No Fragile Family Left Behind - Targeted Indicators of Academic Performance

Anahit Sargsyan¹, Areg Karapetyan¹, Wei Lee Woon¹ and Aamena Alshamsi¹,² *

¹ Department of Computer Science, Masdar Institute (Khalifa University), Abu Dhabi, United Arab Emirates
² The MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA

Abstract

Academic performance is a key component in the development and subsequent empowerment of youth. It is affected by a large number of different factors, such as inherent ability and socioeconomic circumstance, which can vary widely amongst different individuals. In particular, children from disadvantaged families face unique challenges that do not occur in normal families. We analyze the Fragile Families Challenge (FFC) dataset using data science algorithms, and study the relationship between the features reported and GPA scores. We grouped GPA scores into three groups (top, middle and low) and used a random forest classifier to predict the GPA class for each subject. Then, we used a recently developed algorithm—Local Interpretable Model-Agnostic Explanations (LIME)—to cluster subjects into subgroups based on the factors affecting each individual. We further analyzed the clusters to elucidate the differences occurring within different subgroups in the

*Emails: {anahit.sargsyan, areg.karapetyan, wei.woon, aamena.alshamsi}@ku.ac.ae
community of disadvantaged individuals and figure out which set of features matter for each subgroup. Conventional studies seek factors which apply to all members of a population, and often overlook the unique needs of individuals with different backgrounds and characteristics or divide individuals into predetermined subgroups and study which factors affect each subgroup. The approach used here can find correlations which are specific to individual subjects, and can inform the formulation of targeted and effective intervention strategies. Our study contributes to the field of social science by highlighting the differences of indicators of academic performance in different subgroups. Also, our novel data science pipeline contributes to the fields of data science and computational social science.

1 Introduction

Academic performance can have far-ranging effects on the careers and lives of young people. There have been several studies on the indicators of academic success [1]. Academic performance can be path dependent, so performance at an early age of individuals was effective in predicting college GPA [2, 3], while individual characteristics such as intelligence and determination also play a role [4, 5, 6, 7, 8]. Furthermore, some factors are external and are related to the children’s surroundings; these include social, emotional and socioeconomic factors [9, 10, 11, 12]. Special attention was paid to children from disadvantaged backgrounds, who may be impacted by poverty or by being raised in single parent families [13, 14, 15, 16, 17]. Identifying the indicators of academic performance can also be challenging as many indicators must be captured at a very young age, hence can only be attained through longitudinal studies [18, 10].

There have been many interesting studies on disadvantaged children, but existing approaches tend to focus on global patterns of behavior, which may not fully elucidate the nuanced variations between children from different backgrounds. No two subjects are the same, and factors which deeply affect one child may have a lesser impact on a child encountering different circumstances.
Using the FFC dataset as a case study, this paper develops a novel approach which seeks to address the aforementioned shortcoming. The FFC, organized by Princeton University, is based on Fragile Families and Child Wellbeing study that documents the lives of over 4000 non-marital births occurring between 1998 and 2000 in U.S. cities with at least 200,000 population. The interviews capture important information on attitudes, parenting behavior, demographic characteristics and health (both mental and physical), to name a few. A more detailed description of the FFC is provided in [19].

Our proposed approach is centered on the following three key steps.

1. Standard classification algorithms are trained to distinguish between high and low GPA scores, that allowed us to identify factors associated with academic success, such as test scores, attentiveness and financial stability. These findings are consistent with previous studies and provide a measure of validation.

2. To obtain more detailed insights about the factors affecting specific individuals or groups, we use a novel application of a recently developed technique—Locally Interpretable Model Agnostic Explanations (LIME) [20]—to produce custom “explanations” which can reveal the features that are associated with success for each individual child.

