OUTLIER DETECTION

GUIDED BY:
DR. RAJATHILAGAM.B
ASSOCIATE PROFESSOR, CS DEPARTMENT

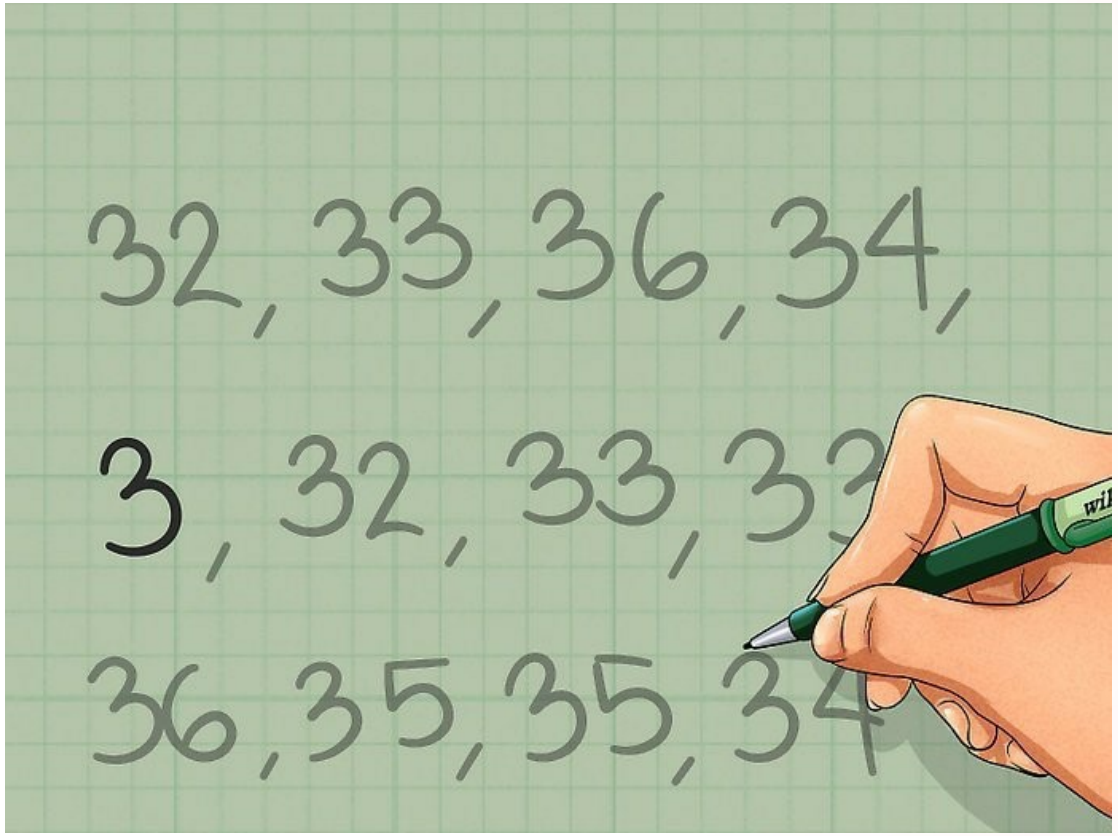
Presented by:
- Ankit.B.Saradagi (CB.EN.P2AID19003)
- Vivek. K. Nair (CB.EN.P2AID19029)
- Jishnu P (CB.EN.P2AID19017)

OUTLIER

- What is an Outlier?

In statistics, an outlier is an observation point that is distant from other observations.

The outliers can be a result of a mistake during data collection or it can be just an indication of variance in the data.



OUTLIERS GOOD OR BAD ??

- Outliers may contain valuable information. Or be meaningless aberrations caused by measurement and recording errors.
- They can cause problems in tests involving optimizing for revenue metrics, like Average Order Value or Revenue Per Visitor.
- Especially in data sets with low sample size, outliers can mess up the model's performance.
- But they are also useful in certain cases where the extreme values are essential for the overall dataset.

