Model Complexity and Racial Biases in Fragile Families Challenge Predictions

Alexis Sursock
Princeton University
asursock@princeton.edu

Tiffany Huang
Princeton University
twhuang@princeton.edu

Abstract

The Fragile Families Challenge [1] offers scientists an opportunity to try to predict life outcomes from a data set of over 4,000 children. This paper is an exploration of which machine learning models perform best on the data set, and which features are most predictive of outcome. We expand upon prior scholars' research and find that the gradient boosting and ensemble models perform best, though all models perform similarly overall. We also find that our gradient boost model performs worse for Black and Hispanic children, and that Black children outperform GPA predictions, perhaps indicating a racial bias in the survey and predictions.

1 Introduction

Much of social science literature surrounds the concept of predicting and understanding social patterns and life outcomes. One such study is the Fragile Families and Child Wellbeing Study, an ongoing study collecting longitudinal data on families with unmarried parents and children born around the year 2000 [1]. In 2017, the Fragile Families Challenge brought together hundreds of scientists to build predictive models for six life outcomes from the study. The results found that even the best predictions, made with the most complex and advanced machine learning methods, were not more accurate than those made by a simple baseline model composed of only a few features. This paper represents our attempt to build upon the results of the challenge [1]. In addition, given growing awareness of racial biases present in social science and academia at-large, we are curious as to the models' prediction accuracies for participants of color.

Here, we review the effectiveness of seven machine learning algorithms' predictions of three continuous life outcomes: "GPA", "grit", and "material hardship". We perform feature selection to identify and keep the top 30% of features that we believe best contribute to outcome predictions. We find that gradient boosting, an ensemble classifier, and the neural network perform best, but are not substantially better than other models we evaluated. We also identify which features are most predictive for each work and find that grit and material hardship share many of the same top features. We also explore what we call "moonshot" children, or children who outperform their predicted outcomes, and find some racial bias in the predictions here. Finally, we find that our predictions see higher error on Black and Hispanic children, potentially highlighting an area for further research.

2 Related Work

This data set represents an encoding of thousands of interviews done with Fragile Families participants over the course of 20 years. Each variable represents a particular interview question and answer (which is represented numerically). Missing answers are represented by negative numbers or NAs. Each family is represented by a "challenge ID"; there are 4,242 families, with 2,121 of them having life outcome data available for challenge participants. The models in the paper were trained on this half of the data, and then evaluated on their predictions for the other half.

We are inspired by prior scholars' work on this challenge; we point to Davidson's success using neural networks to predict "GPA" [2] and Rigobon et al.'s use of gradient-boosting trees in an ensemble method to successfully predict "GPA", "grit", and "layoff" [3] as key motivators. We were also inspired by findings that simple baseline models consisting of linear regression from just a few features perform just as, if not better, than more complex models, and aim to verify this here [1].

While the data set also includes categorical outcomes, we chose to focus our work on predicting the continuous outcomes due to the above scholars' success with regression models. Further explanation on the three continuous outcomes and how they were derived can be found on the Fragile Families website [4].

2.1 Data preprocessing and imputation

We were given a data set containing 13,026 variables and 4,242 families. However, many of the entries of the data set were missing or unusable. We began data preprocessing by removing all features that contain string values, and then replacing all negative values (these are missing values) with "NaNs". We then removed any feature columns that had more than 85% of entries missing (we also tested thresholds of 50% and 70% and found that 85% was best for reducing error), and filled in remaining NaNs with the mode of their respective column. We also removed features that have a variance less than 0.1. Our end result is a data set with 3,687 variables remaining (about 30% of the original data set.

2.2 Regression methods

All of the below methods were optimized using GridSearchCV [5] with several folds of cross validation to tune relevant hyperparameters. Here we specify key information on each model as well as our motivation for using them.

1. Neural Network: used a sequential model built on the Keras and TensorFlow libraries, trained on 100 epochs [6]. Motivated by prior literature [2].
2. Linear Regression on 5 features: performed solely on five features identified as most important by our XGBoost model, forming our baseline model. Motivated by prior literature [1].
3. Linear Regression: performed on data set with all features to compare to linear regression with only a few features.
4. LASSO: we employ this to further reduce our high number of variables, hopefully optimizing performance compared to standard linear regression.
5. Gradient Boosting: Explained further in spotlight section. Motivated by prior literature [3].
6. XGBoost: an optimized version of gradient boosting using the "gbtree" model; identified as being effective by prior literature [3]
7. Ensemble Classifier: combines LASSO, linear regression, neural network, and XGBoost models (weighted equally); motivated by others' success with ensemble models [3].

