Black Box Models and Sociological Explanations:
Predicting High School GPA Using Neural Networks

Thomas Davidson
Cornell University

ABSTRACT

The Fragile Families Challenge provided an opportunity to empirically assess the applicability of black box machine learning models to sociological questions and the extent to which these models can produce interpretable explanations. In this paper I use neural network models to predict high school grade-point average and examine how different variations of basic network architectures affect predictive performance. Using a recently proposed technique, I identified the important variables used by the best-performing model, finding that variables associated with parenting and the child's cognitive and behavioral development are highly predictive, consistent with prior work. I conclude with a discussion of how social scientists can combine prediction and explanation and some of the challenges that must be solved.

INTRODUCTION

The Fragile Families Challenge (FFC) was a competition held in 2017 to bring together social scientists and data scientists in a mass-collaboration to a predict six key outcomes in the Fragile Families survey. For each outcome participants were able to use data from waves 1 through 5, which cover the period from each child's birth until age 9, to the outcome at age 15 (Salganik et al., n.d.). The aim of the FFC was to assess how well these outcomes could be predicted and to help uncover new determinants of these outcomes for further assessment using in-depth interviews.

As part of the FFC, I evaluated the performance of neural network models for predicting self-reported high school grade-point averages (GPA). In this paper I detail the process of constructing, testing and evaluating these models. I also build upon recent calls to reconcile predictions and explanations (Hofman, Sharma, and Watts 2017) by

applying a recently proposed method to help explain the predictions of these black box models (Ribeiro, Singh, and Guestrin 2016). This shows how variables related to parenting, cognitive and behavioral development, as well as the school, the family, and the community are particularly important predictors, consistent with prior research.

In the Background section begin by discussing theoretical and methodological developments on the relationship between prediction and explanation in sociology and other disciplines. I then give a general introduction to neural network models, followed by a short review of the literature on the determinants of educational performance and prior work predicting high school GPA. In the Methodology section I give a detailed technical explanation of the process used to clean and transform the FFC data and then train and evaluate the neural network models. The Results section consists of two parts: the first consists of an evaluation of the predictive performance of the different model architectures tested; in the second I introduce and apply a method to interpret the predictions made by the neural network and assess how the results related to prior work on the determinants of educational attainment. I conclude the paper by reflecting up the results and the current challenges that work combining prediction and sociological explanation must address.

BACKGROUND

Prediction and Sociological Explanations

Watts (2014) argues that common sense thinking is pervasive in sociological reasoning, resulting in a bias towards parsimonious accounts of social life that are commensurable

with our everyday experiences but that may not be scientifically valid. To help ensure the validity of our explanations Watts advocates a shift away from evaluating them based on their understandability and towards their capacity to predict, which can be assessed using out-of-sample testing. Although the focus of Watts' article is on using prediction as a way to evaluate our explanations, prediction can also be an end in itself.

The machine learning models and techniques used by computer scientists and other practitioners provide efficient ways to make predictions with large volumes of data and to reliably assess their accuracy and validity.[1] A limitation of these models is that they are often considered "black boxes",[2] meaning that we do not directly observe how they work or even which variables are associated with the outcome. However their increasing ubiquity and involvement in consequential decisions in domains such as healthcare and criminal justice (Athey 2017) is leading many researchers to develop techniques to help us better understand them. Although most machine learning models are not transparent, in the sense that their inner workings are clear to observers, these researchers are showing that they can still be interpretable (Lipton 2016). For example, a number of methods have recently been proposed to allow us to obtain post-hoc interpretations for any class of model, one of which is used later in this paper (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017; Lakkaraju et al. 2017). Reflecting

[1] Cross-validation and out-of-sample validation is the norm in computer science and the nascent field of data science, but has been neglected by sociologists, despite being known in the field at least since the early 1990s (Camstra and Boomsma 1992).

[2] This usage of the term must be distinguished by the concept of a black box put forward by analytical sociologists. Hedström and Swedberg (1998) argue that regression models, even if they have interpretable coefficients that give us an indication of important variables in a model, are black boxes since they do not provide us with insight into the nature of the *mechanisms* linking the variables to the outcome. In many machine learning models it is difficult to even recover meaningful coefficients and the model is solely evaluated based on its predictive accuracy, so all of the models discussed in this paper, including the interpretations, constituted black boxes under this stricter definition.

upon these trends, Hofman, Sharma, and Watts (2017: 488) conclude their recent *Science* article by highlighting that "the trade-off between predictive accuracy and interpretability may be less severe than once thought". This suggests that the tools may be available to allow us to achieve both high predictive accuracy and explanatory power.

However, there are also major challenges to the reconciling of predictions and explanations. The dominant framework in supervised machine learning is to start with a large set of data and feed it into a model that finds a functional form to optimize predictive accuracy. Mullainathan and Spiess (2017) contend that because these models are optimized to estimate $\hat{y}$values with high accuracy, rather than consistently estimate a set of $\hat{\beta}$s, we cannot reliably use them to interpret the latter as is the norm for the explanatory models used by social scientists. Specifically, they argue that when a predictive model uses a larger number of variables, often more than the number of observations, the most important variables are often a random subset of a larger set of variables that are associated with the outcome. "The very appeal of these algorithms is that they can fit many different functions. But this creates an Achilles' heel: more functions mean a greater chance that two functions with very different coefficients can produce similar prediction quality" (Mullainathan and Spiess 2017: 97-8). Prediction might allow us to better evaluate our explanatory models for $\hat{\beta}$s (Watts 2014), but based on their assessment, post-hoc interpretations of predictive models estimating $\hat{y}$cannot be used to directly generate sociological explanations. Despite this limitation, these models might still prove valuable to sociological inquiry, insofar as they help to identify important features for further examination and the upper bounds for predictive accuracy.

The FFC provided an excellent opportunity to explore these issues in an empirical setting. In this paper I explore how machine learning methods—specifically neural networks—perform at predicting high school GPA, and study the extent to which post-hoc interpretations of these models can contribute to existing sociological explanations.

Neural Networks and Deep Learning

Neural networks have recently shown great promise in many domains including image recognition, language translation, and genomics, out-performing the state-of-the-art in an array of different problems (see LeCun, Bengio, and Hinton 2015). A neural network consists of multiple interconnected layers of nodes, known as neurons; "deep" neural networks include one or more "hidden" layers between the input and output nodes. Figure 1 shows an example where there are two hidden layers, each containing three neurons and a bias unit, analogous to the intercept in a linear regression. The models learn as data are input into the network and passed through these layers. Each neuron calculates a weight that is then used to transform the input data to satisfy an objective function (e.g. to minimize mean squared error). The weighted sum of the inputs to a layer is then passed through an "activation" function, transforming the data, typically by inducing non-linearity, which enables it to be better represented in the output layer, and then passed into the next layer. The data are passed through the network multiple times—each full pass of the dataset is known as an "epoch"—allowing the network to update and improve its predictions. Finally, either after a specified number of epochs or the model fit ceases to improve, the network outputs the predicted values for each observation.

The models are estimated via a process called back-propagation: as the data are passed through each layer, the gradient is calculated to determine whether an increase or decrease in the weight for each feature will increase or decrease the error, and the weights are updated accordingly (Rumelhart et al. 1988; see Hinton 1992 for an overview). This is typically done using stochastic gradient descent, where the weights are adjusted after a small batch of observations have passed through the network, rather than after each epoch, providing a stochastic approximation of the true gradient, which makes the networks more efficient to train.

Work on image-recognition networks has shown how, as the model is trained, individual neurons learn to specialize in certain subtasks, for example one might detect edges of objects while another detects textures (Krizhevsky, Sutskever, and Hinton 2012; LeCun, Bengio, and Hinton 2015).[3] The remarkable success of these models lies in their ability find approximations for a wide range of complex functions (Cybenko 1989; Hornik, Stinchcombe, and White 1989; Lin and Tegmark 2016). Given these successes in other domains, I expect that these models will prove valuable for social scientific research, particularly for predicting outcomes like GPA, where existing approaches are only able to explain a fraction of the variance. Since neural networks are non-parametric they do not make the same assumptions about normality and linearity as linear models, and they tend to be robust towards coding errors, missing data, and noise (Garson 1998: 8-9). They can also automatically fit linear, polynomial, and interactive relationships between the inputs, allowing them to model complex relationships between the inputs and the outcome (Garson 1998: 8). This makes neural networks well suited to the FFC

[3] See LeCun, Bengio, and Hinton (2015) for an overview of neural network models and some major scientific breakthroughs that have been achieved using them.

tasks as they can deal with large volumes of survey data with relatively little cleaning and pre-processing.

Predicting Grade-Point Average

The dependent variable in this analysis is the average self-reported high school GPA at or around age 15 across four subjects, ranging from 4.0, which indicates A grades in all subjects, to 1.0, denoting an average of D or below. The average GPA in the FFC dataset is 2.87, approximately a B letter grade. The FFC includes this variable based on the organizers' observation that social scientists' models of educational attainment typically have low explanatory power,[4] indicating a need for better predictions.

The following sociological studies of educational attainment offer important insights into the correlates of academic performance, although they do not explicitly predict GPA. Demographic variables including race, gender, and ethnicity are associated with differences in outcomes, as are household characteristics including parental education and socioeconomic status (Entwisle, Alexander, and Olson 2005; Rumbaut 2005). Prior work has found that family structure is associated a child's academic performance (Cavanagh, Schiller, and Riegle-Crumb 2006), with children from single parent families (Astone and McLanahan 1991) and with incarcerated fathers (Foster and Hagan 2009) obtaining lower GPA scores than their peers. The nature of parental involvement in children's education can also impact GPA, with both the number of hours spent on homework and parental expectations showing positive associations (Rumbaut 2005). More generally, parenting style can also impact educational outcomes (Roksa and Potter 2011). While high school GPA is often used as a predictor variable or analyzed

[4] See the FFC blog post on why GPA was selected: http://www.fragilefamilieschallenge.org/gpa/.

descriptively in these studies, there is surprisingly little work directly predicting it as an outcome. I now consider all papers I was able to find that explicitly use GPA as a dependent variable.

Baker and Stevenson (1986) find that mothers with higher educational attainment tend to act more strategically to help manage their children's education, resulting in higher GPAs. Parental involvement in the home and in interactions with schools is associated with higher GPAs (Wang and Sheikh-Khalil 2014). Students' interactions with their peers can also directly impact GPA, as well as mediating the effects of parenting (Mounts and Steinberg 1995). Attewell (2001) finds that the school environment can also predict GPA, as schools often devote most resources to a small number of high achieving students, to the detriment of others. Branigan (2017) observes that obesity is associated with a lower GPA, but that the effect only exists for white girls, purportedly due to negative teacher assessments. Research by psychologist has also identified a number of different personality traits including self-efficacy, social-desirability, self-discipline, and motivation strategies can predict GPA (Wolters 1999; Caraway et al. 2003; Duckworth and Seligman 2005).

While all of these studies directly predict GPA, they all use different datasets, GPA is often measured differently, a variety of modelling techniques are used, and model fit statistics, when reported, are all based on in-sample calculations. Moreover, all of these studies focus on testing particular $\hat{\beta}$s rather than predicting $\hat{y}$s so their goal is not to maximize predictive accuracy. Due to these factors, this prior work does not offer a clear baseline against which my predictions can be compared. However this work does provide

an array of potential explanations, which I can assess by comparing them to the predictors identified by my model.

METHODOLOGY

Data Cleaning and Transformations

Missing data were identified using the missing codes provided in the survey documentation; all codes were treated in the same manner due to the time constraints of the FFC. To impute the missing values it was first necessary to identify the type of every column. In the absence of machine-readable information on the column types[5], I used a simple heuristic to ascertain their types, treating any column with less than 50 unique values as categorical and the rest as continuous. I then dropped any columns that were completely empty, had zero variance, or had greater than 70% of observations missing. For the remaining columns I used the fancyimpute (Rubinsteyn and Feldman 2016) implementation of k-Nearest Neighbors to impute missing values, where the values for the five closest neighbors for each observation were used to impute the mean and mode for missing continuous and categorical variables respectively. The k-NN procedure was chosen because it is fast to implement and easily applies to multiple different variable types. I then standardized every continuous variable, by subtracting its mean and dividing it by its standard deviation, which has been shown to speed up the model learning rate (LeCun et al. 2012), and transformed every categorical variable to a set of binary vectors,

[5] This made it impractical to automatically clean and pre-process the entire dataset. This is not usually a problem in traditional sociological analyses where one typically only uses a small fraction of the available variables, which can be cleaned and recoded manually within a reasonable amount of time. In this case, because I use a large proportion of the dataset, such data processing must be performed automatically. As a consequence of the FFC, a new metadata API has been constructed to help facilitate this type of analysis, which I use below to help analyze the results.

or "dummy variables", each denoting the presence or absence of a particular category, a format known as "one-hot encoding".[6] Each column in the final transformed matrix can be considered to be a "feature" that can be used by the model.[7] The changes to the data after each stage in the cleaning and transformation process are shown in a chart in the Supplemental Information (SI). Much of the data manipulation was performed using pandas (McKinney 2013).

Constructing Neural Network Models

To construct the neural network, I used Keras (Chollet 2015), a high-level Python library that acts as an interface with the TensorFlow architecture (Abadi et al. 2016). I restricted the analysis to feed-forward neural networks, where each layer is connected only to the layers adjacent to it and data only flows through the graph in one direction, from input to output. I experimented with a number of different structures to try to identify an appropriate specification for the task and demonstrate how different parameters affect predictive performance.[8]

First, to understand the effects of the network depth, I varied the number of layers, comparing a network with a single neuron and no hidden layers (where the input layer is directly connected to the output layer), a network with a single hidden layer, and "deep" networks with two and three layers. Second, to assess the effects of network breadth I varied the number of neurons in each layer. For each network, I tested hidden

[6] As I was unable to ascertain precise data type for every column, both categorical and ordinal variables were treated in this manner.

[7] I use the term feature to refer to columns in this matrix and variables to refer to variables in the survey

[8] There are many different types of neural networks and different parameters that can be varied with them. Here I have only selected core aspects of the model to introduce the approach to social scientists. Iterating over a larger parameter space would require much more computational power and time than was available, although is also worth noting that the optimal solution to this task may lie outside of the parameter space explored.

layers with 64, 128, and 256 neurons. Third, I tested four different activation functions: the sigmoid, the hyperbolic tangent (tanh), and the rectified linear unit (ReLU), as well as a basic linear activation. The equations for these functions and graphical representations are shown in the SI. The sigmoid and tanh functions are commonly used to induce non-linearity by squashing values into the ranges [0,1] and [-1, 1] respectively. The ReLU activation is a threshold function that sets negative values to zero. It has been used in many applications because it provides good results and is fast to train (Nair and Hinton 2010; LeCun, Bengio, and Hinton 2015). The model with a linear activation function learns the linear combination of the inputs that minimizes the mean squared error and should theoretically produce estimates equivalent to an OLS regression (Kuan and White 1994).

