The University of Texas at Austin

Modeling Student Exam Performance

A study of performance of students in school examinations as a function of social, demographic and behavioral variables

Chris Lawson, Alberto Salvarese, Utkarsh Mujumdar, Jeffrey Gordon

Problem Description

Education is highly regarded as a key factor for a long-term economic achievement.

To guarantee equal opportunities, a society should therefore invest more on the education, developing a school system able to understand and fill the gaps in different students' background, and guarantee a fair educational path

s have improved, there are realities that still
s in the educational system

Goal: Try to evaluate which features contribute the most in determining students' academic success, to have a better understanding on where to intervene to guarantee fairness

Dataset

Source

Grades of students from two secondary schools in Portugal in the subjects of Mathematics and Portuguese language, collected in 2008[1]

Features

Demographic
- School
- Gender
- Age
- House Type
- Family Size

Social/Economic
- Mother's/Father's Education
- Mother's/Father's Occupation
- Parent's Cohabitation
- Internet access
- Family Support

Behavioral
- Studytime
- Freetime
- Participation in a romantic relationship
- Frequency of going out
- Alcohol Consumption

Target

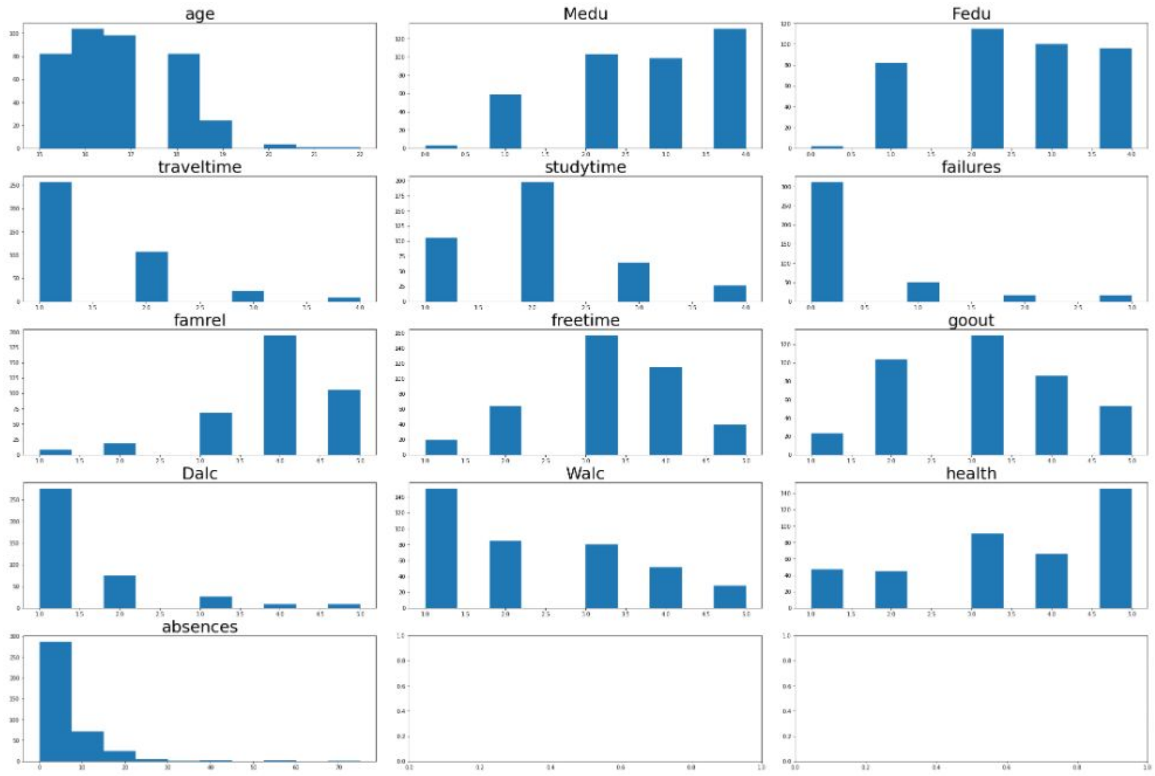
G3

Final Grade obtained in the respective subject (Mathematics/ Portuguese)

1: P.Cortez, A. Silva, *Using Data Mining to Predict Secondary Student Performance*, http://www3.dsi.uminho.pt/pcortez/student.pdf

Feature Exploration

Distributions of numeric variables

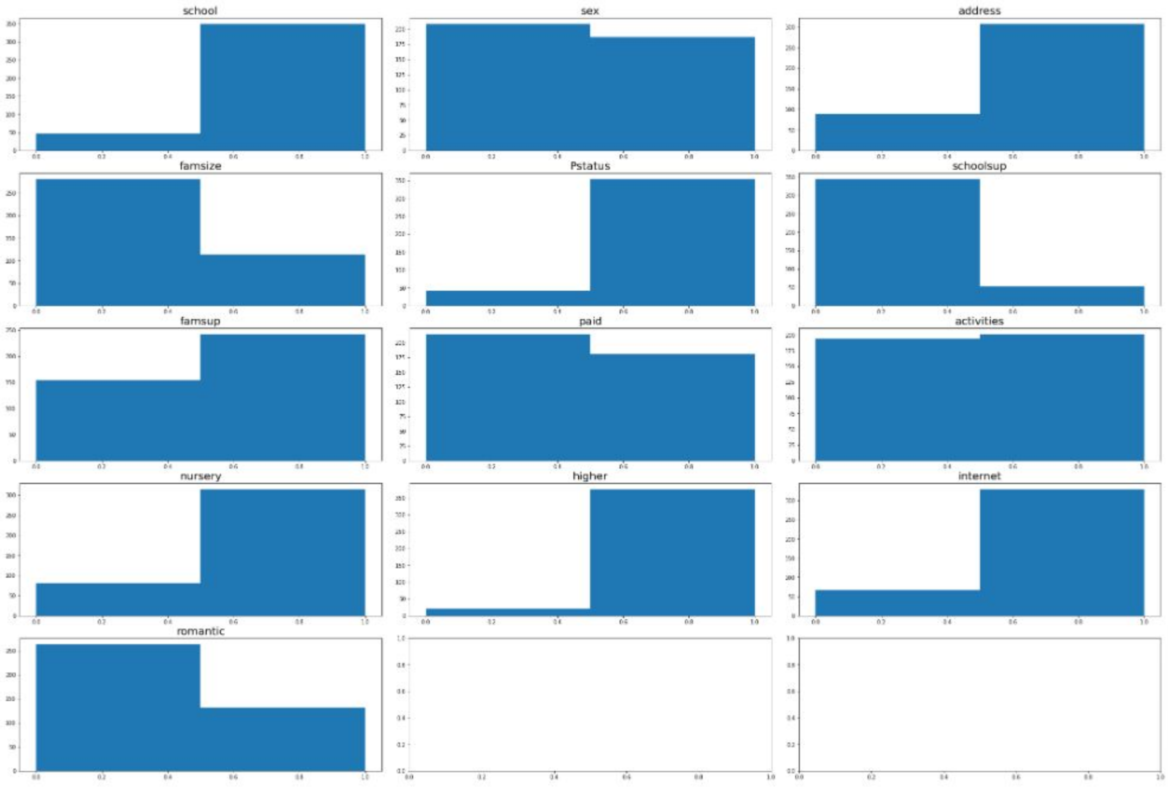


The University of Texas at Austin

WHAT STARTS HERE CHANGES THE WORLD

Feature Exploration

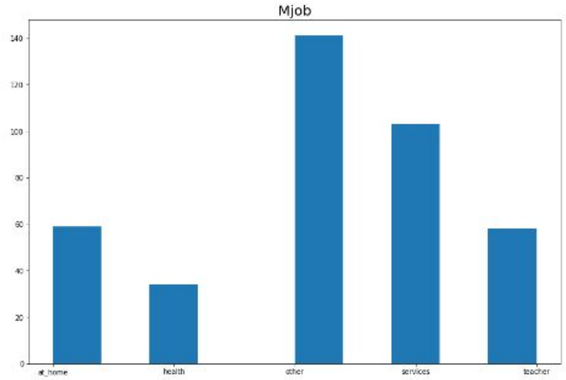
Distributions of categorical variables

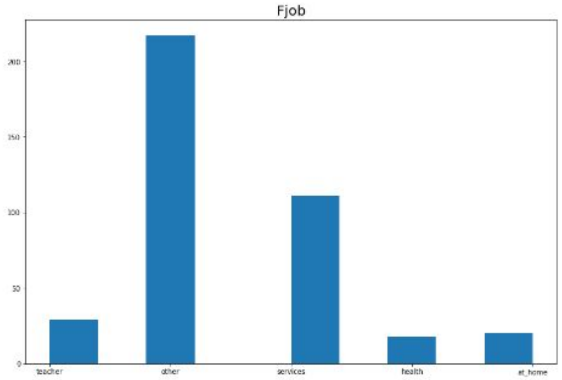


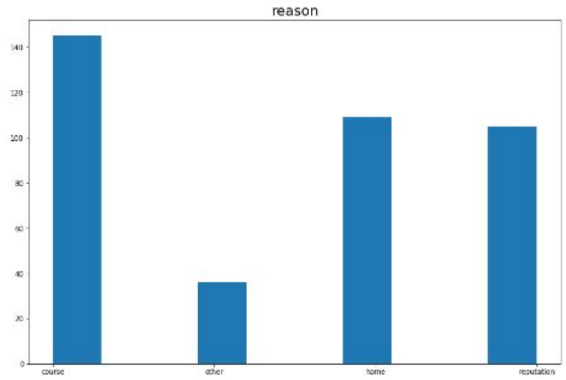
The University of Texas at Austin
WHAT STARTS HERE CHANGES THE WORLD

