Loan Approval Predictor Using Data Science And Machine Learning

Project report submitted in partial fulfillment of the requirement for the degree of Bachelor of Technology

in

Computer Science and Engineering

By

Shail Vatsal Vashist (161480)

Under the supervision of

Dr. Aman Sharma

to

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology Waknaghat, Solan-173234, Himachal Pradesh

Declaration

I hereby declare that the work presented in this report entitled “Loan Approval Predictor” in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering** submitted in the department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from January 2020 to June 2020 under the supervision of Dr. Aman Sharma (Assistant Professor(Grade-II), Computer Science & Engineering and Information Technology).

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Shail Vatsal Vashist

Shail Vatsal Vashist
161480

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Aman Sharma

Dr. Aman Sharma
Assistant Professor(Grade-II)
Computer Science & Engineering and Information Technology
Dated:

ACKNOWLEDGEMENT

First of all, I would like to express my deep gratitude to my project guide **Dr. Aman Sharma***(Assistant Professor (Grade II), Department of Computer Science Engineering & IT)* for providing me an opportunity to work under his supervision and guidance. His constant encouragement at every step was a precious asset during the project work.

I express my deep appreciation and sincere thanks to **Dr. Hemraj Saini** (*Department of Computer Science Engineering & IT*) for providing all kinds of possible help and encouragement during my project work.

I am thankful to the faculty and staff of *Department of Computer Science Engineering & Information Technology*, Jaypee University of Information Technology for providing me with all the facilities required for experimental work.

I would like to thank my family for their continuous support and motivation. Finally I would like to thank those who directly or indirectly helped me in completing this project.

TABLE OF CONTENT

CAPTION **PAGE NO.**

LIST OF FIGURES

ABSTRACT

Lately people depend on bank loans to meet their wishes. The fee of loan packages will increase with a very rapid speed in current years. Risk is constantly involved in approval of loans. The banking officials are very acutely aware of the price of the mortgage quantity by its customers. Even after taking lot of precautions and analyzing the mortgage applicant information, the mortgage approval choices are not continually correct. There is need of automation of this system so that loan approval is much less risky and incur less loss for banks.

Since it is a major activity for the banks,to identify whether a loan of the desired amount should be approved to the applicant or not,the Computer Science is capable of making such a system using Artificial Intelligence,which can make this tough decision accurately and quickly.

Using data science,which is responsible to deal with the large amount of data efficiently,and some algorithms of Machine Learning,a prediction system is made,which,on the basis of some training data set,is capable of identifying if the loan applicant is ideal for the loan approval or not.

Machine Learning algorithms like Decision Tree,Logistic Regression,Random Forest,etc. are used for the analysis.These are efficient algorithms that are followed for data analysis and prediction making.

The system will look into some basic information of the applicant such as his/her profession,age,gender,marital status,etc.,and after analyzing all this information,using visualization and machine learning algorithms,it will come to a decision.

CHAPTER 1:INTRODUCTION

1.1 General introduction:

Information Science is an itemized investigation of the progression of data from the enormous measures of information present in an association's archive. It includes acquiring significant bits of knowledge from crude and unorganised information that is prepared through diagnostic, programming, and business aptitudes.

In a world that is progressively turning into an advanced space, associations influence zetta bytes and yotta bytes of organized and unstructured information daily . Advancing advances have empowered cost reserve funds and more brilliant extra rooms to store basic information. As of now, inside the business, there's a gigantic requirement for talented and approved Data Scientists. They are among the most generously compensated experts in the IT business. As indicated by Forbes, 'the best employment in America is of a Data Scientist with a normal yearly pay of $110,000'. Just a few people can possibly process it and determine important bits of knowledge out of it.

Organizations are overwhelmed with monster measures of information. Accordingly, it's essential to comprehend what to attempt to with this detonating information and the best approach to use it.

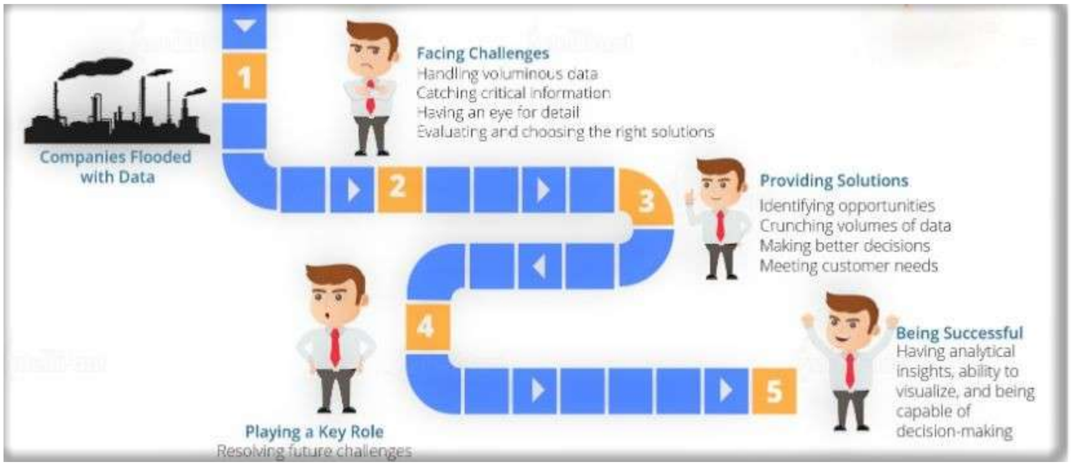

Companies Flooded with Data
1
Facing Challenges
Handling voluminous data
Catching critical information
Having an eye for detail
Evaluating and choosing the right solutions
2
3
Providing Solutions
Identifying opportunities
Crunching volumes of data
Making better decisions
Meeting customer needs
4
Playing a Key Role
Resolving future challenges
5
Being Successful
Having analytical insights, ability to visualize, and being capable of decision-making

Fig 1: Data flow in the industry

It is here, the idea of Data Science comes into the image. Data science along with arithmetics and organisational information allows the organisation to explore approaches to:

- Reduce costs as much as possible
- Explore the new market and get entry into it
- Hold and open to the various demographies
- Get aware of the effective marketing campaign

Come up with the new product or the service that will effect the market.

Furthermore, the rundown is perpetual!

Thus, paying little mind to the business vertical, Information Science is probably going to assume a key job in your association's prosperity.

Take a gander at the beneath infographic, and you will have the option to see how Data Science is making its impression:

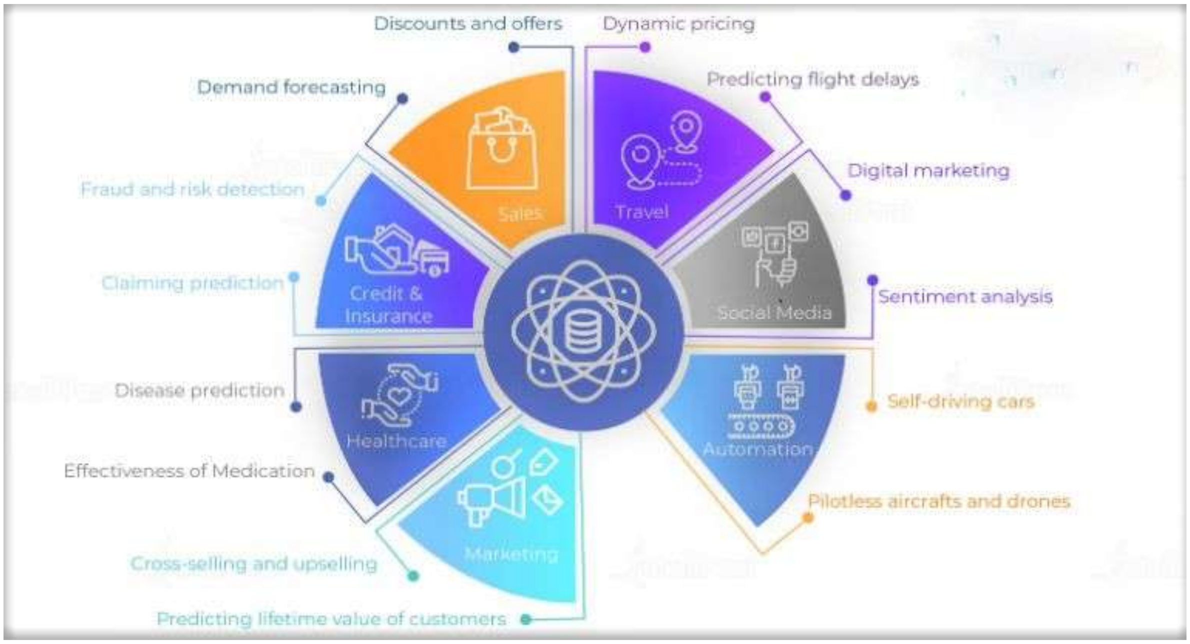


Fig 2:Benefits of using data science techniques

In recent years, *data science* emerged as a new and important discipline.It can also be considered a mixture or combination of some technologies such as data mining and statistics,databases,etc. Existing approaches need to be combined to turn abundantly available data into value for individuals, organizations, and society.

Science means the knowledge acquired by systematic study.Data science basically focused on the data which even may be very large in size,and handling of this data using modern technology.A very big amount of data is generated in the companies every second.This data definitely needs to be handled properly and efficiently.If not handled properly,it could prove dangerous for the company.

Hence every company has the required number of data scientists,who work constantly handling the records which are essential to the company.These data scientists are very heavily paid.The data scientists know all the statistics and have the knowledge of the data which makes itb easy for them to handle it.

Traditional database techniques are not acceptable for knowledge discovery due to the fact they're optimized for instant get right of entry to and summarization of statistics, provided what the user wants to ask, or a question, now not discovery of patterns in big swaths of statistics while customers lack a wellformulated question. Unlike database querying, which asks “What records satisfies this sample (query)?” discovery asks “What patterns fulfill this statistics?” Specifically, our concern is finding interesting and sturdy patterns that fulfill the facts, where “interesting” is usually something sudden and “robust” is a pattern expected to arise within the destiny.

Machine learning uses data science and makes it feasible to generate such models which are able to make accurate predictions,classisfy things into categories,interpretation of images,etc.This makes the machines and robots intelligent as they can learn on their own and there is no need to worry about their accuracy.In today’s world,where eveything is getting automated,there surely is need of the mechanisms which can be eaily trusted for their accuracy.Machine learning along with data science makes this possible.Machine learning is already helping several industries and is well praised for easing the human effort.

Machine learning is a subset of artificial intelligence (AI) wherein algorithms research by way of instance from historical statistics to expect results and uncover patterns which humans cannot spot easily. For instance,ML can screen clients who are probably to churn, possibly fraudulent coverage claims,etc. While ML has been round since the Fifties, latest breakthroughs in low-value compute assets like cloud storage, less complicated collection of data, and the proliferation of information science have made it very a good deal “the next huge thing” in commercial enterprise analytics.

The ML algorithms learns through instance, and then users practice the ones self-gaining knowledge of algorithms to find insights, determine relationships, and make predictions about future tendencies.ML has practical implications throughout industry sectors, together with healthcare, insurance,advertising, manufacturing,etc. When implemented successfully,ML lets in groups to find most appropriate solutions to realistic issues, which results in actual, tangible commercial enterprise value.

Some Top Industries Using Data Science:

Google is so far one of the greatest organization that is on an employing binge for prepared Data Scientists. Since Google is for the most part determined by Data Science nowadays, it offers probably the best datum Science compensations to its representatives.Amazon is a worldwide web based business and distributed computing monster that is recruiting Data Scientists on an exceptionally huge scope. They need Data Scientists to discover client mentality and improve the topographical reach of both web based business and cloud spaces, among different business-driven objectives.

Fig 3:Companies using data science technologies and getting benefits

1.2 Problem Statement:

Loan approval is a completely important procedure for banking businesses. The system approve or reject the mortgage applications. Recovery of loans is a first-rate contributing parameter in the economic statements of a bank. It may be very hard to are expecting the possibility of fee of loan through the purchaser. In current years many researchers worked on mortgage approval prediction structures.ML techniques are very useful in predicting consequences for big quantity of information. In this project some ML algorithms like Logistic Regression, Decision Tree,Random Forest,etc are implemented to are expecting the loan approval for customers. The experimental results conclude that the accuracy of Decision Tree ML algorithm is better in comparison to other algorithms.

1.3 Objectives:

Lately people depend on bank loans to meet their wishes. The fee of loan packages will increase with a very rapid speed in current years. Risk is constantly involved in approval of loans. The banking officials are very acutely aware of the price of the mortgage quantity by its customers. Even after taking lot of precautions and analyzing the mortgage applicant information, the mortgage approval choices are not continually correct. There is need of automation of this system so that loan approval is much less risky and incur less loss for banks.

Artificial Intelligence AI is a rising technology. The utility of AI solves many real world troubles. Machine Learning is an AI method which could be very useful in prediction systems.A model is created from a training data. While making the prediction the model that is evolved by way of training algorithm(ML) is used. The ML algorithm trained the machine the usage of a fragment of the statistics available and the remaining data is tested.

Distribution of the loans is the center commercial enterprise part of almost all banks. The main portion the financial institution’s property is immediately came from the profit earned from the loans allotted via the banks. The high objective in banking environment is to invest their property in secure hands.Today many banks/financial corporations approves loan after a regress procedure of verification and validation but nonetheless there's no surety whether the selected applicant is the deserving right applicant out of all applicants. Through this system we are able to are expecting whether or not that specific applicant is safe or not and the whole process of validation of functions is automatic via ML technique. The downside of this model is that it emphasize exclusive weights to each element however in actual existence sometime loan can be accredited on the idea of single strong component only, which is not feasible via this system. Loan Prediction could be very helpful for worker of banks as well as for the applicant also. The goal of this project is to provide brief,quick and easy way to pick out the deserving candidates. It can offer unique benefits to the financial institution. The Loan Prediction System can can mechanically calculate the load of every features participating in mortgage processing and on new test statistics same capabilities are processed with appreciate to their associated weight .A time restriction can be set for the applicant to test whether his/her mortgage may be sanctioned or not. Loan Prediction System allows leaping to specific utility in order that it is able to be take a look at on the basis of priority.

The ML strategies can be implemented on a sample data first after which can be used in making prediction associated selections. This project applied the ML procedures in solving loan approval trouble of banking area.

1.4Methodology:

Since it is a major activity for the banks,to identify whether a loan of the desired amount should be approved to the applicant or not,the Computer Science is capable of making such a system using Artificial Intelligence,which can make this tough decision accurately and quickly.

