Midterm-1 Project Portion - Version 1

First and last name: Jingwen Zhong // Pair's first and last name: XiaCheng Lu

Submission Date: 03/10/2021

Midterm-1 Project Instruction

Midterm-1 has test and project portions. This is the project portion. Based on what we covered on the modules 1, 2 and 3, you will reflect statistical methods by analyzing data and building predictive models using train and test data sets. The data sets are about college students and their academic performances and retention status, which include categorical and numerical variables.

Throughout the data analysis, we will consider only two response variables, 1) current GPA of students, a numerical response variable, call it **y1=Term.GPA** and 2) Persistence of student for following year, a binary response variable (0: not persistent on the next term, 1:persistent on the next term), call it **y2=Persistence.NextYear**.

Briefly, you will fit regression models on $y1$ and classification models on $y2$ using the subset of predictors in the data set. Don't use all predictors in any model.

A. Touch and Feel the Data - 5 pts

- Import Data Set and Set Up:

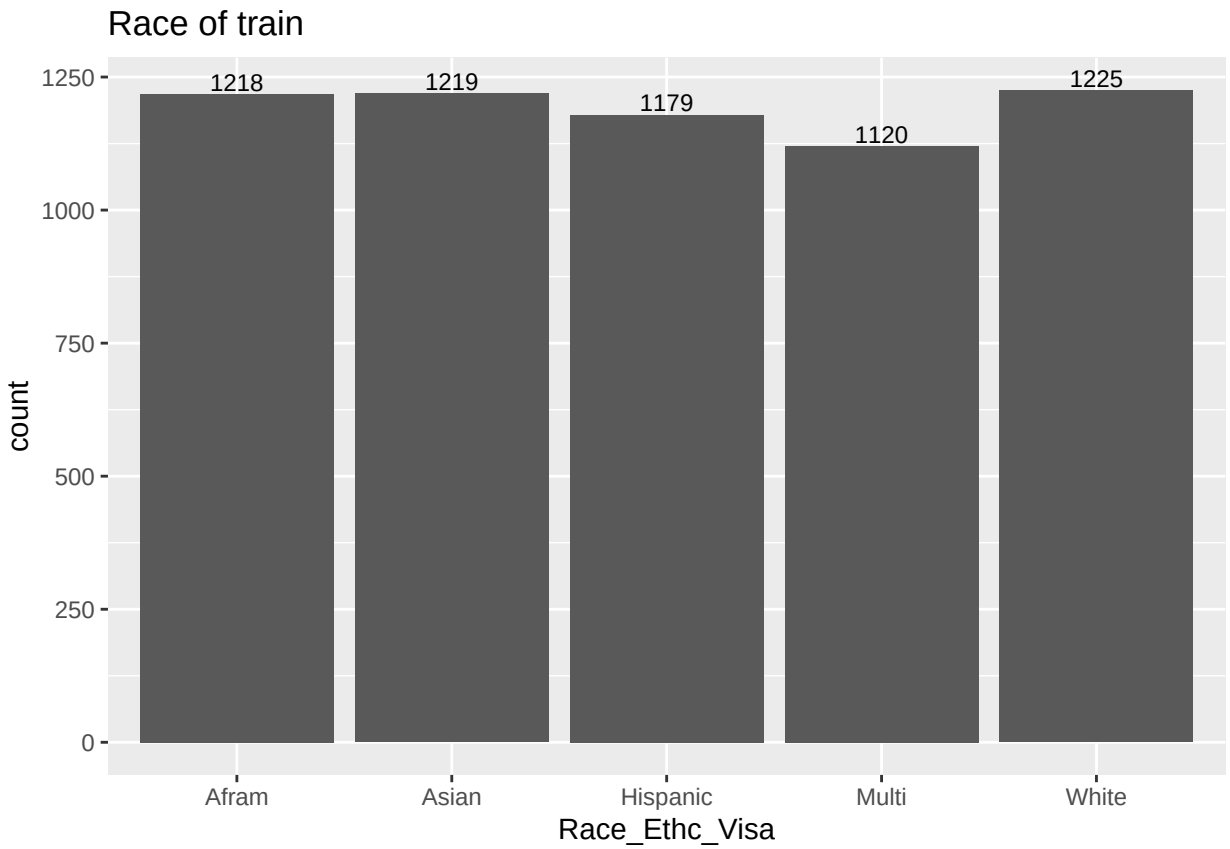
Open the data set **StudentDataTrain.csv**. Be familiar with the data and variables. Start exploring it. Practice the code at the bottom and do the set-up.

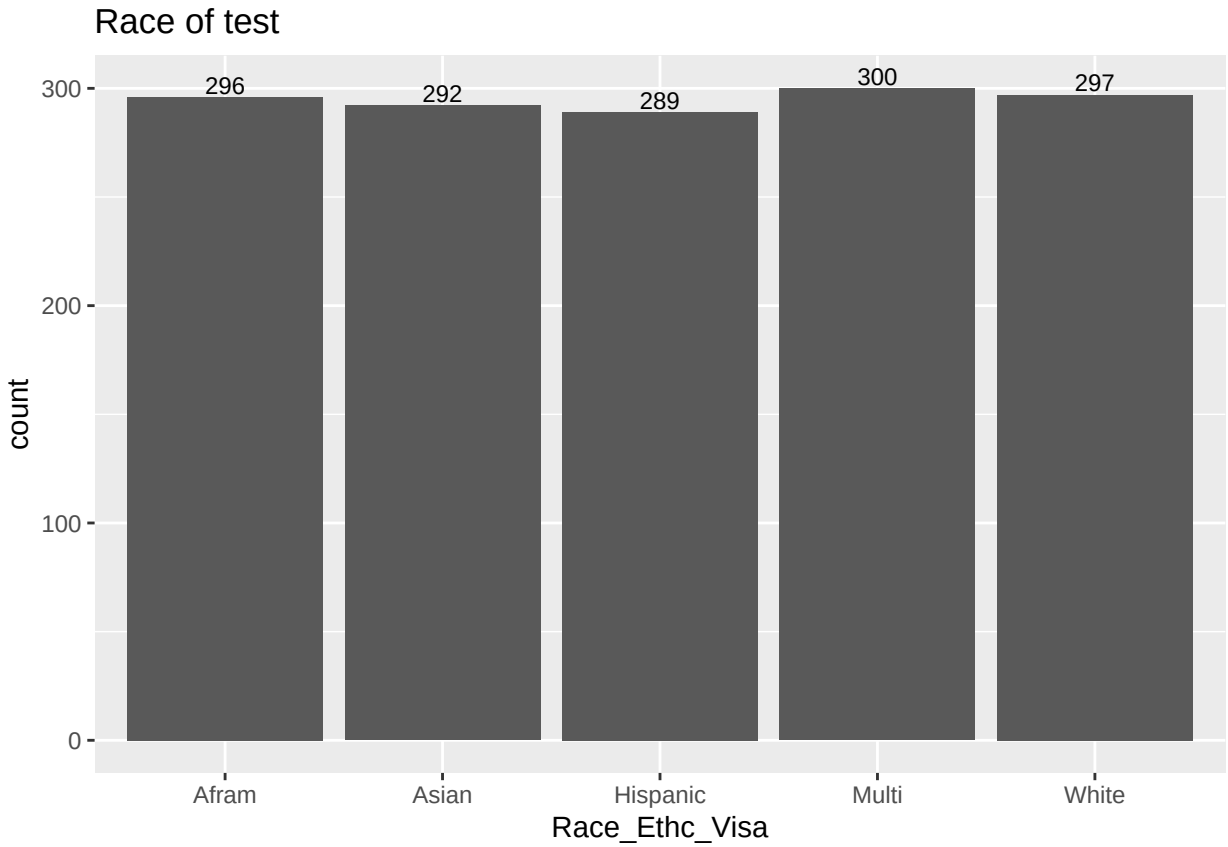
- Do Exploratory Data Analysis:

