MAIN STREET QUANT

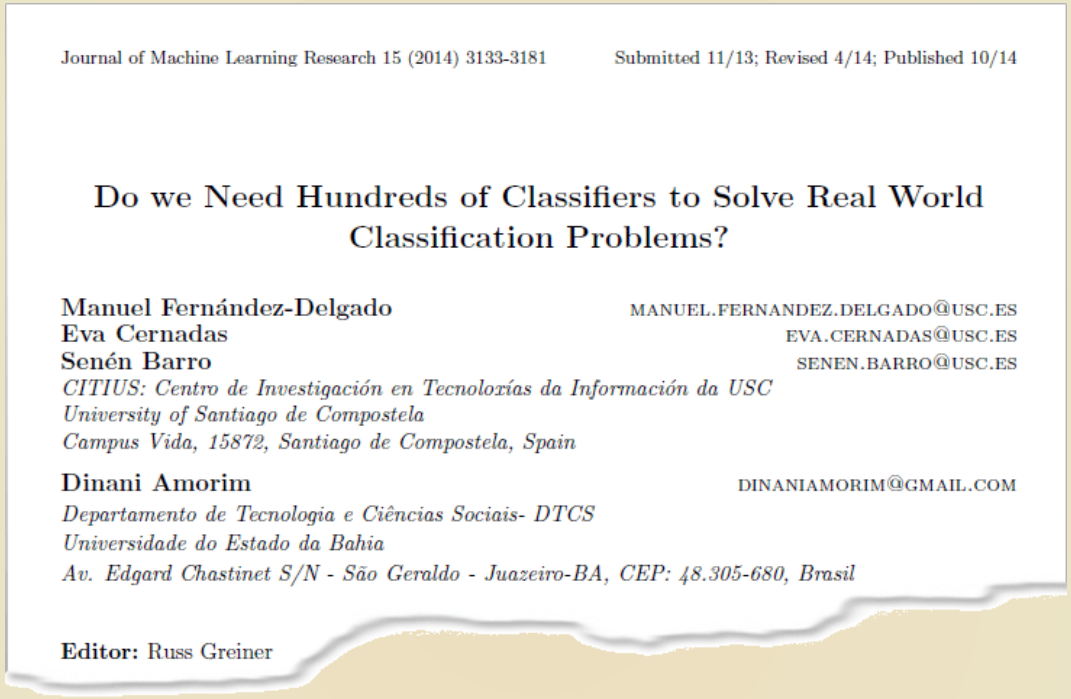
PRESENTS

PHILANTHROPIC ANALYTICS SHOWDOWN

BACKGROUND

- Expansion upon a graduate project / case competition
- Real life data: prominent veteran's charity, 13M+ donors
- Data:
  - Sample for model development, 50/50 balanced response, 60/40 partition
  - Frequency, Recency, Worth, Demographics
  - Costs for each mail piece
- Problem: losing money on the 'spray & pray'
  - Best expected response rate: 5.1%
- Benchmark logistic regression vs. classification tree in SAS JMP to predict likely donors
  - Maximize profit using classification under an asymmetric response
  - Generating the label printing list of likely donors

BACKGROUND



Journal of Machine Learning Research 15 (2014) 3133-3181 Submitted 11/13; Revised 4/14; Published 10/14

Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?

Manuel Fernández-Delgado MANUEL.FERNANDEZ.DELGADO@USC.ES
Eva Cernadas EVA.CERNADAS@USC.ES
Senén Barro SENEN.BARRO@USC.ES
CITIUS: Centro de Investigación en Tecnoloxías da Información da USC
University of Santiago de Compostela
Campus Vida, 15872, Santiago de Compostela, Spain

Dinani Amorim DINANIAMORIM@GMAIL.COM
Departamento de Tecnologia e Ciências Sociais- DTCS
Universidade do Estado da Bahia
Av. Edgard Chastinet S/N - São Geraldo - Juazeiro-BA, CEP: 48.305-680, Brasil

