Harshini Nimmagadda

Project - Data Mining

Contents

[1. Bank Marketing Analysis 5](#_Toc103951738)

[**Executive Summary** 5](#_Toc103951739)

[**Introduction** 5](#_Toc103951740)

[**Data Description** 5](#_Toc103951741)

[**Sample of the Dataset** 5](#_Toc103951742)

[**Table 1. Dataset Sample** 5](#_Toc103951743)

[**1.1.** **Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).** 5](#_Toc103951744)

[**Exploratory Data Analysis** 5](#_Toc103951745)

[**Check for missing values in the dataset** 6](#_Toc103951746)

[**Summary of the dataset** 6](#_Toc103951747)

[**Univariate Analysis** 6](#_Toc103951748)

[**Multivariate Analysis** 10](#_Toc103951749)

[**Heat Map:** 10](#_Toc103951750)

[**Pair Plot:** 11](#_Toc103951751)

[**Checking the Outliers** 11](#_Toc103951752)

[**1.2.** **Do you think scaling is necessary for clustering in this case? Justify** 12](#_Toc103951753)

[**1.3.** **Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them** 13](#_Toc103951755)

[**Table 3. Dataset with Clusters** 14](#_Toc103951756)

[**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.** 14](#_Toc103951757)

[**1.5.** **Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters.** 16](#_Toc103951758)

[**2.** **Insurance Analysis** 17](#_Toc103951759)

[**Executive Summary** 17](#_Toc103951760)

[**Data Description** 17](#_Toc103951761)

[**2.1.** **Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).** 18](#_Toc103951763)

[**Exploratory Data Analysis** 18](#_Toc103951764)

[**Check for missing values in the dataset** 18](#_Toc103951765)

[**Summary of the dataset** 18](#_Toc103951766)

[**Univariate Analysis** 19](#_Toc103951767)

[**Bivariate Analysis** 23](#_Toc103951768)

[**Multivariate Analysis** 25](#_Toc103951769)

[**Pair Plot:** 28](#_Toc103951770)

[**Heat Map:** 29](#_Toc103951771)

[**Checking the Outliers** 29](#_Toc103951772)

[**2.2.** **Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network** 30](#_Toc103951773)

[**Converting Object datatype to Int:** 30](#_Toc103951774)

[**Splitting Dataset in Train and Test Data (70:30)** 30](#_Toc103951775)

[**CART Model** 31](#_Toc103951776)

[**Regularising the Decision Tree** 31](#_Toc103951777)

[**Best Parameters** 31](#_Toc103951778)

[**Variable Importance** 32](#_Toc103951779)

[**Random Forest** 32](#_Toc103951780)

[**Grid Search** 32](#_Toc103951781)

[**Best Parameters** 33](#_Toc103951782)

[**Variable Importance** 33](#_Toc103951783)

[**Artificial Neural Network (ANN)** 33](#_Toc103951784)

[**Grid Search** 34](#_Toc103951785)

[**Best Parameters** 34](#_Toc103951786)

[**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.** 34](#_Toc103951787)

[**CART Model** 34](#_Toc103951788)

[**Classification Report** 34](#_Toc103951789)

[**Confusion Matrix** 35](#_Toc103951790)

[**ROC\_AUC Score and ROC Curve** 35](#_Toc103951791)

[**Cart Conclusion** 35](#_Toc103951792)

[**Random Forest** 36](#_Toc103951793)

[**Classification Report** 36](#_Toc103951794)

[**Confusion Matrix** 36](#_Toc103951795)

[**ROC\_AUC Score and ROC Curve** 37](#_Toc103951796)

[**Random Forest Conclusion** 37](#_Toc103951797)

[**Artificial Neural Network (ANN)** 37](#_Toc103951798)

[**Classification Report** 37](#_Toc103951799)

[**ROC\_AUC Score and ROC Curve** 38](#_Toc103951800)

[**Confusion Matrix** 38](#_Toc103951801)

[**Artificial Neural Network (ANN) Conclusion** 39](#_Toc103951802)

