Midterm-2 Project

First and last name: Jingwen Zhong

Submission Date: 04/09/2021

Midterm-2 Project Instruction

In `Midterm-1 Project`, you have built predictive models using train and test data sets about college students' academic performances and retention status. You fitted four regression models on **Term.GPA** and four classification models on **Persistence.NextYear**. the lowest test score of MSE_{test} achieved on the regression problem was .991 using a simple linear regression, and the highest `accuracy` and `F1` scores obtained were 91.15% and 95.65%, respectively, with the fit of a multiple logistic regression model (equivalently, LDA and QDA give similar performances). Let's call these scores as baseline test scores.

In `Midterm-2 Project`, you will use tree-based methods (trees, random forests, boosting) and artificial neural networks (Modules 5, 6, and 7) to improve the baseline results. There is no any answer key for this midterm: your efforts and justifications will be graded, pick one favorite optimal tree-based method and one optimal ANN architecture for each regression and classification problem (a total of two models for classification and two models for regression), and fit and play with hyperparameters until you get satisfactory improvements in the test data set.

Keep in mind that $Persistence.NextYear$ is not included in as predictor the regression models so use all the predictors except that on the regression. For the classification models, use all the predictors including the term gpa.

First of all, combine the train and test data sets, create dummies for all categorical variables, which include `Entry_Term`, `Gender`, and `Race_Ethc_Visa`, so the data sets are ready to be separated again as train and test. (Expect help on this portion!) You will be then ready to fit models.

Tips

- `Term.gpa` is an aggregated gpa up until the current semester, however, this does not include this current semester. In the modeling of `gpa`, include all predictors except `persistent`.
- The data shows the `N.Ws`, `N.DFs`, `N.As` as the number of courses withdrawn, D or Fs, A's respectively in the current semester.
- Some rows were made synthetic so may not make sense: in this case, feel free to keep or remove.
- It may be poor to find linear association between gpa and other predictors (don't include `persistent` in `gpa` modeling).
- Scatterplot may mislead since it doesn't show the density.
- You will use the test data set to asses the performance of the fitted models based on the train data set.
- Implementing 5-fold cross validation method while fitting with train data set is strongly suggested.
- You can use any packs (`caret`, `Superml`, `rpart`, `xgboost`, or visit to search more) as long as you are sure what it does and clear to the grader.
- Include helpful and compact plots with titles.
- Keep at most 4 decimals to present numbers and the performance scores.

- When issues come up, try to solve and write up how you solve or can't solve.
- Check this part for updates: the instructor puts here clarifications as asked.

