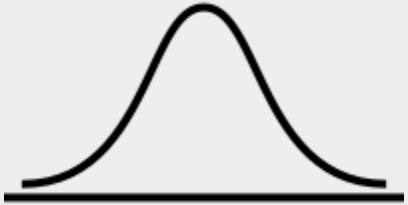
Using Survey Data to Identify Failing Students

Springboard Data Science Capstone Project

Isaac Paulson



I can... (goals for this presentation)

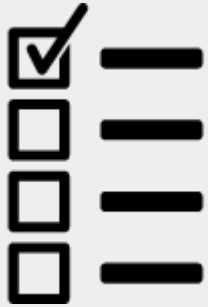
- Communicate data findings
- Understand the data science process
- Use scikit-learn
- Show that I have learned something

Problem:

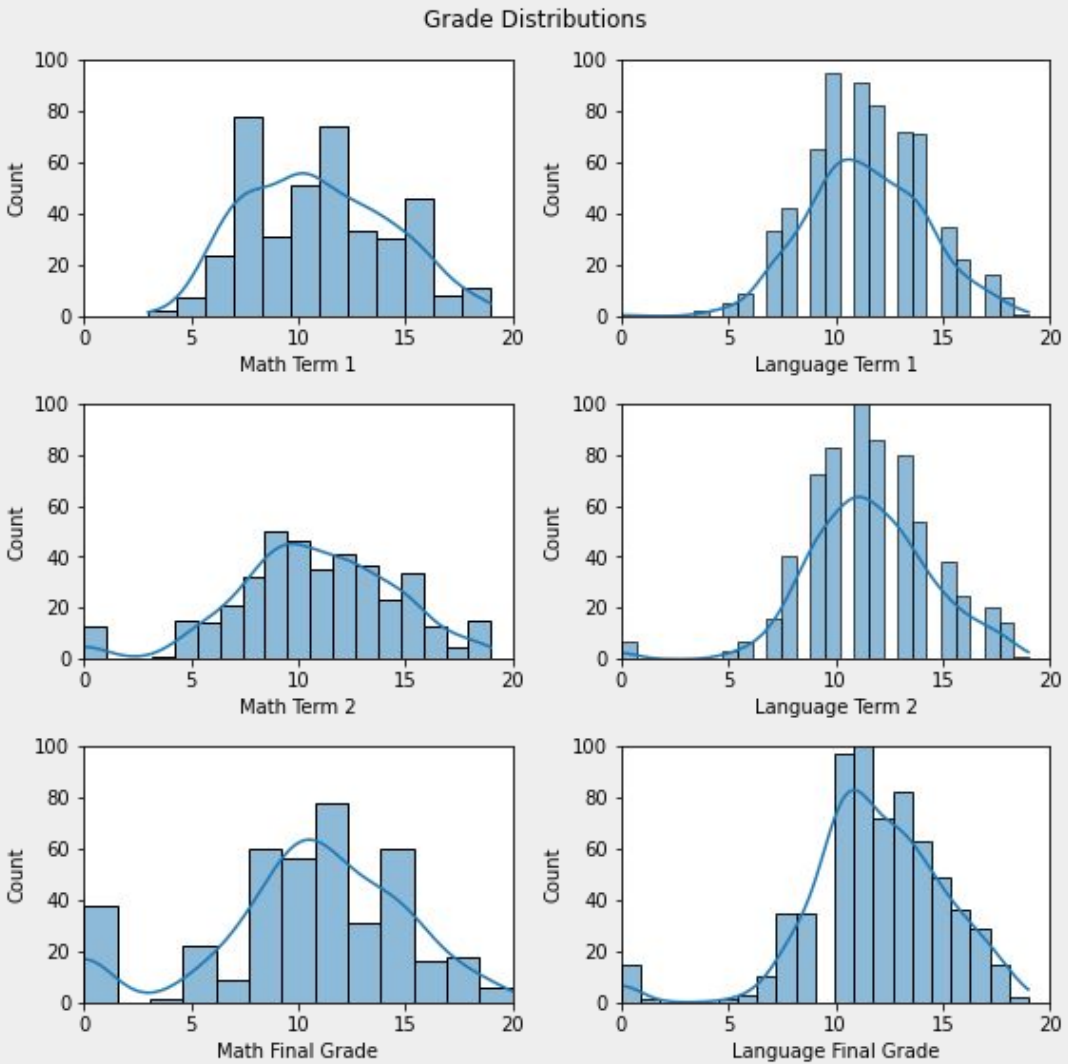
- Students fail
- In educational systems, failure is inefficient
- Teachers think they can identify failing students

The Data:

