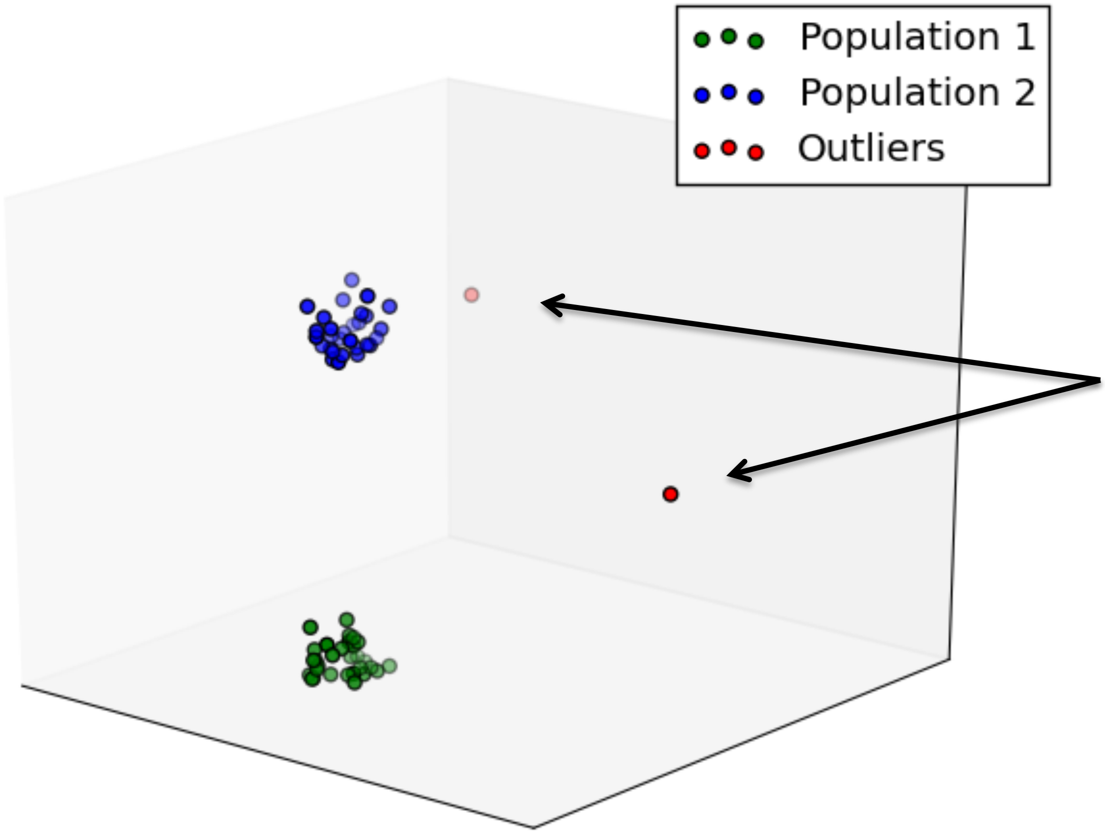
Unsupervised Anomaly Detection using H2O.ai

