COS 424 Homework 2: Predictive Capability of Missing Values on Material Hardship and Job Training

Francisco J. Carrillo
Department of Chemical and Biological Engineering
fjc2@princeton.edu

Abstract

Our ability to address socioeconomic problems in our society has never been higher: we now have the capability to analyze and obtain large data sets in order to identify relevant socioeconomic factors that can be used as future "success" indicators. In this report, we draw upon the data collected by "The Fragile Families and Child Well-being Study" and ask the question: Can missing data be used to predict a child's future "material hardship" or completion of "job training"? To do this, we reduced the complexity of all provided data features into simple Boolean values: "1" for real values, and "0" for missing values. We then compare the performance of this engineered data set with the performance of a minimally-altered data set by doing feature selection and a comparative study of accuracy, precision, and the Receiver Operating Characteristics. We conclude that reluctance or inability to answer a question (or obtain an answer to it) is a good indicator for both future material hardship or job training, even when compared to predictions originating from a traditional data set containing original data values.

1 Introduction

Data science has been shown to be a very powerful tool in predicting human behaviour. This is true not only in the realm of consumer marketing, but also in the realms of socioeconomic and psychological studies . More specifically, many scientific papers have emanated from a particular study named "The Fragile Families and Child Well-being Study", an investigation that follows at-risk children in order to try to investigate what life-factors can predict later "success" in life[2, 8]. Given the complexity of this data, these studies have relied heavily on pre-processing and feature selection in order to fit their models, which is why we will prioritize these steps on this report.

Here we present an analysis on the Fragile Families data set that tries to answer the following question: Can missing data be used to predict a child's future "material hardship (MH) or completion of "job training" (JT)? In order to do this, we studied the effects that pre-processing, feature engineering, and choice of classifier have on our prediction capabilities. To put our new predictive model into context we also tried to achieve similar predictions with a traditional data set, where we look at the actual content of the answers, instead of their existence.

The report goes as follows: First, we start by pre-processing the data: studying basic trends, addressing missing data points, and doing feature engineering. We then proceed to choose the best classifier for each data set by evaluating the performance of three different classifiers. After that, we continue to do feature selection and final performance evaluations. Lastly, we present our conclusions.

2 Related Work and Methods

2.1 Data Description

The data used here is an extensive study that followed about 5000 children from birth until age 9, and then re-connected with them at age 15. This data is comprised of continuous and discrete information (more than 40 million samples spread over roughly 13,000 features) regarding children characteristics such as sex, family relationships, economic status, and race. This raw data was all contained within a 4242 by 12943 matrix. To complement this data, we were also provided with 6 key outcomes measured at age 15 (Grit, GPA, Material Hardship, Job Training, Eviction, and Layoff) in the form of a (2121 by 7) training data set and a (1591 by 7) testing data set.

Although extensive, the provided data (features and results) are far from being complete, as many questions were left unanswered by a variety of reasons (lost contact, reluctance to answer, the question was not understood ext...). In this report, we aim to extract information from the absence of answers by the creation of a new features. An explanation of our resulting data pre-processing procedure now follows.

2.2 Data Processing

In this case, data pre-processing, feature selection, and feature engineering were arguably the most important steps required in the creation of a predictive model. Here we first show our data pre-processing steps and then proceed to explain our reasoning for them. Please note that steps 1 and 2 are mutually exclusive:

1. Create the "Engineered" Data Set: Label each value in the background data as a "0" if the question was not answered (i.e. negative values, NaNs, ext...) and as a a "1" if the question was answered (i.e. 0, 1, continuous values, ext...).
2. Create the "Traditional" Data Set: Fill every missing background data value (negative values, NaNs, ext...) in a given feature with the median feature value.
3. Simplify Continuous Outcome: For each child, re-label all "Material Hardship" outcomes as a "0" if their value is below a desired threshold (i.e. the training set mean) or as a "1" if it is above said threshold.
4. Delete Remaining Missing Values: Remove features in the data sets that still have NaNs. Remove samples in outcome data that have NaNs in the two studied outcomes.
5. Finish "Cleaning" the Data; Remove features that have a variance less than 0.1.

In step 2, we decided to use the median as our replacement value over the mean and the mode for the following reasons: 1) the mean is sensitive to outliers 2) For some continuous variables the mode is sometimes non-existing 3) the median gives is a fair representation of the value distribution and is not affected by the aforementioned problems.

In step 3, we further simplified the data by converting a continuous outcome into a Boolean one. Now, instead of predicting the magnitude of the material hardship a child will endure, we predict if said hardship is above a particular threshold (in this case the mean, but it could be anything). Furthermore, this allows for the consistent use of classifiers instead of continuous regression models (i.e. linear regression).

