DATA MINING PROJECT

S.VISHAL

G3 PGPDSBA JULY

TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| SNO | TITLE | PAGE NO |
| 1 | INTRODUCTION | 4 |
| 2 | DATA DICTIONARY OF PROBLEM 1 | 4 |
| 3 | SOLUTIONS | 5 |
| 4 | CONCLUSION | 21 |
| 5 | DATA DICTIONARY OF PROBLEM 2 | 22 |
| 6 | SOLUTION | 22 |
| 7 | CONCLUSION | 39 |
| 8 | LIST OF TABLES AND FIGURES | 3 |

LIST OF TABLES AND FIGURES

|  |  |  |
| --- | --- | --- |
| SNO | TITLE | PAGE NO |
|  | Sample of the dataset | 6 |
|  | Types of data in Dataset | 7 |
|  | Info of the dataset | 7 |
|  | Shape of the dataset | 7 |
|  | Summary of the Dataset | 8 |
|  | Missing Values in data set | 8 |
|  | Univariate Analysis of spending Variable | 9 |
|  | Univariate Analysis of advance\_payments Variable | 9 |
|  | Univariate Analysis of probability\_of\_full\_payments Variable | 9 |
|  | Univariate Analysis of current\_balance Variable | 10 |
|  | Univariate Analysis of credit\_limit Variable | 10 |
|  | Univariate Analysis of min\_payment\_amt Variable | 10 |
|  | Univariate Analysis of max\_spent\_in\_single\_shopping Variable | 11 |
|  | Box plot to check for outliers | 11 |
|  | Histograms for all the variables | 12 |
|  | Distplot for spending,advance\_payments,probability\_of\_full\_payments,current\_balance | 13 |
|  | Distplot for credit\_limit,min\_payment\_amt,max\_spent\_in\_single\_shopping | 13 |
|  | Pairplot for all the variables | 14 |
|  | Heatmap | 15 |
|  | Before Scaling | 15 |
|  | After Scaling | 16 |
|  | Scaled Data | 16 |
|  | Cluster | 16 |
|  | Dendrogram | 17 |
|  | Optimum Number Of Clusters | 17 |
|  | Cluster Frequency | 17 |
|  | WSS Score | 18 |
|  | Elbow Curve | 18 |
|  | Silhouette Score | 19 |
|  | 3 Group Cluster | 19 |
|  | 3 group Cluster Hierarchical Clustering | 19 |
|  | Cluster frequency Hierarchical | 20 |
|  | Cluster frequency k\_means | 20 |
|  | Sample of the dataset | 22 |
|  | Info of the dataset | 22 |
|  | Shape of the dataset | 23 |
|  | Missing Values | 23 |
|  | Duplicate Entries | 23 |
|  | 5 point Summary | 24 |
|  | Box plot for Age | 24 |
|  | DistPlot and Histograms for Age | 25 |
|  | Box plot for Commision | 25 |
|  | DistPlot and Histograms for Commision | 25 |
|  | Box plot for Duration | 26 |
|  | DistPlot and Histograms for Duration | 26 |
|  | Box plot for sales | 26 |
|  | DistPlot and Histograms for Sales | 27 |
|  | Count Plot for Agency code | 27 |
|  | Count Plot for Type | 27 |
|  | Count Plot for Channel | 28 |
|  | Count Plot for Product Name | 28 |
|  | Count Plot for Destination | 28 |
|  | Pairplot | 29 |
|  | Heat Map | 30 |
|  | New dataset | 30 |
|  | Variable Importance DT | 31 |
|  | Variable Importance RF | 31 |
|  | NN | 31 |
|  | CART AUC and ROC For Training Data | 31 |
|  | RF AUC and ROC | 32 |
|  | NN AUC and ROC | 32 |
|  | Confusion matrix for Decision Tree Test Data | 32 |
|  | Confusion Matrix of Decision Tree Training data | 32 |
|  | Classification Report of Decision Tree test Data | 33 |
|  | Classification Report of Decision Tree Training Data | 33 |
|  | Precision,Recall,f1score of CART Test data | 33 |
|  | Precision,Recall,f1score of CART Training data | 33 |
|  | Confusion Matrix of RF Training data | 33 |
|  | Classification Report of RF training Data | 34 |
|  | Precision,Recall,f1score of RF Training data | 34 |
|  | Confusion Matrix of RF Test data | 34 |
|  | Classification Report of RF test Data | 34 |
|  | Precision,Recall,f1score of RF Test data | 34 |
|  | Confusion Matrix Of NN Training data | 34 |
|  | Classification Report Of NN training Data | 35 |
|  | Precision,Recall,f1score of NN Training data | 35 |
|  | Confusion Matrix Of NN test data | 35 |
|  | Classification Report Of NN test Data | 35 |
|  | Precision,Recall,f1score of NN Test data | 35 |
|  | COMPARISON OF PERFORMANCE METRICES OF ALL 3 MODELS | 37 |
|  | ROC for all 3 MODELS training Data | 37 |
|  | ROC for all 3 MODELS test Data | 38 |
|  | TABLES |  |
| 1.1 | Data Dictionary For problem 1 | 5 |
| 1.2 | Data Dictionary for problem 2 | 22 |
| 1.3 | CART MODEL | 35 |
| 1.4 | RF MODEL | 36 |
| 1.5 | NN MODEL | 36 |
|  |  |  |

