READ ME Data Version 0.2

Content

Introduction 1
Release 1
Changes 2
Data Cleaning changes 2
Anomaly Handling changes 2
Data Preparation 2
UML 3
Parameters 4
Data Cleaning 4
Anomaly Detection 4
Anomaly Handling 4
Detectors with bad days 4
Script output 5

Introduction

This is the second version of the dataset. In this version we focused on some fine tuning and reducing the size of the dataset to make training more efficient.

Release

Samuel Paul 09.11.2024 14:00

Changes

There are changes to the data cleaning, anomaly handling and data preparation.

Data Cleaning changes

- Datapoint that are error flagged are dropped at the beginning
- Data clipping happens now on the column occ and flow, before the traffic value is calculated.
- There is a ne function that drops bad days
- The drop false value (Line detection) function is no longer used
- Some parameters were finetuned

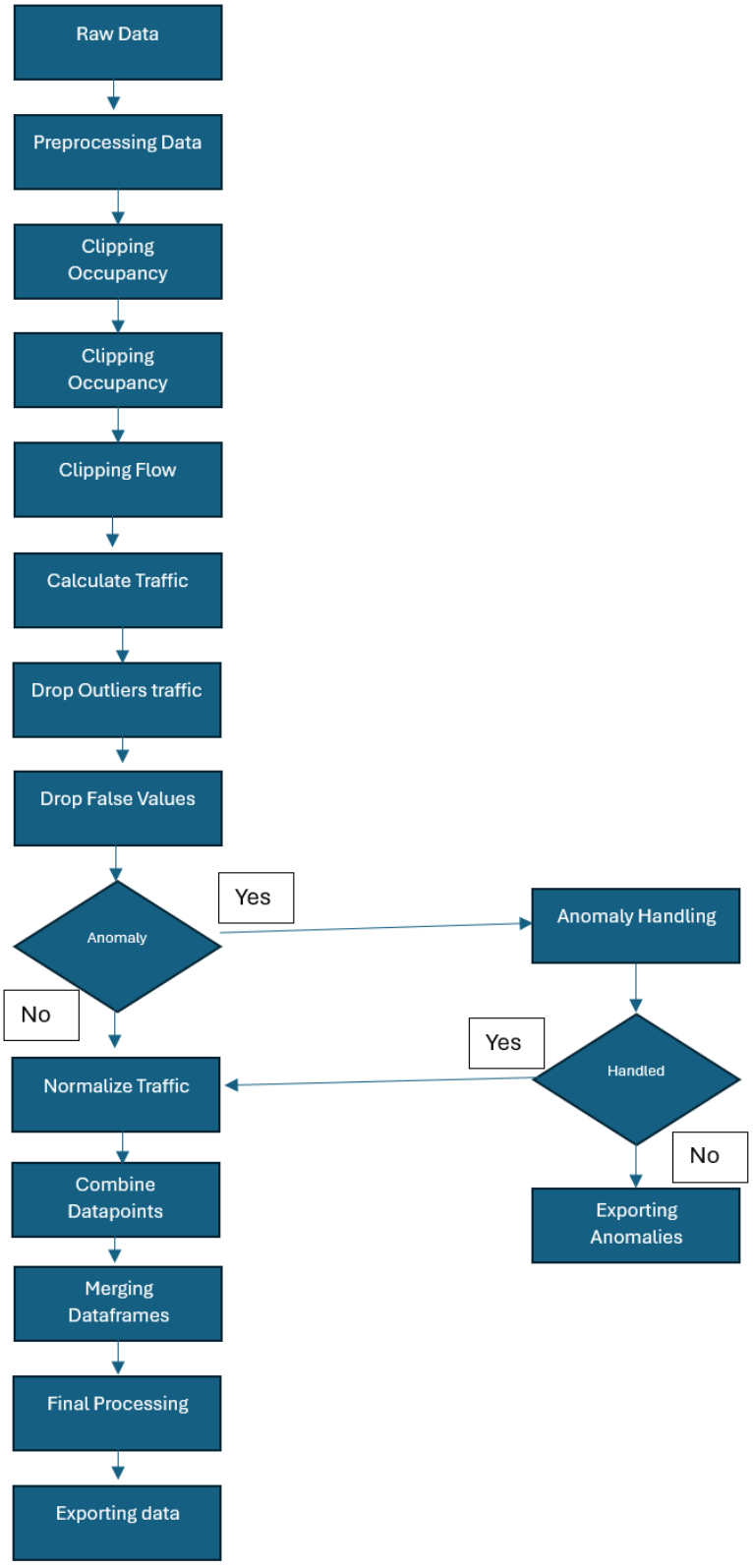
Anomaly Handling changes

- Only finetuning of some parameters

Data Preparation

- Datapoints are now combined to one datapoint per hour. Not longer every 5 minutes.

UML


Raw Data
Preprocessing Data
Clipping Occupancy
Clipping Occupancy
Clipping Flow
Calculate Traffic
Drop Outliers traffic
Drop False Values
Anomaly
Yes
Anomaly Handling
No
Yes
Handled
Normalize Traffic
No
Combine Datapoints
Exporting Anomalies
Merging Dataframes
Final Processing
Exporting data

Parameters

In the create_dataset.py script there are a lot of parameters that can be changed. Does are the ones that got set for this version.

Data Cleaning

Clipping occupancy outlier factor: 3

Clipping flow outlier factor: 3

Dropping traffic outlier factor: 2

Anomaly Detection

Mean out of bound factor: 3

IQR to small, min IQR: 5

Not enough data, min datapoints: 4000

Anomaly Handling

The anomaly handling has now only one function. There will be more in the future.

Detectors with bad days

Minimum datapoints per day: 230

Minimum good days: 10

Minimum one good day for every weekday

Script output

Starting script

Loading data from: C:\Users\samue\OneDrive\AIML\HS2024\Data Sicence Projekt\Data\London_UTD19.csv

Loading data from: C:\Users\samue\OneDrive\AIML\HS2024\Data Sicence Projekt\Data\London_detectors.csv

Data loaded

Preprocessing data

Errors found and dropped: 12234005

Preprocessing data took 9 seconds

Drop bad days

Total outliers detected and removed: 9504

Drop bad days took 137 seconds

Clipping outliers on occ

Total outliers clipped: 81

Clipping outliers on occ took 133 seconds

Clipping outliers on flow

Total outliers clipped: 57

Clipping outliers on flow took 134 seconds

Calculating traffic

Calculating traffic took 0 seconds

Droping outliers on traffic

Total outliers dropped: 661222

Droping outliers on traffic took 21 seconds

Drop false values

Total outliers detected and removed: 0

Drop false values took 132 seconds

Detecting anomalies

Anomalies detected based on IQR: 69

Anomalies detected based on IQR too small: 139

Anomalies detected based on not enough data: 1783

Detecting anomalies took 9 seconds

Handling anomalies

Anomalies with not enough data handled: 686

Total amount of dropeed anomalies: 1163

Handling anomalies took 308 seconds

Exporting anomalies to: C:\Users\samue\OneDrive\AIML\HS2024\Data Sicence Projekt\Data

Exporting anomalies took 0 seconds

Normalizing traffic

Normalizing traffic took 0 seconds

Combine datapoints

Combine datapoints took 6 seconds

Merging dataframes

Merging dataframes took 0 seconds

Final processing

Final processing took 0 seconds

Exporting modified dataset to: C:\Users\samue\OneDrive\AIML\HS2024\Data Sicence Projekt\Data

Exporting modified dataset took 14 seconds

Script finished

Total script execution time: 933 seconds