Predicting Student Success

DS6040 Final Project – Fall 2023 – UVA School of Data Science

Lillian Jarrett, Will Tenpas, Stephen Kullman

Project Repository: Github Link

Problem Description

Education is one of the most important indicators of future earnings[1]. Each degree completed increases future earnings, health outcomes, and financial security. Despite this, education spending in Portugal is decreasing. With less resources devoted to such an important service, it is critical that educators and administrators know how to use these resources to maximize student success.

Thus, the objective of this analysis is to uncover how certain factors influence student success for high school students in Portugal. The goal is to find the most important influencers to academic success, so that educators can implement efficient interventions. To do this, we will use a dataset on student demographics and behaviors to analyze the effect of different predictors on final grade. By understanding the complex relationship between response and predictor, we attempt to contribute to a more informed approach to school spending and educational outcomes.

Data Description

The chosen dataset comes from Kaggle and includes observations from two Portuguese schools coded as GP and MS. The data attributes include student grades, demographic, social, and school related features and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics and Portuguese. It includes both numeric, ordinal, and categorical data. All told, there are over 1000 observations and 33 predictor variables. The response variable is final_grade. Portugal education is split up into trimesters, so the dataset included exam grades for the first and second trimesters, named grade_1 and grade_2, as well.

The first step was data processing. There were no null values in the dataset, but some categorical predictors needed to be encoded. The *Yes/No* questions were remapped to a binary outcome, while the categorical predictors were one-hot encoded. This left us with 64 predictor variables. Because some predictors had wildly different scales, we z-transformed the predictors to ensure each variable was on an equal playing field.

Second, we performed exploratory data analysis. Looking at the numeric variables first, we created a correlation heatmap to see which predictors are correlated. Unsurprisingly, grade_1, grade_2, and final_grade are the most correlated (Figure 1). Class failures were negatively correlated. For the categorical variables, we created bar charts to see if there was a distinct difference in the final grades between the categories. Two indicators of note were internet access and whether the student planned on pursuing college (Figure 2). These findings informed our next steps to building a model, and gave us a starting point to look at significant predictors.

[1] U.S Bureau of Labor Statistics, 2020.

Model Description

Model Goals

Our team sought to conduct a Bayesian regression analysis to build a model that could predict student final grades. If it was successful at prediction, we could also rely on the model as being a successful indicator of which predictors are most important. It was also critical that the model output uncertainty measures to allow us more nuance in our recommendations.

Prior Selection

With this in mind, we started with the model-building process, working our way through several different models and comparing them to see what would be the best fit for the data. The packages PyMC3 and ArviZ were used to accomplish the modeling. The goal was to build two models, the first being a model with all predictors included, and the second a reduced model with the predictors we found most significant.

To select our priors, we first split our predictors into continuous and categorical predictors. Since the continuous variables had more variable distributions and more outliers, we used Student-T priors with uninformative parameters and large tails. For the categorical variables, we used uninformative normal priors with small sigma values since their scale was quite small. We could not use Bernoulli priors because we sought to use variational inference, which relies on a gradient.

Prior Verification

To check if these priors were appropriate, we ran prior predictive checks. (Figure 4). The priors are a little bit skewed to the left of the observed data, but close enough to continue the analysis and let the likelihoods update them.

Posterior Creation

To create the posterior, we used variational inference techniques. The most successful sampling method we used was ADVI. This led to strong convergence (Figure 5) and a strong ELBO plot (Figure 3, left). These VI techniques were more successful than running HMC traces which may be due to the complexity of the probability distribution, which took in 64 predictor variables after encoding.

Posterior Verification

The posterior predictive checks showed us that the model is doing a decent job of approximating the posterior distribution (Figure 5) You can see it underpredicts test scores in part of the domain but does a good job otherwise. From this, we felt confident enough to use the model to make predictions, and interpret the model to understand the most significant predictors.

