Project 2

02450 Introduction to Machine Learning and Data Mining

Authors

Niels Torp Grønskov, **s204510**
Yiming Zhang, **s232896**

Contributions

	Section 1	Section 2	Intro + Discussion	Exam questions
s232896	25%	75%	50%	50%
s204510	75%	25%	50%	50%

Thursday 16th November, 2023

Introduction

This project investigates the machine learning aspects of a data set which consists of various student-related attributes, to predict academic outcomes. Concretely, we will investigate various ML models to predict 'Previous qualification (grade)' via regression and models to classify the 'Target' feature of the data set.

1 Regression

Part a

The initial purpose of this data set was to identify and minimize the amount of students who does not finish a course during their studies. However, one could imagine a similar but slightly different data set were previous performance indicators were not included. For this sake, we will in this regression task try to focus on the variable 'Previous qualification (grade)'.

Our prediction will be based upon all the variables in the data set, excluding the 'target' variable and the 'Admission grade'. This broad spectrum of variables is selected to best possibly encompass the student's characteristics and their educational environments. The 'Admission grade' variable is excluded from this experiment because it is heavily correlated with the 'Previous qualification (grade)' variable.

The categories of the included variables range from personal attributes and family background to economic indicators, details of the application process, and earlier academic performance. Personal demographics provide insights into the students' backgrounds, while parental information might reflect familial socio-economic influences. Economic indicators help in understanding the broader context affecting students, and application details offer a glimpse into the students' initial academic choices. Finally, academic performance metrics are expected to be crucial in gauging students' future engagement and success in their following courses.

A full list of the variables included in the model is here shown:

- Marital status
- Application mode
- Application order
- Course
- Daytime/evening attendance
- Previous qualification
- Nationality
- Mother's qualification
- Father's qualification
- Mother's occupation
- Father's occupation
- Displaced
- Educational special needs
- Debtor
- Tuition fees up to date
- Gender
- Scholarship holder
- Age at enrollment
- International
- Curricular units 1st sem (credited)
- Curricular units 1st sem (enrolled)
- Curricular units 1st sem (evaluations)
- Curricular units 1st sem (approved)
- Curricular units 1st sem (grade)
- Curricular units 1st sem (without evaluations)
- Curricular units 2nd sem (credited)
- Curricular units 2nd sem (enrolled)
- Curricular units 2nd sem (evaluations)
- Curricular units 2nd sem (approved)
- Curricular units 2nd sem (grade)
- Curricular units 2nd sem (without evaluations)
- Unemployment rate
- Inflation rate
- GDP

The goal of the regression is to construct a model which can predict the value of the 'Previous qualification (grade)' variable based on the other features in the data set. Furthermore, the goal is to quantify the performance of this regression model and evaluate how well it will generalize to predict similar values for new unseen data points.

As suggested in the appendix of the original paper [Rea+22] we have done a rank transformation. Further more the data matrix $\mathbf{X}$ have been centered and standardized such that each column has a mean of 0 and a standard deviation of 1.

Regularization parameter λ

The regularization parameter λ is introduced to control the complexity of the model, with the aim of preventing overfitting. By increasing the value of λ, we can penalize the magnitude of the coefficients in the regression model, effectively reducing overfitting but potentially increasing underfitting.

To further understand the generalization errors of the model, we plot the train and validation errors against different λ values on a log-log scale.

