COS 424 Assignment 2: Fragile Families Challenge

Mihika Kapoor mkapoor@

April 15, 2017

Abstract

The Fragile Families Challenge is a study performed on little over 4000 families from inner cities that aims to determine how factors such as parental marital status affect children growing up. The families are referred to as fragile as many have single parents and live in impoverished conditions. This study was motivated to identify how well surveys during the developmental years of these children predict measures of their GPA, grit, material hardship, eviction, job training and layoff at age 15. Using several classifiers and regression models, the most accurate of which was the Random Forest Model, I found was able to predict material hardship, eviction, GPA, grit, layoff and job training with errors of .04249, .5660, .46216, .35875, .20404 and .23968, respectively.

1 Introduction

The Fragile Families Challenge [4] is based on the Fragile Families and Child Wellbeing Study, a study conducted by Princeton University and Columbia University of around 4242 American children who have grown up in cities across the country. Preliminary data of interviews with the mothers and fathers was taken from 1998-2000, with follow ups when the children were of ages 1, 3, 5 and 9. These interviews were primarily focused on development (emotional, health and cognitive) and living environment. The children were subsequently followed up with at the age of 15 in 2014. A large majority (around 75%) of these children have unmarried parents and the resulting domestic fragility that occurs as a result of this and the poverty in which many live is why these are referred to as fragile families.

This paper and study essentially aim to predict 6 factors/outcomes based on early data. There are 3 binary labels: eviction, job loss and job training and 3 continuous labels: GPA, grit and material hardship. I applied multiple classifiers, regression models and feature selection methods to devise a way to predict these outcomes at age 15. I was able to predict material hardship and eviction with the greatest accuracy, indicating that they were more heavily influenced by environmental and familial patterns of children in the surveyed families.

2 Methods and Related Work

2.1 Initial data and pre-processing

I downloaded the initial data from the Fragile Families website, which spanned a study of 4242 families, each with around 12,000 features describing them and their circumstances. I used the MissingDataScript.py file to impute the data and eliminate values such as N/A. I subsequently went through and eliminated all strings with the selectdtypes() method. This eliminated quite a few columns. Following that, I split background.csv into 2 files. 1 file, the important one, had the data from the rows where the challengeID matched those in train.csv.The columns in train.csv served as my training labels and the columns in the matchedbackground.csv served as the features I trained on.

2.2 Classification

I used the following classifiers for the binary data from the SciKit Learn library throughout our analysis tasks: **Naive Bayes** (NB): using multinomial implementation; **Support Vector Machine** with linear kernel (SVML); **Support Vector Machine** with Gaussian kernel (SVMG); **Logistic Regression** (LR); **Decision Tree** (DT); **Random Forest** (RF); **K Nearest Neighbors** (KNN).
I used the following regression models for the continuous data from the SciKit Learn library throughout our analysis tasks: **Random Forest** (RF); **Elastic Net**; **Lasso**; **Ridge**. Since RF performed best, and Ian indicated that probability predictions were better than binary predictions even for the binary labels (as the FFC website looks at mean squared error), I applied RF regression to those 3 as well.

2.3 Evaluation

2.3.1 Accuracy, Precision, Recall

For the binary labels, I compared the performance of the classifiers and features, and feature extraction using accuracy, precision and recall. Accuracy is a simple statistic that provides the fraction of correctly categorized instances (the ratio of total correct to incorrect classifications). Precision is the number of true positives/all predicted positives; recall is the true positive rate (true positives/all positives).

2.3.2 RMSE

For the continuous labels, I compared the performance of the classifiers using calculations of the RMSE (root mean squared error). The RMSE is a measure of the differences in the predicted values and the training value.

2.3.3 Cross validation

For the Random Forest classifier, which performed the best, I evaluated the efficacy of using cross validation techniques. Cross validation compares the class to the classifier

model, and the accuracy returns the mean score and 95% confidence interval across k=10 folds in the data.

3 Spotlight Model: Random Forest Classification

We consider a binary case of Random Forest Classification, the first algorithm of which was created by Tin Kam Ho of AT&T Bell Laboratories[3]. The model was later extended by Leo Breiman and Adele Cutler[2]. RF is an ensemble approach, similar in nature to a nearest neighbor predictor.

In my study, I found the the RF classifier yielded the most accurate results for the binary labels.

An ensemble approach is essentially predicated on the idea that individually, classifiers are weak, but together can work together to learn more effectively. RF begins with a decision tree, which is an example of a poor classifier or "weak learner." Trees are susceptible to a lot of inaccuracy, such as over fitting. RF averages several trees together by training on different parts, splits and folds of a data set, ultimately to reduce variance and increase accuracy.

RF employed bootstrap aggregating (bagging) to learners. Let's say you are given a training set X and testing set Y:

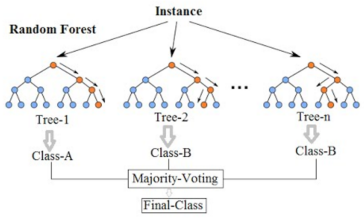
$$X = x_1, x_2, ..., x_n$$

$$Y = y_1, y_2, ..., y_n$$

Upon each iteration of bootstrapping (Assuming we bootstrap n times), sample n training examples from X and Y, and train on that decision tree d_t. Predictions of latent values (x') can be averaged using the following formula:

$$\frac{1}{B}\sum_{i=1}^{n} d_t(x')$$

Here is an visualization of the process:



[1]

One can subsequently test the efficacy of RF using cross validation, as I have to calculate confidence intervals and measure the accuracy.

