Report

On

Credit Card Fraud Detection using machine learning

Submitted in partial fulfillment of the

requirement for the award of the degree of

Computer Science and Engineering

Under The Supervision of
Dr. Sampath Kumar k

Submitted by

Garvit Kashyap
(18SCSE1010473)
Ujjwal Raj
(18SCSE1010660)

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA
DEC , 2021

CANDIDATE'S DECLARATION

I/We hereby certify that the work which is being presented in the thesis/project/dissertation, entitled **"CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING"** in partial fulfillment of the requirements for the award of the B. Tech (CSE) submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of Oct 2021 to Dec 2021, under the supervision of Dr. Sampath Kumar K, Department of Computer Science and Engineering, of School of Computing Science and Engineering , Galgotias University, Greater Noida.

The matter presented in the thesis/project/dissertation has not been submitted by me/us for the award of any other degree of this or any other places.

GARVIT KASHYAP (18SCSE1010473)

UJJWAL RAJ (18SCSE1010473)

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Dr. Sampath Kumar K.

CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of GARVIT KASHYAP (18SCSE1010473) and UJJWAL RAJ(18SCSE1010660) has been held on and his/her work is recommended for the award of B Tech(CSE).

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date: December 2021

Place: Greater Noida

Table of Contents

Abstract

In todays world, the Internet is a part of our life. Due to the extensive use of internet, the popularity of online shopping is growing day by day. Credit Card is the simplest method to do online shopping and paying bills. Thus Credit Card become very popular and convenient mode for online money transaction and is increasing very rapidly. With the increase of Credit Card usage, the opportunities for fraudster to steal credit card details and subsequently commit fraud are also increasing.

Credit Card fraud is the fraud committed by the use of another persons credit card. To support safe credit card usage an efficient fraud detection system is essential. Presently, many modern techniques based on Artificial Intelligence, Sequence Alignment, Data Mining, Fuzzy Logic, Machine Learning, Genetic Programming etc. has been introduced for detecting various credit card fraudulent transactions. This paper presents a survey of various current techniques used in fraud detection mechanism and provides a comprehensive review of different techniques based on certain design criteria.

The prediction analysis is the approach which can predict future possibilities on the current data. When the physical-card based purchasing technique is applied, the card is given by the cardholder to the merchant so that a successful payment method can be performed. The fraudulent transactions are conducted by the attacker by stealing the credit card. When the loss of the card is not noticed by the cardholder, a huge loss can be faced by the credit card company. A very little amount of information is required by the attacker for conducting any fraudulent transaction in online transactions. In this research work, various credit card fraud detection techniques are reviewed in terms of certain parameters.

Keywords: Machine learning, Classification, Credit card fraud Detection

INTRODUCTION

A credit card is a thin handy plastic card that contains identification information such as a signature or picture, and authorizes the person named on it to charge purchases or services to his account - charges for which he will be billed periodically. Today, the information on the card is read by automated teller machines (ATMs), store readers, bank and is also used in online internet banking system. They have a unique card number which is of utmost importance. Its security relies on the physical security of the plastic card as well as the privacy of the credit card number. There is a rapid growth in the number of credit card transactions which has led to a substantial rise in fraudulent activities. Credit card fraud is a wide-ranging term for theft and fraud committed using a credit card as a fraudulent source of funds in a given transaction. Generally, the statistical methods and many data mining algorithms are used to solve this fraud detection problem. Most of the credit card fraud detection systems are based on artificial intelligence, Meta learning and pattern matching. The Genetic algorithms are evolutionary algorithms which aim to obtain the better solutions in eliminating the fraud. A high importance is given to develop efficient and secure electronic payment system to detect whether a transaction is fraudulent or not.

Literature Survey

Literature Review Kuldeep Randhawa et al. proposed a technique using machine learning to detect credit card fraud detection. Initially, standard models were used after that hybrid models came into picture which made use of AdaBoost and majority voting methods. Publically available data set had been used to evaluate the model efficiency and another data set used from the financial institution and analyzed the fraud. Then the noise was added to the data sample through which the robustness of the algorithms could be measured. The experiments were conducted on the basis of the theoretical results which show that the majority of voting methods achieve good accuracy rates in order to detect the fraud in the credit cards. For further evaluation of the hybrid models noise of about 10% and 30% has been added to the sample data. Several voting methods have achieved a good score of 0.942 for 30% added noise. Thus, it was concluded that the voting method showed much stable performance in the presence of noise.

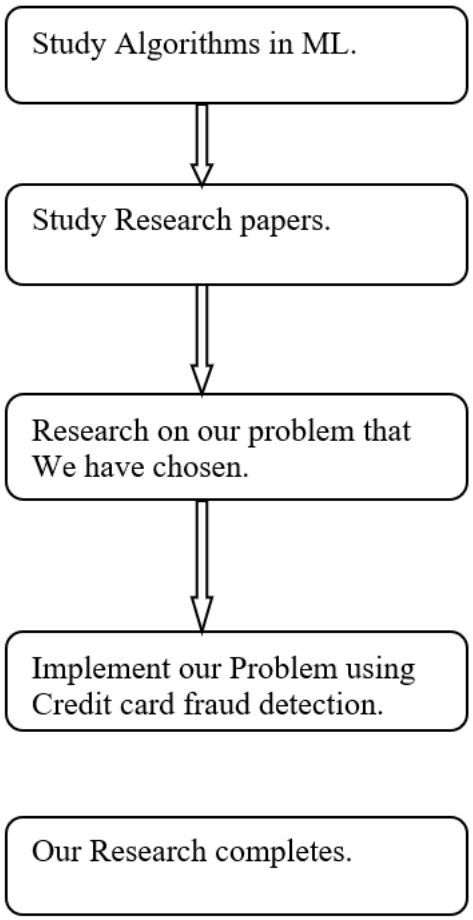
Abhimanyu Roy et al. proposed deep learning topologies for the detection of fraud in online money transaction. This approach is derived from the artificial neural network with in-built time and memory components like long term short term memory and several other parameters. According to the efficiency of these components in fraud detection, almost 80 million online transactions through credit card have been pre-labeled as fraudulent and legal. They have used high performance distributed cloud computing environment. The study proposed by the researchers provides an effective guide to the sensitivity analysis of the proposed parameters as per the performance of the fraud detection. The researchers also proposed a framework for the parameter tuning of Deep Learning topologies for the detection of fraud. This enables the financial institution to decrease the losses by avoiding fraudulent activities.

Sharmistha Dutta et al. [15] presented a study on the commonly found crime within the credit card applications. There are certain issues faced when the existing non-data mining approaches are applied to avoid identity theft. A novel data mining layer of defense is proposed for solving these issues. For detecting the frauds within various applications, two algorithms named Communal Detection and Spike Detection which generate novel layer. There is a large moving window, higher numbers of attributes and numbers of link types available which can be searched by CD and SD algorithms. Thus, results can be generated by the system by consuming a huge amount of time. Since the attackers do not get time to modify their behaviors with respect to the algorithms being deployed in real time, there is no true evaluation achieved even after a regular update of the algorithms. Therefore, it is not possible to properly demonstrate the concept of adaptability. These issues can be resolved by making certain enhancements in the proposed algorithm in future work.

Problem Statement

Credit card frauds are increasing heavily because of fraud financial loss is increasing drastically. Every year due to fraud Billions of amounts lost. To analyze the fraud there is lack of research. Many machine learning algorithms are implemented to detect real world credit card fraud. Logistic Regression algorithms are applied

DFD diagram



What Is Data Science?

Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions. Data science uses complex machine learning algorithms to build predictive models.

The data used for analysis can come from many different sources and presented in various formats.

Now that you know what data science is, let's see why data science is essential to today's IT landscape.

The Data Science Lifecycle

Data science's lifecycle consists of five distinct stages, each with its own tasks:

1. Capture: Data Acquisition, Data Entry, Signal Reception, Data Extraction. This stage involves gathering raw structured and unstructured data.
2. Maintain: Data Warehousing, Data Cleansing, Data Staging, Data Processing, Data Architecture. This stage covers taking the raw data and putting it in a form that can be used.
3. Process: Data Mining, Clustering/Classification, Data Modeling, Data Summarization. Data scientists take the prepared data and examine its patterns, ranges, and biases to determine how useful it will be in predictive analysis.
4. Analyze: Exploratory/Confirmatory, Predictive Analysis, Regression, Text Mining, Qualitative Analysis. Here is the real meat of the lifecycle. This stage involves performing the various analyses on the data.
5. Communicate: Data Reporting, Data Visualization, Business Intelligence, Decision Making. In this final step, analysts prepare the analyses in easily readable forms such as charts, graphs, and reports.

Prerequisites for Data Science

Here are some of the technical concepts you should know about before starting to learn what is data science.

1. Machine Learning

Machine learning is the backbone of data science. Data Scientists need to have a solid grasp of ML in addition to basic knowledge of statistics.

2. Modeling

Mathematical models enable you to make quick calculations and predictions based on what you already know about the data. Modeling is also a part of Machine Learning and involves identifying which algorithm is the most suitable to solve a given problem and how to train these models.

3. Statistics

Statistics are at the core of data science. A sturdy handle on statistics can help you extract more intelligence and obtain more meaningful results.

4. Programming

Some level of programming is required to execute a successful data science project. The most common programming languages are Python, and R. Python is especially popular because it's easy to learn, and it supports multiple libraries for data science and ML.

5. Databases

A capable data scientist needs to understand how databases work, how to manage them, and how to extract data from them.

What Does a Data Scientist Do?

