Assignment 2: Fragile Families Challenge

Ji Won Shin
Department of Computer Science
jwshin@princeton.edu

Prakhar Kumar
Department of Electrical Engineering
prakhark@princeton.edu

Abstract

The project deals with the Fragile Families Challenge which focuses on a group of about 5000 children. The ultimate aim of the challenge is to address policy questions for child well-being and welfare. This challenge aims at leveraging predictive data analysis for real-world social science experiment. The background dataset consists of questions asked to the families from the child's birth year to age 9. This background information forms a large feature set of about 12000 dimensions. Using this background data along with some training data for age of 15, we aim to predict the six key outcomes for the test data of those of age 15. The high dimensionality of the training data along with limited number of training examples requires two-fold feature selection techniques Principal Component Analysis (PCA) and Lasso. Other regression methods such as ELasticNet & support vector regression have also been implemented. The project was extended to implement Bootstrapping in order to get an estimate for the confidence intervals on the MSE of the regression models. This work aims to estimate the performance in terms of GPA for the children at the age of 15.

1 Introduction

Fragile Families Challenge is an inter-disciplinary challenge which aims to take advantage of data science skills and apply them to the real-world social science tasks. This challenge relates to predicting the performance of around 5000 children from families labeled as 'fragile'. Their performance is evaluated on the basis of six key outcomes (in no particular order): GPA, grit, material hardship, eviction, layoff and job training. The first three outcomes have been modelled as continuous variables while the last three are considered as binary , that is either True or False. GPA and grit take values between 1.0 and 4.0, while material hardship ranges from 0 to 1 with higher value indicating greater hardships. One of the major challenges with social science datasets is missing values. Therefore, in order to deal with them, the background data had to be cleaned by using imputation methods. After proper imputation, dimensionality reduction technique using PCA was adopted and regression methodologies like Lasso, ElasticNet and Support Vector Regression (SVR) were implemented to predict the outcomes of GPA. The algorithms were implemented using the SciKit-Learn Python Library [1].

2 Related Work

Many social and data scientists are working towards tackling the Fragile Families Challenge, but "a huge proportion of the variance remains buried in the error term of regression models" [2]. Among many regression methods, in [3], Zou and Hastie showed that elastic net resulted in about 24% lower prediction error than lasso regression. They also proposed that the accuracy of lasso regression decreases when the predictors are highly correlated.

3 Methods

3.1 Dataset and Features

The dataset consists of sample points for 4242 families with 12495 features. The features correspond to the answers to numerous survey questions asked to each family over the period of 15 years in 5 waves. The labeled training data consists of the 6 key outcomes for 2121 families for the year 15 while the test data corresponds to the remaining 2121 families.

3.2 Data Processing and Imputation

Imputation refers to the process of replacing missing values in a dataset with appropriate values. Since, social science datasets generally suffer from the problem of missing data values, the fragile family dataset also had to be imputed for numerous missing values. This was done using the mean-imputation method by replacing the missing feature value with its average across all samples. Moreover, on closer analysis of the data it was observed that the dataset contained certain fields such as "mothid" and "fathid" (ID numbers for the parents) and entries for year and month of interview. Since they do not contribute as characterizing features, these specific columns were removed from the dataset table.

3.3 Dimensionality Reduction

In order to deal with large number of features, Principal Components Analysis (PCA) was used for dimensionality reduction. PCA is an orthogonal transformation that converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. [4] Based on the variance analysis, a subset of the projected features was selected, thereby reducing the dimensionality of the features to the more relevant ones as discussed in Section 4.2

3.4 Bootstrapping

The analysis of the models developed was extended to include the estimation of confidence intervals for the mean squared error of these models. For this purpose, we used bootstrapping technique, where the training data was randomly divided into a training set and test set by sampling from the original training dataset. This process was repeated 100 times to yield 100 different train and test sets. The statistics of the mean squared error was then analyzed over these 100 test sets as described in Section 4.3.

3.5 Regression Methods

3.5.1 Lasso

Given the extremely large number of features, we first performed PCA based dimensionality reduction as described in Section 3.3. After selecting a subset of features based on their relative variance contributions, we further applied Lasso regression to predict the outcomes for the GPA. Lasso was implemented in conjunction with cross-validation so as to select an appropriate value for the penalty hyperparameter λ.

3.5.2 ElasticNet

Elastic Net combines both the ridge and lasso regression by using a linear combination of the L-2 and L-1 penalties. The mathematical formulation of ElasticNet objective function is given as:

$$J(\mathbf{w}) = \sum_{i=1}^{N} (y_i - (w_0 + \mathbf{w^T x_i}))^2 + \lambda_1 \|\mathbf{w}\|_{\mathbf{1}} + \lambda_2 \|\mathbf{w}\|^{\mathbf{2}} \quad (1)$$

Here λ_1 and λ_2 determine the amount of L-1 and L-2 penalties respectively. For this work, we

implemented ElasticNet using cross validation on the features processed by PCA so as to achieve the best fitted model along a regularization path.

