DATA MINING

Fraud Detection

Drilon Hyseni

CONTENTS OF THIS PROJECT

Here’s what you’ll find :

1. Fraud Detection
2. Anomaly detection.
3. Link analysis
4. Rule-based systems
5. Machine learning
6. Social network analysis
7. Predictive modeling
8. Study Case

Fraud Detection

Fraud detection in data mining is the process of using data mining techniques to identify fraudulent activities and prevent financial losses. It involves analyzing large amounts of data from various sources, such as transaction records, customer profiles, and behavioral patterns, to identify patterns and anomalies that may indicate fraudulent activities.

Data mining techniques used for fraud detection include:

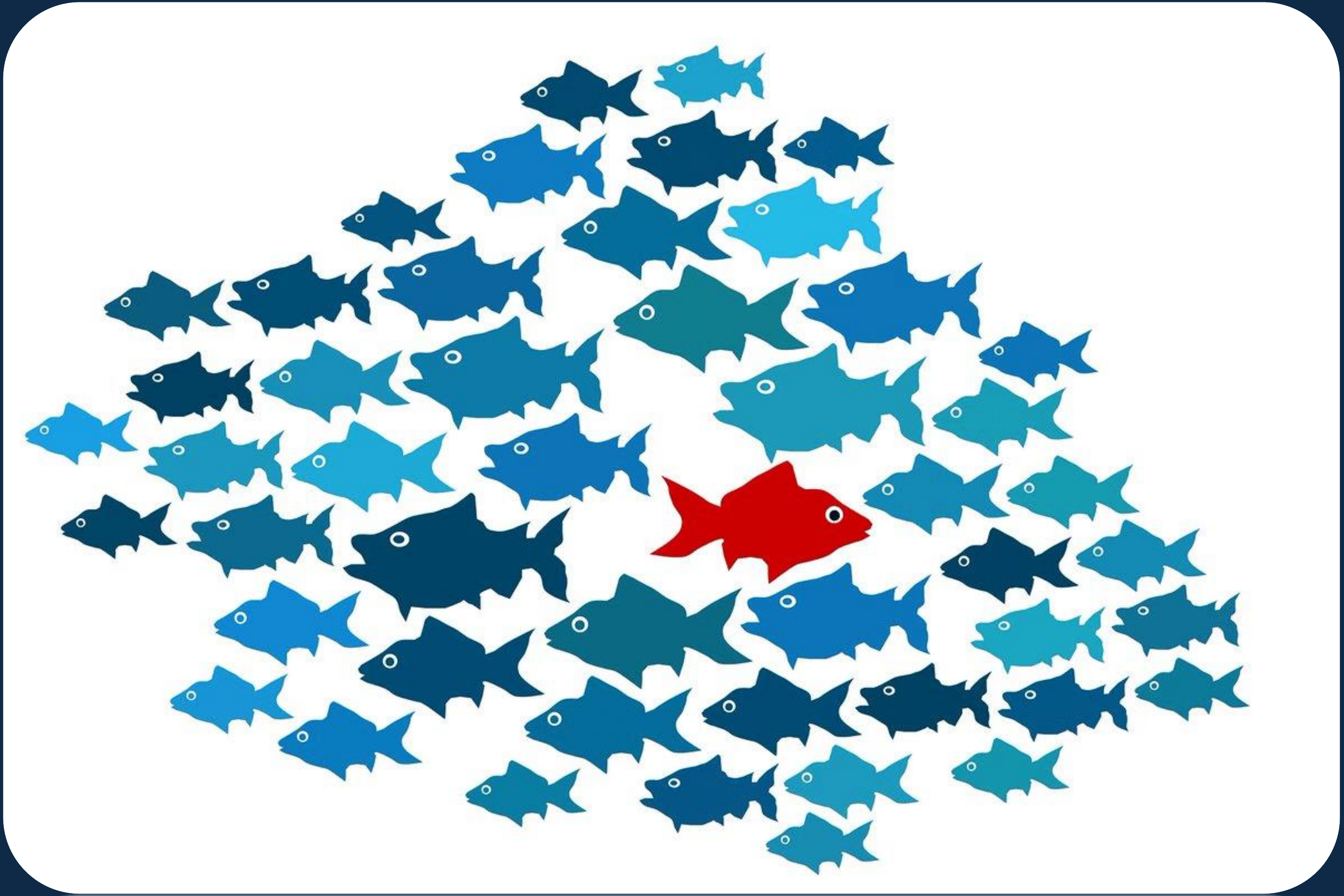
- Anomaly detection: Identifying unusual transactions that deviate from normal patterns.
- Link analysis: Examining relationships between entities involved in fraudulent activities.
- Rule-based systems: Flag transactions that violate predetermined rules or conditions.
- Machine learning: Using algorithms to detect fraud based on patterns in historical data.
- Social network analysis: Analyzing connections between individuals and groups involved in fraudulent activities.
- Predictive modeling: Forecasting potential fraud based on past behaviors and trends.

Fraud detection in data mining is a critical aspect of financial and business operations, as it helps organizations protect their assets and reputation by detecting and preventing fraudulent activities. The goal of fraud detection in data mining is to minimize the financial losses incurred by fraudulent activities and to improve the efficiency of the fraud detection process.

Anomaly Detection

- Anomaly detection is the process of identifying data points that fall outside the normal behaviour. For example, typical data will generally conform to certain shapes, such as a bell curve. By looking at the datasets and points that don’t fit, business stakeholders and data teams can identify patterns or points that may be threats. Companies also use data mining for fraud detection and network intrusion detection activities.
- In the case of fraud detection at a credit card company, anomaly detection algorithms look for unusual data points for cardholders. Say a frugal cardholder’s credit card activity registered several extravagant purchases all in one day. The credit card company would most likely flag that as potential fraud and take action to verify it.
- An Intrusion Detection System (IDS) looks for anomalous attempts to get into a network or any other malicious activity. It will send an alert to an admin when it detects red flags like policy violations or significant changes in network traffic. These basic, real-world rules can make all the difference.

- Data points can cause consequences and disadvantages if they are not found when an anomaly or outlier makes it into the training set. As a result, the algorithm learns the anomaly as normal. Organizations need the capability to identify abnormalities quickly, and that won’t happen if the algorithm’s pattern recognition is skewed.
- Anomaly detection and outlier detection differ based on how their results can be applied. It’s that simple. Anomaly detection can lead to the discovery of a whole new model when the anomaly is found to be part of a different dataset, whereas outlier detection can improve the accuracy of the current model by drawing attention to the treatment of outliers. These are meaningful benefits and causes for including both anomaly and outlier detection in an analytics strategy.
- Other benefits of anomaly detection include automating KPI calculations and analysis, preventing security breaches, and achieving faster results.



Common types of anomalies in data mining

- Point anomalies
- Contextual anomalies
- Collective anomalies

- Update anomaly
- Deletion anomaly
- Insertion anomaly

Point anomalies

A point anomaly occurs when a single data point is extremely far from the normal distribution. For example, one person trying to cash a check for a million dollars would be a point anomaly.

Contextual anomalies

In time series data anomalies (or anomalies in datasets spread over time), the data points that stand out are those that don’t fit the time-tested pattern. For example, if the number of cars going through a single toll booth on a regular Tuesday was to spike from 1,000 to 10,000.

Collective anomalies

The collective anomaly occurs when several data points, or a collection, fall far outside environments of normal behavior, but the individual data points are not anomalous. Everyone repainting their house on the same day is a collective anomaly because it isn’t unusual for people to paint their houses. But it is rare that everyone would do it on the same day.

Update anomaly

This anomaly occurs when the data has become redundant or has been partially updated. For example, pies sold at a bakery can be categorized by both the number of crusts they have and whether or not they have fruit filling. There would be an update anomaly if a double-crust apple pie is updated on the fruit pie list, but not the other.

Deletion anomaly

When data is lost because it has been removed, it can cause a deletion anomaly. An example would be if a double-crust chocolate pudding pie is removed from the inventory, not because it was sold, but because it was deleted by accident.

Insertion anomaly

The insertion anomaly is actually about omission. When data can’t be entered into the dataset because it’s incomplete, an insertion anomaly occurs due to that data not being included. For example, when an experimental crustless pecan pie is not included in the inventory.

Anomaly detection methods

The two main anomaly detection techniques are unsupervised anomaly detection and supervised anomaly detection.