3. LIME algorithm determines localized explanations, so a unique explanation can be generated for each individual presented to the classifier, but this would be too difficult to analyze (since there are over 800 such explanations—one explanation per individual). Hence, we clustered the individuals based on the LIME coefficients - as these coefficients indicates the indicators of academic success in each case, clustering them in this way is intended to group individuals who are similar in terms of their respective success indicators, and not merely who are similar in terms of their childhood experiences or other features. We cluster the subjects based on the respective LIME expla-
nations, and find that the children fall into four main “classes”. Each class is then characterized by the mean of the LIME coefficients for all instances in that class.

While the proposed approach is based on an existing technique, we apply the technique in an entirely novel way by combining it with a clustering algorithm to obtain groups of individuals with similar characteristics or motivations, in the context of academic achievements. We believe that this study contributes to both social science and data science, and resulted in insights that would be difficult to achieve using traditional statistical models.

2 Methodology

This section lays out the data science pipeline developed, which is portrayed as a flowchart in Fig. 1. The process can be divided into 4 general phases: (1) Pre-processing, (2) Feature Selection, (3) GPA Prediction, (4) Explanation of the results through application of the proposed approach. Next, each of these steps will be explained in detail.

2.1 Pre-processing

Due to the nature of the dataset under study, a number of challenges were confronted when analyzing the data. For instance, there were many missing values where respondents either refused to answer or were unavailable to answer certain questions.

These issues were resolved through the judicious use of the following pre-processing techniques, as shown in Fig. 1. First, all missing and negative values were replaced by NaN and the columns with 0 variance were removed. Next, only the columns having at least 400 non-NaN values were retained. Lastly, the variant of $k$NN ($k$-Nearest Neighbors) imputation algorithm [21] implemented in the Python package Fancyimpute was leveraged, with value of 100 for the parameter
\( k \), to estimate the NaN values.

Figure 1: Flowchart of the employed methodology with \( n \) and \( m \) standing for the number of features and entries in the given step, respectively.

* For example, PPVT age equivalent score of ‘4-2’ (4 years, 2 months) is converted to 4.166
2.2 Feature Selection and GPA Prediction

The above steps reduced the dimensionality of the dataset from over 12,900 to 6,610 features. However, this is still a large number given that one of the main aims of this study is to determine the key correlates of GPA scores at later stages of the subject’s life. As such, it was necessary to use feature selection algorithms to exclude redundant features. A wide variety of filter and wrapper based techniques were tested, such as Principal Component Analysis [22], Ridge [23], Lasso [24], Recursive Feature Elimination [25], Gradient Boosting Regression [26] to name a few. Generally, filter based methods use proxy measure (e.g. mutual information) to score a feature subset independent of the learning algorithm, while wrapper based methods use a predictive model to score the subset and the evaluation criteria is the error rate.

However, extensive experiments revealed that no single method produced a satisfactory feature set that maximized the accuracy of GPA prediction. Instead, the best results were obtained using a combination of three feature subsets obtained as follows. Feature importances were estimated using the Extra Trees Regressor algorithm [27] (with 500 estimators) and Randomized Lasso [28], and the top 500 features were retained from each. For the latter, two different values were considered for the regularization parameter $\alpha$, namely 0.004 and 0.000004, thus resulting in two separate feature subsets.

The intersection of these three subsets, containing 69 features, led to improved GPA prediction accuracy. In particular, with the Random Forest algorithm, a mean squared error (MSE) of approximately 0.363 was achieved over the entire dataset allowing these results to be placed in the top quartile of the final FFC scoreboard. This final set of 69 features, tabulated in Table 1 in the Appendix, was then used in all the subsequent analysis.

As such, in the proposed approach the aforementioned prediction subroutine serves as a means of validation for the final feature set selected, thereby solidifying the credibility of the explanations to be derived in the subsequent analysis. In a
sense, this step assesses through MSE the extent of correctness to which the generic performance indicators for the entire dataset were chosen.

