TEACHERS COLLEGE
COLUMBIA UNIVERSITY

Student Retention Analysis

Xi (Ciel) Zhao

Agenda

PURPOSE

DATA WRANGLING

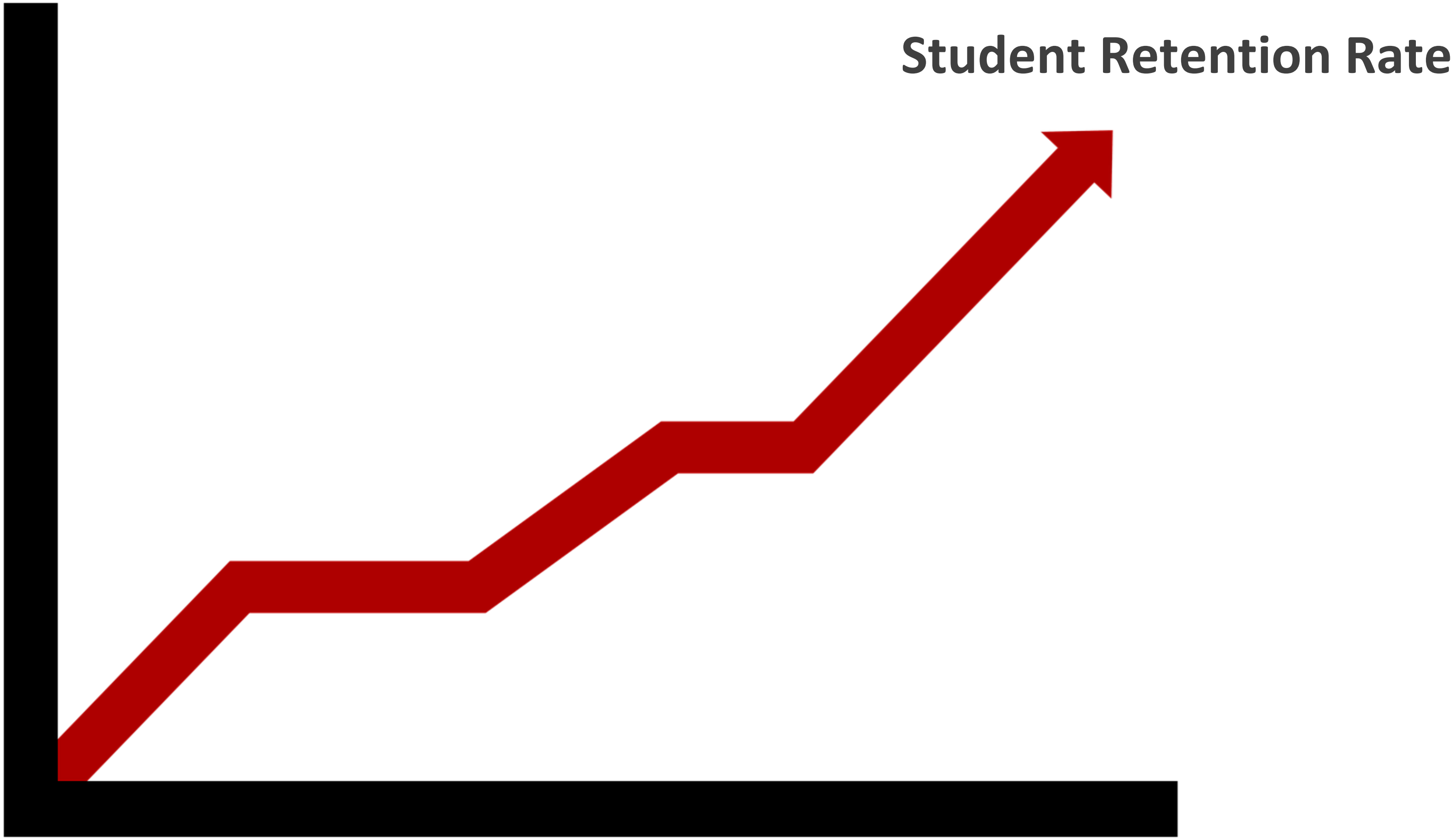
EXPLORATORY DATA ANALYSIS

FEATURE ENGINEERING

MODELLING

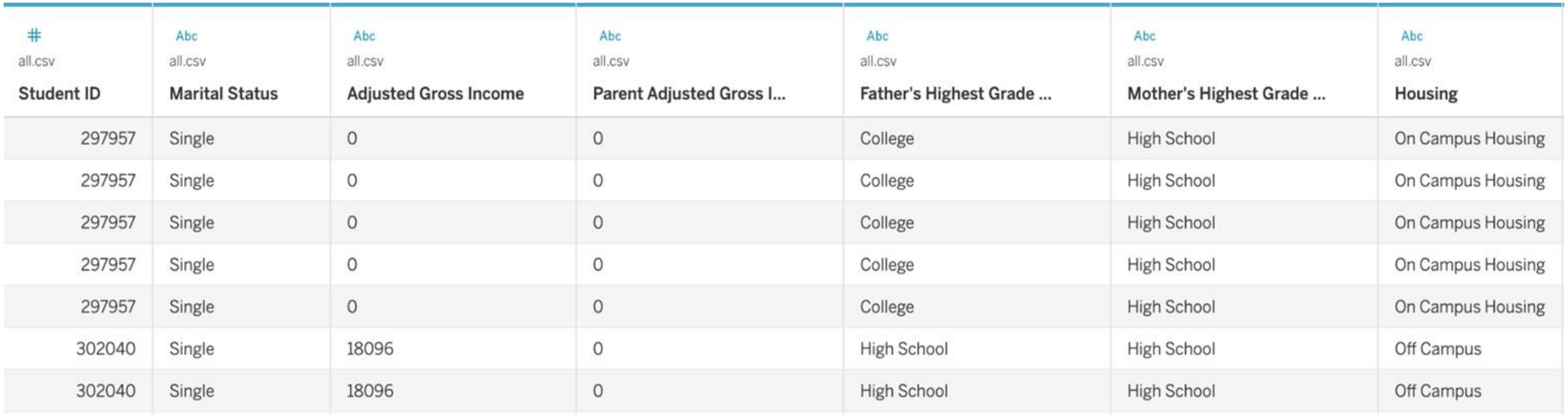
CONCLUSION

Purpose



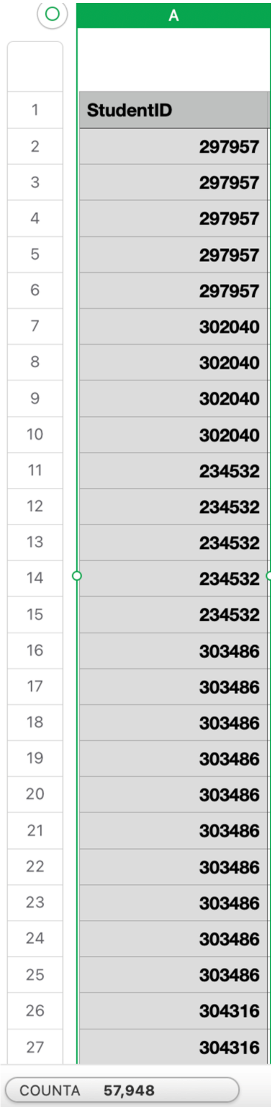
Data Wrangling

Data Merged by Student ID



#	Abc	Abc	Abc	Abc	Abc	Abc
all.csv	all.csv	all.csv	all.csv	all.csv	all.csv	all.csv
Student ID	**Marital Status**	**Adjusted Gross Income**	**Parent Adjusted Gross I...**	**Father's Highest Grade ...**	**Mother's Highest Grade ...**	**Housing**
297957	Single	0	0	College	High School	On Campus Housing
297957	Single	0	0	College	High School	On Campus Housing
297957	Single	0	0	College	High School	On Campus Housing
297957	Single	0	0	College	High School	On Campus Housing
297957	Single	0	0	College	High School	On Campus Housing
302040	Single	18096	0	High School	High School	Off Campus
302040	Single	18096	0	High School	High School	Off Campus

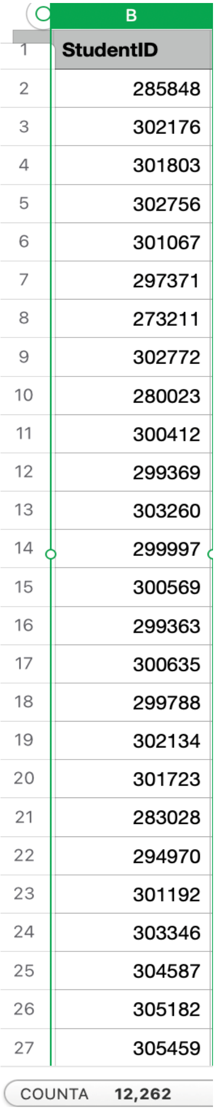
Data Wrangling



	A
1	StudentID
2	297957
3	297957
4	297957
5	297957
6	297957
7	302040
8	302040
9	302040
10	302040
11	234532
12	234532
13	234532
14	234532
15	234532
16	303486
17	303486
18	303486
19	303486
20	303486
21	303486
22	303486
23	303486
24	303486
25	303486
26	304316
27	304316

COUNTA 57,948

→



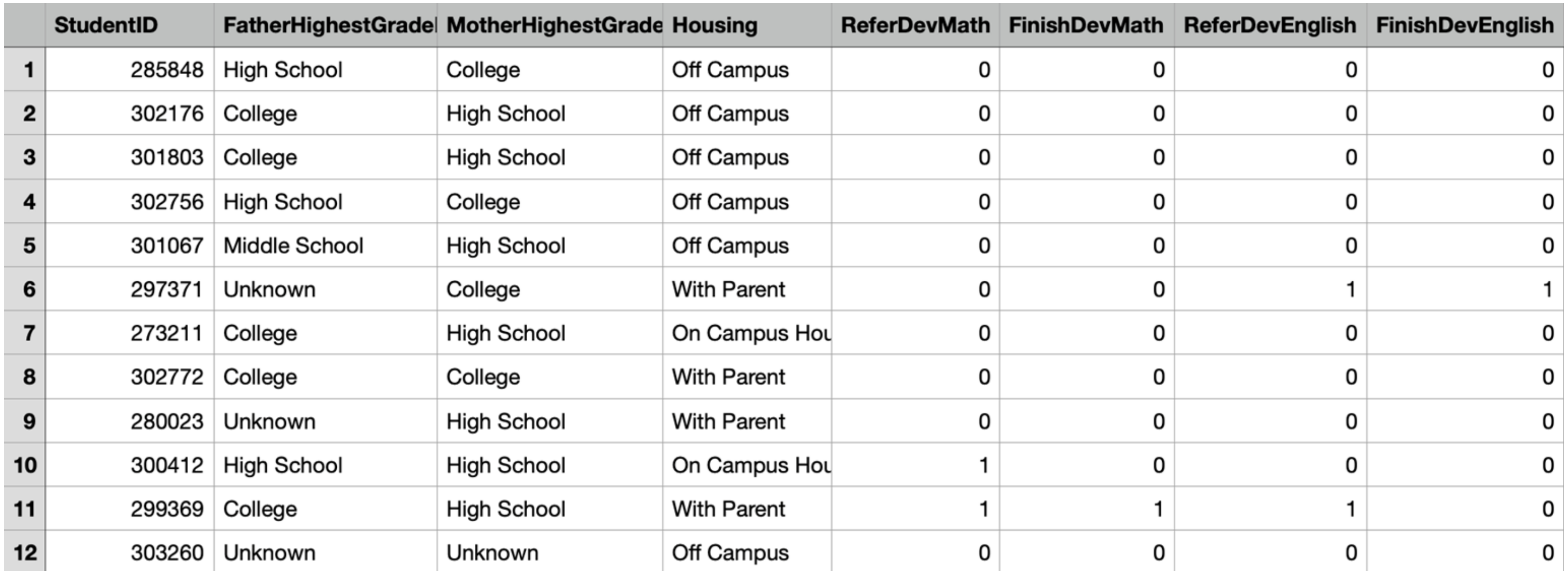
	B
1	StudentID
2	285848
3	302176
4	301803
5	302756
6	301067
7	297371
8	273211
9	302772
10	280023
11	300412
12	299369
13	303260
14	299997
15	300569
16	299363
17	300635
18	299788
19	302134
20	301723
21	283028
22	294970
23	301192
24	303346
25	304587
26	305182
27	305459

