

FEATURE IMPORTANCE AND ENSEMBLE METHODS

A new perspective on feature utilization



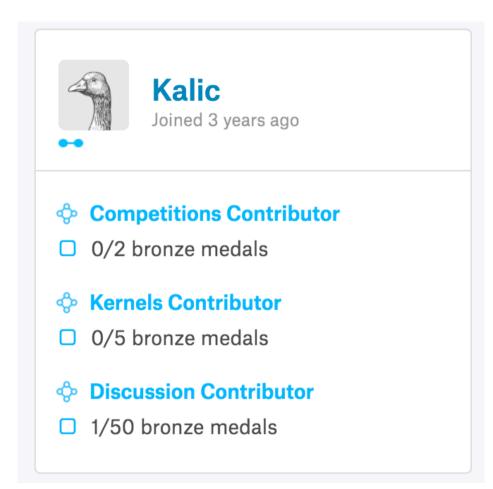
BIG DATA VILNIUS- NOVEMBER 2017 CONSTANT BRIDON

34 AVENUE DE L'OPERA 75002 PARIS > FRANCE > WWW.OCTO.COM





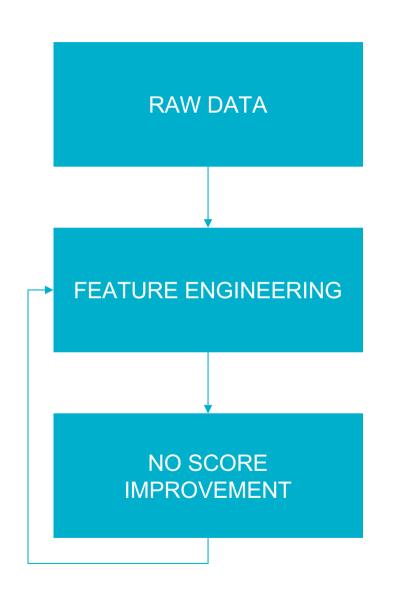
Home of frustration and despair

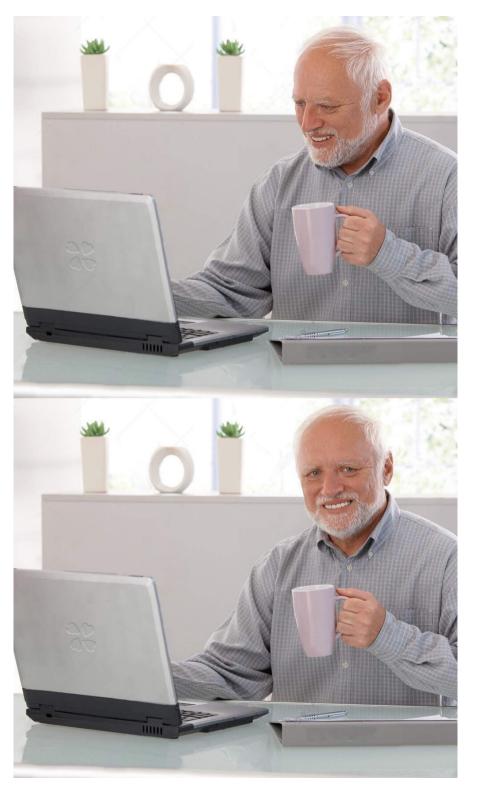




○ < + ○ ▶ ○ < + ○ ▶ ○ < + ○ ▶ ○ < + ○ ▶ ○ < + ○ ▶ ○ < + ○</p>

FEATURE ENGINEERING IS CRUEL







RECALLS & DEFINITIONS 02 > FEATURE IMPORTANCE & ENSEMBLE METHODS

GRADIENT BOOSTING : TREECEPTION

POINT FEATURE IMPORTANCE ?

05 > LIMITS & WAYS FORWARD

06 > ANNEXES

OCTO TECHNOLOGY > THERE IS A BETTER WAY



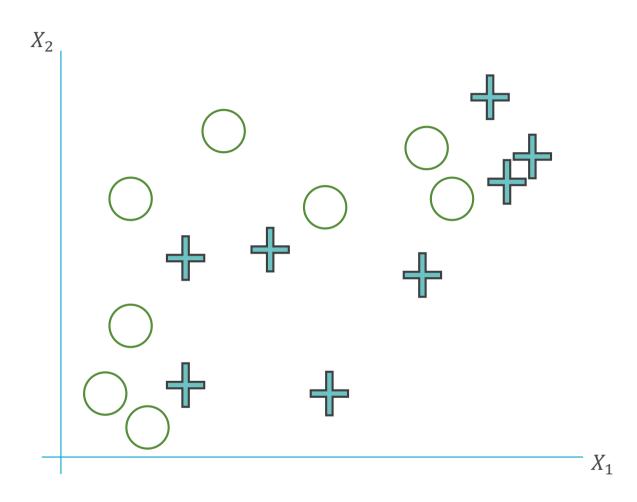
0<+0>0<+0>0<+0>0<+0>0<+0>0<+0<+0

O RECALLS & DEFINITIONS

 $\circ \blacktriangleright \circ < + \circ \vdash \circ < + \circ$

BOOSTING : LEARNING FROM YOUR MISTAKES

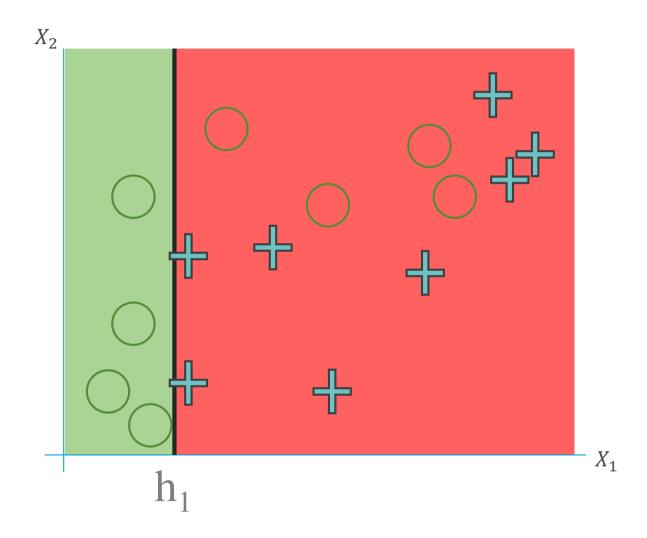
Simple binary classification problem using two variables





BOOSTING- WEAK LEARNERS - ROUND 1

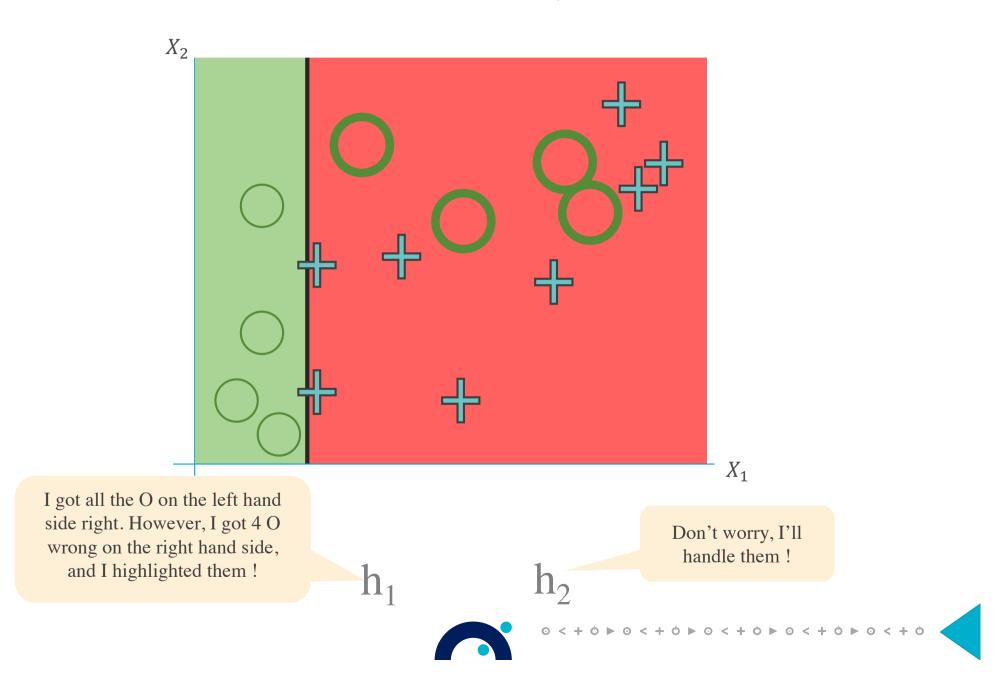
First weak learner on one variable splits the dataset