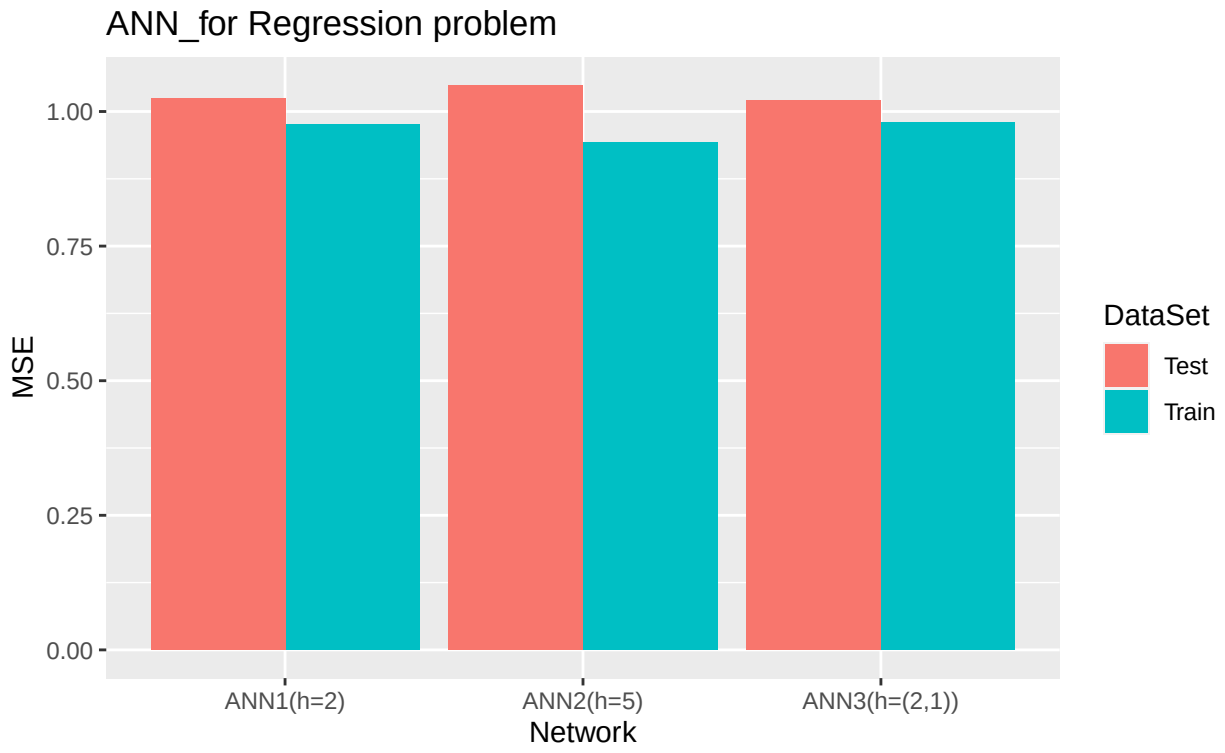
A. Improving Regression Models - 15 pts

- Explore tree-based methods, choose the one that is your favorite and yielding optimal results, and then search for one optimal ANN architecture for the regression problem (so two models to report). Fit and make sophisticated decisions by justifying and writing precisely. Report `the test MSE` results in a comparative table along with the methods so the grader can capture all your efforts on building various models in one table.

Solution for Section A.

Table 1: Optimal Regression Model using Boosting Method

ntree	shrinkage	interaction_depth	train_MSE	test_MSE
1000	0.002	4	0.9828	0.9751



For the tree based regression model, I choose boosting method, and I set range of ntrees 1000,2000,3000,4000 to 5000, shrinkage 0.002, 0.01, 0.025, 0.05, 0.1. Then Run the loop to get the minimal Test MSE

For ANN architecture regression model, I tried hidden layer 2, hidden layer 5, and hidden layer(2,1),and I choose logistic method. From the graph above, for the final optimal model I choose ANN with hidden layer(2,1)

Table 2: Comparison of 2 optimal models for regression

	Train MSE	Test MSE
Boosting	0.9828309	0.9750715
ANN	0.9805923	1.0212423

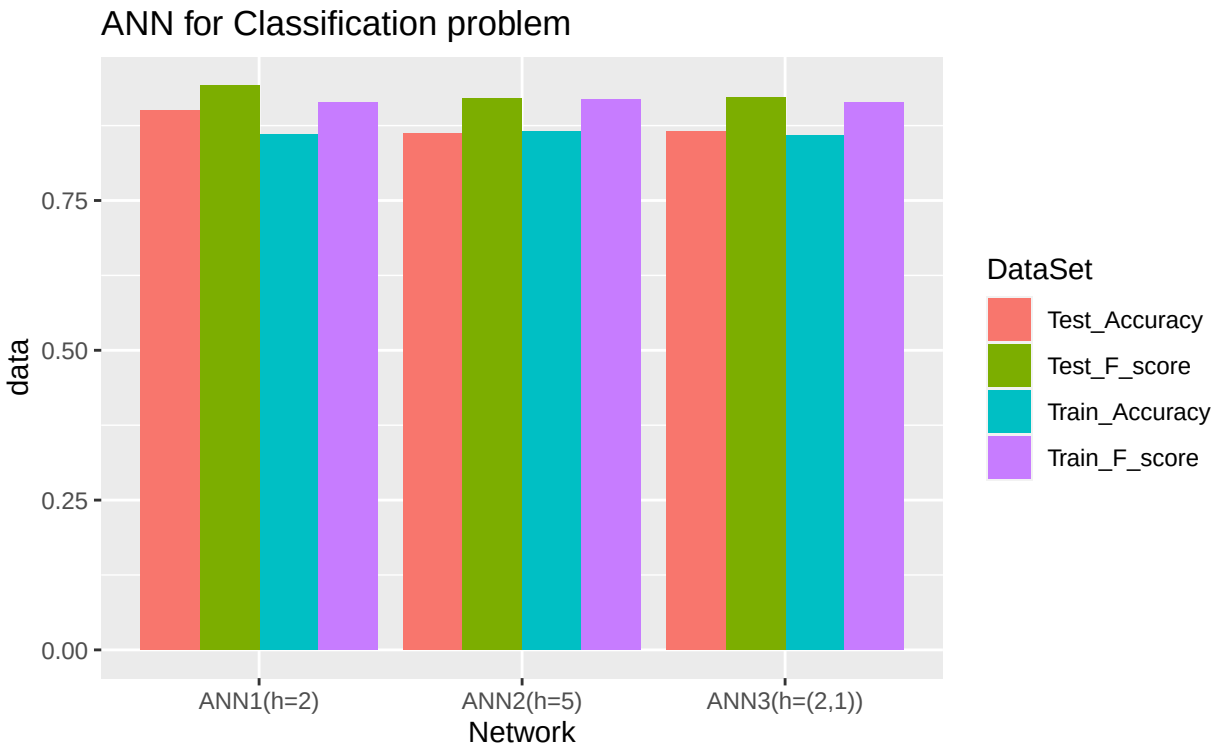
B. Improving Classification Models - 20 pts

