Machine Learning Approach to Credit Risk Prediction: A Comparative Study Using Decision Tree, Random Forest, Support Vector Machine and Logistic Regression

Thesis · March 2023

DOI: 10.13140/RG.2.2.31652.14725

1 author:

Ankit Karmakar
Madras School of Economics
2 PUBLICATIONS **4** CITATIONS

SEE PROFILE

TERM PAPER:

Machine Learning Approach to Credit Risk Prediction: A Comparative Study Using Decision Tree, Random Forest, Support Vector Machine and Logistic Regression

Ankit Karmakar

M.A. Financial Economics
Madras School of Economics

Reg. No: FE/2022/006

Abstract

The banking industry is expanding globally, but banks are encountering a significant challenge in managing credit risk. Credit risk is the risk associated with a borrower defaulting on debt obligations. While there are lots of factors behind bank loan defaults, banking authorities are trying to mitigate the risk. Recently, the advancement of machine learning methods has made it possible to assess credit information and determine if an individual qualifies for credit financing. Even though there are organizations that are offering their credit assessments of customers to banks, researchers are still looking into different machine learning techniques to enhance the precision of credit risk evaluation. Three machine learning models which can predict the eligibility for a loan by analysing specific attributes and will aid financial institutions while they have to select appropriate candidates for loans among the many applicants. I thoroughly investigate 1. Decision Tree (DT), 2. Random Forest, 3. Support Vector Machine (SVM) and 4. Logistic Regression (LR) models and compare them to determine which one is more adaptable for credit risk analysis. The data set was tested on these four algorithms and the results indicate that considerably higher accuracy was produced by the Random Forest algorithm when compared to the SVM, DT and LR.

Keywords: Credit risk, Loan prediction, Random Forest, Machine Learning, Default detection

1. Introduction

Worldwide, individuals rely on banks in one way or the other so as to address their financial difficulties and attain personal objectives via loans disbursed for different purposes. And with the shift in the world economy, the competition in the financial sector is rising and making credit lending an indispensable part of this scenario. To accommodate the need, numerous banking and non-banking financial institutions currently provide credit lending services. Additionally, a significant portion of these institutions' revenue directly comes from the interests earned on loans given.

Even though both the parties enjoy benefits, some considerable risks are involved during the disbursement of loans. ‘Credit Risk’ refers to the risks which stand for the situations when the borrower is unable to pay back the loan amount by the terms which both the lender and borrower had mutually decided on. [3]. So, mitigating the risks became one of the main objectives of these lending institutions. To examine a borrower in the traditional lending process, the banks majorly make use of the ‘5C Principle’ – Capacity, Capital, Character, Conditions and Collateral. [8]. But this process had a lot of limitations, one of them is obviously the subjective judgement. Banks and other financial institutions issue loans after verification and validations but it is heavily dependent on the risk control personnel. Even after the rigorous process, it is not absolutely determined whether the chosen applicant will be able to pay off their debt on time.

Professionals whose only job was to evaluate the individual’s profile and give verdict whether it is safe to give loans to them were hired by banks in the past. At that time, they decided the worthiness of a borrower by a numeric score, also known as ‘Credit Score’. The score helps the authorities to estimate the probability of the borrower paying off the loan within the stipulated timeline and terms as per the credit history and/or payment history of the applicant alongside their background. [1]

With the advent of technology, researchers, banks and other financial institutions have started employing machine learning and deep learning algorithms to train various classifiers that can

predict an applicant's eligibility to get a loan as per their credit history and other data. This process can make it easier to select eligible candidates before approving a loan.

This term paper also focused on machine learning algorithms to find out which model is the best fit in current times to accurately predict any defaults, which will help the banks and the financial institutions. The classifiers I used here are Random Forest, Support Vector Machine and Logistic Regression. They will each be analysed independently for the dataset, find its patterns, and draw conclusions from them. And in the end, based on our analysis we will determine whether a new applicant shall default on a loan or not.

The paper is distributed in the following sections: Starting with Section 2 which talks briefly about our Dataset, cleaning the dataset, then the insights we can get from our data by running a proper analysis and then describes the four algorithms used in this paper, i.e., Decision Tree, Random Forest algorithm, Support Vector Machine & Logistic Regression with the Model evaluation. Section 3 shows the Results we got from our models in the form of Confusion matrix and Classification report by comparative analysis of the four classifiers. Section 4 draws the conclusion of the paper with final remarks and the different ways these methods can be employed in the future.