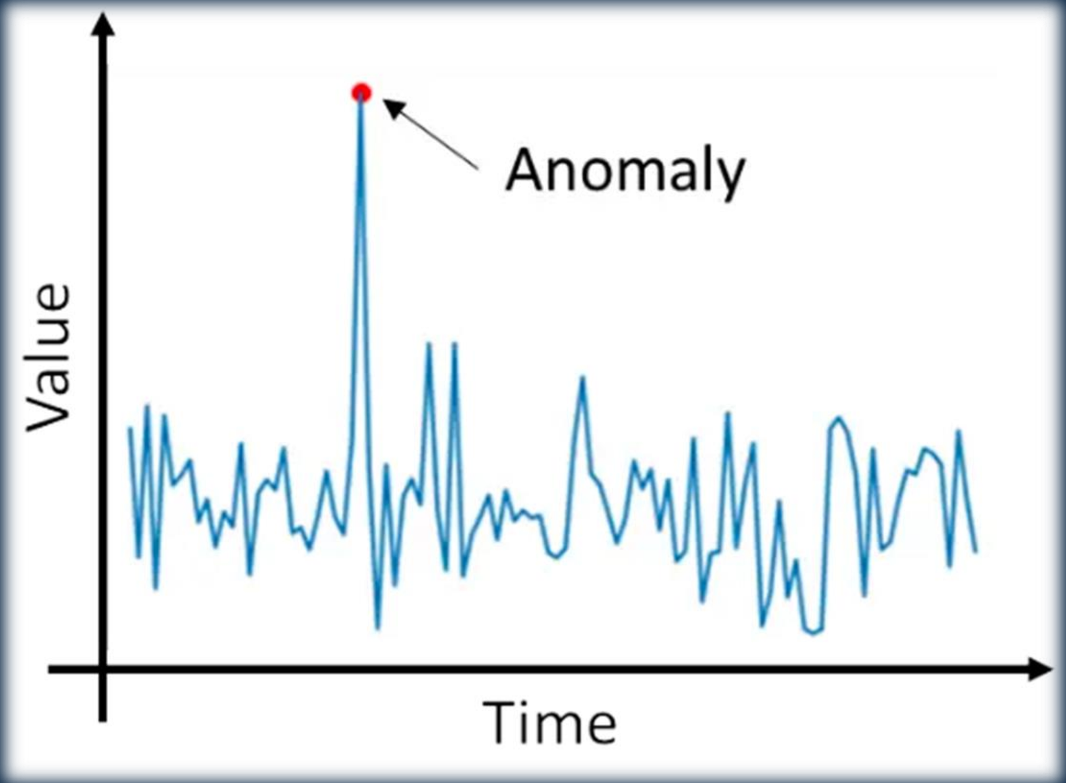
Unsupervised anomaly detection occurs when an algorithm, such as the k-means clustering algorithm, does all the work.

Supervised anomaly detection happens when there is the availability of human interventions or classifications and maybe a Support Vector Machine (SVM) to more efficiently analyse data.

While it may be difficult to have a person or team of people monitoring the machine learning algorithms that drive anomaly detection, people need to be involved to ensure the algorithm’s quality and accuracy. Another method to consider is semi-supervised anomaly detection, which combines supervised and unsupervised anomaly detection.

One practical application of the decision tree algorithm is in the Adaptive Intrusion Detection Systems where anomalous data need to be evaluated quickly and accurately. The decision tree reduces the number of false positives and improves the anomaly detection rate.

The Local Outlier Factor (LOF) is also a popular anomaly detection technique. The idea is to put the distance between a number and its k-nearest neighbours (KNN) into a ratio. The LOF gives increased precision in determining what’s an outlier.


Anomaly
Value
Time

Link Analysis

Link analysis in data mining

What Does Link Analysis Mean?

Link analysis is a data analysis technique used in network theory that is used to evaluate the relationships or connections between network nodes. These relationships can be between various types of objects (nodes), including people, organizations and even transactions.

Link analysis is essentially a kind of knowledge discovery that can be used to visualize data to allow for better analysis, especially in the context of links, whether Web links or relationship links between people or between different entities. Link analysis is often used in search engine optimization as well as in intelligence, in security analysis and in market and medical research.

Techopedia Explains Link Analysis

Link analysis is literally about analysing the links between objects, whether they are physical, digital or relational. This requires diligent data gathering. For example, in the case of a website where all of the links and backlinks that are present must be analysed, a tool has to sift through all of the HTML codes and various scripts in the page and then follow all the links it finds in order to determine what sort of links are present and whether they are active or dead. This information can be very important for search engine optimization, as it allows the analyst to determine whether the search engine is actually able to find and index the website.

In networking, link analysis may involve determining the integrity of the connection between each network node by analyzing the data that passes through the physical or virtual links. With the data, analysts can find bottlenecks and possible fault areas and are able to patch them up more quickly or even help with network optimization.

Link analysis has three primary purposes:

- Find matches for known patterns of interests between linked objects.
- Find anomalies by detecting violated known patterns.
- Find new patterns of interest (for example, in social networking and marketing and business intelligence).

More about link analysis and examples:

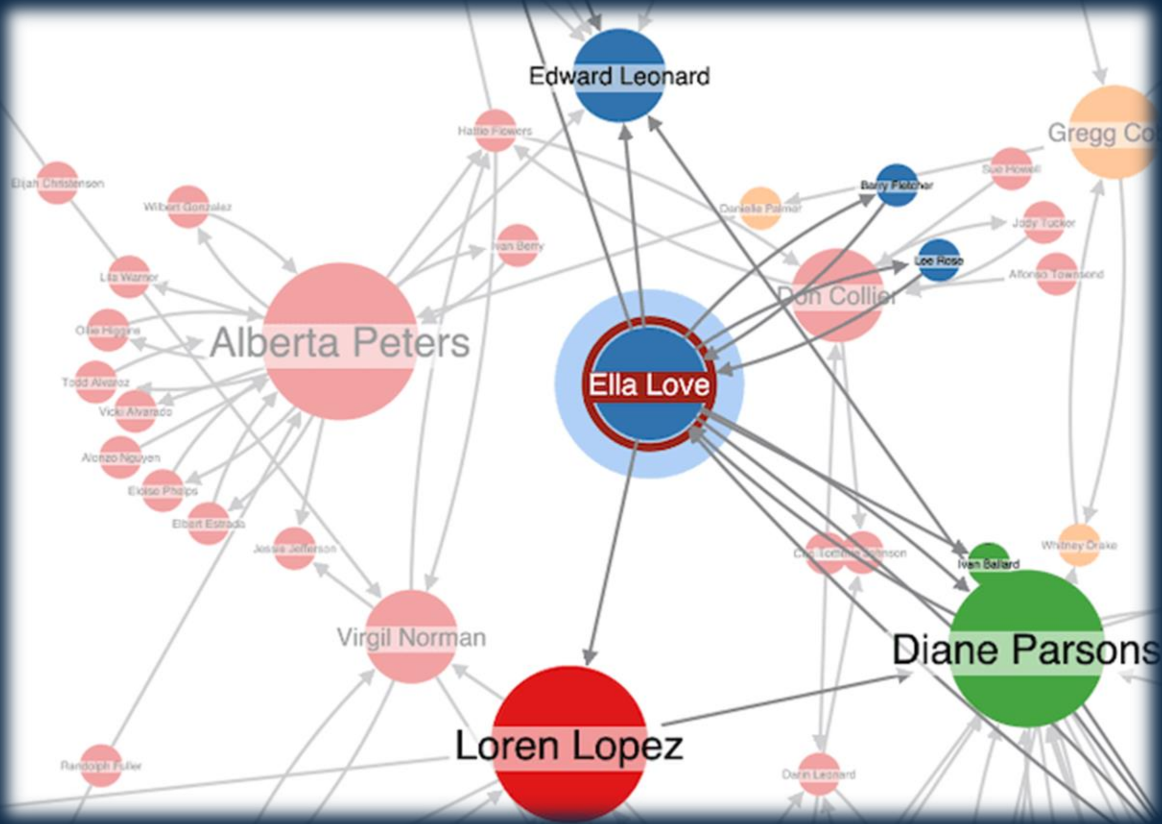
About link analysis

Link analysis uses a network of interconnected links and nodes to identify and analyse relationships that are not easily seen in raw data. Common types of networks include the following:

- Social networks that show who talks to whom
- Semantic networks that illustrate topics that are related to each other
- Conflict networks indicating alliances of connections between players
- Airline networks indicating which airports have connecting flights

Examples

- A crime analyst is investigating a criminal network. Data from cell phone records can be used to determine the relationship and hierarchy between members of the network.
- A credit card company is developing a new system to detect credit card theft. The system uses the known patterns of transactions for each client, such as the city, stores, and types of transactions, to identify anomalies and alert the client of a potential theft.
- A public health analyst is researching the opioid crisis in North America. The analyst uses data on prescriptions and demographics to identify new patterns that are emerging as the crisis spreads.


Edward Leonard
Gregg Co
Hattie Flowers
Sue Howell
Elijah Christenson
Barry Fletcher
Wilbert Gonzalez
Daniella Palmer
Jody Tucker
Ivan Berry
Lila Warner
Lee Rose
Alfonso Townsend
Don Collier
Ollie Higgins
Alberta Peters
Todd Alvarez
Ella Love
Vicki Alvarado
Alonzo Nguyen
Eloise Phelps
Elbert Estrada
Jessie Jefferson
Tommie Johnson
Ivan Ballard
Whitney Drake
Virgil Norman
Diane Parsons
Loren Lopez
Randolph Fuller
Darin Leonard

Rule Based Systems

What is a rule-based system?

