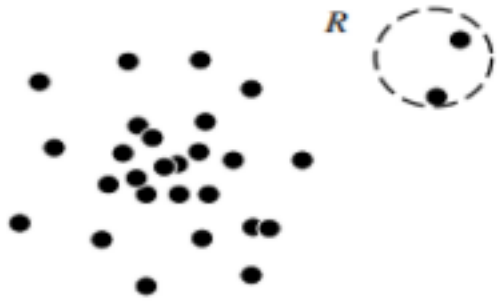
Outlier: An Introduction-:

- ❖ The process of Data Mining and Analytics involves the analysis of data and predicting the information that the data holds. Sometimes the certain object of a dataset deviates from the others. These deviated objects are termed outliers.
- ❖ It is a data object that deviates significantly from the rest of the data objects and behaves in a different manner.
- ❖ An outlier is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution errors. The analysis of outlier data is referred to as outlier analysis or outlier mining.
- ❖ Errors such as computational errors or incorrect entry of an object cause outliers. The differences of outliers to that of noise are:
  - Whenever some random error occurs in some measured variable or there is variance in the measured variable, then it is termed as noise.
  - Before detecting the outliers present in a dataset, it is advisable to remove the noise.

Ex-:



The objects in region *R* are outliers.

- ❖ Here most objects follow a roughly Gaussian distribution. However, the objects in region R are significantly different. It is unlikely that they follow the same distribution as the other objects in the data set. Thus, the objects in R are outliers in the data set.

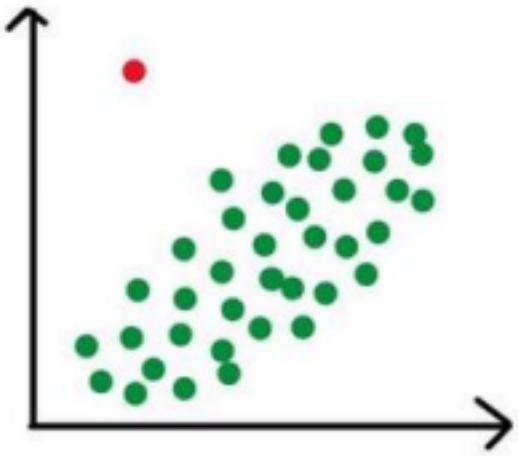
Outliers are of three types, namely –

1. Global (or Point) Outliers
2. Collective Outliers
3. Contextual (or Conditional) Outliers

1. Global Outliers

- ❖ They are also known as *Point Outliers*. These are the simplest form of outliers. If, in a given dataset, a data point strongly deviates from all the rest of the data points, it is known as a global outlier.

❖ Mostly, all of the outlier detection methods are aimed at finding global outliers. ❖ *For example,* In Intrusion Detection System(device or software application that monitors a network for malicious activity), if a large number of packages are broadcast in a very short span of time, then this may be considered as a global outlier and we can say that that particular system has been potentially hacked.

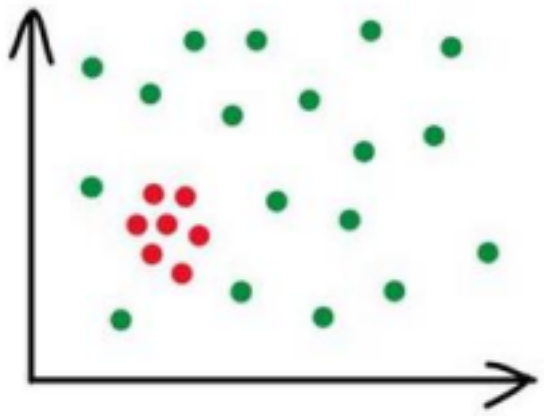


The red data point is a global outlier.

2. Collective Outliers

❖ As the name suggests, if in a given dataset, some of the data points, as a whole, deviate significantly from the rest of the dataset, they may be termed as collective outliers. ❖ Here, the individual data objects may not be outliers, but when seen as a whole, they may behave as outliers.