Using data science,which is responsible to deal with the large amount of data efficiently,and some algorithms of Machine Learning,a prediction system is made,which,on the basis of some training data set,is capable of identifying if the loan applicant is ideal for the loan approval or not.

Machine Learning algorithms like Decision Tree,Logistic Regression,Random Forest,etc. are used for the analysis.These are efficient algorithms that are followed for data analysis and prediction making.

The system will look into some basic information of the applicant such as his/her profession,age,gender,marital status,etc.,and after analyzing all this information,using visualization and machine learning algorithms,it will come to a decision.

CHAPTER-2: LITERATURE SURVEY

2.1 DATA SCIENCE:

Data science is the sphere of study that mixes domain information, programming competencies, and know-how of mathematics and facts to extract significant insights from records. Data scientists follow ML algorithms to numbers, text, pics, video, audio, etc. to generate AI systems to carry out tasks that generally require human intelligence. In turn, those structures generate insights which analysts and users of related field can turn into tangible business value.

Big data is a blanket term for any series of records so large or complicated that it becomes difficult to technique them using traditional records management strategies consisting of, as an instance, the relational database management systems(RDBMS).The extensively adopted RDBMS has lengthy been regarded as a one-length-suits-all solution, but the needs of coping with big records have proven in any other case. Data science involves using techniques to investigate big quantities of information and extract the know-how it carries. You can consider the relationship among big information and data science as being just like the relationship between crude oil and an oil refinery. Data technology and massive facts advanced from facts and traditional facts control but are actually taken into consideration to be different disciplines.

Data science makes it easy to handle big data.

There are data scientists who are professionally sound and are able to handle the big data easily using data science.

2.2 Benefits and uses of data science

Big data and data science are used almost everywhere in both business and noncommercial settings. The variety of use instances is large.Commercial organizations in nearly each industry use big data and data science to gain insights into their customers, tactics, group of workers, completion,and products.Data Science is used by many businesses to offer clients a higher user experience, as well as to cross-sell, up-sell, and customize their offerings.

An example of this is Google AdSense.It generates the advertisements for the users based on their interest and past searches.This makes it easier for the users to get the required items of their interests easily.

Human aid specialists analyse the employees by analysing their moods and behaviours.Relations with the co-workers can also be analysed this way.

Data science is used by financial institutions in prediction of stock markets,decide the risk of lending cash, and discover ways to attract new customers for institution's services.

Maximum trade is nowadays takes place with the help of mechanisms which highly operate on the algorithms of machine learning.These are reliable machines and their outputs can be trustred without any question.

The corporate sectors also uses data science.There are a number of governmental bodies where data scientists have important position as they have to deal with confidential records which are important to the organisation.These records can further be used to gain insights and also in developing applications which are driven with records.

Not only governmental,but there are certain non-governmental organisations also which are responsible to deal with big records.There are a thousands of NGO’s in every country.All these NGO’s have to maintain a number of records,irrespective of their field of work.Hence data science and data scientists are required by these non-governmental organisation also.

Universities use data science in their studies however also to elevate the observe experience of their students. The upward thrust of massive open on line courses (MOOC) produces lots of information, which allows universities to have a look at how this form of learning can complement traditional training.

2.3 Types of Data:

The main categories of data are these:

- Structured
- Unstructured
- Natural language
- Machine-generated
- Graph-based
- Audio, video, and images
- Streaming

2.3.1 Structured data:

Structured data is information that depends on a statistics model and resides in a hard and fast discipline within a report. As such, it’s easy to contain structured records in tables within databases or Excel documents.

SQL (Structured Query Language), is the language which helps us deal with the records contained in the tables within the databases. SQL makes the manipulation of data contained in databases much easier which otherwise would have been a tough task.

2.3.2 Unstructured data:

Unstructured data comprises of data is varying data and hence is not suitable to be used within a data model.Example of unstructured data is email,

There are several structured components in an email like sender name,body,etc. but it is difficult to identify the mails containing the content having same reference sent by a number of people because there are a number of methods to refer to someone.

2.3.3 Natural language data:

Natural language is a unique form of unstructured statistics; it’s processing is tough as it requires know-how of particular data science strategies and linguistics. The natural language processing community has had success in recognition of certain entity,recognition of subject matter, summarization, text completion, and sentiment evaluation, however models trained in a single domain do not generalize nicely to remaining domains.

2.3.4 Machine Generated data:

Machine-generated facts is information which is robotically created by a computerized system,process, software, or different device without human intervention. Machine-generated statistics is turning into a major information useful resource and could maintain to do so.

The analysis of machine information is based on quite scalable gear, due to its high quantity and generation speed. Examples of machine records are web server logs, call records, community event logs, and telemetry.

2.3.5 Graph-based data:

Graph Data may be a confusing term due to the fact any information can be shown in a graph. “Graph” in this case factors to mathematical graph theory. In graph principle, a graph is a mathematical shape to model pair-wise relationships.Graph or network statistics is, in brief, facts that makes a speciality of the relationship or adjacency of items. The graph structures use nodes, edges, and functionalities to visualize and keep graphical facts. Graph-based data is a natural manner to represent social networks, and its structure lets in you to calculate particular metrics consisting of the affect of someone and the shortest direction between humans.

2.3.6 Audio,Image and Video data:

Audio,image, and video are sort of data that pose precise demanding situations to a data scientist. Tasks which might be easy for humans, consisting of spotting objects in photographs, grow to be challenging for computer systems. MLBAM (Major League Baseball Advanced Media) announced in 2014 that they’ll boom video capture to about 7 TB consistent with game for the live, in-sport analytics. High-speed cameras at stadiums will capture actions of players and ball to calculate in real time, for instance, the route taken by a defender relative to two baselines. Recently an organisation called DeepMind succeeded at developing some set of rules that’s able to learning the way to play video video games. The video game screen acts as an input for the algorithm and learns to interpret everything via a complex manner of deep learning. It’s an exceptional feat that triggered Google to buy the enterprise for their very own Artificial Intelligence (AI) development plans. The learning algorithm takes in data because it’s produced via the pc game; it’s streaming information.

2.3.7 Streaming data:

While streaming records can take nearly any of the previous form, it has an additional distinguishing property. The record flows into the machine whilst an event occurs rather than being loaded right into a store in a batch. Although this isn’t truely a one-of-a-kind form of data, we deal with it right here as such due to the fact you want to adapt your process to cope with this form of information.Examples are the “What’s trending” on Twitter, live events like matches and concerts, and the share market.

2.4 The Data Science Process:

The data science process typically consists of six steps:

1. Setting the research goal
2. Retrieving data
3. Data preperation
4. Data exploration
5. Data modeling
6. Presentation and automation

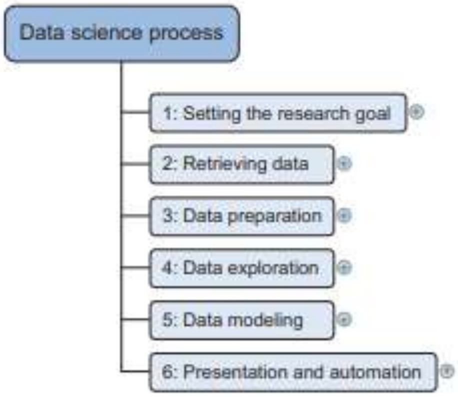


Fig 4:The Data Science Process

2.4.1 Setting the research goal:

Data science is in most cases implemented in the context of a corporation. When the business asks you to carry out a project related to data science and analysis,a project charter will be preapared by one.This charter incorporates statistics including what you're going to research, how the agency advantages from that, what statistics and resources you want, a timetable, and deliverables.

2.4.2 Retrieving data:

The 2nd step is to accumulate statistics.As already stated inside the project charter which records are needed and where to find that data. In this step it is ensured that you can use the statistics to your program, which means checking the existence of, quality, and get access to to the data. Data also can be added by means of third-party businesses and takes many forms starting from Excel spreadsheets to different sorts of databases.

2.4.3 Data preperation:

Collection of data is an errors-susceptible procedure; on this segment you enhance the record quality and get the records ready for use in subsequent steps. This section consists of 3 subphases: Data cleansing removes fake values from a information supply and inconsistencies across sources of records, statistics integration enriches data sources by combining facts from a couple of information sources, and data transformation guarantees that the information is in an appropriate format to be used in the model.

2.4.4 Data exploration:

Data exploration is concerned with building a deeper expertise of records. You try to understand how variables engage with every other, the distribution of the records, and whether there are outliers. To attain this particularly use descriptive statistics, visual techniques, and simple modeling. This step is called Exploratory Data Analysis and regularly is going by the abbreviation EDA.

2.4.5 Data modeling:

In this section models are used, domain knowledge, and insights of the information discovered in the preceding steps to reply the studies query.Pick a way from the fields of statistics,ML, operations studies, and so on. Building a model is an iterative procedure that involves selecting the variables for the model, executing the model, and diagnostics of model.

2.4.6 Presentation and automation:

Finally, you present the outcomes on your business. These outcomes can take many forms, ranging from presentations to analysis reports. Sometimes there would be need to automate the execution of the system because the commercial enterprise will need to use the insights that were acquired in every project or enable an operational technique to apply the final results out of your model.

2.5 MACHINE LEARNING:

ML is an utility of AI that gives systems the ability to learn mechanically and improve from experience without being programmed explicitly. ML centers around the advancement of PC programs that can get to information and use it learn for themselves.

To accomplish ML, specialists create broadly useful calculations that can be utilized on enormous classes of learning issues. At the point when you need to explain a particular undertaking you just need to take care of the calculation progressively explicit information. As it were, you're customizing by model. By and large a PC will utilize information as its wellspring of data and contrast its yield with an ideal yield and afterward right for it. The more information or "experience" the PC gets, the better it becomes at its assigned activity, similar to a human does.

For instance, as a user writes more textual content messages on a mobile phone, the mobile learns the vacabulary that is used quite often in the messages and might expect these words instantly and with high accuracy. In the broader subject of technology,ML is a subfield of AI and is carefully related to arithmetic and statistics.

The learning begins with observations or facts, together with examples, direct experience, or practise, with a view to look for patterns in information and make higher decisions in future based on the earlier instances given. The primary intention is to permit the computers learn robotically with out human interference or help and regulate movements as a result.

ML algorithms are categorized as follows:

- Supervised ML algorithms can follow what has been discovered within the beyond to new statistics using categorised examples to predict events that might occur in future.To begin with,consider the analysis of an acknowledged training dataset, the ML algorithm produces an inferred feature to predict the output values. The system is capable of providing goals for any new I/P after adequate training. These set of rules can also evaluate its output which is comparable to the correct one,expected output and locate errors so one can adjust the model according to the requirements.
- In evaluation, unsupervised ML algorithms are used when the statistics used for training is neither categorised nor classified. Unsupervised learning researches how programs can infer a characteristic to describe a hidden shape from unlabeled records. The system doesn’t discern out the proper output, however it explores the facts and might draw inferences from datasets to describe hidden structures from unlabeled information.
- Semi-supervised ML algorithms lie someplace among supervised and unsupervised methods; when you consider that they use both classified and unlabeled statistics for

getting trained – usually a small amount of categorized records and a big amount of unlabeled data. The programs using this technique are able to appreciably improve correctness of getting learned.Semi-supervised approach is selected when the considered categorized statistics requires professional and relevant sources with a view to getting it trained / learning from it.Else, unlabeled data typically doesn’t require extra sources.

- Reinforcement ML algorithms is technique of getting trained, that interacts with its surroundings via producing moves and discovers mistakes.Trial and blunders seek and behind schedule reward are the maximum applicable traits of reinforcement approach. This approach permits machines and software program retailers to automatically decide the precise behavior inside a selected context if you want to maximize its overall performance. Simple reward comments is needed for the agent to examine which movement is great; that is known as the reinforcement sign.

ML allows large quatities of facts to get analyzed. While it grants faster, more accurate outcomes in an effort to perceive worthwhile possibilities or certain riks, it can also require time beyond regulation and assets to get trained nicely. Combining ML with AI and cognitive technology can prove to be extra efficient in processing huge volumes of records.

2.6 Importance of machine learning:

As statistical evaluation depends mostly upon rule-based selection-making, ML excels at responsibilities which are difficult to outline with actual guidelines.ML can be implemented to several enterprise cases in which final results relies on loads of factors.Elements which might be tough or not possible for a person to analyze.Hence, companies use ML of for predicting mortgage defaults, information elements that result in customer churn, figuring out probable fraudulent transactions, optimizing coverage claims approaches, predicting medical institution readmission, and plenty of different scenarios.

Organizations that correctly put in force ML and different AI technology get great benefits over others . According to a recent report by way of McKinsey & Company, AI will create $50 trillion of fee through 2025.The organizations failing to do so may be not able to compete with folks who include the new frontier.

Historically,ML has proved to be an effort taking method that calls for manual programming,proscribing the potential of corporations to take full gain of the ML algorithms. Without teams of hard-to-locate information scientists at their disposal, organizations are restrained inside the range of models they're able to broaden and check – and regularly those models consumes a great amount of time to get developed, they are old by the time they're developed.

2.7 Applications of machine learning in data science:

Two important concepts that are needed by data scientists are classification and regression. Hence,in order to gain the advantages of these two concepts,data scientists use ML. Some of the popular practical uses of regression and automatic classification are described as following:

- Finding unknown places such as oil fields, gold mines,archeological sites,etc based on existing sites (classification and regression)
- Finding names of persons or places using texts (classification)
- Trying to identify the correct person with the help of pictures or even some voice notes (classification)
- Processing the voices made by the birds and identifying it (classification)
- Predicting whether the loan shoul be approved to a customer or not (regression)
- Identification of profitable customers (regression and classification)
- Identification of parts of vehicles those are likely to malfuntion (regression)
- Identification of various diseases and tumors in bodies (classification)
- Prediction of the expenditure of a person on product X (regression)
- Predicting the number of volcanic eruptions under a certain time period (regression)
- Prediction of some organization's yearly revenue (regression)
- Prediction of the team whose probability of winning the tournament is maximum (classification)

Time-to-time models are built by data scientists (an abstraction of reality) telling how actually certain phenomenon work.Sometimes the intention of a model is iterpretation instead of prediction.This scenario is called root cause analysis.

Here are a few examples:

- Understanding certain processes of organizations and optimising the required processes.For instance,identifying the products that adds value to the product line.
- Discovering the causes of diseases
- Determining what causes traffic jams on roads

This list of machine learning applications can only be seen as an appetizer because it's ever-present within data science.As we have discussed earlier that regression and classification are two important techniques, but the repertoire and the applications don't end.Another important and beneficial technique is clustering. ML techniques prove benficial throughout the process of data science.

