Predicting Student Success & GPA

Adithya Murali - Eshaan Arora - Timmy Ren - Ramzi Kattan

Table of contents

01

Exploring Our Data

Problem Definition

Comprehensive information on 2392 high school students (data obfuscated for confidentiality)

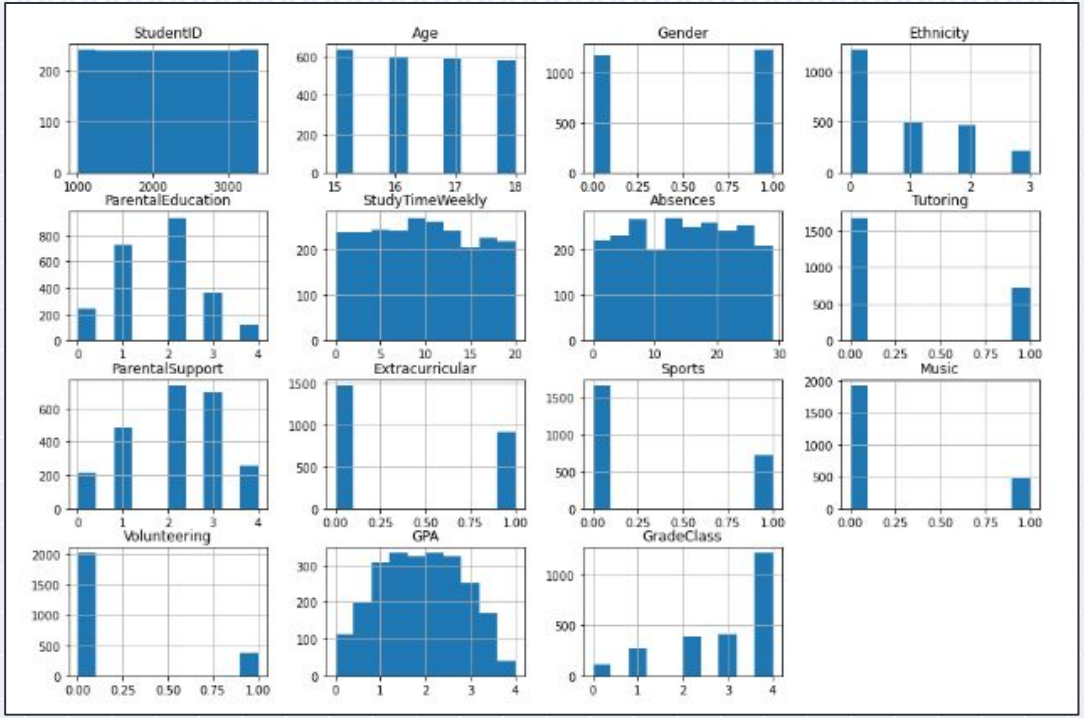
How can the information contained within this dataset assist us in the following? -

1. Predicting an individual student's GPA based on external factors
2. Predicting which students are likely to pass and which are likely to fail their academic course load
3. Providing assistance to at-risk students

Note that this dataset contains information regarding:

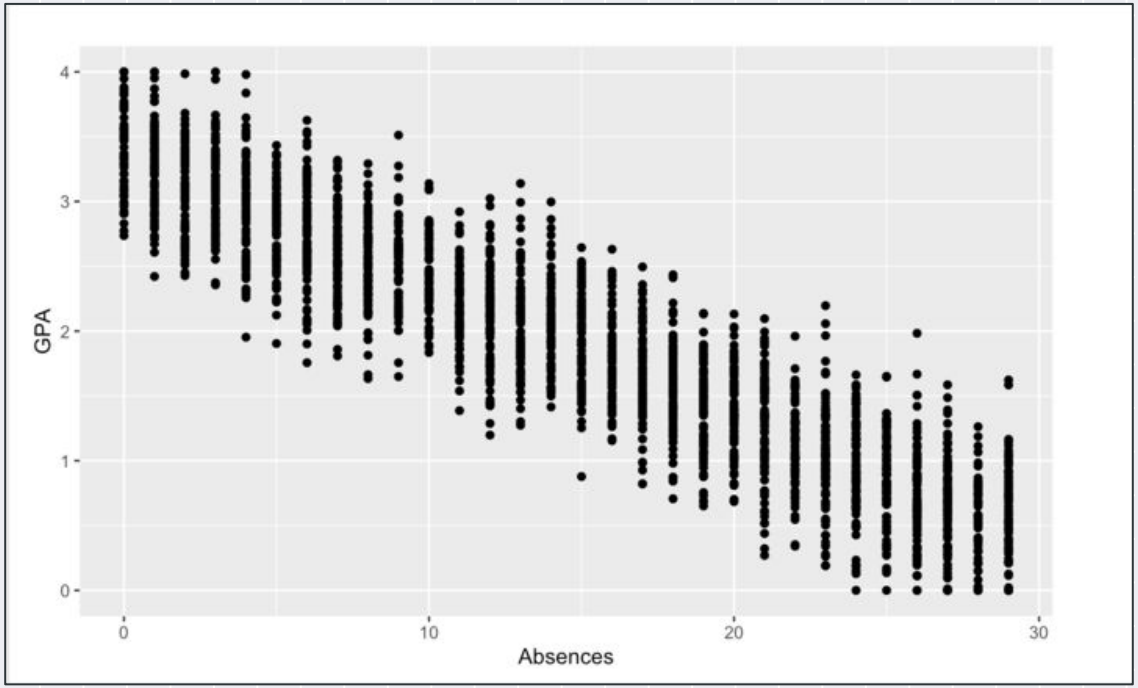
- Demographic Details
- Study Habits
- Parental Involvement
- Extracurriculars
- Academic Performance

Data Cleaning




```
df.isna().sum()

StudentID            0
Age                  0
Gender               0
Ethnicity            0
ParentalEducation    0
StudyTimeWeekly      0
Absences             0
Tutoring             0
ParentalSupport      0
Extracurricular      0
Sports               0
Music                0
Volunteering         0
GPA                  0
```

Absences Correlation



Pass_Fail

Fails (Positive Class)
1274

Pass (Negative Class)
1118

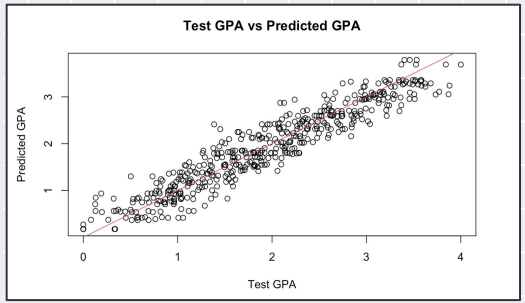
GPA	GradeClass
3.137624	3.0
3.189217	2.0
1.795369	3.0
2.435958	4.0
1.844056	1.0
...	...
3.455509	0.0
3.279150	4.0
1.142333	2.0
1.803297	1.0
2.140014	1.0