- Explore tree-based methods, choose the one that is your favorite and yielding optimal results, and then search for one optimal ANN architecture for the classification problem (so two models to report). Fit and make sophisticated decisions by justifying and writing precisely. Report `the test accuracy` and `the test F1` results in a comparative table along with the methods so the grader can capture all your efforts in one table.

Solution for Section B.

Table 3: Optimal Classification Models Using Random Forest

ntree	mtry	train_accuracy	highest_test_accuracy	train_f_score	test_f_score
150	3	0.9814	0.955	0.9885	0.975



For the tree based classification model, I choose random forest method, and I set range of ntrees from 100 to 400, mtry from 2 to 8 since the square root of length(train is around 4). Then Run the loop to get the highest accuracy and F score.

For ANN architecture regression model, I tried hidden layer 2, hidden layer 5, and hidden layer(2,1), and from the graph above I should choose ANN with hidden layer 2 but the test accuracy is higher than the training one, which is not acceptable, same as ANN with hidden layer(2,1). So I can only choose ANN with hidden layer 5.

Table 4: Comparison of 2 optimal models for classification

	Train accuracy	Test accuracy	Train F score	Test F score
Random Forest	0.9814139	0.9550173	0.9884785	0.975048
ANN	0.8653813	0.8622837	0.9180848	0.919855

C. Importance Analyses - 15 pts

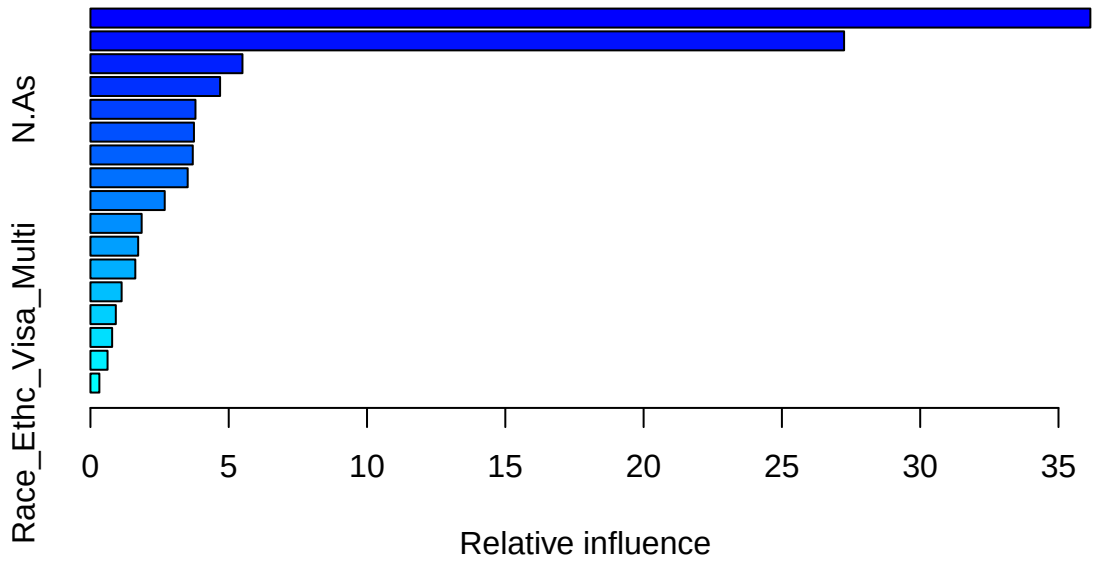
- Part a. Perform an importance analysis on the best regression model: which three predictors are most important or effective to explain the response variable? Find the relationship and dependence of these predictors with the response variable. Include graphs and comments.

