Oxford Internet Institute, University of Oxford

Assignment Cover Sheet

Candidate Number *Please note, your OSS number is NOT your candidate number*	--------
Assignment *e.g. Online Social Networks*	Applied Machine Learning
Term *Term assignment issued, e.g. MT or HT*	HT
Title/Question *Provide the full title, or If applicable, note the question number and the FULL question from the assigned list of questions*	Fragile Families Challenge: Using Shapley Additive Explanations to Improve Model Interpretability
Word Count	4427

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

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

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

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

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

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

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

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

Please remember:

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

Fragile Families Challenge:

Using Shapley Additive Explanations to Improve Model Interpretability

1. Introduction

Data science competitions have become increasingly popular in the past decade. One competition that has received significant attention and that combines data science as well as social science research is the Fragile Families Challenge (Brooks-Gunn et al., 2011; Salganik et al., 2019). This challenge allows competitors to predict life outcomes of vulnerable families using various characteristics related to these families. The dataset contains six life outcomes of over 4200 families, with nearly 13000 predictors obtained from various questionnaires collected after a participant's child was born. The six outcomes of the challenge are related to the child of a fragile family and their primary caregiver. These outcomes include the child's GPA, grit, and material hardship, as well as outcomes related to if the child's primary caregiver was evicted, laid off from their job, or received job training. All outcome variables were obtained when the child turned 15 (Salganik et al., 2019).

The GPA outcome ranges from 1 to 4 and was determined by averaging the child's self-reported GPA from four subjects. These subjects include English, mathematics, social sciences, and natural sciences (Lundberg, 2017a). The grit outcome is used to measure the child's perseverance and was obtained by averaging the response to four questions related to the child's ability to see work through until the end. The variable is on a scale from 1 to 4, where a 4 represents children with the least amount of grit based on their answers to this questionnaire (Lundberg, 2017b). The material hardship outcome is used to measure the poverty levels of children. It was obtained by asking the primary caregiver 11 yes or no questions related to if the child has experienced extreme poverty. The metric is on a scale from 0 to 1 and represents the proportion of questions that the primary caregiver answered that indicate that the child has experienced extreme poverty. A score of 1 represents the children that experienced the highest degree of poverty based on this questionnaire (Lundberg, 2017c). The eviction outcome was obtained by asking the child's primary caregiver if they were evicted from their home since the last questionnaire (when the child turned 9). This variable is categorical, where it can be either true or false based on whether the primary caregiver was evicted (Lundberg, 2017d). The layoff outcome is categorical and was determined by asking the primary caregiver of each family if they were laid off since the last questionnaire (Lundberg, 2017e). The job training outcome was obtained by asking primary caregivers if they had taken classes to improve their job skills since the last questionnaire. This outcome is also binary (Lundberg, 2017f).

Researchers have used various methods to predict these life outcomes. The Fragile Families data is extremely messy, so most existing studies cleaned the data using various techniques before training their models. The part of the data cleaning process that typically receives the most attention is dealing with missing predictors. Many researchers have used simple imputation techniques for dealing with missing data, such as mean and mode imputation (Compton, 2019),

while others use more complex techniques, such as Amelia imputation (Stanescu et al., 2019). The next part of the data science pipeline of many studies is to select relevant features. This process depends on which kind of models were trained. Some researchers extracted hand-crafted features and achieved promising results (Ahearn & Brand, 2019). Other researchers used automatic variable selection techniques such as mutual information (Rigobon et al., 2019) and lasso regression (Stanescu et al., 2019). Several models do not require any feature selection techniques before model training since they automatically select features. Examples of these kinds of models include tree-based models, neural networks, and lasso regression.

Once the relevant features are extracted (if feature selection is necessary), researchers proceed to train their models. A variety of models have been trained to predict the outcomes of the Fragile Families Challenge. Examples include tree-based models (Rigobon et al., 2019), elastic net models (Raes, 2019), support vector machines (Garcia & Ta, 2018), and neural networks (Davidson, 2017). After training their models and reporting their performance, many researchers interpret the behavior of their models to understand why the model makes certain predictions. Simple models, such as linear regression and decision trees are easy to interpret. More sophisticated techniques have been used to interpret black box models. For instance, Davidson (2017) used LIME to interpret a neural network that predicted the GPA of children while competing in this challenge. It appears that few studies use state of the art techniques for model interpretability, such as SHapley Additive exPlanations (SHAP).

In my study I aim to address this gap in literature by interpreting black box models using SHAP. To this end, I predict the six outcomes of the Fragile Families Challenge using various techniques described in existing literature. I use simple data cleaning techniques and train various models that automatically learn their own features. Next, I use hyperparameter tuning techniques to determine optimal hyperparameters of each model. Lastly, I interpret my best performing models using SHAP to identify the predictors that most influence the outcome predictions, and to better understand which features may be related to at-risk families.

2. Methods

2.1. Data Cleaning

Before training my models, the dataset required significant cleaning. First all constant columns were removed from the dataset. Next, all non-response codes (such as integers between -9 and -1) were converted to missing data, and all columns with more than 80 percent missing data were removed. 5268 predictors remained after these cleaning steps. All numerical columns that had more than 20 distinct values were considered continuous variables, and all other types of columns were considered categorical. As the next step of the data cleaning process, missing data for continuous variables was imputed using mean imputation and missing data for categorical variables was imputed using mode imputation. Next, depending on the model type, the variables were ether convert to one hot encodings, or label encodings. The tree-based models performed better with label encodings, whereas the elastic net and neural network models performed better

with dummy variables. Dummy variables that occurred only one time in the entire dataset, as well as dummy variables that occurred all but once (dummy variables that occurred 4241 times) were removed from the dataset to eliminate dummy variables with minimal variance. The label encoded data contained 5268 features, and the dummy encoded data contained 19028. All the models used in my study automatically learn which features are most important for making predictions, so no feature selection was necessary.

After cleaning the data, the training set was split into a training and validation set, where 80 percent of the original training dataset was used for training, and 20 percent was used for model validation. This resulted in 1696 examples for training and 425 for validation. The challenge has a leaderboard and test dataset that contain 1591 and 530 observations, respectively. Not all these examples could be used for training and testing since certain families were missing data for certain outcomes. Therefore, the number of training and testing examples varied slightly between outcome variables.

2.2. Models

Once the data was cleaned, three types of models were trained to predict the six outcomes of the fragile family challenge. These models included tree-based models, elastic net models, and neural networks. The goal of the challenge is to minimize the mean-squared error of the continuous outcomes, and the brier loss of the categorical outcomes on the leaderboard and test set. The brier loss meant that the models used to predict categorical outcomes needed to output the underlying probability of a class, rather than a binary decision related to if the observation belonged to a certain class. The leaderboard accuracy is visible to all participants of the challenge, and the test set results are typically hidden from participants until the end of the competition. Since the competition was already completed several years ago, the ground truth outcomes of the test set were available. Therefore, the performance of each model was evaluated on both the leaderboard and test set.

The first type of models that were trained were tree-based models using Microsoft's Light Gradient Boosting Machine Framework (LGBM) (Machado et al., 2019). Microsoft's LGBM Framework enables the efficient training of various tree-based models. Three tree-based models were trained using this framework, including a traditional Gradient Boosting Decision Tree (GBDT), a Dropouts Multiple Additive Regression Tree (DART), and a Gradient-based One-Side Sampling tree (GOSS). GBDT uses an ensemble of decision trees and leverages a gradient descent-based procedure while adding trees to the ensemble (Friedman, 2001). This model has been shown to perform well for a variety of tasks, including those with relatively small amounts of training data (Touzani et al., 2018; Ben Taieb & Hyndman, 2014). DART uses dropout while training a boosted decision tree to reduce the overfitting caused by adding trees in the late part of the training process that tend to overspecialize to a small amount of training examples (Rashmi & Gilad-Bachrach, 2015). GOSS uses a sampling technique to keep instances in the dataset that have a large gradient and randomly selects examples that have small gradients to significantly decrease the amount of time it takes to train a model while retaining high accuracy (Machado et

al., 2019). The features used for each of these models were the set of features with label-encoded categorical variables.

