Stata Conference

Dario Sansone

2017 User Conference

Baltimore

GEORGETOWN UNIVERSITY

Now You See Me

High School Dropout and Machine Learning

Dario Sansone

Department of Economics
Georgetown University

Thursday July, 27th 2017

GEORGETOWN UNIVERSITY

Introduction

- U.S. High School graduation rate of **82%**, below OECD average. Extensive literature (Murnane, 2013)
- Goal: use ML in Education
- Create an algorithm to **predict which students are going to drop** out using only information available in 9th grade
- Current practices based on few indicators lead to **poor predictions**
- Improvements using **Big Data** and **ML**
- **Microeconomic foundations** of performance evaluations
- **Unsupervised ML** to capture heterogeneity among weak students

GEORGETOWN UNIVERSITY

Machine Learning

- **Econometrics: causal inference**
- **ML: prediction**
- Takes into account the trade-off between bias and variance in the MSE in order to maximize out-of-sample prediction.

- Algorithms can identify **patterns too subtle** to be detected by human observations (Luca et al, 2016)
- ML applications limited in economics, but several **policy-relevant** issues that require accurate predictions (Kleinberg et al., 2015)
- Ml is **gaining momentum**
  Belloni et al (2014), Mullainathan and Spiess (2017)
- Reduce **dropout rates in college**
  Aulck et al (2016), Ekowo and Palmer (2016)

GEORGETOWN UNIVERSITY

Machine Learning - References

Comprehensive review:

- J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, Springer.

MOOCs (w/o Stata):

- A. Ng, *Machine learning*, Coursera and Stanford University.
- J. Leek, R.D. Peng, B. Caffo, *Practical Machine Learning*, Coursera and Johns Hopkins University
- T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning*
- S. Athey and G. Imbens, NBER 2015 Summer Institute

Podcast for economist/policy:

- APPAM – The Wonk
- EconTalk

GEORGETOWN UNIVERSITY

Machine Learning - References

Intro for Economists:

- H.R. Varian, Big data: New tricks for econometrics, Journal of Economic Perspectives, 28(2):3–27, 2014
- S. Mullainathan and J. Spiess. Machine learning: An applied econometric approach. Journal of Economic Perspectives, 31(2):87–106, 2017

ML and Causal Inference:

- A. Belloni, V. Chernozhukov, and C. Hansen, *High-dimensional methods and inference on structural and treatment effects*, Journal of Economic Perspectives, 28(2):29–50, 2014
- S. Athey and G. Imbens, The State of Applied Econometrics: Causality and Policy Evaluation, Journal of Econometric Perspective, 31(2):3-32, 2017

GEORGETOWN UNIVERSITY

Goodness-of-fit

- No single indicator for binary choice model
- Option 1: comparison with a model which contains only a constant (**McFadden-R^2**)
- Option 2: **compare correct and incorrect predictions**

Advantage: clear distinction between type I (wrong exclusion) and type II (wrong inclusion) errors

- **Accuracy**: proportion correct predictions
- **Recall (Sensitivity)**: proportion correct predicted dropouts over all actual dropouts
- **Specificity**: proportion corrected predicted graduates over all actual graduates

