Loan Eligibility Prediction

1st Mehrab Islam 1821183642 2nd Anik Roy Pranto 1921219042 3rd Adnan Samiul Alam 1921476042

***Abstract*—In this day and age of computer technology, where everything is becoming automated, doing something as loan eligibility prediction manually is both inefficient and tedious. Banks can save a ton of resources and time if loan eligibility is done automatically with the help of machine learning. Thousands of people's details and credentials are typically reviewed by loan underwriters each year before deciding whether to grant or deny loans. The main goal of this work is to develop a complete end-to-end loan eligibility prediction system that will not only save time but also help banks minimize foreseen issues such as credit risk. This work starts with collecting relevant data and then combining them into a public dataset and then with EDA and data preprocessing is applied to clean the data and find the correlation between the features. Then the data was trained and tested using seven machine learning models: Logistic Regression, Decision Tree, Random Forest, SVM, KNN, Categorical Naïve Bayes and Gaussian Naïve Bayes. All these models were compared based on their precision, recall, F1 score and accuracy. The results showed that Random Forest was the model with the highest testing accuracy out of the seven models. The model was then trained and extracted to be used to predict loan eligibility from data collected from a website. A hypertext markup language-based web interface was constructed to be used as a front end where users can input their credentials to get the loan eligibility prediction.**

Keywords: loan eligibility, machine learning model, random forest, precision, recall, F1 score, accuracy, hypertext markup language.

I. Introduction

The banking industry, like many other businesses, is increasingly looking to take use of the opportunities provided by modern technologies to enhance their operations, boost productivity, and reduce costs. According to [10], the predictive analytics feature of Machine learning was the most utilized feature for applications in the banking sector worldwide in 2020. The success or failure of most lending platforms largely depends on their ability to evaluate credit risk [11]. Our main objective of this research is to predict the safety of loan [12]. Basically Loan eligibility is simply the process of determining whether or not a potential borrower meets the criteria for a particular loan. This can be based on variables like credit score, employment history, and income level. For example, The ability of borrowers to repay their loans is supported by loan eligibility, which is crucial. Defaulting on a loan can have a number of negative effects, such as lowering one's credit score and making it more challenging to get future financing. As a result, before granting credit, lenders take considerable effort to determine loan eligibility. Borrowers can increase their chances of getting funding by being aware of the criteria that lenders use to assess loan eligibility. Hence we are focused on developing Machine Learning (ML) models to predict loan eligibility, which is vital in accelerating the decision-making process and determining if an applicant gets a loan or not.

Our objectives in this study include;

- Clean and Preprocess the data for modeling
- Perform Exploratory Data Analysis (EDA) on the dataset
- Build various ML models to predict loan eligibility and
- Evaluate and Compare the different Models built

Machine learning is the field of study that gives the computer the ability to learn without being explicitly programmed. [13] It is the study of how to make machines more human-like in their behavior and decisions by allowing them to learn and generate their own programs. In order to resolve the problem, we developed an automated loan prediction system. First, we have collected our data. Then we completed the data pre-processing and exploratory data analysis for getting the data in the most suitable form. Then we will train the system using our collected data on a set of parameters loans were approved. Since the dataset is labeled and the desired output falls into two categories, we used supervised learning and classification algorithms. Consequently, a machine can evaluate and comprehend the process. The system will then look for a suitable applicant and provide the results. The main objective of this project is to give a rapid, easy, and immediate method of selecting the qualified applicants.

Advantages:

- It will take less time to sanction loans.
- The entire process will be computerized to eliminate human error.
- A suitable applicant will immediately receive loan approval.

Automation in banking sector is nothing new and already has been implemented in many ways to reduce human errors and time consumption to minimize cost of operation. Machine learning based research and development in this field such as fraud detection in bank payment, risk management, task automation and of course loan prediction is continuously yielding good results. [14], [15], [16]. These machine learning applications is based on historical data or the data which was provided to the banks by the clients [17]. Data also can be collected from Kaggle on which different machine learning techniques can be applied for load prediction. [18]

This paper implements automated loan prediction based on machine learning with the contributions described below:

- A novel web user interface which is built with optimization, compatibility and responsiveness mind so that it runs well on a plethora of hardware. So that, each branch of the banks can operate from anywhere in the country as long as they have a computer of some sort regardless of the configuration.
- Another contribution is on the data set that was collected from Kaggle where 200 additional instances of data were added to it. Which was provided and verified by a loan underwriter.

II. Related Works

Loan approval requires a bank to have designated qualified and skilled underwriters to go through all the credentials of all the applicants which is costly, time-consuming [10]. While with machine learning, any amount of data can be processed and can be used to predict loan eligibility, minimizing the overall cost [11]. That is why many researchers are taking advantage of data science techniques such as machine learning to automate this entire process [12]. In the following paragraphs, the contribution of those researchers and their developed systems are briefly described.