A rule-based system is a system that applies human-made rules to store, sort and manipulate data. In doing so, it mimics human intelligence.
To work, rule-based systems require a set of facts or source of data, and a set of rules for manipulating that data. These rules are sometimes referred to as ‘If statements’ as they tend to follow the line of ‘IF X happens THEN do Y’.
Automation software like Think Automation is a good example. It automates processes by breaking them down into steps.

- First comes the data or new business event
- Then comes the analysis: the part where the system conditonally processes the data against its rules
- Then comes any subsequent automated follow-up actions

So , a rule-based system is a logical program that uses pre-defined rules to make deductions and choices to perform automated actions.

How does a rule-based system work?

Rule-based systems, unsurprisingly, work based on rules. These rules outline triggers and the actions that should follow (or are triggered). For example, a trigger might be an email containing the word “invoice”. An action might then be to forward the email to the finance team.

These rules most often take the form of if statements. ‘IF’ outlines the trigger, ‘THEN’ specifies the action to complete. So, if you want to create a rule-based system capable of handling 100 different actions, you’d have to write 100 different rules. If you want to then update the system and add actions, then you would need to write new rules.

In short, you use rules to tell a machine what to do, and the machine will do exactly as you tell it. From there, rule-based systems will execute the actions until you tell it to stop.

But remember: if you tell it to do something incorrectly, it will do it incorrectly.

What is a rule-based system not?

Due to early use in the fields, rule-based systems are commonly confused with artificial intelligence and machine learning. However, they are not AI, and they are not machine learning.

It’s easy to confuse the two as they can look very similar. Both involve machines completing tasks, seemingly on their own. The difference is that AI can determine the action to take itself; it can learn and adapt. Meanwhile, rule-based systems do exactly as instructed by a human.

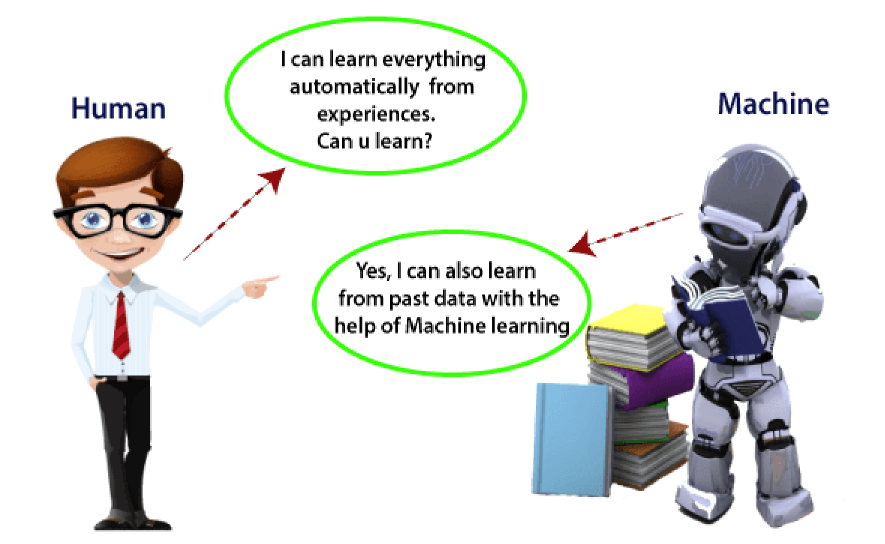
In other words, unlike artificial intelligence and machine learning, the actions carried out by rule-based systems (the rules that they follow) are determined by a human.

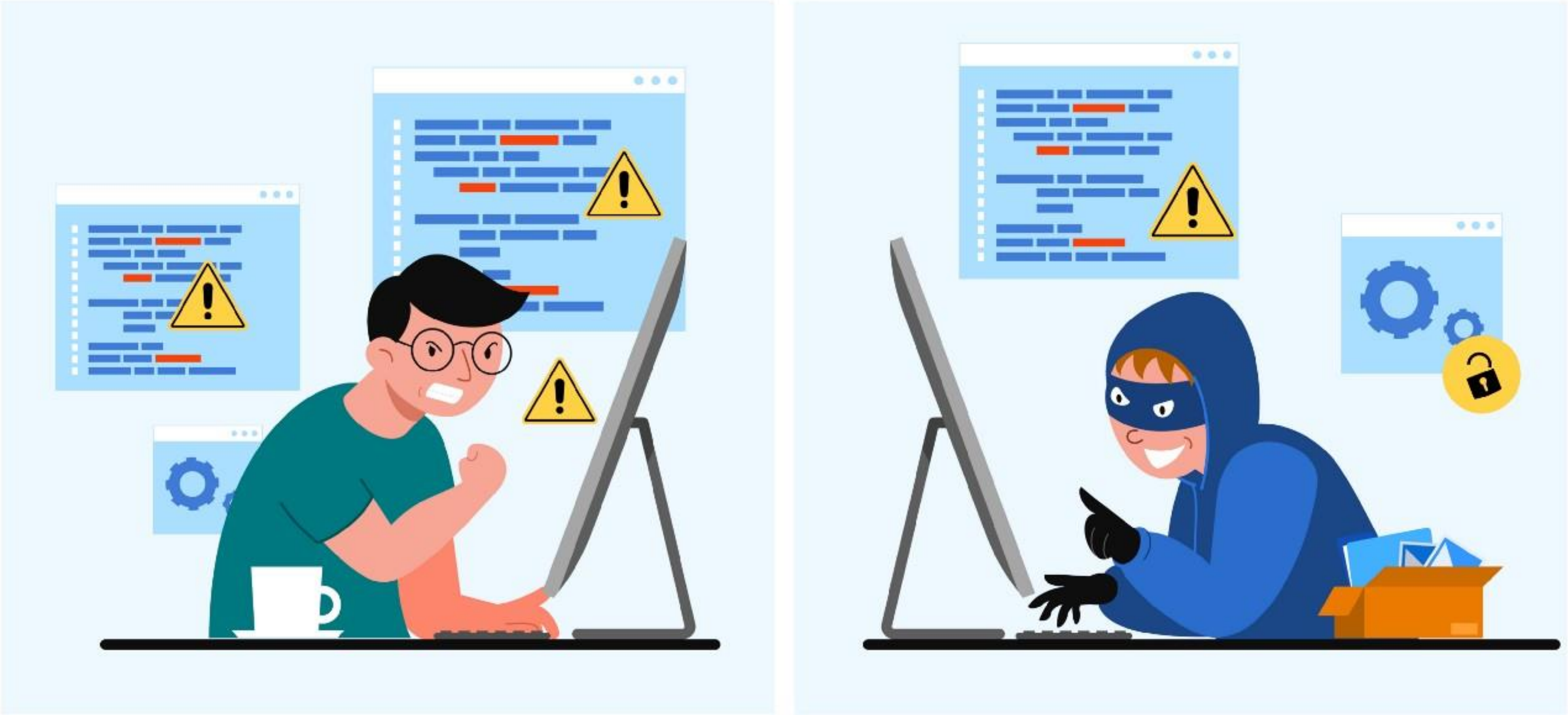
The system doesn’t work it out for itself, or intelligently make decisions.

Machine Learning

What is Machine Learning?

Machine Learning is the study of making computers more human-like in their behaviour and decisions by giving them the capacity to learn and generate their own programming. This is accomplished with little human interaction. The Machine Learning method is automated and refined depending on the machines’ experiences during the process.


Human
I can learn everything automatically from experiences. Can u learn?
Machine
Yes, I can also learn from past data with the help of Machine learning



- Yes, machine learning is a popular data mining technique used for fraud detection. It involves training machine learning algorithms on historical data to identify patterns and relationships that can be used to identify fraudulent activities.
- The algorithms can be supervised or unsupervised, depending on the availability of labeled data. Supervised algorithms, such as decision trees and random forests, require labeled data to train the model, while unsupervised algorithms, such as clustering and anomaly detection, do not require labeled data.

