Oxford Internet Institute, University of Oxford

Assignment Cover Sheet

Candidate Number	1080738
Assignment	Applied Machine Learning
Term	Hilary Term 2024
Title/Question	Combining Machine Learning and Social Theory to Discern GPA Predictors in the Fragile Families Challenge
Word Count	4,879

By placing a tick in this box ✓ I hereby certify as follows:

(a) This thesis or coursework is entirely my own work, except where acknowledgments of other sources are given. I also confirm that this coursework has not been submitted, wholly or substantially, to another examination at this or any other University or educational institution;

(b) I have read and understood the Education Committee's information and guidance on academic good practice and plagiarism at `https://www.ox.ac.uk/students/academic/guidance/skills?wssl=1`.

(c) I agree that my work may be checked for plagiarism using Turnitin software and have read the Notice to Candidates which can be seen at: `http://www.admin.ox.ac.uk/proctors/turnitin2w.shtml`, and that I agree to my work being screened and used as explained in that Notice;

(d) I have clearly indicated (with appropriate references) the presence of all material I have paraphrased, quoted or used from other sources, including any diagrams, charts, tables or graphs.

(e) I have acknowledged appropriately any assistance I have received in addition to that provided by my [tutor/supervisor/adviser].

(f) I have not sought assistance from a professional agency;

(g) I understand that any false claims for this work will be reported to the Proctors and may be penalized in accordance with the University regulations.

Contents

Combining Machine Learning and Social Theory to Discern GPA Predictors in the Fragile Families Challenge

1080738

1 Introduction

Grade Point Average (GPA) is a universally recognised metric of students' academic performance (Suryawan & Putra, 2016). Despite its measurement errors (Wittman, 2022), it has appeal for being simple, objective and recognisable, which facilitates cross-student comparisons crucial to institutional admissions among other uses (Papadogiannis et al., 2023, 513). As a significant predictor of future outcomes, such as graduation likelihood (Gayles, 2012; Gershenfeld, Ward Hood, & Zhan, 2016) and income (French, Homer, Popovici, & Robins, 2015), predicting GPA as early and efficiently as possible stands to optimise educational interventions and student support. This requires an understanding of how factors, such as cognitive ability, learning environment, and socioeconomic circumstance, influence academic performance.

Abundant literature links academic performance to cognitive ability. Previous research has linked high school GPA, standardised test scores and IQ (Frey & Detterman, 2004; Jensen, 1998). However, general metrics, such as IQ, leave much variance in academic performance unaccounted for, so scholars explore specific metrics, including working memory, processing speed, and spatial ability (Colom, Escorial, Shih, & Privado, 2007; Rohde & Thompson, 2007). For instance, spacial ability has been found to be predictive of academic performance in STEM subjects (Berkowitz & Stern, 2018). Performance in working memory tasks, such as Digit Span Forward, Corsi Block Backward and Digit Span Backward tasks, has also been found to predict academic success (Siquara, dos Santos Lima, & Abreu, 2018). Similar findings have been observed with processing speed (Colom et al., 2007).

Many scholars also attribute importance to learning environment (Baker, Bridger, Terry, & Winsor, 1997; Brock, Nishida, Chiong, Grimm, & Rimm-Kaufman, 2008; Flook, Repetti, & Ullman, 2005; Hamre & Pianta, 2005). This is rooted in the theory of self-system processes, wherein people have three needs in a learning environment - competence, autonomy, and relatedness - whose fulfillment enables better academic performance (Brock et al., 2008). The importance of external environment for learning outcomes came into focus during COVID-19 (Käser, Hallinen, & Schwartz, 2017). One important aspect of learning environment predictive of performance is class size (Borland, Howsen, & Trawick, 2005), exemplified by the seminal Tennessee STAR experiment (Nye, Hedges, &

Konstantopoulos, 2000). However, later literature finds that the effect of class size on performance is unclear, as many studies find either conflicting or insignificant relationships (Hanushek, 2002).

Socioeconomic circumstance, such as family wealth, family education and demographic grouping, is another important predictor of academic performance. For instance, Chiu et al. (2016) find that paternal education levels is important for predicting GPA, in an investigation linking family income, parental education, race and academic success. Parental income has also been found to be a highly statistically significant and robust predictor of GPA (Betts & Morell, 1999) at virtually all levels of schooling (Stinebrickner & Stinebrickner, 2000). Findings that income plays a stronger predictive role for students from low-income families than for their high-income counterparts (ibid) suggest that high- and low-income familial environments create different education experiences, so a model predicting GPA needs to be conscious that features can have different importances across different income groups.

Predicting GPA is notoriously difficult (Goldman & Slaughter, 1976). Moreover, most scholars focus on university-level GPA, only using middle/high school GPA as a predictive feature and not an important outcome in itself. Following from the above literature review, I ask the following questions:

> **RQ**$_1$: To what extent can high-school GPA be predicted using childhood-level data?
>
> **RQ**$_2$: To what extent are metrics of a child's cognitive ability, learning environment, and socioeconomic circumstance important predictors of high-school GPA?
>
> **RQ**$_3$: To what extent do these metrics' predictive importances vary between low- and high- income households?

In this report, I explore these questions using data from the *Fragile Families and Child Wellbeing Study* (FFCWS), used in the *Fragile Families Challenge* (FFC) (Reichman, Teitler, Garfinkel, & McLanahan, 2001). I build, optimise and evaluate a predictive model of GPA at 15 years old, positing the following hypotheses:

> **H**$_1$: the childhood data from the FFCWS can be used to reliably predict GPA at 15 years old.
>
> **H**$_2$: metrics of cognitive ability, learning environment, and socioeconomic circumstance will appear among the top 5 most important features in my predictive model.
>
> **H**$_3$: the feature importance space will be different for low-income and high-income households, such that income metrics will have higher feature importance than cognitive ability for low-income student while the opposite would be true for high-income students.

The rest of this research report proceeds as follows. Section 2 reports my methodological approach, including data description, pre-processing, feature selection, imputation and model construction/evaluation. Section 3 will summarise the results. Section 4 will offer a critical discussion of findings, limitations and scope for future improvements. Section 5 will conclude.

2 Methods

This section reports on the data science approach undertaken. The pipeline, shown in Figure 1, consist of four distinct types of tasks: 1) data preprocessing; 2) feature selection; 3) model building; 4) results interpretation.

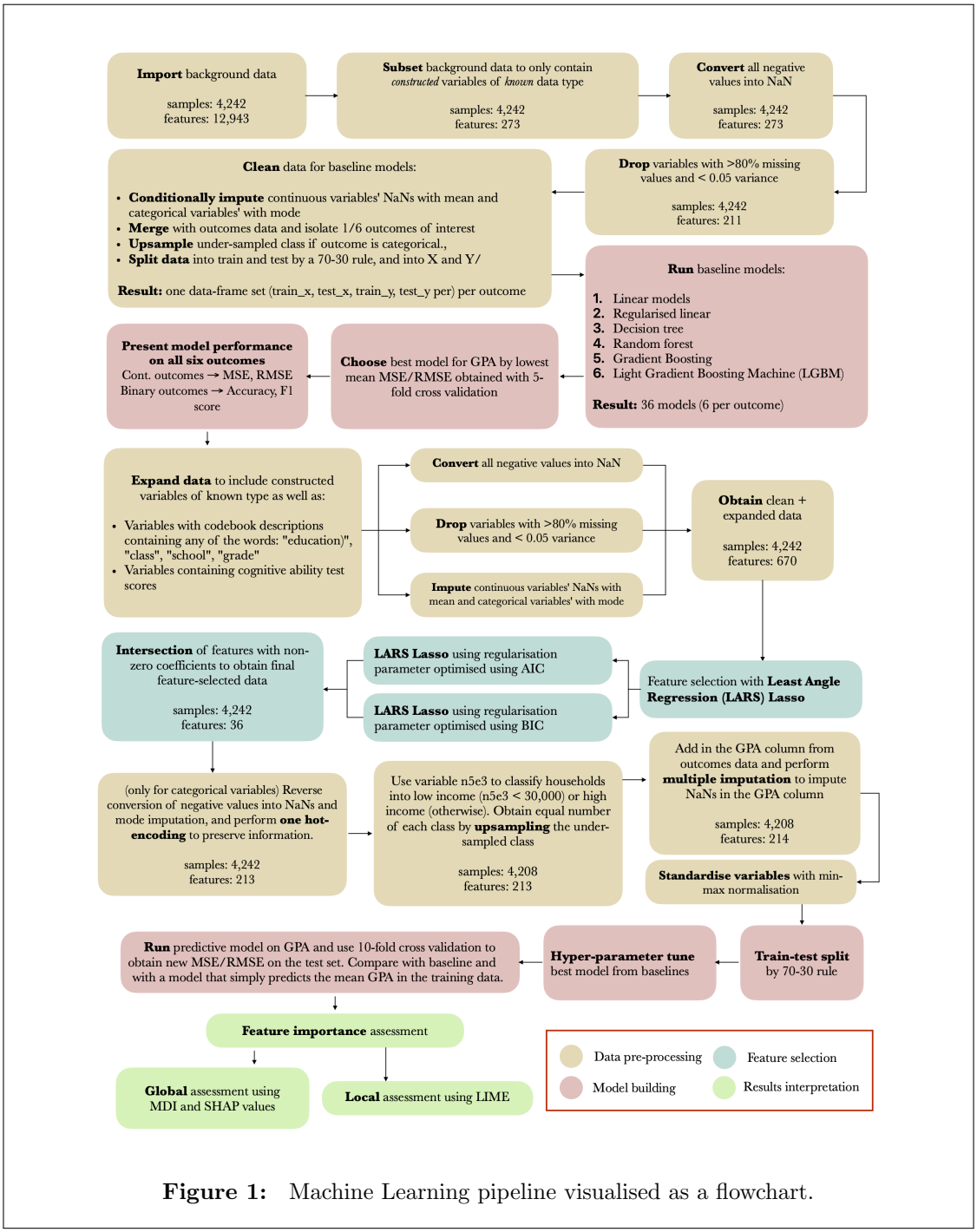

Figure 1: Machine Learning pipeline visualised as a flowchart.

