HOMEWORK 2

DENNIS MULUMBI KYALO

17th September 2021

1. INTRODUCTION

In this homework problem, we shall begin by using the regression data set to discuss the evaluation metrics for regression models and one of the techniques used in regression learning.
Secondly, we shall use the classification data set to explain the five evaluation metrics for classification models and also apply the Classification and Regression Trees (CART) model to the data set.
Thirdly, we shall use examples to discuss feature scaling and its importance and finalize by explaining the Receiver Operating Characteristic curve (ROC) and why it is essential.

2. ANALYSIS AND RESULTS

2.1. REGRESSION

Table 1: Data summary

Name	regression_data
Number of rows	50
Number of columns	5
Column type frequency:	
character	1
numeric	4
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
State	0	1	7	10	0	3	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
R.D	0	1	73721.62	45902.26	0.00	39936.37	73051.08	101602.8	165349.2
Admin	0	1	121344.64	28017.80	51283.14	103730.88	122699.80	144842.2	182645.6
Marketing	0	1	211025.10	122290.31	0.00	129300.13	212716.24	299469.1	471784.1
Profit	0	1	112012.64	40306.18	14681.40	90138.90	107978.19	139766.0	192261.8

The above tables show that the regression data set has 50 samples and five variables. The five variables include State, a factor variable with three states; New York, California, and Florida, the rest four, R.D, admin, marketing, and Profit, are numeric variables. We shall be using this data set to predict the Profit based on the other four predictors.

Table 4: Coefficients of the Best Model

	x
(Intercept)	43014.4907722
R.D	0.7109963
Admin	0.0687453
Marketing	0.0458219
StateFlorida	-1591.5633029
StateNew York	-1085.9317732

We applied a multiple linear regression model to predict Profit based on the four predictors in the dataset. We ended up having the following regression model.

Profit = 3014.4907722 + 0.7109963 R.D + 0.0687453 Admin + 0.0458219 Marketing - 1591.5633029 StateFlorida - 1085.9317732 StateNew York

Table 5: Evaluation Metrics

Method	Value
Correlation	0.9631315
R-Squared	0.8971830
Mean Absolute Error	9443.9527089
Root Mean Squared Error	12937.7497595
Mean Squared Error	167385368.8398886

Various evaluation metrics for regression models help us choose the model that best fits the dataset. In this analysis, We begin by splitting the dataset into a train and test set, in which we then use the training model to predict the profit in the test set. We then compare the actual profit values in the test set with the predicted profit values.
Here, we used the following evaluation metrics as evidenced from the table above:

1. Correlation. There is a high correlation of 0.9631315, which explains a very strong positive relationship between the predicted and actual profit values.

2. R-squared. This metric helps us measure the proportion of variability in the dependent variable Profit, which is explained using the four predictors. It helps us explain the performance of the model. In this analysis, we see that 89.7183% of the Profit is explained using the four predictors. The nearer the value is to 1, the better the model, while the further the value is from 1, the poorer the model.

3. Mean Absolute Error (MAE) This valuation metric calculates the absolute difference between the actual Profit values and the predicted profit values. In our analysis, we have an MAE of 9443.9527089. The lower the MAE value, the better the model.

4. Root Mean Squared Error (RMSE) The RMSE value is calculated by taking the square root of the average of the absolute squared difference between the profit's actual and predicted value. It's simply the square root of the Mean Squared Error. From our analysis, we have an RMSE of 12937.7497595. The lower the RMSE value, the better the model.

5. Mean Squared Error (MSE) We got the MSE by calculating the average of the absolute difference between the predicted values of the profit and its actual values. From the analysis, we have an MSE of 167385368.8398. In this metric, the lower the MSE value, the better the model.

There are several techniques in regression learning; in our analysis, we shall be explaining the backward selection method, which is categorized under the stepwise regression technique.
In this method, we first start by using all the predictors to predict the dependent variable. Then we start eliminating one variable at a time. The reduction happens when the particular variable is not statistically significant to the model compared to the other variables. This process then continues until we have a stable equation that is statistically significant.

