CAB330: Data and Web Analytics

Assessment 1

Team Name: [HDZ]

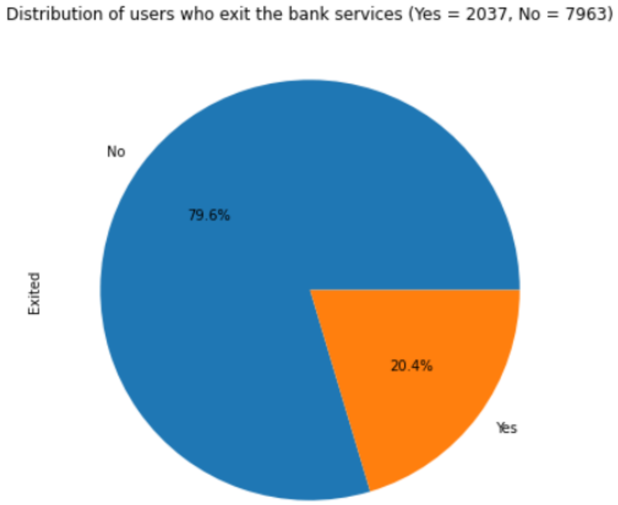
Group No. [9]

Student Name	Student Id
Haonan Jiang	10533001
Shu Du	10505024
Jiyan Zhu	10415483

	Haonan Jiang	Shu Du	Jiyan Zhu
Haonan Jiang	<100 %>	<100 %>	< 100%>
Shu Du	<100 %>	<100 %>	<100 %>
Jiyan Zhu	<100 %>	<100 %>	<100 %>

Task 1

1. What is the proportion of customers who exited and stopped using the banking services?

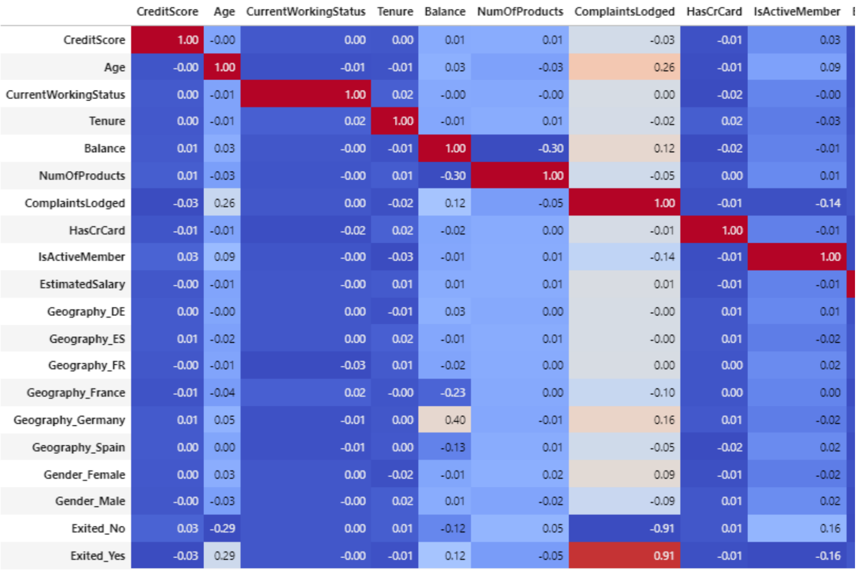


As shown in the pie chart above, 29.4% (2037) of the users have exited the bank services while 79.6% (7963) of the users have stayed with the bank services.

2. The dataset may include irrelevant and redundant variables. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.

By looking at the dataset, it has been found that ***RowNumber, CustomerId*** and ***Surname*** are irrelevant variables for constructing the predictive models as both ***RowNumber*** and ***CustomerId*** are unique numbers for identification purpose, and Surname are string values that has no direct relationships with whether a customer will exit and stop using the banking services.

Also, ***Gender*** and ***Sex*** are redundant variables as they represent the same thing while ***Gender*** is string variable and sex is a binary variable where 0 is male, 1 is female.



	CreditScore	Age	CurrentWorkingStatus	Tenure	Balance	NumOfProducts	ComplaintsLodged	HasCrCard	IsActiveMember
CreditScore	1.00	-0.00	0.00	0.00	0.01	0.01	-0.03	-0.01	0.03
Age	-0.00	1.00	-0.01	-0.01	0.03	-0.03	0.26	-0.01	0.09
CurrentWorkingStatus	0.00	-0.01	1.00	0.02	-0.00	-0.00	0.00	-0.02	-0.00
Tenure	0.00	-0.01	0.02	1.00	-0.01	0.01	-0.02	0.02	-0.03
Balance	0.01	0.03	-0.00	-0.01	1.00	-0.30	0.12	-0.02	-0.01
NumOfProducts	0.01	-0.03	-0.00	0.01	-0.30	1.00	-0.05	0.00	0.01
ComplaintsLodged	-0.03	0.26	0.00	-0.02	0.12	-0.05	1.00	-0.01	-0.14
HasCrCard	-0.01	-0.01	-0.02	0.02	-0.02	0.00	-0.01	1.00	-0.01
IsActiveMember	0.03	0.09	-0.00	-0.03	-0.01	0.01	-0.14	-0.01	1.00
EstimatedSalary	-0.00	-0.01	-0.00	0.01	0.01	0.01	0.01	-0.01	-0.01
Geography_DE	0.00	-0.00	0.00	-0.01	0.03	0.00	-0.00	0.01	0.01
Geography_ES	0.01	-0.02	0.00	0.02	-0.01	0.01	-0.00	0.01	-0.02
Geography_FR	-0.00	-0.01	-0.03	0.01	-0.02	0.00	0.00	0.00	0.02
Geography_France	-0.01	-0.04	0.02	-0.00	-0.23	0.00	-0.10	0.00	0.00
Geography_Germany	0.01	0.05	-0.01	0.00	0.40	-0.01	0.16	0.01	-0.02
Geography_Spain	0.00	0.00	-0.01	0.00	-0.13	0.01	-0.05	-0.02	0.02
Gender_Female	0.00	0.03	0.00	-0.02	-0.01	0.02	0.09	-0.01	-0.02
Gender_Male	-0.00	-0.03	-0.00	0.02	0.01	-0.02	-0.09	0.01	0.02
Exited_No	0.03	-0.29	0.00	0.01	-0.12	0.05	-0.91	0.01	0.16
Exited_Yes	-0.03	0.29	-0.00	-0.01	0.12	-0.05	0.91	-0.01	-0.16

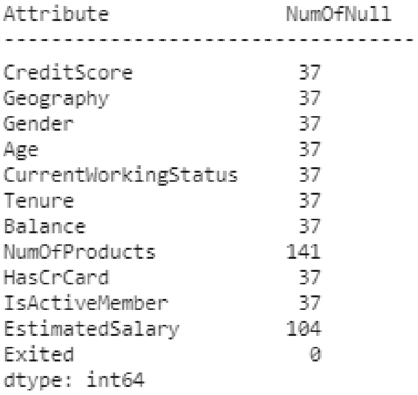
After looking at the correlation of each of the attribute, it has been found that in particular that the attribute ***ComplaintsLodged*** has a very strong correlation with the target variable ***Exited*** while all of the others have a relatively low correlation. Thus, ***ComplaintsLodged*** could potentially be a leaky variable where its strong correlation dominates the target variable dominates the predictions when constructing the models. Therefore, it is further decided to not include the attribute ***ComplaintsLodged.***

Therefore, it has been decided that ***RowNumber, CustomerId*** and ***Surname*** will not be included as it does not relate to the predicting tasks and ***Sex*** is also excluded due to it has the same meaning with ***Gender*** while gender has more valid rows (9963) than ***Sex*** (9805). Lastly, ***ComplaintsLodged*** is also excluded according to it could be a leaky variable potentially.

