**Predicting Students' dropout and academic success rates**

By: Chukwudi Nkamnebe

DS 535 Capstone

Sept 2023

**BACKGROUND OF THE STUDY**

This dataset provides a detailed view of students enrolled in various undergraduate programs offered at a higher education institution.

It includes demographic data, academic performance, and social-economic factors, these are crucial information that can be used to analyze the possible predictors of student dropout and academic success.

This dataset was published on kaggle.com which has 35 variables and 4424 observations.

For this project's sake, a simple random sampling was performed to obtain a sample of 1,500 from the data.

The training data equals 900 observations, and the validation data will be 600 observations.

The X variables are:

* Application mode
* Displaced
* Debtor
* Tuition fees up to date
* Gender
* Scholarship holder
* Curricular unit's 1st sem (enrolled)
* Curricular unit's 1st sem (approved)
* Curricular unit's 1st sem (grade)
* Curricular unit's 2nd sem (enrolled)
* Curricular unit's 2nd sem (approved)
* Curricular unit's 2nd sem (grade)

The Y variable is the Target variable, meaning a student is a dropout, currently enrolled or graduated. The target variable was converted to a numerical variable so that linear regression can be performed on the data.

**AIM OF THE PROJECT**

The project aims at exploring the predictors of academic performance and determining its relationship to the academic outcome (Target = Dropout, Enrolled and Graduate) using the provided data set. The project also aims to build a model of predictors to determine the target outcome of the students. The model should be able to learn from the data and predict the target outcome in any higher education institution given that all other metrics are present.

Also, in this third report, we will be using the logistic regression method to explore the selected variables and determine the most significant variables classifying the Y variable (Target). The following variables namely Debtor, Gender and Curricular unit’s 1sr sem (enrolled) in the previous data because they do not provide any substantial significance in explaining and drawing more insights from the data.

After this we have our new variables listed below:

The X variables are:

* Application mode
* Displaced
* Tuition fees up to date
* Scholarship holder
* Curricular unit's 1st sem (approved)
* Curricular unit's 1st sem (grade)
* Curricular unit's 2nd sem (enrolled)
* Curricular unit's 2nd sem (approved)
* Curricular unit's 2nd sem (grade)

The Y variable is the Target variable.

**VARIABLE CONVERSION**

The Y variable which is target was originally classified as 1 = Dropout, 2 = Enrolled and 3 = Graduate.

The median value of the Target variable was calculated with the value of 2. The reason behind this is because we will be performing logistic regression on the data and will require the Y variable to be in a binary form (0 and 1) for easy interpretation of the data.

The new values of the Target variable are 0 and 1 where 1 represents if a student will Graduate and 0 represents if a student will Dropout.

**FULL MODEL**

For this part, I performed logistic regression analysis on the whole data because we need to distinguish how each variable affects each other and over the individual and collective impact on the data. So, all 10 variables were included in the analysis and with the output shown below for more insight.

A screenshot of a computer code

Description automatically generated

We will explain and draw insights from the X variables and see if we can draw collective insights from the data.

1. Intercept: The intercept is -2.88159. In logistic regression, the intercept represents the log-odds of the event when all predictors are zero and, in this case, it's the log-odds of the event when all categorical variables are at their reference levels and continuous variables are at zero.

2. Application mode: For a one-unit increase in the Application mode, the log-odds of the response variable decrease by -0.03647. Since the p-value is less than 0.05, we consider this variable statistically significant. It shows that as the Application mode increases, the odds of the event happening decrease.

3. Displaced: The coefficient for displaced is 0.10365, but the p-value is greater than 0.05. Therefore, we cannot confidently say that being Displaced has a significant effect on the log-odds of the Target variable stating if a student will graduate or not.

4. Tuition.fees.up.to.date: A one-unit increase in Tuition.fees.up.to.date will cause an increase in the log-odds of the Target variable by 2.32913. This variable is highly significant (p-value < 0.001), showing that higher tuition fees are strongly associated with increased odds of a student graduating. This also makes sense because as students pay their tuition on time, they will have little or no reason to miss classes and will be more serious with their studies so as to get back their value for money.

5. Scholarship.holder: As a student, being a Scholarship holder is naturally linked with serious minded students and from the analysis we can see it is associated with an increase in the log-odds by 0.99596. This variable is highly significant (p-value < 0.001), meaning that Scholarship holders have significantly higher odds at 99.5% of graduating from college compared to non-scholarship holders.

6. Curricular.units.1st.sem.enrolled.: A one-unit increase in Curricular.units.1st.sem.enrolled will result in a decrease in the log-odds by 0.42141. This variable is significant (p-value < 0.001), indicating that higher enrollment in the first semester is associated with decreased odds of a student graduating. This also makes sense because the more courses a student register, they might be overwhelmed and slack off in their academic performance

7. Curricular.units.1st.sem.approved.: A one-unit increase in Curricular.units.1st.sem.approved. will result in an increase in the odds of graduation by 0.65547. This variable is highly significant (p-value < 0.001), suggesting that higher approval of curricular units in the first semester is strongly associated with increased odds of a student graduating college. Also, this can be seen as a positive trend because if you’re approved of the first semester then there’s a higher chance that a student will graduate.

8. Curricular.units.1st.sem.grade.: A one-unit increase in Curricular.units.1st.sem..grade. is associated with a slight decrease in the log odds by 0.13093. This variable is significant at 0.05 (p-value = 0.008699), indicating that higher grades in the first semester are slightly linked with decreased

odds of the student graduating. This can be true if interpreted as such, a student feeling more relaxed about their first semester performance can cause reduced efforts in the second semester causing a slight possibility of poor performance by the student.

9. Curricular.units.2nd.sem.enrolled.: A one-unit increase in Curricular.units.2nd.sem.enrolled. is associated with a decrease in the chances of a student graduating by 0.93907. This variable is highly significant (p-value < 0.001), showing that higher enrollment in the second semester is strongly linked to if a student will graduate or not.

10. Curricular.units.2nd.sem.approved.: A one-unit increase in Curricular.units.2nd.sem.approved. is associated with an increase in the chances of a student graduating by 0.97182. This variable is highly significant (p-value < 0.001), suggesting that higher approval of curricular units in the second semester is strongly associated with increased odds of the event.