A data scientist analyzes business data to extract meaningful insights. In other words, a data scientist solves business problems through a series of steps, including:

- Before tackling the data collection and analysis, the data scientist determines the problem by asking the right questions and gaining understanding.
- The data scientist then determines the correct set of variables and data sets.
- The data scientist gathers structured and unstructured data from many disparate sources—enterprise data, public data, etc.
- Once the data is collected, the data scientist processes the raw data and converts it into a format suitable for analysis. This involves cleaning and validating the data to guarantee uniformity, completeness, and accuracy.
- After the data has been rendered into a usable form, it's fed into the analytic system—ML algorithm or a statistical model. This is where the data scientists analyze and identify patterns and trends.
- When the data has been completely rendered, the data scientist interprets the data to find opportunities and solutions.
- The data scientists finish the task by preparing the results and insights to share with the appropriate stakeholders and communicating the results.

Now we should be aware of some machine learning algorithms which are beneficial in understanding data science clearly.

Why Become a Data Scientist?

According to Glassdoor and Forbes, demand for data scientists will increase by 28 percent by 2026, which speaks of the profession's durability and longevity, so if you want a secure career, data science offers you that chance.

So, if you're looking for an exciting career that offers stability and generous compensation, then look no further!

Where Do You Fit in Data Science?

Data science offers you the opportunity to focus on and specialize in one aspect of the field. Here's a sample of different ways you can fit into this exciting, fast-growing field.

Data Scientist

- Job role: Determine what the problem is, what questions need answers, and where to find the data. Also, they mine, clean, and present the relevant data.
- Skills needed: Programming skills (SAS, R, Python), storytelling and data visualization, statistical and mathematical skills, knowledge of Hadoop, SQL, and Machine Learning.

Data Analyst

- Job role: Analysts bridge the gap between the data scientists and the business analysts, organizing and analyzing data to answer the questions the organization poses. They take the technical analyses and turn them into qualitative action items.
- Skills needed: Statistical and mathematical skills, programming skills (SAS, R, Python), plus experience in data wrangling and data visualization.

Data Engineer

- Job role: Data engineers focus on developing, deploying, managing, and optimizing the organization's data infrastructure and data pipelines. Engineers support data scientists by helping to transfer and transform data for queries.
- Skills needed: NoSQL databases (e.g., MongoDB, Cassandra DB), programming languages such as Java and Scala, and frameworks (Apache Hadoop).

Data Science Tools

The data science profession is challenging, but fortunately, there are plenty of tools available to help the data scientist succeed at their job.

- Data Analysis: SAS, Jupyter, R Studio, MATLAB, Excel, RapidMiner
- Data Warehousing: Informatica/ Talend, AWS Redshift
- Data Visualization: Jupyter, Tableau, Cognos, RAW
- Machine Learning: Spark MLib, Mahout, Azure ML studio

The Basic Skills You Need to Become a Data Scientist

- Mathematical Expertise: There is a commonly circulated meme about grownups realizing that studying algebra was useless because there are no opportunities to use it in everyday life. Surprise! Data scientists need to understand linear algebra, as well as quantitative techniques.

- A Strong Business Acumen: Data scientists are supposed to derive information that is useful to businesses and share it with the appropriate individuals and teams. So, data scientists need to have a solid business understanding so they can have the correct perspective when making these determinations.
- Technology Skills: Data scientists work with sophisticated tools and complex algorithms. They also may be called on to code and develop solutions prototypes rapidly. These expectations mean the data scientist should have proficiency in languages like SQL, R, Python, and SAS, and occasionally in Java, Scala, and Julia.
- Project Management: Data scientists must oversee projects that rely heavily on the data they collect and process. It's up to the data scientists to ensure that things are moving forward and everyone is communicating with each other.

- In computing, data is information that has been translated into a form that is efficient for movement or processing. Relative to today's computers and transmission media, data is information converted into binary digital form. It is acceptable for data to be used as a singular subject or a plural subject. Raw data is a term used to describe data in its most basic digital format.

- The concept of data in the context of computing has its roots in the work of Claude Shannon, an American mathematician known as the father of information theory. He ushered in binary digital concepts based on applying two-value Boolean logic to electronic circuits. Binary digit formats underlie the CPUs, semiconductor memories and disk drives, as well as many of the peripheral devices common in computing today. Early computer input for both control and data took the form of punch cards, followed by magnetic tape and the hard disk.

- Early on, data's importance in business computing became apparent by the popularity of the terms "data processing" and "electronic data processing," which, for a time, came to encompass the full gamut of what is now known as information technology. Over the history of corporate computing, specialization occurred, and a distinct data profession emerged along with growth of corporate data processing.

How data is stored

- Computers represent data, including video, images, sounds and text, as binary values using patterns of just two numbers: 1 and 0. A bit is the smallest unit of data, and represents just a single value. A byte is eight binary digits long. Storage and memory is measured in megabytes and gigabytes.

- The units of data measurement continue to grow as the amount of data collected and stored grows. The relatively new term "brontobyte," for example, is data storage that is equal to 10 to the 27th power of bytes.

- Data can be stored in file formats, as in mainframe systems using ISAM and VSAM. Other file formats for data storage, conversion and processing include comma-separated values. These formats continued to find uses across a variety of machine types, even as more structured-data-oriented approaches gained footing in corporate computing.

- Greater specialization developed as database, database management system and then relational database technology arose to organize information.

Types of data

Growth of the web and smartphones over the past decade led to a surge in digital data creation. Data now includes text, audio and video information, as well as log and web activity records. Much of that is unstructured data.

The term big data has been used to describe data in the petabyte range or larger. A shorthand take depicts big data with 3Vs -- volume, variety and velocity. As web-based e-commerce has spread, big data-driven business models have evolved which treat data as an asset in itself. Such trends have also spawned greater preoccupation with the social uses of data and data privacy.

Data has meaning beyond its use in computing applications oriented toward data processing. For example, in electronic component interconnection and network communication, the term data is often distinguished from "control information," "control bits," and similar terms to identify the main content of a transmission unit. Moreover, in science, the term data is used to describe a gathered body of facts. That is also the case in fields such as finance, marketing, demographics and health.

Data management and use

With the proliferation of data in organizations, added emphasis has been placed on ensuring data quality by reducing duplication and guaranteeing the most accurate, current

records are used. The many steps involved with modern data management include data cleansing, as well as extract, transform and load (ETL) processes for integrating data. Data for processing has come to be complemented by metadata, sometimes referred to as "data about data," that helps administrators and users understand database and other data.

Analytics that combine structured and unstructured data have become useful, as organizations seek to capitalize on such information. Systems for such analytics increasingly strive for real-time performance, so they are built to handle incoming data consumed at high ingestion rates, and to process data streams for immediate use in operations.

Over time, the idea of the database for operations and transactions has been extended to the database for reporting and predictive data analytics. A chief example is the data warehouse, which is optimized to process questions about operations for business analysts and business leaders. Increasing emphasis on finding patterns and predicting business outcomes has led to the development of data mining techniques.

Data professionals

The database administrator profession is an offshoot of IT. These database experts work on designing, tuning and maintaining the database.

The data profession took firm root as the relational database management system (RDBMS) gained wide use in corporations, beginning in the 1980s. The relational database's rise was enabled in part by the Structured Query Language (SQL). Later, non-SQL databases, known as NoSQL databases, arose as an alternative to established RDBMSes.

Today, companies employ data management professionals or assign workers the role of data stewardship, which involves carrying out data usage and security policies as outlined in data governance initiatives.

A distinct title -- the data scientist -- has appeared to describe professionals focused on data mining and analysis. The benefit of presenting data science in an evocative manner has even given rise to the data artist; that is, an individual adept at graphing and visualizing data in creative ways.

What is data cleaning?

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

What is the difference between data cleaning and data transformation?

Data cleaning is the process that removes data that does not belong in your dataset. Data transformation is the process of converting data from one format or structure into another. Transformation processes can also be referred to as data wrangling, or data munging, transforming and mapping data from one "raw" data form into another format for warehousing and analyzing. This article focuses on the processes of cleaning that data.

How do you clean data?

While the techniques used for data cleaning may vary according to the types of data your company stores, you can follow these basic steps to map out a framework for your organization.

Step 1: Remove duplicate or irrelevant observations

Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations. Duplicate observations will happen most often during data collection. When you combine data sets from multiple places, scrape data, or receive data from clients or multiple departments, there are opportunities to create duplicate data. De-duplication is one of the largest areas to be considered in this process. Irrelevant observations are when you notice observations that do not fit into the specific problem you are trying to analyze. For example, if you want to analyze data regarding millennial customers, but your dataset includes older generations, you might remove those irrelevant observations. This can make analysis more efficient and minimize distraction from your primary target—as well as creating a more manageable and more performant dataset.

Step 2: Fix structural errors

Structural errors are when you measure or transfer data and notice strange naming conventions, typos, or incorrect capitalization. These inconsistencies can cause mislabeled categories or classes. For example, you may find “N/A” and “Not Applicable” both appear, but they should be analyzed as the same category.

Step 3: Filter unwanted outliers

Often, there will be one-off observations where, at a glance, they do not appear to fit within the data you are analyzing. If you have a legitimate reason to remove an outlier, like improper data-entry, doing so will help the performance of the data you are working with. However, sometimes it is the appearance of an outlier that will prove a theory you are working on. Remember: just because an outlier exists, doesn’t mean it is incorrect. This step is needed to determine the validity of that number. If an outlier proves to be irrelevant for analysis or is a mistake, consider removing it.

Step 4: Handle missing data

You can't ignore missing data because many algorithms will not accept missing values. There are a couple of ways to deal with missing data. Neither is optimal, but both can be considered.

1. As a first option, you can drop observations that have missing values, but doing this will drop or lose information, so be mindful of this before you remove it.
2. As a second option, you can input missing values based on other observations; again, there is an opportunity to lose integrity of the data because you may be operating from assumptions and not actual observations.
3. As a third option, you might alter the way the data is used to effectively navigate null values.

Step 5: Validate and QA

At the end of the data cleaning process, you should be able to answer these questions as a part of basic validation:

- Does the data make sense?
- Does the data follow the appropriate rules for its field?
- Does it prove or disprove your working theory, or bring any insight to light?
- Can you find trends in the data to help you form your next theory?
- If not, is that because of a data quality issue?

