
Invited Talk @Lenddo
Knowledge Sharing Lunch
29th of October - 12.30pm - New York City

INSTANCE-LEVEL EXPLANATION ALGORITHMS ON BEHAVIORAL AND TEXTUAL DATA: A COUNTERFACTUAL-ORIENTED COMPARISON

Yanou Ramon, David Martens, Foster Provost, Theodoros Evgeniou

29th of October 2019, Invited Talk Lenddo, New York City


PROBLEM
STATEMENT

MOVIE VIEWING DATA (MovieLens)

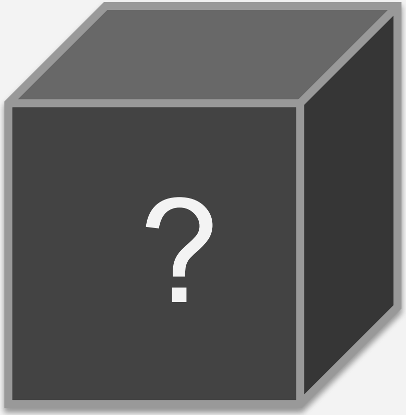
Active feature = "evidence"

6,040 users

	Star wars	Pearl Harbor	Django	...	Home Alone	Target $\hat{y}$ Gender
User 1	1	0	0	…	1	***M***
User 2	1	1	0	…	0	***F***
…	…	…	…	…	…	***…***
User n	1	0	0	…	0	***M***

Sparsity *p* = 95,53%

	Star wars	Pearl Harbor	Django	…	Home Alone	*Target $\hat{y}$ Gender*
User 1	1	0	0	…	1	***M***
User 2	1	1	0	…	0	***F***
…	…	…	…	…	…	…
User n	1	0	0	…	0	***M***



$\hat{y} = 1$ ***if*** *male*
else $\hat{y} = 0$

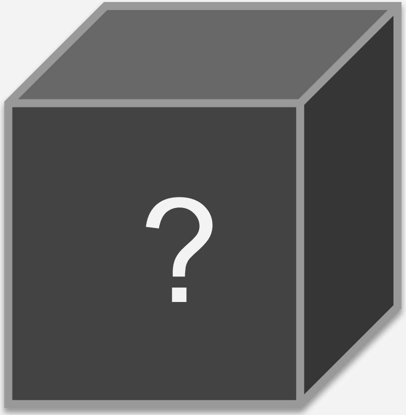
Movie Viewing Data (MovieLens)

Black box model
$\Rightarrow$Thousands of coefficients
$\Rightarrow$Nonlinear techniques

Comprehensibility issues

	Star wars	Pearl Harbor	Django	...	Home Alone	*Target $\hat{y}$ Gender*
User 1	1	0	0	…	1	***M***
User 2	1	1	0	…	0	***F***
…	…	…	…	…	…	***…***
User n	1	0	0	…	0	***M***

Movie Viewing Data (MovieLens)



Black box model

$\Rightarrow$Thousands of coefficients

$\Rightarrow$Nonlinear techniques

$\hat{y} = 1$ ***if*** *male*

else $\hat{y} = 0$

INSTANCE-LEVEL EXPLANATIONS: Why relevant?

WHY EXPLANATIONS?

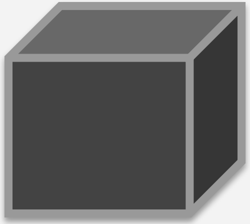
- Improving the model: data leakage, overfitting, misclassifications
- Trust and acceptance
- Detect bias / discrimination
- Formal objectives *vs* ethical objectives
- Compliance (e.g., right to explanations)
- ...

WHY EXPLANATIONS?

- **Improving the model: explain misclassifications**

Example: objectionable web content detection (Martens & Provost, 2013)

Web pages



Black box model

$\hat{y} = 1$ ***if*** *objectionable*

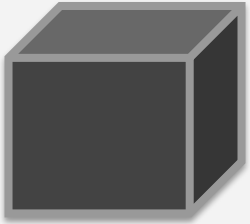
else $\hat{y} = 0$

WHY EXPLANATIONS?

- **Improving the model: explain misclassifications**

Example: objectionable web content detection (Martens & Provost, 2013)

Web page



Black box model

$\hat{y} = 0$

*"Why was this page **NOT** classified as objectionable?"*

WHY EXPLANATIONS?

- **Improving the model: explain misclassifications**

Example: objectionable web content detection (Martens & Provost, 2013)

Misclassified web page: predicted as non-objectionable

IF the word ***"bikini"*** was not on the page, **THEN** the predicted class would change from non-objectionable to objectionable

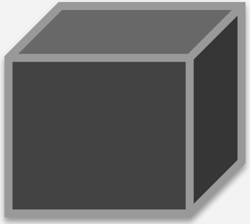
Why?

