Scientia Africana, Vol. 22 (No. 3), December, 2023. Pp 137-144
© Faculty of Science, University of Port Harcourt, Printed in Nigeria *ISSN 1118 – 1931*

DESIGN AND IMPLEMENTATION OF A LOAN DEFAULT PREDICTION SYSTEM USING RANDOM FOREST ALGORITHM

[1]Oghenekaro, L. U., and [2]Chimela, M. C.
[1,2]Computer Science Department, Faculty of Computing, University of Port Harcourt, Nigeria
Emails: linda.oghenekaro@uniport.edu.ng, cmaxwell002@uniport.edu.ng

Received: 20-09-2023
Accepted: 01-11-2023

https://dx.doi.org/10.4314/sa.v22i3.12

Journal Homepage: http://www.scientia-african.uniportjournal.info
Publisher: *Faculty of Science, University of Port Harcourt.*

ABSTRACT

Loan default prediction is a crucial task in the lending industry; it helps financial institutions make informed decisions about granting loans. It is usually a daunting task for the bank or financial institution to predict customers who will default on a loan especially when there are thousands of applicants. This loan default prediction system aimed to improve the Area Under the Curve (AUC) score. This loan default prediction system used various data sources, such as demographic information, credit history, and financial performance to predict the likelihood of a loan being defaulted. The system used a random forest (RF) machine learning algorithm to analyze the data and build predictive models. The model was then used to make predictions about new loan applicants and existing borrowers who may default in the future. The system can be customized to meet the specific requirements of different lending institutions. The system enables lenders to make better decisions on loan approval, interest rate determination, and credit risk, management. The loan default prediction system also provides insights into risk factors that contribute to loan default and helps lenders develop effective strategies to mitigate these risks, making it an indispensable tool for lenders. The resultant system achieved an improved AUC score of 98%.

Keywords: AUC score, Loan Default, Loan Processing, Predictive Model, Random Forest Algorithm

INTRODUCTION

Loan processing is a crucial issue faced by banks in recent years. It is a way of checking if a customer will default on a loan in the process of repayment, and this knowledge will determine if a loan should be granted to the customer or not. Many financial institutions or banks approve and disburse loans following a long authentication and validation process, but there is no assurance that the selected candidate is the most eligible of all applicants (Purohit et al., 2011). Through this process, the bank minimizes the losses that could be incurred from defaults, hence increasing the profit generate from the interest from the loan. Loans produce the largest income but constitute a huge risk and exposure. In order to fund viable projects, banks mobilize deposits and create loans. When loans are of good quality, they generate revenue for the bank and at the same time help to stimulate economic growth (Hussain and Shorouq, 2014). In finance, a loan is the lending of money by one or more individuals, organizations, or other entities to individuals etc. In a lot of instances, the lenders

usually add some charges called interest to the amount borrowed which the debtor must pay while repaying the amount borrowed. The repayment of this loan by the debtor is usually within a fixed time frame maybe months or weeks. At times the debtors do not pay their loan as at when due, resulting in a loan default. This leads to loss of money on the part of the lenders due to the fact that the debtors might end up not paying part of the loan taken. Loan defaulting is a major financial risk for the finance industry as it harms the interest of the financiers and destroys social trust (Twala, 2010). Due to loss on the part of the lender (usually financial institutions) there has been efforts to forecast the outcome of a loan before approval to curb instances of bad debts. In recent times, with the introduction of new technologies data is been generated with every click, and data scientists have been researching and making progress in the finance and banking field (Hamid and Ahmad, 2011). Research has been carried out to build systems that will predict if a customer will pay back his/her loan on time. Before now when the applicant filled out a form to get a loan from the bank, the customer's credit score history was usually analyzed by the loan officers together with other things like the amount to be loaned, the salary of the applicant, reason for applying for loan, amount in the bank currently, and also if the customer is on any loan when he is applying for the new loan, with all this process it was usually time consuming and tasking, especially when the number for loan applicants are more. Currently with a lot of data been generated on daily basis and with the aid of machine learning algorithms, the processing of loan gets faster and more efficient, saving losses as incidence of bad debts are reduced. The traditional system becomes slow as compared to what the speed, and accuracy we could get with the help of machine learning.

