

PROFESSIONAL TRAINING REPORT
AT
Sathyabama Institute of Science and Technology
(Deemed to be University)

Submitted in partial fulfillment of the requirements for the award
of Bachelor of Technology Degree in

Information Technology

By

NAME :SHANMUGAPRIYAN N

(REG NO: 37120069)



DEPARTMENT OF INFORMATION TECHNOLOGY
SCHOOL OF COMPUTING

SATHYABAMA
INSTITUTE OF SCIENCE AND
TECHNOLOGY (DEEMED TO BE
UNIVERSITY)
Accredited with Grade "A" by NAAC
JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI -
600 119

AUGUST 2019



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with "A" Grade by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai - 600 119.

Phone: 044 - 2450 3150 / 51 / 52 / 54 / 55 Fax: 044 - 2450 2344

www.sathyabama.ac.in



SCHOOL OF COMPUTING

BONAFIDE CERTIFICATE

This is to certify that this Professional Training Report is the bonafide work of Shanmugapriyan N (Reg. No. 37120069.) who carried out the project entitled "PREDCTION OF LOAN APPROVAL USING MACHINE LEARNING APPROACH" under our supervision from Jul 2020 to August 2020.

Internal Guide

Dr. JABEZ J., M.E., Ph.D.,

Head of the Department

Dr. R. SUBHASHINI, M.E., Ph.D.,

Submitted for Viva voice Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I am N.Shanmugapriyan hereby declare that the mini project report entitled on “PREDICTION OF LOAN APPRVAL USING MACHINE LEARNING APPROACH” done by me under the guidance of Dr.Jabez j.,ME.,Ph.D., at Sathyabama Institute of Science And Technology, Semmancheri, Chennai – 600 119 and Mr. Sathish, Senior Embedded Engineer, LCS Controls PVT. LTD. is submitted in partial fulfillment of the requirements for the award of Bachelor of Technology degree in Information Technology.

DATE:

PLACE:

N. Shanmugapriyan

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

The satisfaction and elation that accompany the successful completion of any task would be incomplete without the mention of the people who have made it possible. It is our great privilege to express our gratitude and respect to all those who have guided me during the course of my Professional Training. First and foremost, we would express our sincere gratitude to our beloved Founder Chancellor **Col. Dr. JEPPIAAR, M.A., B.L., Ph.D.**, I extend my sincere thanks to our Chancellor **Dr. Mariazeena Johnson, B.E., M.B.A., M. Phil., Ph.D.**, and the President **Dr. Marie Johnson, B.E., M.B.A., M. Phil., Ph.D.**, and for providing me the necessary facilities for the completion of the professional training. I also acknowledge our Vice Chancellor **Dr. T.Sasipraba M.E.,Ph.D.**, and the Pro Vice Chancellor **Dr. E.LOGASHANMUGAM, M.E., Ph.D.**, for their constant support and endorsement.

I would like to express our gratitude to our Registrar **Dr. S. S. Rau, Ph.D.** and Controller of Examinations **Dr. IgniSabasti Prabu M.E., Ph.D.**, for their valuable support offered to complete my professional training successfully.

I like to express our gratitude to **Dr. T. Sasikala, M.E., Ph.D.**, Dean of School of Computing and **Dr. R. Subhashini, M.E., Ph.D.**, Head of the Department of Information Technology, Sathyabama Institute Of Science and Technology for having been a constant source of support and encouragement for the completion of the professional training. I would also like to express our sincere thanks to our internal guide **Dr. J.JABEZ, M.E., Ph.D.**, who guided me in the preparation of the report. Also, I thank the **Almighty** and my **Parents** for supporting me in the completion of the professional Training.

This certificate is computer generated and can be verified by scanning the QR code given below. This will display the certificate from the NPTEL repository, <https://nptel.ac.in/noc/>


Roll No: NPTEL19CS40S21180444

To
SHANMUGAPRIYAN .N
OLD MAHABALIPURAM RD, KAMARAJ NAGAR,
SEMMANCHERI, CHENNAI, TAMIL NADU 600119
CITY
CHENNAI
TAMIL NADU
600119
PH. NO :9488085681



Score	Type of Certificate
≥ 90	Elite+Gold
75-89	Elite+Silver
≥ 60	Elite
40-59	Successfully Completed
< 40	No Certificate

No. of credits recommended by NPTEL:2

An additional 1 credit may be awarded if the University deems it fit, based on the actual student effort involved.



Elite

NPTEL Online Certification

(Funded by the Ministry of HRD, Govt. of India)



This certificate is awarded to

SHANMUGAPRIYAN .N

for successfully completing the course

Programming, Data Structures and Algorithms Using Python

with a consolidated score of **62** %

Online Assignments	21.88/25	Programming Exam	18.75/25	Proctored Exam	21.25/50
--------------------	----------	------------------	----------	----------------	----------

Total number of candidates certified in this course: **2747**



Prof. Devendra Jaliha
Chairman
Centre for Continuing Education, IITM

Jul-Sep 2019
(8 week course)



Prof. Andrew Thangaraj
NPTEL Coordinator
IIT Madras



Indian Institute of Technology Madras



Roll No: NPTEL19CS40S21180444

To validate and check scores: <https://nptel.ac.in/noc>

Certificate of Training

Shanmugapriyan from **Sathyabama Institute of Science and Technology** has successfully completed **Digital Marketing** internship at **Intracure** through Internshala.

A handwritten signature in green ink, appearing to read "Sarvesh".

Sarvesh Agarwal
Founder & CEO

Date of certification: 2020-10-12

Certificate Number: 0787F8741-0A8G-5402-3720EPQ21-DM1012
For certificate authentication please visit https://internshala.com/verify_certificate

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	LIST OF FIGURES	ix
	ABSTRACT	1
1	INTRODUCTION	2
1.1	DOMAIN REVIEW	2
1.2	PREPARING THE DATA SET	3
1.2.1	GATHERING DATA	3
1.3	OUTLINE OF THE PROJECT	4
1.3.1	OVERVIEW OF THE SYSTEM	5
2	AIM AND SCOPE OF THE PROJECT	6
2.1	AIM OF THE PROJECT	6
2.2	PROBLEM DESCRIPTION	6
2.3	SCOPE OF THE PROJECT	7
2.4	APPLICATION	7
3	SYSTEM DETAILS	8
3.1	EXISTING SYSTEM	8
3.1.1	DRAWBACKS	9
3.2	PRPOSED SYSTEM	9
3.2.1	ADVANTAGES	10
3.3	PROJECT GOALS	10
4	IMPLEMENTATION	12
4.1	MODULE DESCRIPTION	12
4.1.1.	VARIABLE IDENTIFICATION PROCESS	12
4.1.2	EXPLORATION DATA ANALYSIS OF VISUALIZATION	14
4.1.3	OUTLIER DETECTION PROCESS	16
4.1.4	COMPARING ALGORITHM WITH PREDCTION IN THE FORM OF BEST ACCURACY RESULT	17
4.1.5	PREDICTIO RESULT BY ACCURACY	18
4.2	ALGORITHM AND TECHNIQUES	19
4.2.1	ALGORITHM EXPLANATION	19
4.2.2	LOGISTIC REGRESSION	20

