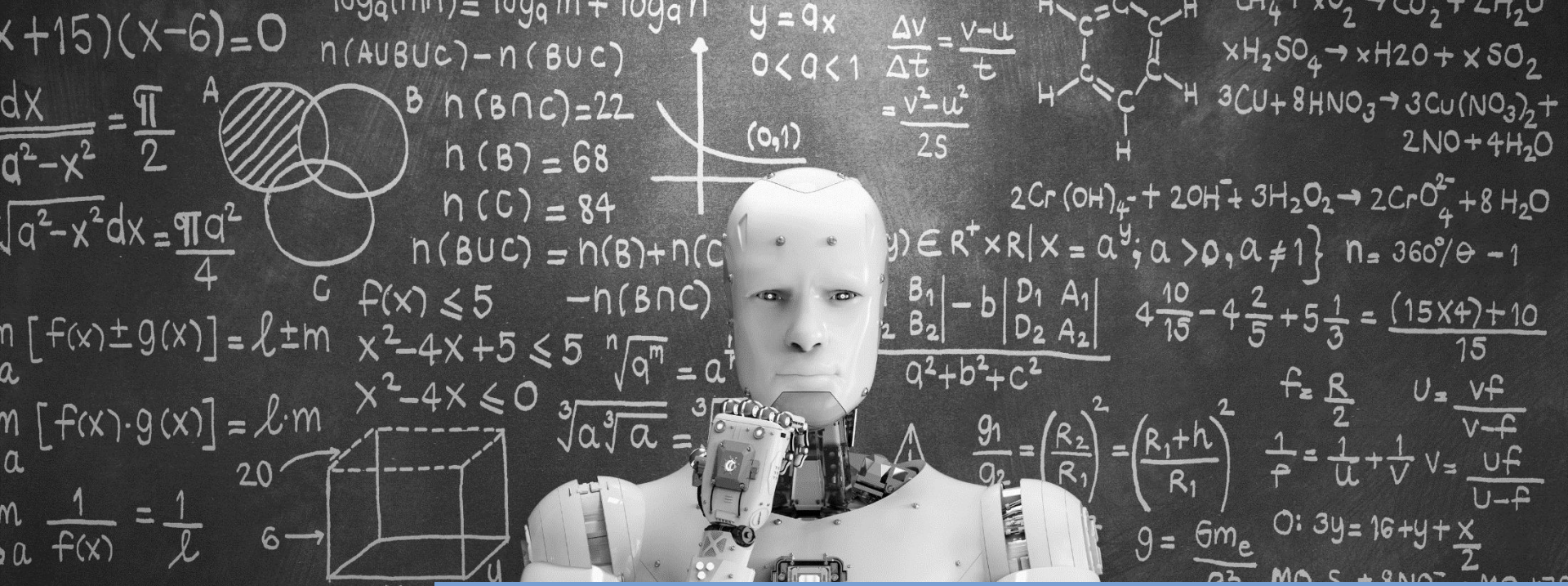
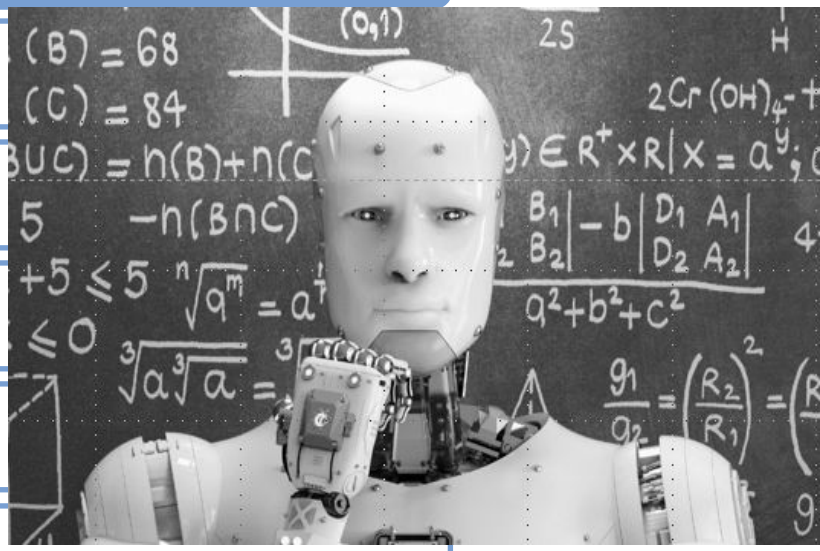