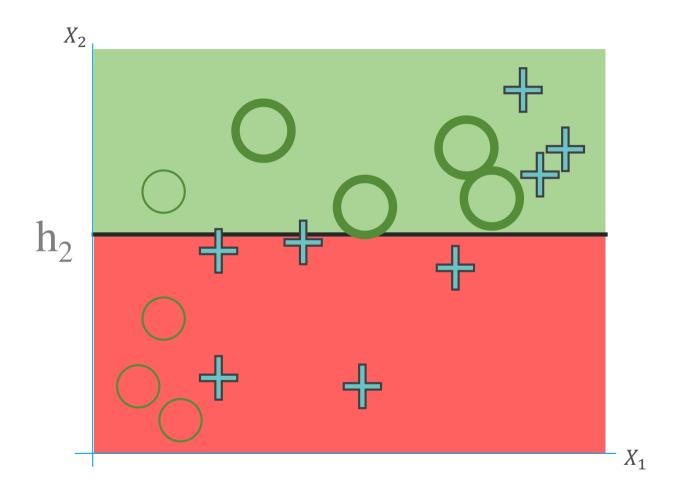
BOOSTING- WEAK LEARNERS - ROUND 2

Since first weak learner is too weak, it asks help from second weak learner



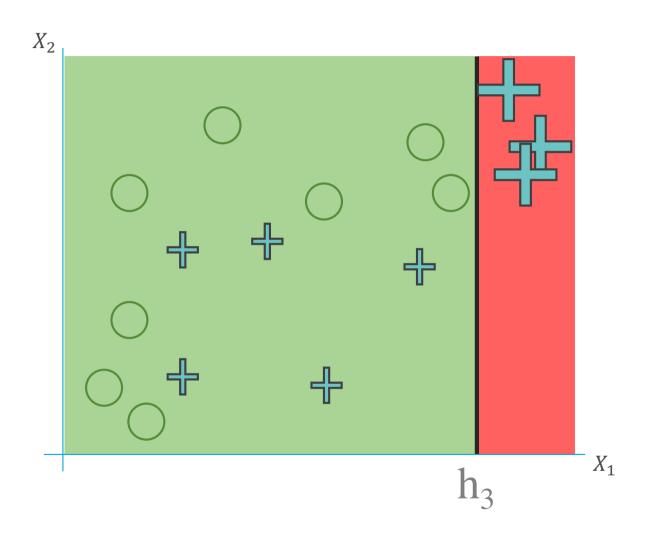
BOOSTING- WEAK LEARNERS - ROUND 2

Second weak learner correctly predicts the mistakes of first weak learner



ADABOOST – WEAK LEARNERS – ROUND 3

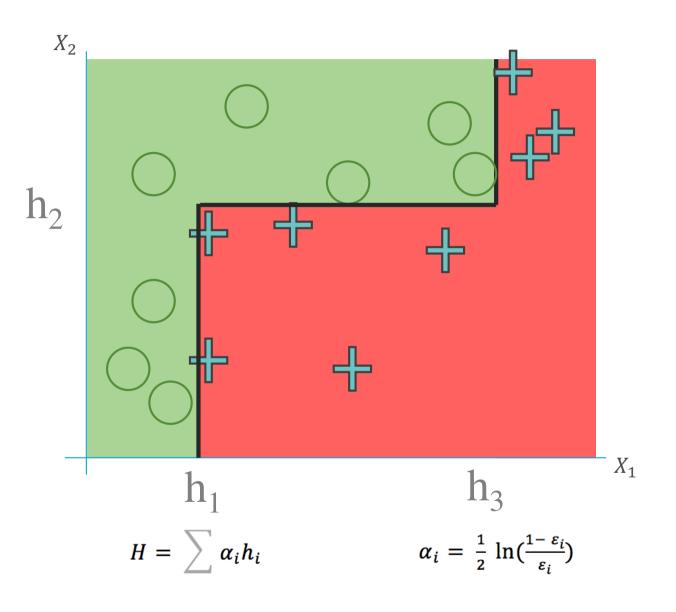
But second weak learner also needs help...



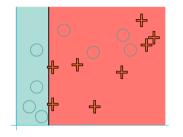


BOOSTING- WEAK LEARNERS - ROUND FINAL

Altogether the three weak learners make a strong learner

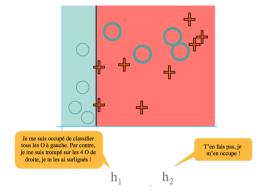


GRADIENT BOOSTING



- At the first stage, the overall hypothesis H is equal to h1
- What is the role of h2 ? Make sure that h1 + h2 is equal to y, i.e

$$\forall i \in (1, \dots, n) \quad H(x_i) + h_2(x_i) \approx y_i$$



$$\forall i \in (1, \dots, n) \qquad h_2(x_i) \approx y_i - H(x_i)$$

SOLUTION $h_2. fit(X, y - H)$

With,

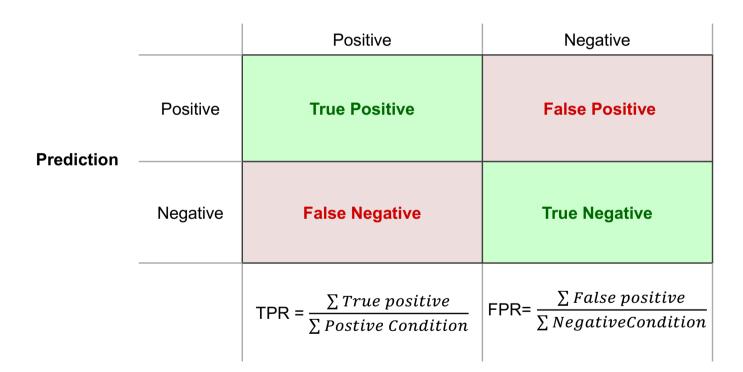
Then,

$$j(y,H(x)) = \frac{(y-H(x))^2}{2} \qquad H(x_i) \coloneqq H(x_i) - \frac{\partial J}{\partial H(x_i)} \quad \forall i \in (1, \dots, n)$$



ROC AUC – A METRIC FOR CLASSIFICATION

Definition of the confusion matrix



Real value



ROC CURVE - A PARAMETRIC CURVE WITH THE THRESHOLD

How to post process your predictions ?

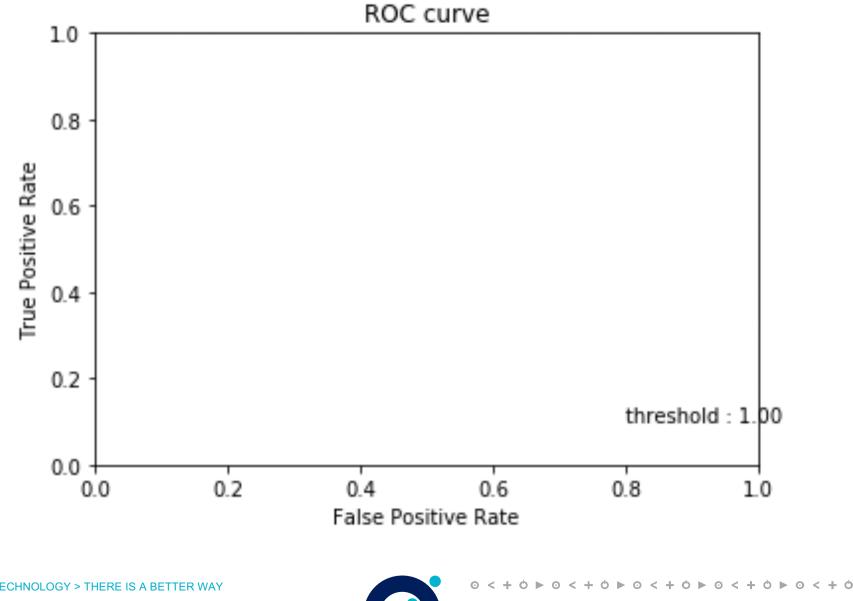
	pred	true
43123	0.229203	0
24080	0.232772	0
70632	0.236211	0
15958	0.239722	0
51413	0.243677	0
61074	0.248302	0
63564	0.253800	0
	0.260722	1
	0.280703	
58454	0.298216	0
13160	0.314317	1
66162	0.333712	1
73577	0.365904	1
44743	0.410899	1
55067	0.469812	1
29718	0.541483	1
71963	0.617832	1
47659	0.713352	1
18209	0.828930	1

 According to the threshold on the predicted probability, I am going to generate False Positive, or False Negative