LITERATURE REVIEW

Almamun et al. (2022) adopted six different machine learning (ML) algorithms to predict if a loan applicant is eligible. The ML algorithms include Random Forest, Adaboost, XGBoost, Decision Tree, K-Nearest Neighbor and Lightgbm. The algorithms were trained with secondary data obtained from kaggle website, the dataset contained 10,128 applicants, 23 attributes and 1 class attribute. The data was preprocessed using missing value handling, feature extraction and categorical variables transformation. The adopted the hold-out approach to validate the dataset, where 70% of the data where for training the algorithm and 30% was for testing. With this approach, the performance of all six machine learning algorithms that were adopted for the work were evaluated under the metrics of precision, accuracy, recall, F1-Score and Area Under Curve (AUC). Of which the Lightgbm recorded the highest accuracy with a score of 0.9189, and decision tree had the lowest accuracy score of 0.8497. In addition to the evaluation of the models in terms of accuracy, the models were also evaluated using the AUC metrics, and AUC graphs were produced for all six classification algorithms. The Lightgbm outperformed other ML algorithms with an AUC score of 75%. Based on the result from the test data, it was concluded that applicants with low credit score should be denied access to loan facility as they have a high probability of defaulting. The results showed that applicants with high income, requesting for small loan amounts were ideal applicants to be granted loan. Their study showed that data features such as gender and marital status were not determining factors for the prediction output.

Wu, W. J. (2022) applied the random forest algorithm and the XGBoost algorithm to build prediction models. Dataset was obtained from Imperial College London, the dataset contained a total of 105,471 records and 778 features. The work employed the variance threshold method at the feature engineering stage, where unimportant features were filtered out of the dataset. Variance inflation factor (VIF) was used to measure multi core linearity of the data set. The pre-processed dataset was randomly separated into 80-20 proportion, where 80% was the training dataset, and 20% was the test dataset. The model demonstrated that though the random forest and the XGBoost algorithms

Scientia Africana, Vol. 22 (No. 3), December, 2023. Pp 137-144
© Faculty of Science, University of Port Harcourt, Printed in Nigeria *ISSN 1118 – 1931*

are decision tree algorithms, the random forest model recorded a prediction accuracy of 0.90657, while XGBoost was 0.90635. The result indicated an insignificant accuracy between the two decision tree algorithms. The study was able to demonstrate that the random forest as well as the XGBoost algorithm are suitable algorithms for loan default prediction.

Uwais & Khaleghzadeh (2022) implemented the machine learning (ML) algorithms preset on the Sparks Big Data Platform, to build loan default prediction models. The work applied six different supervised ML classification algorithms to predict loan default, they include; Decision Tree, Logistic Regression, Gradient Boosted Tree, Random Forest, Linear Support Vector Machine, and Factorization Machine. Secondary dataset was adopted from Kaggle website, the dataset contained 640,000 instances and 14 features. The dataset was randomly separated. Income was plotted against education using a scatter plot to identify correlation between these two features of the dataset, using the pandas matplotlib function of the python language, available on Spark. A positive correlation was seen between applicant's educational level and income, because as level of education increases, the income increases. The work adopted several histograms to visualize the information from the dataset based on minority and gender status. Data was pre-processed by removal of null values and adjustment of attribute data type. The pre-processed data was further prepared using the steps of feature selection, addressing class imbalance problem, converting categorical data to numerical data, and randomly splitting data in 70% training and 30% test data. The six supervised ML algorithms present on Spark MLib were applied to the training data, and used to train the models, while the test data was used to evaluate the model. Of all six ML classifiers, the decision tree and random forest demonstrated best performance with receiver operating characteristic (ROC) curve score of 99.56%, recall 99.2%. F-Score 99.5%, and precision 99.8%. The work demonstrated success in classifying loan defaulters in one of the available two classes.

