ABSTRACT

As the lending industry continues to evolve, it's more important than ever to have accurate loan eligibility predictions. The aim of this project is to develop a loan eligibility prediction model using machine learning algorithms. The dataset is downloaded from Kaggle and preprocessed to remove any inconsistencies or irrelevant features. The analysis is done by checking whether the user is defaulted or not to take the loan. Different preprocessing technique such as data discretization, and label encoding is used to change continuous and categorical data to numerical categorical data. The data initially obtained was not balanced so the SMOTEENN technique is used to balance the dataset. It uses machine learning algorithms such as Decision Tree, Random Forest, AdaBoost, Gradient Boost, and XG Boost. All five models are compared on their accuracy and the best model is chosen. Hyperparameter tuning is applied to the best-performing model to further improve its accuracy and other metrics like precision, recall, and f1-score. All these training processes are tracked using mlflow. The trained model is then deployed on a web app built using Django, which includes login, signup, prediction, and analysis module for customers as well as model training, application update, and login modules for admins. The proposed system aims to predict loan eligibility by analyzing factors such as age, income, age, experience, profession, city, state, etc. The system is expected to provide accurate predictions and improve the overall loan approval process.

Keywords:

Loan Eligibility Prediction, Machine Learning, Ensemble, Bagging, Boosting, Decision Tree, Classification, Prediction, Web-based application

INTRODUCTION

Introduction

The loan is the process of lending money from one individual or organization to another individual or organization. Loans are of different types according to needs and denominations. Loan issuance is one of the riskiest functions for any bank and financial institution because of the creditworthiness of the customer. Loan eligibility prediction is the estimation of the tendency of the customer whether they are eligible to take the loan as per their status and properties. In the past few years, there has been huge growth in the number of people taking a loan from financial institutions or banks. Financial institutions take much information such as income, properties, and salary to be on the safe side. Customers should meet all the requirements and regulations from the institutions to take the loan. For this project, the dataset [1] belongs to the Hackathon organized by “Univ.AI”. This data set contains 280000 instances,11 input features along with 1 target feature. This dataset contains information regarding the people who are eligible for loans or not. It includes features like income, age, profession, experience, etc. This dataset helps to predict the status of people who are eligible to take the loan or not.

Problem Statement

The problem statement of loan eligibility prediction is to develop a machine learning model that can accurately predict whether a loan applicant is eligible for a loan based on various factors such as income, employment history, property, salary, and other relevant financial and personal information. Manual evaluation of loan eligibility based on customer behavior can be time-consuming and prone to errors. The goal is to develop a model that can assist lending institutions in making informed and accurate loan decisions, reducing the risk of loan defaults, improving customer satisfaction, and optimizing the loan portfolio. The loan eligibility prediction model should be able to analyze large amounts of data, identify patterns and trends, and provide insights and recommendations to lending institutions. It should be accurate, reliable, and transparent, and should comply with relevant regulations and ethical standards.

Objectives

The major objectives of this project are as follows:

- To implement an algorithm to predict the customers who are eligible to take loans from financial institutions.
- To train multiple machine learning models and choose the best model from them.
- To provide a positive customer experience.

Literature Review

Goyal and Kaur developed a loan prediction model using various machine learning algorithms in 2017. The dataset with features mainly gender, marital status, education, number of dependants, employment status, and income, loan amount, loan tenure, credit history, existing loan status, and property area. Various ML models adopted in the present method include the Linear model, Decision Tree (DT), Neural Network (NN), Random Forest (RF), SVM, Extreme learning machines, Model tree, Multivariate Adaptive Regression Splines, Bagged Cart Model, NB and Thermogravimetric analysis (TGA). When evaluating 9 of these models using Environment in five runs, TGA resulted in better loan forecasting performance than the other methods. [2]

Several organizations find it beneficial for describing the threshold or lending cut-off as a directive for screening the loans. Here, it is problematical to obtain information regarding the economic value and consequent usage policies from statistical information utilized for reporting the performance of the model. Various financial organizations actively scrutinize loans of different kinds considering certain factors that involve terms, default criteria, age, and lenders from different areas. The prediction of total loans becomes complex if the risk linked with the givenloan is predicted effectively, then the agencies can perform suitable communication and action for preventing loan default if it is effectively predicted. For evaluating the financial and banking industries, a variety ofalgorithms are used. Some of them are discussed. The BCO is a stochastic, random-search technique. This technique uses an analogy between the way in which bees in nature searchfor food, and the way in which optimization algorithms search for an optimum of combinatorial optimization problems. The basic idea behind the BCO is to build a multiagent system (a colony of artificial bees) able to effectively solve difficult combinatorial optimization problems. [3]

Loan eligibility prediction models often involve a combination of machine learning or deep learning techniques, ensemble learning approaches, feature engineering, and explainable AI techniques. The choice of the best model may depend on the specific dataset, problem requirements, and available resources. Further research in this field continues to explore new approaches and techniques to improve the accuracy, interpretability, and reliability of loan eligibility prediction models.