These models have a plethora of hyperparameters, therefore, to tune the hyperparameters I used Bayesian optimization. The best hypermeters were selected as those that minimized the average cross-validated mean-squared error and brier loss on the training set of the continuous and categorical outcomes, respectively. Repeated cross validation and repeated stratified cross validation - with 3 folds and 2 repeats - were used during Bayesian optimization for the continuous and categorical outcomes, respectively. Stratified cross validation was used to preserve the ratio of classes for each fold, which helps preserve the probabilities of each class. I used 20 iterations of Bayesian optimization to determine suitable hyperparameters of each model. Early stopping was used to end the training process once the performance of the model stopped improving for more than five consecutive epochs on the validation set to reduce overfitting.

The distribution of classes for the categorical outcome variables were highly imbalanced. Therefore, I tried training two sets of models for each of these three trees, one with class weights that increased the weight of the minority class, and one without weights. After training the weighted models, it was clear that the models with class weights could not compete with the models that did not use class weights. This was because the class weights biased the output probability of the models towards the minority class, resulting in poor performance in terms of brier score. I tried using calibration techniques to correct the biased probabilities, but none of the weighted models performed as well as the model without weights. It was also clear while reviewing literature that few papers used class balancing techniques for the categorical outcomes. It was therefore clear that class weights and oversampling techniques would not improve performance for my models in terms of brier score, so I decided not to use any class balancing techniques for the rest of my models. A comparison of the performance of my tree-based models using class weights, and no class weights can be found in the appendix.

The next type of model that was trained was an elastic net. Elastic net is a linear model that uses both L1 and L2 regularization to reduce overfitting. The categorical features of this model were one-hot-encoded. The continuous features were standardized with a mean of 0 and a variance of 1. The continuous outcomes were also standardized with zero mean, and unit variance. This type of model has two relevant hyperparameters that affect performance, both are related to the model's L1 and L2 regularization. Due to the small number of hyperparameters, the optimal hyperparameters were determined using grid search. The hyperparameters were chosen as those that minimized the error during cross validation on the training set. Repeated cross validation and repeated stratified cross validation were used on the continuous and categorical variables, respectively.

The final type of model that was tested was a neural network with one hidden layer. The model used the set of features with one-hot-encoded categorical predictors. The continuous features were normalized so that the values of each feature in the training set was between 0 and 1. For continuous outcomes, the final layer of the neural network was a single neuron with a linear activation function, and for categorical outcomes it was a single neuron with a sigmoid activation

function. The model's loss function for continuous outcomes was the mean squared error, and for categorical outcomes it was categorical cross entropy. The models were trained for a maximum of 500 epochs, with early stopping enabled if the validation accuracy did not decrease for more than 20 consecutive epochs. Adam was used as the optimizer for training these models. The hyperparameters that needed to be tuned were the model's learning rate, the number of hidden neurons, and the dropout of the hidden layer.

Various options were considered for tuning the hyperparameters. The Keras neural network tuner does not use cross validation, as it determines hyperparameters that minimize the error on the validation dataset with no option for performing cross validation on the training set. Also, it appears the random search and Bayesian optimization libraries were not compatible with Keras objects. Therefore, the best available option for hyperparameter tuning was grid search. Due to the long run times involved with grid search, the activation function of the hidden layer was set to a sigmoid function. No other activation functions were tested for the hidden layer, which resulted in significantly faster run times. The best hyperparameters were obtained as those that minimized the repeated cross validation MSE for continuous outcomes and the repeated stratified cross validation brier loss for categorical outcomes.

All models were compared to a baseline. For continuous outcomes, the baseline was a model that always predicted the average outcome value of the training set. For categorical outcomes, the baseline always predicted the training dataset's average probability of belonging to the minority class.

After training all the models and evaluating their performance on the leaderboard and test set, I used SHAP values to interpret the best performing models. Two shapely plots were used to interpret the best model of each outcome, the "SHAP feature importance plot" and the "SHAP bee swarm plot". The SHAP feature importance plot is well suited to determine the global importance of each feature for making predictions but does not give information related to if the feature positively or negatively affects the outcome. The SHAP bee swarm plot enables us to visualize the effect of individual feature values on the output, so it is well suited to address the shortcoming of the SHAP feature importance plot.

3. Results

Tables 1 and 2 show the performance of each model for predicting the six outcomes on the leaderboard and test sets, respectively. The small size of the training dataset meant that models were particularly prone to overfitting. The issue of overfitting was much more significant for the neural network models than the tree-based and elastic net models. This was due to the large number of parameters of the neural networks, where the large number of features meant that even a relatively small number of hidden neurons meant that millions of parameters needed to be learned. To mitigate this issue, very high dropout values were included in the hyperparameter search space. It can be seen in Table 3 that the neural network models with optimal hyperparameters tend not to overfit due to high values of dropout, except for the material hardship outcome. Interestingly, the model performance on the leaderboard and test set are still

competitive with other models, so it seems overfitting was not such a significant issue for this outcome. The tree-based and elastic net models did not appear to suffer from significant overfitting. A comparison between the training and validation accuracies of each model (to measure the degree of overfitting) can be found in the appendix.

Model / **Outcome**	Baseline	LGBM (GBDT)	LGBM (DART)	LGBM (GOSS)	Elastic Net	Neural Network
GPA (MSE)	0.3927	0.3776	0.3812	**0.3711**	0.3891	0.3877
Grit (MSE)	0.2200	0.2171	0.2169	**0.2166**	0.2215	0.2205
Material Hardship (MSE)	0.0288	0.0261	**0.0256**	0.0261	0.0277	0.0262
Eviction (Brier Score)	0.0534	**0.0506**	0.0521	0.0519	0.0523	0.0520
Layoff (Brier Score)	0.1744	0.1742	0.1744	0.1744	0.1749	**0.1732**
Job Training (Brier Score)	0.2022	**0.1994**	0.2043	0.2033	0.2019	0.2012

Table 1: Performance of each model on the leaderboard set. The best performing model associated with each outcome is bolded.

Model / **Outcome**	Baseline	LGBM (GBDT)	LGBM (DART)	LGBM (GOSS)	Elastic Net	Neural Network
GPA (MSE)	0.4251	**0.3517**	0.3567	0.3720	0.3633	0.3583
Grit (MSE)	0.2530	0.2446	**0.2440**	0.2486	0.2493	0.2532
Material Hardship (MSE)	0.0249	0.0211	**0.0197**	0.0211	0.0233	0.0213
Eviction (Brier Score)	0.0555	0.0537	**0.0533**	0.0538	0.0534	0.0534
Layoff (Brier Score)	0.1672	0.1671	0.1669	**0.1664**	0.1669	0.1665
Job Training (Brier Score)	0.1853	0.1761	**0.1757**	0.1765	0.1783	0.1801

Table 2: Performance of each model on the test set. The best performing model associated with each outcome is bolded.