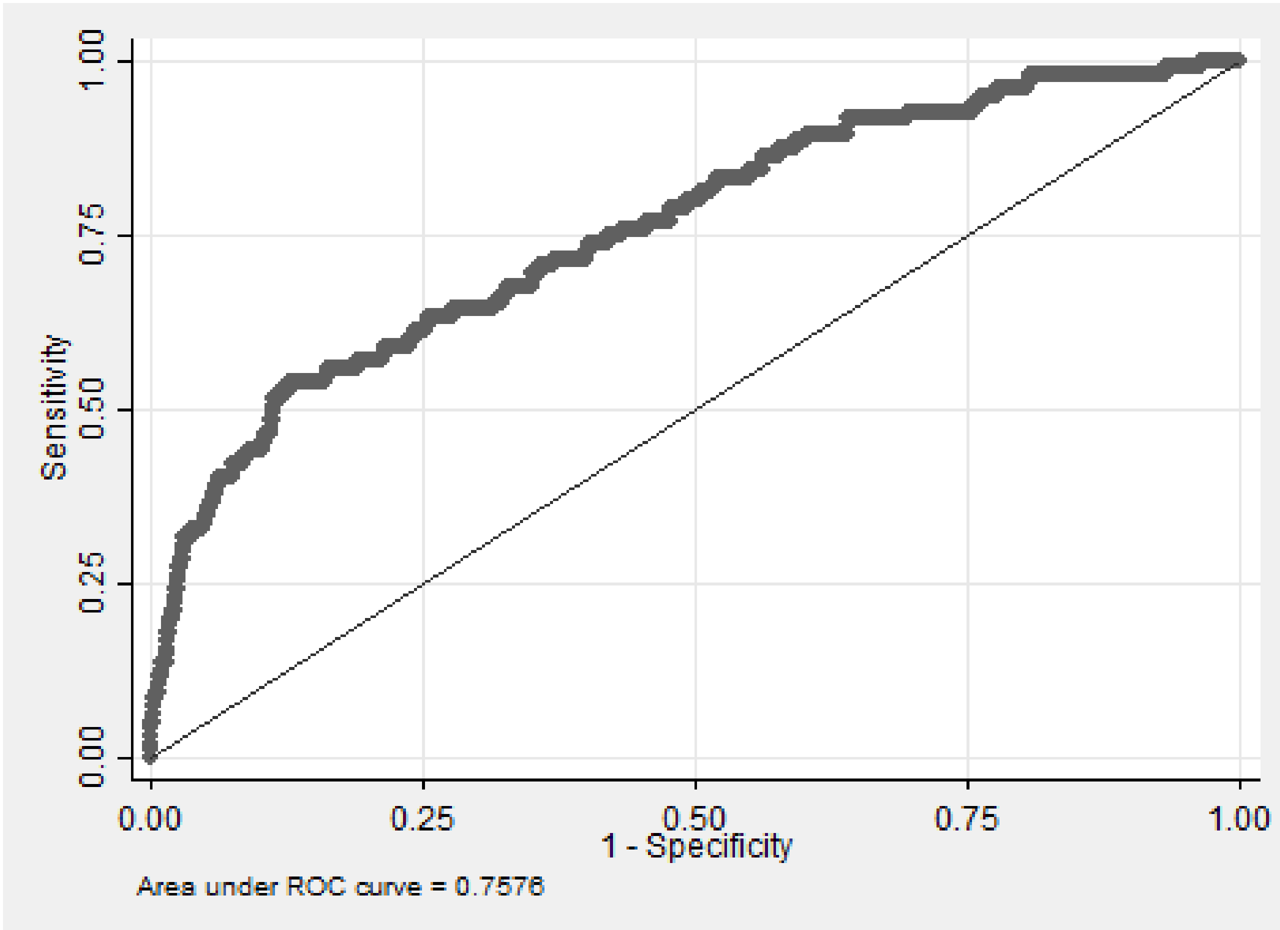
GEORGETOWN UNIVERSITY

ROC curve

- Most algorithms produce by default predicted probabilities
- Usually, predict 1 when probability > 0.5 (in line with Bayes classifier)
- **ROC curve** computes how Specificity and 1-Sensitivity change as the classification threshold changes
- **Area under the curve** used as evaluation criteria
- Stata code:

roctab depvar predicted_probabilities, graph

ROC curve - Example



GEORGETOWN UNIVERSITY

Cross-Validation

- Maximizing in-sample R^2 or Accuracy lead to **over-fitting** (high variance).
- Solution: Cross-Validation (CV). Divide sample in
  - **60% Training** sample: to estimate model
  - **20% CV** sample: to calibrate algorithm (e.g. penalization term)
  - **20% Test** sample: to report out-of-sample performances
- Advantage: easy to compare in-sample and out-of-sample performances (high bias vs. high variance)
- Alternatives: k-fold CV

GEORGETOWN UNIVERSITY

CV - Stata

```
set seed 1234
*generate random numbers
gen random = uniform()
sort random

*split sample in train (60%), CV (20%) and test (20%)
gen byte train = ( _n <= (_N*0.6) )
gen byte cv = ( ((_N*0.6) < _n) & (_n <= (_N*0.8)) )
gen byte test = ( _n > (_N*0.8) )
```

CV – foreach loop

1. For given parameters, estimate algorithm using training sample
2. Measure performances using CV sample
3. Repeat for different values of the parameters
4. Select values of the parameters which max performances in the CV sample
5. Estimate algorithm with selected parameters using training sample
6. Report performances in test sample

