Journal Title
XX(X):1–9

Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Modeling and Decision Making with Social Systems: Appendix

Brian J. Goode[1], Debanjan Datta[1] and Naren Ramakrishnan[1]

Keywords
social systems, reductive leverage, computational social science, model evaluation

A1. Output and Related Survey Questions

All of the output variables were derived from the FFC 15-year survey block. Here we list the related questions that we used in each of the previous surveys.

GPA

GPA only appeared in the 15-year survey. There was no exactly equivalent question in any other or prior year survey.

Grit

Four survey questions form the Grit index:

1. I keep at my schoolwork until I am done with it.
   {D2I}
2. Once I make a plan to get something done, I stick to it.
   {D2K}
3. I finish whatever I begin.
   {D2M}
4. I am a hard worker.
   {D2V}

[1]Discovery Analytics Center, Virginia Tech

Corresponding author:
Brian J. Goode, Discovery Analytics Center Virginia Tech, Arlington, VA 22203, USA
Email: bjgoode@vt.edu

Material Hardship

1. In the past twelve months, did you receive free food or meals?
   {J37}
2. In the past twelve months, were you ever hungry, but didn't eat because you couldn't afford enough food?
   {J37}
3. In the past twelve months, did you ever not pay the full amount of rent or mortgage payments?
   {J37}
4. In the past twelve months, were you evicted from your home or apartment for not paying the rent or mortgage?
   {J37}
5. In the past twelve months, did you not pay the full amount of gas, oil, or electricity bill?
   {J37}
6. In the past twelve months, was your gas or electric services ever turned off, or the heating oil company did not deliver oil, because there wasn't enough money to pay the bills?
   {J37}
7. In the past twelve months, did you borrow money from friends or family to help pay bills?
   {J37}
8. In the past twelve months, did you move in with other people even for a little while because of financial problems?
   {J37}
9. In the past twelve months, did you stay at a shelter, in an abandoned building, an automobile, or any other place not meant for regular housing, even for one night?
   {J37}
10. In the past twelve months, was there anyone in your household who needed to see a doctor or go to the hospital but couldn't go because of the cost?
    {J37}
11. In the past twelve months, was your telephone service (mobile or land line) cancelled or disconnected by the telephone company because there wasn't enough money to pay the bill?
    {J37}

Layoff

1. Since {MONTH AND YEAR COHORT CITY FIELDED IN YR 9}, have you been laid off from your employer for any time?
   { J51 }

Job Training

1. Since {MONTH AND YEAR COHORT CITY FIELDED IN YR 9}, have you taken any classes to improve your job skills, such as computer training or literacy classes?
   { J51 }

Eviction

1. Since {MONTH AND YEAR COHORT CITY FIELDED IN YR 9}, were you evicted from your home or apartment for not paying the rent or mortgage?
   { J51 }