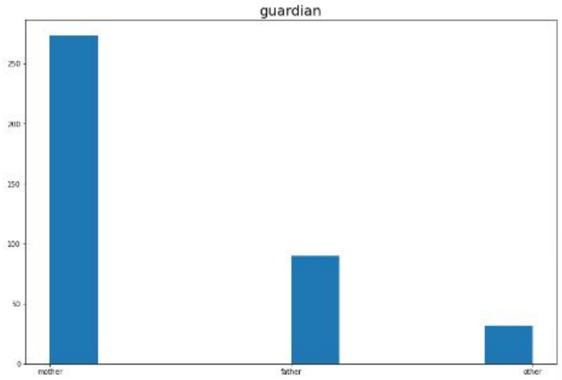
Feature Exploration

Distributions of categorical variables (contd)



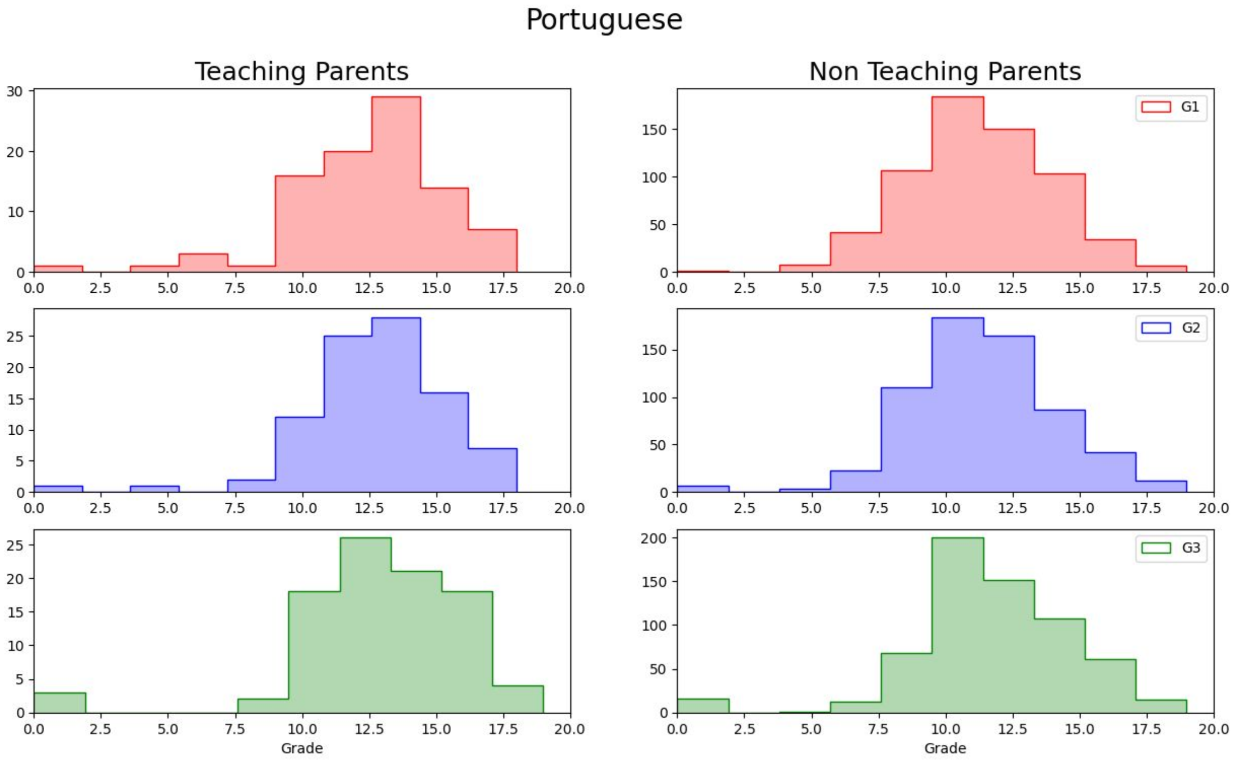






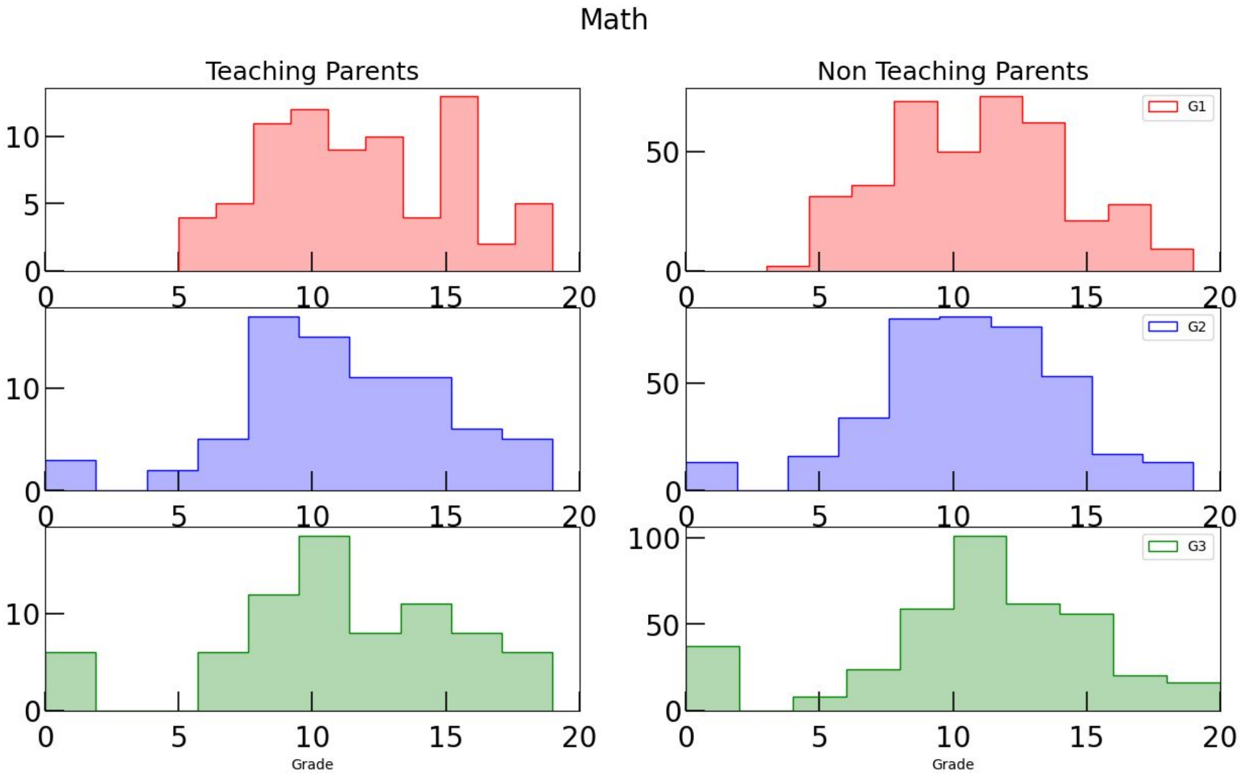
Feature Exploration

Comparison of grades for students having at least one parent who is a teacher vs students having no parents as teachers



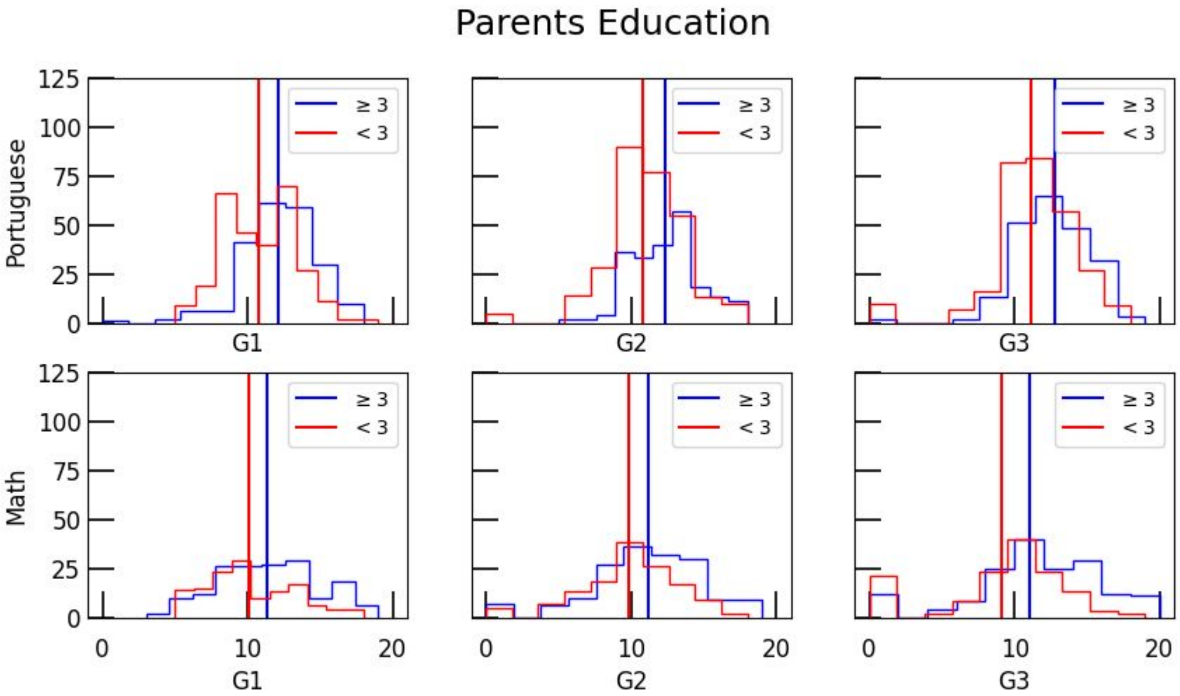
Feature Exploration

Comparison of grades for students having at least one parent who is a teacher vs students having no parents as teachers