Data

- High School Longitudinal Study of 2009 (HSLS:09)
- Panel database **24,000 students** in 9th grade from 944 schools
- 1st round: students, parents, math and science teachers, school administrator, school counselor
- 2nd round: 11th grade (no teachers)
- 3rd round: freshman year in college
- Data on math test scores, HS transcripts, SAT, demographics, family background, school characteristics, expectations
- New perspective on **Millennials** and their educational choices

Dropout programs

- **45% of the students** in schools which have a formal dropout prevention program
- This may include tutoring, vocational courses, attendance incentives, childcare, graduation/job counseling
- How are students selected for these programs?
  - **Poor grades** (93%)
  - **Behind on credits** (89%)
  - Counselor's referral (86%)
  - Absenteeism (83%)
  - Parental request (77%)

Basic Model

- Include past student achievements, demographics, family background and school characteristics
- **Very low performances**

Model	Out-of-Sample		
	Obs	Accuracy	Recall
1- Logit	2,060	91.8%	7%
2- OLS	2,060	91.7%	0.6%
3- Probit	2,060	91.8%	5.3%
4- Logit + Interactions	2,060	91.5%	7%

GEORGETOWN UNIVERSITY

SVM + LASSO

- SVM better than Logit
- SVM + LASSO to **select variables** improves performance

	Out-of-Sample		
Model	Obs	Accuracy	Recall
1- SVM	2,540	80%	47%
2- SVM + LASSO	2,970	86%	50%

GEORGETOWN UNIVERSITY

Stata Code - Preparation

Important: all predictors have to have the same magnitude!

Option 1: **normalization** (consider not to normalize dummy var)

```
foreach var of global PREDICTOR {
    qui inspect `var'
    if r(N_unique)!=2 {
        qui sum `var'
        qui replace `var' = (`var'-r(mean))/r(sd)
    }
}
```

Option 2: **rescaling** (this does not alter dummy variables)

```
foreach var of global PREDICTOR {
    qui sum `var'
    qui replace `var' = (`var'-r(min))/(r(max)-r(min))
}
```

Stata Code – Preparation /2

How to deal with missing data:

- Option 1: **drop observations with missing items**
  - Cons: lose variables
  - Pros: easier to interpret when selecting variables
- Option 2: **impute missing values to zero** and create a dummy variable for each predictor to indicate which items were missing
- Try both!

Stata Code - LASSO

LASSO code provided by C. Hansen

- NO help file!
- Very fast
- Key assumption: sparsity (Most coefficients equal to 0)

Estimator:

$$\hat{\beta}(\lambda) = \underset{\beta \epsilon \mathbb{R}^k}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i' \beta)^2 + \lambda \|\beta\|_1$$

$$\|\beta\|_1 = \sum_{j=1}^{k} |\beta_j|$$

Stata Code – LASSO /2

lassoShooting *depvar indepvars [if] [, options]*

Options:

- lambda: select the penalization term. Use CV with grid-search
  0 is equal to the default (see Belloni et al., RES 2014)
- controls(varlist): specify variables which must be always selected (e.g. time fixed effects)
- lasiter: number of iterations of the algorithm (suggested 100)
- Display options: verbose(0) fdisplay(0)

Post-LASSO:

```
global lassoSel `r(selected)'
regress depvar $lassoSel if train==1
```

Stata Code - SVM

- Stata Journal article: svmachines
- Note: SVM cannot handle missing data
- Objective function similar to Penalized Logit
- **Combination with kernel functions** allow high flexibility (but low interpretability)
- Use grid-search with CV to calibrate algorithm:
  - Kernel: rbf (normal) is the most common. Try also sigmoid
  - C is the penalization term (similar to Lambda in LASSO)
  - Gamma controls the smoothness of the kernel
  - Select C and Gamma to balance trade-off between bias and variance

Stata Code - Boosting

- Stata Journal article: boosting
- Hastie's explanation on YouTube
- Note: cannot handle missing data
- Similar to random forest
- Combination of a sequence of classifiers where at each iterations observations which were misclassified by the previous classifier are given larger weights
- Key idea: **combining simple algorithms** such as regression trees can lead to higher performances than a single more complex algorithm such as Logit
- Works very well with highly nonlinear underlying models
- Works better with large datasets
- Can create graph with the influence of each predictor

GEORGETOWN UNIVERSITY

Additional ML codes

- Least Angle Regression (lars)
- Penalized Logistic Regression (plogit)
- Kernel-Based Regularized Least Squares (krls)
- Subset Variable Selection (gvselect)

