Outlier detection in pension asset data

DeNederlandscheBank

Iris Nonneman

05-03-2024

Goal

Automatically detect outliers in the line-by-line data of pension funds and insurers using machine learning techniques





Agenda

- Why detecting outliers
- Motivation
- Model approach
- From data to model
- Results and performance
- Use cases for actuaries





Detecting outliers?



Natural causes of outliers in data

- 1. Data error: wrong measurement data observation
- 2. Natural occurrence but different than expected



Problems caused by outliers

Outliers in the data influence model fitting (linear models)
 Outliers can inflate metrics which give higher weights to large errors (like RMSE).



Why detecting outliers?



<u>A light bulb surrounded by many black ones</u> <u>photo – Innovation Image on Unsplash</u>

DeNederlandscheBank

Outliers can be informative

1. Outliers that are data errors influence results

2. Outliers that are not errors (anomalies), are informative: show behavior that is different from "expected" / the bulk of data



Background project

DeNederlandscheBank

FTK data



Motivation for using ML algorithm

Characteristics asset data

Large amounts of data

Sensitive to errors

De**Nederlandsche**Bank

Rule based controls





Rule based controls





Rule based controls





Rule based controls are not enough





Motivation for using ML algorithm

Outlier detection model

- Automatically detect errors
- Speed up data cleaning process
- Support rule-based controls
- Next step 'anomaly' detection



De**Nederlandsche**Bank

Models

DeNederlandscheBank

EUROSYSTEEM

. •

Outlier detection

Ensemble learner

- 1. Make groups for the detection of outliers
- 2. Detect outliers with an ensemble of 3 or 3+ methods
- 3. Define a total score for each observation based on the different methods
- 4. Highest scoring observations are potential outliers





Ensemble methods



De**Nederlandsche**Bank

Interquantile range

$$x - (P_{80} - P_{20})/2$$

$$P_{80} - P_{20}$$
Distance from middle
$$P_{20} - P_{80} Range$$



Local outlier factor



Outlier detectie

Nearest neighbourgh distance



19

Kmeans

K-means identifies k number of centroids and allocates every data point to the nearest cluster keeping the centroids as small as possible.

Towards outlier score: Outliers are scored by calculating their **z-score**, which is defined as the observation value minus the centroid divided by the centroid's standard deviation.



XGBoost

Is a gradient-boosting decision tree algorithm. It trains a number of trees sequentially and uses the fit of the previous tree to improve next fit.

It combines all the trees to create the ultimate predicted value.

As predictors, we (can) use several explanatory variables, like sector, credit quality and valuta.





Scoring



Figure 1: An overview of how the different methods would score points on a onedimensional scale



Approach: data to model

DeNederlandscheBank









Data preprocessing

- Define 'waarde per eenheid'
 - Median of previous period
 - Percentage difference with previous period
- Create categorical variables
- Stock splits cleaning

relatienummer	
URI	
aantal	
nominaal_bedrag	
aankoopwaarde	
marktwaarde	
aangegroeide_rente	
land	
valuta	
CIC	
rating	
kredietkwalitei	
interne_rating	
prijs_per_eenheid	
percentage_nominaal_bedrag	
URI_cat	
sector_main	
sector_sub	
sector_sub_sub	
waarde_per_eenheid	
waarde_per_eenheid_dt	















Reinforcement learning

Reinforcement Learning (RL) is a ML technique that enables us to create a model that learns by trial and error through exposure with its environment.



Reinforcement learning

- Our data is not on forehand **classified**: we do not know whether a point is an outlier or not:
 - Do we deal with red or blue fish?