G. Arutjothi and Dr. C. Senthamarai [13] used R as the programming environment and K-Nearest Classifier along with min-max normalization techniques to perform predictions. They retrieved publicly available data of lending club repository data and they took only ten thousand records for sampling. For pre-processing they split in two for training and testing which was performed randomly, 70 % for training and 30 for testing. They used min-max normalization to normalize the dataset. They also use a random sampling method to balance the dataset. Then for the model, they used K-NN (K-Nearest Neighbor) credit scoring model, which correctly classified the labeled data and unlabeled data with 75.08% accuracy in the test dataset. This proved to be the higher accurate model then other classifiers they researched using R package. Dr. T C Thomas, Dr. J P Sridhar, Dr. M J Chandrashekar, Dr. Makarand Upadhyaya and Dr. Sagaya Aurelia developed a machine learning-powered website for loan prediction [14]. They use a load prediction dataset from Kaggle consisting of credentials like loan id, gender, marital status, dependents, education details, employment details, income details, loan amount, credit history, property details, and loan status. They for pre processing they removed the null values, label encoded the categorical variables, and derived a correlation matrix to find the irrelevant data. For the models, they trained and tested the K-Nearest Neighbor(K-NN), the support vector machine algorithm (SVM), and eXtreme Gradient Boosting (XGBoost). And upon getting the confusion matrix from those models, they were able to conclude that K-NN was the lowest performing with only 85% accuracy. The SVM had the second-highest accuracy with 89.16%. And the XGBOOST had the highest accuracy percentage of 91.66%. Then they went ahead and built a website using HTML (hypertext markup language), which was powered by the XGBoost algorithm, which allowed the users to enter their credentials in order to learn if they are eligible for the loan or not. Suliman Mohamed Fati [15] also developed a machine learning-based loan prediction model using the same Kaggle dataset as Dr. T C Thomas, Dr. J P Sridhar, Dr. M J Chandrashekar, Dr. Makarand Upadhyaya and Dr. Sagaya Aurelia [14] but his approach was much different, and he trained and tested different models. For data preprocessing, he used the heatmap technique for missing feature values discovery, used outlier detection using box plots techniques, and derived the correlation between attributes using the heat map. He then trained and tested the data using Logistic Regression, Decision tree, and Random Forest. The results showed that Logistic regression had better performance with 81% accuracy, while Decision tree and Random Forest got 72% and 76% accuracy, respectively, and was validated using the ROC curve. Ugochukwu .E. Orji, Chikodili .H. Ugwuishiwu, Joseph. C. N. Nguemaleu and Peace. N. Ugwuanyi [16] also developed machine learning models and published their results in their research paper based on the same Kaggle dataset as [14] and. For data preprocessing, they applied Synthetic Minority Oversampling Technique (SMOTE) for data balancing. They used One-hot encoding to convert the categorical features into numerical features, and they also performed normalization. They then performed Exploratory Data Analysis (EDA) to get all the detailed information of the dataset, such as the ratio of males and females and missing data. They then substituted missing data with close estimation and then derived the correlation matrix of the key variables in the dataset. They used evaluation matrices (Confusion Matrix and F1 Score) to explain the performance of the models they used. For models, they trained and tested Logistic Regression (LR) Algorithm, K-Nearest Neighbor (KNN) Algorithm, Support Vector Machine (SVM), Decision Tree (DT) Algorithm, and Bagging and Boosting Algorithm (random forest (RF) Algorithm, Gradient Boost (GBM) Algorithm). Upon looking at the results, they found out that Random Forest was the most accurate model with 95.56% followed by K Neighbor's and Gradient Boost's 93.33%, Decision Tree's 91.11%, SVM's 84.44%, and Logistic Regression's 80% accuracy.

Mohamed Alardi Sawsan Hilal [17] developed a machine learning tree-based method for loan approval. Here to forecast a bank's loan approval status, numerous statistical learning classification techniques were used in this study. The focus was on decision trees, random forests with different variants, and boosting. Because the apparent true function for determining loan status is too complex to be represented in a single decision tree, the decision tree approach failed to establish a thorough and meaningful relationship between the attributes and the loan status. However, this failure was not caused by some violation of the approach's fundamental assumptions. Multiple tree strategies consequently turned out to be the most representative methodologies in this field of study. This included random forests, bagging, and boosting. These techniques work well for modeling this kind of data since they simulate several decision trees and finally compute the common vote. Credit history, income, loan amount, geography, marital status, and education were ranked differently by the three methodologies, from highest to lowest importance. However, amongst the implemented multiple tree methods, boosting came in superior according to selection criteria outlined earlier with an accuracy of 98.75%, specificity of 100%, sensitivity of 92.5%, and AUC of 97%. Vishal Singh, Ayushman Yadav Rajat Awasti [18] has used three machine learning algorithms which are used to find out the best possible prediction of the dataset. Hence after implementing all the methods, it finds the prediction accuracy is suitable for both datasets. When a client experiences a calamity, for example, the algorithm may be unable to forecast the best course of action. This can identify possible clients who can pay back the loan, and its accuracy is good. The key elements in determining (whether the client would have been) include loan duration, loan amount, age, and income. The two most crucial variables for determining the loan applicant's category are their zip code and credit history. Mr. Abhiroop Sarkar has got at an accuracy of 80.78%; logistic regression is the most accurate of the three machine learning algorithms, followed closely by random forest at 79.79% and decision tree with 70.51% on Machine Learning techniques for recognizing the loan eligibility [19]. It prefers logistic regression for loan eligibility. Ideas for integrating other machine learning algorithms, such as XGBoost and others, can be compared based on the findings; research for these algorithms is already in action. Therefore, the model is trained to generate results with acceptable accuracy. After that, it generates accurate results on whether or not to lend money to a borrower without the need for tiresome manual work.