2. Data and Methodology

2.1 Dataset Description

For this paper I will be considering a publicly available Credit Risk Dataset [6] on Kaggle platform and modified it a little bit to meet our needs. This data covers approximately 3 hundred million amounts of loans credited to more than 30 thousand people. It consists of 11 features in total, which describes each individual's profile.

I have defined the various features as per our need (in Table 1) and then encoded it further for ease of use (in Table 2 and 3). Please refer to the appendix for the Tables in question.

2.2 Data Cleaning

After getting any data it is necessary to do data cleaning. There are lots of empty cells we have to identify and drop them. We also have to look for relevant feature and drop the rest. Data cleaning is prerequisite of the Exploratory Data Analysis.

```
person_age                  int64
person_income               int64
person_home_ownership       int64
person_emp_length         float64
loan_intent                 int64
loan_amnt                   int64
loan_int_rate             float64
loan_status                 int64
loan_percent_income       float64
default_hist                int64
cred_hist_length            int64
dtype: object
```

Fig. 1 Columns with their respective datatypes

Here the figure 1 describes each data types of each column, except Person Employment Length, Loan Interest rate and Loan to percentage Income all takes integer values. And data set also needs some correction as there are some missing values. So, we dropped all the rows which had null values.

2.3 Data Exploration

Data Exploration played a really interesting part in understanding our Credit Risk dataset. Before going diving into the classification models, it's obviously better to look into the relationship among the data. By diving into the relationships in the data, we gained more insight of the interconnection that we couldn't without this analysis. In this section we will examine distribution of the data in more detail and pose specific queries regarding the dataset's data.

