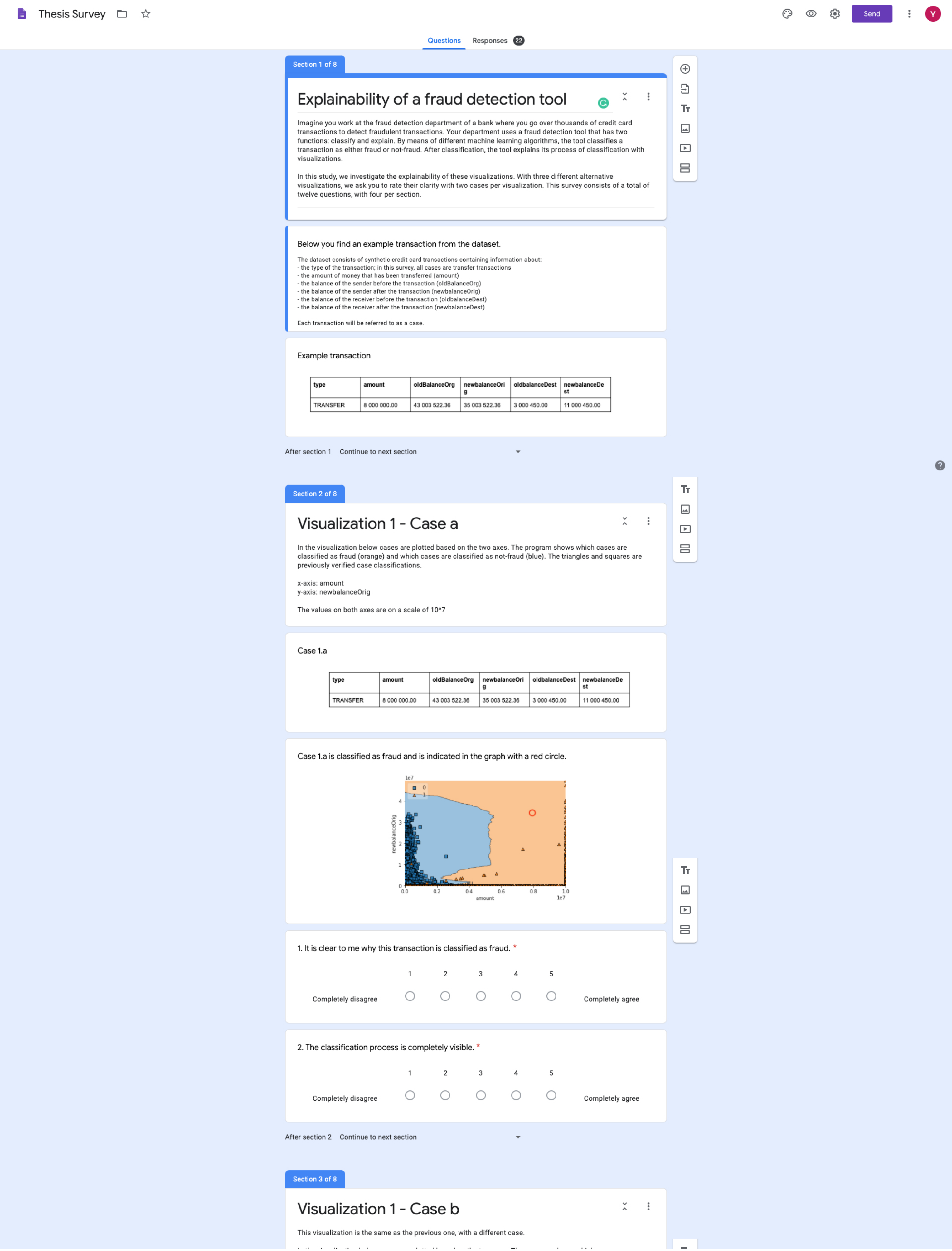


Explainability of a fraud detection tool

Imagine you work at the fraud detection department of a bank where you go over thousands of credit card transactions to detect fraudulent transactions. Your department uses a fraud detection tool that has two functions: classify and explain. By means of different machine learning algorithms, the tool classifies a transaction as either fraud or not-fraud. After classification, the tool explains its process of classification with visualizations.