Mohammad Ahmad Sheikh, Amit Kumar Goel, and Tapas Kumar [20] worked on the approval process based on a set of parameters loans were approved. To prevent missing values in the data set, the data is first cleansed. 1500 examples, 10 numerical attributes, and 8 categorical features were used to train the model. On the dataset, best-case accuracy was obtained 0.811. The major drawback of this model is most of the time, it declined the loan where the applicant's credit score was the worst will due to a higher probability of not paying back the loan amount. But that, applicants with high incomes who request smaller loans are more likely to be approved and more likely to repay their loans.

Another work by Ashwini S. Kadam, Shraddha R. Nikam, Ankita A. Aher, Gayatri V. Shelke , Amar S. Chandgude [21] describes a similar type of conclusion. SVM and Naive Bayes(NB) models were implemented, and the NB model was extremely efficient and gives a better result when compared to other models. MIraz Al Mamun, Afia Farjana, and Muntasir Mamun have predicted bank loan eligibility using machine learning models, and comparison analysis [22]. This model is used for the banking system or anyone who wants to apply for a loan. It is obvious from the analysis of the data that it lessens all frauds perpetrated during the loan approval process. The processes in the prediction process include data cleaning and processing, imputation of missing values, experimental analysis of the data set, model creation, and testing on test data. The original data set's best-case accuracy is 0.9189 on the Data set. After analyzing all the data, it found that the lowest credit scores will be denied a loan because they have a higher risk of defaulting on the loan. Since they are more likely to repay their obligations, candidates with high incomes and smaller loan requests are typically more likely to be granted. The system is trained using the prior training data, but it is possible to alter the software in the future so that it may accept new testing data as well as training data and predict as necessary.

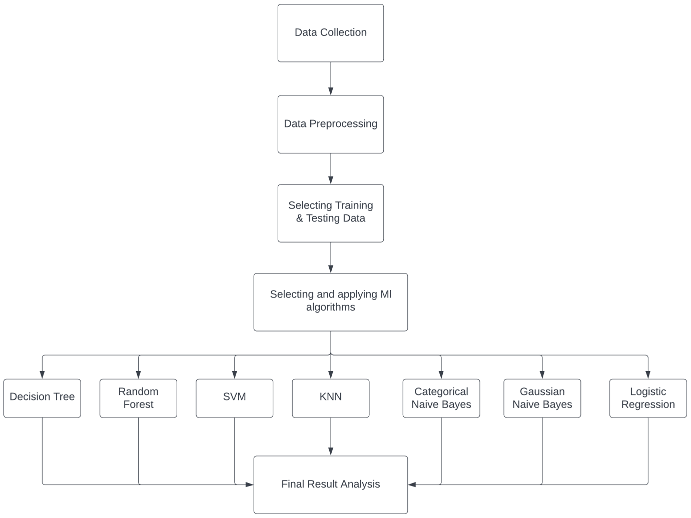


Fig. 1. Project Workflow

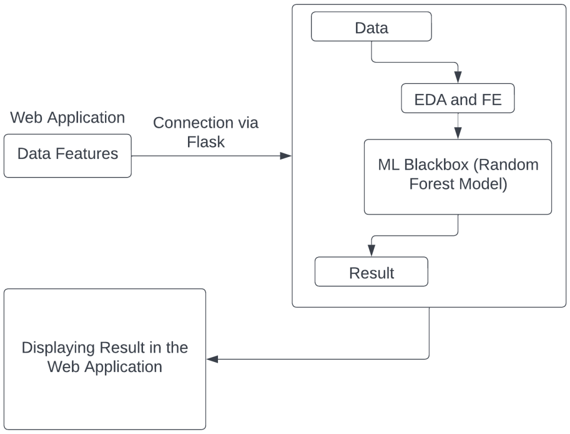

Web Application
Data Features
Connection via Flask
Data
EDA and FE
ML Blackbox (Random Forest Model)
Result
Displaying Result in the Web Application

Fig. 2. Software Architecture

III. Methodology

The methodology section of the study provides a detailed overview of the approaches and procedures used. The flowchart in the following section illustrates the methodology and development process. A visual summary of the methodology can also be found in the accompanying figures:

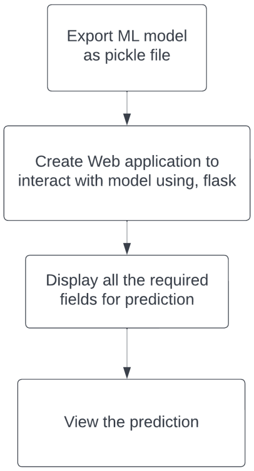

Export ML model as pickle file
Create Web application to interact with model using, flask
Display all the required fields for prediction
View the prediction