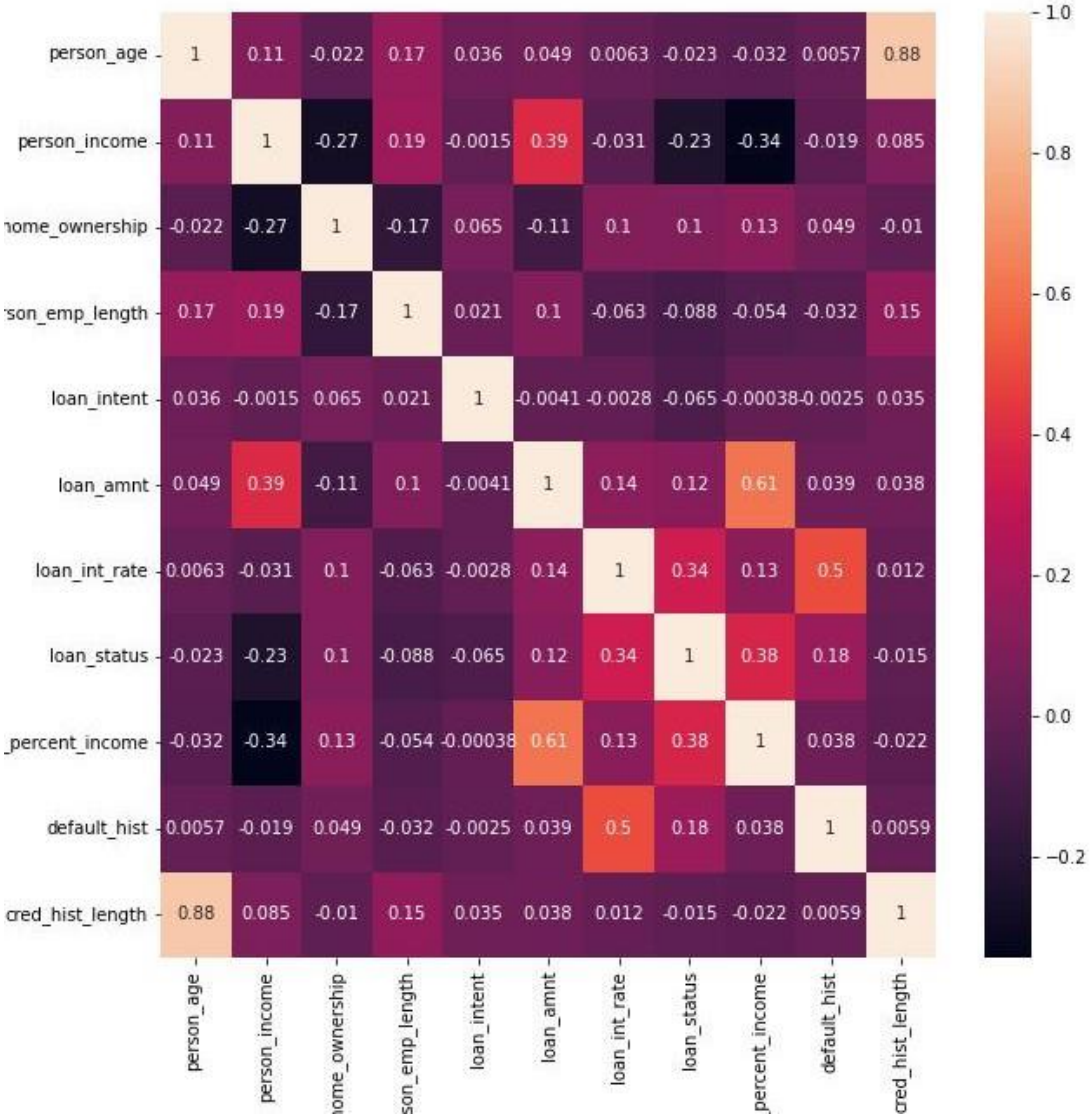

Fig. 2 Heatmap showing the correlation among the features

Heatmaps showed in the Fig.2 helps us to find out any type of correlations are there among the features in the dataset or not. Here anyone can find meaningful relationships like loan_amount and loan_percent_income has a positive relation and has correlation of 0.61.

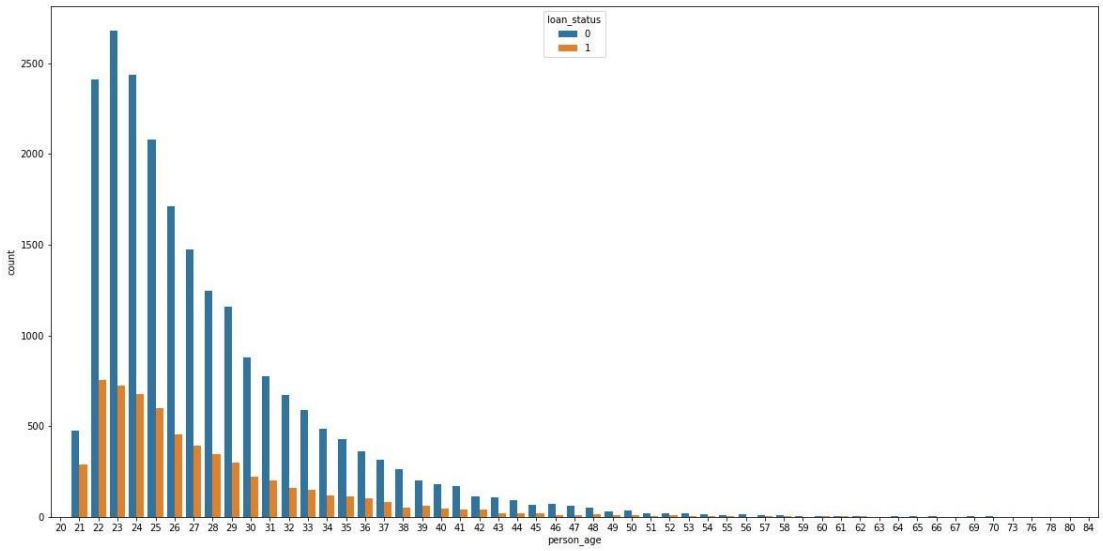

Fig.3 Bar diagram showing Defaults and Non-defaults of loan in different age group

In the above Bar diagram (Fig. 3), I am taking horizontally person's age and vertically the no. of defaults and non-defaults of loan; and here '0' & '1' means defaults and non-defaults respectively. This tells us an interesting insight, younger borrowers have the highest default rates, mostly because their sources of income are less stable.