Peter Schrott, Julian Voelkel | Scalable Data Analysis and Data Mining

Agenda

1. Introduction
2. Problem Statement
3. Methodology
4. Experiments & Results
5. Conclusion

Anomaly Detection



Identify data points that do not fit the pattern of the data

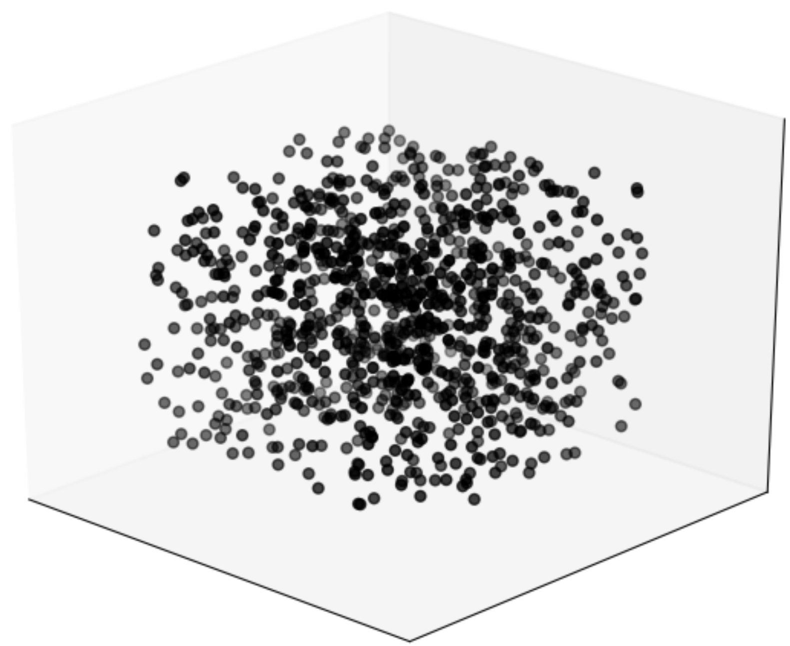
Unsupervised Anomaly Detection

"Anomaly detection is about finding what you don't know to look for."

- Ted Duning[1]

Fundamental assumption:

- Amount of normal data points exceeds the amount of anomalous data points by far



[1] Practical Machine Learning: A new Look at Anomaly Detection

Problem Statement

Is a deep learning auto-encoder well suited for anomaly detection in an unlabeled dataset?

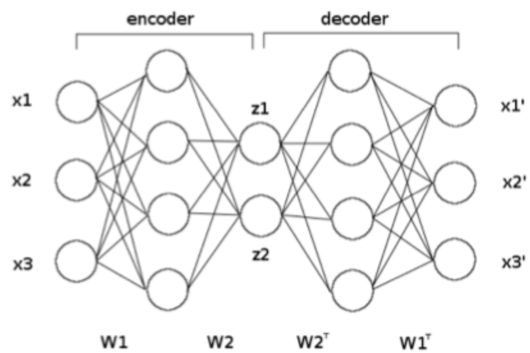
Problem Statement

? ? ?

Is a deep learning auto-encoder well suited for anomaly detection in an unlabeled dataset?

Foundations: Deep Learning Auto-Encoder

3	5	8	17
12	11	–3	12
0	34	11	23
7	–1	–2	9



≈

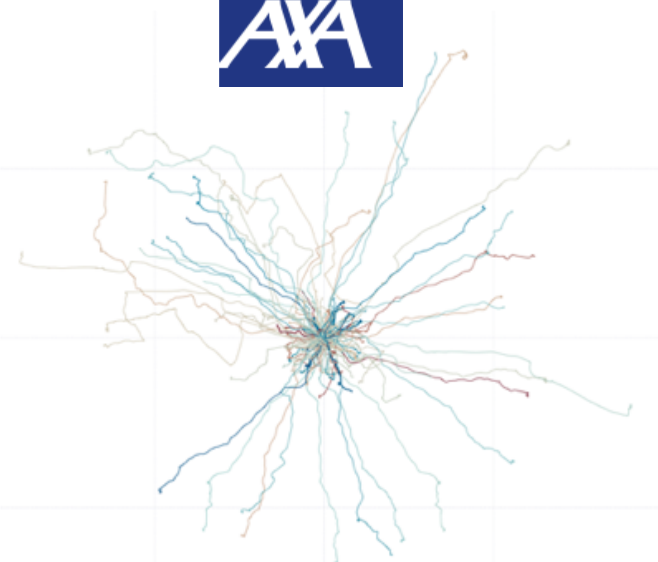
3.2	4.9	8.1	17.4
12.2	10.7	–3.5	12.1
0.4	34.2	11.1	23.3
7.3	–1.06	–2.1	9