Cross-Validation and Model Estimation

I first split the training data into two parts, keeping 80% as a training set and holding out the remaining 20% as an out-of-sample validation set. I used the GridSearchCV function in scikit-learn (Pedregosa et al. 2011) to iterate through the different parameter combinations and to implement 5-fold cross-validation using the training set. Both the out-of-sample validation and the 5-fold cross-validation should help to mitigate the risk of overfitting. When combined with cross-validation the different combinations of parameters resulted in 40 different model specifications and 200 different model fits. The loss function used for each model was negative mean squared error; the objective of training is to maximize this function (which equates to minimizing mean squared error). Each neural network was trained using mini-batch gradient descent via the Adam

algorithm (Kingma and Ba 2014) and ran for 200 epochs, or until 25 epochs passed without any loss function reduction in a subset of the data used for validation (this threshold is set using the "patience" parameter in the model). To further mitigate the risk of over-fitting the training data I used dropout, where the outputs of a random set of neurons from each layer are set to zero (or "dropped") at each stage in the training process (Srivastava et al. 2014). I fixed the amount of dropout to 50% for all models. The models ran consecutively on the CPU of a 2016 MacBook Pro with a 3.3 GHz Intel Core i7 processor and 16 GB of RAM, and took just over 12 hours to run.

The performance of these models is compared against the baseline prediction, which is simply setting all predicted values as the mean, and an OLS regression baseline, as OLS is the most frequently used method in the papers directly predicting GPA that were reviewed above. The regression was implemented using the LinearRegression function in scikit-learn and 5-fold cross-validation was also used.[9]

RESULTS

[Table 1 about here]

Comparing Network Architectures

The table showing the results for all 40 parameter combinations can be found the SI, along with graphs showing the relationship between model architecture and performance. The scores reported are the average mean squared error on the training data across all five folds used in cross-validation. All models without any hidden layers performed rather poorly, all far worse than the baseline OLS model. As the number of layers increases, the

[9] See the SI for a discussion of the problems associated with estimating linear regressions using high-dimensional data.

predictive performance tends to improve, although the best performing model on the training data only has a single layer of hidden neurons. Across all models, there does not appear to be a systematic relationship between the number of layers or neurons and the overall performance of the model. This is supported by regression analyses of the results, presented in the SI, although the sample size is small after removing outlier predictions. The greatest variation in performance is observed due to the activation function used. Networks with hyperbolic tangent and sigmoid activations tend to consistently achieve low MSE scores, with the sigmoid appearing to be the best overall activation function for this task. The ReLU function performs well in some cases but has far higher variability, performing significantly worse on average than the other activation functions, an issue discussed in greater depth in the SI. Looking at the linear activation function, performance is highly also variable. In some cases, the models appear to perform equivalently to the other non-linear functions, while in others they perform poorly. Overall there is quite a high amount of variation in performance across all specifications; 1/4 of the models perform worse than the mean baseline and only 13 of the 40 outperform the OLS baseline.

[Table 1 around here]

To better understand these results, I took the top five best performing sets of parameters on the training data and retrained models,[10] this time reducing the patience parameter to help prevent over-fitting by allowing the training to stop sooner when performance no longer improved on the validation set. These models were then used to predict the GPA for all individuals in the dataset. The scores of these models on the

[10] By default only the best performing model is stored after the grid search so the models needed to be re-computed.

initial training data, the validation data, the leader board data, and the final held-out data are shown in Table 1. With the exception of the model using the ReLU activation, which stopped training before it converged on an appropriate solution, all of the models outperform the OLS baseline on all of the datasets, likely because the OLS overfits the training data, although the mean baseline beats all but the best model on the leaderboard data. When evaluated against the final held-out datasets, the model with 3 layers, 256 neurons in each layer, and a sigmoid activation function performs best, with an MSE of 0.3797 on the leader board data and 0.3866 on the final held-out data (the best model in the competition got an MSE in the final held-out data of 0.3438).

While this model outperforms both baselines and is not too far off the competition winner, the increase in performance compared to the baselines is still relatively small: on the final hold-out data the MSE is only 15% lower than the OLS regression and 10% lower than the mean baseline. Neural networks have dramatically outperformed baselines in other domains such as image recognition and language translation, but based on these results it is unclear if these successes can translate over to predictions using survey data. One possible limitation is the size of the data: after holding out 20% of the available training data for validation only 1696 observations are available to train the model. It may be the case that neural networks require far more data to train effectively. Another limitation is that only a relatively small parameter space could be explored; the optimal model may well lie outside of this space. I now examine the best performing model more closely in an effort to interpret its predictions.

Interpreting Black Box Models

While there do exist methods to identify important features from feed-forward neural networks they are non-trivial to apply and are not available as open-source packages, making them unsuitable for this analysis given time constraints of this challenge (e.g. Goh 1995; Olden and Jackson 2002; Tzeng and Ma 2005). Moreover, unlike the approach taken below, these methods are often not applicable to more complex neural network architectures than those discussed here, so may be less useful for other researchers. To better understand how the model is classifying observations I use the Local Interpretable Model-Agnostic Explanations (LIME) algorithm, which fits a simpler model to attempt to explain the predictions for a subset of the observations obtained from a more complex black box model (Ribeiro, Singh, and Guestrin 2016). LIME works by taking as inputs an individual observation (or row of data) and the predicted value for the observation from the black box model, and constructing a number of permutations of the row by randomly deleting non-zero values, and then fitting a linear model on these permuted data, where each permutation is weighted by its distance from the original observation. This results in a linear approximation of the local decision function of the original model, which can identify important predictive features. The algorithm returns an "explanation," consisting of a user-specified number of features that predict the outcome for that specific observation.

As LIME is computationally expensive, I randomly selected 100 observations and ran the LIME algorithm on each one, with the top 5 most important features returned in each case.[11] This provides approximation of the most important predictive features in the neural network for a subset of the individuals in the dataset, which can give some insight

[11] I experimented with different values but found that values higher than 5 often ran into problems where the algorithm failed to converge and therefore did not produce a local estimated model.

into the model, which is otherwise a black box. As discussed above, these features correspond to particular columns in the training data, for example whether the value is 1 or 0 for a given category of a variable. For the analysis below I match all features to the corresponding variables and focus on the latter (see the SI for more information). Of the one hundred observations selected, there are 419 unique variables that appear as the top predictors identified by LIME;[12] less than 15% of variables occur in more than one individual explanation and none occurs more than 5 times.[13] To examine these variables in more detail, I use the new Fragile Families metadata API to automatically obtain information on the wave, respondent, and overall topic of each variable. A full list of variables and associated metadata can be found in the SI.[14]

Beginning with the topics that the questions relate to,[15] Figure 2[16] shows the proportion of the variables in the explanations that belong to each category. Overall, we see that questions about parenting and the child's cognitive and behavioural development occur most frequently, followed by variables related to home and housing, finances, health and health behaviour, and education and school. Over 80% of the variables in the explanations belong to these six categories. Simply analysing these frequencies, however, may result in spurious conclusions because the distribution of topics across all survey questions is uneven. Some topics may appear more frequently simply by chance. To

[12] The maximum possible is 500, if each explanation consisted of 5 unique variables.

[13] Ribeiro et al. (2016) also propose a method to find the subset of features that best explain a range of different observations, providing a global approximation, but given that most of the features in this solution are unique, it will be unlikely to find a better solution than simply visually inspecting the results and identifying features that recur with high frequency.

[14] I was unable to obtain metadata for 6 of the variables that occurred in these explanations: 'hv3c_c3', 'hv3r10a7', 'hv3cwtalone', 'hv4cflag', 'hv4agemos', 'hv4food_exp'.

[15] Here I focus on the broader "umbrella" topics rather than the more granular topics. Analysis of the latter shows consistent results and is reported in the SI. Note that each variable can belong to up to two different umbrella topics; both are counted in this analysis.

[16] All figures hereafter were produced using Matplotlib (Hunter 2007).

control for the topic prevalence in the survey, I conduct z-tests to assess whether the proportion of variables relating to each topic is different in the LIME explanations and the entire survey (waves 1 through 5). I also show created two graphs ranking the topics by the difference in proportions (Figure 3) and the ratio (Figure 4). Stars in front of the topic names in the figures indicate statistically significant differences in proportions.[17]

[Figures 2, 3, and 4 about here]

Variables related to parenting constitute the highest proportion of those identified using LIME and occur significantly more frequently in these explanations than in the survey itself, with a ratio of over 2.5:1. These include variables that capture the parents' involvement in their child's life, mostly consisting of parental interactions with the with the child, such as reading bedtime stories, whether they play outside together, how the parents discuss problems with the child and if they physically punish them. This category also includes parental assessments of the child's behaviour, such as how often the child watches TV, how many books the child has, and how many mornings a week the child eats breakfast. My findings are consistent with Baker and Stevenson's (1986: 165) emphasis on the importance of parental strategies in promoting academic success: "Parents must do a long series of small things to assist their child toward maximum educational attainment".

The other topic that appears most frequently in the LIME explanations, with the highest ratio of occurrences compared to the survey, is cognitive and behavioral development. Most of the variables in the explanation relate to the child's behavior, such

[17] * = $p < 0.05$, ** = $p < 0.01$, and *** = $p < 0.001$.

as whether they argue a lot, have temper tantrums, tease other children, and have difficulty concentrating in class. Prior work using these data has found that these types of behaviors are associated with poor cognitive development in middle childhood (Turney and McLanahan 2015) and these results indicate that these variables can also predict high school performance.

Variables pertaining to education and the school remain important in the adjusted ranking, accounting for just under 10% of the variables in the explanations. These include both the child's responses about their experiences in school, as well as parental and teacher assessments. This is consistent with prior work that has identified the school context as an important predictor of educational attainment (Alexander and Eckland 1975). Variables related to the family, social support, and the community also occur significantly more frequently in the explanations than in the survey, suggesting the importance of considering social structures and contexts beyond the parent-child relationship and the school.

Looking at the least frequent variables, it is notable that demographic variables only occur 9 times; while these variables are often the mainstay of sociological analyses it appears that fine-grained behavioural data is more useful for prediction. Variables associated with health and healthcare and the legal system—predominantly relating to the father's incarceration—do appear in some explanations, but occur less frequently than expected. This supports prior work that has identified these factors to be important predictors of children's educational outcomes (e.g Haskins 2014; Haskins and Jacobsen 2017; Branigan 2017), but they appear to only be predictive in a small number of cases. Variables relating to employment and childcare appear to be far less frequent than in the

survey, along with paradata and weights, which is the most frequent category in the survey and largely consists of constructed variables, the latter suggesting that the model may be making more use of the raw variables than any pre-engineered features.

Looking at the distribution of LIME identified variables across survey waves, I find that those obtained in later waves appear far more frequently in the explanations, but this effect disappears when taking the proportion in the entire dataset into account. Variables from wave 1 occur significantly more frequently in the explanations than in the survey, with a ratio of 1.5:1, consistent with prior work that shows that early childhood indicators can predict future educational attainment (Alexander, Entwisle, and Horsey 1997; Entwisle, Alexander, and Olson 2005). Variables from wave 3, which contributed the largest number of variables to the survey, occur significantly less frequently, although it is unclear why.

Turning to the respondent associated with each variable, I find that the vast majority of variables derived from survey segments corresponding to the mother, father, and primary caregiver. The proportions for mother and father are not statistically different from the overall survey but that LIME explanations contain more variables associated with the primary caregiver and the child's teacher, the ratios for both respondent types are above 2:1. The primary caregiver, typically the child's mother—unless they live with their father or another relative—provides information about parenting, child health, and development that appears to be particularly useful for prediction.[18] The finding that variables from the teacher interviews occur more frequently than expected is consistent with studies showing how the teacher-pupil relationship and

[18] See the "Introduction to the Fragile Families Public Use Data" for more information about the surveys and the role of the primary caregiver: https://fragilefamilies.princeton.edu/sites/fragilefamilies/files/ff_public_guide_0to5.pdf.

the teacher's assessment of the pupil can help explain differences in educational attainment (Alexander, Entwisle, and Thompson 1987; Branigan 2017).

Although LIME is limited by its computational demands and only provides relatively simple post-hoc explanations, revealing little about the inner workings of the neural network, it can help shed light on the factors associated with the predictions made by the more complex model (Lipton 2016). These results illustrate how complex predictive models can be useful as a data-driven way to identify new factors to consider in our theoretical accounts. The findings suggest that variables related to parenting, the child's behaviour and cognitive development, community, family and social support, and education are particularly important for predicting future GPA. Assessment of the waves of the survey suggest that even information collected very early on can be valuable and that the primary caregiver and teacher are particularly important sources of information. The variables identified are broadly consistent with prior work, suggesting that many of these sociological explanations may have predictive validity.

It is also important to add a caveat, however, that while LIME allows us to gain some insight into the predictions made by black box models, we must still be aware of the danger of interpreting these explanations as if they are causal. It is plausible, and indeed highly likely, that many other possible variables could have similar predictive power (Mullainathan and Spiess 2017) and that we could tell similar stories about them (Watts 2014). These post-hoc explanations are not substitutes for models that aim to carefully estimate the $\hat{\beta}$'s and that can be validated using out-of-sample testing. Nonetheless these results show how predictive models can be interpreted to help us to identify important predictive factors that may help to improve our sociological

explanations. Future work should assess how the predictors identified above, most importantly fine-grained behavioral data, can be integrated into our sociological explanations.

CONCLUSION

The FFC presents a valuable opportunity to assess how black box machine learning methods perform in prediction tasks compared to the models sociologists are more familiar with. This paper finds that while neural networks can perform reasonably well at regression tasks that, in this case, they do not seem to perform particularly better than other types of models. I hope that others build upon this foundation by assessing the extent to which different network architectures can be reliably used for the types of prediction tasks of interest to social scientists. Predictive performance aside, the variables identified by LIME not only provide a window into the model, but are consistent with prior literature. This suggests both that existing sociological explanations of high school GPA, and educational attainment more generally, have some predictive power, and that the model was able to inductively learn that factors identified these literatures were important. Notwithstanding the concerns raised by Mullainathan and Spiess (2017), the results suggest that machine learning can provide a valuable complement to more traditional approaches, potentially allowing us to predict outcomes with relatively high accuracy and to inductively identify important variables by making use of large volumes of data. Examination of how these findings can contribute to sociological explanations requires further research that can explore causal mechanisms by more robustly by

estimating $\hat{\beta}$'s. While my results are not the best of the entrants to the FFC I hope that readers are encouraged to apply these methods more widely and to further explore how they can be used in sociological research. Black box modelling approaches like neural networks deserve more attention from social scientists, particularly as they are becoming more amenable to interpretation.[19]

There are two other important implications of this study. First, despite the attempts to prevent over-fitting by using cross-validation and early-stopping, there are still significant differences in model performance across the training, validation, and held-out test datasets. Not only does this illustrate the difficulty of constructing models that can generalize out-of-sample, but it also casts doubt on the validity of any results that are only reported in-sample, reinforcing Watts' (2014) emphasis on the importance of out-of-sample testing. Second, as machine learning becomes more frequently adopted by social scientists it is important to recognize the nuances of the distinction between explanatory models and predictive models, and when either approach should be used (Mullainathan and Spiess 2017). More work is necessary to identify the best ways of integrating the insights gleaned from predictive modelling with the more traditional causal analyses. Overall, I hope that this paper highlights both the importance and the challenges of reconciling predictions and explanations to advance social scientific inquiry. Mass-collaboration efforts like the FFC provide a valuable opportunity to push us towards this goal by simultaneously advancing both our ability to predict social outcomes and to explain the social world.

