Fragile Families Challenge: Identifying Early Childhood Predictors for Academic Success

Claudia V. Roberts
Princeton University
claudiar@cs.princeton.edu
FFC-id: chicacvr

Abstract

The Fragile Families Challenge is a mass collaboration social science data challenge aimed at learning how various early childhood predictors affect the long-term outcomes of children born into majority single-parent homes. Outcomes include measures of academic achievement, passion and perseverance, and extreme poverty as well as events of housing eviction, caregiver layoff, and caregiver job training. In an effort to identify the features most predictive of academic achievement and to better understand the subtleties of the Fragile Families dataset, I trained eight different models using various regularization and preprocessing techniques such as imputation, feature selection, and scaling. I found the education level of the mother's boyfriend to be the top predictor of academic achievement later in the child's life.

1 Introduction

The Fragile Families and Child Wellbeing Study (FFCWS) is a joint cohort study run by Princeton University and Columbia University [13]. It follows a group of nearly 5,000 American children born into majority single-parent households, or "fragile families," from birth into adulthood. The aim of the study is to characterize the relationships and conditions of unmarried parents and to study the cognitive development, mental and physical health, and social relationships of children born into such families. The Fragile Families Challenge (FFC) is a mass collaboration social science project designed to harness the predictive power of the Fragile Families dataset [14]. The FFC invites community members to use the data to build models that best predict six key outcomes: GPA, grit, material hardship, eviction, job loss, and job-training. In this paper, I focus on predicting GPAs only.

In order to identify the most important early childhood predictors for academic achievement and to better understand the idiosyncrasies and challenges that arise in social science problems such as these, I trained and evaluated eight different model types. I used various combinations of feature selection, data imputation, data scaling, and regularization methods to generate over a hundred uniquely trained predictive models. I present the top ten performing models along with a comparative assessment of their computational performance and predictive ability.

2 Related Work

The task of predicting child GPAs using the Fragile Families (FF) dataset is a multiple regression analysis task. Multiple regression analysis is a predictive technique that seeks to find a relationship between multiple independent variables, or predictors, and a single continuous dependent variable, or response [10]. In the social sciences, the most widely used predictive modeling strategies involve variations of the generalized linear model, such as Ordinary Least Squares linear regression [14]. Most research seeks to extend these models to address the shortcomings that arise in particularly

challenging data settings. For example, in low data settings or settings where the number of parameters far exceeds the number of samples, overfitting becomes a problem, and the learned model loses its ablity to generalize. [16] demonstrated how the regularization method could be used to reduce the complexity or magnitude of parameter estimates and thus, help avoid overfitting [3]. Since then, many regularized regression models have been developed and optimized for such high-dimensional data settings. [9] propose the Least Angle Regression (LARS) algorithm, a computationally efficient version of Forward Stagewise Regression, [15] provide interpretability, stability, and built-in feature selection with their popular least absolute shrinkage and selection operator (lasso) model, and [6] present the Dantzig selector, a statistical estimator for high-dimensional noisy contexts.

2.1 Data processing

The given FFC dataset contains survey responses to 12,943 questions for 4,242 children collected from their birth to age nine. The dataset also contains year-15 data for six key outcomes, including GPA. We are given the reported high school GPAs for 2,121 children. The task then is to predict the GPAs for the remaining 2,121 children.

The first preprocessing step involved removing the samples for which the GPA was "NA" or outside the range 1.0 through 4.0, inclusive. This reduced the sample space from 2,121 to 1,165. The second step involved imputing the missing values in the survey data. I used the SciKitLearn Python library [11] to implement two imputation strategies: median imputation and mode imputation. Respectively, these methods replace the missing values with the median and mode of the observed values for that feature column [4]. Since the SciKitLearn Imputer class removes columns for which all values are missing, this step reduced the feature space from 12,943 to 10,228. I used median and mode since these imputation methods are more robust to outliers than their mean counterpart.