---

solution for Section C. Part a

```
## The importance of predictors in the boosting for regression:

##                                    var   rel.inf
## HSGPA                            HSGPA 36.152039
## SAT_Total                    SAT_Total 27.247875
## N.RegisteredCourse N.RegisteredCourse  5.496849
```



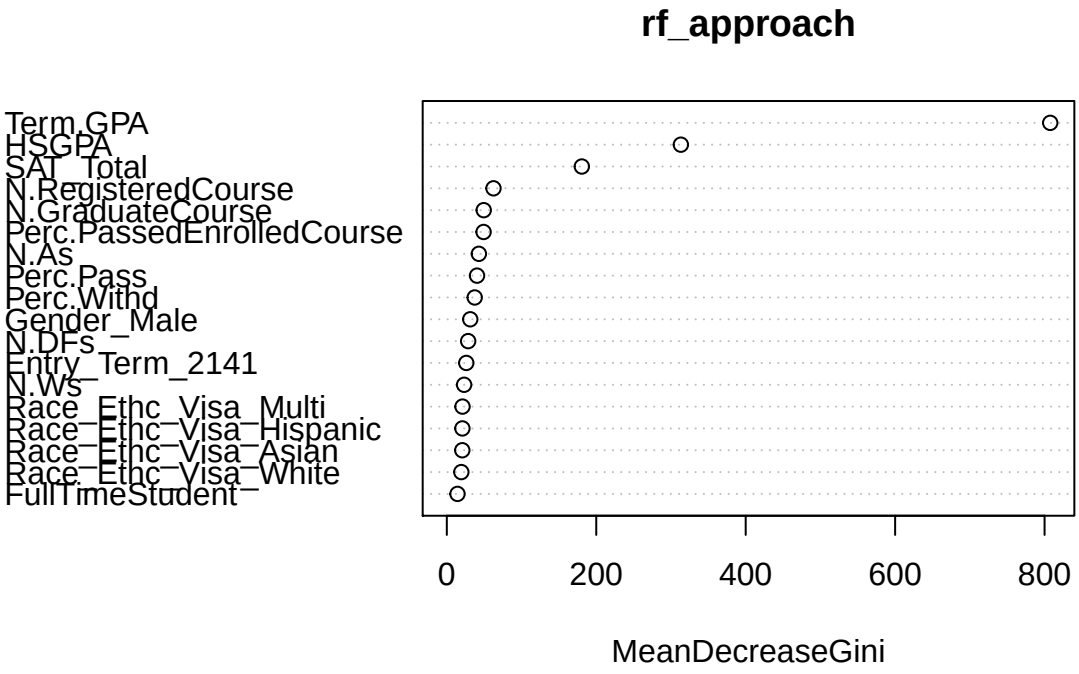
The first three will be SAT_Total, HSGPA and Perc.PassedEnrolledCourse by a function in "Generalized Boosted Models: A guide to the gbm package" page 10 to determine the, and the importance.rel.inf is quite large for the first 2.

- Part b. Perform an importance analysis on the best classification model: which three predictors are most important or effective to explain the response variable? Find the relationship and dependence of these predictors with the response variable. Include graphs and comments.

solution for Section C. Part b

```
## The importance of predictors in the random forests for classification:

##           MeanDecreaseGini
## Term.GPA          807.5606
## HSGPA             313.2227
## SAT_Total         180.6936
```



The first three will be Term.GPA, HSGP, SAT_Total by using Gini index to determine the importance, and the Gini Index is quite large for these three.

- Part c. Write a conclusion paragraph. Evaluate overall what you have achieved. Did the baselines get improved? Why do you think the best model worked well or the models didn't work well? How did you handle issues? What could be done more to get `better` and `interpretable` results? Explain with technical terms.

solution for Section C. Part c

The baseline for regression get improved by using boosting method, but the test MSE does not change that much while it takes so much long time, so I guess it's not worth it. Moreover, it does not have much improvement when using ANN. One of the reasons that my ANN does not work well might be because I didn't try much hyperparameters (because it takes so long to run one ANN). The baselines for classification

get improved by using random forest and also has some problem with ANN. Another reason I think may cause ANN method fails is the scale method I use, maybe I should use the math function in the useful link that professor provided us in the tips. By preprocessing the data, I think there are no big issues in these models(although it still has some issue such like the variables are not correlated to eachother), but I still don't know how to tune parameters, the way I use to find the hyperparameters is try some of them, but I don't think it is a efficient and right way to do it.

I hereby write and submit my solutions without violating the academic honesty and integrity. If not, I accept the consequences.

Write your pair you worked at the top of the page. If no pair, it is ok. List other fiends you worked with (name, last name):

Disclose the resources or persons if you get any help: Chenglu Xia

How long did the assignment solutions take?: 10hrs

References

...