4 Results

4.1 Binary Labels

4.1.1 Classification Models

In Table 2 we can see that the general recall, precision and accuracy for eviction were fairly high. The Random Forest classifier, with the 100 best features outperformed the rest in all 3 measures. SVM (Gaussian Kernel) and RF with 200 features performed nearly identically and second best. DT follows up next, which is interesting because RF is predicated on DT's. For layoff, we find that DT and RF performed the best, with

Classifier	Prec	Recall	Accuracy
NB	.93	.29	.287
LR	.91	.89	.894
SVM (Linear Kernel)	.91	.88	.878
SVM (Gaussian Kernel)	.91	.96	.955
DT	.91	.92	.915
KNN	.91	.89	.894
RF (200 features)	.91	.96	.955
RF (100 features)	.94	.97	.962

Table 1: **Results from classifiers on eviction label.**

RF with 200 features actually outperforming RF with 100 features, indicating layoff predictions improve with increased context. For job training, RF with 100 featured performed best overall.

	Layoff			Job Training		
Classifier	Prec	Recall	Accuracy	Prec	Recall	Accuracy
NB	.80	.31	.308	.79	.73	.729
LR	.78	.78	.781	.73	.75	.746
DT	.81	.77	.769	.73	.75	.746
KNN	.78	.78	.781	.73	.75	.747
RF(200 features)	.77	.88	.878	.68	.83	.826
RF(100 features)	.76	.87	.877	.71	.84	.844

Table 2: **Results from classifiers on layoff and job training.**

4.1.2 Cross Validation with RF Classification

Cross validating the data (Figure 3) indicated the accuracy and 95% confidence intervals of the data. Eviction performed the best.

n	Accuracy	Standard Deviation
eviction	0.96	+/- 0.03
layoff	0.88	+/- 0.03
job training	0.84	+/- 0.05

4.1.3 Non-binary predictions

Ian indicated that non-binary outcomes were actually preferred in the submission even for binary labels, given the way that the submissions were measured. I therefore also ran eviction layoff and job training with an RF Regression model and uploaded that. It outperformed the RF Classification for layoff and job training (see results below).

4.2 Continuous Labels

Here are the RMSE results when the 4 regression models were applied to two of the continuous labels, grit and GPA. Applying RF Regression techniques resulted in by far the least error. I also executed parameter tuning for different values of alpha for Elastic Net, Lasso and Ridge models. Here are sample values for alpha = .1 and .01. GPA predictions were better with the larger alpha while grit predictions were better with the smaller alpha. However, RF outperformed them all consistently, which led me to predict that it would ultimately predict the best results for the FFC website.

	alpha = .1		alpha = .01	
Model	grit	GPA	grit	GPA
RF Regression	.29238	.32807	.29238	.32807
Elastic Net	.55457	.58234	.44253	.62177
Lasso	.50830	.51613	.40651	.59454
Ridge	.52045	.39603	.37516	.49084

Table 3: **RMSE results from models on continuous labels.**

4.3 FFC Website Results

My username on the website was mkapoor and these were the results I got when uploading. Layoff and job training produced errors of .22453 and .27736 when using binary classification prediction models, instead of probabilities.

label	Measure	Model
eviction	0.05660	RF Classification
layoff	0.20404	RF Regression
job training	0.23968	RF Regression
GPA	0.46216	RF Regression
grit	0.35875	RF Regression
material hardship	0.04249	RF Regression

5 Discussion and Conclusion

In this assignment, I used various models to predict the outcomes of 6 characteristics measured at age 15 in the Fragile Families Challenge. Random Forest Classification and Regression produced results that matched the imputed data the most correctly. GPA and grit were the most difficult to predict accurately given the training data, indicating that these operate more independently of the nature/nurture speculations. Further, material hardship and eviction were easier to predict, indicating that circumstances and environments into which one is born do continue to affect children up until the age of 15. Finally job training and layoff were in the middle, yet their errors were highly correlated to each other, which was interesting.

There are several ways to improve results and conduct deeper analysis moving forward, despite the constraints of the sensitive data. First, I would like to use more thorough imputing scripts. I used several features such as selectdtypes() which just eliminated parts of the original data, which may have been critical to predictions. Additionally, it would be interesting to use ML techniques to split up the background data by person who answered (mother, father, child, etc) and see which groups answers was most predictive.

References

[1] Random forest based classification. https://www.youtube.com/watch?v=ajTc5y3OqSQ. Accessed: 2017-04-11.

[2] Leo Breiman. *Random Forests*. University of California, Berkeley, Berkeley, CA 94720, January 2001.

[3] Tim Kam Ho. *Random Decision Forests*. ATT Bell Laboratories, 600 Mountain Avenue, 2C-548C, Murray Hill, NJ 07974.

[4] Ian Lundberg and Matthew J. Salganik. Fragile families challenge. http://www.fragilefamilieschallenge.org/. Accessed: 2017-03-24.