E-Banking-Fraud-Challenge

Eigentümlichkeiten des POC-Datensatzes

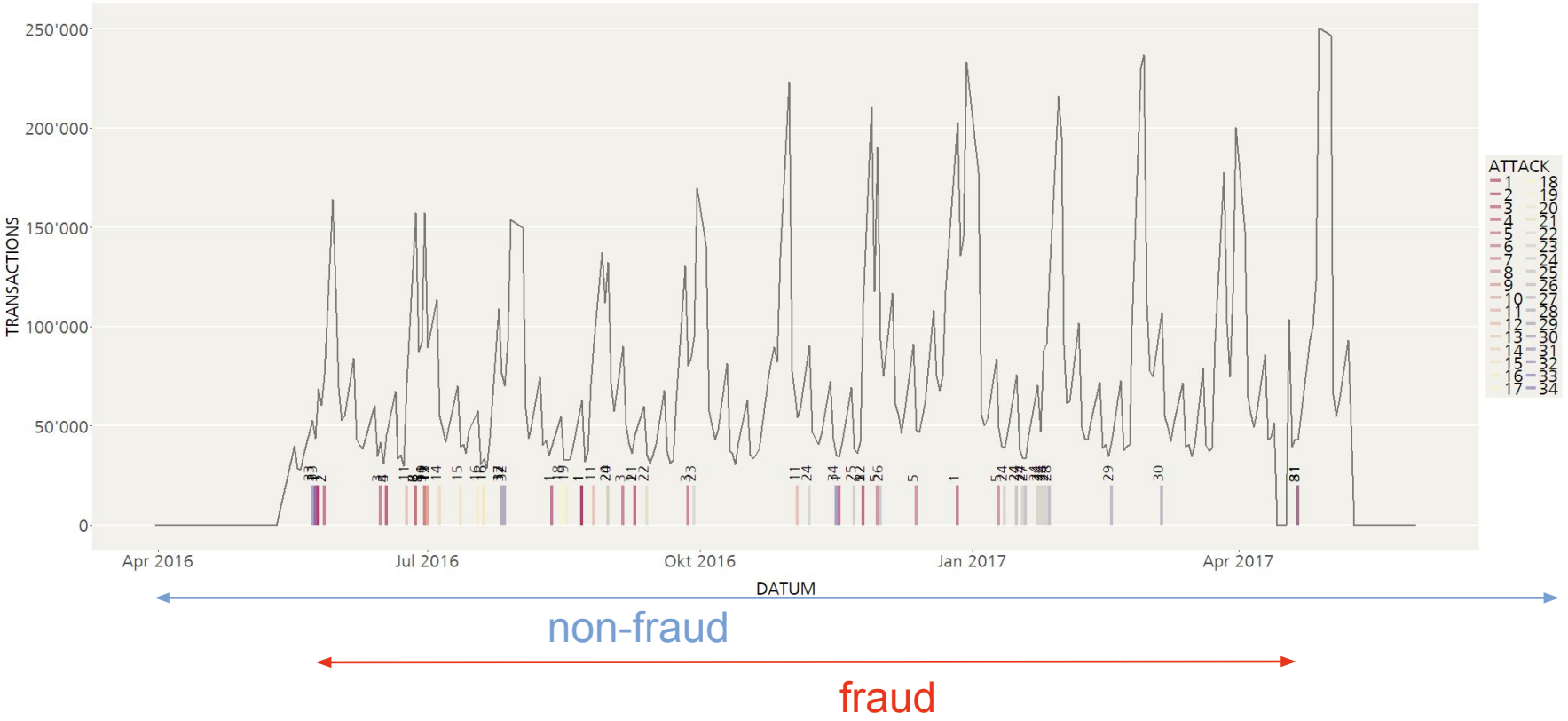
E-Banking Fraud-Detection

1. Datengrundlage
2. Datenmodell
3. Data-Leakage
4. Abhängigkeit der Daten
5. Kreuzvalidierung mit abhängigen Daten
6. Passende Metrik: AUC?



Übersicht Datensatz

Der Datensatz besteht aus 18’164’183 regulären Transaktionen aus dem E-Banking. Dem gegenüber stehen 73 identifizierte Fraud-Fälle. Die regulären Transaktionen verteilen sich über den Zeitraum vom 31.03.2016 bis 31.05.2017. Ab dem 10.05.2017 liegen nur sehr wenige Daten vor.
Die Fraud-Fälle erstrecken sich über den Zeitraum vom 23.05.2016 bis zum 21.04. 2017 und sind in unten stehender Graphik farblich markiert.



Schwierigkeiten des Datensets

Der Datensatz besteht aus 18’164’183 regulären Transaktionen aus dem E-Banking. Dem gegenüber stehen 73 identifizierte Fraud-Fälle. Information über klassische Fraud-Muster kann nur aus den 73 Fraud-Fällen gewonnen werden. Die Schwierigkeit ist, Regeln zu finden, die nicht nur spezifisch für die 73 Fraud-Fälle sind, sondern darüber hinaus auch auf anderen Transaktionen generalisieren.

51'375 mit Betrag 5000

Einzige Zahlung an diesen Empfänger;
Aber: 1'609'233 einmalige Empfänger bei non-fraud

ZAHLUNGSART	MEDIUM	WAEHRUNG	BETRAG	SAMMELAUFTRAG	ADRESSE		GEGENPARTEI_BANK_ADRESSE
Zahlung Schweiz	Einzelzahlung	CHF	5000	S	Maci	)06 Zürich	UBS Switzerland AG 8049 Zürich

12'235'539 mit Sammelauftrag ‘S’

1006 non-frauds an diese Gegenpartei_Bank_Adresse

1’491 mit Betrag 4100

Einzige Zahlung an diesen Empfänger;
Aber: 1'609'233 einmalige Empfänger bei non-fraud

ZAHLUNGSART	MEDIUM	WAEHRUNG	BETRAG	SAMMELAUFTR...	ADRESSE		GEGENPARTEI_BANK_ADRESSE
Zahlung Ausland	Einzelzahlung	CHF	4100	S	RA\	HAILAND	KRUNG THAI BANK PUBLIC COMPANY LIMITED ...

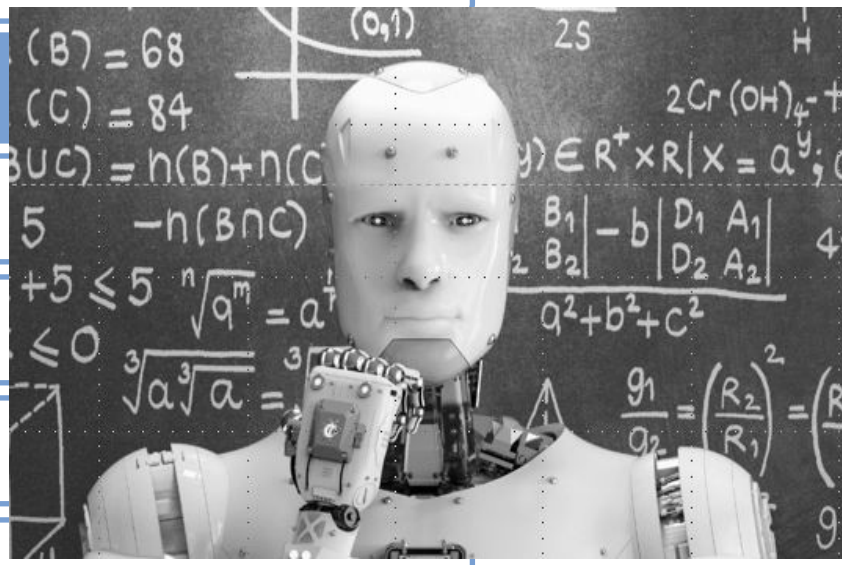
87'955 mit Zahlungsart ‘Ausland’

12'235'539 mit Sammelauftrag ‘S’

