Fragile Families Challenge

Nicholas Kim
Princeton University
nk6@princeton.edu

Nicholas Schmeller
Princeton University
nbs@princeton.edu

Abstract

Helping at-risk children in "fragile families" is an essential concern for any future, but our society has limited resources to ensure that these children have stable and successful adult lives. Numerous informational barriers challenge the well-meaning efforts of those who intend to help, but the Fragile Families dataset offers a vast domain of features with strong informative potential. Six key outcome factors - gpa, grit, material hardship, eviction, job loss, and job training - are reported at the end of the study and offer impactful variables to predict and understand. Comprehending what affects these key life variables would be a critical and invaluable asset in structuring initiatives to offer humanitarian assistance/aid. In this assignment, we received the data of 4242 individuals with 16,993 features at years birth, 1, 3, 5, and 9 as well as the final key variable outcomes at year 15. For our study, we focus on predicting "grit", given its fundamental (and arguably foremost) usefulness and importance in helping individuals achieve their life objectives, even amidst significant barriers. We begin by analyzing all relevant mother and father features over the entire timeframe with a variety of classifiers and proceed by examining the individual predicitivity of the mother and father at differing points. We then analyze a tangential question related to the net-desirability of grit, which is its relation to depression. We find that parental correlation to grit is notably weak, with most classifiers failing to produce more accurate results than a simple mean prediction. The Bayesian Regression classifier performed the best in this instance. For the mother and father at specific time points, there is no notable difference in predictivity between any particular time window. Overall, there seems to be difficulty reaching any strongly predictive or notable conclusions with this data-set.

1 Introduction

Socioeconomic inequity is perverted as it is pervasive. Our society spends billions of dollars seeking to right wrongs that have existed for millenia, but social justice is an uphill battle and progress at times seems nonexistent. A rapidly growing area of social science research is the use of heavily quantitative methods to identify the most effective way to spend money to help the disadvantaged. The Fragile Families study was a 15 year study motivated by these reasons that tracked the development and well-being of thousands of children (as well as their parents and teachers) believed to be "at risk" due to poverty, non-marital births, and other related factors. In particular for this paper, we are interested in grit and understanding mother/father consequences.

This project begins by using eight unique classifiers to predict Grit based on mother/father-relevant features. We focus on analyzing the subset of constructed features related directly to the mother and father for two reasons - 1) As a research question, it poses an interesting/complex set of possible interactions 2) It provides very actionable data from a social intervention perspective should a specific pattern be found in the father and/or mother with regards to Grit. We then further analyze with two classifiers which years are the most predictive for the mother and father respectively. Based on our preconceived understandings of family life in the United States, we hypothesize that mothers will have the most impact earlier in a child's life, whereas fathers will have more impact as the child grows. The implication of this hypothesis being true is more efficiently targeted use of resources, i.e. extra support for mothers in early years or for fathers in later years. In order to test this hypothesis, we used several regression classifiers to predict grit at the age of 15 based on different filters of feature selection to isolate the influence of different parents at varying ages. We also examine a tangential question related to grit's correlation with depression.

2 Related Work

Statistical analysis in the social sciences has been a burgeoning research area in recent years. A substantial literature exists to explain various potential interactions and actionable points of change for unfortunate individuals. In particular, there is a healthy interest in parental-related research. One paper analyzes the impacts of fathers taking time off work after a child's birth, finding significantly increased child wellbeing in such cases [4]. Another paper found analyzed the impacts of parental conflict and maternal stress, finding increased child aggression under such exposure [5]. We seek to better understand the critical area of mother/father impacts from both a general predictivity and insight-oriented perspective.

2.1 Data Processing

We started by downloading all the relevant files from the Fragile Families website. These included: background.csv, background.dta, train.csv, prediction.csv, codebook_FFChallenge.txt, leaderboard.csv, leaderboardUnfilled.csv, test.csv, and various data files. In addition, the MissingDataScript.py was also downloaded from the website. The script was modified from its original form, as we noticed some scenarios where the original script created excessive noise in the data. The existing script replaced NA's with the mode, replaced remaining NA's with 1, and replaced negative values with 1. In particular, we noticed that for numerous features, negative values were essentially the equivalent of an NA with regards to usefulness in classification (they frequently referred to missing data, with the specific number clarifying the circumstance of missing data, such as "refusing to answer", "skipping the question", "not in wave", etc.). As a result, there were many variables for which "1" was a very non-representative value to impute. For instance, household income variables such as "cf1hhinc" would have their usual data points in the 20,000-40,000 range, but had negative values imputed as 1! This was sure to add a lot of noise, so we edited the script to simply fill NA and negative values with the median value of the corresponding feature. In observing the "y_test" Grit results, we dropped all challengeIDs which possessed an "NA" value.

