STUDENT SUCCESS DATA ANALYSIS

Jason Antal, Grace Lin, Sarah Dominguez,
Quinlan O’Connell, Sam Chen, Cole Brown

ROADMAP

01 DESCRIPTION OF PROJECT GOALS

02 EXPLORATORY ANALYSIS

03 SOLUTION AND INSIGHTS

04 LINEAR, LASSO, RIDGE

05 KNN

06 BAGGING AND BOOSTING

07 RANDOM FORESTS

08 RESULTS

DESCRIPTION OF PROJECT GOALS

Description:

- Student performances
- Best predictors for success is school

Importance of the problem:

- Quality of school can only go so far
- Knowing what can help
- Schools, Businesses, and Individuals
- Outliers

Data Context

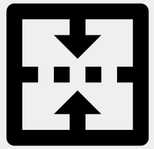
- Target Variable: final_grade
- 1044 rows, 40 columns

- Data stems from students' grades in a math class and a language class
- Portuguese Grade Scale (0-20)
  - 0-9 -> Fail
  - 10-13 -> Passing, sufficient
  - 14-15 -> Passing, good
  - 16-17 -> Passing, very good
  - 18-20 -> Passing, excellent

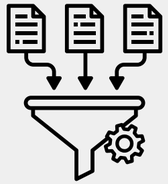
Predictors

- Sex
- Age
- Family Status
- Parent's Education
- Parent's Occupation
- Study Time
- Failures
- Activities
- Extra Paid Classes
- Higher Education Desire
- Internet Access
- Romantic Relationship
- Free Time
- Go Out
- Weekday/End Alcohol
- Health
- Absences
- G1 -> First Period Grade (dropped)
- G2 -> Second Period Grade (dropped)

Data Preparation



- Merged two data files with data from each individual course



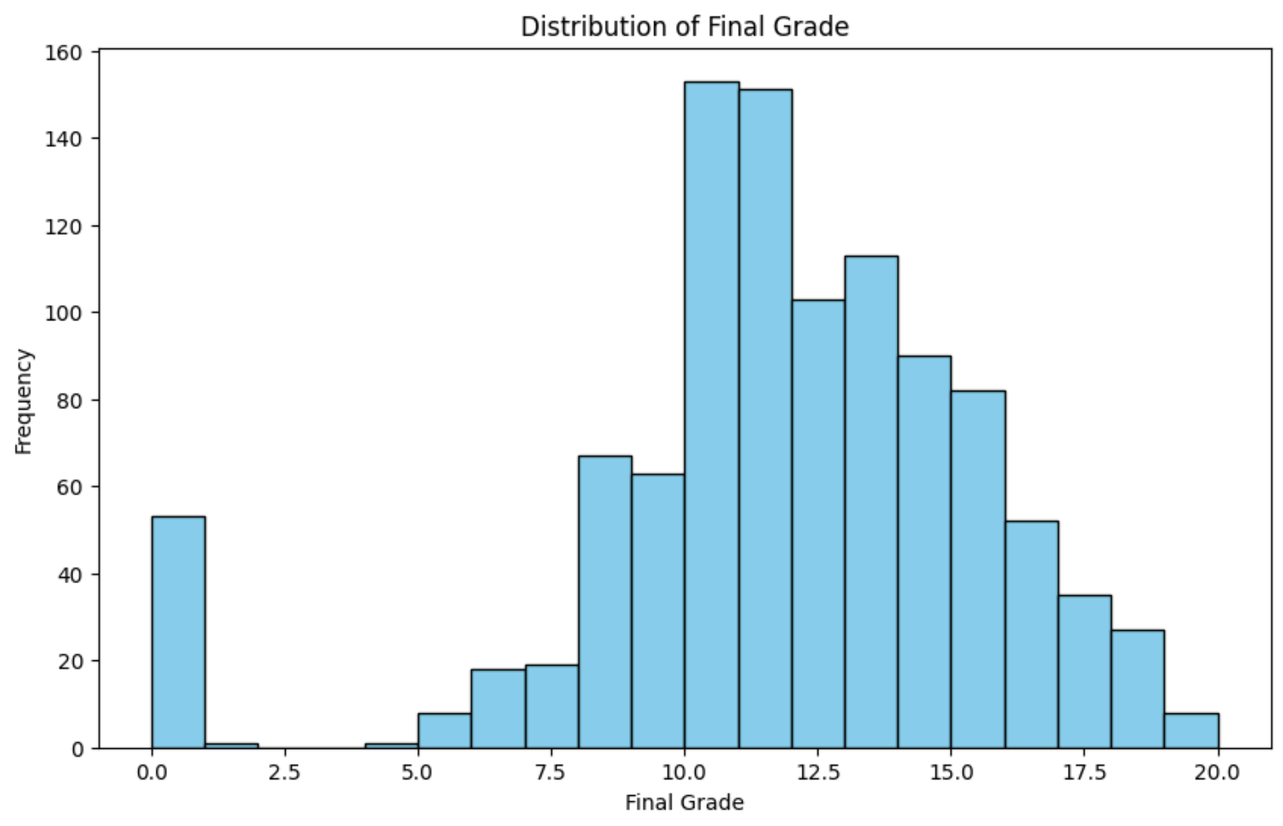
- Dropped explicit predictors G1 & G2 to make predictions more difficult
- Dummy encoded categorical variables
  - School (two schools)
  - Parent’s Occupation
  - Guardian
  - Extra Classes
  - Internet

Predictors

- Sex
- Age
- Family Status
- Parent’s Education
- Parent’s Occupation
- Study Time
- Failures
- Activities
- Extra Paid Classes
- Higher Education Desire
- Internet Access
- Romantic Relationship
- Free Time
- Go Out
- Weekday/End Alcohol
- Health
- Absences
- G1 -> First Period Grade (dropped)
- G2 -> Second Period Grade (dropped)

