Oxford Internet Institute, University of Oxford

Assignment Cover Sheet

Candidate Number	1053722
Assignment	Applied Machine Learning
Term	HT
Title/Question	Can Machine Learning Models and Diagnostics be used to Predict and Understand Child Life Outcomes?
Word Count	4998

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.

Please remember:

- To attach a second relevant cover sheet if you have a disability such as dyslexia or dyspraxia. These are available from the Higher Degrees Office, but the Disability Advisory Service will be able to guide you.

Can Machine Learning Models and Diagnostics be used to Predict and Understand Child Life Outcomes?

1053722

1 Introduction to Fragile Families Challenge

While traditional sociological research is deductive by nature, an abundance of data and modern technological resources have offered opportunities for an inductive approach to understanding relationships in large and complex data sets (Grimmer, Roberts, & Stewart, 2021). Using data from the Fragile Families and Child Wellbeing Study (FFCW), a detailed birth-cohort study, this analysis seeks to use modern machine learning techniques to optimise and understand the predictions of six key life trajectory variables: 'GPA', 'grit', 'material hardship', 'eviction', 'layoff', and 'job training'. The following paper seeks to optimise three models for each outcome variable, assess their performance, and use model diagnostic techniques to determine the extent to which their decision-making criteria intersect with existing sociological literature (Salganik, Lundberg, Kindel, & McLanahan, 2019).

During the 20th century, divorce rates in the United States dramatically increased, leading to more children growing up with one parent or a parent and stepparent (Waldfogel, Craigie, & Brooks-Gunn, 2010). The Fragile Families and Wellbeing Study aims to understand how this shift has influenced the outcomes of children, from a cognitive and behavioural perspective (Waldfogel et al., 2010). Of the complex mechanisms which influence the development of children, Waldfogel et al. (2010) discuss five important factors: parental resource; parental mental health; parental relationship quality; parenting quality; and father involvement. In fragile families, especially those with a single parent, fewer resources are able to be allocated due to economic challenges. This can impact children through fewer books, clothes, or worse schools in disadvantaged neighborhoods (Ryan, Kalil, & Leininger, 2009). Single or cohabitating mothers have also been found to suffer more from depression, which may impact the quality of parenting offered to their children (Friedlander, Weiss, & Traylor, 1986). The conflict stemming from divorce, and the numerous

adjustments occurring in the aftermath, are other factors which influence parenting and may negatively impact a child's environment (Peterson & Zill, 1986). While parental quality is self-explanatory as an influence on the upbringing of a child, father involvement has been linked with improved behavioural quality and improved cognitive and language abilities (Heiland & Liu, 2006).

This paper will begin by describing the FFCW data set and the various cleaning steps undertaken to optimise the data for modelling. Subsequently, a description of the models used on each outcome variable will be provided. The modeling approaches used in this study include both linear and tree-based methods, and the relevance of their various hyperparameters will be explored. The following portion will explain the motivation for using a randomised search hyperparameter optimisation, and the various data preprocessing steps in each model's pipeline prior to training to the data. The preprocessing techniques include imputation, standardisation, feature selection, and imbalanced class procedures for the binary variables. The results will be composed of each model's performance on all six outcome variables relative to the 'baseline model', explained in the Methods Section. The discussion section will use SHAP, a model diagnostic technique to determine the birth cohort variables which are most predictive to each life outcome, and place these findings within the literature with an inductive approach. The closing section of the essay will provide an interpretation of the models' performance, and the limitations of modelling social phenomena with machine learning methods.

2 Methods

The following section will discuss the data, hyperparameter optimisation methods, model classes, model pipelines, and performance metrics, all included in the Code Appendix.

2.1 Data

The FFCW data set originally contains observations from 4,242 families, with 13,027 features for each. Each feature represents a survey response relating to the parents, children, or environment of each family. Removing columns that are entirely constant, thus not providing predictive value to model, there are 10,594 features. The data set is structured so that negative values represent different forms of non-response, such as '-9', which is 'not in wave', and '-3', which is 'missing'. In addition to this, there are also values which have no label ('NaN'), or state 'Other' or 'Missing'. While the negative values do not provide specific information pertaining to

the question, there are two forms of non-response which are potentially predictive information: '-2', representing 'don't know', and '-1', denoting a refusal to answer.

The data cleaning procedure is an essential step when managing large data sets, as one must maximise the inclusion of potentially useful information and thoughtfully classify the feature types, while avoiding redundant features which increase its dimensionality. The columns with 80% or more 'NaN' values are dropped from the data set entirely, deemed to provide too little information, leading the erasure of 34 columns. Subsequently, all values from '-9' to '-3', and 'Other' or 'Missing' values are grouped into a new label, referred to as '-10'. All columns with over 80% of values as this '-10' grouping were removed, leading to a reduction of 5,265 features. While most features are numerical, there are 81 features composed of strings, which are translated to binary columns to allow their inclusion into the models. The survey includes questions that are either continuous, ordinal, categorical. Due to the size of the data set, this analysis employed a heuristic method of determining the structure of the variables. If a feature contained 15 or fewer unique values, and no 'float' values, it was deemed ordinal or categorical. Prior to creating binary classes for each possible feature in this group, a threshold was used to determine if the values contained in these features contained a strong majority class. Rather than specify the specific minority labels, which may have too few observations to contribute meaningful information to the model, this threshold may allow the model to better capture the impact of being in the non-majority. This was calculated as following: if the most common non-negative value in an ordinal or categorical columns represented at least 80% values, a dummy was created if it was in this majority class or the minority group. Values representing '-10', '-2', and '-1' were included as separate dummies. The remaining columns without this value imbalance had all features dummied. To prevent the '-10', '-2', and '-1' values from contaminating the continuous feature, they were set to 'NaN' values and replaced with a binary classification label where present. The final data set after the cleaning and transformation process includes 29,449 variables (code included in Appendix B file).

2.2 Hyperparameter Optimisation

This analysis uses a combination of manual and randomised search to determine the optimal combination of hyperparameters for each model. Manual searches were initially used to develop an intuitive estimation of a broad range of suitable parameters. Following this, random iterations across a wide and granular parameter space were assessed, selecting the model with the lowest mean squared error (continuous case), or brier score (binary case). Randomised search has been found to select models

with equal or better performance as a sequential grid search, with a lower computation time (Bergstra & Bengio, 2012). This is commonly attributed to its ability to search larger and less promising parameter spaces (Bergstra & Bengio, 2012). Especially in the random forest and boosting models, with many hyperparameters, it allowed for a thorough exploration of optimal hyperparameter combinations, which would have been computationally infeasible with a manual or sequential grid search. While it has been found in Bergstra and Bengio (2012) that most data sets can be optimised without tuning all hyperparameters, the relative importance of certain hyperparameters varies based on the data set. For this reason, a large and diverse hyperparameter space was chosen to optimise models on the data set, made feasible by using an iterative random approach.