	MSE or Brier Loss on the Training Set	MSE or Brier Loss on the Validation Set	Optimal Dropout Rate from Grid Search
GPA	0.2266	0.3704	0.75
Grit	0.2412	0.212	0.85
Material Hardship	0.0043	0.0179	0.75
Eviction	0.0428	0.047	0.90
Layoff	0.1339	0.1519	0.95
Job Training	0.101	0.1663	0.80

Table 3: Accuracies on training and validation set for neural network models, as well as dropout values used by each model.

The model performance on the test set is the most important metric of success in literature, as it is normally withheld from participants until the very end of the competition. Therefore, I interpreted the models that performed best on the test set. The SHAP plots are provided in the appendix since they take up a significant amount of space. Figures A1 to A12 show the SHAP feature importance plot and SHAP bee swarm plots of the top 7 predictors of each model. The feature importance plot can be interpreted as the features that on average most affected the outcome predictions on the test set. The bee swarm plot can be interpreted as the contribution of each individual observation towards the shapley values of the feature importance plot. In this plot, the color of the observations allows us to determine if larger or smaller values of the feature contribute most towards changing the outcome.

4. Discussion

It is clear from Tables 1 and 2 that it is very challenging to predict the life outcomes of children and families, since my best performing models only slightly outperform the baseline. Salganik et al. (2019) explains that even the best models in existing literature only perform slightly better than the baseline, likely due to the complexity of the task. Therefore, the performance of my models appears to be sensible. It also appears that my best performing models are tree-based, where the DART model performs particularly well for most outcomes. The neural network model is the best performing model for predicting layoff on the leaderboard set, but the DART model outperforms the neural network for this metric on the test set. As mentioned earlier, overfitting was a serious issue with this dataset, therefore, the strong performance of the DART-based model can be attributed to its ability to limit the effects of overfitting by using dropout.

Figures A1 and A2 highlight that the most important features for predicting GPA are related to standardized test scores, where better performance during these tests appears to increase the model's predicted GPA. The child's household income also seems to strongly influence predictions for this outcome. Betts and Morell (1999) determined that family income affects a

child's academic performance, so this result aligns with existing literature. Figures A3 and A4 show that some of the most important features for predicting the child's grit at age 15 are related to the child's self-reported grit when they were 9, (such as if the child reported following work through to the end during the previous survey). Another important feature is whether the child takes vitamins. It has been shown that vitamin deficiencies can negatively affect energy levels and the fatigue of children (Tardy et al., 2020), so it makes sense that vitamins could affect the perseverance of children by increasing energy levels for those that take vitamins. Other features appear to be related to the child's academic performance, such as the child's standardized test scores and whether the child usually completed their homework. These features are intuitively related to grit, since performing well at school involves following work through until it is completed, even if the child would rather perform other activities.

Figures A5 and A6 illustrate that the most important features for predicting material hardship when the child turned 15 are related to if the primary caregiver experienced some form of material hardship when the child was 9. This result shows that fragile families are more likely to experience extreme poverty if they experienced it before. Another important predictor for material hardship is the primary caregiver's satisfaction with their lives. It has been shown that extreme poverty affects people's satisfaction with their lives (Samman & Santos, 2013), so this result aligns with previous studies. It can be seen from Figures A7 and A8 that the most important features for predicting the caregiver's risk of eviction are similar to those of material hardship (Questions related to if the child experienced poverty in the past). Therefore, it appears that the material hardship and eviction outcomes are highly linked, which intuitively makes sense, since poorer families are at higher risk of not having enough money to pay rent. Another important feature appears to be the amount of money the primary caregiver spends eating out. This result also makes sense, since poor families (those most at risk of eviction) tend to eat more fast food than middle class or rich families (Wise, 2018).

Figures A9 and A10 highlight that the layoff outcome shares many of the same important features as material hardship and the risk of eviction. These are features related to poverty, which makes intuitive sense, since people without jobs are more likely to experience poverty due to their loss of income. Moreover, it has been shown that poor people are at increased risk of losing their jobs during economic downturns (Perry, 2020). Another important feature is related to if the caregiver works overtime or has a second job. It appears that caregivers that work more overtime are more likely to be laid off (since a higher value for this feature in Figure A10 means that a person did not work overtime, and a lower value means that they did). There does not seem to be much literature on the topic of how working overtime or multiple jobs are related to increased risk of being laid off. However, it appears that this feature might be related to poverty, since poorer households need to work more hours due to low-paying jobs. Therefore, this feature appears to be important due to a spurious correlation with poverty.

It can be seen from Figures A11 and A12 that the most important features for predicting if the caregiver received job training when their child turned 15 are related to if the caregiver received job training in the past. Additionally, it appears that parents with intellectually gifted children are more likely to take part in job training. It also appears that transformative religious experiences

may play a part in caregivers seeking job training, which makes intuitive sense since transformative life experiences can also change people's career paths, often requiring additional training (Brown, 1997). Overall, it appears my best performing models are leveraging features that make intuitive sense for predicting each of the six life outcomes of the Fragile Families Challenge.

5. Limitations

My study has various limitations that affect both the generalizability of the findings, and the performance of my models. To begin, I used very simple imputation techniques, so more sophisticated techniques may have improved my model performance. Stanescu et al. (2019) showed that Amelia imputation performed well for predicting each of the six outcomes, so future work could leverage a similar imputation technique. Another limitation that likely affected my model performance was how I dealt with non-continuous features. I treated all non-continuous data as categorical data; However, some non-continuous features have clear ordering, so ordinal encodings would have been more appropriate. The model of Rigobon et al. (2019) achieved state of the art performance by correctly identifying the ordinal variables. However, they used a complex procedure to identify the ordinal variables, and even went through the features one at a time to manually determine some of the variables that were difficult to identify algorithmically. Therefore, due to time constraints, it was impractical to determine these variables. An additional limitation of my study is that I did not create an ensemble model by combining the predictions of each of my five models. This ensemble-based model likely would have improve the performance of my models but was beyond the scope of my study.

In terms of the generalizability of my findings, it is important to note that all data from the Fragile Families Challenge was collected from families in the US, so may not hold in different countries. Moreover, it only examines children born during a specific year, so may not generalize to future generations of children. It is also important to mention that SHAP values help explain the behavior of my model, but do *not* provide any insights related to causality. Therefore, no causal claims can be made about the relationships between the predictors and outcomes in my study. Additionally, the performance of my models was only slightly better than a simple baseline. Therefore, I acknowledge that I interpreted models that provide little predictive power in terms of predicting the life outcomes of fragile families. Salganik et al. (2019) mentioned that this is one of the shortcomings of this dataset, and that policy makers should not make policy decisions based on the results of this challenge alone (Salganik et al., 2019).

6. Conclusion

In this study, I trained five models to predict the six outcomes of the Fragile Families Challenge. Simple imputation techniques were used during the data cleaning process and no feature selection was necessary due to the nature of my chosen models. The best performing models were tree-based, and all the best performing models outperformed a simple baseline model. I

interpreted my best performing model for each outcome using SHAP. Although no causal relationship could be determined from the SHAP values, they enabled me to determine which features most affected the outcome predictions of my models. It was evident that my models were making predictions based on features that were intuitively related to the outcome variables. Despite my study's interesting results, it is clear that additional research needs to be conducted to better understand the factors that affect the life outcomes of fragile families.

References

Brooks-Gunn, J., Garfinkel, I., McLanahan, S. S., & Paxson, C. (2011). Fragile Families and Child Wellbeing Study [Public Use Data]. *ICPSR Data Holdings*. https://doi.org/10.3886/icpsr31622.v1

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

Lundberg, I. (2017a). *Fragile Families Challenge Blog - GPA*. Fragile Families Challenge. https://www.fragilefamilieschallenge.org/gpa/.