EXPLORATORY DATA ANALYSIS

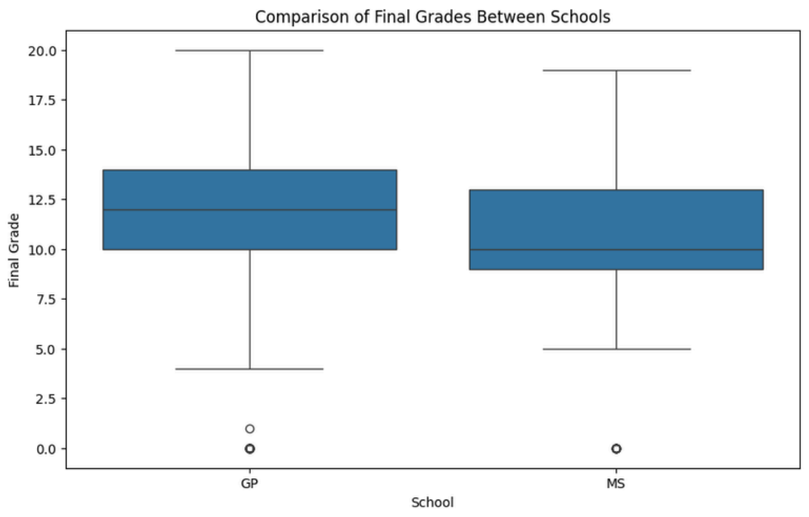
Target Variable Distribution



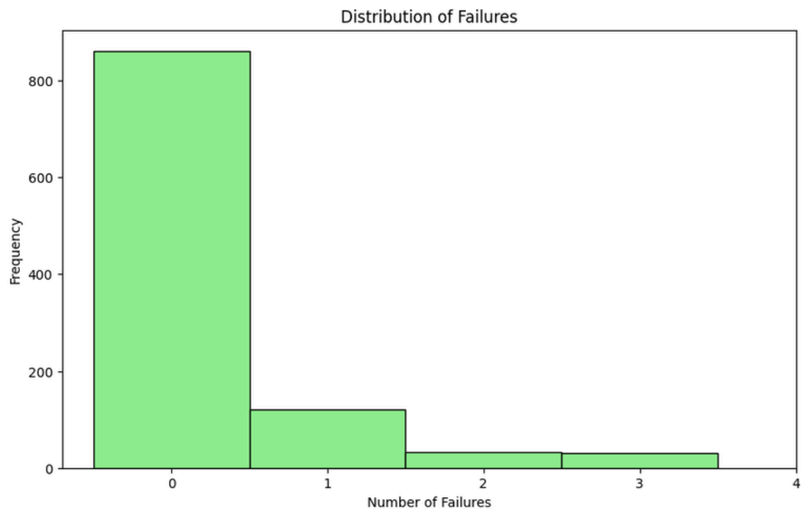
- Majority of students obtain passing final grades, around a 10-15
- Decent population who fails class with a grade of 0.0
- Normal Distribution