Results

Model Predictive Results

To identify significant predictors, we analyzed the posterior plots. These plots and the highest density intervals indicated nine influential factors in predicting final grade. These predictors included grade_2, travel_time_greater_than_1_hour, school_MS (going to one of the two schools),

mother_education_none, father_job_other, father_education_primary_education_4th_grade, father_education_none, sex_M, and parent_status_Apart. Travel time seemed to be the most significant according to the posterior plots (Figure 8).

The Portuguese class dataset was used as our test set, which we cleaned and standardized the data with the same techniques as with the training set (Math class). We extracted and applied the mean and highest density intervals (for predictive uncertainty) to the test data. See Figure 7, which compares the predicted final grades (blue line) versus the actual final grades (red scatter). The HDI lower and upper bounds are plotted as well (yellow band) to show prediction error and uncertainty in predictions. The full model demonstrated better predictive accuracy. The reduced model followed similar trends, but underpredict final grades in this Portuguese class overall. While this model is significantly less complex, we need to make some adjustments to improve the model's predictive accuracy. The adjustments to fix this would include tuning the likelihood so that it has a higher mean. Both of these models follow the patterns in the data quite well, just with some slight offsets.

Model Significance Results

We evaluated and compared the full and reduced model performances using widely available information criteria and leave-one-out (Figure 6). The WAIC is the Bayesian extension of AIC, computed using the entire posterior distribution. The WAIC is helpful for generalizability. It takes into account the uncertainty in the posterior by taking the mean over all of the samples from the posterior. For both WAIC and LOO, the reduced is ranked higher than the full model. Since the effective logwise point density is more negative for the full model (edlp_waic = -1161.16; -1118.0), it suggests that the full model has better predictive accuracy. However, the effective number of parameters is significantly less for the reduced model than the full model (p_waic = 253.64; 27.76). An important tradeoff is finding the balance between model fit and complexity.

Discussion

While we accomplished our goal of the analysis, future work could improve the model. The first potential improvement is the idea of creating a hierarchical model or pooled model. While this seems attractive initially, upon further examination of the data and project there are two main pitfalls to this. The first is fundamental to our goal for the project, which is to find general indicators of student success that educators across Portugal can use. While a hierarchical model may provide more accuracy for the schools in this study, the goal of the project is not to create individualized accurate models. If we had data for every school in Portugal, that could be incredibly useful, but with our data set and goals, a more generalized result is more useful. Second, a pooled model would be subject to issues that come with the unbalanced dataset. There is far more data in the GP school than the MS school, so creating pools would likely hurt the accuracy of the model for the MS school.

The second main improvement would be to shift our likelihood priors up. Since our model routinely underpredicted final grade, we could shift the likelihood mean higher.

One limitation of our analysis was the small dataset. Having data from only two schools for two classes limits the usefulness of our model outside of that context. It not only could introduce bias or inaccurate relationships, it is hard to know if certain predictors are even significant across classes. More

data would reduce our uncertainties for those predictors and give a more holistic view to school administrators.

Conclusions

Based on our results, we are able to make several recommendations to satisfy our initial goal.

Recommendations

Based on the significant predictors we found, we suggest a three pronged intervention. The first is to implement demographic interventions. These interventions should be based on if the student meets certain demographic criteria- like distance from school, divorced parents, or low parental education. These interventions could involve implementing preemptive learning plans. The second prong involves implementing after school enrichment to account for differing levels of enrichment being given at home. It is critical that transportation is provided to these, so that students that live far from school can still attend them. The third prong is to implement mid year interventions for students that are struggling after the first two trimesters. These two indicators were massive indicators of future struggles, so additional resources being devoted to students who struggle early in the year would be well spent.

Overall, our model was able to predict final grade relatively accurately and able to identify the most significant predictors to final grade.

