Oxford Internet Institute, University of Oxford

Assignment Cover Sheet

Candidate Number	1047904
Assignment	Applied Machine Learning
Term	Hilary Term 2021
Title/Question	Home, Sweet Home? An application of machine learning to predict life outcomes for fragile families
Word Count	4996(excl. captions and formulae)

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.

Home, Sweet Home? An application of machine learning to predict life outcomes for fragile families

1047904

1 Introduction

What are the conditions and capabilities of unmarried parents, especially fathers? How do children born into these families fare?

The *Fragile Families and Child Wellbeing Study* (FFCWS), a longitudinal birth cohort study launched in 1998, continues to provide an unprecedented opportunity to address questions in the areas of non-marital childbearing, fatherhood and social welfare on the grounds of an ever-increasing data foundation (Reichman, Teitler, Garfinkel, & McLanahan, 2001; OPR, 2021).

In 2017, 437 researchers allied by the aspiration to improve the odds of upward social mobility for disadvantaged children in the US were granted access to some 54 million data points in order to trial the predictability of 6 life outcomes. This research was conducted in the context of an event which should become a poster child for global, scientific mass collaboration: The *Fragile Families Challenge* (Salganik, Lundberg, Kindel, & McLanahan, 2019; Makhijani, 2017). While sobering results across response variables and machine learning approaches continue to lure researchers to chance their luck in modelling the data set forth by the challenge, unfortunate cases of algorithmic policy enforcement question the euphoric use of predictive models to police families and children at risk (Kleinberg, Ludwig, Mullainathan, & Obermeyer, 2015). Researchers' reliance on data skewed to low-income households and the excessive use of proxy variables recurrently place models such as the *Allegheny Family Screening Tool* (AFST) in the public crosshairs for inaccurate, socio-demographically biased predictions in matters of child safety (Eubanks, 2018). We understand such findings as urgent prompt for further scholarly pursuits to decipher the complex, social dynamics at play with novel techniques.

With one in four children being estimated to suffer from child abuse or neglect in their lifetime (Brown, Yilanli, & Rabbitt, 2020), the topic of child well-being constitutes a long-established item on the societal agenda. Yet, the persistence of the

COVID-19 crisis has motivated health professionals to signal the creeping presence of a "secondary pandemic" (Green, 2020; Adams, 2020). While the risk of child neglect rises under conditions of containment, measures meant to decelerate the disease outbreak cut children off from a broad spectrum of supportive relationships upheld through school, the community and extended familial ties. In times where children's well-being stands and falls with the intactness of the very family construct, concerns about malnutrition, inaccessibility to education and spillover effects of deteriorating mental health on adults' caregiving capacities frame the public debate (ACPHA, 2020).

1.1 Objective

Motivated by the quest to improve on existing *Fragile Families Challenge* (FFC) outcomes and encouraged by the pronounced attention familial dynamics and child well-being should receive in today's times, this study re-discovers the FFC dataset. Using more than 12,000 features together with response data for a fraction of households, this study aims to predict 6 distinct life outcomes recorded at child age 15: (I) child *grade point average* (GPA), (II) child *grit*, (III) household *material hardship* (IV) household *eviction*, (V) caregiver *layoff* as well as (VI) caregiver participation in *job training*.

Findings will provide an entry lane for prospective attempts to explore this dataset while contributing to the broader debate about the predictability of life outcomes using machine learning.

1.2 Related Work

The following section is tripartite in nature: We briefly outline existing research drawn from FFCWS data, refer to publications of scholars who attempted the challenge and conclude by spotlighting a body of literature mobilizing non-FFCWS predictors to study effects of neighborhood, educational access and racial disparities on children's upward social mobility.

The extensive spectrum of indicators collected through the FFCWS has spawned more than 750 studies to date (Salganik et al., 2019), many of which study the influencing factors on well-being years before the FFC was held. Exploring the impact of family structure on the cognitive development of children, Liu and Heiland (2012) discover that marriage in the aftermath of childbirth positively affects children's cognitive development as indicated by *Peabody Picture Vocabulary Test* (PPVT)

scores. Sigle-Rushton and McLanahan's (2002) simulations on how well fragile families would fare if they were married or full-time employed informs the conclusion that differentials between married and unwed couples do not resolve through change in marital status alone. Whereas marriage would pull around 47% of low-income, single mothers over the federal poverty threshold, the creation of full-time employment opportunities for either parent emerges as even more effective lever to poverty reduction. Investigating the effect of family structure on various health outcomes including asthma, obesity and hospitalizations, Bzostek and Beck (2008) find that children raised by single mothers are in worse health, even after controlling for socio-demographic attributes such as ethnicity, maternal age or education.

While previous studies predominantly investigate child well-being using classical statistics, scholars resort to machine learning techniques to predict life outcomes of surveyed families in the context of the FFC. Synthesizing the essence of 12 challenge participant's publications, Salganik et al.(2019) sketch out possible courses of action concerning imputation, variable construal & selection and modelling. While the later served as seminal guidance to plan this study, contributions by Rigobon et al. (2019), Stanescu et al. (2019) and Goode et al. (2019) were of particular inspiration.

Appreciating distinct missingness patterns such as monotonic increases in parental non-response with each successive survey wave, Goode et al. (2019) trial semi-automated imputation strategies which logically infer missing values from related questions within and across survey waves and adjacent respondents. Stanescu et al. (2019) use *Amelia*, an imputation algorithm exploiting the correlation structures between predictors. Spearheading the leaderboard on life outcomes *GPA* , *grit*, and *layoff*, Rigobon et al. (2019) engage in variable selection using *LASSO* and *Mutual Information*. The researchers further infer additional variables to capture latent effects suspected in non-response behaviours. At the end of the challenge, the best performing models only mark single to lower double-digit departures from associated baselines (Salganik et al., 2019). Results resonate with remarks in pre-challenge literature which mourn the prevalence of inconclusive insights into how differentials between traditional and fragile families evolve (Waldfogel, Craigie, & Brooks-Gunn, 2010) while affiliating with more recent scholarship speculating limits to predictive modelling endeavors in complex social systems (Hofman, Sharma, & Watts, 2017).

Eventually, research driven by Harvard scholars Raj Chetty and Nathaniel Hendren demonstrate the utility of both alternative data sources and additional features in reasoning discrepancies in social mobility prospects. Concluding a 3.4 percentile increase in children's income in association with a 10 percentile increase in parental

income from administrative records covering 40 million children and their parents, Chetty et al. (2014) underpin the inheritability of social mobility and the moderating effect of neighborhood. Chetty et al. (2016) further demonstrate that re-locating to low-poverty neighborhoods sooner rather than later coincides with higher college attendance and lower single parenthood rates. Experiences and exposures as early as kindergarden are found to impact future earning potential (Chetty et al., 2011) whereby race continues to rule a pronounced social divide between black and white boys for 99% of US census tracts (Chetty, Hendren, Jones, & Porter, 2020).

Contrastful findings in social science literature paired with moderate successes across previous challenge attempts motivate the formulation of two exploratory research question ahead of analysis:

RQ_1: To which extent are life outcomes predictable based on FFC data?

RQ_2: What are the major determinants of the life outcomes set forth by the FFC?

2 Methods

2.1 Data Source

The data set used for analysis represents a purpose-built version of the *Fragile Families and Child Wellbeing study* (FFCWS). Grounded on a stratified, multi-stage sample of almost 5000 infants born between 1998 and 2000 across 20 major cities in the US, the FFCWS aspires to monitor childrens' life trajectories by structurally interviewing mothers, fathers and primary caregivers in longitudinal fashion. In oversampling births to unmarried parents at a 3:1 ratio, research deliberately focuses on gaining insight into family dynamics of unwed parents and their children (Reichman et al., 2001). Data collection at birth was complemented by repeat surveying at the age of 1,3, 5,9, 15 and 22 (Princeton University, 2021a). Questions cover a broad spectrum of factors researchers consider indicative of child well-being (Salganik et al., 2019).

Data encompassing survey results from childbirth up to preliminary life outcomes measured at the age of 15 was retrieved from the *Office of Population Research* (OPR) data archive upon special permission. The download constitutes the 2018 re-release of the original file collection made available to challenge participants. Subsequent analyses ensued from the following three files: (*1*) *background.csv* containing 4,242 observations (one per child) on 13,026 features, (*2*) *train.csv* containing 2,021 observations (one per child in the training set) across 6 response variables and (*3*) *constantVariables.txt* aiding the early removal of constant features.