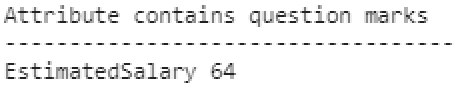
Variables included in the analysis are:

Variable	Roles	Measurement Level Set
CreditScore	Input	Numeric (Continuous)
Geography	Input	Categorical (Nominal)
Gender	Input	Categorical (Binary)
Age	Input	Numeric (Discrete)
CurrentWorkingStatus	Input	Categorical (Binary)
Tenure	Input	Numeric (Discrete)
Balance	Input	Numeric (Continuous)
NumOfProducts	Input	Numeric (Discrete)
HasCrCard	Input	Categorical (Binary)
IsActiveMember	Input	Categorical (Binary)
EstimatedSalary	Input	Numeric (Continuous)
Exited	Target	Categorical (Binary)

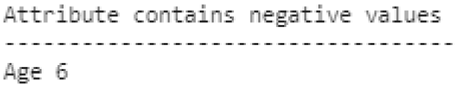
3. Did you have to fix any data quality problems? Detail them. Apply the imputation method(s) to the variable(s) that need it. List the variables that needed it. Justify your choice of imputation if needed.



```
Attribute              NumOfNull
----------------------------------
CreditScore             37
Geography               37
Gender                  37
Age                     37
CurrentWorkingStatus    37
Tenure                  37
Balance                 37
NumOfProducts          141
HasCrCard               37
IsActiveMember          37
EstimatedSalary        104
Exited                   0
dtype: int64
```



```
Attribute contains question marks
----------------------------------
EstimatedSalary 64
```



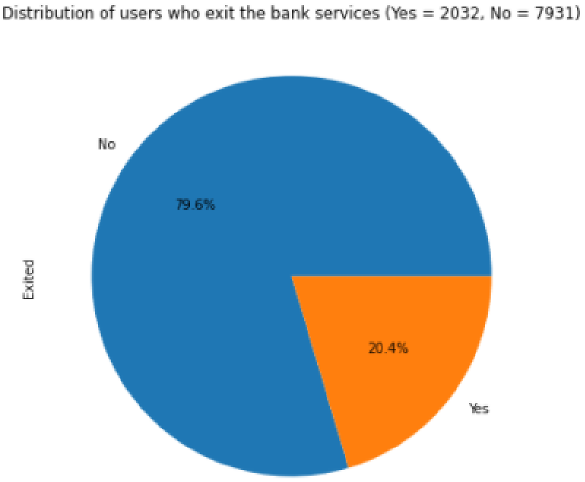
```
Attribute contains negative values
----------------------------------
Age 6
```

After removing the 4 attributes stated above, it has been found that apart from the attribute ***Exited***, all the other attributes contain a different number of null values (as shown below). Furthermore, there exists 64 questions marks in ***EstimatedSalary*** and 6 negative values in ***Age*** which are all considered as invalid.

Thus, it is obvious that ***NumOfProducts***, ***EstimatedSalary*** and ***Age*** contains more invalid data comparing to the others. Therefore, mean age 39 and mean salary 100198.15 are used to replace the invalid data in ***EstimatedSalary*** and ***Age*** respectively. As the ***NumOfProducts*** needs to be a whole integer, mean value would not be a valid imputation method in this case. As both the median and mode for ***NumOfProducts*** is 1, it is further decided to use the 1 to replace all the invalid data within the attribute.

As all the other attributes are containing a relatively low volume of invalid data,

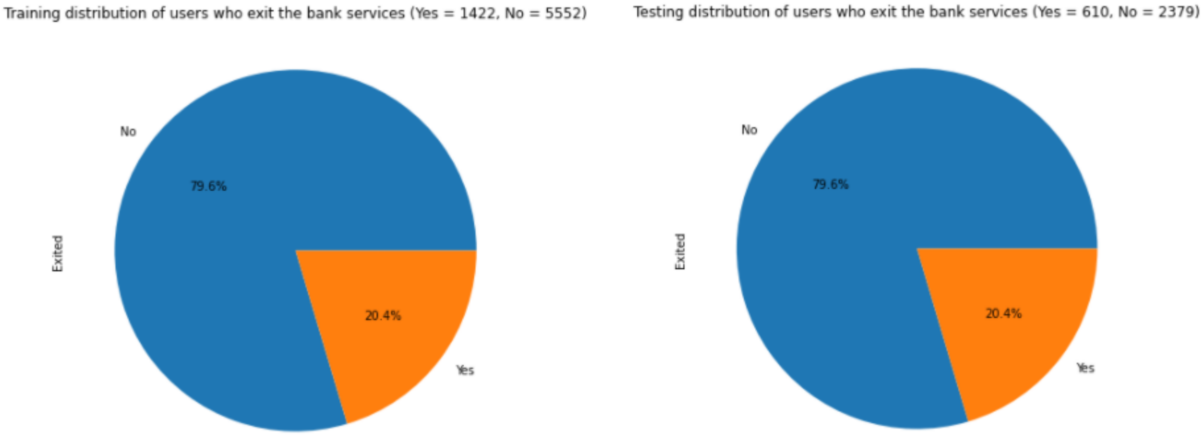
4. Report the proportion of values of the target variable for the dataset after the above-mentioned pre-processing.



After performed all the above-mentioned pre-processing, there are currently 20.4% (2032) have exited the bank services, and 79.6% (7391) users stayed with the bank services.

5. What distribution split between training and test datasets have you used?

By using the train_test_split() function to split the data randomly into 30% testing and 70 % training, the data now distributed as follow:

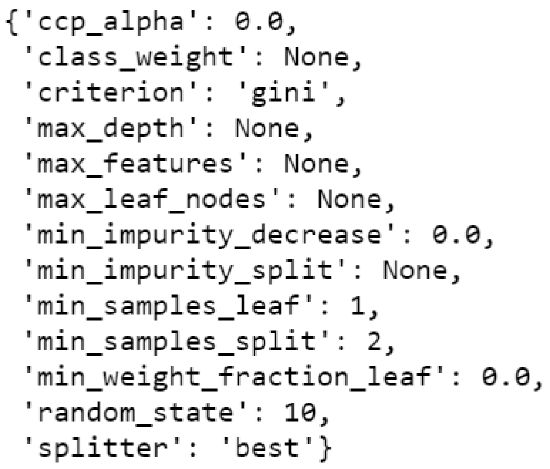


It can be found that both training and testing data has the same proportion of people exited or continue the banking services and

Task 2. Predictive Modeling Using Decision Trees (4 marks)

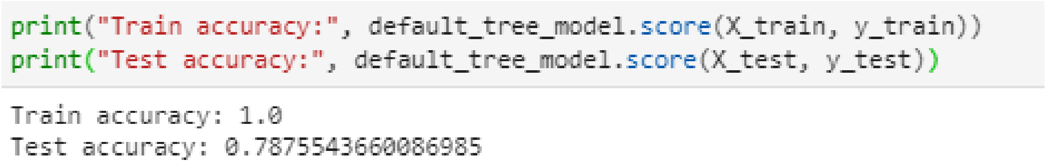
1. Build a decision tree using the default setting. Examine the tree results and answer the following:

a. **What parameters have been used in building the tree? Detail them.**
By using ‘**model.get_params(deep=True)’,** the output below shows all the parameters have been used. They are the default parameters of ‘sklearn.tree.decisionTreeClassifier’.



```
{'ccp_alpha': 0.0,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': None,
 'max_features': None,
 'max_leaf_nodes': None,
 'min_impurity_decrease': 0.0,
 'min_impurity_split': None,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'random_state': 10,
 'splitter': 'best'}
```

b. **What is classification accuracy on training and test datasets?**

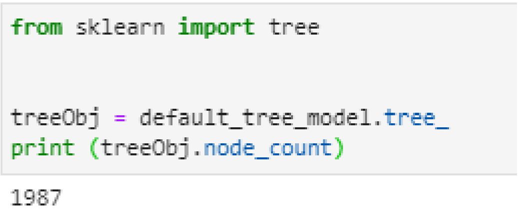


```
print("Train accuracy:", default_tree_model.score(X_train, y_train))
print("Test accuracy:", default_tree_model.score(X_test, y_test))
```

```
Train accuracy: 1.0
Test accuracy: 0.7875543660086985
```

c. **What is the size of the tree (number of nodes and rules)?**
The number of nodes and rules is 1987

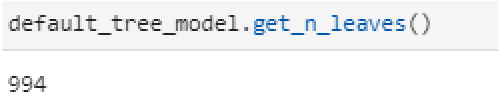
.



```
from sklearn import tree


treeObj = default_tree_model.tree_
print (treeObj.node_count)
```

```
1987
```

The number of leaf nodes is 994.



