**Data Mining**

# Project Report

**Problem 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

1.1 Read the data and do exploratory data analysis. Describe the data briefly.

**Exploratory Data Analysis**

**A screenshot of a cell phone

Description automatically generated**

The data has 7 variables, Spending, Advance payments, Probability of full payment, current balance, credit limit, min\_payment\_amt, max spent in single shopping. The data type for all variables in float. The shape of the data is 210 rows and 7 columns.

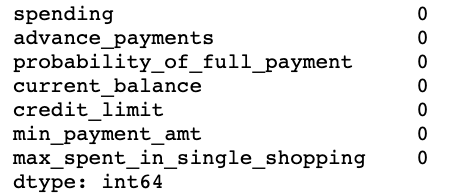
Descriptive Statistics for the data set-

A screenshot of a cell phone

Description automatically generated

We use the df.describe function to get the 5 point summary for all the variables.

We check for null values using the df.isnull().sum() function



We can see that there are no null values in the data set

Using the df.hist function we see the distribution of each of the variables. This helps us to identify if any variables are skewed or not

We also plot the boxplot to identify if the data has any outliers.

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

Yes I do think scaling is necessary for clustering in this case, as the data set provided to us, is in 100’s or 1000’s. While performing clustering we need to make sure that the data is standardized in order for us to get accurate results

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

To scale the data we import the standardscaler module. We then fit and transform the data. From the scipy library we import dendrogram and linkage method to perform hierarchical clustering

A screenshot of a cell phone

Description automatically generated

We then truncate the dendrogram to identify the optimum number of clusters.

A screenshot of a cell phone

Description automatically generated

From the truncated dendrogram we can see that the optimum number of clusters is 2

We then import fclusters and append the respective cluster values to each row

**1.4** Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score.

We import KMeans from sklearn.cluster module. We then define the number of clusters. Based on the dendrogram we mention the n\_clusters as 2

We then calculate the inertia score using the kmeans.inertia\_ command

We calculate the inertia , for multiple number of clusters in order to identify the optimum number of clusters and then plot the elbow curve. We calculate the wss in order to plot the elbow curve. From the elbow curve we can see that there is a significant drop at n=2 and n=3. We can see that drop is more significant from 1 to 2, therefore we can identify the optimum number of clusters as 2.

We then calculate the silhouette score which gives us an idea of the distance between the values in a cluster. The silhouette score for 2 clusters is 0.46

A screenshot of a cell phone

Description automatically generated

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

**A screenshot of a cell phone

Description automatically generated**

**Basis the cluster profiling we can see that**

Cluster 0**:** People with relatively lower per month spending, current balance, credit limit and max spent in single and higher minimum payment amount

Cluster 1: People with higher per month spending current balance, credit limit and max spent in single and lower minimum payment amount

**Basis the cluster profiling, the bank can choose to target both cluster for different promotion categories**

**Cluster 1 can be targeted, for high priced products, with cash back promotions in order to encourage higher spending**

**Cluster 0 can be targeted and be offered promotional interest rates on new credit cards in order for them to get more cards further leading to increased spending**

**For both clusters, the probability of full payment is very high therefore the risk is lower as opposed to having a low probability of full payment**

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

We import the data using the pd.read function. We also import all the necessary libraries to fit the Cart, Random Forest and Neural Network Models, along with Classification reports, ROC-AUC curves

We use the df.head function to see if the data has been imported properly.

Further we look at the summary of the data, using the df.info and df. describe function

df.info function gives us the data types of each variable and number of non-null values

df. describe tells us the 5 point summary of the continuous variables

We also check if there are any duplicate values in the dataset using the df.duplicated function

We can perform univariate analysis by plotting the histogram to understand the distribution of the data. The Df. hist function helps us visualize the distribution of the data and understand if there are any highly skewed variables

We also plot the boxplot to identify if the data has any outliers. From the boxplot we can observe that all the numerical variables have outliers. In case of presence of outliers we need to either impute them or remove them from the data so that it does not affect our insights.

We use the sns.pairplot() function to plot the relationship between all the numerical variables in the dataset. The histograms on the diagonal shows distribution of each variable whereas the scatter plots on the upper and lower triangles show the relationship between two variables. We can see that all the plots show a linear relationship between the variables

We also plot a heatmap to view the correlation between the continuous variables. WE can observe that there is high correlation between sales and commission of 0.77

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

For using any classification model, we first need to split the data into training and testing data in the proportion of 70% train and 30% test data. We use the sklearn.train\_test\_split function to split the data into train and test. The **training** set contains a known output and the model learns on this **data** in order to be generalized to other **data** later on. We have the **test** dataset (or subset) in order to **test** our model's prediction on this subset

To perform the classification, we need to convert the categorical variables into numerical ones, using codes. This assigns numerical values to each unique value in the respective variables, making it easier for us to analyse.

