NATIONAL ECONOMICS UNIVERSITY

Faculty of Mathematical Economics

Analyzing Credit Risk and Predicting Loan Default by using Machine Learning Models

Credit Risk Report

Nguyễn Hoàng Long

Student ID: 11202352

Risk Analysis DSEB 62, 2023

Instructor: Ms.Nguyễn Thị Liên

Table of Contents

ABSTRACT

With the rise of big data and advanced modeling techniques, lenders are increasingly turning to machine learning algorithms to make more accurate predictions of credit risk. The report discusses the use of machine learning models in credit risk analysis and loan default prediction. It examines various machine learning models, including Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine, to predict loan default based on borrower characteristics, financial histories, and economic indicators. The study uses a public dataset about clients and their creditworthiness for data collection, exploratory data analysis, data preprocessing and evaluates the models' performance based on accuracy, precision, recall and F1 score. The report concludes that the random forest model is the most effective in predicting loan default, and the use of machine learning models can help lenders make more informed decisions and reduce financial losses.

I. INTRODUCTION

Credit risk is an ever-present concern in the world of finance. Lenders face the risk that borrowers may default on their loans, resulting in financial losses for the lender. To mitigate this risk, lenders typically employ credit risk analysis to assess the creditworthiness of borrowers before extending loans. Credit risk analysis involves evaluating the borrower's credit history, financial stability, and other relevant factors to determine the likelihood of default. With the rise of big data and machine learning, credit risk analysis has evolved, and lenders are increasingly turning to advanced modeling techniques to make more accurate predictions of credit risk.

There are various techniques and machine learning models that lenders and analysts use to predict credit risk. Traditional techniques include credit scoring models, which use statistical analysis to assess creditworthiness based on historical data. Machine learning models, on the other hand, utilize algorithms that can learn from data and improve over time. These models can analyze vast amounts of data, including borrower characteristics, financial histories, and economic indicators, to predict credit risk. Some common machine learning techniques used in credit risk analysis include decision trees, neural networks, and logistic regression. These techniques have shown promise in improving the accuracy of credit risk prediction and can help lenders make informed decisions about extending credit.

In this report, some machine learning models are used such as Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine. The aim of this report is to investigate the use of machine learning techniques in loan prediction which specify in default prediction. The subject of estimating credit defaulters is considered to employ a basic predictive analytics procedure. Data collection, exploratory data analysis, data preprocessing and choosing the best model for prediction. The results are all portions of the suggested models. This report assembles data from a public dataset about clients and their creditworthiness. To create solid outcomes, the machine learning model is prepared by training on the collected data and used to find the best model to predict loan default and then give some discussions and recommendations.

II. THEORETICAL BACKGROUND

2.1. Theoretical Review

The objective of this section is to examine the literature that covers the conventional methods of evaluating credit risk, as well as the increasing use of machine learning algorithms for credit risk assessment and relevance analysis

What is Machine Learning? Machine Learning is a field that utilizes mathematical principles and powerful algorithms, which are combined with statistical analysis, to forecast or improve upon previous outcomes. It is also capable of updating these outcomes with high accuracy, even when new data is introduced. Banks and financial institutions employ machine learning algorithms to detect patterns and make informed decisions regarding credit card fraud and loan default predictions. This has made the process more efficient and precise. Machine learning is composed of several distinct forms, including supervised, semi-supervised, unsupervised, and reinforcement.

2.2. Models in Machine learning

2.2.1. Logistic Regression

Logistic regression is a statistical technique employed for examining binary or categorical outcomes. It constitutes a category of regression analysis that gauges the correlation between a dependent variable (binary or categorical) and one or more independent variables (categorical or continuous) by approximating the probability of an event of interest. The logistic regression model applies a logistic function to convert the output of a linear regression model into a probability ranging from 0 to 1. The formula for logistic regression can be expressed as follows:

$$p = \frac{1}{1+e^{-z}} \text{, where}$$

- p is the probability of the dependent variable taking the value 1
- e is the base of the natural logarithm
- z is the linear combination of the independent variables

The formula for z is: $z = b_0 + b_1x_1 + b_2x_2 + \cdots + b_2x_2$
where:

- b0 is the intercept
- b1, b2, ..., bn are the coefficients for the independent variables x1, x2, ..., xn

The logistic regression model's coefficients denote the impact of each independent variable on the probability of the dependent variable being assigned a value of 1. These coefficients are derivable through maximum likelihood estimation or alternative optimization algorithms. Upon obtaining the coefficients, the logistic regression model becomes operational for making predictions on fresh data, wherein the independent variables' values are inserted, and the predicted probability is evaluated through the logistic function.

2.2.2. Decision Tree (DT)

A decision tree (DT) is a hierarchical depiction of data attributes and instances. It is employed for both regression and classification tasks. A DT is constructed in a tree structure with a starting point (root node) and decisions made at decision nodes. The most significant benefit of using a DT is its minimal computational time once it is created. However, modifying the root or decision nodes may result in various DTs, and the process of determining the sequence of these nodes is the primary drawback of a DT.

To construct DTs for classification, the Gini and entropy algorithms are commonly used, while the mean squared error (MSE) is frequently utilized for regression problems.

