Loan Approval Prediction Using Machine Learning Algorithms

by

Reak Roy
22301776
Tahsin Alam
19301171
Syed Hafiz Kabir
23241063
Mirza Abyaz Awsaf
20101146
Shadik Ul Haque
23141087

A project submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
October 2024

Declaration

It is hereby declared that

1. The submitted project is our original work carried out towards the fulfillment of the degree requirements at BRAC University.
2. This project does not contain material published where other people are the authors or except those areas which are in accessed in full and accurate reference within the project.
3. This project in whole or in part has not been presented or accepted for any other degree or certificate in any university or institution.
4. All the important sources of help and assistance have been clearly stated and acknowledged.

Student's Full Name & Signature:

Reak Roy	Syed Hafiz Kabir
22301776	23241063
Tahsin Alam	Mirza Abyaz Awsaf
19301171	20101146

Shadik Ul Haque

23141087

Approval

The project titled "Loan Approval Prediction Using Machine Learning Algorithms" submitted by

1. Reak Roy (22301776)
2. Syed Hafiz Kabir (23241063)
3. Tahsin Alam (19301171)
4. Mirza Abyaz Awsaf (20101146)
5. Shadik ul Haque (23141087)

Of Summer, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on October 22, 2024.

Examining Committee:
Supervisor:
(Member)

Dr.Amitabha Chakrabarty,PhD

Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Dr.Md.Golam Rabiul Alam,PhD

Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi,PhD

Chairperson and Associate Professor
Department of Computer Science and Engineering
BRAC University

Abstract

This research describes the potential of several classifiers of classical machine learning and architecture of deep neural networks when predicting the status of a loan application. The data set of 613 observations and 13 features, provided with the information about the applicants and their credit profiles, was utilized together with other techniques, such as bootstrapping, for more data qualityutimaltely leading to 9824 observations. Some imputation strategies were applied to deal with the lack of values, while also features were carefully prepared by employing ANOVA, Mutual Information and Tree based approaches among other statistical methods. For the validation of the model performance, the dataset was split into two parts: training (70%) and testing (30%). Many classical machine learning algorithms were applied including but not limited to Logistic Regression, Support Vector Classifiers(SVC), Decision Trees, Random Forests, Multi-Layer Perceptron, Gradient Boosting machines, K-Nearest Neighbors, etc. Out of all models used in the research, Random Forest Classifier demonstrated the most high values of accuracy of 86.84% and F1-score (0.9043), hence it was the best performing one. Advanced methodologies such as SMOTE (accuracy of 88.16%) and ADASYN (accuracy of 87.07%)were also used to handle the issue of class imbalance, where the performance of K- Nearest Neighbors was impressive acuuracy of 88.16% after resampling. In a different, yet similar analysis, five types of neural network architectures, Simple Recurrent Neural Network(RNN), Long-Short Term Memory(LSTM), Convolutional Neural Networks(CNN), Fully Connvolutional Neural Networks(FCNN) and Fully Connected Neural Networks(FCN) were built with the use of Tensorflow, Scikit-learn, and Numpy running on Google Colaboratory notebooks. The outcomes showed that the Fully Convolutional Network (FCN) has the best validation accuracy of 89.75% and validation loss of 0.2255 among the models built.

Keywords: Loan Approval Prediction, Machine Learning, Neural Networks, Random Forest, K- Nearest Neighbors , Bootstrapping, SMOTE, ADASYN, RNN, LSTM, CNN, FCN, Financial Analytics.

Acknowledgement

In the blessings of the Almighty, we appreciate the strength, wisdom, and endurance bestowed upon us during this period.

Special thanks to our supervisor Dr. Amitabha Chakrabarty, who reliably offered his expertise, time, and support in the completion of this project.

Moreover, our family members deserve our deepest gratitude due to their prayers, care, and motivation, as well as our friends who provided us with valuable assistance and discussions.

Dedication

This project is dedicated to those whose support has been the absolute foundation of our studies as well as our lives.

To our family, for their unceasing support and faith in us which helped us strive to the best of our abilities. Their care and help have motivated us throughout and for that, we are grateful eternally.

To our mentors and professors, for their priceless judgment and hard work in nurturing such work, which has greatly impacted the conception and attitude towards this particular work.

And to our companions, whose care and friendship have made the whole endeavor a lot more enjoyable.

Table of Contents

List of Figures

List of Tables

Nomenclature

The following list contains several symbols and abbreviations that will be used later in the document.

CNN Convolutional Neural Networks

DT Decision Tree

FCNN Fully Connected Neural Networks

FCN Fully Convolutional Neural Network

GNB Gausssian Naive Bayes

LR Logistic Regression

LSTM Long-Short Term Memory

KNN K-Nearest Neighbours

MLP Multilayer Perceptron

RF Random Forest

SMOTE Synthetic Minority Oversampling

ADASYN Adaptive Synthetic Oversampling

RNN Recurrent Neural Networks

SVC Support Vector Classifiers

Chapter 1

Introduction

Loan Approval is essential for employees of banks. The main mission of this project is to give an easy was to select good capable candidates. All loans are handled by finance companies. They can see all local, semi-local, and rural areas. After a corporation or bank confirms a client's eligibility, the client gives application. On basis of information clients supply on the application form, an institution mainly banks wishes to speed up the loan eligibility process. These details can be their age, financial status, banking transaction, source of income withdrawal amount and history of credit.The dataset we have collected, which had a set of parameters loans were approved. In order to get accurate findings, this model is programmed in this way. Our main target of this project is to forecast the safeness of loans. Ada Boost, Gaussian Naïve Bayes, MLP, KNeighbors, DecisionTree, LogisticRegression, RandomForest, GradientBoosting, Support Vector Classifiers or SVC algorithms are used for predicting loan safety. Fist of all the datais cleaned to remove any missing values from the data collection. Loan approval is a really important process for banking associations. The system approves or rejects the loan operations. One of the most significant contributing factors to a bank's financial results is loan recovery. It's actually delicate to predict the eventuality of payment of loan by the client.

1.1 Motivation

Recent studies show that the loan approval process is at the core of most credit systems and is critical to both the credit granting organization and the loan applicant. In practice, whereas the problems of loan application evaluation may be solved by assessing applications by several experts using a set of metrics to meet specific profiles, a decision making system that combines the use of stereotypes and evaluation of financial and demographic data is employed. Even so, these procedures tend to be slow and of personal judgment, and there is always a risk of the presence of the so-called "human factor" in these processes. The growing flow of loan requests will require making more precise, efficient, and most importantly, automated systems for making decisions for them.

Machine learning (ML) and deep learning (DL) have changed the point of view of many industries, including the finance sector as well. These processes have made it possible to work on internal processes such as loan approvals and hence make

decisions more accurately and time saving. In this case ML and DL models focus on the previously known data which in the present case is very complex and involves a lot of interdependencies. When it comes to this application, reducing the risks of default is made easier, the time taken for processing the outcome is reduced and less subjectivity is seen in the decision making processes.

Furthermore, the prospect of big data analytics that comes with dozens of variables on the applicants' profile as well as their banking history, demographic and credit scoring information makes it possible to build sophisticated models for decision making. This study is driven by the urgency of the challenge on how modern machine learning and deep learning techniques can be used for predicting phenomena such as loan approval and aims to enhance the precision, scalability and justice of such predictions to financial institutions and customers.

1.2 Project Contribution

The present project offers a variety of serious proposals to the sphere of predicting loan approval.

Systematic Assessment of Classical ML Modeling Approaches: A number of classical machine learning techniques - LogisticRegression, SupportVectorMachine, DecisionTrees and RandomForests, K-NearestNeighbors and GradientBoostingMachine, among others –are reviewed for predicting loan approval. The results are presented in order to show the advantages and disadvantages of each algorithm particularly with respect to financial datasets and the treatment of highly imbalanced classes.

Adaptation of more sophisticated Models of Deep Learning: The current study also investigates the prediction of loan approval using deep learning models such as Simple Recurrent Network, LongShortTermMemory, ConvolutionalNetwork, FullyConnectedNeuralNetwork and FullyConvolutionalNetworks. These models are tested on their capabilities of understanding intricate relations in time series and other data spaces, such as the financial history of the applicants.

Resampling Techniques to Handle Class Imbalance: This study also addresses the problem of class imbalance, where the majority class consists of loan approved applicants and the remaining class rejected applicants is extremely small. This involves using techniques such as SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (AdaptiveSyntheticSampling) to alter the data distribution towards the positive class to avoid training on extremely imbalanced datasets.

Feature Engineering and Selection: Also, to make the model perform better, high end statistical approaches like ANOVA, MI (MutualInformation) and tree based feature additions are incorporated in the study. Because of these approaches, only important variables like credit rating, income, loan amount and applicant's work history are taken in the model's final version.

Assessment of Performance Using Various Models: The models are assessed through a number of metrics. Thanks to this multi-metric evaluation, which enables the understanding of the performance of each model in more detail especially

when handling imbalanced datasets. The Random Forest classifier records the best performance as far as classical ML algorithms are concerned while FCN records the best results for deep learning models.

1.3 Problem statement

While machine learning (ML) holds immense potential to automate and optimize loan approval processes, its integration into the financial landscape raises pressing concerns about potential algorithmic bias, explainability, fairness, and technical limitations. Historical discrimination embedded in training data can perpetuate unfair outcomes for specific demographics[6]. This can manifest in models that disproportionately reject loan applications from women, minorities, or low income individuals, even if their actual creditworthiness is similar [29].Certain ML models, with their complex decision-making processes, can inherently favor certain features over others, leading to biased results[30].Also, the "black box" nature of many ML models makes it difficult to understand the rationale behind their loan decisions. This lack of transparency creates mistrust and hinders efforts to rectify potential biases or errors in the model's reasoning. Moreover, utilizing sensitive financial data for loan prediction raises concerns about data security and privacy violations. Robust safeguards against unauthorized access and misuse are essential to protect client information [31].Furthermore, measuring fairness in loan prediction remains a complex and evolving task. Traditional metrics like accuracy might not capture biases against specific groups. Developing and deploying fairness-aware metrics is crucial to ensure equitable outcomes. Building accurate and reliable ML models and thoroughly validating and testing ML models before deployment is crucial. Ensuring they generalize well to real world scenarios and perform fairly across different demographics is essential for responsible implementation [32].Finally, models trained on limited data might over fit, leading to poor performance on unseen data. Balancing model complexity with data availability is vital to generalizability and robustness. This System refers to the problems and handles imbalance, Data Pre-processing, Models Selection, evaluation metrics, Training, Testing, validation and results to produce a robust outcome.

1.4 Aims and Objective

While giving Loan to somebody there are always risks. So, its better to work for risk management.The data set after follwed by collection and cleaning its time for model training using ur main target of this project is to forecast the safeness of loans. Ada Boost, Gaussian Naïve Bayes, MLP, KNeighbors,Decision Tree, Logistic Regression, Random Forest, Gradient Boosting, SVC and many more alogirhms were implemented . By using them we can potentially find loan defaulters. Besides, exploring different algorithm for more accuracy is also viable. Employ evaluation metrics to find the predictive models effectiveness. So that we can find out Non-Performing Assets (NPAs) which will lead to most revenue earned. It'll save more time and operational cost for both the corporation and the client. As, the processes will be very rapid . Moreover, we'll be able to predict loan repayment probability, eliminate biases and discrimination, personalize loan offers, understanding customer behavior

and lastly to enable Data-Driven decision making. Thus, the recommendation from the customer may increase.

1.5 Methodology

In this project , the research methodology worked upon in order to achieve a well-structured model, both in terms of its development and validation. The following explains how the procedure works:

Data Collection and Preparation: For the purpose of this study, a dataset consisting of 9,824 records and 13 variables describing individuals including their background, history, amount required, income, and credit worthiness is used. Pre-processing steps include the imputation of missing values, which applies diverse strategies, normalization of continuous variables, encoding of categorical variables . Data quality is also improved using other methods such as through the use of bootstrapping.

Feature Selection and Engineering: Futurology is so important in this study since the models can only learn from significant variables. ANOVA, Mutual Information, and tree-based selection are used here for dealing with dimensionality and improving the dimensions of the model. While tree-based methods offer a means to assess the importance of features, their main aim is to help alleviate the dimensionality issues presented by the dataset.

Model Development and Training: The implementation of the classical machine learning models, comprising Logistic Regression, Decision Trees, Support Vector Machines, K-Nearest Neighbors, Random Forests, and Gradient Boosting is executed. Plus, sophisticated deep learning models namely Simple RNN, LSTM, CNN, Fully Connected Neural Network and Fully Convolutional Network are also constructed using TensorFlow, Scikit-learn and other frameworks.

Handling Imbalance in Classes: The dataset that has been used demonstrates that class imbalance exists since there are several more accepted than rejected loan applications. The minority class which in this case is made up of rejected loans is remapped using SMOTE and ADASYN techniques to create more examples of this class. This helps in ensuring that the dataset is balanced and as such there is no tendency for the models to be biased to the majority class.

Evaluation of Models: Several metrics are used to evaluate the execution of each model. The database is used for splits into training and testing with a percentage of seventy and thirty, respectívely and cross validation is done to ensure that the findings are not only correct but can also apply to different populations. Important metrics encompass accuracy, F1 score, precision, recall and valiladtion loss with an aim of developing an overall notion of the model's ability to predict.

Tools and Frameworks: Tensorflow, Scikit-learn, Pandas and Numpy are examples of Python libraries that a model makes use of. The research is being carried out in Google Colaboratory notebooks, which is an advantage as it is a cloud-based environment suitable for effective computation and model training.

1.6 Summary of the Contribution

The contributions made in this project are laid out in a concise manner:
Classical ML algorithms Compared: This section presented an in-depth approach on the performance of several ordinary machine learning algorithms where Random Forest was the best algorithm in predicting loan approval with an accuracy of 86.84% and its F1 score of 0.939.
Performance of Deep learning Models: Out of the different types of deep learning model architectures that were examined, it was found that FCN had the highest validation accuracy of 89.75% and the lowest validation loss of 0.2255 for loan approval prediction.
Class Imbalance-Resampling Techniques: With the help of SMOTE and ADASYN, class imbalance is effectively addressed by the performance of classifiers such as K-Nearest Neighbors. SMOTE with accuracy of 88.16% and ADASYN accuracy of 87.07%
Feature Engineering and Selection: Advanced feature selection methods were applied in the model so that only the viable features are entered in the model hence enhancing the performance and effectiveness of the model.

1.7 Project Outline

Chapter 1: Introduction - This particular chapter explains the purpose, contributions, and methods used in the research work.

Chapter 2: Related Works - This chapter deals with the existing literature on prediction of loan approval, machine learning techniques and application of deep learning in finance stating where there is need for improvement.

Chapter 3: Proposed Model - The Work flow of the project and the classical ML algorithms and deep learning architectures employed in the study will be covered in this chapter including model configurations, hyperparameters, training procedures employed in the research.

Chapter 4:Dataset –this chapter outlines the dataset used, data preprocessing processes and feature engineering processes carried out in the research work.

Chapter 5:Implementation and Results – This chapter contains the findings of the model evaluations, where various models are assessed along several performance metrics

Chapter 6: Conclusion and Future Work – This chapter offers an overview of the results and their interpretation, addresses the deficiencies identified in this research and offers ideas on where the field of loan approval prediction could be advanced in the future.

Chapter 2

Related Works

This paper presents a new machine learning model aimed at minimizing loan defaults and maintaining privacy by preserving sensitive information from borrowers' financial reports. The model uses machine learning and data mining to predict loan eligibility for users, automating the process by identifying eligible segments from online loan application forms. Decision tree algorithms, widely used in banking for classification and regression tasks, are used for loan prediction and severity forecasting. The R package is used for data mining visualization, but real-time consumer data collections may contain imputed or missing data. For classification and regression problems DT is a supervised learning algorithm, using tree representation for prediction. The analytical process includes data cleansing, missing value imputation, exploratory analysis, and model construction. The best accuracy on the public test set is 0.811, with applicants with poor credit history less likely to get approved and high-income applicants more likely to repay loans [14].

The modern banking system heavily relies on its credit system for income, and risk evaluation is recommended to minimize losses and decrease non-profit assets. Customer information is crucial in estimating loan acceptance, and artificial intelligence techniques are used to provide reliable results. The best model for predicting loan acceptance will be determined by comparing these classifiers. The Dream Housing Finance Company provided the data, which was cleaned and white spaces removed. The banking industry frequently uses the selection tree, a non-parametric supervised machine learning method, to solve classification and regression issues. Logistic regression uses a dual dependent variable to lessen system complexity, employing a larger target variable and a bigger number of samples to find the category. Among the collected data, 70 percent of the data was used for training and 30 percent for testing. The LR Algorithm had an accuracy rate of 83.7percent, while the Decision Tree Algorithm provided 85.4percent accuracy. The proposed model predicts loan acceptance using machine learning methods, with decision tree achieving the most accurate results [15].

Micro credit is a small loan program for impoverished borrowers without collateral or credit history, often rejected by traditional financial institutions. China's online lending market has expanded, providing basic financial services to a large user base. A framework, features, and reinforcement learning-based searching strategy are proposed based on user behavior data from 360 Financial' online system. The data includes interactions between users and the platform, with event id being a unique index. Feature Tools is used to generate discriminated features through a

novel search strategy, aiming to answer questions about feature derived, guidance signal introduction, and feature value calculation. The feature engineering problem is transformed into a reinforcement learning problem using a Markov chain transformation link, aiming to find high-information features through a policy gradient method. Using an actual default problem, the proposed method was evaluated and compared with professional judgment and conventional genetic programming. The method was trained on 100,000 users from the 360 Financial online lending system, showing a nearly four times improvement in both velocity and velocity+ features compared to random policies. The paper proposes a performance-driven framework for automated feature generation from raw data using reinforcement learning, unifying feature structure, interpretation, and calculation logic, reformulating the feature generation problem as reinforcement learning. Experiments show the proposed method improves human effort and avoids local optimums in traditional genetic programming [16].

