|  |
| --- |
|  |
| DM PROJECT REPORT |
| DSBA |
|  |
| **Shripad Anwekar / PGPDSBA Online Mar\_A2021** |
| **7/25/2021** |

|  |
| --- |
|  |

Contents

[1. Problem 1 Clustering 4](#_Toc78037336)

[1.1. Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis). 4](#_Toc78037337)

[Describe 4](#_Toc78037338)

[Data Pre-processing 4](#_Toc78037339)

[Data Visualization 4](#_Toc78037340)

[Bank ['spending'] 5](#_Toc78037341)

[Bank[‘advance\_payments’] 7](#_Toc78037342)

[Bank [‘probability\_of\_full\_payment’] 9](#_Toc78037343)

[Bank [‘current\_balance’] 11](#_Toc78037344)

[Bank [‘credit\_limit’] 13](#_Toc78037345)

[Bank [‘max\_spent\_in\_single\_shopping’] 15](#_Toc78037346)

[Bank [spending Vs advance\_payments] 17](#_Toc78037347)

[Bank [spending Vs probability\_of\_full\_payment] 17](#_Toc78037348)

[Bank [spending Vs current\_balance] 18](#_Toc78037349)

[Bank [spending Vs credit\_limit] 18](#_Toc78037350)

[Bank [spending Vs min\_payment\_amt] 19](#_Toc78037351)

[Bank [spending Vs max\_spent\_in\_single\_shopping] 19](#_Toc78037352)

[1.2. Do you think scaling is necessary for clustering in this case? Justify 20](#_Toc78037353)

[1.3. Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them 21](#_Toc78037354)

[1.4. Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters. 23](#_Toc78037355)

[1.5. Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters. 25](#_Toc78037356)

[2. Problem 2 CART-RF-ANN 29](#_Toc78037357)

[2.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis). 29](#_Toc78037358)

[Describe 29](#_Toc78037359)

[Data Pre-processing 29](#_Toc78037360)

[Data Visualization 29](#_Toc78037361)

[Insurance ['Age'] 30](#_Toc78037362)

[Insurance ['Commision'] 32](#_Toc78037363)

[Insurance ['Duration'] 34](#_Toc78037364)

[Insurance [' Sales'] 36](#_Toc78037365)

[Insurance [Age Vs Agency\_Code] 38](#_Toc78037366)

[Insurance [Age Vs Type] 38](#_Toc78037367)

[Insurance [Age Vs Claimed] 39](#_Toc78037368)

[Insurance [Age Vs Commision] 39](#_Toc78037369)

[Insurance [Age Vs Channel] 40](#_Toc78037370)

[Insurance [Age Vs Sales] 40](#_Toc78037371)

[Insurance [Age Vs Product Name] 41](#_Toc78037372)

[Insurance [Age Vs Destination] 41](#_Toc78037373)

[Insurance [Commision Vs Duration] 42](#_Toc78037374)

[Insurance [Commision Vs Sales] 42](#_Toc78037375)

[Insurance [Commision Vs Product Name] 43](#_Toc78037376)

[Insurance [Commision Vs Destination] 43](#_Toc78037377)

[Insurance [Duration Vs Sales] 44](#_Toc78037378)

[Insurance [Duration Vs Product Name] 44](#_Toc78037379)

[Insurance [Duration Vs Destination] 45](#_Toc78037380)

[Insurance [Sales Vs Product Name ] 45](#_Toc78037381)

[Insurance [Sales Vs Destination] 46](#_Toc78037382)

[2.2 Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network 47](#_Toc78037383)

[Cart Modelling 47](#_Toc78037384)

[Ensemble Random Forest Classifier 49](#_Toc78037385)

[Building Neural Network Model 49](#_Toc78037386)

[2.3 Performance Metrics: Comment and Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score, classification reports for each model. 50](#_Toc78037387)

[CART Model Evaluation 50](#_Toc78037388)

[Random Forest Model Evaluation 51](#_Toc78037389)

[Neural Network Model Evaluation 55](#_Toc78037390)

[Applying above best parameters returned following improved classification report 57](#_Toc78037391)

[2.4 Final Model: Compare all the models and write an inference which model is best / optimized. 59](#_Toc78037392)

[2.5 Inference: Based on the whole Analysis, what are the business insights and recommendations? 60](#_Toc78037393)

[Appendices 61](#_Toc78037394)

[List of Plots 61](#_Toc78037395)

[List of Tables 62](#_Toc78037396)

# Problem 1 Clustering

A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. You are given the task to identify the segments based on credit card usage.

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

### Describe

Supplied Data set has 7 columns, with 7 numerical values. Supplied data dictionary is reviewed against the data set and all Data types are found correct.

Quantity and relevance of data seem to be correct from Exercise stand point.

### Data Pre-processing

***Treat BAD Data***

Data types are in line with the supporting data dictionary, Presense of no Bad Data is noticed

***Treat Anomalies***

There are apparently no noticeable Anomalies, with the definition of Existing Business rules that are available

***Treat Missing Values***

There are no Missing values in the supplied DataSet

***Checking for Duplicates***

There are no Duplicate values in the supplied DataSet

### Data Visualization

#### Univariate Analysis

Following reviews are done for All 7 numeric variables to understand the trend

1. 5 Number Analysis
2. Histogram Review
3. BoxPlot
4. Empirical Rule
5. mean, median and mode

We will go through each of these variables to understand the inference of above reviews as ‘Insight’ for each of the analysis Point.

### Bank ['spending']

* 5 number Summary Suggests that -

Distribution is Right skewed –

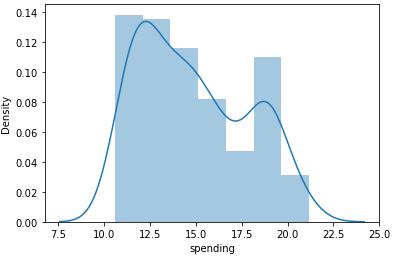
(Q1 - Min) which is 1.68 < Max - Q3 which is 3.88

Following points support the symmetry -

(Q2 - Min) which is 3.77 < (Max - Q2) which is 6.82

(Q2 - Q1) which is 2.09 < (Q3 - Q2 ) which is 2.95

* Histogram Review



***Plot 1.1.1***

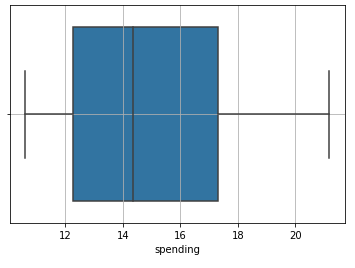
Histogram Review for Bank ['spending'] Shows -

1) 2 Peaks

2) Right Skewed

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 1.1.2***

Bank ['spending'] review of Boxplot suggests that -

1) There are No outliers

2) Distribution is RightSkewed due to longer Right Whisker

* Empirical Rule

Bank ['spending'] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Right Skewed)

2) '2 standard deviation' from Mean is asymmetrical (Right Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Right Skewed)

* Bank ['spending'] review of Mean, Median and Mode suggests that –

In this case Mean > Median > Mode hence the distribution is Right Skewed

**Insights (spending):**

Amount spent by the customer per month (in 1000s) range from 10.59K to 21.18K

75% of the customer spend less than 17.30K .

From above figure, we can say that the Amount spent by the customer per month is right skewed

Average spend by customer is 14.856K which is marginally higher than the median value 14.36K indicating that the distribution is right tailed.

There are no outliers which supports data consistency.

### Bank[‘advance\_payments’]

* 5 number Summary Suggests that -

Distribution is Right skewed –

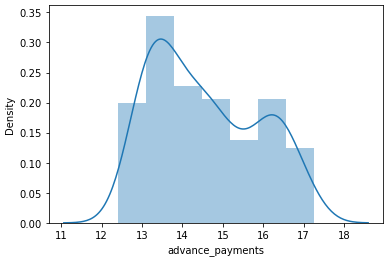
(Q1 - Min) which is 1.04 < Max - Q3 which is 1.54

Following points support the symmetry -

(Q2 - Min) which is 1.91 < (Max - Q2) which is 2.93

(Q2 - Q1) which is 0.87 < (Q3 - Q2) which is 1.39

* Histogram Review



***Plot 1.1.3***

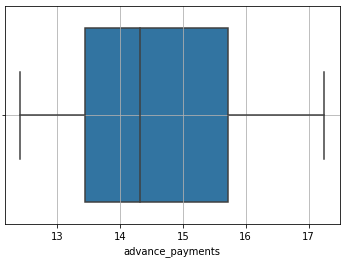
Histogram Review for Bank [‘advance\_payments’] Shows -

1) 2 Peaks

2) Right Skewed

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 1.1.4***

Bank ['advance\_payments'] review of Boxplot suggests that -

1) There are No outliers

2) Distribution is RightSkewed due to longer Right Whisker

* Empirical Rule

Bank ['advance\_payments'] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Right Skewed)

2) '2 standard deviation' from Mean is asymmetrical (Right Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Right Skewed)

* Bank ['advance\_payments'] review of Mean, Median and Mode suggests that –

In this case Mean > Median > Mode hence the distribution is Right Skewed

**Insights (**advance\_payments**):**

Amount paid by the customer in advance by cash range from 1.241K to 1.725K

75% of the customer's advance payment is less than 1.571K.

From above figure, we can say that the Amount spent by the customer per month is right skewed

Average spend by customer is 1.456K which is marginally higher than the median value 1.432K indicating that the distribution is right tailed.

There are no outliers which show data consistency.

