COS 424 Homework 2

Malte Moeser
Department of Computer Science
Princeton University
mmoeser@princeton.edu

Samantha Weissman
Department of Computer Science
Princeton University
srw2@princeton.edu

Abstract

The Fragile Families Challenge is a mass collaboration experiment, intending to combine elements of social science and data science to gain new insights into the determinants for success of disadvantaged children in the US. In this context, this paper contributes a systematic assessment of different machine learning techniques in order to predict outcomes in the Fragile Families dataset. Using a combination of imputation, feature selection, and cross-validated model selection, we achieve prediction scores that are highly competitive among the current set of submissions.

1 Introduction

The Fragile Families and Child Wellbeing Study is a birth cohort panel study that follows a group of 5,000 children born in the US between 1998 and 2000. Waves of the study were conducted at birth and ages one, three, five, and nine, which consisted of interviews with parents, home visits and interviews with the child, and in the later waves, teacher interviews; a followup study was also conducted at age 15 [1]. The Fragile Families Challenge (FFC) is an attempt to use elements of social science and data science to better understand the Fragile Families data and gain insights that can influence policies and programs to better the lives of disadvantaged children in the US [6]. The challenge involves a mass collaboration experiment that encourages participants to use Fragile Families data from the first five waves and a training set from the age 15 followup to predict six key outcomes – grit, GPA, material hardship (MH), eviction, layoff, and job training (JT).

In this paper, we present a structured approach that compares different models and feature selection methods to determine the best method for predicting each of the six outcomes. Additionally, we inspect the features that contributed most significantly to the prediction of each outcome.

2 Data and Methods

The FFC dataset consists of 12,000 features for about 5,000 children from the first five waves of the study. We start by imputing missing values in the data, then perform three different types of feature selection, and train a set of models on each of the datasets.

2.1 Data Preprocessing and Imputation

The FFC dataset contains missing values both in the features as well as in the outcomes. Features may be missing when participants were absent from a wave, questions were skipped, or participants refused to answer; outcomes may be missing for similar reasons. For example, the FFC blog reports that the GPA outcome is reported as NA when children "were not interviewed, reported no grade, refused to answer, did not know, or were homeschooled" [5].

We clean the dataset and impute missing values as follows. First, we remove all ID and character features. After that, we code skipped and refused answers through new binary feature vectors. Then,

we impute all missing values by replacing the categorical values with the mode, and numerical values with the mean of each respective feature. Next, we remove features that contain only a single value, as well as highly correlated features (when two features have a Spearman correlation > 0.6, we remove the first feature). Finally, we convert all categorical values into indicator variables and scale continuous variables to have a mean of 0 and unit variance.

2.2 Feature Selection

Even though we already remove highly correlated features, the preprocessed and imputed dataset still contains 17303 features. At only 4242 observations, this is more than four times than the number of observations (i.e., the dataset is high-dimensional). This poses two major challenges: it increases the computational effort to fit the classifiers, and it may lead to overfitting as many of the features may contain more noise than signal.

We employ two strategies to reduce the dimensionality of the data while preserving the interpretability. This allows us to later reason about the features selected (this means we ignore dimensionality-reduction techniques such as PCA). Our first two approaches are based on selection of features from models that employ $l1$-regularization in order to find a sparse vector of feature weights. We use Lasso regression and elastic net regression. For both, we use cross-validation to find the best hyperparameters as available in the `LassoCV` and `ElasticNetCV` classifiers in scikit-learn. Second, we use cross-validated recursive feature elimination (RFE) based on a support vector machine with a linear kernel and a step size of 5 (each iteration removes the 5 least important features).

2.3 Classification and Regression Methods

We use the following classifiers and regressors (as implemented in scikit-learn version 0.18 [7]).

