Loan default prediction using decision trees and random forest: A comparative study

Loan default prediction using decision trees and random forest: A comparative study

Mehul Madaan[1], *, Aniket Kumar[1], Chirag Keshri[1], Rachna Jain[2] and Preeti Nagrath[2]

[1] Department of Electronics and Communication Engineering, Bharati Vidyapeeth's College of Engineering, GGSIP University, New Delhi, India

[2] Department of Computer Science and Engineering, Bharati Vidyapeeth's College of Engineering, GGSIP University, New Delhi, India

*E-mail: mehul_madaan@ieee.org

Abstract. With the improving banking sector in recent times and the increasing trend of taking loans, a large population applies for bank loans. But one of the major problem banking sectors face in this ever-changing economy is the increasing rate of loan defaults, and the banking authorities are finding it more difficult to correctly assess loan requests and tackle the risks of people defaulting on loans. The two most critical questions in the banking industry are (i) How risky is the borrower? and (ii) Given the borrower's risk, should we lend him/her? In light of the given problems, this paper proposes two machine learning models to predict whether an individual should be given a loan by assessing certain attributes and therefore help the banking authorities by easing their process of selecting suitable people from a given list of candidates who applied for a loan. This paper does a comprehensive and comparative analysis between two algorithms (i) Random Forest, and (ii) Decision Trees. Both the algorithms have been used on the same dataset and the conclusions have been made with results showing that the Random Forest algorithm outperformed the Decision Tree algorithm with much higher accuracy.

Keywords: Credit Risk, Credit Score, Data Analysis, Decision Trees, Loan Prediction, Machine Learning, Random Forest

1. Introduction

Individuals all around the world in some way depend on banks to lend them loans for various reasons to help them overcome their financial constraints and achieve some personal goals. Due to the ever-changing economy and ever-increasing competition in the financial world, the activity of taking a loan has become inevitable. Also, small scale to large scale banking firms depend on the activity of lending out loans to earn profits for managing their affairs and to function smoothly at times of financial constraints. A loan is the major source of income for the banking sector as well as the biggest source of financial risk for banks. Large portions of a bank's assets directly come from the interests earned on loans given.

Though lending loans is quite beneficial for both the parties, the activity does carry great risks. These risks represent the inability of a borrower to pay back the loan by the designated time which was decided mutually by both the lender and the borrower and it is referred to as 'Credit Risk' [1]. For that,

it is highly necessary to assess the clients' credit suitability before authorizing a loan. In the traditional lending process, banking authorities mainly adopt the '5C principle', i.e., Character, Capital, Capacity, Collateral, and Conditions to evaluate a borrower [2]. This evaluation mainly relied on personal experience and knowledge of customer dealing. This method has great limitations. Banks and large financial firms approve loan requests after a verification and validation process of regress, but still, there's no guarantee that the applicant selected after the process will be able to repay the loan in time.

While in the past, banks used to hire highly professional individuals whose sole purpose was to evaluate applicants and after close review decide and tell whether a candidate was eligible for receiving a loan. The worthiness of a candidate for loan approval or rejection was based on a numerical score called 'Credit Score'. Generally, the credit score helps the authorities to compute the probability of borrower repaying the loan by the designated time based on their credit history or payment history along with their background [3].

The process of credit scoring required experts alongside statistical algorithms to accurately predict the creditworthiness of an applicant. However, quite recently, the researchers and the banking authorities have opted for training classifiers based on various machine learning and deep learning algorithms to automatically predict the credit score of an applicant based on their credit history and other historical data and make the process of selecting the eligible candidates a lot easier before the loan is approved.

Therefore, addressing the aforementioned scenario, the goal of this paper is to discuss the application of different machine learning models in the loan lending process and work out the best approach for a financial institution which accurately identifies whom to lend loan to and help banks identify the loan defaulters for much-reduced credit risk. Classifiers that we used to build the model are Random Forest and Decision Trees. They'll be used separately to analyse the dataset and identify the patterns in the dataset and learn from those. Based on that analysis, predict whether a new applicant is likely to default on a loan or not.