INTRODUCTION

1. 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.
   1. Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).
   2. Do you think scaling is necessary for clustering in this case? Justify

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

**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.

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

The dataset provided to us is stored as “bank\_marketing\_part1\_Data.csv”

* 1. DATA DICTIONARY

|  |  |  |
| --- | --- | --- |
| SNO | VARIABLE | 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) |

Table 1: Data Dictionary

PACKAGES USED

* pandas as pd
* numpy as np
* seaborn as sns
* from sklearn.cluster import KMeans
* matplotlib.pyplot as plt
* from sklearn.preprocessing import StandardScaler
* from scipy.cluster.hierarchy import dendrogram, linkage
* from scipy.cluster.hierarchy import fcluster
* from sklearn.metrics import silhouette\_samples, silhouette\_score
* from scipy import stats

1.2 SOLUTIONS

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

Importing the dataset

Given dataset “bank\_marketing\_part1\_Data” is a csv file, imported in jupyter notebook using pd.read\_csv () and stored in “df”. Top 10 rows of the data set are shown below using df.head ( ):

Table

Description automatically generated

Fig 1: Sample of the dataset

OBSERVATIONS: Data looks good based on initial records seen.

Structure and Dimensions of the Dataset

df.shape provides the dimensions of the dataset, df.dtypes provides the datatype of all variables in the dataset and df.info () shows the structure of the dataset

Text

Description automatically generated

Fig 2: Types of Data in the Dataset

Text

Description automatically generated

Fig 3: Info of the Dataset



Fig 4: Shape of the Dataset

OBSERVATIONS: From the above figures we find the shape, types of data and the various information about the datatypes and from the we found out that the dataset has 7 variables and 210 records, no missing records are found, and all the variables are of numeric type

Summary of the Dataset: .describe()

Graphical user interface

Description automatically generated

Fig 5: Summary of the Dataset

OBSERVATIONS: The 5 point summary of the dataset and we see for most of the variables, mean/medium are nearly equal and Standard Deviation is high for Spending

Checking for Missing values:

Missing or null values are to be identified before analysing the data to either remove or replace the values for the ease of evaluation. Missing Values can be computed using. isnull().sum() function

Text

Description automatically generated

Fig 6: Missing Values

OBSERVATIONS: No Missing values are present in the Dataset

Univariate Analysis

The Univariate Analysis is done for all the variables, the distribution of the variable in the dataset and the outliers present in each of the variables are noted. Boxplots,Distplots and Histograms are used for the univariate analysis

