Framework for Bank Loan Re-Payment Prediction and Income Prediction

Chudi Dhruv
Artificial Intelligence, Department of Information Technology
National Institute of Technology Karnataka
Surathkal, India

Deva Paul
Artificial Intelligence, Department of Information Technology
National Institute of Technology Karnataka
Surathkal, India

M Hemanth Kumar
Artificial Intelligence, Department of Information Technology
National Institute of Technology Karnataka
Surathkal, India

Anand Kumar M
Department of Information Technology
National Institute of Technology Karnataka
Surathkal, India

M Shashank Reddy
Artificial Intelligence, Department of Information Technology
National Institute of Technology Karnataka
Surathkal, India

2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC) | 979-8-3503-0071-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSCCC58608.2023.10176363

***Abstract*—This research study aims to develop a predictive framework for income and bank loan repayment prediction. The primary objective is to accurately predict an individual's income and their ability to repay a loan to help them make informed financial decisions. Using a data-driven approach, we collected and analyzed data on various factors that impact income and loan repayment, such as employment history, education, credit score, and demographic information. This data will be used to build predictive models that can provide accurate estimates for both income and loan repayment. The models will be validated using historical data and refined to improve accuracy. The study will focus on developing two separate predictive models: one for income prediction and another for bank loan repayment prediction. The income prediction model will provide individuals with an estimate of their future income based on their individual financial circumstances. The bank loan repayment prediction model will help financial institutions predict the likelihood of loan repayment based on the borrower's financial history and current financial circumstances. This predictive framework will provide valuable insights into the financial stability of individuals and the creditworthiness of borrowers. It will help individuals plan for their financial future, such as saving for retirement or investing in the stock market. It will also assist financial institutions in making informed lending decisions, reducing the risk of loan defaults, and improving the overall health of the financial industry. The development of a predictive framework for income and bank loan repayment prediction will provide valuable insights and tools for both individuals and financial institutions. Accurate predictions of income and loan repayment will enable informed financial decisions, improving the financial stability and well-being of all parties involved. We have created a user-friendly financial bot that can provide basic definitions of financial terms based on user queries.**

***Index Terms*—Income Factors, Loan Eligibility, Income Prediction, Financial queries.**

I. Introduction

Recently, the incorporation of machine learning in the financial industry has seen significant growth and is now considered a vital element in various financial services and applications, such as determining credit scores, approving personal loans, assessing mortgage applications, and categorizing customer risk.

A. Income Prediction

Income prediction using census data helps identify patterns and trends in income levels among different demographic groups. It is used by government organizations, businesses, and non-profits to improve the economic well-being of communities. It can also be used for research purposes to analyze the distribution of income among different demographic groups. Census data is crucial for maintaining the democratic system of government and has a significant impact on the economy. It is used to allocate federal funding from the government to different states and localities.

B. Bank loan eligibility

Bank loan eligibility is a criterion that has been standardized by lenders to evaluate the willingness and ability of a customer to qualify for a loan scheme. Bank loan repayment prediction is important because it helps financial institutions assess the risk of lending money to a borrower. By accurately predicting whether a borrower will be able to repay a loan, banks, and other lenders can make more informed decisions about which loan applications to approve, and at what terms. This can help reduce the risk of default and financial losses for the lender and also help ensure that credit is available to borrowers who are likely to use it responsibly. Additionally, it can also help to identify potentially fraudulent activity.

II. MOTIVATION

The paper "Framework for Bank Loan Re-Payment Prediction and Income Prediction" aims to address the need for accurate prediction models in the banking industry. Accurate predictions of loan approvals and income levels are crucial for banks to make informed decisions and mitigate risks associated with lending. However, the existing prediction models have limitations such as insufficient data, complex algorithms, and lack of transparency, which result in inaccurate predictions and reduce the efficiency of the lending process.

To overcome these limitations, the paper proposes a framework that utilizes machine learning techniques to develop a robust prediction model. The proposed framework leverages various data sources, including customer data, credit history, and economic indicators, to generate accurate predictions of loan approvals and income levels.

The significance of this research lies in its potential to improve the accuracy of loan and income predictions, enabling banks to make informed decisions and minimize the risk of defaults. The proposed framework can also enhance the efficiency of the lending process by reducing the time and resources required to assess loan applications. The framework's transparency and interpretability can increase the trust of customers in the lending process and enable them to make informed decisions about their financial futures.

III. LITERATURE

Navoneel Chakrabarty and Sanket Biswas [1] in their paper proposed a statistical approach to predict the income level of individuals based on demographic and socio-economic factors. The authors used logistic regression and decision trees to develop models for predicting income and evaluated their performance using metrics such as accuracy and precision.