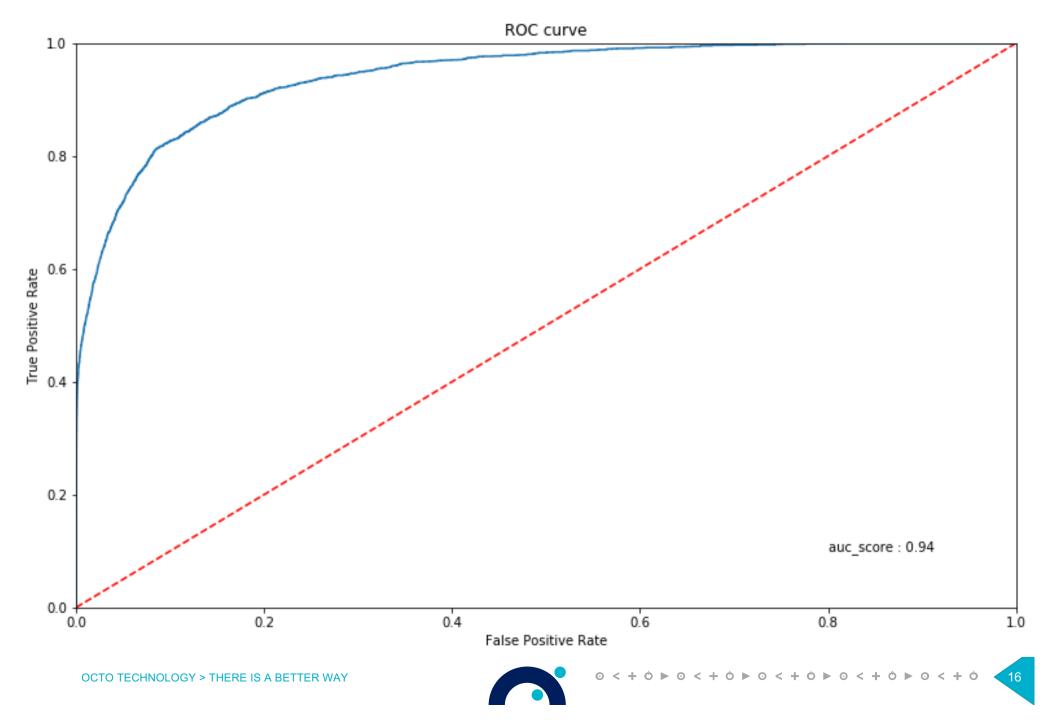


ROC CURVE - A PARAMETRIC CURVE WITH THE THRESHOLD

We can plot the TPR wrt to the FPR according to the threshold



ROC AUC – AREA UNDER THE CURVE

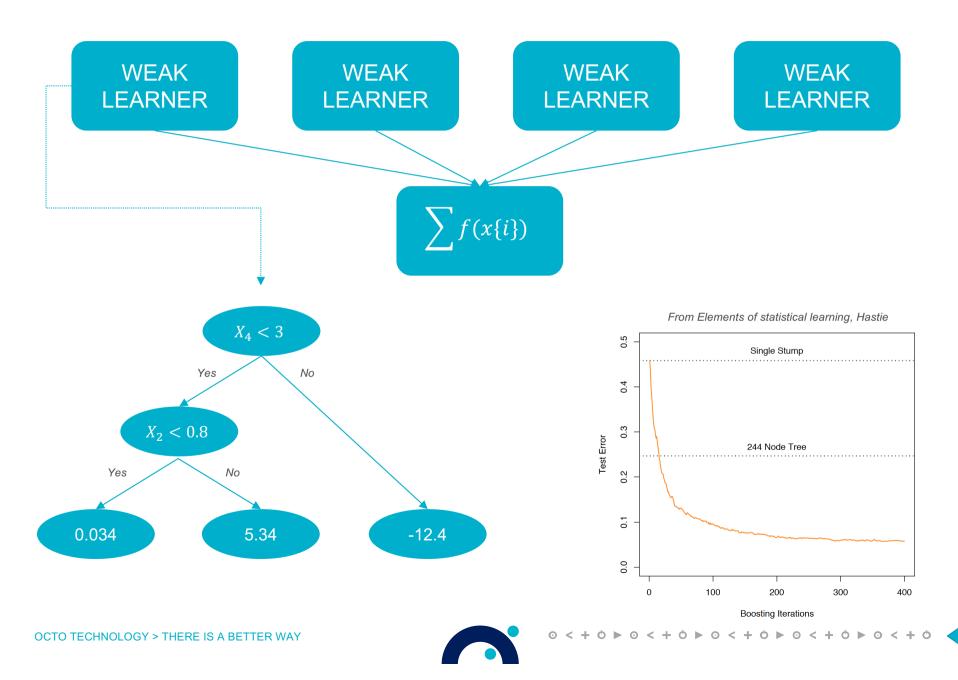




 $\circ \blacktriangleright \circ < + \circ \vdash \circ < + \circ \vdash \circ < + \circ \vdash \circ < + \circ$

ENSEMBLE METHODS : JOIN THE PACK

A collection of weak learners is better than a unique large learner

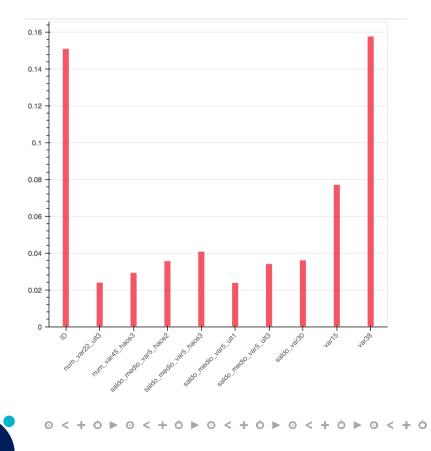


FEATURE IMPORTANCE ? JUST COUNT THEM UP !

Each split is equally important, so feature importance is just a cumulative sum

kaggle	Search kaggle	(۹	Competitions	Datasets	Kernels	Discussion	Jobs	•••	Sign In
	Santander		rs are	stomer Satis happy customers s · a year ago						

- Xgboost binary classification
- 0.83 ROCAUC (cross validated)
- Feature importance computation

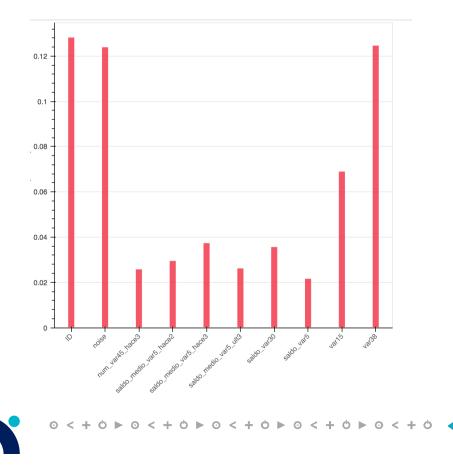


CUMULATIVE SUM IS NOT ROBUST

Adding a noise variable changes the importance of features

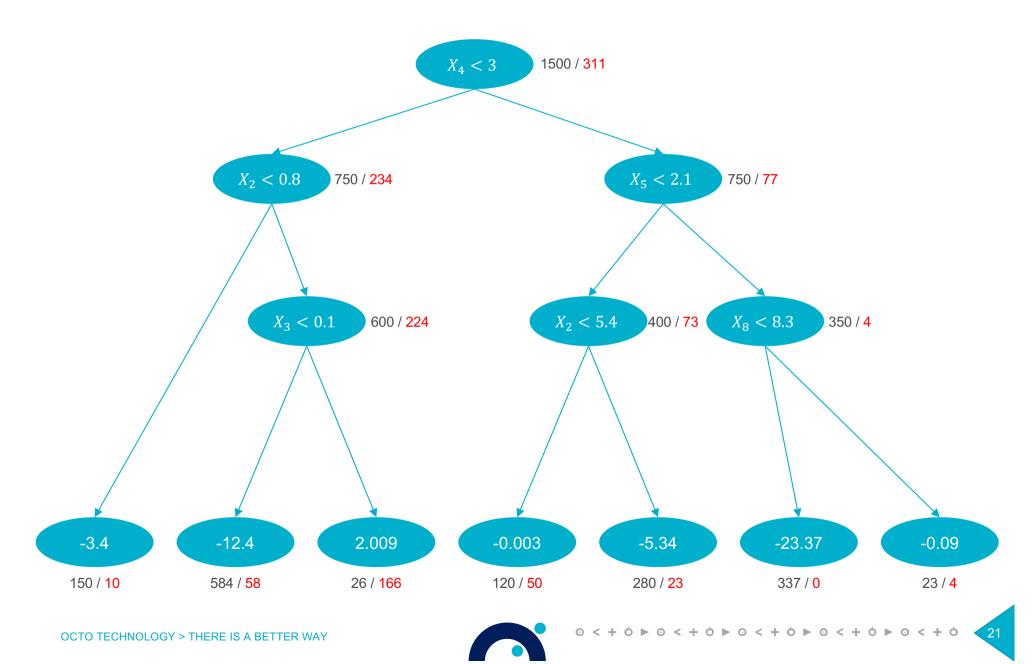
• Without any feature engineering, add random noise (uniform distribution between 0-1)

- Xgboost binary classification
- SAME PERFORMANCE (0.83)
- Feature importance computation



TREES ARE HIERARCHICAL

A feature importance is associated with the number of occurrences it handles





 $\circ \blacktriangleright \circ < + \circ \blacktriangleright \circ < + \circ \blacktriangleright \circ < + \circ \land \circ < + \circ \circ \circ < + \circ \circ \circ < + \circ \circ \circ > = = \circ \circ \circ \circ \circ \circ \circ \circ$

GRADIENT BOOSTING STRUCTURE

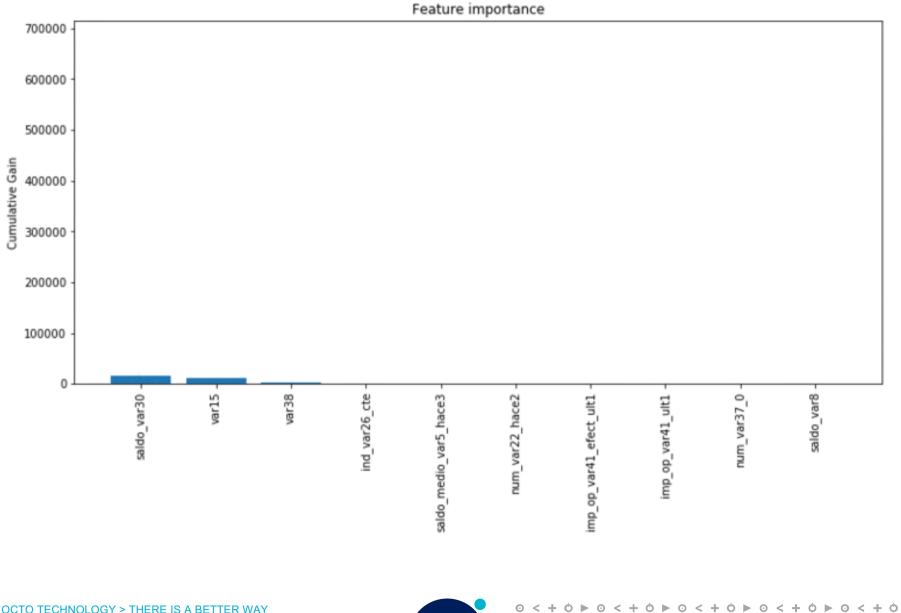
Trees make mistakes to train subsequent trees