2.1 Data

The FFCWS is a longitudinal study tracking 4,242 children and their families in large U.S. cities, collecting 12,942 *independent* variables (at ages 0-9 of the child) and six *dependent* variables (GPA, eviction, grit, material hardship, layoff, and job training, at age 15 of the child) (Carnegie & Wu, 2019; Reichman et al., 2001). The units of analysis vary, with some variables concerning households and caregivers, while others concern non-familial relations and the children themselves (M. J. Salganik, Lundberg, Kindel, & McLanahan, 2019). The FFC challenged scholars to compete in building the best predictive model of the six outcomes using this data. Beyond the direct purpose of optimising predictive power, the highly granular FFCWS data is invaluable for social data science. The below sub-sections outline how I prepared the data for modelling, as it risks biasing results in its raw form due to missingness, imbalances and multicolinearity.

2.1.1 Pre-processing

In line with Rigobon (2019), I drop any features with too many missing values (over 80%), as imputing them could skew results, and those with too low absolute variance (below 0.05), since they would not add substantial information to the model. To identify missingness percentage, I first convert all negative values to null values as well, since that is how valid skips (and other circumstantial missingness values) are coded in the data (see Table i in Appendix B). This leads to information loss, thus as discussed in section 2.1.3, negative values are later re-introduced into the data using one-hot encoding. This is done after feature-selection, as one-hot encoding before feature selection led to convergence issues with the feature selection algorithm discussed in section 2.1.2.

The Quantile-Quantile (QQ) plots in Figure i show that there are non-normally distributed features in the data. Scholars have identified that addressing distribution issues is appropriate for the FFC data, and have recommended using min-max normalisation via the following formula:

$$x' = (x_{\text{new max}} - x_{\text{new min}}) \times \frac{(x_i - x_{\min})}{(x_{\max} - x_{\min})} + x_{\text{new min}} \tag{1}$$

Typically, $x_{\text{new min}}$ and $x_{\text{new max}}$ are set to 0 and 1 so that the feature is re-scaled to a scale of 0-1. This step is implemented before model training.

2.1.2 Feature-selection

Due to the bias-variance trade-off, while more complex (higher feature number) models can capture more variance in the target outcome, they also risks reducing the model's capacity to generalise, fitting on noise rather than true relationships between features and outcome (Compton, 2019). Moreover, dealing with 12,942 requires significant computing power. As such, feature selection is required, for which I implement a two-tier approach, one theory-driven and one data-driven.

By the theory-driven approach, I first reduce the dataset to the 276 constructed variables with known data types, created by researchers to facilitate analysis (I. Lundberg, 2017). To these constructed variables, I manually re-add all non-constructed features which reflect cognitive ability (e.g. test scores), learning environment (e.g. class size) or socioeconomic circumstance (e.g. total household income), in line with theoretical discussions in section 1. Previous research has taken the same approach, finding that it actually leads to a relative improvement in model performance (Raes, 2019, 1). This results in a dataset of 670 features for 4242 households.

By the data-driven approach, I take this reduced dataset and perform Least Absolute Selection and Shrinkage Operator (LASSO) regression fit with Least Angle Regression (LARS), similar to previous research (Rigobon et al., 2019). Combining regularisation from LASSO with forward-selection from LARS, this approach is better than using LASSO alone for various reasons, including efficiency with high-dimensional data and stability (as LASSO could be sensitive to the sequence of feature introduction into the model). I considered alternative approaches, such as Bayesian generalised linear models (BGLM), and Bayesian additive regression trees (BART) (Chipman, George, & McCulloch, 2006), to detect non-linear relationships. However, I chose a LASSO-based approach as linear models are more interpretable.

To select the LASSO regularisation parameter α, I use the `LassoLarsIC` algorithm and find two α values (see Figure iv) - $\alpha = 0.0297$ minimising the *Akaike Information Criterion* (AIC) and $\alpha = 0.0614$ minimising the *Bayesian Information Criterion* (BIC). Ideally a single α would minimise both, but this is unexpected as the AIC and BIC have different formulations. I then fit two iterations of LASSO - one with each α - and identify the features which have non-zero coefficients in both LASSO models. According to Kuha (2004), this is superior to just using a single information criterion. The resultant dataset reduces the feature number to 36, with the features being a mix of metrics of cognitive ability, learning environment and socioeconomic circumstance (see Table ii for a full list of chosen features).

2.1.3 Imputation

Figure ii (see Appendix B) shows that especially after negative values are re-coded as null values, most features have non-zero missingness. Moreover, 45.1% of households (excluding those in the holdout set) have no GPA data. As keeping only complete cases yields a very small sample, and the models would not run with null values, imputation is necessary.

To minimise information loss, independent variables were conditionally imputed, such that continuous features are mean-imputed. Some would argue that to mitigate against uncertainty, multiple imputation is better (Pedersen et al., 2017). However, literature argues that multiple and single imputation in the FFCWS data do not lead to significantly different results, thus allowing the use of single imputation as a less computationally intensive approach to high-dimensional data (Ahearn & Brand, 2019). Salganik (2017) actually shows that mean imputation performs better than model-based imputation on the FFCWS data.

For categorical features, a two-tier approach is undertaken. Used by scholars working with the FFCWS data, one-hot encoding is ideal as it minimises information loss (Rigobon et al., 2019). However, it increases the feature number substantially, which increases compute time and creates convergence issues with the `LassoLarsIC` algorithm as the number of features exceeds samples. To avoid this, before feature selection, categorical features are imputed using single-value (mode) imputation (Ahearn & Brand, 2019). After feature selection, categorical features' (from the chosen 36) missing values are restored, and one-hot encoding is implemented to maximise information available to the final model.

I also adopt a two-tier approach to imputing GPA. For training the final model, rows with no values of GPA undergo multiple imputation, in favour of increasing sample size. However, doing this at the feature selection stage risks inflating coefficients, so the feature selection is ran on a dataset where rows with no values of GPA are removed. The final model is thus trained on data where the continuous features are mean-imputed, categorical features are one-hot encoded, and GPA is multiple-imputed.

2.1.4 Addressing imbalance

The investigation of $\mathbf{RQ}_3$ includes the splitting of all households into one of two categories - low-income and high-income. Using the `n5e3` feature, all households with total income below 30,000 USD are classified as low-income, and all others as high-income, in accordance with the 2015 US federal poverty line (Bishaw & Glassman, 2016). The resultant class imbalance can be dealt with either through under- or over-sampling. In line with scholars that find over-sampling to be more successful for predictive tasks (Bria, Marrocco, & Tortorella, 2020), I bootstrap the under-sampled class (low income) to obtain a balanced dataset.

2.2 Model

A series of baseline models were deployed on each of the six outcomes to gauge performance differentials and choose the best model for predicting GPA. The chosen model undergoes hyper-parameter tuning. The baseline models are evaluated using 5-fold cross validation to obtain average performance statistics on the test set, and the chosen model does the same with 10-fold cross validation. The model is interpreted using three different explainable AI (XAI) techniques.

2.2.1 Baseline Models

The goal of the baseline models is to gauge how well various models perform on predicting each of the six outcome variables in the FFCWS data. An additional step of correcting class imbalances in the three binary outcomes by bootstrapping and over-sampling the under-sampled class was undertaken to ensure a neater baseline not obscured by skews in the data. The baseline models implemented are the following:

Table 1: Baseline models implemented for continuous and categorical outcomes

Model Type	**Continuous Outcomes (GPA, grit, material hardship)**	**Categorical Outcomes (eviction, layoff, job training)**
Mean	Predicting the mean value	Predicting the mode value
Linear	Ordinary Least Squares (OLS) Regression	Logistic Regression
Regularised	Elastic Net Regression	Ridge Logistic Regression
Tree-based	Decision Tree (DT) Regressor Random Forest (RF) Regressor Gradient Boosting (GB) Regressor Light Gradient Boosting Machine (LGBM) Regressor	Decision Tree (DT) Classifier Random Forest (RF) Classifier Gradient Boosting (GB) Classifier Light Gradient Boosting Machine (LGBM) Classifier

These models were chosen for various reasons. The mean/mode model is a widely used baseline (Rigobon et al., 2019). Linear models are another common baseline, whose easily interpretable coefficients make them useful for models aimed at explaining relationships between features and outcomes. Regularised linear models are better suited at working with highly dimensional data, like the FFCWS data, by penalising model complexity (Sirimongkolkasem & Drikvandi, 2019). Tree-based models are also included, as they are able to model non-linear relationships which linear and regularised linear models cannot do (Chen et al., 2019). While neural networks was considered as a further baseline, I ultimately decided against it, as deep neural network models are notoriously non-interpretable (Liu, Wang, & Matwin, 2018), which would make feature importance exploration difficult.