Bramesh S M [2], in his paper compared the performance of various machine learning algorithms, including k-NN, SVM, and random forest, in predicting income based on the census income dataset. The results showed that random forest performed better than the other algorithms.

The paper by Chet Lemon, Chris Zelazo, and Kesav Mulakaluri [3] explores the use of simple classification techniques such as decision trees and Naive Bayes to predict whether an individual's income exceeds 50,000 per year based on census data.

Monika Papouskova and Petr Hajek [4], in their paper proposed a two-stage ensemble learning approach to predict consumer credit risk. The authors used multiple algorithms, including SVM, decision trees, and random forest.

Anshika Gupta and team [5] ,in their paper proposed a machine learning-based approach to predict the likelihood of loan repayment based on various factors, such as income, credit score, and employment status. The authors used various algorithms, including logistic regression and decision trees.

Kumar and Goel [6] ,in their paper proposed a machine learning-based approach to predict loan approval using demographic and financial data. The authors used various algorithms, including logistic regression and SVM.

Sayan Das and team [7] ,in their paper presents a study on salary prediction using regression techniques such as linear regression, polynomial regression, and decision tree regression. The study compares the performance of these techniques on a dataset containing information about employees' years of experience and salaries. The results show that the decision tree regression technique outperforms the other methods in terms of accuracy and efficiency.

Dr. C K Gomathy and team [8] ,in their paper proposed a machine learning-based approach to predict the likelihood of loan approval based on various factors, such as income, credit score, and employment status. The authors used various algorithms, including decision trees and random forests.

The paper by Anuj More, Amay Naik, and Sarita Rathod [9] presents a novel framework for predicting the salary of freshers based on their skills. The framework uses machine learning techniques such as linear regression and decision trees to predict the salary based on the skills possessed by the fresher.

IV. METHODOLOGY

A person's income plays a crucial role in their ability to obtain a bank loan. Higher-income is typically associated with lower credit risk, making loan approval more likely and resulting in more favorable loan terms such as lower interest rates. Thus, a strong income is a key factor for individuals seeking financial assistance from a bank. So, we are predicting the income of an individual first and then using this income to predict whether they are able to repay the bank loan. However, income is not the only factor that affects bank loan repayment. There are many other factors that affect bank loan repayment, as mentioned in the literature section.

The Flow chart in the figure-1 shows the workflow/methodology that had been followed.

A. Data collection

We collected the dataset for predicting income from the census, and for bank loan prediction, we collected a dataset from a reputed bank.

B. Exploratory data analysis

We used feature engineering to find the best parameters for predicting income and loan status.

Analysis of Income census data:

The dataset used in this study includes an "Education" column as shown in Figure 2 with 16 different categories. A thorough examination of the data revealed that there are no missing values in the Education column. Additionally, a significant proportion of individuals in the data set have completed high school (Hs-grad) or some college, as these categories had the highest frequency among all the education levels.

The dataset used in this study includes an "Education_num" column as shown in Figure 3 which represents the number of

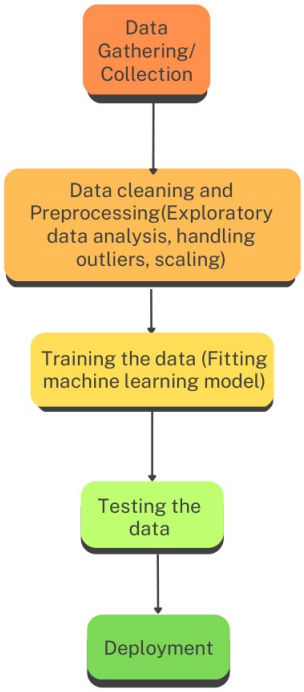

Fig. 1: workflow

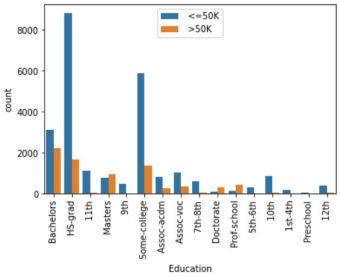

Fig. 2: Education and Income relation

years of education an individual has received. This column has 16 distinct values and provides similar information as the ”Education” column, which categorizes educational levels.

The dataset used in this study includes a “Sex” column as shown in Figure 4 with two categories: Male and Female. An examination of the data revealed that the number of male individuals in the sample is almost double the number of female individuals. Based on this information, it can be hypothesized that males may have a higher probability of earning more than 50K compared to females.

