A Newsletter from Ernst & Young

April 2008

USER MANUAL

OUTLIER DETECTION

USING UNSUPERVISED LEARNING

Contents

[Project Requirement Specifics 3](#_Toc195692248)

[Running the Pyhon Script 3](#_Toc195692248)

[Sales Dataset Format 3](#_Toc195692248)

[Mapping File Format 4](#_Toc195692248)

[Steps for usinig the python script 5](#_Toc195692248)

Project Requirement Specifics

* Language: Python 3.6.1
* Editor: Jupyter Notebook
* Compiler: Anaconda 3
* Extension: ipynb

Uploading the python script

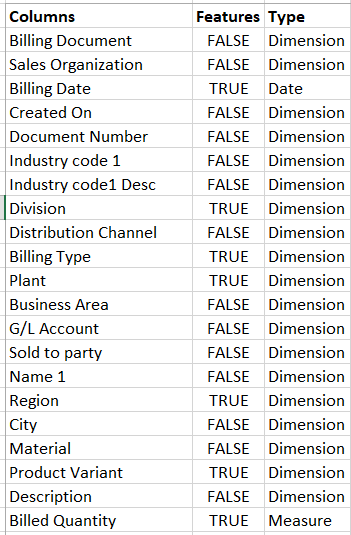
* Link for the anaconda compiler is below:
* <https://www.anaconda.com/distribution/>
* Once the compiler has been downloaded search for **Jupyter Notebook** from the range of editors.
* Open jupyter notebook and click on upload and search for the file **Outlier Detection for Sales using Unsupervised Learning.ipynb** .
* Once it has been uploaded click on it and start running the file by pressing Shift+Enter.

Sales Dataset Format

* The format should be in csv file
* The measures should be numerical format(not general)
* The measures should not have any errors e.g. DIV error
* The dimensions should not have too many categories or the model will not give correct outliers. Make sure the total number of categories are not more than 150-200

Mapping File Format

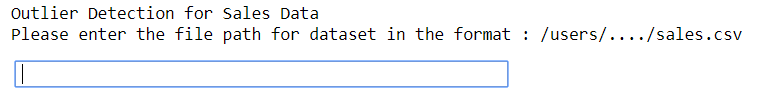
Add a mapping file in the csv extension to the input in the following format:



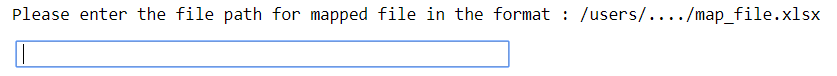
* Columns: This contains all the column names in the dataset
* Features: This contains True/False values where **True** is for that column to be considered as a feature for modelling
* Type: Contains 3 types of values.
* Dimension
* Measure
* Date

Steps for using the Python Script

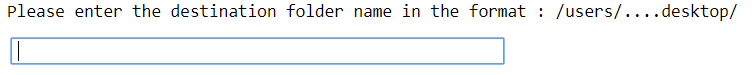
1. Specify the File Path for the dataset

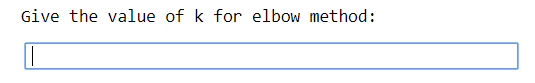


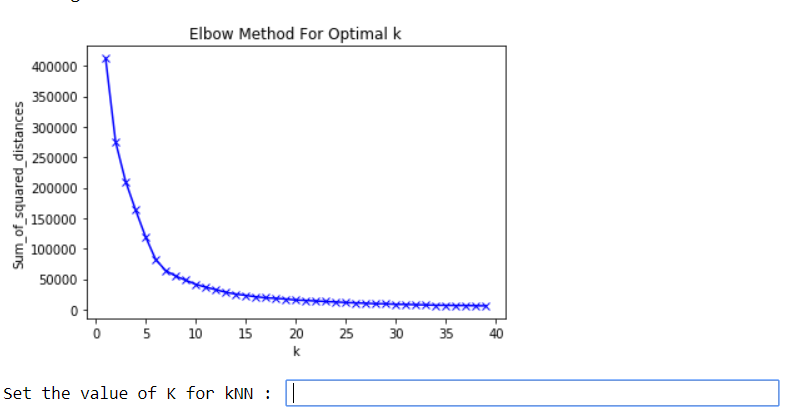
1. Specify the File Path for the mapping file



1. Specify the Destination folder Name



1. Specify the range of k for inertia graph 
2. Specify the Value of K by approximating the elbow from the inertia graph



1. Specify the contamination level for KNN. If contamination is 0.1 then it means 10% of the total records in the dataset are outliers

step 6

1. Specify the name of the output file containing the clusters and the prediction label as 0- normal points and 1-outlier.

step 7

1. Specify the extension of the output file e.g. csv, xlsx etc.

step 8

8. Specify the separator for the output file e.g. ( , )( tab )( | )step 9

After this the output will be stored in the destination folder provided in STEP 3.

**Offices of network firms of   
S.R. Batliboi & Affiliates**

**Kolkata**

22, Camac Street

3rd Floor, Block ‘B’,

Kolkata – 700 016

**Mumbai**

**Ruby**

12th Floor, The Ruby,

29 Senapati Bapat Marg

Dadar (West)

Mumbai – 400 028

**Goregaon**

5th Floor, Block-B-2,

Nirlon Knowledge

Park, Off W E Highway,

Goregaon (East),

Mumbai – 400 063

**New Delhi**

4th Floor, Office 405

World Mark - 2, Asset No. 8

IGI Airport Hospitality District,

Aerocity

New Delhi – 110 037

**Noida**

7th Floor, Plot No. 2B

Tower 3, Sector – 126

Noida – 201 304

Gautam Budh Nagar,

Uttar Pradesh

**Pune**

C-401, Fourth Floor,

Panchshil Tech Park,

Yerwada,

Pune – 411 006

**Ahmedabad**

Shivalik Ishan Building,

2nd Floor, Near CN Vidhyalaya,

Ambavadi,

Ahmedabad – 380 015

**Bengaluru**

12th Floor

“UB City” Canberra Block

No. 24, Vittal Mallya Road

Bengaluru – 560 001

**Chandigarh**

1st Floor, SCO: 166-167,

Madhya Marg, Sector 9-C

Chandigarh – 160 009

**Chennai**

6th Floor – “A” Block

Tidel Park

No. 4, Rajiv Gandhi Salai Taramani

Chennai – 600 113

**Gurgaon**

**GVT, 2nd and 3rd Floor,**

Golf View Tower B,

Sector – 42, Sector Road,

Gurugram – 122 002,

Haryana

**Tower A, Building No. 8,**

1st Floor, DLF Cyber City,

Phase – II, Sector - 25,

Gurgaon – 122 002

Haryana

**Hyderabad**

Plot No.18, OVAL Office

ILABS Centre,

Software Units Layout,

Hitec City, Madhapur

Hyderabad – 500 081

**Kochi**

9th Floor, Abad Nucleus

NH-49, Maradu

Kochi, Kerala – 682 304

**S R B C & Associates LLP**

S R B C & Associates LLP is a limited liability partnership firm of Chartered Accountants, registered with the Institute of Chartered Accountants of India, and having its registered office at 22 Camac Street, 3rd Floor, Block B, Kolkata – 700016.

© 2019 S R B C & Associates LLP. Published in India.

All Rights Reserved.

This publication contains information in summary form and is intended for general guidance only. It is not intended to be a substitute for detailed research or the exercise of professional judgment. Neither S R B C & Associates LLP nor any of its fellow network firms can accept any responsibility for loss occasioned to any person acting or refraining from action as a result of any material in this publication.

This publication is meant for distribution internally, to our clients or distributable on specific requests only.