[**2.4.** **Final Model: Compare all the models and write an inference which model is best/optimized.** 39](#_Toc103951803)

[**2.5.** **Inference: Based on the whole Analysis, what are the business insights and recommendations** 39](#_Toc103951804)

**List of Tables**

|  |  |
| --- | --- |
| **Table Name** | **Page No** |
| Dataset Sample | **5** |
| Scaled Dataset | **12** |
| Dataset with Clusters | **14** |
| Dataset with kmeans | **16** |
| Dataset with sil\_width | **16** |
| Dataset with Freq | **15** |
| Insurance Dataset | **17** |
| Insurance Dataset – Conversion | **30** |
| Model Comparison | **39** |

**List of Figures**

|  |  |
| --- | --- |
| **Figure Name** | **Page No** |
| Univariate - Spending | 6 |
| Univariate – advance\_payments | 7 |
| Univariate –probability\_of\_full\_payment | 7 |
| Univariate – current\_balance | 8 |
| Univariate – credit\_limit | 8 |
| Univariate – min\_payment\_amt | 9 |
| Univariate - max\_spent\_in\_single\_shopping . | 9 |
| Multivariate – Heat Map | 10 |
| Multivariate – Pair Plot | 11 |
| Boxplot – Outlier checking | 12 |
| Clustering | 13 |
| Truncated Clustering | 13 |
| Max Cluster | 13 |
| Clustering- Sets | 14 |
| Elbow Curve | 15 |
| Silhouette score. | 15 |
| Univariate – Age | 19 |
| Univariate – Commission | 19 |
| Univariate – Duration | 20 |
| Univariate – Sales | 20 |
| Univariate – Agency code | 21 |
| Univariate – Type | 21 |
| Univariate – Claimed | 21 |
| Univariate – Channel | 22 |
| Univariate – Product Name | 22 |
| Univariate – Destination | 22 |
| Agency code – Sales | 23 |
| Type – Sales | 23 |
| Channel – Sales | 23 |
| Product Name – Sales | 24 |
| Destination – Sales | 24 |
| Agency code – Sales – Claimed | 25 |
| Boxplot- Agency code – Sales – Claimed | 25 |
| Type– Sales – Claimed | 25 |
| Boxplot-Type– Sales – Claimed | 26 |
| Channel– Sales – Claimed | 26 |
| Boxplot - Channel– Sales – Claimed | 26 |
| Product Name– Sales – Claimed | 27 |
| Boxplot -Product Name– Sales – Claimed | 27 |
| Boxplot -Destination– Sales – Claimed | 28 |
| Pair plot | 28 |
| Heat Map | 29 |
| Boxplot – Outlier checking | 29 |
| Train Test Data | 30 |
| Decision Tree | 31 |
| Regularising Decision Tree | 32 |
| Scaled Trained Data | 33 |
| Scaled Test Data | 33 |
| Classification Report - Train | 34 |
| Classification Report - Test | 34 |
| Confusion Matrix - Train | 35 |
| Confusion Matrix - Test | 35 |
| ROC\_AUC Score and ROC Curve - Train | **35** |
| ROC\_AUC Score and ROC Curve - Test | 35 |
| Classification Report – Train | 36 |
| Classification Report – Test | 36 |
| Confusion Matrix – Train | 36 |
| Confusion Matrix – Test | 36 |
| ROC\_AUC Score and ROC Curve – Train | 37 |
| ROC\_AUC Score and ROC Curve – Test | 37 |
| Classification Report– Test | 37 |
| Classification Report– Train | 38 |
| Confusion Matrix– Train | 38 |
| Confusion Matrix– Test | 39 |

# **Bank Marketing Analysis**

# **Executive Summary**

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.

# **Introduction**

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Cluster the data to identify the segments based on credit card usage. The dataset consists of 210 customers and 7 variables.