Editor: Russ Greiner

- Almost always parallel random forest (R/Caret), if not, then Gaussian SVM (libSVM)
- *"This is consistent with our experience running hundreds of Kaggle competitions: for most classification problems, some variation on ensembles decision trees (random forests, gradient boosted machines, etc.) performs the best."*

  - Ben Hamner, Co-founder & CTO

NARRATIVE

- Imagine that YOU are the Executive Director of your favorite non-profit…
  - Education, health, faith, politics, social good, etc.
  - Maximize donations, minimize costs
- Do you have more than a $1M in your treasury?
  - University foundations
  - Political campaigns
  - Major national charities
- Do you have the means to hire a Fundraising Manager?
  - Not a programmer, uses point-and-click
- Do you have the means to hire a Data Scientist/Programmer?
  - Higher salary

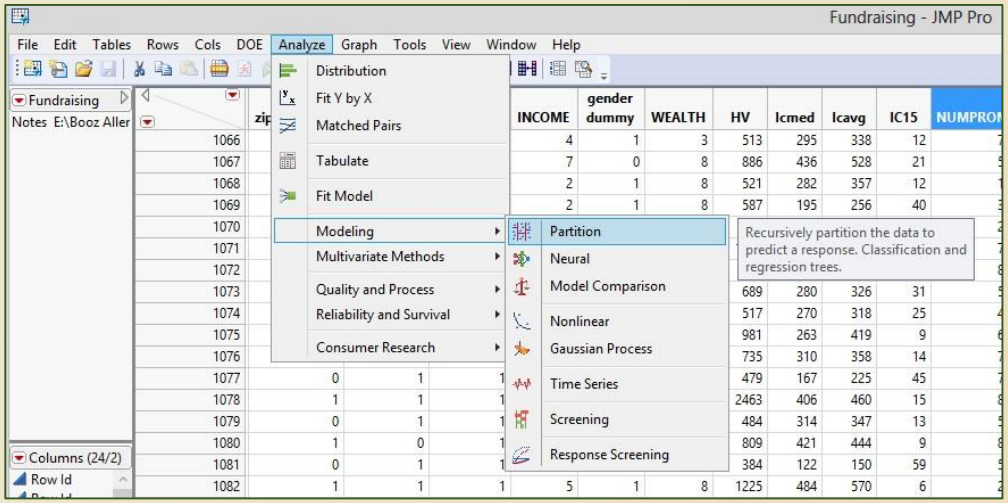
NET PROFIT

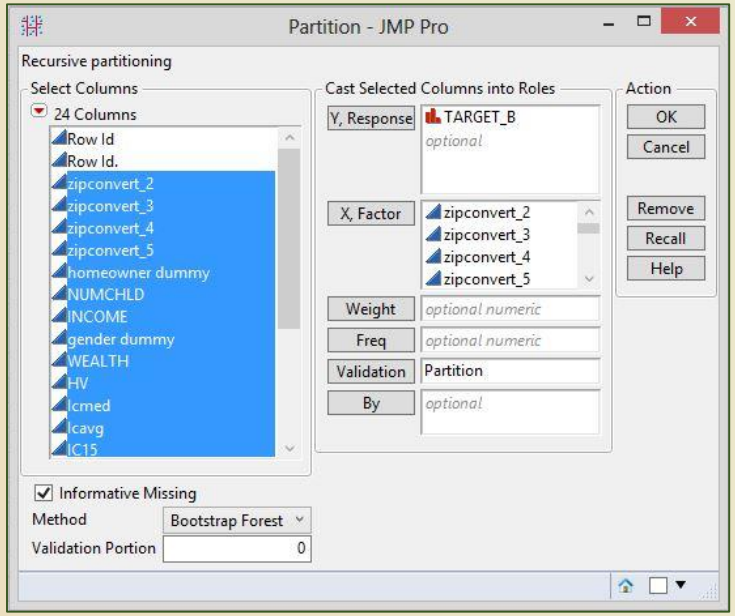
	Baseline (No Sort)
Role	
Projected Take	$8,619,000
Mailer Costs ($0.68 Each)	$8,840,000
Pieces to Send	13,000,000
Expected Response Rate	5.10%
Misclassification Rate	94.90%
Gross Profit Lift	
Gross Profit	-$221,000
Labor	
Burden (50%)	
Software	
Net Profit	-$221,000

