4/16/2022

RISK ANALYSIS REPORT

Credit risk prediction

Author name: Nguyen Thuy Linh

Table of contents

Part 1: Theoretical background

1.1. Introduction to credit risk

1.1.1 What is credit?

In the financial world, credit has many definitions, but it is generally defined as a contract agreement in which a borrower receives a sum of money or something of value and repays the lender at a later date, usually with interest. Credit can also refer to an individual's or a company's creditworthiness or credit history.

1.1.2 Credit risk

Credit risk is a measure of a borrower's creditworthiness. Lenders calculate credit risk by estimating the likelihood that they will recover all of their principal and interest when making a loan. Borrowers with low credit risk are charged lower interest rates. Rating agencies are used by lenders, investors, and other counterparties to assess the credit risk of doing business with a company.

1.1.3. Type of credit risk

1.1.3.1. Good risk

A good risk is an investment that one believes will be profitable. The term most commonly refers to a loan made to a creditworthy individual or business. Good risks are thought to be extremely likely to be repaid.

1.1.3.2. Bad risk

A bad risk loan is one that is unlikely to be repaid due to a poor credit history, insufficient income, or other factors. A bad risk increases the lender's risk and the likelihood of default on the borrower's part.

2.2. Personal credit

2.2.1. Personal credit definition

Personal credit often referred to as consumer credit is the debt taken by an individual to buy goods and services. Personal credit can be in the form of a credit card or any type of personal loan.

2.2.2. Characteristics of personal credit

Individual customer loans are small-scale loans, but they account for a large proportion of commercial bank loans. With the characteristics of loans to meet consumer needs in life or to provide additional capital for business, the volume of these loans is also limited by factors such as ability to repay, loan purpose, and collateral.

Personal credit is a type of credit that carries a high level of risk for the bank because customer information is difficult to obtain completely and honestly. As a result of this information asymmetry between the bank and individual customers, the appraisal process's reliability is low.

Part 2: Methodology

2.1. Introduction to machine learning

Mehryar Mohri & his colleagues (2018) introduced the concept of machine learning as computational methods that use experience to improve performance or to make accurate predictions. In this context, experience refers to previously available information to learners, which is typically in the form of electronic data collected and made available for analysis. This data can take the form of human-labeled digitized training sets or other types of information gleaned from interaction with the environment. In any case, its quality and size are critical factors in the success of the learners' predictions.

2.2. Some Machine Learning methods for the problem

2.2.1 Logistics regression

The logistic regression model is one of the most commonly used methods in most quantitative studies on credit risk classification of individual customers. A logistic regression model is a type of general linear model used to predict the probability classification of binary variables. The logit function is used to estimate the probability of the dependent variable occurring, with its labels replaced by 0 and 1.

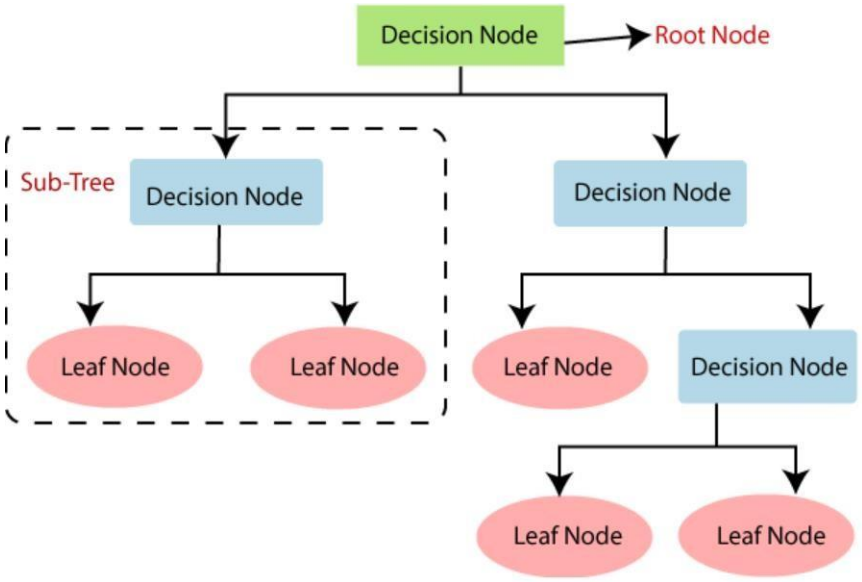
For example, if we want to predict whether a customer has a good or bad credit risk, the dependent variable will be the customer's credit risk classification, which will have two values: good or bad credit risk.

2.2.3. Decision tree

Decision Tree is a Supervised learning technique that can be used to solve classification and regression problems, but it is most commonly used to solve classification problems. It is a tree-structured classifier in which internal nodes represent dataset features, branches represent decision rules, and each leaf node represents the result.

A Decision tree has two nodes: the Decision Node and the Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and have no additional branches.

