Practical Diagnostic Tools for Data Linkage Methods November 2019

MoonJung Cho[*] Justin McIllece[†]

Key Words: Classification Tree, Establishment Survey, Lasso Regularization of Generalized Linear Models, Receiver Operating Characteristic Curve, Quarterly Census of Employment and Wages, Weighted Match Algorithm.

1. Introduction

The Quarterly Census of Employment and Wages (QCEW) program of the U.S. Bureau of Labor Statistics (BLS) registers are compiled from data obtained by state-level Unemployment Insurance (UI) programs and covers nearly the entire universe of U.S. business establishments. The QCEW maintains business establishment registers for all states, which include such information as employment, total paid wages, industry codes and physical location. The QCEW files are used to construct sampling frames for BLS establishment surveys, such as those conducted by the Current Employment Statistics (CES) and Occupational Employment Statistics (OES) programs. On a quarterly cycle, these lists are linked longitudinally and updated with the most recently available administrative and economic data. An analytical linkage method is based on linking variables which were chosen following the administrative definitions of establishment linkage. We consider some diagnostics tools which can be applicable in production.

2. QCEW Record Linkage Procedure

In QCEW, establishments that are continuing operations under the same ownership from one quarter to the next are linked by administrative linkage procedure. In the initial steps of the record linkage system, establishments are linked through a unique combination of state code, UI Number, and Reporting Unit Number. This combined field is called SESA ID. Linkage procedure passes through a series of steps for further comparison iteratively. If a record pair meets the criteria in these steps, it is considered as a match, otherwise, it is considered as a nonmatch. These administrative steps match the majority (about 95%) of the establishments of records (Helfand and McIllece, 2016). For the remaining nonmatched records from the administrative steps, Weighted

[*] Office of Survey Methods Research, U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue NE, Washington, DC 20212

[†] Office of Employment and Unemployment Statistics, U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue NE, Washington, DC 20212

Match is applied at the final step. For some general background on Weighted Match (WM) and the QCEW, see McIllece and Kapani (2014).

Weighted Match (WM), the BLS-developed-algorithm, has replaced the commercial software AutoMatch since 2015. WM closely follows matching criteria of establishment predecessor and successor definition from the QCEW Operating Manual: "A predecessor/successor relationship is defined as one where the successor (the new owner of an establishment) performs similar operations to the predecessor (the previous owner of an establishment) using some or all of the predecessors employees. These operations are frequently, but not necessarily, performed at the same location as the predecessor." The key concepts from this definition are: performing similar operations; using some or all of the same employees; and frequently performing operations at the same location. Predictor variables are chosen based on the key concepts for WM. For example, EIN, LEGAL, TRADE, and RUD for similar operations; EMP and WAGE for retainment of employees; ADDR1, ADDR2, CNTY, and PHONE are chosen for the same location. See Table 5 for the variable description.

The important steps of WM are in sequence: standardizing, blocking, scoring, weighting and matching. The blocking step constructs an initial match file of pairs that meet a baseline matching requirement. Blocking minimizes the computational burden by screening through prescribed matching criterion, or block.

In the scoring step, the similarity of all record pairs is quantified for each of the eleven variables listed in Table 5. For a pair p, a score is computed for each variable. Each variable of a pair receives a matching score between zero and one.

In the final step of weighting and matching, WM assigns larger weights on variables which are considered to have more discriminatory power. Specifically, more weight is given to unique variables that are considered to provide greater discriminatory power for individual establishments: EIN, LEGAL, TRADE, ADDR1, and RUD. It then calculates a linkage grand score D_p, a weighted sum of variable scores of a pair p:

$$D_p = \left(1.00 \times x_p + 0.75 \times x_p^*\right) \Big/ 1.75$$

x_p: normalized score of a pair p across all 11 variables;
x_p^*: normalized score of a pair p across unique 5 variables.

A pair is declared to be a match if the grand score of a pair is above a certain threshold k (i.e., $D_p > k$). See the Appendix section for more detailed formula of a linkage grand score D_p.

3. Data Description

Our data are from establishments between the fourth quarter of 2012 and the first quarter of 2013 on which QCEW compared a new algorithm (WM) with the existing software (AutoMatch). Before QCEW made a decision to adopt WM over AutoMatch, the team compared them on seven test datasets where each dataset represented an individual state. Among those seven datasets, QCEW data experts selected two datasets (Alabama and New York) and reviewed them manually, meaning that QCEW data experts reviewed the ‘matched’ set of WM and ‘matched’ set of AutoMatch. QCEW data experts examined the pairs and graded them as good links, bad links, or indeterminable. A pair received 1 if the data experts identified it as a good link, 0 if they identified it as a bad link, and 0.5 if there was not enough information to judge the pair either way.

The Alabama dataset was small enough that the entire dataset was reviewed. The entire dataset meant a union of matched sets from WM and AutoMatch. 185 pairs from the Alabama dataset were reviewed by QCEW data experts. After deleting pairs with missing score values[1] or undetermined cases, there were 166 pairs with 155 matched and 11 not-matched for this study.