Fig. 3. Web application process

A. Mathematical explanation of the model that has given the best result:

Random Forest: A machine learning method called Random Forest is a component of the supervised learning approach. It can be applied to classification and regression issues. It is derived from ensemble learning, a technique that combines several classifiers to tackle complicated problems and improve the performance of the model. Numerous individual decision trees seen in random forests function as an ensemble. Each tree in the forest gives out a prediction, and the class that has the maximum number of outcomes becomes the model's prediction [24]. Many relatively uncorrelated models running together will perform better than any one of the constituent models alone. Between the models, there is little correlation. Uncorrelated models can generate forecasts in an ensemble that are more accurate than any single prediction. The following are necessary for random forests to operate more effectively: To perform better than random guessing, models must be developed utilizing features that have actual signals. The predictions (and errors) made by each tree should have little association with one another. The greater the number of trees in the forest, the higher the accuracy, and thus, it prevents the situation of overfitting [25]. Random forest is a better algorithm because it takes less training time than other algorithms, and it increases the accuracy even for large datasets [26]. Like bagging, random forests construct several decision trees using bootstrap samples. The distinction is that a random sample of predictors from the entire collection of predictors is picked as split candidates each time a split in a tree is considered by de-correlating the built-in decision trees; this strategy has the advantage of lowering the variance when averaging the trees. So basically, Random Forest stands for bagging instead of boosting. Therefore, we have got 85% accuracy through Random Forest.

Bagging: An example of an ensemble method is bagging. A machine learning technique called an ensemble method combines multiple base models to create a single, ideal predictive model. Considering this definition, bagging builds several complete classification trees using a bootstrap sample. The bagging approach computes each built tree's forecast when a new observation is received, and a prediction is required, then aggregates the results into a majority vote. The overall model prediction is the class that the trees predicted with the highest number of votes. The primary reason why the trees might grow in the building phase without being restricted or pruned without concern for a variance increase is that this majority vote technique naturally lowers the variance of decision trees.

Variable Name	Description	Type
Loan_ID	Unique Loan_ID	Integer
Gender	Male/Female	Character
Married	Applicant	Married(Y/N) Character
Dependents	Number of dependents	Integer
Education	Graduate/Not Graduate	String
Self_Employed	Self Employed	(Y/N) Character
Applicant_Income	Applicant income	Integer
Co_Applicant_Income	Co-Applicant income	Integer
Loan_Amount	Loan amounts in thousands	Integer
Loan_Amount_Term	Term of the loan in months	Integer
Credit_History	Credit history guidelines	Integer
Property_Area	Urban/ Semi Urban/ Rural	String
Loan_Status	Loan Approved(Y/N)	Character

TABLE I
DATASET

B. Dataset Description and Prprocessing

Data collection is the process of gathering and analyzing information on intended modifications to an existing system so that pertinent questions can be answered and outcomes can be assessed. The goal of all data collecting is to gather reliable information that can be used for analysis and to create misleading, concrete responses to the questions that are presented. The dataset has a total of 814 rows and 13 columns, where the dataset has been divided into train dataset and test dataset. Shows in Table 1

1) Data Pre-Processing: The data processing techniques covered data transformation and missing data imputation following the approach adopted in [23]. We can handle the missing data by removing instances that contain missing data in pandas. Removing the entire attribute with missing data and setting values to some statistical measure (zero, mean, median) can handle the missing data. Therefore, after implementing those steps (dropping null values), the dataset has got zero missing data. For the few missing values, we will use the Mean Imputation technique to estimate the missing values. For example, in the gender feature, we found 13 missing values in the training part and 11 missing values in the testing part. So, we must apply mean imputation here to handle the missing data. Then we plotted all the bar plots. Therefore, in figure :4, the relation between gender and loan status is shown where about 410 Males have got the loan and 150 females have got the loan. On the other hand, 160 Males didn't get the loan, and 50 Females didn't get the loan.

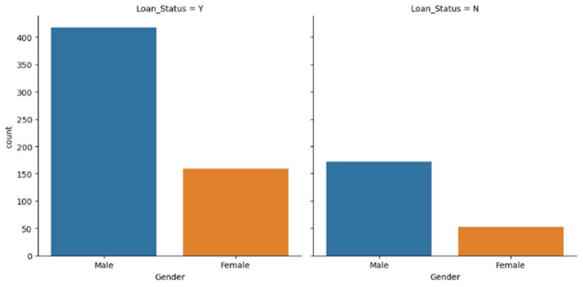


Fig. 4. Relation between gender & loan_status

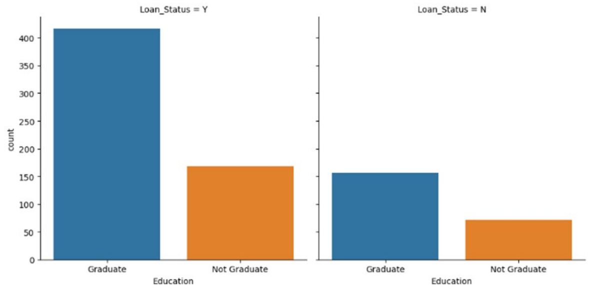


Fig. 5. Relation Between Education and Loan_Status

Likewise, in figure 5, the relation between education and loan status is shown. Where the no. of graduate and not graduate who got the loan and who didn't get the loan has been shown.