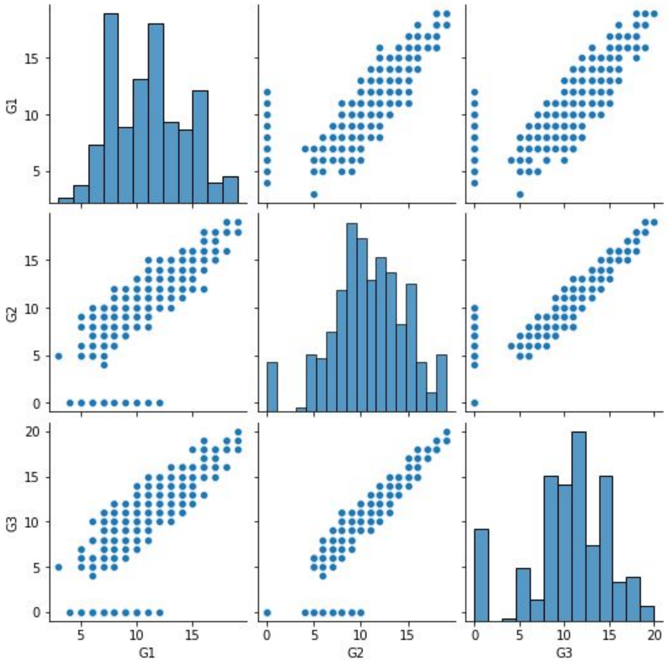
Feature Exploration

Effect of parents education on the student's grades in Math & Portuguese
>=3: Secondary education or Higher
<3: Lower than secondary education



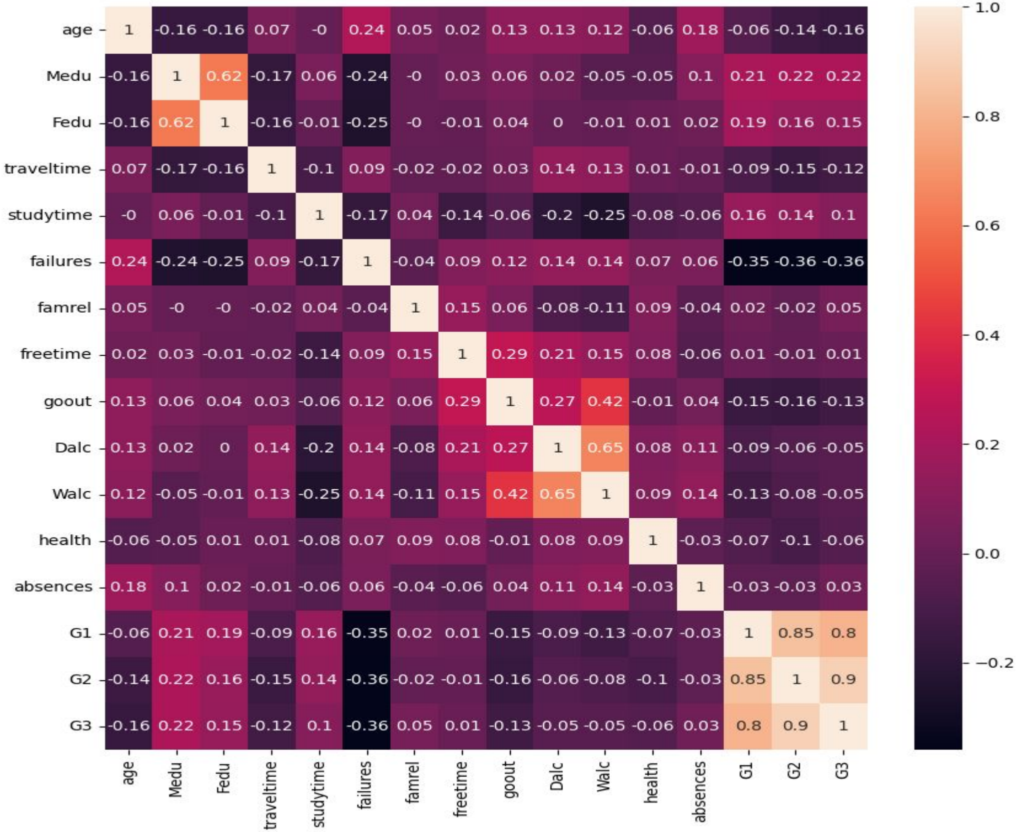
Feature Exploration

Correlations between grades G1, G2 and G3 for the Math subject

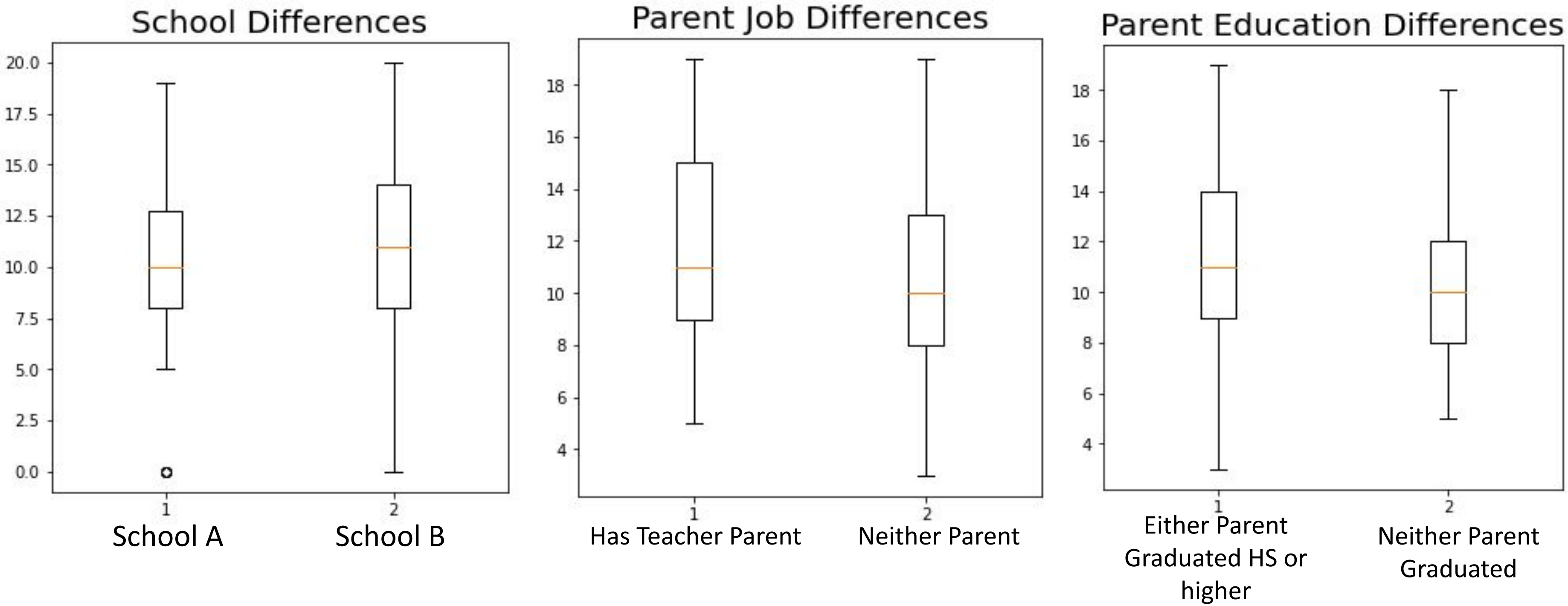


Feature Exploration

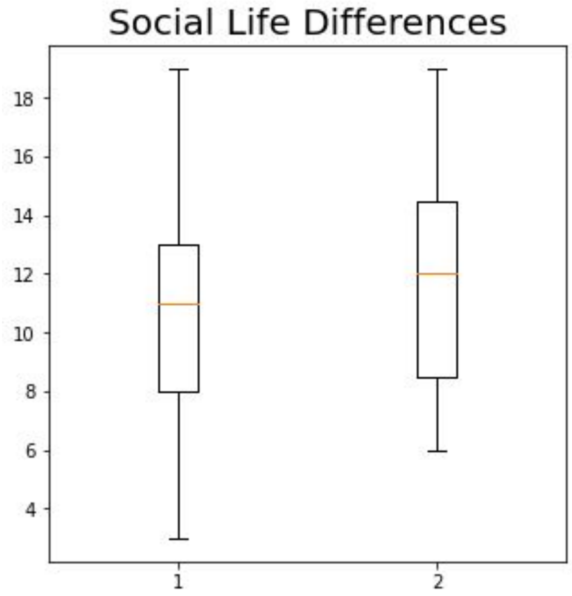
Correlation matrix of all numeric features