2.8 Where machine learning is used in data science process:

Majorly ML is related to the data science'data modeling step.Still ML somehow can be used in every step of data science.

It is necessary to have some qualitative raw data to get the data modeling phase started.Before this step,ML can be used in the data preparation step.

For instance:Suppose list os strings needs to be cleansed.Comparing the strings to spot the spelling errors can be done by placing the similar strings together,which can be done by ML algorithms.

ML is also useful in data exploration step of data science.

For instance:ML algorithms are capable of finding out patterns in data and bring the required data together,which would have been difficult errand otherwise,

It would not be correct to give the full credit to ML alone.Without certain Python libraries it would not have been possible to do these activities.Python libraries allows all the manipulation and processing of large sets of data.

2.9 Python tools used in machine learning:

Python provides a large number of packages,containing a several libraries that makes the ML efficient and beneficial to use.

Below is a figure,fig 5,which shows a number of Python packages and libraries.

- The first type of package allows some basic tasks and fitting of data into the memory.
- The second type of package enables code optimization as sometimes after the protyping is done there arises sever issues related to speed and memory.
- The third type of package enables the use of Python in big data technologies.

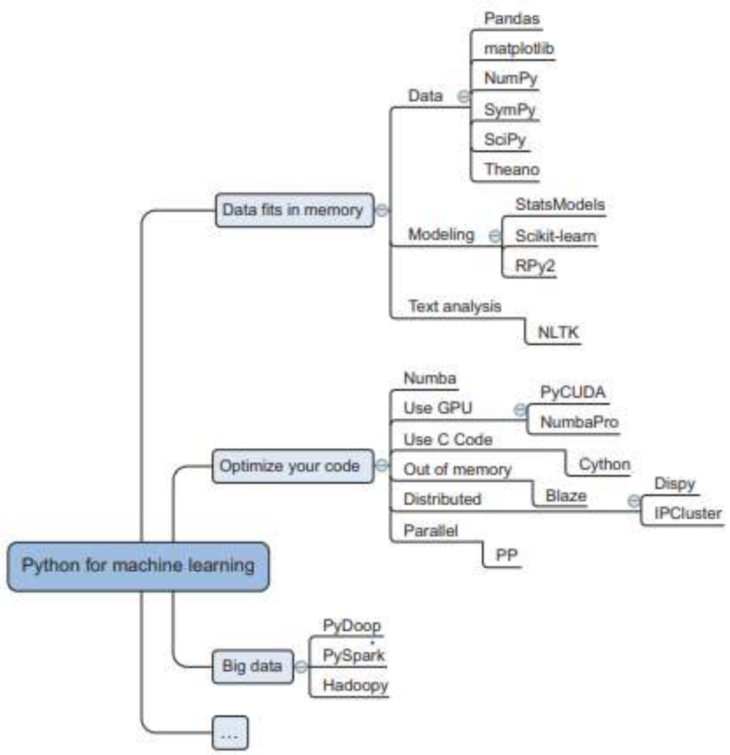


Fig 5:Python packages used in machine learning

2.9.1 Packages for working with data in memory:

These packages provides a number of functionalities and that too which consumes a very less numbers of lines.

- SciPy: A library that is responsible for the integration of some fundamental packages such as NumPy, matplotlib, Pandas, and SymPy.

- NumPy :It gives access to functions related to array functions and linear algebra.
- Matplotlib: This library is often called as 2D plotting package with some 3D functionality.
- Pandas:This library is a very useful high-performance library which is capable of manipulating data easily. It provides dataframes to Python.
- SymPy: It is a package which helps with computational algebra and symbolic mathematics.
- StatsModels: It is a package helpful for statistical methods.
- Scikit-learn:It is a Python library containing a large number of algorithms.
- RPy2 :This permits us to call functionalities of R from within Python.
- NLTK (Natural Language Toolkit) : This Python toolkit focuses on text analytics.

  It is a wise choice to begin with Python using these libraries.The real performance comes into play when a Python code is run at some regular intervals.

2.9.2 Packages for optimizing operations:

As we begin with the production phase of an application, the below listed libraries help to attain the required speed. This even includes connecting to big data infrastructures like Hadoop and Spark.

- Numba and NumbaPro—Compilations written in Python are speeded using these libraries. The GPU can be used intensly by NumbaPro.
- PyCUDA—Library lets us write program which instead of the central processing unit will be executed on the GPU.This makes it capable of applications that require a lot of calculations. An example is examining the strength of predictions by calculating several different outcomes based on a single start state.
- Cython, or C for Python—C language of coding is brought to Python by this.. As it is known that C is a lower-level language,bytecode,which is finally used by the system is used.When the type of variable is known then it makes the system faster.
- Blaze—Allows to work with massive sets of data by providing the appropriate data

structures which are very big(even bigger than the main memory) .

- Dispy and IPCluster—These packages permits the programmer tomake programs those can be shared over a network of computers.
- PP—Allows parallelization of computations on a system or a network of system which would have been run as a single process otherwise.
- Pydoop and Hadoopy—Python and Hadoop are linked with this.
- PySpark—Python and another Big Data framework,Spark are linked together with this.

2.10 The modeling process:

There are following stes in modeling phase:

1. Feature engineering and model selection
2. Training the model
3. Model validation and selection
4. Applying the trained model to unseen data

The starting three steps are iterated through until an appropriate model is not found.
In some cases,the final objective is explanation instead of prediction.In such cases the last step of applying the trained model to unseen data is not present.For example,we want to find the reasons of extinction of animal species but do not predict the name of species that might extinct in some time.

Chaining of models can also be done by making a combination of more than one technique.In chaining,output of one model is treated as an input to the other model.After combination of multiple models training of those models is still done separately.However,their results are combined together.This is called ensemble learning.

There are two major components in a model namely, feature and target variable.The major objective of the model is the prediction of target variable,for instance,high temperature of the next day.The features are the variables that help in making this prediction.Some examples of features for the temperature example are Wind Speed,Today’s Temperature,Movement of clouds.The models that accurately predicts are considered to be best.For this feature engineering is most helpful part of modeling.

2.10.1 Engineering features and selecting a model:

Creation of appropriate predictors for the model is done with engineering features.Since in this step the model recombines these features to attain the required predictions,hence this step is also considered one of the most important step.

Some functions are the variables you get from a records set. In exercise there is a need to discover the features on his/her self, that might be scattered among unique statistics units. In numerous tasks we had to convey collectively more than 20 one of a kind facts resources before we had the raw information we required. Many times there is a need to transform an input until it turns into an accurate predictor or to combine a couple of inputs. An instance of combining more than one inputs would be interaction variables: the impact of both single variable is low, but if both are present their effect becomes high. This is specifically authentic in chemical and environments related to medical science. For instance, even though vinegar and bleach are harmless commonplace family merchandise through themselves, blending them outcomes in poisonous chlorine fuel, a gasoline that killed hundreds all through World War I.

Many-a-times in order to get the features derived,modeling techniques are to be used.One model's output is used in second model.This is commonly done as in text mining.In order to categorize the content,documents first need to be annotated.Counting of number of places or people in the text can also be done.But this process is not that much easy,how it sounds.Firstly,the model is made to recognize some words like names of places or people.This information is then givem to another model that is to be developed.Availability bias is considered a major mistake in process of model making.If the model is having availablity bias then that model will fail while validation as it is seen that it is not an accurate model.

2.10.2 Training the model:

Model training can be done having an idea of efficient modeling technique and using the correct predictors at the correct place.In this step training data is given to the model so that it can learn using this data.The modeling techniques that are famous have all-set implementations ready to be used in nearly all coding languages.Thus,it makes it easier to train the model just by running a few lines of program.Other techniques of data science requires complex calculations and needs to be implemented using modern techniques.After the model gets trained successfully,it is to be checked whether this model is capable to deal with real-word problems or not.

2.10.3 Validating the model:

There are several data modeling techniques in data science.One just has to chose the correct and efficient one.Basically,there are two distinguishing properties of a good model:

- It has a good prediction making power
- It works efficiently and accurately with new data (test data)

In order to get these properties right,error measure needs to be defined which tells the extent to which the model is inaccurate and also a strategy for its validation.

Two of the most prominent error measures in ML are as follows:

- the classification error rate for classification problems: It is the percentage of observations in the testing data that the particular model mislabeled.The lesser this rate is,better the model is considered to be,
- the mean squared error for regression problems: It measures the extent of how big the average error of required prediction is. Results of squaring the average error are:wrong prediction cannot be undone in one direction with some prediction having faults in other direction. For instance, overestimating future turnover for next month by 5,000 is not capable of cancelling out underestimating it by 5,000 for the following month.Secondly,squaring makes big errors way more bigger.

Some of the validation techniques are:

- Partitioning the given data by taking out some percent of the data as training set,This is very common technique.
- K-folds cross validation:According to this technique the given data is divided into k parts and uses each part one time as a test data set the others as a training data set.The advantage of using this technique is that all the data present in the data set is used.
- eave-1 out:This approach is the similar to k-folds having k=1. One observation is always left out and training is done on the rest of the data. This technique is implemented only on datasets that are not too big.
- Regularization is a famous term in ML. While using regularization, a penalty is

givem for using the extra variables during the making of the model.

- A model having minimum possible predictors is ensured with L1 regularization. This ensures the model's robustness.
- The variance between the coefficients of the predictors are kept as minimum as possible using L2 regularization.It becomes difficult to see the actual impact of every predictor if the variance between predictors overlap.If there is no such overlapping the variance is interpreted more easily and clearly.
- Basically regularization prevents a model to use several features and this inturn prevents over-fitting.
- Validation checks if the model is actually working properly with the real-world applications also or not.This makes validation step important.

2.11 MACHINE LEARNING ALGORITHMS FOR PREDICTION MAKING:

- Logistic Regression
- Decision Tree
- Random Forest
- KNN
- Naïve Bayes
- Artificial Neural Network

2.11.1 Logistic Regression:

- In logistic regression,the dependent variable is binary in nature
- 1 refers to true and 0 refers to false
- Goal is to find the best fitting model for independent and dependent variable relationship
- Independent variable can be continous or binary
- It is also called Logit Regression

- Logistic regression is used in machine learning
- Deals with probability to measure the relation between dependent and independent variable

2.11.2 Decision Trees:

One of the most easy and famous classification algorithm is Decision Tree Algorithm.This algorithm helps interpretating and understanding better.

Decision tree algorithms is one of the supervised learning algorithms.

The decision tree algorithm is capable of solving both, classification and regression problems,which distinguishes it from rest of the supervised learning algorithms.

The main objective of using this algorithm is to predict the value/class of the target variable by learning some decision rules.

For predictions using this algorithm,we have to begin with the root node.The value of record,s attribute and root node are compared.This comparision tells what will be the next node that needs to be followed.

Types of decision trees:

There are two types of decision trees.

- Categorical Variable Decision Tree:If the target variable is of categorical type,then the tree is called Categorical Variable Decision Tree.
- Continous Variable Decision Tree:If the target variable is of continous type,then the tree is called Continous Variable Decision Tree.

Terminology:

1. **Root Node:** The first and the top most node of the decision tree from where the other edges and nodes emerge downwards.
2. **Splitting:** The process of dividing nodes into subnodes.
3. **Decision Node:** Makes the splitting of a node into sub nodes after a decision.
4. **Leaf / Terminal Node:** The bottom most nodes of the tree which do not split any further.

5. **Pruning:** It is the process exactly opposite to splitting.In this process the sub nodes from a node are removed.

6. **Branch / Sub-Tree:** It is a subsection of a tree.

7. **Parent and Child Node:** A node from which other nodes emerge is called a parent node,and the nodes ,so emerged, are called child nodes.

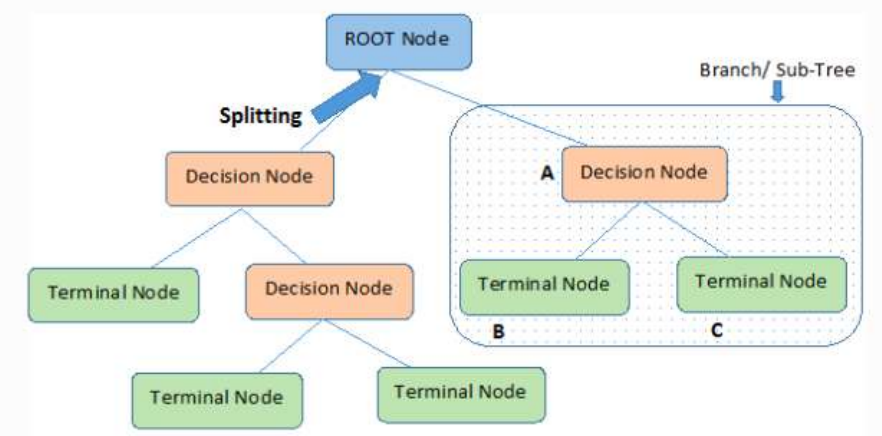


Fig 6:Decision tree terms visualized

Every node acts as a test case for some other attribute,and every edge running down from a node corresponds to a probable answer of that test case.

In the starting the entire train dataset is considered as the root of the decision tree.

Before building tyhis model the continous variables are converted into categorical.

Attribute values decide the distribution of the records.

Our focus will be on the ID3 algorithm used in the decision tree approach.This algorithm follows a top down approach.It is a greedy method.

Steps of ID3 algorithm:

1. The root node corresponds to the entire set S.

2. In each iteration an unused attribute of S is iterated and Entropy and Information Gain of that attribute is calculated.

3. The attribute having either minimum entropy or maximum Information gain is selected.
4. This selected attribute then divides the set S.
5. Similarly this process gets applied on the subset produced by selecting only those attributes that are not considered previously.

2.11.3 Random Forest:

As discussed previously,the VC dimension of decision tree of any said size is infinite. The size of the decision tree is hence restricted.Construction of an ensemble of tress can decrease the overfitting in some way.

The Random Forest illustrated below:

- Ensemble Learning is an approach which tells not to be reliant only on one model to make prediction.Rather take into consideration a number of models, and on the basis of outputs of all such models,come to a conclusion.
- The prediction made using this aaproach is far more accurate than it would have been considering only one model for predicting.
- Random forest is kind of an ensemble classifier which is using decision tree algorithm in a randomized fashion.
- Initially we are provided with a training data set i.e. original data set(OD).
- Using this OD, a new data set is generated i.e. bootstrap data set(BD).This data set is generated by randomly sampling the records from the OD and putting it in BD.Duplicate records are allowed in BD.However,it is better to have more unique data in the bootstrap data set.
- Now,considering the bootstrap data set,a decision tree is to be built.To decide the root node,a subset of the total number of variables present in the bootstrap data set is considered.For instance,if there are total three variables except the target variable in the bootstrap data set,then any two of the total three variables are considered randomly for the root.Out of these variables,one will become the root node of the decision tree.
- Similarly,generate a new bootstrap data set for each decision tree to be built.Build a

number of decision trees in the same way using the subset of total variables present in the bootstrap data set.

