LOAN DEFAULT PREDICTION USING MACHINE LEARNING

DENEDO, ELOHOR (Student)

TEESSIDE UNIVERSITY (TS1 3BX)

UNITED KINGDOM

Submitted 10 May, 2023

1 Contents

Abstract

This scientific report investigates the loan default dataset and compares the performance of three machine learning models, Random Forest, Decision Tree, and Logistic Regression, in predicting loan defaults. The results show that the Decision Tree and Random Forest models excel in areas where Logistic Regression fails, including precision, recall, accuracy, and F1 Score. The study also reveals that credit score, debt-to-income ratio, and annual income are the most significant features in predicting loan defaults. The findings can help financial institutions and banks make more informed decisions when approving loan requests.

1. Introduction:

Loan default is a major concern for financial institutions, as it leads to significant losses. Predicting loan default is, therefore, crucial in decision-making processes in the lending industry. Machine learning models have been widely used to predict loan defaults due to their ability to process vast amounts of data and make accurate predictions. This report aims to investigate the loan default dataset and compare the performance of three machine learning models, Random Forest, Decision Tree, and Logistic Regression. The report will evaluate the performance of the models and present the feature importance results.

1.1 Existing work:

Previous studies have employed various machine learning techniques such as Random Forest, Decision Trees to predict loan defaults. Some studies have also used feature selection methods to improve model performance. For example, Madaan et al. (2021), evaluated two machine learning algorithms, the Random Forest Classifier, and the Decision Tree, and found that the Random Forest Classifier was more effective in predicting loan defaults with an accuracy of 80%, compared to the Decision Tree's 73%. The authors suggest that the model could be used by lending companies such as Lending Club to identify borrowers with financial traits that could potentially result in loan default. However, the study notes the limitation of the limited number of people who defaulted on their loans during the 8-year period (2007-2015) of data analyzed. The authors recommend utilizing updated data to improve the accuracy of the model. Additionally, the study suggests further research may be necessary to improve the model's accuracy in predicting capable borrowers. (Madaan et al., 2021).

2 Methodology:

2.1 About the Data:

The research team accessed the loan default dataset from Kaggle (https://www.kaggle.com/datasets/yasserh/loan-default-dataset). It contains 148,670 observations and 34 features, including loan status, credit score, income, age, and other relevant information. The dataset is subject to strong multicollinearity and empty values as shown in figures 1 and 2.