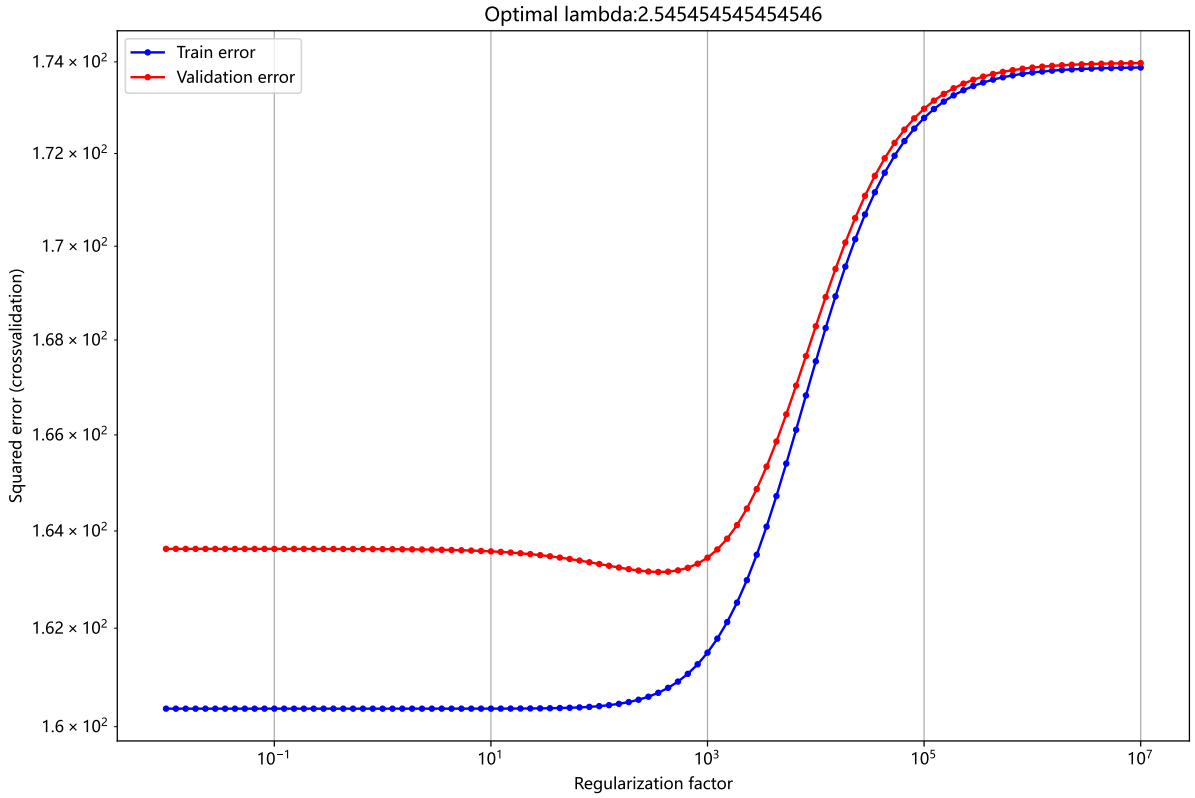


Figure 1: The estimated generalization error as a function of λ

The optimal λ is highlighted in the plot title and is chosen as the one which minimizes the validation error. The plot suggests that as λ increases, both training and validation errors remain stable for a while before slightly decreasing followed by a sharp increase. This indicates generalization improvement followed by a point of underfitting where the model becomes too simple to capture the underlying pattern in the data.

The output of the model

Below is presented a plot which shows mean coefficient values for the different attributes in your data set against different λ values on a log scale.

