Potential Student Dropout

Early Detection

Group 2

Triet Tran, Alex Ordonez, Subaranjana Giridharan, Chris Nicols

[**Introduction 3**](#_heading=h.u0fay1q5sgv8)

[**Project High Level 3**](#_heading=h.i4pj39k9hi74)

[**Dataset 4**](#_heading=h.w516twwoadeg)

[Data Collection 4](#_heading=h.tngdavr1syht)

[Summary 4](#_heading=h.7bx6br9rp5t6)

[**Data Exploration Analysis (DEA) 5**](#_heading=h.evevr8fsi1nf)

[Using Domain Knowledge 5](#_heading=h.pasle0wj1lrg)

[Statistical Methods 6](#_heading=h.9g5e1j7376tp)

[Removing Variables 7](#_heading=h.ng9x91evwc0a)

[Summary of Data Reduction 9](#_heading=h.d2210mg0x3dl)

[**Applying Models 9**](#_heading=h.tx5eb9wa5mn3)

[**Logistic Regression 10**](#_heading=h.pow69dxws7xq)

[Summary Result 10](#_heading=h.y9ryz84nowv6)

[SAS Figures 10](#_heading=h.7uznvu3hcfiy)

[**CART 11**](#_heading=h.nvq26hiahbnr)

[Summary Result 11](#_heading=h.c5gs9huw35eb)

[SAS Figures 12](#_heading=h.ng0yz5nji9i6)

[**Neural Network 13**](#_heading=h.we18kqproutv)

[Summary Result 13](#_heading=h.ovoine4r0a8r)

[SAS Figures 14](#_heading=h.58c5785wd9xp)

[**Comparison Different Model Types: Logistic Regression - CART - Neural Network 15**](#_heading=h.v3whejnwk2p1)

[**Conclusion 15**](#_heading=h.n48gztu4r4gf)

[**Appendix - Sample SAS Code 15**](#_heading=h.h8yyqwrr65di)

[Statistical Methods for EDA. 15](#_heading=h.le58r8lkq3bg)

[Logistic regression 16](#_heading=h.fk1fn6elir9m)

[Backward 16](#_heading=h.99hk0vg6gx7y)

[Forward 17](#_heading=h.ezhiheu5sgkh)

[Stepwise 18](#_heading=h.trolzyuoolc5)

[CART 19](#_heading=h.p9ajp7un6vyb)

[CART using Entropy with VarsEDA 19](#_heading=h.ac8eprv45csv)

[CART using Entropy with reduced list of variables 19](#_heading=h.tnd3qyzh7gq3)

[Cart using Gini with VarsEDA 19](#_heading=h.z1a7ild9qb16)

[CART using Gini with reduced list of variables 19](#_heading=h.aw8id3mzqpry)

[Neural Network 20](#_heading=h.du6kkwys95fn)

[1 Hidden Layer 32 Neurons 20](#_heading=h.cuijcfk4448k)

[2 Hidden Layers 32 & 16 Neurons 20](#_heading=h.jvafhotvxnqb)

[2 Hidden Layers 32 & 48 Neurons 21](#_heading=h.phtr1om65ecd)

[**Reference 21**](#_heading=h.jf6mojui7te8)

# Introduction

A student entering into their college years can be a difficult time as the college environment is a major adjustment from high school. Some students are able to adjust easier than others, and for the students who have a tough time adjusting we see high dropout rates. Universities never want to see students struggle in their classes or dropout entirely. While universities do not want students to drop out of the school, we are seeing an average dropout rate of 32.9% every year in the US. If we use 2022 enrollment data with a total student enrollment of 18.58 million students, that means we are observing a current dropout rate of 6.11 million students per year.

If universities can identify students at risk of dropping out early in their academic journey, they can provide targeted support to help them succeed. Therefore, our goal is to predict whether a student is likely to drop out using only the information available from enrollment through the second semester.

# Project High Level

An overview of the dataset is given, including how data is collected as well as a summary of the dataset.

Due to the sparsity of the dataset, dimension reduction was a strong focus. This is done in Data Exploratory Analysis (EDA) through three steps:

* Using domain knowledge
* Using statistical approach
* Removing variables

The list of reduced variables is referred to as VarsEDA.

Then, Logistic Regression, CART and Neural Network models are considered. Discriminant Analysis is not appropriate due to the large number of categorical variables.

For Logistic Regression, three selection methods are applied, namely backward, forward and stepwise.

With CART, four models are considered

1. CART Entropy with VarsEDA
2. CART Entropy with reduced list of variables from model 1
3. CART Gini with VarsEDA
4. CART Gini with reduced list of variables from model 3

Three Neural Network architectures are considered, all using VarsEDA

* One hidden layer of 32 neurons
* One hidden layer of 32 neurons, one hidden layer of 16 neurons
* One hidden layer of 32 neurons, one hidden layer of 48 neurons

Finally, the comparison of predictive performance between Logistic Regression, CART and Neural Network models are given, followed by a conclusion of the project.

# Dataset

## Data Collection

The dataset includes records of students enrolled in undergraduate courses at the Polytechnic Institute of Portalegre, Portugal. It spans academic years from 2008/09 to 2018/2019 and covers various undergraduate degree programs.

Student information is captured at multiple stages: right after enrollment, after the first and second semesters, and again at the end of the standard duration of their respective programs.

More information about the dataset can be found at reference 2.

## Summary

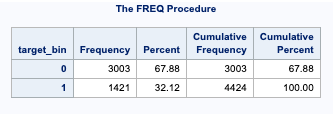
The dataset contained 36 variables and 4424 rows of observations. Of the 36 variables 25 are multiclass categorical variables, 8 variables are binary categorical variables, and 4 were numerical variables. The independent variable contains demographic information of the student at the time of their enrollment, their program, their first semester and second semester.

