Insurance fraud analytics

Knowing me, knowing you: social networks in insurance

Katrien Antonio
LRisk - KU Leuven and ASE - University of Amsterdam

November 27, 2020

Fraud in insurance

Verzekeringsfraude: het opzettelijk misleiden van een verzekeraar bij de totstandkoming en/of uitvoering van een verzekeringsovereenkomst met de bedoeling om onrechtmatig verzekeringsdekking, -uitkering, -prestatie of dienstverlening te krijgen.

Source: Centrum Bestrijding Verzekeringscriminaliteit.

Some examples:

- staged accidents
- fake insurance claims
- ...
- exaggerated claims
- false declarations
- ...

Fraud in insurance

The Netherlands

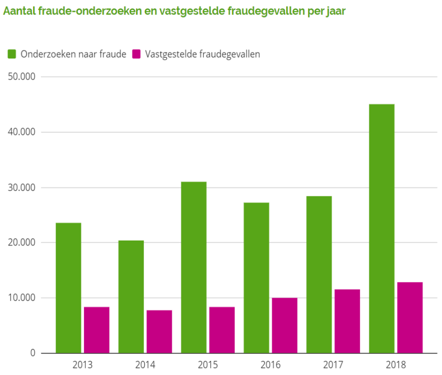
Financial consequences, according to Centrum Bestrijding Verzekeringscriminaliteit:

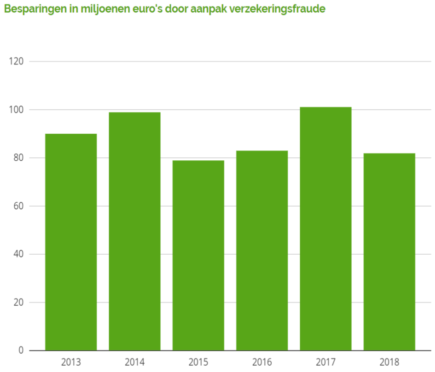
- in 2017: 11 540 confirmed fraudulent cases (on a total of 28 435 investigated cases) for a total of 101M euro
- in 2018: 12 879 confirmed fraudulent cases (on a total of 44 810 investigated cases) for a total of 82M euro
- in 2019: 22 376 confirmed fraudulent cases (on a total of 51 839 investigated cases) for a total of 96M euro.

Source: CBV factsheet September 2018, CBV factsheet October 2019 and CBV factsheet October 2020.

Fraud in insurance

The Netherlands





Source: Verbond van Verzekeraars, October 24, 2019.

Fraud detection cycle

deVolkskrant

ECONOMIE VERZEKERINGSFRAUDE

Waarom het moeilijker wordt om de verzekering te tillen

Source: De Volkskrant, January 29, 2019 of hier.

Fraud detection cycle

- Claim is flagged because of suspicion of fraud:
  - via expert knowledge, business rules
  - via analytical models.
- Fraud inspectors investigate the claim: (cfr. Gedragscode Persoonlijk Onderzoek)
  - confirm fraud or non-fraud.
- Insights used to flag new suspicious claims (Warren & Schweitzer, 2018).

Source illustration: https://www.mikanassociates.com/risk-analytics/.

Fraud detection cycle

Challenges

Baesens et al. (2015) and Van Vlasselaer et al. (2017):

developing fraud detection strategies is challenging, because fraud is:

I. Uncommon
II. Well considered
III. Time evolving
IV. Carefully organised
V. Imperceptibly concealed.

Fraud detection cycle

Challenges

> Verzekeringsfraudeurs komen voor in alle soorten en maten. Zij frauderen met alle denkbare verzekeringsproducten. Het kan gaan om een opportunistische debutant op het verkeerde pad maar ook om een doorgewinterde misdadiger die, geregeld in georganiseerd verband, de verzekeraar probeert op te lichten.
>
> Source: Fraudeurs gevangen in facts en figures, CBV.

(Insurance) Fraud detection

Literature review

- Business rules.
- Model using intrinsic (i.e. local) features (see e.g. Brockett et al., 2002; Artís et al., 2002).
- Model using network-based features (see e.g. Šubelj et al., 2011).
- Combined model GOTCHA! to detect fraud in social security (see e.g. Van Vlasselaer et al., 2017).
- Use of unstructured data (e.g. pictures and their meta data, text).

Insurance fraud detection
EIOPA report

- Thematic review by EIOPA (May, 2019)

  Big data analytics in motor and health insurance: a thematic review

  is an interesting starting point.

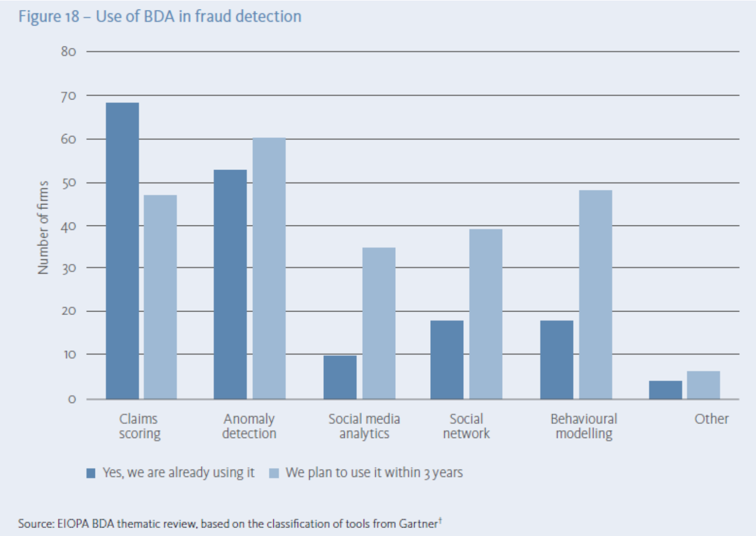


Figure 18 – Use of BDA in fraud detection
80
70
60
50
40
30
20
10
0
Number of firms
Claims scoring
Anomaly detection
Social media analytics
Social network
Behavioural modelling
Other
Yes, we are already using it
We plan to use it within 3 years
Source: EIOPA BDA thematic review, based on the classification of tools from Gartner†