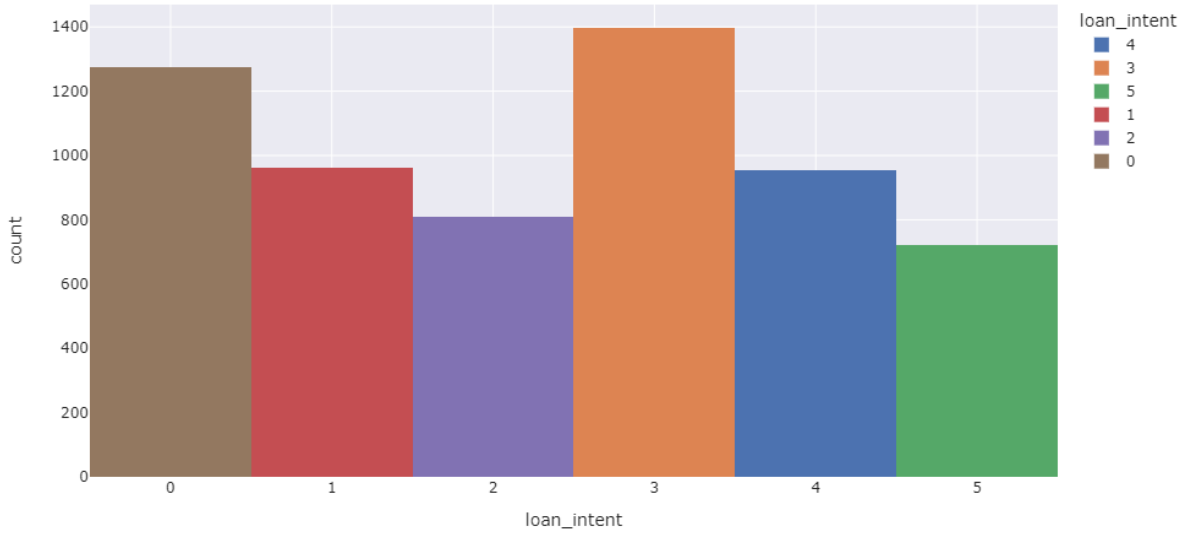

Fig. 4 No. people who took debt and defaulted for different reasons

In this diagram (Fig. 4) I'm showing for which intention borrowers took the loan and defaulted. And as you can see '3' shows the highest data following by '0', which explains that people don't save for any medical conditions and in the end up taking loans. And the later explains that people tend to consolidate the debts so it become easier to manage. Sometimes it occurs when it is more advantageous to take out a loan to pay off debts, transforming a tangle of bills into a single instalment to be paid off.

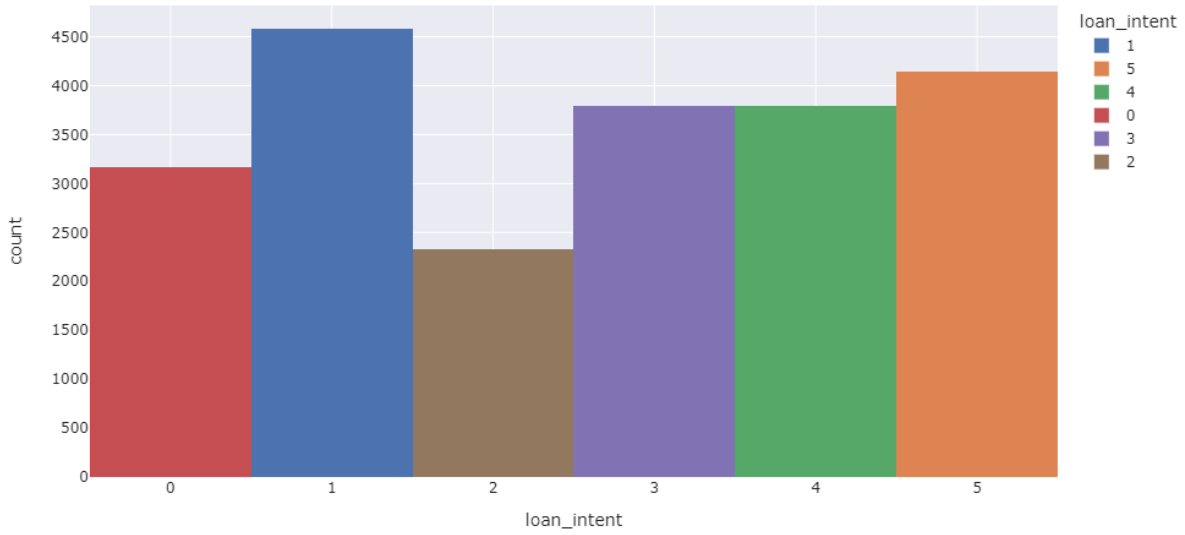

Fig. 5 No. people who took debt and defaulted for different reasons

In Fig. 5 the bar diagram shows the borrower's intention behind taking the loans and paying them back without defaulting. Here you can see Mostly the education ('1') loan is paid back followed by Venture ('5') loans.