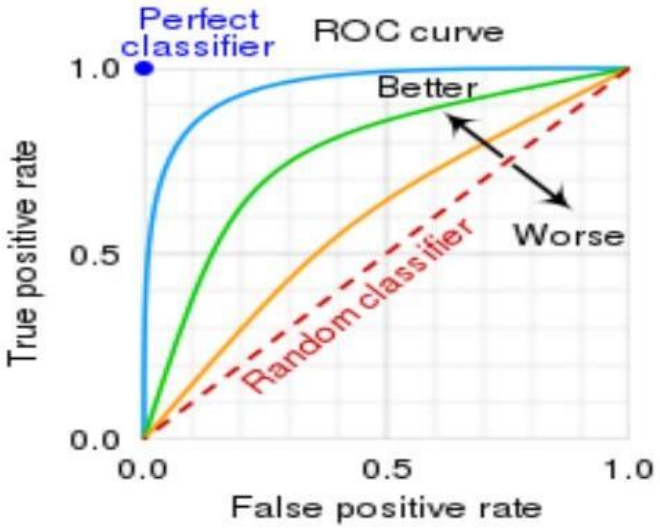
Below diagram explains the general structure of a decision tree:



2.3. Metrics to evaluate model

- AUC-ROC

- AUC - ROC is a method for calculating a classifier model's performance against various classification thresholds. Assume we have a binary classification problem with two classes. Choosing different classification thresholds [0..1] will affect the model's classification ability, and we need to calculate the model's influence. AUC is an abbreviation for Area Under The Curve, and ROC is an abbreviation for Receiver Operating Characteristics. ROC is a probability curve, and AUC is the model's classification level. AUC-ROC is also abbreviated as AUROC (Area Under The Receiver Operating Characteristics).
- The following is an interpretation of the term AUROC: It is the likelihood that a randomly chosen positive sample will outperform a randomly chosen negative sample.
- The formula: AUC = P(score(x+) > score(x-))
- The higher the AUC, the more accurate the model is in classifying classes.
- The ROC curve depicts pairs of indices (TPR, FPR) at each threshold, with TPR acting as the continuous axis and FPR acting as the horizontal axis.



Part 3: Data

3.1. Data description

The dataset pertains to the credit risk of individual customers of a German bank.It contains 1000 entries with 10 categorial/symbolic attributes prepared by Prof. Hofmann. Each entry in this dataset represents a person who obtains credit from a bank. Each individual is classified as a good or bad credit risk based on a set of attributes.

- Declare data and explain the meaning of features

Features	Type	Meaning of variable	Number of Unique Values	% Null	Example
Age	Numeric	Age of customer	53	0	67, 22, 34
Sex	Categorical	Sex of customer	2	0	Male, Female
Job	Categorical	Type of job	4	0	-Unskilled and non-resident -Unskilled and resident - Skilled -Highly-

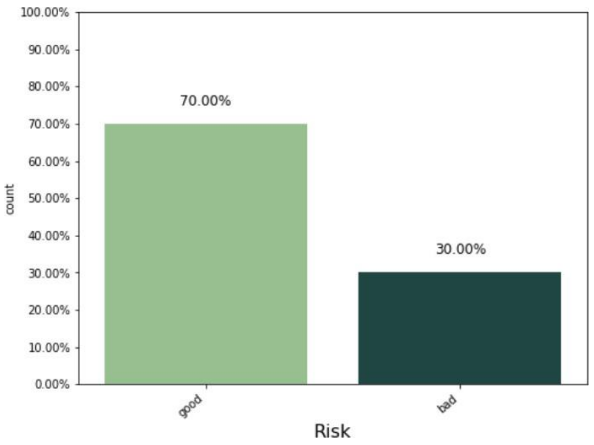
					skilled
Housing	Categorical	Type of housing	3	0	- Own - Rent - Free
Saving accounts	Numeric	Type of saving account	4	18.3	- Little - Moderate - Quite rich - Rich
Checking account	Numeric	Type of checking account	3	39.4	- Little - Moderate - Rich
Credit amount	Numeric	The amount of customer's credit	921	0	1169, 5951, 2096
Duration	Numeric		33	0	6, 48, 12
Purpose	Categorical	The purpose of customer credit	8	0	- Car - Radio/TV …..
Risk	Categorical	Type of customer's credit	2	0	Good/Bad

3.2. Exploratory data analysis

In this data analysis, we will look at the behavior of German borrowers. Questions such as, "Why do German borrowers apply for credit loans?" How many jobs does each borrower have? What patterns (if any) determine whether the loan is good or bad risk? Of course, our in-depth analysis of German credit borrowers will answer many more questions. To make our visualizations more interactive, we will use plotly, an interactive library that will provide us with a better understanding of our data.

