1. **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.

**1.1** **Read the data and do exploratory data analysis (3 pts). Describe the data briefly. Interpret the inferences for each (3 pts). Initial steps like head() .info(), Data Types, etc . Null value check. Distribution plots(histogram) or similar plots for the continuous columns. Box plots, Correlation plots. Appropriate plots for categorical variables. Inferences on each plot. Summary stats, Skewness, Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct.**

|  |
| --- |
| **Read the data and do exploratory data analysis (3 pts)** |
| 1. df.head(): initial 10 head records have no anomalies; it tells us that the data provided to us is of optimal quality. 2. df.info(): Returns a total of 7 columns and 210 rows. All rows are Non-Nulls. The DataType for all columns is float64. No categorical columns were provided to us in the data set. 3. df.isnull().sum() and dum.duplicated().sum(): There are no duplicates or null values in the data set. Provided data set is impressively clean. 4. Spending and advance payments seem to go hand in hand, specifying, user is flexible to make an advance payment in the case if his spending is high. 5. Rest all information is available below. |
| **Describe the data briefly. Interpret the inferences for each (3 pts)** |
| df.describe():   * 1. Values of spending and advance\_payments is high, we get to infer this result using the mean and standard deviation values.   2. IQR and Central Tendency values for the spending and advance\_payments is higher when compared to other columns.   3. Also, probability\_of\_full\_payment is a probability value which hovers around 0-1 and hence the IQR and CT values are substantially smaller.   4. Current\_balance feature seem to have the half value of the spending, specifying the users maintain at par half current balance in their accounts at all times.   5. Credit\_limit feature has to have a higher value than the spending, looks unsure as it has minimum values than Spending. For example, credit\_limit of 4 lakhs meaning, user can spend only until 4 lakhs. This reason, could also be a candidate for Scaling.   6. Min\_payment\_amount of the user for the credit card seems to be fine with min and max ranging from 0 and 8.   7. Max\_spent\_in\_shopping of the user is does not substantial share in the whole pending of the user, specifying user is spending in multiple intervals of the shopping.   From describe, we cannot surely say, if there are any outliers. The values look to be scaled at column level. But however to Cluster the records, we might have to scale it using one of the Scalers. |
| **Box plots, Correlation plots. Appropriate plots for categorical variables. Inferences on each plot. Summary stats, Skewness, Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct** |
| **Box Plot and Histogram (Uni Variate Analysis):**   1. **Spending** has a slight right skewed distribution as the median is to the left of the distribution and most of the data is on the right-hand side of the median, it has no outliers. 2. **Advance\_payments**, like Spending, it has a right skewed distribution as the median is to the left of the distribution and most of the data is on the right-hand side of the median, it has no outliers. 3. **probability\_of\_full\_payment** has a slight left skewed distribution (looks almost normal distribution) with majority of it’s data plotted onto the left side of the median, Yes, it has outliers on the negative side of Lower Quartile Q1. It specifies, for some customers chance of full payment is significantly low, could be an always defaulter. 4. **Current\_balance** is right skewed in nature with majority of it’s data available on the right hand side of the median. This feature does not have any outliers. 5. **Credit\_limit** has an almost normal distribution without any outliers, it specifies among the customers, credit limit feature is equally spread. 6. **min\_payment\_amt** is a right skewed feature with few outliers. This outliers are good in a way, it’s specifying, users were paying more than minimum payment amount across the spendings. 7. **Max\_spent\_in\_single\_shopping** is also a right skewed histogram with majority (60) spending of not so significant amount. This amount if it’s higher then it will help business to reach goal of Customer Spending faster.     **Pair Plot (Bi Variate and Multi Variate Analysis)**   1. Spending is strongly correlated to advance\_payments, current\_balance, credit\_limit and max\_spent \_in\_single\_shopping. It’s not correlated to min\_payment\_amt and weakliy correlated to probability\_of\_full\_payment. 2. Advance\_payments is strongly correlated to max\_spent \_in\_single\_shopping, Spending, current\_balance, credit\_limit, and probability\_of\_full\_payment and it’s weakly correlated to min\_payment\_amt. 3. Current\_balance is strongly correlated to max\_spent \_in\_single\_shopping, credit\_limit, advance\_payments and spending and it’s weakly correlated to probability\_of\_full\_payment, min\_payment\_amount. 4. Probability\_of\_full\_payment is not strongly correlated to any features, however it’s weakly correlated to credit\_limit, advance\_payments and spending. 5. Credit\_limit is weakly correlated to max\_spent\_in\_single\_shopping and not at all correlated to min\_payment\_amt, and it’s strongly correlated to current\_balance, probability\_of\_full\_payment, advance\_payments and spending. 6. Min\_payment\_amt is not correlated with any, it does not have any affect on other features, it’s ok to not focus on this feature while Clustering. 7. Max\_spent\_in\_single\_shopping is strongly correlated to spending, advance\_payments and current\_balance, however it’s weakly correlated to probability\_of\_full\_payment, credit\_limit and no correlation with min\_payment\_amt. 8. Min\_payment\_amt has no value in the data set, that could be deducted as a result here.     **Correlation Plot**  In the support of strong correlations from the above diagram, below values will clarify with strong correlation quotient value. For example, spending is strongly correlated with Advance payments, that was the result deduced from above as well. |