False conclusions because of incorrect or "dirty" data can inform poor business strategy and decision-making. False conclusions can lead to an embarrassing moment in a reporting meeting when you realize your data doesn't stand up to scrutiny. Before you get there, it is important to create a culture of quality data in your organization. To do this, you should document the tools you might use to create this culture and what data quality means to you.

Components of quality data

Determining the quality of data requires an examination of its characteristics, then weighing those characteristics according to what is most important to your organization and the application(s) for which they will be used.

5 characteristics of quality data

1. Validity. The degree to which your data conforms to defined business rules or constraints.
2. Accuracy. Ensure your data is close to the true values.
3. Completeness. The degree to which all required data is known.

4. Consistency. Ensure your data is consistent within the same dataset and/or across multiple data sets.
5. Uniformity. The degree to which the data is specified using the same unit of measure.

Benefits of data cleaning

Having clean data will ultimately increase overall productivity and allow for the highest quality information in your decision-making. Benefits include:

- Removal of errors when multiple sources of data are at play.
- Fewer errors make for happier clients and less-frustrated employees.
- Ability to map the different functions and what your data is intended to do.
- Monitoring errors and better reporting to see where errors are coming from, making it easier to fix incorrect or corrupt data for future applications.
- Using tools for data cleaning will make for more efficient business practices and quicker decision-making.

Data cleaning tools and software for efficiency

Software like Tableau Prep can help you drive a quality data culture by providing visual and direct ways to combine and clean your data. Tableau Prep has two products: Tableau Prep Builder for building your data flows and Tableau Prep Conductor for scheduling, monitoring, and managing flows across your organization. Using a data scrubbing tool can save a database administrator a significant amount of time by helping analysts or administrators start their analyses faster and have more confidence in the data. Understanding data quality and the tools you need to create, manage, and transform data is an important step toward making efficient and effective business decisions. This crucial process will further develop a data culture in your organization. To see how Tableau Prep can impact your organization, read about how marketing agency Tinuiti centralized 100-plus data sources in Tableau Prep and scaled their marketing analytics for 500 clients.

What is Data Exploration?

DATA EXPLORATION DEFINITION: Data exploration refers to the initial step in data analysis in which data analysts use data visualization and statistical techniques to describe dataset characterizations, such as size, quantity, and accuracy, in order to better understand the nature of the data.

Data exploration techniques include both manual analysis and automated data exploration software solutions that visually explore and identify relationships between different data

variables, the structure of the dataset, the presence of outliers, and the distribution of data values in order to reveal patterns and points of interest, enabling data analysts to gain greater insight into the raw data.

Data is often gathered in large, unstructured volumes from various sources and data analysts must first understand and develop a comprehensive view of the data before extracting relevant data for further analysis, such as univariate, bivariate, multivariate, and principal components analysis.

Data Exploration Tools

Manual data exploration methods entail either writing scripts to analyze raw data or manually filtering data into spreadsheets. Automated data exploration tools, such as data visualization software, help data scientists easily monitor data sources and perform big data exploration on otherwise overwhelmingly large datasets. Graphical displays of data, such as bar charts and scatter plots, are valuable tools in visual data exploration.

A popular tool for manual data exploration is Microsoft Excel spreadsheets, which can be used to create basic charts for data exploration, to view raw data, and to identify the correlation between variables. To identify the correlation between two continuous variables in Excel, use the function CORREL() to return the correlation. To identify the correlation between two categorical variables in Excel, the two-way table method, the stacked column chart method, and the chi-square test are effective.

There is a wide variety of proprietary automated data exploration solutions, including business intelligence tools, data visualization software, data preparation software vendors, and data exploration platforms. There are also open source data exploration tools that include regression capabilities and visualization features, which can help businesses integrate diverse data sources to enable faster data exploration. Most data analytics software includes data visualization tools.

Why is Data Exploration Important?

Humans process visual data better than numerical data, therefore it is extremely challenging for data scientists and data analysts to assign meaning to thousands of rows and columns of data points and communicate that meaning without any visual components.

Data visualization in data exploration leverages familiar visual cues such as shapes, dimensions, colors, lines, points, and angles so that data analysts can effectively visualize and define the metadata, and then perform data cleansing. Performing the initial step of data exploration enables data analysts to better understand and visually identify anomalies and relationships that might otherwise go undetected.

What is Exploratory Data Analysis?

Exploratory Data Analysis (EDA), similar to data exploration, is a statistical technique to analyze data sets for their broad characteristics. Visualization tools for exploratory data analysis such as OmniSci's Immerse platform enable interactivity with raw data sets, giving analysts increased visibility into the patterns and relationships within the data.

Data Exploration in Machine Learning

A Machine Learning project is as good as the foundation of data on which it is built. In order to perform well, machine learning data exploration models must ingest large quantities of data, and model accuracy will suffer if that data is not thoroughly explored first. Data exploration steps to follow before building a machine learning model include:

- Variable identification: define each variable and its role in the dataset
- Univariate analysis: for continuous variables, build box plots or histograms for each variable independently; for categorical variables, build bar charts to show the frequencies
- Bi-variable analysis - determine the interaction between variables by building visualization tools
- ~Continuous and Continuous: scatter plots
- ~Categorical and Categorical: stacked column chart
- ~Categorical and Continuous: boxplots combined with swarmplots
- Detect and treat missing values
- Detect and treat outliers

The ultimate goal of data exploration machine learning is to provide data insights that will inspire subsequent feature engineering and the model-building process. Feature engineering facilitates the machine learning process and increases the predictive power of machine learning algorithms by creating features from raw data.

Interactive Data Exploration

Advanced visualization techniques are employed throughout a variety of disciplines to empower users to visualize patterns and gain insight from complex data flows, and make subsequent data-driven decisions. Industries from engineering to medicine to education are learning how to do data exploration.

In big data exploration tools, interactivity is an important component in the perception of data exploration visual technologies and the dissemination of insights. The manner in which users perceive and interact with visualizations can heavily influence their understanding of the data as well as the value they place on the visualization system in general.

Interactive data exploration emphasizes the importance of collaborative work and facilitates human interaction with the integration of advanced interaction and visualization technologies. Accelerated multimodal interaction platforms equipped with graphical user interfaces that prioritize human-to-human properties facilitate big data exploration through visual analytics, accelerate the sharing of opinions, remove the data bottleneck of individual analysis, and reduce discovery time.

What is Data Visualization?

Data visualization is a graphical representation of quantitative information and data by using visual elements like graphs, charts, and maps.

Data visualization convert large and small data sets into visuals, which is easy to understand and process for humans.

Data visualization tools provide accessible ways to understand outliers, patterns, and trends in the data.

In the world of Big Data, the data visualization tools and technologies are required to analyze vast amounts of information.

Data visualizations are common in your everyday life, but they always appear in the form of graphs and charts. The combination of multiple visualizations and bits of information are still referred to as Infographics.

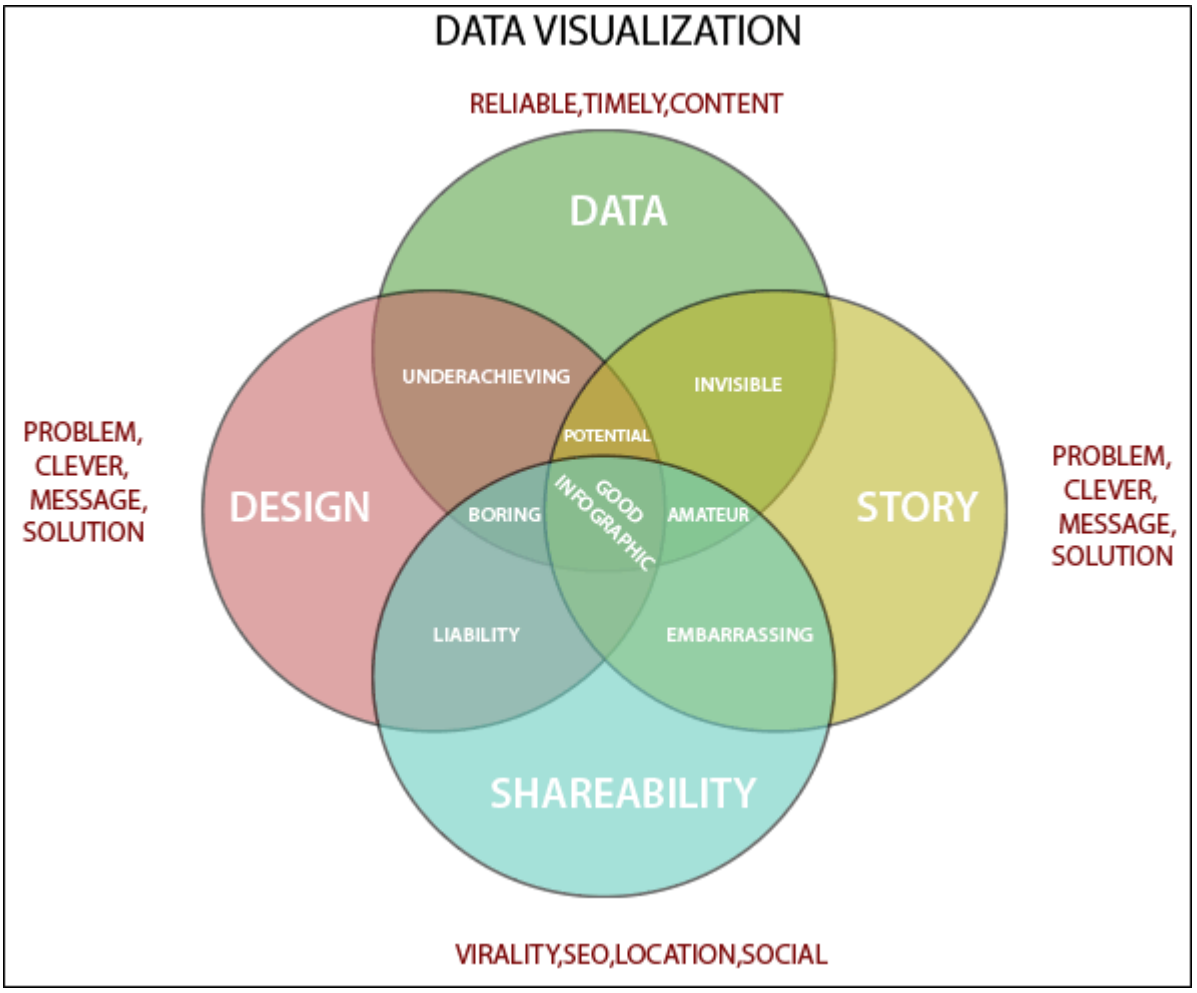
Data visualizations are used to discover unknown facts and trends. You can see visualizations in the form of line charts to display change over time. Bar and column charts are useful for observing relationships and making comparisons. A pie chart is a great way to show parts-of-a-whole. And maps are the best way to share geographical data visually.

