**Question 1.1**

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

Distribution:

Except for probability\_of\_full\_payment which is somewhat normally distributed none of the other variables have normal distribution.

Outliers:

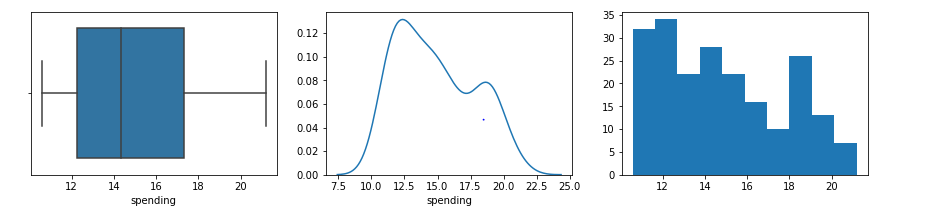
Except for probability\_of\_full\_payment and min\_payment\_amt which has few outliers none of the other variables have outliers.

Please find below the plots that depict above observations .

1. Univariate analysis for spending

Mean is 14.847524, Median is 14.355000, Mode(s) are 11.2300

Column spending does not have outliers

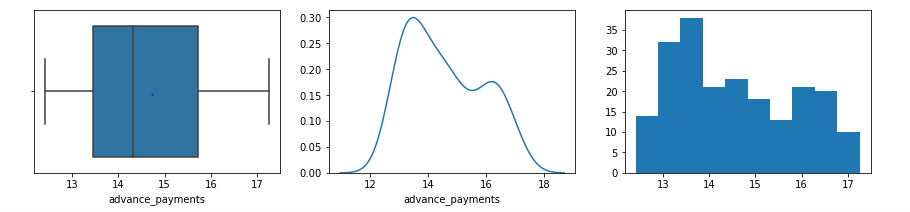


Column spending is not normally distributed

2. Univariate analysis for advance\_payments

Mean is 14.559286, Median is 14.320000, Mode(s) are 13.4700

Column advance\_payments does not have outliers

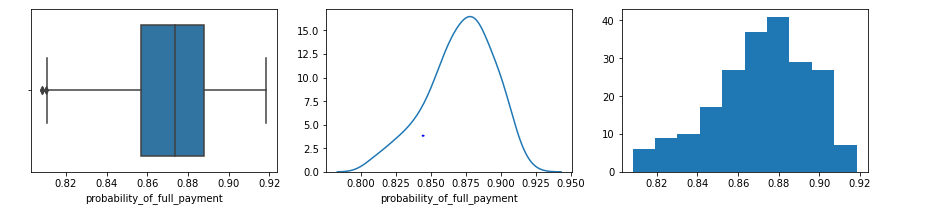


Column advance\_payments is not normally distributed

3. Univariate analysis for probability\_of\_full\_payment

Mean is 0.870999, Median is 0.873450, Mode(s) are 0.8823

Column probability\_of\_full\_payment has outliers

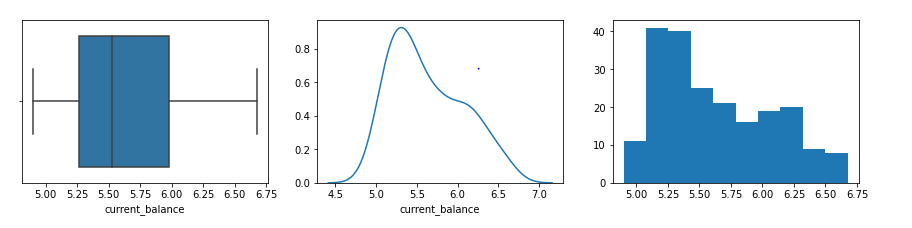


Column probability\_of\_full\_payment is not normally distributed

4. Univariate analysis for current\_balance

Mean is 5.628533, Median is 5.523500, Mode(s) are 5.2360

Column current\_balance does not have outliers

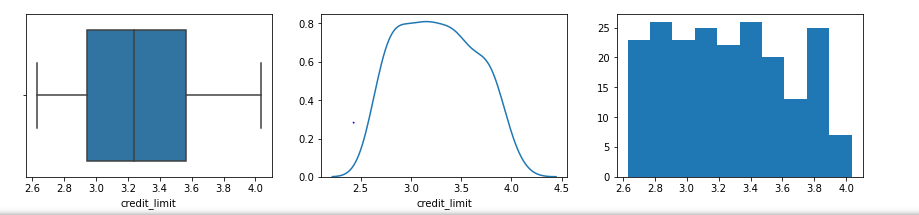