4.2.3	DECISION TREE	20
4.2.4	K-NEAREST NEIGHBOR	21
4.2.5	SUPPRT VECTOR MACHINS	22
5	REQUIREMENT SPECIFICATION	24
5.1	GENERAL	24
5.2	FUNCTIONAL REQUIREMENTS	24
5.3	NON FUNCTIONSL REQUIREMENTS	24
5.4	ENVIRONMENTQL REQUIREMENTS	24
6	PROJECT DESCRIPTION	26
6.1	ARCHITECTURE DIAGRAM	26
6.2	FLOW CHART	27
7	RESULTS AND SCREESHOTS	29
8	CONCLUSION AND FUTURE WORK	31
8.1	CONCLUSION	31
8.2	FUTURE WORK	31

LIST OF FIGURES

FIGURE NO.	FIGURE NAME	PAGE NO.
4.1.1.	VARIABLE IDENTIFICATION PROCESS	12
4.1.2	EXPLORATION DATA ANALYSIS OF VISUALIZATION	14
4.1.3	OUTLIER DETECTION PROCESS	16
4.1.4	COMPARING ALGORITHM WITH PREDCTION IN THE FORM OF BEST ACCURACY RESULT	17
4.1.5	PREDICTIO RESULT BY ACCURACY	18
4.2.1	ALGORITHM EXPLANATION	19
4.2.2	LOGISTIC REGRESSION	20
4.2.3	DECISION TREE	20
4.2.4	K-NEAREST NEIGHBOR	21
4.2.5	SUPPRT VECTOR MACHINS	22

PREDICTION OF LOAN APPROVAL USING MACHINE LEARNING APPROACH

ABSTRACT:

Nowadays loan approved prediction, play an important role for machine learning tasks. The loan is very important products of the banking sectors. All the banks are trying to connect customers for applying their loans. However, there are some customers behave negatively after their application are approved. To prevent this problem, banking sectors have to predict customer's behaviors using machine learning techniques. The analysis of dataset by supervised machine learning technique(SMLT) to capture several information's like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments etc. so, supervised machine learning classification problems of algorithms aim to categorizing data from priority information by using python. The main objective is to predict whether assigning the loan to particular person will be safe or not. Additionally, to compare and discuss the performance of comparative study with finding the best accuracy apply in various supervised machine learning technique from the given dataset with GUI based application by given dataset attributes.

CHAPTER 1

INTRODUCTION

1.1 DOMAIN REVIEW

Machine learning is to predict the future from past data. Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using python. Process of training and prediction involves use of specialized algorithms. It feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. Machine learning can be roughly separated in to three categories. There are supervised learning, unsupervised learning and reinforcement learning. Supervised learning program is both given the input data and the corresponding labeling to learn data has to be labeled by a human being beforehand. Unsupervised learning is no labels. It provided to the learning algorithm. This algorithm has to figure out the clustering of the input data. Finally, Reinforcement learning dynamically interacts with its environment and it receives positive or negative feedback to improve its performance.

Data scientists use many different kinds of machine learning algorithms to discover patterns in python that lead to actionable insights. At a high level, these different algorithms can be classified into two groups based on the way they “learn” about data to make predictions: supervised and unsupervised learning. Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function from input variables(X) to discrete output variables(y). In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc.

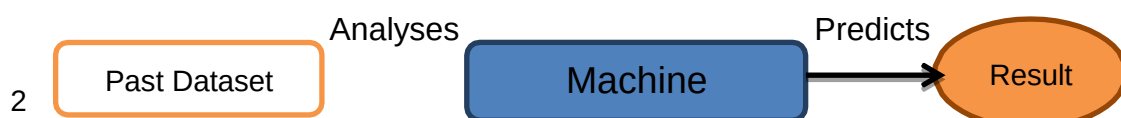




Fig: Process of Machine learning

Supervised Machine Learning is the majority of practical machine learning uses supervised learning. Supervised learning is where have input variables (X) and an output variable (y) and use an algorithm to learn the mapping function from the input to the output is $y = f(X)$. The goal is to approximate the mapping function so well that when you have new input data (X) that you can predict the output variables (y) for that data. Techniques of Supervised Machine Learning algorithms include logistic regression, multi-class classification, Decision Trees and support vector machines etc. Supervised learning requires that the data used to train the algorithm is already labeled with correct answers. Supervised learning problems can be further grouped into Classification problems. This problem has as goal the construction of a succinct model that can predict the value of the dependent attribute from the attribute variables. The difference between the two tasks is the fact that the dependent attribute is numerical for categorical for classification. A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes. A classification problem is when the output variable is a category, such as “red” or “blue”.

Now a day's bank plays a vital role in market economy. The success or failure of organization largely depends on the industry's ability. Before giving the loan to borrowers, bank decides whether the borrower is bad or good. The prediction of borrower status i.e. in future borrower will be cheater or non-cheater is a challenging task for any organization or bank. Basically the loan approval prediction is a binary classification problem. Loan amount, costumer's history governs his creditability for receiving loan. The problem is to classify borrower as fraud or non-fraud. However developing such a model is a very challenging task due to increasing in demands for loans. A prototype of the model is described which can be used by the organizations for making the correct or right decision for approve or reject the request for loan of the customers. This work includes the construction of a collective model by combining three different machine learning models.

Exploring the data by calculating summary and descriptive statistics, and by creating visualizations of the data, several potential relationships between bank accountant characteristics and loan default were identified. Later, a classification model is created to predict loan from its feature. The efficiency of the model is increased by identifying the most

valuable and contributing features among the given features in the data set. Banks struggle a lot to get an upper hand over each other to enhance overall business due to tight competition. Banks have realized that retaining the customers and preventing fraud must be the strategy tool for healthy competition. Credit Risk assessment is a crucial issue faced by Banks nowadays which helps them to evaluate if a loan applicant can be a fraud at a later stage so that they can go ahead and grant the loan or not. The dataset used for analysis contains many inconsistencies like missing values, outliers and inconsistencies and it have to be handled before being used to build the model. To classify if, the applicant is a fraud or not. Loan Prediction is very helpful for employee of banks as well as for the applicant also. To provide quick, immediate and easy way to choose the deserving applicants. It can provide special advantages to the bank. The Loan Prediction System can automatically calculate the weight of each features taking part in loan processing and on new test data same features are processed with respect to their associated weight.

A time limit can be set for the applicant to check whether his/her loan can be sanctioned or not. Loan Prediction System allows jumping to specific application so that it can be check on priority basis. This exclusively for the managing authority of Bank/finance Company, whole process of prediction is done privately no stakeholders would be able to alter the processing. Result against particular Loan Id can be send to various departments of banks so that they can take appropriate action on application. This helps all others department to carried out other formalities.

1.2 PREPARING THE DATASET

The dataset is now supplied to machine learning model on the basis of this data set the model is trained. Every new applicant details filled at the time of application form acts as a test data set. After the operation of testing, model predict whether the new applicant is a fit case for approval of the loan or not based upon the inference it conclude on the basis of the training data sets.