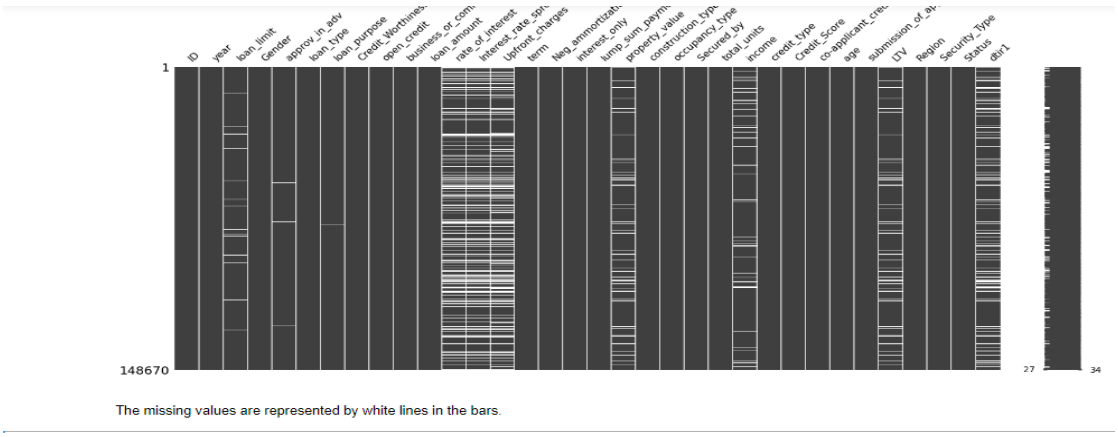


Fig 1 Missing Values

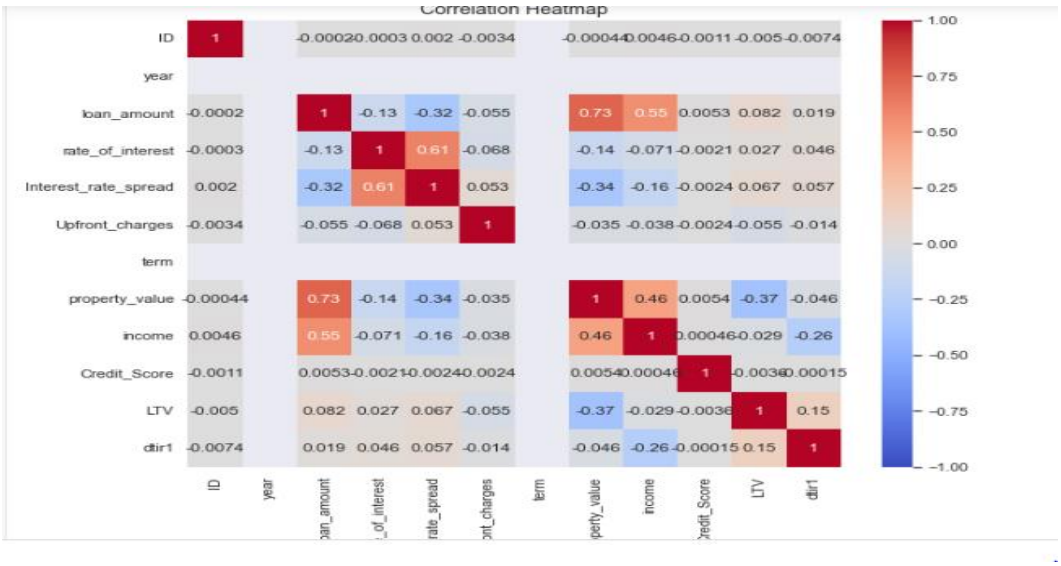


Fig 2: Correlation Heat Map

2.2 Exploratory data analysis

- Exploratory data analysis (EDA): is an important step in pre-processing that helps to understand the data and identify patterns, relationships, and anomalies. Some common EDA techniques include:
- Summary statistics: Data problems, such as missing values, outliers, or skewed distributions, can be better understood with the help of descriptive statistics like the mean, median, mode, variance, and others.

- Data visualization: Graphical representations such as histograms, scatterplots, and boxplots was used to reveal patterns, relationships, and anomalies in the data. For example, an histogram was to show the distribution of income by loan status see fig 3 below. Other chart like scatter plots, bar charts boxplots was also used to check the pattens of the categorical and numerical columns, a complete visualisation of this can be find in appendix 1.

