**A PROJECT REPORT ON**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Title of the Project\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**Submitted to \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* University for the partial fulfillment of the Requirement for the**

**Award of Degree for**

**\*\*\*\*\*\*\*\*\*\*\*Course Name\*\*\*\*\*\*\*\*\*\*\*\*\***

**Done By**

**Mr. /Miss \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*Institute of Management and Computer Science**

**CERTIFICATE**

**This is to certify that Mr., /Miss \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* bearing Roll**

**No. \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* have developed software project**

**Titled \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* For \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Software**

**Solutions as Partial fulfillment for the award of the Degree of**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**Head of the Department Principal**

**\*\*\*\*\*\*College Name\*\*\*\*\*\***

**External**

**ACKNOWLEDGEMENT**

At every outset I express my gratitude to almighty lord for showering his grace and blessings upon me to complete this project.

Although our name appears on the cover of this book, many people had contributed in some form or the other form to this project Development. We could not done this project without the assistance or support of each of the following we thank you all.

I wish to place on my record my deep sense of gratitude to my project guide, **Mr. \*\*\*\*\*\*, \*\*\*\*\* Software Solutions,** for his constant motivation and valuable help through the project work. Express my gratitude to **Mr. \*\*\*\*\*\***, Director of \*\*\*\*\*\*\* **Institute of Management & Computer Sciences** for his valuable suggestions and advices throughout the \*\*\*\*\* course. I also extend my thanks to other Faculties for their Cooperation during my Course.

Finally I would like to thank my friends for their cooperation to complete this project.

\*\*\*\*\*\*\*Your Name\*\*\*\*\*\*\*\*\*

**-:Abstract:-**

Billions of dollars of loss are caused every year by fraudulent credit card transactions. The design of efficient fraud detection algorithms is key for reducing these losses, and more and more algorithms rely on advanced machine learning techniques to assist fraud investigators. The design of fraud detection algorithms is however particularly challenging due to the non-stationary distribution of the data, the highly unbalanced classes distributions and the availability of few transactions labeled by fraud investigators. At the same time public data are scarcely available for confidentiality issues, leaving unanswered many questions about what is the best strategy. In this thesis we aim to provide some answers by focusing on crucial issues such as: i) why and how under sampling is useful in the presence of class imbalance (i.e. frauds are a small percentage of the transactions), ii) how to deal with unbalanced and evolving data streams (non-stationarity due to fraud evolution and change of spending behavior), iii) how to assess performances in a way which is relevant for detection and iv) how to use feedbacks provided by investigators on the fraud alerts generated. Finally, we design and assess a prototype of a Fraud Detection System able to meet real-world working conditions and that is able to integrate investigators’ feedback to generate accurate alerts.

**Index Terms**— Hidden Markov Model, credit card, fraud detection, online shopping, e-commerce (HMM)

**Index**

1. **INTRODUCTION**
   1. INTRODUCTION TO PROJECT
   2. ORGANIZATION PROFILE
   3. PURPOSE OF THE SYSTEM
   4. PROBLEMS IN EXISTING SYSTEM
2. **SYSTEM ANALYSIS**
   1. INTRODUCTION
   2. STUDY OF THE SYSTEM
   3. HARDWARE AND SOFTWARE REQUIRMENTS
   4. PROPOSED SYSTEM
   5. INPUT AND OUTPUT
   6. PROCESS MODULES USED WITH JUSTIFICATION
3. **FEASIBILITY REPORT**
   1. TECHNICAL FEASIBILITY
   2. OPERATIONAL FEASIBILITY
   3. ECONOMICAL FEASIBILTY
4. **SOFTWARE REQUIREMENT SPECIFICATIONS**
   1. FUNCTIONAL REQUIREMENTS
   2. PERFORMANCE REQUIREMENTS

**5. SELECTED SOFTWARE**

5.1INTRODUCTION TO PYTHON

5.2PYTHON

**6. SYSTEM DESIGN**

6.1INTRODUCTION

6.2SYSTEM WORK FLOW

6.3UML DIAGRAMS

**7. OUTPUT SCREENS**

**8. SYSTEM TESTING AND IMPLEMENTATION**

8.1INTRODUCTION

8.2STRATEGIC APPROACH OF SOFTWARE TESTING

8.3UNIT TESTING

**9. SYSTEM SECURITY**

9.1INTRODUCTION

9.2SECURITY IN SOFTWARE

**10. CONCLUSION**

**11. FUTURE ENHANCEMENTS**

**Introduction**

The online shopping growing day to day. Credit cards are used for purchasing goods and services with the help of virtual card and physical card where as virtual card for online transaction and physical card for offline transaction. In a physical-card based purchase, the cardholder presents his card physically to a merchant for making a payment. To carry out fraudulent transactions in this kind of purchase, an attacker has to steal the credit card. If the cardholder does not realize the loss of card, it can lead to a substantial financial loss to the credit card company. In online payment mode, attackers need only little information for doing fraudulent transaction (secure code, card number, expiration date etc.). In this purchase method, mainly transactions will be done through Internet or telephone. To commit fraud in these types of purchases, a fraudster simply needs to know the card details. Most of the time, the genuine cardholder is not aware that someone else has seen or stolen his card information. The only way to detect this kind of fraud is to analyse the spending patterns on every card and to figure out any inconsistency with respect to the “usual” spending patterns. Fraud detection based on the analysis of existing purchase data of cardholder is a promising way to reduce the rate of successful credit card frauds. Since humans tend to exhibit specific behavioristic profiles, every cardholder can be represented by a set of patterns containing information about the typical purchase category, the time since the last purchase, the amount of money spent, etc. Deviation from such patterns is a potential threat to the system.

**ORGANIZATION PROFILE**

Software Solutions is an IT solution provider for a dynamic environment where business and technology strategies converge. Their approach focuses on new ways of business combining IT innovation and adoption while also leveraging an organization’s current IT assets. Their work with large global corporations and new products or services and to implement prudent business and technology strategies in today’s environment.

**Xxxxxxx’s** **RANGE OF EXPERTISE INCLUDES:**

* Software Development Services
* Engineering Services
* Systems Integration
* Customer Relationship Management
* Product Development
* Electronic Commerce
* Consulting
* IT Outsourcing

We apply technology with innovation and responsibility to achieve two broad objectives:

* Effectively address the business issues our customers face today.
* Generate new opportunities that will help them stay ahead in the future.

**THIS APPROACH RESTS ON:**

* A strategy where we architect, integrate and manage technology services and solutions - we call it AIM for success.
* A robust offshore development methodology and reduced demand on customer resources.
* A focus on the use of reusable frameworks to provide cost and times benefits.

They combine the best people, processes and technology to achieve excellent results - consistency. We offer customers the advantages of:

**SPEED:**

They understand the importance of timing, of getting there before the competition. A rich portfolio of reusable, modular frameworks helps jump-start projects. Tried and tested methodology ensures that we follow a predictable, low - risk path to achieve results. Our track record is testimony to complex projects delivered within and evens before schedule.

**EXPERTISE:**

Our teams combine cutting edge technology skills with rich domain expertise. What’s equally important - they share a strong customer orientation that means they actually start by listening to the customer. They’re focused on coming up with solutions that serve customer requirements today and anticipate future needs.

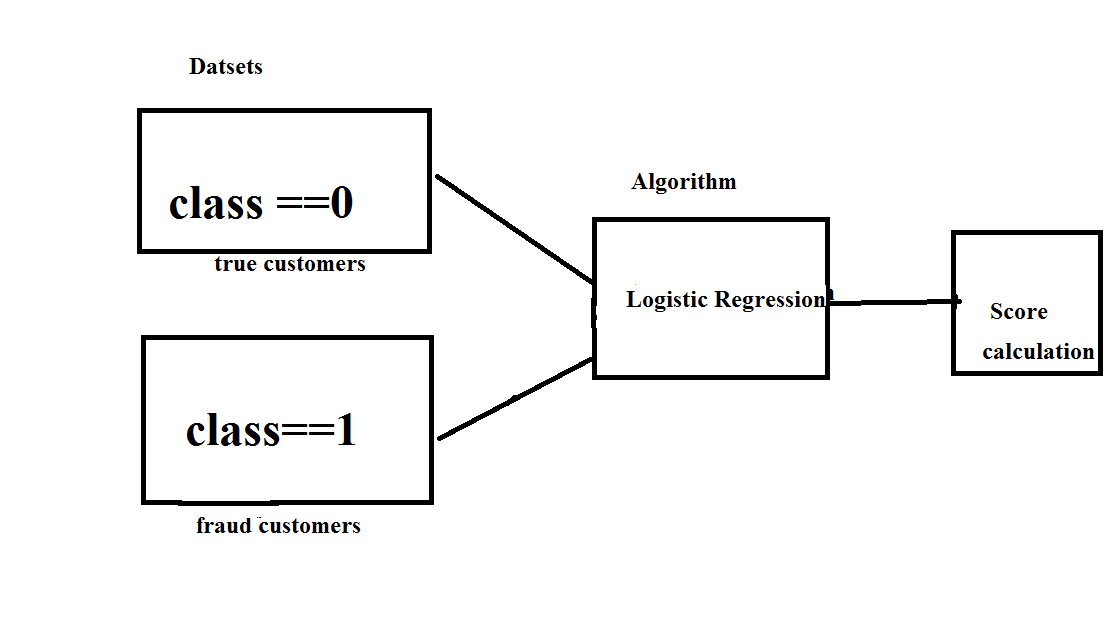
**A FULL SERVICE PORTFOLIO:**

They offer customers the advantage of being able to Architect, integrate and manage technology services. This means that they can rely on one, fully accountable source instead of trying to integrate disparate multi vendor solutions.

**SERVICES:**

Xxx is providing its services to companies which are in the field of production, quality control etc with their rich expertise and experience and information technology they are in best position to provide software solutions to distinct business requirements.

**Design:**

**Problem statement**

Credit card fraud stands as major problem for word wide financial institutions. Annual lost due to it scales to billions of dollars. We can observe this from many financial reports. Such as (Bhattacharyya et al., 2011) 10th annual online fraud report by Cyber Source shows that estimated loss due to online fraud is $4 billion for 2008 which is 11% increase than $3.6 billion loss in 2007and in 2006, fraud in United Kingdom alone was estimated to be £535 million in 2007 and now costing around 13.9 billion a year (Mahdi et al., 2010). From 2006 to 2008, UK alone has lost £427.0 million to £609.90 million due to credit and debit card fraud (Woolsey &Schulz, 2011). Although, there is some decrease in such losses after implementation of detection and prevention systems by government and bank, card-not-present fraud losses are increasing at higher rate due to online transactions. Worst thing is it is still increasing un-protective and un-detective way.

Over the year, government and banks have implemented some steps to subdue these frauds but along with the evolution of fraud detection and control methods, perpetrators are also evolving their methods and practices to avoid detection. Thus an effective and innovative methods need to be develop which will evolve accordingly to the need.

**Existing system**

This was on k-means Algorithm implementation, Only the two features with the most variance were used to train the model. The model was set to have 2 clusters, 0 being non-fraud and 1 being fraud. We also experimented with different values for the hyper parameters, but they all produced similar results. Changing the dimensionality of the data (reducing it to more dimensions than 2) also made little difference on the final values.

**Disadvantages:**

The Clustering doesn’t produce the less accuracy when compared to Regression methods in scenarios like credit card fraud detection. Comparatively with other algorithms k-means produce less accurate scores in prediction in this kind of scenarios

**Proposed System:**

Our goal is to implement machine learning model in order to classify, to the highest possible degree of accuracy, credit card fraud from a dataset gathered from Kaggle. After initial data exploration, we knew we would implement a logistic regression model for best accuracy reports.

Logistic regression, as it was a good candidate for binary classification. Python sklearn library was used to implement the project, We used Kaggle datasets for Credit card fraud detection, using pandas to data frame for class ==0 forno fraud and class==1 for fraud, matplotlib for plotting the fraud and non fraud data, train\_test\_split for data extraction (Split arrays or matrices into random train and test subsets) and used Logistic Regression machine learning algorithm for fraud detection and print predicting score according to the algorithm. Finally Confusion matrix was plotted on true and predicted.

**Advantages:**

* The results obtained by the Logistic Regression Algorithm is best compared to any other Algorithms.
* The Accuracy obtained was almost equal to cent percent which proves using of Logistic algorithm gives best results.
* The plots that were plotted according to the proper data that is processed during the implementation

Hardware Requirements:

•RAM: 4GB and Higher

•Processor :Intel i3 and above

•Hard Disk: 500GB: Minimum

Software Requirements:

•OS: Windows

•Python IDE : python 2.7.x and above

•Pycharm IDE

•setup tools and pip to be installed for 3.6.x and above

Hardware Requirements:

•RAM: 4GB and Higher

•Processor :Intel i3 and above

•Hard Disk: 500GB: Minimum

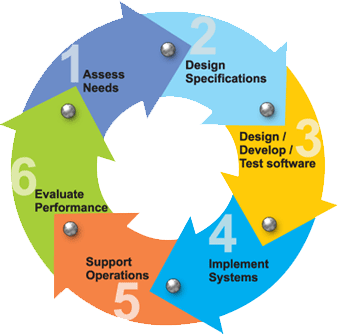
Software Requirements:

•OS: Windows

**2.1 INTRODUCTION**

**Software Development Life Cycle:-**

There is various software development approaches defined and designed which are used/employed during development process of software, these approaches are also referred as "Software Development Process Models". Each process model follows a particular life cycle in order to ensure success in process of software development.



**Requirements**

Business requirements are gathered in this phase.  This phase is the main focus of the project managers and stake holders.  Meetings with managers, stake holders and users are held in order to determine the requirements.  Who is going to use the system?  How will they use the system?  What data should be input into the system?  What data should be output by the system?  These are general questions that get answered during a requirements gathering phase.  This produces a nice big list of functionality that the system should provide, which describes functions the system should perform, business logic that processes data, what data is stored and used by the system, and how the user interface should work.  The overall result is the system as a whole and how it performs, not how it is actually going to do it.

**Design**

The software system design is produced from the results of the requirements phase.  Architects have the ball in their court during this phase and this is the phase in which their focus lies.  This is where the details on how the system will work is produced.  Architecture, including hardware and software, communication, software design (UML is produced here) are all part of the deliverables of a design phase.

**Implementation**

Code is produced from the deliverables of the design phase during implementation, and this is the longest phase of the software development life cycle.  For a developer, this is the main focus of the life cycle because this is where the code is produced.  Implementation my overlap with both the design and testing phases.  Many tools exists (CASE tools) to actually automate the production of code using information gathered and produced during the design phase.

**Testing**

During testing, the implementation is tested against the requirements to make sure that the product is actually solving the needs addressed and gathered during the requirements phase.  Unit tests and system/acceptance tests are done during this phase.  Unit tests act on a specific component of the system, while system tests act on the system as a whole.

So in a nutshell, that is a very basic overview of the general software development life cycle model.  Now let’s delve into some of the traditional and widely used variations.

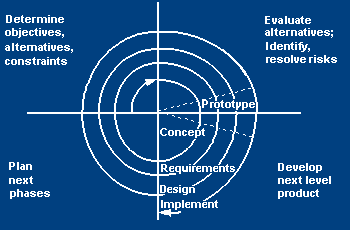
**SDLC METHDOLOGIES**

This document play a vital role in the development of life cycle (SDLC) as it describes the complete requirement of the system. It means for use by developers and will be the basic during testing phase. Any changes made to the requirements in the future will have to go through formal change approval process.

**SPIRAL MODEL** was defined by Barry Boehm in his 1988 article, “A spiral Model of Software Development and Enhancement. This model was not the first model to discuss iterative development, but it was the first model to explain why the iteration models.

As originally envisioned, the iterations were typically 6 months to 2 years long. Each phase starts with a design goal and ends with a client reviewing the progress thus far. Analysis and engineering efforts are applied at each phase of the project, with an eye toward the end goal of the project.

**The following diagram shows how a spiral model acts like:**



The steps for Spiral Model can be generalized as follows:

* The new system requirements are defined in as much details as possible. This usually involves interviewing a number of users representing all the external or internal users and other aspects of the existing system.
* A preliminary design is created for the new system.
* A first prototype of the new system is constructed from the preliminary design. This is usually a scaled-down system, and represents an approximation of the characteristics of the final product.
* A second prototype is evolved by a fourfold procedure:

1. Evaluating the first prototype in terms of its strengths, weakness, and risks.
2. Defining the requirements of the second prototype.
3. Planning an designing the second prototype.
4. Constructing and testing the second prototype.

* At the customer option, the entire project can be aborted if the risk is deemed too great. Risk factors might involved development cost overruns, operating-cost miscalculation, or any other factor that could, in the customer’s judgment, result in a less-than-satisfactory final product.
* The existing prototype is evaluated in the same manner as was the previous prototype, and if necessary, another prototype is developed from it according to the fourfold procedure outlined above.
* The preceding steps are iterated until the customer is satisfied that the refined prototype represents the final product desired.
* The final system is constructed, based on the refined prototype.
* The final system is thoroughly evaluated and tested. Routine maintenance is carried on a continuing basis to prevent large scale failures and to minimize down time.

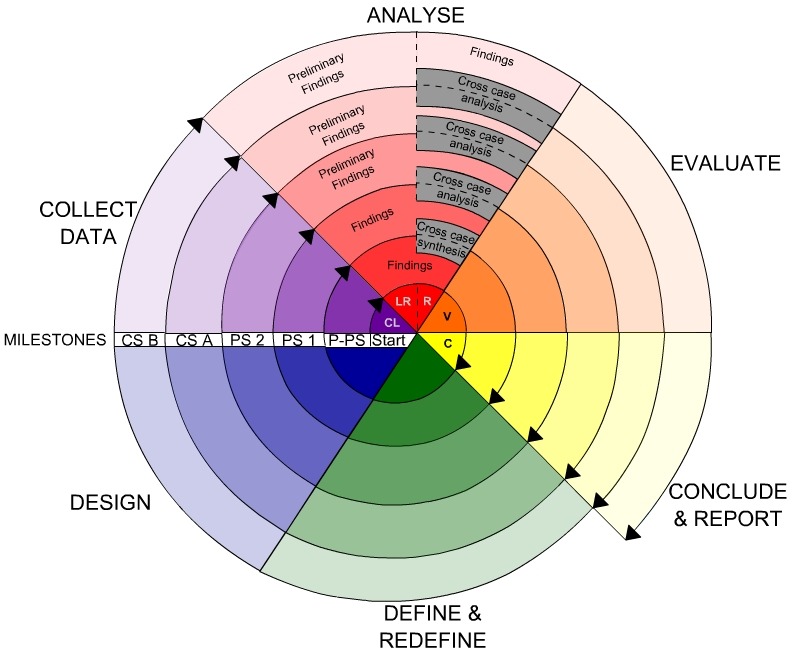
**2.2 STUDY OF THE SYSTEM**

In the flexibility of uses the interface has been developed a graphics concepts in mind, associated through a browser interface. The GUI’s at the top level has been categorized as follows

1. Administrative User Interface Design
2. The Operational and Generic User Interface Design

The administrative user interface concentrates on the consistent information that is practically, part of the organizational activities and which needs proper authentication for the data collection. The Interface helps the administration with all the transactional states like data insertion, data deletion, and data updating along with executive data search capabilities.