The dataset used in this study includes a “Race” column as shown in Figure 5 with five different categories. An examination of the data revealed that the largest proportion of individuals in the sample identifies as ’White’. Based on this information, it can be hypothesized that individuals who

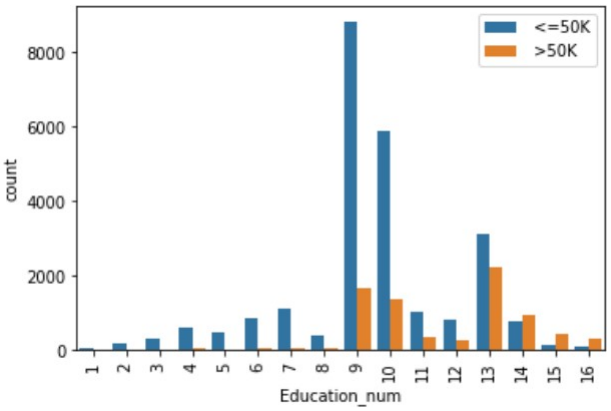

Fig. 3: Education number and Income relation

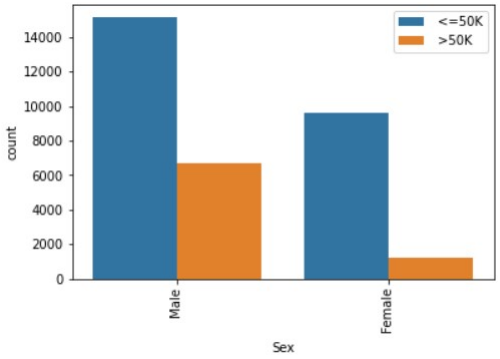

Fig. 4: Sex and Income relation

identify as ’White’ or ’Asian-PacIslander’ may have a higher probability of earning more than 50K.

The dataset used in this study includes a “Marital_Status” column as shown in Figure 6 with seven different categories. An examination of the data revealed that the majority of individuals in the sample have a marital status of ’Married-

