MIS772

Predictive Analytics

Anomaly Detection

Finding the odd ones out

Refer to your textbook by Vijay Kotu and Bala Deshpande, *Data Science: Concepts and Practice*, 2nd ed, Elsevier, 2018.

Anomaly Detection

- **Outliers vs anomalies**
- **Causes of outliers**
- **Anomaly detection techniques**
  Statistical, Distance, Density, Clustering and Classification
- **Anomaly visualisation**
  with PCA and SVD (Singular Value Decomposition)

DEAKIN BUSINESS SCHOOL

AACSB ACCREDITED

EFMD EQUIS ACCREDITED

Outliers vs Anomalies

- *Outliers* are data points that are very rare in a data set.
- Outliers is a stats concept.
- *Anomalies* are data points that are far removed from the established norms.
- Anomalies is a ML concept.
- Individuals that are lonely, dissimilar, abnormal, suspect, criminal or exceptional are considered anomalous.
- Outliers and anomalies are also referred to as deviants, abnormalities, discordants and extremes.
- Sometimes, we will consider *noise* and *errors* as separate from outliers and anomalies.
- *We will use both terms interchangeably*.

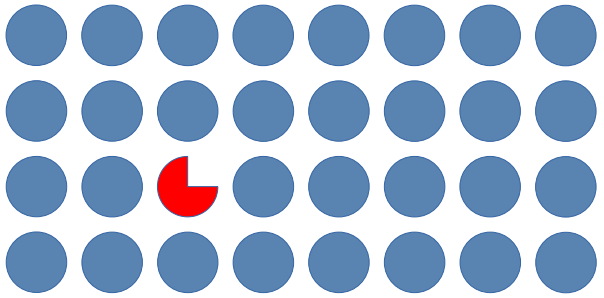
- In the simplest uni-variate case outlier detection is the task to identify extreme values.
- In multi-variate case outliers have low probability of finding their combinations of attribute values, as compared with the rest of data.
- The concept of anomaly is often associated with the notion of measuring distance between data points.
- In this context, anomaly detection is the process of finding data points, which are far away from the main group or the closest group of other data points.

Anomaly ≈ Outlier

Also see KD 13.1.1 on Causes of Outliers

Causes of Outliers

- *Data errors* due to human error (e.g. during data entry), poor data acquisition practices, faulty data storage or comms equipment, etc.
- *Normal variance in data*, due to distribution of its attribute values, e.g. in a normal distribution 0.3% of attribute values are outside 3 standard deviation from the mean, and thus they are very rare.
- *Incorrect assumption about data distribution in population*, e.g. it can be assumed that only 5% of students get HDs, the result deviating from this assumption may be considered an outlier.



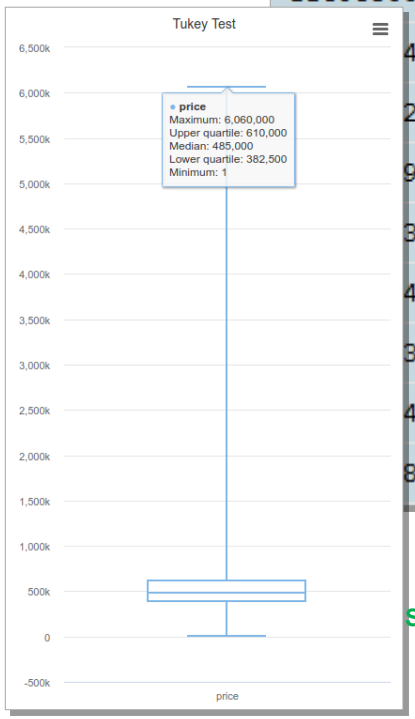
- Other causes of anomalies in data include extreme cases, chance, skewness, flawed theory, bad data distribution (e.g. normal vs poisson), mix of distributions, etc.
- Anomaly detection is important to identify the specific individuals in the population, to correct data, to improve data pre-processing or the model itself.
- There are two main types of anomaly detection, i.e. using *statistical methods* and using *data mining*.
- Outliers can also be determined as out-of-cluster or cluster-boundary data points.
- In some cases, it is possible to train a classification model to learn identifying anomalies, e.g. SVM for global anomaly detection.