- Key Missing: Neural Network

- Some of them are quite slow
- Double-check which criteria are used to calibrate parameters

GEORGETOWN UNIVERSITY

Pivotal Variables

- LASSO can also identify **top predictors**
  - If school wants to use few indicators, select best ones
  - Identify variables **worth collecting** at national level

- GPA 9th grade
- Credits in 9th grade
- Credits in 9th grade * SES
- Gender * vocational school
- Hours with friends * principal teaches
- Hours playing video games * private school
- Hours extra-curricular activities * hours counselors spends assisting students for college
- 9th grader talks with father about college * principal teaches
- Private school * % teachers absent
- Principal: students dropping out problem * lead counselor: counselors expect very little from students

GEORGETOWN UNIVERSITY

Microeconomic Foundation

- Justify using recall rate (φ)

$$\min E[dropout]$$
$$s.t.\ BC$$

- Define p(s,t) as the probability of dropping out for student type s ϵ {0,1} subject to treatment t ϵ {0,1}. φ = Recall Rate

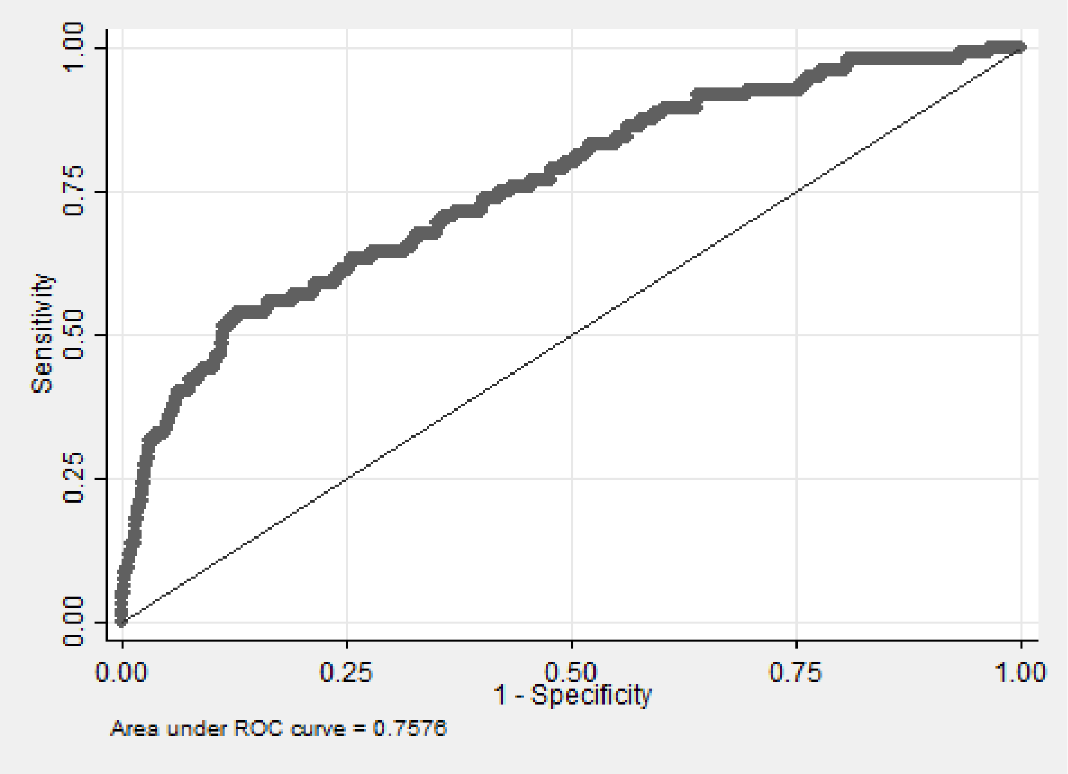
$$\min n_1[(1-\varphi)p(1,0)+\varphi p(1,t)]$$
$$s.t.\ \tau[wr_1+c_1]\leq B$$

- Imposing functional forms

$$\min\ (1-\varphi)$$
$$s.t.\ \tau[wr_1+c_1]\leq B$$

Application

- **Calibrate parameters** in the algorithms to maximize Recall Rate (Sensitivity) while respecting the B.C. (1 – Specificity).