COUNTA 12,262

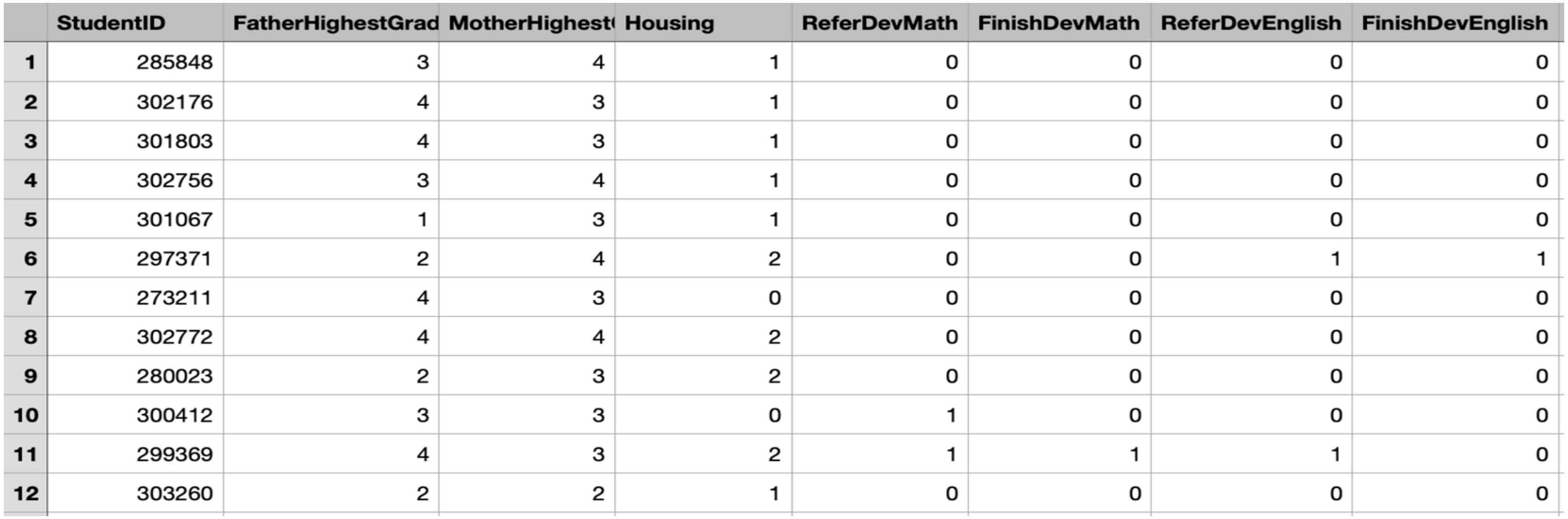
Delete the Duplicate Value

45686 Duplicate Rows

Data Wrangling



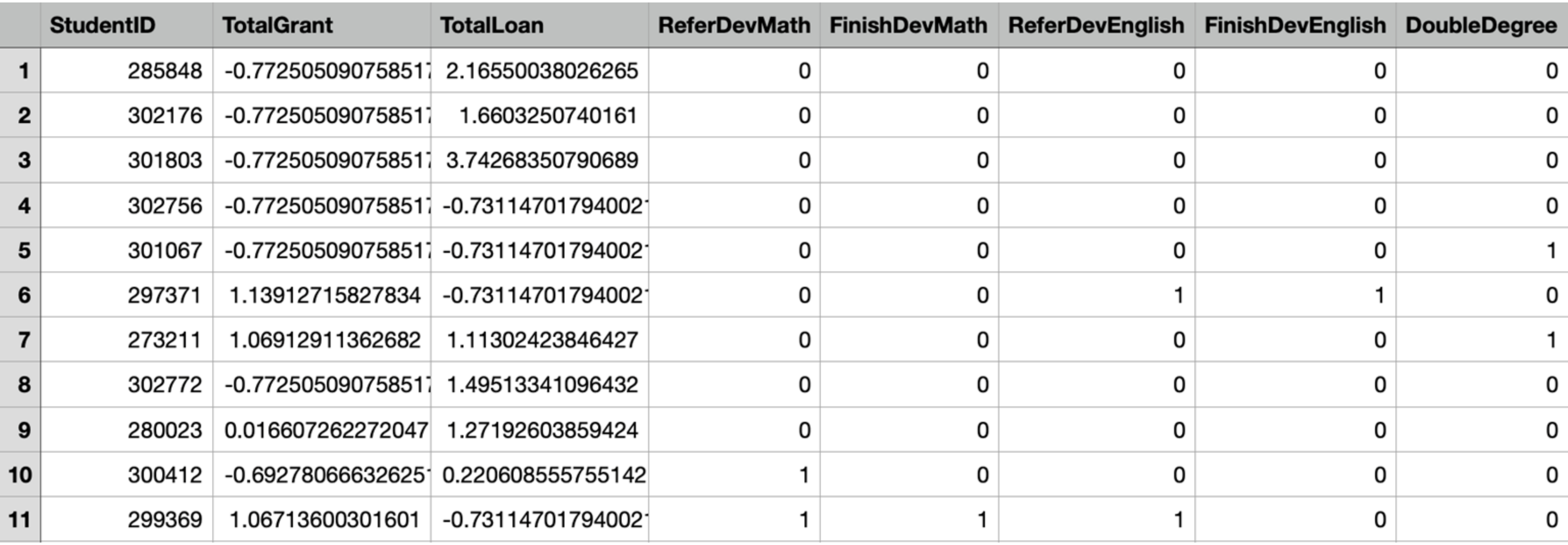
	StudentID	FatherHighestGrade	MotherHighestGrade	Housing	ReferDevMath	FinishDevMath	ReferDevEnglish	FinishDevEnglish
1	285848	High School	College	Off Campus	0	0	0	0
2	302176	College	High School	Off Campus	0	0	0	0
3	301803	College	High School	Off Campus	0	0	0	0
4	302756	High School	College	Off Campus	0	0	0	0
5	301067	Middle School	High School	Off Campus	0	0	0	0
6	297371	Unknown	College	With Parent	0	0	1	1
7	273211	College	High School	On Campus Hou	0	0	0	0
8	302772	College	College	With Parent	0	0	0	0
9	280023	Unknown	High School	With Parent	0	0	0	0
10	300412	High School	High School	On Campus Hou	1	0	0	0
11	299369	College	High School	With Parent	1	1	1	0
12	303260	Unknown	Unknown	Off Campus	0	0	0	0



	StudentID	FatherHighestGrad	MotherHighest	Housing	ReferDevMath	FinishDevMath	ReferDevEnglish	FinishDevEnglish
1	285848	3	4	1	0	0	0	0
2	302176	4	3	1	0	0	0	0
3	301803	4	3	1	0	0	0	0
4	302756	3	4	1	0	0	0	0
5	301067	1	3	1	0	0	0	0
6	297371	2	4	2	0	0	1	1
7	273211	4	3	0	0	0	0	0
8	302772	4	4	2	0	0	0	0
9	280023	2	3	2	0	0	0	0
10	300412	3	3	0	1	0	0	0
11	299369	4	3	2	1	1	1	0
12	303260	2	2	1	0	0	0	0

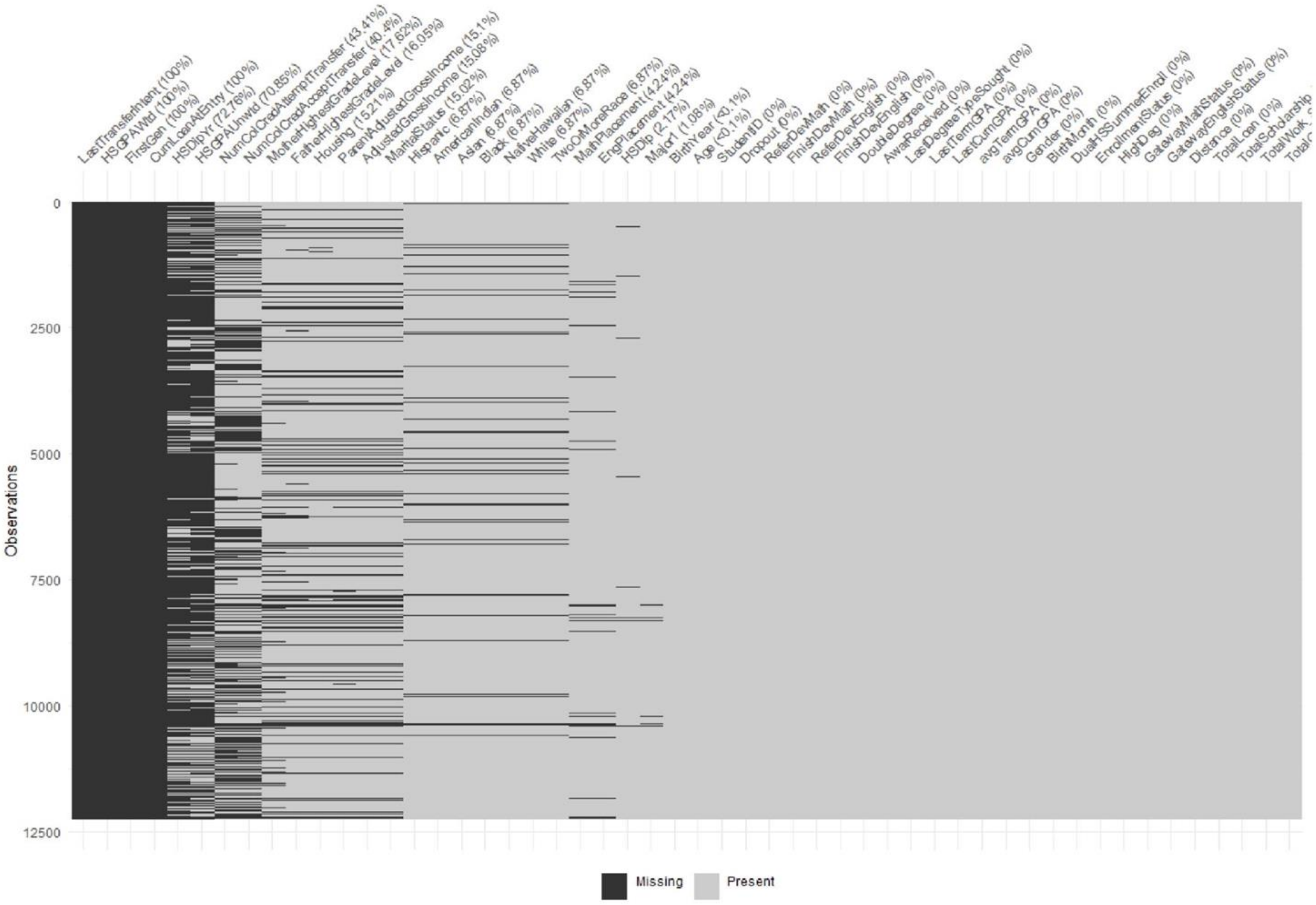
Data Wrangling

New Variables added into Dataset



	StudentID	TotalGrant	TotalLoan	ReferDevMath	FinishDevMath	ReferDevEnglish	FinishDevEnglish	DoubleDegree
1	285848	-0.772505090758517	2.16550038026265	0	0	0	0	0
2	302176	-0.772505090758517	1.6603250740161	0	0	0	0	0
3	301803	-0.772505090758517	3.74268350790689	0	0	0	0	0
4	302756	-0.772505090758517	-0.731147017940021	0	0	0	0	0
5	301067	-0.772505090758517	-0.731147017940021	0	0	0	0	1
6	297371	1.13912715827834	-0.731147017940021	0	0	1	1	0
7	273211	1.06912911362682	1.11302423846427	0	0	0	0	1
8	302772	-0.772505090758517	1.49513341096432	0	0	0	0	0
9	280023	0.016607262272047	1.27192603859424	0	0	0	0	0
10	300412	-0.692780666326251	0.220608555755142	1	0	0	0	0
11	299369	1.06713600301601	-0.731147017940021	1	1	1	0	0

Handling variables with missing value over 30%



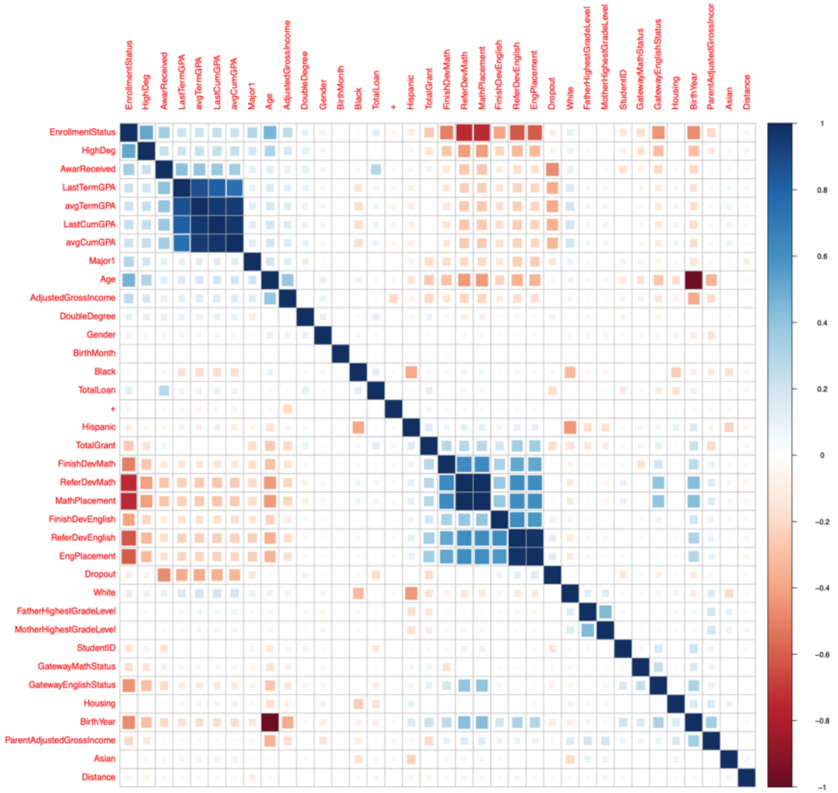
Exploratory Data Analysis:

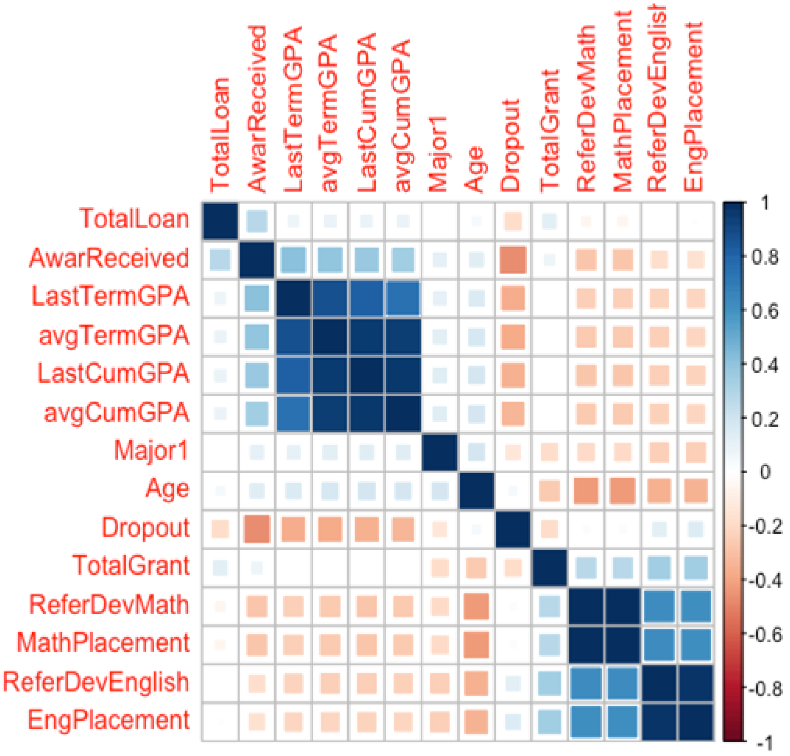
1. **Goal of EDA**
    a. Identifying relationships between variables that are particularly interesting or unexpected
    b. Checking distributional assumptions
    c. Checking for outliers.
    d. Suggesting possible modeling strategy by understanding variable properties :
        i. central trends (mean)
        ii. spread (variance)
        iii. skew outliers