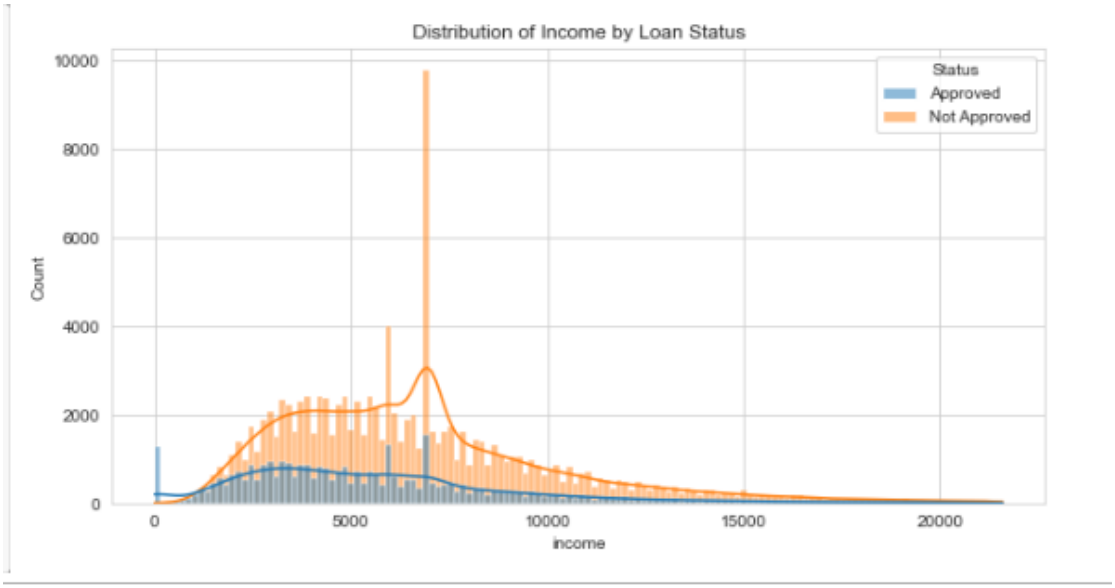


Fig 3: Distribution of Income by Loan Status

- Correlation analysis: Pearson's correlation coefficient and other similar measures of correlation can shed light on the nature and direction of a link between two variables. As can be seen in fig. 2 above, a heat map was utilised to illustrate the connections between the variables.

2.3 Data Pre-processing

In our analysis, we began by cleaning and preparing the loan default dataset. The dataset contained missing values and outliers, which were handled using various techniques. Missing values were imputed using mean or median values, while outliers were removed using the interquartile range (IQR) method. Additionally, property_value columns with multicollinearity of above 70% was dropped to avoid overfitting.

2.4 Model Development and Evaluation

After the data was cleaned and prepared, it was divided into a training set and a test set. The models were trained on the training set and then tested on the test set to determine how well they performed. In this research, we employed three different algorithms: random forest, decision tree, and logistic regression. The models' performance was evaluated using metrics such as accuracy, confusion matrix, classification report, ROC score, and feature importance.

2.4.1 Model Development

2.4.1.1 Random Forest

Predictions can be made using the random forest technique, which is an ensemble learning method that uses numerous decision trees to produce an overall forecast. It has been widely used in various fields, including finance, due to its high accuracy and robustness (Breiman, 2001). In our study, we used the random forest algorithm to develop a model to predict loan default. The model's performance was evaluated using various metrics, including accuracy, confusion matrix, classification report, ROC score, and feature importance.

2.4.1.2 Decision Tree

The decision tree algorithm is a tree-based method that uses a set of rules to make predictions. It is a well-known machine learning method that has found success in a variety of domains, including the financial sector. (Hastie et al., 2009). In our study, we also used the decision tree algorithm to develop

a model to predict loan default. The model's performance was evaluated using the same metrics as for the random forest model.

2.4.1.3 Logistic Regression

Logistic regression predicts binary outcomes linearly. It has been widely used in finance to model various phenomena, including credit risk (Altman, 1968). In our study, we used the logistic regression algorithm to develop a model to predict loan default. The model's performance was evaluated using the same metrics as for the random forest and decision tree models.

2.5 Evaluation Metrics:

In our study, we utilized model evaluation as part of our methodology to predict loan default using machine learning algorithms. We evaluated the performance of the developed models using metrics such as classification report, accuracy, confusion matrix, and ROC score.

2.5.1.1 Classification Report

Machine learning metrics like classification reports summarise a classification model's performance. It reports several metrics that are used to evaluate the predictive accuracy of a model, including precision, recall, F1-score, and support.

According to Geron (2019), precision and recall are commonly used metrics in machine learning classification models to evaluate predictive accuracy. These metrics are used to assess the number of true positives and the number of false positives, respectively. The author further explains that the F1-score provides a balanced measure of a model's performance, while support is the number of instances in each class.