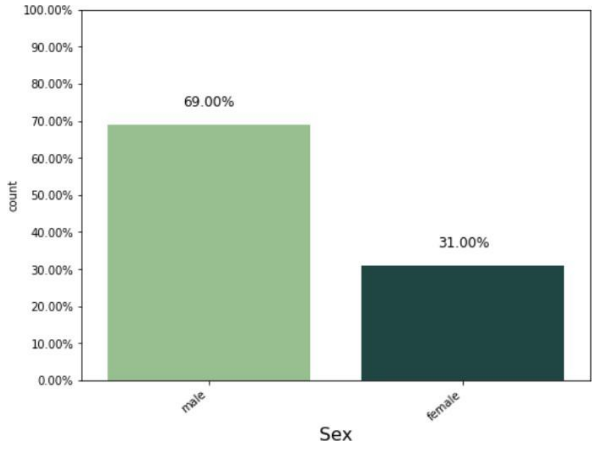
- The general Good/Bad rate in data

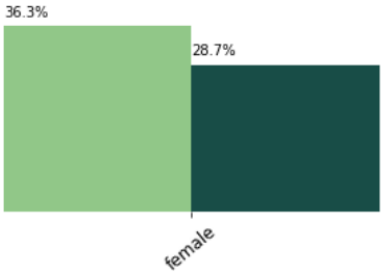
Credit risk is classified as good for up to 70% of customers and bad for 30% of customers. Because Personal loans are quite risky, this is a good sign for the bank. If the number of customers with bad credit is high, the bank will have a difficult time recovering this loan.

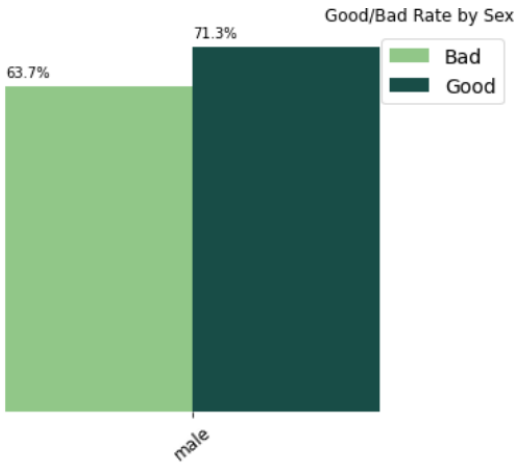


- Good/Bad rate by sex

- In this dataset, we can see that male have a much higher proportion of credit loans than females. It accounts for 69% of the total.
- Males have a lower percentage of bad ratings than females.

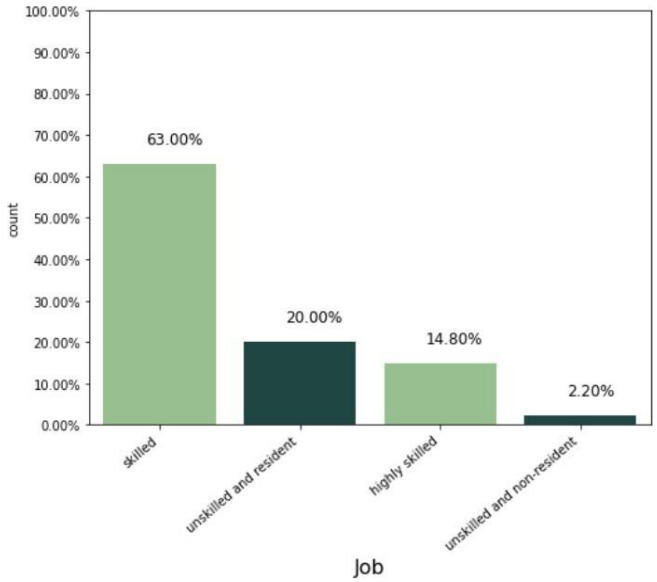




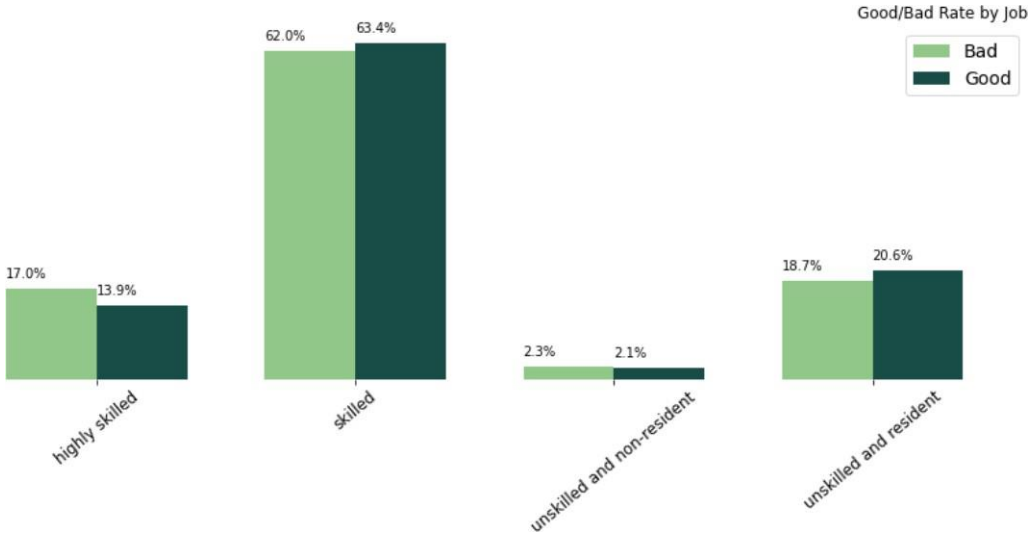


- Good/Bad rate by job
- Occupations were classified into four categories in this data. Skilled - people accounted for the greatest proportion of the population, accounting for 63%. The

next group is of type unskilled and resident , and the last is of type unskilled and non- resident, with a rate of 2.2%.