In this study, we investigate the explainability of these visualizations. With three different alternative visualizations, we ask you to rate their clarity with two cases per visualization. This survey consists of a total of twelve questions, with four per section.

Below you find an example transaction from the dataset.

The dataset consists of synthetic credit card transactions containing information about:
- the type of the transaction; in this survey, all cases are transfer transactions
- the amount of money that has been transferred (amount)
- the balance of the sender before the transaction (oldBalanceOrg)
- the balance of the sender after the transaction (newbalanceOrig)
- the balance of the receiver before the transaction (oldbalanceDest)
- the balance of the receiver after the transaction (newbalanceDest)

Each transaction will be referred to as a case.

Example transaction

type	amount	oldBalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest
TRANSFER	8 000 000.00	43 003 522.36	35 003 522.36	3 000 450.00	11 000 450.00

After section 1 Continue to next section

Visualization 1 - Case a

In the visualization below cases are plotted based on the two axes. The program shows which cases are classified as fraud (orange) and which cases are classified as not-fraud (blue). The triangles and squares are previously verified case classifications.

x-axis: amount
y-axis: newbalanceOrig

The values on both axes are on a scale of 10^7

Case 1.a

type	amount	oldBalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest
TRANSFER	8 000 000.00	43 003 522.36	35 003 522.36	3 000 450.00	11 000 450.00

Case 1.a is classified as fraud and is indicated in the graph with a red circle.

1. It is clear to me why this transaction is classified as fraud. *

	1	2	3	4	5	
Completely disagree	○	○	○	○	○	Completely agree

2. The classification process is completely visible. *

	1	2	3	4	5	
Completely disagree	○	○	○	○	○	Completely agree

After section 2 Continue to next section

Visualization 1 - Case b

This visualization is the same as the previous one, with a different case.

In the visualization below cases are plotted based on the two axes. The program shows which cases are classified as fraud (orange) and which cases are classified as not-fraud (blue). The triangles and squares are previously verified case classifications.

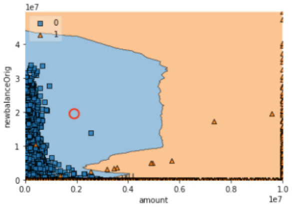
x-axis: amount
y-axis: newbalanceOrig

The values on both axes are on a scale of 10^7

Case 1.b

type	amount	oldBalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest
TRANSFER	2 000 000.00	22 005 070.03	20 005 070.03	355 600.00	2 355 600.00

Case 1.b is classified as not-fraud and is indicated in the graph with a red circle.



3. It is clear to me why this transaction is classified as not-fraud. *

	1	2	3	4	5	
Completely disagree	○	○	○	○	○	Completely agree

4. The classification process is completely visible. *

	1	2	3	4	5	
Completely disagree	○	○	○	○	○	Completely agree

After section 3 Continue to next section

Section 4 of 8

Visualization 2 - Case a

In the visualization below cases follow a path of questions in order to get classified, please zoom-in if the text is not readable. The program has a final class when there is a dead-end. Each final class is indicated with a colour: fraud is blue; not-fraud is orange (pay attention to the fact that the colours of the classes are not the same as in the previous visualization).

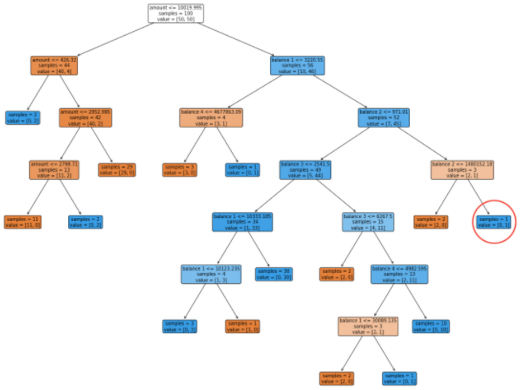
Yes: follow the left branch
No: follow the right branch

The colours of the waypoints can be ignored, only the colours of the endstations are relevant.