A2. Distribution of Error - MSE

Final Results

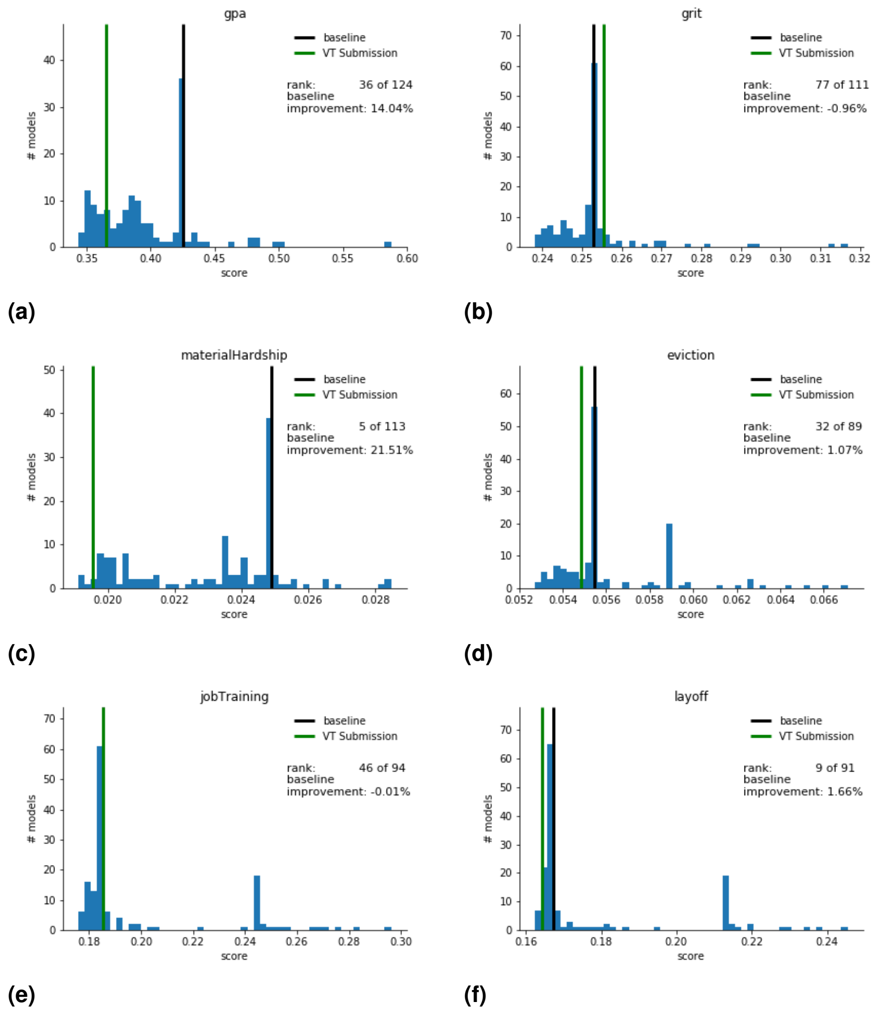


Figure 1. These figures show how our model performed in each of the output categories (green line). The black line shows the performance of the baseline model. The histograms show the performance of all project submissions. With the exception of GPA and Layoff, our model either met or exceeded the baseline result. Our best performing model, Material Hardship, exceeded the baseline by about 14%. This document explores what this means in real terms.

Intermediate (leaderboard) Results

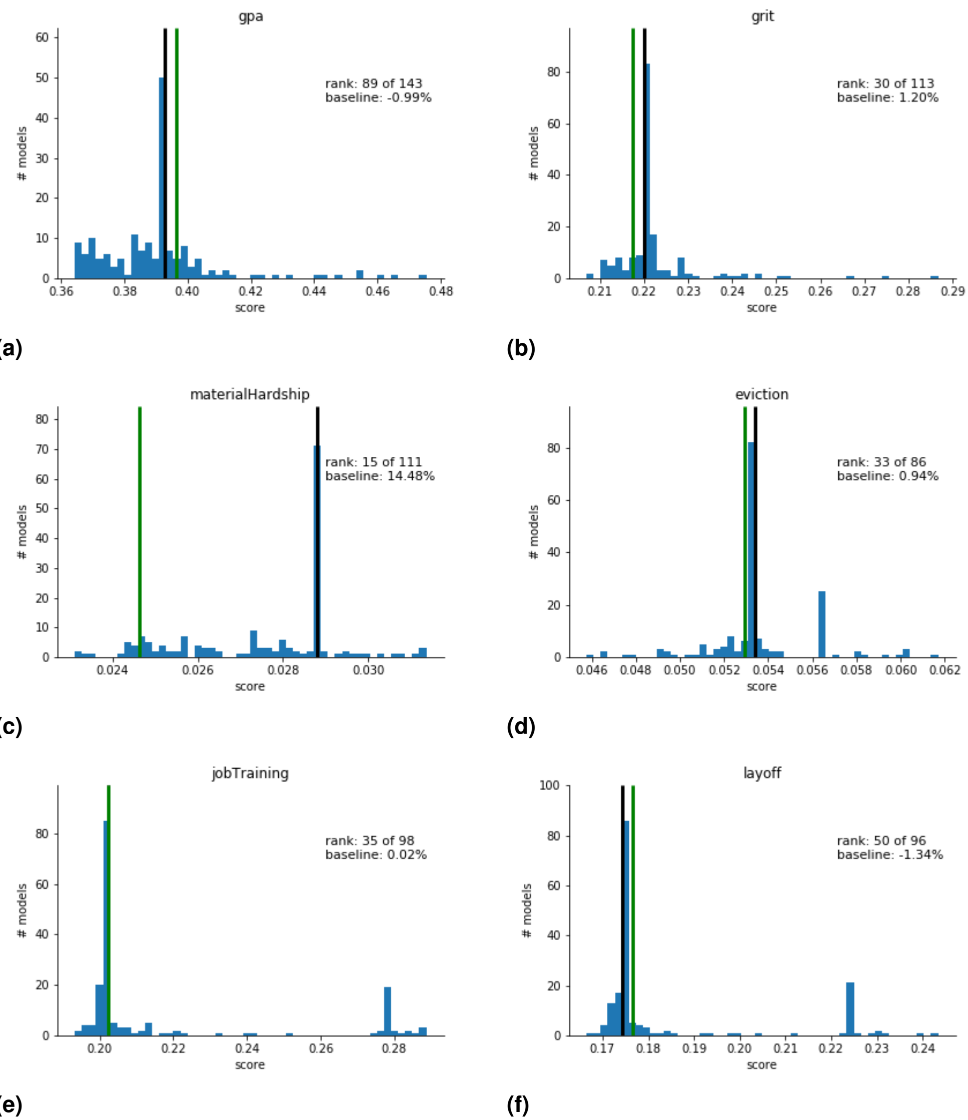
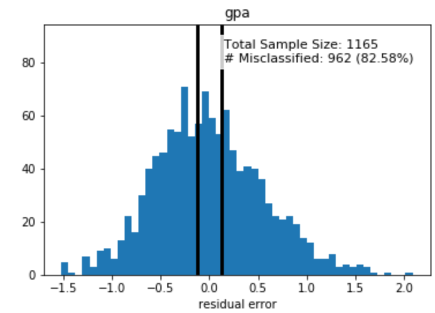
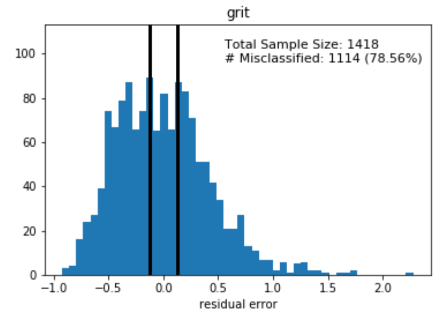