2.2.2 Chosen model

Each model will be evaluated on each outcome using the mean MSE and RMSE from 5-fold cross validation. MSE is a standard metric of for forecasting performance in machine learning (Han, Pei, & Tong, 2022), and RMSE has a more meaningful interpretation as it is measured in the units of the outcome (while MSE is measured in square units of the outcome). The model with the lowest MSE and RMSE on GPA becomes the chosen model for the final training, testing and interpretability stages.

If the chosen model is either regularised linear or tree-based, the model will undergo hyper-parameter tuning using random search. Bergstra and Bengio (2012, 281) find that "random search over the same domain [as pure grid search] is able to find models that are as good or better within a small fraction of the computation time [of pure grid search]". This is because random search is able to search a bigger configuration space (ibid), which becomes clear in Figure 2:

Alternative techniques, such as Bayesian Optimisation (BO) were considered. However, while performing well on low-dimensional data, scaling it up to higher dimensional data is a challenge given "exponentially increasing statistical and computational com-

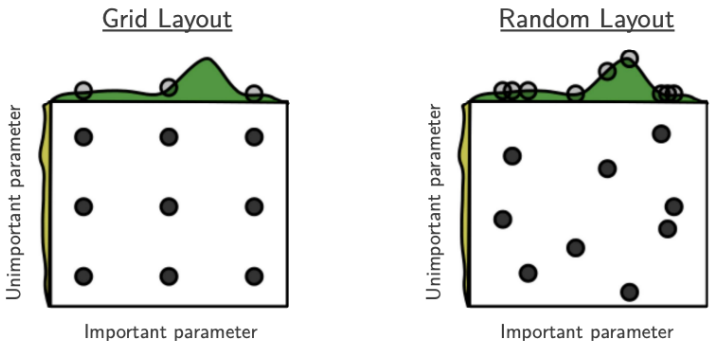

Figure 2: Pure grid and random search diagrams representing nine trials for tuning the hyper-parameters of a function $f(x, y) = g(x) + h(y) \approx g(x)$ with low effective dimensionality. Using grid search, nine trials can only test $g(x)$ in three places on the green curve, while all nine trials explore distinct regions of the green curve if random search is used.

Source: Bergstra and Bengio (2012, 284)

plexity with increasing dimensions" (Malu, Dasarathy, & Spanias, 2021). Remedies for this are outside the scope of this project. The hyper-parameters I would tune for each model (as any of the baseline models could become the chosen model) can be found in Table iii (see Appendix D).

To evaluate the predictive performance of the chosen model, average MSE and RMSE from a 10-fold cross validation will be reported. For interpreting features' importance, the method depends on the chosen model. If the chosen model is linear or regularised linear, then the coefficients on each feature can be used to discern features' importance (M. Salganik, 2017). If the chosen model is tree-based, three explainable AI (XAI) techniques are appropriate:

- **MDI (Mean Decrease in Impurity)**: this method sums the weighted impurity decreases over all nodes that split on a given feature, averaging across all trees in the forest - a high MDI value indicates that the feature is useful in many important operations of the forest's prediction process (Bénard, Da Veiga, & Scornet, 2022, 882). MDI is found to be consistent when features are independent, which is partially confirmed by Figure v (see Appendix E). However, empirical work has highlighted issues with MDI (Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008), so I cross-reference results with the next two XAI techniques.
- **SHAP (SHapley Additive exPlanations)**: this value is defined as the "the Shapley value [average contribution of a feature to a prediction, expressed by comparing predicted values of the model with and without the given feature] for the conditional expected value function" of the given model (Lee, Oh, Kim, & Kim, 2023, 579-580). I employ this method, because it satisfies all three characteristics of *Additive Feature Attribution* techniques (S. M. Lundberg & Lee, 2017) - local accuracy, missingness, and consistency (Lee et al., 2023, 580). Nevertheless, conventional SHAP values are not sensitive to variance in data distribution.

- **LIME (Local Interpretable Model-agnostic Explanations)**: LIME derives feature importances via comparing results of an interpretable linear model it generates with the predictions of the original tree-based model (Främling, Westberg, Jullum, Madhikermi, & Malhi, 2021). It does so by perturbing the input data to observe how predictions change, and variables that cause the most significant changes in predictions across these perturbations are considered more important. I use it given that it is one of the most predominant XAI methods, conscious that its image explainer is more prevalent than its tabular explainer (Dieber & Kirrane, 2020).

3 Results

3.1 Baseline models

Tables 2 and 3 show the results of baseline models for each of the six outcome variables, with MSE and RMSE scores obtained with 5-fold cross validation. The mean model is not included in the baseline results, as it is intended as a baseline comparison with the final chosen model, not as a candidate for being the predictive model itself

variable	MSE	RMSE	model
gpa	0.516309	0.717608	OLS
gpa	0.406696	0.637363	ElasticNet
gpa	0.831760	0.911480	DecisionTree
gpa	0.405152	0.636131	RandomForest
gpa	0.424787	0.651276	Gradient Boosting
gpa	0.430200	0.655291	LGBM
grit	0.283923	0.532277	OLS
grit	0.234083	0.483774	ElasticNet
grit	0.480158	0.692509	DecisionTree
grit	0.235771	0.485484	RandomForest
grit	0.244756	0.494598	Gradient Boosting
grit	0.256216	0.505877	LGBM
materialHardship	0.027377	0.165129	OLS
materialHardship	0.022843	0.150957	ElasticNet
materialHardship	0.045521	0.213245	DecisionTree
materialHardship	0.023513	0.153181	RandomForest
materialHardship	0.024251	0.155552	Gradient Boosting
materialHardship	0.024815	0.157361	LGBM

Table 2: MSE and RMSE across baseline models for each of the three continuous outcomes. Lowest MSE/RMSE model for each outcome highlighted in red.

variable	Accuracy	F1	model
eviction	0.583835	0.571099	Logistic Regression
eviction	0.834195	0.830234	Ridge Logistic Regression
eviction	0.740905	0.682393	Decision Tree
eviction	1.000000	0.999270	Random Forest
eviction	0.705169	0.657223	Gradient Boosting
eviction	0.688713	0.600789	LGBM
layoff	0.461386	0.440247	Logistic Regression
layoff	0.620297	0.613816	Ridge Logistic Regression
layoff	0.588119	0.478366	Decision Tree
layoff	0.824257	0.809864	Random Forest
layoff	0.467327	0.362610	Gradient Boosting
layoff	0.523267	0.396978	LGBM
jobTraining	0.447731	0.428981	Logistic Regression
jobTraining	0.590799	0.584513	Ridge Logistic Regression
jobTraining	0.505371	0.375882	Decision Tree
jobTraining	0.747410	0.737015	Random Forest
jobTraining	0.425315	0.324219	Gradient Boosting
jobTraining	0.481218	0.345751	LGBM

Table 3: Accuracy and F1 score for baseline models on each of the three categorical outcomes. Highest accuracy/F1 model for each outcome highlighted in red.

The RF regressor achieved the best baseline performance in predicting GPA, with MSE of 0.636 (i.e. on average, predicted GPA differs from actual GPA by 0.636). This is very similar to Elastic Net regression (RMSE = 0.637). While the differences in performance are minuscule, using RF is justified as it mostly outperforms other models for the other outcomes. Thus, the chosen model is the RF regressor.

3.2 Random Forest model

Table 4 shows the hyper-parameter ranges investigated using random search for each parameter of the RF regressor. The table also shows the parameters forming the best combination in terms of minimising cross-validation score.

Hyper-parameter	Range	Best Hyper-parameter
n_estimators	10-2000	878
max_features	'auto', 'sqrt'	'sqrt
max_depth	None, 1-10	9
min_samples_split	2-10	5
min_samples_leaf	1-10	7

Table 4: RF regressor hyper-parameter ranges and best parameters found from random search.

As a result of feature selection, imputation, balancing and tuning with the above parameters, the RF regressor saw a sizeable improvement in performance. A 10-fold cross validation using the new RF model produced a RMSE of 0.376 (MSE of 0.141).

This is a 65.2% decrease in MSE. The tuned RF model's MSE is 68.2% lower than that of the baseline model predicting mean GPA (0.443).

I conducted various robustness checks. Firstly, I ran a version of the tuned RF model using only non-imputed data, and a version using non-balanced data, to ensure that the results are not skewed by data pre-processing. The tuned RF model using only non-imputed values generates an average MSE of 0.155, and the model using non-oversampled data generates a MSE of 0.153. Secondly, the performance of the tuned RF model is assessed on the holdout set. While this would not have been an available step to participants in the FFC, if the holdout MSE is a lot higher than the one obtained on the test set, this would serve as evidence of over-fitting. The MSE from the holdout set is 0.206, which represents a 53.4% decrease from the baseline mean model. Finally, this value is re-computed using only the rows of the holdout set for which GPA is not-imputed, to check whether imputation is skewing the results significantly. By this approach, the average MSE is 0.365, representing a 17.5% decrease from the baseline mean model.