Reinforcement learning

- Our data is not on forehand **classified**: we do not know whether a point is an outlier or not:
- Thus, setting optimal **model coefficients** on forehand is difficult
- Let the model learn over time, when we know which outliers identified: using reinforcement learning







Figure 2: An overview of the reinforcement learning algorithm applied to the outlier detection



• The results are shared with the institutions, who file ammended reporting data





• The differences between the initial and amended reporting, and the outlier scores are recorded





• The differences between the initial and amended reporting, and the outlier scores are recorded





• The agent optimises the parameters of the outlier detection algorithm





• Parameters of the outlier detection are updated





Reinforcement Learning Assisted Outlier Detection for Costly to Verify Data

Michiel Nijhuis*

Data Science Hub, De Nederlandsche Bank, Amsterdam, 1000 AB, The Netherlands

Iman P.P. Van Lelyveld

Data Science Hub, De Nederlandsche Bank, Amsterdam, 1000 AB, The Netherlands Department of Finance, VU Amsterdam, Amsterdam, 1081 HV, The Netherlands

Abstract

Within the financial reporting outliers in data are often due to data quality issues. However, some data anomalies are real and are of interest. Often extreme data points can be verified and the ground truth for a data point can be established. With the increasing granularity of data, checking all the data points is time-consuming, moreover the underlying issues leading to



Want to know more?

Read the paper!

DNB PUBLIC 15.22 1960

.v

00

0

0

ŝ

.0.

20.00

40. .07

.00

69-69-

.v.

85C3 C51

40.00 60.00

12.22

11.11

00 40 VV

100.00

16.91

10.10

124

.cv

000

c112

1.000

16.02

50°.27

25.28

665

555 . 150 54 0 V

13.22

55.01

0,20

20.02

N1.26

60.62

17.17

es. v

29011

818.24

.1

091-20

140.28

-02-1-02-

980.01

206.26

60 60 7 7

0960

619 61.00

.20

01004

49.00

05.76

1. 1. V. 1.

6.67

66.69

115,48

50

10.0L

SCU

62.00

.0

220

5-5-50 - 100

0.

120

V2614

.2

25

192.52 VA

VIS VIS

24.69

.0

202.8

151-56

23.22

555 CSL

6.91

001.01

40V 28

2502

.c.

60.000.000 C

55.71

1000

V12-00

.

S.S.S.

.Su

90 . VS

909.52

5.00

53. C2.

100

43.00

100 EB

652.16

16922

50.52

5.15

.0

8199

10

15-50

S. S.

62.06

651 - 651

29.02

2002

25.25

120.01

515-18

910.92

10001

100:19

2960

000 000 000

12.08

STO 8421

296

542.08

So.

180.00

2

010.00

000

.1.

500

600

Results

con

192 6.00 0362 862

1200

695.24

550.050

.0 577

252.28

NY 28

012:08

605

69.99

210,71

200.001

851-14

60000

.10

2005

529.25

823

.20

00000

0200

.c.

.2

4 CAL

809.18

24.21

50°.77

215 1.12

151-69

569.02

25.25

512.19

12.52

000 66.00

192.08

150.10

60.69

12:09

6.69

200 VIL

012.21

0.00

120.00

60.05

110.05

996.92

162.51

020

000.52

110.00

V.V.

10.02

00

C.

11.000

55-10

128.04

200

050.72

V12:09

Sister

Vess 22

162.10

1

.°° 2

5.51

1054

0.64

692.95

000001

980.04

01.54

20

510.22

11.00

105.24

161

020

620.020

.0. 57

·%

012.21

N

161

S.

75.278

10

52.825

20

1.502 1.502

04.00

1758

10.51

1.74

68.971

25.00

12.926

13.291

18.98

N.003

21.00

90.736

82.623

8.

2.50

65.576

66.281

293

10.011

63.129

92.19

1.18

78.503

80.11

100 BB

759

153.982

12

23.668

67.301

14.661

5.515

05:09

120000

60.00

11.05

986.92

162.01

020

.c.

110.00

10.91

5

252.28

24.28 VVL 28

012.08

10.000

026.10

155.2L

V62.V6

3

66.68

230

692.051

17

000001

6.07

OLV.CL

230.40

59.52

80.

15.721

69.306

36.01

92.292

8.000

39.774

18.76

1.981

81.55

224

50.

