

# **PREDECTING DEFAULT LOANS USING MACHINE LEARNING**

A major project report submitted in partial fulfilment of the requirements  
for the award of the degree of  
**BACHELOR OF TECHNOLOGY**  
in  
**COMPUTER SCIENCE AND ENGINEERING**

by  
**B. BHARGAVI** (21S41A0510)  
**D.ABHINAYA** (21S4 1A0531)  
**G. RADHIKA** (21S41A0545)  
**G. PRANEETH** (21S41A0539)

**Under the Guidance of**

**Dr. T. RAVIKUMAR**

(professor)



**Department of Computer Science & Engineering**  
**VAAGESWARI COLLEGE OF ENGINEERING**  
Accredited by NAAC with A+ Grade  
(Affiliated to JNTU Hyderabad & Approved by AICTE New Delhi)  
Ramakrishna colony, Karimnagar-505527

**Department of Computer Science & Engineering**  
**VAAGESWARI COLLEGE OF ENGINEERING**  
Accredited by NAAC with A+ Grade  
(Affiliated to JNTU Hyderabad & Approved by AICTE New Delhi)  
Ramakrishna colony, Karimnagar-505527



### **CERTIFICATE**

This is to certify that the mini project report entitled "**“PREDECETING DEFAULT LOANS USING MACHINE LEARNING”**" submitted by the following students in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology in CSE, and is a Bonafide record of the work performed by

<b>B. BHARGAVI</b>	<b>(21S41A0510)</b>
<b>D.ABHINAYA</b>	<b>(21S41A0531)</b>
<b>G. RADHIKA</b>	<b>(21S41A0545)</b>
<b>G. PRANEETH</b>	<b>(21S41A0539)</b>

**Internal Guide**

**Dr.T. RAVIKUMAR**

**Professor**

**Principal**

**Dr.CH SRINIVAS**

**Head of The Department**

**Dr. N. CHANDRAMOULI**

**Professor**

**External Examiner**

## **DECLARATION**

We hereby declare that the project titled “**PREDICTING DEFAULT LOANS USING MACHINE LEARNING**” submitted to Vaageswari College of Engineering, affiliated to Jawaharlal Nehru Technological University Hyderabad (JNTUH) for the award of the Degree of Bachelor of Technology in CSE is a result of original research carried-out in this work. It is further declared that the report or any part thereof has not been previously submitted to any University or Institute for the award of degree or diploma.

1. Name of the Student: B. BHARGAVI

Hall ticket Number:21S41AO510

2. Name of the Student: D. ABHINAYA

Hall ticket Number:21S41A0531

3. Name of the Student: G. RADHIKA

Hall ticket Number:21S41AO545

4. Name of the Student: G. PRANEETH

Hall ticket Number:21S41AO539

## **ACKNOWLEDGEMENT**

We wish to pay our sincere thanks to **Dr. CH. SRINIVAS**, Principal, Vaageswari College of Engineering, Karimnagar, for providing all required facilities and his support during the Project Work.

We would like to thank **Dr. N. CHANDRAMOULI**, Professor and HOD of the Computer Science and Engineering Department for his valuable suggestions during the Project Work.

We sincerely extend our thanks to the Mini project guide **Dr. T. RAVIKUMAR**, professor of Computer Science and Engineering Department for sparing his valuable time in guiding the Mini project work and giving feedback with many useful suggestions during the Mini project work

We are also conveying our heartfelt thanks to the Institute authority, Department Library and Laboratory staff of Vaageswari College of Engineering for their co-operation during our mini project. We thank our beloved friends for their help and encouragement regarding the concepts and experimentation

<b>B. BHARGAVI</b>	<b>(21S41A0510)</b>
<b>D.ABHINAYA</b>	<b>(21S41A0531)</b>
<b>G. RADHIKA</b>	<b>(21S41A0545)</b>
<b>G. PRANEETH</b>	<b>(21S41A0539)</b>

# **PREDICTING DEFAULT LOAN DEFAULTS**

## **USING MACHINE LEARNING**

### **ABSTRACT**

Accurately predicting bank loan defaults is a critical task for financial institutions, as it directly impacts their risk management and decision-making processes. Traditional methods for credit risk assessment often rely on statistical models and historical financial data, which may not fully capture the complexities and patterns associated with loan defaults. This paper proposes the use of deep learning techniques to enhance the prediction accuracy of bank loan defaults. By leveraging the capabilities of neural networks, particularly deep neural networks (DNNs) and recurrent neural networks (RNNs), the proposed model aims to identify intricate patterns and relationships in the data that traditional models might overlook. The model is trained on a comprehensive dataset comprising various borrower attributes, loan characteristics, and historical default information. Feature engineering techniques are employed to extract meaningful features from raw data, enhancing the model's predictive power. The proposed system integrates multiple deep learning architectures, including convolutional neural networks (CNNs) for feature extraction and long short-term memory (LSTM) networks for sequential data analysis, to capture both static and dynamic aspects of borrower behavior. Evaluation metrics such as accuracy, precision, recall, and the area under the receiver operating characteristic (ROC-AUC) curve are used to assess the model's performance. The results demonstrate that the deep learning approach outperforms traditional statistical methods, providing more accurate and reliable predictions of loan defaults.

# INDEX

<b>TOPICS</b>	<b>Page No's</b>
➤ Certificates	
➤ Acknowledgement	
➤ Abstract	
➤ Figures/Tables	
<b>CHAPTER-1: INTRODUCTION</b>	1
<b>CHAPTER-2: LITERATURE SURVEY</b>	2-3
<b>CHAPTER-3: SYSTEM ANALYSIS</b>	
3.1 Existing System	4-5
3.2 Proposed System	5
<b>CHAPTER-4: SYSTEM REQUIREMENTS</b>	
4.1 Hardware Requirement	6
4.2 Software Requirements	6
<b>CHAPTER-5: SYSTEM STUDY</b>	
5.1 Feasibility Study	7
5.2 Feasibility Analysis	7-8
<b>CHAPTER-6: SYSTEM DESIGN</b>	
6.1 SYSTEM ARCHITECTURE	9
6.2 UML DIAGRAMS	9-15
6.2.1 Use Case Diagram	
6.2.2 Class Diagram	
6.2.3 Sequence Diagram	
6.2.4 Collaboration Diagram	
6.2.5 Activity Diagram	
6.2.6 Data Flow Diagram	

**CHAPTER-7: INPUT AND OUTPUT DESIGN**

7.1 Input Design	16
7.2 Output Design	17

**CHAPTER-8: IMPLEMENTATION**

8.1 Modules	18
8.1.1 Module Description	18

**CHAPTER-9: SOFTWARE ENVIRONMENT**

9.1 Python	19-20
------------	-------

**CHAPTER-10: RESULTS/DISCUSSIONS**

10.1 System Testing	21-24
10.2 Output Screens	25-44

**CHAPTER-11: CONCLUSION**

11.1 Conclusion	45
11.2 Future Scope	45

**CHAPTER-12: REFERENCES**

## **LIST OF FIGURES**

S.NO	TABLES/FIGURES	PAGE NO'S
1	System Architecture	9
2	UML Diagrams	10-15
	2.1 Use Case Diagram	11
	2.2 Class Diagram	11
	2.3 Sequence Diagram	12
	2.4 Collaboration Diagram	13
	2.5 Activity Diagram	14
	2.6 Data FLOW DIAGRAM	15
3	Screenshots	25-44

# **CHAPTER-1**

## **INTRODUCTION**

Predicting bank loan defaults is a critical task in the financial industry, as it helps banks and other financial institutions assess the creditworthiness of potential borrowers and manage their risk exposure. Accurate prediction of loan defaults can lead to better decision-making, improved risk management, and reduced financial losses. Traditional methods for predicting loan defaults typically rely on statistical models that analyze historical financial data and borrower attributes. While these methods have been effective to some extent, they often struggle to capture complex patterns and relationships in the data. Additionally, they may not be able to adapt to changing market conditions or evolving borrower behavior. Deep learning techniques offer a promising approach to improving the accuracy of loan default prediction models. Deep learning models, such as deep neural networks (DNNs) and recurrent neural networks (RNNs), have shown great success in handling complex, high-dimensional data and identifying intricate patterns that are difficult for traditional models to detect. This paper explores the use of deep learning techniques for predicting bank loan defaults. The proposed model leverages the capabilities of deep neural networks to analyze a comprehensive dataset containing various borrower attributes, loan characteristics, and historical default information. By extracting meaningful features from the data and utilizing advanced neural network architectures, the model aims to achieve higher prediction accuracy compared to traditional methods. In recent years, the financial industry has witnessed rapid advancements in predictive modeling, driven largely by the growth of machine learning and deep learning techniques. One of the most critical applications of these technologies lies in predicting credit default, a challenge that holds immense significance for banks and financial institutions. Accurate predictions of loan defaults can help mitigate risk, enhance decision-making, and ensure more efficient credit allocation. This paper delves into the use of deep learning techniques for predicting bank loan defaults, exploring how advanced models such as neural networks and their variants outperform traditional methods. By leveraging large volumes of data and extracting complex patterns that may not be evident in simpler statistical models, deep learning techniques have the potential to revolutionize the credit risk assessment process. Consequently, understanding and implementing these methods can significantly contribute to financial stability and reduce the likelihood of bad debt accumulation.

## CHAPTER-2

### LITERATURE SURVEY

**TITLE:** A Comprehensive Survey on Deep Learning Techniques for Predicting Credit Default in Banking Systems

**ABSTRACT:** In recent years, predicting credit default has become a critical task for financial institutions aiming to reduce financial risks. The advent of deep learning has significantly advanced predictive modeling by leveraging large volumes of structured and unstructured data. This literature survey explores state-of-the-art deep learning models, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and autoencoders, in the context of credit default prediction. Various studies have demonstrated the superiority of these models over traditional machine learning algorithms, primarily due to their ability to capture complex, non-linear relationships in the data. This paper reviews multiple deep learning architectures applied to financial datasets and evaluates their performance metrics, including accuracy, precision, and recall. Additionally, this survey discusses the challenges and future directions in integrating these models into real-world banking systems, with an emphasis on interpretability, data privacy, and regulatory compliance.