WHY EXPLANATIONS?

- **Trust and acceptance**

Example: explainable legal document classification (Chhatwal et al., 2019)

Legal documents



Black box model

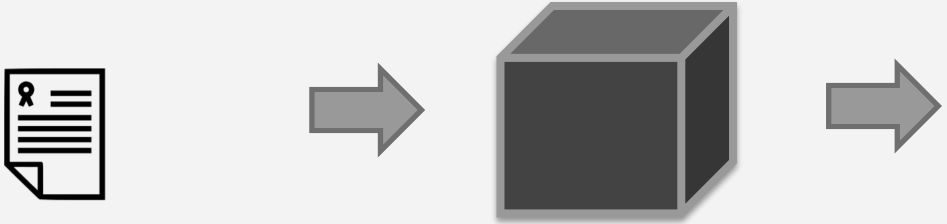
$\hat{y} = 1$ ***if*** *responsive*

else $\hat{y} = 0$

WHY EXPLANATIONS?

- **Trust and acceptance**

Example: explainable legal document classification (Chhatwal et al., 2019)



Legal document

Black box model

$\hat{y} = 1$

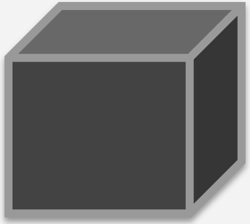
"Why is this document classified as responsive?"

WHY EXPLANATIONS?

- **Generate insights**

Example: Know your customer (e.g., Hall, 2012; Grossnickle, 2001)

Visited URLs



Black box model

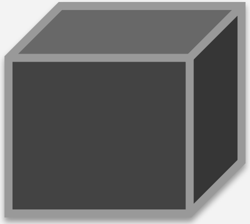
$\hat{y} = 1$ ***if*** *interested in product*
else $\hat{y} = 0$

WHY EXPLANATIONS?

- **Generate insights**

Example: Know your customer (e.g., Hall, 2012; Grossnickle, 2001)

Visited URLs



Black box model

"Who are we targeting?
Why are we targeting them?"

COUNTERFACTUAL EXPLANATIONS

- Instance-level explanation of particular prediction
- Insight into how model works (causality *within* model)
- Rule: a minimal set of features such that the predicted class changes when "removing" them (~setting value to zero)
- Comprehensible and concise
- Argued to be the most intuitive and valuable for humans because they are contrastive *("Why X rather than not-X?"; Miller, 2017)*

COUNTERFACTUAL EXPLANATIONS

Example: gender prediction using movie viewing data

	Star wars	Pearl Harbor	Django	...	Home Alone	*Target $\hat{y}$ Gender*
User 1	1	0	0	...	1	***M***
User 2	1	1	0	...	0	***F***
...	...	...	...	...	...	**...**
User n	1	0	0	...	0	***M***

Sam watched 120 movies
Sam was predicted as 'male'

COUNTERFACTUAL EXPLANATIONS

Example: gender prediction using movie viewing data

	Star wars	Pearl Harbor	Django	...	Home Alone	*Target $\hat{y}$ Gender*
User 1	1	0	0	...	1	***M***
User 2	1	1	0	...	0	***F***
...	...	...	...	...	...	**...**
User n	1	0	0	...	0	***M***

Sam watched 120 movies
Sam was predicted as 'male'

Why?

29th of October 2019, Invited Talk Lenddo, New York City

COUNTERFACTUAL EXPLANATIONS

Example: gender prediction using movie viewing data

Sam watched 120 movies
Sam was predicted as 'male'

IF Sam would not have watched **{Taxi Driver, The Dark Knight, Die Hard, Terminator 2, Now You See Me, Interstellar}**,
THEN the predicted class changes from 'male' to 'female'