2.3 Evaluation

We evaluate our models using the mean squared error metric on predictions of the continuous outcomes. We note that this is the only metric provided to us by the Fragile Families Challenge submission site. In our examination of racial biases in the predictions, we randomly split our training set into a new train and validation set, and again evaluate using several additional metrics on the split.

We define an ideal regressor as one that has low mean squared error on the three continuous outcomes. We expect that the neural network and the gradient-boosting/XGBoost models will perform the best based on prior literature cited above. Rigobon et al. found that the most significant features for GPA, grit, and material hardship are those relating to standardized test scores, child demeanor and environment, and length of interview (respectively); we expect similar findings. Of the three variables, we expect grit to be the hardest to predict because it is more abstractly defined than the others, and because the causes of grit are still unknown [7].

3 Methods

3.1 Spotlight Algorithm: Gradient Boosting

Gradient boosting is derived from the concept of boosting, or modifying weak learners (classification/regression algorithms that perform only slightly better than random chance [8]) to be better. The first boosting algorithm made was AdaBoost; here the weak learners are decision trees, which formulate training data information into a hierarchical structure that is easily understood [9]. The model makes predictions based on an ensemble model that is calculates the weighted sum of predictions made by each tree [10]. Gradient boosting can be seen as a generalization of AdaBoost where "the objective is to minimize the loss of the model by adding weak learners using a gradient descent-like procedure" [8].

We begin by minimizing a loss function (any differentiable loss function may be used; for example, linear regression commonly uses the mean squared error function[8]). In gradient boosting, decision trees are still used as the weak learners. We then mimic gradient descent to move in the direction that minimizes residual loss when adding new weak learners [8]. The below figure explains this procedure mathematically:

Algorithm 16.4: Gradient boosting

1 Initialize $f_0(\mathbf{x}) = \text{argmin}_\gamma \sum_{i=1}^{N} L(y_i, \phi(\mathbf{x}_i; \boldsymbol{\gamma}))$;
2 **for** $m = 1 : M$ **do**
3 Compute the gradient residual using $r_{im} = -\left[\frac{\partial L(y_i, f(\mathbf{x}_i))}{\partial f(\mathbf{x}_i)}\right]_{f(\mathbf{x}_i)=f_{m-1}(\mathbf{x}_i)}$;
4 Use the weak learner to compute $\boldsymbol{\gamma}_m$ which minimizes $\sum_{i=1}^{N}(r_{im} - \phi(\mathbf{x}_i; \boldsymbol{\gamma}_m))^2$;
5 Update $f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \nu\phi(\mathbf{x}; \boldsymbol{\gamma}_m)$;
6 Return $f(\mathbf{x}) = f_M(\mathbf{x})$

Figure 1: the steps of gradient boosting, courtesy of Kevin Murphy's "Machine Learning: A Probabilistic Perspective" [11].

One version of gradient boosting is called XGBoost, which is "an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable" [12]. XGBoost employs decision tree ensembles that contain a set of classification and regression trees (CART) [13]. CARTs differ from decision trees because each leaf contains a prediction score rather than just a decision value, which allows for a more unified approach to optimization [13]. Another key component of XGBoost is that our objective function that we aim to minimize contains both a loss function and a regularization term; this prevents overfitting and improves performance [13]. We employ both regular gradient boosting and XGBoost in our work.

Because we are unable to assess all possible tree combinations within our model, gradient boosting optimizes one level at a time by splitting each leaf in two and assessing if doing so improves our score [13]. However, this can fail for some edge cases, as we can only consider one feature dimension at a time [13]. Here, we assume that our data does not contain these edge cases (i.e. situations where the dataset has perfect symmetry and the model isn't able to split properly; one such situation is outlined here: [14]).

In our work, we train our gradient boosting and XGBoost models on our training set (this is half of our total dataset, and without the features that we have removed because of high missingness or low variance). We employ GridSearchCV [5], a tuning technique that optimizes hyperparameters through 5-fold cross validation, in order to identify the best parameter values for our model (example parameters include learning rate and number of estimators). We have a separate model (with separate parameters) for each outcome we predict.

