Fraud Detection

DETECTION OF FRAUDULENT CREDIT CARD TRANSACTIONS

Benchmark vs. Logistic Regression, Random Forests, Boosted Trees & Neural Networks

UC#20 – 2024/07 (v2.2) xtractis.ai

PROBLEM DEFINITION

GOAL

Design, from only some transactions characteristics, an AI-based decision system that accurately diagnoses credit card transactions, in order to rationally and instantly detect the fraudulent ones to eventually adapt protection measures.

PROS & BENEFITS

- Identify the parameters involved in the fraudulent transactions and enhance knowledge by helping banking specialists understand the causal relationships between these parameters, their combination, and the occurrence of fraud (i.e., understand the scammers' strategies).
- Help the banking sector to make transparent decisions through automatic, explainable detection, while improving the consumer experience.
- Use a detection model with fewer transaction characteristics to speed up protection process.

REFERENCE DATA

Source: Worldline and Machine Learning Group of Université Libre de Bruxelles

Dataset: www.kaggle.com, 2014

Modeling work carried out by engineering students S. Beziat and I. Abou Khachfe (Spring semester 2023) with the UTC-dedicated XTRACTIS® platform

Variable to Predict

The model diagnoses the Transaction: **NORMAL** | **FRAUDULENT**

Potential Predictors

29 variables characterize each transaction: transaction amount, and 28 axes of inertia (dimensionality reduction via Principal Component Analysis to protect user identities and sensitive features).

Remark: to maintain intelligibility, we would have worked directly with the original variables. In the following, intelligibility is assessed in terms of the number of variables and rules retained in each model.

Observations

284,807 reference credit card transactions made by European cardholders within two days of September 2013. Data are divided into a Learning Dataset for model induction using Training and Validation Datasets, and an External Test Dataset to check the top-model's real performance on unknown cases and for benchmarking.

Learning Dataset: 242,085 | 85% transactions 82.35% for Training, 17.65% for Validation	
NORMAL	FRAUDULENT
241,667 | 99.83%	418 | 0.17%

External Test Dataset: 42,722 | 15% transactions	
NORMAL	FRAUDULENT
42,648 | 99.83%	74 | 0.17%

MODEL TYPE Regression | Multinomial Classification | **Binomial Classification** | Scoring

XTRACTIS-INDUCED DECISION SYSTEM

☑ **Intelligible Model, Explainable Decisions**

- The top-model is a decision system composed of 7 gradual rules without chaining.
- Each rule uses from 2 to 5 predictors among the 9 variables that XTRACTIS automatically identified as significant (out of the 29 ones characterizing transactions).
- Only a few rules are triggered at a time to compute the decision.

☑ **High Predictive Capacity** It has very good Real Performance (on unknown data).

☑ **Ready to Deploy** It computes real-time predictions up to 70,000 decisions/second, offline or online (API).

XTRACTIS PROCESS

STEPS

Reference Data → **INDUCTION** → XTRACTIS Top-Model

New Cases → **DEDUCTION** → NORMAL / FRAUDULENT

Automated Decision (detect fraud)

SOFTWARE ROBOTS

XTRACTIS® **REVEAL**
Delivers the decision system + its Structure & Performance Reports

XTRACTIS® **PREDICT**
Delivers the decision + the Prediction Report explaining its reasoning

TOP-MODEL INDUCTION

INDUCTION PARAMETERS

Powered by:
XTRACTIS® REVEAL
v12.2.45294

1. We launch 2,000 inductive reasoning strategies; each strategy is applied to the same single partition of the learning dataset (82.35% Training / 17.65% Validation) to get a reliable assessment of the descriptive and predictive performances, respectively from Training and Validation Datasets.
2. Each strategy thus generates one unitary model called **Individual Virtual Expert** (IVE).
3. Among the 2,000 induced models, the top-IVE selected is the one that has the best predictive performance, close to its descriptive performance, and with the best intelligibility, i.e. with the fewer predictors and rules.

Total number of induced unitary models
2,000 IVEs

Criterion for the induction optimization
F_1-Score

Validation criterion for the top-model selection
F_1-Score

Duration of the process @ Induction Speed FP64
10 days @ 1Tflops

TOP-MODEL STRUCTURE

The top-IVE model has an excellent intelligibility as it has **7 rules** combining **9 predictors**, with 3.9 predictors per rule on average.

