RAMACHANDRA
COLLEGE OF ENGINEERING
NH-16 Bypass Road, Vatluru (V), Eluru-534007, W.G Dt., A.P.

Approved by AICTE
Permanently Affiliated to JNTUK
Recognized by UGC 2(F) & 12(B)
Accredited by NAAC B++
Accredited by NBA (ME, ECE, CSE)
ISO 9001 : 2015 Certified

Consolidate Sheet

1.3.2. Name and Average percentage of courses that include experiential learning through project work/field work/internship during last five years

SL NO	Academic Year	Regulation	Department	Number of Subjects included from total number of Subjects.	% of Courses
1	2021-22	R19	CIV	25/30	83.33%
2	2021-22	R19	ME	24/32	75.00%
3	2021-22	R19	ECE	25/30	83.33%
4	2021-22	R19	EEE	22/31	70.00%
5	2021-22	R19	CSE	21/29	72.00%
6	2021-22	R19	MBA	12/17	70.00%
				Average Numbers of Courses	**75.613%**

Principal/Director

PRINCIPAL
Ramachandra College of Engineering
VATLUR (V), ELURU - 534 007
West Godavari District

Consolidate Sheet(Academic Year Wise)SAMPLE

1.3.2. Name and Average percentage of courses that include experiential learning through project work/field work/internship during last five years

Department : Computer Science & Engineering   **Academic Year: 2021-2022**

SL NO	Regulation(R19)	Name of Subjects Referred from total number of Subjects.
1	R19	Fundamentals of Computer Science
2	R19	Programming for Problem Solving using C
3	R19	Mathematical Foundations of Computer Science
4	R19	Software Engineering
5	R19	Python Programming
6	R19	Data Structures
7	R19	Object Oriented Programming through C++
8	R19	Computer Organization
9	R19	Design and Analysis of Algorithms
10	R19	Java Programming
11	R19	Operating Systems
12	R19	Database Management Systems
13	R19	Formal Languages and Automata Theory
14	R19	Data Warehousing and Data Mining
15	R19	Computer Networks
16	R19	Software Project Management
17	R19	Software Engineering
18	R19	Compiler Design
19	R19	Artificial Intelligence
20	R19	Distributed Systems
21	R19	Cryptography and Network Security
22	R19	Machine Learning
23	R19	Management and Organizational Behavior
24	R19	DevOps

HOD   **Principal**

1.3.2 Average percentage of courses that include experiential learning through project work/field work/internship during last five years

SL NO	COURSE NO	COURSE NAME
1	C.1.	Fundamentals of Computer Science
2	C.2.	Programming for Problem Solving using C
3	C.3.	Mathematical Foundations of Computer Science
4	C.4.	Software Engineering
5	C.5.	Python Programming
6	C.6.	Data Structures
7	C.7.	Object Oriented Programming through C++
8	C.8.	Computer Organization
9	C.9.	Design and Analysis of Algorithms
10	C.10.	Java Programming
11	C.11.	Operating Systems
12	C.12.	Database Management Systems
13	C.13.	Formal Languages and Automata Theory
14	C.14.	Data Warehousing and Data Mining
15	C.15.	Computer Networks
16	C.16.	Software Project Management
17	C.17.	Software Engineering
18	C.18.	Compiler Design
19	C.19.	Artificial Intelligence
20	C.20.	Distributed Systems
21	C.21.	Cryptography and Network Security
22	C.22.	Machine Learning
23	C.23.	Management and Organizational Behavior
24	C.24.	DevOps

HOD

Principal

CUSTOMER CHURN PREDICTION
USING MACHINE LEARNING

The Project work submitted to Jawaharlal Nehru Technological University,

Kakinada in partial fulfillments of the requirements for the award of the
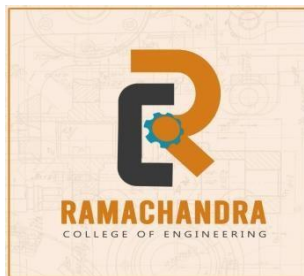
Degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

By

M. Neeraja

18ME1A0563

Under the Esteemed Guidance of

Dr. Satyabrata Dash

Associate Professor

Department of Computer Science and

Engineering Ramachandra College Of Engineering,Vatluru

Approved by AICTE, New Delhi, Affiliated by JNTU,Kakinada

NH-5,Near By Pass,Vatluru(V),Eluru-534007,WestGodavariDist

2018-2022

RAMACHANDRA COLLEGE OF ENGINEERING

(NBA,Accredited by NAAC at B++ Approved by AICTE,

New Delhi Affiliated by JNTUK,Kakinada)Vatluru,Eluru,A.P.

DEPARTMENT OF
COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that **M. Neeraja (18ME1A0563),**Bachelor of Technology in Computer Science and Engineering have successfully completed their project work entitled "CUSTOMER CHURN PREDICTION USING MACHINE LEARNING" atRamachandra college of Engineering,Eluru during the Academic Year 2018- 2022.This document is submitted in the partial fulfillment for the award of the Degree of Bachelor of Technology in Computer Science and Engineering and the same is not submitted elsewhere.

,

Dr. SatyabrataDash
Professor, Department of CSE
Project Guide

Dr.V.Suryanarayana
Professor & HOD, Department of CSE

External Examiner

ii

DECLARATION

I amM.Neeraja(18ME1A0563),hereby declares the project report titled "Customer Churn Prediction Using Machine Learning" under the supervision of Dr.Satyabrata Dash, Associate Professor, in the Department of Computer Science and Engineering is submitted in partial fulfillment of the requirements for the award of the degree of Bachelor Of Technology in Computer Science and Engineering.This is a record of carried out by us and the results embodied in this project have not been reproduced or copied from any source. The results embodied in this project report have not been submitted to any other university or institute for the award any other Degree orDiploma.

ACKNOWLEDGEMENT

I wish to take this opportunity to express our deep gratitude to all the people who have extended their cooperation in various ways during our project work. It is our pleasure and responsibility to acknowledge the help of all those individuals.

I have extended our sincere thanks to**. Dr.Satyabrata Dash**, Associate Professor in the Department of CSE for helping us in successful completion of our project under hissupervision.

I am very grateful to **Dr.V.Suryanarayana**, Head of the Department, Department of Computer Science & Engineering for his assistance and encouragement in all respects in carrying throughout our project work.

I express my deepest gratitude to **Dr. M. Muralidhar Rao**, Principal, Ramachandra College of Engineering, Eluru for his valuable suggestions during preparation of draft in our document.

I express my deepest gratitude to **The Management** of Ramachandra College of Engineering, Eluru for their support and encouragement in completing our project work and providing us necessary facilities.

I sincerely thank all the faculty members and staff of the Department of CSE for their valuable advices, suggestions and constant encouragement which played a vital role in carrying out this projectwork.

Finally, we thank one and all who directly or indirectly helped us to complete our project work successfully.

M. Neeraja
I8ME1A0563

INDEX

CHAPTER 4
METHODOLOGY

CHAPTER 5
RESULTS/OUTPUTS

CHAPTER 6
CONCLUSION

CHAPTER7
REFERENCES

ABSTRACT

All over the world, in different sectors churn prediction plays a very important role in the growth of the organization. For the company's revenue and profit, customer churn is very harmful. The most important step to avoid churn is to detect churn and its reason, accordingly initiate the prevention measures. Nowadays machine learning plays a vital role to get rid of this problem. The objective of this project is to predict the churn in banking sectors, by using well known machine learning techniques like Logistic Regression. The classification model is built by analyzing historical data and then applying the prediction model based on the analysis.Withholding the customer in an organization is one of the primary growth in today's world. Predicting the customer churn rate will help the bank in knowing which category of customers generally tend to leave the bank. Churn is based on various factors, including changing the position to a competitor, canceling their subscription because of poor customer service, or discontinuing all contact with a brand due to insufficient interaction between customers. Being connected for a long period of time with customers is more effective than trying to attract new customers. Dealing to figure out the amiss issues can make the customers happy. In this project finding the major factors that affect the customers to churn and analyze them by using Machine learning algorithms.Then churn gives the information of how many existing customers tend to leave the business, so lowering churn has an immense positive impact on the revenue streams. On the basis of this the Churn rates track lost customers, and growth rates track new customers comparing and analyzing both of these metrics tells exactly how much the business is growing over time. In this predictive process popular models have been used to achieve a decent level of accuracy.

- **C16,C17,C3,C19**

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 ABOUT THEPROJECT

To get and keep loyal customers for every business organization is a big challenge. Correct prediction about a customer is going to churn or not and then successfully convincing him to stay with that company can increase the revenue of that company.Therefore,predicting customer churn, i.e.if a customer is about to leave for a better service, that is an important part for analyzing the customer behavior. The churn model is a representation of various calculations that are built on existing historical data. The customer churn can be defined in other ways also, like low switching cost, deregulation motivates a customer to replace the sector. The churn is also classified into two: voluntary and involuntary churn Voluntary churn is defined as the termination of services by the customer itself, whereas involuntary churn is defined as the termination of services by the bank for fraud, non-payment services.

- **C.3**
- **C19**
- **C16**
- **C7**

The customer churn is very risky if it is not managed carefully, it may bring a company to its knees for poor maintenance services. Cost of customer churn also includes loss of revenue, profit. Previous case study has shown that the cost of maintaining a new customer is higher than the cost of maintaining the old one There are various banks who are suffering with this customer churn problem. So the most defined way to deal with this problem is developing a predictive model that can reliably and easily identify the possible churner. In the recent past the most frequently used technique is data mining to develop such models with satisfactory results Day by day when this churn prediction problem gets importance, much more research efforts are generated towards improving churn prediction rates.

Most frequently used features that have been used previously include credit score, geography, having a credit card or not, active member, estimated salary Due to business privacy and policy it is difficult to find and use public dataset for churn predictionIn this project, a new subset of features has been presented in order to improve the prediction accuracy. Logistic Regression has been used to predict the results.

1.2 PURPOSE

Customer churn prediction is the practice of assigning a churn probability to each customer in the company database, according to a predicted relationship between that customer's historical information and its future churning behavior. Practically, the probability to end the relationship with the company is then used to rank the customers from most to least likely to churn, and customers with the highest propensity to churn receive marketing retention campaigns**.** It reveals the behavior of those customers over a time period, and assesses the prediction quality of the models against the actual outcome of those customers.

All the leading organizations have been working for the customer's best interest. A customer has a choice due to healthy competition among the service providers and there is no end to best services. Shortage of data, targeted sales and up-gradation of companies are the major challenges while attaining the new customers. It is found that customer value and increasing revenue are the factors of current customer retention instead of the new customer acquiring.

The companies know their current existing customers and come into a strong relation with them and have a huge amount of data about them which is the key point to increase the profit and customer value. It is very important to find whether the customer will churn in the near future or stay within the bank, Which affects the revenue streams of the bank.

1.3 MOTIVATION

Churn will continue to exist and customer management is the best way to ensure sustainable business growth for long term profitability rather than capturing new customers.

Moreover, it should be noted that monitoring churn should be a constant concern of any company today, regardless of its industry, as competition is fierce and the digital age has made it easier for customers to move their business elsewhere if so is desired.

It is, therefore, in the best interest of any company to keep track of the behavior of its customers in order to potentially anticipate any signs of dissatisfaction that could eventually lead to churning. Such actions may be instrumental to reach out to those customers and hopefully save the relationship with the bank.

The underlying financial benefits of such efforts are straightforward, as potential churnerschose to maintain their business with the bank.so that the company's marketing

efforts be also channeled towards the retention of its current customers rather than just the recovery of past churners.

CHAPTER 2
LITERATURE SURVEY

In the banking industry, the scope of the term is wide and is currently being utilized within several different fields of the business. Credit card churn occurs when a customer ceases to use its credit card within a specific timeframe. Likewise, network banking churn may be defined as a customer who stops using its internet (home banking) service - Chiang, Wang, Lee, & Lin (2003) covered this topic by measuring the periodicity of transaction time of the users.

Additionally, Glady et al. (2008) defined a churner as a customer with less than 2.500 Euros of assets at the bank (savings, securities, or other kinds of products), and therefore paved the way for two distinct definitions of churn that exist in the organization that will be studied in this paper: the notion of voluntary churn and involuntary churn. The current study will be particularly geared towards tracking the behavioral history of past churners within a specific time frame in order to pinpoint certain patterns that might indicate that a customer is at risk of churning.

As Oyeniyi and Adeyemo (2015) pointed out, "churning is an important problem that has been studied across several areas of interest, such as mobile and telephony, insurance, and healthcare. Other sectors where the customer churn problem has been analyzed includes online social network churn analysis, and the retail banking industries' '. Although the broad or most generally accepted definition of churn refers to the loss of a customer by a specific company, one must analyze the concept with regards to the context in which it is being employed.