02

Regression

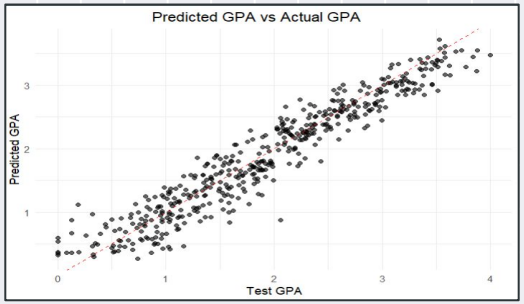
Regression Model Plots & MSE

Regression Tree



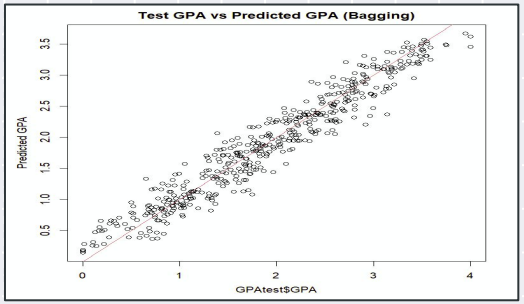
0.08991757

K-Nearest Neighbors



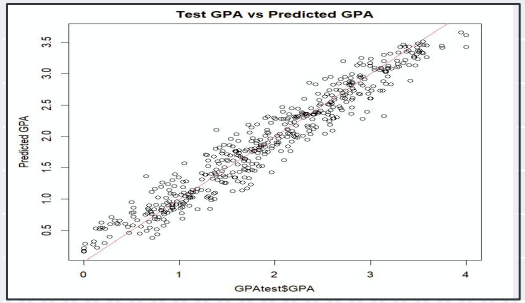
0.07327167

Bagging



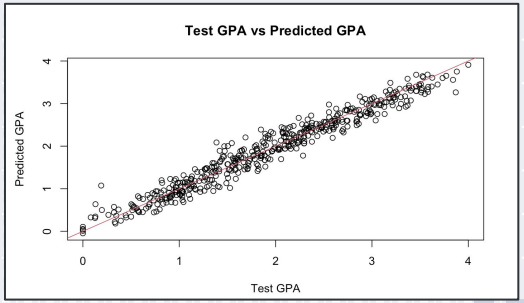
0.05611447

Random Forest



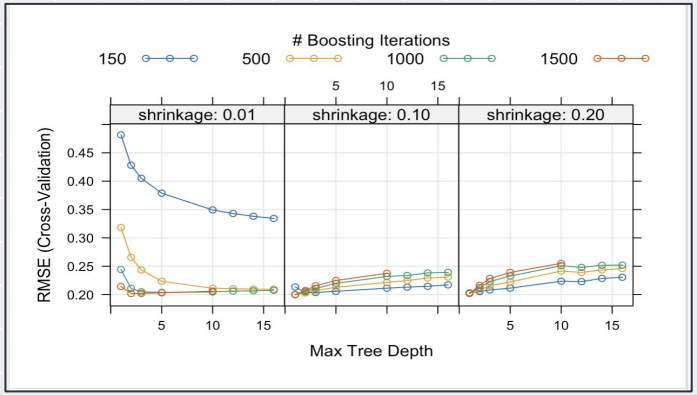
0.053264108

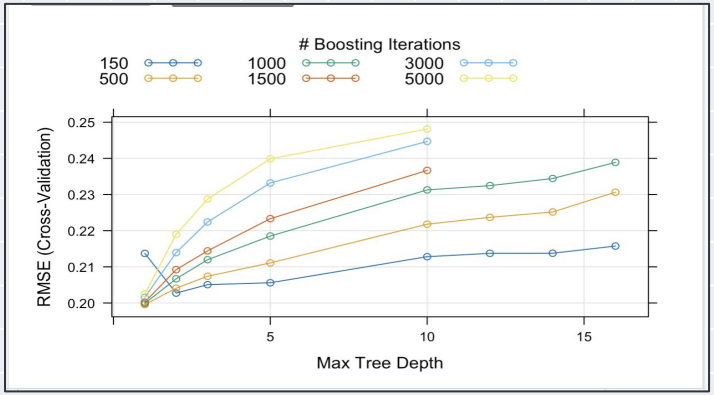
BART

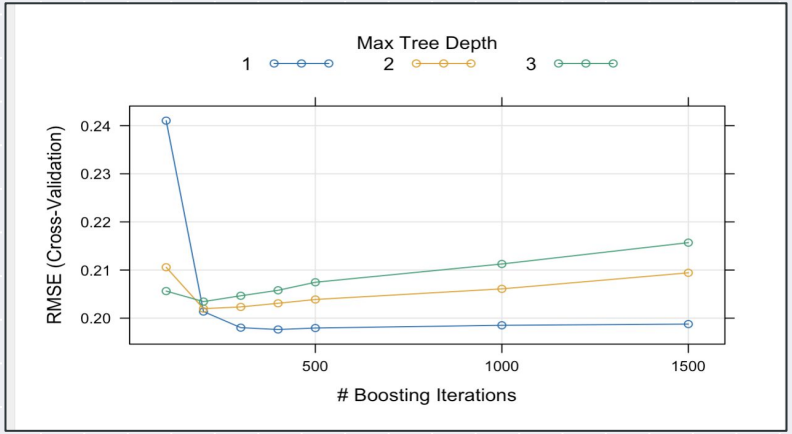


