Assignment 2: Fragile Families Challenge

Peter Chen
Princeton University
xi@princeton.edu

Brandon Lanchang
Princeton University
lanchang@princeton.edu

Abstract

As a part of the Fragile Families Challenge, this assignment allowed us to use multiple machine learning techniques to explore how best to use a given set of data to predict both continuous and binary outcomes. For continuous outcomes, we used regularized linear regressions (Ridge/Lasso), and for binary we used the logistic regression. We tested several supported theories in the social science field regarding the correlation between certain features and our 6 target outcomes (grit, GPA, material hardship, eviction, job loss, and job training). We looked at these subsets to see which theories the Fragile Families Challenge best supports and provides us with the most accurate predictions.

1 Introduction

For this project, we will be using survey data from The Fragile Families & Child Wellbeing Study, which follows a cohort of nearly 5,000 children born in large U.S. cities between 1998 and 2000 roughly three-quarters of whom born to unmarried parents. We use the data from birth to year 9 and some training data from year 15 to predict six key outcomes in the year 15 test data (continuous outcomes grit, GPA, material hardship and binary outcomes eviction, job loss, and job training). We will be only using numeric explanatory variables to simplify the analysis.

2 Related Work

Research attempting to find correlation between varying forms of success and childhood nurturing/household factors are very much ongoing. The Fragile Families & Child Wellbeing Study is a big part of that research, compiling data across families to take into account as many factors as possible. However, some factors may see a greater correlation to the outcomes we seek to predict. It is a possible strategy to attempt to model based off as many factors given, but it is also a proper strategy to isolate factors that may appear more promising due to recent theories of sociology. Therefore, investigating this research may help us tailor our model to more accurate results.

In this assignment, we took it upon ourselves to help test varying theories of correlation between factors found in the given data and the outcomes we seek to predict. One method could be to isolate responses by the child, more specifically regarding their current academic performance, integration, comfort, etc. This includes measures of their academic success as well as less obvious features such as levels of bullying. Though often found more at a collegiate levels, these factors tend to have a significant role in future success regarding the child. This implies isolating responses from later waves though. Despite the possibility of less correlation between this data and more parental focused outcomes like eviction, there could be a strong connection between data and outcomes like GPA. [8]

Another theory to test is on the effects of poverty on our target outcomes. Unlike the previous feature set, intuition and research implies that household poverty greatly impacts student success as well as parental success. Students with households with lower income and socioeconomic standing

suffer from higher drop out rates. There families additionally experience higher rates of negative outcomes like eviction. This set could help improve results for multiple outcomes extending from student-dependent factors like GPA to more parental-dependent factors like eviction and job loss. [1][9]

One final research-supported theory we investigated is how self-reported data can include a bias that would inhibit results and how teachers often have a significant impact on students even from a young age. Researchers across the globe have questioned the validity of self-reported data in instances from drug use to weight. Disregarding human error of self-reporting, if the subject is reporting a negative feature they may be more inclined to provide a biased answer. Such could be the case with features reported in this data set regarding employment, financial circumstances, abuse, etc. Since we want to test results with data less prone to bias, teacher respondent data provides a possible solution. This additionally calls back to the theory that prior academic success and comfort leads to futures success since teacher respondent data answers questions of misbehavior and performance of the child in school as well. Using widespread thought and research as motivation for our subsets of data, we have direction for us to work towards as we attempt to create more accurate predictors. [6][10]

3 Methods

3.1 Missing Data

Missing data comes in the form of item non-response, where respondents refuse to answer a survey question, or survey non-response, where respondents cannot be located or refuse to answer any questions in an entire wave of the survey. We remedy this issue through single imputation using the mode, where we replace missing values with the most common response of that variable (1 if negative or mode not available). [7]

3.2 Motivated Feature Selection

Heavily referencing papers discussed in the Related Work section, we tried several approaches of feature selection to create three separate data subsets to train and test our models that are described below. We were able to use the filtering tool on the Fragile Families website to copy a list of variable names related to each subset we wanted to build. From there we wrote the filterData.py script so that we could use the list of variable names to drop columns we deemed unrelated to the tested theory and craft a new csv file of data to feed into our model. We did this three times to create our feature sets for child respondent data, data related to poverty and finances, and 'unbiased' data from teachers.