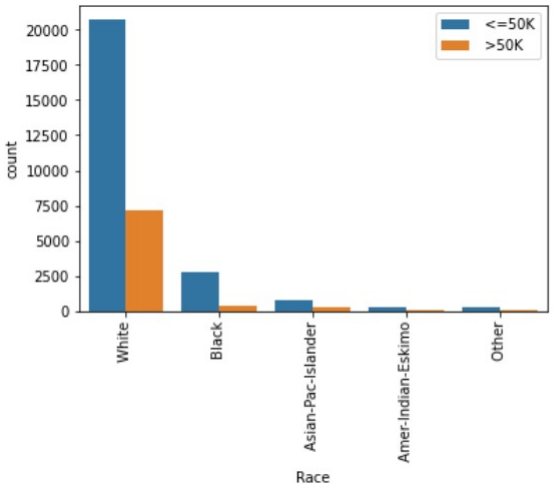

Fig. 5: Race and Income relation

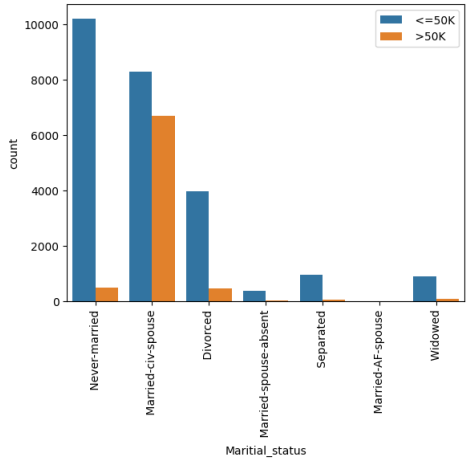

Fig. 6: Marital status and Income relation

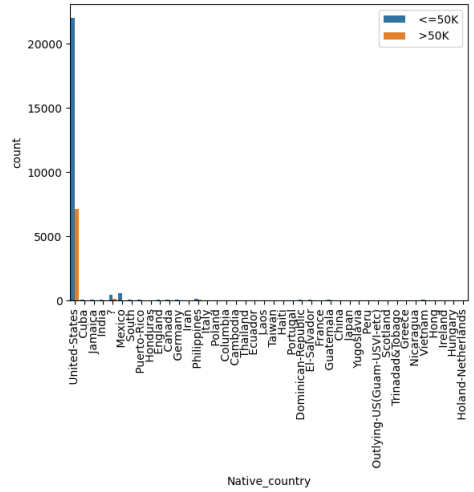

Fig. 8: Native Country and Income relation

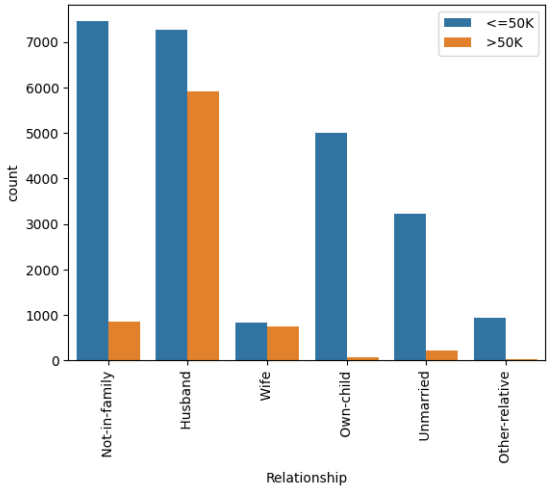

Fig. 7: Relationship in family and Income relation

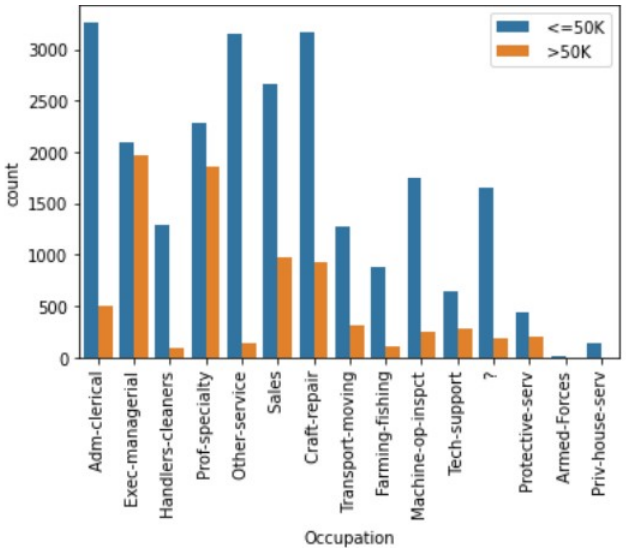

Fig. 9: Occupation and Income relation

civ-spouse', while the least common is 'Married-AF-spouse'. Additionally, a significant proportion of individuals in the sample have never been married.

The dataset used in this study includes a "Relationship" column as shown in Figure 7 with six different categories. An examination of the data revealed that the highest number of individuals in the sample have a relationship status of 'Husband', while the lowest is 'Other-relative'. Additionally, no missing values are present in this column. Based on this information, it can be hypothesized that individuals who have a relationship status of 'Husband' or 'Wife' may have a higher probability of earning more than 50K.

The dataset used in this study includes a "Native_Country" column as shown in Figure 8 with multiple categories. An examination of the data revealed that the highest number of individuals in the sample are from 'United-States', while the rest of the countries have a significantly lower representation in the sample.

The dataset used in this study includes an "Occupation" column as shown in Figure 9 with 14 different categories. An examination of the data revealed that there are some missing values in this column represented by the '?'. To address this issue, the missing values in the 'Occupation' column was replaced by 'Prof-speciality'. This method was chosen to ensure that the dataset is complete and ready for analysis.

An analysis of the correlation between the numeric columns in the dataset as shown in Figure 10 revealed that the income has a moderate correlation with 'Education_num' (33%), 'hours_per_week' and 'age' (23%), and 'Capital_gain' (22%).

Analysis of Bank Loan data:
The dataset used in this study includes a "Years in current job" column as shown in Figure 11 which represents the number

