Assignment 2: Fragile Families Challenge

Kyle Genova
Princeton University
kgenova@cs.princeton.edu

Abstract

We consider the prediction of six key outcome metrics reported in the Fragile Families Challenge [1]. The Fragile Families challenge dataset provides detailed survey responses for 4,242 disadvantaged children taken over a series of years from birth to age 15. The specific task we consider is to predict three continuous variables and the probabilities of three binary variables captured at age 15 using training data from birth to age 9. We use a unified framework based on Elastic Net regression to predict all six outcomes jointly, and demonstrate its efficacy for both continuous and binary variables. At the time of this writeup, all models are in the top 10, four are in the top 3, and the job training ranks 1^{st}. Our approach is also interpretable, and so we present a discussion of the most important survey questions for each of the six key outcomes.

1 Introduction

Predicting outcome variables from the Fragile Families dataset is a practically relevant and thereby an interesting machine learning task for study. The ability to predict the future outcomes of disadvantaged children based on their current survey responses would be highly useful for two reasons. First, it would enable social services agencies to adjust their aid allocations more dynamically in advance of anticipated challenges in a family's situation. Second, it would enable improvements to the approach to intervention by (1) inspecting trained models to determine which categorized factors are most correlated with success and (2) examining outlier children according to a model to interview further to discover unexpected positive or negative correlations to success. We focus on a model capable of predicting all outcome variables. The predicted variables are GPA, a measure of academic success, grit, a measure of perserverence, material hardship, a measure of extreme poverty, eviction, a binary variable estimating whether a family has been evicted since the year 9 survey, layoff, whether a primary care-giver has been laid off since the year 9 survey, and job training, whether the child has participated in any kind of job training since the year 9 survey. We compare a variety of models fit separately to each outcome, and show that by jointly predicting the entire six outcome set in a single model, performance can be improved. Additionally, we interpet our models.

2 Related Work

While the Fragile Families Challenge is a novel step in applying machine learning techniques to social studies problems, there has been a wealth of work on regression models under assumptions appropriate for the Fragile Families Challenge. In particular, linear models with regularization have been found to be generalizable for problems with high-dimensional feature spaces and relatively small datasets [6]. Linear regression and Generalized Linear Models are highly prevalent techniques for regression, and there are entire books and book chapters devoted to variants; see [3, 6] for two such chapters. Specifically, three regularization variants are particularly relevant to this problem, as overfitting in this context is difficult to avoid. The first is ℓ_2 regularization, often known as Ridge regression or Tikhonov regularization. It became known from its applicability to solving integral equations and to avoiding numerical instability, as in [8] and [2]. This regularization seeks to reduce the weights of all correlative variables, avoiding overreliance on a small subset of features. Next, ℓ_1 regularization, also known as Lasso regression, is relevant. It enables generating sparser models by selecting a subset of active constraints. It was first proposed by Tabshirani [9]. This approach has the advantage of improved interpretability, as the resultant linear model relies on a smaller number of features. Finally, there has been prior work on combining the benefits on both Ridge and Lasso

regression by solving models having both penalties; this is known as Elastic Net regression and was proposed in [10].

There has also been much work on data imputation, or the task of estimating or otherwise replacing missing features in training examples: books such as [5] are devoted to the task. One of the critical assumptions in this task is why data is missing; assumptions such as missing at random, missing completely at random, or missing not at random are common variants. When the missing data is assumed to be explainable from only the data that is present, one of the most useful techniques is to train a machine learning model such as logistic regression on the dataset to predict the missing values; this corresponds to the missing at random assumption (see [4]). We adapt these approaches, as described below.

3 Data Processing: Imputation and Encoding

One of the primary challenges in working with the Fragile Families dataset is the large fraction of missing survey responses, as well as the fraction of features missing at least one result. A variety of possible imputations are reasonable, but we opt for a custom approach, rather than using the data imputation script provided. We break the task in two. First, we examine the data booklet and compute some statistics on the missing data and to help inform what might be reasonable (statistics omitted for space). Given this information, we split the features heuristically into categorical, numerical, and irrelevant clusters (only a small number of features known to be irrelevant, such as the id values, were manually removed). Categorical values are re-encoded using a one-hot encoding; here, each missing value code was considered a separate feature (i.e. not applicable, not asked, not in wave are all separate features). Given this one-hot encoding matrix, the second step is imputation of the numerical values. Numerical values are each considered a separate regression task, and are imputed by fitting a simple linear regression model using the one-hot encoding matrix as a feature set. The imputed numerical values and the one-hot-encoded values were concatenated as the final training matrix for all methods discussed further; methods for which the 0-1 and numerical values should be considered separately (such as naive bayes) are split, trained separately with Gaussian versus Bernoulli assumptions, and then combined into an ensemble model.