### Bank [‘probability\_of\_full\_payment’]

* 5 number Summary Suggests that -

Distribution is Left skewed –

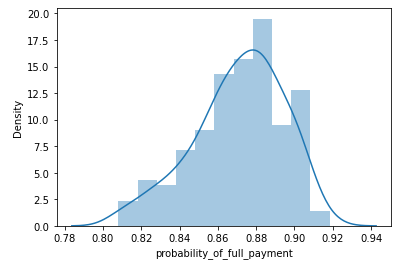
(Q1 - Min) which is 0.05 > Max - Q3 which is 0.03

Following points support the symmetry -

(Q2 - Min ) which is 0.07 > (Max - Q2) which is 0.04

(Q2 - Q1 ) which is 0.02 > (Q3 - Q2 ) which is 0.01

* Histogram Review



***Plot 1.1.5***

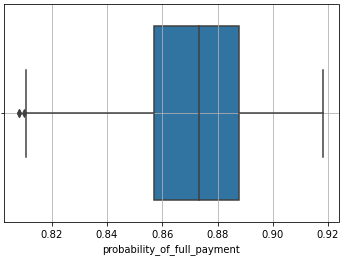
Histogram Review for Bank [‘probability\_of\_full\_payment'] Shows -

1) 1 Peak

2) Left Skewed

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 1.1.6***

Bank [probability\_of\_full\_payment'] review of Boxplot suggests that -

1) There are couple of outliers

2) Distribution is Left Skewed due to longer Left Whisker

* Empirical Rule

Bank [‘probability\_of\_full\_payment’] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Left Skewed)

2) '2 standard deviation' from Mean is asymmetrical (Left Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Left Skewed)

* Bank [‘probability\_of\_full\_payment’] review of Mean, Median and Mode suggests that –

In this case Mean < Median < Mode hence the distribution is Left Skewed

**Insights (probability\_of\_full\_payment):**

**Probability of payment done in full by the customer to the bank Range from 0.81 to 0.92**

**Probability that 75% of the customers will pay in full is less than 0.89 . From above figure, we can say that the Amount spent by the customer per month is Letf skewed**

**Average probability as well as median value of probability that customer will pay in full is same and is 0.87**

**There are couple of Outliers towards lower probability of full payment**

### Bank [‘current\_balance’]

* 5 number Summary Suggests that -

Distribution is Right skewed –

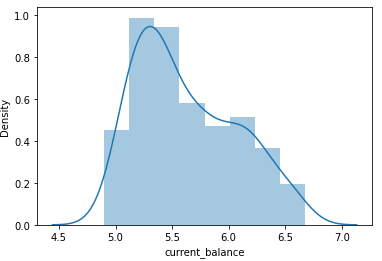
(Q1 - Min) which is 0.36 < Max - Q3 which is 0.70

Following points support the symmetry -

(Q2 - Min) which is 0.62 < (Max - Q2) which is 1.15

(Q2 - Q1) which is 0.26 < (Q3 - Q2) which is 0.46

* Histogram Review



***Plot 1.1.7***

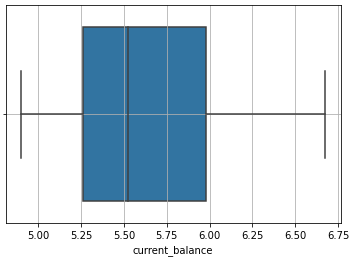
Histogram Review for Bank [‘current\_balance’] Shows -

1) 2 Peaks

2) Right Skewed

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 1.1.8***

Bank [current\_balance'] review of Boxplot suggests that -

1) There are No outliers

2) Distribution is Right Skewed due to longer Right Whisker

* Empirical Rule

Bank [‘current\_balance’] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Right Skewed)

2) '2 standard deviation' from Mean is asymmetrical (Right Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Right Skewed)

* Bank [‘current\_balance'] review of Mean, Median and Mode suggests that –

In this case Mean > Median > Mode hence the distribution is Right Skewed

**Insights (current\_balance):**

**Balance amount left in the account to make purchases (in 1000s)range from 4.90K to 6.67K**

**75% of the customer's balance left is less than 5.98K.**

**From above figure, we can say that the Amount spent by the customer per month is right skewed**

**Average spend by customer is 5.63K which is marginally higher than the median value 5.52K indicating that the distribution is right tailed.**

**There are no outliers Outliers which shows data consistency.**

### Bank [‘credit\_limit’]

* 5 number Summary Suggests that -

Distribution is Right skewed –

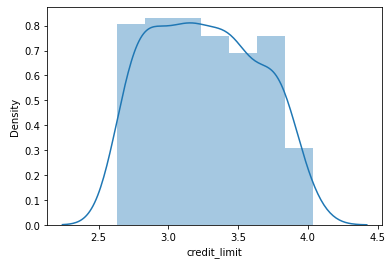
(Q1 - Min) which is 0.31 < Max - Q3 which is 0.47

Following points support the symmetry –

(Q2 - Min) which is 0.61 < (Max - Q2) which is 0.80

(Q2 - Q1) which is 0.29 < (Q3 - Q2) which is 0.32

* Histogram Review



***Plot 1.1.9***

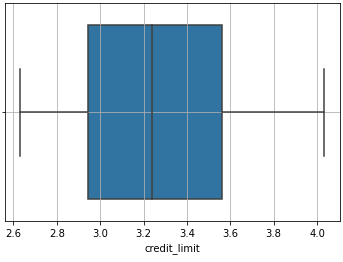
Histogram Review for Bank [‘credit\_limit’] Shows -

1) Flat Top

2) Right Skewed

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 1.1.10***

Bank [‘credit\_limit’] review of Boxplot suggests that -

1) There are No outliers

2) Distribution is Right Skewed due to longer Right Whisker

* Empirical Rule

Bank [‘credit\_limit’] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Right Skewed)

2) '2 standard deviation' from Mean is asymmetrical (Right Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Right Skewed)

* Bank [‘credit\_limit’] review of Mean, Median and Mode suggests that –

In this case Mean > Median > Mode hence the distribution is Right Skewed

**Insights (credit\_limit):**

**Limit of the amount in credit card (10000s) range from 26.3K to 40.03K**

**75% of the customer’s credit limit is less than 35.6K .**

**From above figure, we can say that the Amount spent by the customer per month is right skewed**

**Average spend by customer is 32.6 K which is marginally higher than the median value 32.4K indicating that the distribution is right tailed.**

**There are no Outliers which shows data consistency.**

### Bank [‘max\_spent\_in\_single\_shopping’]

* 5 number Summary Suggests that -

Distribution is Right skewed –

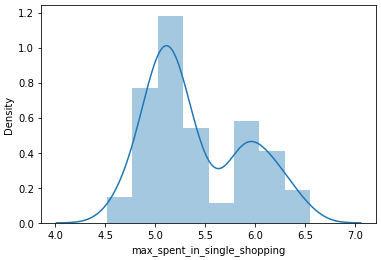
(Q1 - Min) which is 0.53 < Max - Q3 which is 0.67

Following points support the symmetry –

(Q2 - Min ) which is 0.70 < (Max - Q2) which is 1.33

(Q2 - Q1 ) which is 0.18 < (Q3 - Q2 ) which is 0.65

* Histogram Review



***Plot 1.1.11***

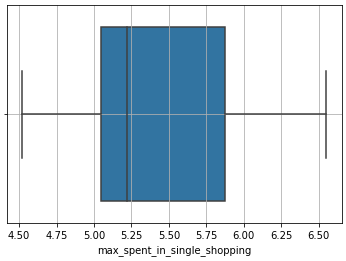
Histogram Review for Bank [‘max\_spent\_in\_single\_shopping’] Shows -

1) 2 Peaks

2) Right Skewed

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 1.1.12***

Bank [‘max\_spent\_in\_single\_shopping’] review of Boxplot suggests that -

1) There are No outliers

2) Distribution is Right Skewed due to longer Right Whisker

* Empirical Rule

Bank [‘max\_spent\_in\_single\_shopping’] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Right Skewed)

2) '2 standard deviation' from Mean is asymmetrical (Right Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Right Skewed)

* Bank [‘max\_spent\_in\_single\_shopping’] review of Mean, Median and Mode suggests that –

In this case Mean > Median > Mode hence the distribution is Right Skewed

**Insights (max\_spent\_in\_single\_shopping):** Minimum paid by the customer while making payments for purchases made monthly (in 100s) range from 0.452K to 0.655K

75% of the customer's credit limit is less than 0.588K .

From above figure, we can say that the Amount spent by the customer per month is right skewed

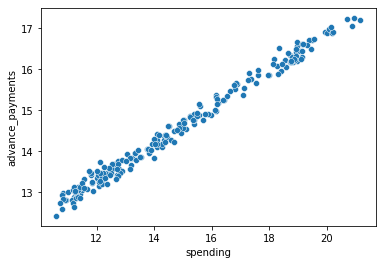
Average spend by customer is 0.541 K which is mirginally higher than the median value 0.522K indicating that the ditribution is right tailed.

There are no outliers which show data consistency.

#### Bivariate Analysis

Please find BiVariate Analysis performed on Attributes pairing target field spending to demonstrate the approach as well as illustrate typical insights that can be drawn. Please note we only have continuous numerical fields as part of the supplied DataSet.