**TITLE:** Deep Learning Models for Bank Loan Default Prediction: A Survey of Techniques and Applications

**ABSTRACT:** Predicting bank loan defaults is a critical concern for financial institutions seeking to manage credit risk effectively. This literature survey reviews the application of various deep learning models, including Long Short-Term Memory (LSTM) networks, deep neural networks (DNNs), and Generative Adversarial Networks (GANs), in predicting bank loan defaults. Numerous studies show that deep learning techniques outperform traditional statistical models, such as logistic regression and decision trees, due to their ability to process high-dimensional data and automatically extract features. This survey also examines the importance of feature engineering, hyperparameter tuning, and the choice of loss functions in improving model performance. Furthermore, it highlights the adoption of hybrid models that combine deep learning with ensemble techniques for improved robustness in prediction accuracy. The paper concludes by outlining the limitations of deep learning in this domain, particularly in terms of model transparency and the potential for overfitting, while suggesting future research directions to address these issues.

**TITLE:** Analyzing Deep Learning Approaches in Predicting Credit Defaults: Techniques, Challenges, and Future Directions

**ABSTRACT:** Deep learning has revolutionized the field of credit risk assessment by offering powerful predictive tools for identifying potential loan defaults. This literature survey provides an in-depth analysis of recent research on deep learning approaches for predicting credit default, focusing on models such as feedforward neural networks (FNNs), deep belief networks (DBNs), and graph neural networks (GNNs). The survey identifies key areas where deep learning models have demonstrated significant predictive power compared to traditional methods like support vector machines (SVMs) and random forests. Moreover, this paper discusses the various data preprocessing techniques, such as imbalanced data handling, normalization, and data augmentation, which play crucial roles in enhancing the performance of deep learning models. The survey also explores the ethical considerations and challenges of implementing these models in practice, including bias mitigation, explainability, and compliance with regulatory standards in the financial industry.

**TITLE:** Advances in Deep Learning for Predicting Bank Loan Defaults: A Survey of Techniques, Datasets, and Model Interpretability

**ABSTRACT:** The rise of deep learning has paved the way for sophisticated models in predicting bank loan defaults, which is critical for mitigating financial risk. This literature survey reviews the latest advancements in deep learning techniques, such as Transformer-based models, autoencoders, and attention mechanisms, that have been applied to bank loan default prediction. By analyzing a range of studies, this survey highlights the advantages of deep learning over traditional statistical methods, especially in capturing intricate patterns in time-series and transaction data. The review also focuses on the types of datasets commonly used, including public datasets like the LendingClub and proprietary bank datasets, and the challenges associated with data quality and availability. Additionally, this paper emphasizes the growing importance of model interpretability in deep learning, examining techniques such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) to address the "black-box" nature of these models, ensuring that predictions can be explained and trusted by stakeholders.

# **CHAPTER-3**

## **SYSTEM ANALYSIS**

### **3.1 EXISTING SYSTEM**

The existing systems for predicting bank loan defaults primarily rely on traditional statistical models and machine learning algorithms. These systems analyze historical financial data and borrower attributes to assess the creditworthiness of potential borrowers and predict the likelihood of loan defaults. While these methods have been effective to some extent, they often struggle to capture complex patterns and relationships in the data, leading to limitations in prediction accuracy.

One common approach in the existing systems is to use logistic regression models to predict loan defaults based on a set of predefined features such as credit score, income, loan amount, and loan-to-value ratio. These models provide a binary classification of whether a borrower is likely to default on a loan or not. While logistic regression is a simple and interpretable model, it may not capture the nonlinear relationships and interactions between variables, limiting its predictive power.

Another approach is to use decision tree-based models such as Random Forests or Gradient Boosting Machines (GBM). These models can capture nonlinear relationships and interactions in the data, leading to improved prediction accuracy compared to logistic regression. However, they may still struggle with capturing complex patterns in the data, especially when dealing with high-dimensional and heterogeneous data such as that found in loan default prediction.

#### **3.1.1 DISADVANTAGES**

Predicting credit defaults using deep learning techniques can be a promising approach due to the ability of deep learning models to automatically learn complex patterns from data. However, there are several drawbacks in existing systems for predicting bank loan defaults that deep learning can potentially address:

1. Limited Feature Representation: Traditional models often rely on a limited set of hand-crafted features, which may not capture the full complexity of the data. Deep learning can automatically learn relevant features from raw data, potentially leading to more robust and accurate predictions.

2. Non-linear Relationships: Deep learning models can capture non-linear relationships between features and the target variable, which may be missed by linear models used in traditional systems.

## 3.2 PROPOSED SYSTEM

In the proposed system for predicting credit default or bank loan defaults using deep learning techniques, several key components are considered. Firstly, a robust dataset containing historical loan information, borrower characteristics, economic indicators, and repayment status is utilized. This dataset is preprocessed to handle missing values, normalize features, and potentially engineer new features that may enhance model performance. Next, various deep learning models such as deep neural networks (DNNs), convolutional neural networks (CNNs), or recurrent neural networks (RNNs) are employed to learn complex patterns and relationships within the data. These models can effectively capture nonlinearities and interactions among features, which are crucial for accurate predictions. To improve the model's performance and generalization ability, techniques such as regularization, dropout, batch normalization, and early stopping are employed. These techniques help prevent overfitting and improve the model's ability to generalize to unseen data. Furthermore, the model's performance is evaluated using metrics such as accuracy, precision, recall, and F1-score. Additionally, techniques such as cross-validation and hyperparameter tuning are employed to optimize the model's performance.

### 3.2.1 ADVANTAGES

The proposed system for predicting credit default or bank loan defaults using deep learning techniques offers several advantages:

**High Accuracy:** Deep learning models can capture intricate patterns and relationships in the data, leading to more accurate predictions compared to traditional methods.

**Feature Learning:** Deep learning models can automatically learn relevant features from the data, reducing the need for manual feature engineering and potentially improving model performance.

**Nonlinearity Handling:** Deep learning models are well-suited for capturing nonlinear relationships in the data, which are often present in financial datasets.