3.3 Cross Validation

We implement K-fold cross validation in all our prediction models, setting K = 10 as commonly used in practice. In doing so, we partitioned our data randomly into K equal disjointed subsets, then for for i = 1, 2, ..., K, we let fold i be the test fold to be held out, fir our model on the other K-1 folds, predict on the test fold, and compute generalization error for each sample (i.e. if we were taking a look at R^2 values, we would end up with K=10 separate R^2 values, one for each subset). [4]

Another important use of k-folds is for tuning hyperparameters. We will be using doing this for α in Ridge and Lasso regularized regressions. We do this by performing the following:

- Split training data into k folds: $S_1, S_2, ..., S_k$
- For hyperparameter C in $C_1, C_2, ...C_m$
  - For i = 1,2,...,k, fold S_i, fit on the remaining k-1 folds, and predict on S_i
  - Compute generalization error E_c from one prediction
- Pick $C* = argmin_c E_c$

[5]

3.4 Linear Regression

To predict for continuous variables 'gpa' 'grit' 'materialHardship', we first use a linear regression. The ordinary least squares regression fits a linear model with coefficients $\hat{\beta}$ to fit a linear relation ship of x_i's on y's in the training data. The equation will be in the following form:

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + ... + \beta_m * x_m$$

where we calculate the estimate $\hat{\beta}$ by minimize the residual sum of squares between the observed and predicted responses, represented as:

$$min||(\beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + ... + \beta_m * x_m) - y||^2$$

We can then use the predicted $\hat{\beta}$'s to predict for $\hat{y}$'s using x's from the testing data.

$$\hat{y} = \hat{\beta_0} + \hat{\beta_1} * x_1 + \hat{\beta_2} * x_2 + ... + \hat{\beta_m} * x_m$$

We evaluate the model with predicted r-squared r^2, which measures correlation between predicted and true response values with higher values (max of 1 which indicates perfect prediction) indicating better predictions. It is calculated as follows:

$$r^2 = 1 - \frac{\text{Regression Sum of Squared Error}}{\text{Total Sum of Squared Error}} = 1 - \frac{\sum(y_i - \hat{y_i})}{\sum(y_i - \bar{y_i})}$$

[2]

3.5 Regularization: Ridge and Lasso Regressions

To improve our predictions, we penalize higher degree polynomials to reduce overfitting, a process called regularization. We do this via ridge and lasso regressions.

In a ridge regression, we apply a L2 (least squares error) penalty. As a result, we prefer small coefficients (but not zero) in this regression. Our objective can thus be formulated as:

$$min||Y - X\beta||_2^2 + \alpha||\beta||_2^2$$

Our α is a regularization/shrinkage parameter and controls the size of the coefficient (the strength of regularization). As α approaches infinity, our coefficients approach 0, and as α approaches 0, our coefficients approach that of OLS.

In lasso regression, we apply a L1 (least absolute deviation) penalty. As a result, we prefer sparsity, which can lead to some zero coefficients.Our objective can thus be formulated as:

$$min||Y - X\beta||_2^2 + \alpha||\beta||_1$$

Our α is a regularization/shrinkage parameter and controls the size of the coefficient (the strength of regularization). As α approaches infinity, our coefficients approach and become 0 (not the case in ridge), and as α approaches 0, our coefficients approach that of OLS.

We select our α in each case by using cross-validation methods noted above. [2]

3.6 Logistic Regression

To classify binary variables ('eviction' 'layoff' 'jobTraining'), we use a logistic regression. The logistic regression model is similar to OLS in that it maximizes the likelihood of observing the sample values we use the train the model, as follows:

$$log\frac{P(y_i = 1|x_i)}{1 - P(y_i = 1|x_i)} = \beta^T x_i,$$

$$P(y = 1|x_1, ..., x_k) = \frac{exp(\beta_0 + \beta_1 x_1 + ... + \beta_m x_m)}{1 + exp(\beta_0 + \beta_1 x_1 + ... + \beta_m x_m)}$$

We evaluate the model using accuracy, which is simply the percentage of predictions that are correct. [3]

3.7 Bootstrap

We use a bootstrapping method to resample the data and calculate confidence intervals on our predictions. Using the ordinary bootstrap method, we perform the following:

- Start with data set $D = (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$.
- Generate K (=1000) bootstrap samples $D_1, D_2, ..., D_k$, each sample has n (=10000) observation pairs randomly sampled from D with replacement.
- Build a model on each bootstrap sample D_i and test using the oirginal dataset D.

It should be noted that the prediction error is underestimated using this method as many observations are used in both training and testing.

4 Results

As an initial test, we ran an OLS regression on all available numeric variables to predict for 'gpa' 'grit' 'materialHardship', obtaining the average r^2 over 10 folds seen in Table 1. Similarly, we ran a logistic regression on all available numeric variables to classify 'eviction' 'layoff' 'jobTraining', obtaining the average accuracy over 10 folds seen in Table 2. Though these results are not terrible, the regression took a very long time to fit given the amount of variables. This is one reason that subsets of the data can benefit our model.