Insurance fraud detection

EIOPA report

USE OF BDA TO PREVENT FRAUD

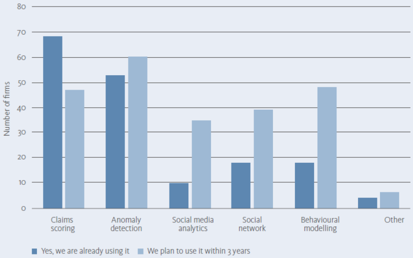
As shown in Figure 17, in claims management BDA is most often used to prevent fraud. Insurance fraud, i.e. intentionally bringing about an insurance event or causing the misconception of the occurrence of an insured event with the intention to receive insurance indemnity from the insurance firm, is a crime typified by the national law of the different Member States. According to Insurance Europe, the European insurance trade association, it is estimated to account for approximately 10% of all consumer claims.*

The expenses incurred by insurance firms in investigating and processing claims are known as loss adjustment expenses. Some insurance firms have special dedicated anti-fraud investigation units, often composed by personnel with a legal background as well as former police officers. In case of signs of consumer fraud, enhanced assessments are performed, which can include the use of private detectives. Insurance firms also commonly collaborate, creating claims and fraud databases within their respective national trade associations or in collaboration with public authorities.

Traditionally, there are two key stages in fraud-prevention: the first stage is prior to the conclusion of the contract; during the quotation process where insurance firms review the information provided by the consumer and cross-check it with internal and external sources of information such as fraud and claims databases or credit references. During the second phase, when processing claims, insurance firms' due diligence includes reviewing the documentation and evidence provided by the consumer to proof the loss and ensure the damages claimed by the consumer are accurate.**

BDA can support the detection of fraudulent claims in different ways. Most insurance firms have claims scoring tools, using ML algorithms in models trained to look for fraud patterns based on hundreds of different attributes (e.g. incident location, contract premium, number of previous claims by the policyholder etc.) and provide a fraud score for each claim. Often in combination with claims scoring techniques, insurance firms also use rule-based algorithms to assess claims, for instance by scanning invoices or images to automatically evaluate if the prices and damages are within the range of predefined/historical values or if they present anomalies. By flagging potentially fraudulent claims, investigators can focus on claims that are likely to be fraudulent and reduce the number of false positives and false negatives.

Figure 18 – Use of BDA in fraud detection



Source: EIOPA BDA thematic review, based on the classification of tools from Gartner†

Social media analytics, social network analytics and behavioural modelling are used less often amongst insurance firms. In this regard one firm stated that it assesses social media to analyse trends, although it does not really use BDA on it. Another firm described the generation of network diagrams in motor insurance, which are reviewed by fraud handlers alongside normal fraud referral processes, in order to help disclose hidden links between claims. Another firm stated that behavioural modelling is central to their health programme; it analyses different characteristics of health using BDA in order to best assess which behaviours best influence the overall health outcome.

* Insurance Europe, http://www.insuranceeurope.eu/fraud

** See EIOPA's fifth consumer trends report: https://europa.eu/!jc69pn

† Classification of types of BDA tools to prevent fraud is based on Gartner's analysis; Market Guide for Insurance Fraud Analytics, Gartner, 2016, https://www.gartner.com/doc/3241821/market-guide-insurance-fraud-analytics

Social network analytics for supervised fraud detection in insurance

Óskarsdóttir, Ahmed, Antonio, Baesens, Dendievel, Donas & Reynkens, 2020 (R&R).

Research goals

- Build an insurance fraud detection model
  - use 'classic' (or: intrinsic, local) features
  - use network data and extract useful features from network (new!)
  - use information from multiple claim types or LoBs: car, liability, fire, etc. (holistic view)
  - apply supervised learning (for now).
- Flag suspicious claims for further investigation.
- Find the working paper here.

The data

- A data set with over two million claims, over a period of six years.
- Each claim has a 'target variable': fraud, non-fraud or unknown.
- Focus in supervised learning on <u>motor insurance cover</u>, with:

    intrinsic (or local) features of claim and policyholder

    claimed amount, was police called?, at fault?, claim history of policyholder, ...

The data

- Resources are limited and fraud inspectors have limited time.
- Only a small fraction of all claims investigated.
- Per year, only 0.2% of all claims go through a fraud investigation with less than half resulting in a known fraud label (Challenge I).

