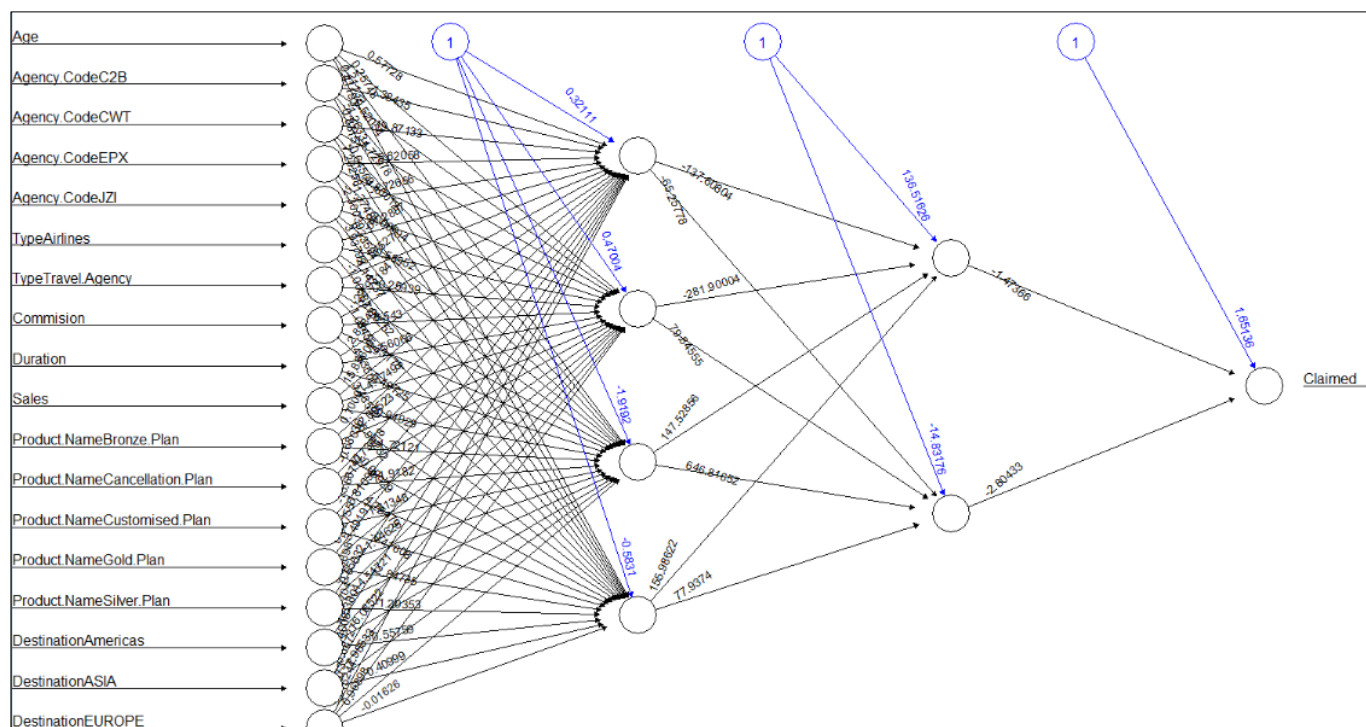
Sales	Product.NameBronze Plan	Product.NameCancellation Plan	Product.NameCustomised Plan	Product.NameGold Plan
-0.8166529	-0.5255432	-0.5405034	1.2809614	-0.194208
-0.5694264	-0.5255432	-0.5405034	1.2809614	-0.194208
-0.7121930	-0.5255432	-0.5405034	1.2809614	-0.194208
-0.4846146	-0.5255432	1.8495101	-0.7804032	-0.194208
-0.5976970	1.9021586	-0.5405034	-0.7804032	-0.194208
-0.2160439	1.9021586	-0.5405034	-0.7804032	-0.194208

Product.NameSilver Plan	DestinationAmericas	DestinationASIA	DestinationEUROPE	Claimed
-0.4074651	-0.3456187	0.4659853	-0.277901	0
-0.4074651	-0.3456187	0.4659853	-0.277901	0
-0.4074651	2.8923967	-2.1452745	-0.277901	0
-0.4074651	-0.3456187	0.4659853	-0.277901	0
-0.4074651	-0.3456187	0.4659853	-0.277901	0
-0.4074651	-0.3456187	0.4659853	-0.277901	1

Below is the structure of the converted data

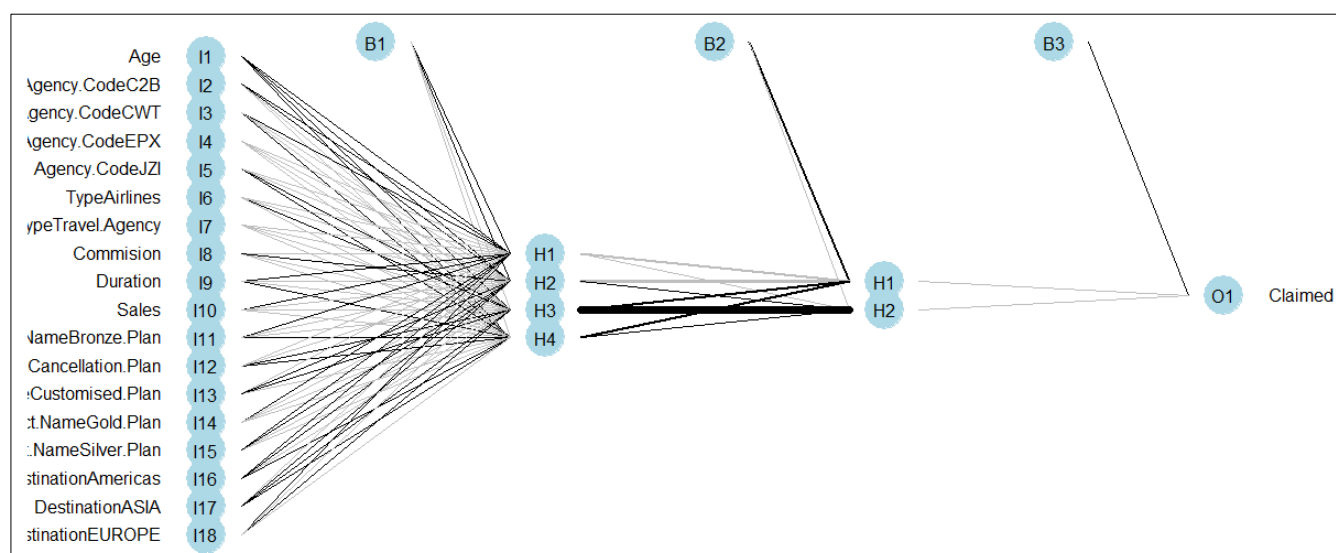
```
'data.frame': 2998 obs. of 19 variables:
 $ Age : num 0.947 -0.2 0.0868 -0.2 -0.4867 ...
 $ Agency.CodeC2B : num 1.499 -0.667 -0.667 -0.667 -0.667 ...
 $ Agency.CodeCWT : num -0.432 -0.432 2.313 -0.432 -0.432 ...
 $ Agency.CodeEPX : num -0.914 1.094 -0.914 1.094 -0.914 ...
 $ Agency.CodeJZI : num -0.294 -0.294 -0.294 -0.294 3.405 ...
 $ TypeAirlines : num 1.258 -0.795 -0.795 -0.795 1.258 ...
 $ TypeTravel Agency : num -1.258 0.795 0.795 0.795 -1.258 ...
 $ Commision : num -0.543 -0.57 -0.337 -0.57 -0.323 ...
 $ Duration : num -0.582 -0.326 -0.619 -0.61 -0.147 ...
 $ Sales : num -0.817 -0.569 -0.712 -0.485 -0.598 ...
 $ Product.NameBronze Plan : num -0.526 -0.526 -0.526 -0.526 1.902 ...
 $ Product.NameCancellation Plan: num -0.541 -0.541 -0.541 1.85 -0.541 ...
 $ Product.NameCustomised Plan : num 1.28 1.28 1.28 -0.78 -0.78 ...
 $ Product.NameGold Plan : num -0.194 -0.194 -0.194 -0.194 -0.194 ...
 $ Product.NameSilver Plan : num -0.407 -0.407 -0.407 -0.407 -0.407 ...
 $ DestinationAmericas : num -0.346 -0.346 2.892 -0.346 -0.346 ...
 $ DestinationASIA : num 0.466 0.466 -2.145 0.466 0.466 ...
 $ DestinationEUROPE : num -0.278 -0.278 -0.278 -0.278 -0.278 ...
 $ Claimed : num 0 0 0 0 1 0 0 0 0 ...
```

We now build the neural network model with the hidden layers as 4 and 2. We have chosen first hidden layer as the square root of the number of variables which is 19. Note that the variable numbers has been increased due to the categorical to numeric conversion using dummies. The square root of 19 is around 4.34 which we are rounding off to 4 and the second hidden layer is taken to be square root of first layer which is two. We are setting the threshold as 0.01 and running the model. We would make use of neuralnet libraries to run the model. Below are the weights and output generated for the neural network after the model converged at the 112028 iterations.

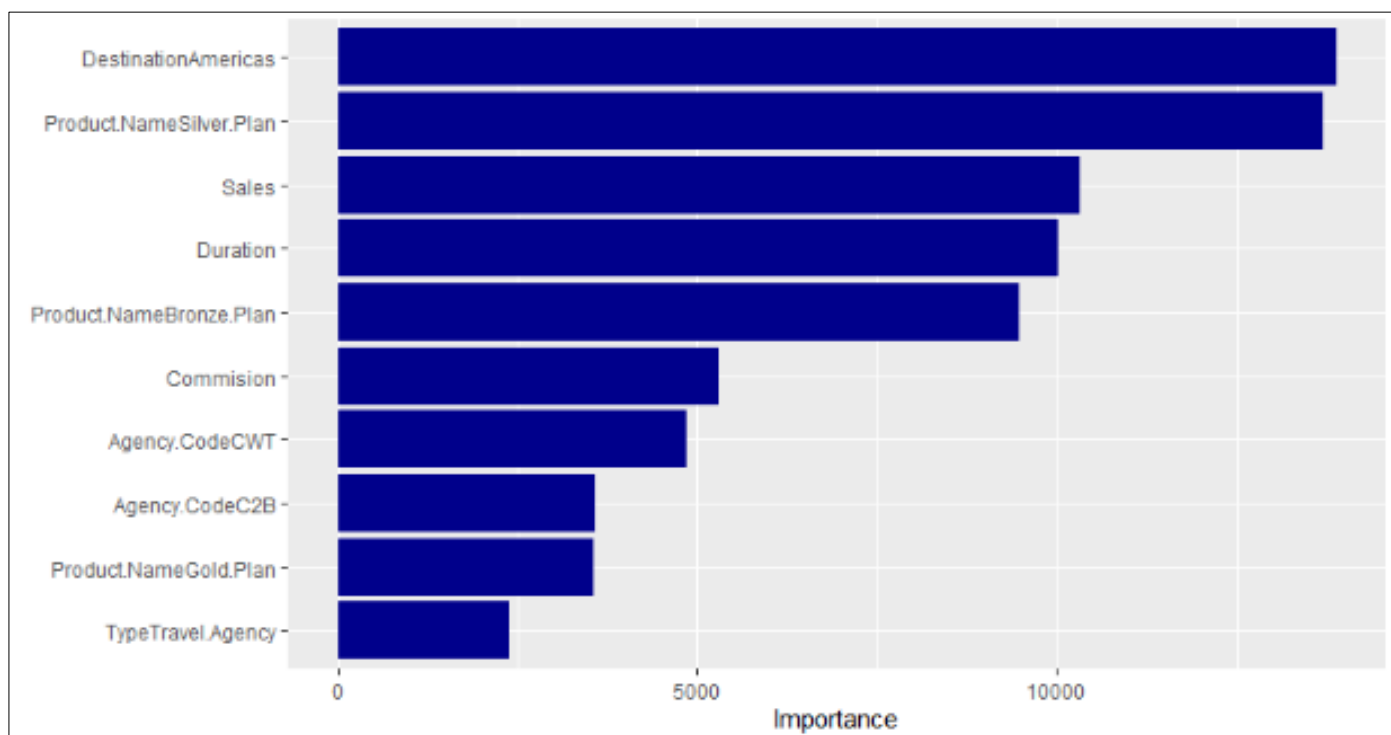
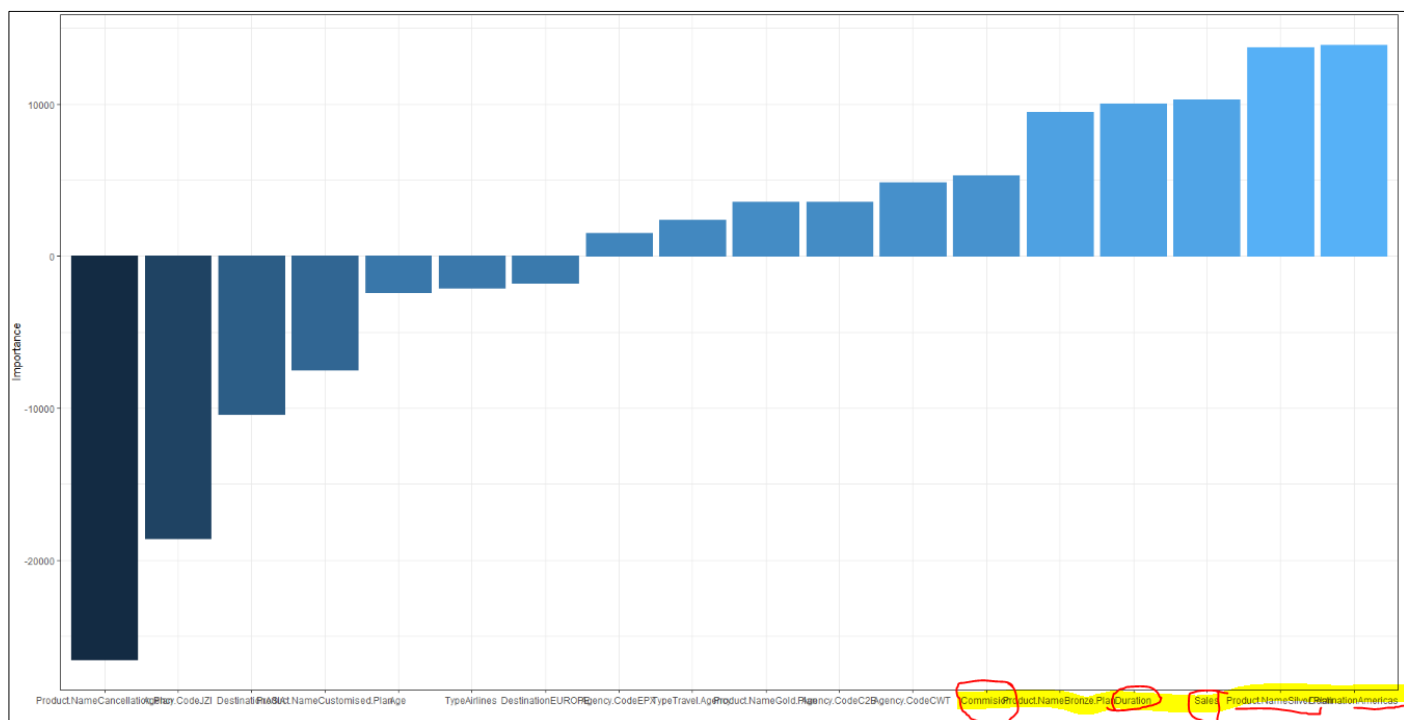


Above plot is the neural network plot with generalized weights of all the layers and as well as the bias nodes.