Huang et al. (2023) attempted to increase the percentage accuracy of predicting loan defaulter by adopting the ensemble learning algorithm.The paper selected Adaboost algorithm as best performing model for loan default prediction. Secondary dataset was from the credit platform provided in a Tianchi competition. The dataset originally contained 1.2 million records and 47 data features. However, considering time factor in processing the huge dataset, a total of 100,000 records were randomly selected for the purpose of model building. The data was cleaned for missing values and outliers, and the feature selection technique was adopted to select relevant features from irrelevant features. At the model construction stage, the initial value of the parameters and the tuned values were tabularized in the work. The proposed model recorded an accuracy of 88%.

Li et al. (2021) aimed to improve prediction accuracy by using the blending method to fuse 3 models; Random forest (RF), CatBoost and Logistics Regression (LR). The blending method involved training a new learner, and the model of the blending method was a two-layer framework. Loan data was obtained from a lending club for Q4 2019, as made publicly available on kaggle website. The data contained 128,262 records and 150 attributes, however, over 40% of the data was removed as they were insignificant to the study. The adaptive synthetic sampling approach (ADASYN) was adopted to address class imbalance problem of the dataset, and solve the problem of performance degradation due to data imbalance. The RF, CatBoost, and LR served as benchmark to the proposed fused model. Validation metrics of accuracy, roc curve, F1-score and recall, demonstrated that the fused model outperformed the other three individual models.

Odegua (2020) adopted the Extreme Gradient Boosting (XGBoost) to build a predictive model to predict loan defaulters. They obtained dataset from Data Science Nigeria, hosted on Zindi platform. The dataset contained 26,897 records and 31 attributes, which underwent data pre-processing and wrangling stages, before being

used for training with the XGBoost classifier algorithm. The system was implemented with python programming language, and the classifier was trained on the cleaned dataset, using the good_bad_flag feature as target. Five metrics; Recall, Accuracy, F1-Score, ROC value, and Precision were used to evaluate the model.

Literature Review has shown several attempts made by researchers to improve the accuracy of predicting loan defaulters automatically.

MATERIALS AND METHOD

The Dataset used in the loan default prediction dataset was compiled by M. Yasser and uploaded to the Kaggle Data Science community data repository. The data contained 148,670 thousand records and 34 features.

The following are the processes used to build the loan default prediction system:

1) Data Loading;
2) Data Cleaning;
3) Data Processing;
4) Feature Extraction;
5) Model Training;
6) Model evaluation.

1. Data Loading

The data was loaded into the Google Colab environment using the read_csv method from the pandas library. It can be seen in figure 1.

```
[ ] loan_dataset = pd.read_csv('/content/drive/MyDrive/Loan_Default.csv')
    loan_dataset.head()
```

	ID	year	loan_limit	Gender	approv_in_adv	loan_type	loan_purpose	Credit_Worthiness	open_cred
0	24890	2019	cf	Sex Not Available	nopre	type1	p1	l1	nc
1	24891	2019	cf	Male	nopre	type2	p1	l1	nc
2	24892	2019	cf	Male	pre	type1	p1	l1	nc
3	24893	2019	cf	Male	nopre	type1	p4	l1	nc
4	24894	2019	cf	Joint	pre	type1	p1	l1	nc

5 rows × 34 columns

Figure1: Data loading using read_csv function

2. Data Cleaning

The data was cleaned from missing values, outliers, to make it fit for training, using the simple imputer method in the scikit-learn library as seen in figure 2.

```
[ ] from sklearn.impute import SimpleImputer
    imputer = SimpleImputer()
    loan_dataset[['rate_of_interest', 'term','property_value','income','dtir1']] = imputer.fit_transform(

[ ] from sklearn.impute import SimpleImputer
    imputer = SimpleImputer(strategy='most_frequent')
    loan_dataset[['age']] = imputer.fit_transform(loan_dataset[['age']])
```