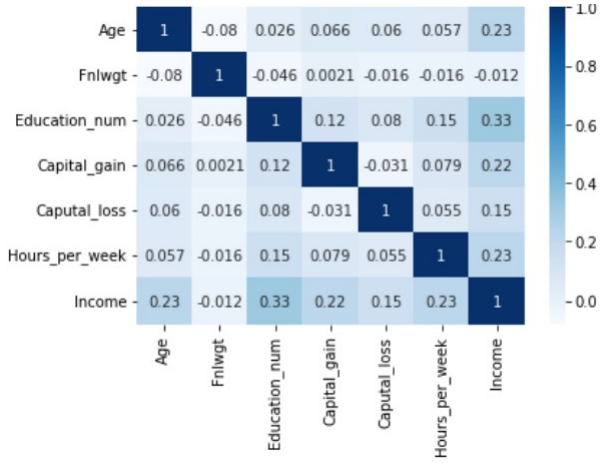

Fig. 10: correlation matrix of numerical features

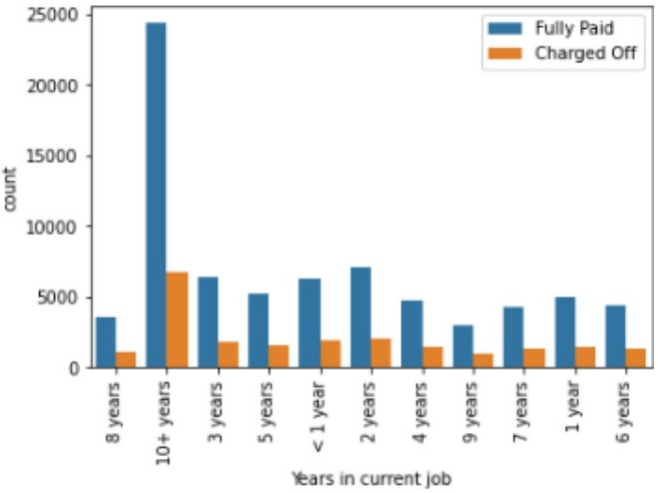

Fig. 11: Years in current job and Loan status relation

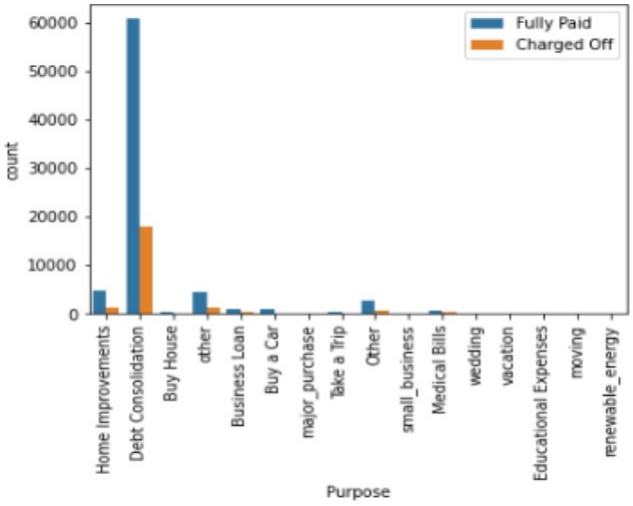

Fig. 12: Purpose and Loan status relation

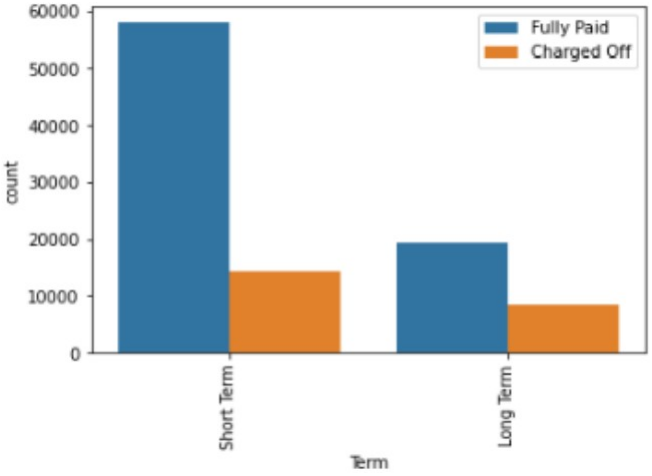

Fig. 13: Term and Loan status relation

of years an individual has been working at their current job. This column has 11 distinct values. An examination of the data revealed that the proportion of loans that have been successfully repaid is higher for individuals who have been working at their current job for more than 10 years when compared to those who have been working for less than 10 years. However, it is also observed that the proportion of loans that have been charged off is also higher for individuals who have been working at their current job for more than 10 years when compared to those who have been working for less than 10 years.

The dataset used in this study includes a ”Purpose” column as shown in figure 12 represents the reason for applying for a loan. This column has 16 distinct categories, however, many of the categories have a very low representation in the data. An examination of the data revealed that the purpose of ”debt consolidation” has the highest proportion of loan repayment and the highest proportion of loans being charged off.

The dataset used in this study includes a ”Term” column as shown in Figure 13 which represents the duration of the loan. This ”Term” column has two categories: one is “Short Term” and the other is ”Long Term”. An examination of the data revealed that in comparison, the proportion of loans that have been successfully repaid is higher for short-term loans, as well as the proportion of loans that have been charged off is also higher for short-term loans.

The dataset used in this study includes a ”Home Ownership” column as shown in Figure 14 which represents the type of home an individual owns or rents. This column has four categories: ”Home Mortgage”, ”Rent”, ”Own Home”, and ”Have Mortgage”. An examination of the data revealed that the proportion of loans that have been successfully repaid is highest in the case of a home mortgage followed by rent, while the proportion of loans that have been charged off is maximum when it is rent followed by a home mortgage.