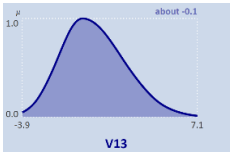
Its Structure Report reveals all the internal logic of the decision system and ensures that the model is understandable by the human expert. It is a transparent model that can be audited and certified before deployment to end-users.

PREDICTORS

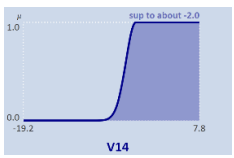
- 9 features out of 29
- Ranked by impact significance (2 strong signals, 6 medium signals, 1 weak signal): #1 V17 /#2 V14…
- Labeled by binary and fuzzy classes.

Examples:

fuzzy number
"about -0.1"



fuzzy interval
"sup. to about -2.0"



RULES

- 7 connective fuzzy rules without chaining (aggregated into 2 disjunctive fuzzy rules)
- 2 to 5 predictors per rule (on average, 3.9 predictors per rule)
- Example: **fuzzy rule R3** uses 4 predictors and concludes "NORMAL". 6 other fuzzy rules complete this model.

IF	V9	IS	superior to about -0.8
AND	V13	IS	about -0.1
AND	V14	IS	superior to about -2.0
AND	V26	IS	inferior to about -0.10
THEN	Transaction	IS	NORMAL

TOP-MODEL PERFORMANCE

The top-IVE performances, measured in Training / Validation, then in External Test on reference data, guarantee the model's predictive and real performances.

Performance Type / Dataset	DESCRIPTIVE 82.35% Training	PREDICTIVE 17.65% Validation	REAL External Test
F_1-Score	87.58%	87.32%	86.96%
Classification Error	0.04%	0.04%	0.04%

EXPLAINED PREDICTIONS FOR 2 UNKNOWN CASES

Powered by: XTRACTIS® PREDICT v12.2.45294

CASE
(from the External Dataset,
i.e., not included in the Learning Dataset)

DEDUCTIVE INFERENCE OF RULES

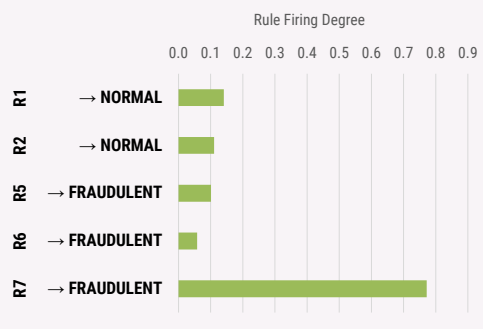
AUTOMATED DECISION

Transaction #6720

actual value = FRAUDULENT	
V1	-0.3
V9	-3.4
V10	-6.8
V11	7.6
...	...
V26	0.52

Real Time

For this transaction, 5 rules are triggered:
R7 is fired at 0.772, **R5** at 0.101 and
R6 at 0.058 to conclude "FRAUDULENT",
R1 is fired at 0.141 and **R2** at 0.111
to conclude "NORMAL"



NUMBER OF TRIGGERED RULES
5 / 7

FUZZY PREDICTION
{ FRAUDULENT | 0.772,
NORMAL | 0.141 }

FINAL PREDICTION
{ FRAUDULENT }

The system detects this fraud correctly
as in the observed situation:

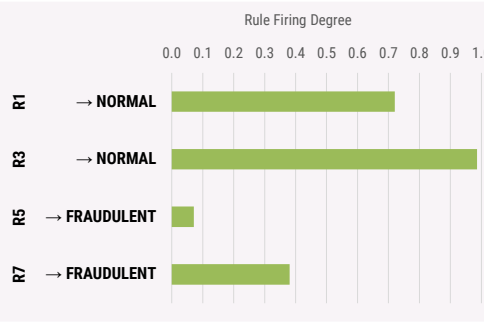
FRAUDULENT

Transaction #224

actual value = NORMAL	
V1	-2.4
V9	0.7
V10	0.2
V11	-1.2
...	...
V26	-0.31

Real Time

For this transaction, 4 rules are triggered:
R3 is fired at 0.986 and **R1** at 0.720
to conclude "NORMAL",
R7 is fired at 0.381 and **R5** at 0.071
to conclude "FRAUDULENT".