2 v 0: [saldo_var3&-2.955] yes=1, no=2, missing=1, gain=223.785, cover=12204 3 v 1: [var15<28] yes=3, no=4, missing=301.257, cover=4276.25 4 v 3: [saldo_var8<-2.085] yes=7, no=8, missing=7, gain=3.61861, cover=2239 5 v 7: [var3&-1380] yes=15, no=16, missing=15, gain=3.6216, cover=2238 15: leaf=-0, cover=1 7 v 16: [imp_op_var39_comer_ult3<1613.43] yes=27, no=28, missing=27, gain=2.73829, cover=2237 27: [var15<26] yes=43, no=46, missing=45, gain=1.56854, cover=2235.75 4 v 17: [var15<26] yes=43, no=46, missing=45, gain=1.56854, cover=2235.75 4 v 17: [var15<26] yes=43, no=46, missing=45, gain=1.56854, cover=2235.75 4 v 19: [var15 8 v 10: leaf=-0, cover=1 7 v 10: leaf=-0, 0012053] yes=99, no=100, missing=99, gain=0.854701, cover=2.25 10: leaf=-0, 101111, cover=1.25 10: leaf=-0, 102529, cover=1886.25 11 v 16: leaf=-0, 1029229, cover=1886.25 12 100: leaf=-0, 1029239, cover=318, 5 73: [var38<47875.7] yes=101, no=102, missing=101, gain=3.33744, cover=346 101: leaf=-0, 103333, cover=27.5 102: leaf=-0, 103333, cover=218.5 74: leaf=-0, 0222222, cover=1.25 28: leaf=-0, cover=1 21 v 4: [var38<417321] yes=9, no=10, missing=9, gain=55.9796, cover=2037.25 9: [var38<47305.9] yes=17, no=18, missing=77, gain=5.3777, cover=636 29: [ind_var30_0<1] yes=47, no=48, missing=74, gain=5, 02247, cover=584.5 29: [ind_var30_0<1] yes=47, no=48, missing=47, gain=5, 02247, cover=584.5 29: [ind_var30_0<1] yes=47, no=48, missing=77, gain=0.141027, cover=12 75: [leaf=-0.13333, cover=11 76: [leaf=-0, 05; cover=1 48: [imp_op_var41_cfect_ult3<405] yes=77, no=78, missing=77, gain=0.141027, cover=572.5 103: [leaf=-0, 01812, cover=426.75 104: [leaf=-0.10812, cover=426.75 78: [lsaldo_umedi_var8_ult1<15.8] yes=105, no=106, missing=105, gain=4.75, cover=3 105: leaf=-0.10812, cover=426.75 78: [lsaldo_umedi_var8_ult1<11.58] yes=105, no=106, missing=105, gain=4.75, cover=3 105: leaf=0, lover8_2 105: leaf=0.1, cover=2 105: leaf=0.1, cover=2 105: leaf=0.01, cover=2 105: leaf=0.01, cover=2 105: leaf=0.01, cover=2 105:	1	booster[0]:
3 × 1: [var15-228] yes=3, no=4, missing=3, gain=301.257, cover=4276.25 4 × 3: [saldo_var8<2.085] yes=7, no=8, missing=7, gain=3.61861, cover=2238	2 🔻	
<pre>3:[saldo_var8-2.085] yes=7, no=8, missing=7, gain=3.61861, cover=2239 7:[var38<13396] yes=7, no=16, missing=15, gain=3.6216, cover=2238 15:leaf=-0, cover=1 16:[imp_op_var39_comer_ult3<1613.43] yes=27, no=28, missing=27, gain=2.73829, cover=2237 27:[var15<26] yes=45, no=46, missing=47, gain=3.6504, cover=2235, 75 27:[var15<26] yes=45, no=46, missing=71, gain=0.635187, cover=1888.5 10 7:[noise<0.00132532] yes=9, no=100, missing=99, gain=0.854701, cover=2.25 11 12 13 13 14 12:leaf=-0, cover=1 13 14 12:leaf=-0, cover=1 15 14 15:leaf=-0, cover=1 16:[into_ise<0.00132532] yes=93, no=700, missing=99, gain=0.854701, cover=2.25 19 100:leaf=-0, cover=1 11 12 13 14 12:leaf=-0, cover=1 15 15 16 17:[var38<47875.7] yes=101, no=120, missing=101, gain=3.33744, cover=346 101:leaf=-0.133333, cover=27.5 102:leaf=-0, 133333, cover=27.5 102:leaf=-0, 133333, cover=12.5 17 18 19 21 22 23 24 24 24 25 29:[var38<473891.9] yes=17, no=48, missing=47, gain=5.9796, cover=2037.25 29:[ind_var30_0x1] yes=47, no=48, missing=47, gain=5.3777, cover=636 29:[ind_var30_0x1] yes=47, no=48, missing=47, gain=55, 9796, cover=584.5 17 18 21 22 23 24 25 25 25 25 26 27 27 27 27 27 27 28 28 29 29:[ind_var30_0x1] yes=47, no=48, missing=47, missing=77, gain=5.86103, cover=572.5 104:leaf=-0.018323, cover=142.75 104:leaf=-0.018323, cover=142.75 104:leaf=-0.01812, cover=426.75 12 13 14 15 15 15 15 15 15 15 15 16 15 16 17 16 17 17 16 16 17 17 16 16 17 17 16 16 17 17 16 16 17 17 16 16 17 17 16 16 17 17 16 16 17 17 16 16 17 17 16 16 17 17 16 16 17 17 16 16 17 17 17 16 16 17 17 17 16 16 17 17 17 16 16 17 17 17 17 17 17 17 17 17 17 17 17 17</pre>		
<pre>5 7 7: [var38+13396] yes=15, no=16, missing=15, gain=3.6216, cover=2238 6 7 7 15: [eaf=-0, cover=1 15: [laf_mp_op_var39_comer_ult3<1613, 43] yes=27, no=28, missing=27, gain=2.73829, cover=2237 7 7: [var15<26] yes=45, no=46, missing=45, gain=1, 65054, cover=2235, 75 9 7 27: [var15<26] yes=45, no=46, missing=45, gain=1, 65054, cover=2235, 75 9 7 45: [noise<0.0013053232] yes=71, no=72, missing=99, gain=0. 635187, cover=2.25 10 11 12 13 14 12 15: [aaf=-0, 101111, cover=1.25 12 14 15: [aaf=-0, 123229, cover=1886.25 15: [aaf=-0, cover=1 15: [aaf=-0, 123222, cover=1866.25 16: [aaf=-0, 123222, cover=1866.25 17: [var38<47875.7] yes=101, no=102, missing=99, gain=0. 654701, cover=346 10: [aaf=-0, 133333, cover=27.5 18 18 19 28: [aaf=-0, cover=1 28: [aaf=-0, cover=1 28: [aaf=-0, cover=1.25 29 29: [var38<73981.9] yes=17, no=18, missing=17, gain=17.467, cover=1539.25 21: [saldo_medio_var5_hace3<0.435] yes=29, no=30, missing=29, gain=15.3777, cover=636 29: [ind_var30_0<fli>cover=1 21: [saldo_medio_var5_hace3<0.435] yes=79, no=70, missing=75, gain=0.141027, cover=12 26 27: [ind_var30_0<fli>cover=1 28: [aaf=-0, 13333, cover=11 27: [saldo_medio_var5_hace3<0.435] yes=77, no=78, missing=77, gain=5.86103, cover=572.5 29: [var38<73981.9] yes=103, no=104, missing=103, gain=4.56021, cover=569.5 29: [var38<73981.9] yes=103, no=104, missing=103, gain=4.56021, cover=569.5 29: [var38<733] yes=103, no=104, missing=103, gain=4.56021, cover=572.5 29: [var38<29, over=10 29: [var38<20; over=11 20: [saldo_medio_var5_var40_cover=142.75 20: [saldo_medio_var5_var40_cover=142.75 20: [var38</fli></fli></pre>	4 ▼	
6 15: leaf=-0, cover=1 16: [imp_op_var39_comer_ult3 16: [imp_op_var39_comer_ult3 7* 27: [var15<26] yes=45, no=46, missing=45, gain=1. 65064, cover=2235. 75	5 🔻	
7 16: [imp_op_var39_comer_ult3<1613.43] yes=27, no=28, missing=27, gain=2. 73829, cover=2237	6	
8 27: [var15<26] yes=45, no=46, missing=45, gain=1.65054, cover=2235.75	7 💌	
<pre>9 45: [noise=0.00132532] yes=71, no=72, missing=71, gain=0.635187, cover=1888.5 10 45: [noise=0.0010589] yes=9, no=100, missing=99, gain=0.854701, cover=2.25 99: leaf=-0.111111, cover=1.25 100: leaf=-0.cover=1 13 72: leaf=-0.192529, cover=1886.25 14 4 46: [ID<151300] yes=73, no=74, missing=73, gain=5.9299, cover=347.25 13 73: [var38-47875.7] yes=101, no=102, missing=101, gain=3.33744, cover=346 16 101: leaf=-0.133333, cover=27.5 102: leaf=-0.0222222, cover=1.25 18 28: leaf=-0, cover=1 19 28: leaf=-0, cover=1.25 20 8: leaf=-0, cover=1 21 4: [var38<117321] yes=9, no=10, missing=9, gain=55.9796, cover=2037.25 9: [var38<73981.9] yes=17, no=18, missing=77, gain=15.3777, cover=636 29: lind_var30_0<1] yes=47, no=48, missing=47, gain=5.02247, cover=584.5 47: [imp_op_var40_comer_ult1<695.445] yes=75, no=76, missing=75, gain=0.141027, cover=12 75: leaf=-0.183333, cover=11 76: leaf=-0.085, cover=1 48: [imp_op_var41_efect_ult3<405] yes=77, no=78, missing=77, gain=5.86103, cover=572.5 103: leaf=-0.12487, cover=142.75 104: leaf=-0.12487, cover=426.75 78: [saldo_medio_var8_ult1<11.58] yes=105, no=106, missing=105, gain=4.75, cover=3 33 105: leaf=0.1, cover=2 </pre>	8 🔻	
<pre>10</pre>	9 🔻	
11 99:leaf=-0.111111, cover=1.25 12 100:leaf=-0.cover=1 13 72:leaf=-0.cover=1 14 40:[ID 15 40:[ID 16 101:leaf=-0.13333, cover=27.5 17 102:leaf=-0.022222, cover=318.5 18 74:leaf=-0.022222, cover=1.25 19 28:leaf=-0.022222, cover=1.25 10 102:leaf=-0.13333], cover=2037.25 18 74:leaf=0.022222, cover=1.25 19 28:leaf=-0.022222, cover=1.25 10 102:leaf=-0.13333], cover=2037.25 11 11.111, cover=0.30, missing=9, gain=55.9796, cover=2037.25 11 11.11, cover=0.30, missing=17, gain=17.467, cover=1539.25 11 11.11, cover=0.44, missing=17, gain=17.467, cover=1539.25 12 12.11, var30, ocil, wes=47, no=48, missing=29, gain=5.02247, cover=636 29.11, ind_var30, ocil, wes=47, no=48, missing=27, gain=5.02247, cover=636 29.11, ind_var30, ocil, wes=47, no=48, missing=17, gain=5.02247, cover=584.5 21 75:leaf=-0.183333, cover=11 76:leaf=-0.05, cover=1 77:lvar15 103:leaf=-0.12487, cover=142.75 103:leaf=-0.12487, cover=142.75 103:leaf=-0.12487, cover=426.75	10 🔻	
12 13 14 V 14 V 14 V 15 V 14 V 16 16 16 10: leaf=-0.192529, cover=1886.25 17; lvar38-47875.7] yes=101, no=102, missing=101, gain=3.33744, cover=346 10: leaf=-0.133333, cover=27.5 102: leaf=-0.133333, cover=27.5 102: leaf=-0.0222222, cover=1.25 19 28: leaf=-0, cover=1 21 V 4: [var38<117321] yes=9, no=10, missing=9, gain=55.9796, cover=2037.25 9: [var38<73981.9] yes=17, no=18, missing=17, gain=17.467, cover=1539.25 17: [saldo_medio_var5_hace3<0.435] yes=29, no=30, missing=29, gain=15.3777, cover=636 29: [ind_var30_0<1] yes=47, no=48, missing=47, gain=5.02247, cover=584.5 47: [imp_op_var40_comer_ult1<695.445] yes=75, no=76, missing=75, gain=0.141027, cover=12 75: leaf=-0.183333, cover=11 76: leaf=-0.05, cover=1 48: [imp_op_var41_efect_ult3<405] yes=77, no=78, missing=77, gain=5.86103, cover=572.5 77: [var15<3] yes=103, no=104, missing=103, gain=4.56021, cover=569.5 103: leaf=-0.10487, cover=142.75 104: leaf=-0.101812, cover=246.75 38 38 38 38 38 38 39 30 30 30 30 30 30 30 30 30 30		
13 72: leaf=-0.192529, cover=1886.25 14 v 46: [II0<151300] yes=73, no=74, missing=7.3, gain=5.9299, cover=347.25		
<pre>14 \ 46: [ID<151300] yes=73, no=74, missing=73, gain=5.9299, cover=347.25 15 \ 73: [var38<47875.7] yes=101, no=102, missing=101, gain=3.33744, cover=346 101: leaf=-0.133333, cover=27.5 102: leaf=-0.183412, cover=318.5 74: leaf=-0.0222222, cover=1.25 28: leaf=-0, cover=1 21 \ 4: [var38<117321] yes=9, no=10, missing=9, gain=55.9796, cover=2037.25 29 \ 9: [var38<73981.9] yes=17, no=18, missing=17, gain=17.467, cover=1539.25 23 \ 9: [var38<73981.9] yes=17, no=18, missing=47, gain=5.02247, cover=584.5 24 \ 29: [ind_var30_0<1] yes=47, no=48, missing=47, gain=5.02247, cover=584.5 25 \ 29: [ind_var30_0<1] yes=47, no=48, missing=47, gain=5.02247, cover=584.5 25 \ 75: leaf=-0.183333, cover=11 27 26 \ 48: [imp_op_var40_comer_ult1<695.445] yes=75, no=76, missing=75, gain=0.141027, cover=12 28 \ 48: [imp_op_var41_efect_ult3<405] yes=77, no=78, missing=77, gain=5.86103, cover=572.5 29 \ 103: leaf=-0.12487, cover=142.75 104: leaf=-0.101812, cover=246.75 38 \ 105: leaf=-0.1000, ar8_ult1<11.58] yes=105, no=106, missing=105, gain=4.75, cover=3 38 \ </pre>		
15 ▼ 73: [var38<47875.7] yes=101, no=102, missing=101, gain=3.33744, cover=346		
16 101: leaf=-0.133333, cover=27.5 17 102: leaf=-0.183412, cover=318.5 18 74: leaf=0.022222, cover=1.25 19 28: leaf=-0, 0222222, cover=1.25 20 8: leaf=-0, 0222222, cover=1.25 21 4: [var38 101: leaf=-0.133333, cover=2037.25 22 9: [var38 9: [var38 17: [saldo_medio_var5_hace3 17: [saldo_medio_var5_hace3 19: [var30_0<1] yes=47, no=48, missing=47, gain=5.02247, cover=584.5		
17 102:leaf=-0.183412, cover=318.5 18 74:leaf=0.0222222, cover=1.25 19 28:leaf=-0, cover=1 20 8:leaf=-0, cover=1 21 4:[var38<117321] yes=9, no=10, missing=9, gain=55.9796, cover=2037.25		
18 74: leaf=0.0222222, cover=1.25 19 28: leaf=-0, cover=1.25 20 8: leaf=-0, cover=1 21 4: [var38<117321] yes=9, no=10, missing=9, gain=55.9796, cover=2037.25		
19 28:leaf=-0.0222222, cover=1.25 20 8:leaf=-0, cover=1 21▼ 4:[var38<117321] yes=9, no=10, missing=9, gain=55.9796, cover=2037.25		
20 8:leaf=-0,cover=1 21 ▼ 4:[var38<117321] yes=9,no=10,missing=9,gain=55.9796,cover=2037.25 22 ▼ 9:[var38<73981.9] yes=17,no=18,missing=17,gain=17.467,cover=1539.25 23 ▼ 17:[saldo_medio_var5_hace3<0.435] yes=29,no=30,missing=29,gain=15.3777,cover=636 29:[ind_var30_0<1] yes=47,no=48,missing=47,gain=5.02247,cover=584.5 29 ▼ 17:[imp_op_var40_comer_ult1<695.445] yes=75,no=76,missing=75,gain=0.141027,cover=12 75:leaf=-0.183333,cover=11 76:leaf=-0.05,cover=1 48:[imp_op_var41_efect_ult3<405] yes=77,no=78,missing=77,gain=5.86103,cover=572.5 77:[var15<3] yes=103,no=104,missing=103,gain=4.56021,cover=569.5 103:leaf=-0.12487,cover=142.75 104:leaf=-0.101812,cover=426.75 78:[saldo_medio_var8_ult1<11.58] yes=105,no=106,missing=105,gain=4.75,cover=3 105:leaf=0.1,cover=2		
<pre>21 ▼ 4: [var38<117321] yes=9, no=10, missing=9, gain=55.9796, cover=2037.25 9: [var38<73981.9] yes=17, no=18, missing=17, gain=17.467, cover=1539.25 17: [saldo_medio_var5_hace3<0.435] yes=29, no=30, missing=29, gain=15.3777, cover=636 24 ▼ 29: [ind_var30_0<1] yes=47, no=48, missing=47, gain=5.02247, cover=584.5 25 ▼ 26 26 27 28 ▼ 29 28 ▼ 29 29 ▼ 29 30 31 32 ▼ 33 42: [var38<117321] yes=9, no=10, missing=105, gain=4.75, cover=3 105: leaf=0.1, cover=2 43: [var30_0</pre>		
<pre>9: [var38<73981.9] yes=17, no=18, missing=17, gain=17.467, cover=1539.25 17: [saldo_medio_var5_hace3<0.435] yes=29, no=30, missing=29, gain=15.3777, cover=636 24 ▼ 29: [ind_var30_0<1] yes=47, no=48, missing=47, gain=5.02247, cover=584.5 47: [imp_op_var40_comer_ult1<695.445] yes=75, no=76, missing=75, gain=0.141027, cover=12 75: leaf=-0.183333, cover=11 76: leaf=-0.05, cover=1 48: [imp_op_var41_efect_ult3<405] yes=77, no=78, missing=77, gain=5.86103, cover=572.5 77: [var15<33] yes=103, no=104, missing=103, gain=4.56021, cover=569.5 103: leaf=-0.12487, cover=142.75 104: leaf=-0.101812, cover=426.75 78: [saldo_medio_var8_ult1<11.58] yes=105, no=106, missing=105, gain=4.75, cover=3 105: leaf=0.1, cover=2</pre>		
23 ▼ 17: [saldo_medio_var5_hace3<0.435] yes=29, no=30, missing=29, gain=15.3777, cover=636	22 🔻	
24 ▼ 29: [ind_var30_0<1] yes=47, no=48, missing=47, gain=5.02247, cover=584.5 47: [imp_op_var40_comer_ult1<695.445] yes=75, no=76, missing=75, gain=0.141027, cover=12 75: leaf=-0.183333, cover=11 76: leaf=-0.05, cover=1 48: [imp_op_var41_efect_ult3<405] yes=77, no=78, missing=77, gain=5.86103, cover=572.5 77: [var15<33] yes=103, no=104, missing=103, gain=4.56021, cover=569.5 103: leaf=-0.12487, cover=142.75 104: leaf=-0.101812, cover=426.75 78: [saldo_medio_var8_ult1<11.58] yes=105, no=106, missing=105, gain=4.75, cover=3 105: leaf=0.1, cover=2		
25 ▼ 47:[imp_op_var40_comer_ult1<695.445] yes=75, no=76, missing=75, gain=0.141027, cover=12 26 27 28 ▼ 48:[imp_op_var41_efect_ult3<405] yes=77, no=78, missing=77, gain=5.86103, cover=572.5 29 ▼ 48:[imp_op_var41_efect_ult3<405] yes=77, no=78, missing=77, gain=5.86103, cover=572.5 77:[var15<33] yes=103, no=104, missing=103, gain=4.56021, cover=569.5 103:leaf=-0.12487, cover=142.75 104:leaf=-0.101812, cover=426.75 32 ▼ 78:[saldo_medio_var8_ult1<11.58] yes=105, no=106, missing=105, gain=4.75, cover=3 105:leaf=0.1, cover=2		
26 27: leaf=-0.183333,cover=11 28 ▼ 48: [imp_op_var41_efect_ult3<405] yes=77,no=78,missing=77,gain=5.86103,cover=572.5 29 ▼ 77: [var15<33] yes=103,no=104,missing=103,gain=4.56021,cover=569.5 30 103: leaf=-0.12487,cover=142.75 31 104: leaf=-0.101812,cover=426.75 32 ▼ 78: [saldo_medio_var8_ult1<11.58] yes=105,no=106,missing=105,gain=4.75,cover=3 33 105: leaf=0.1,cover=2		
<pre>27 28 ▼ 76:leaf=-0.05,cover=1 28 ▼ 48:[imp_op_var41_efect_ult3<405] yes=77,no=78,missing=77,gain=5.86103,cover=572.5 29 ▼ 77:[var15<33] yes=103,no=104,missing=103,gain=4.56021,cover=569.5 30 30 31 32 ▼ 78:[saldo_medio_var8_ult1<11.58] yes=105,no=106,missing=105,gain=4.75,cover=3 33 33 33 33 34 35 35 35 36 36 36 36 36 36 36 36 36 36 36 36 36</pre>		
28 ▼ 48:[imp_op_var41_efect_ult3<405] yes=77,no=78,missing=77,gain=5.86103,cover=572.5 29 ▼ 77:[var15<33] yes=103,no=104,missing=103,gain=4.56021,cover=569.5 30 103:leaf=-0.12487,cover=142.75 31 104:leaf=-0.101812,cover=426.75 32 ▼ 78:[saldo_medio_var8_ult1<11.58] yes=105,no=106,missing=105,gain=4.75,cover=3 105:leaf=0.1,cover=2		
29 ▼ 77:[var15<33] yes=103,no=104,missing=103,gain=4.56021,cover=569.5 30 103:leaf=-0.12487,cover=142.75 31 104:leaf=-0.101812,cover=426.75 32 ▼ 78:[saldo_medio_var8_ult1<11.58] yes=105,no=106,missing=105,gain=4.75,cover=3 33 105:leaf=0.1,cover=2		
30 103:leaf=-0.12487,cover=142.75 31 104:leaf=-0.101812,cover=426.75 32 ▼ 78:[saldo_medio_var8_ult1<11.58] yes=105,no=106,missing=105,gain=4.75,cover=3	29 🔻	
31 104:leaf=-0.101812,cover=426.75 32 ▼ 78:[saldo_medio_var8_ult1<11.58] yes=105,no=106,missing=105,gain=4.75,cover=3		
32 ▼ 78:[saldo_medio_var8_ult1<11.58] yes=105,no=106,missing=105,gain=4.75,cover=3 33 105:leaf=0.1,cover=2		
33 105:leaf=0.1,cover=2	32 🔻	
	33	
	34	106:leaf=-0.1,cover=1
30:[num_op_var41_hace2<43] yes=49,no=50,missing=49,gain=5.43074,cover=51.5	35 💌	
36 ▼ 49:[noise<0.0232753] yes=79,no=80,missing=79,gain=1.47009,cover=50.25		
37 79: leaf=-0.04, cover=1.5	37	
38 ▼ 80:[imp_op_var41_efect_ult3<2850] yes=107, no=108, missing=107, gain=0.0875811, cover=48.75		
39 107: leaf=-0.177436, cover=47.75		
40 108:leaf=-0.05,cover=1		
41 50:leaf=0.0222222,cover=1.25		