Some of the variables are listed below:

| marital\_status | Categorical variable where values represent different marital statuses (1 – single, 2 – married, 3 – widower, 4 – divorced, 5 – facto union, 6 – legally separated) |
| --- | --- |
| nationality | nationality of the applicant |
| daytime/evening\_attendance | Binary variable where 1 indicates daytime attendance and 0 indicates evening attendance |
| previous\_qualification\_grade | Numerical value representing the grade of previous qualification (between 0 and 200) |
| inflation\_rate | Numerical value representing the inflation rate of student’s home country at the time of enrollment as a percentage |
| GDP | Numerical value representing the GDP of student’s home country at the time of enrollment |
| mother's\_qualifications | educational qualification of the applicant’s Mother; containing of more than 20 classes |
| mother's\_occupations | occupation of the applicant’s mother; containing more than 20 classes |
| curricular\_units\_1st\_sem\_credit | # transferred credit for 1st semester |
| curricular\_units\_1st\_sem\_enrolled | # credits enrolled for 1st semester |
| curricular\_units\_1st\_sem\_approved | # credits earned after 1st semester |
| curricular\_units\_1st\_sem\_evaluations | # tests in 1st semester |
| curricular\_units\_1st\_sem\_grade | numerical grade average of the 1st semester between 0 - 20 |

The target variable contains three classes: dropout, enrolled and graduate. Since we are interested in students who dropout, dropout is considered class 1 and both enrolled and graduate are combined into class 0 of the binary target.

The distribution of the target variable is as follows



# Data Exploration Analysis (DEA)

With this sparse dataset, it was important to reduce dimensions as much as possible. This process is carried out through three main steps: relying on domain knowledge, using statistical methods and finally removing variables where appropriate.

The following section describes each step and demonstrates how the process is applied on some variables. Please refer to the excel sheet “Modified Variables” and “Excluded Variables” for a detailed explanation for how each of the variables was processed.

## Using Domain Knowledge

In this step, we tried to understand the meaning of each variable and see if we can group them in any meaningful way based on their original meaning.

For variables with larger groupings like father/mother occupation we broke them down into the base groupings combining partial class groupings with their affiliated main group. We were able to similarly group father/mother qualification, as well as marital status, nationality, and previous qualifications.

For example, this is grouping of the variable father’s\_occupation. The classes from 0 to 10 are representative of other classes. For instance, class 4 (Administrative staff) is representative of classes 141, 143 and 144. Based on this knowledge, the classes are reduced to simply the representative classes from 0 to 10.

**0 - Student** **1 - Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers** **2 - Specialists in Intellectual and Scientific Activities** **3 - Intermediate Level Technicians and Professions** **4 - Administrative staff** **5 - Personal Services, Security and Safety Workers and Sellers** **6 - Farmers and Skilled Workers in Agriculture, Fisheries and Forestry** 7 **- Skilled Workers in Industry, Construction and Craftsmen** **8 - Installation and Machine Operators and Assembly Workers** **9 - Unskilled Workers** **10 - Armed Forces Professions** 90 - Other Situation 99 - (blank) 101 - Armed Forces Officers 102 - Armed Forces Sergeants 103 - Other Armed Forces personnel 112 - Directors of administrative and commercial services 114 - Hotel, catering, trade and other services directors 121 - Specialists in the physical sciences, mathematics, engineering and related techniques 122 - Health professionals 123 - teachers 124 - Specialists in finance, accounting, administrative organization, public and commercial relations 131 - Intermediate level science and engineering technicians and professions 132 - Technicians and professionals, of intermediate level of health 134 - Intermediate level technicians from legal, social, sports, cultural and similar services **135 - Information and communication technology technicians** 141 - Office workers, secretaries in general and data processing operators 143 - Data, accounting, statistical, financial services and registry-related operators 144 - Other administrative support staff 151 - personal service workers 152 - sellers 153 - Personal care workers and the like 154 - Protection and security services personnel **161 - Market-oriented farmers and skilled agricultural and animal production workers** 163 - Farmers, livestock keepers, fishermen, hunters and gatherers, subsistence 171 - Skilled construction workers and the like, except electricians 172 - Skilled workers in metallurgy, metalworking and similar 174 - Skilled workers in electricity and electronics 175 - Workers in food processing, woodworking, clothing and other industries and crafts 181 - Fixed plant and machine operators 182 - assembly workers 183 - Vehicle drivers and mobile equipment operators 192 - Unskilled workers in agriculture, animal production, fisheries and forestry 193 - Unskilled workers in extractive industry, construction, manufacturing and transport 194 - Meal preparation assistants **195 - Street vendors (except food) and street service providers**

