BLOG POST

Authors:

Birwa Galia, galia.b@husky.neu.edu

Milony Mehta, mehta.mil@husky.neu.edu

Shantanu Deosthale, deosthale.s@husky.neu.edu

Heading: Anomaly Detection and Outlier Analysis

Case Study: Credit Card Fraud Detection

This overview is intended for beginners in the fields of data science and machine learning. Almost no formal professional experience is needed to follow along, but the reader should have some basic knowledge of calculus (specifically integrals), the programming language Python, functional programming, and machine learning.

Anomaly detection is a technique used to identify unusual patterns that do not conform to expected behavior, called outliers. It has many applications in business, from intrusion detection (identifying strange patterns in network traffic that could signal a hack) to system health monitoring (spotting a malignant tumor in an MRI scan), and from fraud detection in credit card transactions to fault detection in operating environments.

An outlier is a data point which is significantly different from the remaining data. Hawkins formally defined the concept of an outlier as follows: “An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.” Outliers are also referred to as abnormalities, discordant, deviants, or anomalies in the data mining and statistics literature. In most applications, the data is created by one or more generating processes, which could either reflect activity in the system or observations collected about entities. When the generating process behaves in an unusual way, it results in the creation of outliers.

Presentation Link:

https://www.slideshare.net/ShantanuDeosthale/outlier-analysis-and-anomaly-detection

Google Drive:

Github Link:

https://github.com/shaan06/AnomalyDetection-and-OutlierAnalysis-Case-Study.git

Outlier Analysis:

An outlier is a data point which is significantly different from the remaining data. Hawkins formally defined [205] the concept of an outlier as follows: “An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.”

Outliers are also referred to as abnormalities, discordant, deviants, or anomalies in the data mining and statistics literature. In most applications, the data is created by one or more generating processes, which could either reflect activity in the system or observations collected about entities. When the generating process behaves in an unusual way, it results in the creation of outliers.

They are data points that are considered out of the ordinary or abnormal. When the generating process behaves in an unusual way, it results in the creation of outliers. Therefore, an outlier often contains useful information about abnormal characteristics of the systems and entities, which impact the data generation process. The recognition of such unusual characteristics provides useful application-specific insights. Some examples are,

- **Intrusion Detection Systems:** In many host-based or networked computer systems, different kinds of data are collected about the operating system calls, network traffic, or other activity in the system. This data may show unusual behavior because of malicious
- **Credit Card Fraud:** Credit card fraud is quite prevalent, because of the ease with which sensitive information such as a credit card number may be compromised. This typically leads to unauthorized use of the credit card. In many cases, unauthorized use may show different patterns, such as a buying spree from geographically obscure locations. Such patterns can be used to detect outliers in credit card transaction data.
- **Interesting Sensor Events:** Sensors are often used to track various environmental and location parameters in many real applications. The sudden changes in the underlying patterns may represent events of interest. Event detection is one of the

- **Medical Diagnosis:** In many medical applications the data is collected from a variety of devices such as MRI scans, PET scans or ECG time-series. Unusual patterns in such data typically reflect disease conditions
- **Law Enforcement:** Outlier detection finds numerous applications to law enforcement, especially in cases, where unusual patterns can only be discovered over time through multiple actions of an entity. Determining fraud in financial transactions, trading activity, or insurance claims typically requires the determination of unusual patterns in the data generated by the actions of the criminal entity.
- **Earth Science:** A significant amount of spatiotemporal data about weather patterns, climate changes, or land cover patterns is collected through a variety of mechanisms such as satellites or remote sensing. Anomalies in such data provide significant insights about hidden human or environmental trends, which may have caused such anomalies.

  In all these applications, the data has a "normal" model, and anomalies are recognized as deviations from this normal model. In many cases such as intrusion or fraud detection, the outliers can only be discovered as a sequence of multiple data points, rather than as an individual data point. For example, a fraud event may often reflect the actions of an individual in a particular

  sequence. The specificity of the sequence is relevant to identifying the anomalous event.