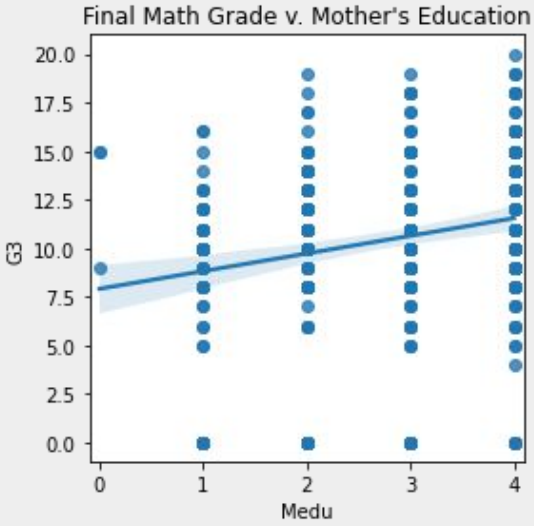
- Math and language grades for Portugeuse students
- Survey questions:
  - Parents' education
  - Study time
  - Other failures
  - Social factors (e.g. dating, alcohol use)

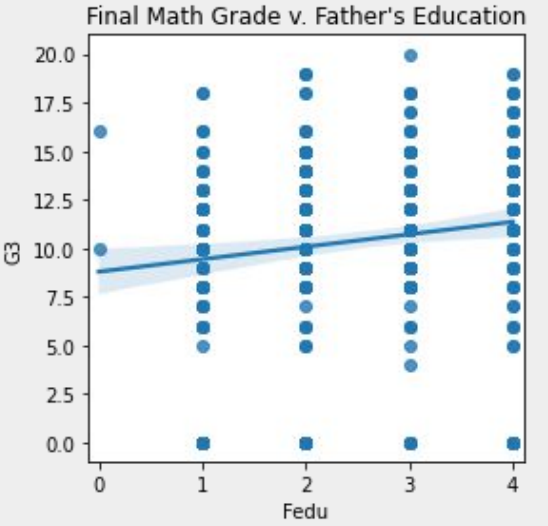


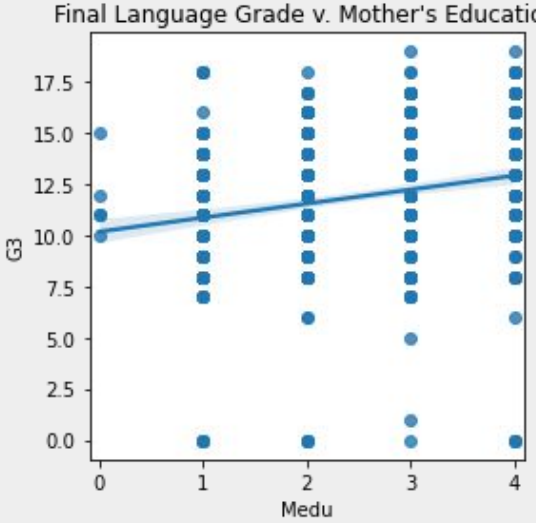
EDA

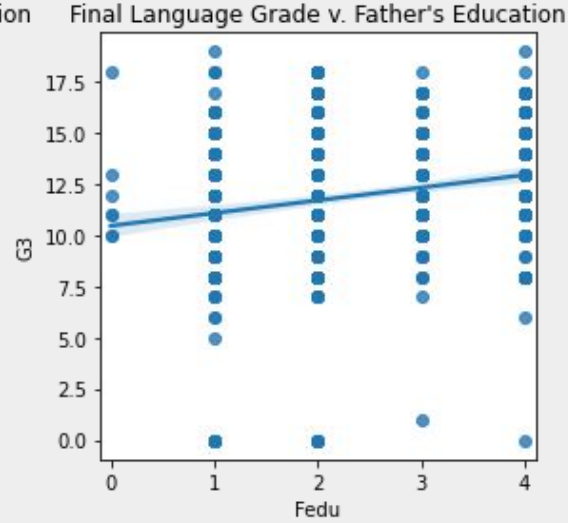

Grade Distributions
Count
Math Term 1
Language Term 1
Math Term 2
Language Term 2
Math Final Grade
Language Final Grade

EDA
(parents' education)