1. **Process of performing EDA**
    a. Propose question
        i. What distribution does my data follow?
        ii. Are there any outliers?
    b. Find correlation
    c. Compare correlation
    d. Find highest top variable
    e. Using summary statistics and relevant plot to maximize insight of dataset, detect outliers and anomalies
        i. Summary statics
        ii. Three type relevant plot
            1. correlation matrix
            2. box plot
            3. histogram

Correlation Matrix

Threshold =0.1

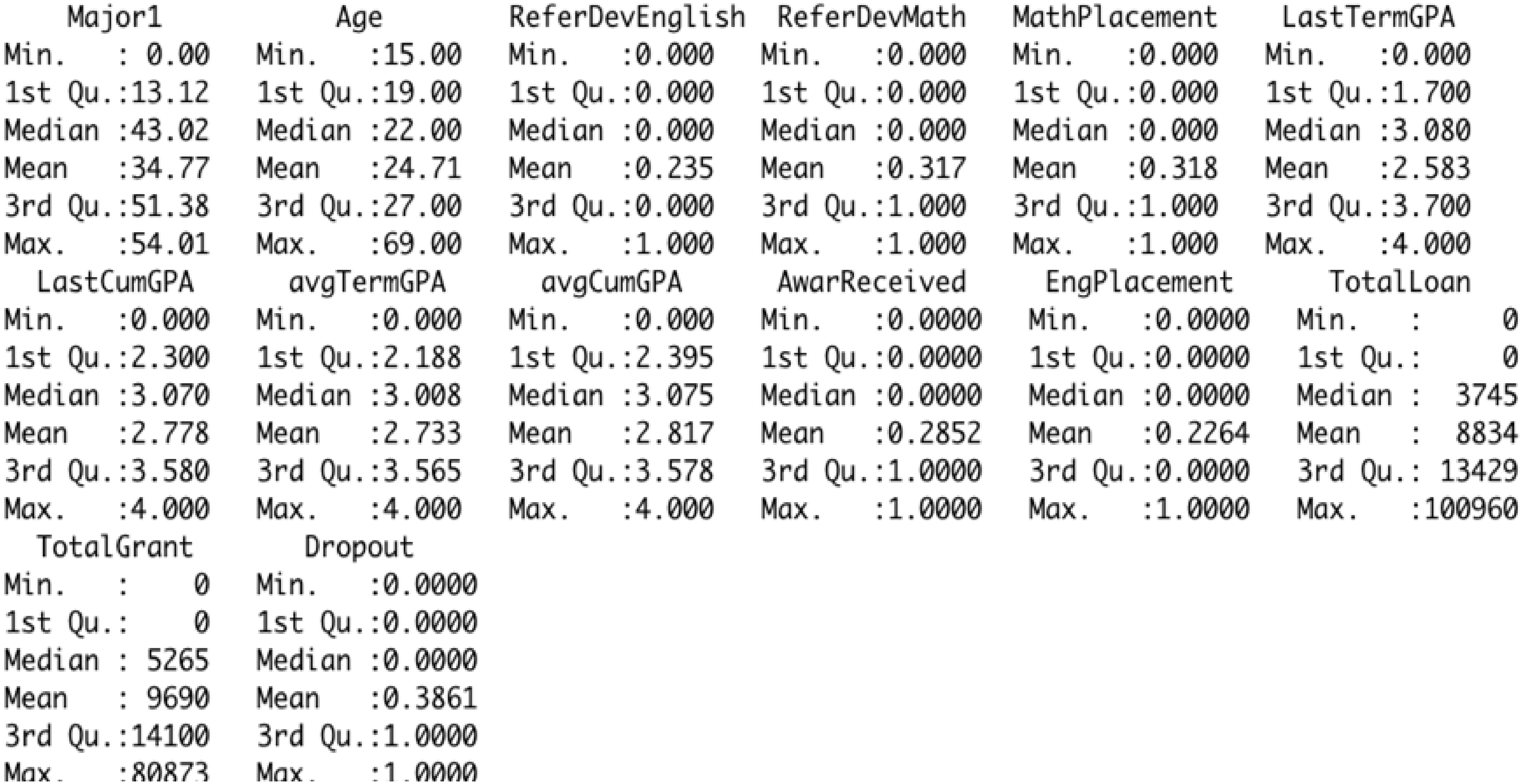




Before: Compare variables by correlation coefficient to find variables that most related to the Dropout rate.

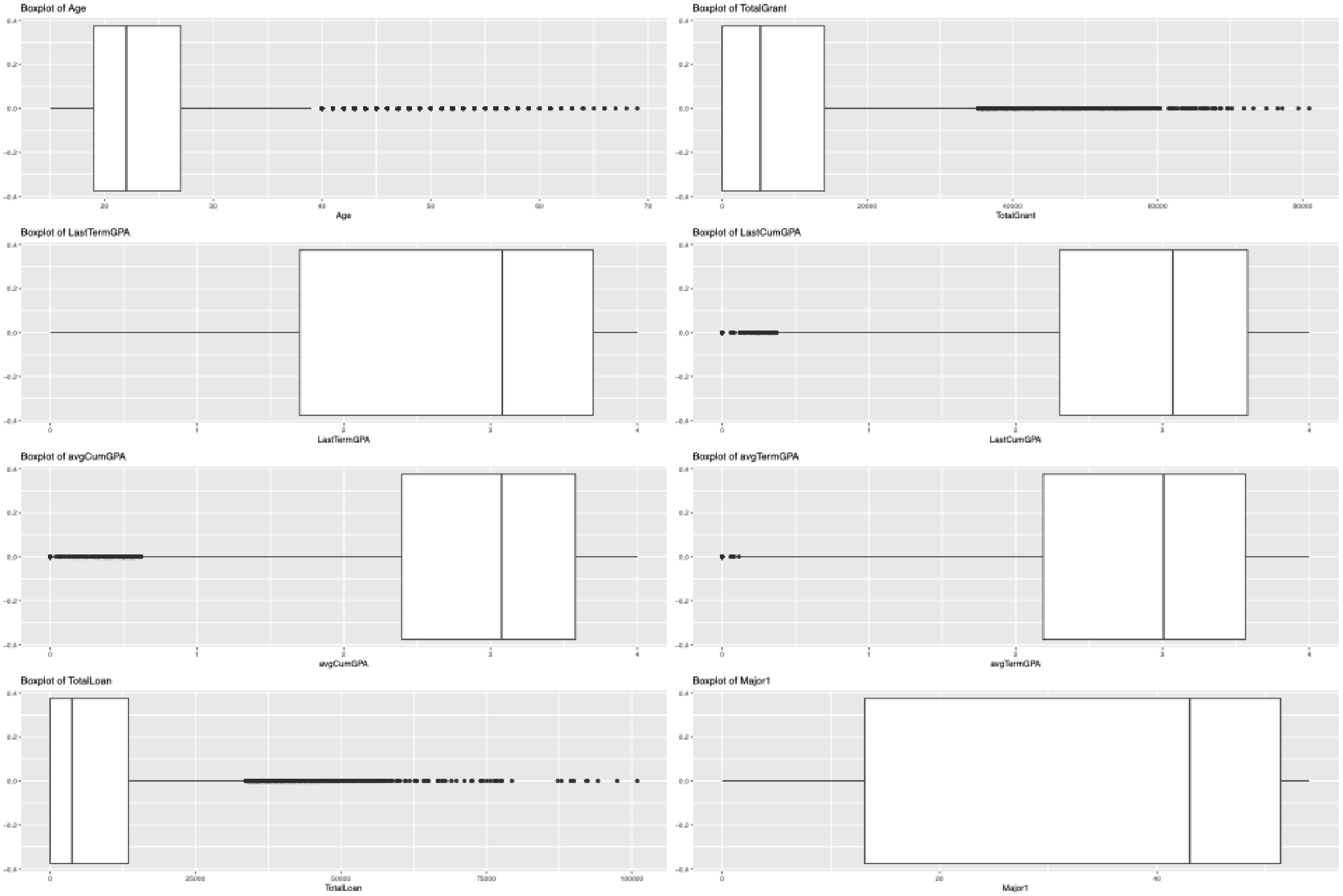
After: Pick top 13 variables that have the highest correlation coefficient and exceed baseline

Summary Statistics



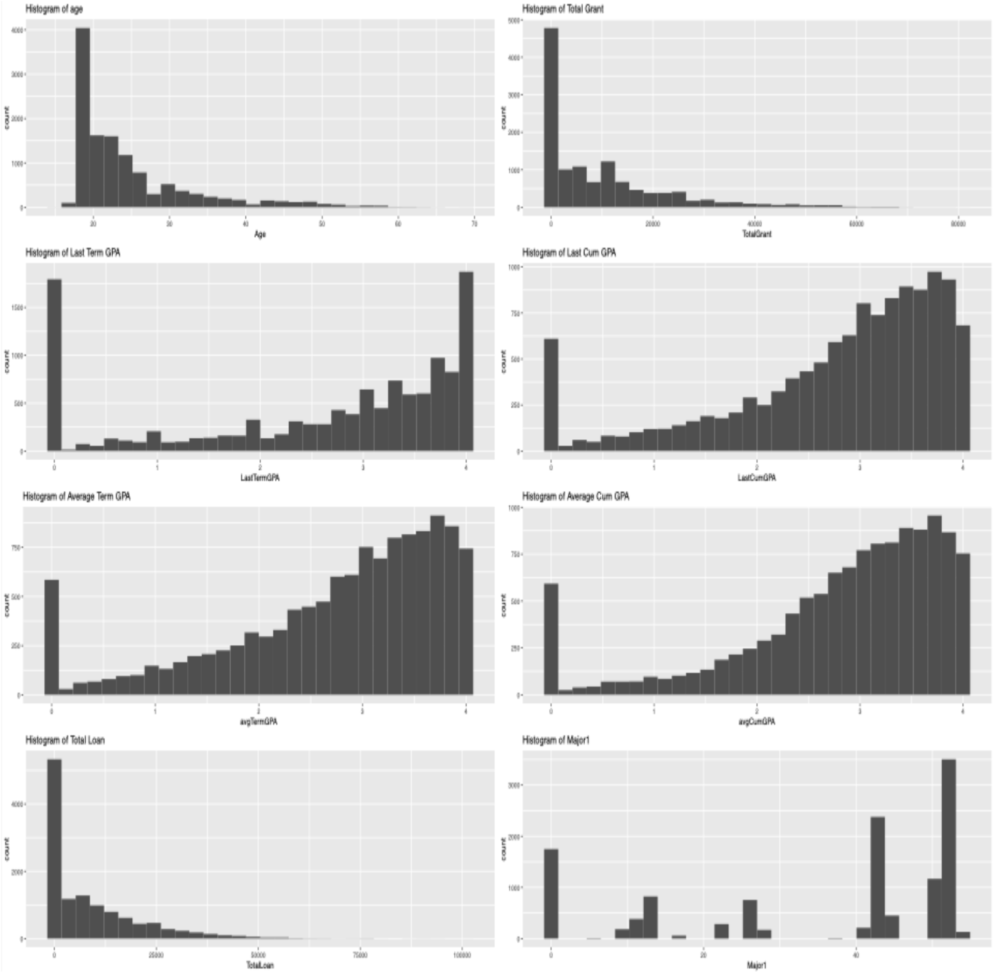
```
     Major1          Age        ReferDevEnglish  ReferDevMath    MathPlacement     LastTermGPA   
 Min.   : 0.00   Min.   :15.00   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000  
 1st Qu.:13.12   1st Qu.:19.00   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:1.700  
 Median :43.02   Median :22.00   Median :0.000   Median :0.000   Median :0.000   Median :3.080  
 Mean   :34.77   Mean   :24.71   Mean   :0.235   Mean   :0.317   Mean   :0.318   Mean   :2.583  
 3rd Qu.:51.38   3rd Qu.:27.00   3rd Qu.:0.000   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:3.700  
 Max.   :54.01   Max.   :69.00   Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :4.000  
   LastCumGPA      avgTermGPA      avgCumGPA      AwarReceived     EngPlacement      TotalLoan     
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :     0  
 1st Qu.:2.300   1st Qu.:2.188   1st Qu.:2.395   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:     0  
 Median :3.070   Median :3.008   Median :3.075   Median :0.0000   Median :0.0000   Median :  3745  
 Mean   :2.778   Mean   :2.733   Mean   :2.817   Mean   :0.2852   Mean   :0.2264   Mean   :  8834  
 3rd Qu.:3.580   3rd Qu.:3.565   3rd Qu.:3.578   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.: 13429  
 Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :1.0000   Max.   :1.0000   Max.   :100960  
   TotalGrant       Dropout      
 Min.   :    0   Min.   :0.0000  
 1st Qu.:    0   1st Qu.:0.0000  
 Median : 5265   Median :0.0000  
 Mean   : 9690   Mean   :0.3861  
 3rd Qu.:14100   3rd Qu.:1.0000  
 Max.   :80873   Max.   :1.0000  
```