Column current\_balance is not normally distributed

5. Univariate analysis for credit\_limit

Mean is 3.258605, Median is 3.237000, Mode(s) are 3.0260

Column credit\_limit does not have outliers

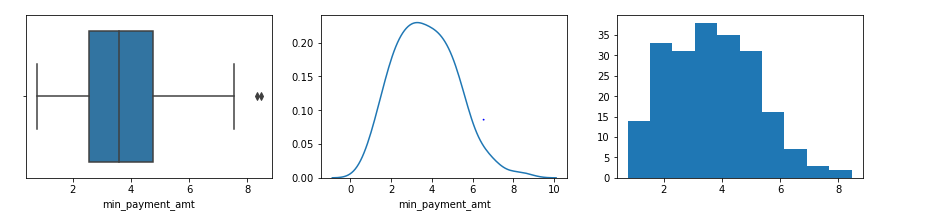


Column credit\_limit is not normally distributed

6. Univariate analysis for min\_payment\_amt

Mean is 3.700201, Median is 3.599000, Mode(s) are 2.1290

Column min\_payment\_amt has outliers

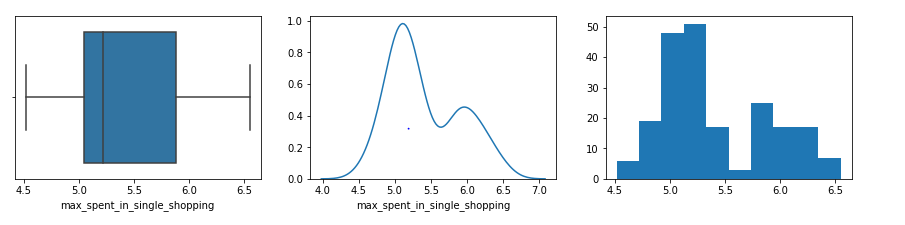


Column min\_payment\_amt is not normally distributed

7. Univariate analysis for max\_spent\_in\_single\_shopping

Mean is 5.408071, Median is 5.223000, Mode(s) are 5.0010

Column max\_spent\_in\_single\_shopping does not have outliers



Column max\_spent\_in\_single\_shopping is not normally distributed

Correlation:

Positive correlations listed incrementally without overlaps among them as below:

* Spending has high positive correlation with max\_spent\_in\_single\_shopping, credit\_limit, current\_balance and advance\_payments
* Advance\_payments has high positive correlation with max\_spent\_in\_single\_shopping, credit\_limit, current\_balance
* Probability\_of\_full\_payment has positive correlation with credit\_limit
* Current\_balance has high positive correlation with max\_spent\_in\_single\_shopping, credit\_limit
* Credit\_limit has positive correlation with max\_spent\_in\_single\_shopping

