# Loan default prediction using Machine Learning Algorithms

R.M. Sanduni Kanishka Chamodani
University of Dalarna
Borlänge, Sweden
e-mail: v23skrmu@du.se

*Abstract*— **The objective of this study is to create accurate predicting model for loan default. Accurate loan default prediction can be beneficial for the financial industry's stability, profitability, and responsible lending practices. It benefits both lenders and borrowers while also contributing to the overall health of the economy. For creating the prediction model, 3 methods were selected such as Random Forest, Logistic Regression and Decision tree. Models were created using those 3 methods and then the models were evaluated. Then cross validation is done to assess the effectiveness and robustness of the models. K fold cross validation is used in here. Random forest model is selected as the best performed model. Finally, important factors which have impact for loan default were selected using Random Forest. According to the study, interest rate, annual income, dti and revol balance have the highest impact on loan default.**

**Keywords-Random Forest; Decision Tree; Logistic Regression; cross validation; loan; analysis; data visualization;**

## I. INTRODUCTION

As the primary source of revenue and profit for banks, loans are of utmost importance. Banks offer loans to people, companies, and organizations so they may fund a variety of things including purchasing a home, growing their business, or investing in projects. The interest on these loans brings in a sizable amount of money for the banks. Giving the right applicants the right amount of loans, nevertheless, is essential. Large amounts of bad debt may result in capital adequacy issues for banks and, in the worst-case scenario, bankruptcy. Therefore, it is necessary to determine whether or not the loan they offered to clients would be in default. To meet that criterion, businesses can choose which clients to lend money to and how much money to lend by using a system that predicts loan default.

However, it is also crucial for the customer to determine if he will be able to repay the loan in full. As a result, he can prevent his economy from becoming bankrupt and decide how much money he may borrow up to before defaulting.

The main objective of this project is to create a loan default prediction model using several methods. In here, Random Forest, Decision Tree and Logistic Regression methods are used.

The following research questions will be addressed in this project:

1. How accurately can above data mining techniques predict the loan default based on above features?
2. Which factors have the biggest impact on loan default?
3. Which data mining technique is most efficient in predicting loan default among Random Forest, Logistic Regression and Decision tree?

The goal of this study is to develop an accurate and trustworthy loan forecasting model utilizing the random forest algorithm, decision trees, and logistic regression, as well as to compare the effectiveness of each method and identify the best model. Additionally, the study aims to pinpoint the major risk factors for loan default and reveal the risk variables that contribute to the loan default.

## II. LITERATURE REVIEW

Predicting loan defaults is a critical problem for the financial sector, having repercussions for lenders, borrowers, and the whole economy. By highlighting important findings, methodology, and trends in the area, this literature review seeks to give a broad overview of the present research on the prediction of loan default. Loan default prediction has been a trending topic for past years.

Some researches were aimed to find the factors that affecting loan defaults. As an example, the paper [1] tries to find out the factors that contribute to student loans. In there the author shows that providing loans to low- and moderate-income students comes with a higher risk of default since these students frequently come from families with a poor credit history and may be more likely to drop out of school or take jobs with lesser pay. Also, he suggested several options to avoid the student loan default in that paper.

The paper [2] focuses on a unique set of microdata on defaulted bank loans from a European bank and performed mortality analysis. The empirical findings include the timing of recoveries on bad and dubious bank loans, the

distribution of cumulative recovery rates, their economic causes, and the direct expenses spent by that bank during recoveries on bad and dubious loans.

The paper [3] proposed an improved random forest algorithm which, during tree aggregation for prediction, assigns weights to decision trees in the forest. Also, this paper uses balanced random forests to deal with imbalanced data. In terms of both balanced and overall accuracy measures, this suggested approach outperforms the original random forest.

Numerical studies focused on online peer to peer lending. The paper [4] suggest a loan default prediction model to predict whether a borrower will default on a loan is of significant concern to platforms and investors in online peer-to-peer (P2P) lending. The author there makes a suggestion for a technique of P2P loan default prediction that combines soft data from textual description. He introduces a topic model to extract useful characteristics from the loan-related descriptive text and builds four default prediction models to show how well these features work for default prediction.