# **Data Description**

1. **spending**: Amount spent by the customer per month (in 1000s).
2. **advance\_payments**: Amount paid by the customer in advance by cash (in 100s).
3. **probability\_of\_full\_payment**: Probability of payment done in full by the customer to the bank.
4. **current\_balance**: Balance amount left in the account to make purchases (in 1000s).
5. **credit\_limit**: Limit of the amount in credit card (10000s)
6. **min\_payment\_amt**: minimum paid by the customer while making payments for purchases made monthly (in 100s).
7. **max\_spent\_in\_single\_shopping**: Maximum amount spent in one purchase (in 1000s).

# **Sample of the Dataset**

**Table

Description automatically generated**

# **Table 1. Dataset Sample**

Dataset has 210 customers with 7 different variables.

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

# **Exploratory Data Analysis**

**Table

Description automatically generated with medium confidence**

There are total 210 rows and 7 columns in the dataset. All the columns are of Float datatype.

# **Check for missing values in the dataset**

Text

Description automatically generated

From the above results we can see that there is no missing value present in the dataset.

# **Summary of the dataset**

Table

Description automatically generated

* The mean & median values almost equal for all the variables.
* There is no large difference between 75% and the Max value.
* By looking at the above, we can say that there are no extreme values in the dataset.

# **Univariate Analysis**

Chart, histogram

Description automatically generated

**Fig 1. Univariate - Spending**

* The 25th quantile of spending is 12.27
* The median or 50th quantile of spending is 14.355
* The 75th quantile of spending is 17.305
* The Inter quantile range (IQR) of spending is 5.035
* Lower limit of Duration is 4.717499999999999
* Upper limit of Duration is 24.8575
* The distribution of spending is right skewed 0.399889191717758

Chart, histogram

Description automatically generated

**Fig 2. Univariate – advance\_payments**

* The 25th quantile of advance\_payments is 13.45
* The median or 50th quantile of advance\_payments is 14.32
* The 75th quantile of advance\_payments is 15.715
* The Inter quantile range (IQR) of advance\_payments is 2.2650000000000006
* Lower limit of Duration is 10.052499999999998
* Upper limit of Duration is 19.1125
* The distribution of advance\_payments is right skewed 0.3865727731912213

Chart, histogram

Description automatically generated

**Fig 3. Univariate – probability\_of\_full\_payment**

* The 25th quantile of probability\_of\_full\_payment is 0.8569
* The median or 50th quantile of probability\_of\_full\_payment is 0.8734500000000
* The 75th quantile of probability\_of\_full\_payment is 0.887775
* The Inter quantile range (IQR) of probability\_of\_full\_payment is 0.03087499996
* Lower limit of Duration is 0.8105875
* Upper limit of Duration is 0.9340875
* The distribution of probability\_of\_full\_payment is left skewed -0.537953728398

Chart, histogram

Description automatically generated

**Fig 4. Univariate – current\_balance**

* The 25th quantile of current\_balance is 5.26225
* The median or 50th quantile of current\_balance is 5.5235
* The 75th quantile of current\_balance is 5.97975
* The Inter quantile range (IQR) of current\_balance is 0.7175000000000002
* Lower limit of Duration is 4.186
* Upper limit of Duration is 7.056000000000001
* The distribution of current\_balance is right skewed 0.52548156013189

Chart, histogram

Description automatically generated

**Fig 5. Univariate – credit\_limit**

* The 25th quantile of credit\_limit is 2.944
* The median or 50th quantile of credit\_limit is 3.237
* The 75th quantile of credit\_limit is 5.97975
* The Inter quantile range (IQR) of credit\_limit is 0.61775
* Lower limit of Duration is 2.017375
* Upper limit of Duration is 4.488375
* The distribution of credit\_limit is right skewed 0.13437824513162136

Chart, histogram

Description automatically generated

**Fig 6. Univariate – min\_payment\_amt**

* The 25th quantile of min\_payment\_amt is 2.5614999999999997
* The median or 50th quantile of min\_payment\_amt is 3.599
* The 75th quantile of min\_payment\_amt is 4.76875
* The Inter quantile range (IQR) of min\_payment\_amt is 2.20725
* Lower limit of Duration is -0.7493750000000006
* Upper limit of Duration is 8.079625
* The distribution of min\_payment\_amt is right skewed 0.40166734329025183