Testing the correlation between data properties is the final phase in the pre-processing process, which aims to identify the most notable aspect of the prediction process. To see the correlation between the variables for this purpose, we utilize a heat map. The heat map for the data collection properties is in figure 7. The most crucial element for loan prediction is immediately apparent from the heat map. Loan_ID has been deleted from the heat map, which is noteworthy because it has no bearing on the prediction procedure.

C. Website

Screenshots of website for this model Showing the prediction result which is YES for this sample input.Figure: 9

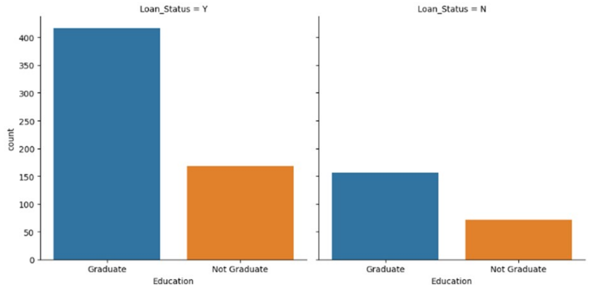


Fig. 6. Relation Between Education and Loan_Status

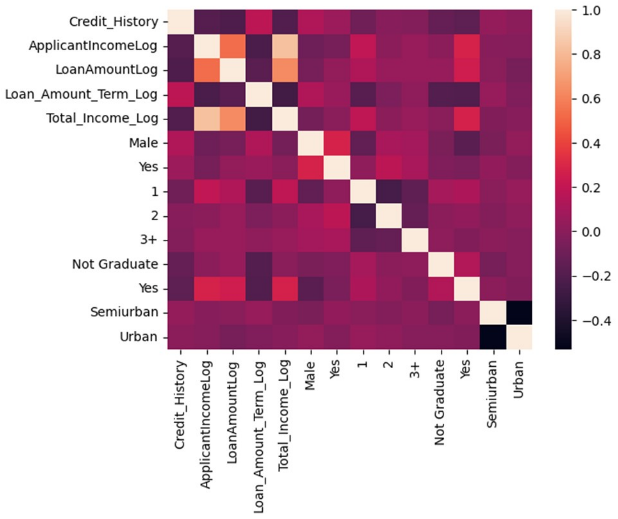
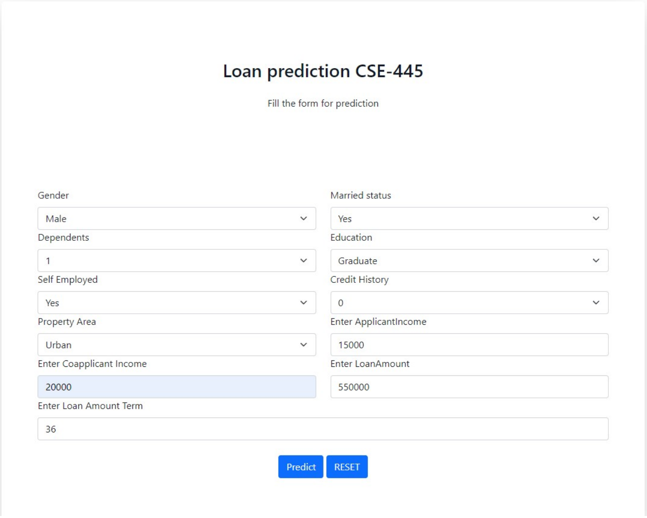


Fig. 7. Representing the correlation between attributes using the heat map



Loan prediction CSE-445

Fill the form for prediction

Gender: Male

Married status: Yes

Dependents: 1

Education: Graduate

Self Employed: Yes

Credit History: 0

Property Area: Urban

Enter ApplicantIncome: 15000

Enter Coapplicant Income: 20000

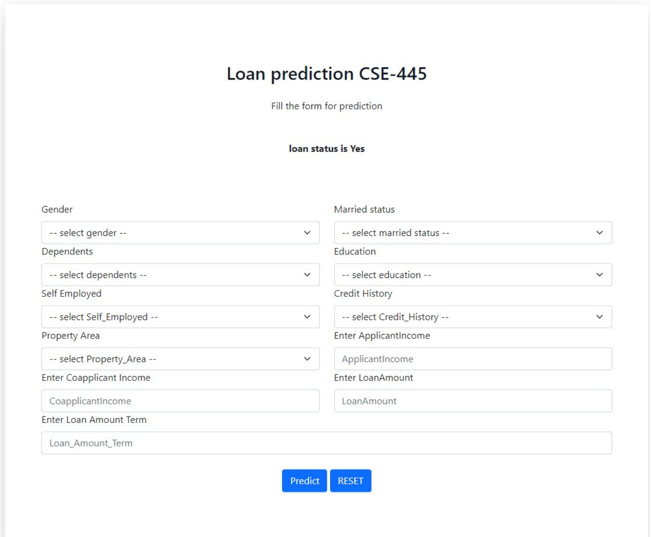
Enter LoanAmount: 550000

Enter Loan Amount Term: 36

Fig. 8. Sample input test case

D. Explanation of the software architecture

- Web Application: This is the frontend of our software architecture built using hypertext markup language where data features required for loan eligibility prediction are taken as inputs and the predictions (Loan approved or not) are given as output.
- Flask: Flask is a micro web framework using which we first load the model that we saved then we pass the data into the model and get the predictions as output. Within flask we one hot encode the categorical features to pass into the model for the model to work and produce predictions as output.
- ML Blackbox (Random Forest): Within this ML Blackbox our model (Random Forest) gets the data already prepared and uses the input data and input then uses the saved model weights to predict whether the person who