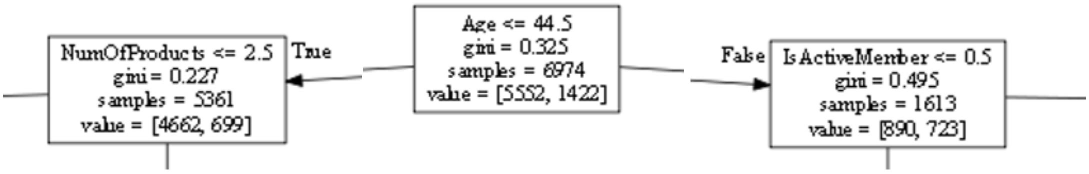
```
default_tree_model.get_n_leaves()
```

```
994
```

d. Which variable is used for the first split? What are the variables that

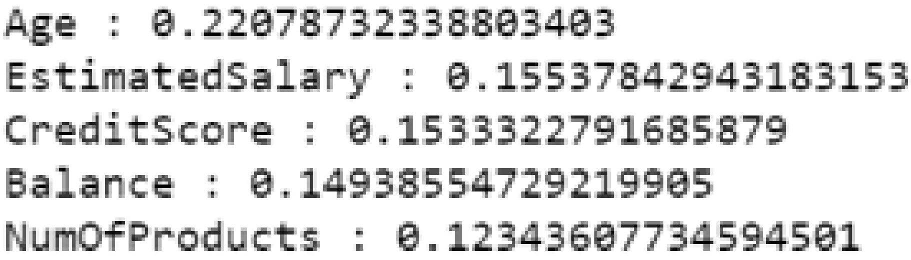
are used for the second split?

From the saved png we can see that variable ‘Age’ is used for the first split. ‘NumOfProducts’(left) and ‘IsActiveMember’ (right) are used for the second split.



d. What are the 5 important variables in building the tree?

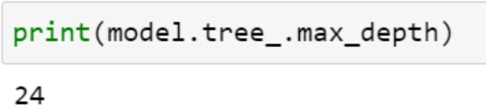
Based on the feature importance, below are the top 5 important variables in building this tree.



```
Age : 0.22078732338803403
EstimatedSalary : 0.15537842943183153
CreditScore : 0.1533322791685879
Balance : 0.14938554729219905
NumOfProducts : 0.12343607734594501
```

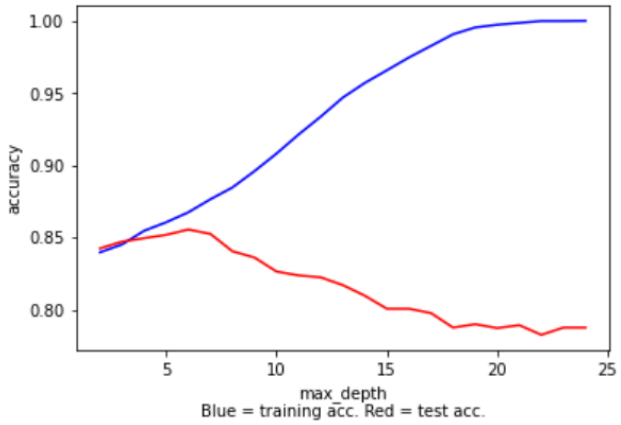
e. Report if you see any evidence of model overfitting.

The max tree depth of the model is 24. From the plot of max depth hyperparameter values with 24 versus training and test accuracy score, it can be seen that the test accuracy starts to drop from depth of 6 while the training accuracy still increases. So, there is overfitting when the depth goes beyond 6.



```
print(model.tree_.max_depth)
```

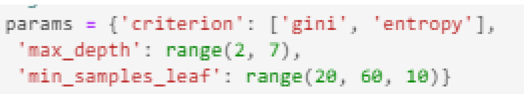
```
24
```



Blue = training acc. Red = test acc.

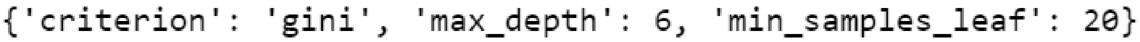
2. Build another decision tree tuned with GridSearchCV. Examine the tree results.

a. What are the optimal parameters for this decision tree?



```
params = {'criterion': ['gini', 'entropy'],
 'max_depth': range(2, 7),
 'min_samples_leaf': range(20, 60, 10)}
```

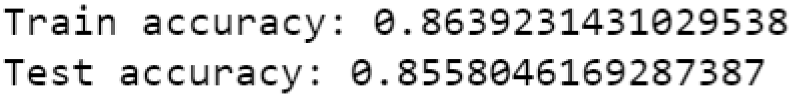
By using the grid search function for the above parameters,
The best parameters are shown as below.



```
{'criterion': 'gini', 'max_depth': 6, 'min_samples_leaf': 20}
```

b. What is classification accuracy on training and test datasets?

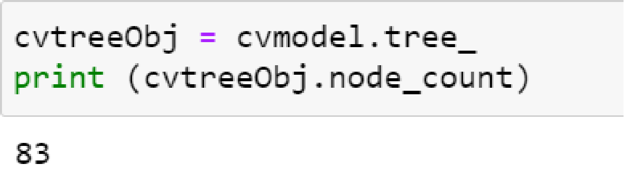
Both the training and test set accuracy are around 0.86.



```
Train accuracy: 0.8639231431029538
Test accuracy: 0.8558046169287387
```

c. What is the size of the chosen tree (number of nodes and rules)?

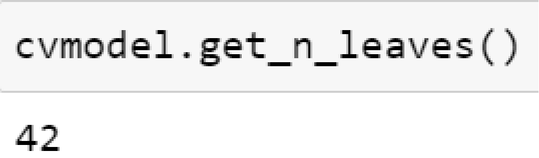
The chosen tree has 83 nodes and rules.



```
cvtreeObj = cvmodel.tree_
print (cvtreeObj.node_count)
```

```
83
```

Number of leaf nodes is 42.

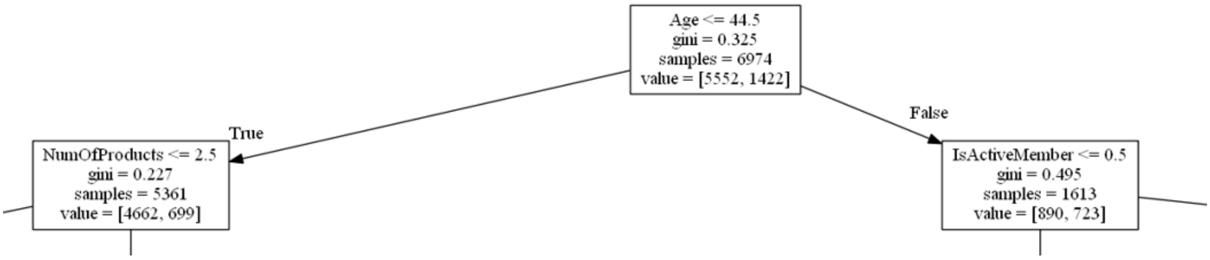


```
cvmodel.get_n_leaves()
```

```
42
```

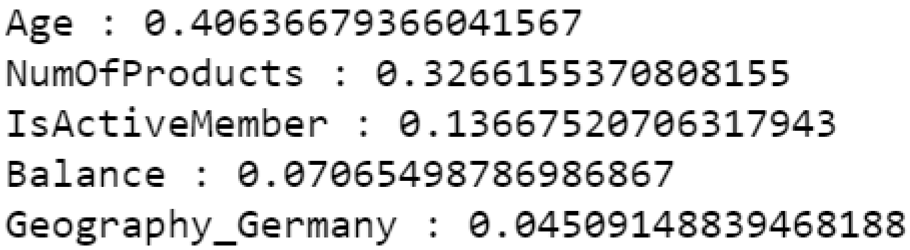
d. Which variable is used for the first split? What are the variables that are used for the second split?

The first split used ‘Age’ , the second split used ‘NumOfProducts’ and ‘IsActiveMember’.