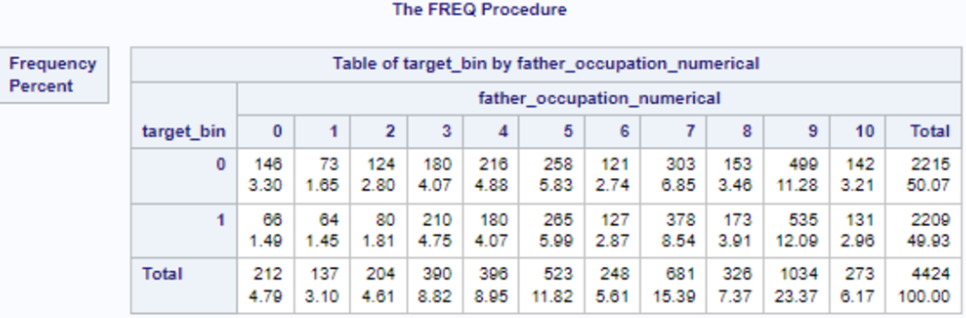
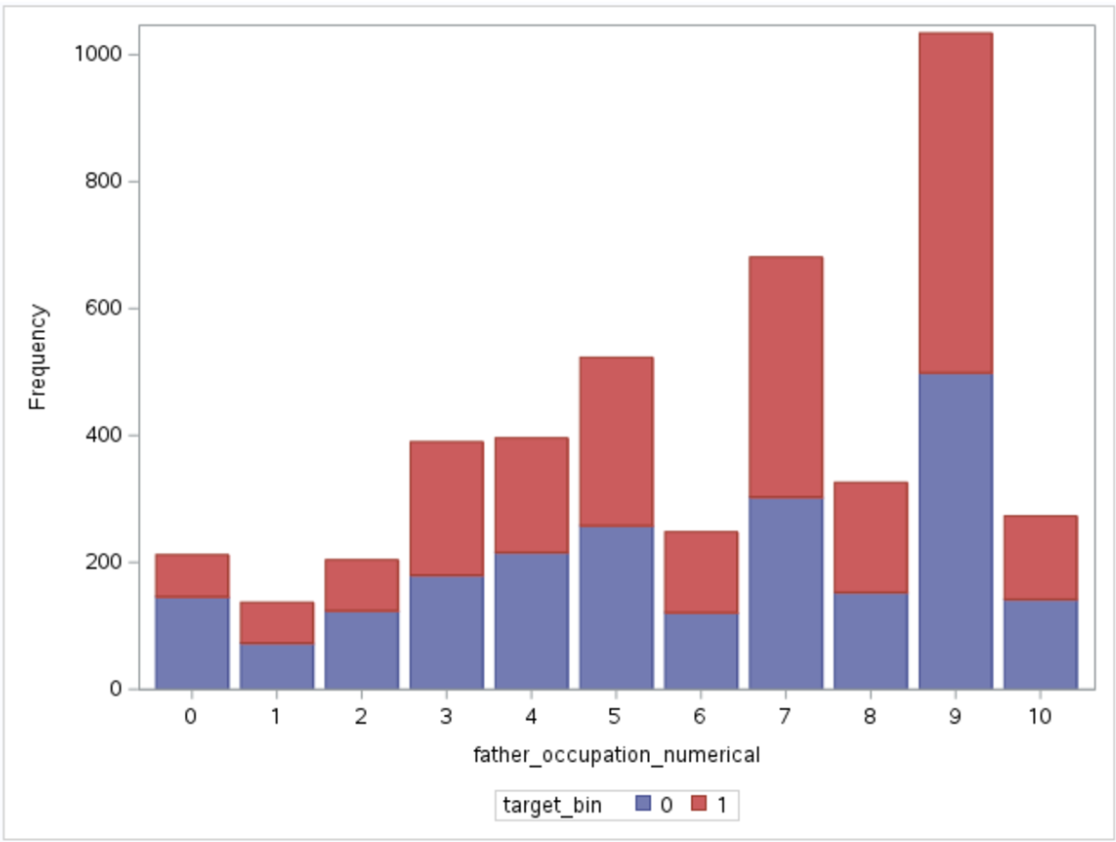
## Statistical Methods

After inspecting the meaning relation amongst individual groups, further grouping of classes in multiclass variables was explored.

Classes that have the same distribution over the two classes of the target variable do not contribute to predicting the target variable and as such are grouped together.

For each multiclass variable, its stacked bar chart over the target variable was examined to spot seemingly similar distributions. Then, exact figures are examined with a frequency table.

As an illustration, below is how this process is applied for the variable father’s\_occupation.



As can be seen from the charts, classes 1, 2 and 4 have similar distributions. The same is true for classes 3, 5, 6, 9 and 10.

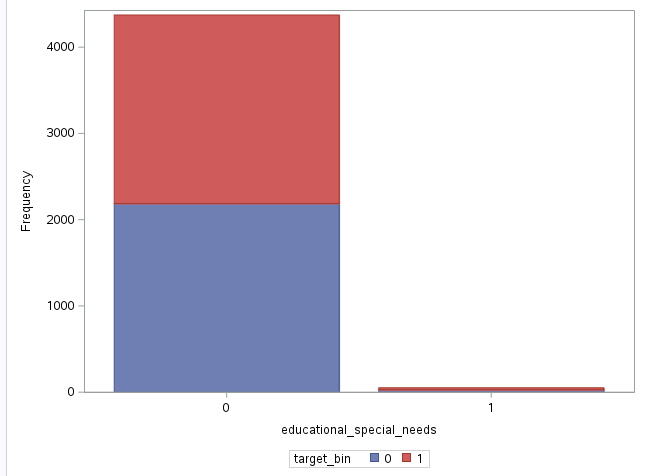
By examining the exact figures from the frequency table, the groupings that were arrived at are (1, 2, & 4), (3, 5, 6, & 8), and (9 & 10), leaving 7 in its own group.

## Removing Variables

Building on the same statistical reasoning above, variables whose classes distribute evenly over the classes of the target variable do not contribute to the prediction of the target variable. Therefore, these variables are not useful predictors and can be removed.

The five variables that were removed during this step are educational\_special\_needs, nationality, international, curricular\_units\_1st\_sem\_credited and curricular\_units\_2nd\_sem\_credited. A sample process of this step for the first three variables are provided below.

From the following bar chart, special needs students were a fractional percentage of the population and their dropout rate was at a near identical level to that of non-special needs students. Therefore it was removed.