Loan prediction CSE-445

Fill the form for prediction

loan status is Yes

Gender: -- select gender --

Married status: -- select married status --

Dependents: -- select dependents --

Education: -- select education --

Self Employed: -- select Self_Employed --

Credit History: -- select Credit_History --

Property Area: -- select Property_Area --

Enter ApplicantIncome: ApplicantIncome

Enter Coapplicant Income: CoapplicantIncome

Enter LoanAmount: LoanAmount

Enter Loan Amount Term: Loan_Amount_Term

Fig. 9. Sample input test case output

input the data is going get a loan or not. Then this output is displayed onto the web application

IV. Results and Discussion

A. Definition of evaluation parameters

Precision, recall, and F1 score are all metrics used to evaluate the performance of a machine learning model.

- Precision: Precision is the proportion of true positive predictions made by the model out of all positive predictions made by the model. In other words, it is the ratio of true positive predictions to the total number of positive predictions. For example, if a model makes 100 positive predictions, and 90 of those predictions are correct, then the precision of the model is 90%. Precision = True Positives / True Positives + False Positives
- Recall: Recall is the proportion of true positive predictions made by the model out of all actual positive cases. In other words, it is the ratio of true positive predictions to the total number of actual positive cases. For example, if there are 1000 actual positive cases, and the model correctly predicts 900 of them, then the recall of the model is 90%. Recall = True Positives / True Positives + False Negatives
- F1 Score: F1 score is a metric that combines precision and recall. It is calculated as the harmonic mean of precision and recall. The F1 score is a useful metric when you want to balance precision and recall. A model with a high F1 score is a model that has a good balance between precision and recall. F1 Score = 2 * Precision * Recall / Precision + Recall

In summary, precision is a measure of the accuracy of the model's positive predictions, recall is a measure of the model's ability to find all the positive cases, and F1 score is a balance between precision and recall.

B. Model Results Table

The tables in this section present the machine learning models that were utilized and their corresponding accuracy results. These results provide insight into the effectiveness of each model in terms of predicting the outcome of training and testing accuracy.

Deployed Model	Train accuracy (%)	Test accuracy (%)
Logistic Regression	80.18	79.14
Decision Tree	100	74.84
Random Forest	100	84.66
SVM	72.35	69.93
KNN	82.18	76.07
Categorical Naive Bayes	76.19	77.3
Gaussian Naive Bayes	78.64	78.52

TABLE II
ACCURACIES FOR EACH MODEL

Deployed Model	Target V	Precision	Recall	f1-score
Logistic Regression	0	0.89	0.35	0.5
	1	0.78	0.98	0.87
Decision Tree	0	0.58	0.57	0.58
	1	0.82	0.82	0.82
Random Forest	0	0.88	0.57	0.69
	1	0.84	0.96	0.9
KNN	0	0.73	0.33	0.45
	1	0.77	0.95	0.85
Categorical Naive Bayes	0	0.8	0.33	0.46
	1	0.77	0.96	0.86
Gaussian Naive Bayes	0	0.79	0.39	0.52
	1	0.78	0.96	0.86

TABLE III
ACCURACY METRICS FOR EACH MODEL

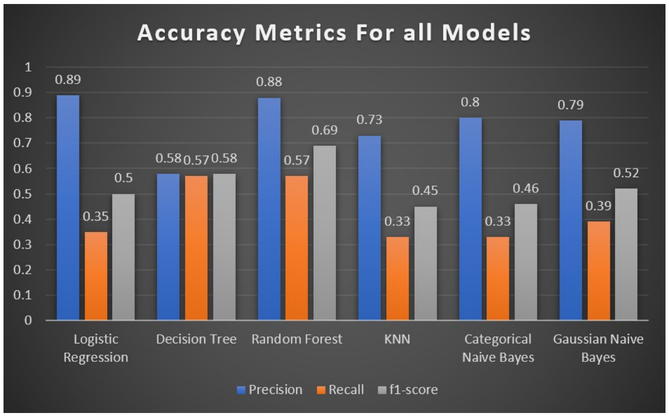


Fig. 10. Accuracy Metrics for models for target value 0

The Performance metrics are used to tell us about the measure of performance across all possible classifications. Here using precision, recall, and f1-score we can observe the model's performance. In figure 7 after setting the target value to 0, we can see that logistic regression has got the highest precision accuracy on the other hand Decision tree has the lowest precision accuracy. Random Forest and Decision tree have the highest recall accuracy and KNN Categorical Naïve Bayes have the lowest recall accuracy. About F1-score