Eichinger, Nauck, and Klawonn (2006) defined customer attrition when a customer is leaving for a competitor. This notion has been backed by Qiasi, Roozbehani, &Minaei-bidgoli (2002) who consider churn when a customer discontinues the use of an organization's products and services in favor of a competitor's products and services. On the other hand, Neslin et al. (2006) described customer churn as the propensity.

M.A.H. Farquad [4] proposed a hybrid approach to overcome the drawbacks of general SVM model which generates a black box model (i.e., it does not reveal the knowledge gained during training in human understandable form). The hybrid approach contains three phases: In the first phase, SVM-RFE (SVM-recursive feature elimination) is employed to reduce the

feature set. In the second phase, dataset with reduced features is then used to obtain SVM model and support vectors are extracted. In the final phase, rules are then generated using Naive Bayes Tree (NBTree which is combination of Decision tree with naive Bayesian Classifier).

The dataset used here is bank credit card customer dataset (Business Intelligence Cup 2004) which is highly unbalanced with 93.24% loyal and 6.76%churned customers. The experimental showed that the model does not scalable to large datsets.

Wouter Verbeke [6] proposed the application of Ant-Miner+ and ALBA algorithms on a publicly available churn prediction dataset in order to build accurate as well as comprehensible classification rule-sets churn prediction models. Ant-Miner+ is a high performing data mining method based on the principles of Ant Colony Optimization which allows to include domain knowledge by imposing monotonicity constraints on the final rule-set. The advantages of Ant-Miner+ are high accuracy, comprehensibility of the generated models and the possibility to demand intuitive predictive models. Active Learning Based Approach (ALBA) for SVM rule extraction is a rule extraction algorithm, which combines the high predictive accuracy of a non-linear support vector machine model with the comprehensibility of the ruleset format.

The results which are benchmarked to C4.5, RIPPER, SVM and logistic regression showed that ALBA, combined with RIPPER, results in the highest accuracy, while sensitivity is the highest for C4.5 and RIPPER applied on an oversampled dataset. Ant-Miner+ results in less sensitive rule-sets, but allows to include domain knowledge, and results in comprehensible rule-sets that are much smaller than the rulesets induced with C4.5. RIPPER also results in small and comprehensible rule-sets, but lead models that violate domain

- **C16,C17,C3,C19**

to unintuitive knowledge.

Ning Lu proposed the use of boosting algorithms to enhance a customer churn prediction model in which customers are separated into two clusters based on the weight assigned by the boosting algorithm. As a result, a high risky customer cluster has been found. Logistic regression is used as a basis learner, and a churn prediction model is built on each cluster, respectively. The experimental results showed that boosting algorithm provides a good separation of churn data when compared with a single logistic regression model.

BenlanHe suggested a customer churn prediction methodology based on SVM model,

and used random sampling method to improve SVM model by considering the imbalance characteristics of customer data sets. A support vector machine constructs a hyper-plane in a high- or infinitedimensional space, which can be used for classification.

Random sampling method can be used to change the distribution of data in order to reduce the imbalance of the dataset. Imbalance in dataset is caused due to the low proportion of churners.

Ssu-Han Chen used a novel mechanism based on the gamma Cumulative SUM (CUSUM) chart in which the gamma CUSUM chart monitors individual customer's Inter Arrival Time (IAT) by introducing a finite mixture model to design the reference value and decision interval of the chart and used a hierarchical Bayesian model to capture the heterogeneity of customers. Recency, another time interval variable which is complementary to IAT, is combined into the model and tracks the recent status of the login behavior. In addition, benefits from the basic nature of control charts, the graphical interface for each customer is an additional advantage of the proposed method. The results showed that the accuracy rate (ACC) for gamma CUSUM chart is 5.2% higher than exponential CUSUM and the Average Time to Signal (ATS) is about two days longer than required for exponential CUSUM.

Koen W. De Bock proposed two rotation-based ensemble classifiers namely Rotation Forest and Rotboost as modeling techniques for customer churn prediction. An ensemble classifier is a combination of several member classifier models into one aggregated model, including the fusion rule to combine member classifiers outputs. In Rotation Forests, feature extraction is applied to feature subsets in order to turn the input data for training base classifiers, while RotBoost combines Rotation Forest with AdaBoost. Four data sets from real-life customer churn prediction projects are used here. The results showed that Rotation Forests outperform RotBoost in terms of area under the curve (AUC) and top-decile lift, while RotBoost demonstrates higher accuracy than Rotation Forests. They also compared three alternative feature extraction algorithms namely: Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Sparse Random Projections (SRP) on classification performance of both RotBoost and Rotation Forest.In general, the performance of rotation-based ensemble classifier depends upon: (i) the performance criteria used to measure classification performance and (ii) the implemented feature extraction algorithm.

Lee et al. focused on building an accurate and succinct predictive model with the purpose of churn prediction by using a Partial Least Squares (PLS) based method on highly correlated data sets among variables. They not only present a prediction model to accurately predict customers churning

behaviour, but also a simple but implementable churn marketing program was employed. The proposed methodology allows the marketing managers to maintain an optimal (atleast a near optimal) level of churners effectively and efficiently through the marketing programs. Here, PLS is employed as the prediction modelling method.

Y.Xie et al., [16] used an improved balance random forest (IBFR) model which is a combination of balanced random forests and weighted random forests in order to overcome the data distribution problem. The nature of IBRF is that the best features are iteratively learned by altering the class distribution and by putting higher penalties on misclassification of the minority class.

2.1 FEASIBILITY STUDY

The preliminary investigation examines project feasibility, the likelihood the application will be useful to the user. The main objective of the feasibility study is to test the Technical, Operational, and Economical feasibility for adding new modules and debugging traditional desktop-centric applications, and porting them to mobile devices. All systems are feasible if they are given unlimited resources and infinite time. There are aspects in the feasibility study portion of the preliminaryinvestigation:

The three major areas one should consider while determining the feasibility of the projectare:

- Technical Feasibility

- EconomicFeasibility

- Operational Feasibility

2.1.1 TECHNICAL FEASIBILITY

Evaluating the technical feasibility is the trickiest part of a feasibility study. This is because, at this point in time, not too many- detailed design of the system, making it difficult to access issues like performance, costs on (on account of the kind of technology to be deployed) etc. A number of issues have to be considered while doing a technical analysis.

1. Understand the different technologies involved in the proposed system: Before commencing the project, we have to be very clear about what are the technologies that are to be required for the development of the new system.

2. Find out whether the organization currently possesses the required technologies:

Is the required technology available with the organization? If so, is the capacity sufficient? For an instance - "Will the current printer be able to handle the new reports and forms

required for the new system?"

2.1.2 ECONOMIC FEASIBILITY

Economic feasibility attempts to weigh the costs of developing and implementing a new system, against the benefits that would accrue 9 from having the new system in place.

This feasibility study gives the top management the economic justification for the new system.

A simple economic analysis which gives the actual comparison of costs and benefits are much more meaningful in this case. These could include increased customer satisfaction, improvement in product quality better decision-making timeliness of information, expediting activities, improved accuracy of operations, better documentation and record keeping, faster retrieval of information, better employee morale.

2.1.3 OPERATIONAL FEASIBILITY

Proposed projects are beneficial only if they can be turned into information systems that will meet the organizations operating requirements. Simply stated, this test of feasibility asks if the system will work when it is developed and installed. Are there major barriersto Implementation? Here are questions that will help test the operational feasibility of a project:

Is there sufficient support for the project from management from users?

If the current system is well liked and used to the extent that persons will not be able to see reasons for change, there may be resistance.

Are the current business methods acceptable to the user?

If they are not, Users may welcome a change that will bring about a more operational and useful systems.

Has the user been involved in the planning and development of the project?

Early involvement reduces the chances of resistance to the system and in general and increases the likelihood of successful project.

Since the proposed system was to help reduce the hardships encountered. In the existing manual system, the new system was considered to be operational feasible.

2.2 EXISTINGSYSTEM

Churn studies have been used for years to achieve probability and to establish a sustainable

customer–company relationship. A customer churn in the banking sector indicates who has closed all their active accounts. A customer who has not used their bank for a few months or one year can be also considered as churned. Organizations developing churn management systems as a part of their customer relationship management.If churn is higher than growth, then business is getting smaller. Cost of acquiring new customers is high.

2.3 PROPOSED SYSTEM

In the proposed system, the models used in this project predict the customers likely to be churn based on the identified characteristics. This idea enables them to take the necessary actions and decreases the churn rates to retain such customers. Our study investigates in this process that the methods which are used in churn prediction have the ability to process huge amounts of customer data. Offering a comprehensive knowledge base can detach stuck users to reach their goals and to hold back the customers for a long period of time. Using a regression model in this project which works well with very large datasets. This model is very efficient to train and gives the direction of association which provides discreet output.
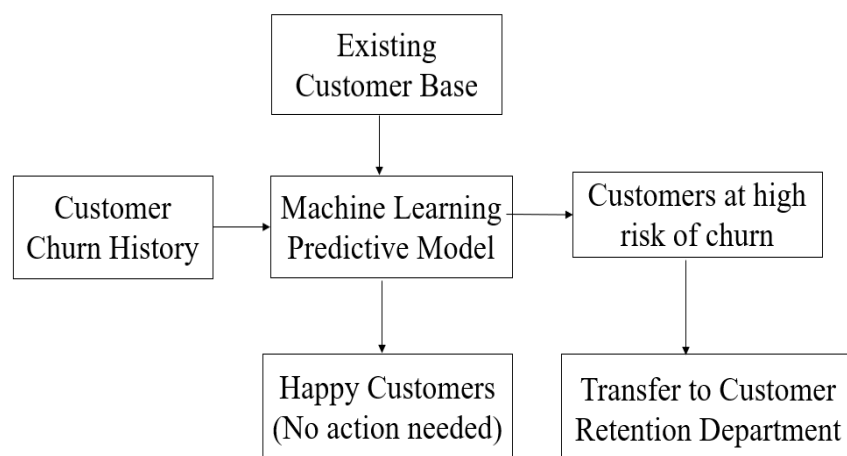
Fig 2.3.1 Block Diagram

CHAPTER 3

SYSTEM ANALYSIS AND DESIGN

3.1 REQUIREMENT SPECIFICATION

A software requirement specification (SRS) is a detailed description of a software system to be developed with its functional and non-functional requirements. The SRS is developed based on the agreement between the customers and contractors. It may include the use cases of how a user is going to interact with the software system. The software requirement specification document consists of all requirements required for project development. To develop the software system, we should have a clear understanding of the software system. To achieve this, we need continuous communication with customers to gather all requirements.

A good SRS defines how a software system will interact with all internal modules, hardware, communication with other programs, and human user interactions with a wide range of real-life scenarios. Using the software requirements specification (SRS) document on the QA lead, managers create a testplan.

Types

There are two types of requirements specification. They are:

- Functional requirementsspecification
- Non-functional requirementsspecification

Functional Requirements Specification

A Functional Requirement (FR) is a description of the service that the software must offer. It describes a software system or its component. A function is nothing but inputs to the software system, its behavior, and outputs. It can be a calculation, data manipulation, business process, user interaction, or any other specific functionality which defines what function a system is likely to perform. Functional Requirements are also called Functional Specification.

In software engineering and systems engineering, a Functional Requirement can range from the highlevel abstract statement of the sender's necessity to detailed mathematical functional requirement specifications. Functional software requirements help you to capture the intended behavior of the system

Non- Functional Requirement Specifications

Non-functional requirement (NFR) specifies the quality attribute of a software system. They judge the software system based on Responsiveness, Usability, Security, Portability, and other non-functional standards that are critical to the success of the software system. An example of a non-functional requirement, "how fast does the website load?" Failing to meet non-functional requirements can result in systems that fail to satisfy user needs.

The non-functional requirement allows you to impose constraints or restrictions on the design of the system across the various agile backlogs. For example, the site should load in 3 seconds when the number of simultaneous users is > 10000. Description of non-functional requirements is just as critical as a functional requirement.

Performance

The system must be interactive and the delays involved must be fewer. So, in every action-response of the system, there are no immediate delays. In the case ofopening windows forms, popping error messages, and saving the settings or sessions there is a delay much below 2 seconds. In the case of opening databases, sorting questions, and evaluation there are no delays, and the operation is performed in less than 2 seconds for opening, sorting, computing, posting > 95% of the files. Also, when connecting to the server the delay is based editing on the distance of the 2 systems and the Configuration between them so there is a high probability that there will be or not a successful connection in less than 20 seconds for sake of good communication.

Reliability

As the system provides the right tools for discussion, problem- solving it must be made sure that the system is reliable in its operations and for   sensitive details.

Safety

Information transmission should be securely transmitted to the server without any information changes.