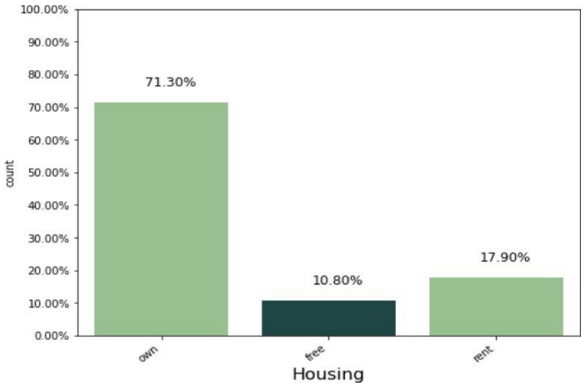


- Looking at the good/bad credit risk ratio chart by job classification, we can see that most groups have a higher good credit risk ratio than a bad credit risk ratio. However, the only group with a higher ratio of bad credit to good credit is highly skilled workers. =>Job does not affect the credit risk rate.

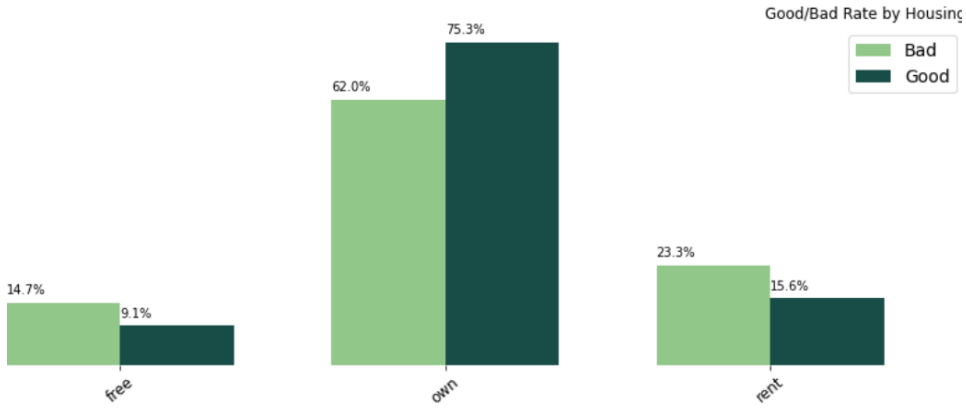


- Good/Bad rate by housing

- This data set contains three types of housing, which are own, rent, and free. Customers who own a home account for the highest proportion, accounting for 71.3%. The lowest is the percentage of customers who stay at home for free; this figure is only 10.8%.

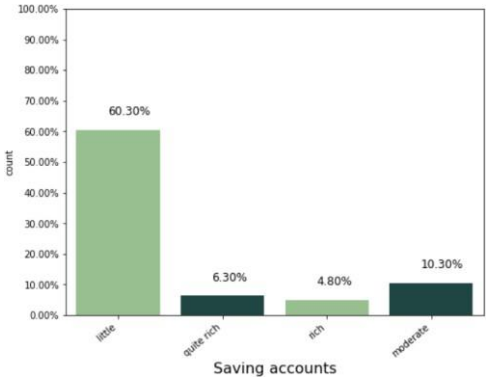


- Homeowners are the only customer group with a higher rate of good credit risk than bad credit risk. The remaining two customer groups have a good/bad ratio greater than one. This could be due to the fact that the group of these customers has mỏe expenses to pay such as housing service fees and housing rent and they pay bank loans late.