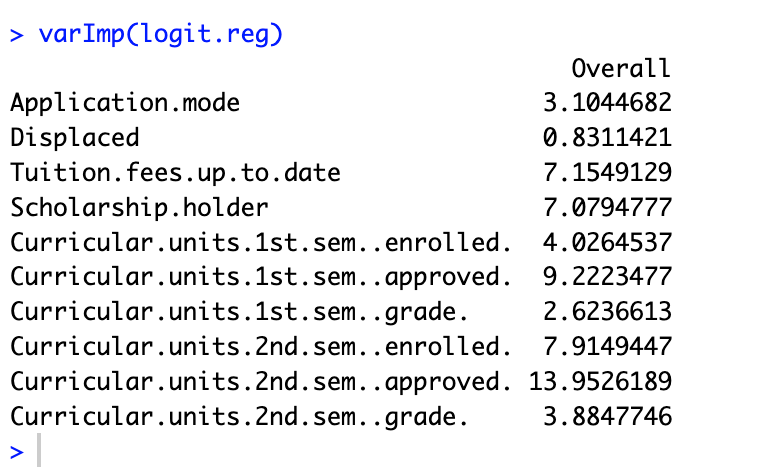
11. Curricular.units.2nd.sem.grade.: A one-unit increase in Curricular.units.2nd.sem.grade. is associated with an increase in the log-odds by 0.19877. This variable is significant (p-value = 0.000102), indicating that higher grades in the second semester are associated with increased chances of a student graduating from college.

**COLLECTIVE EFFECTS**

- The Tuition.fees.up.to.date, Scholarship.holder, and both Curricular.units.1st.sem.approved. and Curricular.units.2nd.sem..approved. variables have a notable positive impact on the log-odds, indicating a higher probability of the event.

- Conversely, Application.mode, Curricular.units.1st.sem..enrolled., Curricular.units.2nd.sem..enrolled., Curricular.units.1st.sem..grade., and Curricular.units.2nd.sem..grade. have negative impacts on the log-odds, showing a lower probability of the event.

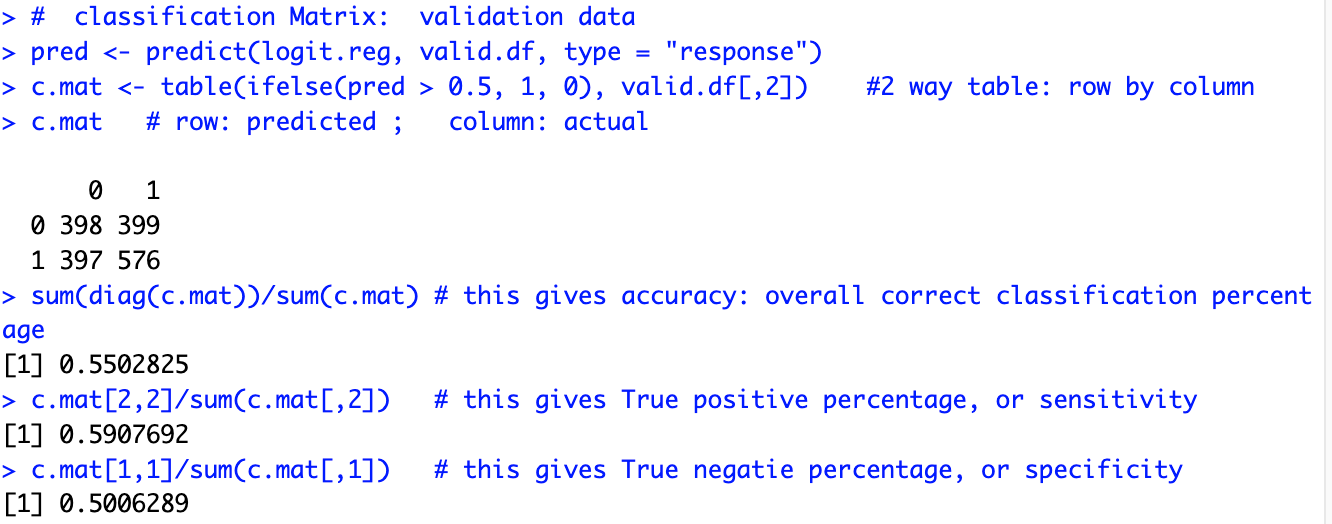
- Displaced does not seem to have a statistically significant impact on the outcome.



We can deduce that, Application mode, Debtor, Curricular 1st semester enrolled, Curricular.units.2nd.sem..grade, Curricular.units.1st.sem..approved, Scholarship.holder, Tuition.fees.up.to.date are moderately significant in predicting if a student graduates or drops out.

Curricular.units.2nd.sem..approved is the variable with the highest importance among all predictors.

**CONFUSION MATRIX - VALIDATION**



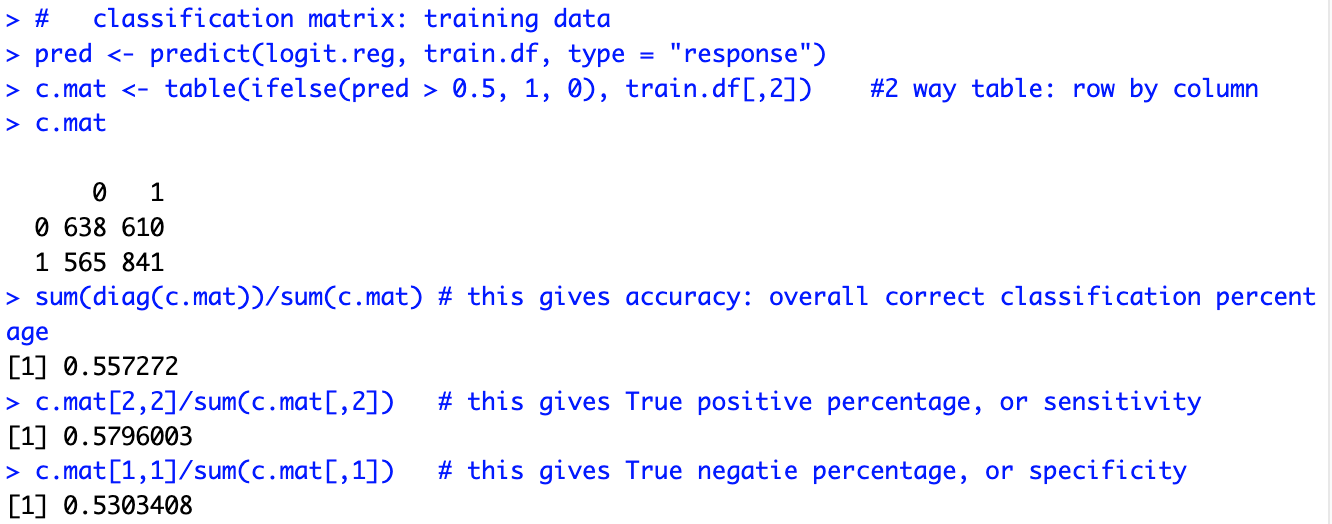
From the validation data, we can see that the confusion matrix shows the counts of true positives (398), false positives (399), true negatives (576), and false negatives (397).

The overall correct classification percentage for the validation data is approximately 55%.

Sensitivity: The model correctly identifies about 59% of the positive cases in the validation data.

Specificity: The model correctly identifies about 50% of the negative cases in the validation data.

**CONFUSION MATRIX - TRAINING**



Similar to the validation data, this matrix shows the counts of true positives, false positives, true negatives, and false negatives for the training data.

