ISSN: 2249-7196
IJMRR/April. 2025/ Volume 15/Issue 2s/42-50
Kadali Anusha / International Journal of Management Research & Review

Predicting Bank Loan Defaults

Kadali Anusha
PG scholar, Department of MCA, DNR College, Bhimavaram, Andhra Pradesh.
A.Durga Devi
(Assistant Professor), Master of Computer Applications, DNR college, Bhimavaram, Andhra Pradesh.

Abstract: In our everyday lives, loan lending plays an important role. It powerfully promotes the economy and the growth of consumption. A loan carries risk, is inevitable, and may result in a financial crisis. As a result, determining whether a person is eligible for the loan is critical. In this research, we use the XGBoost and Random Forest techniques to train the prediction model and compare their accuracy. We use the variance inflation factor and variance threshold approaches in the feature engineering section to filter out unnecessary features before feeding them into XGBoost and Random Forest. In loan default scenarios, there is no difference in prediction accuracy between XGBoost and Random Forest because both attain a high accuracy of roughly 0.9.

Keywords: Prediction Model, Random Forest, Machine Learning, Loan Default

I. INTRODUCTION

The aim of predicting bank loan defaults is to assess the likelihood that a borrower will fail to repay a loan, leading to a default on the loan obligation. As a result, financial institutions can manage their risks and make informed lending decisions.

Our objective is to predict the loan default to build a machine learning-based solution of customers in advance based on features such as term duration, the purpose of the loan, debt-to-income ratio, employment status, income, etc. The objectives of predicting bank loan defaults revolve around managing risk, making informed lending decisions, and maintaining the financial stability of the lending institution.

Predicting bank loan defaults is a critical endeavor in the realm of finance, where the intricate dance between risk and reward shapes the foundation of lending practices. As financial institutions extend loans to individuals and businesses, the potential for borrowers to default on their repayment obligations presents an inherent challenge. This challenge, however, has sparked The development of sophisticated predictive models that harness the power of data and analytics to peer into the future with unprecedented clarity.

In our everyday lives, loan lending plays an important role. It powerfully promotes the economy and the growth of consumption. [1]. For people, taking a loan has been inevitable because, to overcome financial constraints, individuals worldwide depend on loans, and organizations rely on loans to expand their production and achieve their personal goals [2].

Loan lending usually benefits both the lenders and the borrowers. As a result, determining whether a candidate is eligible to receive a loan is critical. Previously, the evaluation was mostly based on manual review, which was labour-intensive and time-consuming [3]. Recently, to automatically predict loan defaults, banks have opted for machine learning approaches since they can greatly enhance the efficiency and accuracy of the predictions.

On the one hand, because of the popularity of mobile payments and online shopping, banks may collect vast amounts of transaction data. Machine learning models are rapidly developing and offer various beneficial applications for predicting loan default, prompting the banking industry to use them. In several loan lending situations, researchers discovered that Random Forest outperformed other models such as support vector machines, decision trees, and logistic regression [4]. We will also apply XGBoost to make a comprehensive comparison to predict loan default because, for machine learning, XGBoost is one of the most advanced methods developed in recent years [5].

In this era of data-driven decision-making, the ability to foresee loan defaults holds immense significance for banks. The repercussions of defaulted loans reverberate through the financial ecosystem, causing not only financial losses for the lender but also unsettling waves across economic stability. The art of predicting loan defaults is, in essence, a pursuit of understanding patterns, deciphering indicators, and untangling the intricate web of factors that contribute to a borrower's creditworthiness.

II. LITERATURE SURVEY

For banks, one of the major income sources is the loan business. For business loans, default problems are a major issue. With the development of machine learning techniques and the introduction of the big data era, we now have more options than manual processing for predicting and classifying loan defaults. Using a real-world dataset from a famous multinational bank, we show that the AdaBoost model can forecast loan default with 100% accuracy, surpassing alternative models such as multilayer perceptrons, k closest neighbours, XGBoost, and random forest. Our findings demonstrate the potential of machine learning techniques in the banking industry. [1]