Chart, histogram

Description automatically generated

Fig 7: Univariate analysis of spending Variable

Chart, histogram

Description automatically generated

Fig 8: Univariate Analysis of advance\_payments variable

Chart, histogram

Description automatically generated

Fig 9: Univariate Analysis of probability\_of\_full\_\_payment variable

Chart, histogram

Description automatically generated

Fig 10: Univariate Analysis of current\_balance variable

Chart, histogram

Description automatically generated

Fig 11: Univariate Analysis of credit\_limit variable

Chart, histogram

Description automatically generated

Fig 12: Univariate Analysis of min\_payment\_amt variable

Chart, histogram

Description automatically generated

Fig 13: Univariate Analysis of max\_spent\_in\_single\_shopping variable

Checking for Outliers

Chart, box and whisker chart

Description automatically generated

Fig 14: Box plot to check for Outliers

OBSERVATIONS: This box plot shows that only the variable” min\_payment\_amt “has a few outliers. Since only one of the seven variables have a very small outlier value, hence there is no need to treat the outliers. This small value will not create any difference in our analysis. We can conclude from the above graphs that the 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

The Histograms is plotted for all the independent variables

Chart, bar chart

Description automatically generated

Fig 15: Histograms for all the variables in the dataset

Skewness is checked for all the variables using distplots and a KDE plot is plotted to find the probability density at different values in a continuous variable

Diagram

Description automatically generated

Fig16: Distplotforspending,advance\_payments,probability\_of\_full\_payments,current\_balance

A picture containing graphical user interface

Description automatically generated

Fig17: Distplot for credit\_limit, min\_payment\_amt, max\_payment\_amt

OBSERVATIONS: Credit limit average is around $3.258(10000s), distribution is skewed to the right tail for all the variable except probability\_of\_full\_payment variable, which has left tail

MULTIVARIATE ANALYSIS

Checking for Multicollinearity

A picture containing shape

Description automatically generated

Fig 18: Pairplot for the variables

OBSERVATIONS: From the pair plot we found that there is a strong correlation between spending & advance\_payments, advance\_payments & current\_balance, credit\_limit & spending, spending& current\_balance, credit\_limit and advance\_payments max\_spent\_in\_single\_shopping, current\_balance

Table

Description automatically generated

Fig 19: Heat map

OBSERVATIONS: 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

• 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

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

Scaling needs to be done as the values of the variables are different. spending, advance\_payments are in different values, and this may get more weightage. Also have shown below the plot of the data prior and after scaling. Scaling will have all the values in the relative same range. I have used zscore to standardise the data to relative same scale -3 to +3.

A picture containing chart

Description automatically generated

Fig 20: Before Scaling

Chart, line chart

Description automatically generated

Fig 21: After Scaling

Graphical user interface, application

Description automatically generated

Fig 22: Scaled Data

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

Clustering is a widely accepted Unsupervised Learning technique in Machine Learning

Clustering can be divided into two categories namely, Hierarchical and K-means clustering. Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly like each other. There are two types of hierarchical clustering, Divisive and Agglomerative.

Chart, histogram

Description automatically generated

Fig 23: Cluster

Clustering of the data has been done with the Ward's method for linkage, part of Agglomerative Hierarchical Clustering method.

To create a Dendrogram using our scaled data we have firstly imported the package dendrogram, linkage from scipy.cluster

Chart, histogram

Description automatically generated

Fig 24: Dendrogram

We use fcluster() to find the optimal number of clusters. With Ward link method, and criterion as maxclust with value as 3 we find the optimal number of clusters. Also, when you look at the dendrogram, it seems that 2 clusters would be optimal. But that is not the only evidence. With distance criterion and value as 20-25, we see that that the optimal number of clusters shall be 3. This cuts a horizontal line at the value mentioned and the number of lines vertical lines cut by this horizontal line gives the optimal number of clusters.