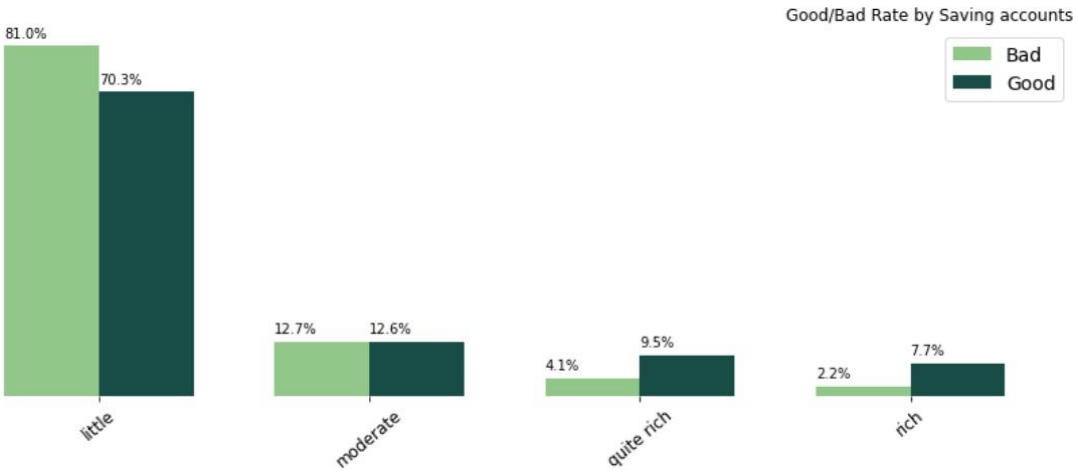


- Saving account

- There are for type of saving account:
+ Little
+ Moderate
+ Quite rich
+ Rich

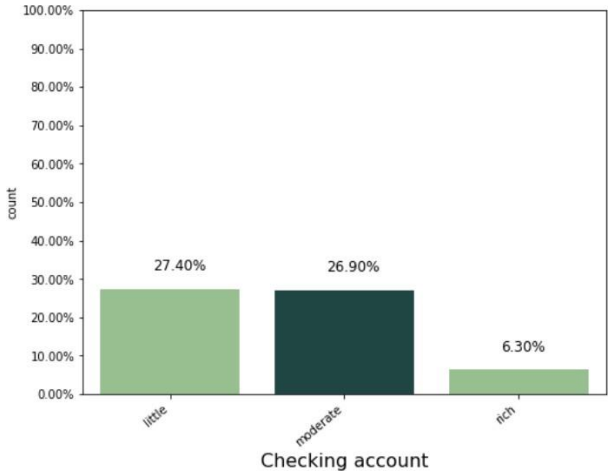


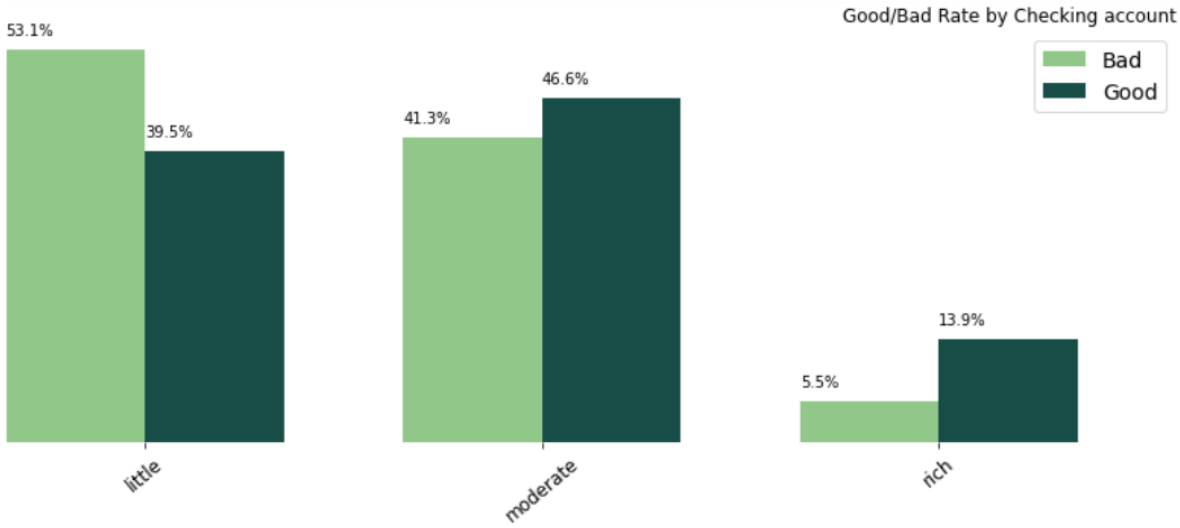
- Most people are in the records have little savings (not rich). People who have rich savings are the lowest on record only accounting for 4.8%.
- In the below bar chart, we can see that the person with more savings means less risk to the bank.



- Checking account

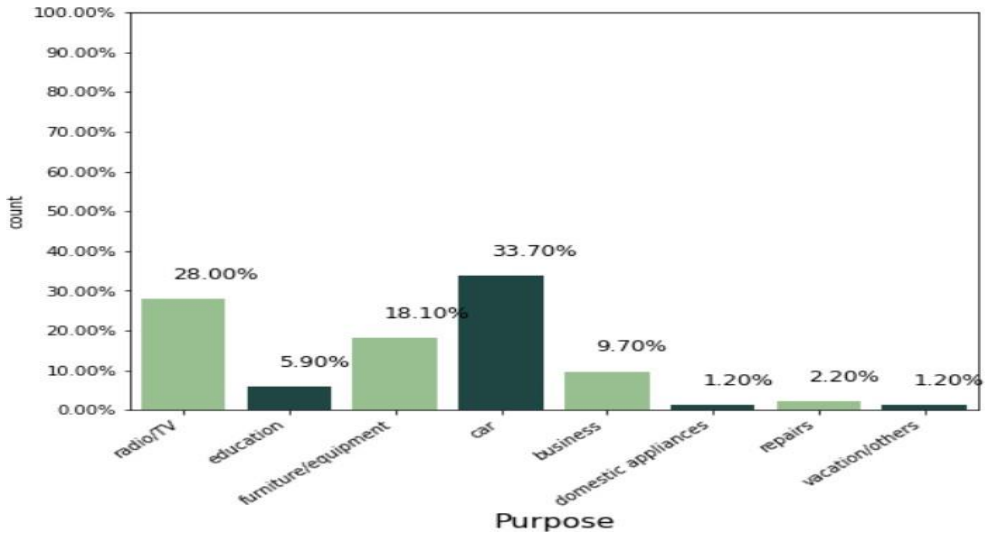
- We divide checking accounts into 3 types.
+ Rich
+ Moderate
+ Little
- The scale is not 100% because the data is missing.
- About half of people who have little checking account are considered as having a bad rating in risk.