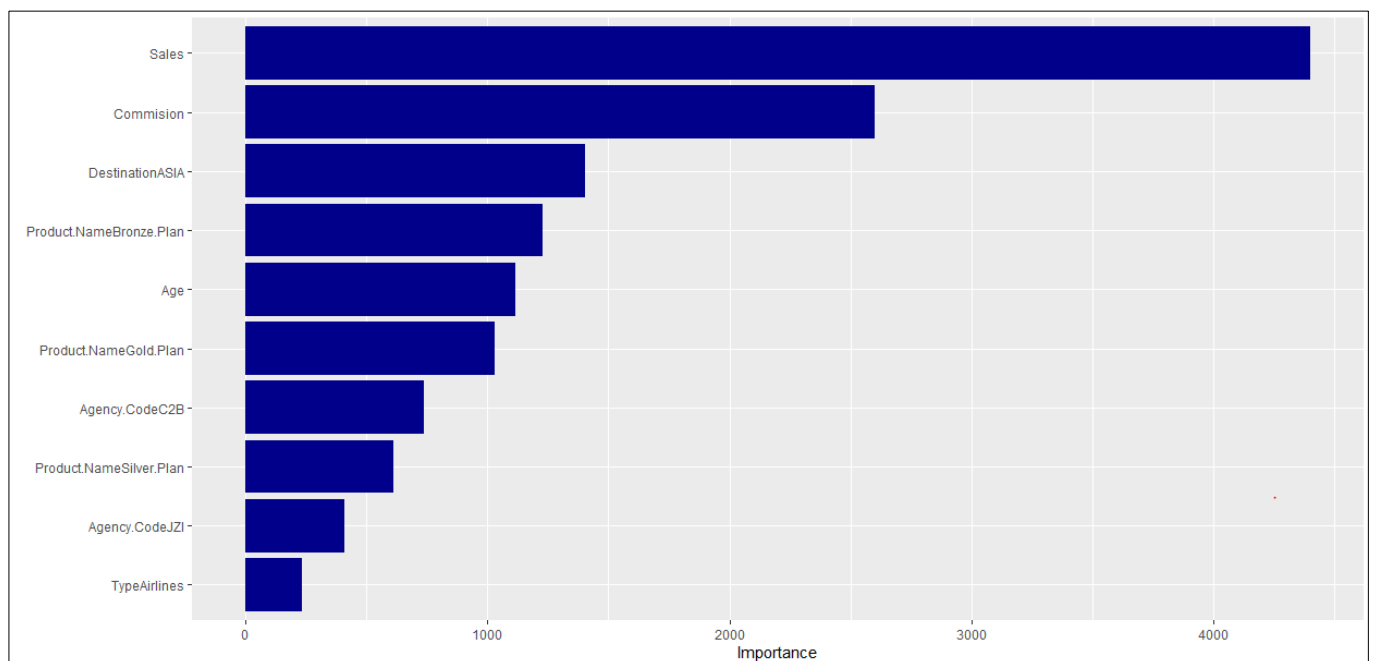
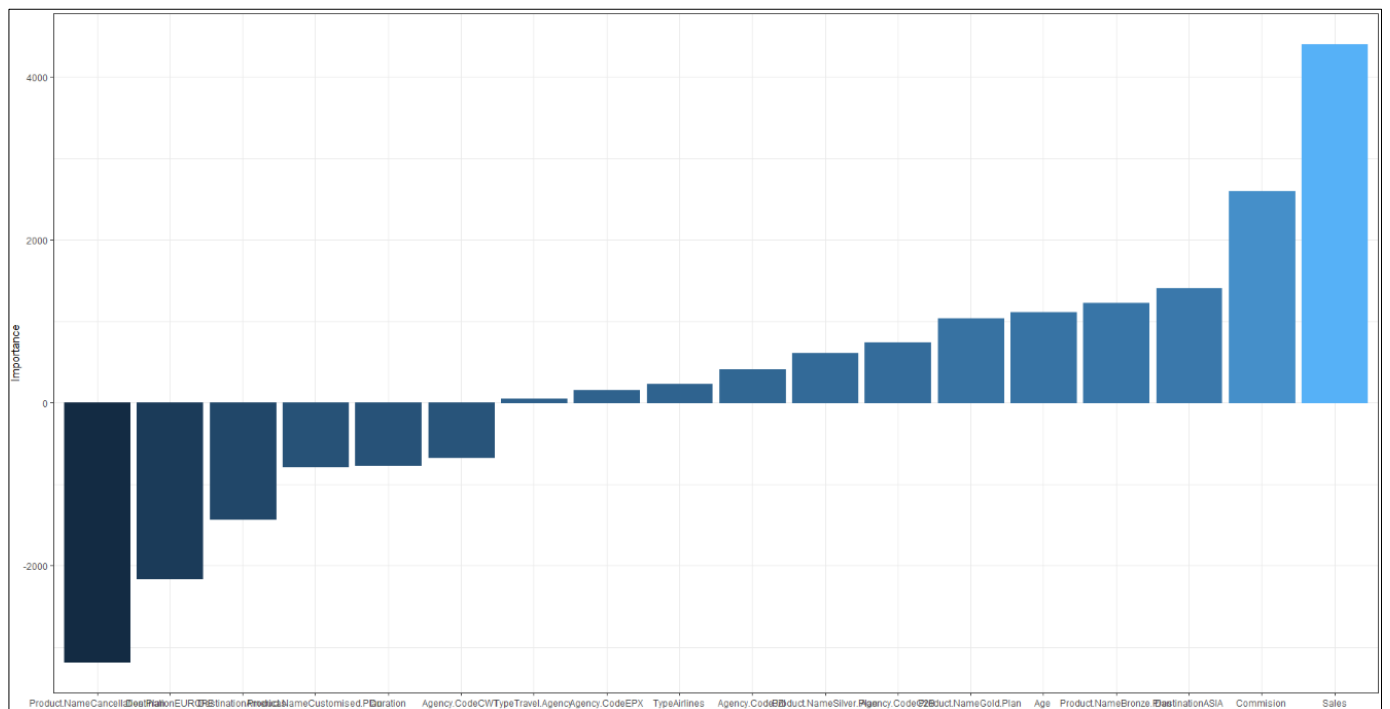
error	1.418277e+02
reached.threshold	9.937679e-03
steps	1.120280e+05



Next we are plotting the variable importance of the model and below is the plot obtained using Neuralnettools & vip package.



For Test Set :



We can see that the product names appear with high importance value in both the train and test set followed by sales variable. Next followed by sales is agency code , commission and the duration of travel. Silver , gold and bronze plan has highest claims compared to other two as seen in below table. Agency code 'C2B' & 'CWT' has high relative importance in insurance claims.

Training Set		Testing Set	
DestinationAmericas	13881.812	Sales	4400.99654
Product.NameSilver.Plan	13709.484	Commision	2601.52031
Sales	10316.035	DestinationASIA	1403.73961
Duration	10022.273	Product.NameBronze.Plan	1227.28264
Product.NameBronze.Plan	9472.399	Age	1114.45587
Commision	5299.613	Product.NameGold.Plan	1031.37281
Agency.CodeCWT	4841.992	Agency.CodeC2B	739.49262
Agency.CodeC2B	3572.779	Product.NameSilver.Plan	610.53914
Product.NameGold.Plan	3556.481	Agency.CodeJZI	407.83733
TypeTravel.Agency	2373.009	TypeAirlines	231.90334
Agency.CodeEPX	1502.746	Agency.CodeEPX	156.41239
DestinationEUROPE	-1756.964	TypeTravel.Agency	47.77833
TypeAirlines	-2080.79	Agency.CodeCWT	-669.77953
Age	-2373.242	Duration	-770.51085
Product.NameCustomised.Plan	-7465.524	Product.NameCustomised.Plan	-783.37919
DestinationASIA	-10414.061	DestinationAmericas	-
Agency.CodeJZI	-18578.349	DestinationEUROPE	-
Product.NameCancellation.Plan	-26520.421	Product.NameCancellation.Plan	-
			3183.28976

Below is the final output of the generalized weights generated by our model which reduces the error as per stochastic gradient descent.

```
$wts$`hidden 1 1`
[1] 0.32111133 0.57727927 1.38435216 19.87132732 -5.62057909 -
8.72656296 -0.22887434 -0.57527069 -22.97183869 2.85347651
[11] 15.85607672 0.70096821 -0.68035778 -13.66741821 0.02755018
9.86898219 3.82704286 -11.14008253 2.09373301

$wts$`hidden 1 2`
[1] 0.4700432 0.2574834 0.5205381 -4.7297750 -0.8001184 -3.0059430 -
0.5895207 -0.2593855 6.5430022 -3.5606594 -4.1749122
[12] -2.0052311 -0.7341812 6.8109553 -3.4919120 0.6563207 1.5380396
0.6872743 -3.1524412

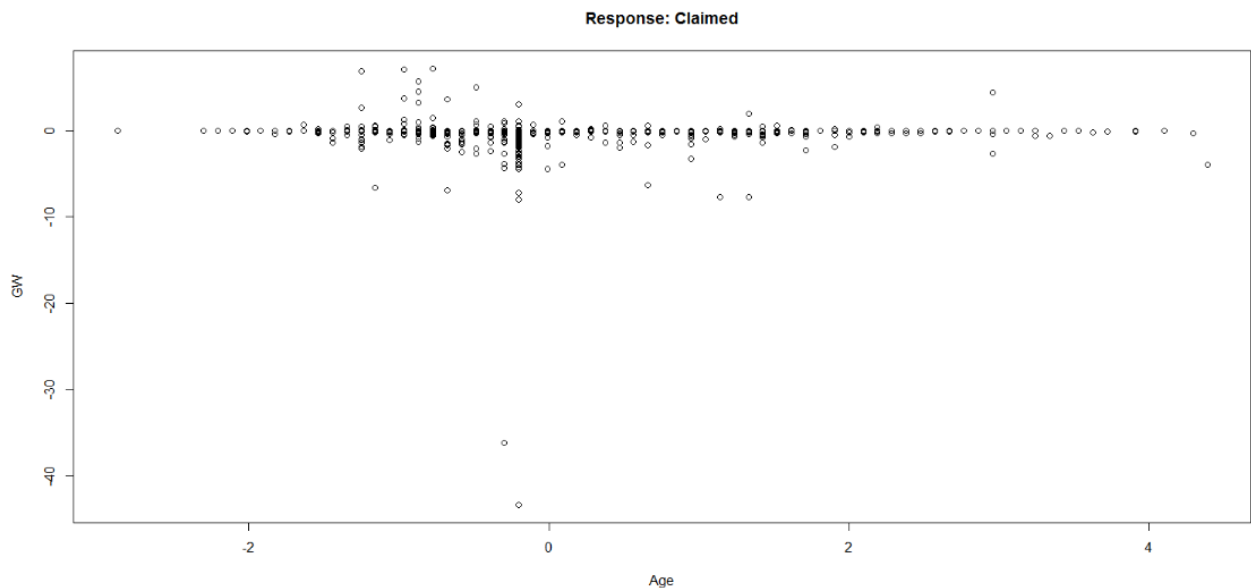
$wts$`hidden 1 3`
[1] -1.9192046 1.2113176 -1.2633052 -0.6456373 -1.3748768 6.3359411
1.1473749 -1.0625215 -5.3317763 -5.4972527 -0.9192867
[12] -4.7212093 11.9181982 1.8134782 -1.4462457 -4.5472113 -6.0852175
2.9853309 0.9689751

$wts$`hidden 1 4`
[1] -0.583101523 0.411926756 -0.831509286 7.202582629 -2.300392313
3.937918691 -1.005865135 -1.084501475 -4.635022309 3.485924452
[11] -6.982316517 0.009289081 4.254788523 -0.416078805 -2.847346543 -
1.203532963 0.557587808 0.409985341 -0.016258794

$wts$`hidden 2 1`
[1] 136.5163 -137.6080 -281.9000 147.5286 155.9862
```

```
$wts$`hidden 2 2`
[1] -14.83176 -65.25778 79.84555 646.81652 77.93740

$wts$`out 1`
[1] 1.651361 -1.473657 -2.804330
```



This model is now used for predicting test data set and the model is evaluated for accuracy.

4.4 Qn 3 : Performance Metrics: Check the performance of Predictions on Train and Test sets using Confusion Matrix

The following are the performance metrics that are derived for the three Machine Learning models and compared.

1) Confusion Matrix :

A confusion matrix is a technique for summarizing the performance of a classification algorithm. It consist of the below set of values.

- True Positives: The predicted positive (yes) and the actual positive is same
- True Negatives: The predicted negative (No) and the actual negative is same
- False Positives (Type 1 Error): The value is predicted as positive (yes) but the actual value is Negative in the data.
- False Negatives (Type 2 Error): The value is predicted as negative (no) but the actual value is positive in the data.