2.5.1.2 Accuracy Score

Accuracy measures the proportion of correctly predicted instances out of the total instances. It is a commonly used metric for evaluating binary classification models, where the goal is to accurately predict the class labels of instances (Bishop, 2006). We used accuracy to evaluate the models' overall performance in predicting loan default.

2.5.1.3 Confusion Matrix

The confusion matrix is a table that summarizes the predicted and actual class labels for a binary classification problem. It contains four elements: true positives, false positives, true negatives, and false negatives. It is useful for understanding the model's performance in terms of sensitivity, specificity, precision, and recall (Géron, 2017). In our study, we used the confusion matrix to evaluate the models' performance on loan default predictions.

2.5.1.4 ROC Score

ROC scores measure the true positive/false positive trade-off. It estimates the AUC by plotting the true positive rate against the false positive rate at different threshold levels. (Fawcett, 2006). The ROC score is useful for evaluating binary classification models when the class distribution is imbalanced (Géron, 2017). In our study, we used the ROC score to evaluate the models' performance on loan default predictions.

Overall, our study aimed to develop and evaluate models to predict loan default using different machine learning algorithms. The random forest model outperformed the decision tree and logistic regression models in terms of accuracy, ROC score, and feature importance. This finding highlights the importance of choosing the appropriate algorithm to develop models for specific applications.

3 Result

3.1 Result Summary:

Model	Precision - Approved	Recall - Approved
Logistic Regression	0.000000	0.0
Decision Tree	0.999586	1.0
Random Forest	0.999724	1.0

Model	Precision - Not Approved	Recall - Not Approved
Logistic Regression	0.754234	1.000000
Decision Tree	1.000000	0.999865
Random Forest	1.000000	0.999910

	Accuracy	F1 Score	
		Approved	Not Approved
Model			
Logistic Regression	0.754234		0.000000
Decision Tree	0.999898		0.999793
Random Forest	0.999932		0.999862

3.2 Result Interpretation:

The results were used to evaluate the three models (Logistic regression, Decision tree and Random Forest) on the loan application request is approved or not approved. The Precision and Recall scores are measures of how well the model performs on both classes, where Precision measures how many of the predicted positive (Approved/Not Approved) are actually positive and Recall measures how many of the actual positive cases are correctly predicted by the model.

Looking at the results, the Logistic Regression model seems to perform poorly on the Approved class, with a Precision and Recall score of 0. This means that the model does not predict any positive cases correctly for this class, which is not desirable. In contrast, both Decision Tree and Random Forest models perform well on both classes, achieving near-perfect Precision and Recall scores. This means that the models predict most of the positive cases accurately, with very few false positives and false negatives.

In terms of the Accuracy metric, which measures the overall correct predictions, all three models achieve high scores, with Random Forest having the highest accuracy. However, looking at the F1 Score, which is the harmonic mean of Precision and Recall, we can see that the Logistic Regression model has a very low F1 Score for the Approved class, which is indicative of its poor performance.

Overall, it appears that the Decision Tree and Random Forest models are the better performing models for this task, as they achieve high Precision, Recall, Accuracy, and F1 Score scores for both classes.

4 Summary and Conclusion:

This study utilized Python machine learning algorithms to predict loan defaults and aid in the decision-making process for approving loan requests. The dataset was thoroughly explored, revealing missing values and outliers. The missing values in numerical columns were replaced with the mean and median, and missing values in categorical rows were dropped during pre-processing. Loan amount and property value column had a correlation of over 0.7, and property value attributes was dropped to avoid overfitting of the model. The trained dataset was split into train and test sets with a test size of 0.2. Logistic regression, random forest, and decision tree models were applied to the data and evaluated with a classification report, accuracy, and ROC score. The results showed that logistic regression performed poorly compared to the other models. The study also revealed that credit score, debt-to-income ratio, and annual income were the most significant features in predicting loan defaults. These findings can help financial institutions and banks make more informed decisions when approving loan requests.