3 Results and Discussions

As previously mentioned, this study is concerned primarily with identifying the factors that correlate with the GPA scores attained by the subjects later in their lives. Towards this end, we first framed the problem as a classification problem and discretized the GPA scores into three classes, Low, Middle and Top in the following manner

\[
GPA \in \begin{cases} 
  \text{Low}, & \text{if } 1 \leq \text{GPA} \leq 2.5 \\
  \text{Middle}, & \text{if } 2.5 < \text{GPA} < 3.25 \\
  \text{Top}, & \text{if } 3.25 \leq \text{GPA} \leq 4.
\end{cases}
\]

Accordingly, only the subjects falling into the Top and Low categories were retained. The underlying motivation is to allow the classification algorithms to focus particularly on the factors that clearly distinguish between high and low performers. Indeed, the factors responsible for “borderline” performances are likely to be the ones with the smallest impact and inferring them might add noise to the results. On the other hand, abstracting away a large group of subjects could possibly lead to the loss of pertinent factors. Thus, the thresholds were set according to the top and bottom 30% percentiles of GPA scores. This is to ensure participation of as many subjects in the analysis as possible while retaining a sizable gap between the two classes.

3.1 General Indicators of Success

Here, two predictive models, namely Logistic Regression and CART Decision Tree, were used to determine the factors which broadly correlate with academic
Figure 2: Visualization of decision nodes of CART decision tree trained with all 69 selected features. The color of a node reflects the majority class at the node (blue is high while orange is low), while the intensity of the color reflects the degree of certainty.

(a) First two layers of decision tree

(b) Layer 3 child of the “Child attends to instructions” node from 2a

(c) Layer 3 child of the “Child science and social studies” node from 2a
performance. Both are widely used algorithms suitable for interpretation and visualization. Their application was carried out through the Python library Scikit-learn under the default parameters.

We first examine the structure of the resulting CART decision tree. For clarity of exposition, only the first two layers of the tree are illustrated in Fig. 2. In addition, in Fig. 2b and Fig. 2c we expand two of the nodes occurring in Fig. 2a.

The following three salient observations were drawn from the analysis.

1. The key indicator appears to be the *Peabody Picture Vocabulary Test (PPVT)* standard score. This is a standardized test designed to measure an individual’s vocabulary and comprehension and provide a quick estimate of verbal ability or scholastic aptitude. It could be deduced that almost $\frac{3}{4}$ of subjects with a standardized score above about 99.25 went on to achieve top GPA scores, underscoring the central importance of academic performance and aptitude.

2. For the subjects with low PPVT scores, the ability to follow instructions provided some relief, and increased the probability of a high GPA score to about 54% (up from 40%).

3. The level 2 features pictured were related to social welfare and the possibility of obtaining a loan, respectively, which suggests that financial stability is also influential in the attainment of academic excellence.

Next, the coefficients of the Logistic Regression model are presented in Fig. 3 (L1 regularization). As before, it is observed that grades and other early indicators of academic performance are crucially important, and are associated with five of the top eight coefficients. Again, many factors relating to the child’s social background and financial stability feature prominently.

The emerging picture is exceedingly complex and multi-faceted. On the one hand, test scores and academic aptitude occupy a central role, which is to be expected. Yet, there are indications that, beyond this, other features reflective of
Figure 3: Logistic Regression coefficients for the selected 69 features, ordered in a non-decreasing value of their waves, and Top/Low classes. Note that many of the features were “one hot encoded” in the dataset and as such the sign of these coefficients is not as important as the magnitude, as an indicator of importance. The figure has been annotated with the coefficients corresponding to the top eight coefficients by absolute value.
social and financial stability could also play a part, which strongly motivates the second, targeted part of this study.

3.2 Targeted Indicators of Success

While the insights devised in the previous section were highly illuminating, they were extracted from the entire dataset and the perspectives obtained were thus quite broad. This is why we sought to elucidate localized (targeted) indicators. In specific cases, were there certain factors which may “tip the balance” between a particular subject achieving high or low GPA scores? Test scores and a supportive home environment are always important, but we may find that the former is of greater significance in some cases yet the inverse is true in others.