2.3 Models

2.3.1 Elastic Net

For the continuous response variables, an elastic net regression is used which linearly combines the feature selection of L1 regularisation and the parameter shrinkage of L2 regularisation (Zou & Hastie, 2005). A linear model is fit to the data which prioritises error minimisation *and* a model coefficient budget specified by a regularisation constant (alpha) and the specified L1 and L2 balance. Each categorical variable is modelled with an elastic net logistic regression with stochastic gradient descent learning. Similarly to an elastic net logistic regression, the L1 ratio must be optimised, which determines the relative balance of L1 (feature selection) versus L2 (shrinkage) regularisation. The second optimised parameter is alpha, which is a regularisation constant, which places greater constraint on the model as it increases. Where this model differs from traditional elastic net logistic regression in its learning method: at each iteration, the loss gradient for a sample of training data is estimated, and the model is updated with a decreasing learning rate until a minimum is found for the logistic loss function.

2.3.2 Random Forest Classifier

A random forest is an ensemble of regression or classification trees that fit numerous trees to the data set and averages the total result. In this analysis, regression trees are fit to the continuous variables and classification trees are fit to the binary outcomes. The main components of single decision trees are nodes and branches, which subdivide the data set into mutually exclusive subsets (Song & Lu, 2015). Responses are predicted by finding an optimal split minimising a constant piece-wise

loss function, then constructing the nodes and branches in a hierarchical manner, with a prediction in the bottom tier. The random forest uses bootstrapping to generate T decision trees, which are fit to the data set using a specified subsample of features. The random selection of features reduces the risk of overfitting to the training set. The average of the predictions from all the generated decision trees is used to generate the random forest prediction. In the continuous case, the outcomes will be the average of all continuous predictions, and in the binary case, it is based on the proportion of predicted classes.

Increasing the number of estimators in a decision tree can help to improve the model's accuracy, but slow down the training process (Probst & Boulesteix, 2017). Due to this constraint, the number of estimators is included in the randomised search to find the number of trees which achieves the accuracy without including more trees than necessary. From a bias-variance perspective, important hyperparameters to tune are the maximum depth of the tree, which increase its complexity, the minimum samples to create a split, which constrains the model as it increases, and the minimum sample to create a 'leaf' at the base of the tree, similarly constraining complexity as it increases.

2.3.3 Gradient Boosting

The third model implemented on each variable is gradient boosting regression (if continuous) or classification (if binary). Similarly to a random forest, boosting involves an ensemble of decision trees. In boosting algorithms, new decision trees, known as weak learners, are added to account for the residuals of prior models. The loss function used to define errors in the continuous case is mean squared error, and a logistic loss function was used in the binary case. The minimisation of these differentiable loss functions is why this model is referred to as *gradient* boosting. The predictions of all trees are then averaged to reach a final estimation.

The learning rate decreases the relative importance of each newly added model, reducing the risk of overfitting to the training set. Maximum depth is included as a hyperparameter to optimise the bias-variance tradeoff. The gamma value is another regularisation parameter that 'prunes' the tree moving upwards based on the gain of a node relative to the parameter threshold. The minimum child weight requires a minimum number of samples to create a new node, limiting the complexity of the model. Subsampling performs a selection of training instances for each iteration, which increases the variance as the sample moves towards the entire data set. Similarly to random forests, another hyperparameter is the fraction of features used in the training set, to increase the bias and limit overfitting. This hyperparameter is

specified at the tree-level and node-level. Compared to random forests, boosting algorithms are more at risk of overfitting as the number of estimators increases, due to the iterative process of fitting residuals (Boehmke & Greenwell, 2019). All of these hyperparameters are specified in the randomised search to select the hyperparameters which best predict the cross-validation set.

2.4 Model Pipeline

The first step in the pipeline used for continuous or categorical response variables is to impute the missing input data. The randomised-search cross-validation technique included both mean and median as imputation strategies, which were chosen at random with the rest of the model parameters. This analysis omitted implementing a regression-based iterative approach to computing missing values, as this approach has been found to be ineffective with high dimensional data, and reduced the performance of the models (Deng, Chang, Ido, & Long, 2016).

The second step in the pipeline is the standardisation of input data. The randomised hyperparameter search included both a min-max scaler and a robust scaler, which were randomly combined with the other hyperparameters. The min-max scaler sets the minimum value for a feature as 0 and the maximum value as 1, linearly mapping each value to fall between them. With the binary variables created in the data cleaning process, this method naturally preserved their values while scaling the other features. The robust scaler is more robust to outliers relative to the min-max scaler, mapping the variables based on each value's relation to the 25th and 75th percentile (Cao, Stojkovic, & Obradovic, 2016).

For continuous response variables, a z-score standardisation of the response label was integrated in each model's pipeline. This standardisation method allowed each response variable to be interpreted in terms of the number of standard deviations from the mean. This method allows the outcome variables to be assessed on a comparable scale for model diagnostics (Caldwell et al., 2019). A quantile transformer designed to normalise the response variables was attempted for skewed variables such as material hardship (Figure 1), but negatively contributed to the models performance, and was therefore omitted from the analysis.

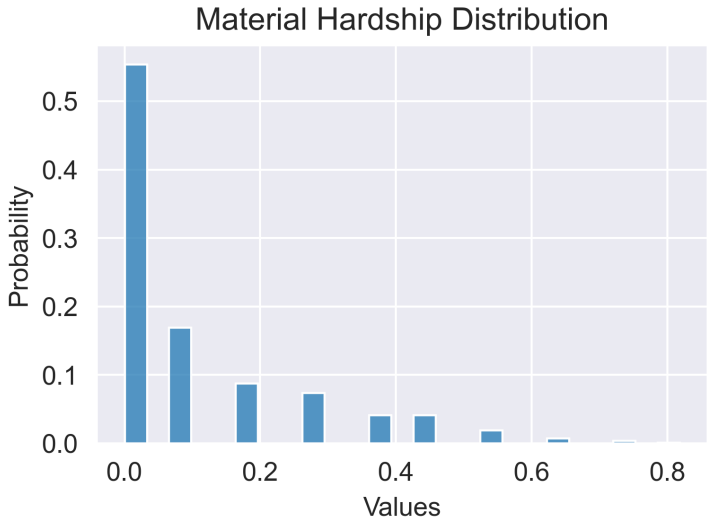

Figure 1: Distribution of Material Hardship Outcome Variable

For categorical response variables, which did not require any standardisation method, this analysis resamples the training set minority class. The binary response variables in the data set display a strong imbalance, shown in Figure 2. This analysis originally attempted two methods for handling the imbalanced labels: upsampling the minority class, and increasing the penalty for a misclassified minority value. With highly imbalanced response variables, both linear and tree-based models tend to be biased towards the majority class, with lower sensitivity towards the minority class (Lee, 2014; Burnaev, Erofeev, & Papanov, 2015). In the hyperparameter search, it was found that resampling produced improved brier scores relative to increasing the penalty for a misclassified minority value.