Other Distributions

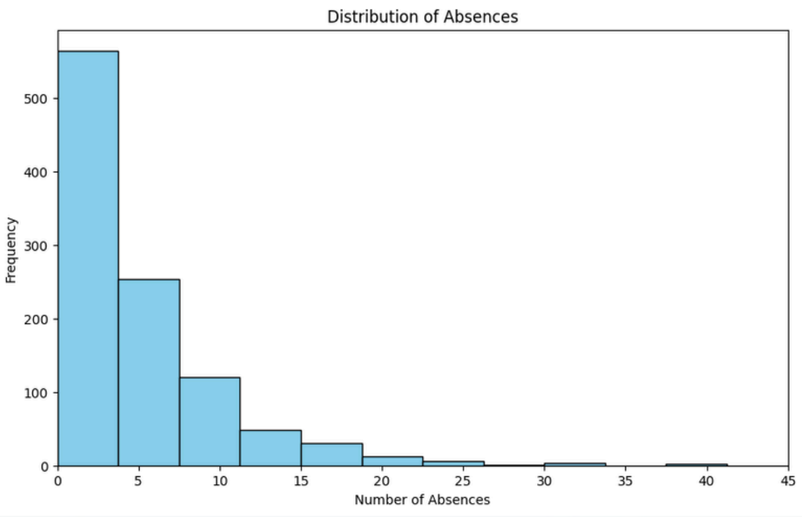
School vs. School



Visualizing Failures

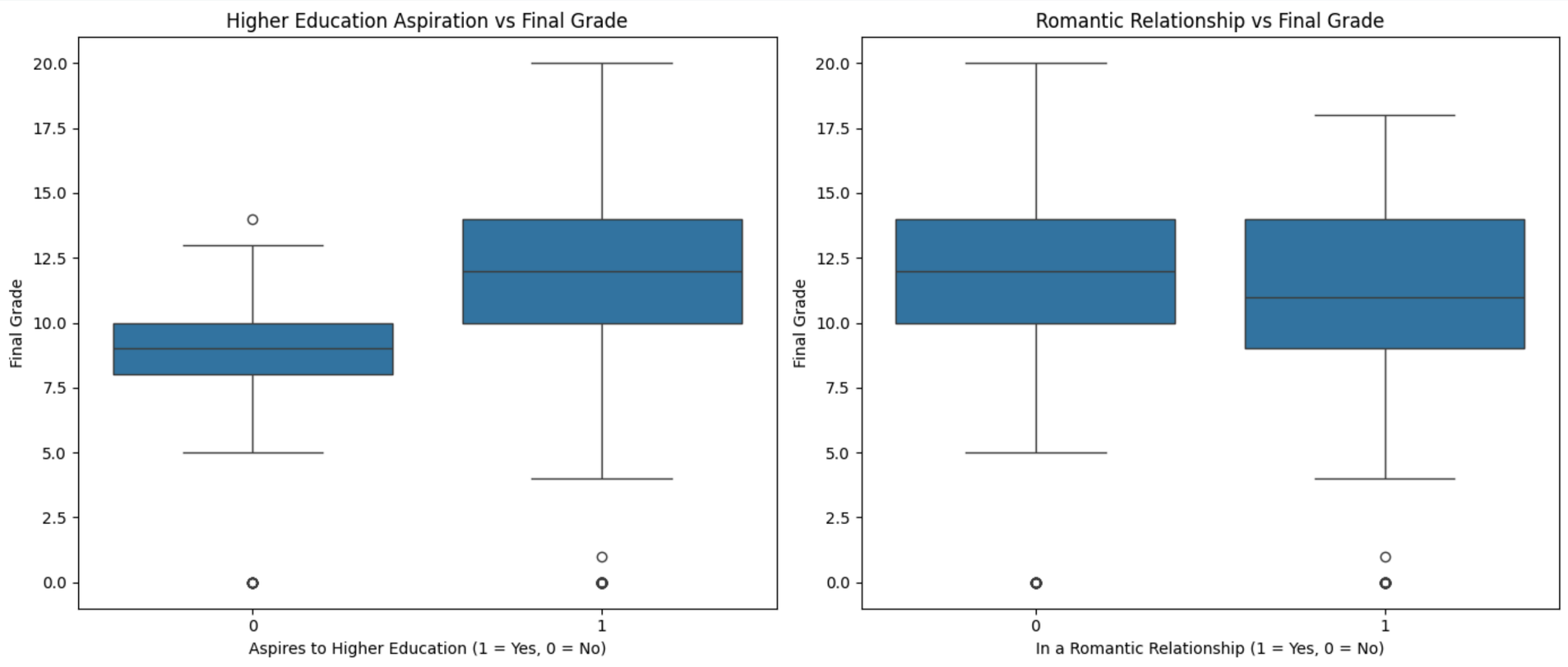


Visualizing Absences



- Most students have never failed a course
- Majority of students fall between 0-10 absences
- Few outliers in absences and failures
- GP school tends to have a higher mean grade, but larger range of grades

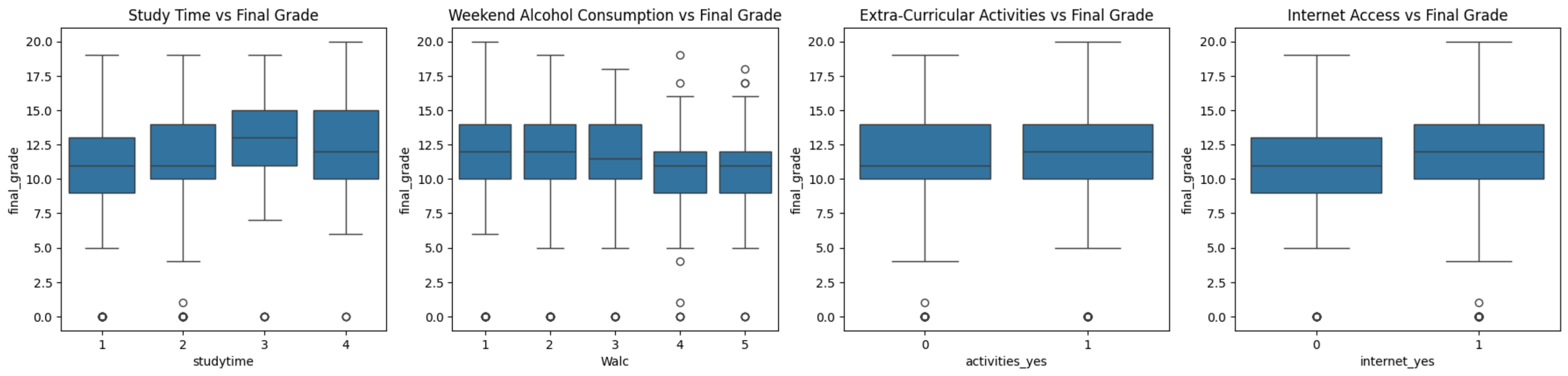
Impactful Predictors



- Aspiration to achieve a higher education leads to a **large** jump in final grade prediction -> harder working students !