2.2 Feature type tagging & type casting

To promote confident decision-making in matters of imputation, feature encoding and scaling, features were annotated by type to start with. Concerns about the recency, implicity and unhandy format of the provisioned codebook (*codebook-FFChallenge.txt*) motivated a full metadata download of the *FFCWS Metadata Explorer* (Princeton University, 2021b) to devise a programmatic lookup routine. While the FCCWS survey methodology distinguishes between 5 data types (*Continuous, Ordered Categorical, Unordered Categorical, Binary & String*), we attempted to qualify each variable as either continuous (supplemented with prefix "CONT_") or categorical (supplemented with prefix "CAT_") following Rigobon et al.'s (2019) taxonomy.

Exact matches were achieved for 98.85% of the 10,594 features after constant feature removal. Scales for an additional 27 of the 121 non-assignable features were recovered through consultation of an open-sourced participant dictionary (Rigobon et al., 2019). A comparatively small remainder of 94 features was eventually discarded to anticipate erroneous treatment during pre-processing as well as to prevent untraceable features from tainting the interpretability of model results. The type-tagging exercise retained 5,787 categorical and 4,713 continuous features for further consideration. Erroneous type castings for 73 continuous variables were rectified. Conspicuous date-like and quasi-binary values were discretized for variables "cf4fint" and "m5b12" in line with metadata information. Aptitude test score attributes "ch5wj9pr" and "ch5wj10pr", erroneously qualified as "string" in the metadata, were converted to continuous to avoid feature set inflation during one-hot encoding.

2.3 Feature Engineering before train-test split

2.3.1 Preliminary Feature Pruning

Having removed constant and untraceable columns, the remaining 10,500 features were investigated for high proportions of missing values and low variance.

Elimination of high NaN features. Variables with more than 80% missing values were discarded (Figure 1). Removals were not solely based on the presence of regular NaNs but factor in occurrences of any of the 9 numerical missingness codes (negative integers from -1 to -9). 5,344 features were discarded in consequence.

Figure 2 contrasts the distribution of missingness codes before (Subplot 1) and after (Subplot 2) feature pruning.

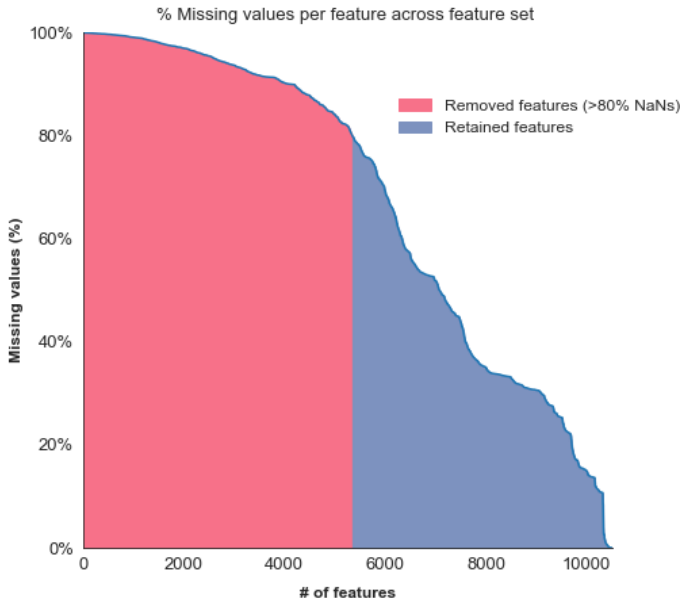


Figure 1: Share of Missing values per feature (sorted in descending order)

Missingness referred to as "not in wave" (missing code "-9" occurring $14.4 * 10^6$ times) and "valid skip" (missing code "-6" occurring $12.9 * 10^6$ times) outnumber ordinary NaNs. Whereas the former epitomizes high levels of survey non-response, the latter rests on hard-to-verify interviewer judgement (Lundberg, 2018).

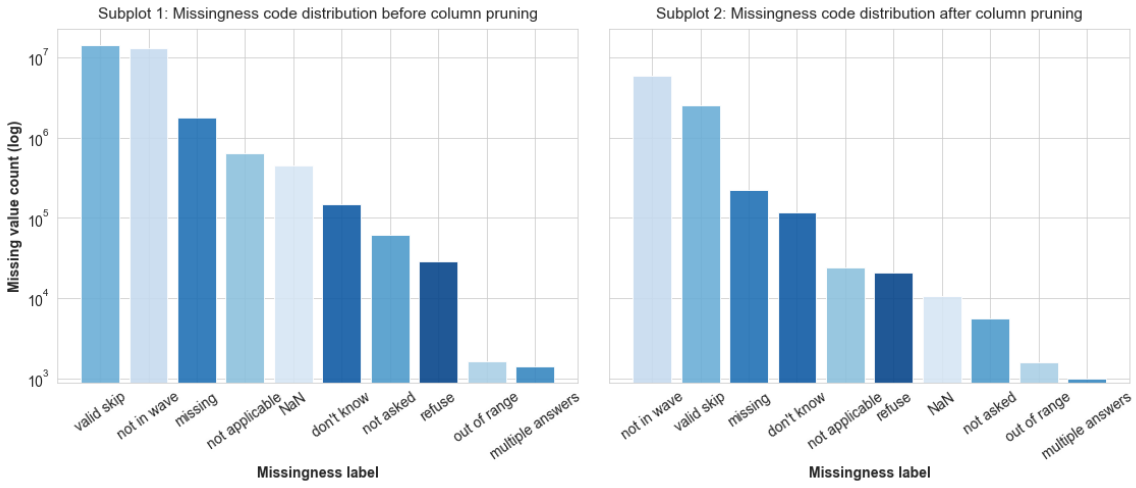


Figure 2: Missing code distribution before and after "high NaN" column pruning

Column elimination considerably mitigated the occurrence of all missingness codes while reshuffling their frequency ranks. The remaining 5,165 features were assessed for low levels of variance.

Elimination of highly invariant features. 65 continuous variables with standard deviations below a uniform threshold of 0.05 were precluded from the dataset. In order to make features comparable, *min-max scaling* was temporarily applied to re-scale distributions into ranges between 0 and 1 before computing standard deviations.

Since the variability inherent in categorical data grounds on the notions of "unalikability" or diversity rather than deviation from the mean (Kader & Perry, 2007), *Shannon Entropy* was calculated to quantify information uncertainty for all categorical features. In dividing each variable's entropy score by its corresponding maximum entropy $log_b n$, we detached measured variabilities from the inflating effect of larger category bandwidths and bounded values by 0 and 1. With n representing the number distinct categories of feature x and $p(x_i)$ denoting the discrete probability of the i^{th} category, normalized *Entropy* $H_n(x)$ is calculated as follows:

$$H_n(x) = -\sum_i \frac{p_i log_b p_i}{log_b n} \tag{1}$$

13 features with normalized entropy scores ≤ 0.2 were removed. Figure 3 provides a visual supplement to the just described variability assessments implying the exclusion of several categorical (Subplot 1) and continuous features (Subplot 2).

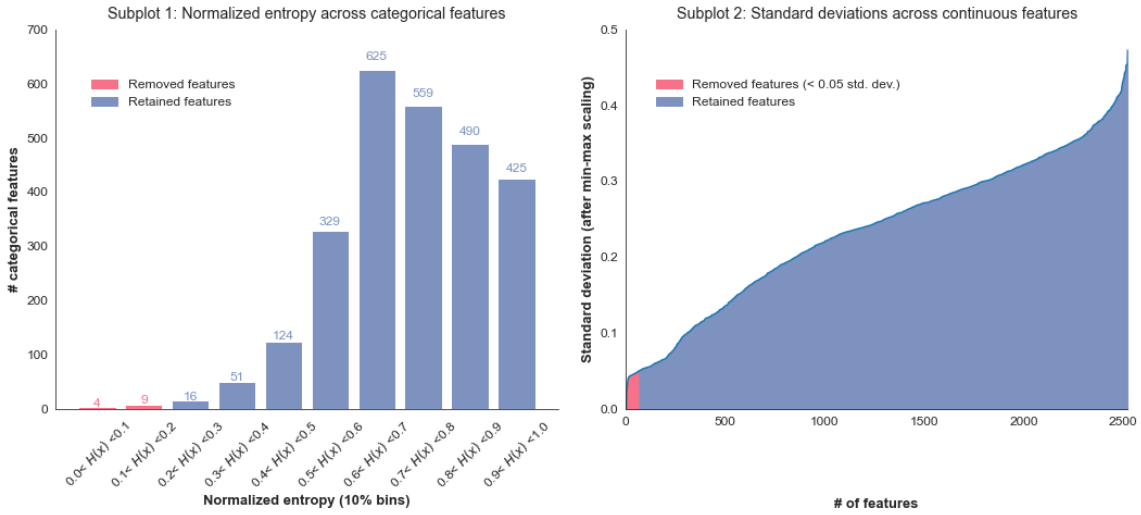


Figure 3: Identification of categorical and continuous features with low variance

In sum, the removal of high NaN and low variance features downsized our feature space from 10,500 to 5,078 variables.