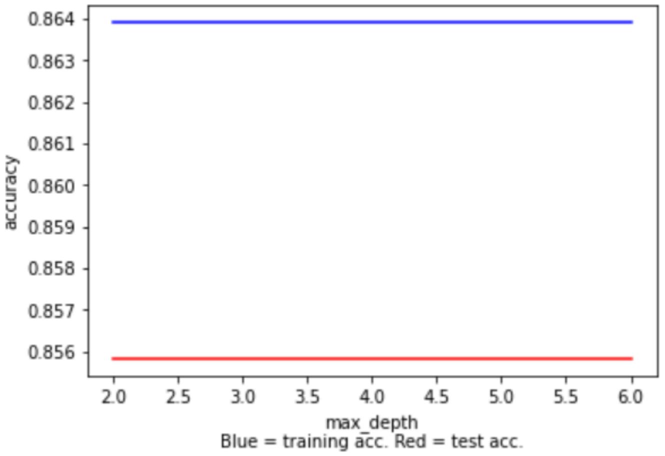
e. What are the 5 important variables in building the tree?

The top 5 important variables are shown below in a descending order.



```
Age : 0.40636679366041567
NumOfProducts : 0.3266155370808155
IsActiveMember : 0.13667520706317943
Balance : 0.07065498786986867
Geography_Germany : 0.04509148839468188
```

f. Report if you see any evidence of model overfitting.
Plotting the accuracy vs tree depth graph again, as the depth increases, both training and test accuracy remains the same. So, there is no overfitting taken place.



3. What is the significant difference do you see between these two decision tree models – default (Task 2.1) and using GridSearchCV (Task 2.2)? How do they compare performance-wise? Explain why those changes may have happened.

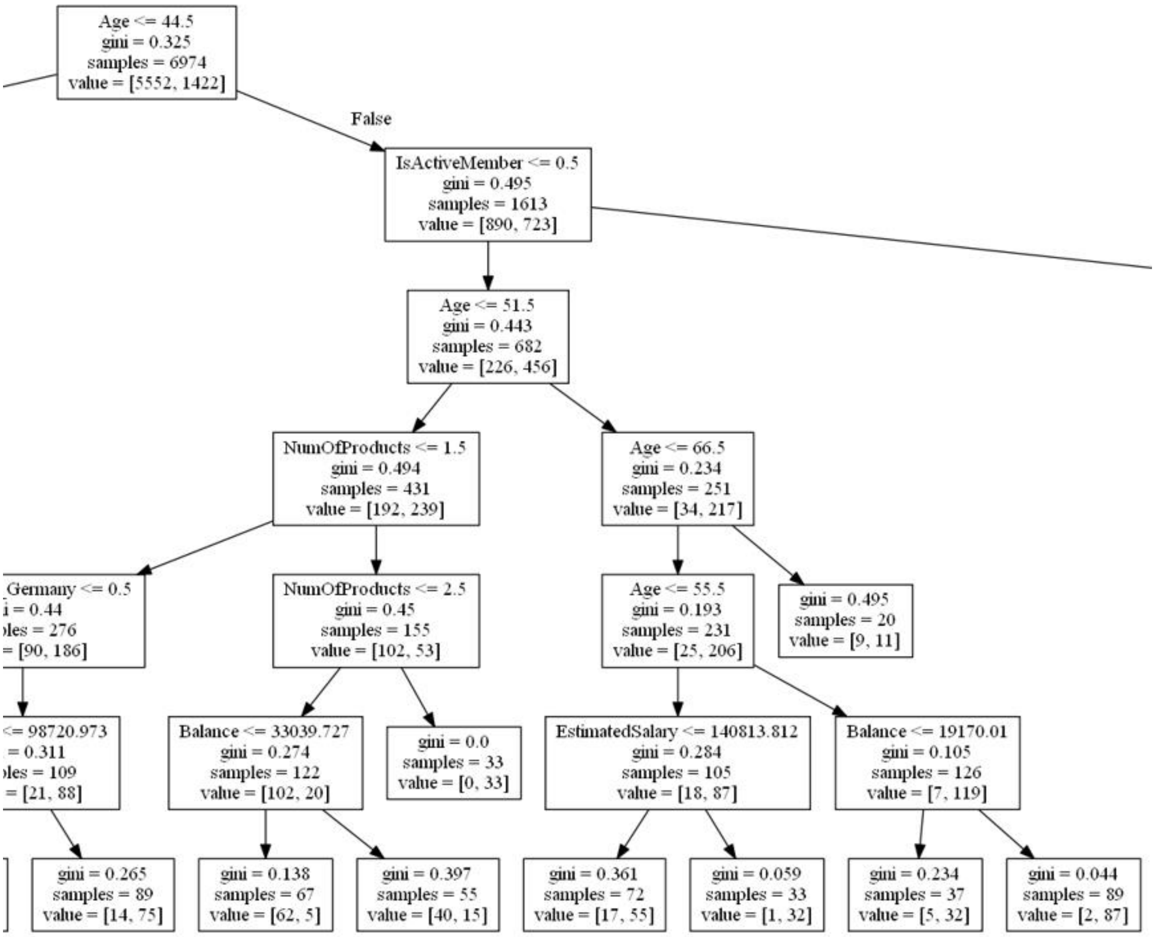
The major difference is that the model complexity has been significantly reduced for the optimal tree. The tree depth was reduced from 24 to 6 and the number of nodes was downsized from 1959 to 83.

By eliminating model overfitting, the optimal tree has a much better performance of 0.86 compared to default tree's 0.79 on test accuracy. The computational speed in using the model is faster for the optimal model.

This is because when doing a grid search, it finds the best hyper parameter combinations and evaluate the performance based on validation dataset (K-fold cross validation in this model). Thus, it eliminates model overfitting by reducing model complexity and improve model performance.

4. From the better model, can you provide a descriptive summary of customers

that most likely exit and stop using the banking services?



Above is a subtree of the optimum model. The numbers of the 'value' attribute stand for not exited (left) and exited (right). The bottom rightmost leaf node gives the highest ratio of customer exited (87) vs not exited (2). By following this tree splits, the customers that most likely exit would be an inactive member aged between 56 and 66 whose balance are over 19170.01.

Task 3. Predictive Modeling Using Regression (5.5 marks)

1. Describe what and why you will have to do additional preparation for variables to be used in regression modelling. List the variables that needed it with the processing detail.

Standardisation is needed for logistic regression model. This is due to that regression models are sensitive to input variables that having different ranges, which will further make it hard to compare between data points. It could also affect the gradient descent where the weights for the larger scale inputs will performing faster updates comparing to the inputs that has a smaller scale, which further leads to a suboptimal model performance.

The variables that need standardisation are variables not scaled between 0 and 1, they are 'CreditScore',' Age', 'Tenure', 'Balance', 'NumOfProducts', 'EstimatedSalary'.

Regression models are also sensitive to extreme or outlying values in the input space. Highly skewed inputs with or inputs distribute in a kurtotic way are often chosen comparing to the inputs that have the better overall predictions. Thus, Logarithmic transformation can be used to regularise the input distributions and improve model performance.

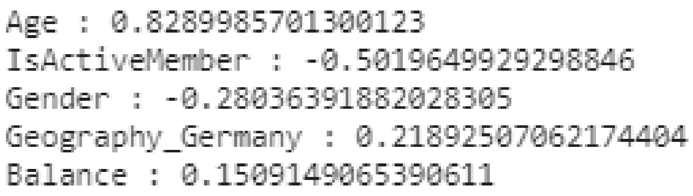
'The variables that need Logarithmic transformation are numerical variables with skewed distribution, they are' Age', 'Balance', 'NumOfProducts'.

2. Build a regression model using the default regression method with all inputs. Once you have completed it, build another model and tune it using GridSearchCV. Answer the following:

a. **Name the Regression function used.**
   Both of the default and grid search model use Logistic Regression.
b. **Report the variables that are included in the regression model.**
   The variables used for the both models are the same: CreditScore, Gender, Age, CurrentWorkingStatus, Tenure Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary Geography_DE, Geography_ES, Geography_FR, Geography_France Geography_Germany Geography_Spain.