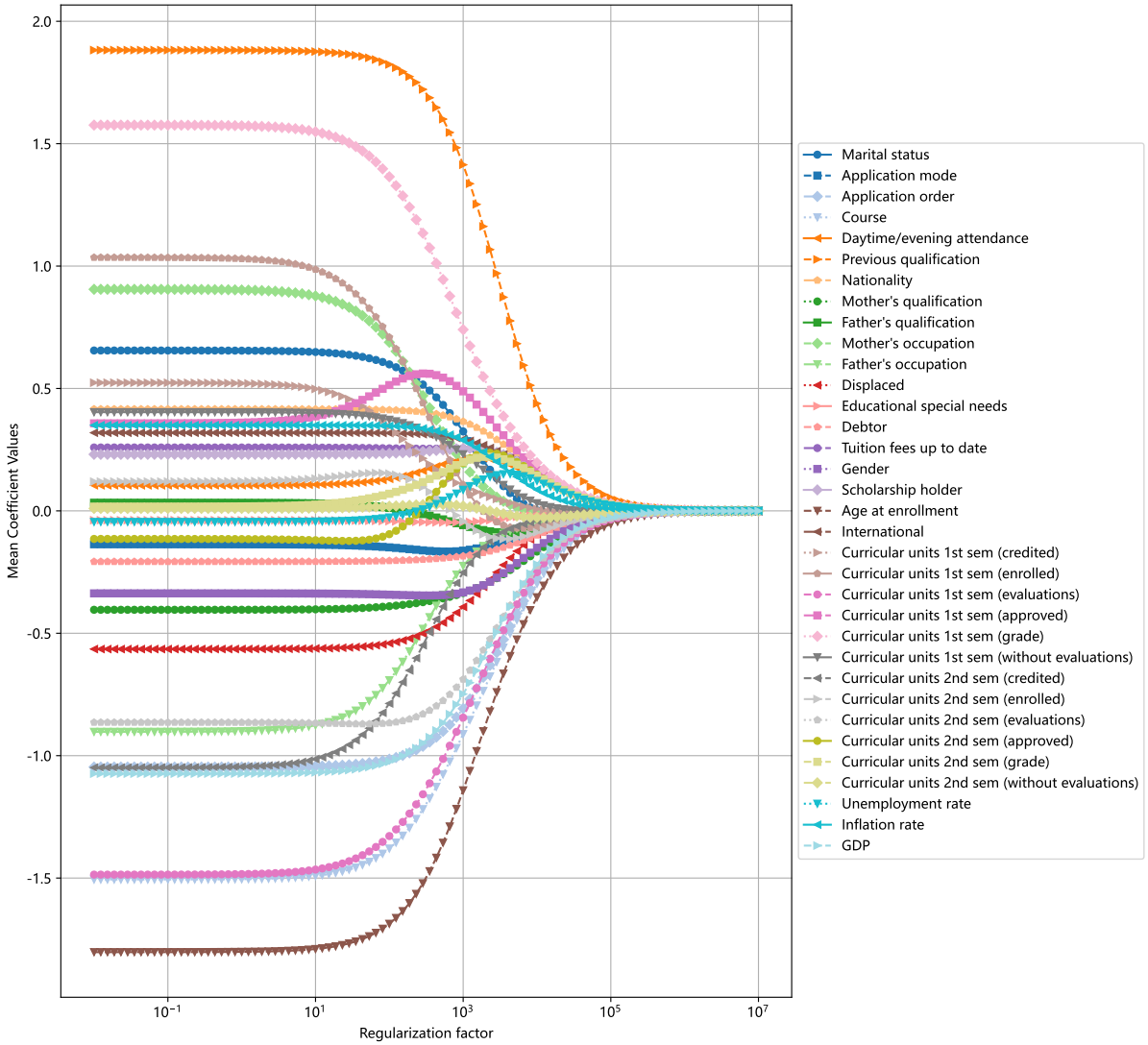


Figure 2: The mean of the coefficient values as a function of the regularization factor λ

This plot helps in understanding how the coefficients diminishes as regularization strength increases. From the coefficient values plot, one can determine the effect of each individual attribute on the output $\mathbf{y}$. A coefficient value indicates the size and direction of the contribution of its corresponding attribute to the model's output. If a coefficient is large and positive, an increase in the corresponding attribute value will result in a significant increase in $\mathbf{y}$. Conversely, a large negative coefficient means that an increase in the attribute value will significantly decrease $\mathbf{y}$. A coefficient close to zero implies that the attribute has little or no effect on the output.

We can here see that, a prominent predictor variable is the 'Previous qualification', while the 'International' feature is the most negative contribution. Overall the plot seem to roughly align with expectations of the feature contributions.

Part b: Model comparison

In this section, we will compare the performance of three models: a regularized linear regression model from the previous section, an artificial neural network (ANN), and a baseline. As a baseline model, we will apply a linear regression model with no features (it computes the mean of y on the training data, and use this value to predict y on the test data).