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
Non-fraudulent	0.11%	0.10%	0.10%	0.11%	0.11%	0.11%
Fraudulent	0.07%	0.06%	0.08%	0.07%	0.09%	0.06%
Unknown	99.8%	99.8%	99.8%	99.8%	99.8%	99.8%

The data

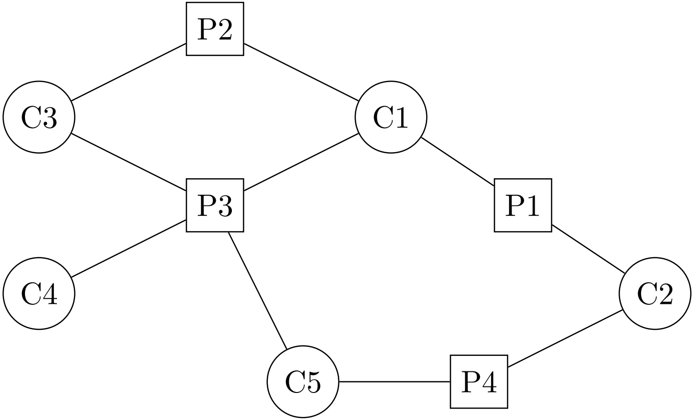
- Take a holistic view:
  - use claims across all available LoBs
  - use parties involved in a claim: policyholders, brokers, experts and garages.
- Go beyond traditional (flat, rectangular) data with target + set of intrinsic features.

The data

Network data

The data

Simplified network data



A sample network (also called: social network) with five claims and four parties.

The data

Simplified network data

- $G = (C \cup P, E)$ a bipartite network of nodes $C \cup P$ and edges E.
- Each edge in E connects one node in C to one node in P.
- The network's edges carry weights to indicate the strength of the connection:

$$\boldsymbol{W} = (w_{ij}),\ \text{where } i \in \{1, \ldots, n_C\}, j \in \{1, \ldots, n_P\},$$

  with n_C rows and n_P columns, the nodes in C and P.
- The network is undirected, with $w_{ij} = w_{ji}, \forall\, i, j \in n_C \cup n_P$.

The data

Simplified network data

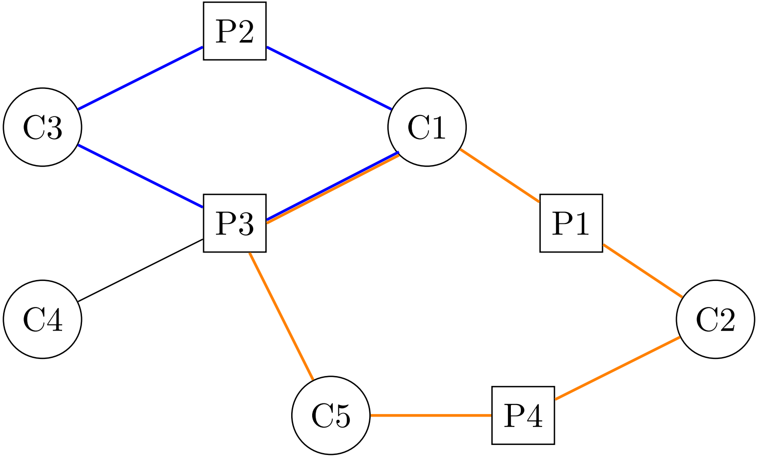
- The *k-th order neighborhood* of a node c_i or p_j,

$$\mathcal{N}_{c_i}^k \text{ or } \mathcal{N}_{p_j}^k,$$

  the set of all nodes that are connected to c_i, via a path of exactly k edges.

- The degree of a node, denoted with d_i for claim c_i or d_j for party p_j, is
  - the number of nodes in the first order neighborhood for an un-weighted network
  - the sum of weights on the edges between the node and the nodes in the first order neighborhood for a weighted network.

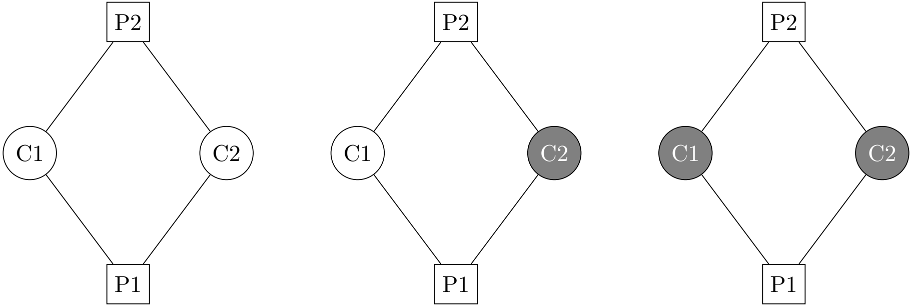
Graph cycles



A diamond (in blue, 4-cycle) and a triangle (in orange, 6-cycle).

Graph cycles

Diamonds

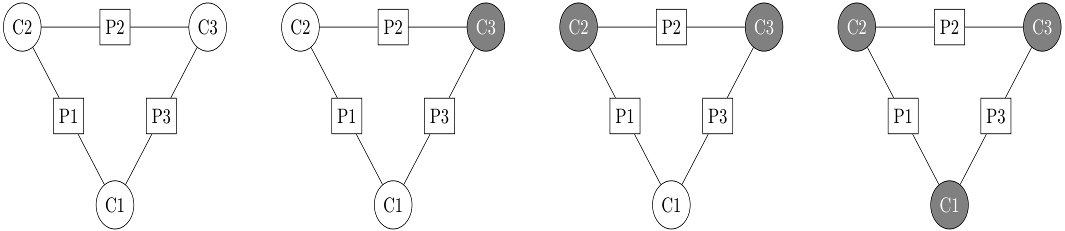


Diamonds (4-cycles) with zero (left), one (middle) or two (right) fraudulent claims.

Fraudulent claims are colored dark gray and non-fraudulent claims are white.

Graph cycles

Triangles



Triangles (6-cycles) with zero, one, two and three fraudulent claims (from left to right).

Fraudulent claims are colored dark gray and non-fraudulent claims are white.