The overall correct classification percentage for the training data is approximately 55.72%.

Sensitivity: The model correctly identifies about 57.96% of the positive cases in the validation data.

Specificity: The model correctly identifies about 53% of the negative cases in the validation data.

**OVERALL ACCURACY**

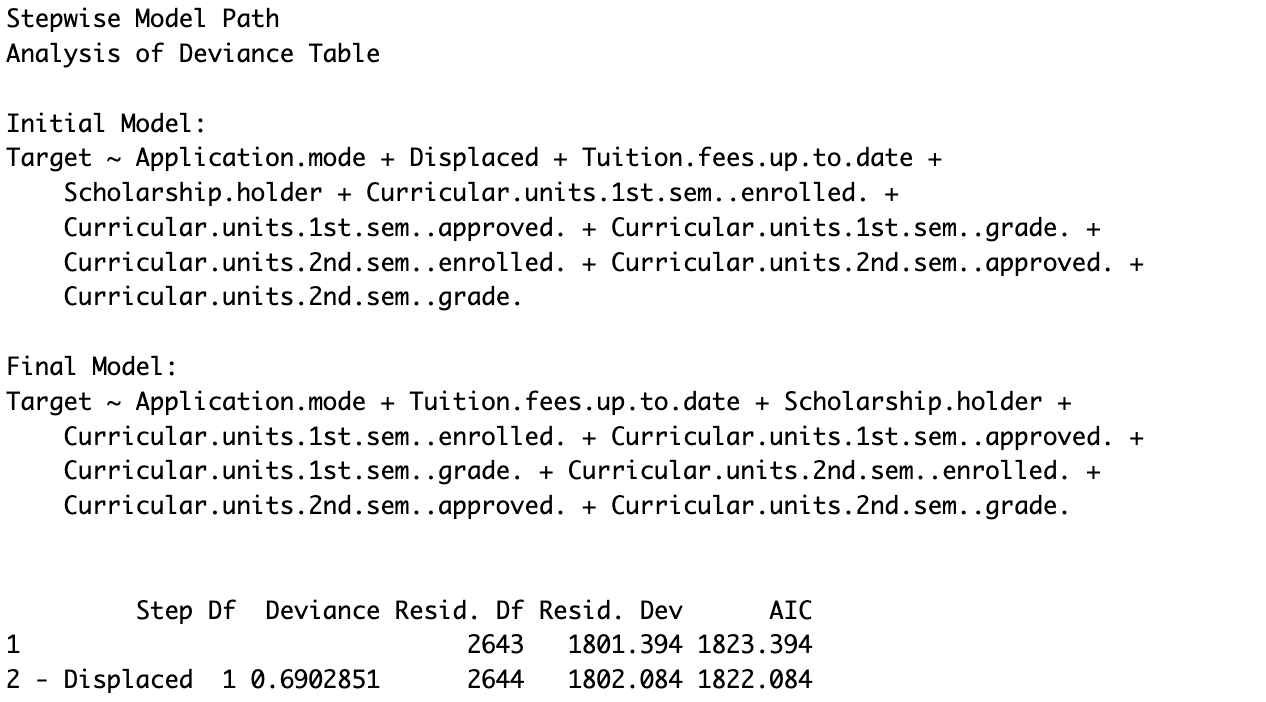
The overall accuracy for both the validation and training datasets is around 54-56% approximately. This suggests that the model's performance is not very high, and there may be room for improvement.

Sensitivity and Specificity:

Sensitivity (True Positive Rate) is relatively mid-range, indicating that the model slightly struggles to correctly identify positive cases. This is evident in both the validation and training datasets.

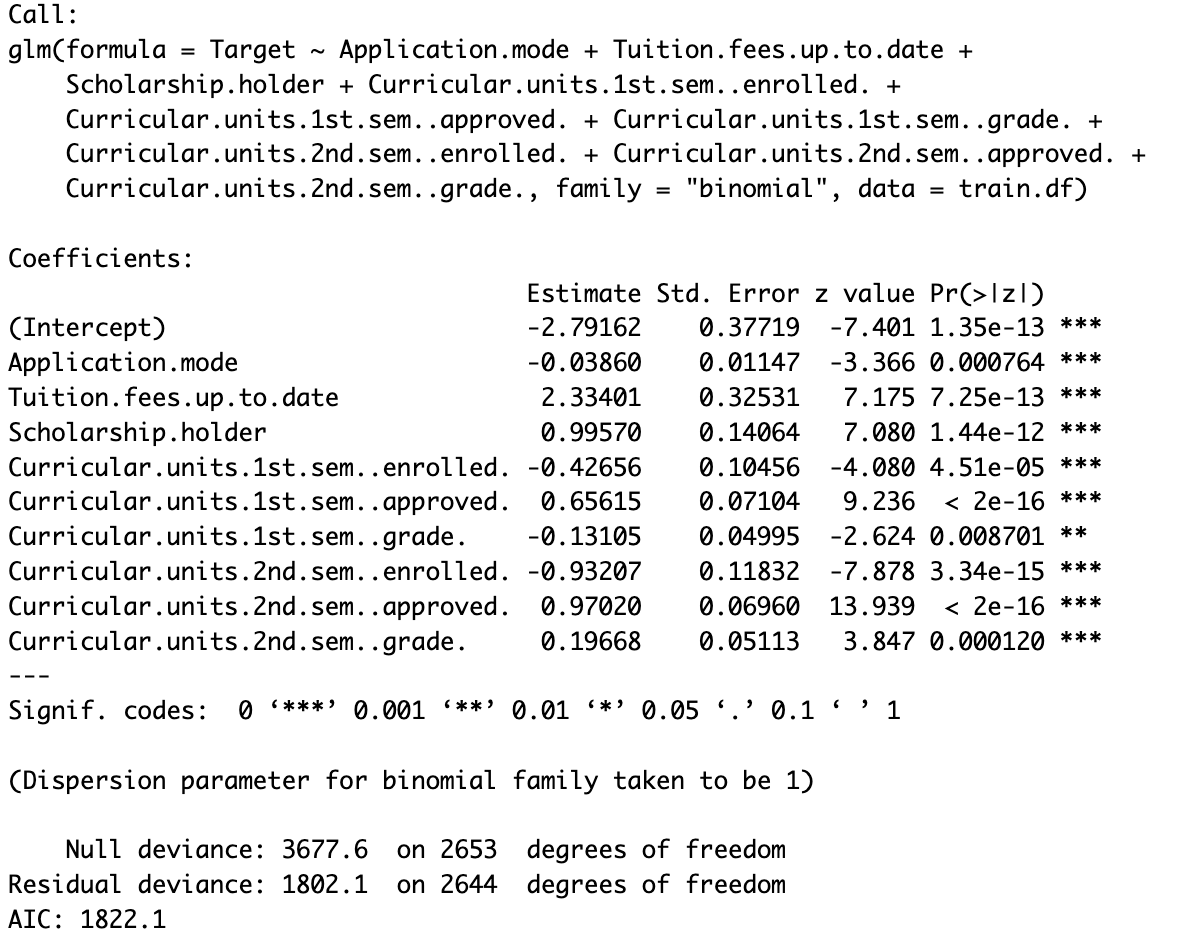
Specificity (True Negative Rate) is somewhat higher, but still not very high. The model has better performance in correctly identifying negative cases.