Also see KD 13.1.2 on Techniques

Statistical Methods

- **Tukey's outlier test** is used with uni-variate data only and is the simplest of the methods.
- It identifies outliers as data points which are beyond the boundaries (Q1 – 1.5*IQR, Q3 + 1.5*IQR). Where Q1=Lower quartile, Q3=Upper quartile, IQR= Inter-quartile range
- For example, the IQR of the Price attribute is $228K (Q1=$382K and Q3=$610K), Q3+1.5*IQR is $952K, so Price=$955K is (just) an outlier.

propertyID	price ↑	TukeyTest_price
119605915	950000	No Outlier
118983663	950000	No Outlier
45	955000	Top Outlier
21	960000	Top Outlier
99	960000	Top Outlier
39	960000	
49	962500	
38	965000	
47	965000	
88	970000	



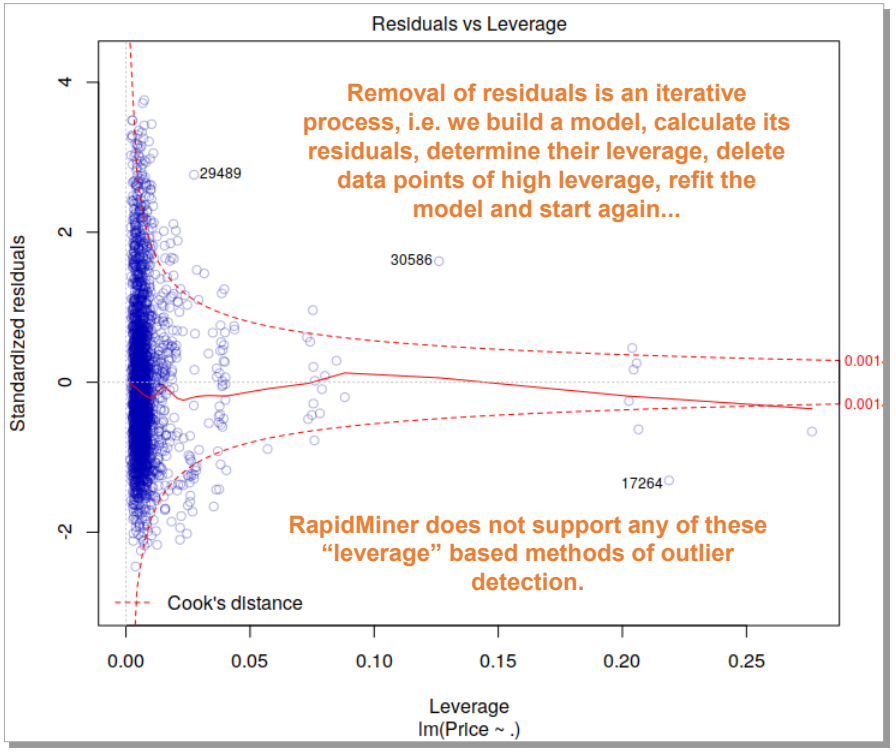
Also see KD 13.1 on Stats Methods

- A histogram approach is similar but for multi-variate examples.
- Here attribute values are split into a fixed number of bins in each attribute and the bin histograms are analysed. The example's outlier score will be high when its attributes belong to the most sparse bins
- For instance, the histogram-based approach scored high the following examples as outliers:
  - Large number of car spaces vs the number of bedrooms
  - Huge price vs page visits
  - No car space vs page visits

propertyID	score ↓	car_spaces	property_type	bedrooms	bathrooms	agency	suburb	price	page_visits	latitude	longitude
119687423	9.559	4	House	2	2	Aussie ...	Toorak	2550000	2420	-37.843	145.009
119202603	9.000	2	Apartment	2	3	Castra...	Kooyong	1610500	943	-37.841	145.032
120615729	7.384	2	Apartment	2	2	Kay & B...	Toorak	2376000	1299	-37.838	145.015
119476319	6.675	1	Apartment	2	2	Melbou...	Prahran	2922000	3500	-37.852	145.011
107452520	6.611	3	Apartment	2	2	Buxton	St Kilda	1630000	580	-37.864	144.974
119613271	6.496	0	House	2	1	Hodges	Bentleigh	1770000	1864	-37.916	145.036
120861005	6.389	1	House	4	1	Wooda...	Carnegie	1607000	1360	-37.889	145.054
105189092	6.211	0	Apartment	1	1	T G Ne...	Caulfield North	201000	8182	-37.867	145.026
120797961	6.140	2	Apartment	2	2	Greg H...	Armadale	1270000	491	-37.854	145.016
119042907	6.072	2	Apartment	2	2	Castra...	Toorak	1180000	1831	-37.836	145.024
110955019	5.896	0	Apartment	1	1	Buxton	St Kilda	200000	7071	-37.860	144.974
105936687	5.888	2	Apartment	2	2	Chishol...	St Kilda	1700000	1177	-37.866	144.975
105608135	5.813	0	Apartment	1	1	Gary P...	Caulfield North	180000	5889	-37.867	145.026
106962237	5.813	0	Apartment	1	1	Biggin ...	Caulfield North	150000	5849	-37.867	145.026
115615915	5.751	1	Apartment	1	1	Barry Pl...	Prahran	110000	5038	-37.850	144.992