The operational and generic user interface helps the users upon the system in transactions through the existing data and required services. The operational user interface also helps the ordinary users in managing their own information helps the ordinary users in managing their own information in a customized manner as per the assisted flexibilities.



**3. Fundamental Concepts on (Domain)**

What is Cloudera?

Cloudera is revolutionizing enterprise data management by offering the first unified Platform for Big Data: The Enterprise Data Hub. Cloudera offers enterprises one place to store, process, and analyze all their data, empowering them to extend the value of existing investments while enabling fundamental new ways to derive value from their data.

Why do customers choose Cloudera?

Cloudera was the first commercial provider of python-related software and services and has the most customers with enterprise requirements, and the most experience supporting them, in the industry. Cloudera’s combined offering of differentiated software (open and closed source), support, training, professional services, and indemnity brings customers the greatest business value, in the shortest amount of time, at the lowest TCO.

**Data Mining**

There is a huge amount of data available in the Information Industry. This data is of no use until it is converted into useful information. It is necessary to analyze this huge amount of data and extract useful information from it. Extraction of information is not the only process we need to perform; data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. Once all these processes are over, we would be able to use this information in many applications such as Fraud Detection, Market Analysis, Production Control, Science Exploration, etc.

What is Data Mining?

Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data. The information or knowledge extracted so can be used for any of the following applications:

Market Analysis

Fraud Detection

Customer Retention

Production Control

Science Exploration

Data Mining Applications

Data mining is highly useful in the following domains: Market Analysis and Management

Corporate Analysis

& Risk Management Fraud Detection

Apart from these, data mining can also be used in the areas of production control, customer retention, science exploration, sports, astrology, and Internet Web Surf-Aid

Market Analysis and Management

Listed below are the various fields of market where data mining is used: Customer Profiling - Data mining helps determine what kind of people buy what kind of products.

Identifying Customer Requirements - Data mining helps in identifying thebest products for different customers. It uses prediction to find the factors that may attract new customers.

Cross Market Analysis - Data mining performs Association/correlations between product sales.

Target Marketing - Data mining helps to find clusters of model customers who share the same characteristics such as interests, spending habits, income, etc.

Determining Customer purchasing pattern - Data mining helps in determining customer purchasing pattern.

Providing Summary Information - Data mining provides us various multidimensional summary reports.

Corporate Analysis andRisk Management

Data mining is used in the following fields of the Corporate Sector: Finance Planning and Asset Evaluation - It involves cash flow analysis and prediction, contingent claim analysis to evaluate assets.

Resource Planning - It involves summarizing and comparing the resources and spending. Competition - It involves monitoring competitors and market directions.

Fraud Detection

Data mining is also used in the fields of credit card services and telecommunication to detect frauds. In fraud telephone calls, it helps to find the destination of the call, duration of the call, time of the day or week, etc. It also analyzes the patterns that deviate from expected norms.

Data mining essential step in the process of knowledge discovery

1.Data cleaning (to remove noise and inconsistent data)

2. Data integration (where multiple data sources may be combined)

3. Data selection (where data relevant to the analysis task are retrieved from the database)

4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)

5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns)

6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

Based on this view, the architecture of a typical data mining system may have the following major components

Database, data warehouse, World Wide Web, or other information repository: This is one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories.

Data cleaning and data integration techniques may be performed on the data. Database or data warehouse server: The database or data warehouse server is responsible for fetching the relevant data, based on the user’s data mining request.

Knowledge base: This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns. Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction. Knowledge such as user beliefs, which can be used to assess a pattern’s interestingness based on its unexpectedness, may also be included. Other examples of domain knowledge are additional interestingness constraints or thresholds, and metadata (e.g., describing data from multiple heterogeneous sources).

Data mining engine: This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.

Pattern evaluation module: This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search toward interesting patterns. It may use interestingness thresholds to filter out discovered patterns. Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used.

For efficient data mining, it is highly recommended to push the evaluation of pattern interestingness as deep as possible into the mining process so as to confine the search to only the interesting patterns.

User interface: This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results. In addition, this component allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

**DATA MINING ON WHAT KIND OF DATA?**

**RELATIONAL DATABASES**

A database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data A relational database is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (records or rows)

Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values A semantic data model, such as an entity-relationship (ER) data model, is often constructed for relational databases. An ER data model represents the database as a set of entities and their relationships Relational data can be accessed by database queries written in a relational query language, such as SQL

**DATA WAREHOUSES**

A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing

**OBJECT-RELATIONAL DATABASES**

Based on an object-relational data model Extends the relational model by providing a rich data type for handling complex objects and object orientation Objects that share a common set of properties can be grouped into an object class. Each object is an instance of its class. Object classes can be organized into class/subclass hierarchies

**ADVANCED DATA AND INFORMATION SYSTEMS**

With the progress of database technology, various kinds of advanced data and information systems have emerged and are undergoing development to address the requirements of new applications handling spatial/temporal data (such as maps) engineering design data (such as the design of buildings, system components, or integrated circuits) hypertext and multimedia data (including text, image, video, and audio data) time-related data (such as historical records or stock exchange data) stream data (such as video surveillance and sensor data, where data flow in and out like streams) the World Wide Web (a huge, widely distributed information repository made available by the Internet)

**THE WORLD WIDE WEB**

The World Wide Web and its associated distributed information services, such as Yahoo! and Google provide rich, worldwide, on-line information services, where data objects are linked together to facilitate interactive access Capturing user access patterns in such distributed information environments is called Web usage mining (or Weblog mining)

Database or data warehouse server responsible for fetching the relevant data, based on the user’s data mining request can be decouples/loose coupled/tightly coupled with the database layer

Knowledge base the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns interestingness constraints or thresholds, metadata, concept hierarchies, etc.

Data mining engine this is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis query languages (DMQL) based on mining primitives to access the data

Pattern evaluation module interacts with the data mining modules so as to focus the search toward interesting patterns may use interestingness thresholds to filter out discovered patterns may be integrated with the mining module

User interface communicates between users and the data mining system allows the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms

**4. System Analysis**

The **Systems Development Life Cycle (SDLC)**, or *Software Development Life Cycle* in [systems engineering](http://en.wikipedia.org/wiki/Systems_engineering), [information systems](http://en.wikipedia.org/wiki/Information_systems) and [software engineering](http://en.wikipedia.org/wiki/Software_engineering), is the process of creating or altering systems, and the models and [methodologies](http://en.wikipedia.org/wiki/Methodologies) that people use to develop these systems.

In software engineering the SDLC concept underpins many kinds of [software development methodologies](http://en.wikipedia.org/wiki/Software_development_methodologies). These methodologies form the framework for planning and controlling the creation of an information system the [software development process](http://en.wikipedia.org/wiki/Software_development_process).

**SOFTWARE MODEL OR ARCHITECTURE ANALYSIS:**

Structured project management techniques (such as an SDLC) enhance management’s control over projects by dividing complex tasks into manageable sections. A software life cycle model is either a descriptive or prescriptive characterization of how software is or should be developed. But none of the SDLC models discuss the key issues like Change management, Incident management and Release management processes within the SDLC process, but, it is addressed in the overall project management. In the proposed hypothetical model, the concept of user-developer interaction in the conventional SDLC model has been converted into a three dimensional model which comprises of the user, owner and the developer. In the proposed hypothetical model, the concept of user-developer interaction in the conventional SDLC model has been converted into a three dimensional model which comprises of the user, owner and the developer. The ―one size fits all‖ approach to applying SDLC methodologies is no longer appropriate. We have made an attempt to address the above mentioned defects by using a new hypothetical model for SDLC described elsewhere. The drawback of addressing these management processes under the overall project management is missing of key technical issues pertaining to software development process that is, these issues are talked in the project management at the surface level but not at the ground level.

**Proposed System:**

Our goal is to implement machine learning model in order to classify, to the highest possible degree of accuracy, credit card fraud from a dataset gathered from Kaggle. After initial data exploration, we knew we would implement a logistic regression model for best accuracy reports.

Logistic regression, as it was a good candidate for binary classification. Python sklearn library was used to implement the project, We used Kaggle datasets for Credit card fraud detection, using pandas to data frame for class ==0 forno fraud and class==1 for fraud, matplotlib for plotting the fraud and non fraud data, train\_test\_split for data extraction (Split arrays or matrices into random train and test subsets) and used Logistic Regression machine learning algorithm for fraud detection and print predicting score according to the algorithm. Finally Confusion matrix was plotted on true and predicted.

**2.5Functional requirements**

Outputs from computer systems are required primarily to communicate the results of processing to users. They are also used to provide a permanent copy of the results for later consultation. The various types of outputs in general are:

* External Outputs, whose destination is outside the organization,.
* Internal Outputs whose destination is within organization and they are the
* User’s main interface with the computer.
* Operational outputs whose use is purely within the computer department.
* Interface outputs, which involve the user in communicating directly.
* Understanding user’s preferences, expertise level and his business requirements through a friendly questionnaire.
* Input data can be in four different forms - Relational DB, text files, .xls and xml files. For testing and demo you can choose data from any domain. User-B can provide business data as input.

**Non-Functional Requirements**

1. Secure access of confidential data (user’s details). SSL can be used.
2. 24 X 7 availability.
3. Better component design to get better performance at peak time
4. Flexible service based architecture will be highly desirable for future extension

Preliminary investigation examine project feasibility, the likelihood the system will be useful to the organization. The main objective of the feasibility study is to test the Technical, Operational and Economical feasibility for adding new modules and debugging old running system. All system is feasible if they are unlimited resources and infinite time. There are aspects in the feasibility study portion of the preliminary investigation:

* Technical Feasibility
* Operational Feasibility
* Economical Feasibility

**3.1. TECHNICAL FEASIBILITY**

The technical issue usually raised during the feasibility stage of the investigation includes the following:

* Does the necessary technology exist to do what is suggested?
* Do the proposed equipments have the technical capacity to hold the data required to use the new system?
* Will the proposed system provide adequate response to inquiries, regardless of the number or location of users?
* Can the system be upgraded if developed?
* Are there technical guarantees of accuracy, reliability, ease of access and data security?

Earlier no system existed to cater to the needs of ‘Secure Infrastructure Implementation System’. The current system developed is technically feasible. It is a web based user interface for audit workflow at NIC-CSD. Thus it provides an easy access to the users. The database’s purpose is to create, establish and maintain a workflow among various entities in order to facilitate all concerned users in their various capacities or roles. Permission to the users would be granted based on the roles specified. Therefore, it provides the technical guarantee of accuracy, reliability and security. The software and hard requirements for the development of this project are not many and are already available in-house at NIC or are available as free as open source. The work for the project is done with the current equipment and existing software technology. Necessary bandwidth exists for providing a fast feedback to the users irrespective of the number of users using the system.

**3.2. OPERATIONAL FEASIBILITY**

Proposed projects are beneficial only if they can be turned out into information system. That will meet the organization’s operating requirements. Operational feasibility aspects of the project are to be taken as an important part of the project implementation. Some of the important issues raised are to test the operational feasibility of a project includes the following: -

* Is there sufficient support for the management from the users?
* Will the system be used and work properly if it is being developed and implemented?
* Will there be any resistance from the user that will undermine the possible application benefits?

This system is targeted to be in accordance with the above-mentioned issues. Beforehand, the management issues and user requirements have been taken into consideration. So there is no question of resistance from the users that can undermine the possible application benefits.

The well-planned design would ensure the optimal utilization of the computer resources and would help in the improvement of performance status.

**3.3. ECONOMICAL FEASIBILITY**

A system can be developed technically and that will be used if installed must still be a good investment for the organization. In the economical feasibility, the development cost in creating the system is evaluated against the ultimate benefit derived from the new systems. Financial benefits must equal or exceed the costs.

The system is economically feasible. It does not require any addition hardware or software. Since the interface for this system is developed using the existing resources and technologies available at NIC, There is nominal expenditure and economical feasibility for certain.

**Abstract:-**

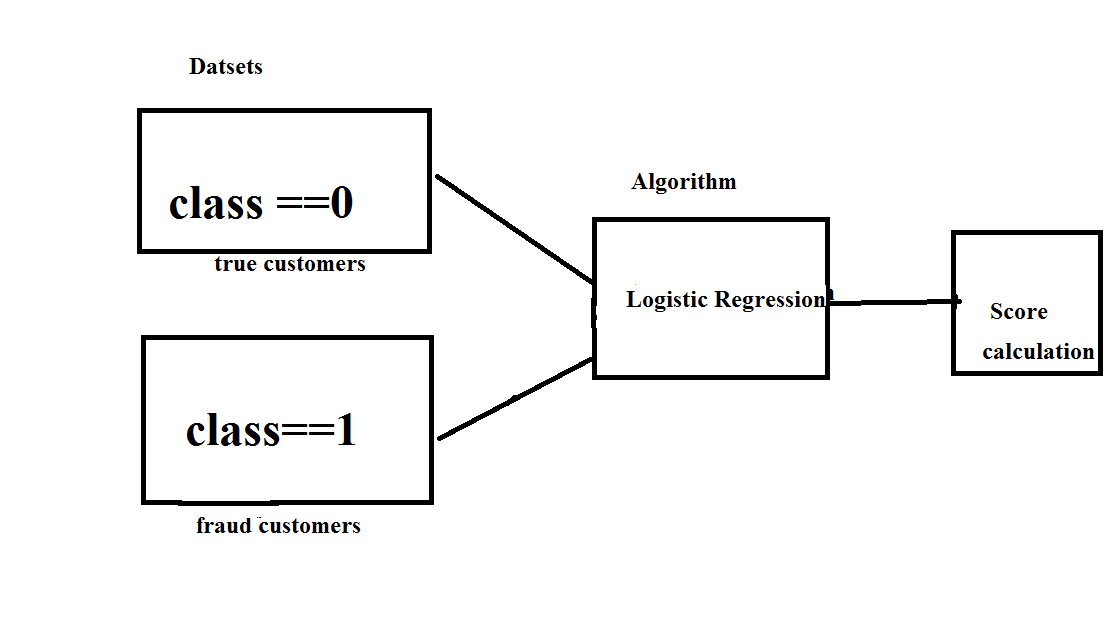
Billions of dollars of loss are caused every year by fraudulent credit card transactions. The design of efficient fraud detection algorithms is key for reducing these losses, and more and more algorithms rely on advanced machine learning techniques to assist fraud investigators. The design of fraud detection algorithms is however particularly challenging due to the non-stationary distribution of the data, the highly unbalanced classes distributions and the availability of few transactions labeled by fraud investigators. At the same time public data are scarcely available for confidentiality issues, leaving unanswered many questions about what is the best strategy. In this thesis we aim to provide some answers by focusing on crucial issues such as: i) why and how under sampling is useful in the presence of class imbalance (i.e. frauds are a small percentage of the transactions), ii) how to deal with unbalanced and evolving data streams (non-stationarity due to fraud evolution and change of spending behavior), iii) how to assess performances in a way which is relevant for detection and iv) how to use feedbacks provided by investigators on the fraud alerts generated. Finally, we design and assess a prototype of a Fraud Detection System able to meet real-world working conditions and that is able to integrate investigators’ feedback to generate accurate alerts.

**Index Terms**— Hidden Markov Model, credit card, fraud detection, online shopping, e-commerce (HMM)

**Introduction**

The online shopping growing day to day. Credit cards are used for purchasing goods and services with the help of virtual card and physical card where as virtual card for online transaction and physical card for offline transaction. In a physical-card based purchase, the cardholder presents his card physically to a merchant for making a payment. To carry out fraudulent transactions in this kind of purchase, an attacker has to steal the credit card. If the cardholder does not realize the loss of card, it can lead to a substantial financial loss to the credit card company. In online payment mode, attackers need only little information for doing fraudulent transaction (secure code, card number, expiration date etc.). In this purchase method, mainly transactions will be done through Internet or telephone. To commit fraud in these types of purchases, a fraudster simply needs to know the card details. Most of the time, the genuine cardholder is not aware that someone else has seen or stolen his card information. The only way to detect this kind of fraud is to analyse the spending patterns on every card and to figure out any inconsistency with respect to the “usual” spending patterns. Fraud detection based on the analysis of existing purchase data of cardholder is a promising way to reduce the rate of successful credit card frauds. Since humans tend to exhibit specific behavioristic profiles, every cardholder can be represented by a set of patterns containing information about the typical purchase category, the time since the last purchase, the amount of money spent, etc. Deviation from such patterns is a potential threat to the system.

**Design:**



**Problem statement**

Credit card fraud stands as major problem for word wide financial institutions. Annual lost due to it scales to billions of dollars. We can observe this from many financial reports. Such as (Bhattacharyya et al., 2011) 10th annual online fraud report by Cyber Source shows that estimated loss due to online fraud is $4 billion for 2008 which is 11% increase than $3.6 billion loss in 2007and in 2006, fraud in United Kingdom alone was estimated to be £535 million in 2007 and now costing around 13.9 billion a year (Mahdi et al., 2010). From 2006 to 2008, UK alone has lost £427.0 million to £609.90 million due to credit and debit card fraud (Woolsey &Schulz, 2011). Although, there is some decrease in such losses after implementation of detection and prevention systems by government and bank, card-not-present fraud losses are increasing at higher rate due to online transactions. Worst thing is it is still increasing un-protective and un-detective way.

Over the year, government and banks have implemented some steps to subdue these frauds but along with the evolution of fraud detection and control methods, perpetrators are also evolving their methods and practices to avoid detection. Thus an effective and innovative methods need to be develop which will evolve accordingly to the need.

**Existing system**

This was on k-means Algorithm implementation, Only the two features with the most variance were used to train the model. The model was set to have 2 clusters, 0 being non-fraud and 1 being fraud. We also experimented with different values for the hyper parameters, but they all produced similar results. Changing the dimensionality of the data (reducing it to more dimensions than 2) also made little difference on the final values.

**Disadvantages:**

The Clustering doesn’t produce the less accuracy when compared to Regression methods in scenarios like credit card fraud detection. Comparatively with other algorithms k-means produce less accurate scores in prediction in this kind of scenarios

**Proposed System:**

Our goal is to implement machine learning model in order to classify, to the highest possible degree of accuracy, credit card fraud from a dataset gathered from Kaggle. After initial data exploration, we knew we would implement a logistic regression model for best accuracy reports.

Logistic regression, as it was a good candidate for binary classification. Python sklearn library was used to implement the project, We used Kaggle datasets for Credit card fraud detection, using pandas to data frame for class ==0 forno fraud and class==1 for fraud, matplotlib for plotting the fraud and non fraud data, train\_test\_split for data extraction (Split arrays or matrices into random train and test subsets) and used Logistic Regression machine learning algorithm for fraud detection and print predicting score according to the algorithm. Finally Confusion matrix was plotted on true and predicted.