0.03952747

Boosting

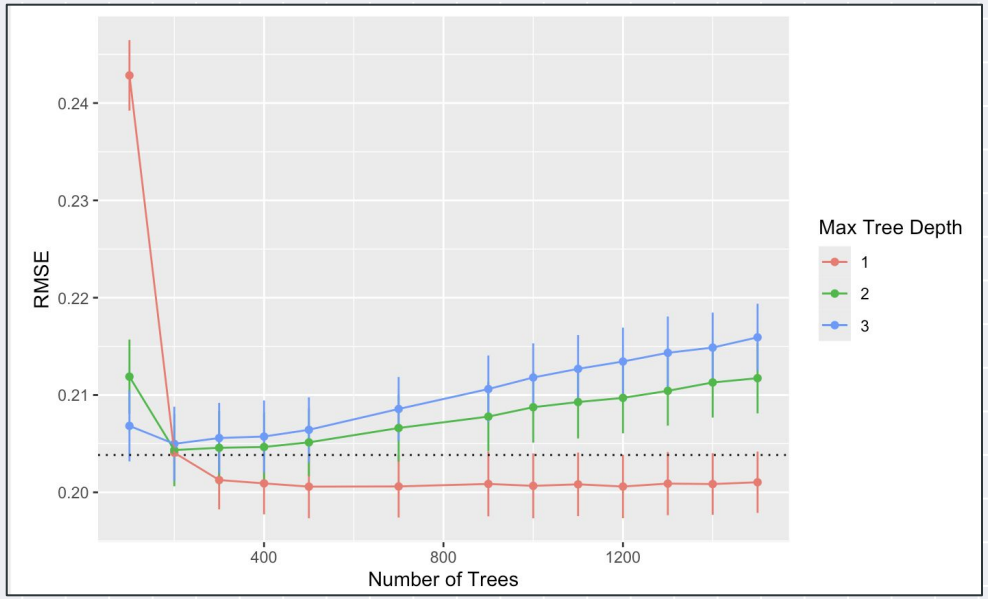

Boosting Iterations
150
500
1000
1500
shrinkage: 0.01
shrinkage: 0.10
shrinkage: 0.20
RMSE (Cross-Validation)
0.45
0.40
0.35
0.30
0.25
0.20
5
10
15
Max Tree Depth


Boosting Iterations
150
500
1000
1500
3000
5000
RMSE (Cross-Validation)
0.25
0.24
0.23
0.22
0.21
0.20
5
10
15
Max Tree Depth


Max Tree Depth
1
2
3
RMSE (Cross-Validation)
0.24
0.23
0.22
0.21
0.20
500
1000
1500
Boosting Iterations