**1.2 Do you think scaling is necessary for clustering in this case? Justify The learner is expected to check and comment about the difference in scale of different features on the bases of appropriate measure for example std dev, variance, etc. Should justify whether there is a necessity for scaling and which method is he/she using to do the scaling. Can also comment on how that method works.**

|  |
| --- |
| 1. Yes, scaling is very necessary for the Given data set. 2. Mean for the columns spending and advance\_payments is staggeringly high when compared to other columns. 3. On the other hand, the probability\_of\_full\_payment’s CT values are very low. 4. For the distance-based algorithms where the cluster distance is measured, it’s more effective and faster to form a cluster if we scale all the values on a uniform-scale. 5. Z-Score Scaling:    1. I would use Z-Score scaling as we focus on standardising the column values to a Normal Standard Distribution with the measure of Mean to be 0 and Standard Deviation to be 1. This can be verified once we apply zscore on the data set.    2. Formula and internal working of Z-Score: z = (x-mU)/sigma, this scaling method will make sure, that, there will not be much of variance with high values. |

**1.3 Apply hierarchical clustering to scaled data (3 pts). Identify the number of optimum clusters using Dendrogram and briefly describe them (4). Students are expected to apply hierarchical clustering. It can be obtained via Fclusters or Agglomerative Clustering. Report should talk about the used criterion, affinity and linkage. Report must contain a Dendrogram and a logical reason behind choosing the optimum number of clusters and Inferences on the dendrogram. Customer segmentation can be visualized using limited features or whole data but it should be clear, correct and logical. Use appropriate plots to visualize the clusters.**

|  |
| --- |
| **Fcluster (“ward” and “eucledian”):**   1. “ward” linkage method is a comprehensive one as it calculates the variance increase and proceeds ahead with the merge. “eucledian” distance metric is a rule of thumb to find the distance between the data points. 2. “ward” linkage method was used to find cluster distance, along with “eucledian distance” metric to calculate distance between individual data points. 3. In “ward” linkage method, within-cluster-variance will be calculated between the clusters. Wherever we find less cluster-variance, we can go ahead with the merge. 4. Dendrogram with ward link method and it was truncated at 15 clusters for better readability. 5. Looking at above Dendrogram, I find, the highest distanced vertical cluster and the next highest in distance and make a horizontal cut. Technically, that will let us settle at 2 clusters. 6. Reason to choose 3 clusters: However, for a business it’s always good to have multiple segmentations, having only 2 segments not help the business and as we want to provide promotional offers to our customers based on their credit card usage, so we can go ahead with having 3 clusters. 7. Cluster Frequency: 8. Bar plot of Segments |
| **Agglomerative Clustering (clusters=3, affinity=eucledian, linkage=ward):**   1. As in FCluster we also use “eucledian” and “ward” here and start with 3 clusters to begin with. 2. “ward” linkage method is a comprehensive one as it calculates the variance increase and proceeds ahead with the merge. “eucledian” distance metric is a rule of thumb to find the distance between the data points. 3. “ward” linkage method was used to find cluster distance, along with “eucledian distance” metric to calculate distance between individual data points. 4. In “ward” linkage method, within-cluster-variance will be calculated between the clusters. Wherever we find less cluster-variance, we can go ahead with the merge. 5. Cluster frequency:     2. Cluster frequency is like the F Cluster, but it’s not sorted by the means.    3. Fcluster 1 will be same as Agglo Cluster 1 and FCluster 2 will be same as Agglo Cluster 2 and FCluster 3 will be same as Agglo cluster 0. 6. Bar plot of Segments |