Confusion Matrix	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

2) Accuracy :

It is a measure of how accurately our model classify the data points in terms of percentage. In general less the value of false prediction rate, the more is the accuracy.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

3) Sensitivity or Recall :

Sensitivity/Recall is the ratio between how much were correctly identified as positive compared to how much were actually positive. Lesser the false negative rates, higher is the sensitivity of the model.

$$\text{Sensitivity} = TP / (FN + TP)$$

4) Specificity :

Specificity of a classifier is the ratio between how much were correctly classified as negative to how much was actually negative.

$$\text{Specificity} = TN / (FP + TN)$$

5) Precision :

It is a measure of how much were correctly classified as positive out of all positives. Among the points identified as Positive by the model, how many are really Positive ones.

$$\text{Precision} = TP / (TP + FP)$$

6) F1 Score

F1 Score might be a better measure to use if we need to seek a balance between Precision and Recall when there is an uneven class distribution (large number of Actual Negatives). The harmonic mean of precision and recall gives f1 score.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

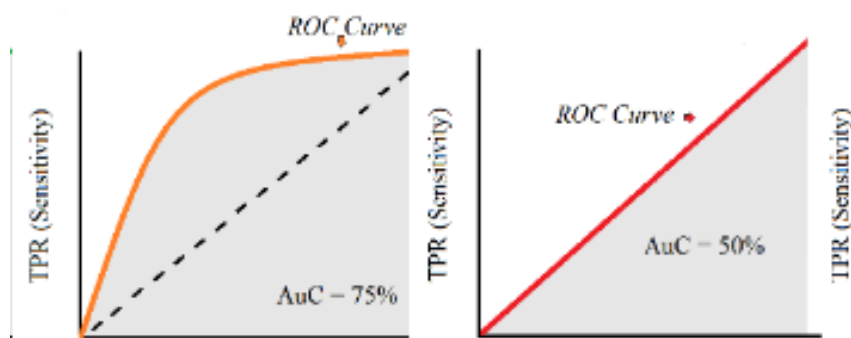
7) Kappa Coefficient :

It basically tells us how much better the classification model is performing over the performance of a ML model that simply guesses at random according to the frequency of each class. It is a coefficient value that vary between 0 to 1 with ranges indicating 0–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1 as almost perfect agreement.

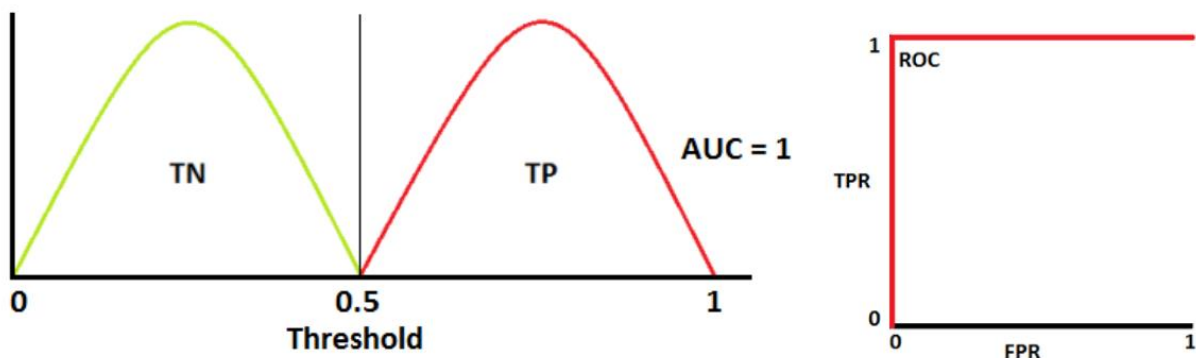
8) Receiver Operating Characteristics (ROC) Curve

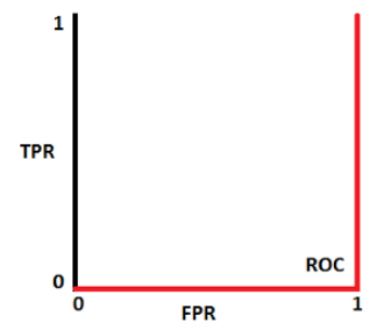
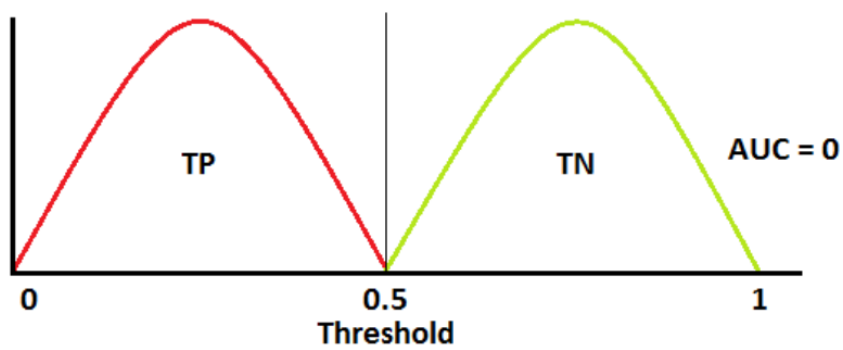
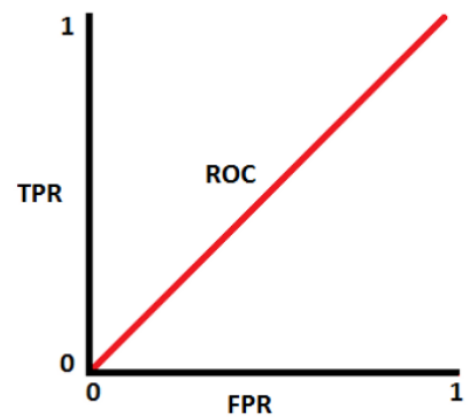
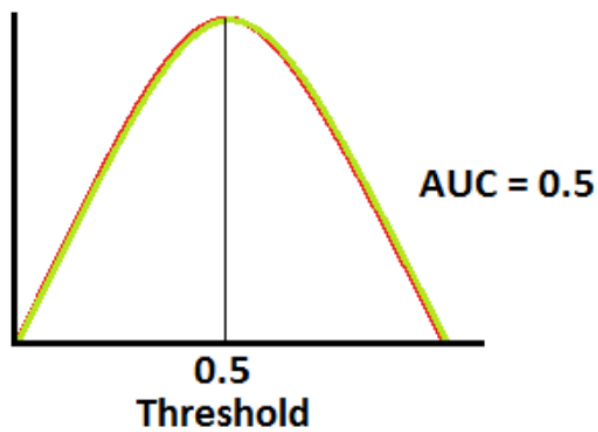
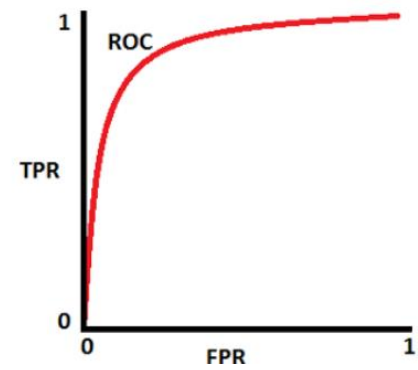
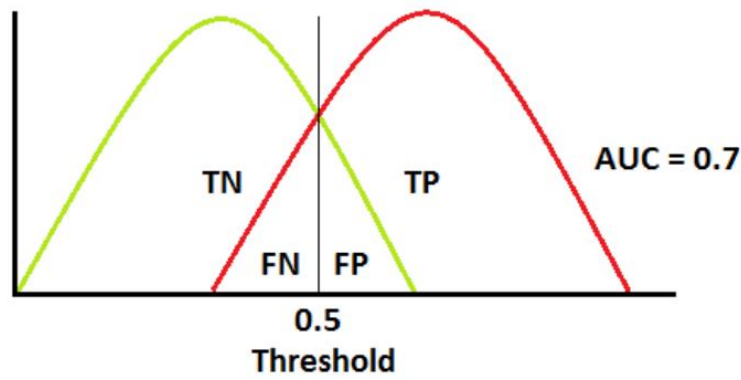
AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s

The ROC curve is plotted with True Positive Rates (TPR) against the False Positive Rates (FPR) where TPR is on y-axis and FPR is on the x-axis



An excellent model has AUC near to the 1 which means it has good measure of separability. A poor model has AUC near to the 0 which means it has worst measure of separability, when AUC is 0.5, it means model has no class separation capacity. Below depicts the various scenarios.





With above explanation on various model evaluation measures below are the calculated values for each model.

CART Model

Below is the confusion matrix for the trained data set, Accuracy of the model is 80.56%

Confusion Matrix and Statistics

```

      Reference
Prediction No  Yes
      No  1309  265
      Yes   143  382

      Accuracy : 0.8056
      95% CI : (0.788, 0.8224)
      No Information Rate : 0.6918
      P-Value [Acc > NIR] : < 2.2e-16

      Kappa : 0.5191

Mcnemar's Test P-Value : 2.093e-09

      Sensitivity : 0.5904
      Specificity : 0.9015
      Pos Pred Value : 0.7276
      Neg Pred Value : 0.8316
      Precision : 0.7276
      Recall : 0.5904
      F1 : 0.6519
      Prevalence : 0.3082
      Detection Rate : 0.1820
      Detection Prevalence : 0.2501
      Balanced Accuracy : 0.7460

      'Positive' Class : Yes
```

Below is the confusion matrix for the test data set , accuracy of the model is 77.31%, accuracy for the test dataset will be always low compared to train set.

Confusion Matrix and Statistics

```

      Reference
Prediction No  Yes
      No   556  138
      Yes   66  139

      Accuracy : 0.7731
      95% CI : (0.7443, 0.8001)
      No Information Rate : 0.6919
      P-Value [Acc > NIR] : 3.671e-08

      Kappa : 0.4264

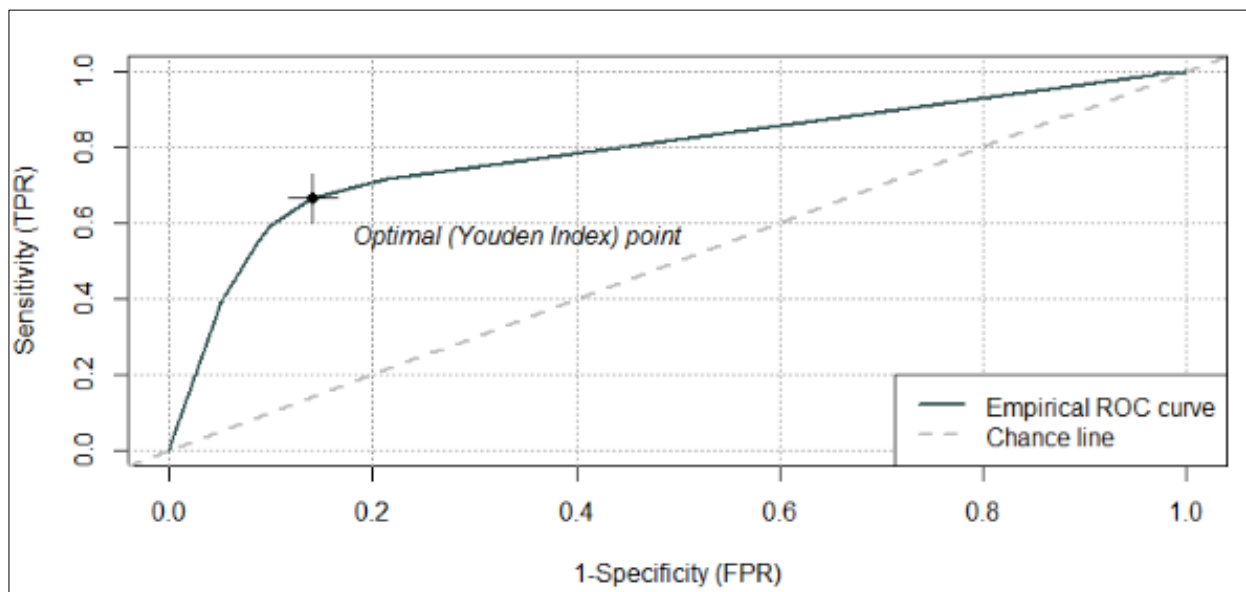
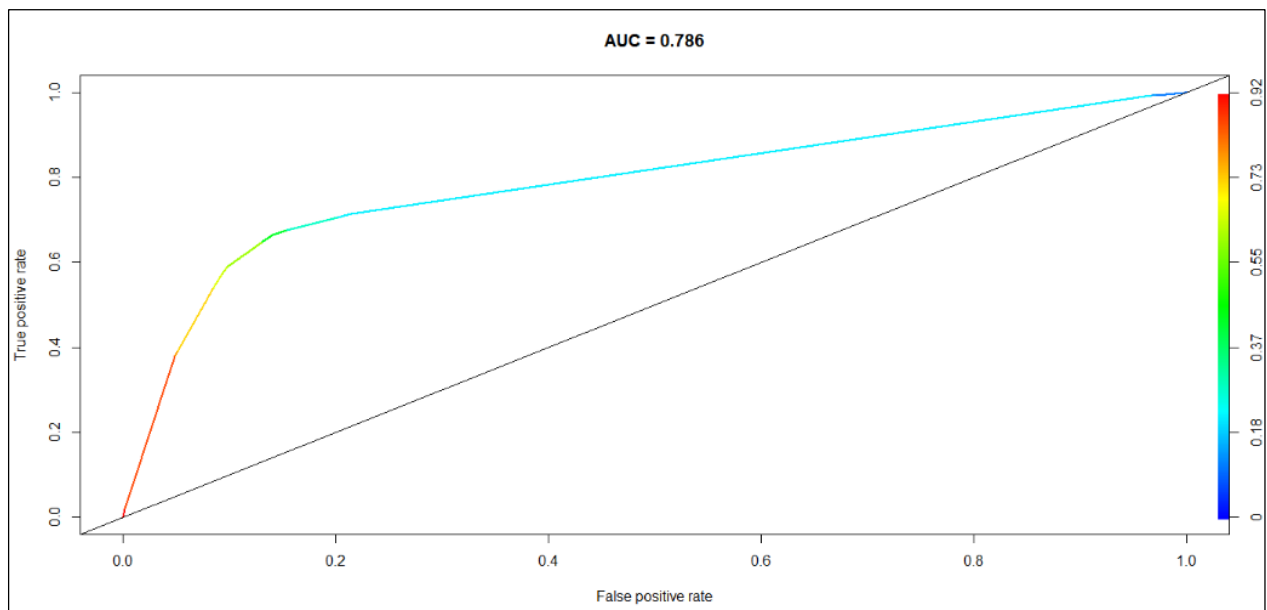
Mcnemar's Test P-Value : 6.661e-07

      Sensitivity : 0.5018
      Specificity : 0.8939
      Pos Pred Value : 0.6780
      Neg Pred Value : 0.8012
      Precision : 0.6780
      Recall : 0.5018
```