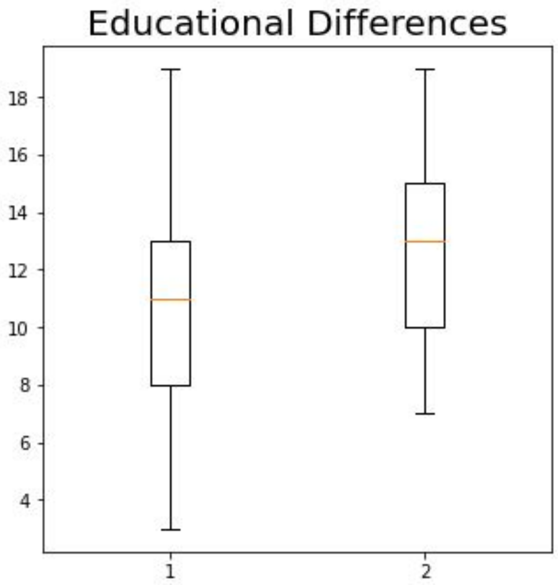
Feature Exploration



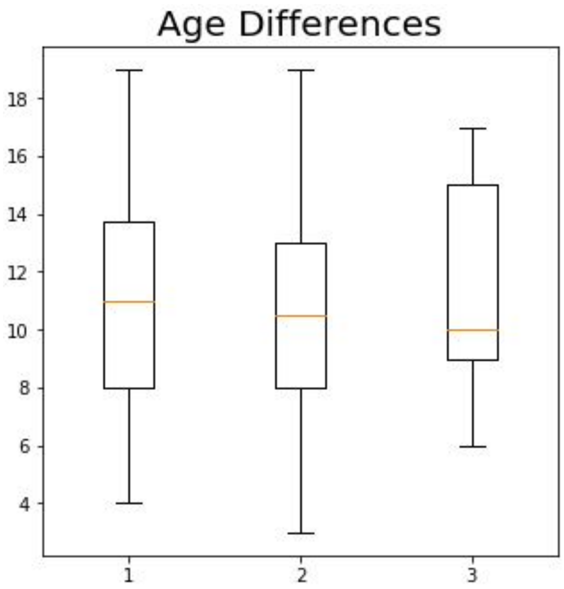
Feature Exploration



Goes out More Goes less More

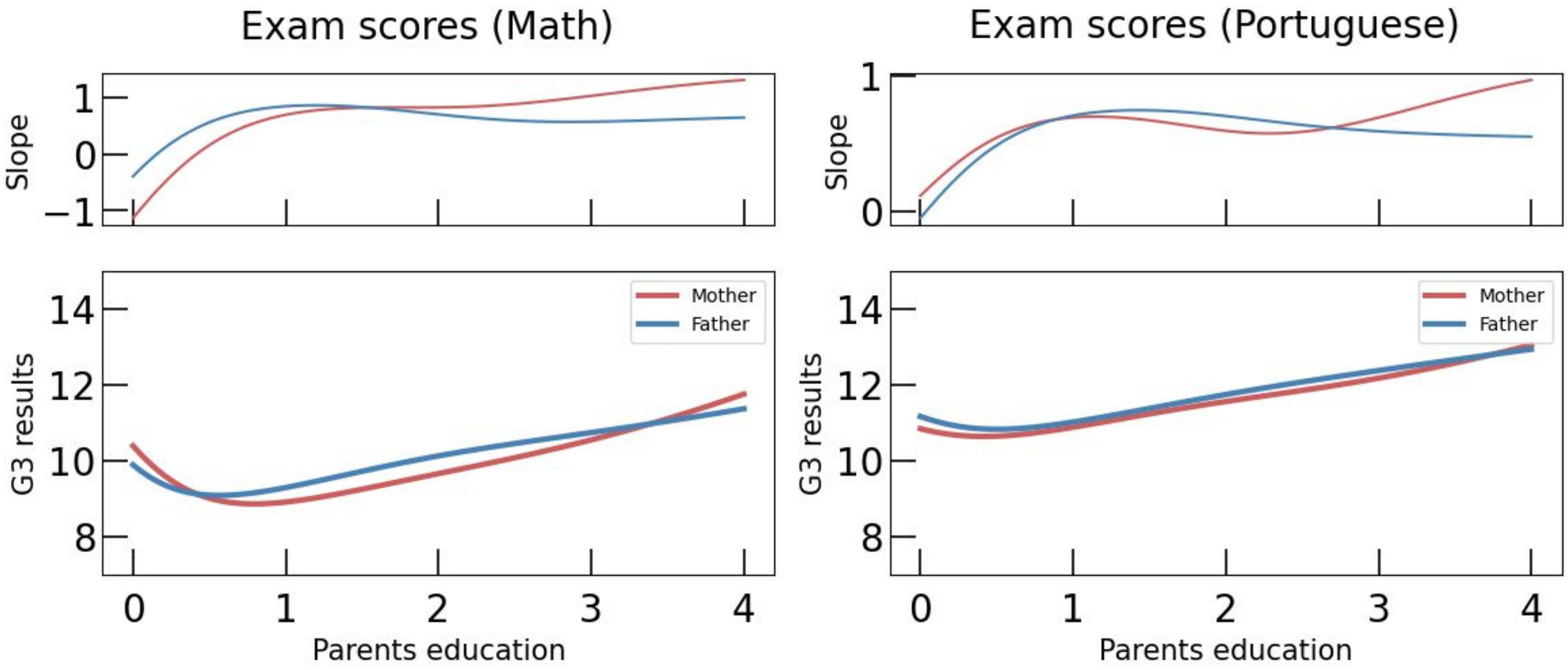


Higher failures, absences Less failure absences

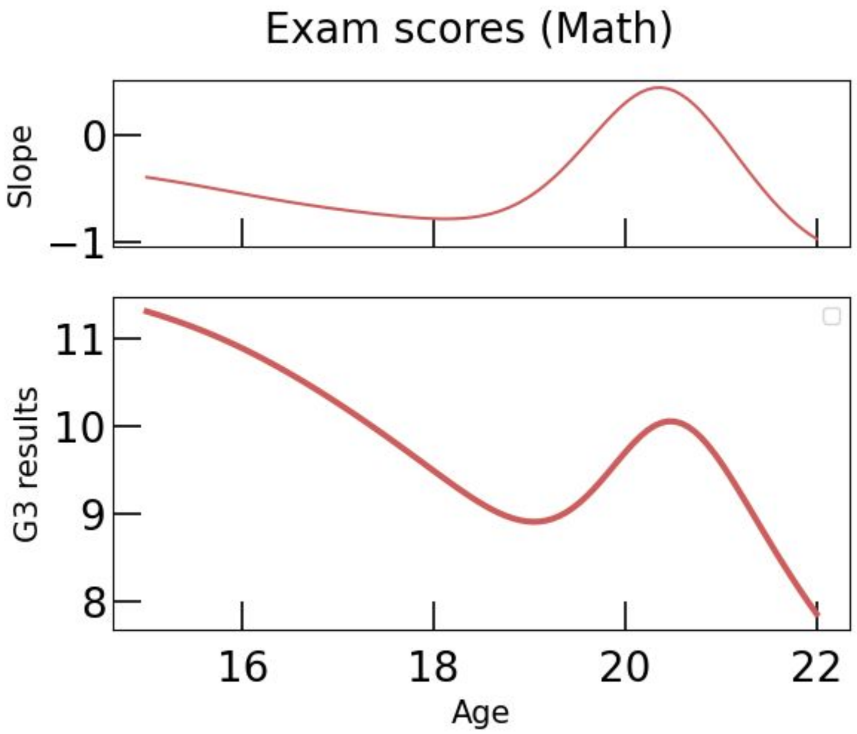


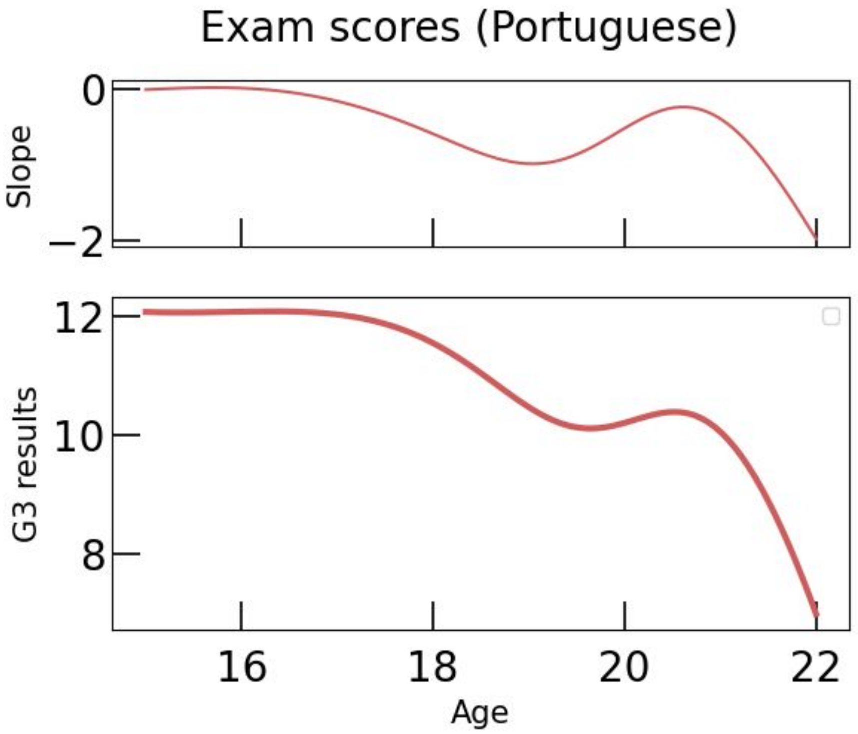
17) [17, 20) [20,

Feature Exploration

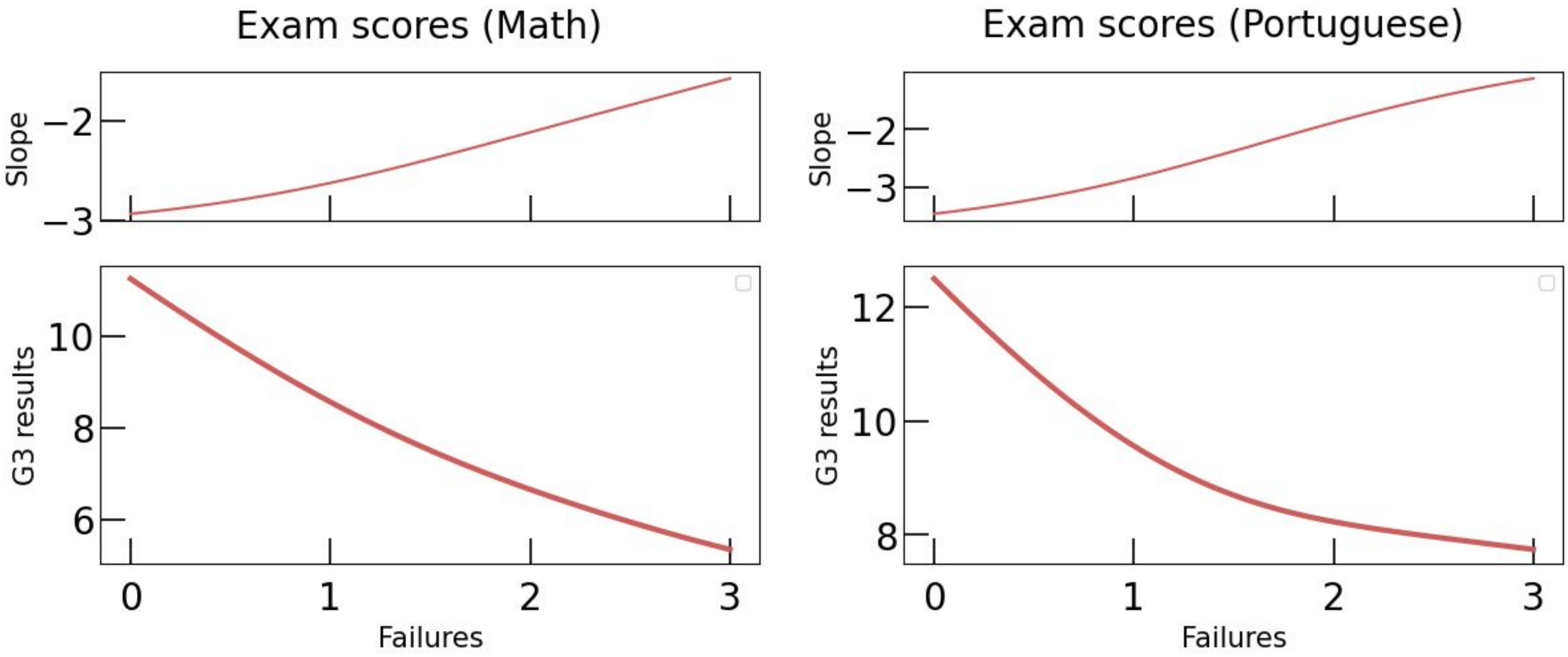


Feature Exploration





Feature Exploration



Feature Engineering

