**CASE STUDY ON**

ANOMALY DETECTION (GERMAN CREDIT RISK)

**Objective of Analysis:**

**Minimization of risk and maximization of profit on behalf of the bank.**

**Data:**

The German Credit Data contains data on 20 variables and the classification whether an applicant is considered a Good or a Bad credit risk for 1000 loan applicants. A predictive model developed on this data is expected to provide a bank manager guidance for making a decision whether to approve a loan to a prospective applicant based on his/her profiles.

**Description of Data:**

Training data consists of credit card applications. The dataset consists of 1000 entries with 20 feature columns and 1 label column. The features are a mix of numeric and categorical types. The original data and the description of individual columns can be found in the [UCI data repository](https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/). In order to train the anomaly detectors, **we will be relying on the 'normal' category (label value = 1) while ignoring the 'risky' category (label value = 2).** However, we will be using both categories for evaluating the detector. Each variable description,

1. Creditability,
2. Account Balance,
3. Duration of Credit (month),
4. Payment Status of Previous Credit,
5. Purpose,
6. Credit Amount,
7. Value Savings/Stocks,
8. Length of current employment,
9. Instalment per cent,
10. Sex & Marital Status,
11. Guarantors,
12. Duration in Current address,
13. Most valuable available asset,
14. Age (years),
15. Concurrent Credits,
16. Type of apartment,
17. No of Credits at this Bank,
18. Occupation,
19. No of dependents,
20. Telephone,
21. Foreign Worker

**Algorithm Used:**

* PCA – Principal component analysis:

Linear dimensionality reduction using Singular Value Decomposition of the data to project it to a lower dimensional space. Basically pca used to pre-process the data.

* Gaussian Mixtures (Clustering algorithm):

A multivariate normal distribution or multivariate Gaussian distribution is a generalization of the one-dimensional Gaussian distribution into multiple dimensions.

* Logistic Regression (A way for binary classification)

**Work Flow:**

Load Dataset

Apply Vector Assembler & Satandar Scalar

Apply PCA algorithm

Apply Random Split [train,test]

Train Dataset

Test Dataset

Apply Gaussian Mixture Classifier

Apply LabeledPoint

Created Model

Predict on test data

Evaluation of data :Plot Scatter plot on predicted and label

Apply Logistic Regress on train data

Created Model

Evaluate on test data :Predict on test data &

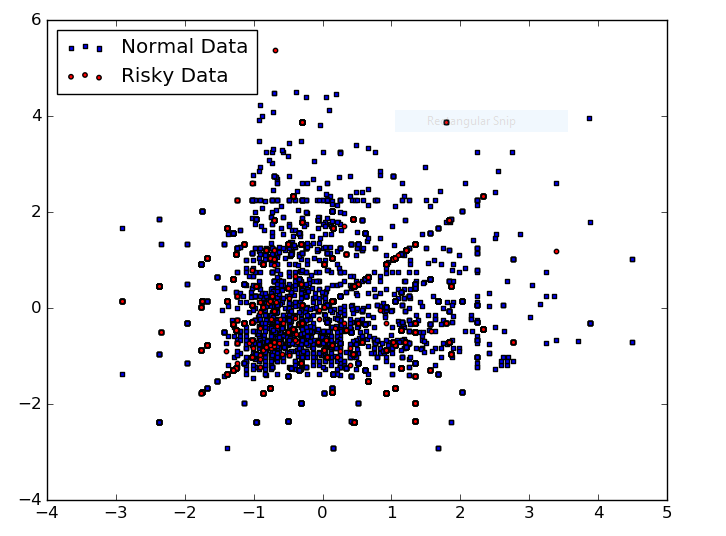
Here we found error & Accuracy of classifier

Pre-Process Data here

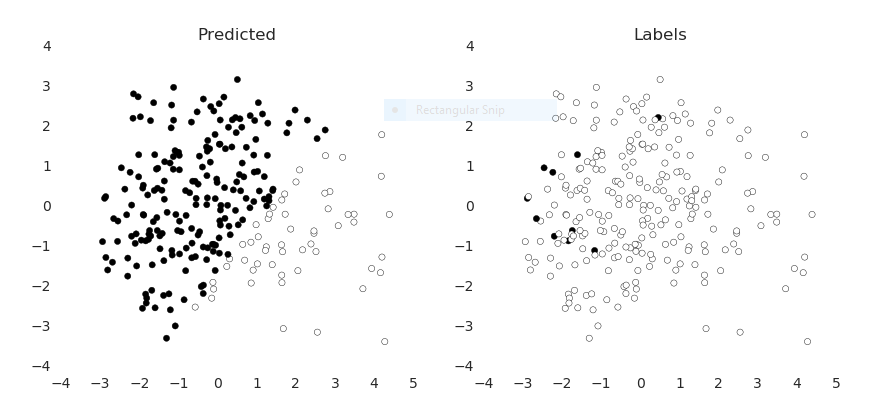
**Final Result : Confusion matrix of the classifier & plot ROC curve**