Table 6: Data summary

Name	titanic
Number of rows	891
Number of columns	6
Column type frequency:	
character	1
numeric	5
Group variables	None

Variable type: character

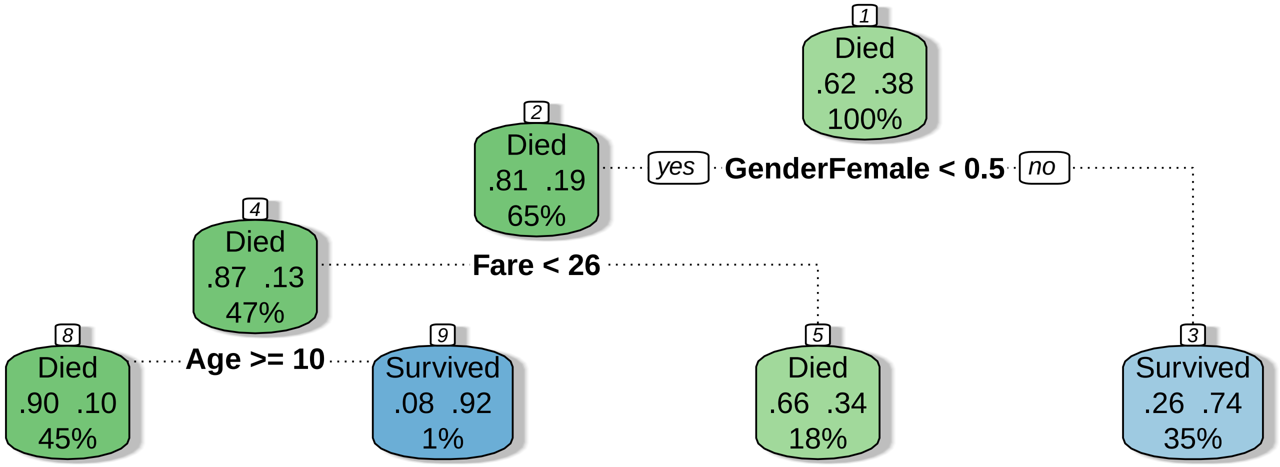
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Gender	0	1	4	6	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Survived	0	1.0	0.38	0.49	0.00	0.00	0.00	1	1.00
Age	177	0.8	29.70	14.53	0.42	20.12	28.00	38	80.00
SibSp	0	1.0	0.52	1.10	0.00	0.00	0.00	1	8.00
Parch	0	1.0	0.38	0.81	0.00	0.00	0.00	0	6.00
Fare	0	1.0	32.20	49.69	0.00	7.91	14.45	31	512.33

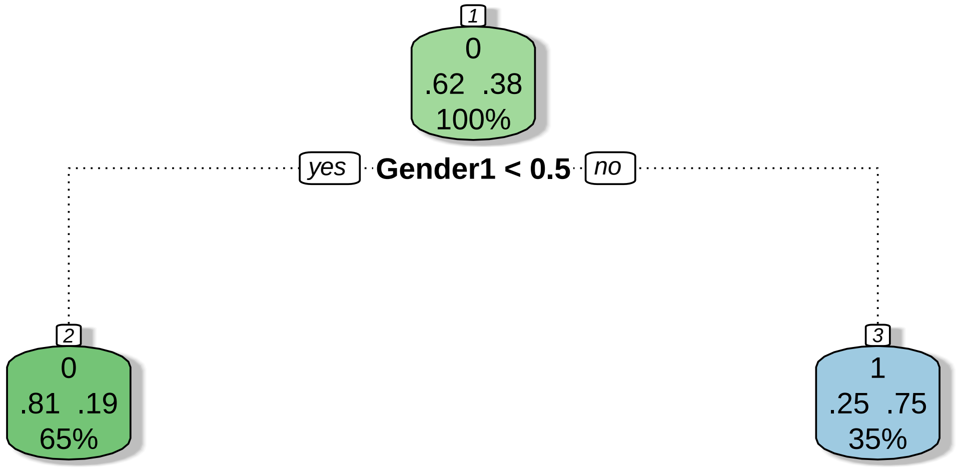
In the classification problem, we shall be using the Titanic's data set to predict the survivors of the catastrophe. From the table above, we have 891 samples and six variables, which include, Gender, Survived, Age, Sibling and Spouse (SibSp), Parents and children (Parch), and Fare. We can also see that we have 177 missing values in the Age variable.