The rest of the paper has the following sections: Section 2 covers a brief literature review of the approaches that have been used on credit risk analysis. Section 3 talks briefly about what machine learning is and discusses the two algorithms used in the paper, i.e., the Random Forest algorithm and Decision Trees. Section 4 presents an introduction to the dataset used to train and test the model. Section 5 introduces our methodology in this work which covers the data analysis, cleaning the dataset, and going through the model we built using the two classifiers. Section 6 presents the results of the model after the training and testing is over and shows the results of the comparative analysis done for the performance difference of the two classifiers. Section 7 concludes the paper with final remarks and future works.

2. Literature Review

This section discusses in brief about some of the work that has already been done on creating ML and DL models using various algorithms to improve the loan prediction process and help the banking authorities and financial firms select an eligible candidate with very low credit risk.

Loan prediction is a much-talked-about subject in the sectors of banking and finance. Credit scoring has become a key tool for the same in this competitive financial world. Furthermore, following the recent improvements in data science and several notable developments in the field of artificial intelligence, this topic has gained more attention and research interest. In recent years, it has attracted more focus towards research on loan prediction and credit risk assessment. Due to the high demands of loan now, demand for further improvements in the models for credit scoring and loan prediction is increasing significantly. A multitude of techniques have been used to assign individuals a credit score and much research has been done over the years on the topic. Unlike previously, where experts were hired and the models depended on professional opinions were used for assessing the individual's creditworthiness, the focus has shifted to an automated way of doing the same job. In recent years, the researchers and

banking authorities have been focused on applying machine learning algorithms and neural networks for credit scoring and risk assessment. Many noteworthy conclusions have been drawn in this regard which serve as stepping-stones for researches and studies.

The Random Forest Algorithm was adopted by Lin Zhu et al. in paper [4] and Nazeeh Ghatasheh in paper [5] to construct a model for loan default prediction. Paper [4] concluded that random forest has much better accuracy (98%) than other algorithms like logistic regression (73%), decision trees (95%), and support vector machines (75%). The results of the paper [5] concluded that the random forest algorithm is one of the best options for credit risk prediction. Paper [5] also talked about the advantages of the algorithms, which are the competitive classification accuracy and simplicity.

Paper [6] reviewed many methods available like logistic regression, k- nearest neighbours, random forest, neural networks, support vector machines, stochastic gradient boosting, Naive Bayes, etc. and concluded that it is nearly impossible to declare one best method of all.

Nikhil Madane and Siddharth Nanda in paper [7] reviewed credit scoring of mortgage loans and made the following conclusions:

- Credit applications that do not pass certain requirements are often not accepted because the probability of them not paying back is high.
- Low-income applicants are more likely to get approval, and they are more likely to pay back their loans in time.

Pidikiti Supriya et al. [8] used Decision Trees as a machine learning tool to implement their model. They started their analysis with data cleaning pre-processing, missing value imputation, then exploratory data analysis, and finally model building and evaluation. The authors on a public test set managed to achieve the best accuracy of 81%.

The conducted tests using the C4.5 algorithm in decision trees in paper [9] showed that the maximum precision value achieved was 78.08% with data partition of 90:10 and the biggest recall value was 96.4% with data partition of 80:20. Therefore partition of 80:20 was concluded to be the best due to its highest accuracy and high recall value.

The authors in paper [10] did an exploratory data analysis. The paper's main purpose was to classify and examine the nature of loan applicants. Seven different graphs were plotted and visualized and using these graphs the authors concluded that most loan applicants preferred short-term loans.

Syed Zamil Hasan Shoumo et al. [11] concluded that Support Vector Machines are capable of outperforming other models like logistic regression, random forest, etc. that have been used in the paper for comparative performance analysis.