2.3.2 Selective Feature Creation

We re-purposed two missingness codes for the inference of new features meant to capture whether or not a respondent refused to answer a given question (missing code "-1") or did not know the answer (missing code "-2"). This decision draws on Rigobon et al.'s (2019) suspicion that missingness immediately attributable to respondent behavior may be indicative of an effect worth preserving from imputation.

Consequently, 1,367 binary features indicating refusal and 1,681 features indicating lack of knowledge are created, raising the feature space to a total of 8,126 features prior to train-test split. Derived features are annotated with prefix "DER_" and suffixes "_refusal" or "_dontknow" to ease identification.

Figure 4 recaps all preparational measures taken up to this point.

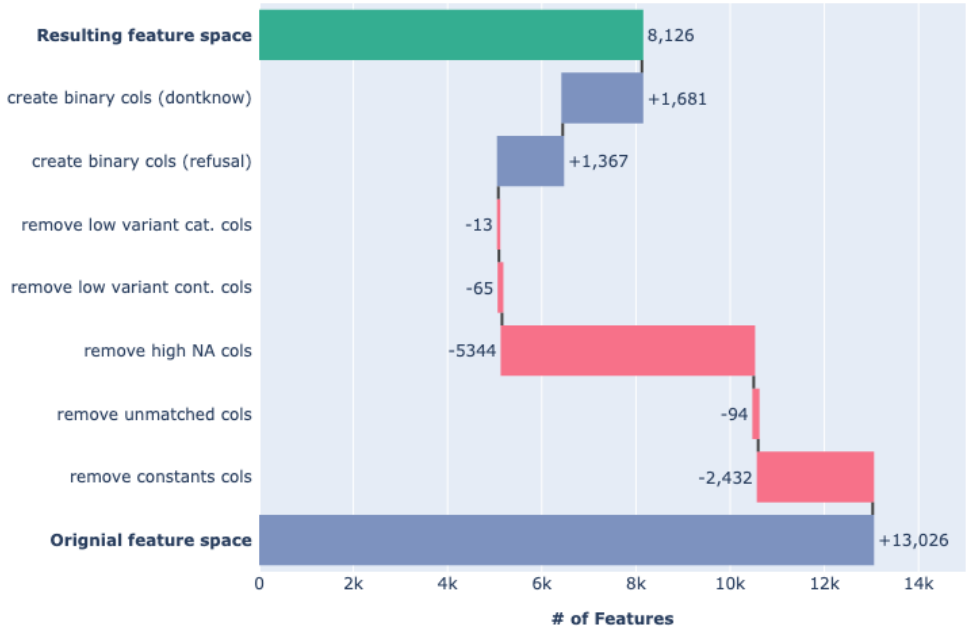


Figure 4: Overview of feature pruning decisions prior to train-test split

2.4 Model Selection

We opted for a Gradient-boosted tree (GBT) classifier to address the 6 prediction tasks. Our choice was motivated by competitive outcomes achieved with the former by Rigobon et al. (2019) during the challenge. Using the *XGBoost* implementation, the researchers topped the leaderboard on the continuous outcomes *gpa* and *grit* while reporting high scores for 5 out 6 response variables in relation to all competing algorithms attempted in preparation of submission. Relative to *Random Forest* where an ensemble of trees is trained concurrently but in isolation from one another, GBT trains successive trees on the residual errors of their respective predecessors. While incremental assembly of a strong learner is computationally taxing, it may translate into higher predictive performance (Salganik et al., 2019). We departed from Rigobon et al.'s (2019) implementation in several ways:

Model Implementation. We utilized the *LightGBM* library to implement GBT. The library's resorts to histogram-based algorithms for continuous feature bucketing promises faster runtime at lower compute (Jin & Agrawal, 2003). While other GBT implementations grow trees in depth-wise fashion, *LightGBM*'s leaf-wise (best-first) approach to tree expansion may translate into additional predictive performance (Shi, 2007; Light GBM, 2021b). In search for a performant yet time-effective classifier, *LightGBM* was eventually favored over *XGBoost* considering available time and resource constraints.Using regressor and classifier implementations of the former, we attempted predictions of continuous and binary response variables respectively.

Hyperparameter Search and Sampling. Rigobon et al.'s (2019) run *Grid Search* over a coarse-grained hyper-parameter response surface where each of the 5 parameters in scope is represented by a list of 2-3 discrete values. As the number of possible combinations grows exponentially with the number of hyper-parameters in scope, discretizing the grid provides computational remedy but comes at the expense of poor coverage in dimensions that are important. Bergstra and Bengio's (2012) find that *Random Search* is more efficient since not all parameters are equally important to adjust. We hence adopted *Random Search* both, to accommodate the inclusion of additional parameters despite computational constrains, and to address limited insight into parameter interaction effects hampering the confident specification of a discrete grid apriori. Since the goodness of search yet heavily depends on the search space defined, we considered Rigobon et al.'s (2019) parameter bandwidths as reliable reference to construct uniform discrete and continuous distributions for parameter sampling. For instance, while the researchers trained models with either 100 or 1000 boosting iterations, we randomly sample values in between these bounds.

In extension to Rigobon et al.'s (2019) 3-fold cross-validation routine, we validate any sampled parameter combination over 5 folds to alleviate concerns about high bias and variance in accordance with empirical benchmarks (James, Witten, Hastie, & Tibshirani, 2013)

Hyperparameters. In order to win in speed and generalizability, Rigobon et al.(2019) uses 5 hyper-parameters to tune their GBT implementations. We reconsider these hyperparamters for our study while adding 3 additional dimensions to the mix. Since *LightGMB*'s leaf-wise splitting procedure may overfit for small datasets, we regularize the minimum number of data points that must fall into a node for it to be added as well as the number of leaves. We clip the distribution of the latter at the upper bound of the maximum tree depth parameter distribution to anticipate adversarial effects on overfitting as outlined in the documentation (Light GBM, 2021c). Appendix 1 briefly outlines all parameters considered.

2.5 Feature Engineering after train-test split

We devised a 6-step pre-processing pipeline to prepare the data for the prediction of continuous or discrete outcomes. Figure 5 outlines the applicability of each step contingent on data type.

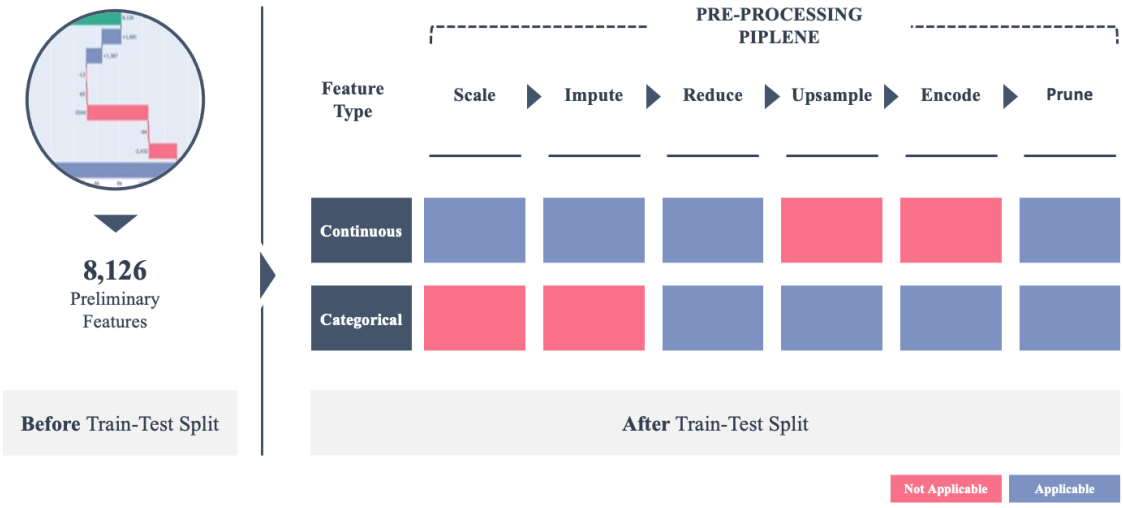


Figure 5: Overview of pre-processing pipeline after train-test split

Each step is briefly outlined in the following:

(1) Scale. To obtain a "honest" estimate of population means and standard deviations, continuous independent and dependent variables were scaled at the outset.