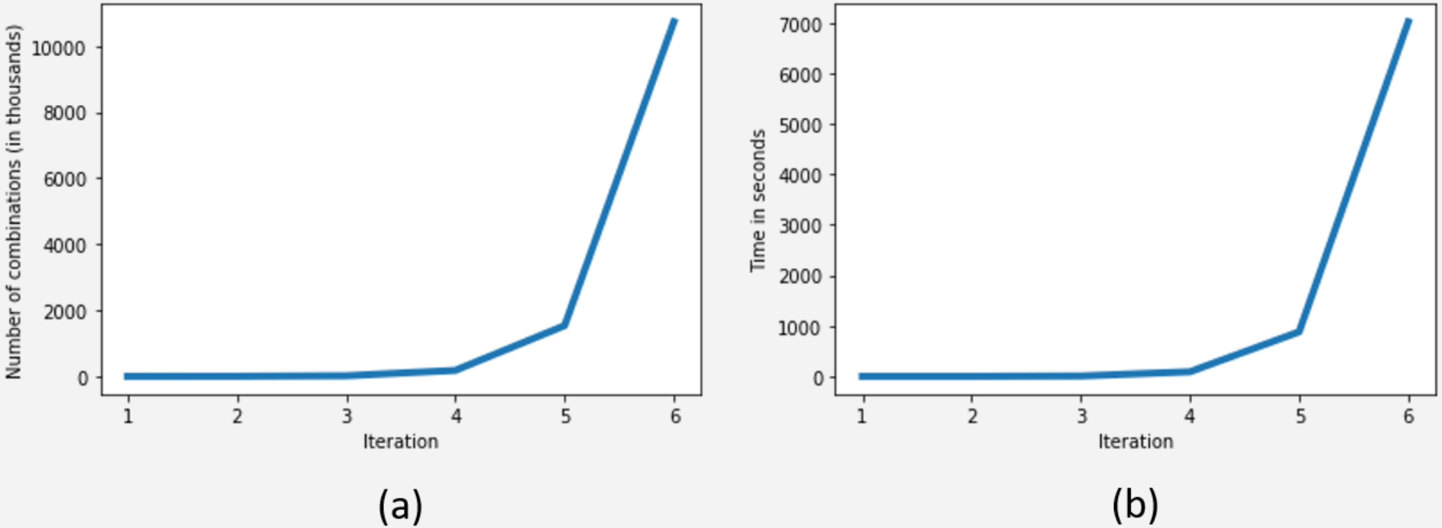
WHY COMPLETE SEARCH FAILS

- Start with removing one feature and increase number of features in the subset until the predicted class changes
- Scales exponentially with active features *m* and required number of features *k* to be removed

  e.g., for an instance with *m* features, a combination of *k* features requires $\frac{m!}{(m-k)!k!}$ evaluations

WHY COMPLETE SEARCH FAILS

Figure 1: Number of combinations (a) and time elapsed (b) per iteration for an instance with 34 active features and a counterfactual of 6 features (*MovieLens data*)




COUNTERFACTUAL
ALGORITHMS

ALGORITHMIC ASSUMPTIONS

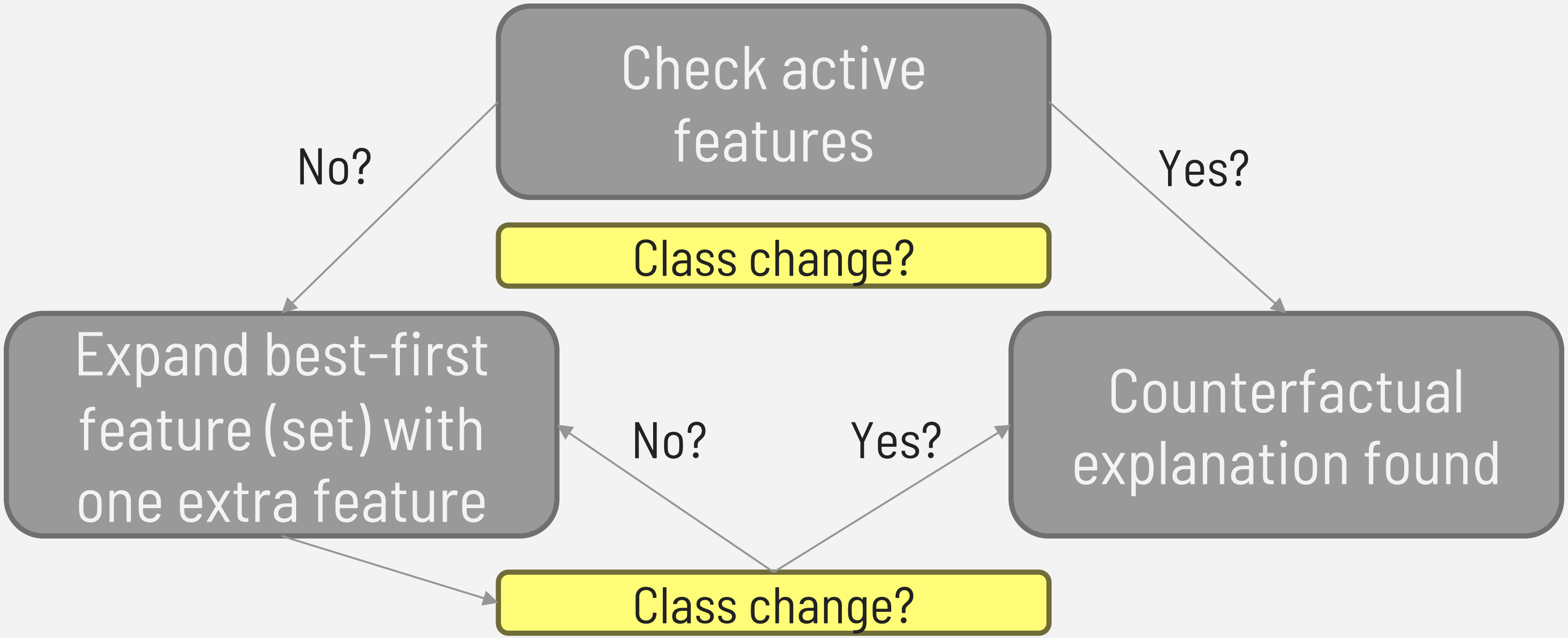
- **Goal**: find counterfactual explanation as fast and as concise as possible (efficiency-effectiveness tradeoff)
- Model-agnostic
- Max. 30 features in explanation
- Max. 5 minutes to compute explanation