In the financial world, loan lending has played a significant role. Though advantageous and profitable for borrowers and lenders, it entails a high credit risk in the loan lending sector. Researchers and industry experts award credit ratings to people worldwide to evaluate risk and creditworthiness. Machine learning algorithms have been used for years to forecast and assess credit risk by examining an individual's past information. This paper examines the current literature on risk prediction models that employ machine learning techniques. [2]

A significant amount of the population has recently looked for bank loans as a result of the growing banking industry and the rising trend of borrowing. For banking authorities, the rising rate of loan defaults makes it challenging. For banks, it is necessary to examine the loan request and address the risks of people defaulting on loans, which is also a major challenge in this dynamic economy. In light of the problems raised, this paper suggests two machine learning models to predict, by evaluating specific attributes, whether an individual should be granted a loan. This will assist the banking authorities by making choosing suitable individuals from a given list of candidates who applied for a loan easier. This study compares and thoroughly analyzes the two algorithms, Decision Tree and RF. [3]

Social lending has become a viable platform through the development of social media and electronic commerce, where borrowers and lenders can conduct business without institutional intermediaries. Recently, social lending has accelerated significantly, with certain platforms quickly achieving multi-billion dollar loan circulation. On the other hand, dependable risk attribution to specific borrowers is critical for such platforms' long-term viability and potential general adoption. For forecasting borrower status we propose a random forest (RF) based classification algorithm. Our findings on information from the well-known social lending platform Lending Club (LC) show that the RF-based strategy better identifies quality borrowers than FICO credit scores and LC grades.[4]

In the 21st century, the Internet, the financial sector, and big data have tremendously developed. The national attention paid to this field has gradually improved as well. For the traditional financial industry, a powerful complement is peer-to-peer (P2P), an innovative borrowing mode. The expected credit default rate is an absolute requirement for ensuring the proper operation of associated financial platforms or initiatives. In this study, we apply a multi-dimensional and multi-observation data cleaning method and XGboost, based on real Lending Club P2P transaction data and in Asia at the end of 2016 the modern machine learning algorithms LightGBM. The platform's default risk is anticipated innovatively and robustly. [5]

To present a Credit Scoring Model is the purpose of this article is utilized by a Microfinance Institution in Herzegovina and Bosnia, and to show how the most relevant criteria for its execution were chosen. To predict default clients the goal of

the developed Credit Scoring Model and reduce the credit risk of Microfinance Institutions by using a data mining algorithm to find patterns for default client recognition and, thus, improve the loan approval decision-making process. The developed model demonstrated high predicted accuracy and confidence and reliable outcomes in feature selection; therefore, the Microfinance institution opted to use it as a decision-making aid. [6]

III. PROPOSED METHOD

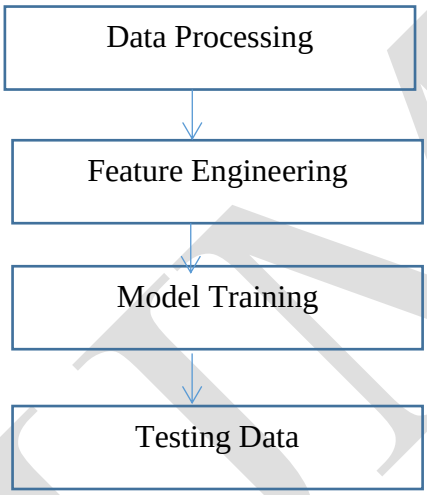


Figure 1. Flowchart

3.1 Data processing:

The dataset from Imperial College London has 105,471 records with 771 columns, 778 characteristics, record loss, and customer ids. The dataset's NAs are then cleaned. NAs are present in 525 columns.

3.2 Feature Engineering

When creating an efficient model, feature selection seeks to eliminate duplicated columns that provide little valuable information and minimize the amount of input characteristics.

First, since such columns don't include useful information for categorization, we rapidly filter them out using the variance threshold method. 760 columns are left after using the variance threshold approach.

To lessen multicollinearity, we employ the Variance Inflation Factor approach. The standard errors of the coefficients are unnecessarily inflated by multicollinearity, and higher common errors suggest that the coefficients of some features may be close to zero, rendering some features irrelevant when they should be significant. The VIF will be 1 if there are no linked features. The VIF indicates how badly the regression coefficients are calculated due to multicollinearity if it is more than 10 [13]. This study excludes all characteristics with a VIF larger than 10.