**USING STEP AIC TO PERFORM VARIABLE SELECTION**

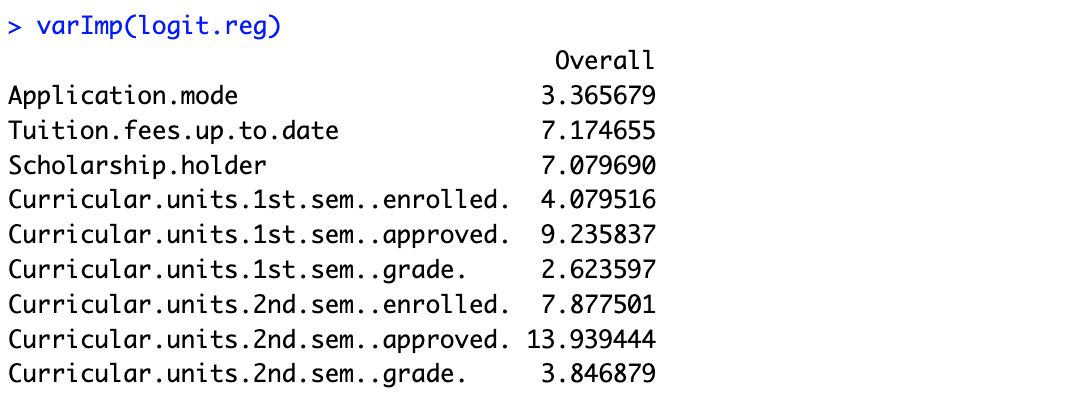


After performing the STEP AIC, we can deduce that the Displaced variable was removed from the original model which confirms the analysis we performed while analyzing the logistic regression model output from the initial analysis.

Furthermore, a new best model will be developed using the results of the STEP AIC and the output will be displayed below.



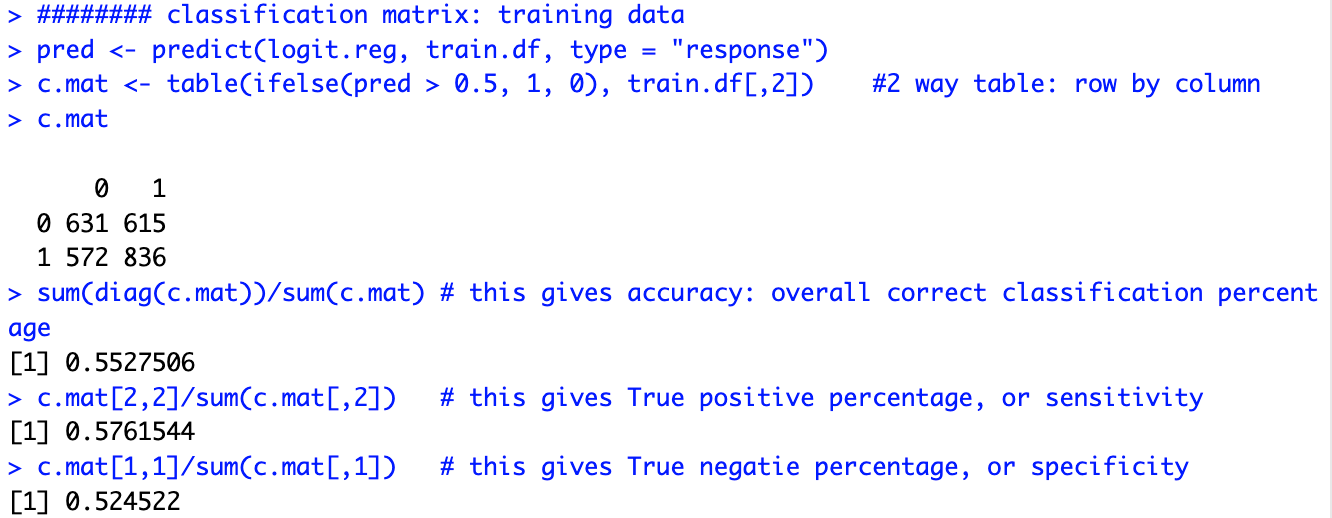
From the final model, which is the best model, we can notice that all the selected variables are statistically significant regardless of their positive and negative impacts on the model which can be seen from the coefficients in the model.



We can also deduce by judging from the variables with the highest T statistics score which includes Tuition up to date, Scholarship holder, Curricular unit's 1st semester approved, Curricular unit’s 2nd semester enrolled, and Curricular unit’s 2nd semester approved with scores 7.174, 7.079, 9.235, 7.877 and 13.939 respectively.

**CLASSIFICATION MATRIX: BEST MODEL TRAINING AND VALIDATION**

**TRAINING DATA;**



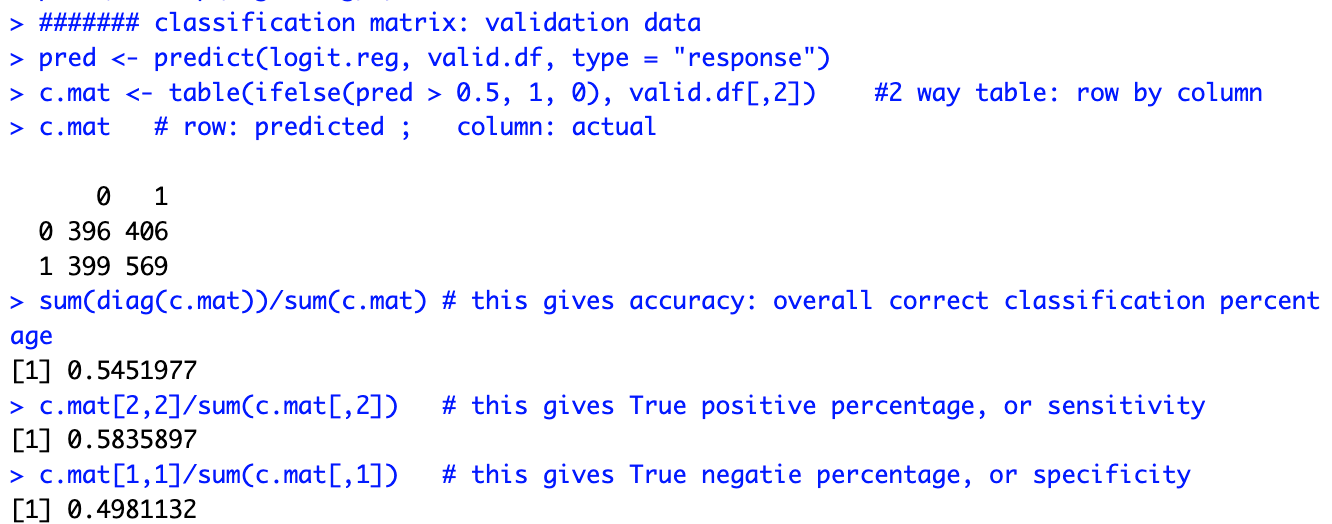
Similar to the validation data, this matrix shows the counts of true positives, false positives, true negatives, and false negatives for the training data.