We now proceed to predict the Survivors with respect to the five variables.



Predicting Survivors using CART model for the entire data set

From the chart above, we used the entire data set to predict the survivors so as to have clear understanding of the survivors based on the four predictors.



Predicting Survivors using CART model from the train and test set

We then proceeded to split the data set into a training and test set. Fitted the model using the training set then used it to predict the test set. We, therefore, have a clear vision of the final model, as evidenced in the above chart.

	x
Sensitivity	0.8394161
Specificity	0.6705882
Pos Pred Value	0.8041958
Neg Pred Value	0.7215190
Precision	0.8041958
Recall	0.8394161
F1	0.8214286
Prevalence	0.6171171
Detection Rate	0.5180180
Detection Prevalence	0.6441441
Balanced Accuracy	0.7550021

	x
Accuracy	0.7747748
Kappa	0.5169292
AccuracyLower	0.7140840
AccuracyUpper	0.8279705
AccuracyNull	0.6171171
AccuracyPValue	0.0000004
McnemarPValue	0.4795001

Table 9: Evaluation Metrics Tables

There are several evaluation metrics for classification models. We shall be using the table above generated from the classification model to discuss the five main evaluation metrics, which are as follows:

1. Accuracy. We have an accuracy of 77.47748%, which explains the number of correct predictions made in the model based on the total number of predictions made.

2. Precision. The model has a precision of 80.41958%, which shows the proportion of survivors who survived the catastrophe, actually survived. In simple terms, this is the proportion of predicted positives that are genuinely positive.

3. Recall or sensitivity. The model's sensitivity, 83.94161%, explains that the algorithm found the proportion of survivors that actually survived as having survived. This metric shows the actual positives that are correctly classified.

4. Specificity. The specificity proportion of 67.05882% explains the survivors who died, which the model predicted as genuinely dead. This metric describes the negative tests on the model that are genuinely negative.

5. F1 Score. We have an F1-score of 82.14286%, which shows the harmonic mean of the precision and the sensitivity.

2.3. FEATURE SCALING

Feature scaling is a technique used to normalize the range of predictors in a data set to have a standard scale within the set variables. This technique is crucial, especially before running a machine learning algorithm on a given dataset. Reason being that certain variables dominate over others due to the variation in magnitudes. And to suppress this effect, we perform feature scaling to ensure that all the variables have the same magnitude level.

To better understand the importance of feature scaling, we predict the survivors from the titanic data set before scaling and after scaling the data set, then use the accuracy as the evaluation metric to compare the difference in the two criteria used.

The table below shows the evaluation metrics before feature scaling.

	x
Sensitivity	0.8978102
Specificity	0.6235294
Pos Pred Value	0.7935484
Neg Pred Value	0.7910448
Precision	0.7935484
Recall	0.8978102
F1	0.8424658
Prevalence	0.6171171
Detection Rate	0.5540541
Detection Prevalence	0.6981982
Balanced Accuracy	0.7606698

	x
Accuracy	0.7927928
Kappa	0.5431690
AccuracyLower	0.7334642
AccuracyUpper	0.8441229
AccuracyNull	0.6171171
AccuracyPValue	0.0000000
McnemarPValue	0.0121928

Table 10: Evaluation Metrics before Feature Scaling

The table below shows the evaluation metrics after applying feature scaling on the age and fare variables in the data set.

	x
Sensitivity	0.9197080
Specificity	0.7176471
Pos Pred Value	0.8400000
Neg Pred Value	0.8472222
Precision	0.8400000
Recall	0.9197080
F1	0.8780488
Prevalence	0.6171171
Detection Rate	0.5675676
Detection Prevalence	0.6756757
Balanced Accuracy	0.8186775

	x
Accuracy	0.8423423
Kappa	0.6564075
AccuracyLower	0.7876182
AccuracyUpper	0.8876623
AccuracyNull	0.6171171
AccuracyPValue	0.0000000
McnemarPValue	0.0425225