Outliers in Linear Models: Leverage of Data Points

- In regression, an outlier is determined by the size of its residual and its influence on the model.
- A data point has high *leverage* when the amount of the total sum of squared errors contributed by that point is above a certain cut-off value.
- *Cook's distance* defines this cut-off range which allows to calculate the "leverage" boundary for residuals, outside of which an example will be considered an outlier.
- Another approaches may involve deleting a data point to determine its influence on the model.
- After removing a data point, we refit the regression model using the remaining data.
- *Deleted residual* for a data point is a difference between its actual and predicted label, when calculated using a model without that point. A large deleted residual for a data point indicates its high influence on the model.



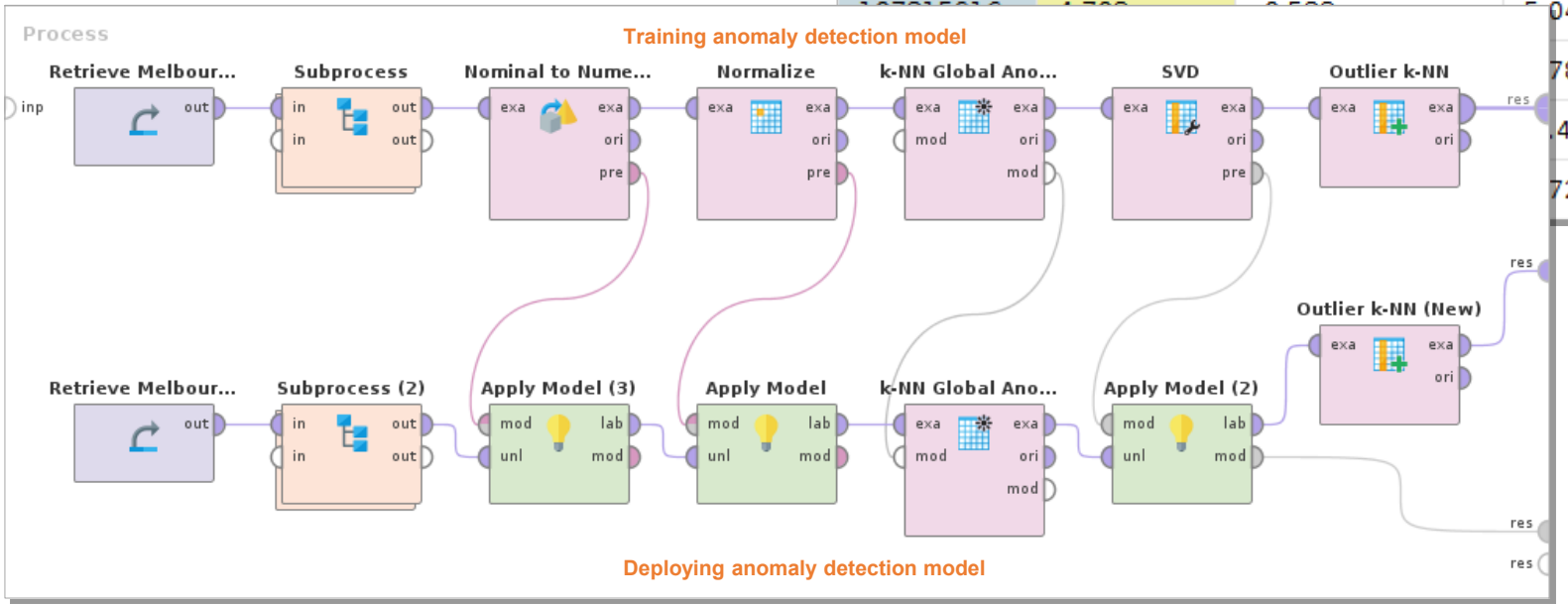
DEAKIN BUSINESS SCHOOL

- ***Distance based approaches***, based on k-NN, are common – they identify anomalies as data points with the maximum sum of distances to their neighbours.
- K-NN distance measures must fit the data, e.g. Cosine measures suit data describing general direction or trend, Euclidean distance suit spatial data, while mixed metrics cater for examples with nominal attributes.
- k-NN based methods provide an outlier score – the higher the score the further anomaly.