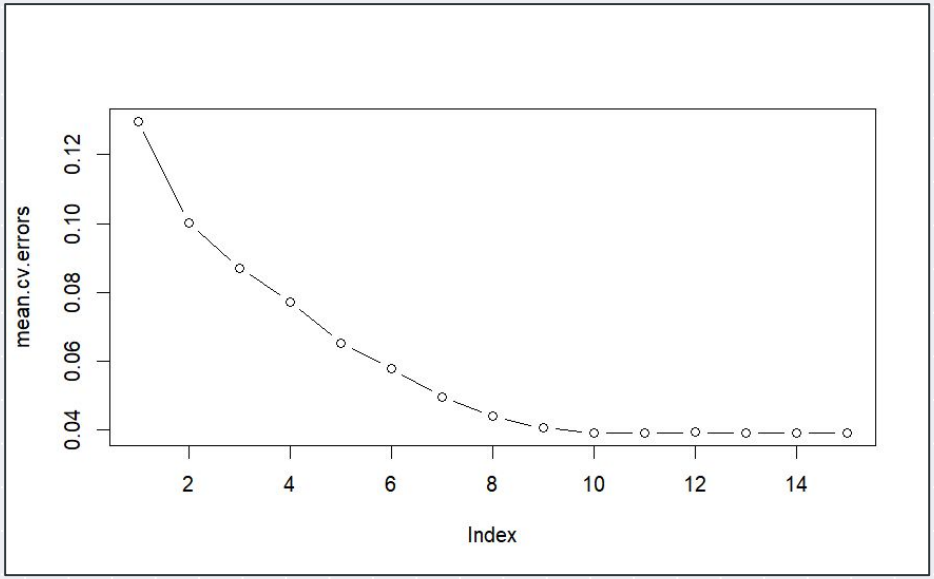
Boosting

# of Trees	300
Interaction Depth	1
Shrinkage	0.1
Min. # of Observations	5



RMSE = 0.1979

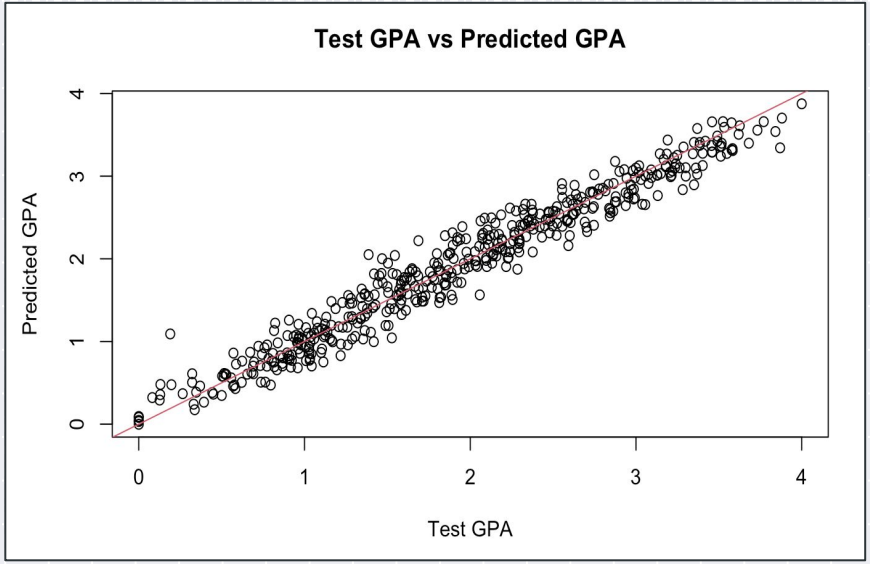
Linear Regression




```
Coefficients:
                     Estimate
(Intercept)         2.5115962
StudyTimeWeekly     0.0285837
Absences           -0.0996945
Tutoring1           0.2492710
ParentalSupport1    0.1667349
ParentalSupport2    0.3068281
ParentalSupport3    0.4616456
ParentalSupport4    0.6171480
Extracurricular1    0.1888150
Sports1             0.1950106
Music1              0.1370141
```