3.3 Feature Importance

Figure vi (see Appendix E) shows the top 15 most important features by MDI. The top 5 features are `n5e3` (total household income), `t5e1` (class size), `m1i6` (What was BF doing most of last week (working, going to school,else)?), `hv5_wj9ss` (Woodcock-Johnson test score) and `k5g2h` (It's hard for me to finish my schoolwork). Income level is a derivative of `n5e3`, hence why I do not list it. Dis-aggregating this into the 15 most important features for kids in low-income versus high income households, Figure vii shows that for low-income students, class size and household income appear to be much more important, while for high-income households, maternal metrics of income (e.g. `m4l1`, `m5j1` and `m4k13`) alongside Woodcock-Johnson test scores were the most important.

Figure viii shows the 15 top features for predicting GPA using SHAP. Maternal income metrics feature alongside metrics such as `t5e6` (Highest level of education completed by aide) and `k5g2h` (It's hard for me to finish my schoolwork). Figure ix shows that for low-income households, class size is once again the most important feature, alongside parental metrics of income, as well as whether the value for GPA was imputed or not. For high-income households, metrics of cognitive ability (i.e. `hv5_wj10ss` and `hv5_ppvtss`) have higher importance alongside income metrics.

Figure x shows the top 15 features derived by summing up the LIME values across all local predictions to approximate global feature importance. For high-income households, variables relating to learning environment (e.g. difficulty completing schoolwork, aide's highest education level, class size) are closer to the top of the most important features list, while for low-income households, variables relating to household income (mostly maternal) dominate the most important features.

4 Discussion

4.1 Findings

The findings in section 3 show partial support for $\mathbf{H}_1$, which hypothesised that 'the childhood data from the FFCWS can be used to reliably predict GPA at 15 years old'. On one hand, the results suggest that on average, my model's predictions were 0.376 points away from the actual GPA for the given challenge ID, which is a substantial difference given that GPA is bounded between 0 and 4 (see Figure iii in Appendix B, calling into question the reliability of the model. On the other hand, it is difficult to determine what qualifies as a reliable prediction - the aim is to get the MSE/RMSE as close as possible to zero. In that respect, given that the model ranked first for predicting GPA in the FFC had an MSE of 0.377 on the holdout set (Fragile Families Challenge Team, 2016), my model's MSE of 0.206 (imputed holdout set) and 0.365 (unimputed holdout set) on the holdout set suggests that it is *relatively* reliable in predicting GPA, performing at least as well as the top-ranked model in 2016.

My results do support $\mathbf{H}_2$, which posited that 'metrics of cognitive ability, learning environment, and socioeconomic circumstance will appear among the top 15 most important features in my predictive model'. MDI-derived feature importance scores show that the top 5 features are total household income (`n5e3`), class size (`t5e1`), a proxy for mother's partner's occupation (`m1i6`), the child's Woodcock-Johnson test score (`hv5_wj9ss`) and difficulty completing homework (`k5g2h`). This is congruent with literature linking class size to academic performance (Borland et al., 2005; Nye et al., 2000) while contributing to clearing up uncertainty on the topic (Hanushek, 2002). This is also congruent with literature showing that higher household income (Betts & Morell, 1999; Chiu et al., 2016; Stinebrickner & Stinebrickner, 2000) and cognitive ability (Colom et al., 2007; Siquara et al., 2018) are both enablers of academic success. Deriving feature importance with SHAP also reveals a similar mix, with additional learning environment metrics such as the child's aide's education (`t5e6`), and more parental income metrics.

My results show partial support for $\mathbf{H}_3$, which theorised that 'the feature importance space will be different for low-income and high-income households, such that income metrics will have higher feature importance than cognitive ability for low-income student while the opposite would be true for high-income students'. By all three XAI techniques, the household income variable (`n5e3`) was consistently ranked as more important for children from low-income households than for those from high-income households. This is aligned with existing findings (Stinebrickner & Stinebrickner, 2000). However, parent-specific metrics of income do appear in the top 5 most important features regardless of household income category, which invites further research to discern whether total household income and parent-specific income have different effects on academic performance.

4.2 Limitations

Firstly, while oversampling was implemented to obtain a sample balanced between low- and high-income households, the FFCWS data is imbalanced on other dimensions, given that participants were selected using a complex sample design wherein members were not selected independently, and without equal probabilities of being chosen (Fragile Families Challenge Team, 2001). For instance, urban households are over-sampled for various reasons (Reichman et al., 2001), as are non-marital births (Fragile Families Challenge Team, 2001). There is reason to believe that other skewing effects are present, as the binary outcomes are also class-imbalanced (Compton, 2019). The data is mired with additional issues for which my approach does not account, such as baseline non-response and attrition. This leaves room for future research to implement my methodology on a *weighted* version of the FFCWS data, using the corrective weights provided (ibid.)

Secondly, various steps in the data pre-processing stage risk introducing bias and skewing results. My approach does not assess model results for sensitivity to perturbations in the feature selection and imputation algorithms. The results reported in section 3 are likely dependent on the feature selection algorithm chosen. Existing research has found that the choice of feature selection algorithm has a significant impact on machine learning model performance metrics (Haury, Gestraud, & Vert, 2011). Similarly, the choice of imputation method has been found to significantly impact model results (Stavseth, Clausen, & Røislien, 2019), which threatens my results' validity given that approximately 45% of households had missing values for GPA. Moreover, a substantial part of my analysis relies on my construction of the binary `income_level` variable. Given the difficulty of discerning household size per challenge ID, and that the US Federal poverty line varies with household size, it is likely that a substantial portion of the households is misclassified in my categorisation, which further threatens the validity of my findings. Furthermore, Figure iii shows that the model does not make GPA predictions below 2.5, which is unlike the real-life GPA distribution, which could be skewing the MSE/RMSE estimates. This all invites future research to examine the sensitivity of my results to different feature selection methods (e.g. using Elastic Net) possibly leveraging the full dataset instead of just the constructed variables, as well as more sophisticated imputation methods such as those described by Goode, Datta and Ramakrishnan (2019). Further research should also improve the `income_level` feature by accounting for household size.

Moreover, there could be loss of information, as well as loss of potentially higher performance, stemming from data-processing, hyper-parameter tuning and interpretability steps undertaken. Due to the high computational cost of working with the full feature set of the FFCWS data, analysis was carried out only on the constructed variables (alongside a few non-constructed variables). This likely removes some potentially important features that would have been uncovered during feature selection. At the hyper-parameter tuning stage, it is possible that a more granular search (e.g. using random search to identify regions that likely contain the optimal hyper-parameter combination, and then using grid-search or another random search layer to explore that region further) could reveal a better hyper-parameter combination that could recover any loss in performance from the hyper-parameters I chose using random search alone. Furthermore, the feature importance analysis in section 3 mostly focuses on global importance, summing up LIME values for example. This disables the analytical framework from identifying specific instances and examining the importance of features on individual children to reveal

more granular insights. Future research should leverage machines with higher computing power to incorporate the full feature space into feature selection, as well as more elaborate hyper-parameter tuning methods, such as the one outlined above or Bayesian optimisation with appropriate adjustments to deal with high-dimensional data (Malu et al., 2021). Future research should also examine SHAP and LIME values at the local level.

Finally, while not a methodological issue, the FFCWS data presents an ethical risk through the threat of re-identification, inherent to all big data sources which deal with sensitive information. There are numerous examples of re-identification in academic research using seemingly-anonymous data. For instance, an MIT graduate student was able to re-identify anonymous medical records by combining date of birth, sex, and ZIP code (all included in the anonymous data) with auxiliary data (Sweeney, 2002). Similarly, DNA records have been matched with hospital discharge records, which contained basic demographic data, to link them to voting records (Malin & Sweeney, 2004). Lundberg, Narayanan, Levy and Salganik (2019) describe multiple avenues for a potential re-identification attack on the FFCWS data. Given the extraordinary volume of available data, a lot of which is highly sensitive (i.e. concerning drug use, child delinquency, etc.), such an attack is definitely possible, and could be detrimental to the life of a participant in the FFCWS study. This all calls into question whether using such data is ethnical, especially in the format of the FFC by which it was broadcast to the wider scientific community.

5 Conclusion

This report detailed the use of data from the *Fragile Families and Child Wellbeing Study* to build a model predicting a child's GPA at 15 years old using a high number of variables collected throughout their childhood. Aside from seeing how well my model performs compared to various benchmarks, as well as the actual winning model for GPA from the Fragile Families Challenge, I explored whether the model's predictions were informed by a mix of features of the child's cognitive ability, learning environment and socioeconomic circumstance, as suggested by relevant literature. I further investigated whether GPA predictions of children from low-income households were more influenced by socioeconomic factors, such as household income, than those for children from high-income households. Across various permutations undertaken to assess robustness, my model performed at least as well as the winning model, both on the self-constructed test set and the FFC-provided holdout set, showing substantial improvements from baseline models. I observed a mix of features congruent with literature in the top most important features for the model's predictions, and found partial evidence that GPA predictions for low-income households rely more on total household income than their high-income counterparts. Mindful of various potential improvements to my methodology, this work contributes to ongoing effort to better predict academic performance in order to optimise the support students receive to thrive during and after high-school.