- After building a number of decision trees,a test data is given for which the value of target variable is to be predicted as output.This data will be tested by each and every decision tree.Keeping a track of all the outputs generated by these trees,a prediction will be made.
- The prediction will definitely be far more accurate than it would have been using only one decision tree.

2.11.4 KNN:

KNN is a rule which learns by memorizing.This algorithm prerequisites containing of the training data.The neighbors are found at the time of testing of the already stored training data. The time complexity of implementation of this algorithm is Θ(d m). This makes the computation expensive at the tie of testing. In case where the value of d is small, there are a number of data structures projected by several computational geometry results which makes the time complexity of this algorithm as (d O(1) log(m)). Anyhow, these data structures have the space complexity of mO(d) , hence making this approach inappropriate for cases where the value of d is large.In order to get rid of this problem,the approximate search should be allowed as this will result in improved searching.

The KNN algorithm is illustrated below:

- In KNN algorithm,Euclidean distance is measured between the training and test data.
- Euclidean distance is calculated as square root of summation of difference between between the data to be tested and training data.
- Suppose we are provided with a data set having a number of values.Let us say we have are given with a value whose category is to be tested.
- We are also provided with a value K,which tells us the number of neighbours which are closest to the test value.
- In order to do so,Euclidean distance between all the similar values from the table are to be calculated.
- K values having least distance with the test value are considered and their category is checked.

- Checking the categories of K nearest values,the value of the test data is predicted.

2.11.5 Naïve Bayes:

This algorithm serves as a primitive manifestation about simpifying the learning process via estimations of parameters and by taking some generative assumptions into consideration. Let us assume having a scenario of predicting a label y ∈ {0, 1} on the basis of a vector of features x = (x1, . . . , xd), where we assume that each xi is in {0, 1}. Recall that the Bayes optimal classifier is

$$h_S(\mathbf{x}) \;=\; \sum_{i=1}^{k} \frac{\rho(\mathbf{x}, \mathbf{x}_{\pi_i(\mathbf{x})})}{\sum_{j=1}^{k} \rho(\mathbf{x}, \mathbf{x}_{\pi_j(\mathbf{x})})}\, y_{\pi_i(\mathbf{x})}.$$

In order to illustrate the probability function P[Y = y|X = x] we need 2d parameters, each of which corresponds to P[Y = 1|X = x] for a certain value of x ∈ {0, 1} d . This shows that there is an exponential growth in the number of required instances with number of features.

In this approach a general assumption is taken into consideration that labels and features are independent of each other. That is,

$$\mathcal{P}[X = \mathbf{x}|Y = y] = \prod_{i=1}^{d} \mathcal{P}[X_i = x_i|Y = y].$$

With this assumption and using Bayes' rule, the Bayes optimal classifier can be further simplified:

$$\begin{aligned} h_{\text{Bayes}}(\mathbf{x}) &= \operatorname*{argmax}_{y\in\{0,1\}} \mathcal{P}[Y = y|X = \mathbf{x}] \\ &= \operatorname*{argmax}_{y\in\{0,1\}} \mathcal{P}[Y = y]\mathcal{P}[X = \mathbf{x}|Y = y] \\ &= \operatorname*{argmax}_{y\in\{0,1\}} \mathcal{P}[Y = y] \prod_{i=1}^{d} \mathcal{P}[X_i = x_i|Y = y]. \end{aligned}$$

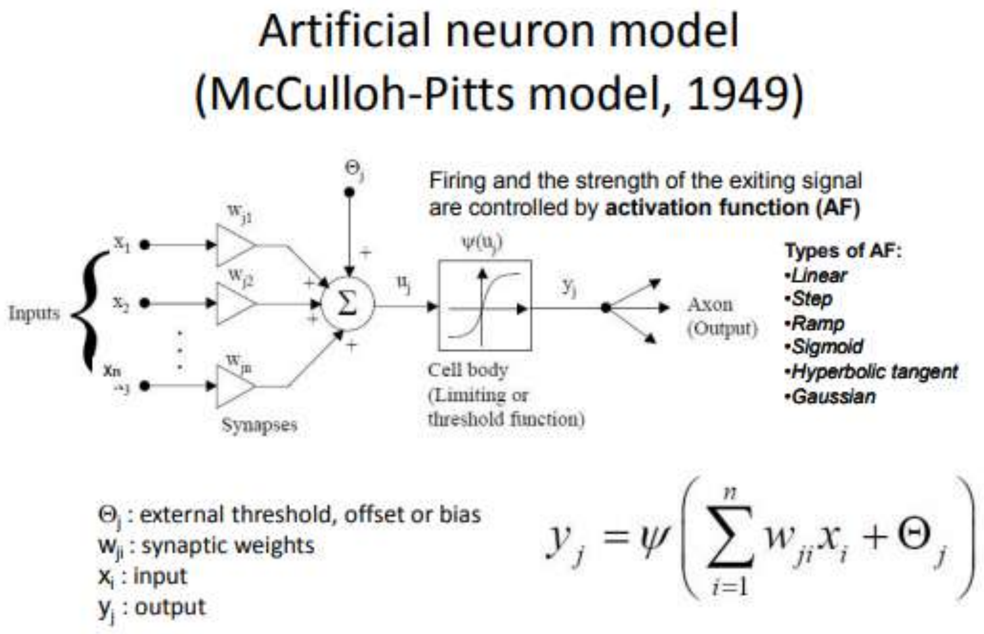
Now we can say that the total number of parameters needed to estimate is 2d + 1. In this the previously made generative assumptions reduced significantly the number of parameters we need to learn. When we also estimate the parameters using the maximum likelihood principle, the resulting classifier is called the Naive Bayes classifier.

2.11.6 Artificial Neural Network:

An algorithm of computer science which is totally comparable to real biological nervous system.A network similar to the network of nerves in the body is created which helps in learning,memorizing and in generating outputs to certain inputs.

Biological Neuron:

- Soma: responsible for processing of input.It contains the nucleus
- Dendrites:endings emerging out of the soma.They serves as input channels
- Axon:It serves as output channel.It is basically a type of link emerging from Soma
- Output of the axon: a pulse of voltage merely for a milisecond
- The Axon endings have the synaptic juction.It is a sort of contact with other neurons and this contact is electrochemical in nature.
- Learning is proportional to the synapse size
- Larger area is referred as excitatory
- smaller area is referred inhibitory



Artificial neuron model
(McCulloh-Pitts model, 1949)

Firing and the strength of the exiting signal are controlled by **activation function (AF)**

Types of AF:
- *Linear*
- *Step*
- *Ramp*
- *Sigmoid*
- *Hyperbolic tangent*
- *Gaussian*

Θ_j : external threshold, offset or bias
w_{ji} : synaptic weights
x_i : input
y_j : output

$$y_j = \psi\left(\sum_{i=1}^{n} w_{ji}x_i + \Theta_j\right)$$

Fig 7:Artificial neuron model

Different NN Types:

- Single-layer NNs
- Multilayer feedforward NNs
- Temporal NNs
- Self-organizing NNs
- Combined feedforward and self-organizing NNs

The ANN Applications:

- Classification
- Pattern matching
- Pattern completion
- Optimization
- Control

- Function approximation/times series modeling
- Data mining

Feedforward networks:

Input layer:The total number of input values is qual to the number of nodes present in this layer.Since the neurons of this layer do not participate in modification of signal,these nodes are refered to as passive nodes.These neurons are responsible for transferring the signal to the hidden layer.

Hidden layer:The hidden layer can have as many number layers as well as ,as many number of neurons too.The nodes present in the hidden layer are also called active nodes as these nodes are responsible for signal modification.

Output layer: The number of values generated by the network as output is equal to the number of neurons present in this layer.These nodes are also called active nodes.

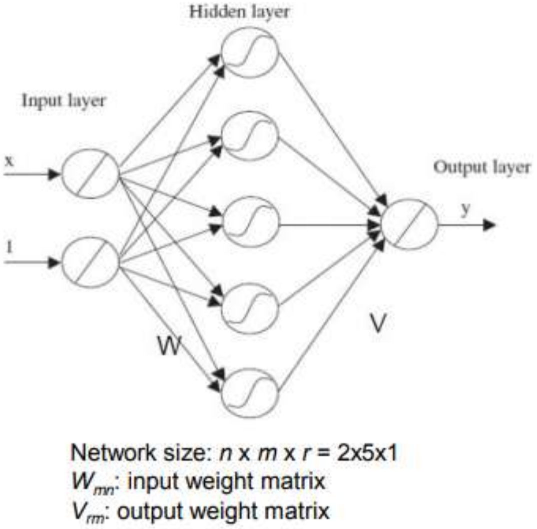

Hidden layer
Input layer
x
1
Output layer
y
W
V
Network size: n x m x r = 2x5x1
W_{mn}: input weight matrix
V_{rm}: output weight matrix

Fig 8:Feedforward Network

Feedback Networks:

The neurons which are used for recognizing pattern acts as a path for the output to be given as input either directly or indirectly.

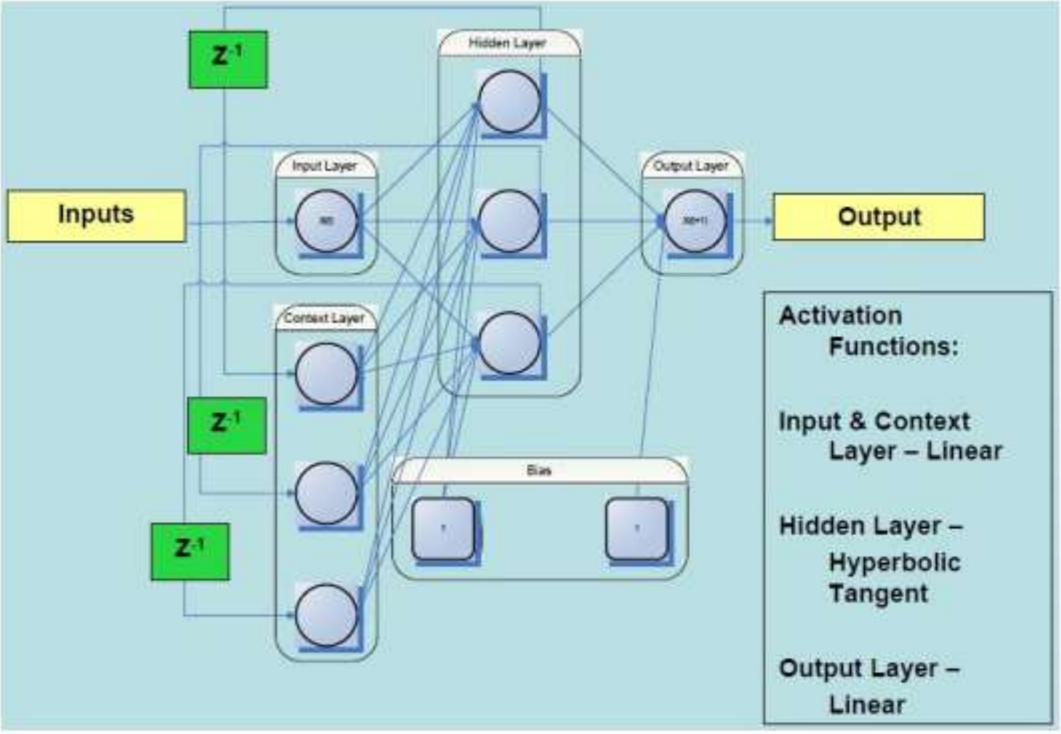


Fig 9:Feedback Network

Lateral Networks:

- Within a single layer,coupling of neurons occur
- Among the layers,there exists no path for essentially explicit feedback
- This can be treated like a compromise among forward and feedback network

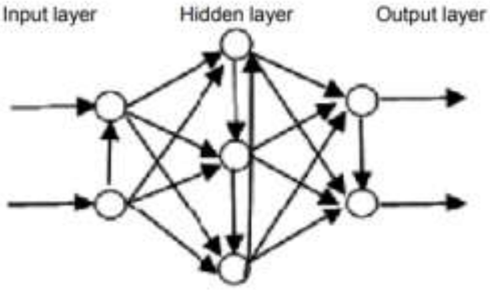


Fig 10:Lateral Network

Optimized values of weights are used in ANN.

Learning is the procedure of acquiring these optimized weights.

It is made to learn by the system,that how it should generate correct outputs for each and every input provided to that system.

After the learning procedure the whole network gets trained.Now,the newly updated weights are provided as inputs and the trained neural network should now generate the outputs for these new input values.The output must be expectedly accurate.

Learning methods

- Supervised learning
- Unsupervised learning
- Reinforcement learning

CHAPTER-3: SYSTEM DEVELOPMENT

3.1 System Requirements:

The system requirements for the algorithms to run efficiently and for the implementation of the whole idea are:

- Windows 10 (64-bit)
- 8 GBRAM
- Intel(R) Core(TM) i5-6200U CPU @ 2.30GHz
- ANACONDA
- Python

3.2 Python Libraries:

3.2.1 numpy:

Numpy or ‘Numerical Python’ is a Python package.This library contain objects of multidimensional array.It also contains,for processing of array,a large collection of routines.

Numpy enables the programmer to implement:

- Certain mathematical and logical operations on arrays.
- Fourier transforms and some routines for shape manipulation.
- Some linear algebra operations. NumPy has in-built functions for linear algebra and random number generation

Some packages such as SciPy and Matplotlib are also used quite frequently with NumPy.SciPy refers to the Scientific Python and Matplotlib serves as a plotting library for visualization.Whenever these two packages are combined together,they can serve the functionalities similar to Matlab,hence proves to be a replacement for the Matlab.This option is used more frequently over the Matlab.Hence this shows how much capable Python truly is.

The NumPy package allows the usage of mathematical operators in various datatypes,be it a list or a dictionary.Numpy proves beneficial in manipulating and doing some calculative work over the columns of the datasets.

3.2.2 Pandas:

One of the most famous library of Python programming language for study and interpretation of data is Pandas.