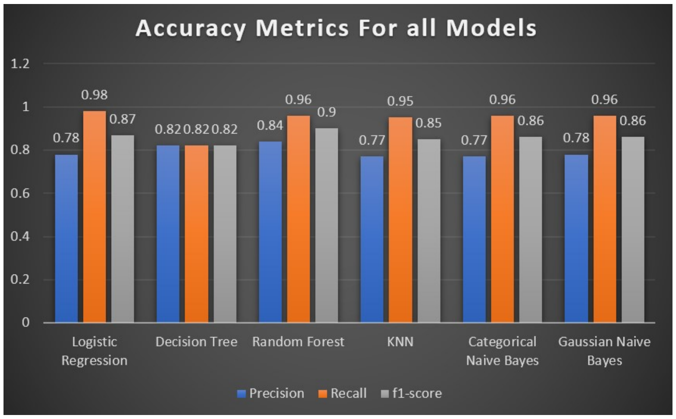


Fig. 11. Accuracy Metrics for models for target value 1

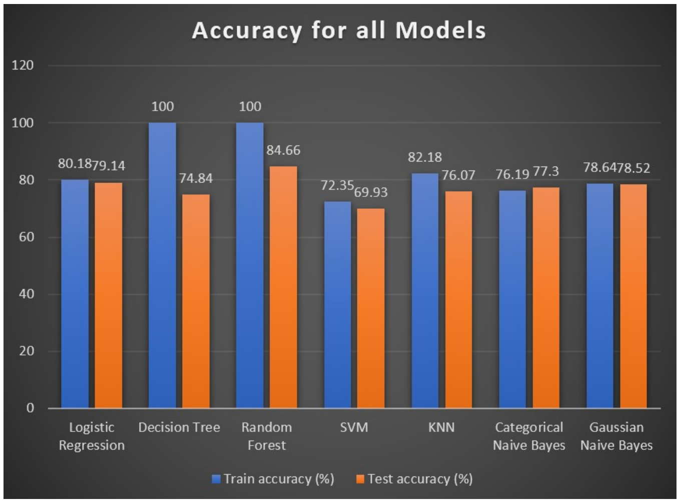


Fig. 12. Train Test Accuracy

Random Forest has got the highest f1 score accuracy and KNN has got the lowest accuracy. In Figure 8 after setting the target value 1 we can observe some changes where Random Forest has got the best accuracy for both precision and f1 score but logistic regression has the best recall accuracy. Therefore, the lowest accuracy observed for precision, recall, and f1 score are KNN & Categorical Naïve Bayes, Decision Tree & Decision Tree respectively. In Figure 9 We can observe the accuracy rate for both training and testing. Here Decision Tree Random Forest both have got the highest training accuracy which is 100% KNN has got the lowest training accuracy which is 72%. On the other hand, Random Forest has got the highest testing accuracy which is 84% and KNN has got the lowest training accuracy which is 69%. We have used seven models where we have got different values from each model. However, from the graphs, we can easily conclude a result. The seven models are Logistic Regression, Decision Tree, Random Forest, SVM, KNN, Categorical Naïve Bayes, Gaussian Naïve Bayes. From those seven models, we have got Random Forest as the best model for predicting the data.

V. Conclusion and Future Work

Loan granting require a lot of resources such as hiring experienced underwriters and necessary office space and equipment. In the literature, many machine learning based approaches are proposed based on different datasets, factors and parameters. In this paper we contributed to a public dataset. In the dataset, there are thirteen features and from the correlation matrix we found out that Credit_History was the most important feature for the loan eligibility prediction. For the preprocessing EDA was performed to understand the data. Then various method such as one hot encoding, normalization was performed to make the dataset trainable and also to clean the noise of the dataset. Then the data was trained and tested using seven machine learning model: Logistic Regression, Decision Tree, Random Forest, SVM, KNN, Categorical Naïve Bayes and Gaussian Naïve Bayes. Out of the seven machine learning models, Random Forest the best result with an accuracy of 84.66 percent, while Logistic Regression was 79.14 percent accurate, Decision Tree was 74.84 percent accurate, SVM was 69.93 percent accurate, KNN was 76.07 percent accurate, Categorical Naïve Bayes was 77.3 percent accurate and Gaussian Naïve Bayes was 78.52 percent accurate. The model was then compared with the related works. This Random Forest model was then exported and used with a hypertext markup language-based web interface to predict loan eligibility from user data.

The future work could be to acquire the user data and prediction to into the model with the help of online learning to make the model more robust and accurate. Also new features can be added to the website, such as a client registration or login system. A client database can be added to the website to keep the client's history and data.

References

[1] A. Kumar, S. Sharma, & M. Mahdavi, ”Machine Learning (ML) Technologies for Digital Credit Scoring in Rural Finance: A Literature Review.” Risks 9.11 (2021): 192.

[2] G. Dorfleitner, E.M. Oswald, & R. Zhang. ”From Credit Risk to Social Impact: On the Funding Determinants in Interest-Free Peer-to-Peer Lending.” J Bus Ethics. 2021 Vol.170, pp. 375–400.

[3] Toby Segaran, “Programming Collective Intelligence: Building Smart Web 2.0 Applications.” O'Reilly Media.