[19] See Garson (1998) for an excellent and prescient discussion of the potential for these models in the social sciences.

REFERENCES

Alexander, Karl L., and Bruce K. Eckland. 1975. "Contextual Effects in the High School Attainment Process." *American Sociological Review* 40 (3): 402. https://doi.org/10.2307/2094466.

Alexander, Karl L., Doris R. Entwisle, and Carrie S. Horsey. 1997. "From First Grade Forward: Early Foundations of High School Dropout." *Sociology of Education* 70 (2): 87. https://doi.org/10.2307/2673158.

Alexander, Karl L., Doris R. Entwisle, and Maxine S. Thompson. 1987. "School Performance, Status Relations, and the Structure of Sentiment: Bringing the Teacher Back In." *American Sociological Review* 52 (5): 665. https://doi.org/10.2307/2095602.

Astone, Nan Marie, and Sara S. McLanahan. 1991. "Family Structure, Parental Practices and High School Completion." *American Sociological Review* 56 (3): 309. https://doi.org/10.2307/2096106.

Athey, Susan. 2017. "Beyond Prediction: Using Big Data for Policy Problems." *Science* 355 (6324): 483–485.

Attewell, Paul. 2001. "The Winner-Take-All High School: Organizational Adaptations to Educational Stratification." *Sociology of Education* 74 (4): 267. https://doi.org/10.2307/2673136.

Baker, David P., and David L. Stevenson. 1986. "Mothers' Strategies for Children's School Achievement: Managing the Transition to High School." *Sociology of Education* 59 (3): 156. https://doi.org/10.2307/2112340.

Branigan, Amelia R. 2017. "(How) Does Obesity Harm Academic Performance? Stratification at the Intersection of Race, Sex, and Body Size in Elementary and High School." *Sociology of Education* 90 (1): 25–46. https://doi.org/10.1177/0038040716680271.

Camstra, Astrea, and Anne Boomsma. 1992. "Cross-Validation in Regression and Covariance Structure Analysis: An Overview." *Sociological Methods & Research* 21 (1): 89–115.

Caraway, Kirsten, Carolyn M. Tucker, Wendy M. Reinke, and Charles Hall. 2003. "Self-Efficacy, Goal Orientation, and Fear of Failure as Predictors of School Engagement in High School Students." *Psychology in the Schools* 40 (4): 417–27. https://doi.org/10.1002/pits.10092.

Cavanagh, Shannon E., Kathryn S. Schiller, and Catherine Riegle-Crumb. 2006. "Marital Transitions, Parenting, and Schooling: Exploring the Link Between Family-Structure History and Adolescents' Academic Status." *Sociology of Education* 79 (4): 329–54. https://doi.org/10.1177/003804070607900403.

Duckworth, Angela L., and Martin EP Seligman. 2005. "Self-Discipline Outdoes IQ in Predicting Academic Performance of Adolescents." *Psychological Science* 16 (12): 939–944.

Entwisle, Doris R., Karl L. Alexander, and Linda Steffel Olson. 2005. "First Grade and Educational Attainment by Age 22: A New Story." *American Journal of Sociology* 110 (5): 1458–1502. https://doi.org/10.1086/428444.

Foster, Holly, and John Hagan. 2009. "The Mass Incarceration of Parents in America: Issues of Race/ Ethnicity, Collateral Damage to Children, and Prisoner

Reentry." *The ANNALS of the American Academy of Political and Social Science* 623 (1): 179–94. https://doi.org/10.1177/0002716208331123.
Goh, A. T. C. 1995. "Back-Propagation Neural Networks for Modeling Complex Systems." *Artificial Intelligence in Engineering* 9 (3): 143–151.
Haskins, Anna. 2014. "Unintended Consequences: Effects of Paternal Incarceration on Child School Readiness and Later Special Education Placement." *Sociological Science* 1: 141–58. https://doi.org/10.15195/v1.a11.
Haskins, Anna R., and Wade C. Jacobsen. 2017. "Schools as Surveilling Institutions? Paternal Incarceration, System Avoidance, and Parental Involvement in Schooling." *American Sociological Review*, 0003122417709294.
Hedström, Peter, and Richard Swedberg. 1998. "Social Mechanisms: An Introductory Essay." In *Social Mechanisms: An Analytical Approach to Social Theory*, edited by Peter Hedström and Richard Swedberg, 1–31. New York, NY: Cambridge University Press. http://professor-murmann.info/images/uploads/Social-mechanism.pdf.
Hofman, Jake M., Amit Sharma, and Duncan J. Watts. 2017. "Prediction and Explanation in Social Systems." *Science* 355 (6324): 486–488.
Hunter, John D. 2007. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering* 9 (3): 90–95.
Lakkaraju, Himabindu, Ece Kamar, Rich Caruana, and Jure Leskovec. 2017. "Interpretable & Explorable Approximations of Black Box Models." *ArXiv:1707.01154 [Cs]*, July. http://arxiv.org/abs/1707.01154.
Lipton, Zachary C. 2016. "The Mythos of Model Interpretability." In *Workshop on Human Interpretability in Machine Learning*. New York, NY. https://arxiv.org/pdf/1606.03490.pdf.
Lundberg, Scott M., and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems*, 4768–4777.
McKinney, Wes. 2013. *Python for Data Analysis*. Beijing: O'Reilly.
Mounts, Nina S., and Laurence Steinberg. 1995. "An Ecological Analysis of Peer Influence on Adolescent Grade Point Average and Drug Use." *Developmental Psychology* 31 (6): 915.
Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106. https://doi.org/10.1257/jep.31.2.87.
Olden, Julian D., and Donald A. Jackson. 2002. "Illuminating the 'Black Box': A Randomization Approach for Understanding Variable Contributions in Artificial Neural Networks." *Ecological Modelling* 154 (1): 135–150.
Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "Why Should i Trust You?: Explaining the Predictions of Any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. ACM. http://dl.acm.org/citation.cfm?id=2939778.
Roksa, Josipa, and Daniel Potter. 2011. "Parenting and Academic Achievement: Intergenerational Transmission of Educational Advantage." *Sociology of Education* 84 (4): 299–321. https://doi.org/10.1177/0038040711417013.

Rubinsteyn, S, and S Feldman. 2016. "Fancyimpute." https://github.com/iskandr/fancyimpute.
Rumbaut, Rubén G. 2005. "Turning Points in the Transition to Adulthood: Determinants of Educational Attainment, Incarceration, and Early Childbearing among Children of Immigrants." *Ethnic and Racial Studies* 28 (6): 1041–86. https://doi.org/10.1080/01419870500224349.
Salganik, Matthew J., Ian Lundberg, Alex Kindel, and Sara S. McLanahan. n.d. "Introduction to the Special Issue on the Fragile Families Challenge."
Turney, Kristin, and Sara McLanahan. 2015. "The Academic Consequences of Early Childhood Problem Behaviors." *Social Science Research* 54 (November): 131–45. https://doi.org/10.1016/j.ssresearch.2015.06.022.
Tzeng, F.-Y., and K.-L. Ma. 2005. "Opening the Black Box-Data Driven Visualization of Neural Networks." In *Visualization, 2005. VIS 05. IEEE*, 383–390. IEEE. http://ieeexplore.ieee.org/abstract/document/1532820/.
Wang, Ming-Te, and Salam Sheikh-Khalil. 2014. "Does Parental Involvement Matter for Student Achievement and Mental Health in High School?" *Child Development* 85 (2): 610–25. https://doi.org/10.1111/cdev.12153.
Watts, Duncan J. 2014. "Common Sense and Sociological Explanations." *American Journal of Sociology* 120 (2): 313–51. https://doi.org/10.1086/678271.
Wolters, Christopher A. 1999. "The Relation between High School Students' Motivational Regulation and Their Use of Learning Strategies, Effort, and Classroom Performance." *Learning and Individual Differences* 11 (3): 281–299.

Table 1: Performance of Top 5 models from the training data

Hidden layers	Neurons (per layer)	Activation function	MSE training (mean)	MSE validation[1]	MSE leader-board	MSE final held-out
1	256	sigmoid	0.2559	0.3022	0.4174	0.3871
1	128	relu	0.2598	0.4823	0.6439	0.6215
3	128	linear	0.2598	0.3178	0.4253	0.4023
2	256	sigmoid	0.2659	0.3157	0.4344	0.3956
3	256	sigmoid	0.2662	0.3164	0.3797	0.3866
Mean baseline			0.3111	0.3407	0.3927*	0.4251*
OLS baseline			0.3005	0.3708	0.4680*	0.4454*

The training scores are the mean MSE scores across all five folds used in cross-validation. For the other columns, the best model from cross-validation is retrained on the entire training set and is used to predict the outcomes on the initial validation dataset, the leaderboard dataset, and the final held-out dataset. *These results were not submitted as part of the FFC competition.

Figure 1: Network diagram

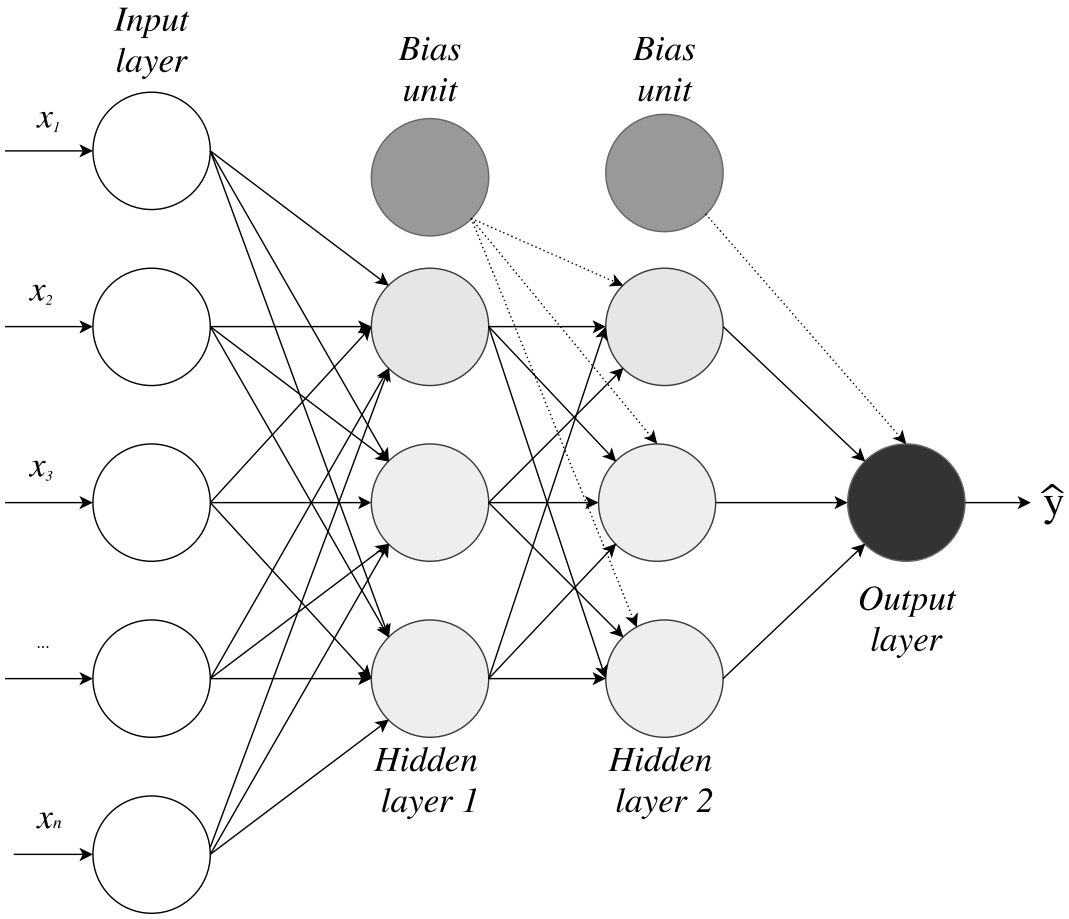


Figure 2

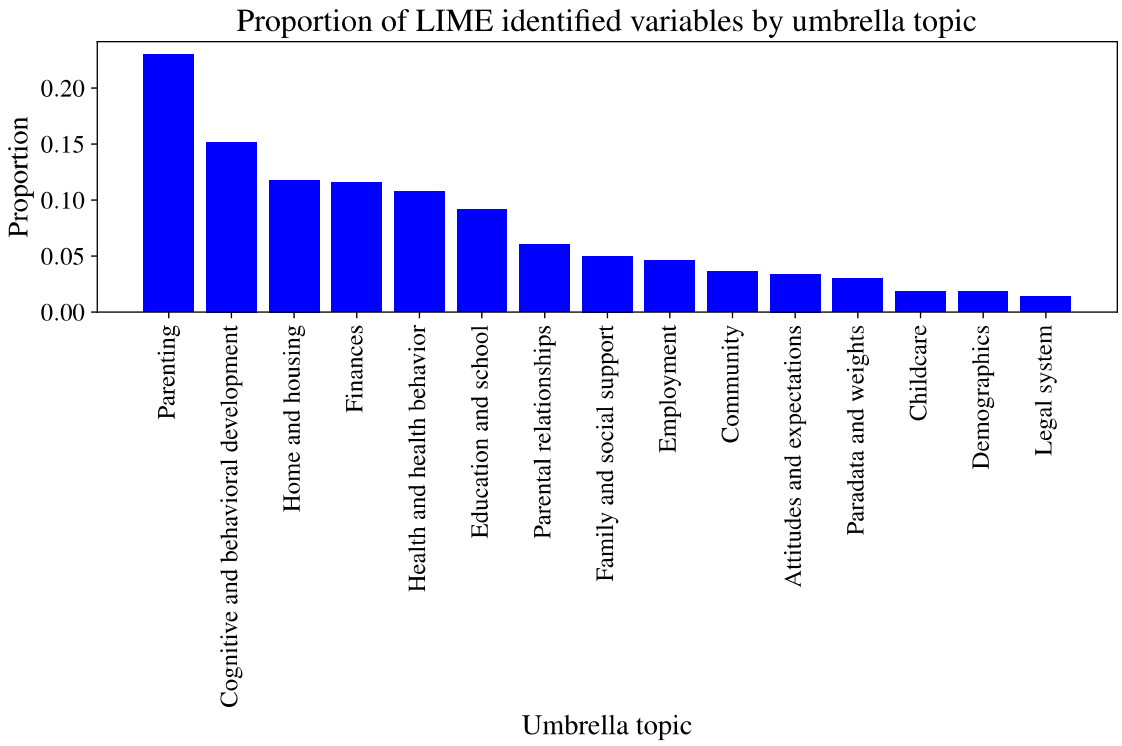

Proportion of LIME identified variables by umbrella topic
Proportion
0.20
0.15
0.10
0.05
0.00
Parenting
Cognitive and behavioral development
Home and housing
Finances
Health and health behavior
Education and school
Parental relationships
Family and social support
Employment
Community
Attitudes and expectations
Paradata and weights
Childcare
Demographics
Legal system
Umbrella topic