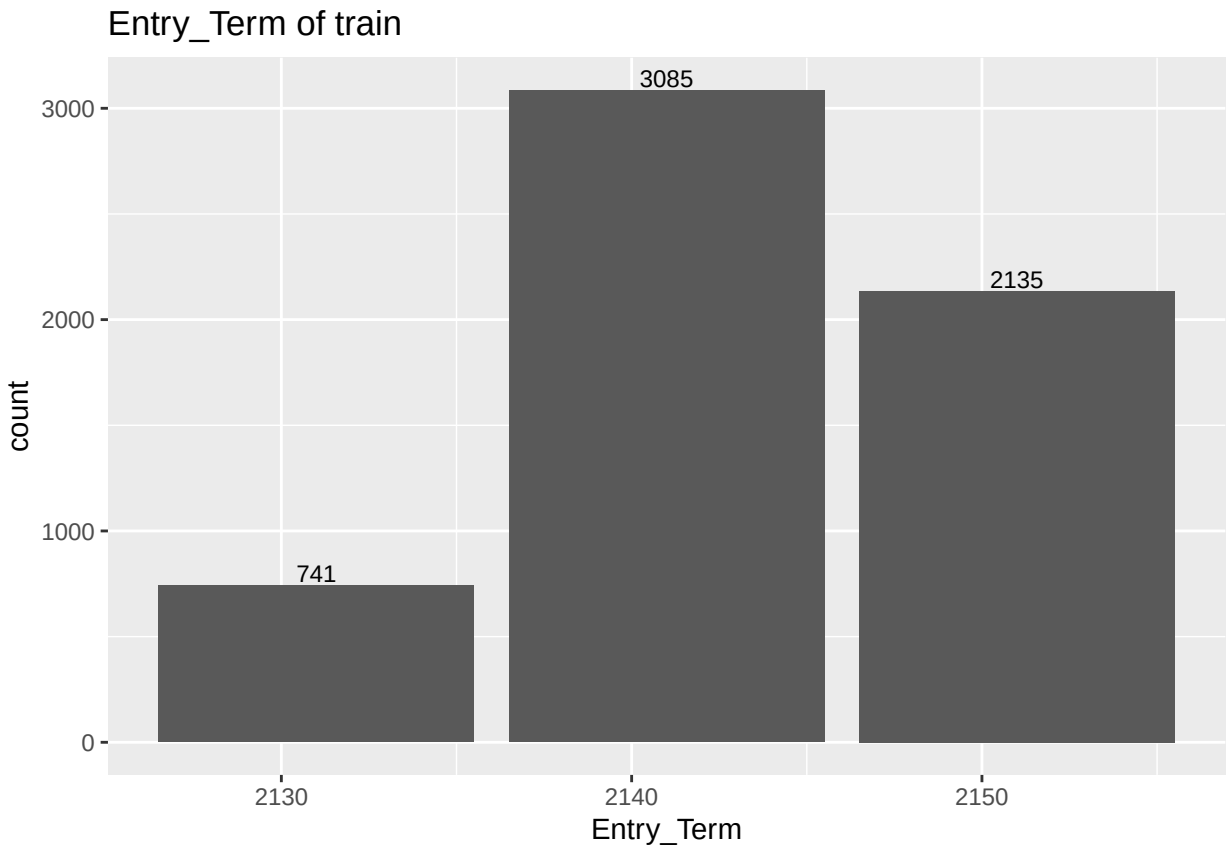
Start with Exploratory Data Analysis (EDA) before running models. Visually or aggregatedly you can include the description and summary of the variables (univariate, and some bivariate analyses). If you keep this part very simple, it is ok.

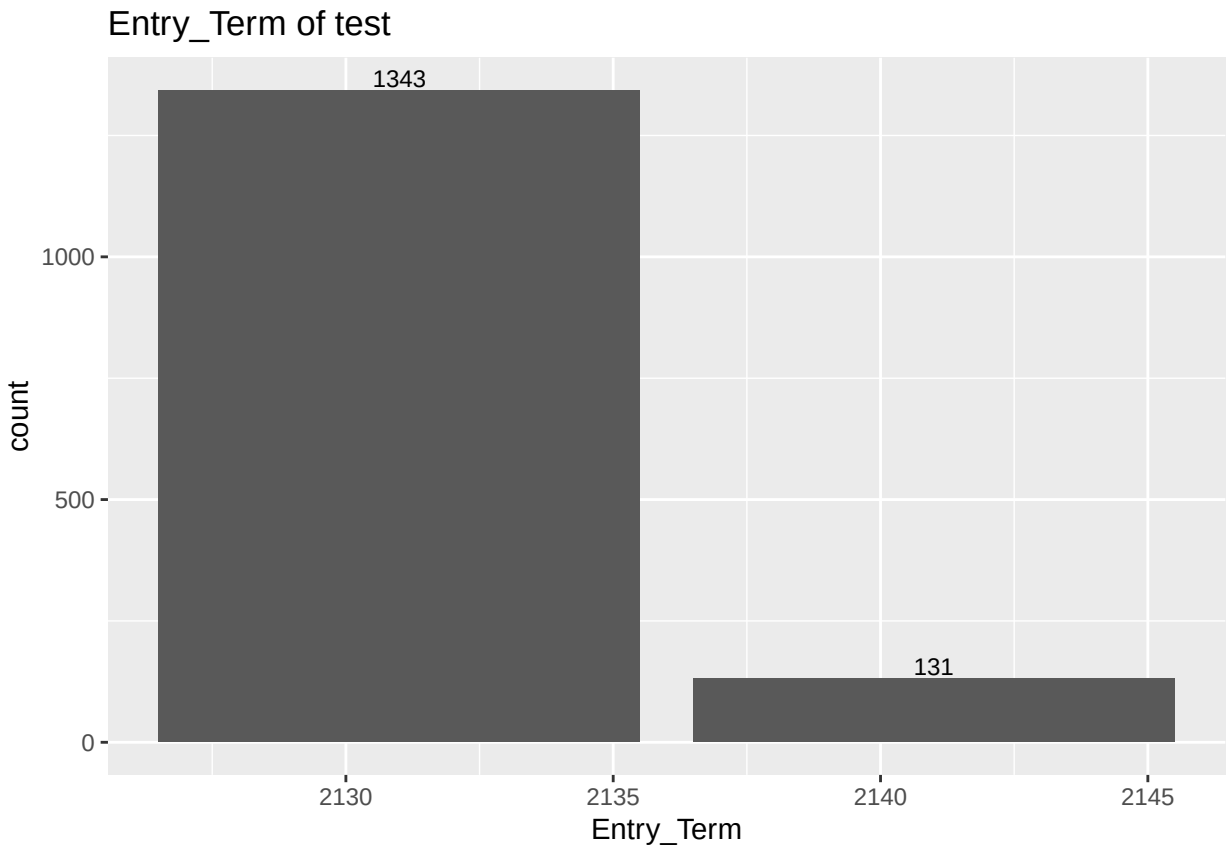
Section A solution

```
## [1] "C:/Users/Jingwen/Desktop"
```