16.021

52.098

278

22.199

10.532

23.662

17.078

29.695

08.913

68.683

23.972

12.751

10.139

050.235

85.05

13.991

01.250

82.894

12.903

35.31

23.164

13.003

29.608

14.555

09.683

72.732

82.104

10.00

76.693

8.202

59.12

25.22.

20.55

17.121

2.50

17.609

01.821

629

9.228

50.00

13.051

7.050

15.014

23.261

1.088 CP

11.924

64.55

83.579

10.814

84.51

66.986

020

10.00

2

72.394

00

0

11.00

04.550

18.003

20.310

17.062

1.50

11.12

29.03

82.1M

1. 22. 6%

02000

6.

100

195.95

6020

00 coss

.0

128.00

.61

619

64.0

.2

224.11

ST 200.11

V62.19

102.18

109.52

200

.2

N

126.06

250.00

oly .c.9

106.04

11.000

84.00

262

NOS VOU

c06.01

89.99

c1.91

851.1V

60.00°

2500

205.2

529

823.51

.2

64.150

84.745

1.283

54.014

15.991

50.78

12.00

0.40

050

200

121.09

.2

-c20 -c20 -c20

6.69

12

V ó

61.64

86.210

200

01.980

322

00

64.64

600

00

191.95

2

19.90

602 126.96

275. 25

531.62

05000

0.

250.00

014

1000

10.001

53.020

3.

25.

192.68

159.10

60.00

DeNederlandscheBank

EUROSYSTEEM

7.10

73:01

81.201

30.32

96.924

23.221

A.A.

5. 2. 00

2

0.

0

S.

69.593

63.8AT

201

5.

2.2.

10.00

92.908

800

10.102

5. 25

15.00

22

Nº

32.029

38.13

20.282

12.01

01:01

7.321

0.988

2.00.101

2.21.062

51.10

77.355

31.27

81.01

18.51

63

9. 9. 5. S.

00

21.664

16.293

13.013

22.211

65.634

15.59

~

10.00

19.430

34:09

20 000 - CE 1 - 2-2

82.3

10.

3

8

798

0.

23.323

12.036

51.345

14.000

1.605

15.026

67.217

63.032

3

S.

18.372

6.129

67.672

81.91

55.794

97.891

81.382

34.018

31.251

28.11°

62.31

22.766

N. 28

50

S.

83.104

77.910

80. MI

6.61.5

1.154

15.15

75.293

22.662

05.151

66.579

8.519

12.960

91.883

032.002

10.080

39.76

1.579

92.23

17 69.1

13.511

N. 1989

19.50

26.00

15.11

20.13

32.17

28.00

10.890

20.148

10.504

18.152

80.00

29.131

76.198

17.653

51.919

Same asset, different value?

Fonds	Periode	aantal	Aankoop- waarde	Marktwaarde	Asset	Prijs per eenheid	Waarde per eenheid
1	2018Q1	2072	€ 27,800.93	€ 41,109.01	French company	€ 18.32	€ 19.8
2	2018Q1	0	€ -	€ 2,100.76	French company	€ 18.32	€ 18.3
3	2018Q1	4598	€ 51,130.96	€ 84,200,23	French company	€ 18.32	€ 18.3

*Examples with fictional data



Did they sell the assets or not?

periode	aantal	aankoopwaarde	marktwaarde	asset	waarde per eenheid
2020Q4	20600	€ 81,701.17	€ 121,942.05	American company	€ 5.92
2021Q1	19400	€ 89,705.56	€ 130,008.09	American company	€ 6.70
2021Q2	18700	€ 70,528.19	€ 130,607.77	American company	€ 6.98
2021Q3	16003	€ 288,343.92	€ 1,310,654.20	American company	€ 81.90
2021Q4	1	€ 1.08	€ 1,200,899.20	American company	€ 1,200,899.20

*Examples with fictional data



Potential of outlier detection

DeNederlandscheBank

Use cases outlier detection for actuaries





Fraud in claims

Outliers in reserving data

Data cleaning



Questions?

DeNederlandscheBank

EUROSYSTEEM

CH0