## **CHAPTER-4**

### **SYSTEM REQUIREMENTS**

#### **4.1 H/W SYSTEM CONFIGURATION:**

- System : i3 or above.
- Ram : 4 GB.
- Hard Disk : 40 GB

#### **4.2 SOFTWARE REQUIREMENTS:**

- Operating System : Windows8 or Above.
- Coding Language : python

## **CHAPTER-5**

### **SYSTEM STUDY**

#### **5.1 FEASIBILITY STUDY**

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

#### **5.2 FEASIBILITY ANALYSIS**

Three key considerations involved in the feasibility analysis are

**ECONOMICAL FEASIBILITY**

**TECHNICAL FEASIBILITY**

**SOCIAL FEASIBILITY**

#### **ECONOMICAL FEASIBILITY**

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

#### **TECHNICAL FEASIBILITY**

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead

to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

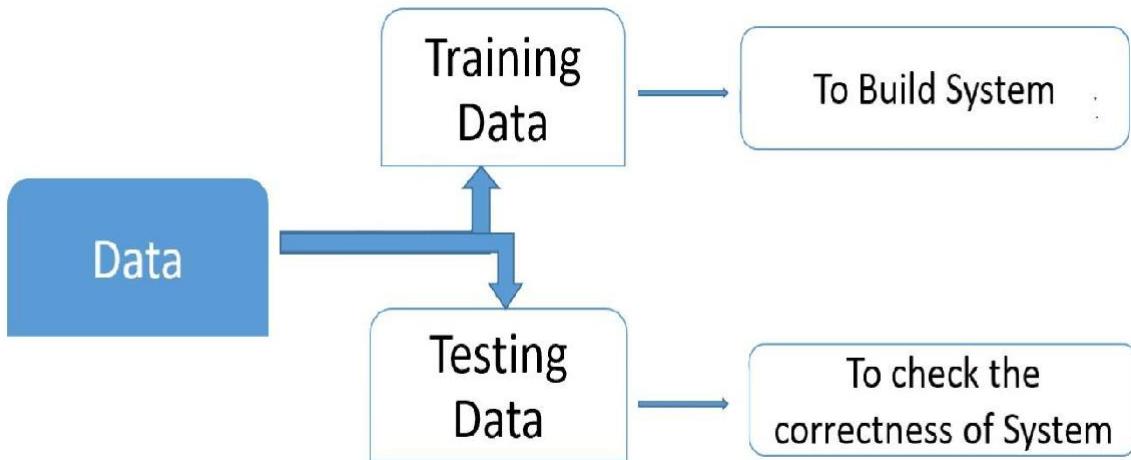
## **SOCIAL FEASIBILITY**

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

## CHAPTER-6

### SYSTEM DESIGN

#### 6.1 SYSTEM ARCHITECTURE



---

#### 6.2 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

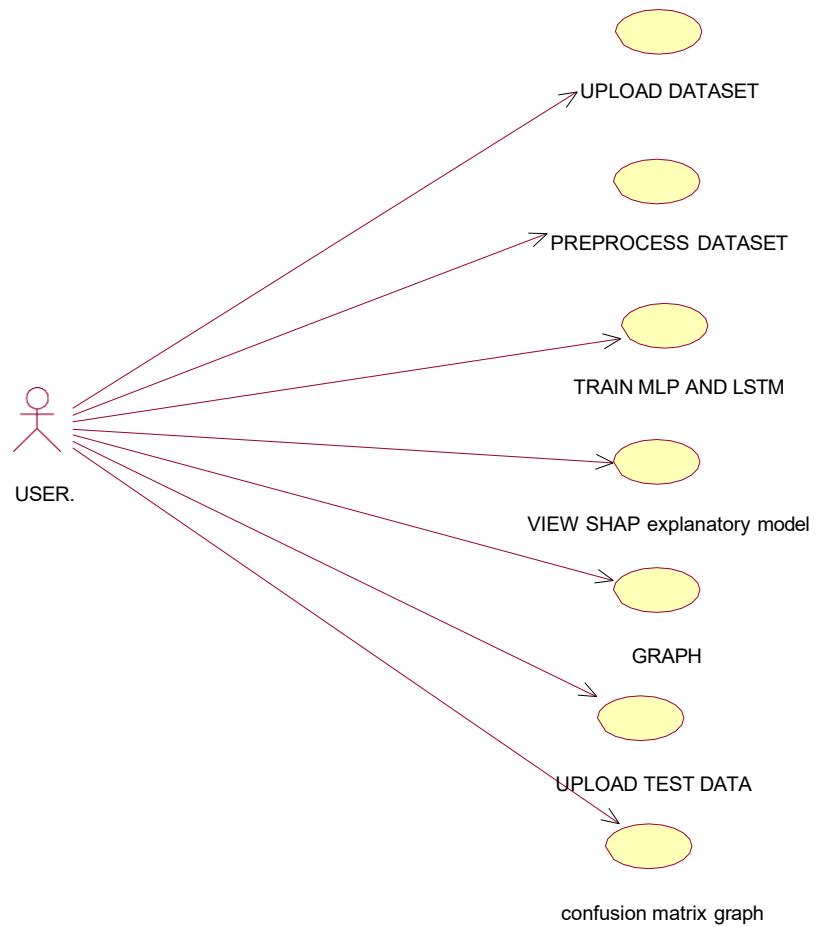
## **GOALS:**

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

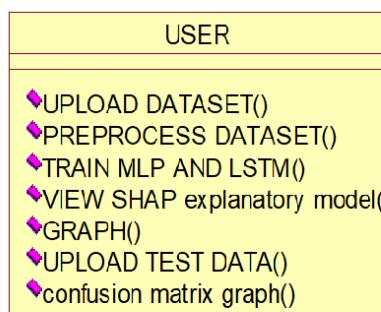
### **6.2.1 USE CASE DIAGRAM**

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.



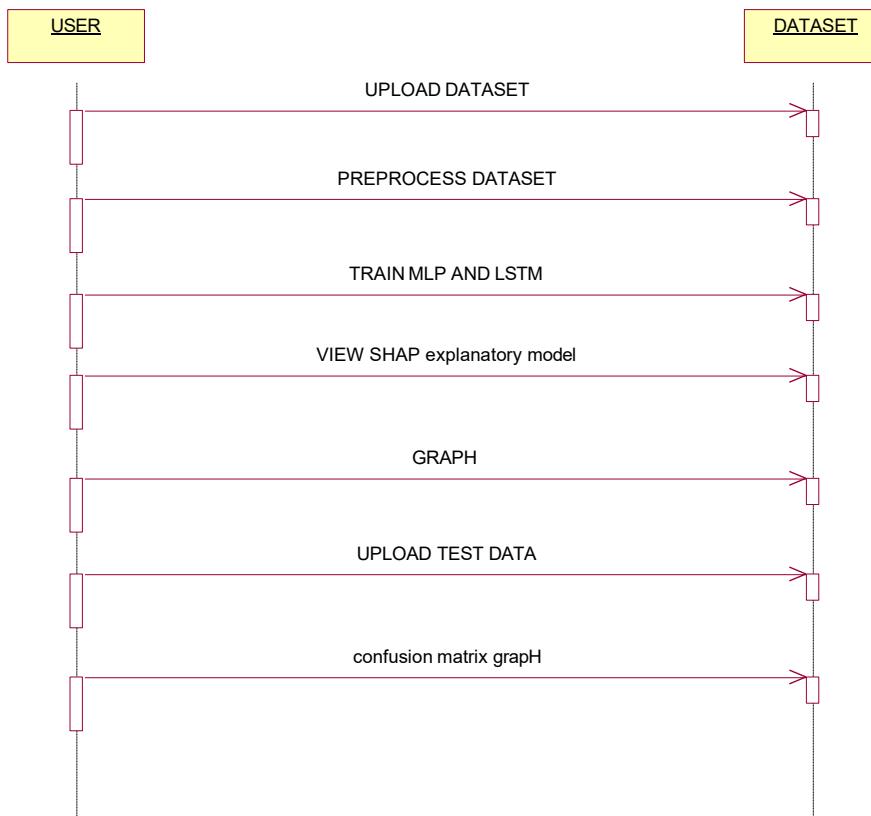
## 6.2.2 CLASS DIAGRAM

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.



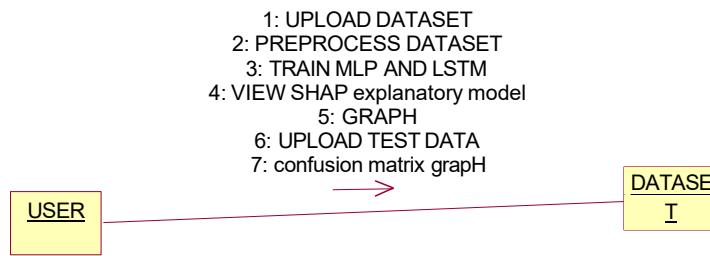
### 6.2.3 SEQUENCE DIAGRAM

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.



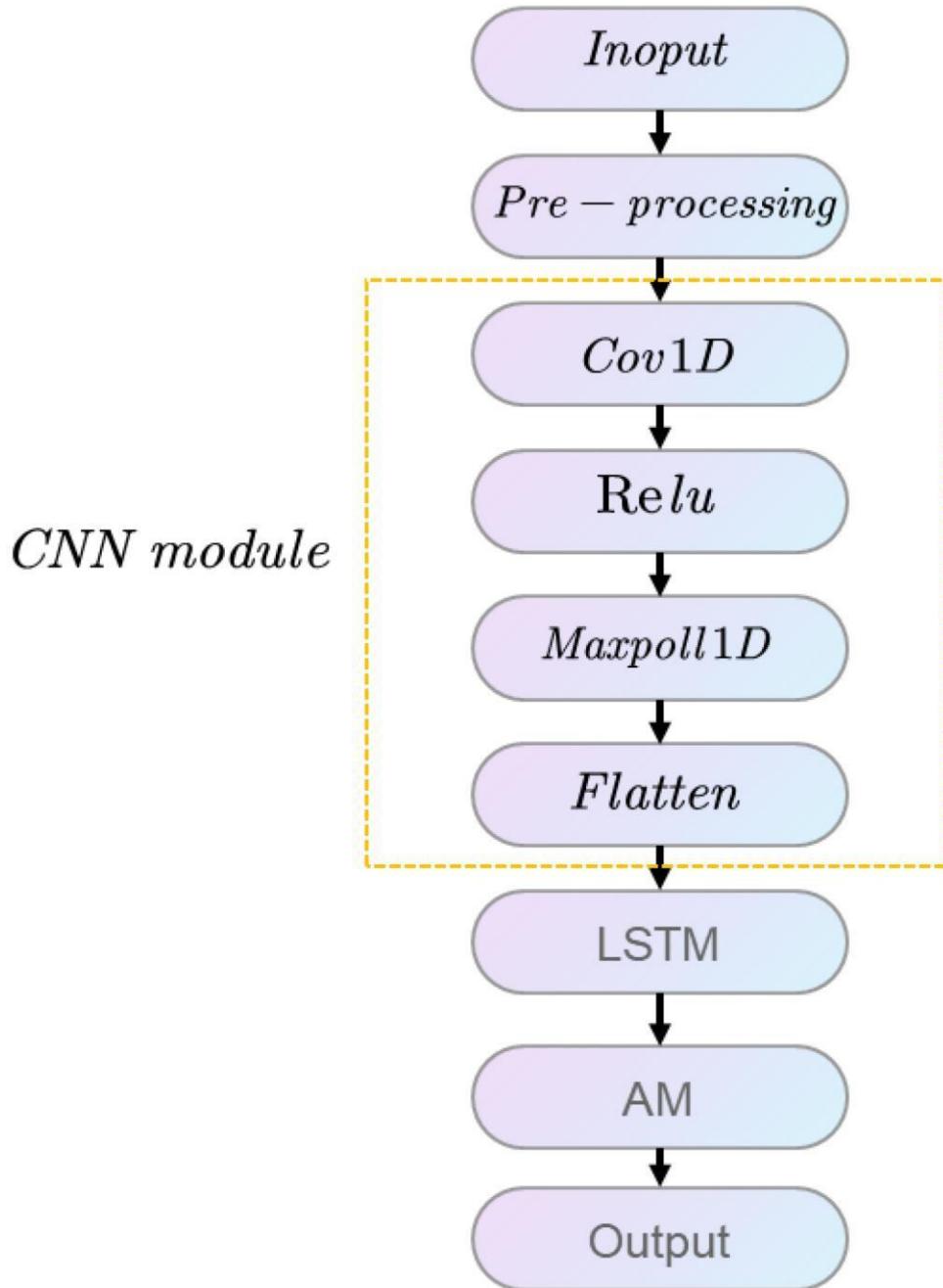
## 6.2.4 COLLABRATION DIAGRAM

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

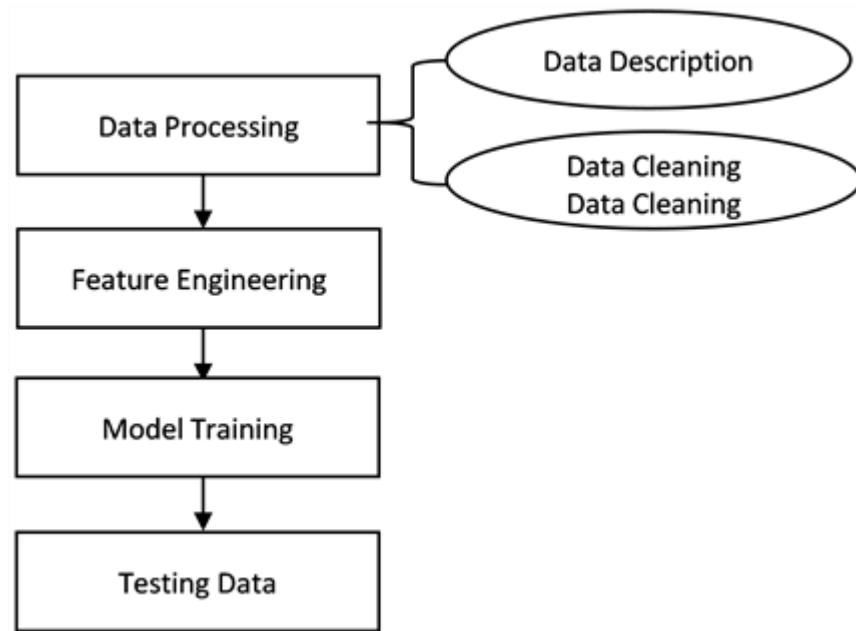


## 6.2.5 ACTIVITY DIAGRAM

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



### 6.2.6 DATA FLOW DIAGRAM



## **CHAPTER-7**

### **INPUT AND OUTPUT DESIGN**

#### **7.1 INPUT DESIGN**

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

##### **7.1.1 OBJECTIVES**

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in

maize of instant. Thus the objective of input design is to create an input layout that is easy to follow.