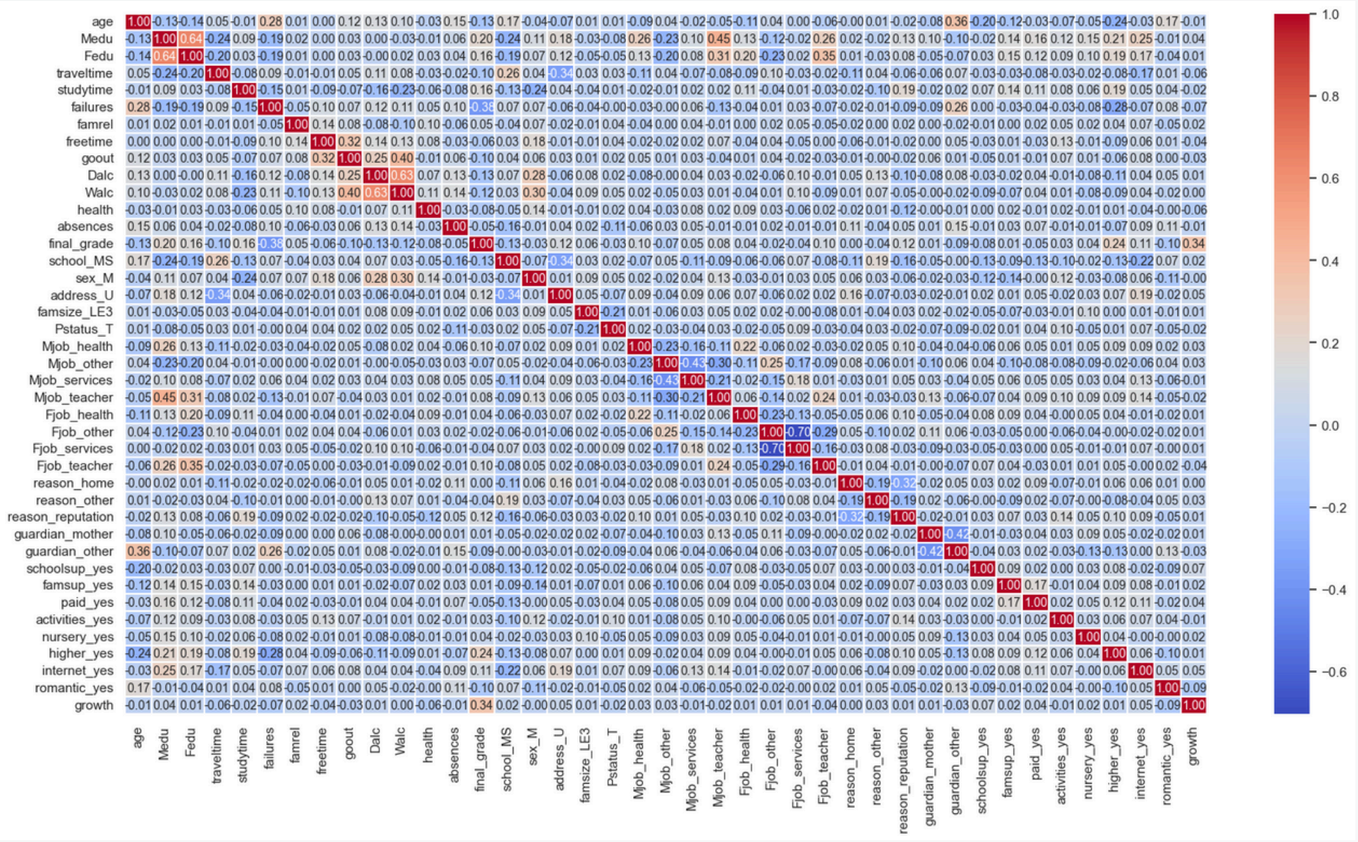
- Being in a relationship slightly decreases grade prediction?

Impactful Predictors



- More study time -> tends to lead to higher grades on average
- Final grade shifts downwards at alcohol consumption of 4 and 5
- Students who participate in extra curriculars have a slightly higher mean final grade
- As expected, internet access has a positive affect on performance

Correlation Matrix



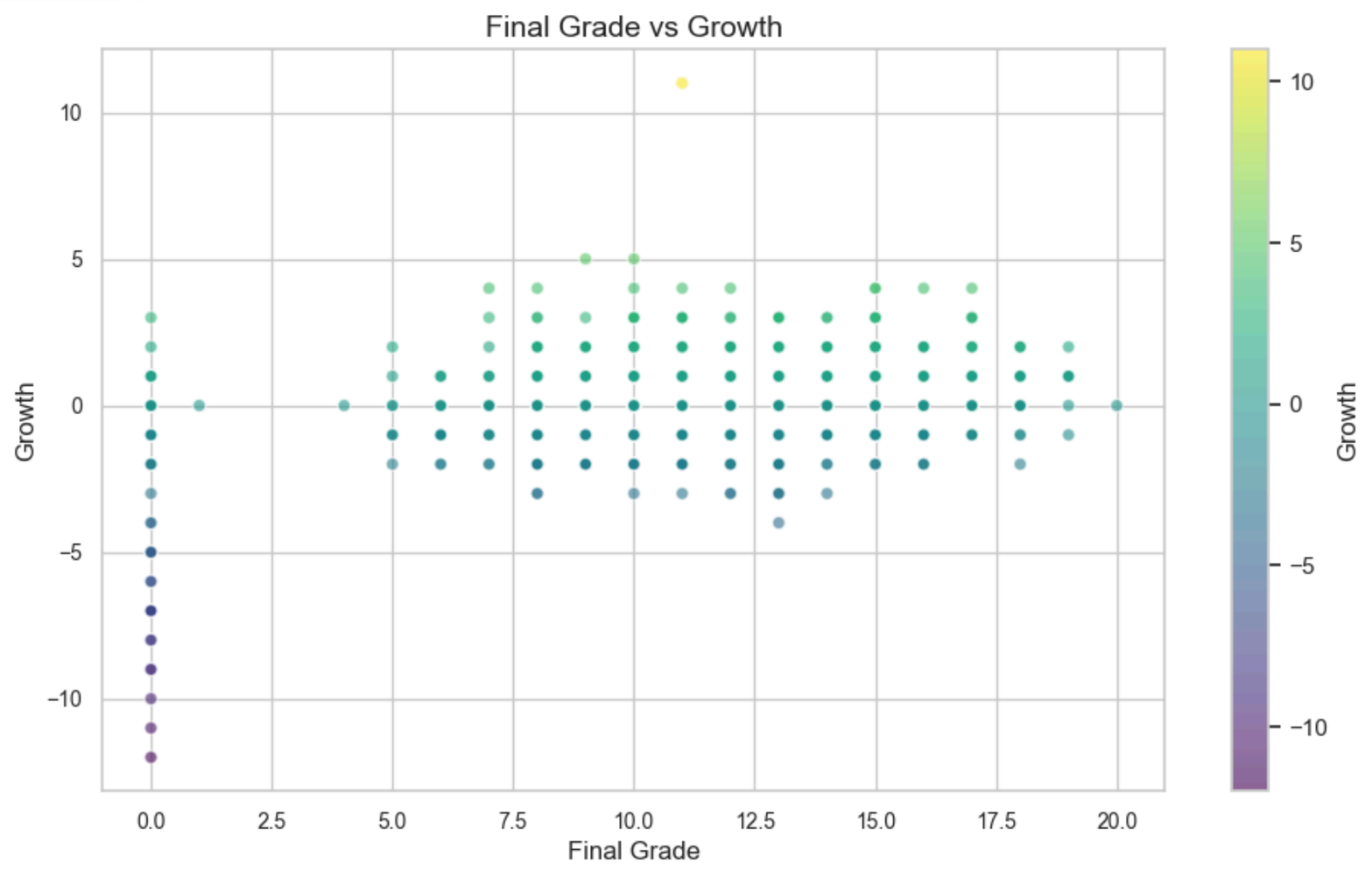
EXTREMELY hard to interpret correlation matrix because of number of predictors