New York dataset had too many pairs to review manually. 161 pairs were selected randomly in such a way that pairs whose scores were close to the threshold had more chance to be selected. After deleting pairs with missing score values or undetermined cases, there were 111 pairs with 72 matched and 39 not-matched.

4. Diagnostic Tools

4.1 Visual Display

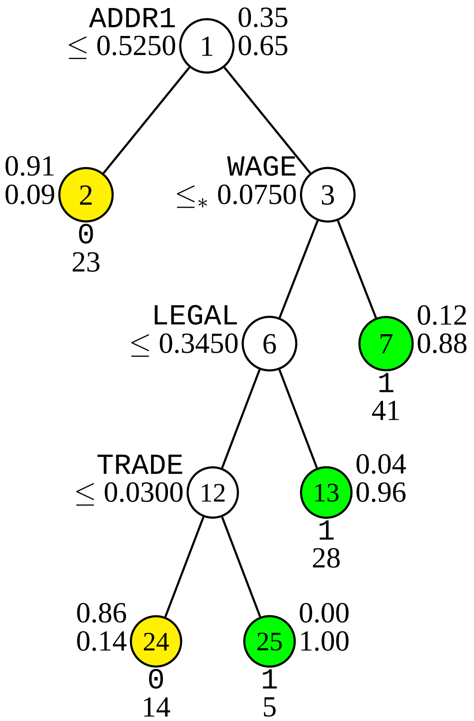
We can visually check how well WM classifier performed. Figure 1 presents the mean variable scores of matched pairs and nonmatched pairs. It shows that the mean scores of 11 variables of matched pairs were consistently higher than mean scores of nonmatched pairs. Figure 2 shows the differences of mean scores between nonmatched and matched pairs. Differences of mean variable scores between matched and nonmatched pairs were much larger for LEGAL, ADDR1, and EMP; smaller for RUD and PHONE. Similar results are shown for median variable scores in Figure 3.

4.2 Variable Selection

We applied GUIDE classification tree to Y variable and scores of 11 variables as predictor variables. GUIDE is a multi-purpose machine learning algorithm for constructing classification and regression trees. GUIDE stands for Generalized, Unbiased, Interaction Detection and Estimation. We applied the classification tree to identify and rank important variables in predicting Y:

$$Y = \begin{cases} 1 & \text{if matched} \\ 0 & \text{if nonmatched.} \end{cases}$$

[1] We are interest in WM algorithm performance, therefore records without WM scores were removed



GUIDE v.30.0 0.50-SE classification tree for predicting Y using estimated priors and unit misclassification costs. Number of observations used to construct tree is 111. Maximum number of split levels is 10 and minimum node sample size is 2. At each split, an observation goes to the left branch if and only if the condition is satisfied. The symbol '$\leq_*$' stands for '$\leq$ or missing'. Predicted classes and sample sizes printed below terminal nodes; class proportions for Y = 0 and 1 beside nodes. Second best split variable at root node is LEGAL.

The tree for predicting Y first splits on ADDR1 (street address of an establishment). Pairs whose ADDR1 score is less than and equal to 0.525 were sent to the left terminal node (Node 2). There were 23 pairs sent to Node 2: 91% of them has been reviewed as nonmatch and 9% as match. Naturally, GUIDE predicts Node 2 as nonmatch. Pairs whose ADDR1 score is larger than 0.525 were sent to the right node where they were split on WAGE. Pairs whose WAGE score is larger than 0.075 were sent to the right terminal node which predicted as match. Pairs whose WAGE score is less than 0.075 or missing were sent to the left node where they were split further by LEGAL and TRADE.

Although RUD has been regarded as one of the unique variables, it ranks lower. Note that GUIDE considers variables with unscaled scores above 1 important. Initially, WAGE and EMP were assigned to general variables which were not considered to offer information specific to an individual establishment. Yet, their unscaled scores were above 1 and ranked higher than some of the unique variables. Since scores of 11 predictor variables were not affected by weighting, we can see more clearly how each variable contributed as shown in a tree diagram.

For the same reason, we excluded a grand score highly compounded with other variables. When

Scaled	*Unscaled*	*Rank*	*Variable*
100.0	4.71	1.00	LEGAL
82.5	3.89	2.00	ADDR1
58.7	2.77	3.00	WAGE
54.0	2.55	4.00	EMP
40.4	1.91	5.00	TRADE
33.0	1.56	6.00	EIN
28.4	1.34	7.00	NAICS
25.5	1.20	8.00	CNTY
5.2	0.24	9.00	RUD
3.4	0.16	10.00	ADDR2
1.3	0.06	11.00	PHONE

Table 1: Predictor Variables Ranked by Importance Scores

a grand score was included in predictor variables, however, a grand score was the first split variable and the most important one according to GUIDE variable rankings.