Once we convert variables, we have to understand the proportion of 0 and 1s in the target variable. In this case the column ‘Claimed’ is our target variable where

Value = 0 – Person has not made claims

Value = 1 – Person has made claims

Once we do all this, we define our x and y by dropping and populating the target variable In x and y respectively

We then split the data into test and train

For each of the classification models, we follow the following steps-

1.We define the parameter grid and assign a variable with the respective classifier (Decision Tree, Random Forest, or Artificial Neural Networks)

2. Perform grid search cross validation, using the classification model as the estimator - – This helps us identify the best parameters to be used for model classification

3. We then fit the grid search variable on the train data, and find the best parameters basis the values mentioned by us in the parameter grid – This helps the train model learn,in order to make predictions for the test data

4. Using the best parameter values we use it to predict on the test and train data

Please refer the appendix for code for each of the steps for respective classification models

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

5. We then get the predicted classes and probabilities on the test data –

6. We then perform the model evaluation using the ROC and AUC for the training and test data respectively – This gives us the area under the curve

7. We then calculate the confusion matrix for train and test data

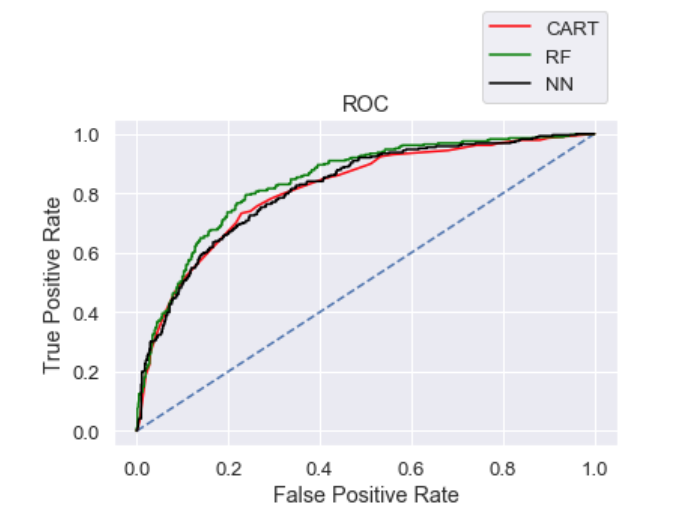
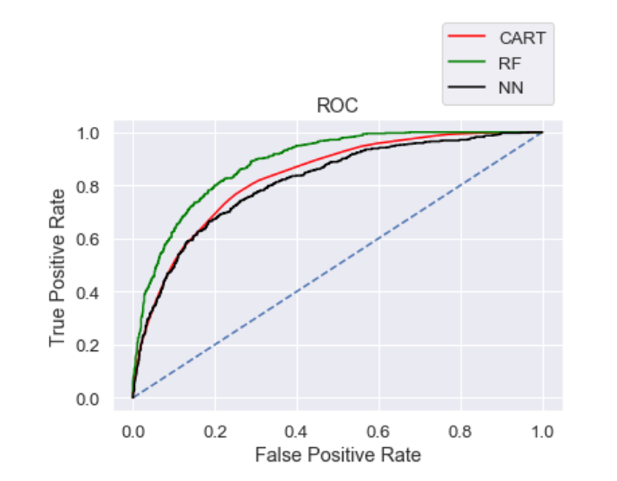
8. Finally we execute the command for the classification report and accuracy for both train and test data

Based on the AUC, Accuracy, sensitivity, precision and f1-score values, of the train and test data, we draw our conclusions about the model

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

A screenshot of a cell phone

Description automatically generated



Basis the table we can see the that CART model has the highest AUC score for both train and test data. However out of the 3 models we see that Random forest has a slightly better performance than the other two models. Overall all the 3 models are not reasonably stable enough to be used for making any future predictions. From Cart and Random Forest Model, the variable ‘Product Name’ is found to be the most useful feature amongst all other features for predicting if a person has diabetes or not. If change is yes, then those patients have more chances of getting diabetes.

**2.5** Inference: Basis on these predictions, what are the business insights and recommendations

Basis these classification models, we can predict whether a claim will be made or not. We also know that product category plays an important role and the company can further assess if a certain type of plan is leading to more claims.

A deeper dive can be done to find out what are the claims being made against and who/ which agency is making these claims. The company can also be prepared whether a claim will be made or not and can make changes to their policies in order to mitigate any uncovered risks