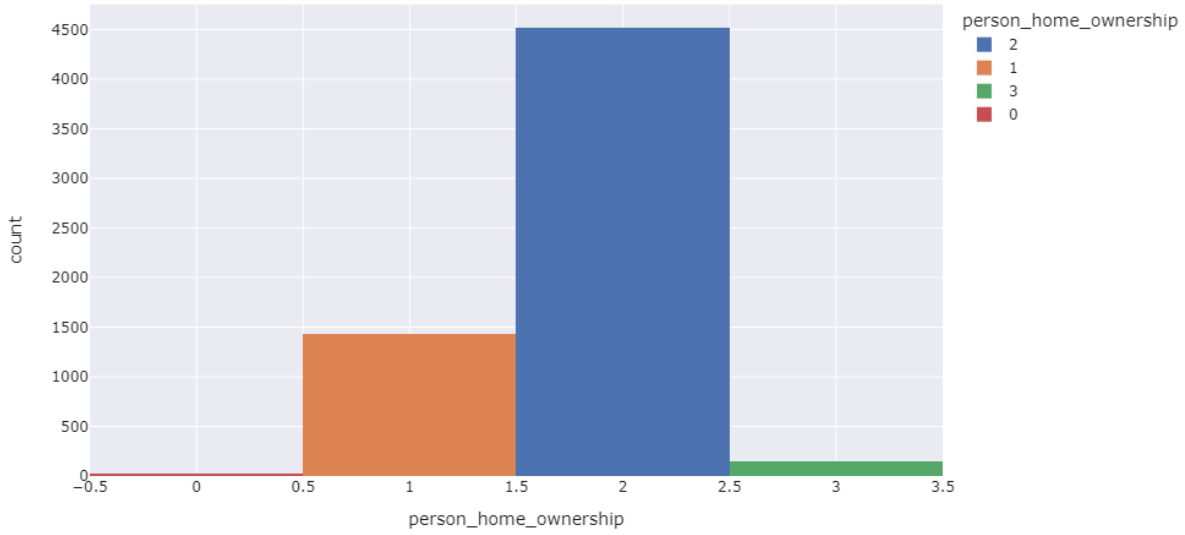

Fig. 6 Type of homeownership of defaulters

In the two diagram (Fig. 6 & Fig. 7), the bar diagrams shows the different type of home ownership of defaulters and non-defaulters respectively. For Fig. 5, we can point out that the borrowers who stays in the rental ('0') houses tends to default loan the highest followed by Mortgage ('1'). Again, in the Fig. 6 we can see that those who live in the mortgage houses paid back the loan the most followed by Rent.

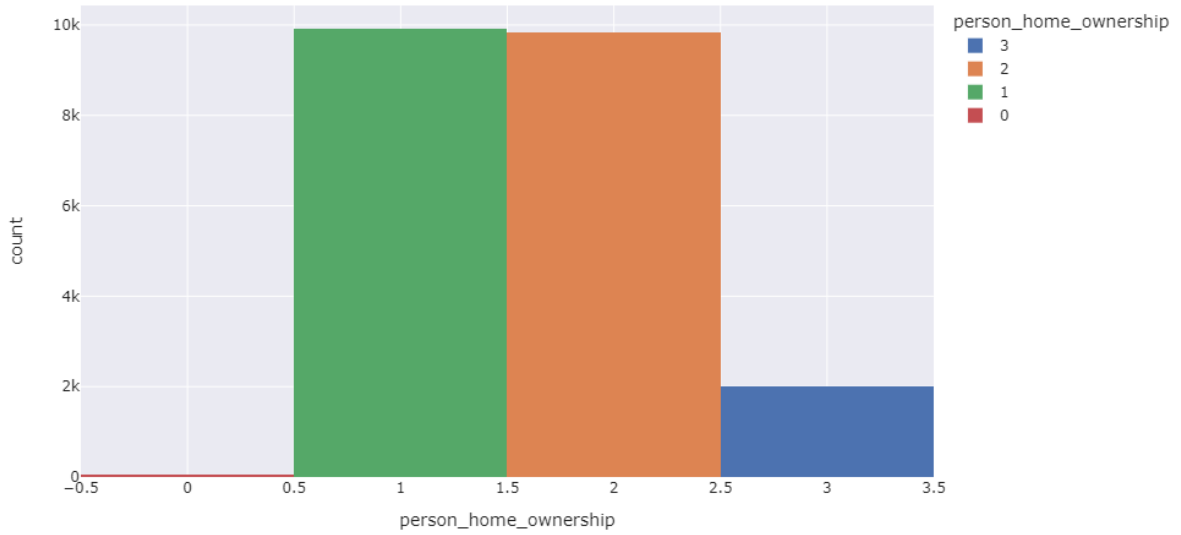

Fig. 7 Type of homeownership of non-defaulters

2.4 Modelling

Under the realm of Artificial Intelligence, there exists Machine Learning which is a method that uses algorithms to analyse previous data to find out any possible patterns and make decision with a reference of previous instances. A system that builds analytical model automatically with minimal human intervention is basically Machine Learning. [13]

The ML model has the ability to enhance its performance with every iteration. Now to measure the efficacy of the models, it is crucial to divide the dataset into separate sets for training and testing. Thus, before we train our models, we have partitioned the Credit Risk dataset into a comparatively smaller dataset for Training, which constituted 80% of the entire dataset, and an even smaller dataset for Testing, which accounted for the remaining 20%.