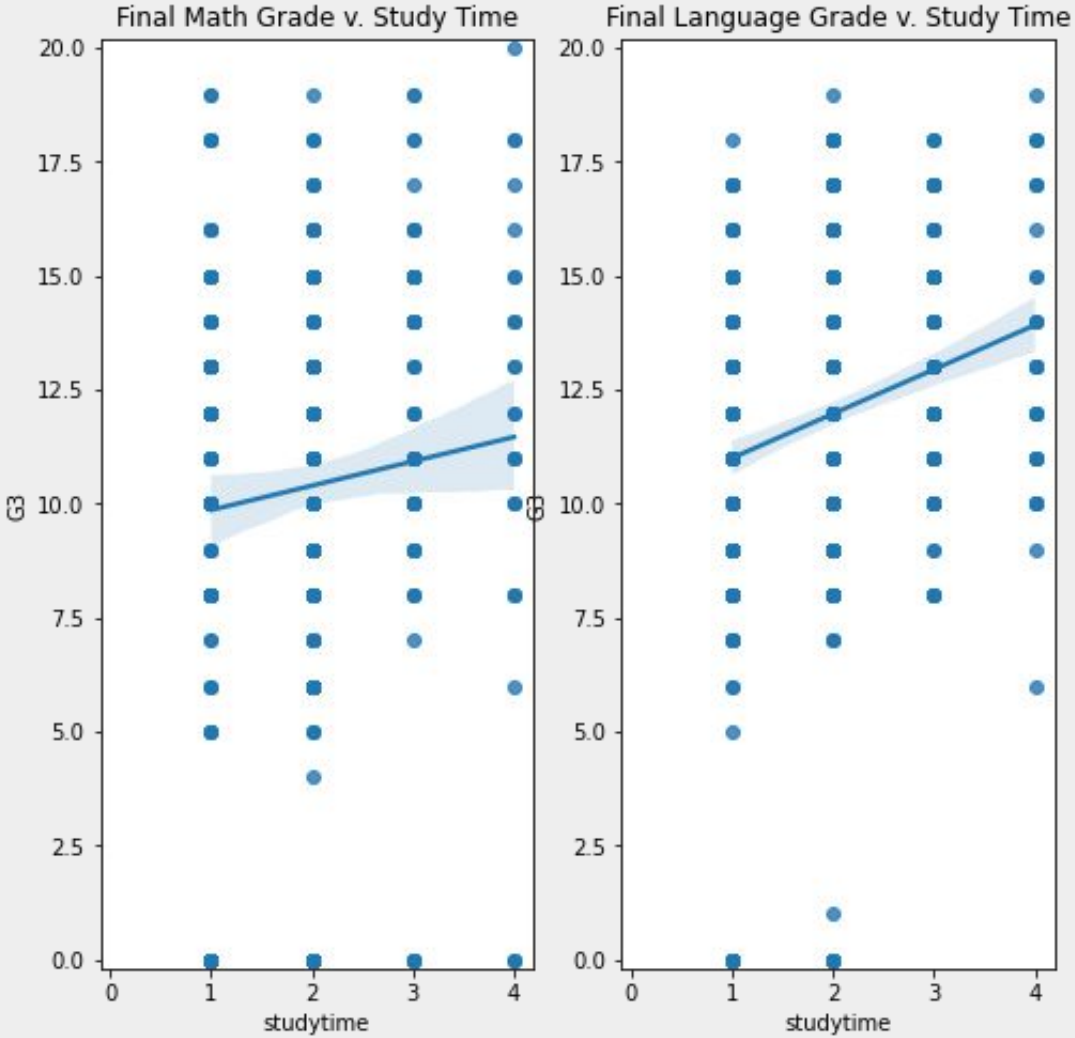




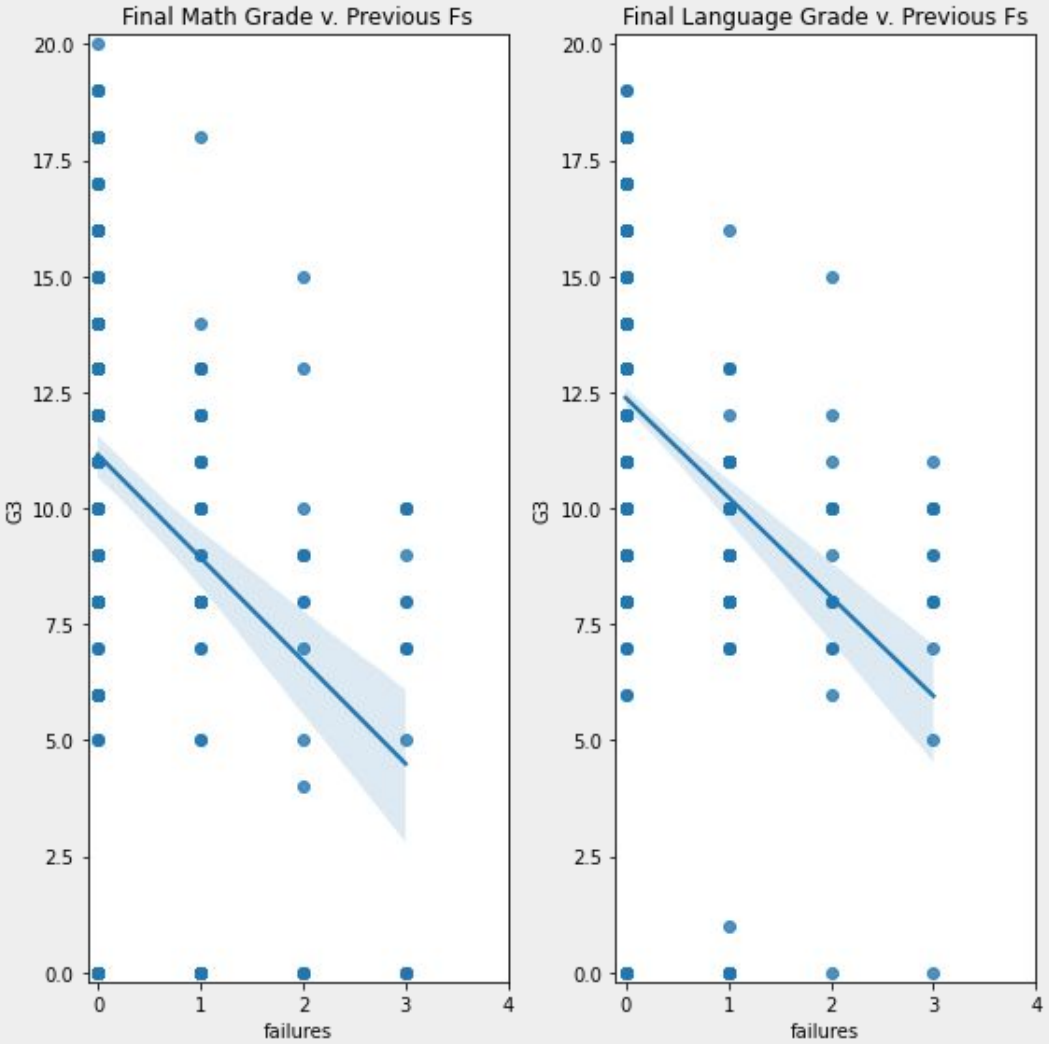