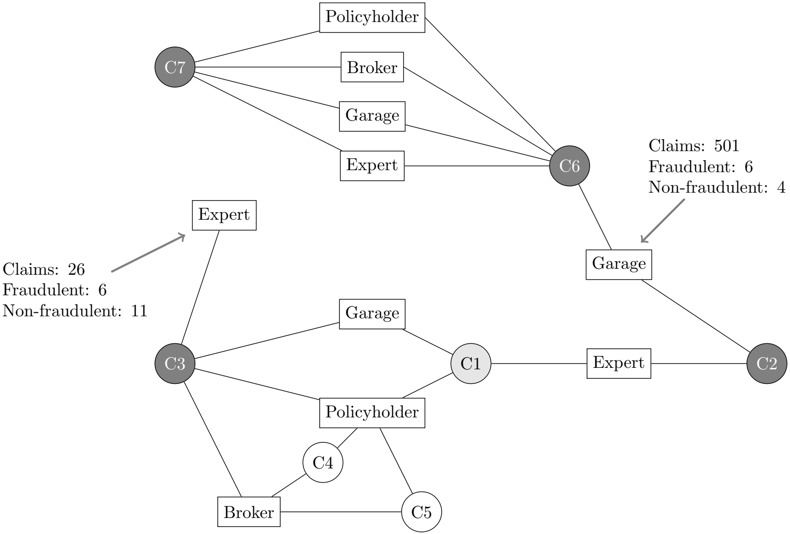
Graph cycles

Empirical findings

- As an analyst you can try to find empirical evidence of structural similarities in the network (homophily): (Challenge IV)
  - among fraudulent claims
  - among non-fraudulent claims.
- Neo4j offers a nice visual exploration of (parts of) the network.
- However, the network is too complex for manual inspection and detection (Challenge V).

The data

Complexity of the network



A supervised learning model for fraud detection

We develop an analytical model for flagging suspicious claims:

1. rank the claims with respect to their exposure to known fraudulent claims (cfr. homophily)

   (BiRank, a personalized PageRank for bipartite networks)

2. extract features from the network and combine with intrinsic features

   (network featurization)

3. use both in a predictive, supervised model to flag the most suspicious claims.

   (Random Forests and logistic regression)

A supervised learning model for fraud detection

We develop an analytical model for flagging suspicious claims:

1. rank the claims with respect to their exposure to known fraudulent claims (cfr. homophily)

   (BiRank, a personalized PageRank for bipartite networks) paus

2. extract features from the network and combine with intrinsic features

   (network featurization)

3. use both in a predictive, supervised model to flag the most suspicious claims.

   (Random Forests and logistic regression)

Ranking algorithm

PageRank

- This algo assigns a PageRank (score, or a measure of importance) to a webpage, invented by Larry Page and Sergei Brin (in 1999), founders of Google.
- A webpage is part of a large network where the nodes (webpages) of the network are linked together by hyperlinks.
- PageRank pictures a random surfer moving through the web:
  - (i) visit a linking webpage at random (with probability d)
  - (ii) pick a next, not necessarily linked, website at random (with probability $1 - d$).
- The PageRank is the long-run fraction of time spent at a webpage.

Ranking algorithm

PageRank

The algorithm assigns the score to each webpage i, based on: (circular first idea)

- linking webpages (j to i)
- do not just count, but weight

  webpages that link to i and have high PageRank scores themselves get more weight

  webpages that link to i but also to many other webpages should be given less weight.

Ranking algorithm

PageRank

- The PageRank (Page & Brin, 1999) of website x, $\text{PR}(x)$:

$$\text{PR}(x) \quad = \quad \frac{1-d}{N} + d \cdot \sum_{y \to x} \frac{\text{PR}(y)}{L(y)},$$

  where d is the damping factor (~ 0.85), the probability a surfer's random walk visits x from a connecting webpage y

  - $\text{PR}(y)$ the PageRank of website y, and
  - $\frac{1}{L(y)}$ the probability he opens the link from y to x, with $L(y)$ the number of outgoing links of webpage y.

Ranking algorithm

PageRank

- The PageRank (Page & Brin, 1999) of website x, $\text{PR}(x)$:

$$\text{PR}(x) \quad = \quad \frac{1-d}{N} + d \cdot \sum_{y \to x} \frac{\text{PR}(y)}{L(y)},$$

where with probability $1-d$ the surfer's random walk picks x at random, from the N available pages, then

$$(1-d)\sum_{y} \frac{\text{PR}(y)}{N} = \frac{1-d}{N},$$

because the $\text{PR}(y)$ over all webpages y define a probability distribution.

Ranking algorithm

PageRank

$$
\begin{aligned}
R &= \frac{1-d}{N}\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + d \cdot \begin{bmatrix} m(x_1,x_1) & \dots & m(x_1,x_N) \\ \vdots & \ddots & \\ m(x_N,x_1) & \dots & m(x_N,x_N) \end{bmatrix} \cdot R \\
&= \frac{1-d}{N} \cdot \mathbf{1} + d \cdot M \cdot R
\end{aligned}
$$

where (mind modification for dangling nodes!)

$$
m(x_i,x_j) = \begin{cases} \frac{1}{L(x_j)} & \text{if there exists a link from } x_j \text{ to } x_i \\ 0 & \text{a link from } x_j \text{ to another } x_k \text{ but not to } x_i \\ \frac{1}{N} & \text{no link from } x_j. \end{cases},
$$

and R the vector of PageRank scores.

Ranking algorithm

PageRank - example, see tutorial in the R Markdown on the workshop homepage

Let's put matrix M together:

- node A has three outgoing links, hence, $L(A) = 3$

  $m(., A) = \frac{1}{3}$ if a link exists from . to A

- node C has two outgoing links, $L(C) = 2$ and $m(., C) = \frac{1}{2}$ if a link exists from . to C

- same for node D

- node B has no outgoing links (dangling node), then $m(., B) = \frac{1}{4}$.