The authors in paper [12] selected 4 different models:

- M1: Logistic Regression model
- M2: Random Forest model
- M3: Gradient Boosting model
- D1-D4: Multilayer Neural Network models (deep learning)

And using these models they showed that data quality check is important, i.e., data analysis and cleaning before modelling to omit redundant variables. The paper also concluded that the choice of features and the algorithm are two major aspects when deciding whether to give an individual a loan or not.

In paper [13], Aboobyda Jafar Hamid, and Tarig Mohammed Ahmed used Data Mining to build a model for classifying loan risk. They used three algorithms for it:

- J48
- Bayes Net
- Naive Bayes

The paper concluded that J48 was the best algorithm for the purpose because of its high accuracy (78.3784%) and low mean absolute error (0.3448).

Aditi Kacheria et al. [14] used the Naive Bayesian algorithm for their model. And to improve the classification accuracy, they used the k-NN and binning algorithms to improve the quality of the data. K-NN was used to deal with the missing values and the binning algorithm was used to remove the anomalies from the data set.

Martin Vojtek and Evzen Kocenda concluded that most local banks in the Czech Republic and Slovakia are using logit method-based models in paper [15]. Other methods like CART or neural networks are primarily used as support tools in the variable selection process or the process of model quality evaluation. The authors also concluded that the k-NN method is not used at all or is very rarely used.

Yu Li in paper [2] did a comprehensive study comparing the XGBoost algorithm's performance with the performance of logistic regression. The paper concluded that the model discrimination and model stability of the XGBoost model was substantially higher than that of the logistic regression model.

3. Theory

Nearly every sector in the world is advancing towards complete automation. Various concepts and methods are being developed every day to achieve this goal and many fields have been under study for many years. One of the most upcoming fields that have grabbed the attention and excitement of scientists, researchers, and technologists is Artificial Intelligence (AI).

AI is the idea of making a computer or a machine to simulate human-like intelligence and behaviour [16]. It dates to the time when computers were first built and has since diversified into various fields like Machine Learning (ML), Neural Networks, Natural Language Processing (NLP), etc. [16].

3.1. What is Machine Learning?

It is a concept that enables machines to learn from real-world interactions and observations and behave like human beings and improve their ability to learn and perform using data given as input [11]. In the recent years, ML has gained a huge focus and interest of researchers and technologists that they are trying to implement various machine learning models and algorithms in fields which will make various important tasks and lives of common man a lot easier. Two popular examples are the banking sector and finance. With the help of various ML models, banking authorities and financial firms are observing patterns and making conclusions in areas like credit card frauds, loan default prediction. It has made the process much easier now and more accurate.

The models mentioned above are based on various machine learning methods. It is almost impossible to compile and provide a list of all the ML methods. Usually, the name given to a model is a combination of data structure, design, estimator, ensemble mechanism, and more [6]. Regarding this paper, the two algorithms in the domain of machine learning used are Random Forest and Decision Trees.

3.2. Decision Trees

They are a versatile algorithm used to perform the tasks of classification and regression [7]. They are one of the most popular algorithms used for classification which comprise several branches, leaf nodes, and root nodes [1]. The algorithm generates a structure like a tree by classifying the instances and utilizing a Recursive Portioning Algorithm (RPA) [1]. A class label is represented by a leaf node and the branches represent test results. These tests are represented by internal nodes for an attribute [1].

3.3. Random Forest

Random Forest belongs to the supervised learning algorithm. Like decision trees, they are also used for classification and regression. A predictor ensemble is built with several decision trees that expand in randomly selected data subspaces [12].

Using Random Forest over other machine learning algorithms has many advantages like:

- Immunity to overfitting.
- Accurate classification or regression.
- More efficient on large databases.

4. Dataset Description

We'll be using the publicly available Lending Club dataset from Kaggle and prepare it accordingly to meet our goals. The data covers approximately 22 lakh loans funded by the platform between 2007 and 2015. The interest rate is provided to us for each borrower.

5. Proposed Model

5.1. Project Pre-Work

Before moving forward with machine learning modelling, a few steps were required to familiarize ourselves with the Lending Club dataset. The first important step was to import the necessary libraries and data files required for the model. And the second step was to do an exploratory data analysis (EDA) of the given data to examine its features and answer the following questions.