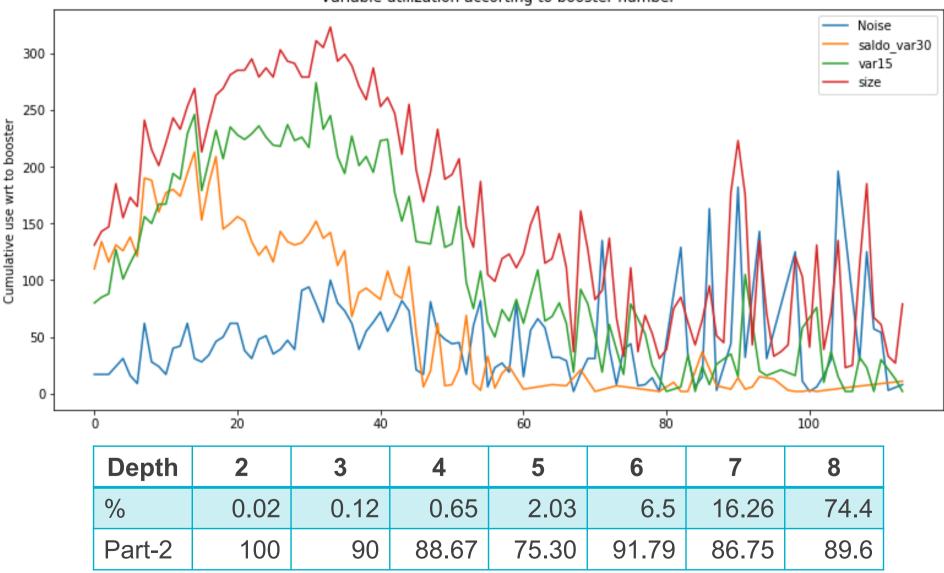
EVOLUTION OF VARIABLES REGARDING ITERATIONS