279 non-frauds an diese Gegenpartei_Bank_Adresse

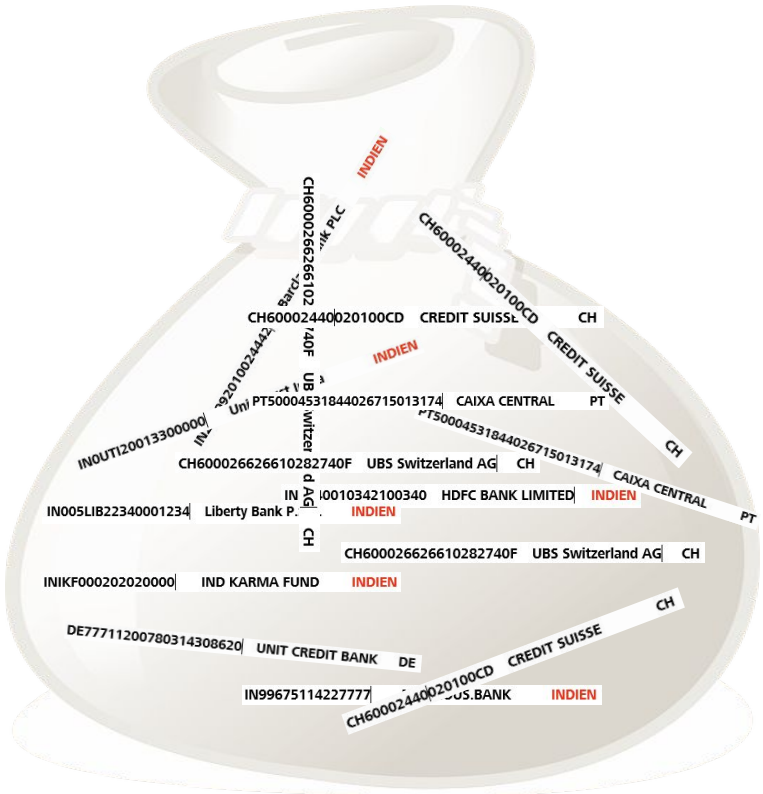
E-Banking Fraud-Detection

1. Datengrundlage
2. Datenmodell
3. Data-Leakage
4. Abhängigkeit der Daten
5. Kreuzvalidierung mit abhängigen Daten
6. Passende Metrik: AUC?



Datenmodell: Bag of transactions (bot)

Dieses Datenmodell führt zu fehlerhafter Berechnung von Variablen. Information, die erst zu einem späteren Zeitpunkt vorliegt fliesst in frühere Daten mit ein («leakage»).
Beispiel: Anzahl frauds pro Land. Wenn jedem fraud-case als Variablewert die Gesamtzahl der Fälle pro Land zugewiesen wird, so missachtet man den sequentiellen Character der Daten. Das Modell generalisiert anschliessend schlechter auf «echte» Daten.



=

Land	# Fraud pro Land
CH	0
PT	0
IN	6
DE	0

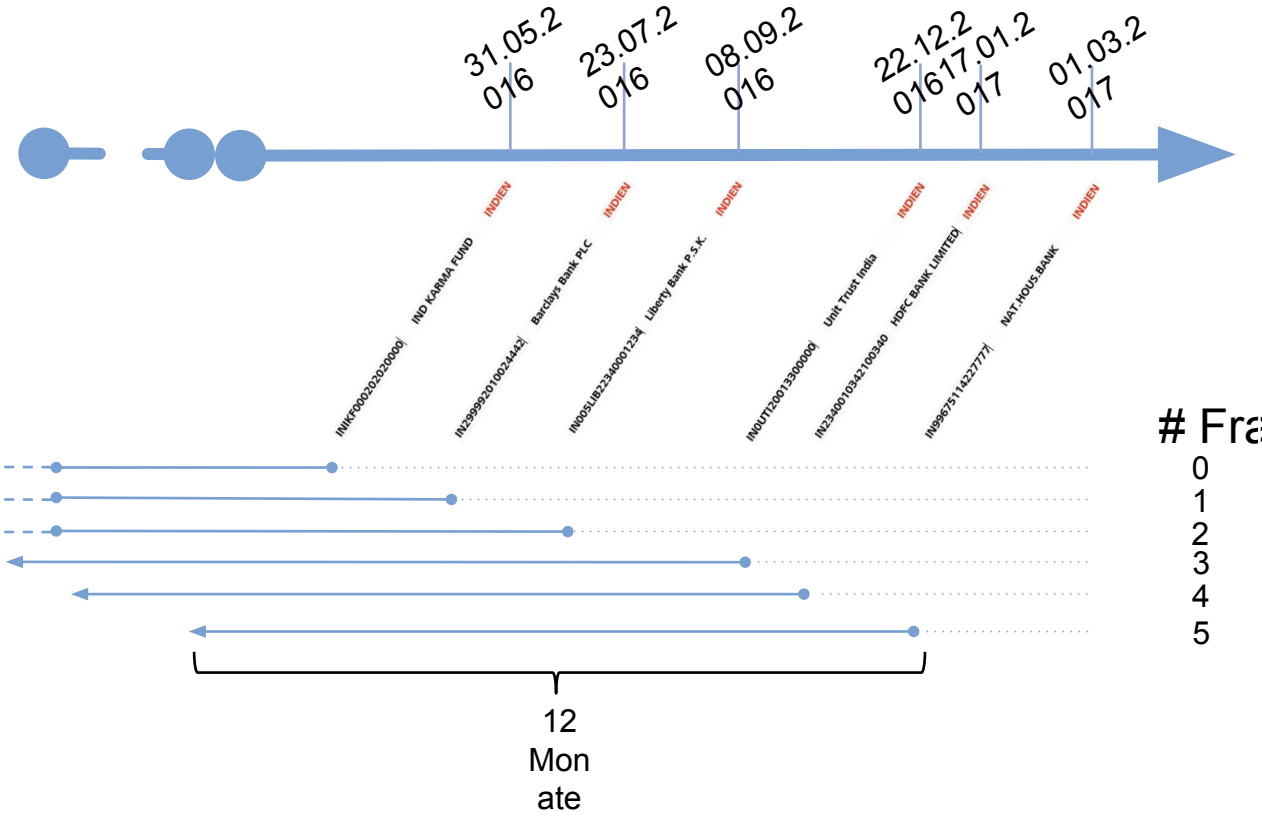
-

Aber die Daten sind sequentiell:

Datum	Bank IN	# Fraud	Falsch
31.05.2016	IND KARMA FUND	0	6
23.07.2016	Barclay PLC	1	6
08.09.2016	Liberty Bank	2	6
22.12.2016	United Trust F.	3	6
17.01.2017	HDFC BANK	4	6
01.03.2017	NAT.HOUS.BANK	5	6

Datenmodell: Sequence of transactions (sot)

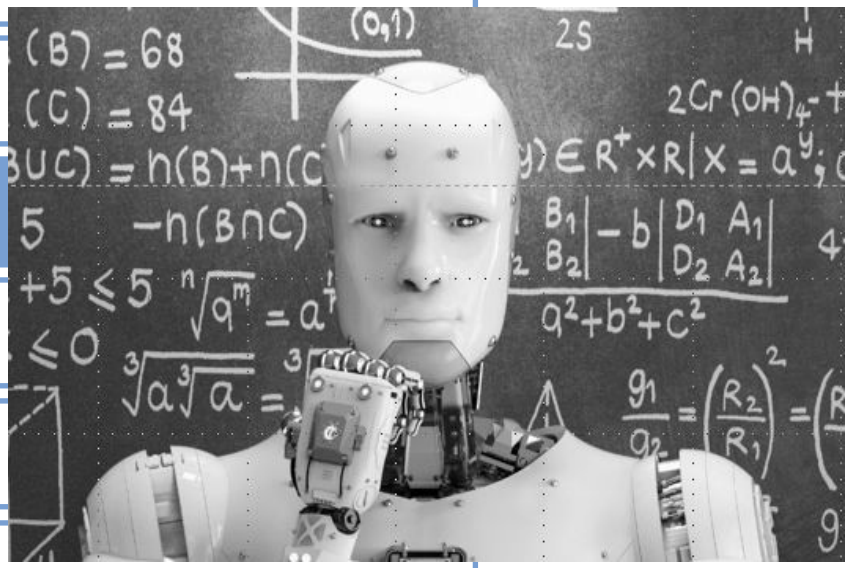
Für dieses Datenmodell werden gleich lange Zeitperioden vor dem interessierenden Ereignis zur Berechnung der Variablen herangezogen. Somit liegt nur Information aus der Vergangenheit vor.
Beispiel: Anzahl frauds pro Land. Wird jedem fraud-case als Variablenwert die Anzahl der vorherigen Fälle zugewiesen, so wird nur Information verwenden, die zum jeweiligen Zeitpunkt auch effektiv vorlag. Das Modell generalisiert ohne Performanz-Verluste.