A screenshot of a computer screen

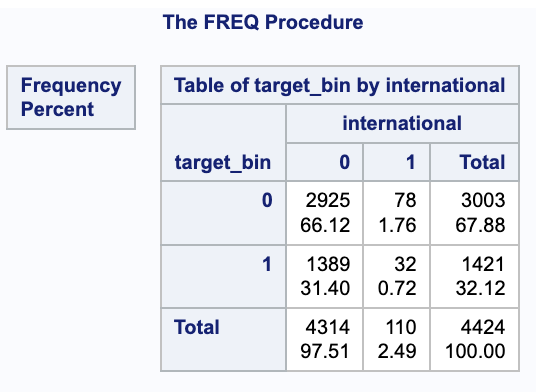
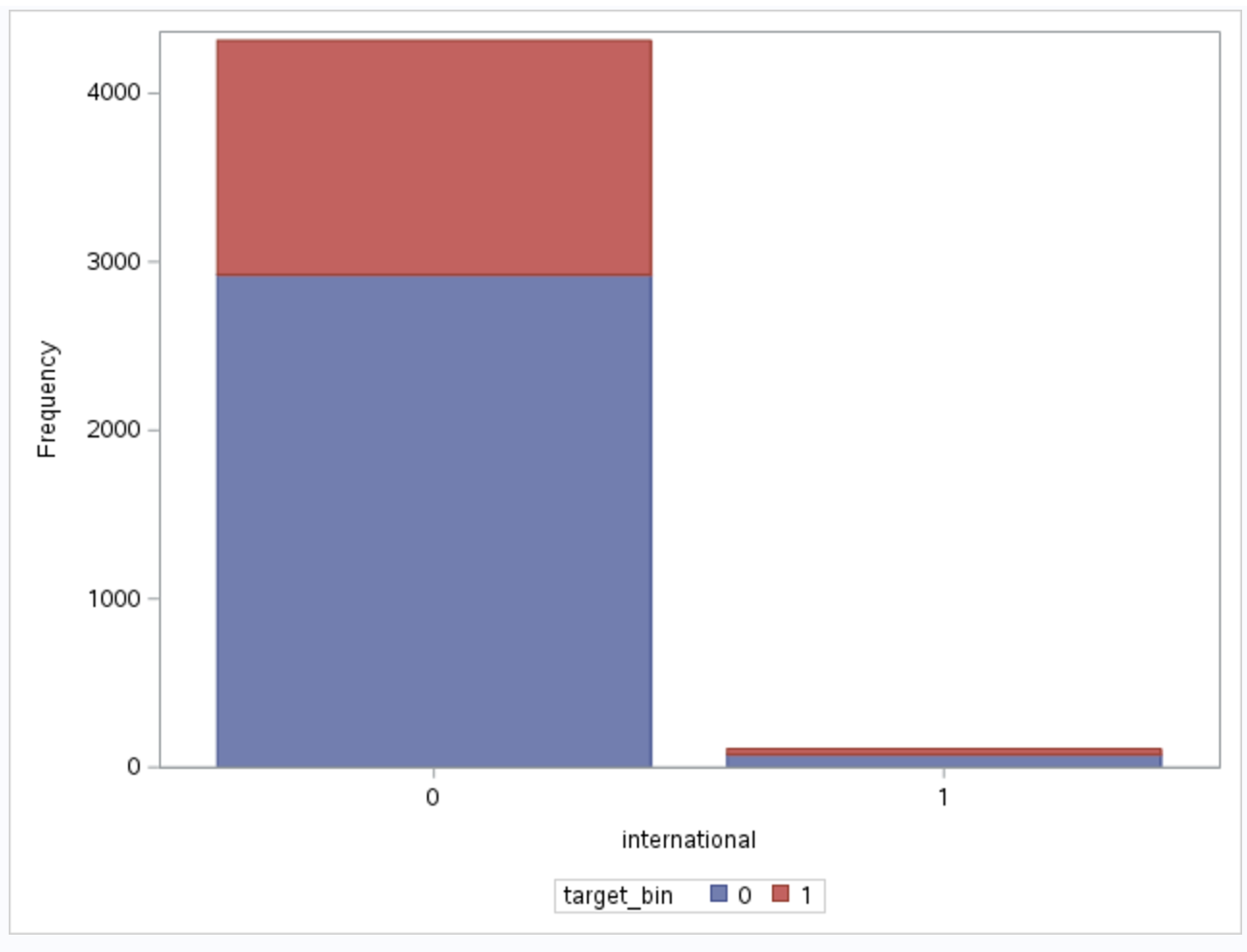
Description automatically generated

The nationality of students was taken out due to the fact that 97% of the data is in one class (Portuguese) and it does not distinguish well between the two target classes.

A screenshot of a graph

Description automatically generated

In the same manner, international students had a dropout percentage that was nearly identical to that of non-international students. Therefore the international independent variable was also removed.



Just as these variables have similar distribution over the classes of the target variables, curricular units (credited) for both semesters also showed little effect on the target\_bin and as such were also removed.

## Summary of Data Reduction

|  | Before Reduction | After Reduction | % Reduction |
| --- | --- | --- | --- |
| # Independent variables | 36 | 30 | 16.67% |
| # independent variables  after one-hot encoding | 517 | 73 | 85.88% |

**This list of reduced variables is referred to as *VarsEDA*.**

**The full list of variables that are removed/transformed can be found in the Variable Information Excel file.**

# Applying Models

After EDA and dimension reduction, the following models that are suitable for binary classification problems are considered: Logistic Regression, CART and Neural Network. Discriminant Analysis is not appropriate due to the large number of categorical variables.

# Logistic Regression

Three models using different selection methods are considered: backward, forward and stepwise. The input data is the dimension-reduced list of variables - VarsEDA.

## Summary Result

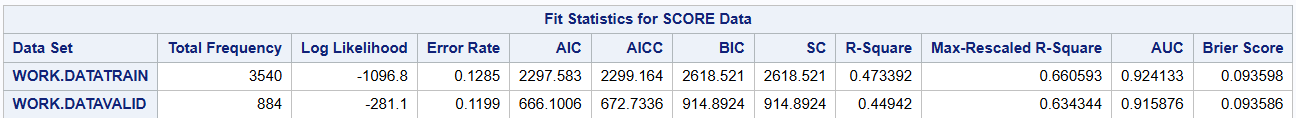
All models were not overfitting. The following results show the performance of each regression model selection method on the validation set.

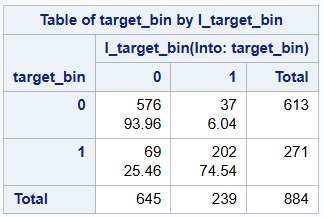
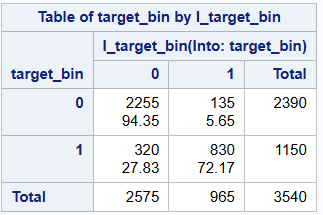
|  | Backward | Forward | Stepwise |
| --- | --- | --- | --- |
| AUC | 91.58 | 91.36 | 91.34 |
| Sensitivity | 74.54 | 74.54 | 73.80 |
| Specificity | 93.96 | 93.96 | 93.47 |
| # Variables | 51 | 49 | 49 |

Forward selection is the best model in terms of measurements and is also the most parsimonious.