The Alabama dataset had one dominant class, i.e., $Y=1$. In a case like this, instead of using estimated priors, the equal priors option may be used to find out which variables are more predictive and how they affect the dependent variable. Although the resulting model should not be used for prediction, it can be used to identify the nodes where the dominant class proportion is much higher or much lower than average (User Manual for GUIDE ver. 31.0).

4.3 Performance and Threshold

Using counts in the positive class, we estimated TP (True Positive) and FN (False Negative) counts. In the same way, using counts in the negative class, we estimated FP (False Positive) and TN (True Negative) counts. All such criteria are described by a 2-by-2 confusion matrix. The confusion matrix is defined as

$$\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix}.$$

Given a threshold, we can compute the true positive rate (TPR) and false positive rate (FPR) of the WM algorithm. The true positive rate is calculated from the cells where QCEW experts declared as matched: TP/ (TP + FN). The true positive rate is also called sensitivity. Similarly, the false positive rate (FPR) is calculated from the cells where QCEW experts declared as nonmatched: FP/ (FP + TN). The false positive rate is also expressed as 1 - specificity.

The Receiver Operating Characteristic (ROC) curve in Figure 4 is a plot of TPR and FPR resulting from continuously varying the threshold. The curve plots TPR against FPR, and the change of color indicates change of threshold values. ROC curve considers all possible thresholds

Threshold	*TPR*	*FPR*	*TPR-FPR*
. . .			
0.59	0.7917	0.1282	0.6635
0.58	0.8333	0.1282	0.7051
0.57	0.8750	0.1538	0.7212
0.56	0.8889	0.1538	0.7350
0.55	**0.9028**	**0.1538**	**0.7489**
0.54	0.9028	0.2308	0.6720
. . .			

Table 2: True Positive Rate and False Positive Rate on varying Threshold (NY Dataset)

and displays quality of the classifier visually. It also provides a performance summary number, the area under curve (AUC). As the AUC measures the overall quality of the classifier, a perfect result with no misclassified points is a right angle to the top left of the plot. Larger AUC values indicate better classifier performance. A diagonal line ($y = x$) represents performance with no discriminating power. For the New York data, ROC curve was well above a diagonal line and AUC measure is 0.9197. Figure 5 displays both TPR and FPR against threshold values. It would be ideal to achieve higher TPR and lower FPR. As Figure 5 illustrates, however, TPR and FPR move in the same direction. As a threshold value becomes lower, TPR is getting higher but also FPR gets higher as shown. Meanwhile, as a threshold value becomes higher, FPR gets lower but TPR also gets lower.

There are several approaches to obtain an optimal threshold value. One option is to find an optimal threshold value where a difference between TPR and FPR is maximum.

Figure 6 displays difference between TPR and FPR against values of threshold. For the New York data, the difference takes its maximum, 0.7489, where threshold was 0.55. It provides easy graphical inspection of the optimal cutoff.

For the Alabama data, the ROC curve is also well above a diagonal line and AUC measure is 0.9601. The difference between TPR and FPR takes its maximum, 0.8317, where threshold is 0.57.

For the combined data of New York and Alabama, the ROC curve is well above a diagonal line and AUC measure is 0.9481. The difference between TP and FP rates takes its maximum, 0.7675, where threshold is 0.57.

We resampled 1000 times through bootstrap from the combined data. The mean of threshold values across 1000 samples was 0.5706 and median was 0.57. Table 3 shows that the most frequently chosen value is 0.58.

5. Modeling

For modeling, we need to balance between fitting the data well and prediction. In other words, we need to balance between fitting the data well with generalizability to new data. In generalized linear models (GLM), the response variable y_i is assumed to follow an exponential family distribution

Threshold	*Count*	*Percent*
0.55	196	19.60%
0.56	207	20.70%
0.57	245	24.50%
0.58	255	25.50%
0.61	86	8.60%
0.62	11	1.10%

Table 3: Distribution of Threshold Values from 1000 Random Samples of Combined data

with mean which is assumed to be some function of $x_i^T\beta$.

$$\begin{aligned}\ell &= \beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_{11}x_{11}\\ \ell &= log(p/(1-p))\\ p &= E(Y|X)\end{aligned}$$

where p is the conditional mean of Y given X. In linear regression, the expected values of the response variable are modeled based on a combination of values taken by the predictors. Lasso (Least Absolute election and Shrinkage Operator) includes a penalty term that constrains the size of the estimated coefficients. As the penalty term increases, it sets more coefficients to zero. This means that the lasso estimator is a smaller model with fewer predictors. It is recommendable to use lasso to identify important predictors and reduce the number of predictors in a model. Lasso is also known to produce estimates with potentially lower predictive errors than ordinary least squares:

$$\hat{\beta}^{\text{LASSO}} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n}(l_i - \mathbf{x}_i^T\beta)^2 + \lambda\sum_{j=i}^{p}|\beta_j|\ .$$

λ is a tuning parameter which controls the penalty. If $\lambda=0$, $\hat{\beta}^{\text{LASSO}}$ becomes linear regression estimate. As λ increases, more coefficients shrunken to 0, and finally $\hat{\beta}^{\text{LASSO}}$ becomes 0 when $\lambda=\infty$.