Figure 2. These figures show how our model performed in each of the output categories (green line). The black line shows the performance of the baseline model. The histograms show the performance of all project submissions. With the exception of GPA and Layoff, our model either met or exceeded the baseline result. Our best performing model, Material Hardship, exceeded the baseline by about 14%. This document explores what this means in real terms.

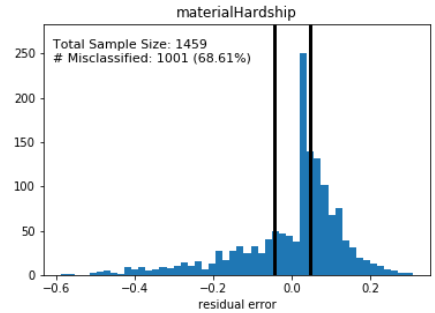
A3. Distribution of Error - Individual Threshold Errors



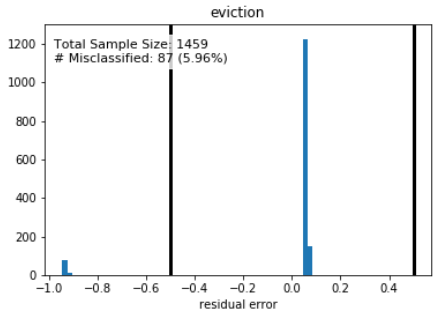
(a)



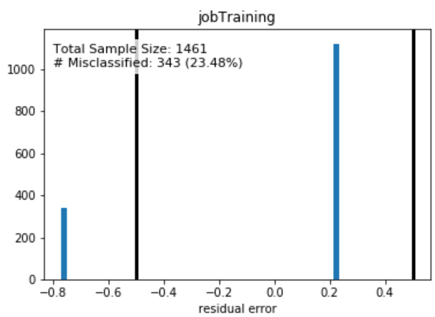
(b)