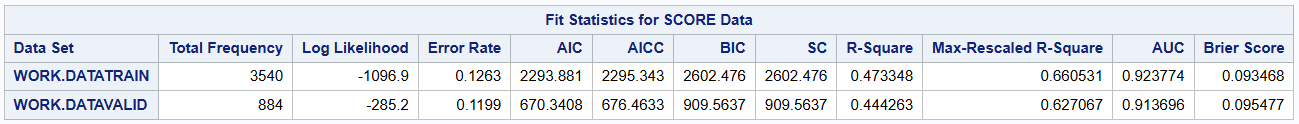
## SAS Figures

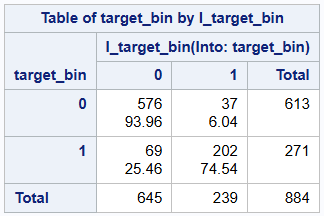
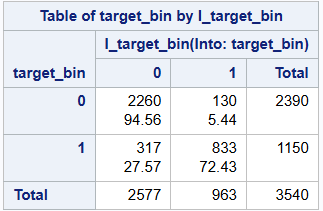
***Backward***



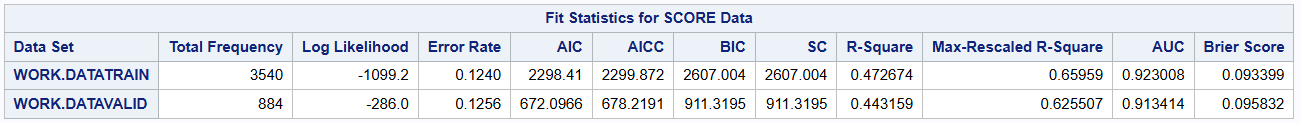
  
training (left) - validation (right) summary

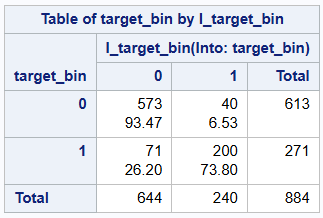
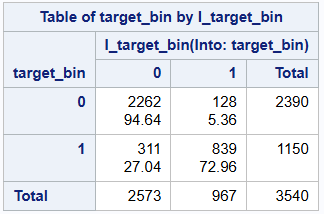
***Forward***



  
training (left) - validation (right) summary

***Stepwise***



  
training (left) - validation (right) summary

# CART

Gini and Entropy growth measures are considered with VarsEDA.

In addition, due to the sparsity of the dataset, it was interesting to see if a model would perform better with an even more reduced dataset than VarsEDA. Therefore, the same growth measure is run again on the list of important variables decided by the previous one.

As such, four CART models are considered in total:

* CART Gini with VarsEDA
* CART Gini with vars reduced by CART Gini
* CART Entropy VarsEDA
* CART Entropy with vars reduced by CART Entropy

## Summary Result

All models were not overfitting. The following results show the performance of each CART model on the validation set.

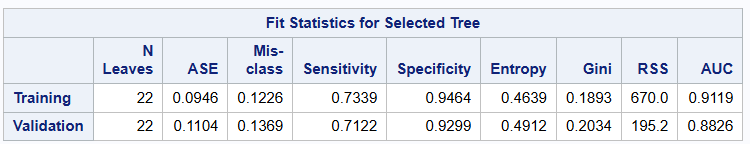
|  | Entropy VarsEDA | Entropy (reduced) | Gini VarsEDA | Gini (reduced) |
| --- | --- | --- | --- | --- |
| AUC | 88.26 | 85.34 | 86.68 | 86.37 |
| Sensitivity | 71.22 | 65.31 | 69.74 | 66.79 |
| Specificity | 92.99 | 94.94 | 93.47 | 94.29 |
| # Leaves | 22 | 15 | 13 | 10 |

Entropy VarsEDA is the best model in terms of measurements whereas GiniVarsEDA is the most parsimonious.

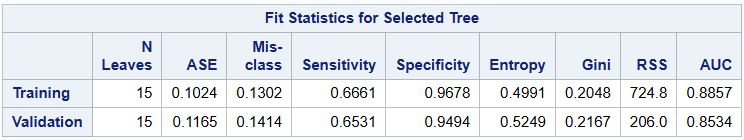
Since the comparison between Logistic Regression - CART - Neural Networks models is based on predictive capability (especially sensitivity measurement), not parsimony, Entropy VarsEDA is considered the best CART model.

## SAS Figures

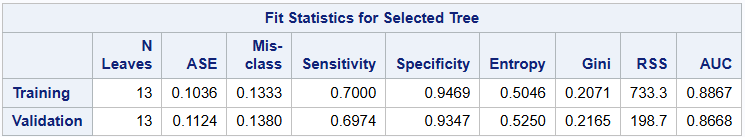
***CART using Entropy with VarsEDA***



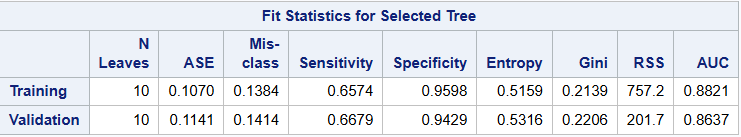
***CART using Entropy with reduced list of variables***



***Cart using Gini with VarsEDA***



***CART using Gini with reduced list of variables***

******

# Neural Network

Three Neural Network architectures are considered, all using VarsEDA

* One hidden layer of 32 neurons
* One hidden layer of 32 neurons, one hidden layer of 16 neurons
* One hidden layer of 32 neurons, one hidden layer of 48 neurons

## Summary Result

All models were not overfitting. The following results show the performance of each neural network on the validation set.