Figure 3

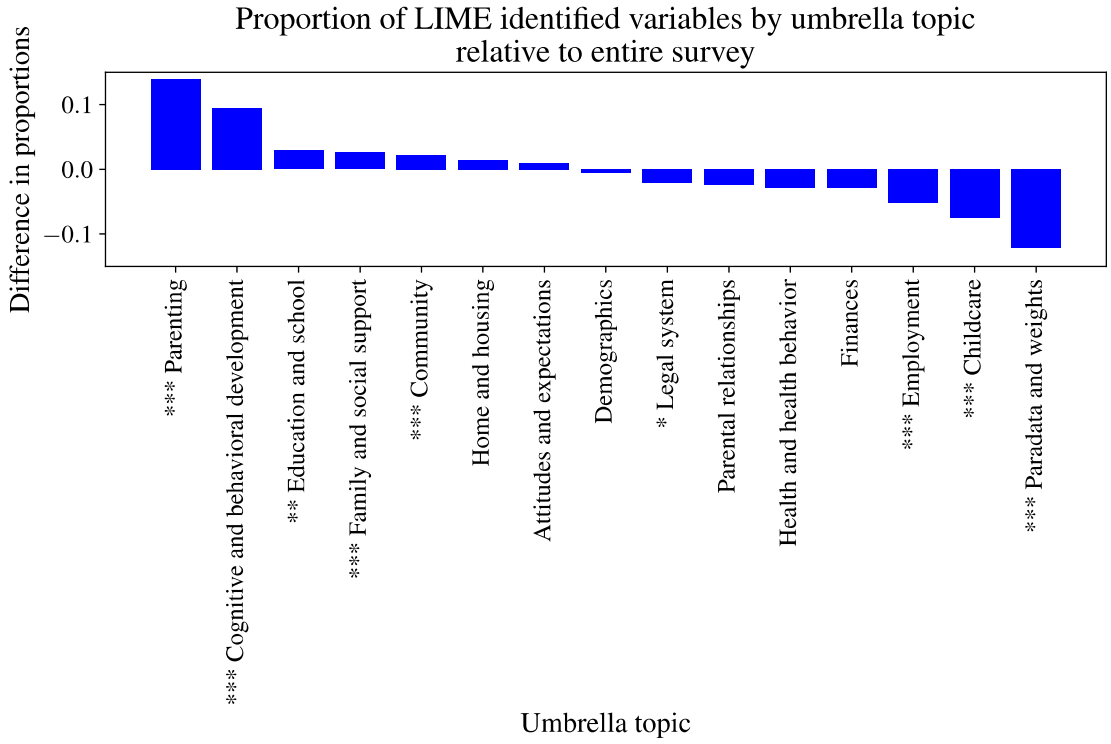
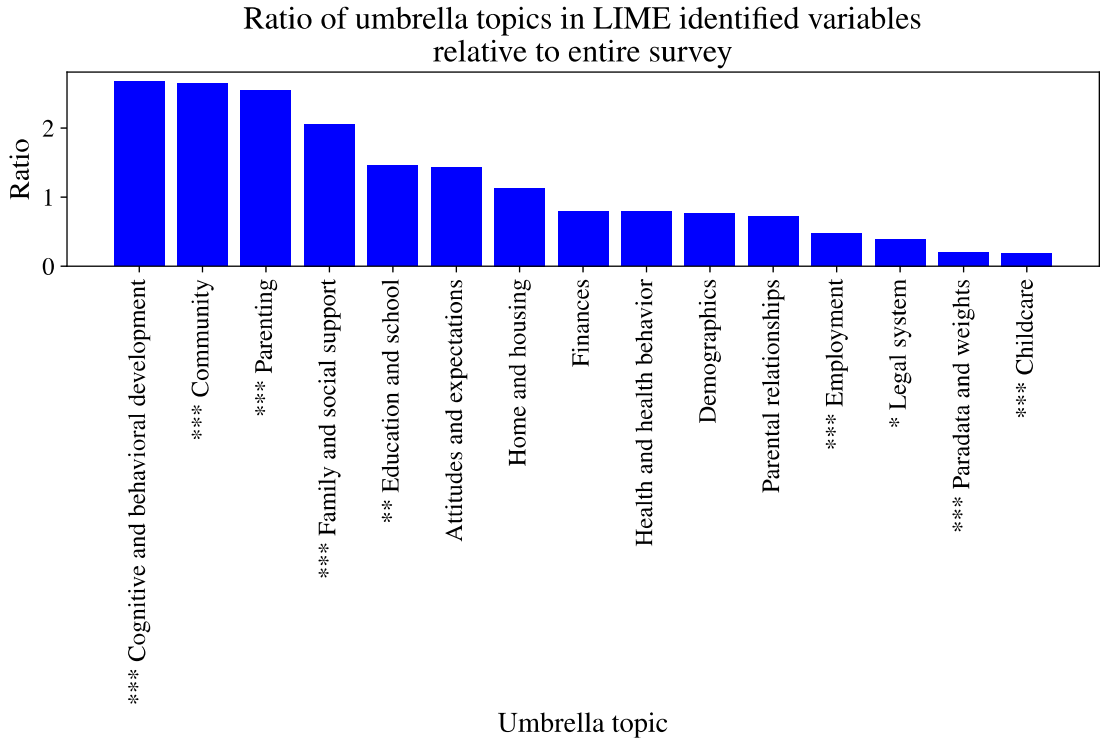

Proportion of LIME identified variables by umbrella topic
relative to entire survey
Difference in proportions
0.1
0.0
−0.1
*** Parenting
*** Cognitive and behavioral development
** Education and school
*** Family and social support
*** Community
Home and housing
Attitudes and expectations
Demographics
* Legal system
Parental relationships
Health and health behavior
Finances
*** Employment
*** Childcare
*** Paradata and weights
Umbrella topic

Figure 4



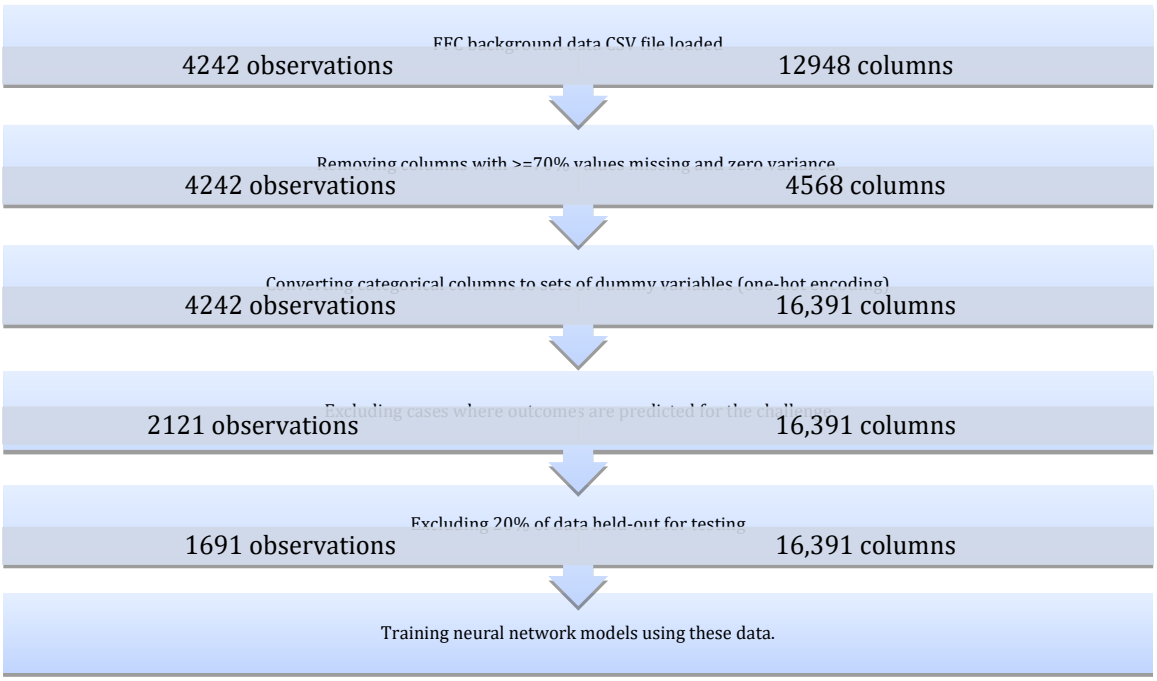
Black Box Models and Sociological Explanations: Predicting GPA Using Neural Networks

SUPPLEMENTAL INFORMATION

A. Data transformations

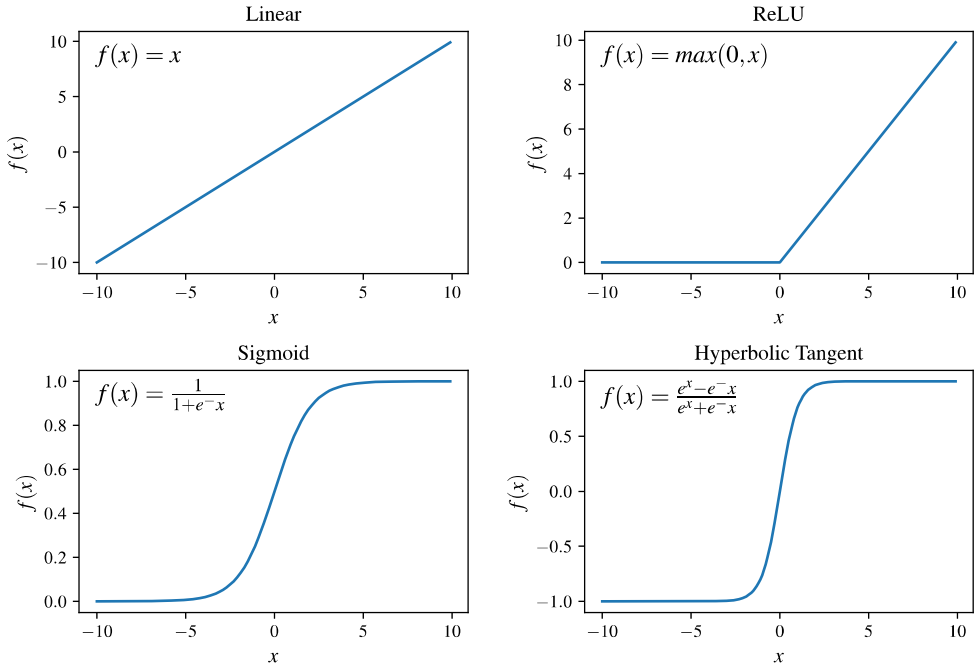
The diagram below shows how the raw data provided for the FFC was transformed before it was modelled on. The initial variable cleaning and imputation is performed using the `clean_data.py` script. The one-hot encoding and the definition of the training data and held-out test set are performed before modelling in `gpa.ipynb`.

Data transformations



B. Activation functions

The figure below shows the four activation functions used in the paper and the transformations they produce for values of *x* ranging from -10 to 10.



C. Regression with high-dimensional data

OLS regression is used as an additional baseline, since multivariate OLS regression is the most frequent method used in past work predicting GPA. My approach creates problems for OLS, however, as there are considerably more variables than observations. The inclusion of such a large number of predictors can result in multicollinearity, leading the standard errors to be inflated, impeding identification. However, since the objective here is to predict $\hat{y}$, the estimated coefficients and standard errors are not important (Hastie, Tibshirani, and Friedman 2009: 46; Mullainathan and Spiess 2017). Of more concern, is the fact that there are substantially more variables than observations, which mean that the linear regression is not well defined because the matrix $X^T X$ is not invertible (recall the

linear regression estimates $\hat{\beta} = X^T X^{-1} X^T y$). As the number of features increase the OLS model can become increasingly sensitive to random errors, resulting in high variance, which can be detrimental to predictive performance.

To try to mitigate these problems I also experimented with LASSO regression, which applies regularization by shrinking some coefficients to zero, effectively selecting only a subset of all of the features, which can improve predictive accuracy (Hastie, Tibshirani, and Friedman 2009: 57-73). However the model always predicted identical values for all observations regardless of the value of the penalty, making it unsuitable for prediction. The same problem occurred when trying to use an ElasticNet, which also incorporates a penalty to reduce the size of coefficients, which should help mitigate the risk of overfitting. Thus only the OLS results are reported.

D. Results of initial analysis

The table below shows the average MSE obtained across all five folds of training data for all 40 model specifications. The scores for the top 5 best performing models are highlighted in bold font.

Hidden layers	Neurons (per layer)	Activation function	MSE training
Mean baseline			0.3927
OLS baseline			0.3005
0	0	linear	2.5847
		relu	7.2273
		sigmoid	3.6258
		tanh	3.6258
1	64	linear	0.2896
		relu	0.4979

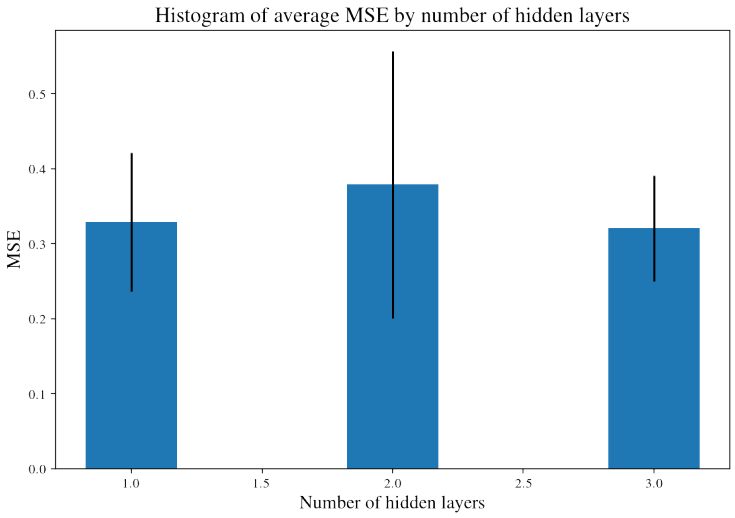
		sigmoid	0.2979
		tanh	0.3104
	128	linear	0.2800
		relu	0.2598
		sigmoid	0.2771
		tanh	0.3092
	256	linear	0.5423
		relu	0.3247
		sigmoid	0.2559
		tanh	0.3042
2	64	linear	0.2756
		relu	0.3509
		sigmoid	0.3058
		tanh	0.3094
	128	linear	0.2694
		relu	0.5654
		sigmoid	0.2862
		tanh	0.3087
	256	linear	1.9546
		relu	0.8518
		sigmoid	0.2659
		tanh	0.3788
3	64	linear	0.3012
		relu	0.3150
		sigmoid	0.2913
		tanh	0.3100
	128	linear	0.2599
		relu	0.3297
		sigmoid	0.2815
		tanh	0.3160
	256	linear	27.5780
		relu	0.5196
		sigmoid	0.2663
		tanh	0.3396

All scores are the average MSE across all five folds used in training. Top 5 best performing models are shown in bold font.