Race of train
1250
1000
750
500
250
0
count
1218
1219
1179
1120
1225
Afram
Asian
Hispanic
Multi
White
Race_Ethc_Visa


Race of test
300
200
100
0
count
296
292
289
300
297
Afram
Asian
Hispanic
Multi
White
Race_Ethc_Visa

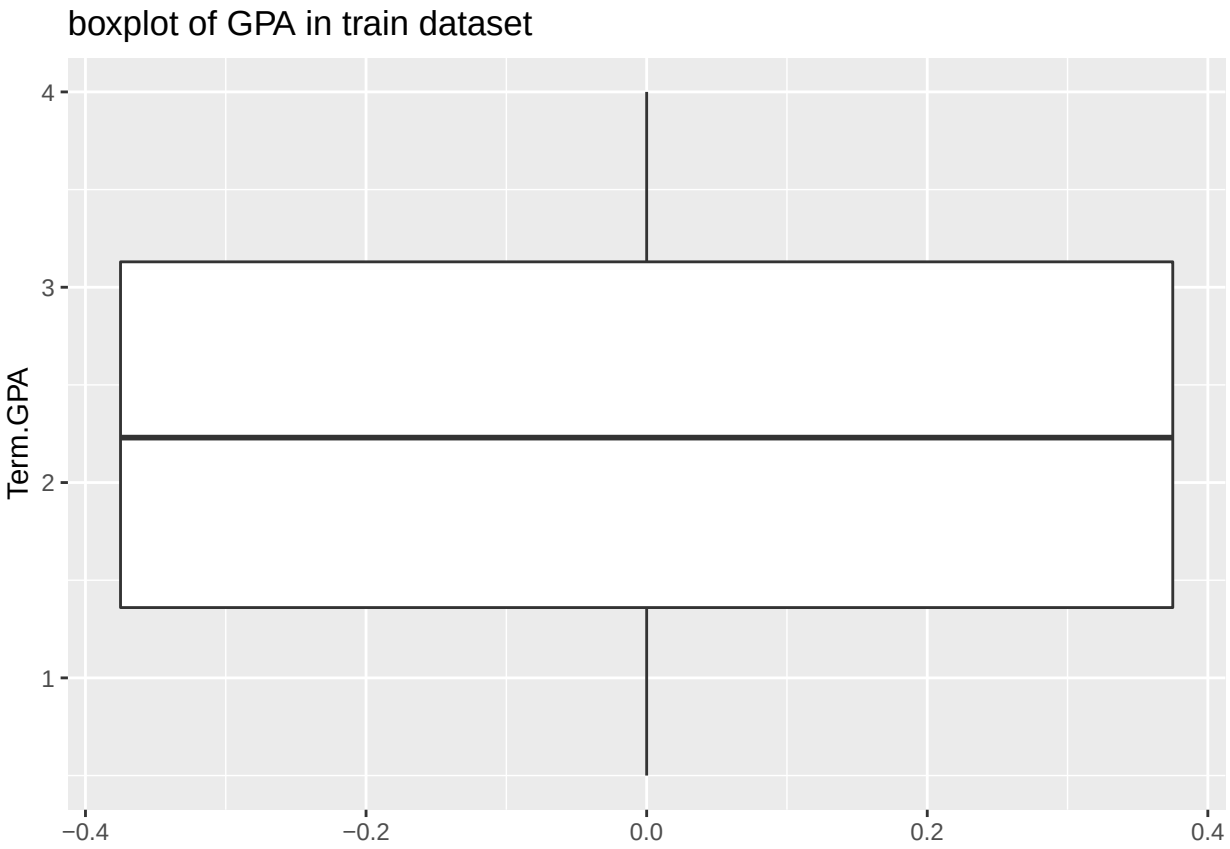

Entry_Term of train
3085
2135
741
3000
2000
1000
0
count
2130
2140
2150
Entry_Term