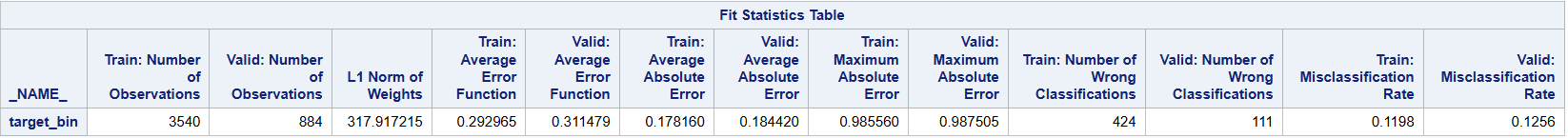
|  | 32 | 32 & 16 | 32 & 48 |
| --- | --- | --- | --- |
| Sensitivity | 73.06 | 70.85 | 73.80 |
| Specificity | 93.80 | 94.78 | 94.45 |
| Misclassification | 12.56 | 12.56 | 11.88 |

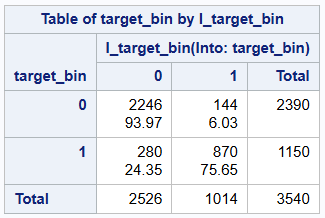
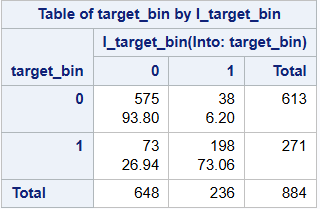
The model with two hidden layers of 32 and 48 neurons is the best model in terms of measurements whereas the one with hidden one layer of 32 neurons is the most parsimonious.

Like the CART models, since the final comparison between different types of models are based on predictive performance only with an emphasis on sensitivity measurement, the model with two hidden layers of 32 and 48 neurons is the best neural network model.

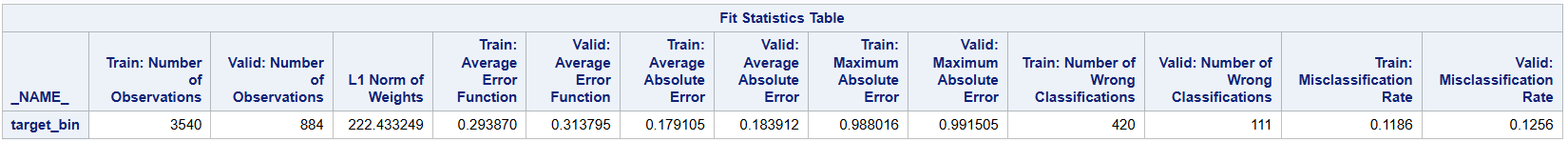
## SAS Figures

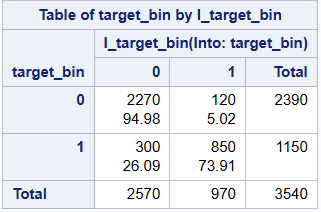
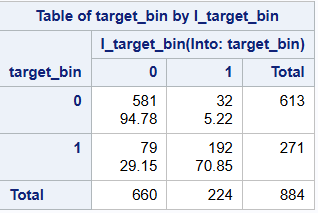
***1 Hidden Layer 32 Neurons***



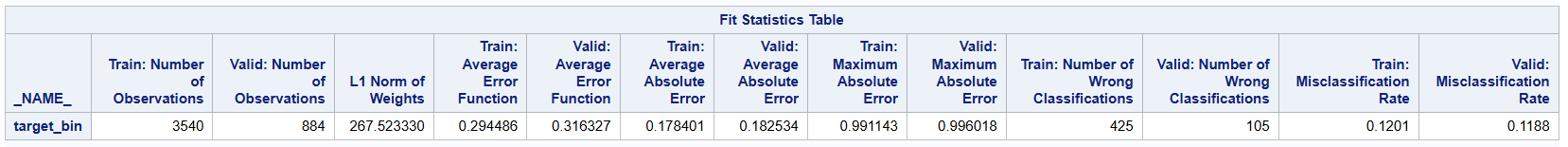
 

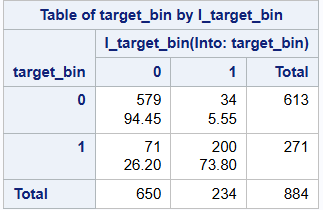
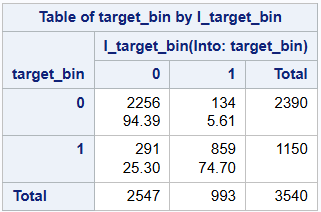
***2 Hidden Layers 32 & 16 Neurons***



***2 Hidden Layers 32 & 48 Neurons***





# Comparison Different Model Types: Logistic Regression - CART - Neural Network

The following results show the performance of the best model of its type on the validation set.

|  | Logistic Regression  Forward Selection | CART Entropy VarsEDA | NN 32 - 48 |
| --- | --- | --- | --- |
| Sensitivity | 74.54 | 71.22 | 73.80 |
| Specificity | 93.96 | 92.99 | 94.45 |
| Misclassification | 11.99 | 13.69 | 11.88 |

Even though the Neural Network model scores best in both Specificity and Misclassification, since Sensitivity is emphasized, Logistic Regression with Forward Selection is the best model for our problem.

# Conclusion

The best model for classification is Logistic Regression. However, in general all models perform well (difference of ~1%).

Even with the best model, sensitivity is only 74.50% which is significantly better than a random guess but not accurate enough for practical use. The model can be used with other information to help universities make judgements on which students should receive early support.

If a user of this model were to consider downsampling methods, that could be a possible way to reduce the effect of imbalance data. Getting as close to a 100% predictor as possible in a university environment would be the most desired outcome. Universities need to be sure they are implementing a model like this that is the most effective in its nature. To combat the student dropout rates, as Universities need a model that is most effective in targeting students at risk of dropping out.