Types of Outlier Analysis:

Univariate and Multivariate Outliers:

In this post we will discuss univariate and multivariate outliers. A univariate outlier is a data point that consists of an extreme value on one

variable. A multivariate outlier is a combination of unusual scores on at least two variables. Both types of outliers can influence the outcome of statistical analyses. Outliers exist for four reasons. Incorrect data entry can cause data to contain extreme cases. A second reason for outliers can be failure to indicate codes for missing values in a dataset. Another possibility is that the case did not come from the intended sample. And finally, the distribution of the sample for specific variables may have a more extreme distribution than normal.

In many parametric statistics, univariate and multivariate outliers must be removed from the dataset. When looking for univariate outliers for continuous variables, standardized values (*z* scores) can be used. If the statistical analysis to be performed does not contain a grouping variable, such as linear regression, canonical correlation, or SEM among others, then the data set should be assessed for outliers as a whole. If the analysis to be conducted does contain a grouping variable, such as MANOVA, ANOVA, ANCOVA, or logistic regression, among others, then data should be assessed for outliers separately within each group. For continuous variables, univariate outliers can be considered standardized cases that are outside the absolute value of 3.29. However, caution must be taken with extremely large sample sizes, as outliers are expected in these datasets. Once univariate outliers have been removed from a dataset, multivariate outliers can be assessed for and removed.

Multivariate outliers can be identified with the use of Mahalanobis distance, which is the distance of a data point from the calculated centroid of the other cases where the centroid is calculated as the intersection of the mean of the variables being assessed. Each point is recognized as an X, Y combination and multivariate outliers lie a given distance from the other cases. The distances are interpreted using a $p <$

.001 and the corresponding χ^2 value with the degrees of freedom equal to the number of variables. Multivariate outliers can also be recognized using leverage, discrepancy, and influence. Leverage is related to Mahalanobis distance but is measured on a different scale so that the χ^2 distribution does not apply. Large scores indicate the case if further out however may still lie on the same line. Discrepancy assesses the extent that the case is in line with the other cases. Influence is determined by leverage and discrepancy and assesses changes in coefficients when cases are removed. Cases > 1.00 are likely to be considered outliers.

Anomaly Detection:

In data mining, **anomaly detection** (also **outlier detection**) is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset. Typically, the anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or errors in a text. Anomalies are also referred to as outliers, novelties, noise, deviations and exceptions

In particular, in the context of abuse and network intrusion detection, the interesting objects are often not *rare* objects, but unexpected *bursts* in activity. This pattern does not adhere to the common statistical definition of an outlier as a rare object, and many outlier detection methods (in particular unsupervised methods) will fail on such data, unless it has been aggregated appropriately. Instead, a cluster analysis algorithm may be able to detect the micro clusters formed by these patterns.

Three broad categories of anomaly detection techniques exist. **Unsupervised anomaly detection** techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the