Chart, histogram

Description automatically generated

**Fig 7. Univariate – max\_spent\_in\_single\_shopping**

* The 25th quantile of max\_spent\_in\_single\_shopping is 5.045
* The median or 50th quantile of max\_spent\_in\_single\_shopping is 5.2230000000000
* The 75th quantile of max\_spent\_in\_single\_shopping is 5.877000000000001
* The Inter quantile range (IQR) of max\_spent\_in\_single\_shopping is 0.83200000000
* Lower limit of Duration is 3.796999999999999
* Upper limit of Duration is 7.125000000000002
* The distribution of max\_spent\_in\_single\_shopping is right skewed 0.561897374954

**Inference:** After plotting the Boxplots for all the variables we can conclude that a few outliers are present in the variable namely, **min\_payment\_amt** which means that there are only a few customers whose minimum payment amount falls on the higher side on an average.

We can conclude from the above graphs that most of the customers in our data have a higher spending capacity, high current balance in their accounts and these customers spent a higher amount during a single shopping event. Majority of the customers have a higher probability to make full payment to the bank.

# **Multivariate Analysis**

# **Heat Map:**

Graphical user interface

Description automatically generated

**Fig 8. Multivariate – Heat Map**

**Inference**: As per the Heat Map, we can conclude that the following variables are highly correlated:

* Spending and advance\_payments, spending and current\_balance, spending and credit\_limit
* Advance payment and current\_balance, advance payment, and credit limit
* Current balance and max spent in single shopping

By this we can conclude that the customers who are spending very high have a higher current balance and high credit limit. Advance payments and maximum expenditure done in single shopping are done by majority of those customers who have high current balance in their bank accounts.

Probability of full payments are higher for those customers who have a higher credit limit.

Minimum payment amount is not correlated to any of the variables; hence, it is not affected by any changes in the current balance or credit limit of the customers.

# **Pair Plot:**

Chart, scatter chart

Description automatically generated

**Fig 9. Multivariate – Pair Plot**

With the help of the above pair plot we can understand the Univariate and Bivariate trends for all the variables in the dataset.

# **Checking the Outliers**

Chart, box and whisker chart

Description automatically generated

**Fig 10. Boxplot – Outlier checking**

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

Yes, scaling is necessary to do Clustering. Scaling is used to eliminate the redundant data and ensures the good quality clusters are generated which can improve the efficiency of clustering algorithms.

For the data given to us, scaling is required as all the variables are expressed in different units such as spending in 1000’s, advance payments in 100’s and credit limit in 10000’s, whereas probability is expressed as fraction or decimal values. Since the other values expressed in higher units will outweigh probabilities and can give varied results hence it is important to Scale the data using Standard Scaler and therefore normalise the values where the means will be 0 and standard deviation 1.

Graphical user interface

Description automatically generated

# **Table 2. Scaled Dataset**

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

After performing the hierarchical clustering to scaled data and following result is obtained. Wardlink method is used to obtain the clusters – Green & Red through dendrogram. We find the maximum number of customers falls under the red cluster.

Chart

Description automatically generated

**Fig 10. Clustering**

We have used truncate function with the value P = 10 to get the clear output of dendrogram.

Chart, histogram

Description automatically generated

**Fig 11. Truncated Clustering**

From the above truncated output, if we try to draw a line between 15 and 20 then we can see 3 vertical lines falls under. By using Max clust and distance method we can see that 3 clusters are good enough.

Text

Description automatically generated

**Fig 12. Max Cluster**

**Table

Description automatically generated**

# **Table 3. Dataset with Clusters**

A picture containing chart

Description automatically generated

**Fig 13. Clustering- Sets**

Cluster 1 having 70 rows, cluster 2 having 67 rows and cluster 3 having 73 rows.