**Advantages:**

* The results obtained by the Logistic Regression Algorithm is best compared to any other Algorithms.
* The Accuracy obtained was almost equal to cent percent which proves using of Logistic algorithm gives best results.
* The plots that were plotted according to the proper data that is processed during the implementation

Hardware Requirements:

• RAM: 4GB and Higher

• Processor :Intel i3 and above

• Hard Disk: 500GB: Minimum

Software Requirements:

• OS: Windows

• Python IDE : python 2.7.x and above

• Pycharm IDE

• setup tools and pip to be installed for 3.6.x and above

Hardware Requirements:

• RAM: 4GB and Higher

• Processor :Intel i3 and above

• Hard Disk: 500GB: Minimum

Software Requirements:

• OS: Windows

• Python IDE : python 2.7.x and above

• Pycharm IDE

• setup tools and pip to be installed for 3.6.x and above

**Hardware Requirements:**

• RAM: 4GB and Higher

• Processor: Intel i3 and above

• Hard Disk: 500GB: Minimum

**Software Requirements:**

* OS: Windows or Linux
* Python IDE : python 2.7.x and above
* Pycharm IDE Required
* Setup tools and pip to be installed for 3.6 and above
* Language : Python Scripting

**Functional requirement:**

Outputs from computer systems are required primarily to communicate the results of processing to users. They are also used to provide a permanent copy of the results for later consultation. The various types of outputs in general are:

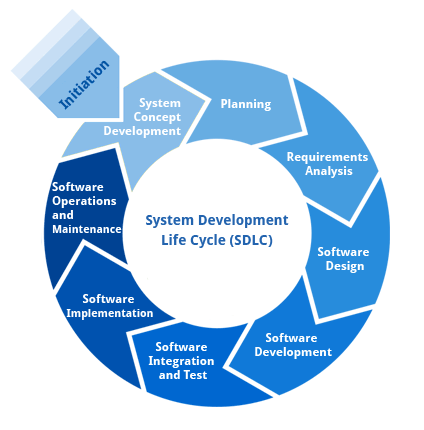
* External Outputs, whose destination is outside the organization,.
* Internal Outputs whose destination is within organization and they are the
* User’s main interface with the computer.
* Operational outputs whose use is purely within the computer department.
* Interface outputs, which involve the user in communicating directly.
* Understanding user’s preferences, expertise level and his business requirements through a friendly questionnaire.
* Input data can be in four different forms - Relational DB, text files, .xls and xml files. For testing and demo you can choose data from any domain. User-B can provide business data as input.

**Non functional requirements:**

* Secure access of confidential data (user’s details). SSL can be used.
* 24 X 7 availability.
* Better component design to get better performance at peak time
* Flexible service based architecture will be highly desirable for future extension

**Software Development Life Cycle**

The **Systems Development Life Cycle (SDLC)**, or Software Development Life Cycle in systems engineering, information systems and software engineering, is the process of creating or altering systems, and the models and methodologies use to develop these systems.



**Requirement Analysis and Design**

Analysis gathers the requirements for the system. This stage includes a detailed study of the business needs of the organization. Options for changing the business process may be considered. Design focuses on high level design like, what programs are needed and how are they going to interact, low-level design (how the individual programs are going to work), interface design (what are the interfaces going to look like) and data design (what data will be required). During these phases, the software's overall structure is defined. Analysis and Design are very crucial in the whole development cycle. Any glitch in the design phase could be very expensive to solve in the later stage of the software development. Much care is taken during this phase. The logical system of the product is developed in this phase.

**Implementation**

In this phase the designs are translated into code. Computer programs are written using a conventional programming language or an application generator. Programming tools like Compilers, Interpreters, and Debuggers are used to generate the code. Different high level programming languages like C, C++, Pascal, Java, .Net are used for coding. With respect to the type of application, the right programming language is chosen.

**Testing**

In this phase the system is tested. Normally programs are written as a series of individual modules, this subject to separate and detailed test. The system is then tested as a whole. The separate modules are brought together and tested as a complete system. The system is tested to ensure that interfaces between modules work (integration testing), the system works on the intended platform and with the expected volume of data (volume testing) and that the system does what the user requires (acceptance/beta testing).

**Maintenance**

Inevitably the system will need maintenance. Software will definitely undergo change once it is delivered to the customer. There are many reasons for the change. Change could happen because of some unexpected input values into the system. In addition, the changes in the system could directly affect the software operations. The software should be developed to accommodate changes that could happen during the post implementation period.

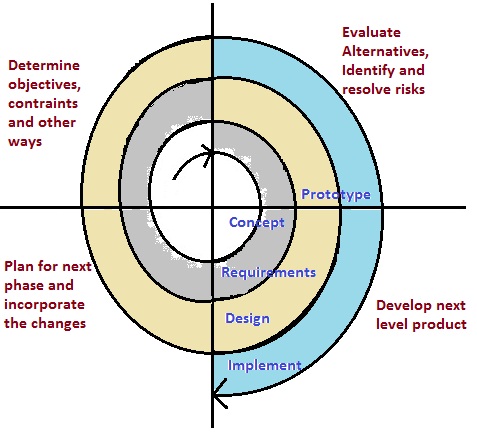
**SDLC METHDOLOGIES**

This document play a vital role in the development of life cycle (SDLC) as it describes the complete requirement of the system. It means for use by developers and will be the basic during testing phase. Any changes made to the requirements in the future will have to go through formal change approval process.

**SPIRAL MODEL** was defined by Barry Boehm in his 1988 article, “A spiral Model of Software Development and Enhancement. This model was not the first model to discuss iterative development, but it was the first model to explain why the iteration models.

As originally envisioned, the iterations were typically 6 months to 2 years long. Each phase starts with a design goal and ends with a client reviewing the progress thus far. Analysis and engineering efforts are applied at each phase of the project, with an eye toward the end goal of the project.

**The following diagram shows how a spiral model acts like:**



The steps for Spiral Model can be generalized as follows:

* The new system requirements are defined in as much details as possible. This usually involves interviewing a number of users representing all the external or internal users and other aspects of the existing system.
* A preliminary design is created for the new system.
* A first prototype of the new system is constructed from the preliminary design. This is usually a scaled-down system, and represents an approximation of the characteristics of the final product.
* A second prototype is evolved by a fourfold procedure:

1. Evaluating the first prototype in terms of its strengths, weakness, and risks.
2. Defining the requirements of the second prototype.
3. Planning an designing the second prototype.
4. Constructing and testing the second prototype.

* At the customer option, the entire project can be aborted if the risk is deemed too great. Risk factors might involved development cost overruns, operating-cost miscalculation, or any other factor that could, in the customer’s judgment, result in a less-than-satisfactory final product.
* The existing prototype is evaluated in the same manner as was the previous prototype, and if necessary, another prototype is developed from it according to the fourfold procedure outlined above.
* The preceding steps are iterated until the customer is satisfied that the refined prototype represents the final product desired.
* The final system is constructed, based on the refined prototype.
* The final system is thoroughly evaluated and tested. Routine maintenance is carried on a continuing basis to prevent large scale failures and to minimize down time.

## *4.1. FUNCTIONAL REQUIREMENTS*

**OUTPUT DESIGN**

Outputs from computer systems are required primarily to communicate the results of processing to users. They are also used to provides a permanent copy of the results for later consultation. The various types of outputs in general are:

* External Outputs, whose destination is outside the organization
* Internal Outputs whose destination is within organization and they are the
* User’s main interface with the computer.
* Operational outputs whose use is purely within the computer department.
* Interface outputs, which involve the user in communicating directly.

**OUTPUT DEFINITION**

# The outputs should be defined in terms of the following points:

* + - Type of the output
    - Content of the output
    - Format of the output
    - Location of the output
    - Frequency of the output
    - Volume of the output
    - Sequence of the output

It is not always desirable to print or display data as it is held on a computer. It should be decided as which form of the output is the most suitable.

**INPUT DESIGN**

Input design is a part of overall system design. The main objective during the input design is as given below:

* To produce a cost-effective method of input.
* To achieve the highest possible level of accuracy.
* To ensure that the input is acceptable and understood by the user.

**INPUT STAGES:**

The main input stages can be listed as below:

* Data recording
* Data transcription
* Data conversion
* Data verification
* Data control
* Data transmission
* Data validation
* Data correction

**INPUT TYPES:**

It is necessary to determine the various types of inputs. Inputs can be categorized as follows:

* External inputs, which are prime inputs for the system.
* Internal inputs, which are user communications with the system.
* Operational, which are computer department’s communications to the system?
* Interactive, which are inputs entered during a dialogue.

**INPUT MEDIA:**

At this stage choice has to be made about the input media. To conclude about the input media consideration has to be given to;

* Type of input
* Flexibility of format
* Speed
* Accuracy
* Verification methods
* Rejection rates
* Ease of correction
* Storage and handling requirements
* Security
* Easy to use
* Portability

Keeping in view the above description of the input types and input media, it can be said that most of the inputs are of the form of internal and interactive. As

Input data is to be the directly keyed in by the user, the keyboard can be considered to be the most suitable input device.

**ERROR AVOIDANCE**

At this stage care is to be taken to ensure that input data remains accurate form the stage at which it is recorded up to the stage in which the data is accepted by the system. This can be achieved only by means of careful control each time the data is handled.

**ERROR DETECTION**

Even though every effort is make to avoid the occurrence of errors, still a small proportion of errors is always likely to occur, these types of errors can be discovered by using validations to check the input data.

**DATA VALIDATION**

Procedures are designed to detect errors in data at a lower level of detail. Data validations have been included in the system in almost every area where there is a possibility for the user to commit errors. The system will not accept invalid data. Whenever an invalid data is keyed in, the system immediately prompts the user and the user has to again key in the data and the system will accept the data only if the data is correct. Validations have been included where necessary.

The system is designed to be a user friendly one. In other words the system has been designed to communicate effectively with the user. The system has been designed with popup menus.

**USER INTERFACE DESIGN**

It is essential to consult the system users and discuss their needs while designing the user interface:

**USER INTERFACE SYSTEMS CAN BE BROADLY CLASIFIED AS:**

1. User initiated interface the user is in charge, controlling the progress of the user/computer dialogue. In the computer-initiated interface, the computer selects the next stage in the interaction.
2. Computer initiated interfaces

In the computer initiated interfaces the computer guides the progress of the user/computer dialogue. Information is displayed and the user response of the computer takes action or displays further information.

**USER\_INITIATED INTERGFACES**

User initiated interfaces fall into tow approximate classes:

1. Command driven interfaces: In this type of interface the user inputs commands or queries which are interpreted by the computer.
2. Forms oriented interface: The user calls up an image of the form to his/her screen and fills in the form. The forms oriented interface is chosen because it is the best choice.

**COMPUTER-INITIATED INTERFACES**

The following computer – initiated interfaces were used:

1. The menu system for the user is presented with a list of alternatives and the user chooses one; of alternatives.
2. Questions – answer type dialog system where the computer asks question and takes action based on the basis of the users reply.

Right from the start the system is going to be menu driven, the opening menu displays the available options. Choosing one option gives another popup menu with more options. In this way every option leads the users to data entry form where the user can key in the data.

**ERROR MESSAGE DESIGN:**

The design of error messages is an important part of the user interface design. As user is bound to commit some errors or other while designing a system the system should be designed to be helpful by providing the user with information regarding the error he/she has committed.

This application must be able to produce output at different modules for different inputs.

**4.2. PERFORMANCE REQUIREMENTS**

Performance is measured in terms of the output provided by the application.

Requirement specification plays an important part in the analysis of a system. Only when the requirement specifications are properly given, it is possible to design a system, which will fit into required environment. It rests largely in the part of the users of the existing system to give the requirement specifications because they are the people who finally use the system. This is because the requirements have to be known during the initial stages so that the system can be designed according to those requirements. It is very difficult to change the system once it has been designed and on the other hand designing a system, which does not cater to the requirements of the user, is of no use.

The requirement specification for any system can be broadly stated as given below:

* The system should be able to interface with the existing system
* The system should be accurate
* The system should be better than the existing system

The existing system is completely dependent on the user to perform all the duties.

**IMPLIMENTATION ON (PYTHON)**

What Is A Script?

Up to this point, I have concentrated on the interactive programming capability of Python.  This is a very useful capability that allows you to type in a program and to have it executed immediately in an interactive mode

**Scripts are reusable**

Basically, a script is a text file containing the statements that comprise a Python program.  Once you have created the script, you can execute it over and over without having to retype it each time.

**Scripts are editable**

Perhaps, more importantly, you can make  different versions of the script by modifying the statements from one file to the next using a text editor.  Then you can execute each of the individual versions.  In this way, it is easy to create different programs with a minimum amount of typing.

**You will need a text editor**

Just about any text editor will suffice for creating Python script files.

You can use *Microsoft Notepad, Microsoft WordPad, Microsoft Word,*or just about any word processor if you want to.

**Difference between a script and a program**

Script:

Scripts are distinct from the core code of the application, which is usually written in a different language, and are often created or at least modified by the end-user. Scripts are often interpreted from source code or byte code, where as the applications they control are traditionally compiled to native machine code.

Program:

The program has an executable form that the computer can use directly to execute the instructions.

The same program in its human-readable source code form, from which executable programs are derived(e.g., compiled)

**Python**

what is Python? Chances you are asking yourself this. You may have found this book because you want to learn to program but don’t know anything about programming languages. Or you may have heard of programming languages like C, C++, C#, or Java and want to know what Python is and how it compares to “big name” languages. Hopefully I can explain it for you.

Python concepts

If your not interested in the the hows and whys of Python, feel free to skip to the next chapter. In this chapter I will try to explain to the reader why I think Python is one of the best languages available and why it’s a great one to start programming with.

• Open source general-purpose language.

• Object Oriented, Procedural, Functional

• Easy to interface with C/ObjC/Java/Fortran

• Easy-ish to interface with C++ (via SWIG)

• Great interactive environment

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

* **Python is Interpreted** − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
* **Python is Interactive** − You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.
* **Python is Object-Oriented** − Python supports Object-Oriented style or technique of programming that encapsulates code within objects.
* **Python is a Beginner's Language** − Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

**History of Python**

Python was developed by Guido van Rossum in the late eighties and early nineties at the National Research Institute for Mathematics and Computer Science in the Netherlands.

Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, SmallTalk, and Unix shell and other scripting languages.

Python is copyrighted. Like Perl, Python source code is now available under the GNU General Public License (GPL).

Python is now maintained by a core development team at the institute, although Guido van Rossum still holds a vital role in directing its progress.

**Python Features**

Python's features include −

* **Easy-to-learn** − Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.
* **Easy-to-read** − Python code is more clearly defined and visible to the eyes.
* **Easy-to-maintain** − Python's source code is fairly easy-to-maintain.
* **A broad standard library** − Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.
* **Interactive Mode** − Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.
* **Portable** − Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
* **Extendable** − You can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.
* **Databases** − Python provides interfaces to all major commercial databases.
* **GUI Programming** − Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
* **Scalable** − Python provides a better structure and support for large programs than shell scripting.

Apart from the above-mentioned features, Python has a big list of good features, few are listed below −

* It supports functional and structured programming methods as well as OOP.
* It can be used as a scripting language or can be compiled to byte-code for building large applications.
* It provides very high-level dynamic data types and supports dynamic type checking.
* IT supports automatic garbage collection.
* It can be easily integrated with C, C++, COM, ActiveX, CORBA, and Java.

**Dynamic vs Static**

Types Python is a dynamic-typed language. Many other languages are static typed, such as C/C++ and Java. A static typed language requires the programmer to explicitly tell the computer what type of “thing” each data value is.

For example, in C if you had a variable that was to contain the price of something, you would have to declare the variable as a “float” type.

This tells the compiler that the only data that can be used for that variable must be a floating point number, i.e. a number with a decimal point.

If any other data value was assigned to that variable, the compiler would give an error when trying to compile the program.

Python, however, doesn’t require this. You simply give your variables names and assign values to them. The interpreter takes care of keeping track of what kinds of objects your program is using. This also means that you can change the size of the values as you develop the program. Say you have another decimal number (a.k.a. a floating point number) you need in your program.

With a static typed language, you have to decide the memory size the variable can take when you first initialize that variable. A double is a floating point value that can handle a much larger number than a normal float (the actual memory sizes depend on the operating environment).

If you declare a variable to be a float but later on assign a value that is too big to it, your program will fail; you will have to go back and change that variable to be a double.

With Python, it doesn’t matter. You simply give it whatever number you want and Python will take care of manipulating it as needed. It even works for derived values.

For example, say you are dividing two numbers. One is a floating point number and one is an integer. Python realizes that it’s more accurate to keep track of decimals so it automatically calculates the result as a floating point number

**Variables**

Variables are nothing but reserved memory locations to store values. This means that when you create a variable you reserve some space in memory.

Based on the data type of a variable, the interpreter allocates memory and decides what can be stored in the reserved memory. Therefore, by assigning different data types to variables, you can store integers, decimals or characters in these variables.

**Standard Data Types**

The data stored in memory can be of many types. For example, a person's age is stored as a numeric value and his or her address is stored as alphanumeric characters. Python has various standard data types that are used to define the operations possible on them and the storage method for each of them.

Python has five standard data types −

* Numbers
* String
* List
* Tuple
* Dictionary

Python Numbers

Number data types store numeric values. Number objects are created when you assign a value to them

## Python Strings

Strings in Python are identified as a contiguous set of characters represented in the quotation marks. Python allows for either pairs of single or double quotes. Subsets of strings can be taken using the slice operator ([ ] and [:] ) with indexes starting at 0 in the beginning of the string and working their way from -1 at the end.

## Python Lists

Lists are the most versatile of Python's compound data types. A list contains items separated by commas and enclosed within square brackets ([]). To some extent, lists are similar to arrays in C. One difference between them is that all the items belonging to a list can be of different data type.

The values stored in a list can be accessed using the slice operator ([ ] and [:]) with indexes starting at 0 in the beginning of the list and working their way to end -1. The plus (+) sign is the list concatenation operator, and the asterisk (\*) is the repetition operator.

## Python Tuples

A tuple is another sequence data type that is similar to the list. A tuple consists of a number of values separated by commas. Unlike lists, however, tuples are enclosed within parentheses.

The main differences between lists and tuples are: Lists are enclosed in brackets ( [ ] ) and their elements and size can be changed, while tuples are enclosed in parentheses ( ( ) ) and cannot be updated. Tuples can be thought of as **read-only** lists.

## Python Dictionary

Python's dictionaries are kind of hash table type. They work like associative arrays or hashes found in Perl and consist of key-value pairs. A dictionary key can be almost any Python type, but are usually numbers or strings. Values, on the other hand, can be any arbitrary Python object.

Dictionaries are enclosed by curly braces ({ }) and values can be assigned and accessed using square braces ([]).

**Different modes in python**

Python has two basic modes: normal and interactive.

The normal mode is the mode where the scripted and finished .py files are run in the Python interpreter.

Interactive mode is a command line shell which gives immediate feedback for each statement, while running previously fed statements in active memory. As new lines are fed into the interpreter, the fed program is evaluated both in part and in whole

# 20 Python libraries

**1.** Requests. The most famous http library written by kenneth reitz. It’s a must have for every python developer.

**2.** Scrapy. If you are involved in webscraping then this is a must have library for you. After using this library you won’t use any other.

