MASTER OF DATA SCIENCE

FACULTY OF COMPUTER AND MATHEMATICAL SCIENCES

STA761 – STATISTICAL DATA MINING

PART B – CASE STUDY (70%)

NAME

NORAZMIERA AYUNIE BINTI AZMAN

GROUP

CS779/2A

LECTURER'S NAME

DR. AHMAD ZIA UL-SAUFIE BIN MOHAMAD JAPERI

DR NURAIN BINTI IBRAHIM

SUBMISSION DATE

10TH DECEMBER 2022

TABLE OF CONTENTS

LIST OF TABLES

TABLE OF FIGURES

1. BUSINESS UNDERSTANDING

In the context of the CRISP-DM process, the business understanding phase is the first step in building a predictive model to predict customer churn. During this phase, the business must define the problem they are trying to solve. In this study, the Yamani Tour & Travels agency wants to predict which customers are at risk of churning and the steps to prevent them from churning.

Our main objective is to build a predictive model that can accurately identify the customers who churned using different models such as logistic regression and decision tree approaches. The most accurate model is said to be the best predictive model for customer churn. The tools used in this study are Python for exploratory data analysis and RapidMiner for data preparation and modeling. Based on the indicators given for Yamani Tour & Travels agency, many factors can influence customer churn.

Overall, the business understanding phase aims to provide a clear and comprehensive understanding of the problem, goals, and data that will be used to build the predictive model for customer churn. This will serve as the foundation for the rest of the CRISP-DM process.

2. DATA UNDERSTANDING

2.1. DATA ACQUISITION

This dataset, named “Customertravel” consists of the Yamani Tour & Travels agency information. It is a supervised learning dataset where it is design to predict whether the customer will churn or not. The dataset contains a total of nine attributes and 954 instances to discover about. The attributes are ID, Age, Gender, Retire, AnnualIncomeClass, ServicesOpted, AccountSyncedToSocialMedia, BookedHotelOrNot and Churn.

Below is the description of the data:

Table 1 Data Acquisition

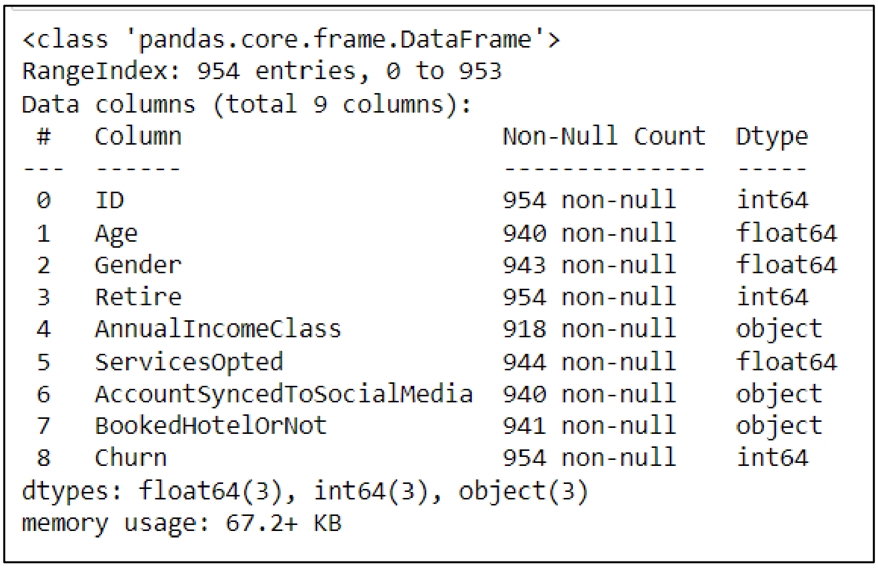
Variable Name	Description	Data Type
ID	ID of the customer	Integer
Age	Age of customer	Integer
Gender	Gender 0 = "Male" 1 = "Female"	Integer
Retire	Retire 0 = "No" 1 = "Yes"	Integer
AnnualIncomeClass	Class of annual income of user	Nominal
ServicesOpted	Number of times services opted during recent years	Integer
AccountSyncedToSocialMedia	Whether Company Account Of User Synchronised to Their Social Media	Nominal
BookedHotelOrNo	Whether the customer book lodgings/Hotels using company services	Nominal
Churn	Churn 0= "Customer Doesn’t Churn" 1 = "Customer Churn"	Integer

2.1. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is designed to help in examining the pattern of the data, the data sanity, and also summarizing the relevant data for future use (Exploratory Data Analysis (EDA) Notebook | Adobe Experience Platform, n.d.). Using Python as the medium, the EDA of this study was conducted.

2.1.1. INFORMATION ABOUT THE DATA

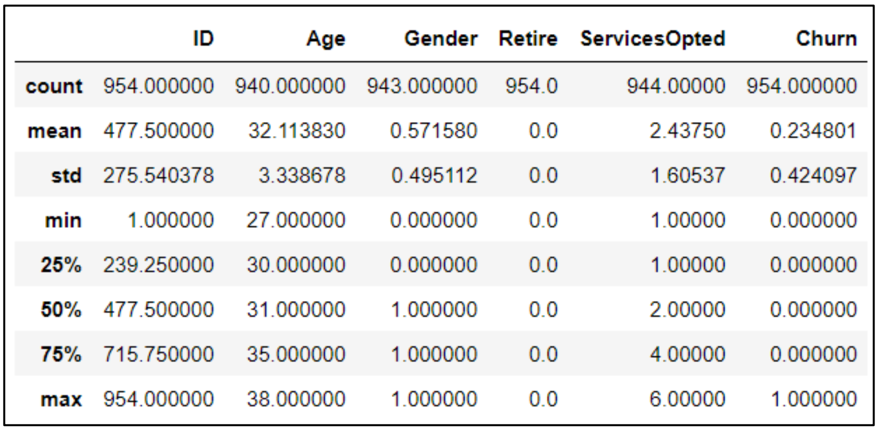
Four basic libraries were imported into Python: pandas, NumPy, seaborn, and matplotlib. pyplot. Before starting any work, it is essential to know the details of the data, both for the basic information and the descriptive statistic for numerical data. The basic information shows 954 entries with 9 attributes in the dataset. The dataset's data type consists of integers, float and object. In addition, the count of non-null rows for each attribute identified.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 954 entries, 0 to 953
Data columns (total 9 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   ID                          954 non-null    int64
 1   Age                         940 non-null    float64
 2   Gender                      943 non-null    float64
 3   Retire                      954 non-null    int64
 4   AnnualIncomeClass           918 non-null    object
 5   ServicesOpted               944 non-null    float64
 6   AccountSyncedToSocialMedia  940 non-null    object
 7   BookedHotelOrNot            941 non-null    object
 8   Churn                       954 non-null    int64
dtypes: float64(3), int64(3), object(3)
memory usage: 67.2+ KB
```

Figure 1 Data Information

A descriptive statistic shows the statistical measure for each numerical attribute, such as count, mean, standard deviation, max, and quartiles. The mean value for “Age” and “ServicesOpted” are 32.113830 and 2.43750 respectively.



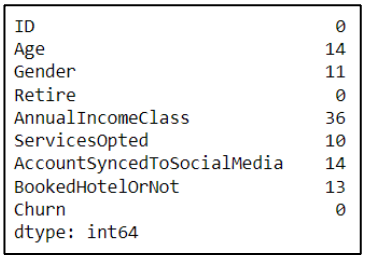
	ID	Age	Gender	Retire	ServicesOpted	Churn
count	954.000000	940.000000	943.000000	954.0	944.00000	954.000000
mean	477.500000	32.113830	0.571580	0.0	2.43750	0.234801
std	275.540378	3.338678	0.495112	0.0	1.60537	0.424097
min	1.000000	27.000000	0.000000	0.0	1.00000	0.000000
25%	239.250000	30.000000	0.000000	0.0	1.00000	0.000000
50%	477.500000	31.000000	1.000000	0.0	2.00000	0.000000
75%	715.750000	35.000000	1.000000	0.0	4.00000	0.000000
max	954.000000	38.000000	1.000000	0.0	6.00000	1.000000

Figure 2 Descriptive Statistic of the Data

2.1.2. DETECT THE MISSING VALUES

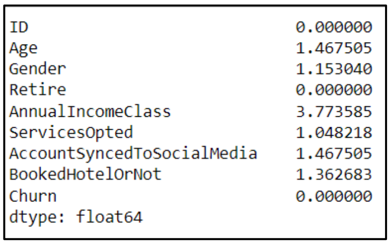
To improve the data quality, detecting the missing values in the dataset is essential, as the missingness is closely correlated with the result. If the missing values are ignored, it will be biased toward the outcome. Below are the steps where the missing values were detected. A representation in a heatmap clearly showed that there are some missing values in the dataset. Also stated below is the missing value percentage representation. From here, the following steps of handling the missing value between dropping the column or imputing the missing values can be decided based on the percentage.

The highest percentage for missing value is AnnualIncomeClass with 3.773685%, while the lowest is ServicesOpted with 1.048218%. Three attributes which are ID, Retire and Churn, do not contain the missing values.



```
ID                            0
Age                          14
Gender                       11
Retire                        0
AnnualIncomeClass            36
ServicesOpted                10
AccountSyncedToSocialMedia   14
BookedHotelOrNot             13
Churn                         0
dtype: int64
```

Figure 3 Count of the Missing Values