## 7.2 OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.
2. Select methods for presenting information.
3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

# **CHAPTER-8**

## **IMPLEMENTATION**

### **8.1 MODULES**

- USER

#### **8.1.1 MODULE DESCRIPTION**

**User Interface:** Develop a user-friendly interface for interacting with the system, allowing users to input text and view the generated summaries along with sentiment analysis results. Design the interface to be intuitive, responsive, and accessible across different devices and platforms.

**Data Preprocessing:** Prepare the textual data by removing noise, such as special characters, punctuation, and stop words. Tokenize the text into sentences or paragraphs to facilitate sentiment analysis and summarization.

**Deployment:** Deploy the implemented system in production environments, considering factors such as scalability, reliability, and security. Ensure proper monitoring and maintenance procedures are in place to address potential issues and ensure continuous performance optimization.

**Feedback Loop:** Establish a feedback loop to gather user feedback and monitor system performance over time. Use feedback to iteratively improve the system's accuracy, usability, and effectiveness based on user requirements and evolving needs.

# CHAPTER-9

## SOFTWARE ENVIRONMENT

### 9.1 PYTHON

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

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

### Python Features

Python's features include:

- **Easy-to-learn:** Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.
- **Easy-to-read:** Python code is more clearly defined and visible to the eyes.
- **Easy-to-maintain:** Python's source code is fairly easy-to-maintain.
- **A broad standard library:** Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.

- **Interactive Mode:** Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.
- **Portable:** Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
- **Extendable:** You can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.
- **Databases:** Python provides interfaces to all major commercial databases.
- **GUI Programming:** Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
- **Scalable:** Python provides a better structure and support for large programs than shell scripting.

Python has a big list of good features:

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

# **CHAPTER-10**

## **RESULTS/DISCUSSION**

### **10.1 SYSTEM TESTING**

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

### **TYPES OF TESTS**

#### **UNIT TESTING**

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

#### **INTEGRATION TESTING**

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

## **FUNCTIONAL TEST**

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

### **Functional testing is centered on the following items:**

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

## **SYSTEM TEST**

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration-oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

## **WHITE BOX TESTING**

White Box Testing is a testing in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is used to test areas that cannot be reached from a black box level.

## **BLACK BOX TESTING**

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

## **UNIT TESTING**

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

## **TEST STRATEGY AND APPROACH**

Field testing will be performed manually and functional tests will be written in detail.

### **Test objectives**

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

### **Features to be tested**

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

## **Integration Testing**

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects. The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

**Test Results:** All the test cases mentioned above passed successfully. No defects encountered.

## **ACCEPTANCE TESTING**

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

**Test Results:** All the test cases mentioned above passed successfully. No defects encountered.

## **10.2 SCREENSHOTS**

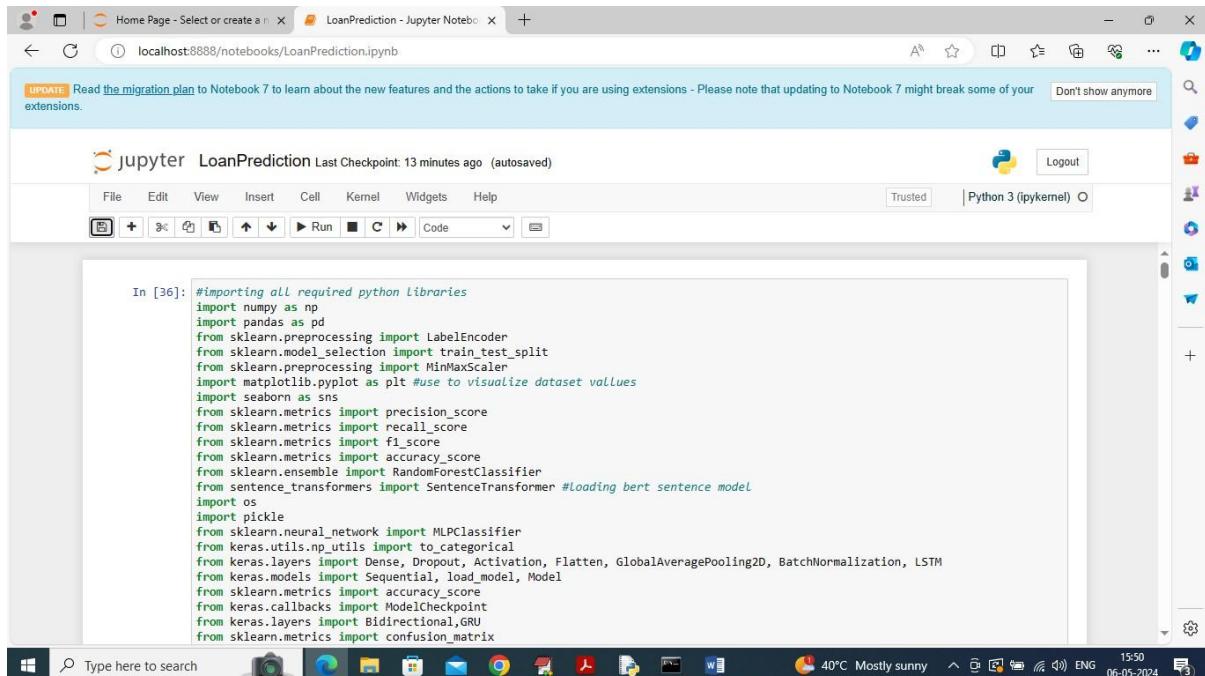
Deep Learning Techniques Based predicting credit default/Predicting Bank Loan Defaults  
In propose work utilizing text based dataset to predict bank loan defaulters and non-defaulters. All bank transaction contains some type of text features and this text features will be converted to vector using scratch technique without using any pre-trained model. Experimenting with pre-trained and scratch generated deep text model performing better compare to pre-trained models.

In this work we are converting all text features into numeric vector model without using any pre-trained model and then comparing with vector model obtained from pre-trained BERT model.

Both scratch and pre-trained vectors get trained with MLP and LSTM algorithms and then tested its performance using various metrics such as Accuracy, precision, Recall, Confusion Matrix and FSCORE.

Both models giving best accuracy using features generated from scratch and then employ SHAP explanatory model to identify features which are helping models in getting better accuracy.

To train all algorithms we have utilized same dataset given by you and then we have coded this project using JUPYTER notebook. Below are the code and output screens with blue colour comments.



The screenshot shows a Jupyter Notebook interface running in a browser window. The title bar says "LoanPrediction - Jupyter Notebook". The main area displays the following Python code:

```
In [36]: #importing all required python libraries
import numpy as np
import pandas as pd
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
import matplotlib.pyplot as plt #use to visualize dataset values
import seaborn as sns
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
from sklearn.metrics import f1_score
from sklearn.metrics import accuracy_score
from sklearn.ensemble import RandomForestClassifier
from sentence_transformers import SentenceTransformer #loading bert sentence model
import os
import pickle
from sklearn.neural_network import MLPClassifier
from keras.utils.np_utils import to_categorical
from keras.layers import Dense, Dropout, Activation, Flatten, GlobalAveragePooling2D, BatchNormalization, LSTM
from keras.models import Sequential, load_model, Model
from sklearn.metrics import accuracy_score
from keras.callbacks import ModelCheckpoint
from keras.layers import Bidirectional, GRU
from sklearn.metrics import confusion_matrix
```

In above screen importing required python classes and packages

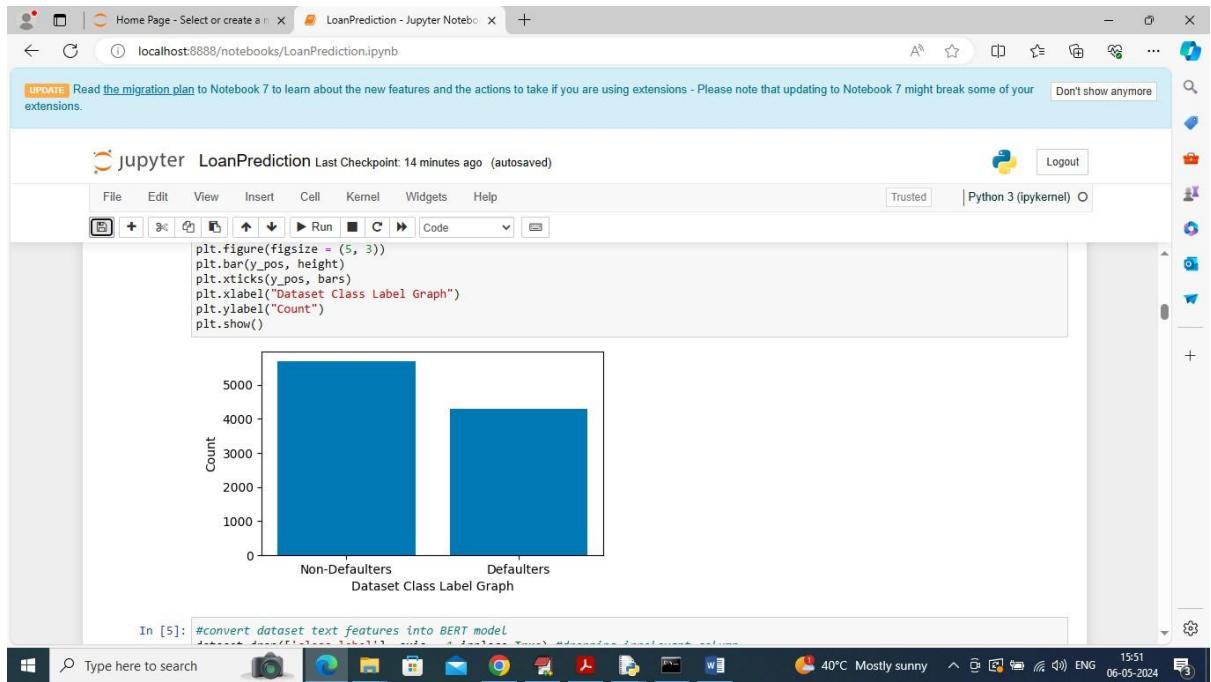
The screenshot shows a Jupyter Notebook running on a Windows desktop. The browser tab is titled 'LoanPrediction - Jupyter Notebook'. The notebook interface includes a toolbar with File, Edit, View, Insert, Cell, Kernel, Widgets, Help, and a Python 3 (ipykernel) kernel selector. A message at the top indicates an update to Notebook 7. The code cell In [2] contains the following Python code:

```
#Loading and displaying Loans dataset
dataset = pd.read_csv("Dataset/loans_full_schema.csv")
dataset
```