It is already known very clearly that before the preperation of any model,it is necessary to set up the data sets for the model.For this very cause,Pandas'role come into play.Pandas library is responsible for the preparation and extraction of adequate data for a data model.

Pandas provides us with useful functionalities which helps in importing files,manipulating those files and enable to implement certain significant concepts.It also enables grouping and seperation of datasets.

- In Pandas,a table of information is also refered to as a DataFrame
- Pandas proves to be the adequate tool for processing and cleaning of tabular data.For instance,the data stored in some database or in spreadsheets can easily be processed using Pandas library.
- Pandas allows working with files of several formats,like,csv,excel,sql,json,etc.
- To import the file into the system,there is a functionality whisch is used as, read_*.
- In the same way, to_* method stores data.
- Pandas also provide methods for grouping,dividing and extracting the data from a file which is imported previously.
- It is also possible to filter out some number of specific rows of records out of a large set of records using Pandas.
- Pandas,using matplotlib enables to visualize the data in a desired way.There are many graphs that can be generated be it bar graph or scattered graph.
- Pandas make it easy to add new column to a Dataframe.
- While making calculations,Pandas saves the effort of going throuhgh all the records of the dataframe.Calculations can also be made columnwise.
- Pandas also allow us to apply basic statistical aggregate operations of

mean,median,min,max,count,etc.These functions can either be applied on the full dataset or by putting on some conditions can be applied on desired part of the table only.

- The basic structure of the table containing information can be changed using Pandas in several ways.For instance,melt() can help in making a wide table tidy by making it long and pivot() can change the table from long to wide.
- Pandas enables concatination of many tables,be it row-wise or column-wise
- Pandas provides a number of tools for working with dates and times.
- Not only numeric data,but Pandas also proves helpful in doing operations on textual datam contained in a table.

3.2.3 Matplotlib:

Matplotlib is a Python library used for generating 2D plots and charts.

Matplotlib is a very popular library for visualization of data.This library allows plotting the data in various formats and visualizing of this plot is an easy task.This visualization of the data make the things more understandable.Data interpretation becomes easy by visualization.

Hence we can say that Matplotlib is valuable for a developer to picturize the information for better understanding.

The pyplot module enables the developers to take a controlover the style of lines,properties of styles of texts,etc.This module helps plotting various forms of grahs such as bars,histograms and many more.

- Matplotlib.pyplot enables the matplotlib work same as Matlab by its functionalities.
- Certain changes to a figure can be made using pyplot: for instance, making a figure, specifying a particular plotting area in the figure, generating certain lines in the specified plotting area, making the plot more readable by adding lables to the plot, etc.
- Matplotlib picturizes/visualizes the data without taking much time.It is quick

3.2.4 Seaborn:

Another important library,Seaborn enables generating graphics that are statistical.This seaborn library works properly in cooperation with data structures of Pandas and is built on top of Matplotlib library.

Functionalities offered by Seaborn library are:

- An API which works in accordance with a dataset for knowing the inter-realtion of the variables
- Provides support for the variables which can be categorizes for visualizing statistrics
- Provides the choice for visualization of distributions be it either univariate or bivariate,and allows their comparision with data subsets.
- Seaborn provides support to linear regression models by estimating and generating plots for various types of features.
- Makes complicated datasets look understandable and readable.
- Enables easy generation of even complex visualizations by abstractions that are of good standards.
- Helps styling figures generated by matplotlib using a number of default themes.
- Some tools for opting color palettes are also provides.The helps in making the plot readable.

The major objective of seaborn library is to prove the importance of visualization in understanding the data.Seaborn's functions are appliable on dataframes and data structures containing multiple datasets.These functions implicitly do the mapping and aggregation to generate valuable plots.

3.3 Scikit Learn:

Scikit Learn is a very important machine learning library.This library is used for traditional machine learning calculations.Scikit Learn is based upon two major libraries of Python,namely,NumPy and SciPy.

Scikit Learn can be used for data mining and processing of data.Hence this makes an essential tool for the machine learning and development.

3.4 Anaconda:

Anaconda serves as a distribution for Python and R programming language.This is an open-source distribution.

Data science and machine learning uses Anaconda.

Anaconda provides more than 300 libraries for data science.This makes Anaconda popular among the developers working with data science.

Management and deployment of simplified packages are assisted by Anaconda.Anaconda provides a very massive variety of tools which are capable of assembling data from several sources with the help of ML algorithms.

Anaconda provides a space which is simple and easily manageable which makes deployement of any project just a click away.

3.5 Jupyter Notebook:

The console-based approach is extended to computing which is interactive in nature in a new path by the notebook.The notebook provides us with an application which is mostly web based which supports the computational process which includes developing,documenting and implementation of code and also the generation of results.

The Jupyter Notebook is responsible for combining the following two components:

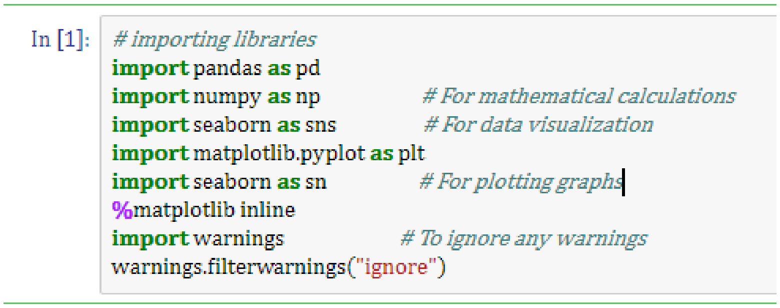
- A web application: a tool which is assisted by the browser in order to author the documents in a better way,which inturn links texts,arithmetic operations and computations and their output.
- Notebook documents: This includes the portrayal of inputs,outputs of the processes,texts,arithmetic operations,media along with all the content that is visible in the web application.

CHAPTER 4 PERFORMANCE ANALYSIS

In this project:

1. The data set needs to be explored initially.
2. Certain models will be created toknow whether the loan should be approved or not.
3. Accuracy scores for the clients will be generated.

The first step is of this project development is to import the required libraries:



```
# importing libraries
import pandas as pd
import numpy as np               # For mathematical calculations
import seaborn as sns            # For data visualization
import matplotlib.pyplot as plt
import seaborn as sn             # For plotting graphs
%matplotlib inline
import warnings                  # To ignore any warnings
warnings.filterwarnings("ignore")
```

Fig 11:Importing libraries

Imported libraries are:

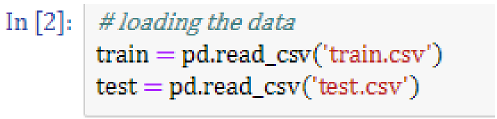
- pandas
- numpy
- seaborn
- matplotlib.pyplot
- warnings

All these libraries have been discussed in the literature survey in detail.

Objects of all these libraries are made.These objects will help in accessing the methods present in these libraries.

The next step is to import the training data set in the object named train and the testing data set in the object named test.

This import is done by pandas method of reading .csv files into the system.

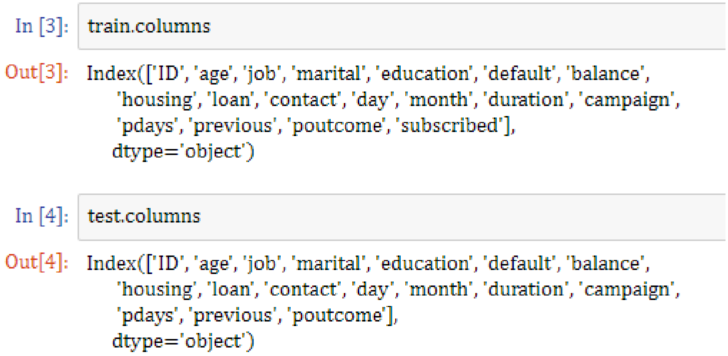


```
In [2]: # loading the data
        train = pd.read_csv('train.csv')
        test = pd.read_csv('test.csv')
```

Fig 12:Reading files into the system

Next we check the the features present in the loaded data sets.

This is done by .columns .



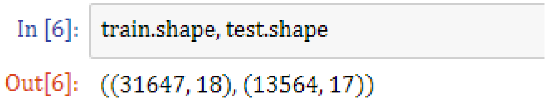
```
In [3]: train.columns

Out[3]: Index(['ID', 'age', 'job', 'marital', 'education', 'default', 'balance',
               'housing', 'loan', 'contact', 'day', 'month', 'duration', 'campaign',
               'pdays', 'previous', 'poutcome', 'subscribed'],
              dtype='object')

In [4]: test.columns

Out[4]: Index(['ID', 'age', 'job', 'marital', 'education', 'default', 'balance',
               'housing', 'loan', 'contact', 'day', 'month', 'duration', 'campaign',
               'pdays', 'previous', 'poutcome'],
              dtype='object')
```

Fig 13:Features in datasets

Subscibed is present only in the training dataset and hence this makes it the target variable.

Shape of the data sets are checked further.

The .shape tells us how many total number of rows and columns are present in a dataset.



```
In [6]: train.shape, test.shape

Out[6]: ((31647, 18), (13564, 17))
```

Fig 14:Shape of datasets

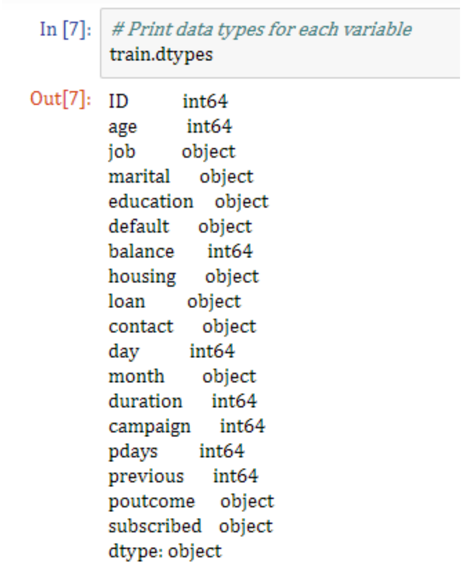
In the training data set there are 17 independent variables and 1 target variable(subscribed). Rest of the features of both the datasets are identical.

The target variable subscribed,will be predicted using model trained with the train data.

Further we will see what categorical and numerical variables are there in our dataset.Datatypes of the variables will be checked.

Now let us check out the datatype of each feature:

The .dtypes is used for this.



In [7]:
```
# Print data types for each variable
train.dtypes
```

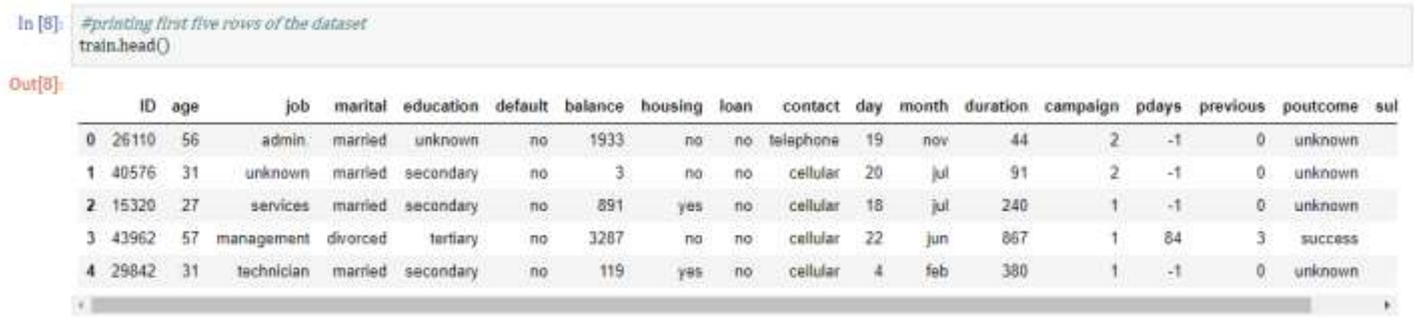
Out[7]:
```
ID            int64
age           int64
job          object
marital      object
education    object
default      object
balance       int64
housing      object
loan         object
contact      object
day           int64
month        object
duration      int64
campaign      int64
pdays         int64
previous      int64
poutcome     object
subscribed   object
dtype: object
```

Fig 15:Checking datatypes of features

It is clearly seen that there are two categories of data types:

1. **object**: Includes the variables that are categorical. job, marital, education, default, housing, loan, contact, month, poutcome, subscribed lie in this category.
2. **int64**: Includes the integer variables. ID, age, balance, day, duration, campaign, pdays, previous lie in this category.

To display some of the records of a dataset for reference, .head() is used.If we specify a number in the parameter of this function,then that many rows are displayed otherwise top five records are displayed.



```
In [8]: #printing first five rows of the dataset
        train.head()
Out[8]:
```

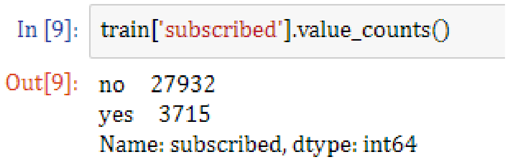
	ID	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	sul
0	26110	56	admin	married	unknown	no	1933	no	no	telephone	19	nov	44	2	-1	0	unknown	
1	40576	31	unknown	married	secondary	no	3	no	no	cellular	20	jul	91	2	-1	0	unknown	
2	15320	27	services	married	secondary	no	891	yes	no	cellular	18	jul	240	1	-1	0	unknown	
3	43962	57	management	divorced	tertiary	no	3287	no	no	cellular	22	jun	867	1	84	3	success	
4	29842	31	technician	married	secondary	no	119	yes	no	cellular	4	feb	380	1	-1	0	unknown	

Fig 16:The .head() method

Univariate analysis:

Let us check the distributionof the target variable subscribed.Since this variable lies in the categorical type,let us generate a frequency table,distribution and bar plot.