- What are the characteristics of each loan?
- What features make them different or similar?
- How to best explain the data?
- What are the most important characteristics for classification purposes?
- Which method would be the most effective to clean the dataset as per our needs?

Working on the above questions will eventually help develop a better understanding of the dataset and will guide an effective machine learning model.

5.2. Data Cleaning

There are many columns with null values in the dataset. It is necessary to identify the percentage of null values in each column to drop certain columns that don't meet a percentage threshold. Data cleaning needs to be done before performing the Exploratory Data Analysis.

Figure 1 shows the datatype of columns left after dropping those columns which were not needed for the model.

```
loan_amnt              int64
term                  object
int_rate             float64
installment          float64
grade                 object
sub_grade             object
emp_length            object
home_ownership        object
issue_d               object
verification_status   object
purpose               object
dti                  float64
delinq_2yrs          float64
loan_status           object
zip_code              object
avg_cur_bal          float64
revol_bal              int64
dtype: object
```

Figure 1. Columns with their respective datatype left after dropping those columns which did not meet the percentage threshold.

```
loan_amnt:0
term:0
int_rate:0
installment:0
grade:0
sub_grade:0
home_ownership:0
purpose:0
dti:1711
loan_status:0
zip_code:1
revol_bal:0
```

Figure 2. Count of null values in each column left after more cleaning.

The dataset is cleaned further. Figure 2 shows the count of null values in each column.

5.3. Exploratory Data Analysis

Exploratory Data Analysis (EDA) played an integral part in understanding the Lending Club dataset. It was vital to get familiar with different relationships within the data through different types of plots before moving towards classification. Analysing these relationships helped us with interpreting the outcomes of the models. Asking questions about these relationships provided us with additional knowledge about relationships that we may not have known existed. This section will further investigate data distribution and ask specific questions about the data lying within the dataset.

Lending Club has nine categories of Loan Status. Figure 3 shows the count values for each Loan Status category

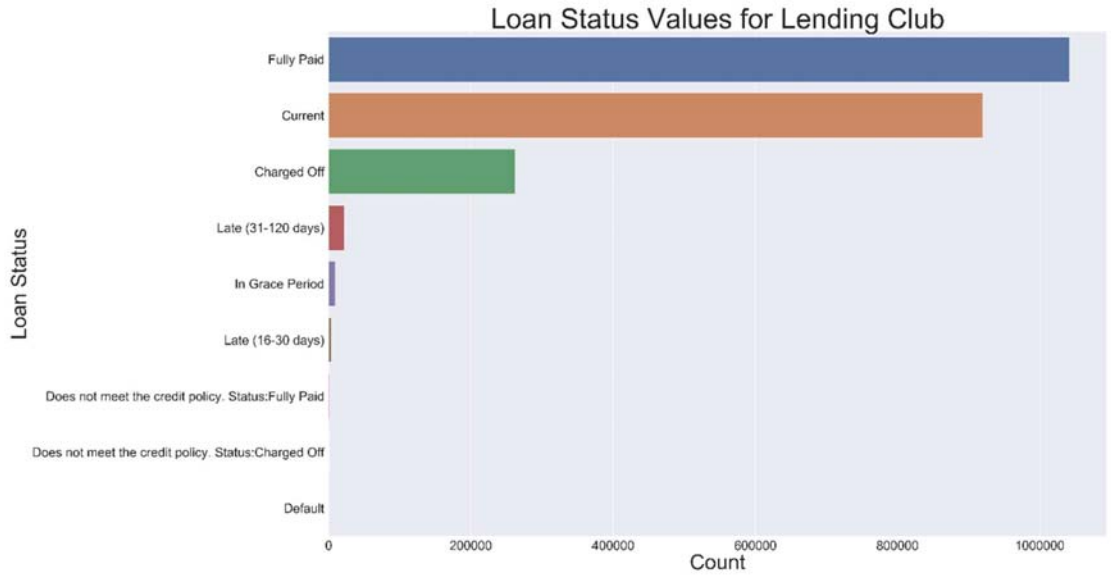

Figure 3. Count plot to show the number of values in each Loan status category.