The overall correct classification percentage for the training data is approximately 55.27%.

Sensitivity: The model correctly identifies about 57.62% of the positive cases in the validation data.

Specificity: The model correctly identifies about 52.45 % of the negative cases in the validation data.

**VALIDATION DATA;**



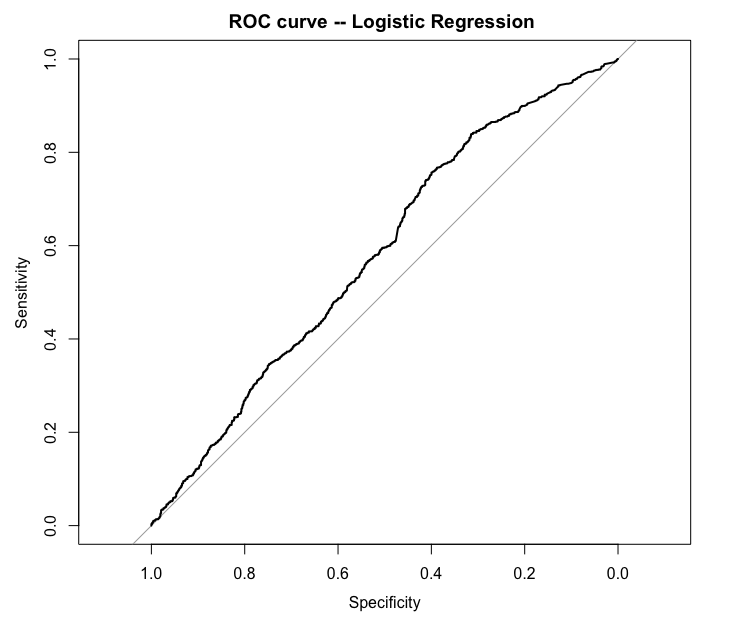
From the validation data of the best model, we can see that the confusion matrix shows the counts of true positives (396), false positives (406), true negatives (569), and false negatives (399).

The overall correct classification percentage for the validation data is approximately 54.51%.

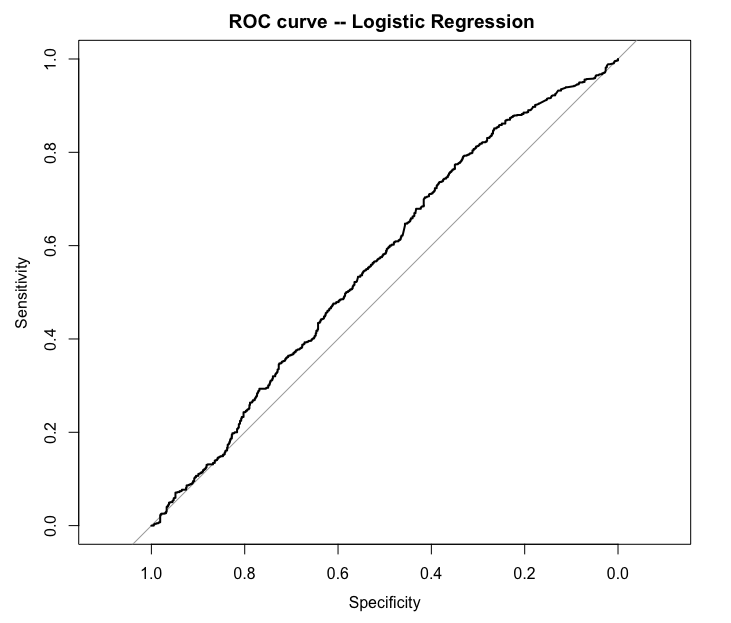
Sensitivity: The model correctly identifies about 58.35% of the positive cases in the validation data.

Specificity: The model correctly identifies about 50% of the negative cases in the validation data.

**ROC CURVES: TRAINING AND VALIDATION DATA**



**This is the ROC curve for the training data with a AUC score of 0.5804 (58.04%).**



**This is the ROC curve of the validation data with an AUC score of 0.5623.**

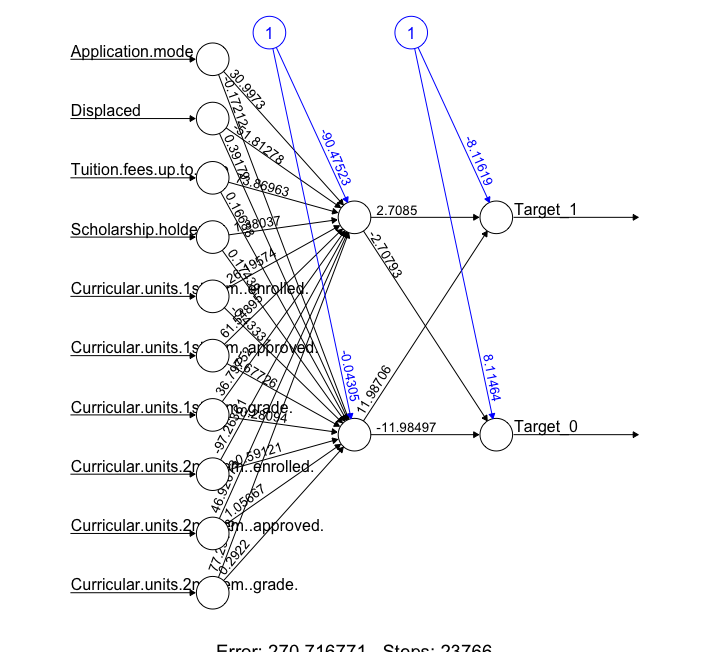
**NEURAL NETWORK MODEL ANALYSIS**

A neural network is an artificial intelligence method that gives computers instructions to evaluate data in a way that closely resembles the structure of the human brain. Deep learning, a type of artificial intelligence technology, mimics the organization of the human brain by using interconnected neurons or nodes arranged in layers. I created a neural network model using the dataset in order to examine the data in greater detail. The following sections provide an explanation of the findings and observations. Similar to how the dataset was divided in the previous analysis, 60% of the data were allocated to the training dataset and 40% to the validation dataset using random sampling with a set seed value of 2 so that the data can be reproduced.