After filtering the data to create our other three subsets (child data, financial data, and teacher data), we ran the subsets through our extended model which yielded the rest of the results covered in Table 1 and 2. You may notice that results for teacher data are not included with these tables. Their results were for the majority subpar to the other subsets with the exception of a few anomalous extremely negative values. Without any indication of what caused these, we found it best to omit them from the final analysis.

Using the rest of the results we decided our best subset was the child data. We then used the bootstrapping technique on that subset and recorded the values found in tables 3 and 4. We also submitted this data and our model to the Fragile Families Challenge website. This gave us the results seen in Table 5, which shows the mean squared error metric the Fragile Families Challenge uses to rank submissions.

		All data	Child data	Financial data
OLS	GPA	-4.3569	-0.0142	-6.6903
	Grit	-4.7375	-0.0044	-12.2823
	Material Hardship	-4.0518	-0.0530	-10.9238
Ridge	GPA	-	0.0586	-0.4057
	Grit	-	0.0281	-0.3259
	Material Hardship	-	0.0057	-0.2167
Lasso	GPA	-	0.0371	0.0625
	Grit	-	0.0350	0.0012
	Material Hardship	-	-0.0029	0.0679

Table 1: Table of r^2 values given varying sets of data run through OLS, with ridge regression, and with lasso regression.

		All data	Child data	Financial data
Accuracy	Eviction	0.8518	0.9342	0.8594
	Layoff	0.6348	0.7758	0.6819
	Job Training	0.6582	0.7486	0.6829

Table 2: Table of accuracy values given varying sets of data

	OLS mean	OLS Std Dev	Ridge Mean	Ridge Std Dev	Lasso Mean	Lasso Std Dev
GPA	0.1524	0.0013	0.1440	0.0012	0.0000	0.0001
Grit	0.1237	0.0014	0.1187	0.0013	-0.0001	0.0002
Material Harship	0.0955	0.0015	0.0900	0.0013	0.0000	0.0001

Table 3: Table of bootstrap values for child data subset across OLS, ridge, and lasso models

	Logistic Mean	Logistic Std Dev
Eviction	0.9342	0.0012
Layoff	0.7947	0.0028
Job Training	0.7707	0.0027

Table 4: Table of bootstrap values for child data subset for logistic regression models

	GPA	Grit	Material Hardship	Eviction	Layoff	Job Training
Submission Results	0.39862	0.21277	0.02846	0.05660	0.22830	0.28679

Table 5: Table of mean squared error results as determined by the Fragile Families Challenge website

5 Discussion and Conclusion

We see that "Child data" gave us substantially better predictions than "Financial data" or using the entire dataset. This indicates that the features for child data quite possibly display a better pattern of correlation with the six incomes we are trying to predict. Something else to consider is the size of the feature set. "Child Data" had the least variables. It is a possibility that the other sets with their additional features also include more noise that masks any pattern our models are trying to identify.

For continuous outcomes (grit, GPA, material hardship), we ran ordinary least squares regressions and regularized using Ridge/Lasso. We notice that Lasso works better on material hardship, and Ridge works better on grit and GPA. From this, we can derive that material hardship in particular prefers sparsity of data in that it works better when certain subsets of features from the model are removed; grit and GPA on the other hand prefer small but nonzero weight vectors. The logistic regression performed well for all our binary outcomes (eviction, job loss, and job training), and eviction in particular. This indicates that our input space can be linearly separated into the two classes.

One extension that would have been particularly helpful given more time is studying the residuals more in depth to analyze why certain participants were difficult to predict (in particular, why do some participant perform way better than expected). Trying out more complex models (such an elastic net regularization that combines Ridge and Lasso) may also yield better predictions.

Acknowledgments

Prof. Barbara Engelhardt, Dr. Xiaoyan Li, who provided the lectures for COS 424.

References

[1] High school dropout and completion rates in the united states: 2007.

[2] sklearn.linear_model scikit-learn 0.19.1 documentation.

[3] sklearn.linear_model.LogisticRegression scikit-learn 0.19.1 documentation.

[4] sklearn.model_selection.KFolds scikit-learn 0.19.1 documentation.

[5] Jason Brownlee. How to tune algorithm parameters with scikit-learn, 2014.

[6] Lana Harrison. The validity of self-reported data on drug use, 1995.

[7] Ian Lundberg. Fragile families challenge: Missing data, 2017.

[8] Kirsten McKenzie and Robert Schwietzer. Who succeeds at university? factors predicting academic performance in first year australian university students, 2018.

[9] Justin Ready, Lorraine Mazerolle, and Elyse Revere. Getting evicted from public housing: An analysis of the factors influencing eviction decisions in six public housing sites, 1998.

[10] RJ Roberts. Can self-reported data accurately describe the prevalence of overweight?, 1995.