# Fraud	falsch
0	6
1	6
2	6
3	6
4	6
5	6

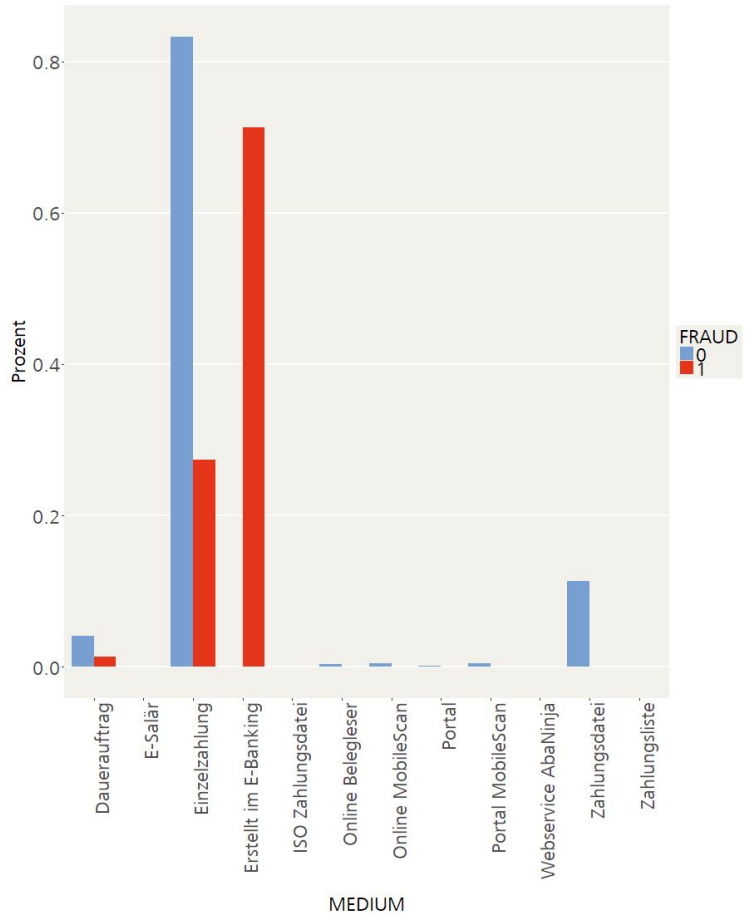
E-Banking Fraud-Detection

1. Datengrundlage
2. Datenmodell
3. Data-Leakage
4. Abhängigkeit der Daten
5. Kreuzvalidierung mit abhängigen Daten
6. Passende Metrik: AUC?



Data - Leakage

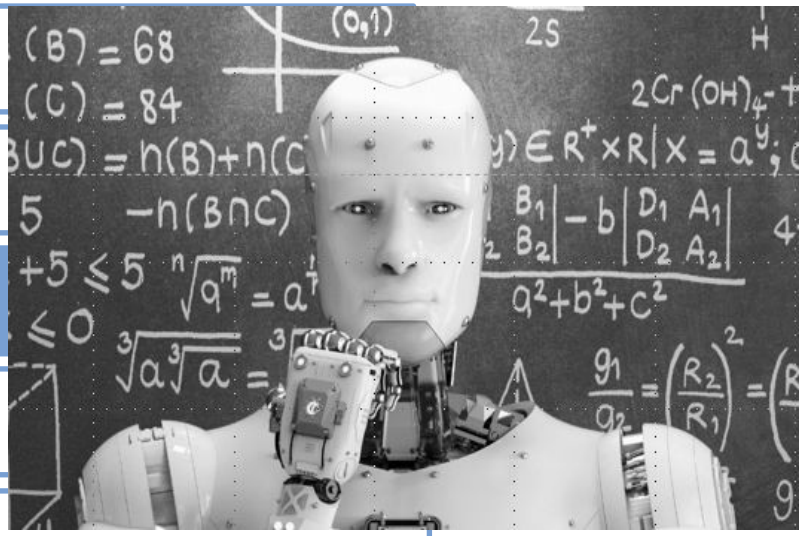
Durch Fehler bei der Erstellung von Daten-Sets kann Information über die Klassenzugehörigkeit entstehen, die unter normalen Bedingungen nicht zur Verfügung steht. Beispielsweise korreliert die Datensatz-ID mit den Klassen, wenn postivie und negative Fälle aneinander gefügt wurden. Oder alle positiven Fälle enstammen nur einem bestimmten Zeitraum. Bei Wettbewerben wird diese Information gezielt gesucht und ausgenutzt um das Preisgeld zu erlangen.



71% der Fraud-Fälle haben auf der Variable «Medium» die Ausprägung **«Erstellt im E-Banking»**. Diese Kategorie kommt bei den regulären Transaktionen kein einziges Mal vor. Die meisten Fraud-Fälle entstammen einer seperaten Excel-Datei in der diese Kategorie in der Spalte «Medium» eingetragen ist. Durch Unachtsamkeit oder gezieltes Ausnutzen erlangt man a priori 71% Klassifikationsgenauigkeit.

E-Banking Fraud-Detection

1. Datengrundlage
2. Datenmodell
3. Data-Leakage
4. Abhängigkeit der Daten
5. Kreuzvalidierung mit abhängigen Daten
6. Passende Metrik: AUC?



Die Fraud-Fälle sind nicht unabhängig

Mehr als die Hälfte der Fraud-Fälle gehört einer Serie von frauds an. Vermutlich kommen solche Serien durch die wiederholte Ausnutzung einer Schwachstelle zustande. Die Anzahl der unabhängigen Fraud-Fälle reduziert sich von 73 auf ca. 36. Somit liegt für das Trainieren von machine learning Algorithmen noch weniger Information vor.
Hier werden einige Beispiele aufgeführt.