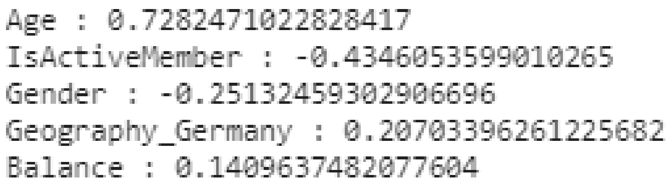
c. **Report the top-5 important variables (in order) in the model.**

Below are the top-5 important variables by looking at the absolute value of coefficients. Both have the same variables in the same order. However, the coefficients varies a little.



```
Age : 0.8289985701300123
IsActiveMember : -0.5019649929298846
Gender : -0.28036391882028305
Geography_Germany : 0.21892507062174404
Balance : 0.1509149065390611
```

Figure 1 Top 5 most important variable in default

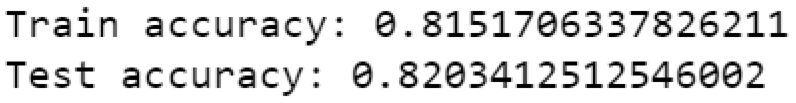


```
Age : 0.7282471022828417
IsActiveMember : -0.4346053599010265
Gender : -0.25132459302906696
Geography_Germany : 0.20703396261225682
Balance : 0.1409637482077604
```

Figure 2 Top 5 most important variable in gridserach

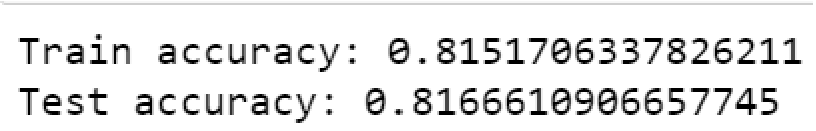
d. What is classification accuracy on training and test datasets? Report any sign of overfitting.

The test accuracy is slightly higher than the training accuracy, so the model is not overfitted.



```
Train accuracy: 0.8151706337826211
Test accuracy: 0.820341251254602
```

Figure 3 Logsitic Regression by default



```
Train accuracy: 0.8151706337826211
Test accuracy: 0.8166610906657745
```

Figure 4 Logistic Regression with Gridsearch

3. Build another regression model using the subset of inputs selected by RFE.

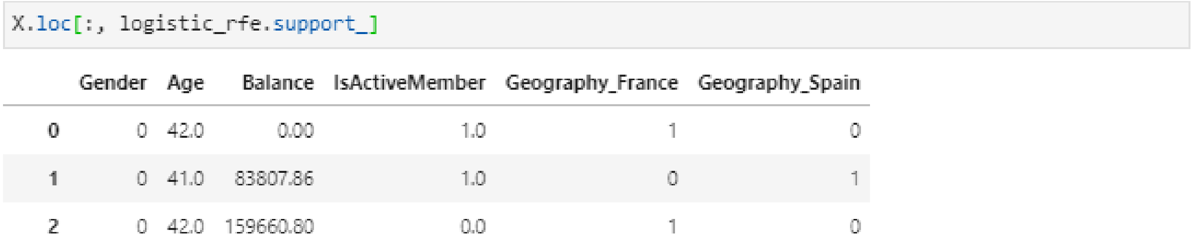
Answer the following:

a. Was dimensionality reduction useful to identify a good feature set for building the accurate model?

Dimensionality reduction using RFE is useful to identify good feature set for the accurate model.

b. Report the variables that are included in the regression model.

RFE eliminated the variables from 16 to 6. The variables included are shown as below.

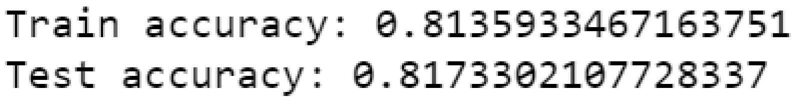


```
X.loc[:, logistic_rfe.support_]
```

	Gender	Age	Balance	IsActiveMember	Geography_France	Geography_Spain
0	0	42.0	0.00	1.0	1	0
1	0	41.0	83807.86	1.0	0	1
2	0	42.0	159660.80	0.0	1	0

c. What is classification accuracy on training and test datasets? Report any sign of overfitting.

The test accuracy has increased to 0.817 from the previous 0.816. As the test accuracy is higher than the training accuracy, there is no overfitting.

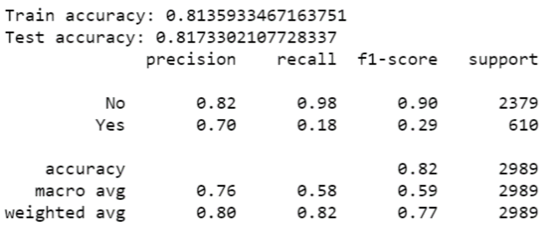


```
Train accuracy: 0.8135933467163751
Test accuracy: 0.8173302107728337
```

4. Using the comparison statistics, which of the regression models appears to be better? Explain why.

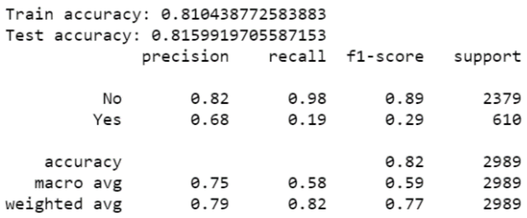
By comparing the accuracy table, the default model on the bottom has the highest recall score of 'Yes', which means it has a higher capability to capture the people who is going to exit the banking services.

The model also has the best test accuracy among the three. This means the default logistic regression model is better on predicting a customer who is likely to leaving the bank.



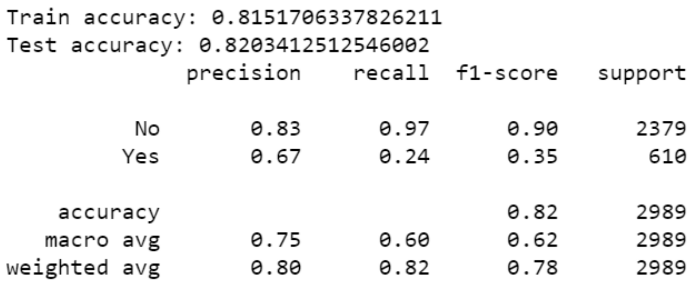
```
Train accuracy: 0.8135933467163751
Test accuracy: 0.8173302107728337
              precision    recall  f1-score   support

          No       0.82      0.98      0.90      2379
         Yes       0.70      0.18      0.29       610

    accuracy                           0.82      2989
   macro avg       0.76      0.58      0.59      2989
weighted avg       0.80      0.82      0.77      2989
```

Figure 5 Logistic Regression with RFE & Gridsearch



```
Train accuracy: 0.810438772583883
Test accuracy: 0.815991970558715З
              precision    recall  f1-score   support

          No       0.82      0.98      0.89      2379
         Yes       0.68      0.19      0.29       610

    accuracy                           0.82      2989
   macro avg       0.75      0.58      0.59      2989
weighted avg       0.79      0.82      0.77      2989
```

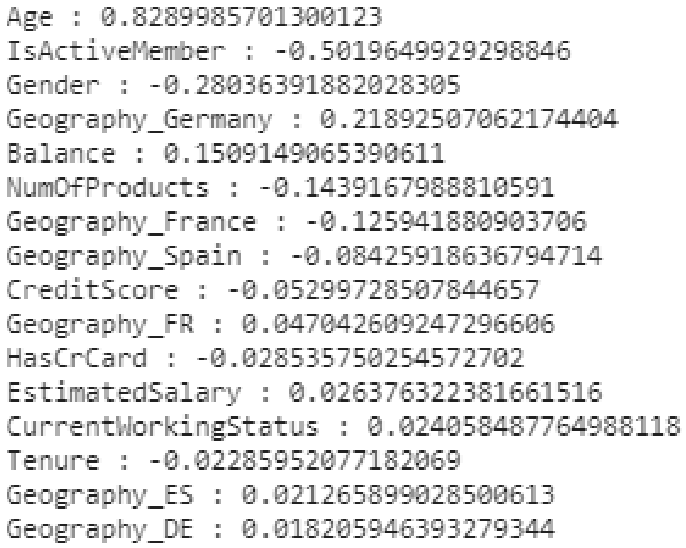
Figure 6 Logistic Regression with Gridsearch



```
Train accuracy: 0.8151706337826211
Test accuracy: 0.8203412512546002
              precision    recall  f1-score   support

          No       0.83      0.97      0.90      2379
         Yes       0.67      0.24      0.35       610

    accuracy                           0.82      2989
   macro avg       0.75      0.60      0.62      2989
weighted avg       0.80      0.82      0.78      2989
```

Figure 7 Logistic Regression by default

5. From the better model, can you provide a descriptive summary of customers that most likely exit and stop using the banking services?



```
Age : 0.8289985701300123
IsActiveMember : -0.5019649929298846
Gender : -0.28036391882028305
Geography_Germany : 0.21892507062174404
Balance : 0.1509149065390611
NumOfProducts : -0.1439167988810591
Geography_France : -0.125941880903706
Geography_Spain : -0.08425918636794714
CreditScore : -0.05299728507844657
Geography_FR : 0.04704260924729660б
HasCrCard : -0.028535750254572702
EstimatedSalary : 0.026376322381661516
CurrentWorkingStatus : 0.024058487764988118
Tenure : -0.02285952077182069
Geography_ES : 0.021265899028500613
Geography_DE : 0.018205946393279344
```

Based on the feature importance of our best default model where the feature coefficients with an absolute value over 0.1: Age (0.829), IsActiveMember(-0.501), Gender(-0.280), Geography_Germany (0.219), Balance(0.151),NumOfProducts (-0.144) , Geography_France(-0.126).