References

Ahearn, C. E., & Brand, J. E. (2019). Predicting layoff among fragile families. *Socius*, *5*, 2378023118809757.

Baker, J. A., Bridger, R., Terry, T., & Winsor, A. (1997). Schools as caring communities: A relational approach to school reform. *School Psychology Review*, *26*(4), 586–602.

Bénard, C., Da Veiga, S., & Scornet, E. (2022). Mean decrease accuracy for random forests: inconsistency, and a practical solution via the sobol-mda. *Biometrika*, *109*(4), 881–900.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2).

Berkowitz, M., & Stern, E. (2018). Which cognitive abilities make the difference? predicting academic achievements in advanced stem studies. *Journal of Intelligence*, *6*(4), 48.

Betts, J. R., & Morell, D. (1999). The determinants of undergraduate grade point average: The relative importance of family background, high school resources, and peer group effects. *Journal of human Resources*, 268–293.

Bishaw, A., & Glassman, B. (2016). *Poverty: 2014 and 2015*. US Department of Commerce, Economics and Statistics Administration, US ....

Borland, M. V., Howsen, R. M., & Trawick, M. W. (2005). An investigation of the effect of class size on student academic achievement. *Education Economics*, *13*(1), 73–83.

Bria, A., Marrocco, C., & Tortorella, F. (2020). Addressing class imbalance in deep learning for small lesion detection on medical images. *Computers in biology and medicine*, *120*, 103735.

Brock, L. L., Nishida, T. K., Chiong, C., Grimm, K. J., & Rimm-Kaufman, S. E. (2008). Children's perceptions of the classroom environment and social and academic performance: A longitudinal analysis of the contribution of the responsive classroom approach. *Journal of School Psychology*, *46*(2), 129–149.

Carnegie, N. B., & Wu, J. (2019). Variable selection and parameter tuning for bart modeling in the fragile families challenge. *Socius*, *5*, 2378023119825886.

Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., ... others (2019). A comparison of linear regression, regularization, and machine learning algorithms to develop europe-wide spatial models of fine particles and nitrogen dioxide. *Environment international*, *130*, 104934.

Chipman, H., George, E., & McCulloch, R. (2006). Bayesian ensemble learning. *Advances in neural information processing systems*, *19*.

Chiu, J., Economos, J., Markson, C., Raicovi, V., Howell, C., Morote, E.-S., & Inserra, A. (2016). Which matters most? perceptions of family income or parental education on academic achievement. *New York Journal of Student Affairs*, *16*(2), 3.

Colom, R., Escorial, S., Shih, P. C., & Privado, J. (2007). Fluid intelligence, memory span, and temperament difficulties predict academic performance of young adolescents. *Personality and Individual differences*, *42*(8), 1503–1514.

Compton, R. (2019). A data-driven approach to the fragile families challenge: Prediction through principal-components analysis and random forests. *Socius*, *5*, 2378023118818720.

Dieber, J., & Kirrane, S. (2020). Why model why? assessing the strengths and limitations of lime. *arXiv preprint arXiv:2012.00093*.

Flook, L., Repetti, R. L., & Ullman, J. B. (2005). Classroom social experiences as predictors of academic performance. *Developmental psychology*, *41*(2), 319.

Fragile Families Challenge Team. (2001). *Fragile families child wellbeing study: A brief guide to using the weights for waves 1-6.* `https://ffcws.princeton.edu/sites/g/files/toruqf4356/files/documents/using_the_fragile_families_weights_waves_1_6_01052021.pdf`. Princeton University.

Fragile Families Challenge Team. (2016). *Fragile families challenge demonstration leaderboard.* `https://codalab.fragilefamieschallenge.org/competitions/36#results`. (Accessed: 22-04-2024)

Främling, K., Westberg, M., Jullum, M., Madhikermi, M., & Malhi, A. (2021). Comparison of contextual importance and utility with lime and shapley values. In *International workshop on explainable, transparent autonomous agents and multi-agent systems* (pp. 39–54).

French, M. T., Homer, J. F., Popovici, I., & Robins, P. K. (2015). What you do in high school matters: High school gpa, educational attainment, and labor market earnings as a young adult. *Eastern Economic Journal*, *41*, 370–386.

Frey, M. C., & Detterman, D. K. (2004). Scholastic assessment or g? the relationship between the scholastic assessment test and general cognitive ability. *Psychological science*, *15*(6), 373–378.

Gayles, J. (2012). Race, late bloomers and first-year gpa: Predicting beyond the freshman year. *Educational Research Quarterly*, *36*(1), 13–29.

Gershenfeld, S., Ward Hood, D., & Zhan, M. (2016). The role of first-semester gpa in predicting graduation rates of underrepresented students. *Journal of College Student Retention: Research, Theory & Practice*, *17*(4), 469–488.

Goldman, R. D., & Slaughter, R. E. (1976). Why college grade point average is difficult to predict. *Journal of Educational Psychology*, *68*(1), 9.

Goode, B. J., Datta, D., & Ramakrishnan, N. (2019). Imputing data for the fragile families challenge: Identifying similar survey questions with semiautomated methods. *Socius*, *5*, 2378023118822647.

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child development*, *76*(5), 949–967.

Han, J., Pei, J., & Tong, H. (2022). *Data mining: concepts and techniques.* Morgan kaufmann.

Hanushek, E. A. (2002). Evidence, politics, and the class size debate. *The class size debate*, 37–65.

Haury, A.-C., Gestraud, P., & Vert, J.-P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PloS one*, *6*(12), e28210.

Jensen, A. R. (1998). *The g factor: The science of mental ability.*

Käser, T., Hallinen, N. R., & Schwartz, D. L. (2017). Modeling exploration strategies to predict student performance within a learning environment and beyond. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 31–40).

Kuha, J. (2004). Aic and bic: Comparisons of assumptions and performance. *Sociological methods & research*, *33*(2), 188–229.

Lee, Y.-G., Oh, J.-Y., Kim, D., & Kim, G. (2023). Shap value-based feature importance analysis for short-term load forecasting. *Journal of Electrical Engineering &*

Technology, *18*(1), 579–588.

Liu, X., Wang, X., & Matwin, S. (2018). Interpretable deep convolutional neural networks via meta-learning. In *2018 international joint conference on neural networks (ijcnn)* (pp. 1–9).

Lundberg, I. (2017). Constructed variables - data dictionary. *Fragile Families Challenge Blog*.

Lundberg, I., Narayanan, A., Levy, K., & Salganik, M. J. (2019). Privacy, ethics, and data access: A case study of the fragile families challenge. *Socius*, *5*, 2378023118813023.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Malin, B., & Sweeney, L. (2004). How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of biomedical informatics*, *37*(3), 179–192.

Malu, M., Dasarathy, G., & Spanias, A. (2021). Bayesian optimization in high-dimensional spaces: A brief survey. In *2021 12th international conference on information, intelligence, systems & applications (iisa)* (pp. 1–8).

Nye, B., Hedges, L. V., & Konstantopoulos, S. (2000). The effects of small classes on academic achievement: The results of the tennessee class size experiment. *American Educational Research Journal*, *37*(1), 123–151.

Papadogiannis, I., Poulopoulos, V., Platis, N., Vassilakis, C., Lepouras, G., & Wallace, M. (2023). First grade gpa as a predictor of later academic performance in high school. *Knowledge*, *3*(3), 513–524.

Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., & Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical epidemiology*, 157–166.

Raes, L. (2019). Predicting gpa at age 15 in the fragile families and child wellbeing study. *Socius*, *5*, 2378023118824803.

Reichman, N. E., Teitler, J. O., Garfinkel, I., & McLanahan, S. S. (2001). Fragile families: Sample and design. *Children and Youth Services Review*, *23*(4-5), 303–326.

Rigobon, D. E., Jahani, E., Suhara, Y., AlGhoneim, K., Alghunaim, A., Pentland, A. & Almaatouq, A. (2019). Winning models for grade point average, grit, and layoff in the fragile families challenge. *Socius*, *5*, 2378023118820418.

Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. *Intelligence*, *35*(1), 83–92.

Salganik, M. (2017). A data pipeline for the fragile families challenge. *Fragile Families Challenge Blog*.

Salganik, M. J., Lundberg, I., Kindel, A. T., & McLanahan, S. (2019). Introduction to the special collection on the fragile families challenge. *Socius*, *5*.

Siquara, G. M., dos Santos Lima, C., & Abreu, N. (2018). Working memory and intelligence quotient: Which best predicts on school achievement? *Psico*, *49*(4), 365–374.

Sirimongkolkasem, T., & Drikvandi, R. (2019). On regularisation methods for analysis of high dimensional data. *Annals of Data Science*, *6*(4), 737–763.

Stavseth, M. R., Clausen, T., & Røislien, J. (2019). How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data. *SAGE open medicine*, *7*, 2050312118822912.