3.1 Overview of Models

We compare the following models, implemented using the Scikit-learn library [7]:

1. Linear Regression, trained independently on each of the six outcomes (on the rows for which that outcome was present, with imputed features as described above).
2. Lasso Regression, trained independently on each of the six outcomes (on the rows for which that outcome was present, with imputed features as described above). Hyperparameters were tuned using a grid with shuffle split cross validation having 5 folds and an 80-20% training-validation split.
3. Ridge Regression, trained independently on each of the six outcomes (on the rows for which that outcome was present, with imputed features as described above). Hyperparameters were tuned using a grid with shuffle split cross validation having 5 folds and an 80-20% training-validation split.
4. Elastic Net Regression, trained independently on each of the six outcomes (on the rows for which that outcome was present, with imputed features as described above). Hyperparameters were tuned using a grid with shuffle split cross validation having 5 folds and an 80-20% training-validation split.
5. Multioutput Elastic Net Regression. Trained jointly to predict the 6-outcome set, on the rows for which all six outcomes were present. Hyperparameters were tuned using a grid with shuffle split cross validation having 5 folds and an 80-20% training-validation split.
6. Soft Ensemble Multioutput Elastic Net and individual Elastic Net Regression. This is a soft-voting ensemble regressor that simply averages the estimates of the Multioutput Elastic Net and the individually-fit Elastic Net. Stacking regressors is generally preferable to averaging, but is not applicable here because of the training dataset discrepancy. The motivation behind this classifier is that the Multioutput classifier is able to take advantage of correlations between the outcome variables, but must rely only on the significantly smaller

dataset comprised of respondents for which all outcomes are known. In contrast, the individually fit elastic net models can take advantage of a larger dataset, but without knowing the correlations between the outcome variables. Thus, an ensemble approach can potentially exploit this disparity.

Note also that many other regressors and classifiers were fit and tested, including: Gaussian Processes, Logistic Regression, Gaussian Naive Bayes, Bernoulli Naive Bayes, Stochastic Gradient Descent, and Decision Trees. Models were first fit using hyperparameter optimization with cross validation where applicable, but with only 80% of the dataset. The remaining holdout set was used for local testing of the models, so that models could be evaluated without submitting to the Fragile Families Challenge. Based on those results, a subset of techniques that seemed worth comparing here were refitted. Once the hyperparameters were chosen via cross validation on the entire dataset, each model was refit to the entire dataset, and then those models were submitted to the Fragile Families Challenge website to get the error results presented here. This why only a subset of attempted methods have detailed results; some were pruned before submission to the challenge website.

4 Spotlight Classifier: Linear Regression with ℓ_1 Regularization

We highlight the Lasso method for regression, which is a particularly important regression technique and which the most successful regression method used (elastic net regression) builds upon.

4.1 Motivation, Goal, and Assumptions of Lasso Regression

The motivation behind Lasso Regression is that a sparse linear combination of predictors is a useful model for two reasons. First, if features are linearly correlated with the output prediction, the model is at least partially interpretable. Furthermore, the fewer predictors that are used, the more interpretable the model. Second, a sparse linear model is extremely efficient to evaluate. Given these motivations, the objective function of Lasso regression can be stated as follows, where $\beta_{1,\ldots,N}$ are the parameters that describe the linear relationship between $x_{1,\ldots,N}$, the input features, and $\bar{y}$, the predicted estimates of the true values y given some regularization constant t:

$$\hat{\beta}^{\text{lasso}} = \arg\max_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$
$$\text{subject to} \sum_{j=1}^{p} |\beta_j| \leq t$$

When broken by lines, the above equation can be explained in terms of two intuitive sections. The first line is the residual sum of squares between the training example points and the training prediction points; without the constraint on the following line, this problem could be solved straightforwardly via ordinary least squares (OLS) using a method such as direct application of the Moore-Penrose Pseudoinverse.