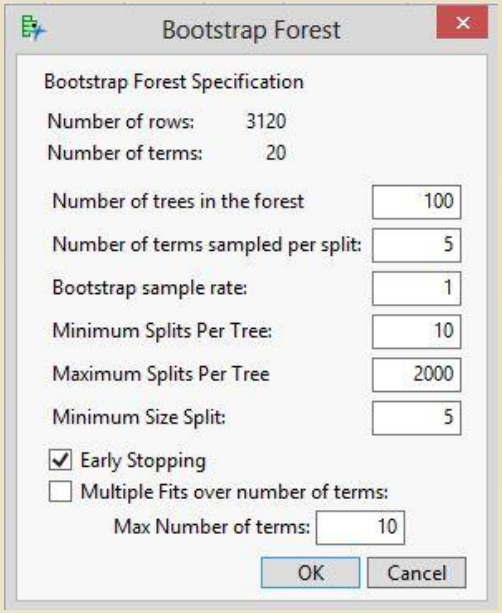
NET PROFIT

	Baseline (No Sort)	Logistic Regression (Excel)	Classification Tree (Excel)
Role		Consultant	
Projected Take	$8,619,000	$3,895,028	$5,679,558
Mailer Costs ($0.68 Each)	$8,840,000	$3,130,948	$4,888,602
Pieces to Send	13,000,000	4,604,334	7,189,120
Expected Response Rate	5.10%	6.51%	6.08%
Misclassification Rate	94.90%	35.99%	53.78%
Gross Profit Lift		445.74%	457.90%
Gross Profit	-$221,000	$764,080	$790,956
Labor		$4,300	$4,300
Burden (50%)			
Software			
Net Profit	-$221,000	$759,780	$786,656

CLICK PATH - JMP







B.STRAP FOREST

Bootstrap Forest for TARGET_B

Specifications

Target Column:	TARGET_B	Training rows:	1872
Validation Column:	Partition	Validation rows:	1248
		Test rows:	0
Number of trees in the forest:	100	Number of terms:	21
Number of terms sampled per split:	5	Bootstrap samples:	1872
		Minimum Splits Per Tree:	10
		Minimum Size Split:	5

Overall Statistics

Measure	Training	Validation	Definition
Entropy RSquare	0.3140	0.0031	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.4705	0.0057	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.4752	0.6900	$\sum$ -Log(ρ[j])/n
RMSE	0.3807	0.4984	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.3742	0.4892	$\sum \lvert y[j]-\rho[j]\rvert/n$
Misclassification Rate	0.0449	0.4503	$\sum$ (ρ[j]≠ρMax)/n
N	1872	1248	n

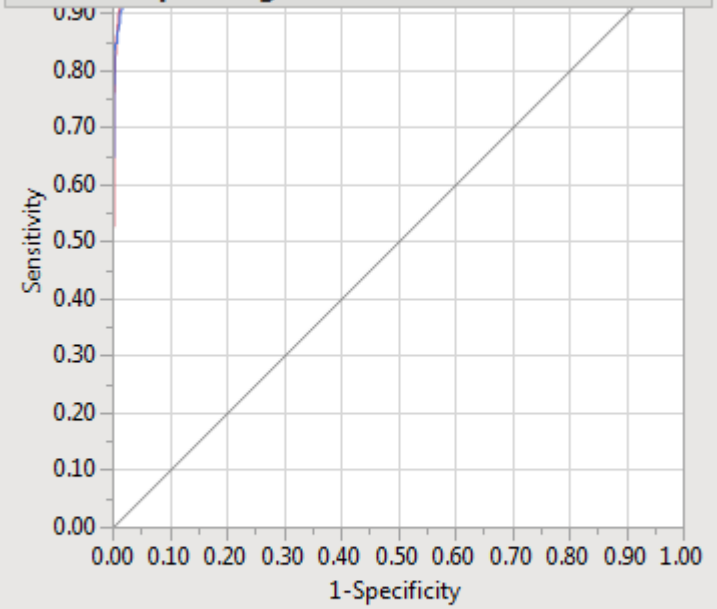
Confusion Matrix

Actual	Predicted		Actual	Predicted	
Training	**0**	**1**	**Validation**	**0**	**1**
0	943	21	0	397	199
1	63	845	1	363	289