To measure and properly evaluate the accuracy of our model's predictions, we have to use some sort of performance sets. It's not sufficient to rely only on model accuracy to assess performance. So considered the F1 score, recall and confusion matrix as significant metrics for analysis.

2.4.A Decision Tree

Under all the machine learning algorithms, one of the simplest algorithms is Decision tree. This algorithm is a classifier which is extensively used for classification and regression purposes. A framework very similar to a tree is generated by this algorithm. A recursive portioning algorithm is utilised in the process. [4] A class label is represented by each leaf node in the tree while all the outcomes for the test are represented by the branches in the tree. The internal nodes for an attribute represent these tests. [9] In between all the techniques that are used for classification this is the most popular and widely used for credit scoring.

2.4.B Random Forest

A combination of tree predictors such that each tree depends on the values of an independently sampled random vector and with the same distribution for all trees in the forest is called a

random forest. [5] This is nothing but a collection of decision trees, where trees are different from the others slightly. This is favoured over other algorithms because it minimizes the time required for managing and preparing data with greater accuracy. It also has various applications such as assessing high credit risk customers, identifying fraud, and solving option pricing problems.

2.4.C Support Vector Machine

Support Vector Machine or SVM is also a classifier machine learning algorithm but a supervised one. In the domain of Neural Networks and ML, this is a highly efficient contender for classification purposes. Study shows that, to analyse binary classes concepts of most learning algorithms are used by supervised model like SVM [2]. An SVM model comprises of a hyperplane which is used to separate the different observations for classification purpose [11].

2.4.D Logistic Regression

In the world of finance Logistic regression is undoubtedly the most popular and used statistical technique. The main advantages of a LR model can be found in its straightforward concept, reliable performance with not so complicated implementation process [10]. Moreover, LR performs better than linear regression by addressing and resolving multiple problems. Logistic regression overcomes the limitation of Linear regression's non-positive and greater than value 1 scenario. By assigning a continuous grade between 0 and 1 and while keeping the output limited to values between 0 and 1, LR solves the problem [4].

2.5 Model Evaluation

Here we will evaluate the models by their performance which will use consist some constraints of underfoot or overfit of the model. We will talk about the parameters to evaluate the performance of our models such as Confusion metrics, Accuracy, Precision, Recall, F1 score etc.

2.5.1 Confusion Matrix

A confusion matrix is a way to display visually the various outcomes of a classification problem, showing how well the predictions match the actual results. It is presented in a table format and provides a visual representation of the performance of a classifier, by mapping out the predicted and actual values.

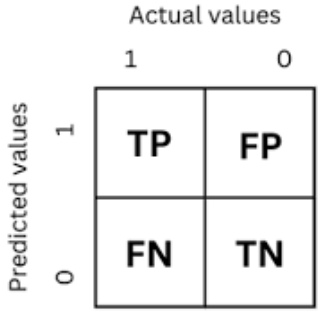

Fig. 8 Confusion Matrix

Here TP, FP, FN & TN means True Positive, False Positive, False Negative & True Negative respectively.

2.5.2 Accuracy

The model's accuracy has been evaluated using predetermined metrics. When applied to a balanced class, the model exhibits a high level of accuracy. However, when used on an unbalanced class, the accuracy is considerably lower.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

2.5.3 Precision

The Precision value is derived from dividing truly predicted positive values by the total predicted positive values in a percentage sense. The denominator over here tells us about the total predicted positive values from the dataset. The precision value explains the perfectness of the model.

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

2.5.4 Recall

Percentage ratio of positive instances with actual total positive instances is recall value. It talks about accuracy of our model is out of those predicted positive, how many of them are actual positive.

$$\text{Recall} = \frac{TP}{TP + FN}$$

2.5.5 F1 Score

For getting the F1 Score, we have to take the Harmonic mean of Precision and Recall. A model is considered as good model if it shows higher F1 score. Numerator over here is the product of precision and recall, and hence these two are positively related with the final F1 score.

$$\text{F1-score} = \frac{2 \text{ x } (\text{precision x recall})}{(\text{precision} + \text{recall})}$$

3. Results

In this paper I have used four machine learning algorithms - Decision Tree, Random Forest, Support Vector Machine and the Logistic Regression to find a best suited model for loan prediction and credit risk assessment.