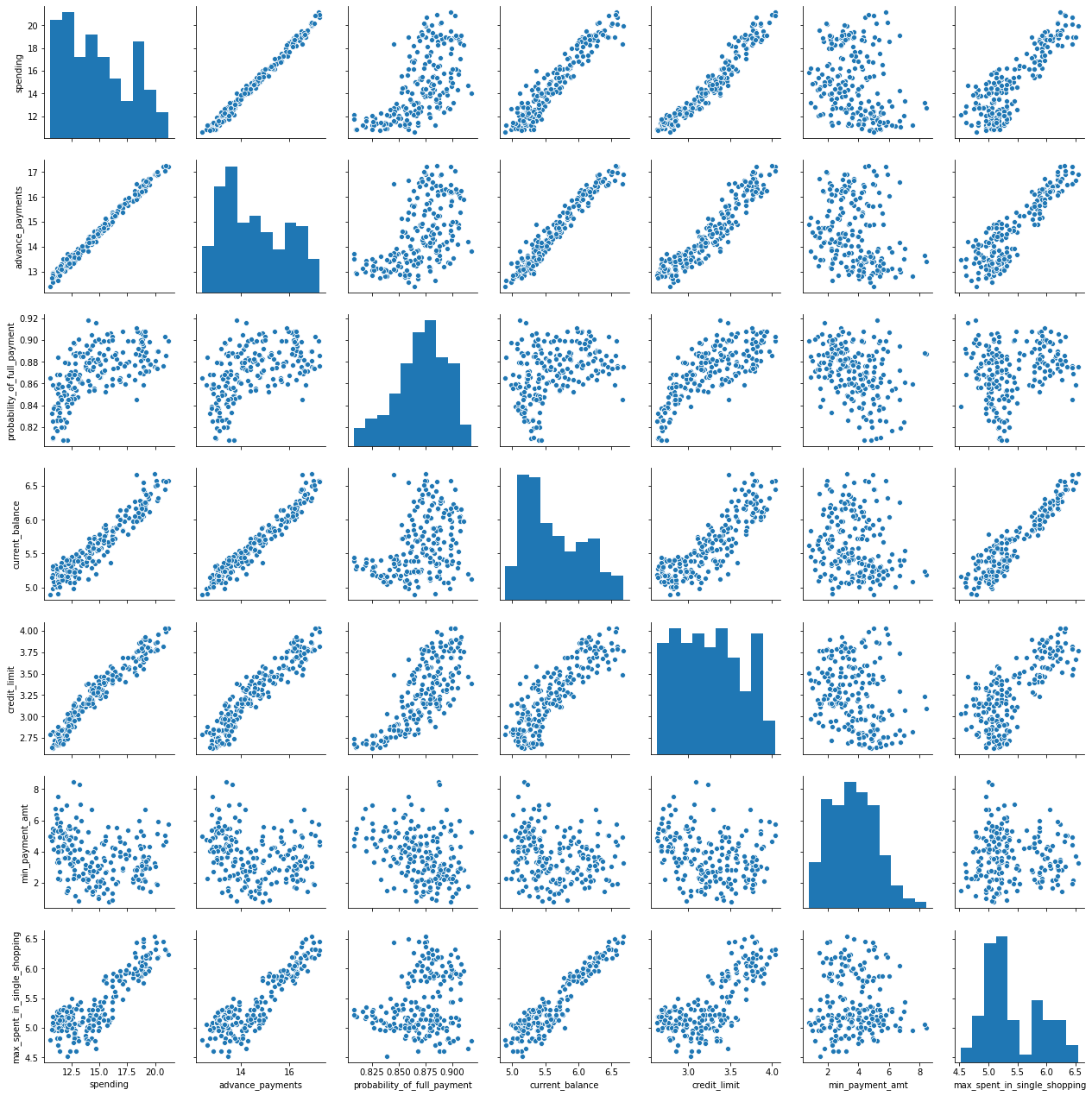
Negative correlation

* No negative correlation across variables has been observed.

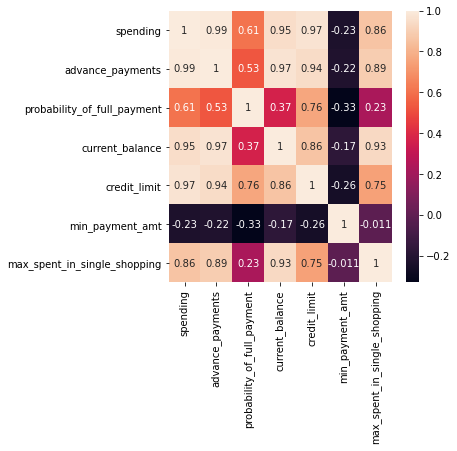
Note: Min\_payment\_amt does not show noticeable correlation with other attributes.

