Applying Bayesian Methods for Final Grade Prediction

Group: Rhea Agarwal, Ami Kano, Tatev Kyosababyan

Problem Description

Conventionally, academic success is evaluated by the grades obtained throughout the educational career. Whether a letter grade or a numeric value, it is meant to assess how a student has performed while in school. Despite the possible unfairness due to ignorance towards other vital factors, grades often define a person and are seen as indicators linked to future success and opportunities. However, it is not necessarily the case that a child's academic success solely depends on their motivation and effort. In our project, we analyze the extrinsic factors impacting the academic success of middle school students in Portugal. This is a critical age for children since, besides academic growth, it is also a crucial developmental period of time, which can also impact one's grades significantly. Considering the numerous environmental and developmental contributing factors, we aim to implement Bayesian techniques to examine the uncertainty surrounding academic success for secondary students.

Approach

The data was obtained from *Kaggle*; it has about 400 entries and 33 features. It is survey data about students attending one of the two participating secondary schools in Portugal. The feature values indicate demographic information about students and other factors like the number of absences, final grades, alcohol consumption, extracurricular activities, involvement in a romantic relationship, parents' or guardians' education level, occupation, etc. After removing the redundant columns, we were left with 29 features, out of which only three were continuous: number of absences, final grades, and age.

As we can see in the heatmap in Graph 1, the numeric values fail to show a significant amount of correlation. These were later scaled for generalization purposes in our further steps. The histogram on Graph 2 emphasizes the skewness of the number of absences. We can clearly see that the majority of the students did not have any absences throughout the given period of time. The boxplot shown in Graph 3 is our outcome variable: the final grade for each individual student. The grading is on a scale of 0 to 20.

For our analysis, we will be using Bayesian Linear Regression along with Bayesian Additive Regression Trees (BART).

Probability Model

As briefly stated in the Approach section, we chose a simple Bayesian linear regression model to predict the grades of the students in the dataset. Initially, we created a full model with 29 predictors used to estimate the value of the grade variable *G3*, using sampling methods. For numerical or multinomial predictor variables, we used No U-Turn Sampling (NUTS), which is an implementation of Hamiltonian Monte Carlo Sampling. For binomial predictors, we used the basic Metropolis Sampling. This resulted in a posterior distribution (Graph 7) that was vastly different from the true distribution of the response variable *G3* (Graph 9). From the distribution plot of the posterior, we deemed the full model ineffective, and attempted to create another model with a different combination of variables.

Out of the 29 predictor variables, more than 10 had possibly insignificant coefficients - as in, their 3% and 97% values of Highest Density Intervals had a 0 in between them. From this observation, we chose to create a reduced model with less predictor variables. After omitting predictor variables based on collinearity and "common sense," we ended up with 14 predictor variables. Because there were no binomial variables within the chosen predictors, all sampling was done with NUTS. The resulting posterior (Graph 8) resembled the true distribution of the response variable (Graph 9) slightly more, but it was far from satisfactory.

In an attempt to further improve the model, we attempted to make an ensemble with the full and reduced model through Bayesian Model Averaging (BMA). To effectively average the two models, we calculated the weight with which each model should be measured. The weight was calculated with two metrics: the Widely Applicable Information Criterion (WAIC) and Leave-One-Out cross-validation (LOO). However, with both WAIC and LOO, the reduced model was deemed to be weighed 100%. This meant that an ensemble created from the full and reduced model would simply be the same model as the reduced model. This led us to believe that a BMA would not be a viable method for improving model performance.