- In model deployment, all data transformations applied in data pre-processing, anomaly detection and visualisation must be saved and applied in exactly the same way to any newly collected data (see a sample model below).

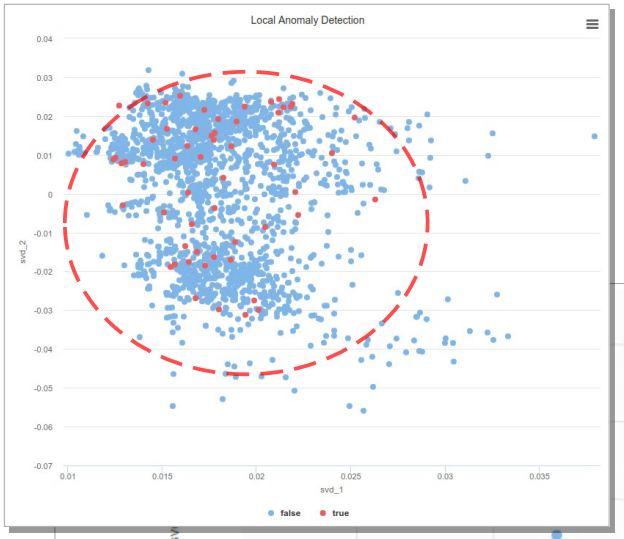
Also see KD 13.2 on Distance Methods

propertyID	outlier ↓	property_type	agency	su
106758513	7.789	9.381	5.584	0.1
119687423	6.490	1.277	6.372	4.4
112797151	6.031	-0.523	-0.577	-0.
106760252	5.840	6.679	-0.648	7.3
109744981	4.706	2.178	-0.111	-0.
			047	7.5
			784	1.4
			.434	-0.
			721	1.4



Distance-Based AD: Global vs Local

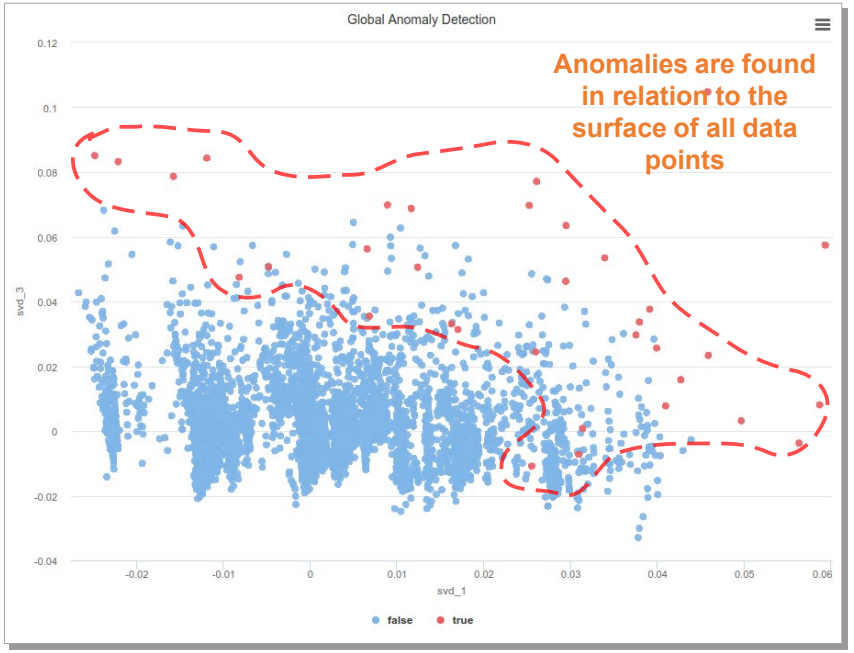
- ***Global anomaly detection*** **methods consider all points as potential neighbours.**
- ***Local anomaly detection*** **identifies anomalies as data points furthest removed within a neighbourhood.**
- **Global anomaly detection is easier to visualise in 2D (below).**
- **Local anomaly methods can be more precise as they find anomalies in sparse areas of data.**
- **However, they are more difficult to project into a 2D plane (right).**