While assuming Gaussian distributions across all 2,459 continuous features and the 3 responses constitutes a daunting claim that we are not able to validate in breadth nor depth given feature set size and limitations in domain knowledge, standardization was preferred of normalization. Both approaches aid modelling by preventing features from dominating objective functions for reasons of scale and endorse the learning smaller weights in favor of model stability. Yet, standardization to mean 0 and unit variance is found to better preserve information about outliers without suffering from their presence to the extent min-max normalization does when re-applying previously learned scaling to unseen data (Raschka, 2015). While tree-based models are exceptionally robust to arbitrarily scaled data, scaling was deemed necessary, not only to address the effects of skew found for some responses (Figure 6) but also to establish amenable grounds for different algorithms.

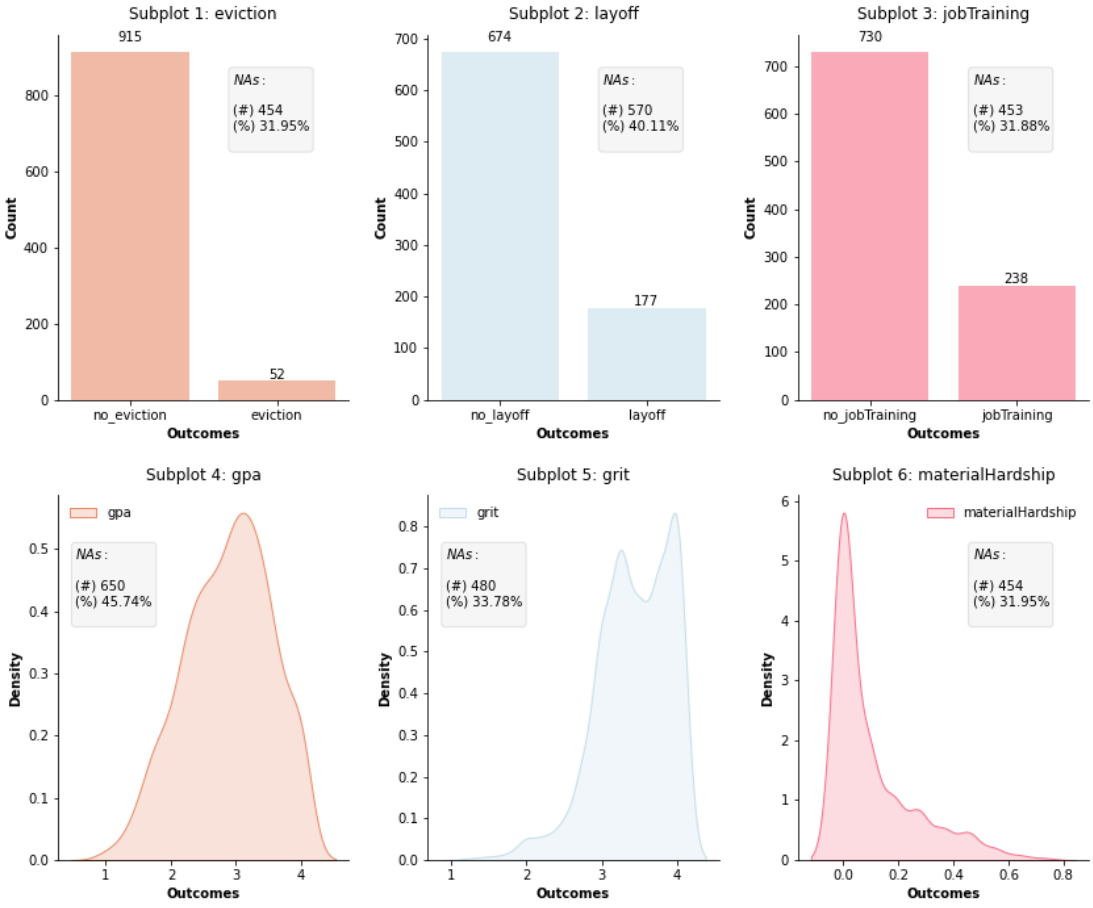


Figure 6: Overview of response variable distributions in train set

Appendix 2 contrasts the influence of different transformation on continuous response variable distributions for supplementary reference.

(2) Impute. Having replaced all missingness codes with "NaN" placeholders, continuous features were imputed using the median to anticipate distorting effects of potential outliers. In suspicion of more complex patterns of missingness among covariates, we also experimented with multivariate imputation (Kenward & Carpenter, 2007). Initializing missing values with median, a regressor (X, y) was fit to predict missing values in a given column y based on its 5 nearest neighbouring features X over 10 iterations. Response variables were not subjected to imputation.

(3) Reduce. Contingent on the response variable due for prediction, training and validation sets were reduced to observations with non-missing response values. Considering missingness rates between 30% to 45% across responses (Figure 6), input spaces to learn from substantially shrunk across prediction tasks.

(4) Upsample. Minority class upsampling was used to prevent one-sided learning without compromising on sample size in light of salient class imbalances across discrete response variables (Figure 6).

(5) Encode. Categorical features were one-hot encoded to prevent models from erroneously assuming natural orderings in absence of true ordinal relationships. Contrary to the common recommendation in statistics to drop reference categories in order to prevent *multicollinearity* (James et al., 2013), resulting dummy encodings were preserved in their entirety. The motivation is to persevere the flexibility to represent categories in the test set not encountered when fitting the encoder on the training data. While one-hot encoded columns remediate newly encountered categories with zero-entries during transform, the encoder loses its ability to differentiate between reference category and unseen categories if the reference category is dropped (Martin, 2019).

(6) Prune. After one-hot encoding, the ratio of number of observations to number of features is almost 1:8 in the train set. While internal feature selection approaches at play for tree-based ensemble techniques like *Gradient Boosting* are known to safeguard against unimportant variables, we experiment with preliminary feature pruning using *Mutual Information* (MI) to ease runtime, further endorse model generalizability and anticipate overfitting due to data sparsity in high-dimensional space (Subramanian & Simon, 2013). Evaluating the joint distribution (X, Y) relative to the product of the two marginal distributions in the event of independence, MI measures relatedness between two random variables (X and (Y) based on the following formula:

$$I(X,Y) = \sum_{y_{\in} X} \sum_{y \in Y} p(x,y) log \left(\frac{p_{X,Y}(x,y)}{p_X(x) p_Y(y)} \right) \tag{2}$$

We used *sklearn*'s MI-based feature selection implementations to accommodates differences in selection routines between continuous and discrete-values features (Ross, 2014).

3 Results

Remark on replicability: Pre-processing and modelling was spread across two separate *Jupyter Notebooks* for performance reasons. To replicate reported analyses, both notebooks need to placed in the same location as data folder FFChallenge_v5 and run in sequence for notebook "CODE_Explore" to write out the pre-processed data and associated dytpe dictionary. Both files will be picked up by notebook "CODE_Predict" from the outset. In addition, file "FFMetadata_v07.csv" needs to be placed in *FFChallenge_v5*.

After splitting the dataset into 67% train and 33% test, all data was pre-processed in line with the aforementioned workflow. The application of different feature pruning thresholds and alternative imputation strategies yielded 6 distinct modelling baselines per prediction task (Appendix 3). A total of 36 datasets were created. On each pre-processed training set, a *LightGBM* model was fitted. Using *Random Search*, we sampled 15 hyper-parameter settings per regression task from previously defined distributions. Substantially longer run times of the *LightGBM* classifier on binary outcomes constrained parameter sampling to 10 search iterations. Parameter settings were evaluated over 5-fold cross-validation using R^2 and $F1$ score as scoring functions for continuous and discrete responses respectively.

In response to RQ_1, Figure 7 summaries modelling results across all prediction tasks. Subplots depict the mean cross-validation scores achieved by each model across parameter search iterations ranked in ascending order per prediction task. The model yielding the highest cross-validation performance was assessed against the associated test set portion held out for final evaluation and is highlighted in green. Winning parameters and the pre-processing convention underlying the data baseline of the winning model are reported to the right of each plot. Test set results are color-coded in red.