There are 419 columns left after features are filtered using the variance threshold approach and the VIF.

3.3 Model Training & Testing Data

This study examines the performance of the XGBoost and Random Forest methods for model training and prediction accuracy. According to empirical studies, we should use between 70% and 80% of the data for training and between 20% and 30% for testing to get the best results [14]. So, we split the dataset into two pieces at random. The first section comprises the training dataset, which contains 80% of the data, and the testing dataset, includes the 20%.

3.4 The Prediction Accuracy

0.90635 is the prediction accuracy of the XGBoost model while the prediction accuracy of the Random Forest model is 0.90657.

IV. RESULT

In this project, we have added all given algorithms, and all possible FEATURES engineering concepts and we have coded this

project in a JUPYTER notebook. Below are the output screens,

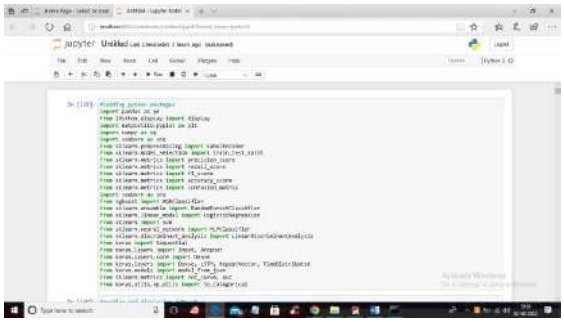


Fig.2 Jupyter Notebook

In above screen loading all python packages

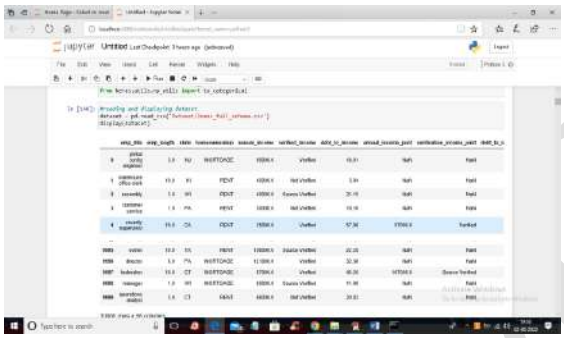


Fig.3 Dataset

In the above screen reading and displaying dataset values, we can see that it contains both non-numerical and numeric values, so by using the ENCODING algorithm and we need to convert non-numerical into numeric.

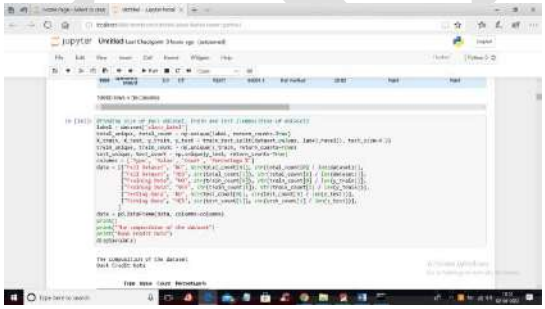


Fig.4 calculating dataset composition

In above screen we are calculating dataset composition and below is the composition output

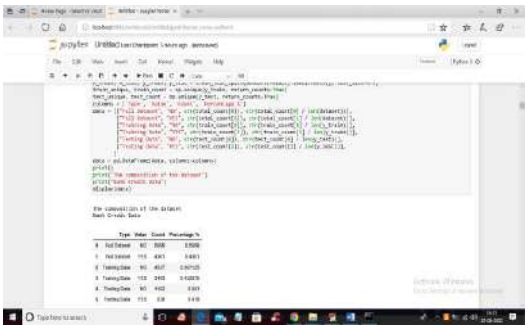


Fig. 5 composition output

In above screen you can see dataset composition and in below screen we are displaying USED VARIABLES NAMES