**3.** wxPython. A gui toolkit for python. I have primarily used it in place of tkinter. You will really love it.

**4.** Pillow. A friendly fork of PIL (Python Imaging Library). It is more user friendly than PIL and is a must have for anyone who works with images.

**5.** SQLAlchemy. A database library. Many love it and many hate it. The choice is yours.

**6.** BeautifulSoup. I know it’s slow but this xml and html parsing library is very useful for beginners.

**7.** Twisted. The most important tool for any network application developer. It has a very beautiful api and is used by a lot of famous python developers.

**8.** NumPy. How can we leave this very important library ? It provides some advance math functionalities to python.

**9.** SciPy. When we talk about NumPy then we have to talk about scipy. It is a library of algorithms and mathematical tools for python and has caused many scientists to switch from ruby to python.

**10.** matplotlib. A numerical plotting library. It is very useful for any data scientist or any data analyzer.

**11.** Pygame. Which developer does not like to play games and develop them ? This library will help you achieve your goal of 2d game development.

**12.** Pyglet. A 3d animation and game creation engine. This is the engine in which the famous [python port](https://github.com/fogleman/Minecraft) of minecraft was made

**13.** pyQT. A GUI toolkit for python. It is my second choice after wxpython for developing GUI’s for my python scripts.

**14.** pyGtk. Another python GUI library. It is the same library in which the famous Bittorrent client is created.

**15.** Scapy. A packet sniffer and analyzer for python made in python.

**16.** pywin32. A python library which provides some useful methods and classes for interacting with windows.

**17.** nltk. Natural Language Toolkit – I realize most people won’t be using this one, but it’s generic enough. It is a very useful library if you want to manipulate strings. But it’s capacity is beyond that. Do check it out.

**18.** nose. A testing framework for python. It is used by millions of python developers. It is a must have if you do test driven development.

**19.** SymPy. SymPy can do algebraic evaluation, differentiation, expansion, complex numbers, etc. It is contained in a pure Python distribution.

**20.** IPython. I just can’t stress enough how useful this tool is. It is a python prompt on steroids. It has completion, history, shell capabilities, and a lot more. Make sure that you take a look at it.

**Numpy**

NumPy’s main object is the homogeneous multidimensional array. It is a table of elements (usually numbers), all of the same type, indexed by a tuple of positive **integers. In NumPy dimensions are called axes. The number of axes is rank.**

**• Offers Matlab**-ish capabilities within Python

• Fast array operations

• 2D arrays, multi-D arrays, linear algebra etc.

**matplotlib**

• High quality plotting library.

**Python class and objects**

These are the building blocks of OOP. class creates a new object. This object can be anything, whether an abstract data concept or a model of a physical object, e.g. a chair. Each class has individual characteristics unique to that class, including variables and methods. Classes are very powerful and currently “the big thing” in most programming languages. Hence, there are several chapters dedicated to OOP later in the book.

The class is the most basic component of object-oriented programming. Previously, you learned how to use functions to make your program do something.

Now will move into the big, scary world of Object-Oriented Programming (OOP). To be honest, it took me several months to get a handle on objects.

When I first learned C and C++, I did great; functions just made sense for me.

Having messed around with BASIC in the early ’90s, I realized functions were just like subroutines so there wasn’t much new to learn.

However, when my C++ course started talking about objects, classes, and all the new features of OOP, my grades definitely suffered.

Once you learn OOP, you’ll realize that it’s actually a pretty powerful tool. Plus many Python libraries and APIs use classes, so you should at least be able to understand what the code is doing.

One thing to note about Python and OOP: it’s not mandatory to use objects in your code in a way that works best; maybe you don’t need to have a full-blown class with initialization code and methods to just return a calculation. With Python, you can get as technical as you want.

As you’ve already seen, Python can do just fine with functions. Unlike languages such as Java, you aren’t tied down to a single way of doing things; you can mix functions and classes as necessary in the same program. This lets you build the code

Objects are an encapsulation of variables and functions into a single entity. Objects get their variables and functions from classes. Classes are essentially a template to create your objects.

Here’s a brief list of Python OOP ideas:

• The class statement creates a class object and gives it a name. This creates a new namespace.

• Assignments within the class create class attributes. These attributes are accessed by qualifying the name using dot syntax: ClassName.Attribute.

• Class attributes export the state of an object and its associated behavior. These attributes are shared by all instances of a class.

• Calling a class (just like a function) creates a new instance of the class.

This is where the multiple copies part comes in.

• Each instance gets ("inherits") the default class attributes and gets its own namespace. This prevents instance objects from overlapping and confusing the program.

• Using the term self identifies a particular instance, allowing for per-instance attributes. This allows items such as variables to be associated with a particular instance.

**Inheritance**

First off, classes allow you to modify a program without really making changes to it.

To elaborate, by subclassing a class, you can change the behavior of the program by simply adding new components to it rather than rewriting the existing components.

As we’ve seen, an instance of a class inherits the attributes of that class.

However, classes can also inherit attributes from other classes. Hence, a subclass inherits from a superclass allowing you to make a generic superclass that is specialized via subclasses.

The subclasses can override the logic in a superclass, allowing you to change the behavior of your classes without changing the superclass at all.

Operator Overloads

Operator overloading simply means that objects that you create from classes can respond to actions (operations) that are already defined within Python, such as addition, slicing, printing, etc.

Even though these actions can be implemented via class methods, using overloading ties the behavior closer to Python’s object model and the object interfaces are more consistent to Python’s built-in objects, hence overloading is easier to learn and use.

User-made classes can override nearly all of Python’s built-in operation methods

**Exceptions**

I’ve talked about exceptions before but now I will talk about them in depth. Essentially, exceptions are events that modify program’s flow, either intentionally or due to errors.

They are special events that can occur due to an error, e.g. trying to open a file that doesn’t exist, or when the program reaches a marker, such as the completion of a loop.

Exceptions, by definition, don’t occur very often; hence, they are the "exception to the rule" and a special class has been created for them. Exceptions are everywhere in Python.

Virtually every module in the standard Python library uses them, and Python itself will raise them in a lot of different circumstances.

Here are just a few examples:

• Accessing a non−existent dictionary key will raise a KeyError exception.

• Searching a list for a non−existent value will raise a ValueError exception

. • Calling a non−existent method will raise an AttributeError exception.

• Referencing a non−existent variable will raise a NameError exception.

• Mixing datatypes without coercion will raise a TypeError exception.

One use of exceptions is to catch a fault and allow the program to continue working; we have seen this before when we talked about files.

This is the most common way to use exceptions. When programming with the Python command line interpreter, you don’t need to worry about catching exceptions.

Your program is usually short enough to not be hurt too much if an exception occurs.

Plus, having the exception occur at the command line is a quick and easy way to tell if your code logic has a problem.

However, if the same error occurred in your real program, it will fail and stop working. Exceptions can be created manually in the code by raising an exception.

It operates exactly as a system-caused exceptions, except that the programmer is doing it on purpose. This can be for a number of reasons. One of the benefits of using exceptions is that, by their nature, they don’t put any overhead on the code processing.

Because exceptions aren’t supposed to happen very often, they aren’t processed until they occur.

Exceptions can be thought of as a special form of the if/elif statements. You can realistically do the same thing with if blocks as you can with exceptions.

However, as already mentioned, exceptions aren’t processed until they occur; if blocks are processed all the time.

Proper use of exceptions can help the performance of your program.

The more infrequent the error might occur, the better off you are to use exceptions; using if blocks requires Python to always test extra conditions before continuing.

Exceptions also make code management easier: if your programming logic is mixed in with error-handling if statements, it can be difficult to read, modify, and debug your program.

User-Defined Exceptions

I won’t spend too much time talking about this, but Python does allow for a programmer to create his own exceptions.

You probably won’t have to do this very often but it’s nice to have the option when necessary.

However, before making your own exceptions, make sure there isn’t one of the built-in exceptions that will work for you.

They have been "tested by fire" over the years and not only work effectively, they have been optimized for performance and are bug-free.

Making your own exceptions involves object-oriented programming, which will be covered in the next chapter

. To make a custom exception, the programmer determines which base exception to use as the class to inherit from, e.g. making an exception for negative numbers or one for imaginary numbers would probably fall under the Arithmetic Error exception class.

To make a custom exception, simply inherit the base exception and define what it will do.

**Python modules**

Python allows us to store our code in files (also called modules). This is very useful for more serious programming, where we do not want to retype a long function definition from the very beginning just to change one mistake. In doing this, we are essentially defining our own modules, just like the modules defined already in the Python library.

To support this, Python has a way to put definitions in a file and use them in a script or in an interactive instance of the interpreter. Such a file is called a module; definitions from a module can be imported into other modules or into the main module.

**Testing code**

As indicated above, code is usually developed in a file using an editor.

To test the code, import it into a Python session and try to run it.

Usually there is an error, so you go back to the file, make a correction, and test again.

This process is repeated until you are satisfied that the code works. T

he entire process is known as the development cycle.

There are two types of errors that you will encounter. Syntax errors occur when the form of some command is invalid.

This happens when you make typing errors such as misspellings, or call something by the wrong name, and for many other reasons. Python will always give an error message for a syntax error.

## Functions in Python

## It is possible, and very useful, to define our own functions in Python. Generally speaking, if you need to do a calculation only once, then use the interpreter. But when you or others have need to perform a certain type of calculation many times, then define a function.

## You use functions in programming to bundle a set of instructions that you want to use repeatedly or that, because of their complexity, are better self-contained in a sub-program and called when needed. That means that a function is a piece of code written to carry out a specified task.

## To carry out that specific task, the function might or might not need multiple inputs. When the task is carred out, the function can or can not return one or more values.There are three types of functions in python:

## help() ,min() ,print().

## Python Namespace

Generally speaking, a **namespace** (sometimes also called a context) is a naming system for making names unique to avoid ambiguity. Everybody knows a namespacing system from daily life, i.e. the naming of people in firstname and familiy name (surname).

An example is a network: each network device (workstation, server, printer, ...) needs a unique name and address. Yet another example is the directory structure of file systems.

The same file name can be used in different directories, the files can be uniquely accessed via the pathnames.   
Many programming languages use namespaces or contexts for identifiers. An identifier defined in a namespace is associated with that namespace.

This way, the same identifier can be independently defined in multiple namespaces. (Like the same file names in different directories) Programming languages, which support namespaces, may have different rules that determine to which namespace an identifier belongs.

Namespaces in Python are implemented as Python dictionaries, this means it is a mapping from names (keys) to objects (values). The user doesn't have to know this to write a Python program and when using namespaces.

Some namespaces in Python:

* **global names** of a module
* **local names** in a function or method invocation
* **built-in names**: this namespace contains built-in functions (e.g. abs(), cmp(), ...) and built-in exception names

**Garbage Collection**

Garbage Collector exposes the underlying memory management mechanism of Python, the automatic garbage collector. The module includes functions for controlling how the collector operates and to examine the objects known to the system, either pending collection or stuck in reference cycles and unable to be freed.

**Python XML Parser**

XML is a portable, open source language that allows programmers to develop applications that can be read by other applications, regardless of operating system and/or developmental language.

What is XML? The Extensible Markup Language XML is a markup language much like HTML or SGML.

This is recommended by the World Wide Web Consortium and available as an open standard.

XML is extremely useful for keeping track of small to medium amounts of data without requiring a SQL-based backbone.

XML Parser Architectures and APIs The Python standard library provides a minimal but useful set of interfaces to work with XML.

The two most basic and broadly used APIs to XML data are the SAX and DOM interfaces.

Simple API for XML SAX : Here, you register callbacks for events of interest and then let the parser proceed through the document.

This is useful when your documents are large or you have memory limitations, it parses the file as it reads it from disk and the entire file is never stored in memory.

Document Object Model DOM API : This is a World Wide Web Consortium recommendation wherein the entire file is read into memory and stored in a hierarchical tree − based form to represent all the features of an XML document.

SAX obviously cannot process information as fast as DOM can when working with large files. On the other hand, using DOM exclusively can really kill your resources, especially if used on a lot of small files.

SAX is read-only, while DOM allows changes to the XML file. Since these two different APIs literally complement each other, there is no reason why you cannot use them both for large projects.

**Python Web Frameworks**

A web framework is a code library that makes a developer's life easier when building reliable, scalable and maintainable web applications.

## Why are web frameworks useful?

Web frameworks encapsulate what developers have learned over the past twenty years while programming sites and applications for the web. Frameworks make it easier to reuse code for common HTTP operations and to structure projects so other developers with knowledge of the framework can quickly build and maintain the application.

Common web framework functionality

Frameworks provide functionality in their code or through extensions to perform common operations required to run web applications. These common operations include:

1. URL routing
2. HTML, XML, JSON, and other output format templating
3. Database manipulation
4. Security against Cross-site request forgery (CSRF) and other attacks
5. Session storage and retrieval

Not all web frameworks include code for all of the above functionality. Frameworks fall on the spectrum from executing a single use case to providing every known web framework feature to every developer. Some frameworks take the "batteries-included" approach where everything possible comes bundled with the framework while others have a minimal core package that is amenable to extensions provided by other packages.

## Comparing web frameworks

There is also a repository called [compare-python-web-frameworks](https://github.com/mattmakai/compare-python-web-frameworks) where the same web application is being coded with varying Python web frameworks, templating engines and object.

## Web framework resources

* When you are learning how to use one or more web frameworks it's helpful to have an idea of what the code under the covers is doing.
* Frameworks is a really well done short video that explains how to choose between web frameworks. The author has some particular opinions about what should be in a framework. For the most part I agree although I've found sessions and database ORMs to be a helpful part of a framework when done well.
* what is a web framework? is an in-depth explanation of what web frameworks are and their relation to web servers.
* Django vs Flash vs Pyramid: Choosing a Python web framework contains background information and code comparisons for similar web applications built in these three big Python frameworks.
* This fascinating blog post takes a look at the  code complexity of several Python web frameworks by providing visualizations based on their code bases.
* Python’s web frameworks benchmarks  is a test of the responsiveness of a framework with encoding an object to JSON and returning it as a response as well as retrieving data from the database and rendering it in a template. There were no conclusive results but the output is fun to read about nonetheless.
* What web frameworks do you use and why are they awesome? is a language agnostic Reddit discussion on web frameworks. It's interesting to see what programmers in other languages like and dislike about their suite of web frameworks compared to the main Python frameworks.
* This user-voted question & answer site asked "What are the best general purpose Python web frameworks usable in production?". The votes aren't as important as the list of the many frameworks that are available to Python developers.

## Web frameworks learning checklist

1. Choose a major Python web framework (Django or Flask are recommended) and stick with it. When you're just starting it's best to learn one framework first instead of bouncing around trying to understand every framework.
2. Work through a detailed tutorial found within the resources links on the framework's page.
3. Study open source examples built with your framework of choice so you can take parts of those projects and reuse the code in your application.
4. Build the first simple iteration of your web application then go to the [deployment](https://www.fullstackpython.com/deployment.html)section to make it accessible on the web.

**Python-Data Base Communication**

Connector/Python provides a connect() call used to establish connections to the MySQL server. The following sections describe the permitted arguments for connect() and describe how to use option files that supply additional arguments.

A database is an organized collection of data. The data are typically organized to model aspects of reality in a way that supports processes requiring this information.

The term "database" can both refer to the data themselves or to the database management system. The Database management system is a software application for the interaction between users database itself.

Databases are popular for many applications, especially for use with web applications or customer-oriented programs

Users don't have to be human users. They can be other programs and applications as well. We will learn how Python or better a Python program can interact as a user of an SQLdatabase.   
  
This is an introduction into using SQLite and MySQL from Python.

The Python standard for database interfaces is the Python DB-API, which is used by Python's database interfaces.

The DB-API has been defined as a common interface, which can be used to access relational databases.

In other words, the code in Python for communicating with a database should be the same, regardless of the database and the database module used. Even though we use lots of SQL examples, this is not an introduction into SQL but a tutorial on the Python interface.

SQLite is a simple relational database system, which saves its data in regular data files or even in the internal memory of the computer, i.e. the RAM.

It was developped for embedded applications, like Mozilla-Firefox (Bookmarks), Symbian OS or Android.

SQLITE is "quite" fast, even though it uses a simple file. It can be used for large databases as well.

If you want to use SQLite, you have to import the module sqlite3. To use a database, you have to create first a Connection object.

The connection object will represent the database. The argument of connection - in the following example "companys.db" - functions both as the name of the file, where the data will be stored, and as the name of the database. If a file with this name exists, it will be opened.

It has to be a SQLite database file of course! In the following example, we will open a database called company.

MySQL Connector/Python enables Python programs to access MySQL databases, using an API that is compliant with the Python Database API Specification v2.0 (PEP 249). It is written in pure Python and does not have any dependencies except for the Python Standard Library.

For notes detailing the changes in each release of Connector/Python, see MySQL Connector/Python Release Notes.

MySQL Connector/Python includes support for:

* Almost all features provided by MySQL Server up to and including MySQL Server version 5.7.
* Converting parameter values back and forth between Python and MySQL data types, for example Python datetime and MySQL DATETIME. You can turn automatic conversion on for convenience, or off for optimal performance.
* All MySQL extensions to standard SQL syntax.
* Protocol compression, which enables compressing the data stream between the client and server.
* Connections using TCP/IP sockets and on Unix using Unix sockets.
* Secure TCP/IP connections using SSL.
* Self-contained driver. Connector/Python does not require the MySQL client library or any Python modules outside the standard library

**6.1. INTRODUCTION**

Software design sits at the technical kernel of the software engineering process and is applied regardless of the development paradigm and area of application. Design is the first step in the development phase for any engineered product or system. The designer’s goal is to produce a model or representation of an entity that will later be built. Beginning, once system requirement have been specified and analyzed, system design is the first of the three technical activities -design, code and test that is required to build and verify software.

The importance can be stated with a single word “Quality”. Design is the place where quality is fostered in software development. Design provides us with representations of software that can assess for quality. Design is the only way that we can accurately translate a customer’s view into a finished software product or system. Software design serves as a foundation for all the software engineering steps that follow. Without a strong design we risk building an unstable system – one that will be difficult to test, one whose quality cannot be assessed until the last stage. The purpose of the design phase is to plan a solution of the problem specified by the requirement document. This phase is the first step in moving from the problem domain to the solution domain. In other words, starting with what is needed, design takes us toward how to satisfy the needs. The design of a system is perhaps the most critical factor affection the quality of the software; it has a major impact on the later phase, particularly testing, maintenance. The output of this phase is the design document. This document is similar to a blueprint for the solution and is used later during implementation, testing and maintenance. The design activity is often divided into two separate phases System Design and Detailed Design.

System Design also called top-level design aims to identify the modules that should be in the system, the specifications of these modules, and how they interact with each other to produce the desired results. At the end of the system design all the major data structures, file formats, output formats, and the major modules in the system and their specifications are decided.

During, Detailed Design, the internal logic of each of the modules specified in system design is decided. During this phase, the details of the data of a module is usually specified in a high-level design description language, which is independent of the target language in which the software will eventually be implemented.