**1.4 Apply K-Means clustering on scaled data and determine optimum clusters (2 pts). Apply elbow curve and silhouette score (3 pts). Interpret the inferences from the model (2.5 pts). K-means clustering code application with different number of clusters. Calculation of WSS(inertia for each value of k) Elbow Method must be applied and visualized with different values of K. Reasoning behind the selection of the optimal value of K must be explained properly. Silhouette Score must be calculated for the same values of K taken above and commented on. Report must contain logical and correct explanations for choosing the optimum clusters using both elbow method and silhouette scores. Append cluster labels obtained from K-means clustering into the original data frame. Customer Segmentation can be visualized using appropriate graphs.**

|  |
| --- |
| **K Means Clustering:**   1. WSS or inertia:    1. WSS score will let us know how many Clusters are ideal for the given data set.    2. For 1-15 clusters below are the WSS Scores    4. Reasoning: I can say that the drop is very significant from 1 to 2 and 2 to 3 is moderate and 3-4 is not so significant, from there on the WSS score is very minimal. We can conclude the major drop at 3 and later that diminishing WSS score is no longer worth to be considered. Top 3 clusters will explain lot of variation in the data set.    5. WSS Plot        2. In the above plot elbow is bent at Clusters 2 and 3, however, at 2 the elbow is significant than 3. We will proceed with 3 clusters as the significance of WSS diminishes post 3 clusters.    6. Silhouette Score:       1. For 2 Clusters, Sil Score stands at 0.47 when rounded to 2 places.       2. For 3 clusters, Sil score stands at 0.4 when rounded to 2 places       3. Overall for Cluster size 2 and 3 the Silhouette Score is ~0.4 which is near to zero, it specifies that the Clusters are not well separated from each other.       4. Minimum of Silhouette Sample for 2 and 3 clusters are 0, i.e. there is a chance the clusters may be overlapping each other and not well separated.    7. Cluster Profiling/Cluster Frequency/Customer Segmentation:          4. Above information helps us to understand, all segments are distributed almost equally. Segment 2 has greater Spendings and Advance Payments and high probability of full payment.       6. Pairplot KMeans\_Clusters s hue will give overall insight on the trend. Can see low, medium and high trend for all clusters and where they stand in the correlations. |

**1.5 Describe cluster profiles for the clusters defined (2.5 pts). Recommend different promotional strategies for different clusters in context to the business problem in-hand (2.5 pts ). After adding the final clusters to the original dataframe, do the cluster profiling. Divide the data in the finalyzed groups and check their means. Explain each of the group briefly. There should be at least 3-4 Recommendations. Recommendations should be easily understandable and business specific, students should not give any technical suggestions. Full marks will only be allotted if the recommendations are correct and business specific. variable means. Students to explain the profiles and suggest a mechanism to approach each cluster. Any logical explanation is acceptable.**

|  |
| --- |
| **Cluster Profiling:**    2. Inferences:    1. Cluster 2:       1. Customers in Cluster 2 have higher Spending ability and they pay their bills in Advance without going to wait for the billing period to lapse, thereby they maintain **good credit score**.       2. 88% customers of this Cluster pay their bills on time, which is ~90%. This factor will credit to the good credit scores of this group.       3. Credit\_limit is high for these customers; it makes the customer more important to the Organization and earns Respect of the customer as we value his spending ability and increase the limit appropriately.       4. Minimum payment due is neither high nor low in comparison to the other clusters. However, we can reduce the minimum payment as majority of the customers already do an advance payment.       5. Recommendations:          1. Better Promotional Offers to be given for Cluster-2 customers as they have higher spending of the lot, followed by Cluster-0 and Cluster-1.          2. Cluster-2’s spending, probability of payment and credit limit is higher compared to others. Cluster-2 is bound to have good credit score, Organization can provide offers on high cost items like iphones, macbook other electronics No-Cost-Emi for longer duration. This will improve customer loyalty and recognition. Offers to repeat the purchase and free subscription offers on Streaming services could be a cheaper way for the organization to gain customer loyalty.          3. Cluster-0’s customer is very loyal to the Organization. His Spending and Advance Payments are good, he pays before time and his probability of payment is also high. Min\_due could also be the reason for customer to pay the bills on time. This ensures the customer to have good credit score. Organization should improve the spending capacity of the customer by increasing the Credit\_limit and Credit\_Balance. Customer can be provided with good offers on high priced items to improve his spending with exchange value, this will make the customer want for new premium products. Improving the credit limit in one way can make the customer to buy to the limit.          4. Cluster-1 customers are good in paying the bills in advance and in excess too, this will increase the cash inflow for the organization, however the spending is less compared to other Clusters. Probability of paying the bills on time is good, however, this group has high minimum due payment, this could be one reason customer is not willing to spend any. We must reduce the min\_payment\_amt and introduce promotional offers with add-on vouchers to have the customer to repeat the purchases. This way, we can ensure his spendings will increase in consistent with other Clusters.          5. Overall, organization can increase the credit limit across the segments and provide credit card offers (like VIP lounge access, Co-Work Access, Payback points) for users with a higher advance payment, in our case Cluster-2. The same will apply to Cluster-1 as well, though the spendings are less, they pay more than the bill amount everytime, this should be rewarded. These steps will make customer feel important as he is proactively paying all due on the credit card. |