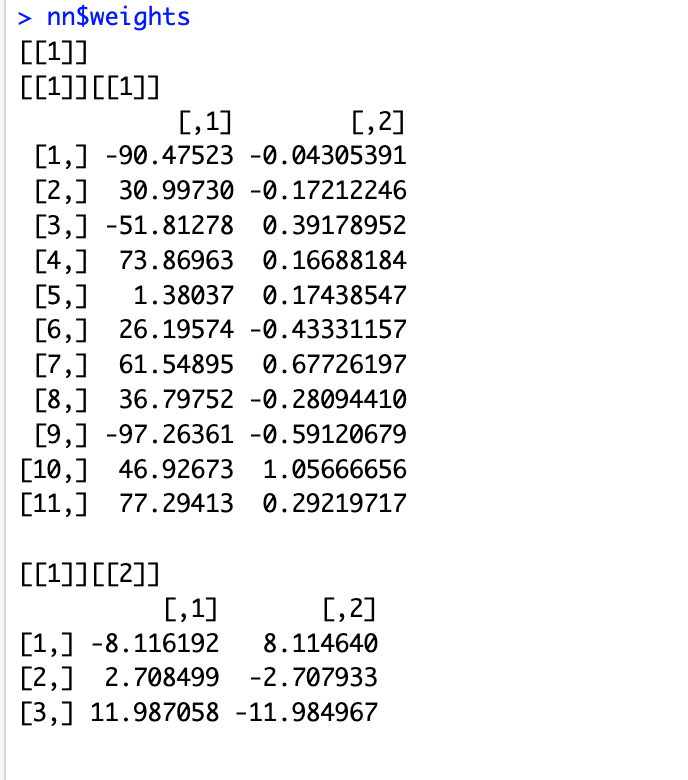
In the previous report we focused on performing logistic regression on the dataset to derive more insights from the data and now we will be performing neural networks model estimation to discover better insights in the data. First, we prepare the data for the analysis which includes normalizing the predictor variables in the data. The reason behind the normalization is simply to make the model estimation run faster.

Also, two logical variables Target\_1 and Target\_0 was created using the information on the Y variable (Target) to run NN model, they are added to the end of the data frame with logical variables that have values True or False.

The neural network estimation will be performed on all the variables, and we can see the output of the full model with one layer and two neurons below:



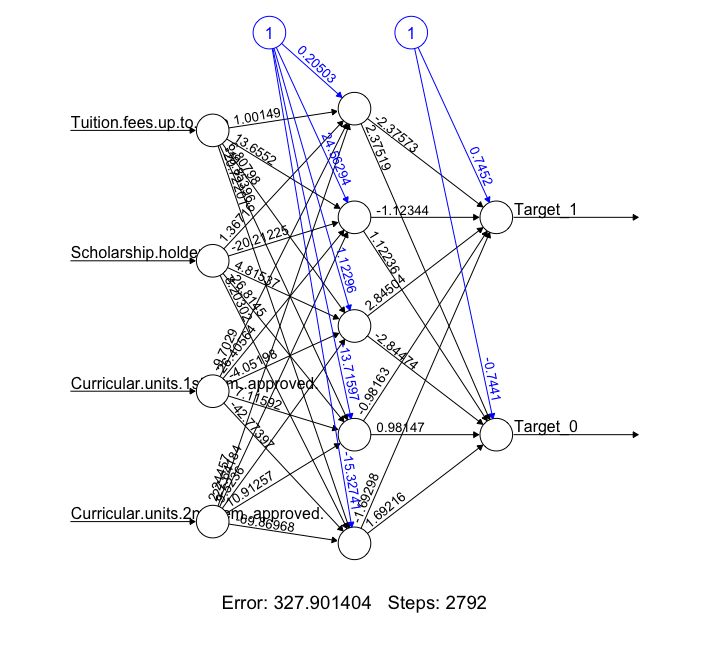
**FULL MODEL REGRESSOR COEFFICIENTS**



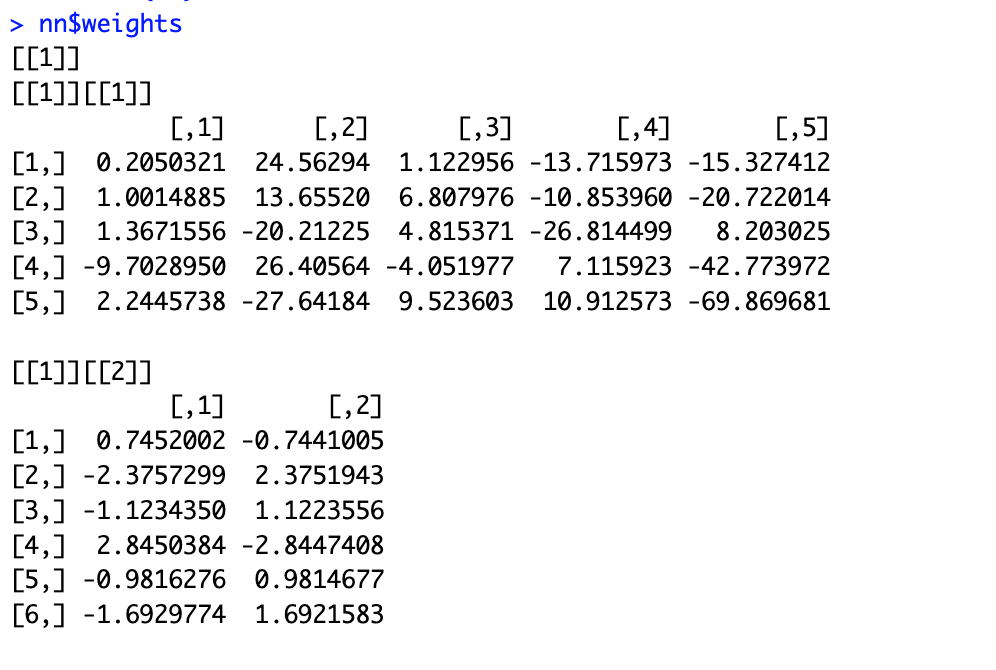
Also, we have an output of the individual weights of the entire neural network shown below:

**MODEL 2 NEURAL NETWORK AND REGRESSOR WEIGHTS**

In model 2, the following variables were selected for the analysis: Tuition fees up to date, Scholarship holder, Curriculaer units 1st semester approved and curricular units second semester approved. There are the models in the best model derived from the logistic regression performed in the previous part. This model consists of one layer and 5 neurons.