Banks distribute loans; their main asset is the earnings from such loans. By predicting applicant safety and automating feature validation, machine learning may aid in the development of a Loan Prediction System. Both bank workers and loan applicants gain from this quick and easy process, which gives loan applicants a window of time to approve their loans. The probable methodology will be the collection and the deployment of the data set followed by the training of the model on the training dataset and the test the model on the testing data set and after that the results will be analysed. The authors of this paper in 2016 suggested there will be six machine models that will be used as per their paper[1].The models are, DT classifiers, RF Classifers, SVM Classifiers, Linear Model, Neural Networks and Adaboost. The first model used in the paper is the extension of C4.5 algorithm, the basic algorithm of this tree is that it requires all the attribute to be discredited. The random forest is just the group of learning system that works by building large numbers of decision tree. The linear model highlights its use for both distinct quality and multiple quantity factors even if it is mathematically identical to other models.

In the paper[2], the authors suggested a model that is also made based upon the decision tree. The decision tree are widely used in the banking sector due to its highly accurate results. Moreover, its ability make a statistical model makes it more desire able. The decision tree also effectively completes the classification and regression task[1]. The methodology used in the paper[2] is the collection of the data and the preprocessing of the data followed by the building of the classification model using decision tree and finally the prediction from the results[2].The data that has been collected may have inconsistency, preprocessing of the data will be needed to make the algorithm more efficient. Decision tree algorithm is used for loan defaulter and non-defaulter problem prediction, its tendency to provide better results and its intuitive implementation, interpret able predictions, unbiased estimated error, easy tuning, and highest accuracy makes it perfect for implementation in the project. The analytical process involved data cleaning and processing, missing value imputation with mice package followed by building the model that have accuracy of 0.811 from public testing. The results concluded that people with high income and low loan request are more likely to get approved as they may repay bac their loan easily. However, the basic characteristics like gender, martial status are not taken in consideration by the company.

In the paper[3] the author used data set provided by Xiamen International Bank, by using the data and various machine learning models including XBoost, RandomForest, Adaboost, knearestneighbours, multilayerperceptions. The use of the data and models the authors predicts weather the loan will be approved. The data set included 132029 records the can be divided into three groups 1.User basic attributes 2.Loan related information 3.Information related to user credit reporting The featuring extraction was done by removal of useless data and classification of data was done. The models were then used on the data, the base model of XGBoost, random forest and adaboost is mainly a decision tree. In the boosting the base learner was initially trained, according to the performance of the base learner the training samples were distributed. The based on the distribution sample another base learner was trained, the process is repeated until it reaches to the T value. The k-nearestneighbour make prediction based on the k neighbours information, the Multilayer perception is an artificial neural network, each neuron has a series of parameters that can be learned and uses nonlinear function as the activation function, the introduction of the nonlinear function can make MLP more effective. However, the results shows us that RF give 0.5010 acurracy adaboost gives 1.0000 accuracy, XGBoost gives 0.7166 accuracy the kNN give 0.5036 accuracy and MLP gives us 0.5000 accuracy. So from the results we can see that ADAboost gives us 100 percent accurate results which is greater than all the models used
Banks primarily generate revenue from credit lines, which depend on loan repayment and client default rates. Predicting loan defaulters can help lower Non-Performing Assets. Research shows various methods for loan default control, with Logistic regression models being a crucial predictive analytics tool. Kaggle data is used for analysis and prediction.
Small loans are crucial for aspiring entrepreneurs, but they also carry the risk of default. This is a common issue in the financial industry, and banks often offset the loss with other fully paid loans. Peer-to-peer lending companies like Lending Club provide a platform for borrowers to create small unsecured personal loans, with investors choosing which loans to invest in. This shifts the burden of loss from a single bank to several individual investors, requiring diversification to avoid winners and losers. Machine learning, a subfield of artificial intelligence, automates data processing and creates analytical models with minimal human intervention.

In this paper, we used decisiontree, naivebayesclassification,ordinary leastsquareregression, logisticregression, supportvectormachine (svm) and clusteringalgorithms as machine learning algorithms. In result we got, the public test set, data cleaning, missing value imputation, exploratory analysis, and model development yielded the greatest accuracy of 0.811 [8].

For over a century, banks have relied on accurate default risk prediction. With the availability of massive data sets and open-source data, interest in risk prediction has grown. Automating loan approval procedures can expand financing options for small firms and individuals, promoting equitable access to loans. P2P lending, with sites like Lending Club lent over 45 billion dollar, has gained popularity in less developed economies.
Neural Network was applied but to default the prediction here only.L2 regularization was the most commonly used regularization strategy in grid search for LR and SVMs to prevent overfitting.The study used recall and AUC metrics for result validation,

considering credit risk and rating in relation to other loans. Logistic regression was applied to combined data, and a hyper parameter grid search was optimized to maximize the average unweighted recall. The recall macro was prioritized over AUC to avoid over fitting rejected classes, as AUC weights accuracy over forecasts.
In results, automated P2P loan acceptance and default prediction, with high rejection and default recalls. The methodology could reduce defaults to 10 percent, improving market efficiency, and using Logistic Regression for approval and Deep Neural Networks for default [9].

The banking industry's credit lending sector is facing rapid expansion and competition from new start-ups, leading to negative credit losses. To address this, research is needed to design effective models that exploit existing data and provide strong predictive prototypes. This will help banks maximize profits by understanding applicant tendencies, money usage, and default predictions.Here in the data set contains 850 bank default payment records, which were preprocessed using techniques like cleaning, integration, formatting, and normalization. The predictive model's accuracy was assessed using methods like MLR, DT, SVM, Random Forest and other algorithms. The study used a dichotomous default payment as a dependent variable and compared categorization findings to the destination's score. The research was implemented using Python on a local machine using the Jupyter Kernel. Eight major explanatory factors were identified, including age, educational background, employment status, address, income, debt, credit to debt ratio, and other factors.15 Python-based classification methods, including LR, SVM, and Naive Bayes, after data pre-processing. To evaluated many other metrics such as CM, Precision metrics, Recalling metrics and FI which enhance the likelihood of identifying and fixing algorithm errors, leading to improved results.This paper uses classification algorithms to predict bank loan defaults, focusing on job experience and debt income. Python performance indicators are used to identify problematic consumers, improving credit approval. This was the outcome [10].

This research uses data from former bank clients to predict loan safety using a machine learning model trained on 1500 examples, 10 numerical characteristics, and 8 categorical features. Factors like CIBIL Score, Business Value, and client assets are considered when deciding whether to credit a loan. A well-liked machine learning approach for classification issues that focuses on predictive analysis is logistic regression. It presents data and clarifies connections between independent nominal, ordinal, and ratio level variables and binary variables. The sigmoid function is used in the model's development with binary outcomes as the aim. Bank clients' data is split into training and test sets, and any missing values are filled in using the mean, median, or mode. For exploratory data analysis, Feature Engineering approaches are applied with a focus on loan-eligible consumers. Data preparation, processing, imputation, experimental analysis, model construction, assessment, and testing are all steps in prediction process. The best accuracy scenario is 0.811. Loan applications with modest loan amounts are more likely to be granted than those with excellent credit scores and lower credit limits. Gender and marital status are not taken into account [4].

Distribution of loans is a basic business function of banks, and credit risk assessment is essential for banks globally. The main goal is to place assets in trustworthy hands. There is no assurance that the applicant picked is the appropriate one,

despite the fact that many banks approve loans following a rigorous process of verification and validation. A loan prediction system can swiftly and simply identify worthy candidates, giving banks a particular edge. The system permits priority application checks, time constraints for applicants to verify loan sanctioned status, and computerized calculation of the weight of characteristics in loan processing. The conclusions of this document, which is intended only for management authorities at banks and financial institutions, may be forwarded to relevant departments for necessary action. Vaidya, Ashlesha's article forecasts loan approvals using logistic regression as a machine learning method. Power terms and nonlinear effects may be accommodated by the model, but parameter estimation needs independent variables and a sizable sample. The application of artificial neural network as an early warning system for identifying loan hazard is covered by Yang, Baoan, et al. In a prediction model for dynamic lending, genetic algorithms are employed to optimize profit and reduce loan approval mistakes. Modeling accuracy for Logistic Regression, Decision Trees, and Random Forest are 80.945 percent; 93.648 percent; and 83.388 percent when Cross Validation Results are 80.945 percent; 72 This means that although decision tree based model achieves best accuracy with the data set, random forest is more interpret able and generalized, despite only having slightly higher cross-validation score when compared to logistic regression [5].

A key component of a bank's operations is loan distribution, with the main purpose being to place assets in trustable hands. There is no guarantee that the chosen candidate is the deserving one, despite the fact that many banks grant loans using a similar procedure. A Loan Prediction System employs machine learning to automate feature validation and forecast applicant safety. This approach offers rapid, immediate, and simple ways to choose suitable candidates, which is advantageous for bank workers and applicants. It establishes time limitations for applicants, determines the weight of characteristics in loan processing, and provides priority review of particular applications. Only management authorities of banks and financial institutions may use this system. Customer segmentation, high-risk loan applications, anticipating default payments, promotions, collateral monitoring, asset grading, regular sales management, stock holdings management, cash management, and cross-selling are just a few financial industries where data mining is an essential tool. It is essential for managing client profiles and transaction data in banking, enabling users to make informed choices. A person's likelihood of repaying financial obligations depends on their credit score, which categorizes applicants into those with excellent credit and those with low credit. While credit rating distinguishes between present and future customers, credit evaluation links a customer's characteristics to previous borrowers. It's crucial for both banks and clients to monitor default vulnerability. An algorithm for classification is also used to forecast results based on data. The credit validity forecast framework filters through advance solicitations from a current bank data set with a 66 percentage preparation set and a 34 percentage test set using a choice tree and computed relapse enlistment information mining. The application helps banks anticipate credit status and make informed decisions, reducing bad loans and cut-offs. It uses AI calculations and packages to analyze data and make informed decisions. This technology aids in identifying necessary data from vast information, reducing bad credit issues. It also aids in attracting new clients, maintaining credit, avoiding extortion, identifying misrepresentations, offering customer-based products, and enhancing customer relationships.[6]

With the development and application of diverse concepts, the technical world is moving closer to automation. An important characteristic that attempts to imitate intellect similar to that of humans in computers is artificial intelligence. In this modern age, technologists want to collaborate with people to bring forth new discoveries. Some of these include machine learning, neural networks, fuzzy logic, NLP (natural language processing), and expert systems. Advanced industries are leveraging machine learning to boost sales growth because being able to build an analytic model with much less coding is what makes it a powerful technology to begin with. machine learning offers a way to replace some of the exploitation of humans with something else for as long as there is growing "big" data in a big data economy? Data sets are used by money lending companies to decide which applicants will be granted loans. These files provide pertinent data on things like gender, education, income, and property type. Using logistic regression and variables including education, credit history, self-employment, and property area, the model forecasts the likelihood of a loan being approved. The model must adhere to the requirements of money lending companies and be accurate and quick. The logistic regression parameters are used to compute the likelihood that the loan will be authorized, and if the likelihood is larger than 0.5, the loan will be approved. The loan will not be granted if the likelihood is less than 0.5. The logistic regression model, a statistical machine learning model used for predictive analysis, is covered in this work. It emphasizes precision, handling of non-linear effects, and power terms. However, logistic regression has drawbacks such a high sample size, reliance on independent variables, and incapability to deliver continuous outputs like forecasts of temperature rise [7].

2.1 Machine Learning / Classifier

Machine Learning augmentation has redefined the ability to analyze large and complex data sets. Also, the ability to learn different patterns among the data set to find out the relation between them. By implementing various Machine Learning Classifiers it can help with the automation which can reduce the need for manual human input. Thus, reducing human error also will lead to more accurate results. Furthermore, Machine Learning Classifiers can handle the high dimensional data to find out connections from various features. These classifiers can be trained and tested thoroughly for more optimized results. Giving a flexible framework to accurately predict loan approval.

The use of machine learning for loan approval prediction and credit card fraud detection are explained in this paper [17]. The banking sector aims to secure assets through verification processes, but this process can be time-consuming and ineffective. A system has been developed to predict loan applicant suitability using machine learning algorithms, achieving 92% accuracy using the Random Forest Algorithm. This paper used an online banking transaction repository data set to analyze and classify transactions if it is fraudulent or normal. A New Web Application for Predicting Loan Approvals and Detecting Fraudulent Transactions has been implemented It used Random Forest Algorithm and Support Vector Machine Learning Algorithm for improved accuracy. The paper also highlights the need of finding and protecting fraudulent transactions, and it utilized the Support Vector Machine Algorithm to analyze and preprocess data. It was then tested on a 615-row

training data set, achieving 92% accuracy for loan approval predictions. Finally, it was tested on a 30,000-customer data set, achieving 94% accuracy for credit card fraud detection.

In this paper [18] Machine learning models, such as XGBoost, random forest classifier, and support vector machine classifier, are used to predict loan approval. Many people are unable to back loans to banks, leading to losses for banks. The main reason for getting a loan is to fulfill the needs of something, such as business growth or a loss. The problem arises because not everyone can loan, and if they can't return, the lender, company, or bank gets in the loss. Main intention will be the loan will be given not. After testing XGBoost gives 77.7778% accuracy. Random Forest gives 76.3889% accuracy and Decision Tree 64.58% accuracy.

To enhance predictions and minimize defaults, a logistic regression model, utilizing Kaggle data, considers not only checking account information but also personal customer attributes in this paper [19]. Sensitivity and specificity are compared between models, revealing marginal improvement. Using a Logistic regression algorithm with data from previously approved loans. The data set works on 1500 cases with 10 numerical and 8 categorical attributes, including CIBIL Score, Business Value, and customer assets. Parameters such as qualification, income, loan amount, and credit history contribute to the model's efficacy. Logistic regression model with a sigmoid function is employed. The preprocessing phase, involving exploratory data analysis and feature engineering, consumes significant time. Two separate datasets are fed into the model for robust predictions. Imputation, feature engineering, data mining and cleaning are used for a better result. Evaluation methods such as confusion metrics, accuracy, precision, recall, and F1 score guide the selection process. The chosen model must meet the stakeholder requirements and constraints. So, the best case using this data set can obtain an accuracy rate of 81.10%.

Due to fierce rivalry, banks frequently struggle to gain the upper hand over one another and improve overall company. The vast amount of data that is readily available. The establishment of knowledge bases and their effective use have enabled banks to create effective delivery channels. Data mining can be used to optimize business choices. The main source of risk that the banking business faces is credit risks, which include the risk of loss and loan defaults. The primary feature of this loan credibility prediction system is its usage of the Decision Tree Induction Data Mining Algorithm for loan request screening and filtering. An Assemblage, a preexisting bank dataset with 4520 records and 17 attributes, is mined for data to create a Tree. The final dataset is split into a 34% test set and a 66% training set after preprocessing. The classifier's final output is validated using the test set. This section presents the findings from the experimental analysis used to estimate the loan repayment capacity. Our suggested paradigm has been implemented in ASP.NET-MVC5 The prediction was made using a bank dataset that already existed. For the experimental analysis, a somewhat sized bank dataset (4520) was utilized. Following the pre-processing stage, the dataset was manually reduced to 3271 by performing dimensionality reduction. The manual addition and use of Information Gain as an attribute evaluator and Ranker as a search yields the ranks of the attributes [20].

Both customers and bank representatives find Loan Prediction to be of great use. This project's goal is to provide a quick, easy, and expedient method for selecting

the primary client. The purpose of the Loan Prediction System is to enable prompt application so that it may be verified based on need. This project is exclusively intended for the bank's or account organization's supervisory authority. The primary objective of data mining research is to obtain a large amount of obtained data, making it a very active and significant field of study. Data mining is becoming more popular mainstream in a financial sector given that effective investigative methods exist for separating obscure. By applying the covering technique, ascribes emerged as fundamental components among the absolute of 31 attributes of the variable importance chart collection of media transmissions. The tribute forecast model is said to be greatly impacted by these 21 attributes. Accuracy foreseeing a customer's purchasing behavior through a disorganized grid with precision 91.36. ROC twists are often applied equally to requests to take into account the classifier's yield. The true positive rate is determined by the Y-hub, while the x-pivot shows the false positive rate, which ranges in value from 0.1 to 1.0. Accuracy in anticipating a customer's purchasing behavior using a disorganized grid with precision 92.18Typically, ROC twists are applied equally to investigate the classifier's yield. The Y-hub shows the true positive rate, while the x-pivot establishes the false positive rate, with a value ranging from 0.1 to 1.0. Consequently, the tendency is for the exactness of the two models to almost remain unchanged [21].

Artificial intelligence algorithms and machine learning models have applications in a variety of industries, including education, healthcare, entertainment, and other professions. Credit ratings and loan conditions are the characteristics that most likely influence the outcome, as we found during this investigation. A training set (80%) and a test set (20%) were then created from the dataset. We utilized MATLAB to train twenty-seven different machine learning models. Bayesian optimization was used to three models in order to determine the optimal hyperparameters with the least amount of error. Our validation methodology was 5-fold cross-validation. The dataset was split into a training set (80% of the data, or 3416 observations) and a test set (20% of the data, or 853 observations) prior to training. Using the training set, we used MATLAB R2023a to train 27 machine learning classification models. Bayesian optimization was used to optimize the hyperparameters of three of the models. Table 1 provides a summary of the research findings. It is evident that the optimal outcome, 98.45% accuracy on the training set (validation), and narrow neural networks were used to achieve 98.83% on the test set. The training set (validation) accuracy of an optimized ensemble classification model was 98.42%, and the test set accuracy was 98.83% determined. Optimized ensemble model was the second model in Table 1 that achieved 98.42% validation accuracy and 98.83% test accuracy. We all determined the optimal hyperparameters for the model using Bayesian optimization. The minimum classification error plot is presented in Fig. 10. It is evident that during the optimization process, the classification error dropped to 0.015809 [22].