AUSFUEHRUNGSDA...	ADRESSE		BUSINESS_...	GEGENPARTEI_BANK_ADRESSE
20.07.2016 00:00:00	three hi	700017 kol...	9466	HDFC BANK LIMITED HDFC BANK LTD P34-AP...
18.07.2016 00:00:00	windov	ukee	9466	PNC BANK, N.A. 15219 PITTSBURGH,PA Verein...
20.07.2016 00:00:00	windov	ukee	9466	PNC BANK, N.A. 15219 PITTSBURGH,PA Verein...

Wiederholte betrügerische Überweisungen bei gleichem Business-Partner.

AUSFUEHRUNGSD...	ADRESSE		BUSINESS_PA...	GEGENPARTEI_BANK_ADRESSE
25.11.2016 00:00:00	atlar	nited gb...	162	BARCLAYS BANK PLC 1 CHURCHILL PLACE E14 5...
30.11.2016 00:00:00	atlar	nited gb...	285	BARCLAYS BANK PLC 1 CHURCHILL PLACE E14 5...
10.01.2017 00:00:00	dudl	om air bn...	678	BARCLAYS BANK PLC 1 CHURCHILL PLACE E14 5...
17.06.2016 00:00:00	loftc	d gb-bd7...	705	BARCLAYS BANK PLC 1 CHURCHILL PLACE E14 5...
13.12.2016 00:00:00	atlar	nited gb...	835	BARCLAYS BANK PLC 1 CHURCHILL PLACE E14 5...

Überweisungen zur selben Gegenpartei-Bank-Adresse, teilweise auch zu identischen Gegenpartei-Adressen.

AUSFUEHRUNGSDATUM	ADRESSE		BUSINESS_...	GEGENPARTEI_BANK_ADRESSE
26.01.2017 00:00:00	vinc	k-5270 odense.n.	1664	NORDEA BANK DANMARK A/S STRANDGADE 3 9...
07.11.2016 00:00:00	vinc	enseg dense.n. 5270	1741	NORDEA BANK DANMARK A/S STRANDGADE 3 9...
24.01.2017 00:00:00	ravi	and	2318	KRUNG THAI BANK PUBLIC COMPANY LIMITED 35 ...
12.01.2017 00:00:00	vinc	enseg dense.n. 5270	2318	NORDEA BANK DANMARK A/S STRANDGADE 3 9...
31.08.2016 00:00:00	vinc	enseg dense.n. 5270	2318	NORDEA BANK DANMARK A/S STRANDGADE 3 9...
16.01.2017 00:00:00	stef	stria	2318	Raiffeisen Regionalbank Moedling eGen Hauptstra...
16.01.2017 00:00:00	stef	stria	2318	Raiffeisen Regionalbank Moedling eGen Hauptstra...
18.01.2017 00:00:00	stef	triche	2318	Raiffeisen Regionalbank Moedling eGen Hauptstra...
23.01.2017 00:00:00	stef	triche	2318	Raiffeisen Regionalbank Moedling eGen Hauptstra...
25.01.2017 00:00:00	stef	stria	2318	UniCredit Bank Austria AG Schottengasse 6-8 1010 ...

Mehrfache betrügerische Überweisungen bei gleichem Business-Partner. Andere Überweisungen zu identischen Gegenpartei-Adressen.

Die Fraud-Fälle sind nicht unabhängig

Mehr als die Hälfte der Fraud-Fälle gehört einer Serie von frauds an. Vermutlich kommen solche Serien durch die wiederholte Ausnutzung einer Schwachstelle zustande. Die Anzahl der unabhängigen Fraud-Fälle reduziert sich von 73 auf ca. 36. Somit liegt für das Trainieren von machine learning Algorithmen noch weniger Information vor.
Hier werden einige Beispiele aufgeführt.

AUSFUEHRUNGSDATUM	ADRESSE	BUSINESS_P...	GEGENPARTEI_BANK_ADRESSE
21.04.2017 00:00:00	mateo technologies llc us-queen cre...	2651	BANK OF AMERICA, N.A. 94104 SAN FRANCIS...
21.04.2017 00:00:00	adyen client management fou 1011dj	2651	Deutsche Bank AG Zürich Branch 8021 Zürich
21.04.2017 00:00:00	adyen client management fou 1011dj	2651	Deutsche Bank AG Zürich Branch 8021 Zürich
21.04.2017 00:00:00	adyen client management fou 1011dj	2651	Deutsche Bank AG Zürich Branch 8021 Zürich

Wiederholte betrügerische Überweisungen bei gleichem Business-Partner.

AUSFUEHRUNGSDATU...	ADRESSE	BUSINESS_PART...	GEGENPARTEI_BANK_ADRESSE
17.11.2016 00:00:00	mr stfach ch-809...	1267	UBS Switzerland AG 8098 Zürich
22.08.2016 00:00:00	kaa	1381	PostFinance AG 3030 Bern
22.08.2016 00:00:00	ma ietikon	1381	UBS Switzerland AG 8098 Zürich
24.05.2016 00:00:00	eric gnon	1777	PostFinance AG 3030 Bern
25.05.2016 00:00:00	ale paderborn	1777	SPARKASSE HERFORD AUF DER...
25.05.2016 00:00:00	ger quier-montbarry	1777	UBS Switzerland AG 1630 Bulle
12.08.2016 00:00:00	ack mingerstrasse 2...	2137	PostFinance AG 3030 Bern
25.11.2016 00:00:00	san	2428	PostFinance AG 3030 Bern
27.12.2016 00:00:00	ber es 31 1227 car...	2534	PostFinance AG 3030 Bern
09.09.2016 00:00:00	frai 1 8127 forch	5352	PostFinance AG 3030 Bern
01.07.2016 00:00:00	ma en	8116	PostFinance AG 3030 Bern
30.06.2016 00:00:00	ma en	8116	PostFinance AG 3030 Bern

Wiederholte betrügerische Überweisungen bei gleichem Business-Partner. Teilweise Überschneidung der Gegenpartei-Bank-Adressen.

AUSFUEHRUNGSDATUM	ADRESSE	BUSINESS_P...	GEGENPARTEI_BANK_ADRESSE
27.06.2016 00:00:00	mra sse 8/15 at-10...	1055	BAWAG P.S.K. Bank fuer Arbeit und Wirtschaft und ...
21.04.2017 00:00:00	niko gasse 49/5	1969	BAWAG P.S.K. Bank fuer Arbeit und Wirtschaft und ...

Gleiche Gegenpartei-Bank-Adresse

Die Fraud-Fälle sind nicht unabhängig

Mehr als die Hälfte der Fraud-Fälle gehört einer Serie von frauds an. Vermutlich kommen solche Serien durch die wiederholte Ausnutzung einer Schwachstelle zustande. Die Anzahl der unabhängigen Fraud-Fälle reduziert sich von 73 auf ca. 36. Somit liegt für das Trainieren von machine learning Algorithmen noch weniger Information vor.
Hier werden einige Beispiele aufgeführt.