BEST-FIRST SEARCH (SEDC)

- Explaining document classifications (Martens & Provost, 2013)
- Model-agnostic algorithm SEDC: heuristic best-first search (*lin-SEDC*: linear implementation)
- Optimal for linear models

Implementation on *https://github.com/yramon/edc*

BEST-FIRST SEARCH (SEDC)



NOVEL HYBRID ALGORITHMS

Additive Feature Attribution methods:

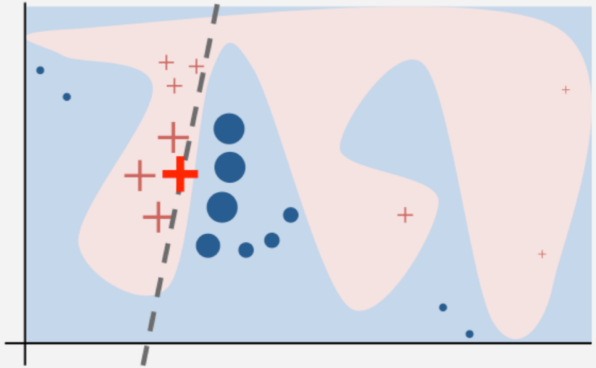
- LIME: Local Model-agnostic Explainer (Ribeiro et al., 2016)
- SHAP: Shapley Additive Explanations (Lundberg et al., 2018)

Output: importance-ranked list

NOVEL HYBRID ALGORITHMS

LIME / SHAP

- Sparse, linear explanation model
- Approximates original model in neighbourhood of instance
- Perturbed instances

