A picture containing logo

Description automatically generated

**Fundamental of Data Science**

**Assignment THREE**

Reyad Melies

Part A: Clustering

1. K-Means Clustering

Using only the Sex and Age fields



ensure you standardize Age.



a) Perform k‐means clustering on the selected attributes, specifying k = 4 clusters and plot.

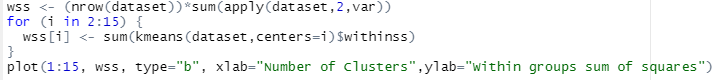
Text

Description automatically generated with medium confidence

Chart

Description automatically generated

b) Apply the elbow method to determine the best k and plot



A picture containing histogram

Description automatically generated

Best K is equal to 4 same as the figure above

c) Evaluate the quality of the clusters using the Silhouette Coefficient method.

Diagram

Description automatically generated

Text

Description automatically generated

For 4 clusters silhouette score: 0.4387456

hierarchical agglomerative clustering with single linkage:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 10 | 20 | 40 | 80 | 85 | 121 | 160 | 168 | 195 |
| 10 | 0 |  |  |  |  |  |  |  |  |
| 20 | 10 | 0 |  |  |  |  |  |  |  |
| 40 | 30 | 20 | 0 |  |  |  |  |  |  |
| 80 | 70 | 60 | 40 | 0 |  |  |  |  |  |
| 85 | 75 | 65 | 45 | 5 | 0 |  |  |  |  |
| 121 | 111 | 101 | 81 | 41 | 36 | 0 |  |  |  |
| 160 | 150 | 140 | 120 | 80 | 75 | 39 | 0 |  |  |
| 168 | 158 | 148 | 128 | 88 | 83 | 47 | 8 | 0 |  |
| 195 | 185 | 175 | 155 | 115 | 110 | 74 | 35 | 27 | 0 |

Chart

Description automatically generated

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 10 | 20 | 40 | 80-85 | 121 | 160 | 168 | 195 |
| 10 | 0 |  |  |  |  |  |  |  |
| 20 | 10 | 0 |  |  |  |  |  |  |
| 40 | 30 | 20 | 0 |  |  |  |  |  |
| 80-85 | 70 | 60 | 40 | 0 |  |  |  |  |
| 121 | 111 | 101 | 81 | 36 | 0 |  |  |  |
| 160 | 150 | 140 | 120 | 75 | 39 | 0 |  |  |
| 168 | 158 | 148 | 128 | 83 | 47 | 8 | 0 |  |
| 195 | 185 | 175 | 155 | 110 | 74 | 35 | 27 | 0 |

Diagram

Description automatically generated

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 10 | 20 | 40 | 80-85 | 121 | 160-168 | 195 |
| 10 | 0 |  |  |  |  |  |  |
| 20 | 10 | 0 |  |  |  |  |  |
| 40 | 30 | 20 | 0 |  |  |  |  |
| 80-85 | 70 | 60 | 40 | 0 |  |  |  |
| 121 | 111 | 101 | 81 | 36 | 0 |  |  |
| 168-160 | 158 | 148 | 128 | 83 | 47 | 8 | 0 |
| 195 | 185 | 175 | 155 | 110 | 74 | 27 | 0 |

A picture containing text, whiteboard

Description automatically generated

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 10-20-40 | 80-85 | 121 | 160-168 | 195 |
| 10-20-40 | 0 |  |  |  |  |
| 80-85 | 40 | 0 |  |  |  |
| 121 | 81 | 36 | 0 |  |  |
| 160-168 | 120 | 75 | 39 | 0 |  |
| 195 | 155 | 110 | 74 | 27 | 0 |

Chart, line chart

Description automatically generated

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 10-20-40 | 80-85 | 121 | 160-168-195 |
| 10-20-40 | 0 |  |  |  |
| 80-85 | 40 | 0 |  |  |
| 121 | 81 | 36 | 0 |  |
| 160-168-195 | 120 | 75 | 39 | 0 |