Stratifying data with respect to three variables: ‘Student quality’, ‘School’, and ‘Absences’

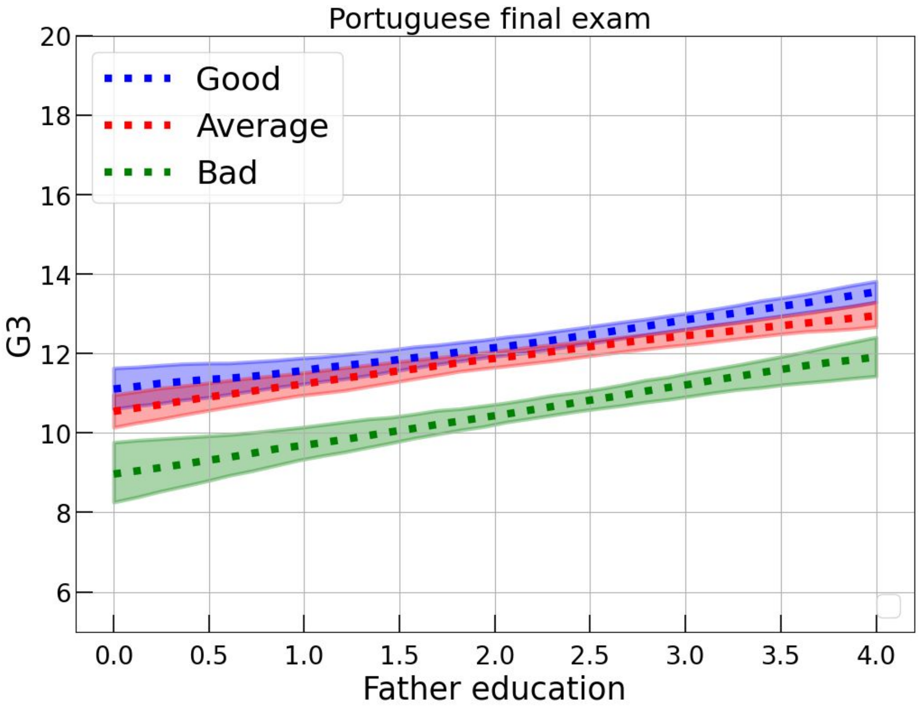
Student quality: sum of (normalized) absences, hours going out, daily alcohol, and weekend alcohol consumption

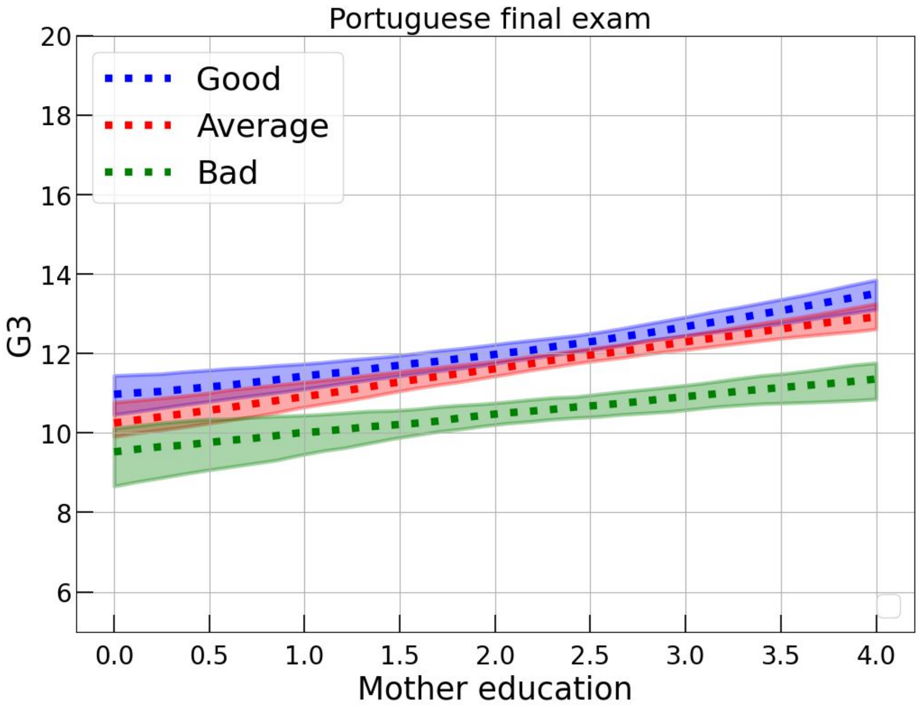
‘**Good student**’: from 0 to 1; ‘**Average**’: from 1.1 to mean+std; ‘**Bad**’: above mean+std

School: students enrolled in Gabriel Pereira or Mousinho da Silveira

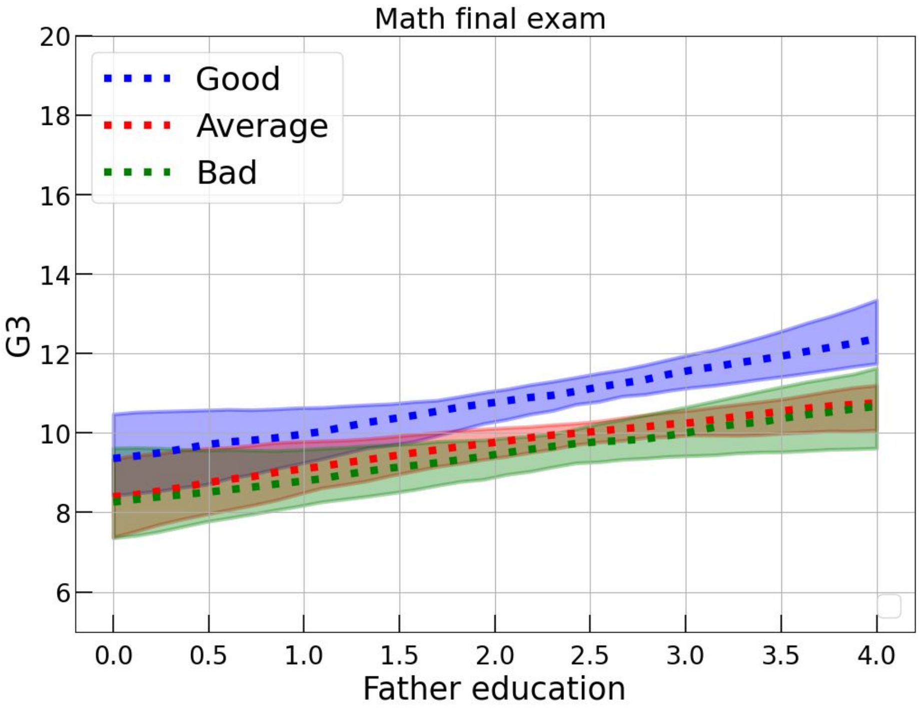
Absences: from 0 to 5; from 5 to 20; above 20