A picture containing text, outdoor

Description automatically generated

Fig 25: Optimum number of clusters

A picture containing table

Description automatically generated

Fig 26: Cluster Frequency

* 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 one of the unsupervised machine learning algorithms. K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. For the dataset we will be using K-means clustering on scaled data and identify the clusters formed and use them further to devise tools to target each group separately. Firstly, we will calculate the value of inertia and store it in WSS as mentioned in the image below: Calculated within sum of squares variance for the clusters ranging from 2 to 11:

Text

Description automatically generated

Fig 27: WSS score

we will now plot an elbow curve to determine the optimal number of clusters(k) to be used for our clustering using within sum of squares(wss) method

Chart, line chart

Description automatically generated

Fig 28: 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.

Silhouette score Calculation:

The silhouette score is a measure of how close each point in one cluster is to points in the neighbouring clusters. From sklearn.metrics, we import silhouette\_samples, silhouette\_score function to do this.

A picture containing text, bottle, document

Description automatically generated

Fig 29: Silhouette Score

The silhouette scores are calculated using silhouette\_samples and silhouette\_score package from sklearn.metrics. The average silhouettes score is coming to be 0.400 . The silhouette score ranges from -1 to +1 and higher the silhouette score better the clustering.

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

Text

Description automatically generated

Fig 30: 3 Group cluster Kmeans

Table

Description automatically generated

Fig 31: 3 Group cluster hierarchical Clustering

Graphical user interface, application

Description automatically generated

Fig 32: Cluster Frequency Hierarchical

Table

Description automatically generated with medium confidence

Fig 33: Cluster Frequency K\_Means

# Cluster Group Profiles

# Group 1: High Spending

# Group 3: Medium Spending

# Group 2: Low Spending

# Promotional strategies for each cluster

## Group 1: High Spending Group

* + Giving any reward points might increase their purchases.
  + maximum max\_spent\_in\_single\_shopping is high for this group, so can be offered discount/offer on next transactions upon full payment
  + Increase their credit limit and
  + Increase spending habits
  + Give loan against the credit card, as they are customers with good repayment record.
  + Tie up with luxury brands, which will drive more one\_time\_maximun spending

## Group 3 : Medium Spending Group

* + They are potential target customers who are paying bills and doing purchases and maintaining comparatively good credit score. So we can increase credit limit or can lower down interest rate.
  + Promote premium cards/loyalty cars to increase transactions.
  + Increase spending habits by trying with premium ecommerce sites, travel portal, travel airlines/hotel, as this will encourage them to spend more

## Group 2: Low Spending Group

* + Customers should be given remainders for payments. Offers can be provided on early payments to improve their payment rate.
  + Increase their spending habits by tying up with grocery stores, utilities (electricity, phone, gas, others)

1.3 CONCLUSION

Those who are clustered under 3rd cluster are seen as elite customers who spend more monthly and all other parameters are also relatively high compared to members of other clusters. It is seen that minimum paid by the customer while making payments for purchases made monthly (in 100s) is high for those under the second cluster. The credit limit for those under 2nd can be increased as a promotional strategy as their averages of all parameters are relatively same compared to others. Max amount spent on a single purchase is the same for those under 1 and 2 clusters. Customers can spend on different products. Customers under third cluster seem to spend more, but their minimum payment is less than those of second cluster customers. Bank shall increase their slab of min payment amount. Advance payments made by second cluster customers are less, which the bank shall investigate

**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).

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

**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.

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

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

2.1 DATA DICTIONARY OF PROBLEM 2

|  |  |
| --- | --- |
| SNO | ATTRIBUTE |
| 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) |

