**Outlier Analysis on Labeled Outlier Data**

**Note** :

TP -> Data points which are actually outliers detected as outliers , TN -> Data points which are not outliers not detected as outliers , FP -> Data points which not actually outliers detected as outliers, FN -> Data points which are actually outliers not detected as outliers

In order to check multivariate normality Shapiro wilk test and Jarque bera test has been used to obtain results.

1. **Simulated Outlier Dataset**

Total Data Points : 11000

Outliers : 10 % (1100)

Multivariate Normality : No

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Algorithm | Parameter | TP | FP | TN | FN | Precision | Recall |
| Isolation Forest | Contamination= 0.1 | 660 | 440 | 9460 | 440 | 0.92 | 0.6 |
| Contamination = ‘auto’ | 1100 | 2519 | 7381 | 0 | 0.771 | 1.0 |
| Local Outlier Factor | Contamination= 0.1,  n\_neighbors = 20 | 451 | 649 | 9251 | 649 | 0.882 | 0.41 |
| Contamination=’auto’,  n\_neighbors = 20 | 308 | 220 | 9680 | 792 | 0.908 | 0.28 |
| Contamination= 0.1,  n\_neighbors = 50 | 902 | 198 | 9702 | 198 | 0.964 | 0.82 |
| Contamination= 0.1,  n\_neighbors = 100 | 968 | 132 | 9768 | 132 | 0.976 | 0.88 |
| One Class SVM | Kernel= ‘rbf’, Nu = 0.1 | 451 | 638 | 9262 | 649 | 0.883 | 0.41 |
| Kernel = ‘linear’,Nu = 0.1 | 154 | 957 | 8943 | 946 | 0.827 | 0.14 |

1. **Forest Cover Dataset**

Total Data Points : 286048

Outliers : 0.9 % (2747)

Multivariate Normality : No

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Algorithm | Parameters | TP | FP | TN | FN | Precision | Recall |
| Isolation Forest | Contamination = ‘auto’ | 2392 | 37125 | 246176 | 355 | 0.869 | 0.871 |
| Contamination = 0.009 | 233 | 2342 | 280959 | 2514 | 0.983 | 0.085 |
| Local Outlier Factor | Contamination = 0.009, n\_neighbors = 20 | 125 | 2450 | 280851 | 2622 | 0.982 | 0.046 |
| Contamination = ‘auto’, n\_neighbors = 20 | 51 | 847 | 282454 | 2696 | 0.988 | 0.019 |
| Contamination = 0.009, n\_neighbors =50 | 104 | 2471 | 280830 | 2643 | 0.982 | 0.038 |
| Contamination = 0.009, n\_neighbors =100 | 114 | 2461 | 280840 | 2633 | 0.982 | 0.041 |
| One Class SVM | Kernel = ‘linear’, Nu = 0.009 | 1376 | 1198 | 282103 | 1371 | 0.991 | 0.501 |
| Kernel = ‘rbf’, Nu = 0.09 | 188 | 2387 | 280914 | 2559 | 0.983 | 0.068 |

1. **Shuttle Dataset**

Total Data Points : 49097

Outliers : 7 % (3511)

Multivariate Normality : No

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Algorithm | Parameters | TP | FP | TN | FN | Precision | Recall |
| Isolation Forest | Contamination = 0.07 | 3239 | 196 | 45390 | 272 | 0.99 | 0.923 |
| Contamination = ‘auto’ | 3452 | 1855 | 43731 | 59 | 0.961 | 0.983 |
| Local Outlier Factor | Contamination = 0.07, n\_neighbors = 20 | 383 | 3054 | 42532 | 3128 | 0.874 | 0.109 |
| Contamination = ‘auto’, n\_neighbors = 20 | 214 | 1031 | 44555 | 3297 | 0.912 | 0.061 |
| Contamination = 0.07, n\_neighbors =50 | 554 | 2883 | 42703 | 2957 | 0.881 | 0.158 |
| Contamination = 0.07, n\_neighbors =100 | 697 | 2740 | 42846 | 2814 | 0.887 | 0.199 |
| One Class SVM | Kernel = ‘linear’, Nu = 0.07 | 16 | 3421 | 42165 | 3495 | 0.859 | 0.005 |
| Kernel = ‘rbf’, Nu = 0.07 | 1972 | 1467 | 44119 | 1539 | 0.939 | 0.562 |

1. **Satellite Dataset**

Total Data Points : 6435

Outliers : 32 % (2036)

Multivariate Normality : No

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Algorithm | Parameters | TP | FP | TN | FN | Precision | Recall |
| Isolation Forest | Contamination = ‘auto’ | 931 | 188 | 4211 | 1105 | 0.799 | 0.457 |
| Contamination = 0.32 | 1119 | 940 | 3459 | 917 | 0.711 | 0.55 |
| Local Outlier Factor | Contamination = 0.32, n\_neighbors = 20 | 760 | 1299 | 3100 | 1276 | 0.6 | 0.373 |
| Contamination = ‘auto’, n\_neighbors = 20 | 69 | 62 | 4337 | 1967 | 0.685 | 0.034 |
| Contamination = 0.32, n\_neighbors =50 | 135 | 134 | 4265 | 1901 | 0.684 | 0.066 |
| Contamination = 0.32, n\_neighbors =100 | 200 | 216 | 4183 | 1836 | 0.681 | 0.098 |
| One Class SVM | Kernel = ‘linear’, Nu = 0.32 | 736 | 1324 | 3075 | 1300 | 0.592 | 0.361 |
| Kernel = ‘rbf’, Nu = 0.32 | 1079 | 979 | 3420 | 957 | 0.699 | 0.53 |