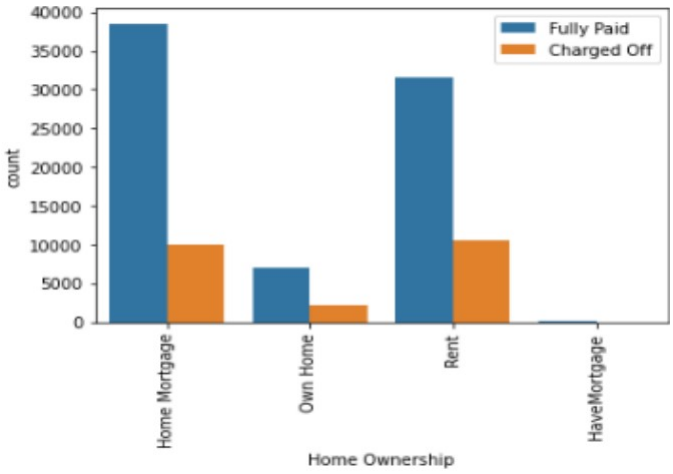

Fig. 14: Home Ownership and Loan status relation

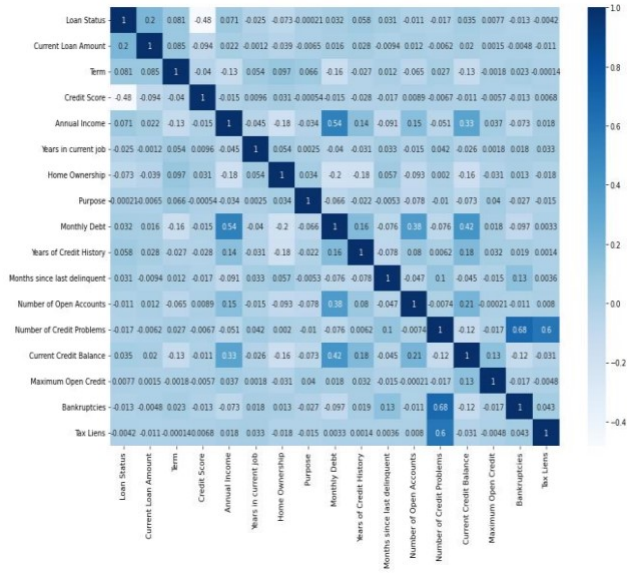

Fig. 15: correlation of numerical features in the dataset

An analysis of the correlation between the numeric columns in the dataset as shown in Figure 15 revealed that the 'Loan Status' has a moderate correlation with 'Current Loan Amount' (20%), 'Term' (8.1%), 'Annual Income' (7.1%), 'Monthly Debt' (3.2%), 'Years of Credit History' (5.8%), 'Months since last delinquent' (3.1%), 'Current Credit Balance' (3.5%) and 'Maximum Open Credit' (0.77%).

C. Handling outliers and Imbalanced data:

Handling outliers and imbalanced data are very important to improve the performance of the model.

In the Income dataset,
the fnlwgt column, capital gain, and capital loss have outliers, and we have handled them. The income column is the target feature of the Income dataset with 2 classes '≤50K' and '>50K'.The '≤50K' class has 24719 values and the '>50K' class has 7841 values, suggesting that our data set is imbalanced with respect to target features.
We have found out the features Work-class, Occupation, and Native country have missing data, which we handled by replacing the missing values with the most commonly occurring ones.

In the Bank Loan data set,
The target feature- loan status, has two classes, fully paid, and charged off respectively. There are 43179 values in the fully paid class and 12568 values in the charged-off class. This implies that the data set is imbalanced with respect to the target feature. There are missing values in credit score, annual income, years in current job, months since last delinquency, and bankruptcies.
We handled these missing values by removing the respective rows using the function drop ().

D. Encoding and Scaling the data

we utilized label encoder from sci-kit for encoding categorical values as numerical values. Additionally, we applied StandardScaler from sci-kit as a pre-processing technique to standardize the features of our dataset by subtracting the mean and dividing by the standard deviation. This normalization process is necessary to ensure that all features are on a similar scale and possess similar properties, thereby enhancing the model's performance.

E. Machine Learning Models

The accuracy of different models are observed in Table-I and Table-II shown below.
Income Prediction using Census Data,

1) **SVM**
Support Vector Machine (SVM) [10][11][12] is a widely used supervised learning algorithm that leverages a hyperplane to partition data into separate classes based on their attributes. The algorithm determines the optimal hyperplane by maximizing the margin between the two classes, resulting in superior classification accuracy and the ability to handle noisy data.

2) **Random Forest**
Random Forest [10][11][12] is an ensemble learning algorithm that merges multiple decision trees for the purpose of classifying or regressing data. It creates diverse trees by randomly selecting features and data samples, thereby reducing overfitting while improving the model's accuracy, generalization, and ability to handle complex data patterns. .

3) **Gradient Boosting**
Gradient Boosting [10][11][12] is a commonly used ensemble learning technique that combines weak learners, typically decision trees, in order to enhance prediction accuracy. The algorithm iteratively trains new models, concentrating on the errors made by previous models, which reduces the residual error and results in

TABLE I: Analysis of Models used based on the accuracy, F1 score, Recall score and Precision Score for Income data set.