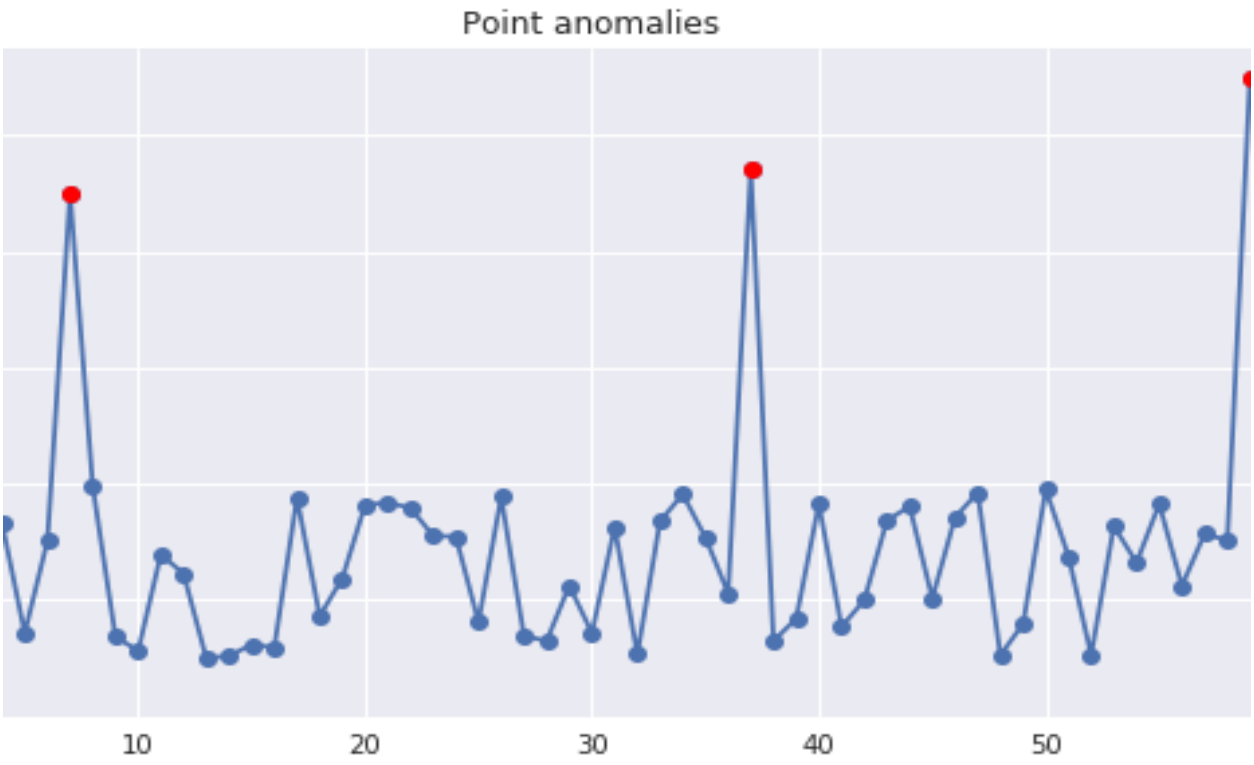
instances in the data set are normal by looking for instances that seem to fit least to the remainder of the data set. **Supervised anomaly detection** techniques require a data set that has been labeled as "normal" and "abnormal" and involves training a classifier (the key difference to many other statistical classification problems is the inherent unbalanced nature of outlier detection). **Semi-supervised anomaly detection** techniques construct a model representing normal behavior from a given *normal* training data set, and then testing the likelihood of a test instance to be generated by the learnt model.

Point Anomaly:

In an instance is anomalous compared with the rest of instances, the anomaly is considered as point anomaly.

- *Business use case:* Detecting credit card fraud based on "amount spent“

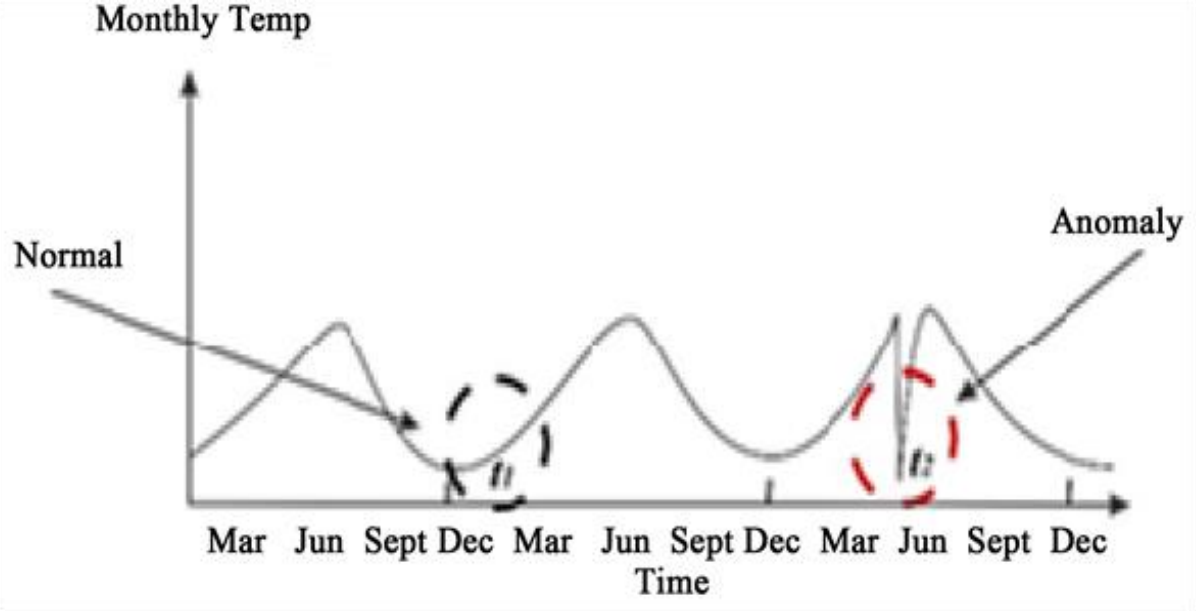
Purchase with large transaction value, Transaction of $50000 with no previous record of transactions more than that $1000