### Bank [spending Vs advance\_payments]

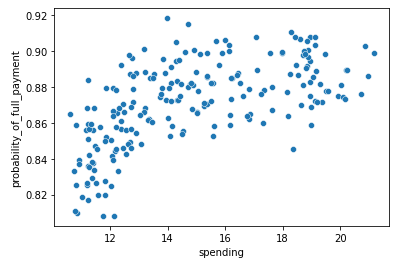


**INSIGHT**

Advance Payment and Spending have a linear relationship, i.e. customers who spend more tend to make advance payment in then proportion of the spend

***Plot 1.1.13***

### Bank [spending Vs probability\_of\_full\_payment]

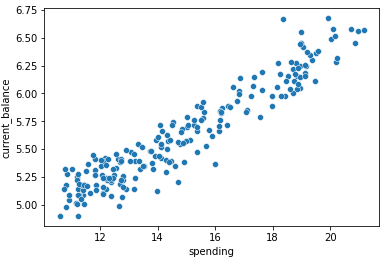


**INSIGHT**

Probability\_of\_full\_payment increases with spend but then becomes steady as spend increase

***Plot 1.1.14***

### Bank [spending Vs current\_balance]

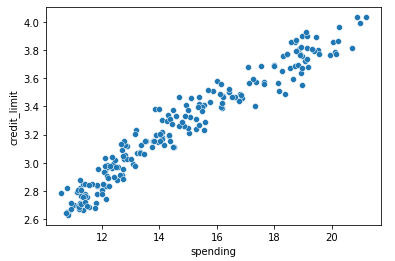


**INSIGHT**

Current balance and Spending have a linear relationship, i.e. customers who spend more tend to have current balance in the proportion of the spend

***Plot 1.1.15***

### Bank [spending Vs credit\_limit]

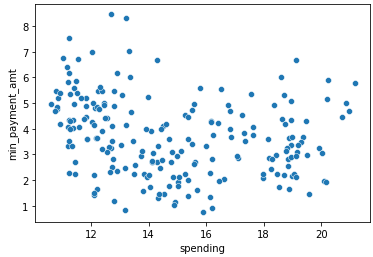


**INSIGHT**

Credit limit and Spending have a linear relationship, i.e. customers who spend more tend to have credit limit in the proportion of the spend

***Plot 1.1.16***

### Bank [spending Vs min\_payment\_amt]

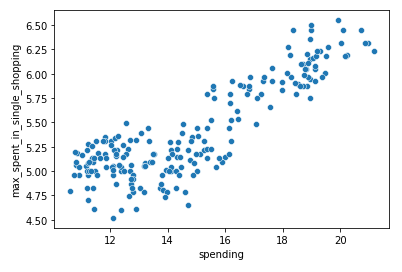


**INSIGHT**

Minimum Payment Amount and Spending have a really scattered relationship, i.e. customers who spend less can make more payment as minimum payment and vice a versa

***Plot 1.1.17***

### Bank [spending Vs max\_spent\_in\_single\_shopping]



**INSIGHT**

Customers who spend more also tend to have spent maximum in single shopping, which is as expected behaviour

***Plot 1.1.18***

## Do you think scaling is necessary for clustering in this case? Justify

Clustering calculate a new segmentation based on variance/Variability (SD Deviation) of the variable in consideration. This means variables with higher (SD Deviation) will have higher weightage. In the normalized data where we have 'Zero' as mean and 'One' as Std deviation we will have same weight magnitude and the Clustering calculates relevant axis to consider.

Secondly the unit of measurement of the variables includes Numbers having a lot of difference within their magnitude

To Summarize, Reason that we must scale the data is

* In the un-scaled data standard deviation varies from 0.02 to 2.91.
* In the un-scaled data Mean varies from 0.87 to 14.85.
* Variable and their varied Measurement of Units

|  |  |
| --- | --- |
| **Variable** | **Multiplier** |
| spending | 1000 |
| advance\_payments | 100 |
| probability\_of\_full\_payment | 1 |
| current\_balance | 1000 |
| credit\_limit | 10000 |
| min\_payment\_amt | 100 |
| max\_spent\_in\_single\_shopping | 1000 |

***Table 1.2.1***

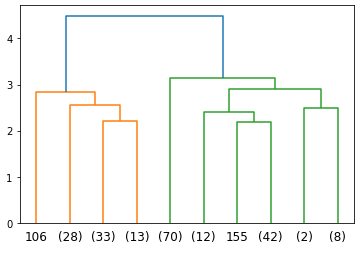
## Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them

Following steps are followed to conclude on the number of Clusters that will logically segment the observations around the target attribute-

* Link Method is decided as Average, as it offers a compromise between
  + The sensitivity of complete-link clustering to outliers
  + The tendency of single-link clustering to form long chains
* A Dendrogram was plotted with link Method as ‘Averag’ on the scaled data.

We apply **Truncate\_mode** as ‘ lastp’ to Truncation condense the dendrogram for better

Read



***Plot 1.3.1***

* From the Dendogram it is evident that there are two clear segments distinguished hence we can assign 2 clusters however second cluster shows the possibility of division into further two though there is no significant distance noticed. So for further Analysis it is advisable to conclude and segment the given observations into **3 CLUSTERS**
* The criterion to use in forming flat clusters is decided as ‘maxclust’, to ensure no more than 3 clusters are formed.
* A column with assigned value for a particular observation is added to store the assigned ‘cluster no’ to unscaled (Bank) data set.
* Following is the distribution brief of the given data set post assignment of Clusters

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **clusters** | **spending** | **advance\_payments** | **probability\_of\_full\_payment** | **current\_balance** | **credit\_limit** | **min\_payment\_amt** | **max\_spent\_in\_single\_shopping** | **Freq** |
| **2** | 11.92 | 13.29 | 0.85 | 5.26 | 2.85 | 4.62 | 5.12 | 70 |
| **3** | 14.22 | 14.2 | 0.88 | 5.44 | 3.25 | 2.77 | 5.06 | 65 |
| **1** | 18.13 | 16.06 | 0.88 | 6.14 | 3.65 | 3.65 | 5.99 | 75 |

***The outcome, inference and strategy as result of above clustering exercise will be discussed in Question 1.5***

***Table 1.3.1***

## Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters.

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable **K**. ... Data points are clustered based on feature similarity.

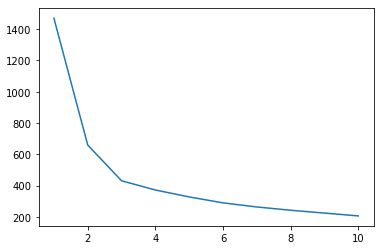
Inertia measures how well a dataset was clustered by K-Means. It is calculated by measuring the distance between each data point and its centroid, squaring this distance, and summing these squares across one cluster. A good model is one with low inertia AND a low number of clusters (K).

In the given scenario, we have experimented to find the Inertia for 11 clusters to compare and conclude on the optimal ‘K’ value-

|  |  |  |  |
| --- | --- | --- | --- |
| **K' Value** | **Inertia** | **Difference between Inertia** | **% Drop from Previous 'K'** |
| 1 | 1470.0 | 1470.0 |  |
| 2 | 659.1 | 810.9 | 55% |
| 3 | 430.5 | 228.6 | 28% |
| 4 | 371.5 | 59.0 | 26% |
| 5 | 327.2 | 44.3 | 75% |
| 6 | 289.2 | 38.0 | 86% |
| 7 | 263.0 | 26.2 | 69% |
| 8 | 241.9 | 21.1 | 81% |
| 9 | 223.9 | 18.0 | 85% |
| 10 | 205.7 | 18.2 | 101% |

***Table 1.4.1***

Inertia doesn’t drop sharply beyond K=3, We further plotted the trend and elbow curve on WSS supports that K Value of 3 is optimal for the dataset in question.



***Plot 1.4.1***

Following is the summary of silhouette\_score indicating that K Value of 3 is optimum for the given dataset for further segmentation

|  |  |  |
| --- | --- | --- |
| Cluster | silhouette\_score | Diff with Earlier |
| 2 | 0.46578 | 0.0 |
| 3 | 0.40086 | 0.065 |
| 4 | 0.32939 | 0.071 |
| 5 | 0.28648 | 0.043 |
| 6 | 0.28888 | -0.002 |
| 7 | 0.27956 | 0.009 |
| 8 | 0.26633 | 0.013 |
| 9 | 0.25828 | 0.008 |
| 10 | 0.24532 | 0.013 |
| 11 | 0.25434 | -0.009 |

***Table 1.4.2***

From Above two tables it is clear that K=3 is the optimal number that K-Means clustering can be used for segmentation

## Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters.

Following table shows the comparison between Hierarchical and K- Means clustering.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| K- Means | **Clus\_kmeans4** | **spending** | **advance\_payments** | **probability\_of\_full\_payment** | **current\_balance** | **credit\_limit** | **min\_payment\_amt** | **max\_spent\_in\_single\_shopping** | **Freq** |
| **1** | 11.86 | 13.25 | 0.85 | 5.23 | 2.85 | 4.74 | 5.1 | 72 |
| **0** | 14.44 | 14.34 | 0.88 | 5.51 | 3.26 | 2.71 | 5.12 | 71 |
| **2** | 18.5 | 16.2 | 0.88 | 6.18 | 3.7 | 3.63 | 6.04 | 67 |
| Hierarchical | **clusters** | **spending** | **advance\_payments** | **probability\_of\_full\_payment** | **current\_balance** | **credit\_limit** | **min\_payment\_amt** | **max\_spent\_in\_single\_shopping** | **Freq** |
| **2** | 11.92 | 13.29 | 0.85 | 5.26 | 2.85 | 4.62 | 5.12 | 70 |
| **3** | 14.22 | 14.2 | 0.88 | 5.44 | 3.25 | 2.77 | 5.06 | 65 |
| **1** | 18.13 | 16.06 | 0.88 | 6.14 | 3.65 | 3.65 | 5.99 | 75 |

***Table 1.5.1***

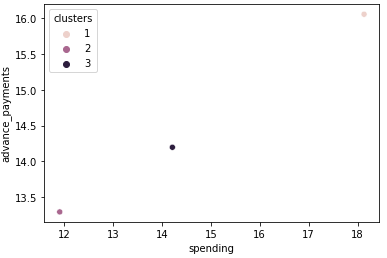
Apparently K – Means has come up with less customer count in the higher spending bracket.