```
ID                           0.000000
Age                          1.467505
Gender                       1.153040
Retire                       0.000000
AnnualIncomeClass            3.773585
ServicesOpted                1.048218
AccountSyncedToSocialMedia   1.467505
BookedHotelOrNot             1.362683
Churn                        0.000000
dtype: float64
```

Figure 4 Percentage of the Missing Values

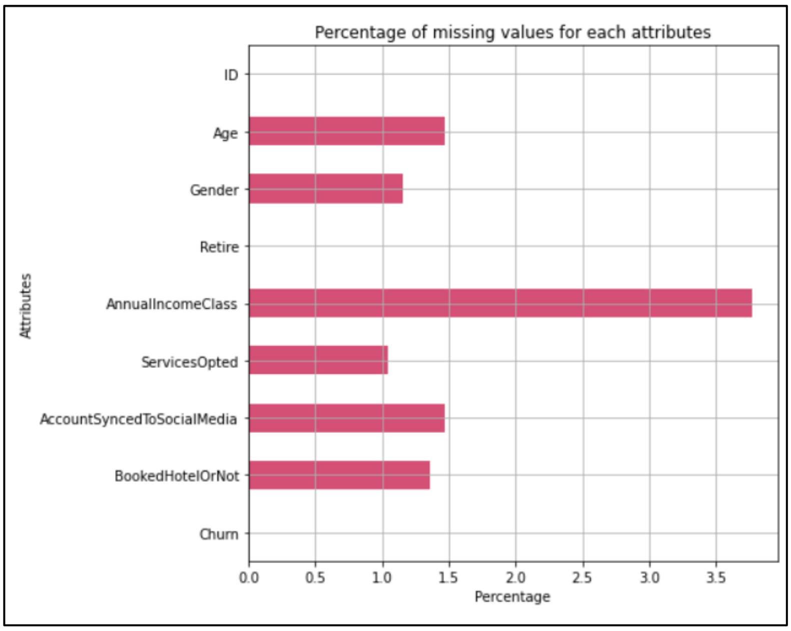
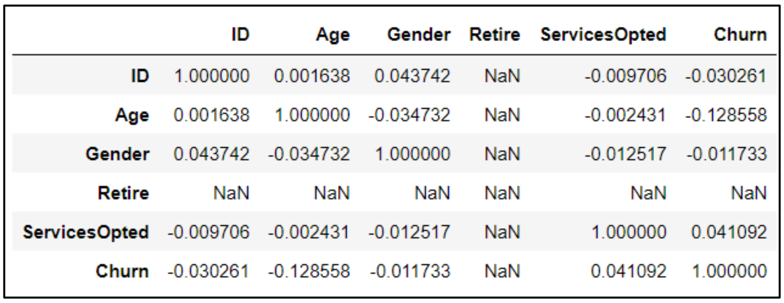


Figure 5 Bar chart for Percentage of the Missing Values

2.1.3. CORRELATION BETWEEN THE ATTRIBUTES

In Python, the correlation function can be used to find the correlation strength among the attributes. The correlation matrix ranges from +1 to -1, where +1 is highly and positively correlated, and -1 will be highly negatively correlated. For the churn attribute, ServicesOpted positively correlated with 0.041092 compared to the other attributes with negative values.



	ID	Age	Gender	Retire	ServicesOpted	Churn
ID	1.000000	0.001638	0.043742	NaN	-0.009706	-0.030261
Age	0.001638	1.000000	-0.034732	NaN	-0.002431	-0.128558
Gender	0.043742	-0.034732	1.000000	NaN	-0.012517	-0.011733
Retire	NaN	NaN	NaN	NaN	NaN	NaN
ServicesOpted	-0.009706	-0.002431	-0.012517	NaN	1.000000	0.041092
Churn	-0.030261	-0.128558	-0.011733	NaN	0.041092	1.000000

Figure 6 Correlations between the Attributes

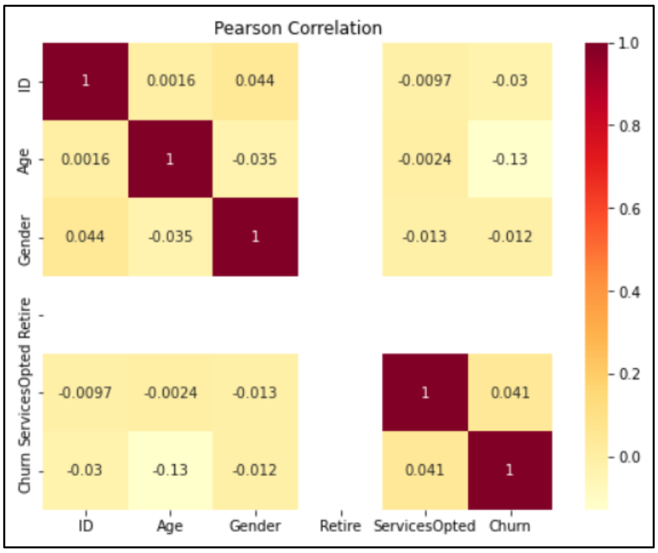


Figure 7 Correlation Heatmap between the Attributes

2.1.4. DISTRIBUTION FOR EACH ATTRIBUTE

First, the age distribution was distributed from age 27 until age 38. However, there is no single entry at age 32. The customer is mostly in the age of 30 while least in age 33.

Most of the Yamani Tour & Travels agency's customers are female, with 539 customers, while the male is 404 customers.

Based on the data, the total of non-retired customers is 954. It is believed that most non-retired customers are comfortable with services provided by Yamani Tour & Travels agency.

Customers who travel with Yamani Tour & Travels agency came from three different income classes: high, middle and low. Most of the customers are in middle income while only some of the customers came from high income annually.

In recent years, some the customers opted from the services. However, the highest number of customer who opted for our services are 399, which is one time. This shows that many of the customers are first-timers using the service. There is also a total of 63 who opted for the services as many as six times. This shows that the customer is happy and love the services provided by Yamani Tour & Travels agency.

Next, most of the company accounts of the user does not synchronized to their social media, with a 583 count. Most customers also did not book hotels or lodging using the company services. They probably use the Yamani Tour & Travels agency, excluding the accommodation.

Lastly, the majority of the customer most likely does not churn and are loyal to the services provided by Yamani Tour & Travels agency. In contrast, 224 customers will most likely churn from the service.

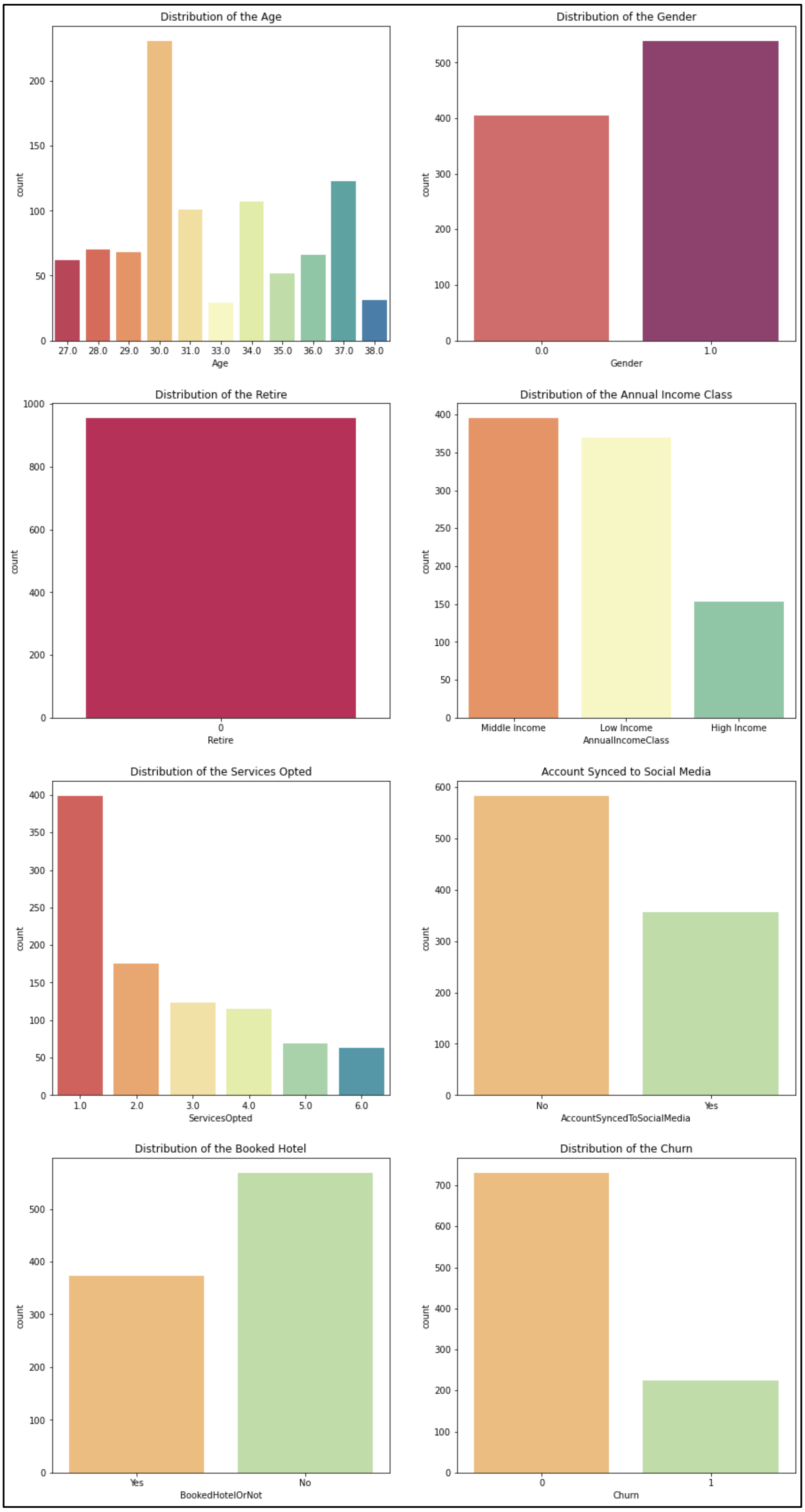


Figure 8 Bar Chart for Each Attributes Count

3. DATA PREPARATION

The next section for this study is the data preparation process. There are a total of three steps of the process used for this dataset: generating the attributes, removing useless attributes and dropping the missing value.

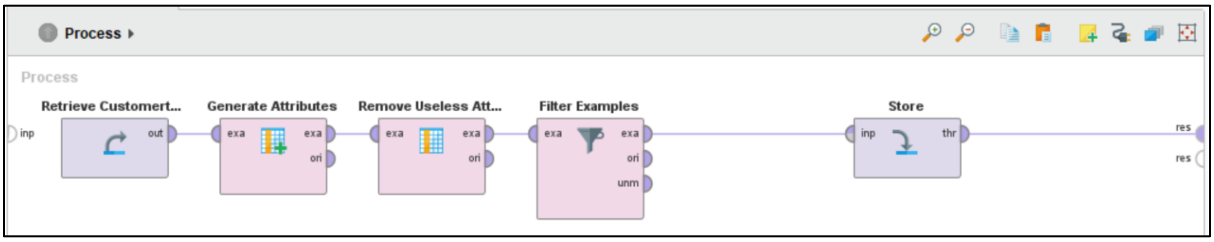
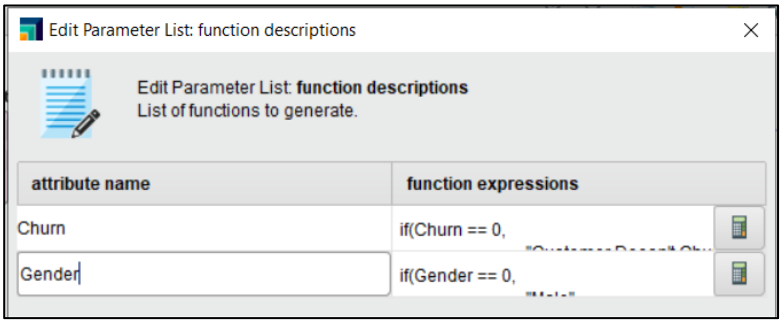


Figure 9 Full Process of Data Preparation

3.1. GENERATE ATTRIBUTES

Firstly, generating the attributes. Since the Churn and Gender annotates with 0 and 1 values, a new attribute created and replaced the existing ones. For Churn, 0 defines the Customer Doesn't Churn while 1 defines the Customer Churn. As for Gender, 0 is for Males while one is for Female. Below is the expression used to generate the attributes for both Gender and Churn attributes. By doing this process, the original data type of both attributes also changed from integer to nominal data type.

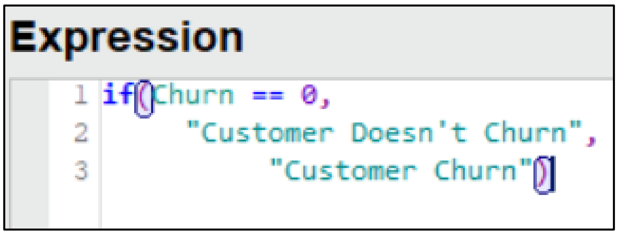


Edit Parameter List: function descriptions

Edit Parameter List: **function descriptions**
List of functions to generate.

attribute name	function expressions
Churn	if(Churn == 0,
Gender	if(Gender == 0,

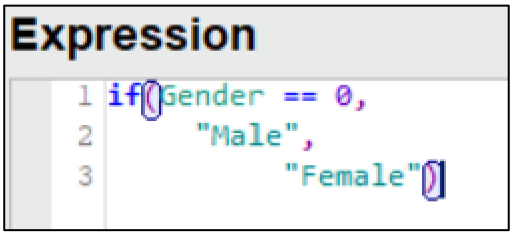
Figure 10 Function Express for Generate Attributes



Expression

```
if(Churn == 0,
        "Customer Doesn't Churn",
                "Customer Churn")
```

Figure 11 Expressions for Churn



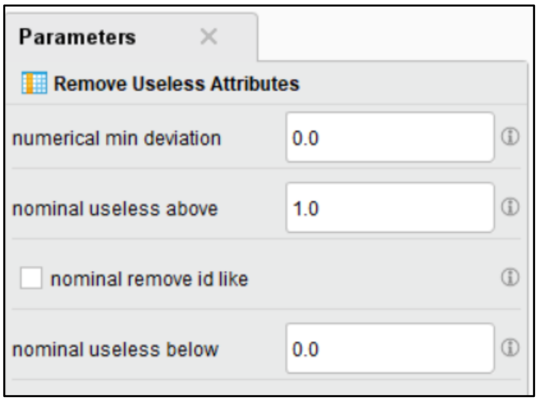
Expression