Contextual Anomaly:

It is specific-context based anomaly. Observation that is unusual in a certain context but not in entire context as a whole

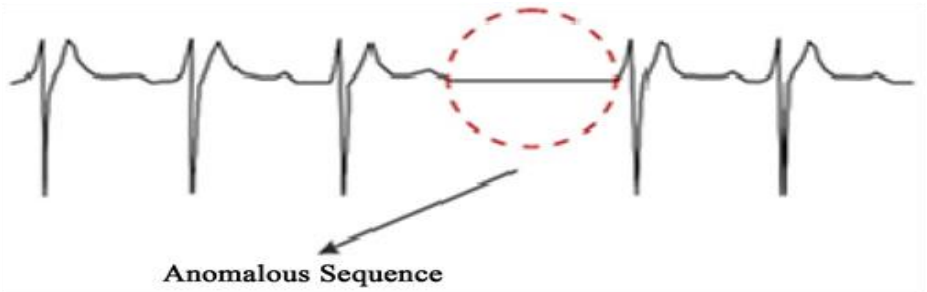
- *Business use case:* Spending $100 on food every day during the holiday season is normal, but may be odd otherwise
- As you see from the image the t1 anomaly is allowed, as it the same in the whole context of the data flow, while temperature t2 is an anomaly which is not the same in the whole context hence, this kind of anomaly is known as Contextual Anomaly



Collective Anomaly:

A set of data instances collectively helps in detecting anomalies.

- *Business use case:* Someone is trying to copy data form a remote machine to a local host unexpectedly, an anomaly that would be flagged as a potential cyber attack.
- Multiple Buy Stock transactions and then a sequence of sell transactions around an earnings release date may be anomalous and may indicate insider trading
- Multiple http requests from an ip address may indicate a probable web attack.



Applications of Anomaly Detection:

- **Intrusion Detection**

  It was introduced to detect unknown attacks, in part due to the rapid development of malware. The basic approach is to use machine learning to create a model of trustworthy activity, and then compare new behavior against this model. Although this approach enables the detection of previously unknown attacks, it may suffer from false positives: previously unknown legitimate activity may also be classified as malicious.

- **Fraud Detection**

  It was introduced to detect fraud data from large collection of data. The basic approach is to use machine learning algorithm which can classify the abnormal data from normal data, example can be Credit card fraud detection.

- **Fault Detection**

  It was introduced to detect false or invalid data from large collection of data. It uses machine learning algorithm which classifies the incorrect data from normal data, example can be Wireless Sensor Fault Detection

- **Detecting Ecosystem Disturbances**

  It was introduced to detect large-scale ecosystem disturbances with natural and anthropogenic causes using low-resolution data, and to detect changing land cover patterns using high-resolution data.