In system design the focus is on identifying the modules, whereas during detailed design the focus is on designing the logic for each of the modules. In other works, in system design the attention is on what components are needed, while in detailed design how the components can be implemented in software is the issue.

Design is concerned with identifying software components specifying relationships among components. Specifying software structure and providing blue print for the document phase. Modularity is one of the desirable properties of large systems. It implies that the system is divided into several parts. In such a manner, the interaction between parts is minimal clearly specified.

During the system design activities, Developers bridge the gap between the requirements specification, produced during requirements elicitation and analysis, and the system that is delivered to the user.

Design is the place where the quality is fostered in development. Software design is a process through which requirements are translated into a representation of software.

**Data Flow Diagrams:**

A graphical tool used to describe and analyze the moment of data through a system manual or automated including the process, stores of data, and delays in the system. Data Flow Diagrams are the central tool and the basis from which other components are developed. The transformation of data from input to output, through processes, may be described logically and independently of the physical components associated with the system. The DFD is also know as a data flow graph or a bubble chart.

DFDs are the model of the proposed system. They clearly should show the requirements on which the new system should be built. Later during design activity this is taken as the basis for drawing the system’s structure charts. The Basic Notation used to create a DFD’s are as follows:

**1. Dataflow:** Data move in a specific direction from an origin to a destination.

**2. Process:** People, procedures, or devices that use or produce (Transform) Data. The physical component is not identified.

**3. Source:** External sources or destination of data, which may be People, programs, organizations or other entities.

**4. Data Store:** Here data are stored or referenced by a process in the System.

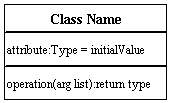
**What is a UML Class Diagram?**

Class diagrams are the backbone of almost every object-oriented method including UML. They describe the static structure of a system.

**Basic Class Diagram Symbols and Notations**

Classes represent an abstraction of entities with common characteristics. Associations represent the relationships between classes.

Illustrate classes with rectangles divided into compartments. Place the name of the class in the first partition (centered, bolded, and capitalized), list the attributes in the second partition, and write operations into the third.



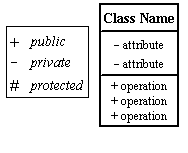
**Active Class**

Active classes initiate and control the flow of activity, while passive classes store data and serve other classes. Illustrate active classes with a thicker border.



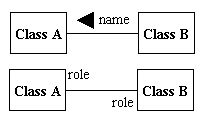
**Visibility**

Use visibility markers to signify who can access the information contained within a class. Private visibility hides information from anything outside the class partition. Public visibility allows all other classes to view the marked information. Protected visibility allows child classes to access information they inherited from a parent class. [.](http://www.smartdraw.com/resources/tutorials/Text-and-Tables)



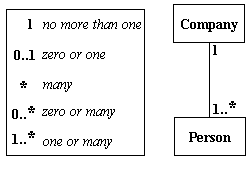
**Associations**

Associations represent static relationships between classes. Place association names above, on, or below the association line. Use a filled arrow to indicate the direction of the relationship. Place roles near the end of an association. Roles represent the way the two classes see each other.  
***Note:*** It's uncommon to name both the association and the class roles.



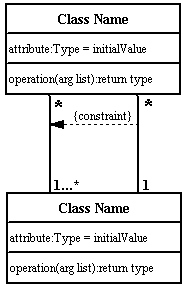
**Multiplicity (Cardinality)**

Place multiplicity notations near the ends of an association. These symbols indicate the number of instances of one class linked to one instance of the other class. For example, one company will have one or more employees, but each employee works for one company only.



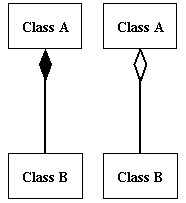
**Constraint**

Place constraints inside curly braces {}.

http://wc1.smartdraw.com/resources/tutorials/images/uml_constraint.gif*Simple Constraint* 

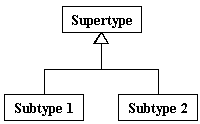
**Composition and Aggregation**

Composition is a special type of aggregation that denotes a strong ownership between Class A, the whole, and Class B, its part. Illustrate **composition** with a filled diamond. Use a hollow diamond to represent a simple **aggregation** relationship, in which the "whole" class plays a more important role than the "part" class, but the two classes are not dependent on each other. The diamond end in both a composition and aggregation relationship points toward the "whole" class or the aggregate



**Generalization**

Generalization is another name for inheritance or an "is a" relationship. It refers to a relationship between two classes where one class is a specialized version of another. For example, Honda is a type of car. So the class Honda would have a generalization relationship with the class car.



In real life coding examples, the difference between inheritance and aggregation can be confusing. If you have an aggregation relationship, the aggregate (the whole) can access only the PUBLIC functions of the part class. On the other hand, inheritance allows the inheriting class to access both the PUBLIC and PROTECTED functions of the super class.

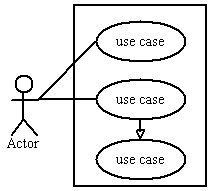
**What is a UML Use Case Diagram?**

Use case diagrams model the functionality of a system using actors and use cases. Use cases are services or functions provided by the system to its users.

**Basic Use Case Diagram Symbols and Notations**

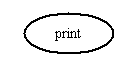
**System**

Draw your system's boundaries using a rectangle that contains use cases. Place actors outside the system's boundaries.



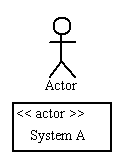
**Use Case**

Draw use cases using ovals. Label with ovals with verbs that represent the system's functions.



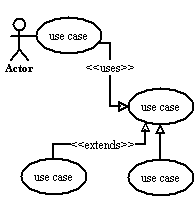
**Actors**

Actors are the users of a system. When one system is the actor of another system, label the actor system with the actor stereotype.



**Relationships**

Illustrate relationships between an actor and a use case with a simple line. For relationships among use cases, use arrows labeled either "uses" or "extends." A "uses" relationship indicates that one use case is needed by another in order to perform a task. An "extends" relationship indicates alternative options under a certain use case.



**Sequence Diagram**

Sequence diagrams describe interactions among classes in terms of an exchange of messages over time.

**Basic Sequence Diagram Symbols and Notations**

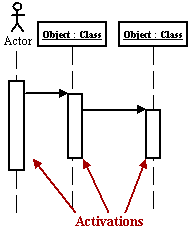
**Class roles**

Class roles describe the way an object will behave in context. Use the UML object symbol to illustrate class roles, but don't list object attributes.

Class roles

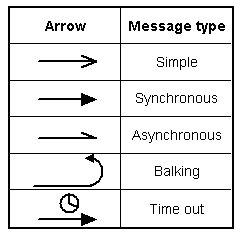
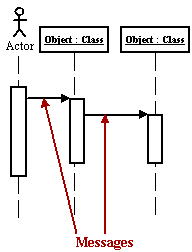
**Activation**

Activation boxes represent the time an object needs to complete a task.



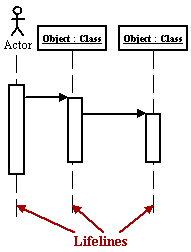
**Messages**

Messages are arrows that represent communication between objects. Use half-arrowed lines to represent asynchronous messages. Asynchronous messages are sent from an object that will not wait for a response from the receiver before continuing its tasks.

  
*Various message types for Sequence and Collaboration diagrams*

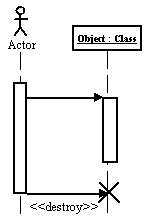
**Lifelines**

Lifelines are vertical dashed lines that indicate the object's presence over time.



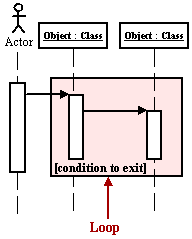
**Destroying Objects**

Objects can be terminated early using an arrow labeled "<< destroy >>" that points to an X.



**Loops**

A repetition or loop within a sequence diagram is depicted as a rectangle. Place the condition for exiting the loop at the bottom left corner in square brackets [ ].



**Collaboration Diagram**

A collaboration diagram describes interactions among objects in terms of sequenced messages. Collaboration diagrams represent a combination of information taken from class, sequence, and use case diagrams describing both the static structure and dynamic behavior of a system.

**Basic Collaboration Diagram Symbols and Notations**

**Class roles**

Class roles describe how objects behave. Use the UML object symbol to illustrate class roles, but don't list object attributes.

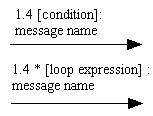
Class roles

**Association roles**

Association roles describe how an association will behave given a particular situation. You can draw association roles using simple lines labeled with stereotypes.  
Association roles

**Messages**

Unlike sequence diagrams, collaboration diagrams do not have an explicit way to denote time and instead number messages in order of execution. Sequence numbering can become nested using the Dewey decimal system. For example, nested messages under the first message are labeled 1.1, 1.2, 1.3, and so on. The a condition for a message is usually placed in square brackets immediately following the sequence number. Use a \* after the sequence number to indicate a loop.  
[Learn how to add arrows to your lines.](http://www.smartdraw.com/resources/tutorials/Lines)



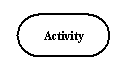
**Activity Diagram**

An activity diagram illustrates the dynamic nature of a system by modeling the flow of control from activity to activity. An activity represents an operation on some class in the system that results in a change in the state of the system. Typically, activity diagrams are used to model workflow or business processes and internal operation. Because an activity diagram is a special kind of state chart diagram, it uses some of the same modeling conventions.

**Basic Activity Diagram Symbols and Notations**

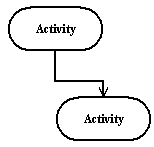
**Action states**

Action states represent the non interruptible actions of objects. You can draw an action state in Smart Draw using a rectangle with rounded corners.



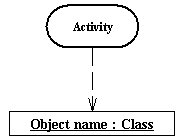
**Action Flow**

Action flow arrows illustrate the relationships among action states.



**Object Flow**

Object flow refers to the creation and modification of objects by activities. An object flow arrow from an action to an object means that the action creates or influences the object. An object flow arrow from an object to an action indicates that the action state uses the object.



**Initial State**

A filled circle followed by an arrow represents the initial action state.

Initial State

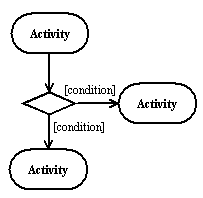
**Final State**

An arrow pointing to a filled circle nested inside another circle represents the final action state.

Final State

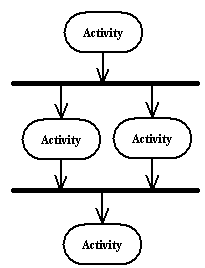
**Branching**

A diamond represents a decision with alternate paths. The outgoing alternates should be labeled with a condition or guard expression. You can also label one of the paths "else."



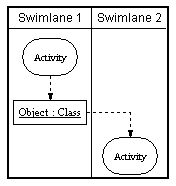
**Synchronization**

A synchronization bar helps illustrate parallel transitions. Synchronization is also called forking and joining.



**Swimlanes**

Swimlanes group related activities into one column.

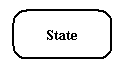
**State chart Diagram**

A state chart diagram shows the behavior of classes in response to external stimuli. This diagram models the dynamic flow of control from state to state within a system.

**Basic State chart Diagram Symbols and Notations**

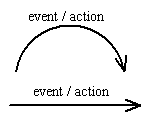
**States**

States represent situations during the life of an object. You can easily illustrate a state in Smart Draw by using a rectangle with rounded corners.



**Transition**

A solid arrow represents the path between different states of an object. Label the transition with the event that triggered it and the action that results from it.



**Initial State**

A filled circle followed by an arrow represents the object's initial state.

Initial State

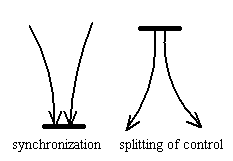
**Final State**

An arrow pointing to a filled circle nested inside another circle represents the object's final state.

Final State

**Synchronization and Splitting of Control**

A short heavy bar with two transitions entering it represents a synchronization of control. A short heavy bar with two transitions leaving it represents a splitting of control that creates multiple states.



**STATE CHART DIAGRAM:**

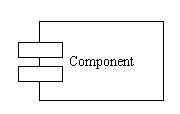
**What is a UML Component Diagram?**

A component diagram describes the organization of the physical components in a system.

**Basic Component Diagram Symbols and Notations**

**Component**

A component is a physical building block of the system. It is represented as a rectangle with tabs.  
[Learn how to resize grouped objects like components.](http://www.smartdraw.com/resources/tutorials/Objects)



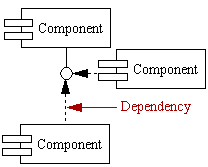
**Interface**

An interface describes a group of operations used or created by components.

Interface

**Dependencies**

Draw dependencies among components using dashed arrows.  
[Learn about line styles in SmartDraw.](http://www.smartdraw.com/resources/tutorials/Lines)



**COMPONENT DIAGRAM:**

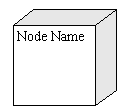
**What is a UML Deployment Diagram?**

Deployment diagrams depict the physical resources in a system including nodes, components, and connections.

**Basic Deployment Diagram Symbols and Notations**

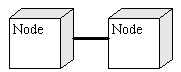
**Component**

A node is a physical resource that executes code components.  
[Learn how to resize grouped objects like nodes.](http://www.smartdraw.com/resources/tutorials/Objects)



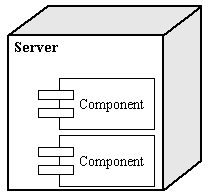
**Association**

Association refers to a physical connection between nodes, such as Ethernet.  
[Learn how to connect two nodes.](http://www.smartdraw.com/resources/tutorials/Lines)

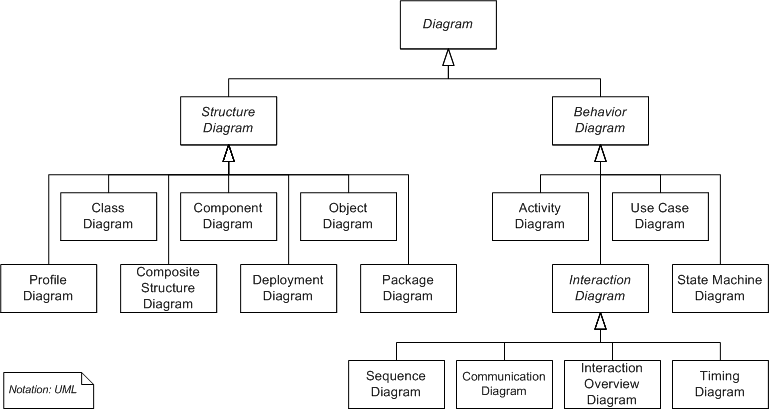


**Components and Nodes**

Place components inside the node that deploys them.



UML Diagrams Overview



UML combines best techniques from data modeling (entity relationship diagrams), business modeling (work flows), object modeling, and component modeling. It can be used with all processes, throughout the software development life cycle, and across different implementation technologies. UML has synthesized the notations of the Booch method, the Object-modeling technique (OMT) and Object-oriented software engineering (OOSE) by fusing them into a single, common and widely usable modeling language. UML aims to be a standard modeling language which can model concurrent and distributed systems.

UMLS

ON

CREDIT FRAD DETECTIONS

**Class diagram:**



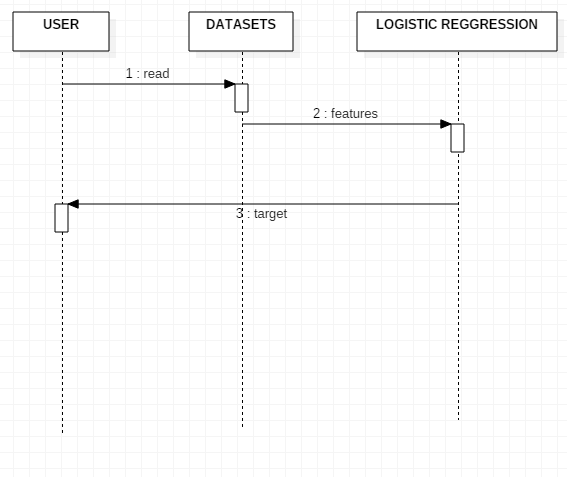
**Activity diagram:**

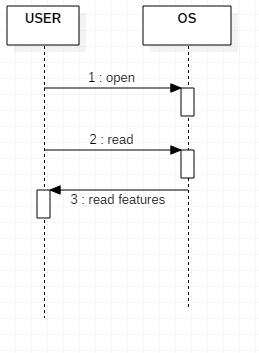
Activity diagrams are graphical representations of Workflows of stepwise activities and actions with support for choice, iteration and concurrency.In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.



**Sequence diagram:**

A **sequence diagram** in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. A sequence diagram shows, as parallel vertical lines ("lifelines"), different processes or objects that live simultaneously, and, as horizontal arrows, the messages exchanged between them, in the order in which they occur. This allows the specification of simple runtime scenarios in a graphical manner.





**Component diagram**



**Deployement diagram:**



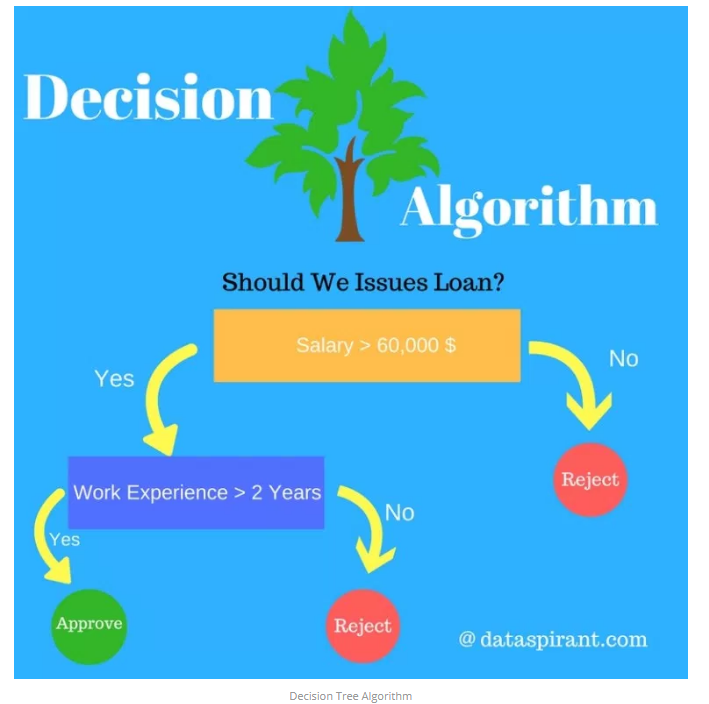
**Credit card Fraud Detection using Predictive Modeling**

**INTRODUCTION:**

Banks collect a lot of historical records corresponding to millions of customer’s transactions. They are credit card and debit card operations, but unfortunately, only a small portion, if any, is open access. Fraud detection is a critical problem affecting large financial companies that have increased due to the growth in credit card transactions [1]. The proposed method consists of the Predictive modeling and Logistic Regression. Now a day’s bank transactions as well as credit card frauds increased. One of the most target frauds are credit card fraud, the fraud can occur any type of credit products, such products are retail, home loan and personal loan. During the last few decades as technology has changed, dramatically the face of fraud also changed. To detect credit card fraud, data mining techniques Predictive modeling and Logistic Regression are used. In prediction model to predict the continuous valued functions. Credit card of CSV files will be analyzed to predict the outcome.