1. **Satimage-2 Dataset**

Total Data Points : 5803

Outliers : 1.2 % (71)

Multivariate Normality : No

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Algorithm | Parameters | TP | FP | TN | FN | Precision | Recall |
| Isolation Forest | Contamination = 0.012 | 62 | 8 | 5724 | 9 | 0.997 | 0.873 |
| Contamination = ‘auto’ | 68 | 498 | 5234 | 3 | 0.914 | 0.958 |
| Local Outlier Factor | Contamination = 0.012, n\_neighbors = 20 | 5 | 65 | 5667 | 66 | 0.977 | 0.07 |
| Contamination = ‘auto’, n\_neighbors = 20 | 11 | 127 | 5605 | 60 | 0.968 | 0.155 |
| Contamination = 0.012, n\_neighbors =50 | 4 | 66 | 5666 | 67 | 0.977 | 0.056 |
| Contamination = 0.012, n\_neighbors =100 | 5 | 65 | 5667 | 66 | 0.977 | 0.07 |
| One Class SVM | Kernel = ‘linear’, Nu =0.012 | 0 | 69 | 5663 | 71 | 0.976 | 0.0 |
| Kernel = ‘rbf’, Nu = 0.012 | 20 | 48 | 5684 | 51 | 0.983 | 0.282 |

1. **Mammography Dataset**

Total Data Points : 11183

Outliers : 2.32 % (260)

Multivariate Normality : No

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Algorithm | Parameters | TP | FP | TN | FN | Precision | Recall |
| Isolation Forest | Contamination = 0.0232 | 54 | 206 | 10717 | 206 | 0.963 | 0.208 |
| Contamination = ‘auto’ | 157 | 1277 | 9646 | 103 | 0.877 | 0.604 |
| Local Outlier Factor | Contamination = 0.0232, n\_neighbors = 20 | 50 | 210 | 1071 | 210 | 0.962 | 0.192 |
| Contamination = ‘auto’, n\_neighbors = 20 | 40 | 149 | 1077 | 220 | 0.967 | 0.154 |
| Contamination = 0.0232, n\_neighbors =50 | 65 | 195 | 1072 | 195 | 0.965 | 0.25 |
| Contamination = 0.0232, n\_neighbors =100 | 61 | 199 | 1072 | 199 | 0.964 | 0.235 |
| One Class SVM | Kernel = ‘linear’, Nu= 0.0232 | 141 | 1052 | 400 | 191 | 0.048 | 0.542 |
| Kernel= ‘rbf’, Nu = 0.0232 | 64 | 3516 | 7407 | 196 | 0.668 | 0.246 |

1. **Ann thyroid Dataset**

Total Data Points : 7200

Outliers : 7.42 % (534)

Multivariate Normality : No

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Algorithm | Parameters | TP | FP | TN | FN | Precision | Recall |
| Isolation Forest | Contamination = 0.0742 | 180 | 355 | 6311 | 354 | 0.902 | 0.337 |
| Contamination = ‘auto’ | 199 | 501 | 6165 | 335 | 0.884 | 0.373 |
| Local Outlier Factor | Contamination = ‘auto’, n\_neighbors = 20 | 92 | 193 | 6473 | 442 | 0.912 | 0.172 |
| Contamination = 0.0742, n\_neighbors = 20 | 157 | 378 | 6288 | 377 | 0.895 | 0.294 |
| Contamination = 0.0742, n\_neighbors =50 | 168 | 367 | 6299 | 366 | 0.898 | 0.315 |
| Contamination = 0.0742, n\_neighbors =100 | 174 | 361 | 6305 | 360 | 0.9 | 0.326 |
| One Class SVM | Kernel = ‘linear’, Nu= 0.0742 | 47 | 488 | 6178 | 487 | 0.865 | 0.088 |
| Kernel = ‘rbf’, Nu= 0.0742 | 97 | 438 | 6228 | 437 | 0.878 | 0.182 |

1. **Thyroid Dataset**

Total Data Points : 3772

Outliers : 2.5 % (93)

Multivariate Normality : No

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Algorithm | Parameters | TP | FP | TN | FN | Precision | Recall |
| Isolation Forest | Contamination = 0.025 | 50 | 45 | 3634 | 43 | 0.977 | 0.538 |
| Contamination = ‘auto’ | 85 | 270 | 3409 | 8 | 0.926 | 0.914 |
| Local Outlier Factor | Contamination = 0.025, n\_neighbors = 20 | 18 | 77 | 3602 | 75 | 0.96 | 0.194 |
| Contamination = ‘auto’, n\_neighbors = 20 | 24 | 126 | 3553 | 69 | 0.948 | 0.258 |
| Contamination = 0.025, n\_neighbors =50 | 26 | 69 | 3610 | 67 | 0.964 | 0.28 |
| Contamination = 0.025, n\_neighbors =100 | 37 | 58 | 3621 | 56 | 0.97 | 0.398 |
| One Class SVM | Kernel = ‘linear’, Nu = 0.025 | 17 | 77 | 3602 | 76 | 0.959 | 0.183 |
| Kernel = ‘rbf’, Nu = 0.025 | 21 | 75 | 3604 | 72 | 0.961 | 0.226 |