In study [5] , a brand-new profit-driven prediction model is put forward, using a profit indicator as the Bayesian optimization's optimization goal to enhance the predictor-categorical boosting's hyperparameters. The link between the input variables and the predicted values is then further analyzed by calculating the Shapley additive explanations (SHAP) value.

In paper [6] , the author compares the performances of five machine learning models from different families, including XGBoost, random forest (RF), AdaBoost, k nearest neighbors (kNN), and multilayer perceptrons (MLP), to determine whether loan default will occur or not based on real-world data provided by Xiamen International Bank. In this study, the author not only analyzes the prediction outcome but also presents the preprocessing procedure.

The application of data mining techniques for the prediction and categorization of loan failures was the primary focus of research [7]. New methods such as KDD, CRISP-DM, and SEMMA were the techniques employed in this work.

## III. METHOD DESCRIPTION

### *A. The Dataset*

Dataset for loan default prediction was taken from Kaggle. The dataset consists of 37426 rows and 22 columns. Columns are as follows:

Table 1: Feature description

| Variable | Variable Type | Has null values? |
|---|---|---|
| | | |
| loan_amnt | numerical | yes |
| funded_amnt | numerical | yes |
| funded_amnt_inv | numerical | yes |
| term | numerical | no |
| int_rate | numerical | no |
| installment | numerical | yes |
| emp_length | categorical | yes |
| home_ownership | categorical | no |
| annual_inc | numerical | yes |
| verification_status | categorical | no |
| purpose | categorical | no |
| dti | numerical | no |
| delinq_2yrs | numerical | yes |
| open_acc | numerical | yes |
| pub_rec | categorical | yes |
| revol_bal | numerical | yes |
| revol_util | numerical | yes |
| total_acc | numerical | yes |
| total_pymnt | numerical | yes |
| total_pymnt_inv | numerical | yes |
| repay_fail | categorical | no |
| | | |

Python is the programming language used in this research. Below are the libraries that used in this study.

1. numpy : This brings the computational power to the programming language.
2. pandas: Pandas is used for analyzing, cleaning, exploring, and manipulating data.
3. sklearn: Stands for Scikit-learn and used to implement machine learning models.
4. Matplotlib: Used to create plots such as box plots, correlation heat maps and so on

### *B. Data Preparation and EDA*

It is crucial to prepare the data for analysis since it enhances data quality, removes missing values, outliers, and inconsistencies, and makes the data more reliable. Exploratory data analysis is crucial because it aids in understanding the data, reveals patterns, correlations, and trends, and produces insights that inform decision-making and further analysis.

For the data preparation, first the data frame was checked for missing values and removed those records. Then the duplicate rows were dropped from the data frame. Then data visualization was done for several variables to get the insight of data.