Lending Club has classified loans into seven grades, A - G. And each grade represents a certain level of risk associated.

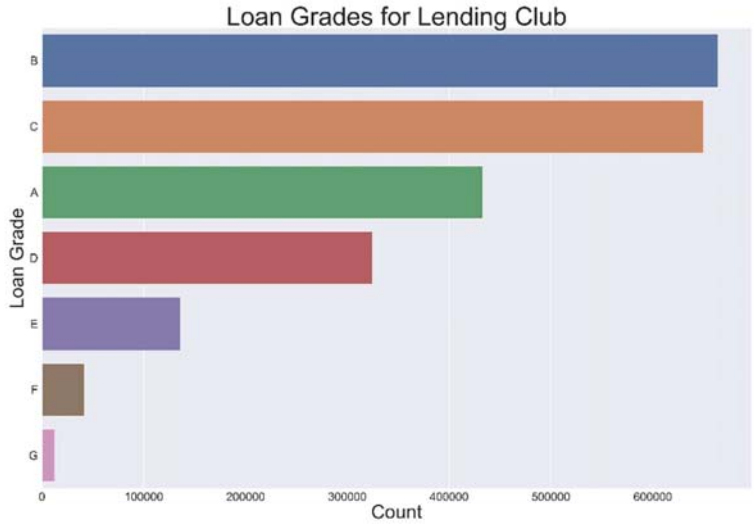

Figure 4. Count plot for Loan Grades showing the number of values in each grade.

From figure 4, we can infer that the most popular grades are B and C. Grade A comes in third. And the least popular grades are F and G.

Lending Club has also classified borrowers into six home-ownership types.

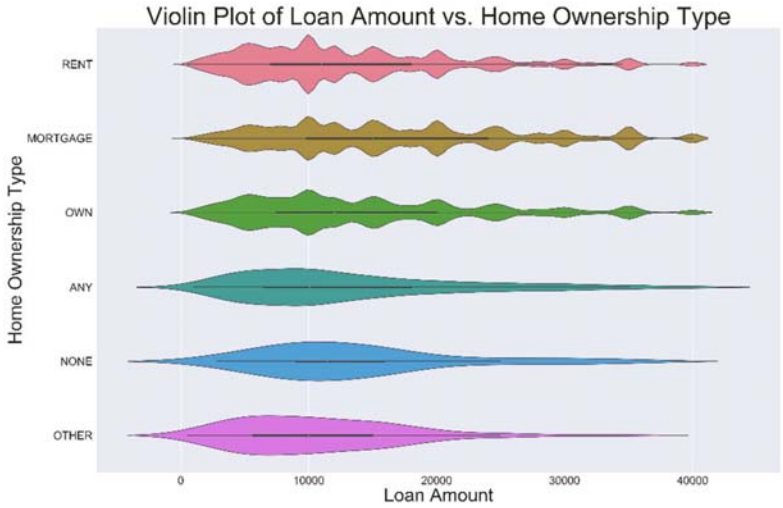

Figure 5. Violin plot of Loan Amount vs Home Ownership type.

The violin plot in figure 5 shows the mean and density distribution for loan amounts for each home-ownership type. Borrowers with mortgages have the highest mean loan amount and borrowers with rent have the lowest mean loan amount.

The plot shown in figure 6 shows a box plot of Loan amount vs Loan Purpose. According to the plot, there are 14 different purposes according to the Lending Club dataset individuals apply for loans.