- Neural Net
- Desired output = input

- Auto-encoder learns a compressed representation of input

➢ Most commonly used for dimensionality reduction purposes

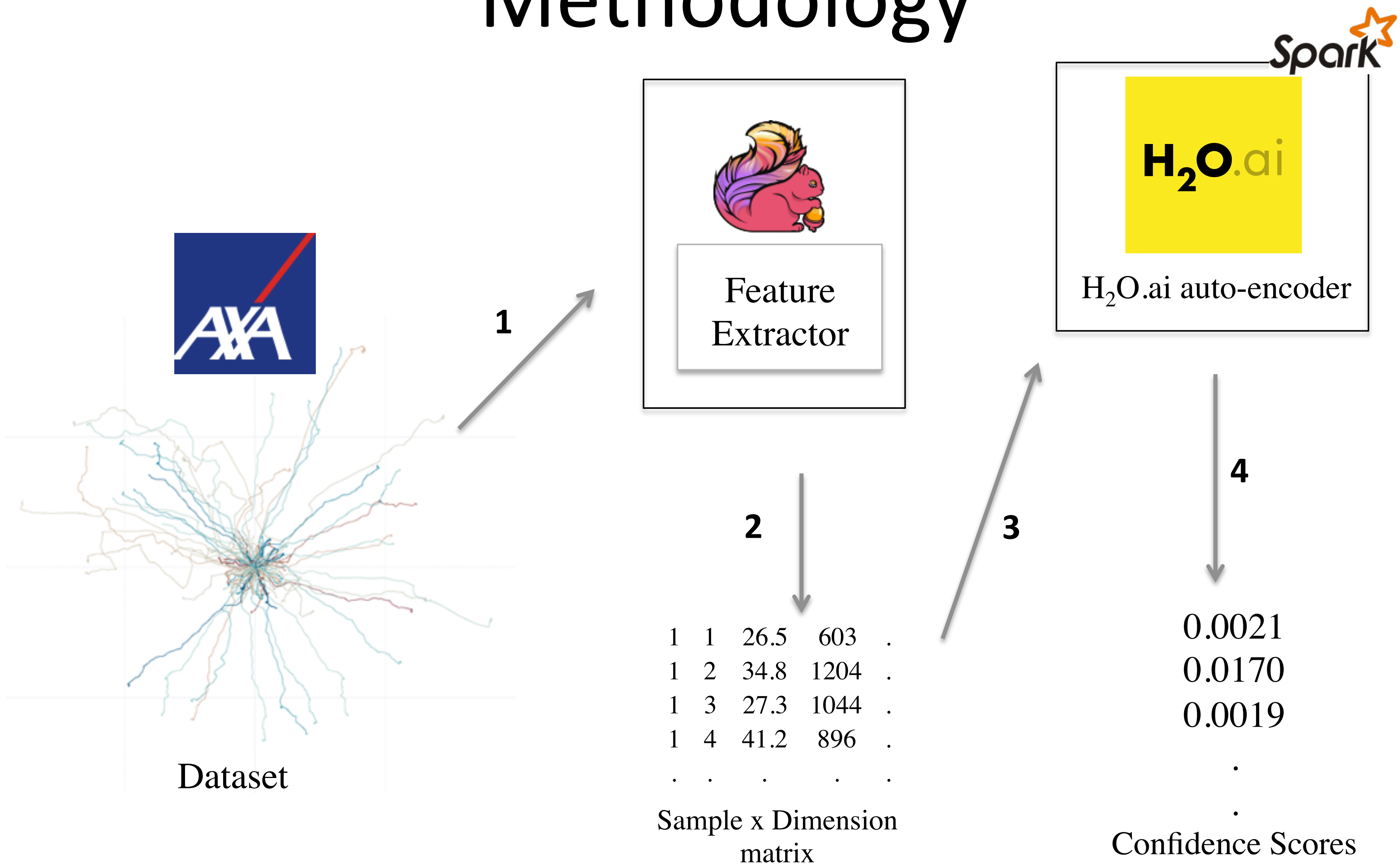
Dataset



Dataset

- Provided by Kaggle.com
- 5.92 GB
- 2736 Driver
- 200 Trips each
- X- / Y-coordinates
- Folder- / File- based structure
- Anonymized by cropping & rotation
- Trips start at (0,0)