Boxplot: Continues Variables

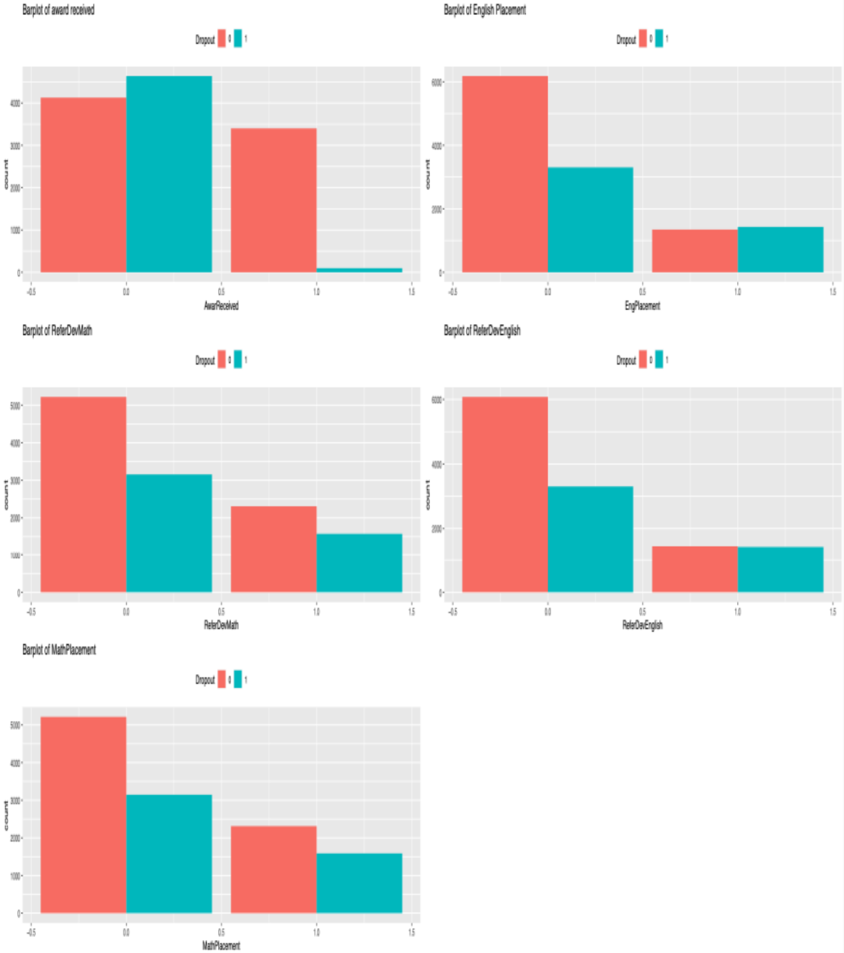


HIstogram: Continues VS. Discrete

Continues Variables



Discrete Variables



Feature Engineering

1. Filter Features

- compute the correlation between each feature and the Dropout
- filter top 13 variables that have the highest correlation coefficient

2. Filter Near - Zero - Variance

- **Zero-variance features**: only have a unique value
- not carry any meaningful information → cause the model crash or become unstable

- **near-zero-variance features** : have a few unique values that occur very rarely
- mislead the model training or even become zero-variance when splitting the data into multiple subsets

- `step_nzv`： remove these variables that are highly sparse and unbalanced

Feature Engineering

3. Impute via k-nearest neighbors

- missing values exist in selected features
- step_impute_knn： impute missing data using nearest neighbors

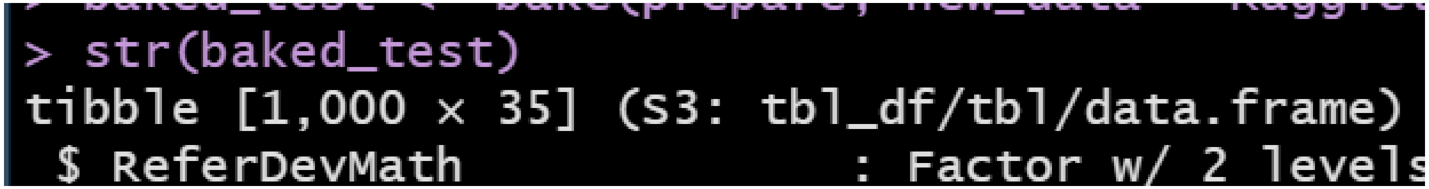
4. Normalization and Standardization

- skewed + differ wildly in scale
- degrade the model's ability to describe typical cases as it has to deal with rare cases on extreme values (especially regression based models)

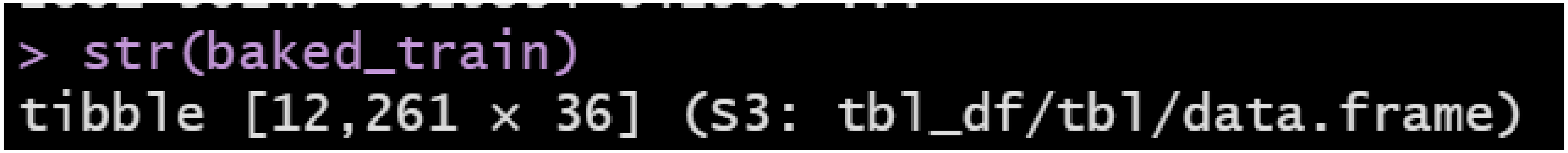
  →transform these features to normal distribution

- step_center： normalize numeric data to have a mean of zero
- step_scale： normalize numeric data to have a standard deviation of one

Variables

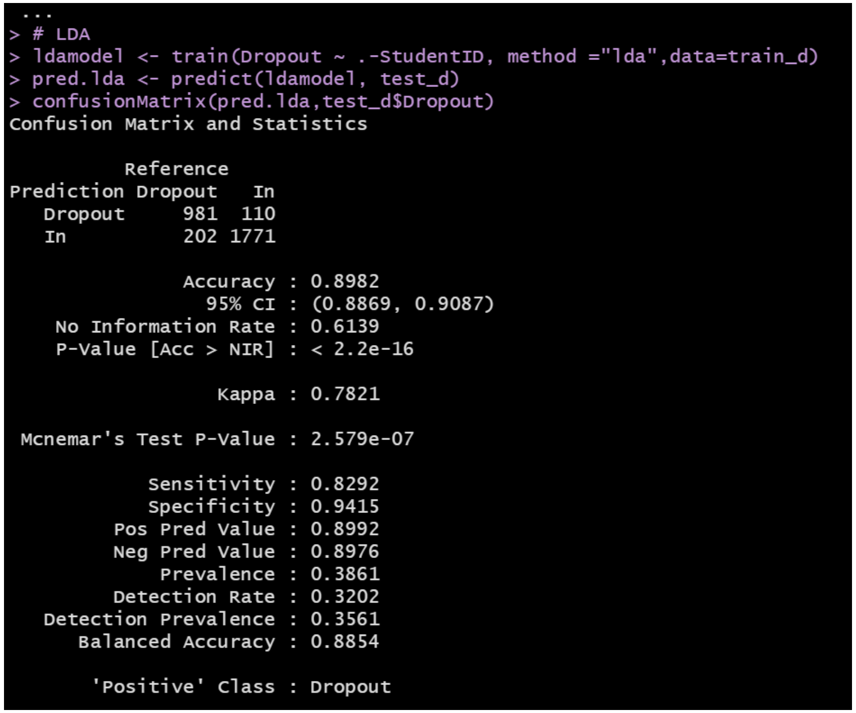


```
> baked_test <- bake(prepare, new_data = Kaggle
> str(baked_test)
tibble [1,000 × 35] (S3: tbl_df/tbl/data.frame)
 $ ReferDevMath                : Factor w/ 2 levels
```

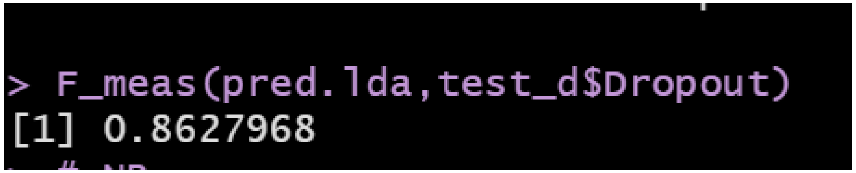


```
> str(baked_train)
tibble [12,261 × 36] (S3: tbl_df/tbl/data.frame)
```

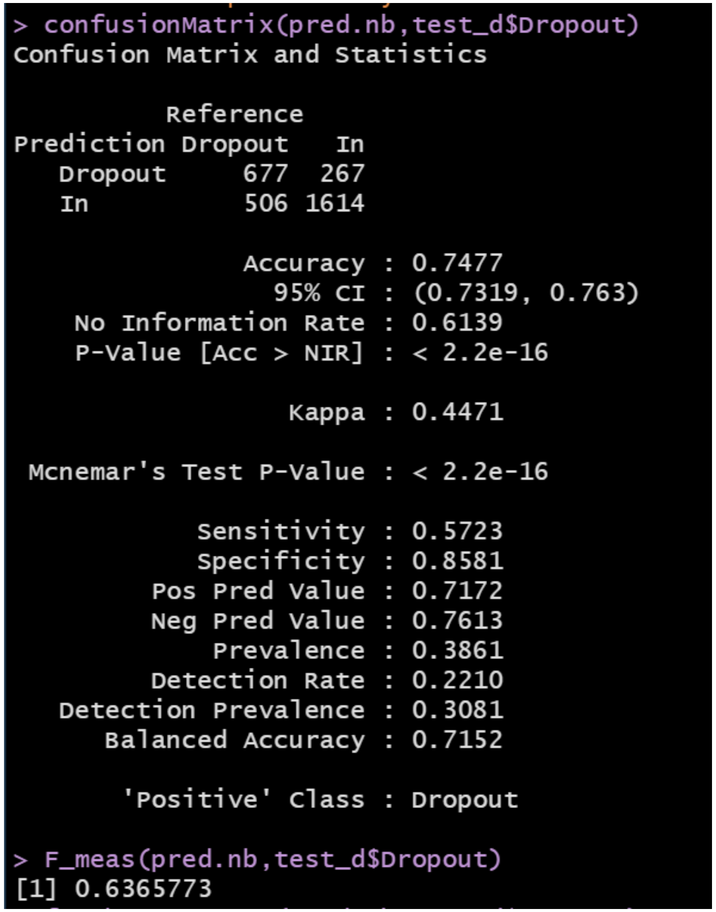
Modeling



```
 ...
> # LDA
> ldamodel <- train(Dropout ~ .-StudentID, method ="lda",data=train_d)
> pred.lda <- predict(ldamodel, test_d)
> confusionMatrix(pred.lda,test_d$Dropout)
Confusion Matrix and Statistics

          Reference
Prediction Dropout   In
   Dropout     981  110
   In          202 1771

               Accuracy : 0.8982
                 95% CI : (0.8869, 0.9087)
    No Information Rate : 0.6139
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7821

 Mcnemar's Test P-Value : 2.579e-07

            Sensitivity : 0.8292
            Specificity : 0.9415
         Pos Pred Value : 0.8992
         Neg Pred Value : 0.8976
             Prevalence : 0.3861
         Detection Rate : 0.3202
   Detection Prevalence : 0.3561
      Balanced Accuracy : 0.8854

       'Positive' Class : Dropout
```

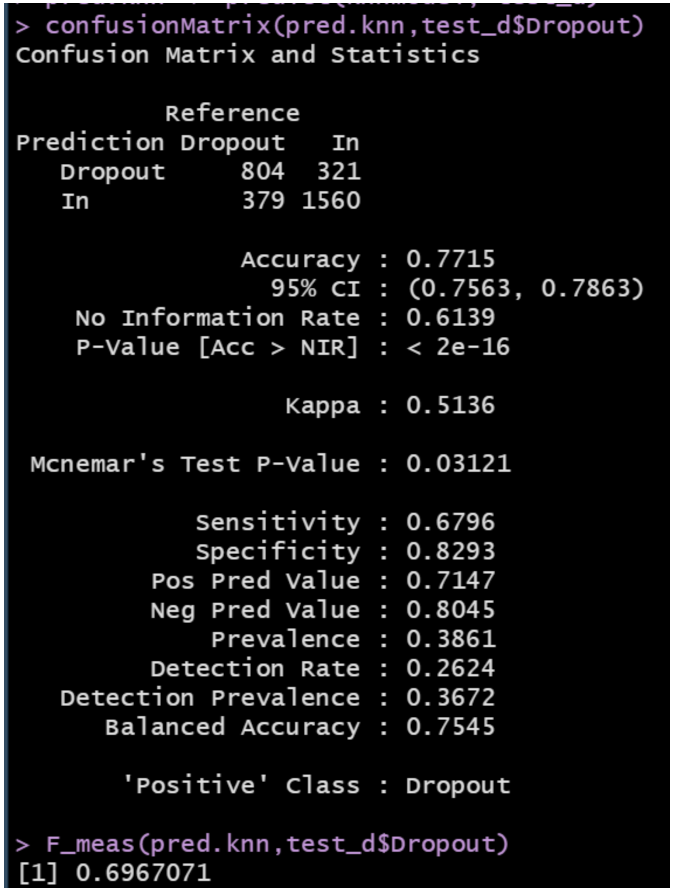


```
> F_meas(pred.lda,test_d$Dropout)
[1] 0.8627968
```

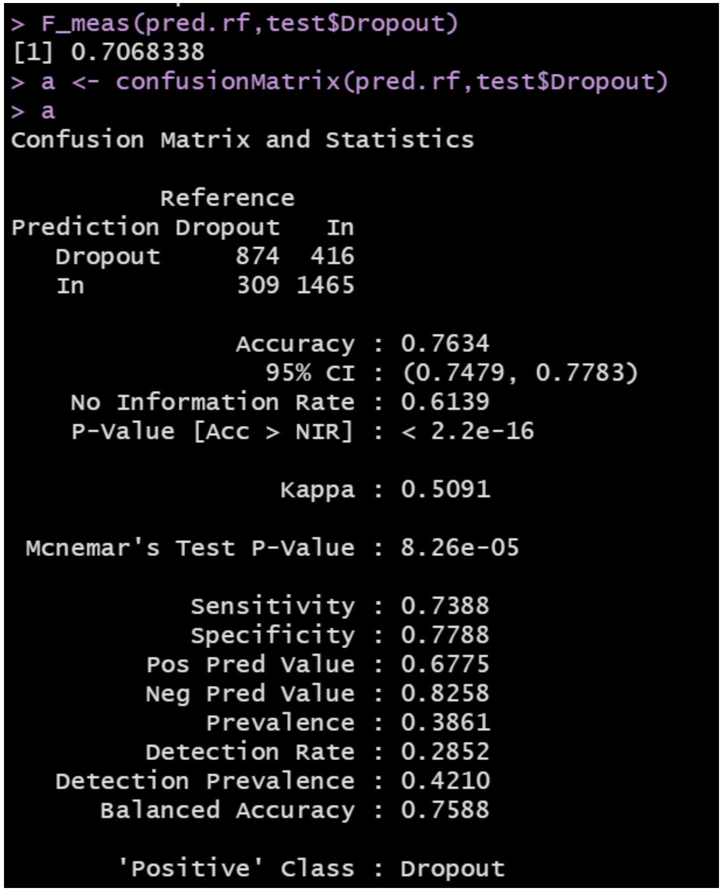


```
> confusionMatrix(pred.nb,test_d$Dropout)
Confusion Matrix and Statistics

          Reference
Prediction Dropout   In
   Dropout     677  267
   In          506 1614

               Accuracy : 0.7477
                 95% CI : (0.7319, 0.763)
    No Information Rate : 0.6139
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4471

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.5723
            Specificity : 0.8581
         Pos Pred Value : 0.7172
         Neg Pred Value : 0.7613
             Prevalence : 0.3861
         Detection Rate : 0.2210
   Detection Prevalence : 0.3081
      Balanced Accuracy : 0.7152

       'Positive' Class : Dropout

> F_meas(pred.nb,test_d$Dropout)
[1] 0.6365773
```