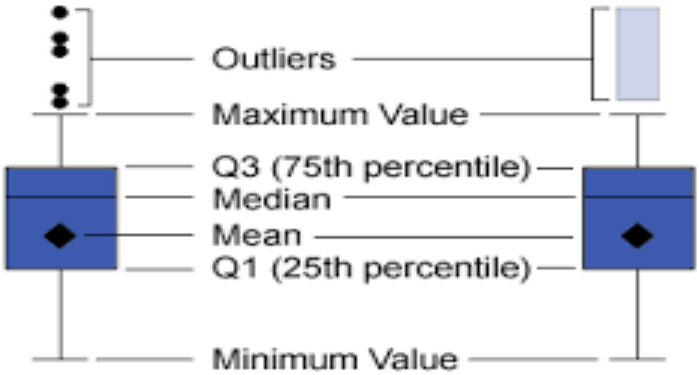
Methodologies to Anomaly detection:

- **Graphical Approach**

  Graphical methods utilize extreme value analysis, by which outliers correspond to the statistical tails of probability distributions. There are various plots which can be done to detect anomalies.

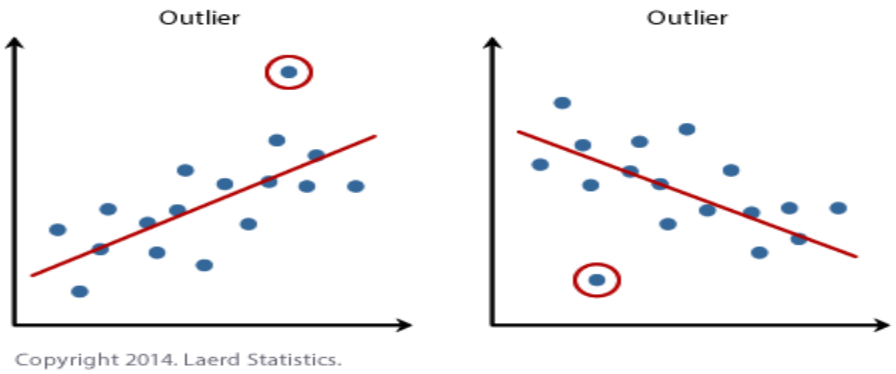
1. **Box plot**

   In this kind of plot, the data is divided into quartiles which include minimum, first quartile, median, third quartile, and maximum. The points lying between minimum and maximum are not outliers. The points outside can be termed as outlier which lead to anomaly.



2. **Scatter plot**

   A mathematical diagram, which uses Cartesian coordinates for plotting ordered pairs to show the correlation between typically two random variables. An outlier is defined as a data point that doesn't seem to fit with the rest of the data points.



- **Statistical Approach**

  Statistical methods utilizes Mathematical concepts, formulas, models, techniques to detect which data is an outlier among the dataset.

  1. Hypothesis Test
     a. **Chi-square Test**

        Chi-square test performs a simple test for detecting outliers of univariate data based on Chi-square distribution of squared difference between data and sample mean. Chi-square test helps us identify the lowest and highest values, since outliers can exist in both tails of the data.

     b. **Grubb's Test**

        Its test for outliers for univariate data sets assumed to come from a normally distributed population. Grubbs' test detects one outlier at a time. This outlier is expunged from the dataset and the test is iterated until no outliers are detected.

  2. Scores
     a. **Z-score**

        In Z-score, the data is given in units of how many standard deviations it is from the mean. Z-scores to identify possible

outliers to detect anomaly, this can be misleading (particularly for small sample sizes) due to the fact that the maximum Z-score is at most (n−1)/sqrt(n) where n is total no. of data.

b. IQR score

The interquartile range (IQR), also called the mid-spread or middle 50%, or technically H-spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles IQR = Q3 − Q1. Outlier is any data point more than 1.5 interquartile ranges (IQRs) below the first quartile or above the third quartile which can be helpful in detecting anomaly.