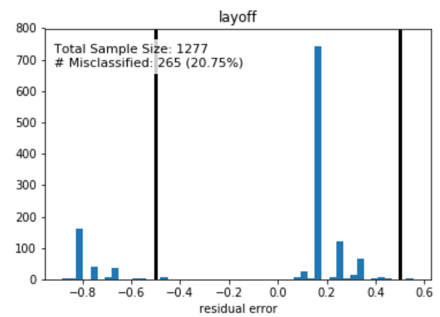
(c)



(d)



(e)



(f)

Figure 3. The individual residual error, $r = \hat{y}_i - y_i$, is presented for each of the model outputs. The vertical bars show the threshold error allowed for correct classification given the definition of the output. Error counts within the threshold will be correctly partitioned. Error counts outside of the bars represent incorrect partitioning.

A4. Derivation (statistical mechanics for social scientists)

Macro/Micro-state measures of error

In this section we show the simple process by which one can view the aggregate mean squared error measure in the Fragile Families Challenge as a macrostate measure. The corresponding microstates in this case, individual sample errors (residuals), are shown to be a spherical level set. The mean squared (aggregate) error measure is defined as

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

where $e \in \mathbb{R}$ is the individual sample error and N is the number of samples considered. The values, $\hat{y} \in \mathbb{R}$, are model predictions. The true values are y_i. The residual is defined by $r_i = \hat{y}_i - y_i$ and gives the individual difference in prediction for a given sample, i. The vector of residuals is therefore $\mathbf{r} \in \mathbb{R}^N$.

As the measure of model performance the mean squared error is an L^2-norm: $\mathbf{r} \mapsto e$. This is an injective function with more than one residual vector, $\mathbf{r}$, that maps to a given error, e.

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

$$e = \frac{1}{N}\langle \mathbf{r}, \mathbf{r}\rangle \tag{3}$$

$$eN = \langle \mathbf{r}, \mathbf{r}\rangle \tag{4}$$

We can see that the aggregate error, e is a macrostate measure of the model error, because there are multiple $\mathbf{r}$ that satisfy this equation for a given N and e. The $\langle \cdot, \cdot \rangle$ operator is the dot product. Therefore if we make the substitution, $R^2 = eN$,

$$R^2 = \langle \mathbf{r}, \mathbf{r}\rangle \tag{5}$$

which shows that the valid solutions for $\mathbf{r}$ form a spherical level set in an N-dimensional space. We term these solutions, $\mathbf{r}_e$, which correspond to a given error e and sample size N specification equal to $dim(\mathbf{r})$. The solution manifold for individual errors is then given by

$$\mathbf{r}_e \in \mathcal{S}_N(R) \tag{6}$$

where $\mathcal{S}_N(R)$ is an N-dimensional sphere with radius, R. The difference between the aggregate error and the individual error is subtle in practice, but in terms of topology it creates demonstrable changes in the mathematics. Namely, the topology as we have just shown, changes from being a point measure in aggregate to a level-surface in N-dimensions for individual samples. Any residual vector on the spherical manifold can result in having the same aggregate measure. Without any prior knowledge of the data or models, then we can only assume that each residual vector is just as likely as any other with a uniform probability.

Manifold probability and individual samples: uniform thresholding.

In the paper, we discuss thresholding individual samples and the probability of being incorrect without using a data-defined baseline. This section shows how the probability of incorrectly thresholding an individual sample is defined as an integral over a manifold.

The number of partition errors, n_e, for a uniform threshold, τ, is found by evaluating the function, $\phi(\mathbf{r})$ where,

$$n_e = \phi(\mathbf{r}) = \sum_{i=1}^{N} \mathbb{1}(|r_i| > \tau) \tag{7}$$

and represents the number of samples out of N that will be incorrectly classified. We note that this is a two-sided count around the threshold centroid - including the boundaries. This simplifies the problem by not having to consider both positive and negative thresholds, but may not always account appropriately for boundary effects. For example, at the lowest threshold, a model may output a value $\hat{y}_i \ll \tau$, yet it may still be classified in the lowest bin depending on the system in use. In this model, we assume that any threshold crossing counts toward the error.

We now want to find the probability, $P(n_e)$, that there are n_e samples that cross threshold τ.

$$P(n_e) = \frac{\sigma^N(\{\mathbf{r} \in \mathcal{S}_N(R) : \phi(\mathbf{r}) = n_e\})}{\sigma^N(\mathcal{S}_N(R))} \tag{8}$$

where $\sigma^N(\cdot)$ is a spherical measure.