Security

The main security concern is for the user's account hence proper login mechanism should be used to avoid hacking. The tablet id registration is a way to spam check for increasing security. Hence, security is provided from the unwanted use of recognition software.

Availability

If the internet service gets disrupted while sending information to the server, the information can be sent again for verification.

Usability

As the system is easy to handle and navigates most expectedly with no delays. In that case, the system program reacts accordingly and transverses quickly between its states.

Portability

It is the usability of the same software in different environments. The pre-requirement for portability is the generalized abstraction between theapplication logic and system interfaces. When software with the same functionality is produced for severalcomputing platforms, portability is the keyissue for development costreduction.

Testability

Software testability is the degree to which a software artifact (i.e. a software system, software module, requirements- or design document) supports testing in a given test context. If the testability of the software artifactis high, then finding faults in the system (if it has any) through testing.

3.2 SOFTWAREREQUIREMENTS

OperatingSystem   **:**Windows 10

Coding Language **:** Python

Front end   **:**HTML, CSS,

Specification   **:** InternetBrowser

3.3 HARDWARE REQUIREMENTS

Processor   **:**11th Gen Intel(R) Core(TM) i5

Hard Disk   **:**1 TB.

RAM   **:**4GB

3.4 SYSTEMDESIGN

System design is the process or art of defining the architecture, components, modules, interfaces, and data for a system to satisfy specified requirements. One could see it as the application of systems theory to product development. There is some overlap and synergy with the disciplines of systems analysis, systems architecture, and systems engineering.

3.4.1 SYSTEM ARCHITECTURE

A coherent computer-aided diagnosis system is built for brain tumordetection.Computer-aideddiagnosis(CAD)iseasyfordoctorsto identify the cancerous cells accurately. This system mainly deals with the identification of the brain cancer stage. The proposed method for brain tumor detection follows three basic tasks namely, preprocessing, feature extraction, andclassification.

As stated above, the acquired MRI scan image is preprocessed. Next, we extract the features from the segmented image. At last, we classify the image based upon the extracted features and area. In our proposed algorithm we have tried to solve the problems that we come across in the existed system.

This system is identified whether the tumor is cancerous or not. If the tumor is cancerous, it produces the results as Cancer. If the tumor is non- cancerous, it produces the results as No Tumor. Based on this information the tumor is curable by giving the proper treatment bythedoctors.So,thepatientcanbecurable fromthetumoratanearly stage oflife.

3.4.2Input Design

Input design plays a vital role in the life cycle of software development, it requires very careful attention of developers. The input design is to feed data to the application as accurately as possible. So, inputs are supposed to be designed effectively so that the errors occurring while feeding as minimized. According to the software engineering concepts, the input forms or screens are designed to provide to have a validation control over the input limit, range, and other relatedvalidations.

This system has input screens in almost all the modules. Error messages are developed to alert the user whenever he commits some mistakes and guides him in the right way so thatinvalid entries are not made. Let us see deeply about this under module design.

Input design is the process of converting the user-created input into a computer-basedformat. The goal of the input design is to make the data entry logical and free from errors.

Theerror is in the input are controlled by the input design. The application has been developed ina user-friendly manner. The forms have been designed in such a way during the processing

the cursor is placed in the position where must be entered. The user is also provided with an option to select an appropriate input from various alternatives related to the field in certaincases.

Validations are required for each data entered. Whenever a user enters erroneous data, an error message is displayed and the user can move on to the subsequent pages aftercompleting all the entries on the current page.

3.4.3 Output Design

The output from the computer is required to mainly create an efficient method of communication within the company primarily among the project leader and his team members, in other words, the administrator and the client. The output of VPN is the system which allows the project leader to manage his clients in term of creating new clients and assigning new projects to them, maintaining a record of the project validity and providing folder level access to each client on the user side depending on the project allotted to him. After completion of a project, a new project may be assigned to the client. User authentication procedures are maintained at the initial stages themselves. A new user may be created by theadministrator himself or a user can himself register as a new user but the task of assigning projects and validation a new user sets with the administrator only.

The application starts running when it is executed for the first time. The server has tobe started. The project will run on the local area network so the server machine will serve as the administrator while the other connected systems can act as the clients. The developed system is highly user-friendly and can be easily understood by anyone using it even for the firsttime.

3.5 UML DIAGRAMSINTRODUCTION

The UML stands for Unified modeling language, is a standardized general-purpose visual modeling language in the field of Software Engineering. It is used for specifying, visualizing, constructing, and documenting the primary artifacts of the software system. It helps in designing and characterizing, especially those software systems that incorporate the concept of Object orientation. It describes the working of both the software and hardware systems.

The UML was developed in 1994-95 by Grady Booch, Ivar Jacobson, and James Rumbaugh at the Rational Software. In 1997, it got adopted as a standard by the Object Management Group (OMG).

The Object Management Group (OMG) is an association of several companies that controls the open standard UML. The OMG was established to build an open standard that mainly supports the interoperability of object-oriented systems. It is not restricted within the boundaries, but it can also be utilized for modeling the non-software systems. The OMG is best recognized for the Common Object Request Broker Architecture (CORBA) standards.

UML CONCEPTS

The Unified Modeling Language (UML) is a standard language for writing software blue prints. The UML is a language for • Visualizing 26 • Specifying • Constructing • Documenting the artifacts of a software intensive system. The UML is a language which provides vocabulary and the rules for combining words in that vocabulary for the purpose of communication. A modeling language is a language whose vocabulary and the rules focus on the conceptual and physical representation of a system. Modeling yields an understanding of a system.

UML DIAGRAMS

UML is a way of visualizing a software program using a collection of diagrams. The notation has evolved from the work of Grady Booch, James Rumbaugh, Ivar Jacobson, and the Rational Software Corporation to be used for object-oriented design, but it has since been extended to cover a wider variety of software engineering projects. Today, UML is accepted

by the Object Management Group (OMG) as the standard for modeling software development.

BEHAVIORAL DIAGRAMS

3.5.1 Use case diagram

To model a system, the most important aspect is to capture the dynamic behavior. Dynamic behavior means the behavior of the system when it is running/operating. Only static behavior is not sufficient to model a system rather dynamic behavior is more important than staticbehavior.

These internal and external agents are known as actors. Use case diagrams consist of actors, use cases, and their relationships. The diagram is used to model the system/subsystem of an application. A single and their relationships. The diagram is used to model the system/subsystem of an application. A single-use case diagram captures a particular functionality of a system.

Purpose of Use Case Diagrams

The purpose of the use case diagram is to capture the dynamic aspect of a system. However, this definition is too generic to describe the purpose, as the other four diagrams (activity, sequence, collaboration, and State chart) also have the same purpose. We will look into some specific purpose, which will distinguish it from the other four diagrams. When the initial task is complete, use case diagrams are modeled to present the outside view.

In brief, the purposes of use case diagrams can be said to be as follows–

- Used to gather the requirements of asystem.
- Used to get an outside view of asystem.
- Identify the external and internal factors influencing thesystem.

i.Use cases

A use case describes a sequence of actions that provide something of measurable value to an actor and is drawn as a horizontal ellipse.

ii. Actors

An actor is a person, organization, or external system that plays a role in one or more interactions with the system.
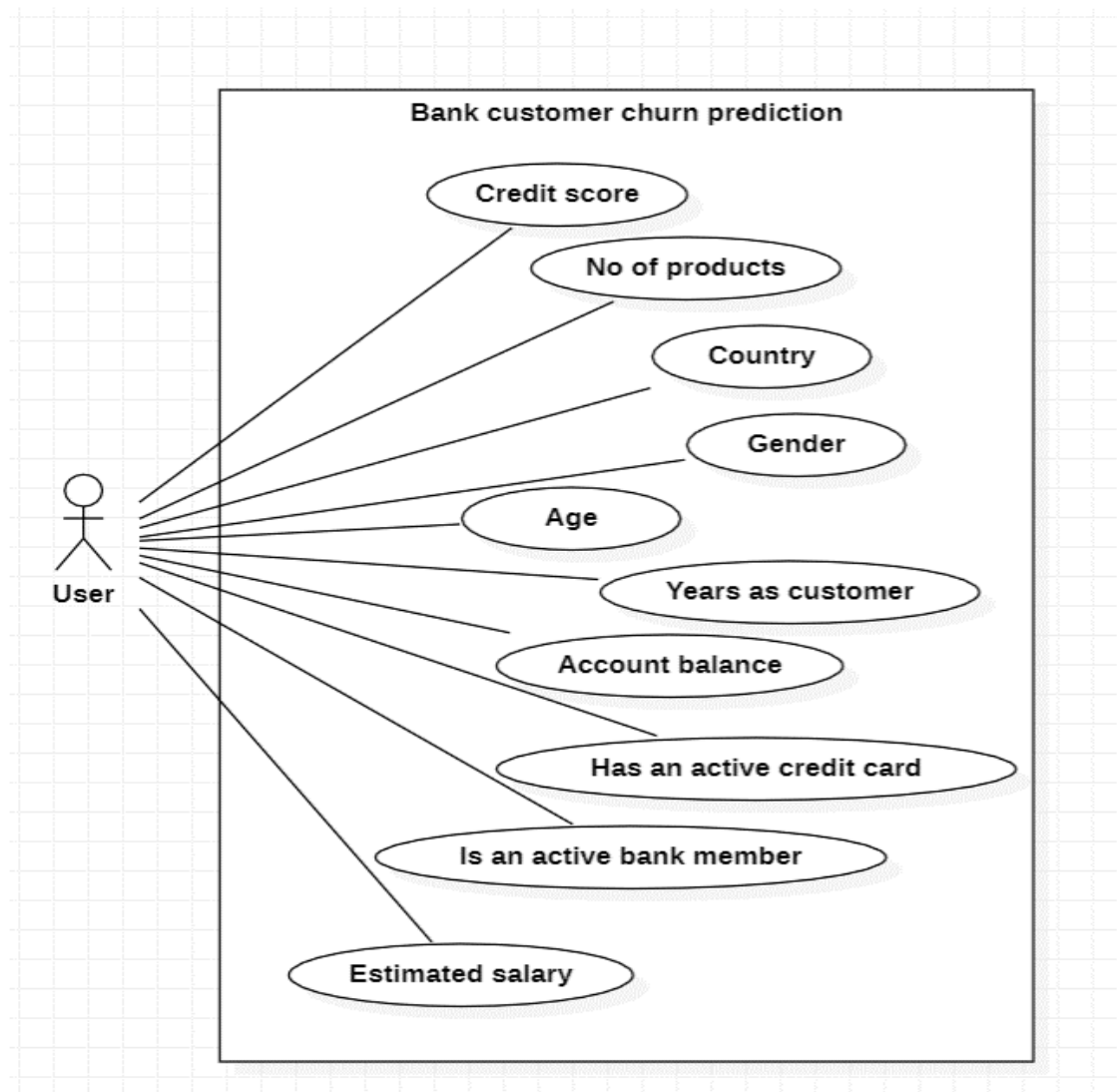