Figure 2: Data Cleaning

Scientia Africana, Vol. 22 (No. 3), December, 2023. Pp 137-144
© Faculty of Science, University of Port Harcourt, Printed in Nigeria *ISSN 1118 – 1931*

3. Data Processing

The data was processed to remove duplicate columns or features. One-hot encoding was done as seen in figure 3, to convert categorical columns into numerical columns, filling those columns with 0's and 1's since the random forest classifier that will be used to train on the data cannot find patterns in categorical values.

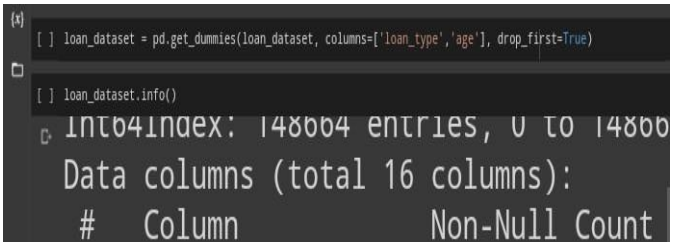


Figure 3: Code for data processing

4. Feature Extraction

Some features were expunged in this phase since they had little or no effect on the target (label) or they were duplicates. In this phase, a total of twenty-four features were dropped, and ten remained as seen in figure 4.

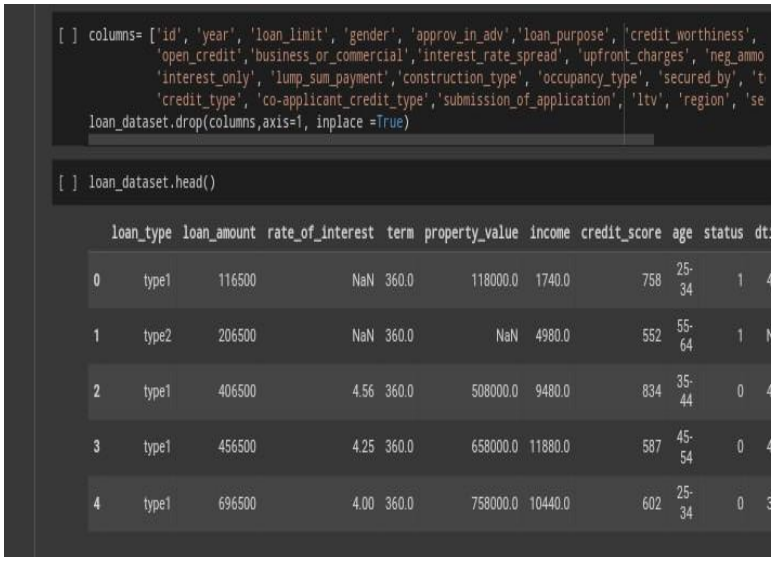
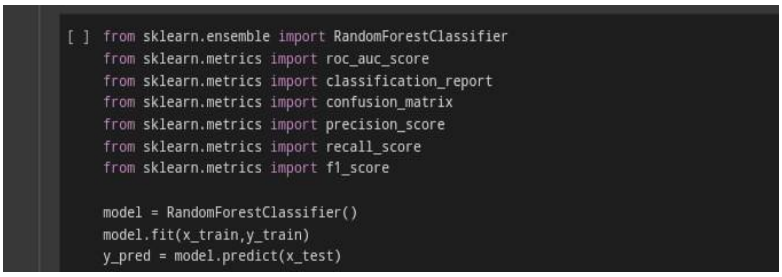


Figure 4: Code for Feature Extraction

5. Model Training

In this phase the data was fed into the random forest classifier in the scikit learn library in python. The data was trained using 70% of the data set. The codes nippet can be seen in figure 5.