Ranking algorithm

PageRank - example

Let's put matrix M together:

$$M = \begin{pmatrix} 0 & 0.25 & 0.5 & 0.5 \\ 1/3 & 0.25 & 0 & 0 \\ 1/3 & 0.25 & 0 & 0.5 \\ 1/3 & 0.25 & 0.5 & 0 \end{pmatrix}.$$

Node B has only one incoming link (not so important).

Nodes A, C and D have two incoming links, but the links going from A are spread among 3 nodes (B, C and D) (thus?).

Ranking algorithm

PageRank as a Markov chain

- Then,

$$\begin{aligned} R &= \frac{1-d}{N} \cdot 1 + d \cdot M \cdot R \\ &= G \cdot R, \end{aligned}$$

  where $G = \left(\frac{1-d}{N} \cdot E + d \cdot M\right) \cdot R$, and E the matrix of 1s.

- The entries of R, the PageRanks, define a probability distribution.
- G is the (Google) transition matrix of a Markov chain.
- Find R, the unique stationary distribution, called PageRank, to which the chain converges.

Ranking algorithm

PageRank algebraic

For time step $k \to \infty$ we could say that

$$R \approx (I - d \cdot M)^{-1} \cdot \frac{1-d}{N} \cdot 1,$$

where I is the $N \times N$ identity matrix.

Ranking algorithm

PageRank iterative

With an iterative strategy:

- at time $k = 0$ initialize a probability distribution, $PR(x) = \frac{1}{N}$,
- iterate

$$R_k = \frac{1-d}{N} \cdot 1 + d \cdot M \cdot R_{k-1}.$$

- the iteration ends when for a sufficiently small ε and a large k it holds that $|R_k - R_{k-1}| < \varepsilon$.

Ranking algorithm

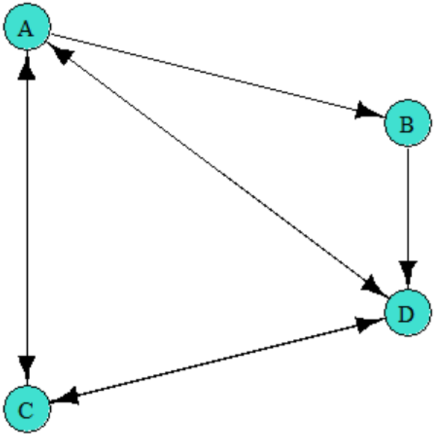
PageRank - example revisited

Calculating the PageRank scores for this example:

- PR(A) $= 0.3012950$
- PR(B) $= 0.1560215$
- PR(C) $= 0.2713417$
- PR(D) $= 0.2713417$

Ranking algorithm

PageRank - Your Turn!



How would you rank the nodes in this graph (from small to large) using the PageRank

Ranking algorithm

Personalized PageRank

- Simple PageRank algorithm gives each node an equal probability to be chosen by the random surfer.
- Personalized PageRank brings out nodes in a network that are most central from the perspective of a set of specific source nodes.
- Personalize the ranks of nodes in a network towards these source nodes (e.g. fraudsters).
- The random surfer jumps to nodes that belong to the set of specific source nodes, with probabilities stored in a teleportation vector.

Ranking algorithm

Personalized PageRank

In matrix notation:

$$
\begin{aligned}
R &= (1-d)\cdot\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix} + d\cdot\begin{bmatrix} m(x_1,x_1) & \dots & m(x_1,x_N) \\ \vdots & \ddots & \\ m(x_N,x_1) & \dots & m(x_N,x_N) \end{bmatrix}\cdot R \\
&= (1-d)\cdot V + d\cdot M\cdot R,
\end{aligned}
$$

where V is the teleportation vector.

Ranking algorithm

Personalized PageRank

With the matrix M defined as

$$m(x_i, x_j) = \begin{cases} \frac{1}{L(x_j)} & \text{if link from } x_j \text{ to } x_i \\ 0 & \text{if link from } x_j \text{ to another } x_k \text{ but not to } x_i \\ v_i & \text{if no link from } x_j. \end{cases}$$

In a network of N claims with F known fraudulent claims, the elements of the teleportation vector V are (e.g.)

$$v_i = \begin{cases} \frac{1}{F} & \text{if } x_i \text{ is fraudulent} \\ 0 & \text{if } x_i \text{ is not fraudulent.} \end{cases}$$

Ranking algorithm

BiRank

- Many variants of (personalized) PageRank exist.
- BiRank (He et al., 2017) is a personalized PageRank algorithm specifically designed for biparite networks, where nodes of the same type cannot be connected.
- Now return to this bipartite network $G = C \cup P$, with edges E and corresponding weights in W.

Ranking algorithm

Time weighting on edges and nodes

- We investigate time weighting in the BiRank: (Challenge III)
  - on edges

$$w_{i,j} = \begin{cases} \exp(-\gamma h_i) & \text{if relationship between claim } i \text{ and party } j \\ 0 & \text{otherwise.} \end{cases}$$

    with h_i the time since claim and γ the decay constant,

  - on fraud restart vector

$$v_i = \begin{cases} \exp(-\beta h_i) & \text{if node } i \text{ is a claim and fraudulent} \\ 0 & \text{otherwise,} \end{cases}$$

    with β the decay constant.