1.2.1 Gathering Data: (Sample dataset attributes)

In the first step of accumulating information, data from previously approved loan datasets from online sources are gathered together. These datasets are merged to form a common dataset, on which analysis will be done.

Table shows details of the datasets:

Variable	Description
Loan_ID	Unique loan ID

Gender	Male / Female
Married	Applicant Married (Yes/No)
Dependents	Number of dependents
Education	Applicant Education (Graduate/ Not Graduate)
Self Employed	Self Employed (Yes/No)
Application Income	Applicant Income
Co-applicant Income	Co-Applicant Income
Loan Amount	Loan amount details
Loan_Amount_Term	Terms of loan in months
Credit_History	Credit history meets guidelines
Property_Area	Urban / Rural / Semi urban
Loan_Status	Loan approved (Yes/No)

1.3 OUTLINE OF THE PROJECT

1.3.1 Overview of the system

Loan Prediction is very helpful for employee of banks as well as for the applicant also. The aim is to provide quick, immediate and easy way to choose the deserving applicants. It can provide special advantages to the bank. The Loan Prediction System can automatically calculate the weight of each features taking part in loan processing and on new test data same features are processed with respect to their associated weight. A time limit can be set for the applicant to check whether his/her loan can be sanctioned or not. Loan Prediction System allows jumping to specific application so that it can be check on priority basis and exclusively for the managing authority of Bank/finance Company, whole process of prediction is done privately no stakeholders would be able to alter the processing. Result against particular Loan Id can be send to various departments of banks so that they can take appropriate action on application. This helps all others department to carried out other formalities. It have to find Accuracy of the training dataset, Accuracy of the testing dataset, Specification, False Positive rate, precision and recall by comparing algorithm using python code. The following Involvement steps are,

- Define a problem
- Preparing data
- Evaluating algorithms
- Improving results
- Predicting results

The steps involved in Building the data model is depicted below.

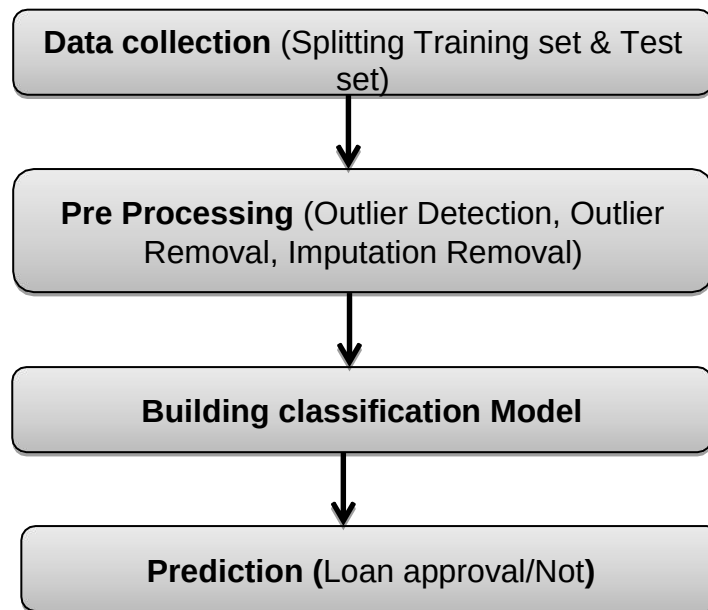


Fig: data flow diagram for Machine learning model

Distribution of the loans is the core business part of almost every bank. The main portion the bank's assets are directly came from the profit earned from the loans distributed by the banks. The prime objective in banking environment is to invest their assets in safe hands where it is. Today many banks/financial companies approves loan after a regress process of verification and validation but still there is no surety whether the chosen applicant is the deserving right applicant out of all applicants. Through this system we can predict whether that particular applicant is safe or not and the whole process of validation of features is automated by machine learning technique. The disadvantage of this model is that it emphasizes different weights to each factor but in real life sometime loan can be approved on the basis of single strong factor only, which is not possible through this system.

CHAPTER 2

AIM AND SCOPE

2.1 AIM

- This analysis aims to observe which features are most helpful in predicting loan approval and to see the general trends that may help us in model selection and hyper parameter selection. The goal is to classify whether the approved or not with identifying the fraud persons.
- Apply the fundamental concepts of machine learning from an available dataset and Evaluate and interpret my results and justify my interpretation based on observed dataset.
- Create notebooks that serve as computational records and document my thought process and investigate applications of statistics for loan to analyses the data set.
- Evaluate and analyses statistical and visualized results, which find the standard patterns for all regiments.

2.2 PROBLEM DESCRIPTION

Importance of loans in our day-to-day life has increased to a great extent. People are becoming more and more dependent on acquiring loans, be it education loan, housing loan, car loan, business loans etc. from the financial institutions like banks and credit unions. However, it is no longer surprising to see that some people are not able to properly gauge the amount of loan that they can afford. In some cases, people undergo sudden financial crisis while some try to scam money out of the banks. Assessing the risk, which is involved in a loan application, is one of the most important concerns of the banks for survival in the highly competitive market and for profitability. These banks receive number of loan applications from their customers and other people on daily basis. Not everyone gets approved. Most of the banks use their own credit scoring and risk assessment techniques in order to analyze the loan application and to make decisions on credit/loan approval.

Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online/offline application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the bank accountant those are eligible for loan amount so that they can specifically target these customers. In the

banking sector lots of people are applying for bank loans but the bank has its limited resources which it has to grant to limited people only, so finding out to whom the loan can be granted which will be a safer option for the bank is a typical process. In that limited resources have fraud happened.

2.3 SCOPE:

The scope of this project is to investigate a dataset of loan records for banking sector using machine learning technique. To identifying fraud behavior person is more difficult. we try to reduce this risk factor behind selecting the safe person so as to save lots of bank efforts and assets and to predict whether assigning the loan to particular person will be safe or not.

2.4 APPLICATION:

A bank wants to automate the loan eligibility process (real time) based on the customer detail provided while filling online/ offline application form. To automate this process, they have given a problem to identify the bank accountant. Those are eligible for loan approved with getting expected amount.

CHAPTER 3

SYSTEM DETAILS

3.1 EXISTING SYSTEM

The check pointing cost in Shadow at first, to analyze the effectiveness of using the pre-copying approach for system downtime reduction. The system downtime reductions of pre-copy based check pointing and Shadow, compared to the stop-and copy check pointing. Check pointing is a widely-used technique in many application domains, such as fault tolerance, application debugging and transactional memory systems. The performance of hard-disk based check pointing is limited by low I/O bandwidth between main memory and backup storage. Due to the increasing system scale and memory footprint, frequent check pointing can significantly degrade the system performance. Recently, there have on reducing the check pointing cost by leveraging emerging non-volatile memory technologies. Those are demonstrating that storing checkpoint data in Non-Volatile Memory (NVM) can significantly reduce the check pointing cost compared to the traditional disk-based approaches. As the pre-copy based check pointing and co-running applications contend with each other for the shared memory banks, the inter-application memory interference can cause bank conflicts and application performance degradation. In order to characterize the bank-level memory interference, it developed a pre-copy based check pointing system and run multi-programmed workloads generated and Problem Based Benchmarks Suite on Gem5 simulator.