To calculate the Gini index, I will first calculate the Gini index, the Gini index calculated at each node.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

- C is the number of classes to be classified,
- pi=niN
- ni is the number of elements in the ith class
- N is the total number of elements in that node

After calculating the gini index at the parent node and the 2 children are calculated, we can calculate the gini index:

$$gini\_index = gini(p) - \sum_{i=1}^{K} \frac{m_k}{M} gini(c_k)$$

- gini(p) is the gini index of the parent node
- K is the number of split child nodes
- gini(ck) is the number of gini at the kth child node
- M is the number of elements at node p
- mi is the number of elements in the i-th child node

2.2.3. Random Forest

Random forest (RF) is a type of ensemble method based on decision trees that creates multiple decision trees during training and optimizes the results by considering the average regression of each individual tree. Random forest is widely used in both regression and classification tasks.

2.2.4. Support Vector Machine (SVM)

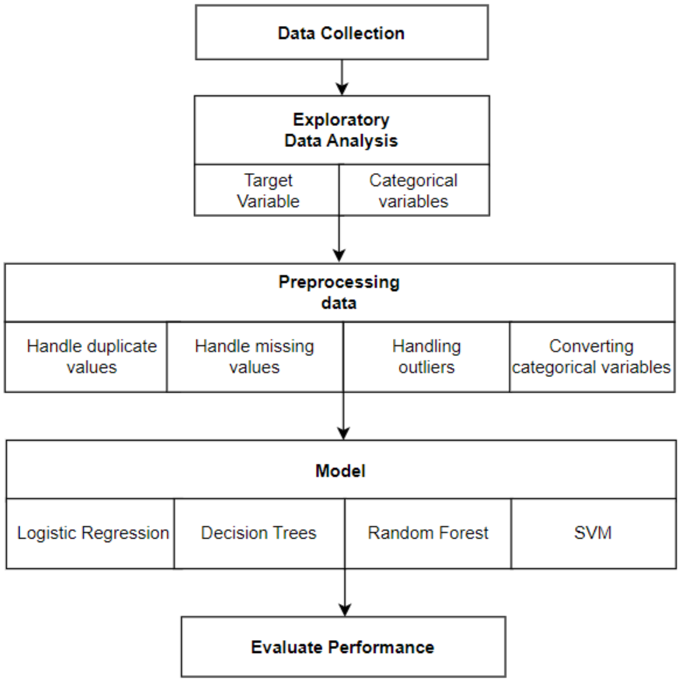
SVM is a machine learning algorithm used for classification and regression analysis. The goal of SVM is to find a hyperplane that can best separate the data into different classes. The hyperplane is chosen in such a way that it maximizes the margin between the classes, which is the distance between the hyperplane and the nearest data points from each class.

SVM works by mapping the data points into a high-dimensional feature space and finding the hyperplane that best separates the classes in this space. The algorithm uses a kernel function to compute the inner products of the data points in this high-dimensional space, without actually having to explicitly compute the coordinates of each data point in that space.

2.3. Methodology

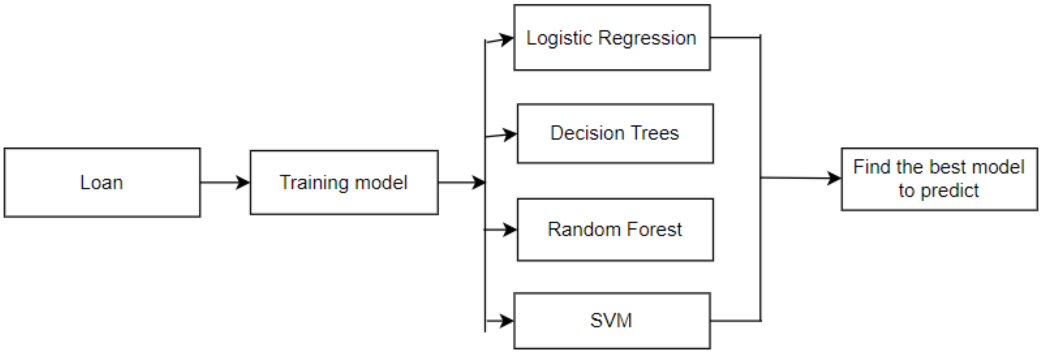
2.3.1. Research Designs

The difficulty of picking the important variables in classification is the focus of this research effort. The categorization purpose is to predict if a certain borrower is likely to default on their loan based on the information gathered throughout the loan application process. Data collection, exploratory data analysis, data preprocessing, and model development using decision trees, logistics regression, random forest, and SVM techniques are all part of the design. Observing and evaluating the models' performance.



2.3.2. Proposed Model

The program will be able to forecast if a loan application would default on a particular loan. The following diagram depicts the system architecture.



There are three steps in here:

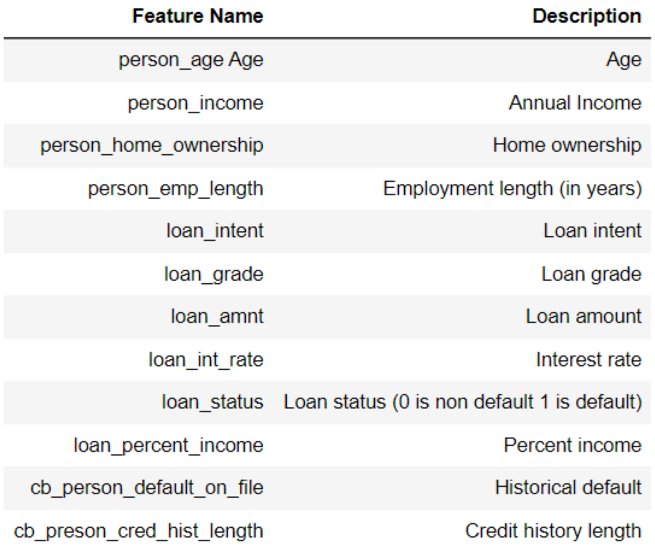
- The loan application is processed by the trained model, which employs three classification algorithms.
- The machine learning with the best performance in accuracy is selected
- The machine learning algorithm is applied to the loan application.