The majority of bank revenues come from loans. Financial banks value loan approval. As rates rise, banks struggle to appropriately assess requests and mitigate risks when predicting consumer loan payments. Numerous researchers have studied loan approval system prediction in recent years. Machine learning is useful for forecasting large data sets. Loans are banks' principal income and risk. Many of a bank's assets come from loan interest. Risks include borrowers not repaying loans

on schedule. The term is "credit risk". Loan approval or denial credibility was determined. This paper attempts to explain Machine Learning techniques that accurately identify loan beneficiaries and let banks detect loan defaulters, decreasing credit risk. Our models include Random Forest, Decision Tree, Naive Bayes, and Logistic Regression. The process of analysis begins with data purification and missing value processing, followed by exploratory analysis, model creation, and model evaluation. Higher accuracy and other performance criteria indicate the public test set has the best accuracy. This document can help predict if a candidate will receive a bank loan or not [23].

A bank loan is a credit offer offered to a customer or business by a bank. One of a bank's fundamental financial products is lending, and interest on loans creates the majority of its profits. After an accurate sequence of verification and validation, the loan corporations grant a loan. However, they are still unsure if a particular application would be able to repay their debt. In banking operations, manual procedures are typically employed to decide whether an applicant is qualified for a loan from their bank. This project's principal objective was to analyse if an application is acceptable for a loan by collecting information from numerous sources and employing machine learning algorithms to extract essential data. This would enable banks and lending organisations to decide on the best course of action for each loan approval. The field of aritficilal intellingence or AI known as machine learning or ML is dedicated to teaching computers how to learn without the need for predetermined, explicit guidelines.Models employed here include LogisticRegression, DT classifier, RF Classifier,XGboost are used which got 77.8%,68.1%,73.5% and 76.7% accuracy correspondingly. To conclude, Logistic regression provided the best results. Perhaps in the near future, this prediction module and the automated processing system module will be integrated [24].

With the rise in the bank business, a big group of people are requesting bank loans. It only issues loans to a restricted number of applicants due to its limited resources, thus deciding who would be the best candidate for a loan and which will be more financially sound for the bank is a frequent procedure. We therefore strive to minimise the risk factor in identifying the safe people in this research in order to save a substantial amount of bank resources and labor. Nowadays, getting loan is very tough. We can assess whether a certain thing is safe or not using this way, and machine learning technology has automated the entire feature validation operation. The models that are used include e RandomForest (RF), SVM and Treemodel with GeneticAlgorithm (TGA). This application matches with all Banker requirements and functions as planned. This section is straightforward to tie into a variety of other systems. It is highly accurate, satisfies all banker criteria, and is interoperable with various other systems. Multiple computer failures, content difficulties, and weight fixing in computerized prediction systems were identified. In the near future, banking software that connects with an automated processing unit may be more dependable, accurate, and dynamic. Numerous instances of content errors, computer breakdowns, and most critically, fixed feature weights in automated prediction systems that give more dynamic, safe, and consistent weight modification [25].

In the paper 'Loan analysis Predicting Defaulters' the authors used dataset from kaggle, the dataset consists 855969 numbers of data and among which are 46467

data of failed loans. The LAPD is a credit risk scoring model that uses historic data to predict future defaulters by identifying patterns. Further in the process the data were pre-processed and label encoder technique was used to convert the variable data to numerical value. Later, the data was splitted into test and train sets in the ratio of 7:3, the 70% was used for training and the rest of the 30% was used for testing the model. Thee algorithm used by the authors were LogisticRegression, DecisionTree, RandomForest and AdaBoost. The results were, LogisticRegression gave an accuracy of 62%, the DecisionTree gave an 90% accuracy result whereas Ada Boost gave 86% accuracy and random forest gave 92% accuracy which is the highest the author concluded [26].

In the paper 'Loan Prediction Using random Forest and Decision Trees', the authors collected the data sets from the banking sector, it consists of 12 attributes. The data was splitted into testing and training sets, after the pre processing were done by the authors. Later the data was given to the model training set to be specific to train the model. Then the testing set was given to the model to see weather the prediction are right. Two machine learning classification model was used by the authors the Random Forest classifiers and the Decision Tree classifiers. The Decision Tree is an extension of C4.5 classification algorithm, the experiment was done by the use of J48 Decision Tree which is an implementation of C4.5 Descion Tree. However, with a confidence factor of 0.15 the accuracy was 62.12% and if the confidence is 0.25 the accuracy is 63.39%, so if the confidence factor is high the accuracy is high. The Random Forest had been experimented several ways with different parameters each time the best results without all attribute selection was 87.75% [27].

The authors mentions that the data taken for the paper was taken from housing company finance, the dataset consists of both demographic and socioeconomic characteristics of individual borrower.Moreover, it was retrieved from Kaggle data repository. Later the data was preprocessed and splitted into training and testing sets. The algorithm used here were SVM, LR, and Naive Bayes.Here the data was splitted into 70% for training and 30% for testing. The results are presented in two tables one with the attribute of Early_R and other without the attribute of Early_R, the results are different of each table. However, from the first table the results shows that the LR has an accuracy of 92%, SVM has accuracy of 83.6% and the Naive Bayes has accuracy of 91.8%. Furthermore, the second table without the attribute of Early_R has less accuracy that the table before, the Lr has 87% accuracy, SVM has accuracy of 83.6 and the Naive Bayes has accuracy of 85.6%. Therefore it is concluded that for both the cases the LR gives the best results [28].

2.2 Machine Learning / Classifier Summary

Table 2.1: **Classifer Summary Table**

Ref	Task	Classifier/Model	Data set	Accuracy
17	Web application	SVM, Random Forest	N/A	92%
18	loan approval prediction model for the banking sector using machine learning classifiers .	XGBoost, random forest, SVM, and decision tree	N/A	64.58%
19	Streamline loan approval prediction by prioritizing data preparation and optimizing accuracy	Logistic Regression	Kaggle	81.10%
20	Loan Credibility Prediction System Based on Decision Tree Algorithm	Decision Tree	N/A	66% training, 34% test.
21	A Comparative Analysis of Feature Selection for Loan Prediction Model	Random Forest classifier, Boruta classifier	N/A	92.18%
22	Comparing ML Classification Models on a Loan Approval Prediction Dataset	Machine Learning Classification, Ensemble Model	N/A	98.83%
23	Loan approval Prediction Based on Random Forest Algorithm.	RandomForest algorithm, DecisionTree algorithm, NaiveBayes algorithm, LogisticRegression	N/A	Needs to be tested.
24	Collecting eligible application by utilizing machine learning techniques.	Logistic regression, Decision tree, Random forest, XG boost	N/A	77.8%, 68.1%, 73.5%, 76.7%
25	Determining the best loan candidate ensuring financial soundness of the bank	Random Forest (RF), SVM and Tree model with Genetic Algorithm (TGA)	N/A	Needs to be Tested
26	Building LAPD(Loan Prediction System) and Integrating it with Web Application	Logistic Regression, Decision Tree, Random Forest, Ada Boost	Kaggle	62%, 90%, 92%, 86%

Continued on next page

Table 2.1 – *Continued from previous page*

Ref	**Task**	**Classifier/Model**	**Data set**	**Accuracy**
27	Making a Loan prediction system using random forest and decision trees	Random Forest, Decision tree	Banking Sector	At confidence factor 0.15 accuracy 62.12%., at 0.25 the accuracy 63.39%

Chapter 3

Proposed Model

3.1 Workplan

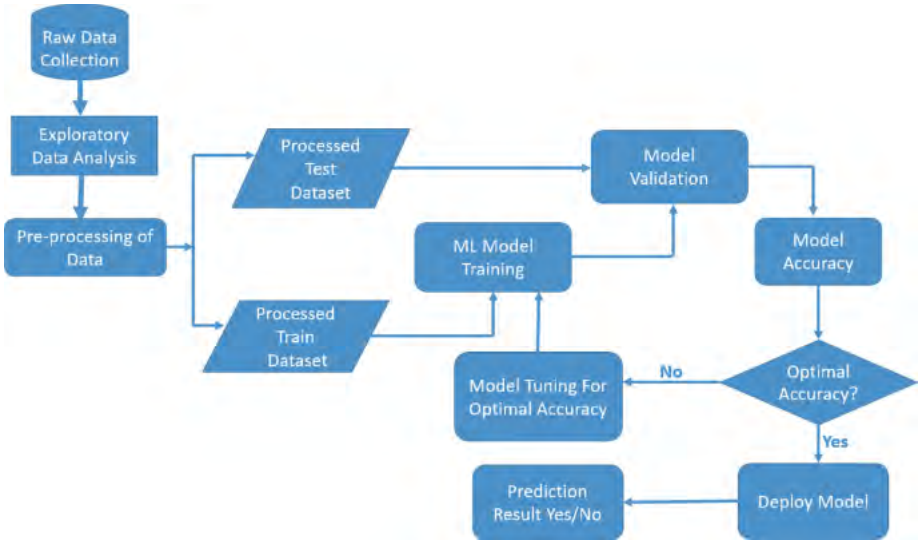

Figure 3.1: Workplan

The proposed model utilizes comprehensive datasets encompassing applicant demographics, financial history, credit scores, loan characteristics, and approval outcomes. Addresses missing values, inconsistencies, and class imbalances using appropriate techniques. Extract meaningful features from raw data. Experiment with different MLalgorithms that includes DecisionTree,RandomForests, LogisticRegression and more. The metrics such as Accuracy, Precision, F1 and finally the Recall to validate models.

3.2 The Nine Classification Model

Total nine classification models are used in our suggested model. Logistic regression, ADABoost, SupportVectorClassifier (SVC), Multi-layerPerceptron Classifier (MLP), RandomForestClassifier(RFC), DecisionTree, K-Neighbours, GaussianNaiveClassifier and Gradient Boosting are some of the algorithms we have suggested. We used classifiers so that we could have a comparative analysis and see which one produces the,optimal results

3.2.1 Logistic Regression

The LogisticRegression is a form of regression analysis which is used for predicting the outcome of a categorical dependent variable.Thus is mainly based on one or predictor variables.The model is popular when it comes for traditional learning, besides it is great for classification problems [53].

The equation 3.1 describes the logistic function. This function is used to model the relationship between the independent and the dependent variables in logistic regression

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{3.1}$$

This function outputs values between 0 and 1, making it suitable for probability estimation.
Prediction: The output of the sigmoid function, $\sigma(z)$, is interpreted as the probability that the input $\mathbf{x}$ belongs to the positive class (usually labeled as 1):

$$\sigma(z) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)}} \tag{3.2}$$

To conclude, logistic regression uses a linear model to calculate a value z and applies the logistic function to convert z into a probability, and then it uses this probability to classify the input. The model is trained by adjusting the coefficients to minimize the log-loss function.

3.2.2 ADAboost

It is an ensemble learning approach that combines several weak classifiers to produce a strong classfier. It is the abbreviation for adaptive boosting. ADAboost's primary goal is to increase prediction accuracy by concentrating on hard to classify objects.By concentrating on examples that are challenging to classify, the strong and adaptive ADAboost algorithm improves the performance of weak classifiers and produces a final model that is more accurate and reliable. Finding a more precise weak learning condition in multiclass problems and obtaining a more fixed generalization error restriction are two examples of boosting theory. ADAboost stopping circumstances, anti noise capability increase, and accuracy improvement through base clasifier variety dhould all be thoroughly investigated.

3.2.3 DecisionTree

The decisiontree and a well-known ML algorithm which is used for classification, regression related work. Decision tree is a tool that recursively divides the data from splitting. Leaf nodes consisting of the values for your input features (which grows to be a tree) so as to segment data or make predictions based on the characteristics efficiently.
Every internal node is a choice based on the value of one feature, and every leaf node in presents a pre-set class label or continuous value. The tree is divided into the following sub sections,RootNodeInternal NodeLeafnodeValue the branches. The tree building involves choosing the best feature to consider for cutting the data in each node. G ionin:This is especially used in case of classification where we¯t to find the least homogenous splits use the impurity and Information Gain, where Gi in impuity denote that a lower value of gi would be better lower split and higher formation gain depict a better split[50].

$$Equation for Gini Impurity : Gini(D) = 1 - \sum_{i=1}^{C} p_i^2 \tag{3.3}$$

$$Equation for Information Gain : IG(D, A) = Entropy(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} Entropy(D_v) \tag{3.4}$$

3.2.4 Random Forest

RandomForest is a commonly used ensemble learning method for tasks. Regression and Classification entering the scene And to run, it constructs tons decision trees in the process of training. An answer to this question can either be the latter, take an mean of predictions in case of regression outcomes Output: For a classification problem, it provides the class to which an individual tree belongs. Random Forest is a strong and very versatile algorithm able to generate high accurate predictivesions with many domain and applications. Big ability to manage dataset with higher dimension[39]

3.2.5 K-Neighbours

The K-NearestNeighbours algorithm is a non-parametric technique used in both regression and classification.The KNN method, or K-Nearest Neighbour, has been extensively utilised in data analysis and ML because it is straightforward but incredibly practical with unique execution.After training sample data, classification is used to forecast the labels of test data points. Although various classification techniques have been proposed by researchers in the last few decades, KNN remains one of the most often used techniques for classifying data sets. The input is made up of the k nearest examples in each area; its is selected from an array of objects or objects with similar attributes; This collection all items might be referred to as the dataset for training purposes [40].

3.2.6 Multilayer Perceptron

Dense layers that are entirely linked convert any input dimension to the required dimension.Multi-layer perception refers to a neural network that has multiple layers. Neural networks are made up of connected neurones, several of which outputs serve as inputs for other neural networks.Any amount of hidden layers and nodes can be identified in each hidden layer of multi-layer perceptrons.They have a hidden layer with an arbitrary number of nodes for each output, an input layer with a single neurone (or node) for each input, and an output layer with a single node for each output [41].

3.2.7 GaussianNaiveBayes

Gaussian Naive Bayes (GNB) is a machine learning algorithm mostly used in continuous data, because of its simple structure and effectiveness in classification tasks. It is another type of the Naive Bayes classifier which simplifies computations while providing reliable results, assumes the features follow a Gaussian normal distribution. As it follows Bayes' theorem, the algorithm depends on prior knowledge of related conditions to calculate the probability of a certain event.
In GNB, the posterior probability of a class given the feature set , the prior probability of the class P(A), the probability of the features given the class P(BA), and the P(B) is the total probability of the features.

In GNB, the features are mostly independent of one another, which allows for faster calculations as the probability computations are simplified. When applied to normally distributed data, GNB assumes that the likelihood of each feature given a class follows a Gaussian distribution, characterized by its mean and standard deviation . One of the prime assumptions in GNB is that they follow a bell-shaped curve, on which the data features are normally distributed [55].

3.2.8 GradientBoostingClassifier

The Gradient Boosting Classifier works to form a stronger and more accurate model by sequentially building an average of weak models like decision trees. First of all, the algorithm starts with an initial prediction, like predicting the average value or class probability and after this, the algorithm calculates the previous errors made by other classifiers, identifying where the predictions are incorrect. After that, in order to capture the patterns missed by the previous model a new weak learner, usually a shallow decision tree, is trained on these errors. As each new learner is added, it corrects the mistakes made by the earlier models to improve the overall prediction. The predictions from the new learner are then combined with the previous predictions which updates the ensemble to be more accurate. This process continues iteratively, with each step focusing on reducing the errors from the previously used learners. Throughout the process, the model gets better with each iteration as the algorithm ensures that it is gradually descending along the gradient of the error. Regularization techniques like shrinkage or subsampling may also be applied to control the learning process and prevent overfitting. Lastly, once it reaches a targeted number of iterations or when the improvement in accuracy almost riches the expected result the algorithm stops. As it incrementally improves the model and

learning from it's recent mistakes, the Gradient Boosting Classifier builds a highly accurate predictive model over each iteration [54].

3.2.9 Support Vector Classifiers

The Support Vector Classifier of more commnly known as SVC, it is mainly used for fiiting the data that has been provided.The SVC mainly provides us with the best fit plane that helps us tho catagorize the data we have [52].
The SVC finds the hyper-plane which efficiently divides different classes of the SVC. The hyperplane should be chosen so that the distance between it and the nearest datapoints from each class is as large as possible in order to best utilise the margin. It shows the hyperplane's equation in the equation 3.5 [53].

$$w^T x + b = 0 \tag{3.5}$$

3.3 Implementation of the Neural Models

Neural Networks are very reliable in predicting loan approval mostly because of their capability to learn complex and high dimensional relationships even in non-linear financial data. Their self-learning ability eliminates the need for elaborate feature engineering as they capture feature interactions quite easily. While these systems are capable of operating over big data, they also have retraining capabilities which allows them to learn new information which enhance their predictive performance. The reliability of such models allows them to better perform with respect to the training data without overfitting. These models can also have more than one target variable and can work with other models for more complex predictions. Overall, owing to the ease with which such networks can be trained to provide high prediction accuracy makes neural networks ideal for approval prediction of loans.

RecurrentNeuralNetwork (RNN)

A Structured RNN introduces a novel way of modeling data that has a sequence by retaining the hidden state along time instances [42] RNNs incorporate the use of feedback loops within their features, as opposed to feedforward neural network systems, thus enabling the retention of a hidden state of the system in various time instances [47]. This property allows RNNs to understand the sequences both temporally and in context.
The RNN represents:

$$h_t = \sigma(W_{xh} x_t + W_{hh} h_{t-1} + b_h) \tag{3.6}$$

here the h_t is the hidden state at time t and x_t is the input at time t, and W_{xh} and W_{hh} are weight matrices.