4 Results

4.1 Model Evaluation and Result Analysis

We find that our best-performing models are the gradient boosting (for GPA) and the ensemble regressor (for grit and material hardship) (see Table 1). By summing each model's MSE scores across all three outcomes, we find that the best overall model is the gradient boosting regressor, closely followed by the neural network (see Table 5 in Appendix). We find that the worst model overall is LASSO; however, it is important to note that the differences in mean squared error between models are not extremely substantial and that the models perform similarly overall.

Models	GPA	Grit	Material Hardship
Neural Network	0.380	0.220	0.029
Linear Regression	0.393	0.226	0.026
Linear Regression (5 features)	0.389	0.222	0.028
Lasso	0.402	0.223	0.027
Gradient Boosting	**0.375**	0.220	0.029
XGboost	0.379	0.228	0.026
Ensemble	0.380	**0.219**	**0.025**
Best Performance	**0.375**	**0.219**	**0.025**

Table 1: A comparison of mean squared error for selected classifiers (table format source: [15]

We also find that linear regression on just the five best features performs better on GPA and Grit than linear regression on all features, which confirms prior findings that simple baseline models can be more effective at predicting on this dataset [1]. Surprisingly, the XGBoost regressor performs worse overall than the gradient boosting regressor; the difference in performance is likely negligible and further research is required to elucidate this.

It is clear that predictions for material hardship have significantly lower error values for all models. We note that material hardship had far less variance than the grit and GPA (0.023 compared to 0.225 and 0.429, respectively), which explains its lower error values. While we initially predicted that grit would be the most difficult to predict, in reality the outcome with the most mean squared error is GPA; this is also the outcome with the highest variance, so it is difficult to say if this outcome was truly harder to predict or not (further exploration is required here).

4.2 Significant Features

To identify the most significant features in predicting each outcome, we employ XGBoost's feature importance function to calculate each feature's F-score, which represents how many times a split was made on that particular feature [16]. We then plot the top five features and their F-score (see Appendix). For GPA, we find that the top two features are the child's scores on the Woodstock Johnson test (which tests cognitive ability); this is as predicted. Interestingly, we find that the fourth-most important feature for GPA prediction is a child's Body-Mass-Index, which perhaps confirms prior findings that obesity is linked with lower GPA [17].

Additionally, we find that the top five best features for grit and material hardship share three variables in common: household income, monthly amount of money paid for groceries, and biological mother's height. Given that grit is defined as perseverance in the face of hardship, it makes sense that the two outcomes can be predicted by similar features. These identified features only align somewhat with prior scholars' findings [3], indicating that different models may identify different "top" features.

Finally, we find that even though among the features that remain after removing 70% of our initial variables, there are still several that stand out in feature importance (as calculated by XGBoost); see Figure 5 in the Appendix. This seems to indicate that further feature selection may benefit model accuracy (at least for XGBoost).

4.3 Racial Bias and Residual Analysis

Using our spotlight algorithm, XGBRegressor, we evaluated the racial bias of our model and analyzed the residuals for different racial and ethnic groups using different validation sets. We found that Hispanics and African Americans, although representing the majority of the dataset, were disproportionately difficult to predict with regards to GPA and Material Hardship. Asian Americans were the group with the highest predicted values, and highest error (possibly due to their lack of representation in the data set), and the only group whose predicted GPA exceeded actual grades 2 3 4. This can be potentially explained by prior research on complex racial differences in marital outcomes and variations in the "composition and stability of fragile families"[18] [19]. Additional findings can be found in the Appendix 6.2.

GPA	Hispanic	White	Black	Asian
Avg Signed Error	0.0793	0.0337	0.0839	-0.1105
Avg Squared Error	0.3819	0.2937	0.4169	0.8797
Avg Prediction	2.8397	3.1650	2.7560	3.2355
Avg in Validation	2.9191	3.1987	2.8400	3.1250

Table 2: A comparison of average metrics for different racial and ethnic backgrounds. Note: White, African American, and Asian American specifically represent the non-Hispanic subgroups of said races. Average signed error refers to the predicted outcome minus the actual outcome, summed over each child.