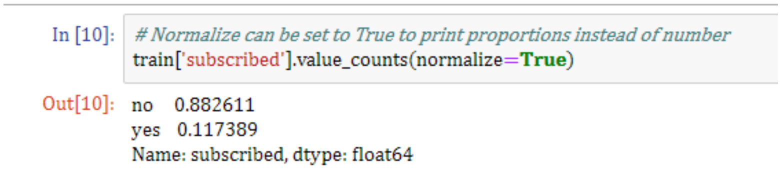
First of all let us see how many yes and how many no are there in the the target variable subscribed.



```
In [9]: train['subscribed'].value_counts()
Out[9]: no     27932
        yes     3715
        Name: subscribed, dtype: int64
```

Fig 17:Counting values in target variable “subscribed”

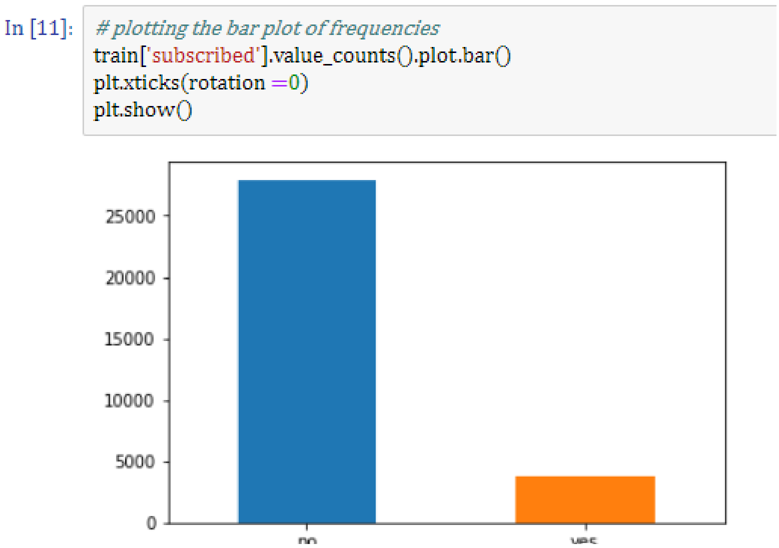
In order to see yes and no in proportion format,we set normalize to true.



```
In [10]: # Normalize can be set to True to print proportions instead of number
         train['subscribed'].value_counts(normalize=True)
Out[10]: no     0.882611
         yes    0.117389
         Name: subscribed, dtype: float64
```

Fig 18:Normalizing

Now we will generate a plot of frequencies of the target variable subscribed using the object of matplotlib.



```
In [11]: #plotting the bar plot of frequencies
         train['subscribed'].value_counts().plot.bar()
         plt.xticks(rotation =0)
         plt.show()
```

Fig 19:Generating plot of features of target variable

It is concluded that there are 3715 yes,which means this many users have subscribed.It is 12%of the total number of users.

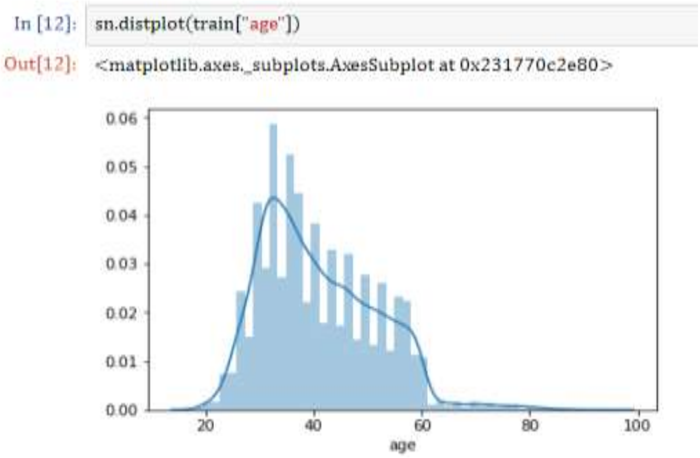
Different variables will be explored so that the entire data set is better understood.

Univariate analysis of variables will help analysing the variables individually.

Bivariate analysis would enable us to see the relation of different variables with the target variable.

In order to see which variable influences the target variable the most,correlation plot will be used.

The distribution of age will show the number of people belonging to a similar age group.



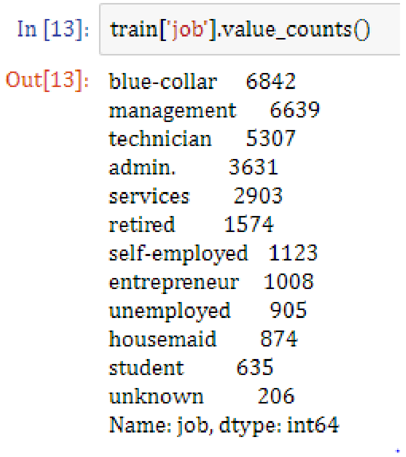
```
In [12]: sn.distplot(train["age"])
Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0x231770c2e80>
```

Fig 20:Distribution of age

It is visible that maximum people lie under the 20-60 group.

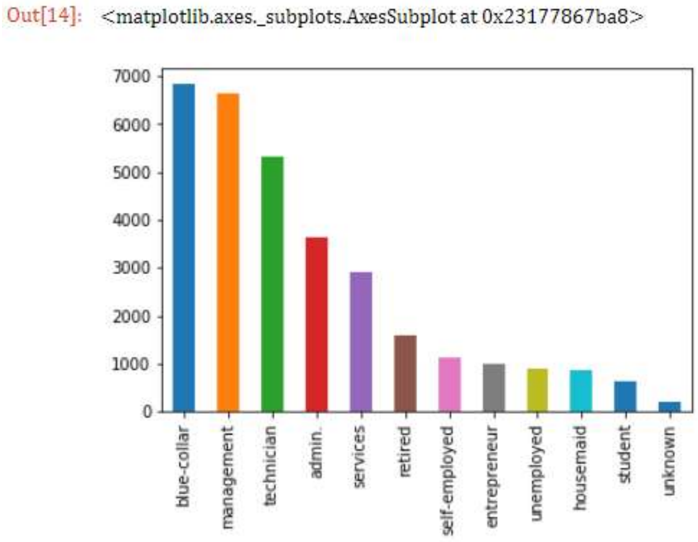
Next,we should see what are the professions of people using the job variable.

Frequency table of job:



```
In [13]: train['job'].value_counts()

Out[13]: blue-collar     6842
         management      6639
         technician      5307
         admin.          3631
         services        2903
         retired         1574
         self-employed   1123
         entrepreneur    1008
         unemployed       905
         housemaid        874
         student          635
         unknown          206
         Name: job, dtype: int64
```

Fig 21:Frequency of job



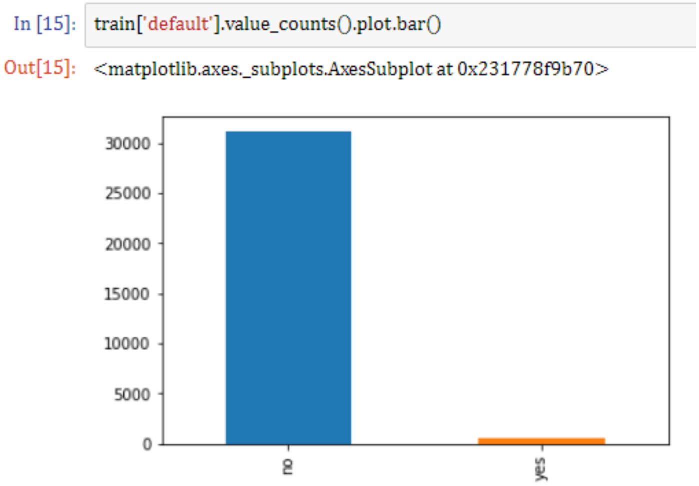
```
Out[14]: <matplotlib.axes._subplots.AxesSubplot at 0x23177867ba8>
```

Fig 22:Plot of features of job

It is visible in the plot,that maximum people are employed at blue-collar job.

It is also seen that number of students is very less.It is because generally students do not apply for a loan.

Now let us see how many users are having a default history.



```
In [15]: train['default'].value_counts().plot.bar()
Out[15]: <matplotlib.axes._subplots.AxesSubplot at 0x231778f9b70>
```

Fig 23:Plot of default

Less than 10% users have a default history.

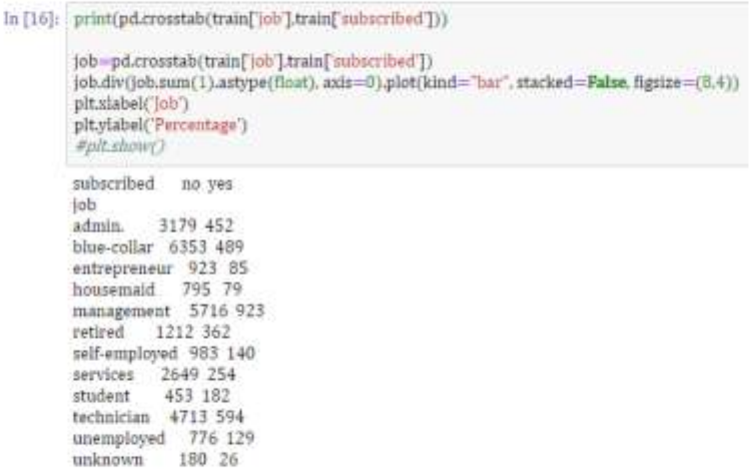
Bivariate analysis:

In bivariate analysis,we will see the relationship of target variable subscribed with other independent variables.

For numeric/continous variable,scatter plots will be used.

For categorical type of variables,crosstabs would be used.

Let's start with job and subscribed variable.



```
In [16]: print(pd.crosstab(train['job'],train['subscribed']))

         job=pd.crosstab(train['job'],train['subscribed'])
         job.div(job.sum(1).astype(float), axis=0).plot(kind="bar", stacked=False, figsize=(8,4))
         plt.xlabel('Job')
         plt.ylabel('Percentage')
         #plt.show()

subscribed     no yes
job
admin.       3179 452
blue-collar   6353 489
entrepreneur   923  85
housemaid     795  79
management    5716 923
retired      1212 362
self-employed  983 140
services     2649 254
student      453 182
technician    4713 594
unemployed     776 129
unknown      180  26
```

Fig 24:Job-Subscribed Crosstab

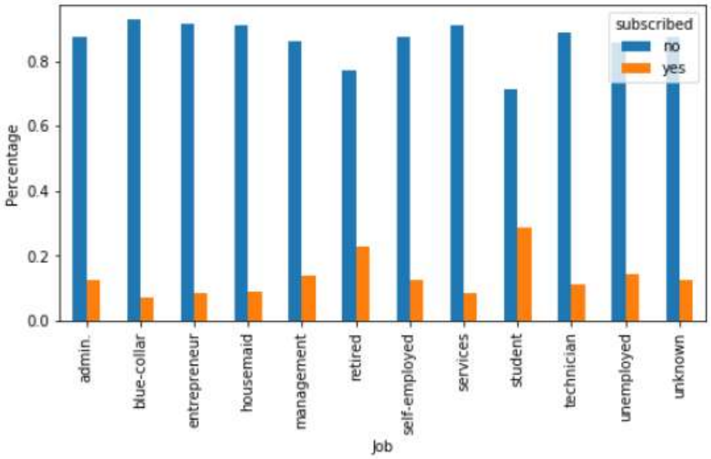
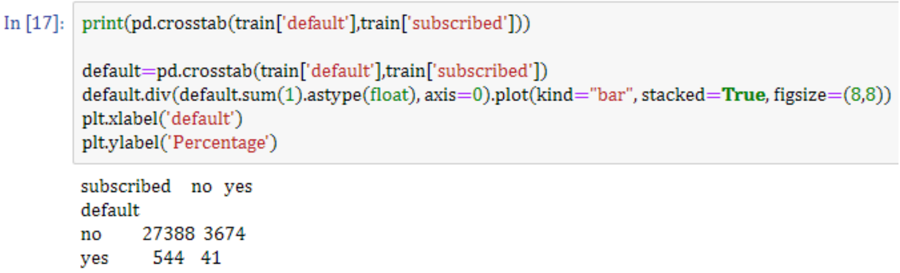


Fig 25:Plot of Job-Subscribed Crosstab

The above graph shows that students and retired people have more tendency to take a loan.This is surprising because students in general do not apply for a loan.So,the reason for this is that total number of students is very less in our data set when compare to other types.In comparision to other professions,a little more number of students have applied for the loan.

Now, let's see the relationship of default variable and the subscribed variable.



```
In [17]: print(pd.crosstab(train['default'],train['subscribed']))

         default=pd.crosstab(train['default'],train['subscribed'])
         default.div(default.sum(1).astype(float), axis=0).plot(kind="bar", stacked=True, figsize=(8,8))
         plt.xlabel('default')
         plt.ylabel('Percentage')
```

```
subscribed   no  yes
default
no        27388 3674
yes         544   41
```

Fig 26: Default-Subscribed Crosstab

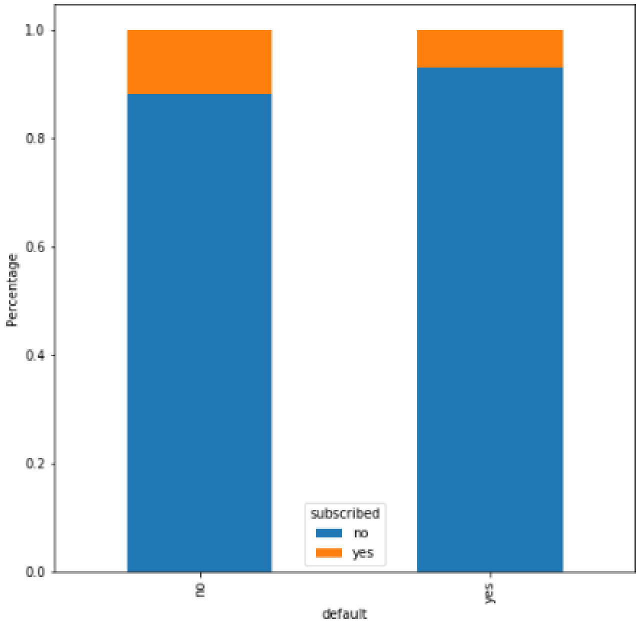


Fig 27:Plot of Default-Subscribed Crosstab

It is visible,that people who do not have a past loan history,have higher chances of applying for a loan.

Now we will look into the correlation between the numerical variables of data set.

The variables having higher positive/higher negative values are correlated.

This can tell,which variables have more tendency to affect the target variable subscribed.For this analysis,the target variable first needs to be conerted into numerical type.