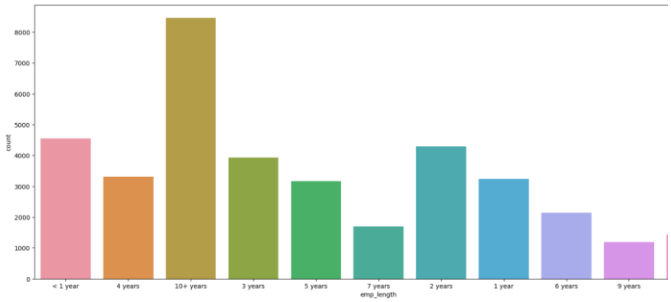

Figure 1: Record counts vs Employment length

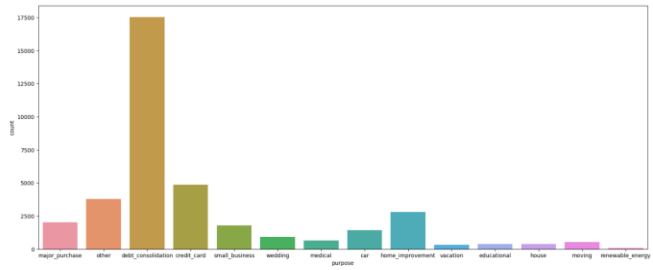

Figure 2: Record counts vs purpose

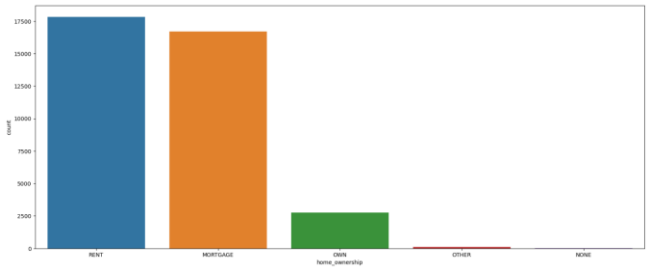

Figure 3: Record counts vs home ownership

Figure 1, 2, 3 shows the data visualizations for some features of the dataset. Figure 1 was graphed using the employment length of the dataset. According to that graph, the highest count of values that consist of dataset are 10+ years and the least count of values are 9 years. Figure 2 was drawn to visualize the purpose of the loan. According to that, it is obvious majority of people took the loan for debt consolidation and credit cards. The minority of people took loan for renewable energy. Figure 3 shows the behavior of home ownership. According to that, the majority of people who got the loan are mentioned as rent for the home ownership. Like this, we can find out the behavior of each feature and get an idea of them.

Then the categorical values were converted to the numerical values by label encoding. In here, 'term', 'emp_length', 'home_ownership', 'verification_status', purpose and 'revol_util' are converted to numerical format. Then the boxplots are drawn to find out the outliers.

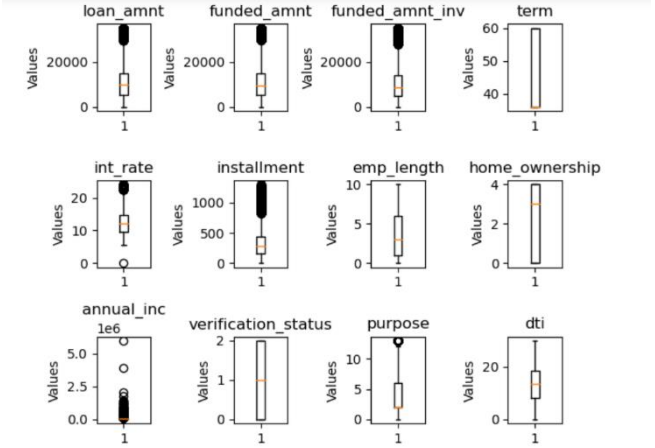


Figure 4: Box plot of variables - 1

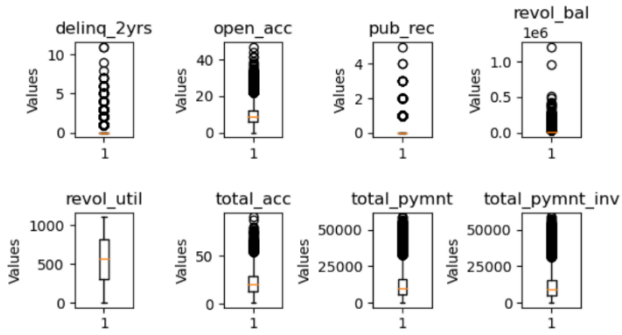


Figure 5: Box plot of variables – 2

According to the figure 4 and figure 5, it is obvious that loan_amnt, funded_amnt, funded_amnt_inv, int_rate, installment, annual_inc, open_acc, revol_bal, toatal_acc, total_pymnt and total_pymnt_inv have outliers. Outliers can have a disproportionate effect on statistical results, which can result in misleading interpretations. Therefore, they have to be removed. The interquartile range (IQR) technique is used to eliminate the outliers. The first and third quantile values are used in this approach to determine if an observation is an outlier or not. If an observation is 1.5 times the interquartile range higher than the third quartile or 1.5 times the interquartile range lower than the first quartile, it is considered an outlier.

The correlation plot was created as the following stage. The correlation plot was shown in figure 6. It can be used to find duplicate or strongly associated variables. Strongly correlated variables may produce redundant information and overfitting or lower interpretability in a model if both are included. According to the correlation plot,

1. 'installment' is highly corelated with 'loan_amnt', therefore 'installement' was removed.
2. 'total_pymnt' and 'total_pymnt_inv' are corelated with 'loan_amnt'. Therefore, both two were removed.

3. Also, ‘funded_amnt’ and ‘funded_amnt_inv’ are corelated with each other and also corelated with ‘loan_amnt’. Therefore, those two were removed.