- *AdaBoost:* We use AdaBoost based on decision trees with 200 estimators
- *Gaussian Process:* We use Gaussian processes with up to 10 restarts
- *Linear:* We use cross-validated elastic net regression for continuous outcomes and cross-validated logistic regression for categorical outcomes
- *Random forest:* We use a random forest with Gini impurity and 200 trees
- *Support vector machine (SVM):* We optimize hyperparameters as described below, and compare it with a baseline model using a linear kernel and default parameters

2.4 Cross-Validation and Hyperparameter Tuning

For many of the models we used, scikit-learn provides classes that will perform cross-validation to find optimal hyperparameters. We use these classes with 10-fold cross-validation (stratified for binary outcomes). For the SVM, we use a grid search and cross-validation to select the best hyperparameters. We evaluate both a linear and a Gaussian kernel (rbf) and a suitable range of parameters. The best hyperparameters are reported in Appendix A.1 in the appendix.

3 Lasso Regression

A recurring theme in social science as well as in data science is the problem of high-dimensionality. When low storage combined with scalable means of collecting large amounts of information meet a scientists desire to capture as many potentially relevant factors in a study as possible, we often end up with datasets that contain vastly more features (i.e. properties) than observations.

Simple regression models, such as ordinary least-squares regression, are ill-suited to provide consistent predictions since they try to use all the information in order to build a model. This can lead to overfitting as well as imprecise predictions.

Lasso regression [8] is a regression model that tries to improve upon these issues by imposing a penalty on the number of feature weights that are not equal to zero. Given a vector of outcomes y, a matrix of features X and a vector β of feature weights, we can express the problem of Lasso

Table 1: Number of features selected for each outcome based on different selection strategies

Dataset	GPA	Grit	Material Hardship	Eviction	Layoff	Job Training
Elastic Net Model	93	11	135	112	42	3
Lasso Model	89	11	131	110	42	3
RFE (with SVM)	8	1	38	8	1	3

regression as finding the weight vector $\hat{\beta}$ such that the combined sum of prediction error and sum of absolute weights is minimized.

$$\hat{\beta} = \arg\min \sum_{i=1}^{n} (y_i - \sum_{j=1} \beta_j x_{i,j})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1}$$

Lasso is also referred to as $l1$-regularization, as $\sum_{j=1}^{p} |\beta_j|$ represents the l_1 norm that allows us to approximate the l_0 norm, which represents the number of non-zero elements – intuitively the measure we would like to minimize to create sparse feature weight vectors [2]. The choice of λ in Equation 1 allows us to specify the weight this normalization is given; larger values of λ lead to less non-zero values β_j.

We can compute possible Lasso solutions using Forward Stepwise Regression (FSR) [2]. Starting with a model that does not contain any non-zero weights, the impact of each feature on the quality of the model is evaluated using a measure such as the Akaike or the Bayesian information criterion. The feature that improves the model the most is added, and this procedure is repeated until the model no longer improves in quality. Based on these models, we then compute $\hat{\beta}$.

Lasso regression is closely related to Ridge regression, which uses the l_2 norm $\sum_{j=1}^{p} \beta_j^2$ instead of the l_1 norm as a penalty. As a consequence, Ridge regression achieves weight vectors with small coefficients instead of sparse vectors. Both types of regularization can be linearly combined into elastic net regularization, which has a number of benefits over pure Lasso regularization, such as it's enhanced ability to deal with groups of highly correlated features [3, p. 662 ff.]. In this paper, we employ both techniques for feature selection and use elastic net regression as our linear regressor.

4 Results

4.1 Feature Selection

In Table 1 we report the number of features that remain after performing feature selection. Recursive feature elimination selects the sparsest vector of features; the outcomes for grit and layoff are especially noteworthy as for both outcomes only one feature remains.

We briefly inspect the features in those cases where 3 or less features were chosen by the model. For the grit outcome with RFE, the feature reports on the frequency the child feels jealous about how their dad treats his other children (question code `k5c5`). For the layoff outcome with RFE, the feature refers to the importance for the mother's boyfriend to provide regular financial support to children (`k5c5`). For the job training with RFE, the features selected were whether the child participates in a "gifted and talented" program (`p5l13f`), whether the father refused to answer how much he earned (`f1j8a`), and the reason for why legal paternity has not been established (`f4c10a1`). For the job training outcome in the lasso and elastic net model, the features selected were the highest grade the mother has achieved in school (`m1i1`), how many regular jobs she had for 2 or more weeks (`m3k22`) as well as the amount earned from all regular jobs in the past 12 months (`m5i19a`). Clearly, many of these features relate to the outcomes of interest.