We applied LASSO-GLM to NY dataset. For λ_{min} which obtained minimum deviance, ADDR2, NAICS, and WAGE were dropped from 11 variables. For $\lambda_1=\lambda_{\text{min}}+(1\times\text{s.e.})$, additional EIN, PHONE, and RUD were dropped. λ_1 makes the smaller model with relatively low mean squared error (MSE). Figure 7 presents an estimate of the MSE on new data fitted by LASSO-GLM per λ and error bars for the estimates. There are two specific λ values with green and blue dashed lines: a green, dashed line indicates λ_{min} with a minimum cross-validated MSE; a blue, dashed line indicates λ_1 that is within one standard error of the minimum MSE.

Table 4 shows five variables which LASSO-GLM (λ_1) model selected: three from unique variables and two from general variables. Note that GUIDE also considered those five variables important in its variable ranking. AUC measure for fitted values of LASSO-GLM (λ_1) model was 0.93, and optimal threshold which obtained the maximum difference between TPR and FPR was 0.58.

Variable	*Coef Est*
Intercept	-2.51
ADDR1	2.59
LEGAL	2.01
TRADE	1.05
EMP	0.40
CNTY	0.25

Table 4: LASSO-GLM Coefficients for λ_1 (NY)

6. Summary

We considered diagnostics tools which can be applicable in production The decisions from the QCEW manual review were used as a gold standard. Graphical displays help us to understand the data easily and to achieve better communication among practitioners. We presented mean and median variable scores of matched pairs and nonmatched pairs. By comparing visually, we could check how well the WM classifier performed. A performance curve offers more information and allows us to examine the classifier performance across a range of thresholds. For example, we could examine a performance summary number (AUC) and locate the threshold that maximizes the classification accuracy. Classification tree is simple and powerful method. We applied classification tree to rank variables and examine data structure. Finally, we applied LASSO-GLM for variable selection and performance comparison.

Acknowledgments

The authors thank David Talan for his helpful comments on the Quarterly Census of Employment and Wages (QCEW) program. The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the U.S. Bureau of Labor Statistics.

Appendix

$$\begin{aligned}
D_p &= \left(1.00 \times x_p + 0.75 \times x_p^*\right) \Big/ 1.75 \\
&= \left(1.00 \times \sqrt{\frac{\sum\limits_{i=1}^{11} s_{pi}^2}{11 - I_p[single]}} + 0.75 \times \sqrt{\frac{\sum\limits_{u=1}^{5} s_{pu}^2}{5 - I_p[single]}} \right) \Big/ 1.75
\end{aligned}$$

s_{pi}: a score of each variable i for a pair p

s_{pi}: a score of unique variable u for a pair p

$I_p[single]$: an indicator of a single establishment:
1 if the establishment is flagged a single establishment and 0 otherwise.

REFERENCES

Helfand, J. and McIllece, J. (2016), “Implementation and Results of a New Administrative Record Linkage Methodology in the Quarterly Census of Employment and Wages,” *Proceedings of the American Statistical Association*, Government Statistics Section [CD-ROM], 939-947.

Loh, W.-Y. (2009), “Improving the precision of classification trees,” *Annals of Applied Statistics*, vol. 3, 1710-1737.

Loh, W.-Y. (2019), *User Manual for GUIDE ver. 31.0*, Available at http://pages.stat.wisc.edu/ loh/treeprogs/guide/guideman.pdf (accessed September 2019).

Martinez, W.L. and Cho, M.J. (2015), *Statistics In Matlab A Primer*, CRC Press, New York.

McIllece, J. and Kapani, V. (2014), “A Simplified Approach to Administrative Record Linkage in the Quarterly Census of Employment and Wages,” *Proceedings of the American Statistical Association*, Survey Research Methods [CD-ROM], 4392- 4404.

Robertson, K., Huff, L., Mikkelson, G., Pivetz, T., and Winkler, A. (1997), “Improvements in Record Linkage Process for the Bureau of Labor Statistics Business Establishment List, *Record Linkage Workshop and Exposition Proceedings*, 212-221.

US Bureau of Labor Statistics, *BLS Handbook of Methods*, Business Employment Dynamics website, Available at https://www.bls.gov/bdm/ (accessed September 2019).