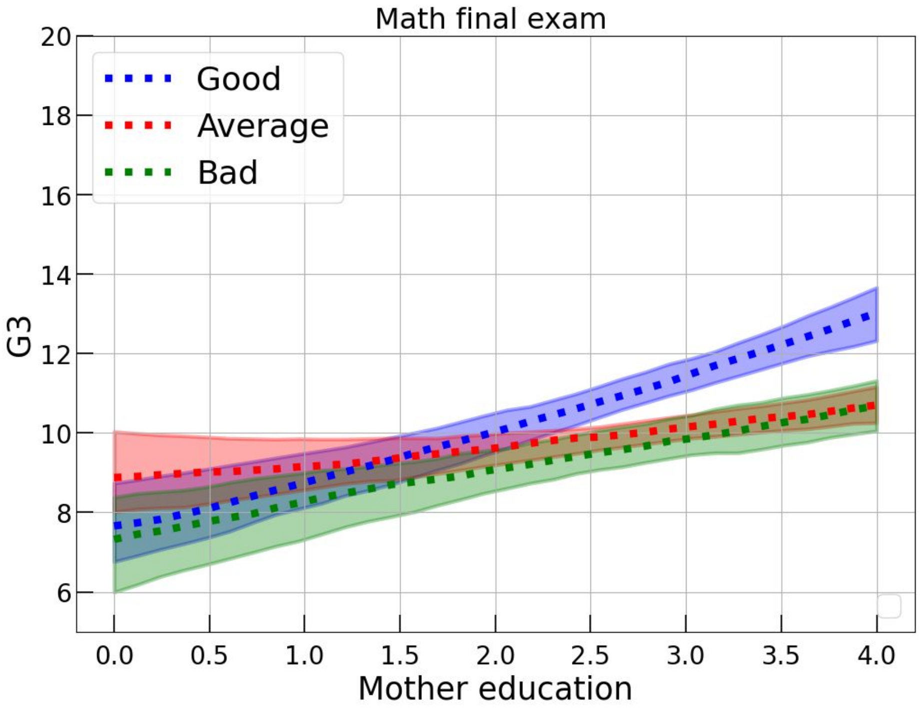
Feature Exploration



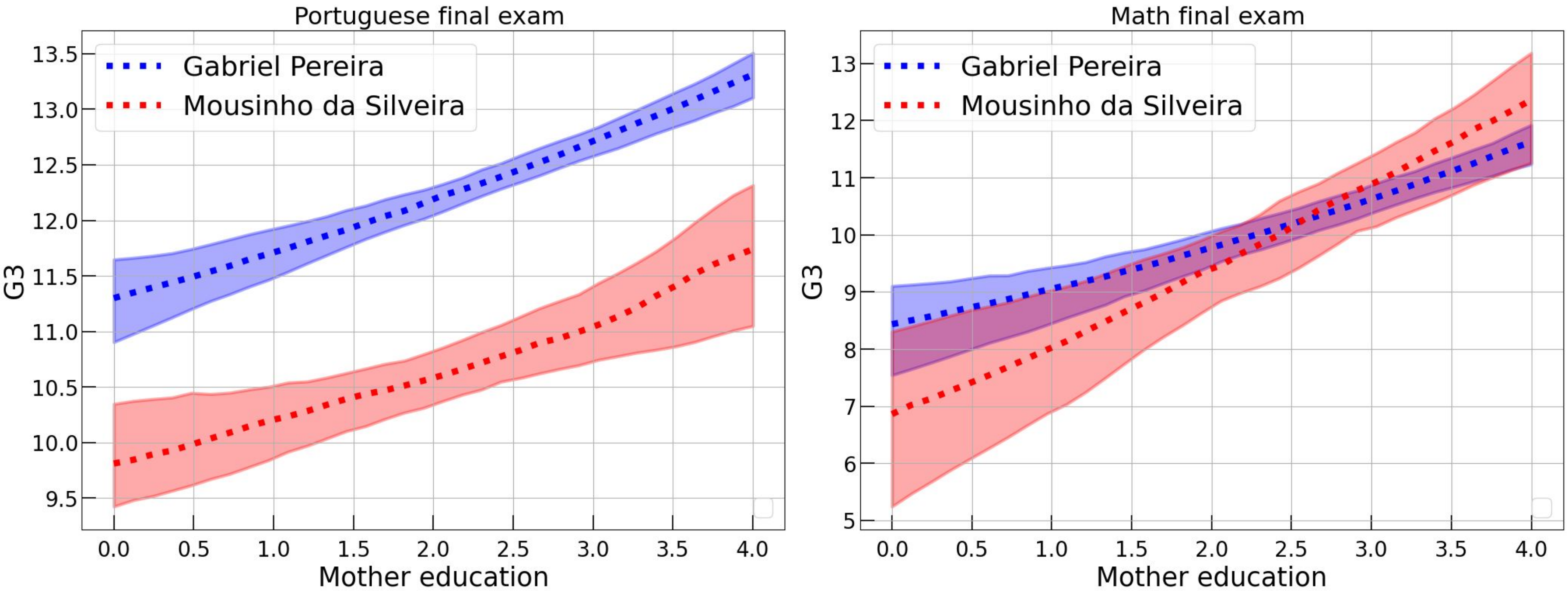


Feature Exploration

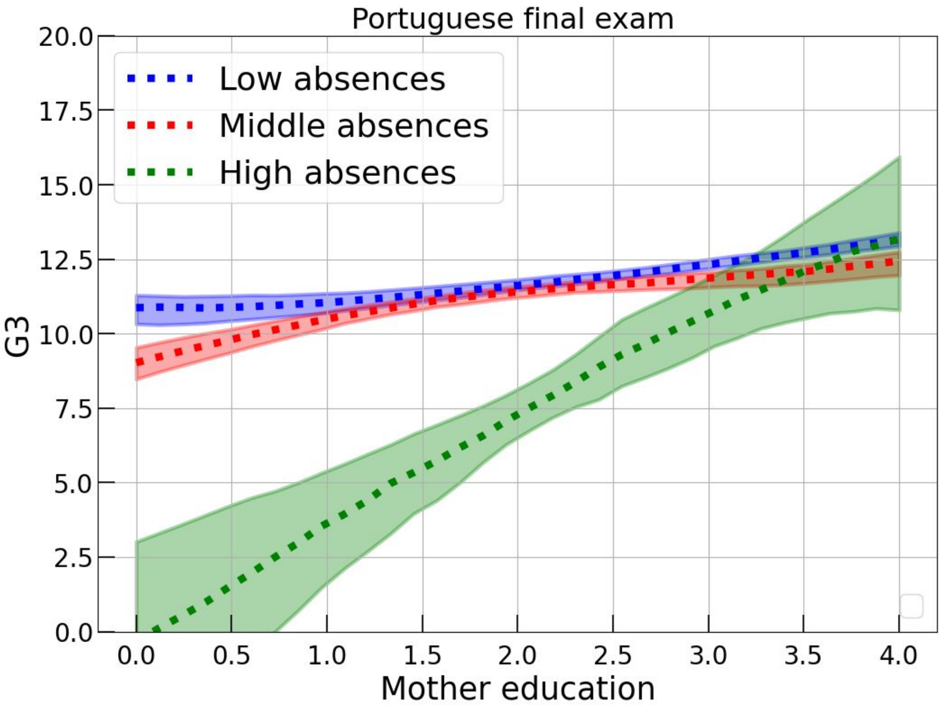


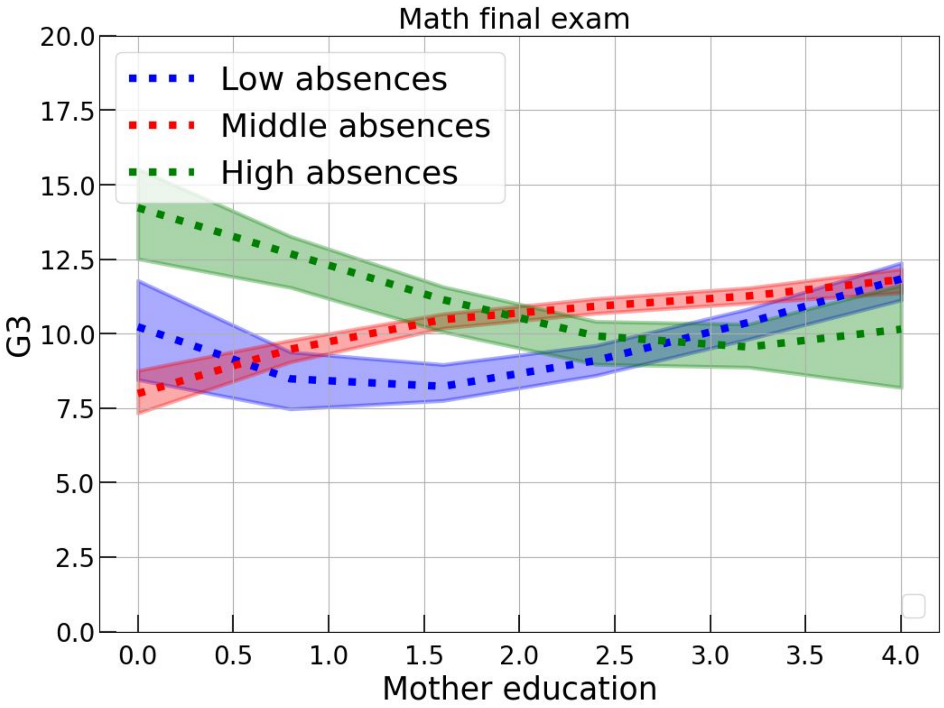


Feature Exploration



Feature Exploration





Project Hypothesis

- Has the school students are enrolled in an impact on their results?