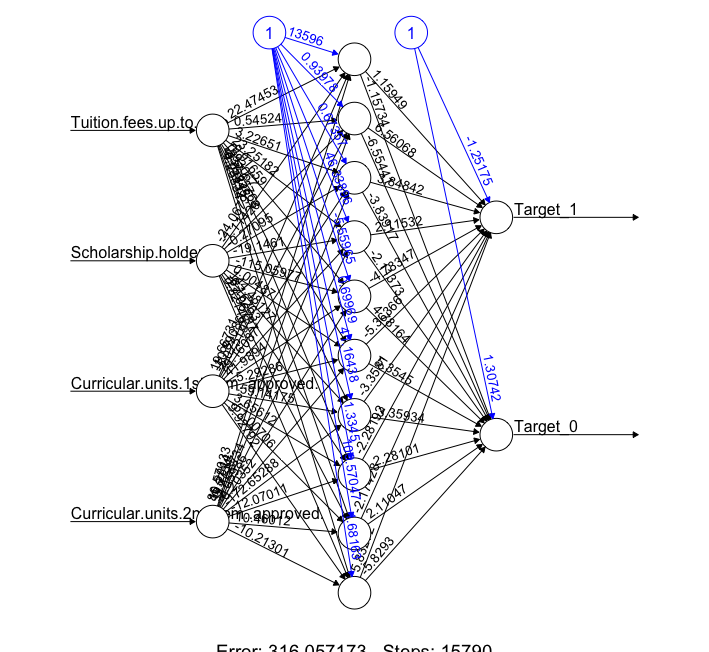


**MODEL 2 REGRESSOR COEFFICIENTS**

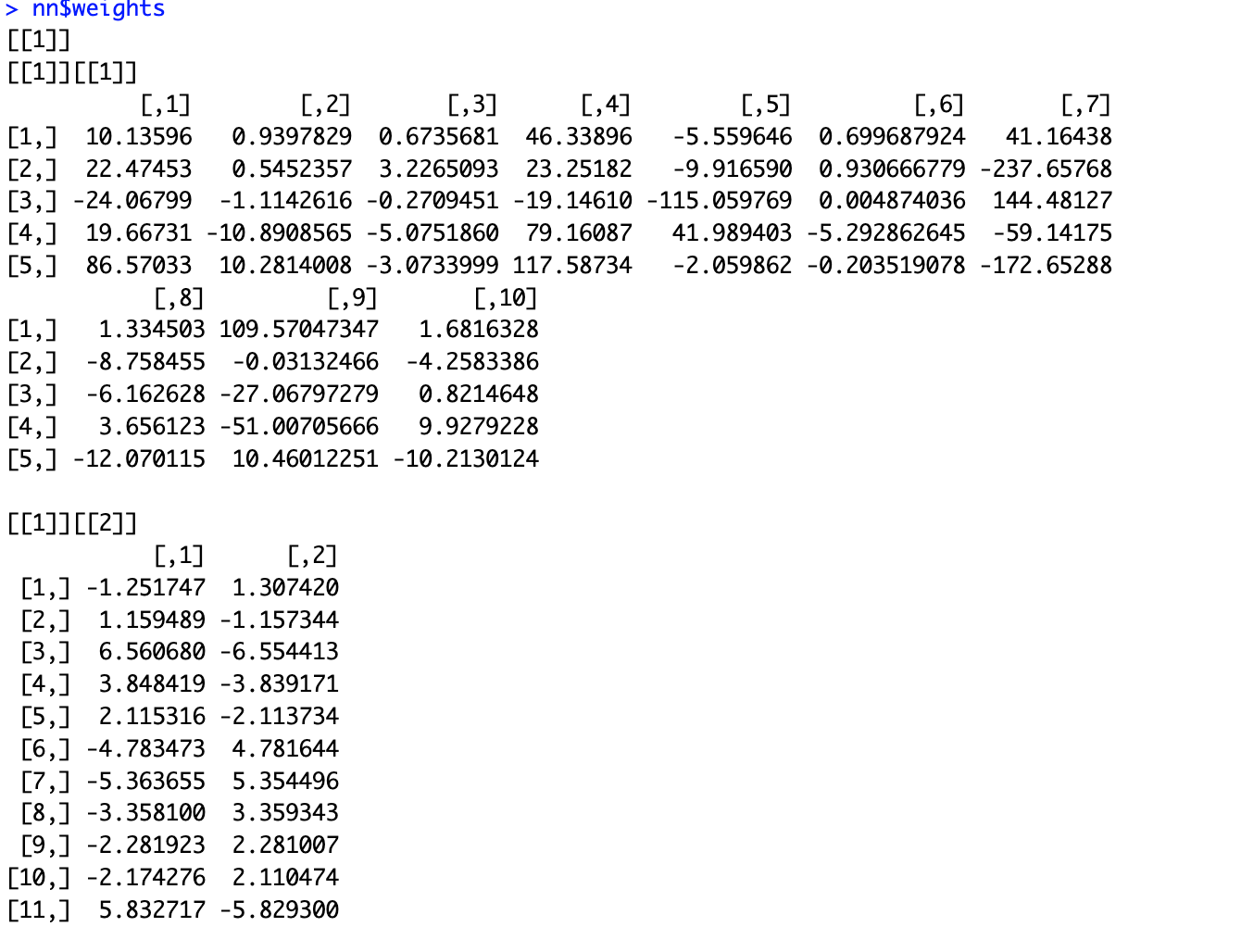


**MODEL 3 NEURAL NETWORK AND REGRESSOR WEIGHTS**

In model 3, the following variables were selected for the analysis: Tuition fees up to date, Scholarship holder, Curriculaer units 1st semester approved, and curricular units second semester approved. There are the variables in the best model derived from the logistic regression performed in the previous part. This model consists of one layer and 10 neurons.