NUMBER OF TRIGGERED RULES
4 / 7

FUZZY PREDICTION
{ NORMAL | 0.986,
FRAUDULENT | 0.381 }

FINAL PREDICTION
{ NORMAL }

The system delivers a correct diagnosis
compared to the observed case:

NORMAL

TOP-MODELS BENCHMARK

MODELING PARAMETERS

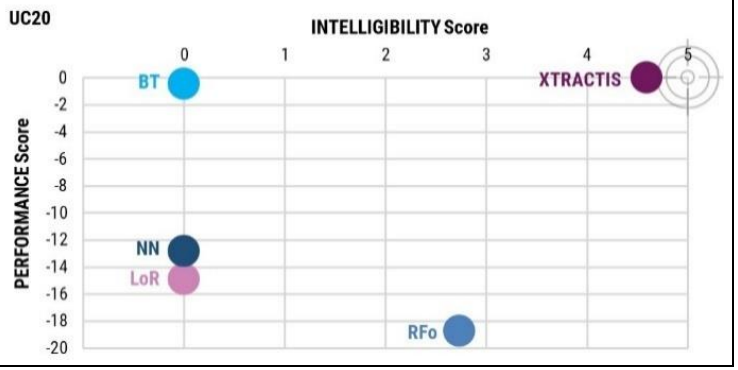
	XTRACTIS	LOGISTIC REGRESSION	RANDOM FOREST	BOOSTED TREE	NEURAL NETWORK
MODELS RELEASE	2023/06	2023/06	2023/06	2023/06	2023/06
ALGORITHM VERSION	XTRACTIS **REVEAL** 12.2.45294	Python 3.9.10 | Scikit-Learn 1.1.2	Python 3.9.10 | LightGBM 3.3.2	Python 3.9.10 | LightGBM 3.3.2	Python 3.9.10 | TensorFlow 2.10.0 | Keras 2.10.0
CROSS-VALIDATION TECHNIQUE	All explored strategies for all algorithms use the same single-split of the Learning Dataset: 82.35% Training | 17.65% Validation				
NUMBER OF EXPLORED STRATEGIES(1)	2,000 induction strategies	2,000 data analysis strategies	2,000 ML strategies	2,000 ML strategies	2,000 ML strategies
TOP-MODEL SELECTION(2)	Top-IVE among 2,000 IVEs	Top-IVE among 2,000 IVEs	Top-IVE among 2,000 IVEs	Top-IVE among 2,000 IVEs	Top-IVE among 2,000 IVEs

TOP-MODEL STRUCTURE

	XTRACTIS	LOGISTIC REGRESSION	RANDOM FOREST	BOOSTED TREE	NEURAL NETWORK
NUMBER OF PREDICTORS (out of 29 Potential Predictors)	**9**	**29**	**26**	**29**	**29**
AVERAGE NUMBER OF PREDICTORS PER RULE OR EQUATION	**3.9** per rule	**29.0** per equation	**4.9** per rule	**4.6** per rule	**22.5** per equation
STRUCTURE OF THE DECISION SYSTEM	**7** fuzzy rules without chaining (aggregated into 2 disjunctive rules) Only a few rules are triggered at a time to compute a decision	**1** linear equation	**12** trees **210** binary rules	**1** chain of **171** trees **2,217** binary rules Tree #N corrects the error of the N-1 previous trees	4 hidden layers | **44** hidden nodes **45** equations 44 unintelligible synthetic variables

TOP-MODEL SCORES

	Random(3)	XTRACTIS	LoR	RFo	BT	NN
INTELLIGIBILITY Score(4)		4.59	0.00	2.73	0.00	0.00
IVE Real Performance (F_1-Score) in External Test	*4.05*	86.96	72.05	68.24	86.52	74.14
Gap to IVE Leader in External Test		0.00	-14.91	-18.72	-0.44	-12.82
PERFORMANCE Score(4)		0.00	-14.91	-18.72	-0.44	-12.82



(1) For all algos: on the same Learning Dataset. All models are optimized according to their Validation F_1-Score.
(2) All top-models are selected according to their validation F_1-Score while checking that it remains close to their Training F_1-Score.
(3) Baseline performances that models must exceed to perform better than chance (P-value = 0.001; 100,000 models generated by random permutation of the output values). The value of each performance criterion is generally achieved by a different random model.
(4) See Appendices for explanations and detailed results. Performance Scores are calculated on all available unknown data.