Cumulative gain is stable through iterations



NOISE ADDITION IN GRADIENT BOOSTING PREDICTION

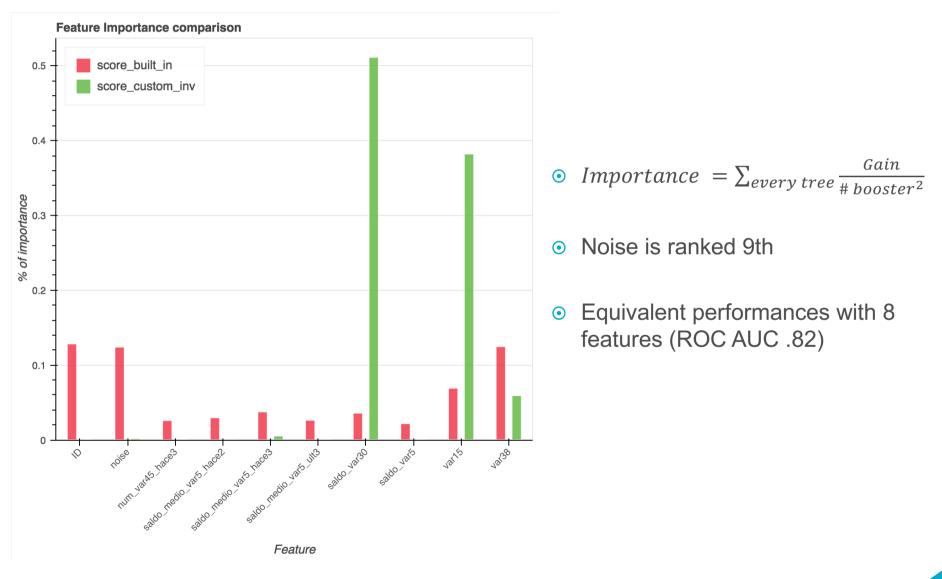
Noise is used only deeply in the trees with small discrimination power



Variable utilization accorting to booster number

A ROBUST METRIC NOISE-PROOF

It cancels noise, it must be good !



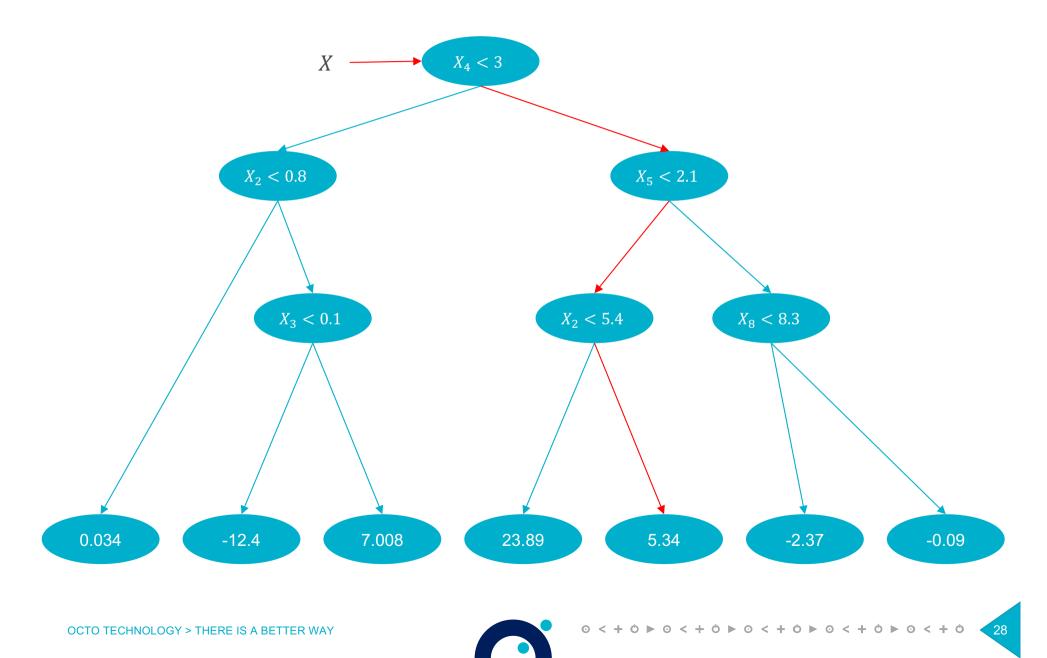


03 > POINT FEATURE IMPORTANCE ?

○ ► ○ < + ○ ► ○ < + ○ ► ○ < + ○ ► ○ < + ○

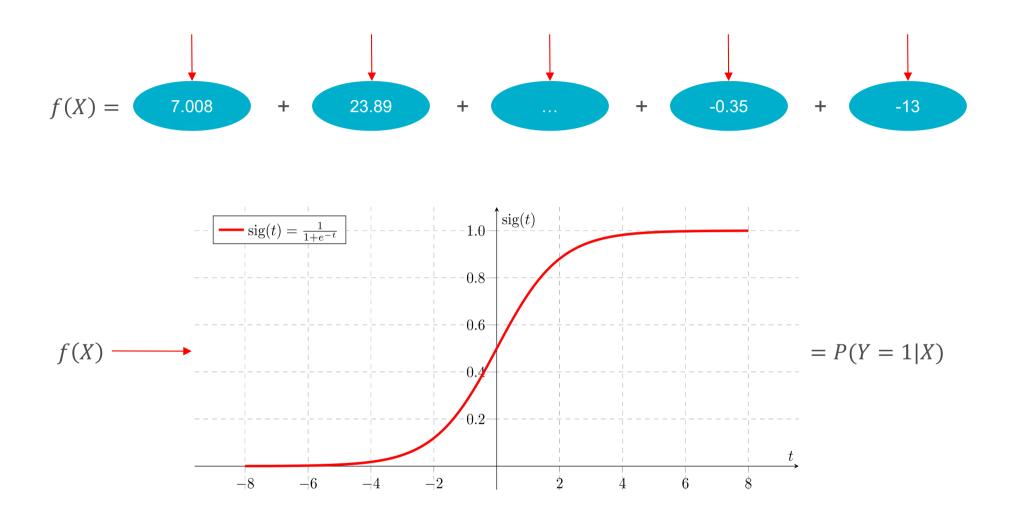
IS IT REALLY IMPORTANT FOR EVERYONE ?