The results of the four models are shown and compared with their Classification report and Confusion matrices to get the better idea of which model is more accurate and precise.

3.1 Decision Tree

For our dataset Decision tree algorithm secures an accuracy of 85%

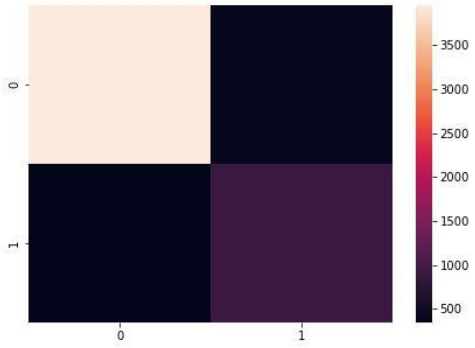

Fig. 9 Confusion Matrix of Decision Tree

	precision	recall	f1-score	support
0	0.92	0.89	0.91	4394
1	0.65	0.72	0.68	1190
accuracy			0.86	5584
macro avg	0.79	0.81	0.80	5584
weighted avg	0.86	0.86	0.86	5584

Fig. 10 Classification Report of Decision Tree

3.2. Random Forest

For our dataset Random Forest algorithm secures an accuracy of 91%

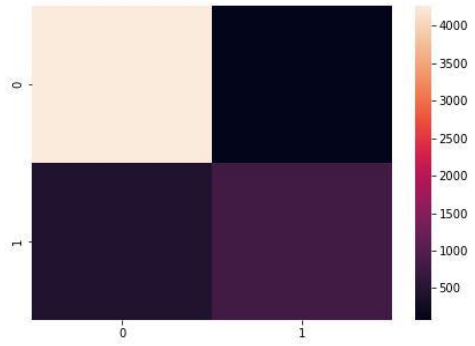

Fig. 11 Confusion Matrix of Random Forest

	precision	recall	f1-score	support
0	0.91	0.97	0.94	4394
1	0.87	0.67	0.75	1190
accuracy			0.91	5584
macro avg	0.89	0.82	0.85	5584
weighted avg	0.90	0.91	0.90	5584

Fig. 12 Classification Report of Random Forest

3.3 Support Vector Machine

For our dataset Support Vector Machine secures an accuracy of 89%

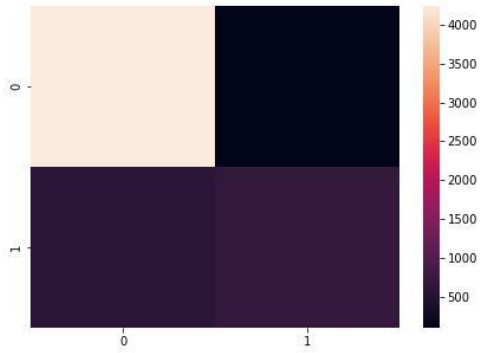

Fig. 13 Confusion Matrix of Support Vector Machine

	precision	recall	f1-score	support
0	0.89	0.97	0.93	4394
1	0.85	0.56	0.68	1190
accuracy			0.89	5584
macro avg	0.87	0.77	0.80	5584
weighted avg	0.88	0.89	0.88	5584

Fig. 14 Classification report of Support Vector Machine

3.4 Logistic Regression

For our dataset Logistic Regression secures an accuracy of 89%

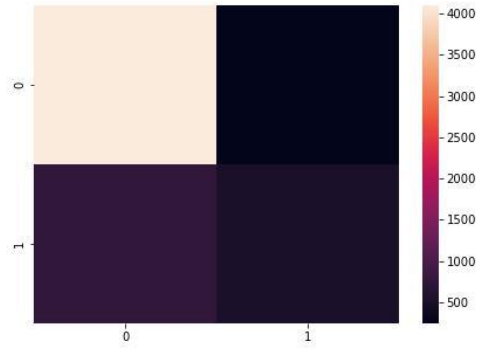

Fig. 15 Confusion Matrix of Logistic Regression

	precision	recall	f1-score	support
0	0.89	0.97	0.93	4394
1	0.85	0.56	0.68	1190
accuracy			0.89	5584
macro avg	0.87	0.77	0.80	5584
weighted avg	0.88	0.89	0.88	5584

Fig. 16 Classification Report of Logistic Regression

4. Conclusion

The objective of this study was to investigate, examine, and construct a machine learning model that can accurately determine if an individual, based on certain attributes, is likely to default on a loan. This type of usage of machine learning algorithms can help the Banks and other financial Institution to recognize specific financial characteristics of prospective borrowers that may indicate the risk of defaulting and failing to repay their loan within the specified timeframe.