In step 4, the outcomes for both the test and labeled data sets were further pre-proceed by removing the samples that had missing outcome values. We did this for the reason: simply inputting of median or mean outcome would introduce bias in the training model and in the performance metrics. Furthermore, given Boolean outcomes that are already skewed towards one result (some features are overwhelmingly "False"), any sort of mean/median/mode data inputting would exacerbate said skew and reduce our real predictive capability.

This resulting data sets had the following dimensions: Engineered Data Set: (1817 by 4804), Traditional Data Set: (1817 by 4970), Training Outcomes: (1014 by 7), Testing Outcomes: (803 by 7). Although some of these are relatively smaller than their original versions, we believe all these have statistically significant sizes.

2.3 Classification Models and Feature Selection

Our resulting analysis involved the following 3 classification methods: Logistic Regression, Random Forest, and Naive Bayes Classifiers. The algorithms for all of these were obtained from the SciKitLearn Python Libraries [7] and used with standard values unless specified otherwise.

Initial feature selection for Logistic Regression was done through L1 regularization (see Spotlight Classifier Section) and then improved by Recursive Feature Elimination (an algorithm that obtains the "n" most representative features by doing recursive fitting and recursive performance evaluation) [5]. During L1 regularization for Logistic Regression, the hyper-parameter "Lambda 1" was optimized through a cross validation algorithm obtained from the aforementioned python libraries. Feature Selection with Random Forest Classifiers was done by obtaining the features with the highest magnitude coefficients and ranking them accordingly.

2.4 Evaluation Criteria

The effects of pre-processing, fitting, and model performance were evaluated by the following metrics: Accuracy ($\frac{\text{Correct Samples}}{\text{Total Samples}}$), Precision ($\frac{\text{True Positives}}{\text{True and False Positives}}$), and the Area Under Curve (AUC) of the Receiver Operating Characteristic. Accuracy was chosen because it gives an overall idea of performance, but does not take into account the effects of true positives (Precision and Recall) or true negatives (Specificity). For example, when it comes to evaluating outcomes such as "layoff" accuracy can give a false sense of security: a classifier that assigns a label of "0" to every child would still have an accuracy of about 0.7 (most outcomes are False) while having a precision score of 0. The AUC combines both Recall ($\frac{\text{True Positives}}{\text{Real Positives}}$) and Specificity ($\frac{\text{True Negatives}}{\text{Real Negatives}}$) values through graphical means, resulting in a relatively robust performance metric [1] [3]. Even though these are all valid statistics, there is no single metric that can give the "whole picture". We believe that these choice of metrics gives a nice balance of seeing the overall trend while still allowing us to make sure that positives are true positives and negatives are true negatives. Note that ROC curves are not shown in the report for brevity (but their AUC values are).

3 Spotlight Classifier: Logistic Regression

Logistic Regression is a discriminative binary classification method based on the assumption that the data set $D(x_i, y_i)$ is linearly separable and that the probability of the model's response can be modeled through the linear combination of the features' values with a set of fitted coefficients β [4].

Just like linear regression, logistic regression seeks to obtain $p(y|x)$ given a labeled data set $D(x_i, y_i)$. However, instead of using a Gaussian distribution to obtain a fit to continuous response, logistic regression is fitted to binary-classified data by using a Bernoulli distribution and the logistic function:

$$p(y_i|x_i, \beta) \propto Bernoulli(\mu(\beta^T x_i)$$

$$p(y_i = 1|x_i, \beta) = E[y_i = 1|x_i, \beta] = \mu(\beta^T x_i) = \frac{1}{1 + e^{-\beta^T x_i}}$$

The inclusion of the logistic function (the last term on the right) into the Bernoulli distribution still produces a continuous probability response, but constraints most of the resulting values between zero and one. It does this while still allowing the linear combination of the covariates $\beta^T x_i$ to be main input function. Furthermore, the previous equation can be easily kernelized.

Logistic regression uses a linear separator in feature space. In this case, the position of the classification boundary is dictated by the point at which the continuous response reaches a probability of 0.5 (where $\beta^T x_i = 0$). This position can be easily modified/fit by the addition of an intercept, such that $\beta^T x_i + \beta_0 = 0$. In other, words, $||\beta^T x_i + \beta_0||$ is directly proportional to the distance to the linear separator.

