Prediction of social outcomes of fragile families

Mayank Mahajan
Department of Computer Science
Princeton University
mmahajan@princeton.edu
FFC Username: mayank.mahajan

Akash Kapoor
Department of Computer Science
Princeton University
kapoor@princeton.edu
FFC Username: kapoor

Abstract

The emergence of statistical models in social science has been motivated by a common goal to understand and predict social outcomes from a collected data set. In this study we explore a variety of imputation methods and machine learning classifiers to predict six social outcomes of children from the Fragile Families survey. The Randomized LASSO feature selection technique was also implemented to reduce complexity. We achieved ranks of 31st, 24th, and 15th on the grit, GPA, and materialHardship predictions respectively. On the binary outcome (Job Training), we achieved a rank of 27. We showed how the LASSO and Logistic Regression classifiers actually outperformed more complex methods like the AdaBoost classifier and overly simple ones like OLS. The results demonstrate how these trained classifiers could be useful in both accurate prediction and the design of future surveys with much fewer questions.

1 Introduction

In the field of social science, information is being produced at an increasingly rapid pace, from social media outlets to online and in person surveys [1]. With the rise of technology in everyday life, it has become much easier to collect large amounts of this observational data, which motivates the need for intelligent techniques to extract meaningful dependencies between current knowledge and future outcomes. Thus there exists much opportunity in the space of big data and machine learning to find these descriptive inferences that can shed light on the observations we see in the data. As opposed to clearly defined problems like classification or prediction, social scientists have used machine learning techniques in the past usually to gauge a certain quality present in the world by creating a model of causality and judging the fit of the data to that model. Not only can this data tell us about what the current state of the world is today, but it can also provide key insights about what the state of society may be in the future, which is very lofty yet fascinating prospect for social science.

Statistical regression in particular is of interest to social scientists because of its ability to accurately predict outcomes with only a small amount of data, a constant concern when data is collected manually [2]. Most commonly, these mechanistic models are used to create and test hypotheses about the data. The intersection of intelligent models and social sciences is quite large, with applications ranging from evaluating simulations to predicting future trends. In this study, machine learning methods were employed in three different ways: 1) to account for missing data through accurate imputation techniques, 2) to determine the ability of the results of a survey to accurately predict six social outcomes of children later in life, and 3) to explore potential subsets of features that may also predict these outcomes well while sharply reducing the required complexity of the survey.

2 Related Work

2.1 Imputation and Filtering

A common problem with manually collected data as is often the case with surveys is missing data. For any number of reasons (death, privacy, or other unavailability) participants may not have provided answers to all of the required questions, which precipitates both regressor and response variables that are not defined for all participants. For each response variable analyzed, the data was filtered for non-NA values. This ensured that there would be no problem with an unknown response variable. For example, there were only 1418 participants out of the original 2121 members of the training set had a non-null value for the "grit" response. Then, imputation and filtering was performed to update the columns of the regression matrix. In this study, three imputation techniques were used: 1) Median 2) MICE [3] and 3) K-Nearest Neighbors. In cases of multiple imputation, the MICE method has frequently been noted as an effective technique because it iteratively cycles through all of the other columns treating them as dependent variables a high level of confidence is reached. In Hollenbach et. al's 2014 study, the MICE technique was chosen as a benchmark for newer imputation techniques, so it was reasonable to include this method in our analyses[4]. As well, the K-Nearest Neighbors method has been thoroughly studied as a potential imputation technique, and has been shown to accurately impute values with a non-trivial value of k and a high proportion of complete samples, even when the amount of missing data is large[5].

There were separate pipelines designed so that both simpler and more complex imputation techniques could be tested. For the simple method, we employed the median technique, where each of the missing values in the columns was filled with the median of the column. Because this method was not computationally intensive, it was feasible to impute every single column of the data before feature selection. However, more sophisticated methods required a preliminary level of feature selection that allowed the algorithm to narrow in on columns whose missing data it would be valuable to compute. For each of these techniques (MICE and K-Nearest Neighbors), the full set of non-imputed data was passed through a Randomized LASSO algorithm (detailed in Section 3) at a relaxed threshold parameter of $\lambda = 10^{-3}$ so that all potentially employed columns would have no missing data. Imputation was then performed on the support columns before being used to train the classifier.

2.2 Dimensionality Reduction

The data was downloaded on March 10th, 2016 from the Fragile Families website. In total, there were 4242 samples, each with 12,945 regression features and six outcome variables. Three of these outcomes, grit, GPA, and materialHarship, were continuous variables. The other three outcomes, layoff, jobTraining, and eviction, were binary outcomes that represented whether the parents had been laid off, trained in their last job, and been evicted respectively. We focused only on jobTraining out of the binary outcomes for this study.