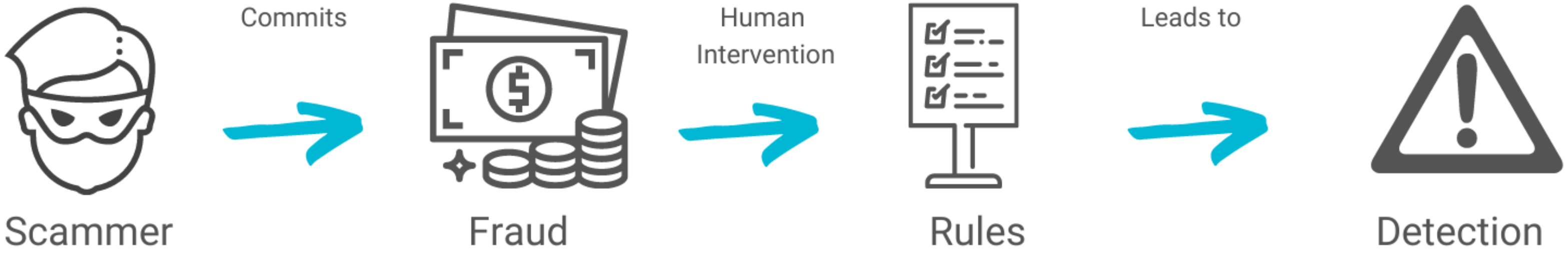
- In fraud detection, machine learning algorithms are trained on historical data to learn the patterns and characteristics of fraudulent transactions. The algorithms can then be used to flag transactions that deviate from these patterns as potentially fraudulent. Machine learning algorithms can also be used to detect fraudulent activities in real-time, by analyzing live transaction data.
- Machine learning is an effective technique for fraud detection as it can handle large amounts of data, identify patterns and relationships that are not easily detected by humans, and improve accuracy and efficiency over time as more data becomes available. However, it is important to validate the results of machine learning algorithms to ensure that they are not flagging legitimate transactions as fraudulent, and to continually update and improve the algorithms as new types of fraud emerge.

Key Features of Machine Learning

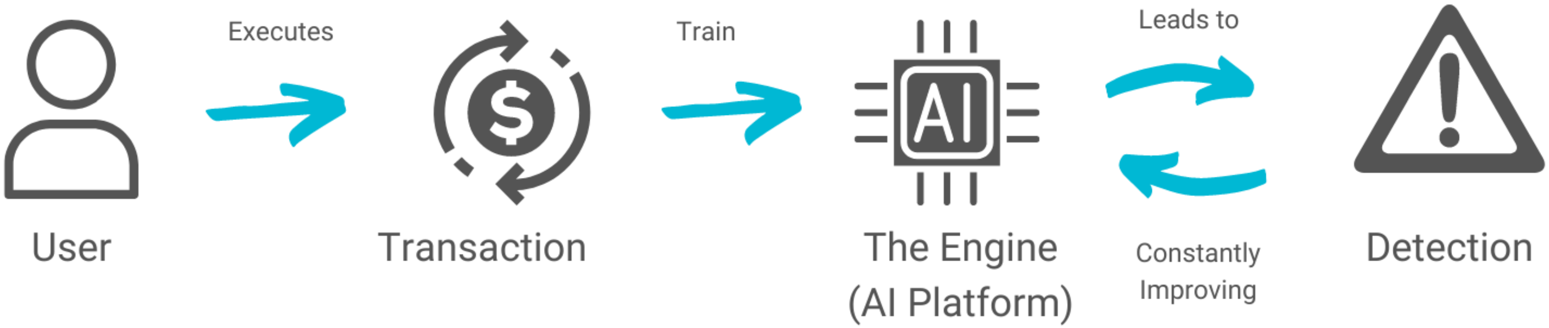
Machine Learning has a lot of power that can be understood by its features. In today's data-rich environment, there are a lot of examples that mirror the characteristics of Machine Learning. Here are some key features of Machine Learning:

1. Automated Data Visualization: Machine Learning provides a variety of techniques that generate rich data snippets that can be used for both unstructured and structured data. Businesses may get many fresh insights to boost efficiency in their operations by utilizing user-friendly automated Data Visualization tools in Machine Learning.
2. Better Customer Engagement: Machine Learning is crucial in helping organizations or businesses to start more effective customer engagement dialogues. Machine Learning techniques examine certain words, phrases, sentences, and material styles that appeal to a specific audience.
3. Better Analysis: With the help of Machine Learning, people can quickly and efficiently process enormous amounts of data. Machine Learning may create correct analysis and outcomes by designing rapid and efficient algorithms and data-driven models for real-time Data Analysis.
4. Improved Business Intelligence: When Machine Learning features are combined with Data Analytics work, they can produce extraordinary Business Intelligence. This has helped several companies in making strategic initiatives.

TRADITIONAL RULE-BASED APPROACH



MACHINE LEARNING APPROACH



Social Network Analysis

Types of Social Networks Analysis

Social networks are the networks that depict the relations between people in the form of a graph for different kinds of analysis. The graph to store the relationships of people is known as Sociogram. All the graph points and lines are stored in the matrix data structure called Sociomatrix. The relationships indicate of any kind like kinship, friendship, enemies, acquaintances, colleagues, neighbours, disease transmission, etc.

Social Network Analysis (SNA) is the process of exploring or examining the social structure by using graph theory. It is used for measuring and analysing the structural properties of the network. It helps to measure relationships and flows between groups, organizations, and other connected entities. We need specialized tools to study and analyse social networks.

Basically, there are two types of social networks:

- Ego network Analysis
- Complete network Analysis

Ego Network Analysis

Ego network Analysis is the one that finds the relationship among people. The analysis is done for a particular sample of people chosen from the whole population. This sampling is done randomly to analyze the relationship. The attributes involved in this ego network analysis are a person’s size, diversity, etc.

This analysis is done by traditional surveys. The surveys involve that they people are asked with whom they interact with and their name of the relationship between them. It is not focused to find the relationship between everyone in the sample. It is an effort to find the density of the network in those samples. This hypothesis is tested using some statistical hypothesis testing techniques.

The following functions are served by Ego Networks:

- Propagation of information efficiently.
- Sensemaking from links, For example, Social links, relationships.
- Access to resources, efficient connection path generation.
- Community detection, identification of the formation of groups.
- Analysis of the ties among individuals for social support.

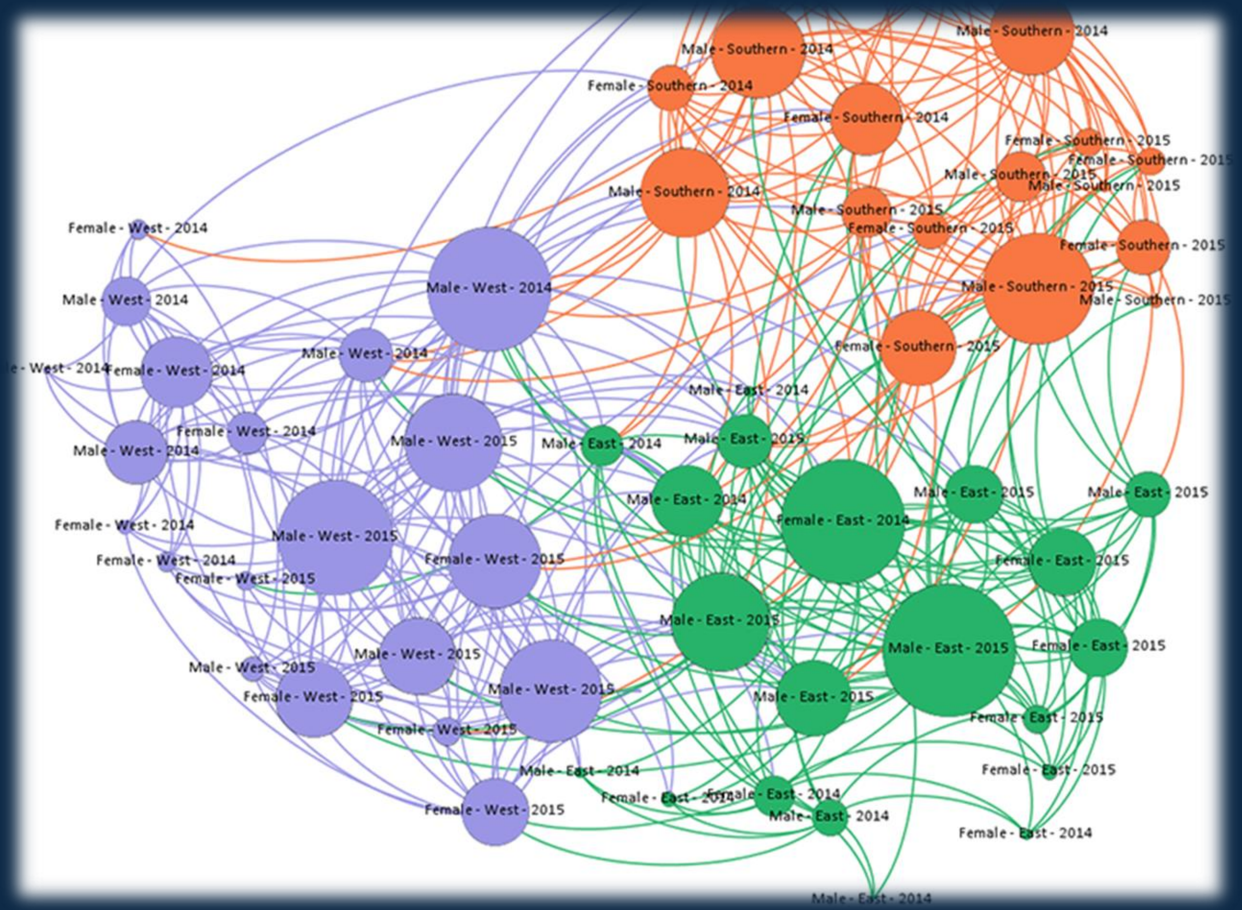
Complete Network Analysis

Complete network analysis is the analysis that is used in all network analyses. It analyses the relationship among the sample of people chosen from the large population. Subgroup analysis, centrality measure, and equivalence analysis are based on the complete network analysis. This analysis measure helps the organization or the company to make any decision with the help of their relationship. Testing the sample will show the relationship in the whole network since the sample is taken from a single set of domains.