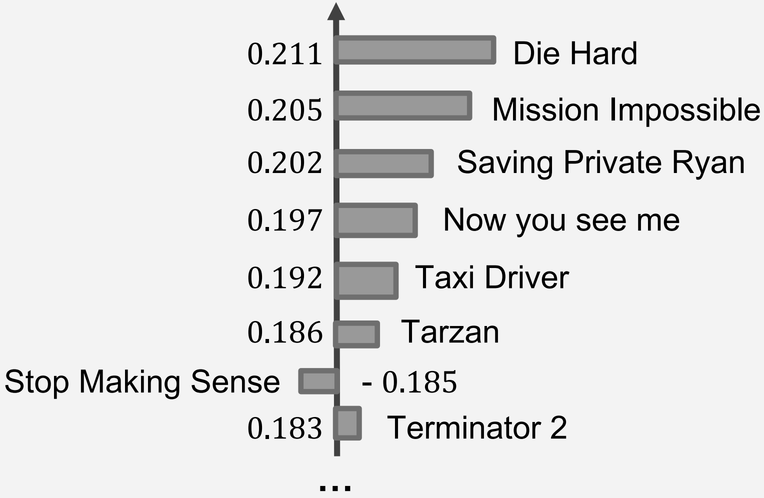


Ribeiro et al., 2016

NOVEL HYBRID ALGORITHMS

LIME / SHAP

Example: gender prediction using movie viewing data



NOVEL HYBRID ALGORITHMS

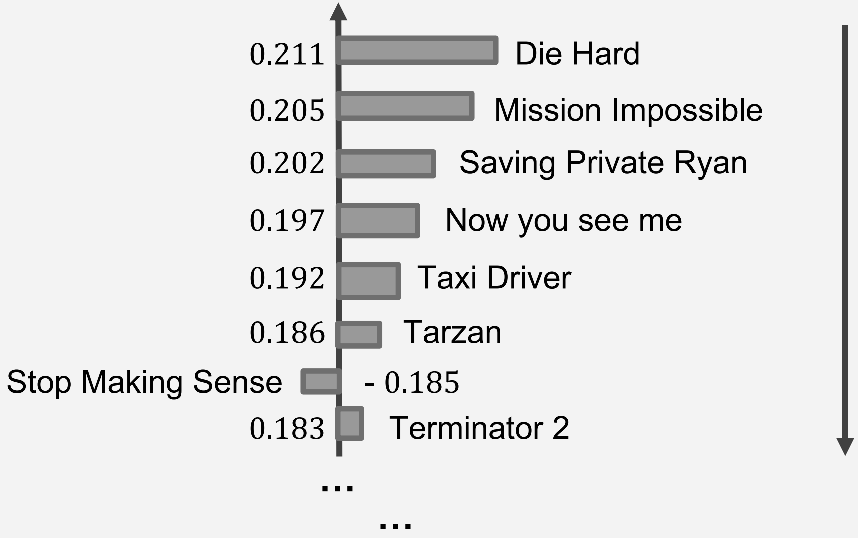
Originality: importance rankings may be "intelligent" starting point for efficiently searching counterfactuals

$\Rightarrow$ Novel algorithms: **LIME–C** and **SHAP-C**

NOVEL HYBRID ALGORITHMS

LIME-C / SHAP-C

Example: gender prediction using movie viewing data



Remove features with positive importance weight until the class changes

CONTRIBUTIONS

- Two novel model-agnostic algorithms (LIME-C / SHAP-C)
- Define quantitative evaluation criteria
- Evaluate performance against existing SEDC algorithm and make practical recommendations


EXPERIMENTAL SETUP

Collect data sets and build models

Textual data: linear/rbf SVM

Behavioral data: LR/MLP

Table 1: Data sets and characteristics

Dataset	Type	Target	Instances	Features	b	p	Test set (%)	$\dot{m}_{lin}$	$\dot{m}_{nonlin}$	ref
Flickr*	B	comments	100,000	190,991	36.91%	99.99%	20,000 (20%)	2.02	2.96	[38]
Ecommerce*	B	gender	15,000	21,880	21.98%	99.99%	3,000 (15%)	2.60	2.67	[3]
Airline*	T	sentiment	14,640	5,183	16.14%	99.82%	2,928 (15%)	7.81	8.21	[2]
Twitter	T	topic	6,090	4,569	9.15%	99.74%	1,218 (10%)	9.52	9.35	[5]
Fraud*	B	fraudulent	858,131	107,345	6.4e-5%	99.99%	171,627(1%)	11.83	14.09	n.a.
YahooMovies*	B	gender	7,642	11,915	28.87%	99.76%	1,529 (20%)	25.24	25.00	[6]
TaFeng*	B	age	31,640	23,719	45.23%	99.90%	6,328 (15%)	44.32	37.24	[23]
KDD2015*	B	dropout	120,542	4,835	20.71%	99.67%	24,109 (20%)	49.01	46.40	[4]
20news	T	atheism	18,846	41,356	4.24%	99.84%	3,770 (5%)	67.96	62.77	[1]
Movielens_100k	B	gender	943	1,682	28.95%	93.69%	189 (25%)	68.73	73.42	[21]
Facebook*	B	gender	386,321	122,924	44.57%	99.94%	77,265 (30%)	83.03	84.55	[9]
Movielens_1m*	B	gender	6,040	3,706	28.29%	95.53%	1,208 (25%)	168.46	153.46	[21]
Libimseti*	B	gender	137,806	166,353	44.53%	99.93%	27,562 (30%)	229.16	226.97	[8]

Collect data sets and build models

Generate explanations for test instances

Evaluation

Textual data: linear/rbf SVM

Behavioral data: LR/MLP

SEDC

LIME-C

SHAP-C

Positively-predicted test instances

max. 5 minutes
max. 30 features

SEDC: max 50 iterations
LIME/SHAP-C: 5000 samples

EVALUATION CRITERIA

The **goal** is to find a small-sized counterfactual as fast as possible ➔ **tradeoff** between

- **Effectiveness**
  - Percentage explained
  - Switching point: # features in explanation
- **Efficiency**
  - Computation time in seconds