Case 2.a

type	amount	balance 1	balance 2	balance 3	balance 4
TRANSFER	8 000 000.00	43 003 522.36	35 003 522.36	3 000 450.00	11 000 450.00

Case 2.a is classified as fraud and its endstation is indicated in the tree with a red circle.


amount <= 10019.995 samples = 100 value = [50, 50]
amount <= 426.32 samples = 44 value = [40, 4]
balance 1 <= 3220.55 samples = 56 value = [10, 46]
samples = 2 value = [0, 2]
amount <= 2952.985 samples = 42 value = [40, 2]
balance 4 <= 4677863.09 samples = 4 value = [3, 1]
balance 2 <= 971.01 samples = 52 value = [7, 45]
amount <= 2798.71 samples = 13 value = [11, 2]
samples = 29 value = [29, 0]
samples = 3 value = [3, 0]
samples = 1 value = [0, 1]
balance 3 <= 2541.5 samples = 49 value = [5, 44]
balance 2 <= 1480152.18 samples = 3 value = [2, 1]
samples = 11 value = [11, 0]
samples = 2 value = [0, 2]
balance 1 <= 10333.185 samples = 34 value = [1, 33]
balance 3 <= 6267.5 samples = 15 value = [4, 11]
samples = 2 value = [2, 0]
samples = 1 value = [0, 1]
balance 1 <= 10123.235 samples = 4 value = [1, 3]
samples = 30 value = [0, 30]
samples = 2 value = [2, 0]
balance 4 <= 4982.595 samples = 13 value = [2, 11]
samples = 3 value = [0, 3]
samples = 1 value = [1, 0]
balance 1 <= 30089.135 samples = 3 value = [2, 1]
samples = 10 value = [0, 10]
samples = 2 value = [2, 0]
samples = 1 value = [0, 1]

5. It is clear to me why this transaction is classified as fraud. *

1 2 3 4 5

Completely disagree ○ ○ ○ ○ ○ Completely agree

6. The classification process is completely visible. *

	1	2	3	4	5	
Completely disagree	○	○	○	○	○	Completely agree

After section 4 Continue to next section

Section 5 of 8

Visualization 2 - Case b

This visualization is the same as the previous one, with a different case.

In the visualization below cases follow a path of questions in order to get classified, please zoom-in if the text is not readable. The program has a final class when there is a dead-end. Each final class is indicated with a colour: fraud is blue; not-fraud is orange (pay attention to the fact that the colours of the classes are not the same as in the previous visualization).

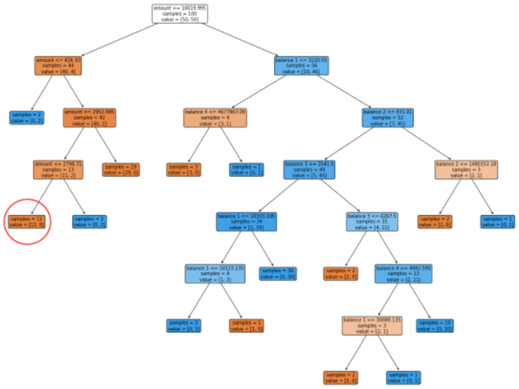
Yes: follow the left branch
No: follow the right branch

The colours of the waypoints can be ignored, only the colours of the endstations are relevant.

Case 2.b

type	amount	balance 1	balance 2	balance 3	balance 4
TRANSFER	2 500.00	17 356.77	19 856.77	390 037.43	392 537.43

Case 2.b is classified as not-fraud and its endstation is indicated in the tree with a red circle.