Variable	Full Name	Description	Type
ADDR1	Address1	Physical street address of an establishment	Text
ADDR2	Address2	Physical city and zip code of an establishment	Text
CNTY	County	Physical county code of an establishment	Binary
EIN	Employment Identification Number	Federal Tax Identification Number	Binary
EMP	Employment	Count of establishments average quarterly employment	Num
LEGAL	Legal Name	Legal Name	Text
NAICS	North American Industrial Classification System	6-Digit North American Industrial Classification System Code	Cat
PHONE	Phome Number	Establishment phone number	Binary
RUD	Reporting Unit DescriptionNumber	Additional operational or locational information for some multi-establishment firms	Text
TRADE	Trade Name	Trade Name	Text
WAGE	Wage	Count of total quarterly wages paid to employees	Num

Table 5: Variable Description

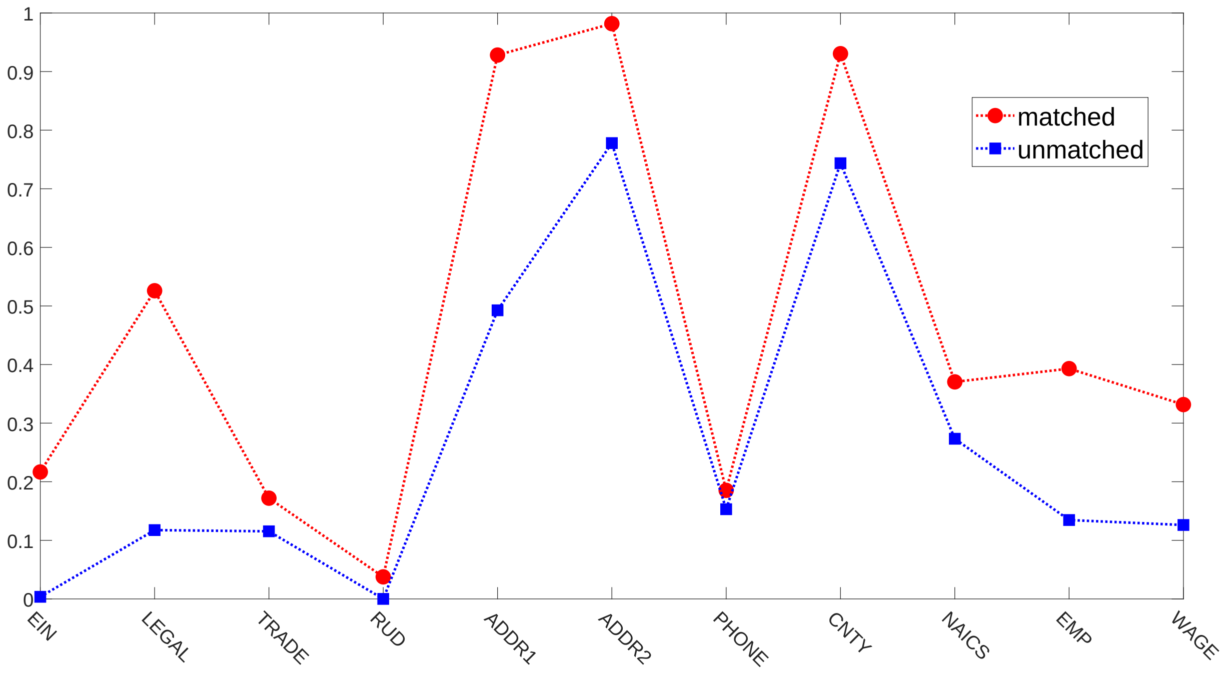


Figure 1: **Mean Scores of Variables (NY)**

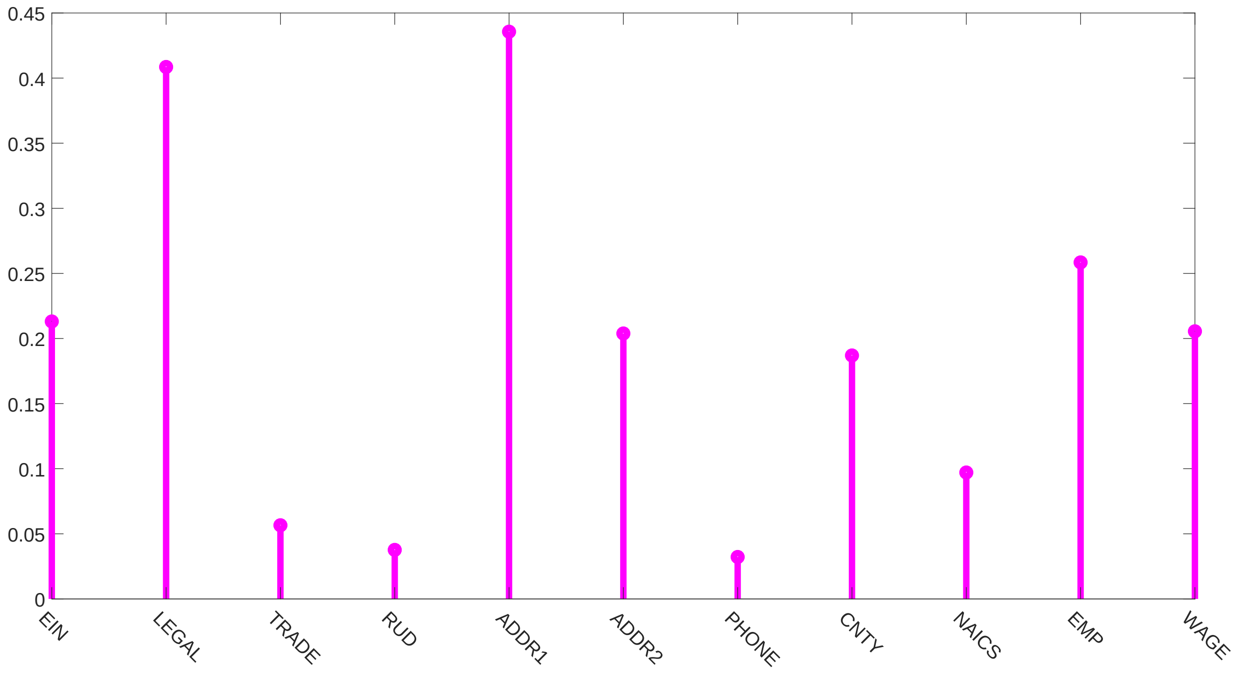


Figure 2: **Difference of Mean Scores (NY)**

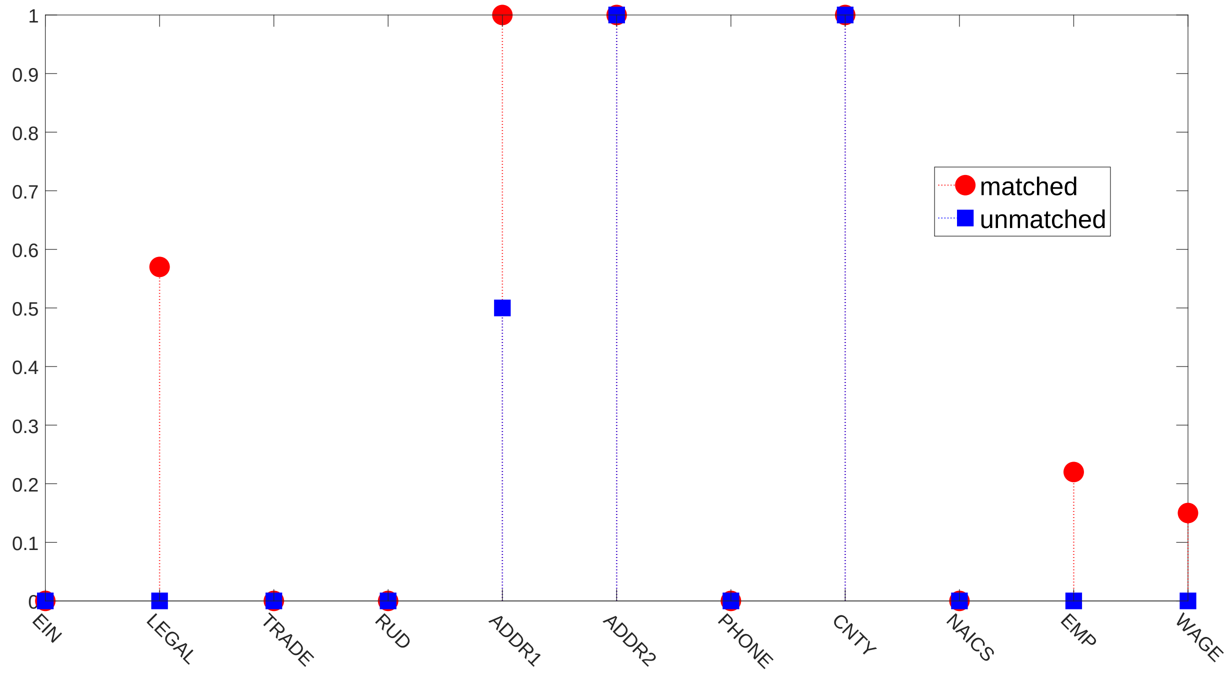


Figure 3: **Median Scores of Variables (NY)**

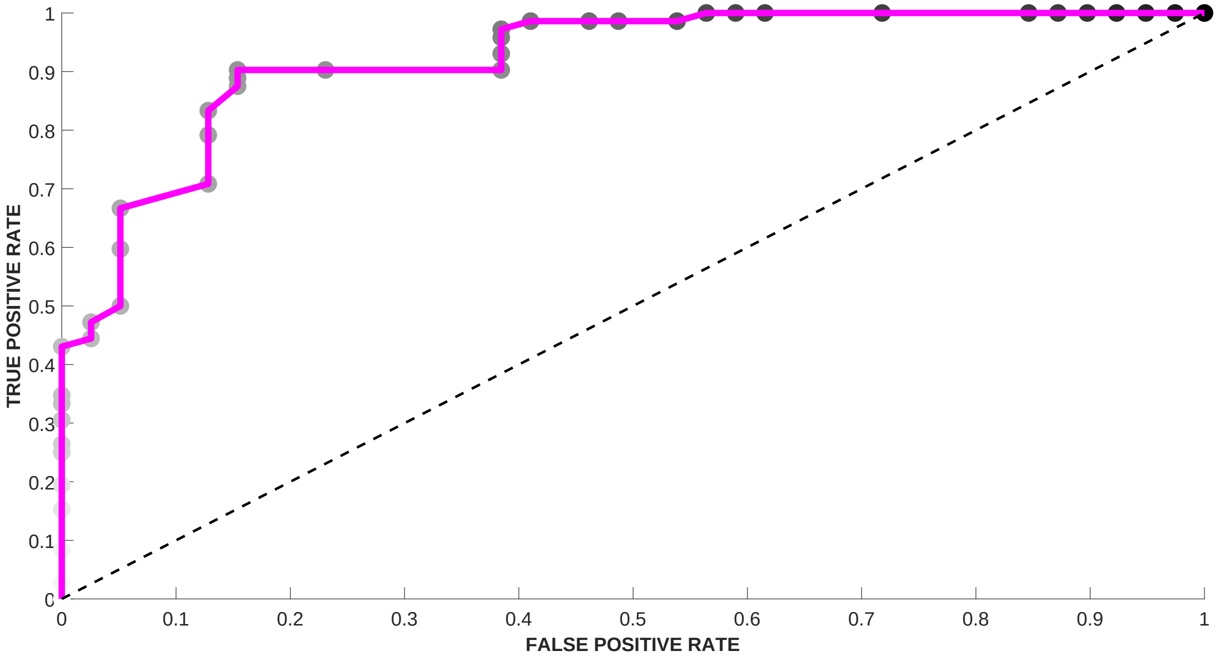