Table 1.2 Data Dictionary for Problem 2

2.2 SOLUTIONS

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

Graphical user interface

Description automatically generated with medium confidence

Fig 34: Sample of the dataset

SHAPE AND INFO OF THE DATA SET

Text

Description automatically generated with medium confidence

Fig 35: Info of the dataset

Graphical user interface, text, application

Description automatically generated

Fig 36: Shape of the dataset

OBSERVATIONS: From the above figures we now know the shape and types of data present in the given data set. The data set has 3000 observations and 10 variables in the data set .Out of the 10 variables, 5 are categorical, 4 are Numerical and remaining 1 is Boolean.9 independent variable and one target variable – Claimed

CHECKING FOR MISSING VALUES

A picture containing table

Description automatically generated

Fig 37: Missing values

OBSERVATIONS: From the above figure we find that the given dataset does not have any missing values

CHECKING FOR DUPLICATE ENTRIES

Graphical user interface

Description automatically generated

Fig 38: Duplicate entries

OBSERVATIONS: There are 139 duplicate entries

• We don't have any Customer Id or Customer name or any other unique value to analyse whether these are real duplicates or not

• Travel company can sell the same kind of tour package to similar demography

• Benefit of the doubt I will not be dropping these duplicates

STATISTICAL SUMMARY

Table

Description automatically generated

Fig 39: 5 Point Summary

OBSERVATIONS: Information on the Statistical Summary:

• Mean and Median are not similar for all the variables

• All the variables have higher standard deviation

• Duration minimum is in negative, looks like wrong entry

• Age range starting from 8 to 84

UNIVARIATE ANALYSIS

Chart, box and whisker chart

Description automatically generated

Fig 40: Box Plot for age variable

Chart, histogram

Description automatically generated

Fig 41: Distplot and Histogram for Age

Chart, box and whisker chart

Description automatically generated

Fig 42: Boxplot for Commision

Chart, histogram

Description automatically generated

Fig 43: Distplot and Histogram for Commision

Chart

Description automatically generated

Fig 44: Box plot for Duration

Chart, bar chart, histogram

Description automatically generated

Fig 45: Distplot and Histogram for duration

Chart, box and whisker chart

Description automatically generated

Fig 46: Boxplot for Sales

Chart, histogram

Description automatically generated

Fig 47: Distplot and Histogram for Sales

CATEGORICAL VARIABLES

Chart, bar chart

Description automatically generated

Fig 48: Countplot for Agency\_code

Chart, bar chart

Description automatically generated

Fig 49: Countplot for Type

Chart, bar chart

Description automatically generated

Fig 50: Countplot for Channel

Chart, bar chart

Description automatically generated

Fig 51: Countplot for Product Name

Chart, bar chart

Description automatically generated

Fig 52: Countplot for Destination

Chart, scatter chart

Description automatically generated

Fig 53: Pairplot

OBSERVATIONS:

• All the variables are positively skewed

• There is no strong positive correlation

Chart

Description automatically generated

Fig 54: Heat Map

OBSERVATIONS:

