# Anomaly detection

## Preprocessing

The given dataset has 1567 examples and 591 features. Ratio between normal examples and anomalies is 1:14. All variables are numeric. Firstly, missing values have been filled with mean value for each feature. Secondly, it's important to reduce dimensionality. I've found about hundred features with constant values for every example. I have also dropped every feature with variance lower than 0.03 and only 53 features remained.

## Anomaly detection system

There are lots of anomaly detecion techniques, such as: classification based, NN based, clustering based, etc. Firstly, I chose algorithm from clustering based technique called DBSCAN. It's very powerful density-based algorithm but it didn't perform well on my dataset. The results were poor so I had to change the algorithm. The much better results were after applying One-class SVM algorithm. It's important to note that accuarcy is not a good measure. The most important measure for this problem is recall for anomalies because it's important to detect the fault in a timely manner. Classification reports for these two implemented algorithms are shown in next two tables (table 1 and table 2).

Table – Classification report for DBSCAN algorithm

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| Outlier | 0.07 | 0.51 | 0.13 | 104 |
| Not outlier | 0.94 | 0.54 | 0.69 | 1463 |
| Avg/total | 0.88 | 0.54 | 0.65 | 1567 |

Table – Classification report for One-class SVM algorithm

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| Outlier | 0.58 | 1.00 | 0.74 | 104 |
| Not outlier | 1.00 | 0.95 | 0.97 | 1463 |
| Avg/total | 0.97 | 0.95 | 0.96 | 1567 |

Providing the normal training data, One-class SVM algorithm creates a model of this data. If new example is too different from other (training) examples, model will label this new example as anomaly.