2.2 Methods

With 1,165 samples and 10,228 predictors, the FFC dataset is a high-dimensional dataset. This motivated my decision to train and test the following eight models, which in this case, were used to perform regression analysis. Models marked with "*" had their hyperparameters tuned using cross-validation. LARS, Ridge, and Lasso models had their α parameter tuned, Elastic Net had its regularization and shrinkage parameters tuned, and the SVR model had its penalty parameter C and ε parameter tuned.

1. *Ordinary least squares linear regression* (LR): simple modeling technique, assumes the predictors are independent of each other
2. *Least-angle regression** (LARS): numerically efficient in high-dimensional contexts
3. *Ridge regression** (Ridge): linear least squares with ℓ_2 regularization
4. *Elastic Net** (EN): provides combined ℓ_1 and ℓ_2 priors as regularizer
5. *Orthogonal Matching Pursuit* (OMP): simple and fast implementation [5]
6. *Lasso regression** (Lasso): uses ℓ_1 prior as regularizer, offers automatic dimensionality reduction
7. *Decision Tree regression* (DTR) using Mean Squared Error score: results are interpretable
8. *ε-Support Vector Regression with linear kernel** (SVR): effective in high-dimensional spaces [8]

To reduce computation time and avoid overfitting, I used univariate feature selection using the F-step and mutual information scoring, selecting top 10% and top 20% scored features. To ensure all the predictors contribute equally to the regression anlaysis, regardless of scale, I standardized the data.

2.3 Pipeline

To get an idea of the predictive performance and generalization ability of the models prior to submitting prediction results to the FFC web portal, I made a 80%/20% split of the dataset. This allowed me to have an internal, representative validation set on which to more carefully analyze the features and prediction uncertainty [1]. I used the training set (80% split) to assess goodness-of-fit of 124

uniquely trained models, and used cross-validation (CV) to pick optimal hyperparameters for each. I then picked the top 10 performing models, re-trained them using the optimal hyperparameters, and evaluated their performance against my internal validation set. Finally, I used these 10 models to make predictions on the unseen FFC submission samples.

2.4 Evaluation metrics

I used the following evaluation metrics to perform a comparative assessment of the various regression model, imputation, scaling, and feature selection combinations: 10-fold CV mean Mean Squared Error (MSE) score with 95% confidence interval, wall clock time for training and testing (seconds), R^2 score, residual plots, prediction error plots, and MSE score. I chose these metrics because they are the most popular measures for quantifying the quality of model predictions.

3 Spotlight: Linear ε-Support Vector Regression

The Support Vector Regression Machine (SVR) is a version of Support Vector Machines (SVM) that can be used for regression problems. Similar to SVM classifiers, SVRs only use a subset of the training data to build the model, making them memory efficient. Given a set of training samples $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}$, where $\boldsymbol{x}_i$ is an N-dimensional feature vector and $y_i \in \mathbb{R}$ is the response variable, SVRs seek to find a function $f(\boldsymbol{x})$ that deviates at most ε from the observed response variables y_i for each training point $\boldsymbol{x}_i$ [12]. In other words, SVR finds a function for which all predictions are within a margin of tolerance ε. The case of the linear prediction function takes the following form:

$$f(\boldsymbol{x}) = \langle w, x \rangle + b \text{ with } w \in \chi, b \in \mathbb{R}$$

Where χ denotes the space of the input patterns and $\langle ., . \rangle$ denotes the dot product in χ [12].

Training the SVR means minimizing error and solving the following convex optimization problem:

$$\text{minimize } \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{subject to } \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon \end{cases}$$

However, in some cases, solving this problem is not feasible, at which point, slack variables can be introduced to allow for error.

$$\text{minimize } \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{\ell}(\xi_i + \xi_i^*)$$

$$\text{subject to } \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

This formulation allows for deviations larger than ε to be tolerated. The amount of tolerance, meaning the amount the error can deviate by, is dictated by the constant $C > 0$.