Looking at the clusters and frequencies: -

* Cluster 1 has maximum power of spending and does maximum shopping in single visit as well as pays the amount in advance. The customer from has the highest probability of paying the full amount to the bank. As the current balance as well as credit limit is high, due to it increases the power of purchase. People in cluster 1 which is 70 out of 210 have more spending power as compared to other clusters
* Overall, the three clusters are divided into high/medium/low spending based on maximum spent in single shopping

# **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.**

We are fitting the scaled data into K Means model. We are able to see that the cluster mapping for the variables.

* k\_means.inertia\_ for Clusters =1 - 1469.9999999999998.
* k\_means.inertia\_ for Clusters =2 - 659.1474009548498.
* k\_means.inertia\_ for Clusters =3 - 430.298481751223.
* k\_means.inertia\_ for Clusters =4 - 371.221763926848.
* k\_means.inertia\_ for Clusters =5 - 326.8846407681858.

From Cluster 1 to 2 we have a significant drop close to 900 points.

From Cluster 2 to 3 we have a good drop close to 240 points.

From Cluster 3 to 4 it isn’t significant only 50 points drop.

From Cluster 4 to 5 it isn’t significant only 50 points drop.

By looking at the drop we can say 3 is optimal for us.

Chart, line chart

Description automatically generated

**Fig 13. Elbow Curve**

As per the above plot i.e. within sum of squares (wss) method we can conclude that the optimal number of clusters to be taken for k-means clustering is 3 since as per the elbow method it can be easily seen in the curve that after 3 the curve gets flat.

Graphical user interface, text, application, email

Description automatically generated

**Fig 14. Silhouette score.**

Average silhouette scores the highest Average score is corresponding to k=3. Hence, as per both the methods i.e. within sum of squares and silhouette method we can conclude that the optimal number of k or clusters that needs to be taken for k-means clustering is 3.

**Table

Description automatically generatedTable 4. Dataset with kmeans**

Table

Description automatically generated with low confidence

**Table 5. Dataset with sil\_width**

The average silhouettes score is 0.400 and minimum silhouette score is 0.002. The silhouette score ranges from -1 to +1 and higher the silhouette score better the clustering.

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

**Table

Description automatically generated with medium confidence**

**Table 6. Dataset with Freq**

Looking at the above summary, we can conclude:

**K-means Cluster 0**: This segment has higher spending per month, high current balance and credit limit. This is the **Prosperous or Upper class** with majorly higher income. This segment can be targeted using various offers such as cards with rewards and loyalty points for every spent.

**K-means Cluster 1**: This segment has the lowest spending per month, lowest current balance and credit limit. This is the **Financially Stressed Class** with very low income on an average. This segment can be targeted with cards with offers such as zero annual charges and lurking them with benefits such as free coupons or tickets and waivers on a variety of places.

**K-means Cluster 2**: This segment has must lower spending per month with low current balance and lower credit limit. This is the **Middle Class** with low incomes. This segment can be targeted with cards that have lower interest rates so as to encourage more spending.

# **Insurance Analysis**

# **Executive Summary**

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.

# **Data Description**

1. Target: Claim Status (Claimed)
2. Code of tour firm (Agency Code)
3. Type of tour insurance firms (Type)
4. Distribution channel of tour insurance agencies (Channel)
5. Name of the tour insurance products (Product)
6. Duration of the tour (Duration in days)
7. Destination of the tour (Destination)
8. Amount worth of sales per customer in procuring tour insurance policies in rupees (in 100’s)
9. The commission received for tour insurance firm (Commission is in percentage of sales)
10. Age of insured (Age)

# **Sample of the Dataset**

Table

Description automatically generated

**Table 7. Insurance Dataset**

Dataset has 3000 rows and 10 columns.

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

# **Exploratory Data Analysis**

A picture containing text, receipt

Description automatically generated

There are total 210 rows and 10 columns in the dataset. 6 columns are of Object datatype, 2 columns are of float datatype and 2 columns are of Integer datatype.

# **Check for missing values in the dataset**

Table

Description automatically generated

From the above results we can see that there is no missing value present in the dataset.

# **Summary of the dataset**

Table

Description automatically generated

* The variable Duration has extreme high and Max values.
* The variable Age has large difference between the 75% and the Max values and also the mean and median.