3.5.3 Support Vector Regression (SVR)

Even though support vector machines (SVMs) are generally used for classification tasks, a variation of SVM can be used for regression tasks in the form of SVR. It uses the principle of maximum margin to identify the support vectors and builds a regression model with radial basis function kernel, which performs regression in a higher dimensional space. Given that kernels are adept at handling high dimensional data, it can be pre-empted to perform better than lasso and ElasticNet.

3.6 Spotlight Classifier: Lasso Regression

Lasso (Least absolute shrinkage and selection operator) is a shrinkage and selection method used for linear regression [5]. Lasso minimizes the sum of squared errors with a bound on the sum of the absolute values of the coefficients and introduces a penalty on the sum of L-1 norm of the coefficients. This penalty leads to many coefficients being set to zero leading to a sparse coefficient vector. The sparsity in coefficients results in inherent feature selection. Due to this characteristic, Lasso is especially useful for datasets with high-dimensional features and small number of sample points. Regularization allows complex models to be trained on datasets of limited size without suffering from overfitting. Since Lasso builds upon a standard a linear regressor, we first describe linear regression and then extend it to Lasso.

Consider the data given as $(x_i, y_i), i = 1, 2, \ldots, N$, where y_i is the response variable and x_i is the D-dimensional predictor variable (or simply features). The aim of basic linear regression is to find the coefficients **w**=$\{w_1, ..w_D\}$ to minimize the Residual Sum of Squares (RSS). If the predicted variable is given by $\hat{y_i}$, such that

$$\hat{y_i} = w_0 + \mathbf{w^T x_i} \tag{2}$$

, where w_0 is the intercept term, then the RSS is calculated as :

$$RSS(\mathbf{w}) = \sum_{i=1}^{N} (y_i - \hat{y_i})^2 = \sum_{i=1}^{N} (y_i - (w_0 + \mathbf{w^T x_i}))^2 \tag{3}$$

The coefficients **w** can be fitted to the data using the Maximum Likelihood Estimate (MLE) such that it minimizes the RSS. However, the problem with MLE is that can overfit especially in case of noisy data and those cases where the number of sample points is smaller than the feature dimensionality. In order to ameliorate this issue, we would prefer smaller weights. The two commonly used approaches are ridge regression and lasso regression. The difference between ridge and lasso is that the former introduces a L-2 penalty while the latter uses a L-1 penalty in our objective function. Therefore, the new objective function J(**w**) for Lasso that needs to be minimized is given as:

$$J(\mathbf{w}) = \sum_{i=1}^{N} (y_i - (w_0 + \mathbf{w^T x_i}))^2 + \lambda \|\mathbf{w}\|_\mathbf{1} \tag{4}$$

,where $\|\mathbf{w}\|_\mathbf{1} = \sum_{i=1}^{N} |w_i|$.

Here, λ controls the strength of the L-1 penalty such that when $\lambda = 0$, then it will be a regular linear regression. For $\lambda = \infty$, all the coefficients are set to zero. It should also be emphasized that the intercept term w_0 is not regularized by the objective function in equation 4. Moreover, it is necessary to ensure that the various predictor variables are on the same scale.
The solutions for the coefficients of Lasso regression are calculated using the quadratic programming methodology. The kinds of weights selected by the lasso regression can be seen from the estimation picture as shown in Figure 1. Assuming only two weights β_1 and β_2, it can be seen that the contours for the constraint functions and the error functions may intersect at a point where one of the weights is zero.

4 Results

4.1 Evaluation metrics

The performance of the various regression methodologies was analyzed using multiple evaluation metrics. We employed standard metrics such as mean-squared error (MSE), coefficient of

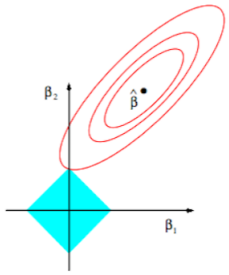

Figure 1: Estimation figure for lasso regression. The red ellipses represent the contours for the least squares error function. The blue region represents the region satisfying the weights constraint function.The weights are subject to the constraint $|\beta_1| + |\beta_2| \leq t$. [6]

determination(R^2), and the explained variance scores.

$$\text{MSE}(y, \hat{y}_i) = \frac{\sum_{i=1}^{N}(y - \hat{y}_i)^2}{N} \tag{5}$$

$$R^2(y, \hat{y}_i) = 1 - \frac{\sum_{i=1}^{N}(y - \hat{y}_i)^2}{\sum_{i=1}^{N}(y - \bar{y}_i)^2}, \bar{y}_i = \frac{\sum_{i=1}^{N}(y_i)}{N} \tag{6}$$

$$\text{Explained\_Variance}(y, \hat{y}) = 1 - \frac{Var\{y - \hat{y}\}}{Var\{y\}} \tag{7}$$

MSE measures the average of the squares of the difference between the estimated value based on regression and the actual value. Therefore, the closer MSE of a regression method is to 0, the better estimator the method is. R^2 is the square of the correlation between the actual value and the predicted value of a regression method. For both R^2 and explained variance regression score, value 1 signifies the best possible estimator while lower values signify less accurate models.