Today's data visualization tools go beyond the charts and graphs used in the Microsoft Excel spreadsheet, which displays the data in more sophisticated ways such as dials and gauges, geographic maps, heat maps, pie chart, and fever chart.

What makes Data Visualization Effective?

Effective data visualization are created by communication, data science, and design collide. Data visualizations did right key insights into complicated data sets into meaningful and natural.

American statistician and Yale professor Edward Tufte believe useful data visualizations consist of ?complex ideas communicated with clarity, precision, and efficiency.



Importance of Data Visualization

Data visualization is important because of the processing of information in human brains. Using graphs and charts to visualize a large amount of the complex data sets is more comfortable in comparison to studying the spreadsheet and reports.

Data visualization is an easy and quick way to convey concepts universally. You can experiment with a different outline by making a slight adjustment.

Why Use Data Visualization?

1. To make easier in understand and remember.
2. To discover unknown facts, outliers, and trends.
3. To visualize relationships and patterns quickly.
4. To ask a better question and make better decisions.
5. To competitive analyze.
6. To improve insights.

What is Tableau?

Tableau is a powerful and fastest growing data visualization tool used in the Business Intelligence Industry. It helps in simplifying raw data in a very easily understandable format. Tableau helps create the data that can be understood by professionals at any level in an organization. It also allows non-technical users to create customized dashboards.
Data analysis is very fast with Tableau tool and the visualizations created are in the form of dashboards and worksheets.

The best features of Tableau software are

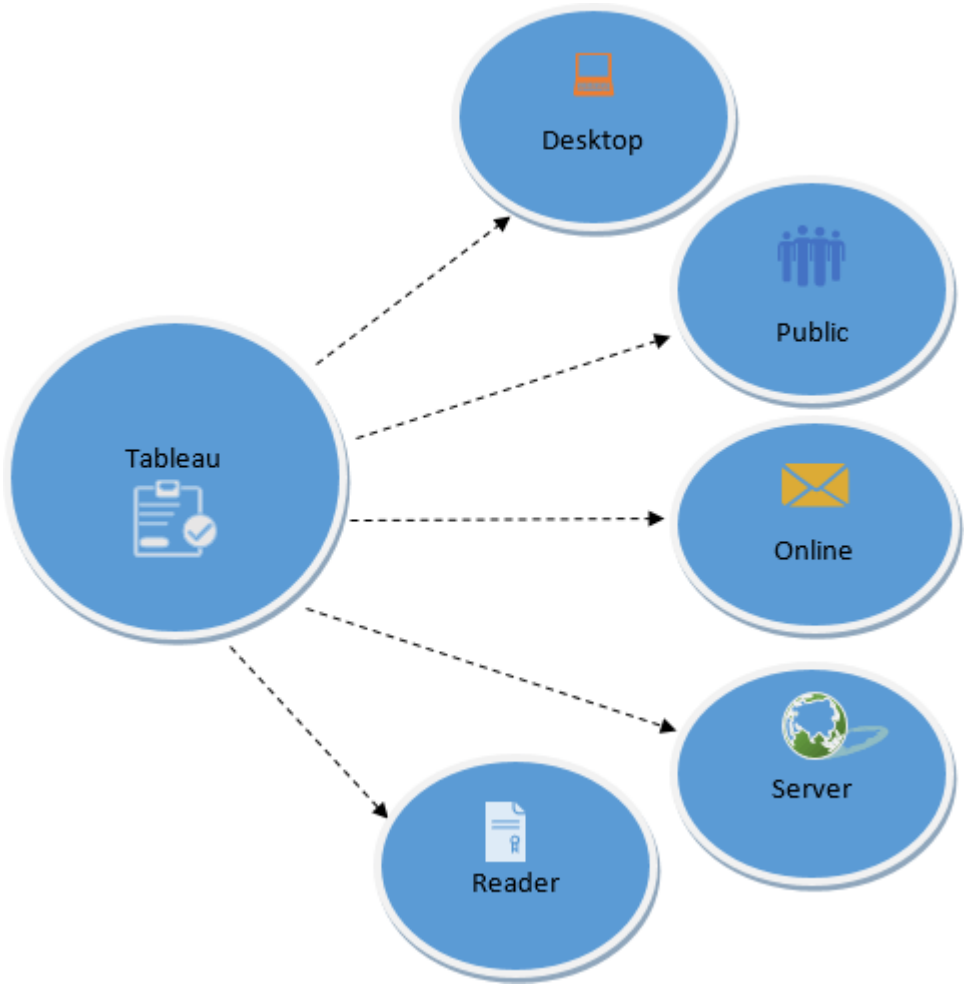
- Data Blending
- Real time analysis
- Collaboration of data

The great thing about Tableau software is that it doesn't require any technical or any kind of programming skills to operate. The tool has garnered interest among the people from all sectors such as business, researchers, different industries, etc.

Tableau Product Suite

The Tableau Product Suite consists of

- Tableau Desktop
- Tableau Public
- Tableau Online
- Tableau Server
- Tableau Reader



For a clear understanding, data analytics in Tableau tool can be classified into two section.

1. Developer Tools: The Tableau tools that are used for development such as the creation of dashboards, charts, report generation, visualization fall into this category. The Tableau products, under this category, are the Tableau Desktop and the Tableau Public.
2. Sharing Tools: As the name suggests, the purpose of these Tableau products is sharing the visualizations, reports, dashboards that were created using the developer tools. Products that fall into this category are Tableau Online, Server, and Reader.

Tableau Desktop

Tableau Desktop has a rich feature set and allows you to code and customize reports. Right from creating the charts, reports, to blending them all together to form a dashboard, all the necessary work is created in Tableau Desktop.

For live data analysis, Tableau Desktop provides connectivity to Data Warehouse, as well as other various types of files. The workbooks and the dashboards created here can be either shared locally or publicly.

Based on the connectivity to the data sources and publishing option, Tableau Desktop is classified into

- Tableau Desktop Personal: The development features are similar to Tableau Desktop. Personal version keeps the workbook private, and the access is limited. The workbooks cannot be published online. Therefore, it should be distributed either Offline or in Tableau Public.
- Tableau Desktop Professional: It is pretty much similar to Tableau Desktop. The difference is that the work created in the Tableau Desktop can be published online or in Tableau Server. Also, in Professional version, there is full access to all sorts of the datatype. It is best suitable for those who wish to publish their work in Tableau Server.

Tableau Public

It is Tableau version specially build for the cost-effective users. By the word "Public," it means that the workbooks created cannot be saved locally; in turn, it should be saved to the Tableau's public cloud which can be viewed and accessed by anyone.

There is no privacy to the files saved to the cloud since anyone can download and access the same. This version is the best for the individuals who want to learn Tableau and for the ones who want to share their data with the general public.

Tableau Server

The software is specifically used to share the workbooks, visualizations that are created in the Tableau Desktop application across the organization. To share dashboards in the Tableau Server, you must first publish your work in the Tableau Desktop. Once the work has been uploaded to the server, it will be accessible only to the licensed users.

However, It's not necessary that the licensed users need to have the Tableau Server installed on their machine. They just require the login credentials with which they can check reports via a web browser. The security is high in Tableau server, and it is much suited for quick and effective sharing of data in an organization.

The admin of the organization will always have full control over the server. The hardware and the software are maintained by the organization.

Tableau Online

As the name suggests, it is an online sharing tool of Tableau. Its functionalities are similar to Tableau Server, but the data is stored on servers hosted in the cloud which are maintained by the Tableau group.

There is no storage limit on the data that can be published in the Tableau Online. Tableau Online creates a direct link to over 40 data sources that are hosted in the cloud such as the MySQL, Hive, Amazon Aurora, Spark SQL and many more.

To publish, both Tableau Online and Server require the workbooks created by Tableau Desktop. Data that is streamed from the web applications say Google Analytics, Salesforce.com are also supported by Tableau Server and Tableau Online.

Tableau Reader

Tableau Reader is a free tool which allows you to view the workbooks and visualizations created using Tableau Desktop or Tableau Public. The data can be filtered but editing and modifications are restricted. The security level is zero in Tableau Reader as anyone who gets the workbook can view it using Tableau Reader.

If you want to share the dashboards that you have created, the receiver should have Tableau Reader to view the document.

How does Tableau work?

Tableau connects and extracts the data stored in various places. It can pull data from any platform imaginable. A simple database such as an excel, pdf, to a complex database like Oracle, a database in the cloud such as Amazon webs services, Microsoft Azure SQL database, Google Cloud SQL and various other data sources can be extracted by Tableau.

When Tableau is launched, ready data connectors are available which allows you to connect to any database. Depending on the version of Tableau that you have purchased the number of data connectors supported by Tableau will vary.

The pulled data can be either connected live or extracted to the Tableau's data engine, Tableau Desktop. This is where the Data analyst, data engineer work with the data that was pulled up and develop visualizations. The created dashboards are shared with the users as a static file. The users who receive the dashboards views the file using Tableau Reader.