Activation Function

The sigmoid activation function is introduced in order to fit any real number value into a number between 0 and 1. This can be expressed mathematically as:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3.7}$$

BinaryCross-EntropyLoss

The Binary Cross-Entropy loss is suitable for this RNN model because it allows for probabilistic interpretation of outputs, making it effective for training purposes.

$$L = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{3.8}$$

LongShort-TermMemory (LSTM)

LongShort-TermMemory (LSTM) networks can be classified under the umbrella of RNNs because they are able to model and remember complicated patterns in data over relatively long sequences [42]. The issue of short-term dependency, which has been a setback to most recurrent neural networks, is addressed by the LSTM architecture. An LSTM unit contains gates as follows:

- **Forget Gate:** Determines what information is allowed to leave the cell state.
- **Input Gate:** Allows new data to enter the cell state.
- **Output Gate:** Determines what part or how much of the hidden state can be output.

These gates can be represented as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{3.9}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3.10}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{3.11}$$

The binary cross Entropy loss function is being used for binary classification that is also similar to that mentioned in equation number 3.8.

Convolutional Neural Network (CNN)

Convolutional neural networks (CNNs) are increasingly utilized for the analysis of time series data due to their ability to perform feature extraction from raw input without any hand designing [46]. More specifically, in contrast to the conventional approaches, CNNs use convolutional layers with the help of learnable filters which systematically scan over the data and are able to capture the temporal characteristics of the data quite efficiently [44]. Time series data convolved with Convolutional Neural Networks (CNNs) consist of several important steps. To begin with the convolution step takes a time series input, applies a filter to it and produces feature maps. Another one comes a non linearity such as ReLU which is often used as the activation function after convolution i.e. ReLu activation function. Another layer used at the end of the network called the expansion stage, fully connected

layers, also applies weights sums to the output from the previous layers after the dimensions of the previous layers have been flattened. To conclude the structure, a sigmoid function is then applied for the purpose of predicting two class labels. The training of the model is done through the binary cross-entropy that predicts the target variable and true class labels which requires the adjustment of network parameters through backpropagation which is usually done in layers [47].These can be expressed mathematically as:

Convolutional Operation: The convolution operation applies a filter (or kernel) to the input data. For one-dimensional time-series data, the equation is:

$$y[n] = (x * w)[n] = \sum_{m=0}^{M-1} x[m] \cdot w[n-m] \tag{3.12}$$

Activation Function: To indroduc non-linearity an activation function is intoduced, ReLU or also konwn as Rectified Linear Unit is a coomon choice

$$a[n] = \max(0, y[n]) \tag{3.13}$$

FullyConnectedLayer: After all the layers, the output is flattened and passed to a fully connected layer. The output of a fully connected layer can be represented as:

$$z = W \cdot v + b \tag{3.14}$$

Output Layer : For binary classification, the output layer often uses a sigmoid activation function:

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}} \tag{3.15}$$

Fully Connected Neural Networks (FCNNs)

FullyConnectedNeural Networks (FCNNs) are powerful tools for binary classification, leveraging interconnected neurons across multiple layers [47]. The output of each neuron is a weighted sum calculated with an external nonlinear function, which in this case assists in computing class probabilities. Training the model is done via backpropagation, which regresses the binary cross-entropy loss of the model performance. This task has a binary decision making learning from the input data and thus, FCNNs are efficient in this tasks [48]. The basic functions can mathematically be represented in the following ways.

Fully Connected (Dense) Layers: The fully connected layers perfors matrix multiplication between the input and the weight matrix and finally add a bias term. Then an activation function is applied, which introduces non-linearity into the model. The equation for a fully connected layer is:

$$y = \sigma(W \cdot x + b) \tag{3.16}$$

The binary cross entropy is calculated just like the CNN.

Activation Function: ReLu or RectifiedLinearUnit is used in the hidden layer to introduce the non-linearity and to prevent the model from collapsing.The Sigmoid-Function is used in the output layer to convert the raw output into probability for binaryclassification.

Fully Convolutional Neural Networks (FCNs)

In binary tasks like loan prediction, Fully Convolutional Neural Networks (FCNs) have come up as a novel architecture. These networks are made up of neurons with convolutional layers only [44]. Thus, they can learn spatial characteristics of the presented data effectively without much feature engineering.

$$h_i^{(l)} = \sigma\left(\sum_{k=1}^{K} W_k^{(l)} x_{i+k-1}^{(l-1)} + b^{(l)}\right) \tag{3.17}$$

After the convolutional layers and pooling, the data is flattened into a 1D vector to feed into fully connected layers. This transformation is done mathematically by reshaping the output of the previous layers into a single vector:

$$x_{\text{flat}} = \text{Flatten}(h) \tag{3.18}$$

Where h is the output from the last pooling layer.
3. Dense Layers (FullyConnectedLayers) The output of the flattened layer is passed through dense (fullyconnected) layers. Each fullyconnected layer computes:

$$y = \sigma(W \cdot x + b) \tag{3.19}$$

Chapter 4

Dataset

4.1 Data Description

In this section we'll discuss the datasets that have been collected. A Loan Approval Prediction Dataset typically consists of structured data containing information about individuals or entities applying for loans, along with the outcome of their loan applications. The dataset is used to train and evaluate machine learning models that predict loan approved or not based on the provided features. The goal is to develop different models that accurately predict whether an applicant is likely to be approved for a loan, assisting financial institutions in making informed lending decisions.

4.2 Data Preprocessing:

The dataset has 613 observations with 13 factors like Gender, Married Status, Applicant Income, Loan Amount, Credit History and the target variable of Seeks loan. We created additional data using bootstrapping and resampling techniques. To provide more data points for model training, this dataset was increased to 9,824 rows using bootstrapping. In the process of preparing the data sets for analysing the Label Encoder technique was utilized in order to transform the categorical features into numerical values. Moreover, in the later process the data sets were splitted into two parts one is the training set which will be used to train the models and the other is the test set which will be later used to test weather our models are giving accurate results. Furthermore, the missing values were also explored and by utilizing the missingno library the data patterns were visualized.

4.2.1 Handing Missing Values

We used different imputation methods, replacing missing values with mode or mean based on feature categorical or numerical. Categorical Variables: Categorical columns were encoded using label encoding and directly replacing column values to make them compatible with machine learning algorithms.

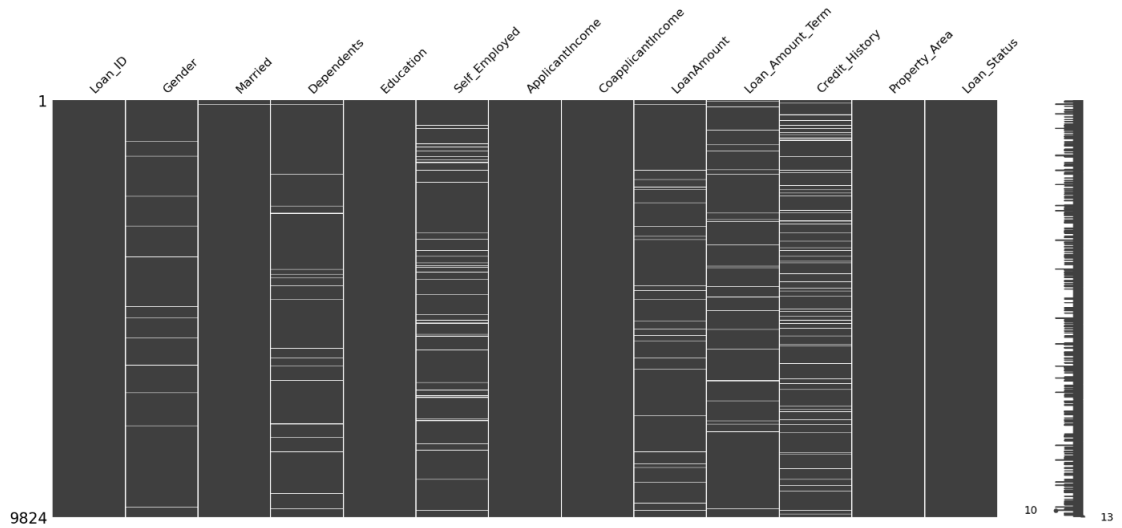

Figure 4.1: Missing values before mean and mode

As is it shown in figure 4.1, there are too many columns missing with a small amount of null values, therefore we used mean and mode to replace with NaN values. The Y values with 1 and N values with 0 were replaced as well and the same for other Boolean types of columns. Then by the use of Label Encodes some specified categorical columns in both the training and testing data sets will be replaced with numerical representations. So, the figure 4.2 shows that there isn't any missing value in the data set

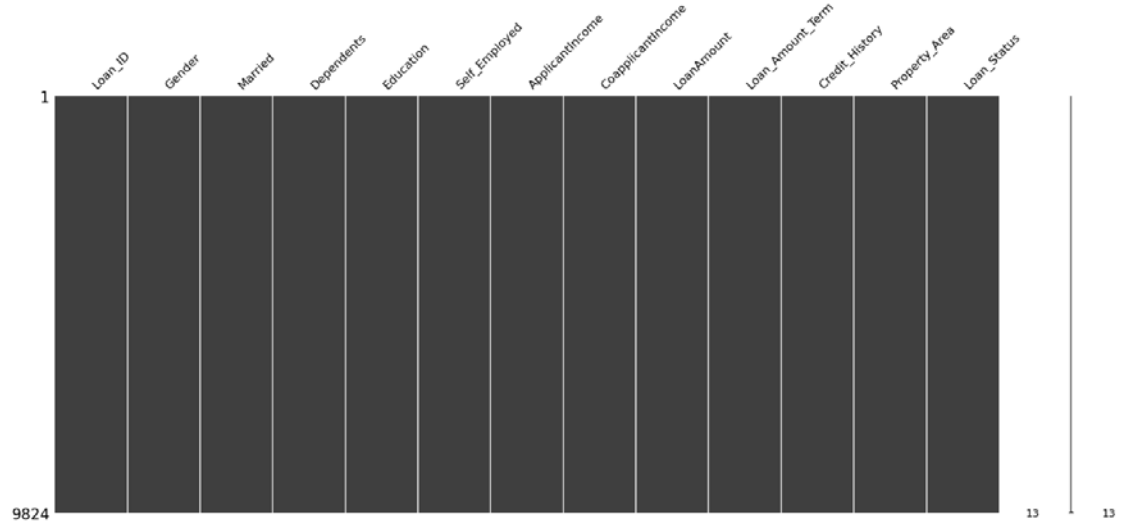

Figure 4.2: Missing values afteer mean and mode

4.2.2 SMOTE and ADASYN

With the help of SMOTE and ADASYN, we are presently implementing two different methods for resampling and data augmentation in order to overcome the problem of imbalanced class distribution in a binary classification task.

SMOTE: This is a method that targets the minority class and involves generating new samples based on the blending of current minority class samples. This method ensures that the number of minority class samples is on par with that of the majority class samples and this helps to eliminate the imbalance present in the data set.

ADASYN: This is also a technique for oversampling like SMOTE. Nevertheless it focuses more on areas of the feature space where there are very few controllable instances of the minority class which are difficult classify, particularly along the boundary. The objective is to create synthetic instances in order to balance the classes although balance is probably not symmetrical in practice as depicted in the output.

Before resampling:

Class 1 (minority class): 7017 instances
Class 0 (majority class): 2807 instances
There are more instances of class one as compared to class zero. Above one case is more than two times cases and hence there is an extreme disparity between the two cases.

SMOTE Resampling:

After SMOTE resampling:
Class 1: 7017 instances
Class 0: 7017 instance

SMOTE does this by adding new synthetic samples specifically for the minority class (class 0) until both classes have same sized samples (i.e, class 1).

ADASYN Resampling:

After ADASYN resampling:
Class 1: 7017 instances
Class 0: 7025 instances

As with ADASYN, this technique also creates a few extra synthetic samples for the less produced class, hence leading to a distribution that is very close to but not exactly equal.The figure 4.3 shows us the balance of the dataset after both technique was applied. Moreover the table 4.1 also shows the number of instances after SMOTE and ADAYSN is applied

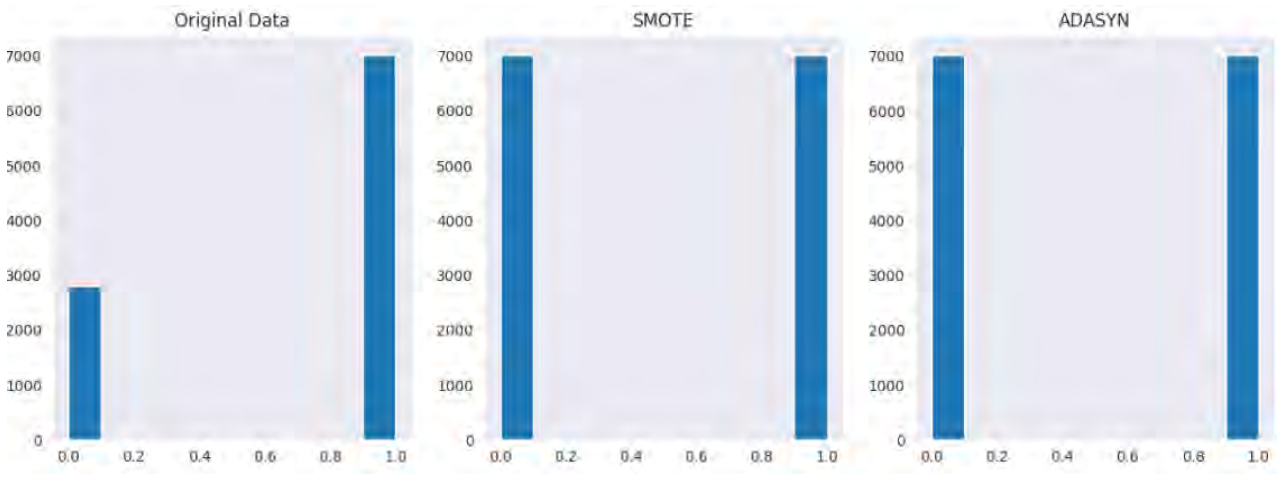

Figure 4.3: Balancing the Dataset

Resampling Method	Class 1 Instances	Class 0 Instances	Total Instances
Original	7017	2807	9824
SMOTE	7017	7017	14034
ADASYN	7017	7025	14042

Table 4.1: Class Distribution Before and After Resampling

4.2.3 Feature Selection:

Several dataset-dependent feature selection techniques were used to evaluate the significance and relevance of features. First of all an ANOVA F-test was used to assess the statistical significance of each feature. Next, Mutual Information was applied to determine the dependency of features on the target variable. Additionally, tree-based methods such as Random forest were utilized to compute feature importance, providing insights into the acknowledgement to attribute to the its performance. Finally, the Chi-Squared test was conducted to evaluate the statistical relevance of categorical features, ensuring a comprehensive feature selection process.

Moreover,Data Standardization Final standardization was performed on selected numeric features to ensure consistent scaling across the dataset. It was clear from the feature selection mechanisms the same features are playing an important role in predicting loan status i.e CreditHistory, LoanAmount, Education,ApplicantIncome and CoapplicantIncome. These findings underline the necessity for feature engineering when performing predictive modeling.The following Figure 4.s shows us the visual representation of the feature plot graph of the dataset

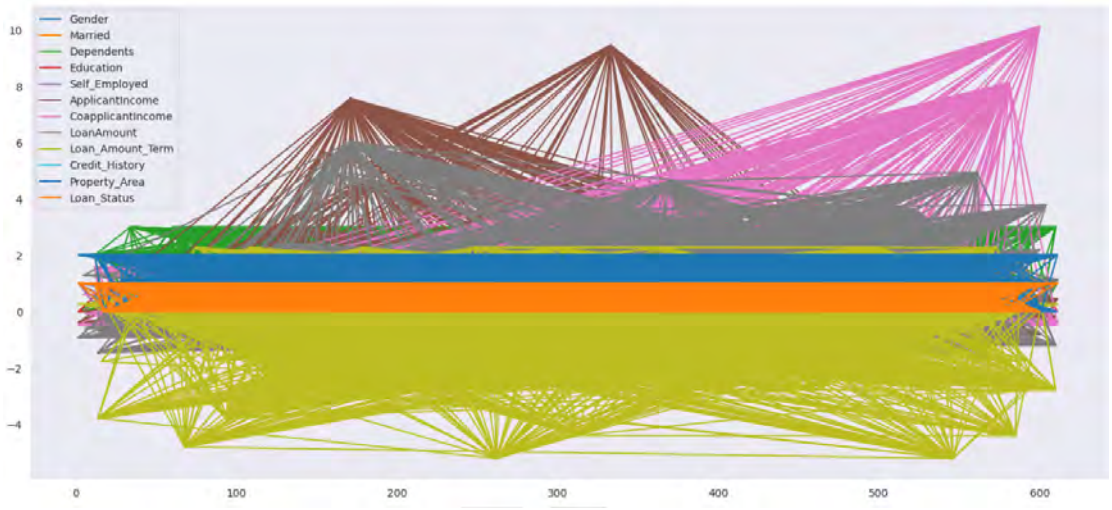

Figure 4.4: The Feature Plot graph of the dataset

4.3 Data Prepraration

The Dataset used contains features relevant to loan decision making which are also typically part of applicant profile. for example:

Loan Amount: The amount of loan requested by the applicants

Education: The educational status or qualification of the applicants

ApplicantIncome: How much does each applicant earn every year

CoapplicantIncome: How much does each coapplicant earn every year

Credit History: Do borrowers have some record of successfully repaying loans

LoanStatus, the target variable it is a binary variable indicating whether the loan was approved (1) or not (0) The first step we took in our process was to split our dataset into training and testing subsets. The training subset consists of 70% of the data in all, or about 6,876 records.