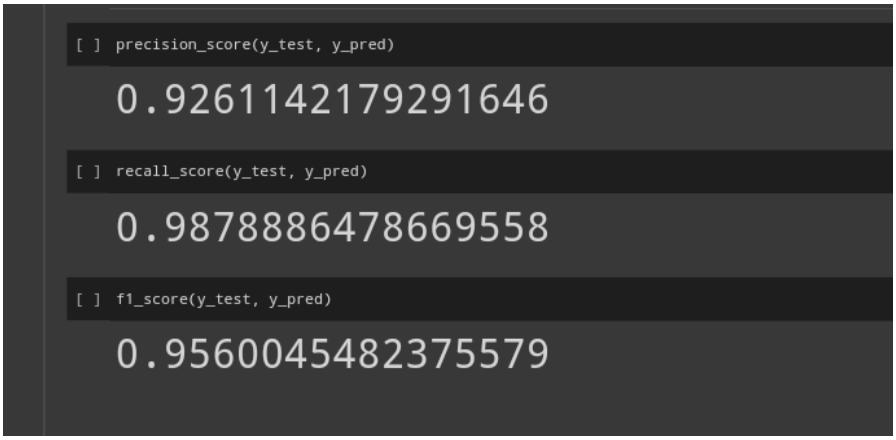
```
[ ] from sklearn.ensemble import RandomForestClassifier
    from sklearn.metrics import roc_auc_score
    from sklearn.metrics import classification_report
    from sklearn.metrics import confusion_matrix
    from sklearn.metrics import precision_score
    from sklearn.metrics import recall_score
    from sklearn.metrics import f1_score

    model = RandomForestClassifier()
    model.fit(x_train,y_train)
    y_pred = model.predict(x_test)
```

Figure 5: Code for model training

6. Model Evaluation

The model was tested with the test dataset and evaluated using the area under the curve score, recall and precision. The following evaluation scores were generated as demonstrated in figure 6.



```
[ ] precision_score(y_test, y_pred)
0.9261142179291646
[ ] recall_score(y_test, y_pred)
0.9878886478669558
[ ] f1_score(y_test, y_pred)
0.9560045482375579
```

Figure 6. Code for model evaluation

RESULT DISCUSSION

Figure 7 shows the confusion matrix of the loan default prediction system. The number of true positives where 32,664 observations which implies that the number of those instances that will not default and were predicted as such were 32,664 observations. The true negative where 10,930 observations meaning the number of those instances that will default and were correctly predicted as default were 10,930 observation. Table 1 shows some performance metrics of the model; such as precision, Recall, and F1_score. The F1_score is interpreted as the harmonic mean of precision and recall, where an F1 score reaches it best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The F1_Score of 0.98 means that the model is close to being optimal. Recall is the ratio true positive to the sum of true positive and false negative. This is the ability of the classifier to classify all positive observation as positive, the recall of the proposed system is 0.9965. The ability of the model to classify all positive observation was 99.6% accurate. Precision is the ratio of true positive to the sum of true positive and false positive. It represents the ability of this loan default classifier not to label as non–default, a sample that is default. The precision for the trained random forest model is 0.9742, showing that the model classifies about 97 occurrences out of 100 correctly. AUCscore represents the area under the curve. The AUC Score reflects how well a model predicts the correct category a loan will fall into. The Area Under Curve score for the RF model was evaluated to be 0.9823. This represents the ability of the loan default system to accurately make prediction, and gives additional indication of the quality of prediction made by the model.