In order to handle training data that is not linearly separable, like SVM classifiers, SVRs can exploit the kernel trick. Kernel functions and the kernel trick allow similarity between samples to be used as features in a computationally efficient way [2]. The kernel solution involves projecting the features to a higher dimensional space and using a linear function in the projected feature space. The most common kernels supported by SVRs are polynomial, radial basis function (rbf), and sigmoid.

Unlike generalized linear models such as linear regression, that rely on the assumption that the response variable is a linear combination of predictor variables that are themselves independent from each other, SVRs do not make such strong assumptions.

4 Discussion and Conclusion

Below are the top 7 predictors for academic achievement, based on the top two performing models. The coefficients are in parenthesis. Since the data were standardized, I was able to compare predictor coefficients according to their relative impact on the response variable.

Figure 1: Shows the effects of different feature selection, imputation, and scaling methods on mean MSE scores across all 124 models (the lower the score the better). F-Reg feature selection is robust to outliers, median imputation performs three orders of magnitude better than mode, and no scaling performs better than standardization since it is robust to outliers.

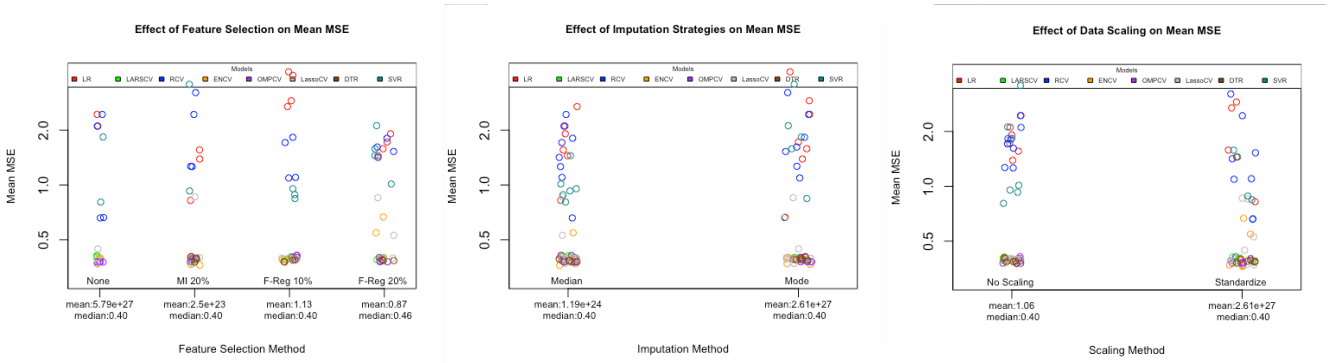
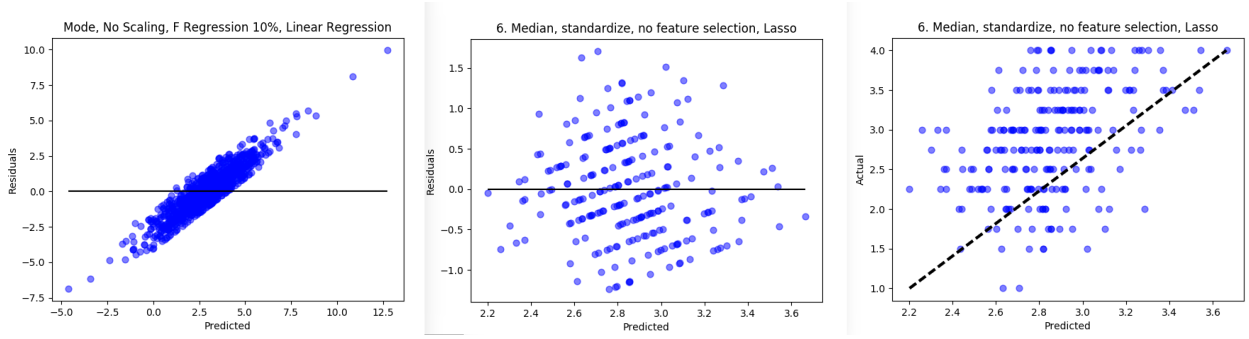