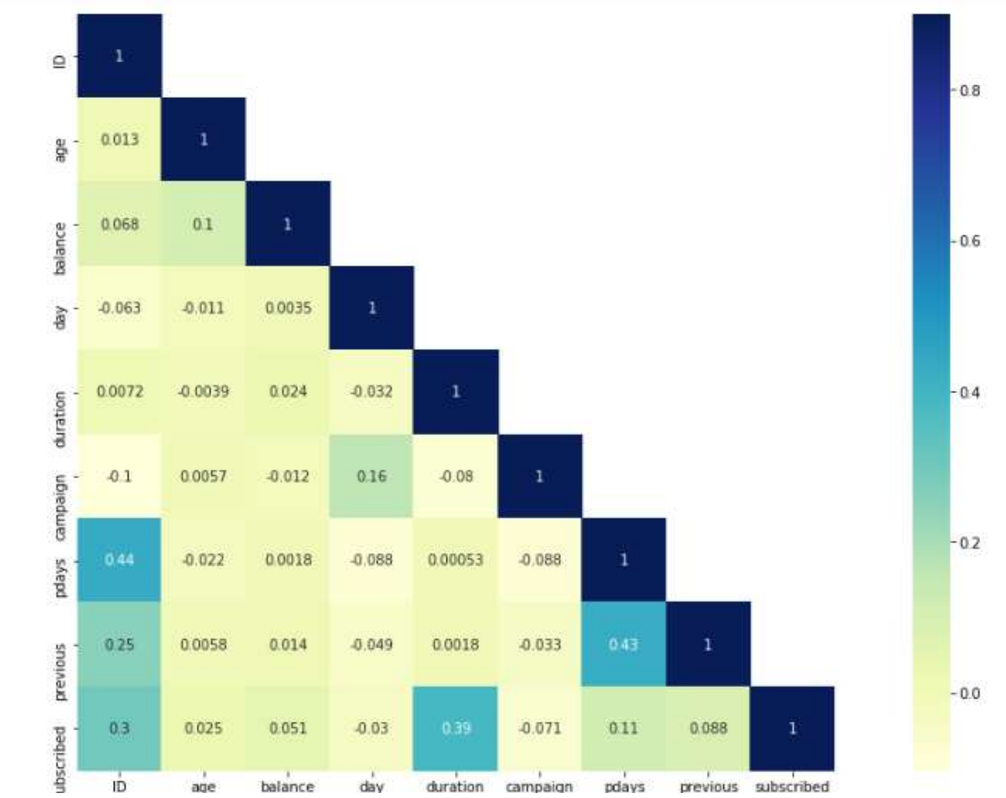


Fig 28:Correlation between numeric variables

It is visible that target variable subscribed and duration are highly correlated.

This is evident as the client having a higher call duration might be showing more interest in the loan scheme.This implies, that particular client might get ready to apply for the loan.

Next we will look for any missing values in the dataset.

There are no missing values in the train dataset.

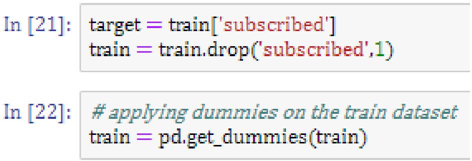
Model Building:

Now,the development of a model for predicting if the user will apply for a loan or not will start.

Dummies will be used for converting categorical variables into numerical variables because sklearn models allows only numerical inputs.

ID, being the unique valued variable will be removed before applying dummies.

The target variable subscribed will also be removed.

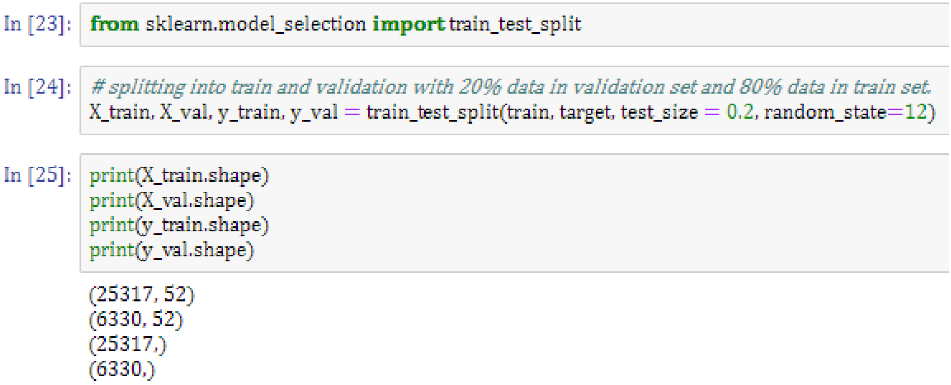


```
In [21]: target = train['subscribed']
         train = train.drop('subscribed',1)

In [22]: # applying dummies on the train dataset
         train = pd.get_dummies(train)
```

Fig 29:the .get_dummies() method

Now, we will start building the model.

The train data set will be divided into two parts,80% of the data will act as training data and the remaining 20% data will be the validation data.

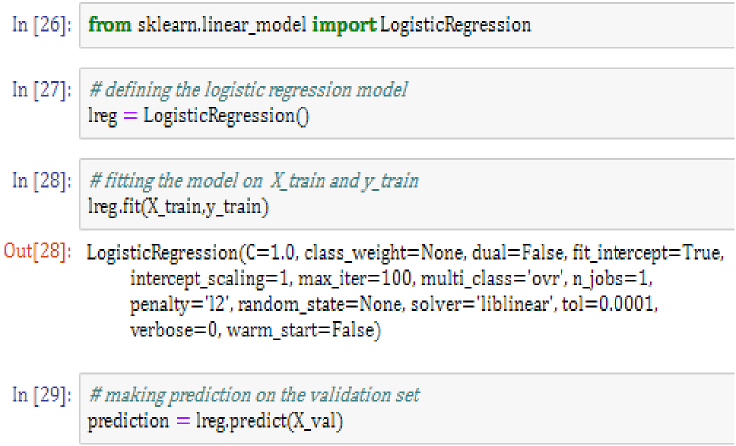


```
In [23]: from sklearn.model_selection import train_test_split

In [24]: # splitting into train and validation with 20% data in validation set and 80% data in train set.
         X_train, X_val, y_train, y_val = train_test_split(train, target, test_size = 0.2, random_state=12)

In [25]: print(X_train.shape)
         print(X_val.shape)
         print(y_train.shape)
         print(y_val.shape)

         (25317, 52)
         (6330, 52)
         (25317,)
         (6330,)
```

Fig 30:Splitting the data

Logistic Regression:

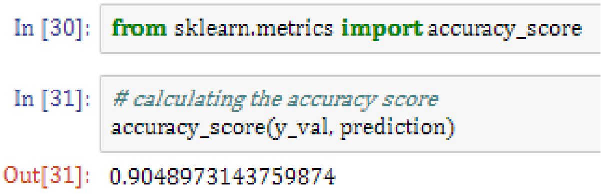
We will first build a Logistic Regression model since logistic regression is used for classification problems.



```
In [26]: from sklearn.linear_model import LogisticRegression

In [27]: # defining the logistic regression model
         lreg = LogisticRegression()

In [28]: # fitting the model on  X_train and y_train
         lreg.fit(X_train,y_train)

Out[28]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
                   penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
                   verbose=0, warm_start=False)

In [29]: # making prediction on the validation set
         prediction = lreg.predict(X_val)
```

Fig 31: Fitting the model

Now the accuracy of the predictions will be checked. Calculating accuracy on the validation set.



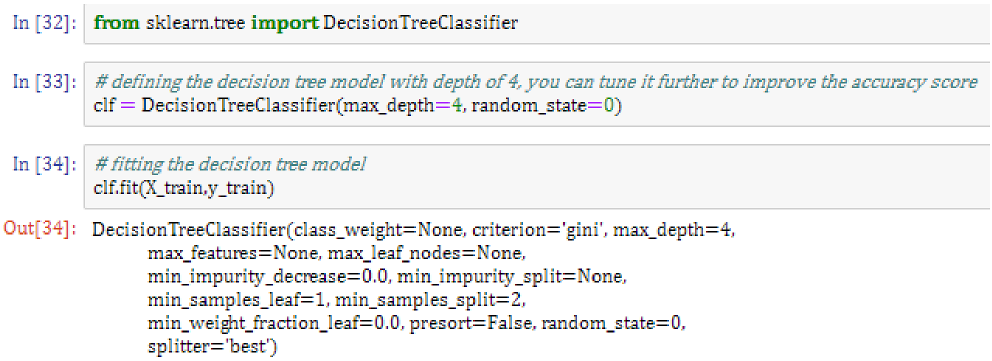
```
In [30]: from sklearn.metrics import accuracy_score

In [31]: # calculating the accuracy score
         accuracy_score(y_val, prediction)

Out[31]: 0.9048973143759874
```

Fig 32:Generating result of logistic regression

Accuracy on the validation set came out to be 90.4%.

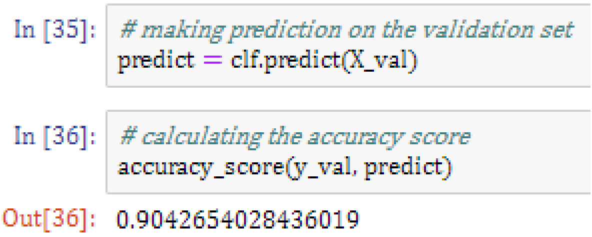
Decision Tree algorithm:

Let's try decision tree algorithm now to check if we get better accuracy with that.



```
In [32]: from sklearn.tree import DecisionTreeClassifier

In [33]: # defining the decision tree model with depth of 4, you can tune it further to improve the accuracy score
         clf = DecisionTreeClassifier(max_depth=4, random_state=0)

In [34]: # fitting the decision tree model
         clf.fit(X_train,y_train)

Out[34]: DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=4,
                     max_features=None, max_leaf_nodes=None,
                     min_impurity_decrease=0.0, min_impurity_split=None,
                     min_samples_leaf=1, min_samples_split=2,
                     min_weight_fraction_leaf=0.0, presort=False, random_state=0,
                     splitter='best')
```

Fig 33:Fitting Decision Tree Model



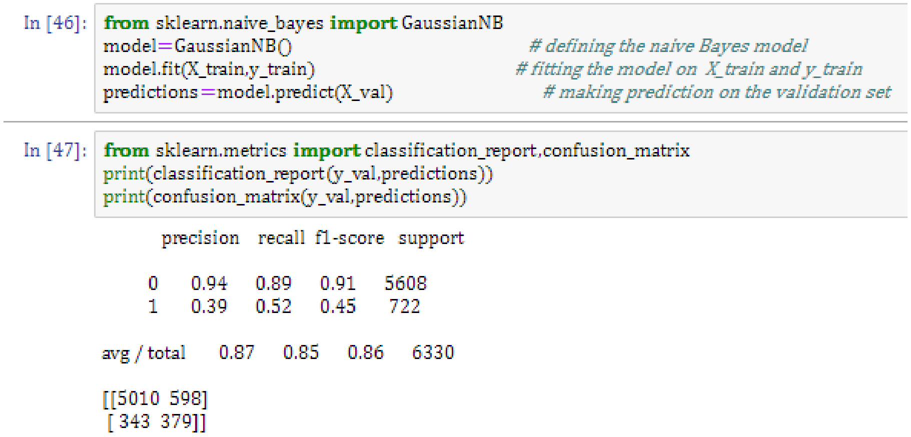
```
In [35]: # making prediction on the validation set
         predict = clf.predict(X_val)

In [36]: # calculating the accuracy score
         accuracy_score(y_val, predict)

Out[36]: 0.9042654028436019
```

Fig 34:Decision Tree Output

We got an accuracy of more than 90.4% on the validation set.

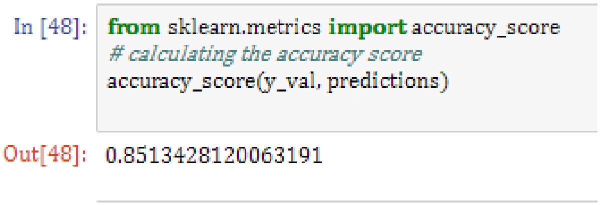
Naïve Bayes algorithm:

Now let us check what accuracy will be generated by Naïve Bayes algorithm.



```
In [46]: from sklearn.naive_bayes import GaussianNB
         model=GaussianNB()                              # defining the naive Bayes model
         model.fit(X_train,y_train)                     # fitting the model on  X_train and y_train
         predictions=model.predict(X_val)                 # making prediction on the validation set

In [47]: from sklearn.metrics import classification_report,confusion_matrix
         print(classification_report(y_val,predictions))
         print(confusion_matrix(y_val,predictions))

                      precision    recall  f1-score   support

                   0       0.94      0.89      0.91      5608
                   1       0.39      0.52      0.45       722

         avg / total       0.87      0.85      0.86      6330

         [[5010  598]
          [ 343  379]]
```

Fig 35:Defining and fitting Naïve Bayes Model



```
In [48]: from sklearn.metrics import accuracy_score
         # calculating the accuracy score
         accuracy_score(y_val, predictions)

Out[48]: 0.8513428120063191
```

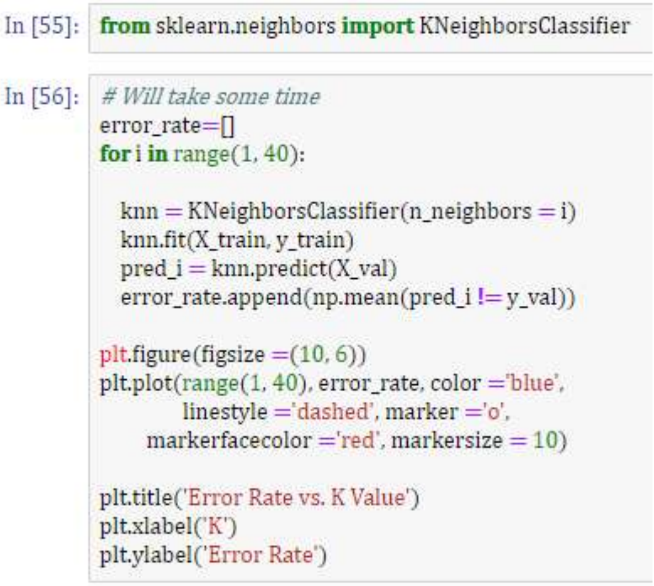
Fig 36:Generating prediction using Naïve Bayes algorithm

We got 85.13% accuracy using this algorithm.

KNN algorithm:

Now we will see the prediction results made by the KNN algorithm.