❖ To detect these types of outliers, we might need background information about the relationship between those data objects showing the behavior of outliers. ❖ *For example:* In an Intrusion Detection System, a DOS (denial-of-service) package from one computer to another may be considered as normal behavior. However, if this happens with several computers at the same time, then this may be considered as abnormal behavior and as a whole they can be termed as collective outliers.



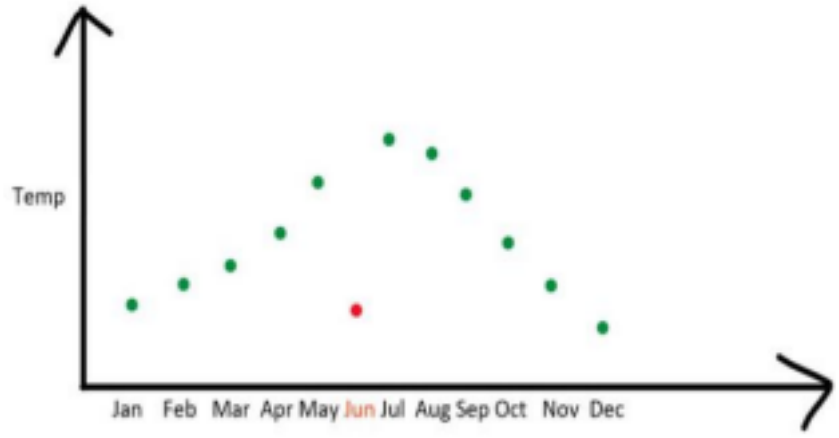
The red data points as a whole are collective outliers.

3. Contextual Outliers

- ❖ They are also known as *Conditional Outliers.* Here, if in a given dataset, a data object deviates significantly from the other data points based on a specific context or condition only. A data point may be an outlier due to a certain condition and may show normal behavior under another condition.
- ❖ Therefore, a context has to be specified as part of the problem statement in order to identify contextual outliers.

❖ Contextual outlier analysis provides flexibility for users where one can examine outliers in different contexts, which can be highly desirable in many applications. ❖ The attributes of the data point are decided on the basis of both contextual and behavioral attributes.

- ❖ *For example:* A temperature reading of 40°C may behave as an outlier in the context of a "winter season" but will behave like a normal data point in the context of a "summer season".



A low temperature value in June is a contextual outlier because the same value in December is not an outlier.

Challenges of Outlier Detection-:

Outlier detection is useful in many applications yet faces many challenges such as

Modeling normal objects and outliers effectively

- ❖ Outlier detection quality highly depends on the modeling of normal (non outlier) objects and outliers. Often, building a comprehensive model for data normality is very challenging, if not impossible.
- ❖ This is partly because it is hard to enumerate all possible normal behaviors in an application.
- ❖ The border between data normality and abnormality (outliers) is often not clear cut. Instead, there can be a wide range of gray area.
- ❖ Consequently, while some outlier detection methods assign to each object in the input data set a label of either "normal" or "outlier," other methods assign to each object a score measuring the "outlier-ness" of the object.

Application-specific outlier detection

- ❖ Technically, choosing the similarity/distance measure and the relationship model to describe data objects is critical in outlier detection. Unfortunately, such choices are often application-dependent.
- ❖ Different applications may have very different requirements.
- ❖ For example, in clinic data analysis, a small deviation may be important enough to justify an outlier. In contrast, in marketing analysis, objects are often subject to larger fluctuations, and consequently a substantially larger deviation is needed to justify an outlier.

Handling noise in outlier detection

❖ As mentioned earlier, outliers are different from noise. It is also well known that the quality of real data sets tends to be poor.
❖ Noise often unavoidably exists in data collected in many applications.
❖ Noise may be present as deviations in attribute values or even as missing values. ❖ Low data quality and the presence of noise bring a huge challenge to outlier detection. ❖ They can distort the data, blurring the distinction between normal objects and outliers.

Understandability

❖ To meet the understandability requirement, an outlier detection method has to provide some justification of the detection.
❖ For example, a statistical method can be used to justify the degree to which an object may be an outlier based on the likelihood that the object was generated by the same mechanism that generated the majority of the data.
❖ The smaller the likelihood, the more unlikely the object was generated by the same mechanism, and the more likely the object is an outlier.