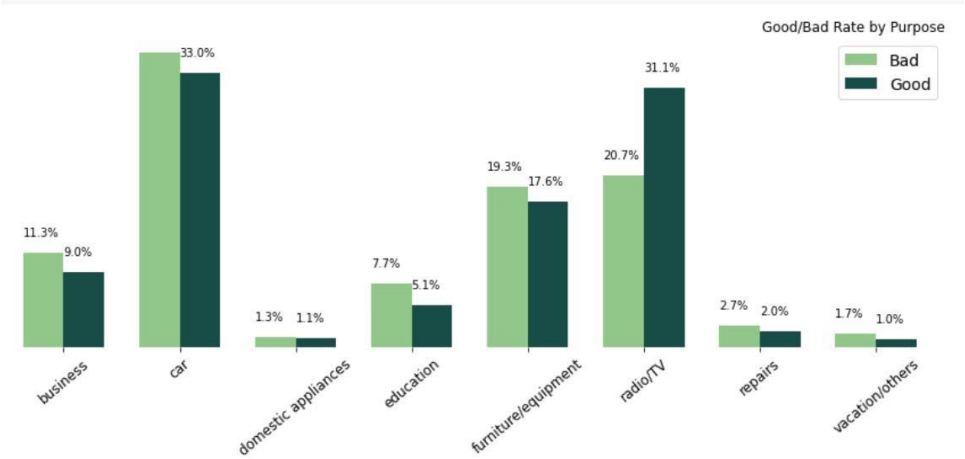


- Purpose

  This chart shows that most people use credit loan for buying car, radio and TV and furniture/equipment.

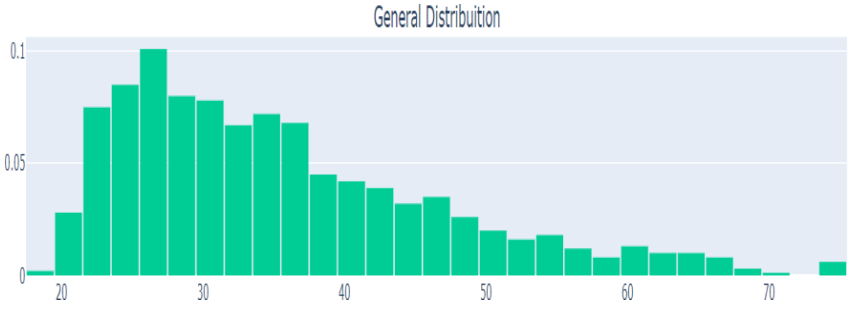


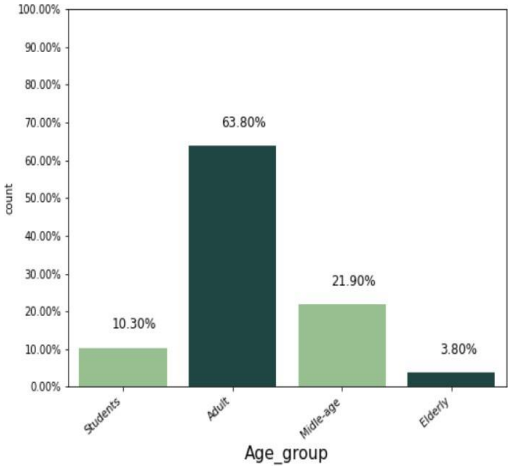
- The majority of these credit loans have a higher bad credit risk to good credit risk ratio. Credit loans for the purpose of purchasing ratio/TV have a higher good-to-bad ratio. This can be explained by the fact that these items are inexpensive to purchase, so customers can easily return them to the bank.



- Age

- In this dataset, customers with bank loans are aged between 19 and 75 years old



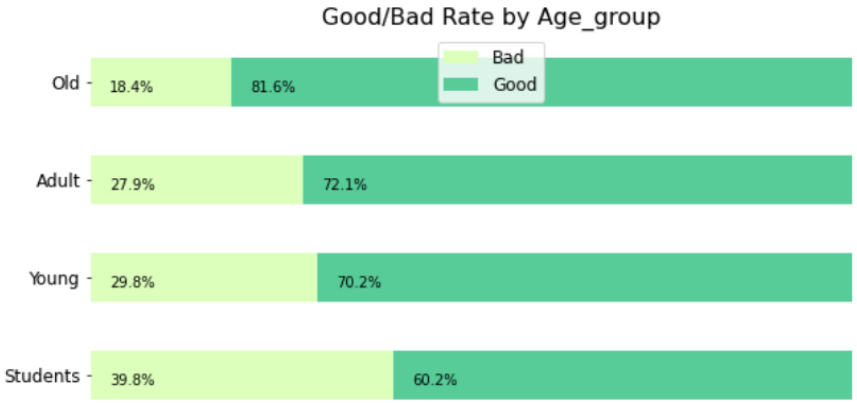


- We proceed to divide this age into 4 large groups in accordance with reality. It includes:

