MIS772

Predictive Analytics

Workshop: Anomaly Detection

Data preparation, anomaly detection and visualisation with SVD

DEAKIN BUSINESS SCHOOL

Workshop Plan

Objectives:

Your task is to create an anomaly detection model, and to isolate (and possibly eliminate) anomalies in a dataset.

Data Set:

Use file "Melb Real Train.csv"

Method:

Attend the seminar, follow the tutor's demo and instructions, take notes. Note that the class and online seminar will be recorded and their videos linked to the CloudDeakin topic for later access and study.

1. **Acquire data for anomaly detection**
   (a) Load the real estate data and unzip
   (b) Read and explore the data set, and store
2. **Create a k-NN Global Anomaly Score model**
   (a) Select a sub-set of attributes
   (b) Undertake data pre-processing
   (c) Add "k-NN Global Anomaly Score" operator
   (d) Create an outlier-flag
   (e) Run and investigate, save
3. ***Optional:* Use PCA/SVD to visualise anomalies**
   (a) Adapt the previous process for anomaly visualisation
   (b) Add PCA/SVD
   (c) Plot anomalies using PCA/SVD, save
4. ***Optional:* Consider how you would apply anomaly detection to new data**

DEAKIN BUSINESS SCHOOL

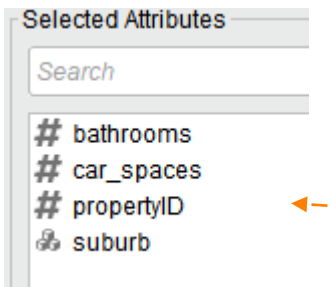
k-NN Glob Anomaly Data Prep

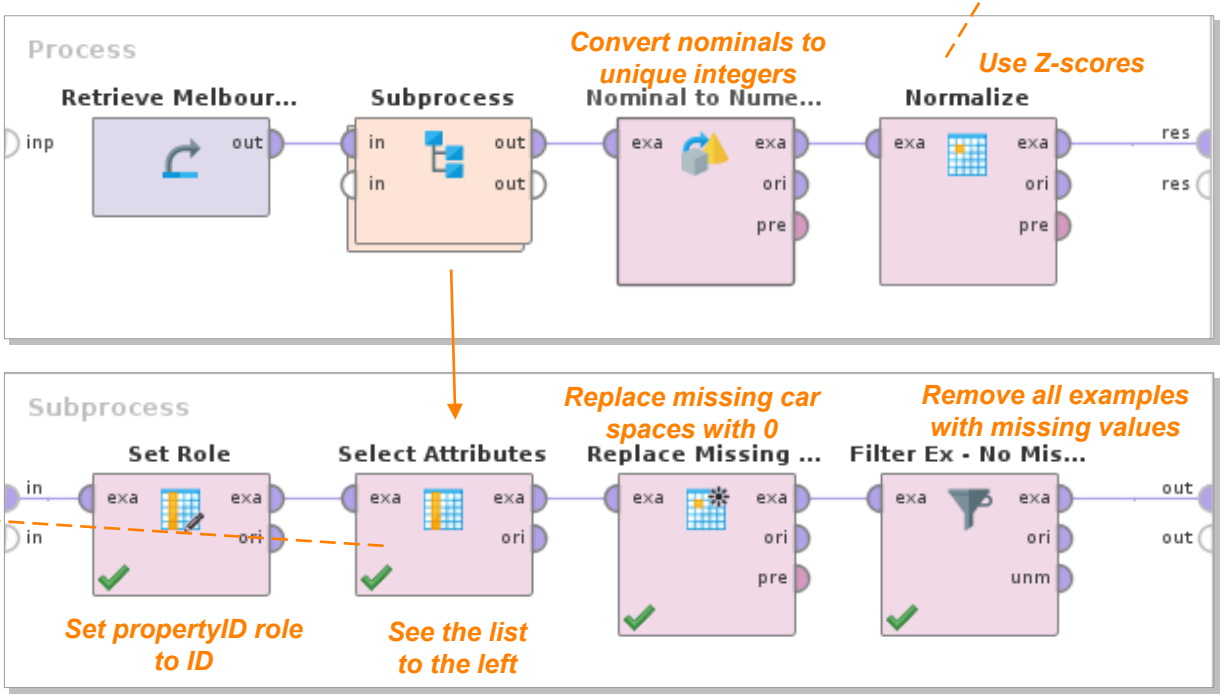
propertyID	property_type	agency	suburb	car_spaces	bedrooms	bathrooms	page_visits
102381458	-0.523	-0.648	-1.017	-1.985	-1.482	-0.312	-1.188
102934609	0.377	-0.613	-0.827	0.018	0.672	-0.312	-1.380
103363090	0.377	-0.613	-0.637	0.018	-1.482	-0.312	-1.380
103733163	-0.523	-0.577	-0.447	0.018	-1.482	-0.312	0.910
103800838	1.277	-0.541	-0.258	-1.985	0.672	-0.312	1.286
104084264	-0.523	-0.648	-0.068	0.018	0.672	-0.312	1.252
104100807	-0.523	-0.648	0.122	0.018	0.672	-0.312	-0.750
104100816	-0.523	-0.648	0.312	0.018	0.672	-0.312	0.550
104100820	0.377	-0.648	0.122	0.018	0.672	-0.312	0.176