Two-level cross-validation

To improve the data fitting of the models, we have implemented two-level cross-validation. Here we especially focus on the complexity limiting factors, namely λ and h_i^* (hidden units in ANN).

The chosen test ranges are $\lambda = 10^{-8}, ..., 10^8$ where 1000 numbers are evenly sampled and $h_i^* \in [1, 8, 16, 24, 32, 40, 48, 56, 64, 96, 128]$.

Outer Fold	**ANN**		**Linear Regression**		**Baseline**
i	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	32	164.1	2.55	169.3	182.4
2	40	145.6	2.44	162.3	176.7
3	32	144.8	2.60	155.6	159.7
4	32	132.1	2.58	140.4	151.5
5	32	152.8	2.46	157.4	166.2
6	40	166.7	2.62	169.7	183.5
7	32	158.7	2.51	164.0	175.6
8	32	166.7	2.59	180.4	194.0
9	32	149.1	2.56	152.7	159.2
10	32	170.4	2.57	180.8	190.1

Table 1: Two-level cross-validation table used to compare the three models in the classification problem

For the ANN, we find that the best number of hidden units is 32 since more or less leads to poorer generalization. Interestingly, the best regularization parameter λ is here found to be around approx. 2.5 (mean value), which is the same as in the previous section were a simpler analysis were done. As expected the baseline model performs the worst.

Statistical evaluation

For the statistical evaluation of the performance of these three models, we used the paired t-test method described in Box 11.3.4. Our null hypothesis is "two models have the same performance".

When comparing the errors of the Linear Regression with the Baseline model, the p-value is evidently above the threshold 0.05, thus the null-hypothesis can not be rejected and there is no significant difference between the models.

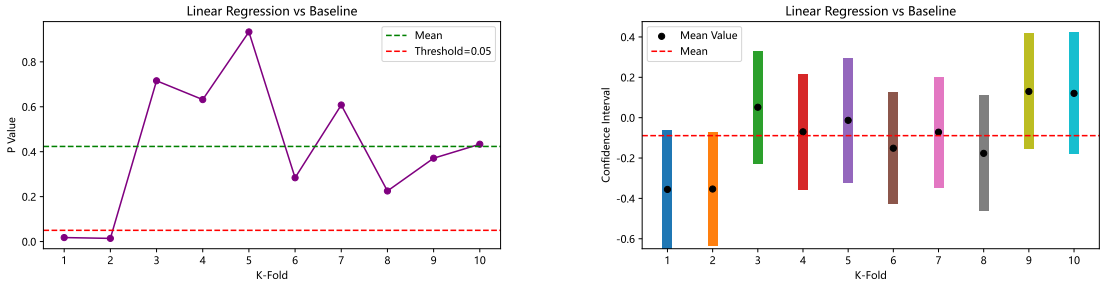


Figure 3

Similarly, when comparing the errors of the ANN with Baseline model, in the most outer fold, the p-value and its mean are greater than 0.05, thus the null hypothesis can not be rejected.

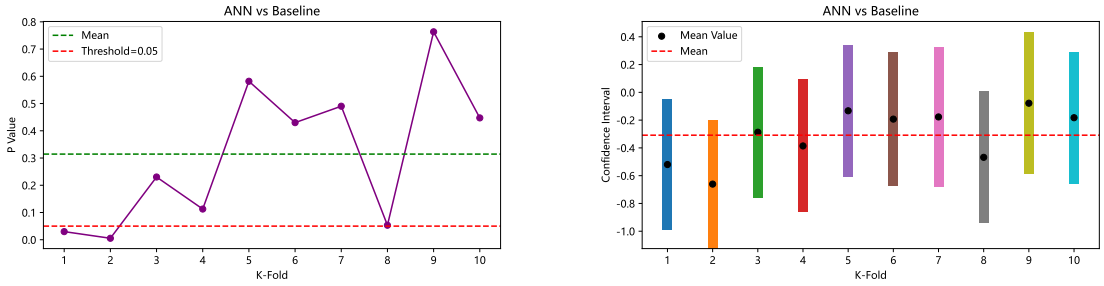


Figure 4

Similarly, when comparing ANN with Linear Regression, p-value and its mean are greater than 0.05, thus the null hypothesis can not be rejected.