Methodology

This section describes the methodology adopted for the loan eligibility prediction system developed in this project. The methodology follows an **iterative approach** that includes four main iterations: data preprocessing, model training, and comparison, hyperparameter tuning, and web app development and deployment.

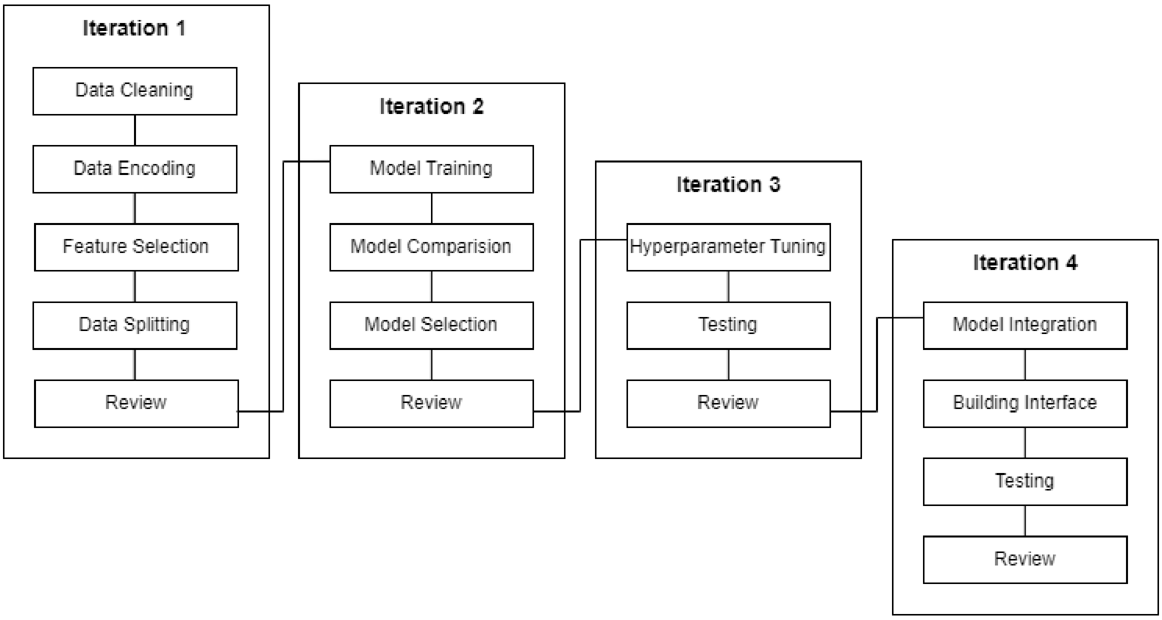


Figure 1: Development Methodology

1. Data Preprocessing

The first iteration involved the preprocessing of the loan eligibility dataset downloaded from Kaggle. The dataset was first analyzed to remove any inconsistencies or irrelevant features. The continuous and categorical data were then converted to numerical categorical data using data discretization and label encoding techniques. The dataset was found to be imbalanced, with more non-defaulters than defaulters. To address this imbalance, the Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbor (SMOTEENN) technique was used to balance the dataset.

2. Model Training and Comparison

In the second iteration, the loan eligibility prediction models were trained and compared. The machine learning algorithms used for this iteration were Decision Tree, Random Forest, AdaBoost, Gradient Boost, and XG Boost. The training process was tracked using mlflow. The models were compared based on their accuracy, precision, recall, and f1-score. The best-performing model was chosen based on these metrics.

3. Hyperparameter Tuning

The third iteration involved hyperparameter tuning for the best-performing model. The hyperparameters for the chosen model were optimized to further improve its accuracy, precision, recall, and f1-score. The metrics used to evaluate the tuned model included area under the curve (AUC), receiver operating characteristic (ROC) curve, and confusion matrix. The results of the hyperparameter tuning process were analyzed to identify areas for improvement.