The second component, the constraint, enables both feature selection and some weight decrease to avoid overfitting. The hyperparameter t is a given constant; the smaller it is, the more regularization is present. The constraint itself states that the ℓ_1 norm of the weight vector must be at most this constant t. To understand why a constraint on the ℓ_1 norm results in sparse features, consider the subspace of the parameter domain that is feasible under the ℓ_1 constraint: it forms a cornered subspace, where the corners are sparse feature space locations. This feasible region is much more likely to intersect the object at such a corner, because the corner reaches further from the origin for smaller t. Thus, we may conclude that a solution to the above objective function will result in model with small least squares and sparse features, for appropriate choice of t.

4.2 Computing the Lasso Regression

It is possible to solve Lasso regression using a blackbox quadratic programming solver, as the above formulation of the objective can be straightforwardly transformed to a quadratic programming problem (the only step is to replace the absolute values with two dummy variables). However, it is possible to improve on this approach via a technique known as Least Angle Regression. In order to fit a Lasso regression in this way, the following forward stepwise regression is used [3]:

1. Standardize features to have zero mean and unit norm. Set the parameters $\beta := 0$ and the initial residual r to the difference between the prediction with these parameters $\bar{y}$ and the true values y.
2. Select the feature i most correlated with r, add it to the currently empty active set , $A := \{i\}$, and update the currently active parameters β_A by taking a step in the direction $\delta = (X_A^T X_A)^{-1} X_A^T r$ of chosen size α, where X_A is the feature matrix filtered to the active set. That is, set $\beta_A := \beta + \alpha(X_A^T X_A)^{-1} X_A^T r$. This steps the parameters in the direction towards which the most correlated feature pulls it.
3. Continue stepping in this direction until another feature also becomes the most correlated feature, and continue adding this feature to the active set. If at any point a parameter β_k becomes 0, remove it from the active set. Continue until convergence.

While the algorithm is guaranteed to converge to the Lasso solution, the proof is beyond the scope of this one-page writeup. However, the algorithm itself is intuitive: repeatedly step in a direction that increases the weight of the parameters most correlated with the residual, maintaining an active set of nonzero parameters.

5 Results

First, we compare the MSE of each of the methods when submitted to the Fragile Families website. These are shown in the following table. More specifically, the value shown for the three continuous variables, GPA, Grit, and MH (Material Hardship), is the Mean Squared Error, while the value shown for the binary variables is the Brier Loss [1]:

$$MSE := \frac{1}{n}\sum_i (Y_i - \hat{Y}_i)^2 \qquad\qquad Brier := \frac{1}{n}\sum_i (Y_i - \hat{P}(Y_i = 1))^2$$

These both quantify the average of the squared distances between the predicted values and their true values. The Brier Loss is similar to Mean Squared Error, but is defined instead for probability estimates of binary variables. Importantly, the use of this error is why the binary variables were treated like regression problems in this writeup; we regress the probability of the event, rather than classifying true or false.

Method	GPA	Grit	MH	Eviction	Layoff	JT
Lasso	0.37494	0.22521	0.02511	0.05188	0.17435	0.20037
Ridge	0.44323	0.31415	0.02716	0.05714	0.21985	0.25414
Linear Regression	0.44328	0.31449	0.02717	0.05714	0.21992	0.25467
Elastic Net	0.37109	0.21963	**0.02461**	0.05224	0.17396	0.19842
MultiTask Elastic Net	**0.36837**	**0.21502**	0.02562	**0.05112**	0.1737	**0.19382**
Elastic Net + MultiTask	0.36924	0.21581	0.02497	0.05137	**0.17336**	0.19531

Table 1: FFC Error for each of the six key outcomes. Lower is better.

Above, we first note that, broadly, the MultiTask Elastic Net classifier is the highest performing classifier. This tells us that there must be some significant correlation between the predicted outputs, and further that it is more valuable to the task than the additional data available by training elastic net models separately. One possible explanation for this, however, is that data elements for which predicted variables are missing are more likely to be outliers, affecting the model if they are included.

Beyond this, we see that other methods perform mostly as expected. Linear regression performs quite poorly: this makes sense, as given the relative feature to data dimensionality, any model without severe regularization will be prone to overfitting. Ridge regression also performs poorly, which makes sense- the feature encoding process was messy, as a lot of imputation was done, some features had more than half of their elements missing, and likely a few numerical versus continuous variables were even misclassified. There are also features which will undoubtedly be very noisy or irrelevant to the outcomes. Thus, a model like ridge regression that cannot set feature weights to zero is predisposed to perform poorly. Finally, we note that the averaged ensemble technique was largely unsuccessful- this is unsurprising, though a more sophisticated ensemble model seems very promising given that performance was able to increase in one instance by simple averaging.