Mean of rest of the columns is almost the same in both the tried approaches.

For Illustration purposes here is the analysis of the expected Usage (Spending) attribute against the remaining to suggest the deviations and to propose the strategy.

Also Count across three Segments appears to be the same and for illustration purpose, **70** are considered as count.

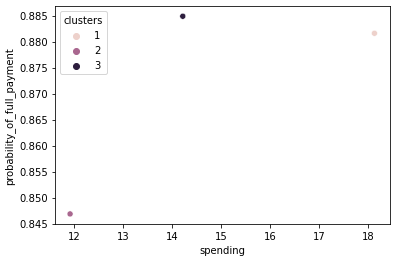
Following Plots Depict trend of attributes against ‘USAGE (Spending)’ for the segments, we will use the same for drawing insight as well as deriving the strategy



**INSIGHT**

**Spending** across three clusters shows linear relationship with **‘advance\_payments’**

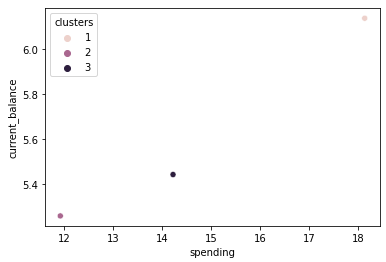
***Plot 1.5.1***



**INSIGHT**

**Spending** across three clusters show almost same probability of full payment.

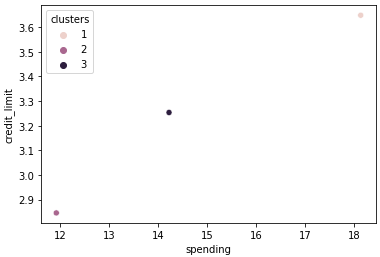
***Plot 1.5.2***



**INSIGHT**

**Spending** across three clusters shows linear relationship with **‘current\_balance’.**

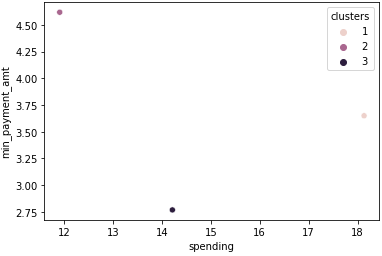
***Plot 1.5.3***



**INSIGHT**

**Spending** across three clusters shows linear relationship with **‘credit\_limit’.**

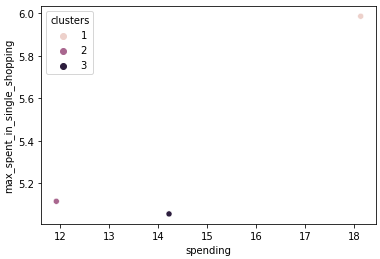
***Plot 1.5.4***



**INSIGHT**

**Spending** across three clusters shows non symmetrical behaviour across. **Minimum Payment Amount** is maximum for the segment with least average spending and it is minimum for average Spending cluster

***Plot 1.5.5***

***Plot 1.5.6***

**INSIGHT**

**Spending** across three clusters shows non symmetrical behaviour across. “Maximum spent in single shopping” is minimum for the segment with average spending segment.

**CONCLUSION and STRATEGY on Clustering**

* Both Agglomerative and K-Means resulted in to same number of observations divided across 3 groups of 70 observations.
* Approximate Average Spends in the three groups is 12K, 14K and 18K.
* Attributes advance\_payments, probability\_of\_full\_payment, current\_balance and credit\_limit form a linear progression against the spending in all 3 segments.
* Spending across three clusters shows non symmetrical behaviour in case of minimum\_payment\_amount. Minimum Payment Amount is maximum for the segment with least average spending and it is minimum for average Spending cluster.
* Spending across three clusters shows non symmetrical behaviour across. “Maximum spent in single shopping” is minimum for the segment with average spending segment.
* Strategy is that reducing minimum payment amount for Cluster-2 customers will help them spend more. We essentially need to revise the minimum payment amount for all three segments.
* Implementation of above point will also help customers from Cluster-3 (average spending segment) and this will improve “Maximum spends in single shopping for Average spending segment.

# Problem 2 CART-RF-ANN

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

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

### Describe

Supplied Data set has 10 columns, with 4 numerical values. Supplied data dictionary is reviewed against the data set and all Data types are found correct.

Quantity and relevance of data seem to be correct from Exercise stand point.

### Data Pre-processing

***Treat BAD Data***

Data types are in line with the supporting data dictionary; Preens of no Bad Data is noticed

***Treat Anomalies***

There are apparently no noticeable Anomalies, with the definition of Existing Business rules that are available

***Treat Missing Values***

There are no Missing values in the supplied Data Set

***Checking for Duplicates***

There are 139 Duplicate values in the supplied Data Set however it is not advisable to remove them since considering the business duplicates are possible

### Data Visualization

#### Univariate Analysis

Following reviews are done for 4 numeric variables to understand the trend

1. 5 Number Analysis
2. Histogram Review
3. BoxPlot
4. Empirical Rule
5. mean, median and mode

We will go through each of these variables to understand the inference of above reviews as ‘Insight’ for each of the analysis Point.

### Insurance ['Age']

* 5 number Summary Suggests that -

Distribution is Right skewed –

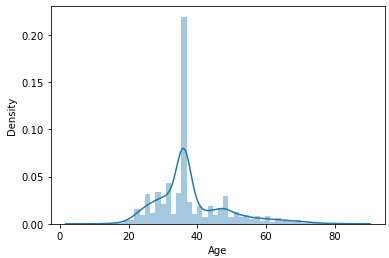
(Q1 - Min) which is 24.00 < Max - Q3 which is 42.00

Following points support the symmetry –

(Q2 - Min) which is 28.00 < (Max - Q2) which is 48.00

(Q2 - Q1) which is 4.00 < (Q3 - Q2 ) which is 6.00

* Histogram Review



***Plot 2.1.1***

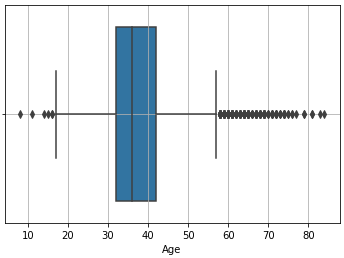
Histogram Review for Insurance ['Age'] Shows -

1) 2 Peaks

2) Right Skewness

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 2.1.2***

Insurance ['Age'] review of Boxplot suggests that -

1) There are many outliers

2) Distribution is Right Skewed due to longer Right Whisker

* Empirical Rule

Insurance ['Age'] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Right Skewed)

2) '2 standard deviation' from Mean is asymmetrical (Right Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Right Skewed)

* Insurance['Age'] review of Mean, Median and Mode suggests that –

In this case Mean > Median > Mode hence the distribution is Right Skewed

**Insights ('Age'):**

**Age of insured range from 8 year to 84 years**

**75% of the insured is less than 42 years**

**From above figure, we can say that the Age of insured is right skewed**

**Average age of insured is 38 years which is mirginally higher than the median value 36 years indicating that the ditribution is right tailed.**

**There are so many Outliers which shows data inconsistency.**

### Insurance ['Commision']

* 5 number Summary Suggests that -

Distribution is Right skewed -

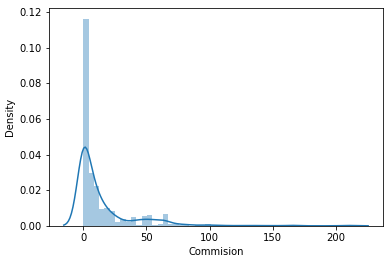
(Q1 - Min) which is 0.00 < Max - Q3 which is 192.98

Following points support the symmetry -

(Q2 - Min) which is 4.63 < (Max - Q2) which is 205.58

(Q2 - Q1) which is 4.63 < (Q3 - Q2) which is 12.61

* Histogram Review



***Plot 2.1.3***

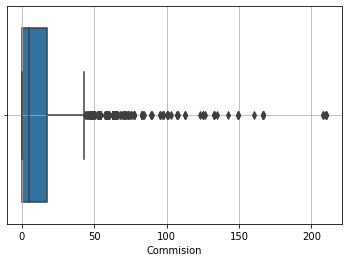
Histogram Review for Insurance ['Commision'] Shows -

1) 1 Peaks

2) Right Skewness

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 2.1.4***

Insurance ['Commision'] review of Boxplot suggests that -

1) There are many outliers

2) Distribution is Right Skewed due to longer Right Whisker

* Empirical Rule

Insurance ['Commision'] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Right Skewed)

2) '2 standard deviation' from Mean is asymmetrical (Right Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Right Skewed)

* Insurance ['Commision'] review of Mean, Median and Mode suggests that –

In this case Mean > Median > Mode hence the distribution is Right Skewed

**Insights (' Commision '):**

The commission received for tour insurance firm range from 0 to 210.21

For 75% of the insured, The commission received for tour insurance is less than 17.23

From above figure, we can say that the Age of insured is right skewed

Average commission received for tour insurance is 14.53 which is higher than the median value 4.63 indicating that the ditribution is right tailed.