**Basic Visualization of Data:**

**1. Normal & Risky data for German credit dataset:**



**2. Result on evaluation data of test After Applying PCA & Gaussian mixtures on this dataset (basically focused on features)**

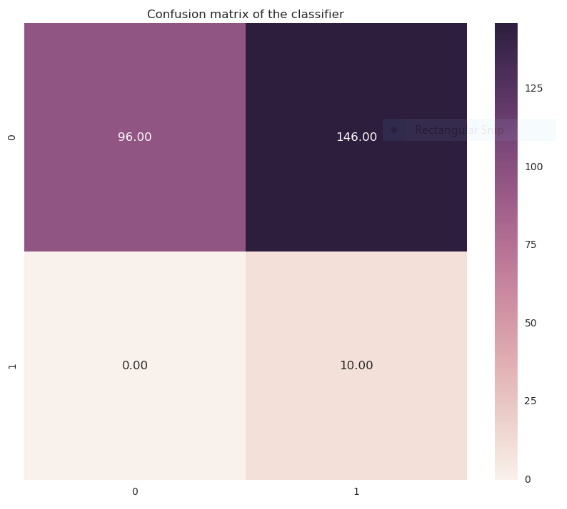


* When applied Logistic Regression on binary data and proceed with the same process. So the result we get the error value in our data.

**Error = 0.579365079365**

**3. Final result:**

* Confusion matrix of the classifier: A confusion matrix is a table that is often used to **describe the performance of a classification model** (or "classifier") on a set of test data for which the true values are known.



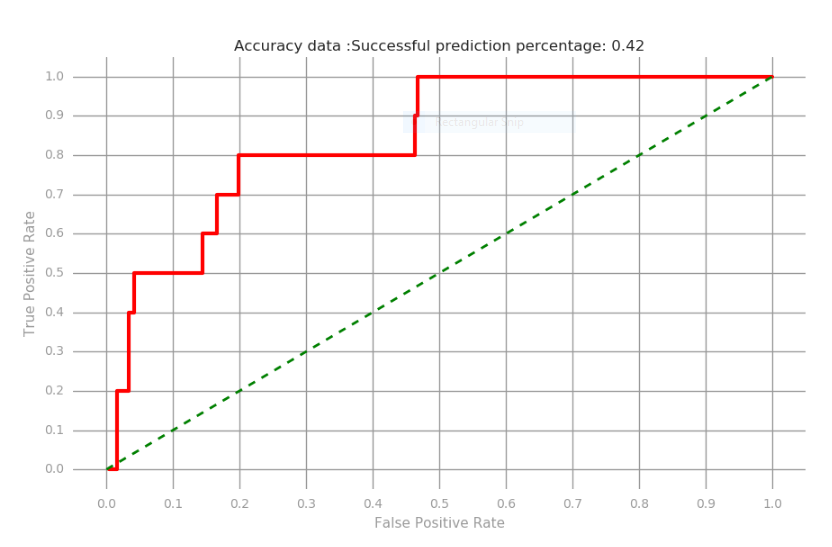
Let’s go through the Confusion Matrix Result what it will describe here,

|  |  |  |
| --- | --- | --- |
| **N=252** | **Predicted NO:** | **Predicted: Yes** |
| **Actual NO:** | **96** | **146** |
| **Actual Yes** | **0** | **10** |

**Form these values we can calculate Error & Accuracy**

* Accuracy data: Successful prediction percentage: **0.4206**
* Error: Error value in our data: **0.579365079365**

**Evaluate** Anomaly detectors can be evaluated on the same metrics as binary classifiers. Area under the ROC curve provides a good way to measure the discriminatory power of the anomaly detectors.



**Conclusions:**

PCA is a method for multivariate data analysis and it is used in many fields to extract relevant information from confusing data sets. Also, PCA provides a way of identifying patterns in data, and of expressing the data in a way that highlights their similarities and differences between them.

An advantage of using PCA consists in quantifying the importance of each dimension for describing the variability of a data set. This method reduces the number of dimensions, without much loss of information.

PCA is preferable for large scale data.