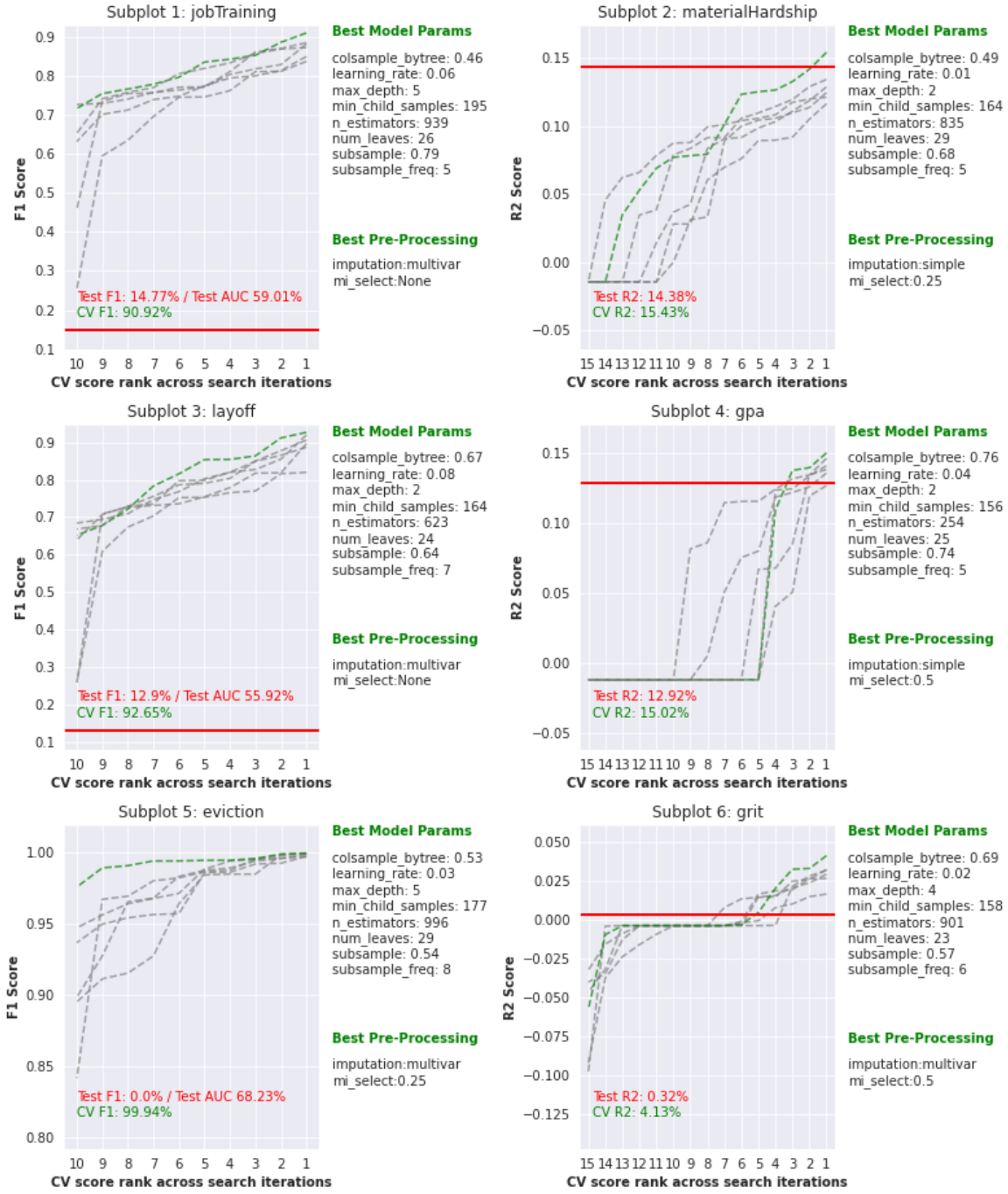


Figure 7: Prediction outcomes across discrete (l.) and continuous (r.) outcomes

3.1 Discrete Outcomes

Predictive performance is generally weak across response variables. The best $F1$ score on the test set was achieved for *jobTraining* (14.77%, Subplot 1), followed by *layoff* (12.9%, Subplot 3). Retrieving confusion matrix results to reason off-the-grid $F1$ results of 0 for *eviction* (Subplot 5) confirms the classifier's inability to predict

the minority class in the heavily skewed test set (7% positive class occurences). Despite upsamling, the model yielded false negatives without exception (Appendix 4)

Pronounced deteriorations between cross-validation ($F1$ scores $> 90\%$) and test set performances ($F1$ scores $< 15\%$) hint at considerable overfitting notwithstanding the inclusion of several regularizing parameters. Yet, AUC scores between 55.9% (*layoff*, Subplot 3) and 68.23% (*eviction*, Subplot 5) indicate that models possess some capacity to discriminate between class labels provided the optimal decision threshold is set.

Best cross-validation results were achieved on input datasets subjected to multivariate imputation across all classification tasks. *LightGBM* Classifiers tended to perform best when exposed to the entire feature bandwidth. MI-based feature pruning only proved "effective" for predicting *eviction*. Pre-processing strategies yielding the best performing model for a given classification task proved quasi-superior across all search iterations.

3.2 Continuous Outcomes

Predictive performance is generally weak across response variables. The best R^2 on the test set was achieved for *material hardship* (14.38%, Subplot 2), followed by *gpa* (12.92%, Subplot 4). Small differences between cross-validation and test scores alleviate concerns about overfitting. Predictive performance stays behind for *grit* (Subplot 6), both in magnitude and absolute discrepancy in R^2 achieved during cross-validation versus testing. Dropping from 4.13% to 0.32%, the model is able to explain less than 1% of the variation in unseen data.

Best cross-validation results were achieved on input datasets subjected to MI-based feature pruning apriori. While *LightGBM* performs best with as few as 25% preselected features for *material hardship*, 50% of features suffice to achieve peak performances for *gpa* and *grit*. We further note a tendency towards simple rather than multivariate imputation.

Yet, none of the datasets yielding the best performing model for either regression task is found to consequently outperform alternatively pre-processed baselines across all 15 search iterations. Since results do not indicate the unequivocal superiority of a particular pre-processing strategy, previous conclusions are tentative. At the same time, segments of plateauing performances become evident for *gpa* and *grit* and to a minor extent for *material hardship*. The extent of recorded hyperparemeter insensitivities seem to vary across and within prediction experiments.

The best performing model for the prediction of *gpa* was insensitive to as many as 11 out 15 randomly sampled hyperparameter combinations (Subplot 4).

Figure 8 exemplifies accomplished random search coverage by plotting sampled values for 2 out of 8 tuned hyperparameters for the winning model per prediction task.

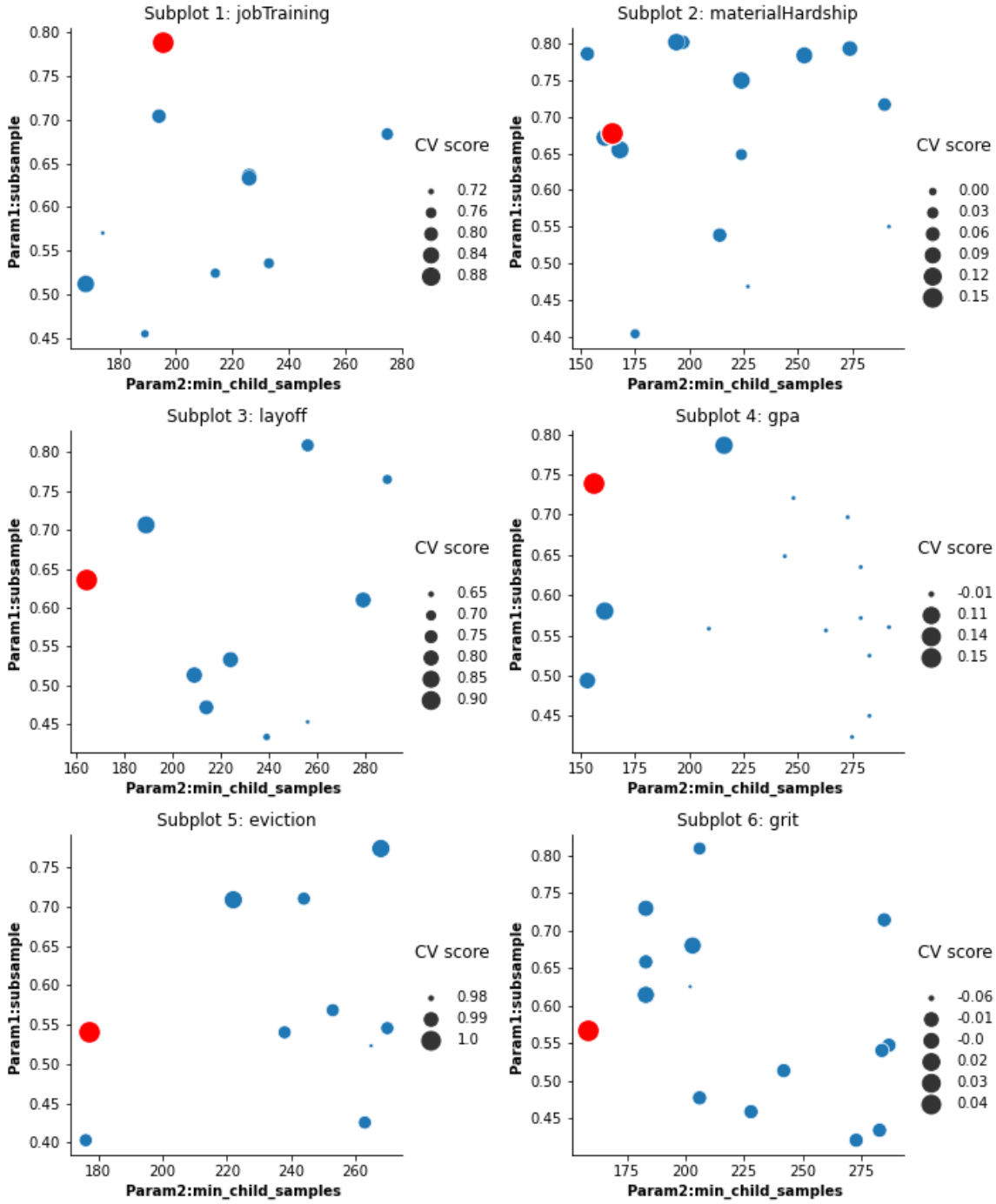


Figure 8: Sampling coverage for combinations of *subsample* and *min_child_smaples*