Collect data sets and build models

Generate explanations for test instances

Evaluation

Textual data: linear/rbf SVM

Behavioral data: LR/MLP

SEDC

LIME-C

SHAP-C

Positively-predicted test instances

max. 5 minutes
max. 30 features

SEDC: max 50 iterations
LIME/SHAP-C: 5000 samples

Percentage explained

Switching point

Computation time

RESULTS & CONCLUSION

EFFECTIVENESS

Table 2: Percentage explained

	Linear			Nonlinear		
Dataset	**SEDC** (%)	**LIME-C** (%)	**SHAP-C** (%)	**SEDC** (%)	**LIME-C** (%)	**SHAP-C** (%)
Flickr	**100**	**100**	**100**	**28.67**	28.33	**28.67**
Ecommerce	**100**	97.33	**100**	95.00	97.00	**99.67**
Airline	**100**	**100**	**100**	**100**	**100**	**100**
Twitter	**100**	**100**	**100**	**100**	**100**	**100**
Fraud	**100**	**100**	81.67	**100**	**100**	75
YahooMovies	**100**	**100**	**100**	98.67	**100**	**100**
TaFeng	**100**	**100**	**100**	93.33	**100**	**100**
KDD2015	**100**	**100**	**100**	99.67	**100**	99.67
20news	99.47	99.47	**100**	99.47	98.94	**100**
Movielens_100k	**100**	**100**	**100**	**100**	**100**	**100**
Facebook	**95.67**	95.00	95.00	70.33	**92.67**	89.67
Movielens_1m	**98.67**	**98.67**	**98.67**	88.33	95.00	**95.67**
Libimseti	**92.67**	90.33	88.67	77.00	**81.67**	72.33
Average	**98.96**	98.52	97.23	88.49	**91.82**	89.28
# wins	**12**	9	10	5	**9**	**9**

EFFECTIVENESS

Table 3: Switching point in # features (Median + Interquantile range)

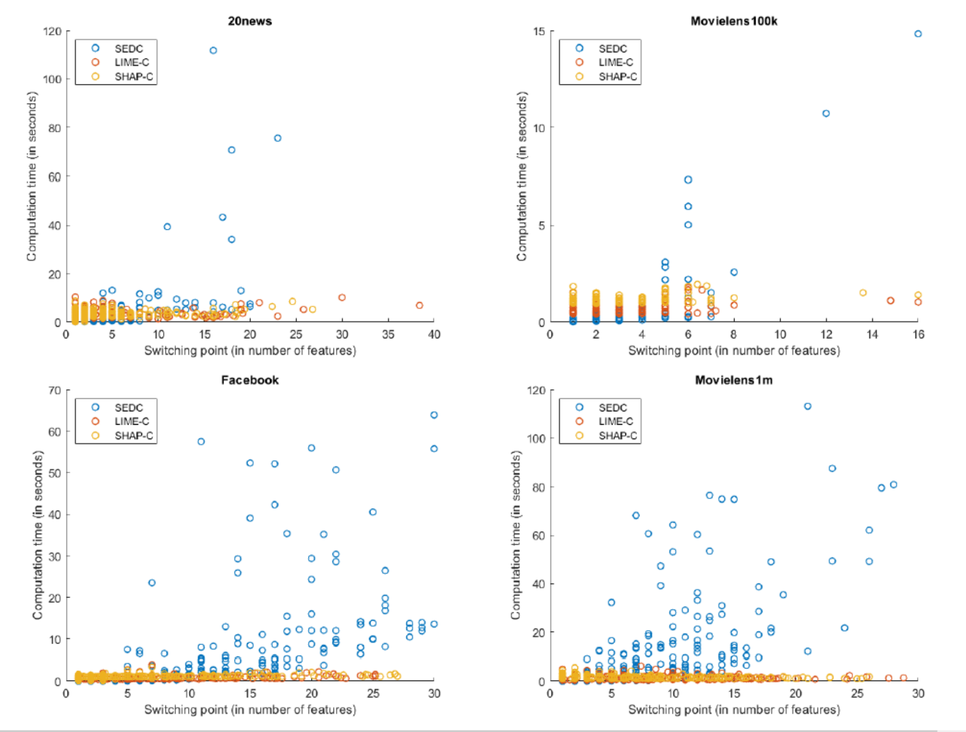
	Linear				Nonlinear			
Dataset	**SEDC**	**LIME-C**	**SHAP-C**	**Random**	**SEDC**	**LIME-C**	**SHAP-C**	**Random**
Flickr	**1(1-1)**	**1(1-1)**	**1(1-1)**	**1(1-1)**	**1(1-1)**	**1(1-1)**	**1(1-1)**	**1(1-2)**
Ecommerce	**1(1-1)**	**1(1-1)**	**1(1-1)**	**1(1-2)**	**1(1-1)**	**1(1-1)**	**1(1-1)**	**1(1-1)**
Airline	**1(1-2)**	**1(1-2)**	**1(1-2)**	2(1 – 3)	**1(1-1)**	**1(1-1)**	**1(1-1)**	2(1 – 3)
Twitter	**2(1-3)**	**2(1-3)**	**2(1-3)**	3(2 – 5)	**1(1-1)**	**1(1-1)**	**1(1-1)**	3(2 – 5.5)
Fraud	**1(1-1)**	**1(1-1)**	**1(1-1)**	**1(1-1)**	**1(1-1)**	**1(1-1)**	**1(1-1)**	**1(1-2)**
YahooMovies	**2(1-4)**	**2(1-4)**	**2(1-4)**	4(2 – 7)	**1(1-3)**	2(1 – 3)	2(1 – 3)	4(2 – 12)
TaFeng	**2(1-4)**	**2(1-4)**	**2(1-4)**	5(3 – 11)	**2(1-8)**	**2(1-3)**	**2(1-3.05)**	6(3 – 17)
KDD2015	**3(1-7)**	**3(1-7)**	**3(1-7)**	8.5(3 – 17.25)	**2(1-3)**	**2(1-3.95)**	**2(1-4.5)**	5(2 – 9)
20news	**2(1-4)**	**2(1-4)**	**2(1-4)**	11(4 – 23.5)	**1(1-3)**	**1(1-3)**	**1(1-3)**	8(3 – 18)
Movielens_100k	**2(1-4)**	**2(1-4)**	**2(1-4)**	5.5(3 – 10)	**2(1-4)**	**2(1-4)**	**2(1-4)**	5(2 – 9.25)
Facebook	**3(2-8)**	**3(2-8)**	**3(2-8)**	8(4 – 20)	4(1 – 13)	**3(1-4.4)**	**3(1.2-5)**	9(4.5 – 19.5)
Movielens_1m	**3(2-7)**	**3(2-7)**	**3(2-7)**	9(4 – 19.25)	**3(1-5)**	**3(1-6)**	**3(1-6)**	7(3 – 14)
Libimseti	**3(2-6)**	**3(2-6.2)**	**3(2-6.2)**	29(13 – 52)	**2(1-5)**	4.2(1.8 – 8.8)	5(2.5 – 11.2)	19(8 – 38.5)
# wins	**13**	**13**	**13**	3	**12**	11	11	3