It can mitigate the memory access interference with the applications by scoring the pre-copying memory requests to fully exploit the idle periods of each bank. The key idea seems simple but requires addressing several challenging problems as follows:

- How to predict the length of bank idle periods and utilize them to serve the memory requests of the pre-copying process.
- How to handle bank conflicts between applications and the pre-copying process in case that all banks are busy.
- How to leverage the hardware features of the multichannel memory controller to implement a low-cost data copying mechanism on a DRAM/NVM hybrid memory system.

Shadow adopts a pre-copy based check pointing mechanism to reduce the system downtime. It supports fine-grained and incremental check pointing at frequencies up to 100 times per second. Under this context, the check pointing can significantly degrade application performance due to memory contention between applications and the check pointing process. Previous check pointing mechanisms on hybrid memory systems have focused on the

performance of check pointing, and have overlooked the impact of memory contention on the application performance. It mitigate the memory contention at the bank level by carefully scheduling memory requests to fully leverage the idle time slots of different memory banks. Moreover, if bank conflicts are unavoidable, Shadow promotes the priority of applications memory requests to lessen their access latencies. By redesigning the memory controllers of DRAM and NVM, we implement a hardware-assisted check pointing mechanism that can directly transfer data from working memory to the checkpoint in NVM, without any intervention of CPUs. Our evaluation shows that Shadow can reduce memory bank conflicts between applications and check pointing by 75%, and decrease applications memory read request latency by 28% on average compared to the pre-copy based check pointing. Moreover, Shadow can also reduce check pointing overhead by 42% and 16% on average compared to the stop-and-copy and pre-copy based check pointing approaches, respectively. It describes the execution phases of the check pointing in Shadow. Shadow conducts iterative pre-copying at the beginning of each check pointing interval and used a dirty bitmap to track all pages modified in each iteration of pre-copying. Shadow copies the modified pages to NVM according to the bitmap in the next round of pre-copying. When the size of remaining dirty memory is reduced to a very small value, or the iterative pre-copying cannot reduce the size of dirty memory anymore, Shadow terminates the pre-copying and begins the stop-and-copy phase. The system is suspended to copy the remaining dirty pages, processors states, and metadata of check pointing to NVM. Finally, the system is restored to continue the applications execution.

3.1.1 Drawbacks:

- Its only describes improvement of memory based development of data storage of checkpoint. But, it's helpful to analysis of data science with historical data storage.
- Its cannot work on top features and find-out the accuracy, Recall, Precision, Confusion matrix and compare it with our old result.
- It's cannot work on using the popular machine learning algorithm to find out the features importance.

3.2 PROPOSED SYSTEM

This analysis is not meant to be providing a final conclusion on the reasons leading to bank accountant's loan approval as it doesn't involve using any inferential statistics techniques/machine learning algorithms. Machine learning supervised classification algorithms will be used to give the loan-approved dataset and extract patterns, which would help in predicting the likely defaulters/ Not-Fraudulent, thereby helping the banks for making

better decisions in the future. Multiple datasets from different sources would be combined to form a generalized dataset, and then different machine learning algorithms would be applied to extract patterns and to obtain results with maximum accuracy.

3.2.1 Advantages

- The goal of this problem is to predict the status of fraud detection in loan approval of test dataset as accurate as possible and time limit can be set for the applicant to check whether his/her bank accountant can be fraud or not and to provide quick, immediate and easy way to choose the deserving applicants.
- To reduce this risk factor behind selecting the safe person so as to save lots of bank efforts and reduce the risk of bank manager.

3.3 Project Goals

The primary goal is to extract patterns from a common loan approved dataset, and then build a model based on these extracted patterns, in order to predict the likely loan defaulters by using classification machine learning supervised algorithms. The historical data of the customers like their age, income, loan amount, employment etc. It will be used in order to do the analysis. Later on, some analysis will also be done to find the most relevant attributes, the factors that affect the prediction result the most.

- Exploration data analysis of variable identification
 - Loading the given dataset
 - Import required libraries packages
 - Analyze the general properties
 - Find duplicate and missing values
 - Checking unique and count values
- Uni-variate data analysis
 - Rename, add data and drop the data
 - To specify data type
- Exploration data analysis of bi-variate and multi-variate
 - Plot diagram of pairplot, heatmap, bar chart and Histogram
- Method of Outlier detection with feature engineering
 - Pre-processing the given dataset
 - Splitting the test and training dataset

- Comparing the Decision tree and Logistic regression model and random forest
- Comparing algorithm to predict the result
 - Based on the best accuracy

CHAPTER 4

IMPLEMENTATION

4.1 MODULES DESCRIPTION

List of Modules:

- Variable Identification Process
- Exploration data analysis of visualization
- Outlier detection process
- Prediction of GUI based Application results

4.1.1 Variable Identification Process

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters. The following diagram is given dataset.

The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers use this data to fine-tune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model. For example, time series data can be analyzed by regression algorithms; classification algorithms can be used to analyze discrete data. For example to show the data type format of given dataset.

A number of different data cleaning tasks using Python's Pandas library and specifically, it focus on probably the biggest data cleaning task, missing values and it able to more quickly clean data. It wants to spend less time cleaning data, and more time exploring and modeling. Some of these sources are just simple random mistakes. Other times, there can be a deeper reason why data is missing. It's important to understand these different types of missing data from a statistics point of view. The type of missing data will influence how to deal with

filling in the missing values and to detect missing values, and do some basic imputation and detailed statistical approach for dealing with missing data. Before, joint into code, it's important to understand the sources of missing data. Here are some typical reasons why data is missing:

- User forgot to fill in a field.
- Data was lost while transferring manually from a legacy database.
- There was a programming error.
- Users chose not to fill out a field tied to their beliefs about how the results would be used or interpreted.

Variable identification with Uni-variate, Bi-variate and Multi-variate analysis:

- import libraries for access and functional purpose and read the given dataset
- General Properties of Analyzing the given dataset
- Display the given dataset in the form of data frame
- show columns
- shape of the data frame
- To describe the data frame
- Checking data type and information about dataset
- Checking for duplicate data
- Checking Missing values of data frame
- Checking unique values of data frame
- Checking count values of data frame
- Rename and drop the given data frame
- To specify the type of values
- To create extra columns

Data Validation/ Cleaning/Preparing Process

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from

dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

4.1.2 Exploration data analysis of visualization

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

Sometimes data does not make sense until it can look at in a visual form, such as with charts and plots. Being able to quickly visualize of data samples and others is an important skill both in applied statistics and in applied machine learning. It will discover the many types of plots that you will need to know when visualizing data in Python and how to use them to better understand your own data.

- How to chart time series data with line plots and categorical quantities with bar charts.
- How to summarize data distributions with histograms and box plots.
- How to summarize the relationship between variables with scatter plots.

Data Visualization types:

- Introduction to Matplotlib
- Line Plot
- Bar Chart
- Histogram Plot
- Box and Whisker Plot
- Scatter Plot

There are many excellent plotting libraries in Python and it recommend exploring them in order to create presentable graphics. For quick and dirty plots intended for your own use, it recommends using the matplotlib library. It is the foundation for many other plotting libraries and plotting support in higher-level libraries such as Pandas. The matplotlib provides a context, one in which one or more plots can be drawn before the image is shown or saved to file and

context can be accessed via functions on pyplot.