Given the large number of features, we decided to limit ourselves to constructed, continuous variables relevant to the father/mother (for which we would manipulate subsets of for exploration). To start, the constructed variables are mentioned as being more meaningful, and given 15,000+ features, we believed there would be a high risk of noise should we try to deviate or directly, manually explore the numerous features. The continuous variables presented digestable data for regression classifiers while also appearing to encompass particularly meaningful metrics relative to other categories (binary/categorical/unordered categorical), including statistics such as household income, parental age, and a poverty ratio. The father/mother specificity corresponds to our central research question of interest of understanding the particular, relative roles of the two. To filter for these features, we used the "Advanced Search" feature of the Fragile Families website, applying the following queries: Focal Person/Respondent as "Mother"+"Father", source as "constructed", and variable type as "continuous".

2.2 Classification Methods

For this classification task, we explored 8 different classification methods from the SciKitLearn Python libraries.[3] Parameters are default unless otherwise noted.

1. *Linear Regression (LR) with OLS penalty*: non standardized
2. *Bayesian Regression (BR)*: 300-iter (max) Ridge, regularized for precision of weights/noise
3. *Support Vector Regression (SVR)*: with rbf kernel and libvsm implementation
4. *Stochastic Gradient Descent (SGD)*: squared loss, l2 regularizer
5. *Decision Tree (DT)*: mean-squared-error and best split criterion
6. *Gradient Boosting (GB)* : least-square loss, 100 estimators
7. *K Nearest Neighbors (KNN)*: 5 nearest neighbors with weighting by distance
8. *Neural Network (NN)*: relu activation and hidden layer size 100

With this being a regression problem, each classifier was trained using variant schemes that could output a continuous float value as opposed to a classification category. In the interest of maintaining consistency and comparative relevance with the preexisting challenge, classifier outcomes were not cross-validated across shuffled instances of the dataset. The training data and holdout data were kept as in the original, explicit form as contained/divided in the challenge.

2.3 Evaluation

We analyze each methods effectiveness using R-squared, mean squared error, mean absolute error, and median absolute error scores. Defining the evaluation terms, we have [3]

R-squared = $1 - \frac{\sum (y.actual - y.predicted)^2}{\sum (y.actual - y.mean)^2}$

Mean squared error = $\frac{1}{n} * \sum (y.actual - y.predicted)^2$

Mean absolute error = $\frac{1}{n} * \sum |y.actual - y.predicted|$

Median absolute error = median of set($absolute\ value\ of\ prediction\ errors$)

Each evaluation term serves a unique insight and applicability. The R-squared value essentially compares how well the prediction does relative to a flat mean prediction for each instance. A value of 1 corresponds to perfect prediction, a value of 0 corresponds to an equivalent of the mean prediction, and negative values correspond to poorer-than-mean prediction (essentially saying there is little point in using such a classifier). MSE is a standard for gauging the error of a regression classifier, with the squaring of errors being useful to strongly penalize outlier classifications. It fits and applies well to most standard distributions of data. Mean absolute error and Median absolute error on the other-hand do not strongly penalize outliers, so they may be of good use in non-normally distributed data or datasets where outliers are frequent and expected to a degree.

3 Methods

3.1 Spotlight Classifier: Support Vector Regression

Support vector regression is an application of the support vector algorithm, which in turn originated from the *generalized portrait* algorithm. This algorithm and the fundamentals for support vector applications were developed by Vladimir Vapnik and Alexey Chervonenkis in 1963, and are the result of three decades of statistical learning theory developed by those authors.[2] The support vector machine (SVM) proper was developed at Bell Labs in the 1990s by Vapnik and other American researchers for the purposes of optical character recognition, but its modern applications are present in a plurality of data science and machine learning projects.