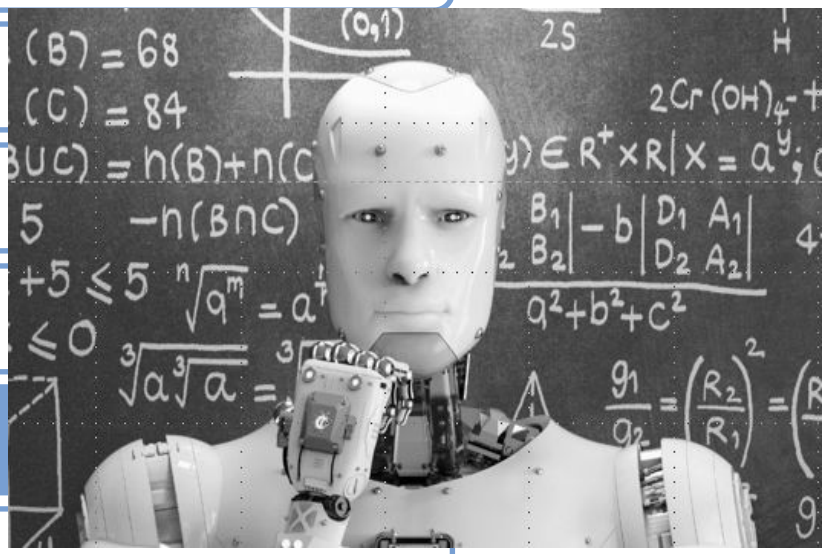
AUSFUEHRUNGS_DATUM	ADRESSE		BUSINESS_PARTNER	GEGENPARTEI_BANK_ADRESSE
23.05.2016 00:00:00	Ma	ol 8006 Zürich	1206	UBS Switzerland AG 8049 Zürich
24.05.2016 00:00:00	Ma	ol 8006 Zürich	1206	UBS Switzerland AG 8049 Zürich
31.08.2016 00:00:00	Vin	ensegyden 23 ...	2318	NORDEA BANK DANMARK A/S ...
12.01.2017 00:00:00	Vin	ensegyden 23 ...	2318	NORDEA BANK DANMARK A/S ...
16.01.2017 00:00:00	ste	ria	2318	Raiffeisen Regionalbank Moedli...
16.01.2017 00:00:00	STE	ustria	2318	Raiffeisen Regionalbank Moedli...
18.01.2017 00:00:00	ste	iche	2318	Raiffeisen Regionalbank Moedli...
23.01.2017 00:00:00	ste	iche	2318	Raiffeisen Regionalbank Moedli...
25.01.2017 00:00:00	ste	ria	2318	UniCredit Bank Austria AG Scho...
21.04.2017 00:00:00	Ad	nent Fou 10...	2651	Deutsche Bank AG Zürich Branc...
21.04.2017 00:00:00	Ad	nent Fou 10...	2651	Deutsche Bank AG Zürich Branc...
21.04.2017 00:00:00	Ad	nent Fou 10...	2651	Deutsche Bank AG Zürich Branc...
30.06.2016 00:00:00	Ma	ittisellen	8116	PostFinance AG 3030 Bern
01.07.2016 00:00:00	Ma	ittisellen	8116	PostFinance AG 3030 Bern
18.07.2016 00:00:00	Wi	53172 South ...	9466	PNC BANK, N.A. 15219 PITTSBU...
20.07.2016 00:00:00	Wi	53172 South ...	9466	PNC BANK, N.A. 15219 PITTSBU...

Echte Dubletten:
- Business_Partner
- Gegenpartei

D.h., der Datensatz beinhaltet 14% echte Dubletten und verkleinert die Anzahl der frauds auf 63 (anstatt 73)

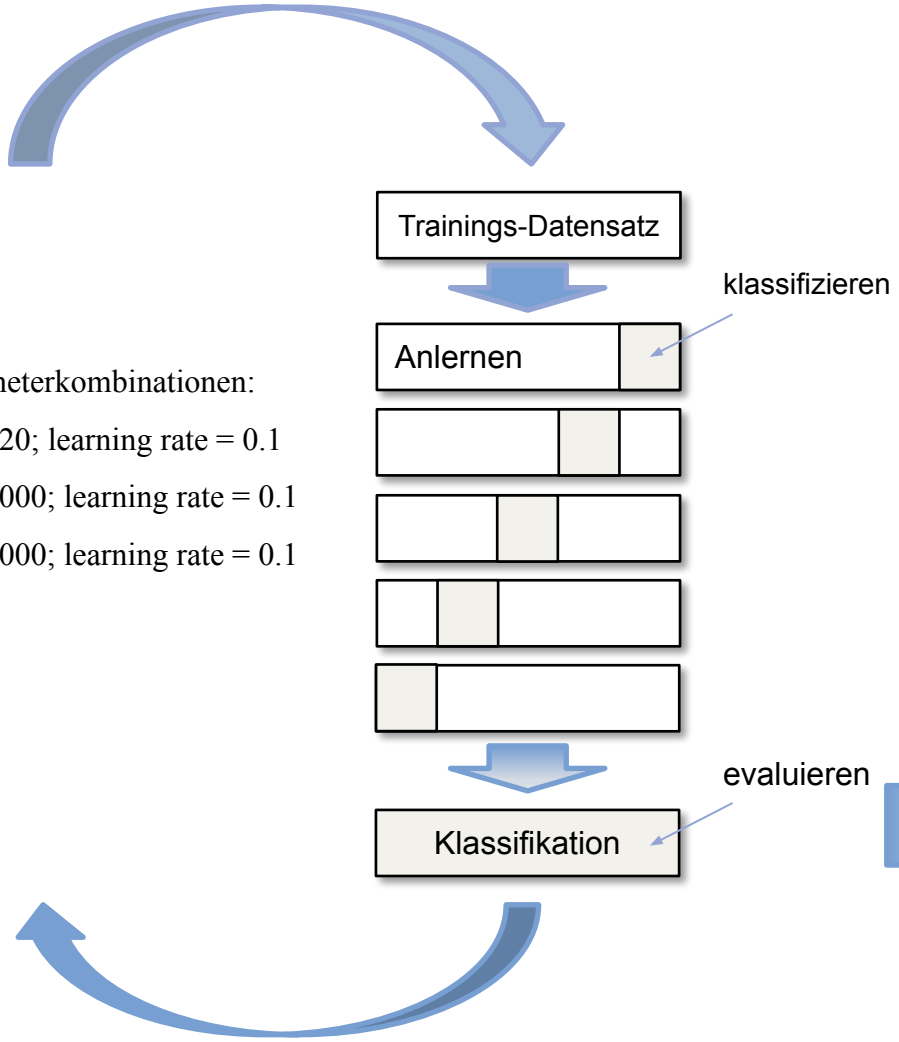
E-Banking Fraud-Detection

1. Datengrundlage
2. Datenmodell
3. Data-Leakage
4. Abhängigkeit der Daten
5. Kreuzvalidierung mit abhängigen Daten
6. Passende Metrik: AUC?