The output cell Out[2] displays the first 10 rows of the dataset as a pandas DataFrame:

	emp_title	emp_length	state	homeownership	annual_income	verified_income	debt_to_income	annual_income_joint	verification_income_joint	debt_to_in
0	global config engineer	3.0	NJ	MORTGAGE	90000.0	Verified	18.01	NaN	NaN	NaN
1	warehouse office clerk	10.0	HI	RENT	40000.0	Not Verified	5.04	NaN	NaN	NaN
2	assembly	3.0	WI	RENT	40000.0	Source Verified	21.15	NaN	NaN	NaN
3	customer service	1.0	PA	RENT	30000.0	Not Verified	10.16	NaN	NaN	NaN
4	security supervisor	10.0	CA	RENT	35000.0	Verified	57.96	57000.0	Verified	NaN
...	...	...	...	...	...	...	...	...	...	...
9995	owner	10.0	TX	RENT	108000.0	Source Verified	22.28	NaN	NaN	NaN
9996	director	8.0	PA	MORTGAGE	121000.0	Verified	32.38	NaN	NaN	NaN
9997	toolmaker	10.0	CT	MORTGAGE	67000.0	Verified	45.26	107000.0	Source Verified	NaN
9998	manager	1.0	WI	MORTGAGE	80000.0	Source Verified	11.99	NaN	NaN	NaN

In above screen loading and displaying dataset values and can see dataset contains both text and numeric features



In above screen displaying graph of defaulter and non-defaulter class labels found in dataset where x-axis represents class labels and y-axis represents count of those class labels.

The screenshot shows a Jupyter Notebook interface running on a Windows desktop. The notebook title is "LoanPrediction - Jupyter Notebook". A message at the top says "UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions." Below this, the code cell In [5] contains Python code for converting dataset text features into BERT model embeddings:

```
#convert dataset text features into BERT model
dataset.drop(['class_label'], axis = 1,inplace=True) #dropping irrelevant column
data = dataset.values
if os.path.exists("model/text.npy"):
    textData = np.load("model/text.npy")
    Y = np.load("model/Y.npy")
else:
    for i in range(len(dataset)):
        data = ""
        for j in range(len(dataset[i])):#Loop and read all text data from dataset
            if dataset[i,j] is not None:
                value = str(dataset[i, j]).strip()
                if len(value) > 0 and value != 'nan':
                    data += value+" "
        textData.append(data.strip().lower())
textData = np.asarray(textData)
np.save("model/text", textData)
np.save("model/Y", Y)
if os.path.exists("model/bert.npy"):
    bert_X = np.load("model/bert.npy")
else:
    embeddings = bert.encode(textData, convert_to_tensor=True)#convert text data into bert vector
    bert_X = embeddings.numpy()
    np.save("model/bert", bert_X)
print("BERT Vector from text Data")
```

In above screen using pre-trained BERT model converting all text transaction features into numeric vector and after executing this block will get below output

The screenshot shows a Jupyter Notebook interface running on a Windows desktop. The notebook title is "LoanPrediction - Jupyter Notebook". A message at the top says "UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions." Below the message, the notebook header includes "jupyter LoanPrediction Last Checkpoint: 15 minutes ago (autosaved)" and "Python 3 (ipykernel)". The main area displays Python code for generating BERT vectors from text data:

```
if os.path.exists("model/bert.npy"):
    bert_X = np.load("model/bert.npy")
else:
    embeddings = bert.encode(textData, convert_to_tensor=True)#convert text data into bert vector
    bert_X = embeddings.numpy()
    np.save("model/bert", bert_X)
print("BERT Vector from text Data")
print(bert_X)
```

The output shows the generated BERT vectors:

```
BERT Vector from text Data
[[ 0.27395287  0.23365359 -0.12183109 ... -0.64672446 -0.31620488
-0.32734865]
 [ 0.3153612   0.2655095   0.22826457 ... -0.53634703 -0.24058835
-0.33399954]
 [ 0.18211041  0.11876082 -0.39152658 ... -0.06244954 -0.04253473
-0.24840376]
 ...
 [ 0.37628374  0.13677596 -0.15356751 ... -0.27453268 -0.15418768
-0.3865813 ]
 [ 0.3509853   0.24931842  0.10075459 ... -0.45178473 -0.42381537
0.06397581]
 [ 0.20309126  0.1081479  -0.00264408 ... -0.44929758 -0.07972104
-0.20787609]]
```

In [6]: #now create deep text vector by processing dataset from scratch  
dataset.fillna(0, inplace = True)  
label encoder = []

The taskbar at the bottom shows various application icons, and the system tray indicates it's 40°C, Mostly sunny, with the date 06-05-2024.

In above screen displaying some BERT vector features generated from TEXT data

In [6]:

```
#now create deep text vector by processing dataset from scratch
dataset.fillna(0, inplace = True)
label_encoder = []
columns = dataset.columns
types = dataset.dtypes.values
for i in range(len(types)):
    name = types[i]
    if name == 'object': #converting all text column into numeric vector using scratch technique
        le = LabelEncoder()
        dataset[columns[i]] = pd.Series(le.fit_transform(dataset[columns[i]].astype(str)))#encode all str columns to numeric
        label_encoder.append([columns[i], le])
#handling missing values
dataset.fillna(0, inplace = True)
X = dataset.values
sc = MinMaxScaler((0,1))
X = sc.fit_transform(X)
print("Deep Text Vector from SCRATCH")
print(X)
```

Deep Text Vector from SCRATCH

[0.37502637 0.3	0.6122449	...	0.0246035	0.24076946 0.	]
[0.98628981 1.	0.2244898	...	0.00871575	0.03569125 0.	]
[0.0438726 0.3	0.95918367	...	0.00438425	0.02524167 0.	]
...					
[0.94600295 1.	0.12244898	...	0.05955875	0.53088625 0.	]
[0.5085425 0.1	0.95918367	...	0.062047	0.09283424 0.	]
[0.50842402 0.2	0.12244898	...	0.02062025	0.10278775 0.	]

In above screen converting all text features from dataset into numeric vector using scratch technique without using any pre-trained model and then displaying generated vector values

The screenshot shows a Jupyter Notebook interface running on a Windows desktop. The notebook title is "LoanPrediction". The code in cell [7] splits the dataset into training and testing sets for both BERT and scratch features. The output shows the shapes of the resulting arrays: 80% BERT vectors for training and 20% for testing, and similarly for the scratch vectors.

```
[0.37502637 0.3 0.6122449 ... 0.0246035 0.24076946 0.]
[0.98628981 1. 0.2244898 ... 0.00871575 0.03569125 0.]
[0.0438726 0.3 0.95918367 ... 0.00438425 0.02524167 0.]
...
[0.94600295 1. 0.12244898 ... 0.05955875 0.53088625 0.]
[0.5085425 0.1 0.95918367 ... 0.0026047 0.09283424 0.]
[0.58426492 0.3 0.12244898 ... 0.03062925 0.10278576 0.]]
In [7]: #split both deep text scratch and BERT data into train and test
bert_X_train, bert_X_test, bert_y_train, bert_y_test = train_test_split(bert_X, Y, test_size=0.2) #split BERT data into train and test
scratch_X_train, scratch_X_test, scratch_y_train, scratch_y_test = train_test_split(X, Y, test_size=0.2) #split SCRATCH data into train and test
print("80% BERT vector used to train algorithms : "+str(bert_X_train.shape[0]))
print("20% BERT vector used to test algorithms : "+str(bert_X_test.shape[0]))
print("80% SCRATCH vector used to train algorithms : "+str(scratch_X_train.shape[0]))
print("20% SCRATCH vector used to test algorithms : "+str(scratch_X_test.shape[0]))
4
80% BERT vector used to train algorithms : 8000
20% BERT vector used to test algorithms : 2000
80% SCRATCH vector used to train algorithms : 8000
20% SCRATCH vector used to test algorithms : 2000
In [14]: #define global variables to save accuracy and other metrics
accuracy = []
precision = []
recall = []
```

In above screen splitting and displaying both BERT and scratch features into train and test where application using 80% dataset for training and 20% for testing

The screenshot shows a Jupyter Notebook interface running on a Windows desktop. The browser tab is titled 'LoanPrediction - Jupyter Notebook'. The notebook has two cells visible:

```
In [14]: #define global variables to save accuracy and other metrics
accuracy = []
precision = []
recall = []
fscore = []

In [15]: #function to calculate all metrics
def calculateMetrics(algorithm, testY, predict):
    p = precision_score(testY, predict, average='macro') * 100
    r = recall_score(testY, predict, average='macro') * 100
    f = f1_score(testY, predict, average='macro') * 100
    a = accuracy_score(testY, predict)*100
    accuracy.append(a)
    precision.append(p)
    recall.append(r)
    fscore.append(f)
    print(algorithm+" Accuracy : "+str(a))
    print(algorithm+" Precision : "+str(p))
    print(algorithm+" Recall : "+str(r))
    print(algorithm+" FSCORE : "+str(f))
    conf_matrix = confusion_matrix(testY, predict)
    fig, axs = plt.subplots(1,2,figsize=(10, 3))
    ax = sns.heatmap(conf_matrix, xticklabels = labels, yticklabels = labels, annot = True, cmap="viridis" ,fmt = "g", ax=axs[0]);
    ax.set_xlim(0, len(labels));