Chart

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
|  | 10-20-40 | 121-80-85 | 168-168-195 |
| 10-20-40 | 0 |  |  |
| 80-121-85 | 40 | 0 |  |
| 160-168-195 | 120 | 39 | 0 |

Chart, line chart

Description automatically generated

|  |  |  |
| --- | --- | --- |
|  | 10-20-40 | 160-168-195-121-80-85 |
| 10-20-40 | 0 |  |
| 160-168-195-121-80-85 | 40 | 0 |

Chart, line chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

5 8 10 20 27 36 39 40

hierarchical agglomerative clustering with complete linkage

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 10 | 20 | 40 | 80 | 85 | 121 | 160 | 168 | 195 |
| 10 | 0 |  |  |  |  |  |  |  |  |
| 20 | 10 | 0 |  |  |  |  |  |  |  |
| 40 | 30 | 20 | 0 |  |  |  |  |  |  |
| 80 | 70 | 60 | 40 | 0 |  |  |  |  |  |
| 85 | 75 | 65 | 45 | 5 | 0 |  |  |  |  |
| 121 | 111 | 101 | 81 | 41 | 36 | 0 |  |  |  |
| 160 | 150 | 140 | 120 | 80 | 75 | 39 | 0 |  |  |
| 168 | 158 | 148 | 128 | 88 | 83 | 47 | 8 | 0 |  |
| 195 | 185 | 175 | 155 | 115 | 110 | 74 | 35 | 27 | 0 |

Chart

Description automatically generated

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 10 | 20 | 40 | 80-85 | 121 | 160 | 168 | 195 |
| 10 | 0 |  |  |  |  |  |  |  |
| 20 | 10 | 0 |  |  |  |  |  |  |
| 40 | 30 | 20 | 0 |  |  |  |  |  |
| 80-85 | 75 | 65 | 45 | 0 |  |  |  |  |
| 121 | 111 | 101 | 81 | 41 | 0 |  |  |  |
| 160 | 150 | 140 | 120 | 80 | 39 | 0 |  |  |
| 168 | 158 | 148 | 128 | 88 | 47 | 8 | 0 |  |
| 195 | 185 | 175 | 155 | 115 | 74 | 35 | 27 | 0 |

Diagram

Description automatically generated

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 10 | 20 | 40 | 80-85 | 121 | 160-168 | 195 |
| 10 | 0 |  |  |  |  |  |  |
| 20 | 10 | 0 |  |  |  |  |  |
| 40 | 30 | 20 | 0 |  |  |  |  |
| 80-85 | 75 | 65 | 45 | 0 |  |  |  |
| 121 | 111 | 101 | 81 | 41 | 0 |  |  |
| 160-168 | 158 | 148 | 128 | 88 | 47 | 0 |  |
| 195 | 185 | 175 | 155 | 115 | 74 | 35 | 0 |

A picture containing text, whiteboard

Description automatically generated

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 10-20 | 40 | 80-85 | 121 | 160-180 | 195 |
| 10-20 | 0 |  |  |  |  |  |
| 40 | 30 | 0 |  |  |  |  |
| 80-85 | 75 | 45 | 0 |  |  |  |
| 121 | 111 | 81 | 41 | 0 |  |  |
| 160-168 | 158 | 128 | 88 | 47 | 0 |  |
| 195 | 185 | 155 | 115 | 74 | 35 | 0 |

A picture containing text, whiteboard

Description automatically generated

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 10-20-40 | 80-85 | 121 | 160-180 | 195 |
| 10-20-40 | 0 |  |  |  |  |
| 80-85 | 75 | 0 |  |  |  |
| 121 | 111 | 41 | 0 |  |  |
| 160-168 | 158 | 88 | 47 | 0 |  |
| 195 | 185 | 115 | 74 | 35 | 0 |

A picture containing text, whiteboard

Description automatically generated

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 10-20-40 | 80-85 | 121 | 160-168-195 |
| 10-20-40 | 0 |  |  |  |
| 80-85 | 75 | 0 |  |  |
| 121 | 111 | 41 | 0 |  |
| 160-168-195 | 185 | 115 | 74 | 0 |