# **Univariate Analysis**

Chart, box and whisker chart

Description automatically generated

**Fig 15. Univariate – Age**

* The 25th quantile of Age is 31.0
* The median or 50th quantile of Age is 36.0
* The 75th quantile of Age is 43.0
* The Inter quantile range (IQR) of Age is 12.0
* Lower limit of Age is 13.0
* Upper limit of Age is 61.0
* The distribution of Age is right skewed 1.1031446044352335

Chart

Description automatically generated

**Fig 16. Univariate – Commission**

* The 25th quantile of Commision is 0.0
* The median or 50th quantile of Commision is 5.63
* The 75th quantile of Commision is 17.82
* The Inter quantile range (IQR) of Commision is 17.82
* Lower limit of Commision is -26.73
* Upper limit of Commision is 44.55
* The distribution of Commision is right skewed 3.1047406576922842

Chart

Description automatically generated with medium confidence

**Fig 17. Univariate – Duration**

* The 25th quantile of Duration is 12.0
* The median or 50th quantile of Duration is 28.0
* The 75th quantile of Duration is 66.0
* The Inter quantile range (IQR) of Duration is 54.0
* Lower limit of Duration is -69.0
* Upper limit of Duration is 147.0
* The distribution of Duration is right skewed 13.786096073249146

Chart, histogram

Description automatically generated

**Fig 18. Univariate – Sales**

* The 25th quantile of Sales is 20.0
* The median or 50th quantile of Sales is 33.5
* The 75th quantile of Sales is 69.3
* The Inter quantile range (IQR) of Sales is 49.3
* Lower limit of Sales is -53.94999999999999
* Upper limit of Sales is 143.25
* The distribution of Sales is right skewed 2.3446426921667585

**Categorical Univariate Analysis**

**Chart, bar chart

Description automatically generated**

**Fig 19. Univariate – Agency code**

**Chart, bar chart

Description automatically generated**

**Fig 20. Univariate – Type**

**Chart, bar chart

Description automatically generated**

**Fig 21. Univariate – Claimed**

**Chart, bar chart

Description automatically generated**

**Fig 22. Univariate – Channel**

**Chart, bar chart

Description automatically generated**

**Fig 23. Univariate – Product Name**

**Chart, bar chart

Description automatically generated**

**Fig 24. Univariate – Destination**

# **Bivariate Analysis**

**Chart, bar chart

Description automatically generated**

**Fig 25. Agency code – Sales**

**Chart, bar chart

Description automatically generated**

**Fig 26. Type – Sales**

**Chart, bar chart

Description automatically generated**

**Fig 27. Channel – Sales**

**Chart, bar chart

Description automatically generated**

**Fig 28. Product Name – Sales**

**Chart, bar chart

Description automatically generated**

**Fig 29. Destination – Sales**

# **Multivariate Analysis**

**Chart, bar chart

Description automatically generated**

**Fig 30. Agency code – Sales – Claimed**

**Chart, box and whisker chart

Description automatically generated**

**Fig 31. Boxplot- Agency code – Sales – Claimed**

**Chart, bar chart

Description automatically generated**

**Fig 32. Type– Sales – Claimed**

**Chart, box and whisker chart

Description automatically generated**

**Fig 33. Boxplot-Type– Sales – Claimed**

**Chart, bar chart

Description automatically generated**

**Fig 34. -Channel– Sales – Claimed**

**Chart, box and whisker chart

Description automatically generated**

**Fig 35. -Boxplot - Channel– Sales – Claimed**

**Chart, bar chart

Description automatically generated**

**Fig 36. -Product Name– Sales – Claimed**

**Chart, box and whisker chart

Description automatically generated**

**Fig 37. Boxplot -Product Name– Sales – Claimed**

**Chart, bar chart

Description automatically generated**

**Fig 38. -Destination– Sales – Claimed**

**Chart, box and whisker chart

Description automatically generated**

**Fig 39. Boxplot -Destination– Sales – Claimed**

# **Pair Plot:**

**Chart, diagram

Description automatically generated**

**Fig 40. Pair plot**

**Inference**: After plotting the Boxplots for all the numerical variables we can conclude that a very high number of outliers are present in the variables namely, **Age, Commision, Duration and Sales**.