**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 and do exploratory data analysis (4 pts). Describe the data briefly. Interpret the inferences for each (2 pts). Initial steps like head() .info(), Data Types, etc . Null value check. Distribution plots(histogram) or similar plots for the continuous columns. Box plots, Correlation plots. Appropriate plots for categorical variables. Inferences on each plot. Summary stats, Skewness, Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct.** |
| **Data Brief:**   1. Df.head(): Returned top 4 rows, straight away I could identify some 0’s for Comission, more to be identified in the describe stage. 2. Df.info(): Out of 10 columns, all are non-null and there are 3000 rows in the data set and 10 columns. In 10 columns, 4 columns are Numeric type and other 6 are Object type. 3. Df.describe():    1. Duration column has a negative minimum, it looks like an anomaly.    3. Insurance Sale count is 0, this needs to be investigated.    4. Commission is 0 for 25% of the data, needs to be investigated.    5. Age feature seems to be fine.    6. No nulls or na values exist for 6 Categorical Columns.    7. Claim percentage is 69% - No and 31% Yes, Minority is Yes. If we can have a better recall score then we will have a better prediction on when the Claim could be Completed Successfully. 4. Nulls and Anomalies:    1. No nulls or na values exist for any feature.    2. 3 Anomalies exist for duration <= 0. A trip cannot have 0 or negative as duration. This anomaly is removed as no of rows is small.    3. Comission is 0 for many records, but it is 0 when there was No Claim, this does not affect the organization and it’s possible in real time, so it’s not considered as an anomaly.    4. 0 Sales are possible from a Travel Agency for a Specific insurance agency, I did not consider it as an anomaly. 5. Visualization:     2. Age feature has a normal distribution with outliers existing on either ends of IQR.    3. Agency\_Code has high value of the agency EPX followed by C2B, CWT and JZI    4. Type feature has major insurance bookings from Travel Agency and some from Airlines.    5. Majority of the insurance Claims were not claimed and minority portion were only Claimed.    6. Commission is a heavily right skewed feature with huge tail of outliers and one extreme outlier.    7. Channel, for the most time the travel agency insurance was procured only Online.    8. Duration is a right skewed feature with a tail of outliers and with extreme outliers.    9. Product Name, customized plan is the one which was Insurer able to sell in more quantity.    10. Most travelled and insured destination was Asia.     13. Age feature is not correlated with any features.    14. Commission and Sales are mutually Correlated to each other and not with any other features.    16. Commission and Sales are only positively correlated and next comes the commission and duration. |
| **2.2 Data Split: Split the data into test and train(1 pts), build classification model CART (1.5 pts), Random Forest (1.5 pts), Artificial Neural Network(1.5 pts). Object data should be converted into categorical/numerical data to fit in the models. (pd.categorical().codes(), pd.get\_dummies(drop\_first=True)) Data split, ratio defined for the split, train-test split should be discussed. Any reasonable split is acceptable. Use of random state is mandatory. Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Apply grid search for each model and make models on best\_params. Feature importance for each model.** |
| 1. 6 Categorical Features [Agency\_Code, Type, Claimed, Channel, Product Name and Destination] were converted into Numeric Type. 3. Proportions of Claimed status 5. Data Split: Data split ratio was set to 30%, i.e. the training data will be the 70% and the testing data will be the other 30% , x\_train is the training part of 70% and x\_test is the testing part of 70%, y\_train is the training part of 30% and y\_test is the testing part of 30%, 6. Best Model Comparison:     2. For Cart, depth of trees were limited to 10 as the webgraphviz denoted it’s good to prune till 10 levels. min\_samples\_leaf and min\_sample\_split were set to 50 and 150 as a rule of thumb, one is 3 times the other.     5. For RandomForestClassifier max depth ranges between 20 and 40, this model will have a much deeper range than CART and feature split was chosen between 4 and 6 as we have 10 features. Estimators were set to 300.    7. Hidden layers were 600 and max iterations were 5000 and stochiastic gradient descent was used as a solver and a tolerance of 0.001 was applied. 7. Random\_state is set to 1 randomness. |