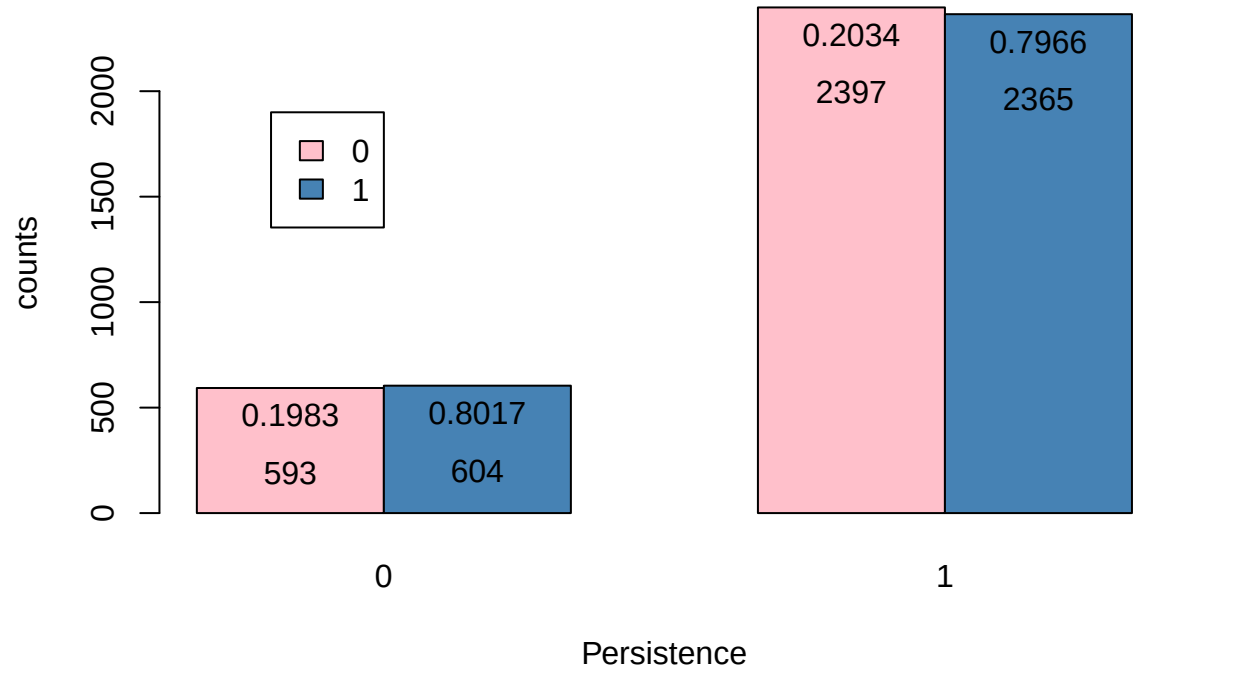


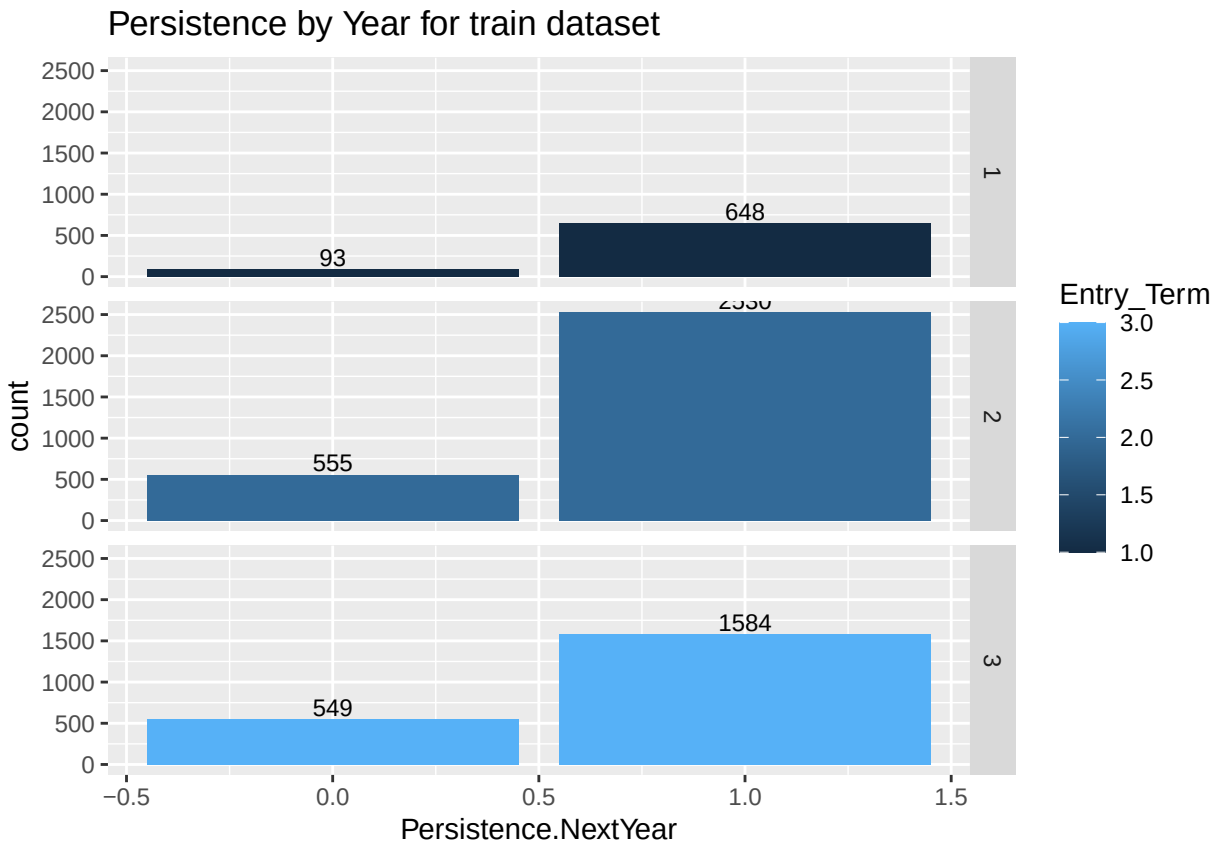

```
## Belew are  some graphs of (response variables mostly) training set
```

Table 1: Student Count with Persistence and Gender for training dataset

	0	1	Sum
0	593	604	1197
1	2397	2365	4762
Sum	2990	2969	5959


boxplot of GPA in train dataset
Term.GPA
4
3
2
1
−0.4
−0.2
0.0
0.2
0.4

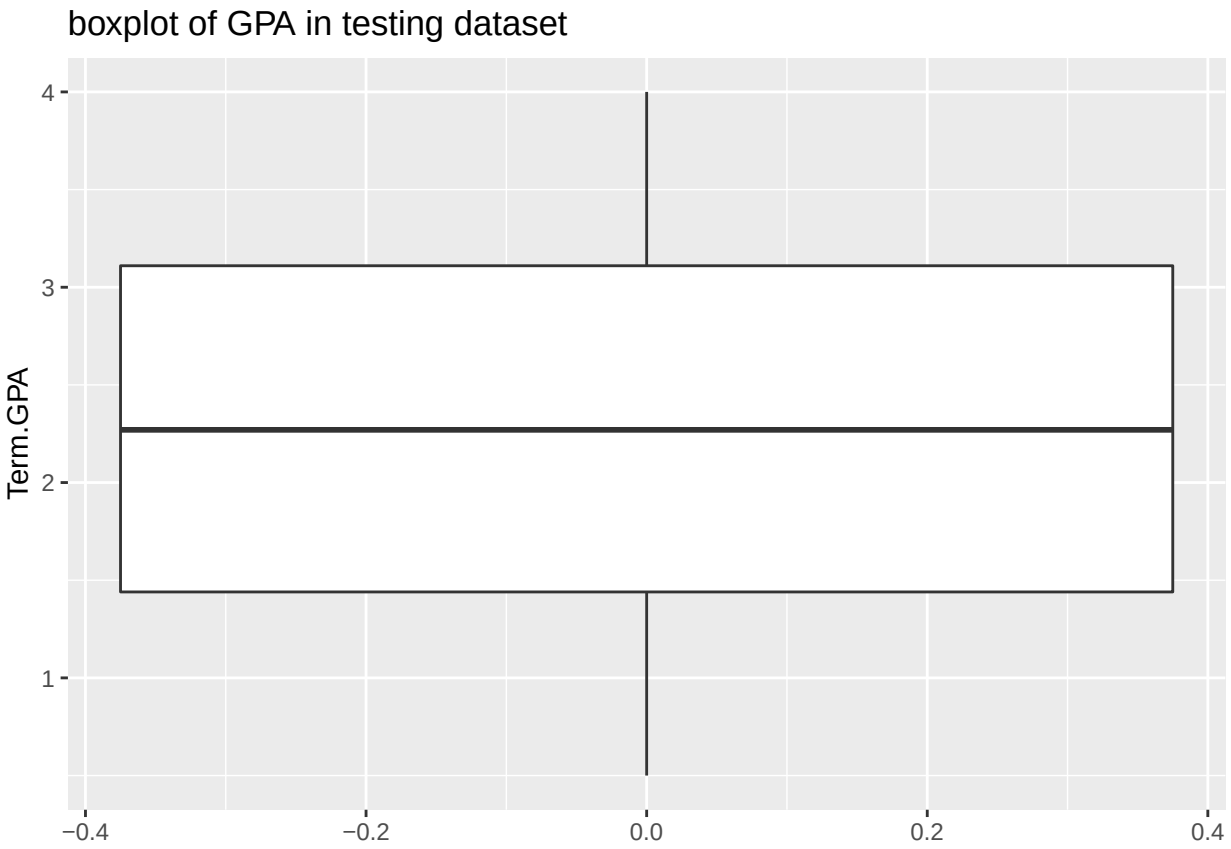

Student Count and percentage with Persistence across Gender for trai
counts
0
500
1000
1500
2000
0
1
0.1983
593
0.8017
604
0.2034
2397
0.7966
2365
0
1
Persistence