More Use Cases:
xtractis.ai/use-cases/

APPENDIX 1 — Calculation of the Intelligibility × Performance

AI Technique #i	T_i	i∈[1 ; n] n = number of AI Techniques benchmarked in terms of data-driven modeling = 5
Benchmark #k	B_k	k∈[1 ; p] p = number of Benchmarks for the Use Case ∈ {1, 2, 3}

Remarks:

- In case of a small number of reference data, a CVE model (College of Virtual Experts) is generated by each explored strategy of T_i, generally via an N×K-fold cross validation. In this case, a Benchmark is led with the top-CVE on the External Test Dataset (ETD, composed of unknown reference cases). Then, a top-IVE model (Individual Virtual Expert) is generated from the top-CVE, through the XTRACTIS® reverse-engineering process, or for the other T_i, by applying the top-strategy, which has generated the top-CVE, on the Training and Validation Datasets. And a second Benchmark is led with this top-IVE on the same ETD.
- In case of a huge number of reference data, an IVE is generated by each explored strategy of T_i, via a 1-split validation. In this case, Benchmarks are led with the top-IVE on the Test Dataset (TD, composed of unknown reference cases) and on the available ETDs.
- Each Benchmark uses the latest versions of the following algorithms available at the date of the benchmark. XTRACTIS®: REVEAL; Logistic Regression: Python, Scikit-Learn; Random Forest & Boosted Tree: Python, LightGBM; Neural Network: Python, TensorFlow, Keras.
- Each B_k uses exactly the same TD and ETD for each T_i model.
- No Regression models can be obtained by Logistic Regression. So, this Data Analysis technique is benchmarked only for Classification or Scoring problems.
- The Holy Grail for critical AI-based decision systems is to obtain a model with the highest Performance and the highest Intelligibility scores (top-right corner of the graph).

PERFORMANCE Score

For each B_k, we calculate the values of the Performance Criterion (PC) on the same ETD for all the T_i top-CVEs; and on the same TD and ETDs for all the T_i top-IVEs. The PC is: RMSE in percentage for a Regression; F_1-Score for a Binomial Classification; Average F_1-Score or Average F_2-Score for a Multinomial Classification; Gini index for a Scoring.
Then, we compare the value of the PC of each T_i top-CVE (resp. top-IVE) to the best value of this PC reached by the best T_i top-CVE (resp. top-IVE) on ETD (resp. on TD and ETDs).

For Regression, we calculate for each T_i top-model (CVE and IVE): $PS(T_i, B_k) = Best\_PC(B_k) - PC(T_i, B_k)$.

For Classification and Scoring, we calculate for each T_i top-model: $PS(T_i, B_k) = PC(T_i, B_k) - Best\_PC(B_k)$.

Performance Score of T_i

$$PS(T_i) = Mean\ (PS(T_i, B_k))_{\ k \in [1\ ;\ p]}$$

Remark:

- Each PS varies theoretically from -100 (Lowest Score) to 0 (Highest Score), but practically between -50 and 0.

INTELLIGIBILITY Score

We consider the T_i top-IVE. Its Intelligibility Score $IS(T_i)$ is valued from 0.00 to 5.00 regarding the structure of the model: number of predictors, classes, rules, equations, trees, synthetic variables, modalities to predict for classifications (or numeric variables to predict for regressions or scoring). The more compact the model, the higher its IS.

The IS of each T_i is obtained by accumulating the following five penalty values to the ideal IS value of 5.00 (each penalty has a null or a negative value):

- Penalty 1 (logarithmic penalty regarding the number of predictors):

  $Pen1(T_i) = \min(0\ ,\ 1 - \log_{10} number\ of\ predictors)$

  Examples: *Pen1 = 0.00 for up to 10 predictors*
  Pen1 = −3.00 for 10.000 predictors

- Penalty 2 (linear penalty regarding the average number of rules or equations per modality to predict):

  $Pen2(T_i) = \min\left(0\ ,\ 0.01 - \frac{average\ number\ of\ rules\ or\ equations\ per\ modality\ to\ predict}{100}\right)$

  Examples: *Pen2 = 0.00 for 1 rule or equation per modality to predict on average*
  Pen2 = −3.00 for 301 rules or equations per modality to predict on average

- Penalty 3 (linear penalty regarding the average number of predictors per rule or equation):

  $Pen3(T_i) = \min\left(0\ ,\ \frac{9 - 3 \times average\ number\ of\ predictors\ per\ rule\ or\ equation}{7}\right)$