```

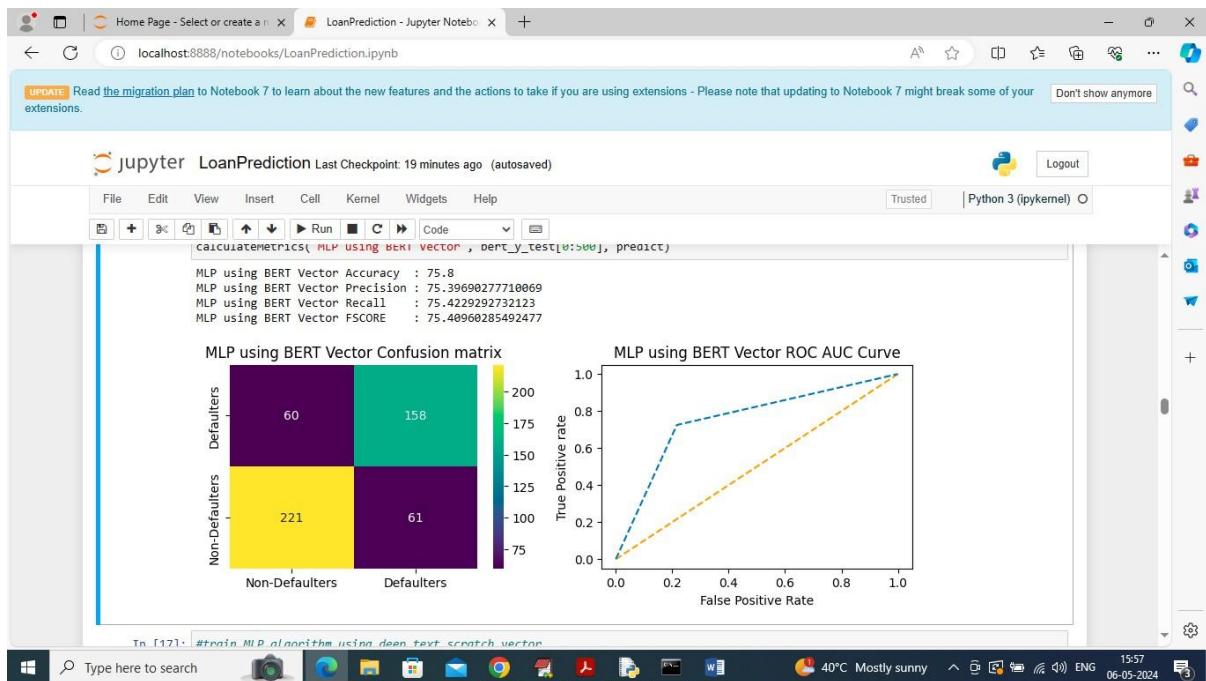
In above screen defining function to calculate accuracy and other metrics

The screenshot shows a Jupyter Notebook interface running on a Windows desktop. The notebook title is "LoanPrediction - Jupyter Notebook". The code cell In [16] contains Python code for training an MLP classifier on BERT features and calculating metrics. The output cell shows the accuracy, precision, recall, and F1-score.

```
#train MLP algorithm on BERT vector
bert_mlp = MLPClassifier(max_iter=500)
bert_mlp.fit(bert_X_train[0:3000], bert_y_train[0:3000])
#perform prediction on test data
predict = bert_mlp.predict(bert_X_test[0:500])
#call this to calculate accuracy and other metrics
calculateMetrics("MLP using BERT Vector", bert_y_test[0:500], predict)
```

MLP using BERT Vector Accuracy : 75.8  
MLP using BERT Vector Precision : 75.39600277710069  
MLP using BERT Vector Recall : 75.422020272122  
MLP using BERT Vector FSCORE : 75.40960285492477

In above screen training MLP algorithm using BERT features and then in blue colour text can see BERT MLP got 75% accuracy and can see other metrics like precision, recall and FSCORE.  
In below screen can see confusion and ROC graph



in above confusion matrix graph x-axis represents Predicted Labels and y-axis represents True labels and then different color boxes like green and yellow represents correct prediction count and all blue boxes represents incorrect prediction count which are very few. In ROC graph x-axis represents False Positive Rate and y-axis represents True Positive Rate and if blue lines comes below orange line then all predictions are incorrect or false and if goes above orange line then all predictions are correct or true.

The screenshot shows a Jupyter Notebook interface on a Windows desktop. The notebook title is "LoanPrediction". The code cell contains Python code for training an MLP classifier and calculating metrics. The output cell shows the following results:

```

scratch_mlp = MLPClassifier(max_iter=500)
scratch_mlp.fit(scratch_X_train[0:3000], scratch_y_train[0:3000])
#perform prediction on test data
predict = scratch_mlp.predict(scratch_X_test[0:500])
#call this to calculate accuracy and other metrics
calculateMetrics("MLP using Deep Text Scratch Vector", scratch_y_test[0:500], predict)

c:\users\admin\appdata\local\programs\python\python37\lib\site-packages\sklearn\neural_network\_multilayer_perceptron.py:571: ConvergenceWarning: Stochastic Optimizer: Maximum iterations (500) reached and the optimization hasn't converged yet.
  % self.max_iter, ConvergenceWarning)

```

MLP using Deep Text Scratch Vector Accuracy : 95.6  
MLP using Deep Text Scratch Vector Precision : 95.5151515151515  
MLP using Deep Text Scratch Vector Recall : 95.59178255168284  
MLP using Deep Text Scratch Vector FSCORE : 95.55188923395445

MLP using Deep Text Scratch Vector Confusion matrix    MLP using Deep Text Scratch Vector ROC AUC Curve

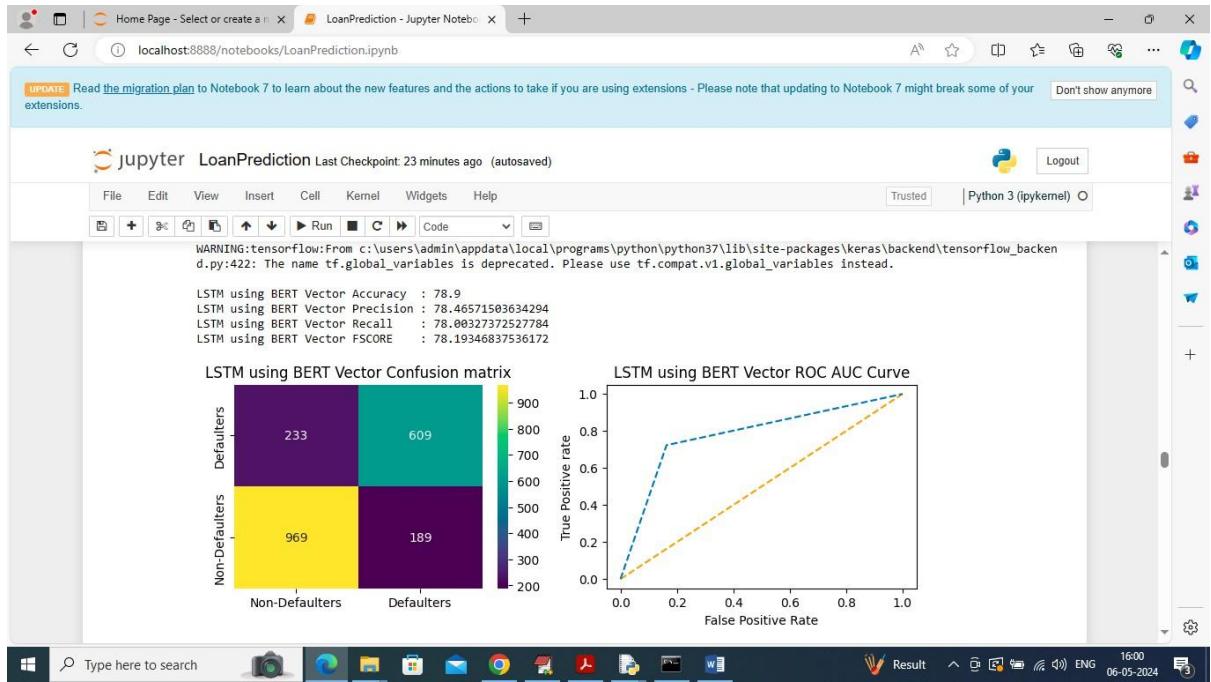
	Defaulters	Users
Defaulters	10	213
Users	10	213

In above screen training MLP on scratch features and then it got 95% accuracy and can see other output and graphs also

The screenshot shows a Jupyter Notebook interface running on localhost:8888/notebooks/LoanPrediction.ipynb. The notebook title is "jupyter LoanPrediction Last Checkpoint: 22 minutes ago (autosaved)". The code in cell [19] is as follows:

```
In [19]: bert_y_train = to_categorical(bert_y_train)
bert_y_test = to_categorical(bert_y_test)
bert_X_train = np.reshape(bert_X_train, (bert_X_train.shape[0], 48, 16))
bert_X_test = np.reshape(bert_X_test, (bert_X_test.shape[0], 48, 16))
#train LSTM algorithm using BERT vector features
lstm_bert = Sequential()#defining deep learning sequential object
#adding LSTM Layer with 100 filters to filter given input X train data to select relevant features
lstm_bert.add(LSTM(100, input_shape=(bert_X_train.shape[1], bert_X_train.shape[2])))
#adding dropout Layer to remove irrelevant features
lstm_bert.add(Dropout(0.5))
#adding another Layer
lstm_bert.add(Dense(100, activation='relu'))
#defining output layer for prediction
lstm_bert.add(Dense(bert_y_train.shape[1], activation='softmax'))
#compile LSTM model
lstm_bert.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
if os.path.exists("model/lstm_bert.hdf5") == False:
    model_checkpoint = ModelCheckpoint(filepath='model/lstm_bert.hdf5', verbose = 1, save_best_only = True)
    hist = lstm_bert.fit(bert_X_train, bert_y_train, batch_size = 32, epochs = 15, validation_data=(bert_X_test, bert_y_test), callbacks=[model_checkpoint])
    f = open('model/lstm_bert.pkl', 'wb')
    pickle.dump(hist.history, f)
    f.close()
else:
    lstm_bert.load_weights("model/lstm_bert.hdf5")
#perform prediction on test data
```