Correlation	ID
0.00959	n5f0
0.00926	f1e7a
0.00812	n5e3a
0.00737	n5b3c
0.00665	f1f10b

Correlation	ID
-0.00922	hv3a16c1_6
-0.00814	hv3a16d3_3
-0.00582	f5g16e
-0.00528	f5f26b2_9
-0.00526	m2e9

Correlation	ID
0.00538	f4k9b
0.00444	m4b9
0.00406	f4k6
0.00388	m4c13a2
0.00376	t5f5a

Table 2: Questions most predictive of GPA, Grit, and Material Hardship

Correlation	ID
0.00584	f4k14a1
0.00497	f4k8
0.0046	m4j2a_2
0.00365	cf3alc_case
0.00331	o5a6b

Correlation	ID
0.00142	m3k3a_4
0.00141	p5i37
-0.00133	m2f2c1
-0.00124	f4a6a2
-0.00123	m3d7d

Correlation	ID
0.00849	f5i26d
0.00714	m4c37e2
0.00586	f4natwt_rep6
0.00461	m5a5e06
0.00428	cf4tele

Table 3: Questions most predictive of Eviction, Layoff, and Job Training

We conclude this section by advertising that at the time of this writeup, the results submitted to the competition are excellent: out of 67 submissions, the ranks are: 1^{st} (Job Training), 2^{nd} (Material Hardship), 3^{rd} (Eviction), 3^{rd} (GPA), 7^{th} (Grit), and 10^{th} (Layoff). We also note that the lead held by the job training model over other submissions is much higher than that for any other category: the gap between our model and 2^{nd} place is larger than that between 2^{nd} place and the score from predicting the expected value. Our submission ID is `kgenova`.

5.1 Interpretation of Results

Because our models rely primarily on a few features, we can compute the predictors having the largest impact on our output predictions. In Tables 2 and 3, we show the top 7 question ids for each of the six independently fitted Elastic Net models. Note that in interpreting these models, it is important to consider how feature encoding relates to the questions. Numerical values were encoded in a 1:1 ratio with questions. However, a one-hot encoding was used for categorical variables. Thus, for categorical variables, we use the individual feature within the group associated with the question that is most correlated, positively or negatively, to the outcome. We now highlight the single question most associated with each category (reworded from the survey only for clarity):

- **GPA**: To Child, Year 9: *"What is your satisfaction with life overall?"*
- **Grit**: To Primary Care Giver, Year 3 *"Has your child had an eye injury?"*
- **Material Hardship**: To Father, Year 5: *"In what year did you last work a regular job for at least 2 weeks to receive a regular paycheck?"*
- **Eviction**: To Father, Year 5: *"At your primary job, do you regularly work weekdays?"*
- **Layoff**: To Mother, Year 3: *"Have you completed nursing school?"*
- **Job Training**: To Father, Year 5: *"Did your job pay with cash or something else?"*

These results are intriguing: We see that GPA is associated with personal happiness, Grit is associated with childhood injury, Material Hardship, Eviction, and Job Training are associated with the father's occupational habits, and Layoff is associated with the mother's educational attainment. We also see that year 3 and 5 questions were the most correlated.

6 Conclusion and Future Work

We have shown that an Elastic Net Regression model is suitable for predicitng the six key outcome variables of the Fragile Families Challenge, and that by traing jointly on all six outcomes the prediction capabilities of such a model can be improved. The use of these particular models was justified theoretically and empirically, and we also highlight our model for data imputation using regression and one-hot encoding. We have also presented a brief interpretation of our models, and showcased the surprisingly relevant survey questions. In the future, one of the most useful steps to take would be a more in-depth analysis of the learned model correlations. Examining which questions are most predictive could help drive both policy changes and future survey design. Furthermore, it would be extremely useful to examine outliers according to these models to better determine which variables are insufficiently modeled in the current feature vectors.

References

[1] Fragile families challenge. `http://www.fragilefamilieschallenge.org`. Accessed: 2017-04-04.

[2] John B. Bell. *Mathematics of Computation*, 32(144):1320–1322, 1978.

[3] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[4] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[5] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[6] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[7] 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.

[8] David L Phillips. A technique for the numerical solution of certain integral equations of the first kind. *Journal of the ACM (JACM)*, 9(1):84–97, 1962.

[9] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[10] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.