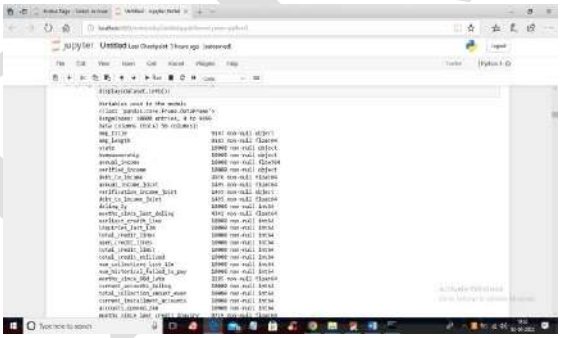


Fig.6 Variable names

In above screen displaying all variables names

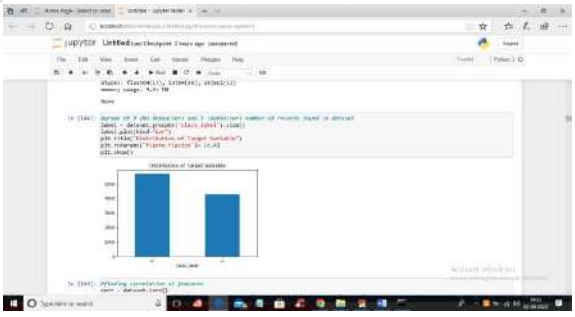


Fig.7 Total Records

In above graph we are finding total records of Non-Defaulter (0) and Defaulters (1).

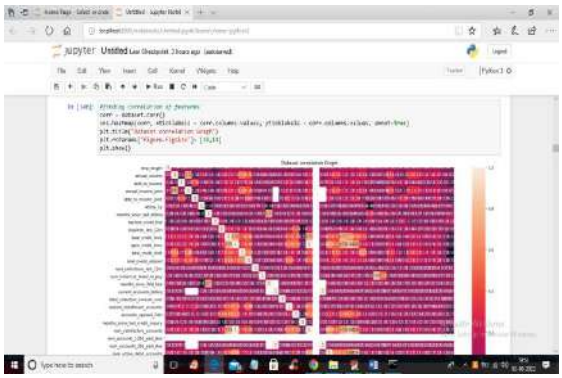


Fig.8 correlation of dataset features

In above screen we are finding correlation of dataset features

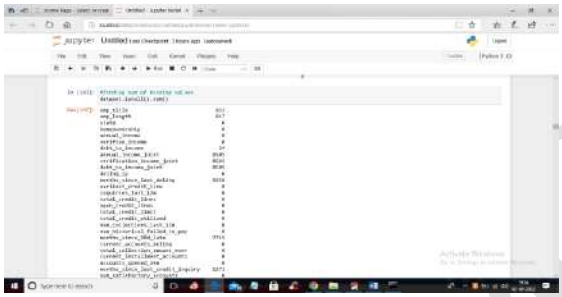


Fig.9 finding sum of missing values

In above screen we are finding SUM of all missing values for each column

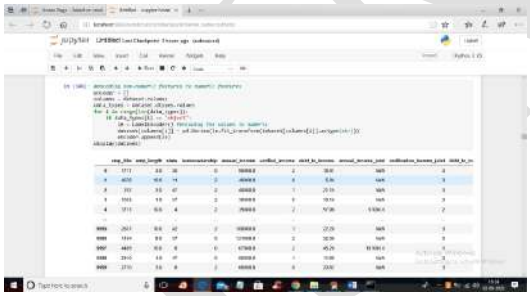


Fig.10 Use encoder algorithm

In above screen using encoder algorithm we are converting all non-numeric data into numeric values

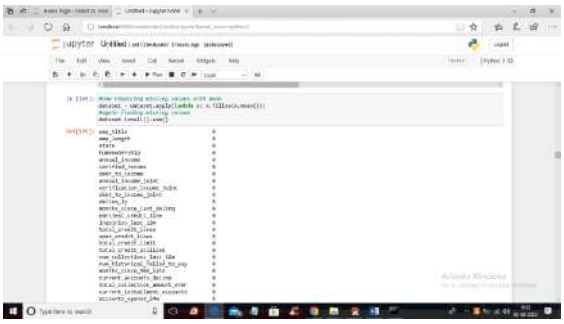


Fig.11 all missing values with MEAN of each column

In above screen we are replacing all missing values with MEAN of each column and after replacing we got missing count values as 0.