In regression applications, some kind of "universal solution" represented by the solution vector $\mathbf{w}$ is learned from a training dataset $\{(x_1, y_1), ..., (x_l, y_l)\} \subset \mathcal{X}$ which returns a prediction $y \in \mathbb{R}$ for a d-dimensional input vector $\mathbf{x} \in \mathbb{R}^d$. In support vector applications, $\mathbf{w}$ is the minimum argument for a special loss function $J(\mathbf{w})$ which encodes a sparse subset of the initial training data by only learning from training data vectors $\mathbf{x}_i$ which for the evaluating metric

$$\hat{y}_i = \mathbf{w}^T \mathbf{w}_i + w_0$$

return a real-valued prediction y_i that is within a certain margin of error ϵ from a general predetermined value $\hat{y}$, and where w_0 is a constant normalizing term for the training data. There may also be a "slack value" ξ outside of ϵ where vectors which are farther away from the "ideal" margin of error ϵ have less weight in the solution vector $\mathbf{w}$. Vectors which satisfy this constraint are called *support vectors*. This constraint can be visualized in the form of a "tube" around the prediction values of the solution $\mathbf{w}$, shown in Figure 1. There are many loss functions which are acceptable objective functions for SVMs, but any SVM loss function should encode this sparsity to allow only support vectors to influence the solution vector. This quality of SVMs is probabilistically quirky because the sparsity is encoded in the loss function (instead of the easier-to-adjust prior) and because this preselected sparsity does not result in probabilistic outputs, i.e. every output value is treated equally even though some may be more certain than others, which may make predictions difficult to understand.

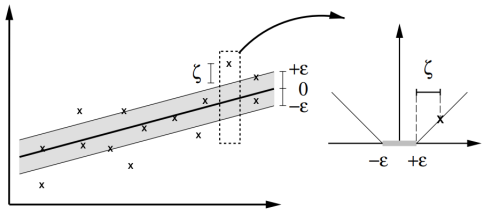


Figure 1: An example ϵ-tube centered around accepted data. x labels denote support vectors with varying weights on the solution vector $\mathbf{w}$, which may be inside the ideal margin of error (shaded region) or in the slack region (distance from shaded region enclosed by dashed line). Sloped lines on the rightmost graph represent increasing α values as the slack value ξ increases. Figure from Smola and Scholkopf 2002.[2]

After loss function optimization, Smola and Scholkopf show that the optimal solution $\hat{\mathbf{w}}$ is

$$\hat{\mathbf{w}} = \sum_i \alpha_i \mathbf{x}_i$$

where $\alpha_i \geq 0$. In this case, α is sparse to eliminate non-support vectors from $\hat{\mathbf{w}}$, aka support vectors are the x_i for which $\alpha_i > 0$. Once $\hat{\mathbf{w}}$ has been learned, a prediction $\hat{y}$ can be made using the function

$$\hat{y}(\mathbf{x}) = \hat{w}_0 + \hat{\mathbf{w}}_T \mathbf{w}$$

An important component of SV analysis is kernelization of the training data and prediction inputs, which allows for more efficient computational complexity of SVM algorithms. SVMs are widely used because of efficiency from the kernel trick, but kernelization hides the explanation for SVM results behind an additional layer of computation and makes them difficult to understand.

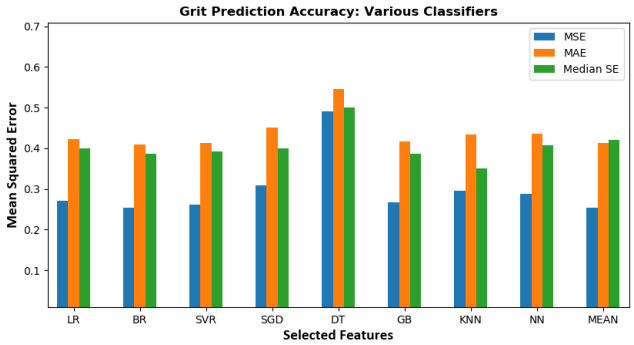
4 Results

4.1 Evaluation Results

(Note: "MEAN" refers to predicting the mean for every instance of the target)