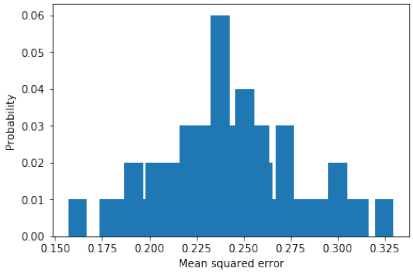

Figure 2: Probability distribution of MSE

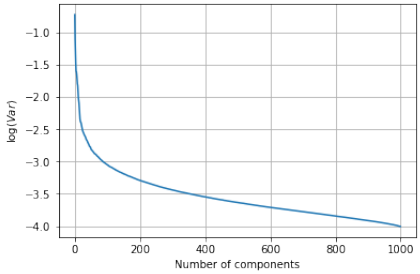

Figure 3: Proportion of variance contributed by top 1000 components

4.2 Scores from the FFC submission

GPA	Grit	Material hardship	Eviction	Layoff	Job training
0.58828 (32)	0.21997 (13)	0.02880 (21)	0.05341 (8)	0.17435 (7)	0.20224 (5)

(The username for the FFC website is prakhar.)

4.3 Analysis

Bootstrapping was performed as an extension to this project to gain an understanding of the performance of the regression model using the training data. The evaluation of performance for the estimators was done on the basis of a subset of the training set. The results for confidence intervals for the mean, variance, and standard deviation of the mean squared error obtained from bootstrapping have been summarized in Table 1. The probability density function of MSE obtained from using bootstrapping on ElasticNet model is shown in Figure 2.

Metric calculated using Bootstrap	5% value	Central value	95% value
Mean	0.2389	0.2448	0.2507
Std. deviation	0.0261	0.0299	0.0345
Variance	0.0006	0.0009	0.0011

Table 1: Confidence interval for mean, standard deviation & variance of mean squared error

Regression Model	Mean-squared error	Coeff. of determination	Explained variance score
Lasso	0.184	0.301	0.301
ELasticNet	0.185	0.5835	0.2984
SVM regressor	0.01403	0.9467	0.9467

Table 2: MSE, R^2, and explained variance score for different regression models

It can be seen that the 90% confidence interval for the average of the MSE lies in 0.2389 to 0.2507. Prior to applying Lasso regression, the first step was dimensionality reduction using PCA so as to aid in the successive feature selection by Lasso. A variance curve of the top 1000 components was used to select a subset of the entire component set. The log(variance) plot is shown in Figure 3. A log plot was chosen instead of a linear scale to highlight the variance contribution by the components.

The performance of the various regression models implemented for this project has been quantified using various metrics such as mean-squared error, coefficient of determination, and explained variance score. All the relevant statistics have been summarized in the Table 2.

It can be seen from Table 2 that Support Vector Regression out-performs the Lasso and ElasticNet models due to its ability to handle high dimensional feature spaces more efficiently. SVR reports a significantly lower MSE and higher R^2 than both Lasso and ElasticNet. Table 3 shows the top 8 features selected by Lasso regression. It can be observed that these features convey information about the economic status of a family and how it affects the education of a child, thereby inherently affecting his/her GPA.

Feature code	Description of feature in Codebook
m2l11	Do you have a bank account?
f5i4	You did regular work for pay last week ?
m3l6a	Can you rely on vehicle to get you to school/work/other?
m4l6a	Can you rely on the vehicle to get you to school, work, or other places?
m2g13	How often does child see your parents?
cf5md_case_con	Father meets depression criteria (conservative) at nine-year
m3i0l	How important is it: to vote in elections?
m5e3a	Mother could count on someone to loan her $1000 during the next year?

Table 3: Top 8 features selected by Lasso

5 Discussion and Conclusion

This work discusses different regression models used for predicting the performance of children in the Fragile Families Challenge at age 15 in terms of their GPA. The task poses many interesting problems such as missing data, high dimensional features, and limited number of sample points. Initially the data was cleaned by removing non-informative entries, such as mother's ID, father's ID, and month and year of interviews, followed by the second step of data imputation to remove any missing values. In order to build reliable models we followed a two fold approach for feature selection. The first step was dimensionality reduction using PCA to select only a subset of the components which contributed significantly to the variance. This reduced dimensionality feature set was then used for all regression purposes. Lasso inherently performs the feature selection by generating a sparse weight vector while ElasticNet tries to balance both the characteristics of lasso and general regularized regression. We carried out the evaluation of the regression models on the basis of mean-squared error, R^2 (coefficient of determination), and explained variance score.

As an extension to the basic analysis, we implemented Bootstrapping in order to estimate the confidence interval for the mean squared error (MSE) of the regression model. The statistics for the mean, standard deviation, and variance of the MSE along with their 90% confidence interval were analyzed. It can be concluded from the performance analysis of the models that Support Vector Regression is a significantly better estimator for GPA than Lasso and ElasticNet.

References

[1] 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.

[2] Ian Lundberg. GPA. Website blog.

[3] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[4] Wikipedia. Principal component analysis — wikipedia, the free encyclopedia, 2017. [Online; accessed 28-March-2017].

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

[6] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2016.