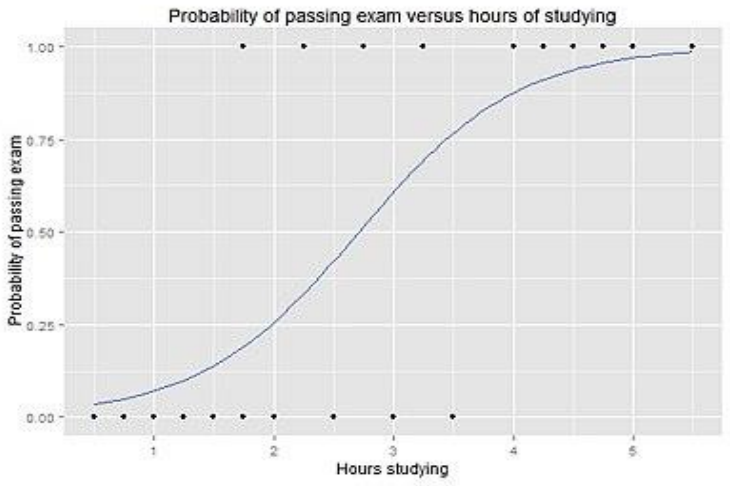
- **Machine Learning Approach**

a. Supervised Learning

This algorithm consists of a target / outcome variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables). Using these set of variables, we generate a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data. Examples of Supervised Learning: Regression, Decision Tree, Random Forest, KNN, Logistic Regression.

Logistic Regression:

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes

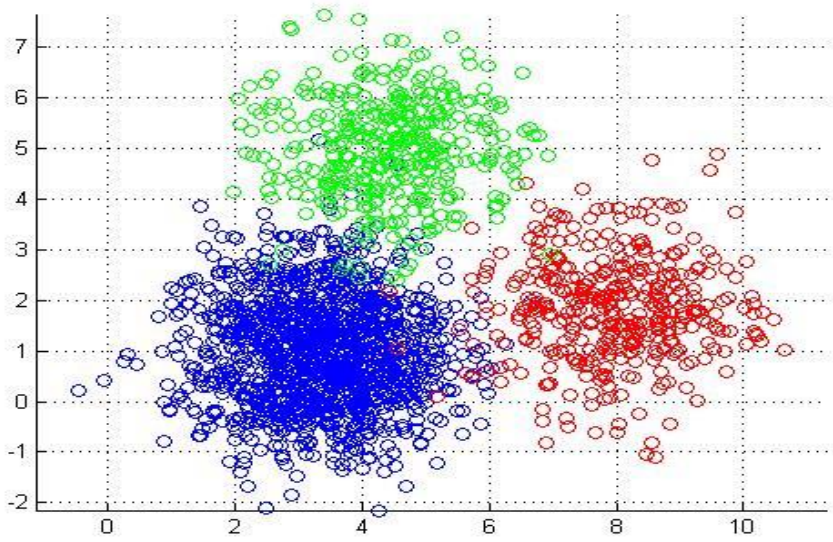


b. Unsupervised Learning

In this algorithm, we do not have any target or outcome variable to predict / estimate. It is used for clustering population in different groups, which is widely used for segmenting customers in different groups for specific intervention. Examples of Unsupervised Learning: Apriori algorithm, K-means.

K-means (Clustering Algorithm)

It is a type of unsupervised algorithm which solves the clustering problem. Its procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). Data points inside a cluster are homogeneous and heterogeneous to the peer groups.



c. Semi-supervised Learning

Using this algorithm, the machine is trained to make specific decisions. It works this way: the machine is exposed to an environment where it trains itself continually using trial and error. This machine learns from past experience and tries to capture the best possible knowledge to make accurate business decisions.

d. Gradient Boosting Algorithms

1. XGBoost

The XGBoost has an immensely high predictive power which makes it the best choice for accuracy in events as it possesses both linear model and the tree learning algorithm, making the algorithm almost 10x faster than existing gradient booster techniques. The support includes various objective functions, including regression, classification and ranking.