Stinebrickner, T. R., & Stinebrickner, R. (2000). *The relationship between family income and schooling attainment: evidence from a liberal arts college with a full tuition subsidy program* (Tech. Rep.). Research Report.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics*, *9*, 1–11.

Suryawan, A. D., & Putra, E. (2016). Analysis of determining factors for successful student's gpa achievement. In *2016 11th international conference on knowledge, information and creativity support systems (kicss)* (pp. 1–7).

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, *10*(05), 557–570.

Wittman, D. (2022). Average rank and adjusted rank are better measures of college student success than gpa. *Educational Measurement: Issues and Practice*, *41*(4), 23–34.

A Code

All code written for this assessment can be found in the file '1080738_AML_code_final.ipynb' inside the following GitHub repository: `https://github.com/1080738/applied-machine-learning/tree/main`. For the data, I used the FFChallenge_v2 folder from the Princeton Office of Population Research archive website.

B Data

Table i: Missing values codes in the FFCWS data

Value	**Explanation**
-9	Not in wave - Did not participate in survey/data collection component
-8	Out of range - Response not possible; rarely used
-7	Not applicable (also -10/-14) - Rarely used for survey questions
-6	Valid skip - Intentionally not asked question; question does not apply to respondent
-5	Not asked "Invalid skip" - Respondent not asked question in the version of the survey they received
-3	Missing - Data is missing due to some other reason; rarely used
-2	Don't know - Respondent asked question; Responded "Don't Know".
-1	Refuse - Respondent asked question; Refused to answer question

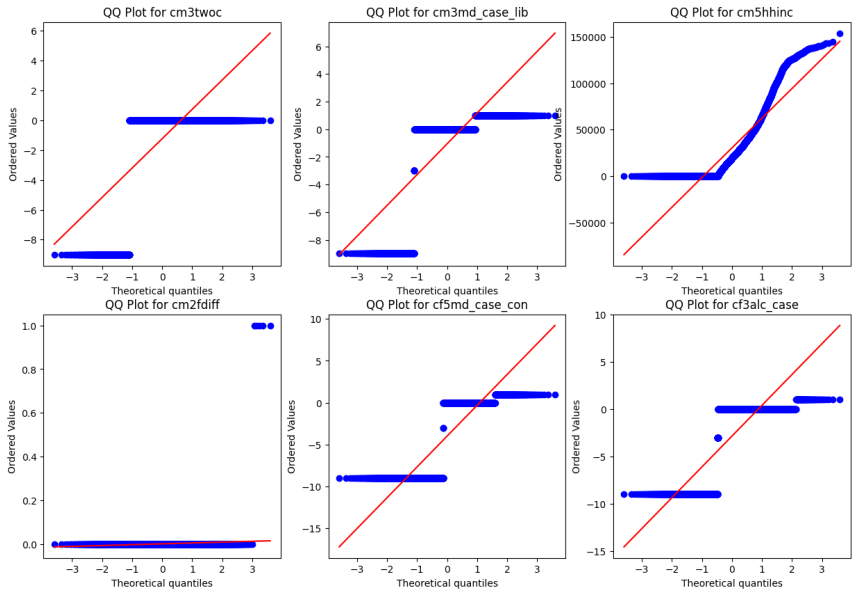

Figure i: Quantile-Quantile (QQ) plots for a random subset of six features from the data, clearly demonstrating deviations from the red line, which occur when the variable is distributed non-normally.

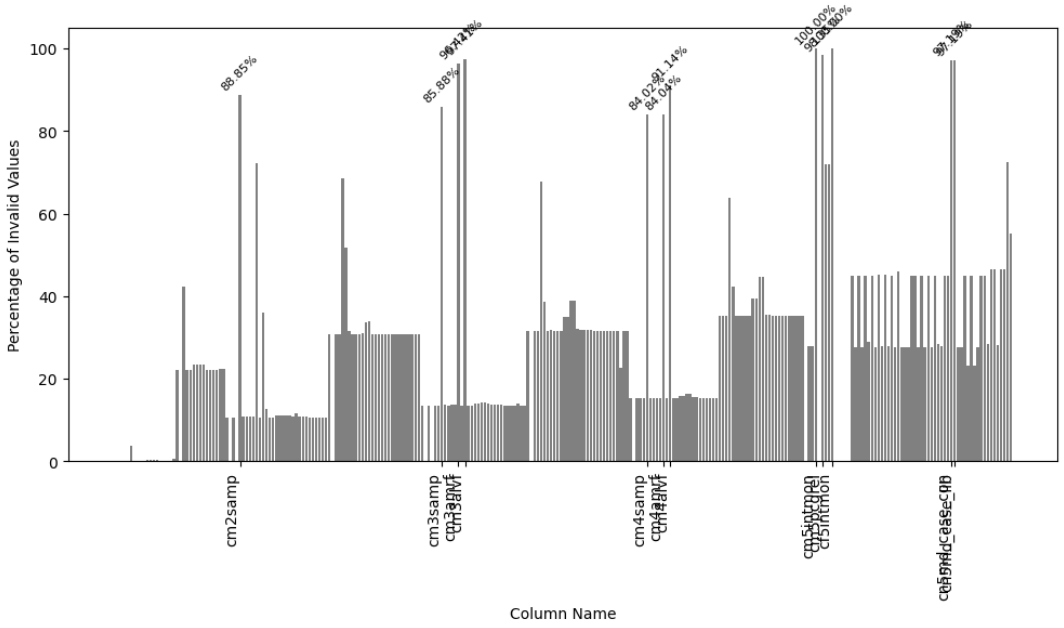

Figure ii: Chart showing what percentage of values, for each column in the reduced (constructed-only) dataset, are either NaNs or negative values. Those with more than 80% invalid values are highlighted using labels.

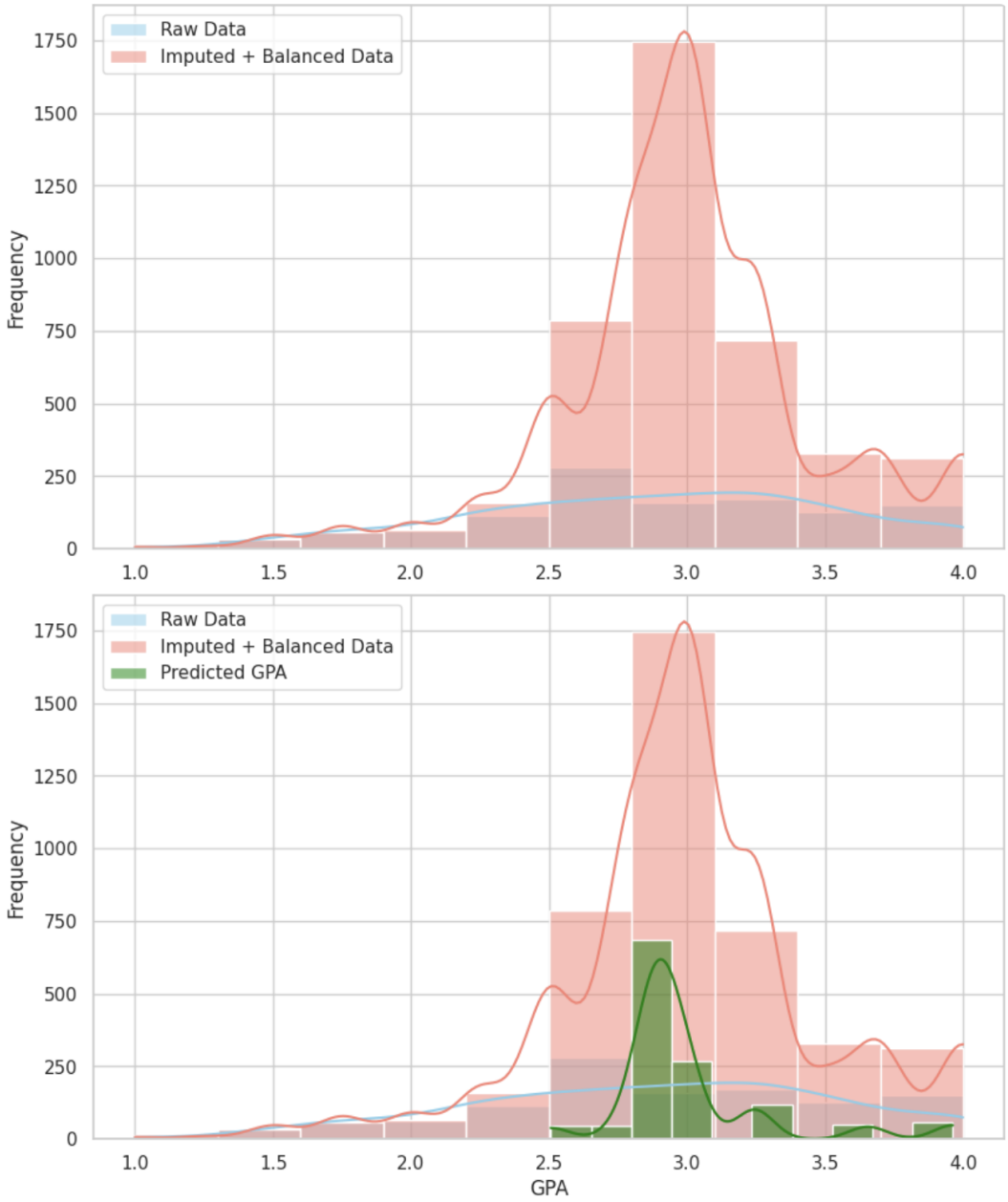

Figure iii: Top: Distribution of GPA in raw outcomes data (NaNs removed) and in the clean data (after Multiple Imputation and balacing on household income category). The figure shows that the cleaning process did keep GPA bounded between 0 and 4.0, and that the distribution peaks around the same mean. Bottom: distribution of predicted GPA values on the test set included. The figure shows that there was no over-prediction (i.e. maximum was smaller than 4). However, the model does not predict values below 2.5.