Let us start with setting up the model for KNN.



```
In [55]: from sklearn.neighbors import KNeighborsClassifier
```

```
In [56]: # Will take some time
         error_rate=[]
         for i in range(1, 40):

             knn = KNeighborsClassifier(n_neighbors = i)
             knn.fit(X_train, y_train)
             pred_i = knn.predict(X_val)
             error_rate.append(np.mean(pred_i != y_val))

         plt.figure(figsize =(10, 6))
         plt.plot(range(1, 40), error_rate, color ='blue',
                         linestyle ='dashed', marker ='o',
                 markerfacecolor ='red', markersize = 10)

         plt.title('Error Rate vs. K Value')
         plt.xlabel('K')
         plt.ylabel('Error Rate')
```

Fig 37:Setting up for KNN

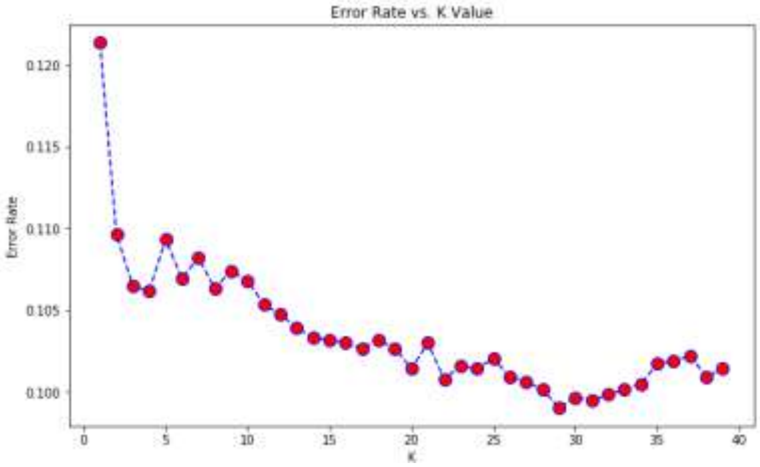
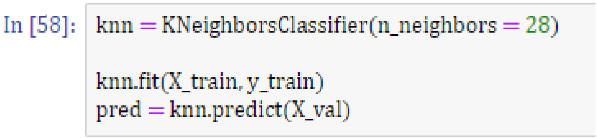


Fig 38:Error rate VS K Value

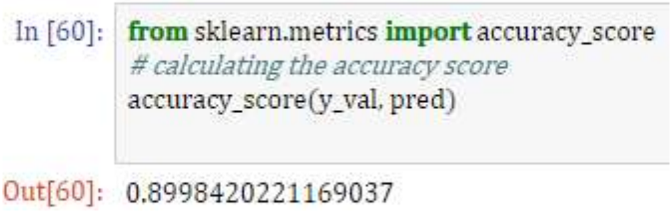
From the above chart the value of k for minimum error must be 28 or 29.

Let us consider k=28.



```
In [58]: knn = KNeighborsClassifier(n_neighbors = 28)

         knn.fit(X_train, y_train)
         pred = knn.predict(X_val)
```

Fig 39:Considering K=28



```
In [60]:  from sklearn.metrics import accuracy_score
           # calculating the accuracy score
           accuracy_score(y_val, pred)

Out[60]:  0.8998420221169037
```

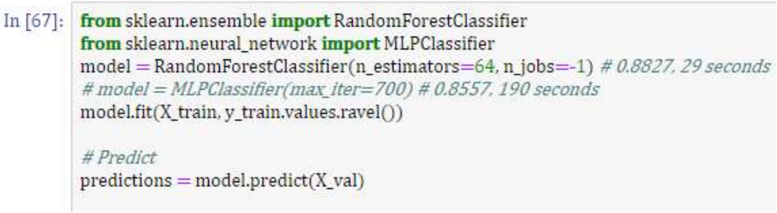
Fig 40:Accuracy using KNN algorithm.

Using the KNN algorithm,we got the accuracy of nearly 89.9 percent.

Random Forest Algorithm:

Now let us calculate the accuracy using another algorithm called Random Forest.

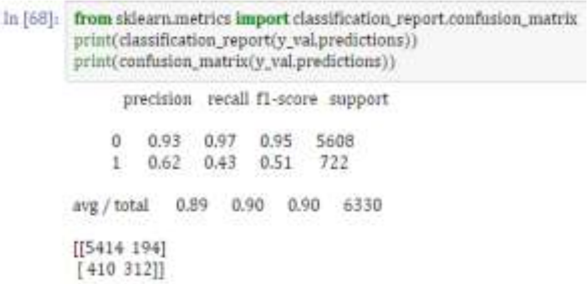
First we have to set up the model for this algorithm.



```
In [67]:  from sklearn.ensemble import RandomForestClassifier
           from sklearn.neural_network import MLPClassifier
           model = RandomForestClassifier(n_estimators=64, n_jobs=-1) # 0.8827, 29 seconds
           # model = MLPClassifier(max_iter=700) # 0.8557, 190 seconds
           model.fit(X_train, y_train.values.ravel())

           # Predict
           predictions = model.predict(X_val)
```

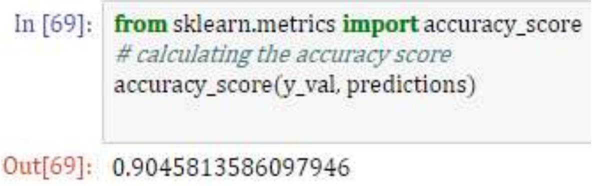
Fig 41:Setting up for Random Forest algorithm

Next we will generate a confusion matrix.



```
In [68]:  from sklearn.metrics import classification_report,confusion_matrix
           print(classification_report(y_val,predictions))
           print(confusion_matrix(y_val,predictions))

              precision   recall  f1-score   support

          0      0.93      0.97      0.95      5608
          1      0.62      0.43      0.51       722

avg / total      0.89      0.90      0.90      6330

[[5414  194]
 [ 410  312]]
```

Fig 42:Confusion Matrix



```
In [69]:  from sklearn.metrics import accuracy_score
           # calculating the accuracy score
           accuracy_score(y_val, predictions)

Out[69]:  0.9045813586097946
```

Fig 43:Accuracy using Random Forest Algorithm

An accuracy of about 90.4 percent is obtained using Random Forest Algorithm.

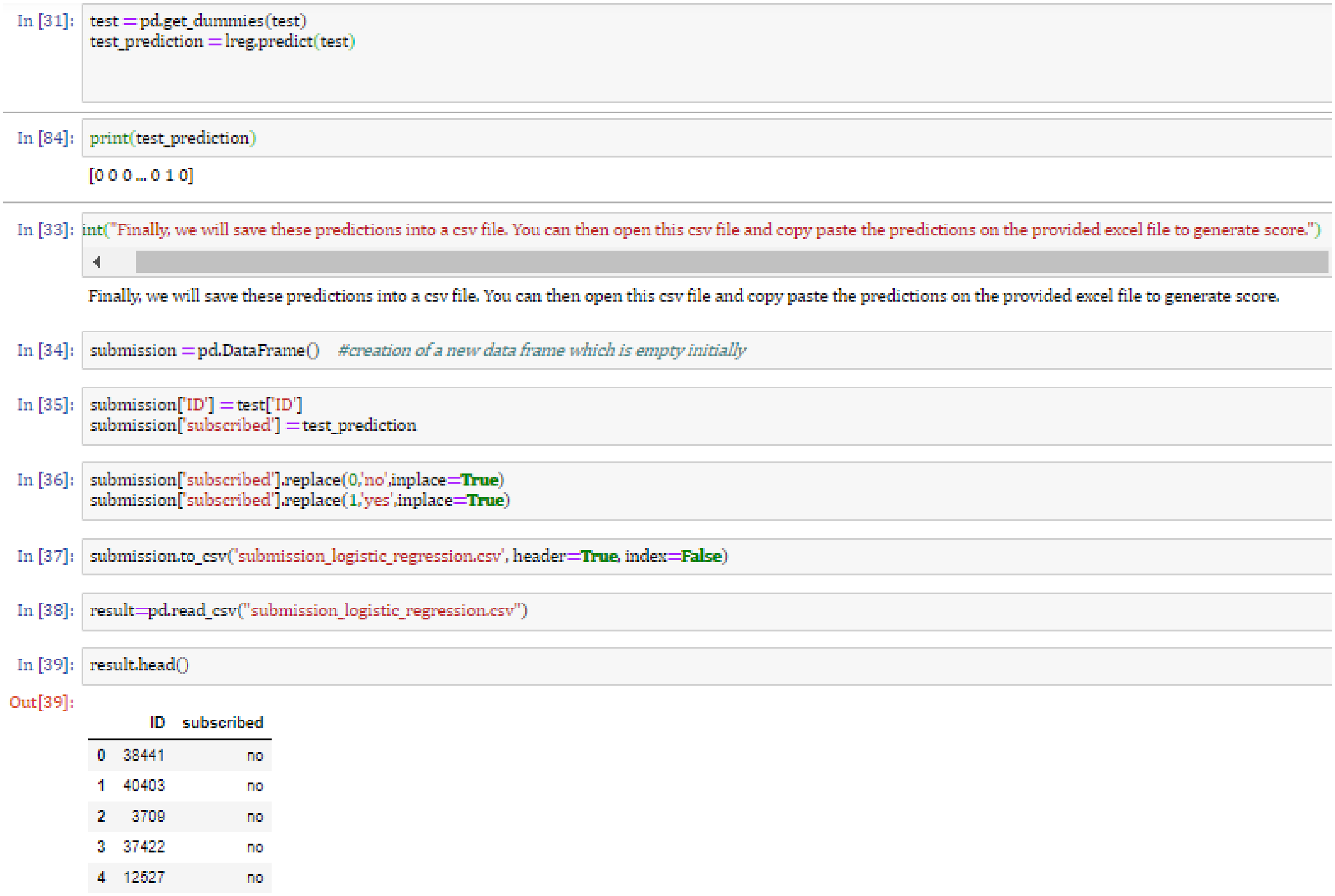
Once the prdiction is made then next step is to give the test data set to the model and the predictions made will be stored in a variable,say,”test_prediction”.

Next, an empty data frame will be created using pandas.DataFrame().

There we will make two columns in this data frame,ID and Subscribed.

The value of IDs will be same as those of training data set and values of subscribed will be taken as the output “test_prediction”.

This newly created data frame will be converted to csv file using .to_csv() method.



```
In [31]: test = pd.get_dummies(test)
         test_prediction = lreg.predict(test)

In [84]: print(test_prediction)

         [0 0 0 ... 0 1 0]

In [33]: int("Finally, we will save these predictions into a csv file. You can then open this csv file and copy paste the predictions on the provided excel file to generate score.")

         Finally, we will save these predictions into a csv file. You can then open this csv file and copy paste the predictions on the provided excel file to generate score.

In [34]: submission = pd.DataFrame()   #creation of a new data frame which is empty initially

In [35]: submission['ID'] = test['ID']
         submission['subscribed'] = test_prediction

In [36]: submission['subscribed'].replace(0,'no',inplace=True)
         submission['subscribed'].replace(1,'yes',inplace=True)

In [37]: submission.to_csv('submission_logistic_regression.csv', header=True, index=False)

In [38]: result=pd.read_csv("submission_logistic_regression.csv")

In [39]: result.head()

Out[39]:
              ID  subscribed
         0 38441          no
         1 40403          no
         2  3709          no
         3 37422          no
         4 12527          no
```

Fig 44:Saving the result

Similarly after checking the accuracy of every algorithm,the test data prediction will be made and result will be stored in a new csv file.

CHAPTER 5:CONCLUSION

Lately people depend on bank loans to meet their wishes. The fee of loan packages will increase with a very rapid speed in current years. Risk is constantly involved in approval of loans. The banking officials are very acutely aware of the price of the mortgage quantity by its customers. Even after taking lot of precautions and analyzing the mortgage applicant information, the mortgage approval choices are not continually correct. There is need of automation of this system so that loan approval is much less risky and incur less loss for banks.

Artificial Intelligence AI is a rising technology. The utility of AI solves many real world troubles. Machine Learning is an AI method which could be very useful in prediction systems.A model is created from a training data. While making the prediction the model that is evolved by way of training algorithm(ML) is used. The ML algorithm trained the machine the usage of a fragment of the statistics available and the remaining data is tested.

In this project some ML algorithms like Logistic Regression, Decision Tree,Random Forest,etc are implemented to expect the loan approval for customers. The experimental results conclude that the accuracy of Decision Tree ML algorithm is better in comparison to other algorithms.

As it is certainly a very important procedure for a bank to check whether an applicant,applying for a term loan should be approved with the loan or not,this project focuses on making the tedious task mechanized.

All the required algorithms were implemented successfully and accurate results were generated.

REFERENCES

1. Data science, “Benefits and uses of data science” https://www.simplilearn.com/why-and-how-data-science-matters-to-business-article

2. Data science, “Types of data” https://www.wintellect.com/beginning-statistics-for-data-science-types-of-data

3. Data science, “The data science process” https://www.kdnuggets.com/2016/03/data-science-process.html

4. Machine learning, “Introduction” https://www.digitalocean.com/community/tutorials/an-introduction-to-machine-learning

5. Machine learning,“Application in data science”https://www.simplilearn.com/importance-of-machine-learning-for-data-scientists-article

6. Python, “Libraries used in machine learning” https://www.geeksforgeeks.org/best-python-libraries-for-machine-learning/

7. Machine learning “Machine learning algorithms” https://www.geeksforgeeks.org/best-python-libraries-for-machine-learning/

ORIGINALITY REPORT

14%	10%	2%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	b-ok.cc Internet Source	5%
2	manning.com Internet Source	2%
3	slidelegend.com Internet Source	1%
4	Submitted to MVJ College of Engineering, Bangalore Student Paper	1%
5	Submitted to Brunel University Student Paper	
6	Submitted to Visvesvaraya Technological University Student Paper	<1%
7	Submitted to IIT Delhi Student Paper	<1%
8	link.springer.com Internet Source	<1%

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

PLAGIARISM VERIFICATION REPORT

Date: 14-07-2020

Type of Document (Tick): PhD Thesis | M.Tech Dissertation/ Report | [✓ B.Tech Project Report] | Paper

Name: SHAIL VATSAL VASHIST **Department:** CSE **Enrolment No** 161480

Contact No. 8219925829/9418588800 **E-mail.** shailvatsal98@gmail.com

Name of the Supervisor: DR. AMAN SHARMA

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): LOAN APPROVAL PREDICTOR USING DATA SCIENCE AND MACHINE LEARNING.

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

- Total No. of Pages = 63 (EXCLUDING PLAGIARISM REPORT)
- Total No. of Preliminary pages = 8
- Total No. of pages accommodate bibliography/references = 1

ShailVatsalVashist
(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found **Similarity Index** at 14 (%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

Aman Sharma

(Signature of Guide/Supervisor)

Signature of HOD

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

Copy Received on	Excluded	Similarity Index (%)	Abstract & Chapters Details	
	• All Preliminary Pages • Bibliography/ Images/Quotes • 14 Words String		Word Counts	
Report Generated on			Character Counts	
		Submission ID	Page counts	
			File Size	

Checked by
Name & Signature

Librarian

Please send your complete Thesis/Report in (PDF) & DOC (Word File) through your Supervisor/Guide at plagcheck.juit@gmail.com