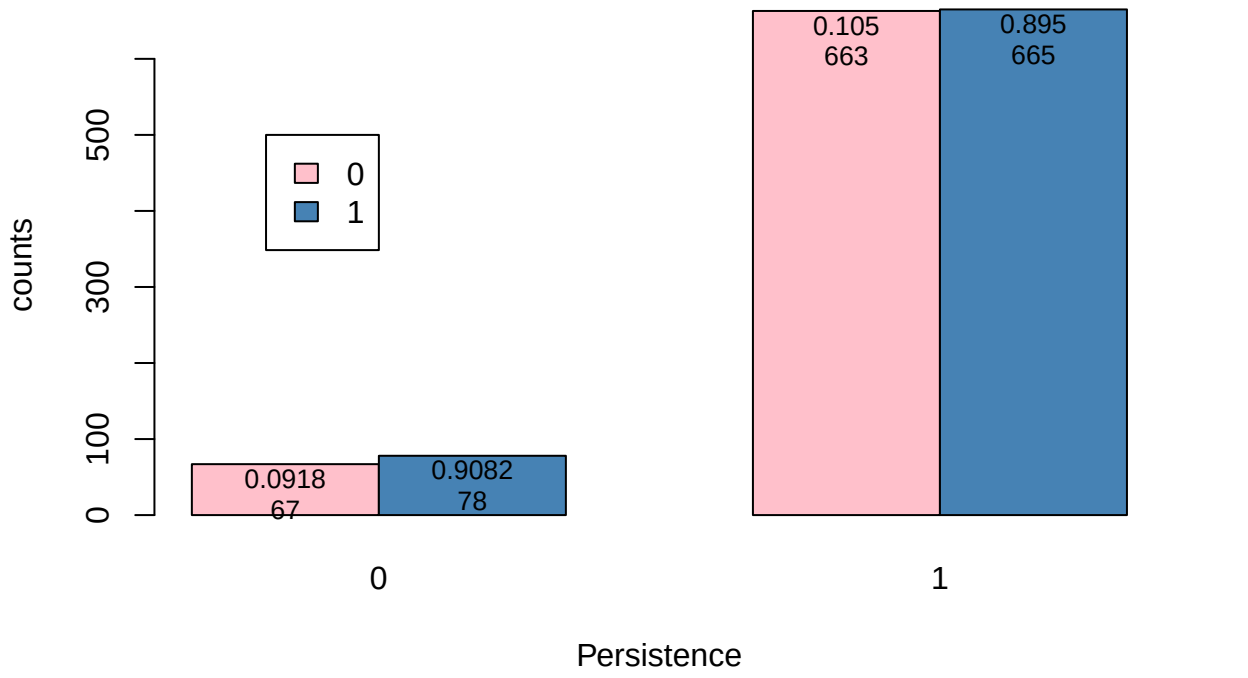


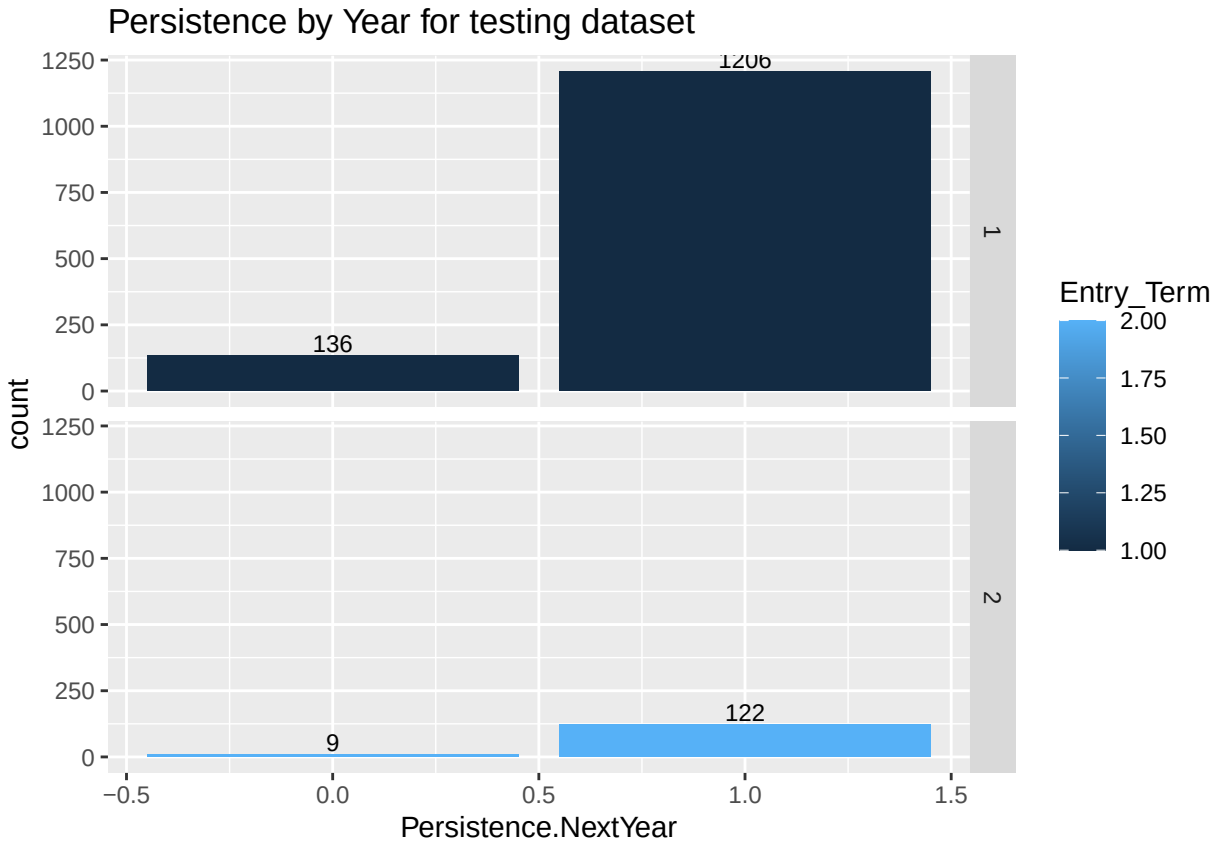

```
## Belew are  some graphs of (response variables mostly) testing set
```

Table 2: Student Count with Persistence and Gender for testing

	0	1	Sum
0	67	78	145
1	663	665	1328
Sum	730	743	1473


boxplot of GPA in testing dataset
Term.GPA
4
3
2
1
−0.4
−0.2
0.0
0.2
0.4


Student Count and percentage with Persistence across Gender for tes
counts
0
100
300
500
0
1
0.0918
67
0.9082
78
0.105
663
0.895
665
0
1
Persistence



Note:

$$Perc.PassedEnrolledCourse = \frac{N.PassedCourse}{N.RegisteredCourse}$$

$$Perc.Pass = \frac{N.PassedCourse}{N.CourseTaken}$$

$$Perc.Withd = \frac{N.Ws}{N.RegisteredCourse}$$

So later we don't need to use N.PassedCourse and N.CourseTaken

I implement all missing value of Perc.Pass to 0, since in that situation, the couseTaken of the students is 0, the passed coursed and the percentage should also be 0.

I implement missing value of HSGPA and SAT_Total to their median value, and delete the rows that missing value of Gender.

I also transfer Gender to a dummy variable Gender(male=1 and female=0)

I create 4 new dummy variables Afram(True=1, False=0), Asian(True=1, False=0), Hispanic(True=1, False=0), Multi(True=1, False=0), (if race before are all 0, then this student will be White), which are numerical value and transferred from Race_Ethc_Visa.

I also transfer Entry term(2131,2141,2151) to factor, then changed it to (2131=1,2141=2,2151=3), since I think as term is a continuous categorical predictor.