Methodology



1st Experiment

- One general model for all drivers
- Normalization of resulting confidence scores

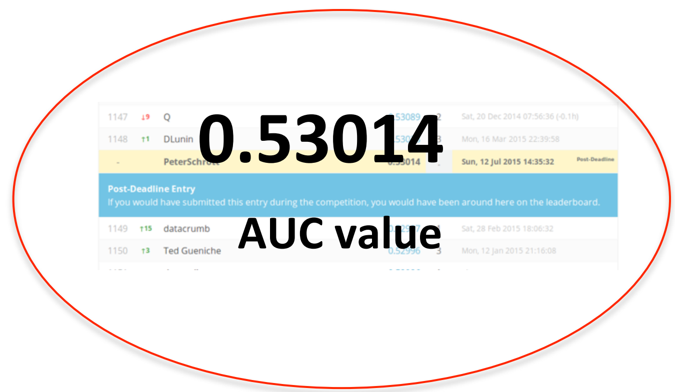
1	1	26.5	603	.
1	2	34.8	1204	.
1	3	27.3	1044	.
1	4	41.2	896	.
.	.	.	.	.

Sample x Dimension matrix

H_2O.ai auto-encoder

0.0021
0.0170
0.0019
.
.

Confidence Scores



Conclusions 1st Experiment

1148	↑1	DLunin	0.53028	3	Mon, 16 Mar 2015 22:39:58	
-		**PeterSchrott**	**0.53014**	-	**Sun, 12 Jul 2015 14:35:32**	**Post-Deadline**

Post-Deadline Entry
If you would have submitted this entry during the competition, you would have been around here on the leaderboard.

1149	↑15	datacrumb	0.52997	1	Sat, 28 Feb 2015 18:06:32

- Generalization of model too high
- Wide variance between drivers
- Finding optimal parameters is tough
  - No labels, means no feedback

2^{nd} Experiment

- **One single model for each driver**
- Normalization of resulting confidence scores

1	1	26.5	603	.
1	2	34.8	1204	.
1	3	27.3	1044	.
1	4	41.2	896	.
.	.	.	.	.

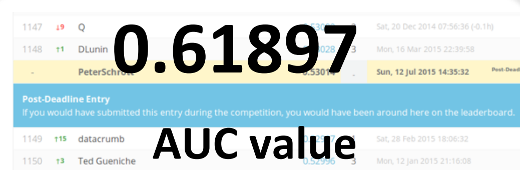
Sample x Dimension matrix

Spark

H_2O.ai auto-encoder

0.0021
0.0170
0.0019
.
.