E. Graphical examination of the effect of model architecture on MSE

The graphs below show the relationships between different architectures and model performance. All graphs are shown with 6 outlier results with MSE scores greater than 1 removed, since these skew the results.[20]

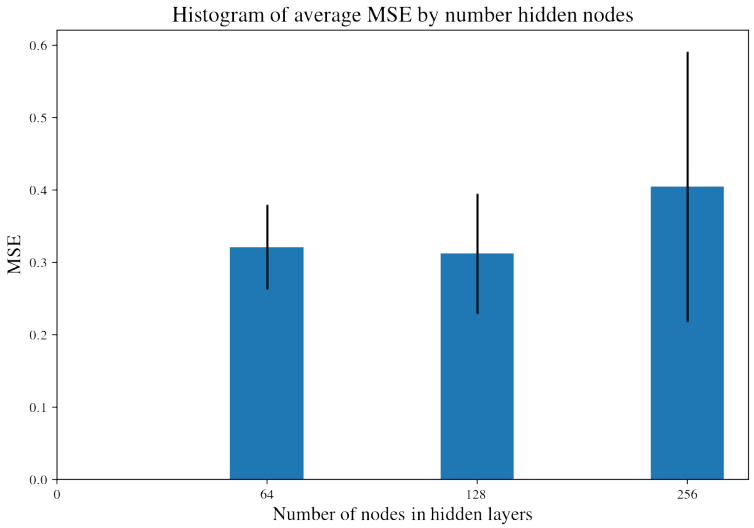
The first plot shows the average MSE by the number of hidden layers. Note that there is no bar for 0 layers since none of these models achieved an MSE below the cut-off.



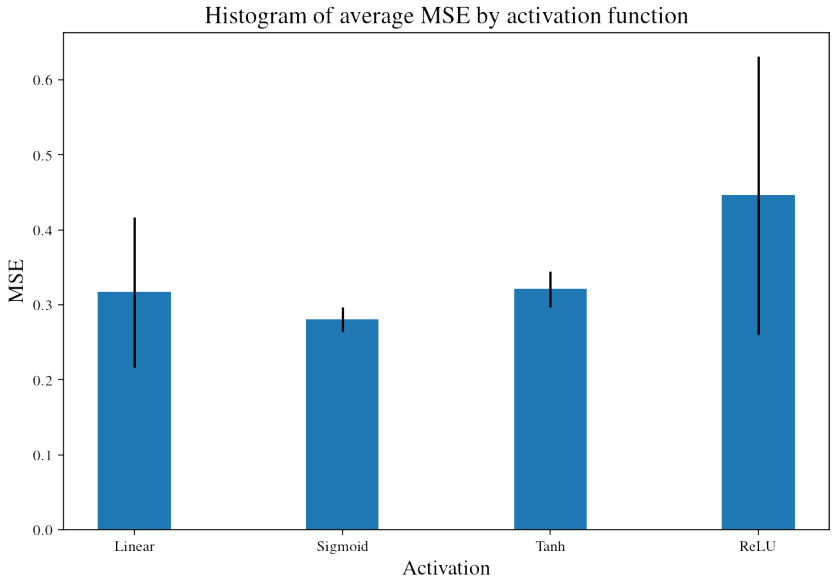
The black bars indicate one standard deviation around the mean. Overall there appears to be no consistent pattern to these results.

The next graph shows the MSE by the number of hidden nodes. Similarly those with zero hidden layers, and hence zero hidden nodes, are removed. Again we see no clear pattern, although the variance is slightly higher for the networks with 256 hidden nodes in each layer.

[20] The graphs including these results can be viewed in [code omitted for peer review].



The results are slightly clearer for the activation function used, as we see the models using the sigmoid and tanh activation tend to perform better and with far less variance than the other two activations. The ReLU activation has the worse performance on average and the highest variance. To better understand these results I now turn to a regression analysis of the relationship between these different architectures and the model performance.



F. Evaluating the effect of network architecture on MSE using regression[21]

To assess whether different network structures and activation functions are associated with statistically significant changes in model performance, we can specify an OLS regression, where the dependent variable is the mean squared error of the model and the independent variables are the number of hidden layers, the number of nodes in each layer, and dummy variables for the activation function (with the linear activation as the reference category). The results of this regression are shown in Table 1. Model 1 shows the results with just the basic features. Overall we do not see any statistically significant predictors of MSE. Note the large intercept value of 2.804 and the low R-squared, both of which suggest this model is a poor fit. This does not change when the interaction between the number of hidden layers and the size of each hidden layer is included. In Model 2 an interaction term is added to assess whether there is a multiplicative relationship between the number of layers and the size of each layer. The term is not statistically significant and makes no substantive change to the model. In Model 3 I removed the 6 outlier observations where MSE > 1.0. We can see that the model fit has improved considerably, with a R-squared of 0.374. Hidden layer size now has a small but statistically significant positive association with MSE, suggesting that the inclusion of too many neurons may be detrimental to model performance. As expected we also see that there is a positive association between the ReLU activation function and MSE, as models using this activation tended to perform considerably worse than the linear activation. The coefficients for the other two activation functions are negative but not significant. Again, we see no major difference in Model 4 where the interaction term is included, although

[21] The code to reproduce these analyses can be found in the notebook "Assessing architecture and performance.ipynb".

note that the coefficient for layer size is no longer significant in this model. Overall these models only explain around a third of the variance in MSE. This lack of explanatory power may be partially a function of the low sample size, but is also likely a consequence of stochastic variation within the different networks that is not easily reducible to their basic architectures. Further work is therefore necessary to establish precisely how these different features affect the predictive performance and in which cases different types of architectures should be preferred.

Table 1: OLS regression predicting MSE by model architecture

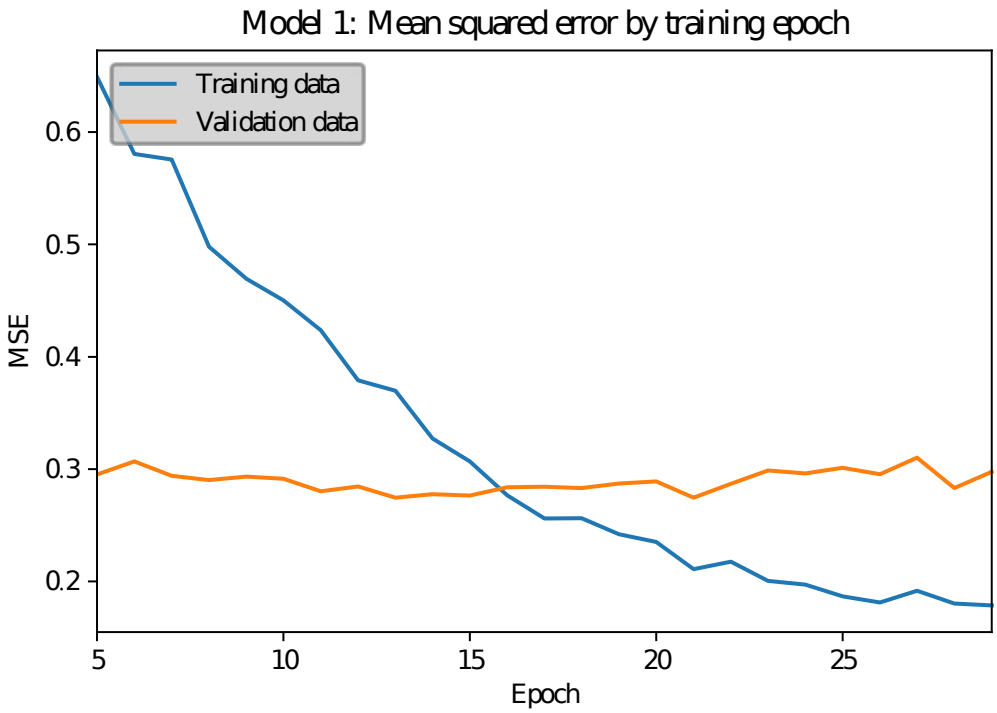
	Model 1	Model 2	Model 3	Model 4
Number of hidden layers	-0.015	-0.011	-0.002	-0.001
	(0.775)	(0.518)	(0.022)	(0.015)
Number of neurons (per layer)	0.005	0.008	0.0005*	0.001
	(0.009)	(0.260)	(0.0002)	(0.007)
Num. hidden layers * Num. neurons		-0.003		-0.0004
		(0.258)		(0.007)
ReLU	-2.309	-2.309	0.116**	0.116**
	(2.041)	(2.041)	(0.053)	(0.053)
Sigmoid	-2.818	-2.818	-0.049	-0.049
	(2.041)	(2.041)	(0.053)	(0.053)
Tanh	-2.782	-2.782	-0.010	-0.010
	(2.041)	(2.041)	(0.053)	(0.053)
Intercept	2.804	2.804	0.266***	0.266***
	(2.090)	(2.090)	(0.064)	(0.064)
N	40	40	34	34
R^2	0.079	0.079	0.374	0.374
R^2-adj	-0.056	-0.056	0.263	0.263
F	0.585	0.585	3.352	3.352

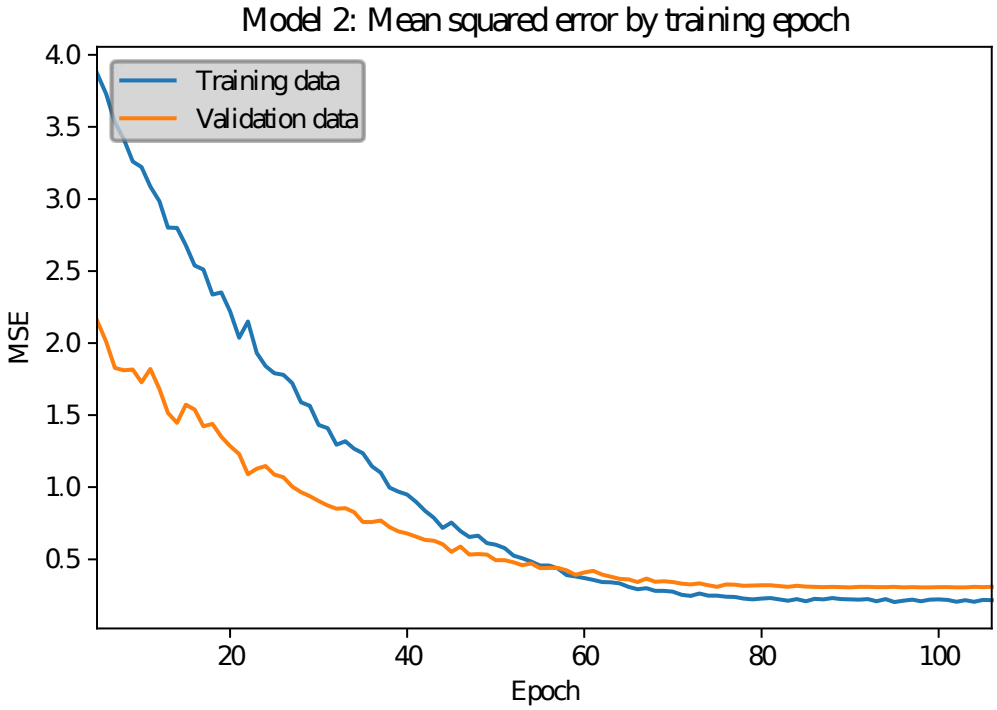
Standard errors are reported in parentheses. * = p <0.05, ** = p <0.01, *** = p < 0.001.

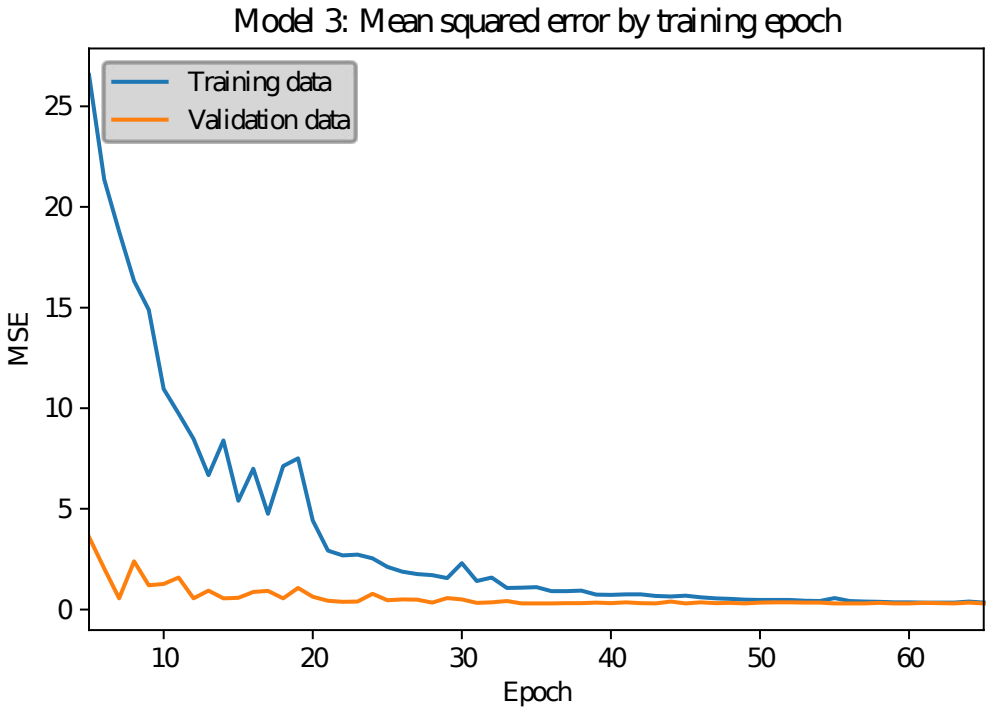
G. Examining model loss during training to observe learning and over-fitting

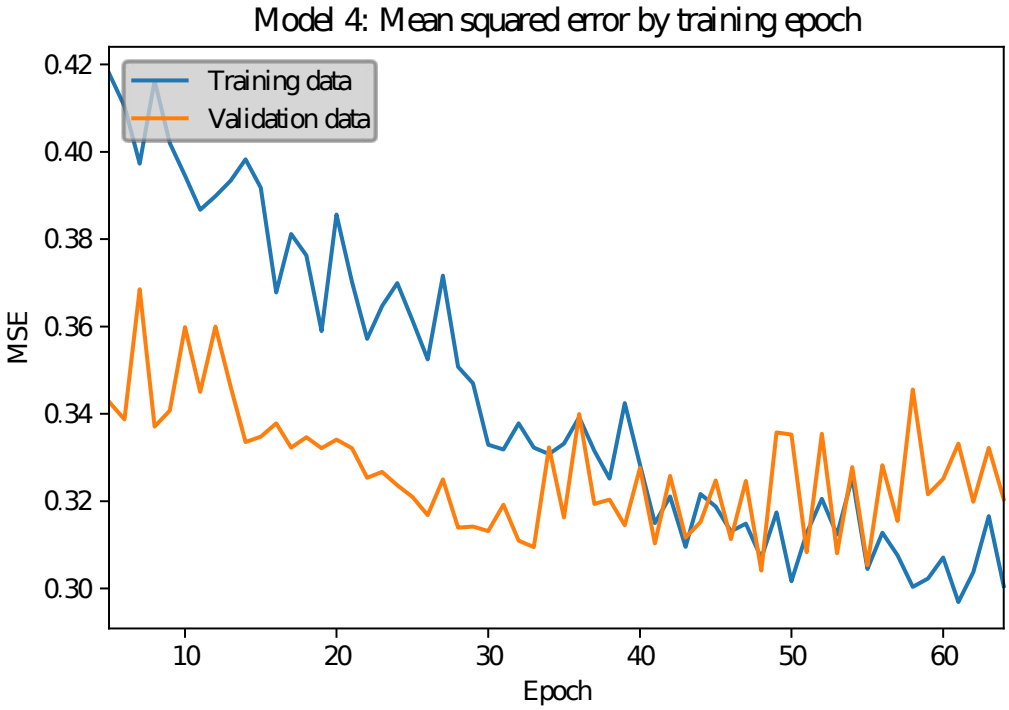
At each epoch in the training process we can observe each how the model performs in terms of mean squared error on both the training data and the validation data, the latter of which is never used to update the model weights. This provides valuable insights into the training process as it allows us to see how quickly the model is learning, and the extent to which it is learning functions that generalize well across both sets or is over-fitting to the training data. In an ideal case, as the model trains the two lines will converge and overlap perfectly, as the model learns a universal function that performs equally well on the training data and the unseen validation data. However, if the performance on the training set continues to improve while the performance on the validation set does not, then it is likely that the model is over-fitting the training data. The plots below show the mean squared error at each epoch for the five best performing models. Note that the x-axis differs across graphs due to different stopping times (all models stopped due to early-stopping rather than reaching 200 epochs, the maximum number allowed) and the y-axis differs based on the range of MSE values that occurred in both sets of data. In all cases we see that the models improve over time.[22]