We can conclude from the above graphs that most of the customers doing a claim in our data belong to age group of 25-40 with the type of Tour Agency firm being Travel Agency, Channel being Online, Product name being Customised Plan and Destination being Asia.

# **Heat Map:**

**Chart, treemap chart

Description automatically generated**

**Fig 41. Heat Map**

As interpreted from the above heat map, there is no or extremely low correlation between the variables given in the dataset.

# **Checking the Outliers**

**Chart, box and whisker chart

Description automatically generated**

**Fig 42. Boxplot – Outlier checking**

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

# **Converting Object datatype to Int:**

**A screenshot of a computer

Description automatically generated with low confidence**

For our analysis and building Decision tree and Random Forest, we must convert the variables which have ‘object’ datatype and convert them into integer.

Table

Description automatically generated

**Table 8. Insurance Dataset – Conversion**

# **Splitting Dataset in Train and Test Data (70:30)**

For building the models we will now have to split the dataset into Training and Testing Data with the ratio of 70:30. These two datasets are stored in X\_train and X\_test with their corresponding dimensions as follows

Graphical user interface, text, application, email

Description automatically generated

**Fig 43. Train Test Data**

# **CART Model**

Using the Train Dataset(X\_train) we will be creating a CART model and then further testing the model on Test Dataset(X\_test)

Graphical user interface, text, application

Description automatically generated

Diagram

Description automatically generated

**Fig 44. Decision Tree**

* As we can see the decision tree has overgrown and too many branches are grown. We need to prune the model.
* The depth of the tree is large, and we are unable to predict.
* By using the Grid search method, we can find the best parameters and best estimator.

# **Regularising the Decision Tree**

**Text

Description automatically generated with medium confidence**

# **Best Parameters**

**Text

Description automatically generated with medium confidence**

These best grid parameters are henceforth used to build the regularised or pruned Decision tree.

**Diagram

Description automatically generated**

**Fig 45. Regularising Decision Tree**

# **Variable Importance**

**Table

Description automatically generated with medium confidence**

The agency code is the most important variable, and it is the root node.

# **Random Forest**

Using the Train Dataset(X\_train) we will be creating a Random Forest model and then further testing the model on Test Dataset(X\_test).

# **Grid Search**

**Text

Description automatically generated with low confidence**

# **Best Parameters**

**Text

Description automatically generated**

Using these best parameters evaluated using GridSeachCV a Random Forest Model is created which is further used for model performance evaluation.

# **Variable Importance**

**Table

Description automatically generated with low confidence**

The agency code is the most important variable, and it is the root node.

# **Artificial Neural Network (ANN)**

Firstly, we will have to Scale the two datasets.

Table

Description automatically generated

**Fig 46. Scaled Trained Data**

**Table

Description automatically generated with medium confidence**

**Fig 47. Scaled Test Data**

# **Grid Search**

**Text

Description automatically generated**

# **Best Parameters**

****

# **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.**

To check the Model Performances of the three models created above certain model evaluators are used i.e., Classification Report, Confusion Matrix, ROC\_AUC Score and ROC Plot. They are calculated first for train data and then for test data.

# **CART Model**

# **Classification Report**

Table

Description automatically generated

**Fig 48. Classification Report - Train**

**Table

Description automatically generated**

**Fig 49. Classification Report - Test**

# **Confusion Matrix**

**Text

Description automatically generated**

**Fig 50. Confusion Matrix - Train**

**Graphical user interface, text

Description automatically generated with medium confidence**

**Fig 51. Confusion Matrix - Test**

# **ROC\_AUC Score and ROC Curve**

**Graphical user interface, text, application, chat or text message, email

Description automatically generated** **Chart, line chart

Description automatically generated**

**Fig 52. ROC\_AUC Score and ROC Curve - Train**

**Graphical user interface, text, application

Description automatically generated** **Chart, line chart

Description automatically generated**

**Fig 53. ROC\_AUC Score and ROC Curve - Test**

# **Cart Conclusion**

**Text

Description automatically generated** **Text

Description automatically generated**

* Training and Testing set results are almost similar, and with the overall measures high, the model is good.
* Agency code is the most important variable for pending “Claimed”