To derive these detailed insights, the LIME technique was incorporated. For every instance, LIME produces a localized explanation of the classifier output by perturbing the feature values in order to generate a set of synthetic data points in the vicinity of the true instance. The posterior probability for each data point is estimated using the trained classifier, and a linear regression model is trained using the synthetic points as the inputs, and the posterior probabilities as the targets. The localized regression coefficients obtained in this way can then be interpreted as the importance of each feature, and is estimated separately for each subject. In other words, LIME specifies which features matter the most for each subject in order to determine whether the GPA of the subject will be high or low.

This technique was adopted, and extended, for the present context as follows.

1. As LIME requires a trained classifier capable of producing posterior probabilities, a Random Forest classifier is trained on the cases with Top and Low GPA scores.

2. Afterwards, LIME is applied to produce feature weights specific to each subject, which are then clustered using $k$-means clustering. The subjects in the resulting clusters are then assumed to share certain backgrounds or
behavioral traits as they share the same set of performance indicators.

3. Each cluster is then characterized by the centroid of the LIME coefficients of instances in the cluster (see Figure 4).

3.3 Analysis and Discussion

In the preceding analysis, k-means clustering was employed to divide the subjects into a set of disjoint clusters. The cluster centers are depicted in Figure 4. As Fig. 4 suggests, characteristics of subjects vary significantly over clusters. Initially, five clusters were obtained, but two of these (clusters 0 and 2) were highly similar and were subsequently combined, leaving 4 clusters in the final pool.

Finally, separate logistic regression models (L1 regularization) were trained for each of the four clusters. For each of these, we list below the features that were statistically significant, sorted in order of decreasing magnitude of the corresponding coefficient.

- **Clusters 0 and 2 (155 subjects):** Child’s attention and earlier performance
  (feature 52: Child attends to your instructions, feature 50: PPVT standard score).

- **Cluster 1 (224 subjects):** Father’s education and child’s earlier performance
  (feature 7: What is the highest grade/years of school that BF have completed?, feature 41: PPVT percentile rank)

- **Cluster 3 (228 subjects):** Social and financial support and father’s education
  (feature 58: You could ask friends/neighbors/co-workers for help/advice, feature 8: What is the highest grade/years of school that you have completed?, feature 6: Who gave you financial support during pregnancy, other?)

- **Cluster 4 (254 subjects):** Financial Support (feature 6: Who gave you financial support during pregnancy, other?)
Figure 4: Horizontal axis represents the selected 69 features sorted by their wave values in a non-decreasing order. Vertical axis stands for the corresponding weights of features for the five clusters generated according to LIME. Above each subfigure, for a given cluster, minimum, maximum, mean, mode and counts of minimum and maximum are reported.
These results are deeply interesting in a number of ways. First, observe that the test scores (PPVT) only appear in the first two clusters (which account for less than half of the subjects). As such, while they continue to be an important factor, they are not as central as in the global models. However, the other features in these two clusters are also related to academic aptitude or attention, which in turn implies that learning ability is the operative factor here, even if it is not always manifested in test scores.

On the other hand, note that financial stability is the most crucial factor in clusters 3 and 4. The feature who gave you financial supporting during pregnancy appears in both clusters 3 and 4, and in fact is the only significant feature in cluster 4. Cluster 3 also contains two other associated features - You could ask friends/neighbors/co-workers for help/advice and What is the highest grade/years of school that you have completed?. This possibly implies that the underlying requirement is for security (which can come in the form of financial resources or in social support), though for cluster 3, academic aptitude is still a prominent factor.