Fig3.5.1.1 usecase diagram

3.5.2 Sequence diagram

Sequence diagram simply depicts interaction between objects in a sequential order i.e. the order in which these interactions take place. It describes how and in what order the objects in a system function.
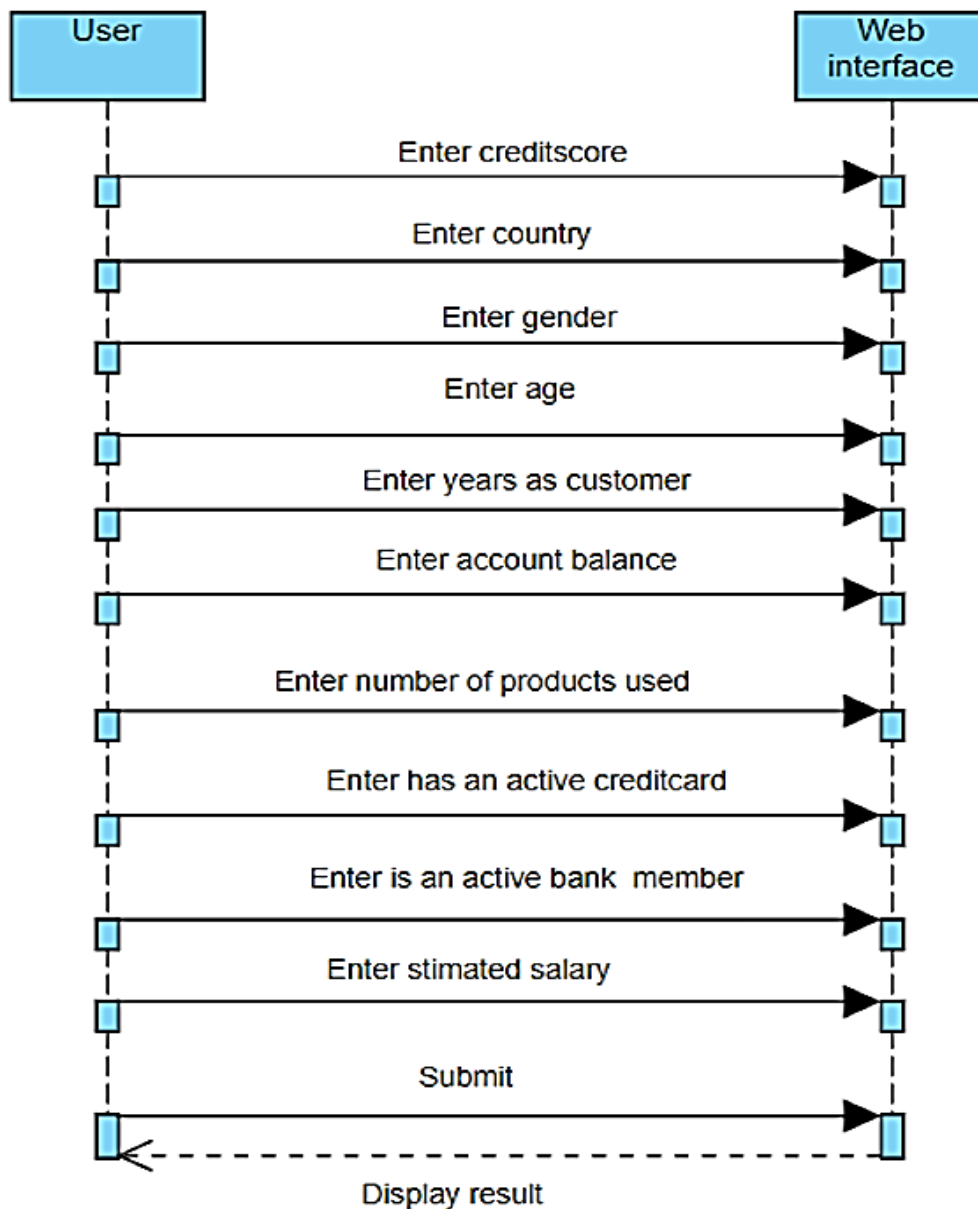
Fig3.5.2.1 sequence diagram

3.5.3 Statechart diagram

The name of the diagram itself clarifies the purpose of the diagram and other details. It

describes the different states of a component in a system. The states are specific to a component/object of a system. A Statechart diagram describes a state machine. A state machine can be defined as a machine that defines different states of an object and these states are controlledby external or internal event.

Purpose of Statechart Diagrams

Statechart diagram is one of the five UML diagrams used to model the dynamic natureofa system. They define different states of an object during its lifetime and these states are changed by events. Statechart diagrams are useful to model reactive systems. Reactive systems can be defined as a system that responds to external or internal events. Statechart diagrams are also used for forward and reverse engineering of a system. However, the main purpose is to model the reactivesystem.

Following are the main purposes of using Statechart diagrams –

- To model the lifetime of a reactivesystem.

- To describe different states of an object during itslifetime.
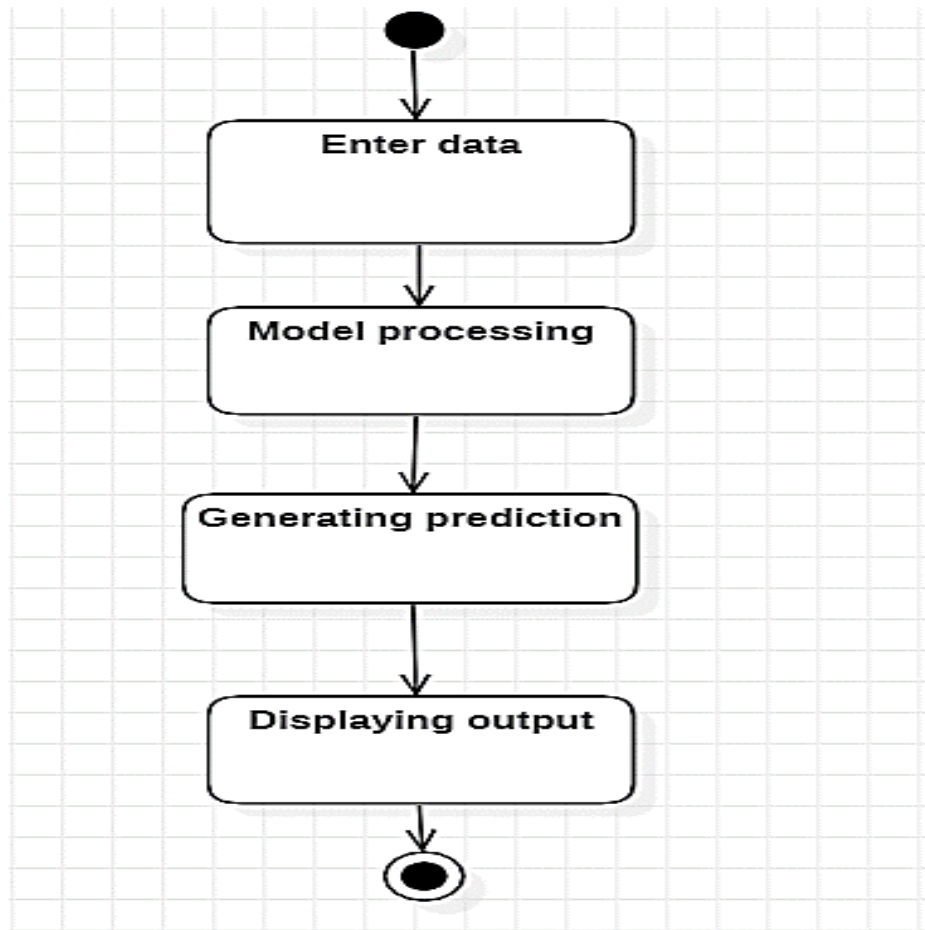
- Define a state machine to model the states of anobject.

Fig3.5.3.1 statechart diagram

3.5.4 Activity Diagram

The activity diagram is another important diagram in UML to describe the dynamic aspects of the system.

An activity diagram is a flowchart to represent the flow from one activity to another activity. The activity can be described as an operation of the system.

Purpose of Activity Diagrams

The basic purposes of activity diagrams are similar to the other four diagrams. It captures the dynamic behavior of the system. The other four diagrams are used to show the message flow from one object to another but the activity diagram is used to show message flow from one activity toanother.

The purpose of an activity diagram can be described as –

- Draw the activity flow of asystem.

- Describe the sequence from one activity toanother.

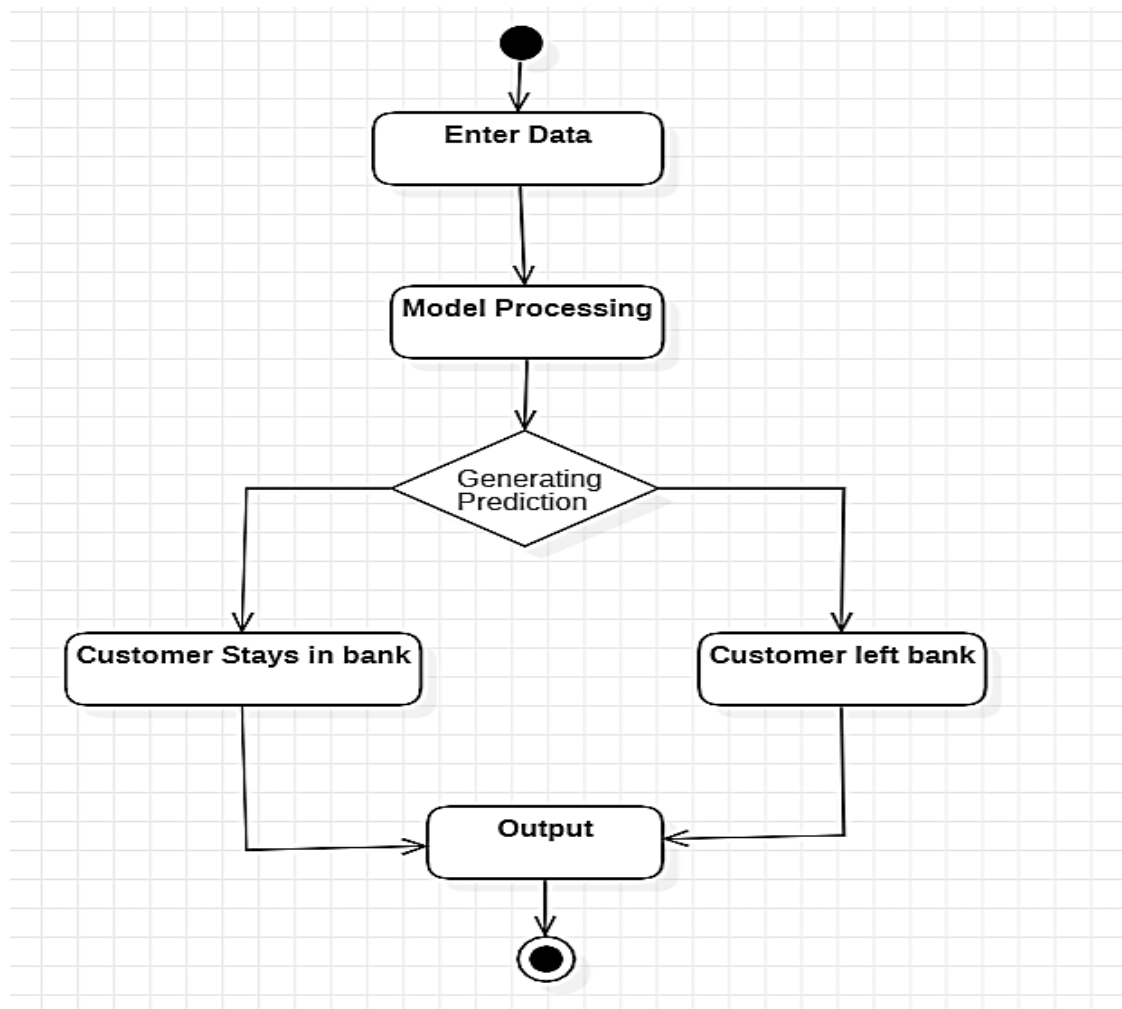- Describe the parallel, branched, and concurrent flow of thesystem.

Fig3.5.4.1 Activity diagram

3.6 SOFTWAREENVIRONMENT

Software environment emerged in the middle of the midrange era as a means of improving software quality and productivity through automation. A software environment may be described as an'operating system environment and a collection of tools or subroutines. A slightly better definition of a software environment is a 'coordinated collection of software tools organized to support some approach to software development or conform to some software process model', where software tools are defined as 'computer programs that assist engineers with the design and development of computer-basedsystems'.

Structured programming environments were created as a means of improving softwarereliability and productivity using guidelines, code libraries, structured coding, top-down development, chief programmer teams, standards, procedures, documentation, education, andmetrics.

Software factories were soon created to introduce discipline and repeatability, software visualization tools, the capture of customer needs or requirements, automated software testing, and software reuse. Computer-assisted software engineering or CASE was also created to enhance software productivity and reliability by automating document production, diagram design, code compilation, software testing, configuration management, management reporting, and sharing of data by multiple developers.

3.6.1 PYTHON PROGRAMMING

- **C16,C17,C3,C19**

Python is one of the most dynamic and versatile programming languages available in the industry today. Since its inception in the 1990s, Python has become hugely popular. Python is an interpreted, interactive, object-oriented programming language. It is also usable as an extension language for applications that need a programmable interface. Finally, Python is portable: it runs on many Unix variants, on the Mac, and on Windows 2000 and later. When he began implementing Python, Guido van Rossum was also reading the published scripts from "Monty Python's Flying Circus", a BBC comedy series from the 1970s. Van Rossum thought he needed a name that was short, unique, and slightly mysterious, so he decided to call the language Python.

Features of Python

1. Easy to Learn and Use

Python is easy to learn as compared to other programming languages. Its syntax is straightforward and much the same as the English language. There is no use of the semicolon or curly-bracket, the indentation defines the code block. It isthe recommended programming language for beginners.

2.Expressive Language

Python can perform complex tasks using a few lines of code. A simple example, the hello world program you simply type print("Hello World"). It will take only one line to execute, while Java or C takes multiple lines.

3. Interpreted Language

Python is an interpreted language; it means the Python program is executed one line at a time. The advantage of being interpreted language, it makes debugging easy and portable.

4.Cross-platform Language

Python can run equally on different platforms such as Windows, Linux, UNIX, and Macintosh, etc. So, we can say that Python is a portable language. It enables programmers to develop the software for several competing platforms by writing a program only once.