```
if(Gender == 0,
        "Male",
                "Female")
```

Figure 12 Expression for Gender

3.2. DROP USELESS ATTRIBUTE

Next, remove the useless attributes. This process uses the Remove Useless Attributes operator. As a result, an attribute named Retire was removed from the dataset. Retire was removed because the attributes only contain “0” as their value. The amount of 0 values in the dataset will eventually affect the data and make the data biased



Parameters

Remove Useless Attributes

numerical min deviation 0.0

nominal useless above 1.0

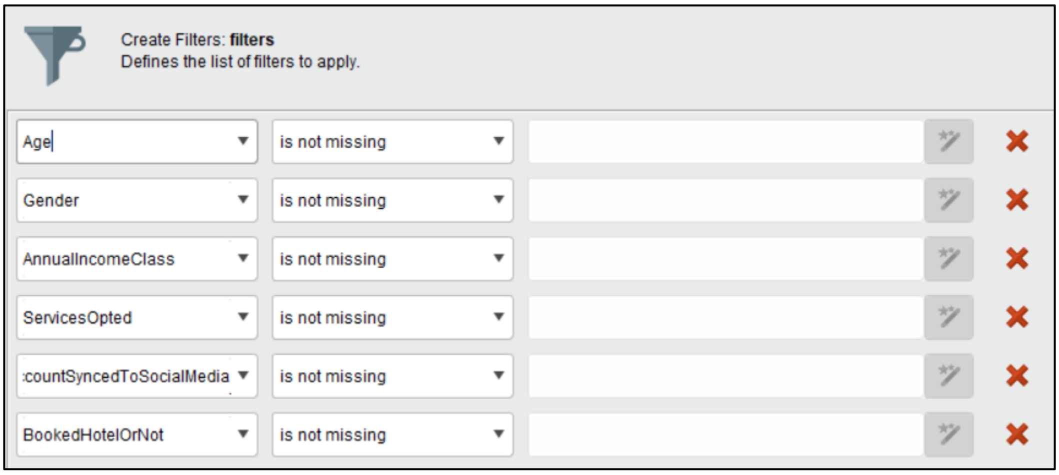
□ nominal remove id like

nominal useless below 0.0

Figure 13 Parameter for Remove Useless Attributes

3.3. DROP THE MISSING VALUES

Lastly, drop the missing value using Filter Example Operator. There are many ways to handle the dataset's missing value, such as dropping and replacing the missing with mode, median, and mean. However, this study chooses to drop the missing value. It is because the percentage of missing values is lower for all dataset attributes. The highest percentage of the missing value is 3.77% and it still does not over the limit of maximum percentage missing values which is 30%.



Create Filters: **filters**
Defines the list of filters to apply.

Age	is not missing
Gender	is not missing
AnnualIncomeClass	is not missing
ServicesOpted	is not missing
:countSyncedToSocialMedia	is not missing
BookedHotelOrNot	is not missing

Figure 14 Filters for Drop the Missing Value

4. MODELLING

4.1. LOGISTIC REGRESSION

4.1.1. LOGISTIC REGRESSION

Logistic Regression is a classification algorithm used to assign observations to a discrete set of classes. In this class, our set of classes are “Customer Doesn’t Churn” and “Customer Churn”. Model builds a regression model to predict the probability of events to success. A binary logistic regression was built for our study because the dependent variable is dichotomous. Logistic regression also allows the mixture of qualitative and quantitative independents.

Below is the full process of Logistic Regression in RapidMiner. There are a total of 6 operators used in the process.

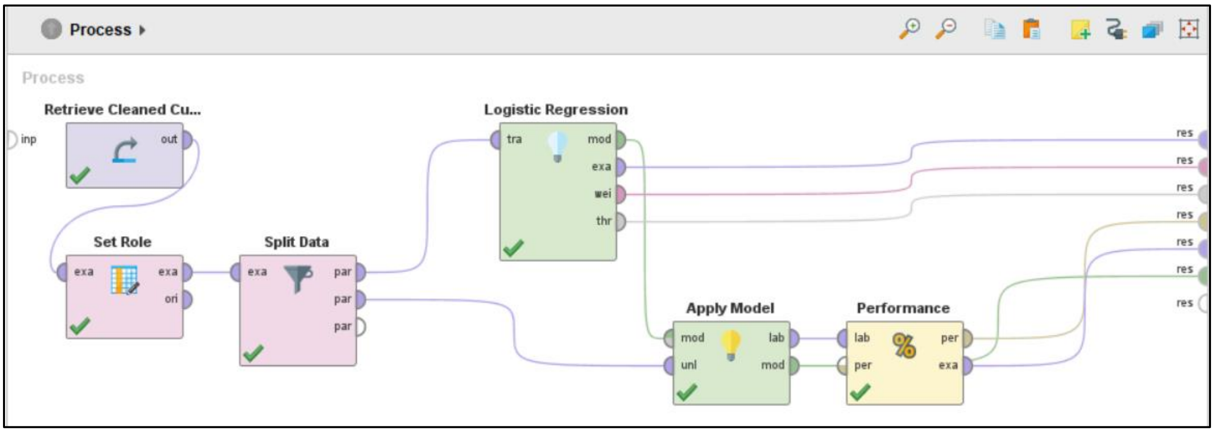
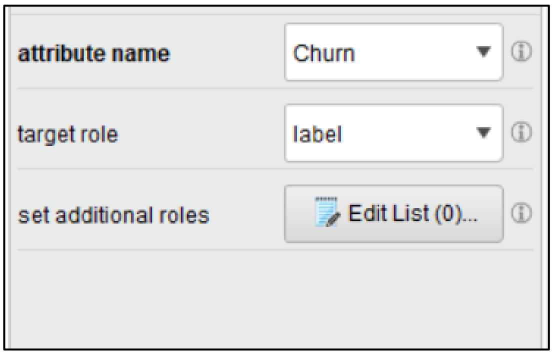


Figure 15 Full Process for Logistic Regression Modelling

Firstly, the cleaned data from the previous process is retrieved into the process space, and connected with the set role operator. In a set role operator, the target attribute “Churn” was labeled as a label role. Thus, the attribute becomes a special attribute. By setting a label for the target attribute, the evaluation of prediction were made based on the attribute. This allows the model to make and compare the predictions in performance of the model.



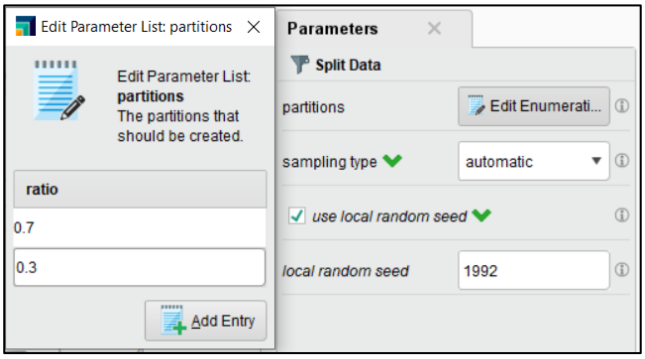
attribute name Churn

target role label

set additional roles Edit List (0)...

Figure 16 Set Role Operator

By using only one set of datasets, the data need to be split into training and testing data. In supervised learning, a prediction on the dataset is made using the train the model on labeled training data. A ratio of 70:30 used in splitting this data set into training and testing, respectively. The training set is used to train the model, while the testing set is used to evaluate the trained model's performance.



Edit Parameter List: partitions

Edit Parameter List: **partitions**
The partitions that should be created.

ratio
0.7
0.3

partitions

sampling type: automatic

☑ *use local random seed*

local random seed: 1992

Figure 17 Split Data

Next, a Logistic Regression model operator and Apply Model used for the modelling process. Meanwhile, the Performance operator is where the performance of the model executed. In this study, it executed the accuracy and the confusion matrix of the model.

4.1.2. LOGISTIC REGRESSION (FORWARD ELIMINATION)

Another way to perform feature selection in logistic regression is by using a forward selection technique. Forward selection starts with an empty model and then adds one feature at a time, incrementally improving the model until the performance reaches a certain threshold or no further improvements can be made. This allows the model to choose only the most relevant features for prediction.

Overall, feature selection can help improve the performance and interpretability of logistic regression models by selecting only the most relevant features for prediction.

The same process were used in this model with the addition of “Forward Selection” operator.

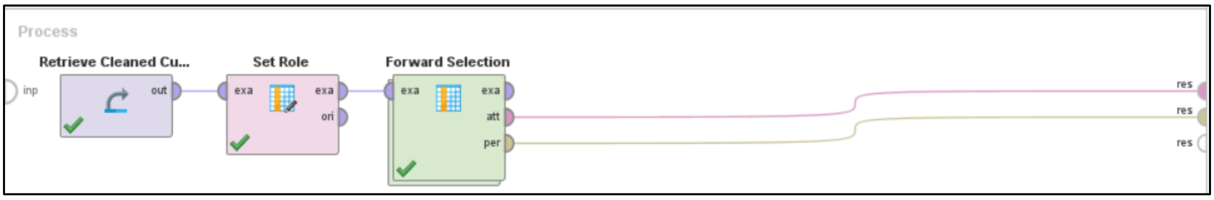


Figure 18 Full Process of Logistic Regression in Forward Selection

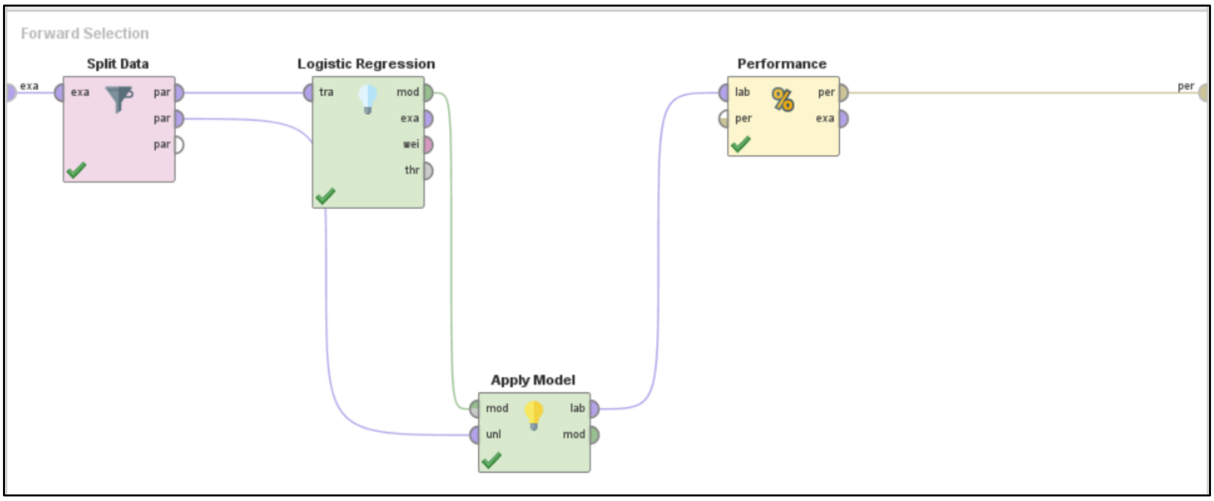


Figure 19 Full Process of Logistic Regression in Forward Selection

4.1.3. LOGISTIC REGRESSION (BACKWARD ELIMINATION)

Backward elimination is a stepwise procedure that starts with all the features included in the model and then removes the least significant feature at each step, until only the most relevent features remain. Backward elimination helps improve the performance and interpretability of logistic regression models by selecting only the most relevant features for prediction. It can also help reduce the risk of overfitting by removing irrelevant or redundant features.

Here is the process of the logistic regression using backward elimination as the feature selection:

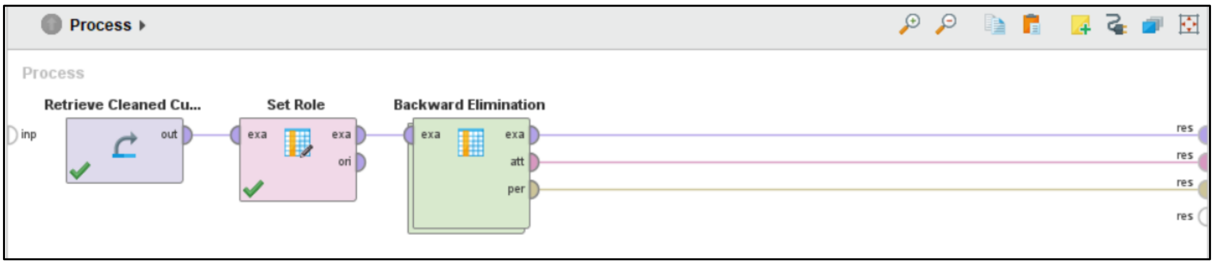


Figure 20 Full Process of Logistic Regression in Backward Elimination

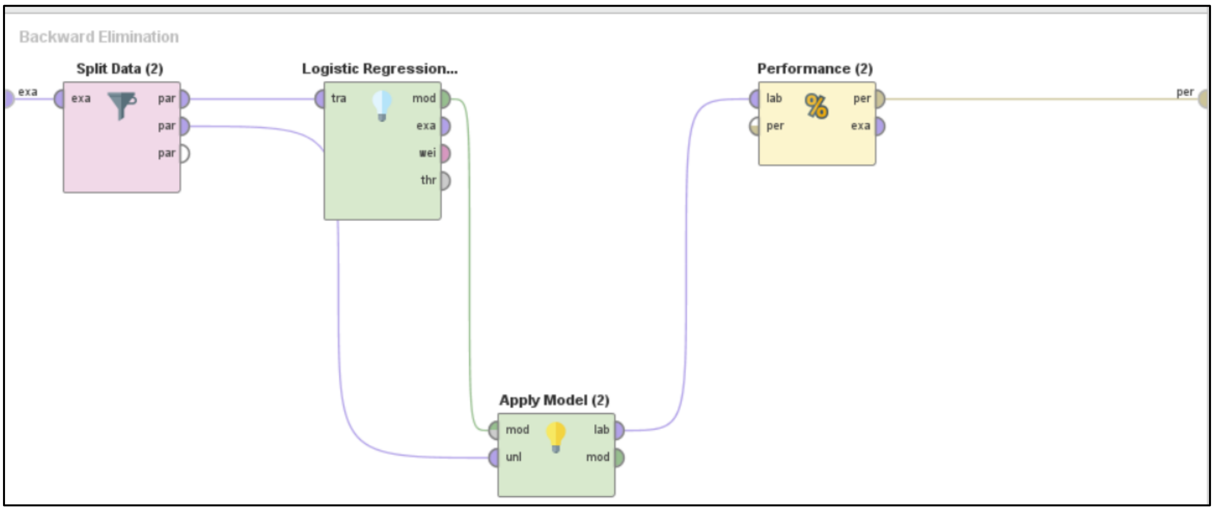


Figure 21 Full Process of Logistic Regression in Backward Elimination

4.1.4. LOGISTIC REGRESSION (OPTIMIZE SELECTION – EVOLUTIONARY)

The next feature selection used is optimization selection (evolutionary). This selection works by selecting the most effective predictor variables for a logistic regression model and then applying a set of rules to evolve over multiple variables until a satisfactory evaluation between the predictor variables and the binary outcome is found. Using an evolutionary algorithm for feature selection can help optimize the selection of features for logistic regression, leading to improved model performance and interpretability.

Here is the process of the logistic regression using the optimize selection as the feature selection:

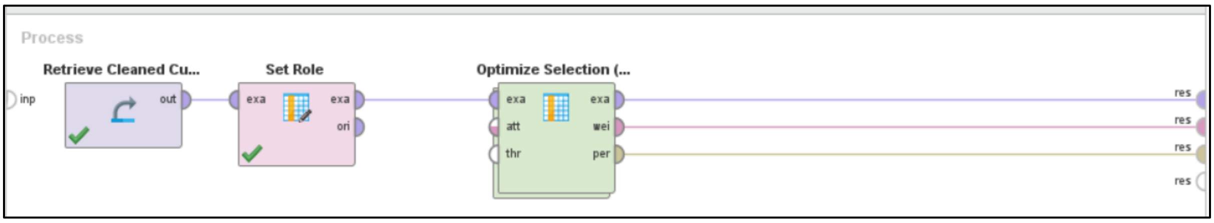


Figure 22 Full Process of Logistic Regression in Optimized Selection (Evolutionary)

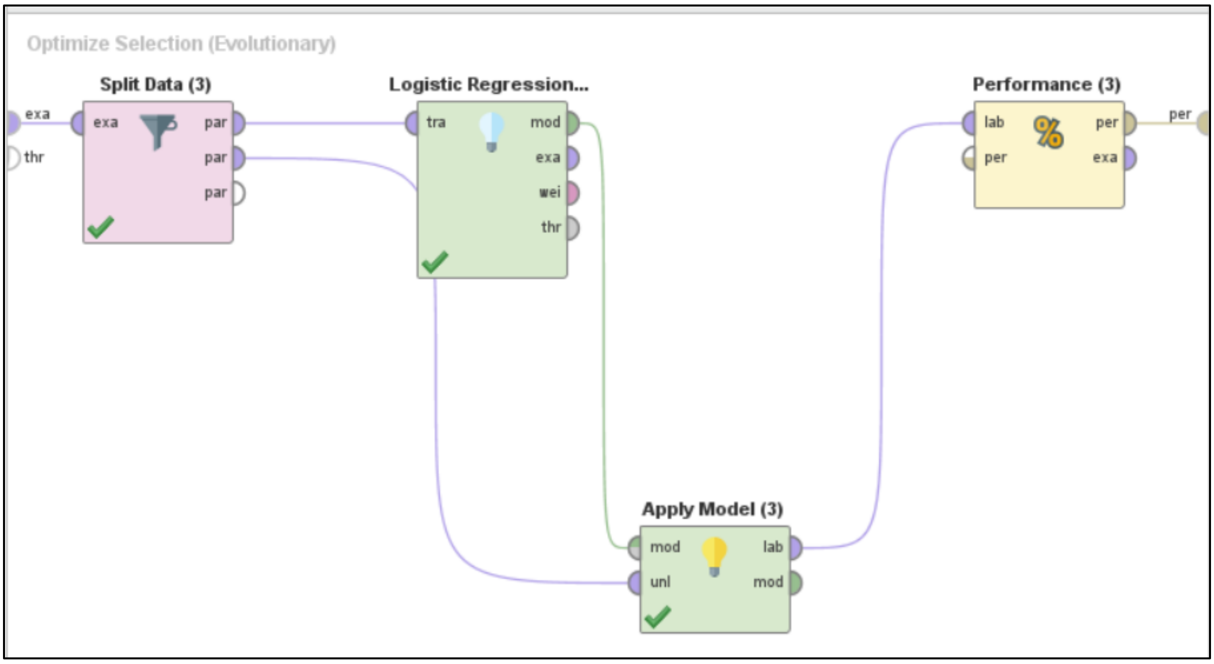


Figure 23 Full Process of Logistic Regression in Optimized Selection (Evolutionary)

4.2. DECISION TREE

A decision tree (DT) algorithm is commonly used in a prediction. It works by dividing the training data into smaller and smaller subsets based on the values of the features until each subset contains samples from only one class. There are three options for splitting in DT: Boolean, nominal, and continuous.

Below is the full process of Decision Tree modeling in RapidMiner. There are a total of 6 operators used in the process.

The same process used in Logistic Regression started with retrieving the cleaned dataset, set the label role for the target variable, “Churn.” Next, split the data into 70:30 for the training and testing set. An operator of Decision Tree is used in this modeling with different splitting criterions such as Gini index, entropy (information gain), and gain ratio. The model is applied using the Apply Model operator and evaluated utilizing the Performance operator.

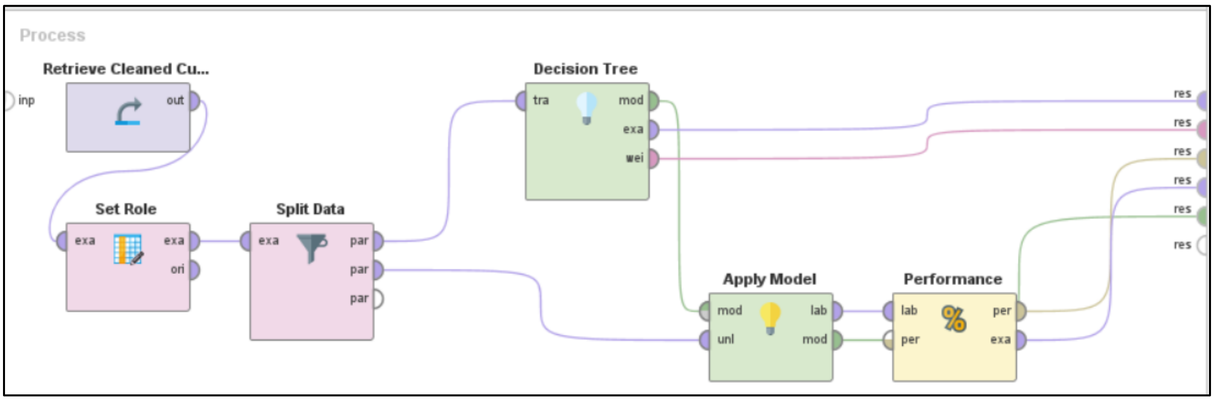
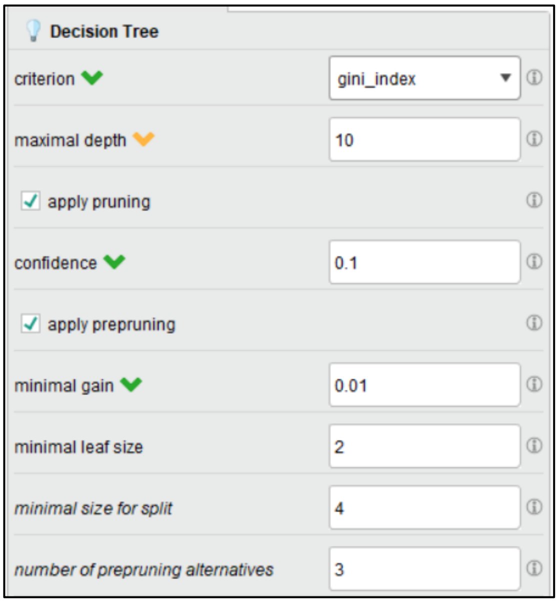


Figure 24 Full Process of Decision Tree Modelling

4.2.1. DECISION TREE (GINI INDEX)

The Gini index measures how well a decision tree can separate the training data into distinct classes. It is commonly used as a metric for evaluating the performance of a decision tree. Gini Index measures the inequality of distribution of label characteristics. Using the Gini index as a metric for decision trees can help ensure that the tree can accurately predict the class for each sample. It can also help avoid overfitting by penalizing decision trees that use splits too specific to the training data.



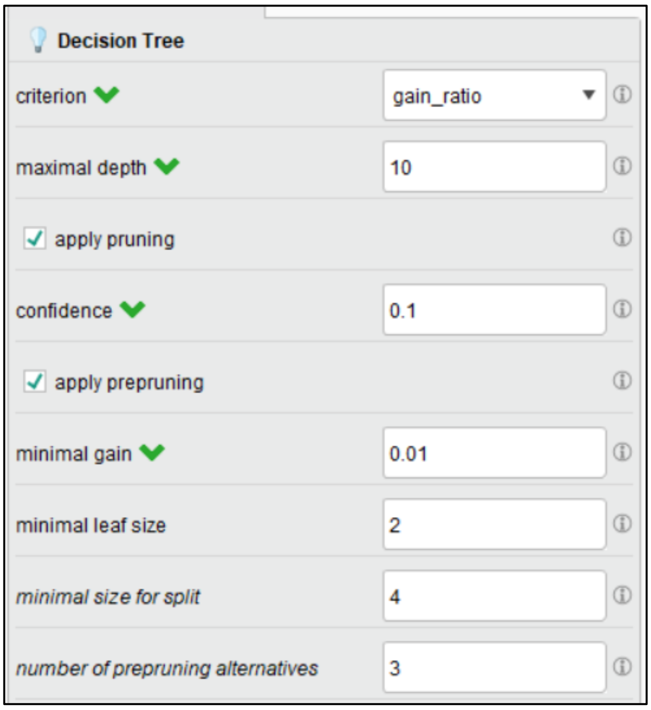
Decision Tree

criterion ✓	gini_index ▾
maximal depth ✓	10
☑ apply pruning	
confidence ✓	0.1
☑ apply prepruning	
minimal gain ✓	0.01
minimal leaf size	2
minimal size for split	4
number of prepruning alternatives	3

Figure 25 Decision Tree Modelling with Gini Index Criterion

4.2.2. DECISION TREE (GAIN RATIO)

The gain ratio criterion is a variant of information gain that adjust the information gain for each attribute to reduce the biased of the data. The gain ratio modified the problem with information gain based on the number of branches that would result before doing the split.



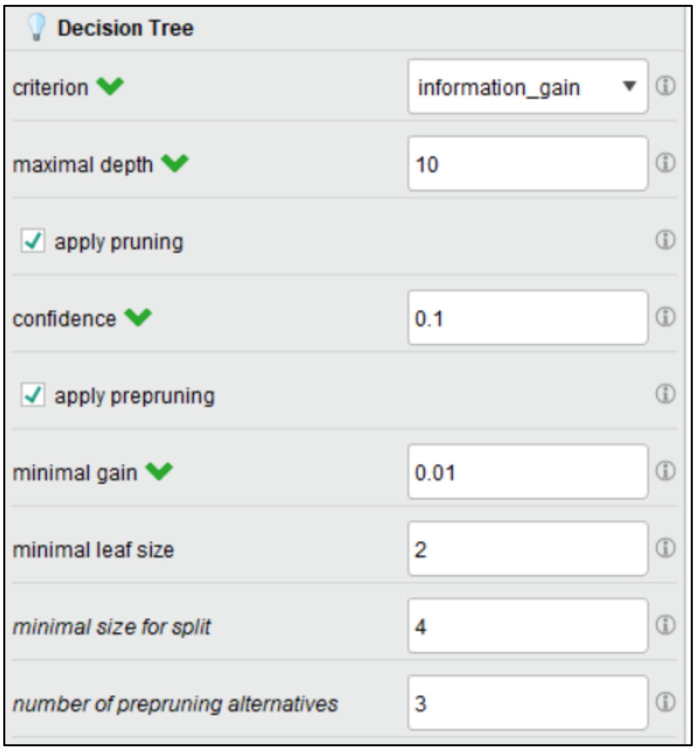
Decision Tree

criterion ✓	gain_ratio ▾
maximal depth ✓	10
☑ apply pruning	
confidence ✓	0.1
☑ apply prepruning	
minimal gain ✓	0.01
minimal leaf size	2
minimal size for split	4
number of prepruning alternatives	3

Figure 26 Decision Tree Modelling with Gain Ratio Criterion

4.2.3. DECISION TREE (INFORMATION GAIN)

The entropies of the attributes are calculated, and attributes with the least entropies or highest information gains are selected for a split for the current node. A low entropy indicates that the data labels are uniform.



Decision Tree

criterion	information_gain
maximal depth	10
☑ apply pruning	
confidence	0.1
☑ apply prepruning	
minimal gain	0.01
minimal leaf size	2
minimal size for split	4
number of prepruning alternatives	3

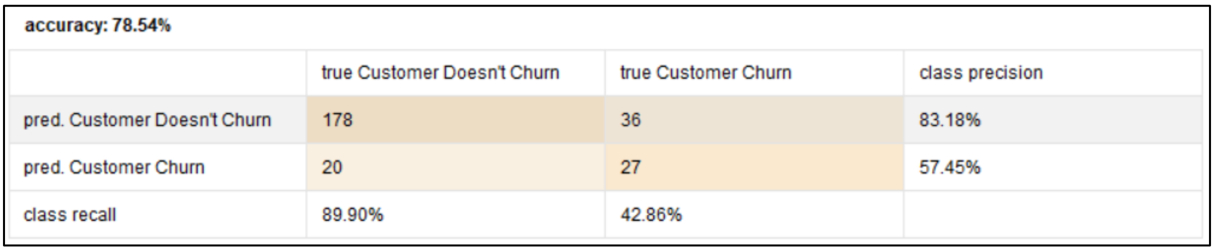
Figure 27 Decision Tree Modelling with Information Gain Criterion

5. EVALUATION

5.1. LOGISTIC REGRESSION

5.1.1. LOGISTIC REGRESSION

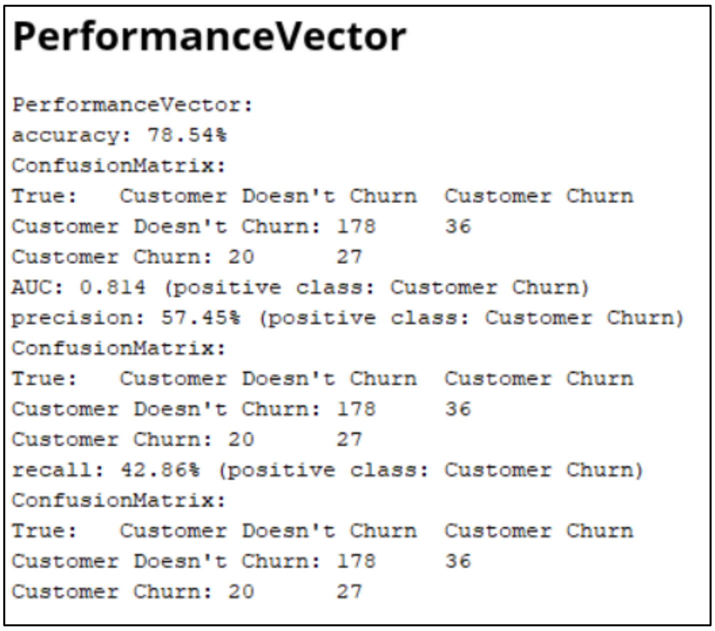
This is the result of the Logistic Regression. The accuracy of the prediction of customer churning is 78.54%, while the precision and recall are 57.45% and 42.86%, respectively. The AUC score is 0.814.



accuracy: 78.54%

	true Customer Doesn't Churn	true Customer Churn	class precision
pred. Customer Doesn't Churn	178	36	83.18%
pred. Customer Churn	20	27	57.45%
class recall	89.90%	42.86%	

Figure 28 Accuracy for Logistic Regression



PerformanceVector