Model	Accuracy(%)	Precision score(%)	Recall score(%)	F1 score(%)
SVM	83.93	51.86	74.38	61.11
Gradient Boosting	86.30	59.11	79.33	67.75
Logistic Regression	82.11	44.41	71.25	54.72
Random Forest	85.22	62.20	73.09	67.21
Gaussian NB	80.33	34.06	69.58	45.74
K-Nearest Neighbours	81.89	56.40	64.68	60.26
Decision Tree	80.76	62.39	60.08	61.21

TABLE II: Analysis of Models used based on the accuracy, F1 score, Recall score and Precision Score for Bank Loan data set.

Model	Accuracy(%)	Precision score(%)	Recall score(%)	F1 score(%)
SVM	85.54	100	91.73	84.72
Gradient Boosting	85.58	99.93	94.80	91.74
Logistic Regression	85.55	100	84.73	91.73
Random Forest	85.44	99.50	84.93	91.64
Gaussian NB	70.63	73.77	87.65	80.11
K-Nearest Neighbours	83.37	95.89	85.22	90.25
Decision Tree	79.47	87.84	86.73	87.28

the development of a robust predictor.

4) **Logistic Regression**
Logistic Regression [10][11][12] is a widely used statistical method that is employed for solving binary classification problems. The technique operates by predicting the probability of an event taking place based on input variables. Logistic Regression models the relationship between the input variables and the binary output using the logistic function. The logistic function maps input values to values ranging between 0 and 1, representing the probability of the positive class.

5) **Decision Tree**
The Decision Tree [10][11][12] is a widely adopted machine learning algorithm that is utilized for both classification and regression tasks. It operates by dividing the data into a hierarchical structure resembling a tree, which is based on the input features. The algorithm uses a series of if-else conditions to split the data recursively, ultimately resulting in the creation of leaf nodes that hold the predicted class or value based on the majority of the samples.

6) **K- nearest Neighbour**
K-Nearest Neighbors (KNN) [10][11][12] is a widely used machine learning algorithm that is utilized in both classification and regression tasks. The algorithm works by identifying the k nearest training examples to a given test example and using them to make a prediction.

7) **Gaussian NB**
Gaussian Naive Bayes (GaussianNB) [10][11][12] is a commonly used probabilistic machine learning algorithm that is primarily employed in classification tasks. It works by determining the conditional probability of each input feature based on the class label, and then applies Bayes' theorem to compute the posterior probability of each class. GaussianNB assumes that input features are independent and follow a Gaussian distribution.

Bank Loan Re-Payment Prediction, the same models have been used on the bank loan data, which gave the following results as shown in Table-2:

F. Streamlit

Streamlit is a freely available framework that enables the creation of interactive web-based applications for Machine Learning and Data Science purposes. With its user-friendly interface, it provides support for widgets and charts to facilitate data visualization, and it can be deployed either locally or on cloud-based servers.
We used a streamlit framework to build an income prediction using census data and a bank loan repayment prediction webpage.

G. Financial Queries

We used the pyttsx3 library to initialize a text-to-speech (TTS) engine using the 'sapi5' driver, which is a Microsoft