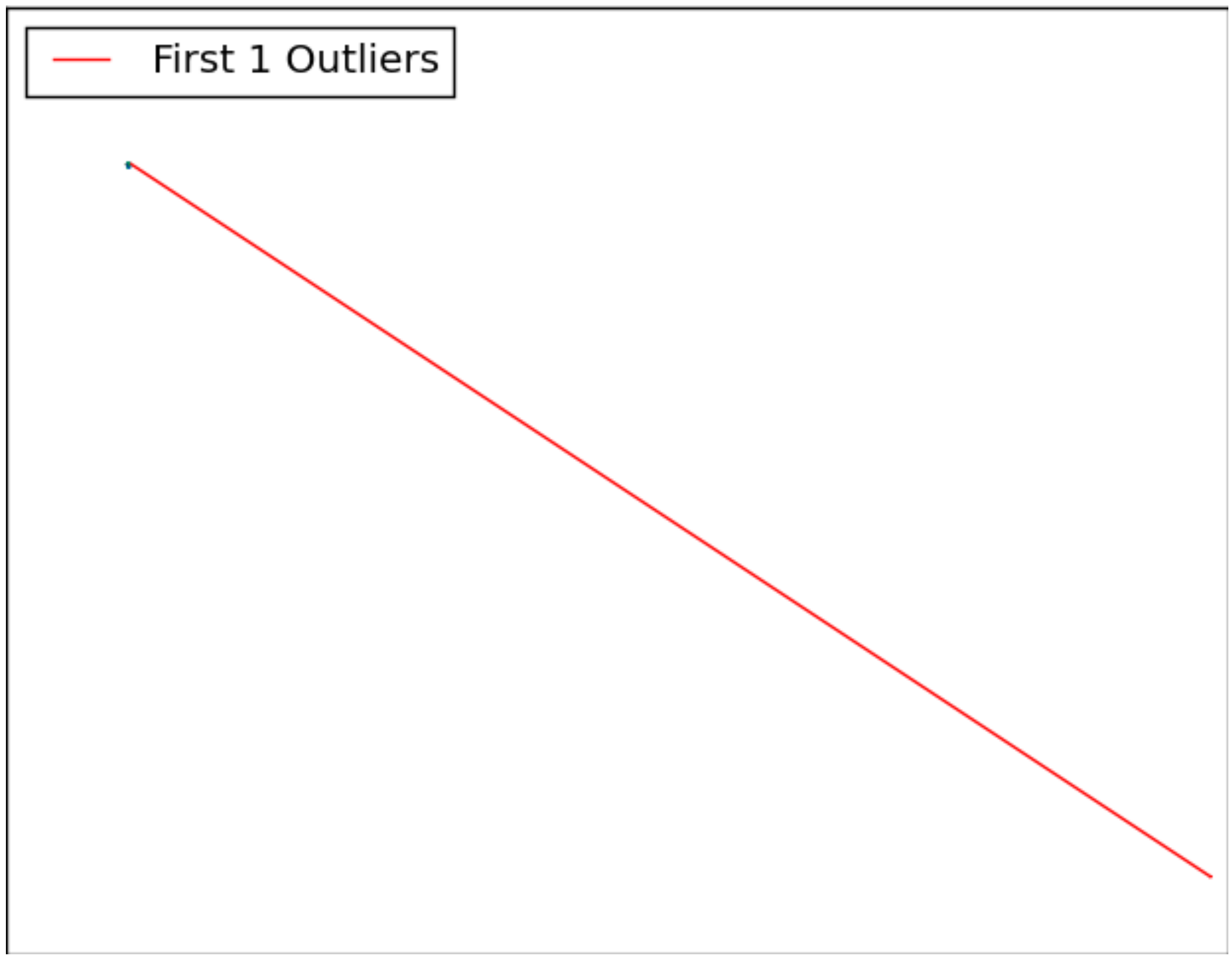
Confidence Scores



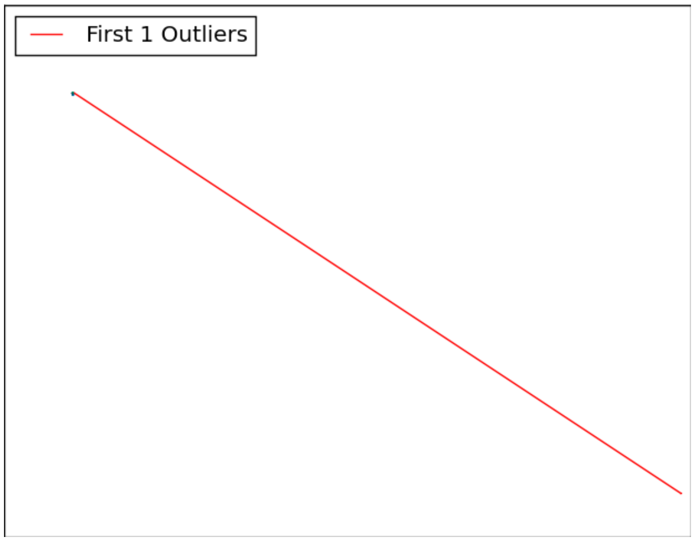
Visual Exploration

- Find 15 trips with highest reconstruction error
- Visualize trips with matplot-lib