Difference between Ego network analysis and Complete network analysis:

The difference between ego and complete network analysis is that the ego network focus on collecting the relationship of people in the sample with the outside world whereas, in Complete network, it is focused on finding the relationship among the samples.

The majority of the network analysis will be done only for a particular domain or one organization. It is not focused on the relationships between the organization. So many of the social network analysis measure uses only Complete network analysis.


Male - Southern - 2014
Male - Southern - 2014
Female - Southern - 2014
Female - Southern - 2014
Female - Southern - 2015
Female - Southern - 2015
Male - Southern - 2015
Male - Southern - 2015
Male - Southern - 2014
Male - Southern - 2015
Female - Southern - 2015
Female - West - 2014
Female - Southern - 2015
Male - Southern - 2015
Male - West - 2014
Male - West - 2014
Male - Southern - 2015
Female - Southern - 2015
Male - West - 2014
Female - West - 2014
Male - West - 2014
Male - East - 2014
Female - West - 2014
Male - West - 2015
Male - East - 2014
Male - East - 2015
Male - West - 2014
Male - East - 2015
Male - East - 2015
Male - East - 2014
Female - East - 2014
Female - West - 2014
Male - West - 2015
Female - East - 2015
Female - West - 2014
Female - West - 2015
Female - West - 2015
Male - East - 2015
Male - East - 2015
Female - East - 2015
Male - West - 2015
Male - West - 2015
Female - West - 2015
Male - West - 2015
Male - East - 2015
Female - East - 2015
Female - West - 2015
Male - East - 2014
Female - East - 2015
Female - East - 2014
Female - East - 2014
Male - East - 2014
Female - West - 2015
Female - East - 2014
Male - East - 2014

Predictive Modeling

What Is Predictive Modelling?

Predictive modelling is a method of predicting future outcomes by using data modelling. It’s one of the premier ways a business can see its path forward and make plans accordingly. While not fool proof, this method tends to have high accuracy rates, which is why it is so commonly used.

In short, predictive modelling is a statistical technique using machine learning and data mining to predict and forecast likely future outcomes with the aid of historical and existing data. It works by analysing current and historical data and projecting what it learns on a model generated to forecast likely outcomes. Predictive modeling can be used to predict just about anything, from TV ratings and a customer’s next purchase to credit risks and corporate earnings.

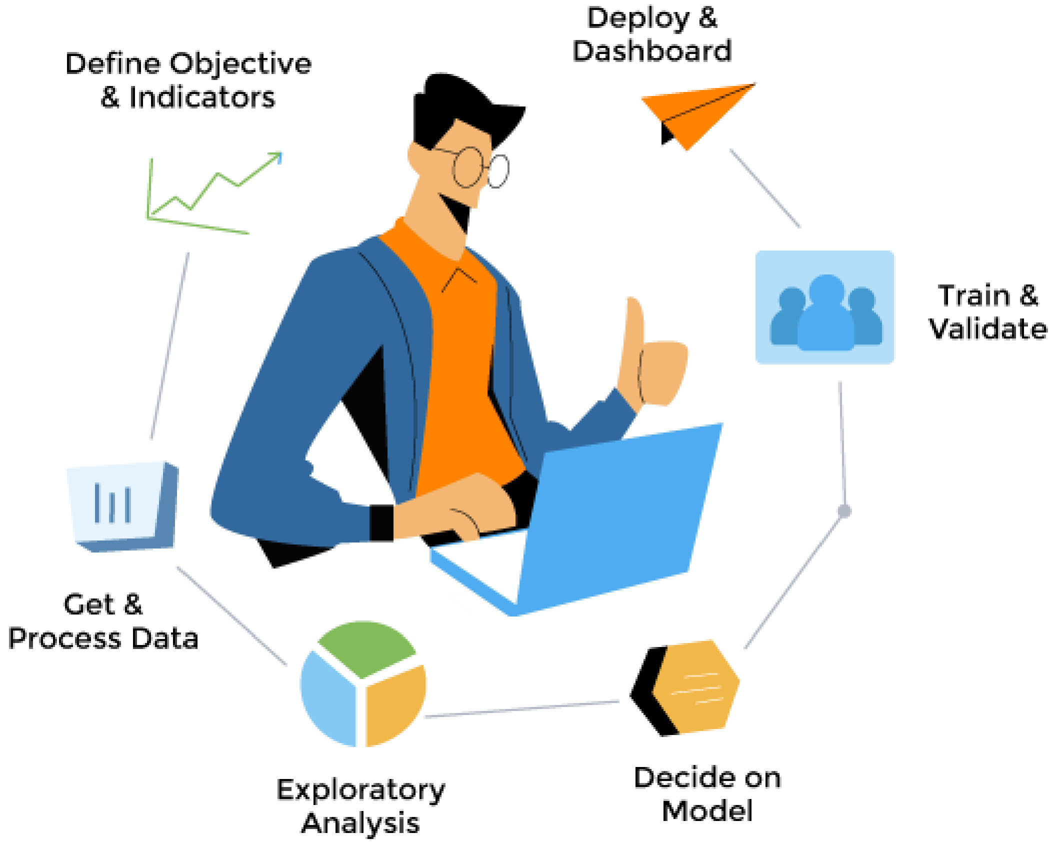
Top 5 Types of Predictive Models

1. **Classification model:** Considered the simplest model, it categorizes data for simple and direct query response. An example use case would be to answer the question “Is this a fraudulent transaction?”
2. **Clustering model:** This model nests data together by common attributes. It works by grouping things or people with shared characteristics or behaviours and plans strategies for each group at a larger scale. An example is in determining credit risk for a loan applicant based on what other people in the same or a similar situation did in the past.
3. **Forecast model:** This is a very popular model, and it works on anything with a numerical value based on learning from historical data. For example, in answering how much lettuce a restaurant should order next week or how many calls a customer support agent should be able to handle per day or week, the system looks back to historical data.
4. **Outliers model:** This model works by analysing abnormal or outlying data points. For example, a bank might use an outlier model to identify fraud by asking whether a transaction is outside of the customer’s normal buying habits or whether an expense in a given category is normal or not. For example, a $1,000 credit card charge for a washer and dryer in the cardholder’s preferred big box store would not be alarming, but $1,000 spent on designer clothing in a location where the customer has never charged other items might be indicative of a breached account.
5. **Time series model:** This model evaluates a sequence of data points based on time. For example, the number of stroke patients admitted to the hospital in the last four months is used to predict how many patients the hospital might expect to admit next week, next month or the rest of the year. A single metric measured and compared over time is thus more meaningful than a simple average.

Common Predictive Algorithms

1. Random Forest: This algorithm is derived from a combination of decision trees, none of which are related, and can use both classification and regression to classify vast amounts of data.
2. Generalized Linear Model (GLM) for Two Values: This algorithm narrows down the list of variables to find “best fit.” It can work out tipping points and change data capture and other influences, such as categorical predictors, to determine the “best fit” outcome, thereby overcoming drawbacks in other models, such as a regular linear regression.
3. Gradient Boosted Model: This algorithm also uses several combined decision trees, but unlike Random Forest, the trees are related. It builds out one tree at a time, thus enabling the next tree to correct flaws in the previous tree. It’s often used in rankings, such as on search engine outputs.
4. K-Means: A popular and fast algorithm, K-Means groups data points by similarities and so is often used for the clustering model. It can quickly render things like personalized retail offers to individuals within a huge group, such as a million or more customers with a similar liking of lined red wool coats.
5. Prophet: This algorithm is used in time-series or forecast models for capacity planning, such as for inventory needs, sales quotas and resource allocations. It is highly flexible and can easily accommodate heuristics and an array of useful assumptions.

Predictive Modeling



Case Study

Fraud Detection in Python
Credit Card

- A typical organization loses an estimated 5% of its yearly revenue to fraud. In this case study, we learn to fight fraud by using data. Apply supervised learning algorithms to detect fraudulent behavior based upon past fraud, and use unsupervised learning methods to discover new types of fraud activities.
- Fraudulent transactions are rare compared to the norm. As such, learn to properly classify imbalanced datasets.