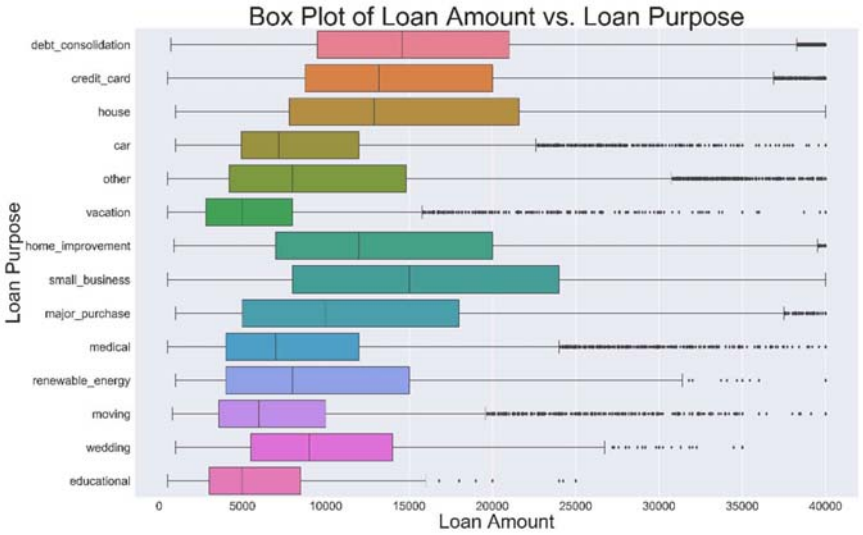

Figure 6. Box plot of Loan Amount vs Loan purpose.

We also plotted a correlation between the features to get a better understanding if any other alteration must be made to make the model perform better.

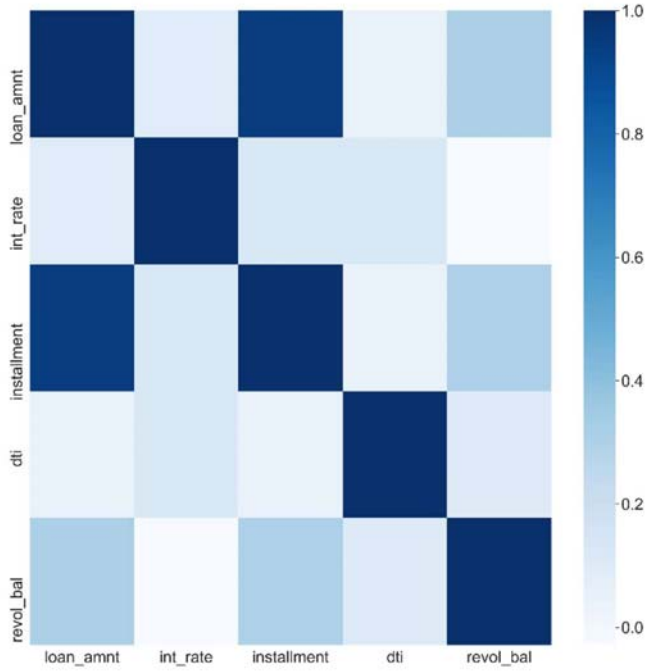

Figure 7. Correlation heatmap between all the features.

From figure 7, we deduced that there was a strong correlation between instalment values and loan amount. This multicollinearity had to be fixed to prevent overfitting of the model because these two features explain the data in a very similar manner. Most machine learning models carry assumptions that call for little multicollinearity.

5.4. Modelling

Machine learning is about predicting and recognizing patterns and generate suitable results after understanding them. ML algorithms study patterns in data and learn from them. An ML model will learn and improve on each attempt. To gauge the effectiveness of a model, it's vital to split the data into training and test sets first. So before training our models, we split the Lending Club data into Training set which was 70% of the whole dataset and Test set which was the remaining 30%.

Then it was important to implement a selection of performance metrics to the predictions made by our model. In this case, we tried to identify whether an individual is going to default on a loan or not. Model accuracy might not be the sole metric to identify how our model performed- the F1 score and confusion matrix should be important metrics to analyse as well. What is important is that the right performance measures are chosen for the right situations.

We used 2 algorithms for our modelling purpose:

5.4.1. Decision Tree. It was the first algorithm we used for the model. The first thing we did was import the necessary libraries using Scikit-Learn and create a variable for the decision tree classifier. And then fit the data accordingly to train the decision tree model followed by prediction on the test data. The number of nodes in the decision tree formed was 355489 and the depth of the tree was 241.