5. Free and Open Source

Python is freely available for everyone. It has a large community across the world that is dedicatedly working towards make new python modules and functions. Anyone can contribute to the Python community. The open-source means, "Anyone can download its source code without paying any penny."

6.Object-Oriented Language

Python supports object-oriented language and concepts of classes and objects come into existence. It supports inheritance, polymorphism, and encapsulation, etc. The object-oriented procedure helps to programmer to write reusable code and develop applications in less code.

7. Extensible

It implies that other languages such as C/C++ can be used to compile the code and thus it can be used further in our Python code. It converts the program into byte code, and any platform can use that byte code.

8.Large Standard Library

It provides a vast range of libraries for the various fields such as machine learning, web developer, and also for the scripting. There are various machine learning libraries, such as Tensor flow, Pandas, Numpy, Keras, and Pytorch, etc. Django, flask, pyramids are the popular framework for Python web development.

9.GUI Programming Support

Graphical User Interface is used for the developing Desktop application. PyQT5, Tkinter, Kivy are the libraries which are used for developing the web application.

10.Integrated

It can be easily integrated with languages like C, C++, and JAVA, etc. Python runs code line by line like C, C++ Java. It makes easy to debug the code.

11. Embeddable

The code of the other programming language can use in the Python source code. We can use Python source code in another programming language as well. It can embed other language into our code.

12. Dynamic Memory Allocation

In Python, we don't need to specify the data-type of the variable. When we assign some value to the variable, it automatically allocates the memory to the variable at run time. Suppose we are assigned integer value 10 to **x,** then we don't need to write **int x = 10.** Just write **x =10.**

Applications of Python

     1. AI and Machine Learning

     2. Data Analytics

     3. Data Visualization

     4. Programming applications

     5. Web development

     6. Game development

     7. Language development

     8. Finance

     9. Search Engine Optimisation

Advantages of Python

1. Extensive Libraries

Python downloads with an extensive library and *c*ontains code for various purposes like regular expressions, documentation-generation, unittesting, web browsers, threading, databases, CGI, email, image manipulation, and more. So, we don't have to write the complete code for that manually.

2.Extensible

Python can be extended to other languages. You can write some of your code in languages like C++ or C. This comes in handy, especially in projects.

3. Embeddable

Complimentary to extensibility, Python is embeddable as well. You can put your Python code in your source code of a different language, like C++.

4.Improved Productivity

The language's simplicity and extensive libraries render programmers more productive than more productive than Java and C++ do.

5. IOT Opportunities

Since Python forms the basis of new platforms like Raspberry Pi, it finds the future bright for the Internet of Things.

6.Object-Oriented

This language supports both the procedural and object - oriented programming paradigms. While, python functions help us with code reusability, classes and objects let us model the real world. A class allows the encapsulation of data and functions into one.

3.6.2 PYTHON – IDLE

Every Python installation comes with an Integrated Development and Learning Environment, which you'll see shortened to IDLE or even IDE. These are a class of applications that help you write code more efficiently. While there are many IDEs for you to choose from, Python IDLE is very bare-bones, which makes it the perfect tool 15 for a beginning programmer. It is included in Python installations on Windows and Mac. If you're a Linux user, then you should be able to find and download Python IDLE using your package manager.

Some of the key features it offers are:

- Python shell with syntax highlighting
- Multi-window text editor
- Code auto completion
- Intelligent indenting
- Program animation and stepping which allows one line of code to run at a time helpful for debugging
- Persistent breakpoints, etc.

Exploring IDLE Software Features

- Remember that it is not advisable to write multiple lines of code that havefunctions/classes in the IDLE shell

- In such cases, you can go to the File option of IDLE and click on New file.

- IDLE can be customized using the options present in the Format, Edit and Options menu.

Fig3.6.2.1 : Python IDLE features

Usage of Python IDLE

The shell is the default mode of operation for Python IDLE. When you click on the icon to open the program, the shell is the first thing that you see:
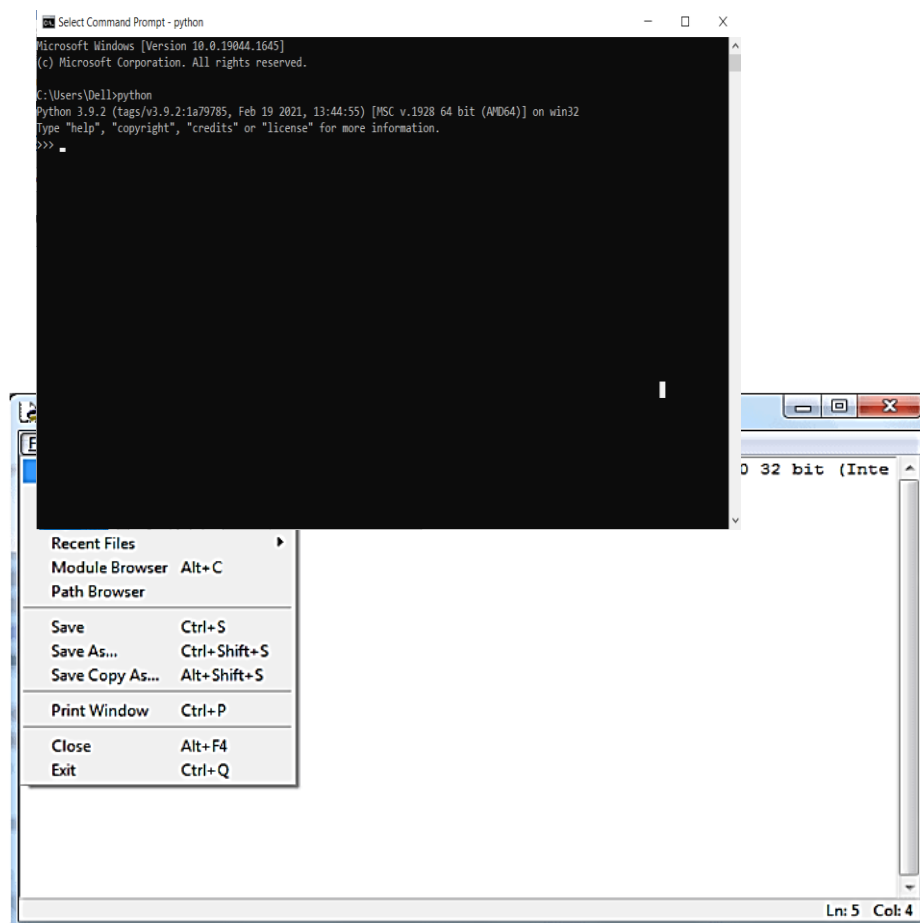
Fig3.6.2.2 : Python Shell, a default mode

This is a blank Python interpreter window. You can use it to start interacting with Python immediately.

3.6.3 VISUAL STUDIO CODE

Visual Studio Code is an integrated development environment (IDE) from Microsoft. It is used to develop console and graphical user interface applications along with Windows Forms or WPF applications, web sites, web applications, and web services in both native code together with managed code for all platforms supported by Microsoft Windows, Windows Mobile, Windows CE, .NET Framework, and so on. It supports different programming languages by means of language services, which allow the code editor and debugger to support (to varying degrees) nearly any programming language, provided a language-specific service exists.

Visual Studio also includes a web-site editor and designer that allow web pages to be authored by dragging and dropping widgets. It is used for developing VB.NET application efficiently to get input and output design easiest one. It will be run at windows application based on services provided by the user.

Features:

Built with love for the Web

VS Code includes enriched built-in support for Node.js development with JavaScript and TypeScript, powered by the same underlying technologies that drive Visual Studio. VS Code also includes great tooling for web technologies such as JSX/React, HTML, CSS, SCSS, Less, and JSON.

Robust and extensible architecture

Architecturally, Visual Studio Code combines the best of web, native, and language-specific technologies. Using Electron, VS Code combines web technologies such as JavaScript and Node.js with the speed and flexibility of native apps. VS Code uses a newer, faster version of the same industrial-strength HTML-based editor that has powered the "Monaco" cloud editor, Internet Explorer's F12 Tools, and other projects. Additionally, VS Code uses a tools service architecture that enables it to integrate with many of the same technologies that power Visual Studio, including Roslyn for .NET, TypeScript, the Visual Studio debugging engine, and more.

Visual Studio Code includes a public extensibility model that lets developers build and use extensions, and richly customize their edit-build-debug experience.

Ready, set, code!

If you prefer a code editor-centric development tool or are building cross-platform web and cloud applications, we invite you to try out Visual Studio Code and let us know what you th
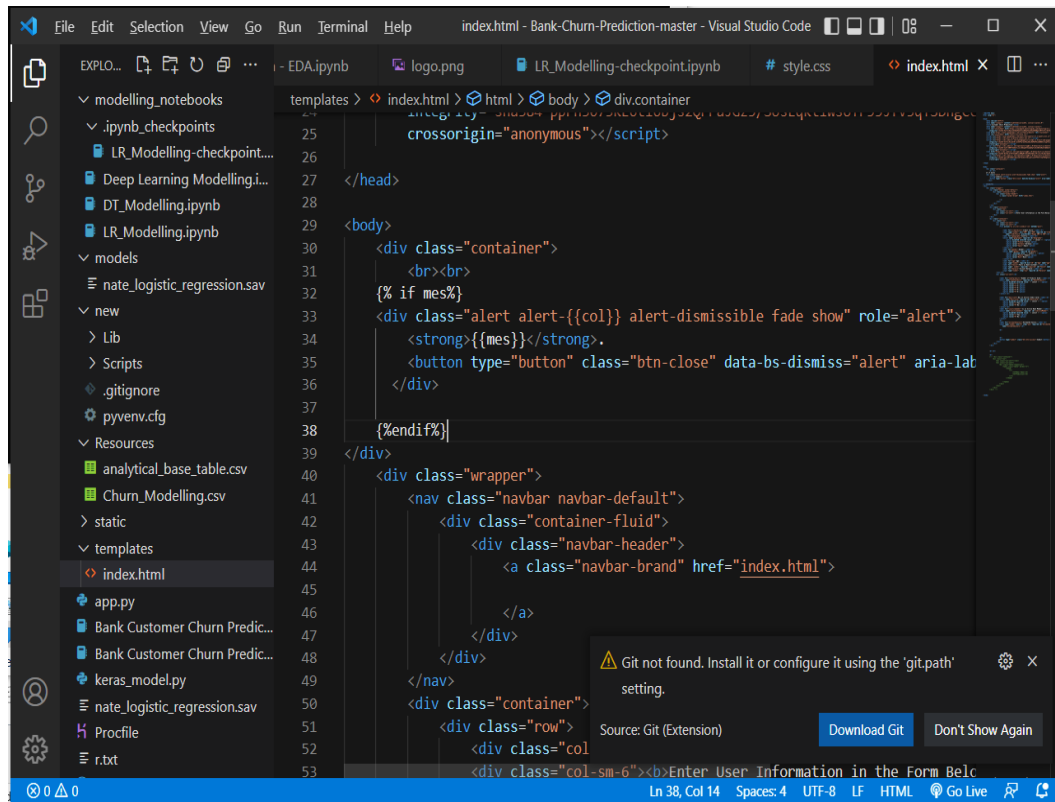
Fig3.6.3.1 Visual Studio Code Environment

3.6.4 WEB TECHNOLOGIES

HTML

- HTML is an acronym which stands for Hyper Text Markup Language which is used for creating web pages and web applications. Let's see what is meant by Hypertext Markup Language, and Web page.

- Hyper Text: HyperText simply means "Text within Text." A text has a link within it, is a hypertext. Whenever you click on a link which brings you to a new webpage, you have clicked on a hypertext. HyperText is a way to link two or more web pages (HTML documents) with each other.

- Markup language: A markup language is a computer language that is used to apply layout and formatting conventions to a text document. Markup language makes text more interactive and dynamic. It can turn text into images, tables, links, etc.

- Web Page: A web page is a document which is commonly written in HTML and translated by a web browser. A web page can be identified by entering an URL. A Web page can be of the static or dynamic type. With the help of HTML only, we can create static web pages.

- Hence, HTML is a markup language which is used for creating attractive web pages with the help of styling, and which looks in a nice format on a web browser. An HTML document is made of many HTML tags and each HTML tag contains different content.

Uses of HTML

Building Web Pages

With purely HTML, one can create a simple webpage. However, The World Wide Web Consortium recommends that HTML be used concurrently with CSS. Since 1999, HTML has been the official standard for displaying web pages. The most common web browsers have inbuilt web development tools that help modify and display web pages written in HTML.

Navigating the Internet

It looks seamless when navigating to the home page by clicking the home button or from one web page to another. Navigation is made possible by HTML anchor tags that link web pages together. HTML became outstanding because of this Hypertext concept. At the moment, it is possible to access any document on the world wide web by the use of a hypertext link, the

anchor tag.