```
PerformanceVector:
accuracy: 78.54%
ConfusionMatrix:
True:   Customer Doesn't Churn  Customer Churn
Customer Doesn't Churn: 178     36
Customer Churn: 20      27
AUC: 0.814 (positive class: Customer Churn)
precision: 57.45% (positive class: Customer Churn)
ConfusionMatrix:
True:   Customer Doesn't Churn  Customer Churn
Customer Doesn't Churn: 178     36
Customer Churn: 20      27
recall: 42.86% (positive class: Customer Churn)
ConfusionMatrix:
True:   Customer Doesn't Churn  Customer Churn
Customer Doesn't Churn: 178     36
Customer Churn: 20      27
```

Figure 29 Performance Vector for Logistic Regression

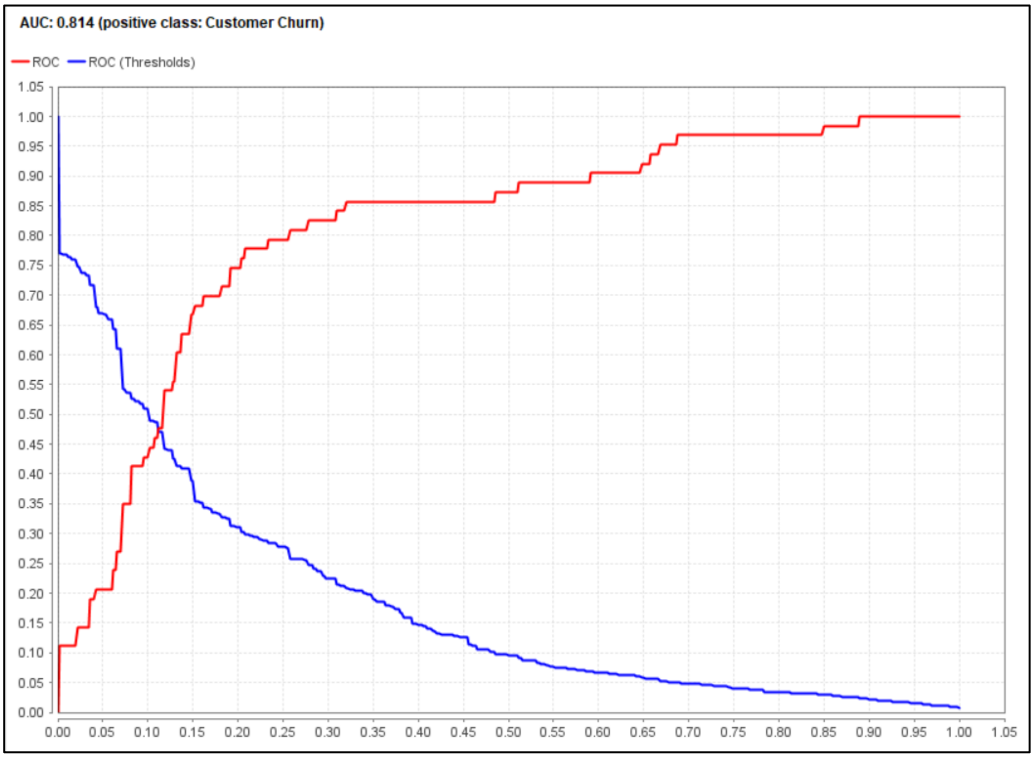
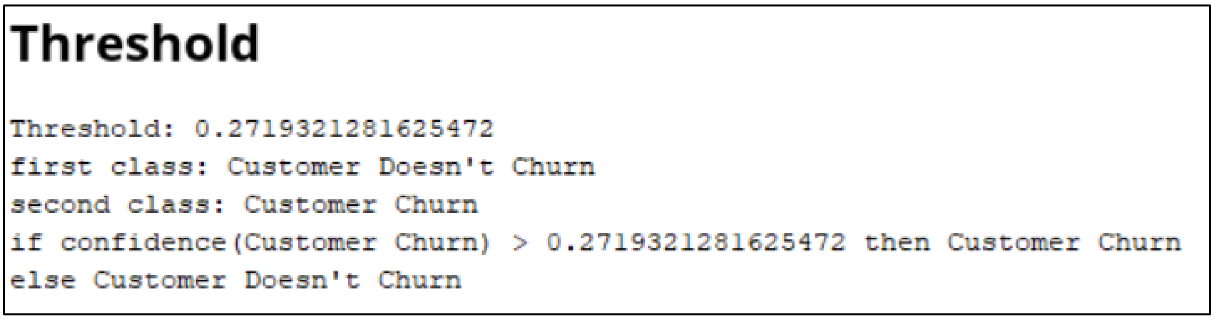


Figure 30 AUC score and ROC Curve for Logistic Regression

The threshold output for the logistic regression model is 0.27. If the value is more than 0.27, the new customer will churn. All of the independent variables are significantly influenced the "Churn" except for Gender.Female because the p-Value is 0.917 above the threshold value. In this model, the final model of logistic regression is as per below:

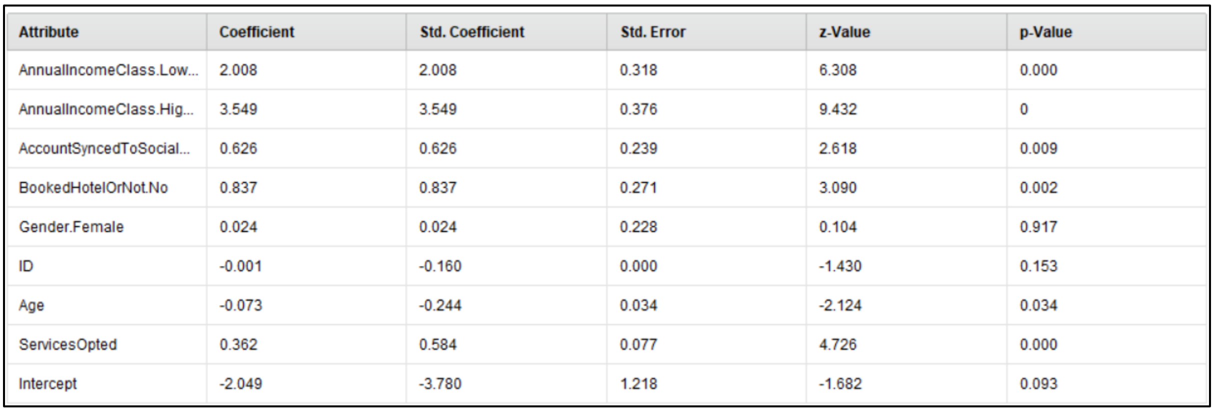
$$\ln\left(\frac{p}{1-p}\right) = -2.049 + 2.008Low + 3.549High + 0.626AccountSyncedToSocialMedia \\ + 0.837BookedHotelOrNot - 0.001ID - 0.073Age + 0.362ServicesOpted$$



Threshold