Introduction to fraud detection

- Types:
  - Insurance
  - Credit card
  - Identity theft
  - Money laundering
  - Tax evasion
  - Healthcare
  - Product warranty
- e-commerce businesses must continuously assess the legitimacy of client transactions
- Detecting fraud is challenging:
  - Uncommon; < 0.01% of transactions
  - Attempts are made to conceal fraud
  - Behavior evolves
  - Fraudulent activities perpetrated by networks - organized crime

Introduction to fraud detection

- Fraud detection requires training an algorithm to identify concealed observations from any normal observations
- Fraud analytics teams:
  - Often use rules based systems, based on manually set thresholds and experience
  - Check the news
  - Receive external lists of fraudulent accounts and names

    - suspicious names or track an external hit list from police to reference check against the client base
  - Sometimes use machine learning algorithms to detect fraud or suspicious behavior

    - Existing sources can be used as inputs into the ML model

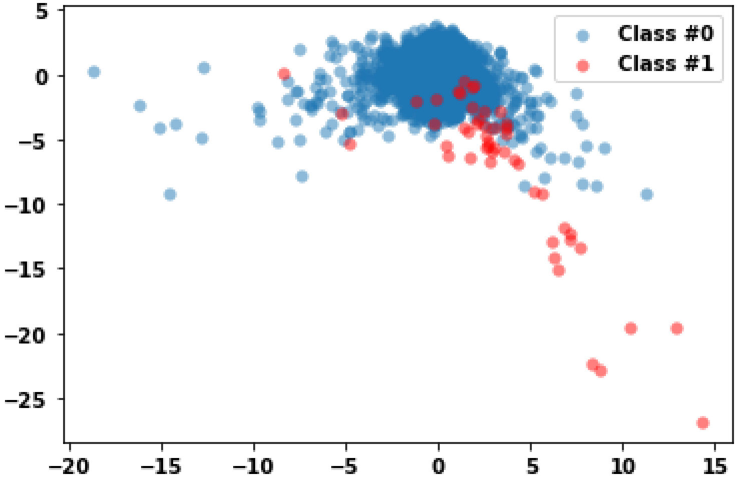
    - Verify the veracity of rules based labels

Creditcard.csv

- The dataset contains transactions made by credit cards in September 2013 by European cardholders.
  This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.
- It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.
- Given the class imbalance ratio, we recommend measuring the accuracy using the Area Under the Precision-Recall Curve (AUPRC). Confusion matrix accuracy is not meaningful for unbalanced classification.

Creditcard.csv

```
Ratio of fraudulent cases: (Class1) 0.009900990099009901
Ratio of non-fraudulent cases: (Class0) 0.9900990099009901
Occurrence:
Class 1: 50
Class 0: 5000
```



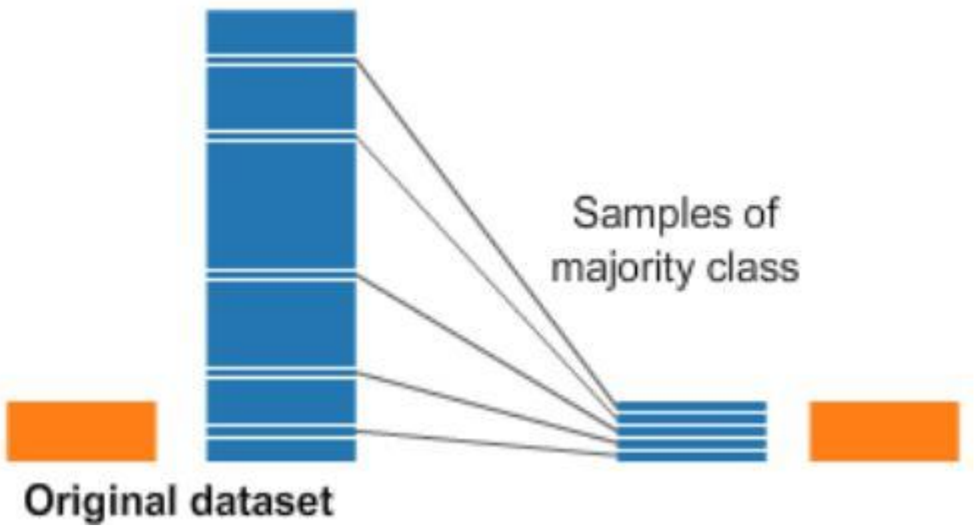
Creditcard.csv

- **Increase successful detections with data resampling**
- resampling can help model performance in cases of imbalanced data sets

  Undersampling the majority class (non-fraud cases)

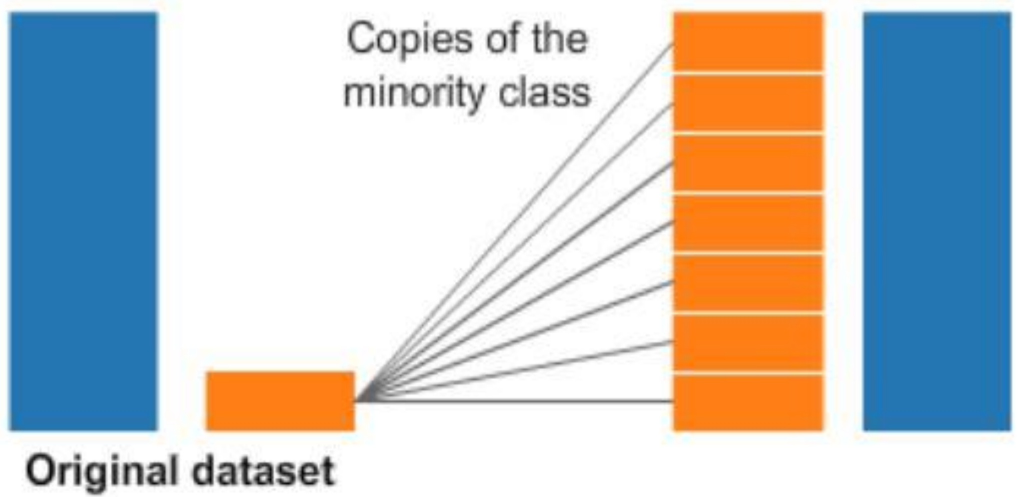
  Straightforward method to adjust imbalanced data

  Take random draws from the non-fraud observations, to match the occurrences of fraud observations (as shown in the picture)

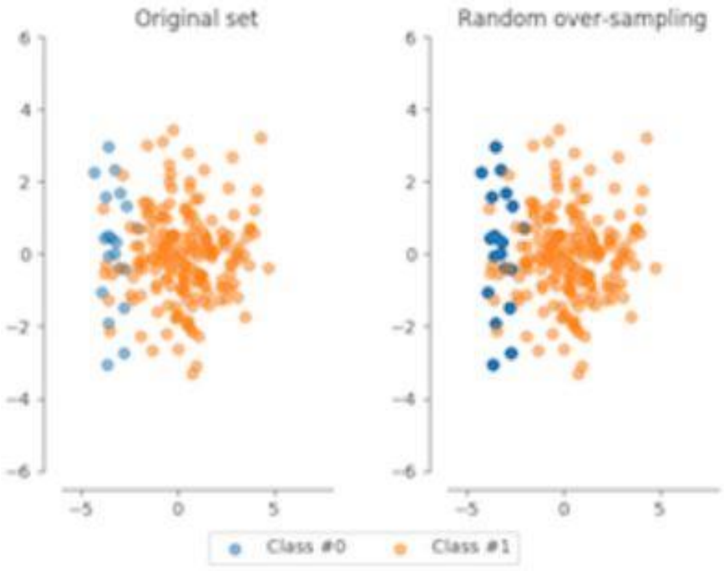


Oversampling

- Oversampling the minority class (fraud cases)
  - Take random draws from the fraud cases and copy those observations to increase the amount of fraud samples
- Both methods lead to having a balance between fraud and non-fraud cases
- Drawbacks
  - with random undersampling, a lot of information is thrown away
  - with oversampling,
  - the model will be trained on a lot of duplicates



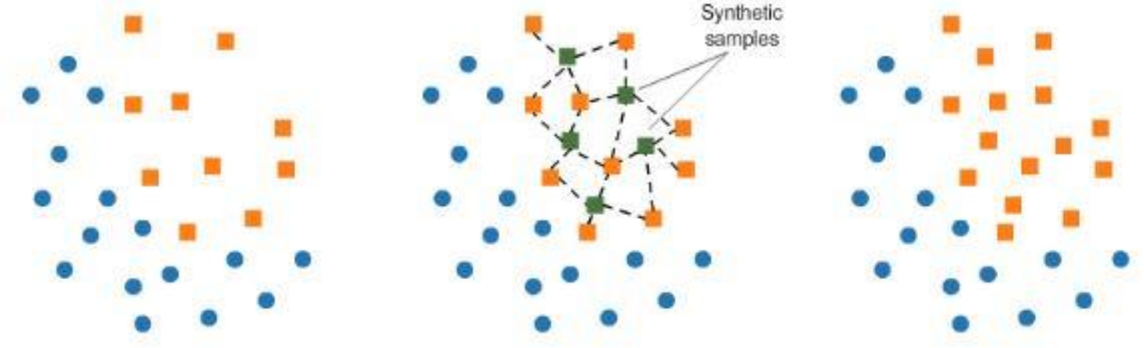
The darker blue points reflect there are more identical data