Scientia Africana, Vol. 22 (No. 3), December, 2023. Pp 137-144
© Faculty of Science, University of Port Harcourt, Printed in Nigeria *ISSN 1118 – 1931*

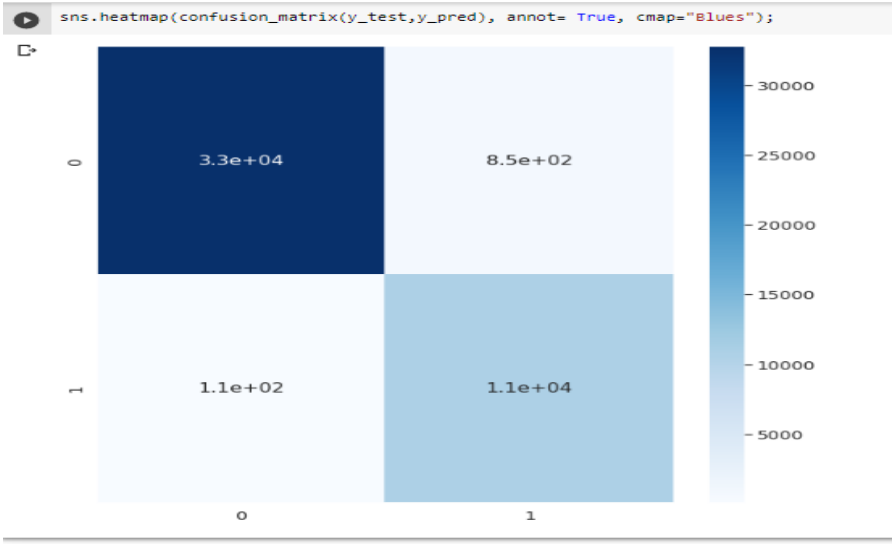


Figure 7: Confusion matrix of the model

CONCLUSION

The study was aimed at achieving a higher AUC score by adopting the random forest algorithm in building the predictive model for predicting loan defaulters. Secondary data was sourced for the research, and the data was preprocessed, and used to train the algorithm. The resultant model was evaluated using performance metrics and area under curve score. The results revealed that the predictive system built with the random forest algorithm recorded high performance percentage both in accuracy metrics and AUC score. Further works can be done, in the aspect of creating a graphic user interface for the application, to make the system more user-friendly.

REFERENCES

Almamun, M., Farjana, A., Mamun, M. (2022). Predicting Bank Loan Eligibility uing Machine Learning Models and Comparison Analysis, *Proceedings of the 7th North American International Conference on Industrial Engineering and Operations Management, Florida*. 1423 – 1432.

Hamid,E. N. andAhmad, N (2011). A New Approach for Labeling the Class of Bank Credit Customers via Classification Method in Data Mining, *International Journal of Information and Education Technology*, 1(2): 150-155.

Huang, Y., Shao, Y., Tang, D., Huang, J., and Chen, S. (2023). Loan Default Prediction Based on Ensemble Learning, *International Journal of Innovation and Research in Educational Sciences*, 10(3): 149 – 159.

Hussain, A.B. and Shorouq,F.K.E. (2014). Credit risk assessment model for Jordanian commercial banks: Neural scoring approach", *Review of Development Finance, Elsevier*, 4(10): 20–28.

Li, X., Ergu, D., Zhang, D.,Qiu, D., Cai, Y. and Ma, B. (2021) Prediction of Loan Default Based on Multi-model Fusion, *Procedia Computer Science*.

Odegua, R. (2020) Predicting Bank Loan Default with Extreme Gradient Boosting, *Preprint Cornell University.*

Purohit, S. U., Mahadevan, V. and Kulkarni, A. N. (2011) Credit Evaluation Models of Loan Proposals for Indian Banks, *International Journal of Modelling and Optimization*. 2(4): 529 – 534.

Twala, B. (2010) Multiple classifier Application to Credit Risk Assessment,

Expert Systems with Applications. 37(4): 3326–3336.

Uwais, A. M. and Khaleghzadeh, H. (2022) Loan Default Prediction using Spark Machine Learning Algorithms, *AIAI 29th Irish Conference on Artificial Intelligence and Cognitive Science, Dublin*, 118-129.

Wu, W. J. (2022) Machine Learning Approaches to Predict Loan Default, *Intelligent Information Management*. 14(3), 157-164.