Chart, histogram

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
|  | 10-20-40 | 80-85-121 | 160-168-195 |
| 10-20-40 | 0 |  |  |
| 80-85-121 | 111 | 0 |  |
| 160-168-195 | 185 | 115 | 0 |

Chart

Description automatically generated

|  |  |  |
| --- | --- | --- |
|  | 10-20-40-80-85-121 | 160-168-195 |
| 10-20-40-80-85-121 | 0 |  |
| 160-168-195 | 185 | 0 |

Chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

5-8-10-30-35-41-111-185

**Part B: Model Evaluation & Performance Improvement**

1. Partition the data set using the holdout method, so that 67% of the records are included in the training data set and 33% are included in the test data set. Use a bar graph to confirm your proportions.

Graphical user interface, text, application, email

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

1. Identify the total number of records in the training data set and how many records in the training data set have a churn value of true (or 1). Calculate how many true churn records you need to resample to have 30% of the rebalanced data set have true churn values. (Changed to 30 % as the announcemnet )

Text

Description automatically generated

Train records: 4713 records

Training data set have a churn value of true: 1253 records.

true churn records you need to resample to have 30%: 161 records.

1. Perform the rebalancing described in (b) and confirm that 30% of the records in the rebalanced data set have true churn values. (Changed to 30 % as the announcemnet )

Text

Description automatically generated with medium confidence

1. Use predictors you think are appropriate and obtain the predicted value.

Create a temp tree and use feature importance.

Text

Description automatically generated

Chart, histogram

Description automatically generatedA picture containing text, receipt

Description automatically generated

And dropped Id as it does not help the model.

Text

Description automatically generated

Create a decision tree model that can predict Churn using the data set given.

Text

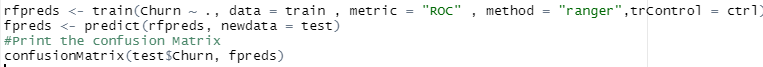
Description automatically generated

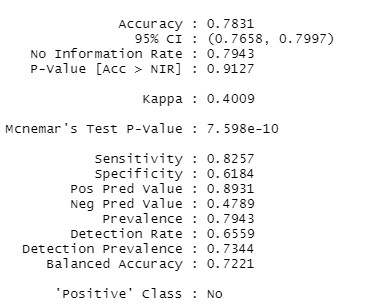
Diagram

Description automatically generated

A picture containing text, receipt

Description automatically generated

1. Use an ensemble method (e.g., Random Forest, Adaboost) to obtain the predicted value of Churn.



1. Tune the hyper-parameters (e.g., node size, max depth, max terminal nodes, etc.) of the ensemble model and compare against the initial model.

Text, letter

Description automatically generated

A picture containing text, receipt, screenshot

Description automatically generated

1. Using a confusion matrix, compare the evaluation measures from the ensemble method with the decision tree model based on the following criteria: Accuracy, Sensitivity and Specificity. Identify the model that performed best and worst according to each criterion.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Accuracy | Sensitivity | Specificity |
| decision tree | 0.721 | 0.7672 | 0.4548 |
| Random forest | 0.721 | 0.7672 | 0.4548 |
| Tuned Random Forest  Best(min node size=2 &mtry=6) | 0.7947 | 0.8369 | 0.6406 |

Model with best accuaracy : Tuned Random Forest Best(min node size=2 &mtry=6)

Model with worst accuaracy : decision tree and are equal.

Model with best Sensitivity: Tuned Random Forest Best(min node size=2 &mtry=6)

Model with worst Sensitivity: decision tree and are equal

Model with best Specificity: Tuned Random Forest Best(min node size=2 &mtry=6)

Model with worst Specificity: decision tree and are equal

g) Carry out a ROC analysis to compare the performance of the ensemble method with the decision tree technique. Plot the ROC graph of the models.

Text

Description automatically generated

Chart

Description automatically generated

Best model is Tuned Random Forest

Best(min node size=2 &mtry=6) then normal random forest and then decision tree