5.4.2. Random Forest. The second algorithm used for the model. The first step we did was import the necessary libraries using the Scikit-Learn library and create a variable for the random forest classifier. We set the estimator count to 100 for the random forest model. And then fit the data accordingly using the fit function on the classifier followed by prediction on the test data.

We further compared the efficiency of the two models which will be shown in the next section.

6. Results

In this paper, we used two machine learning algorithms, the Random Forest and Decision Trees to work out a model for loan prediction and credit risk assessment.

The results of both the model are shown below with their classification report and confusion matrix to get a better understanding of the accuracy and other scores of the two models.

6.1. Decision Tree

The Decision Tree classifier gave us an accuracy score of 73%.

Table 1. Classification Report for Decision Tree.

	Precision	Recall	F1 score	Support
0	0.82	0.85	0.83	312588
1	0.29	0.24	0.26	78504
Accuracy			0.73	391092
Macro Avg	0.55	0.55	0.55	391092
Weighted Avg	0.71	0.73	0.72	391092

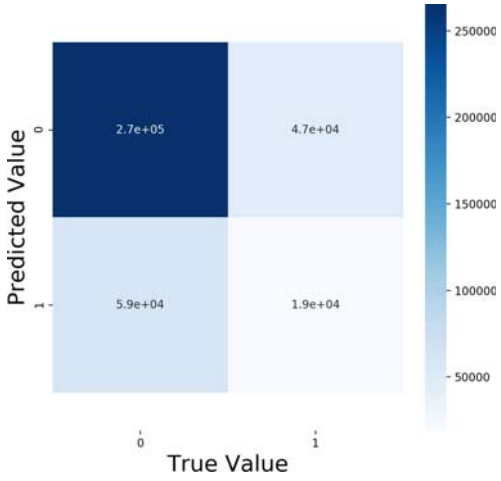

Figure 8. Confusion Matrix of Decision Tree.

6.2. Random Forest

The Random Forest Classifier gave us an accuracy score of 80%.

Table 2. Classification Report for Random Forest.

	Precision	Recall	F1 score	Support
0	0.81	0.98	0.89	312588
1	0.50	0.08	0.14	78504
Accuracy			0.80	391092
Macro Avg	0.65	0.53	0.51	391092
Weighted Avg	0.75	0.80	0.74	391092

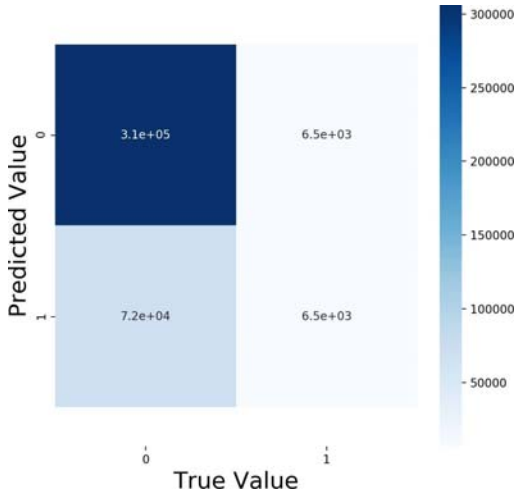

Figure 9. Confusion Matrix of Random Forest.

Looking at the above confusion matrices and classification reports for both the models, one can easily say that the Random Forest algorithm is a much better option over Decision Trees for loan prediction on the given dataset.

7. Conclusion

This paper aimed to explore, analyse, and build a machine learning algorithm to correctly identify whether a person, given certain attributes, has a high probability to default on a loan. This type of model could be used by Lending Club to identify certain financial traits of future borrowers that could have the potential to default and not pay back their loan by the designated time.

The Random Forest Classifier provided us with an accuracy of 80% while the Decision Tree method provided us with an accuracy of 73%. Hence, the Random Forest model appears to be a better option for such kind of data.