EFFICIENCY

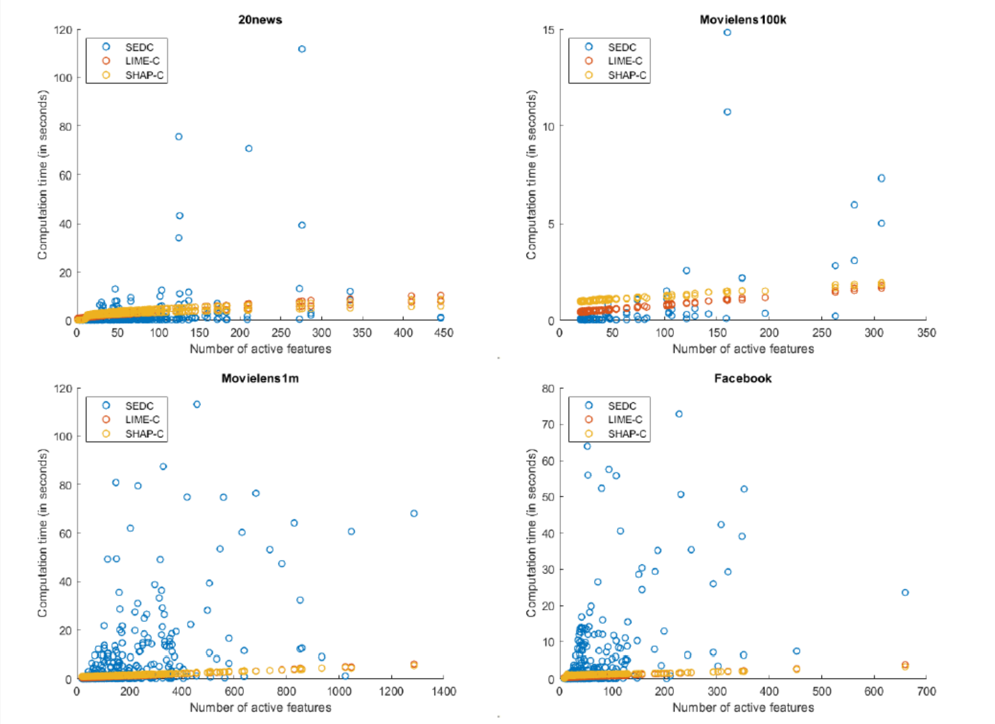
Table 4: Computation time in seconds (Median + Interquantile range)