III. DATA

3.1. Data Collection

The dataset contains 32581 entries rows and 12 columns, which includes dealing with missing values and outliers, which entails utilizing relevant characteristics and removing duplicate values of less important variables. Here is an overview of a dataset. The data is a public dataset named Credit Risk Dataset and it is provided by Kaggle. The information was about individuals and their creditworthiness. The dataset contains information on various attributes such as age, income, education, employment status, loan amount, loan purpose, loan status, and credit history of individuals. Extraction of key features, treatment of missing values, and handling of outliers are all part of the data preparation process.

Data description:



Feature Name	Description
person_age Age	Age
person_income	Annual Income
person_home_ownership	Home ownership
person_emp_length	Employment length (in years)
loan_intent	Loan intent
loan_grade	Loan grade
loan_amnt	Loan amount
loan_int_rate	Interest rate
loan_status	Loan status (0 is non default 1 is default)
loan_percent_income	Percent income
cb_person_default_on_file	Historical default
cb_preson_cred_hist_length	Credit history length

3.2. Exploratory Data Analysis

3.2.1. Target Variable

About the target variable: It is the "loan_status" feature in Dataset, which is about loan status of people (0 is non default, 1 is default)

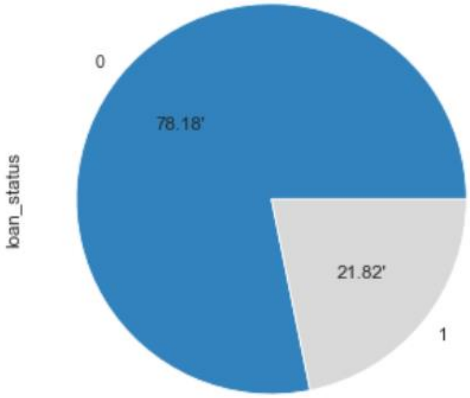
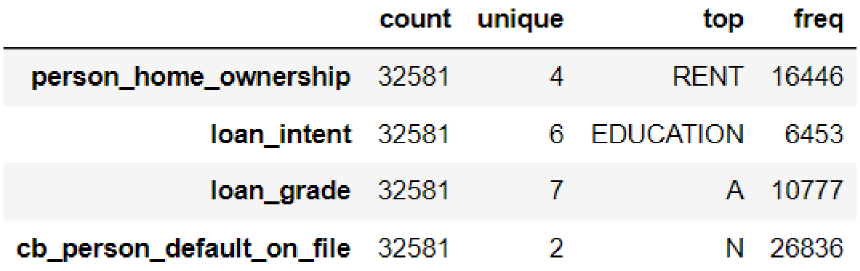


Figure 1: Loan status of people

We can see that 21.82% of the loans are default and 78.18% of loans are non-default, it shows that the number of people who can not pay the loan is low.

3.2.2. Categorical variables

In this process, we will explore some valuable insights from categorical features:



	count	unique	top	freq
person_home_ownership	32581	4	RENT	16446
loan_intent	32581	6	EDUCATION	6453
loan_grade	32581	7	A	10777
cb_person_default_on_file	32581	2	N	26836

Figure 2: Insights of some categorical features

Observations:

- Home ownership: The number of people who use loan living in RENT houses is the highest
- Loan intent: The number of people who intend to use the loan for Education is the highest
- Loan grade: The number of people has a grade A is the highest
- Historical default: The number of people who have history to be non-default is higher than the one for default

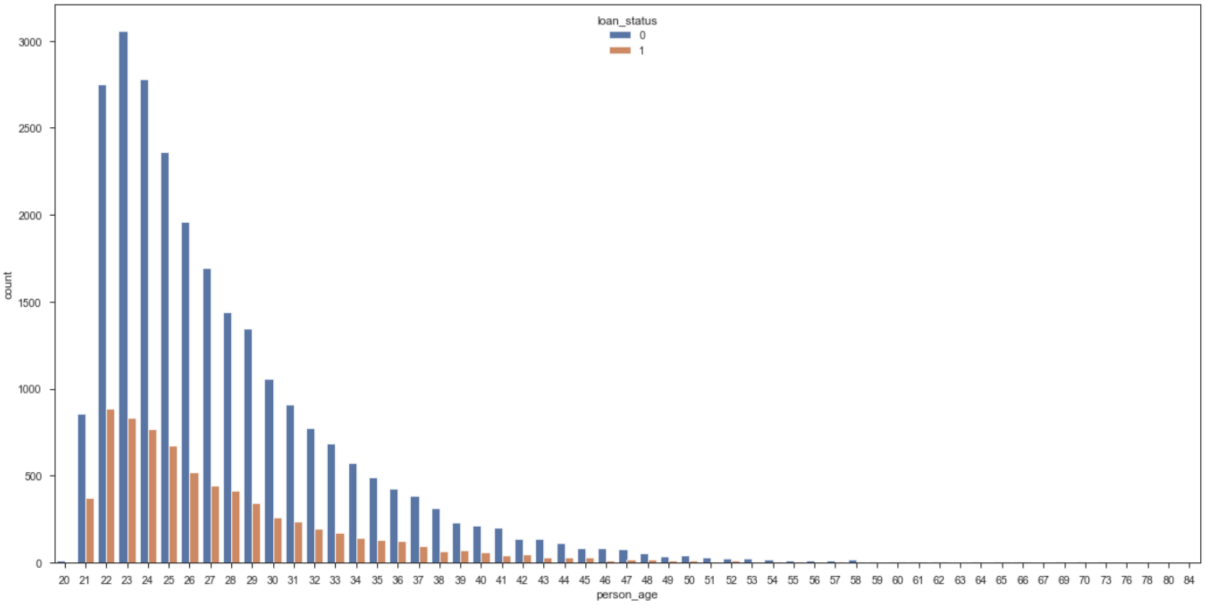


Figure 3: Loan status according to age

Observation: We can observe that people who are younger have a tendency not to pay the loan, 0 paid and 1 did not. The greatest default is among the youngest.