B.STRAP FOREST

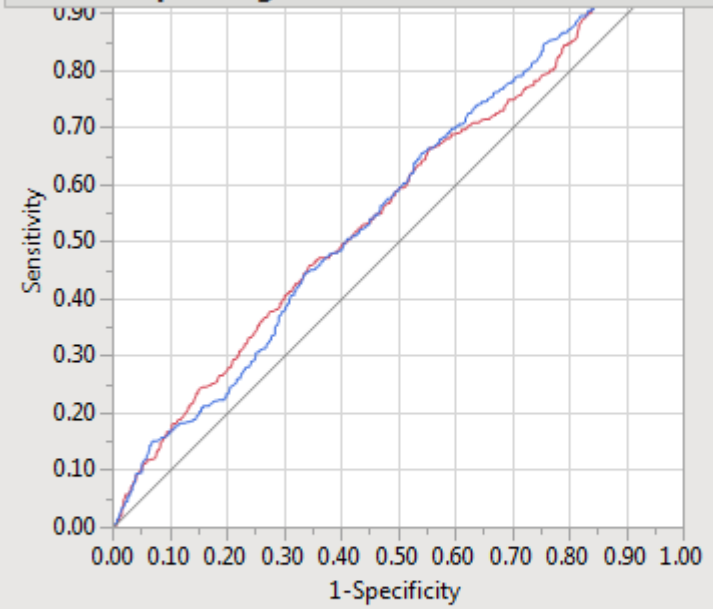
Bootstrap Forest for TARGET_B

Receiver Operating Characteristic



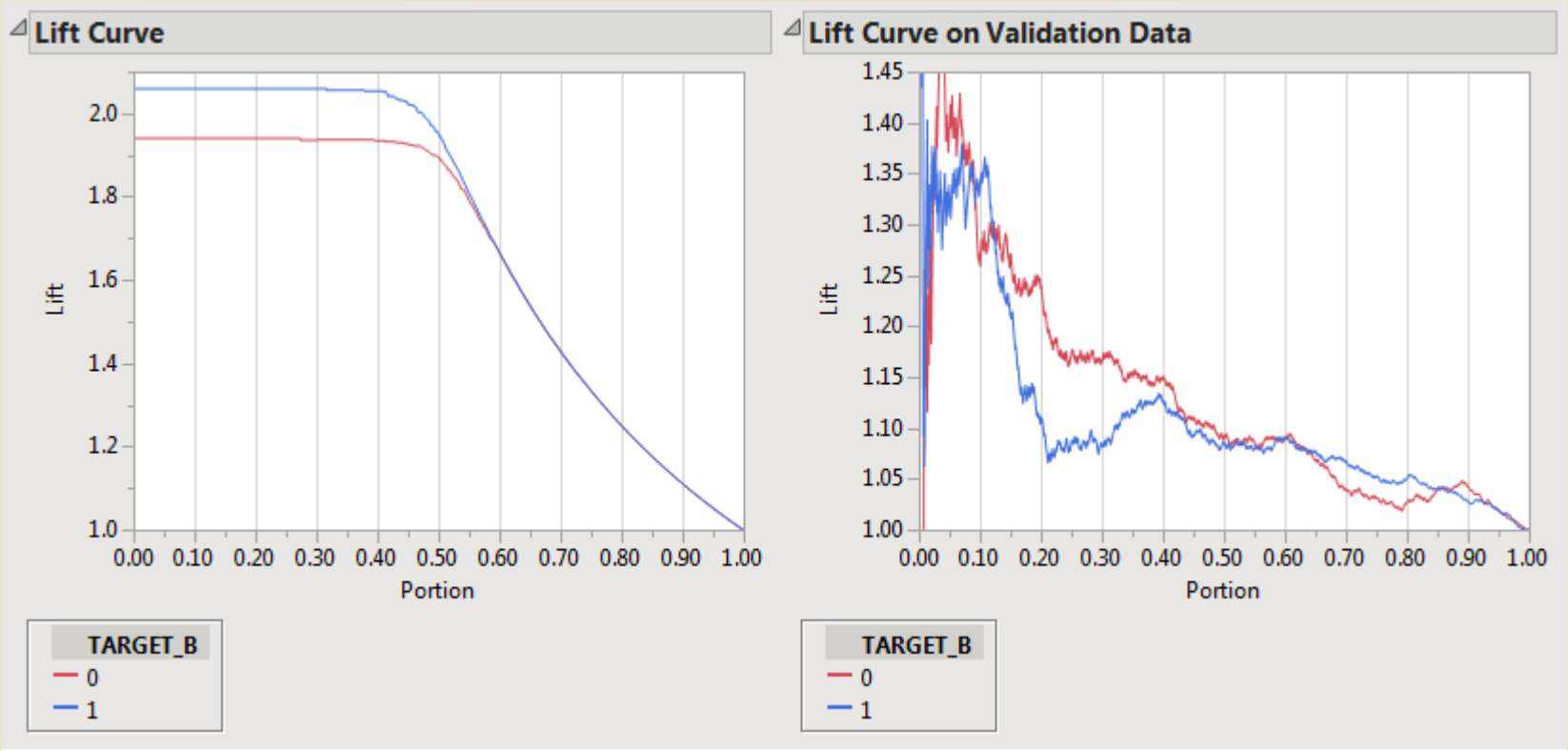
TARGET_B	Area
0	0.9942
1	0.9942

Receiver Operating Characteristic on Validation Data



TARGET_B	Area
0	0.5712
1	0.5712

B.STRAP FOREST



NET PROFIT

	Baseline (No Sort)	Logistic Regression (Excel)	Classification Tree* (SAS JMP)	Bootstrap Forest* (SAS JMP)
Role		Consultant		Fundraiser
Projected Take	$8,619,000	$3,895,028	$5,679,558	$6,265,193
Mailer Costs ($0.68 Each)	$8,840,000	$3,130,948	$4,888,602	$1,601,830
Pieces to Send	13,000,000	4,604,334	7,189,120	2,359,966
Expected Response Rate	5.10%	6.51%	6.08%	20.42%
Misclassification Rate	94.90%	35.99%	53.78%	15.26%
Gross Profit Lift		445.74%	457.90%	2,208.76%
Gross Profit	-$221,000	$764,080	$790,956	$4,660,363
Labor		$4,300	$4,300	$48,500
Burden (50%)				$24,250
Software				$11,000
Net Profit	-$221,000	$759,780	$786,656	$4,576,613

MATRICES