4.2 Prediction

Table 2 contains the evaluation of our models as determined by the FFC scoreboard. We can see that the full dataset does not, in general, produce worse results than the reduced data sets, and in many

Table 2: Scores as retrieved from the FFC scoreboard.

	Regression			Classification		
Data / Classifier	GPA	Grit	MH	Eviction	Layoff	JT
Full						
Adaboost	0.37936	0.22681	0.03303	0.22129	0.24538	0.24786
Gaussian Process	0.39273	0.21997	0.02880	0.25000	0.25000	0.25000
Linear	0.37892	**0.21717**	0.02701	0.05428	**0.17288**	*0.19962*
Random Forest	**0.37340**	0.21750	**0.02619**	0.05294	0.17708	0.20994
SVM (baseline)	0.50996	0.30100	0.02821	*0.05289*	0.17436	0.20343
SVM (tuned)	0.37776	0.21780	0.02858	0.05299	0.17439	0.20267
Elastic Net						
Adaboost	0.38212	0.23315	0.03969	0.22842	0.24768	0.24915
Gaussian Process	0.39324	>1	0.02883	0.24762	0.24528	0.20391
Linear	0.44535	*0.21771*	0.02867	0.05619	0.19901	**0.19960**
Random Forest	*0.38182*	0.25335	*0.02750*	**0.05231**	0.19710	0.22639
SVM (baseline)	0.46548	0.22480	0.03037	0.05319	0.18330	0.20202
SVM (tuned)	0.40947	0.21877	0.02751	0.05331	*0.18231*	0.20650
Lasso						
Adaboost	0.38321	0.23315	0.04085	0.22842	0.24768	0.24915
Gaussian Process	0.39320	> 1	0.02882	0.24760	0.24528	0.20391
Linear	0.44580	*0.21771*	0.02867	0.05688	0.19901	**0.19960**
Random Forest	*0.38099*	0.25335	*0.02755*	0.05296	0.19710	0.22639
SVM (baseline)	0.46755	0.22480	0.03039	0.05344	0.18733	0.20226
SVM (tuned)	0.42459	0.21877	0.02759	*0.05291*	*0.18181*	0.20227
RFE						
Adaboost	*0.40445*	0.23959	0.03291	0.24564	0.24907	0.25014
Gaussian Process	0.41285	0.22179	0.03239	0.05271	*0.17415*	0.20541
Linear	0.41161	*0.22178*	0.02815	*0.05270*	0.17435	0.20555
Random Forest	0.40804	0.22184	0.03117	0.05528	0.17427	0.20729
SVM (baseline)	0.41605	0.22384	*0.02717*	0.05698	0.17435	0.20224
SVM (tuned)	0.42098	0.22384	0.02718	0.05341	0.17435	*0.20223*

Regression score is squared error loss, classification score is Brier loss [4]
Italic values are best among a single dataset, **bold values** are best among all datasets.

cases contains some of the best scores among all datasets. On the other hand, the reduced datasets often only perform slightly worse, even the RFE data that often selected only few features. This means that in fact only a handful of features might be most relevant to predict these outcomes, and whether the the additional features provide better predictions may depend on the model's ability to prevent overfitting (e.g., the linear models include cross-validation as well as $l1$-regularization in the case of the elastic net regression). Furthermore, the reduced datasets enable training the models substantially faster than the full dataset.

Inspecting the performance of different types of classifiers and regressors, we see that the linear models as well as the random forest perform consistently well among all data sets. The effectiveness of the hyperparameter optimization for the SVM (i.e. comparing the baseline SVM with the tuned SVM) depends on the type of the problem (i.e. it seems more effective for regression) and also varies between the different data sets. It is also noteworthy that the AdaBoost classifier produces class probabilities around 50%, whereas the other classifiers produce predictions from 0-100%. Due to the evaluation with the Brier loss, this makes this classifier perform comparably worse than the others.