Lending Club must be careful when identifying potential borrowers who fit certain criteria. For example, borrowers who do not own a home and are applying for a small business or wedding loan, this could be a negative combination that results in the borrower defaulting on a loan.

One of the drawbacks is simply the limited number of people who defaulted on their loan in the 8 years of data (2007-2015). We could use an updated data frame that consists of the next 3 years' values (2015-2018) and see how many of the current loans were paid off, defaulted, or even charged off. Then, these new data points can be used for prediction or and training new models for better and more accurate results.

Since the algorithm puts some of the non-defaulters in the default class, we might want to look further into this issue to help the model accurately predict capable borrowers.

References

[1] 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

[2] Li Y 2019 Credit risk prediction based on machine learning methods *The 14th Int. Conf. on Computer Science & Education (ICCSE)* pp 1011–3

[3] 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

[4] Zhu L, Qiu D, Ergu D, Ying C and Liu K 2019 A study on predicting loan default based on the random forest algorithm *The 7th Int. Conf. on Information Technol. and Quantitative Management (ITQM)* **162** pp 503–13

[5] Ghatasheh N 2014 Business analytics using random forest trees for credit risk prediction: a comparison study *Int. Journal of Advanced Science and Technol.* **72** pp 19–30

[6] Breeden J L 2020 *A survey of machine learning in credit risk*

[7] Madane N and Nanda S 2019 Loan prediction analysis using decision tree *Journal of The Gujarat Research Society* **21** p p 214–21

[8] Supriya P, Pavani M, Saisushma N, Kumari N V and Vikas K 2019 Loan prediction by using machine learning models *Int. Journal of Engineering and Techniques* **5** pp 144–8

[9] Amin R K, Indwiarti and Sibaroni Y 2015 Implementation of decision tree using C4.5 algorithm in decision making of loan application by debtor (case study: bank pasar of yogyakarta special region) *The 3rd Int. Conf. on Information and Communication Technol. (ICoICT)* pp 75–80

[10] Jency X F, Sumathi V P and Sri J S 2018 An exploratory data analysis for loan prediction based on nature of the clients *Int. Journal of Recent Technol. and Engineering (IJRTE)* **7** pp 176–9

[11] Shoumo S Z H, Dhruba M I M, Hossain S, Ghani N H, Arif H and Islam S 2019 Application of machine learning in credit risk assessment: a prelude to smart banking *TENCON 2019 – 2019 IEEE Region 10 Conf. (TENCON)* pp 2023–8

[12] Addo P M, Guegan D and Hassani B 2018 Credit risk analysis using machine and deep learning models *Risks* **6** p 38

[13] Hamid A J and Ahmed T M 2016 Developing prediction model of loan risk in banks using data mining *Machine Learning and Applications: An Int. Journal (MLAIJ)* **3** pp 1–9

[14] Kacheria A, Shivakumar N, Sawkar S and Gupta A 2016 Loan sanctioning prediction system *Int. Journal of Soft Computing and Engineering (IJSCE)* **6** pp 50–3

[15] Vojtek M and Kocenda E 2006 Credit scoring methods *Finance a uver - Czech Journal of Economics and Finance* **56** pp 152–167

[16] Russel S and Norvig P 1995 *Artificial intelligence - a modern approach*

[17] Alshouiliy K, Alghamdi A and Agrawal D P 2020 AzureML based analysis and prediction loan borrowers creditworthy *The 3rd Int. Conf. on Information and Computer Technologies (ICICT)* **1** pp 302–6

[18] Li M, Mickel A and Taylor S 2018, "Should this loan be approved or denied?": a large dataset with class assignment guidelines *Journal of Statistics Education* **26** pp 55–66

[19] Vaidya A 2017 Predictive and probabilistic approach using logistic regression: application to prediction of loan approval *The 8th Int. Conf. on Computing, Communication and Networking Technologies (ICCCNT)* **1** pp 1–6

[20] Murphy K P 2012 *Machine learning: a probabilistic approach*