Appendix

Figure 1: Correlation Heatmap

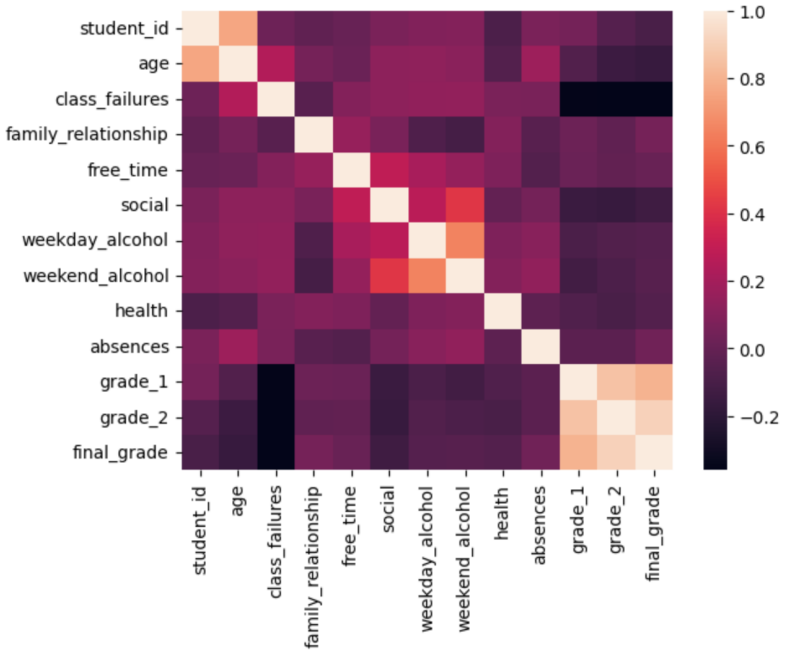


Figure 2: Select Categorical Variable Plots

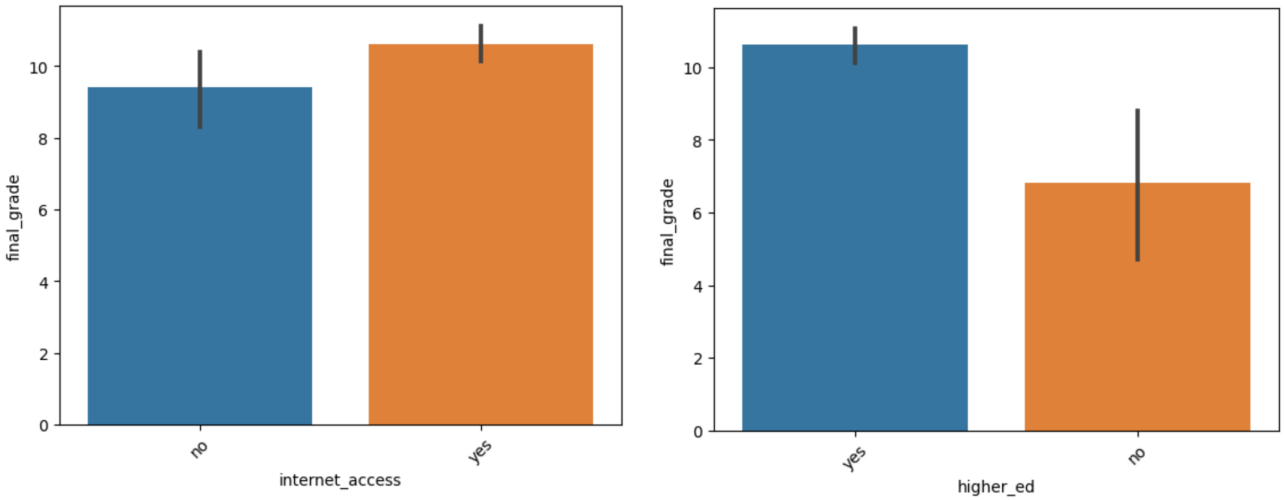


Figure 3: Model Diagnostics (ADVI)