Modeling

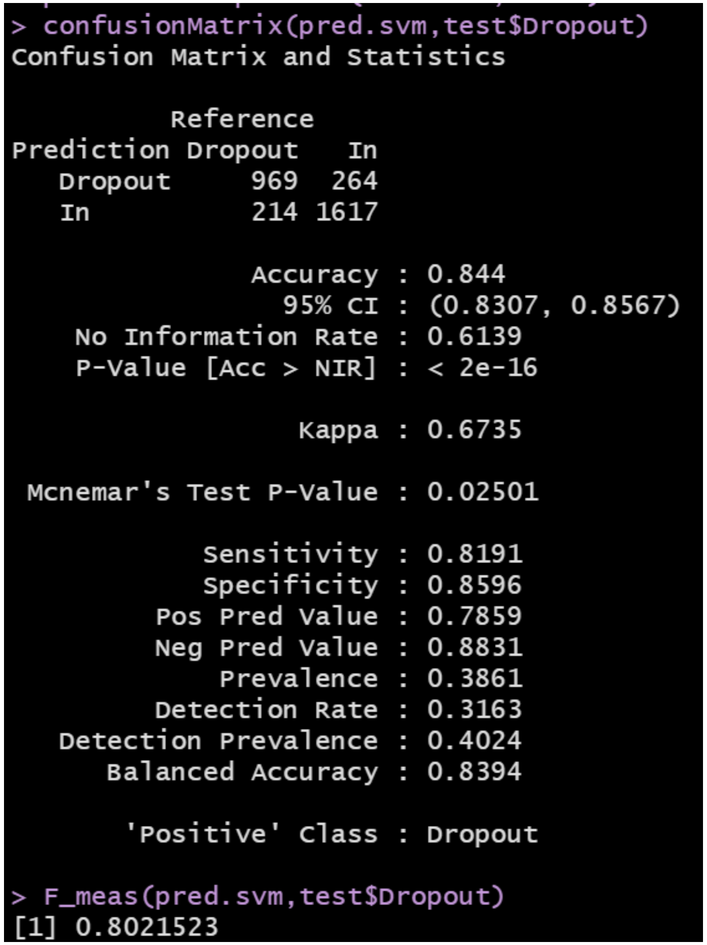


```
> confusionMatrix(pred.knn,test_d$Dropout)
Confusion Matrix and Statistics

          Reference
Prediction Dropout   In
   Dropout     804  321
   In          379 1560

               Accuracy : 0.7715
                 95% CI : (0.7563, 0.7863)
    No Information Rate : 0.6139
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.5136

 Mcnemar's Test P-Value : 0.03121

            Sensitivity : 0.6796
            Specificity : 0.8293
         Pos Pred Value : 0.7147
         Neg Pred Value : 0.8045
             Prevalence : 0.3861
         Detection Rate : 0.2624
   Detection Prevalence : 0.3672
      Balanced Accuracy : 0.7545

       'Positive' Class : Dropout

> F_meas(pred.knn,test_d$Dropout)
[1] 0.6967071
```

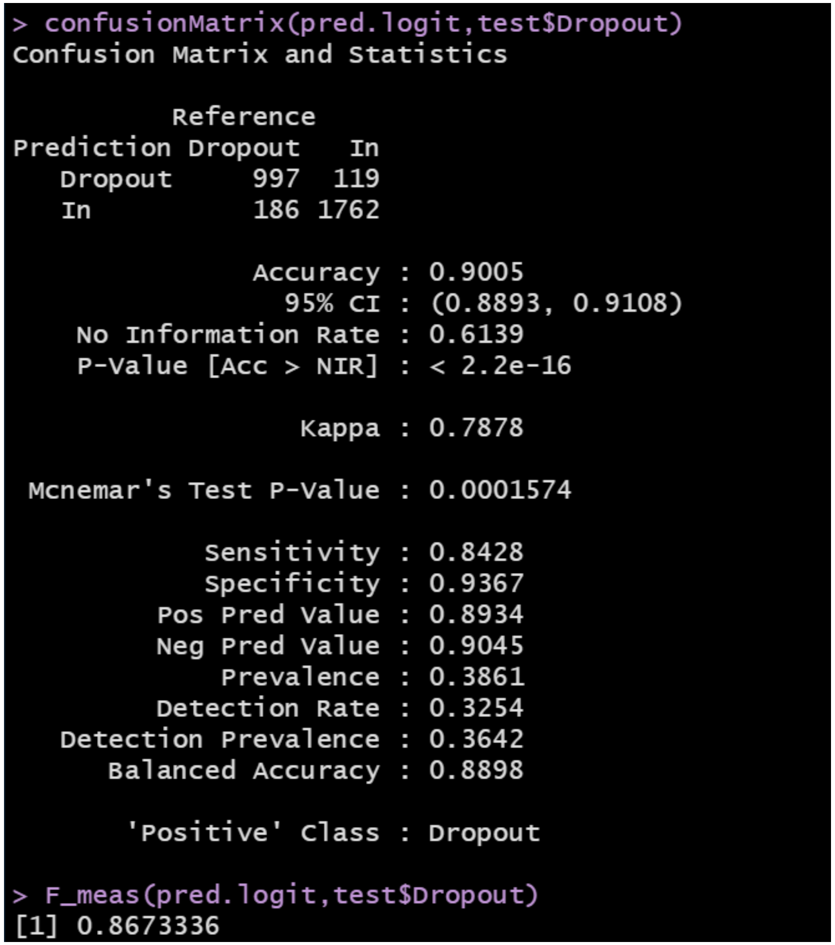


```
> F_meas(pred.rf,test$Dropout)
[1] 0.7068338
> a <- confusionMatrix(pred.rf,test$Dropout)
> a
Confusion Matrix and Statistics

          Reference
Prediction Dropout   In
   Dropout     874  416
   In          309 1465

               Accuracy : 0.7634
                 95% CI : (0.7479, 0.7783)
    No Information Rate : 0.6139
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5091

 Mcnemar's Test P-Value : 8.26e-05

            Sensitivity : 0.7388
            Specificity : 0.7788
         Pos Pred Value : 0.6775
         Neg Pred Value : 0.8258
             Prevalence : 0.3861
         Detection Rate : 0.2852
   Detection Prevalence : 0.4210
      Balanced Accuracy : 0.7588

       'Positive' Class : Dropout
```

Modeling



```
> confusionMatrix(pred.svm,test$Dropout)
Confusion Matrix and Statistics

          Reference
Prediction Dropout   In
   Dropout     969  264
   In          214 1617

               Accuracy : 0.844
                 95% CI : (0.8307, 0.8567)
    No Information Rate : 0.6139
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.6735

 Mcnemar's Test P-Value : 0.02501

            Sensitivity : 0.8191
            Specificity : 0.8596
         Pos Pred Value : 0.7859
         Neg Pred Value : 0.8831
             Prevalence : 0.3861
         Detection Rate : 0.3163
   Detection Prevalence : 0.4024
      Balanced Accuracy : 0.8394

       'Positive' Class : Dropout

> F_meas(pred.svm,test$Dropout)
[1] 0.8021523
```



```
> confusionMatrix(pred.logit,test$Dropout)
Confusion Matrix and Statistics

          Reference
Prediction Dropout   In
   Dropout     997  119
   In          186 1762

               Accuracy : 0.9005
                 95% CI : (0.8893, 0.9108)
    No Information Rate : 0.6139
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7878

 Mcnemar's Test P-Value : 0.0001574

            Sensitivity : 0.8428
            Specificity : 0.9367
         Pos Pred Value : 0.8934
         Neg Pred Value : 0.9045
             Prevalence : 0.3861
         Detection Rate : 0.3254
   Detection Prevalence : 0.3642
      Balanced Accuracy : 0.8898

       'Positive' Class : Dropout

> F_meas(pred.logit,test$Dropout)
[1] 0.8673336
```

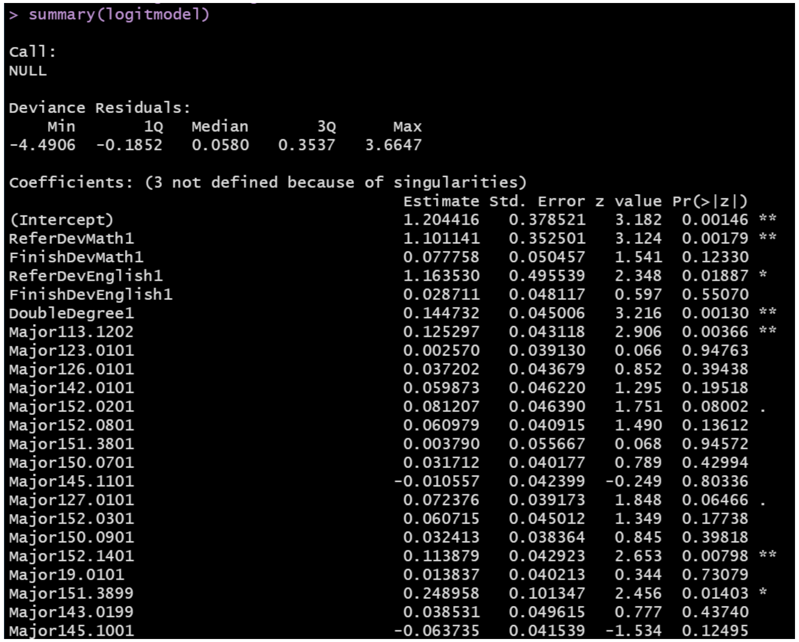
Modeling

Ranking:

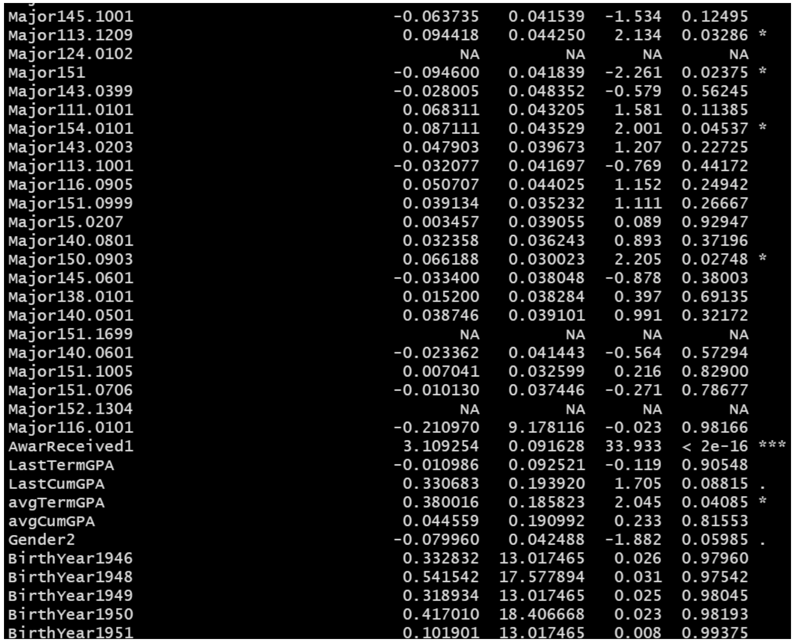
nb < knn < decision tree < svm < lda < logistic

Modeling

logistic

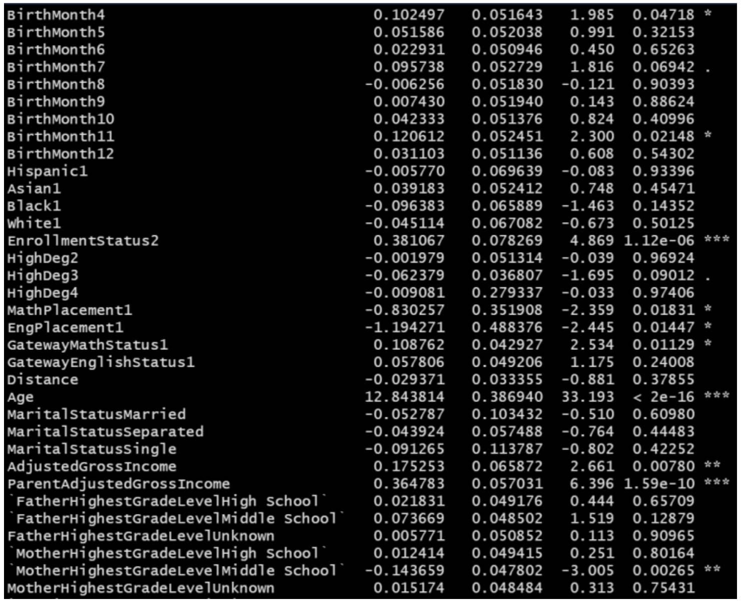


```
> summary(logitmodel)

Call:
NULL

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-4.4906  -0.1852   0.0580   0.3537   3.6647  

Coefficients: (3 not defined because of singularities)
                                         Estimate Std. Error z value Pr(>|z|)    
(Intercept)                              1.204416   0.378521   3.182  0.00146 ** 
ReferDevMath1                            1.101141   0.352501   3.124  0.00179 ** 
FinishDevMath1                           0.077758   0.050457   1.541  0.12330    
ReferDevEnglish1                         1.163530   0.495539   2.348  0.01887 *  
FinishDevEnglish1                        0.028711   0.048117   0.597  0.55070    
DoubleDegree1                            0.144732   0.045006   3.216  0.00130 ** 
Major113.1202                            0.125297   0.043118   2.906  0.00366 ** 
Major123.0101                            0.002570   0.039130   0.066  0.94763    
Major126.0101                            0.037202   0.043679   0.852  0.39438    
Major142.0101                            0.059873   0.046220   1.295  0.19518    
Major152.0201                            0.081207   0.046390   1.751  0.08002 .  
Major152.0801                            0.060979   0.040915   1.490  0.13612    
Major151.3801                            0.003790   0.055667   0.068  0.94572    
Major150.0701                            0.031712   0.040177   0.789  0.42994    
Major145.1101                           -0.010557   0.042399  -0.249  0.80336    
Major127.0101                            0.072376   0.039173   1.848  0.06466 .  
Major152.0301                            0.060715   0.045012   1.349  0.17738    
Major150.0901                            0.032413   0.038364   0.845  0.39818    
Major152.1401                            0.113879   0.042923   2.653  0.00798 ** 
Major19.0101                             0.013837   0.040213   0.344  0.73079    
Major151.3899                            0.248958   0.101347   2.456  0.01403 *  
Major143.0199                            0.038531   0.049615   0.777  0.43740    
Major145.1001                           -0.063735   0.041539  -1.534  0.12495    
```