Please find below the pair plot and the related heat map depicting the above observations.

Pair plot:



Heat map:



**Question 1.2**

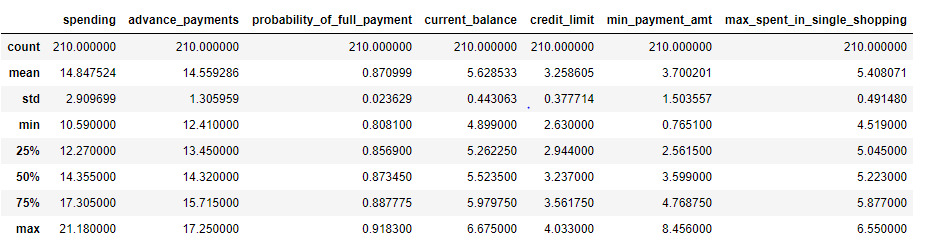
Do you think scaling is necessary for clustering in this case? Justify

Yes, scaling is necessary for clustering.

Scaling the dataset is one of the important data pre processing steps required for hierarchical clustering especially because of its distance based calculation approach during the processing of clustering the data.

Looking at the mean values across the variables in the given dataset, it is observed that there is a difference in magnitude of data across different columns. Hence given that there isn’t significant variance among the data we could go for scaling the data through standardization.

Please find below the summary of statistics for the dataset depicting the mean and standard deviation (a derivative of variance) for the dataset.



**Question 1.3**

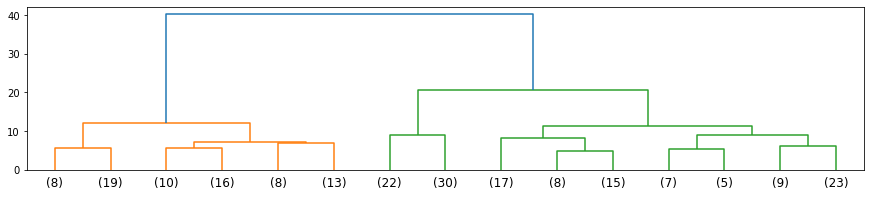
Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them

Below is the dendrogram of the scaled dataset for top 15 clusters following ward linkage method. The dendrogram is depicting the hierarchical clustering outcome based on the agglomerative clustering approach.

However based on the below image we can clearly identify that the data makes two differentiated clusters as suggested by orange and green colours.

We could have pretty much created this dendrogram for top 10 clusters instead. However elongating it further to 15 clusters reassures the inference that the datasets remains clearly divided into two clusters only despite increasing the depth.

Hence 2 clusters will be optimal for the given dataset.



Please find below the 2 clusters created with the respective mean values across all the attributes.



Total count of records across clusters:

Cluster 1: 70 records

Cluster 2: 140 records.

**Question 1.4**

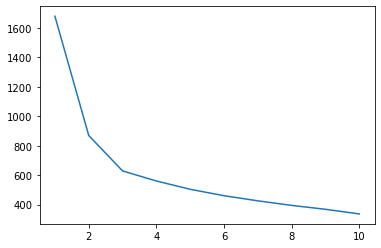
Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score.

By definition, K-means as name suggests refers to averaging the data which in other words is about finding the relevant centroid across K clusters by iterating the clustering technique for the given data set until we could end up with K clusters whose constituent data(s) are at the closest distance to their centroids when compared to other centroids adopting eucledian distance approach.

Determining the optimal cluster count: Once the clusters are created the model can be evaluated using WSS plot. WSS plot/distortion plot helps to know how many clusters are needed as output in K-means clustering. WSS plot stands for within sum of squares plot. also known as distortion plot or error plot. This operates with the notion that within sum of squares (variance) will be high for K=1 but when k begins to increase the sum of within sum of squares across clusters will start reducing from K=1. However as number of clusters increases (i.e incrementing K ) there could be a point where the drop of WSS may not be significant.