Logistic Regression (REWEIGHTED)

	Training			Validation		
	Predicted 1	Predicted 0	Total	Predicted 1	Predicted 0	Total
Actual 1	418	490	908	287	365	652
Actual 0	6,240	10,656	16,896	3,889	8,244	12,133
Total	6,658	11,146	17,804	4,176	8,609	12,785

Tree Partition (REWEIGHTED)

	Training			Validation		
	Predicted 1	Predicted 0	Total	Predicted 1	Predicted 0	Total
Actual 1	601	307	908	427	225	652
Actual 0	9,658	7,238	16,896	6,230	5,903	12,133
Total	10,259	7,545	17,804	6,657	6,128	12,785

Bootstrap Forest (REWEIGHTED)

	Training			Validation		
	Predicted 1	Predicted 0	Total	Predicted 1	Predicted 0	Total
Actual 1	845	63	908	289	363	652
Actual 0	368	16,528	16,896	4,051	8,082	12,133
Total	1,213	16,591	17,804	4,340	8,445	12,785

Misclassification rates (re-weighted)

Logistic Regression:	Training	37.80%
	Validation	33.27%
	Meta	35.99%
Classification Tree:	Training	55.97%
	Validation	50.49%
	Meta	53.78%
Bootstrap Forest:	Training	2.42%
	Validation	34.53%
	Meta	15.26%