Comparing our scores to the current FFC scoreboard (as of 3/27/2017), we see that our best predictions are amongst the top predictions in the challenge so far. In particular, we hold rank 6 for GPA, rank 1 for grit, rank 7 for material hardship, rank 4 for eviction and layoff, and rank 1 for job training.

Table 3: Top five most important features for each outcome

Outcome	1	2	3	4	5
GPA	**hv5_ppvtss**	**f1i1**	**hv5_wj10raw**	**p5q3bw_1**	**hv3m2b_2**
Grit	hv5_ppvtss	f3k25e	hv5_wj9raw	**k5g1b_3**	**p5q1n**
MH	**m5f23e_1**	**m5f23e_2**	**m5f23k_1**	**m5g0_1**	**m4i8c2**
Eviction	f4r1_1	f3k25c	m4j13a	m4b8a	hv3j7
Layoff	hv3a23	m1f15_1	hv4pverr	cmf5fevjail_1	m4c29
JT	**m3k19**	**m5i3b_1**	**m5i2_15_0**	**m4k3a_12**	m3f2c3

4.3 Feature Importance

Next, we take a further look at the most important features in predicting each of the six outcomes based on a Random Forest. Table 3 contains the codes for the top five features for each outcome (the corresponding questions are listed in Appendix A.2).

Features in bold "make sense" as strong predictors for the given variable. GPA, MH, and JT had the most intuitive top predictive features. For GPA, the top predictors correspond to scholastic aptitude and intelligence test scores, father's education level, and the child's ability to focus. For MH, each of the top five predictive features relate to the mother's financial situation with respect to implications on the household (e.g. paying for gas/oil/electricity, telephone service, food stamps), all of which are reasonable predictors for a child's material hardship as many children in the study come from broken families and likely live with primarily their mothers. For JT, top indicators include questions regarding mother's earnings and participation in job training programs, which indicate a priority placed on job training as well as a child's ability to access job training resources.

Identifying feature importance with RF brought out different results than those from our implementation of feature selection – we did not find any common features between the results of the different methods. However, there were some similarities between the results of the two implementations (e.g., the features for layoff in RFE and RF included questions about child support, the features for job training in lasso/elastic net and RF included comparable questions about mother's earnings).

5 Discussion and Conclusion

In this paper we described a systematic approach towards identifying relevant features for predicting outcomes in the Fragile Families Challenge. We evaluated three different feature selection methods and, compared against other participants, achieve highly competitive results.

We plan to extend this work in multiple ways. While we started to improve upon the provided imputation script, better imputation strategies (e.g., incorporating textual features and using multiple imputation) may help to achieve better scores. Some of the classifiers used show strange behavior during prediction, and we will further evaluate the underlying reasons. For example, the Gaussian process sometimes produces high error rates for the grit outcome. To improve predictions, it would be possible to aggregate all the individual classifiers into one single prediction using ensemble methods such as stacking.

Finally, we must consider how much further we can improve our predictions with "dumb" machine learning techniques while ignoring the underlying sociological phenomena in the data. We suspect that at some point, more drastic improvement to our models will come not from applying better machine learning techniques to the data, but rather by integrating social scientist-like approaches to the problem and actually getting a better understanding of the features.

Acknowledgments

Our username on the FFC scoreboard website is `samantha_malte`.

References

[1] *About the Fragile Families and Child Wellbeing Study*. 2017. URL: http://fragilefamilies.princeton.edu/about/ (visited on 03/27/2017).

[2] Barbara Engelhardt. *Regularized Regression*. COS 424/SML 302: Fundamentals of Machine Learning – Lecture Notes. 2017.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York, 2009.

[4] Ian Lundberg. *Evaluating Submissions*. 2017. URL: http://www.fragilefamilieschallenge.org/evaluating-submissions/ (visited on 03/27/2017).