Bubble sizes are proportional to average cross-validation-performances. Combinations underlying the best model per outcome are color-coded in red. While results feature elevated row subsampling fractions ($> 50\%$) paired with rather weakly constrained leaf creation (< 180 minimum samples a leaf) as winning theme, it becomes evident that the small number of search iterations dedicated to tuning across prediction tasks scarcely covered the hyper-parameter response surface. Apart from sampling results for *gpa* which seem to demarcate an area of inadvisable configurations to the right (Subplot 4), no discernable patterns emerge for the response variable remainder. Reverting to Figure 7, best models further appear to take advantage of more boosting iterations ("n_estimators") while appreciating larger feature fractions ("colsample_by_tree") within defined distributional bounds (Appendix 1).

3.3 Feature Importances

In response to RQ_2, feature importances were retrieved from the best performing *LightGBM* model per prediction task. Importance values in *LightGBM* mirror a model's capacity to achieve more homogeneous partitions by splitting on a given feature. The score is akin to measuring *Mean Decrease in Impurity* (MDI) but without normalizing for the number of samples contained in a split (Lee, 2017; Perrier, 2015). The top 5 features ranked by number of times each is used as splitting criterion are reported in Figure 9.

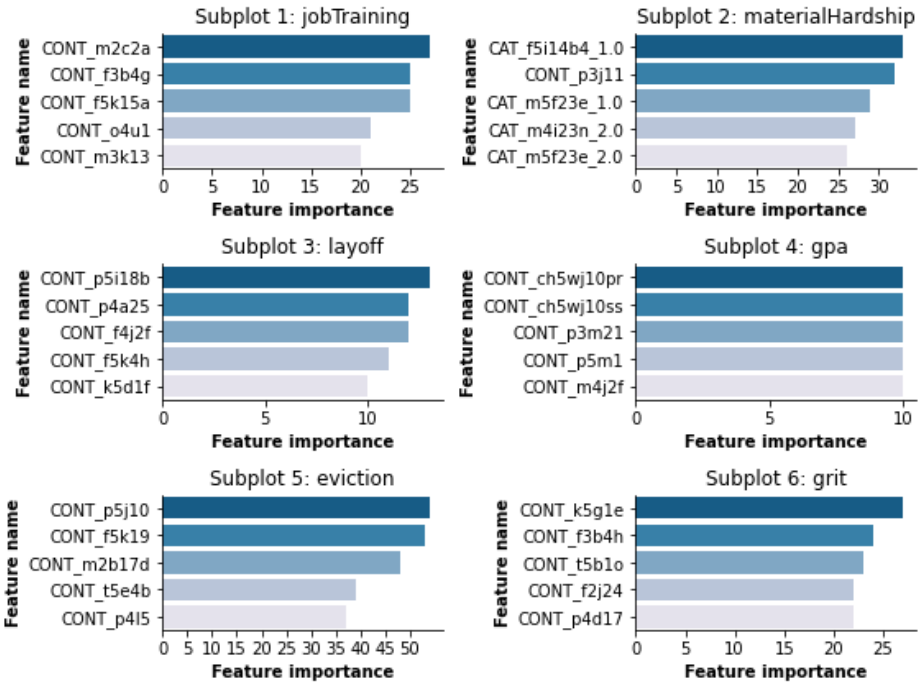


Figure 9: Top 5 most important features across response variables

While content-related interpretations of former remain reserved to the final chapter of this paper, we prepone the discussion of two salient structural findings.

With the exception of *material hardship* (Subplot 2), best models predicting the remaining life outcomes exclusively feature continuous variables in the upper ranks. Findings reflect a well-acknowledged weakness among tree-based learners in dealing with categorical features which introduce harmful amounts of sparsity into the dataset. Problems are particularly pronounced for binary variables which only supply one possible splitting point to the partitioning algorithm, thus promoting unilateral growth in the direction of 0's. Since one-hot encoded features derived from a categorical variable are perceived as independent by the splitting routine, associated purity gains are often marginal in size which puts continuous variables at a structural advantage over sparse binary features when determining the best possible split criterion (Ravi, 2019).

Second, while 4 out of 6 best models were trained on MI feature-pruned baselines, there is only minuscule overlap between the 5 most important features qualified ex-post modelling and the 5 features considered most informative based on *Mutual Information* scores (Appendix 5). With the exception of features "CONT_ch5wj10pr" and "CONT_ch5wj10ss" for *gpa*, no further congruencies emerge. Findings resonate with Rigobon et al.'s (2019) remark that feature pruning outcomes are not necessarily indicative of feature importance or predictive power on unseen data.

4 Discussion

We open the discussion with an assessment of best model feature importances reported in Figure 9. The analysis confines itself to the top 3 features for each response. To ease readers into subsequent elaborations, we matched referenced feature codes with their survey item descriptions for supplementary reference in Appendix 6. Above all, results considerably vary in interpretability.

Comprehensible associations emerge for child *gpa* and *grit*. *Woodcock Johnson Test* cognitive scores and percentile ranks ("CONT_ch5wj10ss" and "CONT_ch5wj10pr") emerge as important determinants of school performance whereas a child's self-appraisal whether he/she follows things through to the end ("CONT_k5g1e") coincides with strength of character.

More subtle connections emerge for outcomes *material hardship* and *job training*. While the need to work weekends ("CAT_f5i14b4_1.0") and the ability to pay utility bills ("CAT_m5f23e_1.0") may be considered pointers of family fortune, care-

givers' persistence demonstrated towards their children through gestures like regularly telling stories ("CONT_f3b4g"), paying visits ("CONT_m2c2a") and dropping them at school ("CONT_f5k15a") manifest a general attitudinal characteristic which may frame caregivers' willingness to work. Lastly, several spurious relationships emerge such as children's computer exposure ("CONT_p5i18b") on the event *layoff* or the regularity with which a cargiver engages in phone calls with a child on household *eviction*.

Eventually,we must treat unearthed importance scores and provided interpretations with caution for several reasons. Mullainathan and Spiess (2017) urge to resist the temptation of drawing hasty conclusions about the data-generating process from fitted functions and frown on libraries that conveniently put the good-faith retrieval of importance scores at scientists' fingertips. The researchers show that variables qualified as important considerably vary when fitting to different partitions of larger samples in the presence of highly correlated, substitutive predictors. Under such circumstances, the odds of fitting comparatively predictive functions on different features rises. By evaluating each candidate feature in isolation for its utility to reduce impurity, *boosting* models are able to discard irrelevant features but remain short-sighted of multicollinearities. At the same time, our feature pruning scheme based on univariate *Mutual Information* suffers from the same limitation. We encourage future research to slot in statistical techniques such as *Variance Inflation Factor* analysis to seize and alleviate just-mentioned concerns in favor of improved interpretability.

Second, one-hot encodings might have obscured the true order of feature importances as dummy-encoded representations of categorical variables may not be selected as splitting criterion even if the "parent" categorical variable is predictive of the outcome for reasons outlined towards the end of the results chapter. Future research should trial *LightGBM*'s categorical feature encoding scheme to improve on recorded results. Grounded on Fisher's (1958) maximum homogeneity algorithm, the former is reported – leaving aside sporadic, skeptical appraisals (Lisovyi, 2018) – to outperform one-hot encoding at many occasions (Light GBM, 2021a).