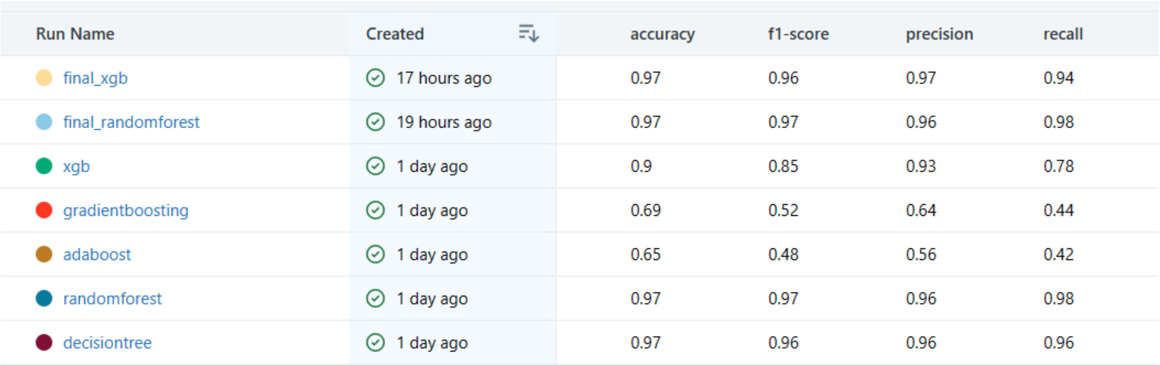
4. Web App Development and Deployment

The final iteration involved the development and deployment of the loan eligibility prediction system as a web application using the Django framework. The web app includes modules for user login, signup, prediction, and analysis, as well as admin modules for application update, and login. The challenges faced during this iteration included integrating the machine learning model with the web app and testing the system on different browsers and devices. The results of this iteration were analyzed to identify areas for improvement.

In conclusion, the iterative approach adopted in this project helped to develop an accurate and reliable loan eligibility prediction system. The methodology involved preprocessing the data, training and comparing the models, hyperparameter tuning, and web app development and deployment. The results of each iteration were analyzed to identify areas for improvement, and the success of the methodology is reflected in the performance of the final system.

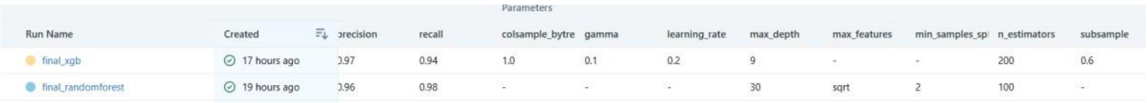
Results and Discussion

The result we obtained from the trained model are:



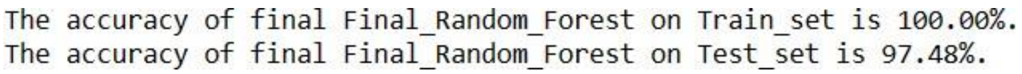
Run Name	Created	accuracy	f1-score	precision	recall
● final_xgb	✓ 17 hours ago	0.97	0.96	0.97	0.94
● final_randomforest	✓ 19 hours ago	0.97	0.97	0.96	0.98
● xgb	✓ 1 day ago	0.9	0.85	0.93	0.78
● gradientboosting	✓ 1 day ago	0.69	0.52	0.64	0.44
● adaboost	✓ 1 day ago	0.65	0.48	0.56	0.42
● randomforest	✓ 1 day ago	0.97	0.97	0.96	0.98
● decisiontree	✓ 1 day ago	0.97	0.96	0.96	0.96

With parameters:

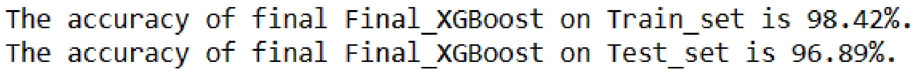


				Parameters							
Run Name	Created	precision	recall	colsample_bytre	gamma	learning_rate	max_depth	max_features	min_samples_sp	n_estimators	subsample
● final_xgb	✓ 17 hours ago	0.97	0.94	1.0	0.1	0.2	9	-	-	200	0.6
● final_randomforest	✓ 19 hours ago	0.96	0.98	-	-	-	30	sqrt	2	100	-

From the above result, random forest is the better model, but if we look into the accuracy of train and test data:



```
The accuracy of final Final_Random_Forest on Train_set is 100.00%.
The accuracy of final Final_Random_Forest on Test_set is 97.48%.
```



```
The accuracy of final Final_XGBoost on Train_set is 98.42%.
The accuracy of final Final_XGBoost on Test_set is 96.89%.
```

Although the random forest model gave higher accuracy, it seemed like overfitting, so we chose the XG-Boost model, which has less variance between the train set and the test set.