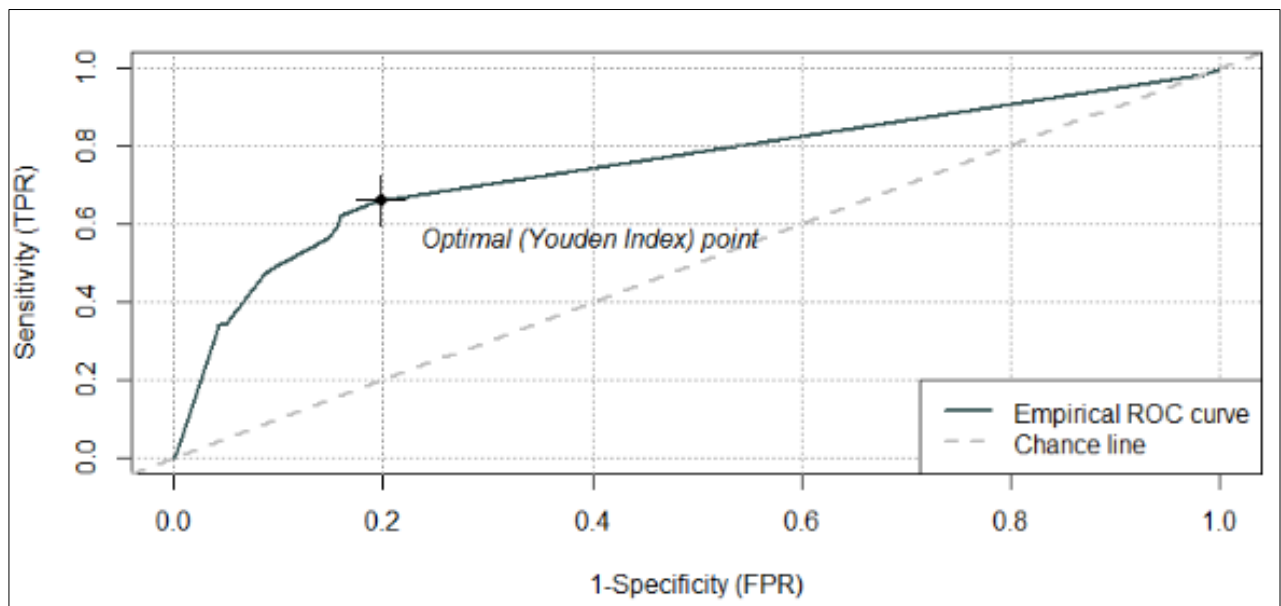
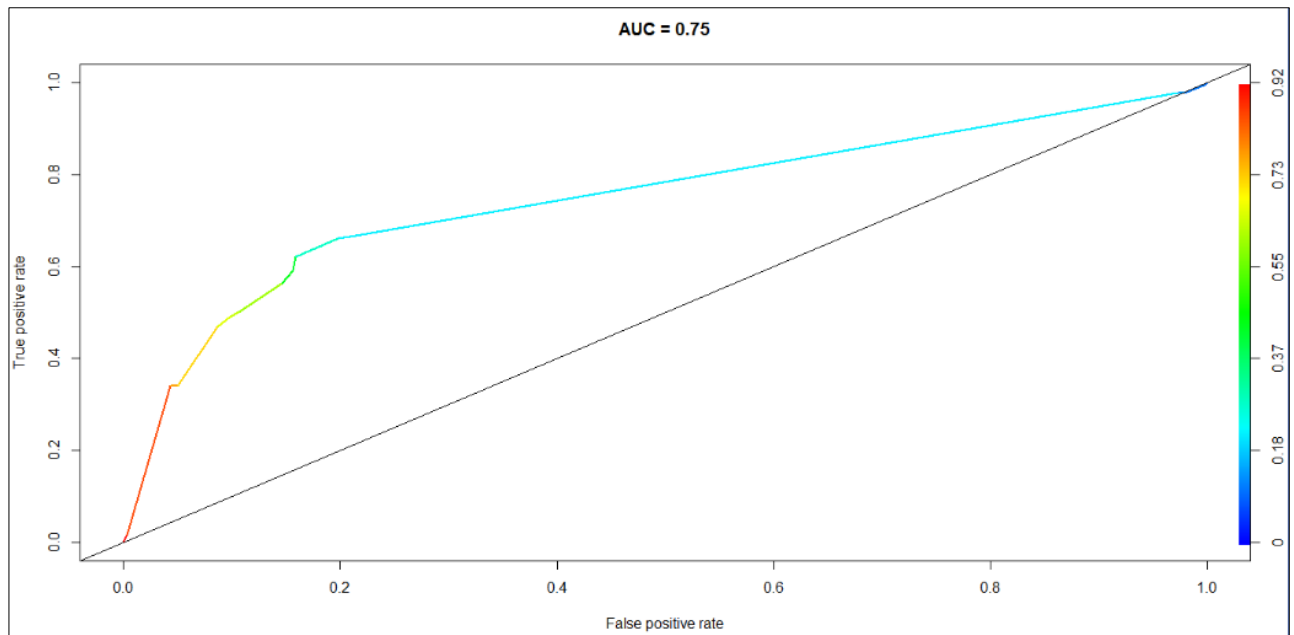
```
F1 : 0.5768
Prevalence : 0.3081
Detection Rate : 0.1546
Detection Prevalence : 0.2280
Balanced Accuracy : 0.6978
```

'Positive' Class : Yes

The ROC curve and AUC value for the train data set is shown below



The ROC curve and AUC value for the test data set is shown below



Random Forest:

Below is the confusion matrix for the trained data set, Accuracy of the model is 83.52%

Confusion Matrix and Statistics

	No	Yes
No	1316	210
Yes	136	437

Accuracy : 0.8352
95% CI : (0.8186, 0.8508)
No Information Rate : 0.6918
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.6008

Mcnemar's Test P-Value : 8.691e-05

Sensitivity : 0.6754
Specificity : 0.9063
Pos Pred Value : 0.7627
Neg Pred Value : 0.8624
Precision : 0.7627
Recall : 0.6754
F1 : 0.7164
Prevalence : 0.3082
Detection Rate : 0.2082
Detection Prevalence : 0.2730
Balanced Accuracy : 0.7909

'Positive' Class : Yes

Below is the confusion matrix for the test data set, Accuracy of the model is 78.09%

Confusion Matrix and Statistics

	No	Yes
No	541	116
Yes	81	161

Accuracy : 0.7809
95% CI : (0.7524, 0.8075)
No Information Rate : 0.6919
P-Value [Acc > NIR] : 1.59e-09

Kappa : 0.4674

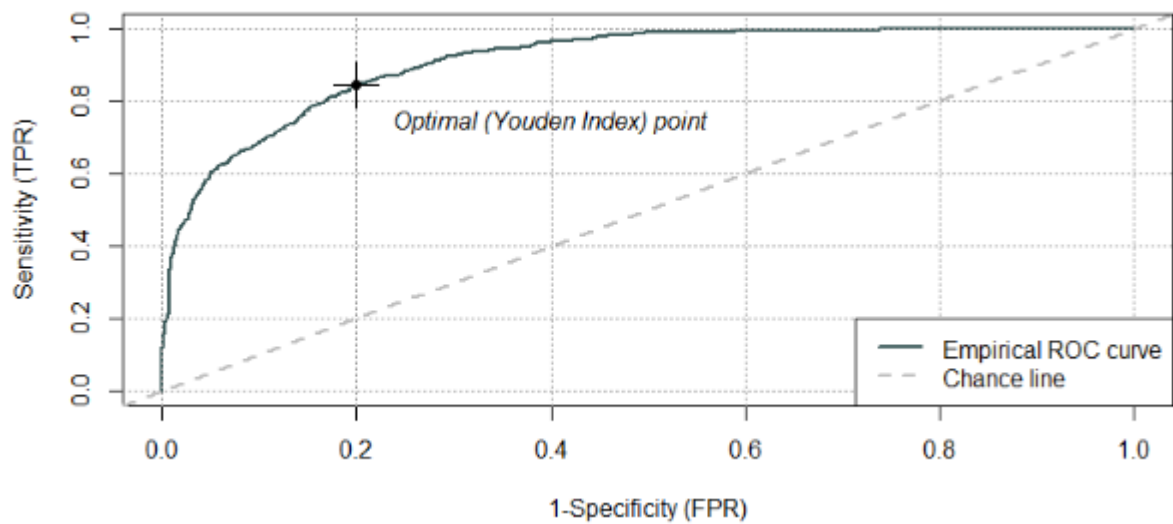
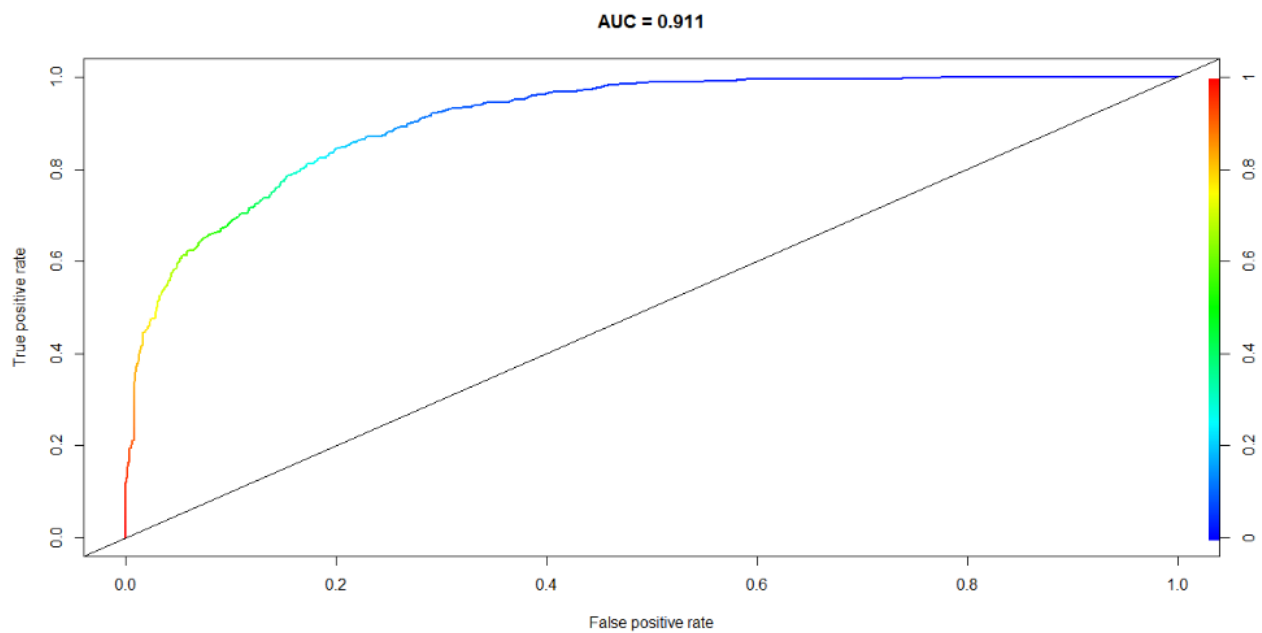
Mcnemar's Test P-Value : 0.01542

Sensitivity : 0.5812
Specificity : 0.8698
Pos Pred Value : 0.6653
Neg Pred Value : 0.8234
Precision : 0.6653
Recall : 0.5812
F1 : 0.6204

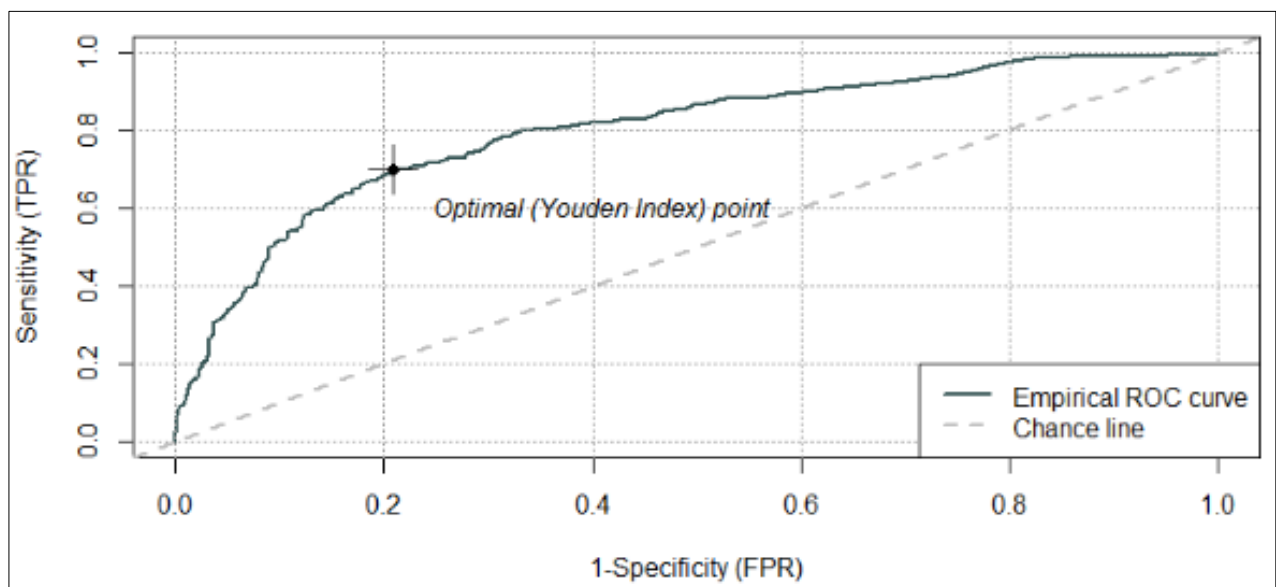
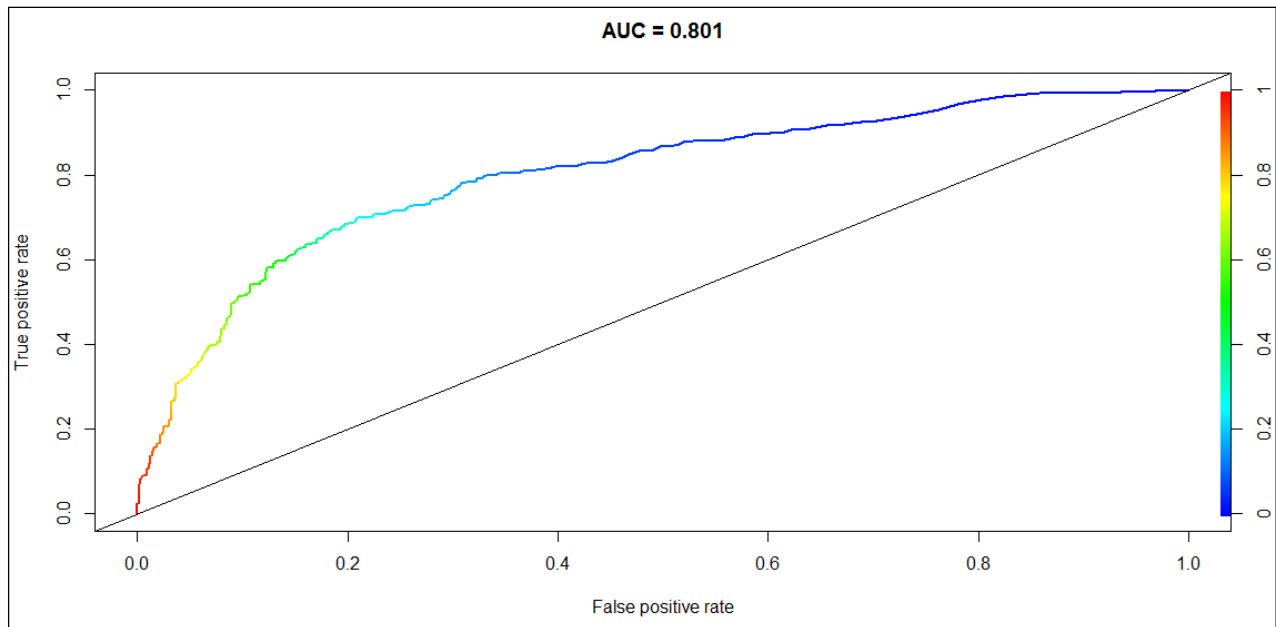
```
Prevalence : 0.3081
Detection Rate : 0.1791
Detection Prevalence : 0.2692
Balanced Accuracy : 0.7255

'Positive' Class : Yes
```

Below is the ROC curve and AUC values for the training dataset.