	final_grade
age	0.125282
Medu	0.201472
Fedu	0.159796
traveltime	0.102627
studytime	0.161629
failures	0.383145
Dalc	0.129642
Walc	0.115740
school_MS	0.127114
address_U	0.117696
Mjob_health	0.101349
Fjob_teacher	0.101361
reason_reput…	0.121303
higher_yes	0.236578
internet_yes	0.107064
growth	0.341705

Sorted to find high correlations with final grade

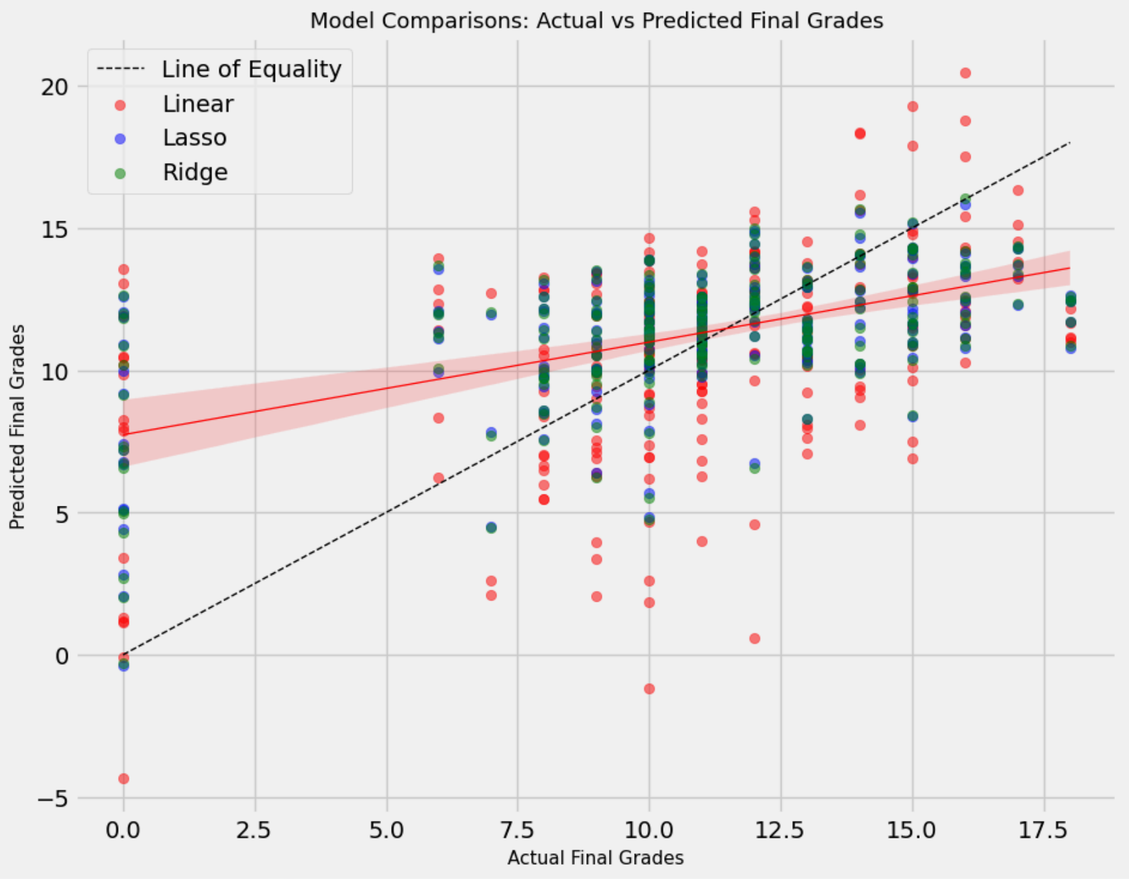
Feature Engineering

Growth: The difference between the 2nd period and the 1st period semester grades



METHODS

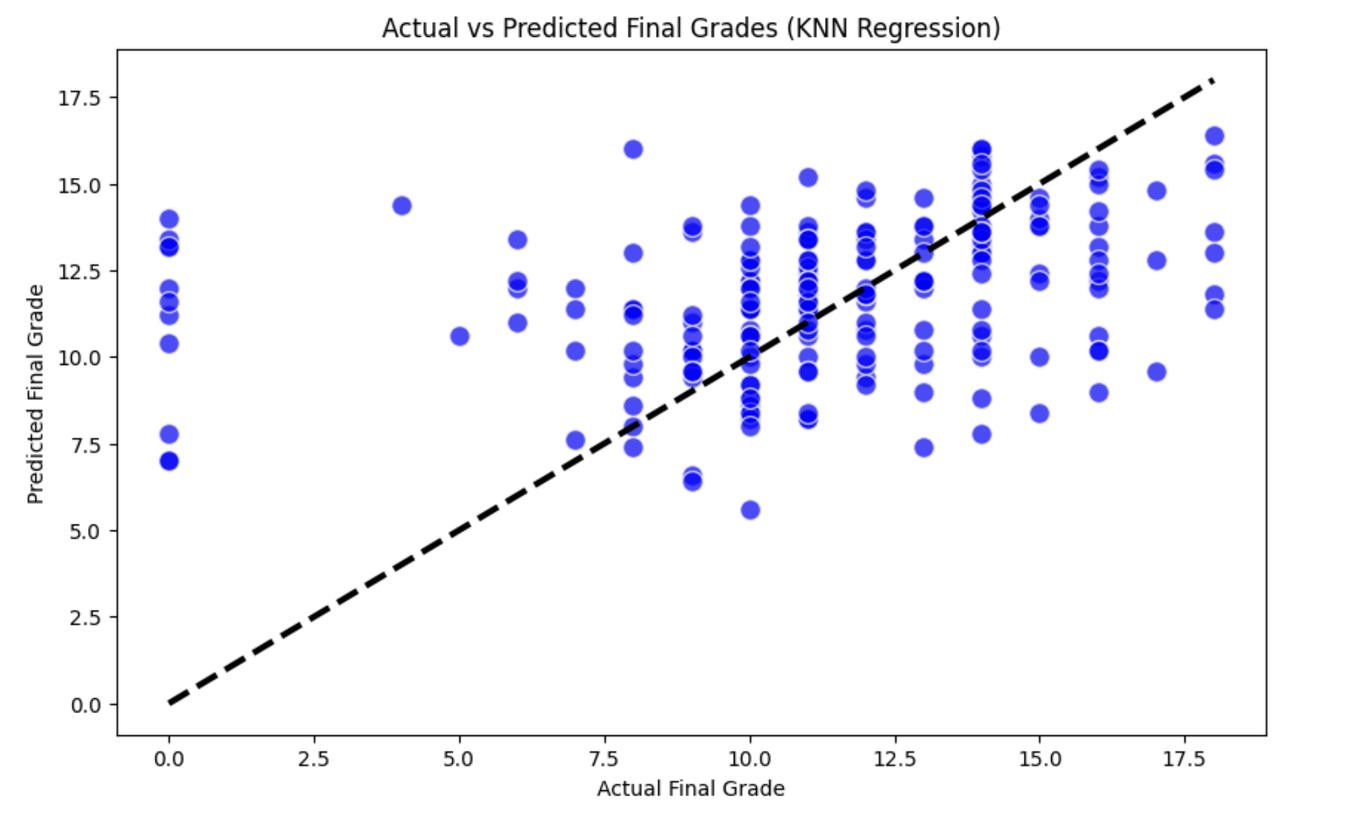
LINEAR, LASSO, RIDGE



		RMSE
Linear		3.422
Lasso	**alpha = 0.01**	3.421
Ridge	**alpha = 0.01**	3.422

Feature	Importance
growth	1.199394
higher_yes	0.406322
studytime	0.292265
Fjob_teacher	0.260954
Medu	0.217811