This prediction process begins with data cleaning and processing, which involves imputing missing values and performing experimental analysis of the dataset, followed by Model building to evaluation of those models. For our given dataset the best accuracy obtained is almost 92% by Random Forest Classifier followed by Support Vector Machine (89%), Decision Tree (85%) and Logistic Regression (83%).

Although for this given dataset, we got this type of accuracy but all the models are proved to be accurate and useful in different settings. We shouldn't train a model on a particular dataset of particular domain and test that model on a completely different domain, as that may cause loss of accuracy. This problem is being explored by various researchers by employing ensemble techniques. [14] Single classifiers are proven to be outperformed by ensemble techniques. [7,12,15]

Most of the studies related to Credit risk prediction are focused on accuracy of the model, but we shouldn't ignore the repercussions of the False Negatives which pose a greater threat to any lending company. Hence, future researchers should have a keen eye towards false negatives while working on problems related to loan lending and prediction.

5. Appendix

Table 1: Definition of Data Features

Feature	Definition
person_age	Age of the individual
person_income	The annual income of the individual
person_home_ownership	Type of home ownership – rental, mortgage, rent, own or other
person_emp_length	Length of employment of the individual (in years)
loan_intent	Intention behind the loan – refer to Table
loan_amnt	Amount reimbursed towards the borrower
loan_int_rate	Interest rate towards the loan
loan_status	Status of the loan repayment (0 is non-default, 1 is default)
loan_percent_income	Percentage of loan amount by total income
default_hist	History of the defaults (whether any) done by the individual
cred_hist_length	Length of the credit history of the individual

Table 2: Loan Intention

Replaced as	Data Name
0	Debt Consolidation
1	Education
2	Home Improvement
3	Medical
4	Personal
5	Venture

Table 3: Personal Home Ownership

Replaced as	Data Name
0	Other
1	Mortgage
2	Rent
3	Own

6. References

1. Ahmed M S I and Rajaleximi P R 2019 An empirical study on credit scoring and credit scorecard for financial institutions Int. Journal of Advanced Research in Computer Engineering & Technol. (IJARCET) 8 275–9
2. Arun, K., Ishan, G. and Sanmeet, K., 2016. Loan Approval Prediction based on Machine Learning Approach. National Conference on Recent Trends in Computer Science and Information Technology (NCRTCSIT-2016), pp.18–21.
3. Aslam U, Aziz H I T, Sohail A and Batcha N K 2019 An empirical study on loan default prediction models Journal of Computational and Theoretical Nanoscience 16 pp 3483–8
4. Baesens, B., Roesch, D. and Scheule, H., 2016. Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS. United States, John Wiley & Sons.
5. Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001).
6. Credit Risk Dataset link from Kaggle: https://www.kaggle.com/code/juniorbueno/analyzing-credit-default/notebook#Data-Exploration
7. Hung C, Chen J-H (2009) A selective ensemble based on expected probabilities for bankruptcy prediction. Expert Syst Appl 36:5297–5303, 04
8. Li Y 2019 Credit risk prediction based on machine learning methods The 14th Int. Conf. On Computer Science & Education (ICCSE) pp 1011–3
9. Marqués, A.I., García, V. and Sánchez, J.S., 2012. Exploring the behaviour of base classifiers in credit scoring ensembles. Expert Systems with Applications, 39(11), pp.10244–10250.
10. Nali´c, J. and Švraka, A., 2018. Using Data Mining Approaches to Build Credit Scoring Model: Case Study—Implementation of Credit Scoring Model in Microfinance Institution. 2018 17th International Symposium Infoteh-Jahorina (INFOTEH), IEEE. pp.1–5.
11. Nehrebecka, N., 2018. Predicting the default risk of companies. Comparison of credit scoring models: LOGIT versus support vector machines. Econometrics, 22(2), pp.54–73.
12. Wang G, Hao J, Ma J, Jiang H (2011) A comparative assessment of ensemble learning for credit scoring. Expert Syst Appl 38(1):223–230

13. Wanga Y, Zhanga Y, Lua Y, Yua X A comparative Assessment of Credit Risk Model Based on Machine Learning - a case study of bank loan data Procedia Computer Science Volume 174, 2020, Pages 141-149
14. Xia Y, Liu C, Li YY, Liu N (2017) A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring. Expert Syst Appl 78:02
15. Yu L, Wang S, Lai KK (2008) Credit risk assessment with a multistage neural network ensemble learning approach. Expert Syst Appl 34(2):1434–1444