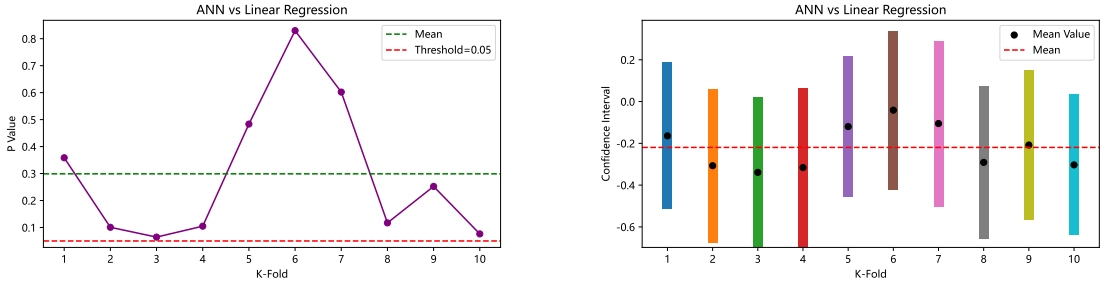


Figure 5

Conclusion on results

According to the statistical evaluations above, these 3 models have same performance. However, summarizing the table and images above, we can conclude that, although ANN and linear regression model do not perform significantly better than baseline, they appear to a slightly lower E^{test} compared to baseline and ANN performs better than linear regression. These results provide us with the knowledge, that further improvements and analysis of the models would be beneficial.

2 Classification

Chosen problem

Following the spirit of the original paper [Rea+22], we will in this section setup models for a multi-class classification where is it predicted if a student will be `Enrolled`, `Dropout` or `Graduate` after the normal period of a course. This prediction variable is also called the 'Target'.
In these models, every feature except the 'Target' variable will be included. A full list of the included variables of the models are here shown:

- Marital status
- Application mode
- Application order
- Course
- Daytime/evening attendance
- Previous qualification
- Previous qualification (grade)
- Nationality
- Mother's qualification
- Father's qualification
- Mother's occupation
- Father's occupation
- Admission grade
- Displaced
- Educational special needs
- Debtor
- Tuition fees up to date
- Gender
- Scholarship holder
- Age at enrollment
- International
- Curricular units 1st sem (credited)
- Curricular units 1st sem (enrolled)
- Curricular units 1st sem (evaluations)
- Curricular units 1st sem (approved)
- Curricular units 1st sem (grade)
- Curricular units 1st sem (without evaluations)
- Curricular units 2nd sem (credited)
- Curricular units 2nd sem (enrolled)
- Curricular units 2nd sem (evaluations)
- Curricular units 2nd sem (approved)
- Curricular units 2nd sem (grade)
- Curricular units 2nd sem (without evaluations)
- Unemployment rate
- Inflation rate
- GDP

This classification will be evaluated with three different models: a logistic regression model, ANN, and a baseline where the baseline will compute the largest class on the training data, and predict everything in the test-data as belonging to that class.