Furthermore, by solving for $\beta^T x_i$ we can conclude that this term is equal to the log of the odds ratio $log(\frac{p(y_i=1|x_i)}{p(y_i=0|x_i)})$. This means we can easily relate probabilities to model behaviour, helping with data interpretation (i.e a negative log odds ratio shows x and y are inversely proportional).

Logistic regression is fitted to $D(x_i, y_i)$ by calculating the β values that maximize the conditional log likelihood. This is the same as minimizing the residuals/errors by minimizing the following function (refer to [4] for mathematical proof). Unfortunately there is no analytical solution for such process, so we need to rely on numerical methods (such as stochastic gradient decent).

$$L(D, \beta) = \sum_{i}^{n} (y_i - \beta^T x_i) x_{ij}$$

Fortunately, this classifier can be easily extended to handle multiple covariates, it is just a matter of replacing the linear combination $\beta^T x_i$ into a sum of linear combinations $\sum_{j=1}^{p} \beta^T x_{i,j}$ within the previous equations.

However, when it comes to large data sets (like the one we will study here), it is necessary to introduce cost functions to $L(D, \beta)$ in order to minimize both the magnitude (Ridge Regression) and quantity (Lasso Regression) of the linear coefficients. This results in the following equation, which we can minimize in the same way we did with the previous equation, where the lambdas correspond to the regularization coefficients. [6]:

$$Ł(D, \beta) = \sum_{i}^{n} (y_i - \beta^T x_i) x_{ij} + \lambda_{ridge} \frac{1}{2} \sum_{j}^{m} \beta_j^2 + \lambda_{lasso} \sum_{j}^{m} ||\beta_j||$$

4 Results and Discussion

4.1 Initial Classifier Comparison

Data fitting involved 4 different Data-Outcome pairs: JT and MH for the engineered data set, and JT and MH for the traditional data set. Here we investigate the performance of 3 classifiers on said pairs:

Legend: (Accuracy/Preccison/AUC)				
	Engineered Data Set		Traditional Data Set	
	Material Hardship	Job Training	Material Hardship	Job Training
Naïve Bayes	0.53/0.21/0.51	0.57/0.32/0.53	0.52/0.37/0.53	0.61/0.32/0.54
Random Forest	0.67/0.37/0.59	0.70/0.33/0.52	**0.69/0.44/0.62**	0.68/0.19/0.52
Logistic Regression	**0.72/0.64/0.71**	**0.72/0.44/0.58**	0.61/0.37/0.59	**0.60/0.30/0.53**

Figure 1: Initial Comparison of Classifier Using all Features

Figure 1 shows that Logistic Regression work best for the initial fitting of 3/4 of our tested pairs, with Random Forest being the best for JT in the traditional data set, and Naive Bayes significantly under performing in all of them. Even with no feature selection, we can see than the engineered data set performs significantly better than the traditional one. Furthermore, we noticed that the fitting and prediction for all classifiers ran significantly faster for the engineered data set (this makes sense, as the engineered data is only composed of Boolean values).

4.2 Feature Selection and Its Effects on Performance

Having chosen the best classifier for each data-outcome pair we then proceeded to obtain the 20 most predictive features following the steps described earlier. After this step, we proceeded to reduce the number of features as much as possible, stopping at the minimum number of features at which precision $\neq 0$ (at which point the classifier is unable to predict true positives). Figure 2 shows the performance of the classifiers after these two steps.

We were able to reduce the number of features of all 4 data-outcome pairs to about 5 features. However, the choice of the optimal number of features now becomes arbitrary, and depends on which performance indicator we prefer. In this case, we believe that our chosen least number of features represents the optimal solution for each data-outcome pair, as data gathering can be expensive and time consuming. Furthermore (and surprisingly), the engineered data set was able to give consistently better predictions with a lower number of features when compared to the traditional data set

Material Hardishp						
	Engineered Data Set			Traditional Data Set		
	Accuracy	Precision	AUC	Accuracy	Precision	AUC
All Features	0.72	**0.64**	**0.71**	0.69	0.44	0.62
Top 20 Features	**0.73**	0.61	0.70	**0.71**	**0.55**	**0.68**
Min. # of Features	0.71	0.54	0.60	0.69	0.41	0.61
	Min. # of Features = **1**			Min. # of Features = **2**		

Job Training						
	Engineered Data Set			Traditional Data Set		
	Accuracy	Precision	AUC	Accuracy	Precision	AUC
All Features	0.72	0.44	**0.58**	0.60	0.30	**0.53**
Top 20 Features	0.71	0.36	0.54	0.71	0.42	0.51
Min. # of Features	**0.72**	**0.52**	0.57	**0.72**	**0.67**	0.51
	Min. # of Features = **4**			Min. # of Features = **5**		

Figure 2: Feature Selection Effects on Classifier Performance for MH and JT

(even when only using 1 feature!). This gives a positive answer to our initial question: Missing data CAN be used to predict JT and MT. For this reason our future discussion will only focus on the results obtained from the engineered data set.