The data from the Tableau Desktop can be published to the Tableau server. This is an enterprise platform where collaboration, distribution, governance, security model, automation features are supported. With the Tableau server, the end users have a better experience in accessing the files from all locations be it a desktop, mobile or email.

Tableau Uses

Following are the main uses and applications of Tableau:

- Business Intelligence
- Data Visualization
- Data Collaboration
- Data Blending
- Real-time data analysis
- Query translation into visualization
- To import large size of data
- To create no-code data queries
- To manage large size metadata

What Is Machine Learning?

Machine learning is the study of using algorithms and data that allow computers to perform tasks without instructions or input from human users. Different experts have created their own definitions to describe machine learning, but at its core, machine learning is characterized by

computers performing autonomous improvement using real-world examples and data to do so, rather than a continual human input.

AI vs Machine Learning

At first glance, machine learning seems to have an almost interchangeable definition with "artificial intelligence" (AI). After all, Merriam-Webster defines AI as "a branch of computer science dealing with the simulation of intelligent behavior in computers; the capability of a machine to imitate intelligent human behavior." However, upon closer inspection, it's clear that these two terms refer to entirely distinct things.

AI encompasses many different processes and practices, including things like neural networks and image processing; machine learning is one of these subsets of AI. So while AI can take many different forms, such as a self-driving car or a digital assistant like Siri or Alexa, machine learning describes a particular aspect of AI function: computers learning autonomously.

How Do Computers Learn?

Put simply, a human user puts data into the computer, which then analyzes the data and looks for patterns in it. When the computer finds a pattern, it adjusts how it processes or manages data to reflect what it found. After the computer finds enough patterns, it can begin to make predictions. Generally, if a larger amount of training data is put in, the computer will become more accurate, faster.

In practice, machine learning is more complicated than that and there are two main forms: supervised and unsupervised learning. Each form requires large amounts of input data to train the machine learning algorithm, but they differ in how they interact with the data.

- Supervised learning: For this form of machine learning, the training data is categorized or labeled with the "correct" outcome. The computer is then given unlabeled

information to process. It will compare the new data with the old and then determine the outcome based on the previous example. Supervised learning is especially well-suited to classifying items into different categories and regression when the output is a real value such as “dollars” or “pounds.” This is the most common form of machine learning and it’s generally more reliable than unsupervised learning.

- Unsupervised learning: In unsupervised learning, the data is not labeled before being put into the algorithm. The machine then attempts to find patterns in the data on its own. Unsupervised learning works particularly well when identifying similarities in groups and clustering them together, as well as identifying anomalies or abnormalities in data. It’s more difficult to measure the accuracy of an unsupervised learning algorithm since there are no training data to compare it to, but it can still provide valuable results and insights.

How Is Machine Learning Used?

Machine learning has a variety of applications in modern life. We’ve already found uses for it in industries ranging from healthcare to cybersecurity, and as this technology continues to develop, we’ll likely find many more. Other common uses of machine learning include:

Machine Learning in Marketing

Machine learning may prove to be especially useful in marketing because of its ability to identify patterns in data that humans might not otherwise notice. This can be particularly helpful when looking at user behavior; machine learning algorithms can analyze massive amounts of user data from multiple sources, such as social media pages and interactions with a website, to better determine how marketers and brands can engage with customers.

Among its many uses in marketing, machine learning can help marketers decide what ads to display to certain customers, identify the best time to send out individual offers or incentives, and even help improve the overall customer experience.

Search Engines

Search engines typically use algorithms to organize relevant results when users input a query. Machine learning is often used to update and refine these algorithms to improve the quality of results given to users. It can also be used to better understand searchers' queries, classify users to make searches more personalized, and determine the best rate for crawling different websites or data sets.

Machine Learning in Weather Prediction and Climate Science

Machine learning also has uses in meteorology and climatology, as it can be used to analyze and predict weather patterns. Websites and apps that use APIs to scrape weather data are useful for consumers, but that kind of technology also empowers weather prediction by autonomous computer systems, providing them with ever more information. Further, machine learning can also play a role in obtaining and analyzing larger climate patterns, which can aid in the development of more accurate and detailed climate prediction models.

Machine Vision: Recognizing Text, Images, and Faces

A specialized type of machine learning, machine or computer vision is a computer's ability to "see," inspect and analyze images or videos. By analyzing images and converting visual elements into data, machine vision can recognize text in an image, identify faces, and even improve or generate images. High-powered computers aren't the only devices capable of this kind of machine learning; using the right kind of API with optical character recognition, you can even turn your cell phone into an image reader, pull text out of physical images, even

perform live translation of written text. Machine vision is being used in a variety of industries for a number of purposes, including social media and law enforcement.

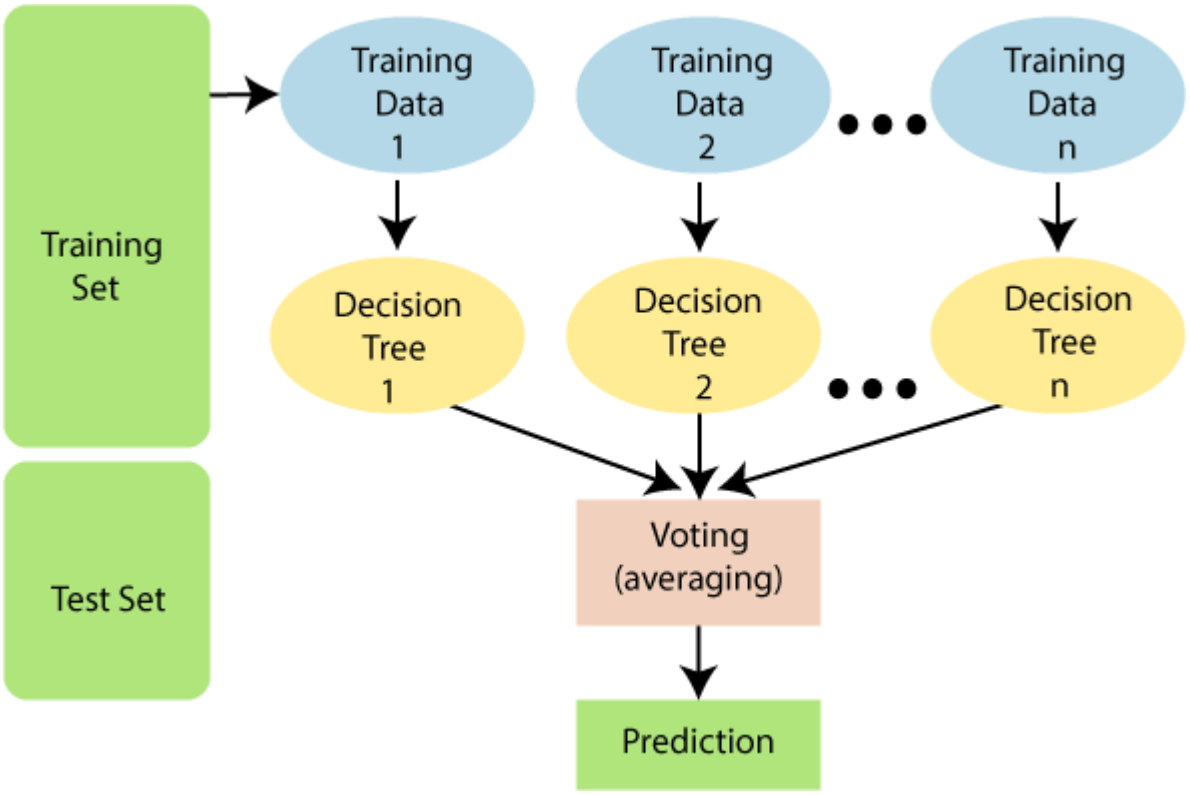
Machine learning is an exciting new field in the world of AI, and though we've made great strides in developing this technology, it has many applications that have yet to be explored. As computer scientists continue to refine the capabilities of machine learning, we'll determine even more ways in which it can be used, and it may become even more important to daily life than it already is.

Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of COMBINING MULTIPLE CLASSIFIERS TO SOLVE A COMPLEX PROBLEM AND TO IMPROVE THE PERFORMANCE OF THE MODEL.

As the name suggests, *"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."* Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.



Assumptions for Random Forest

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

Why use Random Forest?

Below are some points that explain why we should use the Random Forest algorithm:

<="" li="">

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

How does Random Forest algorithm work?

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

The working of the algorithm can be better understood by the below example:

Example: Suppose there is a dataset that contains multiple fruit images. So, this dataset is given to the Random forest classifier. The dataset is divided into subsets and given to each decision tree. During the training phase, each decision tree produces a prediction result, and when a new data point occurs, then based on the majority of results, the Random Forest classifier predicts the final decision.

Applications of Random Forest

There are mainly four sectors where Random forest mostly used:

1. Banking: Banking sector mostly uses this algorithm for the identification of loan risk.
2. Medicine: With the help of this algorithm, disease trends and risks of the disease can be identified.
3. Land Use: We can identify the areas of similar land use by this algorithm.
4. Marketing: Marketing trends can be identified using this algorithm.

Advantages of Random Forest

- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the overfitting issue.

Disadvantages of Random Forest

- Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

Python Implementation of Random Forest Algorithm

Now we will implement the Random Forest Algorithm tree using Python. For this, we will use the same dataset "user_data.csv", which we have used in previous classification models. By using the same dataset, we can compare the Random Forest classifier with other classification models such as Decision tree Classifier,

KNN,

SVM,

Logistic Regression,

etc.

Implementation Steps are given below:

- Data Pre-processing step
- Fitting the Random forest algorithm to the Training set
- Predicting the test result
- Test accuracy of the result (Creation of Confusion matrix)
- Visualizing the test set result.