Lundberg, I. (2017b). *Fragile Families Challenge Blog - Grit*. Fragile Families Challenge. https://www.fragilefamilieschallenge.org/grit/.

Lundberg, I. (2017c). *Fragile Families Challenge Blog – Material Hardship*. Fragile Families Challenge. https://www.fragilefamilieschallenge.org/ material-hardship /.

Lundberg, I. (2017d). *Fragile Families Challenge Blog - Eviction*. Fragile Families Challenge. https://www.fragilefamilieschallenge.org/eviction/.

Lundberg, I. (2017e). *Fragile Families Challenge Blog - Layoff*. Fragile Families Challenge. https://www.fragilefamilieschallenge.org/layoff/.

Lundberg, I. (2017f). *Fragile Families Challenge Blog – Job Training*. Fragile Families Challenge. https://www.fragilefamilieschallenge.org/ job-training /.

Compton, R. (2019). A Data-Driven Approach to the Fragile Families Challenge: Prediction through Principal-Components Analysis and Random Forests. *Socius: Sociological Research for a Dynamic World*, *5*, 237802311881872. https://doi.org/10.1177/2378023118818720

Stanescu, D., Wang, E., & Yamauchi, S. (2019). Using LASSO to Assist Imputation and Predict Child Well-being. *Socius: Sociological Research for a Dynamic World*, *5*, 237802311881462. https://doi.org/10.1177/2378023118814623

Rigobon, D. E., Jahani, E., Suhara, Y., AlGhoneim, K., Alghunaim, A., Pentland, A. "S., & Almaatouq, A. (2019). Winning Models for Grade Point Average, Grit, and Layoff in the Fragile Families Challenge. *Socius: Sociological Research for a Dynamic World*, *5*, 237802311882041. https://doi.org/10.1177/2378023118820418

Raes, L. (2019). Predicting GPA at Age 15 in the Fragile Families and Child Wellbeing Study. *Socius: Sociological Research for a Dynamic World*, *5*, 237802311882480. https://doi.org/10.1177/2378023118824803

Garcia, C., & Ta, A. (2018). Fragile Families.

Davidson, T. (2017). Black Box Models and Sociological Explanations: Predicting GPA Using Neural Networks. https://doi.org/10.31235/osf.io/7nsrf

Ahearn, C. E., & Brand, J. E. (2019). Predicting Layoff among Fragile Families. *Socius: Sociological Research for a Dynamic World*, *5*, 237802311880975. https://doi.org/10.1177/2378023118809757

Machado, M. R., Karray, S., & de Sousa, I. T. (2019). LightGBM: an Effective Decision Tree Gradient Boosting Method to Predict Customer Loyalty in the Finance Industry. *2019 14th International Conference on Computer Science & Education (ICCSE)*. https://doi.org/10.1109/iccse.2019.8845529

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Touzani, S., Granderson, J., & Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings*, *158*, 1533–1543. https://doi.org/10.1016/j.enbuild.2017.11.039

Ben Taieb, S., & Hyndman, R. J. (2014). A gradient boosting approach to the Kaggle load forecasting competition. *International Journal of Forecasting*, *30*(2), 382–394. https://doi.org/10.1016/j.ijforecast.2013.07.005

Rashmi, K. V., & Gilad-Bachrach, R. (2015). DART: Dropouts meet Multiple Additive Regression Trees.

Betts, J. R., & Morell, D. (1999). The Determinants of Undergraduate Grade Point Average: The Relative Importance of Family Background, High School Resources, and Peer Group Effects. *The Journal of Human Resources*, *34*(2), 268. https://doi.org/10.2307/146346

Tardy, A.-L., Pouteau, E., Marquez, D., Yilmaz, C., & Scholey, A. (2020). Vitamins and Minerals for Energy, Fatigue and Cognition: A Narrative Review of the Biochemical and Clinical Evidence. *Nutrients*, *12*(1), 228. https://doi.org/10.3390/nu12010228

Samman, E., & Santos, M. E. (2013). Poor and dissatisfied? Income poverty, poverty transitions and life satisfaction in Chile. *Journal of Poverty and Social Justice*, *21*(1), 19–31. https://doi.org/10.1332/175982713x664038

Wise, J. (2018). Britain's deprived areas have five times as many fast food shops as rich areas. *BMJ*. https://doi.org/10.1136/bmj.k4661

Perry, S. (2020). Low income workers at greater risk of unemployment and mounting ill-health as furlough scheme unwinds.

Brown, M. P. (1997). A study of transformative aspects of career change experiences and implications for current models of career development.

Appendix

Model / **Outcome**	GBDT (No class Weights)	GBDT (Weighted)	DART (No Class Weights)	DART (Weighted)	GOSS (No Class weights)	GOSS (Weighted)
Eviction	**0.050633**	0.052569	**0.052073**	0.053157	**0.051928**	0.052061
Layoff	**0.174197**	0.175055	**0.174404**	0.17664	**0.174404**	0.176173
Job Training	**0.19941**	0.211877	**0.204283**	0.211803	**0.203294**	0.204138

Table A1: Comparing the performance of tree-based models on categorical outcomes using class weights and no classs weights. It is clear from this figure that class weights decrease model performance

	MSE or Brier Loss on Training Set	MSE or Brier Loss on Training Set
GPA	0.2903	0.3724
Grit	0.1889	0.2013
Material Hardship	0.0211	0.0168
Eviction	0.0484	0.0454
Layoff	0.1677	0.1545
Job Training	0.1627	0.1607

Table A2: Training and validation accuracies of the GBDT model.

	MSE or Brier Loss on Training Set	MSE or Brier Loss on Training Set
GPA	0.1880	0.3657
Grit	0.1307	0.2096
Material Hardship	0.0154	0.0160
Eviction	0.0364	0.0462
Layoff	0.1456	0.1550
Job Training	0.1168	0.1619

Table A3: Training and validation accuracies of the DART model.

	MSE or Brier Loss on Training Set	MSE or Brier Loss on Training Set
GPA	0.3477	0.3838
Grit	0.2302	0.2092
Material Hardship	0.0193	0.0172
Eviction	0.0514	0.0469
Layoff	0.1456	0.1550
Job Ttraining	0.1393	0.1647

Table A4: Training and validation accuracies of the GOSS model.

	MSE or Brier Loss on Training Set	MSE or Brier Loss on Training Set
GPA	0.3286	0.3633
Grit	0.1535	0.2493
Material Hardship	0.00414	0.0233
Eviction	0.04918	0.0533
Layoff	0.1548	0.1658
Job Ttraining	0.1606	0.1782

Table A5: Training and validation accuracies of the elastic net model.