For classification models with heavily imbalanced classes, re-sampling the minority class to appear at equal frequencies as the majority can lead to greater sensitivity towards the minority class (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). The resampling function in this analysis is known as SMOTE: synthetic minority oversampling technique. Rather than a random sample and duplication of minority observations, SMOTE operates by generating synthetic data that is similar to minority class examples (Chawla et al., 2002). It selects the nearest neighbors of a given observation, finds the difference between the two observations, and selects a random point along the line segment between them (Chawla et al., 2002). SMOTE generalises the minority class' decision region, a potentially helpful application when introducing new data points from the test set (Chawla et al., 2002).

The final step in the pipeline, prior to the training of models, was a feature selection component. Due to the ability of elastic net classifiers to perform both

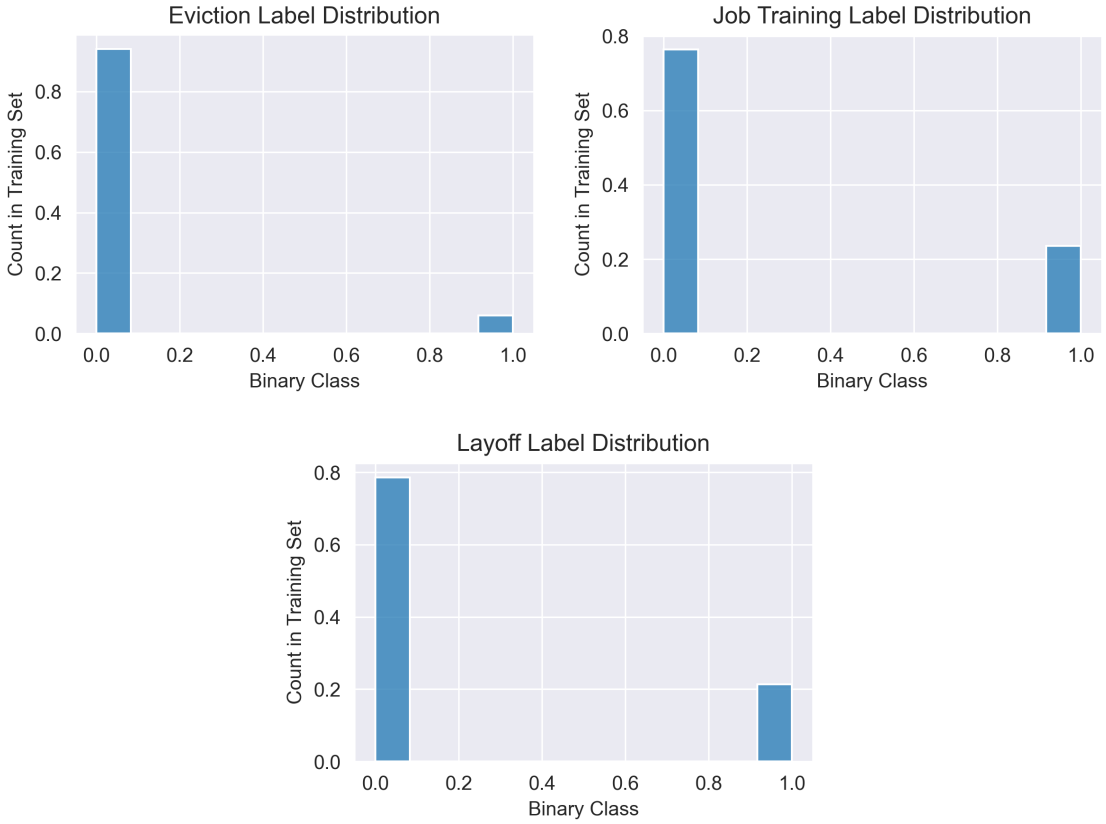

Figure 2: Distribution of Binary Outcomes for 'Eviction', 'Layoff', 'Jobtrainig'

shrinkage *and* large-scale feature selection according to Marafino, John Boscardin, and Adams Dudley (2015), an additional feature selection was not included for that class of models. While boosting and random-forest models have been found to perform well with high dimensional data, their performance tends to degrade when the high-dimensional data is noisy (Capitaine, Genuer, & Thiébaut, 2020; Liu & Tsang, 2017). For this reason, an additional parameter in the randomised search selected the k-best input features based on their ANOVA f-score relative to the training output. The randomisation of this parameter allowed for an assessment of model performance with various feature space sizes, along with the other hyperparameters.

2.5 Performance Metrics

For cross-validation and hyperparameter optimisation, the training set included in the Fragile Families Challenge was used. The final assessment of mean-square error and R^2 was performed on a holdout 'test set', which is presented in the results. Calculated for each model is the mean squared error (MSE) and R^2 of each model. The mean squared error is defined with the following equation, where e is the residual

of each model output relative to the true value (Boehmke & Greenwell, 2019):

$$MSE = \frac{1}{n}\sum_{j=1}^{n} e_j^2 \tag{1}$$

The R^2 value is a separate metric that calculates the percentage of the total variation in the response variable explained by the input variables (Hamilton, Ghert, & Simpson, 2015). An R^2 value of 1 implies that all variation in the output variable is captured by the model, and 0 implies that no outcome variability is explained by the model (Hamilton et al., 2015).

For each classification model, three scores are provided: The Brier Score, precision, and recall. The Brier score is an accuracy metric that calculates the error of estimated probabilities compared to the true value labels (Redelmeier, Bloch, & Hickam, 1991). Its formula is represented by the following equation, where o_t is the true label value, and f_t is the classifier's estimated probability:

$$BS = \frac{1}{N}\sum_{t=1}^{N} (f_t - o_t)^2 \tag{2}$$

Precision is represented by the following equation:

$$Precision = \frac{true\ positives}{true\ positives + false\ positives} \tag{3}$$

It represents the proportion of correctly classified positive cases, divided by all cases labelled as positive. Recall is calculated from the following equation, representing the number of positive cases that were classified as such, relative to the total number of positive cases.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{4}$$

The precision and recall calculations require a predicted label. Due to the up-sampling used to create equal class balances in the cross-validated training data, a threshold of 50% was in the first set of precision and recall values (denoted P (50%, R 50%). Subsequently, an approach to optimise each model's threshold based on the training data was used to generate P** and R**. Each model's respective threshold value was optimised on a portion of the training set held out prior to cross-validation. The threshold chosen maximises the geometric mean, which is calculated as the $\sqrt{(TruePositiveRate) * (1 - FalsePositiveRate)}$. An example of the g-mean optimised point on the ROC curve is shown in Figure 3, using the gradient boosting model for 'Job Training' as an example.