There are so many Outliers which shows data inconsistency, most of the outliers are on higher side

### Insurance ['Duration']

* 5 number Summary Suggests that –

Distribution is Right skewed -

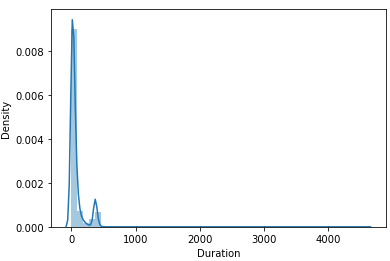
(Q1 - Min) which is 12.00 < Max - Q3 which is 4517.00

Following points support the symmetry -

(Q2 - Min) which is 27.50 < (Max - Q2) which is 4553.50

(Q2 - Q1) which is 15.50< (Q3 - Q2) which is 36.50

* Histogram Review



***Plot 2.1.5***

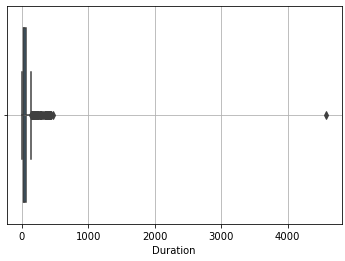
Histogram Review for Insurance ['Duration'] Shows -

1) 1 Peaks

2) Right Skewness

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 2.1.6***

Insurance ['Duration'] review of Boxplot suggests that –

1) There are many outliers

2) Distribution is Right Skewed due to longer Right Whisker

* Empirical Rule

Insurance ['Duration'] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Right Skewed)

2) '2 standard deviation' from Mean is asymmetrical (Right Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Right Skewed)

* Insurance['Duration'] review of Mean, Median and Mode suggests that –

In this case Mean > Median > Mode hence the distribution is Right Skewed

**Insights (**'Duration'**):**

Duration of the tour range from 0 days to 4580 Days

75% of the insured tours are for less than 63 days

From above figure, we can say that the Duration of the tour is right skewed

Average age of insured is 70 days which is higher than the median value 26.50 days indicating that the ditribution is right tailed.

There are so many Outliers which shows data inconsistency.

### Insurance [' Sales']

* 5 number Summary Suggests that –

Distribution is Right skewed -

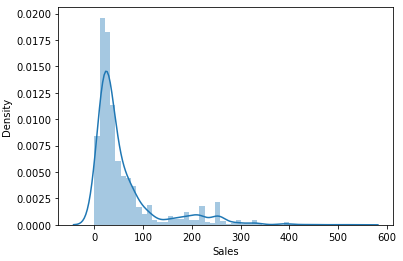
(Q1 - Min) which is 20.00 < Max - Q3 which is 470.00

Following points support the symmetry -

(Q2 - Min) which is 33.00 < (Max - Q2) which is 506.00

(Q2 - Q1) which is 13.00 < (Q3 - Q2) which is 36.00

* Histogram Review



***Plot 2.1.7***

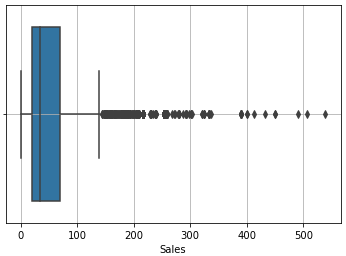
Histogram Review for Insurance ['Sales'] Shows -

1) 1 Peaks

2) Right Skewness

3) It doesn’t Show a perfect 'Bell Curve'

* BoxPlot Review



***Plot 2.1.8***

Insurance ['Sales'] review of Boxplot suggests that –

1) There are many outliers

2) Distribution is Right Skewed due to longer Right Whisker

* Empirical Rule

Insurance ['Sales'] review of Empirical Rule suggest that –

1) '1 standard deviation' from Mean is asymmetrical (Right Skewed)

2) '2 standard deviation' from Mean is asymmetrical (Right Skewed)

3) '3 standard deviation' from Mean is asymmetrical (Right Skewed)

* Insurance['Sales'] review of Mean, Median and Mode suggests that –

In this case Mean > Median > Mode hence the distribution is Right Skewed

**Insights (**'Sales'**):**

Amount of sales of tour insurance policies range from 0 to 539.00

75% of the policies get is less than 69 Amount of sales

From above figure, we can say that the Age of insured is right skewed

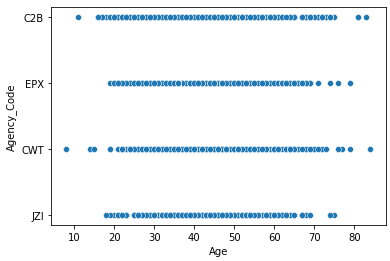
Average Amount of sales of tour insurance policies is 60.25 which is higher than the median value 33.00 indicating that the ditribution is right tailed.

There are so many Outliers which shows data inconsistency.

#### Bivariate Analysis

Please find Bivariate Analysis performed on Attributes to demonstrate the approach as well as illustrate typical insights that can be drawn.

### Insurance [Age Vs Agency\_Code]



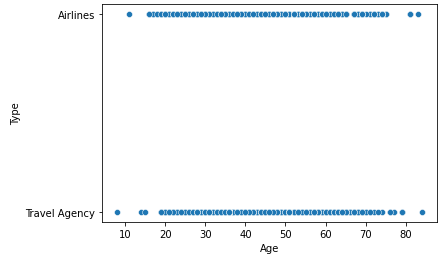
**INSIGHT**

Maximum age range is served by Agencies C2B and CWT

All four Agencies seem to be serving almost similar count of policy holders.

***Plot 2.1.9***

### Insurance [Age Vs Type]



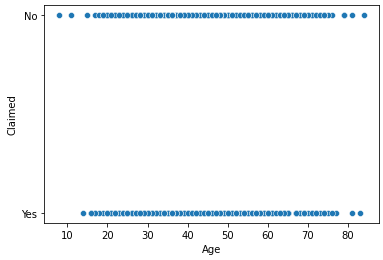
**INSIGHT**

Both Type of tour insurance firms serve the entire age range.

***Plot 2.1.10***

### 

### Insurance [Age Vs Claimed]

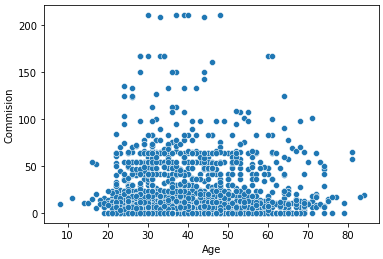


**INSIGHT**

There is similar count of observations in both claimed (Y/N) across the policyholder age.

***Plot 2.1.11***

### Insurance [Age Vs Commision]



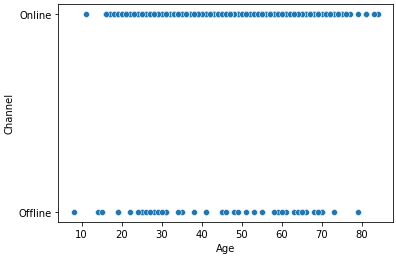
**INSIGHT**

Commission is similar for all the age insured however there are cases where Type of tour insurance firms is higher in handful of cases.

***Plot 2.1.12***

### 

### Insurance [Age Vs Channel]

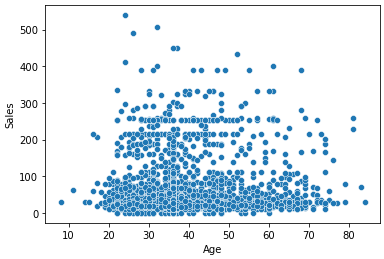


**INSIGHT**

There are more insurers registered online than offline however Insurers chose both Online and Offline channels for registration.

***Plot 2.1.13***

### Insurance [Age Vs Sales]

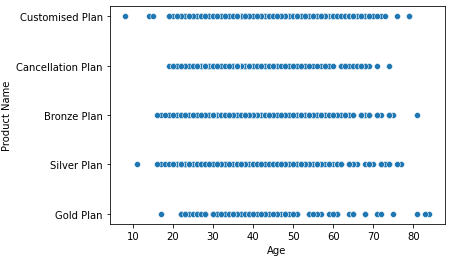


**INSIGHT**

Sales are similar for all the age insured however there are cases where Sales is higher in handful of cases.

***Plot 2.1.14***

### Insurance [Age Vs Product Name]

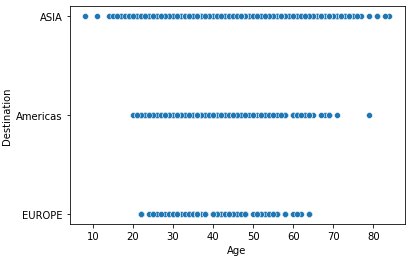


**INSIGHT**

There is similar amount of subscription for all the plans by the Insurers Age group.

***Plot 2.1.15***

### Insurance [Age Vs Destination]



**INSIGHT**

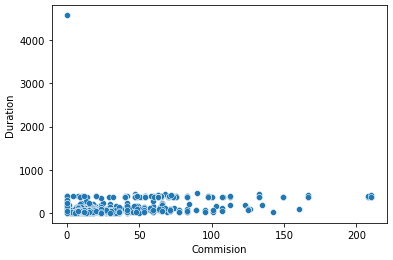
Destination opted differs, there are less insurers who are Elderly or Younger and travelling to Europe.

ASIA apparently has insurers from all the age groups

***Plot 2.1.16***

### 

### Insurance [Commision Vs Duration]

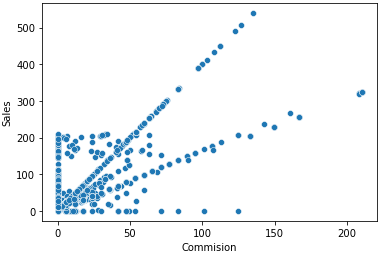


**INSIGHT**

Commission is similar for all the Durations however there are cases where commission is higher in handful of cases.

***Plot 2.1.17***

### Insurance [Commision Vs Sales]

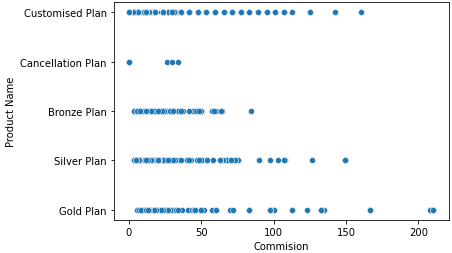


**INSIGHT**

Commision is lower for lower Sale however it increases exponentially as Sales go higher.