Parameter choices

ANN	**Parameter**
Hidden units number	[1,8,16,24,32,40,48,56,64,96,128]
Batch formalization	added
Drop out	added, drop out rate = 0.2
Activation function	ReLU
Units number of out put layer	3
Activation function(output layer)	Softmax
Loss function	CrossEntropyLoss
Optimizer	Adam
Weight decay	added, weight decay=0.0005
Logistic Regression	**Parameter**
λ	np.logspace(-8,8,1000)

Table 2: Parameter setting

Two-level cross-validation

Outer Fold	ANN		Logistic Regression		Baseline
i	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	24	0.191	0.72	0.216	0.510
2	32	0.246	0.72	0.257	0.534
3	40	0.227	0.24	0.246	0.498
4	24	0.180	1.53	0.198	0.480
5	32	0.253	−0.40	0.246	0.495
6	32	0.226	0.56	0.217	0.504
7	24	0.244	0.24	0.248	0.495
8	32	0.226	1.53	0.230	0.486
9	32	0.217	−8.00	0.210	0.472
10	40	0.246	1.53	0.253	0.527

Table 3: Two-level cross-validation table used to compare the three models in the classification problem

Statistical evaluation

For the statistical evaluation of the performance of these three models, we used the McNemera's test described in Box 11.3.2. Our null hypothesis is "two models have the same performance".

When comparing the errors of the Linear Regression and the Baseline model, the p-value and its mean are significantly less than 0.05, thus we reject H_0 meaning that there is a difference between the models.

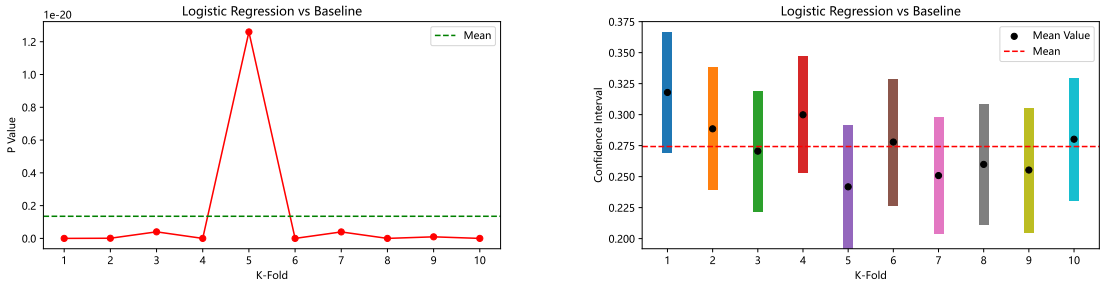


Figure 6

Similarly, when comparing ANN with Baseline, the p-value and its mean are significantly less than 0.05, meaning that we reject H_0 and there is a difference between the models.

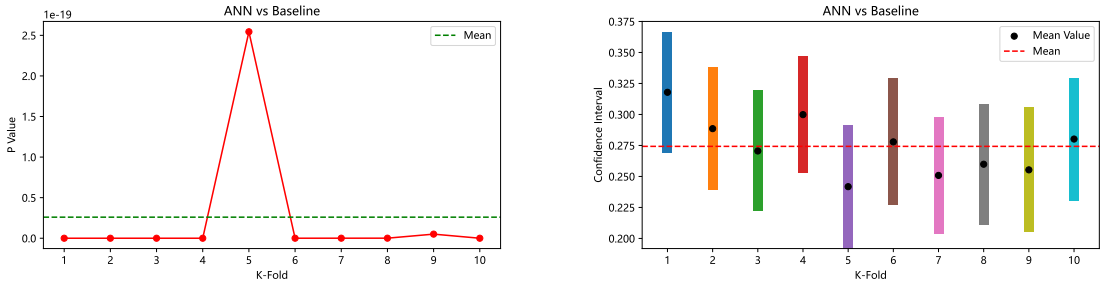
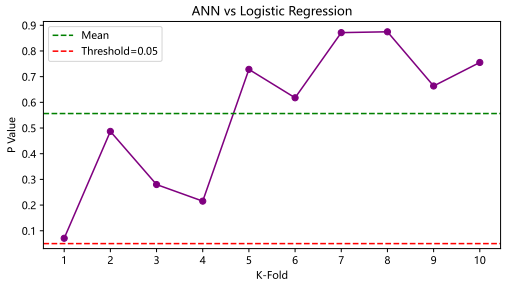


Figure 7

When comparing ANN with Linear Regression, the p-value are greater than 0.05, thus there is no we can not reject H_0 meaning there is no significant difference between the models.