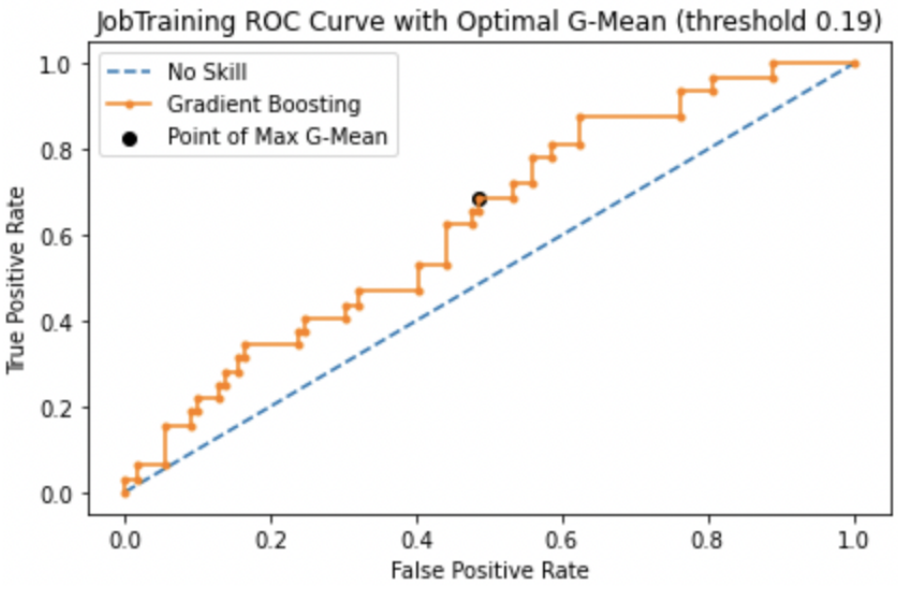

Figure 3: Job Training ROC Curve with Optimal G-Mean Point (Gradient Boosting)

The baseline MSE and Brier Score provided in the results are the scores of a simple linear or logistic regression with four predictor variables selected by an expert in the field, and are included for comparison with this analysis' model performance (Salganik et al., 2020).

3 Results

3.1 Continuous Response Variables

For 'GPA', the elastic net model's MSE of 0.38 and R^2 of 0.08 were the lowest performing metrics of all three models. The random forest model's MSE of 0.35 and R^2 of 0.16 were marginally outperformed by the gradient boosting model's MSE of 0.34 and R^2 of 0.18. All three models outperformed the baseline MSE.

For 'grit', the gradient boosting was the highest performing on MSE, with a value of 0.23, compared to the elastic net's value of 0.25 and random forest's MSE of 0.35. The random forest's R^2 of 0.16 was an improvement on both the elastic net (0.005) and the gradient boosting model (0.05). None of the three models produced an R^2 greater than the baseline.

For 'Material Hardship', the third continuous variable, all three mean squared errors were highly similar, with the random forest and gradient boosting models both outputting 0.019, and the elastic net 0.02. The elastic net's R^2 was the lowest of all three (0.034), with the random forest and elastic net displaying 0.2 and 0.19,

respectively. The elastic net, random forest, and gradient boosting all produced higher R^2 relative to the baseline.

Variable	Model	Baseline MSE	MSE	R2
GPA	ElasticNet	0.39	0.38	0.08
	Random Forest		0.35	0.16
	Gradient Boosting		0.34	0.18
Grit	ElasticNet	0.21	0.25	0.005
	Random Forest		0.35	0.16
	Gradient Boosting		0.23	0.05
Material Hardship	ElasticNet	0.028	0.02	0.034
	Random Forest		0.019	0.2
	Gradient Boosting		0.019	0.19

Table 1: Results for Continuous Predictor Algorithms – MSE and R^2; **Submitted to Class Challenge Leaderboard**

3.2 Binary Response Variables

The 'Eviction' gradient boosting model had the lowest (best performing) Brier Score, with a value of 0.052 – the only model outperforming the baseline. For grit, the random forest and gradient boosting models have a Brier Score of 0.168, relative to the baseline of 0.17. For 'Job Training', the random forest model has the lowest Brier Score with a value of 0.16, relative to the baseline of 0.2.

The precision and recall values vary substantially based on whether the 50% or optimal threshold is used. For the optimal threshold on 'Eviction', the elastic net and gradient boosting have the two highest precision values of 0.18, and random

forest has the highest recall of 0.69. For the 'Layoff' variable, the gradient boosting model has the highest optimised precision value at 0.86, and the random forest has the highest recall at 0.3. For 'Job Training', the elastic net has the highest optimised precision value at 0.76, and the highest recall at 0.24. With the 50% classification threshold, the results demonstrate that 4 out of the 9 models had 0 precision and 0 recall.

Variable	Model	Baseline Brier Score	Brier Score	P (50%)	R (50%)	P **	R **
Eviction	ElasticNet	0.053	0.065	0.190	0.061	0.18	0.1
	Random Forest		0.054	0.0	0.0	0.12	0.69
	Gradient Boosting *		0.052	0.15	0.18	0.18	0.26
Layoff	ElasticNet	0.17	0.37	0.51	0.21	0.31	0.19
	Random Forest		0.168	0.0	0.0	0.29	0.3
	Gradient Boosting *		0.166	0.0	0.0	0.86	0.23
Job Training	ElasticNet	0.2	0.22	0.295	0.329	0.76	0.24
	Random Forest		0.16	0.0	0.0	0.28	0.23
	Gradient Boosting		0.177	0.05	0.48	0.48	0.05

Table 2: Results for Continuous Predictor Algorithms – Brier Score, Precision, Recall (50% and optimised threshold**); **Submitted to Class Challenge Leaderboard**; * Highest Accuracy on Class Challenge Leaderboard

4 Model Diagnostics and Discussion

The first section of this discussion will use model diagnostic techniques to determine the variables most important to each outcome prediction, and place the findings within the context of existing sociological literature. The approach used in this analysis is SHAP (SHapley Addition exPlanations), which will be applied to the model with the lowest MSE or Brier Score for each respective variable. SHAP is a kernel-based estimation method for determining Shapley values, based on local surrogate models (Molnar, 2019). Each feature in the data set is assigned a contribution to the prediction, which allows for an analysis of important features to traditionally 'black box' models, such as random forest and gradient boosting algorithms (Molnar, 2019). The relevant features and their relationship to the outcome variable are discussed in each section, with a total list of variable names and descriptions in appendix A.

4.0.1 GPA: Gradient Boosting

The SHAP analysis for 'GPA' uses the gradient boosting model, with the results shown in 4. As all outcome variables in this analysis are standardised, the baseline value shown in the figure is 0.00, the mean, and the values around the mean are in terms of standard deviations. The variables are ranked in descending order of importance. The first three variables are the 'Woodcock Johnson Test Score', 'PPVT Standard Score', and the 'PPVT Raw Score' (ch5wj10ss, ch5ppvtss, ch5ppvtraw). These variables, which relate to the testing ability of the child, are intuitively positively related to their 'GPA', represented in Figure 4. The fourth highest variable is a dummy representing whether the father had a college or graduate education (cf1edu_4), which is positively related to 'GPA'. The fifth variable is a dummy of whether or not the child had above-average math skills (t5c13c_4). The last three variables are all related to the income and poverty level of the child (cf5povcob, cm5hhinc, cf5hhincb)

The performance of the student on standardised tests and their mathematical abilities are clear indicators of general success in school. That these variables out of the 29,499 were highlighted is useful for assessing the model's ability to select important features. The importance of a father's education on the 'GPA' of the child is concurrent with literature on determinants of child outcomes. Dubow, Boxer, and Huesmann (2009) find that when controlling for other socio-economic factors, the educational attainment of a child's parent is predictive of their child's academic attainment and performance. The other variables are all income-related, communi-

cating that families with more income have children with a higher 'GPA'. This is in line with the theories of child outcomes offered in the introduction, potentially through fewer educational resources or schools in poorer neighborhoods. Additionally, poverty can manifest in poorer parenting quality, child well-being, and cognitive development (Dubow et al., 2009).