EDA
(Study Time)

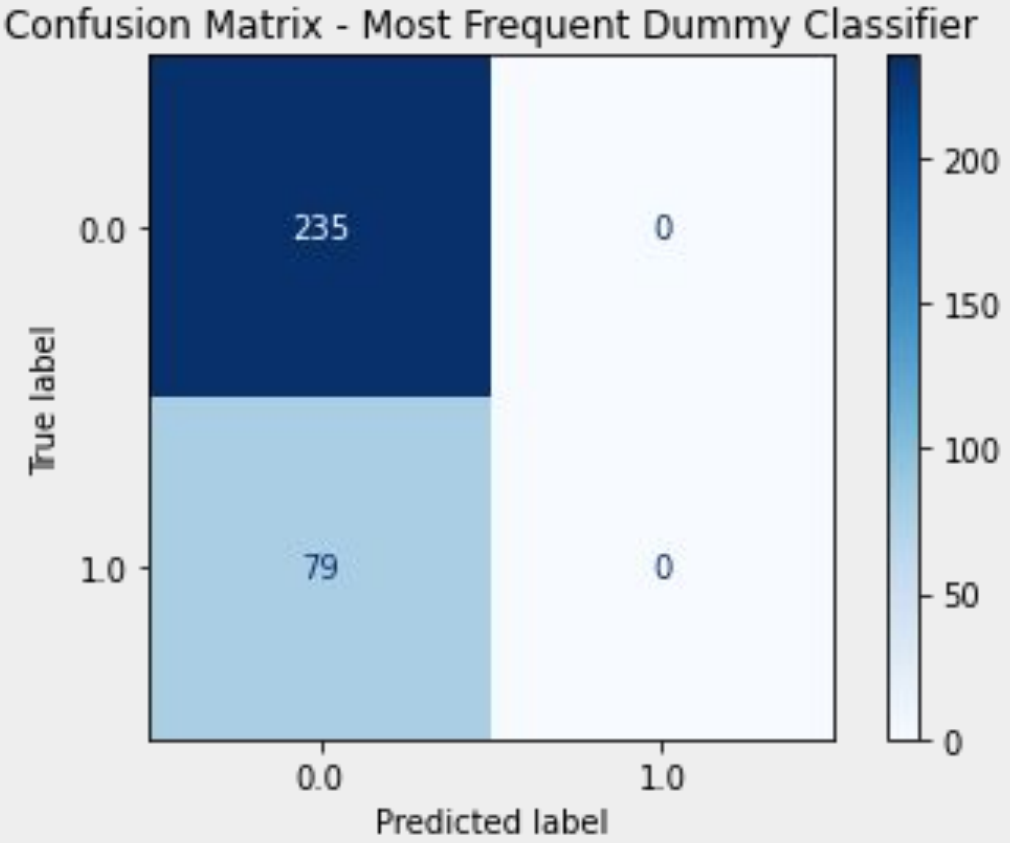


EDA
(Other Failures)