A supervised learning model for fraud detection

We develop an analytical model for flagging suspicious claims:

1. rank the claims with respect to their exposure to known fraudulent claims (cfr. homophily)

   (BiRank, a personalized PageRank for bipartite networks)

2. extract features from the network and combine with intrinsic features

   (network featurization)

3. use both in a predictive, supervised model to flag the most suspicious claims.

   (Random Forests and logistic regression)

Network featurization

Score-based

Name	Order	Description
score	0	The node's fraud score
n1.q1	1	The first quartile of emp. distr. of fraud scores in the node's first order neighborhood
n1.med	1	The median
n1.max	1	The maximum
n2.q1	2	The first quartile of emp. distr. of fraud scores in the node's second order neighborhood
n2.med	2	The median
n2.max	2	The maximum

Mind multicollinearity issues!

Network featurization

Neighborhood-based

Name	Order	Description
n1.size	1	The number of nodes in node's first order neighborhood
n2.size	2	The number of nodes in node's second order neighborhood
n2.RatioFraud	2	The number of known fraudulent claims in node's second order neighborhood divided by n2.size
n2.RatioNonFraud	2	The number of known non-fraudulent claims in node's second order neighborhood divided by n2.size
n2.BinFraud	2	1 if there is a known fraudulent claim in node's second order neighborhood

Mind multicollinearity issues!

A supervised learning model for fraud detection

We develop an analytical model for flagging suspicious claims:

1. rank the claims with respect to their exposure to known fraudulent claims (cfr. homophily)

   (BiRank, a personalized PageRank for bipartite networks)

2. extract features from the network and combine with intrinsic features

   (network featurization)

3. use both in a predictive, supervised model to flag the most suspicious claims.

   (Random Forests and logistic regression)

Supervised learning model for fraud

- Available features:

$$x^{\texttt{intr}},\ x^{\texttt{score}} \text{ and } x^{\texttt{nbh}}.$$

- Target variable:

$$y_i^{\texttt{known}} = \begin{cases} 1 & \Leftrightarrow l_i \in \{\text{fraud, non-fraud}\} \\ 0 & \Leftrightarrow l_i \in \{\text{unknown}\}, \end{cases},$$

or

$$y_i^{\texttt{fraud}} = \begin{cases} 1 & \Leftrightarrow l_i \in \{\text{fraud}\} \\ 0 & \Leftrightarrow l_i \in \{\text{non-fraud, unknown}\}. \end{cases},$$

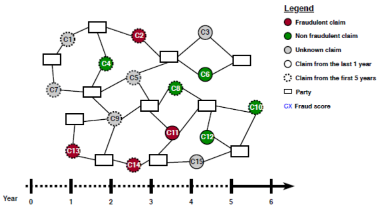
where l_i is the original label (or target) of claim i.

Supervised learning model for fraud

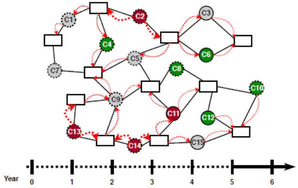
- We use logistic regression to predict fraud (simple, to get started).
- BUT: features used are pre-selected via random forests and variable importance plots.
- We evaluate model performance out-of-time via:
  - AUROC
  - precision-recall
  - top-decile lift: how does incidence in the 10% claims with the highest model predictions compare to the overall incidence?

Supervised learning model for fraud

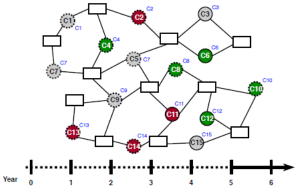
Model evaluation



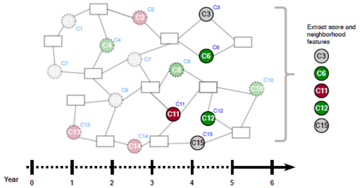
(a) The whole network



(b) Application of the Birank alogrithm



(c) Nodes with fraud scores



(d) The most recent nodes

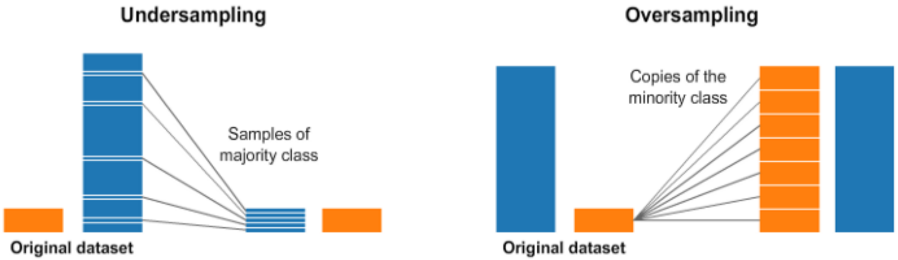
Supervised learning model for fraud

Model building

- Our focus is on the claims filed in the last observed historical year.
- Split these into
  - training (70%) set
  - test (30%) set.
- Both have a high class imbalance (Challenge I), with 4.9% and 1.8% minority class rate in composed training and test sets $\mathcal{D}^{\texttt{known}}$ and $\mathcal{D}^{\texttt{fraud}}$ (see paper).