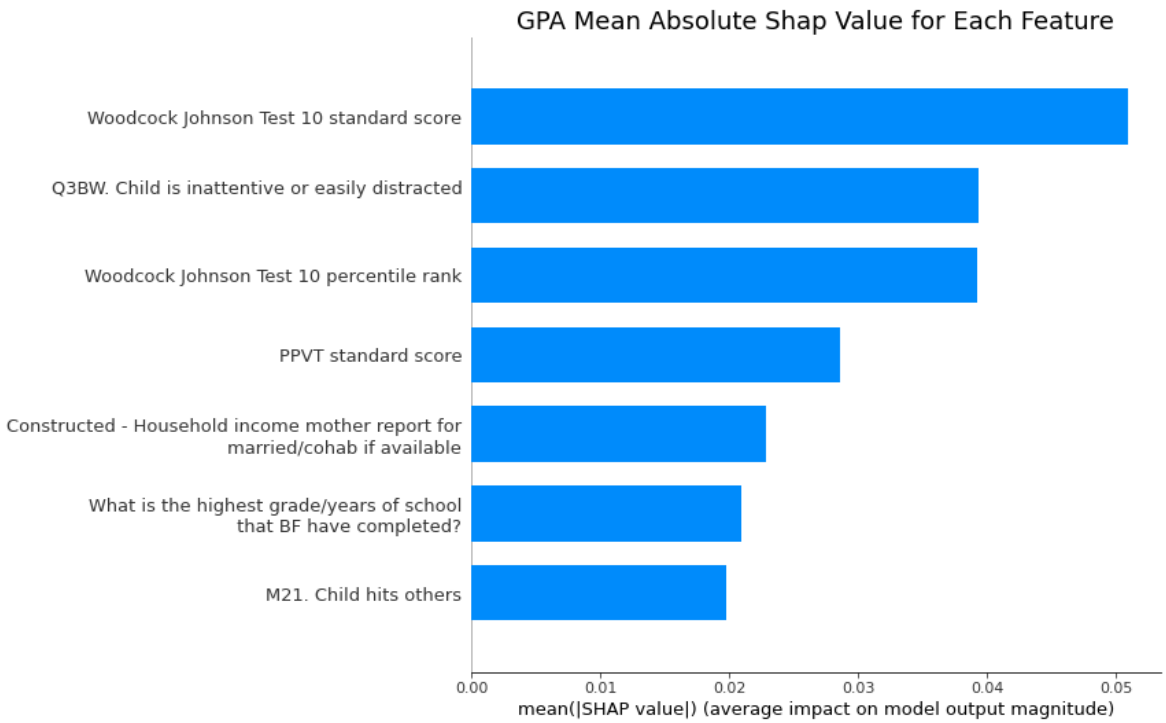

Figure A1: SHAP feature importance plot for GPA.

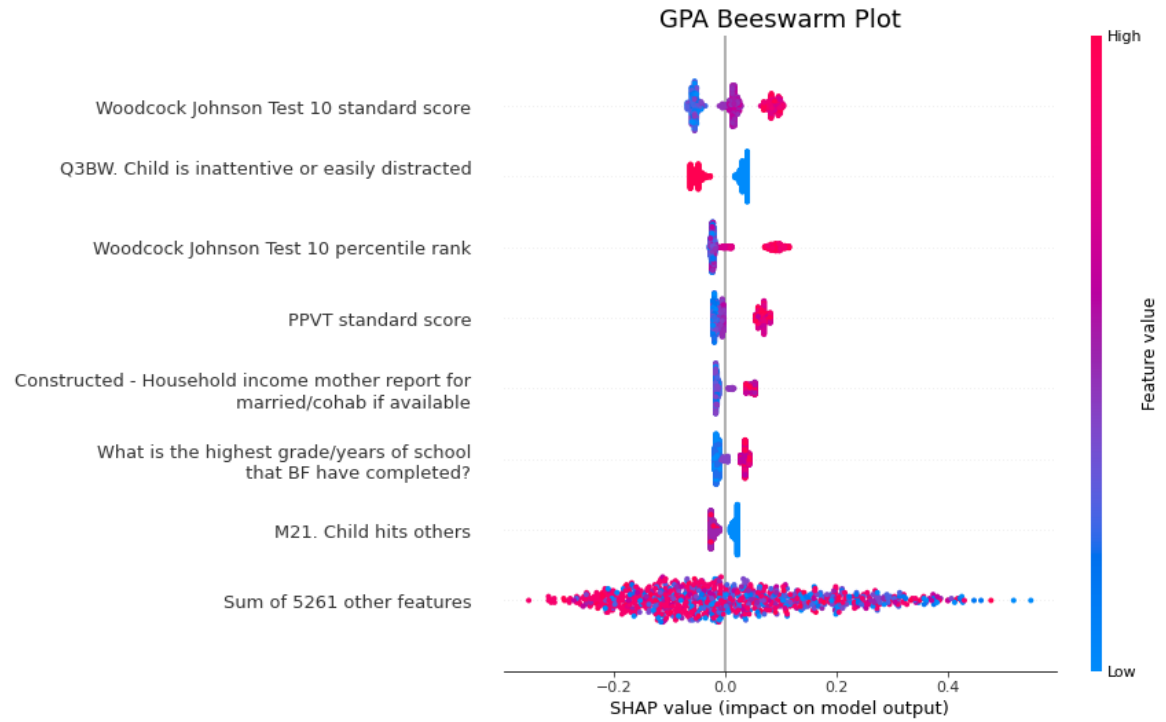

Figure A2: SHAP Bee Swarm plot of GPA

Figure A3: SHAP feature importance plot for Grit.