# Appendix - Sample SAS Code

## Statistical Methods for EDA.

The following code shows stacked barcharts of classes of educational\_special\_needs over the classes of the target class. In addition, the exact figures are shown with a frequency distribution table. The same kind of code is used for every categorical variable.

proc freq data=data;

table target\_bin \* educational\_special\_needs / nocol norow out=prepared\_data;

run;

proc sgplot data=data;

vbar "educational\_special\_needs"n / group=target\_bin;

run;

## Logistic regression

### Backward

proc logistic data=datatrain outmodel=backout;

model target\_bin(event="1")=student\_secondary\_edu -- gdp / selection=backward;

run;

proc logistic inmodel=backout;

score data=datatrain fitstat out=dataout;

score data=datavalid fitstat out=bankvalidout;

run;

proc freq data=dataout;

table target\_bin\*i\_target\_bin/nocol nopercent;

run;

proc logistic data=datatrain outmodel=backout1;

model target\_bin(event="1")="1st\_7\_enrolled"n "1st\_5\_enrolled"n "1st\_6\_enrolled"n "1st\_gt\_7\_enrolled"n "2nd\_5\_enrolled"n "2nd\_6\_enrolled"n "2nd\_7\_enrolled"n "2nd\_8\_enrolled"n "2nd\_gt\_8\_enrolled"n "age\_lt20"n "age\_lt25\_gt20"n "age\_lt30\_gt25"n "CU1sem\_approv1grp\_5"n "CU1sem\_approv1grp\_lt5"n "CU1sem\_approv1grp\_gt5"n "cu1st\_sem\_without\_eval\_0"n "cu1st\_sem\_without\_eval\_1\_2"n "CU2sem\_Eval1grp\_lt6"n "CU2sem\_Eval1grp\_gt9"n "CU2sem\_Eval1grp\_679"n "CU2sem\_Eval1grp\_8"n "CU2sem\_approv1grp\_0"n "CU2sem\_approv1grp\_1234"n "CU2sem\_approv1grp\_gt5"n "CU2sem\_approv1grp\_5"n "mother\_occupation\_grp0"n "mother\_occupation\_grp168"n "mother\_occupation\_grp2357"n "mother\_occupation\_grp4910"n "father\_occupation\_grp0"n "father\_occupation\_grp124"n "father\_occupation\_grp3568"n "father\_occupation\_grp7"n “father\_occupation\_grp910"n "Course\_33"n "Course\_171"n "Course\_9003"n "Course\_9070"n "Course\_9085"n "Course\_9119"n "Course\_9130"n "Course\_9147\_9670"n "Course\_9238\_8014"n "Course\_9254"n "Course\_9500"n "Course\_9556"n "Course\_9773"n "Course\_9853"n "Course\_9991"n "curricular\_units\_2nd\_sem\_grade"n "2nd\_sem\_approved\_all\_enrolled"n "displaced"n "debtor"n "tuition\_fees\_up\_to\_date"n "gender"n "scholarship\_holder"n unemployment\_rate;

run;

proc logistic inmodel=backout1;

score data=datatrain fitstat out=datatrainb;

score data=datavalid fitstat out=datavalidb;

run;

proc freq data=datatrainb;

table target\_bin\*i\_target\_bin/nocol nopercent;

run;

proc freq data=datavalidb;

table target\_bin\*i\_target\_bin/nocol nopercent;

run;

### Forward

proc logistic data=datatrain outmodel=forwardred;

model target\_bin(event="1")=student\_secondary\_edu -- gdp / selection=forward;

run;

proc logistic inmodel=forwardred;

score data=data fitstat out=dataout2;

run;

proc freq data=dataout2;

table target\_bin\*i\_target\_bin/nocol nopercent;

run;

proc logistic data=datatrain outmodel=forward1;

model target\_bin(event="1")="1st\_5\_enrolled"n "1st\_6\_enrolled"n "1st\_7\_enrolled"n "1st\_gt\_7\_enrolled"n app\_phase app\_older23 app\_inter age\_lt20 age\_lt25\_gt20 age\_lt30\_gt25 CU1sem\_Eval1grp\_678 CU1sem\_Eval1grp\_lt6 CU1sem\_Eval1grp\_gt8 CU1sem\_approv1grp\_5 CU1sem\_approv1grp\_lt5 CU1sem\_approv1grp\_gt5 CU2sem\_Eval1grp\_679 CU2sem\_Eval1grp\_8 CU2sem\_Eval1grp\_lt6 CU2sem\_Eval1grp\_gt9 CU2sem\_approv1grp\_5 CU2sem\_approv1grp\_0 CU2sem\_approv1grp\_1234 CU2sem\_approv1grp\_gt5 mother\_occupation\_grp0 mother\_occupation\_grp168 mother\_occupation\_grp2357 mother\_occupation\_grp4910 father\_occupation\_grp0 father\_occupation\_grp124 father\_occupation\_grp3568 father\_occupation\_grp7 father\_occupation\_grp910 Course\_33 Course\_171 Course\_9003 Course\_9070 Course\_9085 Course\_9119 Course\_9130 Course\_9147\_9670 Course\_9238\_8014 Course\_9254 Course\_9500 Course\_9556 Course\_9773 Course\_9853 Course\_9991 "curricular\_units\_2nd\_sem\_grade"n "2nd\_sem\_approved\_all\_enrolled"n displaced debtor tuition\_fees\_up\_to\_date gender scholarship\_holder unemployment\_rate;

run;

proc logistic inmodel=forward1;

score data=datatrain fitstat out=datatrainf;

score data=datavalid fitstat out=datavalidf;

run;

proc freq data=datatrainf;

table target\_bin\*i\_target\_bin/nocol nopercent;

run;

proc freq data=datavalidf;

table target\_bin\*i\_target\_bin/nocol nopercent;

run;

### Stepwise

proc logistic data=datatrain outmodel=stepout;

model target\_bin(event="1")=student\_secondary\_edu -- gdp / selection=stepwise;

run;

proc logistic inmodel=stepout;

score data=data fitstat out=dataout4;

run;

proc freq data=dataout4;

table target\_bin\*i\_target\_bin/nocol nopercent;

run;

proc logistic data=datatrain outmodel=stepout1;