Below is the ROC curve and AUC value for the testing dataset.



Artificial Neural Network

Below is the confusion matrix for the trained data set, Accuracy of the model is 81.13%

```
Prediction      0      1
              0 1248   192
              1  204   455

              Accuracy : 0.8113
                95% CI : (0.7939, 0.8279)
    No Information Rate : 0.6918
    P-Value [Acc > NIR] : <2e-16

              Kappa : 0.5599

McNemar's Test P-Value : 0.5804

              Sensitivity : 0.7032
              Specificity : 0.8595
              Pos Pred Value : 0.6904
              Neg Pred Value : 0.8667
              Precision : 0.6904
              Recall : 0.7032
               F1 : 0.6968
              Prevalence : 0.3082
              Detection Rate : 0.2168
    Detection Prevalence : 0.3140
      Balanced Accuracy : 0.7814

    'Positive' Class : 1
```

Below is the confusion matrix for the trained data set, Accuracy of the model is 79.53 %

```
Confusion Matrix and Statistics

              Reference
Prediction      0      1
              0  534   96
              1   88  181

              Accuracy : 0.7953
                95% CI : (0.7674, 0.8213)
    No Information Rate : 0.6919
    P-Value [Acc > NIR] : 2.03e-12

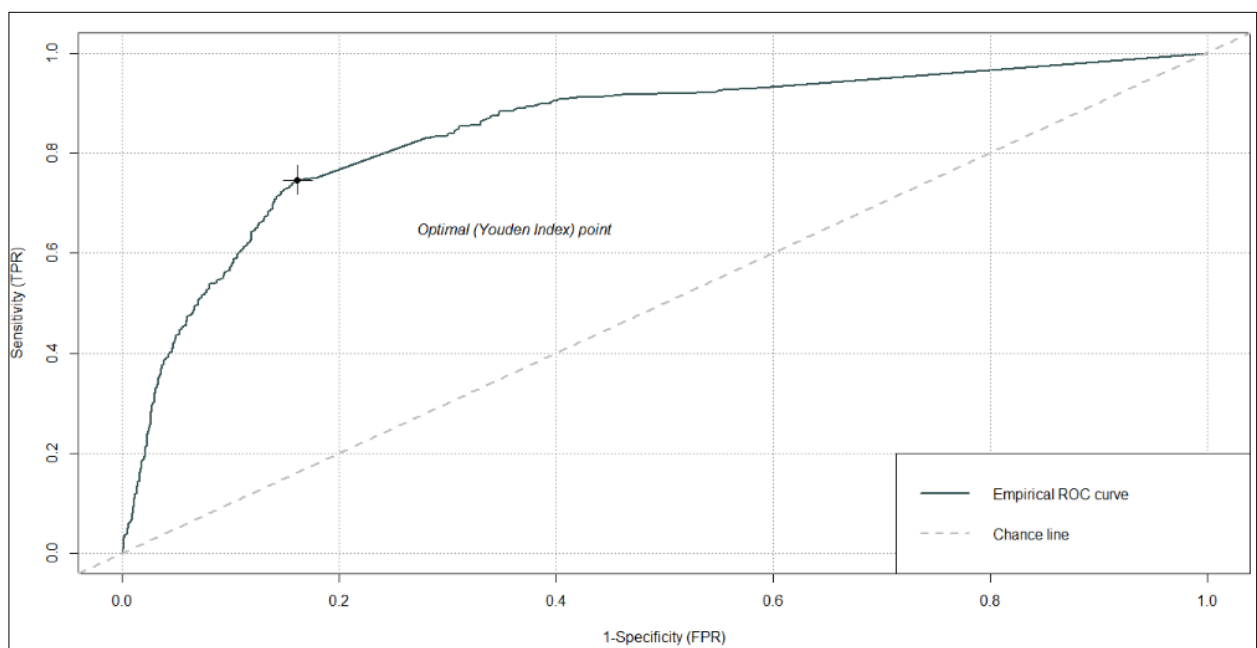
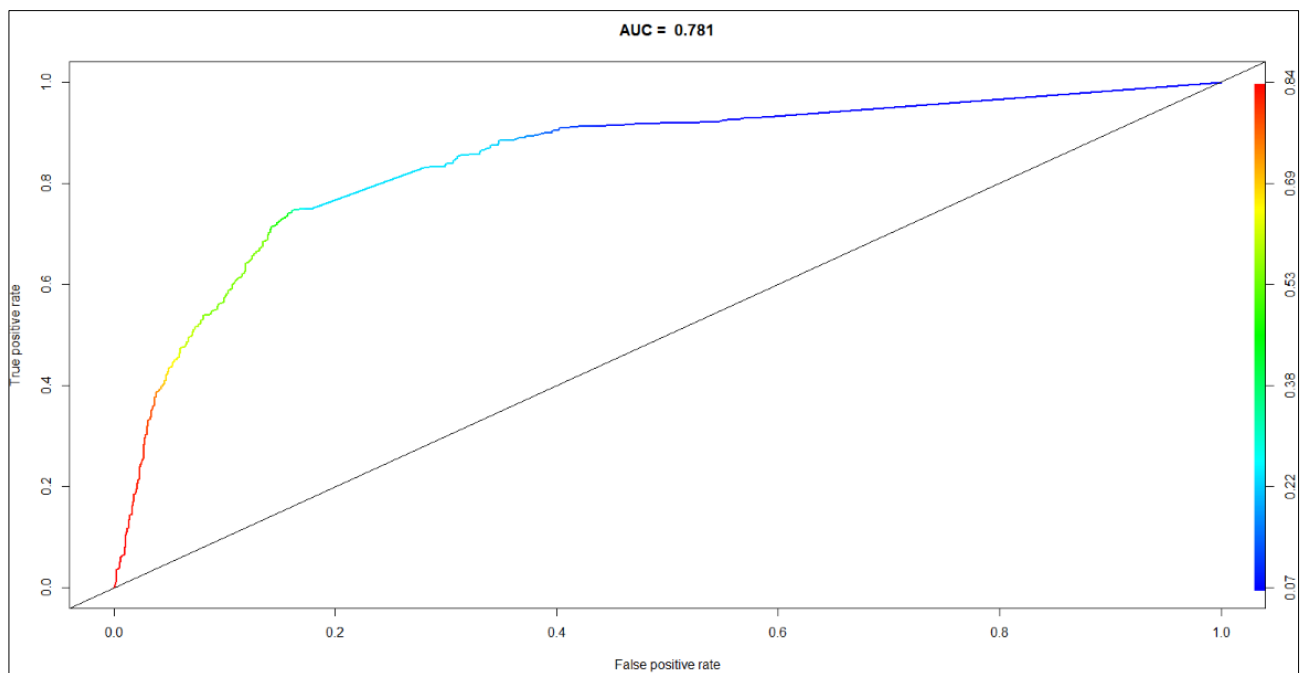
              Kappa : 0.5161

McNemar's Test P-Value : 0.6058

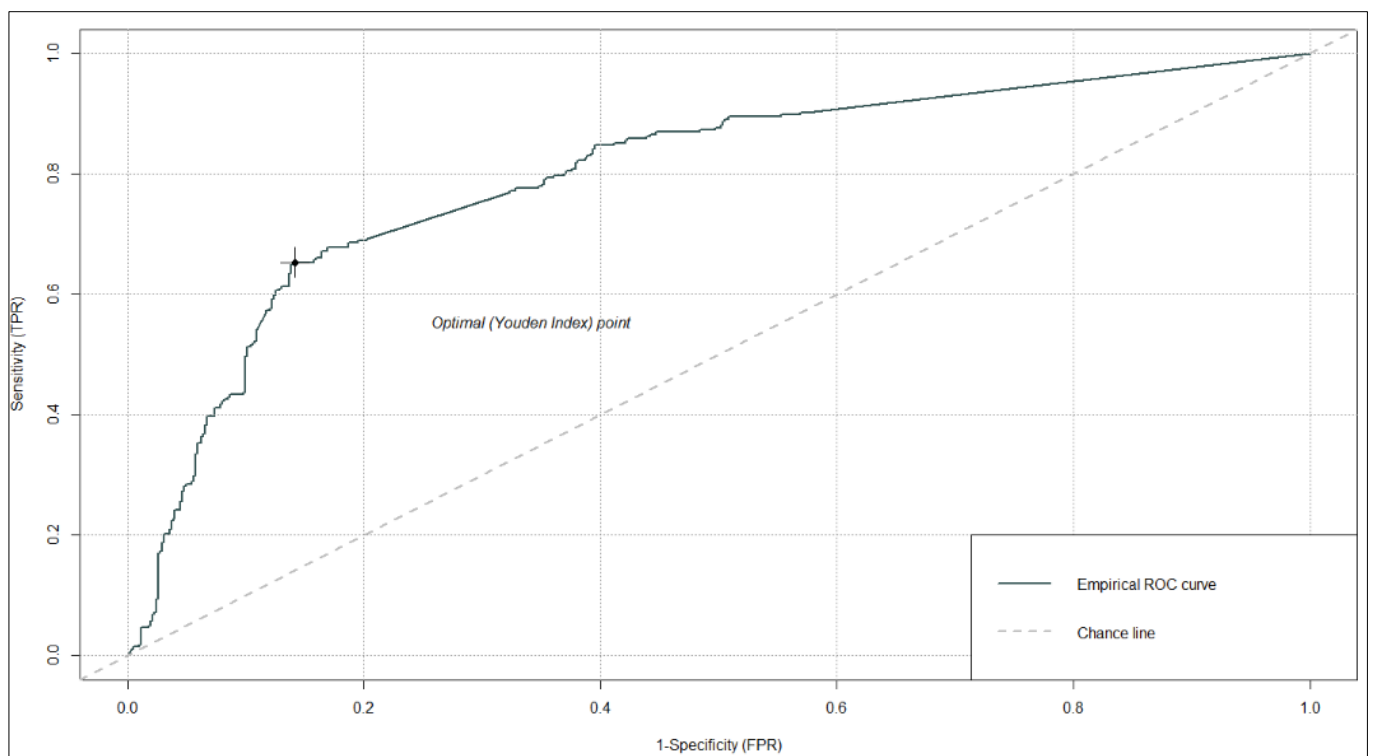
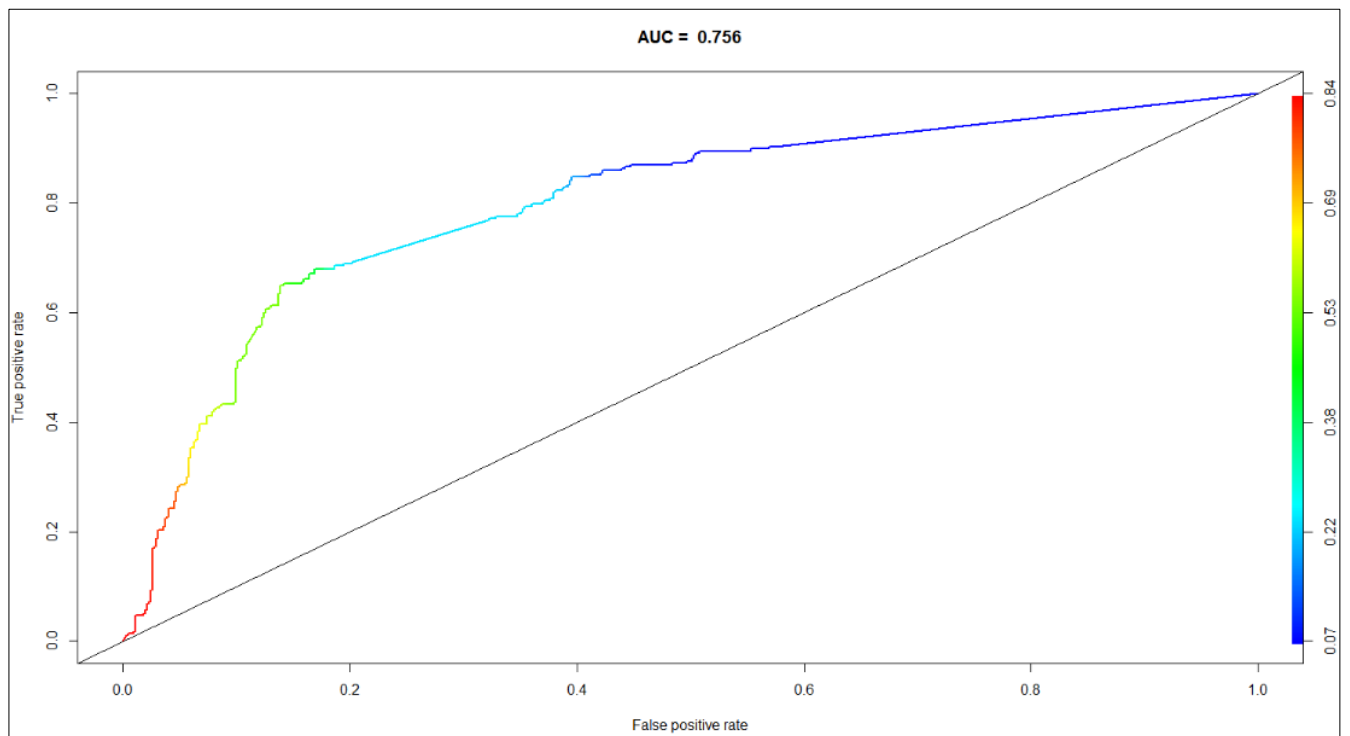
              Sensitivity : 0.6534
              Specificity : 0.8585
              Pos Pred Value : 0.6729
              Neg Pred Value : 0.8476
              Precision : 0.6729
              Recall : 0.6534
               F1 : 0.6630
              Prevalence : 0.3081
              Detection Rate : 0.2013
    Detection Prevalence : 0.2992
      Balanced Accuracy : 0.7560
```

'Positive' Class : 1

ROC curve and AUC value for the trained dataset is shown below.