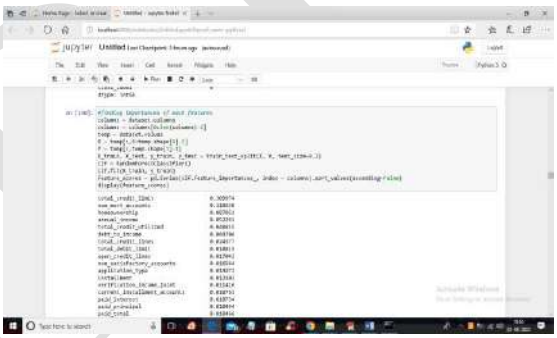


Fig.12 calculating importances of features

In above screen using Random Forest we are calculating importances of features and below is the graph

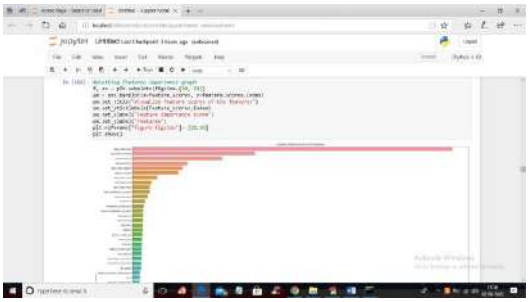


Fig.13 names of features

In above graph we can see names of features which has high value or importance

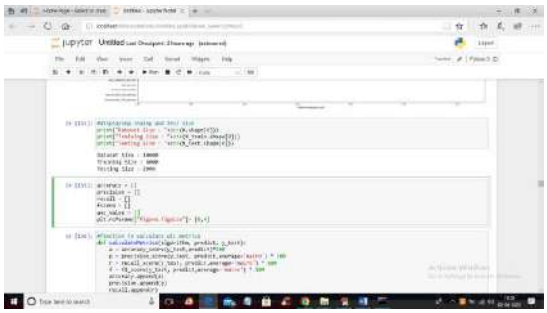


Fig.14 Size of dataset

In above screen we are displaying dataset size, train and test size

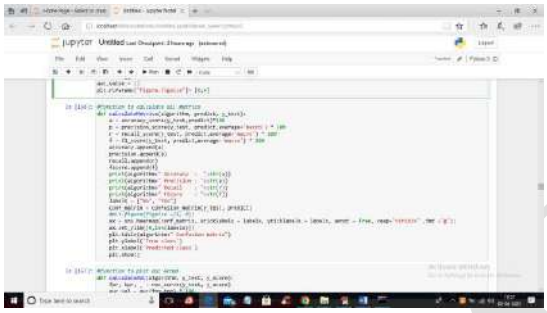


Fig.15 Calculating accuracy

Using above screen code we are calculating accuracy and other metric for each algorithm

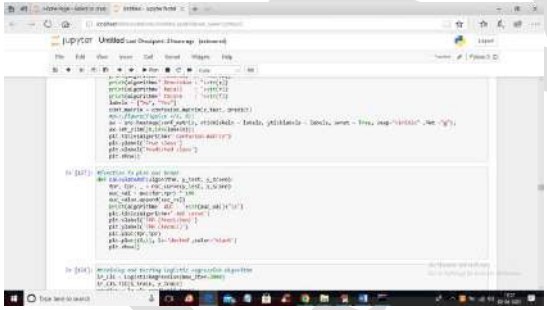


Fig.16 calculating and plotting AUC values

By using above screen function we are calculating and plotting AUC values

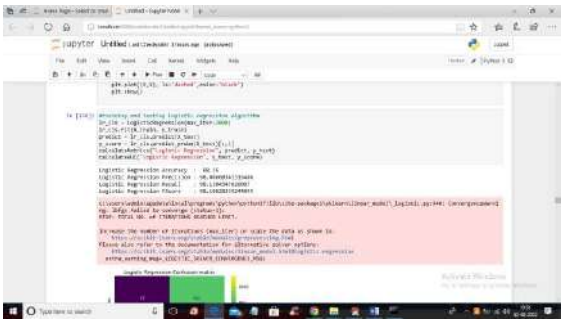


Fig.17 training Logistic Regression

In above screen we are training Logistic Regression got its accuracy as 98.35% and below is the AUC graph and confusion matrix graph