  Examples: *Pen3 = 0.00 for up to 3.0 predictors per rule or equation on average*
  Pen3 = −3.00 for 10.0 predictors per rule or equation on average

- Penalty 4 (linear penalty regarding the number of trees per chain, here for BT only):

  $Pen4(T_i) = \min(0\ ,\ 1 - number\ of\ trees\ per\ chain)$

  Examples: *Pen4 = 0.00 for 1 tree per chain*
  Pen4 = −3.00 for 4 trees per chain

- Penalty 5 (maximum penalty due to unintelligibility of synthetic variables, here for NN only):

  $Pen5(T_i) = -5$

Intelligibility Score of T_i

$$IS(T_i) = \max(0.00\ ,\ 5.00 + (Pen1+Pen2+Pen3+Pen4+Pen5))$$

Remarks:

- For the difference between the Intelligibility and the Explainability of a model, please see the XTRACTIS® Brochure, page 7.
- The real complexity of the process/phenomenon under study is intrinsic, i.e., it could not be reduced or simplified, but only discovered; thus, the top-model will be complex if the process/phenomenon turns out to be complex [Zalila 2017]. Consequently, for some complex process/phenomenon, IS can be equal to 3.00 or less, even if Ti natively produces intelligible models (XTRACTIS, Random Forest).
- For similar structures, the Boosted Tree model is always less intelligible than the Random Forest one, as it is composed of chains of trees, instead of a college of trees (see Penalty 4).
- Neural Network model has always the lowest IS of 0.00, because it uses synthetic unintelligible variables (hidden nodes) in addition to all the potential predictors (see Penalty 5).

APPENDIX 2 — Use Case Results (all Performance criteria of all Top-Models)

Performance Criterion	Classification Error	Min. Sensitivity Specificity	Sensitivity	Specificity	PPV	NPV	**F_1-Score**	Refusal
RANDOM MODEL								
Number of Random Permutations (P-value) = 100,000 (0.001%)								
Performance against chance	*0.33%*	*4.05%*					*4.05%*	
XTRACTIS TOP-MODEL								
IVE - Descriptive Performance (Training)	0.04%	81.98%	81.98%	99.99%	94.00%	99.97%	87.58%	1 (0.00%)
IVE - Predictive Performance (Validation)	0.04%	83.78%	83.78%	99.99%	91.18%	99.97%	87.32%	0 (0.00%)
IVE - Real Performance (External Test)	0.04%	81.08%	81.08%	99.99%	93.75%	99.97%	**86.96%**	0 (0.00%)
LOGISTIC REGRESSION TOP-MODEL								
IVE - Descriptive Performance (Training)	0.09%	81.69%	81.69%	99.94%	70.43%	99.97%	75.64%	
IVE - Predictive Performance (Validation)	0.09%	83.78%	83.78%	99.94%	70.45%	99.97%	76.54%	
IVE - Real Performance (External Test)	0.11%	78.38%	78.38%	99.93%	66.67%	99.96%	**72.05%**	
RANDOM FOREST TOP-MODEL								
IVE - Descriptive Performance (Training)	0.11%	77.62%	77.62%	99.92%	64.03%	99.96%	70.17%	
IVE - Predictive Performance (Validation)	0.15%	85.14%	85.14%	99.88%	54.31%	99.97%	66.32%	
IVE - Real Performance (External Test)	0.13%	78.38%	78.38%	99.91%	60.42%	99.96%	**68.24%**	
BOOSTED TREE TOP-MODEL								
IVE - Descriptive Performance (Training)	0.01%	99.99%	100.00%	99.99%	100.00%	97.45%	99.99%	
IVE - Predictive Performance (Validation)	0.03%	86.49%	86.49%	99.99%	99.98%	90.78%	99.99%	
IVE - Real Performance (External Test)	0.04%	82.43%	82.43%	99.99%	99.97%	**86.52%**	**99.99%**	
NEURAL NETWORK TOP-MODEL								
IVE - Descriptive Performance (Training)	0.07%	76.16%	76.16%	99.97%	80.12%	99.96%	78.09%	
IVE - Predictive Performance (Validation)	0.07%	78.38%	78.38%	99.96%	79.45%	99.96%	78.91%	
IVE - Real Performance (External Test)	0.09%	75.68%	75.68%	99.95%	72.73%	99.96%	**74.14%**	