In above screen training LSTM algorithm on BERT features and after executing above block will get below output

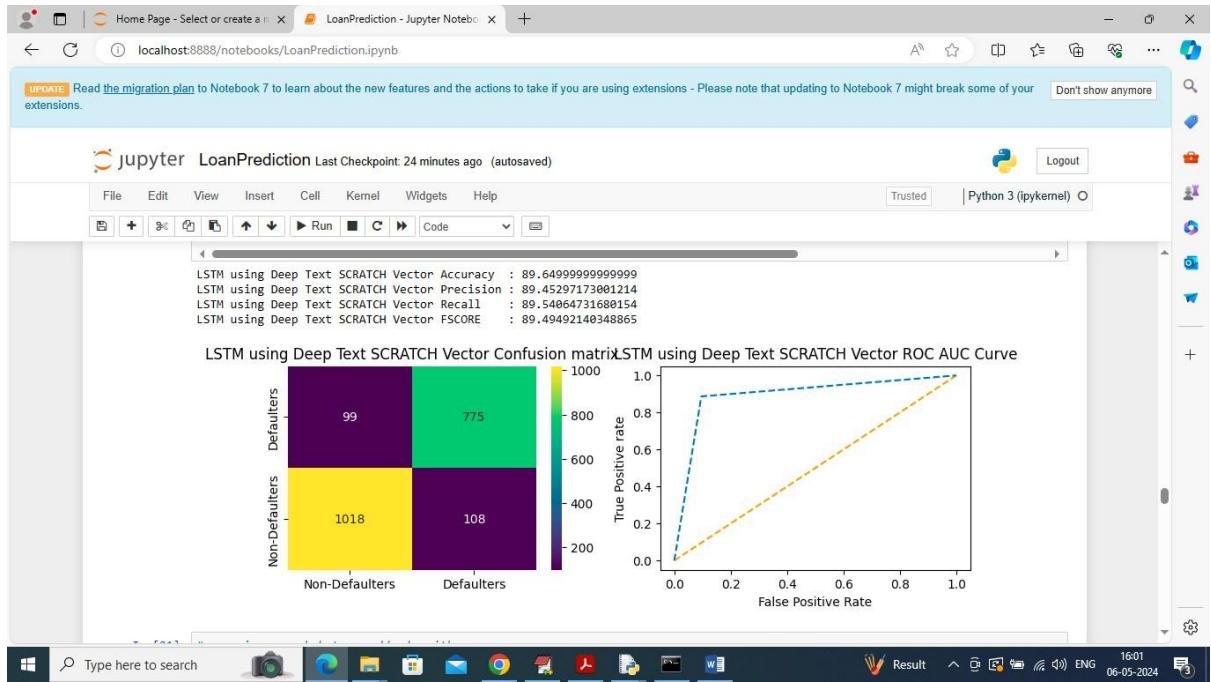


In above screen LSTM on BERT features got 79% accuracy and can see other metrics output also

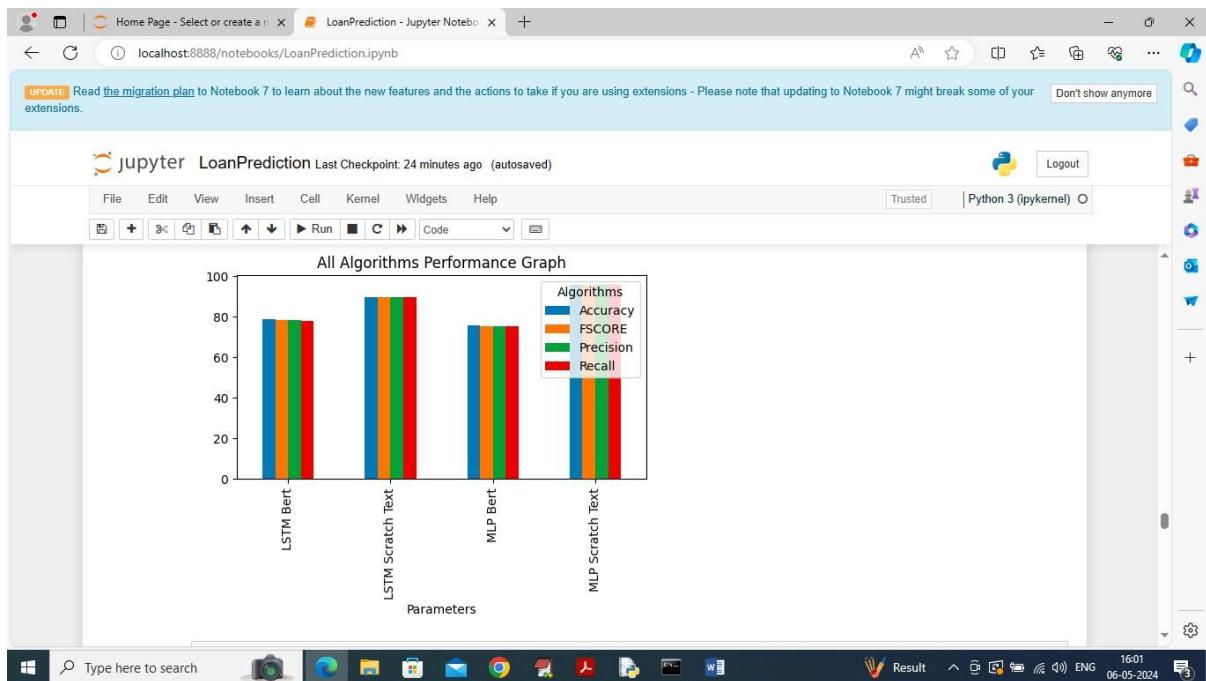
The screenshot shows a Jupyter Notebook interface running on a Windows desktop. The notebook title is "LoanPrediction - Jupyter Notebook". A message at the top says "UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions." Below the message, the notebook content is displayed in a cell titled "In [20]". The code in the cell is as follows:

```
In [20]: #train LSTM algorithm using Deep Text SCRATCH vector features
scratch_y_train = to_categorical(scratch_y_train)
scratch_y_test = to_categorical(scratch_y_test)
scratch_X_train = np.reshape(scratch_X_train, (scratch_X_train.shape[0], scratch_X_train.shape[1], 1))
scratch_X_test = np.reshape(scratch_X_test, (scratch_X_test.shape[0], scratch_X_test.shape[1], 1))
lstm_scratch = Sequential()#defining deep learning sequential object
#adding LSTM Layer with 100 filters to filter given input X train data to select relevant features
lstm_scratch.add(LSTM(100, input_shape=(scratch_X_train.shape[1], scratch_X_train.shape[2]), return_sequences=True))
lstm_scratch.add(Dropout(0.2))
lstm_scratch.add(Bidirectional(GRU(80, return_sequences=True)))
#adding dropout layer to remove irrelevant features
lstm_scratch.add(Dropout(0.2))
lstm_scratch.add(Bidirectional(GRU(64)))
lstm_scratch.add(Dropout(0.2))
#adding another Layer
lstm_scratch.add(Dense(100, activation='relu'))
#defineing output Layer for prediction
lstm_scratch.add(Dense(scratch_y_train.shape[1], activation='softmax'))
lstm_scratch.compile(optimizer = 'adam', loss = 'categorical_crossentropy', metrics = ['accuracy'])
if os.path.exists("model/lstm_scratch.hdf5") == False:
    model_checkpoint = ModelCheckpoint(filepath='model/lstm_scratch.hdf5', verbose = 1, save_best_only = True)
    hist = lstm_scratch.fit(scratch_X_train, scratch_y_train, batch_size = 32, epochs = 15, validation_data=(scratch_X_test, scratch_y_test))
    f = open('model/lstm_scratch.pkl', 'wb')
    pickle.dump(hist.history, f)
    f.close()
else:
```

In above screen training LSTM using SCRATCH features and after executing above block will get below output



In above screen LSTM on scratch features got 89% accuracy



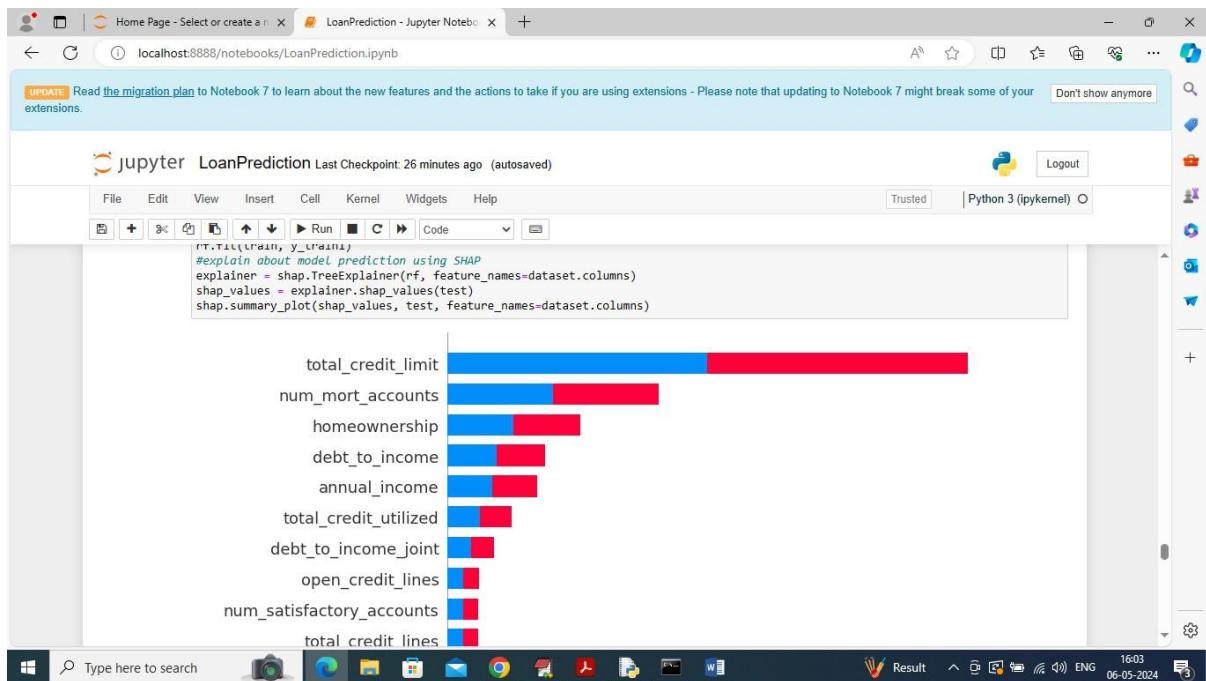
In above screen showing comparison graph of both LSTM and MLP algorithms on BERT and SCRATCH features and can see both algorithms got high accuracy on scratch features. In above graph x-axis represents algorithm names and y-axis represents accuracy and other metrics in different color bars