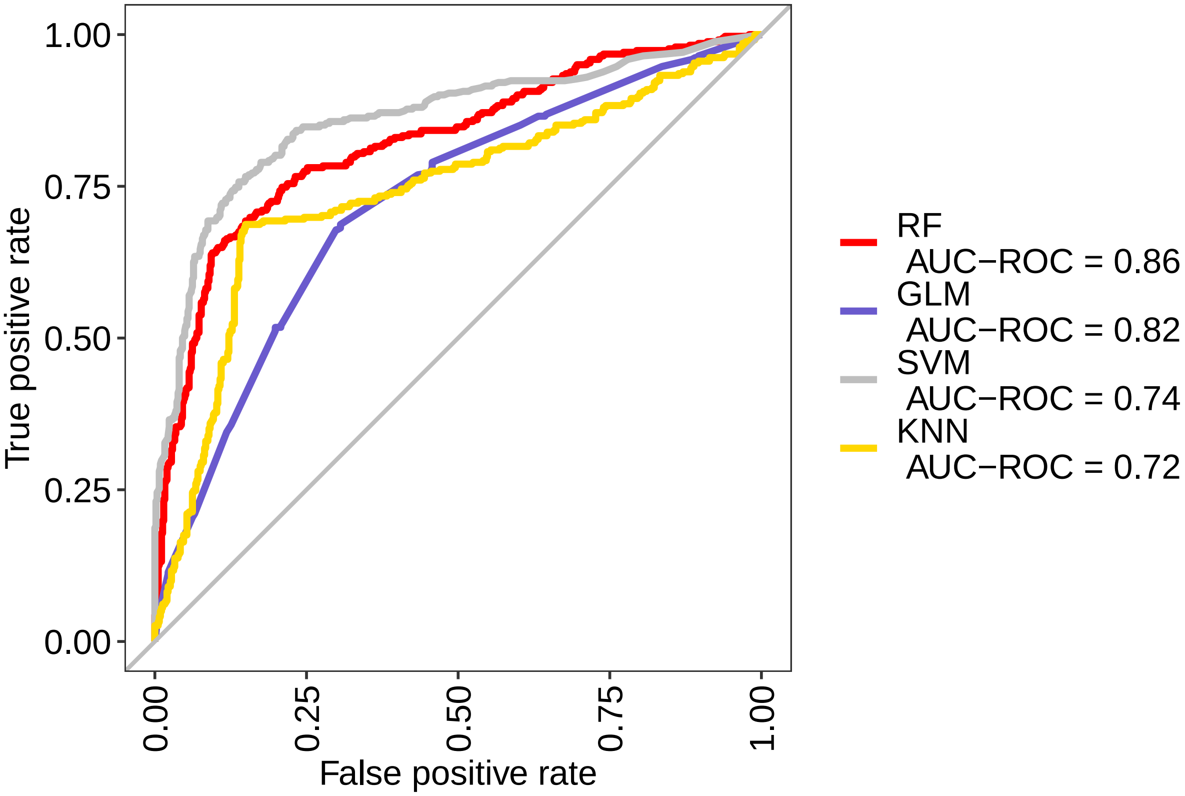
Table 11: Evaluation Metrics after Feature Scaling

Based on the two criteria, it is evident that after applying feature scaling, the accuracy increased from 79.27928% without feature scaling to 84.23423% after using feature scaling. This is a remarkable improvement in accurately predicting the survivors.

2.4. RECEIVER OPERATING CHARACTERISTIC CURVE (ROC)

The Receiver Operating Characteristic (ROC) curve is a graph of the true positive rate (TPR) against the false positive rate (FPR), which helps get a clear understanding of how well the classification models perform.
We apply four supervised classification algorithms to predict the survivors in the dataset and compare their accuracies based on the ROC curve.

The algorithms used are:
1. Random Forest (RF)
2. Generalized Linear Models (GLM)
3. Support Vector Machines (SVM)
4. K-Nearest Neighbor (KNN)



The above ROC curve shows that random forest had the highest classification accuracy of 86% compared to the rest.

3. REFERENCE

An Introduction to Statistical Learning 1st edition, by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, Springer, New York.