A data point does not cross the entireness of the Forest



GRADIENT BOOSTING PREDICTION IS A SUM

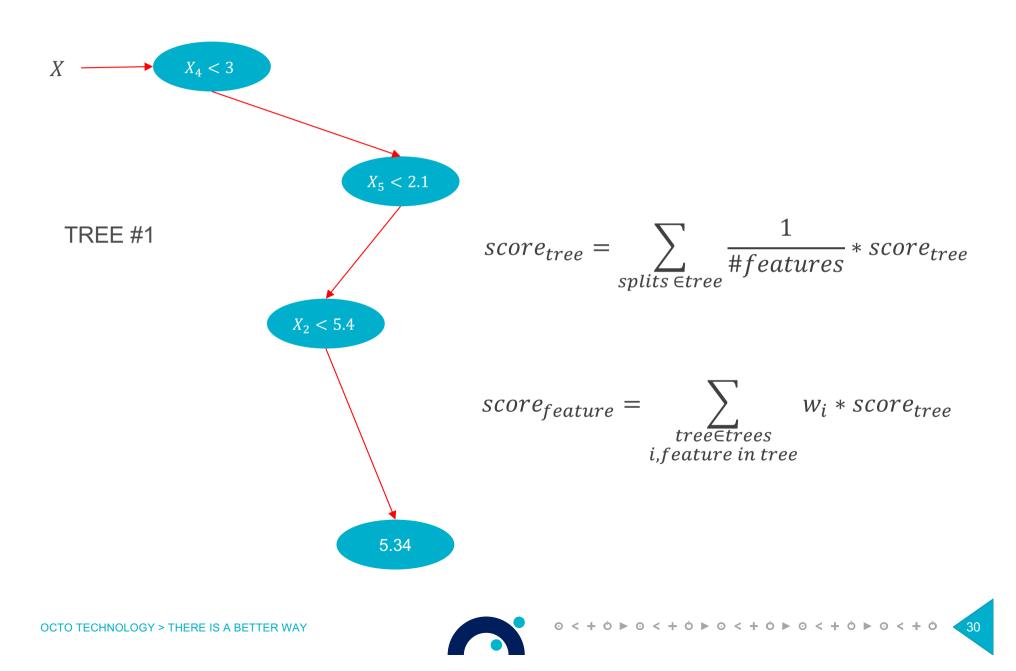
Aggregation of the scores for each leaf





.PREDICT : FOLLOW THE PATHS

The score of a leaf is obtained by going through the « gates » of the splits



AGGREGATION OF SEVERAL POINTS

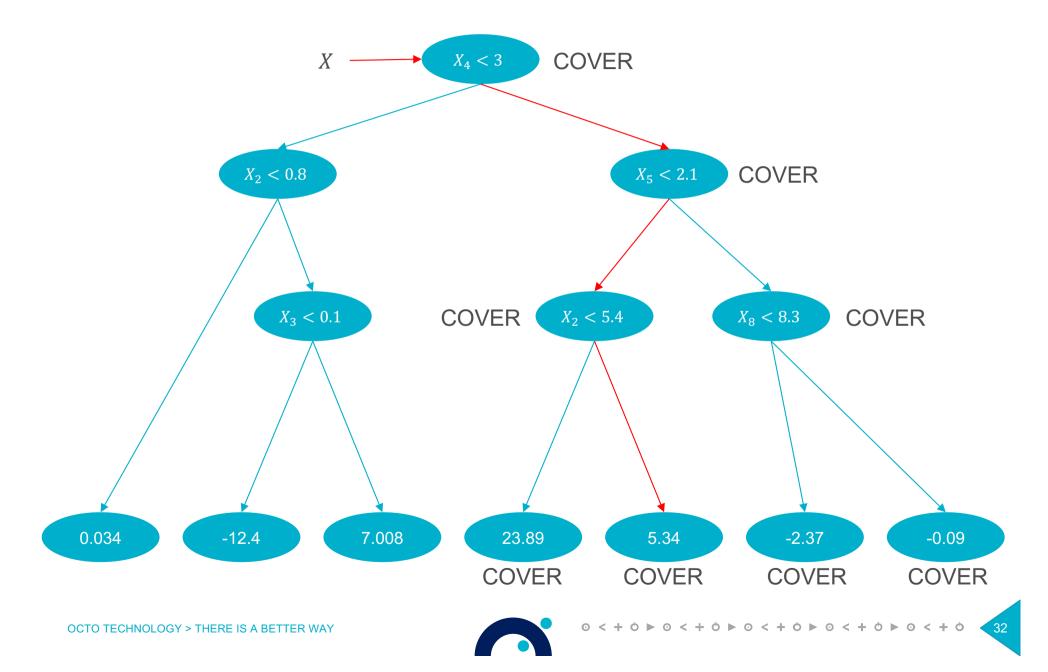
Residual interpretations

- Test
 - > **10 most False Positives** : not a high probability of being zero
 - > Top 5 contributers toward a positive value (misclassification)
- Results
 - > **Two variables come for 4 of the 10 data points**: saldo_medio_var5_hace3, var3
 - > ID is in the top : generation of noise
 - > Lead for investigation **BUT** does not tell what threshold is faulty



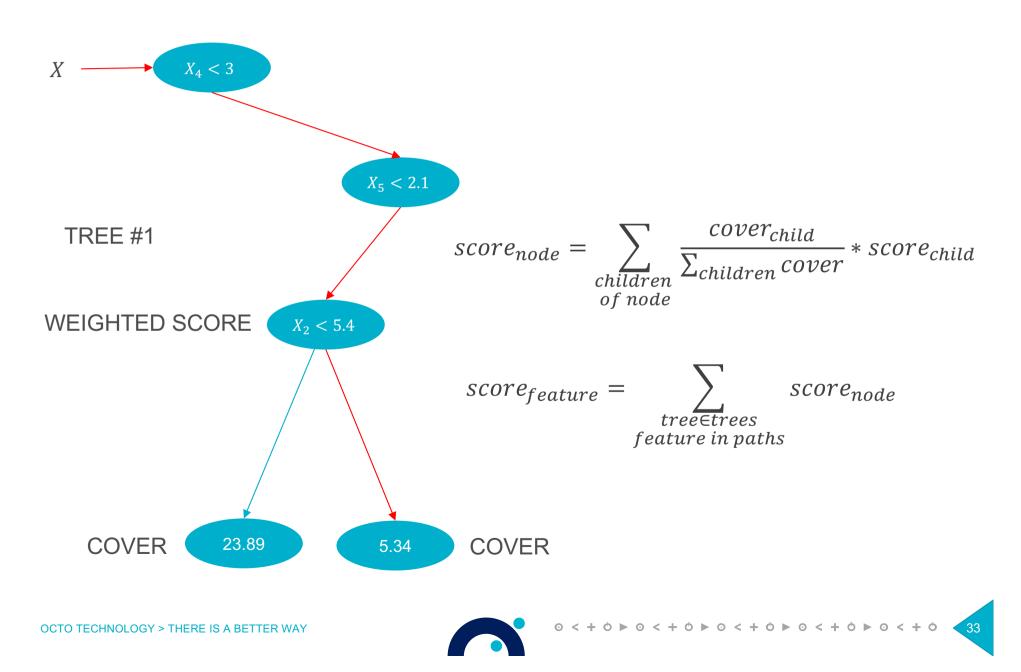
HOW MUCH A NODE IS RESPONSIBLE FOR THE FINAL SCORE ?

Using the data point distribution in a leaf, we can back propagate the score



.PREDICT : STILL THE SAME THE PATHS

Same same, but different, but still same



AGGREGATION OF SEVERAL POINTS

Residual interpretations

- Test
 - > **10 most False Positives** : not a high probability of being zero
 - > Top 5 contributers toward a positive value (misclassification)
- Results
 - > Not the same variables pointed out
 - > Killer Features are main contributers
 - > Maybe look at the hypercube generated in terms of separation in space



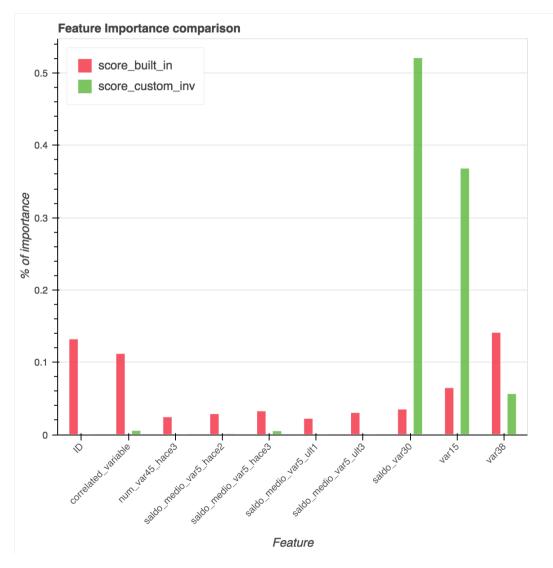


Ó ► O < + Ó ► O < + Ó ► O < + Ó ► O < + Ó</p>

CORRELATED VARIABLES

Machine Learning is just correlation, not causality

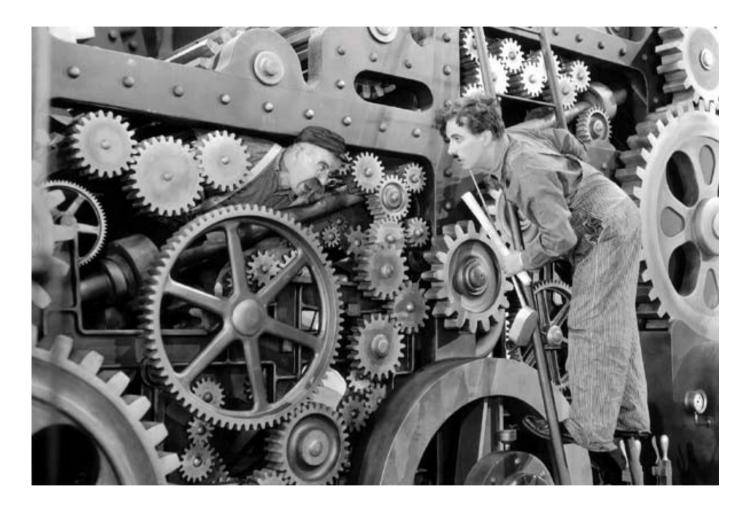
• Without any feature engineering, add highly correlated variable to var15 (95 %)