In this paper, we propose to detect credit card transaction using available data set and data mining techniques of predictive modeling, Decision tree, and Logistic Regression. Predictive modeling splits the data into two partitions 70% of testing and 30%of training check output class distribution to predict the outcome. The decision tree to get the result as a tree with root node describes the best predictor in the data, the combination of two or more branches is denoted by decision node (non leaf nodes) and each branch represents a value for the attribute which is tested. The leaf node may be 1 in the case of fraud and 0 otherwise. Logistic regression or logistic model is a regression model, where the dependent variable is categorical of a linear generalized model.The rest of the paper is organized as explained. Section II describes fraud detection methods. Section III explains Dataset description for credit card transaction. Section IV consists of experimental results of fraud detection methods, and finally, the conclusion of this work included in section V.

# DECISION TREE ALGORITHM:



# Introduction to Decision Tree Algorithm

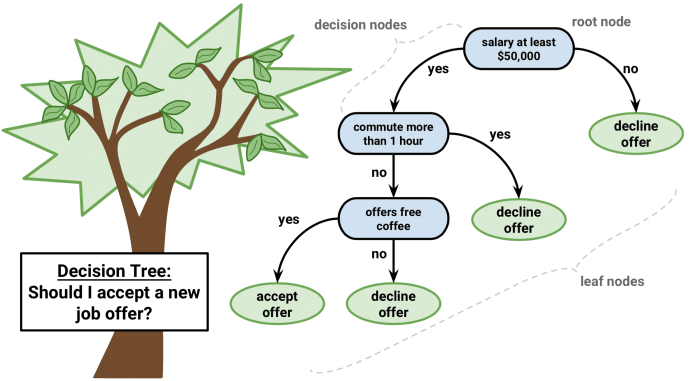
Decision Tree algorithm belongs to the family of [supervised learning algorithms](https://dataaspirant.com/2014/09/19/supervised-and-unsupervised-learning/). Unlike other supervised learning algorithms, decision tree algorithm can be used for solving [**regression and classification**](https://dataaspirant.com/2014/09/27/classification-and-prediction/)**problems** too.

The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by **learning decision rules** inferred from prior data(training data).

The understanding level of Decision Trees algorithm is so easy compared with other classification algorithms. The decision tree algorithm tries to solve the problem, by using tree representation. Each **internal node** of the tree corresponds to an attribute, and each **leaf node** corresponds to a class label.

### ****Decision Tree Algorithm Pseudocode****

1. Place the best attribute of the dataset at the **root** of the tree.
2. Split the training set into **subsets**. Subsets should be made in such a way that each subset contains data with the same value for an attribute.
3. Repeat step 1 and step 2 on each subset until you find **leaf nodes** in all the branches of the tree.



Decision Tree classifier, **Image credit:** www.packtpub.com

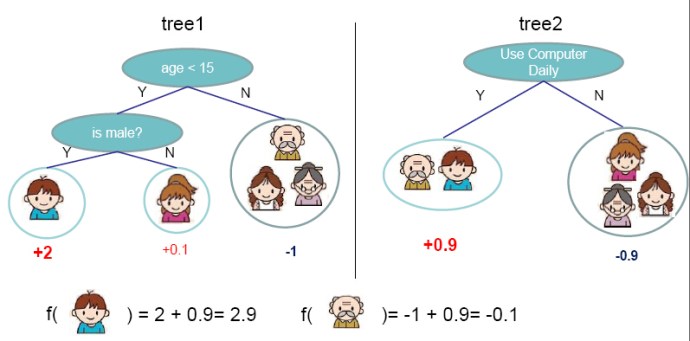
In decision trees, for predicting a class label for a record we start from the **root** of the tree. We compare the values of the root attribute with record’s attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

We continue comparing our record’s attribute values with other **internal nodes** of the tree until we reach **a leaf node** with predicted class value. As we know how the modeled decision tree can be used to predict the target class or the value. Now let’s understanding how we can create the decision tree model.

### Assumptions while creating Decision Tree

The below are the some of the assumptions we make while using Decision tree:

* At the beginning, the whole training set is considered as the **root.**
* Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.
* Records are **distributed recursively** on the basis of attribute values.
* Order to placing attributes as root or internal node of the tree is done by using some statistical approach.



Decision tree model example Image Credit: **http://zhanpengfang.github.io/**

Decision Trees follow **Sum of Product (SOP)** representation. For the above images, you can see how we can predict **can** **we accept the new job offer?  and Use computer daily?**from traversing for the root node to the leaf node.

It’s a sum of product representation. The Sum of product(SOP) is also known as **Disjunctive Normal Form**. For a class, every branch from the root of the tree to a leaf node having the same class is a conjunction(product) of values, different branches ending in that class form a disjunction(sum).

The primary challenge in the decision tree implementation is to identify which attributes do we need to consider as the root node and each level. Handling this is know the attributes selection. We have different attributes selection measure to identify the attribute which can be considered as the root note at each level.

**The popular attribute selection measures:**

* Information gain
* Gini index

### Attributes Selection

If dataset consists of **“n”** attributes then deciding which attribute to place at the root or at different levels of the tree as internal nodes is a complicated step. By just randomly selecting any node to be the root can’t solve the issue. If we follow a random approach, it may give us bad results with low accuracy.

For solving this attribute selection problem, researchers worked and devised some solutions. They suggested using some criterion like **information gain, gini index,** etc. These criterions will calculate values for every attribute. The values are sorted, and attributes are placed in the tree by following the order i.e, the attribute with a high value(in case of information gain) is placed at the root.

While using information Gain as a criterion, we assume attributes to be categorical, and for gini index, attributes are assumed to be continuous.

### Information Gain

By using information gain as a criterion, we try to estimate the information contained by each attribute. We are going to use some points deducted from [information theory](https://en.wikipedia.org/wiki/Information_theory).  
To measure the randomness or uncertainty of a random variable X is defined by **Entropy**.

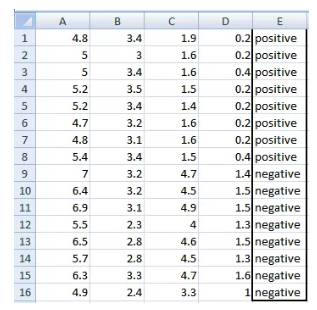
For a binary classification problem with only two classes, positive and negative class.

* If all examples are positive or all are negative then entropy will be zero i.e, low.
* If half of the records are of positive class and half are of negative class then entropy is one i.e, high.

By calculating **entropy measure** of each attribute we can calculate their **information gain**. Information Gain calculates the expected reduction in entropy due to sorting on the attribute. Information gain can be calculated.

To get a clear understanding of calculating **information gain & entropy**, we will try to implement it on a sample data.

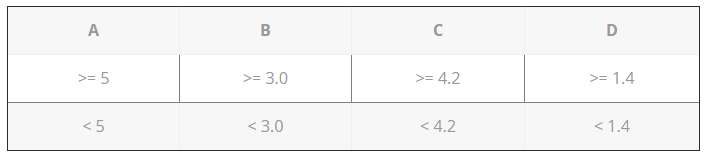
#### Example: Construct a Decision Tree by using “information gain” as a criterion



We are going to use this data sample. Let’s try to use information gain as a criterion. Here, we have 5 columns out of which 4 columns have continuous data and 5th column consists of class labels.

A, B, C, D attributes can be considered as predictors and E column class labels can be considered as a target variable. For constructing a decision tree from this data, we have to convert continuous data into categorical data.

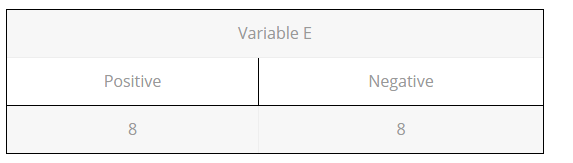
We have chosen some random values to categorize each attribute:



There are **2 steps for calculating information gain** for each attribute:

1. Calculate entropy of Target.
2. Entropy for every attribute A, B, C, D needs to be calculated. Using information gain formula we will subtract this entropy from the entropy of target. The result is Information Gain.

**The entropy of Target:** We have 8 records with negative class and 8 records with positive class. So, we can directly estimate the entropy of target as 1.



**Calculating entropy using formula:**

E(8,8) = -1\*( (p(+ve)\*log( p(+ve)) + (p(-ve)\*log( p(-ve)) )  
= -1\*( (8/16)\*log2(8/16)) + (8/16) \* log2(8/16) )  
= 1

#### Information gain for Var A

Var A has value >=5 for 12 records out of 16 and 4 records with value <5 value.

* For Var A >= 5 & class == positive: 5/12
* For Var A >= 5 & class == negative: 7/12
  + Entropy(5,7) = -1 \* ( (5/12)\*log2(5/12) + (7/12)\*log2(7/12)) = 0.9799
* For Var A <5 & class == positive: 3/4
* For Var A <5 & class == negative: 1/4
  + Entropy(3,1) =  -1 \* ( (3/4)\*log2(3/4) + (1/4)\*log2(1/4)) = 0.81128

Entropy(Target, A) = P(>=5) \* E(5,7) + P(<5) \* E(3,1)  
= (12/16) \* 0.9799 + (4/16) \* 0.81128 = 0.937745

\textrm{Information Gain(IG) = E(Target) - E(Target,A) = 1- 0.9337745 = 0.062255}  

#### Information gain for Var B

Var B has value >=3 for 12 records out of 16 and 4 records with value <5 value.

* For Var B >= 3 & class == positive: 8/12
* For Var B >= 3 & class == negative: 4/12
  + Entropy(8,4) = -1 \* ( (8/12)\*log2(8/12) + (4/12)\*log2(4/12)) = 0.39054
* For VarB <3 & class == positive: 0/4
* For Var B <3 & class == negative: 4/4
  + Entropy(0,4) =  -1 \* ( (0/4)\*log2(0/4) + (4/4)\*log2(4/4)) = 0

Entropy(Target, B) = P(>=3) \* E(8,4) + P(<3) \* E(0,4)  
= (12/16) \* 0.39054 + (4/16) \* 0 = 0.292905

\textrm{Information Gain(IG) = E(Target) - E(Target,B) = 1- 0.292905= 0.707095}  

#### Information gain for Var C

Var C has value >=4.2 for 6 records out of 16 and 10 records with value <4.2 value.

* For Var C >= 4.2 & class == positive: 0/6
* For Var C >= 4.2 & class == negative:  6/6
  + Entropy(0,6) = 0
* For VarC < 4.2 & class == positive: 8/10
* For Var C < 4.2 & class == negative: 2/10
  + Entropy(8,2) = 0.72193

Entropy(Target, C) = P(>=4.2) \* E(0,6) + P(< 4.2) \* E(8,2)  
= (6/16) \* 0 + (10/16) \* 0.72193 = 0.4512

\textrm{Information Gain(IG) = E(Target) - E(Target,C) = 1- 0.4512= 0.5488}  

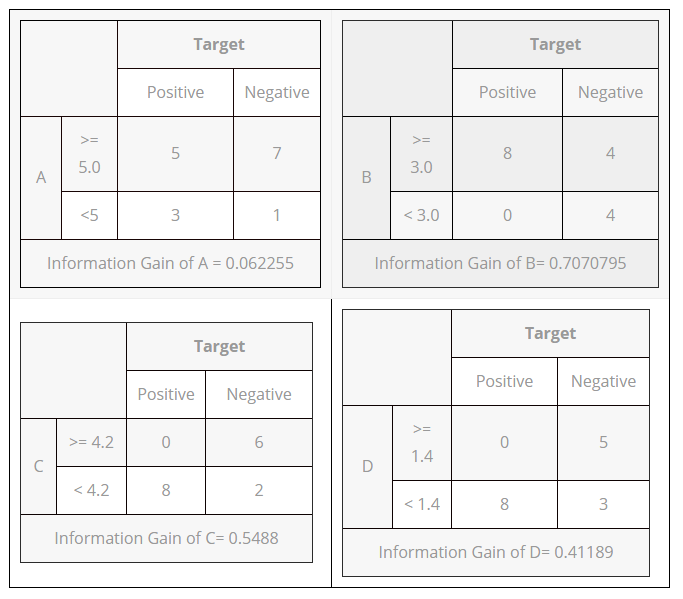
#### Information gain for Var D

Var D has value >=1.4 for 5 records out of 16 and 11 records with value <5 value.

* For Var D >= 1.4 & class == positive: 0/5
* For Var D >= 1.4 & class == negative: 5/5
  + Entropy(0,5) = 0
* For Var D < 1.4 & class == positive: 8/11
* For Var D < 14 & class == negative: 3/11
  + Entropy(8,3) =  -1 \* ( (8/11)\*log2(8/11) + (3/11)\*log2(3/11)) = 0.84532

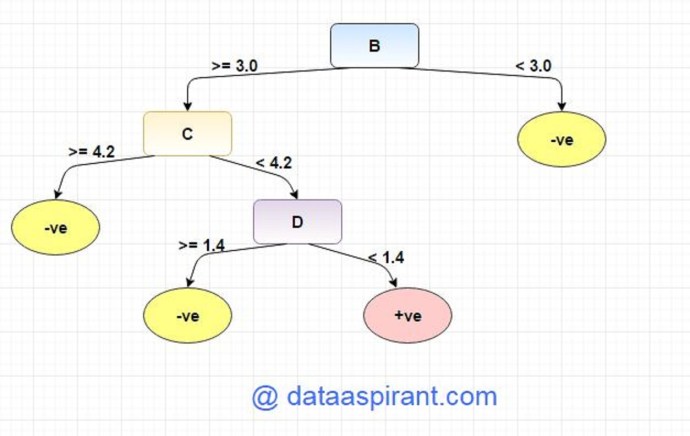
Entropy(Target, D) = P(>=1.4) \* E(0,5) + P(< 1.4) \* E(8,3)  
= 5/16 \* 0 + (11/16) \* 0.84532 = 0.5811575

\textrm{Information Gain(IG) = E(Target) - E(Target,D) = 1- 0.5811575 = 0.41189}  



From the above Information Gain calculations, we can build a decision tree. We should place the attributes on the tree according to their values.

An Attribute with better value than other should position as root and A branch with entropy 0 should be converted to a leaf node. A branch with entropy more than 0 needs further splitting.



### Decision Tree Algorithm Advantages and Disadvantages

#### Advantages:

1. Decision Trees are easy to explain. It results in a set of rules.
2. It follows the same approach as humans generally follow while making decisions.
3. Interpretation of a complex Decision Tree model can be simplified by its visualizations. Even a naive person can understand logic.
4. The Number of hyper-parameters to be tuned is almost null.

#### Disadvantages:

1. There is a high probability of overfitting in Decision Tree.
2. Generally, it gives low prediction accuracy for a dataset as compared to other machine learning algorithms.
3. Information gain in a decision tree with categorical variables gives a biased response for attributes with greater no. of categories.
4. Calculations can become complex when there are many class labels.

**Nearest neighbors:**

## Algorithm

[](https://en.wikipedia.org/wiki/File:KnnClassification.svg)

Example of *k*-NN classification. The test sample (green circle) should be classified either to the first class of blue squares or to the second class of red triangles. If *k = 3* (solid line circle) it is assigned to the second class because there are 2 triangles and only 1 square inside the inner circle. If *k = 5* (dashed line circle) it is assigned to the first class (3 squares vs. 2 triangles inside the outer circle).

The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the [feature vectors](https://en.wikipedia.org/wiki/Feature_vector) and class labels of the training samples.

In the classification phase, *k* is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label which is most frequent among the *k* training samples nearest to that query point.

A commonly used distance metric for [continuous variables](https://en.wikipedia.org/wiki/Continuous_variable) is [Euclidean distance](https://en.wikipedia.org/wiki/Euclidean_distance). For discrete variables, such as for text classification, another metric can be used, such as the overlap metric (or [Hamming distance](https://en.wikipedia.org/wiki/Hamming_distance)). In the context of gene expression microarray data, for example, *k*-NN has also been employed with correlation coefficients such as Pearson and Spearman.[[3]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-3) Often, the classification accuracy of *k*-NN can be improved significantly if the distance metric is learned with specialized algorithms such as [Large Margin Nearest Neighbor](https://en.wikipedia.org/wiki/Large_Margin_Nearest_Neighbor) or [Neighbourhood components analysis](https://en.wikipedia.org/wiki/Neighbourhood_components_analysis).

A drawback of the basic "majority voting" classification occurs when the class distribution is skewed. That is, examples of a more frequent class tend to dominate the prediction of the new example, because they tend to be common among the *k* nearest neighbors due to their large number.[[4]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-Coomans_Massart1982-4) One way to overcome this problem is to weight the classification, taking into account the distance from the test point to each of its *k* nearest neighbors. The class (or value, in regression problems) of each of the *k* nearest points is multiplied by a weight proportional to the inverse of the distance from that point to the test point. Another way to overcome skew is by abstraction in data representation. For example, in a [self-organizing map](https://en.wikipedia.org/wiki/Self-organizing_map) (SOM), each node is a representative (a center) of a cluster of similar points, regardless of their density in the original training data. *K*-NN can then be applied to the SOM.

## Parameter selection[[edit](https://en.wikipedia.org/w/index.php?title=K-nearest_neighbors_algorithm&action=edit&section=3)]

The best choice of *k* depends upon the data; generally, larger values of *k* reduces effect of the noise on the classification,[[5]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-5) but make boundaries between classes less distinct. A good *k* can be selected by various [heuristic](https://en.wikipedia.org/wiki/Heuristic_(computer_science)) techniques (see [hyperparameter optimization](https://en.wikipedia.org/wiki/Hyperparameter_optimization)). The special case where the class is predicted to be the class of the closest training sample (i.e. when *k* = 1) is called the nearest neighbor algorithm.

The accuracy of the *k*-NN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance. Much research effort has been put into [selecting](https://en.wikipedia.org/wiki/Feature_selection) or [scaling](https://en.wikipedia.org/wiki/Feature_scaling) features to improve classification. A particularly popular[[*citation needed*](https://en.wikipedia.org/wiki/Wikipedia:Citation_needed)] approach is the use of [evolutionary algorithms](https://en.wikipedia.org/wiki/Evolutionary_algorithm) to optimize feature scaling.[[6]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-6) Another popular approach is to scale features by the [mutual information](https://en.wikipedia.org/wiki/Mutual_information) of the training data with the training classes.[[*citation needed*](https://en.wikipedia.org/wiki/Wikipedia:Citation_needed)]

In binary (two class) classification problems, it is helpful to choose *k* to be an odd number as this avoids tied votes. One popular way of choosing the empirically optimal *k* in this setting is via bootstrap method.[[7]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-HPS2008-7)

## The 1-nearest neighbor classifier

The most intuitive nearest neighbour type classifier is the one nearest neighbour classifier that assigns a point *x* to the class of its closest neighbour in the feature space, that is {\displaystyle C\_{n}^{1nn}(x)=Y\_{(1)}}.