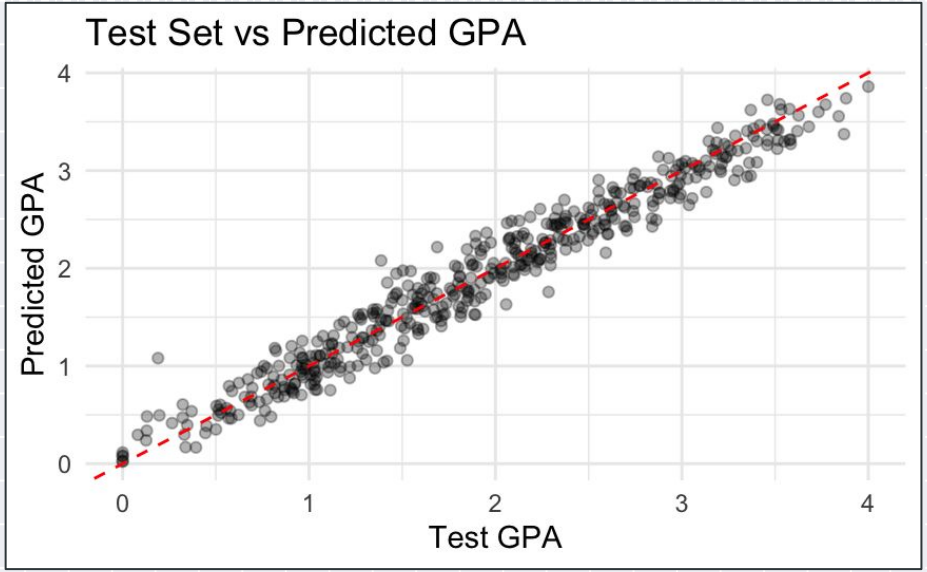
RMSE: 0.1931

Comparison



0.197887

Boosting



0.193078

Linear Regression

Linear Regression

Error Comparison

Linear regression and Boosting RMSEs were comparable

Interpretability

Given that the other models are ensemble methods, it is a lot easier for us to determine the relationships between the parameters and GPA

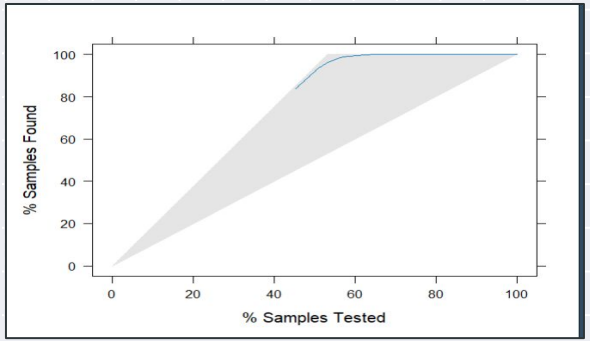
```
Coefficients:
                    Estimate
(Intercept)        2.5115962
StudyTimeWeekly    0.0285837
Absences          -0.0996945
Tutoring1          0.2492710
ParentalSupport1   0.1667349
ParentalSupport2   0.3068281
ParentalSupport3   0.4616456
ParentalSupport4   0.6171480
Extracurricular1   0.1888150
Sports1            0.1950106
Music1             0.1370141
```

03

Classification

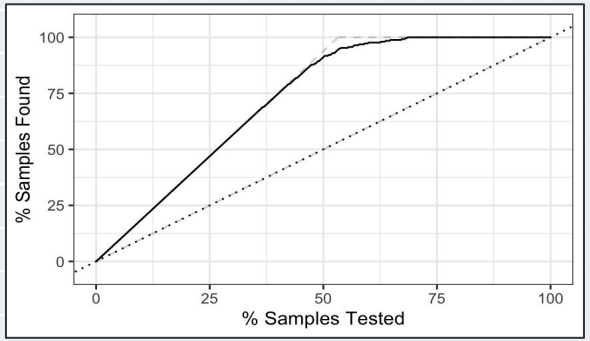
Classification Models Plots & Samples Needed for 95%