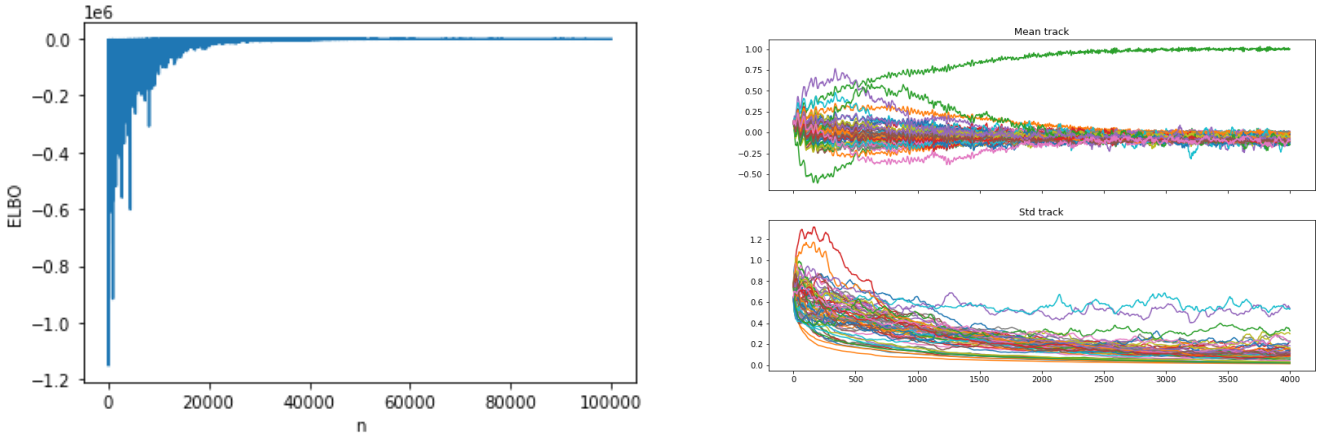


Figure 4: Prior Predictive Checks

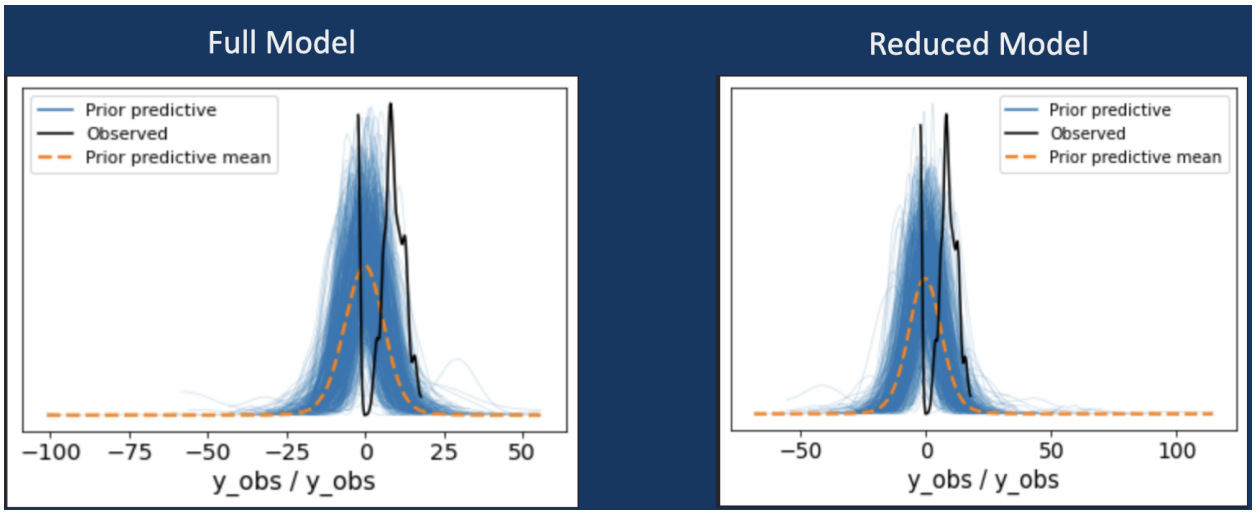


Figure 5: Posterior Predictive Checks

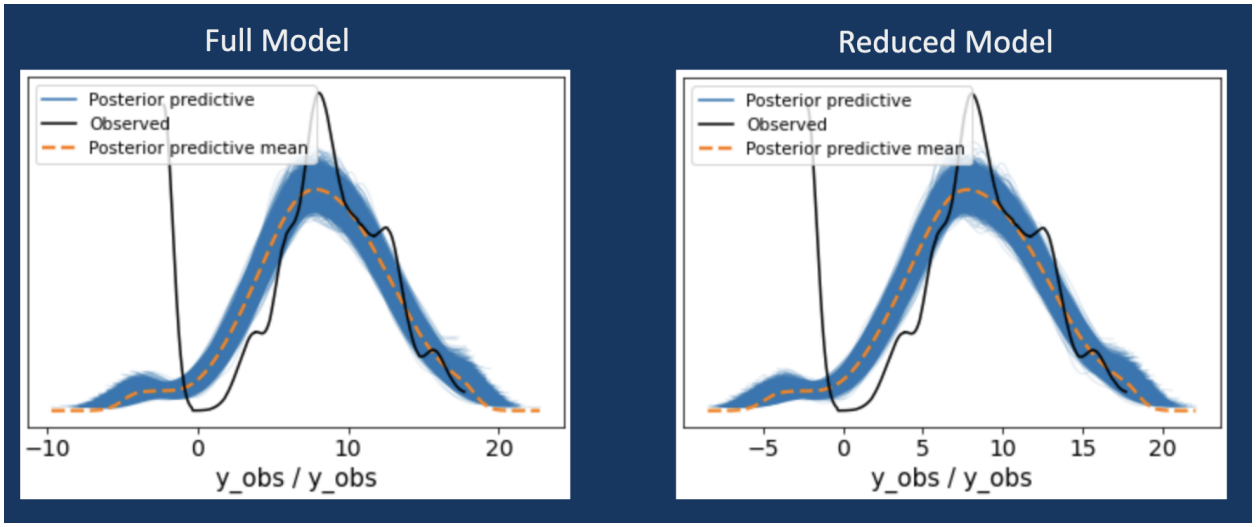


Figure 6: Model Comparison

	rank	elpd_waic	p_waic	elpd_diff	weight	se	dse	warning	scale
Reduced	0	-1118.163558	27.759369	0.000000	0.749205	119.863481	0.00000	True	log
Full	1	-1161.998453	253.636978	43.834895	0.250795	98.802761	38.40039	True	log

	rank	elpd_loo	p_loo	elpd_diff	weight	se	dse	warning	scale
Reduced	0	-1118.443958	28.039769	0.000000	0.755852	119.872946	0.000000	False	log
Full	1	-1165.429258	257.067783	46.985301	0.244148	98.102405	39.007069	True	log