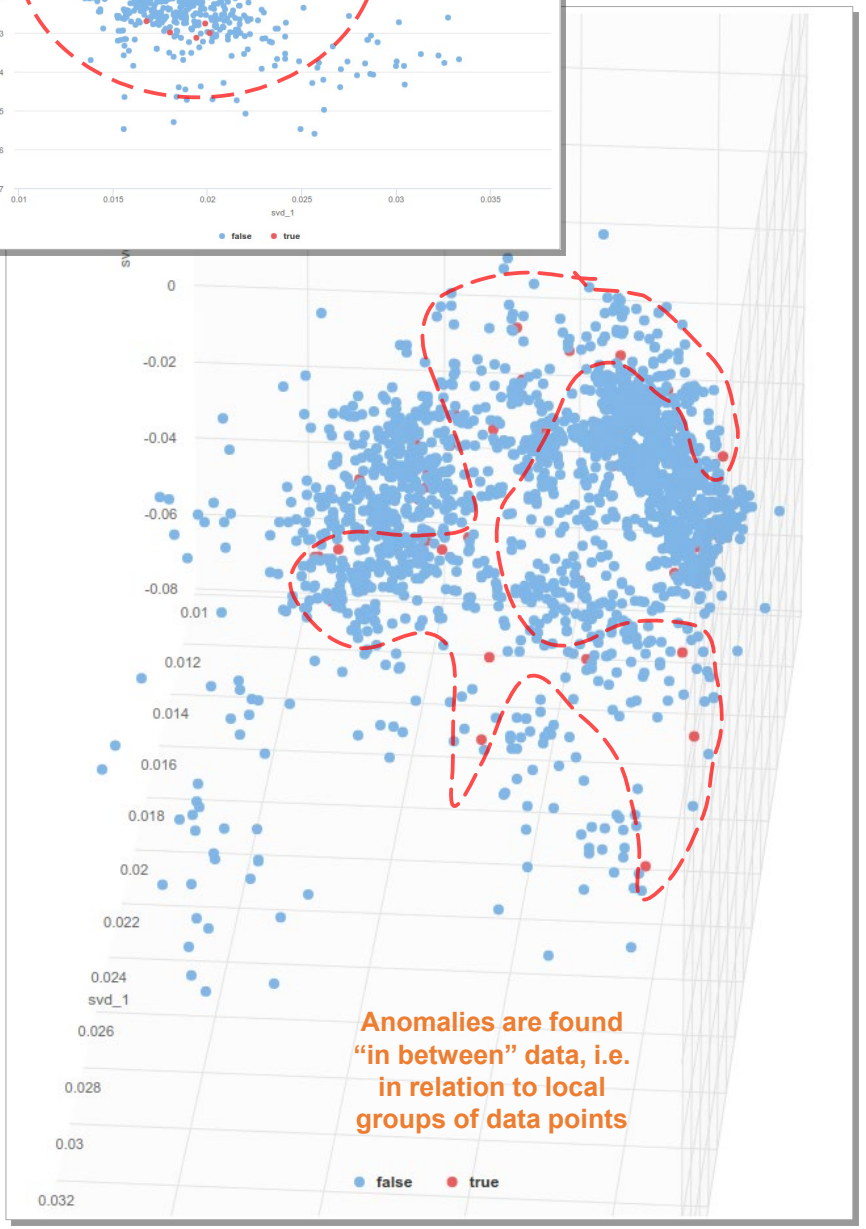


Local anomaly detection is difficult to visualise in 2D

However, some projections may reveal the precision of the method

Global anomaly detection





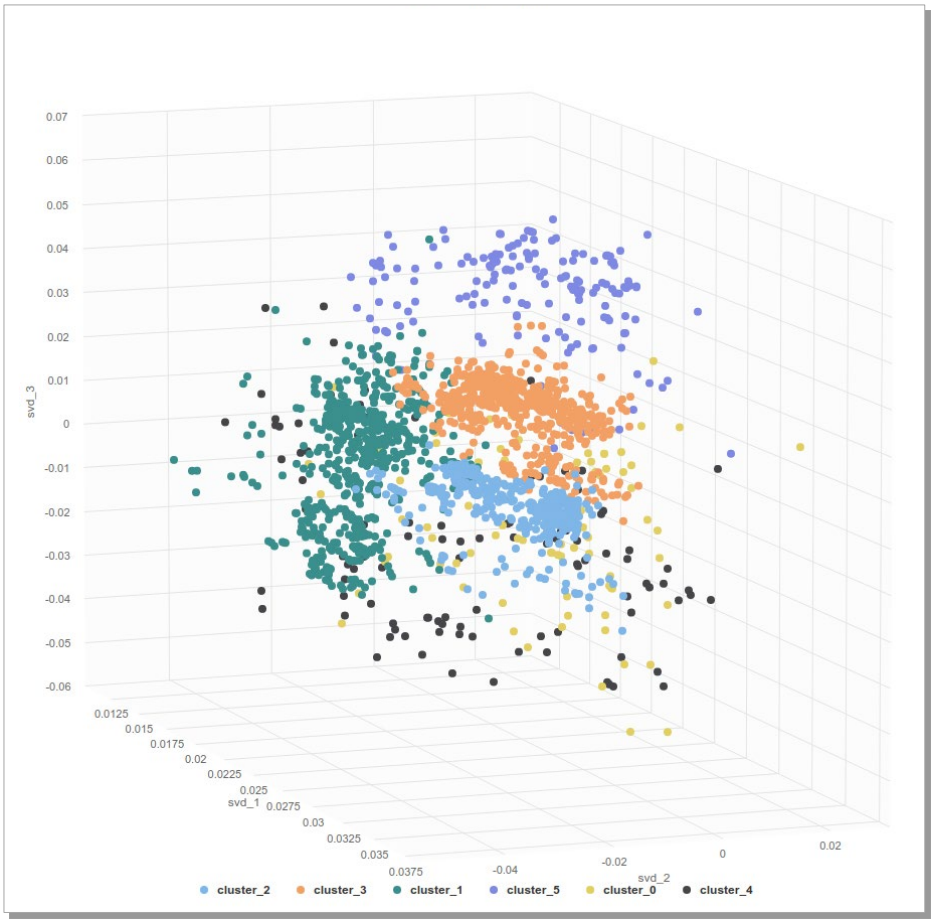
Local anomaly detection

DEAKIN BUSINESS SCHOOL

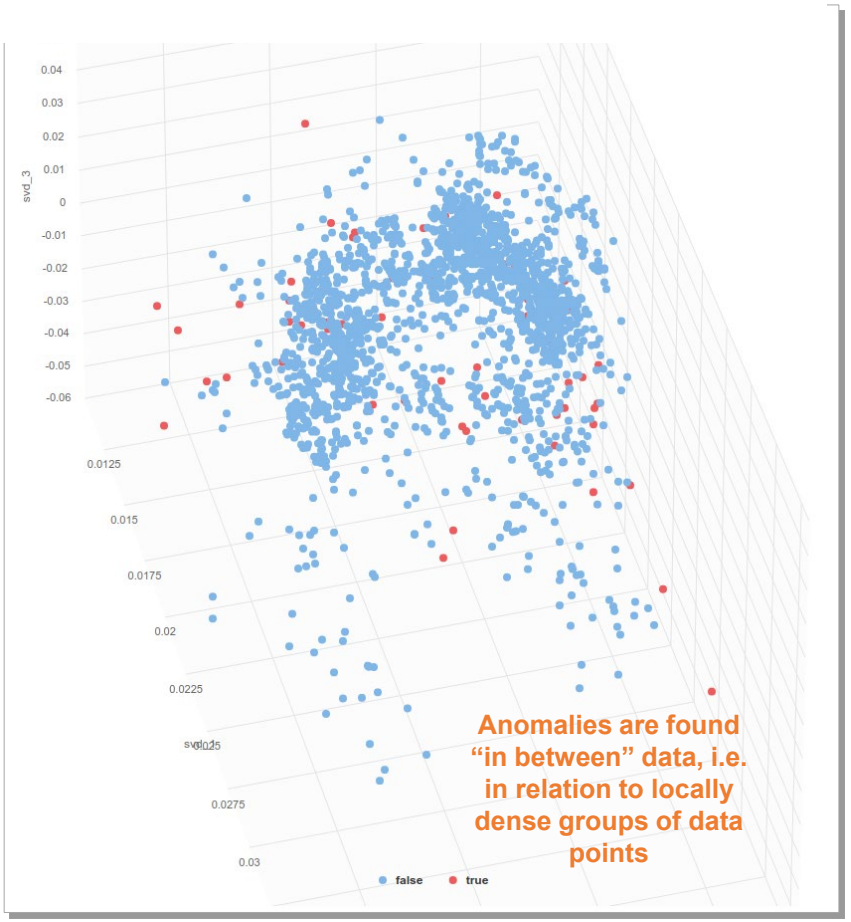
Clustering and Density-Based Anomaly Detection

- ***Density-based anomaly detection*** **methods are similar to the local anomaly detection. They focus on data density in a given "neighbourhood" space.**
- **The neighbourhood can be defined mathematically or by the previously constructed data clustering model.**