```
Threshold: 0.2719321281625472
first class: Customer Doesn't Churn
second class: Customer Churn
if confidence(Customer Churn) > 0.2719321281625472 then Customer Churn
else Customer Doesn't Churn
```

Figure 31 Threshold value for Logistic Regression



Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
AnnualIncomeClass.Low...	2.008	2.008	0.318	6.308	0.000
AnnualIncomeClass.Hig...	3.549	3.549	0.376	9.432	0
AccountSyncedToSocial...	0.626	0.626	0.239	2.618	0.009
BookedHotelOrNot.No	0.837	0.837	0.271	3.090	0.002
Gender.Female	0.024	0.024	0.228	0.104	0.917
ID	-0.001	-0.160	0.000	-1.430	0.153
Age	-0.073	-0.244	0.034	-2.124	0.034
ServicesOpted	0.362	0.584	0.077	4.726	0.000
Intercept	-2.049	-3.780	1.218	-1.682	0.093

Figure 32 Logistic Regression Statistical Measure

In logistic regression, each attribute or feature is assigned a weight that indicates its importance in predicting the model's outcome. The greater the weight, the more important the attribute is in predicting the outcome. The weights are determined during the training phase of the model, using an optimization algorithm. The weights are updated iteratively as the algorithm learns from the training data, and the final weights are used to make predictions on new data.

In this model, all attributes are positively weighted except for Age and ID are negatively weighted. Only six attributes were considered in this model.

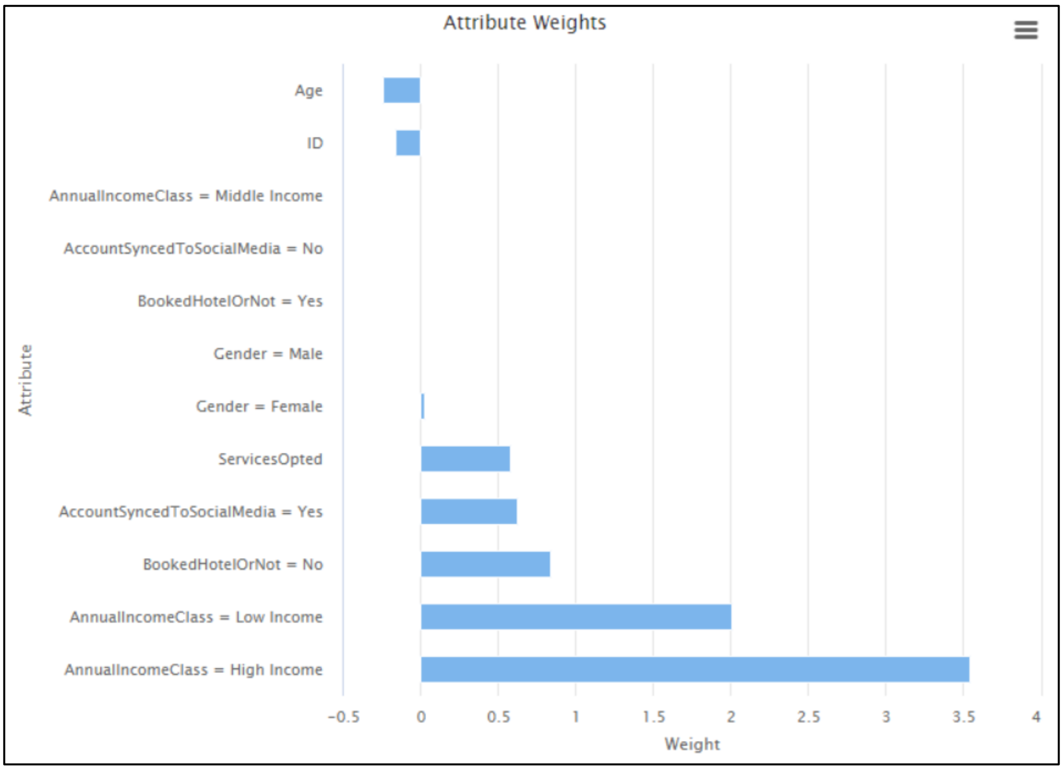
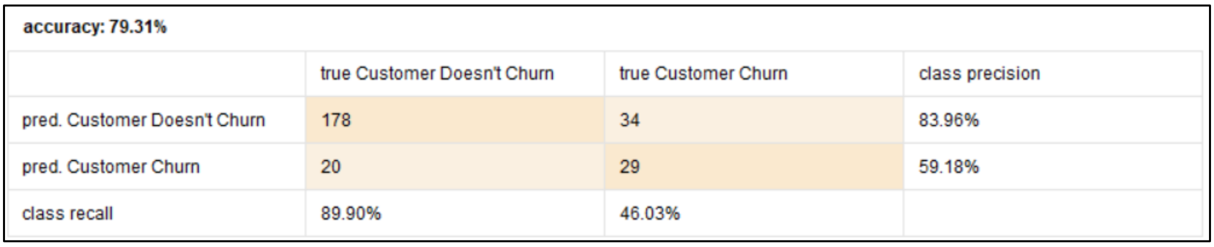


Figure 33 Attribute Weights of Logistic Regression

5.1.2. LOGISTIC REGRESSION (FORWARD SELECTION)

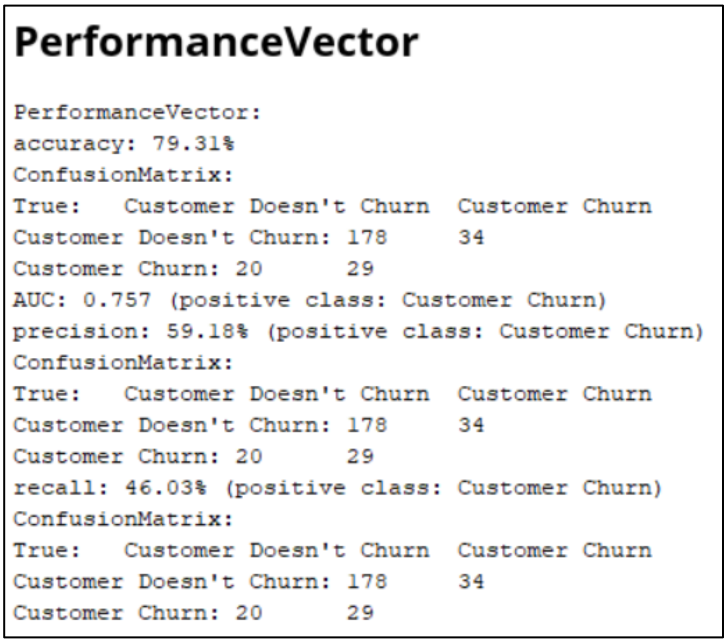
This is the evaluation for logistic regression by using the forward selection feature. The model's accuracy is 79.31%, while the precision and recall values are 59.119% and 46.03%. The AUC score for this model is 0.757. For the attribute weights, only AnnualIncomeClass is weighted for the attribute. Thus, AnnualIncomeClass is the most crucial attribute considered in this model.



accuracy: 79.31%

	true Customer Doesn't Churn	true Customer Churn	class precision
pred. Customer Doesn't Churn	178	34	83.96%
pred. Customer Churn	20	29	59.18%
class recall	89.90%	46.03%	

Figure 34 Accuracy for Logistic Regression with FS



PerformanceVector

```
PerformanceVector:
accuracy: 79.31%
ConfusionMatrix:
True:   Customer Doesn't Churn  Customer Churn
Customer Doesn't Churn: 178     34
Customer Churn: 20      29
AUC: 0.757 (positive class: Customer Churn)
precision: 59.18% (positive class: Customer Churn)
ConfusionMatrix:
True:   Customer Doesn't Churn  Customer Churn
Customer Doesn't Churn: 178     34
Customer Churn: 20      29
recall: 46.03% (positive class: Customer Churn)
ConfusionMatrix:
True:   Customer Doesn't Churn  Customer Churn
Customer Doesn't Churn: 178     34
Customer Churn: 20      29
```

Figure 35 Performance Vector of Logistic Regression with FS

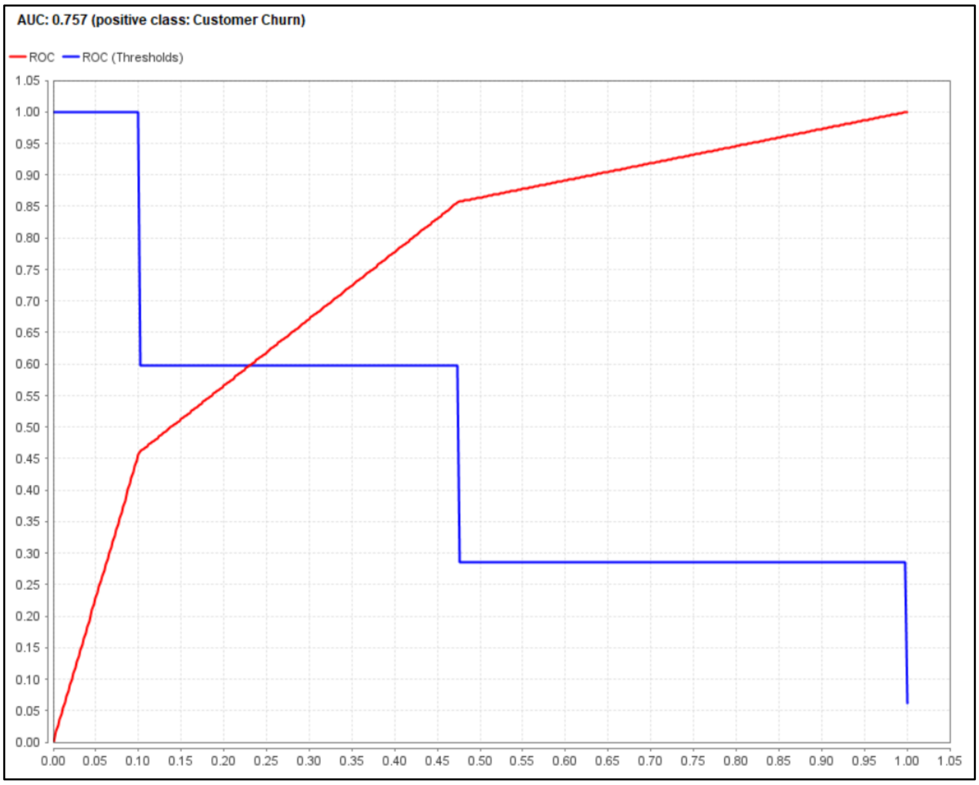


Figure 36 AUC score and ROC Curve for Logistic Regression with FS

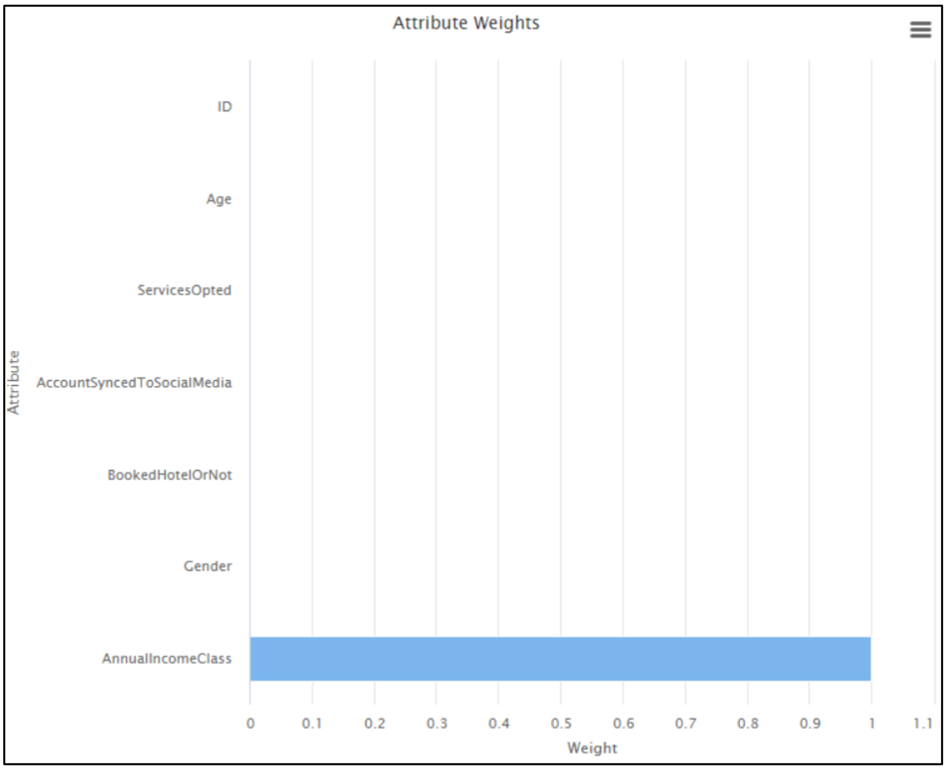
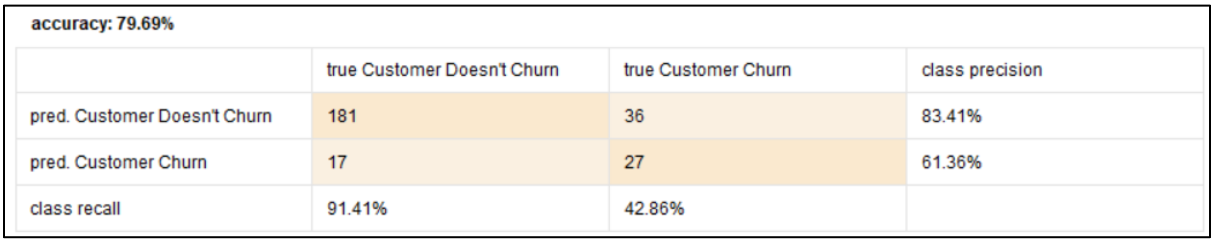


Figure 37 Attribute Weights of Logistic Regression with FS

5.1.3. LOGISTIC REGRESSION (BACKWARD ELIMINATION)

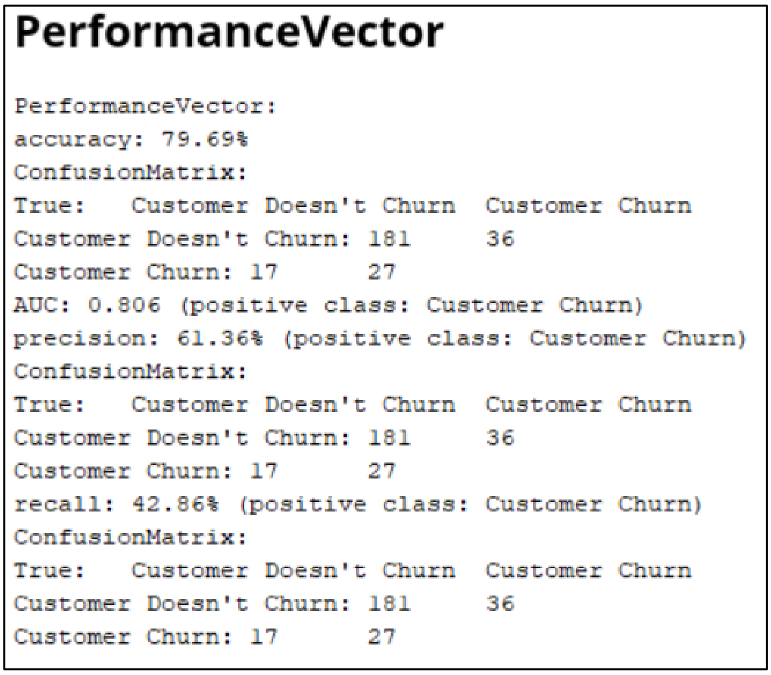
This is the evaluation for logistic regression by using the backward elimination feature selection. The model's accuracy is 79.69%, while the precision and recall are 61.36% and 42.86%, respectively. The AUC score is 0.806. For the attribute weights, only ID, Age, AnnualIncomeClass, ServicesOpted, AccountSyncedToSocialMedia, and Gender were positively weighted for the attribute weight. Thus, most of the attributes were important in this model.



accuracy: 79.69%

	true Customer Doesn't Churn	true Customer Churn	class precision
pred. Customer Doesn't Churn	181	36	83.41%
pred. Customer Churn	17	27	61.36%
class recall	91.41%	42.86%	

Figure 38 Accuracy for Logistic Regression with BE



PerformanceVector