GEORGETOWN UNIVERSITY

Unsupervised ML

- Divide weak students into clusters
- HS dropout is a **multi-dimensional** issue
- Possible applications:
  - Identify subpopulations and design **targeted treatments**
  - Measure **heterogeneity** treatment among subpopulations
- **Hierarchical clustering** identifies four groups:
  - All have low math achievements, low expectations
  - 1: HH without mother
  - 2: difficult environment
  - 3: poor Hispanic male students
  - 4: Blacks, repeated 9^{th} grade, difficult HH background

GEORGETOWN UNIVERSITY

Hierarchical clustering

1. n distinct groups, one for each observations
2. Two closest observations merged together (n-1 groups)
3. Closest two groups merged together (n-2 groups)
4. Repeat until all the observations are merged into one large group.

- The output: **hierarchy of groupings** from one group to n groups.
- **Four decisions** involved in this procedure
  - Measuring distance between observations
  - Measuring distance between groups
  - Selecting the number of observable variables
  - Selecting the optimal number of groups

GEORGETOWN UNIVERSITY

Hierarchical clustering - Stata

cluster *linkage [varlist] [if] [in] [, cluster_options]*

- **Distance between observation**: Euclidean (default in option *measure*)
- **Distance between groups**. Most common are:
  - Single Linkage: measure distance between two closest observations between groups
  - Complete Linkage: measure distance between two farthest observations between groups
  - Centroid Linkage: measure distance between two group means
  - **Average Linkage**: average distance between each point in one cluster to every point in the other cluster. More robust

GEORGETOWN UNIVERSITY

Number of groups

cluster stop *[clname] [, options]*

- General idea: ask whether splitting one cluster would reduce a certain measure of fit.
- Two criteria:
  - **Caliński and Harabasz pseudo-F index** *rule(calinski)*
  - **Duda-Hart Je(2)/Je(1) index** with pseudo-T^2 *rule(duda)*
- Distinct clustering is signaled by
  - High Caliński and Harabasz pseudo-F index
  - Large Je(2)/Je(1) index associated with a low pseudo-T^2 surrounded by much larger pseudo-T^2 values

GEORGETOWN UNIVERSITY

Caliński and Harabasz

It compares the sum of squared distances within the partitions - the distances between clusters - to that in the unpartitioned data, taking account of the number of clusters and number of cases. With q groups (C1,..., Cq) and n observations:

$$pseudoF_{CH} = \frac{trace(B_q)/(q-1)}{trace(W_q)/(n-q)}$$

$$B_q = \sum_{k=1}^{q} \mid C_k \mid \|\bar{c}_k - \bar{x}\|^2$$

$$\mid C_k \mid = \sum_{i=1}^{n} \mathbb{1}[x_i \in C_k]$$

$$W_q = \sum_{k=1}^{q} \sum_{i=1}^{n} \mathbb{1}[x_i \in C_k] \|x_i - \bar{c}_k\|^2$$

Where $\bar{x}$ is the centroid of the data, $\bar{c}_k$ is the centroid of the generic cluster C_k, and x_i is the vector of characteristics for individual *i*. B_q is the between-group dispersion matrix for the data clustered into *q* clusters, $|C_k|$ is the number of elements in cluster C_k, and W_q is the within-group dispersion matrix for the data clustered into *q* clusters.

GEORGETOWN UNIVERSITY

Duda-Hart

The Duda-Hart Je(2)/Je(1) index is literally the sum of squared errors within clusters in the two derived clusters (C_h and C_l) J(2), divided by the sum of squared errors in the combined original cluster (C_m) J(1).

$$Duda-Hart = \frac{J(2)}{J(1)} = \frac{W_h + W_l}{W_m}$$

Where *W* is defined as in the Caliński and Harabasz pseudo-F index.
The Duda-Hart T^2 statistic takes account of the number of observations in both clusters (n_h and n_l):

$$\frac{1}{J(2)/J(1)} = 1 + \frac{T^2}{n_h + n_l - 2}$$

GEORGETOWN UNIVERSITY

Policy Implications

- Early prediction → **Early intervention**
- **Efficient** use of data available to schools
- Suggest vocational tracks (Goux et al, 2016)
- ML can identify **top predictors** worth collecting when resources are scarce (developing countries)
- Include **inexpensive** alternative to the tests used to sort students
- Unsupervised ML to **personalize treatment**

GEORGETOWN UNIVERSITY

Thank you!

GEORGETOWN UNIVERSITY