Please find below the WSS plot for the given dataset for 1 to 10 clusters. The elbow suggests that the recommended number of plots could mostly be 2 and may be 3 clusters too. Exact clusters can be further recommended based on model evaluation subsequently as below.

WSS plot:



WSS (Within sum of squares) for K=1 to 10 as below



Model evaluation:

Hence to narrow down the most optimal number of clusters we could use techniques to validate if the mapping of observations to the clusters are valid or not. Silhouette scoring technique is one such. This includes computation of distance between each observations i.e distance between itself and every established centroid (i.e Silhouette width) to determine if there are any other centroids outside current cluster whose distance could be the shortest for the given observation. The average of such silhouette width will be referred to as silhouette score whose values would be within the range of -1 to +1. A positive score that yield a score closest to +1 will qualify the optimal number of clusters suggesting that the clusters are well seperated.

Please find below the silhouette scores for various clusters.



Above suggests creating 2 clusters on the given datasets will be the optimal one as it has the highest silhouette score suggesting well separated clusters making clear distinction them for the data it carries.

Please find below the clusters created using K means clustering (2 clusters) with respective mean values across all the attributes.



**Question 1.5**

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

Cluster 1 above indicates customers in high spending category end up paying in full with high probability. This segment also suggests customers probably are maximizing the card utilization given that the current balance is comparatively lesser while making their payment in full by month end.

Recommendation 1: So bank can consider offering balance transfer for these customers at a competitive interest rate to improve interest income as well as with the reduced risk of bad debts from this customer group who is expected to pay on time while having the spending ability.

Cluster 0 suggests that these customers probably are conservative spenders while enjoying staggered payments resulting in interest income to the bank.