Figure 4: **ROC Curve with Varying Threshold Values (NY)**

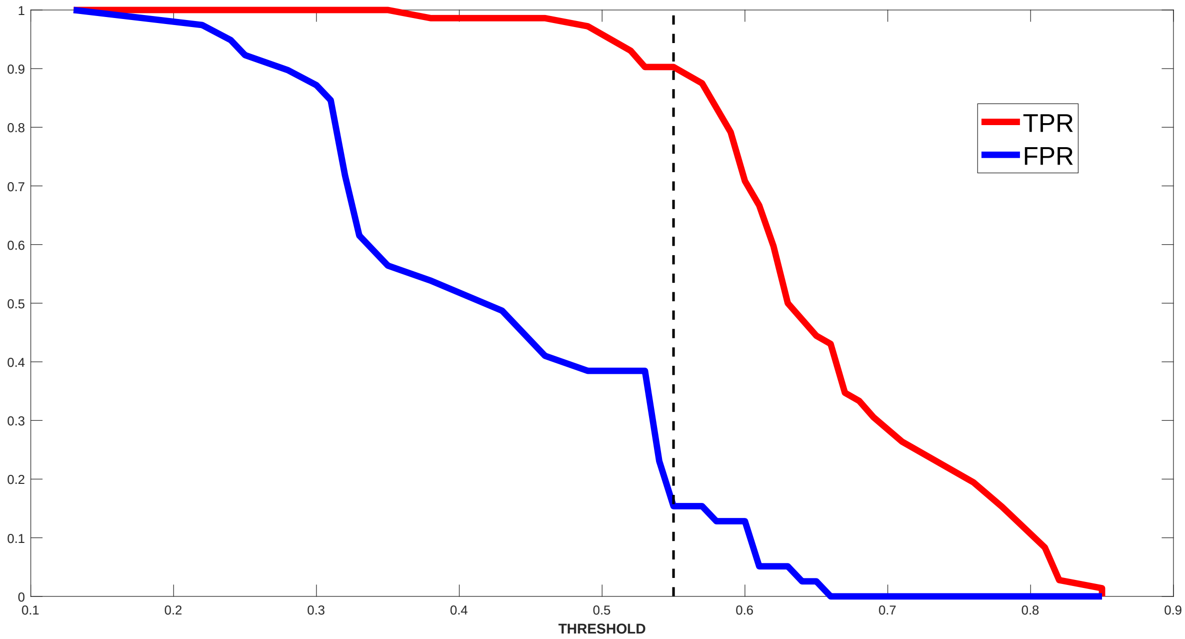


Figure 5: **TPR and FPR against Threshold (NY)**

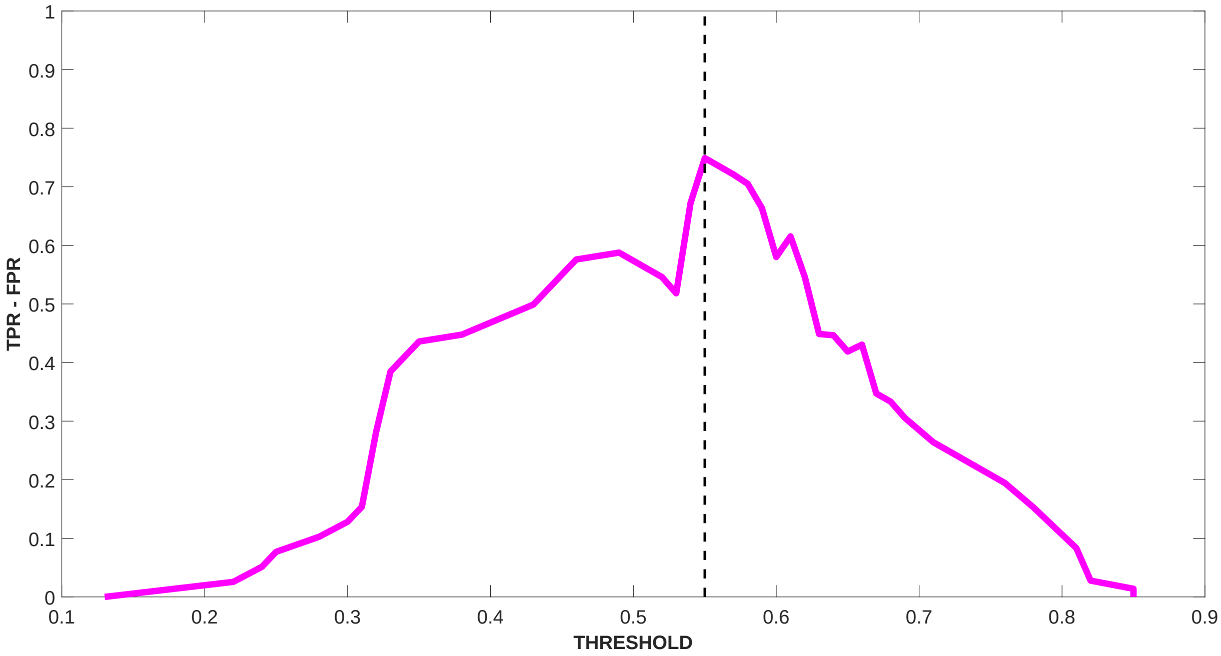


Figure 6: Difference between TPR and FTR against Threshold (NY)

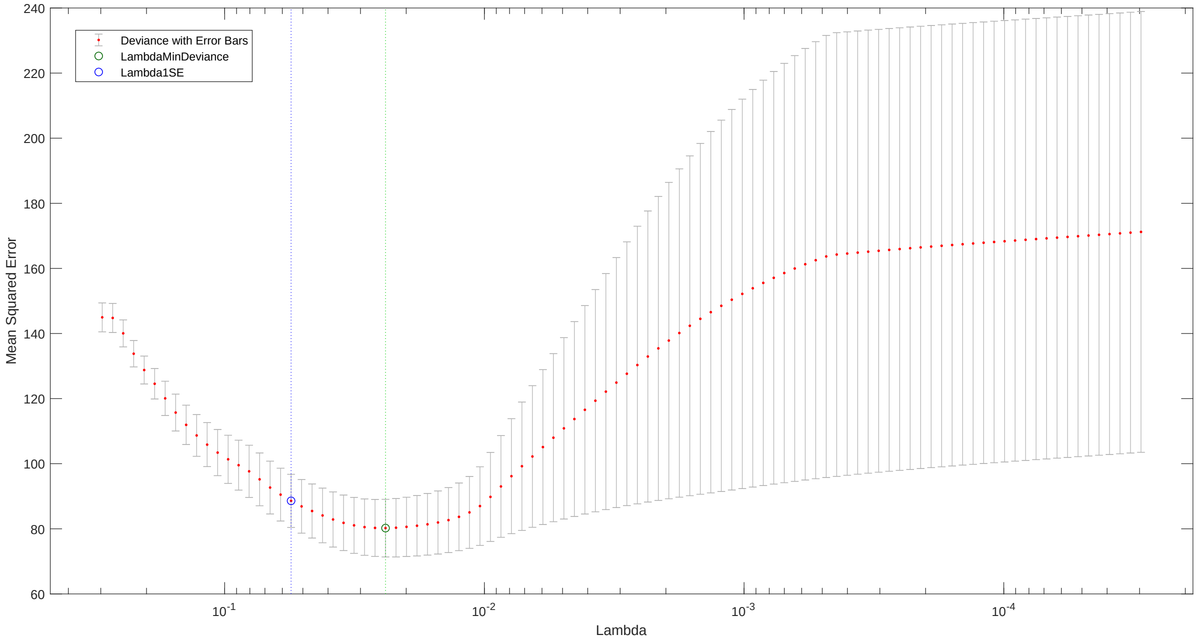


Figure 7: Cross-Validated Deviance of Lasso Fit (NY)