KNN MODEL

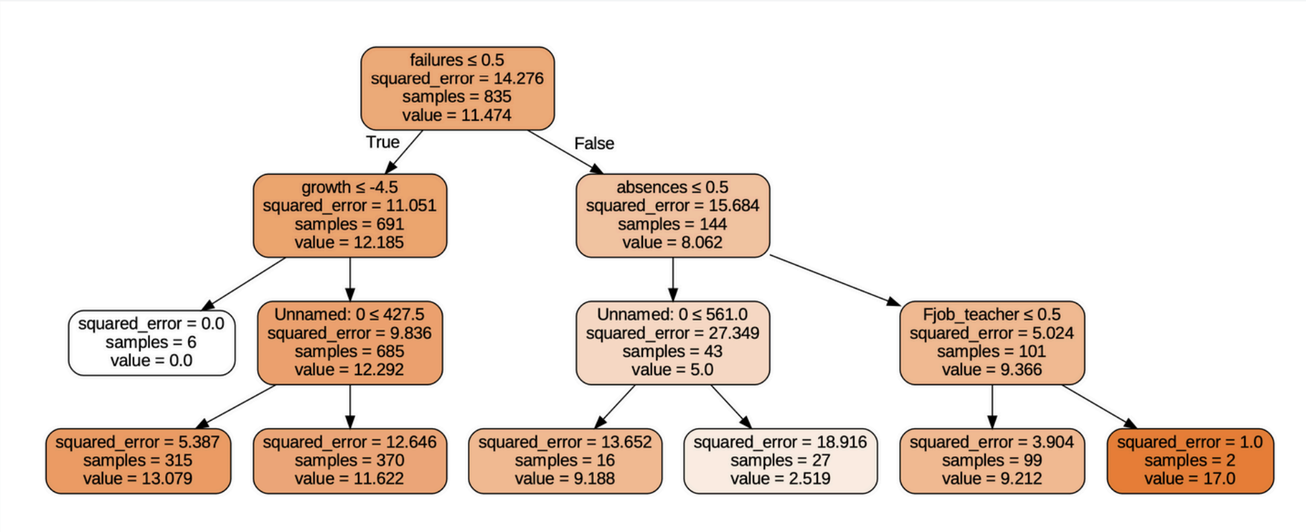


Mean Squared Error (MSE): 11.744

Root Mean Squared Error (RMSE): 3.427

R-Squared: 0.409

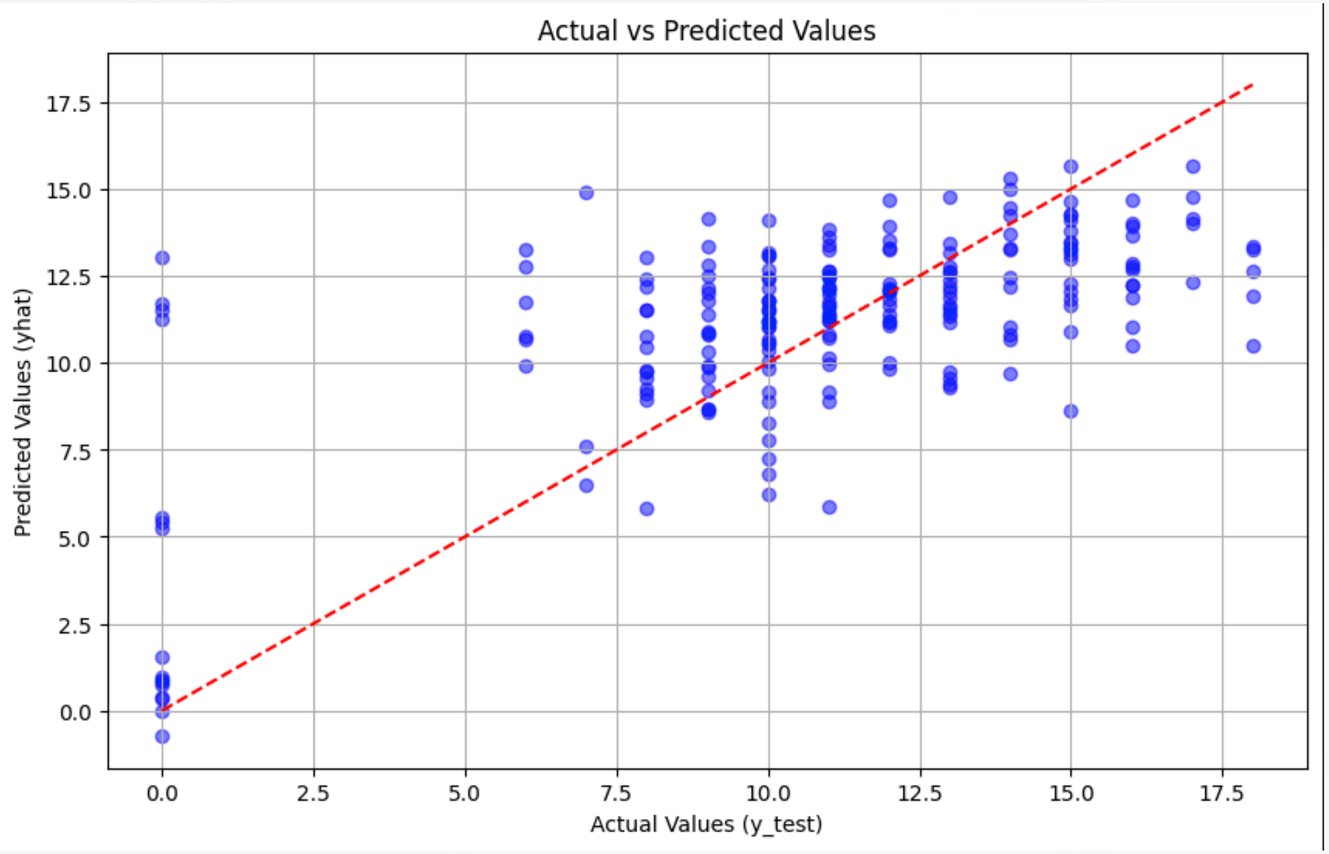
REGRESSION TREE



	Feature	Importance
5	failures	0.372397
39	growth	0.243460
12	absences	0.106371
1	Medu	0.067008
24	Fjob_services	0.064785

- Fitting an initial tree of size = 5, yields an RMSE of 3.35