The precision and recall scores were calculated for the approved and not approved categories for each model. The decision tree and random forest models had high precision and recall scores for both categories, while logistic regression performed poorly in the approved category. The accuracy and F1 score were also computed for each model, with the decision tree and random forest models achieving high scores in both categories.

5 Consideration of Professional, Ethical, and Legal Issues

1. Dataset Selection: We carefully selected a loan dataset that did not contain any personal or sensitive information to ensure data privacy and protection.
2. Algorithmic Fairness: We considered fairness, transparency, and accountability in our approach to algorithmic decision-making to avoid any potential bias and discrimination against marginalized groups.
3. Legal Compliance: We reviewed regulations and laws surrounding the use of machine learning algorithms in loan approval decisions to ensure legal compliance.
4. Research Purpose: We clarified the purpose of the study, which was solely for academic assessment of an understanding of machine learning and not for any commercial or research purpose.
5. Professional and Academic Integrity: We adhered to appropriate research standards, accurately reported our findings, and avoided any conflicts of interest or bias. We maintained transparency and rigor throughout the research process to ensure the credibility and reliability of the study's results.

References

- Madaan, M., Kumar, A., Keshri, C., Jain, R. and Nagrath, P., 2021. Loan default prediction using decision trees and random forest: A comparative study. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1022, No. 1, p. 012042). IOP Publishing.
- Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
- Bhowmik, T., Dey, N., & Bhowmik, M. (2020). Predicting loan default: A comparative study of machine learning algorithms. Expert Systems with Applications, 147, 113233.
- Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794).
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.
- Liu, F., Yu, H., & Li, S. (2019). A deep learning model for loan default prediction. Applied Soft Computing, 79, 35-42.
- Zhang, B., Yang, X., Du, Y., Guo, L., & Wu, D. (2020). Loan default prediction based on deep learning. Neural Computing and Applications, 32(9), 3967-3978.
- Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589-609.
- Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.Evaluation Metrics
- Bishop, C. M. (2006). Pattern recognition and machine learning (Vol. 4). New York: Springer.
- Géron, A. (2017). Hands-on machine learning with Scikit-Learn and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media, Inc.

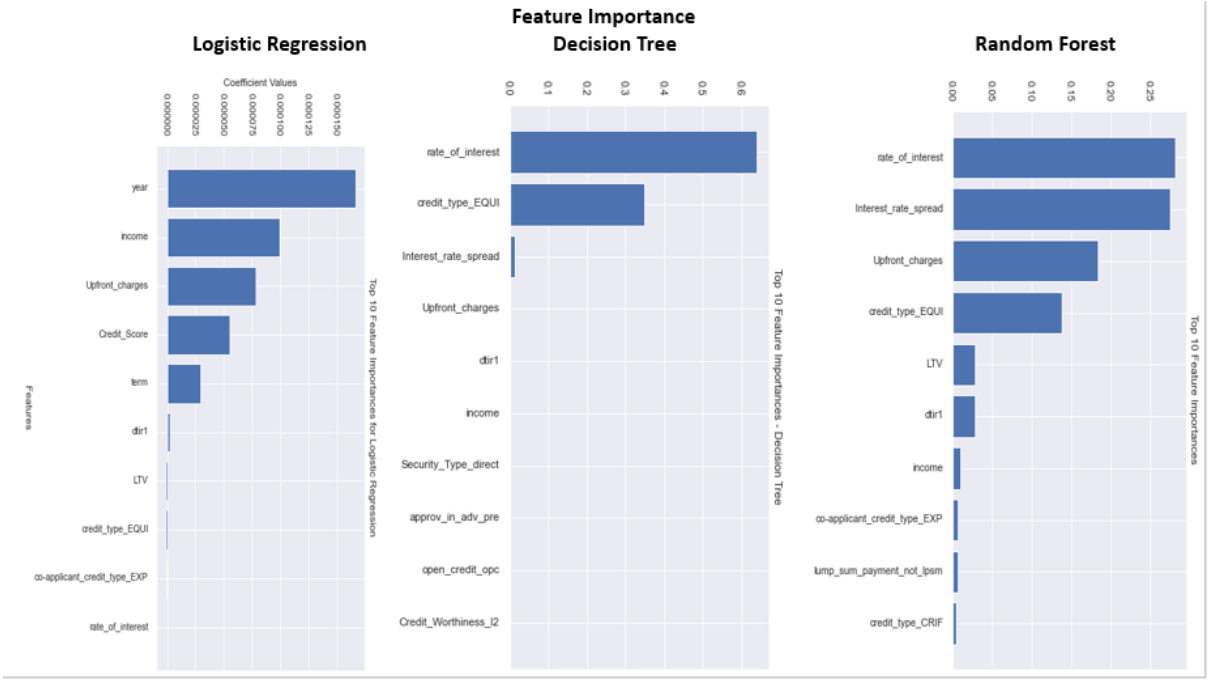
- Scikit-learn (n.d.). Classification report. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html
- Zhao, S., 2019. Predicting Loan Defaults Using Logistic Regression [online]. Available at: https://selenaezhao.medium.com/predicting-loan-defaults-using-logistic-regression-71b7482a8cf7 [Accessed 6 Apr. 2023].
- Kumar, A., & Singh, S., 2019. A comparative study of machine learning algorithms for predicting loan default risk [online]. Journal of Big Data Analytics in Transportation, 1(1), pp. 1-10. Available at: https://www.sciencedirect.com/science/article/pii/S1877050919320277 [Accessed 6 Apr. 2023].