Outlier Detection Methods-:

❖ Here, we present two orthogonal ways to categorize outlier detection methods. ❖ First, we categorize outlier detection methods according to whether the sample of data for analysis is given with domain expert–provided labels that can be used to build an outlier detection model.
❖ Second, we divide methods into groups according to their assumptions regarding normal objects versus outliers.

Supervised Methods-:

❖ Supervised methods model data normality and abnormality. Domain experts examine and label a sample of the underlying data.
❖ The task is to learn a classifier that can recognize outliers. The sample is used for training and testing.
❖ In some applications, the experts may label just the normal objects, and any other objects not matching the model of normal objects are reported as outliers.
❖ Other methods model the outliers and treat objects not matching the model of outliers as normal.
❖ Although many classification methods can be applied, challenges to supervised outlier detection include

1. The two classes (i.e., normal objects versus outliers) are imbalanced. That is, the population of outliers is typically much smaller than that of normal objects. Therefore, methods for handling imbalanced classes may be used, such as

oversampling (i.e., replicating) outliers to increase their distribution in the training set used to construct the classifier. The lack of outlier samples can limit the capability of classifiers built as such. To tackle these problems, some methods "make up" artificial outliers.

2. In many outlier detection applications, catching as many outliers as possible (i.e., the sensitivity or recall of outlier detection) is far more important than not mislabeling normal objects as outliers. Consequently, when a classification method is used for supervised outlier detection, it has to be interpreted appropriately so as to consider the application interest on recall.

❖ Morally, supervised methods of outlier detection must be careful in how they train and how they interpret classification rates due to the fact that outliers are rare in comparison to the other data samples.

Unsupervised Methods-:

❖ In some application scenarios, objects labeled as "normal" or "outlier" are not available. Thus, an unsupervised learning method has to be used.

❖ Unsupervised outlier detection methods make an implicit assumption: The normal objects are somewhat "clustered." In other words, an unsupervised outlier detection method expects that normal objects follow a pattern far more frequently than outliers.

❖ Normal objects do not have to fall into one group sharing high similarity. Instead, they can form multiple groups, where each group has distinct features. However, an outlier is expected to occur far away in feature space from any of those groups of normal objects.

❖ For instance, in some intrusion detection and computer virus detection problems, normal activities are very diverse and many do not fall into high-quality clusters. In such scenarios, unsupervised methods may have a high false positive rate—they may mislabel many normal objects as outliers (intrusions or viruses in these applications), and let many actual outliers go undetected.

❖ Due to the high similarity between intrusions and viruses (i.e., they have to attack key resources in the target systems), modeling outliers using supervised methods may be far more effective.

❖ Many clustering methods can be adapted to act as unsupervised outlier detection methods.

❖ The central idea is to find clusters first, and then the data objects not belonging to any cluster are detected as outliers. However, such methods suffer from two issues. First, a data object not belonging to any cluster may be noise instead of an outlier. Second, it is often costly to find clusters first and then find outliers.

❖ It is usually assumed that there are far fewer outliers than normal objects. ❖ The latest unsupervised outlier detection methods develop various smart ideas to tackle outliers directly without explicitly and completely finding clusters.

Semi-Supervised Methods-:

- ❖ In many applications, although obtaining some labeled examples is feasible, the number of such labeled examples is often small.
- ❖ We may encounter cases where only a small set of the normal and/or outlier objects are labeled, but most of the data are unlabeled. Semi-supervised outlier detection methods were developed to tackle such scenarios.
- ❖ For example, when some labeled normal objects are available, we can use them, together with unlabeled objects that are close by, to train a model for normal objects. The model of normal objects then can be used to detect outliers—those objects not fitting the model of normal objects are classified as outliers.
- ❖ If only some labeled outliers are available, semi-supervised outlier detection is trickier. A small number of labeled outliers are unlikely to represent all the possible outliers. ❖ Therefore, building a model for outliers based on only a few labeled outliers is unlikely to be effective. To improve the quality of outlier detection, we can get help from models for normal objects learned from unsupervised methods.

Statistical Methods-:

As discussed previously, outlier detection methods make assumptions about outliers versus the rest of the data. According to the assumptions made, we can categorize outlier detection methods into three types: statistical methods, proximity-based methods, and clustering-based methods.