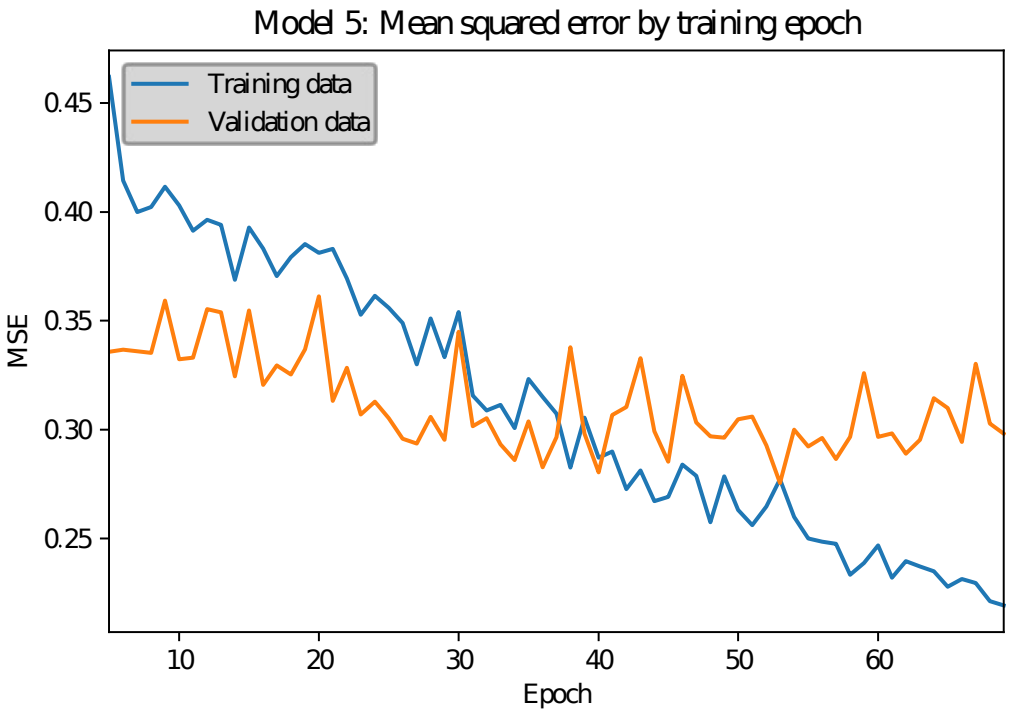
[22] Note that the first two epochs have been removed from each plot, as the first couple of epochs tend to have very high MSE values that extend the vertical axis and render the rest of the plot invisible (the untruncated plots can be viewed in the accompanying notebook).


Model 1: Mean squared error by training epoch
Training data
Validation data
MSE
0.6
0.5
0.4
0.3
0.2
5
10
15
20
25
Epoch


Model 2: Mean squared error by training epoch
Training data
Validation data
MSE
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
20
40
60
80
100
Epoch


Model 3: Mean squared error by training epoch
Training data
Validation data
MSE
Epoch
0
5
10
15
20
25
10
20
30
40
50
60


Model 4: Mean squared error by training epoch
Training data
Validation data
MSE
Epoch
0.30
0.32
0.34
0.36
0.38
0.40
0.42
10
20
30
40
50
60



A problem I noticed with the use of early-stopping is that in some cases the early-stopping is triggered before a model has had sufficient time to emerge at a stable solution. This seems to be the case for a number of the models using the ReLU activation function, which tended take a considerable number of epochs to reach reasonable estimates. For example the MSE for Model 2 does not drop below 1.0 until almost the 40th epoch. Conversely, I also observed the opposite problem with some of the other models, where the patience period led to the models over-fitting the training data for a number of epochs before they were stopped. This is particularly evident in Model 5, where we see that the MSE decreases in the training data at an almost constant rate throughout training, while the MSE for the validation data plateaus. This suggests that a one-size-fits all approach is probably inappropriate and that other parameters should be optimized based on the type of activation function used.

H. Assessing LIME explanations

The "explanations" produced by the LIME algorithm consist of a set of features/variables, the values of these for a particular observation (e.g. X <= 1, or 0 < Y <1), estimated coefficients for these variables in the local model, and point estimates for the outcome variable, high school GPA, derived from the local model. The goal of LIME is to provide a simplified approximation of the relationship between the inputs—in this case information about an individual respondent—and the predicted outcome. Moreover, the explanations provided by the algorithm are intended to be interpretable, with local-fidelity, giving an understandable insight into how the model is behaving in the vicinity of a given observation. Although the results of LIME are a simplification of the neural network model and are not as readily interpretable as the coefficients for a linear regression, they provide us with some insight into what is otherwise a black box and can help us to assess whether the model is performing as expected and what features are important to the prediction.

For example, here is the output obtained for observation 114:

```
{'f1g9g_6.0 <= 0.00': 0.072983960408029208,
 'f2k5_1.0 <= 0.00': 0.013595649885988848,
'm3i0f_1.0 <= 0.00': -0.010205085228747721,
'm5e9_7_1.0 <= 0.00': 0.018694376377378762,
 'p5j2g_2.0 <= 0.00': 0.011027617946481282}
```

This explanation contains five different variables, specific values of the variables, and thresholds, and then corresponding coefficients from the local model. For example, the first row shows the question with the code f1g9g[23], which corresponds to the question "In the past week, how often did you feel fearful?" asked to the father in the baseline survey.

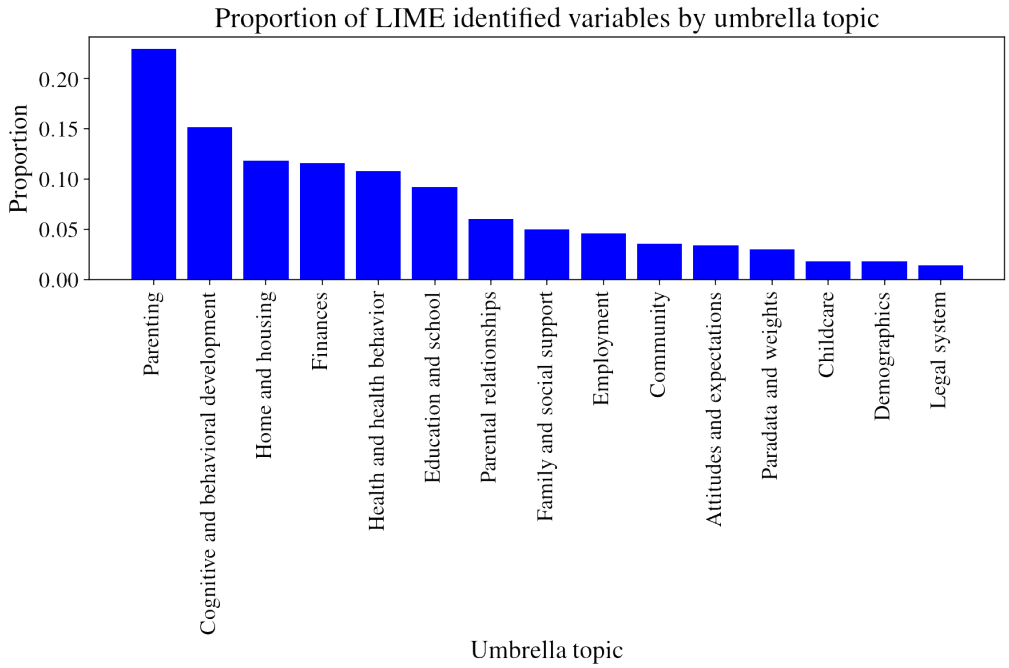
[23] View information about this question on the metadata website: http://browse.fragilefamiliesmetadata.org/variables/f1g9g

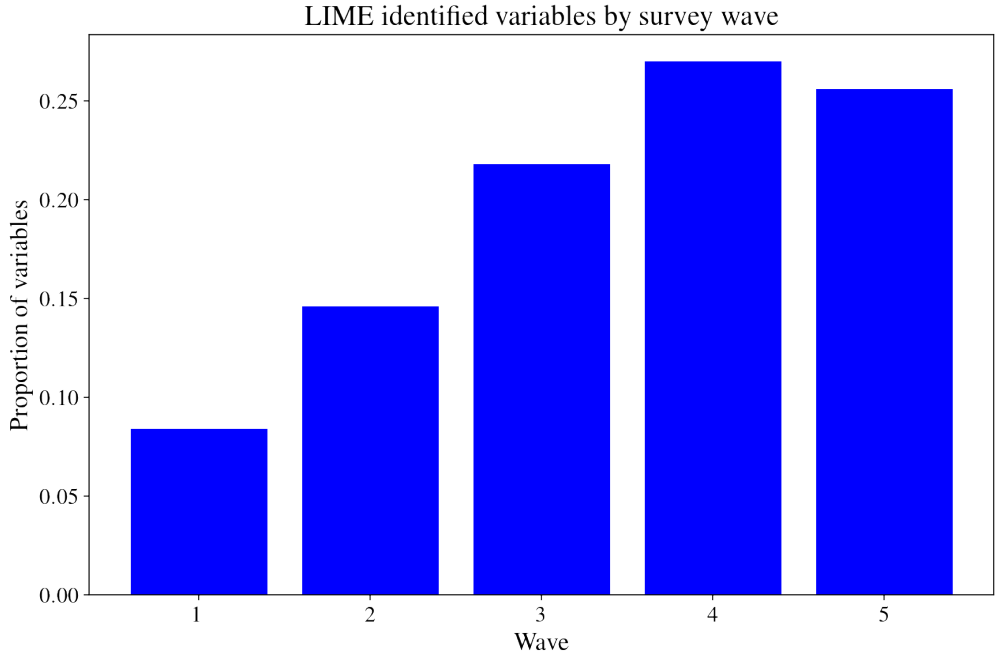
The value 6 indicates that the respondent reported that they felt fearful 6 days a week. The "<= 0" string after this indicates that the value for this respondent is 0 or less, meaning that they did not respond positively to this category of the question. The coefficient indicates the weight the local model learned by LIME put on this answer. For the analyses presented in the paper I decided to focus solely on the variables themselves, rather than their specific values, as well as the respondents and waves they correspond to.

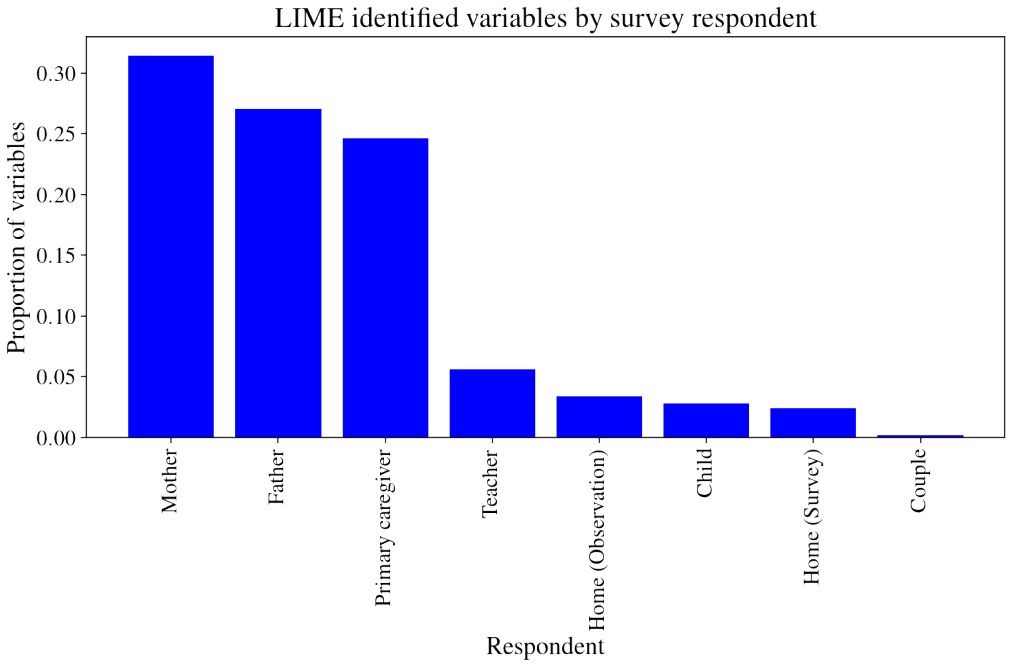
The Fragile Families metadata API allowed me to obtain information about almost all of the variables in the dataset (this was unavailable for 6 of them), including information about the general topic of each variable, which was included by Fragile Families staff after the conclusion of the challenge. Two different types of topics are available: "umbrella topics", which are broad categories such as "parenting" and "education and school", and "topics", which correspond to slightly more fine-grained distinctions such as "parenting behavior" and "parent relationship status". Each variable is associated with either 1 or 2 of each type of topics. Here and in the paper I restrict the analysis to umbrella topics, although results for the other topics are reported in the table below and can be viewed in the corresponding notebook [link to notebook redacted for peer review].