HTML Styling

Without styles, websites would be dull. HTML style element allows for styling of the web page. One can style the font family, colour, and size of the text.

Responsive Design

Responsive design is an integral part of web development. HTML images have an attribute known as 'srcset.' This attribute references the images that the browser will parse and their respective sizes. These media queries can make the selected images to be responsive.

- C.2
- C4
- C5
- C6,C10
- C16,C17

Storage in the Browser

HTML features such as localStorage and IndexDB have transformed the approach to the storage of user data. HTML5 brought in these new features, and most browsers support them. Depending on user permission, these features can be more useful when collecting and storing data.

Offline Web Applications

Nowadays, it is possible to access web applications while offline. HTML allows the user to operate with their data. HTML uses a cache manifest file to determine which data to store while offline. Users can also generate PDF from HTML to use while offline.

CSS

CSS stands for Cascading Style Sheets. It is a style sheet language which is used to describe the look and formatting of a document written in markup language. It provides an additional feature to HTML.

Uses of CSS

Solves a big problem

Before CSS, tags like font, color, background style, element alignments, border and size had

to be repeated on every web page. This was a very long process. For example: If you are developing a large website where fonts and color information are added on every single page, it will be become a long and expensive process. CSS was created to solve this problem.

Saves a lot of time

CSS style definitions are saved in external CSS files so it is possible to change the entire website

by changing just one file.

Provide more attributes

CSS provides more detailed attributes than plain HTML to define the look and feel of the website.

3.7 MACHINE LEARNING

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: The ability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect.

- **C.3**
- **C19**
- **C16**
- **C7**

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow, the market demand for data scientists will increase, requiring them to assist in the identification of the most relevant business questions and subsequently the data to answer them.

Types of Machine Learning

Machine learning is a subset of AI, which enables the machine to automatically learn from data, improve performance from past experiences, and make predictions. Machine learning contains a set of algorithms that work on a huge amount of data. Data is fed to these algorithms to train them, and on the basis of training, they build the model & perform a specific task.

Supervised Machine Learning

As its name suggests, Supervised machine learning is based on supervision. It means in the supervised learning technique, we train the machines using the "labelled" dataset, and based on the training, the machine predicts the output. Here, the labelled data specifies that some of the inputs are already mapped to the output. More preciously, we can say; first, we train the

machine with the input and corresponding output, and then we ask the machine to predict the output using the test dataset.

Let's understand supervised learning with an example. Suppose we have an input dataset of cats and dog images. So, first, we will provide the training to the machine to understand the images, such as the shape & size of the tail of cat and dog, Shape of eyes, colour, height (dogs are taller, cats are smaller), etc. After completion of training, we input the picture of a cat and ask the machine to identify the object and predict the output. Now, the machine is well trained, so it will check all the features of the object, such as height, shape, colour, eyes, ears, tail, etc., and find that it's a cat. So, it will put it in the Cat category. This is the process of how the machine identifies the objects in Supervised Learning.

The main goal of the supervised learning technique is to map the input variable(x) with the output variable(y). Some real-world applications of supervised learning are Risk Assessment, Fraud Detection, Spam filtering, etc.

Applications of Supervised Learning

- ImageSegmentation:
  Supervised Learning algorithms are used in image segmentation. In this process, image classification is performed on different image data with pre-defined labels.

- MedicalDiagnosis:
  Supervised algorithms are also used in the medical field for diagnosis purposes. It is done by using medical images and past labelled data with labels for disease conditions. With such a process, the machine can identify a disease for the new patients.

- Fraud Detection - Supervised Learning classification algorithms are used for identifying fraud transactions, fraud customers, etc. It is done by using historic data to identify the patterns that can lead to possible fraud.

- Spam detection - In spam detection & filtering, classification algorithms are used. These algorithms classify an email as spam or not spam. The spam emails are sent to the spam folder.

- Speech Recognition - Supervised learning algorithms are also used in speech recognition. The algorithm is trained with voice data, and various identifications can be done using the same, such as voice-activated passwords, voice commands, etc.

Unsupervised Machine Learning

Unsupervised learning is different from the Supervised learning technique; as its name suggests, there is no need for supervision. It means, in unsupervised machine learning, the machine is trained using the unlabeled dataset, and the machine predicts the output without any supervision.

In unsupervised learning, the models are trained with the data that is neither classified nor labelled, and the model acts on that data without any supervision Applications of Unsupervised Learning

- Network Analysis: Unsupervised learning is used for identifying plagiarism and copyright in document network analysis of text data for scholarly articles.
- Recommendation Systems: Recommendation systems widely use unsupervised learning techniques for building recommendation applications for different web applications and e-commerce websites.
- Anomaly Detection: Anomaly detection is a popular application of unsupervised learning, which can identify unusual data points within the dataset. It is used to discover fraudulent transactions.
- Singular Value Decomposition: Singular Value Decomposition or SVD is used to extract particular information from the database. For example, extracting information of each user located at a particular location.

Semi-Supervised Learning

Semi-Supervised learning is a type of Machine Learning algorithm that lies between Supervised and Unsupervised machine learning. It represents the intermediate ground between Supervised (With Labelled training data) and Unsupervised learning (with no labelled training data) algorithms and uses the combination of labelled and unlabelleddatasets during the training period.

Although Semi-supervised learning is the middle ground between supervised and unsupervised learning and operates on the data that consists of a few labels, it mostly consists of unlabelled data. As labels are costly, but for corporate purposes, they may have few labels. It is completely different from supervised and unsupervised learning as they are based on the presence & absence of labels.

To overcome the drawbacks of supervised learning and unsupervised learning algorithms, the concept of Semi-supervised learning is introduced. The main aim of semi-supervised learning is to effectively use all the available data, rather than only labelled data like in supervised learning. Initially, similar data is clustered along with an unsupervised learning algorithm, and further, it helps to label the unlabeled data into labelled data. It is because labelled data is a comparatively more expensive acquisition than unlabeled data.

Reinforcement Learning

Reinforcement learning works on a feedback-based process, in which an AI agent (A software component) automatically explore its surrounding by hitting & trail, taking action, learning from experiences, and improving its performance. Agent gets rewarded for each good action and get punished for each bad action; hence the goal of reinforcement learning agent is to maximize the rewards.

In reinforcement learning, there is no labelled data like supervised learning, and agents learn from their experiences only.

The reinforcement learning process is similar to a human being; for example, a child learns various things by experiences in his day-to-day life. An example of reinforcement learning is to play a game, where the Game is the environment, moves of an agent at each step define states, and the goal of the agent is to get a high score. Agent receives feedback in terms of punishment and rewards.

Due to its way of working, reinforcement learning is employed in different fields such as Game theory, Operation Research, Information theory, multi-agent systems.

Regression Analysis in Machine learning

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as temperature, age, salary, price, etc.

- C.3
- C19
- C16
- C7

Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables. It is mainly used for prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables.Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables.. It is mainly used for prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables.

In Regression, we plot a graph between the variables which best fits the given datapoints, using this plot, the machine learning model can make predictions about the data. In simple words, "Regression shows a line or curve that passes through all the datapoints on target-predictor graph in such a way that the vertical distance between the datapoints and the regression line is minimum." The distance between datapoints and line tells whether a model has captured a strong relationship or not.

Some examples of regression can be as:
- Prediction of rain using temperature and other factors
- Determining Market trends
- Prediction of road accidents due to rash driving.

Terminologies Related to the Regression Analysis:

Dependent Variable: The main factor in Regression analysis which we want to predict or

understand is called the dependent variable. It is also called target variable

Independent Variable: The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variable, also called as a predictor.

Outliers: Outlier is an observation which contains either very low value or very high value in comparison to other observed values. An outlier may hamper the result, so it should be avoided.

Multicollinearity: If the independent variables are highly correlated with each other than other variables, then such condition is called Multicollinearity. It should not be present in the dataset, because it creates problem while ranking the most affecting variable.

Underfitting and Overfitting: If our algorithm works well with the training dataset but not well with test dataset, then such problem is called Overfitting. And if our algorithm does not perform well even with training dataset, then such problem is called underfitting.

Why do we use Regression Analysis?

As mentioned above, Regression analysis helps in the prediction of a continuous variable. There are various scenarios in the real world where we need some future predictions such as weather condition, sales prediction, marketing trends, etc., for such case we need some

technology which can make predictions more accurately.

So for such case we need Regression analysis which is a statistical method and used in machine learning and data science.

Below are some other reasons for using Regression analysis:

- Regression estimates the relationship between the target and the independent variable.
- It is used to find the trends in data.
- It helps to predict real/continuous values.
- By performing the regression, we can confidently determine the most important factor, the least important factor, and how each factor is affecting the other factors.

Types of Regression

There are various types of regressions which are used in data science and machine learning. Here we are discussing some important types of regression which are given below:

- **Linear Regression**

40

- **Logistic Regression**
- **Polynomial Regression**
- **Support Vector Regression**
- **Decision Tree Regression**
- **Random Forest Regression**
- **Ridge Regression**
- **Lasso Regression**

3.7.1 Logistic Regression:

- Logistic regression is another supervised learning algorithm which is used to solve the classification problems. In classification problems, we have dependent variables in a binary or discrete format such as 0 or 1.
- Logistic regression algorithm works with the categorical variable such as 0 or 1, Yes or No, True or False, Spam or not spam, etc.
- It is a predictive analysis algorithm which works on the concept of probability.
- Logistic regression is a type of regression, but it is different from the linear regression algorithm in the term how they are used.
- Logistic regression uses sigmoid function or logistic function which is a complex cost function. This sigmoid function is used to model the data in logistic regression. The function can be represented as:

Regression Analysis in Machine learning

f(x)= Output between the 0 and 1 value.

x= input to the function

e= base of natural logarithm.
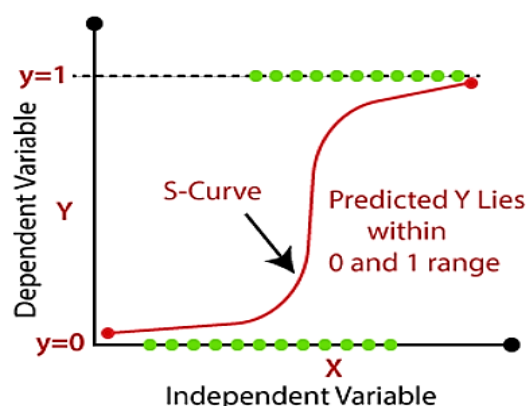
When we provide the input values (data) to the function, it gives the S-curve as follows:

41

Fig.3.7.1.1 Logistic regression

It uses the concept of threshold levels, values above the threshold level are rounded up to 1, and values below the threshold level are rounded up to 0.

There are three types of logistic regression:

- **Binary(0/1, pass/fail)**
- **Multi(cats, dogs, lions)**
- **Ordinal(low, medium, high)**

Logistic Function (Sigmoid Function):

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.

It maps any real value into another value within a range of 0 and 1.The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

Assumptions for Logistic Regression:

- The dependent variable must be categorical in nature.
- The independent variable should not have multi-collinearity.

Logistic Regression Equation:

- The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:
- We know the equation of the straight line can be written as:

$y=b_0+b_1x_1+b_2x_2+.................+bnxn$

- In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$y/1-y$; 0 for y=0, and infinity for y=1

- But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$\log[y/1-y]=b_0+b_1x_1+b_2x_2+b_3x_3+........b_nx_n$

The above equation is the final equation for Logistic Regression

Why Logistic Regression ?

➢ Based on the problem statement, we need a predictive model that can do a binary classification or predict Yes/No or 1/0 type of output variable.

➢ One predictive model commonly implemented for binary classification and prediction of binary outcome is Logistic Regression.

➢ Logistic regression is a binary classification algorithm belonging to the generalized linear regression model.

➢ It can also be used to solve problems with more than 2 classes.

➢ It is possible to use logistic regression to create a model using the customer churn data and use it to predict if a particular customer of a set of customers will discontinue the service.

➢ For example, one of the variables in the data is can be the "annual income". Another variable is the "gender" of the customer.

➢ The outcome of the logistic regression function will tell us how income and or gender determine the probability of service discontinuation by the customer.

Advantages of logistic regression:

• Logistic regression is easier to implement, interpret, and very efficient to train.

• It makes no assumptions about distributions of classes in feature space.

• It can easily extend to multiple classes(multinomial regression) and a natural probabilistic view of class predictions.

• It not only provides a measure of how appropriate a predictor(coefficient size)is, but also its direction of association (positive or negative).

• It is very fast at classifying unknown records.

• Good accuracy for many simple data sets and it performs well when the dataset is linearly separable.

• It can interpret model coefficients as indicators of feature importance.