SPYDER

It is always necessary to have interactive environments to create software applications and this fact becomes very important when you work in the fields of Data Science, engineering, and scientific research. The Python Spyder IDE has been created for the same purpose. In this article, you will be learning how to install and make use of Spyder or the Scientific Python and Development IDE.

Before moving on, let’s take a look at all the topics that are discussed over here:

- What is Python Spyder IDE?
- Features of Spyder
- Python Spyder IDE Installation
- Creating a file/ Starting a Project

- Writing the Code
- Variable Explorer
- File Explorer
- Configuring Spyder
- Help

What is Python Spyder IDE?

Spyder is an open-source cross-platform IDE. The Python Spyder IDE is written completely in Python. It is designed by scientists and is exclusively for scientists, data analysts, and engineers. It is also known as the Scientific Python Development IDE and has a huge set of remarkable features which are discussed below.

Features of Spyder

Some of the remarkable features of Spyder are:

- Customizable Syntax Highlighting
- Availability of breakpoints (debugging and conditional breakpoints)
- Interactive execution which allows you to run line, file, cell, etc.
- Run configurations for working directory selections, command-line options, current/ dedicated/ external console, etc
- Can clear variables automatically (or enter debugging)
- Navigation through cells, functions, blocks, etc can be achieved through the Outline Explorer

- It provides real-time code introspection (The ability to examine what functions, keywords, and classes are, what they are doing and what information they contain)
- Automatic colon insertion after if, while, etc
- Supports all the IPython magic commands
- Inline display for graphics produced using Matplotlib
- Also provides features such as help, file explorer, find files, etc.

About Python

Python is one of those rare languages which can claim to be
both SIMPLE and POWERFUL. You will find yourself pleasantly surprised to see how easy it is to concentrate on the solution to the problem rather than the syntax and structure of the language you are programming in.

The official introduction to Python is:

Python is an easy to learn, powerful programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python's elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms.

I will discuss most of these features in more detail in the next section.

Story behind the name

Guido van Rossum, the creator of the Python language, named the language after the BBC show "Monty Python's Flying Circus". He doesn't particularly like snakes that kill animals for food by winding their long bodies around them and crushing them.

Features of Python

Simple

Python is a simple and minimalistic language. Reading a good Python program feels almost like reading English, although very strict English! This pseudo-code nature of Python is one of its greatest strengths. It allows you to concentrate on the solution to the problem rather than the language itself.

Easy to Learn

As you will see, Python is extremely easy to get started with. Python has an extraordinarily simple syntax, as already mentioned.

Free and Open Source

Python is an example of a FLOSS (Free/Libre and Open Source Software). In simple terms, you can freely distribute copies of this software, read its source code, make changes to it, and use pieces of it in new free programs. FLOSS is based on the concept of a community which shares knowledge. This is one of the reasons why Python is so good - it has been created and is constantly improved by a community who just want to see a better Python.

High-level Language

When you write programs in Python, you never need to bother about the low-level details such as managing the memory used by your program, etc.

Portable

Due to its open-source nature, Python has been ported to (i.e. changed to make it work on) many platforms. All your Python programs can work on any of these platforms without requiring any changes at all if you are careful enough to avoid any system-dependent features.

You can use Python on GNU/Linux, Windows, FreeBSD, Macintosh, Solaris, OS/2, Amiga, AROS, AS/400, BeOS, OS/390, z/OS, Palm OS, QNX, VMS, Psion, Acorn RISC OS, VxWorks, PlayStation, Sharp Zaurus, Windows CE and PocketPC!

You can even use a platform like Kivy to create games for your computer AND for iPhone, iPad, and Android.

Interpreted

This requires a bit of explanation.

A program written in a compiled language like C or C++ is converted from the source language i.e. C or C++ into a language that is spoken by your computer (binary code i.e. 0s and 1s) using a compiler with various flags and options. When you run the program, the linker/loader software copies the program from hard disk to memory and starts running it.

Python, on the other hand, does not need compilation to binary. You just RUN the program directly from the source code. Internally, Python converts the source code into an intermediate form called bytecodes and then translates this into the native language of your computer and then runs it. All this, actually, makes using Python much easier since you don't have to worry about compiling the program, making sure that the proper libraries are linked and loaded, etc. This also makes your Python programs much more portable, since you can just copy your Python program onto another computer and it just works!

Object Oriented

Python supports procedure-oriented programming as well as object-oriented programming (OOP). In PROCEDURE-ORIENTED languages, the program is built around procedures or functions which are nothing but reusable pieces of programs. In OBJECT-ORIENTED languages, the program is built around objects which combine data and functionality. Python has a very powerful but simplistic way of doing OOP, especially when compared to big languages like C++ or Java.

Extensible

If you need a critical piece of code to run very fast or want to have some piece of algorithm not to be open, you can code that part of your program in C or C++ and then use it from your Python program.

Embeddable

You can embed Python within your C/C++ programs to give SCRIPTING capabilities for your program's users.

Extensive Libraries

The Python Standard Library is huge indeed. It can help you do various things involving regular expressions,documentation generation, unit testing, threading, databases, web browsers, CGI, FTP, email, XML, XML-RPC, HTML, WAV files, cryptography, GUI (graphical user interfaces), and other system-dependent stuff. Remember, all this is always available wherever Python is installed. This is called the BATTERIES INCLUDED philosophy of Python.

Besides the standard library, there are various other high-quality libraries which you can find at the Python Package Index.

Broadly, there are 3 types of Machine Learning Algorithms

1. Supervised Learning

How it works: This algorithm consist 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 etc.

2. Unsupervised Learning

How it works: 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.

3. Reinforcement Learning:

How it works: 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. Example of Reinforcement Learning: Markov Decision Process

List of Common Machine Learning Algorithms

Here is the list of commonly used machine learning algorithms. These algorithms can be applied to almost any data problem:

1. Linear Regression
2. Logistic Regression
3. Decision Tree
4. SVM
5. Naive Bayes
6. kNN
7. K-Means
8. Random Forest
9. Dimensionality Reduction Algorithms
10. Gradient Boosting algorithms
    1. GBM
    2. XGBoost
    3. LightGBM
    4. CatBoost

1. Linear Regression

It is used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s). Here, we establish relationship between independent and dependent variables by fitting a best line. This best fit line is known as regression line and represented by a linear equation Y= a *X + b.

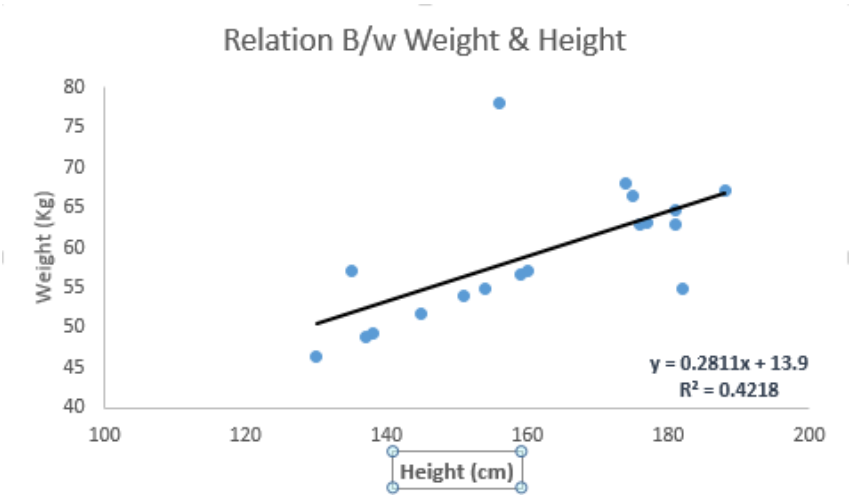
The best way to understand linear regression is to relive this experience of childhood. Let us say, you ask a child in fifth grade to arrange people in his class by increasing order of weight, without asking them their weights! What do you think the child will do? He / she would likely look (visually analyze) at the height and build of people and arrange them using a combination of these visible parameters. This is linear regression in real life! The child has actually figured out that height and build would be correlated to the weight by a relationship, which looks like the equation above.

In this equation:

- Y – Dependent Variable
- a – Slope
- X – Independent variable
- b – Intercept

These coefficients a and b are derived based on minimizing the sum of squared difference of distance between data points and regression line.

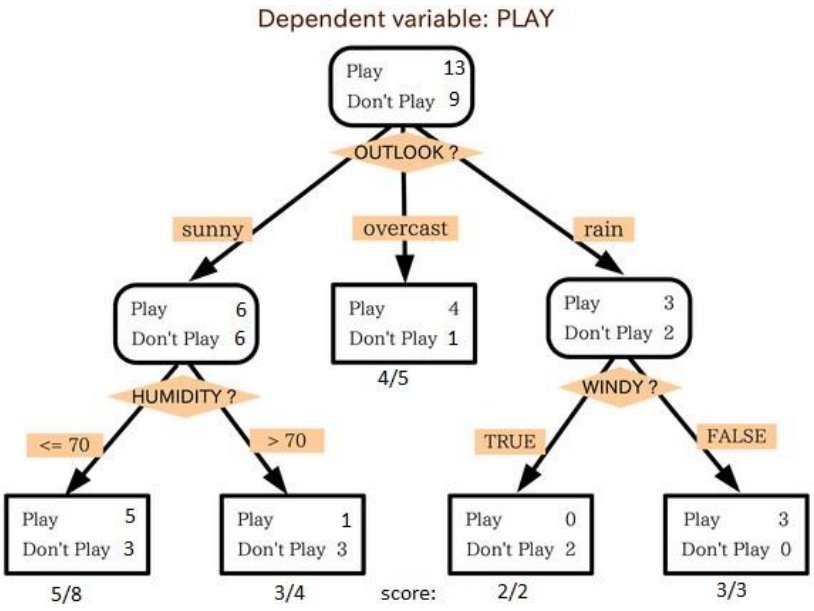
Look at the below example. Here we have identified the best fit line having linear equation y=0.2811x+13.9. Now using this equation, we can find the weight, knowing the height of a person.