While the overall factors of success remain the same as in Figure 3, the relative importance of the features differs amongst the children. For some (i.e., clusters (0,2) and 1), test scores and scholastic aptitude seem to be more important, while financial security and social support are more important for others (i.e., clusters 3 and 4). Apparently, these findings are still preliminary and deeper insights can be obtained with more studies and data. However, the key point is that localized models such as this are important if we are to obtain a more nuanced view of what the actual indicators of success are for specific children and families.

4 Conclusions

In this study, a novel data science pipeline is proposed which was used to shed light on the importance of finding the specific features which were associated with success in different types of individuals. A data-driven approach was used to group
these families, then targeted success indicators were extracted from each of these
groups and analyzed. We note that these findings are based on a technique (LIME)
which was proposed only relatively recently, and as such should be treated as pre-
liminary. However, if and when superior methods are proposed, they can similarly
be incorporated into the workflow presented here, and used to produce even more
illuminating results.

Our findings suggest that the children of fragile families can be given the best
chance of success by using interventions that are tailored to the individual needs of
specific groups of families, e.g. in some families, a small home loan could be the
difference between a star student and a dropout, while in others a free mentoring
scheme could be more valuable.

5 Acknowledgments

The results in this paper were created with software written in Python 2.7 using
the following packages: Pandas [29], NumPy [30], Scikit-learn [31], LIME [20],
PyLab [32]. Funding for the Fragile Families and Child Wellbeing Study was pro-
vided by the Eunice Kennedy Shriver National Institute of Child Health and Human
Development through grants R01HD36916, R01HD39135, and R01HD40421 and
by a consortium of private foundations, including the Robert Wood Johnson Foun-
dation. Funding for the Fragile Families Challenge was provided by the Russell
Sage Foundation.

6 Appendix

Table 1: The set of selected 69 features listed in a non-decreasing order of the wave
values.