```
PerformanceVector:
accuracy: 79.69%
ConfusionMatrix:
True:   Customer Doesn't Churn  Customer Churn
Customer Doesn't Churn: 181     36
Customer Churn: 17      27
AUC: 0.806 (positive class: Customer Churn)
precision: 61.36% (positive class: Customer Churn)
ConfusionMatrix:
True:   Customer Doesn't Churn  Customer Churn
Customer Doesn't Churn: 181     36
Customer Churn: 17      27
recall: 42.86% (positive class: Customer Churn)
ConfusionMatrix:
True:   Customer Doesn't Churn  Customer Churn
Customer Doesn't Churn: 181     36
Customer Churn: 17      27
```

Figure 39 Performance Vector of Logistic Regression with BE

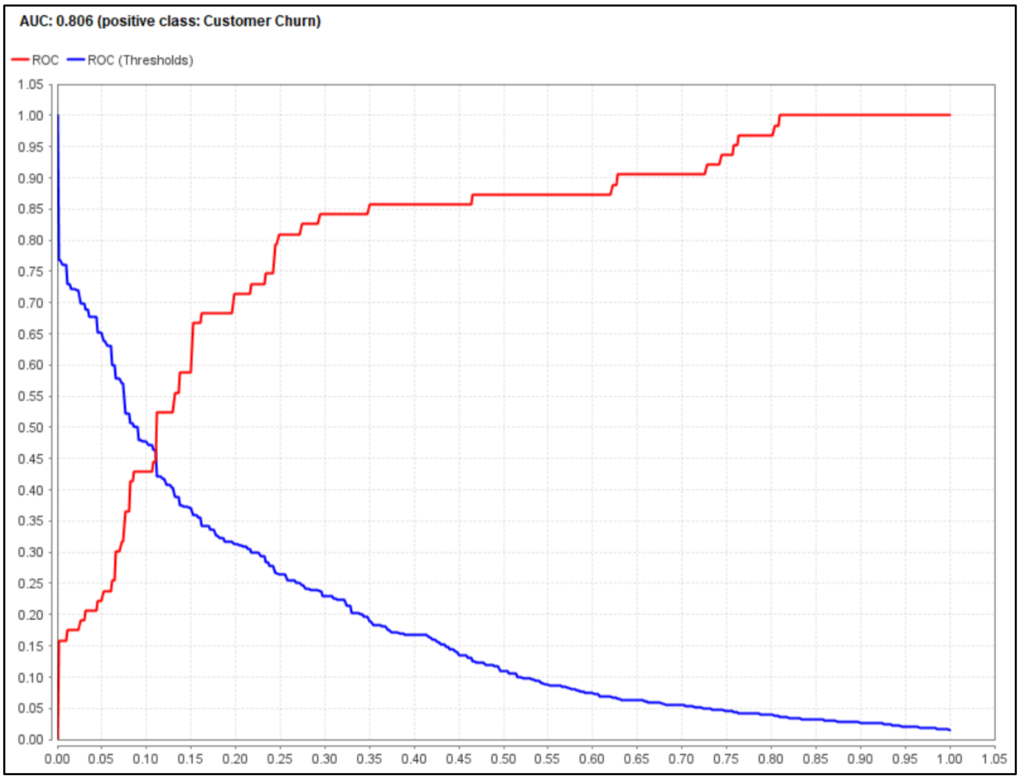


Figure 40 AUC score and ROC Curve for Logistic Regression with BE

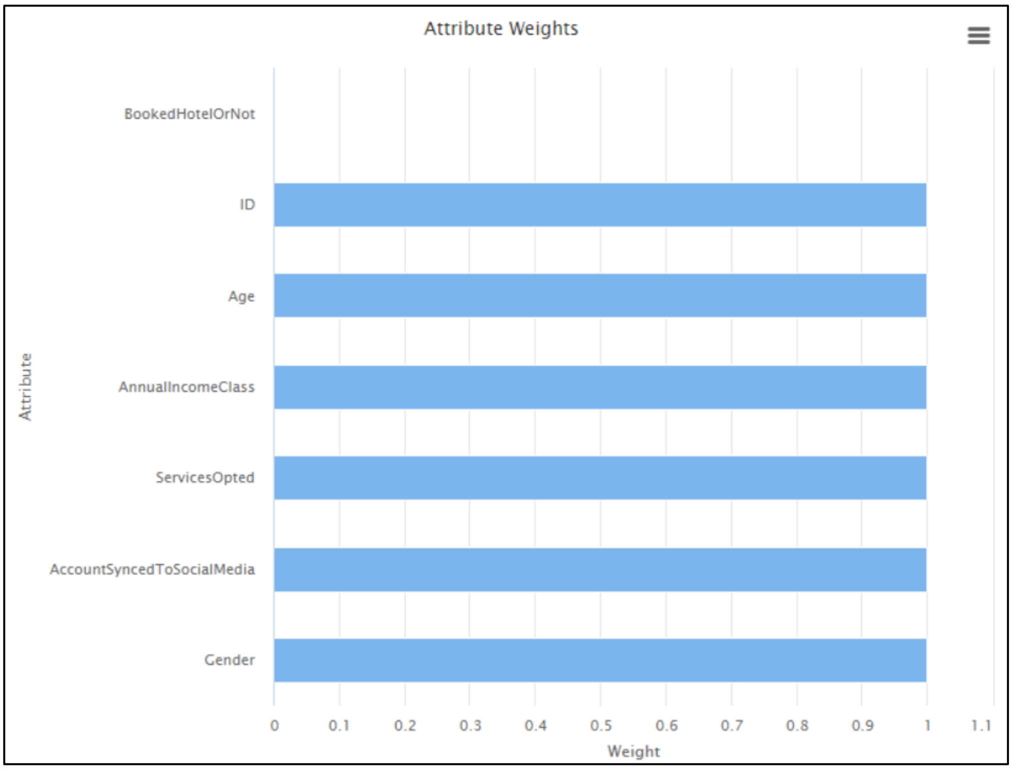
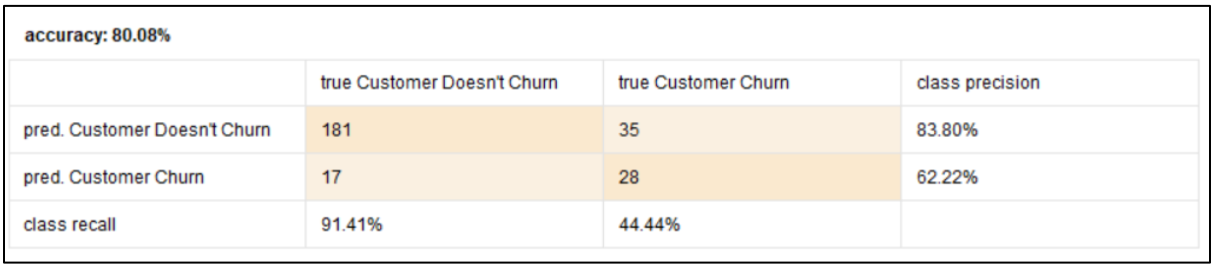


Figure 41 Attribute Weights of Logistic Regression with BE

5.1.4. LOGISTIC REGRESSION (OPTIMIZED SELECTION – EVOLUTIONARY)

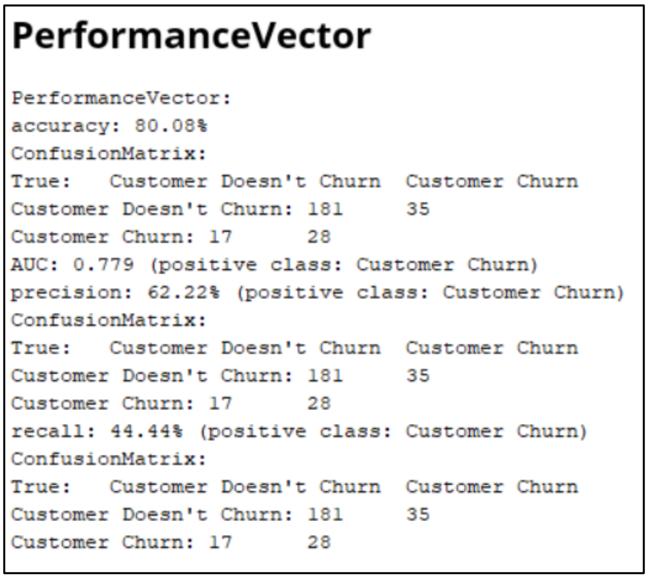
This is the evaluation for logistic regression by using the optimized selection feature selection. The model's accuracy is 80.08%, while the precision and the recall values are 62.22% and 44.44%. The AUC score for this model is 0.779. Four attributes are weighted as 1 in this model: ID, AnnualIncomeClass, AccountSyncedToSocialMedia and Gender.



accuracy: 80.08%

	true Customer Doesn't Churn	true Customer Churn	class precision
pred. Customer Doesn't Churn	181	35	83.80%
pred. Customer Churn	17	28	62.22%
class recall	91.41%	44.44%	

Figure 42 Accuracy for Logistic Regression with OS



PerformanceVector

```
PerformanceVector:
accuracy: 80.08%
ConfusionMatrix:
True:   Customer Doesn't Churn  Customer Churn
Customer Doesn't Churn: 181     35
Customer Churn: 17      28
AUC: 0.779 (positive class: Customer Churn)
precision: 62.22% (positive class: Customer Churn)
ConfusionMatrix:
True:   Customer Doesn't Churn  Customer Churn
Customer Doesn't Churn: 181     35
Customer Churn: 17      28
recall: 44.44% (positive class: Customer Churn)
ConfusionMatrix:
True:   Customer Doesn't Churn  Customer Churn
Customer Doesn't Churn: 181     35
Customer Churn: 17      28
```

Figure 43 Performance Vector of Logistic Regression with OS

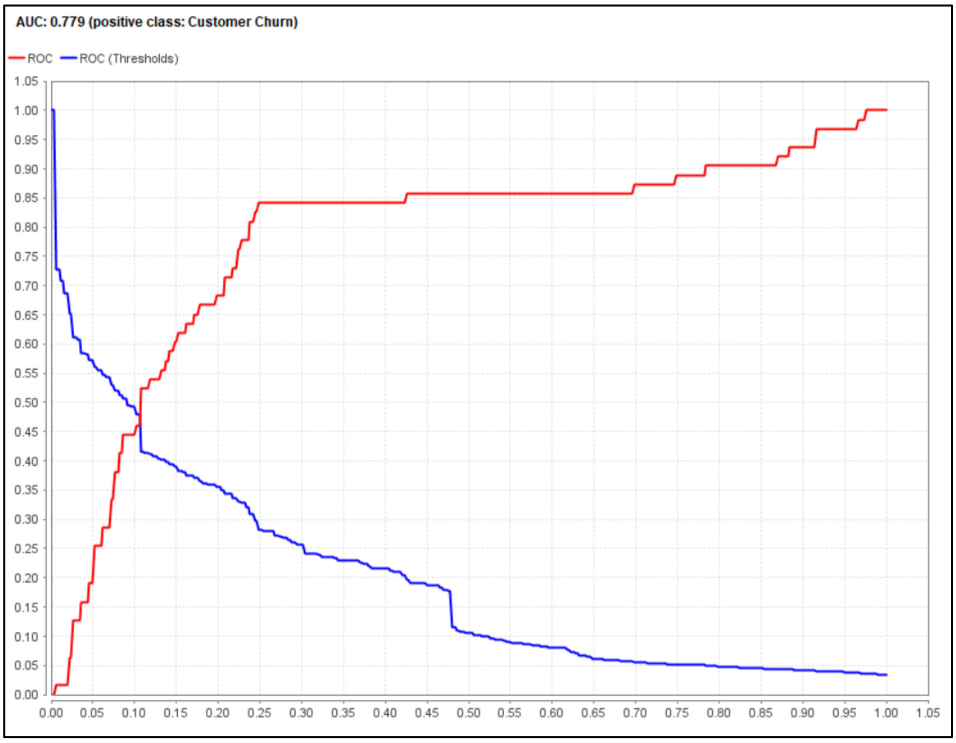


Figure 44 AUC score and ROC Curve for Logistic Regression with OS

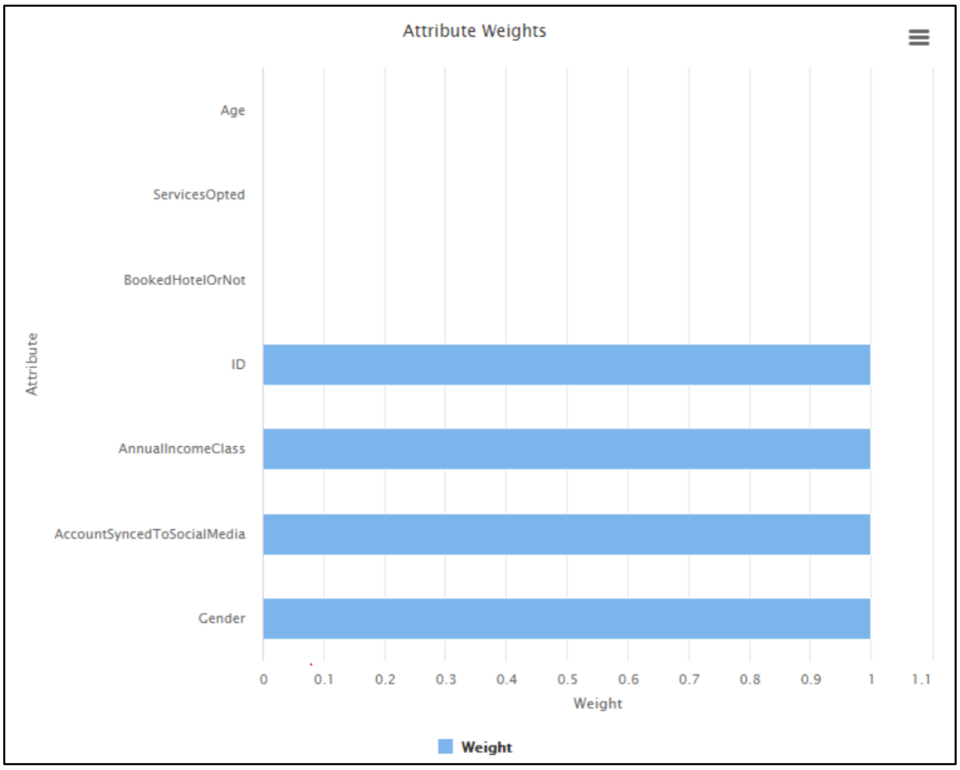


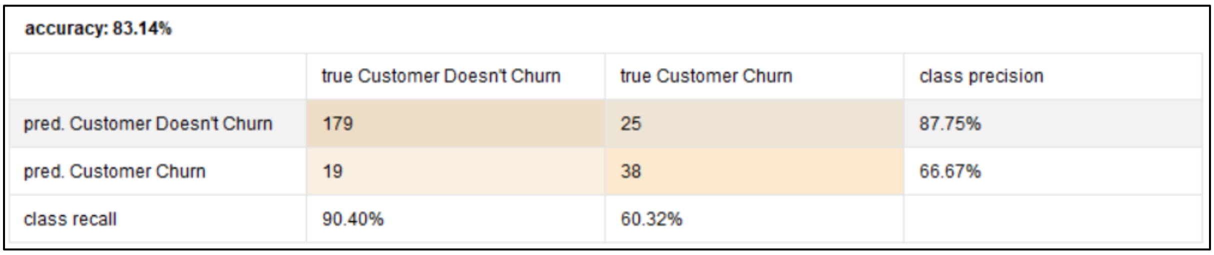
Figure 45 Attribute Weights of Logistic Regression with OS

5.2. DECISION TREE

The accuracy of a decision tree in RapidMiner is a measure of how well the tree can make predictions based on the data it has been trained on. It is calculated as the number of correct predictions made by the tree divided by the total number of predictions made. A decision tree with a high accuracy score is able to make more accurate predictions than a tree with a lower accuracy score.

5.2.1. DECISION TREE (GINI INDEX)

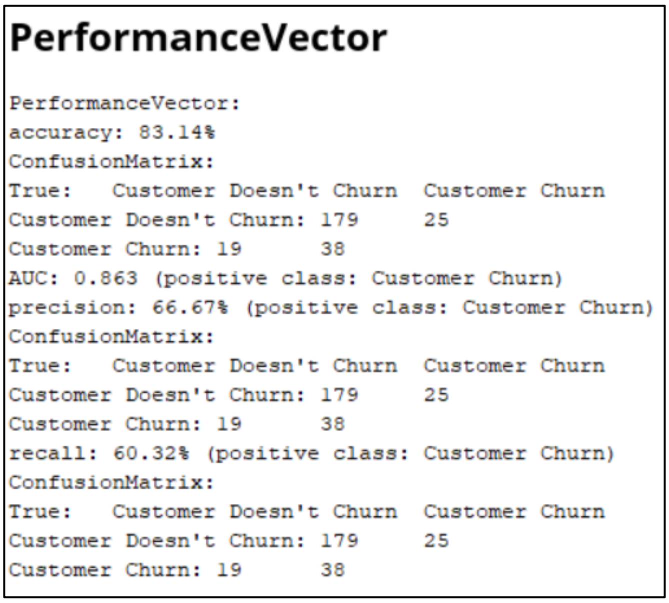
The accuracy for the decision tree using the Gini index is 83.14%, while the precision and recall are 66.67% and 60.32%, respectively. The AUC score for this model is 0.863. Most of the attributes are positively weighted in the model. These weights indicate the relative importance of each attribute in making the predictions.



accuracy: 83.14%

	true Customer Doesn't Churn	true Customer Churn	class precision
pred. Customer Doesn't Churn	179	25	87.75%
pred. Customer Churn	19	38	66.67%
class recall	90.40%	60.32%	

Figure 46 Accuracy of DT with Gini Index



PerformanceVector