Statistical methods

- ❖ Statistical methods (also known as model-based methods) make assumptions of data normality.
- ❖ They assume that normal data objects are generated by a statistical (stochastic) model, and that data not following the model are outliers.
- ❖ The effectiveness of statistical methods highly depends on whether the assumptions made for the statistical model hold true for the given data. There are many kinds of statistical models.
- ❖ For example, the statistic models used in the methods may be parametric or nonparametric.
- ❖ Statistical methods for outlier detection are discussed in statistical approaches.

Proximity-Based Methods-:

- ❖ Proximity-based methods assume that an object is an outlier if the nearest neighbors of the object are far away in feature space, that is, the proximity of the object to its neighbors significantly deviates from the proximity of most of the other objects to their neighbors in the same data set.
- ❖ The effectiveness of proximity-based methods relies heavily on the proximity (or distance)

measure used. In some applications, such measures cannot be easily obtained. Moreover, proximity-based methods often have difficulty in detecting a group of outliers if the outliers are close to one another.

Clustering-Based Methods

- ❖ Clustering-based methods assume that the normal data objects belong to large and dense clusters, whereas outliers belong to small or sparse clusters, or do not belong to any clusters.
- ❖ Clustering is an expensive data mining operation. A straightforward adaptation of a clustering method for outlier detection can be very costly, and thus does not scale up well for large data sets.

Statistical Approaches-:

❖ This is the analysis of raw data using mathematical formulas, models, and techniques. ❖ As with statistical methods for clustering, statistical methods for outlier detection make assumptions about data normality.

- ❖ They assume that the normal objects in a data set are generated by a stochastic process (a generative model). Consequently, normal objects occur in regions of high probability for the stochastic model, and objects in the regions of low probability are outliers.
- ❖ The general idea behind statistical methods for outlier detection is to learn a generative model fitting the given data set, and then identify those objects in low-probability regions of the model as outliers.
- ❖ However, there are many different ways to learn generative models. In general, statistical methods for outlier detection can be divided into two major categories: parametric methods and nonparametric methods, according to how the models are specified and learned.
- ❖ A parametric method assumes that the normal data objects are generated by a parametric distribution with parameter θ. The probability density function of the parametric distribution $f(x, \theta)$ gives the probability that object x is generated by the distribution. The smaller this value, the more likely x is an outlier.
- ❖ A nonparametric method does not assume an a priori statistical model. Instead, a nonparametric method tries to determine the model from the input data. Note that most nonparametric methods do not assume that the model is completely parameter free. Non parametric methods often take the position that the number and nature of the parameters are flexible and not fixed in advance.

Parametric Methods

❖ Here we first discuss methods for univariate data based on normal distribution. We then discuss how to handle multivariate data using multiple parametric distributions. ❖ Data

involving only one attribute or variable are called univariate data. For simplicity, we often choose to assume that data are generated from a normal distribution. We can then learn the parameters of the normal distribution from the input data, and identify the points with low probability as outliers.

Ex-:

- **Univariate outlier detection using maximum likelihood-:** Suppose a city's average temperature values in July in the last 10 years are, in value-ascending order, 24.0◦C, 28.9◦C, 28.9◦C, 29.0◦C, 29.1◦C, 29.1◦C, 29.2◦C, 29.2◦C, 29.3◦C, and 29.4◦C. Let's assume that the average temperature follows a normal distribution, which is determined by two parameters: the mean, μ, and the standard deviation, σ. we can use the maximum likelihood method to estimate the parameters μ and σ. That is, we maximize the log likelihood function.

$$\ln \mathcal{L}(\mu,\sigma^2) = \sum_{i=1}^{n} \ln f(x_i|(\mu,\sigma^2)) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

- Where n is the total number of samples, which are 10 in this example. Taking derivatives with respect to μ and σ^2 and solving the resulting system of first order conditions leads to the following maximum likelihood estimates:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

In this example, we have

$$\hat{\mu} = \frac{24.0+28.9+28.9+29.0+29.1+29.1+29.2+29.2+29.3+29.4}{10} = 28.61$$

$$\begin{aligned}\hat{\sigma}^2 = ((24.1-28.61)^2 &+ (28.9-28.61)^2 + (28.9-28.61)^2 + (29.0-28.61)^2 \\ &+ (29.1-28.61)^2 + (29.1-28.61)^2 + (29.2-28.61)^2 + (29.2-28.61)^2 \\ &+ (29.3-28.61)^2 + (29.4-28.61)^2)/10 \simeq 2.29.\end{aligned}$$

Accordingly, we have σˆ = √ 2.29 = 1.51.