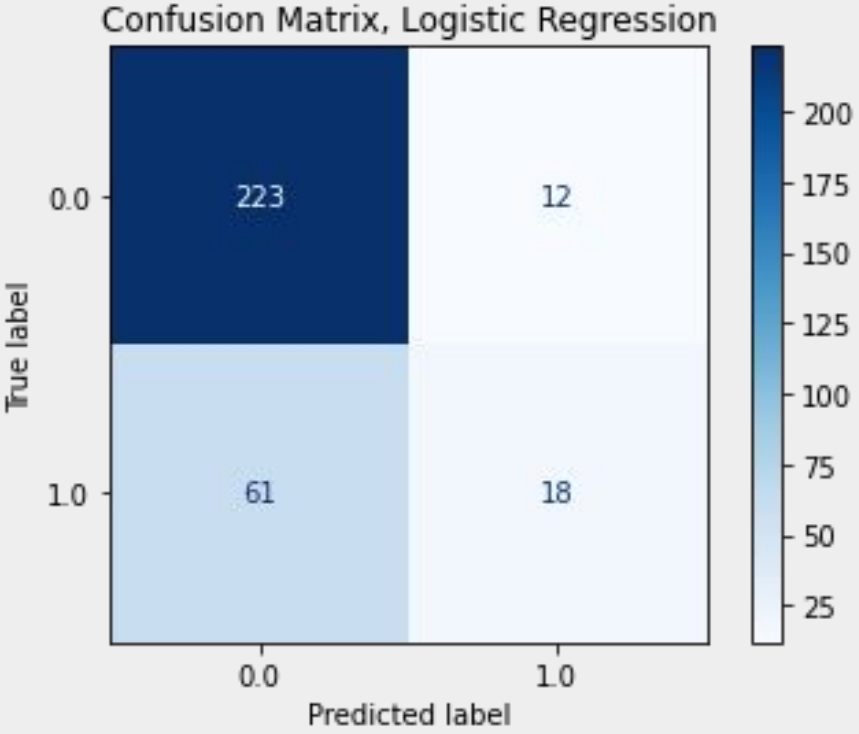


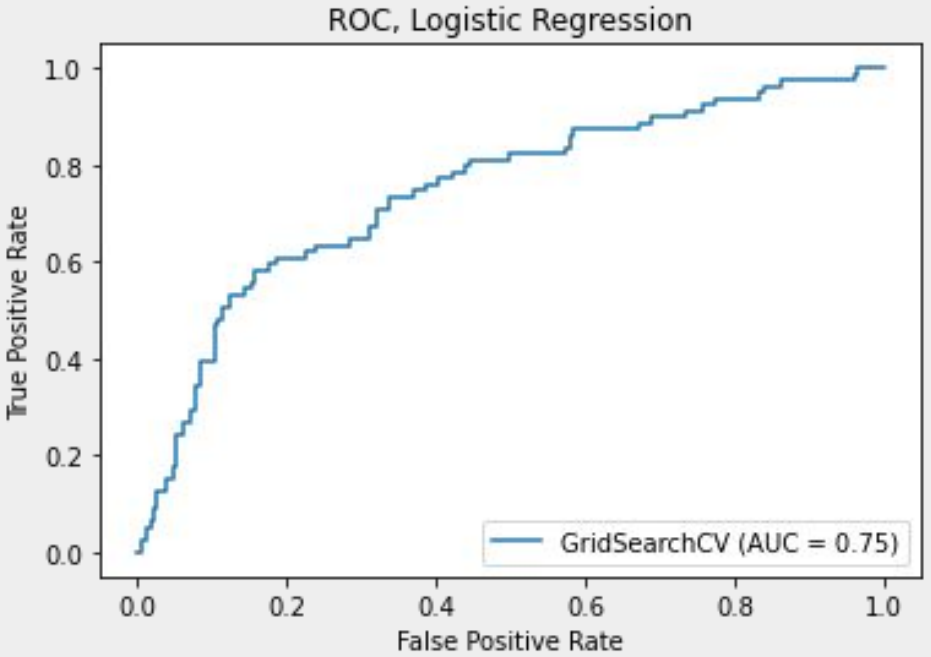
Dummy Classifier

Accuracy: .748



Logistic Regression

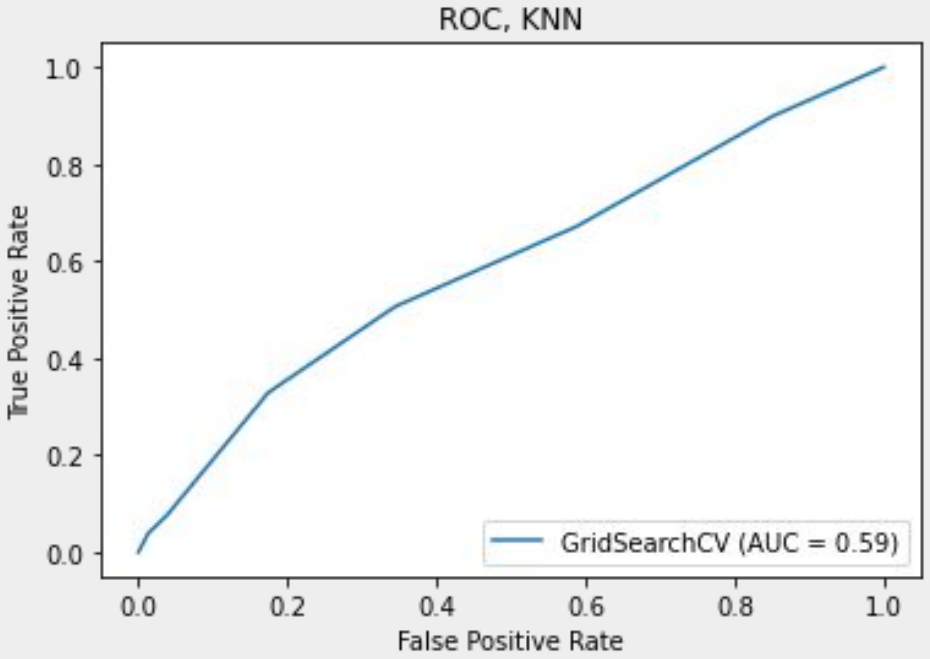


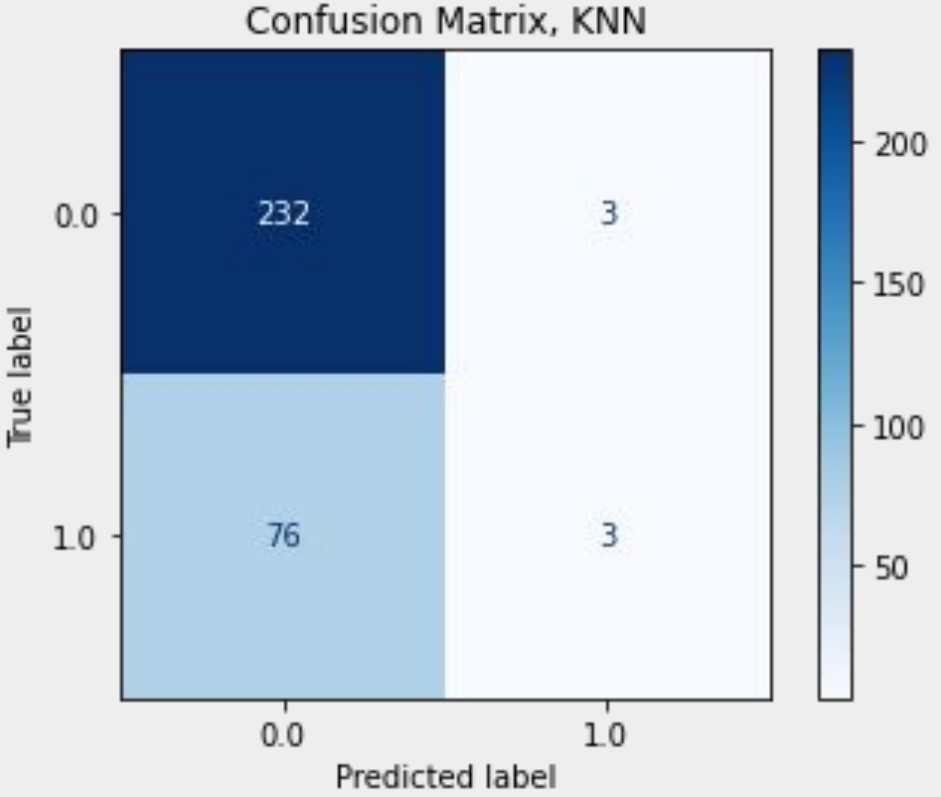


Accuracy: .768

Recall: .228