Bayesian additive regression trees (BART) were implemented as a method to induce variable selection. BART inherently includes regularization so if there are any nonlinearities amongst the variables, BART will perform variable selection itself (*Bayesian additive regression trees: Introduction* 2022). Another advantage is that it can accurately model complex response surfaces as it has a flexible functional form between the predictors and response (Hill et al., *Annual Review of Statistics and its Application: Bayesian Additive Regression Trees: A Review and Look Forward* 2020). A disadvantage is that it can be computationally intensive when fitting large datasets; however since our data is relatively small at around 300 records, this did not pose an issue for the analysis. Another drawback to this approach is that since BART is a complex, non parametric model that can make use of multiple nonlinear interactions between the predictors and response, it can be difficult to understand the exact relationship between predictors and response variables captured by the model (et al., *Bayesian modeling and computation in python* 2021). Furthermore, it is sensitive to the choice of hyperparameters, which can affect the model's performance thus making it difficult to achieve optimal performance on a dataset.

Results

As discussed in the Probability Model section, the two Bayesian Linear Regression models had poor performance in predicting the response variable *G3*; The posterior distribution of the full model (Graph 7) and the reduced model (Graph 8) barely resembled the true distribution (Graph 9). However, the comparatively high performance of the reduced model led us to believe that the model performance can be further improved with even better feature selection, which would ultimately allow us to explore the uncertainty surrounding a student's demographic characteristics and their effects on the student's academic success.

Following the poor performance of the full and reduced models, the results of the feature selection were visualized through a BART method. A variable importance plot was created (Graph 4). This shows the relative importance of each predictor normalized so that the sum adds up to one. Based on the plot created, the variables that were deemed the most important were failures, absences, and the job category for the mother. The least important variable determined was workday alcohol consumption.

To characterize the uncertainty related to the absences variable, a 94% High Density Interval (HDI) plot was produced (Graph 5). This graph indicates which points cover most of the distribution as both the final grade and absences have been standardized. The darker orange band represents the 50% HDI and the lighter one is the 94% HDI. Here the final grade in the course has a downwards slope up to a certain number of absences but then interestingly appears to increase afterwards. Since the 94% interval is very large, there are some indications that that analysis is conservative and uncertain. Therefore, this may not be the most predictive variable for final grade. A potential drawback to the absences predictor is that it was heavily skewed in the middle indicating that there were either virtually no absences by students or many. Looking at the distribution statistics for each parameter shows that a good mix does not include 0 in the credible interval of 3% HDI to 97% HDI (Table 1). A promising sign is that the r_hats for many of the parameters are 1 indicating convergence in the estimates.

Conclusions

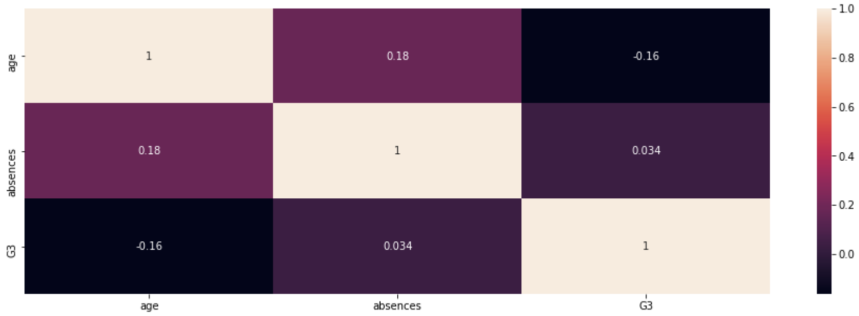
Our goal was achieved in the sense that we were able to measure the uncertainty surrounding the influence of the extrinsically related predictor variables on final grade; however, our model does a poor job accurately predicting those values. Another potential method to use moving forward would be Bayesian model averaging; however, we do not believe this would provide new pertinent information nor would be useful since the BART model is already heavily weighted towards 1 compared to the full and reduced models. Some limitations to the data includes the lack of information regarding the detailed location of where the student's home resident resides. Information like this could reveal whether there are any differences in grades from students who live in rural vs. urban areas and thus have access to varying external resources which could be a potential hierarchical modeling problem. Furthermore, there was skewness towards some values in our predictors, and the data set did not account for other intrinsic factors such as motivation and work ethic in student's grades.