K N N



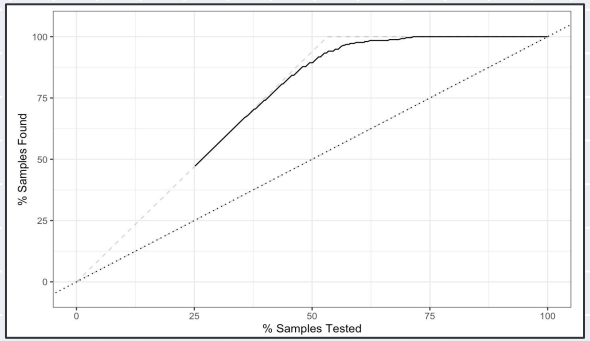
249

Logistic Regression



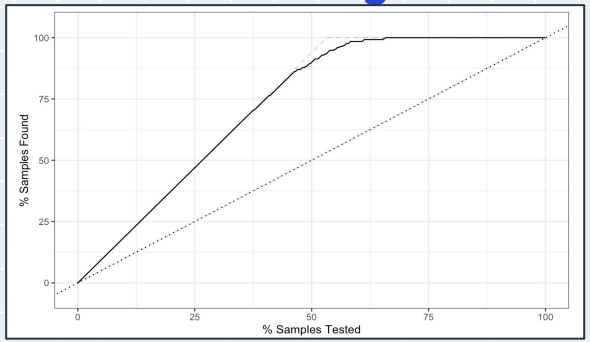
257

Bagging



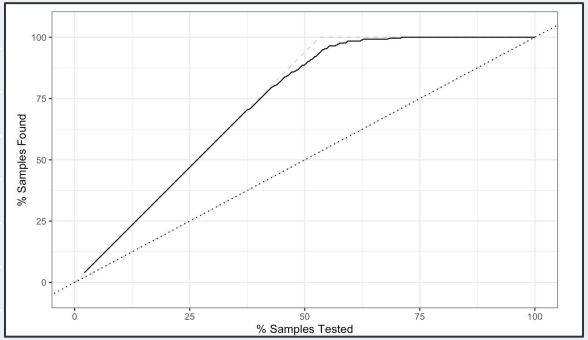
259

Boosting



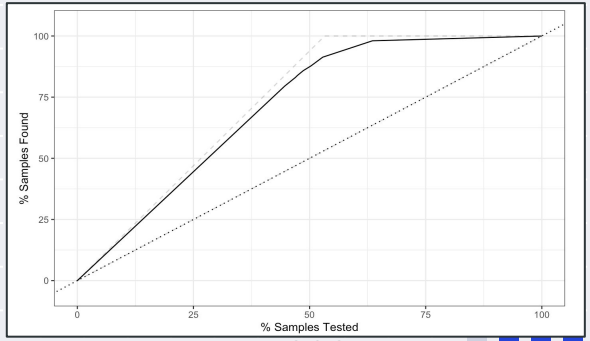
262

Random Forest



266

Decision Tree



288

K - Nearest Neighbors

Error Comparison

K-nearest neighbors had the best lift, on the test set it reached the 95% threshold with only 8 negative cases (241 positive cases)

Interpretability

We decided to focus less on interpretability because the regression model can give an intuitive understanding. We focused more on predictive power.

04

Recommendation

Takeaways

Student Success Indicators (SSIs)

High parental support, study time weekly, & few absences are some of the best indicators of a high GPA student.

Accurate Predictor of Pass / Fail

Enables the school to target specific at-risk students by boosting their SSIs.

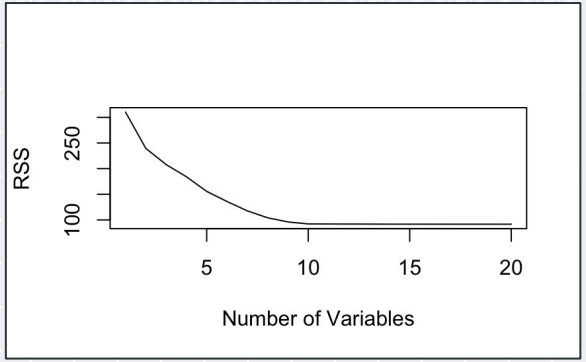
95% At-Risk Student Capture

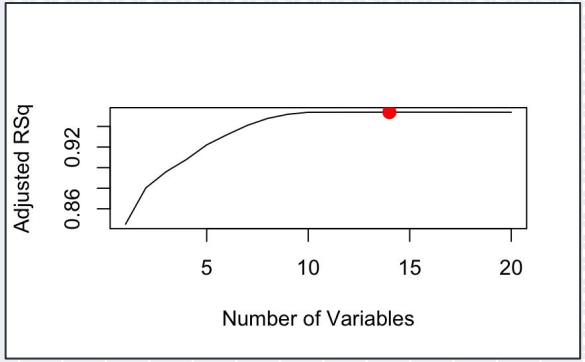
Emphasis on capturing 95% of the at-risk students while minimizing cost of intervention (tutoring, parent-teacher meetings, etc.)