R CODE

```
library("e1071", lib.loc="~/R/win-library/3.1")
training <- read.csv("training.csv")
testing <- read.csv("testing.csv")
svmfit <- svm(TARGET_B~.,data=training)
predict(svmfit,testing)
write.csv(trainsvmresult, file = "trainsvmresult.csv")
write.csv(testsvmresult, file = "testsvmresult.csv")
```

NET PROFIT

	Baseline (No Sort)	Logistic Regression (Excel)	Classification Tree* (SAS JMP)	Bootstrap Forest* (SAS JMP)	Radial SVM (R)
Role		Consultant		Fundraiser	Programmer
Projected Take	$8,619,000	$3,895,028	$5,679,558	$6,265,193	$4,552,486
Mailer Costs ($0.68 Each)	$8,840,000	$3,130,948	$4,888,602	$1,601,830	$2,787,921
Pieces to Send	13,000,000	4,604,334	7,189,120	2,359,966	4,099,883
Expected Response Rate	5.10%	6.51%	6.08%	20.42%	8.54%
Misclassification Rate	94.90%	35.99%	53.78%	15.26%	31.08
Gross Profit Lift		445.74%	457.90%	2,208.76%	898.45%
Gross Profit	-$221,000	$764,080	$790,956	$4,660,363	$1,764,565
Labor		$4,300	$4,300	$48,500	$80,000
Burden (50%)				$24,250	$40,000
Software				$11,000	
Net Profit	-$221,000	$759,780	$786,656	$4,576,613	$1,644,565

MATRICES

Logistic Regression (REWEIGHTED)

Training	Predicted 1	Predicted 0	Total
Actual 1	418	490	908
Actual 0	6,240	10,656	16,896
Total	6,658	11,146	17,804

Validation	Predicted 1	Predicted 0	Total
Actual 1	287	365	652
Actual 0	3,889	8,244	12,133
Total	4,176	8,609	12,785

Tree Partition (REWEIGHTED)

Training	Predicted 1	Predicted 0	Total
Actual 1	601	307	908
Actual 0	9,658	7,238	16,896
Total	10,259	7,545	17,804

Validation	Predicted 1	Predicted 0	Total
Actual 1	427	225	652
Actual 0	6,230	5,903	12,133
Total	6,657	6,128	12,785

Bootstrap Forest (REWEIGHTED)

Training	Predicted 1	Predicted 0	Total
Actual 1	845	63	908
Actual 0	368	16,528	16,896
Total	1,213	16,591	17,804

Validation	Predicted 1	Predicted 0	Total
Actual 1	289	363	652
Actual 0	4,051	8,082	12,133
Total	4,340	8,445	12,785

Support Vector Machine (REWEIGHTED)

Training	Predicted 1	Predicted 0	Total
Actual 1	537	371	908
Actual 0	4,487	12,409	16,896
Total	5,024	12,780	17,804

Validation	Predicted 1	Predicted 0	Total
Actual 1	287	365	652
Actual 0	4,336	7,797	12,133
Total	4,623	8,162	12,785

Misclassification rates (re-weighted)

Logistic Regression:	Training	37.80%
	Validation	33.27%
	Meta	35.99%
Classification Tree:	Training	55.97%
	Validation	50.49%
	Meta	53.78%
Bootstrap Forest:	Training	2.42%
	Validation	34.53%
	Meta	15.26%
Support Vector Machine:	Training	27.29%
	Validation	36.77%
	Meta	31.08%

TAKE-AWAYS

- Even small improvements in misclassification rates can lead to big financial gains.
- Expensive services do no necessarily yield the best results.
- Expensive software does not necessarily yield the best results.
- The world is attempting to automate and democratize statistical functions presently executed with programming:
    - Pro: Saves time and effort
    - Con: Greater use can lead to greater misuse. To wit:
        - Data Cleaning
        - Checking for Normality, Heteroskedacity, Multicolinearity, Endogeneity, Variable Reduction
        - Drawing Statistical Inference from Machine Learning

GITHUB REPO

https://github.com/JD-Freeman/Philanthropic-Analytics-Showdown