It can tell that if a customer is an aged female with a higher account balance who is currently an inactive member from Germany or not from France, she would have a higher chance to exit the banking services.

Task 4. Predictive Modelling Using Neural Networks (5.5 marks)

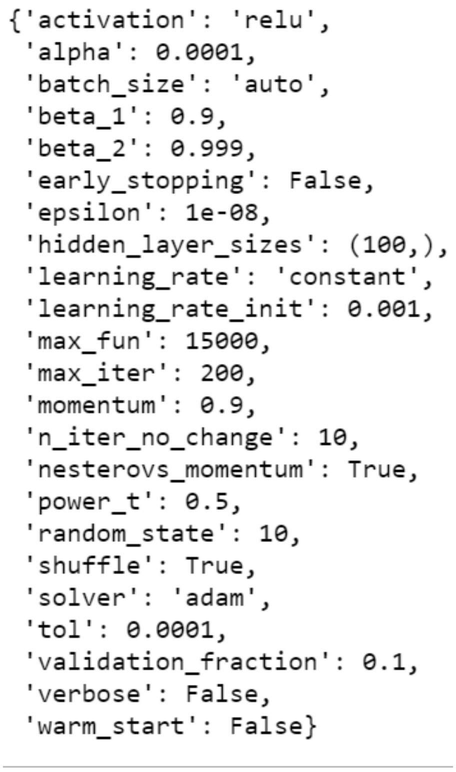
1. Build a Neural Network model using the default setting. Answer the

following:

a. What are the parameters used, e.g., network architecture, iterations,

activation function, etc?

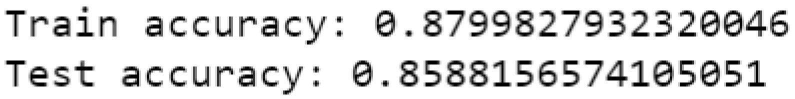
Below are the default parameters, hidden layer size is 100, iterations are mixture of 200, active function is relu.



```
{'activation': 'relu',
 'alpha': 0.0001,
 'batch_size': 'auto',
 'beta_1': 0.9,
 'beta_2': 0.999,
 'early_stopping': False,
 'epsilon': 1e-08,
 'hidden_layer_sizes': (100,),
 'learning_rate': 'constant',
 'learning_rate_init': 0.001,
 'max_fun': 15000,
 'max_iter': 200,
 'momentum': 0.9,
 'n_iter_no_change': 10,
 'nesterovs_momentum': True,
 'power_t': 0.5,
 'random_state': 10,
 'shuffle': True,
 'solver': 'adam',
 'tol': 0.0001,
 'validation_fraction': 0.1,
 'verbose': False,
 'warm_start': False}
```

b. What is the classification accuracy on training and test datasets?

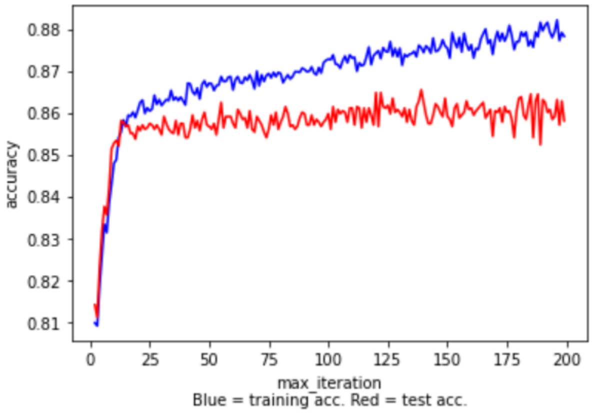
The input data used for training and testing has been logarithmic transformed and standardised. The training accuracy is about 0.88 and test accuracy is around 0.86.



```
Train accuracy: 0.8799827932320046
Test accuracy: 0.8588156574105051
```

c. Comment on the training process concerning underfitting, overfitting or good fitting.

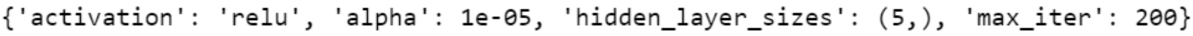
By plotting the max iterations vs training and test accuracy, it is obvious that the test accuracy converges at 10 iterations while the training accuracy keeps rising. This indicates that the model is a little bit start overfitting. However, as the variance between train and test is minimal (0.02 approximately). It can also consider the model is still a reasonable good fit.



2. Refine this network by tuning it with GridSearchCV. Answer the following:

a. What are the parameters used, e.g., network architecture, iterations, activation function, etc?

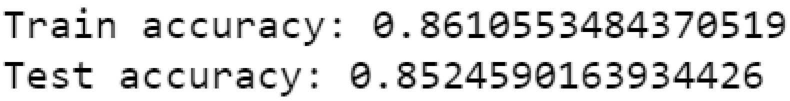
After tuning with GridSearchCV, the best parameters are shown below.



```
{'activation': 'relu', 'alpha': 1e-05, 'hidden_layer_sizes': (5,), 'max_iter': 200}
```

b. What is the classification accuracy on training and test datasets?

The training accuracy is 0.86 and test accuracy is 0.85



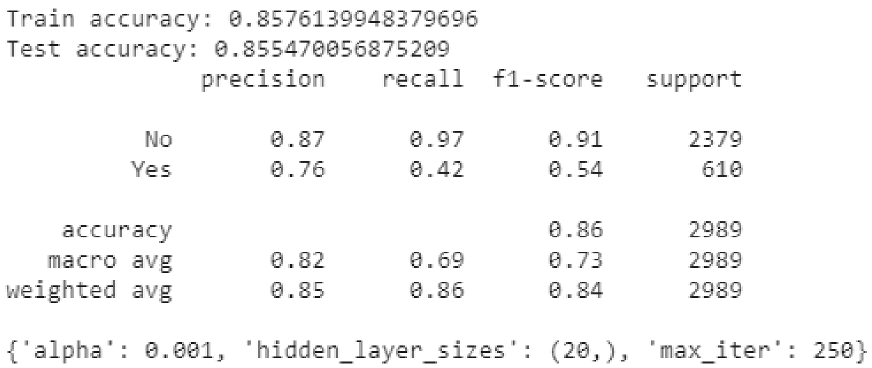
```
Train accuracy: 0.8610553484370519
Test accuracy: 0.8524590163934426
```

c. Comment on the training process concerning underfitting, overfitting or good fitting.

Comparing to the previous model, the grid search has successfully reduced the train accuracy to 0.86 while remains the test accuracy as 0.85. Thus, the gap between the training accuracy and test accuracy is minimal, so this is not overfitted, and it is a good fit..