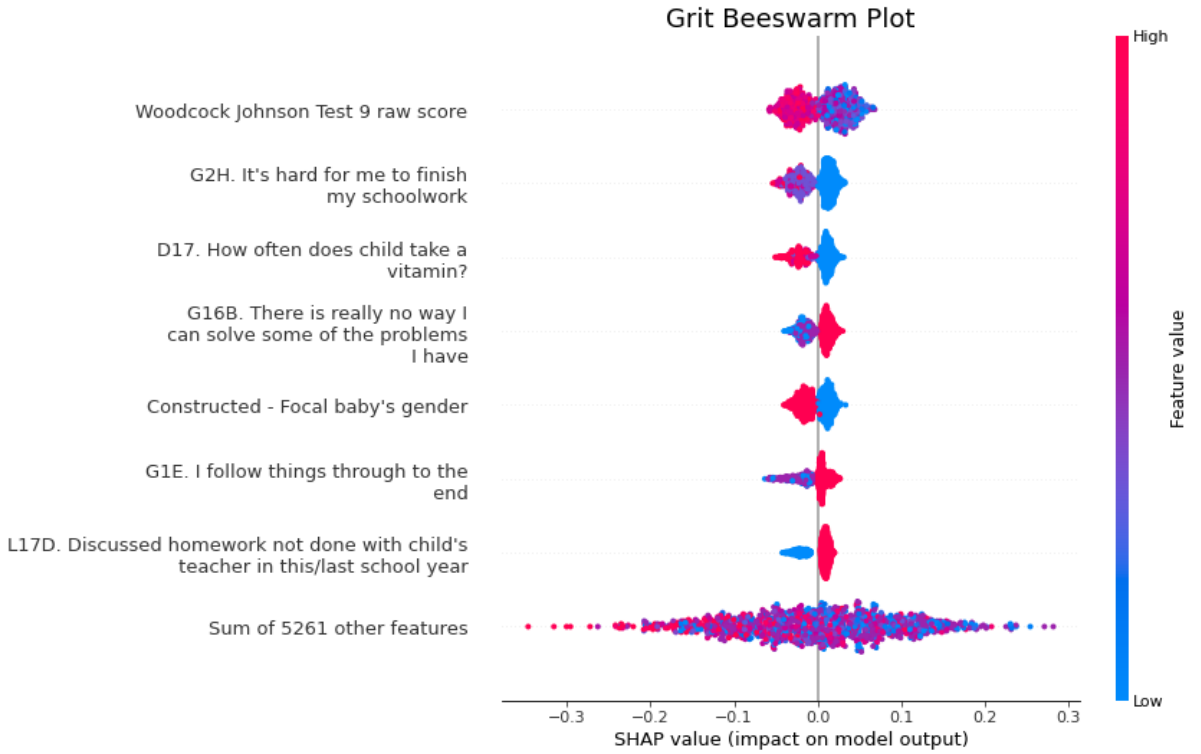

Figure A4: SHAP Bee Swarm plot of grit

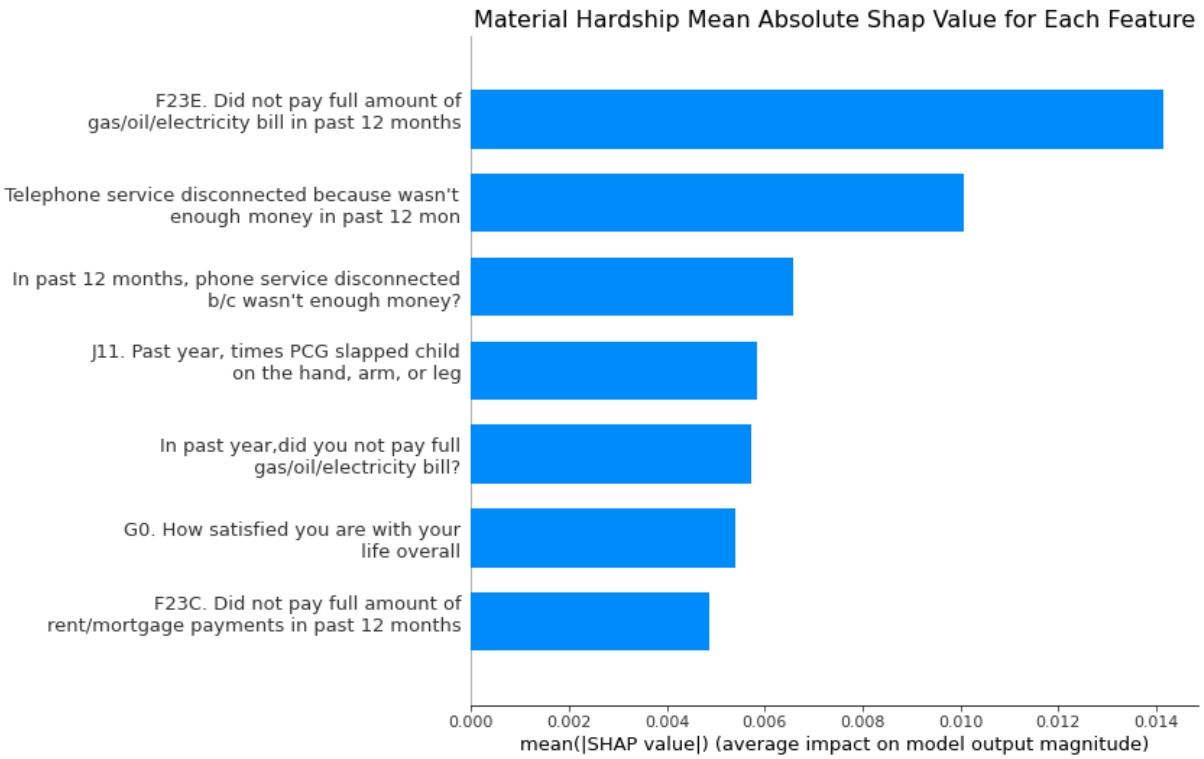

Figure A5: SHAP feature importance plot for material hardship.

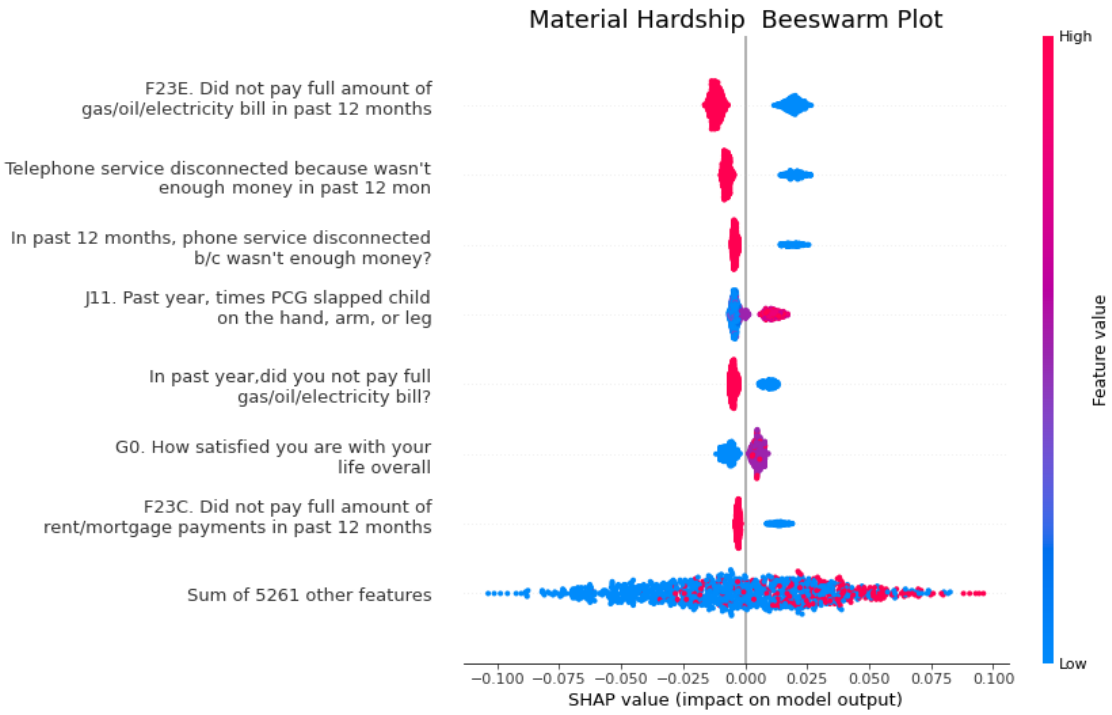

Figure A6: SHAP Bee Swarm plot of material hardship

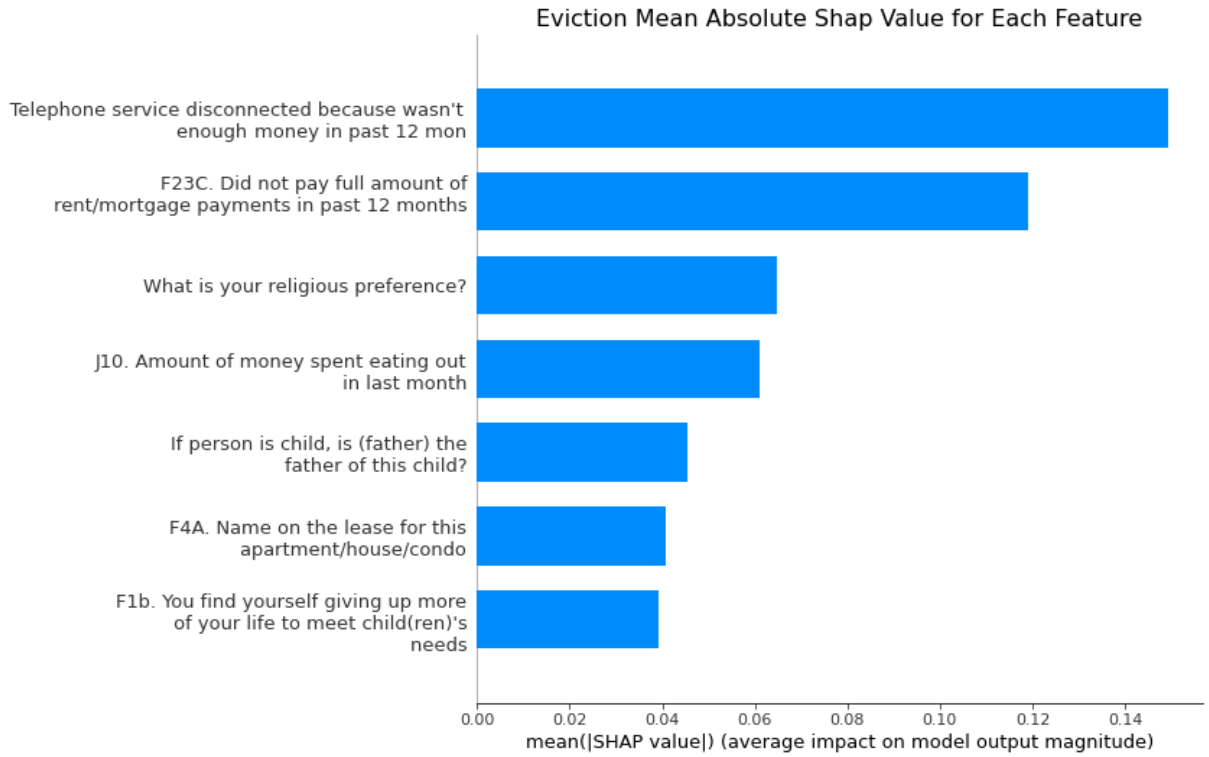

Figure A7: SHAP feature importance plot for eviction.