Most importantly, low levels of predictive power achieved across outcome variables curtails our ability to draw reliable conclusions about observed results. Findings are not only reminiscent of literature advocating the incapacity of models to encapsulate the full meaning of social settings in sober arithmetic (Hofman et al., 2017; Selbst, Boyd, Friedler, Venkatasubramanian, & Vertesi, 2019) but also fuel a selective review of implementational shortcomings in conclusion of this paper.

While non-negligible amounts of missing values for all response variables delimited our baseline to learn from at the outset, endeavors to alleviate serious class imbalances for discrete outcomes yielded strikingly poor test set performances. Since oversampling results in the duplication of minority class observations, balancing out heavy skews for outcomes such as *eviction* where the positive class occurs at a 1:17 ratio might have caused classifiers to overfit (Brownlee, 2021). Since downsampling comes with its very own pitfalls, future studies might want to experiment with best-of-both-world approaches which improve biases to either class by combining moderate rates of up-and downsampling. Similar limitations arise from missing value treatment applied to the feature set. The prevalence of missing values coerced us into omitting more than 5,000 potentially meaningful features while calling for imputation on the feature remainder. Although multivariate imputation proved superior for 4 out of 6 outcomes, devised imputation strategies most likely undermine the complexity of true non-response patterns. Available compute further constrained the number of imputations rounds and the number of neighboring features to fit the regressor.

Last but not least, we refrain from definitive conclusions about best possible parameter settings and unanimously superior pre-processing strategies in light of diagnosed parameter sampling scarcity. Figure 8 provides an important testimony that 10-15 search iterations per model barely scratch the surface considering the wealth of possible parameter combinations in the 8-dimensional parameter space, not to speak of the laundry list of additional parameters available for *LightGBM* but not tuned in the context of this study. While *Random Search* dominates *Grid Search* under such circumstances, *Bayesian Optimization* might prove superior to both by striking a conscious balance between exploration and exploitation. The latter provides a fruitful avenue to navigate complex hyperparameter surfaces under resource constraints whereas *Random Search* remains a guessing game when carrying out small scale searches in high-dimensional space.

We conclude in advocacy of further scientific collaboration. The many shortcomings noted for the FFC data should not leave us with a sense of defeat but fuel the search for creative remedies in cooperation with researchers versed in addressing similar questions with different data. Enriching survey data baselines with community-level features revolving around neighborhood characteristics and administrative records whose analysis proved effective across numerous studies administered by Harvard scholars Chetty et al. (2014, 2016, 2020) constitutes one out many possibilities to further assist the scholarly pursuit of improving the odds for children in fragile families with the help of machine learning. At the same time, results urge us to

acknowledge the limitations of algorithms, be mindful about where and to which extent reliance on ML-assisted insights benefits those under scrutiny and preserve our appreciation for human decision makers who are able to read complex dynamics at play through experience and empathy.

5 Appendices

5.1 Appendix 1: LightGBM Hyperparameters in scope

Hyperparameter Name	Hyperparameter Description	Included by Rigobon et al. (2019)	Value Range considered for Random Search CV	Expected Benefit / Impact
colsample_bytree	randomly subset fraction of features to grow a given tree	Yes	continuous uniform [0.2 0.8]	speed-up / regularization
subsample	fraction of rows subsampled without resampling	Yes	continuous uniform [0.2 0.8]	speed-up / regularization
subsample_freq	bagging frequency (refers to row subsampling)	No	discrete uniform [1 10]	speed-up / regularization
num_leaves	maximum number of leaves in a tree	No	discrete uniform [20 31]	regularization
min_child_samples	minimum number of data in one leaf	No	discrete uniform [150 300]	regularization
n_estimators	number of boosting iterations	Yes	discrete uniform [0.2 0.8]	training accuracy / generalizability
learning_rate	reactivity to error / loss	Yes	continuous uniform [0.01 0.05]	training accuracy
max_depth	maximum distance between root and leaf node in a tree	Yes	discrete uniform [2 6]	regularization/ speed-up

Figure: Overview of hyperparameters in scope for *LightGBM* implementations

5.2 Appendix 2:Outcome variable transforms in comparison

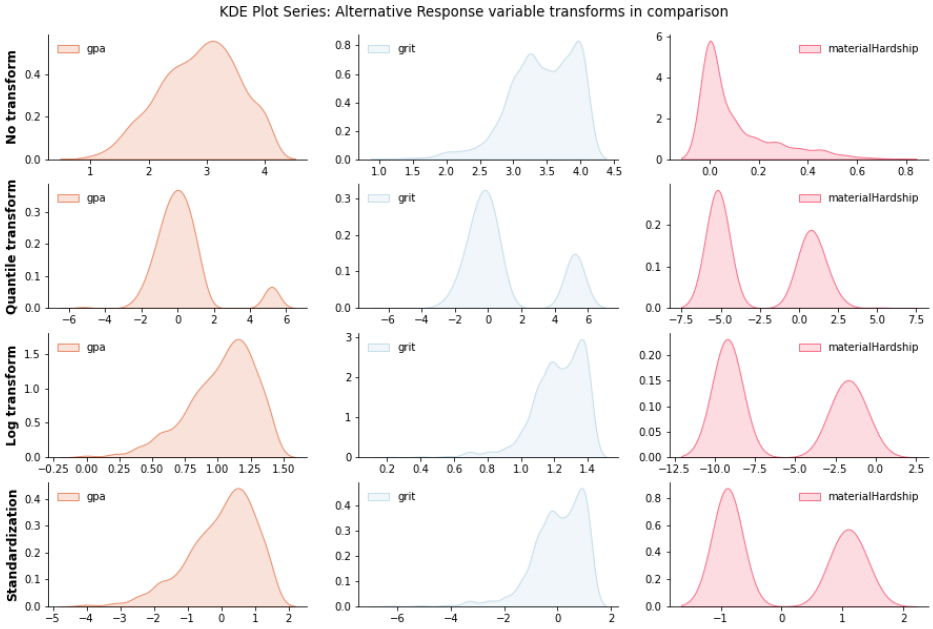


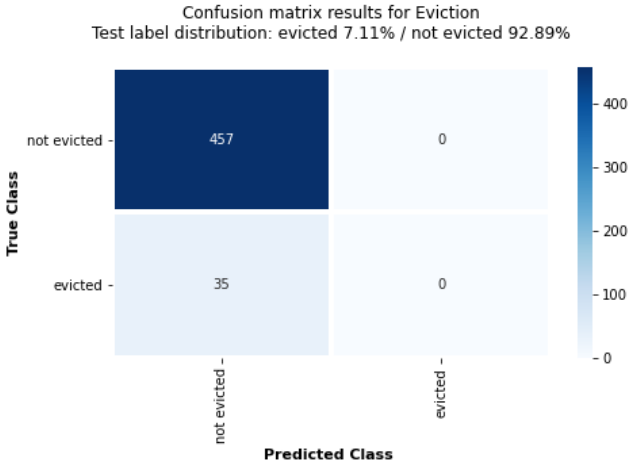
Figure: Response variable distributions (KDE plots) - *Top to bottom:* non-transformed, quantile transform, log transform and standardization ($\mu = 0; \sigma = 1$)

5.3 Appendix 3: Overview of modelling baselines

Config	transform	imputation	oh-encoding	MI feature selection
1	True	simple	True	None
2	True	simple	True	0.25
3	True	simple	True	0.5
4	True	multivar	True	None
5	True	multivar	True	0.25
6	True	multivar	True	0.5

Table: Overview of modelling baselines resulting from different pre-processing choices

5.4 Appendix 4: Confusion matrix results for eviction



5.5 Appendix 5: Top 5 most informative features across best models

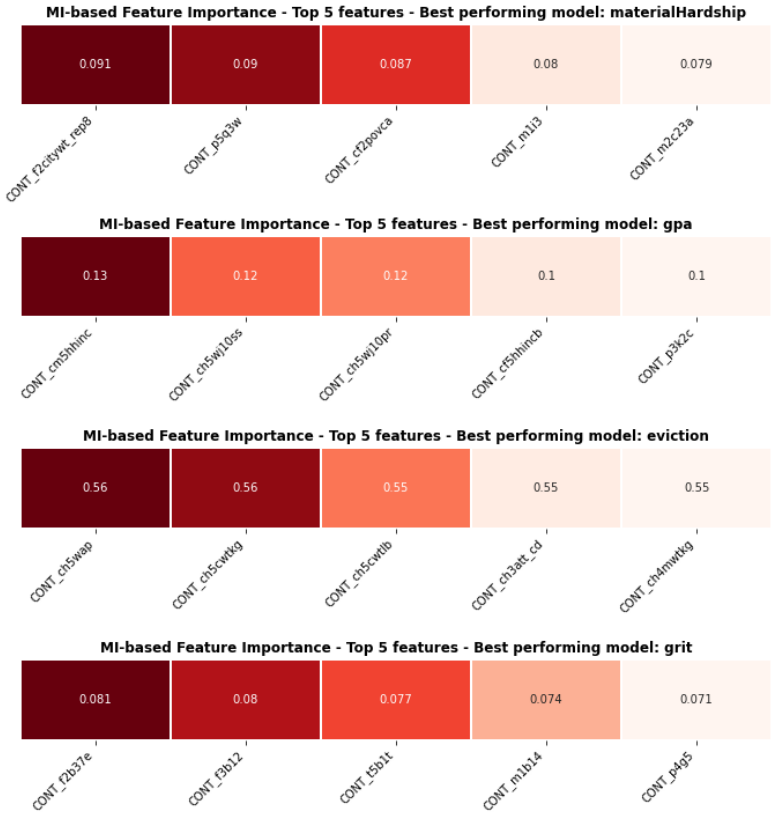


Figure: Overview of top 5 most informative features across best performing models grounded on *Mutual Information*-based feature pruning