	Linear			Nonlinear		
Dataset	**SEDC**	**LIME-C**	**SHAP-C**	**SEDC**	**LIME-C**	**SHAP-C**
Flickr	**0.01(0.00-0.02)**	0.34(0.33 – 0.35)	0.08(0.08 – 0.08)	**0.02(0.00-0.02)**	0.39(0.39 – 0.42)	0.12(0.09 – 0.25)
Ecommerce	**0.02(0.00-0.02)**	0.34(0.33 – 0.36)	**0.02(0.02-0.03)**	**0.02(0.00-0.02)**	0.39(0.38 – 0.41)	0.03(0.03 – 0.03)
Airline	**0.02(0.02-0.02)**	0.94(0.81 – 1.08)	0.09(0.03 – 0.60)	**0.02(0.02-0.02)**	1.35(1.17 – 1.51)	0.13(0.04 – 0.82)
Twitter	**0.03(0.02-0.05)**	0.61(0.56 – 0.64)	0.18(0.06 – 0.46)	**0.02(0.01-0.02)**	0.67(0.63 – 0.69)	0.15(0.06 – 0.47)
Fraud	**0.01(0.00-0.02)**	0.38(0.36 – 0.39)	0.07(0.06 – 0.08)	**0.01(0.01-0.01)**	0.43(0.42 – 0.44)	0.09(0.07 – 0.17)
YahooMovies	**0.03(0.02-0.08)**	0.44(0.43 – 0.49)	0.96(0.90 – 1.00)	**0.06(0.03-0.20)**	0.82(0.79 – 0.85)	1.35(1.28 – 1.39)
TaFeng	**0.05(0.02-0.22)**	0.50(0.45 – 0.59)	1.03(0.97 – 1.08)	**0.04(0.02-0.40)**	0.51(0.46 – 0.59)	1.01(0.95 – 1.06)
KDD2015	**0.11(0.02-0.79)**	0.52(0.47 – 0.61)	1.04(0.99 – 1.09)	**0.14(0.04-0.56)**	0.84(0.78 – 0.94)	1.37(1.31 – 1.45)
20news	**0.19(0.05-1.34)**	3.12(2.09 – 4.18)	3.65(2.74 – 4.49)	**0.09(0.03-0.68)**	2.16(1.49 – 2.95)	2.53(1.99 – 3.09)
Movielens_100k	**0.06(0.03-0.30)**	0.49(0.44 – 0.69)	0.87(0.83 – 1.04)	**0.09(0.04-0.35)**	0.55(0.50 – 0.83)	1.10(1.02 – 1.27)
Facebook	**0.12(0.03-1.17)**	0.55(0.46 – 0.75)	1.11(1.04 – 1.23)	**0.19(0.02-2.20)**	0.51(0.46 – 0.59)	1.06(1.00 – 1.12)
Movielens_1m	**0.37(0.06-3.09)**	0.74(0.52 – 1.21)	1.21(1.05 – 1.53)	**0.39(0.07-1.56)**	0.76(0.59 – 1.12)	1.29(1.16 – 1.54)
Libimseti	**0.36(0.14-2.26)**	1.07(0.92 – 1.38)	1.37(1.27 – 1.52)	**0.39(0.09-1.56)**	1.02(0.91 – 1.23)	1.42(1.35 – 1.53)
# wins	13	0	1	13	0	0

EFFICIENCY: time *vs* switching point



EFFICIENCY: time *vs* active features



CONCLUSION

- **SEDC** most efficient and effective for small instances, *however*
  - computation time very sensitive to switching point
  - flaw in heuristic best-first for some nonlinear models

- **SHAP-C** overall good performance, *however*
  - problems with highly unbalanced data
  - computation time more sensitive to # active features than LIME-C

⇒ **LIME-C** most favourable search algorithm: best tradeoff
  - low computation times
  - least sensitive to switching point and # active features
  - stable performance in terms of effectiveness criteria

FURTHER RESEARCH

- More data sets and models
- Study efficiency-effectiveness tradeoff of the algorithms
- Evaluate other hybrid algorithms
- Other objectives of the algorithm