Bar Chart:

A bar chart is generally used to present relative quantities for multiple categories. The x-axis represents the categories and are spaced evenly. The y-axis represents the quantity for each category and is drawn as a bar from the baseline to the appropriate level on the y-axis.

Box and Plot:

A box and whisker plot, or boxplot for short, is generally used to summarize the distribution of a data sample. The x-axis is used to represent the data sample, where multiple boxplots can be drawn side by side on the x-axis if desired. The boxplot is a graphical technique that displays the distribution of variables. It helps us see the location, skewness, spread, tile length and outlying points. The boxplot is a graphical representation of the Five Number Summary.

Heat map

A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors. It is a bit like looking a data table from above. It is really useful to display a general view of numerical data, not to extract specific data point. It is quite straight forward to make a heat map, as shown on the examples below.

However be careful to understand the underlying mechanisms. You will probably need to normalize your matrix, choose a relevant color palette, use cluster analysis and thus permute the rows and the columns of the matrix to place similar values near each other according to the clustering.

Probability of Loan Analysis

When loading the given dataset, it will predictably have to learn about probability. It is easy to formulate and theory behind probability, but it has essential uses and explore how statistics relates to probability. At the most basic level, probability seeks to answer the question, "What is the chance of an event happening?" An event is some outcome of interest. To calculate the chance of an event of loan approval, it also needs to consider all the other events that can occur.

Probability for loan analysis:

- Preprocessing from label encoder technique
- To find percentage of loan approval
- To calculate Probability for loan

Actually, most of the binary classification models will give the prediction of probability first and then assign the probabilities to 1 or 0 based on the default threshold of 0.5. To improve the recall of the model, it can use the probabilities predicted by the model and set threshold by ourselves. The threshold is set based on several factors such as bank/business objectives. It is different case by case.

In the bank loan behavior prediction, for example, banks want to control the loss to an acceptable level, so they may use a relatively low threshold. This means more customers will be grouped as “potential bad customers” and their profiles will be checked carefully later by the credit risk management team. In this way, banks can detect the default behaviors in the earlier stage and conduct the corresponding actions to reduce the possible loss.

4.1.3 Outlier detection process

Many machine learning algorithms are sensitive to the range and distribution of attribute values in the input data. Outliers in input data can skew and mislead the training process of machine learning algorithms resulting in longer training times, less accurate models and ultimately poorer results.

Even before predictive models are prepared on training data, outliers can result in misleading representations and in turn misleading interpretations of collected data. Outliers can skew the summary distribution of attribute values in descriptive statistics like mean and standard deviation and in plots such as histograms and scatterplots, compressing the body of the data. Finally, outliers can represent examples of data instances that are relevant to the problem such as anomalies in the case of fraud detection and computer security.

It couldn't fit the model on the training data and can't say that the model will work accurately for the real data. For this, we must assure that our model got the correct patterns from the data, and it is not getting up too much noise. Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the

complementary subset of the data-set.

The three steps involved in cross-validation are as follows:

1. Reserve some portion of sample data-set.
2. Using the rest data-set train the model.
3. Test the model using the reserve portion of the data-set.

Advantages of train/test split:

1. This runs K times faster than Leave One Out cross-validation because K-fold cross-validation repeats the train/test split K-times.
2. Simpler to examine the detailed results of the testing process.

Advantages of cross-validation:

1. More accurate estimate of out-of-sample accuracy.
2. More “efficient” use of data as every observation is used for both training and testing.

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. To achieving better results from the applied model in Machine Learning method of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format, for example, Random Forest algorithm does not support null values. Therefore, to execute random forest algorithm null values have to be managed from the original raw data set. And another aspect is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in given dataset.

4.1.4 Comparing Algorithm with prediction in the form of best accuracy result

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be

on unseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created. When have a new dataset, it is a good idea to visualize the data using different techniques in order to look at the data from different perspectives. The same idea applies to model selection. You should use a number of different ways of looking at the estimated accuracy of your machine learning algorithms in order to choose the one or two to finalize. A way to do this is to use different visualization methods to show the average accuracy, variance and other properties of the distribution of model accuracies.

In the next section you will discover exactly how you can do that in Python with scikit-learn. The key to a fair comparison of machine learning algorithms is ensuring that each algorithm is evaluated in the same way on the same data and it can achieve this by forcing each algorithm to be evaluated on a consistent test harness.

In the example below 5 different algorithms are compared:

- Logistic Regression
- Random Forest
- K-Nearest Neighbors
- Decision tree
- Support Vector Machines

The K-fold cross validation procedure is used to evaluate each algorithm, importantly configured with the same random seed to ensure that the same splits to the training data are performed and that each algorithm is evaluated in precisely the same way. Before that comparing algorithm, Building a Machine Learning Model using install Scikit-Learn libraries. In this library package have to done preprocessing, linear model with logistic regression method, cross validating by KFold method, ensemble with random forest method and tree with decision tree classifier. Additionally, splitting the train set and test set. To predicting the result by comparing accuracy.

4.1.5 Prediction result by accuracy:

Logistic regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. We need the output of the algorithm to be classified variable data. Higher accuracy predicting result is logistic regression model by comparing the best accuracy.

Over-fitting is a common problem in machine learning which can occur in most models. K-fold cross-validation can be conducted to verify that the model is not over-fitted. In this method, the data-set is randomly partitioned into *kmutually exclusive* subsets, each approximately equal size and one is kept for testing while others are used for training.

GUI Application:

We have often been asked: "Is there no Tk for Python?" or "Is Tkinter the same as Tk?" Of course, there is Tk for Python. Without Tk Python would be less attractive to many users. Tk is called Tkinter in Python, or to be precise, Tkinter is the Python interface for Tk. Tkinter is an acronym for "Tk interface". If there were a single package which might be called the "standard" GUI toolkit for Python, it would be [Tkinter](#). Tkinter is a wrapper around [Tcl/Tk](#), a popular graphical interface and language

4.2 ALGORITHM AND TECHNIQUES

4.2.1 Algorithm Explanation

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc. In Supervised Learning, algorithms learn from labeled data. After understanding the data, the algorithm determines which label should be given to new data based on pattern and associating the patterns to the unlabeled new data.

Used Python Packages:

sklearn:

- In python, sklearn is a machine learning package which include a lot of ML algorithms.
- Here, we are using some of its modules like `train_test_split`, `DecisionTreeClassifier` or `Logistic Regression` and `accuracy_score`.

NumPy:

- It is a numeric python module which provides fast maths functions for calculations.
- It is used to read data in numpy arrays and for manipulation purpose.

Pandas:

- Used to read and write different files.
- Data manipulation can be done easily with data frames.

Matplotlib:

- Data visualization is a useful way to help with identify the patterns from given dataset.
- Data manipulation can be done easily with data frames.

4.2.2 Logistic Regression

It is a statistical method for analysing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.).