Recommendation 2: So bank can consider offering 0% interest rate for new purchases towards increasing their spending ability while being able to increase the market share in terms of volume of transaction. This would also fetch transaction based commission income for the bank. However the risk of defaulting customers needs to be analyzed.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Question 2.1**  Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an  inference on it.  Please note that the object type categorical variables such as Agency\_Code, Type, Claimed, Channel,  Product\_Name and Destination needs to be converted to integer values as a part of data pre processing and the below descriptive statistics has been performed on the continuous variables in the dataset such as Age, Commision, Duration and the Sales.  Distribution:  Following are the continuous variables in the dataset: Age, Commision, Duration and Sales.  None of the variables are normally distributed while age is somewhat observed to be closer to normal  distribution.  Outliers:  Except Duration which has minimal outlier all other variables namely Age, Commision and Sales have  good amount of outliers. All the outliers are beyond the maximum whisker.  Null Check:  There are no nulls across all the columns.  PFB the five number summary;   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **Age** | **Commision** | **Duration** | **Sales** | | **count** | 3000 | 3000 | 3000 | 3000 | | **mean** | 38.091 | 14.529203 | 70.001333 | 60.24991 | | **std** | 10.463518 | 25.481455 | 134.05331 | 70.73395 | | **min** | 8 | 0 | -1 | 0 | | **25%** | 32 | 0 | 11 | 20 | | **50%** | 36 | 4.63 | 26.5 | 33 | | **75%** | 42 | 17.235 | 63 | 69 | | **max** | 84 | 210.21 | 4580 | 539 |   The above summary suggests that across all the variables outliers are significantly affecting the mean  given the extent of difference between median and the max especially the Duration column.  PFB plots from univariate analysis.  1. Univariate analysis for Age  Mean is 38.091000, Median is 36.000000, Mode(s) are 36.0000  Column Age has outliers    Column Age is not normally distributed  2. Univariate analysis for Commision  Mean is 14.529203, Median is 4.630000, Mode(s) are 0.0000  Column Commision has outliers    Column Commision is not normally distributed  3. Univariate analysis for Duration  Mean is 70.001333, Median is 26.500000, Mode(s) are 8.0000  Column Duration has outliers    Column Duration is not normally distributed  4. Univariate analysis for Sales  Mean is 60.249913, Median is 33.000000, Mode(s) are 20.0000  Column Sales has outliers    Column Sales is not normally distributed.  Pair plot as below:  In [70]:    Heat map:    Above pair plot and heatmap suggests that Commision is highly corelated with Sales.  The pair plot also suggests that some of the maximum sales and commission tend to decline along with  age of the insured.  Apart from above statistics we also have following profiling of the categorical columns.  Value counts as below for each Feature:  Feature: Agency\_Code  EPX :1365  C2B :924  CWT :472  JZI :239  Feature: Type  Travel Agency :1837  Airlines :1163  Feature: Claimed  No :2076  Yes :924  Feature: Channel  Online :2954  Offline :46  Feature: Product Name  Customised Plan :1136  Cancellation Plan :678  Bronze Plan :650  Silver Plan :427  Gold Plan :109  Feature: Destination  ASIA :2465  Americas :320  EUROPE :215 |  |
| **Question 2.2**  Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial  Neural Network  Model building considerations:  Split the dataset into independent and dependent variables.  Following would be the independent variables on which we will be building all the 3 models (i.e CART,  Random Forest, Artificial Neural Network)   * Age, * Agency\_Code * Type, * Commision, * Channel, * Duration, * Sales, * Product Name, * Destination   Following would be the dependent/target variable.   * Claimed   All of the categorical/object type independent variables needs to be converted to numeric for any of the  model building and hence following variables to be converted with numeric encoding to replace the textual  values accordingly.   * Agency\_Code * Type * Channel * Product Name * Destination   Neural networks perform better with scaled data and hence scaling to be applied for neural networks model  building.  Divide the data into training and test data with 70:30 ratio. Follow same random state across all 3 models.  PFB the summary on the training data:    PFB the summary on the test data:    Parameters to be considered for each model:  CART:  Splitting criteria: Gini (Decision tree uses gini index and gini gain to pick the best independent variable to  split the data at each node)  Based on the decision tree visualization built using export\_graphviz API it could be narrowed down that  CART can be built with following parameters.  maximum depth : 5 - This parameter controls the depth of hierarchy from the root node at an optimal point  beyond which it could result in over fitment of the model while training the data.  max\_sample\_leaf : 60 – This parameter control minimum observations require to create a new terminal or  decision node.  max\_samples\_split : 180 – This parameter controls minimum observations from the dataset required to split a  particular decision node/parent node into further nodes.  Based on the model build below are the feature importances:   |  |  | | --- | --- | | **Feature** | **Importance** | | Duration | 0.240804 | | Sales | 0.223446 | | Agency\_Code | 0.203293 | | Age | 0.182107 | | Commision | 0.07257 | | Product Name | 0.043855 | | Destination | 0.023909 | | Channel | 0.007948 | | Type | 0.002067 |   PFB the prediction on the training data set. Listed below are the counts by target class.    PFB the prediction on the test data set. Listed below are the counts by target class.    Random Forest:  Based on the grid search below are the best parameters on which this model is built.  