```
PerformanceVector:
accuracy: 83.14%
ConfusionMatrix:
True:   Customer Doesn't Churn  Customer Churn
Customer Doesn't Churn: 179     25
Customer Churn: 19      38
AUC: 0.863 (positive class: Customer Churn)
precision: 66.67% (positive class: Customer Churn)
ConfusionMatrix:
True:   Customer Doesn't Churn  Customer Churn
Customer Doesn't Churn: 179     25
Customer Churn: 19      38
recall: 60.32% (positive class: Customer Churn)
ConfusionMatrix:
True:   Customer Doesn't Churn  Customer Churn
Customer Doesn't Churn: 179     25
Customer Churn: 19      38
```

Figure 47 Performance Vector of DT with Gini Index

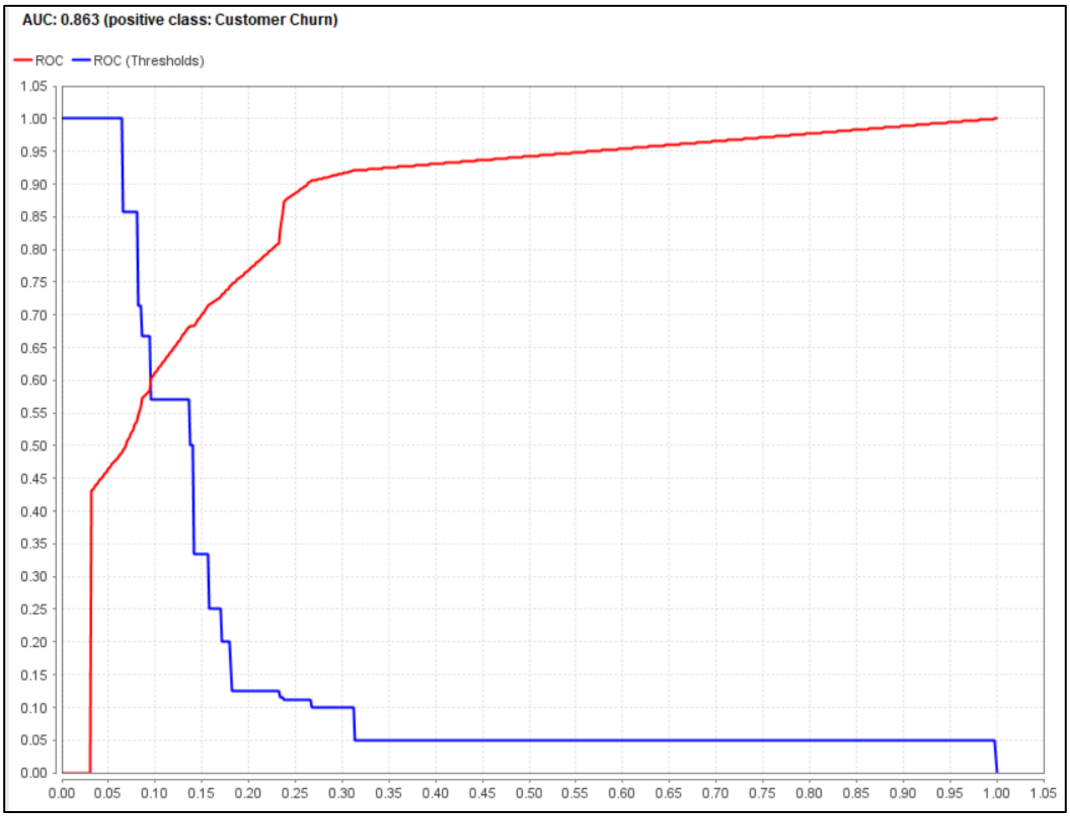


Figure 48 AUC Score and ROC Curve of DT with Gini Index

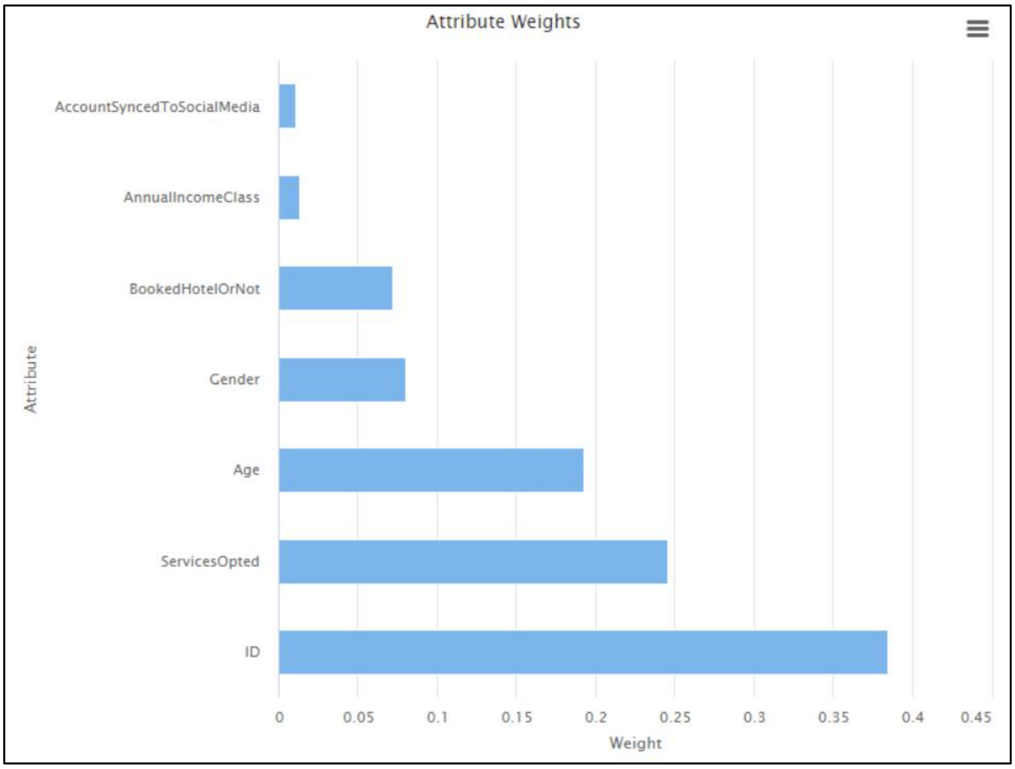
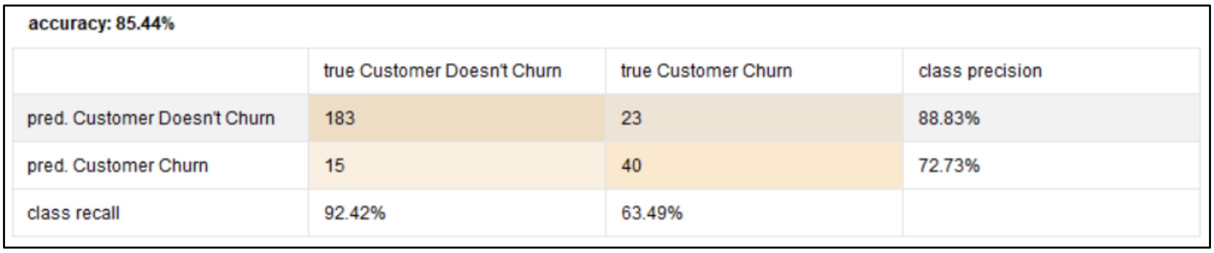


Figure 49 Attribute Weights of DT with Gini Index

5.2.2. DECISION TREE (GAIN RATIO)

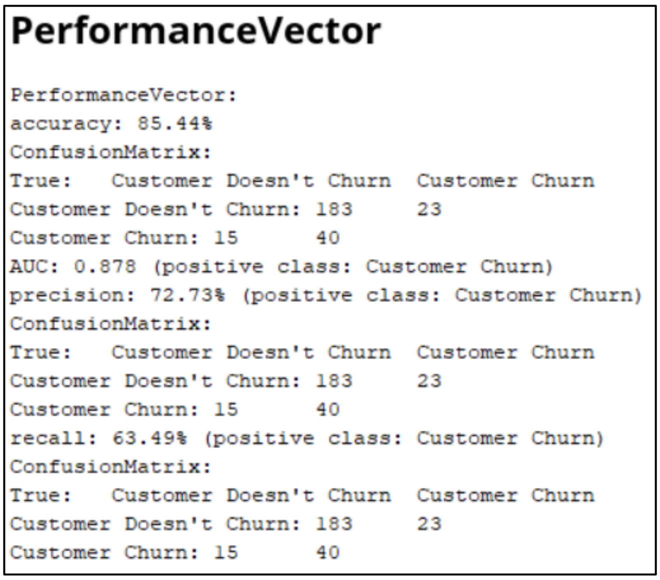
The accuracy for the decision tree using the gain ratio criterion is 85.44%, while the precision and recall are 72.73% and 63.49%, respectively. The AUC score for this model is 0.878. AccountSyncedToSocialMedia, AnnualIncomeClass, Gender, BookedHotelOrNot, Age, ServiesOpted and ID were positively weighted as all of the attributes are important in predicting the model.



accuracy: 85.44%

	true Customer Doesn't Churn	true Customer Churn	class precision
pred. Customer Doesn't Churn	183	23	88.83%
pred. Customer Churn	15	40	72.73%
class recall	92.42%	63.49%	

Figure 50 Accuracy of DT with Gain Ratio



PerformanceVector