In other words, the logistic regression model predicts $P(Y=1)$ as a function of X . Logistic regression Assumptions:

- Binary logistic regression requires the dependent variable to be binary.
- For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.
- Only the meaningful variables should be included.
- The independent variables should be independent of each other. That is, the model should have little.
- The independent variables are linearly related to the log odds.
- Logistic regression requires quite large sample sizes.

4.2.3 Decision Tree

It is one of the most powerful and popular algorithm. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables. Assumptions of Decision tree:

- At the beginning, we consider the whole training set as the root.

- Attributes are assumed to be categorical for information gain, attributes are assumed to be continuous.
- On the basis of attribute values records are distributed recursively.
- We use statistical methods for ordering attributes as root or internal node.

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision node has two or more branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. Decision tree builds classification or regression models in the form of a tree structure. It utilizes an if-then rule set which is mutually exclusive and exhaustive for classification. The rules are learned sequentially using the training data one at a time. Each time a rule is learned, the tuples covered by the rules are removed.

This process is continued on the training set until meeting a termination condition. It is constructed in a top-down recursive divide-and-conquer manner. All the attributes should be categorical. Otherwise, they should be discretized in advance. Attributes in the top of the tree have more impact towards in the classification and they are identified using the information gain concept. A decision tree can be easily over-fitted generating too many branches and may reflect anomalies due to noise or outliers.

4.2.4 K-Nearest Neighbor (KNN)

K-Nearest Neighbor is a supervised machine learning algorithm which stores all instances correspond to training data points in n-dimensional space. When an unknown discrete data is received, it analyzes the closest k number of instances saved (nearest neighbors) and returns the most common class as the prediction and for real-valued data it returns the mean of k nearest neighbors. In the distance-weighted nearest neighbor algorithm, it weights the contribution of each of the k neighbors according to their distance using the following query giving greater weight to the closest neighbors.

Usually KNN is robust to noisy data since it is averaging the k-nearest neighbors. The k-nearest-neighbors algorithm is a classification algorithm, and it is supervised: it takes a bunch of labeled points and uses them to learn how to label other points. To label a new point, it looks at the labeled points closest to that new point (those are its nearest neighbors), and has those neighbors vote, so whichever label the most of the neighbors have is the label for the new point

(the “k” is the number of neighbors it checks). Makes predictions about the validation set using the entire training set. KNN makes a prediction about a new instance by searching through the entire set to find the k “closest” instances. “Closeness” is determined using a proximity measurement (Euclidean) across all features.

Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees’ habit of over fitting to their training set. Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name “Random Forest”. The random forest algorithm can be used for both regression and classification tasks.

The following are the basic steps involved in performing the random forest algorithm:

- Pick N random records from the dataset.
- Build a decision tree based on these N records.
- Choose the number of trees you want in your algorithm and repeat steps 1 and 2.
- In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output). The final value can be calculated by taking the average of all the values predicted by all the trees in forest. Or, in case of a classification problem, each tree in the forest predicts the category to which the new record belongs. Finally, the new record is assigned to the category that wins the majority vote.

4.2.5 Support Vector Machines

A classifier that categorizes the data set by setting an optimal hyper plane between data. I chose this classifier as it is incredibly versatile in the number of different kernelling functions that can be applied and this model can yield a high predictability rate. Support Vector Machines are perhaps one of the most popular and talked about machine learning algorithms. They were extremely popular around the time they were developed in the 1990s and continue to be the go-to method for a high-performing algorithm with little tuning.

- How to disentangle the many names used to refer to support vector machines.
- The representation used by SVM when the model is actually stored on disk.
- How a learned SVM model representation can be used to make predictions for new data.
- How to learn an SVM model from training data.
- How to best prepare your data for the SVM algorithm

CHAPTER 5

REQUIREMENT SPECIFICATION

5.1 GENERAL:

Requirements are the basic constraints that are required to develop a system. Requirements are collected while designing the system. The following are the requirements that are to be discussed.

1. Functional requirements
2. Non-Functional requirements
3. Environment requirements
 - A. Hardware requirements
 - B. software requirements

5.2 FUNCTIONAL REQUIREMENTS:

The software requirements specification is a technical specification of requirements for the software product. It is the first step in the requirements analysis process. It lists requirements of a particular software system. The following details to follow the special libraries like sk-learn, pandas, numpy, matplotlib and seaborn.