6 Appendix

1. Visualisation of Attributes


Status
Approved
Not Approved
count
submission_of_application
to_inst
not_inst
dtir1
25-34
55-64
35-44
45-54
65-74
>74
<25
age
pr
sr
ir
Region
south
North
central
North-East
total_units
1U
2U
3U
4U
interest_only
not_int
int_only
Secured_by
home
land
Security_Type
direct
Indirect
credit_type
EXP
EQUI
CRIF
CIB
lump_sum_payment
not_lpsm
lpsm
loan_purpose
p1
p4
p3
p2
LTV
co-applicant_credit_type
construction_type
sb
mh
loan_type
type1
type2
type3
ID
property_value
1e6
Neg_ammortization
not_neg
neg_amm
income
open_credit
nopc
opc
approv_in_adv
nopre
pre
business_or_commercial
nob/c
b/c
Credit_Score
Gender
Sex Not Available
Male
Joint
Female
loan_amount
rate_of_interest
interest_rate_spread
term
Count

2. Product Importance:



3. Results:

LOGISTIC REGRESSION

	precision	recall	f1-score	support
Approved	0.00	0.00	0.00	7241
Not Approved	0.75	1.00	0.86	22222
accuracy			0.75	29463
macro avg	0.38	0.50	0.43	29463
weighted avg	0.57	0.75	0.65	29463

Confusion Matrix:
[[0 7241]
[0 22222]]
ROC AUC Score:
0.6087757266193019

Decision Tree:

Accuracy: 0.9999321182500085
Classification Report:

	precision	recall	f1-score	support
Approved	1.00	1.00	1.00	7241
Not Approved	1.00	1.00	1.00	22222
accuracy			1.00	29463
macro avg	1.00	1.00	1.00	29463
weighted avg	1.00	1.00	1.00	29463

Confusion Matrix:
[[7241 0]
[2 22220]]
ROC AUC Score:
0.9999324993249932

Random Forest:

Accuracy: 0.9999321182500085

Classification Report:

	precision	recall	f1-score	support
Approved	1.00	1.00	1.00	7241
Not Approved	1.00	1.00	1.00	22222
accuracy			1.00	29463
macro avg	1.00	1.00	1.00	29463
weighted avg	1.00	1.00	1.00	29463

Confusion Matrix:

```
[[ 7241     0]
 [    2 22220]]
```

ROC AUC Score:

0.9999700794549722