[4] M. A. Chowdhury, A. Apon, and K. Dey, Data Analytics for Intelligent Transportation Systems. Amsterdam: Elsevier, 2017.

[5] P. Ranjan, K. Santhosh, A. Kumar and S. Kumar, ”Fraud Detection on Bank Payments Using Machine Learning,” 2022 International Conference for Advancement in Technology (ICONAT), 2022, pp. 1-4, doi: 10.1109/ICONAT53423.2022.9726104.

[6] M. S. Thekkethil, V. K. Shukla, F. Beena and A. Chopra, ”Robotic Process Automation in Banking and Finance Sector for Loan Processing and Fraud Detection,” 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2021, pp. 1-6, doi: 10.1109/ICRITO51393.2021.9596076.

[7] A. Gupta, V. Pant, S. Kumar and P. K. Bansal, ”Bank Loan Prediction System using Machine Learning,” 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART), 2020, pp. 423-426, doi: 10.1109/SMART50582.2020.9336801.

[8] C. Naveen Kumar, D. Keerthana, M. Kavitha and M. Kalyani, ”Customer Loan Eligibility Prediction using Machine Learning Algorithms in Banking Sector,” 2022 7th International Conference on Communication and Electronics Systems (ICCES), 2022, pp. 1007-1012, doi: 10.1109/ICCES54183.2022.9835725.

[9] 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), 2020, pp. 490-494, doi: 10.1109/ICESC48915.2020.9155614.

[10] “How Automation Can Improve Your Loan Origination Process,” www.moodysanalytics.com.

[11] “Loan Prediction Using Machine Learning: How Much Does It Really Cost?,” GiniMachine, Sep. 29, 2021. https://www.moodysanalytics.com/articles/2018/maximize-efficiency-how-automation-can-improve-your-loan-origination-process

[12] I. Publication, “Bank Loan Approval Prediction Using Data Science Technique (ML,” International Journal for Research in Applied Science Engineering Technology (IJRASET), Jan. 2022, Accessed: Dec. 19, 2022. [Online].

[13] G. Arutjothi and C. Senthamarai, ”Prediction of loan status in commercial bank using machine learning classifier,” 2017 International Conference on Intelligent Sustainable Systems (ICISS), 2017, pp. 416-419, doi: 10.1109/ISS1.2017.8389442.

[14] T. Thomas, J. Sridhar, M. Chandrashekar, M. Upadhyaya, and S. Aurelia, “Developing a website for a bank’s Machine Learning- Based Loan Prediction System,” International Journal of Mechanical Engineering, vol. 7, no. 3, 2022, Accessed: Dec. 19, 2022.

[15] S. M. Fati and S. M. Fati, “Machine Learning-Based Prediction Model for Loan Status Approval,” Journal of Hunan University Natural Sciences, vol. 48, no. 10, 2021, [Online]. Available: Machine Learning-Based Prediction Model for Loan Status Approval

[16] U. E. Orji, C. H. Ugwuishiwu, J. C. N. Nguemaleu and P. N. Ugwuanyi, ”Machine Learning Models for Predicting Bank Loan Eligibility,” 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON), 2022, pp. 1-5, doi: 10.1109/NIGERCON54645.2022.9803172.

[17] Mohamed Alaradi and Sawsan Hilal, “Tree-Based Methods for Loan Approval”, 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), doi: 10.1109/ICDABI51230.2020.9325614

[18] Vishal Singh, Ayushman Yadav Rajat Awasthi, ”Prediction of Modernized Loan Approval System Based on Machine Learning Approach”, 2021 International Conference on Intelligent Technologies (CONIT), doi: 10.1109/CONIT51480.2021.9498475

[19] Mr. Abhiroop Sarkar, ”MACHINE LEARNING TECHNIQUES FOR RECOGNIZING THE LOAN ELIGIBILITY”, Volume-03 Isuue-12, December-2021

[20] 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), 2020, pp. 490-494, doi: 10.1109/ICESC48915.2020.9155614.

[21] I. R. J. E. T. Journal, “Prediction for loan approval using machine learning algorithm,” IRJET, 15-Sep-2021. [Online].

[22] M. F. AKÇA and O. SEVLİ, “Predicting acceptance of the bank loan offers by using support Vector Machines,” International Advanced Researches and Engineering Journal, 15-Aug-2022. [Online].

[23] X. Francis Jency, V.P. Sumathi, and J. Shiva Sri, “An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients” International Journal of Recent Technology and Engineering, vol. 7, 2018.

[24] S. Rasoul Safavian and David Landgrebe, ”A Survey of Decision Tree Classifier Methodology”, Transactions on systems, man and cyber netics, vol. 21, No. 3, May/June 1991

[25] Anthony J. Myles, Robert N. Feudale, Yang Liu, Nathaniel A. Woody, Steven D. Brown, “An introduction to decision tree modeling”, Journal of Chemometrics, 2004, DOI: https://doi.org/10.1002/cem.873

[26] M. Pal, “Random Forest classifier for remote sensing classification”, International Journal of Remote Sensing, 2007, DOI: https://doi.org/10.1080/01431160412331269698