The most deviating value, 24.0◦C, is 4.61◦C away from the estimated mean. We know

that the $\mu \pm 3\sigma$ region contains 99.7% data under the assumption of normal distribution. Because 4.61/1.51 = 3.04 > 3, the probability that the value 24.0◦C is generated by the normal distribution is less than 0.15%, and thus can be identified as an outlier.

- ❖ Now we will discuss Detection of Multivariate Outliers. Data involving two or more attributes or variables are multivariate data.
- ❖ Many univariate outlier detection methods can be extended to handle multivariate data. The central idea is to transform the multivariate outlier detection task into a univariate outlier detection problem. Here, we use two examples to illustrate this idea.

Ex-: Multivariate outlier detection using the Mahalanobis distance-: For a multivariate data set, let o^- be the mean vector. For an object, o, in the data set, the Mahalanobis distance from o to o^- is

$$MDist(o, o^-) = (o - {}^-o)^T S^{-1}(o - {}^-o)$$

Where S is the covariance matrix.

MDist (o, o^-) is a univariate variable, and thus Grubb's test can be applied to this measure. Therefore, we can transform the multivariate outlier detection tasks as follows:

1. Calculate the mean vector from the multivariate data set.

2. For each object o, calculate MDist (o, o^-), the Mahalanobis distance from o to o^-. 3.

Detect outliers in the transformed univariate data set, {MDist (o, o^-)|o ∈ D}. 4. If MDist

(o, o^-) is determined to be an outlier, then o is regarded as an outlier as well

Our second example uses the χ 2 -statistic to measure the distance between an object to the mean of the input data set.

Ex-: Multivariate outlier detection using the χ 2 –statistic-: The χ 2 -statistic can also be used to capture multivariate outliers under the assumption of normal distribution. For an object, o, the χ 2 -statistic is

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - E_i)^2}{E_i}$$

Where o_i is the value of o on the ith dimension, E_i is the mean of the i-dimension among all objects, and n is the dimensionality. If the χ 2 -statistic is large, the object is an outlier.

Nonparametric Methods

❖ In nonparametric methods for outlier detection, the model of "normal data" is learned from the input data, rather than assuming one a priori.

❖ Nonparametric methods often make fewer assumptions about the data, and thus can be applicable in more scenarios.

Statistical data mining-:

❖ Data mining techniques are designed for the efficient handling of huge amounts of data that are typically multidimensional and possibly of various complex types. ❖ There are, however, many well-established statistical techniques for data analysis, particularly for numeric data.

❖ These techniques have been applied extensively to scientific data (e.g., data from experiments in physics, engineering, manufacturing, psychology, and medicine), as well as to data from economics and the social sciences. Some of these techniques, such as principal components analysis and clustering. Some of them are

- Regression: In general, these methods are used to predict the value of a response (dependent) variable from one or more predictor (independent) variables, where the variables are numeric.
- Generalized linear models: These models, and their generalization (generalized additive models), allow a categorical (nominal) response variable (or some transformation of it) to be related to a set of predictor variables in a manner similar to the modeling of a numeric response variable using linear regression. Generalized linear models include logistic regression and Poisson regression.
- Analysis of variance: These techniques analyze experimental data for two or more populations described by a numeric response variable and one or more categorical variables (factors). In general, an ANOVA (single-factor analysis of variance) problem involves a comparison of k population or treatment means to determine if at least two of the means are different.
- Mixed-effect models: These models are for analyzing grouped data—data that can be classified according to one or more grouping variables. They typically describe relationships between a response variable and some covariates in data grouped according to one or more factors.
- Factor analysis: This method is used to determine which variables are combined to generate a given factor. For example, for many psychiatric data, it is not possible to measure a certain factor of interest directly (e.g., intelligence); however, it is often possible to measure other quantities (e.g., student test scores) that reflect the factor of interest.
- Discriminant analysis: This technique is used to predict a categorical response variable. Unlike generalized linear models, it assumes that the independent

variables follow a multivariate normal distribution. The procedure attempts to determine several discriminant functions (linear combinations of the independent variables) that discriminate among the groups defined by the response variable. Discriminant analysis is commonly used in social sciences.