| Feature name                                      | Respondent | Wave No. |
|--------------------------------------------------|------------|----------|
| 0 Who gave you fin. supp. during preg., (BF) family? | mother     | 0        |
| 1 Why did rom. rel. end with (BF), Other?         | mother     | 0        |
|   | Question                                                                 | Respondent          | Score |
|---|------------------------------------------------------------------------|----------------------|-------|
| 2 | Are you and BM living together now?                                     | father               | 0     |
| 3 | People who currently live in your HH - 1st gender?                      | mother               | 0     |
| 4 | Father baseline education (combined report)                             | father               | 0     |
| 5 | How imp for successful marriage, wife has steady job?                  | mother               | 0     |
| 6 | Who gave you fin. supp. during preg., other?                           | mother               | 0     |
| 7 | What is the highest grade/years of school that BF have completed?      | mother               | 0     |
| 8 | What is the highest grade/years of school that you have completed?     | mother               | 0     |
| 9 | Are BM & BF living together?                                           | father               | 0     |
|10 | Is third person male or female?                                        | mother               | 2     |
|11 | Could you count on someone to co-sign for a loan for $ 5000?           | mother               | 2     |
|12 | In what country/territory was your father born?                        | father               | 2     |
|13 | How oft. dur. last mon./rel. did mother-withhold/try to control your money? | father               | 2     |
|14 | What were you required to do?--Attend school or training               | mother               | 2     |
|15 | Are all visible rooms of house/apartment dirty or not reasonably cleaned? | home visit           | 3     |
|16 | (He/she) can’t stand waiting, wants everything now                      | home visit           | 3     |
|17 | Her/She has angry moods                                               | home visit           | 3     |
|18 | In addition, do/did you sometimes work: weekends?                      | mother               | 3     |
|19 | Check contact sheet: do mother and father currently live together?     | home visit           | 3     |
|20 | In past year, did you think you were eligible for welfare at any time? | mother               | 3     |
|21 | Is there someone to co-sign for a bank loan with you for $ 1,000?      | mother               | 3     |
|22 | He/She hits others                                                     | home visit           | 3     |
|23 | How many times have you been apart for a week or more?                 | father               | 3     |
|24 | Maternal weight missing                                                | home visit           | 3     |
|25 | What about co-signing for $ 5,000?                                     | father               | 3     |
|26 | How many of the families on your block would you say that you know well? | mother               | 4     |
|27 | (He/She) feels (he/she) has to be perfect                              | home visit           | 4     |
|28 | Does respondent live with family or friends but pay no rent?           | father               | 4     |
|29 | (He/She) fears that (he/she) might think or do something bad           | home visit           | 4     |
|30 | Did you ever get pregnant in any of these relationships?               | mother               | 4     |
|31 | Does exterior of building have broken or cracked windows?              | home visit           | 4     |
|32 | In last 2 years, have you lived together                               | mother               | 4     |
|33 | Who gave help: other relatives of father                               | father               | 4     |
|34 | Age when you had sexual intercourse for the first time                 | father               | 5     |
|35 | Visible rooms are dirty or not reasonably clean                         | other                | 5     |
|36 | Did not pay full amount of rent/mortgage payments in past 12 months    | mother               | 5     |
|37 | Number of child’s close friends you know by sight, first and last name| primary caregiver    | 5     |
|38 | It’s hard for me to pay attention                                      | kid                  | 5     |
|39 | Child can’t concentrate, can’t pay attention for long                   | primary caregiver    | 5     |
|40 | Woodcock Johnson Test 10 percentile rank                               | home visit           | 5     |
|   |   |   |
|---|---|---|
| 41 | PPVT percentile rank | home visit 5 |
| 42 | How often PCG knows what you do during your free time | kid 5 |
| 43 | You were stopped by police but not picked up/arrested since last interview | mother 5 |
| 44 | Number of days per week you drop off or pick up child | primary caregiver 5 |
| 45 | Child is disobedient at home | primary caregiver 5 |
| 46 | Child is inattentive or easily distracted | primary caregiver 5 |
| 47 | Being a parent is harder than I thought it would be | father 5 |
| 48 | Father’s parents currently living together | mother 5 |
| 49 | How well you and your mom share ideas or talk about things that matter | kid 5 |
| 50 | PPVT standard score | home visit 5 |
| 51 | I do not get along well with members of child’s mother’s family | father 5 |
| 52 | Child attends to your instructions | teacher 5 |
| 53 | Child repeated 4th grade | primary caregiver 5 |
| 54 | Woodcock Johnson Test 10 age equivalency | home visit 5 |
| 55 | Father has spent any time in jail | mother 5 |
| 56 | Child has participated in Title I ESL/bilingual | teacher 5 |
| 57 | Number of families on block know well | primary caregiver 5 |
| 58 | You could ask friends/neighbors/co-workers for help/advice | mother 5 |
| 59 | Child ignores peer distractions when doing class work | teacher 5 |
| 60 | Child shows anxiety about being with a group of children | teacher 5 |
| 61 | Father’s parents currently living together | father 5 |
| 62 | Who usually initiated the contact | primary caregiver 5 |
| 63 | Father has talked to doctor about child in last year | primary caregiver 5 |
| 64 | Child’s science and social studies | teacher 5 |
| 65 | It’s hard for me to finish my schoolwork | kid 5 |
| 66 | I get in trouble for talking and disturbing others | kid 5 |
| 67 | Frequency you address developing | teacher 5 |
| 68 | Interior of home is dark | other 5 |

**References**

[1] M Scott DeBerard, Glen Spielmans, and Deana Julka. Predictors of academic achievement and retention among college freshmen: A longitudinal study. *College student journal*, 38(1):66–80, 2004.

[2] Thomas R Coyle and David R Pillow. Sat and act predict college gpa after removing g. *Intelligence*, 36(6):719–729, 2008.
[3] Marisa Salanova, Wilmar Schaufeli, Isabel Martínez, and Edgar Bresó. How obstacles and facilitators predict academic performance: The mediating role of study burnout and engagement. *Anxiety, stress & coping*, 23(1):53–70, 2010.