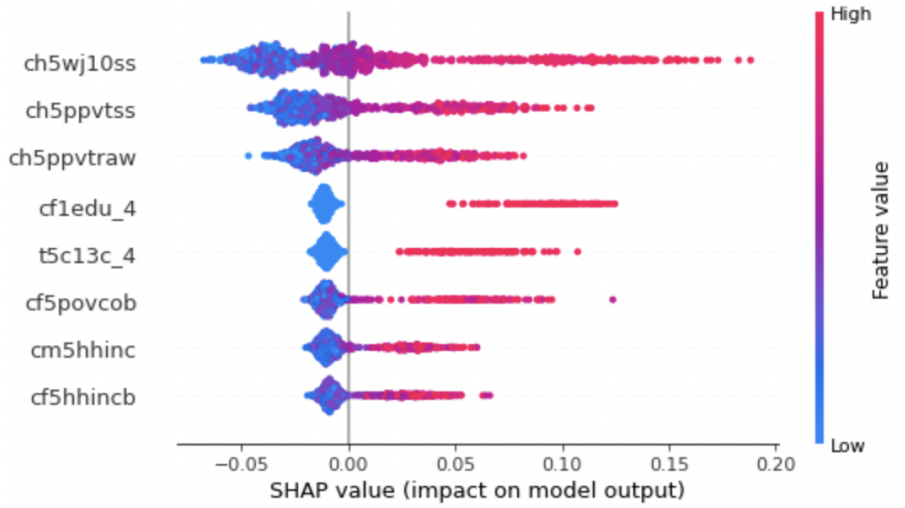

Figure 4: SHAP Analysis – 8 Most Important Features for GPA Prediction

4.0.2 Grit: Gradient Boosting

As displayed in Figure 5, the three most important variables for 'grit' predictions all relate to test scores, displaying that lower scores are associated with higher 'grit' (ch5wj9raw, ch5ppvtss, ch5wj10raw). Additionally, having no trouble paying attention (k5g2d_0), not getting distracted easily (k5g2f_0), and the parent knowing what the child does in free time (p5i26_4) are all positively associated with 'grit'. The dummy variables indicating worrying a little bit about school (k5g21_1) is negatively associated with 'grit', along with sometimes following things through to the end (k5g1e_2).

The literature on 'grit' in children proposes a weak to moderate relationship with educational variables, whereas the gradient boosting model's top three values were all educational related (Christopoulou, Lakioti, Pezirkianidis, Karakasidou, & Stalikas, 2018). According to Christopoulou et al. (2018), perseverance plays a key role, potentially explaining the high SHAP importance of not getting distracted easily or not paying attention. Additionally, the survey response of sometimes following things through to the end may demonstrate low perseverance, explaining its negative relationship to 'grit'.

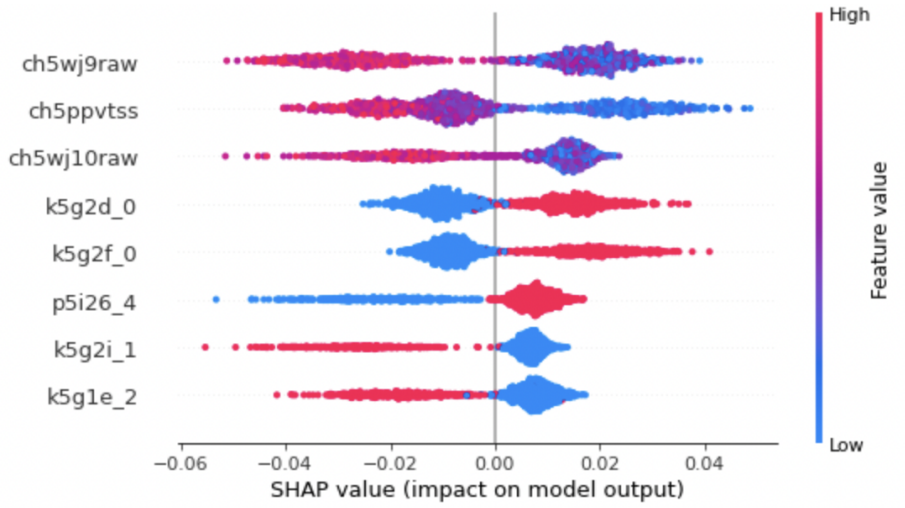

Figure 5: SHAP Analysis – 8 Most Important Features for Grit Prediction

4.0.3 Material Hardship: Random Forest

The variable with the most influence is not being able to pay the electricity bill, which is positively related to hardship, and being able to pay the bill, which is negatively associated (m5f23e_1, m5f23e_2). The third most important is the telephone service being disconnected due to unpaid bills (m5f23k_1), and a negative association with not receiving food stamps in the prior 12 months (m5i20_2). The next variable is not paying full rent/mortgage because there was not enough money (m5g0_1). Having a very positive life satisfaction is also negatively associated with 'material hardship' (m4i23d_1). The final two variables on the list were receiving a free meal in the last 12 months (m5f23a_1), and household income (cm4hhinc).

Of the eight variables most important to the random forest output, seven of them are unambiguously related to income. The topics are related to family troubles due to insufficient funds, an inability to pay bills, or receiving free meals. The variable on life satisfaction can be explained from Bannink, Pearce, and Hope (2016), who discusses that family income and *perception* of family income both contribute to lower self-esteem and life satisfaction.

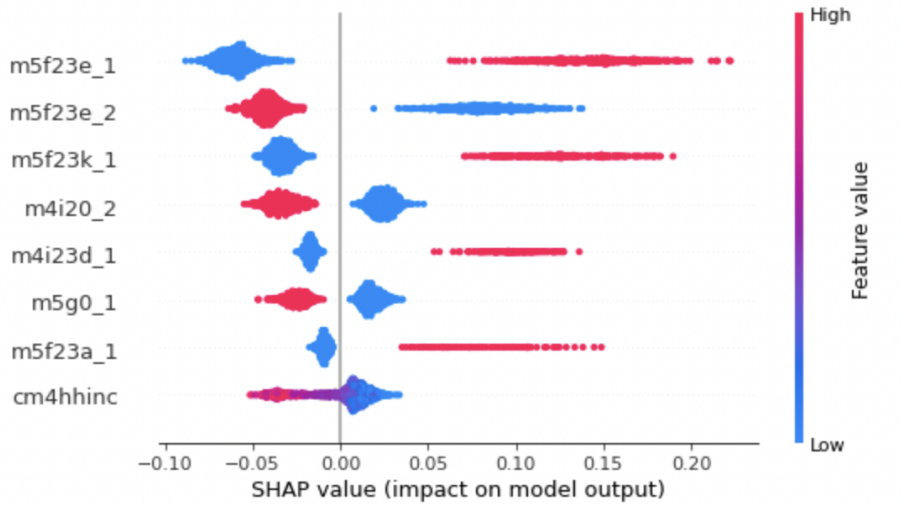

Figure 6: SHAP Analysis – 8 Most Important Features for Material Hardship Prediction

4.0.4 Eviction: Gradient Boosting

The gradient boosting SHAP diagnostic explains that the most influential variables for 'eviction' are not paying rent in the past 12 months (m4i23d_1), and having the telephone service disconnected (m5f23k_1). The third is the money spent eating out (p5j10), which is negatively related to 'eviction'. Subsequently, 'eviction' is positively related to families strongly believing that the bible should be interpreted literally (f3rf_1). The fifth variable is having the telephone service ever disconnected (m3i6a_1), followed by a negative association with the money received from public welfare (m1j2b). The final associations in descending importance are a positive relationship with receiving help from a welfare office or job placement in the last 12 months (m5f7b_1), and a positive relationship with the child 'getting into everything' (p3m18a_a).