model target\_bin(event="1")=student\_higher\_edu student\_secondary\_edu "2nd\_5\_enrolled"n "2nd\_6\_enrolled"n "2nd\_7\_enrolled"n "2nd\_8\_enrolled"n "2nd\_gt\_8\_enrolled"n app\_phase app\_older23 app\_inter age\_lt20 age\_lt25\_gt20 age\_lt30\_gt25 CU1sem\_approv1grp\_5 CU1sem\_approv1grp\_lt5 CU1sem\_approv1grp\_gt5 CU2sem\_Eval1grp\_679 CU2sem\_Eval1grp\_8 CU2sem\_Eval1grp\_lt6 CU2sem\_Eval1grp\_gt9 CU2sem\_approv1grp\_5 CU2sem\_approv1grp\_0 CU2sem\_approv1grp\_1234 CU2sem\_approv1grp\_gt5 mother\_occupation\_grp0 mother\_occupation\_grp168 mother\_occupation\_grp2357 mother\_occupation\_grp4910 father\_occupation\_grp0 father\_occupation\_grp124 father\_occupation\_grp3568 father\_occupation\_grp7 father\_occupation\_grp910 Course\_33 Course\_171 Course\_9003 Course\_9070 Course\_9085 Course\_9119 Course\_9130 Course\_9147\_9670 Course\_9238\_8014 Course\_9254 Course\_9500 Course\_9556 Course\_9773 Course\_9853 Course\_9991 "curricular\_units\_2nd\_sem\_grade"n "2nd\_sem\_approved\_all\_enrolled"n debtor tuition\_fees\_up\_to\_date gender scholarship\_holder unemployment\_rate;

run;

proc logistic inmodel=stepout1;

score data=datatrain fitstat out=datatrains;

score data=datavalid fitstat out=datavalids;

run;

proc freq data=datatrains;

table target\_bin\*i\_target\_bin/nocol nopercent;

run;

proc freq data=datavalids;

table target\_bin\*i\_target\_bin/nocol nopercent;

run;

## CART

### CART using Entropy with VarsEDA

proc hpsplit data=data\_part seed=12345;

partition rolevar=selected(train='1' validate='0');

class "1st\_sem\_approved\_all\_enrolled"n -- target\_bin ;

model target\_bin(event="1")=previous\_qualification\_grade -- fquali;

grow entropy;

prune cc;

run;

### CART using Entropy with reduced list of variables

proc hpsplit data=data\_part seed=12345;

partition rolevar=selected(train='1' validate='0');

class target\_bin CU2sem\_approv1 tuition\_fees\_up\_to\_date course CU1sem\_approv1

"2nd\_sem\_approved\_all\_enrolled"n amode mquali age unemployment\_rate;

model target\_bin(event="1")=CU2sem\_approv1 curricular\_units\_2nd\_sem\_grade

tuition\_fees\_up\_to\_date course CU1sem\_approv1 "2nd\_sem\_approved\_all\_enrolled"n

amode mquali age unemployment\_rate;

grow entropy;

prune cc;

run;

### Cart using Gini with VarsEDA

proc hpsplit data=data\_part seed=12345;

partition rolevar=selected(train='1' validate='0');

class "1st\_sem\_approved\_all\_enrolled"n -- target\_bin ;

model target\_bin(event="1")=previous\_qualification\_grade -- fquali;

grow gini;

prune cc;

run;

### CART using Gini with reduced list of variables

proc hpsplit data=data\_part seed=12345;

partition rolevar=selected(train='1' validate='0');

class target\_bin tuition\_fees\_up\_to\_date CU2sem\_approv1

course CU1sem\_approv1 age unemployment\_rate;

model target\_bin(event="1")=curricular\_units\_2nd\_sem\_grade tuition\_fees\_up\_to\_date

CU2sem\_approv1 course CU1sem\_approv1 age unemployment\_rate;

grow gini;

prune cc;

run;

## Neural Network

### 1 Hidden Layer 32 Neurons

proc hpneural data=data\_part;

partition rolevar=selected(train=1);

target target\_bin/level=nom;

input "1st\_sem\_approved\_all\_enrolled"n -- fquali/level=nom;

input previous\_qualification\_grade -- curricular\_units\_2nd\_sem\_grade/level=int;

hidden 20;

train numtries=10 maxiter=1000;

id target\_bin selected;

score out=neuralout1;

run;

proc freq data=neuralout1;

table target\_bin\*i\_target\_bin/nopercent nocol;

where selected=1;

run;

proc freq data=neuralout1;

table target\_bin\*i\_target\_bin/nopercent nocol;

where selected=0;

run;

### 2 Hidden Layers 32 & 16 Neurons

proc hpneural data=data\_part;

partition rolevar=selected(train=1);

target target\_bin/level=nom;

input "1st\_sem\_approved\_all\_enrolled"n -- fquali/level=nom;

input previous\_qualification\_grade -- curricular\_units\_2nd\_sem\_grade/level=int;

hidden 32;

hidden 16;

train numtries=10 maxiter=1000;

id target\_bin selected;

score out=neuralout2;

run;

proc freq data=neuralout2;

table target\_bin\*i\_target\_bin/nopercent nocol;

where selected=1;

run;

proc freq data=neuralout2;

table target\_bin\*i\_target\_bin/nopercent nocol;

where selected=0;

run;

### 2 Hidden Layers 32 & 48 Neurons

proc hpneural data=data\_part;

partition rolevar=selected(train=1);

target target\_bin/level=nom;

input "1st\_sem\_approved\_all\_enrolled"n -- fquali/level=nom;

input previous\_qualification\_grade -- curricular\_units\_2nd\_sem\_grade/level=int;

hidden 32;

hidden 48;

train numtries=10 maxiter=1000;

id target\_bin selected;

score out=neuralout3;

proc freq data=neuralout3;

table target\_bin\*i\_target\_bin/nopercent nocol;

where selected=1;

run;

proc freq data=neuralout3;

table target\_bin\*i\_target\_bin/nopercent nocol;

where selected=0;

run;

# Reference

1. Kaggle dataset: <https://www.kaggle.com/datasets/syedfaizanalii/predict-students-dropout-and-academic-success>
2. Original source of dataset with related paper: <https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success>
3. Current College Dropout Rates

Bouchrika, Imed. “College Dropout Rates: 2022 Statistics by Race, Gender &

Income.” *Research.com*, 26 Sept. 2022, retrieved from:

research.com/universities-colleges/college-dropout-rates.

Hanson, Melanie. “College Dropout Rates.” *Education Data Initiative*, 17 June

2022, retrieved from: educationdata.org/college-dropout-rates.