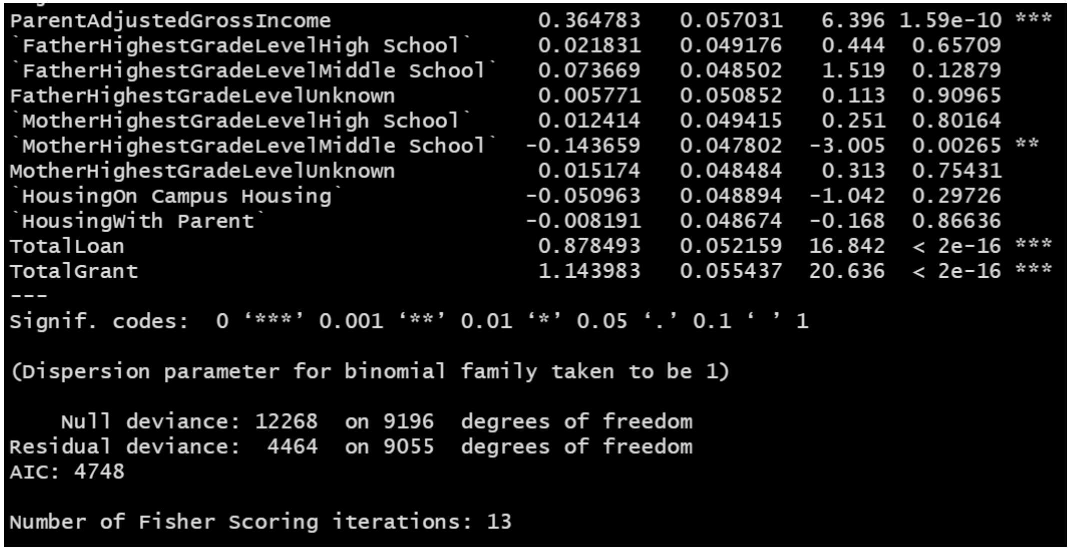


```
Major145.1001                           -0.063735   0.041539  -1.534  0.12495    
Major113.1209                            0.094418   0.044250   2.134  0.03286 *  
Major124.0102                                   NA         NA      NA       NA    
Major151                                -0.094600   0.041839  -2.261  0.02375 *  
Major143.0399                           -0.028005   0.048352  -0.579  0.56245    
Major111.0101                            0.068311   0.043205   1.581  0.11385    
Major154.0101                            0.087111   0.043529   2.001  0.04537 *  
Major143.0203                            0.047903   0.039673   1.207  0.22725    
Major113.1001                           -0.032077   0.041697  -0.769  0.44172    
Major116.0905                            0.050707   0.044025   1.152  0.24942    
Major151.0999                            0.039134   0.035232   1.111  0.26667    
Major15.0207                             0.003457   0.039055   0.089  0.92947    
Major140.0801                            0.032358   0.036243   0.893  0.37196    
Major150.0903                            0.066188   0.030023   2.205  0.02748 *  
Major145.0601                           -0.033400   0.038048  -0.878  0.38003    
Major138.0101                            0.015200   0.038284   0.397  0.69135    
Major140.0501                            0.038746   0.039101   0.991  0.32172    
Major151.1699                                   NA         NA      NA       NA    
Major140.0601                           -0.023362   0.041443  -0.564  0.57294    
Major151.1005                            0.007041   0.032599   0.216  0.82900    
Major151.0706                           -0.010130   0.037446  -0.271  0.78677    
Major152.1304                                   NA         NA      NA       NA    
Major116.0101                           -0.210970   9.178116  -0.023  0.98166    
AwarReceived1                            3.109254   0.091628  33.933  < 2e-16 ***
LastTermGPA                             -0.010986   0.092521  -0.119  0.90548    
LastCumGPA                               0.330683   0.193920   1.705  0.08815 .  
avgTermGPA                               0.380016   0.185823   2.045  0.04085 *  
avgCumGPA                                0.044559   0.190992   0.233  0.81553    
Gender2                                 -0.079960   0.042488  -1.882  0.05985 .  
BirthYear1946                            0.332832  13.017465   0.026  0.97960    
BirthYear1948                            0.541542  17.577894   0.031  0.97542    
BirthYear1949                            0.318934  13.017465   0.025  0.98045    
BirthYear1950                            0.417010  18.406668   0.023  0.98193    
BirthYear1951                            0.101901  13.017465   0.008  0.99375    
```

Modeling



```
BirthMonth4                                 0.102497   0.051643   1.985  0.04718 *
BirthMonth5                                 0.051586   0.052038   0.991  0.32153
BirthMonth6                                 0.022931   0.050946   0.450  0.65263
BirthMonth7                                 0.095738   0.052729   1.816  0.06942 .
BirthMonth8                                -0.006256   0.051830  -0.121  0.90393
BirthMonth9                                 0.007430   0.051940   0.143  0.88624
BirthMonth10                                0.042333   0.051376   0.824  0.40996
BirthMonth11                                0.120612   0.052451   2.300  0.02148 *
BirthMonth12                                0.031103   0.051136   0.608  0.54302
Hispanic1                                  -0.005770   0.069639  -0.083  0.93396
Asian1                                      0.039183   0.052412   0.748  0.45471
Black1                                     -0.096383   0.065889  -1.463  0.14352
White1                                     -0.045114   0.067082  -0.673  0.50125
EnrollmentStatus2                           0.381067   0.078269   4.869 1.12e-06 ***
HighDeg2                                   -0.001979   0.051314  -0.039  0.96924
HighDeg3                                   -0.062379   0.036807  -1.695  0.09012 .
HighDeg4                                   -0.009081   0.279337  -0.033  0.97406
MathPlacement1                             -0.830257   0.351908  -2.359  0.01831 *
EngPlacement1                              -1.194271   0.488376  -2.445  0.01447 *
GatewayMathStatus1                          0.108762   0.042927   2.534  0.01129 *
GatewayEnglishStatus1                       0.057806   0.049206   1.175  0.24008
Distance                                   -0.029371   0.033355  -0.881  0.37855
Age                                        12.843814   0.386940  33.193  < 2e-16 ***
MaritalStatusMarried                       -0.052787   0.103432  -0.510  0.60980
MaritalStatusSeparated                     -0.043924   0.057488  -0.764  0.44483
MaritalStatusSingle                        -0.091265   0.113787  -0.802  0.42252
AdjustedGrossIncome                         0.175253   0.065872   2.661  0.00780 **
ParentAdjustedGrossIncome                   0.364783   0.057031   6.396 1.59e-10 ***
`FatherHighestGradeLevelHigh School`        0.021831   0.049176   0.444  0.65709
`FatherHighestGradeLevelMiddle School`      0.073669   0.048502   1.519  0.12879
FatherHighestGradeLevelUnknown              0.005771   0.050852   0.113  0.90965
`MotherHighestGradeLevelHigh School`        0.012414   0.049415   0.251  0.80164
`MotherHighestGradeLevelMiddle School`     -0.143659   0.047802  -3.005  0.00265 **
MotherHighestGradeLevelUnknown              0.015174   0.048484   0.313  0.75431
```



```
ParentAdjustedGrossIncome                0.364783   0.057031   6.396 1.59e-10 ***
`FatherHighestGradeLevelHigh School`     0.021831   0.049176   0.444  0.65709
`FatherHighestGradeLevelMiddle School`   0.073669   0.048502   1.519  0.12879
FatherHighestGradeLevelUnknown           0.005771   0.050852   0.113  0.90965
`MotherHighestGradeLevelHigh School`     0.012414   0.049415   0.251  0.80164
`MotherHighestGradeLevelMiddle School`  -0.143659   0.047802  -3.005  0.00265 **
MotherHighestGradeLevelUnknown           0.015174   0.048484   0.313  0.75431
`HousingOn Campus Housing`              -0.050963   0.048894  -1.042  0.29726
`HousingWith Parent`                    -0.008191   0.048674  -0.168  0.86636
TotalLoan                                0.878493   0.052159  16.842  < 2e-16 ***
TotalGrant                               1.143983   0.055437  20.636  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12268  on 9196  degrees of freedom
Residual deviance:  4464  on 9055  degrees of freedom
AIC: 4748

Number of Fisher Scoring iterations: 13
```

Modeling

Significant factors:

- **AwarRecevied1**
- **EnrollmentStatus2**
- **Age**
- **ParentAdjustedGrossIncome**
- **TotalLoan**
- **TotalGrant**

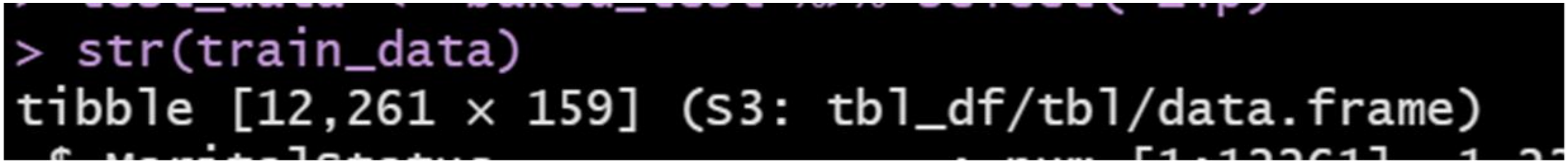
Conclusion 1

What we could do to prevent dropout?

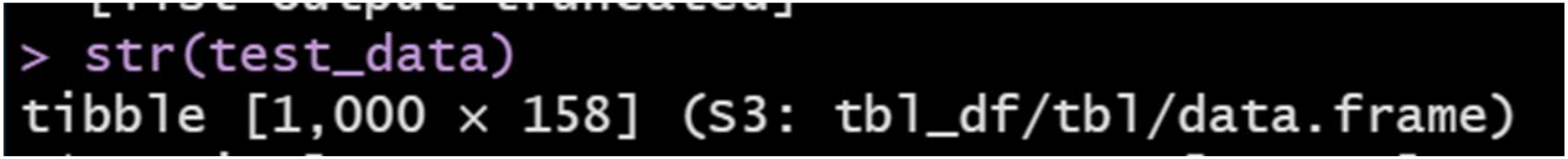
- **Give students more awards to encourage them to learn**
- **Offer more financial aid for students with financial difficulties**
- **Set up special courses for students according to age groups, such as setting up career-oriented courses for older students**
- **Increase total loan amount for students to help them complete studies**

Modeling

Variables

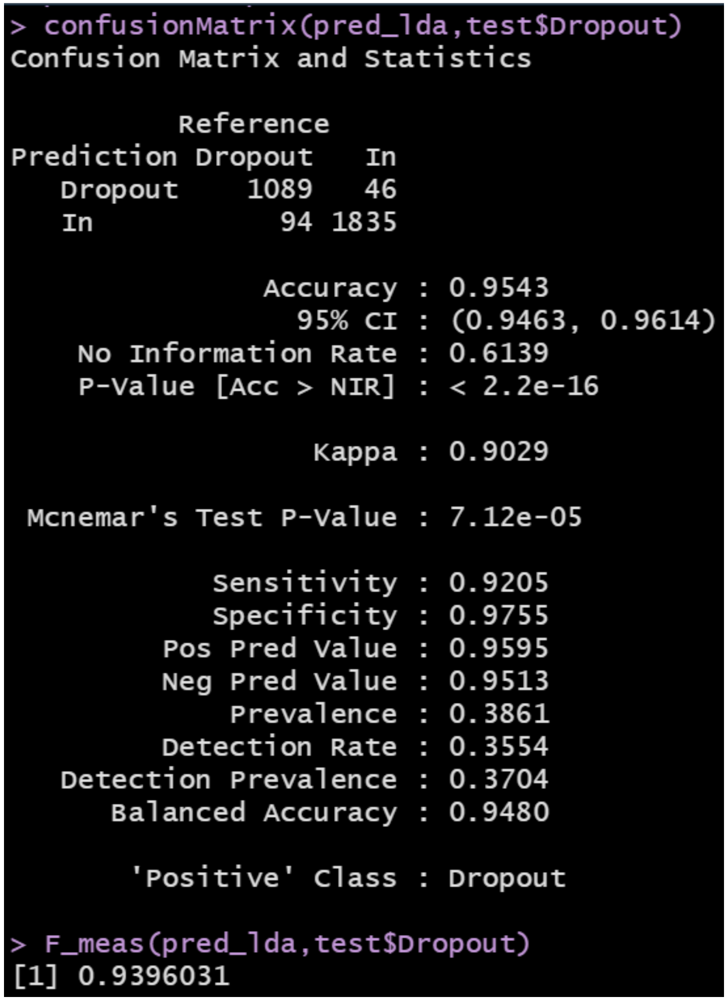


```
> str(train_data)
tibble [12,261 × 159] (S3: tbl_df/tbl/data.frame)
```