• There is no strong correlation

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

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

Table

Description automatically generated

Fig 55: New dataset

Text

Description automatically generated

Fig 56: Variable Importance Decision Tree

Text

Description automatically generated

Fig 57: Variable Importance Random Forest

Text

Description automatically generated

Fig 58: Neural Network

**2.3** Performance Metrics: Comment and Check the performance of Predictions on Train and Test sets using Accuracy,

Chart, line chart

Description automatically generated

Fig 59: CART AUC and ROC for TRAINING DATA

Chart, line chart

Description automatically generated

Fig 60: RF AUC and ROC

Chart, line chart

Description automatically generated

Fig 61: NN AUC and ROC

Text

Description automatically generated with medium confidence

Fig 62: Confusion Matrix Decision Tree test data



Fig 63: Confusion Matrix Decision Tree training data

Table

Description automatically generated

Fig 64: Classification Report Decision Tree of Test data

Table

Description automatically generated

Fig 65: Classification Report Decision Tree of Training data

Text

Description automatically generated with low confidence

Fig 66: Precision,Recall,f1 score of CART MODEL Test Data

Text

Description automatically generated

Fig 67: Precision,Recall,f1 score of CART MODEL Training Data

Text

Description automatically generated

Fig 68: Confusion Matrix for RF Training data

Table

Description automatically generated

Fig 69: Classification report for RF of Training data

A picture containing text

Description automatically generated

Fig 70: Precision,Recall,f1 score of RF MODEL Training Data

Text

Description automatically generated

Fig 71: Confusion Matrix for RF Test data

Table

Description automatically generated

Fig 72: Classification report for RF of Test data

A picture containing text

Description automatically generated

Fig 73: Precision,Recall,f1 score of RF MODEL Test Data

Text

Description automatically generated with low confidence

Fig 74: Confusion Matrix of NN Training data

Table

Description automatically generated

Fig 75: Classification Report Of NN Training data

A screenshot of a computer

Description automatically generated with medium confidence

Fig 76: Precision, Recall,f1 score of NN Training Data

Text

Description automatically generated with medium confidence

Fig 77: Confusion Matrix Of NN Test Data

Table

Description automatically generated

Fig 78: Classification report Of NN Test data

A screenshot of a computer

Description automatically generated with medium confidence

Fig 79: Precision, Recall,f1 score of NN Test data

|  |  |
| --- | --- |
| [Cart Conclusion](http://localhost:8888/notebooks/Downloads/insurance-cart-rf-ann-models.ipynb#Cart-Conclusion) |  |
|  |  |
| [Train Data:](http://localhost:8888/notebooks/Downloads/insurance-cart-rf-ann-models.ipynb#Train-Data:) | [Test Data:](http://localhost:8888/notebooks/Downloads/insurance-cart-rf-ann-models.ipynb#Test-Data:) |
|  |  |
| - AUC: 82% | - AUC: 80% |
| - Accuracy: 79% | - Accuracy: 77% |
| - Precision: 70% | - Precision: 80% |
| - f1-Score: 60% | - f1-Score: 84% |

Table 1.3 CART MODEL

OBSERVATION: Training and Test set results are almost similar, and with the overall measures high, the model is a good model. Change is the most important variable for predicting diabetes

|  |  |
| --- | --- |
| Random Forest Conclusion¶ |  |
|  |  |
| [Train Data:](http://localhost:8888/notebooks/Downloads/insurance-cart-rf-ann-models.ipynb#Train-Data:) | [Test Data:](http://localhost:8888/notebooks/Downloads/insurance-cart-rf-ann-models.ipynb#Test-Data:) |
|  |  |
| - AUC: 86% | - AUC: 82% |
| - Accuracy: 80% | - Accuracy: 78% |
| - Precision: 72% | - Precision: 68% |
| - f1-Score: 66% | - f1-Score: 62 |

Table 1.4 RF MODEL

OBSERVATION: Training and Test set results are almost similar, and with the overall measures high, the model is a good model. Change is again the most important variable for predicting diabetes

|  |  |
| --- | --- |
| [Neural Network Conclusion](http://localhost:8888/notebooks/Downloads/insurance-cart-rf-ann-models.ipynb#Neural-Network-Conclusion) |  |
|  |  |
| [Train Data:](http://localhost:8888/notebooks/Downloads/insurance-cart-rf-ann-models.ipynb#Train-Data:) | [Test Data:](http://localhost:8888/notebooks/Downloads/insurance-cart-rf-ann-models.ipynb#Test-Data:) |
|  |  |
| - AUC: 82% | - AUC: 80% |
| - Accuracy: 78% | - Accuracy: 77% |
| - Precision: 68% | - Precision: 67% |
| - f1-Score: 59 | - f1-Score: 57% |

Table 1.5 NN MODEL

OBSERVATION: Training and Test set results are almost similar, and with the overall measures high, the model is a good model. Change is again the most important variable for predicting diabetes