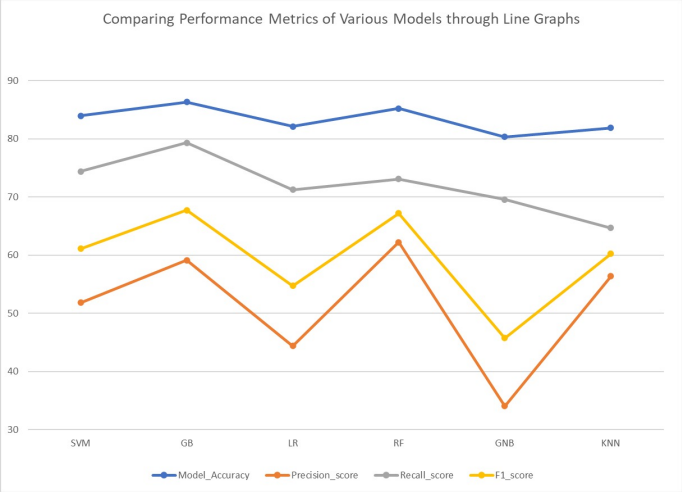

Fig. 16: Comparison of Performance Metrics of various Models for Income dataset

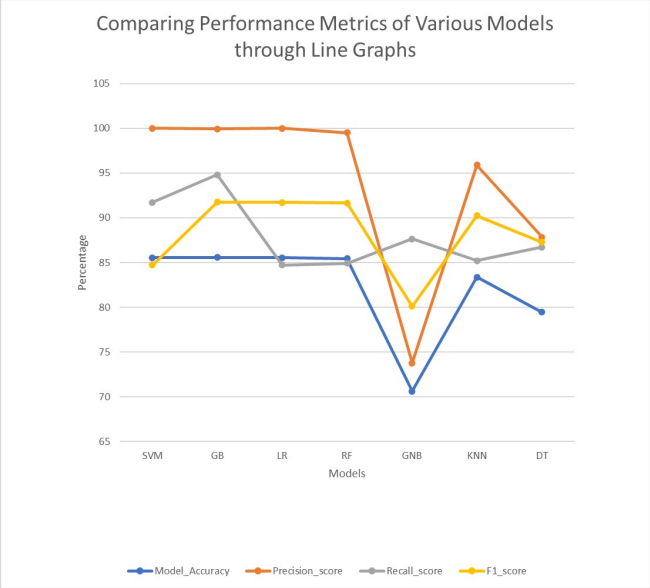

Fig. 17: Comparison of Performance Metrics of Various Models for Bank Loan Dataset

speech API. We then used the 'voices' variable to get a list of available voices from the TTS engine, which was used in the Introduction part of the web application. Additionally, we used the Wikipedia API in Python to search for a Wikipedia page based on a user input query, and it returns the content of that Wikipedia page.

V. CONCLUSION AND FUTURE SCOPE

This paper presents a study on bank loan repayment based on Machine Learning in Finance. We built a website that contains three different options: Income prediction using census data, Bank loan repayment prediction, and Answers to Financial queries. Figures 16 and 17 show the Income Prediction using Census data and Bank loan re-payment prediction web pages, respectively.

This is the github link for the code https://github.com/MogilipalemHemanthKumar/FRAMEWORK-FOR-BANK-LOAN-RE-PAYMENT-AND-INCOME-PREDICTION.

For income prediction, we used a gradient-boosting classifier due to its higher accuracy and f1 score. For bank loan repayment prediction, we used SVM as it also has higher accuracy and an f1 score. We converted these models into a .sav file and deployed a web application using Streamlit. To answer financial queries, we used the Wikipedia API module in Python, which returns the content in Wikipedia if that page exists. We aim to improve the accuracy of these models using hyperparameter tuning techniques such as GridSearchCV and RandomizedSearchCV.

We used the Wikipedia API module in Python to answer the user queries in the financial queries part of the web application. We also aim to build an interactive Finance bot to answer user queries related to finance using Natural Language Processing.

REFERENCES

[1] Navoneel Chakrabarty,Sanket Biswas:"A Statistical Approach to Adult Census IncomeLevel Prediction", International Conference on Advances in Computing, communication control and Networking (ICACCCN2018).
[2] Bramesh S M:"Comparative Study of Machine Learning Algorithms on Census Income Data Set"(IRJET)-(2019).
[3] Lemon, Chet, Chris Zelazo, and Kesav Mulakaluri. "Predicting if income exceeds 50,000 per year based on 1994 US Census Data with Simple Classification Techniques."—2019
[4] Monika Papouskova, Petr Hajek, "Two-stage consumer credit risk modelling using heterogeneous ensemble learning," Decision Support Systems, Volume 118, 2019, Pages 33-45, ISSN 0167-9236, https://doi.org/10.1016/j.dss.2019.01.002.
[5] Anshika Gupta, Moradabad,Vinay Pant,Sudhanshu Kumar,Pravesh Kumar Bansal "Bank Loan Prediction system using Machine Learning", 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART).
[6] Kumar, S., Goel, A. K. (2020):"Prediction of loan approval using machine learning technique. International Journal of Advanced Science and Technology, 29(9s)".
[7] Sayan Das,Rupashri Barik,Ayush Mukherjee:"Salary Prediction Using Regression Techniques",Industry Interactive Innovations in Science, Engineering Technology (I3SET2K19)(2020).
[8] Dr.C K Gomathy, Ms.Charulatha, Mr.AAkash ,Ms.Sowjanya:"THE LOAN PREDICTION USING MACHINE LEARNING ",International Research Journal of Engineering and Technology (IRJET)Volume: 08 Issue: 04 — Apr 2021 .
[9] Anuj More , Amay Naik,Sarita Rathod: "PREDICT-NATION Skills Based Salary Prediction for Freshers":International Conference on Advances in Science Technology (ICAST2021).
[10] Géron, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. " O'Reilly Media, Inc.", 2022.
[11] https://towardsdatascience.com/comparative-study-of-classifiers-in-predicting-the-income-range-of-a-person-from-a-census-data-96ce60ee5a10, accessed on 24 Apr 2023.
[12] https://medium.com/mlreview/gradient-boosting-fromscratch-1e317ae4587d, accessed on 24 Apr 2023.