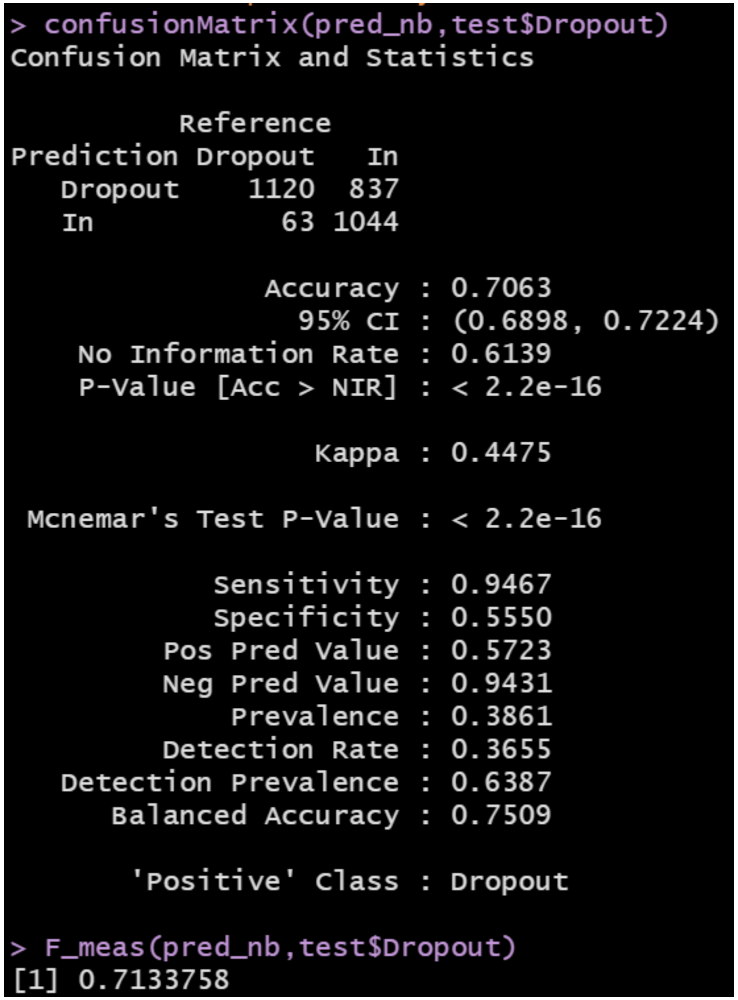


```
> str(test_data)
tibble [1,000 × 158] (S3: tbl_df/tbl/data.frame)
```

Modeling

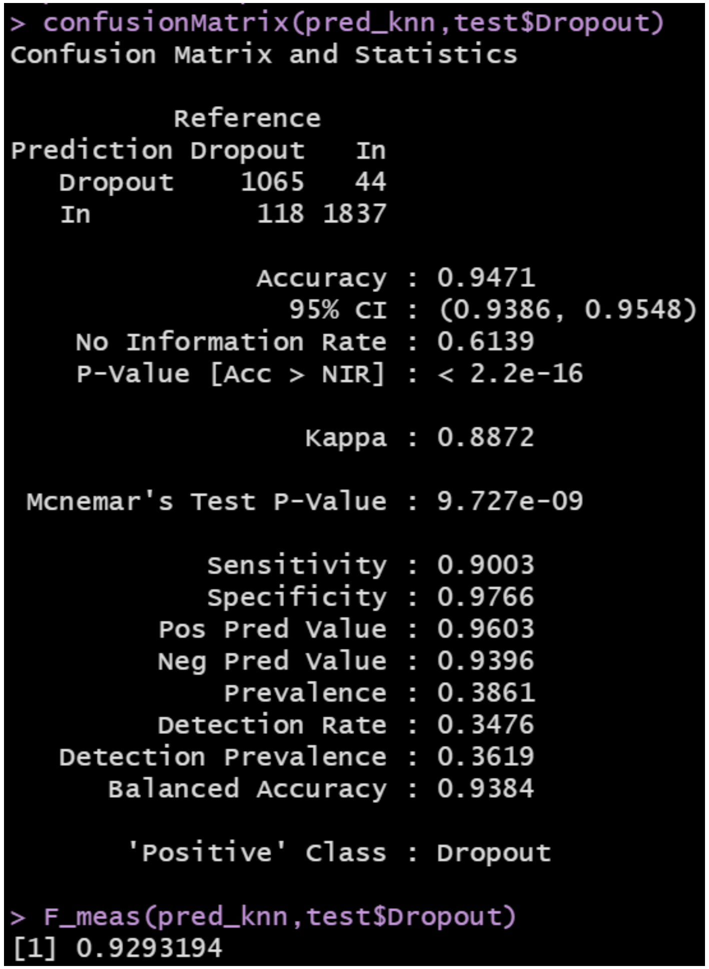


```
> confusionMatrix(pred_lda,test$Dropout)
Confusion Matrix and Statistics

          Reference
Prediction Dropout   In
   Dropout    1089   46
   In           94 1835

               Accuracy : 0.9543
                 95% CI : (0.9463, 0.9614)
    No Information Rate : 0.6139
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9029

 Mcnemar's Test P-Value : 7.12e-05

            Sensitivity : 0.9205
            Specificity : 0.9755
         Pos Pred Value : 0.9595
         Neg Pred Value : 0.9513
             Prevalence : 0.3861
         Detection Rate : 0.3554
   Detection Prevalence : 0.3704
      Balanced Accuracy : 0.9480

       'Positive' Class : Dropout

> F_meas(pred_lda,test$Dropout)
[1] 0.9396031
```

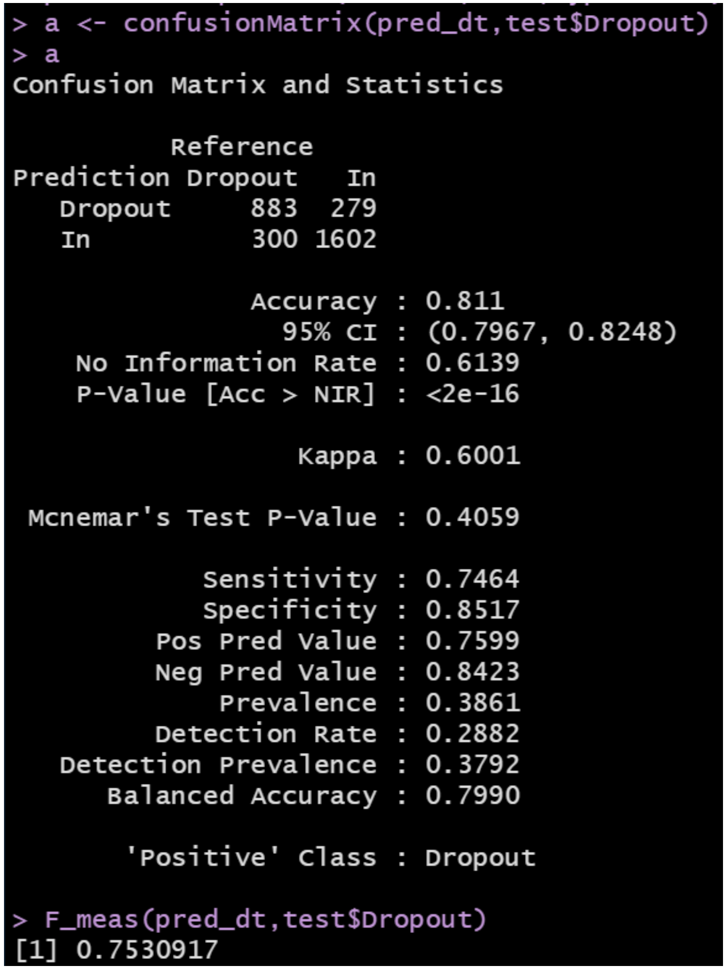


```
> confusionMatrix(pred_nb,test$Dropout)
Confusion Matrix and Statistics

          Reference
Prediction Dropout   In
   Dropout    1120  837
   In           63 1044

               Accuracy : 0.7063
                 95% CI : (0.6898, 0.7224)
    No Information Rate : 0.6139
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4475

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9467
            Specificity : 0.5550
         Pos Pred Value : 0.5723
         Neg Pred Value : 0.9431
             Prevalence : 0.3861
         Detection Rate : 0.3655
   Detection Prevalence : 0.6387
      Balanced Accuracy : 0.7509

       'Positive' Class : Dropout

> F_meas(pred_nb,test$Dropout)
[1] 0.7133758
```

Modeling

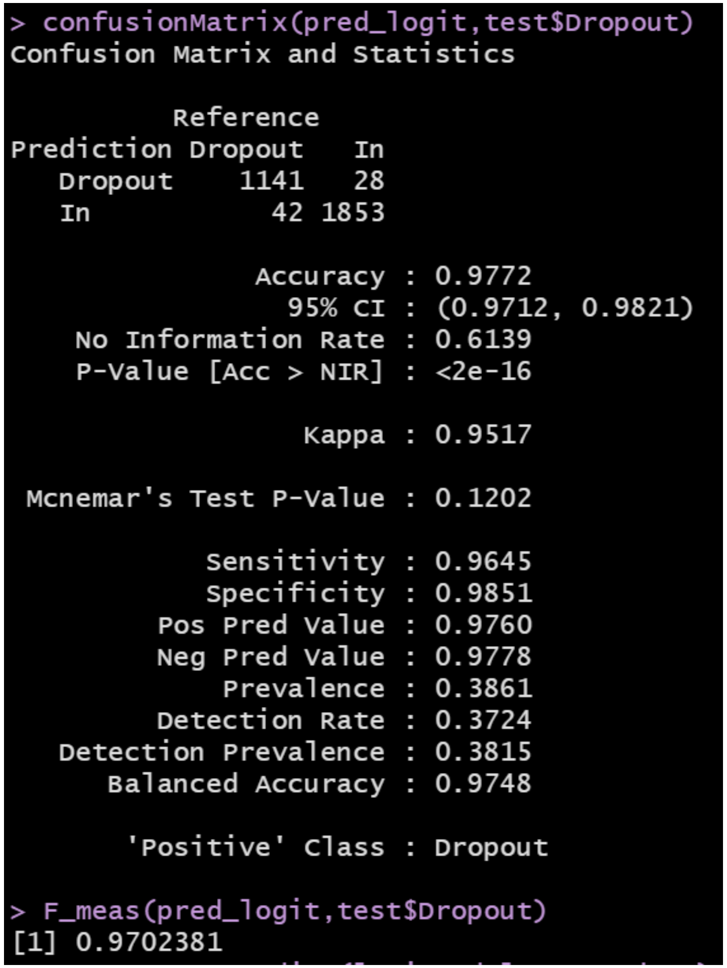


```
> confusionMatrix(pred_knn,test$Dropout)
Confusion Matrix and Statistics

          Reference
Prediction Dropout   In
   Dropout    1065   44
   In          118 1837

               Accuracy : 0.9471
                 95% CI : (0.9386, 0.9548)
    No Information Rate : 0.6139
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8872

 Mcnemar's Test P-Value : 9.727e-09

            Sensitivity : 0.9003
            Specificity : 0.9766
         Pos Pred Value : 0.9603
         Neg Pred Value : 0.9396
             Prevalence : 0.3861
         Detection Rate : 0.3476
   Detection Prevalence : 0.3619
      Balanced Accuracy : 0.9384

       'Positive' Class : Dropout

> F_meas(pred_knn,test$Dropout)
[1] 0.9293194
```

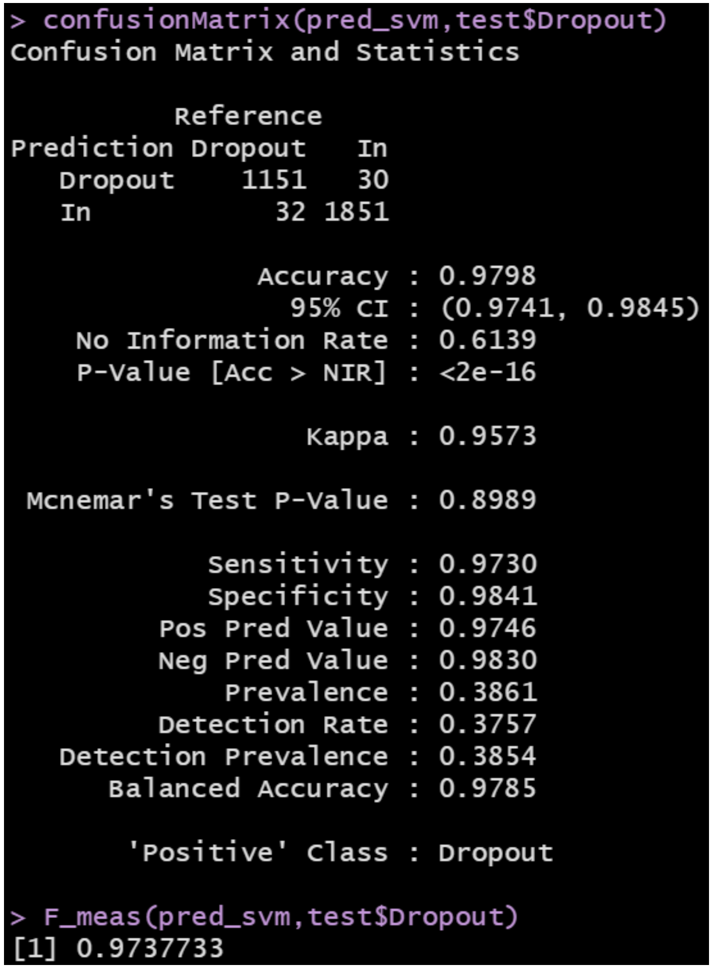


```
> a <- confusionMatrix(pred_dt,test$Dropout)
> a
Confusion Matrix and Statistics

          Reference
Prediction Dropout   In
   Dropout     883  279
   In          300 1602

               Accuracy : 0.811
                 95% CI : (0.7967, 0.8248)
    No Information Rate : 0.6139
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6001

 Mcnemar's Test P-Value : 0.4059

            Sensitivity : 0.7464
            Specificity : 0.8517
         Pos Pred Value : 0.7599
         Neg Pred Value : 0.8423
             Prevalence : 0.3861
         Detection Rate : 0.2882
   Detection Prevalence : 0.3792
      Balanced Accuracy : 0.7990

       'Positive' Class : Dropout

> F_meas(pred_dt,test$Dropout)
[1] 0.7530917
```