In the graph of “loan_intent” we can analyze the people who are in debt, what were the reasons for the loan.

Debtor (loan_intent)

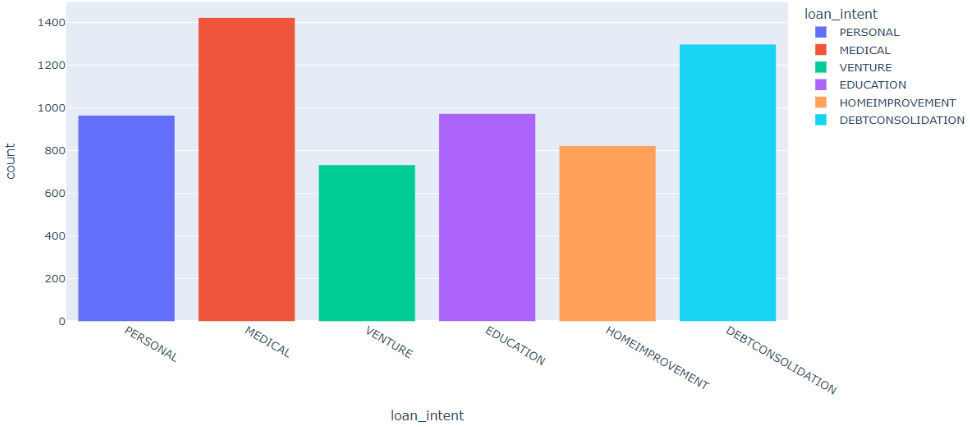


Figure 4: Loan intent of default people

An interesting fact is that the people most likely to default on the loan are the youngest, and the biggest expense on loans is for medical expenses, at least that's what they declare to the financial institution. One of the reasons may be that many do not have health insurance and, in an emergency, end up borrowing money. Debt consolidation is a unification of all debts of an individual. In general, it occurs when it is more advantageous to take out a loan to pay off debts – all of which you owe –, transforming a tangle of bills into a single installment to be paid off.

In the following graph of ‘loan_intent’, we see that in **Figure 5**, those who pay the loan used the amount to pay the student loan, the education factor ends up being an interesting factor for the payment.

No Debtor (loan_intent)

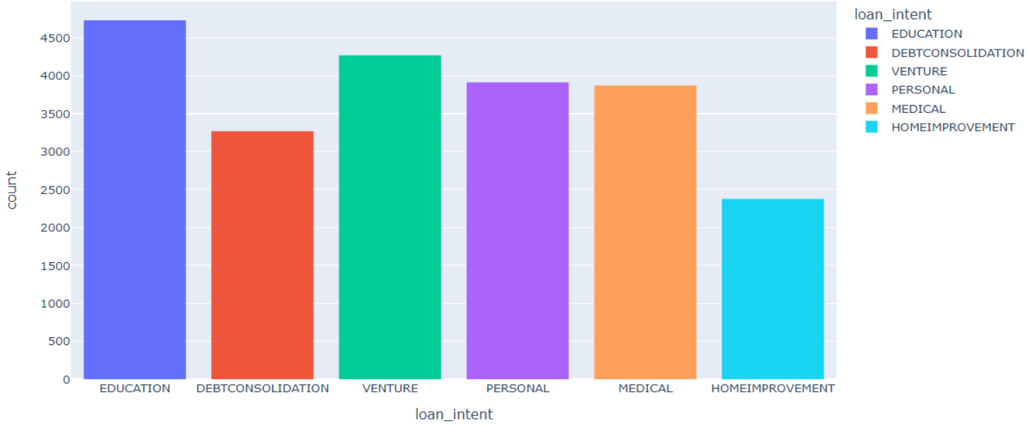


Figure 5: Loan intent of non-default people

Debtor (person_home_ownership)

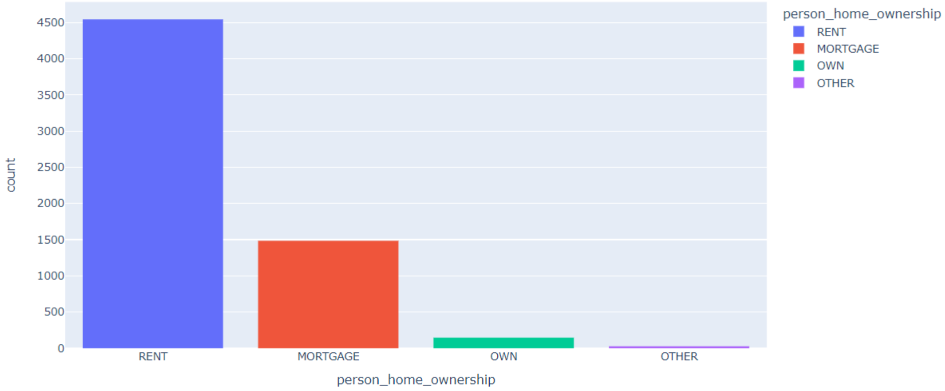


Figure 6: Home ownership of default people

No Debtor (person_home_ownership)