As the size of training data set approaches infinity, the one nearest neighbour classifier guarantees an error rate of no worse than twice the [Bayes error rate](https://en.wikipedia.org/wiki/Bayes_error_rate) (the minimum achievable error rate given the distribution of the data).

## The weighted nearest neighbour classifier

The *k*-nearest neighbour classifier can be viewed as assigning the *k* nearest neighbours a weight {\displaystyle 1/k} and all others *0* weight. This can be generalised to weighted nearest neighbour classifiers. That is, where the *i*th nearest neighbour is assigned a weight {\displaystyle w\_{ni}}, with {\displaystyle \sum \_{i=1}^{n}w\_{ni}=1}. An analogous result on the strong consistency of weighted nearest neighbour classifiers also holds.[[8]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-Stone-8)

Let {\displaystyle C\_{n}^{wnn}} denote the weighted nearest classifier with weights {\displaystyle \{w\_{ni}\}\_{i=1}^{n}}. Subject to regularity conditions on to class distributions the excess risk has the following asymptotic expansion[[9]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-Samworth12-9)

{\displaystyle {\mathcal {R}}\_{\mathcal {R}}(C\_{n}^{wnn})-{\mathcal {R}}\_{\mathcal {R}}(C^{Bayes})=\left(B\_{1}s\_{n}^{2}+B\_{2}t\_{n}^{2}\right)\{1+o(1)\},}

for constants {\displaystyle B\_{1}} and {\displaystyle B\_{2}} where {\displaystyle s\_{n}^{2}=\sum \_{i=1}^{n}w\_{ni}^{2}} and {\displaystyle t\_{n}=n^{-2/d}\sum \_{i=1}^{n}w\_{ni}\{i^{1+2/d}-(i-1)^{1+2/d}\}}.

The optimal weighting scheme {\displaystyle \{w\_{ni}^{\*}\}\_{i=1}^{n}}, that balances the two terms in the display above, is given as follows: set {\displaystyle k^{\*}=\lfloor Bn^{\frac {4}{d+4}}\rfloor },

{\displaystyle w\_{ni}^{\*}={\frac {1}{k^{\*}}}\left[1+{\frac {d}{2}}-{\frac {d}{2{k^{\*}}^{2/d}}}\{i^{1+2/d}-(i-1)^{1+2/d}\}\right]} for {\displaystyle i=1,2,\dots ,k^{\*}} and

{\displaystyle w\_{ni}^{\*}=0} for {\displaystyle i=k^{\*}+1,\dots ,n}.

With optimal weights the dominant term in the asymptotic expansion of the excess risk is {\displaystyle {\mathcal {O}}(n^{-{\frac {4}{d+4}}})}. Similar results are true when using a [bagged nearest neighbour classifier](https://en.wikipedia.org/wiki/Bootstrap_aggregating).

## Properties

*k*-NN is a special case of a [variable-bandwidth, kernel density "balloon" estimator](https://en.wikipedia.org/wiki/Variable_kernel_density_estimation) with a uniform [kernel](https://en.wikipedia.org/wiki/Kernel_(statistics)).[[10]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-Terrell_Scott1992-10) [[11]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-Mills2010-11)

The naive version of the algorithm is easy to implement by computing the distances from the test example to all stored examples, but it is computationally intensive for large training sets. Using an approximate [nearest neighbor search](https://en.wikipedia.org/wiki/Nearest_neighbor_search) algorithm makes *k-*NN computationally tractable even for large data sets. Many nearest neighbor search algorithms have been proposed over the years; these generally seek to reduce the number of distance evaluations actually performed.

*k-*NN has some strong [consistency](https://en.wikipedia.org/wiki/Consistency_(statistics)) results. As the amount of data approaches infinity, the two-class *k-*NN algorithm is guaranteed to yield an error rate no worse than twice the [Bayes error rate](https://en.wikipedia.org/wiki/Bayes_error_rate) (the minimum achievable error rate given the distribution of the data).[[12]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-12) Various improvements to the *k*-NN speed are possible by using proximity graphs.[[13]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-13)

For multi-class *k-*NN classification, [Cover](https://en.wikipedia.org/wiki/Thomas_M._Cover) and [Hart](https://en.wikipedia.org/wiki/Peter_E._Hart) (1967) prove an upper bound error rate of

{\displaystyle R^{\*}\ \leq \ R\_{k\mathrm {NN} }\ \leq \ R^{\*}\left(2-{\frac {MR^{\*}}{M-1}}\right)}

where {\displaystyle R^{\*}}is the Bayes error rate (which is the minimal error rate possible), {\displaystyle R\_{kNN}} is the *k-*NN error rate, and *M* is the number of classes in the problem. For {\displaystyle M=2} and as the Bayesian error rate {\displaystyle R^{\*}} approaches zero, this limit reduces to "not more than twice the Bayesian error rate".

## Error rates[[edit](https://en.wikipedia.org/w/index.php?title=K-nearest_neighbors_algorithm&action=edit&section=7)]

There are many results on the error rate of the *k* nearest neighbour classifiers.[[14]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-PTPR-14) The *k*-nearest neighbour classifier is strongly (that is for any joint distribution on {\displaystyle (X,Y)}) [consistent](https://en.wikipedia.org/wiki/Bayes_classifier) provided {\displaystyle k:=k\_{n}} diverges and {\displaystyle k\_{n}/n} converges to zero as {\displaystyle n\to \infty }.

Let {\displaystyle C\_{n}^{knn}} denote the *k* nearest neighbour classifier based on a training set of size *n*. Under certain regularity conditions, the [excess risk](https://en.wikipedia.org/wiki/Bayes_classifier) yields the following asymptotic expansion[[9]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-Samworth12-9)

{\displaystyle {\mathcal {R}}\_{\mathcal {R}}(C\_{n}^{knn})-{\mathcal {R}}\_{\mathcal {R}}(C^{Bayes})=\left\{B\_{1}{\frac {1}{k}}+B\_{2}\left({\frac {k}{n}}\right)^{4/d}\right\}\{1+o(1)\},}

for some constants {\displaystyle B\_{1}} and {\displaystyle B\_{2}}.

The choice {\displaystyle k^{\*}=\lfloor Bn^{\frac {4}{d+4}}\rfloor } offers a trade off between the two terms in the above display, for which the {\displaystyle k^{\*}}-nearest neighbour error converges to the Bayes error at the optimal ([minimax](https://en.wikipedia.org/wiki/Minimax)) rate {\displaystyle {\mathcal {O}}(n^{-{\frac {4}{d+4}}})}.

## Metric learning

The K-nearest neighbor classification performance can often be significantly improved through (supervised) metric learning. Popular algorithms are [neighbourhood components analysis](https://en.wikipedia.org/wiki/Neighbourhood_components_analysis) and [large margin nearest neighbor](https://en.wikipedia.org/wiki/Large_margin_nearest_neighbor). Supervised metric learning algorithms use the label information to learn a new [metric](https://en.wikipedia.org/wiki/Metric_(mathematics)) or [pseudo-metric](https://en.wikipedia.org/wiki/Pseudometric_space).

## Feature extraction[[edit](https://en.wikipedia.org/w/index.php?title=K-nearest_neighbors_algorithm&action=edit&section=9)]

When the input data to an algorithm is too large to be processed and it is suspected to be redundant (e.g. the same measurement in both feet and meters) then the input data will be transformed into a reduced representation set of features (also named features vector). Transforming the input data into the set of features is called [feature extraction](https://en.wikipedia.org/wiki/Feature_extraction). If the features extracted are carefully chosen it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input. Feature extraction is performed on raw data prior to applying *k*-NN algorithm on the transformed data in [feature space](https://en.wikipedia.org/wiki/Feature_space).

An example of a typical [computer vision](https://en.wikipedia.org/wiki/Computer_vision) computation pipeline for [face recognition](https://en.wikipedia.org/wiki/Facial_recognition_system) using *k*-NN including feature extraction and dimension reduction pre-processing steps (usually implemented with [OpenCV](https://en.wikipedia.org/wiki/OpenCV)):

1. [Haar](https://en.wikipedia.org/wiki/Haar_wavelet) face detection
2. [Mean-shift](https://en.wikipedia.org/wiki/Mean-shift) tracking analysis
3. [PCA](https://en.wikipedia.org/wiki/Principal_Component_Analysis) or [Fisher LDA](https://en.wikipedia.org/wiki/Linear_discriminant_analysis) projection into feature space, followed by *k*-NN classification

## Dimension reduction[[edit](https://en.wikipedia.org/w/index.php?title=K-nearest_neighbors_algorithm&action=edit&section=10)]

For high-dimensional data (e.g., with number of dimensions more than 10) [dimension reduction](https://en.wikipedia.org/wiki/Dimension_reduction) is usually performed prior to applying the *k*-NN algorithm in order to avoid the effects of the [curse of dimensionality](https://en.wikipedia.org/wiki/Curse_of_Dimensionality). [[15]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-15)

The curse of dimensionality in the *k*-NN context basically means that [Euclidean distance](https://en.wikipedia.org/wiki/Euclidean_distance) is unhelpful in high dimensions because all vectors are almost equidistant to the search query vector (imagine multiple points lying more or less on a circle with the query point at the center; the distance from the query to all data points in the search space is almost the same).

[Feature extraction](https://en.wikipedia.org/wiki/Feature_extraction) and dimension reduction can be combined in one step using [principal component analysis](https://en.wikipedia.org/wiki/Principal_Component_Analysis) (PCA), [linear discriminant analysis](https://en.wikipedia.org/wiki/Linear_discriminant_analysis)(LDA), or [canonical correlation analysis](https://en.wikipedia.org/wiki/Canonical_correlation) (CCA) techniques as a pre-processing step, followed by clustering by *k*-NN on [feature vectors](https://en.wikipedia.org/wiki/Feature_(machine_learning)) in reduced-dimension space. In [machine learning](https://en.wikipedia.org/wiki/Machine_learning) this process is also called low-dimensional [embedding](https://en.wikipedia.org/wiki/Embedding).[[16]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-16)

For very-high-dimensional datasets (e.g. when performing a similarity search on live video streams, DNA data or high-dimensional [time series](https://en.wikipedia.org/wiki/Time_series)) running a fast approximate *k*-NN search using [locality sensitive hashing](https://en.wikipedia.org/wiki/Locality_Sensitive_Hashing), "random projections",[[17]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-17) "sketches" [[18]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-18) or other high-dimensional similarity search techniques from the [VLDB](https://en.wikipedia.org/wiki/VLDB) toolbox might be the only feasible option.

## Decision boundary

Nearest neighbor rules in effect implicitly compute the [decision boundary](https://en.wikipedia.org/wiki/Decision_boundary). It is also possible to compute the decision boundary explicitly, and to do so efficiently, so that the computational complexity is a function of the boundary complexity.[[19]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-19)

## Data reduction

Data reduction is one of the most important problems for work with huge data sets. Usually, only some of the data points are needed for accurate classification. Those data are called the *prototypes* and can be found as follows:

1. Select the *class-outliers*, that is, training data that are classified incorrectly by *k*-NN (for a given *k*)
2. Separate the rest of the data into two sets: (i) the prototypes that are used for the classification decisions and (ii) the *absorbed points* that can be correctly classified by *k*-NN using prototypes. The absorbed points can then be removed from the training set.

### Selection of class-outliers

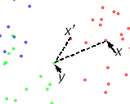
A training example surrounded by examples of other classes is called a class outlier. Causes of class outliers include:

* random error
* insufficient training examples of this class (an isolated example appears instead of a cluster)
* missing important features (the classes are separated in other dimensions which we do not know)
* too many training examples of other classes (unbalanced classes) that create a "hostile" background for the given small class

Class outliers with *k*-NN produce noise. They can be detected and separated for future analysis. Given two natural numbers, *k>r>0*, a training example is called a *(k,r)*NN class-outlier if its *k* nearest neighbors include more than *r* examples of other classes.

### CNN for data reduction

Condensed nearest neighbor (CNN, the [*Hart*](https://en.wikipedia.org/wiki/Peter_E._Hart)*algorithm*) is an algorithm designed to reduce the data set for *k*-NN classification.[[20]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-20) It selects the set of prototypes *U* from the training data, such that 1NN with *U* can classify the examples almost as accurately as 1NN does with the whole data set.

[](https://en.wikipedia.org/wiki/File:BorderRAtio.PNG)

Calculation of the border ratio.

[https://upload.wikimedia.org/wikipedia/commons/e/e7/PointsTypes.png](https://en.wikipedia.org/wiki/File:PointsTypes.png)

Three types of points: prototypes, class-outliers, and absorbed points.

Given a training set *X*, CNN works iteratively:

1. Scan all elements of *X*, looking for an element *x* whose nearest prototype from *U* has a different label than *x*.
2. Remove *x* from *X* and add it to *U*
3. Repeat the scan until no more prototypes are added to *U*.

Use *U* instead of *X* for classification. The examples that are not prototypes are called "absorbed" points.

It is efficient to scan the training examples in order of decreasing border ratio.[[21]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-MirkesKnn-21) The border ratio of a training example *x* is defined as

*a*(*x*) = ||*x'-y*||/||*x-y*||

where ||*x-y*|| is the distance to the closest example *y* having a different color than *x*, and ||*x'-y*|| is the distance from *y* to its closest example *x'* with the same label as *x*.

The border ratio is in the interval [0,1] because ||*x'-y*||never exceeds ||*x-y*||. This ordering gives preference to the borders of the classes for inclusion in the set of prototypes *U*. A point of a different label than *x* is called external to *x*. The calculation of the border ratio is illustrated by the figure on the right. The data points are labeled by colors: the initial point is *x* and its label is red. External points are blue and green. The closest to *x* external point is *y*. The closest to *y* red point is *x'* . The border ratio *a*(*x*) = ||*x'-y*|| / ||*x-y*||is the attribute of the initial point *x*.

Below is an illustration of CNN in a series of figures. There are three classes (red, green and blue). Fig. 1: initially there are 60 points in each class. Fig. 2 shows the 1NN classification map: each pixel is classified by 1NN using all the data. Fig. 3 shows the 5NN classification map. White areas correspond to the unclassified regions, where 5NN voting is tied (for example, if there are two green, two red and one blue points among 5 nearest neighbors). Fig. 4 shows the reduced data set. The crosses are the class-outliers selected by the (3,2)NN rule (all the three nearest neighbors of these instances belong to other classes); the squares are the prototypes, and the empty circles are the absorbed points. The left bottom corner shows the numbers of the class-outliers, prototypes and absorbed points for all three classes. The number of prototypes varies from 15% to 20% for different classes in this example. Fig. 5 shows that the 1NN classification map with the prototypes is very similar to that with the initial data set. The figures were produced using the Mirkes applet.[[21]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-MirkesKnn-21)

* CNN model reduction for k-NN classifiers
* [](https://en.wikipedia.org/wiki/File:Data3classes.png)

Fig. 1. The dataset.

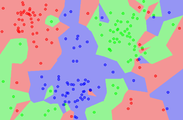
* [](https://en.wikipedia.org/wiki/File:Map1NN.png)

Fig. 2. The 1NN classification map.

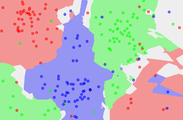
* [](https://en.wikipedia.org/wiki/File:Map5NN.png)

Fig. 3. The 5NN classification map.

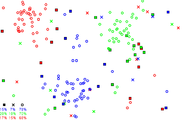
* [](https://en.wikipedia.org/wiki/File:ReducedDataSet.png)

Fig. 4. The CNN reduced dataset.

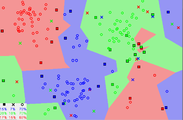
* [](https://en.wikipedia.org/wiki/File:Map1NNReducedDataSet.png)

Fig. 5. The 1NN classification map based on the CNN extracted prototypes.

FCNN[[22]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-22) (for Fast Condensed Nearest Neighbor) is a variant of CNN, which turns out to be one of the fastest data set reduction algorithms for k-NN classification.[[23]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-23)

## *k*-NN regression[[edit](https://en.wikipedia.org/w/index.php?title=K-nearest_neighbors_algorithm&action=edit&section=15)]

In *k*-NN regression, the *k*-NN algorithm is used for estimating continuous variables. One such algorithm uses a weighted average of the *k* nearest neighbors, weighted by the inverse of their distance. This algorithm works as follows:

1. Compute the Euclidean or [Mahalanobis distance](https://en.wikipedia.org/wiki/Mahalanobis_distance) from the query example to the labeled examples.
2. Order the labeled examples by increasing distance.
3. Find a heuristically optimal number *k* of nearest neighbors, based on [RMSE](https://en.wikipedia.org/wiki/RMSE). This is done using cross validation.
4. Calculate an inverse distance weighted average with the *k*-nearest multivariate neighbors.

## *k*-NN outlier

The distance to the *k*th nearest neighbor can also be seen as a local density estimate and thus is also a popular outlier score in [anomaly detection](https://en.wikipedia.org/wiki/Anomaly_detection). The larger the distance to the *k*-NN, the lower the local density, the more likely the query point is an outlier.[[24]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-24) To take into account the whole neighborhood of the query point, the average distance to the *k*-NN can be used.[[25]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-25) Although quite simple, this outlier model, along with another classic data mining method, [local outlier factor](https://en.wikipedia.org/wiki/Local_outlier_factor), works quite well also in comparison to more recent and more complex approaches, according to a large scale experimental analysis.[[26]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-CamposZimek2016-26)

## Validation of results

A [confusion matrix](https://en.wikipedia.org/wiki/Confusion_matrix) or "matching matrix" is often used as a tool to validate the accuracy of *k*-NN classification. More robust statistical methods such as [likelihood-ratio test](https://en.wikipedia.org/wiki/Likelihood-ratio_test) can also be applied.

Import modules:-

import numpy as np

#import sklearn python machine learning module

import sklearn as sk

#import pandas dataframes

import pandas as pd

#import matplotlib for plotting

import matplotlib.pyplot as plt

#import datasets and linear\_model from sklearn module

from sklearn import datasets, linear\_model

#import Polynomial features from sklearn module

from sklearn.preprocessing import PolynomialFeatures

#import train\_test\_split data classification

from sklearn.model\_selection import train\_test\_split

#import ConfusionMatrix from pandas\_ml

from pandas\_ml import ConfusionMatrix

# Loading the dataset:-

dataframe = pd.read\_csv('C:/Python27/creditcard.csv', low\_memory=False)

#dataframe.sample Returns a random sample of items from an axis of object.

#The frac keyword argument specifies the fraction of rows to return in the random sample, so frac=1 means return all rows (in random order).