- Trimming down the tree down to a size of 3 yields a lower RMSE of aroun 3.30

GRADIENT BOOST



- RMSE after conducting grid search: 3.095
- R-squared: 0.441

	Feature	Importance
5	failures	0.413103
12	absences	0.129897
31	schoolsup_yes	0.082456
36	higher_yes	0.046717
7	freetime	0.038209

BAGGING MODEL

Model Performance:

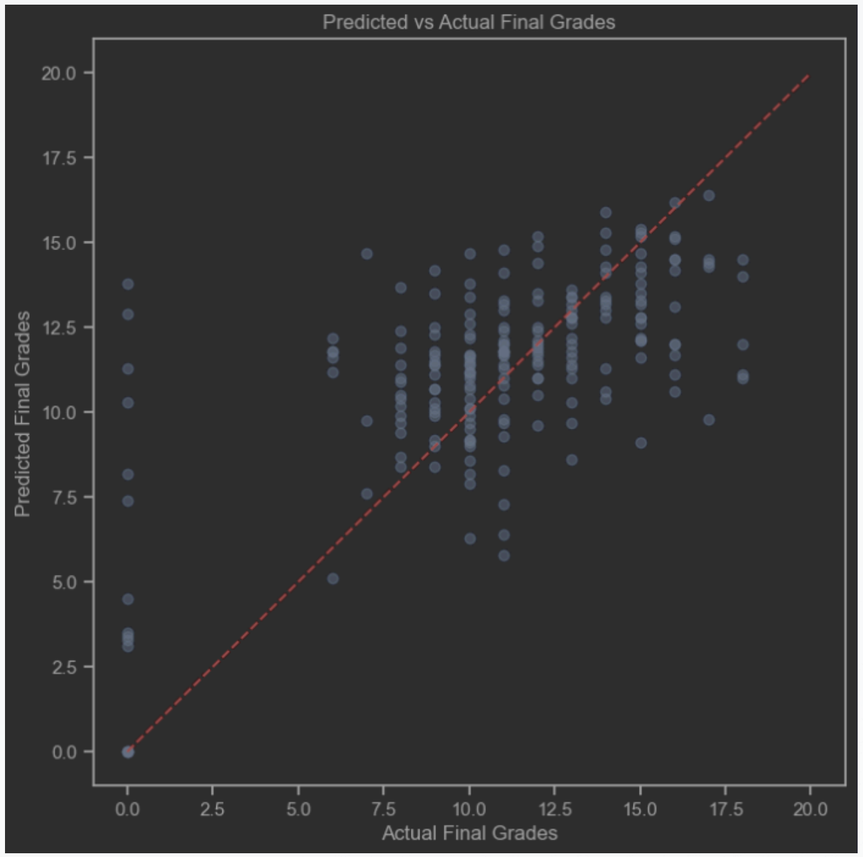
RMSE: 3.098

R-Squared: 0.440

Feature Imp.