Figure 2: Shows residual plots for one of the worst models as well as the best model. The left-most residual plot for the worst model shows a linear pattern and is biased and homoscedastic. It appears it is not capturing some important signals. In contrast, the residual plot for the best model in the center is unbiased and homoscedastic. The data points are randomly scattered along the zero line. The right most graph shows the prediction error plot for the best model. The data points are loosely clustered along the diagonoal, indicating that while the model is good, there are still many errors.



1. m1i3: What is the highest grade/years of school that BF have completed? (0.046)
2. t5b1w: B1W. Child attends to your instructions (0.042)
3. hv5_ppvtpr: PPVT percentile rank (0.042)
4. f1b20: Int chk: Are BM & BF living together? (-0.037)
5. m1i1: What is the highest grade/years of school that you have completed? (0.030)
6. hv5_wj10pr: Woodcock Johnson Test 10 percentile rank (0.029)
7. p5m1: M1. Number of families on block know well (-0.024)

Considering that the task was to predict GPAs, the top five predictors are quite intuitive. It makes sense that there is a positive correlation between how obedient a teacher thinks a child is and the child's performance in school. Similarly, it makes sense that a child's GPA is positively correlated to their performance on tests such as the Peabody Picture Vocabulary Test (PPVT) and Woodcock-Johnson Tests of Cognitive Abilities. The better a child performs on a set of scholastic aptitude and intelligence tests, the better their GPAs are expected to be. Additionally, biological mothers who do not live with their boyfriends seems to negatively impact child GPAs.

Interestingly, the top predictor of a child's GPA is the education level of the mother's boyfriend. The more education the boyfriend has, the better the child's GPA. This points to the importance of having positive male role-models in a child's life. Moreover, this predictor and the fifth predictor give us a hint of why the linear regression model performed so poorly. Research indicates that the education

Model	Impute	Scale	Feat. Sel.	Mean MSE 95% Confid	Train(s)
ENCV	Med	Stand	MI 20%	-0.36(+/- 0.09)	5002.97
ENCV	Mode	Stand	MI 20%	-0.36(+/- 0.10)	2384.4
ENCV	Med	Stand	None	-0.36(+/- 0.09)	3208.61
ENCV	Mode	Stand	None	-0.36(+/- 0.07)	3186.79
LassoCV	Mode	Stand	MI 20%	-0.37(+/- 0.08)	2399.12
...	...	...	...	...	...
LR	Mode	Stand	None	-3.95e+16(+/- 2.37e+17)	30.06
SVR	Mode	Stand	MI 20%	-3.91e+23(+/- 2.34e+24)	1037.62
SVR	Med	Stand	MI 20%	-7.63e+24(+/- 4.57e+25)	2092.40
SVR	Med	Stand	None	-6.61e+25(+/- 3.73e+26)	224.21
SVR	Mode	Stand	None	-1.62e+29(+/- 9.71e+29)	164.32

Table 1: **Mean MSE score for Top 5 and bottom 5 models** The top four performing models use EN regression (the higher the score the better). The top 5 models all used data standardization. Tuning the hyperparameters using CV took between 40 and 80 minutes for EN. The worst four performing models use SVR. While SVR is advertised as being efficient in high-dimensional settings, in cases such as the FFC where the number of parameters is much greater than the number of samples, SVR performs very poorly. However, the data suggests that SVR models are indeed more computationally performant than EN and Lasso. While not shown fully in this table, LR performed the worst after SVR. This indicates that the data might violate some of the assumptions made by LR, such as non-multicollinearity. The drastically low MSE scores for the bottom 5 models suggest that SVR and LR are not robust outliers, unlike the regularized EN and Lasso regression models.