max\_depth : 7 – This parameter controls the depth of hierarchy from  the root node at an optimal point beyond which it could result in over  fitment of the model while training the data.  max\_features : 4 – This feature which is part of bootstrapping in  ensemble technique. Bootstrapping or bagging enables cross validation technique  to ensure we train the model for the unseen data.  min\_samples\_leaf : 10 – This parameter control minimum observations require  to create a new terminal or decision node.  min\_samples\_split : 30 – This parameter controls minimum observations from  the dataset required to split a particular decision node/parent node into  further nodes.  n\_estimators : 201 – This parameter controls how many decision trees  needs to be created as a part of random forest model.  Note: Feature importances for Random forest matches with the CART model.  PFB the prediction on the training data set. Listed below are the counts by target class.    PFB the prediction on the test data set. Listed below are the counts by target class.    Neural network:  In order for neural network to perform better the data set can be normalized.  Based on the various trial and error approach consider some of the observations from grid search inclusive  below were the most optimal hyper parameters considered for this model build.   * hidden\_layer\_sizes : (201,101,51) – Number of neurons in a single hidden layer. Please note   three hidden layer considered in this model build post optimization exercise.   * max\_iter :15 - number of epocs end to end across the layers between input and output.   Also the iterations stopped even before reaching 15 at 13th iteration as the training loss reached  global minima.   * activation :'relu' - It's a Mechanism by which the artificial neuron processes incoming   information and passes it throughout the network.   * solver :'adam' – This parameter specifies the loss function used towards updating the   weights during the iterations. This enables to progress along the parabolic curve by deriving error  values optimally to reach the point of global minima.   * tol :0.1 – This parameter specifies tolerance for optimization which means that   when the loss or score is not improving by at least the specified measure model shall be considered  to be converged and hence the iterations can be stopped.  PFB the prediction on the training data set. Listed below are the counts by target class.    PFB the prediction on the testing data set. Listed below are the counts by target class. |  |
| **Question 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  Model performance comparison:  AUC:   |  |  |  |  | | --- | --- | --- | --- | |  | CART | Random Forest | Neural network | | AUC Training data | 0.837 | 0.868 | 0.82 | | AUC Testing data | 0.801 | 0.821 | 0.80 |   ROC:   |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  |  |  |  |  |  |  |  |  |  |  | | Confusion matrix across all 3 models for training and testing data as below side by side: |  |  |  |  |  |  |  |  |  |  |   Classification report across all models for training and testing data depicted below in next few tables.      Out of all the 3 models Random forest has resulted in better accuracy but other two models have come  closer to that with minor difference falling below the Random forest model. It can also be noted that given  both classes are equally important apart from accuracy even the weighted average has performed marginally  better in the random forest and hence we could narrow down on this model for productionizing the model. |  |
| **Question 2.4**  Final Model: Compare all the model and write an inference which model is best/optimized.  Out of all the 3 models Random forest has resulted in better accuracy but other two models have come  closer to that with minor difference falling below the Random forest model. It can also be noted that given  both classes are equally important apart from accuracy even the weighted average has performed marginally  better in the random forest and hence we could narrow down on this model for productionizing the model.  Why go by accuracy?  Given the fact that predicting both classes of customer's who will end up claiming and those who will not  claim correctly are important towards being able to come up with appropriate recommendation, accuracy  becomes an important factor here. This way the model that best controls both type 1 and type 2 error  would enable efficient overall classification from prediction perspective.  Accordingly, Random forest model fairs better comparatively here with 78% accuracy on the test data. |  |
| **Question 2.5**  Inference: Basis on these predictions, what are the business insights and recommendations  Based on the feature importances as below it is observed that “Duration” plays the highest role among the  variables along with other variables such as Sales, Agency Code and the Age. Hence the  classification shall be analyzed towards the profit and loss of current packages accordingly to restrategize it  for maximum benefit for the customers and insurance firm.   |  |  | | --- | --- | | **Feature** | **Importance** | | Duration | 0.240804 | | Sales | 0.223446 | | Agency\_Code | 0.203293 | | Age | 0.182107 | | Commision | 0.07257 | | Product Name | 0.043855 | | Destination | 0.023909 | | Channel | 0.007948 | | Type | 0.002067 |   Exploring the data further based on the Random forest outcome, top two Duration that yields to higher claim  Include 141 and 2. Please find below the part of the count table in descending order.    Firm can evaluate the kind of service providers for the tour or the nature of the tour to assess these  segments to re evaluate the clauses for underwriting or revise the package configuration to make it an  reasonably affordable package for a win win situation with the customer. If required some of  these segments such as Duration 2 which has almost 83% of the insurance being claimed may not be  a sustainable model at all and may need a revisit to reconfigure the plan to minimize risk while retaining  benefit proposition to the customer. |  |