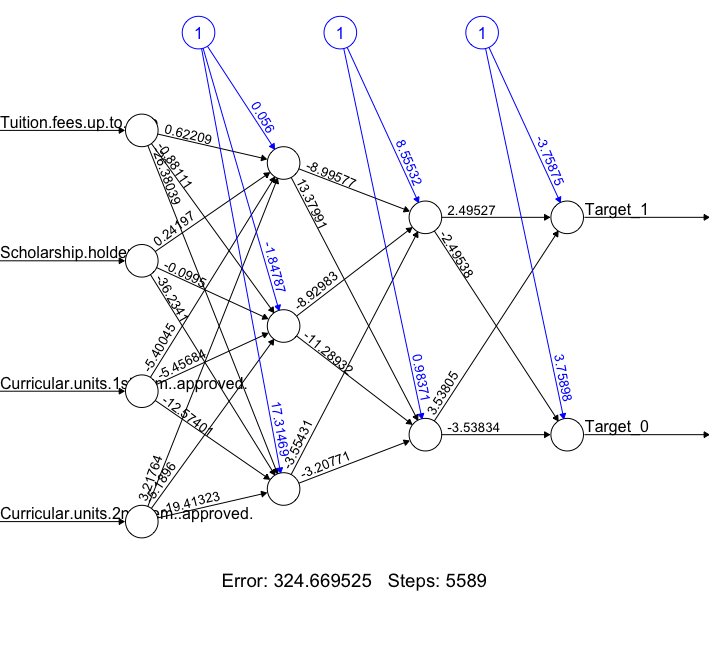


**MODEL 3 REGRESSOR COEFFICIENTS**

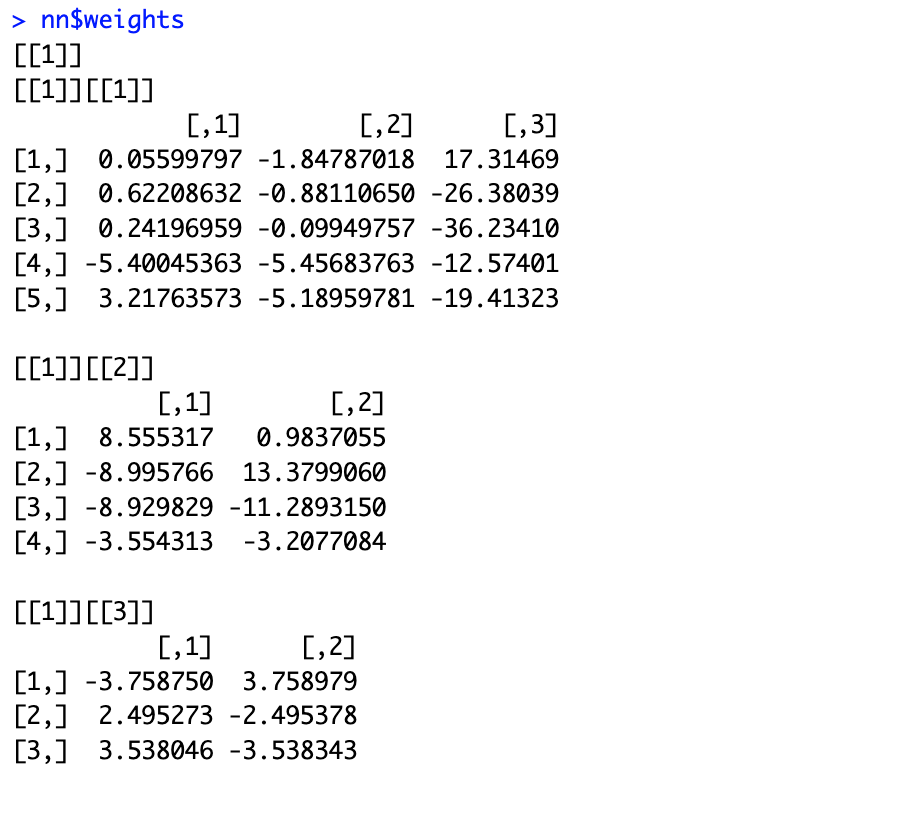


**MODEL 4 NEURAL NETWORK AND REGRESSOR WEIGHTS**

In model 4, the following variables were selected for the analysis: Tuition fees up to date, Scholarship holder, Curricular unit's 1st semester approved, and curricular units second semester approved. There are the variables in the best model derived from the logistic regression performed in the previous part. This model consists of three layers and 2 neurons.



MODEL 4 REGRESSOR COEFFICIENTS



**CLASSIFICATION TABLE OF ALL MODELS**

1. Full Model TD/VD

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 0 | | 1 | |
|  | TD | VD | TD | VD |
| 0 | 1115 | 688 | 141 | 121 |
| 1 | 245 | 167 | 1153 | 794 |

TD: Accuracy: 85.45% Specificity: 89.10% Sensitivity: 81.98%

VD: Accuracy: 83.73% Specificity: 86.78% Sensitivity: 80.47%

1. Model 2 TD/VD

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 0 | | 1 | |
|  | TD | VD | TD | VD |
| 0 | 1068 | 667 | 161 | 110 |
| 1 | 292 | 188 | 1133 | 805 |

TD: Accuracy: 82.93% Specificity: 87.56% Sensitivity: 78.53%

VD: Accuracy: 83.16% Specificity: 87.98% Sensitivity: 78%

1. Model 3 TD/VD

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 0 | | 1 | |
|  | TD | VD | TD | VD |
| 0 | 1072 | 646 | 158 | 113 |
| 1 | 288 | 209 | 1136 | 802 |

TD: Accuracy: 83.19% Specificity: 87.78% Sensitivity: 78.82%

VD: Accuracy: 81.81% Specificity: 87.65% Sensitivity: 75.5%

1. Model 4 TD

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 0 | | 1 | |
|  | TD | VD | TD | VD |
| 0 | 1088 | 659 | 176 | 124 |
| 1 | 272 | 196 | 1118 | 791 |

TD: Accuracy: 83.12% Specificity: 86.40% Sensitivity: 80%

VD: Accuracy: 81.92% Specificity: 86.45% Sensitivity: 77%

**Classification Matrix report for the neural network validation models**

* Full Model VD

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 688 | 121 |
| 1 | 167 | 794 |

Accuracy: 83.73% Specificity: 86.78% Sensitivity: 80.47%

* Model 2 VD

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 667 | 110 |
| 1 | 188 | 805 |

Accuracy: 83.16% Specificity: 87.98% Sensitivity: 78%

* Model 3 VD

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 646 | 113 |
| 1 | 209 | 802 |

Accuracy: 81.81% Specificity: 87.65% Sensitivity: 75.5%

* Model 4 VD

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 659 | 124 |
| 1 | 196 | 791 |

Accuracy: 81.92% Specificity: 86.45% Sensitivity: 77%

Comparing the following models, we can see that Model 3 and Model 4 have accuracy values of 81.81% and 81.92 respectively while the full Model and Model 2 has the highest accuracy meaning that the model correctly classifies about 83.73% and 83.16% of the instances in the model. For the case of this study, I chose Model 2 to be the best model considering that its accuracy rate is marginally close to that of the full model. The Model 2 has more neurons present in its network meaning more complex processing compared to the full model. Also, we can see that the accuracy of models 3 and 4 are close to the best model showing that the models show better performance at correctly classifying instances of the model.

Furthermore, we can observe that across all 4 models, they correctly identify a high percentage of the true negatives and positives instances in the model. Also, we can see that Model 2 has a specificity of 87.98% which is the highest amongst the 3 models, it simply shows that 87.98% of the actual negatives were correctly identified by the model. Finally, model 2 has a sensitivity of 78% meaning that 78% of the true positives were correctly identified by the model. Higher values of sensitivity and specificity of a model shows a better performance at correctly classifying the true positives and negatives in the model therefore model 2 will be the best model.

**SUMMARY AND CONCLUSION**

A neural network is an application of artificial intelligence that instructs computers on how to interpret data in a way that resembles the way the human brain does.

Overall, the neural network's output surpassed the logistic regression in most cases. The best model demonstrates its usefulness in accurately classifying true positives, true negatives, and producing more accurate forecasts, respectively, with its excellent performance in terms of specificity, sensitivity, and accuracy.

We may conclude that, in comparison to the logistic regression, the neural network performs better with our data.