First, we will create a process responsible for data preparation for k-NN based Global Anomaly Detection Method.

Retrieve data. Create a sub-process to undertake pre-processing which does not depend on the knowledge of training data; e.g., set role of the property ID to ID, select specific attributes, replace missing car spaces with zero, filter out all missing values.

Then perform pre-processing that needs knowledge of data; e.g., convert nominals to numeric and Z transform attributes.





DEAKIN BUSINESS SCHOOL

k-NN Anomaly Detection

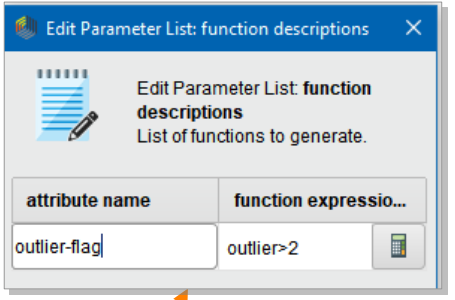
Add a k-NN Global Anomaly Score. Set the number of k-NN neighbours k=10, retain default settings. Add a new attribute outlier_flag to mark outliers true / false.

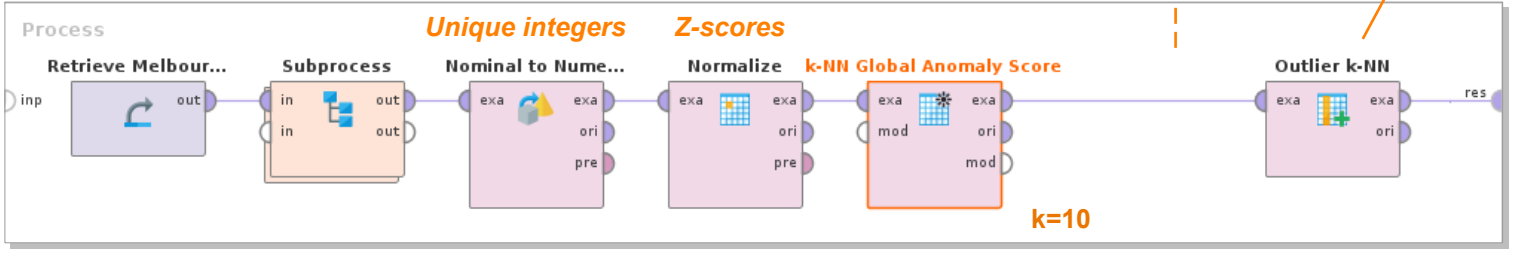
The flag value will be determined by the k-NN outlier score above certain level. The level will be defined by running the process, sorting the results by score in descending order, deciding how many anomalies are to be discovered, and writing an appropriate formula for the outlier flag.

Outlier score

propertyID	outlier ↓	suburb	car_spaces	bathrooms	outlier-flag
110400609	3.680	-0.169	2.058	7.066	true
109744981	3.474	-0.564	6.170	3.386	true
107215916	2.789	6.929	0.002	3.386	true
118905679	2.129	6.140	0.002	3.386	true
119179711	2.071	3.577	-2.054	3.386	true
120514209	1.904	1.408	6.170	-0.294	false
106475799	1.892	-0.958	4.114	3.386	false
107452520	1.890	-0.564	4.114	3.386	false
106760252	1.719	6.731	-2.054	-0.294	false
119614603	1.662	6.337	2.058	-0.294	false
120507829	1.535	6.337	-2.054	-0.294	false

Generate a new attribute to set an outlier flag true when outlier score is above some level, which you need to determine by running the process and assessing how many outliers you want to detect.