There is a clear link between the lack of income available in a family and 'eviction', evidenced in the rent, telephone bill, and eating out survey responses. The positive relationship with religiosity is not supported by the prior work of Desmond and Kimbro (2015a), who finds that religious attendance has a statistically insignificant relationship to 'eviction' in low-income American families. Interestingly, money received from welfare is negatively associated with 'eviction', yet receiving help or work from a welfare office is positively associated with 'eviction'. With US housing rates generally rising and welfare stipends stagnating, less 'eviction' based on greater welfare money is a conceivable outcome (Desmond & Kimbro, 2015b). For the finding that greater help from a welfare office relates to greater values for 'eviction', Desmond and Gershenson (2017) find that drawing the attention of welfare officers

attracts unwanted state attention, making landlords more likely to evict families with children.

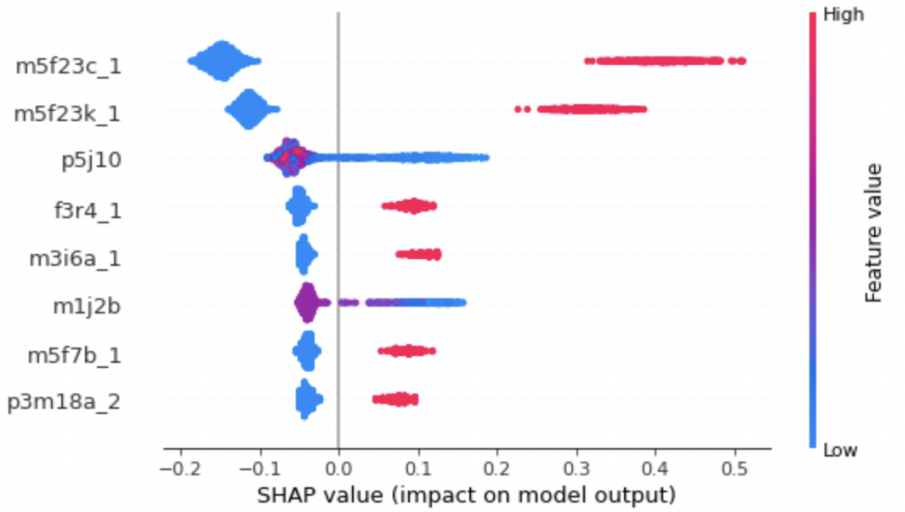

Figure 7: SHAP Analysis – 8 Most Important Features for Eviction Prediction

4.0.5 Layoff: Gradient Boosting

The most influential variable represents the principal component of 9 variables relating to the strength of the relationship between parents (m4d6). The second most important is the '-10' value for family savings, which is a generated non-response label from the cleaning stage (m5j6h_-10). Subsequently, being in excellent health is strongly negatively related to 'layoff' (f4j1_1), along with a father never being in jail (m5b30_2). The following most important variable is not knowing ('-10') whether parents attend informal meetings (t4a10_-10). The next most important variables, which are positively associated with 'layoff', are not receiving help from a non-welfare office (f5f7c_2), feeling very good about oneself as a mother (m3b1_1), and having enough money to see the doctor (f5f23j_2).

The first variable is the first principal component of a range of questions relating to the parent's relationship, with no clear meaning. Similarly, the ambiguity of the term 'informal meetings' leaves interpretation difficult. Interestingly, poor health was found to be associated with being laid off, concurrent with the research of Jusot, Khlat, Rochereau, and Serme (2008), yet *being able* to pay for needed medical assistance is associated with 'layoff' – the two results seeming in opposition.

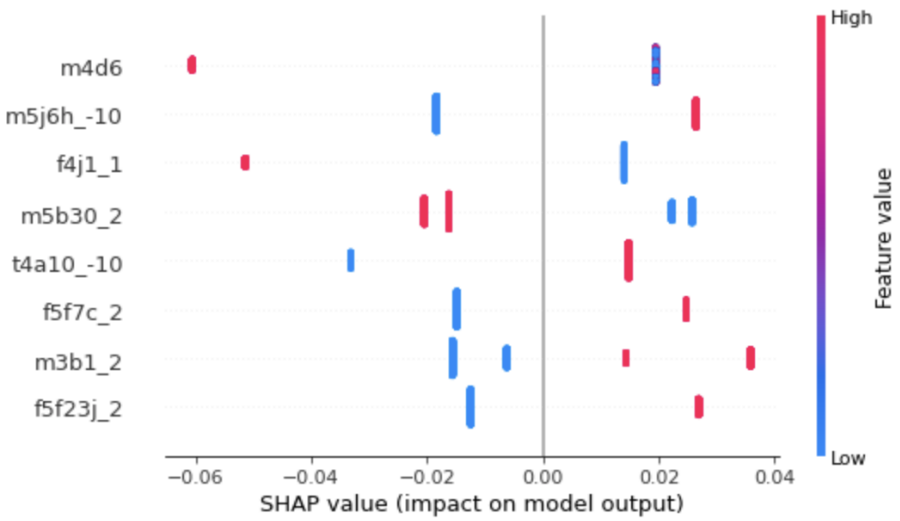

Figure 8: SHAP Analysis – 8 Most Important Features for Layoff Prediction

4.0.6 Job Training: Random Forest

For the gradient boosting model, questions specifically related to the individual's 'job training' were the three most important in the model. Such questions include whether one has taken a class to improve job skills (m4k3b_1), is currently attending any school/training/programs (m5i1_1), or has taken a 'job training' class since the last interview (m5i3b_1). Other positively associated variables include whether one was in the gifted and talented program or not (pfL13f_1, pfL13f_2), their job earnings in the last 12 months (m5i19a, m3k19). Also positively associated with 'job training' is whether the mother had a technical college education (cm5edu_-3).

While the 'job training' variables are intuitively related to the outcome, this diagnosis communicates that individuals who are more gifted are likely to have received 'job training', along with individuals who earn more money. Interestingly, while the father's education was a key variable in 'GPA', it is the mother's education which associates with the child receiving 'job training'. Heinrich (2014) discuss the importance of parents as role models, and that a parent's educational experience may influence a child's decision further job prospects. This is especially true in fragile families, where work may replace welfare, an unattractive alternative, leading to greater motivation to receive adequate 'job training' (Heinrich, 2014).