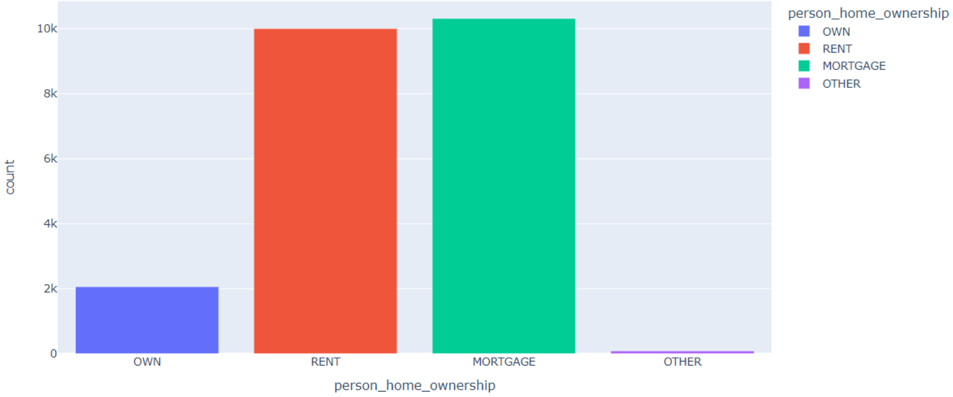


Figure 7: Home ownership of non-default people

We can see from **Figure 6** and **Figure 7** that people who pay rent on their homes are both those who don't pay the loan and those who pay the loan. In addition, the number is much higher in those who pay the loan when related to mortgage.

3.3. Preprocessing data

3.3.1. Handle duplicate values

This step will check the number of duplicate values in dataset. After this step, it shows that there are 165 duplicate values, and then we will drop all these values from the dataset

3.3.2. Handle missing values

Data cleaning is done in this preprocessing stage by looking for and removing any missing values because they have an impact on the model's accuracy. This is accomplished by either using a mean or mode function to fill in the missing values.
We can observe that:

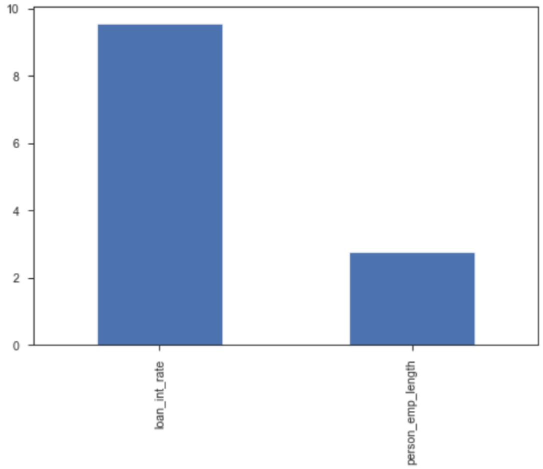


Figure 8: Missing values of two columns “loan_int_rate” and “person_emp_length”

It can be seen that only two columns of data contains NaN and “person_emp_length” column contains 2.75% NaN and loan_int_rate contains 9.56% NaN

After that, it is decided that “person_emp_length” is the person employment history, to be more conservative, the nan values are replaced with mode, which is 0 year and “loan_int_rate” is the loan income rate, to be more conservative, the nan values are replaced with 10.99, which is the median.