4.4 Outlier Children Analysis

We performed additional tests to identify what we call "moonshots", or gifted children that beat the odds and outperformed our GPA predictions by over one point. Using our XGBRegressor model, we found that nearly 62% of all moonshots are of African-American descent, a significantly disproportionate amount. Interestingly over 19% of moonshots do not identify themselves as any of the 4 race options, White, Black, Asian, and American Indian. This further supports the idea that racial bias may have been present amongst those conducting the survey, especially given that many variable values were constructed by the scientists themselves, allowing room for biased interpretations. Future researchers are encouraged to delve deeper into this group of outliers to identify possible reasons to their astonishing accomplishments, and help others replicate their success stories.

5 Discussion and Conclusion

This paper's main findings relate to both the efficacy of different models, and the racial disparities in the accuracy of our models, in predicting three continuous outcome variables: GPA, grit, and material hardship. Regarding the former, we find that complex models, such as neural networks, XGBoost, and ensemble classifiers, barely, if it all, outperform more basic models such as linear regression and LASSO. We also find that optimized versions of some models do not outperform their unoptimized counterparts (see XGBoost vs. gradient boosting and LASSO vs. linear regression). Furthermore, by analyzing the data set's most significant features, we find that a handful of them are much more significant and predictive for our outcome variables. We verified this hypothesis by training a Linear Regression model solely on the 5 most significant features for every outcome, which ended up outperforming the Linear Regression model trained on the full data set in two out of three outcome variables. We suggest that further feature selection and a "less is more" approach is key to improving performance.

Additionally, we analyzed the racial disparities of residuals of our XGBoost model. We found that the model's predictive power varies depending on the child's race and ethnicity. Similarly, we also found that the children which strongly outperform our model's GPA expectations tend to be Black. We recommend that future researchers analyze more closely these "moonshot" children to gather further insights on children who "beat the odds".

Acknowledgments

We want to thank Dr. Li for guiding us to Princeton Research Computing to run our Neural Network on the Adroit Server. We also want to acknowledge Siena Dumas Ang, Deniz Oktay, Yaniv Ovadia, aAndy Jones, and Sulin Liu for helping us in Office Hours, on Ed, and via Email.

References

[1] M. J. Salganik, I. Lundberg, A. T. Kindel, C. E. Ahearn, K. Al-Ghoneim, A. Almaatouq, D. M. Altschul, J. E. Brand, N. B. Carnegie, R. J. Compton, D. Datta, T. Davidson, A. Filippova, C. Gilroy, B. J. Goode, E. Jahani, R. Kashyap, A. Kirchner, S. McKay, A. C. Morgan, A. Pentland, K. Polimis, L. Raes, D. E. Rigobon, C. V. Roberts, D. M. Stanescu, Y. Suhara, A. Usmani, E. H. Wang, M. Adem, A. Alhajri, B. AlShebli, R. Amin, R. B. Amos, L. P. Argyle, L. Baer-Bositis, M. Büchi, B.-R. Chung, W. Eggert, G. Faletto, Z. Fan, J. Freese, T. Gadgil, J. Gagné, Y. Gao, A. Halpern-Manners, S. P. Hashim, S. Hausen, G. He, K. Higuera, B. Hogan, I. M. Horwitz, L. M. Hummel, N. Jain, K. Jin, D. Jurgens, P. Kaminski, A. Karapetyan, E. H. Kim, B. Leizman, N. Liu, M. Möser, A. E. Mack, M. Mahajan, N. Mandell, H. Marahrens, D. Mercado-Garcia, V. Mocz, K. Mueller-Gastell, A. Musse, Q. Niu, W. Nowak, H. Omidvar, A. Or, K. Ouyang, K. M. Pinto, E. Porter, K. E. Porter, C. Qian, T. Rauf, A. Sargsyan, T. Schaffner, L. Schnabel, B. Schonfeld, B. Sender, J. D. Tang, E. Tsurkov, A. van Loon, O. Varol, X. Wang, Z. Wang, J. Wang, F. Wang, S. Weissman, K. Whitaker, M. K. Wolters, W. L. Woon, J. Wu, C. Wu, K. Yang, J. Yin, B. Zhao, C. Zhu, J. Brooks-Gunn, B. E. Engelhardt, M. Hardt, D. Knox, K. Levy, A. Narayanan, B. M. Stewart, D. J. Watts, and S. McLanahan, "Measuring the predictability of life outcomes with a scientific mass collaboration," *Proceedings of the National Academy of Sciences*, vol. 117, no. 15, pp. 8398–8403, 2020. [Online]. Available: https://www.pnas.org/content/117/15/8398

[2] T. Davidson, "Black box models and sociological explanations: Predicting gpa using neural networks," in *SocArXiv*, 2017.