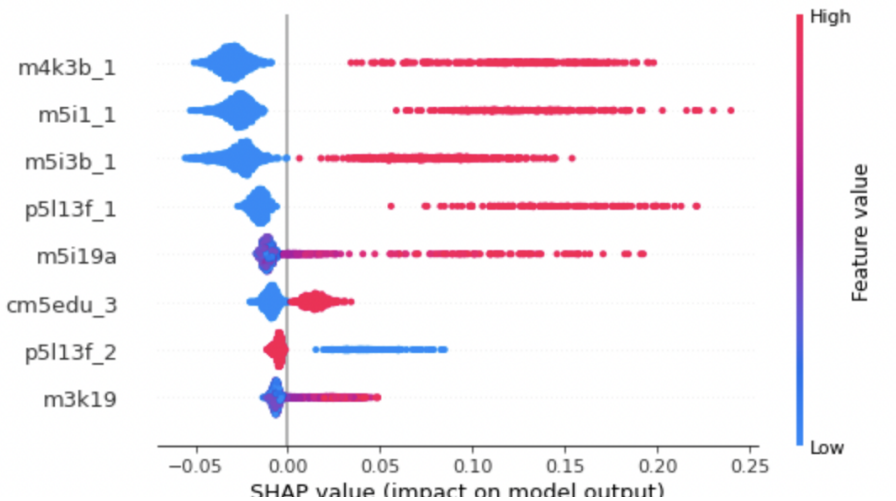

Figure 9: SHAP Analysis – 8 Most Important Features for Job Training Prediction

Despite the ability of the SHAP diagnostic techniques to discern some relationships related to the existing literature, the results section demonstrates the difficulty with which complex social phenomena can be accurately predicted with machine learning techniques. The results for the models, despite their varied approaches to predicting outcomes, do not provide substantial improvements on the baseline scores. Across all continuous variables, the most variation explained by a model is 20%, generated by the random forest on 'job training'. Similarly, in the binary response variables, no model is able to significantly improve on the baseline Brier Scores, or achieve high precision and recall – even when using optimised classification thresholds. Despite the richness of the FFCW data set and apparent model detection of relevant variables, a high proportion of the outcomes remain unexplained.

Using a wide array of data cleaning, pre-processing, hyperparameter optimisation, and modeling techniques, this analysis largely confirms the findings of (Salganik et al., 2020) (2020), that sophisticated machine learning models provide minor improvements, if any, on predicting life outcomes. In the original Fragile Families challenge with 160 submissions, many complex models performed as well or worse than the baseline, with similarly mixed results found in this analysis. The paper has demonstrated that training machine learning models and using model diagnostic techniques to inductively discern relationships may not overcome the inherent uncertainty in real-world phenomena. While an inductive approach to answering questions of social scientific relevance may help to explore relationships in large and complex data sets, this analysis conveys that adequate care must be taken to ensure the models accurately capture the outcome variables if they are to be trusted.

5 Appendix – (Code in Submission Folder)

A Full Variable Names for SHAP Analysis

Feature Name	Description
ch5wj10ss	Woodcock Johnson Test 10 standard score
ch5ppvtss	PPVT standard score
ch5ppvtraw	PPVT raw score
cf1edu_4	Father baseline education – col or grad
t5c13c_4	Child's mathematical skills – above average
cf5povcob	Father's household income/poverty threshold at 9-year
cm5hhinc	Mother's Household income
cf5hhincb	Household income mother report for married/cohab if available
ch5wj9raw	Woodcock Johnson Test 9 raw score
ch5wj10raw	Woodcock Johnson Test 10 raw score
k5g2d_0	It's hard for me to pay attention – not at all true
k5g2f_0	I get distracted easily – not at all true
p5i26_4	Frequency you know what child does during free time – always
k5g2i_1	I worry about doing well in school – a little bit true
k5g1e_2	I follow things through to the end – sometimes
m5f23e_1	Did not pay full amount of gas/oil/electricity bill in past 12 mo – yes
m5f23e_2	Did not pay full amount of gas/oil/electricity bill in past 12 mo – no
m5f23k_1	telephone disconnected in last 12 months – yes
m4i20_2	Time in past 12 mo. you thought might be eligible for food stamps – no
m4i23d_1	12 mo. did not pay full rent/mortgage payments b/c wasn't enough – yes
m5g0_1	How satisfied you are with your life overall – very satisfied
m5f23a_1	Received free food or meals in past 12 months – yes
cm4hhinc	Household income

m5f23c_1	Not paid rent last 12mo - yes
m5f23k_1	Did not pay full about of rent/mortgage 12mo - yes
p5j10	Money spent eating out
f3r4_1	Strong belief that Bible should be literally interpreted
m3i6a_1	Telephone service disconnected 12mo - yes
m1j2b	How much money from pub. assis/welfare?
m5f7b_1	Received help from welfare office/job placement 12mo - yes
p3m18a_2	Child gets into everything – Very True
m4d6	First principal component scale created from m4d6a-i
m5j6h_-10	You or your husband/partner have savings – missing/skip/not in wave
f4j1_1	In general, how is your health? – excellent
m5b30_2	Father has spent any time in jail – no
t4a10_-10	parents attend informal meeting – -10
f5f7c	In the past twelve months, you received help from any other agency – no
m3b1_2	How do you feel about yourself as a mother to child? – very good
f5f23j_2	Someone in hh needed to see doctor but couldn't – no
m4k3b_1	In the last 2y, have you taken any classes to improve your job skills? – yes
m5i1_1	You are currently attending any school/trainings program/classes – yes
m5i3b_1	You have taken classes to improve job skills since last interview – yes
p5L13f_1	Gifted and talented program – yes
m5i19a	Amount earned from all regular jobs in past 12 months
cm5edu_3	Mother's education – some coll, tech
pfL13f_2	Gifted and talented program – no
m3k19	How much did you earn from all regular jobs in past year?

Table 3:

References

Bannink, R., Pearce, A., & Hope, S. (2016). Family income and young adolescents perceived social position: associations with self-esteem and life satisfaction in the UK Millennium Cohort Study. *Archives of Disease in Childhood*, *101*(10), 917–921. Retrieved from https://adc.bmj.com/content/101/10/917 doi: 10.1136/archdischild-2015-309651

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(10), 281-305. Retrieved from http://jmlr.org/papers/v13/bergstra12a.html

Boehmke, B. C., & Greenwell, B. M. (2019). Hands-on machine learning with r..