Linear Regression is mainly of two types: Simple Linear Regression and Multiple Linear Regression. Simple Linear Regression is characterized by one independent variable. And, Multiple Linear Regression(as the name suggests) is characterized by multiple (more than 1) independent variables. While finding the best fit line, you can fit a polynomial or curvilinear regression. And these are known as polynomial or curvilinear regression.

3. Decision Tree

This is one of my favorite algorithm and I use it quite frequently. It is a type of supervised learning algorithm that is mostly used for classification problems. Surprisingly, it works for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets. This is done based on most significant attributes/ independent variables to make as distinct groups as possible. For more details, you can read: Decision Tree Simplified.

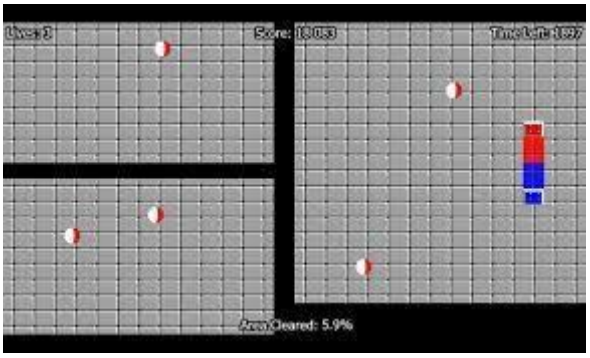


source: statsexchange

In the image above, you can see that population is classified into four different groups based on multiple attributes to identify 'if they will play or not'. To split the population into different

heterogeneous groups, it uses various techniques like Gini, Information Gain, Chi-square, entropy.

The best way to understand how decision tree works, is to play Jezzball – a classic game from Microsoft (image below). Essentially, you have a room with moving walls and you need to create walls such that maximum area gets cleared off with out the balls.

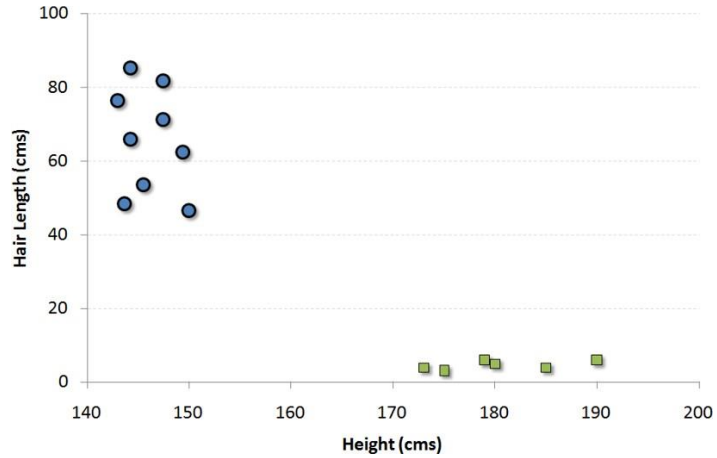


So, every time you split the room with a wall, you are trying to create 2 different populations with in the same room. Decision trees work in very similar fashion by dividing a population in as different groups as possible.

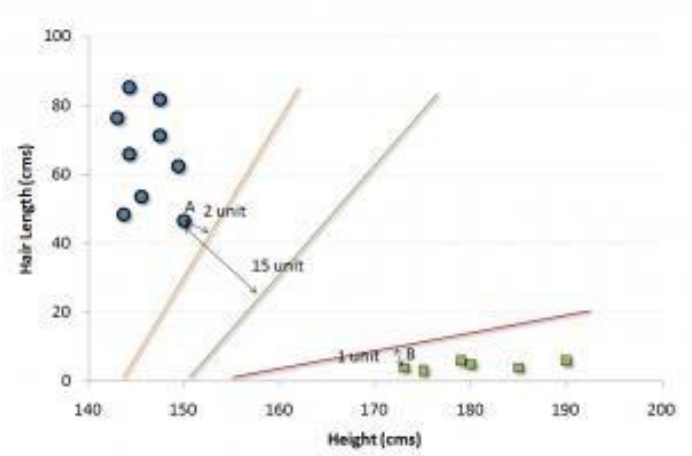
4. SVM (Support Vector Machine)

It is a classification method. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate.

For example, if we only had two features like Height and Hair length of an individual, we'd first plot these two variables in two dimensional space where each point has two co-ordinates (these co-ordinates are known as Support Vectors)



Now, we will find some LINE that splits the data between the two differently classified groups of data. This will be the line such that the distances from the closest point in each of the two groups will be farthest away.



In the example shown above, the line which splits the data into two differently classified groups is the BLACK line, since the two closest points are the farthest apart from the line. This line is our classifier. Then, depending on where the testing data lands on either side of the line, that's what class we can classify the new data as.

More: Simplified Version of Support Vector Machine

Think of this algorithm as playing JezzBall in n-dimensional space. The tweaks in the game are:

- You can draw lines/planes at any angles (rather than just horizontal or vertical as in the classic game)
- The objective of the game is to segregate balls of different colors in different rooms.
- And the balls are not moving.

Logistic Regression

Logistic regression works with sigmoid function because the sigmoid function can be used to classify the output that is dependent feature and it uses the probability for classification of the dependent feature. This algorithm works well with less amount of data set because of the use of sigmoid function if value the of sigmoid function is greater than 0.5 the output will 1 if the output the sigmoid function is less than 0.5 then the output is considered as the 0. But this sigmoid function is not suitable for deep learning because the if deep learning when we back tracking from the output to input we have to update the weights to minimize the error in weight update. we have to do differentiation of sigmoid activation function in middle layer neuron then results in the value of 0.25 this will affect the accuracy of the module in deep learning.

IMPLEMENTATION

The Data

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.

CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING

By Garvit Kashyap & Ujjwal Raj

IMPORTING THE DEPENDENCIES

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

LOADING THE DATASET TO A PANDAS DATA FRAME

credit_card_data = pd.read_csv('creditcard.csv')

FIRST FIVE ROWS OF THE DATASET

```
In [3]: # first 5 rows of the dataset
        credit_card_data.head()
```

Out[3]:

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	V2
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.12853
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.339846	0.16717
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412	-0.689281	-0.32764
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321	-1.175575	0.64737
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458	0.141267	-0.20601

LAST FIVE ROWS OF THE DATASET

```
In [4]: credit_card_data.tail()
```

Out[4]:

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24
284802	172786.0	-11.881118	10.071785	-9.834783	-2.066656	-5.364473	-2.606837	-4.918215	7.305334	1.914428	...	0.213454	0.111864	1.014480	-0.509348
284803	172787.0	-0.732789	-0.055080	2.035030	-0.738589	0.868229	1.058415	0.024330	0.294869	0.584800	...	0.214205	0.924384	0.012463	-1.016226
284804	172788.0	1.919565	-0.301254	-3.249640	-0.557828	2.630515	3.031260	-0.296827	0.708417	0.432454	...	0.232045	0.578229	-0.037501	0.640134
284805	172788.0	-0.240440	0.530483	0.702510	0.689799	-0.377961	0.623708	-0.686180	0.679145	0.392087	...	0.265245	0.800049	-0.163298	0.123205
284806	172792.0	-0.533413	-0.189733	0.703337	-0.506271	-0.012546	-0.649617	1.577006	-0.414650	0.486180	...	0.261057	0.643078	0.376777	0.008797

5 rows × 31 columns

DATASET INFORMATION

```
credit_card_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 284807 entries, 0 to 284806
Data columns (total 31 columns):
 #   Column  Non-Null Count   Dtype  
---  ------  --------------   -----  
 0   Time    284807 non-null  float64
 1   V1      284807 non-null  float64
 2   V2      284807 non-null  float64
 3   V3      284807 non-null  float64
 4   V4      284807 non-null  float64
 5   V5      284807 non-null  float64
 6   V6      284807 non-null  float64
 7   V7      284807 non-null  float64
 8   V8      284807 non-null  float64
 9   V9      284807 non-null  float64
 10  V10     284807 non-null  float64
 11  V11     284807 non-null  float64
 12  V12     284807 non-null  float64
 13  V13     284807 non-null  float64
 14  V14     284807 non-null  float64
 15  V15     284807 non-null  float64
 16  V16     284807 non-null  float64
 17  V17     284807 non-null  float64
 18  V18     284807 non-null  float64
 19  V19     284807 non-null  float64
 20  V20     284807 non-null  float64
 21  V21     284807 non-null  float64
 22  V22     284807 non-null  float64
 23  V23     284807 non-null  float64
 24  V24     284807 non-null  float64
 25  V25     284807 non-null  float64
 26  V26     284807 non-null  float64
 27  V27     284807 non-null  float64
 28  V28     284807 non-null  float64
 29  Amount  284807 non-null  float64
 30  Class   284807 non-null  int64  
dtypes: float64(30), int64(1)
memory usage: 67.4 MB
```

Pre-processing the data

Label Encoding. In the dataset, there are 13 predictors. 2 of them are numerical variables while rest of them are categorical. In order to apply machine learning models, we need

numeric representation of the features. Therefore, all non-numeric features were transformed into numerical form.

Train the data. In this process, 20% of the data was split for the test data and 80% of the data was taken as train data.

Scaling the Data. While exploring the data in the previous sections, it was seen that the data is not normally distributed. Without scaling, the machine learning models will try to disregard coefficients of features that has low values because their impact will be so small compared to the big value features.