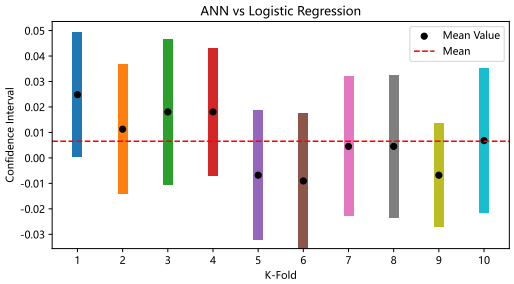


Figure 8

Conclusion on results

Summarizing the table and evaluations above, we can conclude that, ANN and logistic regression perform significantly better than baseline in all fold because they have lower E^{test}. However it seems that ANN and logistic regression have similar performance for this data set.

Discussion

Conclusion from regression and classification

We can conclude that for this data set, there is evidence that comprehensive machine learning models can make predictions about the 'Target' feature. Here the AAN model and logistic regression performed equally good. Further we found that there was a lack of evidence that our models could predict the 'Previous qualification (grade)' feature via regression methods.

In general, we expect that machine learning models can be helpful for analysis of this data set. However one would have to further investigate which model is the best for the specific use case. We can moreover appreciate the significance of each feature in this data set since many of the features contribute very little to the predictions made by the models.

Previous studies of the data set

We found a additional paper [KYK23] citing this dataset for machine learning. This paper examined university dropout predictions using academic, demographic, socioeconomic, and macroeconomic data types in a similar fashin as we did in our studies.

In addition, they performed associated factor analysis to analyze which type of data had the greatest impact on the performance of the machine learning models in predicting graduation and dropout status. Four binary classifiers were trained using these features to determine whether students would graduate or drop out. They tried various machine learning classifiers and found that Random Forests had the best performance. When they excluded all academic-related features from the dataset, the average ROC-AUC score dropped from 0.935 to 0.811. Thus, they found that the data type that had the greatest impact on the model's performance was the academic data. Their preliminary results indicate that a correlation does exist between data types and dropout status.

However, one would have to be careful in comparing our analysis with the results of their paper as their methodology is vastly different. For example, their prediction targets are not the same, while also dividing the features into four groups, i.e. academic, demographic, socioeconomic, and macroeconomic and trained four classification models. In addition, they exclude the status "Enrolled" from the targets, so theirs is a binary classification task while ours is a multi-classification task.

But for the choice of *method*2 in classification, we didn't use a Random Forest model, which the authors of the other paper consider to the best model. In general, the results of this paper still align with our conclusions from earlier; there seem to be evidence of machine learning potential but one would have to investigate further to find the best suited model

References

[KYK23] Sean Kim, Eliot Yoo, and Samuel Kim. "Why Do Students Drop Out? University Dropout Prediction and Associated Factor Analysis Using Machine Learning Techniques". In: *arXiv preprint arXiv:2310.10987* (2023).

[Rea+22] Valentim Realinho et al. "Predicting Student Dropout and Academic Success". In: *Data* 7.11 (Nov. 2022). Number: 11 Publisher: Multidisciplinary Digital Publishing Institute, p. 146. ISSN: 2306-5729. DOI: 10.3390/data7110146. URL: https://www.mdpi.com/2306-5729/7/11/146 (visited on 09/07/2023).

Exam problems

Question 1.

According to:

$$TPR = \frac{TP}{TP + FN}$$
$$FPR = \frac{FP}{FP + TN}$$

Now let's assume that: $\hat{y} > 0.5$ means Predicted Positive, $\hat{y} < 0.5$ means Predicted Negative. All black circles here are Actual Negative. All red crosses here are Actual Positive. So we can plot a table about TP, FN, FP, TN.