Model	Impute	Scale	Feat. Sel.	R^2	MSE	FFC-MSE	Train(s)	Test(s)
EN	Med	Stand	MI 20%	0.17	0.38	0.91	136.44	0.03
EN	Mode	Stand	MI 20%	0.19	0.37	0.40	151.90	0.02
EN	Med	Stand	None	0.18	0.37	**0.37**	0.64	0.02
EN	Mode	Stand	None	0.18	0.37	0.50	0.81	0.02
Lasso	Mode	Stand	MI 20%	0.15	0.38	0.68	149.75	0.02
Lasso	Med	Stand	None	0.18	0.37	**0.37**	0.62	0.01
OMPCV	Med	Stand	None	0.13	0.39	0.39	9.48	0.02
OMPCV	Med	None	None	0.13	0.39	0.39	8.95	0.003
DTR	Med	None	F-Reg 10%	0.02	0.44	0.41	0.21	0.005
DTR	Med	None	MI 20%	-0.03	0.47	0.41	126.60	0.009

Table 2: **Evaluation results for top 10 models on validation set and unseen FFC samples** The model that performed the best on the validation set performed the worst on the unseen FFC dataset. This indicates that this model had poor generalization ablity. Models 3 and 6 performed the best on the unseen data and their validation MSE scores matched that of the unseen FFC MSE scores. While one model used EN and the other used Lasso, both used median imputation, data standardization, and no feature selection. Based on the train times, models that use MI 20% feature selection take much longer to train than the models that use no feature selection or F-Reg 10% feature selection.

level of a mother and her partner are not completely independent. Census data shows that "71% of college graduates married another college graduate" [7]. Nonetheless, these two predictors point to a well-known reality: without parents as role-models, it takes luck, the strength of the community, and sheer grit to break out of the cycle of poverty.

Using the prediction error plots for the top model, I identified the top child who performed better than predicted and one who performed worse. The child who did worse had a mother who completed technical training and who's boyfriend had a graduate degree. For the child who beat the odds, the mother and her boyfriend did not have a high school diploma. The rest of their answers were similar, so what happened? What are we missing? For this, we would need to conduct in-depth interviews and collect more data.

Future extensions include strategic manual feature engineering and using smarter imputation methods such as regression and last observation carried forward. I would also explore automated and manual imputation methods for identifying and imputing values that are missing not at random.

References

[1] Cos 424 assignment 2: Fragile families challenge. [Online; accessed 3-April-2017].

[2] Cos 424 lecture 5: Features and kernels. [Online; accessed 27-February-2017].

[3] Cos 424 lecture 8: Regularized regression. [Online; accessed 3-April-2017].

[4] Cos 424 precept 6: Imputation methods and bootstrapping. [Online; accessed 3-April-2017].

[5] T. T. Cai and L. Wang. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information Theory*, 57(7):4680–4688, July 2011.

[6] Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n (pkg: p2313-2404). *The Annals of Statistics*, 35(6):2313–2351, 2007.

[7] Philip Cohen. College graduates marry other college graduates most of the time. `https://www.theatlantic.com/sexes/archive/2013/04/college-graduates-marry-other-college-graduates-most-of-the-time/274654/`.

[8] Harris Drucker, Chris J. C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines, 1996.

[9] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[10] Karl Pearson and Alice Lee. On the generalised probable error in multiple normal correlation. *Biometrika*, 6(1):59–68, 1908.

[11] 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.

[12] Alex J. Smola and Bernhard Scholkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, August 2004.

[13] The Fragile Families and Child Wellbeing Study. Fragile families and child wellbeing study. `http://www.fragilefamilies.princeton.edu/about`.

[14] The Fragile Families and Child Wellbeing Study. Fragile families challenge. `http://www.fragilefamilieschallenge.org`.

[15] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[16] A. N. Tikhonov. On the solution of incorrectly formulated problems and the regularization method. *Dokl Akad Nauk SSSR*, 151:501–504, 1963.