As discussed in the paper, each LIME explanation contains five variables, resulting in 419 unique variables across all 100 of the sampled respondents, out of a maximum possible of 500. For each variable I identify the associated topics, counting the number of times that variables associated with each topic appear. I then divide these counts by 500 to get a measure of the proportion of variables in the LIME explanations

that correspond to each topic. The same process is repeated for the survey wave and respondent. The results are shown below:



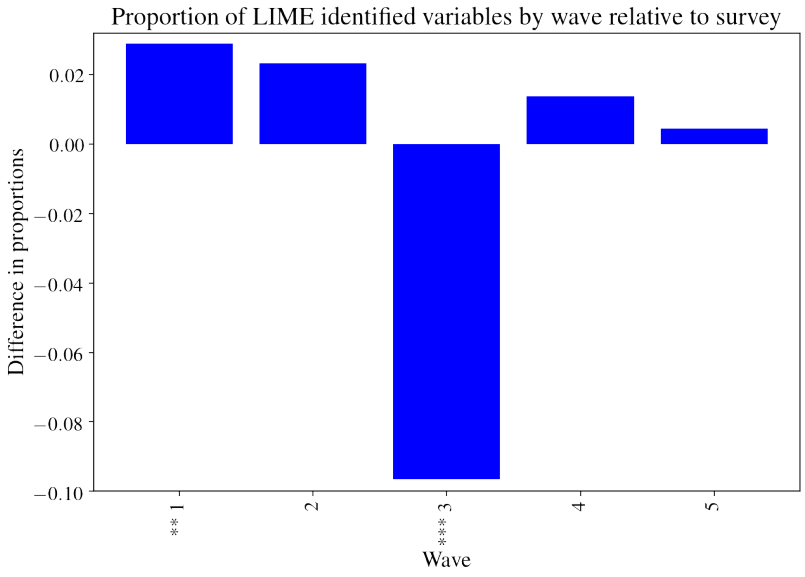


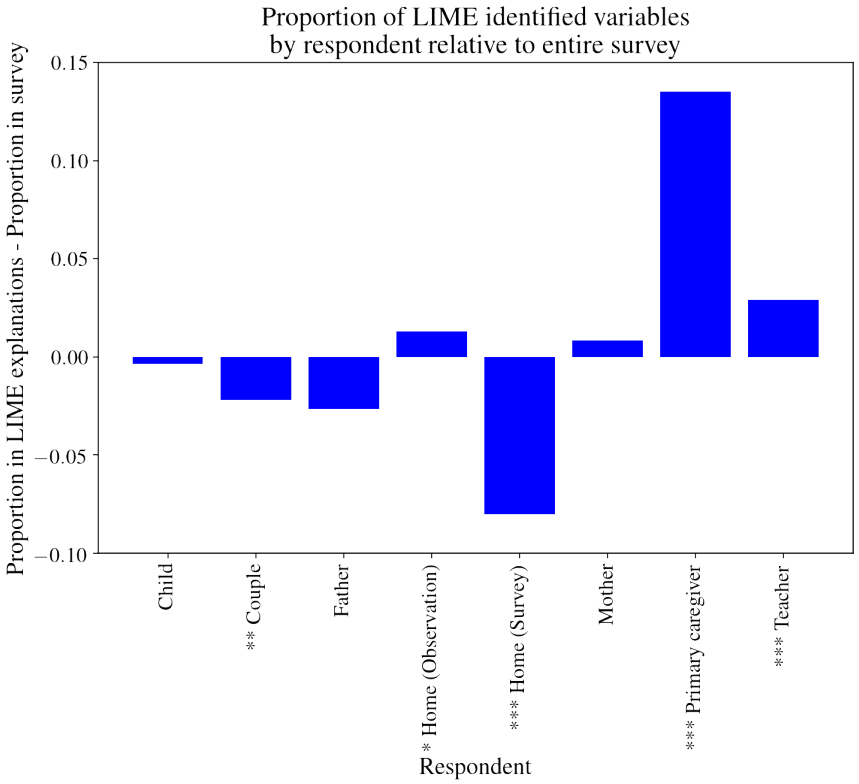


As discussed in the main paper, the issue with interpreting these plots is that the underlying distribution of topics, waves, and respondents across the survey is non-uniform.[24] If we took a random draw of 500 variables from the survey we would see that some categories are overrepresented purely due to their frequency. To assess how the observed results compare to what we would expect by chance I (a) use z-tests to compare the proportions of each category in the LIME explanations to the proportions in the overall survey (waves 1 through 5), (b) plot the difference in proportions, and (c) plot the ratio of the categories in the explanations to the survey.

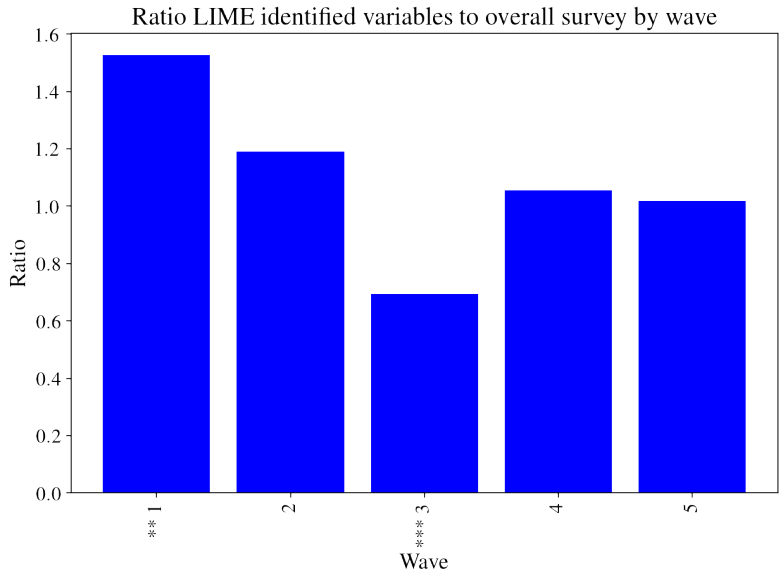
The set of plots below shows the differences in proportions for waves and respondents, respectively (topics are shown in Figure 3 of the main paper). These results are discussed in detail in the main paper but the figures were omitted. Statistically significant differences from the z-tests are indicated by stars next to the category names, where * = $p < 0.05$, ** = $p < 0.01$, and *** = $p < 0.001$.

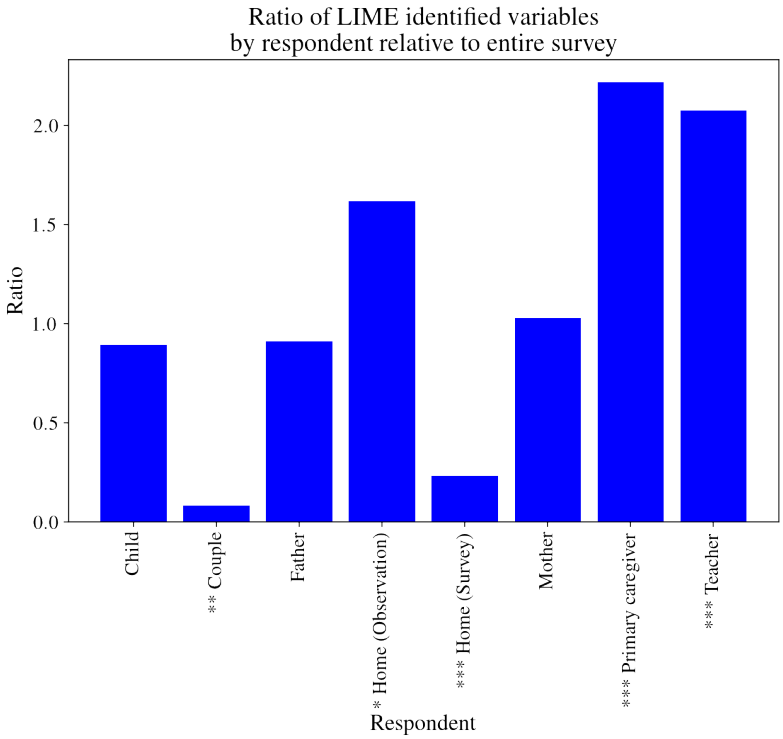
[24] These distributions are plotted in the notebook [removed for peer-review].





Next I show the ratios of categories in the LIME explanations to the overall survey for the waves and respondents respective. Overall we see very similar results:





I. Full list of variables identified by LIME and metadata:

More information about each variable can be obtained by searching for the variable name on the Fragile Families metadata website: http://browse.fragilefamiliesmetadata.org/.

Variable name	# LIME exp.	Wave	Respondent	Umbrella topic(s)
cf1edu	1	1	Father	Education and school
cf1kids	1	1	Father	Home and housing
cf1povca	1	1	Father	Finances
cf2hhimpb	1	2	Father	Finances
cf2povca	1	2	Father	Finances
cf3cogsc	1	3	Father	Cognitive and behavioral development
cf3finjail	1	3	Father	Legal system
cf3povcob	1	3	Father	Finances
cf4b_age	1	4	Father	Demographics & Paradata and weights
cf5marm	1	5	Father	Parental relationships
ch3pvbasal_m	1	3	Home (Survey)	Cognitive and behavioral development
ch3pvnbasal_m	1	3	Home (Survey)	Cognitive and behavioral development
ch3pvnceil_m	1	3	Home (Survey)	Cognitive and behavioral development
ch4ppvtage	1	4	Home (Survey)	Demographics & Paradata and weights
ch4pvceil	1	4	Home (Survey)	Cognitive and behavioral development
ch4pvceilr	1	4	Home (Survey)	Cognitive and behavioral development & Paradata and weights
ch5dsraw	1	5	Home (Survey)	Cognitive and behavioral development
ch5dsss	1	5	Home (Survey)	Cognitive and behavioral development
ch5ppvtpr	1	5	Home (Survey)	Cognitive and behavioral development
ch5ppvtraw	1	5	Home (Survey)	Cognitive and behavioral development
ch5ppvtss	1	5	Home (Survey)	Cognitive and behavioral development
cm1inpov	1	1	Mother	Finances
cm1lenhr	1	1	Mother	Paradata and weights
cm2cfinjail	1	2	Mother	Legal system
cm3b_age	1	3	Mother	Demographics & Paradata and weights
cm3kids	1	3	Mother	Home and housing
cm4age	1	4	Mother	Demographics & Paradata and weights
cm4b_age	2	4	Mother	Demographics & Paradata and weights
cm5finjail	1	5	Mother	Legal system
f1a6a	1	1	Father	Health and health behavior
f1b10	1	1	Father	Parental relationships
f1d1c	1	1	Father	Attitudes and expectations
f1e1b1	1	1	Father	Home and housing
f1e1b3	1	1	Father	Home and housing
f1e3	1	1	Father	Family and social support
f1e4a	1	1	Father	Family and social support
f1f6	2	1	Father	Community

f1f7	1	1	Father	Demographics
f1g9a	1	1	Father	Health and health behavior
f1g9g	1	1	Father	Health and health behavior
f1g9l	1	1	Father	Health and health behavior
f1i1	2	1	Father	Education and school
f1j6	3	1	Father	Employment
f1j8a	1	1	Father	Finances
f1k3	1	1	Father	Finances
f2a3	1	2	Father	Home and housing
f2b36b	1	2	Father	Parenting
f2b36c	1	2	Father	Parenting
f2b36f	1	2	Father	Parenting
f2b36h	1	2	Father	Parenting
f2b37a	2	2	Father	Cognitive and behavioral development
f2b37d	1	2	Father	Cognitive and behavioral development
f2b37e	1	2	Father	Cognitive and behavioral development
f2b38d	1	2	Father	Parenting
f2c6c	1	2	Father	Legal system
f2d2b	2	2	Father	Parental relationships & Parenting
f2d2c	1	2	Father	Parental relationships & Parenting
f2d2d	1	2	Father	Parental relationships & Parenting
f2g16	1	2	Father	Family and social support
f2g3	1	2	Father	Family and social support
f2h7c	2	2	Father	Finances
f2j22	1	2	Father	Cognitive and behavioral development
f2k15ap	1	2	Father	Finances
f2k5	1	2	Father	Education and school
f2k7a	1	2	Father	Education and school
f3a11	1	3	Father	Health and health behavior
f3a2	1	3	Father	Home and housing
f3b4e	1	3	Father	Parenting
f3b4g	1	3	Father	Parenting
f3b4h	1	3	Father	Parenting
f3b4k	1	3	Father	Parenting
f3b4m	1	3	Father	Parenting
f3b6d	1	3	Father	Parenting
f3b7	1	3	Father	Childcare
f3c3e	1	3	Father	Parenting
f3c3i	1	3	Father	Parenting
f3c3j	1	3	Father	Parenting
f3c7b	1	3	Father	Parental relationships
f3d1d	1	3	Father	Parental relationships & Parenting
f3d3a_2	1	3	Father	Family and social support

f3d3a_7	1	3	Father	Family and social support
f3d4	1	3	Father	Parental relationships
f3f1	1	3	Father	Home and housing
f3f2c1	1	3	Father	Home and housing
f3f2d3	3	3	Father	Home and housing
f3i20	1	3	Father	Finances
f3j2e2	1	3	Father	Health and health behavior
f3k13p	1	3	Father	Finances
f3k23	1	3	Father	Employment
f3l3	1	3	Father	Parental relationships & Parental relationships
f3r0a	1	3	Father	Attitudes and expectations
f3r10	2	3	Father	Community
f4a4	1	4	Father	Parental relationships
f4a8c	1	4	Father	Parental relationships & Home and housing
f4b3	1	4	Father	Home and housing
f4b4b13	2	4	Father	Cognitive and behavioral development
f4b4b18	1	4	Father	Cognitive and behavioral development & Health and health behavior
f4b4b19	1	4	Father	Cognitive and behavioral development
f4b6a	3	4	Father	Parenting
f4b6b	1	4	Father	Parenting
f4c27	1	4	Father	Health and health behavior
f4c3f	1	4	Father	Parenting
f4c43a	1	4	Father	Health and health behavior
f4d1g	1	4	Father	Parental relationships & Parenting
f4d4	1	4	Father	Parental relationships
f4d8	1	4	Father	Parental relationships
f4f2c2	1	4	Father	Home and housing
f4f2c3	1	4	Father	Home and housing
f4f2d1	2	4	Father	Home and housing
f4h1p	1	4	Father	Family and social support
f4h1q	1	4	Father	Family and social support
f4h2	1	4	Father	Finances
f4i0g	1	4	Father	Education and school & Family and social support
f4i0m4	1	4	Father	Community
f4i0n2	1	4	Father	Community
f4i22	1	4	Father	Finances
f4i23p2	1	4	Father	Attitudes and expectations & Finances
f4i23p3	2	4	Father	Attitudes and expectations & Finances
f4i23p5	1	4	Father	Attitudes and expectations & Finances
f4j25b1	1	4	Father	Cognitive and behavioral development
f4k10	3	4	Father	Employment
f4k12	1	4	Father	Employment