B. Build Regression Models - 20 pts - each model 5 pts

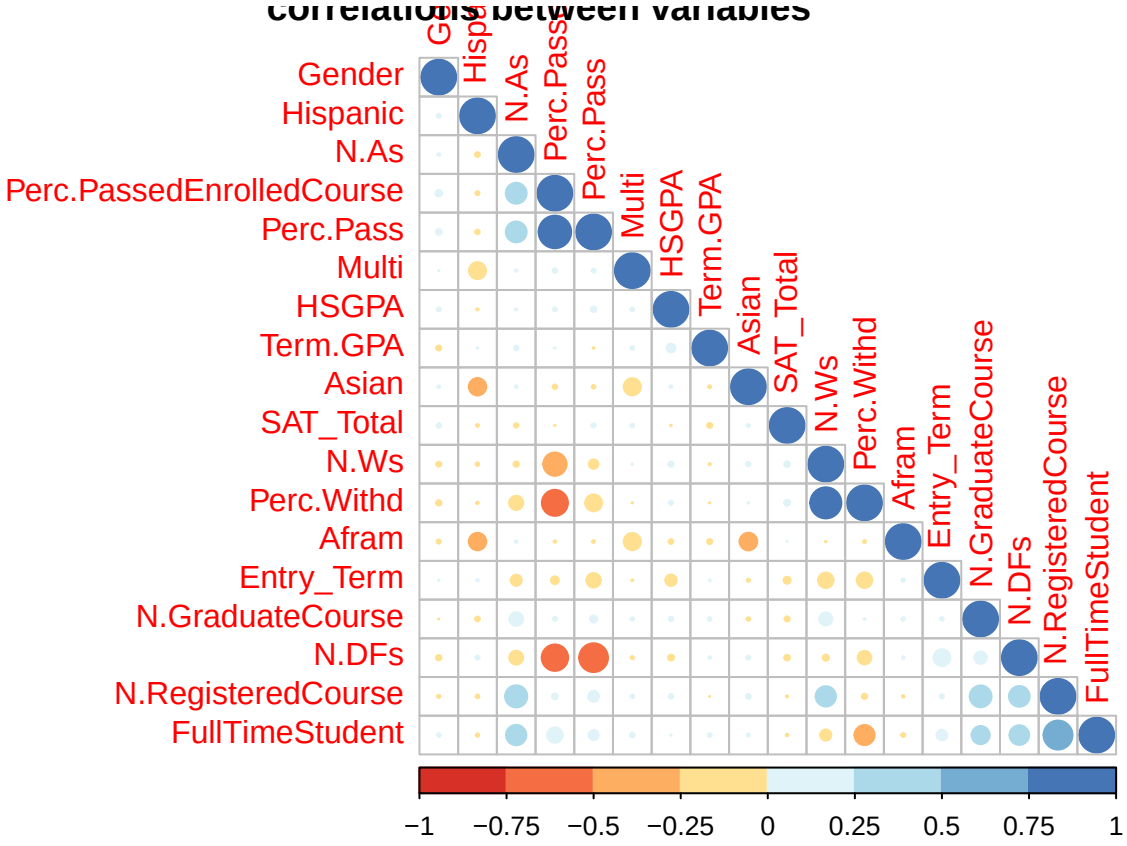
Build linear regressions as listed below the specific four models to predict $y1$ with a small set of useful predictors. Please fit all these by justifying why you do (I expect grounding justifications and technical terms used), report the performance indicators in a comparative table, MSE_{train}, MSE_{test}, $R^2_{adj,train}$ and $R^2_{adj,test}$ using train and test data sets. The regression models you will fit:

1. Best OLS SLR
2. Best OLS MLR using any best small subset of predictors (using any selection methods)
3. Best MLR Ridge with any best small subset of predictors
4. Best MLR Lasso with any best small subset of predictors

For tuning parameter, justify with statistical methods/computations why you choose.

Section B solutions

```
##  [1] "Gender"                  "HSGPA"
##  [3] "SAT_Total"               "Entry_Term"
##  [5] "Term.GPA"                "N.RegisteredCourse"
##  [7] "N.Ws"                    "N.DFs"
##  [9] "N.As"                    "Perc.PassedEnrolledCourse"
## [11] "Perc.Pass"               "Perc.Withd"
## [13] "N.GraduateCourse"        "FullTimeStudent"
## [15] "Afram"                   "Asian"
## [17] "Hispanic"                "Multi"
```



Model 1.

```
## The best OLS SLR using predictor: HSGPA
```

	Predictor	MSE_train	MSE_test	adj.r.squared_train	adj.r.squared_test
The best SLR	HSGPA	1.0282	0.9862	0.0038	0.0021

Model 2.

```
## The best OLS MLR uses: 1 predictor(s)

## Predictor(s) and coefficient(s):

## (Intercept)       HSGPA
##  1.84944247  0.00512087
```

Table 4: The best OLS MLR

	Number of Predictor	MSE_train	MSE_test	adj.r.squared_train	adj.r.squared_test
The best OLS MLR	1	1.0282	0.9862	0.0038	0.0021

####Model 3.

Table 5: Full model fitted by ridge regression

	Best_lambda	MSE_train	MSE_test	adj.r.squared_train	adj.r.squared_test
ridge regression model	0.6097	1.0269	0.989	0.0024	-0.0117

Model 4.

Table 6: Predictor(s) and coefficient(s)

Intercept	HSGPA
2.003933	0.0031011

Table 7: Full model fitted by lasso regression

	Best_lambda	MSE_train	MSE_test	adj.r.squared_train	adj.r.squared_test
lasso regression model	0.0252	1.0289	0.9865	0.0032	0.0018

Tuning parmeter: best_lamda, I use cross validation to choose best_lamda

C. Build Classification Models - 20 pts - each model 5pts

Build four classification models as below. Please fit all these, include performance indicators for train and test data sets, separately. Include confusion matrix for each. For each `train` and `test` data set, report: `accuracy`, `recall`, `precision`, and `f1` in a cooperative table. For LR or LDA, include ROC curve, area and interpretation. The classification models you will fit:

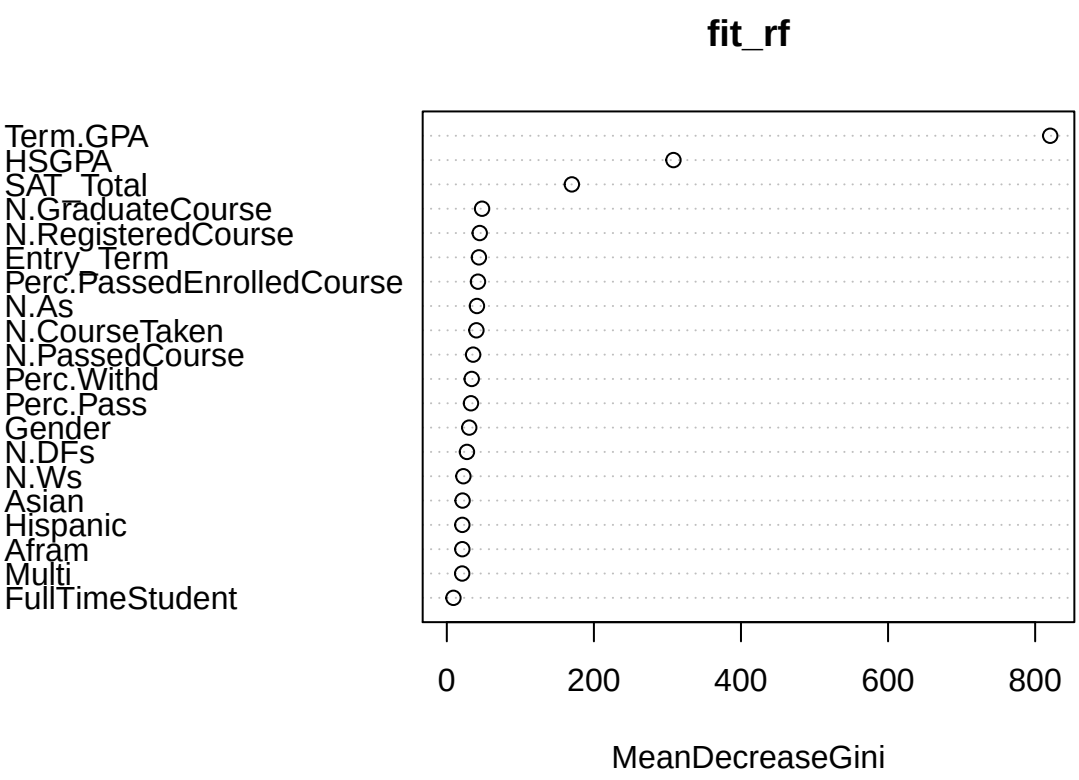
1. Logistic Regression (LR) with any best small subset of predictors
2. KNN Classification with any best small subset of predictors
3. Linear Discriminant Analysis (LDA) with any best small subset of predictors
4. Quadratic Discriminant Analysis (QDA) with any best small subset of predictors

Justify why you choose specific K in KNN with a grid search or CV methods.