[3] D. Rigobon, E. Jahani, Y. Suhara, K. AlGhoneim, A. Alghunaim, A. Pentland, and A. Almaatouq, "Winning models for grade point average, grit, and layoff in the fragile families challenge," *Socius: Sociological Research for a Dynamic World*, vol. 5, p. 237802311882041, 01 2019.

[4] I. Lundberg. Fragile families blog posts. https://www.fragilefamilieschallenge.org/blog-posts/.

[5] Gridsearchcv. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

[6] F. Chollet. Keras. https://github.com/fchollet/keras.

[7] I. Lundberg. Grit. https://www.fragilefamilieschallenge.org/grit/.

[8] J. Brownlee. A gentle introduction to the gradient boosting algorithm for machine learning. https://machinelearningmastery.com/ gentle-introduction-gradient-boosting-algorithm-machine-learning/.

[9] G. Seif. A guide to decision trees for machine learning and data science. https: //www.kdnuggets.com/2018/12/guide-decision-trees-machine-learning-data-science.html.

[10] H. Singh. Understanding gradient boosting machines. https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab.

[11] K. P. Murphy, *Machine Learning*. MIT Press, 2012.

[12] Xgboost: Scalable and flexible gradient boosting. https://xgboost.ai/about.

[13] Introduction to boosted trees. https://xgboost.readthedocs.io/en/latest/tutorials/model.html.

[14] M. Filho. Can gradient boosting learn simple arithmetic? https://www.mariofilho.com/can-gradient-boosting-learn-simple-arithmetic/.

[15] B. E. Engelhardt, "Fast classification of newsgroup posts," 2021.

[16] M. Filho. F-score. https://en.wikipedia.org/wiki/F-score.

[17] A. R. Branigan, "(how) does obesity harm academic performance? stratification at the intersection of race, sex, and body size in elementary and high school." *Sociology of Education*, 2017.

[18] R. A. Hummer and E. R. Hamilton, “Race and ethnicity in fragile families,” 2010.

[19] C. Ellerbe, ““racial differences in marital outcomes among unmarried mothers: The influence of perceived marital benefits and expectations”,” 2018.

6 Appendix

6.1 Racial Bias and Residuals for Grit and Material Hardship

GRIT	Hispanic	White	Black	Asian
Avg Signed Error	0.0573	0.0508	-0.0297	-0.1272
Avg Squared Error	0.1660	0.2452	0.2239	0.0518
Avg Prediction	3.4426	3.3402	3.4717	3.3772
Avg in Validation	3.500	3.3910	3.4420	3.2500

Table 3: A comparison of average metrics for different racial and ethnic backgrounds. Note: White, African American, and Asian American specifically represent the non-Hispanic subgroups of said races.

MATERIAL HARDSHIP	Hispanic	White	Black	Asian
Avg Signed Error	0.01637	-0.0018	0.0283	-0.0463
Avg Squared Error	0.0263	0.0128	0.0288	0.0053
Avg Prediction	0.0932	0.0600	0.1156	0.0463
Avg in Validation	0.1096	0.0582	0.1440	0.0000

Table 4: A comparison of average metrics for different racial and ethnic backgrounds. Note: White, African American, and Asian American specifically represent the non-Hispanic subgroups of said races.

6.2 Significant Features Across All Outcome Variables

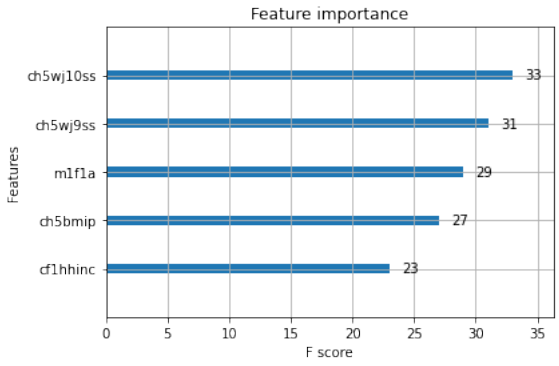

Figure 2: GPA: 5 most important features and their F score, evaluated using our XGBoost model. Top features: ”ch5wj10ss”: Woodcock Johnson Test 10 standard score; ”ch5wj9ss”: Woodcock Johnson Test 9 standard score; ”m1f1a”: How long have you lived in neighborhood - Years; ”ch5bmip”: Child’s Body Mass Index percentile; ”cf1hhinc”: Constructed - Household income (with imputed values)