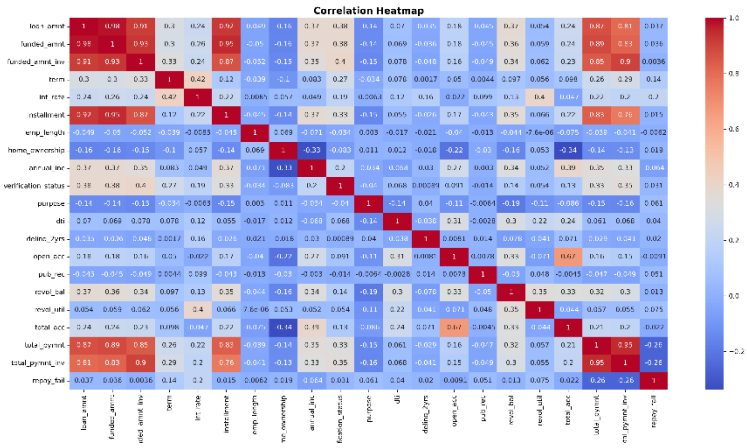


Figure 6: Correlation Map

After that, numerical variables were normalized since doing so ensures that characteristics of different scales or units are equivalent, allowing for fair comparisons and prevents the study from being dominated by any specific element. It also improves the convergence of machine learning optimization algorithms by establishing features on comparable sizes.

### *C. Data Mining Method*

For data modeling, first data was splitted into 2 parts such as training and testing data. 80% of data was given for training and 20% of data was given for testing. Below methods were used for modeling.

1. Random forest
   - RandomForestClassifier in sklearn. ensemble library is used to train the model using random forest algorithm
2. Decision tree
   - DecisionTreeClassifier in sklearn.tree library is used to train the model using decision tree algorithm
3. Logistic regression
   - LogisticRegression in sklearn.linear_model library is used to train the model using logistic regression [8] algorithm

Then data were trained using these models.

Then to measure the performance, confusion matrix is created. Also, accuracy_score, mean_squared_error, f1_score and classification_report is calculated. Finally, the factors that have biggest impact on loan default were selected using the model.

### *D. Cross Validation*

In order to assess the effectiveness and robustness of the models, cross-validation is a crucial stage in modeling tasks. Cross-validation aims to identify issues like overfitting or selection bias and to provide insight into how the model will generalize to an independent dataset by testing the model's propensity to predict fresh data that was not included in its estimation. In this project, k fold cross validation method is used to perform cross validation. The input data are divided into k subsets of data (also known as folds) for k-fold cross-validation. The model is trained on all of the subsets except for one (k-1) before being evaluated on the subset that wasn't utilized for training. A distinct subset is put aside for evaluation (and left out of training) for each of the k times this process is repeated. Results obtained from each model are discussed in the next chapter.

## IV. Results and Analysis

All the models were tested using test data to calculate the accuracy and performance. This analysis helps to find out the performance and the reliability of models. Accuracy of each model are as follows

Table 2: Accuracy of models

| Model | Accuracy |
|---|---|
| | |
| Random Forest | 84.84% |
| Decision Tree | 75.53% |
| Logistic Regression | 84.29% |

In here, accuracies were calculated comparing the actual and predicted values. According to the table 2, all the models have good accuracy, but random forest and logistic regression have significant highest accuracies. Among 3 models, Random Forest has the highest accuracy which is 84.84%.

Table 3: Performance of models

| Model | MSE | F1 Score | K-fold average accuracy |
|---|---|---|---|
| | | | |
| Random forest | 0.1506 | 0.245 | 85.16% |
| Decision Tree | 0.2445 | 0.2231 | 84.91% |
| Logistic Regression | 0.1508 | 0.02 | 85.11% |

According to the table 3, it is obvious that when the K-fold cross validation is done, accuracies of all the models are incresed. Among them, decision tree got a significant increase of accuracy which is 84.91%. After done the cross validation, still Random forest has the highest accuracy which is 85.16%. Also it has the minimum mean squared

error(MSE) among others which is 0.1506. Since the accuracy is high and MSE is low, it implies that this model is an accurate model. Also, F1 score implies that this model has achieved a relatively good level of performance in terms of both precision and recall. With a score of 0.245, the model has demonstrated good precision and recall performance. It implies that a suitable trade-off between limiting false positives (precision) and minimizing false negatives (recall) has been accomplished by the model.