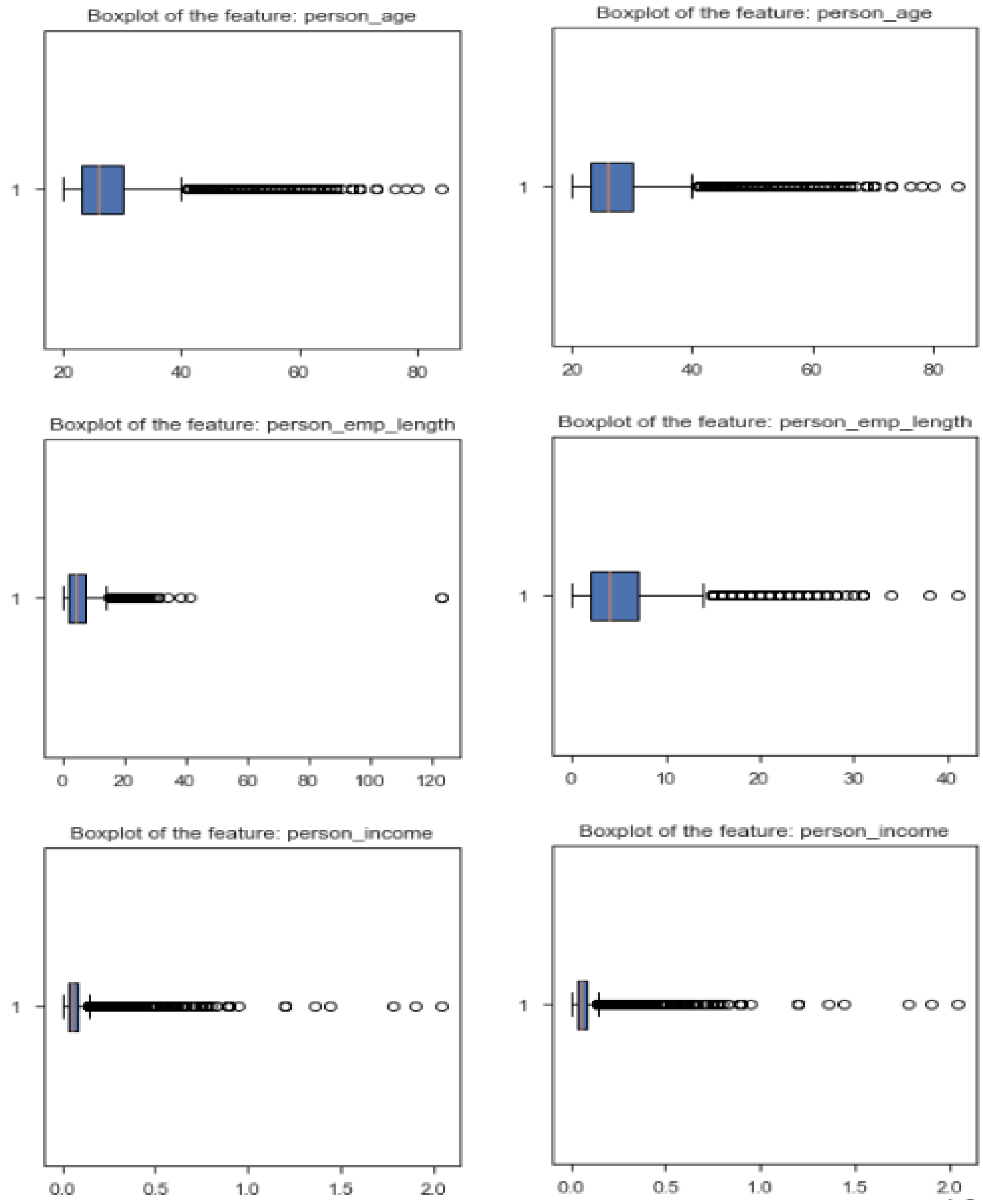
3.3.3. Handle outliers

Outliers are data points that differ significantly from other data points in a dataset, often due to measurement errors or other experimental inaccuracies. Machine learning algorithms are sensitive to the range and distribution of attribute values and may be adversely affected by the presence of outliers. Outliers can result in longer training times, fewer accurate models, and poorer results. Data visualization can be used to detect outliers in a dataset.

There are four possible methods to handle outliers. One way is to remove the outlier records entirely. Another approach is to limit the range of data values to exclude outliers. A third option is to assign a new value if the outlier is not representative of the intended variable. Outliers can have a significant impact on a dataset, but they may also contain valuable information that needs to be incorporated as inputs to train a model. Therefore, only the most extreme and noticeable outliers are removed from the dataset during this process.

Most outliers are retained to train the model. The boxplot and histogram plot indicate that many numeric variables are skewed and contain outliers. However, only extremely noticeable outliers from certain columns are manually removed from the dataset using a specific threshold.

- person_age: Most people are 20 to 60 years old. In the following analysis, to be more general, people age > 100 will be dropped.
- person_emp_length: Most people have less than 40 years of employment. People with employment > 50 years will be dropped.
- person_income: It seems that there are outliers which have to be removed (> 3 million).
- For all other variables, the distribution is more uniform across the whole range, thus they will be kept.



Before dropping outliers

After dropping outliers

Figure 9: Outliers of three columns 'person_age', 'person_income', 'person_emp_length'

We can observe in the graph that outliers of 'person_age' and 'person_income' did not have a dramatic change, while that of 'person_emp_length' decreased significantly.

3.3.4. Converting categorical variables

Before this step, the dataset will be divided between predictors and class. One variable we will use just to store the predictor attributes that will be person_age, loan_amnt, and person_income, and the other variable just to store the class. We drop the target value “loan_status” from the dataset

There are two different ways to convert categorical variables: : Label Encoder and One hot Encoder. In this dataset, we use Label Encoding with the following binary variable ‘cb_person_default_on_file', and apply one-hot encoding to the following categorical variables: 'person_home_ownership', 'loan_intent', 'loan_grade'. For example, ‘cb_person_default_on_file' has only two variables (Y, N), then it will be converted into (1, 0).

3.4. Models

3.4.1. Splitting data

The purpose of splitting the data is to evaluate the performance of a model on new, unseen data. Typically, the dataset is split into two subsets: the training set and the testing set. The model is trained on the training set and then evaluated on the testing set. Before training our models, we split data into Training set which was 70% of the whole dataset and Test set which was the remaining 30%.

3.4.2. Model to train

In this process, we used five different algorithms for our modeling purpose. Logistic Regression, Decision tree, Random Forest, SVM. After that, we check which model has the best performance and then make predictions predict a loan applicant is likely to default or not.

3.5. Performance metrics

In order to evaluate the predictions made by the model, it was necessary to use a range of performance indicators. Specifically, we were trying to predict loan defaults for individuals. In this context, relying solely on model accuracy may not be sufficient to determine its effectiveness. Instead, other metrics such as F1 score, precision score, recall, and confusion matrix should also be taken into account. The key is to select the appropriate performance metrics based on the specific circumstances and requirements of the task at hand.

3.5.1. Accuracy

It is the percentage ratio of the number of correct predictions to the total number of samples. It is a metric that measures the model's ability to classify data points correctly.

$$accuracy = \frac{true\ positives + true\ negatives}{true\ positives + true\ negatives + false\ negatives + false\ positives}$$

3.5.2. Precision

It is the ratio of the number of true positives to the total number of predicted positives. Precision measures how many of the predicted positive cases are actually positive.