5.6 Appendix 6: Survey items associated with top 3 most important features across responses

response_var	feature	feature_text
jobTraining	CONT_m2c2a	During the past 30 days, how many days has father seen child?
jobTraining	CONT_f3b4g	Days/week: tell stories to child?
jobTraining	CONT_f5k15a	Average number of times/month you have taken child to or from school/progr
materialHardship	CAT_f5i14b4_1.0	You sometimes also work weekends
materialHardship	CONT_p3j11	Past year, times PCG slapped child on the hand, arm, or leg
materialHardship	CAT_m5f23e_1.0	Did not pay full amount of gas/oil/electricity bill in past 12 months
layoff	CONT_p5i18b	Number of hours per day child uses computer
layoff	CONT_p4a25	How frequently does child ride in a car, van, or other vehicle?
layoff	CONT_f4j2f	How much does mother weigh?
gpa	CONT_ch5wj10pr	Woodcock Johnson Test 10 percentile rank
gpa	CONT_ch5wj10ss	Woodcock Johnson Test 10 standard score
gpa	CONT_p3m21	Child hits others
eviction	CONT_p5j10	Amount of money spent eating out in last month
eviction	CONT_f5k19	Frequency you talk on telephone with child
eviction	CONT_m2b17d	On a scale of 1-(least like) to 5-(most like) - Child gets upset easily
grit	CONT_k5g1e	I follow things through to the end
grit	CONT_f3b4h	Days/week: play inside with toys with child?
grit	CONT_t5b1o	Child gives compliments to peers

Table: Original Survey item descriptions for most important features per outcome

References

ACPHA. (2020). *Covid-19: Protecting children from violence, abuse and neglect in the home.* Retrieved from `https://www.unicef.org/media/68711/file/COVID-19-Protecting-children-from-violence-abuse-and-neglect`

Adams, C. (2020). *Is a secondary pandemic on its way?* Retrieved from `https://ihv.org.uk/news-and-views/voices/is-a-secondary-pandemic-on-its-way/`

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

Brown, C. L., Yilanli, M., & Rabbitt, A. L. (2020). Child physical abuse and neglect. *StatPearls [Internet]*.

Brownlee, J. (2021). *Random oversampling and undersampling for imbalanced classification.* Retrieved from `https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/`

Bzostek, S., Beck, A., et al. (2008). *Family structure and child health outcomes in fragile families* (Tech. Rep.).

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? evidence from project star. *The Quarterly journal of economics*, *126*(4), 1593–1660.

Chetty, R., Hendren, N., Jones, M. R., & Porter, S. R. (2020). Race and economic opportunity in the united states: An intergenerational perspective. *The Quarterly Journal of Economics*, *135*(2), 711–783.

Chetty, R., Hendren, N., & Katz, L. F. (2016). The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. *American Economic Review*, *106*(4), 855–902.

Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? the geography of intergenerational mobility in the united states. *The Quarterly Journal of Economics*, *129*(4), 1553–1623.

Eubanks, V. (2018). *A child abuse prediction model fails poor families:why pittsburgh's predictive analytics misdiagnoses child maltreatment and prescribes the wrong solution.* Retrieved from `https://www.wired.com/story/excerpt-from-automating-inequality/`

Fisher, W. D. (1958). On grouping for maximum homogeneity. *Journal of the American statistical Association*, *53*(284), 789–798.

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.

Green, P. (2020). *Risks to children and young people during covid-19 pandemic.* British Medical Journal Publishing Group.

Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, *355*(6324), 486–488.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Jin, R., & Agrawal, G. (2003). Communication and memory efficient parallel decision tree construction. In *Proceedings of the 2003 siam international conference on data mining* (pp. 119–129).

Kader, G. D., & Perry, M. (2007). Variability for categorical variables. *Journal of Statistics Education*, *15*(2).

Kenward, M. G., & Carpenter, J. (2007). Multiple imputation: current perspectives. *Statistical methods in medical research*, *16*(3), 199–218.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, *105*(5), 491–95.

Lee, C. (2017). *Feature importance measures for tree models — part 1.* Retrieved from `https://medium.com/the-artificial-impostor/feature-importance-measures-for-tree-models-part-i-47f187c1a2c3`

Light GBM, . (2021a). *Light gbm: Advanced topics: Categorial feature support.* Retrieved from `https://lightgbm.readthedocs.io/en/latest/Advanced-Topics.html?highlight=feature%20importance#categorical-feature-support`

Light GBM, . (2021b). *Light gbm: Features.* Retrieved from `https://lightgbm.readthedocs.io/en/latest/Features.html`

Light GBM, . (2021c). *Light gbm: Parameters tuning.* Retrieved from `https://lightgbm.readthedocs.io/en/latest/Parameters-Tuning.html`

Lisovyi, M. (2018). *Beware of categorical features in lgbm!* Retrieved from `https://www.kaggle.com/mlisovyi/beware-of-categorical-features-in-lgbm`

Liu, S. H., & Heiland, F. (2012). Should we get married? the effect of parents' marriage on out-of-wedlock children. *Economic Inquiry*, *50*(1), 17–38.

Lundberg, I. (2018). *Missing data in the fragile families study.* Retrieved from `https://www.fragilefamieschallenge.org/missing-data/`

Makhijani, P. (2017). *Fragile families challenge uses 'big data' to answer big questions.* Retrieved from `https://www.princeton.edu/news/2017/11/13/fragile-families-challenge-uses-big-data-answer-big-questions`

Martin, D. (2019). *Are you getting burned by one-hot encoding?* Retrieved from `https://kiwidamien.github.io/are-you-getting-burned-by-one-hot-encoding.html`

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

OPR. (2021). *The fragile families and child wellbeing study and fragile families challenge.* Retrieved from https://opr.princeton.edu/archive/FF/

Perrier, A. (2015). *Feature importance in random forests.* Retrieved from https://alexisperrier.com/datascience/2015/08/27/feature-importance-random-forests-gini-accuracy.html

Princeton University, . (2021a). *About the fragile families and child wellbeing study.* Retrieved from https://fragilefamilies.princeton.edu/about

Princeton University, . (2021b). *Welcome to the ffcws metadata explorer!* Retrieved from http://metadata.fragilefamilies.princeton.edu

Raschka, S. (2015). *Python machine learning.* Packt publishing ltd.

Ravi, R. (2019). *One-hot encoding is making your tree-based ensembles worse, here's why?* Retrieved from https://towardsdatascience.com/one-hot-encoding-is-making-your-tree-based-ensembles-worse-heres-why-d64b282b5769

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.

Ross, B. C. (2014). Mutual information between discrete and continuous data sets. *PloS one*, *9*(2), e87357.

Salganik, M. J., Lundberg, I., Kindel, A. T., & McLanahan, S. (2019). Introduction to the special collection on the fragile families challenge. *Socius*, *5*, 2378023119871580.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 59–68).

Shi, H. (2007). *Best-first decision tree learning* (Unpublished doctoral dissertation). The University of Waikato.

Sigle-Rushton, W., & McLanahan, S. (2002). For richer or poorer? marriage as an anti-poverty strategy in the united states. *Population*, *57*(3), 509–526.

Stanescu, D., Wang, E., & Yamauchi, S. (2019). Using lasso to assist imputation and predict child well-being. *Socius*, *5*, 2378023118814623.

Subramanian, J., & Simon, R. (2013). Overfitting in prediction models–is it a problem only in high dimensions? *Contemporary clinical trials*, *36*(2), 636–641.

Waldfogel, J., Craigie, T.-A., & Brooks-Gunn, J. (2010). Fragile families and child

wellbeing. *The Future of children/Center for the Future of Children, the David and Lucile Packard Foundation*, *20*(2), 87.