Synthetic minority Oversampling Technique (SMOTE)

Resampling strategies for Imbalanced Data Sets

Another way of adjusting the imbalance by oversampling minority observations SMOTE uses characteristics of nearest neighbors of fraud cases to create new synthetic fraud cases avoids duplicating observations



Determining the best resampling method is situational

Random Under sampling (RUS):

If there is a lot of data and many minority cases, then under sampling may be computationally more convenient. In most cases, throwing away data is not desirable

Random Oversampling (ROS):

Straightforward

Training the model on many duplicates

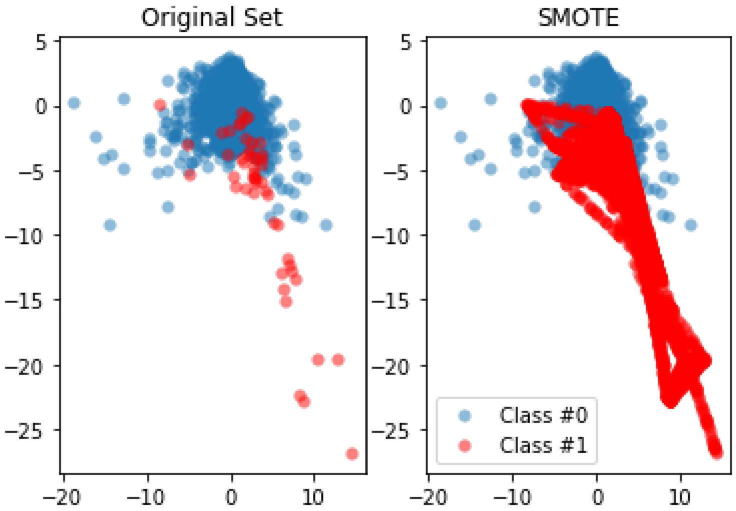
SMOTE:

- more sophisticated
- realistic data set
- training on synthetic data
- only works well if the minority case features are similar

if fraud is spread through the data and not distinct, using nearest neighbors to create more fraud cases, introduces noise into the data, as the nearest neighbors might not be fraud cases

Applying Synthetic Minority Oversampling Technique (SMOTE)

In this exercise, you're going to re-balance our data using the **Synthetic Minority Over-sampling Technique** (SMOTE). Unlike ROS, SMOTE does not create exact copies of observations, but **creates new, synthetic, samples** that are quite similar to the existing observations in the minority class. SMOTE is therefore slightly more sophisticated than just copying observations, so let's apply SMOTE to our credit card data. The dataset df is available and the packages you need for SMOTE are imported. In the following exercise, you'll visualize the result and compare it to the original data, such that you can see the effect of applying SMOTE very clearly.



Fraud detection algorithms in action

Rules Based Systems

- Might block transactions from risky zip codes
- Block transactions from cards used too frequently (e.g. last 30 minutes)
- Can catch fraud, but also generates false alarms (false positive)
- Limitations:

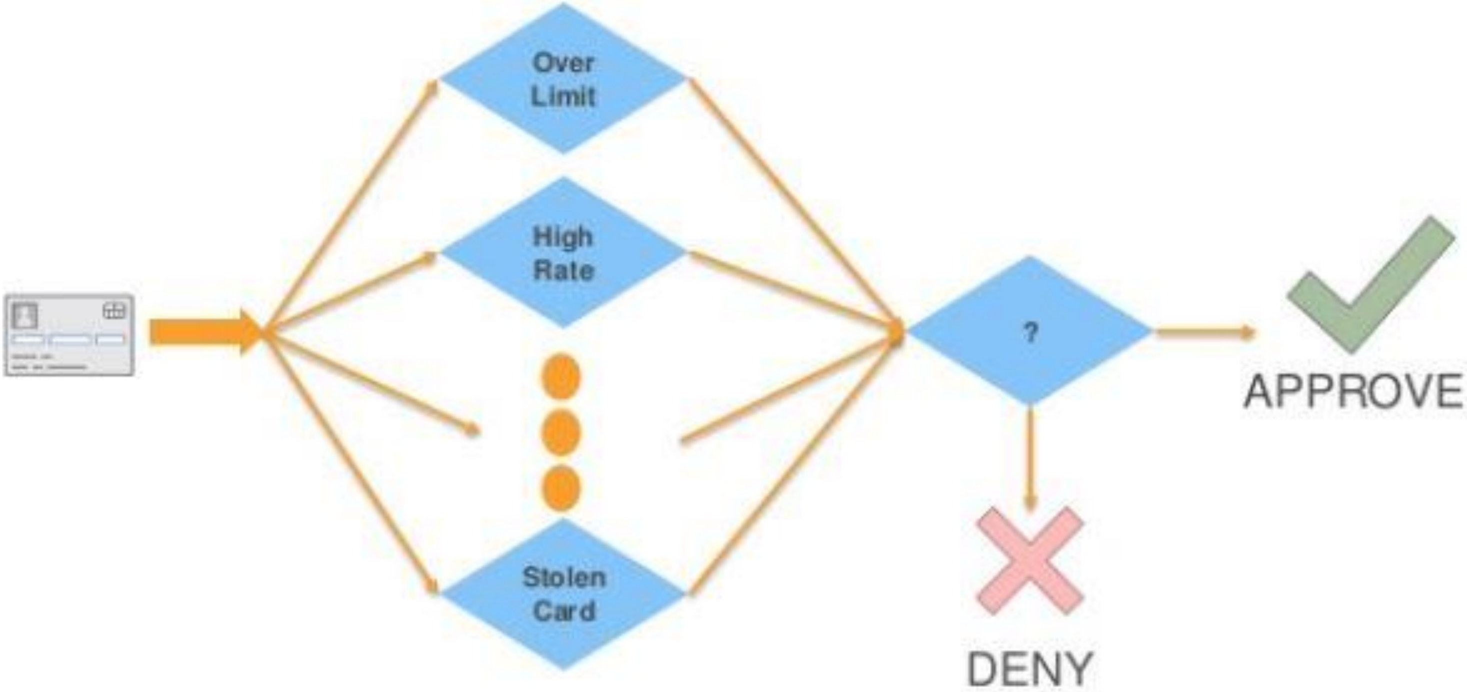
Fixed threshold per rule and it's difficult to determine the threshold; they don't adapt over time

Limited to yes / no outcomes, whereas ML yields a probability

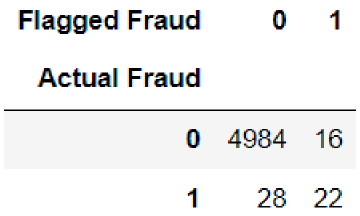
probability allows for fine-tuning the outcomes

(i.e. rate of occurences of false positives and false negatives)

- Fails to capture interaction between features
  - Ex. Size of the transaction only matters in combination to the frequency\


Over Limit
High Rate
Stolen Card
?
APPROVE
DENY

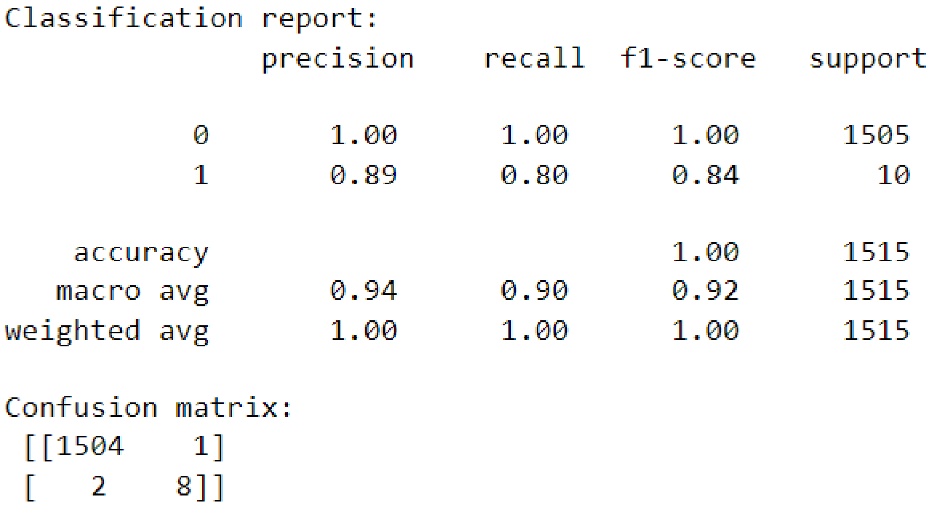
Exploring the traditional method of fraud detection