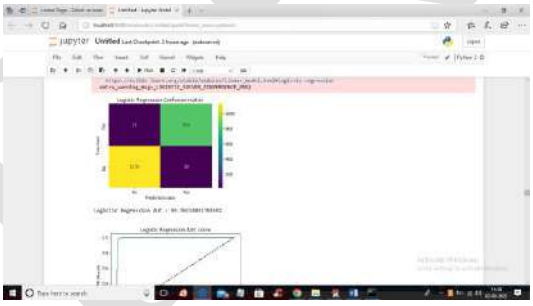


Fig.18 confusion matrix of logistic regression

In the above confusion matrix of LR, the y-axis represents true classes, and the x-axis represents predicted classes and all counts in blue color boxes are wrong predictions, so logistic regression predicted 37 wrong records out of 2000.

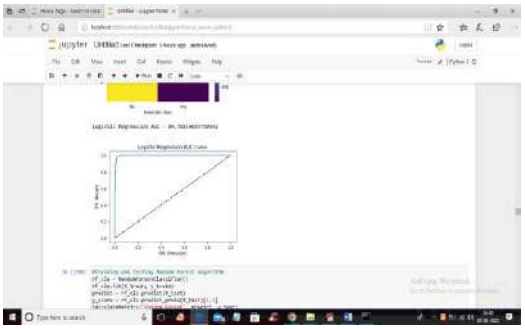
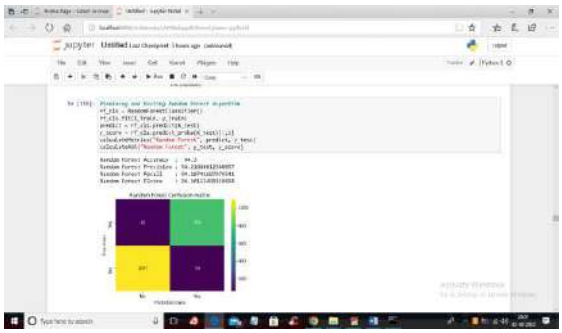
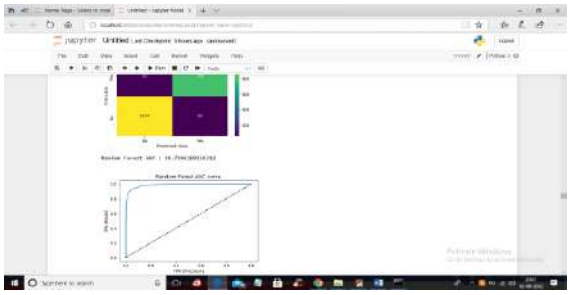


Fig.19 AUC graph

In above AUC graph we can see blue line starts from 0 and reached closer to 1. In below screen showing Random Forest output





In below screen showing XGBOOST output

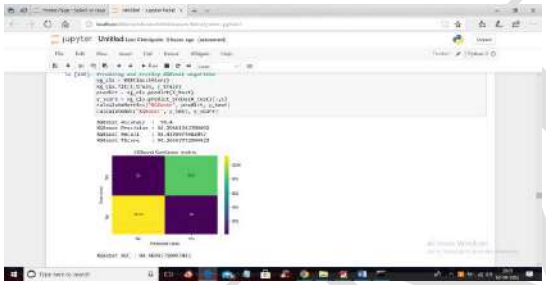
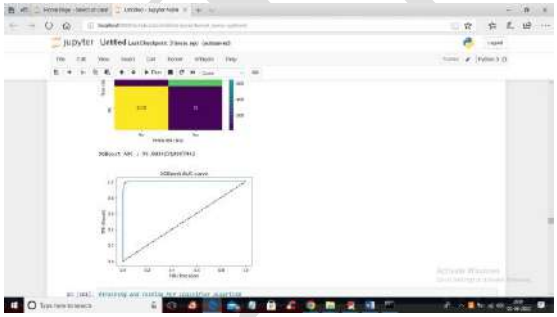