- Does the parents’ education have an influence on the grades students receive?

- Does having parents that worked as a teacher (one or both) have a positive influence on the grades they received?

- Does the age the students have an influence on their exam results?

- Is there a correlation between any of the after school variables (romantic, freetime, going out etc) that have an impact on student performance?

Modeling

First round of regression models applied:

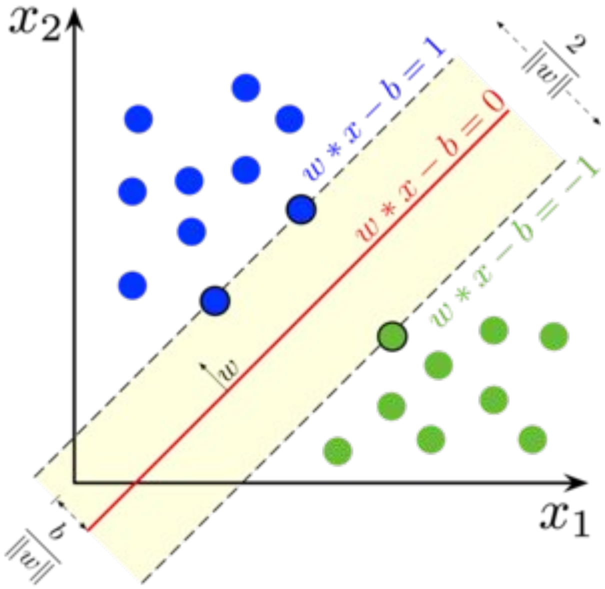
- Linear
- SVM

Kernel Type Models

- Gaussian Process
- Kernel Ridge

Ensemble type Models

- Decision Tree
- Random Forest



Modeling

Feature Manipulation:

1) Categorical variables were OHEd

2) Features dropped:
   a) Reason
   b) School
   c) Guardian
   d) Father Education
   e) Weekday Alcohol Consumption

3) Will only use G3 as a target variable

Data Manipulation:

1) Base Models with Specific Hypertuning

2) Scaling Features

3) Feature Reduction (PCA)

Each step included modifications from previous steps

Modeling

Model Specific Hypertuning parameters:

- Gaussian Process:
  - Default Kernel (Constant * RBF)
  - Dot + White kernels
  - RBF + White kernels
- Kernel Ridge:
  - Linear kernel
  - Polynomial kernel
  - RBF kernel
  - Sigmoid kernel
  - Alpha between 0 and 1 in 0.1 increments
- SVM:
  - Same kernels as KR
  - Degree from 0 to 12
- Decision Tree
  - Max Depth between 5 and 20 in 5 increments
  - Max Features between 5 and 15 in 5 increments
  - Min leaf samples between 5 and 20 in 5 increments
- Random Forest
  - Same as Decision Tree

GAN + Predictor Network

- **GAN**: Generator and Discriminator Neural Network Models
- **Generator (4 Feed Forward + ReLU)**
  - Receives random noise as input
  - When trained, outputs fake data similar to the actual data
- **Discriminator (4 Feed Forward + ReLU + Sigmoid)**
  - Receives samples as input
  - Classifies the data as fake/generated, or real
- **Minimax Game** - Trained Simultaneously using one another
- **Predictor Network (4 Feed Forward + ReLU)**
  - Receives samples as input
  - Outputs the predicted G3 Score
- **Goal**: See if a deep learning approach would give us any significant results/decent models, which would provide us with more insight into the data as a whole

Validation Techniques

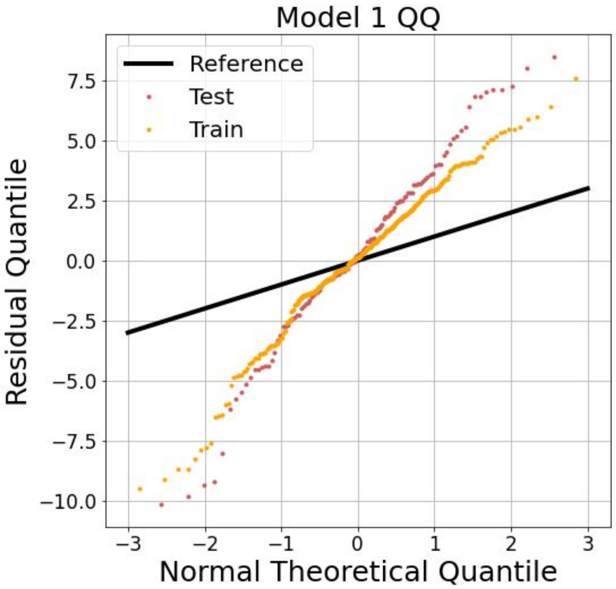
For each model the following validation techniques were used:

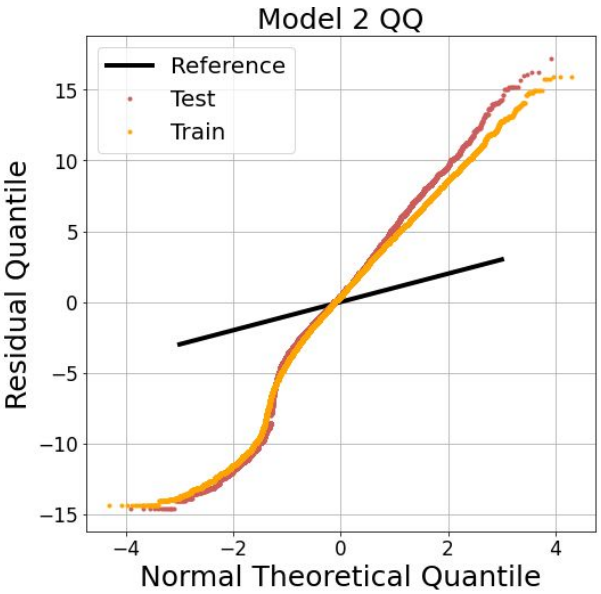
- 10-fold validation
- QQ plots
- Scoring Metric: MSE

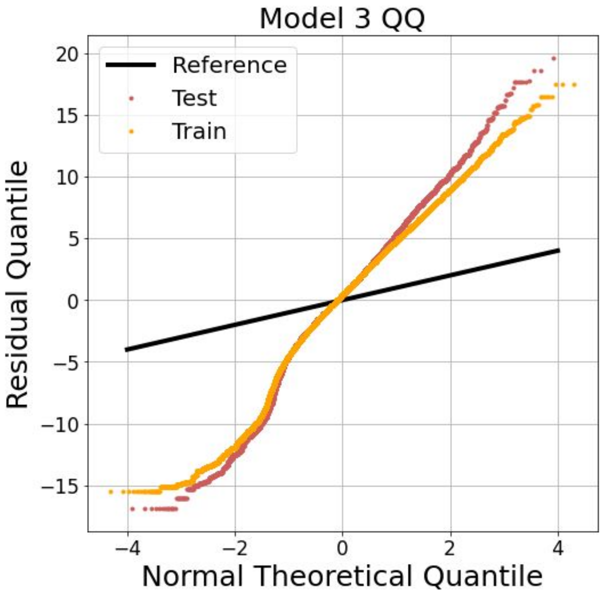
Results

Model + Dataset	MSE	Standard Deviation
Random Forest (Scaled)	17.403	6.398
Gaussian Process (Base)	18.814	7.977
Gaussian Process (All)	19.228	8.000
Gaussian Process (Scaled)	19.228	8.000
Linear (Base)	19.44	8.030
Linear (Scaled)	19.45	8.036