# **Random Forest**

# **Classification Report**

**Table

Description automatically generated**

**Fig 54. Classification Report – Train**

**Table

Description automatically generated**

**Fig 55. Classification Report – Test**

# **Confusion Matrix**

**Text

Description automatically generated**

**Fig 56. Confusion Matrix – Train**

**Graphical user interface, text, application

Description automatically generated**

**Fig 57. Confusion Matrix – Test**

# **ROC\_AUC Score and ROC Curve**

**Graphical user interface, text, application

Description automatically generated** **Chart, line chart

Description automatically generated**

**Fig 58. ROC\_AUC Score and ROC Curve – Train**

**Graphical user interface, text, application

Description automatically generated** **Chart, line chart

Description automatically generated**

**Fig 59. ROC\_AUC Score and ROC Curve – Test**

# **Random Forest Conclusion**

**Text

Description automatically generated** **Text

Description automatically generated**

* Training and Testing set results are almost similar, and with the overall measures high, the model is good.
* Agency code is the most important variable for pending “Claimed”

# **Artificial Neural Network (ANN)**

# **Classification Report**

**Table

Description automatically generated**

**Fig 60. Classification Report– Test**

**Table

Description automatically generated**

**Fig 61. Classification Report– Train**

# **ROC\_AUC Score and ROC Curve**

**Graphical user interface, text, application, email

Description automatically generated** **Chart, line chart

Description automatically generated**

**Fig 62. ROC\_AUC Score and ROC Curve– Test**

**Graphical user interface, text, application

Description automatically generated** **Chart, line chart

Description automatically generated**

**Fig 63. ROC\_AUC Score and ROC Curve– Train**

# **Confusion Matrix**

**Graphical user interface, text, application

Description automatically generated**

**Fig 64. Confusion Matrix– Train**

**A picture containing application

Description automatically generated**

**Fig 65. Confusion Matrix– Test**

# **Artificial Neural Network (ANN) Conclusion**

**Text

Description automatically generated** **Text

Description automatically generated**

* Training and Testing set results are almost similar, and with the overall measures high, the model is good.
* Agency code is the most important variable for pending “Claimed”

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

**Table

Description automatically generated**

**Table 9. Model Comparison**

From the above table, comparing the model performance evaluators for the three models it is quite clear that the Random Forest Model is performing well as compared to the other two as it has high precisions for both training and testing data and although the AUC Score is the same for all the three models for training data but for testing data it is the highest for Random Forest Model. Choosing Random Forest Model is the best option in this case as it will exhibit very less variance as compared to a single decision tree or a multi – layered Neural Network.

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

Business insights and analysis:

* JZI and EPX agency has less sales as compared to other agencies hence they need to do more marketing promotion as well as need to be trained from other agencies.
* There are more sales for Airlines as compared to Travel agency, hence travel agency should work on their sales
* People prefer online channel as compared to offline and online channel has better sales
* Streamlining online experience will also increase traffic on website as well as customer satisfaction, ease of doing business or mode of business will increase the sales and will increase the commission too
* People prefer Gold plan and silver plan as compared to other plans
* Gold and silver plan gives more sales; hence marketing promotion need to be done for other plans as well as more features to the insurance schemes to increase the sales for the other plans
* There are maximum sales is in America and other continents like Asia and Europe need little more promotion and add on features to the insurance schemes increase the sales. More variable (external information) are required like states, paying insurance behaviour patterns etc. to the dataset
* The best model (Random Forest) gives an 81 % accuracy hence the above-mentioned points can improve the accuracy
* Other ways the insurance company can increase the sales via reduce cycle time, improve customer satisfaction, optimize claims recovery, and try to improve present insurance features or compare the features from their competitors and try to give better offers.