Probability of any threshold crossings. A useful way to think about the probability integral over the surface is in terms of a threshold N-cube. The uniform threshold, τ, applied to each sample forms an N-cube, $\mathcal{C}_N(2\tau)$, constraint manifold with edge length 2τ. As noted, without any priors based on data or models, for a given e, any r_e is equally as valid. The problem becomes finding the region of the N-sphere that is not in the intersection of subcubes of an inscribed N-cube, $\mathcal{C}_N(2\tau)$. For example, the probability of any threshold crossings is given by

$$P(n_e > 0) = \frac{\sigma^N(\mathcal{S}_N(R) \backslash \mathcal{C}_N(2\tau))}{\sigma^N(\mathcal{S}_N(R))} \tag{9}$$

We can gain insight into the individual error terms, by finding the bounds of the probability as a function of τ and R.

1. $P(n_e > 0) = 0$, $R \leq \tau$
2. $P(n_e > 0) = 1$, $R \geq \tau\sqrt{N}$
3. $P(n_e > 0) \in (0, 1)$, otherwise

Here, we see that $P(n_e)$ is dynamic over the range $R \in (\tau, \tau\sqrt{N})$. There will be no individual errors if $R \leq \tau$, and there is guaranteed to be at least 1 individual threshold error if $R \geq \tau\sqrt{N}$.

Binomial partitioning of $\mathcal{S}_N(R)$. Here, we show how $P(n_e)$ is related to the binomial distribution, and is a useful way of calculating $\sigma^N(\{\mathbf{r} \in \mathcal{S}_N(R) : \phi(\mathbf{r}) = n_e\})$. In the previous section, we saw how the N-cube could be used to determine whether or not any errors would occur. We will continue to use the

N-cube as the basis for further partitioning $\mathcal{S}_N(R)$ into regions of different error crossings.

$$\mathcal{C}_N^d(\tau, R) := \{[\tau, R]^d \times [0, \tau]^{N-d}\} \tag{10}$$

where d is the number of threshold crossings along each of the dimensions. The set of all $\mathcal{C}_N^d(\tau, R)$ covers only the strictly positive $\mathbf{r}$, but both the N-cube and N-sphere are symmetric about each basis at the origin, meaning the probability still holds. Therefore, we can cast the probability as

$$P(n_e) = 2^N \binom{n_e}{N} \frac{\sigma^N(\mathcal{S}_N(R) \cap \mathcal{C}_N^{n_e}(\tau, R))}{\sigma^N(\mathcal{S}_N(R))} \tag{11}$$

This formulation corresponds to Figure 3 in the document which shows the partitions for N=2. With this form, the binomial coefficient results from symmetries in the cube partitioning, and the 2^N results from symmetries about the origin. We can see a large similarity to the binomial distribution, except that there is no constant probability due to the curvature of the sphere in each of the partitions.

Practical methods to solve for $P(n_e)$ - Direct Sampling.

For very high dimensions, the integral becomes very difficult due to the discontinuities arising at each threshold crossing between the N-sphere and the N-cube partitions. Therefore, we solve the integral numerically using direct sampling. The first step is to sample k points from $\mathcal{S}_N(R)$. The uniformly sampled points are generated using

$$\hat{s}_i = R\frac{\mathbf{x}}{\|\mathbf{x}\|}, i \in [1, k] \tag{12}$$

$$\mathbf{x} = [\mathcal{N}_1(0,1), \mathcal{N}_2(0,1), \ldots, \mathcal{N}_N(0,1)] \tag{13}$$

where $\mathcal{N}(0, 1)$ are independent and normally distributed samples with zero mean and unit variance. This proceedure produces k points on the surface of the N-sphere of radius R. The estimator of the probability, $P(n_e)$, is then given as

$$\hat{P}(n_e) = \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}\left(\phi(\hat{s}_i) = n_e\right) \tag{14}$$

The maximum likelihood estimate of the probability of making n_e errors is

$$\bar{P}(n_e) = \langle \hat{P}(n_e) \rangle_{\hat{s}} \tag{15}$$

Our current procedure uses $k = 10000$. The code is given with this submission.