Fig.20 output of XGBOOST



In below screen showing MLP output

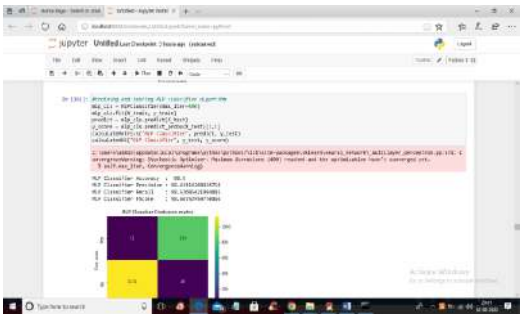
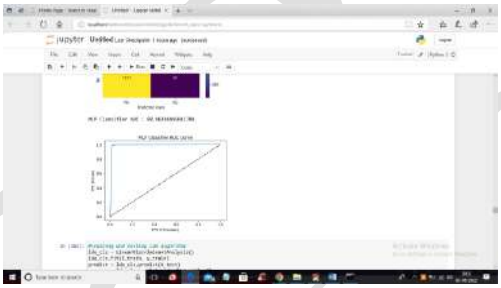


Fig.21 MLP output



In below screen showing LDA output

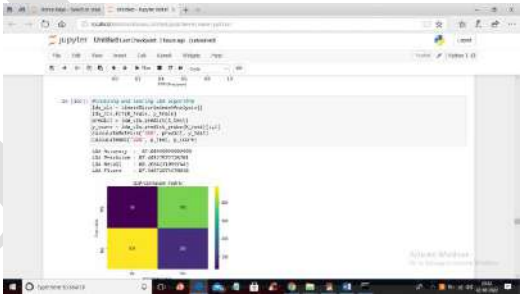
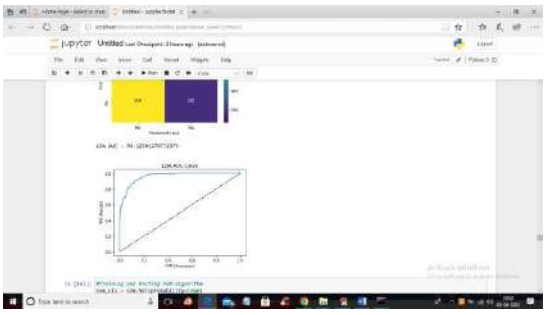


Fig.22 LDA output



In below screen showing SVM output

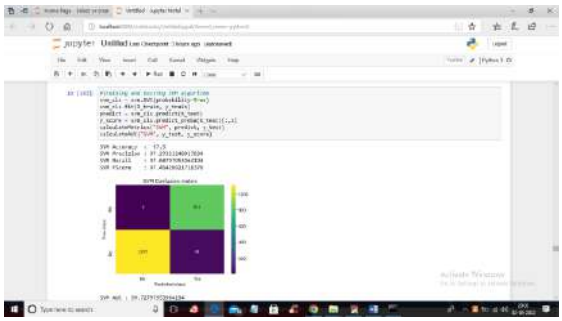
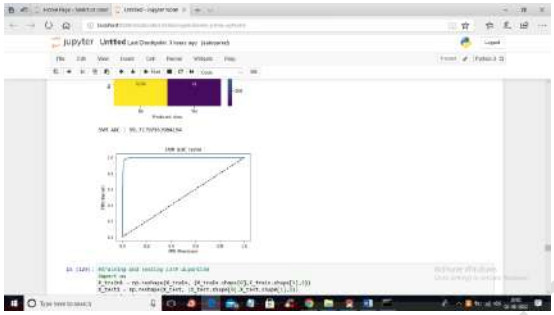


Fig.23 SVM output



In below screen showing LSTM output

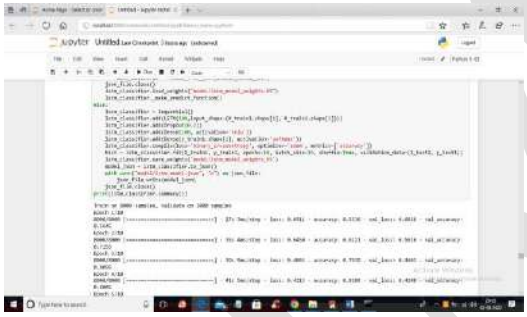
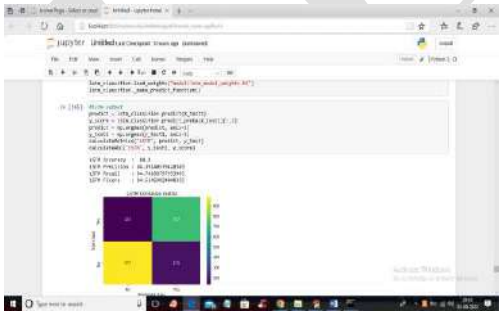
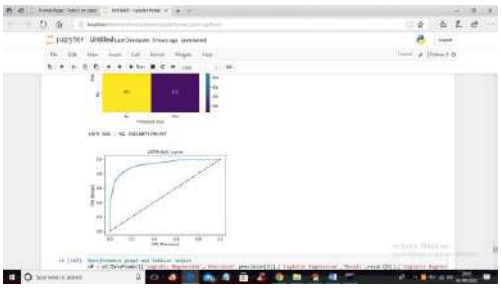