[5] Ian Lundberg. *GPA*. 2017. URL: http://www.fragilefamilieschallenge.org/gpa/ (visited on 03/27/2017).

[6] *Overview of the Fragile Families Challenge*. 2017. URL: http://www.fragilefamilieschallenge.org/ (visited on 03/27/2017).

[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". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[8] Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.

A Appendix

A.1 Optimal Hyperparameters for SVM Regression and Classification

Data / Feature	Epsilon	C	Tolerance	Gamma	Kernel
Lasso					
GPA	0.2	1	0.001	0.001	rbf
Grit	0.2	5	0.001	0.0001	rbf
Material Hardship	0.1	2	0.001	0.0001	rbf
Eviction		5	0.0001		linear
Layoff		2	0.001	0.01	rbf
Job Training		0.1	0.001		linear
Elastic Net					
GPA	0.1	5	0.001	0.0001	rbf
Grit	0.2	5	0.001	0.0001	rbf
Material Hardship	0.1	2	0.001	0.0001	rbf
Eviction		5	0.001		linear
Layoff		2	0.001	0.01	rbf
Job Training		0.1	0.001		linear
RFE					
GPA	0.5	2	0.001		linear
Grit	0.1	2	0.001	0.01	rbf
Material Hardship	0.1	5	0.001		linear
Eviction		0.1	0.001		linear
Layoff		0.1	0.001		linear
Job Training		1	0.001		linear
Full					
GPA	0.1	0.5	0.001	0.0001	rbf
Grit	0.1	0.2	0.0001	0.0001	rbf
Material Hardship	0	0.5	0.00001	0.0001	rbf
Eviction		0.1	0.001		linear
Layoff		0.1	0.001		linear
Job Training		0.1	0.001		linear

A.2 Selected Feature Codes

Code	Question	Outcome
cmf5fevjail_1	Constructed - mother report, father has spent time in jail by mother nine-year interview	Layoff
f1i1	Highest grade of regular school that you completed?	GPA
f3k25c	In past year, how many hrs/week did you work at own business?	Eviction
f3k25e	How much money did you receive in past 12 months from your business?	Grit
f4r1_1	Agree/disagree? - My religious faith is an important guide for my daily life	Eviction
hv3a23	Counting yourself, how many people in your house smoke?	Layoff
hv3j7	Spanked child on the bottom with your bare hand	Eviction
hv3m2b_2	Child can't sit still, is restless or hyperactive	GPA
hv4pverr	Total number of errors (child's PPVT)	Layoff
hv5_ppvtss	PPVT standard score	GPA, Grit
hv5_wj10raw	Woodcock Johnson Test 10 raw score	GPA
hv5_wj9raw	Woodcock Johnson Test 9 raw score	Grit
k5g1b_3	Even when a task is difficult, I want to solve it anyway	Grit
m1f15_1	If BF doesn't want to marry the mother, could he be required to pay child supp.?	Layoff
m3f2c3	What is third person's age? (years)	JT
m3k19	How much did you earn from all regular jobs in past year?	JT
m4b8a	What type of program does child attend most?	Eviction
m4c29	How much child support you receive from other father(s) in past 12 months?	Layoff
m4i8c2	Approximately how much did you receive last month from food stamps?	MH
m4j13a	About how much weight did you gain/lose during these two weeks	Eviction
m4k3a_12	What program/school completed: other type of training	JT
m5f23e_1	Did not pay full amount of gas/oil/electricity bill in past 12 months	MH
m5f23e_2	Did not pay full amount of gas/oil/electricity bill in past 12 months	MH
m5f23k_1	Telephone service disconnected because wasn't enough money in past 12 months	MH
m5g0_1	How satisfied you are with your life overall	MH
m5i2_15_0	Range of amount earned from all regular jobs in past 12 months	JT
m5i3b_1	You have taken classes to improve job skills since last interview	JT
p5q1n	Parent has called child dumb/lazy/some other name like that	Grit
p5q3bw_1	Child is inattentive or easily distracted	GPA