The ROC curve and AUC value for the Test Dataset



4.5 Qn 4 : Final Model: Compare all the model and write an inference which model is best/optimized

Comparing all the three models the random forest seems to be best optimized and has high accuracy rate for training dataset compared to the other two. Also other values such as precision, F1 score and kappa coefficients are better for the random forest model. Hence we can conclude that the random forest is best suitable for our classification analysis.

	CART	Random Forest	ANN	Comments
Accuracy	80.56	83.52	81.13	
Sensitivity	59	67.54	70.32	Low Type II Error
Specificity	90.15	90.63	85.95	Higher value of true negative and lower false positive rate
Precision	72.76	76.27	69	Low Type I Error
F1 score	65.19	71.64	69.68	
Kappa	51.91	60.08	55.99	

For the Test dataset, ANN maintains high accuracy but specificity and precision is slightly higher in CART model compared to ANN then followed by random forest.

	CART	Random Forest	ANN	Comments
Accuracy	77.31	78.09	79.53	
Sensitivity	50.18	58.12	65.34	Low Type II Error
Specificity	89.39	86.98	85.85	Higher value of true negative and lower false positive rate
Precision	67.8	66.53	67.29	Low Type I Error
F1 score	57.68	62	66.3	
Kappa	42.64	46.74	51.61	

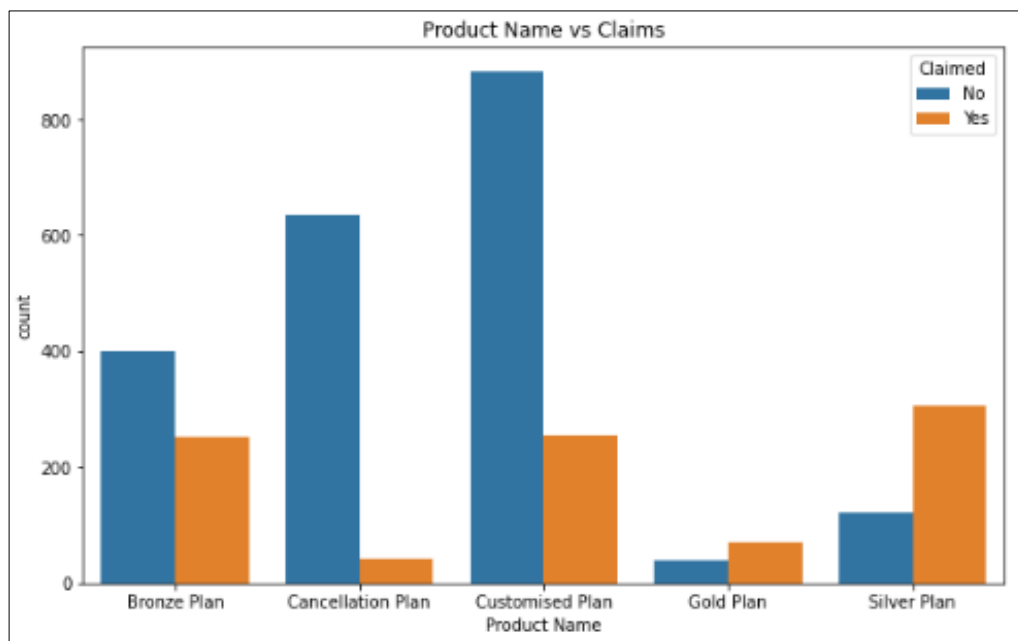
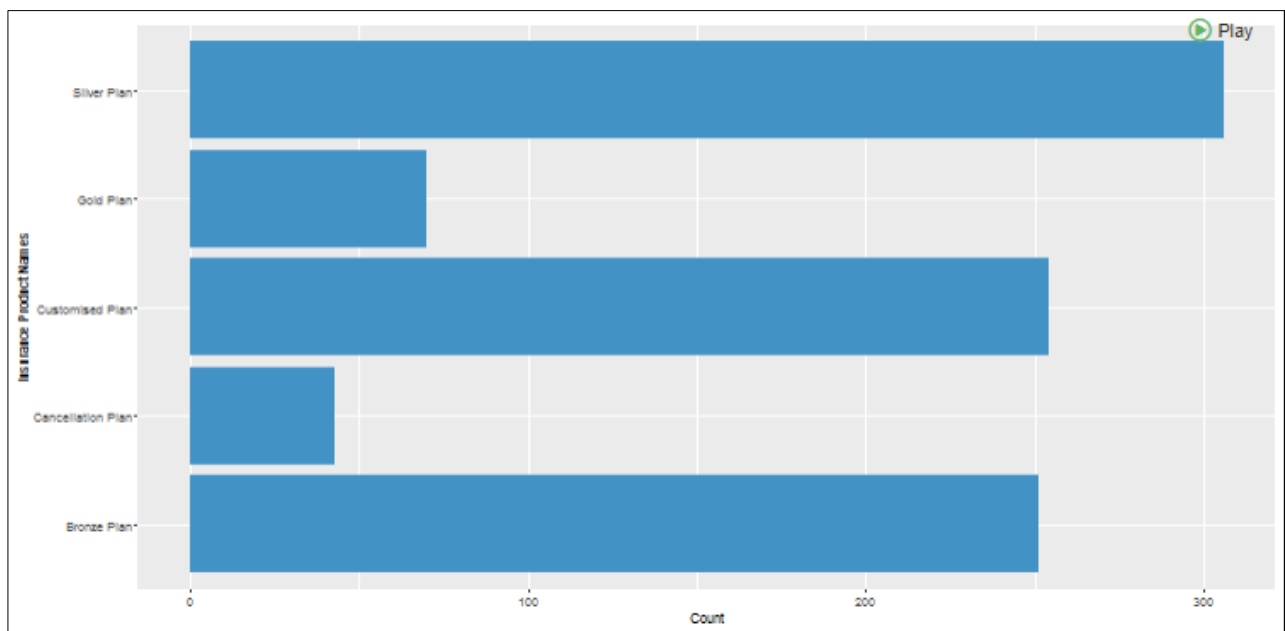
Hence on overall comparison, we can say that the Random forest has high accuracy for training set and ANN model for the testing set hence both Random Forest and ANN are best suited for the insurance claim classification model.

4.6 Qn 5 : Inference: Basis on these predictions, what are the business insights and recommendations

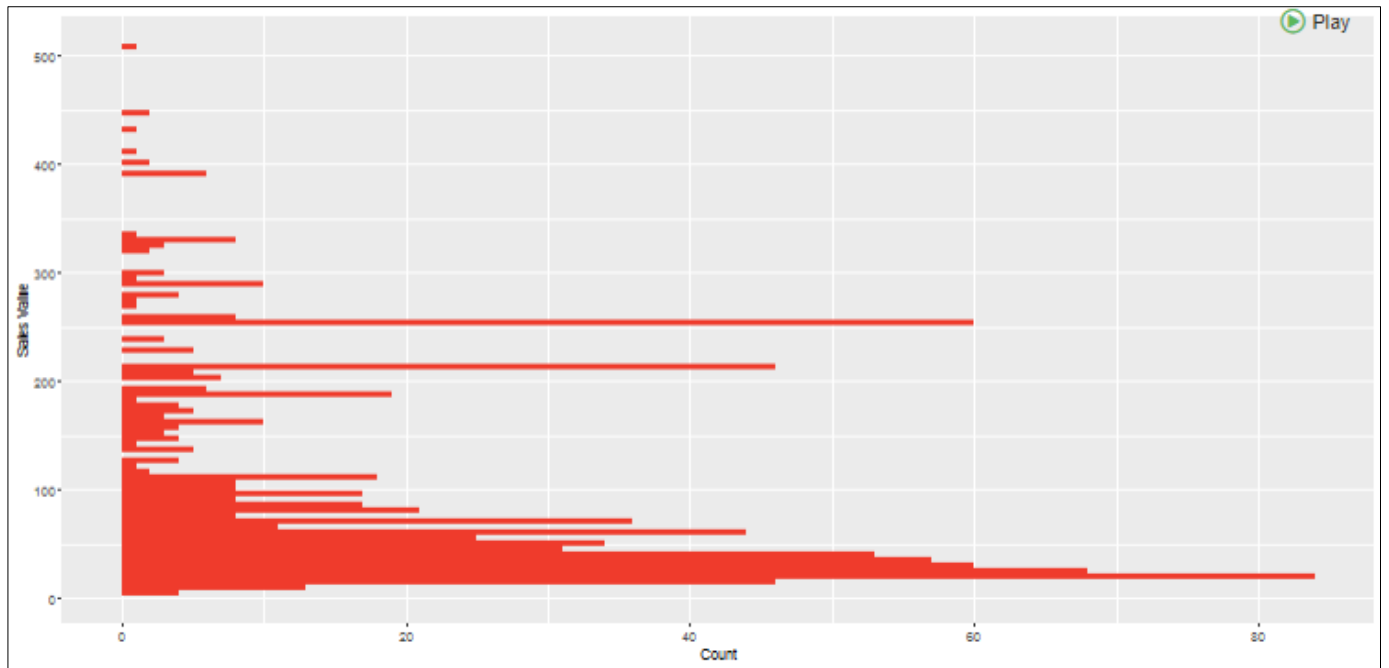
From the three machine learning models we have derived the variables or factors that are leading to higher claim for the company. We could see that the Product name, sales, agency code, commission and duration are common for all three models which are contributing to maximum claims for the insurance company with product type/name being top contributing factor for the claim. Others like age, destination and type are the least contributing factor for claims.

By sub - setting the original dataset with claims as yes, we have split the data to 924 records (all the variables having only claims as yes) and below is the analysis visualized.

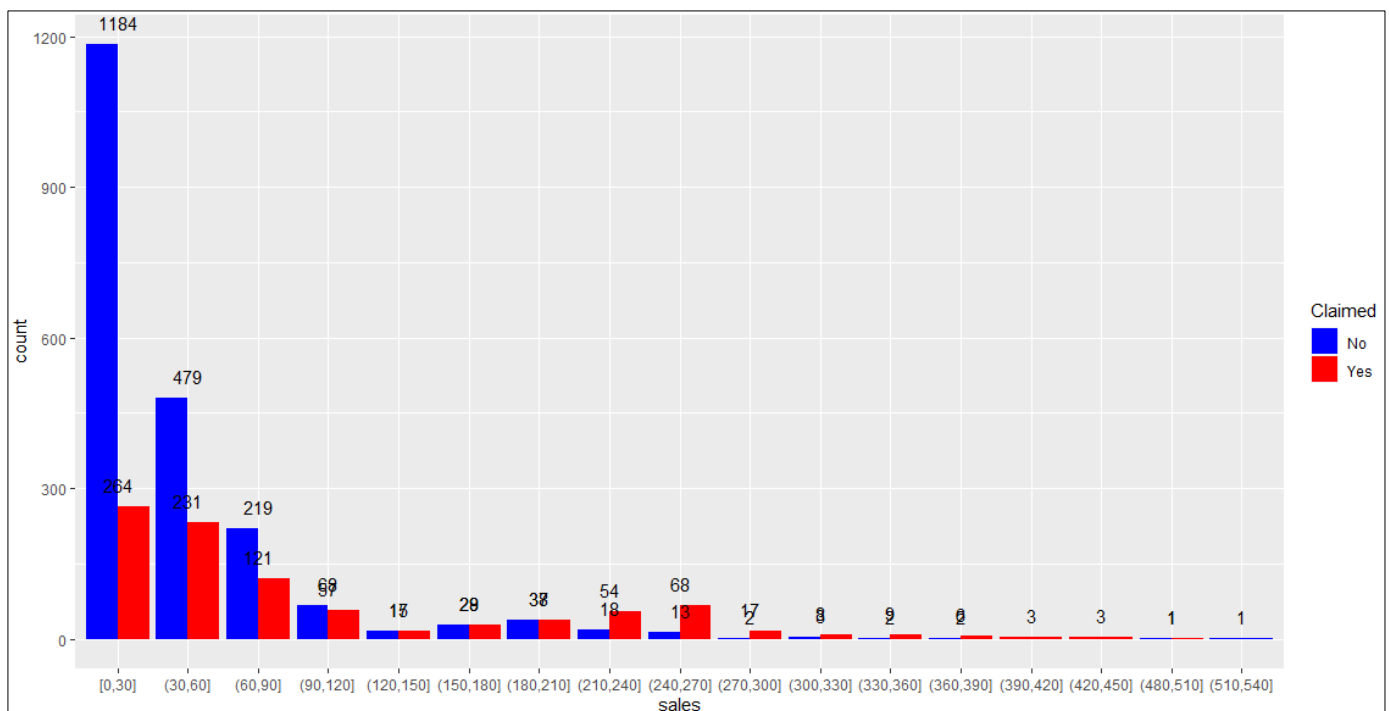
As per the ML models, the Insurance Product type held highest importance for claims hence from below graph we can see that the Silver plan has the highest claims followed by Bronze plan and gold plan of the tour insurance products.