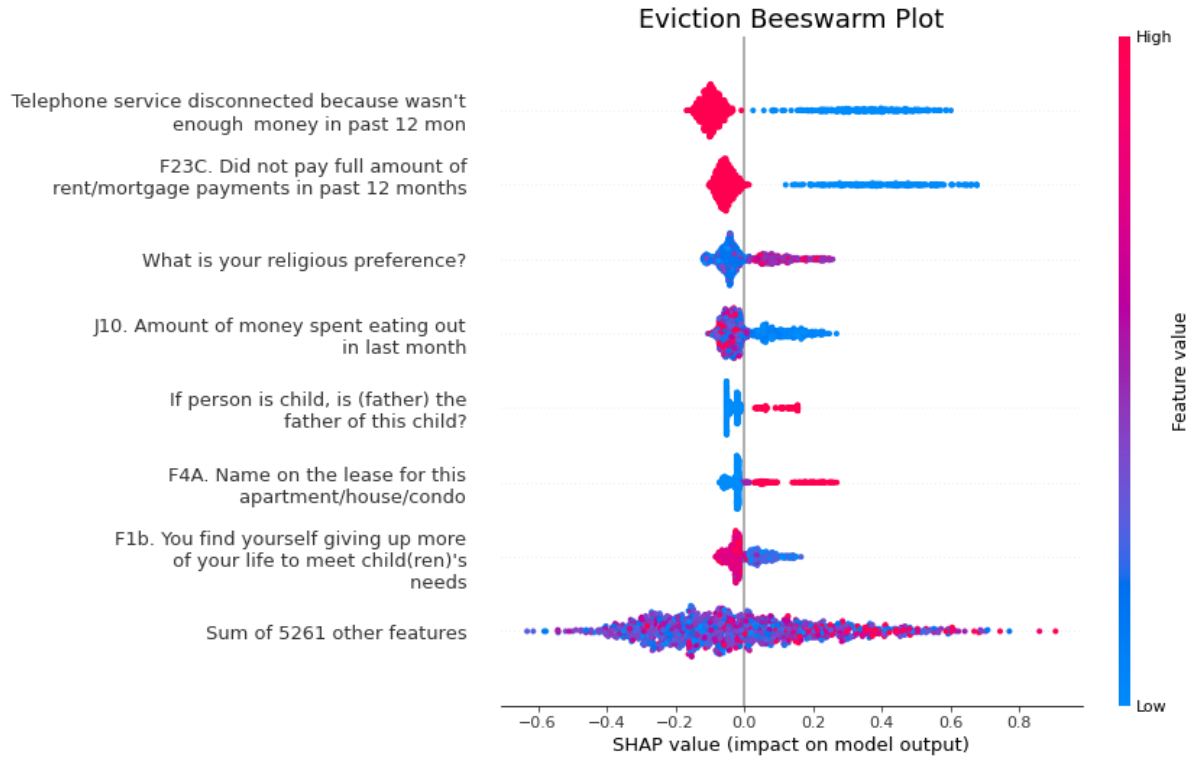

Figure 8: SHAP Bee Swarm plot of eviction

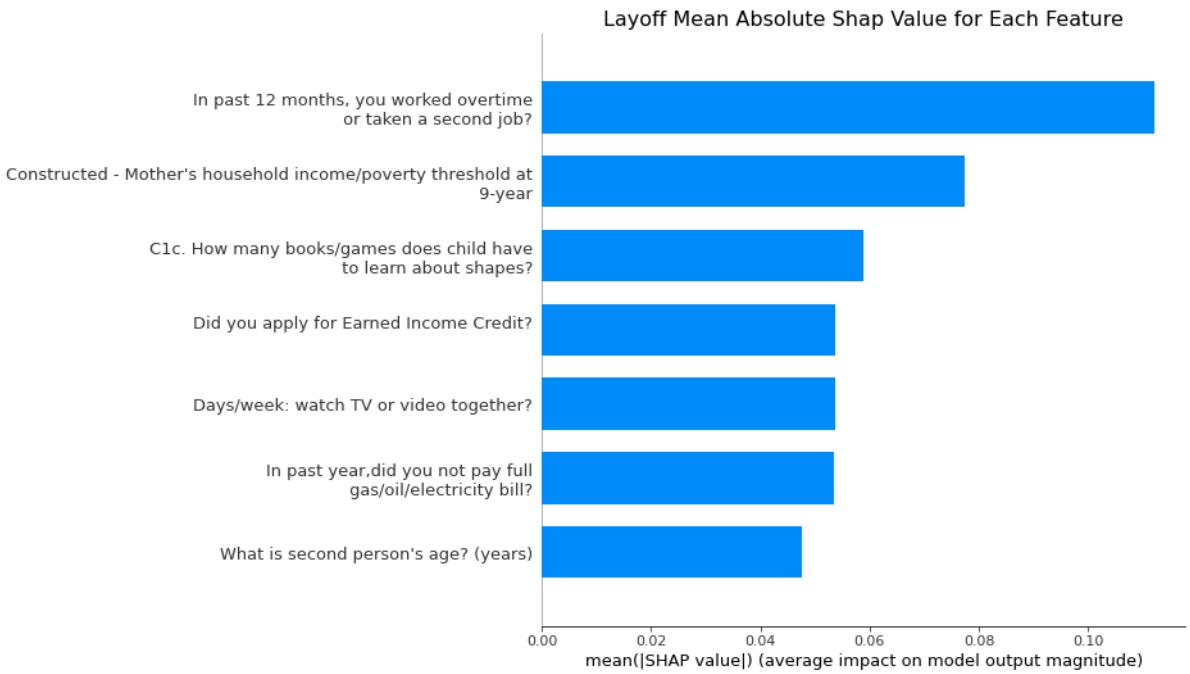

Figure 9: SHAP feature importance plot for layoff.