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

3.5.3. Recall

It is the ratio of the number of true positives to the total number of actual positives. Recall measures how many of the actual positive cases are correctly predicted as positive.

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

Precision and recall are closely linked, as seen from their definitions.

3.5.4. F1-Score

It is the harmonic mean of precision and recall, and provides a balance between the two metrics. It is a metric that combines both precision and recall into a single score, providing an overall measure of the model's performance.

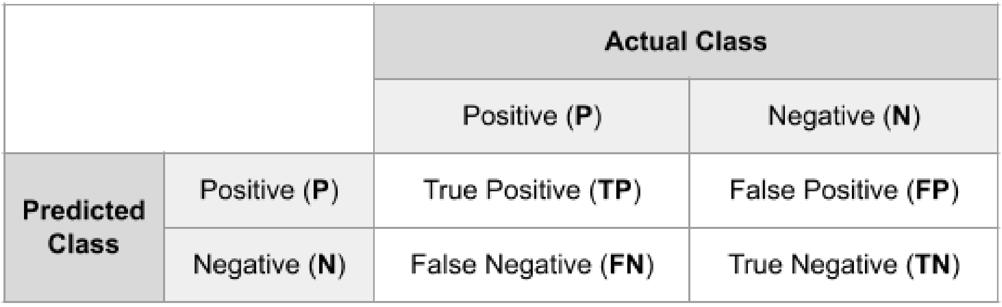
$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

3.5.5. Confusion matrix

A confusion matrix is a tabular representation that summarizes the classification model's performance in predicting examples from different classes. The confusion matrix consists of two axes: the predicted label axis and the actual label axis. It provides a summary of how accurately the model is able to predict the labels for each class.

Now, to fully understand the confusion matrix for this binary class classification problem, we first need to get familiar with the following terms:

- True Positives (TP): Number of samples *correctly* predicted as “positive.”
- True Negatives (TN): Number of samples *correctly* predicted as “negative.”
- False Positives (FP): Number of samples *wrongly* predicted as “positive.”
- False Negatives (FN): Number of samples *wrongly* predicted as “negative.”



		Actual Class	
		Positive (**P**)	Negative (**N**)
Predicted Class	Positive (**P**)	True Positive (**TP**)	False Positive (**FP**)
	Negative (**N**)	False Negative (**FN**)	True Negative (**TN**)

IV. RESULTS AND DISCUSSION

4.1. Results and discussion

The results of models have shown on the table and from this evidence, it is said that Random Forest gives the best results for all metrics with Accuracy, Precision, Recall, and F1-score is approximately 0.93, and it is also the best model to predict loan default, then the classification algorithm is able to identify a loan as default and non-default. The next model ranks second is Decision Tree after tuning parameters with results are a bit lower than Random Forest, followed by SVM and Logistic Regression.

	accuracy	**precision**	**recall**	**f1-score**
Logistic Regression	0.81	0.79	0.81	0.75
Decision Trees	0.89	0.89	0. 89	0.89
Random Forest	0.93	0.93	0.93	0.93
SVM	0.81	0.80	0.81	0.77

Figure 10: Table of results of all metrics from all models

The results of choosing the Random Forest model as the best model for predicting loan default indicate that this model is highly effective in analyzing credit risk and identifying borrowers who are likely to default on their loans. This model uses a combination of decision trees and bootstrapping techniques to analyze large datasets and capture complex relationships between variables, which makes it well-suited for predicting credit risk. The high accuracy, precision, recall, and F1-score of the Random Forest model suggest that it can help lenders make informed decisions about whether to extend credit to borrowers or not. This model can be used to identify high-risk borrowers and adjust lending policies accordingly, which can ultimately help lenders reduce their losses from defaulted loans and improve their overall profitability.

However, it is important to note that no model is perfect, and there is always a risk of errors and false positives in any credit risk analysis. It is also important to consider other factors beyond the model's predictions, such as the borrower's character and intentions, the economic conditions, and the lender's risk appetite, when making lending decisions. Therefore, while the Random Forest model may be an effective tool for predicting loan default, it should be used in conjunction with other risk management practices and expert judgement to make informed lending decisions.

Now, we can check how this model perform on the test set. We plot a confusion matrix to visualize the results for all models (0 is non-default, 1 is default).

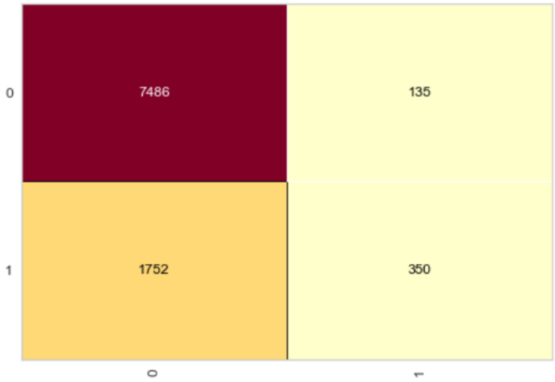
Here's what our model predicts:

True positive: Cases that we predict to be non-default are actually non-default

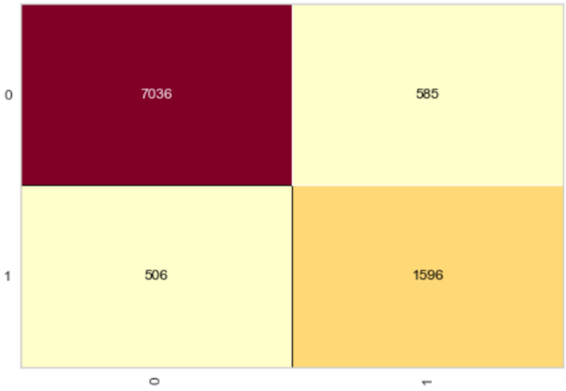
True negative: Cases that we predict to be default are actually default