- Survival analysis: Several well-established statistical techniques exist for survival analysis. These techniques originally were designed to predict the probability that a patient undergoing a medical treatment would survive at least to time t. Methods for survival analysis, however, are also commonly applied to manufacturing settings to estimate the life span of industrial equipment.
- Quality control: Various statistics can be used to prepare charts for quality control, such as Shewhart charts and CUSUM charts (both of which display group summary statistics). These statistics include the mean, standard deviation, range, count, moving average, moving standard deviation, and moving range.

Data mining and recommender systems-:

- ❖ Today's consumers are faced with millions of goods and services when shopping online. Recommender systems help consumers by making product recommendations that are likely to be of interest to the user such as books, CDs, movies, restaurants, online news articles, and other services.
- ❖ Recommender systems may use either a content based approach, a collaborative approach, or a hybrid approach that combines both content-based and collaborative methods.
- ❖ The content-based approach recommends items that are similar to items the user preferred or queried in the past. It relies on product features and textual item descriptions. ❖ The collaborative approach (or collaborative filtering approach) may consider a user's social environment. It recommends items based on the opinions of other customers who have similar tastes or preferences as the user.
- ❖ Recommender systems use a broad range of techniques from information retrieval, statistics, machine learning, and data mining to search for similarities among items and customer preferences.
- ❖ An advantage of recommender systems is that they provide personalization for customers of e-commerce, promoting one-to-one marketing. Amazon, a pioneer in the use of collaborative recommender systems, offers "a personalized store for every customer" as part of their marketing strategy. Personalization can benefit both consumers and the company involved.
- ❖ The recommendation problem considers a set, C, of users and a set, S, of items. Let u be a utility function that measures the usefulness of an item, s, to a user, c. The utility is commonly represented by a rating and is initially defined only for items previously rated by users.
- ❖ For example, when joining a movie recommendation system, users are typically asked to

rate several movies. The space C × S of all possible users and items is huge. ❖ The recommendation system should be able to extrapolate from known to unknown ratings so as to predict item–user combinations. Items with the highest predicted rating/utility for a user are recommended to that user.

❖ Recommender systems face major challenges such as scalability and ensuring quality recommendations to the consumer. For example, regarding scalability, collaborative recommender systems must be able to search through millions of potential neighbors in real time.

Data mining for financial data analysis-:

❖ Most banks and financial institutions offer a wide variety of banking, investment, and credit services (the latter include business, mortgage, and automobile loans and credit cards). Some also offer insurance and stock investment services.

❖ Financial data collected in the banking and financial industry are often relatively complete, reliable, and of high quality, which facilitates systematic data analysis and data mining. Here we present some cases.

- **Design and construction of data warehouses for multidimensional data analysis and data mining**: Like many other applications, data warehouses need to be constructed for banking and financial data. Multidimensional data analysis methods should be used to analyze the general properties of such data. For example, a company's financial officer may want to view the debt and revenue changes by month, region, and sector, and other factors, along with maximum, minimum, total, average, trend, deviation, and other statistical information. Data warehouses, data cubes (including advanced data cube concepts such as multi feature, discovery-driven, regression, and prediction data cubes), characterization and class comparisons, clustering, and outlier analysis will all play important roles in financial data analysis and mining.
- **Loan payment prediction and customer credit policy analysis**: Loan payment prediction and customer credit analysis are critical to the business of a bank. Many factors can strongly or weakly influence loan payment performance and customer credit rating. Data mining methods, such as attribute selection and attribute relevance ranking, may help identify important factors and eliminate irrelevant ones. For example, factors related to the risk of loan payments include loan-to-value ratio, term of the loan, debt ratio (total amount of monthly debt versus total monthly income), payment-to-income ratio, customer income level, education level, residence region, and credit history.
- **Classification and clustering of customers for targeted marketing:** Classification and clustering methods can be used for customer group identification and targeted marketing. For example, we can use classification to

identify the most crucial factors that may influence a customer's decision regarding banking.