failures: 0.1744
growth: 0.1582
absences: 0.0964
freetime: 0.0374
goout: 0.0357
schoolsup_yes: 0.0354

Predicted vs. Actual



RANDOM FOREST

Model Performance:

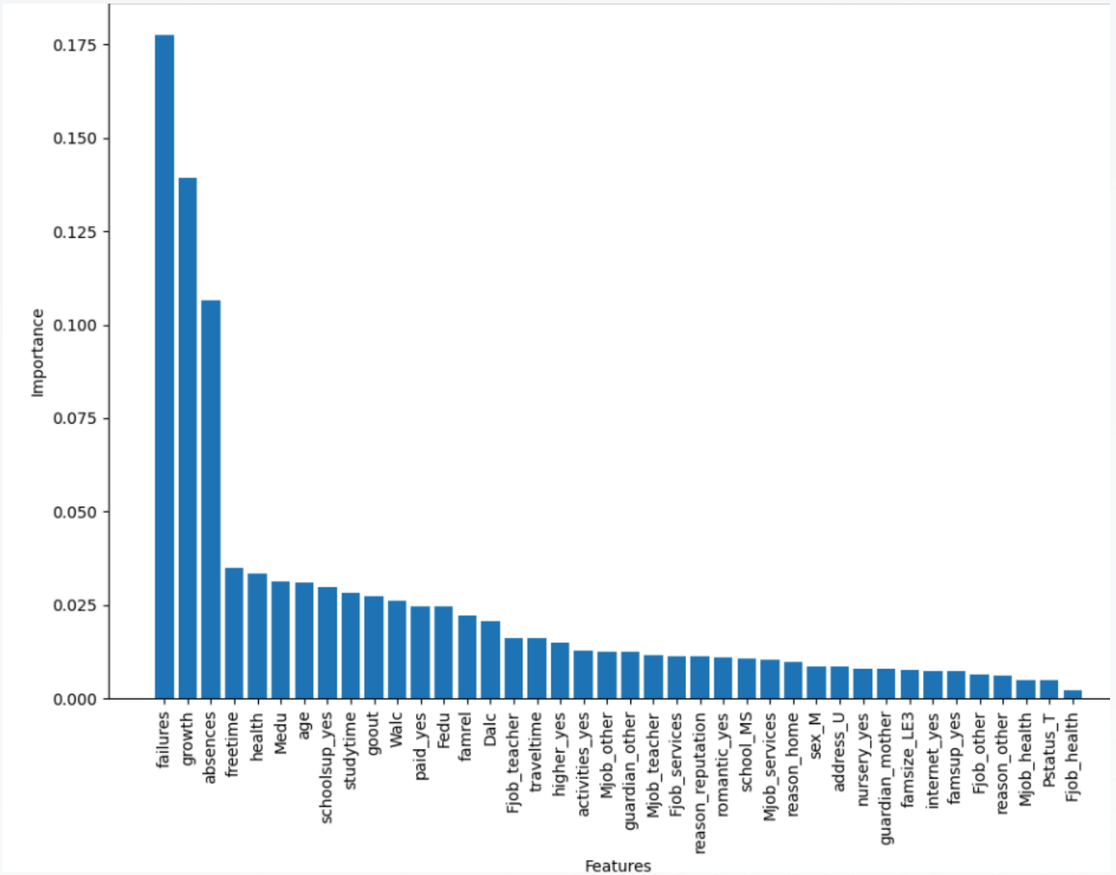
RMSE: 3.084

R-Squared: 0.446

Feature Imp.

Failures: 0.177
Growth: 0.139
Absences: 0.106
Freetime: 0.035
Health: 0.033
Medu: 0.031

Strongest predictors



Model Comparison

Model Type	RMSE	R-Squared
Random Forest	3.084	0.446
Boosting	3.095	0.441
Bagging	3.098	0.440
KNN	3.427	0.409
Linear, Lasso, Ridge Regression	Linear: 3.422, Lasso: 3.421, Ridge: 3.422	Linear: 0.318, Lasso: 0.318, Ridge: 0.318

ADDITIONAL ANALYSIS

The data includes two variables relating to a students grade in the 1st third and 2nd third of the course

Wanted to explore the impact of explicitly including the grade

Many schools collect grades at the “midpoint” and throughout the term

ADDITIONAL ANALYSIS

Addition of G1:

Model Type	RMSE	R^2
Boosting	2.48	0.64
Random Forest	2.431	0.655
Bagging	2.606	0.604

Variable Importance:

```
G1: 0.6657
absences: 0.0690
health: 0.0203
goout: 0.0185
failures: 0.0165
activities_yes: 0.0159
```

Addition of G1 + G2:

Model Type	RMSE	R^2
Boosting	1.66	0.84
Random Forest	1.561	0.858
Bagging	1.661	0.839

Variable Importance:

```
G2: 0.8131
absences: 0.0531
health: 0.0134
studytime: 0.0100
G1: 0.0096
Dalc: 0.0079
```

CONCLUSION

Best Performing Model without Semester Grade Data: Random Forest
Best Performing Model *with* Semester Grade Data: Random Forest

- Schools can monitor student's past failures and absences to get a good read on how they might be performing
- Schools could utilize mid semester data to make accurate predictions of which students are most in danger of not passing
- Schools could emphasize the importance of higher education to potentially motivate students performance further

THANK YOU!