False Positive: Cases that we predict to be non-default but they are actually default

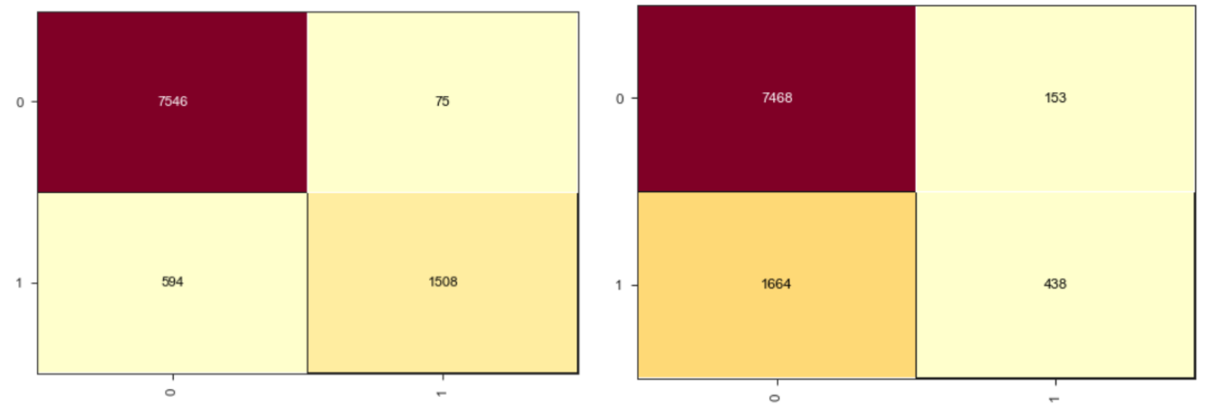
False Negative: Cases that we predict to be default but they are actually non-default



Logistic Regression



Decision Tree



Random Forest SVM

Figure 11: Confusion matrices of all models

As a result, the confusion matrix is calculated based on metrics such as accuracy, precision, recall, and F1 score.

The efficiency of the Random Forest model for loan default prediction can be concluded as high, especially when its False Positive rate is the lowest among the models, with a value of 75. This indicates that the model is better at correctly identifying the borrowers who are likely to default on their loans, while minimizing the chances of rejecting loan applications from creditworthy borrowers. False positives can result in missed lending opportunities and revenue loss for the lender, while false negatives can lead to increased losses from defaulted loans.

With the help of the Random Forest model, lenders can analyze large datasets and identify patterns and relationships between variables that are predictive of loan defaults. This model can help lenders make more informed decisions when assessing creditworthiness and risk, leading to better lending practices and improved profitability.

However, it is important to keep in mind that the effectiveness of any predictive model can be affected by various factors, such as the quality and quantity of data, the model's assumptions and limitations, and changes in the economic and regulatory environment. Therefore, lenders should continue to monitor the performance of the Random Forest model over time and consider other risk management practices and expert judgement when making lending decisions.

4.2. Conclusion:

In conclusion, credit risk analysis is a critical process for lenders, and the use of advanced machine learning models can significantly improve the accuracy of credit risk prediction. This report investigated four machine learning models - logistic regression, decision tree, random forest, and SVM - to determine which one is best suited for predicting loan default.

After careful analysis and evaluation of the models' performance, we found that the random forest model outperformed the other models, with the highest accuracy, precision, recall and F1 score. This indicates that random forest is an effective model for predicting loan default, thanks to its ability to handle large datasets and complex relationships between variables.

It is important to note that the choice of model depends on the specific context and the goals of the lender. However, the results of this report demonstrate the potential of machine learning models in credit risk analysis and their ability to help lenders make more informed decisions.

In summary, machine learning models have revolutionized credit risk analysis, and their use can lead to more accurate predictions of loan default. The findings of this report suggest that the random forest model is a suitable choice for lenders who seek to improve their credit risk analysis processes. Moreover, there are also some recommendations that lenders should care about other factors.

V. APPENDIX

The link provides access to the code file used to back up the findings of this study: https://github.com/Hoanglong14902/Risk-management.git

VI. REFERENCES

1. Shekhar, Shashank. “A Machine Learning Approach to Credit Risk Assessment,” Towards Data Science, March 9, 2021, doi: 10.1016/j.jbankfin.2019.01.022.
2. Silwal, Durga. “Confusion Matrix, Accuracy, Precision, Recall, F1 Score: Measures Explained,” LinkedIn, August 12, 2021. https://www.linkedin.com/pulse/confusion-matrix-accuracy-precision-recall-f1-score-measures-silwal?trk=pulse-article_more-articles_related-content-card.
3. Nguyen, Hoang-Anh, et al. “A Novel Approach for Credit Scoring Using Machine Learning Techniques and Credit Bureau Data,” Applied Sciences, vol. 11, no. 22, November 2021, doi: 10.3390/app112210907.
4. V7Labs. “F1 Score: What It Is and How to Calculate It,” V7Labs Blog, June 29, 2021. https://www.v7labs.com/blog/f1-score-guide.
5. Nguyen, Hieu. “Khái niệm về Confusion Matrix trong Machine Learning,” Math2IT, January 15, 2022. https://math2it.com/hieu-confusion-matrix/.