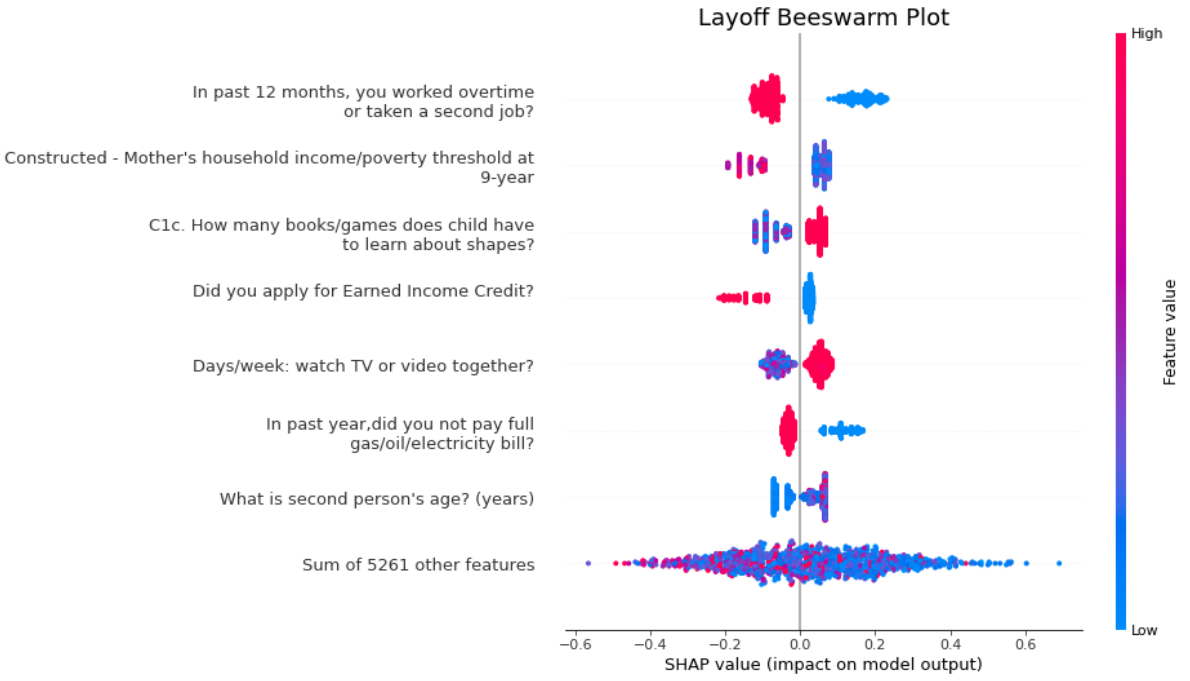

Figure 10: SHAP Bee Swarm plot of layoff

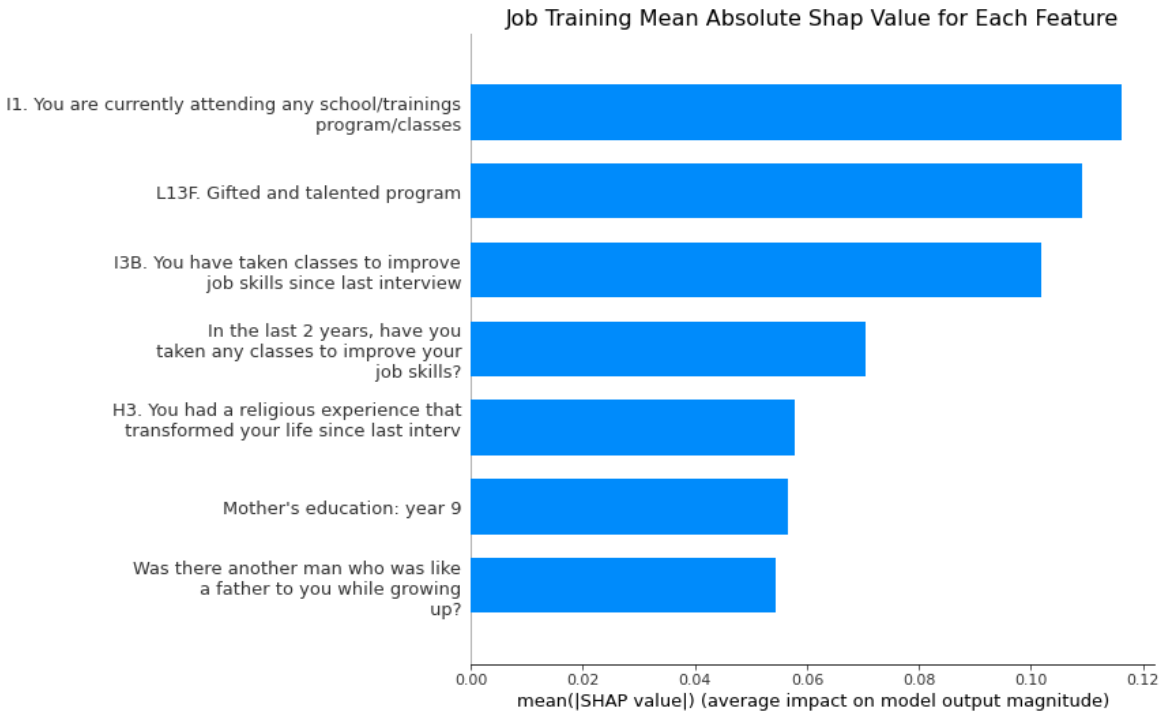

Figure 11: SHAP feature importance plot for job training.

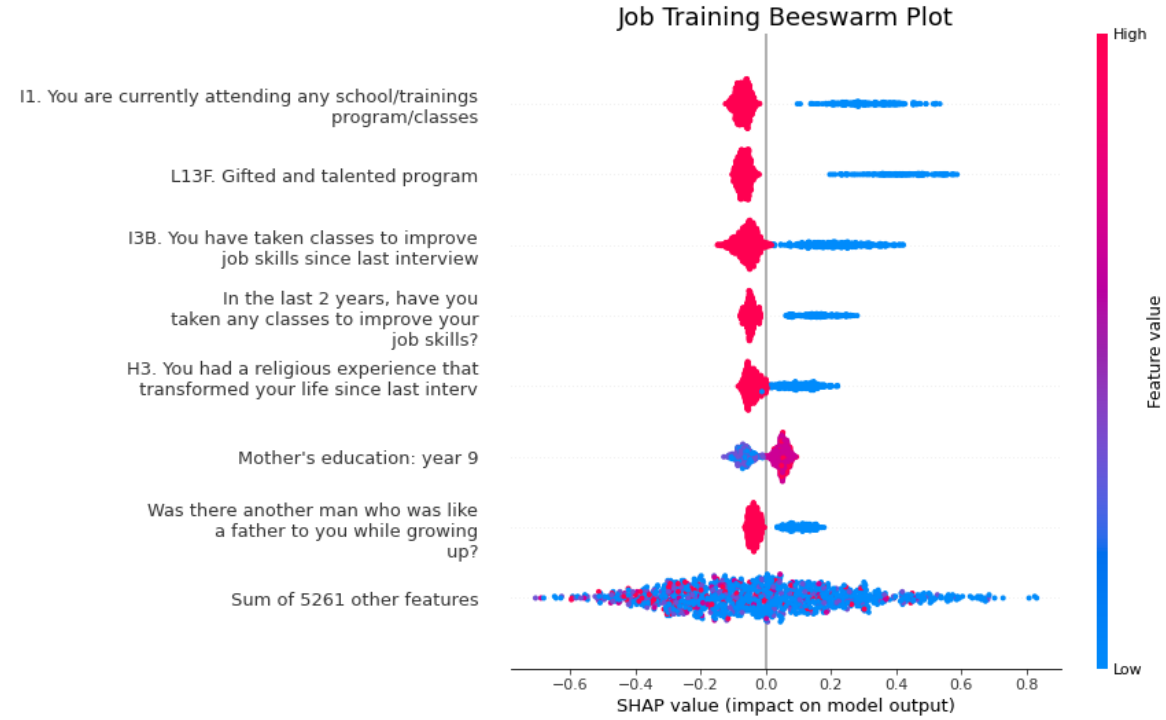

Figure 12: SHAP Bee Swarm plot of job training