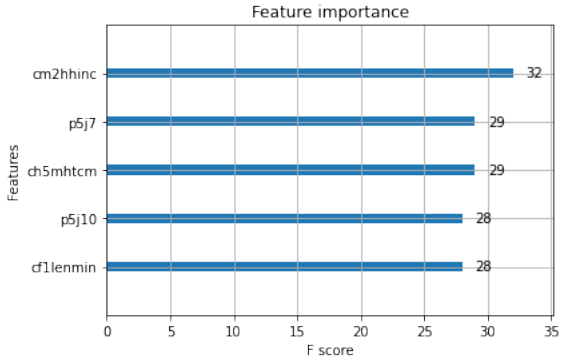

Figure 3: Grit: 5 most important features and their F score, evaluated using our XGBoost model. Top features: ”cm2hhinc”: Constructed - Household income (with imputed values); ”p5j7”: Amount paid for groceries or food used at home in last month; ”ch5mhtcm”: BioMom’s height in cm; ”p5j10”: Amount of money spent eating out in last month; ”cf1lenmin”: What was the total length of interview - Minutes

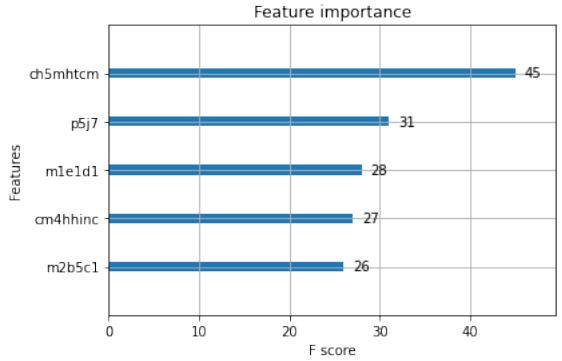

Figure 4: Material Hardship: 5 most important features and their F score, evaluated using our XGBoost model. Top features: ”ch5mhtcm”: BioMom’s height in cm; ”p5j7”: Amount paid for groceries or food used at home in last month; ”M1e1d1” - People who currently live in your HH - 1st age?; ”cm2hhinc”: Constructed - Household income (with imputed values); ”m2b5c1”: How much did child weigh on that day?-(Pounds)

Model	Overall Mean Square Error
Neural Network	0.629
Linear Regression	0.645
Lasso	0.652
Gradient Boosting	0.624
XGboost	0.633
Ensemble	0.639
Linear Regression (5 features)	0.639

Table 5: Sum of three mean squared error values for each model.

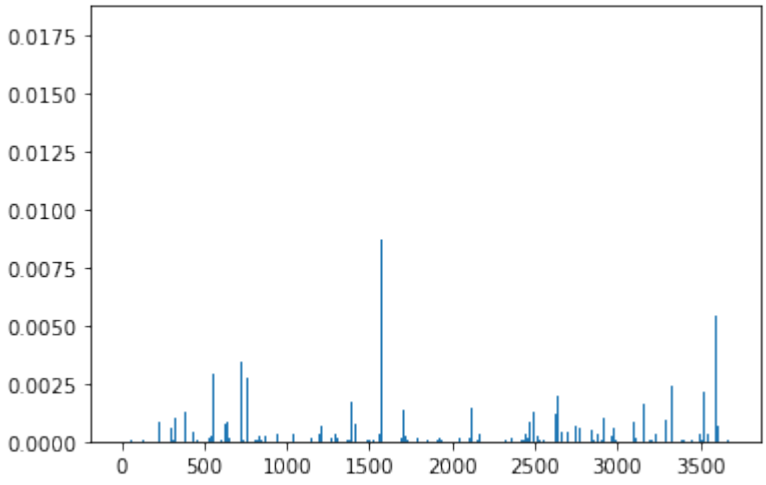

Figure 5: All features and their feature importance score as calculated by XGBoost, highlighting how most of the features are not very predictive of the chosen outcome (in this case, GPA)