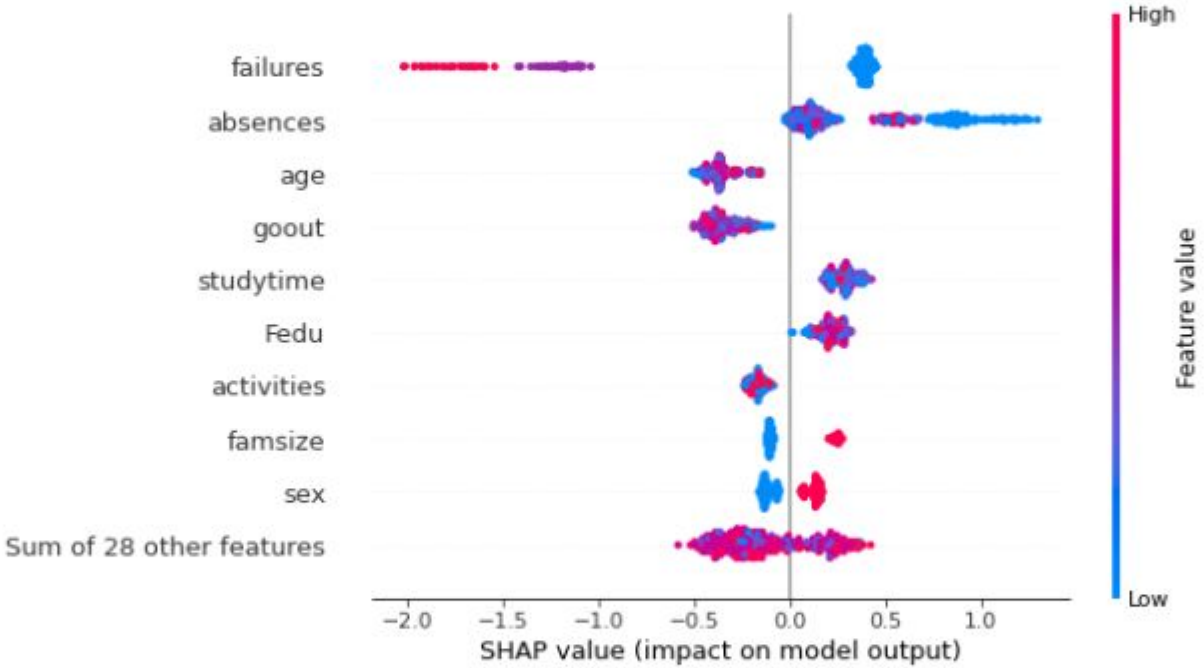
Results


Model 1 QQ
Reference
Test
Train
Residual Quantile
Normal Theoretical Quantile


Model 2 QQ
Reference
Test
Train
Residual Quantile
Normal Theoretical Quantile

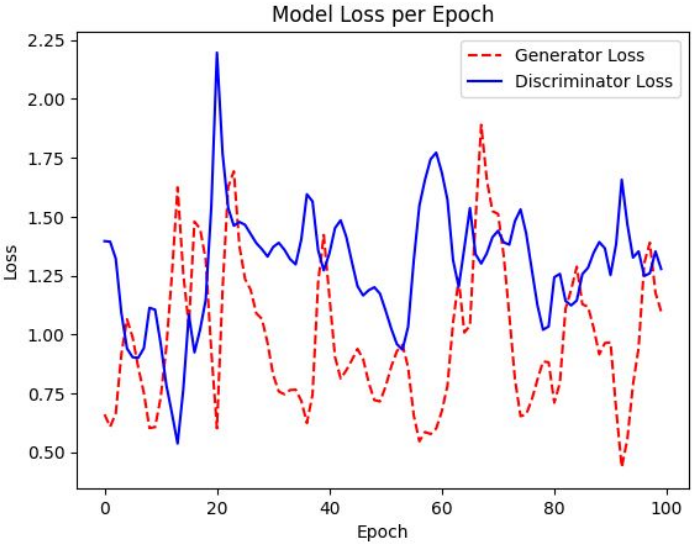

Model 3 QQ
Reference
Test
Train
Residual Quantile
Normal Theoretical Quantile

Results - SHAP

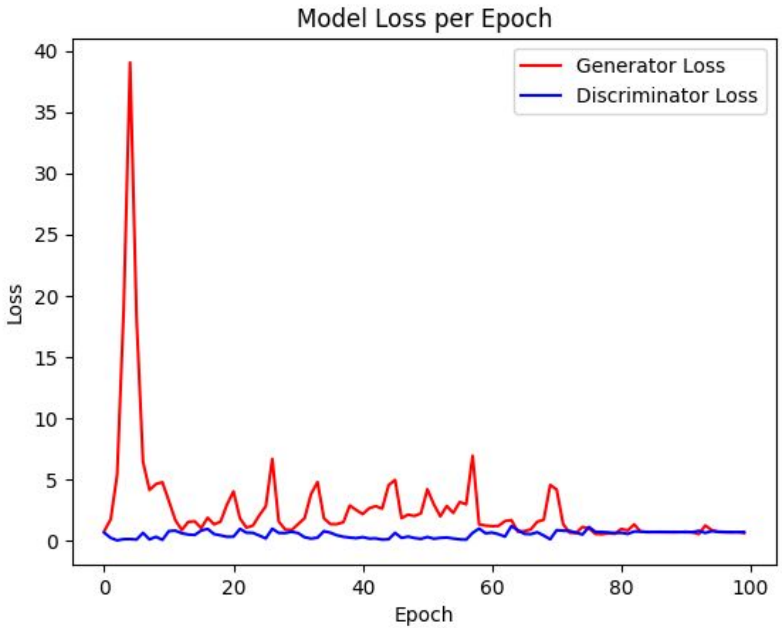


GAN Results

Initial GAN Loss Curves



Final GAN Curve

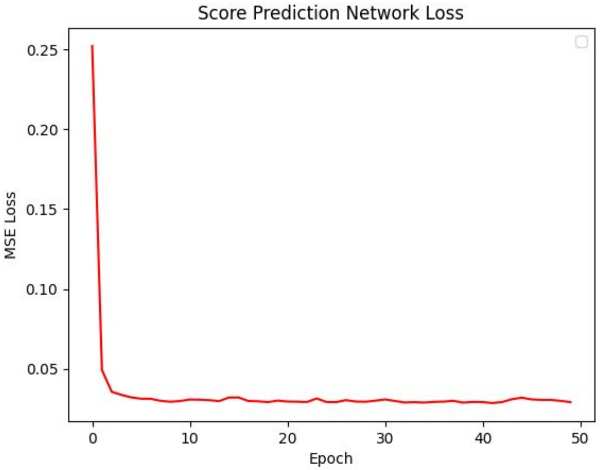


GAN Interpretation

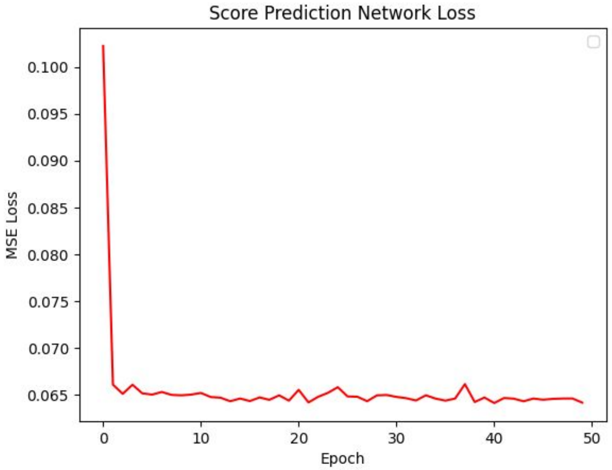
- Best BCE Loss found was in the .59 - .69 range for both models
- Large oscillation and divergence due to the minimax between models being one sided most of the time (discriminator was improved much faster than generator or vice versa, hinders training)
- You generally want to see a small bit of divergence, until both models reach an equilibrium
- Our model reaching equilibrium while still having a decently high BCE Loss tells us that the generator couldn’t find the underlying distribution of the real data, so there's nothing pushing the discriminator to improve

Predictor Network

Trained on Real Data (.029 Loss)



Trained on Real + Fake Data (.064 Loss)



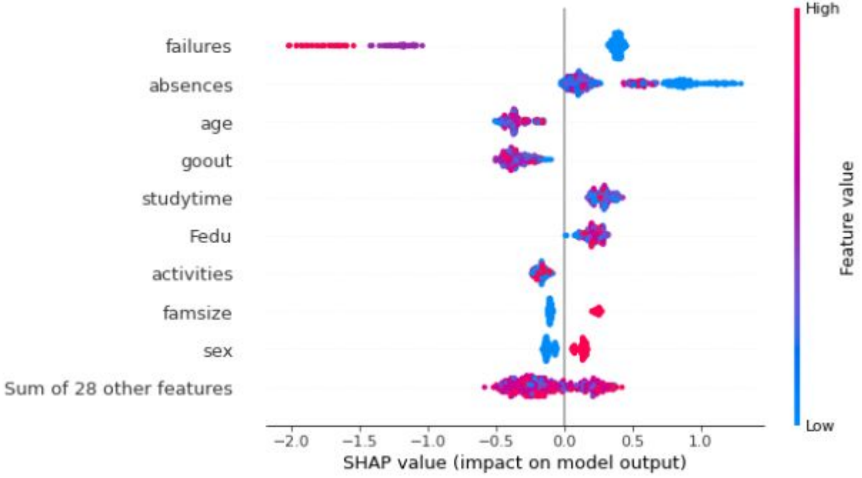
- Good loss value w/ real data, slightly worse with real and generated data, as expected
- The steep and early loss decrease given a very low learning rate tells us:
  - Model had too easy of a time learning to predict the score
  - Model is too complex and overfitting, or the features aren't good predictors of the score
  - However with simple models, convergence didn't happen

Thoughts

- All in all, the neural network models performed mediocre and is most likely not a good fit for the data
- Generator could be used to generate more diverse data, therefore giving the networks something to find patterns in
- Future Improvements:
  – Cross Validation
  – More generated data
  – More optimal models through hyperparameter tuning and training methods

Discussion

- Best model - Random Forest
  - Could have potentially have been overfit
- Questionable Q-Q plots
- SHAP corresponds with our hypothesis
- Limitations
  - Treating categorical variables as numerical
  - Disproportionate group sizes



Conclusion

- Best model
  - Random Forest

- Hypothesis 1 - School Influence
  - **Plausible**
- Hypothesis 2 - Parent's Education
  - **True**
- Hypothesis 3 - Parent's Job
  - **Plausible**
- Hypothesis 4 - Student Age
  - **True**
- Hypothesis 5 - After-school life
  - **Plausible**
- Final Conclusion:
  - Academic Features are most important

Important Features coincide with paper over this dataset

Contribution

Name	Contribution
Alberto Salvarese	100
Chris Lawson	100
Jeffrey Gordon	100
Utkarsh Mujumdar	100