References

1. Martin, Kumar, and Lao, “Bayesian modeling and computation in python,” *7. Bayesian Additive Regression Trees - Bayesian Modeling and Computation in Python*, 2021. [Online]. Available: https://bayesiancomputationbook.com/markdown/chp_07.html. [Accessed: 11-Dec-2022].
2. J. Hill, A. Linero, and J. Murray, “Annual Review of Statistics and its Application: Bayesian Additive Regression Trees: A Review and Look Forward,” *https://par.nsf.gov/servlets/purl/10181031* , 2020.
3. pymc-devs, “Bayesian additive regression trees: Introduction,” *PyMC*, 2022. [Online]. Available: https://www.pymc.io/projects/examples/en/latest/case_studies/BART_introduction.html. [Accessed: 11-Dec-2022].
4. U. C. I. M. Learning, “Student Alcohol Consumption,” *Kaggle*, 19-Oct-2016. [Online]. Available: https://www.kaggle.com/datasets/uciml/student-alcohol-consumption. [Accessed: 11-Dec-2022].

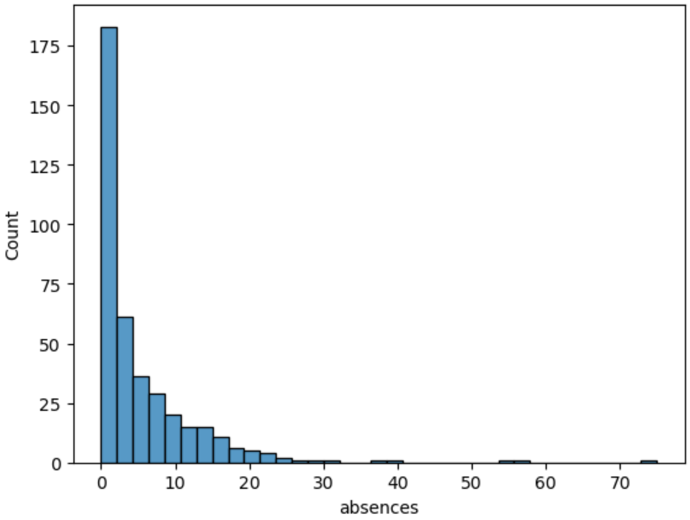
Appendices

Heatmap of Continuous Features



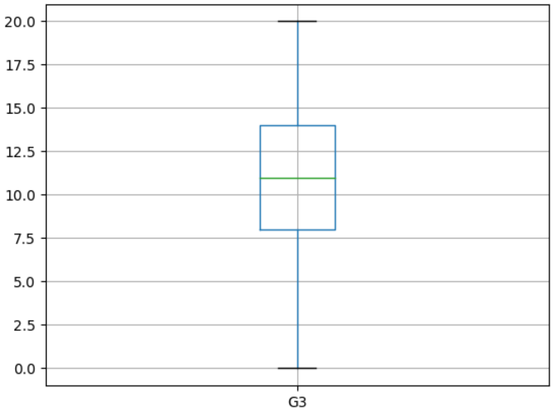
Graph 1

Histogram of Absences



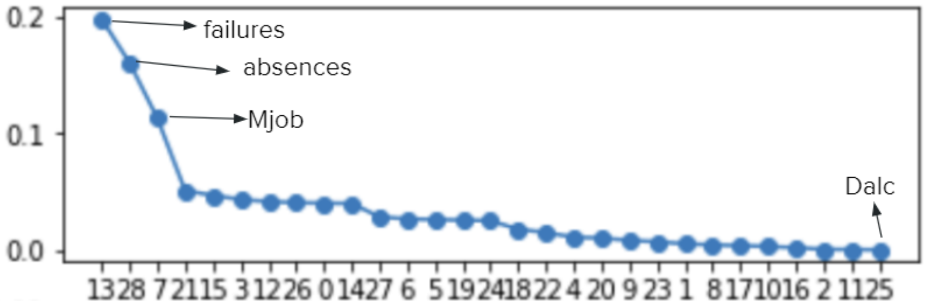
Graph 2

Boxplot of The Outcome Variable: Final Grade



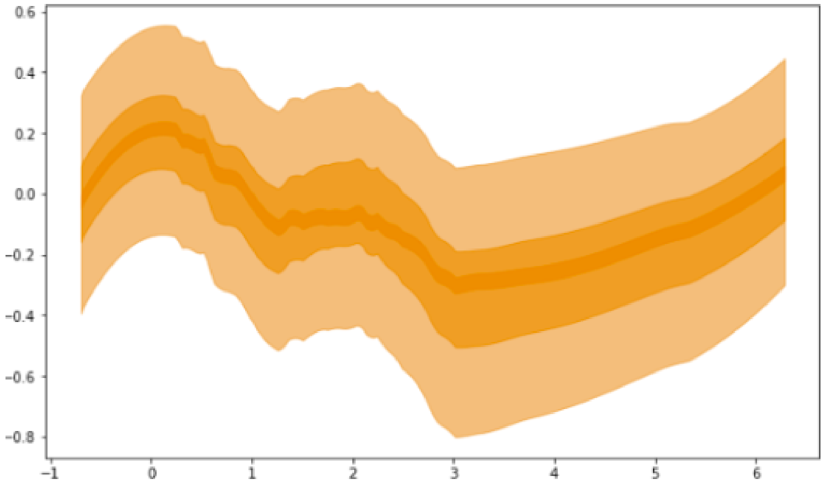
Graph 3

Variable Importance Plot for BART Model



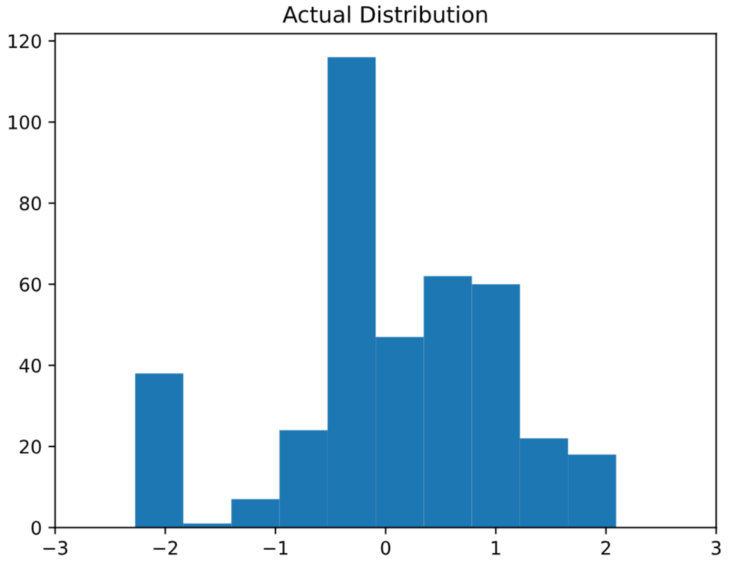
Graph 4

94% HDI for Absences vs. Final Grade



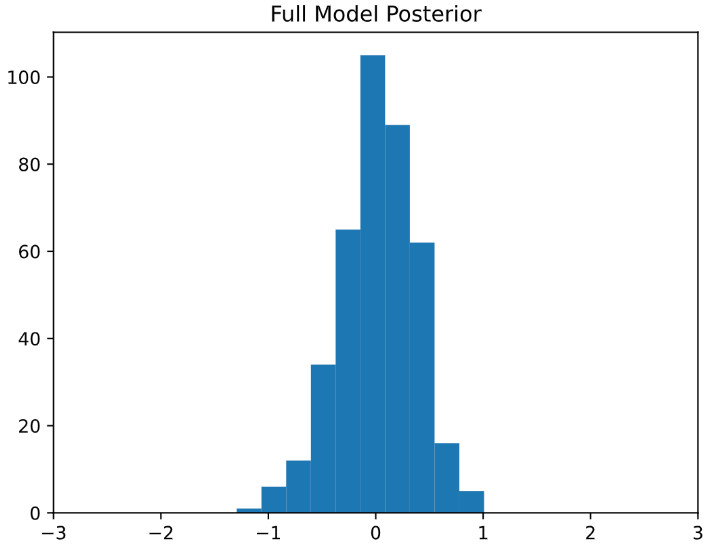
Graph 5

True Distribution of Grade (Scaled) Obtained from the Data



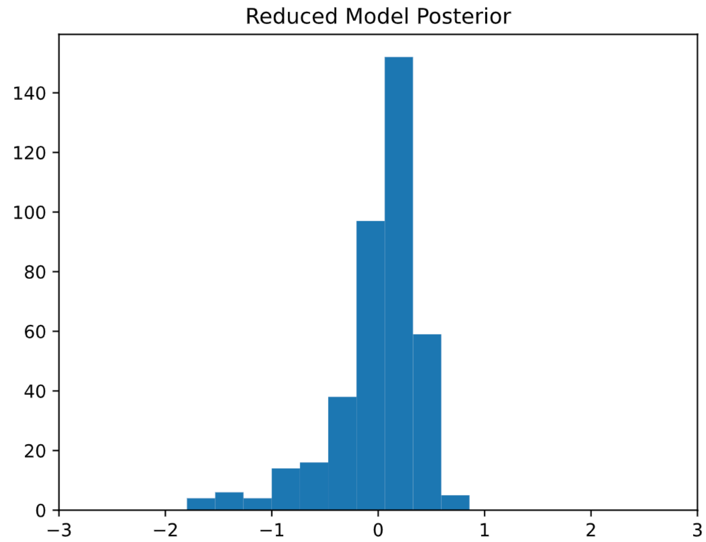
Graph 6

Distribution of Predicted (Scaled) Grades from Full Model



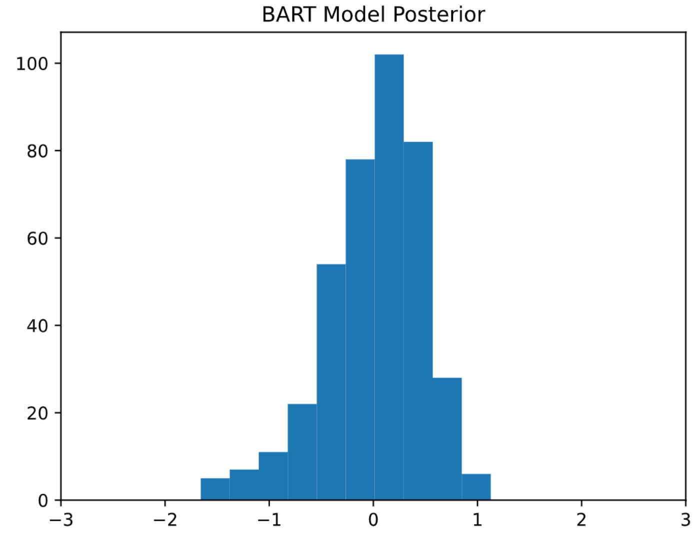
Graph 7

Distribution of Predicted (Scaled) Grades from Reduced Model