Wie funktioniert Kreuzvalidierung?

Beispiel: 5-fache Kreuzvalidierung



verschiedene Parameterkombinationen:

depth = 4; trees = 120; learning rate = 0.1

depth = 5; trees = 2000; learning rate = 0.1

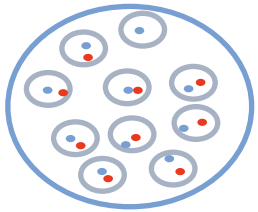
depth = 6; trees = 2000; learning rate = 0.1

Zweck:
«overfitting» verhindern; Algorithmus soll allgemeingültige Regeln finden und nicht Besonderheiten der Trainingsdaten lernen;

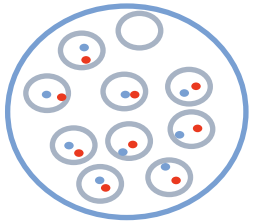
beste Parameterkombination wird anschliessend gewählt

Gedankenexperiment

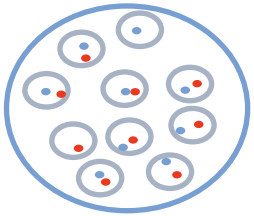
Angenommen die Fraud-Fälle tauchen immer paarweise auf. Hat man 10 solcher Paare, wie hoch ist die Wahrscheinlichkeit, dass 2 zufällig gezogene Fraud-Fälle vom Lerndatensatz abhängig sind? Abhängig sind sie, wenn ein Partner im Lerndatensatz verbleibt.



Nach dem ersten Mal Ziehen verbleiben 19 Datenpunkte.



Mit einer Wahrscheinlichkeit von 1/19 wird beim zweiten Mal Ziehen der Partner gezogen. Kein gezogener Fraud-Fall hat nun einen verbleibenden Partner im Lerndatensatz



Mit einer Wahrscheinlichkeit von 18/19 wird nicht der Partner gezogen. Beide gezogenen Datenpunkte haben nun einen Partner im Lerndatensatz

Folglich sind mit einer Wahrscheinlichkeit von 18/19 ≈ 0.95 beide Datenpunkte im hold-out set abhängig vom Lerndatensatz.

Wie funktioniert Kreuzvalidierung?

Im Folgenden wurde für den Fraud-Datensatz simuliert, wie hoch die Wahrscheinlichkeit ist, dass ein Fraud-Fall im hold-out set der Kreuzvalidierung einen entsprechenden Datenpunkt im Lerndatensatz hat. Die Basis hierfür bilden die gefundenen Fraud-Serien. Mit abnehmender Anzahl von Fraud-Fällen im hold-out set steigt diese Wahrscheinlichkeit.

25-fache Kreuzvalidierung

Anlernen | klassifiziere n

Lerndatensatz:

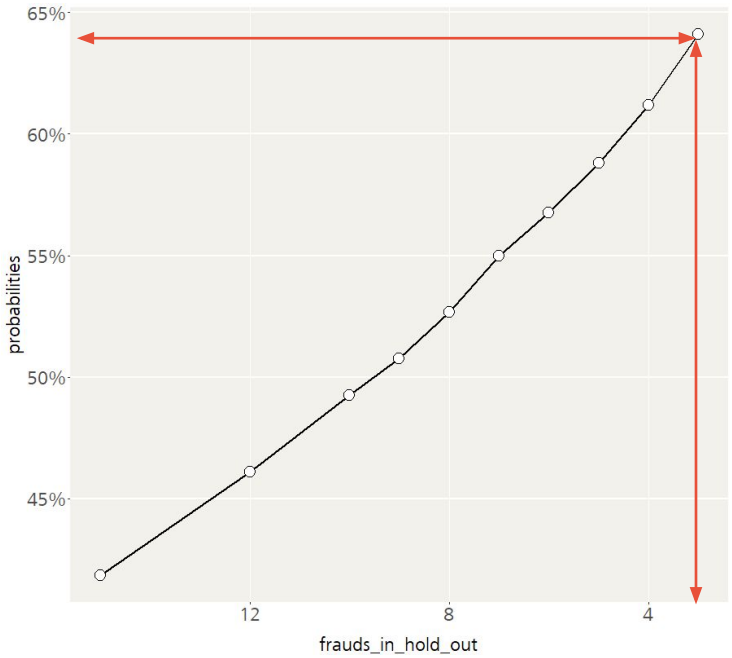
17’280’000	non-fraud
70	fraud

hold-out set:

720’000	non-fraud
3	fraud

Durchschnittlich über 60% der Fraud-Fälle im hold-out set sind abhängig.

prozentualer Anteil abhängiger Daten im hold-out set



3

Was passiert nun wenn die Daten aus dem Lerndatensatz und dem hold-out set nicht unabhängig sind? Zum Beispiel weil sie aus der selben Attacke (fraud) stammen?

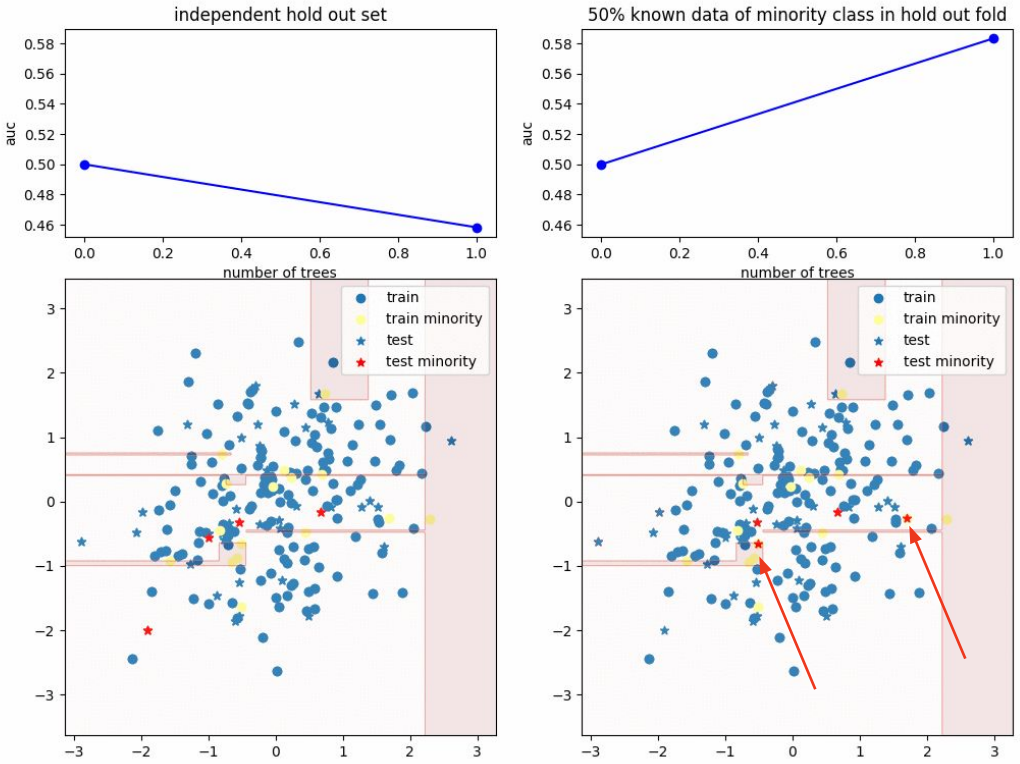
- auswendig lernen des Lerndatensatzes führt auch zu besseren Ergebnissen im hold-out set
- Parameter werden falsch geschätzt (Anzahl der trees im xgboost wird zu hoch geschätzt, genauso wie depth)