C Feature Selection

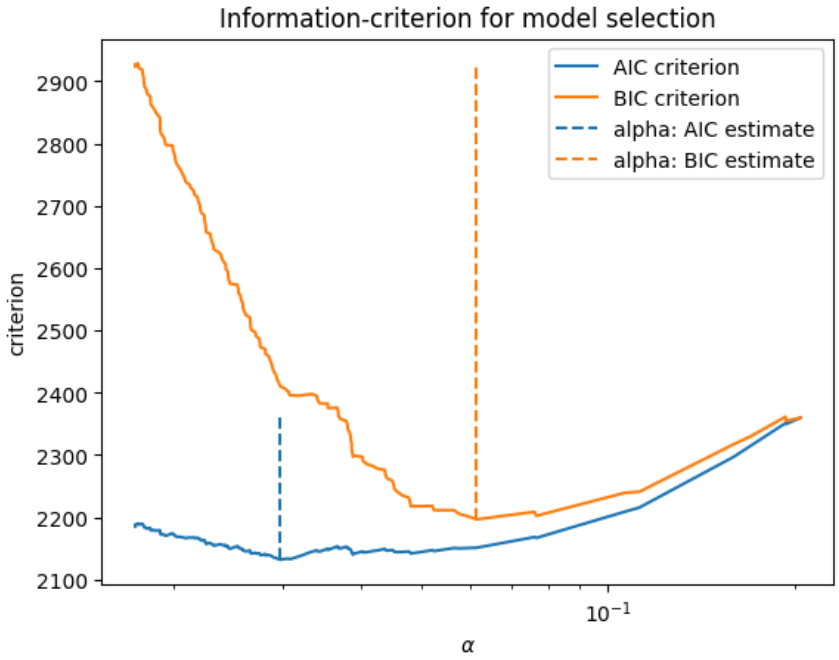

Figure iv: Graph showing the AIC and BIC curves corresponding to a range of α parameters, highlighting the values at which each curve is at a minimum.

variable	label
challengeID	NaN
n5e3	E3. Total income of household
m1i1	What is the highest grade/years of school that you have completed?
m1i3	What is the highest grade/years of school that BF have completed?
m1i6	What was BF doing most of last week (working, going to school,else)?
m4k2$_1$2	What school/program attending: other type of training
p5l5b	L5B. Is/Was this chosen public school . . . ?
hv5_ppvtss	PPVT standard score
hv5_wj9ss	Woodcock Johnson Test 9 standard score
hv5_wj10ss	Woodcock Johnson Test 10 standard score
m4k13	How much you usual earn in job, before taxes and deductions?
m4l1	In past year, total household income from all sources before taxes/deduct?
m3k3a$_1$2	What program/school completed: other training?
f3c41	Is mother currently working, in school, or unemployed?
f5j1	J1. Total household income before taxes/deductions in past 12 months
f5j1a	J1A. Range of total household income before taxes/deductions in past 12 months
f2g3	What is highest grade of school biological father completed?
f2k3k	What kind of school/program are you attending?-Other school
t5b1g	B1G. Child finishes class assignments with time limits
t5b1u	B1U. Child ignores peer distractions when doing class work
f4k3a$_1$2	What program/school completed: other training
f2k5a11	What program/schooling have you completed?-Other school
f4k13	How much you usual earn in job, before taxes and deductions?
m4b9a	Including all child care (not school), how many hrs/wk was he/she in care?
f4l1	In past yr, total household income from all sources before taxes/deduct?
m4c36	What father doing last week-working a reg job, school, or something else?
t5e1	E1. Number of students taught in child's class
m4e7	What current partner doing last week-working reg job, school, or sthg else?
t5e6	E6. Highest level of education completed by aide
k5g2h	G2H. It's hard for me to finish my schoolwork
f2k7b11	What program/schooling did you attend?-Other school
m3c41	Is father currently working, in school, or unemployed?
m5i13	I13. Amount you usually earn in this job before taxes/deductions
m5i13p	I13P. Unit: Amount you usually earn in this job before taxes/deductions
m5j1	J1. Total household income before taxes/deductions in past 12 months
m5j1a	J1A. Range of total household income before taxes/deductions in past 12 months

Table ii: Features chosen from the two-tier feature selection process

D Hyper-parameter Tuning

Model	Hyper-parameters
OLS	None
Elastic Net	alpha, l1_ratio
DT	max_depth, min_samples_split, min_samples_leaf, max_features, min_weight_fraction_leaf
RF	n_estimators, max_features, max_depth, min_samples_split, min_samples_leaf
GB	n_estimators, learning_rate, max_depth
LGBM	boosting, lambda_l1, bagging_fraction, bagging_freq, num_leaves, feature_fraction, max_depth, max_bin, num_iterations, learning_rate, bagging_freq, verbosity, min_data_in_leaf

Table iii: Hyper-parameters to tune for each baseline model should it be chosen for the final training stage

E Feature Importance

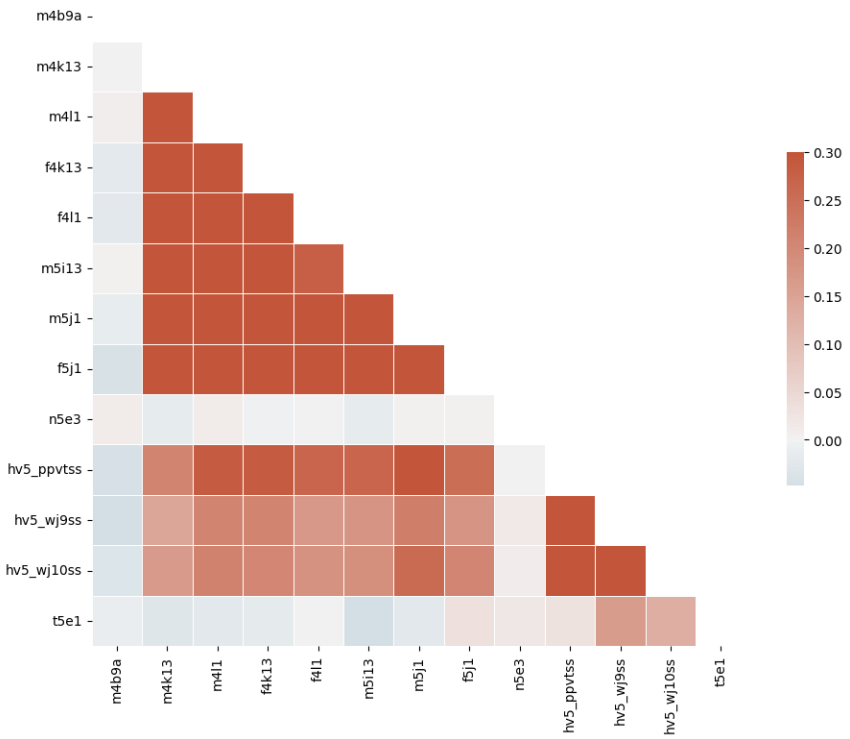

Figure v: Correlation matrix for non-categorical features post-feature-selection, serving as evidence that features are weakly correlated.

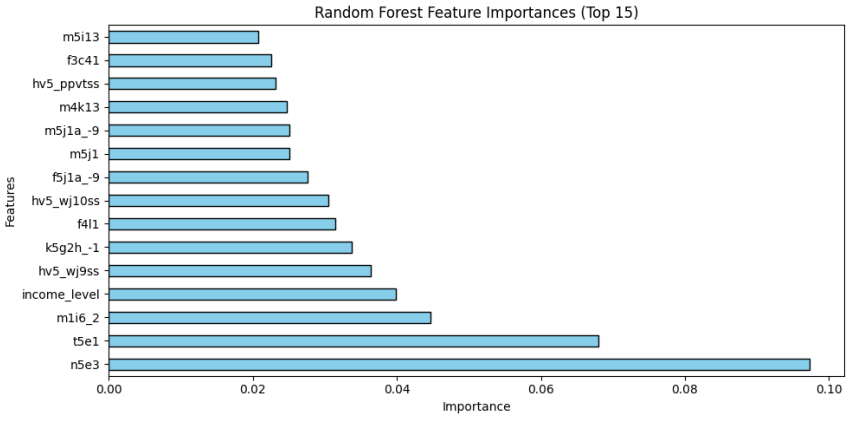

Figure vi: MDI-derived top 15 most important features for predicting GPA using the tuned RF model. The top 5 features (excluding the constructed household income variable as it is a derivative of `n5e3`, are `n5e3` (total household income), `t5e1` (class size), `m1i6` (What was BF doing most of last week (working, going to school,else)?), `hv5_wj9ss` (Woodcock-Johnson test score) and `k5g2h` (It's hard for me to finish my schoolwork).