Burnaev, E., Erofeev, P., & Papanov, A. (2015). Influence of resampling on accuracy of imbalanced classification. In A. Verikas, P. Radeva, & D. Nikolaev (Eds.), *Eighth international conference on machine vision (icmv 2015)* (Vol. 9875, pp. 423–427). SPIE. Retrieved from https://doi.org/10.1117/12.2228523 doi: 10.1117/12.2228523

Caldwell, J. A., Niro, P. J., Farina, E. K., McClung, J. P., Caron, G. R., & Lieberman, H. R. (2019, aug). A Z-score based method for comparing the relative sensitivity of behavioral and physiological metrics including cognitive performance, mood, and hormone levels. *PLOS ONE*, *14*(8), e0220749. Retrieved from https://doi.org/10.1371/journal.pone.0220749

Cao, X. H., Stojkovic, I., & Obradovic, Z. (2016, sep). A robust data scaling algorithm to improve classification accuracies in biomedical data. *BMC bioinformatics*, *17*(1), 359. doi: 10.1186/s12859-016-1236-x

Capitaine, L., Genuer, R., & Thiébaut, R. (2020, aug). Random forests for high-dimensional longitudinal data. *Statistical Methods in Medical Research*, *30*(1), 166–184. Retrieved from https://doi.org/10.1177/0962280220946080 doi: 10.1177/0962280220946080

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002, June). Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, *16*(1), 321–357.

Christopoulou, M., Lakioti, A., Pezirkianidis, C., Karakasidou, E., & Stalikas, A. (2018). The Role of Grit in Education: A Systematic Review. *Psychology*, *09*, 2951–2971. doi: 10.4236/psych.2018.915171

Deng, Y., Chang, C., Ido, M. S., & Long, Q. (2016). Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data. *Scientific Reports*, *6*(1), 21689. Retrieved from https://doi.org/10.1038/srep21689 doi: 10.1038/srep21689

Desmond, M., & Gershenson, C. (2017). Who gets evicted? Assessing individual, neighborhood, and network factors. *Social Science Research*, *62*, 362–377. Retrieved from https://www.sciencedirect.com/science/article/pii/S0049089X16300977 doi: https://doi.org/10.1016/j.ssresearch.2016.08.017

Desmond, M., & Kimbro, R. T. (2015a). Eviction's Fallout: Housing, Hardship, and Health. *Social Forces*, *94*(1), 295–324. Retrieved from https://doi.org/10.1093/sf/sov044 doi: 10.1093/sf/sov044

Desmond, M., & Kimbro, R. T. (2015b). Eviction's Fallout: Housing, Hardship, and Health. *Social Forces*, *94*(1), 295–324. Retrieved from https://doi.org/10.1093/sf/sov044 doi: 10.1093/sf/sov044

Dubow, E. F., Boxer, P., & Huesmann, L. R. (2009, jul). Long-term Effects of Parents' Education on Children's Educational and Occupational Success: Mediation by Family Interactions, Child Aggression, and Teenage Aspirations. *Merrill-Palmer quarterly (Wayne State University. Press)*, *55*(3), 224–249. Retrieved from https://pubmed.ncbi.nlm.nih.gov/20390050https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2853053/ doi: 10.1353/mpq.0.0030

Friedlander, S., Weiss, D. S., & Traylor, J. (1986). Assessing the influence of maternal depression on the validity of the child behavior checklist. *Journal of abnormal child psychology*, *14*(1), 123–133.

Grimmer, J., Roberts, M., & Stewart, B. (2021). Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science*, *24*. doi: 10.1146/annurev-polisci-053119-015921

Hamilton, D. F., Ghert, M., & Simpson, A. H. R. W. (2015, sep). Interpreting regression models in clinical outcome studies. *Bone joint research*, *4*(9), 152–153. Retrieved from https://pubmed.ncbi.nlm.nih.gov/26392591https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4678365/ doi: 10.1302/2046-3758.49.2000571

Heiland, F., & Liu, S. H. (2006). Family structure and wellbeing of out-of-wedlock children: The significance of the biological parents' relationship. *Demographic Research*, *15*, 61–104.

Heinrich, C. J. (2014, may). Parents' Employment and Children's Wellbeing. *The Future of Children*, *24*(1), 121–146. Retrieved from http://www.jstor.org/stable/23723386

Jusot, F., Khlat, M., Rochereau, T., & Serme, C. (2008). Job loss from poor health, smoking and obesity: a national prospective survey in France. *Journal of Epidemiology & Community Health*, *62*(4), 332–337. Retrieved from https://jech.bmj.com/content/62/4/332 doi: 10.1136/jech.2007.060772

Lee, P. H. (2014). Resampling Methods Improve the Predictive Power of Modeling in Class-Imbalanced Datasets. *International Journal of Environmental Research and Public Health*, *11*(9), 9776–9789. Retrieved from https://www.mdpi.com/1660-4601/11/9/9776 doi: 10.3390/ijerph110909776

Liu, W., & Tsang, I. W. (2017). Making decision trees feasible in ultrahigh feature and label dimensions. *Journal of Machine Learning Research*, *18*(81), 1-36. Retrieved from http://jmlr.org/papers/v18/16-466.html

Marafino, B. J., John Boscardin, W., & Adams Dudley, R. (2015). Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. *Journal of Biomedical Informatics*, *54*, 114–120. Retrieved from https://www.sciencedirect.com/science/article/pii/S1532046415000210 doi: https://doi.org/10.1016/j.jbi.2015.02.003

Molnar, C. (2019). *Interpretable machine learning.* (https://christophm.github.io/interpretable-ml-book/)

Peterson, J. L., & Zill, N. (1986). Marital disruption, parent-child relationships, and

behavior problems in children. *Journal of Marriage and the Family*, 295–307.

Probst, P., & Boulesteix, A.-L. (2017, January). To tune or not to tune the number of trees in random forest. *J. Mach. Learn. Res.*, *18*(1), 6673–6690.

Redelmeier, D. A., Bloch, D. A., & Hickam, D. H. (1991). Assessing predictive accuracy: How to compare brier scores. *Journal of Clinical Epidemiology*, *44*(11), 1141–1146. Retrieved from https://www.sciencedirect.com/science/article/pii/089543569190146Z doi: https://doi.org/10.1016/0895-4356(91)90146-Z

Ryan, R. M., Kalil, A., & Leininger, L. (2009). Low-income mothers' private safety nets and children's socioemotional well-being. *Journal of Marriage and Family*, *71*(2), 278–297.

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., ... McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, *117*(15), 8398–8403. Retrieved from https://www.pnas.org/content/117/15/8398 doi: 10.1073/pnas.1915006117

Salganik, M. J., Lundberg, I., Kindel, A. T., & McLanahan, S. (2019). Introduction to the Special Collection on the Fragile Families Challenge. *Socius: Sociological Research for a Dynamic World*, *5*, 237802311987158. doi: 10.1177/2378023119871580

Song, Y.-Y., & Lu, Y. (2015, apr). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130–135. doi: 10.11919/j.issn.1002-0829.215044

Waldfogel, J., Craigie, T.-A., & Brooks-Gunn, J. (2010). Fragile families and child wellbeing. *The Future of children*, *20*(2), 87–112. Retrieved from https://pubmed.ncbi.nlm.nih.gov/20964133https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074431/ doi: 10.1353/foc.2010.0002

Zou, H., & Hastie, T. (2005, may). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *67*(2), 301–320. Retrieved from http://www.jstor.org/stable/3647580