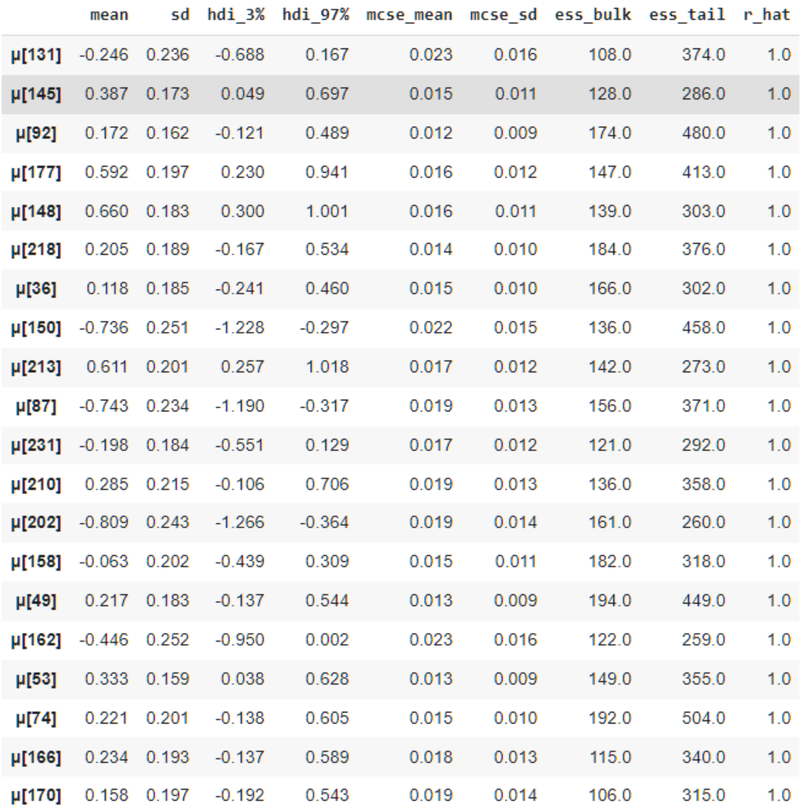
Graph 8

Distribution of Predicted (Scaled) Grades from BART Model



Graph 9

Summary Statistics For Subset of Predictors in BART Model*



	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
μ[131]	-0.246	0.236	-0.688	0.167	0.023	0.016	108.0	374.0	1.0
μ[145]	0.387	0.173	0.049	0.697	0.015	0.011	128.0	286.0	1.0
μ[92]	0.172	0.162	-0.121	0.489	0.012	0.009	174.0	480.0	1.0
μ[177]	0.592	0.197	0.230	0.941	0.016	0.012	147.0	413.0	1.0
μ[148]	0.660	0.183	0.300	1.001	0.016	0.011	139.0	303.0	1.0
μ[218]	0.205	0.189	-0.167	0.534	0.014	0.010	184.0	376.0	1.0
μ[36]	0.118	0.185	-0.241	0.460	0.015	0.010	166.0	302.0	1.0
μ[150]	-0.736	0.251	-1.228	-0.297	0.022	0.015	136.0	458.0	1.0
μ[213]	0.611	0.201	0.257	1.018	0.017	0.012	142.0	273.0	1.0
μ[87]	-0.743	0.234	-1.190	-0.317	0.019	0.013	156.0	371.0	1.0
μ[231]	-0.198	0.184	-0.551	0.129	0.017	0.012	121.0	292.0	1.0
μ[210]	0.285	0.215	-0.106	0.706	0.019	0.013	136.0	358.0	1.0
μ[202]	-0.809	0.243	-1.266	-0.364	0.019	0.014	161.0	260.0	1.0
μ[158]	-0.063	0.202	-0.439	0.309	0.015	0.011	182.0	318.0	1.0
μ[49]	0.217	0.183	-0.137	0.544	0.013	0.009	194.0	449.0	1.0
μ[162]	-0.446	0.252	-0.950	0.002	0.023	0.016	122.0	259.0	1.0
μ[53]	0.333	0.159	0.038	0.628	0.013	0.009	149.0	355.0	1.0
μ[74]	0.221	0.201	-0.138	0.605	0.015	0.010	192.0	504.0	1.0
μ[166]	0.234	0.193	-0.137	0.589	0.018	0.013	115.0	340.0	1.0
μ[170]	0.158	0.197	-0.192	0.543	0.019	0.014	106.0	315.0	1.0

Table 1

*Due to the large number of predictors, every statistic for each predictor is not included. Instead I've attached a snapshot of the subset of predictors that converge to one with their credible intervals.