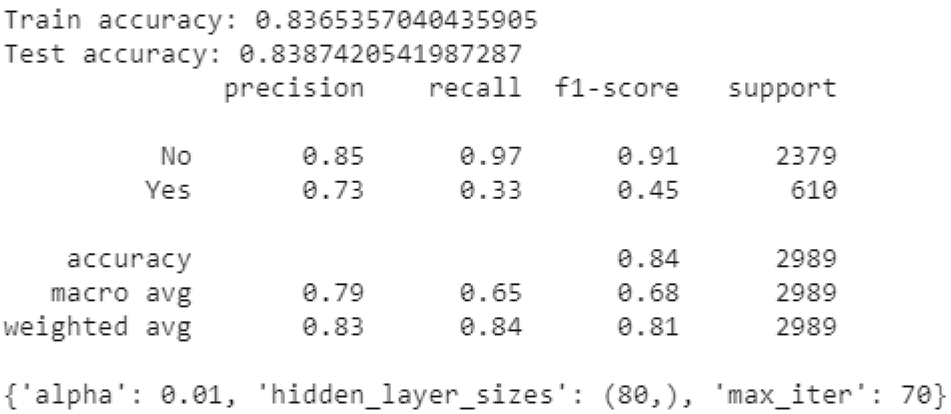
3. Would feature selection help here? Build another Neural Network model with inputs selected from RFE with regression (use the best model generated in Task 3) and from the decision tree (use the best model from Task 2). Tune the model with GridSearchCV to find the best parameters setting. Answer the following for the best neural network model:

a. Did feature selection help here? Which method of feature selection produced the best result? Any change in the network architecture? What inputs are being used as the network input?



```
Train accuracy: 0.8576139948379696
Test accuracy: 0.855470056875209
              precision    recall  f1-score   support

          No       0.87      0.97      0.91      2379
         Yes       0.76      0.42      0.54       610

    accuracy                           0.86      2989
   macro avg       0.82      0.69      0.73      2989
weighted avg       0.85      0.86      0.84      2989

{'alpha': 0.001, 'hidden_layer_sizes': (20,), 'max_iter': 250}
```

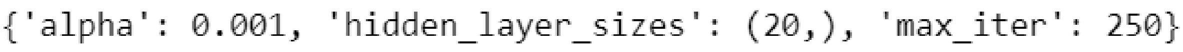
Figure 8 Input selected from best decision tree model



```
Train accuracy: 0.8365357040435905
Test accuracy: 0.8387420541987287
              precision    recall  f1-score   support

          No       0.85      0.97      0.91      2379
         Yes       0.73      0.33      0.45       610

    accuracy                           0.84      2989
   macro avg       0.79      0.65      0.68      2989
weighted avg       0.83      0.84      0.81      2989

{'alpha': 0.01, 'hidden_layer_sizes': (80,), 'max_iter': 70}
```

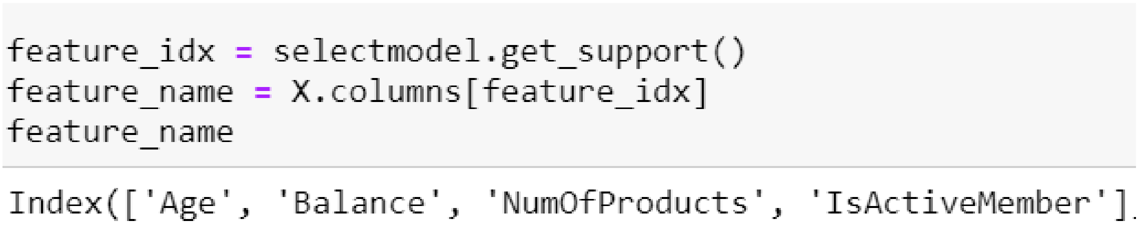
Figure 9 Input selected from best RFE logistic regression model

The Decision Tree feature selection produced the best result with a test accuracy of 0.855 while RFE with regression produces a test accuracy of 0.839. Thus, feature selection did help to improve model from a test accuracy of 0.852 to 0.855. The hidden layer size is increased to 20, max iterations rised from 120 to 250, alpha droped from 0.01 to 0.001, the activation function is still relu.



```
{'alpha': 0.001, 'hidden_layer_sizes': (20,), 'max_iter': 250}
```

The inputs used are Age, Balance, IsActiveMember and NumOfProducts, as shown below.



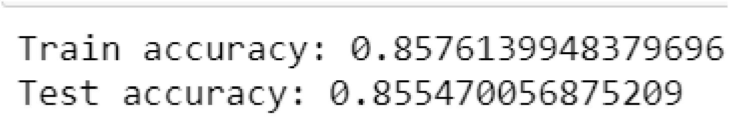
```
feature_idx = selectmodel.get_support()
feature_name = X.columns[feature_idx]
feature_name
```

```
Index(['Age', 'Balance', 'NumOfProducts', 'IsActiveMember']
```

b. What is classification accuracy on training and test datasets? Is there

any improvement in the outcome?

Comparing with the new accuracies below, test accuracy has been improved from of 0.852 to 0.855 while train accuracy has been reduced from 0.861 to 0.857. Thus, it has further suggested that the model constructed with decision tree feature selection has successfully improved the model accuracy and reduced overfitting in the meantime.



```
Train accuracy: 0.8576139948379696
Test accuracy: 0.855470056875209
```

c. How many iterations are now needed to train this network?

250 iterations are now needed to train this network.

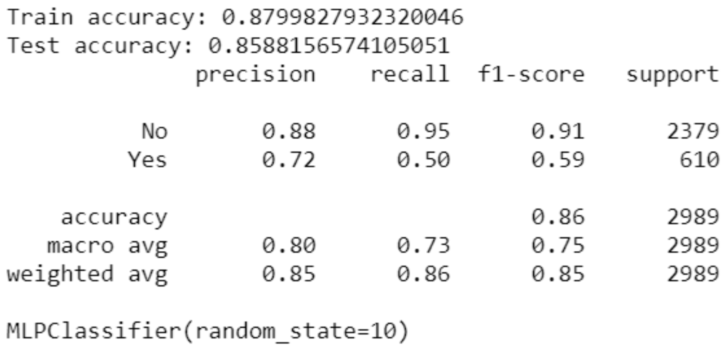
d. Comment on the training process concerning underfitting, overfitting

or good fitting.

As stated above, the test accuracy has improved with the decrease of training accuracy comparing to the model from Q2. Thus, it is clearly that the model is a god fit where there is only a tiny difference (0.003) between the training and testing accuracy.

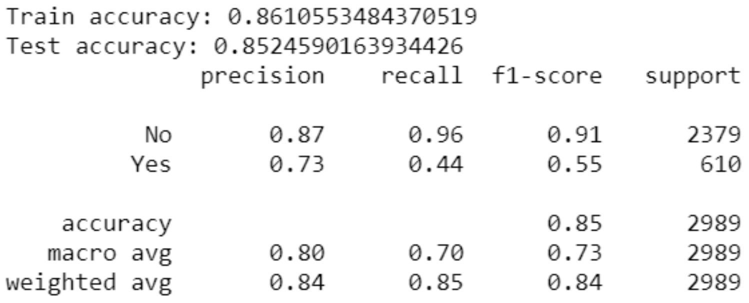
4. Using the comparison statistics, which of the Neural Network models appears to be better?

By comparing all the accuracy metrics, the default Network model on the top has the highest recall on 'Yes' and overall test accuracy, which makes it better than the others. Again, as the main purpose of the task is to predict the people who is going to exit the bank services. Thus, a better recall further suggest that the model is performing better to predict the people who exits.