Working with such a large number of features motivated the use of dimensionality techniques to reduce the complexity of overly complicated models that would result from training classifiers on the full set of regressors. In addition to the shape of the matrix, many of the features were found to be highly correlated with each other, which presented the danger of arbitrarily choosing features that might not have actually been part of the optimal set of columns for regression. Given these constraints, the Randomized LASSO algorithm and the Recursive Feature Elimination (RFE) algorithm showed promise in discerning the most discriminative features from a set of highly similar variables while still significantly decreasing the number of features in the model. Randomized LASSO has previously been shown to perform particularly well in the case of correlated covariates [6].Guyon et. al employed the RFE algorithm to classify cancer by pruning redundant genes and retaining ones that were known to be biologically relevant to cancer proliferation [7]. These prior results showed that dimensionality reduction via Randomized LASSO and RFE could potentially extract meaningful survey questions that could be used to streamline future studies without compromising accuracy.

2.3 Classification methods

Because the problems of predicting continuous outcomes and binary outcomes are fundamentally different, we used different methods to train and predict the six desired properties in the Fragile Families data set. We employed four different regression-based classification methods for the continuous outcomes (grit, GPA, and material hardship), and generalized prediction models for the binary outcomes (jobTraining). All of the classifiers came from the SciKitLearn and StatsModels Python libraries [8, 9]. All parameterizations are the default unless specified.

For the continuous variable outcomes, we used

1. *Ordinary least squares regression* (OLS)
2. *Ridge regression* (Ridge): ℓ_2 penalty was tuned separately for each run
3. *LASSO regression* (LASSO): ℓ_1 penalty was tuned separately for each run
4. *Elastic Net* (Elastic): ℓ_1 ratio = 0.5, α tuned separately for each run

For the binary variable outcomes, we explored

1. *Logistic regression with ℓ_2 penalty* (LR): using stochastic gradient descent
2. *AdaBoost regression* (AB): Estimators & learning rate tuned for each run.
3. *Multinomial Naives Bayes* (NB)

2.4 Evaluation

Evaluation was handled similarly for the continuous and binary outcomes. For each outcome, multiple forms of feature selection with λ thresholds and imputation were performed according to the process above to select multiple subsets of features with high discriminative power for the regression step. Then, for the continuous classifiers, different models were evaluated with the reduced, imputed data across a variety of α parameters. These tests were run with cross-validation to determine the optimal α parameter for each classifier. Each of the fitted classifiers was evaluated using the mean squared error (MSE) on the set of labelled training data to find the best combination of feature selection threshold and classifier for each outcome. Predictions from each of the classifiers for each outcome were uploaded to the Fragile Families website to determine the generalization error on unseen data.

Binary outcomes followed almost exactly the same process with the Logistic Regression, AdaBoost, and Multinomial NB classifiers, where the learning rates were cross-validated and tuned for each run as well.

3 Spotlight Technique: RandomizedLasso Feature Selection

3.1 Overview of LASSO

The Least Absolute Shrinkage and Selection Operator (or LASSO) regression technique builds heavily upon the canonical form of linear regression used to fit a model to existing features and continuous outcomes, Ordinary Least Squares regression. While OLS can be a simple and effective model, one of its major drawbacks is an inability to extract a meaningful subset of features from the ones provided to the model, which can lead to fits with many redundant variables. Not only does this overly complicate the model, but this disadvantage also leads to one of the most common fallacies of regression, which is overfitting. Indeed, it is highly likely that those coefficients are small in magnitude would be highly sensitive to training data and lead to poor generalization error on unseen examples.

The LASSO technique solves this problem by introducing a regularization term on the regression. Thus, the optimization problem for the regression becomes:

$$\min_{\beta} \frac{1}{N}\|y - X\beta\|_2^2 + \alpha\|\beta\|_1$$

where α is the parameter used to tune the ℓ_1 penalty. A visual comparison of the LASSO optimization to the Ridge Regularized Regression demonstrates how the LASSO regression is more effective at removing redundant features completely.

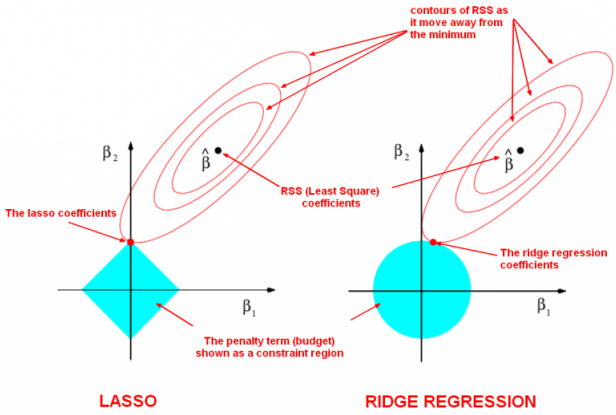

Figure 1. **Comparison of LASSO and Ridge Regression optimization**

The level curves of $\frac{1}{N}\|y - X\beta\|_2^2$ are more likely to intersect with the level curves of $\|\beta\|_1$ at the corners, where some of the features will be zero. This extends easily to many features, and the advantageous outcome of LASSO regression is that larger and less discriminative features are pruned out from the data.