INTERPRETATION IS NOT A MIRACLE

Prediction is achievable, but not prescription



« There will always be a compromise between predictive power and interpretabiliy »

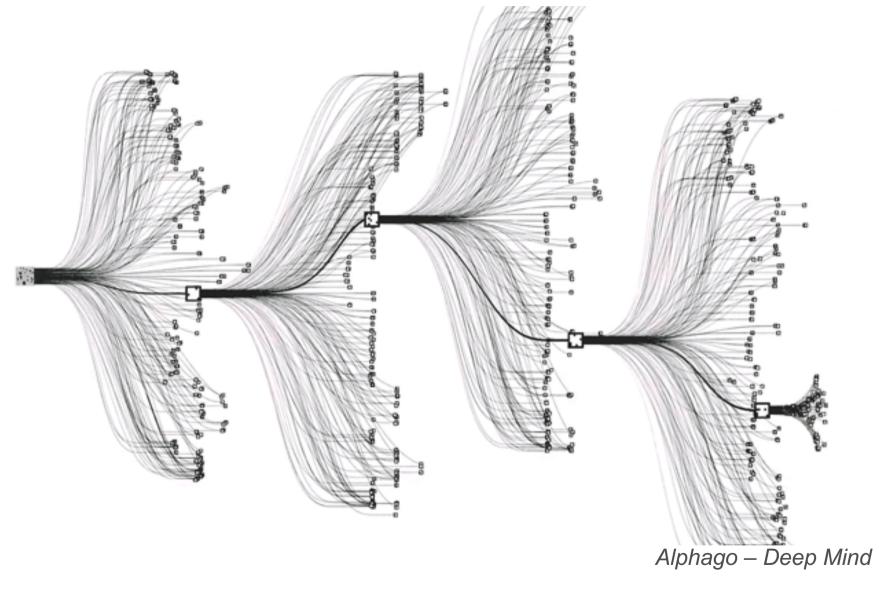
OCTO TECHNOLOGY > THERE IS A BETTER WAY





INTERPRETATION IS MORE IMPORTANT THAN PERFORMANCE

A 90% accuracy interpretable model is better than 99% accuracy black box





MULTI CLASS / REGRESSION / FINE TUNING

Few leads for the next step

- Multi-class classificiation generates a huge amount of trees with gradient boosting
- Regression split gains are different
- Regression scores are inherently different
- What is the influence of the cost function (L1 L2)?
- Taking the threshold into account could lead to hypercube definitions for data segmentation
- Everything is on my github : Cnstant/feature_importance_gbm



THANKS FOR YOUR ATTENTION

• Questions ?





 $\dot{\mathbf{O}} \blacktriangleright \mathbf{O} < + \dot{\mathbf{O}} \vdash \mathbf{O} < + \dot{\mathbf{O}} \vdash \mathbf{O} < + \dot{\mathbf{O}} \vdash \mathbf{O} < + \dot{\mathbf{O}}$

RECURSIVITY FOR PATH ENUMERATION

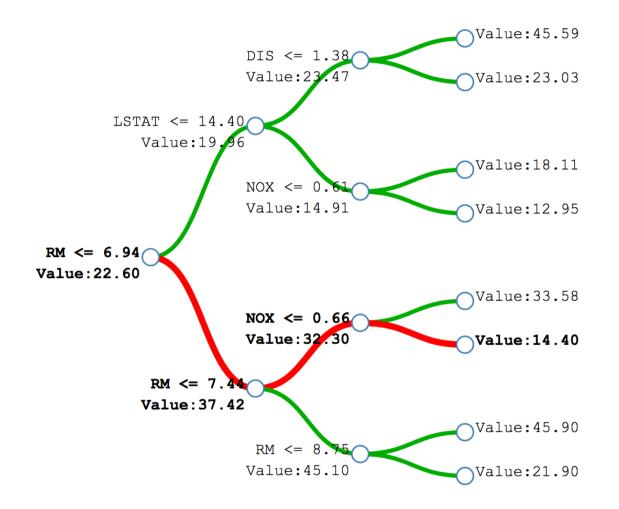
Implementation is to be open sourced

```
def get paths(node, path=None):
    #import pdb; pdb.set trace()
   path = path or []
    # Copy of the actual path if there is a right child
    if node.right:
        right path = copy(path)
    # Left child inherits the current path
   path.append(node)
   paths = get paths(result[node.left], path) if node.left else [path]
    # Right child is second and extend the result
    if node.right:
        right path.append(node)
        right paths = get paths(result[node.right], right path)
        paths.extend(right paths)
    # Pass the results back
    return paths
```

○ < + ○ ▶ ○ < + ○ ▶ ○ < + ○ ▶ ○ < + ○ ▶ ○ < + ○ ▶ ○ < + ○</p>

TREE INTERPRETER FOR RANDOM FOREST

Random Forest are only horizontal



Prediction: $14.40 \approx 22.60$ (trainset mean) + 14.82(gain from RM) - 5.12(loss from RM) - 17.9(loss from NOX)



TREE INTERPRETER FOR RANDOM FOREST

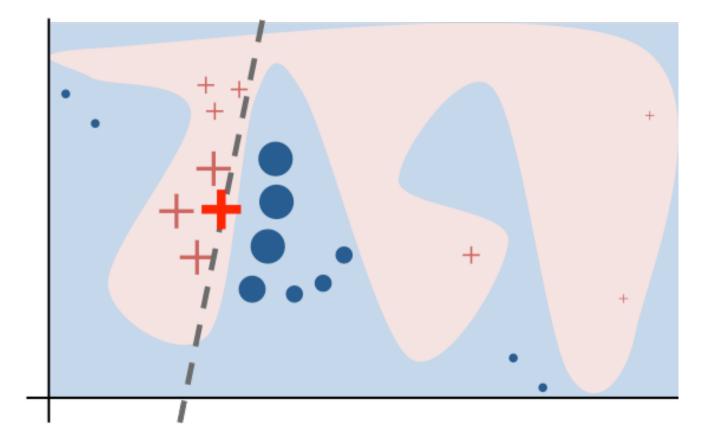
Can we turn prediction into prescription ?

ompliar	nce Simulation						
1. Choose 2. Examine	value from your scores in three steps a line ID e actionable variables best simulation to minimize/maximiz						
Select	your line ID Get Causes				2 Se v38,v34	elect actionable	Variables Get Best Simulation !
Informa	tions	💠 Insights			🔺 Nex	xt best action	
10 ¢ e	ntries Search:	Show 10 \$ entries	Search:		Show 10	+ entries Search	κ
now 10 ¢ e Name	ntries Search:	Show 10 + entries Variable name	Search:	%	Show 10	current	Best
Name				%	Show 10 v38		
Name core_id	Value	Variable name				Current	Best
Name core_id ource	Value 1.0	Variable name v38		29%	v38	Current 4	Best
Name core_id ource	Value 1.0 0.0	Variable name v38 v72		29% 8%	√38 √34	Current 4 8.3874314833	Best 0 4.7480449677
Name core_id ource f rediction	Value 1.0 0.0 1.0	Variable name v38 v72 v12		29% 8% 6%	v38 v34 v14 Score	Current 4 8.3874314833 10.7607964422 0.4212318664	Best 0 4.7480449677 8.7284574509 0.3750294077
Name core_id ource d rediction 18	Value 1.0 0.0 1.0 0.421231886448	Variable name v38 v72 v12 v34		29% 8% 6% 6%	v38 v34 v14	Current 4 8.3874314833 10.7607964422	Bost 0 4.7480449677 8.7284574509
Name core_id ource d rediction 18 19	Value 1.0 0.0 1.0 0.0 2.02868433538	Variable name v38 v72 v12 v34 v14		29% 8% 6% 6% 4%	v38 v34 v14 Score	Current 4 8.3874314833 10.7607964422 0.4212318664	Best 0 4.7480449677 8.7284574509 0.3750294077
	Value 1.0 0.0 1.0 0.421231866448 2.02688433538 0.23372161772	Variable name v38 v72 v12 v34 v14 v10		29% 8% 6% 6% 4% 4%	v38 v34 v14 Score First	Current 4 8.3874314833 10.7607964422 0.4212318664 Previous	Best 0 4.7480449677 8.7284574509 0.3750294077 Next Last
Name core_id ource d wediction 18 19 12	Value 1.0 0.0 1.0 0.421231866448 2.02688433538 0.23372161772 6.61828949496	Variable name v38 v72 v12 v34 v14 v10 v129		29% 8% 6% 6% 4% 4% 3%	v38 v34 v14 Score First	Current 4 8.3874314833 10.7607964422 0.4212318664 Previous t score 0.421. core achievable	Best 0 4.7480449677 8.7284574509 0.3750294077 Next Last

44

LIME

Don't interpret the model, interpret the stability of the prediction





OCTO TECHNOLOGY > THERE IS A BETTER WAY

LIME

Feature importance is how much you need to change your value to change class