• Logistic regression is less inclined to over-fitting but it can overfit in high dimensional datasets.

• One may consider Regularization (L1 and L2) techniques to avoid over-fittingin these scenarios.

CHAPTER 4
METHODOLOGY

4.1 PROJECT DESCRIPTION

- **C16,C17,C3,C19**

The primary objective of the customer churn predictive model is to retain customers at the highest risk of churn by proactively engaging with them. Within the current state of customer service, it is no longer enough for organizations to only react to churn. Prediction and prevention of customer churn requires a holistic approach to advanced analytics that provides further understanding into the customer journey. Organizations that effectively utilize customer journey analytics to reduce churn will not only improve the customer experience but also differentiate their brand from competitors.
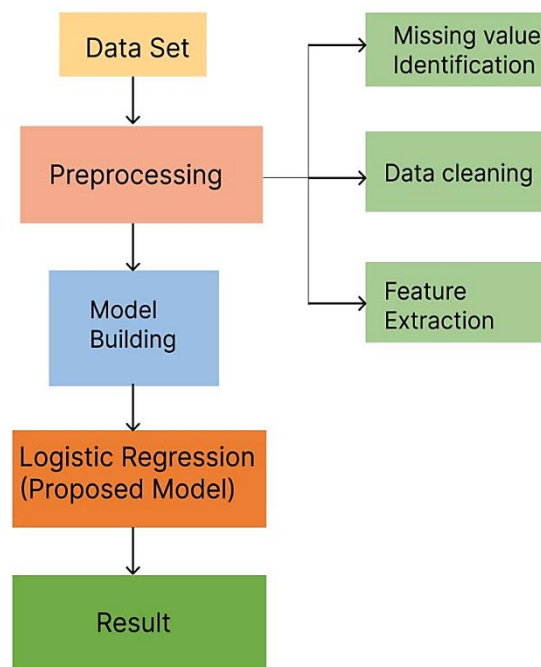
Fig 4.1.1 Flowchart of proposed System

4.2 Dataset

In this Study, a bank data is considered where a huge number of customers are leaving the bank. Almost 10000 records of the bank, collected from Kaggle repository, are going to help the model to investigate and predict which of the customers are about to leave the bank soon. To test and evaluate the features the total dataset sliced into two subsets, training and testing dataset. Training dataset can be used to define the statistical model and the testing dataset can be used to predict the result and calculation of accuracy metrics for determining the model accuracy.For validation of the model accuracy of the classifier is calculated on the basis Confusion Matrix:

- True positive (TP): correctly predicted positive cases.
- False negative (FN): positive cases wrongly predicted as negative.
- False positive (FP): negative cases wrongly predicted as positive.
- True negative (TN): negative cases correctly predicted.

4.3 Data sampling

The dataset set containing 10000 customers data in 14 columns, those are customerid, surname, credit score, gender, age, balance, geography, has credit card, Tenure, Num of Product, Is Active Member, Estimated Salary and exited (the value indicates 1 when the customer churned).

4.4Pre-Processing

In this phase different prepossessing techniques like handling the missing values, data cleaning and feature extraction process have been performed. To identify the missing values in the dataset imputation technique will be used to impute the blank and null values. Noisy data, irrelevant attributes are removed. Those attributes that are not so much important areremoved for model building . Finally for determining the performance of predictive models , feature extraction plays an important role for correct prediction. Here are some important features with description that can be useful for model construction as shown in above Fig.

5Attributes considered from dataset

- **C16,C17,C3,C19**

Sl.No	Parameter	Description
1	RowNumber	Corresponds to the record(row) number.
2	CustomerId	Customer ID is unique for each customer.
3	CreditScore	Can have an effect on customer churn, since a customer with a higher credit score is less likely to leave the bank.
4	Age	This is certainly relevant, since older customers are less likely to leave their bank than younger ones.
5	Tenure	Refers to the number of years that the customer has been a client of the bank. Normally, older clients are more loyal and less likely to leave a bank.
6	Balance	Also a very good indicator of customer churn, as people with higher balance in their accounts are less likely to leave the bank compared to those with lower balances.
7	NumOfProducts	Refers to the number of products that a customer has purchased through the bank.
8	HasCrCard	Denotes whether a customer has credit card.
9	IsActiveMember	Active customers are less likely to leave the bank.
10	EstimatedSalary	People with low salaries are more likely to leave the bank compared to those with higher salaries.
11	Exited	Whether or not the customer left the bank.

In below Fig4.5.2, box plots of some important attributes are given. When it comes to the distribution of all data points over mean, box plots are used to identify the median and also respective qualities inwell-structured manner.
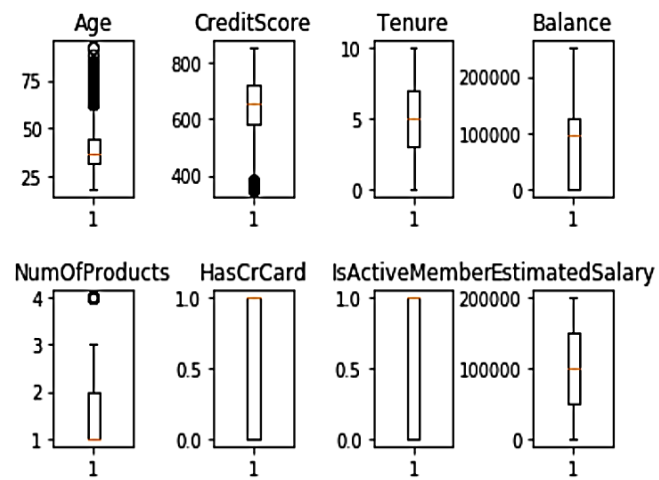
Fig 4.5.2Box-Plots of Important Attributes

Fig 5.3 Data Score

SAMPLE CODE

Introduction

Coding is the process of designing, writing, testing, debugging, and maintaining the source code of computer programs. This source code is written in one or more programming languages. The purpose of programming is to create a set of instructions that computers use to perform specific operations or to exhibit desired behaviors. The process of writing source code often requires expertise in many different subjects, including knowledge of the application domain, specialized algorithms, and formallogic.

Coding
Front-End Template(Main.html):-

```html
<!DOCTYPE html>
<html lang="en">

<head>
<meta charset="UTF-8">
<meta name="viewport" content="width=device-width, initial-scale=1.0">
<title>Customer Churn Prediction</title>
<meta name="viewport" content="width=device-width, initial-scale=1">
<link rel="stylesheet"
href="https://stackpath.bootstrapcdn.com/bootstrap/4.5.0/css/bootstrap.min.css"
     integrity="sha384-
9aIt2nRpC12Uk9gS9baDl411NQApFmC26EwAOH8WgZl5MYYxFfc+NcPb1dKGj7Sk"
crossorigin="anonymous">
<link href="https://fonts.googleapis.com/css?family=Ubuntu" rel="stylesheet">
<link rel="stylesheet" href="static/css/style.css">
<script src="https://ajax.googleapis.com/ajax/libs/jquery/3.5.1/jquery.min.js"></script>
<script src="https://cdn.plot.ly/plotly-latest.min.js"></script>
<script src="https://cdn.jsdelivr.net/npm/@popperjs/core@2.11.5/dist/umd/popper.min.js"
     integrity="sha384-
Xe+8cL9oJa6tN/veChSP7q+mnSPaj5Bcu9mPX5F5xIGE0DVittaqT5lorf0EI7Vk"
crossorigin="anonymous"></script>
<script src="https://cdn.jsdelivr.net/npm/bootstrap@5.2.0-beta1/dist/js/bootstrap.min.js"

integrity="sha384kjU+l4N0Yf4ZOJErLsIcvOU2qSb74wXpOhqTvwVx3OElZRweTnQ6d31fX
EoRD1Jy"
crossorigin="anonymous"></script>
```

```html
<link href="https://cdn.jsdelivr.net/npm/bootstrap@5.2.0-beta1/dist/css/bootstrap.min.css" rel="stylesheet"
    integrity="sha384-
0evHe/X+R7YkIZDRvuzKMRqM+OrBnVFBL6DOitfPri4tjfHxaWutUpFmBp4vmVor"
crossorigin="anonymous">
<script src="https://cdn.jsdelivr.net/npm/bootstrap@5.2.0-beta1/dist/js/bootstrap.bundle.min.js"
    integrity="sha384-
pprn3073KE6tl6bjs2QrFaJGz5/SUsLqktiwsUTF55Jfv3qYSDhgCecCxMW52nD2"
crossorigin="anonymous"></script>

</head>

<body>
<div class="container">
<br><br>
   {% if mes%}
<div class="alert alert-{{col}} alert-dismissible fade show" role="alert">
<strong>{{mes}}</strong>.
<button type="button" class="btn-close" data-bs-dismiss="alert" aria-label="Close"></button>
</div>

   {%endif%}
</div>
<div class="wrapper">
<nav class="navbar navbar-default">
<div class="container-fluid">
<div class="navbar-header">
<a class="navbar-brand" href="index.html">

</a>
</div>
</div>
</nav>
<div class="container">
<div class="row">
<div class="col-sm-3"></div>
<div class="col-sm-6"><b>Enter User Information in the Form Below</b></div>
</div>
</div>
<div class="container">
<div class="row">
<div class="col-sm-3"></div>
```

```html
<div class="col-sm-3"><br>
<form action="{{ url_for('predict')}}" method="post">

<label for="creditScore">Credit Score:</label><br>
<input type="number" min="300" max="850" required id="CreditScore"
                name="creditScore" placeholder="Range 300 to 850"><br>
<label for="geography">Country:</label><br>
<select name="geography" required id="Geography">
<option disabled selected value> -- select -- </option>
<option value="France">France</option>
<option value="Germany">Germany</option>
<option value="Spain">Spain</option>
</select><br>
<label for="gender">Gender:</label><br>
<select name="gender" required id="gender">
<option  selected value> -- select -- </option>
<option value="Female">Female</option>
<option value="Male">Male</option>
</select><br>
<label for="age">Age:</label><br>
<input type="number" required min="18" id="Age" name="age" placeholder="minimum
18"><br>
<label for="tenure">Years as Bank Customer:</label><br>
<input type=" number" required="required" id="Tenure" name="tenure"><br>
<label for="balance">Account Balance:</label><br>
<input type="number" step="any" required id="Balance" name="balance"><br>
</div><br>
<div class="col-sm-3"><br>

<label for="numofproducts">Number of Products Used:</label><br>
<select name="numofproducts" required id="NumofProducts">
<option disabled selected value> -- select -- </option>
<option value="1">1</option>
<option value="2">2</option>
<option value="3">3</option>
<option value="4">4</option>
<option value="5">5</option>
</select><br>

<label for="hascrcard">Has an Active Credit Card:</label><br>
<select name="hascrcard" required id="HasCrCard"><br>
```

50

```html
<option disabled selected value> -- select -- </option>
<option value="1">Yes</option>
<option value="0">No</option>
</select><br>
<label for="isactivemember">Is an Active Bank Member:</label><br>
<select name="isactivemember" required" id="isActiveMember">
<option disabled selected value> -- select --</option>
<option value="1">Yes</option>
<option value="0">No</option>
</select>
<br>
<label for="estimatedsalary">Estimated Salary:</label><br>
<input type="number" min="0" step="any" required id="EstimatedSalary"
            name="estimatedsalary"><br>



<br>
<br>
<button type="submit" class="btnbtn-success">Submit</button></form>
</div>
</div>
</div>


<br><br>

<br>
<!--<div class="container">
<div class="row">
<div class="col-sm-3"></div>
<div class="col-sm-6">
<div class="table-responsive">
<table class="table" border="2">
<thead>
<tr>
<th>Model Name</th>
<th>Prediction</th>
</tr>
</thead>
```

```
</table>
</div>
</div>
</div>
</div> -->


</body>
```