The testing subset has 30 %, 2,948 records. This ensures that models have enough information to learn from while still having enough unseen data for evaluation.

Few of the factors which are considered for loan approval prediction are like Co-applicant Income, Applicant Income, Credit History, Education and Loan Amount.

Chapter 5

Implementation and results

5.1 The Result Analysis of Implemented Classical Models

In this section of the study, we evaluate the performance of various machine learning classifiers for predicting loan approval using a set of features related to applicants. The focus is on comparing accuracy, precision, recall, F1 score for different classifiers to determine the best-performing model for this task.
In this study, prediction of loan approval was attained using different machine learning classifiers such as Desicion Tree, Spport Vector Classifier(SVC), Random Forest and many more.Performance results obtained using Imblearn, Scikit-learn, NumPy and other libraries in Google Colaboratory.

5.1.1 Evaluation Metrics

The models were evaluated on performance using the following measures:

Accuracy: Represents the percentage of instances which were correctly put in their respective classes over all the instances.

Precision: Out of all the positive predictions made, how many were actually positive.

Recall: Out of all the actual positive samples, how many were correctly identified as positive.

F1-score: This is the average of precision and recall whose aim is to find a middle ground between the two metrics.

5.1.2 Simple Train And Test Result

Research suggests that multiple classifiers trained on the available loan status prediction dataset exhibit respective train accuracy and test accuracy scores. The outcomes portrayed here indicate how effective is each model with respect to the training data (to check if the model has been able to 'learn' the patterns correctly) and to the test data (to see how well the model can perform outside the training set).The dataset was splitted into 70% for training and 30% for testing.

Training Score:
This shows the performance of the model on the sample of data, which was used for training. Typically, high training accuracy indicates that the model learned the patterns contained in data. However, if training accuracy is very high, it usually means that the model was overfitted and learned irrelevant data patterns that may not be useful for predicting new data.

Accuracy Score:
This shows the performance of the model on data that was not used for training (test set). The more similar the test accuracy to the training accuracy, the better is the model's generalization. The Score results are shown below in table 5.1 and figure 5.1

Model	Train Accuracy	Test Accuracy
Logistic Regression	0.811082	0.7805
AdaBoostClassifier	0.810791	0.7890
SupportVectorClassifiers	0.884380	0.7815
DecisionTree	0.884380	0.8609
RandomForest	0.860384	0.8375
Multi-Layer Perceptron	0.848604	0.8280
K-Nearest Neighbors	0.874055	0.8392
Gaussian Naive Bayes	0.811227	0.7815
GradientBoostingClassifier	0.813118	0.7890

Table 5.1: Score Results of test and train for classic models

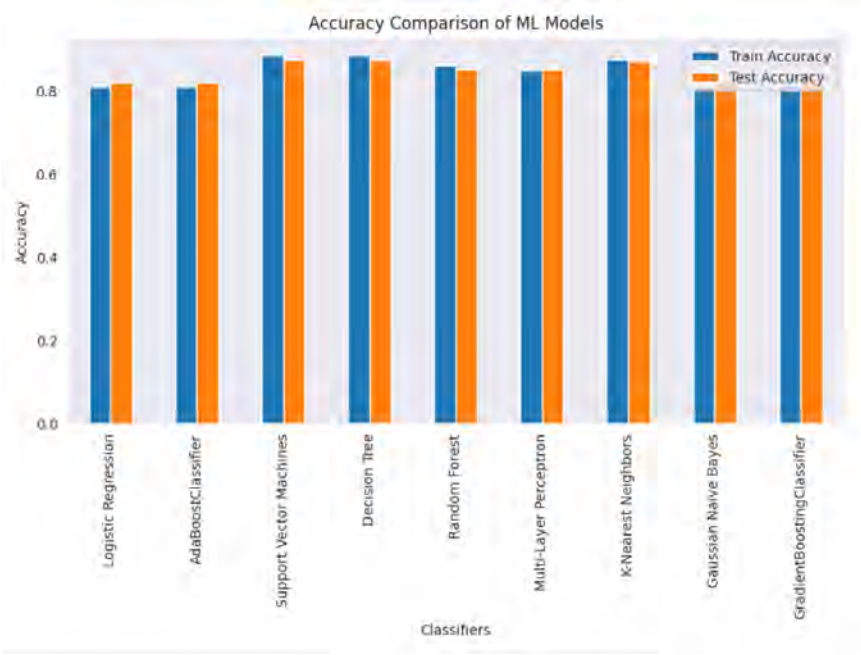

Figure 5.1: Accuracy Comaprision Of the Classic ML Modles

5.1.3 Test Performance for Imbalanced No Resampling Data:

For both the Support Vector Machine and Decision Tree models, they exhibit high training and test performance with similar accuracy on the training and test datasets respectively as shown in the table 5.1. This indicates that these types of models do not suffer from overfitting as they are able to achieve good performance on unseen datasets.

The Random Forest and K-Nearest Neighbors (KNN) classifiers also yield good results, though the evaluative accuracy for the former is slightly less than the corresponding training, suggesting a small degree of overfitting, yet a reasonable rendition.

Logistic Regression, AdaBoost, and Gaussian Naive Bayes models perform notably lesser than the SVM and Decision Algorithm in terms of accuracies, albeit showing consistent performances in the Train and Test datasets, respectively. This suggests that these models have a lower risk of overfitting as they might be simpler in structure compared to fitting complex problems.

The Performance of Gradient Boosting is fair with accuracy in the range of the high seventies. Its performance might improve if hyperparameter tuning is done.
The performance of Multi-Layer Perceptron leaves room for improvement since the difference between the train accuracy and test accuracy is wider than for other models hinting at possible overfitting.
The best training and testing accuracy is observed in the Decision Tree and SVM models. Some models such as Logistic Regression and Naive Bayes are very basic and nonlinear patterns may not be expected from them, however, they are stable in their work. Other models such as Random Forest and KNN have an accurate performance level however they are more complex and adjustment in their use is necessary to prevent overfitting.
Performance Metrics
The following table 5.2 summarizes the performance of each classifier in terms of accuracy, precision, recall, F1-score, and confusion matrix statistics.The figure 5,1 and figure 5.2 shows the Accuracy and the Evaluation Metrics of the Classical Models

Classifiers	**Accuracy**	**Precision**	**Recall**	**F1**
LogisticRegression	0.7805	0.7646	0.9822	0.8599
SVC	0.7815	0.7625	0.9896	0.8613
DecisionTreeClassifier	0.8592	0.8601	0.9490	0.9024
RandomForestClassifier	0.8609	0.8559	0.9584	0.9043
MLPClassifier	0.8375	0.8343	0.9520	0.8893
GradientBoostingClassifier	0.8280	0.8186	0.9624	0.8847
KNeighborsClassifier	0.8392	0.8495	0.9302	0.8880
GaussianNB	0.7815	0.7625	0.9896	0.8613
AdaBoostClassifier	0.7890	0.7829	0.9579	0.8616

Table 5.2: Classifier performance comparison.

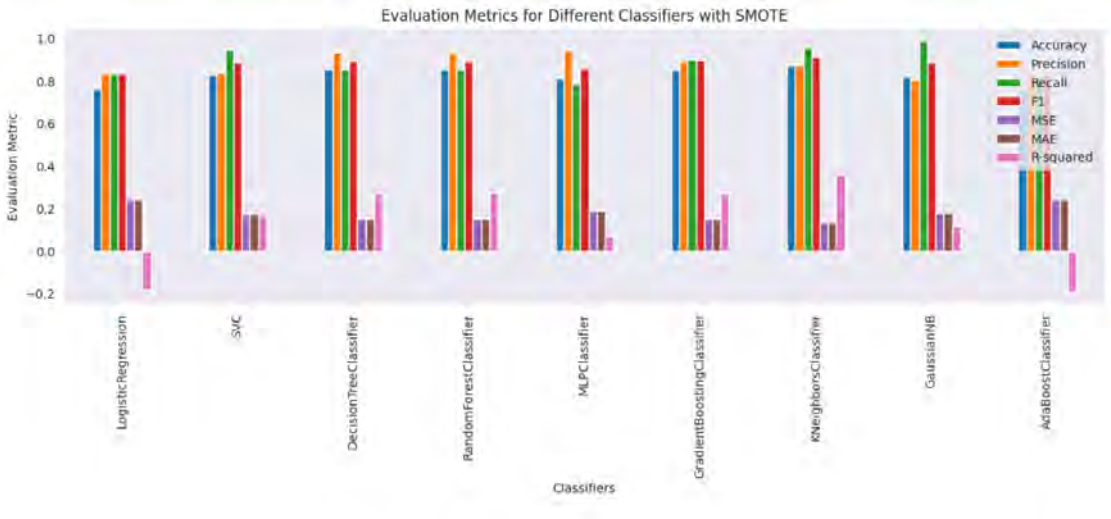

Figure 5.2: Evaluation Metrics Of the Classic ML Modles

5.1.4 Visualization of the Confusion Matrix

Here the results of the classical models are given below, the figure 5.3 is the visualization of the confusion matrixes for the implemented classical models. The table 5.3 shows the values that has been obtained from the confusion matrixes for the classical models that were used:

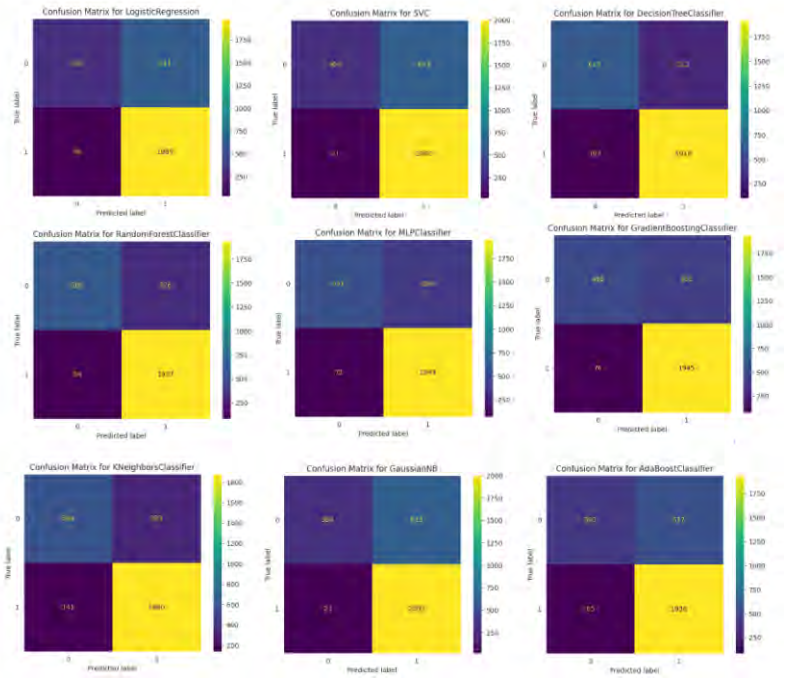

Figure 5.3: Confusion matrix for the Classic models

Model	TruePositives	FalsePositives	FalseNegatives	TrueNegatives
Logistic Regression	1985	611	316	36
SVC	2000	623	304	21
Decision Tree Classifier	1918	312	615	103
Random Forest Classifier	1937	326	601	84
MLP Classifier	1949	396	531	72
Gradient Boosting Classifier	1945	431	496	76
K-Neighbors Classifier	1880	333	594	141
GaussianNB	2000	623	304	21
AdaBoost Classifier	1936	537	390	85

Table 5.3: Confusion Matrices for Different Classifiers

Key Insights:

According to the findings from table 5.2, RandomForestClassifier proved to be the most accurate model. The accuracy obtained was 0.8609 and the F1 score was 0.9043 which means the model is the most efficient and balanced for this dataset. DecisionTreeClassifier also showed strong performance results with high accuracy (0.8592) and precision (0.8601) and the F1 score was 0.9024. This particular model was able to provide the best balance between precision and recall, making it ideal for applications where both need to be achieved. Likewise, SVC and GaussianNB have the highest recall (0.9896), which means they are designed for tasks where there is a greater concern about false negatives than false positives.

Acoording to the results of the table 5.2, RandomForest Classifier offers the best overall performance, particularly in terms of accuracy, F1 score. DecisionTreeClassifier also performed well, with a strong balance of precision, recall, and accuracy. For applications where recall is critical, SVC and GaussianNB may be preferable due to their ability to minimize false negatives. Future work can explore further model optimization and alternative performance metrics based on domain-specific requirements.

Moreover, from confusion matrix It can be seen that all the classifiers have a high True Positive ratio which is indicative of correctly classifying positive classes and the best performance is recorded by the SVC and Gaussian Naive Bayes classifiers respectively.

The DecisionTreeClassifier has the provision for the most Observe .True Negative in distress hence it is a classifier that predicts negative classes effectively. Good TN is also practiced in Random Forest.

There have been cases especially in the curves and SVC where these models exhibit high False Positive rates hence can lead to serious reclassification in very important functions.
From the tblae 5.3 we see that the SVC is noted to have the least False Negative making her positive instance sensitive which works well in situations where pay-off for false negative is high.

Based on the confusion matrices from table 5.3 out of all RandomForestclassifier performed the best. It does well by having very high true positives (1937), adequate true negatives (601) and few false negatives (84).

The analysis of the performance metrics shows that it is very rare for actual negatives to be classified as positives but assures the positives will be detected. This mixture

is important due to the fact that in very many instances both types of errors can be very costly.

RandomForestAlgorithm did best as classification accuracy since it is able to maintain good records of true negatives and at the same time, true positives are very high.

5.2 Evaluation of Performance with Different Classifiers After SMOTE

SMOTE has been applied to adjust the class imbalance within the data set and the classifiers; performance has been assessed and compared using several metrics - Accuracy, Precision, Recall, F1 Score.The following figure 5.4 shows the evaluation metrics that have been obtained after the SMOTE was applied. The later figure 5.5 and table 5.4 shows the confusion matrix of the model after SMOTE has been applied.

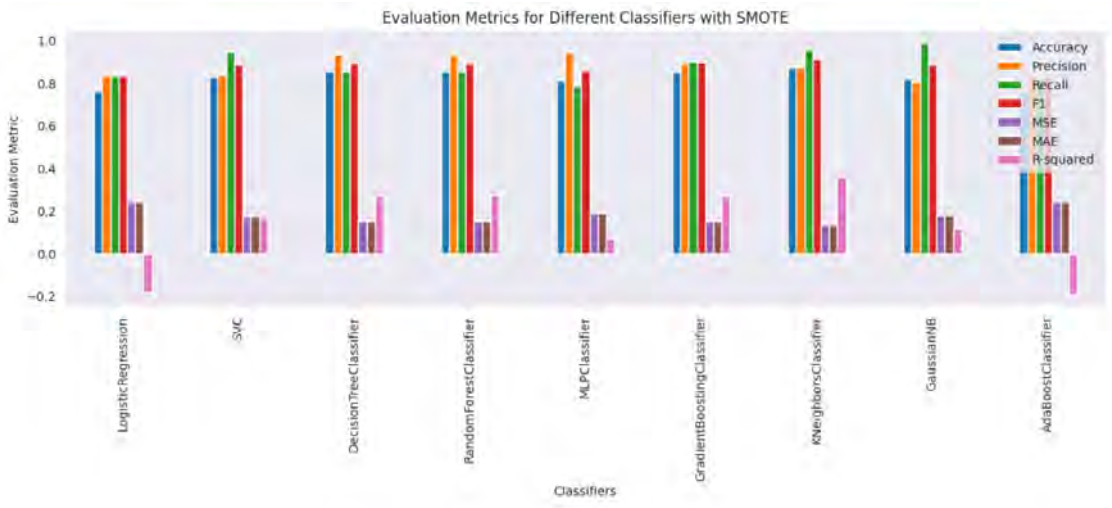

Figure 5.4: Evaluation Metrics of the Classical model after SMOTE

Model	TruePositives	FalsePositives	FalseNegatives	TrueNegatives
Logistic Regression	480	476	75	1917
SVC	480	476	21	1971
Decision Tree Classifier	848	108	280	1712
Random Forest Classifier	848	108	280	1712
MLP Classifier	874	78	430	1562
Gradient Boosting Classifier	675	281	200	1792
K-Neighbors Classifier	722	234	115	1877
GaussianNB	479	477	87	1905
AdaBoost Classifier	560	396	131	1861

Table 5.4: Confusion Matrices for Different Classifiers after SMOTE

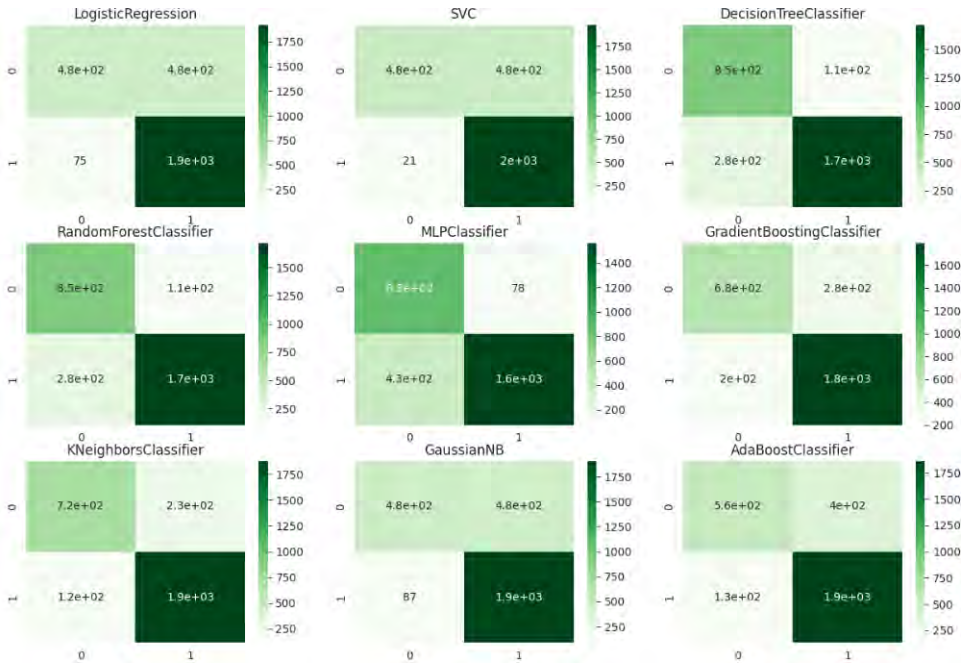

Figure 5.5: Confusion matrix of the Classical model after SMOTE

Classifiers	Accuracy	Precision	Recall	F1
LogisticRegression	0.813094	0.801087	0.962349	0.874344
SupportVectorClassifiers	0.831411	0.805476	0.942269	0.888038
DecisionTreeClassifier	0.868385	0.940659	0.859438	0.873873
RandomForestClassifier	0.868385	0.940659	0.859438	0.898216
MLPClassifier	0.852782	0.939121	0.836345	0.884758
Gradient Boosting Classifier	0.836839	0.864448	0.899598	0.881673
K-NeighborsClassifier	0.881615	0.889152	0.989458	0.939305
GaussianNB	0.808684	0.799748	0.956325	0.871056
AdaBoostClassifier	0.821235	0.824546	0.934237	0.875971

Table 5.5: Classifier Performance Metrics SMOTE

So, based on the exploration of the confusion matrix and key metrics we can decide:

Best Classifier (Overall):From the Table:5.5 we can seee that the K-Nearest Neighbors (KNN) performs particularly well on recall (0.989) and has a high F1 score (0.939). It has an impressive accuracy (0.882) while retaining an almost equal precision and recall. It has less number of false negative cases most of the cases than the other models.