	TP	**FN**	**FP**	**TN**
A	3	1	4	0
B	4	0	3	1
C	3	1	4	0
D	4	0	3	1

For A: $TPR = \frac{TP}{TP+FN} = \frac{3}{3+1} = 0.75$, $FPR = \frac{FP}{FP+TN} = \frac{4}{4+0} = 1$
For B: $TPR = \frac{TP}{TP+FN} = \frac{4}{4+0} = 1$, $FPR = \frac{FP}{FP+TN} = \frac{3}{3+1} = 0.75$
For C: $TPR = \frac{TP}{TP+FN} = \frac{3}{3+1} = 0.75$, $FPR = \frac{FP}{FP+TN} = \frac{4}{4+0} = 1$
For D: $TPR = \frac{TP}{TP+FN} = \frac{4}{4+0} = 1$, $FPR = \frac{FP}{FP+TN} = \frac{3}{3+1} = 0.75$
Therefore the option B,D are wrong. Let's continue with assuming: $\hat{y} > 0.62$ means Predicted Positive, $\hat{y} < 0.62$ means Predicted Negative.

	TP	**FN**	**FP**	**TN**
A	2	2	3	1
C	3	1	2	2

For A: $TPR = \frac{TP}{TP+FN} = \frac{2}{2+2} = 0.5$, $FPR = \frac{FP}{FP+TN} = \frac{3}{3+1} = 0.75$
For C: $TPR = \frac{TP}{TP+FN} = \frac{3}{3+1} = 0.75$, $FPR = \frac{FP}{FP+TN} = \frac{2}{2+2} = 0.5$
Option C corresponding to the coordinates of Figure 1.
Therefore the option C is correct

Question 2.

According to

$$ClassError(v) = 1 - \underset{c}{maxp}(c|v)$$

and

$$\triangle = I(r) - \sum_{K=1}^{K=2} \frac{N(v_k)}{N(r)} I(v_k)$$

We can get:

$$I(r) = 1 - \frac{33+4}{33+4+28+2+1+30+3+29+5} = \frac{98}{135}$$
$$I(v_1) = 1 - \frac{33+4}{33+4+28+2+30+3+29+5} = \frac{97}{134}$$
$$I(v_2) = 1 - \frac{1}{1} = 0$$
$$\triangle = I(r) - \sum_{K=1}^{K=2} \frac{N(v_k)}{N(r)} I(v_k) = \frac{98}{135} - \frac{134}{135} * \frac{97}{134} - \frac{1}{135} * 0 = \frac{1}{135} \approx 0.0074$$

Therefore the option C is correct

Question 3.

According to the information provided, we can get:
hidden layer: 7 input features * 10 hidden units + 10 biases = 70 + 10 = 80
output layer: 10 hidden units * 4 output classes + 4 biases = 40 + 4 = 44
Total number of parameters = 80 + 44= 124
Therefore the option A is correct

Question 4.

We select a data point in the top *Congestion level 1* square and follow the rules from option D in the tree starting from the root A.

Is $b_1 \geq -0.76$; True. Therefore next rule is C. Is $b_1 \geq -0.16$; False. Therefore next rule is D. Is $b_2 \geq 0.01$; True. Therefore the point is classified correctly as *Congestion level 1* and this is not true for any of the other options.

Therefore the option D is correct

Question 5.

It can be seen from the table that $K_1 = 5$. We have that $K_2 = 4$ and each model has 5 different parameters to choose from. This gives the following calculations.

$$params = 5$$

$$nn_{train} = 20, \quad nn_{test} = 5, \quad nn = nn_{test} + nn_{train}$$

$$reg_{train} = 8, \quad reg_{test} = 1, \quad reg = reg_{train} + reg_{test}$$

$$models = reg + nn$$

Outer loop runs K_1 times containing the inner loops and and a training and evaluation of the models. The inner loops are firstly for each parameter and then for each inner fold, where the models are trained and tested.

$$time = 3570 = K_1 \cdot ((params \cdot K_2 \cdot models) + models)$$

Therefore the option C is correct