- **Detection of money laundering and other financial crimes:** To detect money laundering and other financial crimes, it is important to integrate information from multiple, heterogeneous databases (e.g., bank transaction databases and federal or state crime history databases), as long as they are potentially related to the study. Multiple data analysis tools can then be used to detect unusual patterns, such as large amounts of cash flow at certain periods, by certain groups of customers. Useful tools include data visualization tools (to display transaction activities using graphs by time and by groups of customers), linkage and information network analysis tools (to identify links among different customers and activities), classification tools (to filter unrelated attributes and rank the highly related ones), clustering tools (to group different cases), outlier analysis tools (to detect unusual amounts of fund transfers or other activities), and sequential pattern analysis tools (to characterize unusual access sequences). These tools may identify important relationships and patterns of activities and help investigators focus on suspicious cases for further detailed examination.

Data Mining for Intrusion Detection and Prevention-:

- The extensive growth of the Internet and the increasing availability of tools and tricks for intruding and attacking networks have prompted intrusion detection and prevention to become a critical component of networked systems.
- An intrusion can be defined as any set of actions that threaten the integrity, confidentiality, or availability of a network resource (e.g., user accounts, file systems, system kernels, and so on).
- Intrusion detection systems and intrusion prevention systems both monitor network traffic and/or system executions for malicious activities.
- However, the former produces reports whereas the latter is placed in-line and is able to actively prevent/block intrusions that are detected. The main functions of an intrusion prevention system are to identify malicious activity, log information about said activity, attempt to block/stop activity, and report activity.
- The majority of intrusion detection and prevention systems use either signature based detection or anomaly-based detection.
  - **Signature-based detection:** This method of detection utilizes signatures, which are attack patterns that are preconfigured and predetermined by domain experts. A signature-based intrusion prevention system monitors the network traffic for matches to these signatures. Once a match is found, the intrusion detection system will report the anomaly and an intrusion prevention system will take additional appropriate actions. The drawback of such system is the detection mechanism can

only identify cases that match the signatures. That is, it is unable to detect new or previously unknown intrusion tricks.

- **Anomaly-based detection:** This method builds models of normal network behavior (called profiles) that are then used to detect new patterns that significantly deviate from the profiles. Such deviations may represent actual intrusions or simply be new behaviors that need to be added to the profiles. The main advantage of anomaly detection is that it may detect novel intrusions that have not yet been observed. Typically, a human analyst must sort through the deviations to ascertain which represent real intrusions. A limiting factor of anomaly detection is the high percentage of false positives. New patterns of intrusion can be added to the set of signatures to enhance signature-based detection.

❖ Data mining methods can help an intrusion detection and prevention system to enhance its performance in various ways as follows.

- **New data mining algorithms for intrusion detection:** Data mining algorithms can be used for both signature-based and anomaly-based detection. In signature based detection, training data are labeled as either "normal" or "intrusion." A classifier can then be derived to detect known intrusions.
- **Association, correlation, and discriminative pattern analyses help select and build discriminative classifiers**: Association, correlation, and discriminative pattern mining can be applied to find relationships between system attributes describing the network data. Such information can provide insight regarding the selection of useful attributes for intrusion detection.
- **Analysis of stream data:** Due to the transient and dynamic nature of intrusions and malicious attacks, it is crucial to perform intrusion detection in the data stream environment. Moreover, an event may be normal on its own, but considered malicious if viewed as part of a sequence of events.
- **Distributed data mining:** Intrusions can be launched from several different locations and targeted to many different destinations. Distributed data mining methods may be used to analyze network data from several network locations to detect these distributed attacks.
- **Visualization and querying tools:** Visualization tools should be available for viewing any anomalous patterns detected. Such tools may include features for viewing associations, discriminative patterns, clusters, and outliers. Intrusion detection systems should also have a graphical user interface that allows security analysts to pose queries regarding the network data or intrusion detection results.