f4k13	1	4	Father	Finances
f4l5b	1	4	Father	Finances & Finances
f4r2	1	4	Father	Community
f5a6a	2	5	Father	Health and health behavior
f5a8	1	5	Father	Health and health behavior
f5b22c	1	5	Father	Finances
f5c2b	1	5	Father	Parental relationships & Parenting
f5e1i	2	5	Father	Family and social support & Family and social support
f5i13	1	5	Father	Finances
f5i22	2	5	Father	Employment
f5k14d	1	5	Father	Parenting
f5k14e	1	5	Father	Parenting
h5m13	1	5	Home (Survey)	Paradata and weights
k5a1a	3	5	Child	Parenting
k5a1c	1	5	Child	Parenting
k5a3c	1	5	Child	Parenting
k5b1b	1	5	Child	Parenting
k5c1	1	5	Child	Cognitive and behavioral development
k5d1a	1	5	Child	Cognitive and behavioral development
k5d1c	1	5	Child	Cognitive and behavioral development
k5e1b	1	5	Child	Education and school
k5e1c	1	5	Child	Education and school
k5e2b	1	5	Child	Education and school
k5g1b	1	5	Child	Cognitive and behavioral development
k5g2i	1	5	Child	Cognitive and behavioral development
m1b12c	1	1	Mother	Parental relationships
m1b15f	2	1	Mother	Attitudes and expectations
m1c1a	1	1	Mother	Attitudes and expectations
m1e1b1	1	1	Mother	Home and housing
m1e1b2	1	1	Mother	Home and housing
m1e3e	1	1	Mother	Childcare & Family and social support
m1f10a	1	1	Mother	Attitudes and expectations & Finances
m1f10b	1	1	Mother	Attitudes and expectations & Finances
m1f11b	2	1	Mother	Finances
m1f12	1	1	Mother	Attitudes and expectations & Finances
m1f2	1	1	Mother	Home and housing
m1g1	1	1	Mother	Health and health behavior & Health and health behavior
m1i6	1	1	Mother	Employment & Education and school
m1j2b	2	1	Mother	Finances
m2a7	1	2	Mother	Parental relationships
m2a8e	1	2	Mother	Parental relationships
m2b13b	1	2	Mother	Health and health behavior

m2b17d	1	2	Mother	Cognitive and behavioral development
m2b17f	1	2	Mother	Cognitive and behavioral development
m2b18b	1	2	Mother	Parenting
m2b18c	1	2	Mother	Parenting
m2b18d	1	2	Mother	Parenting
m2b18h	1	2	Mother	Parenting
m2b20b	1	2	Mother	Parenting
m2b27	3	2	Mother	Childcare
m2b29	1	2	Mother	Childcare & Finances
m2b8	2	2	Mother	Health and health behavior
m2c23c	1	2	Mother	Finances
m2c24a	1	2	Mother	Health and health behavior
m2c3f	1	2	Mother	Parenting
m2c3h	2	2	Mother	Parenting
m2c3i	1	2	Mother	Parenting
m2citywt_rep6	1	2	Mother	Paradata and weights
m2citywt_rep8	1	2	Mother	Paradata and weights
m2d2	1	2	Mother	Parenting
m2f2b4	1	2	Mother	Home and housing
m2f2d1	2	2	Mother	Home and housing
m2f2e1	1	2	Mother	Home and housing
m2f6	1	2	Mother	Health and health behavior
m2g1b	1	2	Mother	Family and social support & Demographics
m2g3	1	2	Mother	Family and social support
m2g5a8	2	2	Mother	Finances
m2g6a1	1	2	Mother	Family and social support
m2g6d	1	2	Mother	Family and social support
m2h19c	1	2	Mother	Finances
m2h2	2	2	Mother	Home and housing
m2h8g	1	2	Mother	Finances
m2h9a3	1	2	Mother	Finances
m2j4b3	1	2	Mother	Health and health behavior
m2k16	1	2	Mother	Employment
m2k6	1	2	Mother	Employment
m2k8a	1	2	Mother	Employment & Parenting
m2l2	1	2	Mother	Finances
m3a10	1	3	Mother	Health and health behavior
m3a13	1	3	Mother	Parental relationships
m3a2	1	3	Mother	Home and housing
m3a4	1	3	Mother	Parental relationships
m3b23	1	3	Mother	Childcare
m3b3	2	3	Mother	Parenting

m3b4a	1	3	Mother	Parenting
m3b4c	1	3	Mother	Parenting
m3b4d	1	3	Mother	Parenting
m3b4j	1	3	Mother	Parenting
m3b4m	2	3	Mother	Parenting
m3b6a	1	3	Mother	Parenting
m3c1d	1	3	Mother	Legal system & Home and housing
m3c30d	1	3	Mother	Finances
m3c3j	1	3	Mother	Parenting
m3c41	2	3	Mother	Legal system & Employment
m3d1f	1	3	Mother	Parental relationships & Parenting
m3d3	1	3	Mother	Family and social support
m3d8	1	3	Mother	Parental relationships
m3f2d3	2	3	Mother	Home and housing
m3h6a	1	3	Mother	Family and social support
m3i0f	1	3	Mother	Community
m3i21	1	3	Mother	Finances
m3j18a	1	3	Mother	Health and health behavior
m3j44a	1	3	Mother	Cognitive and behavioral development
m3k11	1	3	Mother	Employment & Employment
m3k3b	1	3	Mother	Finances
m3l4a	1	3	Mother	Finances
m4a13	3	4	Mother	Parental relationships
m4a6	1	4	Mother	Parental relationships
m4b3	2	4	Mother	Home and housing
m4b4a2	1	4	Mother	Parenting
m4b4a3	1	4	Mother	Parenting
m4b6c	1	4	Mother	Parenting
m4b8a	1	4	Mother	Childcare
m4b8c	1	4	Mother	Childcare
m4c27	1	4	Mother	Health and health behavior
m4c35b	2	4	Mother	Home and housing
m4c36	2	4	Mother	Employment & Education and school
m4c3e	1	4	Mother	Parenting
m4c3g	1	4	Mother	Parenting
m4f2d1	1	4	Mother	Home and housing
m4f2d3	2	4	Mother	Home and housing
m4f2d4	1	4	Mother	Home and housing
m4h1l	1	4	Mother	Family and social support
m4h4	1	4	Mother	Family and social support
m4i0n1	1	4	Mother	Community
m4i2	2	4	Mother	Home and housing
m4i4	4	4	Mother	Finances

m4i4a	1	4	Mother	Home and housing
m4j25a1	1	4	Mother	Cognitive and behavioral development
m4j2e	1	4	Mother	Health and health behavior
m4k9b	1	4	Mother	Employment & Employment
m4r3	1	4	Mother	Demographics
m5a2	1	5	Mother	Parenting
m5a51	1	5	Mother	Home and housing
m5a5d04	3	5	Mother	Home and housing
m5b3	1	5	Mother	Parenting
m5c1d	1	5	Mother	Attitudes and expectations & Parenting
m5c2b	1	5	Mother	Parental relationships & Parenting
m5c3c	1	5	Mother	Parental relationships & Parenting
m5e2	1	5	Mother	Finances
m5e9_7	1	5	Mother	Family and social support
m5f20	1	5	Mother	Finances
m5f4b	1	5	Mother	Home and housing
m5i12a_code	1	5	Mother	Paradata and weights
m5i4	1	5	Mother	Employment
m5j5b	1	5	Mother	Finances
m5j6b1	1	5	Mother	Finances
o3u2	1	3	Home (Observation)	Cognitive and behavioral development
o3v2	1	3	Home (Observation)	Cognitive and behavioral development
o3v3	1	3	Home (Observation)	Paradata and weights
o4p7b	1	4	Home (Observation)	Home and housing
o4r10a_3	1	4	Home (Observation)	Home and housing
o4r12	2	4	Home (Observation)	Home and housing
o4r13	1	4	Home (Observation)	Home and housing
o4r2	1	4	Home (Observation)	Home and housing
o4u1	1	4	Home (Observation)	Cognitive and behavioral development
o4v2	1	4	Home (Observation)	Cognitive and behavioral development
o4v6b	1	4	Home (Observation)	Cognitive and behavioral development
o5a3	1	5	Home (Observation)	Community
o5a8	1	5	Home (Observation)	Community
o5f2	1	5	Home (Observation)	Cognitive and behavioral development

o5g2	1	5	Home (Observation)	Cognitive and behavioral development
o5g8_3	1	5	Home (Observation)	Paradata and weights
p3a1	1	3	Primary caregiver	Health and health behavior & Health and health behavior
p3a21	3	3	Primary caregiver	Home and housing & Health and health behavior
p3a6a	3	3	Primary caregiver	Health and health behavior
p3b1	2	3	Primary caregiver	Parenting
p3b3	1	3	Primary caregiver	Parenting
p3c6a	2	3	Primary caregiver	Parenting
p3c6e	1	3	Primary caregiver	Parenting & Community
p3g1l	1	3	Primary caregiver	Health and health behavior
p3j13	1	3	Primary caregiver	Parenting
p3j15	1	3	Primary caregiver	Parenting
p3j19	1	3	Primary caregiver	Health and health behavior & Parenting
p3j2	1	3	Primary caregiver	Parenting
p3j23b	2	3	Primary caregiver	Parenting
p3j23e	2	3	Primary caregiver	Parenting
p3j23h	2	3	Primary caregiver	Parenting
p3j3	1	3	Primary caregiver	Parenting
p3j4	1	3	Primary caregiver	Parenting
p3j9	2	3	Primary caregiver	Parenting
p3k1a	1	3	Primary caregiver	Community
p3k1c	1	3	Primary caregiver	Community
p3k3b	1	3	Primary caregiver	Community
p3k3f	1	3	Primary caregiver	Community
p3m16	1	3	Primary caregiver	Cognitive and behavioral development
p3m17	1	3	Primary caregiver	Cognitive and behavioral development
p3m2	1	3	Primary caregiver	Cognitive and behavioral development
p3m34	1	3	Primary caregiver	Cognitive and behavioral development
p4a14	1	4	Primary caregiver	Health and health behavior & Health and health behavior
p4a24	1	4	Primary caregiver	Health and health behavior
p4a30	1	4	Primary caregiver	Health and health behavior
p4b1	1	4	Primary caregiver	Parenting
p4b14	1	4	Primary caregiver	Parenting
p4b2	1	4	Primary caregiver	Parenting
p4b24	3	4	Primary caregiver	Cognitive and behavioral development & Health and health behavior
p4b6	2	4	Primary caregiver	Parenting
p4c14	1	4	Primary caregiver	Home and housing
p4c17b	1	4	Primary caregiver	Parenting
p4c1e	1	4	Primary caregiver	Parenting
p4d1b	1	4	Primary caregiver	Finances

p4d2	1	4	Primary caregiver	Finances
p4e3	1	4	Primary caregiver	Home and housing
p4f1b	1	4	Primary caregiver	Attitudes and expectations
p4f1e	1	4	Primary caregiver	Attitudes and expectations & Parenting
p4f2a1	1	4	Primary caregiver	Parenting
p4f2b1	1	4	Primary caregiver	Parenting
p4f3b	1	4	Primary caregiver	Attitudes and expectations & Health and health behavior
p4g1	3	4	Primary caregiver	Parenting
p4g10	1	4	Primary caregiver	Parenting
p4g13	1	4	Primary caregiver	Parenting
p4g21	1	4	Primary caregiver	Home and housing
p4g23a	1	4	Primary caregiver	Parenting
p4g23b	1	4	Primary caregiver	Parenting
p4g23d	1	4	Primary caregiver	Parenting
p4g23h	1	4	Primary caregiver	Parenting
p4g23i	1	4	Primary caregiver	Parenting
p4g6	1	4	Primary caregiver	Parenting
p4g8	1	4	Primary caregiver	Parenting
p4j1	1	4	Primary caregiver	Parenting
p4l1	1	4	Primary caregiver	Cognitive and behavioral development
p4l11	1	4	Primary caregiver	Cognitive and behavioral development
p4l22	1	4	Primary caregiver	Cognitive and behavioral development
p4l23	1	4	Primary caregiver	Cognitive and behavioral development
p4l43	1	4	Primary caregiver	Cognitive and behavioral development
p4l57	1	4	Primary caregiver	Cognitive and behavioral development
p4l66	1	4	Primary caregiver	Cognitive and behavioral development
p5h15c	2	5	Primary caregiver	Health and health behavior
p5h17b	1	5	Primary caregiver	Health and health behavior
p5h6	1	5	Primary caregiver	Health and health behavior
p5h8	1	5	Primary caregiver	Health and health behavior
p5i12	1	5	Primary caregiver	Health and health behavior & Parenting
p5i1d	1	5	Primary caregiver	Parenting
p5i1g	1	5	Primary caregiver	Parenting
p5i21b	1	5	Primary caregiver	Cognitive and behavioral development
p5i23	1	5	Primary caregiver	Family and social support
p5i24	1	5	Primary caregiver	Family and social support
p5i28	1	5	Primary caregiver	Parenting & Health and health behavior
p5i3	2	5	Primary caregiver	Cognitive and behavioral development
p5i32a	1	5	Primary caregiver	Education and school
p5i4	1	5	Primary caregiver	Cognitive and behavioral development
p5i6	1	5	Primary caregiver	Cognitive and behavioral development
p5j2g	1	5	Primary caregiver	Health and health behavior

p5j9	5	5	Primary caregiver	Finances
p5j9a	1	5	Primary caregiver	Finances
p5l10	1	5	Primary caregiver	Education and school
p5l12d	1	5	Primary caregiver	Home and housing & Education and school
p5l13g	1	5	Primary caregiver	Education and school
p5l18	1	5	Primary caregiver	Education and school
p5l3h	1	5	Primary caregiver	Education and school
p5m3a	1	5	Primary caregiver	Community
p5q1k	1	5	Primary caregiver	Parenting
p5q3a	1	5	Primary caregiver	Cognitive and behavioral development
p5q3ba	1	5	Primary caregiver	Health and health behavior & Health and health behavior
p5q3bb1	1	5	Primary caregiver	Health and health behavior
p5q3bb8	1	5	Primary caregiver	Health and health behavior
p5q3be	1	5	Primary caregiver	Cognitive and behavioral development
p5q3bo	1	5	Primary caregiver	Cognitive and behavioral development
p5q3bs	1	5	Primary caregiver	Cognitive and behavioral development
p5q3cg	1	5	Primary caregiver	Cognitive and behavioral development
p5q3cn	1	5	Primary caregiver	Cognitive and behavioral development
p5q3g	2	5	Primary caregiver	Cognitive and behavioral development
p5q3u	1	5	Primary caregiver	Cognitive and behavioral development
p5q4	1	5	Primary caregiver	Parenting
q2citywt_rep8	1	2	Couple	Paradata and weights
t5b1aa	2	5	Teacher	Cognitive and behavioural development & Education and school
t5b1s	1	5	Teacher	Cognitive and behavioral development & Education and school
t5b1u	1	5	Teacher	Cognitive and behavioral development & Education and school
t5b1w	2	5	Teacher	Cognitive and behavioral development & Education and school
t5b3c	1	5	Teacher	Cognitive and behavioral development & Education and school
t5b3e	1	5	Teacher	Cognitive and behavioral development & Education and school
t5b3f	1	5	Teacher	Cognitive and behavioral development & Education and school
t5b3j	1	5	Teacher	Cognitive and behavioral development & Education and school
t5b4l	1	5	Teacher	Cognitive and behavioral development & Education and school
t5c12	1	5	Teacher	Education and school
t5c16	1	5	Teacher	Education and school
t5c7c	1	5	Teacher	Education and school
t5d5	2	5	Teacher	Education and school
t5e10a2	1	5	Teacher	Education and school

t5e14_4	1	5	Teacher	Education and school
t5e15f	1	5	Teacher	Education and school
t5e17a	1	5	Teacher	Education and school
t5e20	1	5	Teacher	Education and school
t5e3	1	5	Teacher	Education and school
t5e9c	1	5	Teacher	Education and school
t5e9d	2	5	Teacher	Education and school
t5f1a	1	5	Teacher	Attitudes and expectations & Education and school
t5f4b	1	5	Teacher	Community & Education and school
t5f5d	1	5	Teacher	Education and school