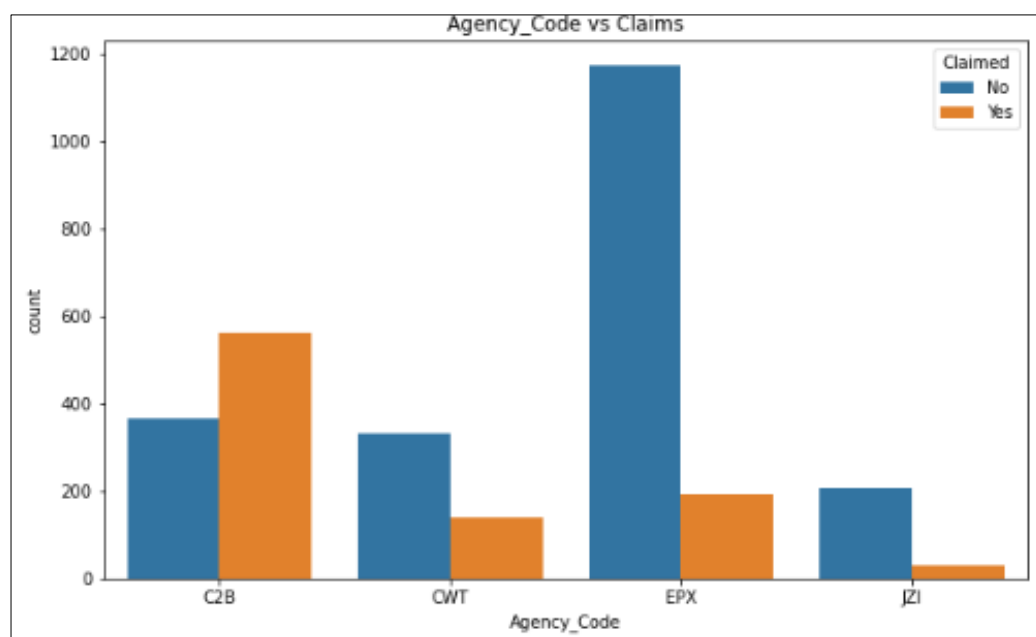
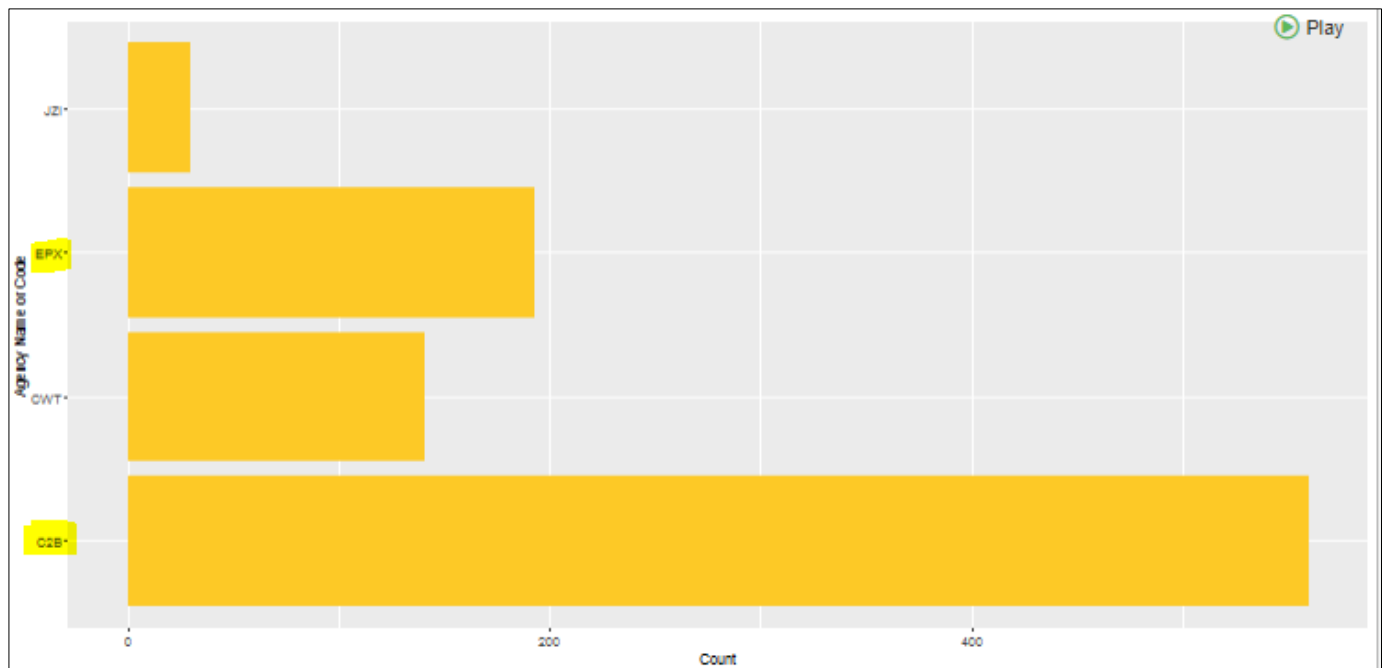
Next important variable is sales and as per the below visuals we can see that the claims are maximum for the sales value in the range between 0 to 100 units (\$). Hence we can say that the maximum claim is from the customers of low sales segment or low travel budget insurance policy segment.



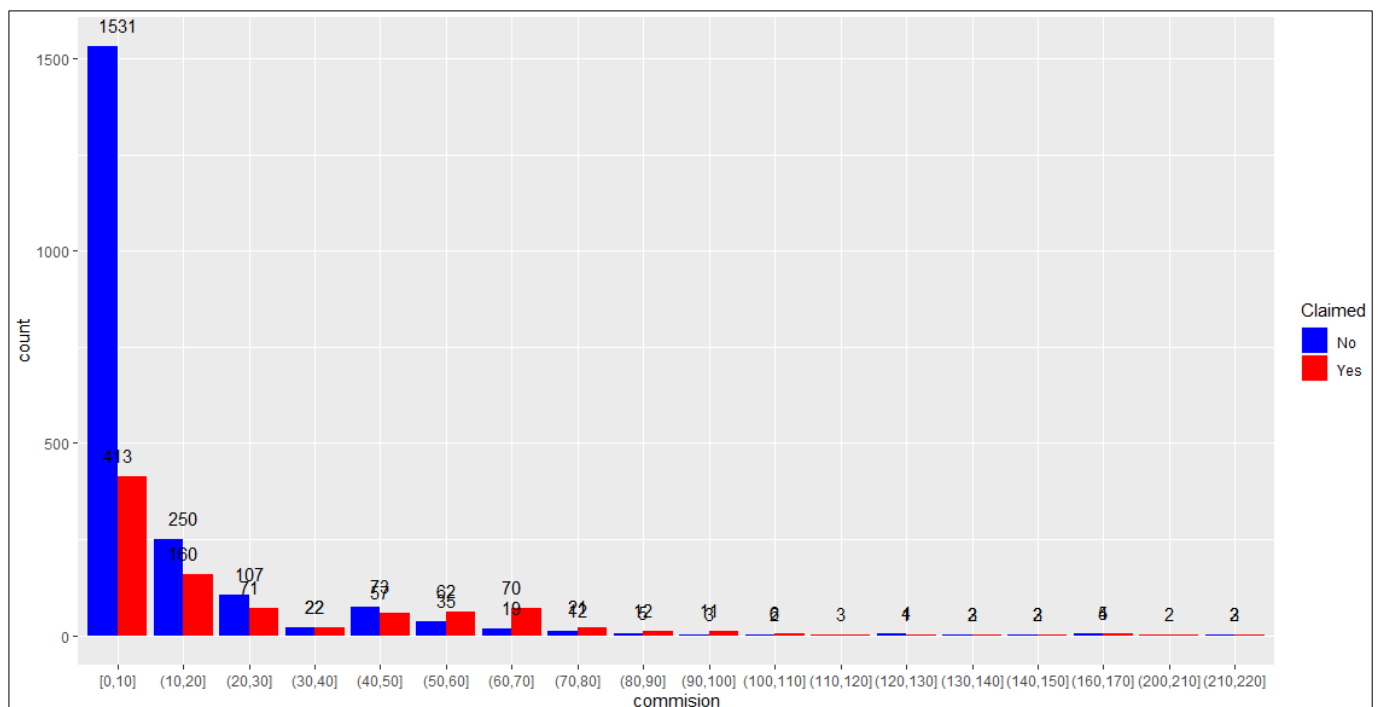
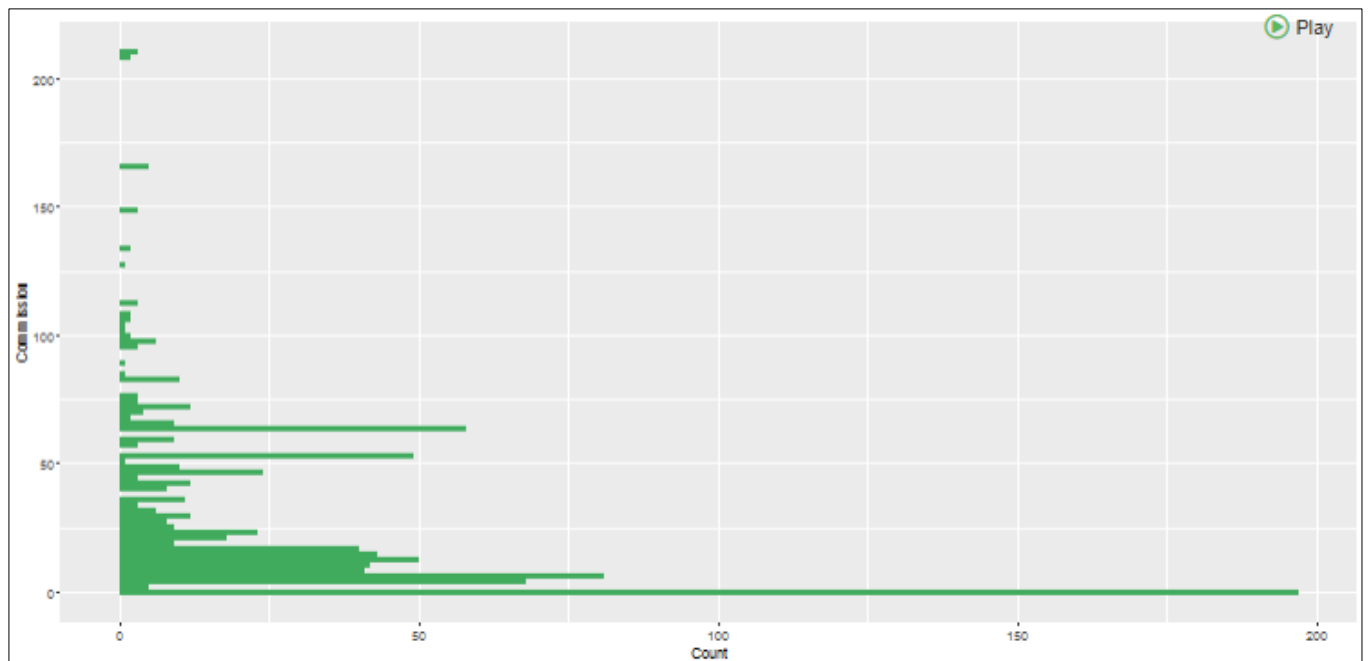
Also the sales between 200 to 300 units (\$) witnessed maximum claims.



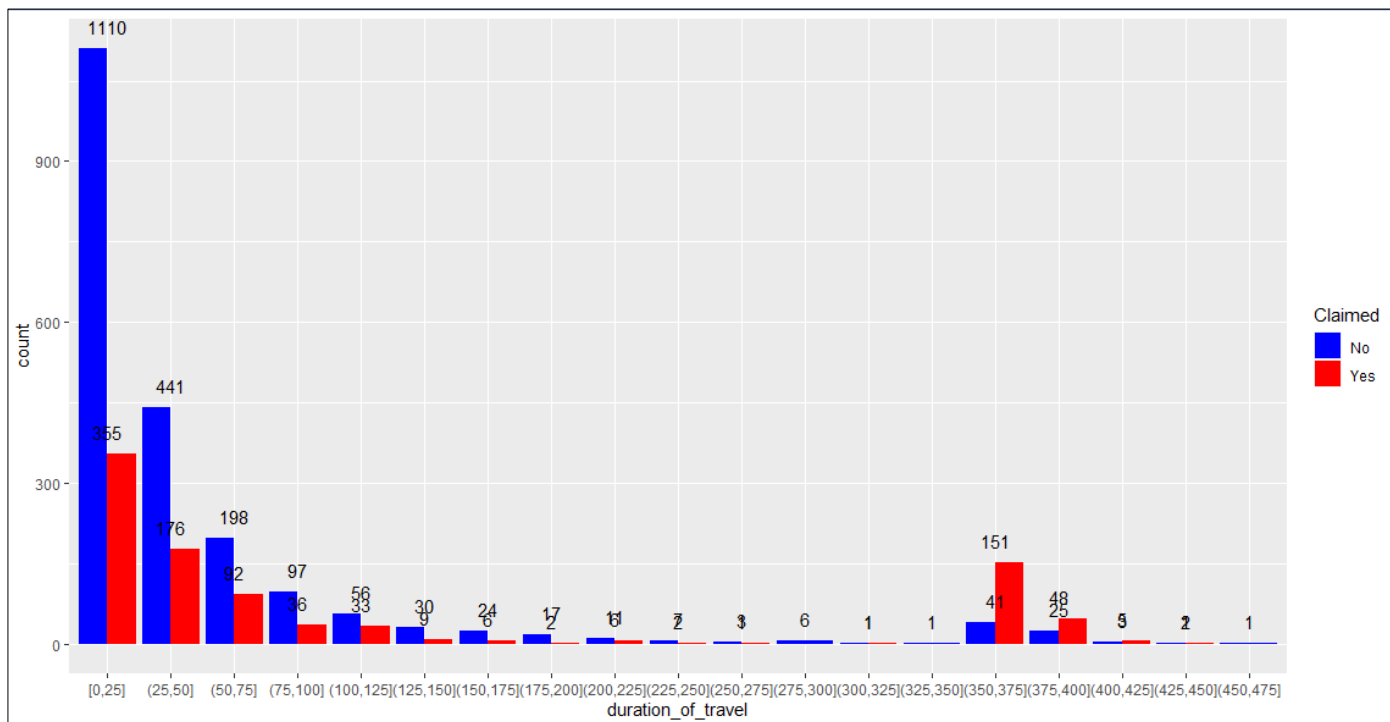
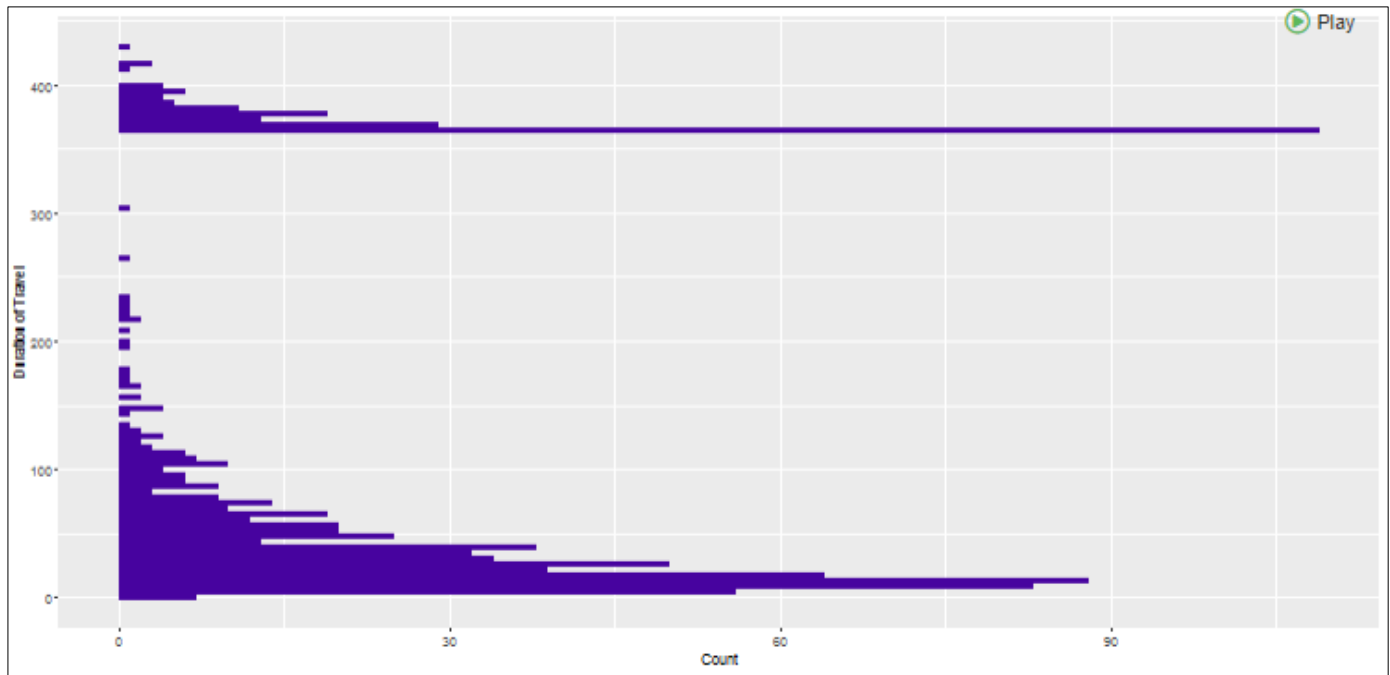
Next important factor is the agency code or Code of the tour firm, we can see that the maximum claims is from agency code '**C2B**' and surpasses all the other three agencies, hence the company needs to investigate why this tour firm has been having most highest claim frequency which may be due to several factors like injury, theft, lost baggage, illness of customer etc with agency location or any other process at their end that leads to higher claims to the insurance firm.