[4] Stuart A Tross, Jeffrey P Harper, Lewis W Osher, and Linda M Kneidinger. Not just the usual cast of characteristics: Using personality to predict college performance and retention. *Journal of College Student Development*, 41(3):323, 2000.

[5] E Ashby Plant, K Anders Ericsson, Len Hill, and Kia Asberg. Why study time does not predict grade point average across college students: Implications of deliberate practice for academic performance. *Contemporary Educational Psychology*, 30(1):96–116, 2005.

[6] Carolyn W Kern, Nancy S Fagley, and Paul M Miller. Correlates of college retention and gpa: Learning and study strategies, testwiseness, attitudes, and act. *Journal of College Counseling*, 1(1):26–34, 1998.

[7] Frank Pajares, James Hartley, and Giovanni Valiante. Response format in writing self-efficacy assessment: Greater discrimination increases prediction. *Measurement and evaluation in counseling and development*, 33(4):214, 2001.

[8] Roberto Colom, Sergio Escorial, Pei Chun Shih, and Jesús Privado. Fluid intelligence, memory span, and temperament difficulties predict academic performance of young adolescents. *Personality and Individual differences*, 42(8):1503–1514, 2007.

[9] Mary E Pritchard and Gregory S Wilson. Using emotional and social factors to predict student success. *Journal of college student development*, 44(1):18–28, 2003.
[10] Paulo A Graziano, Rachael D Reavis, Susan P Keane, and Susan D Calkins. The role of emotion regulation in children’s early academic success. *Journal of school psychology, 45*(1):3–19, 2007.

[11] Vonnie C McLoyd. Socioeconomic disadvantage and child development. *American psychologist, 53*:185, 1998.

[12] Gillian Considine and Gianni Zappalà. The influence of social and economic disadvantage in the academic performance of school students in australia. *Journal of Sociology, 38*(2):129–148, 2002.

[13] Linda A Jackson, Alexander Von Eye, Frank A Biocca, Gretchen Barbatsis, Yong Zhao, and Hiram E Fitzgerald. Does home internet use influence the academic performance of low-income children? *Developmental psychology, 42*(3):429, 2006.

[14] Kathryn Harker Tillman. Family structure pathways and academic disadvantage among adolescents in stepfamilies. *Sociological Inquiry, 77*(3):383–424, 2007.

[15] Priscilla Dass-Brailsford. Exploring resiliency: academic achievement among disadvantaged black youth in south africa: general’section. *South African Journal of Psychology, 35*(3):574–591, 2005.

[16] Timothy J Biblarz and Greg Gottainer. Family structure and children’s success: A comparison of widowed and divorced single-mother families. *Journal of Marriage and Family, 62*(2):533–548, 2000.

[17] Douglas B Downey. The school performance of children from single-mother and single-father families: Economic or interpersonal deprivation? *Journal of family issues, 15*(1):129–147, 1994.

[18] Kaia Laidra, Helle Pullmann, and Jüri Allik. Personality and intelligence as predictors of academic achievement: A cross-sectional study from elemen-
tary to secondary school. *Personality and individual differences*, 42(3):441–451, 2007.

[19] Salganik M., Lundberg I., Kindel A., and McLanahan S. Introduction to the special issue on the fragile families challenge. *Socius: Special issue on the Fragile Families Challenge*, 2018.

[20] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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*, pages 1135–1144. ACM, 2016.

[21] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[22] Ian T Jolliffe. Principal component analysis and factor analysis. In *Principal component analysis*, pages 115–128. Springer, 1986.

[23] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[24] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[25] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[26] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
[27] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.

[28] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[29] Wes McKinney. Data structures for statistical computing in python, 2010.

[30] Stfan van der Walt, S. Chris Colbert, and Gal Varoquaux. The numpy array: A structure for efficient numerical computation, 2011.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[32] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.