Classifier	LR	BR	SVR	SGD	DT	GB	KNN	NN	MEAN
R^2	-0.071	0.0007	-0.033	-0.220	-0.937	-0.058	-0.169	-0.135	0.0
MSE	0.271	0.253	0.261	0.309	0.490	0.268	0.296	0.287	0.253
MAE	0.422	0.409	0.413	0.451	0.546	0.417	0.433	0.436	0.413
Median SE	0.400	0.386	0.393	0.400	0.500	0.387	0.350	0.407	0.420

The eight classifiers each had their share of unique characteristics and variability in results. For the most part however, MSE appeared somewhat consistent across all the classifiers. Of immediate note was also the poor accuracy increases offered by classifiers over simply predicting the mean, and there was only a single classifier, Bayesian Ridge Regression, that emerged with a positive R-squared value. While these results were initially troubling, upon comparison to the prize-winners for Grit during the challenge, we saw that the top performing teams still only managed an MSE of 2.40, which is only a 5% or so increase over the mean prediction baseline. From this, we say that this dataset is not an overly predictive or informative one, at least in the form it is initially presented in. Almost all of the classifiers had an MSE in the approximately 0.250-0.300 range, except for the decision tree regression, which had by far the highest MSE at 0.490. We hypothesize that this could have something to do with the assumption of interactions among variables, since decision trees split only on a single feature at a time. Given a shortage of particularly meaningful relationships/information in the features, such a process is likely to add more noise than it removes by being forced to sequentially split data on unrelated features.

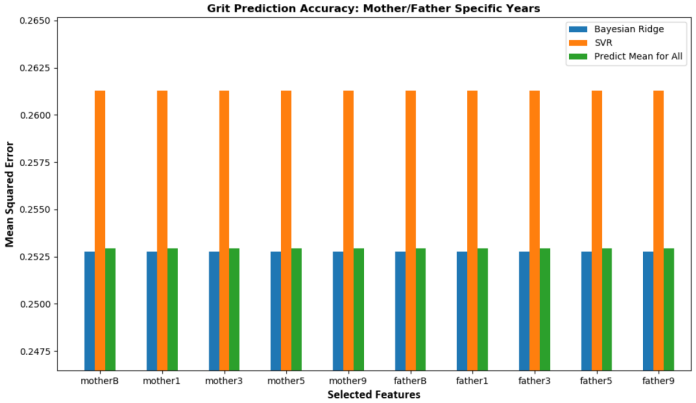


The poor relative performance of most classifiers relying on complex model bases would seem to indicate impactful, high-level, and structurally robust information in the dataset. The strong performance of Bayesian Ridge Regression is interesting in the sense that the strengths of BR in application here are not readily apparent. Its principal advantages include the ability to include a prior (which was not done) and in situations of limited data, which may imply the fragile families dataset is in some respects "limited" with regards to the 6 labels. In analyzing the best metric for this task, it would seem that mean squared error is the most informative, given that there is nothing particularly outlying or abnormal regarding the range/distribution of Grit values. The values for Grit are all contained in the relatively small range of [0,4], and values are not overly clustered in one particular region nor are they plagued by unusually high numbers of outliers. As such, mean-absolute-error, which thrives in such scenarios, is not of special importance here. Median-absolute-error, whose strength is also in handling outlier-heavy datasets, also loses its primary utility. R-squared, while not directly comparable to the error metrics, usefully illustrates in this case that the majority of classifiers struggle to produce meaningful data, seeing as the negative R-squared values correspond to being no better than a simple mean prediction for every outcome. With MSE as the chosen primary metric, we then consider **Bayesian Regression** as the most useful classifier for the task.

Interestingly, using LR and only the suggested features from the Fragile Families "quick-start" page (cm1ethrace, cf1ethrace, cm1edu, cf1edu, and cm1relf), the evaluation metrics are R2 = -0.003, MSE = 0.254. This is in fact, better than how almost all our classifiers did with numerous mother/father features. This would imply a general lack of predictiveness/high amount of noise in most features.

4.2 Mother Father Dynamics

In analyzing the effect of mother/father characteristics on grit, we wanted to better understand the different ways the mother/father characteristics at different years predicted grit. We hypothesized a variety of possible scenarios, such as one where a mother was most strongly predictive in the younger years (from biological roles/social closeness) and the father becoming gradually more predictive over the time. The results however, did not corroborate such theories and are as follows -



Testing with Bayesian Regression and SVR, the spotlight classifier, we can clearly see that the data demonstrates no notable difference in the predictive powers of any of the years for either the mother or father in this dataset. We also see that BR continues to outperform other classifiers even in other contexts.