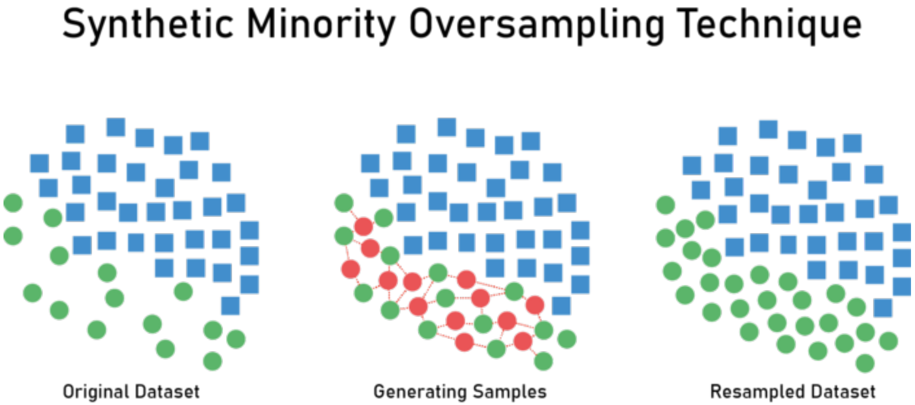
Supervised learning model for fraud

Resampling methods for imbalanced data



Picture taken from What to do you when your data set is imbalanced?

Supervised learning model for fraud



Picture taken from Resampling to Properly Handle Imbalanced Datasets in Machine Learning

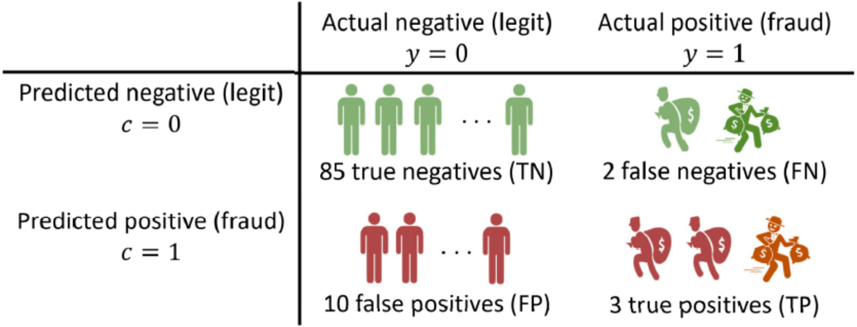
Supervised learning model for fraud

Model building

- Use SMOTE (Chawla et al., 2002) on the training data to create a better balanced training set $\Rightarrow$ increase to 15% of minority class.
- Use this newly sampled training dataset to evaluate the feature importance using random forests.
- Use ten-fold cross-validation to tune hyperparameters.
- Find per group of features the most important ones, separately for $\mathcal{D}^{\texttt{known}}$ and $\mathcal{D}^{\texttt{fraud}}$.

Supervised learning model for fraud

Performance measures - confusion matrix



Supervised learning model for fraud

Performance measures

	Actual negative (legit) $y = 0$	Actual positive (fraud) $y = 1$
Predicted negative (legit) $c = 0$	85 true negatives (TN)	2 false negatives (FN)
Predicted positive (fraud) $c = 1$	10 false positives (FP)	3 true positives (TP)

Supervised learning model for fraud

Performance measures

	Actual negative (legit) $y = 0$	Actual positive (fraud) $y = 1$
Predicted negative (legit) $c = 0$	 85 true negatives (TN)	 2 false negatives (FN)
Predicted positive (fraud) $c = 1$	 10 false positives (FP)	 3 true positives (TP)

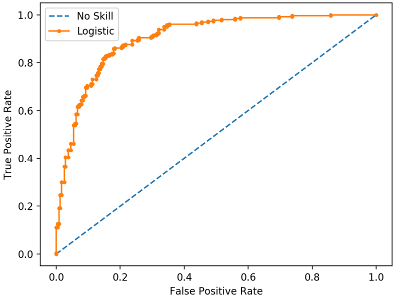
$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN} = \frac{85 + 3}{100} = 88\%$$

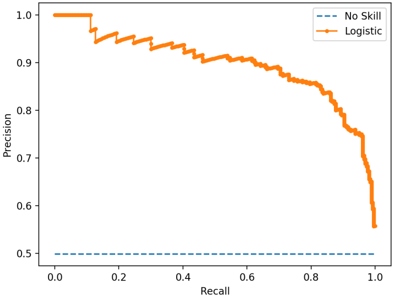
$$Precision = \frac{TP}{TP + FP} = \frac{3}{13} = 23\%$$