K-nearest Neighbors

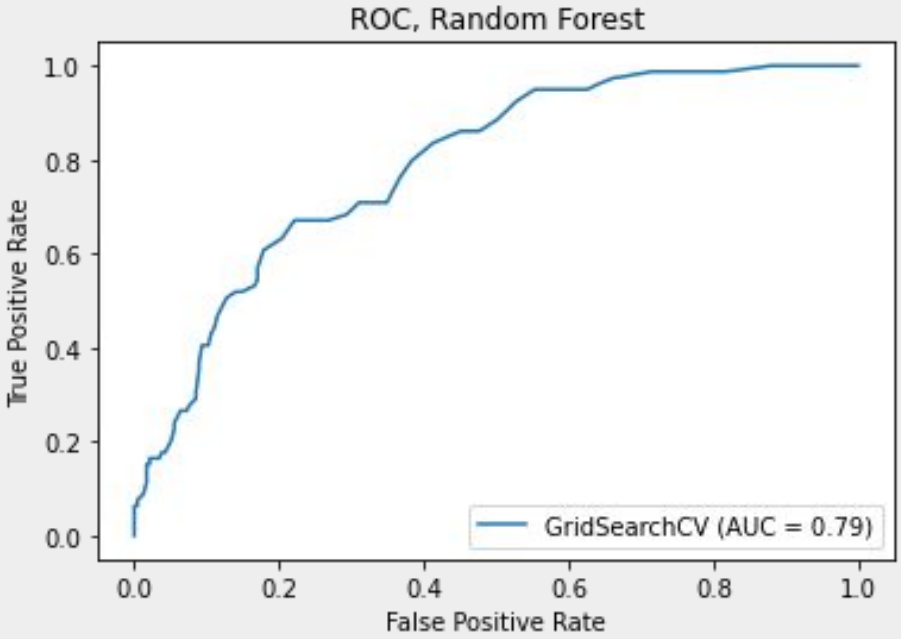


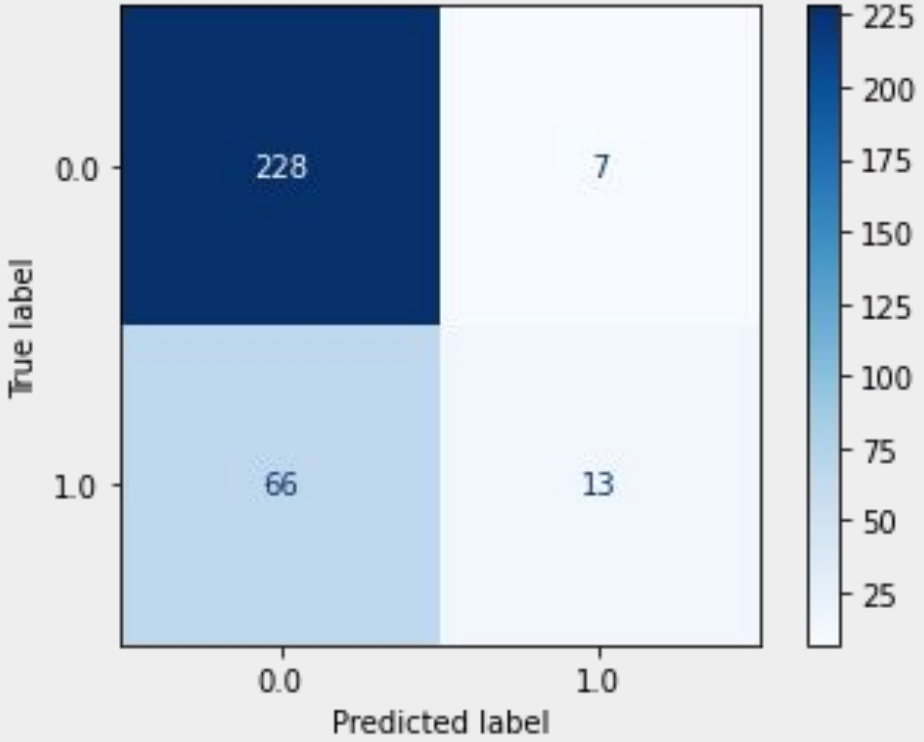


Accuracy: .748

Recall: .038

Random Forest

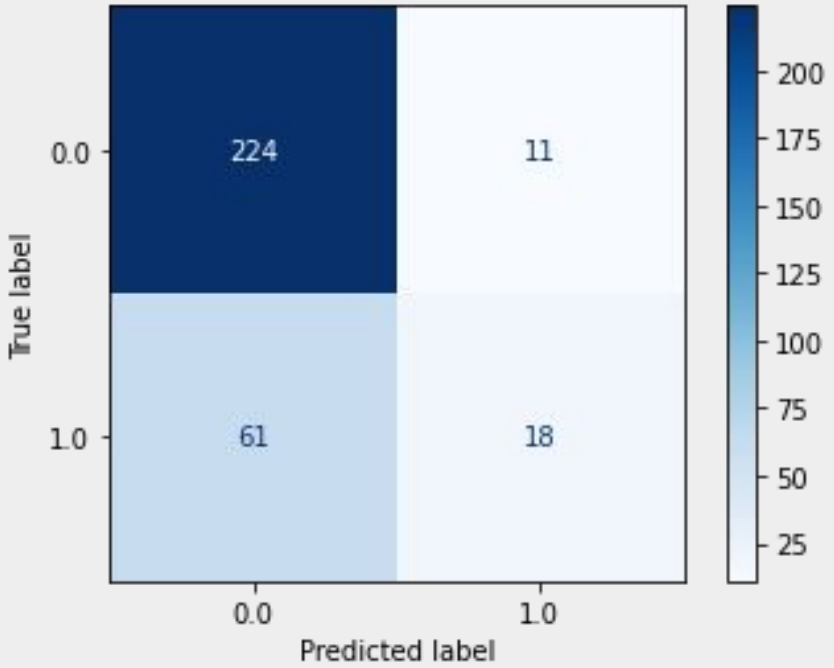


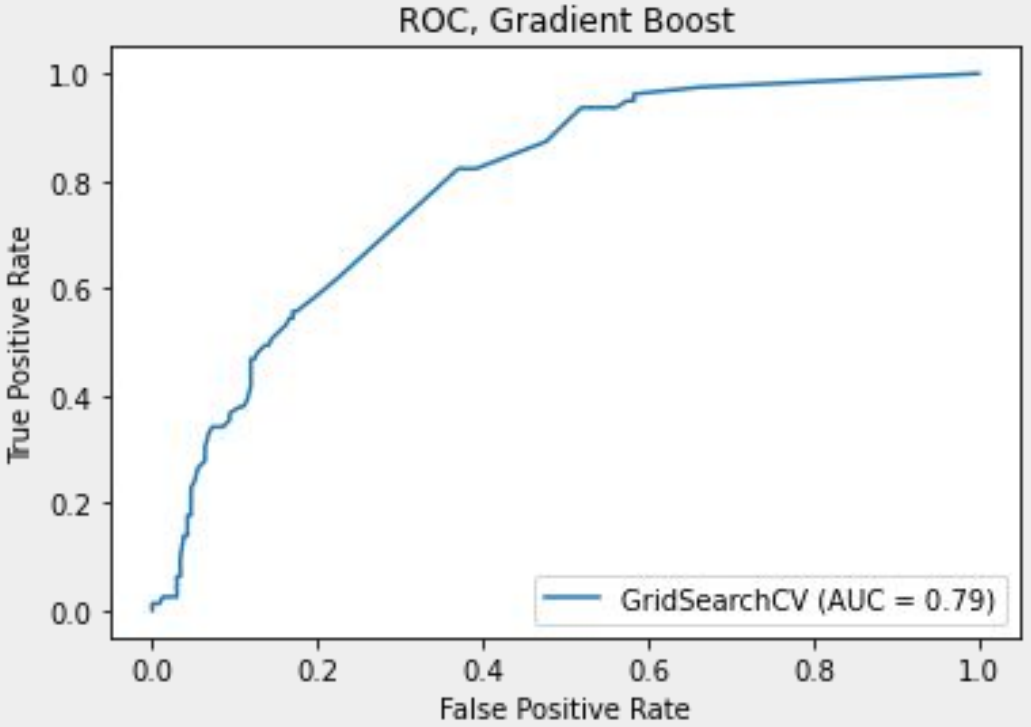


Accuracy: .768

Recall: .165

Gradient Boosting





Accuracy: .771

Recall: .228

Conclusion:

- Need more data
- Even the ‘best’ models are just guessing all students will pass
- Teachers should assume all students will pass!

Next Steps:

- Review literature for similar studies with successful results
- Increase n size or use different types of data
- If available, use techniques for imbalanced datasets (e.g. SMOTE Non-continuous)

What I Learned:

- n size is important, obviously
- Consider the whole process
- Modeling is more than hoping for the best!