Frame Title

Predicting Exam Results

Jonathan Lewis

April 14 2025

Question

What Habits for Students are Most Important to Pass Exams?

We would like to set out and answer two main questions:

- Can we model student performance and predict whether they will pass or fail?
- Can we interpret these models in a way that presents some factors or habits as better than others?

Data

Source: `https://www.kaggle.com/datasets/lainguyn123/student-performance-factors`

Data Description

- Data includes exam score data, along with information about the students.
- Columns include study habits, parental factors, and extracurricular data.

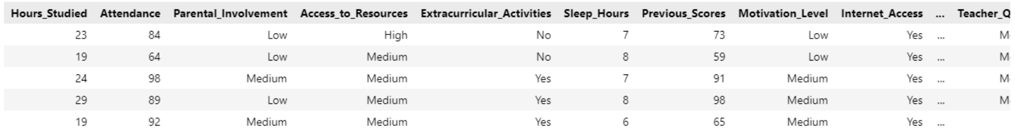
Use of Data

- A binary variable was created determining whether a student had passed the exam. Pass threshold $= 70\%$

Data

Data Format

Below is the head of the data



Hours_Studied	Attendance	Parental_Involvement	Access_to_Resources	Extracurricular_Activities	Sleep_Hours	Previous_Scores	Motivation_Level	Internet_Access	...	Teacher_Q
23	84	Low	High	No	7	73	Low	Yes	...	M
19	64	Low	Medium	No	8	59	Low	Yes	...	M
24	98	Medium	Medium	Yes	7	91	Medium	Yes	...	M
29	89	Low	Medium	Yes	8	98	Medium	Yes	...	M
19	92	Medium	Medium	Yes	6	65	Medium	Yes	...	

Description of Data

For training purposes, all of the columns are included. There is no nesting, and the data was loaded from a single-table SQL database as a pandas dataframe.

Methods

Predictive Models

- The target variable (whether a student passed) is binary. Logistic Regression will be the baseline for prediction, and will be used for variable importance.
- LASSO regression, random forest, and XGBoost will also be used and compared to the baseline model.

ROC Curves and AUC

- ROC curves can be represented visually and reflect the tradeoff at different thresholds. Also allows for easy model comparison.
- AUC is a measure of overall accuracy across all of the thresholds.

Methods

Programming Language

- The coding language used was Python. It contains the most out-of-the-bag models and packages for various measures.
- The libraries used include SCI-KIT LEARN, XGBoost, Pandas, Numpy, and Sqlite3.

Logistic Regression

- Because the coefficients reflect the change in odds, the logistic regression model will be the primary use in investigating the importance of habits and factors.

Results

AUCs

- Logistic Regression AUC: **0.985**
- LASSO AUC: **0.981**
- Random Forest AUC: **0.853**
- Boosting AUC: **0.923**

General Findings

- Logistic Regression performed the best. Likely because data is very linear.
- Random forest performed the poorest. Could be due to lack of parameter tuning.

Results

Visualizing AUC through ROC Curves

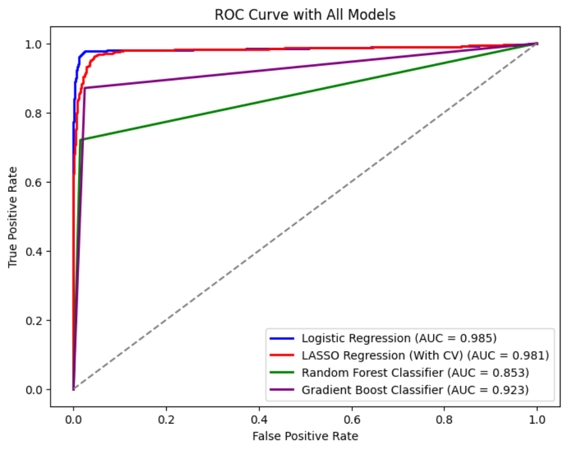
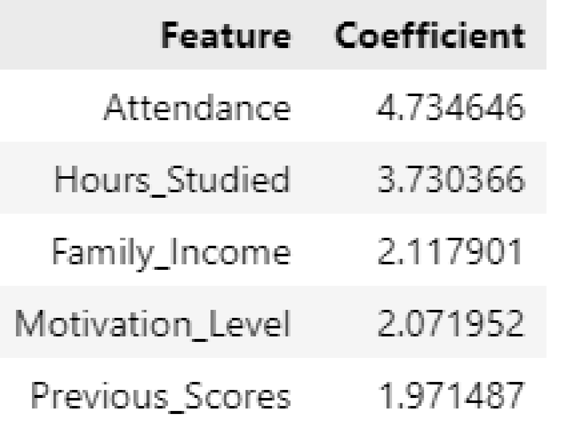


Figure: ROC curves comparing model performance (further up and to the left is better).

Results

Ranked Coefficients for Logistic Regression



Feature	Coefficient
Attendance	4.734646
Hours_Studied	3.730366
Family_Income	2.117901
Motivation_Level	2.071952
Previous_Scores	1.971487

- Attendance rate is considered more important than studying.

Results

LASSO Results

- This model came closest to logistic regression. Cross validation was used to find the best parameter.

Variables Eliminated by LASSO

Two variables were eliminated by the LASSO model:

- School type (public or private)
- Gender

Conclusion

Findings

- Whether a student will pass is very predictable, and the models performed extremely well.
- The models also provided insight into what factors were most prevalent.

Code and data can be found here:
`https://github.com/Staticy01/Student_Grades`