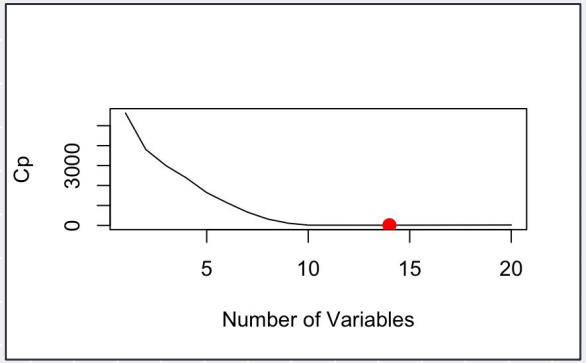
Any Questions?

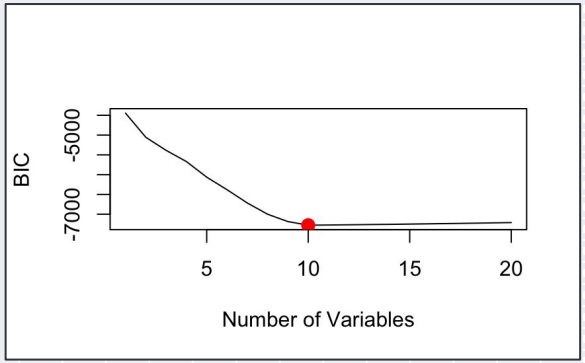
Appendix

Best Subset Linear Regression









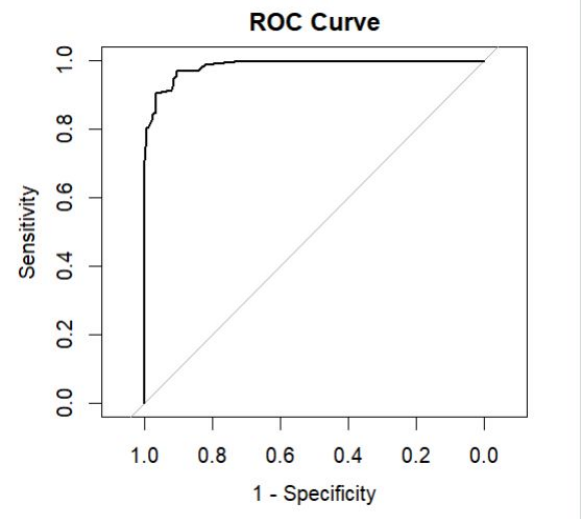
Linear Regression - Best Subset

```
Call:
lm(formula = GPA ~ StudyTimeWeekly + Absences + Tutoring + ParentalSupport +
    Extracurricular + Sports + Music, data = gpa_train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.61593 -0.14018  0.00227  0.14730  0.61468

Coefficients:
                   Estimate Std. Error  t value Pr(>|t|)
(Intercept)       2.5115962  0.0197121  127.414   <2e-16
StudyTimeWeekly   0.0285837  0.0008057   35.475   <2e-16
Absences         -0.0996945  0.0005339 -186.717   <2e-16
Tutoring1         0.2492710  0.0099190   25.131   <2e-16
ParentalSupport1  0.1667349  0.0183611    9.081   <2e-16
ParentalSupport2  0.3068281  0.0173145   17.721   <2e-16
ParentalSupport3  0.4616456  0.0174438   26.465   <2e-16
ParentalSupport4  0.6171480  0.0206805   29.842   <2e-16
Extracurricular1  0.1888150  0.0092992   20.305   <2e-16
Sports1           0.1950106  0.0097753   19.949   <2e-16
Music1            0.1370141  0.0114844   11.930   <2e-16
```

KNN Performance Metricas

```
          Reference
Prediction   0   1
         0 237   6
         1  17 217
```

```
               Accuracy : 0.9518
                 95% CI : (0.9285, 0.9692)
    No Information Rate : 0.5325
    P-Value [Acc > NIR] : < 2e-16
```




```
            Sensitivity : 0.9331
            Specificity : 0.9731
         Pos Pred Value : 0.9753
         Neg Pred Value : 0.9274
             Prevalence : 0.5325
         Detection Rate : 0.4969
   Detection Prevalence : 0.5094
      Balanced Accuracy : 0.9531
```

"AUC: 0.983536951378835"