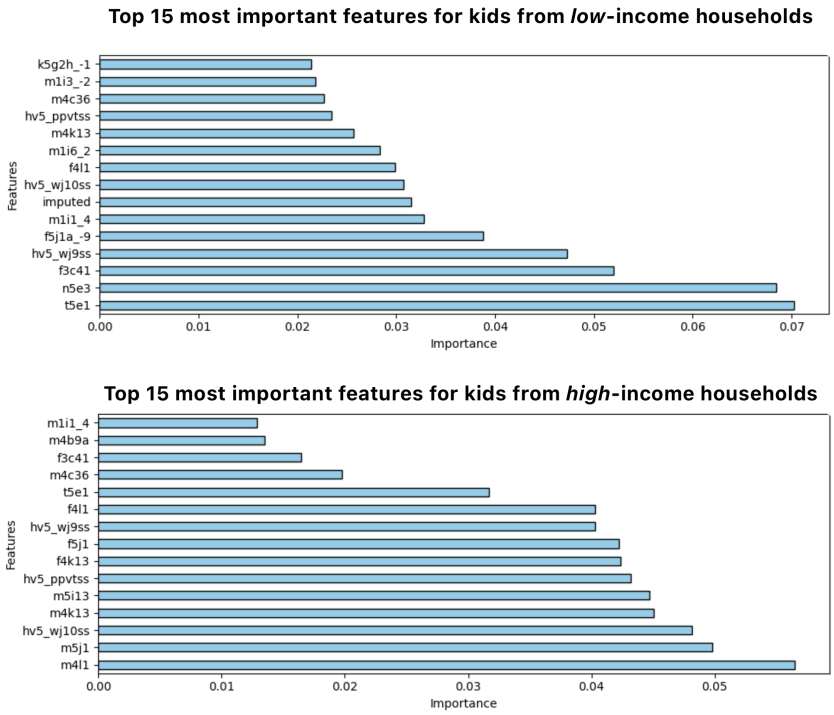

Figure vii: MDI-derived top 15 most important features for predicting GPA using the tuned RF model, for each income class (low income and high income). For low-income search, class size and household income appear to be much more important, while for high-income households maternal metrics of income (e.g. `m4l1`, `m5j1` and `m4k13`) alongside Woodcock-Johnson test scores were the most important.

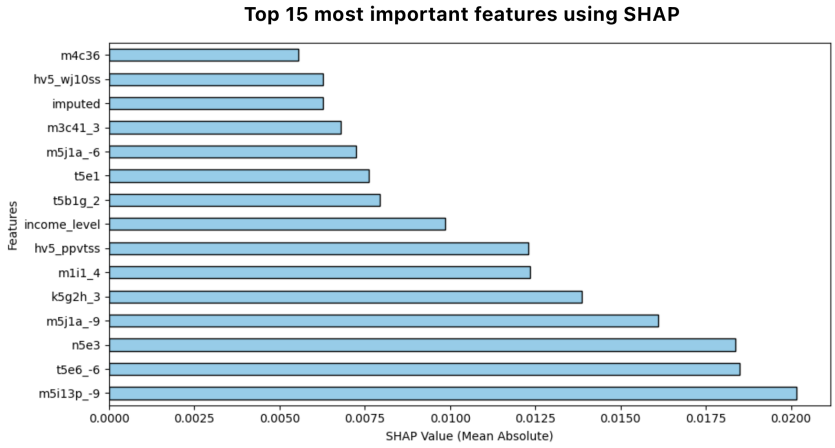

Figure viii: SHAP-derived top 15 most important features for predicting GPA using the tuned RF model. Maternal income metrics feature alongside metrics such as `t5e6` (Highest level of education completed by aide) and `k5g2h` (It's hard for me to finish my schoolwork).

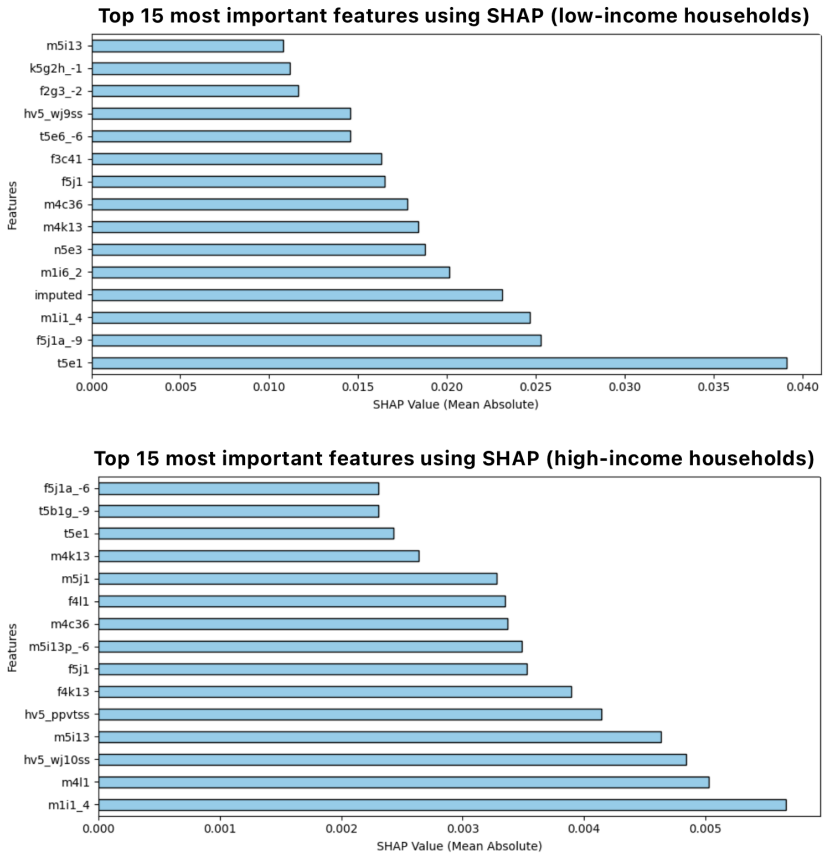

Figure ix: SHAP-derived top 15 most important features for predicting GPA using the tuned RF model, for each income class (low income and high income). For low-income households, class size is once again the most important feature, alongside parental metrics of income, as well as whether the value for GPA was imputed or not. For high-income households, metrics of cognitive ability (i.e. `hv5_wj10ss` and `hv5_ppvtss`) have higher importance alongside income metrics.

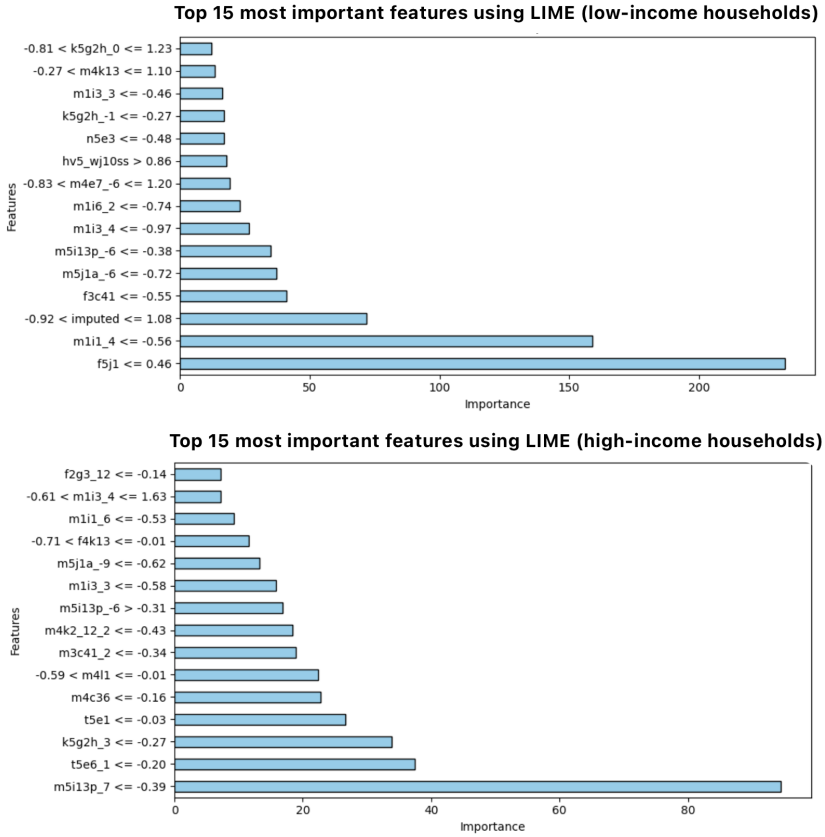

Figure x: LIME-derived top 15 most important features for predicting GPA using the tuned RF model, for each income class (low income and high income). For high-income households, variables relating to learning environment (e.g. difficulty completing schoolwork, aide's highest education level, class size) are closer to the top of the most important features list, while for low-income households, variables relating to household income (mostly maternal) dominate the most important features.