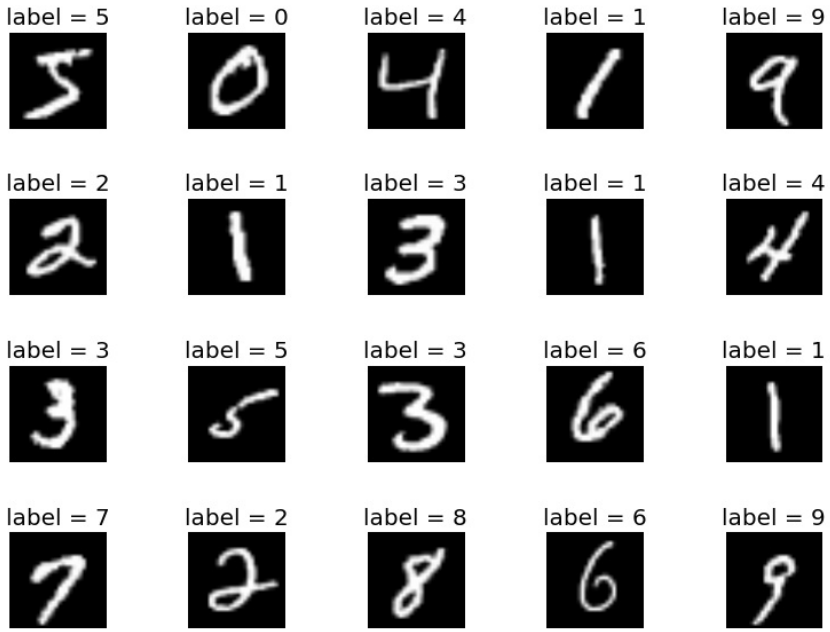
OUTLIER RATIO

- **Outlier Ratio = Outlier data points / Total data points**
- Usually, no. of Outlier data points will be negligible when compared to inliers. So, traditional methods eliminate the outliers.
- But there are cases where there will be a no. of Outlier categories. So, in total, they will make a good portion of the dataset.
- These are called **High Outlier Ratio** datasets & eliminating the outliers here will affect overall model performance.

EXISTING MODEL

- Consider datasets based on the outlier ratio

  Instead of "few" and "different" approach used by traditional methods.
- Based on image dataset.(mnist dataset).
- High robustness and cheaper computations for High Outlier Ratio datasets.



EXISTING MODEL ctd..

- Uses Low Rank Based Efficient Outlier Detection approach.
- Utilizes the low-rank structure embedded in the similarity matrix and evaluates outliers/inliers equally based on such low-rank structure, which lays the foundation to preserve .
- Good robustness under high outlier ratios with much cheaper computations.
- Uses Supervised learning.

OBJECTIVE

- To check the scope for further optimisation in the existing model.
- Analysing the existing model with different types of datasets and comparing their overall performance.

PERFORMANCE OF EXISTING MODEL

LEOD basic

```
Average results for outlier ratio 0.100000: Precision: 0.995706 Recall: 0.974835 F1: 0.985090 Time: 6.862240
Average results for outlier ratio 0.200000: Precision: 0.992965 Recall: 0.987283 F1: 0.990096 Time: 8.365227
Average results for outlier ratio 0.300000: Precision: 0.988212 Recall: 0.987700 F1: 0.987922 Time: 9.169225
Average results for outlier ratio 0.400000: Precision: 0.974506 Recall: 0.986860 F1: 0.980480 Time: 10.533128
Average results for outlier ratio 0.500000: Precision: 0.932941 Recall: 0.987330 F1: 0.958089 Time: 13.993142
Average results for outlier ratio 0.600000: Precision: 0.839016 Recall: 0.987645 F1: 0.901718 Time: 23.003132
```

LEOD fast

```
Average results for outlier ratio 0.100000: Precision: 0.987484 Recall: 0.992739 F1: 0.990054 Time: 0.602090
Average results for outlier ratio 0.200000: Precision: 0.991042 Recall: 0.993452 F1: 0.992226 Time: 0.708948
Average results for outlier ratio 0.300000: Precision: 0.980900 Recall: 0.991109 F1: 0.985845 Time: 0.858002
Average results for outlier ratio 0.400000: Precision: 0.958361 Recall: 0.991280 F1: 0.973931 Time: 1.070418
Average results for outlier ratio 0.500000: Precision: 0.932441 Recall: 0.990586 F1: 0.959367 Time: 1.470761
Average results for outlier ratio 0.600000: Precision: 0.811468 Recall: 0.991126 F1: 0.887475 Time: 2.067596
```

ANALYSIS OF THE PERFORMANCE

- Since the F1 score is good even in high outlier ratio , No need of further optimization.
- Analysed the performance of the LEOD model in different datasets.
- Dataset with Categorical Data
  - Wine_Quality Dataset
  - Titanic Dataset
- Time Series Data
  - Weather Dataset
- Compared the overall performance of both LEOD Basic and LEOD Fast algorithms.

LEOD Fast VS LEOD Basic

- WineQuality Dataset (wine_Quality.csv)

```
Average results for outlier ratio 0.100000: Precision: 0.990921 Recall: 0.872232 F1: 0.927597 Time: 1.837461
Average results for outlier ratio 0.200000: Precision: 0.980069 Recall: 0.871063 F1: 0.922237 Time: 2.227436
Average results for outlier ratio 0.300000: Precision: 0.943053 Recall: 0.838561 F1: 0.887739 Time: 2.481664
Average results for outlier ratio 0.400000: Precision: 0.879930 Recall: 0.878050 F1: 0.876128 Time: 2.537621
Average results for outlier ratio 0.500000: Precision: 0.747129 Recall: 0.816864 F1: 0.769235 Time: 2.711564
Average results for outlier ratio 0.600000: Precision: 0.648859 Recall: 0.746846 F1: 0.672201 Time: 2.931045
```

```
Average results for outlier ratio 0.100000: Precision: 0.991003 Recall: 0.825313 F1: 0.900500 Time: 219.039332
Average results for outlier ratio 0.200000: Precision: 0.980983 Recall: 0.907041 F1: 0.941900 Time: 307.735844
Average results for outlier ratio 0.300000: Precision: 0.965505 Recall: 0.817502 F1: 0.885030 Time: 325.745750
Average results for outlier ratio 0.400000: Precision: 0.908919 Recall: 0.912338 F1: 0.906541 Time: 327.947656
Average results for outlier ratio 0.500000: Precision: 0.765683 Recall: 0.828848 F1: 0.784037 Time: 351.317703
Average results for outlier ratio 0.600000: Precision: 0.626239 Recall: 0.661556 F1: 0.629495 Time: 373.955055
```