7. It is clear to me why this transaction is classified as not-fraud. *

	1	2	3	4	5	
Completely disagree	○	○	○	○	○	Completely agree

8. The classification process is completely visible. *

	1	2	3	4	5	
Completely disagree	○	○	○	○	○	Completely agree

After section 5 Continue to next section

Section 6 of 8

Visualization 3 - Case a

In the visualization below cases are plotted based on the two axes. The graph is compiled by multiple layers consisting of neurons that receive input from the data and process this with mathematical calculation into weights which are visualized through the boldness of the edges. The bolder the edge, the higher its weight.
The program shows which cases are classified as fraud (orange) and which cases are classified as not-fraud (blue) (pay attention to the fact that the colours of the classes are not the same as in the previous visualization).
The orange and blue data points are previously verified case classifications.

x-axis: amount
y-axis: oldBalanceOrig

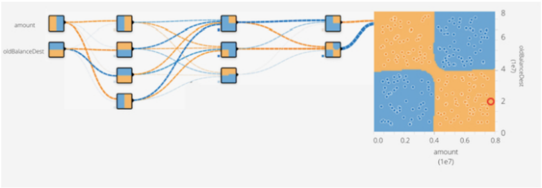
The values on both axes are on a scale of 10^7

Case 3.a is classified as fraud and is indicated in the graph with a red circle.

type	amount	oldBalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest

TRANSFER	7 500 000.00	8 000 039.98	500 039.98	20 000 572.00	27 500 572.00

Case A: This is fraud



9. It is clear to me why this transaction is classified as fraud. *

	1	2	3	4	5	
Completely disagree	○	○	○	○	○	Completely agree

10. The classification process is completely visible. *

	1	2	3	4	5	
Completely disagree	○	○	○	○	○	Completely agree

After section 6 Continue to next section

Section 7 of 8

Visualization 3 - Case b

This visualization is the same as the previous one, with a different case.

In the visualization below cases are plotted based on the two axes. The graph is compiled by multiple layers consisting of neurons that receive input from the data and process this with mathematical calculation into weights which are visualized through the boldness of the edges. The bolder the edge, the higher its weight. The program shows which cases are classified as fraud (orange) and which cases are classified as not-fraud (blue) (pay attention to the fact that the colours of the classes are not the same as in the previous visualization). The orange and blue data points are previously verified case classifications.

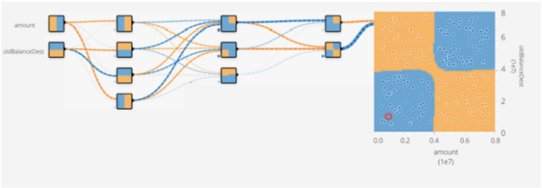
x-axis: amount
y-axis: oldBalanceOrig

The values on both axes are on a scale of 10^7

Case 3.b is classified as not-fraud and is indicated in the graph with a red circle.

type	amount	oldBalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest
TRANSFER	1 000 000.00	4 000 098.33	3 000 098.33	10 000 725.30	11 000 725.30

Case B: This is not fraud



11. It is clear to me why this transaction is classified as not-fraud. *

	1	2	3	4	5	
Completely disagree	○	○	○	○	○	Completely agree

12. The classification process is completely visible. *

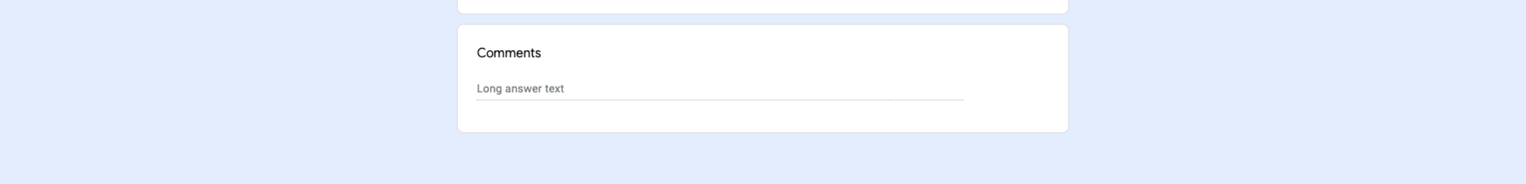
	1	2	3	4	5	
Completely disagree	○	○	○	○	○	Completely agree

After section 7 Continue to next section

Section 8 of 8

Additional comments

This final section is optional. If you have any additional comments regarding your answers or the explainability of the visualizations.



Comments

Long answer text