Also see KD 13.4 on Density Methods



Cluster Model

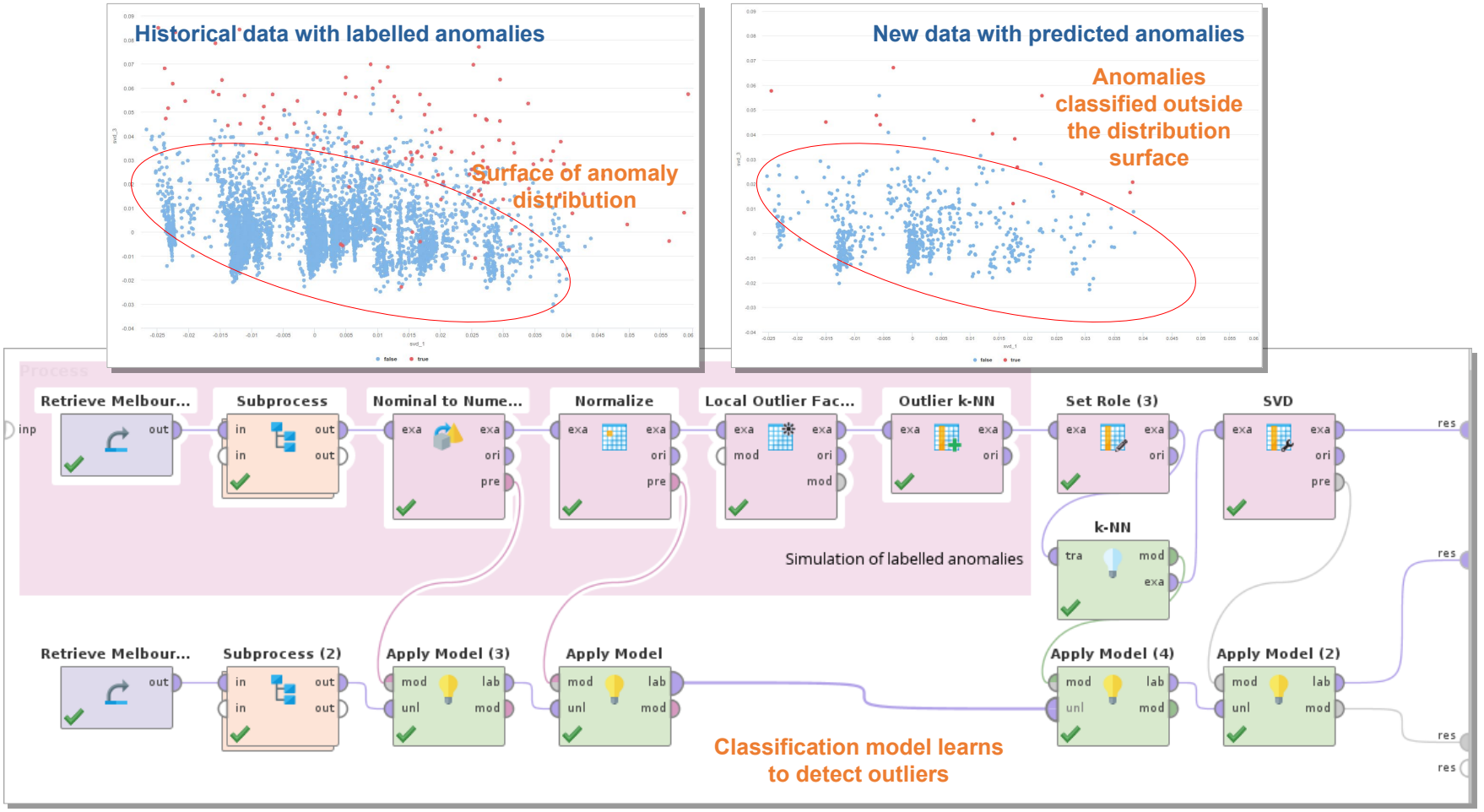


Cluster-Density Anomaly Detection

DEAKIN BUSINESS SCHOOL

Classification-Based Anomaly Detection

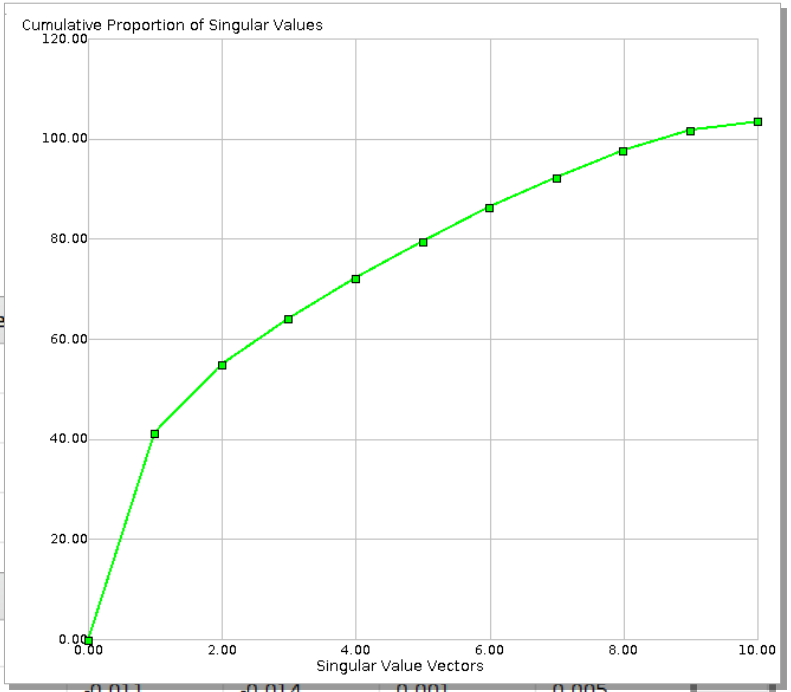
- ***Classification-based anomaly detection*** **methods assume that anomalies have been previously labelled, e.g. identified in the normal business practice or product use.**
- **As any labelled data, it can be used to train a predictive model, e.g. k-NN or a Decision Tree, to identify future anomalies.**
- **The choice of the anomaly classifier depends on the characteristics of the surface of non-anomalous data.**
- **If global anomaly distribution is in place then k-NN (with large k) may be suitable.**
- **If the labels suggest a density style distribution a k-NN (with small k) or a decision tree may be preferred.**



DEAKIN BUSINESS SCHOOL

Singular Value Decomposition (SVD)

- ***Singular Value Decomposition (SVD)***, **is a more general case of PCA and treats examples as a linear combinations of correlated attributes.**
- **SVD finds new uncorrelated components of data called *singular values* such that:**
  - **Singular values are *orthonormal*, i.e. uncorrelated and of unit length**
  - **Each singular values "explain" an amount of variance in data**
  - **The *cumulative variance plot* depicts how singular values contribute to the overall variance**
- **In RapidMiner the SVD cumulative variance is not normalised.**
- **Note that unlike in the PCA, SVD does not need data to be centred or normalised.**