Fig.24 LSTM output





In below screen showing all algorithms performance graph

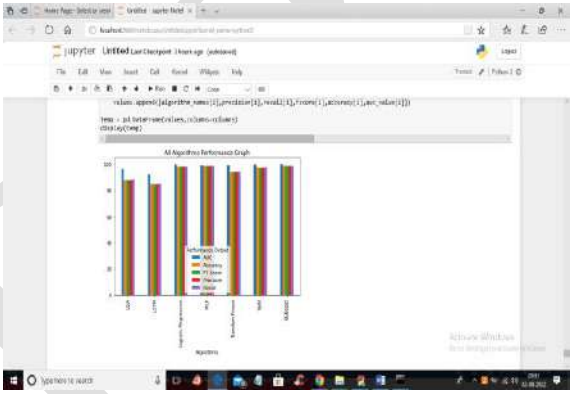


Fig.25 Performance Graph

In below screen showing all algorithms performance in tabular format

Fig.26 all algorithms performance in tabular format

V. CONCLUSION

In this paper to predict loan default verifies the ability of as XGBoost and RF. In conclusion, the practice of predicting bank loan defaults stands as a pivotal pillar of modern financial management. As banks and lending institutions navigate the intricate landscape of lending, the ability to anticipate which

borrowers are likely to default on their loans emerges as a powerful tool for risk mitigation, sound decision-making, and the preservation of financial stability. In the feature engineering section, we utilize the variance inflation factor approach and the variance threshold method to filter out unnecessary features before feeding those features into the XGBoost and Random Forest models. XGBoost has a minor difference in prediction accuracy because both have high accuracy in loan default cases. In the future, we will compare several sophisticated machine learning algorithms, for loan default prediction such as MLP, Neural Network, and KNN or a combination of them, to establish the best model.

REFERENCES

1. Lai, L. (2020) developed by "Loan Default Prediction with Machine Learning Techniques." 2020 International Conference on Computer Communication and Network Security (CCNS), Xi'an, 21-23 August 2020, 5-9. https://doi.org/10.1109/CCNS50731.2020.00009
2. Aslam, U., Tariq Aziz, H.I., Sohail, A., et al. (2019) are developed by "An Empirical Study on Loan Default Prediction Models.' Journal of Computational and Theoretical Nanoscience, 16, 3483-3488
3. Madaan, M., Kumar, A., Keshri, C., et al. (2021) are developed by "Loan Default Prediction Using Decision Trees and Random Forest: A Comparative Study." IOP Conference Series: Materials Science and Engineering, 1022, 012042. https://doi.org/10.1088/1757-899X/1022/1/012042
4. Malekipirbazari, M. and Aksakalli, V. (2015) are developed by "Risk Assessment in Social Lending via Random Forests." Expert Systems with Applications, 42, 4621-4631. https://doi.org/10.1016/j.eswa.2015.02.001
5. Ma, X., Sha, J., Wang, D., et al. (2018) are developed by "Study on a Prediction of P2P Network Loan Default Based on the Machine Learning LightGBM and XGboost Algorithms According to Different High Dimensional Data Cleaning." Electronic Commerce Research and Applications, 31, 24-39
6. Nalić, J. and Švraka, A. (2018) are developed by "Using Data Mining Approaches to Build Credit Scoring Model: Case Study—Implementation of Credit Scoring Model in Microfinance Institution" 2018 17th International Symposium Infoteh-Jahorina (INFOTEH), East Sarajevo, 21-23 March 2018, 1-5