Next factor is the commission factor, we could see that the claims are higher for the commissions of the range between 0 to 50 units, with maximum claim frequency from customers where no commission (0 to 5 units) is charged for the customers. There is also larger claim between 60 to 70 units.

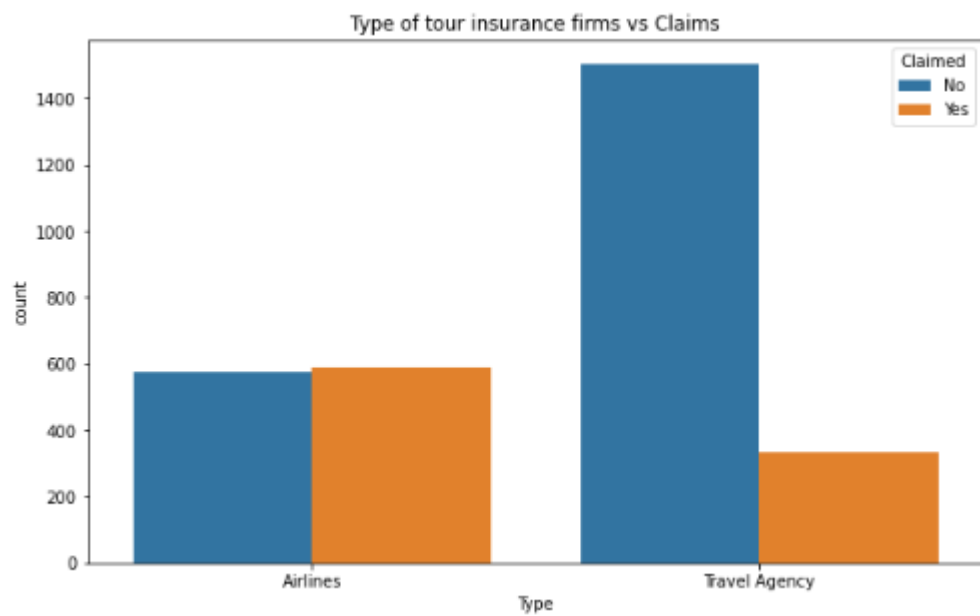
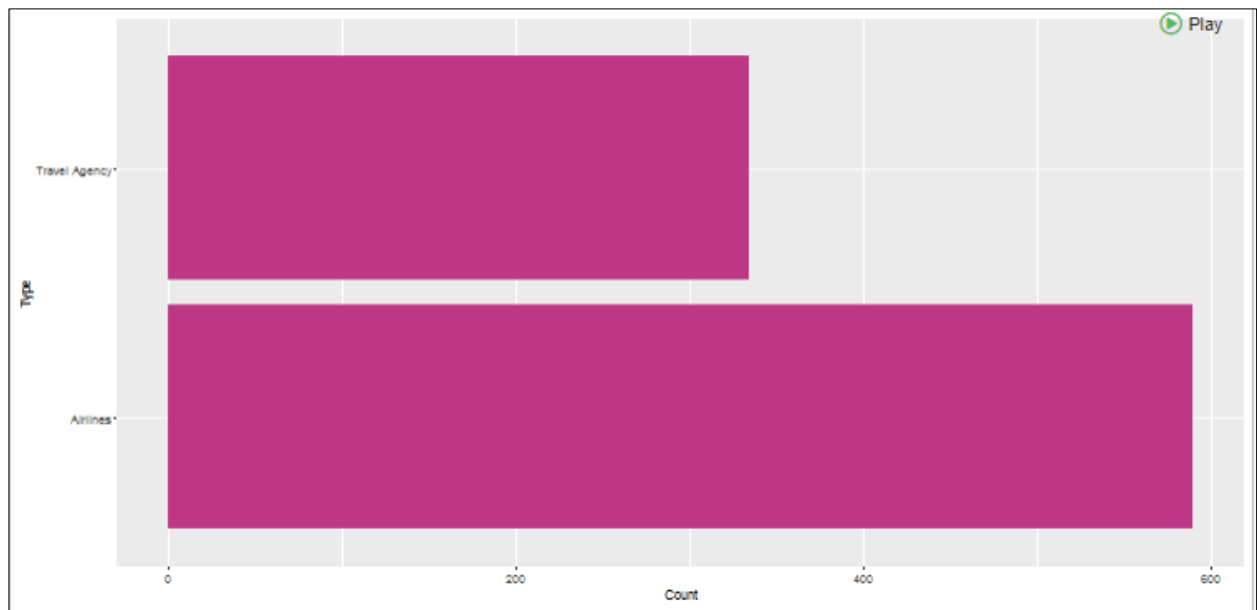


The next factor is duration of travel, we can see that the travel duration of the range between zero to fifty have the highest claims. This may be due to cancellation of travel plans or not showing up, Personal accident, missed or delayed flights . The next highest is of travel duration 370 to 400 hours which demands more claims, this can be attributed to the fact that longer duration travel increases risk hence more claims sought.

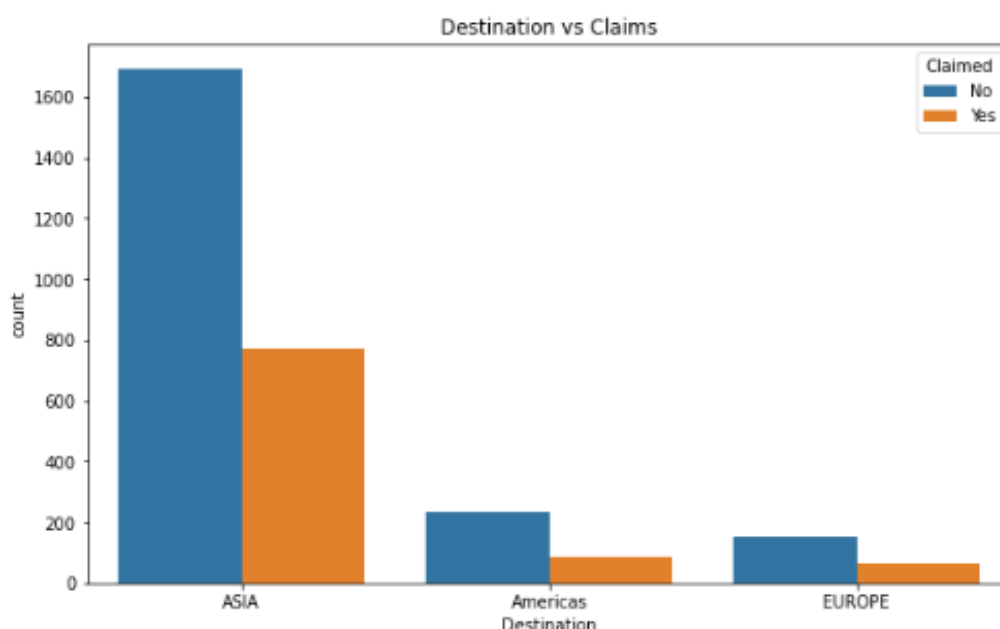
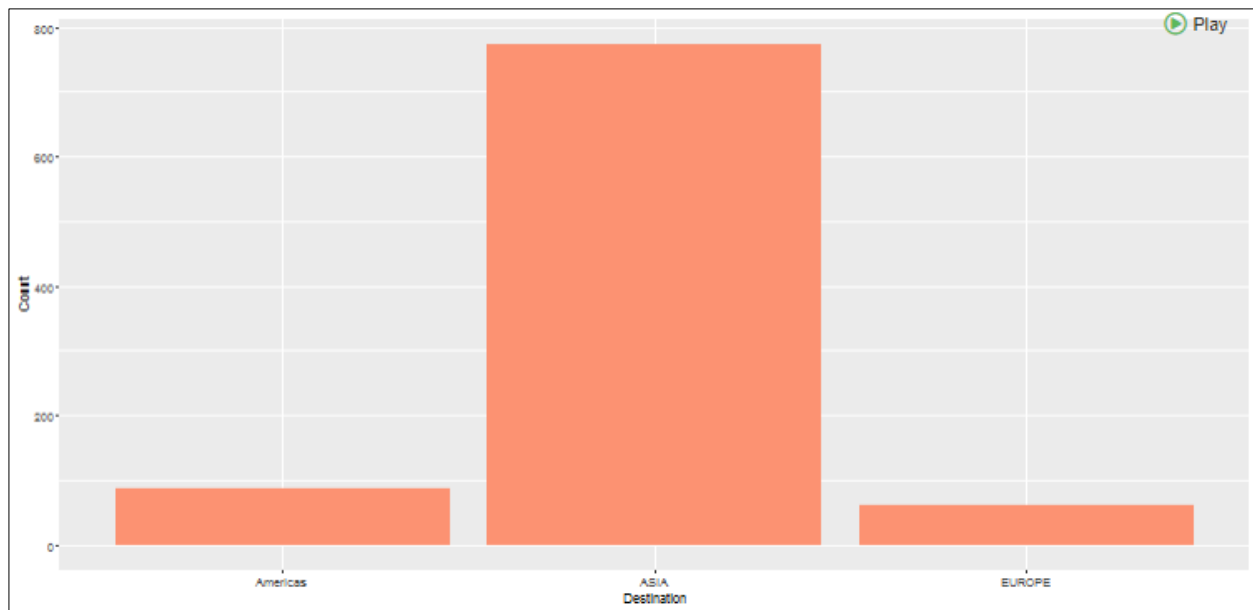


It is evidently clear that the customers travelling for duration 375 to 400 hours are likely to raise claim requests.

Next we can see that the maximum claims are from Insurance firm Type "Airlines" which is quite evident that air travel has more risk, hence more claims for this category.



The other least factors are destination, we can see maximum claims is from ASIA destination



Recommendations and Insights:

The insights gained is that the age is not to be considered for the reason for claims and that the company should primary focus on insurance products which are leading to higher claims. The insurance products may not address clearly the coverage and disclaimers hence the products can be redesigned by getting feedback from customers on the reasons why claims have been highest for products like Silver Plan, Bronze Plan and Customized plan. The reasons can be attributed to many such as medical emergency like an accident or contract an illness on the trip, victim of crime, theft of luggage and belongings such as cash , credit cards, passports etc. The company can get feedback on the profile of customer for the three products to see why these three have higher claims. Feedback and survey can help in improving the claim reduction of the products.

Next the company needs to investigate or audit why the tour firm **C2B** has the highest of the claims. IT can be due to internal factors or external factors of the firm. The insurance company can audit the firm or reach out to customers who buy the insurance from the agency to determine the reasons for the

claims. Through feedback and interviews the company can find out the causes.

The company can next look at the low sales category and the reason for the claims why the low travel budget customers have been seeking claims which may be due to crime, damages , theft etc in these travel destinations or it can be due to other factors.

Last the company can investigate the claims based on duration of travel. For low durations the company can find if the customers are raising claim request due to missed flight or missed travel plans, cancellation due to personal reasons, cancellation due to medical emergencies etc. Based on that we can better predict future claims based on past data and customer claims pattern.

5 Conclusion

We studied how the three Machine learning models can be effective in predicting the target variable and evaluated all the three models to check which was the best model . We also saw how same dataset produced nearly same prediction results and evaluated the accuracy of all the three models. We solved the business problem of a travel insurance facing claims and we proved that the four factors of product type (scheme name), agency firm , sales , duration and commission were the top contributors to the higher claims and via machine learning we assisted the firm to investigate further on the factors contributing to claims. We used the classification model to predict the claims for the training data and build the best models, we have seen that ANN & random forest was the chosen one for future predictions based on test data.

6 Appendix A – Source Code

Attached separately as file name “ML main proj CART,Random,ANN . R”