CHECHKING THE NUMBER OF MISSING VALUES IN EACH COLUMN

credit_card_data.isnull().sum()

```
Time      0
V1        0
V2        0
V3        0
V4        0
V5        0
V6        0
V7        0
V8        0
V9        0
V10       0
V11       0
V12       0
V13       0
V14       0
V15       0
V16       0
V17       0
V18       0
V19       0
V20       0
V21       0
V22       0
V23       0
V24       0
V25       0
V26       0
V27       0
V28       0
Amount    0
Class     0
dtype: int64
```

DISTRIBUTION OF LEGIT AND FRAUD TRANSACTION

credit_card_data['Class'].value_counts()

```
Out[7]: 0    284315
        1       492
        Name: Class, dtype: int64
```

This Dataset is highly unblanced

0 --> Normal Transaction

1 --> fraudulent transaction

SEPRATING THE DATA FOR ANALYSIS

```
legit = credit_card_data[credit_card_data.Class == 0]
fraud = credit_card_data[credit_card_data.Class == 1]
```

```
In [9]: print(legit.shape)
        print(fraud.shape)
```

```
(284315, 31)
(492, 31)
```

STATISTICAL MEASURE OF THE DATA

legit.Amount.describe()

```
Out[10]: count    284315.000000
         mean         88.291022
         std         250.105092
         min           0.000000
         25%           5.650000
         50%          22.000000
         75%          77.050000
         max       25691.160000
         Name: Amount, dtype: float64
```

```
In [11]: fraud.Amount.describe()
```

```
Out[11]: count     492.000000
         mean      122.211321
         std       256.683288
         min         0.000000
         25%         1.000000
         50%         9.250000
         75%       105.890000
         max      2125.870000
         Name: Amount, dtype: float64
```

COMPARING THE VALUES FOR BOTH TRANSACTION

```
credit_card_data.groupby('Class').mean()
```

Out[12]:

Class	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V20	V21	V22	V
0	94838.202258	0.008258	-0.006271	0.012171	-0.007860	0.005453	0.002419	0.009637	-0.000987	0.004467	...	-0.000644	-0.001235	-0.000024	0.000
1	80746.806911	-4.771948	3.623778	-7.033281	4.542029	-3.151225	-1.397737	-5.568731	0.570636	-2.581123	...	0.372319	0.713588	0.014049	-0.040

2 rows × 30 columns

Under-Sampling

Build a sample dataset containing similar distribution of normal transactions and Fraudulent Transactions

Number of Fraudulent Transactions --> 492

In [13]:
```
legit_sample = legit.sample(n=492)
```

Concatenating two DataFrames

In [14]:
```
new_dataset = pd.concat([legit_sample, fraud], axis=0)
```

In [15]:
```
new_dataset.head()
```

Out[15]:

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24
67470	52549.0	1.138856	0.176357	0.393964	1.402253	-0.146174	-0.171439	0.096815	-0.027635	0.265254	...	-0.091823	-0.050500	-0.075740	0.099906
221343	142530.0	-0.100242	0.698772	0.290204	-0.862570	0.757746	-0.496475	1.071998	-0.237056	0.036956	...	-0.284810	-0.605977	-0.044413	-0.522943
252113	155657.0	-0.009898	1.036865	-0.462605	-0.504273	0.831856	-0.806814	1.041227	-0.246557	-0.031422	...	-0.265514	-0.829885	0.105388	0.622603
163196	115736.0	2.123232	0.124924	-2.201047	-0.004618	1.155875	0.188879	0.101285	-0.046693	0.217869	...	-0.389081	-1.010386	0.162083	-0.774889
39013	39653.0	1.127748	-2.072190	1.126754	-0.850534	-2.292021	0.616748	-1.711136	0.346555	-0.484379	...	-0.367821	-0.622492	-0.098361	-0.017575

5 rows × 31 columns

In [16]:
```
new_dataset.tail()
```

Out[16]:

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24
279863	169142.0	-1.927883	1.125653	-4.518331	1.749293	-1.566487	-2.010494	-0.882850	0.697211	-2.064945	...	0.778584	-0.319189	0.639419	-0.294885
280143	169347.0	1.378559	1.289381	-5.004247	1.411850	0.442581	-1.326536	-1.413170	0.248525	-1.127396	...	0.370612	0.028234	-0.145640	-0.081049
280149	169351.0	-0.676143	1.126366	-2.213700	0.468308	-1.120541	-0.003346	-2.234739	1.210158	-0.652250	...	0.751826	0.834108	0.190944	0.032070
281144	169966.0	-3.113832	0.585864	-5.399730	1.817092	-0.840618	-2.943548	-2.208002	1.058733	-1.632333	...	0.583276	-0.269209	-0.456108	-0.183659
281674	170348.0	1.991976	0.158476	-2.583441	0.408670	1.151147	-0.096695	0.223050	-0.068384	0.577829	...	-0.164350	-0.295135	-0.072173	-0.450261

```
In [17]: new_dataset['Class'].value_counts()
```

```
Out[17]: 0    492
         1    492
         Name: Class, dtype: int64
```

```
In [18]: new_dataset.groupby('Class').mean()
```

Out[18]:

Class	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V20	V21	V22	V2
0	90742.863821	-0.071344	0.106861	-0.002285	-0.018632	0.055807	-0.086126	0.024539	0.027101	0.005133	...	-0.012815	-0.015038	-0.071839	0.00938
1	80746.806911	-4.771948	3.623778	-7.033281	4.542029	-3.151225	-1.397737	-5.568731	0.570636	-2.581123	...	0.372319	0.713588	0.014049	-0.0403(

2 rows × 30 columns

SPLITTING THE DATA INTO FEATURES & TARGETS

```
X = new_dataset.drop(columns='Class', axis=1)
Y = new_dataset['Class']
```

```
[ ] print(X)
```

```
             Time        V1        V2  ...       V27       V28  Amount
203131  134666.0 -1.220220 -1.729458  ...  0.173995 -0.023852  155.00
95383    65279.0 -1.295124  0.157326  ...  0.317321  0.105345   70.00
99706    67246.0 -1.481168  1.226490  ... -0.546577  0.076538   40.14
153895  100541.0 -0.181013  1.395877  ... -0.229857 -0.329608  137.04
249976  154664.0  0.475977 -0.573662  ...  0.058961  0.012816   19.60
...          ...       ...       ...  ...       ...       ...     ...
279863  169142.0 -1.927883  1.125653  ...  0.292680  0.147968  390.00
280143  169347.0  1.378559  1.289381  ...  0.389152  0.186637    0.76
280149  169351.0 -0.676143  1.126366  ...  0.385107  0.194361   77.89
281144  169966.0 -3.113832  0.585864  ...  0.884876 -0.253700  245.00
281674  170348.0  1.991976  0.158476  ...  0.002988 -0.015309   42.53

[984 rows x 30 columns]
```

```
[ ] print(Y)
```

```
203131    0
95383     0
99706     0
153895    0
249976    0
         ..
279863    1
280143    1
280149    1
281144    1
281674    1
Name: Class, Length: 984, dtype: int64
```

SPLITTING THE DATA INTO TRAINING AND TESTING DATA

```
In [22]: X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, stratify=Y, random_state=2)
```

```
In [23]: print(X.shape, X_train.shape, X_test.shape)
```

```
(984, 30) (787, 30) (197, 30)
```

MODEL TRAINING:

LOGISTIC REGRESSION ALGORITHM

```
In [24]: model = LogisticRegression()

In [25]: # training the Logistic Regression Model with Training Data
         model.fit(X_train, Y_train)

Out[25]: LogisticRegression()

         Model Evaluation

         Accuracy Score

In [26]: # accuracy on training data
         X_train_prediction = model.predict(X_train)
         training_data_accuracy = accuracy_score(X_train_prediction, Y_train)

In [27]: print('Accuracy on Training data : ', training_data_accuracy)

         Accuracy on Training data :  0.9351969504447268

In [28]: # accuracy on test data
         X_test_prediction = model.predict(X_test)
         test_data_accuracy = accuracy_score(X_test_prediction, Y_test)

In [29]: print('Accuracy score on Test Data : ', test_data_accuracy)
```

```
Accuracy score on Test Data :  0.9238578680203046
```

CONCLUSION

Credit card fraud is without a doubt an act of criminal dishonesty. This article has listed out the most common methods of fraud along with their detection methods and reviewed recent findings in this field. This paper has also explained in detail, how machine learning can be applied to get better results in fraud detection along with the algorithm, pseudocode, explanation its implementation and experimentation results .While the algorithm does reach over 99.6% accuracy, its precision remains only at 28% when a tenth of the data set is taken into consideration. However, when the entire dataset is fed into the algorithm, the precision rises to 33%. This high percentage of accuracy is to be expected due to the huge imbalance between the number of valid and number of genuine transactions.

Since the entire dataset consists of only two days' transaction records, its only a fraction of data that can be made available if this project were to be used on a commercial scale. Being based on machine learning algorithms, the program will only increase its efficiency over time as more data is put into it.

REFERENCES

[1] “Credit Card Fraud Detection Based on Transaction Behaviour -by John Richard D. Kho, Larry A. Vea” published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017

[2] CLIFTON PHUA1, VINCENT LEE1, KATE SMITH1 & ROSS GAYLER2 “ A Comprehensive Survey of Data Mining-based Fraud Detection Research” published by School of Business Systems, Faculty of Information Technology, Monash University, Wellington Road, Clayton, Victoria 3800, Australia

[3] “Survey Paper on Credit Card Fraud Detection by Suman” , Research Scholar, GJUS&T Hisar HCE, Sonepat published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 3, March 2014

[4] “Research on Credit Card Fraud Detection Model Based on Distance Sum – by Wen-Fang YU and Na Wang” published by 2009 International Joint Conference on Artificial Intelligence

[5] “Credit Card Fraud Detection through Parenclitic Network AnalysisBy Massimiliano Zanin, Miguel Romance, Regino Criado, and SantiagoMoral” published by Hindawi Complexity Volume 2018, Article ID 5764370, 9 pages

[6] “Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy” published by IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 29, NO. 8, AUGUST 2018

[7] “Credit Card Fraud Detection-by Ishu Trivedi, Monika, Mrigya, Mridushi” published by International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 1, January 2016

[8] David J.Wetson,David J.Hand,M Adams,Whitrow and Piotr Jusczak “Plastic Card Fraud Detection using Peer Group Analysis” Springer, Issue 2008.