Strong Performers: Random Forest and Gradient Boosting Classifiers present good precision, recall, and balance overall. These are best when the precision is primary the Random forest performs quite well as it has an excellent f score and less of false identification than the rest.

Models to Avoid: When it comes to models like AdaBoost and Logistic Regression, they do not perform well in situations with an overwhelming amount of false

negatives, which means they tend to miss more positive cases than any other models and as such, they become useless in the case of imbalanced dataset.

It can be summarized that KNN, Random Forest, and Gradient Boosting models, when applied, perform effectively with SMOTE-resampled data.

5.3 Evaluation of Performance with Different Classifiers After ADASYN

ADASYN has been applied to adjust the class imbalance within the data set and the classifiers; performance has been assessed and compared also using several metrics -Accuracy, Precision, Recall, F1 Score.The following figure 5.6 shows the evaluation metrics that have been obtained after the ADASYN was applied. The later figure 5.7 an table number 5.6 shows the confusion matrix of the model after ADASYN has been applied respectively.

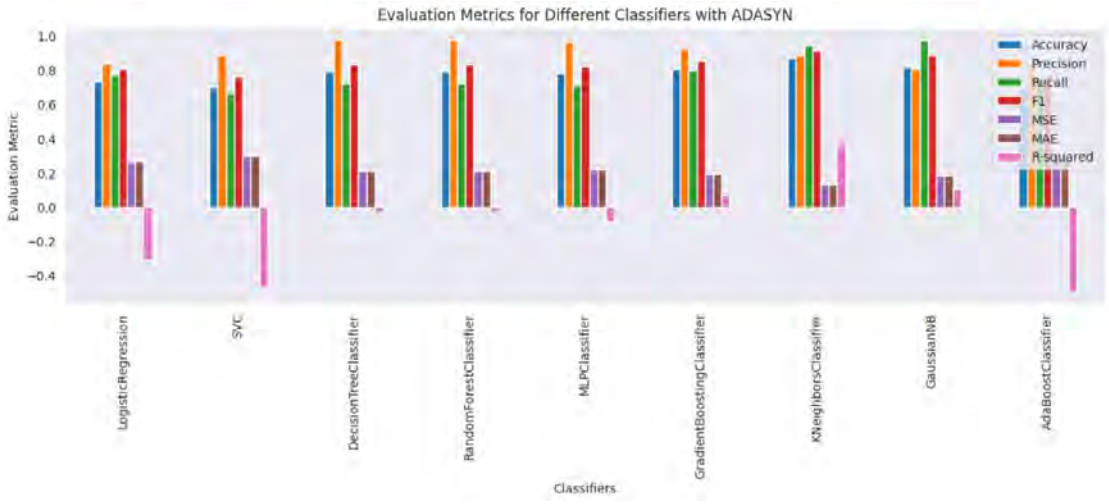

Figure 5.6: Evaluation Metrics of the Classical model after ADASYN

Model	TruePositives	FalsePositives	FalseNegatives	TrueNegatives
Logistic Regression	644	312	507	1485
SVC	781	175	558	1434
DecisionTree Classifier	886	70	418	1574
Random Forest Classifier	886	70	418	1574
MLP Classifier	872	76	432	1560
Gradient Boosting Classifier	799	157	472	1520
K-Neighbors Classifier	765	191	190	1802
GaussianNB	493	463	178	1814
AdaBoost Classifier	704	252	591	1401

Table 5.6: Confusion Matrices for Different Classifiers after ADASYN

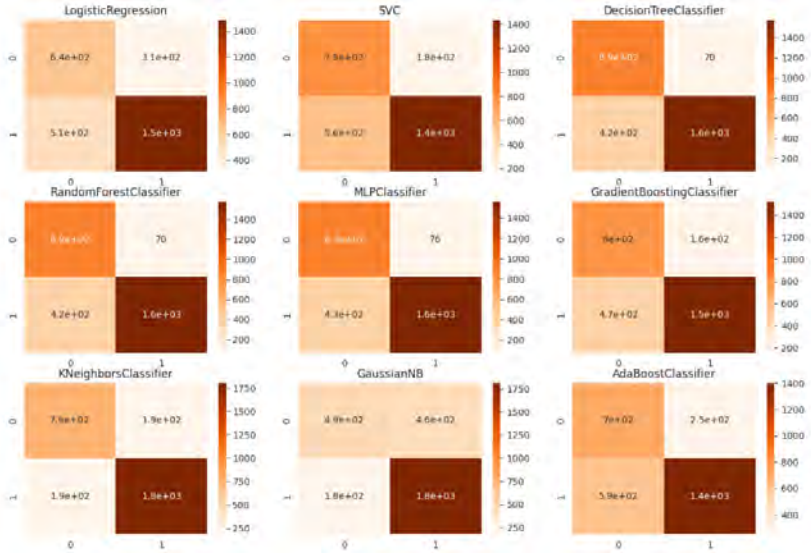

Figure 5.7: Confusion matrix of the Classical model after ADASYN

Classifier	Accuracy	Precision	Recall	F1
LogisticRegression	0.722185	0.826377	0.745482	0.783848
SVC	0.751357	0.891237	0.758618	0.824927
Decision Tree Classifier	0.834464	0.957421	0.790161	0.865787
RandomForestClassifier	0.834464	0.957421	0.790161	0.865787
MLPClassifier	0.827001	0.947464	0.787651	0.860197
Gradient Boosting Classifier	0.786635	0.906380	0.763052	0.828564
K-NeighborsClassifier	0.870760	0.904165	0.904618	0.904615
GaussianNB	0.782564	0.796662	0.910643	0.849848
AdaBoostClassifier	0.714043	0.847550	0.703313	0.768724

Table 5.7: Classifier Performance Metrics ADASYN

Best Classifier: The Table:5.7 K-Nearest Neighbors (KNN) has the highest recall of 0.905 and the bes F1 score of 0.905, which stands him out. It also boasts of the best accuracy score of 0.871, which assures a good balance between precision and recall.

Strong Performers: Gradient Boosting and Gaussian Naive Bayes also exhibit good performance, GaussianNB having superior recall most often, at the risk of having morefalse positives and Gradient boosting performing relatively equally on precision andrecall.

Models to Avoid: AdaBoost and SVC have the least recall and accuracy coupled with the high levels of false negatives thus making these models ineffective for the given problem's skewed classifier's data.

With comparision with SMOTE the K-Nearest Neighbors is reported to deliver good results using both methods and thus it can be concluded that KNN is not adversely affected by class imbalance regardless of the type of application.

The following table 5.8 shows us the comparison of the models:

Metric	No Resampling	SMOTE	ADASYN
Accuracy	0.8439	0.8218	0.7756
Precision	0.8310	0.8055	0.7904
Recall	0.8104	0.7999	0.7674
F1 Score	0.7900	0.7915	0.7663

Table 5.8: Performance Metrics for Different Sampling Techniques

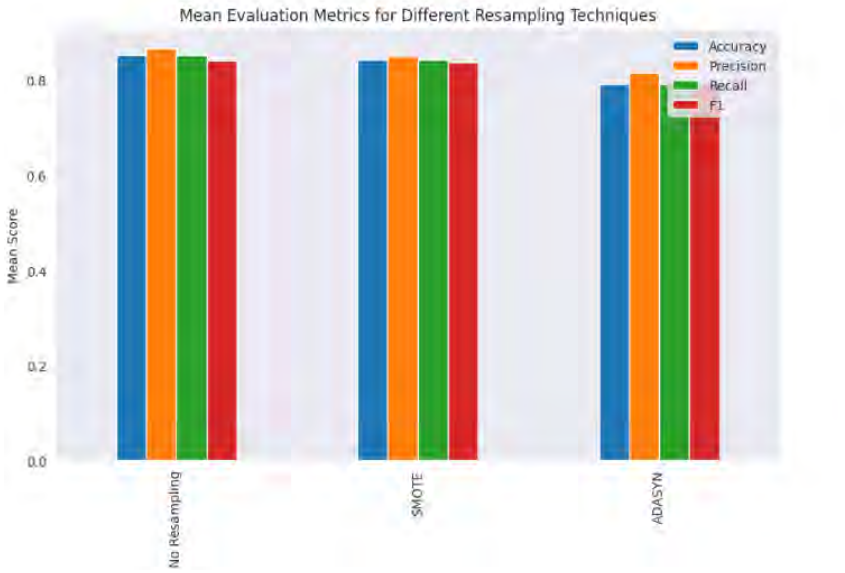

Figure 5.8: Mean Evaluation Metrics of the different Resampling technique

As indicated by the results in table 5.8 and figure 5.8 in of the Accuracy test, there is No Resampling which emphasizes that it is the most accurate as it has the highest accuracy value of 0.8439. It makes the highest number of correct predictions on average. No Resampling comes first followed by SMOTE at 0.8218, while ADASYN is rated last 0.7756 in terms of accuracy.

As indicated by the results of the Accuracy test, there is No Resampling which emphasizes that it is the most accurate as it has the highest accuracy value of 0.8439. It makes the highest number of correct predictions on average. No Resampling comes first followed by SMOTE at 0.8218, while ADASYN is rated last 0.7756 in terms of accuracy.

The No Resampling method scored the highest in precision which was 0.8310. It also depicts that it had the lowest number of false positives when compared to SMOTE and ADASYN. The second position is taken by SMOTE which had a precision level of 0.8055, while ADASYN was slightly lower than this with a level of 0.7904. The concept of recall or sensitivity came out clearly in No Resampling and SMOTE equal results of 0.8104 and 0.7999 respectively. This shows how effective they are in detecting the true positive cases. Again the last winner is ADASYN scoring the lowest.

SMOTE (0.7915) has the best F1 Score. This means that SMOTE has the most appropriate balance between Precision and Recall. This performance is almost similar to that of No Resampling (0.7900) while that of ADASYN (0.7663) is below the average level. It is noticed that The no resampling approach demonstrates the accuracy and precision although its F1 score is slightly lower than that of SMOTE. For priority of achieving overall correctness (accuracy) while minimizing false positives (high precision) this option would be the best. The SMOTE method gives the highest F1 score, this implies that it slightly improves precision and recall. It may be a better strategy to employ in cases where both precision and recall are critical, particularly in skewed datasets.

In all metrics, ADASYN scored the lowest, hence , this one is the least preferred among the three. All in all, one cannot go wrong whether they chose No Resampling or SMOTE for that matter but rather; No Resampling if the priority is on the high level of accuracy and precision. SMOTE if the most critical aspect to consider is the F1 Score (i.e. combination of precision and recall).

5.4 Neural Networks Implementation and Result

The evaluation of loan requests has become an integral aspect of risk management within financial institutions. ML models can be used to minimize the risk of approving loan applicants likely to default or aid in fast-tracking the approval of ideal applicants by understanding previous data. In this particular work, we are interested in analyzing and correlating the performance of five main architectures of neural networks:Recurrent NeuralNetworks (RNN), Long ShortTerm Memory (LSTM), Convolutional NeuralNetworks (CNN), Fully Connected NeuralNetworks (FCNN), and Fully Convolutional NeuralNetworks (FCN) in order to predict loan approval. We describe the models, how they are trained,done for evaluate, and an explanation of the performance results obtained using TensorFlow, Scikit-learn, NumPy and other libraries in Google Colaboratory.

5.5 Implementation and Training Methodology

The dataset consists of various features that influence loan approval decisions, including but not limited to applicant income, credit history, loan amount, etc.n order to evaluate the models, all the data was split into training data (70%) and testing data (30%).
Using the Adam optimizer and binary cross-entropy loss all the models were trained for 100 epochs with a batch size of 32.

5.5.1 Experimental Results

After implmenting our model we have got the results of each model. The performance metrics for each model at epoch 100 are summarized in table 5.9:

Model	Accuracy	Precision	Recall	F1 Score	Confusion Matrix (TP, FP, FN, TN)
Simple RNN	0.8504	0.8552	0.9372	0.8944	(1867, 316, 125, 640)
LSTM	0.8558	0.8429	0.9669	0.9006	(1926, 359, 66, 597)
CNN	0.8823	0.8831	0.9518	0.9162	(1896, 251, 96, 705)
FCNN	0.8803	0.8796	0.9533	0.9150	(1899, 260, 93, 696)
FCN	0.8975	0.8634	0.9804	0.9182	(1953, 309, 39, 647)

Table 5.9: Performance comparison of different neural models

Model Loss and Validation Accuracy

In addition to performance metrics, the training and validation results at epoch 100 provide further insight into each model's behavior. The table 5.10 shows us the Training and Validation Loss compared to the Validation Accuracy of the Diffeent Neutral models used in the project

Model	Training Loss	Validation Loss	Validation Accuracy
Simple RNN	0.3267	0.3328	0.8504
LSTM	0.3130	0.3123	0.8558
CNN	0.2583	0.2450	0.8823
FCNN	0.2527	0.2483	0.8803
FCN	0.2616	0.2255	0.8975

Table 5.10: Training and Validation Loss with Validation Accuracy for Different Models

The findings and visulaizations of our results from figure 5.9 to figure 5.18 it shows the Accuracy and Loss of the implemented neural models. Moreover, from figure 5.19 to figure 5.23 it shows the confusion matrixes of the models