$$Recall = \frac{TP}{TP + FN} = \frac{3}{5} = 60\%$$

Supervised learning model for fraud

Performance curves

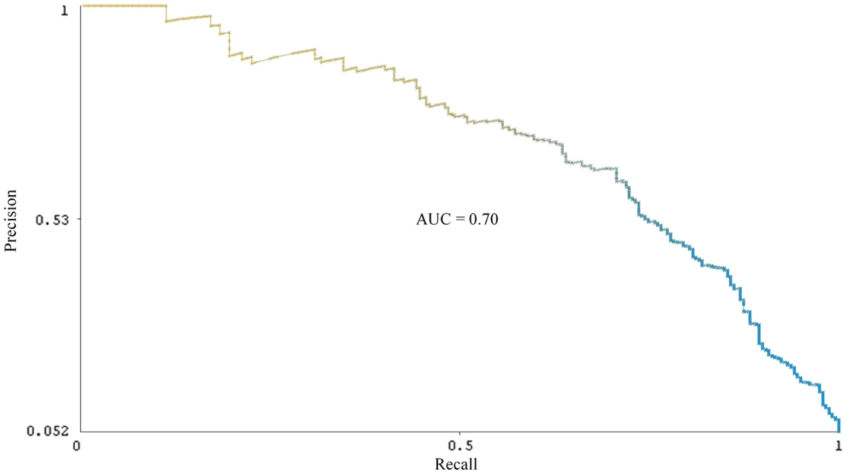




Supervised learning model for fraud

Performance curves

From: Towards scaling Twitter for digital epidemiology of birth defects



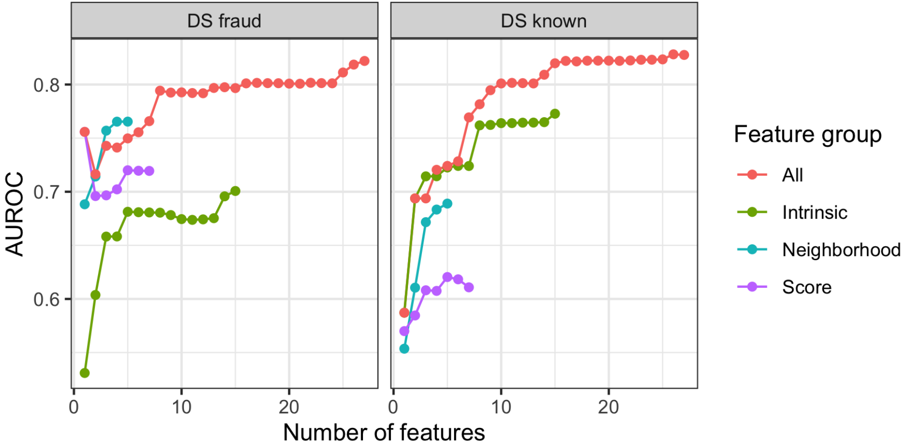
Supervised learning model for fraud

Performance curve measures

- AUROC:
  - between 0.5 (random) and 1 (perfect)
  - capability of model to separate 0/1.
- AUPR:
  - between actual incidence rate (random) and 1 (perfect)
  - more relevant with class-imbalanced data
  - capability of model to predict class 1.

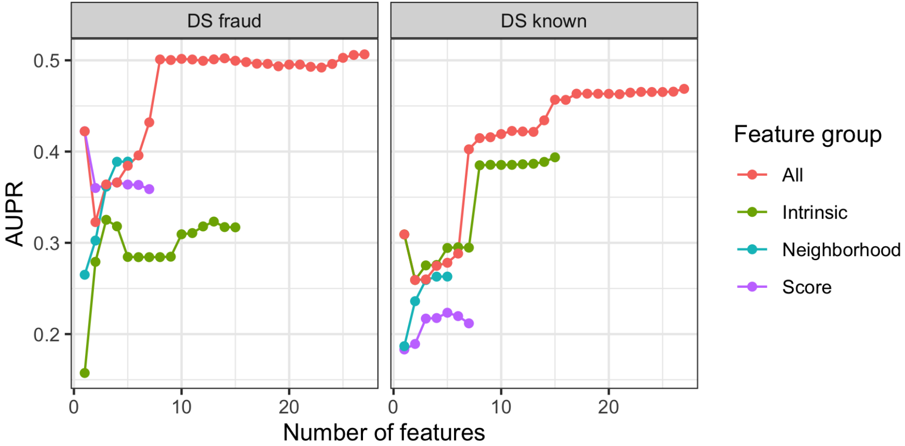
Supervised learning model for fraud

Model building - 10-fold CV, logistic regression



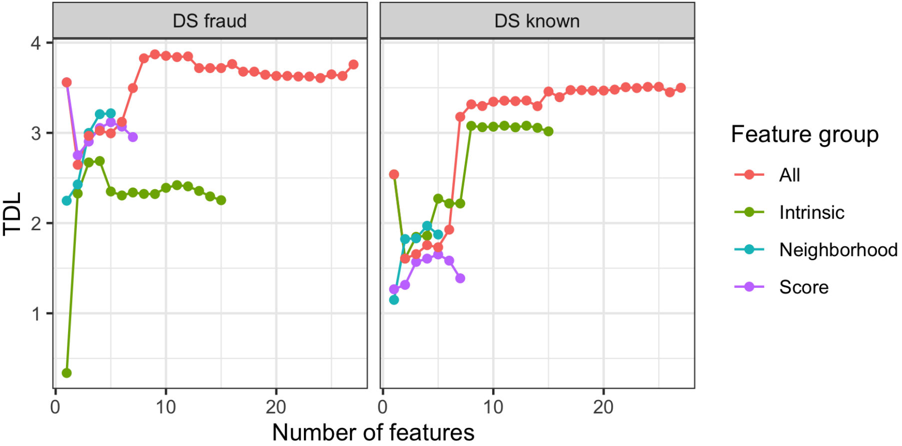
Supervised learning model for fraud

Model building - 10-fold CV, logistic regression



Supervised learning model for fraud

Model building - 10-fold CV, logistic regression



Supervised learning model for fraud

Test set predictions

	DS known			DS fraud		
Features	AUROC	AUPR	TDL	AUROC	AUPR	TDL
Intrinsic	0.691	0.1214	2.85	0.662	0.0301	2.137
Score	0.634	0.0883	2.25	0.660	0.0402	2.812
Neighborhood	0.681	0.1051	2.65	0.719	0.0481	3.262
All	0.725	0.1312	3.457	0.792	0.0810	3.824

Findings

- We leverage the insurance company's database of claims, policyholders, brokers, experts and garages to build a bipartite network.
- The classical intrinsic features are good at distinguishing claims with a known label.
- The combined set of features helps to detect fraudulent claims.
- No convincing evidence for improved performance with time-weighted edges and fraud.

Want to read more?

Working paper available from my website:

Social network analytics and supervised learning for insurance fraud detection

by María Óskarsdóttir, Waqas Ahmed, Katrien Antonio, Bart Baesens, Rémi Dendievel, Tom Donas & Tom Reynkens.