⇒ Overfitting (see for example: create good validation sets)

Veranschaulichung

Eine scheinbar kleine Überschneidung von Datenpunkten zwischen dem Lern-Set und dem hold-out-Set führt zu falschen Schätzungen der optimalen Parameter. Die optimale Anzahl von trees für einen xgboost classifier würde massiv überschätzt werden. Die Konsequenz ist ein ebenso massives overfitting.

Simulation mit 0% (links) bzw. 50% (rechts) Überschneidung der seltenen Fälle (fraud) zwischen Lerndatensatz und hold-out set



Der Betrag an overfitting ergibt sich aus der Differenz zwischen dem besten Wert für das unabh. Datenset (links) und dem besten Wert für den Fall von Abhängigkeit (rechts).

Die Datensets unterscheiden sich nur anhand von 2 Punkten. Im rechten Datenset sind 2 Punkte sowohl im Lernset (gelber Punkt) und im hold-out set (roter Stern). Im linken Datenset gibt es keine Überschneidung zwischen Lernset und hold-out set. Beide Klassen entstammen der selben Grundgesammheit, d.h. es gibt nichts zu lernen. **Der Erwartungswert ist auc = 0.5.**

Der random forest classifier schneidet sukzessive Flächen mit den gelben Punkten aus der Punktewolke

Zusammenfassung

- Highly unbalanced classes
- Data-Leakage durch falsches Datenmodell (BOT vs. SOT)
- Data-Leakage ‘MEDIUM’
- Abhängige Daten in minority class

18’164’183 vs **73**; ratio: **0.000004**

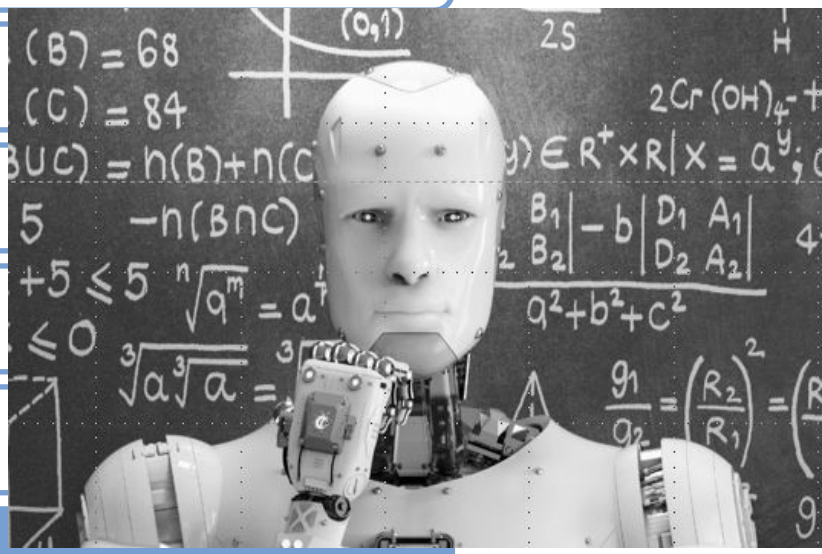
Information durch zeitl. spätere Ereigniss fliesst ein

52 fraud-Fälle haben eineindeutigen Wert

- Kreuzvalidierung unter Missachtung der Abhängigkeitsstruktur führt zu massivem overfitting.
- 25-fache Kreuzvalidierung bedingt, dass ca. 60% der Fraud-Fälle im hold-out set abhängig sind.
- xgboost mit n_trees = 2000 und max_depth=6 bedeutet 2000 Klassifikationsbäumen mit minimal 6 und maximal 63 splits (Verzweigungen) pro Baum:
  - Bei n_trees = 2000, durchschnittlich 20 splits sind das 40’000 splits auf 60 Variablen um weniger als 63 (nur Dubletten entfernt) unabh. Datenpunkte zu klassifizieren.
  - Das sind 635 Variablen-Splits pro Fraud-Fall
  - oder 666 splits pro Variable insgesamt.
  - Für jeden Fraud-Fall wird jede der 60 Variablen durchschnittlich 10 mal in der Klassifikation verzweigt

E-Banking Fraud-Detection

1. Datengrundlage
2. Datenmodell
3. Data-Leakage
4. Abhängigkeit der Daten
5. Kreuzvalidierung mit abhängigen Daten
6. Passende Metrik: AUC?



AUC: was soll optimiert werden?

Die verwendete Metrik, AUC (area under the curve) ist unter Umständen nicht optimal für ungleichgrosse Klassen. Traditionell wird sie auf der ROC – Metrik (true postive rate, false negative rate) berechnet; Besser wäre es die precision-recall curve zu verwenden; Diese Möglichkeit ist in xgboost nicht gegeben.

	positive	**negative**
predicted positive	TP	FP
predicted negative	FN	TN

$$\text{recall} = \text{true positive rate} = \frac{TP}{TP+FN}$$

$$\text{precision} = \frac{TP}{TP+FP}$$

$$\text{false positive rate} = \frac{FP}{FP+TN}$$

Beispiel:

	predicted fraud	real fraud
Algorithmus 1	2000	50
Algorithmus 2	200	50

	true postive rate	false positive rate
Algorithmus 1	0.6849315	0.0001073543
Algorithmus 2	0.6849315	0.00000825802
		0.000099

	recall	precision
Algorithmus 1	0.6849315	0.025
Algorithmus 2	0.6849315	0.25
		0.225