+ Student group: 19-22
+ Adult group: 23-40
+ Middle-aged group: 41-60
+ Elderly group: 61-75

- With more than 30%, the adult age group has the highest credit rate. The adult age group comes next, and the elderly age group comes last. This distinction can be explained by two factors:

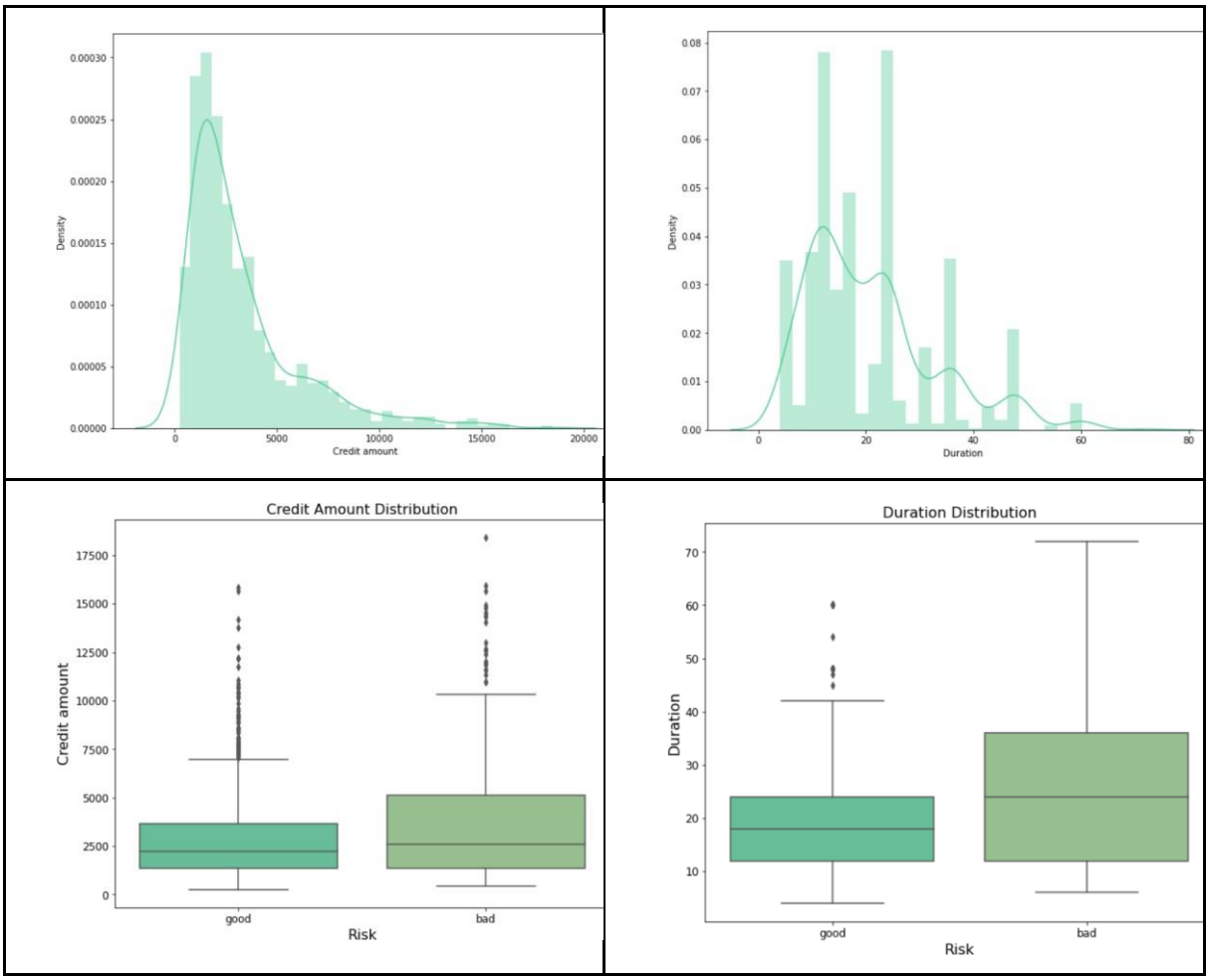
+ To begin with, the adult and middle-aged groups: they are still of working age, have children, have many expenses to spend, and must care for their family, so they use credit more frequently.
+ Second, the elderly have less credit because they are retired, have a salary to spend, and do not have much consumption demand at this age.



- About Good/Bad rate by age. We can see that students have the highest bad credit rate. Due to not having a full-time job, they can’t pay the credit loan on time.