4.3 Grit and Depression

Grit can be thought of as a mental strength to adversity. An abundance/gravity of adverse experiences is, anecdotally, what leads to the creation of such "mental strength" in individuals. We hypothesize that the presence of depression in a family may increase the likelihood of grit in a child, since they undergo accelerated maturation and resilience development. A correlation, if found, would be interesting to approach from an intervention perspective, since then activists (who from a simple level are interested in maximizing "desirable" qualities) must be wary of the balance between the positivity of grit itself and the negativity of factors that may generate it. We created a feature to indicate the presence of depression in the mother and father, ranging from 0-4, and correlated it with grit

Grit-Depression correlation score: **-0.0625** (p-value 0.00177)

Interestingly, it seems that the opposite of the hypothesis is true, where Grit is more common in less-depressed families. This would advocate a positive-reinforcement oriented approach and placates fears that a single-minded "grit improving" approach might occur at the expense of overall wellbeing with regards to depression. An additional note is that the website documentation of a few features was wrong (like for cf5md_case_lib, 0 = no, 1 = yes, though the website said 1 = yes, 2 = no. This was evident by observing the data manually, since no 2s were present but many 0s were), which brings up general questions regarding reliability of data.

5 Discussion and Conclusion

In this work, we used eight different regression classifiers to attempt to understand the influence mothers and fathers have at different stages of an at-risk child's early years. While we discerned signal in some isolated cases, we were ultimately unable to support our hypothesis or emerge with any strongly predictive classifiers. Out of the eight tested, Bayesian Regression had the lowest MSE, which may be a result of the sparseness/uncorrelation of the data in relation to the intended predictors - a scenario that bayesian probablistic approaches may fare better than others in. Since our classifiers as a group were generally less accurate than simply predicting the mean, we can draw either or both of two conclusions: that different parents do not significantly influence development, or that the Fragile Families dataset is too sparse in its original form to discern signal (without significant transformations) that would validate our hypothesis. While one almost certainly expects there to be parental influences on development, it was not strong enough to show through the noise present in the imputed dataset.

We also experimented with isolating features of the father and mother at years birth, 1, 3, 5, and 9 in the hopes of observing differences in predicitivity at particular years. However, the results showed an equal MSE for all years, leaving the data unable to validate our previous hypothesis on the topic (strong initial mother predictivity, growing father predictivity over time). While the results speak for what they are, we also believe that the extreme equality of predictivity at all years points to an aspect of the dataset itself moreso than many other potential factors (accounting for general lack of predicitivity in other areas and even in the prize winners, which were often scarcely more predictive than the mean, or r-squared = 0).

Given these conclusions, there are several remaining directions to go toward to exploring the the data, its information, and our hypothesis. One is to improve the imputation script beyond what we already did, being careful not to inject artificial data from our own biases and more carefully formatting categorical/binary distinctions. Incorporating more diverse features outside mother and father surveys could also potentially unveil a stronger source of signal hidden in the dataset. chi2 feature selection could be explored, as well as transformations on features and the creation of interaction terms. In general, the vast number of features would appear to be a barrier to qualitative feature selection, so perhaps algorithmic transformations/feature relations could be run to generate larger improvements upon the currently lackluster MSE and R^2 values.

References

[1] Salganik. Prize winners. Fragile Families Challenge. 15 September 2017. http://www.fragilefamilieschallenge.org/prize-winners/

[2] Smola and Scholkopf. A Tutorial on Support Vector Regression. NeuroCOLT Technical Report. 30 September 2003. https://alex.smola.org/papers/2003/SmoSch03b.pdf

[3] Pedregosa,Varoquaux, Gramfort, Michel, Thirion, Grisel, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 12 (2001) 2825-2830.

[4] Richard J Petts, Chris Knoester, Are Parental Relationships Improved if Fathers Take Time Off of Work After the Birth of a Child?, Social Forces, March 2019 , soz014, https://doi.org/10.1093/sf/soz014

[5] Xiafei Wang, Qiong Wu, Susan Yoon. Scikit-learn: Pathways from Father Engagement during Infancy to Child Aggression in Late Childhood. Child Psychiatry and Human Development. February 2019