The confusion matrix for this model is:

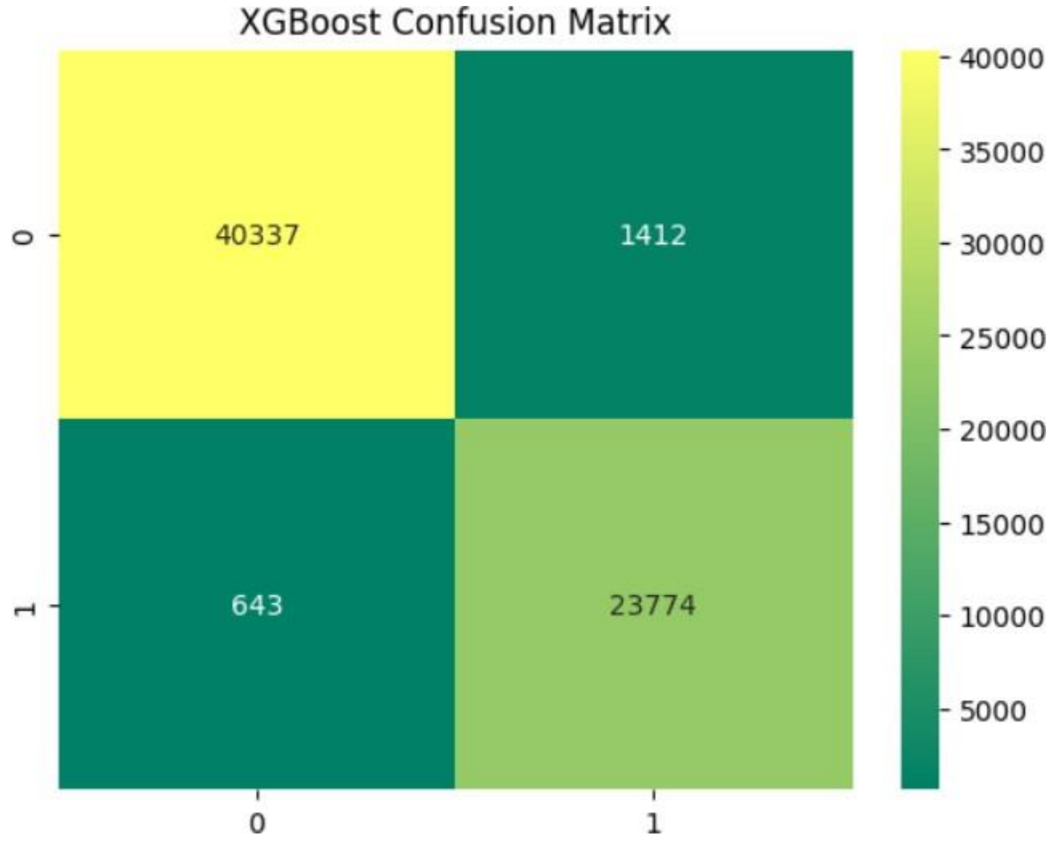


Figure: Confusion matrix

Conclusion

By analyzing customer information such as income, age, experience, ownership, profession, city, and state, institutions can use these models to make informed decisions about approving or denying loan applications. These models can also help enhance transparency and fairness in lending practices by providing explanations for loan approval or denial decisions to borrowers.

Limitation

- Many loan eligibility prediction models use complex algorithms that can be difficult to understand for both parties.
- It may not be able to address the constant economic changes.
- The plans and policies of financial institutions may change over time, which can affect the accuracy of the loan eligibility prediction model.

Future Recommendation

There is still a wide scope of enhancements that can be done to the project. The following are some of the recommendations:

- One potential recommendation for this project is to explore the addition of more features to the loan eligibility prediction model that may be relevant to loan approval, such as credit score, debt-to-income ratio, and employment history.
- It is important to regularly update the model with new data to maintain accuracy and effectiveness.
- Ongoing monitoring of the model's performance can help identify any limitations or areas for improvement, allowing for continual refinement of the model.

References

[1] S. Surana, 2021. [Online]. Available: https://www.kaggle.com/datasets/subhamjain/loan-prediction-based-on-customer-behavior.

[2] Ambika Biradar and Santosh Biradar, "IJARSCT," 1 May 2021. [Online]. Available: https://ijarsct.co.in/Paper1165.pdf.

[3] A. J. Patrick, "Loan Eligibility Prediction Using Adaptive Hybrid Optimization," *Expert System,* vol. 224, 2023.