***Plot 2.1.18***

### Insurance [Commision Vs Product Name]



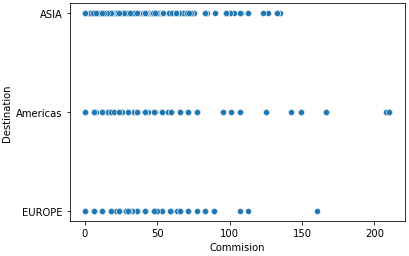
**INSIGHT**

Order of Commision / product is as follows

* Gold Plan
* Customised Plan
* Silver Plan
* Bronze Plan
* Cancellation plan

***Plot 2.1.19***

### Insurance [Commision Vs Destination]



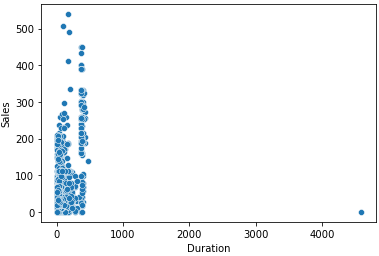
**INSIGHT**

Order of Commision / Destination is

* Americas
* ASIA
* Europe

***Plot 2.1.20***

### Insurance [Duration Vs Sales]



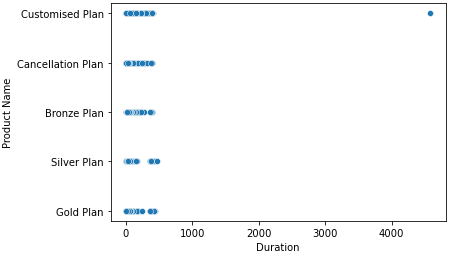
**INSIGHT**

There is higher Sale of a specific duration of Holidays.

However there are instances of higher sales for the some duration based on the destination

***Plot 2.1.21***

### Insurance [Duration Vs Product Name]



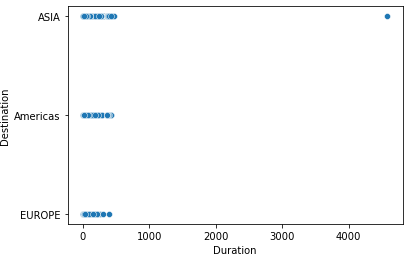
**INSIGHT**

Duration is similar across following products

* Gold Plan
* Customised Plan
* Silver Plan
* Bronze Plan
* Cancellation plan

***Plot 2.1.22***

### Insurance [Duration Vs Destination]



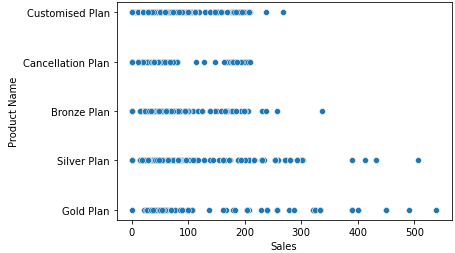
**INSIGHT**

Duration is similar across following Destinations

* ASIA
* Americas
* Europe

***Plot 2.1.23***

### Insurance [Sales Vs Product Name ]



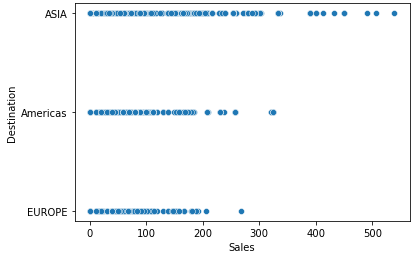
**INSIGHT**

Order of Sales / product is as follows

* Gold Plan
* Silver Plan
* Bronze Plan
* Customised Plan
* Cancellation plan

***Plot 2.1.24***

### Insurance [Sales Vs Destination]



**INSIGHT**

Sales is in order across following Destinations

* ASIA
* Americas
* Europe

***Plot 2.1.25***

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

Cart Modelling

* Prior to performing the Modelling; as part of data preparation; we have converted the object data type variables into categorical variables.
* Feature: Agency\_Code: ['C2B', 'CWT', 'EPX', 'JZI'] mapped to [0 2 1 3]
* Feature: Type: ['Airlines', 'Travel Agency'] mapped to [0 1]
* Feature: Claimed: ['No', 'Yes'] mapped to [0 1]
* Feature: Channel: ['Offline', 'Online'] mapped to [1 0]
* Feature: Product\_Name : ['Bronze Plan', 'Cancellation Plan', 'Customised Plan', 'Gold Plan', 'Silver Plan'] mapped to [2 1 0 4 3]
* Feature: Destination: ['ASIA', 'Americas', 'EUROPE'] mapped to [0 1 2]
* There are 69% of observations that have not Claimed and rest 31% seem to have made claims
* Observations are split into Training set of 70% and rest 30% as Test Set.
* Supported criteria is selected as “gini” for the Gini impurity function to measure the quality of Training Vs Test split
* To Create a Tree File , Names of each of the features is picked from training set and class\_names is taken as ['No', 'Yes']
* Tree File thus created is visualized using <http://webgraphviz.com/>
* Before Regularization of Tree ; variable importance is as follows –

Variable Imp

Duration 0.26

Sales 0.20

Agency\_Code 0.20

Age 0.19

Commision 0.10

Product\_Name 0.04

Destination 0.02

Channel 0.01

Type 0.00

***Table 2.2.1***

* Following Adding Tuning Parameters are added to generate a regularized Model basis the review of original model.
* **max\_depth** The maximum depth of the tree is set to **30** as this gives us sufficient level of purity.
* **min\_samples\_leaf** The minimum number of samples required to be at a leaf node is set to **60.**  A split point at any depth will only be considered if it leaves at least 60 training samples in each of the left and right branches. This will smoothen the model, especially in regression
* **min\_samples\_split: The minimum number of samples required to split an internal node is set to 250 which is around 12% of training data set.**
* After Regularization of Tree ; variable importance is as follows –

Variable Imp

Agency\_Code 0.63

Sales 0.26

Product\_Name 0.06

Duration 0.03

Age 0.02

Commision 0.00

Type 0.00

Channel 0.00

Destination 0.00

***Table 2.2.2***

* It is verified that the predicted test classes and the probabilities are in line
* Predicted classes for first 10 observations

0, 1, 1, 1, 0, 0, 0, 0, 0, 1

* Predicted Probabilities for first 10 observations

|  | **0** | **1** |
| --- | --- | --- |
| **0** | 0.89 | 0.11 |
| **1** | 0.44 | 0.56 |
| **2** | 0.44 | 0.56 |
| **3** | 0.30 | 0.70 |
| **4** | 0.94 | 0.06 |
| **5** | 0.68 | 0.32 |
| **6** | 0.88 | 0.12 |
| **7** | 0.56 | 0.44 |
| **8** | 0.68 | 0.32 |
| **9** | 0.44 | 0.56 |

***Table 2.2.3***

* We will Evaluate the performance of this CART model as part of response to Question 2.3

Ensemble Random Forest Classifier

* We have already split the dataset as part of CART modelling we will utilize the same going forward
* There are 100 number of trees opted in the forest
* Model is Built with 9 Features
* Train Accuracy of Training Data is 0.9947619047619047
* We will Evaluate the performance of this RF model as part of response to Question 2.3

Building Neural Network Model

* We have already split the dataset as part of CART modelling we will utilize the same going forward
* Scaled the training data
* Applied the transformation on the test data
* Multi-layer Perceptron classifier is applied to optimize the log-loss function using LBFGS or stochastic gradient descent.
* hidden\_layer\_sizes = 100, The ith element represents the number of neurons in the ith hidden layer.
* max\_iter=5000, Maximum number of iterations. The solver iterates until convergence (determined by ‘tol’) or this number of iterations. For stochastic solvers (‘sgd’), note that this determines the number of epochs (how many times each data point will be used), not the number of gradient steps.
* solver='sgd', The solver for weight optimization.‘sgd’ refers to stochastic gradient descent.
* verbose=True, to print progress messages to stdout.
* random\_state=1,
* tol=0.01, Tolerance for the optimization. When the loss or score is not improving by atleast this number, convergence is considered to be reached and training stops.

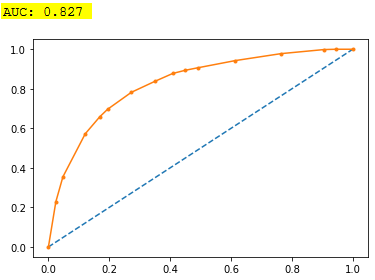
We will evaluate the performance of Neural Network model as part of response to Question 2.3

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

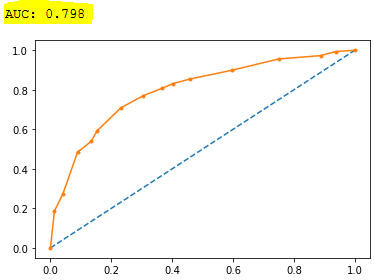
### CART Model Evaluation

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CART | | | | | | | | | |
|  | TRAIN | | | |  | TEST | | | |
|  | precision | recall | f1-score | support |  | precision | recall | f1-score | support |
| 0 | 0.83 | 0.88 | 0.85 | 1471 |  | 0.78 | 0.91 | 0.84 | 605 |
| 1 | 0.67 | 0.57 | 0.62 | 629 |  | 0.73 | 0.48 | 0.58 | 295 |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| accuracy |  |  | 0.79 | 2100 |  |  |  | 0.77 | 900 |
| macro avg | 0.75 | 0.73 | 0.73 | 2100 |  | 0.75 | 0.7 | 0.71 | 900 |
| weighted avg | 0.78 | 0.79 | 0.78 | 2100 |  | 0.76 | 0.77 | 0.76 | 900 |
| AUC | 0.827 |  |  |  |  | 0.798 |  |  |  |