# If you wish to shuffle your dataframe in-place and reset the index

dataframe = dataframe.sample(frac=1).reset\_index(drop=True)

#dataframe.head(n) returns a DataFrame holding the first n rows of dataframe.

dataframe.head()

print dataframe

# Checking the target classes:-

fraud\_class = dataframe.loc[dataframe['Class'] == 1]

#here in dataframe class with 1 label is selected for non\_fraud\_class

non\_fraud\_class = dataframe.loc[dataframe['Class'] == 0]

## Splitting the data:-

X = dataframe.iloc[:,:-1]

y = dataframe['Class']

#Finding the length of X and y

print("X and y sizes, respectively:", len(X), len(y))

#Splitting the training and Testing data

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.35, random\_state=500)

Train:-

logistic = linear\_model.LogisticRegression(C=1e5)

#Fitting the Algorithm for X\_train and y\_train

logistic.fit(X\_train, y\_train)

Test:-

print("Score: ", logistic.score(X\_test, y\_test))

print("Number of frauds on y\_test:", len(y\_test.loc[dataframe['Class'] == 1]), len(y\_test.loc[dataframe['Class'] == 1]) / len(y\_test))

Predict The data:-

y\_predicted = np.array(logistic.predict(X\_test))

y\_right = np.array(y\_test)

#print y\_test

#The confusion matrix (or error matrix) is one way to summarize the performance of a classifier

# for binary classification tasks. This square matrix

# consists of columns and rows that list the number of instances as absolute or

# relative "actual class" vs. "predicted class" ratios.

#Plotting the Confusion matrix for y\_right and y\_predicted

Finally got confusion matrics:-

confusion\_matrix = ConfusionMatrix(y\_right, y\_predicted)

print("Confusion matrix:",confusion\_matrix)

confusion\_matrix.plot(normalized=True)

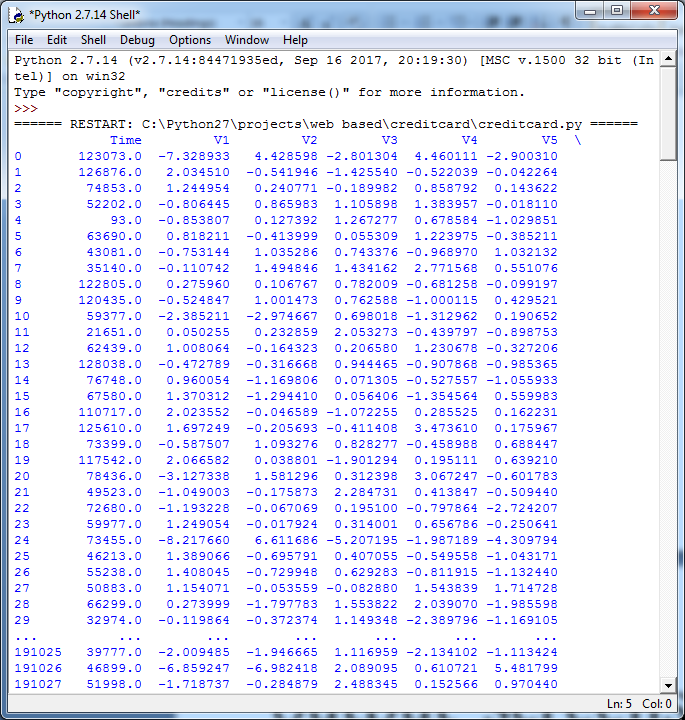
plt.show()

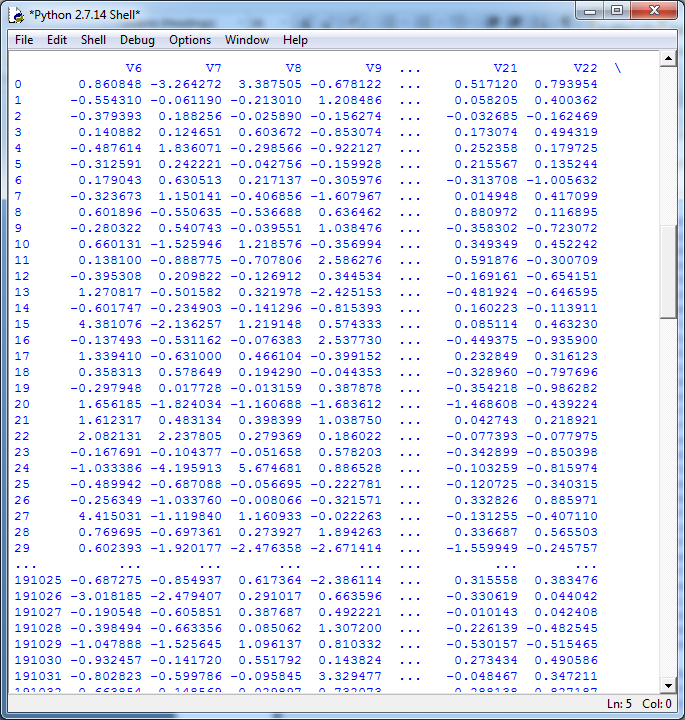
#printing the stats of Confusion matrix

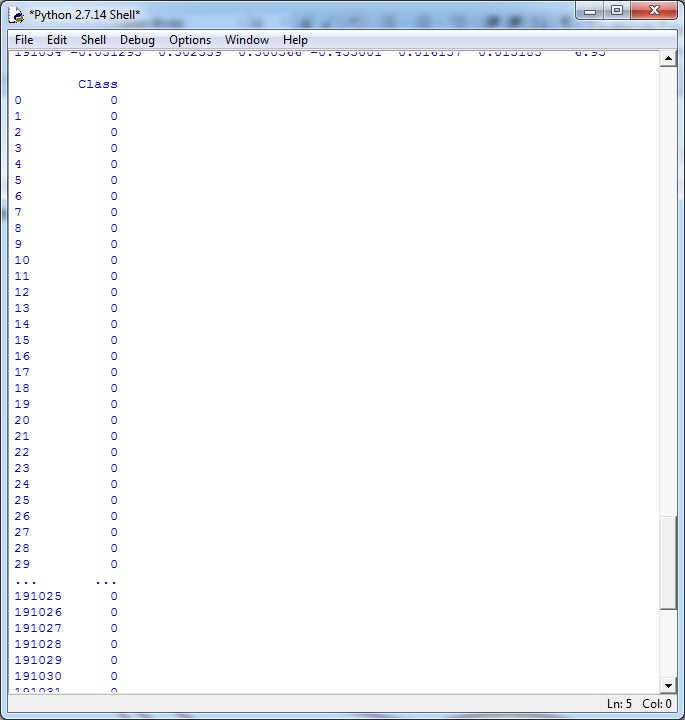
confusion\_matrix.print\_stats()

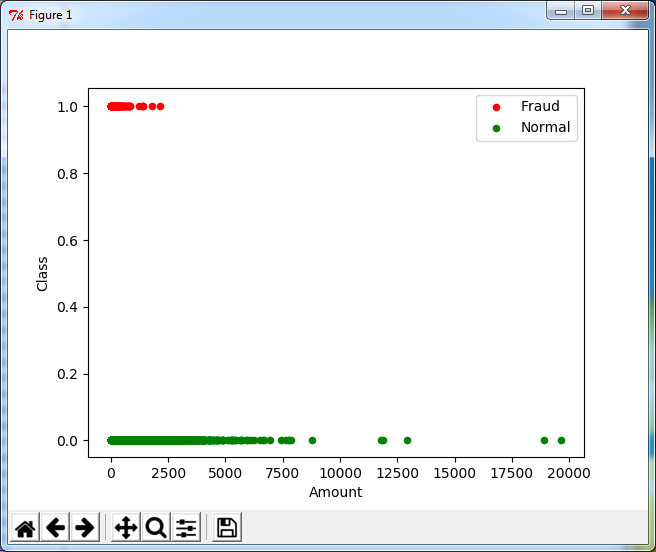
output screens:-

step1:-

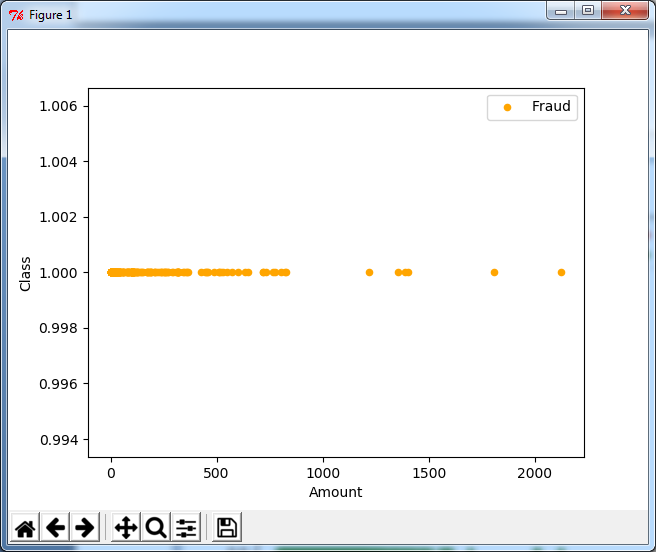




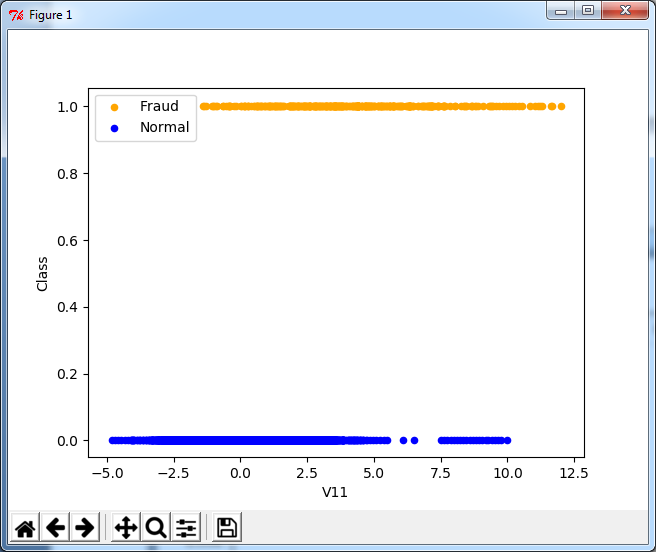




Step 2:-

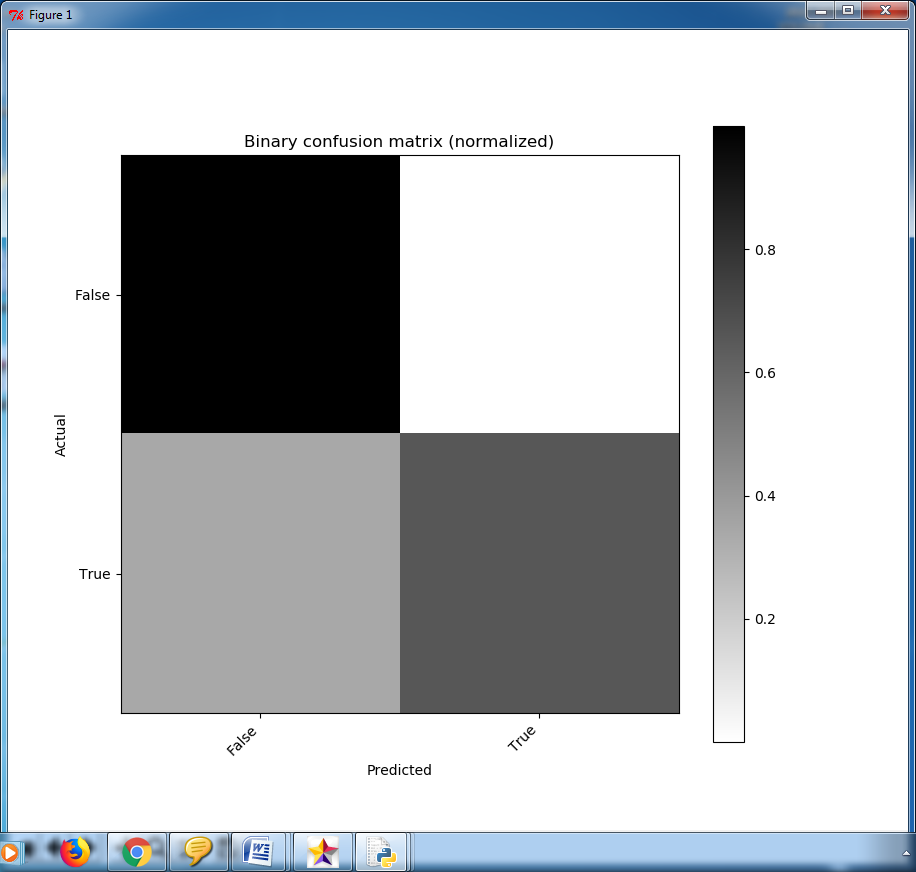


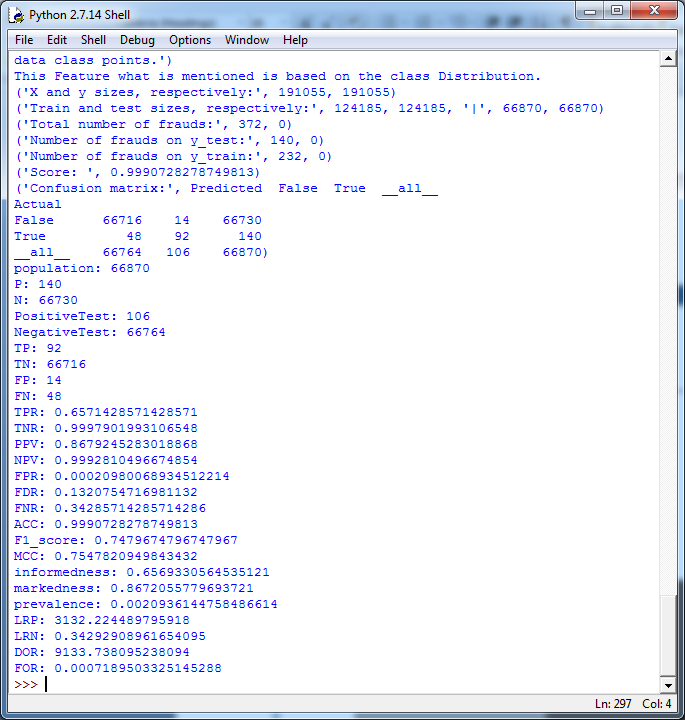
Step 3:-



Step 4

Finally got confusion metrics and high accuracy





S**YSTEM TESTING AND IMPLEMENTATION**

* 1. **INTRODUCTION**

Software testing is a critical element of software quality assurance and represents the ultimate review of specification, design and coding. In fact, testing is the one step in the software engineering process that could be viewed as destructive rather than constructive.

A strategy for software testing integrates software test case design methods into a well-planned series of steps that result in the successful construction of software. Testing is the set of activities that can be planned in advance and conducted systematically. The underlying motivation of program testing is to affirm software quality with methods that can economically and effectively apply to both strategic to both large and small-scale systems.

**8.2. STRATEGIC APPROACH TO SOFTWARE TESTING**

The software engineering process can be viewed as a spiral. Initially system engineering defines the role of software and leads to software requirement analysis where the information domain, functions, behavior, performance, constraints and validation criteria for software are established. Moving inward along the spiral, we come to design and finally to coding. To develop computer software we spiral in along streamlines that decrease the level of abstraction on each turn.

A strategy for software testing may also be viewed in the context of the spiral. Unit testing begins at the vertex of the spiral and concentrates on each unit of the software as implemented in source code. Testing progress by moving outward along the spiral to integration testing, where the focus is on the design and the construction of the software architecture. Talking another turn on outward on the spiral we encounter validation testing where requirements established as part of software requirements analysis are validated against the software that has been constructed. Finally we arrive at system testing, where the software and other system elements are tested as a whole.

UNIT TESTING

MODULE TESTING

SUB-SYSTEM TESING

SYSTEM TESTING

ACCEPTANCE TESTING

Component Testing

Integration Testing

User Testing

**8.3. UNIT TESTING**

Unit testing focuses verification effort on the smallest unit of software design, the module. The unit testing we have is white box oriented and some modules the steps are conducted in parallel.

**1. WHITE BOX TESTING**

This type of testing ensures that

* All independent paths have been exercised at least once
* All logical decisions have been exercised on their true and false sides
* All loops are executed at their boundaries and within their operational bounds
* All internal data structures have been exercised to assure their validity.

To follow the concept of white box testing we have tested each form .we have created independently to verify that Data flow is correct, All conditions are exercised to check their validity, All loops are executed on their boundaries.

**2. BASIC PATH TESTING**

Established technique of flow graph with Cyclomatic complexity was used to derive test cases for all the functions. The main steps in deriving test cases were:

Use the design of the code and draw correspondent flow graph.

Determine the Cyclomatic complexity of resultant flow graph, using formula:

V(G)=E-N+2 or

V(G)=P+1 or

V (G) =Number Of Regions

Where V (G) is Cyclomatic complexity,

E is the number of edges,

N is the number of flow graph nodes,

P is the number of predicate nodes.

Determine the basis of set of linearly independent paths.

**3. CONDITIONAL TESTING**

In this part of the testing each of the conditions were tested to both true and false aspects. And all the resulting paths were tested. So that each path that may be generate on particular condition is traced to uncover any possible errors.

**4. DATA FLOW TESTING**

This type of testing selects the path of the program according to the location of definition and use of variables. This kind of testing was used only when some local variable were declared. The *definition-use chain* method was used in this type of testing. These were particularly useful in nested statements.

**5. LOOP TESTING**

In this type of testing all the loops are tested to all the limits possible. The following exercise was adopted for all loops:

* All the loops were tested at their limits, just above them and just below them.
* All the loops were skipped at least once.
* For nested loops test the inner most loop first and then work outwards.
* For concatenated loops the values of dependent loops were set with the help of
* connected loop.Unstructured loops were resolved into nested loops or concatenated loops and tested as above.

# 9.1 INTRODUCTION

The protection of computer based resources that include hardware, software, data, procedures and people against unauthorized use or natural Disaster is known as System Security.

System Security can be divided into four related issues:

* Security
* Integrity
* Privacy
* Confidentiality

**SYSTEM SECURITY** refers to the technical innovations and procedures applied to the hardware and operation systems to protect against deliberate or accidental damage from a defined threat.

**DATA SECURITY** is the protection of data from loss, disclosure, modification and destruction.

**SYSTEM INTEGRITY** refers to the power functioning of hardware and programs, appropriate physical security and safety against external threats such as eavesdropping and wiretapping.

**PRIVACY** defines the rights of the user or organizations to determine what information they are willing to share with or accept from others and how the organization can be protected against unwelcome, unfair or excessive dissemination of information about it.

**CONFIDENTIALITY** is a special status given to sensitive information in a database to minimize the possible invasion of privacy. It is an attribute of information that characterizes its need for protection.

**9.3 SECURITY SOFTWARE**

It is the technique used for the purpose of converting communication. It transfers message secretly by embedding it into a cover medium with the use of information hiding techniques. It is one of the conventional techniques capable of hiding large secret message in a cover image without introducing many perceptible distortions.

NET has two kinds of security:

* Role Based Security
* Code Access Security

The Common Language Runtime (CLR) allows code to perform only those operations that the code has permission to perform. So CAS is the CLR's security system that enforces security policies by preventing unauthorized access to protected **resources** and **operations**. Using the Code Access Security, you can do the following:

* Restrict what your code can do
* Restrict which code can call your code
* Identify code

This machine learning fraud detection tutorial showed how to tackle the problem of credit card fraud detection using machine learning. It is fairly easy to come up with a simple model, implement it in Python and get great results for the Credit Card Fraud Detection task on Kaggle.