Section C solutions

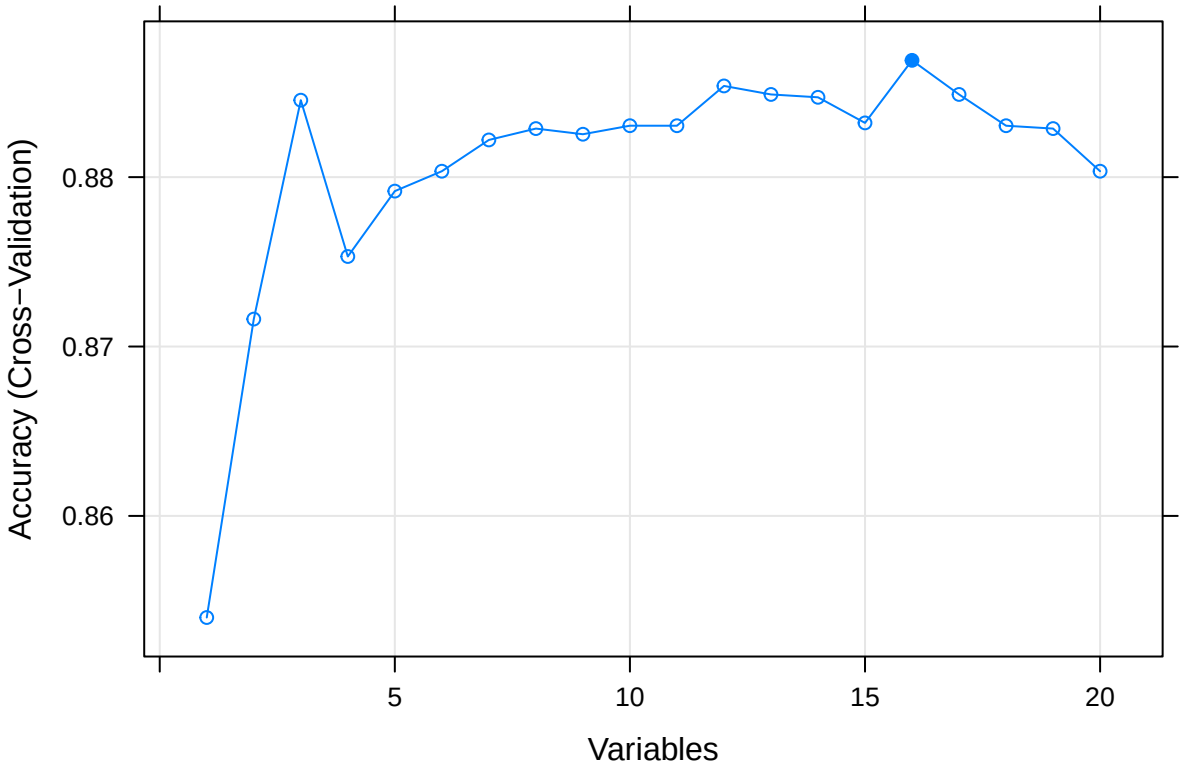
```
##  [1] "Gender"                    "HSGPA"
##  [3] "SAT_Total"                 "Entry_Term"
##  [5] "Term.GPA"                  "N.RegisteredCourse"
##  [7] "N.Ws"                      "N.DFs"
##  [9] "N.As"                      "Perc.PassedEnrolledCourse"
## [11] "Perc.Pass"                 "Perc.Withd"
## [13] "N.GraduateCourse"          "FullTimeStudent"
## [15] "Afram"                     "Asian"
## [17] "Hispanic"                  "Multi"
```

```
## The importance of predictors:
```




```
## How many numbers of feature should we select:

## x is Number of variables
```



Model 1.

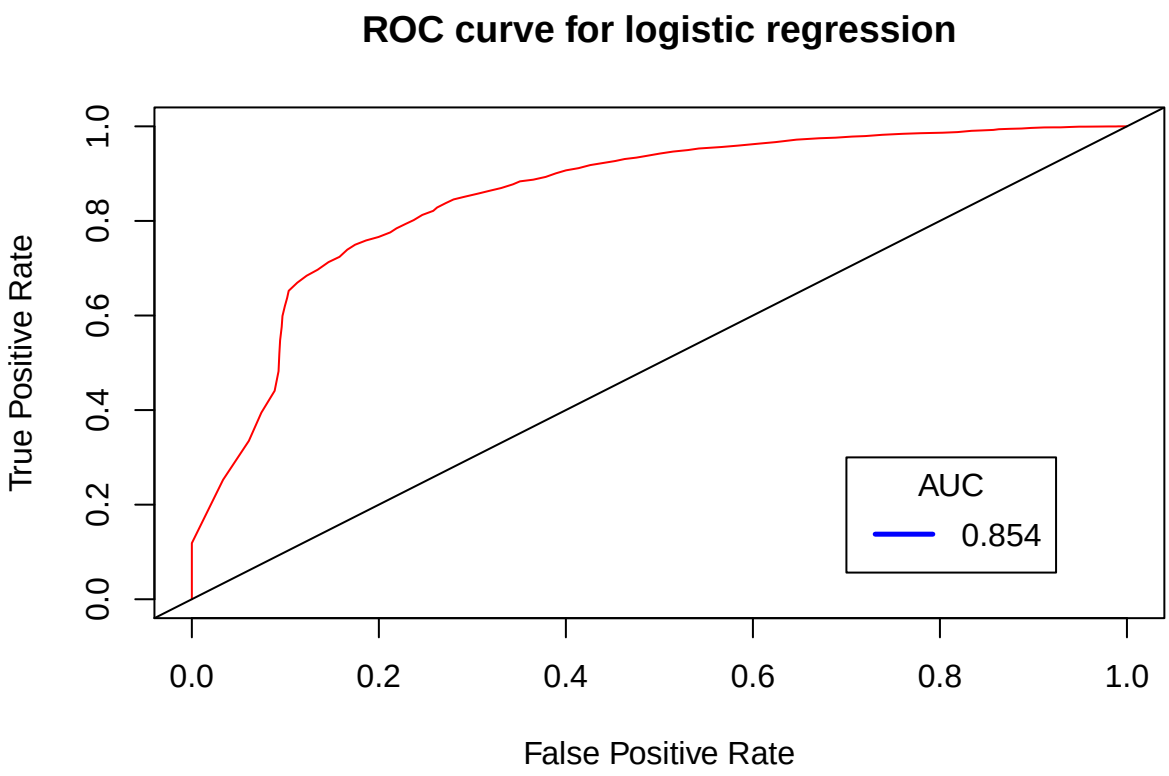
Table 8: Training set Report of logistic regression

Metrics	Values
Accuracy	85.4
Recall	94.67
Precision	87.98
F1	91.2

Table 9: Testing set Report of logistic regression

Metrics	Values
Accuracy	91.45
Recall	93.98
Precision	96.45
F1	95.19

```
## plot:
```



An ideal ROC curve will hug the top left corner, and the larger area under the ROC curve the better the classifier, above graph tells that this logistic regression model works ok.the ROC curve for the logistic regression model fit to these data is virtually indistinguishable from ROC for the LDA model.