```
PerformanceVector:
accuracy: 85.44%
ConfusionMatrix:
True:   Customer Doesn't Churn  Customer Churn
Customer Doesn't Churn: 183     23
Customer Churn: 15      40
AUC: 0.878 (positive class: Customer Churn)
precision: 72.73% (positive class: Customer Churn)
ConfusionMatrix:
True:   Customer Doesn't Churn  Customer Churn
Customer Doesn't Churn: 183     23
Customer Churn: 15      40
recall: 63.49% (positive class: Customer Churn)
ConfusionMatrix:
True:   Customer Doesn't Churn  Customer Churn
Customer Doesn't Churn: 183     23
Customer Churn: 15      40
```

Figure 51 Performance Vector of DT with Gain Ratio

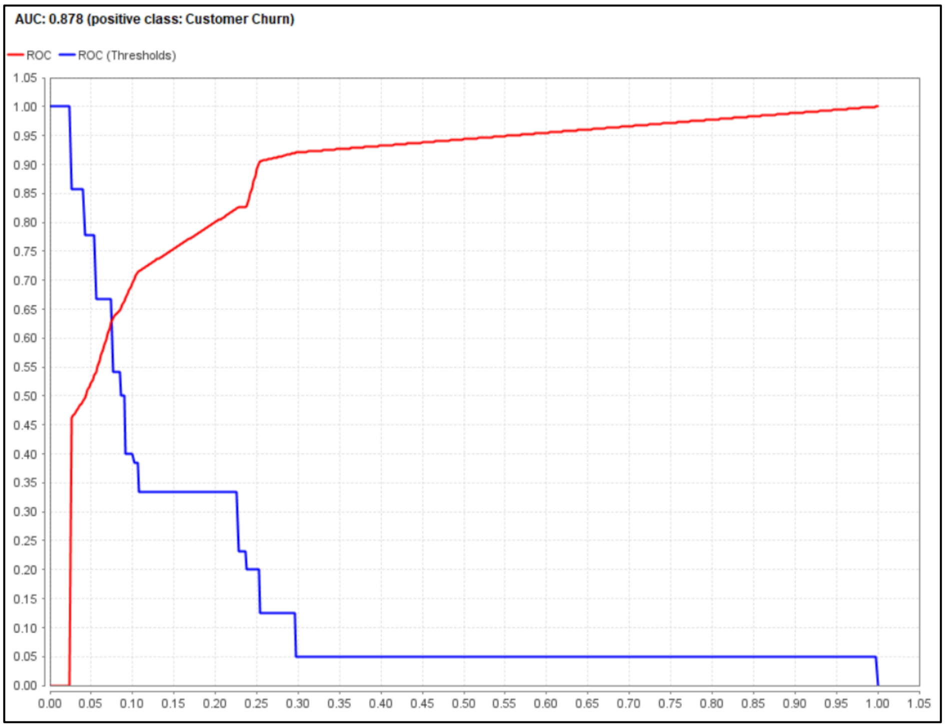


Figure 52 AUC Score and ROC Curve of DT with Gain Ratio

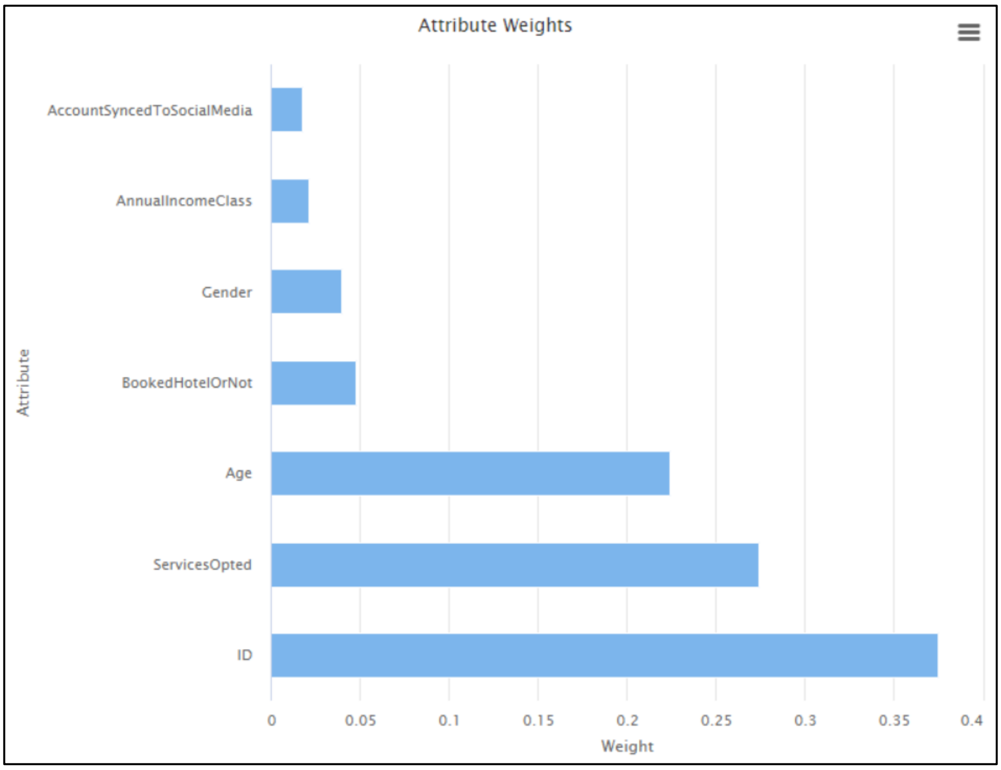
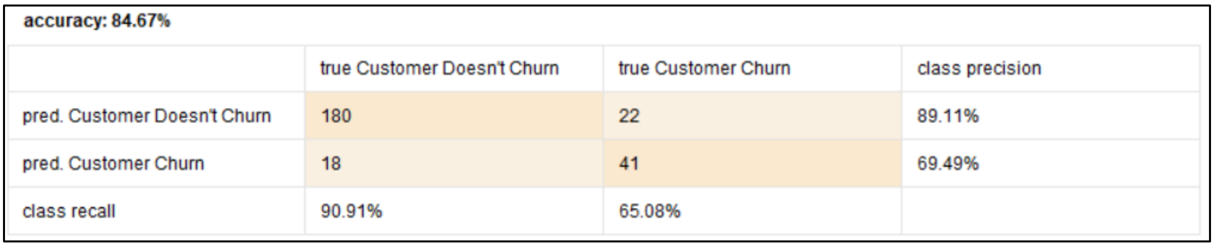


Figure 53 Attribute Weights of DT with Gain Ratio

5.2.3. DECISION TREE (INFORMATION GAIN)

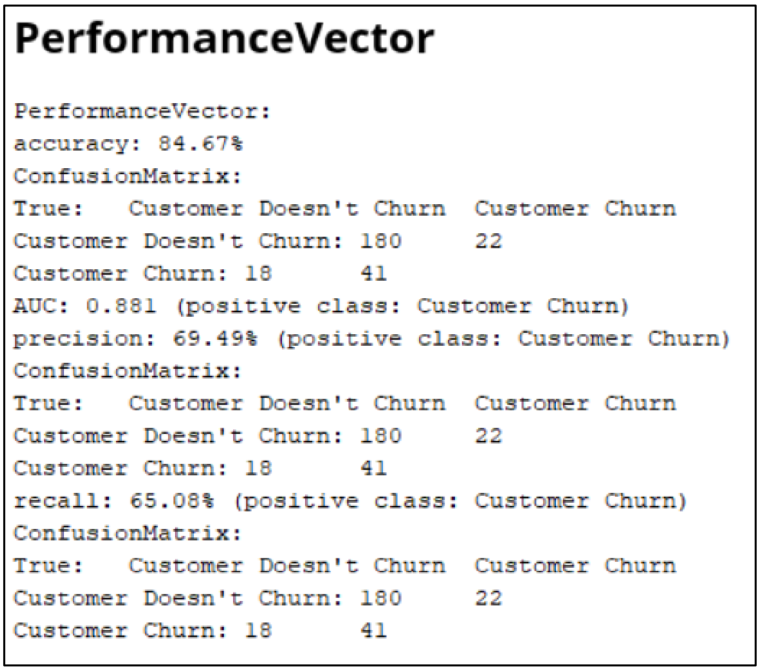
Lastly, the accuracy of the decision tree information gain criterion is 84.67%, while the precision and recall are 69.49% and 65.08%, respectively. The AUC score for this model is 0.881. Age, ID, and ServicesOpted attributes with a slight difference of weights are the most important attributes in this model followed by the least important ones, which are AccountSyncedToSocialMedia, AnnualIncomeClass, Gender, and BookedHotelOrNot.



accuracy: 84.67%

	true Customer Doesn't Churn	true Customer Churn	class precision
pred. Customer Doesn't Churn	180	22	89.11%
pred. Customer Churn	18	41	69.49%
class recall	90.91%	65.08%	

Figure 54 Accuracy of DT with Information Gain



PerformanceVector

```
PerformanceVector:
accuracy: 84.67%
ConfusionMatrix:
True:   Customer Doesn't Churn  Customer Churn
Customer Doesn't Churn: 180     22
Customer Churn: 18      41
AUC: 0.881 (positive class: Customer Churn)
precision: 69.49% (positive class: Customer Churn)
ConfusionMatrix:
True:   Customer Doesn't Churn  Customer Churn
Customer Doesn't Churn: 180     22
Customer Churn: 18      41
recall: 65.08% (positive class: Customer Churn)
ConfusionMatrix:
True:   Customer Doesn't Churn  Customer Churn
Customer Doesn't Churn: 180     22
Customer Churn: 18      41
```

Figure 55 Performance Vector of DT with Information Gain

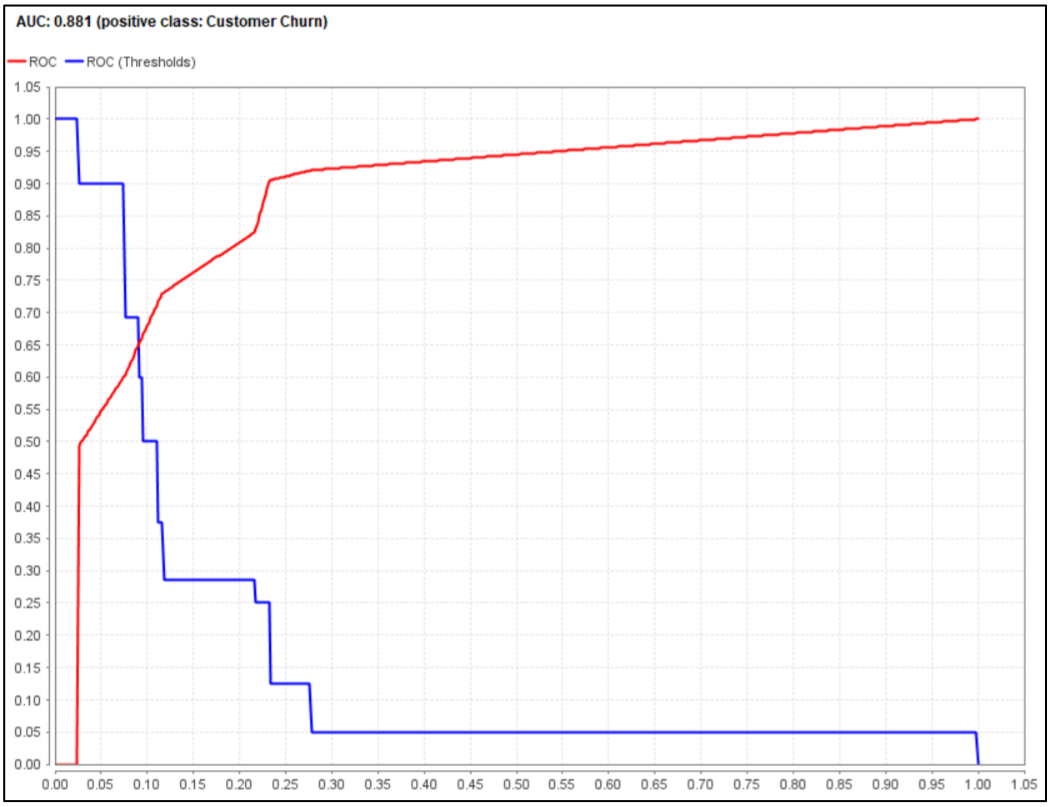


Figure 56 AUC Score and ROC Curve of DT with Information Gain

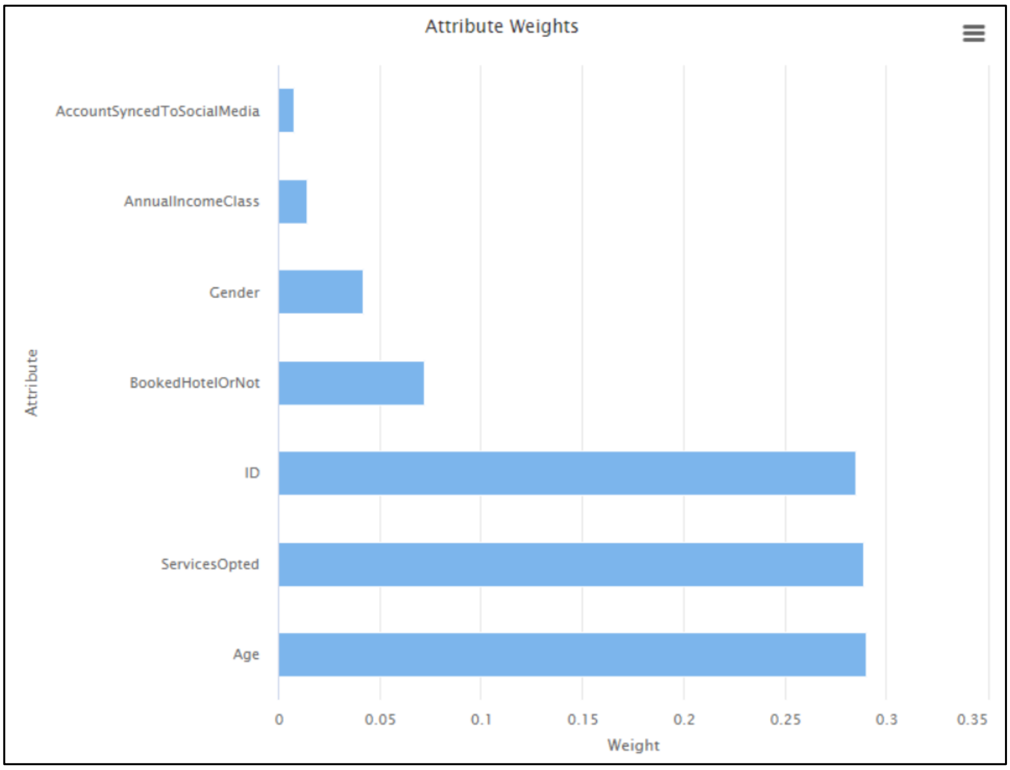


Figure 57 Attributes Weight of DT with Information Gain

5.3. COMPARISONS BETWEEN THE MODELS

Accuracy, precision, recall, and AUC (Area Under the Curve) are all metrics used to evaluate a model's performance. They are all related to one another, but they measure different aspects of a model's performance.

Accuracy is a measure of how well a model correctly predicts the classes of a dataset. It is calculated as the number of correct predictions made by the model divided by the total number of predictions made. Precision measures how well a model only predicts the correct class. It is calculated as the number of true positives divided by the number of true positives plus the number of false positives. Recall measures how well a model can find all of the positive instances in a dataset. It is calculated as the number of true positives divided by the number of true positives plus the number of false negatives.AUC (Area Under the Curve) measures a model's ability to distinguish between positive and negative classes. It is calculated as the area under the ROC (Receiver Operating Characteristic) curve, which plots the true positive rate (recall) against the false positive rate. AUC ranges from 0 to 1, with a higher value indicating a better-performing model.

Generally, a model with high accuracy, precision, recall, and AUC is considered a good model.

Table 2 Comparison Table for Logistic Regression Model

Logistic Regression Model	Accuracy (%)	Precision (%)	Recall (%)	AUC score
Normal	78.54	57.45	42.86	0.814
Forward Selection Feature	79.31	59.18	46.03	0.757
Backward Elimination Feature	79.69	61.36	42.86	0.806
Optimized Feature (Evolutionary)	80.08	62.22	44.44	0.779

The highest accuracy and precision are logistic regression with optimized feature (evolutionary) with 80.08% and 62.22%. Meanwhile, the highest percentage of recall (46.03%) for logistic regression with the forward selection feature. Lastly, the highest AUC score is normal logistic regression, with 0.814.

Based on the four models use, which are normal logistic regression and logistic regression with various feature selections, the logistic regression model with optimized feature (evolutionary) is the best model for predicting customer churn. It is because both of the important metric measures for modelling are the accuracy, and the precision of this model is higher compared to the other model with 80.08% and 62.22%. The percentage of the recall for optimized feature (evolutionary) is 44.44%.

Table 3 Comparison Table for Decision Tree Models

Decision Tree Model	Accuracy (%)	Precision (%)	Recall (%)	AUC score
Gini Index	83.14	66.67	60.32	0.863
Gain Ratio	85.44	72.73	63.49	0.878
Information Gain	84.67	69.49	65.08	0.881

The highest accuracy and precision are decision trees with gain ratio criterion of 85.44% and 72.73%. Meanwhile, the highest percentage of recall (65.08%) and AUC score (0.881) are from decision tree with information gain.

Based on the comparison table of decision tree with different criterions above, decision tree model with the gain ratio is the best model in predicting the customer’s churn. Two out of four metric measure for predictive modelling for this model are the highest with 85.44% and 72.73% for accuracy and precision metrics. This model's recall percentage is 63.49% while the AUC score is 0.878.

6. CONCLUSION

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a widely used data mining methodology that provides a structured approach to planning, conducting, and evaluating data mining projects. The methodology consists of six major phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Each phase includes specific steps and tasks that help guide the data mining process and ensure that the results are reliable and useful.

In this study, the business understanding is about the customers' churn. The objective of building the predictive modelling in predicting the customers' churn. The most accurate model is the best predictive model for customer churn. The tools used in this study is Python for exploratory data analysis and RapidMiner for data preparation and modelling.

The information and exploratory data analysis are mentioned in the data understanding process. Next, some of the process being done in the data preparation process, such as generating attributes, removing useless attributes and dropping the missing values.

In the next phase, two major modelling were built: logistic regression and decision tree. Logistion regression modelling with the addition of various feature selections was built. In decision tree, three criteria of decision tree use are gini index, information gain and gain ratio.

The results of the modelling process were evaluated in a evaluation phase. Logistic regression with optimized selection (evolutionary) is the best model of logistic regression while decision tree with gain ratio is the best model between all of the decision tree modellings.

Overall, the CRISP-DM help Yamani Tour & Travels agency effectively and efficiently use data mining to make data-driven decisions and solve complex business problems of customer churning.