Figure 7: Prediction Plots

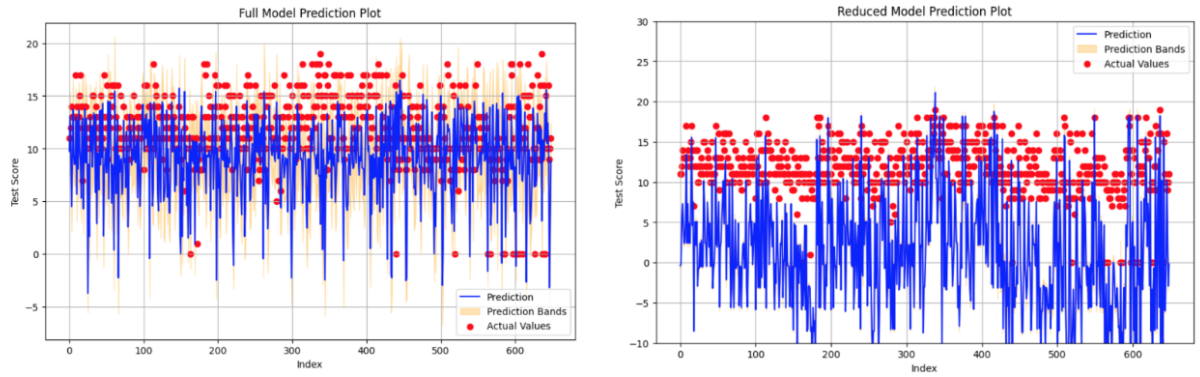
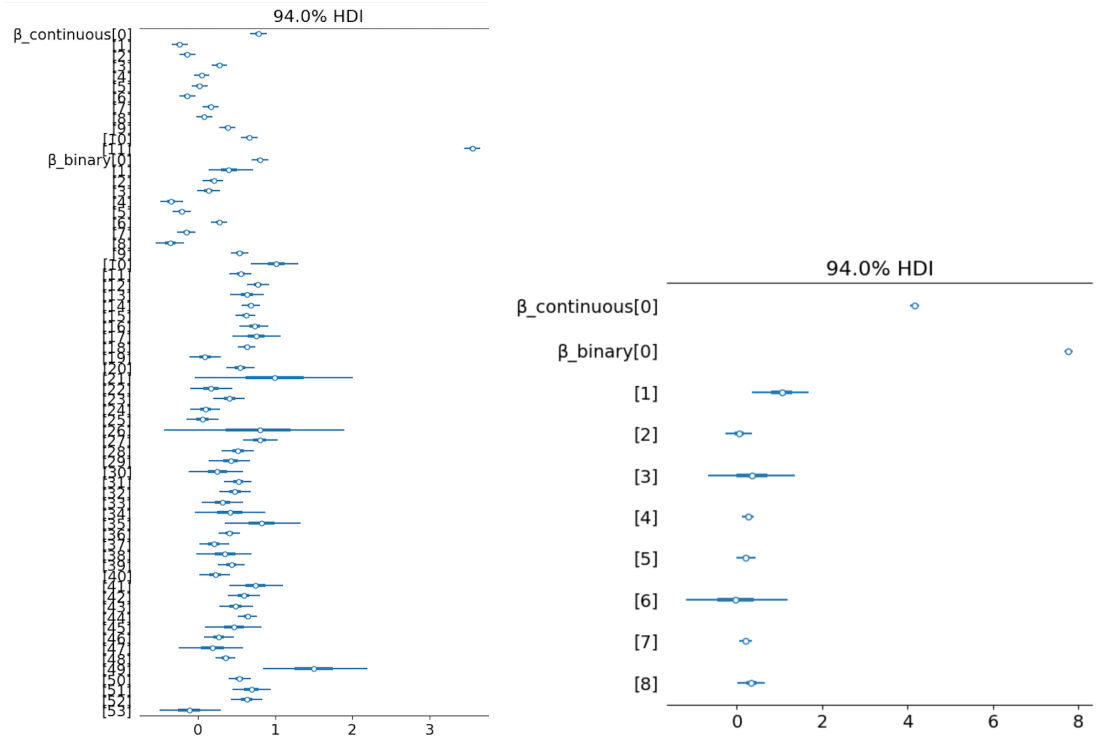


Figure 8: Forest Models



Citations

Myrick, D. (2023, November 10). High School Student Performance & Demographics. Kaggle. https://www.kaggle.com/datasets/dillonmyrick/high-school-student-performance-and-demographics

Portugal education spending 1987-2023. MacroTrends. (n.d.-a). https://www.macrotrends.net/countries/PRT/portugal/education-spending

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

U.S. Bureau of Labor Statistics. (n.d.). Education pays, 2020 : Career outlook. U.S. Bureau of Labor Statistics. https://www.bls.gov/careeroutlook/2021/data-on-display/education-pays.htm