● Credit amount and Duration

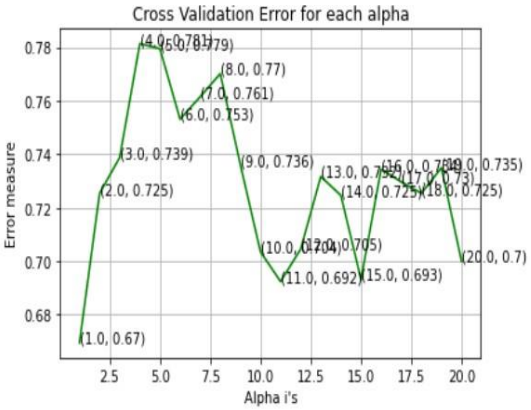
- The higher credit amount and longer duration means higher risk to the bank.



3.3. Data processing

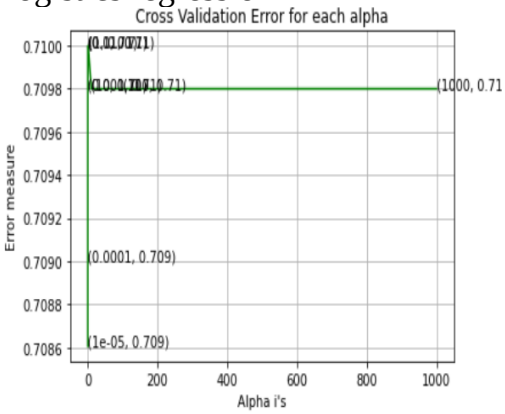
- Check for null in data
- Fill nan by no_inf
- To fit the model for the data, we proceed to convert the categorical variables to numeric using the OneHotEncoder method. The processed features will include 29 columns instead of 9 columns as before
- Scale the data using the Data scaling method
- Split data into Train, Test, Cross-validation sets
- Fit model

Decision tree




```
For values of best alpha =  4.0 The train AUC is: 0.840
For values of best alpha =  4.0 The cross validation AU
For values of best alpha =  4.0 The test AUC is: 0.7609
```

Logistics regression




```
For values of best alpha =  0.01 The train AUC is: 0.7806
For values of best alpha =  0.01 The cross validation AUC
For values of best alpha =  0.01 The test AUC is: 0.79021
```

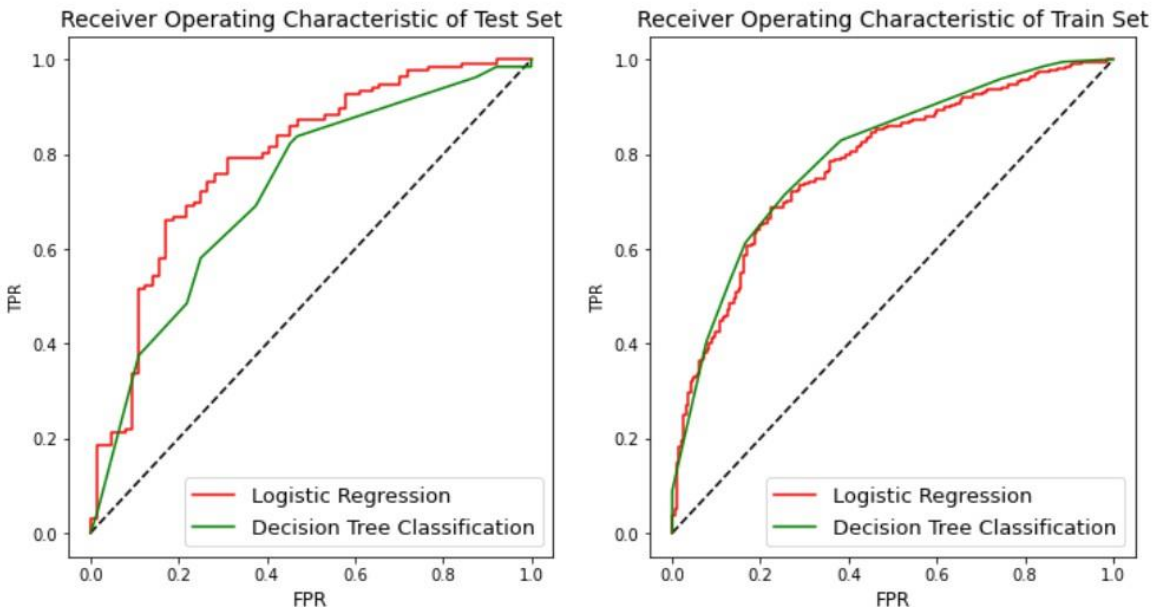
3.4. Modeling

- Evaluation of the accuracy of two logistic models and decision tree

- AUC score

```
roc_auc_score for DecisionTree of test set:  0.723920036764706
roc_auc_score for Logistic Regression of test set:  0.7889476102941176
```

```
roc_auc_score for DecisionTree of train set:  0.7978405257850262
roc_auc_score for Logistic Regression of train set:  0.7810388195337947
```



On the test set, the AUC SCORE of the Logistic Regression model is approximately 0.78 higher than that of the Decision Tree model, which is only 0.72.

Conclusion: We chose the Logistic Regression model to classify customer credit based on the above result.

Appendix: link code https://www.kaggle.com/datasets/uciml/german-credit