Model 2.

```
## the best k is: 35
```

Table 10: Training set Report of KNN

Metrics	Values
Accuracy	80.03
Recall	100
Precision	80.01
F1	88.89

Table 11: Testing set Report of KNN

Metrics	Values
Accuracy	90.22
Recall	100

Metrics	Values
Precision	90.22
F1	94.86

I choose the best k with gird search, and I choose the one with the lowest test error and also the one with lowest train error while I keep test error the lowest.

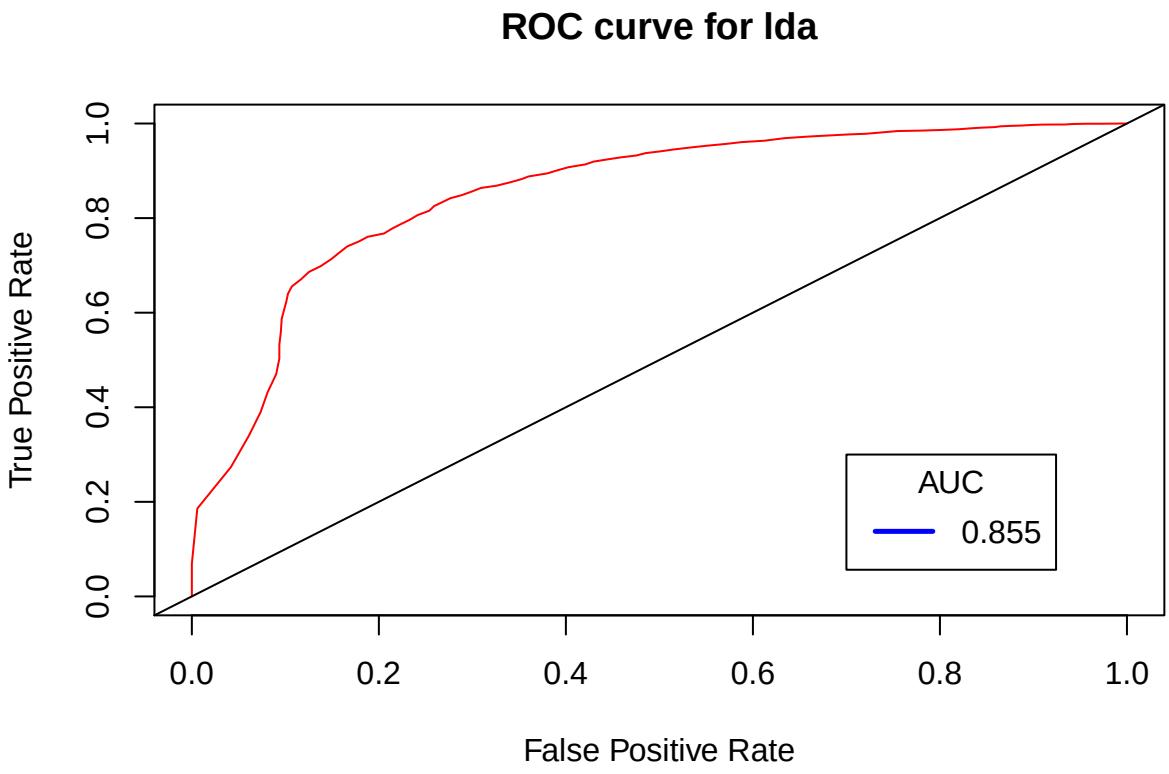
Model 3.

Table 12: Training set Report of LDA

Metrics	Values
Accuracy	85.27
Recall	94.5
Precision	87.96
F1	91.11

Table 13: Testing set Report of LDA

Metrics	Values
Accuracy	91.04
Recall	93.67
Precision	96.28
F1	94.96



An ideal ROC curve will hug the top left corner, and the larger area under the ROC curve the better the classifier. The overall performance of a classifier, giving by AUC in above graph tells that this LDA model works well since its over 0.5.

Model 4.

Table 14: Training set Report of LDA

Metrics	Values
Accuracy	84.41
Recall	92.92
Precision	88.2
F1	90.5

Table 15: Testing set Report of LDA

Metrics	Values
Accuracy	88.93
Recall	91.42
Precision	96.12
F1	93.71

D. Overall Evaluations and Conclusion - 5 pts

Briefly, make critiques of the models fitted and write the conclusion (one sentence for each model, one sentence for each problem - regression and classificaton problems we have here). Also, just address one of these: diagnostics, violations, assumptions checks, overall quality evaluations of the models, importance analyses (which predictors are most important or effects of them on response), outlier analyses. You don't need to address all issues. Just show the reflection of our course materials.

Section D solution

diagnostics, violations, assumptions checks, overall quality evaluations of the models

- Regression Problem

The correlations between Term.GPA and other variables are very low, even the highest correlation (HSGPA and Term.GPA) is very small (about 0.06), and the most important response in these 4 models is HSGPA. The results between the 4 models are very similar.

1. Best OLS SLR: I choose HSGPA to be the variable of my single linear regression since with it, the model with testing data has the lowest MSE.
2. Best OLS MLR: Using forward stepwise selection with OLS to select the best subset selection, the best MLR model is actually with only 1 variable(HSGPA) and which makes it the same model as the Best OLS SLR and have the same result.
3. Best MLR Ridge: My ridge regression model has a negative adjust R square, and a smaller value of adjusted R square indicates a model with a large test error and which makes Ridge regression model in this case a worst model. I guess one of the reasons is because ridge can only make the coefficient close to 0, can't eliminate the variable, so ridge regression will include all p predictors in the final model.
4. Best MLR Lasso: Lasso regression performs better than the ridge regression, and all the coefficients are 0 except HSGPA, which means only HSGPA is used in this model, and the adjust R square is positive in this case.

- Classification Problem

For classification, I first conduct an importance analyses on responses with RANDOM FOREST (shows in the graph in the beginning of Part C), and the graph ranks the importance the responses from high to low, and also I use REF method with random forest algothrim and cross validation to choose how many attributes are the best, and it shows 3 attributes are the best. Therefore for all the classification problems, I use Term.GPA, HSGPA, and SAT_Total.

1. Logistic Regression (LR): The result of my logistic regression is the best among these 4 models, it has the highest F1 score and accuracy. The assumptions of LR are all matched.
2. KNN Classification: The performance of KNN is worse than LR, but its acceptable. Hence KNN is a completely non-parametric approach, no assumptions are made about the shape of the decision boundary. Also, KNN works better when it's nonlinear relationship.
3. Linear Discriminant Analysis (LDA): My LDA model has the similar results as LR, tho I think it does not meet all Gaussian assumptions since LR still performans better than LDA.
4. Quadratic Discriminant Analysis (QDA): QDA gives the worst results in this case, I think its because the boundary of the classes is most likely linear (i.e not nonliear). Since LDA performs better than QDA, we can assume that K classes share a common covariance matrix

---

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: ...

How long did the assignment solutions take?: 10 hrs

References

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Retrieved 1st Mar. 2021 from https://dataaspirant.com/feature-selection-techniques-r/