Back End Flask Code (app.py):-

```python
import numpy as np

from flask import Flask, request, jsonify, render_template

app = Flask(__name__)

# Load saved models

#dt_model = joblib.load('models/nate_decision_tree.sav')

# dl_model = joblib.load('models/imblearn_pipeline.sav')

# dl_model.named_steps['kerasclassifier'].model = load_model('models/keras_model.h5')

# newmodel=pickle.load(open('nate_logistic_regression.sav', 'rb'))

# # knn_model = joblib.load('models/nate_knn.sav')

# lr_model = joblib.load('nate_logistic_regression.sav')

# rf_model = joblib.load('models/nate_random_forest.sav')

# svm_model = joblib.load('models/SVM_model.sav')

# xgb_model = joblib.load('models/XGBoost_model.sav')

# Dictionary of all loaded models

# loaded_models = {

#     'lr': lr_model

# }

# Function to decode predictions

def decode(pred):

    if pred == 1: return 'Customer Exits'

    else: return 'Customer Stays'

@app.route('/')
```

```python
def home():
    # Initial rendering
    result = [



        #{'model':'Deep Learning', 'prediction':' '},
#{'model': 'K-nearest Neighbors', 'prediction': ' '},
        {'model': 'Logistic Regression', 'prediction': ' '}
#{'model': 'Random Forest', 'prediction': ' '},
#{'model': 'SVM', 'prediction': ' '},
        ]
# Create main dictionary
    # maind = {}
    # maind['customer'] = {}
    # maind['predictions'] = result
return render_template('index.html')
@app.route('/predict', methods=['POST'])
def predict():
  # List values received from index
    values = [x for x in request.form.values()]
# new_array - input to models
    # new_array = np.array(values).reshape(1, -1)
    # print(new_array)
    print(values)


    # Key names for customer dictionary custd
    cols = ['CreditScore',
        'Geography',
        'Gender',
        'Age',
        'Tenure',
```

```python
        'Balance',

        'NumOfProducts',

        'HasCrCard',

        'IsActiveMember',

        'EstimatedSalary']
    f=1
    if (values[8]=="0" or 300<=int(values[0])<=400 or int(values[3])>=60):
        f=0
     # Create customer dictionary
    def create() :
custd = {}
        for k, v in  zip(cols, values):
custd[k] = v


    # Convert 1 or 0 to Yes or No



    # Loop through 'loaded_models' dictionary and
    # savepredictiond to the list
predl = []
    # for m in loaded_models.values():
    #     predl.append(decode(m.predict(new_array)[0]))



    if f==1 :
        value="Customer Stays in the bank..!"
        col="success"
    else:
        value="Customer will left the bank..!"
```

```python
            col="danger"

        # Create main dictionary

        return render_template('index.html',mes=value

    ,col=col

if __name__ == "__main__":

        app.run(debug=True)
```

CHAPTER 5
EXPERIMENTAL RESULTS

The data set contains details of a bank's customers and the target variable is a binary variable reflecting the fact whether the customer left the bank (closed his account) or he continues to be a customer.It consists of 10,000 records with demographic and bank history information from customers from three countries, France, Germany and Spain.Continuing with splitting the data into separate training and test sets. 30% ofobservations will be set aside for the test setthe rest, 70%, will be used as the training set.

In the output, it displaysan application containing the customer details to be filled.The first output is the application given input by the user with the appropriate details to predict the result.

Web application HTML page, also known as a web app, provides point-and-click control to the web application. We had developed GUI in HTML using Flask as the backend. We had used various components provided by the HTML Tags like Buttons, input files.ThecompleteGUIDEwindowforoursystemMain page is shown in fig.5.2.1. Our system consists of aapplication. The main window is shown infig.5.2.1.

The results predicted are customer stays in bank and customer will left are shown in fig.45.2.3 and fig 5.2.4 the results are predicted for a bank customer
The proposed Logistic regression was able to predict theCustomer churn in dataset as stays in bank and left the bank . This logistic regression predicts the customer churn with an accuracy of 71.00%.
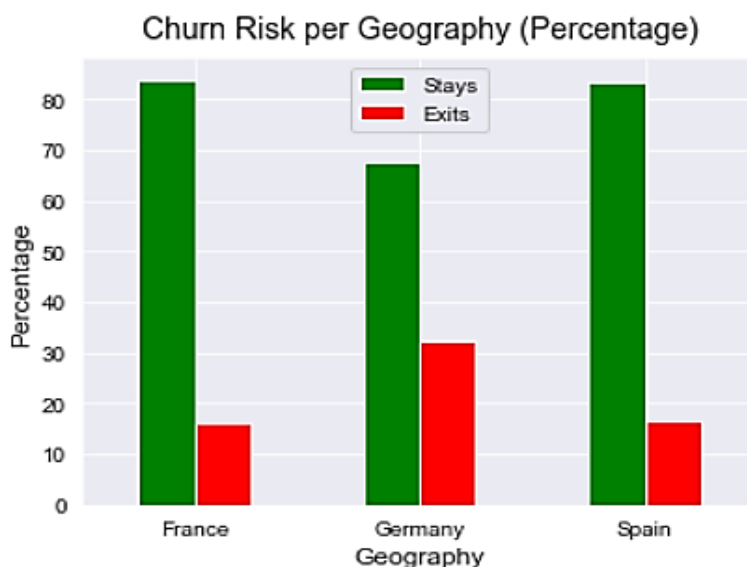
Fig 5.1(a) Churn distribution by Geography

Evaluate models using metrics

Evaluating the performance of the model using different metrics is integral to every data science project. Here is what you have to keep an eye on:

Accuracy

Accuracy is a metric for how much of the predictions the model makes are true. The higher the accuracy is, the better. However, it is not the only important metric when you estimate the performance.

$$Accuracy \equiv \frac{True\ Positives + True\ Negatives}{True\ Positives + False\ Positives + True\ Negatives + False\ Negatives}$$

Loss

Loss describes the percentage of bad predictions. If the model's prediction is perfect, the loss is zero; otherwise, the loss is greater.

Precision

The precision metric marks how often the model is correct when identifying positive results. For example, how often the model diagnoses cancer to patients who really have cancer.

$$Precision \equiv \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Recall

This metric measures the number of correct predictions, divided by the number of results that should have been predicted correctly. It refers to the percentage of total relevant results correctly classified by your algorithm.

$$Recall \equiv \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Confusion matrix

A confusion matrix is an $N \times N$ square table, where N is the number of classes that the model needs to classify. Usually, this method is applied to classification where each column represents a label. For example, if you need to categorize fruits into three categories: oranges, apples, and bananas, you draw a 3×3 table. One axis will be the actual label, and the other will be the predicted one.
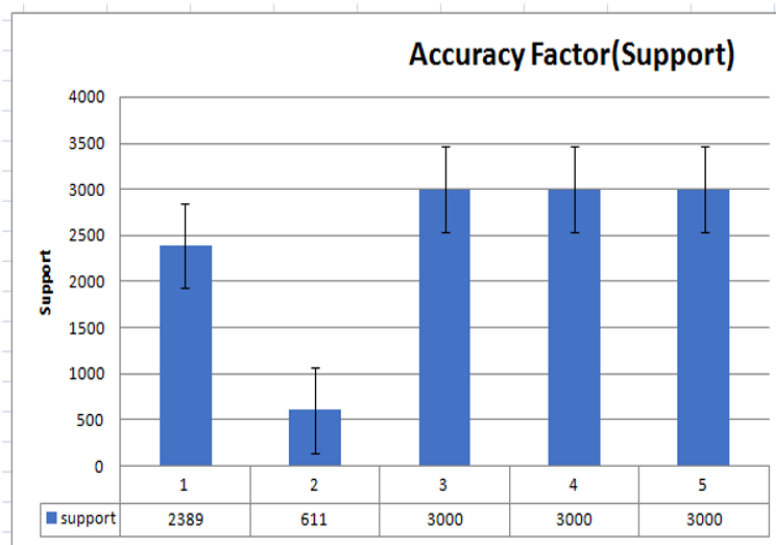
Fig 5.1(b) Accuracy Factor(support)
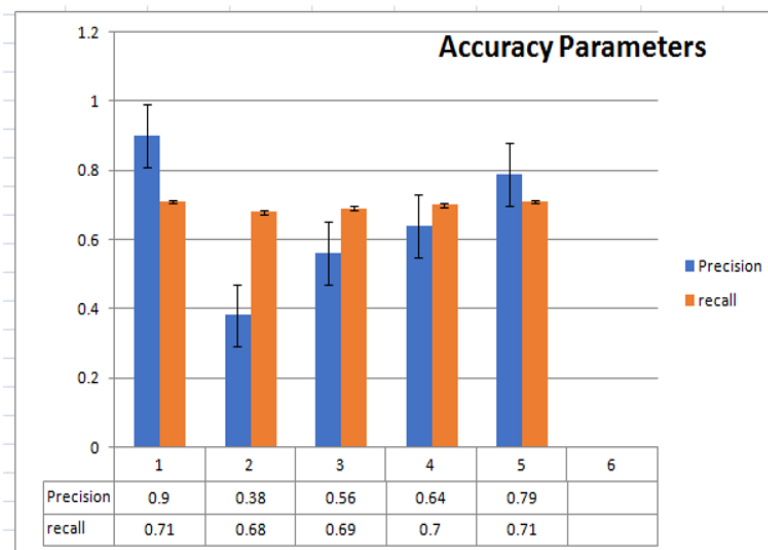
Fig 5.1(c) Accuracy Parameter

Fig 5.1(d) Accuracy Factor(F1-Score)

5.2 OUTPUT SCREENSHOTS

Fig 5.2.1 Loading Main Page (HTML) of our proposed system

Fig 5.2.2Shows an error message.

Fig5.2.3 Displaysthe output as Customer Stays in the bank

Fig 5.2.4 Displays the output as customer will left the bank

60

CHAPTER 6

CONCLUSION

In this project, we proposed an algorithm that can predict Customer Churn and gives the outcomes, contrasting other used techniques. This algorithm which deals with the large amount of datasets. Based on the given customer details it predicts whether the customer will stays in the bank or left the bank.. It gives an accuracy of 71.00 %. The proposed system would be helpful in predicting if a customer will stays or left the bank. And also helps in the growth of a company.

FUTURE ENHANCEMENTS

Further, we need to extract some features for this bank customer churn prediction project. That is we will provide a user friendly interface and in future we are going to add signup and login pages for the user to provide the security that only the authenticated user can access the application and can see the result of churn prediction. In future we are in an idea to improve the user interaction that allows the user to upload their own dataset by themselves only in the application and can see the prediction of each customer in the application itself. We also planning to do the same application which can produce the churn prediction not only for bank customers but also for all kinds of business firms. This application will be a great help for the businesses to improve their business and to retain their customers by doing their own strategy of promotions.

.

CHAPTER 7

REFERENCES

Here, are some of the references for the project from various sites, journals, papers,etc..

1. Seo D, Ranganathan C, Babad Y. Two-level model of customer retention in the US mobile telecommunications servicemarket. Telecommun Policy 2008; 32 (3):182–96.

2. G. M. Weiss, "Mining with rarity: a unifying framework," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 7–19, 2004.

3. J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," Expert Systems with Applications, vol. 36, no. 3, pp. 4626–4636, 2009.

4. M. T. K. a. B. B. B. Huang, "Customer churn prediction in telecommunications," Expert Systems with Applications, vol. 39, no. 1, pp. 1414-1425, 2012.

5. W.-H. a. C. K. C. a. Y. X. Au, "A novel evolutionary data mining algorithm with applications to churn prediction," IEEE transactions on evolutionary computation, vol. 7, no. 6, pp. 532-545, 2003.

6. M.A.H. Farquad, Vadlamani Ravi, S. Bapi Raju "Churn prediction using comprehensible support vector machine: An analytical CRM application", Applied Soft Computing 19 (2014) 31–40.

7. Wouter Verbeke, David Martens, Christophe Mues, Bart Baesens "Building comprehensible customer churn prediction models with advanced rule induction techniques", Expert Systems with Applications 38 (2011) 2354–2364.

8. Ning Lu, Hua Lin, Jie Lu, Guangquan Zhang "A Customer Churn Prediction Model in Telecom Industry Using Boosting", IEEE Transactions on Industrial Informatics, vol. 10, no. 2, may 2014.

9. Benlan He, Yong Shi, Qian Wan, Xi Zhao "Prediction of customer attrition of commercial banks based on SVM model", Proceedings of 2nd International Conference on Information Technology and Quantitative Management (ITQM), Procedia Computer Science 31 (2014) 423 – 430.

10. https://www.kaggle.com/filippoo/deep-learning-az-annn

11. Chu, B. H., Tsai, M. S., and Ho, C. S., "Towards a hybrid data mining model for customer retention", Knowledge-Based Systems, 20, 2007, pp. 703–718.

12. Shui Hua Han a, ShuiXiuLu a, Stephen C.H. Leung., "Segmentation of telecom customers based on customer value by decision tree model", Expert Systems with Applications, 39, 2012, 3964–3973

13. https://learnpython.com/ blog/python-customer-churn-prediction/

14. https://neptune.ai/blog/how-to-implement-customer-churn-prediction

15. Koen W. De Bock, Dirk Van den Poel, "An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction", Expert Systems with Applications 38 (2011) 12293–12301.

16. H. Lee, Y. Lee, H. Cho, K. Im, Y.S. Kim, "Mining churning behaviors and developing retention strategies based on a partial least squares (PLS) model", Decision Support System 52 (2011) 207–216.

17. Yaya Xie, Xiu Li, E.W.T. Ngai, Weiyun Ying, "Customer churn prediction using improved balanced random forests", Expert Systems with Applications 36 (2009) 5445–5449.