RNN

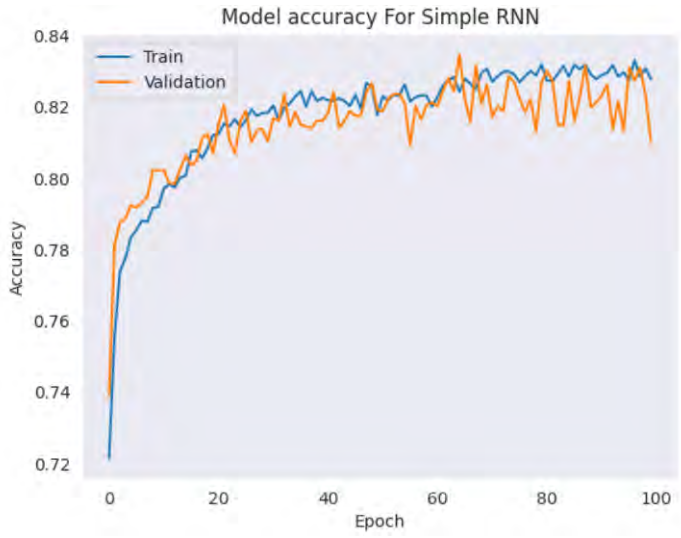

Figure 5.9: RNN Accuracy

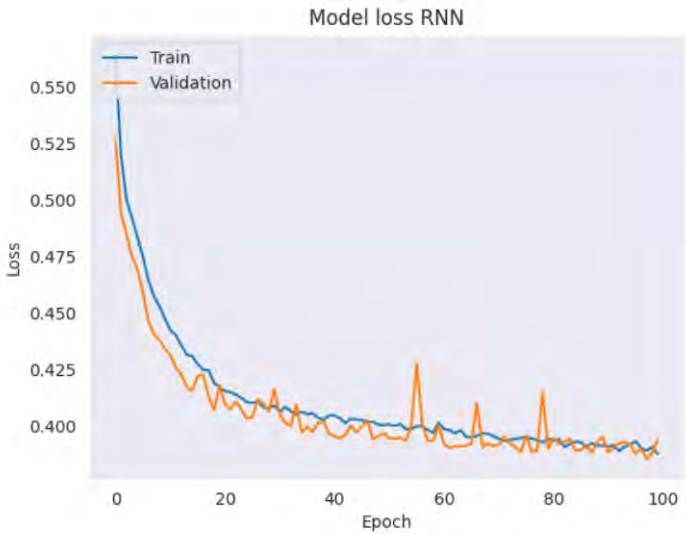

Figure 5.10: RNN Loss

LSTM

Figure 5.11: LSTM Accuracy

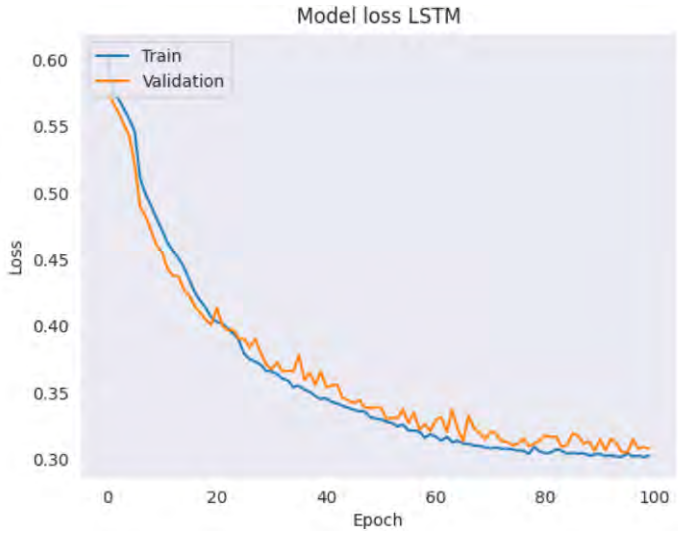

Figure 5.12: LSTM Loss

Convolutioal Neural Networks

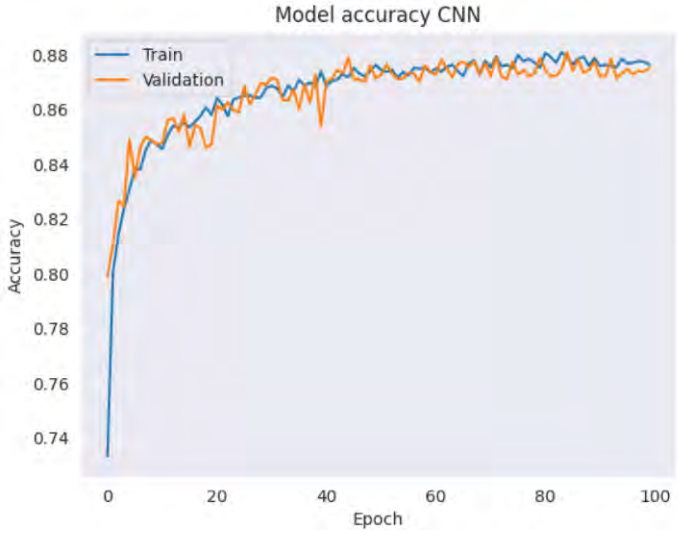

Figure 5.13: CNN Accuracy

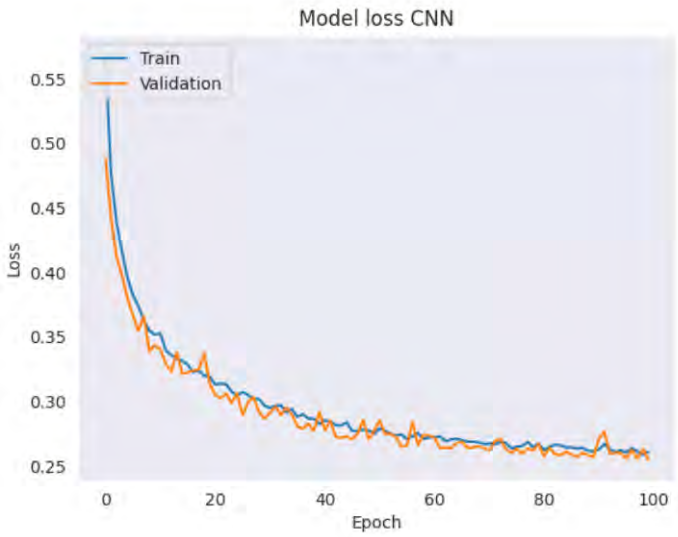

Figure 5.14: CNN Loss

Fully Connected Neural Networks

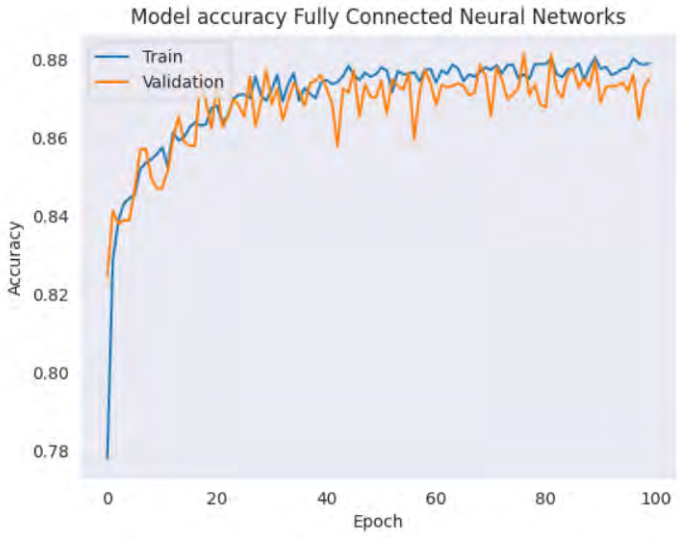

Figure 5.15: FCNN Accuracy

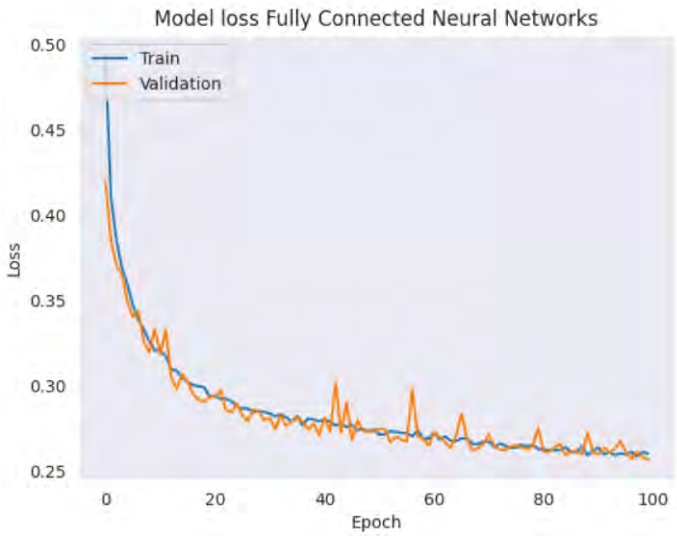

Figure 5.16: FCNN Loss

Fully Convolutional Neural Networks

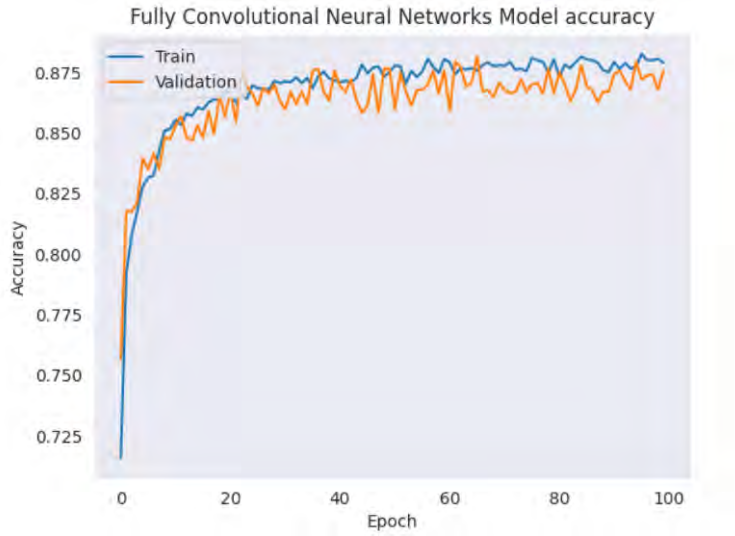

Figure 5.17: FCN Accuracy

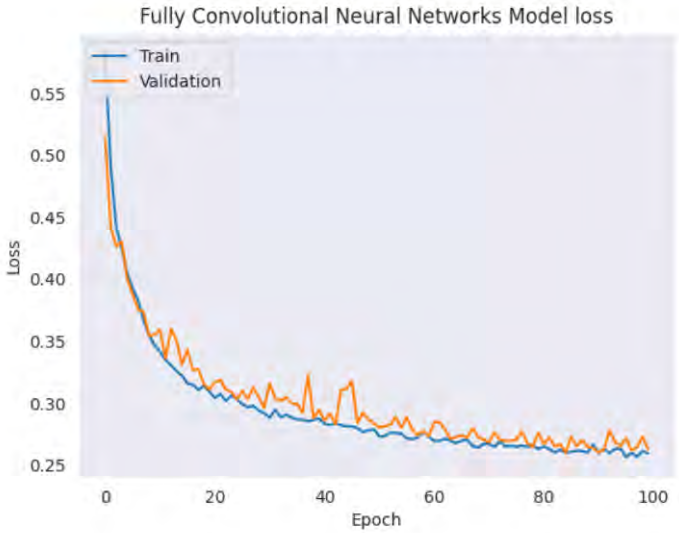

Figure 5.18: FCN Loss

Confusion Matrix For Neural Networks models

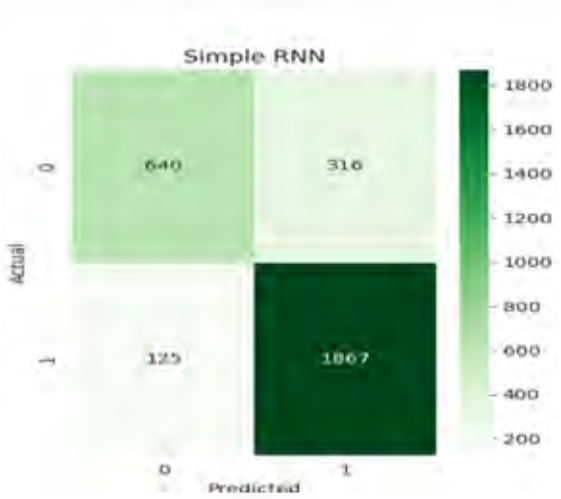

Figure 5.19: Confusion matrix for RNN

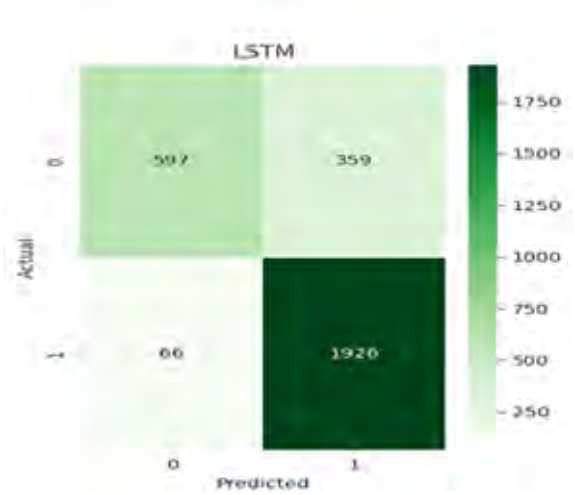

Figure 5.20: Confusion matrix for LTSM

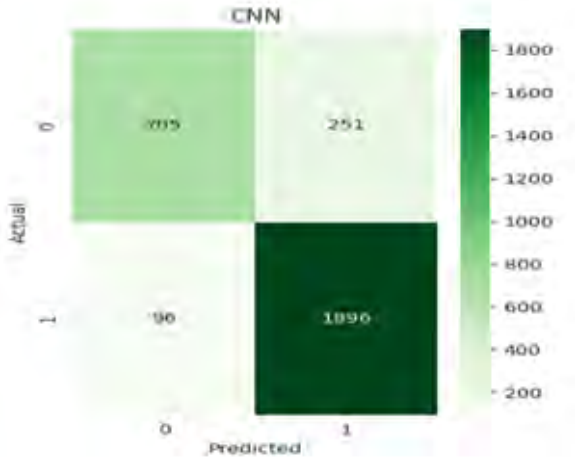

Figure 5.21: Confusion matrix for CNN

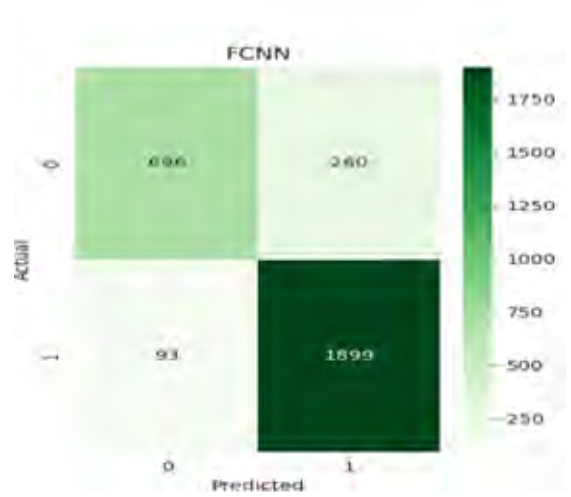

Figure 5.22: Confusion matrix for FCNN

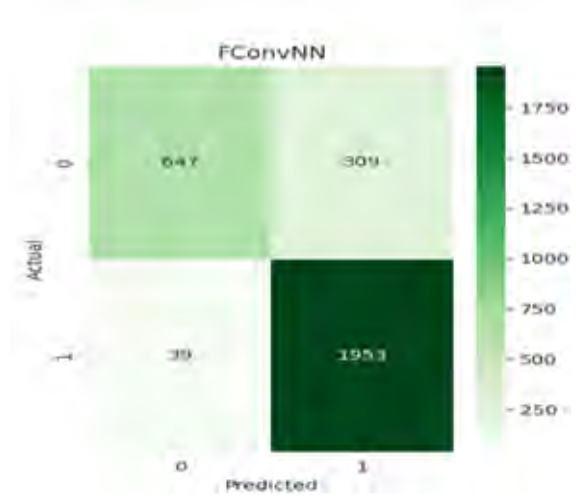

Figure 5.23: Confusion matrix for FCN

5.6 Comparative analysis

5.6.1 Simple RNN

Performance : The Simple RNN performs relatively well with its accuracy at 0.8504, very high recall (0.9372) and F1-score (0.8944), but moderate precision (0.8552) indicates that this model suffers from many false positives compared to ever models.

Loss and Generalization: The model has also proven its also unsatisfactory for deployment due to the very high validation loss, which is paused at about 0.3328.

5.6.2 LSTM

Performance: LSTM has also demonstrated a slight improvement in accuracy scores (0.8558) and recall scores (0.9669) as opposed to Simple RNN indicating the ability of the model to adapt to long term features.

Loss and Generalization: The validation loss of the model is also lower than that of Simple RNN (0.3123), which means it generalizes more effective. However, precision (0.8429) is lower than that of both CNN and FCN leading to more false positives.

5.6.3 Convolutional Neural Network (CNN)

Performance: CNN posts the second highest accuracy and impressive balance of precision (0.8831), recall (0.9518), and F1 scores (0.9162) with the accuracy level being 0.8823. its confusion matrix indicates a healthy range of true positive values and very few false positively(251)and false negative values (96)

Loss Generalization: Although the CNN model yields a higher validation loss (0.2450) than, that indicates a lower generalization than LSTM, Hell gatt is suitable to the loan approval task since its validation accuracy is high at 0.8823.

5.6.4 Fully Connected Neural Network (FCNN)

Performance: FCNN achieves the highest training accuracy (0.8832), though its validation accuracy (0.8803) is slightly lower than CNN. Its precision (0.8796), recall (0.9533), and F1-score (0.9150) are all competitive, although CNN marginally outperforms FCNN in these areas.

Loss and Generalization: Although its validation loss is low (0.2483) implying the network will generalize correctly, the FCNN model does not defeat the CNN model on generalization capability or accuracy.

5.6.5 Fully Convolutional Neural Network (FCN)

Performance: FCN is rated as the best best-suited method, achieving the hyghest validation accuracy (0.8975) and recall (0.9804) and an equally high F1-score (0.9182). The matrix of confusion shows that the fewest false negatives happened (39) indicating excellent predictive capability of the model.

Loss and Generalization: It is also worth noting that FCN has the least validation loss among the models, which is 0.2255, indicating that it is better than other models in predicting unseen data. This further reinforces the view that the FCN is the most complex model which works well in offering balanced precision and recall characteristics whilst reducing errors.

5.6.6 Trade-Offs

RNN: Recorded lower precision and F1-scores when compared to advanced architectures such as LSTM and CNN.

LSTM: Provided better recall which is an advantage in use cases where false negations can be expensive.

CNN and FCN: These models return the best results for all the metrics, indicating that the convolutional layers are useful for this kind of prediction task.

Efficiency in Time and Resource:
In order to the predictive accuracy, the time and resources for computation that was entailed in training each model dynamic was also noted:

Simple RNN: This RNN model is moderate to light in weight allowing for quick training but is hampered by the vanishing gradient effect which in turn results to low performance.

LSTM: This model is advantageous as it can hold longer dependencies but the amount of computation and time needed for training is highly excessive when compared with non-complex structures like RNN or CNN.

CNN and FCNN: These models have a lower training time than LSTM, especially with datasets that exhibit little time variation. The way CNN is structured allows it to be highly optimised and trained within a short span of time without compromising on the predictive results.

FCN: If the performance is to be maximized, the fully convolutional approach should be employed. However, such a model is likely to be more computations intensive than simpler alternatives. It is designed to grasp both local and global context so it does tend to use more resources.