***Table 2.3.1***



***Plot 2.3.1***



***Plot 2.3.2***

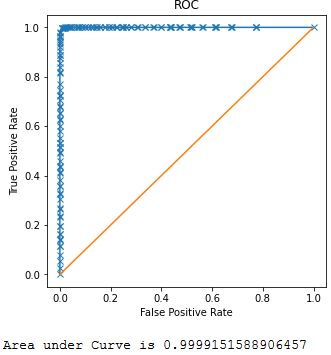
* Accuracy on the Training Data: 79%
* Accuracy on the Test Data: 77%
* AUC on the Training Data: 82.7%
* AUC on the Test: 79.8%

Accuracy, AUC, Precision and Recall for test data is almost in line with training data.  
This proves no over fitting or under fitting has happened, and overall the model is a good model for classification

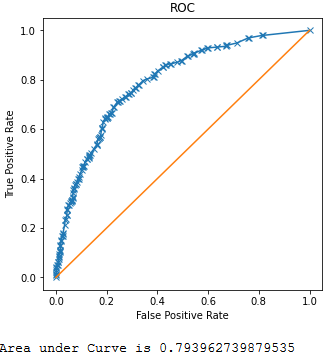
### Random Forest Model Evaluation

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| RF - Original | | | | | | | | | |
|  | TRAIN | | | |  | TEST | | | |
|  | precision | recall | f1-score | support |  | precision | recall | f1-score | support |
| 0 | 0.99 | 1 | 1 | 1471 |  | 0.77 | 0.89 | 0.83 | 605 |
| 1 | 1 | 0.99 | 0.99 | 629 |  | 0.67 | 0.45 | 0.54 | 295 |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| accuracy |  |  | 0.99 | 2100 |  |  |  | 0.75 | 900 |
| macro avg | 0.99 | 0.99 | 0.99 | 2100 |  | 0.72 | 0.67 | 0.68 | 900 |
| weighted avg | 0.99 | 0.99 | 0.99 | 2100 |  | 0.74 | 0.75 | 0.73 | 900 |
| AUC | 1 |  |  |  |  | 0.8 |  |  |  |

***Table 2.3.2***



***Plot 2.3.3***



***Plot 2.3.4***

Area under the curve on the training data is 100%, which indicates very high performance that all classes have been correctly classified. Whereas on the test data model performance is average with AUC 80%, which is very less compare to the performance of the training data.

If a person not having a claim, is incorrectly predicted to have a claim, in this situation, the cost and other impact is severe.

From the Random Forest model, looking at the Accuracy, Sensitivity, Septicity, Recall and AUC, we have 100% results on the training data, whereas on the Test data, performance is lesser, especially in predicting **Class 1**.

This is because over fitting has happened on the training data, and therefore the model is weak in generalizing and predicting any new data.

In this model, we have hard-coded the hyper parameter values. We can optimize/fine-tune the random forest model, by trying different values for the hyper parameters to see if the model performance is improving.

* Grid Search for finding out the optimal values for the hyper parameters

For a 10 Fold Cross Validation Grid Search returned following as best Parameter

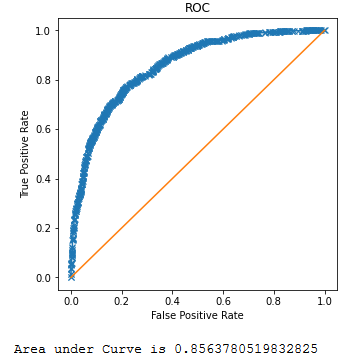
|  |  |  |
| --- | --- | --- |
| **Parameters** | **Option** | **Best Parameter** |
| max\_depth | [5,7,10] | 7 |
| max\_features | [4,6] | 4 |
| min\_samples\_leaf | [5,10] | 5 |
| min\_samples\_split | [50,100] | 50 |
| n\_estimators | [100,200,300] | 100 |

***Table 2.3.3***

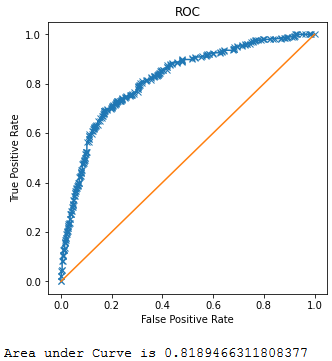
Modelling based on the Best Parameters once applied returned following output which apparently has removed the Over fitting, Training and Test Dataset seem to be closure

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| RF - Best Parameter | | | | | | | | | |
|  | TRAIN | | | |  | TEST | | | |
|  | precision | recall | f1-score | support |  | precision | recall | f1-score | support |
| 0 | 0.84 | 0.89 | 0.87 | 1471 |  | 0.79 | 0.91 | 0.84 | 605 |
| 1 | 0.71 | 0.61 | 0.66 | 629 |  | 0.72 | 0.49 | 0.58 | 295 |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| accuracy |  |  | 0.81 | 2100 |  |  |  | 0.77 | 900 |
| macro avg | 0.78 | 0.75 | 0.76 | 2100 |  | 0.75 | 0.7 | 0.71 | 900 |
| weighted avg | 0.8 | 0.81 | 0.8 | 2100 |  | 0.76 | 0.77 | 0.76 | 900 |
| AUC | 0.85638 |  |  |  |  | 0.818947 |  |  |  |

***Table 2.3.4***



***Plot 2.3.5***



***Table 2.3.6***

* Accuracy on the Training Data: 81%
* Accuracy on the Test Data: 77%
* AUC on the Training Data: 85%
* AUC on the Test: 82%

Accuracy, AUC, Precision and Recall for test data is almost in line with training data.  
This proves no over fitting or under fitting has happened, and overall the model is a good model for classification

### Neural Network Model Evaluation

In this model, we have hard-coded the hyper parameter values. We can optimize/fine-tune the random forest model, by trying different values for the hyper parameters to see if the model performance is improving.

Supplying following set of parameters -

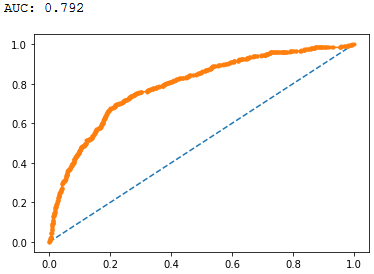
|  |  |
| --- | --- |
| **Parameters** | **Best Parameter** |
| hidden\_layer\_sizes | 100 |
| max\_iter | 5000 |
| solver | sgd |
| tol | 0.01 |

***Table 2.3.5***

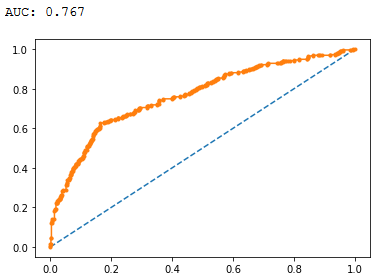
Following classification **report** is returned based on above parameters

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Neural Network Model | | | | | | | | | | | | | | | |
|  | TRAIN | | | | |  | | TEST | | | | | | | |
|  | precision | recall | f1-score | | support | |  | | precision | | recall | | f1-score | | support | |
| 0 | 0.76 | 0.95 | 0.85 | 1471 | |  | | 0.73 | | 0.96 | | 0.83 | | 605 | |
| 1 | 0.72 | 0.31 | 0.44 | 629 | |  | | 0.77 | | 0.27 | | 0.4 | | 295 | |
|  |  |  |  |  | |  | |  | |  | |  | |  | |
|  |  |  |  |  | |  | |  | |  | |  | |  | |
| accuracy |  |  | 0.76 | 2100 | |  | |  | |  | | 0.73 | | 900 | |
| macro avg | 0.74 | 0.63 | 0.64 | 2100 | |  | | 0.75 | | 0.61 | | 0.61 | | 900 | |
| weighted avg | 0.75 | 0.76 | 0.72 | 2100 | |  | | 0.74 | | 0.73 | | 0.69 | | 900 | |
| AUC | 0.792 |  |  |  | |  | | 0.767 | |  | |  | |  | |

***Table 2.3.6***



***Plot 2.3.7***



***Plot 2.3.8***

Above classification\_report show relatively lower accuracy values as well as AUC values, we will try with identification of best parameters using Grid Search for further analysis.