5.3 NON FUNCTIONAL REQUIREMENTS:

Process of functional steps,

1. Problem define
2. Preparing data
3. Evaluating algorithms
4. Improving results
5. Prediction the result

5.4 ENVIRONMENTAL REQUIREMENTS:

1. Software Requirements:

Operating System : Windows

Tool : Anaconda with Jupyter Notebook

2. Hardware requirements:

Processor : Pentium IV/III

Hard disk : minimum 80 GB

RAM : minimum 2 GB

Frontend: GUI Application

Backend: Anaconda tool

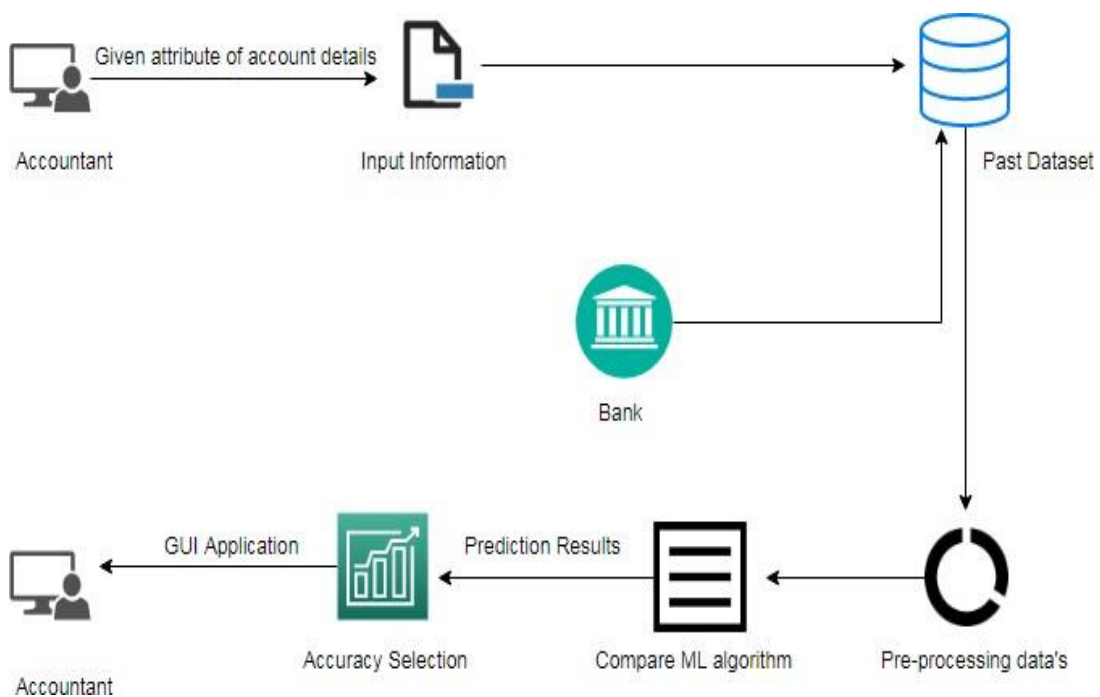
Language: Python

CHAPTER 6

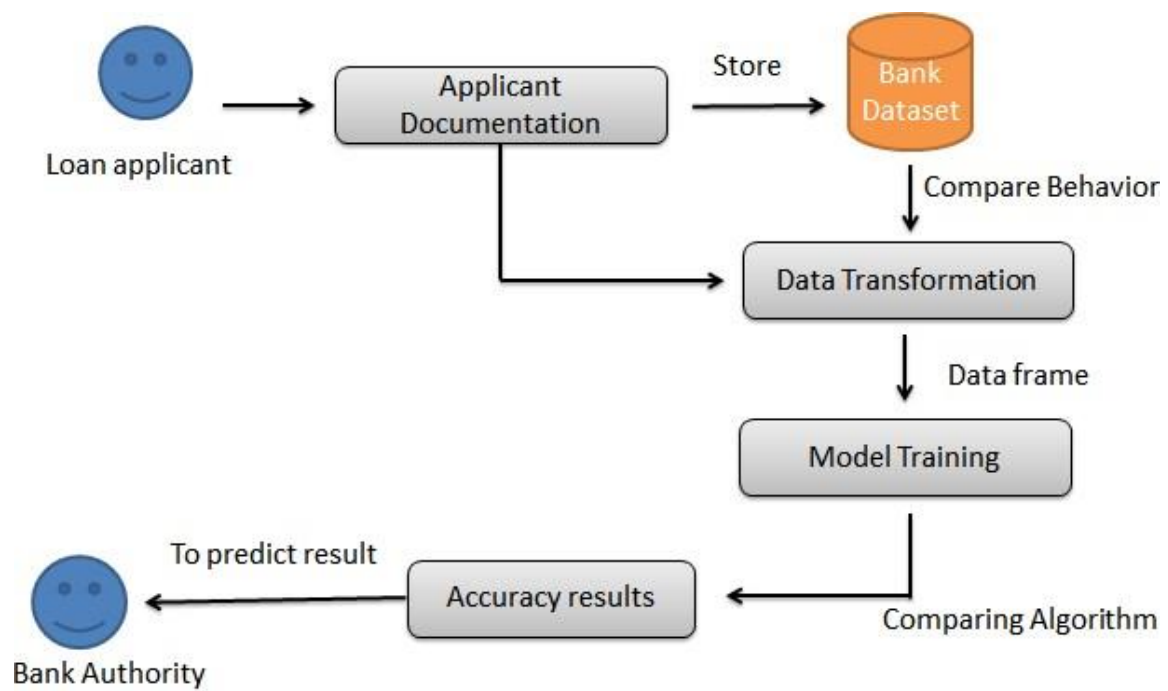
PROJECT DESCRIPTION

6.1 ARCHITECTURE DIAGRAM:

System Architecture / Business Architecture / Overall Architecture



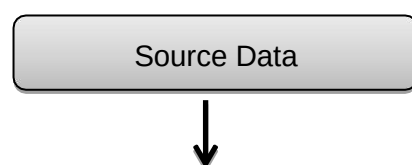
Design Architecture:



Design Architecture

6.2 FLOW CHART

Work flow diagram



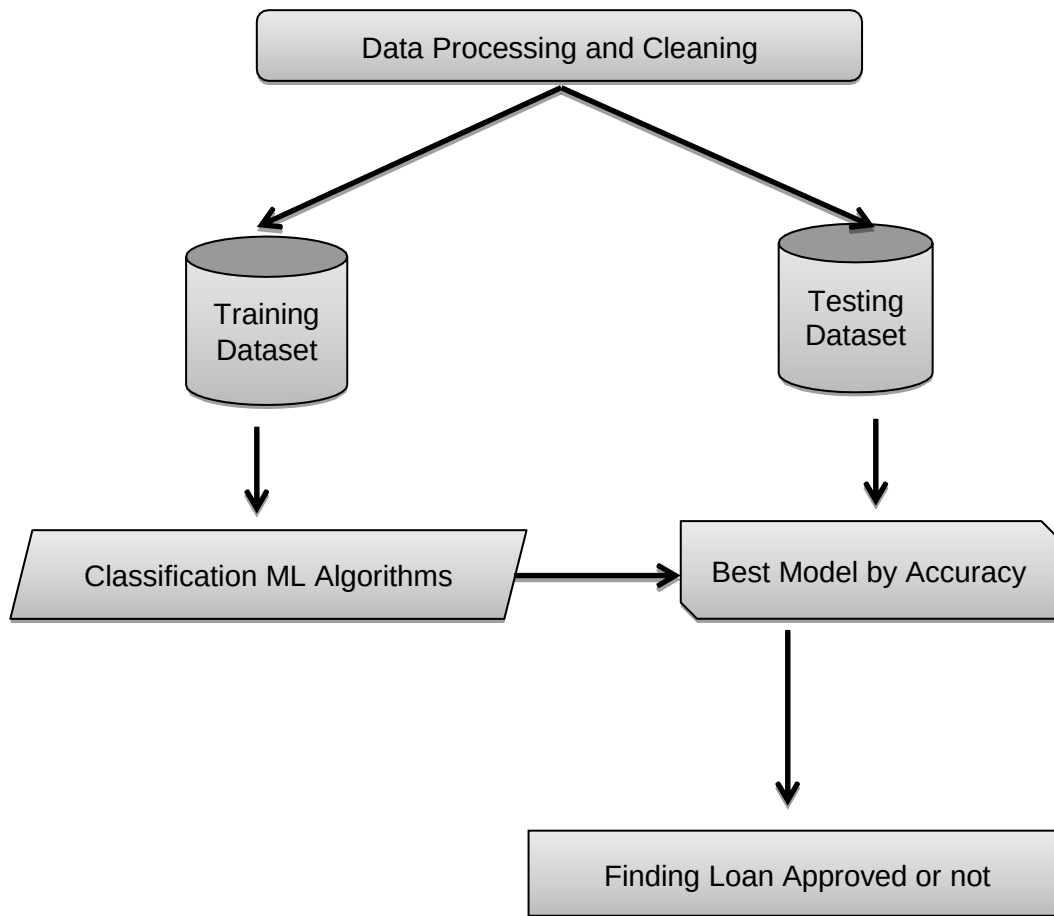


Fig: Workflow Diagram

CHAPTER 7

RESULTS AND SCREESSHOTS

Screen Shots:

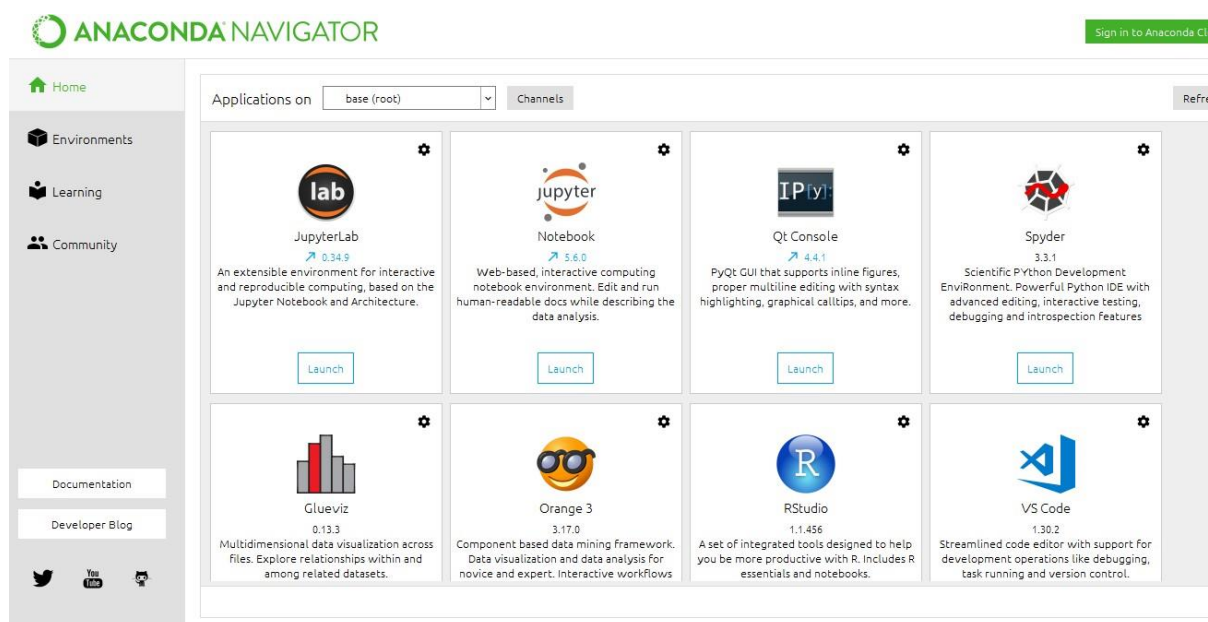


Fig: Open the anaconda navigator

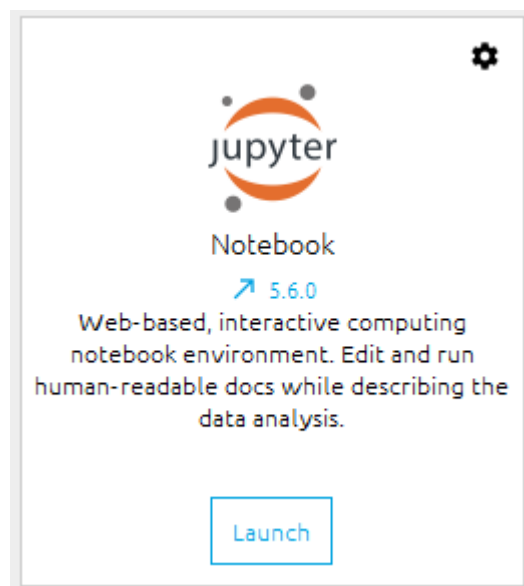


Fig: Launch the jupyter notebook platform

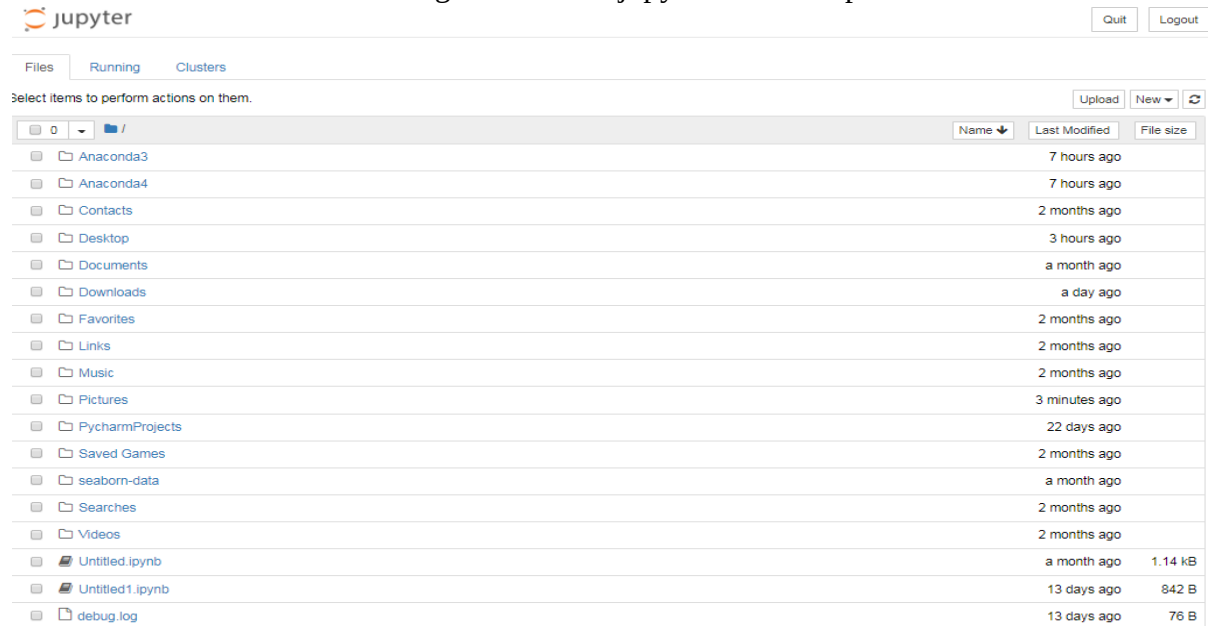


Fig: Open the correspondent result folder

In [3]: #Display the given dataset in the form of dataframe
df.head(10)

Out[3]:

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0
5	LP001011	Male	Yes	2	Graduate	Yes	5417	4196.0	267.0	360.0	1.0
6	LP001013	Male	Yes	0	Not Graduate	No	2333	1516.0	95.0	360.0	1.0
7	LP001014	Male	Yes	3+	Graduate	No	3036	2504.0	158.0	360.0	0.0
8	LP001018	Male	Yes	2	Graduate	No	4006	1526.0	168.0	360.0	1.0
9	LP001020	Male	Yes	1	Graduate	No	12841	10968.0	349.0	360.0	1.0

Fig: data frame of demo dataset

CHAPTER 8

CONCLUSION AND FUTURE WORK

8.1 CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score will be find out for future with comparing accuracy calculations. This brings some of the following insights about loan approval. Applicants with credit history not passing guidelines mostly fail to get approved, probably because that they have a higher probability of not paying back. Most of the time, applicants with high income, sanctioning low amount is more likely to get approved, which makes sense those applicants are more likely to pay back their loans. This application can help banks in predicting the future of loan and its status and depends on that they can take action in initial days of loan. Using this application banks can reduce the number of bad loans and from incurring sever losses. Using this machine learning methodology bank can easily identify the required information from huge amount of data sets and helps in successful loan prediction to reduce the number of bad loan problems.

8.2 FUTURE WORK

- Bank wants to automate the detecting the fraud persons from eligibility process (real time) based on the account detail.
- To automate this process by show the prediction result in web application or desktop application.
- To optimize the work to implement in Artificial Intelligence environment.

1.