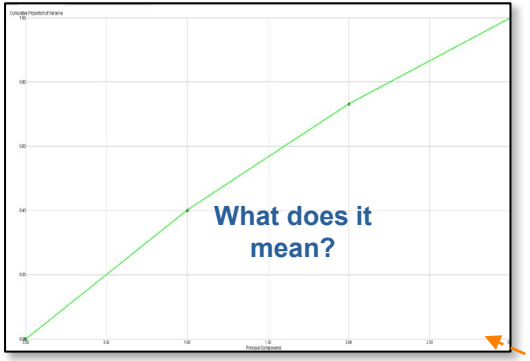


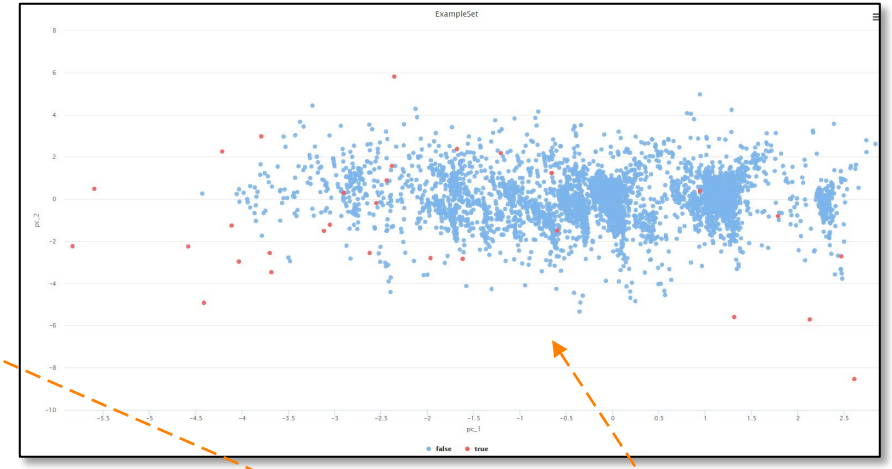


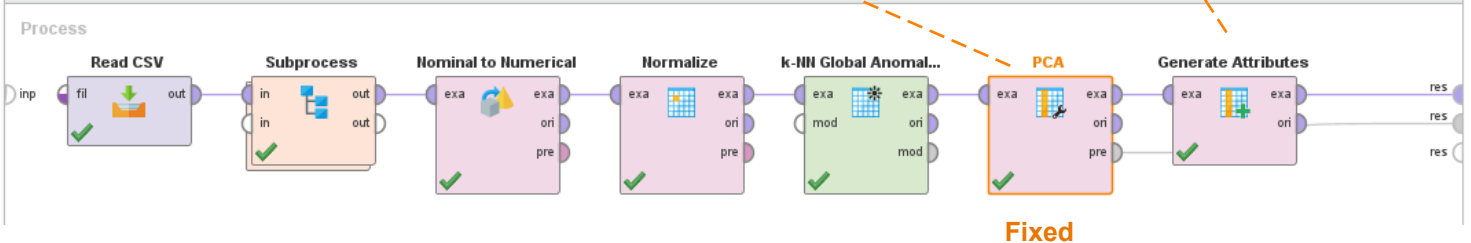
Optional: Anomaly Visualisation

Add a principal components analysis (PCA) operator to the process. Fix the number of dimensions to, say 3 (experiment with the number of dimensions, given the amount of variance captured in PCA dimensions, by referring to the cumulative plot)
This technique is useful when you have multiple dimensions in your data, making visualisation complex.
Because our data is centred and standardised, the cumulative plot explains the "variance" in data, and you can highlight the outliers.

Row No.	propertyID	outlier	pc_1	pc_2	pc_3	outlier-fl...
1245	119179711	2.071	1.601	3.474	-3.721	true
2139	110400609	3.680	6.319	-1.238	-3.567	true
2893	118905679	2.129	3.501	5.670	-2.182	true
3915	107215916	2.789	3.647	6.444	-2.154	true
5150	109744981	3.474	6.529	-1.877	1.924	true
1	102381458	0	-1.805	-0.589	-1.272	false







Fixed dimensions=3

DEAKIN BUSINESS SCHOOL