| **2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy (1 pts), Confusion Matrix (2 pts), Plot ROC curve and get ROC\_AUC score for each model (2 pts), Make classification reports for each model. Write inferences on each model (2 pts). Calculate Train and Test Accuracies for each model. Comment on the validness of models (overfitting or underfitting) Build confusion matrix for each model. Comment on the positive class in hand. Must clearly show obs/pred in row/col Plot roc\_curve for each model. Calculate roc\_auc\_score for each model. Comment on the above calculated scores and plots. Build classification reports for each model. Comment on f1 score, precision and recall, which one is important here.** |
| 1. GridSearchCV for RandomForestClassifier Training Data: 2. Grid Search CV for RandomForest Testing Data 3. Above models are made with the best params of Random Forest Classifier, The training model and test model give encouraging results with ROC\_AUC Accuracy Score ~80. 4. However the recall score is not impressive with a chance to have Type2 Error i.e. False Negative is 50%. 5. Precision values to predict Claim’s No status is impressive with 81% percent, but the Claim Yes. This specified a low Type 1 error. 6. The Predict test accuracy i.e. F1 Score stands at 85% for majority i.e. Claim-No and 58% for Claim-Yes. Recall-Score should be strongest point of the model. 7. Summary: Model is strong enough to predict Claimed-No status for the given data, but Claimed-Yes is only 60% confident to predict correctly. 8. This model is underfitted, more parameter tuning is required, I was able to increase performance, but due to computing constraints couldn’t proceed. 9. GridSearchCV for Neural Neworks Training Data: 11. GridSearchCV for Neural Neworks Testing Data: 13. This model is underfitted, more parameter tuning is required, I was able to increase performance, but due to computing constraints couldn’t proceed. 14. Above 2 models are for RandomForestClassifier, the model above is not completely tuned, due to **computer performance limitations** I am not able to put forward a best model. 15. Training and testing models are near to each other. 16. Models have Accuracy of ~77% Area under the curve. 17. Claim-Yes recall score is not good with only 46% prediction rate of Claim-Yes. 18. Claim-No score prediction is very good, model can be used to predict Claim-No. 19. F1-Score stands at 54% which neither good nor bad performance, so we cannot surely tell if the test results will be accurate. 0 is the worst model performance and 1 to be the best model performance. 20. Recall has a sound importance! Reason, the strength of the predict model has to be over the minority proportion of the Target Variable in our case Claimed-Yes. None of our models which I made were successful in getting a better score. Having more hidden layers and increasing the tolerance might improve learning rate and could provide with a better recall score, but due to my laptop constraints I couldn’t do it. |
| **2.4 Final Model - Compare all models on the basis of the performance metrics in a structured tabular manner (2.5 pts). Describe on which model is best/optimized (1.5 pts ). A table containing all the values of accuracies, precision, recall, auc\_roc\_score, f1 score. Comparison between the different models(final) on the basis of above table values. After comparison which model suits the best for the problem in hand on the basis of different measures. Comment on the final model.**     1. **Training Data:** 2. **Testing Data:**   **Summary:** From the above, conclusion is, it’s derived that RandomForest produced a better AUC and Recall Score. However, it’s possible to tune NeuralNetworks to produce a better result. |
| **2.5 Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective. There should be at least 3-4 Recommendations and insights in total. Recommendations should be easily understandable and business specific, students should not give any technical suggestions. Full marks should only be allotted if the recommendations are correct and business specific.** |
| **Recommendations:**   1. **For 25% of the data Commission was not collected from Customer, this is a major leak of income. Should collect at least a minimum.** 2. **As the duration increases the Commission collected also should increase, the data does not speak the same. Organization can use this to improve on income.** 3. **From the model standpoint, Organization is facing Claim Frequency, but most of the claims are being rejected. Claim-yes must increase to at least 40-45% to gain customer vendor interest to apply travel insurance from the company. Having a high No rate, does not impress airlines or agents to subscribe customers to Organizations travel insurance.** 4. **Commission and Duration should positively corelate and in this way more commission can be charged for more tour duration, current data doesn’t specify it, correlation matrix will tell about it.** |