Therefore, random forest model is used to select the important features for loan default.

| | Feature | Importance |
|---|---|---|
| 2 | int_rate | 0.124635 |
| 8 | dti | 0.112114 |
| 12 | revol_bal | 0.111225 |
| 13 | revol_util | 0.109797 |
| 5 | annual_inc | 0.107823 |
| 0 | loan_amnt | 0.091788 |
| 14 | total_acc | 0.085510 |
| 10 | open_acc | 0.068500 |
| 3 | emp_length | 0.058134 |
| 7 | purpose | 0.045565 |
| 6 | verification_status | 0.024552 |
| 4 | home_ownership | 0.021146 |
| 1 | term | 0.015893 |
| 9 | delinq_2yrs | 0.014678 |
| 11 | pub_rec | 0.008639 |

Figure 7: Feature Importance

Figure 7 and Figure 8 shows the important features that affect to the loan default. According to those figures, interest rate and dti have highest importance while delinq_2_years and pub_rec have least importance for predicting loan default.

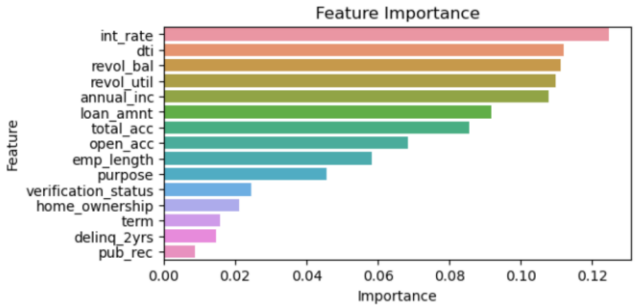


Figure 8: Feature Importance Graph

## V. CONCLUSION

In this study, we were able to create several models that accurately predicted loan default. Among them random forest model is selected as the best performing model. The model performed successfully, obtaining high levels of accuracy and F1 score. Therefore, it is obvious that a random forest model can accurately predict the chance of loan default based on a collection of loan details. Also, those scores implies that this model performed well.

We identified important variables, including int_rate, dti, revol_bal, revol_util, annual_inc that have a substantial impact on the prediction of loan default by feature importance analysis.

The findings of this study have important implications for the bank industry and day to day life. Accurate loan default prediction can be beneficial for the financial industry's stability, profitability, and responsible lending practices. It benefits both lenders and borrowers while also contributing to the overall health of the economy. Therefore, developing and using robust predictive models for assessing credit risk is a good practice for the financial sector.

More improvements can be discovered, such as adding new features, collecting larger and more varied datasets, and utilizing cutting-edge approaches like hyperparameter tuning or ensemble methods to increase the model's performance .

Overall, this study contributes to the growing body of knowledge with regard to the use of machine learning in the financial industry.

## VI. REFERENCES


[1] Gross, Jacob P. K.; Cekic, Osman; Hossler, Don; Hillman, Nick, "What Matters in Student Loan Default: A Review of the Research Literature," 2009. [Online]. Available: https://eric.ed.gov/?id=EJ905712.

[2] J. D. a, "Bank loan losses-given-default: A case study," *ScienceDirect,* vol. 30, no. 4, pp. Pages 1219-1243, 2006.

[3] L. Zhou, "Loan Default Prediction on Large Imbalanced Data Using Random Forests," *ResearchGate,* vol. 10, no. 6, pp. 1519-1525, 2012.

[4] C. Jiang, "Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending," *Springer Link,* p. 511–529, 2017.

[5] L. Zhang, "What should lenders be more concerned about? Developing a profit-driven loan default prediction model," *Elsevier,* vol. 213, 2023.

[6] Xi'an, "Loan Default Prediction with Machine Learning Techniques," *International Conference on Computer Communication and Network Security (CCNS),* 2020.

[7] H. I. T. Aziz, "Loan Default Prediction Model Using Sample, Explore, Modify, Model, and Assess

(SEMMA)," *Journal of Computational and Theoretical Nanoscience,* vol. 8, no. 16, pp. 3489-3503, 2019.

[8] [Online]. Available: https://rameshbanjade.medium.com/exploratory-analysis-and-visualization-of-loan-dataset-153e11859ed8.

.