- **SVD can also be used to reduce dimensionality of data, which is commonly relied on in 2D data visualisation.**



Attribute	SVD Vector 1	SVD Vector 2	SVD Vector 3	SVD Vector 4	SVD Ve
property_type	0.069	-0.118	0.134	0.003	-0.058
agency	0.111	0.053	-0.138	0.449	0.289
suburb	0.217	-0.072	-0.225	0.730	-0.388
car_spaces	0.337	-0.072	0.251	-0.199	0.354
bedrooms	0.316	-0.051	0.670	-0.014	-0.555
bathrooms	0.067				
price	0.118				
page_visits	0.132				
latitude	0.705				
longitude	0.435				

propertyID	svd_1	svd_2	svd_3				
103733163	0.015	0.025	-0.016				
104084264	0.017	0.009	0.010	-0.011	-0.014	0.001	0.005
104100830	0.020	-0.030	-0.000	-0.009	-0.010	0.004	-0.012
104116990	0.015	-0.016	-0.018	-0.020	0.021	0.010	-0.005
104146492	0.013	0.014	-0.013	-0.009	0.010	0.014	0.014
104240560	0.020	-0.009	0.000	-0.006	-0.014	0.007	-0.003
104308597	0.020	0.007	0.000	-0.005	-0.016	0.009	0.000

Also see Readings

DEAKIN BUSINESS SCHOOL

Summary / Revision

- What are outliers?
  What are anomalies?
- What is the aim of anomaly / outlier detection?
- What are the common causes of anomalies?
- What are the main types of anomaly detection?
- Describe two distinct statistical approaches to anomaly detection.
- What is a variable leverage?
  Is it good when it is high?
- What is Cook's Distance?
- What are deleted residuals?
  What is their purpose?
- Describe a distance-based approach to anomaly detection.
- What is the difference between local and global anomaly detection approaches?
- Describe a density-based anomaly detection method.
- Why is it difficult to visualise anomalies based on local or density anomaly detection?
- What is SVD?
- How is SVD used in anomaly detection?