* Grid Search for finding out the optimal values for the hyper parameters

For a 10 Fold Cross Validation Grid Search returned following as best Parameter

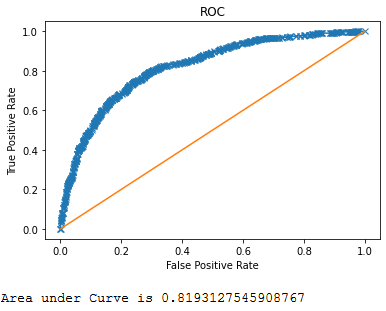
|  |  |  |
| --- | --- | --- |
| **Parameters** | **Option** | **Best Parameter** |
| hidden\_layer\_sizes | [100,200,300,500] | 300 |
| max\_iter | [5000,2500,7000,6000] | 5000 |
| solver | ['sgd','adam'] | adam |
| tol | [0.01] | 0.01 |

***Table 2.3.7***

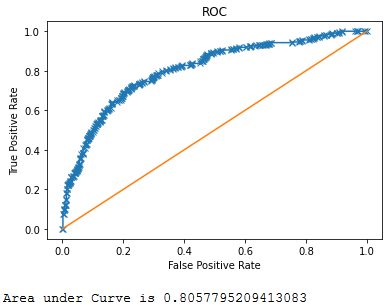
## Applying above best parameters returned following improved classification report

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Neural Network Model - Revised | | | | | | | | | |
|  | TRAIN | | | |  | TEST | | | |
|  | precision | recall | f1-score | support |  | precision | recall | f1-score | support |
| 0 | 0.81 | 0.9 | 0.85 | 1471 |  | 0.77 | 0.93 | 0.84 | 605 |
| 1 | 0.68 | 0.5 | 0.58 | 629 |  | 0.74 | 0.42 | 0.54 | 295 |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| accuracy |  |  | 0.78 | 2100 |  |  |  | 0.76 | 900 |
| macro avg | 0.75 | 0.7 | 0.71 | 2100 |  | 0.75 | 0.67 | 0.69 | 900 |
| weighted avg | 0.77 | 0.78 | 0.77 | 2100 |  | 0.76 | 0.76 | 0.74 | 900 |
| AUC | 0.81931 |  |  |  |  | 0.80578 |  |  |  |

***Table 2.3.8***



***Plot 2.3.9***



***Plot 2.3.10***

* Accuracy on the Training Data: 78%
* Accuracy on the Test Data: 76%
* AUC on the Training Data: 82%
* AUC on the Test: 80.5%

Accuracy, AUC, Precision and Recall for test data is almost in line with training data.  
This proves no over fitting or under fitting has happened, and overall the model is a good model for classification

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

Please refer to the answer 2.3 that contains data pointers to support following comparison amongst the Models i.e. CART, RF and NN

1. Random Forest creates multiple CART trees based on "bootstrapped" samples of data and then combines the predictions. Usually, the combination is an average of all the predictions from all CART models. A bootstrap sample is a random sample conducted with replacement.
2. Random Forest has better predictive power and accuracy than a single CART model (because of random forest exhibit lower variance).
3. Unlike the CART model, Random Forest's rules are not easily interpretable.
4. Random Forest inherits properties of CART-like variable selection, missing values and outlier handling, nonlinear relationships, and variable interaction detection.
5. Random Forest performed accurately on training data set as compared to CART model. But it was over fitted However the drop in prediction accuracy after validation is more in random forest than in CART model.
6. Also AUC is more In case of RF for both Training and Test Data Set comparing to CART
7. When we compare RF matrix Against Neural Network Model, Accuracy apparently is same in both.
8. Also AUC is more In case of RF for both Training and Test Data Set comparing to Neural Network Model
9. Random Forest is less computationally expensive and does not require a GPU to finish training. A random forest can give a different interpretation of a decision tree but with better performance. Neural Networks will require much more data than an everyday person might have on hand to actually be effective. The neural network will simply decimate the interpretability of your features to the point where it becomes meaningless for the sake of performance. While that may sound reasonable to some, it is dependent on each project.
10. Here the goal is to create a prediction model with due care for the variables, hence neural network is less preferred than RF

Considering the Data from Answer 2.3 and Point 5,6,7 and 8 order of preference is

* CART - Stable but Least preferred
* Neural Network – Best results but doesn’t give importance to variable so not a best choice.
* Random Forest – Best Choice , for the same accuracy over ANN

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

There are 3 Parameters that relate to Finances around selling of Insurance Policies-

* Amount of sales of tour insurance policies (Sales)
* The commission received for tour insurance firm (Commission)
* Claim Status (Claimed)
* ‘Sales’ is receivable amount.
* ‘Commission’ is pay out to the Insurance firm
* ‘Claimed’ if made is a pay out to the insurer.
* ‘Sales’ and ‘Commission’ progresses in the same direction so they will not make any impact on the profitability if commission is reduced so business should continue to offer Commission to improve on profitability.
* ‘Claimed’ is the parameter that drives the profitability of the business.
* In the supplied data set 31% of the cases make claims.
* We cannot control the claim as they are regulated.
* We have to devise strategy such that there is balance between Sale and Claim
* Prediction of ‘Claimed’ should be used to revise the Premium / Commission based on following Variables

|  |  |
| --- | --- |
| **Variable** | **Impacted Payable / Receivable** |
| Code of tour firm (Agency\_Code). | Commission |
| Type of tour insurance firms (Type) | Commission |
| Distribution channel of tour insurance agencies (Channel) | Commission |
| Name of the tour insurance products (Product) | Commission |
| Duration of the tour (Duration) | Premium |
| Destination of the tour (Destination) | Premium |
| Age of insured (Age) | Premium |

***Table 2.3.9***

# Appendices

## List of Plots

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Plot Details** | | | | |
| Plot 1.1.1 | spending Variable Histogram |  | Plot 2.1.5 | Insurance ['Duration'] Variable Histogram |
| Plot 1.1.2 | spending Variable BoxPlot |  | Plot 2.1.6 | Insurance ['Duration'] Variable BoxPlot |
| Plot 1.1.3 | advance\_payments Variable Histogram |  | Plot 2.1.7 | Insurance ['Sales'] Variable Histogram |
| Plot 1.1.4 | advance\_payments Variable BoxPlot |  | Plot 2.1.8 | Insurance ['Sales'] Variable BoxPlot |
| Plot 1.1.5 | probability\_of\_full\_payment Variable Histogram |  | Plot 2.1.9 | Age Vs Agency\_Code |
| Plot 1.1.6 | probability\_of\_full\_payment Variable BoxPlot |  | Plot 2.1.10 | Age Vs Type |
| Plot 1.1.7 | current\_balance Variable Histogram |  | Plot 2.1.11 | Age Vs Claimed |
| Plot 1.1.8 | current\_balance Variable BoxPlot |  | Plot 2.1.12 | Age Vs Commision |
| Plot 1.1.9 | credit\_limit Variable Histogram |  | Plot 2.1.13 | Age Vs Channel |
| Plot 1.1.10 | credit\_limit Variable BoxPlot |  | Plot 2.1.14 | Age Vs Sales |
| Plot 1.1.11 | max\_spent\_in\_single\_shopping Variable Histogram |  | Plot 2.1.15 | Age Vs Product Name |
| Plot 1.1.12 | max\_spent\_in\_single\_shopping Variable BoxPlot |  | Plot 2.1.16 | Age Vs Destination |
| Plot 1.1.13 | spending Vs advance\_payments |  | Plot 2.1.17 | Commision Vs Duration |
| Plot 1.1.14 | spending Vs probability\_of\_full\_payment |  | Plot 2.1.18 | Commision Vs Sales |
| Plot 1.1.15 | spending Vs current\_balance |  | Plot 2.1.19 | Commision Vs Product Name |
| Plot 1.1.16 | spending Vs credit\_limit |  | Plot 2.1.20 | Commision Vs Destination |
| Plot 1.1.17 | spending Vs min\_payment\_amt |  | Plot 2.1.21 | Duration Vs Sales |
| Plot 1.1.18 | spending Vs max\_spent\_in\_single\_shopping |  | Plot 2.1.22 | Duration Vs Product Name |
| Plot 1.3.1 | Dendrogram |  | Plot 2.1.23 | Duration Vs Destination |
| Plot 1.4.1 | KMeans elbow curve |  | Plot 2.1.24 | Sales Vs Product Name |
| Plot 1.5.1 | spending Vs advance\_payments (Cluster ) |  | Plot 2.1.25 | Sales Vs Destination |
| Plot 1.5.2 | spending Vs probability\_of\_full\_payment (Cluster ) |  | Plot 2.3.1 | AUC - Training CART |
| Plot 1.5.3 | spending Vs current\_balance (Cluster) |  | Plot 2.3.2 | AUC - Test CART |
| Plot 1.5.4 | spending Vs credit\_limit (Cluster) |  | Plot 2.3.3 | AUC - Training RF |
| Plot 1.5.5 | spending Vs min\_payment\_amt (Cluster) |  | Plot 2.3.4 | AUC - Test RF |
| Plot 1.5.6 | spending Vs max\_spent\_in\_single\_shopping (Cluster) |  | Plot 2.3.5 | AUC - Training RF- Best Parameter |
| Plot 2.1.1 | Insurance ['Age'] Variable Histogram |  | Plot 2.3.6 | AUC - Test RF- Best Parameter |
| Plot 2.1.2 | Insurance ['Age'] Variable BoxPlot |  | Plot 2.3.7 | AUC - Training NN |
| Plot 2.1.3 | Insurance ['Commision'] Variable Histogram |  | Plot 2.3.8 | AUC - Test NN |
| Plot 2.1.4 | Insurance ['Commision'] Variable BoxPlot |  | Plot 2.3.9 | AUC - Training NN- Best Parameter |
|  |  |  | Plot 2.3.10 | AUC - Test NN- Best Parameter |

## List of Tables

|  |  |
| --- | --- |
| **Table Details** | |
| Table 1.2.1 | Unit Of Measurement (Scaling) |
| Table 1.3.1 | hierarchical clustering - Frequency |
| Table 1.4.1 | K-Means Inertia |
| Table 1.4.2 | K-Means silhouette\_score |
| Table 1.5.1 | Hierarchical Vs K- Means clustering |
| Table 2.2.1 | Variable Importance CART |
| Table 2.2.2 | Regularized Variable Importance CART |
| Table 2.2.3 | Predicted Probabilities CART |
| Table 2.3.1 | Confusion Matrix - CART (Training Vs Test) |
| Table 2.3.2 | Confusion Matrix - RF (Training Vs Test) |
| Table 2.3.3 | Best Parameter Table - RF |
| Table 2.3.4 | Confusion Matrix - RF Best Parameter (Training Vs Test) |
| Table 2.3.5 | Param Table - NN |
| Table 2.3.6 | Confusion Matrix -NN (Training Vs Test) |
| Table 2.3.7 | Best Parameter Table - NN |
| Table 2.3.8 | Confusion Matrix - NN Best Parameter (Training Vs Test) |
| Table 2.3.9 | Strategy Impacted Variables |

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*End of Document \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*