Flagged Fraud	0	1
Actual Fraud		
0	4984	16
1	28	22

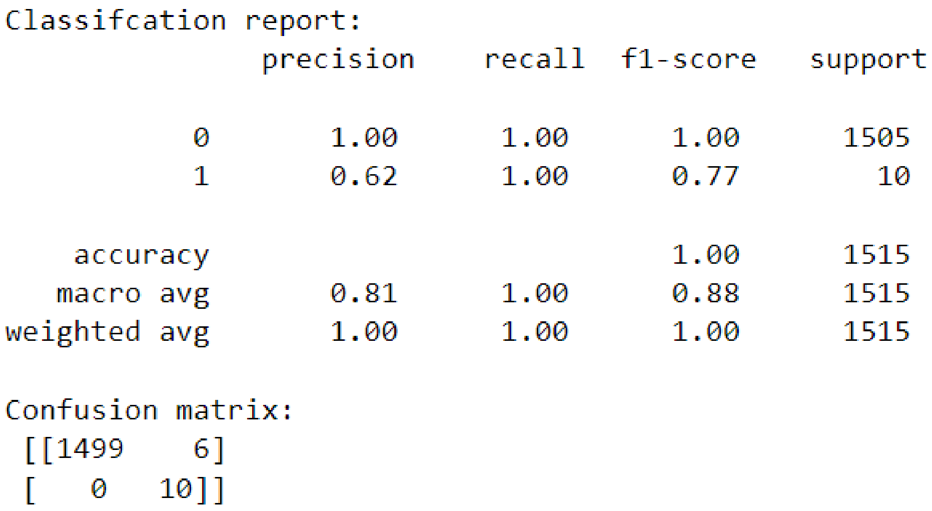
With this rule, 22 out of 50 fraud cases are detected, 28 are not detected, and 16 false positives are identified.

Using ML classification to catch fraud



```
Classification report:
               precision    recall  f1-score   support

           0       1.00      1.00      1.00      1505
           1       0.89      0.80      0.84        10

    accuracy                           1.00      1515
   macro avg       0.94      0.90      0.92      1515
weighted avg       1.00      1.00      1.00      1515

Confusion matrix:
 [[1504    1]
 [   2    8]]
```

Logistic regression with SMOTE



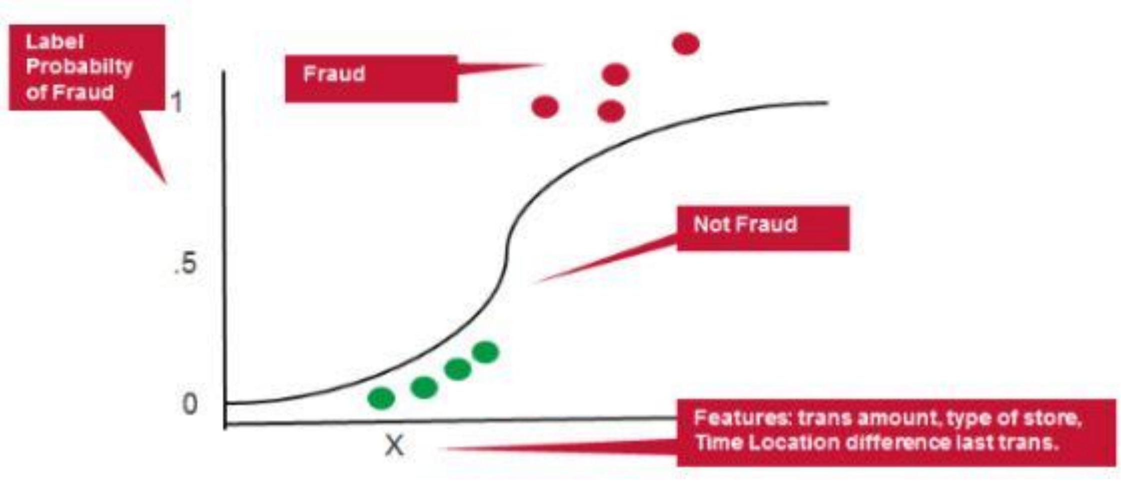
```
Classifcation report:
               precision    recall  f1-score   support

           0       1.00      1.00      1.00      1505
           1       0.62      1.00      0.77        10

    accuracy                           1.00      1515
   macro avg       0.81      1.00      0.88      1515
weighted avg       1.00      1.00      1.00      1515

Confusion matrix:
 [[1499    6]
 [   0   10]]
```

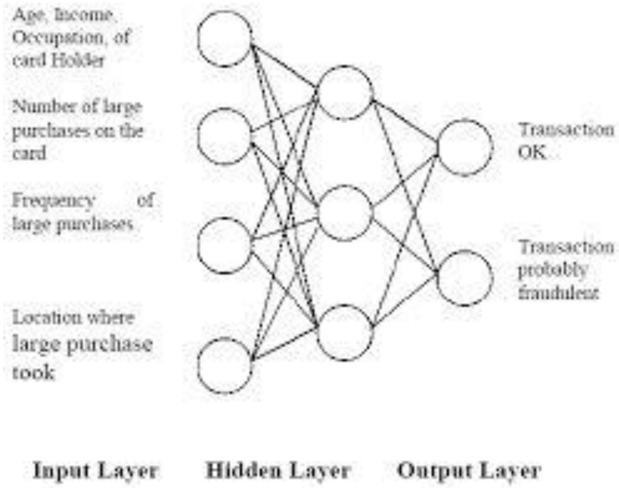
As we can see, the SMOTE slightly improves our results. We now manage to find all cases of fraud, but we have a slightly higher number of false positives, albeit only 7 cases. Resampling doesn't necessarily lead to better results. When the fraud cases are very spread and scattered over the data, using SMOTE can introduce a bit of bias. Nearest neighbors aren't necessarily also fraud cases, so the synthetic samples might 'confuse' the model slightly

Fraud detection using labeled data

Classification methods

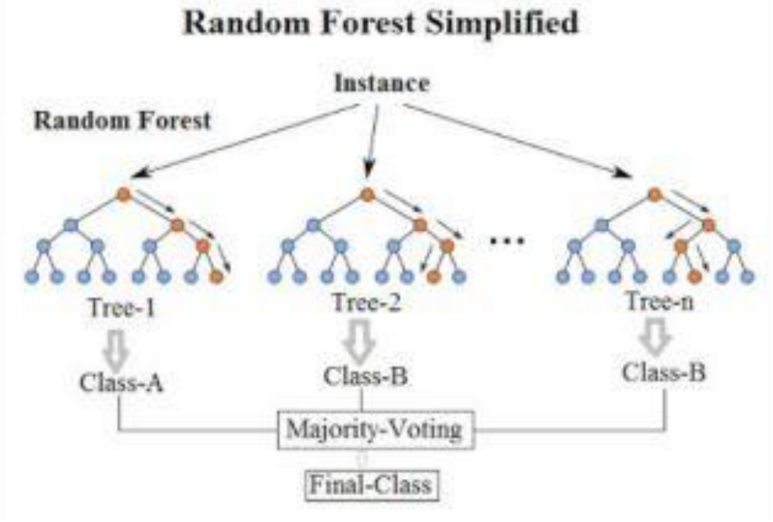
Logistic Regression

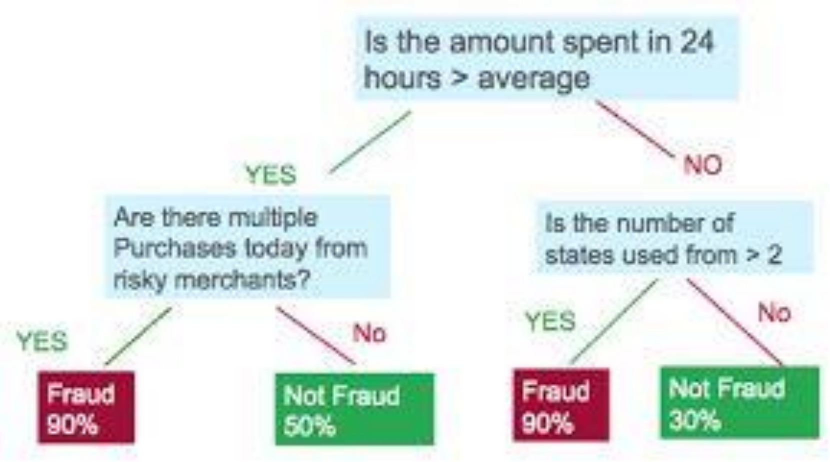



Age, Income, Occupation, of card Holder
Number of large purchases on the card
Frequency of large purchases
Location where large purchase took
Transaction OK
Transaction probably fraudulent
Input Layer
Hidden Layer
Output Layer

Neural Network

Random Forest




Is the amount spent in 24 hours > average
YES
NO
Are there multiple Purchases today from risky merchants?
Is the number of states used from > 2
YES
No
YES
No
Fraud 90%
Not Fraud 50%
Fraud 90%
Not Fraud 30%

Decision Tree

Thank you!