Visual Exploration

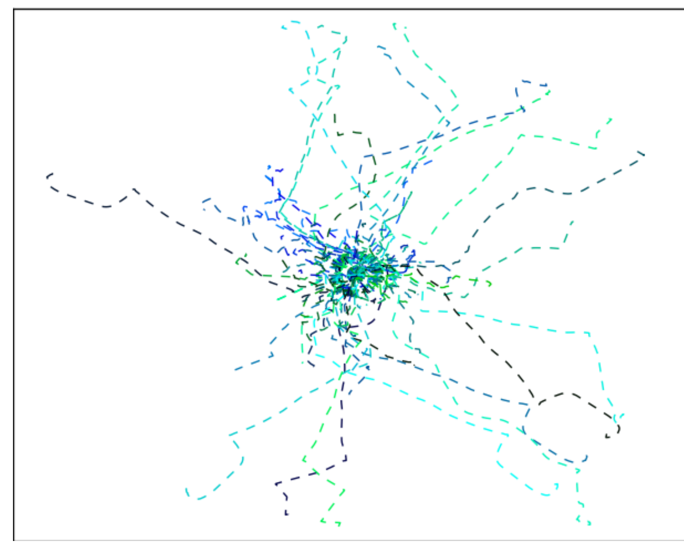


Visual Exploration



All trips of driver 1634

- Dataset apparently still contains junk drives…



All trips but trip #136

- … that seem to be spoiling the model and / or predictions

Conclusions 2nd Experiment

960	↑2	NiNot	0.61904	21	Mon, 16 Mar 2015 19:53:13 (-2.4d)	
-		**PeterSchrott**	**0.61897**	-	**Tue, 14 Jul 2015 07:58:04**	**Post-Deadline**

Post-Deadline Entry
If you would have submitted this entry during the competition, you would have been around here on the leaderboard.

961	↓2	Alvaro	0.61884	9	Sun, 15 Mar 2015 15:20:29 (-23.2h)

- Noisy data seems to influence our model
- Noise is recognized as outliers, but seems to distort the confidence scores and thus probabilities

$$\hat{x}_i = \frac{x_i - min(X)}{max(X) - min(X)}$$

3rd Experiment

- One single model for each driver
- Normalization of resulting confidence scores
- **Discounting highest two confidence scores per driver**

1	1	26.5	603	.
1	2	34.8	1204	.
1	3	27.3	1044	.
1	4	41.2	896	.
.	.	.	.	.

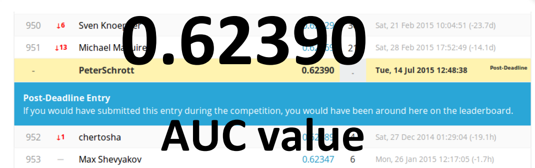
Sample x Dimension matrix

Spark

H_2O.ai auto-encoder

0.0021
0.0170
0.0019
.
.

Confidence Scores



0.62390

AUC value

952th out of 1,529

Conclusions 3rd Experiment

950	↓6	Sven Knoepfler	0.62529	36	Sat, 21 Feb 2015 10:04:51 (-23.7d)	
951	↓13	Michael Maguire	0.62469	21	Sat, 28 Feb 2015 17:52:49 (-14.1d)	
-		**PeterSchrott**	**0.62390**	-	**Tue, 14 Jul 2015 12:48:38**	**Post-Deadline**

Post-Deadline Entry
If you would have submitted this entry during the competition, you would have been around here on the leaderboard.

- Better results after accounting for an arbitrary number of junk drives per driver
- Noise is recognized as outliers, but seems to distort the confidence scores and thus probabilities

Conclusion

- AUC score of 0.62390 far from reliable outlier detection
- But still better than the 0.5 benchmark
- Spend more time on cleaning the data

Recommendations for future work:

- Use cleaner dataset
- If possible with features already given

Thank you for your attention!

Were you surprised by any of your findings?

I was surprised at the number of junk runs there were. Some of the drivers had 30, 40 or more junk runs!

- Scott Hartshorn (2nd place in the AXA Driver Telematics challenge on Kaggle.com)