3.2 Bootstrapping

While the LASSO is effective at removing redundant features, one of its shortfalls is its arbitrary choice of features to keep during the optimization when there are many highly correlated features that can be useful for the regression. Depending on the computational implementation of the algorithm, the LASSO may select almost at random and lead to suboptimal results due to its inability to discern the best features among similar signals. To solve this issue for the purpose of feature selection, bootstrapping is employed to perform LASSO on a subset of the features repeatedly and track the cumulative selection rates of each of the features. By repeatedly sampling a subset, say 75%, of all features, the Randomized LASSO algorithm enables the independent comparison of potentially correlated features such that they are not measured completely in relation to each other. It is important to perform a substantial number of iterations for convergence, but the number of iterations can be much smaller than the overall number of features p. The final step, when only the features that appear in at least a certain proportion of cases are retained, is called **stability selection** and is a well-known method for consistently extracting useful features [10].

4 Results

4.1 Evaluation Results

There was considerable variability in the performance of the different methods used for predicting the outcomes. For continuous variable, OLS performed the worst whereas LASSO, RIDGE and ElasticNet had comparable performances (Table 1). On the other hand, for the binary outcomes Logistic Regression performed the best, and AdaBoost and Multinomial Naive Bayes performed reasonably well (Table 1).

4.2 Feature Selection

The given data set had more than 12000 features, we decided to perform feature selection to prevent over-fitting and to make the computation tractable. For the continuous outcomes, we experimented

Continuous Outcomes				Binary Outcomes	
Classifier	GPA	Grit	Material Hardship	Classifier	Job Training
OLS	0.86	0.67	0.0300	AB	0.302
LASSO	0.396	0.228	0.0261	LR	0.28679
RIDGE	0.397	0.229	0.0264	NB	0.2941
ElasticNet	0.3969	0.2287	0.02639		

Table 1: **Results from using outcome prediction methods on testing data set.**. For each outcome, we report the scores obtained from the Fragile Families Challenge Result page for the submitted predictions.

with RandomizedLasso and RFE feature selection and for nominal outcomes, we tried Mutual Information Gain and CHI2.

After performing feature selection with relaxed thresholds, we were left at most with 43 features for the continuous variables and 100 features for binary outcomes. Upon testing these policies on the test data, we found out that RandomizedLasso v/s RFE and Mutual Information Gain v/s CHI2 resulted in similar performance.

Our results in Table 1 show, how we got decent performance by using a less than 0.33% of features for the continuous characteristics and 0.77% of the features for the binary characteristics.

5 Discussion and Conclusion

We compared different Regression methods along with different imputation and feature selection methods for predicting the outcomes. We found that for continuous outcomes LASSO with Median imputation and Randomized LASSO feature selection worked the best. Whereas, for binary outcomes Logistic Regression with Median imputation and Mutual Information Gain feature selection worked the best.

It was also surprising to see how complicated imputation methods, such as MICE and KNN-based imputation, did not help much. After performing a preliminary feature selection, to make the imputation tractable, there were only 8 values that needed to be imputed. Thus, marginalizing the impact of imputation and stressing on the importance of good feature selection.

Overall, this project reiterates that for the majority of the data analysis tasks, a simpler model is overall a better one. We showed how the median based imputation model performs at-par with the preliminary-feature selection using complicated imputation models. Similarly, for predicting outcomes, simple LASSO-based regression and Logistic Regression-based classification performed as well as other complicated algorithms such as ElasticNet and Adaboost. However, we also demonstrated the importance of regularization and the benefits of parameter tuning evidenced by the relatively poor performance of the OLS classifier. It stands that there does seem to be an extremely small set of features or survey questions that can just as easily predict these six outcomes, and that not only does this study demonstrate the power of the Fragile Families survey, but also the potential to improve future studies to be shorter in time and complexity without sacrificing the ability to perform causal inference.

5.1 Future Work

If given more time, we would have performed imputation after carefully analyzing and understanding qualities for each feature without or with even more relaxed preliminarily feature selection. We would have also tried modelling the data as a time-series and then using that model to make predictions. Thirdly, we could have explored some more feature selectors, regression-methods and classifiers.

References

[1] Grimmer J. We are all social scientists now: how big data, machine learning, and causal inference work together. PS: Political Science & Politics. 2015;48(01):80–83.

[2] Holme P, Liljeros F. Mechanistic models in computational social science. Frontiers in Physics. 2015;3:78. doi:10.3389/fphy.2015.00078.

[3] van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011;45(3).

[4] Hollenbach FM, Metternich NW, Minhas S, Ward MD. Fast Easy Imputation of Missing Social Science Data. ArXiv e-prints. 2014;.

[5] Jonsson P, Wohlin C. An Evaluation of k-Nearest Neighbour Imputation Using Likert Data. In: Proceedings of the Software Metrics, 10th International Symposium. METRICS '04. Washington, DC, USA: IEEE Computer Society; 2004. p. 108–118. Available from: http://dx.doi.org/10.1109/METRICS.2004.10.

[6] Bach F. Model-Consistent Sparse Estimation through the Bootstrap; 2009. Available from: https://hal.archives-ouvertes.fr/hal-00354771.

[7] Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning. 2002;46(1):389–422. doi:10.1023/A:1012487302797.

[8] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.

[9] Seabold JS, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. In: Proceedings of the 9th Python in Science Conference; 2010.

[10] Meinshausen N, Buehlmann P. Stability Selection. ArXiv e-prints. 2008;.