4.3 Discussion: Top Two Features for MH and JT

For MH, the fact the following two questions were left unanswered was very sigificant: 1) In past 12 months, have you received free food or meals because there wasn't enough money? (ID M4i23o) and 2) : Have you stayed at shelter not meant for regular housing? (ID M5f23l). We speculate that families with Material Hardship may be ashamed of not having enough food or having stayed at shelters, which is why a statistically significant number or them chose to not answer this question.

For JT, the fact that the following two questions were left unanswered was very significant: 1) Have you given money as loan to family or friends? (ID m4l2a) and 2) Are you attending a program to help get a job? (ID m5i2-13). Just as we speculated before, we believe the reason these questions were left unanswered may be because of shame or as a way to "protect" their family members. Even if we don't know the exact reason why the absence of data from these features is representative, we do know that they are able to predict future MH or JT with surprising accuracy with only 4 or 1 feature(s), respectively (See Figure 2). If it is of interest, please refer to the appendix for a list of the top 20 features of each.

5 Conclusions

In this report we have concluded that missing answers in some key questions can be used as good predictors for future MH and JT of children in fragile families. We proved this by fitting our predictive models to an engineered data set only containing Boolean values representing if a feature's value was present or not. Initial fitting showed that logistic regression worked as the best classifier. We also compared said analysis to models fitted to a traditional data set containing the actual data values. For all of our models, our performance metrics (accuracy, precision, and AUC) showed that the engineered data set performed better than its counterpart. Furthermore, through feature selection algorithms (RFE and L1 Regularization), we were able to reduce the number of representative features to 1 and 4 for MH and JT, respectively. Finally, we hypothesized why missing data for certain survey questions was a good indicator for future outcomes: some families may be ashamed of answering some very personal questions regarding their ability to provide for their families, leading them to skip or ignore the question altogether.

Future directions would involve the following: 1) Analyze how the type of missing data affects our predictions (did the subject skip it? did (s)he understand the question? was (s)he not contacted?) 2) Investigate if missing data can be used to predict other outcomes such as "eviction" or "gpa". 3) Combine missing data analysis with actual data values to increase model performance. 4) Apply a similar missing data analysis to the outcome data to understand if we can predict which children will not be available to answer questions by year 15.

References

[1] J. Brownlee. How and When to Use ROC Curves and Precision-Recall Curves for Classification in Python, 2018.

[2] Marcia Carlson, Sara McLanahan, and Paula England. Union Formation in Fragile Families. *Demography*, 41(2):237–261, 2005.

[3] Barbara Engelhardt. Precept #3.

[4] Barbara Engelhardt. Lecture #9: Logistic Regression. pages 1–42, 2019.

[5] Scikit learn developers. RFE Feature Selection.

[6] Michał Oleszak. Regularization: Ridge, Lasso and Elastic Net (article) - DataCamp, 2018.

[7] Pedregosa Fabian, Vincent Michel, Grisel OLIVIER, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Jake Vanderplas, David Cournapeau, Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Bertrand Thirion, Olivier Grisel, Vincent Dubourg, Alexandre Passos, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[8] Nancy E. Reichman, Julien O. Teitler, Irwin Garfinkel, and Sara S. McLanahan. Fragile Families: Sample and Design. *Children and Youth Services Review*, 23(4-5):303–326, apr 2001.

6 Data Source Acknowledgment

Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) of the National Institutes of Health under award numbers R01HD36916, R01HD39135, and R01HD40421, as well as a consortium of private foundations. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health

7 Appendix

Job Training	Material Hardship
m1b8	f2j11
m1b14a	f2l2
f1e1c2	m3i4a
f1f9	m3i6
m2j6a	f3c3c
m3r2	f3c4
m3r5	m4c3c
m3k6	m4c3h
m4c25c	m4c43b
m4k3a_11	m4i23o
m4l2a	f4b3
f4l9	f4c7b
k5c1	f4l1
m5e1n	m5e6a
m5j6c	m5f23l
m5i2_13	p5q3bb5
f5a51	p5q3dk
p5m2d	t5b1t
ck5saliva	t5d3
t5f4c	t5e2

Figure 3: Top 20 Features (In No Particular Order) for Models Fitted to the Engineered Data Set