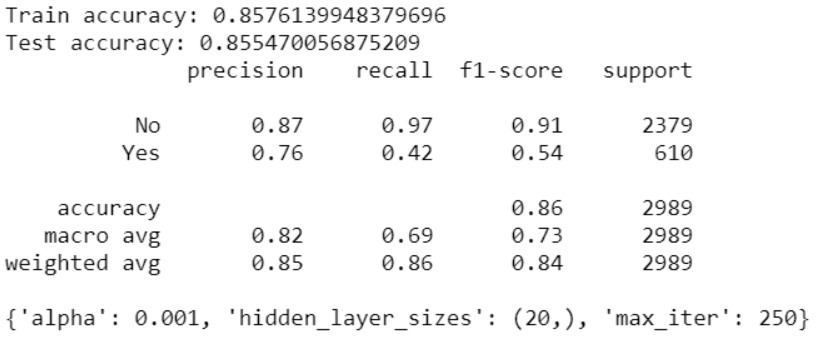
```
Train accuracy: 0.8799827932320046
Test accuracy: 0.8588156574105051
              precision    recall  f1-score   support

          No       0.88      0.95      0.91      2379
         Yes       0.72      0.50      0.59       610

    accuracy                           0.86      2989
   macro avg       0.80      0.73      0.75      2989
weighted avg       0.85      0.86      0.85      2989

MLPClassifier(random_state=10)
```

Figure 10 Network model by default



```
Train accuracy: 0.8610553484370519
Test accuracy: 0.8524590163934426
              precision    recall  f1-score   support

          No       0.87      0.96      0.91      2379
         Yes       0.73      0.44      0.55       610

    accuracy                           0.85      2989
   macro avg       0.80      0.70      0.73      2989
weighted avg       0.84      0.85      0.84      2989
```

Figure 11 Network model Gridsearch



```
Train accuracy: 0.8576139948379696
Test accuracy: 0.855470056875209
              precision    recall  f1-score   support

          No       0.87      0.97      0.91      2379
         Yes       0.76      0.42      0.54       610

    accuracy                           0.86      2989
   macro avg       0.82      0.69      0.73      2989
weighted avg       0.85      0.86      0.84      2989

{'alpha': 0.001, 'hidden_layer_sizes': (20,), 'max_iter': 250}
```

Figure 12 Networkmodel with Decision Tree feature selection

5. From the better model, can you provide a descriptive summary of

customers that most likely exit and stop using the banking services?

Neural network has very less interpretability which makes it difficult to provide a descriptive summary by this model.

Task 5. Comparing Predictive Models (4 marks)

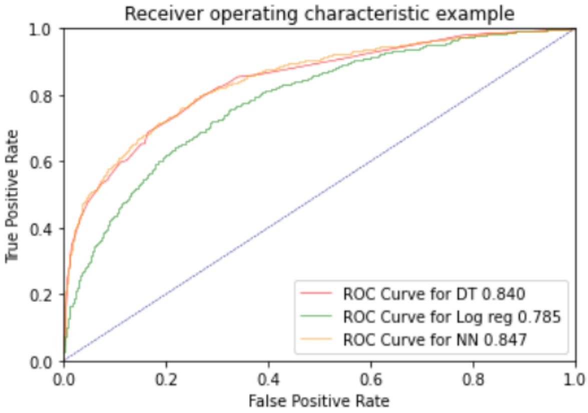
1. Use the comparison statistics to compare the best decision tree model, the

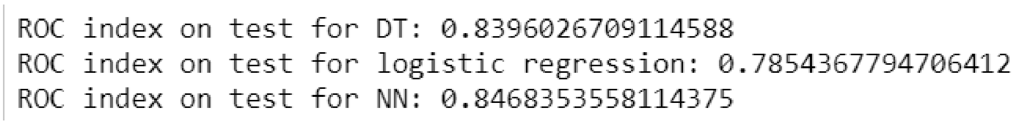
best regression model, and the best neural network model.

a. Discuss the findings led by:

(i) ROC Chart and Index;

Neural network model produces the best ROC and AUC score. This further means that the Neural network model performs in a better way comparing to the other two models based on varied discrimination threshold, t.

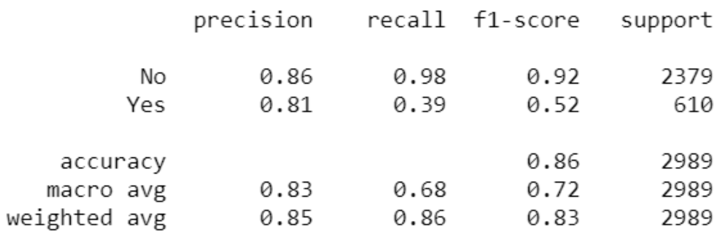




```
ROC index on test for DT: 0.8396026709114588
ROC index on test for logistic regression: 0.7854367794706412
ROC index on test for NN: 0.8468353558114375
```

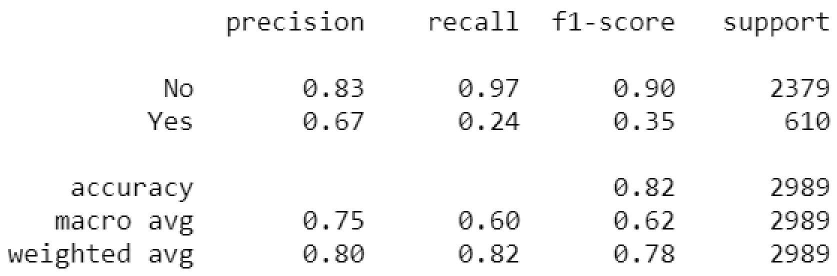
(ii) Accuracy Table;

Because the aim of this project is on finding a customer that would potentially leave the bank, so the recall score of ‘Yes’ is an important measure as it is the proportion of left customers that were retrieved. The model has the highest recall is the Decision Tree model with a score of 0.39, comparing to Logistic Regression(0.24) and Neural Network(0.35).



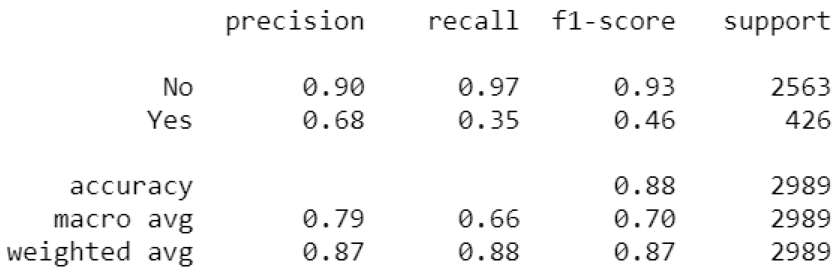
```
              precision    recall  f1-score   support

          No       0.86      0.98      0.92      2379
         Yes       0.81      0.39      0.52       610

    accuracy                           0.86      2989
   macro avg       0.83      0.68      0.72      2989
weighted avg       0.85      0.86      0.83      2989
```

Figure 13 Decision Tree Accuracy Table



```
              precision    recall  f1-score   support

          No       0.83      0.97      0.90      2379
         Yes       0.67      0.24      0.35       610

    accuracy                           0.82      2989
   macro avg       0.75      0.60      0.62      2989
weighted avg       0.80      0.82      0.78      2989
```

Figure 14 Logistic Regression Accuracy Table



```
              precision    recall  f1-score   support

          No       0.90      0.97      0.93      2563
         Yes       0.68      0.35      0.46       426

    accuracy                           0.88      2989
   macro avg       0.79      0.66      0.70      2989
weighted avg       0.87      0.88      0.87      2989
```

Figure 15 Neural Network Accuracy Table

b. Which model would you use in deployment based on these findings?

Discuss why?

I would use the best grid search Decision Tree model. In the ROC plot, although the Neural Network model has the largest AUC score, but it only wins the Decision Tree model by 0.007. The curves for Neural Network and Decision Tree are nearly the same. However, the Decision Tree model has a much higher recall on ‘Yes’, so the Decision Tree model is voted as the best model for this project.

2. Can you summarise the positives and negative aspects of each predictive modelling method based on this data analysis exercise?

Decision tree has the best interpretability and a fast speed on training and predicting on this data.

Logistic Regression has the second-best interpretability as it gives some reveals by its coefficients. Also, it is fast on training and predicting the data.

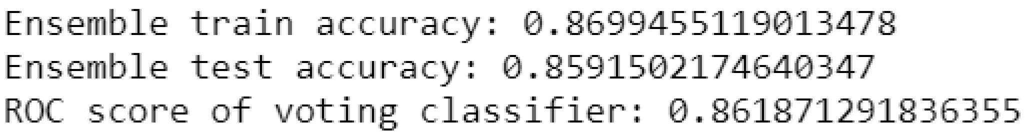
Neural Network is basically not interpretable in this case, and takes longer to train and predict, but it gives a decent accuracy as the result.

3. Finally, can you build an ensemble model combining all models? Does it

produce better/equal/worse performance and why?

Combining the best 3 models, we build a voting-based bagging model our ensemble model. Soft voting is chosen as it can predict the target variable based on the probabilities of each model, which is also a recommendation from sklearn.

It produces a better performance. Because Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to **decrease variance** (bagging), **bias** (boosting), or **improve predictions** (stacking).



```
Ensemble train accuracy: 0.8699455119013478
Ensemble test accuracy: 0.8591502174640347
ROC score of voting classifier: 0.861871291836355
```