LEOD Fast VS LEOD Basic

- Titanic Dataset

```
Average results for outlier ratio 0.100000: Precision: 0.914772 Recall: 0.564704 F1: 0.681528 Time: 0.162488
Average results for outlier ratio 0.200000: Precision: 0.822557 Recall: 0.609411 F1: 0.690462 Time: 0.176029
Average results for outlier ratio 0.300000: Precision: 0.719622 Recall: 0.525585 F1: 0.603374 Time: 0.182762
Average results for outlier ratio 0.400000: Precision: 0.639682 Recall: 0.514737 F1: 0.563404 Time: 0.186276
Average results for outlier ratio 0.500000: Precision: 0.601073 Recall: 0.628149 F1: 0.586686 Time: 0.189356
```

```
Average results for outlier ratio 0.100000: Precision: 0.901067 Recall: 0.665624 F1: 0.748695 Time: 1.271389
Average results for outlier ratio 0.200000: Precision: 0.831515 Recall: 0.635381 F1: 0.707113 Time: 1.359694
Average results for outlier ratio 0.300000: Precision: 0.730401 Recall: 0.550240 F1: 0.624165 Time: 1.425564
Average results for outlier ratio 0.400000: Precision: 0.634465 Recall: 0.495013 F1: 0.549309 Time: 1.647950
Average results for outlier ratio 0.500000: Precision: 0.563667 Recall: 0.487123 F1: 0.511863 Time: 1.628552
```

LEOD Fast VS LEOD Basic

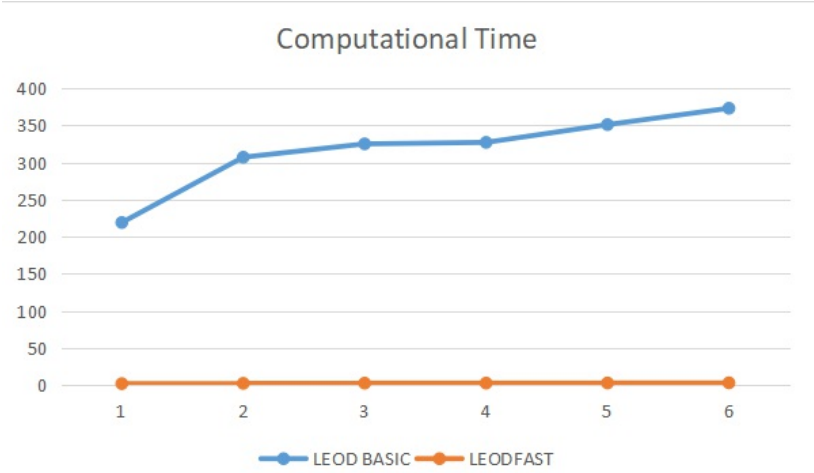
- Weather Data

```
Average results for outlier ratio 0.100000: Precision: 0.955547 Recall: 0.629574 F1: 0.751457 Time: 0.724012
Average results for outlier ratio 0.200000: Precision: 0.886188 Recall: 0.662829 F1: 0.753919 Time: 0.843768
Average results for outlier ratio 0.300000: Precision: 0.846905 Recall: 0.613004 F1: 0.708076 Time: 0.808599
Average results for outlier ratio 0.400000: Precision: 0.770785 Recall: 0.565330 F1: 0.647966 Time: 0.880490
Average results for outlier ratio 0.500000: Precision: 0.690610 Recall: 0.548385 F1: 0.599561 Time: 0.863801
```

```
Average results for outlier ratio 0.100000: Precision: 0.965250 Recall: 0.650421 F1: 0.771590 Time: 34.623352
Average results for outlier ratio 0.200000: Precision: 0.915675 Recall: 0.651262 F1: 0.751620 Time: 41.728278
Average results for outlier ratio 0.300000: Precision: 0.884316 Recall: 0.650589 F1: 0.736600 Time: 41.589248
Average results for outlier ratio 0.400000: Precision: 0.826395 Recall: 0.645849 F1: 0.704787 Time: 42.158036
Average results for outlier ratio 0.500000: Precision: 0.734961 Recall: 0.582307 F1: 0.624054 Time: 44.261546
```

SUMMARY

- LEOD-fast is computationally faster than LEOD-basic.
- Upto 60% outlier ratio datasets(IMAGE) - efficiently detects outlier.
- Upto 40% outlier ratio datasets(Category) - efficiently detects outlier.
- Upto 40% outlier ratio datasets(Time series) - efficiently detects outlier.



FUTURE SCOPE ...

- The current model can use some improvement when dealing with datasets involving categorical & time-series data.

THANK YOU !