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

Table

Description automatically generated

Fig 80: COMPARISON OF PERFORMANCE METRICES OF ALL 3 MODELS

Chart, line chart

Description automatically generated

Fig 81: ROC FOR ALL 3 MODELS Training Data

Chart, line chart

Description automatically generated

Fig 82: ROC FOR ALL 3 MODELS Test Data

OBSERVATIONS:

Out of the 3 models, Random Forest has slightly better performance than the Cart and Neural network model. Overall, all the 3 models are reasonably stable enough to be used for making any future predictions. From Cart and Random Forest Model, instead of creating a single Decision Tree it can create multiple decision trees as Random Forest have much less variance than a single decision tree and hence can provide the best claim status from the data. The Agency\_Code is found to be the most useful feature amongst all other features for predicting if a person will claim for insurance or not.

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

• The Accuracy, Precision, F1 Score are computed using Classification Report. The confusion matrix, AUC\_ROC Scores and ROC plot are computed for each model separately and compared. All the three models have performed well but to increase our accuracy in determining the claims made by the customers we can choose the Random Forest Model. Instead of creating a single Decision Tree it can create multiple decision trees and hence can provide the best claim status from the data. 34

• For all the models i.e., CART, Random Forest and ANN have performed exceptionally well. Hence, we can choose either of the models but choosing Random Forest Model is a great option as even though they exhibit the same accuracy but choosing Random Forest over Cart model is way better as they have much less variance than a single decision tree.

• By performing the 3 models, we can conclude that - the data set is well balanced to conduct the modelling. - The data set is containing significant outliers - The Agency code has significant importance

• Claims are Higher for Online Distribution channel of tour insurance agencies. Claims are very low for Offline Distribution channel of tour insurance agencies. Reason might be, in recent time many people are preferring Online purchase which is very easier. So, Management can think of promoting offline Distribution channel of tour insurance agencies in order to reduce claims. Offline Purchase can be made more attractive by offering extra discounts or additional benefits.

• Higher Claims are observed for Agency Code C2B. So, Management needs to check why claim state is high for this agency. Reason might be lack of knowledge to the insurance representative on insurance policies which leads loopholes. This might be leading to high claims. There is also a possibility that people might be purposely taking insurance from this agency for the reason of easy claims. This agency services can be compared with other agencies with are leading to fewer number of claims.

• Claims are higher for Airlines as Type of tour insurance firms. There are many factors involved in this like flight delay, baggage delays poor service recorded by the Airline Service or connection flight missing. These are usual; hence terms and conditions can be added or delay time frame can be increased. For baggage loss or any other delays, we can increase the premium. Or claim value can be reduced by some percentage.

• Claims are higher for silver plan which is one of the Name of the tour insurance products. We need to recheck for which reason are we getting claims from this silver plan, or this plan can be stopped for some time to confirm if this reduces the claims. Other plans can be promoted in order to reduce the claims.

• Claims are higher for Destination of the tour as ASIA. We need to dig out the reason why there are high claims for ASIA. Several reasons are present for it. People are negligent may be in ASIA. Terms and Conditions need to be increased for ASIA. There is also possibility that we can increase the premium for insurances related to ASIA in order to compensate for the claims. If customers are checking for ASIAN travels, we can give them more interesting options of other places to travel.

• The category for which company is getting more claims can be added as addons at extra cost. And the category for which company is getting less claims can be kept as basic facilities. And this can be varied based on location of travel.

• Management can also think of keeping the purchase procedure simple, but we need to increase the complexity of the Claim Procedure. E.g., Claiming can be done only by offline mode, so as to verify the genuineness of the claim.

2.3 CONCLUSION

CART, RF & ANN are used to compare the models' performances in train and test sets.RF model was found as the superior model than the CART and ANN due to the less variance than a single decision tree and hence can provide the best claim status from the data. The Agency\_Code is found to be the most useful feature amongst all other features for predicting if a person will claim for insurance or not. The inferences were made, and recommendations were given to the company