Therefore, the results of this study indicate that Fully Convolutional Neural Networks (FCN) provide the best performance for loan approval prediction. FCN's high validation accuracy and strong generalization capabilities make it an ideal choice for deployment in financial institutions. CNN also shows excellent promise, providing a balance between performance and efficiency.

5.7 Preview of the System

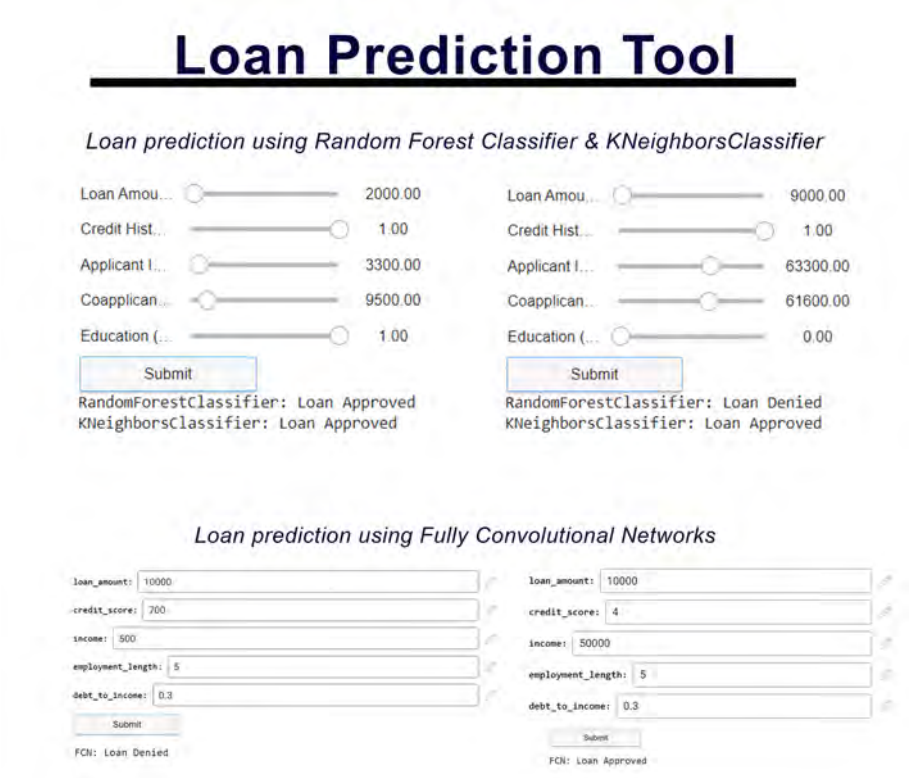

Figure 5.24: The Loan Prediction Tool

Brief Descrpition:The Loan Prediction Tool employs RandomForest (RF) and K-NearestNeighbors(KNN) classifiers among others Fully ConvolutionalNetworks (FCNs) to anticipate the approval status of loans based on the input given by the user.

Chapter 6

Conclusion and Future Works

6.1 Conclusion

This research highlights the importance of employing machine learning classifiers for the prediction of loan approval status which will be beneficial for financial institutions aiming at improving their processes of making decisions. Our extensive analysis demonstrates that suitable data pre-processing and feature selection play critical roles in attaining the best performance of the model. As a result, the Random Forest Classifier has also been shown to be the most effective method with high accuracy and F1 score values while striking a good balance of precision and recall. Its strength makes it a good candidate for finance-related applications in the real world. Also, the tactical use of SMOTE and ADASYN proved to reduce the problem of class imbalance to a greater extent, enabling classifiers such as K-Nearest Neighbors to still perform satisfactorily. The set of evaluation metrics that were applied distinguished the capabilities each model hold for and against each other and this sets a basis for further explorations in this area.

Out of the five types of neural networks analyzed, the architecture of Fully Convolutional Network (FCN) had the best performance in prediction at 89.75% validation accuracy, 0.9804 recall, and 0.2255 being the least validation loss. Hence, this makes FCN the best option for use considering real-time occurrences when accuracy of decisions is highly sought. Both CNN and FCNN showed good performance but the comparative study also showed important benefits and drawbacks of precision and recall against time taken to run the model.

On the other hand, the less sophisticated architectures such as Simple RNNs and LSTMs, although useful in some aspects, did not perform the best as compared to the sophisticated models. This suggests that there is still much room for development and improvement of machine learning applications in the financial space. Further studies may focus on the application of ensemble techniques, more extensive hyperparameter tuning, and these models extending their scope to propensities predicting loan approval among other factors to improve the prediction and generalization in approval predictions.

6.2 Future Works

The increasing application of machine learning and deep learning technologies in ascertaining loan default risk in the financial sector also presents certain hurdles. These barriers should be surmounted to increase the performance and the trustworthiness of the predictive models and also create a better business environment in the given industries. Some of the issues we have encountered as well as strategies that can be applied in the coming years are outlined below.

6.2.1 Quality and Availability of Data

Problem: It is often said that the success of any prediction model is mainly determined by the amount and the quality of data available. Most of the times for eg: banks and other financial institutions face problems as there exists data which are incomplete, irrelevant or inconsistent. One such case is when an individual's credit history is not properly documented because he possesses one from elsewhere and this depicts a futile profile that is of no assistance in deductions made.

Solution: In the future, it is advisable that the research should be directed toward the problem of constructing operative data gathering and data purification methods. More data augmentation methods can also be used to add in more synthetic data for the occassional challenges so that the model performs well without overfitting. Gaining access to such datasets may entail establishing joint ventures with certain consumer-focused financial institutions in order to assist in bettering the model with historical records of applicants and the applications dealt with [41].

6.2.2 Interpretability of Models

Problem: Many modern techniques, deep learning networks for instance, can be regarded as black-boxes by the stakeholders since people cannot see how a prediction is made. In the case of loan approvals, such decisions have effects on people's lives; thus, if a model cannot be explained, the applicants and even the regulators will not be able to trust orthe system.

Solution: Emphasis must be placed on the techniques of Explainable Artificial Intelligence (XAI), which means that an effort to shed light on and clarify the decision reached out of the model is made available. There are, for instance, methods known as LIME (Local Interpretable Model- Agnostic Explanations) and SHAP (SHapley Additive exPlanations) which aid in determining the role played by every feature towards the prediction of a single example. This approach will boost the confidence of the users in the modern technological aided decision making methods as they will appreciate the logic behind the decision taken [42] .

6.2.3 Bias and Fairness

Problem: Predictive models are often influenced by prejudice in the training datasets; this, thus, leads to biased results when discriminating loan applicants. For instance, certain demographic sections may have been favored in the age-old practices of offering credits which will end up desiring models that are biased.

Solution: There lies a very important aspect in ensuring the equity of completion of model at the prediction execution phase of the prometed works. In the future, studies may work on how to detect biases in outcomes and how to reduce or eliminate them, if they exist. This include, for instance, employing techniques for context-sensitive adversarial training, or applying fairness criteria during fitted model training so that predictions do not discriminate against any group based on the group defined characteristics [43].

6.2.4 Harmonization with Current Processes

Problem: Outfitting advanced machine learning models into the current financial systems proves difficult due to the likes of legacy systems which may be incompatible with new technologies. Such integration procedures can be quite time-consuming and costly.

Solution: Additional research may explore the design of modular and deployable structures enabling smoother embedding of the prediction models into the current processes. The implementation of the cloud solutions is favorable to the incorporation of machine learning techniques in the existing systems with minimal changes to the legacy systems[44].

Bibliography

[1] Kumar Arun, Garg Ishan, Kaur Sanmeet(2016). Loan Approval Prediction based on Machine Learning Approach. IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. I (May-Jun. 2016), PP 79-81 www.iosrjournals.org

[2] Pidikiti Supriya , Myneedi Pavani , Nagarapu Saisushma, Namburi Vimala Kumari , K Vikas(2019). Loan Prediction by using Machine Learning Models. International Journal of Engineering and Techniques - Volume 5 Issue 2, Mar-Apr 2019

[3] Lai, L. (2020). Loan Default Prediction with Machine Learning Techniques. 2020 International Conference on Computer Communication and Network Security (CCNS).

[4] Mohammad Ahmad Sheikh; Amit Kumar Goel; Tapas Kumar(2020) An Approach for Prediction of Loan Approval using Machine Learning Algorithm

[5] AFRAH KHAN, EAKANSH BHADOLA, ABHISHEK KUMAR and NIDHI SINGH(2021) LOAN APPROVAL PREDICTION MODEL A COMPARATIVE ANALYSIS

[6] HV Ramachandra; G Balaraju; R Divyashree; Harish Patil(2021) Design and Simulation of Loan Approval Prediction Model using AWS Platform

[7] Ashlesha Vaidya(2017)Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval

[8] Mohammad Ahmad Sheikh; Amit Kumar Goel; Tapas Kumar(2020) An Approach for Prediction of Loan Approval using Machine Learning Algorithm

[9] Jeremy D. Turiel, Tomaso Aste, P2P LOAN ACCEPTANCE AND DEFAULT PREDICTION WITH ARTIFICIAL INTELLIGENCE

[10] Mayank Anand, Arun Velu, Pawan Whig,Prediction of Loan Behaviour with Machine Learning Models for Secure Banking

[11] S. Vimala, K.C. Sharmili, —Prediction of Loan Risk using NB and Support Vector Machine‖, International Conference on Advancements in Computing Technologies (ICACT 2018), vol. 4, no. 2, pp. 110-113, 2018.

[12] Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma, Namburi Vimala Kumari, Vikash,"Loan Prediction by using Machine Learning Models", InternationalJournalofEngineering andTechniques.Volume 5 Issue 2, Mar-Apr 2019

[13] Nikhil Madane, Siddharth Nanda,"Loan Prediction using Decision tree", Journal of the Gujarat Research History, Volume 21 Issue 14s, December 2019.

[14] Rutika Pramod Kathe, Sakshi Dattatray Panhale, Pooja Prakash Avhad, Punam Laxman Dapse, Ghorpade Dinesh B. Prediction Of Loan Approval Using Machine.Learning Algorithm: A Review Paper. International Journal Of Creative Research Thoughts(IJCRT).

[15] S. Sobana, P. Jasmine Lois Ebenezer. A COMPARATIVE STUDY ON MACHINE LEARNING ALGORITHMS FOR LOAN APPROVAL PREDICTION ANALYSIS.International Research Journal of Modernization in Engineering Technology and Science.

[16] Mengnan.Song,Jiasong.Wang,Tongtong.Zhang, Guoguang.Zhang,Ruijun.Zhang, Suisui. Su. Effective Automated Feature Derivation via Reinforcement Learning for Microcredit Default Prediction.

[17] Arunkumar, G. A., Panchuram, C. R., Afzal, K. M. A., Yadav, N. S., Goradiya, U. (2023). Predictive Analysis in Banking using Machine Learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Volume 10 (Issue 2), Page Number 434-439. DOI: https://doi.org/10.32628/CSEIT2390247

[18] Singh, V., Yadav, A., Awasthi, R. (2021). Prediction of modernized loan approval system based on machine learning approach. In Proceedings of the 2021 International Conference on Intelligent Technologies (CONIT) (pp. 1-6). Institute of Electrical and Electronics Engineers (IEEE). DOI: https://doi.org/10.1109/CONIT51480.2021.9498475

[19] M.A. Sheikh, A.K. Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm", 2020 International Conference On Electronics and Sustainable Communication Systems (ICESC), pp. 490-494, 2020. DOI: https://doi.org/10.1109/ICESC48915.2020.9155614

[20] Sivasree M S ,Rekha Sunny T(2015), Loan Credibility Prediction System Based on Decision Tree Algorithm, Volume 4, Issue 9

[21] Karthikeyan S.M, Pushpa Ravikumar(2021), A Comparative Analysis of Feature Selection for Loan Prediction Model, Volume 174, No. 11

[22] Ladislav Végh , Krisztina Czakóová and Ondrej Takáč(2023), Comparing Machine Learning Classification Models on a Loan Approval Prediction Dataset, Volume 7, pp. 98-103

[23] Subhiksha, Vaishnavi, Shalini, Mr. N. Manikandan(2022) Bank Loan Approval Prediction Using Data Science Technique (ML)

[24] Shruti Mishra, Shailki Sharma and Shreyansh Singh(2022) Loan approval prediction

[25] Deepak Ishwar Gouda, Ashok Kumar , Anil Manjunatha Madivala, Dilip Kumar, Dr.Ravikumar(2021) LOAN APPROVAL PREDICTION BASED ON MACHINE LEARNING

[26] Mudit Manish Agarwal, Harshal Mahendra Shirke, Vivek Prafullbhai Vadhiya, Manya Gidwani,'Loan Analysis Predicting Defaulters',2022 Volume 9, Issue 4

[27] Kshitiz Gautam, Arun Pratap Singh, Keshav Tyagi, Mr.Suresh Kumar,'Loan Prediction Using Decision Tree and Random Forest' 2020 Volume 7, Issue 8

[28] Awuza Abdulrashid Egwa,Habeebah Adamu Kakudi, Ahmad Ajiya Ahmad, Abubakar Muhammad Bichi, Muhammad Alhaji Madu,'Prediction Model for Loan default using Machine Learning', February 2022, Volume 10, Issue 2

[29] Lundberg, S. M., Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Machine Learning (ICML) (pp. 477-487). (Discusses SHAP values for interpretability of various models)

[30] Darrell, T., O'Gorman, K. D. (2020). Supervised learning without explanations: Can it be ethical? In Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency (pp. 30-40).

[31] Hardt, M., Price, E., Srebro, N. (2016, June). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems (pp. 3315-3323).

[32] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1), 1929-1958.

[33] The dataset we have collected

[34] Joshi, R. D., Dhakal, C. K. (2021). Predicting Type 2 diabetes using logistic regression and machine learning approaches. International Journal of Environmental Research and Public Health/International Journal of Environmental Research and Public Health, 18(14), 7346. https://doi.org/10.3390/ijerph18147346

[35] Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., Wong, T. Y., Cheng, C. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. Journal of Clinical Epidemiology, 122, 56–69. https://doi.org/10.1016/j.jclinepi.2020.03.002

[36] Sufriyana, H., Husnayain, A., Chen, Y. L., Kuo, C. Y., Singh, O., Yeh, T. Y., Wu, Y. W., Su, E. C. Y. (2020). Comparison of multivariable logistic regression and other machine learning algorithms for prognostic prediction studies in Pregnancy Care: Systematic Review and Meta-Analysis. JMIR Medical Informatics, 8(11), e16503. https://doi.org/10.2196/16503

[37] Charbuty, B., Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(01), 20–28. https://doi.org/10.38094/jastt20165

[38] Ying CAO , Qi-Guang MIAO , Jia-Chen LIU , Lin GAO. (June 2013). Acta Automatica Sinica Volume 39, Issue 6, Pages 745-758. https://doi.org/10.1016/S1874-1029(13)60052-X

[39] Li Lang, Liang Tiancai, Ai Shan, Tang Xiangyan (May 2021). An improved random forest algorithm and its application to wind pressure prediction Volume 37, Issue 2 , pages: 1802-1802. https://doi.org/10.1002/int.22448

[40] Amit Pandey, Achin Jain,"Comparative Analysis of KNN Algorithm using VariousNormalization Techniques", International Journal of Computer Network and Information Security(IJCNIS), Vol.9,No.11, pp.36-42, 2017.DOI: 10.5815ijcnis.2017.11.04

[41] Chan, K. Y., Abu-Salih, B., Qaddoura, R., Al-Zoubi, A. M., Palade, V., Pham, D., Del Ser, J., Muhammad, K. (2023). Deep neural networks in the cloud: Review, applications, challenges and research directions. Neurocomputing, 545, 126327. https://doi.org/10.1016/j.neucom.2023.126327

[42] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory"; "Neural Computation", vol. 9, no. 8, pp. 1735-1780, 1997.

[43] F. A. Gers, J. Schmidhuber, and F. Cummins, quot;Learning to Forget: Continual Prediction with LSTM.quot;

[44] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning"

[45] D. P. Kingma and J. Ba,"Adam: A method for stochastic optimization";

[46] C. M. Bishop, 'Pattern Recognition and Machine Learning';

[47] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning"; , MIT Press, 2016.

[48] Haykin, S. (2009). Neural Networks and Learning Machines (3rd ed.). Prentice Hall.

[49] Long, J., Shelhamer, E., amp; Darrell, T. (2015). "Fully convolutional networks for semantic segmentation". "Proceedings of the IEEE Conference on Computer Vision and Pattern" Recognition, 3431-3440.

[50] Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., Wong, T. Y., Cheng, C. (2020). "Logistic regression was as good as machine learning for predicting major chronic diseases". Journal of Clinical Epidemiology, 122, 56–69. https://doi.org/10.1016/j.jclinepi.2020.03.002

[51] Kashishdafe. (2024b, March 23). Gaussian Naive Bayes: Understanding the basics and applications. Medium. https://medium.com/@kashishdafe0410/gaussian-naive-bayes-understanding-the-basics-and-applications-52098087b963

[52] Charbuty, B., Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(01), 20–28. https://doi.org/10.38094/jastt20165

[53] Shanmugavadivel, K., S, M. D. M., R, M. T., Al-Shehari, T., Alsadhan, N. A., Yimer, T. E. (2024). Optimized polycystic ovarian disease prognosis and classification using AI based computational approaches on multi-modality data. BMC Medical Informatics and Decision Making, 24(1). https://doi.org/10.1186/s12911-024-02688-9

[54] Natekin, A., Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in Neurorobotics, 7. https://doi.org/10.3389/fnbot.2013.00021

[55] Kashishdafe. (2024b, March 23). Gaussian Naive Bayes: Understanding the basics and applications. Medium. https://medium.com/@kashishdafe0410/gaussian-naive-bayes-understanding-the-basics-and-applications-52098087b963