Modeling

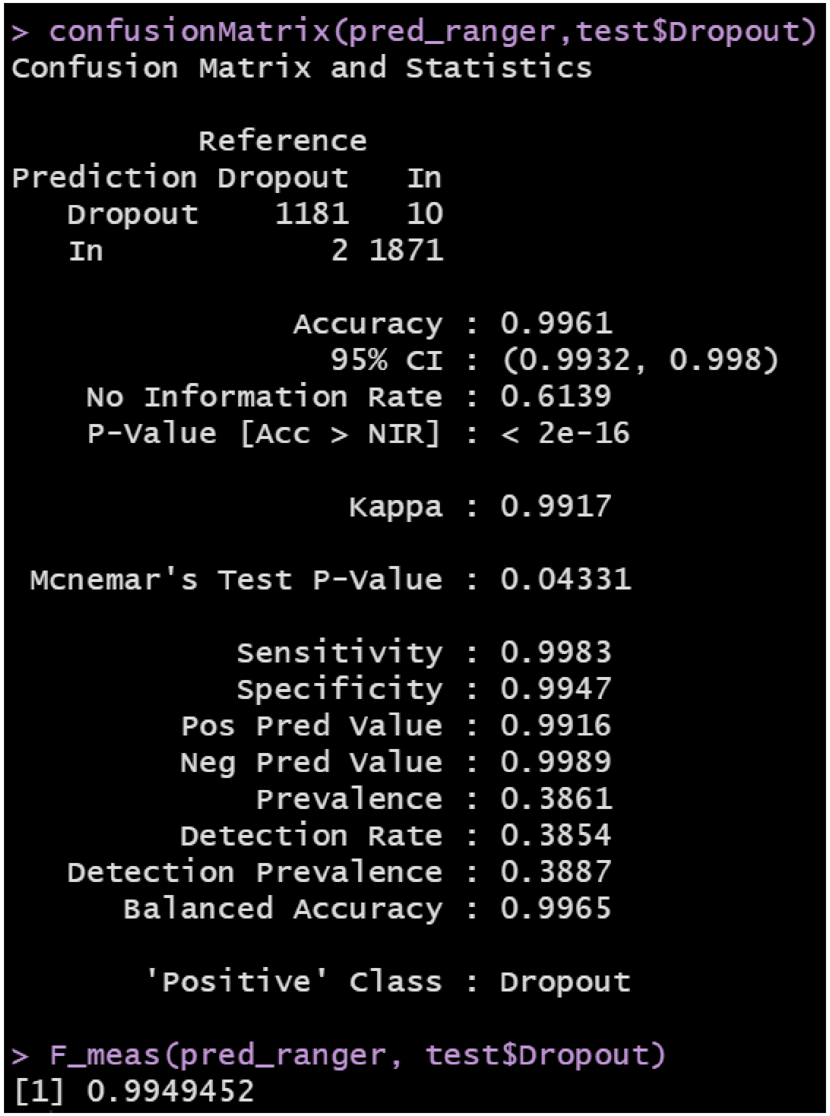


```
> confusionMatrix(pred_logit,test$Dropout)
Confusion Matrix and Statistics

          Reference
Prediction Dropout   In
   Dropout    1141   28
   In           42 1853

               Accuracy : 0.9772
                 95% CI : (0.9712, 0.9821)
    No Information Rate : 0.6139
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9517

 Mcnemar's Test P-Value : 0.1202

            Sensitivity : 0.9645
            Specificity : 0.9851
         Pos Pred Value : 0.9760
         Neg Pred Value : 0.9778
             Prevalence : 0.3861
         Detection Rate : 0.3724
   Detection Prevalence : 0.3815
      Balanced Accuracy : 0.9748

       'Positive' Class : Dropout

> F_meas(pred_logit,test$Dropout)
[1] 0.9702381
```

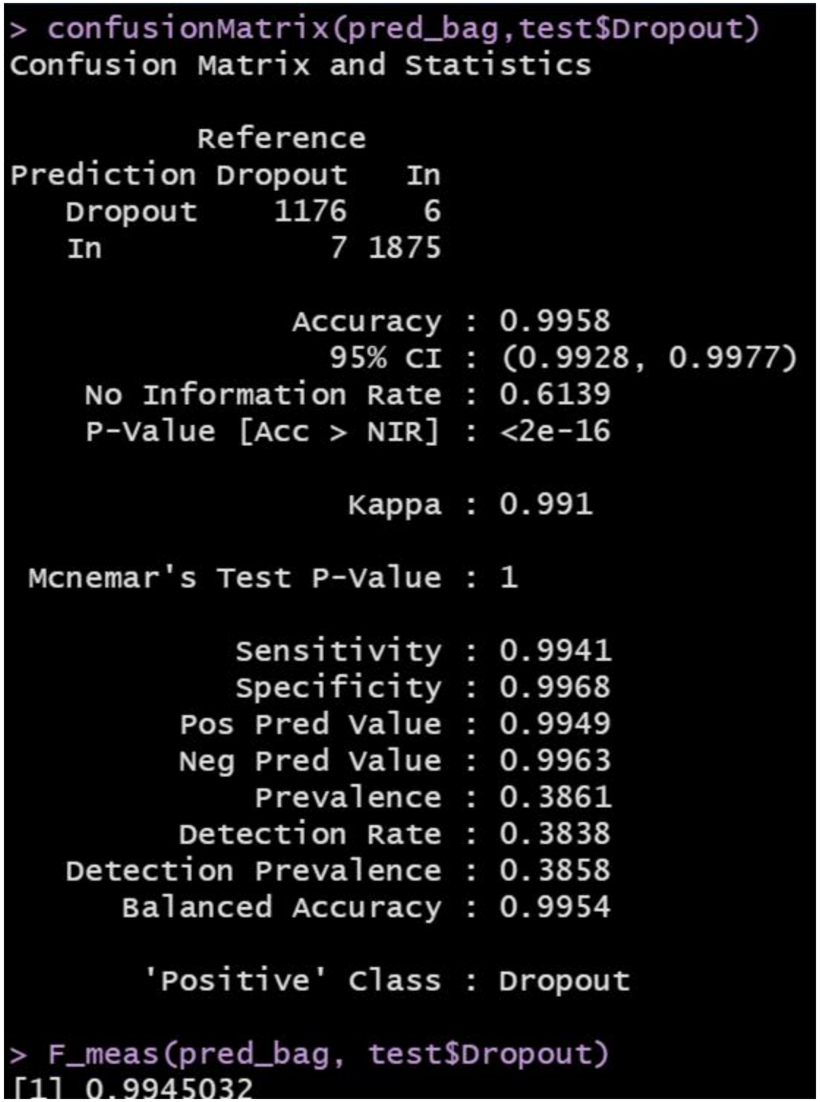


```
> confusionMatrix(pred_svm,test$Dropout)
Confusion Matrix and Statistics

          Reference
Prediction Dropout   In
   Dropout    1151   30
   In           32 1851

               Accuracy : 0.9798
                 95% CI : (0.9741, 0.9845)
    No Information Rate : 0.6139
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9573

 Mcnemar's Test P-Value : 0.8989

            Sensitivity : 0.9730
            Specificity : 0.9841
         Pos Pred Value : 0.9746
         Neg Pred Value : 0.9830
             Prevalence : 0.3861
         Detection Rate : 0.3757
   Detection Prevalence : 0.3854
      Balanced Accuracy : 0.9785

       'Positive' Class : Dropout

> F_meas(pred_svm,test$Dropout)
[1] 0.9737733
```

Modeling



```
> confusionMatrix(pred_ranger,test$Dropout)
Confusion Matrix and Statistics

          Reference
Prediction Dropout   In
   Dropout    1181   10
   In            2 1871

               Accuracy : 0.9961
                 95% CI : (0.9932, 0.998)
    No Information Rate : 0.6139
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.9917

 Mcnemar's Test P-Value : 0.04331

            Sensitivity : 0.9983
            Specificity : 0.9947
         Pos Pred Value : 0.9916
         Neg Pred Value : 0.9989
             Prevalence : 0.3861
         Detection Rate : 0.3854
   Detection Prevalence : 0.3887
      Balanced Accuracy : 0.9965

       'Positive' Class : Dropout

> F_meas(pred_ranger, test$Dropout)
[1] 0.9949452
```



```
> confusionMatrix(pred_bag,test$Dropout)
Confusion Matrix and Statistics

          Reference
Prediction Dropout   In
   Dropout    1176    6
   In            7 1875

               Accuracy : 0.9958
                 95% CI : (0.9928, 0.9977)
    No Information Rate : 0.6139
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.991

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9941
            Specificity : 0.9968
         Pos Pred Value : 0.9949
         Neg Pred Value : 0.9963
             Prevalence : 0.3861
         Detection Rate : 0.3838
   Detection Prevalence : 0.3858
      Balanced Accuracy : 0.9954

       'Positive' Class : Dropout

> F_meas(pred_bag, test$Dropout)
[1] 0.9945032
```

Modeling

Ranking:

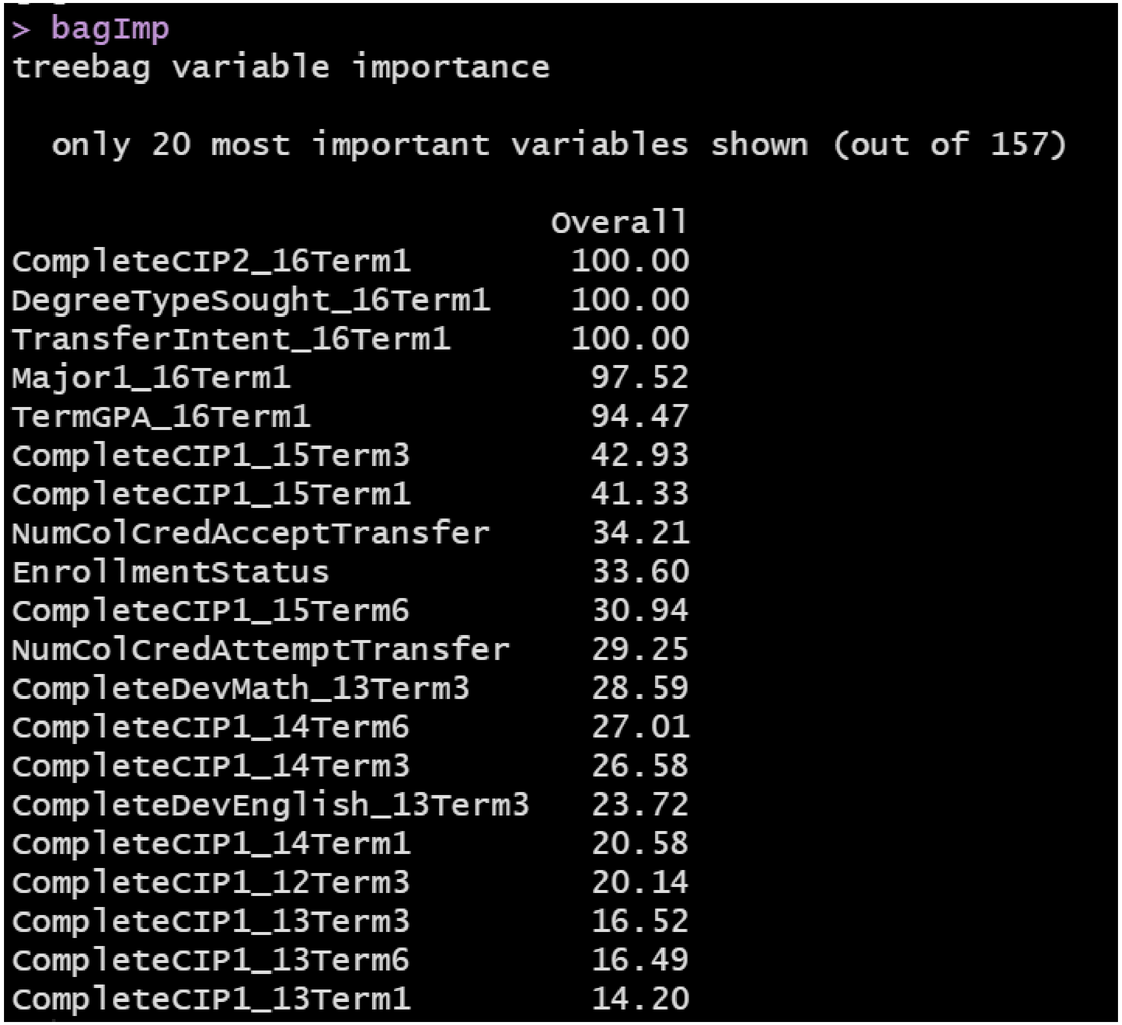
nb < decision tree < knn < lda < logistic < svm < bagging < ranger

Modeling

Ranking:

nb < knn < decision tree < svm < lda < logistic

Modeling



```
> bagImp
treebag variable importance

  only 20 most important variables shown (out of 157)

                           Overall
CompleteCIP2_16Term1        100.00
DegreeTypeSought_16Term1    100.00
TransferIntent_16Term1      100.00
Major1_16Term1               97.52
TermGPA_16Term1              94.47
CompleteCIP1_15Term3         42.93
CompleteCIP1_15Term1         41.33
NumColCredAcceptTransfer     34.21
EnrollmentStatus             33.60
CompleteCIP1_15Term6         30.94
NumColCredAttemptTransfer    29.25
CompleteDevMath_13Term3      28.59
CompleteCIP1_14Term6         27.01
CompleteCIP1_14Term3         26.58
CompleteDevEnglish_13Term3   23.72
CompleteCIP1_14Term1         20.58
CompleteCIP1_12Term3         20.14
CompleteCIP1_13Term3         16.52
CompleteCIP1_13Term6         16.49
CompleteCIP1_13Term1         14.20
```

Conclusion 2

What we could do to prevent dropout?

- **Pay more attention to students who are at their last year.**
- **Give students more supports to help them improve their GPA.**
- **Give students more supports to help them with developmental Math and English.**
- **Pay more attention to students' major status in their last year.**
- **Do further research on the relationship between students' enrollment status and dropout.**
- **Support students to transfer their prior college credits.**

Reference

STEP_NZV: Near-zero variance filter. RDocumentation. (n.d.). Retrieved November 15, 2022, from https://www.rdocumentation.org/packages/recipes/versions/1.0.1/topics/step_nzv

Centering numeric data - step_center. - step_center • recipes. (n.d.). Retrieved November 15, 2022, from https://recipes.tidymodels.org/reference/step_center.html

Scaling numeric data - step_scale. - step_scale • recipes. (n.d.). Retrieved November 15, 2022, from https://recipes.tidymodels.org/reference/step_scale.html

Bravo, H. C. (2020, April 26). *Lecture notes: Introduction to data science*. 21 Exploratory Data Analysis: Summary Statistics. Retrieved November 15, 2022, from https://www.hcbravo.org/IntroDataSci/bookdown-notes/exploratory-data-analysis-summary-statistics.html#spread

Quantifying health. QUANTIFYING HEALTH. (n.d.). Retrieved November 15, 2022, from https://quantifyinghealth.com/stepwise-selection/

THANK YOU