The screenshot shows a Jupyter Notebook interface running on a Windows desktop. The notebook title is "LoanPrediction - Jupyter Notebook". The code cell In [22] contains Python code to display algorithm performance. The output Out[22] is a tabular representation of the data:

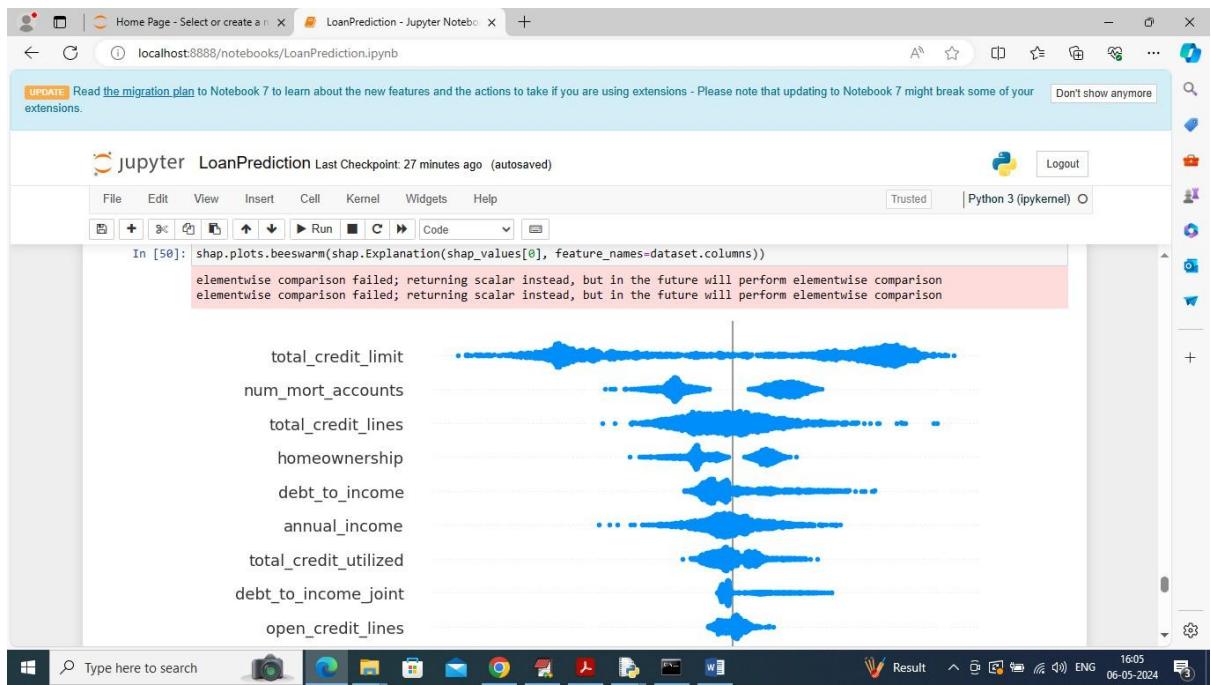
	Algorithm Name	Accuracy	Precision	Recall	FSCORE
0	MLP Bert	75.80	75.396903	75.422929	75.409603
1	MLP Deep Text Scratch	95.60	95.515152	95.591783	95.551889
2	LSTM BERT	78.90	78.465715	78.003274	78.193468
3	LSTM Deep Text Scratch	89.65	89.452972	89.540647	89.494921

The code cell In [48] contains code for interpreting a hybrid random forest model, which is currently commented out.

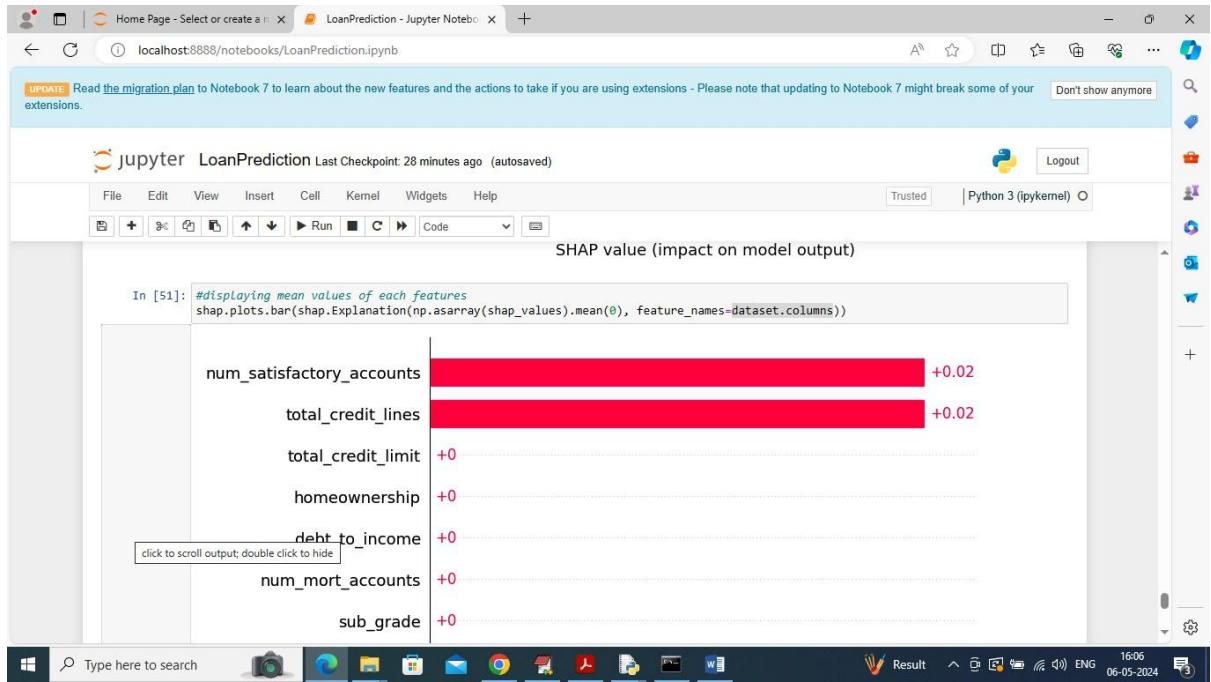
In above screen showing all algorithm performance in tabular format



In above screen explaining and plotting model prediction using SHAP. In above graph y-axis represents feature names and then SHAP displaying in graph which features contributes most in predicted correct class label. Features which is contributing most will have high value



In above screen explaining model using SHAP Beep Swarm plot and this also indicated which features contribute most in correct prediction



In above screen explain about features using mean values.

So in above screen we have trained and tested both MLP and LSTM models using BERT and SCRATCH features and then explain model using SHAP.

# **CHAPTER-11**

## **CONCLUSION**

### **11.1 CONCLUSION**

In conclusion, utilizing deep learning techniques for predicting credit default or bank loan defaults offers a promising approach with several advantages. These techniques can leverage the power of neural networks to learn complex patterns and relationships within the data, leading to more accurate predictions compared to traditional methods. Furthermore, deep learning models can automatically learn relevant features from the data, reducing the need for manual feature engineering and potentially improving model performance. Despite these advantages, it is important to note that deep learning models can be computationally expensive and may require large amounts of data for training. Additionally, interpretability of these models can be challenging, which may be a concern in highly regulated industries such as finance. Overall, while deep learning techniques show great promise for predicting credit default or bank loan defaults, it is important to carefully consider the specific requirements and constraints of the problem at hand before deciding to adopt these techniques.

### **11.2 FUTURE SCOPE**

The title "Deep Learning Techniques Based Predicting Credit Default" represents an advanced approach to tackling the critical problem of predicting bank loan defaults. The future scope of this research direction is immense, as financial institutions continue to rely on more sophisticated predictive models to minimize risk, improve decision-making, and enhance operational efficiency.

One of the key areas of future development includes the integration of more complex and specialized deep learning architectures, such as transformers, graph neural networks (GNNs), and reinforcement learning, to better capture intricate patterns within financial datasets. These models could evolve to handle diverse and high-dimensional data such as transactional histories, customer profiles, and even external factors like economic conditions, enabling more accurate and personalized credit assessments.

## CHAPTER-12 REFERENCES

1. "Deep Learning for Credit Card Fraud Detection" by Q. Wang, L. Zhao, L. Dong, and Y. Yang. This paper discusses the use of deep learning for fraud detection in credit card transactions, which is closely related to credit default prediction.
2. "A Deep Learning Approach for Credit Scoring in Peer-to-Peer Lending" by J. Zhang, J. Zhang, and J. Su. This paper explores the use of deep learning for credit scoring in peer-to-peer lending platforms, which involves predicting the likelihood of default for individual borrowers.
3. "Deep Learning for Loan Risk Prediction" by J. S. Lee, S. Y. Kim, and Y. W. Chang. This paper presents a deep learning approach for predicting loan default risk, focusing on the use of long short-term memory (LSTM) networks.
4. "Deep Credit Risk Analysis" by J. Thomas and D. Seres. This book chapter provides an overview of deep learning techniques for credit risk analysis, including applications in predicting loan defaults.
5. "Machine Learning for Credit Card Fraud Detection - Practical Handbook" by S. K. S. Gupta. This handbook covers various machine learning techniques, including deep learning, for credit card fraud detection, which can be adapted for credit default prediction.